**KATHOLIEKE UNIVERSITEIT LEUVEN**
FACULTEIT INGENIEURSWETENSCHAPPEN
DEPARTEMENT ELEKTROTECHNIEK
Kasteelpark Arenberg 10, 3001 Leuven (Heverlee)

# CLUSTERING OF SCIENTIFIC FIELDS BY INTEGRATING TEXT MINING AND BIBLIOMETRICS

Promotoren:

Prof. dr. ir. B. De Moor
Prof. dr. ir. K. Debackere

Proefschrift voorgedragen tot
het behalen van het doctoraat
in de ingenieurswetenschappen

door

**Frizo JANSSENS**

Mei 2007

**KATHOLIEKE UNIVERSITEIT LEUVEN**
FACULTEIT INGENIEURSWETENSCHAPPEN
DEPARTEMENT ELEKTROTECHNIEK
Kasteelpark Arenberg 10, 3001 Leuven (Heverlee)

# CLUSTERING OF SCIENTIFIC FIELDS BY INTEGRATING TEXT MINING AND BIBLIOMETRICS

Jury:

Prof. dr. ir. Y. Willems, voorzitter
Prof. dr. ir. B. De Moor, promotor
Prof. dr. ir. K. Debackere, co-promotor
Prof. dr. ir. H. Blockeel
Prof. dr. ir. V. Blondel (UCL)
Prof. dr. W. Daelemans (UA)
Prof. dr. W. Glänzel
Prof. dr. M.-F. Moens

Proefschrift voorgedragen tot
het behalen van het doctoraat
in de ingenieurswetenschappen

door

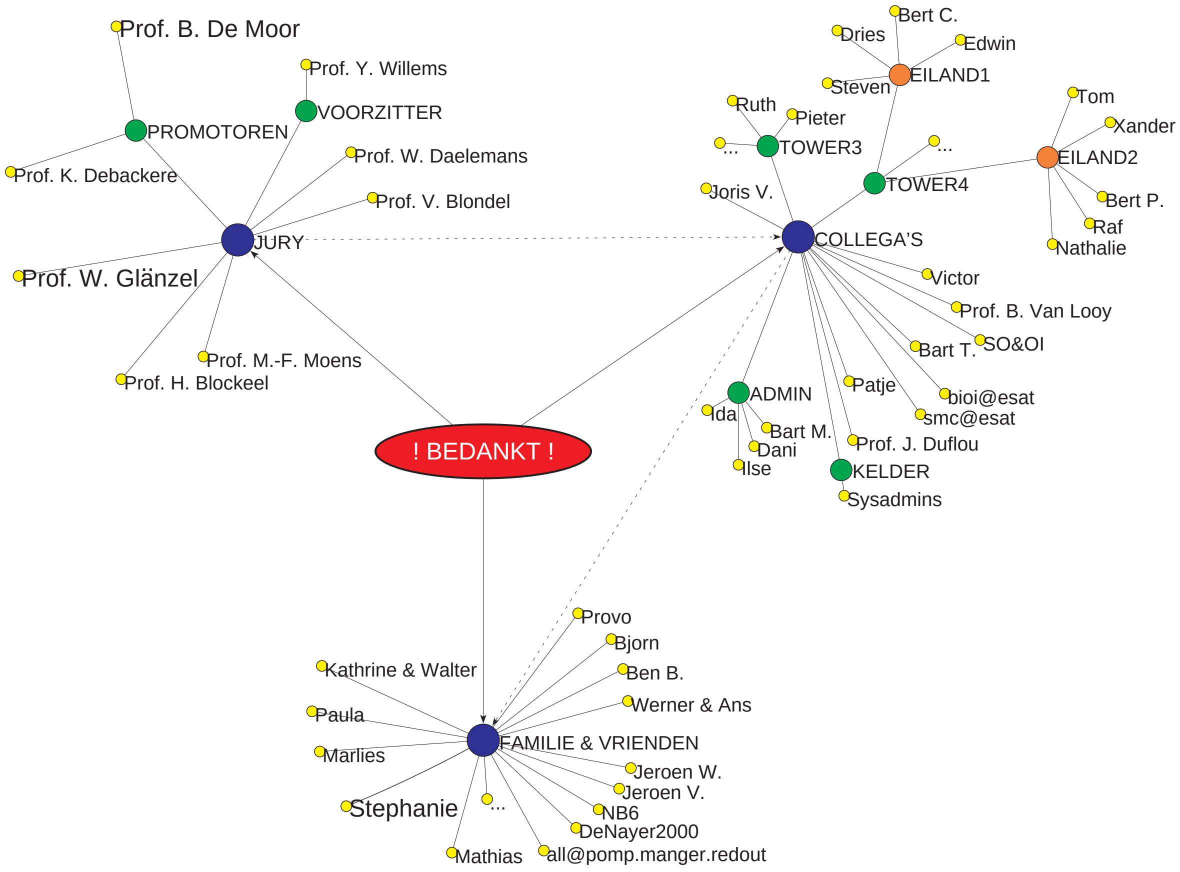
**Frizo JANSSENS**

U.D.C. 681.3*I2

Mei 2007

D/2007/7515/44

ISBN 978-90-5682-810-3


! BEDANKT !
JURY
PROMOTOREN
Prof. B. De Moor
Prof. K. Debackere
VOORZITTER
Prof. Y. Willems
Prof. W. Daelemans
Prof. V. Blondel
Prof. W. Glänzel
Prof. M.-F. Moens
Prof. H. Blockeel
COLLEGA'S
EILAND1
Bert C.
Dries
Edwin
Steven
TOWER3
Ruth
Pieter
...
TOWER4
...
EILAND2
Tom
Xander
Bert P.
Raf
Nathalie
Joris V.
Victor
Prof. B. Van Looy
Bart T.
SO&OI
Patje
bioi@esat
smc@esat
ADMIN
Ida
Bart M.
Dani
Ilse
Prof. J. Duflou
KELDER
Sysadmins
FAMILIE & VRIENDEN
Provo
Bjorn
Ben B.
Werner & Ans
Kathrine & Walter
Paula
Marlies
Stephanie
...
Mathias
all@pomp.manger.redout
Jeroen W.
Jeroen V.
NB6
DeNayer2000

# Abstract

Increasing dissemination of scientific and technological publications via the Internet, and their availability in large-scale bibliographic databases, has led to tremendous opportunities to improve classification and bibliometric cartography of science and technology. This metascience benefits from the continuous rise of computing power and the development of new algorithms. Paramount challenges still remain, however.

This dissertation verifies the hypothesis that accuracy of clustering and classification of scientific fields is enhanced by incorporation of algorithms and techniques from text mining and bibliometrics. Both textual and bibliometric approaches have advantages and intricacies, and both provide different views on the same interlinked corpus of scientific publications or patents. In addition to textual information in such documents, citations between them also constitute huge networks that yield additional information. We incorporate both points of view and show how to improve on existing text-based and bibliometric methods for the mapping of science.

The dissertation is organized into three parts.

Firstly, we discuss the use of text mining techniques for information retrieval and for mapping of knowledge embedded in text. We introduce and demonstrate our text mining framework and the use of agglomerative hierarchical clustering. We also investigate the relationship between the number of Latent Semantic Indexing factors, the number of clusters, and clustering performance. Furthermore, we describe a combined semi-automatic strategy to determine the optimal number of clusters in a document set.

Secondly, we focus on analysis of large networks that emerge from many individual acts of authors citing other scientific works, or collaborating in the same research endeavor. These networks of science and technology can be analyzed with techniques from bibliometrics and graph theory in order to rank important and relevant entities, for clustering or partitioning, and for extraction of communities.

Thirdly, we substantiate the complementarity of text mining and bibliometric methods and we propose schemes for the sound integration of both worlds. The performance of unsupervised clustering and classification significantly improves by deeply merging textual content of scientific publications

with the structure of citation graphs. Best results are obtained by a clustering method based on statistical meta-analysis, which significantly outperforms text-based and citation-based solutions.

Our hybrid strategies for information retrieval and clustering are corroborated by two case studies. The goal of the first is to unravel and visualize the concept structure of the field of library and information science, and to assess the added value of the hybrid approach. The second study is focused on bibliometric properties, cognitive structure and dynamics of the bioinformatics field. We develop a methodology for dynamic hybrid clustering of evolving bibliographic data sets by matching and tracking clusters through time.

To conclude, for the complementary text and graph worlds we devise a hybrid clustering approach that jointly considers both paradigms, and we demonstrate that with an integrated stance we obtain a better interpretation of the structure and evolution of scientific fields.

# Korte inhoud

De toenemende verspreiding van wetenschappelijke en technologische publicaties via het internet, en de beschikbaarheid ervan in grootschalige bibliografische databanken, leiden tot enorme mogelijkheden om de wetenschap en technologie in kaart te brengen. Ook de voortdurende toename van beschikbare rekenkracht en de ontwikkeling van nieuwe algoritmen dragen hiertoe bij. Belangrijke uitdagingen blijven echter bestaan.

Dit proefschrift bevestigt de hypothese dat de nauwkeurigheid van zowel het clusteren van wetenschappelijke kennisgebieden als het classificeren van publicaties nog verbeterd kunnen worden door het integreren van tekstontginning en bibliometrie. Zowel de tekstuele als de bibliometrische benadering hebben voor- en nadelen, en allebei bieden ze een andere kijk op een corpus van wetenschappelijke publicaties of patenten. Enerzijds is er een schat aan tekstinformatie aanwezig in dergelijke documenten, anderzijds vormen de onderlinge citaties grote netwerken die extra informatie leveren. We integreren beide gezichtspunten en tonen hoe bestaande tekstuele en bibliometrische methoden kunnen verbeterd worden.

De dissertatie is opgebouwd uit drie delen.

Ten eerste bespreken we het gebruik van tekstontginningstechnieken voor informatievergaring en voor het in kaart brengen van kennis vervat in teksten. We introduceren en demonstreren het raamwerk voor tekstontginning, evenals het gebruik van agglomeratieve hiërarchische clustering. Voorts onderzoeken we de relatie tussen enerzijds de performantie van het clusteren en anderzijds het gewenste aantal clusters en het aantal factoren bij latent semantische indexering. Daarnaast beschrijven we een samengestelde, semi-automatische strategie om het aantal clusters in een verzameling documenten te bepalen.

Ten tweede behandelen we netwerken die bestaan uit citaties tussen wetenschappelijke documenten, en netwerken die ontstaan uit onderlinge samenwerkingsverbanden tussen auteurs. Dergelijke netwerken kunnen geanalyseerd worden met technieken van de bibliometrie en de grafentheorie, met als doel het rangschikken van relevante entiteiten, het clusteren en het ontdekken van gemeenschappen.

Ten derde tonen we de complementariteit aan van tekstontginning en bibliometrie en stellen we mogelijkheden voor om beide werelden op correcte wijze te integreren. De performantie van ongesuperviseerd clusteren en van classificeren verbetert significant door het samenvoegen van de tekstuele inhoud van wetenschappelijke publicaties en de structuur van citatienetwerken. Een methode gebaseerd op statistische meta-analyse behaalt de beste resultaten en overtreft methoden die enkel gebaseerd zijn op tekst of citaties.

Onze geïntegreerde of hybride strategieën voor informatievergaring en clustering worden gedemonstreerd in twee domeinstudies. Het doel van de eerste studie is het ontrafelen en visualiseren van de conceptstructuur van de informatiewetenschappen en het toetsen van de toegevoegde waarde van de hybride methode. De tweede studie omvat de cognitieve structuur, bibliometrische eigenschappen en de dynamica van bio-informatica. We ontwikkelen een methode voor dynamisch en geïntegreerd clusteren van evoluerende bibliografische corpora. Deze methode vergelijkt en volgt clusters doorheen de tijd.

Samengevat kunnen we stellen dat we voor de complementaire tekst- en netwerkwerelden een hybride clustermethode ontwerpen die tegelijkertijd rekening houdt met beide paradigma's. We tonen eveneens aan dat de geïntegreerde zienswijze een beter begrip oplevert van de structuur en de evolutie van wetenschappelijke kennisgebieden.

# Nederlandse samenvatting

## Clusteren van wetenschappelijke kennisgebieden door integratie van tekstontginning en bibliometrie

### Inleiding

Sinds de aanvang van het informatietijdperk en het toenemende belang van de kenniseconomie is de hoeveelheid digitale informatie enorm gegroeid en dit met steeds grotere snelheid. Reeds enkele jaren geleden werd het aantal online documenten geschat op 550 miljard [17], goed voor een totaal van 7,5 petabyte aan data beschikbaar op websites en in publieke databanken[1]. Dat is vier keer meer dan de ruimte die nodig is om alle informatie van alle Amerikaanse academische bibliotheken digitaal op te slaan [173].

Om 7,5 petabyte aan informatie te kunnen bevatten, zou een stapel documenten met ongeveer 2500 tekens per blad 300 000 km hoog moeten zijn en bijgevolg bijna tot de maan reiken, of 7,5 maal de omtrek van de aarde meten (1 byte per teken en 1 cm voor 100 pagina's). Een persoon die 1 pagina per minuut leest, zou wel 5,7 miljoen jaar nodig hebben om de hele stapel te lezen! Gelukkig komen technieken van informatievergaring, tekstontginning en netwerkanalyse de arme lezer te hulp bij deze sisyfusarbeid.

De verspreiding van wetenschappelijke en technologische publicaties via het internet en in grootschalige bibliografische databanken is gemeengoed geworden. Figuur 0.1 toont de jaarlijkse groei van de MEDLINE[2] databank die informatie bevat over publicaties in onder andere de medische wetenschappen. Een andere belangrijke databank is de *ISI* Web of Science[3] (WoS), waarin alle bibliografische informatie van de 9300 belangrijkste tijdschriften ter wereld opgenomen is. De volledige WoS databank bevat vandaag gegevens over meer dan 36 miljoen artikels en ze groeit met ongeveer 1,1 miljoen records per jaar, afkomstig uit meer dan 230 disciplines.

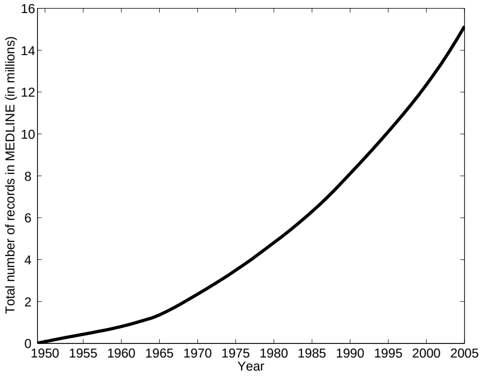
Voor een individu of een bedrijf leidt deze overweldigende hoeveelheid data tot grote moeilijkheden wanneer relevante informatie en kennis gezocht en ver-

[1] Eén petabyte bevat $10^{15}$ bytes.

[2] http://www.pubmed.org, bezocht in januari 2007.

[3] http://scientific.thomson.com/products/wos/, bezocht in januari 2007.

werkt moet worden. Zoekmachines zijn onontbeerlijk maar geven vaak ook een hoop irrelevante resultaten. Wil men meer dan een gewone zoektocht naar interessante documenten, dan dient informatievergaring uitgebreid met andere algoritmen.



**Figuur 0.1:** Groei van MEDLINE, de belangrijkste databank van de U.S. National Library of Medicine (NLM) met voornamelijk wetenschappelijke informatie over geneeskunde. Het totaal aantal wetenschappelijke publicaties in de databank is aangeduid per jaar (in miljoen). Vandaag bevat MEDLINE gegevens over ongeveer 15 miljoen publicaties [49].

## Algemene context

Dit proefschrift handelt over het ***in kaart brengen*** van wetenschappelijke en technologische kennisgebieden met behulp van ***clusteralgoritmen*** en technieken van ***bibliometrie*** en ***tekstontginning***.

***Tekstontginning*** behelst het automatisch en intelligent analyseren van teksten door een computer en heeft als doel het vinden van interessante feiten, relaties en kennis in grote hoeveelheden tekst. Voor dit doel maakt *text mining* gebruik van technieken en algoritmen uit *data mining*, informatievergaring, statistiek, wiskunde, machineleren en computerlinguïstiek.

De ***bibliometrie*** is een interdisciplinaire wetenschap waarbij men gebruik maakt van statistische en wiskundige indicatoren, methoden en modellen voor het bestuderen van geschreven wetenschappelijke communicatie, meestal verzameld in grote databanken met wetenschappelijke publicaties of patenten.

Kennisgebieden worden ***in kaart gebracht*** om de structuur en de evolutie ervan te begrijpen, evenals de relaties met andere domeinen, en dit op basis van publicaties of andere digitale bestanden. Dergelijke documenten bevatten een schat aan informatie en worden beschouwd als indirecte maar ware reflecties van wetenschappelijke kennis en activiteit. Onderzoeksdomeinen kunnen getypeerd worden op basis van belangrijke publicaties en tijdschriften, productieve auteurs, belangrijke concepten, instellingen, landen enz. Voor bedrijven, onderzoeksinstellingen en voor de overheid is kennis over de activiteitsgraad in verschillende domeinen en kennis van nieuwe, opkomende en convergerende gebieden heel belangrijk. Kwantitatieve informatie kan gebruikt worden bij het evalueren van onderzoeksperformantie en als ondersteuning voor het wetenschaps- en technologiebeleid en innovatiemanagement. Een goed beleid is cruciaal wil men de competitieve positie behouden en verbeteren.

***Clusteren*** is een multivariate statistische techniek voor het automatisch indelen van een verzameling objecten in groepen, waarbij elke groep of cluster zo homogeen mogelijk is. De bedoeling is dus dat alle elementen in eenzelfde cluster gelijkaardige kenmerken vertonen, terwijl objecten in verschillende clusters zo veel mogelijk van elkaar verschillen. Het clusteren van documenten heeft bijvoorbeeld tot doel documenten te groeperen die over hetzelfde onderwerp handelen. Eenvoudig gesteld kijkt het algoritme hiervoor naar het aantal gemeenschappelijke woorden.

De belangrijkste hypothese die in dit proefschrift vooropgesteld en geverifieerd wordt, luidt dat de performantie van zowel het clusteren van wetenschappelijke kennisgebieden als het classificeren van publicaties kan verbeterd worden door het integreren van heterogene informatie. Dit betekent dat bibliometrische citatiegegevens geïncorporeerd worden met de wetenschappelijke inhoud van publicaties. De performantie van het clusteren wordt gemeten met behulp van formules die op statistische wijze nagaan hoe 'gelukkig' geclusterde documenten zijn met de toewijzing aan een bepaalde cluster. Met andere woorden: in welke mate is het onderwerp gerelateerd aan dat van andere documenten in dezelfde cluster, en dit in contrast met de mate waarin documenten even goed in een andere cluster zouden kunnen thuishoren. De nauwkeurigheid van classificatie wordt gekwantificeerd door vergelijking met een bestaande 'correcte' of 'gouden standaard' classificatie die gebaseerd is op expertkennis vervat in *Medical Subject Headings* (MeSH[4]), dit zijn termen die geannoteerd zijn aan publicaties.

## Motivatie: tekst- en netwerkwereld

Het onderscheid tussen *tekstwereld* en *netwerk-* of *grafenwereld* verwijst naar de verschillende manieren waarop men een bibliografische databank kan bekijken. Enerzijds is er een schat aan tekstinformatie aanwezig in dergelijke documenten, anderzijds vormen de onderlinge citaties grote netwerken die extra informatie leveren. Zo goed als elke publicatie verwijst namelijk naar eerder gepubliceerde artikels waarop ze gebaseerd is, of naar artikels die op één of andere manier relevant zijn voor het onderwerp. Deze *citaties* staan vermeld in de bibliografie (de lijst van *geciteerde referenties*). Hoewel men andere literatuur om uiteenlopende redenen kan citeren, suggereert een citatie meestal het goedkeuren of aanraden van het voorgaande werk. Citaties tussen publicaties vormen enorme netwerken, net zoals het wereldwijde web bestaat uit *hyperlinks* tussen webpagina's.

[4] http://www.nlm.nih.gov/mesh/, bezocht in Januari 2007.

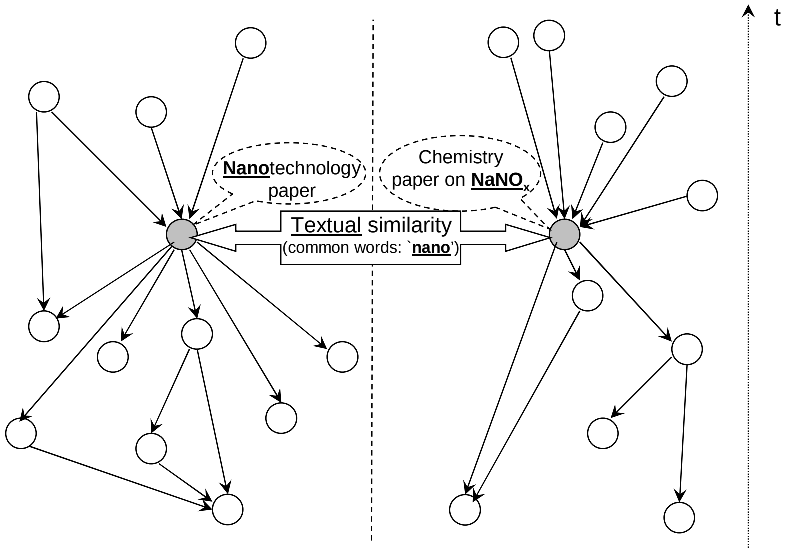
Zowel de tekstuele als de bibliometrische benadering hebben voor- en nadelen, en allebei bieden ze een andere kijk op een corpus van wetenschappelijke publicaties of patenten. Zo bieden beide zienswijzen bijvoorbeeld een verschillende perceptie van de similariteit van documenten of groepen documenten, evenals verschillende methoden voor het observeren van de dynamica van evoluerende databanken. We integreren beide gezichtspunten en tonen hoe bestaande tekstuele en bibliometrische methoden kunnen verbeterd worden bij het in kaart brengen van kennisgebieden.

Tekstuele informatie kan inderdaad overeenkomsten in onderwerp aan het licht brengen die niet zichtbaar zijn voor bibliometrische methoden. Wanneer men enkel tekst beschouwt, kan similariteit echter even goed verborgen blijven door verschillen in woordgebruik. Valse overeenkomsten kunnen eveneens geïntroduceerd worden door voorbewerking van de tekst of door polyseme woorden (met meerdere betekenissen) of woorden met weinig semantische waarde. Zo kunnen documenten over muziekvergaring (*music information retrieval*) verkeerdelijk in verband gebracht worden met patentonderzoek omwille van het voorkomen van gemeenschappelijke woorden die in beide contexten gebruikt worden, zoals *title, record, creative,* en *business.*

Figuur 0.2 toont nog een illustratief voorbeeld. Cirkels stellen wetenschappelijke artikels voor (*nodes* in het citatienetwerk), citaties ertussen worden voorgesteld door pijlen. Hoewel beide artikels in grijze kleur over een verschillend onderwerp handelen (het ene handelt over nanotechnologie en het andere over chemie), kunnen tekstontginningsalgoritmen ze toch verkeerdelijk als gerelateerd aanzien door het regelmatig voorkomen van dezelfde stam *nano* in beide teksten (na voorbewerking). Gelukkig blijkt uit observatie van het citatienetwerk dat beide publicaties zich in andere domeinen bevinden.

Informatievergaring biedt ook voorbeelden waarbij de tekst- en netwerkwerelden complementair zijn en waarbij een gecombineerde benadering een groot voordeel oplevert. Zoekmachines uit de beginjaren van het internet gebruikten enkel de tekstuele inhoud van webpagina's om te bepalen welke daarvan relevant waren voor een bepaalde zoekopdracht. Pas sinds het einde van vorig millennium buiten grootschalige zoekmachines ook de linkstructuur van het web uit. Het bekendste voorbeeld is het PageRank algoritme van *Google*, dat hyperlinks in rekening brengt om de *kwaliteit* van webpagina's te bepalen. Een webpagina waarnaar veel wordt verwezen door andere goede webpagina's is waarschijnlijk een autoriteit op een bepaald gebied en hoort dus op een hoge plaats in de vaak lange lijst met resultaten.

Hybride methoden die zowel de tekst- als de connectie-analyse uitbuiten, worden dus verondersteld tot betere resultaten te leiden dan technieken die louter de tekst of citaties gebruiken. In dit proefschrift demonstreren we de complementaritiet van beide paradigma's. We stellen ook een hybride aanpak voor die deze beide werelden tegelijk bekijkt, en we beweren dat een geïntegreerde benadering leidt tot een beter begrip van de structuur en van de dynamische eigenschappen van grootschalige corpora met wetenschappelijke publicaties of patenten.



**Figuur 0.2:** Illustratie van de motivatie om geïntegreerde (hybride) algoritmen te ontwikkelen. Een klein extract van een citatienetwerk wordt getoond. Cirkels stellen wetenschappelijke publicaties of patenten voor. Citaties ertussen worden voorgesteld door een pijl van de citerende naar de geciteerde publicatie. We bekijken de twee publicaties in grijze kleur, de ene handelt over nanotechnologie en de andere over chemie (natriumnitraat of $NaNO_3$). Automatische tekstontginningsprocedures zouden beide artikels verkeerdelijk kunnen beschouwen als aan elkaar gerelateerd omdat ze allebei vaak dezelfde belangrijke term *nano* bevatten. Door automatische voorbewerkingsmethoden zou de chemische formule $NaNO_x$ herleid kunnen worden tot dezelfde stam 'nano'. Door het bekijken van het volledige citatienetwerk wordt echter duidelijk dat beide publicaties niet gerelateerd zijn aangezien ze zich in verschillende omgevingen of gemeenschappen van het citatienetwerk bevinden. Er zijn geen gemeenschappelijke referenties en geen gemeenschappelijke citerende artikels in beide omgevingen. Een hybride analyse van zowel de *tekstwereld* als de *netwerkwereld* draagt dus bij tot een juistere perceptie van de (dis)similariteit van beide publicaties.

Figuur 0.3 geeft een meer gedetailleerde introductie tot de tekstwereld. Ze bevat een schematisch overzicht van enkele belangrijke stappen uit het raamwerk voor tekstontginning waarbij tekstinformatie voorgesteld wordt in het vectorruimtemodel.

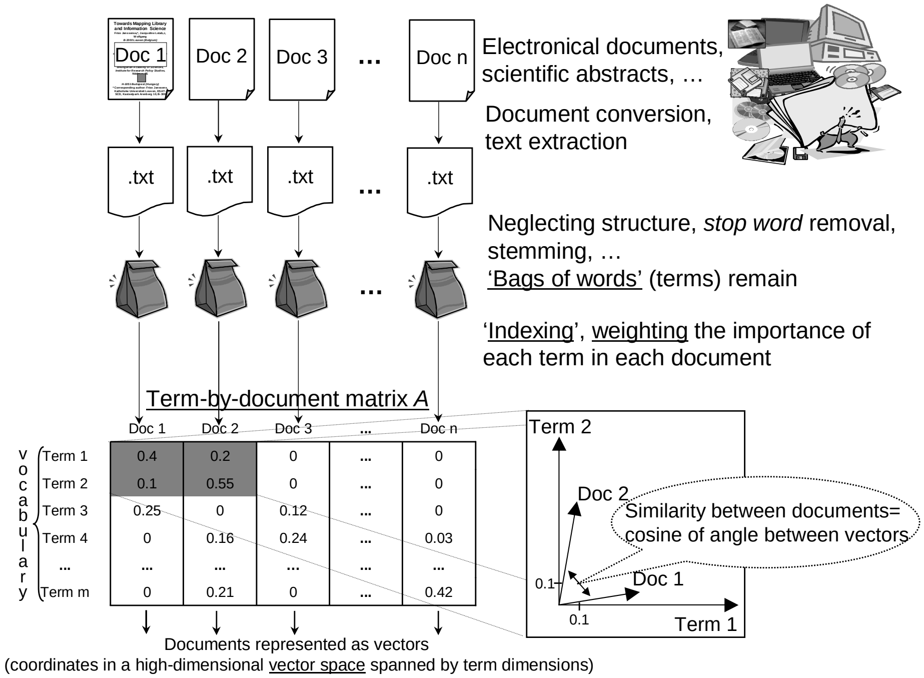
De similariteit van twee documenten, m.a.w. hoe sterk de onderwerpen met elkaar te maken hebben, kan gekwantificeerd worden door de cosinus van de hoek tussen de vectorvoorstellingen van beide documenten. Hoe kleiner deze hoek, en dus hoe groter de cosinus, hoe meer de onderwerpen van beide documenten gerelateerd zijn [9]. Deze cosinussimilariteit of correlatiecoefficient levert een waarde tussen 0 en 1 en wordt als volgt berekend:

$$Sim(\vec{d_1}, \vec{d_2}) = cos(\widehat{\vec{d_1}\vec{d_2}}) = \frac{\vec{d_1} \cdot \vec{d_2}}{\|\vec{d_1}\| \cdot \|\vec{d_2}\|} = \frac{\sum_i w_{i,1} \cdot w_{i,2}}{\sqrt{\sum_i w_{i,1}^2} \cdot \sqrt{\sum_i w_{i,2}^2}}, \tag{0.1}$$

waarbij $d_1$ en $d_2$ twee documenten voorstellen en $w_{i,j}$ het gewicht van term $t_i$ in document $d_j$. De *afstand* tussen beide documenten verkrijgt men door het complement $(1-)$ van de cosinus te nemen.

Naast de toenemende beschikbaarheid van elektronische documenten wordt onze wereld ook gekenmerkt door een steeds hogere mate van onderlinge verbondenheid in vele verschillende soorten netwerken. Er bestaan uiteraard enorme infrastructurele netwerken voor transport van o.a. goederen, personen en elektriciteit, maar evenzeer zijn informatie- en communicatienetwerken van groot belang in onze maatschappij. De groei van het internet en van draadloze netwerken is opmerkelijk. Daarnaast participeren wij als sociale wezens in verschillende vormen van *sociale netwerken*. Netwerken kunnen ook opgebouwd zijn uit communicatieverrichtingen, zoals telefoongesprekken en e-mailberichten, uit kennis (bv. *Wikipedia*), of uit verschillende vormen van biologische en biochemische interacties (bv. neurale systemen of netwerken van proteïne-interacties).

Technieken van de bibliometrie en de grafentheorie kan men gebruiken om netwerken te analyseren die bestaan uit citaties tussen wetenschappelijke documenten, of netwerken die ontstaan uit onderlinge samenwerkingsverbanden. Het doel van dergelijke analyses kan bijvoorbeeld het rangschikken van relevante entiteiten zijn, of het clusteren en ontdekken van gemeenschappen. De wetenschap van evoluerende netwerken kan zelfs bijdragen tot het detecteren van opkomende trends en convergerende wetenschappelijke specialiteiten, alsook van nieuwe technologieën en *hot topics*. Er is reeds veel onderzoek verricht naar de statistische en dynamische eigenschappen van grootschalige netwerken [250, 227, 6, 71, 198, 200, 35, 178]. Algoritmen voor netwerkanalyse worden gebruikt in data-ontginning, patroonherkenning, trenddetectie, strategische positionering, fraudedetectie, analyse van financiële netwerken, epidemiologisch onderzoek, maar ook door inlichtingendiensten enz.


Doc 1
Doc 2
Doc 3
…
Doc n
Electronical documents, scientific abstracts, …
Document conversion, text extraction
.txt
Neglecting structure, stop word removal, stemming, …
'Bags of words' (terms) remain
'Indexing', weighting the importance of each term in each document
Term-by-document matrix A
vocabulary
Term 2
Term 1
Doc 2
Doc 1
0.1
0.1
Similarity between documents= cosine of angle between vectors
Documents represented as vectors
(coordinates in a high-dimensional vector space spanned by term dimensions)

| | Doc 1 | Doc 2 | Doc 3 | ... | Doc n |
|---|---|---|---|---|---|
| Term 1 | 0.4 | 0.2 | 0 | ... | 0 |
| Term 2 | 0.1 | 0.55 | 0 | ... | 0 |
| Term 3 | 0.25 | 0 | 0.12 | ... | 0 |
| Term 4 | 0 | 0.16 | 0.24 | ... | 0.03 |
| ... | ... | ... | ... | ... | ... |
| Term m | 0 | 0.21 | 0 | ... | 0.42 |

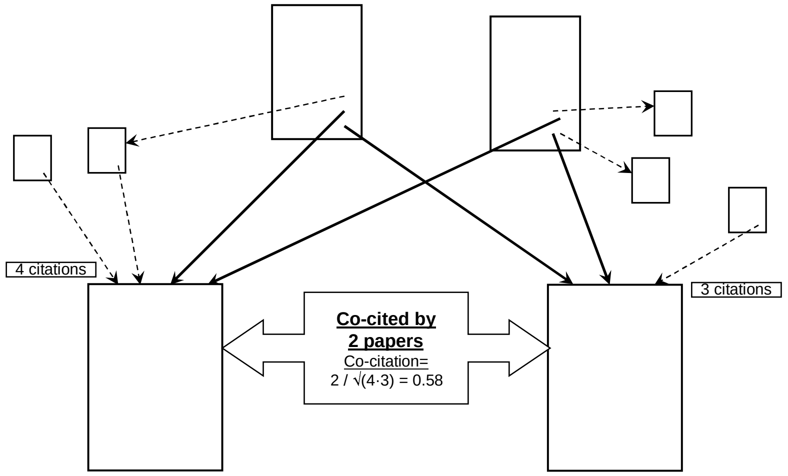
**Figuur 0.3:** Automatische verwerking van digitale documenten en hun voorstelling in het *vectorruimtemodel*. De tekst van alle $n$ documenten bovenaan de figuur wordt op automatische wijze geëxtraheerd. De volgorde van woorden en de stuctuur van zinnen wordt genegeerd, vandaar de naam *bag of words* voorstelling. Men telt alle woorden in een document (tijdens het *indexeren*) en de resulterende aantallen worden bewaard in een *term* × *document* matrix. Elke rij stelt een term (of woord) voor, en elke kolom een document. Alle $m$ woorden die in ten minste één document voorkomen, vormen het *vocabularium*, het *lexicon* of de *thesaurus*. Een waarde $w_{i,j}$ op rij $i$ en kolom $j$ in de matrix stelt het aantal keer voor dat woord $i$ voorkomt in document $j$, meestal gewogen door een extra *wegingsschema*. Elk document (kolom) kan voorgesteld worden als een vector, punt of coördinaat in een hoogdimensionale vectorruimte waarin elke dimensie één term voorstelt. Bijvoorbeeld, rechts onderaan de figuur worden de vectoren van de eerste twee documenten getoond in de tweedimensionale ruimte opgetrokken door de eerste twee termen. Een computerprogramma kan de similariteit (overeenkomst in onderwerp) van beide documenten bepalen door het berekenen van de hoek tussen beide vectoren. Hoe kleiner de hoek, hoe meer gerelateerd het onderwerp van de documenten. Door het grote aantal beschikbare documenten kan een term × document matrix zeer groot worden. De orde van grootte van $n$ kan tientallen miljoenen zijn. De grootte $m$ van het vocabularium is begrenst door het aantal unieke woorden of andere tekenreeksen (zoals bijvoorbeeld namen of projectnummers) die voorkomen in de tekstverzameling. Het totale vocabularium kan honderduizenden 'termen' bevatten, maar de uiteindelijke grootte hangt sterk af van de voorbewerkingsstrategie. We hebben de abstracten van miljoenen publicaties en patenten geïndexeerd, maar het grootste aantal dat we gebruiken voor domeinstudies in dit proefschrift is ongeveer tienduizend, met een vocabularium van twintigduizend termen.

De analyse van citatienetwerken is één van de belangrijkste toepassingen van de bibliometrie. Onderzoekers dragen hun bevindingen bij aan de wetenschappelijke gemeenschap waarvan zij verscheidene vormen van erkenning krijgen, bijvoorbeeld in de vorm van citaties [115]. Omdat de grote meerderheid van publicaties nooit geciteerd wordt, terwijl enkele publicaties enorm veel citaties krijgen, wijst de analyse van citatiegegevens op erg scheve verdelingen [5]. Publicaties die veel geciteerd worden, genieten meer aandacht van andere wetenschappers, waardoor de kans op nóg meer citaties nog vergroot [180, 4].

Alle citaties tussen een verzameling wetenschappelijke publicaties kunnen voorgesteld worden in een citatie- of literatuurnetwerk. In een *co-citatie*netwerk zijn twee publicaties verbonden wanneer beide geciteerd werden door eenzelfde derde publicatie. De onderliggende assumptie is dat co-citatie wijst op gerelateerde onderwerpen. De symmetrische co-citatiesterkte is een waarde tussen 0 en 1 en wordt berekend met behulp van *Saltons* cosinussimilariteit (zie Figuur 0.4, [236]). De co-citatiesterkte $CC(x,y)$ tussen twee artikels $x$ and $y$ is:

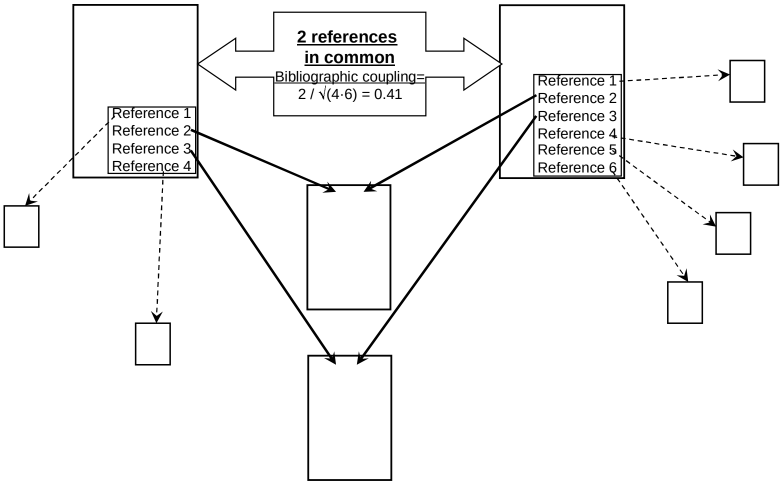
$$CC(x,y) = \frac{N_{xy}}{\sqrt{N_x \cdot N_y}}, \tag{0.2}$$

waarbij $N_x$ het totaal aantal citaties voorstelt dat artikel $x$ gekregen heeft, $N_y$ het totaal aantal keer dat artikel $y$ geciteerd werd, en $N_{xy}$ het aantal publicaties dat zowel artikel $x$ als artikel $y$ geciteerd heeft (dus het aantal bibliografieën dat referenties bevat naar beide artikels).



**Figuur 0.4:** Co-citatie. De onderste twee publicaties zijn respectievelijk 4 en 3 keer geciteerd. Twee keer werden beide artikels door eenzelfde publicatie geciteerd. Bijgevolg is de co-citatiesterkte gelijk aan $\frac{2}{\sqrt{4 \cdot 3}} = 0.58$.

In een netwerk op basis van *bibliografische koppeling* zijn twee publicaties verbonden als ze beide ten minste éénzelfde derde publicatie citeren [147] (zie Figuur 0.5). De koppelingssterkte $BC(x, y)$ wordt eveneens berekend met *Saltons* maat voor cosinussimilariteit. Bovenstaande formule kan dus toegepast worden, maar dan met $N_x$ en $N_y$ de aantallen referenties in artikel $x$ en artikel $y$, en $N_{xy}$ het aantal referenties gemeenschappelijk aan beide bibliografieën.

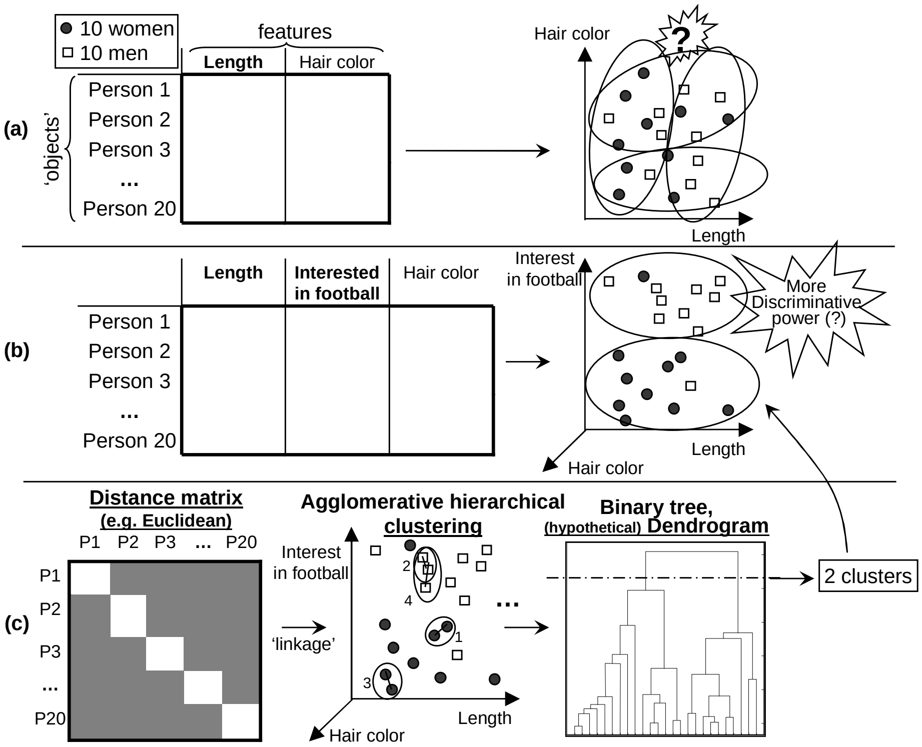


**Figuur 0.5:** Bibliografische koppeling. De bibliografieën van beide publicaties bovenaan bevatten respectievelijk 4 en 6 referenties. In de bibliografieën komen twee identieke referenties voor. Bijgevolg is de sterkte van bibliografische koppeling gelijk aan $\frac{2}{\sqrt{4 \cdot 6}} = 0.41$.

Een voordeel van bibliografische koppeling ten opzichte van co-citatie is dat bij bibliografische koppeling geen tijd nodig is voor het verkrijgen van een voldoende aantal citaties. Alle nodige informatie (referenties) is immers beschikbaar wanneer een artikel gepubliceerd wordt, wat een belangrijk voordeel oplevert voor doeleinden zoals opkomende-trenddetectie. Recente publicaties die onderling sterk gerelateerd zijn op basis van bibliografische koppeling kunnen momentopnames voorstellen van vroege stadia in de ontwikkeling van een specialiteit [96].

## Clustering

Onze inspanningen om de tekst- en netwerkwerelden te combineren in een hybride analyse zijn voornamelijk gericht op clusteralgoritmen. Figuur 0.6 biedt een overzicht van enkele belangrijke aspecten van clustering. Clusteren is een vorm van *ongesuperviseerd leren* omdat het algoritme objecten indeelt zonder voorgaande kennis in verband met het aantal groepen dat er is, en zonder voorbeelden van objecten die tot de groepen behoren. Classificatie daarentegen werkt op een *gesuperviseerde* manier: het algoritme krijgt informatie over de groep waartoe objecten in de *trainingverzameling* behoren.



**Figuur 0.6:** Overzicht van agglomeratieve hiërarchische clustering. Stel dat we 20 mensen willen indelen in twee groepen (clusters), één met vrouwen en één met mannen, maar dat het geslacht van de personen niet gekend is (vaak is zelfs het aantal gewenste groepen onbekend). Het doel van een clusteralgoritme is in dit geval het automatisch indelen van de personen in clusters, gebaseerd op gegevens die wel gekend zijn. Personen met gelijkaardige eigenschappen moeten dus in dezelfde groep terechtkomen en de verschillen tussen de groepen moeten zo groot mogelijk zijn. In **(a)** zijn enkel de eigenschappen *lengte* en *haarkleur* gekend voor elke persoon. Het is zeer moeilijk om op basis van deze gegevens homogene groepen te vinden omdat *haarkleur* geen onderscheid biedt tussen mannen en vrouwen en *lengte* onvoldoende. In **(b)** is ook de eigenschap *geïnteresseerd in voetbal* gekend. Deze eigenschap biedt meer informatie om onderscheid te maken tussen mannen en vrouwen. Natuurlijk zijn er nog steeds uitzonderingen: sommige mannen houden helemaal niet van voetbal terwijl dit voor sommige vrouwen juist wel geldt. **(c)**. De meeste clusteralgoritmen berekenen paarsgewijze afstanden (bv. *Euclidische*) tussen alle 'objecten' op basis van een selectie van gekende eigenschappen. Deze afstanden worden bewaard in een *afstandsmatrix*. Agglomeratieve hiërarchische clustering vertrekt van *singleton* clusters, waarbij elk afzonderlijk object in een aparte cluster zit, en groepeert iteratief *die* objecten of clusters waartussen de afstand het kleinst is (volgens een bepaald afstandscriterium). Dit iteratief samenbrengen gaat door tot alle objecten zich in één grote cluster bevinden. Een *dendrogram* is een visualisatie van dit proces. Zo'n hiërarchische boom kan 'afgesneden' worden op verschillende plaatsen om verschillende aggregatieniveaus te bekomen waarop de objecten onderverdeeld worden in meer of minder groepen. In dit voorbeeld is de boom afgeknipt op 2 clusters.

## Combinatie van tekstontginning en bibliometrie

Door seriële combinatie van tekstontginning en bibliometrie onderzoeken we in welke mate ze elkaar kunnen aanvullen om bij het in kaart brengen van wetenschap en technologie de individuele benaderingen te verbeteren. Documentgroepen gevonden dankzij tekstontginning bieden duidelijk additionele informatie om structuren gevonden met de hulp van bibliometrie uit te breiden, te verbeteren en te verklaren en vice versa. Publicaties met gelijkaardige inhoud kunnen verschillende bibliometrische eigenschappen hebben afhankelijk van de doelgroep en het applicatiedomein. Anderzijds kunnen bibliometrische indicatoren gebaseerd op referenties helpen om tekstgebaseerde clusters te verfijnen. Seriële combinatie van tekstontginning en bibliometrie blijkt een geschikte manier om cognitieve structuur te ontrafelen en te begrijpen. Daarom willen we beide informatiebronnen ook vroeger in het segmentatieproces integreren.

## Hybride clustering door integratie van tekst en bibliometrische informatie

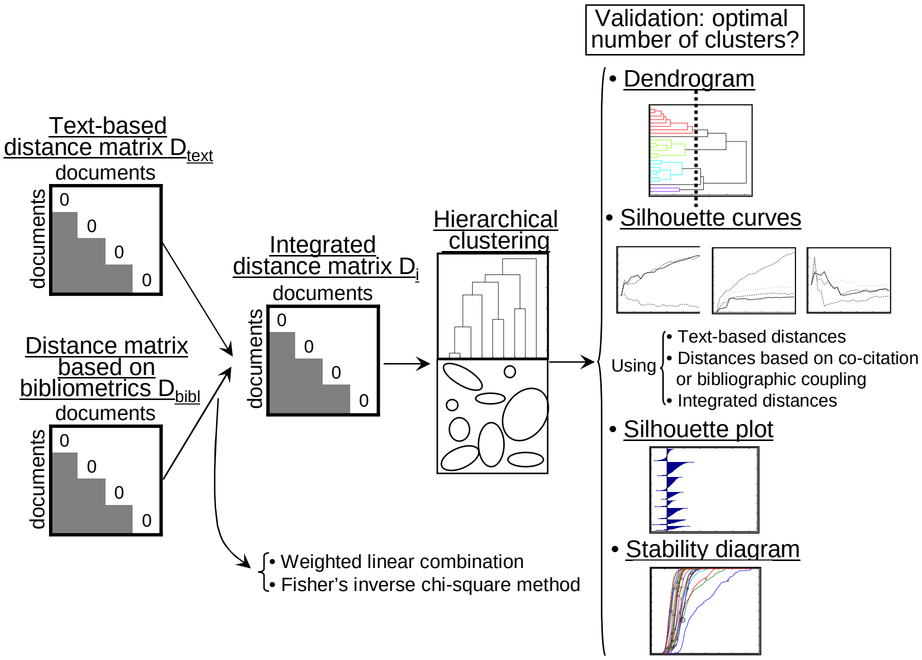
In deze dissertatie ontwikkelen we een methode voor het integreren van tekstontginning en bibliometrie. Meerdere informatiebronnen worden geïncorporeerd vóór toepassing van een clusteralgoritme. De eigenlijke integratie gebeurt door het combineren van ongelijksoortige afstanden tussen eenzelfde paar documenten, maar berekend met behulp van verschillende afstandsmaten die een andere blik op de documenten werpen. Paarsgewijze afstanden kunnen namelijk gebaseerd zijn op de tekstuele inhoud van documenten, maar ook op citaties (bv. gelijkenis tussen referentielijsten) of op andere bibliometrische eigenschappen.

Figuur 0.7 illustreert het integreren van afstanden. Belangrijk bij de meeste clusteralgoritmen is het bepalen van het aantal clusters waarmee de aanwezige onderwerpen zo goed mogelijk worden weergegeven. We maken gebruik van vier methoden voor clusterevaluatie. Voor het integreren van afstandsmatrices maken we gebruik van gewogen lineaire combinaties en van een methode gebaseerd op statistische meta-analyse. We stellen ook een methode voor gebaseerd op *Random Indexing*, waarvoor we een veelbelovend resultaat tonen. Voor illustratieve doeleinden beperken we het aantal databronnen tot twee, maar ook meerdere databronnen kunnen geïntegreerd worden. We combineren tekstuele inhoud en citaties aanwezig in een verzameling bio-informatica documenten, maar ook andere bibliometrische indicatoren kunnen samengevoegd worden.

Voor elke databron, zoals een genormaliseerde term × document matrix $A$ of een genormaliseerde referentie × document matrix $B$, kan een vierkante afstandsmatrix geconstrueerd worden als volgt:

$$\begin{aligned} D_t &= O_N - A^T \cdot A \\ D_{bc} &= O_N - B^T \cdot B \end{aligned} \tag{0.3}$$

met $N$ het aantal documenten en $O_N$ een vierkante matrix van dimensionaliteit $N$ gevuld met één'tjes. $bc$ verwijst naar bibliografische koppeling.


Validation: optimal number of clusters?
Text-based distance matrix D_text
documents
documents
0
0
0
0
Distance matrix based on bibliometrics D_bibl
documents
documents
0
0
0
0
Integrated distance matrix D_i
documents
documents
0
0
0
0
Hierarchical clustering
• Weighted linear combination
• Fisher's inverse chi-square method
• Dendrogram
• Silhouette curves
Using
• Text-based distances
• Distances based on co-citation or bibliographic coupling
• Integrated distances
• Silhouette plot
• Stability diagram

**Figuur 0.7:** Geïntegreerde hiërarchische clustering en evaluatie van resultaten om het aantal clusters te bepalen. Bij *hybride* of geïntegreerd clusteren zijn de paarsgewijze afstanden tussen documenten gebaseerd op informatie van zowel de tekstwereld (cf. figuur 0.3) als de netwerkwereld. De afstanden worden eerst berekend in beide werelden afzonderlijk, waarna ze geïntgreerd worden vóór toepassing van het clusteralgoritme. Afstandsmatrices gebaseerd op de tekst en op de netwerkstructuur worden op wiskundige en op statistische wijze gecombineerd alvorens ze gebruikt worden bij het clusteren. Om afstandsmatrices te integreren maken we in dit proefschrift voornamelijk gebruik van gewogen lineaire combinaties en van een methode gebaseerd op *Fisher's inverse chi-square method* (*Fishers* inverse chi-kwadraatmethode). Het aantal clusters in een documentverzameling, en dus het aantal voorgestelde onderwerpen, wordt bepaald met behulp van verschillende methoden. Enkele methoden evalueren de homogeniteit en de spreiding van clusters met behulp van statistische formules die rekening houden met alle afstanden binnen en tussen de clusters. Een andere methode evalueert de statistische stabiliteit van clusters door na te gaan of dezelfde objecten steeds in dezelfde clusters terechtkomen wanneer de clustering meerdere keren herberekend wordt voor een telkens lichtjes gewijzigde documentverzameling.

### Gewogen lineaire combinatie van afstandsmatrices

De afstandsmatrices $D_t$ en $D_{bc}$ kan men samenvoegen tot een geïntegreerde afstandsmatrix $D_i$ met behulp van een gewogen lineaire combinatie (linco):

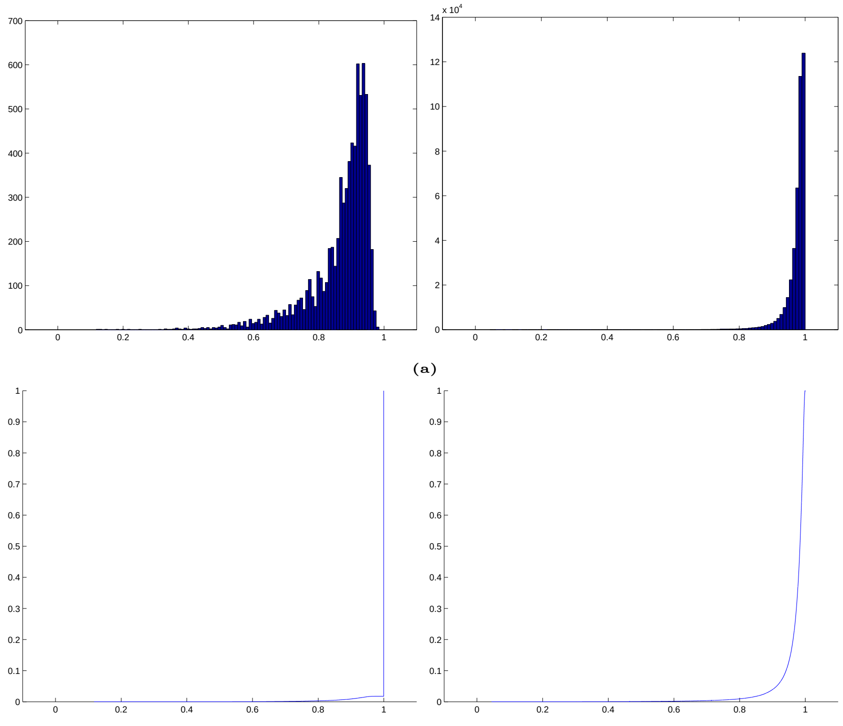
$$D_i = \alpha \cdot D_t + (1 - \alpha) \cdot D_{bc} \tag{0.4}$$

De resulterende $D_i$ kan men dan gebruiken in algoritmen voor clustering of classificatie. Hoewel dit een aantrekkelijke, eenvoudige en relatief schaalbare integratiemethode is, moet men er voorzichtig mee omspringen aangezien een lineaire combinatie belangrijke verschillen in distributionele eigenschappen van databronnen negeert. Figuur 0.8(a) toont de histogrammen met paarsgewijze afstanden (kleiner dan 1) tussen documenten gebaseerd op bibliografische koppeling (links) en tekstinformatie (rechts). Hoewel het gebruik van dezelfde afstandsmaat in dit geval leidt tot hetzelfde interval van mogelijke afstanden, verschillen de afstandsverdelingen van elkaar. Figuur 0.8(b) toont de empirische cumulatieve distributiefuncties van alle paarsgewijze afstanden (inclusief die gelijk aan 1). De verschillen worden nog duidelijker. De karigheid (*sparseness*) van bibliografische koppeling is zichtbaar door het grote aantal afstanden gelijk aan 1 ($> 95\%$). Deze verschillen in eigenschappen van verdelingen worden genegeerd door lineaire combinaties.

De discrepantie in distributionele eigenschappen wordt nog groter wanneer men andere informatiebronnen in aanmerking neemt. We hebben bijvoorbeeld ook tekstgebaseerde afstanden gecombineerd met artificiële Euclidische afstanden, berekend in een tweedimensionale ruimte bepaald door twee bibliometrische indicatoren. Verschillende afstandsmatrices (zoals term × document en indicator × document) kunnen inderdaad een verschillende afstandsmaat vereisen. Verschillen in overeenkomstige distributies kunnen een ongelijke of oneerlijke bijdrage van beide databronnen veroorzaken in de uiteindelijke geïntegreerde data. Dat kan leiden tot inferieure resultaten door het impliciet bevoorrechten van tekstuele inhoud of van bibliometrische eigenschappen. Valse of overdreven (dis)similarieiten kunnen correcte relaties, zichtbaar gemaakt door de andere databron, vernietigen.

### *Fishers* inverse chi-kwadraatmethode

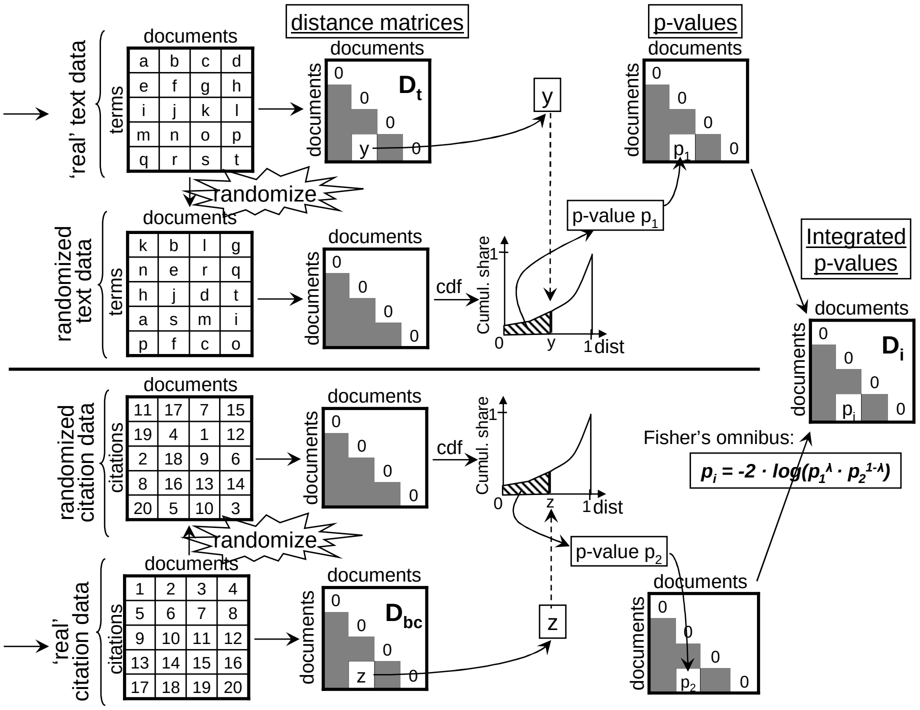
Behalve *vroege* integratiemethoden die data integreren vóór het berekenen van afstanden (bv. door aaneenvoegen van vectoren), en behalve een nieuwe methode om tekstuele inhoud en citaties te integreren met behulp van *Random Indexing*, ontwerpen we ook een methode gebaseerd op statistische meta-analyse. Figuur 0.9 illustreert het concept van afstandsintegratie door *Fishers* inverse chi-kwadraatmethode. Dat is een omnibusstatistiek om $p$-waarden van verschillende origine te combineren in een nieuwe $p$-waarde [123]. In tegenstelling tot de gewogen lineaire combinatie kan deze methode werken met afstanden afkomstig van verschillende metrieken en met verschillende distributionele eigenschappen. Ze vermijdt bovendien dat één informatiebron de andere domineert.



**(a)**

**(b)**

**Figuur 0.8:** Voor bibliografische koppeling (links) en tekstinformatie (rechts) bevat **(a)** histogrammen met alle paarsgewijze afstanden tussen documenten kleiner dan 1, en **(b)** de empirische cumulatieve distributiefuncties van alle paarsgewijze afstanden. De afstandsverdelingen verschillen duidelijk van elkaar (let ook op de verschillende schaal op de $Y$-as in (a)). Deze verschillen in parameters van de verdelingen worden genegeerd door lineaire combinaties.

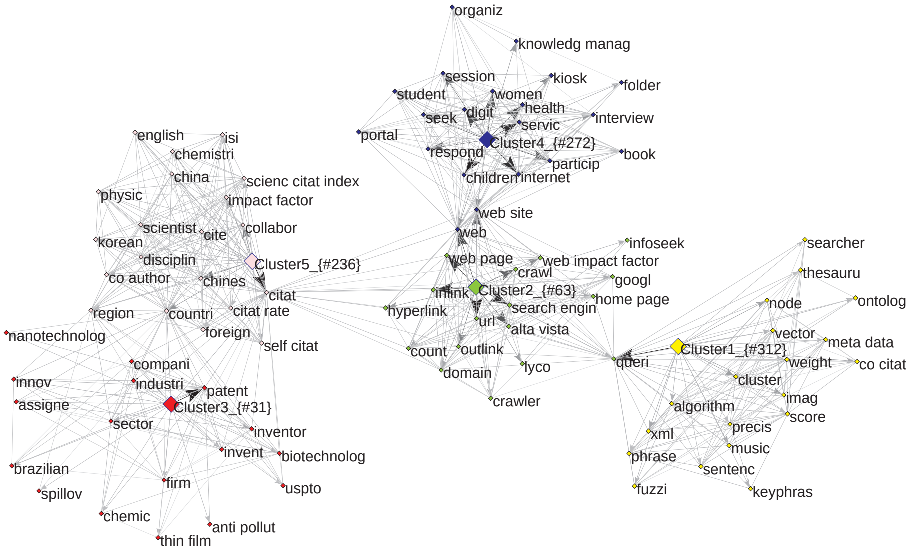


**Figuur 0.9:** Integratie van paarsgewijze afstanden tussen documenten met behulp van *Fishers* inverse chi-kwadraatmethode. Alle tekstgebaseerde afstanden in de afstandsmatrix $D_t$ en alle afstanden in $D_{bc}$ gebaseerd op citaties worden omgezet naar $p$-waarden ten opzichte van de empirische cumulatieve distributiefunctie van afstanden tussen gerandomiseerde data. Randomisatie gebeurt door het willekeurig herverdelen van woorden en citaties over alle documenten, terwijl karakteristieke eigenschappen bewaard blijven (bv. het gemiddeld aantal documenten waarin een bepaald woord voorkomt). Deze randomisatie is noodzakelijk voor het bekomen van geldige $p$-waarden. Een $p$-waarde betekent in deze context de kans dat de similariteit tussen twee documenten ten minste even groot zou kunnen zijn door louter toeval alleen. Door gebruik te maken van *Fishers* inverse chi-kwadraatmethode kan een geïntegreerde statistiek $p_i$ berekend worden op basis van de $p$-waarden voor de tekstdata ($p_1$) en de citatiegegevens ($p_2$). De resulterende matrix met geïntegreerde $p$-waarden is de nieuwe afstandsmatrix die men kan gebruiken in algoritmen voor clusteren of classificeren. Deze methode laat toe om afstanden te integreren die afkomstig zijn van verschillende metrieken met sterk verschillende distributies, en ze voorkomt dominantie van één van de informatiebronnen.

### Hybride studie van bibliotheek- en informatiewetenschappen

Dankzij de hybride clustering op basis van *Fishers* inverse chi-kwadraatmethode verkrijgen we een beter beeld van het domein van bibliotheek- en informatiewetenschappen, in kwantitatieve en kwalitatieve zin, in vergelijking met de tekstgebaseerde clustering en de lineaire combinatie. Twee clusters in verband met bibliometrie worden samengenomen, waardoor het domein ingedeeld wordt in 5 clusters. Er treedt een duidelijke verbetering op aangezien verschillende artikels in een meer relevante cluster terechtkomen door het gebruik van zowel tekst als citaties. Figuur 0.10 toont termnetwerken met voor elke cluster de 20 beste woordstammen uit titels en abstracten.

Hoewel lineaire combinatie enerzijds een eenvoudige en schaalbare methode is en er anderzijds in een eerder experiment geen significant verschil met *Fishers* inverse chi-kwadraatmethode kon worden vastgesteld, behaalt deze laatste in deze domeinstudie betere resultaten dan de lineaire combinatie.



**Figuur 0.10:** Termnetwerken met voor elk van de 5 clusters de 20 beste woordstammen.

## Bibliometrische informatievergaring

Een combinatie van tekstuele en bibliometrische componenten kan ook gebruikt worden in het kader van informatievergaring. Een belangrijke uitdaging in elke domeinstudie is het afbakenen van een vaak complex onderzoeksdomein zoals nanotechnologie of bio-informatica. Dit is verre van triviaal omwille van het interdisciplinaire karakter van veel wetenschappelijke deelgebieden en gezien de verspreiding van wetenschappelijke resultaten via verschillende kanalen (bv. multidisciplinaire tijdschriften). Om te voorkomen dat zoekopdrachten enkele bladzijden lang moeten zijn om alle relevante publicaties uit bibliografische databanken te verzamelen, maken we gebruik van *bibliometrische informatievergaring.* Dit is een uitbreiding van traditionele informatievergaring met componenten gebaseerd op bibliografische koppeling, referenties en citaties.

## Dynamische, hybride analyse van bio-informatica

De *bibliometrische informatievergaring* werd toegepast om bio-informatica af te bakenen, een domein gekenmerkt door een exponentiële groei in aantal publicaties gedurende de laatste twee decennia. Hierbij werd een verzameling samengesteld van 7401 relevante publicaties. In een bibliometrische analyse bestuderen we de groei van het domein, de internationale samenwerkingsverbanden, de patronen van nationale publicatie-activiteit en de citatie-impact. Vervolgens onderzoeken we de cognitieve structuur zoals waargenomen door het hybride clusteralgoritme.

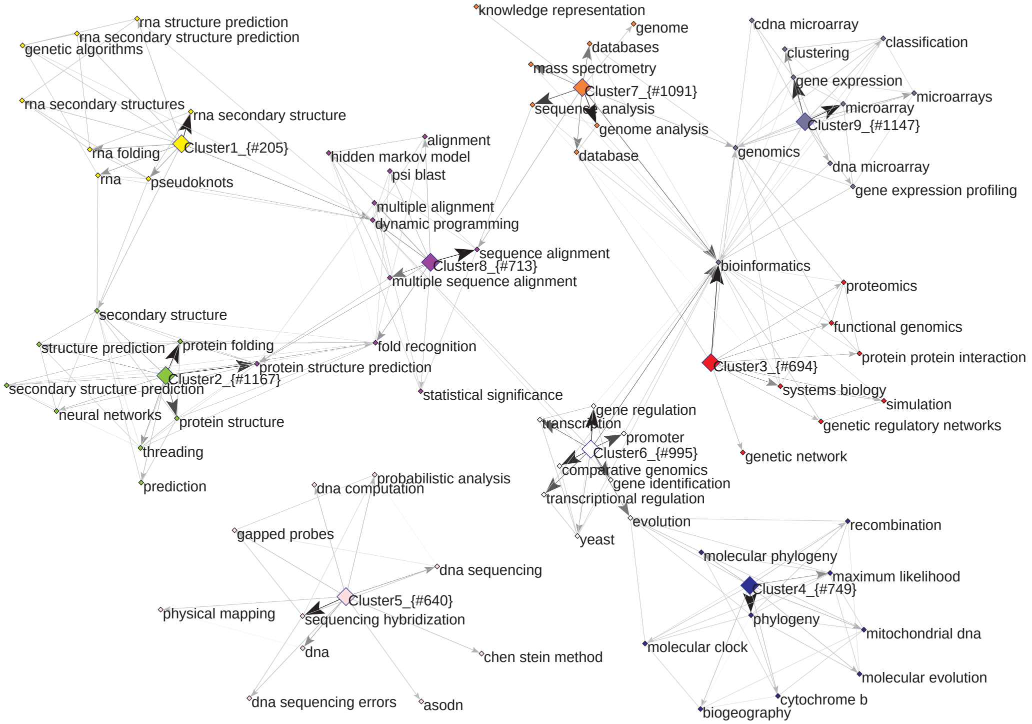
### Hybride clustering van bio-informatica

Om de bio-informatica artikels in groepen in te delen, maken we gebruik van agglomeratieve hiërarchische clustering gebaseerd op *Fishers* inverse chi-kwadraatmethode. De gecombineerde strategie om het aantal clusters te bepalen wijst op 9 clusters. Voor elke cluster tonen we term- en samenwerkingsnetwerken, representatieve publicaties, het relatieve belang voor de 5 meest actieve landen, citatiepatronen, en de 'naïeve dynamica' van de cluster.

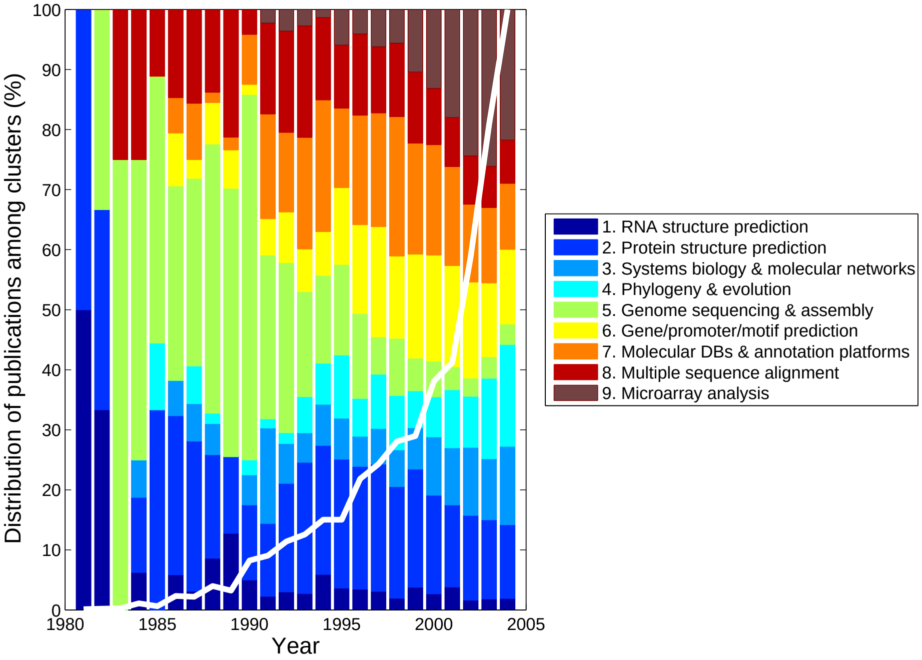
In tabel 0.1 geven we voor elke cluster de Engelse naam, het aantal documenten en de automatisch gedetecteerde belangrijkste woorden. Cluster 1 is met 205 publicaties de kleinste; alle andere bevatten meer dan 600 en minder dan 1200 artikels. Figuur 0.11 toont de cognitieve structuur van bio-informatica met behulp van termnetwerken die voor elke gevonden cluster de 10 beste termen weergeven. Belangrijke, alom gewaardeerde bio-informatica publicaties kunnen in elk deeldomein geïdentificeerd worden door analyse van het citatienetwerk. We gebruiken hiervoor de connectie-gebaseerde algoritmen HITS [149] en *Google*'s PageRank [37]. Verder bekijken we ook het (gemiddeld) aantal citaties en de *ISI* Impact Factor [89].

### *Naïeve* dynamica

Figuur 0.12 geeft een beeld van de populariteit van verschillende deelgebieden binnen bio-informatica gedurende de laatste twee decennia.



**Figuur 0.11:** Termnetwerken met voor elk van de negen clusters de 10 belangrijkste concepten (automatisch geïdentificeerd). Elke cluster wordt voorgesteld door een centrale *node* in de vorm van een ruit, die ook het aantal documenten in de cluster weergeeft. Elke centrale node wijst naar de beste termen voor een cluster. Wanneer een term tot de beste descriptors behoort voor meerdere clusters, dan wordt de term maar één keer herhaald maar is hij verbonden met meerdere centrale nodes. De grijswaarde en dikte van een pijl duiden het belang aan van een woord voor een bepaalde cluster. Twee woorden zijn verbonden als beide samen voorkomen in één of meerdere publicaties in een cluster; hoe frequenter ze samen voorkomen, hoe dichter de woorden bij elkaar staan.



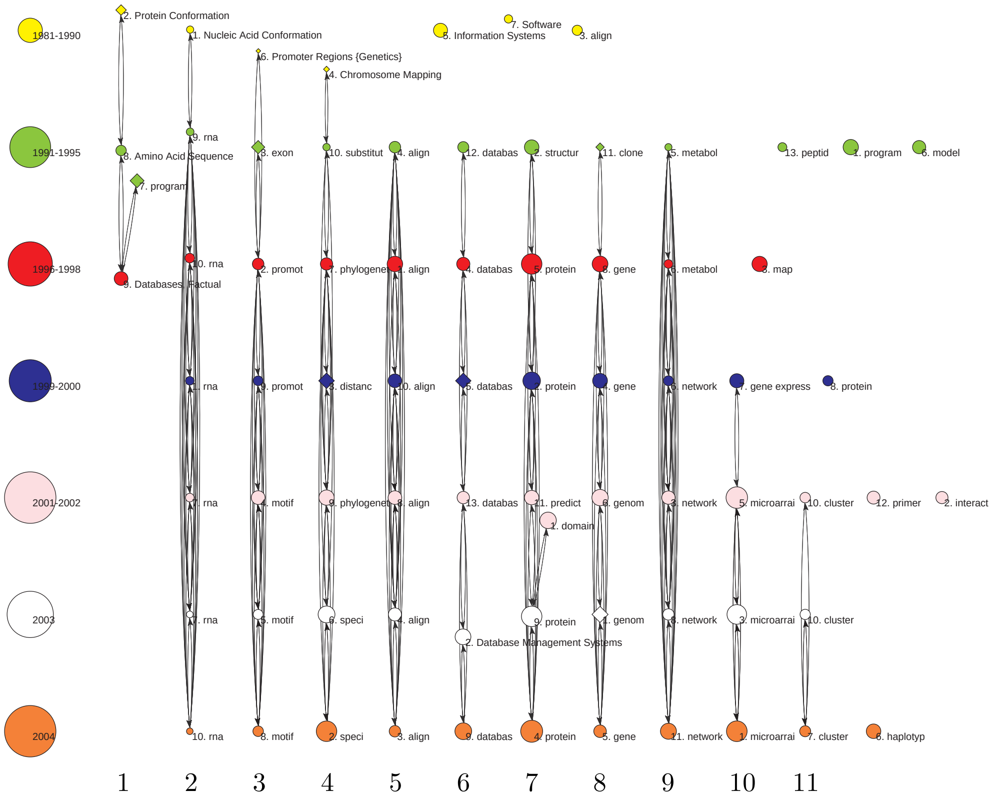
**Figuur 0.12:** *Naïeve* dynamica van de 9 clusters waarmee we zicht krijgen op de hoeveelheid aandacht die de bio-informaticagemeenschap doorheen de tijd aan de verschillende deelgebieden geschonken heeft. De term *naïeve* wijst erop dat tijdsinformatie genegeerd werd tijdens het clusteren, maar dat de jaartallen in rekening gebracht werden na het opdelen van de volledige verzameling publicaties. Met verschillende kleuren worden percentages weergegeven van de totale jaarlijkse publicatie-output die tot de verschillende clusters behoren. De witte lijn duidt per jaar het relatieve aantal publicaties aan ten opzichte van het aantal in 2004 (1455). Deze figuur toont het relatieve groeien en krimpen van de verschillende deelgebieden binnen de bio-informatica. Een stijgende trend kan toegeschreven worden aan de clusters *Microarray analysis* (#9; microroosteranalyse), *Phylogeny & evolution* (#4; fylogenie en evolutie) en *Systems Biology & molecular networks* (#3; Systeembiologie & moleculaire netwerken). Dat zijn duidelijk deelgebieden waarin vandaag veel onderzoek verricht wordt. Cluster #4 (*Phylogeny & evolution*) is een relatief oud onderzoeksdomein, maar nieuwe ontwikkelingen binnen de bio-informatica hebben voor een heropleving gezorgd. Sommige clusters, zoals *Genome sequencing & assembly* (#5; genoomsequentie en assemblage), stellen duidelijk oudere deelgebieden voor die in relatieve zin minder en minder aandacht krijgen.

**Tabel 0.1:** De 9 clusters binnen bio-informatica.

| Cluster | Naam | Aantal publicaties | Beste *author keyword* | Beste term in titels en abstracten | Beste MeSH term |
|---|---|---|---|---|---|
| 1 | RNA structure prediction | 205 | rna secondary structure | RNA | Nucleic Acid Conformation |
| 2 | Protein structure prediction | 1167 | protein structure prediction | protein | Proteins/chemistry |
| 3 | Systems biology & molecular networks | 694 | bioinformatics | network | Models, Biological |
| 4 | Phylogeny & evolution | 749 | phylogeny | phylogenet | Phylogeny |
| 5 | Genome sequencing & assembly | 640 | sequencing hybridization | base sequenc | Base Sequence |
| 6 | Gene/promoter/motif prediction | 995 | gene regulation | gene | Sequence Analysis, DNA/methods |
| 7 | Molecular DBs & annotation platforms | 1091 | genome analysis | databas | Databases, Factual |
| 8 | Multiple sequence alignment | 713 | sequence alignment | align | Sequence Alignment/methods |
| 9 | Microarray analysis | 1147 | microarray | microarrai | Oligonucleotide Array Sequence Analysis/methods |
|  | Alle bio-informatica publicaties | 7401 | bioinformatics | protein | Algorithms |

### Dynamisch clusteren

Figuur 0.13 illustreert de strategie die we uitgewerkt hebben voor het dynamisch clusteren van een evoluerende documentcollectie door het vergelijken en volgen van clusters doorheen de tijd. Dit is belangrijk voor het detecteren van opkomende trends, convergerende clusters en *hot topics*. Er werden zeven opeenvolgende perioden gedefinieerd voor een dynamische analyse. In elke periode werd een aparte, hybride, hiërarchische clustering uitgevoerd, waarbij het aantal clusters bepaald werd met de gecombineerde methode. Vervolgens werd een *complete graaf* gebouwd met als knopen alle clustercentra van elke periode, en als gewichten op de verbindingen de paarsgewijze cosinussimilariteiten. Nadien leidden twee stappen tot het vormen van *clusterkettingen*. Eerst werden enkel die verbindingen weerhouden die similariteiten van meer dan 95% voorstelden. Alle andere verbindingen werden verwijderd. Na toepassing van deze strenge voorwaarde waren de meeste clusterkettingen reeds gevormd. Bij een tweede stap werden clusters die met geen enkele andere cluster een similariteit boven 95% vertoonden toch in een ketting opgenomen als de similariteit met alle clusters in die ketting groter was dan 80%. Dergelijke clusters zijn weergegeven als een ruit in plaats van een cirkel. We analyseren de structuur, de evolutie en verschillende statistieken van elke clusterketting. ‘Dynamische’ termnetwerken laten toe om verschuivingen in samenwerkingspatronen en in terminologie te observeren. Tenslotte onderzoeken we de evolutie in citatiepatronen tussen clusterkettingen, alsook de jaarlijkse impact van elke clusterketting.



**Figuur 0.13:** Dynamisch clusteren: vergelijken en volgen van clusters doorheen de tijd. Elk horizontaal niveau stelt een periode voor zoals aangeduid in de linker kolom. De grootte van een cirkel stelt het aantal publicaties voor. Voor elke cluster is de beste term weergegeven, herleid tot de *stam* met behulp van de *Porter stemmer* [225]. Elf clusterkettingen werden gedetecteerd.

## Besluit

In dit proefschrift onderzoeken we of algoritmische en multivariate statistische verwerking van grote collecties wetenschappelijke literatuur toelaat om de inhoud, samenstelling en interactie van wetenschappelijke deelgebieden in kaart te brengen. Onze belangrijkste bijdragen zijn de volgende:

- ***Hybride clustering.*** Door seriële combinatie van tekstontginning en bibliometrie tonen we de complementariteit aan van de tekstuele inhoud van wetenschappelijke publicaties en de bibliometrische analyse van citaties. In het algemeen blijkt tekstinformatie krachtiger dan citaties voor zowel clustering als classificatie. De kwaliteit stijgt sterk door dimensionaliteitsreductie met behulp van singuliere-waardenontbinding (SWO), vooral indien toegepast op tekstinformatie. De beste resultaten worden echter behaald met geïntegreerde datatypes.

  We ontwerpen hybride methoden voor het clusteren van wetenschappelijke deelgebieden waarbij we tegelijkertijd rekening houden met de tekst en met de structuur van citatienetwerken. We tonen aan dat correcte statistische integratie bijdraagt tot de kwaliteit van het resultaat, en dat de geïntegreerde data een beter begrip opleveren van de structuur van wetenschappelijke kennisgebieden. De performantie van ongesuperviseerd clusteren en van classificeren verbetert significant door de integratie. Een clustermethode gebaseerd op statistische meta-analyse behaalt de beste resultaten en overtreft zowel methoden die enkel gebaseerd zijn op tekst of citaties, als integratiemethoden gebaseerd op aaneenvoegen van matrices. Paarsgewijze afstanden tussen documenten worden omgezet in $p$-waarden ten opzichte van de afstanden tussen gerandomiseerde data, en *Fishers* inverse chi-kwadraatmethode wordt vervolgens gebruikt om de $p$-waarden van verschillende origine te combineren. Deze methode laat toe om afstanden samen te voegen die afkomstig zijn van verschillende metrieken met sterk verschillende distributies, en voorkomt dominantie van één van de informatiebronnen. Maar deze methode bleek niet altijd significant verschillend van overeenkomstige lineaire combinaties van afstandsmatrices waarbij ook SWO gebruikt werd. Omwille van de complexiteit van *Fishers* inverse chi-kwadraatmethode en een gereduceerde schaalbaarheid, is een gewogen lineaire combinatie een eenvoudigere en eveneens effectieve oplossing voor het integreren van tekst- en citatie-informatie, op voorwaarde dat LSI gebruikt wordt. In een domeinstudie leverde *Fishers* inverse chi-kwadraatmethode evenwel betere resultaten op.

  Een combinatie van tekstuele en bibliometrische componenten helpt ook bij het afbakenen van complexe, interdisciplinaire wetenschappelijke deelgebieden zoals bio-informatica. Het afbakenen behelst de toepassing van verschillende strategieën voor informatievergaring om een collectie samen te stellen van publicaties die zo relevant mogelijk zijn voor het onderwerp. Dit is verre van triviaal omwille van het interdisciplinaire karakter van veel wetenschappelijke deelgebieden en de verspreiding van wetenschappelijke resultaten via verschillende kanalen (bv. multidisciplinaire tijdschriften).

- ***Dynamische, hybride clustering.*** We ontwikkelen een methode voor hybride dynamische analyse van groeiende bibliografische corpora door het vergelijken en volgen van clusters doorheen de tijd. Dit soort clustering biedt een kijk op de evolutie van bestaande deelgebieden en op de aandacht die in verschillende perioden uitgaat naar verschillende onderwerpen. Dit draagt bij tot het ontdekken van opkomende of convergerende clusters en *hot topics*.

- ***Aantal clusters in een documentcollectie.*** Het aggregatieniveau waarop een documentcollectie ingedeeld moet worden in groepen is moeilijk te achterhalen. Verschillende algoritmen en formules voor evaluatie en validatie zijn voorhanden, maar vaak is er geen eenduidig antwoord. Desondanks illustreren we dat het gebruik van verschillende methoden duidelijke indicaties oplevert voor een correct aantal clusters. We beschrijven een samengestelde, semi-automatische strategie voor het bepalen van het aantal clusters. Het betreft een combinatie van methoden gebaseerd op afstanden en op stabiliteit. Een eerste indicatie wordt geleverd door een aangewezen afsnijpunt in het dendrogram. Daarnaast gebruiken we curves met gemiddelde Silhouettewaarden (gebaseerd op tekst en citaties) voor verschillende aantallen clusters. De tekst- en netwerkwerelden bieden complementaire informatie voor het bepalen van het aantal clusters. Tenslotte evalueren we de kwaliteit van een clustering met de stabiliteitsmethode voorgesteld door *Ben-Hur et al.* [16].

- ***Aantal factoren voor Latent Semantische Indexering.*** Latent Semantische Indexering (LSI) is een techniek voor dimensionaliteitsreductie gebaseerd op de singuliere-waardenontbinding van een term × document matrix. Een interessant effect van LSI is dat synoniemen of verschillende woordcombinaties die hetzelfde betekenen impliciet gerelateerd worden als gevolg van de gemeenschappelijke context waarin ze meestal voorkomen, zelfs wanneer deze woorden nooit samen voorkomen in eenzelfde document. Een zoekmachine kan dus documenten vinden die de zoektermen niet letterlijk bevatten. De zoekopdracht *auto* zou bijvoorbeeld ook documenten kunnen opleveren waarin enkel over *wagen* geschreven wordt, en dit zonder enig gebruik van een woordenboek. Een ander belangrijk voordeel van LSI is dat reductie van het aantal dimensies in een vectorruimte de performantie van clustering en classificatie verbetert. Het is echter zeer moeilijk om het aantal te weerhouden dimensies te bepalen. We tonen aan dat een goede keuze een sterke invloed heeft op de nauwkeurigheid van de resultaten. We onderzoeken de relatie tussen enerzijds de performantie van het clusteren en anderzijds het gewenste aantal clusters en het aantal factoren voor LSI. De nauwkeurigheid van het clusteren van bio-informatica documenten, gemeten met de *Silhouette coefficient*, is significant hoger voor een lager aantal factoren. Hoewel in de literatuur vaak een waarde tussen 100 en 300 genomen wordt voor het aantal factoren, tonen we aan dat een zeer bescheiden aantal (bv. 10) de beste resultaten biedt, op voorwaarde dat het aantal LSI factoren niet kleiner is dan het gewenste aantal clusters. Dit dient echter verder onderzocht voor andere datacollecties.

- ***Domeinstudie bibliotheek- en informatiewetenschappen.*** Het doel van deze eerste domeinstudie is het ontrafelen en visualiseren van de bibliotheek- en informatiewetenschappen. In eerste instantie analyseren we de tekst in 938 publicaties uit 5 tijdschriften, waarbij we alle bibliografische en bibliometrische componenten negeren. Dit levert zes clusters op. Maar dankzij de hybride clustering worden twee clusters in verband met bibliometrie samengenomen en krijgen we een beter beeld van het domein, zowel in kwantitatieve als kwalitatieve zin.
- ***Structurele en bibliometrische domeinstudie van bio-informatica.*** Onze procedure voor geïntegreerd clusteren gebaseerd op *Fishers* inverse chi-kwadraatmethode wordt ingezet voor het onderzoeken en visualiseren van bio-informatica. Het afbakenen van het domein (7401 publicaties) gebeurt met behulp van *bibliometrische informatievergaring*. De gecombineerde strategie voor het bepalen van het aantal clusters suggereert 9 deelgebieden. Voor elke cluster genereren we term- en samenwerkingsnetwerken en representatieve publicaties. Bovendien onderzoeken we de belangrijkste tijdschriften, de evolutie van publicatie-output en citatie-impact, het belang van deelgebieden voor de 5 meest actieve landen, en de samenwerking op verschillende niveaus van aggregatie. Daarnaast analyseren we ook de *naïeve dynamica* van elke cluster, waarmee bedoeld wordt dat we het jaartal van publicatie niet in aanmerking nemen tijdens het clusteren, maar enkel achteraf. Tenslotte definiëren we zeven opeenvolgende perioden voor een dynamische analyse.

# Publication list

## International journal papers

- P. Glenisson, W. Glänzel, F. Janssens, and B. De Moor. Combining full text and bibliometric information in mapping scientific disciplines. *Information Processing & Management*, 41(6):1548–1572, 2005.
- N. L. M. M. Pochet, F. A. L. Janssens, F. De Smet, K. Marchal, J. A. K. Suykens, and B. L. R. De Moor. M@CBETH: a microarray classification benchmarking tool. *Bioinformatics*, 21(14):3185–3186, 2005.
- F. Janssens, J. Leta, W. Glänzel, and B. De Moor. Towards mapping library and information science. *Information Processing & Management*, 42(6):1614–1642, 2006.
- V. Rodriguez, F. Janssens, K. Debackere, and B. De Moor. Do material transfer agreements affect the choice of research agendas? The case of biotechnology in Belgium. *Scientometrics*, 71(2):239–269, 2007.
- V. Rodriguez, F. Janssens, K. Debackere, and B. De Moor. Material transfer agreements and collaborative publication activity: The case of a biotechnology network. Accepted for publication in Research Evaluation, 2007.

## International conference papers

- J. Vertommen, F. Janssens, B. De Moor, and J. Duflou. Advanced personalization and document retrieval techniques in support of efficient knowledge management. In *Proceedings of the 2nd International Seminar on Digital Enterprise Technology (DET2004)*, Seattle, Washington, USA, Sep. 2004.
- F. Janssens, P. Glenisson, W. Glänzel, and B. De Moor. Co-clustering approaches to integrate lexical and bibliographical information. In P. Ingwersen and B. Larsen, editors, *Proceedings of the 10th international conference of the International Society for Scientometrics and Informetrics*

*(ISSI)*, volume 1, pages 284–289, Stockholm, Sweden, July 2005. Karolinska University Press.

- N. Pochet, F. A. L. Janssens, F. De Smet, K. Marchal, I. Vergote, J. A. K. Suykens, and B. De Moor. M@CBETH: Optimizing clinical microarray classification. In *Proceedings of the 2005 IEEE Computational Systems Bioinformatics Conference (CSB2005)*, Stanford, California, USA, Aug. 2005, pages 89–90.

- F. Janssens, V. Tran Quoc, W. Glänzel, and B. De Moor. Integration of textual content and link information for accurate clustering of science fields. In *Proceedings of the I International Conference on Multidisciplinary Information Sciences & Technologies (InSciT2006).* Current Research in Information Sciences and Technologies, volume I, pages 615–619, Mérida, Spain, October 2006.

- W. Glänzel, F. Janssens, and B. Thijs. A comparative analysis of publication activity and citation impact based on the core literature in bioinformatics. In *Proceedings of the 11th International Conference of the International Society for Scientometrics and Informetrics (ISSI)*, Madrid, Spain, 2007.

- F. Janssens, W. Glänzel, and B. De Moor. A hybrid mapping of information science. In *Proceedings of the 11th International Conference of the International Society for Scientometrics and Informetrics (ISSI)*, Madrid, Spain, 2007.

## Technical reports

- F. Janssens and B. De Moor. Application of HITS algorithms to detect terms and sentences with high saliency scores. Technical Report 04-29, ESAT-SISTA, K.U.Leuven, Leuven, Belgium, 2004.

- V. Rodriguez, F. Janssens, K. Debackere, and B. De Moor. Material transfer agreements and interorganisational collaboration: Biotechnology network of co-authorship and co-assigneeship. Technical Report 06-206, ESAT-SISTA, K.U.Leuven, Leuven, Belgium, 2006.

- J. Vertommen, F. Janssens, J. Duflou, and B. De Moor. Multiple-vector user profiles for knowledge management systems. Technical Report 06-22, ESAT-SISTA, K.U.Leuven, Leuven, Belgium, 2006.

- T. Van Herpe, K. Pelckmans, J. De Brabanter, F. Janssens, B. De Moor, and G. Van den Berghe. Assessing the accuracy of glycemia sensors: The GLYCENSIT procedure. Paper in revision for *Clinical Chemistry*, Technical Report 06-135a, ESAT-SISTA, K.U.Leuven, Leuven, Belgium, 2006.

# List of acronyms

**AUC** Area Under the ROC Curve

**BC** Bibliographic Coupling

**BR** Bibliometric Retrieval

**Candecomp** Canonical Decomposition

**DAG** Directed Acyclic Graph

**DB** Database

**dBC** dense Bibliographic Coupling

**DEDICOM** DEcomposition into DIrectional COMponents

**DOC** Microsoft Word file format

**EPO** European Patent Office

**ETD** Emerging Trend Detection

**GNU** GNU's Not Unix

**GPL** General Public License

**GSVD** Generalized Singular Value Decomposition

**ICT** Information and Communication Technology

**IR** Information Retrieval

**ISI** Institute for Scientific Information

**kNN** $k$-Nearest Neighbor

**LDA** Latent Dirichlet Allocation

**LIS** Library and Information Science

**LSI** Latent Semantic Indexing

**MCL** Markov CLuster algorithm

**MDS** MultiDimensional Scaling

**MECR** Mean Expected Citation Rate

**MeSH** Medical Subject Headings

**MOCR** Mean Observed Citation Rate

**MRA** Mean Reference Age

**MTA** Material Transfer Agreement

**NLM** National Library of Medicine

**NLP** Natural Language Processing

**NMF** Non-negative Matrix Factorization

**OCR** Optical Character Recognition

**PARAFAC** PARAllel FACtors

**PDF** Portable Document Format

**PLSI** Probabilistic Latent Semantic Indexing

**PHITS** Probabilistic HITS

**QSVD** Quotient Singular Value Decomposition

**RI** Random Indexing

**ROC** Receiver Operating Characteristic

**S&T** Science and Technology

**SCIE** Science Citation Index Expanded

**SVC** Silhouette Value per Clustering

**SVD** Singular Value Decomposition

**TF** Term Frequency

**TF-IDF** Term Frequency - Inverse Document Frequency

**URL** Uniform Resource Locator

**UPGMA** Unweighted Pair Group Method using arithmetic Averaging

**USPTO** United States Patent and Trademark Office

**VSM** Vector Space Model

**WoS** Web of Science

# Contents

# Chapter 1

# Introduction

Since the information age and the knowledge economy, the availability of information in digital format has tremendously grown and is continuously increasing. Figure 1.1 shows the upward trend in the yearly estimated number of Web sites on the World Wide Web. A few years ago, rough estimates already mentioned 550 billion online documents [17], with a total size of 7.5 petabyte of data on Web sites and in public databases.[1] This is 4 times more than the space needed to store all information available in all U.S. academic research libraries [173]. In order to store 7.5 petabyte of information, a pile of plain text documents with about 2500 characters per page and 1 byte per letter or character, would be as high as 300 000 km and would consequently almost reach the moon, or traverse the circumference of the earth 7.5 times (1 cm for 100 pages). A person reading 1 page each minute would need to keep on reading for almost 5.7 million years to read it all! Fortunately, techniques from information retrieval, text mining and link or network analysis are here to save the poor reader from this *Sisyphean* challenge...

In addition, the dissemination of scientific and technological publications via the Internet and in large-scale bibliographic databases has become standard practice. Figure 1.2 shows the yearly growth of MEDLINE[2], which covers fields such as medicine, nursing and dentistry. Today, MEDLINE contains approximately 15 million journal articles in life sciences. Another database of major importance is the *ISI* Web of Science[3], which stores all bibliographic information from nearly 9300 of the most prestigious research journals in the world. Today, the complete WoS database contains over 36 million records and provides over 1.1 million records per year from more than 230 disciplines in science, social sciences, arts and humanities (see Figure 1.3). Patent databases grow as well. Figure 1.4 provides the yearly total number of patent applications filed by the

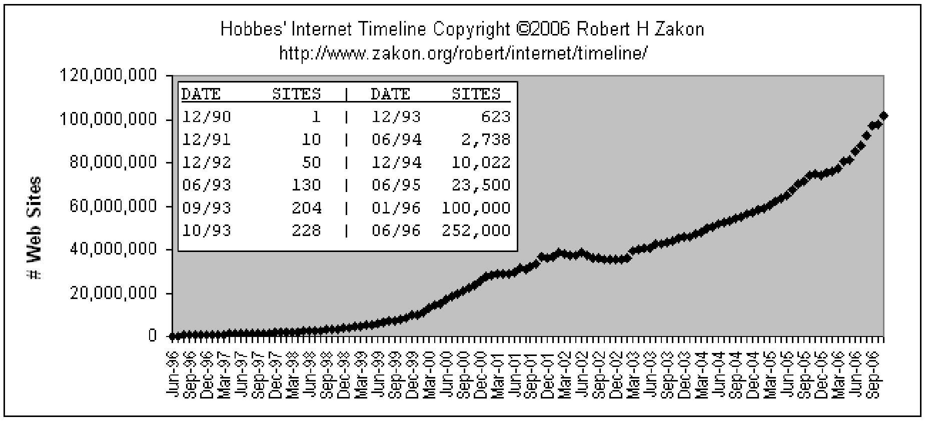
[1] One petabyte contains $10^{15}$ bytes.

[2] http://www.pubmed.org, visited in January 2007.

[3] http://scientific.thomson.com/products/wos/, visited in January 2007.

*European Patent Office*[4](EPO). The EPO has access to 56 million documents from over 70 countries.

For individuals and organizations alike, this overwhelming amount of digital data leads to major difficulties to find and process relevant information and knowledge. Search engines are essential to find relevant information, but often return a mass of irrelevant results within very long result lists. Information retrieval should be complemented with other algorithms to move beyond the mere finding of interesting documents. Existing classifications of science are inherently outdated because of the pace at which scientific knowledge advances.



Hobbes' Internet Timeline Copyright ©2006 Robert H Zakon
http://www.zakon.org/robert/internet/timeline/

| DATE | SITES | DATE | SITES |
|---|---|---|---|
| 12/90 | 1 | 12/93 | 623 |
| 12/91 | 10 | 06/94 | 2,738 |
| 12/92 | 50 | 12/94 | 10,022 |
| 06/93 | 130 | 06/95 | 23,500 |
| 09/93 | 204 | 01/96 | 100,000 |
| 10/93 | 228 | 06/96 | 252,000 |

**Figure 1.1:** The estimated number of Web sites on the World Wide Web. More precisely, the number of servers is counted. Hence, the actual number of sites will be larger since one server may host multiple sites. Moreover, most Web sites contain a multitude of Web pages or documents. *Google* currently indexes more than eight billion pages! The total number of static and dynamic Web pages is even many times larger and continuously increasing.

[4] http://www.european-patent-office.org/index.en.php, visited in January 2007.

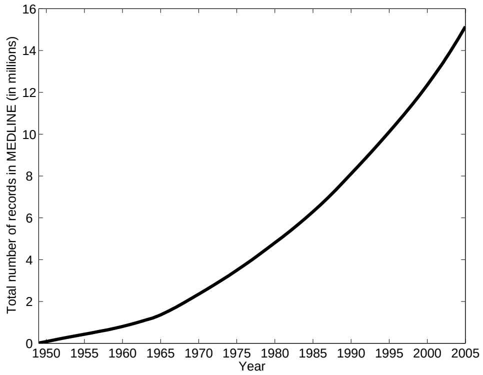


**Figure 1.2:** Growth of MEDLINE, the U.S. National Library of Medicine (NLM) premier bibliographic database covering the fields of medicine, nursing, dentistry, veterinary medicine, the health care system and preclinical sciences. The total number of scientific publications (in millions) is indicated for each year. Today, MEDLINE contains approximately 15 million unique records about journal articles in life sciences. This figure was constructed using data published by NLM [49].

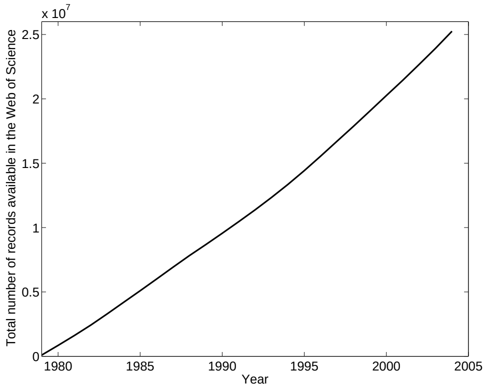


**Figure 1.3:** Total number of records in the local copy of the *ISI* Web of Science database that is available at the *Steunpunt O&O Indicatoren* (Leuven, Belgium). The Web of Science contains current and retrospective information (to the year 1900), derived from nearly 9300 of the most prestigious, high impact research journals in the world. Today, the complete WoS database contains over 36 million records and provides over 1.1 million records per year from more than 230 disciplines in science, social sciences, arts and humanities. The WoS provides access to all significant items within each research journal covered, including articles, bibliographies, reviews, editorials, letters and notes.

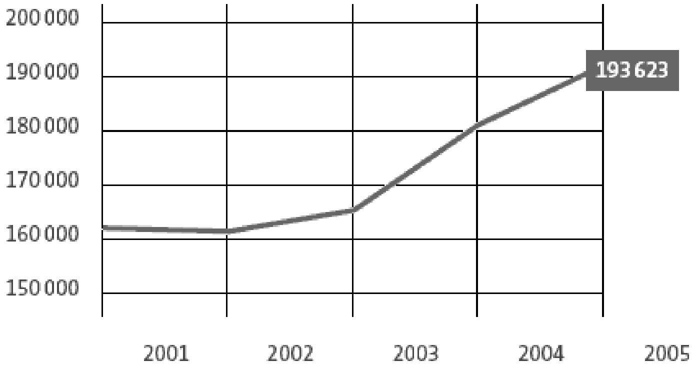


**Figure 1.4:** The total number of patent applications filed yearly by the *European Patent Office* (EPO). When carrying out patent searches, the EPO has access to 56 million documents from over 70 countries. Diagram taken from the EPO Annual Report 2005 [79].

## 1.1 General context

The general scope of this dissertation is the ***mapping*** of scientific and technological fields by using ***clustering*** algorithms and techniques from ***bibliometrics*** and ***text mining***.

***Text mining*** comprises the intelligent automated analysis of textual data and aims for extraction of interesting facts and relationships and discovery of knowledge from large amounts of texts. For this purpose, text mining employs techniques and algorithms from disciplines such as data mining, information retrieval, statistics, mathematics, machine learning and computational linguistics.

***Bibliometrics*** is an interdisciplinary science in which statistical and mathematical indicators, methods and models are used to study written scientific communication, mostly collected in large databases containing scientific publications or patents. Although somewhat more general in scope, *bibliometrics* is today often used synonymously with *scientometrics*.

The purpose of ***mapping***, *charting* or *cartography* of scientific fields is to understand the structure and evolution of various research areas and of their relationships with other fields, based on scientific publications. Whether available in full-text documents or as records stored in bibliographic databases, such publications contain a wealth of information and are considered to be indirect, but true reflections of scientific knowledge and activity. Research fields can be profiled in terms of prolific authors, major concepts, important publications and journals, institutions, regions and countries, etc. Knowledge about the amount of activity in various fields and about new, emerging and converging fields is important to organizations, research institutions and nations. Quantitative information can be used for evaluation of research performance and to support innovation management and science and technology policies (for example, what fields should be supported through funding?). Such policies are crucial to maintain and improve the competitive position of nations and organizations.

***Clustering*** is a multivariate statistical technique to automatically subdivide a set of objects into groups. The purpose is to make each group (or cluster) as homogeneous as possible in the sense that all objects in it have similar properties, while objects in different clusters should be as dissimilar as possible. For example, in the case of documents, the occurrence of a lot of common words might indicate that both documents are similar and discuss the same subjects.

A few examples are given in Figures 1.5–1.8. Figure 1.5 visualizes a modest literature network that was built by using bibliographic information from the Web of Science. Distinct groups of publications that were found by a link-based clustering algorithm are identified with a different color. The algorithm succeeded in finding several homogeneous clusters of publications on well-defined topics. In Figure 1.6, a visualization is provided of topical similarity of scientific documents, represented as mutual distances in a two-dimensional space. Each dot represents a scientific publication and the distance between two dots represents how (dis)similar the subject of the documents is. A small distance indicates that both documents are on comparable subjects. Next, Figure 1.7 shows term or concept networks that can be used to comprehend the content of large sets of documents. Nine clusters were found in a set of 7401 bioinformatics publications. Finally, Figure 1.8 visualizes international collaboration in bioinformatics, based on affiliations indicated on the same 7401 publications.

The **main hypothesis** to be verified in this dissertation states that the performance of clustering of scientific fields and the classification of scientific publications improves by integration of heterogeneous information. More specifically, citation-based bibliometric data is incorporated with textual content of scientific publications or patents. Clustering performance is computed with statistical measures for how 'happy' clustered documents are in their own cluster versus how 'happy' they would be in another cluster. Performance of classification is quantified by contrasting results with a 'ground truth' or 'gold standard' classification that is based on expert knowledge contained in Medical Subject Headings that are annotated to publications in MEDLINE (MeSH[5] terms).

[5] http://www.nlm.nih.gov/mesh/, visited in January 2007.

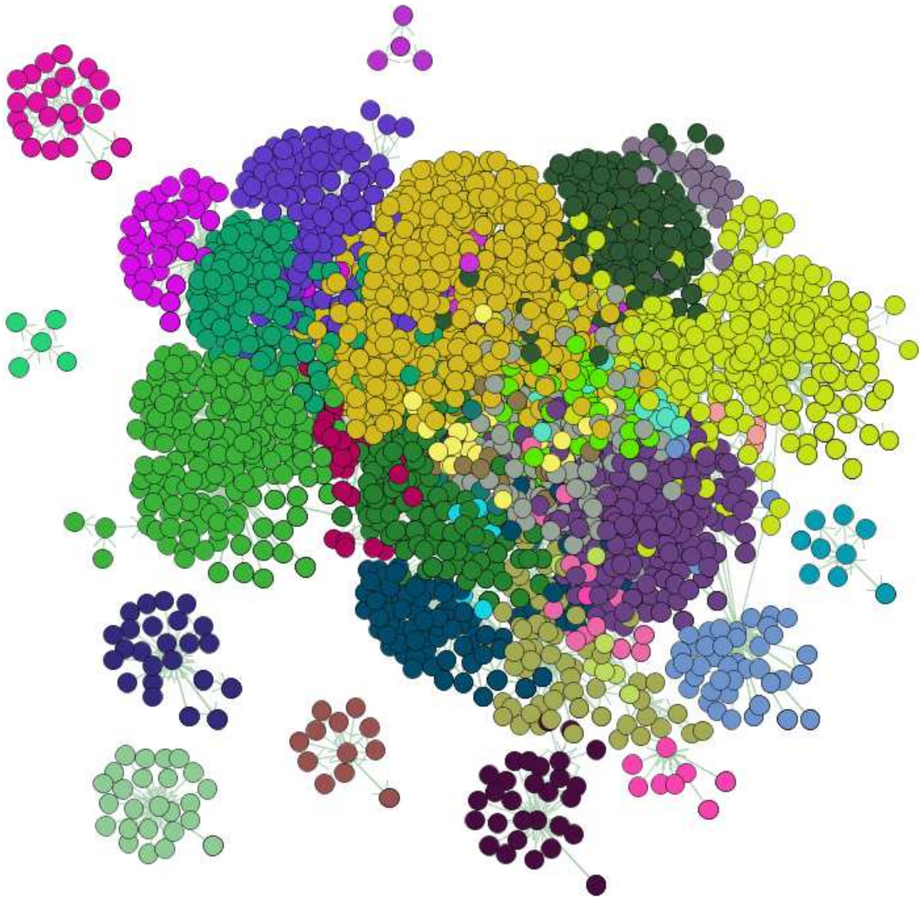


**Figure 1.5:** For the occasion of the Honorary Doctoral Degree awarded to Professor *Lennart Ljung* [168] at the Workshop on System Identification and Data Modeling (October 12–13, 2004, *Katholieke Universiteit Leuven*, Belgium), we built a modest literature network by using bibliographic information from the Web of Science. The network was constructed using as seed papers all 138 papers of *Ljung L* known to the Web of Science, and by extending the network with all cited and all citing publications. The resulting directed graph contains 4943 nodes or vertices (publications represented as small circles) and 6216 edges or links (citations represented as small arrows). In this graph, distinct groups of publications that were found by a link-based clustering algorithm are identified with a different color. The algorithm succeeded in finding several homogeneous clusters of publications on well-defined topics, and could also identify 13 papers among the seed publications that were written by another author *Ljung L*. We used Biolayout Java by *Enright* and *Ouzounis* to visualize the network [78].

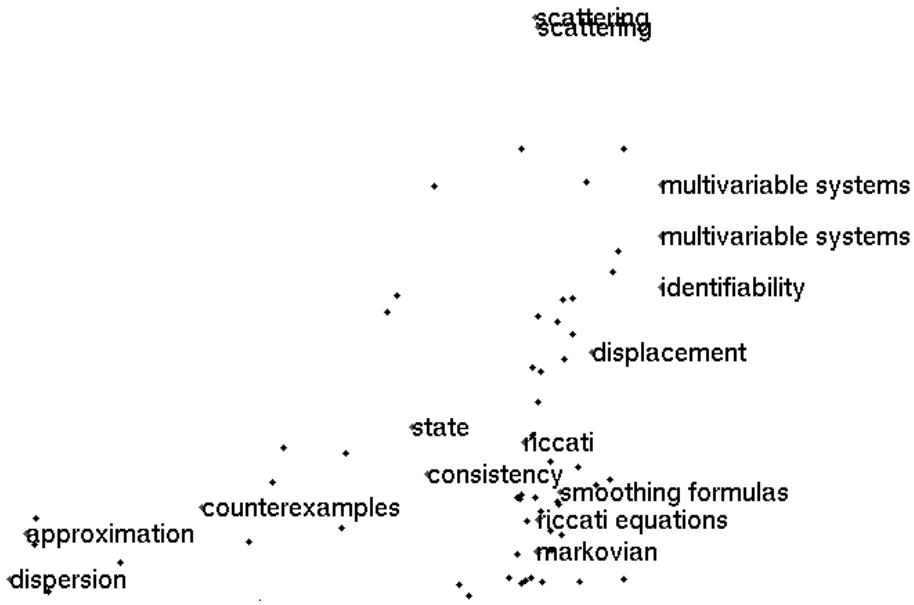


**Figure 1.6:** Each dot represents a scientific publication by *Lennart Ljung* [168]. For about half of the publications, a word is shown that describes the content, as automatically determined by text mining. The distance between documents (dots) in this two-dimensional figure was computed by a specific algorithm (Multidimensional scaling, MDS) and represents how (dis)similar the contents of the documents are. A small distance indicates that both documents are on comparable subjects. For example, in the upper part of the figure, two documents are very close and hence are perceived as very similar in content. Indeed, the best term *scattering* is the same for both documents. The same goes for two documents on *multivariable systems*.

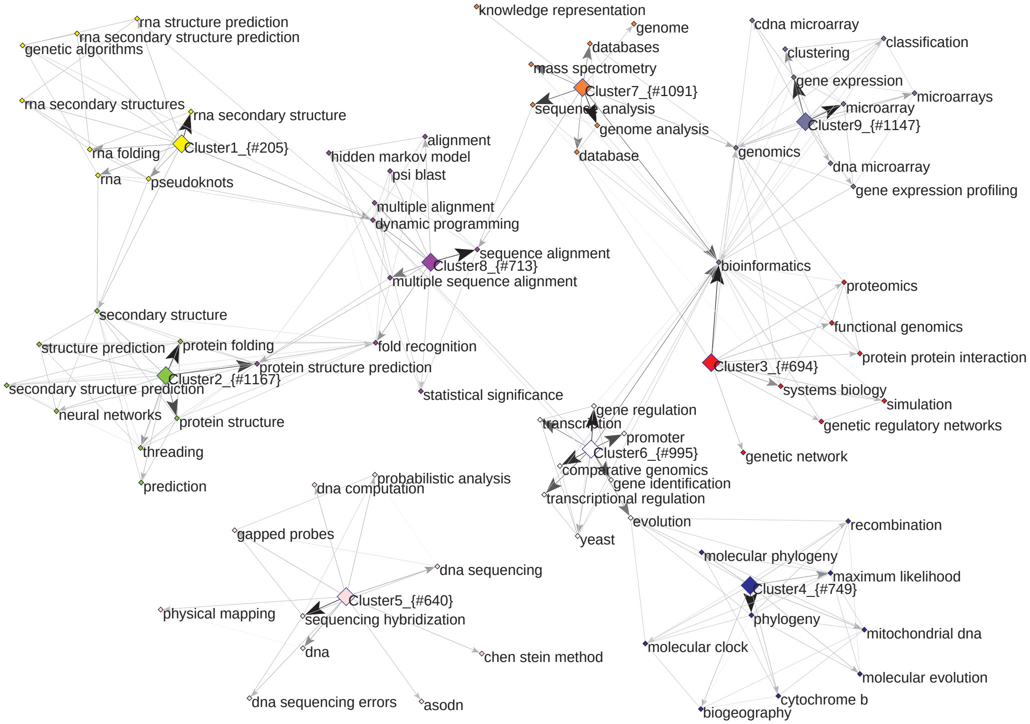


**Figure 1.7:** Term or concept networks that describe the content of nine clusters (nine groups of documents) that were found in a set of 7401 bioinformatics publications. Each cluster is represented by a 'central node' (a diamond), which also indicates the number of documents in the cluster. Each central node points to the best 10 keywords that describe the content of the corresponding cluster. When a keyword is among the best 10 for more than one group, it is only repeated once but connected to all corresponding central nodes. The gray level and thickness of an arc reflect the importance of a keyword for a cluster. Two terms are connected if both co-occur in one or more papers of the same cluster; the more co-occurrences, the closer the terms.

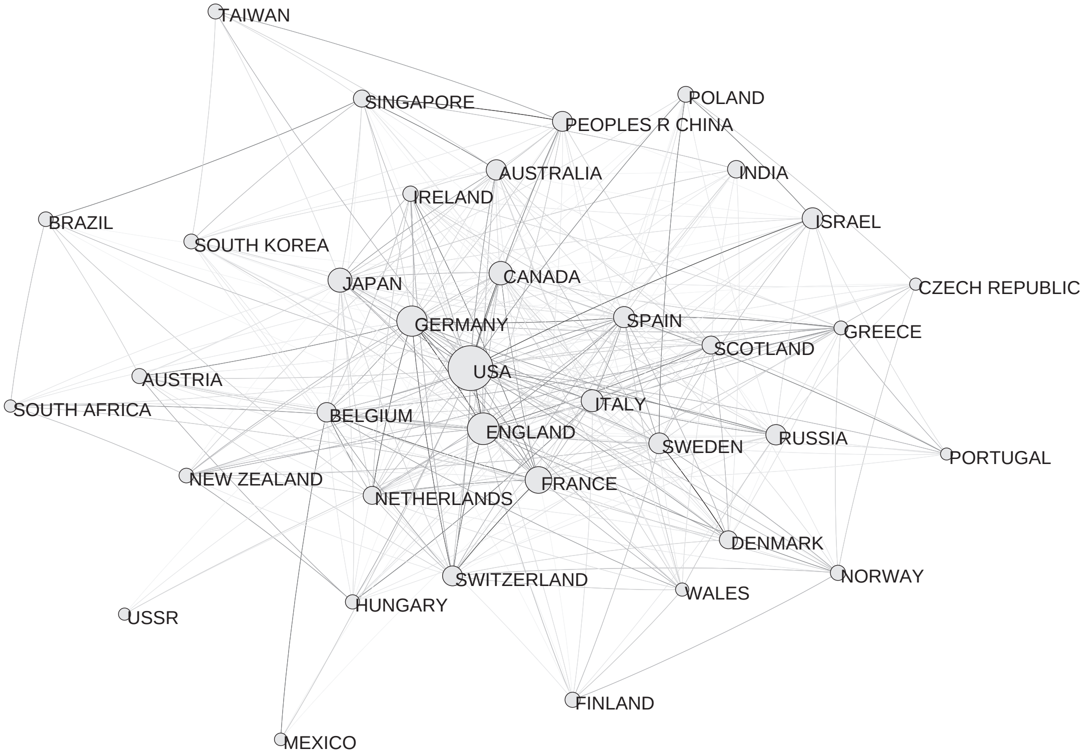


**Figure 1.8:** International collaboration in bioinformatics (based on 7401 publications). The larger a node that represents a country, the more publications the country has contributed to. The length of an edge (connection) between two countries represents the number of publications for which both countries have collaborated (mutual co-authorship). In general, countries that are close to each other in the network, have intensely collaborated. The *big* countries, USA, UK, Germany, France and Japan can be found in the center of the diagram. Since the figure is based on all bioinformatics papers retrieved for 1980–2004, both the USSR and Russia appear in the diagram.

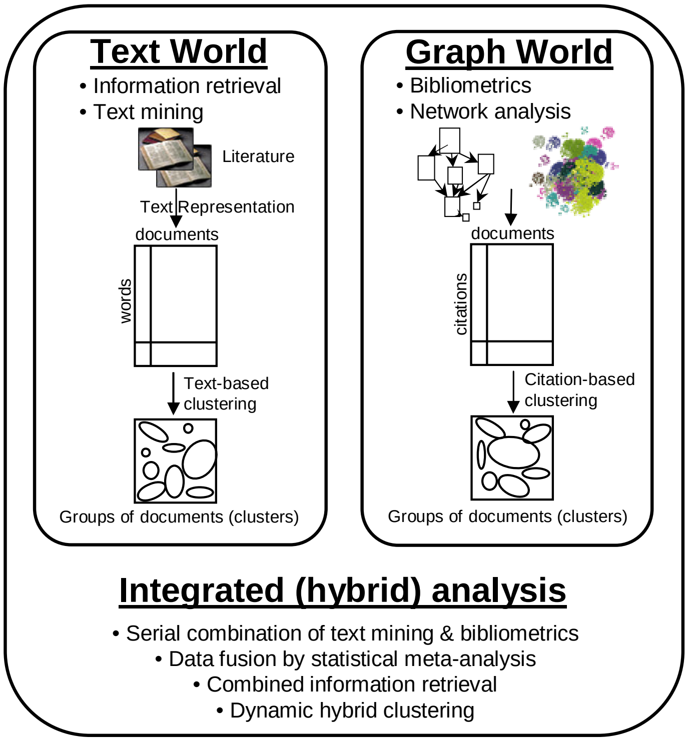

Text World
• Information retrieval
• Text mining
Literature
Text Representation
documents
words
Text-based clustering
Groups of documents (clusters)
Graph World
• Bibliometrics
• Network analysis
documents
citations
Citation-based clustering
Groups of documents (clusters)
Integrated (hybrid) analysis
• Serial combination of text mining & bibliometrics
• Data fusion by statistical meta-analysis
• Combined information retrieval
• Dynamic hybrid clustering

**Figure 1.9:** The general scope of this dissertation is the mapping of scientific and technological fields by using clustering algorithms and techniques from text mining and bibliometrics. The distinction between *text world* and *graph world* refers to different views on a collection of interlinked publications. In addition to textual information in such documents, citations between them also constitute large networks that yield additional information. Both complementary approaches have advantages and intricacies and both can be used to subdivide groups of publications in clusters or groups of documents. In an integrated or hybrid analysis we incorporate both points of view and show how to improve on existing text-based and graph analytic (or bibliometric) methods by deeply merging textual content with the structure of the citation graph. In subsequent chapters we will show the same figure while highlighting relevant parts.

## 1.2 Motivation: text world vs. graph world

The distinction between *text world* and *graph world* refers to different views that one can cast on an interlinked data collection such as the World Wide Web and bibliographic databases containing written scientific communications. On the one hand, these documents contain textual information that can be mined for knowledge by using text mining techniques. On the other hand, each document refers to other documents that are in some way related (see also Figure 1.9).

Most scientific work indeed cites previous research on which it is based or which is considered to be relevant for the subject. These *citations* are collected in the *bibliography*[6] of a publication. Although various reasons are conceivable for citing other work, citations usually imply endorsement or recommendation of the previous work.

All citations among publications or hyperlinks among Web pages constitute extremely large networks, of which the World Wide Web is the biggest example. In Figure 1.10, a very small network contains a few scientific papers and some citations. Contrary to the Web, in which each Web page can have hyperlinks to any other page, a citation network or literature network is (approximately) a directed *acyclic* graph (DAG). Citations and hyperlinks each have a direction (they *point* from one entity to another), but citations are not reciprocal and no directed cycles occur in the citation graph. Usually, a scientific paper only cites documents that have already been published before.

A lot of patents have references as well, either to other patents or *non-patent references* to scientific publications or other reports. The interesting connection between science and technology is, however, beyond the scope of this dissertation. Because of the duality in the term *citation* (each *cited reference* entails a received *citation* for the cited entity), a distinction is generally made between *backward* and *forward* citations in patent analysis. References in patents are usually made by both inventors and examiners [181]. In webometrics, hyperlinks are usually referred to as *out-links* and *in-links.*

Both the textual and graph-based approaches have advantages and intricacies and provide other views on the same data; for example, different perceptions of similarity between documents or groups of documents, and different methods to observe dynamics in evolving databases. We incorporate both viewpoints and claim to improve on existing text-based and graph analytic or bibliometric methods to science and technology mapping. Textual information can indeed indicate similarities that are not visible to bibliometric techniques. Based on text alone, true document similarity can be obscured by differences in vocabulary use, or spurious similarities might be introduced as a result of textual pre-processing, or because of polysemous words (a word with several meanings) or words with little semantic value. For instance, documents about music information retrieval might erroneously be linked to patent-related research based on common terms that are used in both contexts, such as *title, record, creative,* and *business.*

Another illustrative example is given in Figure 1.11, in which scientific papers (*nodes* or *vertices* in the citation network) are represented as circles. Although both highlighted papers are not related in subject, automatic text mining algorithms might yet perceive similarity because of a lot of occurrences of the same term *nano* in both papers (after pre-processing). Fortunately, by observing the neighborhood graph of both publications, it is obvious that both reside in different subject domains. Likewise, if two competing organizations both

[6] A bibliography contains the list of *cited references.*

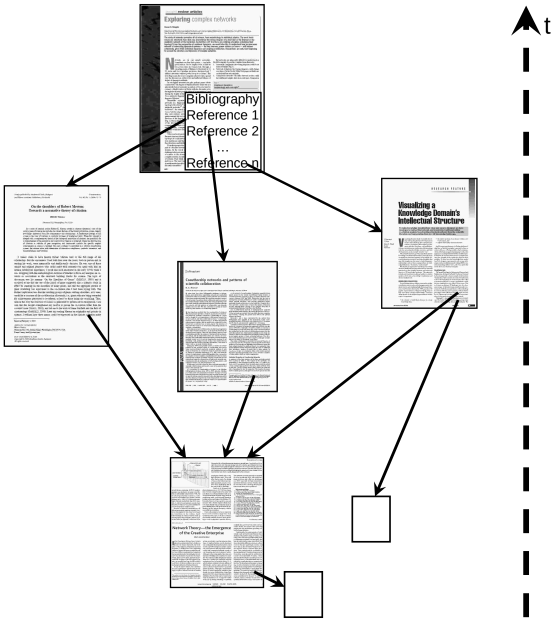

Exploring complex networks
Bibliography
Reference 1
Reference 2
...
Reference n
On the shoulders of Robert Merton: Towards a normative theory of citation
Coauthorship networks and patterns of scientific collaboration
Visualizing a Knowledge Domain's Intellectual Structure
Network Theory—the Emergence of the Creative Enterprise
t

**Figure 1.10:** Small illustrative extract of a citation network consisting of scientific publications and citations among them. Citations have a direction and are represented as an arrow from one publication to another. A scientific paper can only cite documents that have been written in the past. Hence, when the citation network grows in time ($t$), a hierarchy of papers is formed. The figure shows only 7 publications and 8 citations, but real networks can grow extremely large. For the field of bioinformatics, we considered a citation network with about 8000 publications and 67 000 citations. The network was also extended with all publications that cited at least one of these bioinformatics publications, and with all publications that were cited by those articles. The resulting network contained about 261 000 publications and 586 000 citations.

publish related articles, but never cite each other's work, text-based methods can correctly identify similarity.

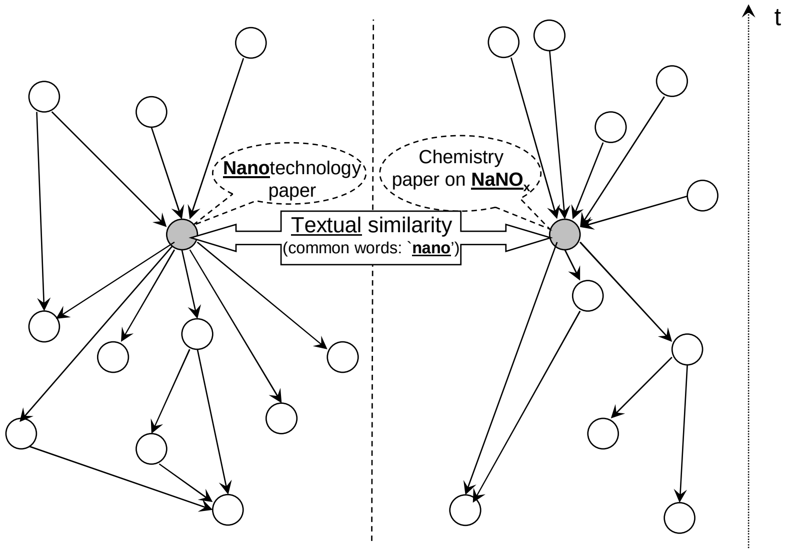


**Figure 1.11:** Illustration of the motivation for our quest for integrated (hybrid) mining algorithms. A small extract of a citation network is shown. Circles represent scientific publications or patents and arrows represent citations between them. We consider two publications in gray, one is about nanotechnology and the other one is a paper about chemistry (sodium nitrate or $NaNO_3$). Automatic text mining procedures might consider both publications to be related in subject since both contain the same keyword *nano*. Indeed, after automatic pre-processing, the chemical formula $NaNO_3$ might be reduced to the same term 'nano'. However, by considering the citation network, it is clear that both highlighted papers are probably not related since they reside in a different neighborhood or community of the graph. There are no common references or common citing papers between both neighborhoods. Hence, a hybrid analysis of both the *text world* and the *graph world* might provide more accurate perception of topical similarity of publications.

Information Retrieval (IR) provides yet another example in which the text world and graph world have complementary qualities and for which a combined approach clearly yields a great advantage. IR algorithms used by early Internet search engines only considered the textual content of Web pages in order to determine relevance with respect to a user's query. Only since the end of the previous millennium, large-scale search engines started to exploit the link structure of the Web as well, the most famous example is *Google*'s PageRank algorithm which considers hyperlinks to determine the *quality* of Web pages. A Web page that is referred to by many other good Web pages can be considered an authority in its subject domain and should thus be ranked higher in the often very large result sets.

In conclusion, hybrid methods that exploit both text and link analysis are assumed to achieve better results than pure text-based or link-based methods. In this dissertation, we demonstrate the complementarity of both paradigms, we devise a hybrid approach that considers both worlds and we claim that with the integrated stance we attain a better interpretation of the underlying structure and dynamic properties of large-scale *corpora* containing publications. The following subsections give a more detailed introduction to the text world and the graph world.

### 1.2.1 Text world

The use and power of text mining techniques for automated retrieval of information and for mapping or charting of knowledge embedded in texts is the subject of Chapter 2. These techniques are becoming increasingly important in the light of the overwhelming amount of textual information available, even more so since the advent of the Internet, massive databases, email archives, powerful search engines, and recent phenomena such as semantic wikis, blogs, e-books and machine generated data. Today, text mining is even used for emerging trend detection, policy-making processes, intelligence services, press monitoring to automatically detect breaking news, marketing, data protection, law enforcement, personalized advertising, etc.

Given the enormous and ever-increasing size of contemporary databases, the applicability of mining algorithms on large collections is a matter of concern. With scalability and complexity issues in mind, and given the fact that statistical and mathematical methods can provide surprisingly good results when turning data into information into knowledge, we do not make use of advanced or ‘deep’ natural language processing (NLP) techniques. The adopted algorithms will mainly consider the (co-)occurrence of words in texts and as such will neglect parsing down to the level of the clause. The linguistic structure of sentences, word order and other important aspects of human discourse are thus disregarded. Despite these rather naive simplifications, we demonstrate that the applied statistical and mathematical techniques are very powerful and scalable. We do, however, make use of *shallow parsing* techniques as a means to filter important terms and to detect phrases or composite terms. Shallow parsing is an NLP technique to algorithmically analyze sentences and to annotate words and word groups with *part of speech* tags that identify nouns, verbs, adjectives, etc.

### A concise overview of the application of quantitative linguistics in informetrics and bibliometrics

Quantitative linguistics dates back to at least the middle of the 19th century [114]. However, the classical theoretical work by *Zipf* (1949) is considered pioneering in quantitative linguistic analysis [277]. Since the 1970s, a remarkable increase in interest has been observed for this topic of information science. *Wyllys*' study is one of the first in its application to scientific literature [271].

At present, the most frequent techniques are co-word, co-heading and co-author clustering. They are based on analysis of co-occurring keywords, terms extracted from titles, abstracts and/or full text, subject headings or cited authors. The method was developed by *Callon et al.* more than two decades ago, for purposes of evaluating research [44]. The methodological foundation of co-word analysis is the idea that the co-occurrence of words describes the contents of documents. By measuring the relative intensity of these co-occurrences, simplified representations of a field's concept networks and their evolution can be illustrated [43].

*van Raan* and *Tijssen* have discussed the potential of bibliometric mapping or charting based on co-word analysis [261]. Many researchers have used this methodology to investigate concept networks in different fields, among others, *de Looze* and *Lemarie* in plant biology [57], *Bhattacharya* and *Basu* in condensed matter physics [26], *Peters* and *van Raan* in chemical engineering [261], *Ding et al.* in information retrieval [69] and *Onyancha* and *Ocholla* in medicine [209]. Co-heading analysis was introduced by *Todorov* and *Winterhager* [253].

The extension of co-word analysis towards the full texts of large sets of publications was possible as early as large textual databases became available in electronic form. The descriptive power of controlled terms or of the vocabulary used by authors to summarize their work in title and abstract, makes it possible to use text mining and co-word analysis as sophisticated tools both in structural [252] and dynamic bibliometrics [278, 279]. Nonetheless, the added value of full text with respect to title and abstract information can be high; *Glenisson et al.* [110, 109] and *Shah et al.* [241] have found that the use of full text included more relevant phrases for interpretation.

Co-word analysis has recently also become the preferred tool for the mapping of science at CWTS (Leiden, the Netherlands), where bibliometric mapping is used within a science policy and research management context [206]. The shift from co-citation analysis to co-word techniques allows application to non-citation indexes as well.

The statistical analysis of natural language has a long history. *Manning* and *Schütze* have provided a comprehensive introduction [174], *Berry* has provided a survey of text mining research [20], and *Moens* has discussed the automatic indexing and abstracting of document texts [187]. For science and technology research, *Leopold et al.* have given an overview of data and text mining fundamentals [163]. *Porter* and *Newman* coined the term 'tech mining' for text mining of collections of patents to support technology management [224].

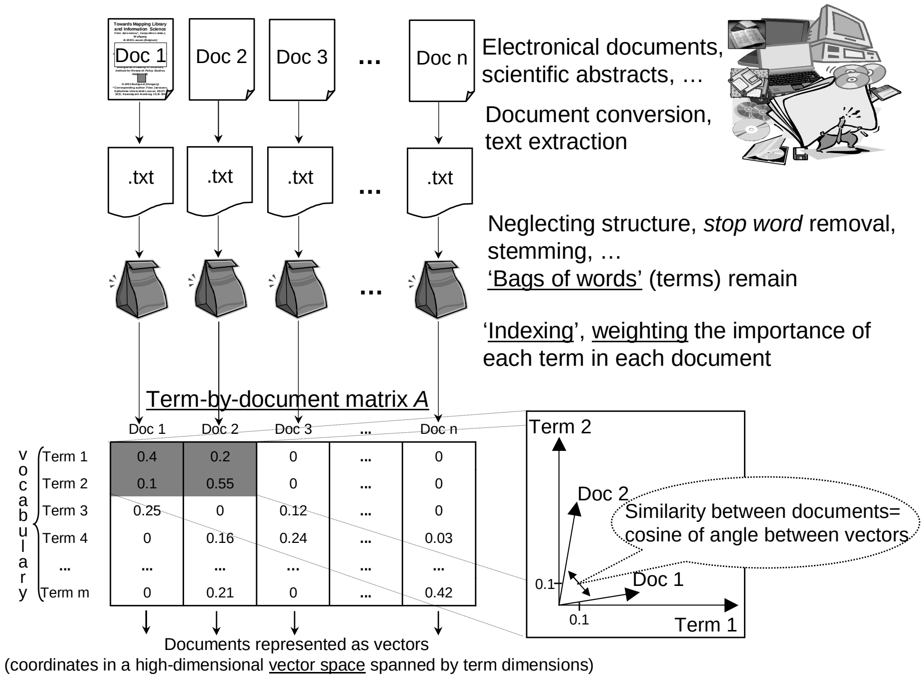
**Representation of textual data**

The ability to mine vast amounts of text presupposes that the textual data is represented in a machine-readable format. Once a suitable representation has been defined, it can be used to hold information and to extract knowledge from a plethora of different formats of textual entities (such as collections of e-mail messages, Web pages, documents written by humans or generated by machines, and abstracts or full-texts of patents and scientific publications). These data sets or corpora might be stored locally or be accessible via networks. The downside of this versatile applicability of text mining techniques is the huge amount of pre-processing steps needed before the actual algorithms can be applied to a specific corpus. In our experience, the labor put in the pretreatment of textual documents usually encompasses the largest part of the total mining effort. Fortunately, the process is modular, so reusability of components is guaranteed. For example, a convertor for text extraction from various file formats. In a certain project, we analyzed a data set containing 19 940 documents (34 GB) available on the intranet of a company. Extraction of the text from different file formats was the most lengthy task. On a machine with a 2.8 Ghz processor and 4 GB of internal memory, text extraction took about 14 hours. The use of more sophisticated tools can speed this up since we used freely available software. Conversion of *.pdf* or *.ps* documents took about 0.3 seconds on average, but conversion of *.doc* or *.rtf* documents took 15 seconds per file. For final indexing of the complete data set just about 16 minutes were needed.

Figure 1.12 provides a schematic visualization of part of the text mining methodology that is adopted in our research. Almost all of the techniques are tailored to the analysis of unstructured data, except for the occasional separate analysis of different fields available in database records. *Unstructured data* denotes data that is not in a predefined format, particular template or fixed classification. It is not restricted to terms from a closed taxonomy or ontology, but possibly residing in a chaotic environment such as a company's intranet or on the Internet. Of course, structure in texts can always be neglected and our mining techniques can thus also be applied on structured data, but we do not make use of algorithms specifically directed towards structured data. An advantage is that no categories nor any rigid form need to be defined beforehand, but the complementary power of structured data is of course beyond dispute. A gain in strength might be made possible by combinations of both structured and unstructured methods, for example, in the form of dynamic, adaptive structures.

### 1.2.2 Graph world

Our contemporary world is characterized by ever increasing interlinking. Networks can be of various types, such as physical, infrastructural networks for transportation and for provision of electricity, gas and water, that become more and more observable and controllable by cheap sensors and actuators. In addition, telephone networks and other ICT networks are of paramount importance


Doc 1
Doc 2
Doc 3
...
Doc n
Electronical documents, scientific abstracts, …
Document conversion, text extraction
.txt
.txt
.txt
...
.txt
Neglecting structure, *stop word* removal, stemming, …
'Bags of words' (terms) remain
'Indexing', weighting the importance of each term in each document
Term-by-document matrix *A*
vocabulary
Term 2
Doc 2
Similarity between documents= cosine of angle between vectors
0.1
Doc 1
0.1
Term 1
Documents represented as vectors
(coordinates in a high-dimensional vector space spanned by term dimensions)

| | Doc 1 | Doc 2 | Doc 3 | ... | Doc n |
|---|---|---|---|---|---|
| Term 1 | 0.4 | 0.2 | 0 | ... | 0 |
| Term 2 | 0.1 | 0.55 | 0 | ... | 0 |
| Term 3 | 0.25 | 0 | 0.12 | ... | 0 |
| Term 4 | 0 | 0.16 | 0.24 | ... | 0.03 |
| ... | ... | ... | ... | ... | ... |
| Term m | 0 | 0.21 | 0 | ... | 0.42 |

**Figure 1.12:** Automatic processing of digital documents and their representation in the *Vector Space Model*. The text from all $n$ documents at the upper part of the figure is automatically extracted. All word order is neglected, which can be interpreted as putting all words of each document in a separate bag (hence the name *bag of words* representation). All words in each bag (document) are counted (in a process called *indexing*) and the resulting numbers are stored in a *term-by-document* matrix. In such a matrix, each row represents a term (or word), and each column represents one of the documents. The total set of $m$ words that occur in any document is referred to as the *vocabulary*. A value $w_{i,j}$ on row $i$ and column $j$ in the matrix represents the number of times word $i$ occurs in document $j$, usually weighted by an extra *weighting scheme*. Each document (column) can be considered as a vector or coordinate in a high-dimensional vector space in which each dimension represents one term. For example, in the lower right corner of the figure, the vectors for the first two documents are shown in the two-dimensional space spanned by the first two terms. A computer can then measure topical similarity of both documents by calculating the angle enclosed by both vectors. The smaller the angle, the more related the documents are. Given the astronomic number of digital documents available, a *term-by-document* matrix can be huge. $n$ can be in the order of millions. The size $m$ of the vocabulary is bounded by the total number of distinct words or other tokens, such as project numbers or names, that are encountered in the text collection. The total vocabulary can contain hundreds of thousands of items, but strongly depends on pre-processing strategies. We have indexed abstracts of millions of scientific publications and patents, but the largest number of documents used for case studies in this dissertation is about ten thousand, with a vocabulary size of about twenty thousand.

in our daily lives. The growth of the Internet and of wireless communication networks is striking. On the other hand, in our networked society, we, as social beings, are plugged into and actively collaborate in various other types of *social* networks. In scientific communication, for instance, networks emerge from collaborations among (groups of) people. Other examples of networks represent communication acts (phone, email, etc.), social communities on the WWW, organizational networks, networks of knowledge such as *Wikipedia*, and many biological and biochemical networks (for example, neural systems or protein interactions). Networks can be conceptualized for any type of collaborative, transactional, or affinity information and can thus also be built from textual information.

Techniques from bibliometrics and graph theory can be used to analyze networks that emerge from many individual acts of authors reading and citing other scientific works. These extremely large networks can be examined in order to rank relevant entities, for clustering, extraction of communities, collaborative filtering, etc. The science of evolving networks can even contribute to detection of emerging and converging clusters representing scientific specialties, new technologies and hot topics.

## 1.3 Clustering

Our efforts to combine text and graph worlds into a hybrid analysis will mainly focus on clustering algorithms. Figure 1.13 presents an overview of some important aspects of clustering. *Clustering* is a multivariate statistical technique for automated grouping of objects (for instance, vector representations of documents) such that similar objects are put in the same group or *cluster*, while dissimilar objects end up in different clusters as much as possible. The similarity of two objects is defined by an objective formula that considers known properties of each object. For documents, similarity is measured by considering the amount of words two documents have in common. In case of a lot of shared words, it is assumed that both documents discuss the same topic (see Fig. 1.12).

Clustering belongs to the *unsupervised learning* paradigm in the sense that the algorithm tries to partition objects in the optimal way according to some validation measure, merely based on data representations without any knowledge of group membership. *Classification* algorithms work in a supervised manner. They are presented with the correct class information for objects in a *training set*, which they use to learn a model for classifying previously unseen examples. The use of *validation* and *test sets* is very important for correct evaluation of the classifier. Otherwise, *overfitting* might occur, which means that the trained classifier is very good at classifying the objects in the data set at hand, but does not generalize well to unseen cases [184].

We are interested in unsupervised clustering rather than building an optimal classifier for assigning documents to predefined categories, since an accurate classification of scientific articles—needed for training a classifier—is not avail-

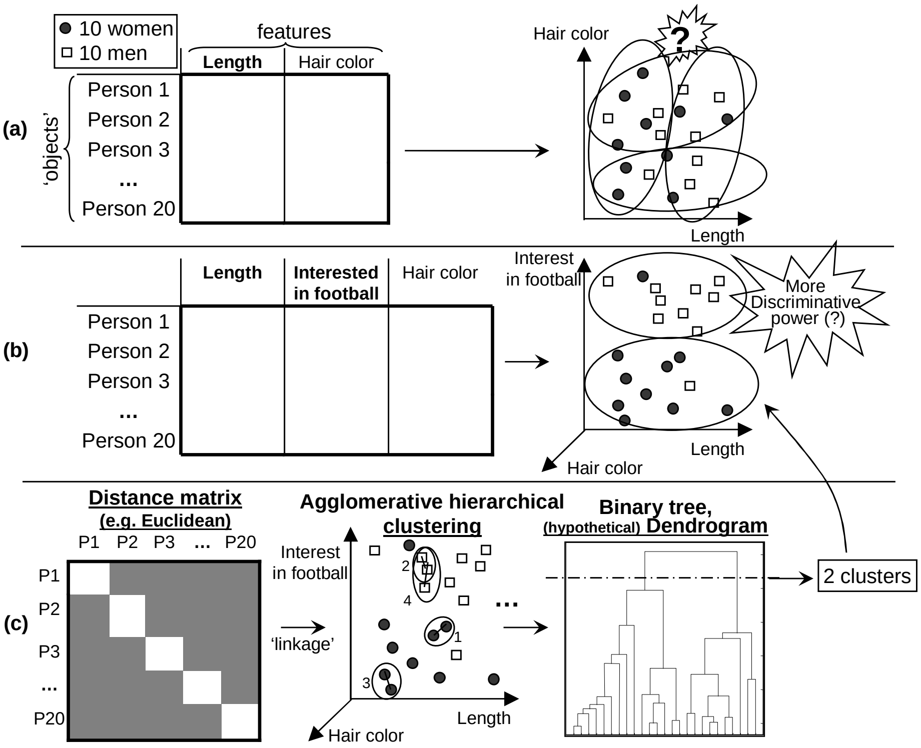


**Figure 1.13:** Overview of agglomerative hierarchical clustering. Suppose we would like to find two groups (clusters) of people, 10 women and 10 men, but that gender information is not known. Usually, even the number of groups to find is not known in advance. The goal of a clustering algorithm is to automatically divide the 20 persons into groups based on known *features* (dimensions) and to make the groups as homogeneous as possible. People with similar features should be put in the same group and the dissimilarity between groups should be as high as possible. In **(a)**, only the features *length* and *hair color* are known about each person. It is very difficult to find homogeneous groups based on these features since *hair color* provides no information to discriminate between women and men and *length* does not provide sufficient information. In **(b)**, the feature *interested in football* is also known. It has more *discriminative power*: the distinction between men and women is much more clear. Of course, there still are *outliers*: some women do like football very much, while some men don't. **(c)**. Most clustering algorithms compute pairwise (mutual) distances between all 'objects' based on their features. These distances are stored in a *distance matrix*. The *Euclidean* distance that we use in our daily lives can also be used to measure distance between objects described by a set of selected features. There are other *distance measures* as well. Starting from singleton clusters (each object represents one cluster), *agglomerative hierarchical clustering* proceeds by iteratively grouping those objects or clusters that are least distant from each other according to a *linkage* criterion (see Figure 1.15). This iterative merging continues until all objects are in one big cluster. A *dendrogram* provides a visualization of this process. This hierarchical tree can be 'cut off' at any level to provide different levels of aggregation at which the objects can be subdivided into groups. In this example, the tree is cut off at 2 clusters.

able. Existing classifications are indisputably outdated because of the dynamic nature of contemporary science and technology. However, in some experiments we do consider a classification setting since it offers a well-grounded basis for assessing relative performance.

### Data representation

The first necessary condition for clustering is the availability of a suitable abstract representation of all objects, for example, by encoding in the Vector Space Model (see Figure 1.12). Here, each object (for example, a document) is characterized by a set of weighted features (terms), indicating the importance of each feature for each object. Selection of the most valuable features and construction of new features prior to clustering are important issues and can have an influence on the quality of the outcome that should not be underestimated.

Next, for many clustering algorithms the requisite input includes mutual distances between all objects, stored in a symmetric, square distance matrix. An appropriate distance metric is needed for measuring dissimilarity between the mathematical representations of a pair of objects. Figure 1.14 provides a visualization of similarity or correlation matrix calculation. The distance matrix can then be obtained by subtracting each entry from 1.

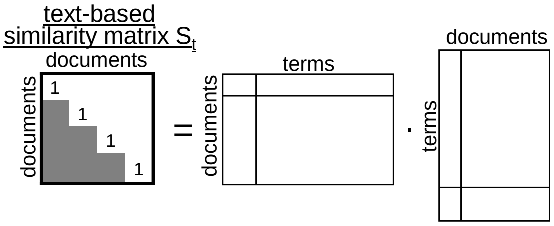


**Figure 1.14:** Construction of a text-based document similarity matrix $S_t$ from a *term-by-document* matrix with normalized columns. Each element in $S_t$ is the cosine of the angle between two document vectors.

### Algorithms

Hierarchical clustering algorithms group objects in an iterated manner to construct a binary tree, either starting from singleton clusters to the trivial cluster containing all objects (*agglomerative* clustering), or vice versa (*divisive* clustering). The leafs of the tree represent the objects (documents), whereas the different branches show the grouping of objects or sub-clusters into larger clusters. The strategy used to measure the distance between clusters and hence to determine which objects or clusters to group in each iteration affects the result (see Figure 1.15). *Single linkage* (nearest neighbor) defines the distance between two clusters as the smallest distance between any two points from both clusters, whereas *complete linkage* (furthest neighbor) considers the maximal distance between any two points from both clusters. The more advanced UPGMA (unweighted pair group method using arithmetic averaging), also referred to as *group average*, calculates the distance between clusters as the weighted average of all mutual distances between objects from both clusters. The result is *unweighted* because of the equal contribution of each distance. In the even more complicated method of *Ward*, at each iteration step those objects are grouped such that the increase in total within-cluster sum of squares over all clusters is minimized [267, 133, 146]. We will mainly use *Ward*'s method and UPGMA.

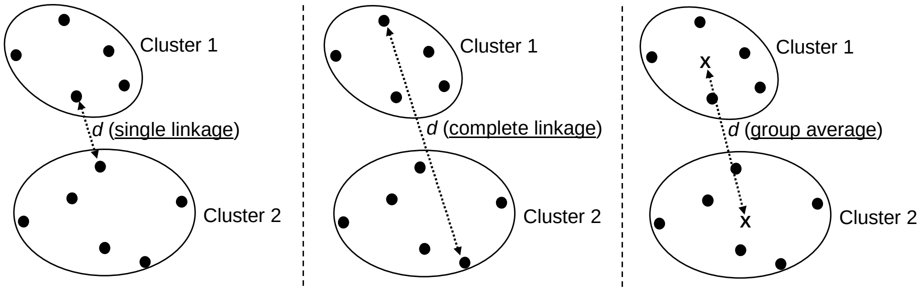


**Figure 1.15:** Linkage in hierarchical clustering. *Single linkage* (nearest neighbor) defines the distance $d$ between two clusters as the smallest distance between any two points from both clusters, whereas *complete linkage* (furthest neighbor) considers the maximal distance between any two points from both clusters. The computationally more expensive *group average* method calculates the distance between clusters as the weighted average of all mutual distances between objects from both clusters. This mean distance between all possible pairs of elements of both clusters is visualized as the distance between the cluster centers ($x$).

### Hybrid clustering

In this dissertation we devise a methodology for deeply combining text mining and bibliometrics or network analysis. In particular, multiple information sources are incorporated before the clustering algorithm is applied. The actual integration is achieved by combining various distances between the same pair of documents. Each distance results from possibly different distance measures exploiting different views on the documents. Mutual document distances can be based on textual content, on citations or bibliometric indicators, or on a combination of any of these information sources.

We describe weighted linear combination of distance matrices as well as an integration method based on *Fisher*'s inverse chi-square. Both methods can be considered *intermediate* integration methods: mutual document similarities are calculated in separate spaces, but integrated before application of the clustering algorithm. We also experiment with *early* integration methods that incorporate data even before distance calculation (for example, by appending vectors).

Hybrid clustering by intermediate integration is illustrated in Figure 1.16.

Important in any clustering effort is validation of results and determination of the number of clusters that best capture existing subjects. As indicated in the figure, we use a combination of four clustering evaluation methods.

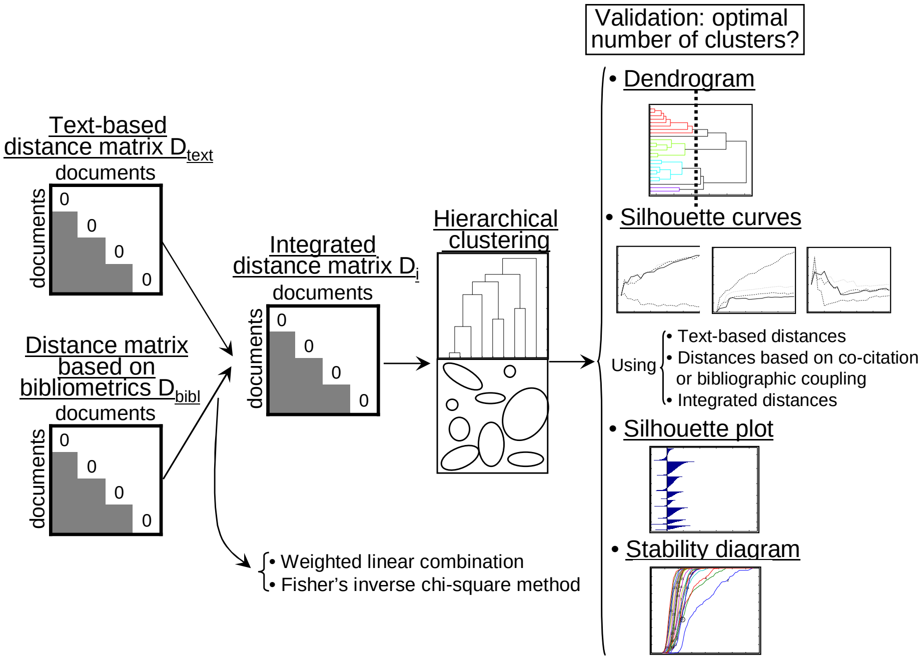

Text-based distance matrix $D_{text}$
documents
documents
0
0
0
0
Distance matrix based on bibliometrics $D_{bibl}$
documents
documents
0
0
0
0
Integrated distance matrix $D_i$
documents
documents
0
0
0
0
Hierarchical clustering
• Weighted linear combination
• Fisher's inverse chi-square method
Validation: optimal number of clusters?
• Dendrogram
• Silhouette curves
Using
• Text-based distances
• Distances based on co-citation or bibliographic coupling
• Integrated distances
• Silhouette plot
• Stability diagram

**Figure 1.16:** Hybrid hierarchical clustering by *intermediate integration* and evaluation of results to determine the most natural number of clusters. In *hybrid* or integrated clustering, pairwise distances between documents are based on information from the text world (text-based distances, cf. Figure 1.12) as well as on information from the graph world. *Intermediate integration* refers to the fact that distances are first computed in both worlds separately, after which they are integrated, before application of the hierarchical clustering algorithm. Hence, text-based and link-based distance matrices are combined in a mathematical or statistical way before being used as input for the hierarchical clustering algorithm (see Figure 1.13 for information on hierarchical clustering). We will mainly use weighted linear combination of distance matrices as well as a method based on Fisher*'s inverse chi-square* to integrate distance matrices. The number of clusters (that represent different subjects) in a document set is determined by using various methods. The first set of methods evaluate the homogeneity of clusters and the separation between clusters based on statistical formulas that consider all distances within and between clusters. Another method considers statistical *stability* of clusters by measuring whether the same objects always end up in the same cluster when the clustering is computed multiple times on slightly different data sets. Each of these methods will be treated in detail in chapter 2.

## 1.4 Contributions

We investigate whether algorithmic and multivariate statistical processing of large collections of scientific literature can uncover and describe the topic structure of scientific fields at different levels of aggregation, to provide insight into how different subfields of science interact.

Challenges addressed in this dissertation comprise the development of a data mining framework able to manage corpora of structured and unstructured scientific information. Algorithms should be able to cope with high dimensionality in terms of number of publications as well as the number of features used to describe them. For instance, the total number of words that occur in all documents.

Our main contributions are the following:

- ***Hybrid clustering.*** In previous sections we have discussed the complementarity of textual content and the structure of literature networks in the context of clustering of scientific fields. The availability of heterogeneous information is a great asset and we believe that it can be exploited in a robust integrated manner to improve on individual techniques. However, it is a major challenge to properly use both information sources in an integrated analysis. Primitive integration schemes might neglect important different properties of both worlds. We hypothesize that careful statistical integration contributes to the quality of the ultimate result. We test and assert this by assessing the quality of clustering and classification results by measuring objective statistical homogeneity of clusters (*Silhouette coefficient*) and by contrasting classification results with a 'ground truth' or 'gold standard' classification based on expert knowledge contained in Medical Subject Headings that are annotated to publications (MeSH[7] terms).

  Hence, we demonstrate the complementarity of text mining and bibliometric methods, and propose schemes for a sound integration of both worlds. We confirm the hypothesis that such an integrated or hybrid clustering approach allows better comprehension of content structure and dynamic properties of textual corpora and thus provides more accurate mappings. In particular, we assert that the performance of unsupervised clustering and classification of scientific publications is significantly improved by merging textual content and citations.

  Besides fusing text and citations by means of Random Indexing, we propose a hybrid clustering method in which pairwise (mutual) distances between documents are converted to $p$-values with regard to randomized data, and in which *Fisher's inverse chi-square method* is consecutively used to combine $p$-values from various origins. This method can incorporate distances stemming from different metrics with highly dissimilar distributional characteristics and avoids domination of any information

[7] http://www.nlm.nih.gov/mesh/, visited in January 2007.

source. This hybrid clustering approach integrates text mining and bibliometrics and significantly outperforms text-based and citation-based solutions.

A combination of text-based and bibliometric components also improves the complex *delineation* or *demarcation* of interdisciplinary research fields such as bioinformatics. The delineation of a research field involves the application of several information retrieval strategies in order to construct a set of publications highly relevant for the topic of interest. This is far from trivial given the interdisciplinary nature of scientific fields and dissemination via various channels (possibly multidisciplinary journals).

- ***Dynamic hybrid clustering.*** We develop a flexible methodology for hybrid dynamic analysis of evolving bibliographic data sets by matching and tracking clusters through time.

- ***Optimal number of clusters.*** The level of aggregation at which a document set should be subdivided into groups is difficult to determine. Various algorithms and several statistical measures for evaluation and validation are available for this purpose, but they do not necessarily agree on the 'best' number of clusters. Nevertheless, we believe that combined use of various methods provides useful indications for the natural number of clusters. We describe a combination of several strategies, which comprises distance-based and stability-based methods. Text world and graph world also provide complementary means for evaluation of the number of clusters.

- ***Number of factors in Latent Semantic Indexing.*** Latent Semantic Indexing (LSI) is a dimensionality reduction technique based on the Singular Value Decomposition (SVD) of a *term-by-document* matrix. The reduction of the number of features in a vector space improves the performance of clustering and classification algorithms since a lot of algorithms are meaningless in high dimensional spaces. However, it is not straightforward to decide on the number of dimensions to retain. We believe and assert that an appropriate choice can highly affect and improve clustering performance. We contribute to an important open research problem in LSI research, namely the debate about the number of factors. We investigate the relationship between number of factors, number of clusters and clustering performance. In our data sets, the quality of clustering proves significantly higher for a smaller number of factors, when quality is measured with the *Silhouette coefficient*. In spite of the fact that a number of dimensions between 100 and 300 is often assumed in literature to be a good choice, we show that a very modest number of factors (e.g., 10) can provide the best clustering performance, on condition that there are no fewer LSI factors than the desired number of clusters. This should, however, be further assessed using other corpora as well.

- ***Mapping library and information science (LIS).*** First, we focus on the full textual content of 938 publications from a set of 5 journals, excluding any bibliographic or bibliometric component which might influence the quantitative linguistic analysis of the scientific text. A number of 6 clusters seemd to be the optimum solution. Nevertheless, hybrid clustering yields a better mapping in quantitative and qualitative sense by merging two clusters on bibliometrics.

- ***A structural and bibliometric domain study of bioinformatics.*** Our hybrid clustering procedure based on *Fisher*'s inverse chi-square method is adopted to unravel and visualize the concept structure of bioinformatics. The delineation of the field (7401 publications) is achieved by *bibliometric retrieval*. The strategy for defining the number of clusters suggests nine subdisciplines. For each cluster we provide term and collaboration networks as well as the most representative publications. In addition, we investigate journal coverage, evolution of publication output, evolution of citation impact, cluster representation of the 5 most active countries, as well as collaboration at different levels of aggregation. Next, we analyze *'naive' dynamics* of each cluster, which means that publication years are not considered during clustering, but only afterwards. Finally, seven consecutive periods with approximately the same number of publications are defined for dynamic hybrid clustering.

In conclusion, statistical and mathematical techniques from text mining, bibliometrics and link analysis prove to deliver powerful methods for mapping knowledge embedded in bibliographic databases. The proposed hybrid clustering algorithms which exploit information from both *text world* and *graph world*, will provide accurate means to unravel the structure and evolution of dynamic document sets. The strategy of tracking clusters through time facilitates detection of emerging or converging clusters and hot topics.

In this dissertation we mainly focus on clustering of scientific and technological fields. Although not further mentioned explicitly, most of the techniques and algorithms are generic and can be used in other application domains as well. During the years of this Ph.D. research they have been applied in a corporate knowledge management project [263, 262]. A straightforward extension is the analysis of sets of Web pages connected by hyperlinks. Other examples for which link structure can be analyzed in combination with textual content are networks of knowledge such as *Wikipedia*, semantic wikis, Web logs, newsgroups, and e-mail archives. Web pages or documents often consulted together might also be arranged in a network allowing an integrated analysis.

Technical and methodological issues are in the foreground. We provide a set of tools and solutions that are based on existing techniques, new for their combined application in bibliometrics. Various exploratory experiments demonstrate the application and power of algorithms, but experts remain indispensable for interpretation of results and for deciding on data collection and preprocessing strategies.

We have also contributed to clinical classification of microarray data by developing the M@CBETH Web service (a MicroArray Classification BEnchmarking Tool on a Host server), offering the microarray community a simple tool for making optimal two-class predictions. M@CBETH aims at finding the best prediction among different classification methods by using randomizations of the benchmarking data set [223].

## 1.5 Dissertation structure

An overview of the several chapters in this book is given in Figure 1.17. The current chapter (**Chapter 1**) gives an introduction and motivation by distinguishing the text and graph worlds and by introducing the hybrid clustering methodology. Main contribitions of this work are discussed in this chapter as well.

Chapters 2 and 3 describe both respective worlds in more detail. **Chapter 2** discusses the use of text mining techniques for mapping knowledge domains. Representation of textual data in the Vector Space Model and the text mining framework are explained. The *curse of dimensionality* and the necessity of dimensionality reduction by feature selection and methods such as Latent Semantic Indexing (LSI), Random Indexing or multidimensional scaling (MDS) will receive ample treatment. Next, the adopted hierarchical clustering algorithm and validation measures are addressed, as well as a combined strategy for detecting the number of clusters by distance-based and stability-based methods. We contribute to the debate about how to choose the number of latent semantic factors in LSI, in relation to the number of clusters and clustering performance. Introduced algorithms are demonstrated in two case studies, in which any bibliographic or bibliometric components are not taken into account as they might influence the quantitative linguistic analysis. The first study is about biotechnology in Belgium, the second one treats the field of library and information science (LIS).

**Chapter 3** focuses on the analysis of networks that emerge from authors citing other scientific works or collaborating in the same research endeavor. First, we introduce and briefly describe a selection of bibliometric indicators, graph analytic techniques and types of networks that we consider. Next, we discuss the HITS and PageRank algorithms used in information retrieval for identification of important authorities and hubs, and for ranking result sets. We touch upon graph partitioning and detection of communities. Finally, a selection of bibliometric techniques is applied to the bioinformatics field.

**Chapter 4** investigates possible ways to incorporate bibliometrics of Chapter 3 with text mining algorithms of Chapter 2, in order to come up with a hybrid methodology for information retrieval and for mapping of fields of science. First, complementarity of both techniques is shown by a serial combination. Next, we devise a methodology for combining text mining and bibliometrics by integrating text-based and bibliometric information early in the mapping process.

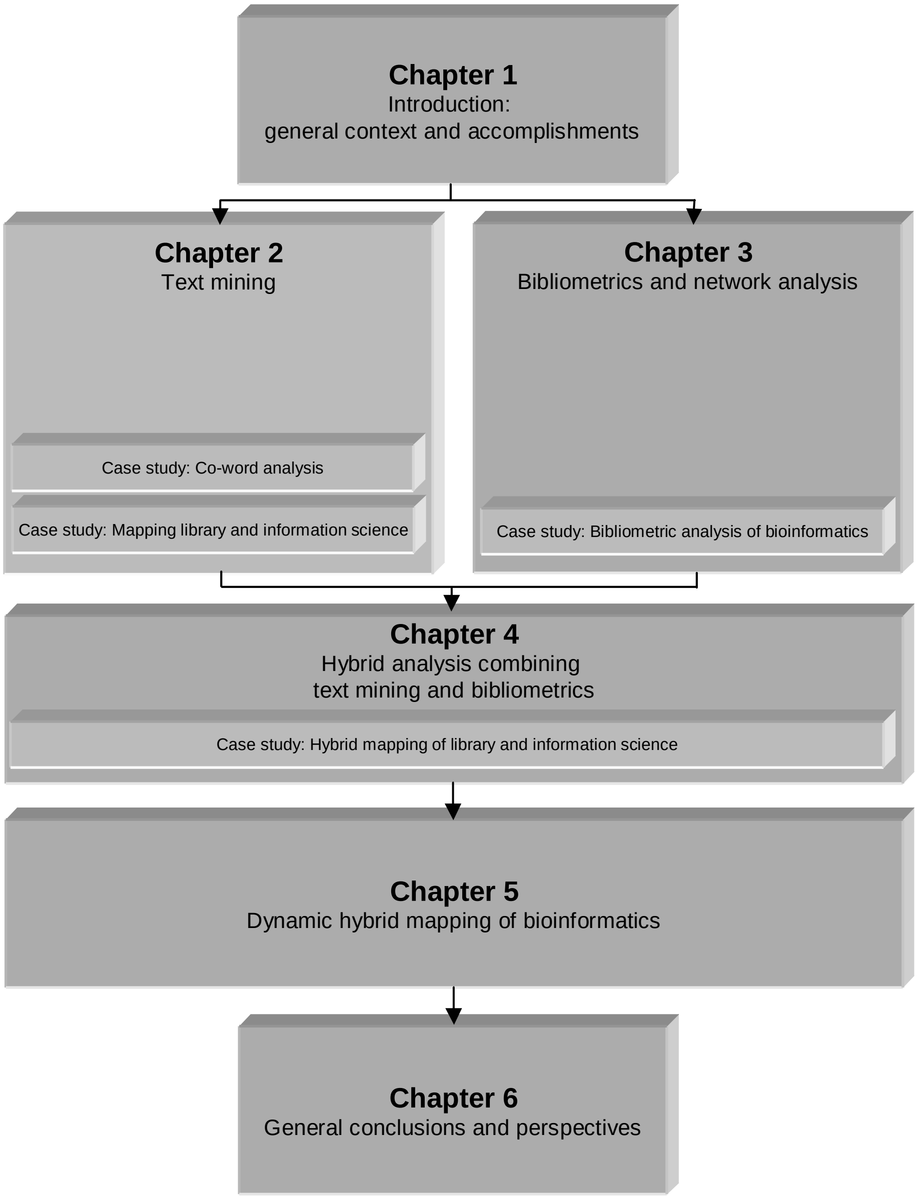

Chapter 1
Introduction:
general context and accomplishments
Chapter 2
Text mining
Case study: Co-word analysis
Case study: Mapping library and information science
Chapter 3
Bibliometrics and network analysis
Case study: Bibliometric analysis of bioinformatics
Chapter 4
Hybrid analysis combining
text mining and bibliometrics
Case study: Hybrid mapping of library and information science
Chapter 5
Dynamic hybrid mapping of bioinformatics
Chapter 6
General conclusions and perspectives

**Figure 1.17:** Structure of the dissertation. The current chapter gives an introduction and motivation by distinguishing the text and graph worlds. Chapter 2 discusses the text world in more detail, whereas the graph or bibliometrics world is treated in Chapter 3. In the figure, both are at the same level since they provide different views of equal value on the same bibliographic data set. In Chapter 4, both views are integrated. Resulting hybrid procedures are demonstrated in a case study, as well as in Chapter 5. Finally, Chapter 6 will give general conclusions and perspectives.

We mathematically and statistically combine document similarity matrices from textual information with similarity matrices that are based on network structure or bibliometric indicators. We demonstrate that performance of unsupervised clustering and classification of scientific papers is significantly improved and that best results are obtained by integration. We revisit the mapping of LIS by using hybrid methods and assess added value compared to the outcome of the text-only clustering of Chapter 2. Finally, a hybrid information retrieval strategy consisting of textual and bibliometric components is described and applied to delineate core literature in bioinformatics.

In **Chapter 5**, the bioinformatics field is further analyzed, focusing on cognitive structure as perceived by our hybrid clustering algorithm with *Fisher*'s inverse chi-square method, which provides integrated analysis of both text and citation worlds. For each cluster we provide term and collaboration networks, most representative publications, relative importance for the 5 most active countries, as well as citation patterns and *'naive' dynamics* of the cluster. The term *naive* refers to the fact that publication years are not considered during clustering, but only afterwards, when clusters are already formed. Subsequently, we introduce *dynamic hybrid clustering* for matching and tracking clusters through time. The resulting *cluster chains*, their structure and evolution, and various statistics are analyzed and compared with clusters found by *static* hybrid clustering of the complete bioinformatics set.

In closing, **Chapter 6** gives general conclusions, ideas for further research and some perspectives.

# Chapter 2

# Text mining

This chapter demonstrates the use and power of text mining techniques for automated retrieval of information and mapping of knowledge embedded in an overwhelming amount of digital texts (see also Section 1.2.1). *Text mining* comprises the intelligent analysis of textual data and aims for extraction and discovery of interesting facts, relationships and knowledge. Fast text mining algorithms are needed for keeping up with the dynamic character of science, with the immense volume of new articles published each year, and to discover interesting, unknown connections between various scientific research areas.

An overview of the application of text mining or quantitative linguistics in the context of informetrics and bibliometrics was given in Section 1.2.1. In the present chapter, Section 2.1 presents the text mining framework that we have adopted. It is not intended to be exhaustive as it only touches upon the text representation models that have been used in subsequent applications. The Vector Space Model is discussed, as well as preceding processing and indexing pipelines, detection of phrases (composite terms), and weighting schemes.

Next, Section 2.2 introduces the *curse of dimensionality* for data mining tasks and the necessity of dimensionality reduction by feature selection and methods such as Latent Semantic Indexing (LSI) or Random Indexing (RI). Interestingly, these techniques will to some extent also model semantics or meaning by mere mathematical processing. Section 2.2 is concluded with multidimensional scaling (MDS).

The adopted hierarchical clustering algorithm and validation measures are introduced in Section 2.3. A combination of several strategies for detection of the number of clusters is described (Section 2.3.3), which is based on dendrograms, text-based and citation-based Silhouette values and stability diagrams. Section 2.3.3 contributes to the debate about how to choose the number of latent semantic factors in LSI, in relation to the number of clusters and clustering performance. The introduction of *'second-order similarities'* concludes Section 2.3.

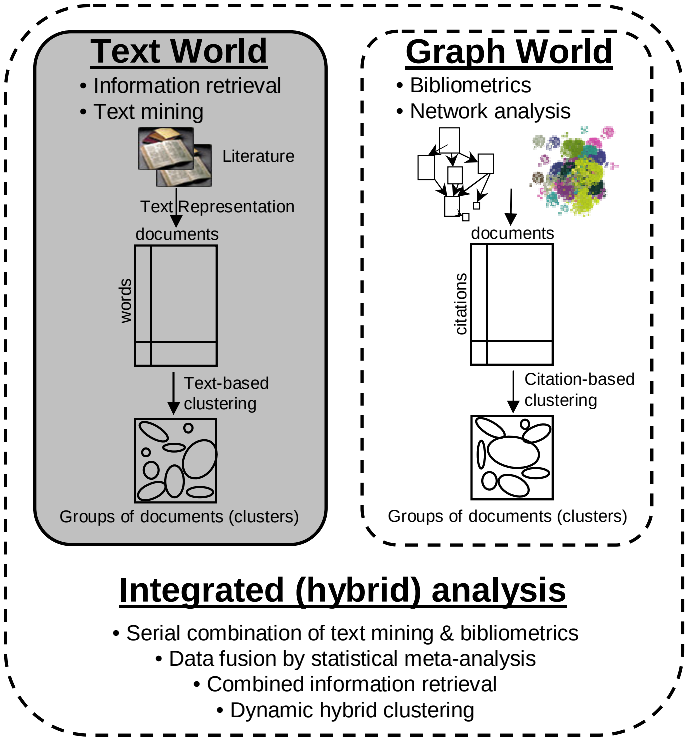

Text World
• Information retrieval
• Text mining
Literature
Text Representation
documents
words
Text-based clustering
Groups of documents (clusters)
Graph World
• Bibliometrics
• Network analysis
documents
citations
Citation-based clustering
Groups of documents (clusters)
Integrated (hybrid) analysis
• Serial combination of text mining & bibliometrics
• Data fusion by statistical meta-analysis
• Combined information retrieval
• Dynamic hybrid clustering

**Figure 2.1:** The focus of this chapter is on the *text world*. We introduce and demonstrate our text mining framework and the use of agglomerative hierarchical clustering. Next, we describe strategies to determine the number of clusters in a document set. Introduced algorithms are subsequently demonstrated in two case studies. The first study is about biotechnology in Belgium, the second one treats the field of library and information science (LIS).

Subsequently, the algorithms are demonstrated in two case studies in Sections 2.4 and 2.5. One uses co-word analysis as a means to examine whether and to what extent material transfer agreements (MTAs) influence research agendas in biotechnology. The other study uses a spectrum of mining techniques to unravel the cognitive structure of the field of library and information science (LIS).

## 2.1 Representation of textual data

In this section we give an overview of the text mining framework that was developed for case studies discussed throughout this thesis. The discipline of text-based information retrieval has a long history and various models have already been proposed. We only touch upon the most important aspects of the Vector Space Model that was used in this work. For a general overview we refer to Figure 1.12 on page 17. A graphical representation of most components of the document processing and indexing pipeline can be found in Figure 2.2. The full text of publications from 5 journals was mined in order to get a view on the structure of the scientific field under study. We discuss various processing steps within the framework of this application.

### 2.1.1 Text extraction

The first necessary step to constructing a mathematical representation of the textual information contained in documents is the extraction of the text. For full-text papers in Microsoft Word format (*.doc*) we used the StarOffice Software Development Kit[1] to extract the content. Files in the Portable Document Format (*.pdf*) were extracted by making use of Xpdf PDF text extractor[2], licensed under the GNU General Public License[3] (GPL). Unfortunately, text extraction is not always possible, particularly when a file only contains graphical scanned images of a document. For handling such documents, we used Optical Character Recognition (OCR) techniques from Scansoft's commercial package Omnipage 14. Note that text extraction or OCR techniques can introduce errors, special characters and sometimes even strange, very long strings.

### 2.1.2 Vector Space Model

In the Vector Space Model (VSM), an entity such as a document is represented by a vector or point in a (often) high-dimensional space. The dimensions constituting the vector space usually represent the set of all different words that can be found throughout a document collection, i.e., the *vocabulary*, *lexicon* or *thesaurus*.

[1] http://www.sun.com/software/star/staroffice/sdk/, visited in January 2007.

[2] http://www.foolabs.com/xpdf/, visited in January 2007.

[3] http://www.gnu.org/licenses/gpl.txt, visited in January 2007.

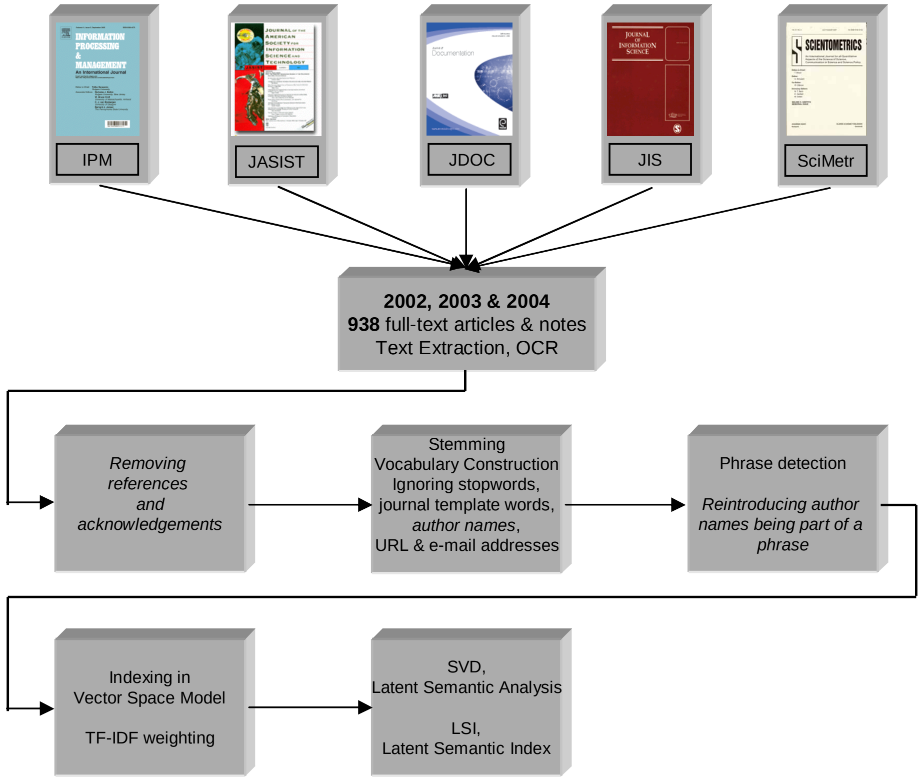


**Figure 2.2:** The full text of 938 articles and notes from 3 publication years of 5 journals is analyzed in various successive steps that are discussed in detail throughout this chapter. The arrows indicate the flow of the textual information through different components of the analysis pipeline. Terms in *italics* represent specific subtasks that are not crucial to the text mining framework and will consequently not be discussed here.

When processing a document to define its vector, all punctuation, word order and structure of the text is typically discarded. The vector will get values $w_i$ different from zero for each dimension corresponding to a word that is encountered in the text. Various weighting schemes are defined to estimate the importance of each word to each document. Because all structure is neglected, the VSM is frequently referred to as the *bag-of-words* representation.

### Matrix representations

Once a vector has been defined for each document in the corpus, they can be collected in a *term-by-document* matrix $A$ in which each row represents a word (or *term*) and each column represents a document. In general, a term-by-document matrix is extremely *sparse*, meaning that most values are zero. This is obvious as most documents only contain a very small fraction of terms from the global vocabulary. Figure 1.12 on page 17 presents pre-processing and indexing steps, the resulting term-by-document matrix $A$ and a two-dimensional visualization of the vectors for the first two documents, in the space spanned by the first two terms.

**Table 2.1:** Small extract from a term-by-document matrix constructed from a bioinformatics corpus. The dimensionality of the full matrix is 18163 × 7401. Only ten documents are shown (columns $d_1$–$d_{10}$) and only nine terms (rows). The set of nine terms was made up by choosing from each of 10 documents the term with the highest value.

| | $d_1$ | $d_2$ | $d_3$ | $d_4$ | $d_5$ | $d_6$ | $d_7$ | $d_8$ | $d_9$ | $d_{10}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| align | 0 | 0 | 0 | 0.04 | 0 | 0 | 0.05 | 0 | 0 | 0 |
| bind site | 0 | 0.07 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| fold | 0 | 0.38 | 0.45 | 0 | 0 | 0 | 0 | 0 | 0 | 0.07 |
| microarrai | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.25 | 0.42 |
| network | 0 | 0 | 0 | 0 | 0.53 | 0.61 | 0 | 0 | 0.06 | 0 |
| parasit | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.46 | 0 | 0 |
| predict | 0 | 0.06 | 0.20 | 0.25 | 0.03 | 0 | 0 | 0 | 0 | 0 |
| rbcl | 0 | 0 | 0 | 0 | 0 | 0 | 0.39 | 0.10 | 0 | 0 |
| rna | 0.60 | 0.45 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

The availability of a matrix allows the use of clustering techniques to subdivide documents or terms into groups (see also Section 1.3). *Clustering* is a multivariate statistical technique for automated grouping of objects such that similar objects are put in the same group or *cluster*, while dissimilar objects end up in different clusters as much as possible. A cluster then represents a set of documents that are similar in subject. We can approach the matrix from either a document viewpoint, or a term viewpoint. In the former case, the purpose is to group similar documents based on their term profiles. The goal of the latter is to cluster terms based on documents in which they occur. Other authors have also adopted an integrated viewpoint by *co-clustering* terms and documents, among others, *Dhillon* [63], but we do not further consider this option.

### Distance

Both the use of a search engine for Information Retrieval (IR) and the use of a clustering algorithm for grouping of documents involve computing mutual document distances. Various distance measures exist, among which the Euclidean distance metric and the complement of cosine similarity are frequently used.

**Euclidean distance** The Euclidean distance between two documents $d_1$ and $d_2$ is defined as

$$d(\vec{d_1}, \vec{d_2}) = \sqrt{\sum_i (w_{i,1} - w_{i,2})^2}, \tag{2.1}$$

where $w_{i,j}$ is the weight of term $t_i$ in document $d_j$. Euclidean distance is a true metric since it fulfills four required conditions for any three vectors $x$, $y$ and $z$:

1. $d(x, y) \geq 0$
2. $d(x, x) = 0$
3. $d(x, y) = d(y, x)$
4. $d(x, z) \leq d(x, y) + d(y, z)$

The length of a document has a large influence when using Euclidean distances. Long documents can be very similar merely by virtue of document length. This is an undesirable property as the similarity of two documents on a specific subject should not depend on their respective lengths. Therefore, in text mining applications it is certainly advisable to normalize all document vectors before application of the Euclidean distance measure.

**Cosine similarity** The cosine similarity measure computes the cosine of the angle between vector representations (correlation), resulting in a value between 0 and 1 (for the standard VSM) [236]. A distance measure can be obtained by subtracting each similarity value from 1. More formally, the similarity of two documents $d_1$ and $d_2$ can be computed as:

$$Sim(\vec{d_1}, \vec{d_2}) = cos(\widehat{\vec{d_1}\vec{d_2}}) = \frac{\vec{d_1} \cdot \vec{d_2}}{\|\vec{d_1}\| \cdot \|\vec{d_2}\|} = \frac{\sum_i w_{i,1} \cdot w_{i,2}}{\sqrt{\sum_i w_{i,1}^2} \cdot \sqrt{\sum_i w_{i,2}^2}}, \tag{2.2}$$

where $w_{i,j}$ is the weight of term $t_i$ in document $d_j$. The underlying hypothesis of the model states that the smaller the angle, and thus the higher the cosine similarity between two document vectors, the more semantically related the documents are [9]. The cosine similarity is insensitive to vector norms such that it does not discriminate between long and short documents. When dealing with length normalized vectors, division by the denominator is superfluous and the cosine similarity is equal to the inner product.

When a user enters a *query* in a search engine, it is converted to another vector in the same high-dimensional vector space, a *pseudo document*, of which the similarity to all other documents can be computed. The returned documents are then ranked according to these similarities. For normalized vectors, the Euclidean distance and the complement of cosine similarity give the same ranking of documents. The performance of a retrieval system is typically measured by *precision* and *recall*, or by the combined *F-measure* to balance the trade-off between both (precision and recall are inversely related). Precision is the ratio of the number of relevant documents retrieved to the total number of retrieved documents. It measures the percentage of documents within the search result set that are indeed relevant for the search topic. Recall is the ratio of the number of relevant documents retrieved to the total number of relevant documents. This ratio measures how many percent of all documents relevant for the query are indeed found by the search engine (some relevant documents might be missed).

As a final remark, it should be noted that the Vector Space Model can as well be applied to model people, institutions, journals, document clusters, genes, etc., by simply constructing a weighted average of the vectors related to relevant textual pieces of information. For instance, a certain employee can be modeled by linearly combining vectors that represent work authored by him or her. The VSM can also encode non-textual features in a vector space, such as the structure of a citation network, as will be discussed in the next chapter.

### 2.1.3 Indexing

Indexing is the process during which the extracted text is split into distinct terms or *tokens*, according to a predefined set of delimiters, and during which all occurrences of tokens in documents are counted [21, 19]. For the core indexing task we made use of Jakarta Lucene[4] [120], which is a high-performance, open source, full-featured text search engine library written in Java and ported to other languages as well. For specific needs, we extended Lucene with an inhouse java package called Textpack, which contains a set of extra filters that enable a more intelligent textual pre-processing. Examples are: *phrase filter*, *synonym filter*, *special character filter*, *regular expressions to ignore filter* and *not-to-fragment-patterns filter*.

[4] http://lucene.apache.org, visited in January 2007.

### Neglecting of stop- and template words, URLs and e-mail addresses

*Stop words* are words with little or no semantic value, such as *'the'* or *'and'*, typically collected in a *stop list*. Plain stop words, month and day names, terms with one or two characters and all terms containing non-alphabetical characters could safely be discarded during indexing. Template terms, i.e., terms printed on each page of a document such as journal-specific information, were also neglected as they might influence results. In addition, pattern matching was performed to match and ignore all e-mail addresses and URLs throughout the indexing process. These addresses introduce a lot of specific tokens that, if included, might also influence results (such as clustering based on institution or domain names).

### Stemming

On all remaining terms a language-specific stemmer can be applied, such as the *Porter* stemmer [225]. Stemming involves removal of word affixes such as plurals, verb tenses and deflections, and replacement of a term by the canonized equivalent. The *Porter* stemmer uses a simple rule-based scheme to process the most common English words. An advantage of stemming is the equation of different forms of the same word, resulting in a reduced dimensionality of the vector space and thus lessening computational costs and the *curse of dimensionality*. A disadvantage is possible loss of morphological information necessary for discerning between different meanings of two similar words.

### Phrase detection

We devoted a lot of time to the detection of domain-specific phrases, which are composite terms consisting of several words, such as *artificial intelligence*, that should be treated as one concept. Although external phrase lists are available for particular domains, often one has to resort to Natural Language Processing (NLP) techniques to automatically detect them. Since the best phrase candidates can be found in noun phrases, the programs LT POS and LT CHUNK[5] have first been applied to detect all noun phrases in the complete document collection. In addition, multi-word author keywords or Medical Subject Headings (MeSH[6]) were considered candidates. LT POS is a part-of-speech tagger that uses a lexicon and a hidden Markov model disambiguation strategy. LT CHUNK is a syntactic chunker or partial parser. It uses part-of-speech information provided by LT POS and employs mildly context-sensitive grammars to detect boundaries of syntactic groups. For the scoring of phrase candidates, *Dunning*'s *log*-likelihood method for detection of bigrams was followed to detect bigrams, trigrams, and tetragrams [73, 174]. The likelihood ratio tests the hypothesis that terms occur independently in a corpus. When rejected, the words are presumed to be corre-

[5] http://www.ltg.ed.ac.uk/software/pos/index.html, visited in January 2006.

[6] http://www.nlm.nih.gov/mesh/, visited in January 2007.

lated. It is a parametric statistical analysis based on the binomial or multinomial distribution and may lead to more accurate results than other text analyses that, often unjustifiably, assume normality, which limits the ability to analyze rare events. To detect trigrams, we considered for each occurring sequence of three words the first two and consequently the last two tokens as one single entity. Then we calculated the average *log*-likelihood score over both cases. The ranking of bigrams remains the same, but this *trick* aids to consider trigrams as well. For example, *hidden markov model* was correctly ranked higher than both *hidden markov* and *markov model*. An analogous extension was used for detecting tetragrams. However, a careful manual check of the phrase list was needed. Memory-based language processing techniques, which are more recent and advanced methods for, among other purposes, part-of-speech tagging, are described by *Daelemans* and *van den Bosch* [55].

#### Synonym resolution

Synonyms are different words or terms that carry the same meaning. Normalization of synonyms to one representative usually improves the statistical information about a word or term, provided that an external list or automatic detection procedure is available.

### 2.1.4 Weighting

In most applications, frequency values in the term-by-document matrix $A$ are weighted according to some weighting scheme during or after indexing, in order to increase accuracy of the VSM for IR or clustering tasks. Various popular weighting schemes apply a *local* and a *global* weighting component. These are proportional to various document-related and collection-related statistics, respectively. In this section, we outline the weighting schemes that have been used in our applications.

#### Boolean and TF models

In the boolean model, vector weights are binary. A word either occurs in a document or does not, indicated by 1 or 0, respectively. A measure for relative importance of words is not included. In the term frequency (TF) model, each word weight represents the frequency of that word in the document and is calculated as follows:

$$w_{i,j} = \frac{f_{i,j}}{max_j(f_{i,j})}, \tag{2.3}$$

where $w_{i,j}$ represents the weight of index term $t_i$ in document $d_j$. $f_{i,j}$ is the number of occurrences of $t_i$ in $d_j$. The rationale behind this approach is that words with high frequency are important and define the content of a document accurately. However, a problem of this model is that words with high frequency can also be words that bear not much content, in addition to the predefined list of stop words.

#### TF-IDF model

The TF-IDF (*term frequency - inverse document frequency*) weighting scheme has shown to be very effective in information retrieval for determining the most relevant documents to a user's query and the most important terms in documents. It represents the relevance or importance of terms in a document by counting the frequency of every word, like the TF model does, but by also taking into account the occurrence of a particular word in the entire document collection. TF-IDF values are computed as follows:

$$w_{i,j} = f_{i,j} \cdot log\Big(\frac{N}{n_i}\Big), \tag{2.4}$$

where $f_{i,j}$ is the term frequency, i.e., the number of occurrences of term $t_i$ in document $d_j$, $N$ represents the total number of documents, and $n_i$ is the number of documents containing term $t_i$.

The TF-IDF weight of a term in a document is high if the term frequently occurs in that document, but only if it occurs in just a few other documents as well, i.e., having a low document frequency and consequently a high IDF. As a result, terms that occur in a lot of documents are considered common terms and are down-weighted. For example, in a corpus containing nothing but computer science publications the term *computer* is presumably not a good term to discriminate between documents.

## 2.2 Dimensionality reduction and semantics

Even when dealing with modestly-sized document collections of a few tens of thousands of documents, the total number of words (or tokens in general) encountered throughout the corpus can easily reach values of the order of tens to hundreds of thousands. The next subsection discusses inherent problems associated with such a high dimensionality, which makes dimensionality reduction an indispensable pre-processing step. The remaining subsections mention possible reduction methods. Apart from reducing the number of dimensions, Sections 2.2.3, 2.2.4 and 2.3.4 also introduce latent semantics into the mining process.

### 2.2.1 Curse of dimensionality

The *curse of dimensionality* refers to the exponential growth of the 'hyper volume' with increasing dimensionality [84, 3, 117]. The ratio of the volume of the unit hyper sphere to the volume of the hyper cube in which it is embedded, decreases with increasing dimensionality. The result is a decline in the performance of algorithms to discern between documents that are close to a given document and the majority of other documents in a collection. When the dimensionality rises, distance measures become increasingly meaningless as all objects seem to be almost equidistant from each other. The search for *nearest neighbors* becomes very unstable. The number of training examples should also increase exponentially in order to counter the inherent sparsity associated with high dimensionality and to maintain accuracy.

Because our endeavors to map or chart scientific and technological fields rely on clustering algorithms that, in turn, rely on a distance measure, a prior reduction of dimensionality becomes indispensable [210, 66, 215]. As a consequence, computational cost of the actual clustering algorithm will often be reduced as well, and the interpretability of clusters enhanced. If the dimensionality is further reduced to two or tree dimensions, data can be visualized directly and one can benefit from human pattern recognition abilities. Unfortunately, a severe reduction of dimensionality might as well destroy important aspects of data and cause loss of information.

### 2.2.2 Feature selection

In the Vector Space Model, dimensionality is determined by the number of documents and the size of the vocabulary, i.e., the amount of distinct terms or phrases that occur throughout the document collection.

One approach towards reducing the vocabulary size, and thus circumventing the curse of dimensionality, is selection of the most relevant terms (features) that probably carry a lot of meaning, and by neglecting other terms. A primitive, limited way of achieving this is by using a *stop list*, as described earlier. In some application areas this method can be reversed by neglecting all tokens except for those listed in an application-specific *domain vocabulary*, hand-crafted or derived from, for instance, an ontology. However, the latter approach, while possibly an effective dimension reduction technique, should only be used with care as it is very drastic and might lead to overlooking essential information nuggets. Moreover, detection of important new concepts in a collection is no longer within bounds of possibility. Such a harsh restriction of terms is not a recommendable option in general-purpose retrieval tasks.

Besides stemming, another popular and effective method for reducing dimensionality is to cut off *Zipf*'s curve [277]. If all words that occur in a document set are sorted in decreasing order of frequency $f$, and if those frequencies are multiplied with the rank $r$ (position in the list), the result will approximately be a constant $C$, i.e., $C = r \cdot f$. The frequency of a word is thus roughly inversely proportional to its rank in the frequency list. This is formulated in the famous law of *Zipf*. Words in the tails of the curve can be considered to bear less content than terms in the middle of the curve. Hence, 'cutting off *Zipf*'s curve' by neglecting terms or phrases that, for instance, only occur once or in more than 50% of all documents, is a form of feature selection to retain the most important dimensions.

Although out of scope of this dissertation, other existing univariate or multivariate statistical feature selection methods are based on *t-statistics*, *correlation*, *chi-square*, *information gain*, *mutual information*, *entropy* or *Bayesian* techniques [1, 254, 130, 249, 64, 254, 236, 273].

### 2.2.3 Latent Semantic Indexing

Latent Semantic Indexing (LSI) is a mathematical technique based on the truncated Singular Value Decomposition (SVD) of a matrix [111], which analyzes the term-by-document matrix $A$ in order to find the major associative patterns of word usage in the document collection and to get rid of the 'noise variability' in it. It assumes that there is some underlying or latent structure in word usage that is partially obscured by variability in word choice [62, 23]. LSI makes it possible to compose a matrix $A_k$ that is an optimal approximation of $A$ (in least squares sense), but with rank $k$ much lower than the term or document dimension of $A$.

LSI is a 'feature transformation technique' that creates factors from linear combinations of the original term dimensions. Instead of the huge number of rows in the matrix $A$ (equal to the total number of terms), only $k$ statistically derived orthogonal indexing *factors* or *pseudo concepts* remain in $A_k$. The vocabulary is thus replaced by a much smaller set of pseudo concepts that are present in the document set. A document has a weight for each pseudo concept that indicates how much the 'concept' is represented in the document. If the weight indicates a highly positive correlation, it means that it is a relevant 'concept' for the document.

An interesting effect of LSI is that synonyms or different term combinations describing the same idea are mapped onto the same factor based on the common context in which they generally appear, even for terms that do not co-occur in any document. Besides the implicit relating of synonyms, also the problem of polysemy (a word with different meanings depending on the context) is partly addressed by LSI. In an information retrieval scenario where a search engine is used in order to find documents relevant to a certain query, the main advantage of LSI is that documents are found even if they do not literally contain the query words. The query 'car' might, for instance, also retrieve documents that only talk about 'automobile', without the use of any form of dictionary. Indeed, *Kontostathis* and *Pottenger* have shown that LSI captures higher-order term co-occurrence information for terms that never co-occur in the same document, but that can yet cognitively be linked by co-occurrence with the same other terms. They observed a strong correlation between second-order term co-occurrence and the values produced by SVD [151].

One drawback of LSI besides the computational load is the loss of the reduced matrices' sparseness, which results in a less efficient use of memory with respect to the dimensionality and size of the data set. Another disadvantage of LSI is that it is an off-line technique which is unable to incorporate new documents appearing in a dynamic collection. A limited number of new documents can be 'folded' into an existing LSI, but if a lot of new documents should be considered, the model might become inaccurate and a recalculation of the complete index should be performed.

### Singular Value Decomposition

LSI uses the truncated SVD to approximate a term-by-document matrix $A$ with a matrix $A_k$ of lower rank $k$ [111, 23]. The SVD of a given matrix $A$ of size $m \times n$ $(m \geq n)$ and of rank $r$, is written as

$$A = U \cdot \Sigma \cdot V^T, \tag{2.5}$$

where the diagonal, real elements of $\Sigma$ are the *singular values*, $\Sigma = diag(\sigma_1, ..., \sigma_n)$ with $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_n$, and $\forall i \leq r : \sigma_i > 0$, $\forall j > r : \sigma_j = 0$. These singular values are the nonnegative square roots of the $n$ eigenvalues of $A^T \cdot A$.

$U$ and $V$ are orthonormal matrices, i.e., $U^T \cdot U = V^T \cdot V = I_n$, and their respective first $r$ columns, the left and right *singular vectors*, are the orthonormal eigenvectors associated with the $r$ nonzero eigenvalues of $A \cdot A^T$ and $A^T \cdot A$ [111, 23].

Following a famous theorem by *Eckart* and *Young* [75], $A_k$ is the closest rank-$k$ approximation to $A$, in least squares sense, and can be constructed as follows:

$$A_k = SVD_k(A) = U_k \cdot \Sigma_k \cdot V_k^T = \textstyle\sum_{i=1}^{k} u_i \cdot \sigma_i \cdot v_i^T \tag{2.6}$$

Figure 2.3 gives the mathematical representation of $A_k$ and shows the construction of the *concept-by-document* matrix (LSI).

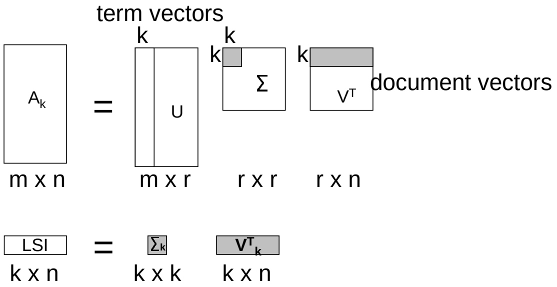


**Figure 2.3:** Mathematical representation of the matrices $A_k$ and the LSI.

For information retrieval purposes, a user's query vector $q$ can be compared with the latent semantic vectors of all documents by first projecting $q$ in the same space of reduced dimensionality:

$$q_p = q^T \cdot U_k \cdot \Sigma_k^{-1} \tag{2.7}$$

### Probabilistic LSI and Latent Dirichlet Allocation

Besides LSI, other concept indexing methods exist but are outside the scope of our investigation [65, 125]. Probabilistic Latent Semantic Indexing (PLSI) [126, 125] evolved from LSI and is based on a statistical latent class model for factor analysis of co-occurrence data, which associates an unobserved class variable with each observation. The so-called *aspect* (factor) *model* is fitted from a training corpus of text documents, for example by expectation maximization. A joint probability model over documents and words is defined by a mixture. Documents are characterized by a specific mixture of weighted factors. Each term in a document is generated from a single topic from the mixture, while different words from the same document may result from different topics. In some aspects PLSI is similar to LSI. For instance, LSI factors correspond to the mixture components of the aspect model, and the mixing proportions in PLSI substitute the singular values in LSI. However, the objective function used to determine the optimal approximation is different. PLSI aims at maximization of the predictive power of the model by maximizing the likelihood of observed term frequencies in order to learn the factors as well as the mixing weights for each document.

An advantage of PLSI is that the factors have a clear probabilistic meaning as multinomial word distributions, whereas LSI factors are difficult to interpret as they can contain negative values as well. A possible drawback of PLSI is a higher computational complexity compared to LSI. Another disadvantage is that the expectation maximization algorithm is only guaranteed to find a local maximum of the likelihood function. Furthermore, the number of parameters grows linearly with the number of documents. A related problem is overfitting and consequently a problematic generalization to documents not considered for training. According to *Blei*, *Ng*, and *Jordan* [28], PLSI is incomplete as it does not provide a well-defined generative model at the level of documents; probabilities can not be assigned to previously unseen documents. They introduced Latent Dirichlet Allocation (LDA) to counter disadvantages of PLSI.

LDA provides a generative probabilistic model of text collections. Documents are represented as random finite mixtures over latent topics, where each topic is characterized by a distribution over words. Hence, to generate a document, a distribution over topics is chosen and then each term in the document is chosen from a topic selected according to this distribution. In contrast to PLSI, LDA provides a well-defined generative model and generalizes easily to new documents. It has better scaling properties and less issues with overfitting. *Griffiths* and *Steyvers* [112] have applied LDA to a collection of abstracts from the *Proceedings of the National Academy of Sciences of the USA* (PNAS). They presented a Markov chain Monte Carlo algorithm for inference in this model and they used Bayesian model selection to decide on the number of topics. Furthermore, they also explored basic temporal dynamics to identify *hot topics*.

### Number of LSI factors

How to choose the rank $k$ remains an open question. Often, a *scree plot* with the decay of singular values is observed to look for a good cut-off point, or 'elbow', where most of the information is explained by the $k$ retained singular values, whereas the added value of additional singular values is relatively low. A related heuristic is to only retain those factors for which the singular value is larger than the average value. The use of *amended parallel analysis* as a means to selecting the number of factors was introduced by *Efron* [76]. Typical values for $k$ found in literature range from 100 [62] (for less than 2000 abstracts and 7000 terms) to about 300 [24] (for about 12 000 documents and 40 000 terms).

Admittedly, the absence of a straightforward rule for determining the number of factors renders any selection slightly arbitrary in the sense that other values might be appropriate as well. In addition, *Kostoff* and *Block* stated that interpretation of a scree plot is partly subjective and that the plot exhibits a 'fractal-like behavior' [154]. Consequently, depending on the resolution of the scree plot, a different value for $k$ might be perceived as the best choice.

Nonetheless, according to *Deerwester et al.*, the number of retained factors might be crucial to the success of LSI [62]. Choosing too few factors might result in loss of important information, whereas too many might lead to overfitting of the model.

We believe that choosing a good value for the number of latent semantic factors might seriously improve the performance of subsequent clustering tasks. Perhaps it is even more important for clustering than it is for information retrieval problems given that clustering often involves iterative distance calculations and that the outcome is affected by cumulative choices.

In Section 2.3.3, we will investigate clustering performance for various numbers of clusters and LSI factors and provide some insight into the relation between number of LSI factors, number of clusters, and clustering performance. There, it will be shown that for our data sets a very modest number of factors, e.g., 10, can provide a local maximum in clustering performance. This observation is in line with observations by *Kontostathis* [152]. She has shown in a retrieval setting that a small, fixed dimensionality reduction parameter ($k = 10$) can be used to capture the term relationship information in a corpus.

## 2.2.4 Random Indexing

Random Indexing (RI) has been proposed as a simple and scalable alternative to LSI with interesting advantages [143, 144, 234]. The method is mathematically equivalent to Random Mapping [145] or Random Projections [214]. RI is motivated by the *Johnson-Lindenstrauss* Lemma which states that the projection of points in a randomly selected subspace of sufficiently high dimensionality preserves mutual distances [142]. One of the most interesting properties of RI is the fixed dimensionality of, for instance, 1800 dimensions, that need not be increased when new data is observed. This in contrast to the traditional

VSM, in which the dimensionality of the vocabulary might increase with each observed new document. Moreover, RI eliminates the need for an off-line computationally expensive dimensionality reduction step such as LSI, while offering comparable *latent semantics* by taking the context of words into account. RI is thus an *incremental* method, meaning that it is very flexible to the addition of new data and suitable for dynamic document sets with changing and growing information. There is no need for recalculating the index, nor for 'folding in' new documents. Although an RI is only an approximation to an original term-by-document matrix, mutual distances between documents are very well preserved.

A random index can be constructed in the following manner. First, a fixed dimensionality is chosen as a parameter, e.g., $d = 1800$. Secondly, for each document, paragraph or sentence, a sparse random *index vector* is constructed by setting 4 randomly chosen dimensions to $+1$ and 4 random dimensions to $-1$, while leaving all other dimensions zero. Next, for all terms a *context vector* is built by simply adding all *index vectors* (contexts) in which the term occurs. For a new document, only one extra index vector needs to be defined and added to the context vectors of all terms that occur in it. Terms that had never occurred before get a new context vector equal to the document's index vector. These context vectors can readily be used to compare or cluster terms based on all contexts in which they appear. Following *Sahlgren*, in an extra step documents can be the subject of analysis by adding all (weighted) concept vectors of terms that occur in it [235]. This *bag-of-concepts* brings in higher-order co-occurrence information and thus latent semantics as is achieved by LSI.

### 2.2.5 Multidimensional scaling

Multidimensional scaling (MDS) represents high-dimensional points (for example, documents) in a lower dimensional space by explicitly requiring that the pairwise distances between the points approximate the original high-dimensional distances as precisely as possible [175, 113]. If the dimensionality is reduced to two or three dimensions, these mutual distances can directly be visualized. We have used classical metric MDS in order to get a view of science fields. See Figures 1.6 on page 7 and 2.13 on page 67 for some examples. It should however be stressed that interpretations concerning such a low-dimensional approximation of very high-dimensional distances must be handled with care.

## 2.3 Clustering

For a general introduction to clustering we refer to Section 1.3. In the present section, we mainly discuss evaluation and validation, and a combination of several strategies for detection of the number of clusters.

### 2.3.1 Algorithm

Clustering is a difficult task. Different solutions are possible by applying various algorithms or even by choosing other parameterizations or validation measures for the same algorithm. For a lot of algorithms there is in fact no guarantee that the optimal solution is within reach for a certain set of parameters or initializations. Algorithms can get stuck in local minima. For example, some algorithms that minimize intra-cluster variance can not guarantee to find a global minimum.

A wealth of clustering algorithms have already been proposed for text and data mining. We mostly opt for agglomerative hierarchical clustering using *Ward*'s method or UPGMA (see Section 1.3), but we also report on experiments with complete linkage and with the $k$-means partitioning algorithm.

Useful surveys on clustering have been composed by *Jain, Murty* and *Flynn* [134], *Berkhin* [18], and *Xu* and *Wunsch* [272]. These works treat various linkage methods and different clustering methodologies such as, among other, divisive and partitional clustering, nearest neighbour, density-based, grid-based, fuzzy, and model-based clustering.

Like any other algorithm, deterministic agglomerative hierarchical clustering has advantages and weaknesses. Although it is computationally heavy, the agglomerative hierarchical method has the advantage that a hierarchical tree (a dendrogram) can be inspected visually to determine a suitable number of clusters. The tree can be cut off at different levels, tuning the granularity of categorizations, without the need for reclustering. One disadvantage is that wrong choices (merges) that are made by the algorithm in an early stage can never be repaired [146]. Another reason why the clustering results leave room for improvement is that in this chapter we only make use of text and neglect other information. In Chapter 4 we asses the performance of different schemes for integration of textual and bibliometric information to obtain even better clustering results.

### 2.3.2 Evaluation and validation

Cluster quality can be assessed by *internal* or *external* validation measures. Internal validation solely considers statistical properties of data and clusters, whereas external validation compares the clustering result to a known gold standard partitioning. *Halkidi*, *Batistakis*, and *Vazirgiannis* gave an overview of quality assessment of clustering results and cluster validation measures [116]. In the following paragraphs we describe the measures that we adopted.

#### Silhouette coefficient

The Silhouette value $S(i)$ for a document $i$ ranges from -1 to +1 and measures how similar the document is to documents in its own cluster vs. documents in other clusters [232, 133]. $S(i)$ is defined as follows:

$$S(i) = \frac{min\big(B(i,C_j)\big) - W(i)}{max\Big[min\big(B(i,C_j)\big), W(i)\Big]}, \tag{2.8}$$

where $W(i)$ is the average distance from document $i$ to all other documents within its cluster, and $B(i,C_j)$ is the average distance from document $i$ to all documents in another cluster $C_j$.

The mean Silhouette value over all documents in a cluster is an indication of cluster quality. The average for the complete data set gives an intrinsic measurement of the overall quality of the clustering result. As Silhouette values are based on distances, depending on the applied distance measure different Silhouettes can be calculated.

Evaluation of clustering results can be a time consuming operation. Because of the huge amount of calculations involving *Silhouette* values, we could greatly reduce the computational cost by considering the (well-known) fact that the average distance of a document to all documents of a cluster is exactly the same as the distance of the document to the centroid of that cluster. The standard algorithm in *Matlab* could be sped up by making this modification. The altered implementation scales roughly linear with the number of documents.

### Jaccard similarity coefficient

The *Jaccard index* is the ratio of the cardinality of the intersection of two sets and the cardinality of their union. The *Jaccard similarity coefficient* is an extension of the *Jaccard* index and can be used as a measure for external cluster validation. It is used to compare a clustering result $C = \{C_1, C_2, \cdots, C_k\}$ with an external partitioning $P = \{P_1, P_2, \cdots, P_l\}$, where $k$ and $l$ represent the number of clusters and partitions, respectively. For $C$ and $P$ we define $n \times n$ matrices $M^C$ and $M^P$, where $n$ is the total number of documents. Each Boolean value $M^C_{ij}$ indicates whether documents $i$ and $j$ belong to the same cluster in $C$, while $M^P$ indexes all documents that are in the same partition in $P$.

Let $N_{00}$ represent the number of pairs of documents that do not belong to the same cluster in $C$, nor in $P$, i.e., the number of elements $(i,j)$ for which $M^C_{ij} = M^P_{ij} = 0$. Likewise, $N_{01}$ counts the number of elements for which $M^C_{ij} = 0$ and $M^P_{ij} = 1$. $N_{10}$ and $N_{11}$ are defined analogously.

The *Jaccard* coefficient is then defined as follows:

$$J(C,P) = \frac{N_{11}}{N_{11} + N_{01} + N_{10}} \tag{2.9}$$

The resulting value between 0 and 1 quantifies the correlation between the two binary matrices, while disregarding negative agreements ($N_{00}$).

### Rand index

The *Rand* index is another external validation measure to quantify the correspondence between a clustering outcome $C$ and a ground-truth categorization $P$ [133]. In contrast to the *Jaccard* coefficient, the *Rand* index does take into account negative matches as follows:

$$R(C,P) = \frac{N_{11} + N_{00}}{N_{11} + N_{01} + N_{10} + N_{00}} \tag{2.10}$$

The result is also a value between 0 and 1, with 1 indicating that $C$ and $P$ are identical.

A problem with the *Rand* index is that the expected value for the agreement between two random partitions is not a constant value and can be rather high. *Hubert* and *Arabie* therefore proposed the *adjusted Rand* index [131, 274]. It assumes the generalized hypergeometric distribution as a model of randomness to compute the expected value $E\big(R(C,P)\big)$. The *adjusted Rand* index is then computed as:

$$R_{adj}(C,P) = \frac{R(C,P) - E\big(R(C,P)\big)}{max\big(R(C,P)\big) - E\big(R(C,P)\big)} \tag{2.11}$$

*Milligan* and *Cooper* have recommended the *adjusted Rand* index as the external validation measure of choice [183].

## 2.3.3 Optimal number of clusters

A lot of clustering algorithms require the number of clusters as a predefined parameter, otherwise another parameter is mostly used for tuning granularity. Determination of the number of clusters in a data set is a difficult issue and depends on the adopted validation and chosen similarity measures, as well as on data representation. The strategy we use throughout this dissertation for determining the number of clusters is a combination of distance-based and stability-based methods. For deciding on the number of clusters, the following sections discuss the use of a dendrogram, Silhouette curves and Silhouette plots, and the stability-based method of *Ben-Hur, Elisseeff* and *Guyon* [16]. The combined strategy is illustrated on a document set containing 7401 bioinformatics publications.

### Dendrogram

A first judgment is offered by a dendrogram, which provides a visualization of distances between (sub-)clusters. It shows iterative grouping or splitting of clusters in a hierarchical tree. For instance, Figure 2.4 depicts a dendrogram that resulted from clustering 938 scientific full-text documents about library and information science. The dendrogram is cut off on the left-hand side at 25 clusters.

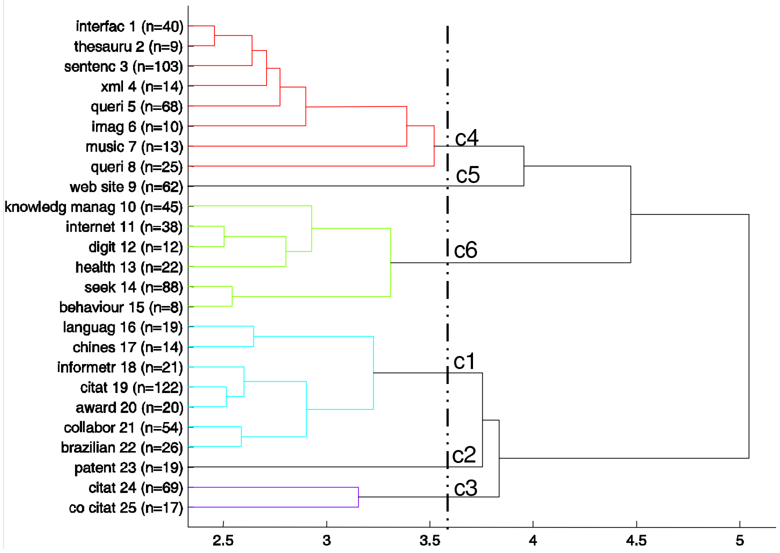


**Figure 2.4:** Dendrogram, cut off at 25 clusters on the left-hand side and at 6 clusters (c1–c6) at the vertical line, for hierarchical clustering of 938 papers on library and information science. For each of 25 clusters, the best mean TF-IDF term is shown.

The horizontal lines connect clusters in a hierarchical tree and the line length represents the distance between two connected clusters. At each leaf node, the term representing the cluster has the highest mean TF-IDF value in the sub-set. A candidate number of clusters can be determined visually by looking for a cut-off point where an imaginary vertical line would cut the tree such that resulting clusters are well separated. In Figure 2.4, an appropriate cut-off level is visible for 6 clusters. Because of the difficulty to define the most natural cut-off point on a dendrogram [133], we complement this method with other techniques. For the bioinformatics publications, Figure 2.5 depicts the dendrogram that resulted from hybrid hierarchical clustering, cut off at 9 clusters on the left-hand side.

### Silhouette curves

A second appraise for the number of clusters is given by the mean Silhouette curve [232, 146] (see Section 2.3.2). The best combination of number of clusters and number of LSI factors depends on the document collection at hand and on the objectives of the study. In order to investigate clustering performance for various numbers of clusters and LSI factors, Figure 2.6 presents clustering performance measured by mean Silhouette coefficient for 2 to 50 clusters, for different numbers of factors and for the standard VSM. For this experiment we again use the bioinformatics set. Note that for the sake of comparability each clustering was evaluated with Silhouette values computed from the original term-by-document matrix $A$ on which SVD had not been applied.

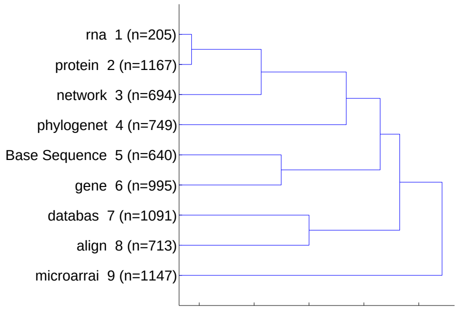


**Figure 2.5:** Dendrogram, cut off at 9 clusters on the left-hand side, for hierarchical clustering of 7401 bioinformatics publications. For each of 9 clusters, the number of publications and the best mean TF-IDF term or phrase are shown.

Figure 2.6 demonstrates that for this data set, in general, clustering performance is higher for a lower number of LSI factors ($k$). Nevertheless, performance seems to drop quickly if the number of clusters is higher than the number of factors. Thus, in this case there clearly is a connection between number of factors and number of clusters. An explanation might be that it is a harder task to discern a certain number of clusters encoded in a lesser amount of dimensions. Hence, as a heuristic, it might be advisable to use a number of factors at least as high as the desired number of clusters. However, these observations should be further assessed using other corpora as well.

When looking for a coarse-grained clustering solution, a very modest number of factors, e.g., $k = 10$, seems to provide the best clustering performance and also has direct advantages in terms of storage needs and processing time. Indeed, Figure 2.6 shows that for any number of clusters less than 10, 10 LSI factors provide the highest Silhouette values. Next, for more than 10 clusters, 15 factors take the lead, whereas 30 factors do best for finer-grained clustering solutions with more than 15 clusters. Again, from 31 clusters onwards, the next smallest number of factors in line, 50, is the winning number. Main observations are summarized in Table 2.2. Perhaps a good strategy might be to calculate the SVD with a number of factors equal to the desired maximal number of clusters, and to use a solution with less factors for obtaining coarser-grained clustering solutions. The drawback is then that the clustering needs to be recomputed for different levels of granularity, which is unnecessary in standard hierarchical clustering of a single LSI reduced matrix with a fixed number of factors.

Although being all positive, overall mean Silhouette values in Figure 2.6 each seem low, hinting at groups of documents that are not clearly separable according to the original classification of *Rousseeuw* [232]. This is probably due to the very high dimensionality of the original vector space in which the Silhouette values are computed (for comparability, as mentioned above), in contrast to the low-dimensional problems discussed by *Rousseeuw*. In addition, the nature of natural language usage might be of influence. When dealing with documents in

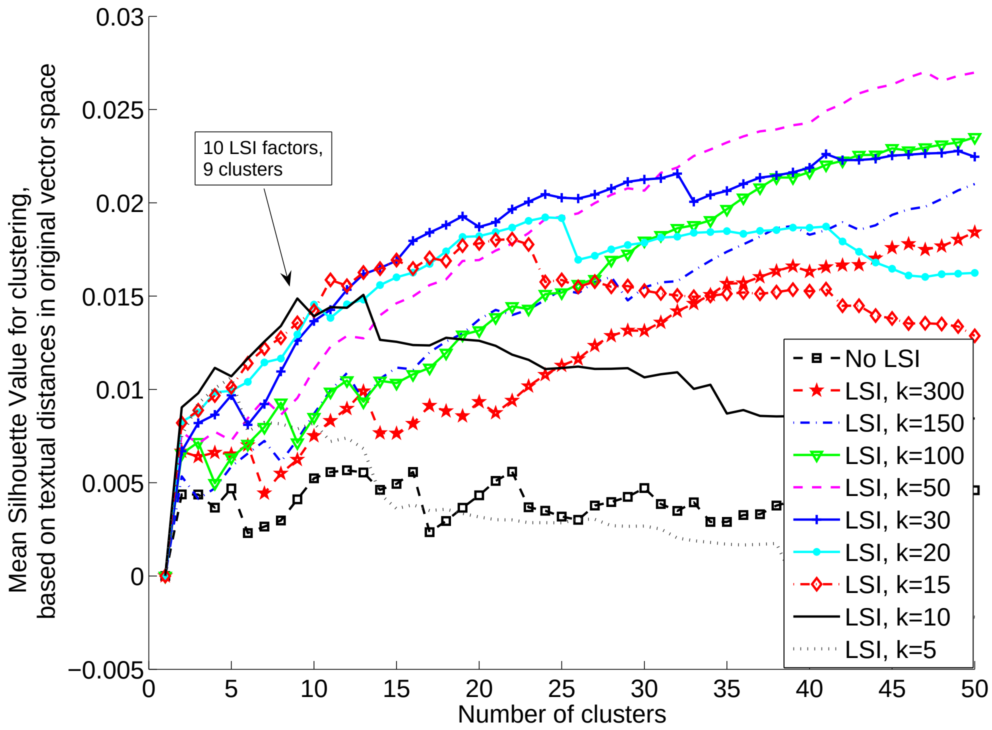


**Figure 2.6:** Silhouette curves with mean Silhouette coefficient for text-based clustering solutions (using *Ward*'s method) of 2 up to 50 clusters, for the original term-by-document matrix ('No LSI') and for derived latent semantic indices ('LSI') with different numbers of factors $k$. The arrow indicates the chosen combination of 9 clusters and 10 LSI factors for the bioinformatics set. In general, for this data set the quality of clustering proves significantly higher for a smaller number of factors. A modest number of factors (e.g., 10) seems to provide best clustering performance, on condition that there are no fewer LSI factors than the desired number of clusters.

**Table 2.2:** Main observations regarding the best number of LSI factors for different numbers of clusters, chosen from the set $\{5, 10, 15, 20, 30, 50, 100, 150, 300\}$ (cf. Figure 2.6). In general, the lower the number of LSI factors $k$, the higher clustering performance for this data set. However, $k$ should be larger than or equal to the number of clusters $c$.

| Number of clusters $c$ | Best number of LSI factors $k$ |
|---:|:---|
| $c < 10$ | 10 |
| $c = 10$ | 20 or 15 |
| $10 < c \leq 13$ | 15 |
| $13 < c \leq 15$ | 15 or 20 |
| $15 < c \leq 30$ | 30 |
| $30 < c$ | 50 |

comparable subject areas, the amount of overlapping words between different papers is, of course, considerable. Hence, documents in different clusters are likely to have terms in common as well.

For the bioinformatics document set, 10 LSI factors and 9 clusters seem to be the best combination (cf. the indicated local maximum). One might argue that other solutions with more clusters and more factors can provide higher Silhouette values, but we are rather looking for a local maximum. For instance, an expert will most likely have some ideas about a range of possible numbers of clusters to look for and will probably not be interested in 50 clusters within the bioinformatics field. Additional evidence is given by the Silhouette curve for link-based clustering using bibliographic coupling, which also shows a clear local maximum at 9 clusters (see Figure 2.7). For more information on bibliographic coupling we refer to Section 3.2.5.

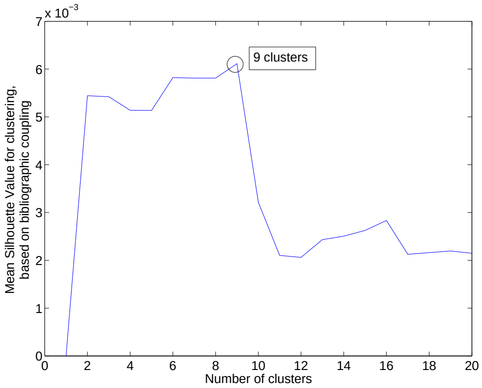


**Figure 2.7:** Silhouette curve with mean citation-based Silhouette coefficient for link-based (bibliographic coupling) clustering with 2 up to 20 clusters. The Silhouette values based on citation information also suggest 9 clusters.

### Silhouette plot

When a data set is divided in a specific number of clusters, quality of these clusters can be visualized in a Silhouette plot. In a Silhouette plot (see Figure 2.8), the sorted Silhouette values of all members of each cluster are indicated with horizontal lines. The more the Silhouette profile of a cluster is to the right of the vertical line at the value 0, the more coherent the cluster is.

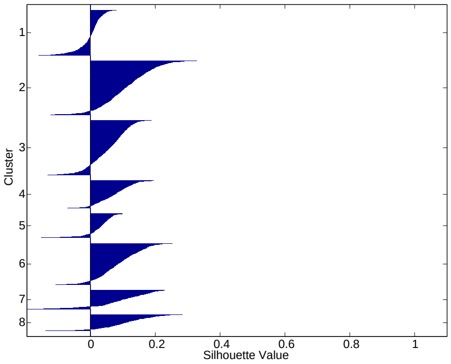


**Figure 2.8:** Example Silhouette plot for eight clusters. The sorted Silhouette values of all members of each cluster are indicated with horizontal lines. Silhouette profiles with mainly positive values indicate coherent clusters, whereas negative values indicate that the corresponding objects should rather belong to another cluster.

### Stability

Even more evidence for our 9 clusters within the bioinformatics field is provided by the stability-based method of *Ben-Hur, Elisseeff* and *Guyon* [16], which allows to visually and quantitatively detect the most stable number of clusters from a stability diagram. The method can be used with any clustering algorithm and can also detect lack of structure in data. The main idea is that perceived structure should remain stable if only a subsample of objects is available, or if noise objects are added to the data set.

Multiple subsamples (e.g., 200) are randomly drawn from the set, each comprising for instance 85% of objects. Then, a clustering algorithm subdivides each subsample into different numbers of clusters (e.g., 2 to 25 clusters). When a hierarchical method is used, only one run of the algorithm is needed because each level of the binary tree represents a different number of clusters. Next, the overlap between each pair of clustered subsamples is quantified by using the *Jaccard* coefficient (for a specific number of clusters).

The diagram of Figure 2.9 shows, for 2 up to 25 clusters, the cumulative distribution of pairwise *Jaccard* similarities, between 200 pairs of clustered random subsamples, each comprising 6291 bioinformatics publications (sampling ratio of 85%). Each number of clusters thus leads to one curve in the stability diagram. The more a curve is to the right of the diagram, the higher the pairwise similarities between clustered subsamples, and the more stable the clustering solutions with that specific number of clusters. A point on a curve representing a certain number of clusters can be interpreted as the fraction of subsample pairs ($Y$-axis) that have *Jaccard* values lower than or equal to the corresponding value on the $X$-axis. As explained above, the number of clusters is chosen such that partitioning different subsamples leads to quite stable structures. In practice, a transition curve to the band of distributions on the left-hand side of the figure is selected.

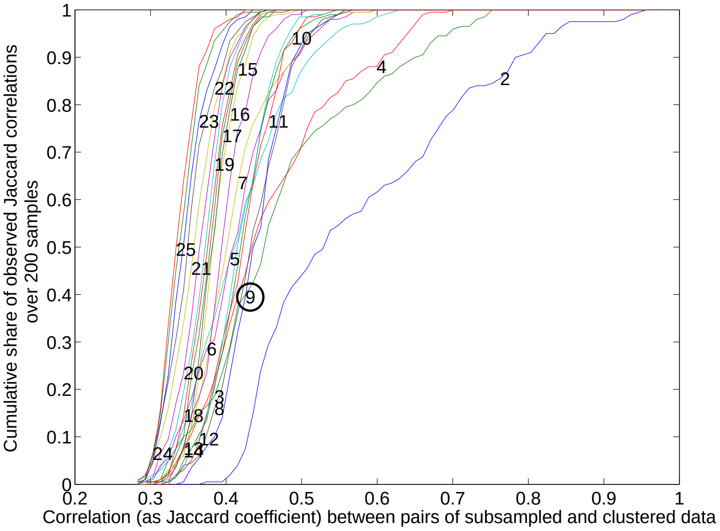


**Figure 2.9:** Stability diagram for determination of the number of clusters according to *Ben-Hur et al.* [16]. The most stable solution is obtained for 2 clusters. Nine clusters prove much more stable than 5, 6 or 7 clusters, and compete with solutions of 3 and 4 clusters.

Although the most stable solution is obtained for partitioning the bioinformatics papers into two clusters, we are looking for a finer-grained clustering. Nine clusters prove much more stable than 5, 6 or 7 clusters, and compete with solutions of 3 and 4 clusters. In particular, nine clusters are more stable than 4 clusters for 55% of subsample pairs (higher *Jaccard* values), and more stable than 3 clusters for 40%. The other pairs are in favor of 3 or 4 clusters. 8 clusters are almost as stable as 9 clusters.

**Conclusion**

Our semi-automatic strategy for determining the number of clusters is based on interpreting dendrograms, Silhouette curves, and stability diagrams. Although the number of clusters remains a difficult to define parameter, our experience is that different strategies often agree on a certain local maximum of performance. For the bioinformatics field we found 9 clusters.

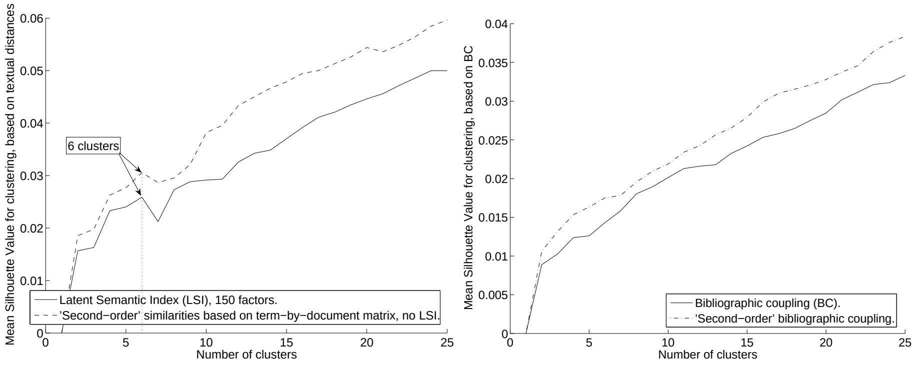
### 2.3.4 Second-order similarities

Mutual similarities between documents are usually stored in a square, symmetric matrix $S_t$ (see Figure 1.14 on page 20). With 'second-order similarities' we mean that the similarity profiles of each document to all other documents (i.e., row or column of $S_t$) are used as input for an extra step of pairwise similarity calculation:

$$S_t^2 = S_t \cdot S_t \tag{2.12}$$

Hence, the ultimate similarity of two documents is based on their respective similarities with all other documents.

This method is related to LSI in the sense that higher-order co-occurrences of terms are also taken into account. It even seems to outperform LSI with 150 factors (see Figure 2.10(a)). Note that a local maximum is present at 6 clusters for both curves in (a). Figure 2.10(b) presents the analogous application to citation information (bibliographic coupling). Somewhat better performance is obtained by taking into account second-order linkages. For more information on bibliographic coupling we refer to Section 3.2.5.



(a)

(b)

**Figure 2.10:** Silhouette curves with mean Silhouette coefficient for clustering solutions of 2 up to 25 clusters. **(a)**. Second-order text-based similarities vs. LSI with 150 factors. **(b)**. Bibliographic coupling vs. second-order bibliographic coupling.

Whereas the performance of LSI clearly drops when the number of clusters exceeds the number of factors (cf. Section 2.3.3), we have observed that the good performance of second-order similarities seems to hold for a wider range of cluster numbers.

However, we do not further investigate this method as it has no good scaling properties, and it requires a lot of memory storage and processing time. Nevertheless, when dealing with a moderately-sized data set (less than 10 000 documents), it is worth to be considered as an alternative to LSI, without associated difficulties of choosing the number of factors.

## 2.4 Co-word analysis. MTA and research agenda setting

### 2.4.1 Introduction

**Do material transfer agreements affect the choice of research agendas? The case of biotechnology in Belgium**

In this case study we examine whether and to what extent material transfer agreements (MTAs) influence research agenda setting in biotechnology.[7] MTAs are signed when proprietary research materials are exchanged between laboratories. Research agendas are mapped through patents, articles, letters, reviews, and notes. Three groups are sampled: (1) documents published by government and industrial organizations that have used research materials received through those agreements, (2) documents published by government and industrial organizations that have used in-house materials, (3) documents published by academia. The challenge is to detect the effect of MTAs on research agenda setting. Methodologically, a co-word analysis is performed to detect if there is a difference in underlying structure between the first two groups of documents. Research agendas are represented by co-word clusters found in titles and abstracts of the sampled documents. This study was inspired by the following questions: Is the research agenda choice modified because of MTAs? Can MTAs encroach on the flow of scientific information and distort the content of research programs? Can a research line be eroded or changed because research laboratories sign MTAs?

A caveat must be stated. The purpose of this study is to demonstrate the use of co-word analysis; technical and methodological issues are in the foreground. We provide a set of techniques, but results should be interpreted with care. Especially when associating sociological hypotheses and implications with the outcome of such statistical methods. Another issue is the small sample size of documents related to material transfer agreements; co-word analysis and clustering techniques rely on statistics that presuppose a sufficient amount of documents. Further research is needed to assess what kind of questions can reliably be answered with these techniques, and what conditions should be met

[7] This study has been published in the journal *Scientometrics* [231].

regarding collection, pre-processing and subdivision of the document set. Experts are indispensable for validating several decisions made with respect to the processing of data as well as for interpretation of the outcome of algorithms. In this study we could not detect an influence of MTA's on research agenda setting and this result also complies with the outcome of a parallel survey. However, we can not make strong claims. The results depend on the document collection at hand and on the assignment of publications to different groups of documents which have or have not used research materials received through MTAs. To decide on these assignments, authors were asked whether or not their publications were related to MTAs. Results also depend on these assertions. Hence, we do not provide strong answers, but give some indications and in fact pose additional questions.

In the following we give some general information about Material Transfer Agreements. Next, we describe the data set and methodology, particularly co-word analysis, clustering, strategic diagrams and the synchronic and diachronic analysis of common terms. Finally, we discuss results and give some remarks that will conclude this case study.

### 2.4.2 Material transfer agreements

Pioneered by industry, MTAs are increasingly used by government and academia. If some provisions are not followed, the contract is breached and the wronged party has the right to bring action against the other, such as suing for damages. Unlike patents or copyrights, MTAs do not rest upon codified legal statutes defining specific rights and obligations [230]. Although there has been no formal agreement on a format when a for-profit entity is providing research material to a non-profit organization, a draft text was compiled in 1992 in the USA. Although MTAs are sometimes necessary, academic researchers as well as policymakers have suggested that the trend towards standardization of MTAs might impede the progress of science and technology by constraining the choice of research agendas. This limitation might be caused by the lack of research materials. Another restriction might be the absence of recipients' freedom to continue a line of research because they no longer own inventions made through the use of the material. Finally, delays or denials to publish research results for which the received material was used, might hamper research agendas.

### 2.4.3 Data

We focus on the core set of *articles*, *letters*, *notes*, *reviews* and *patents* from the database created by *Glänzel et al.* [104]—further on mentioned as documents—which have or have not used research materials received through MTAs. The documents were disclosed between 1992 and 2000 by industry, government, and academia in Belgium (Table 2.3). We represent research agendas through co-word clusters from titles and abstracts of sampled documents. In order to study the effect of MTAs on research agenda setting, we would have to detect whether

**Table 2.3:** Sample of Belgian biotech documents between 1992 and 2000. Note: in our sample, industry is formed by for-profit corporations; government is composed by public research institutes; and academia is constituted by universities and colleges.

| 1992–2000 | Patents | Articles, letters, notes, and reviews | Total documents | Document producers |
|---|---|---|---|---|
| Industry and government | 241 | 255 | 496 | 20 |
| Academia | 88 | 6952 | 7040 | 17 |

there is a difference between documents which used materials through MTAs and documents which used in-house materials. The first group ($F_1$) are documents published by industry and government which used research materials received through MTAs. The second group ($F_2$) is made up of documents published by industry and government which used in-house research materials. The third group ($F_3$) is a set of documents published by academia. Table 2.4 compiles the number of documents corresponding to three different time periods for each group in the sample.

**Table 2.4:** Distribution of documents between 1992 and 2000.

| Period | $F_1$ | $F_2$ | $F_3$ |
|---|---|---|---|
| 1992–1994 | 11 | 113 | 2070 |
| 1995–1997 | 11 | 135 | 2547 |
| 1998–2000 | 20 | 206 | 2423 |

The problem of explaining the role of MTAs in the development of research programs is a difficult one. In this study we have considered co-word analysis as a potential means to address it. A research system dynamically evolves as a result of decisions taken by its parts to engage their activity in a given direction. The co-word analysis technique was developed to assess the degree of convergence of these decisions by analysis of a publications database.

### 2.4.4 Methodology

Titles and abstracts of articles, letters, notes, reviews and patents are transformed into a set of co-word clusters. Co-word analysis offers a flexible way to enter into and to unravel the content structure of a scientific or technological domain. We assume here that cognitive aspects can to some extent be treated quantitatively. The ability to identify themes in a research area by clustering terms from titles and abstracts allows the creation of maps based on cognitive relations between themes [206].

We indexed titles and abstracts of 496 documents in industry and government, and 7040 in academia. To maintain the most important terms for analysis, only terms occurring in noun phrases were kept. The result of indexing was a 2125 × 496 term-by-document matrix for industry and government, and 15 019 × 7040 for academia.

### Term co-occurrence analysis

As it is necessary to have a minimal number of documents to execute the statistical analysis, we opted to group documents in 6 sub-sets as shown in Table 2.4. Each of the sub-sets, composed of documents of a specific group $g$ ($g = 1$ for $F_1$, $g = 2$ for $F_2$, and $g = 3$ for $F_3$) in one of the three periods $p$, was filtered as follows. From the global term-by-document matrix containing binary values ($A_b$), a sub-matrix $A_{b,g,p}$ was constructed for each group and period. Only those documents (columns) that belonged to the sub-set were retained as well as only those terms (rows) that appeared in at least two documents. In addition, only terms having a TF-IDF value larger than or equal to 5 in at least one document of the complete set were kept. From each sub-matrix $A_{b,g,p}$, a term co-occurrence matrix $C_{g,p}$ was constructed by multiplying $A_{b,g,p}$ with its transpose:

$$C_{g,p} = A_{b,g,p} \cdot A_{b,g,p}^T \tag{2.13}$$

Then, following *Callon* [43], each $C_{g,p}$ was converted into an *equivalent index matrix* $E_{g,p}$ by transforming the co-occurrence frequency for two terms $i$ and $j$ into their equivalence or association index $e_{ij}$, by applying the following function:

$$\begin{cases} e_{ij} = 0, & \text{if } c_i = 0 \text{ or } c_j = 0 \text{ or } c_{ij} = 0 \\ e_{ij} = c_{ij}^2/(c_i \cdot c_j), & \text{otherwise,} \end{cases} \tag{2.14}$$

in which $c_i$ and $c_j$ are the respective document frequencies of terms $i$ and $j$ in the sub-set and $c_{ij}$ is their co-occurrence frequency in that sub-set. A last term filter was applied by requiring that the largest equivalence index of a term in a sub-set be higher than 0.2 in order to drop terms that have no strong association with others in the sub-set. Subtracting each equivalence index matrix $E_{g,p}$ from 1, results in a distance matrix that can be used as input for a clustering algorithm.

### Clustering

We applied hierarchical clustering (see Section 2.3, [133]) by considering as input the distance matrix with the complement of equivalence indices for each sub-set. To determine the number of clusters in each sub-set, we inspected four diagrams: dendrograms, stability diagrams, mean Silhouette curves and silhouette plots (see Section 2.3.2). We obtained the cluster numbers shown in Table 2.5

**Table 2.5:** Number of clusters

| Sub-set | 1992–1994 | 1995–1997 | 1998–2000 |
|---|---|---|---|
| $F_1$ | 4 | 3 | 3 |
| $F_2$ | 31 | 30 | 23 |
| $F_3$ | 14 | 21 | 19 |

### Strategic diagrams and dynamics

Once the number of co-word clusters for each sub-set of documents was determined, each cluster was featured by an index of *centrality* and *density*, and plotted in a *strategic diagram* (see Figure 2.11 for an example). *Density* is defined as the mean of the equivalence indices $e_{ij}$ over all term pairs in a cluster (internal links) and *centrality* is the mean of $e_{ij}$ for all possible pairs of words of which one is an element of the cluster and the other is not (external links). These two measures constitute powerful instruments for studying the dynamics of a research network. They enable us to characterize research themes given: (i) their degree of development, i.e., whether or not topics are solidly constituted; (ii) their positions in the network, i.e., whether or not topics are obligatory passage points. A *strategic diagram* can be split into four quadrants based on the classification developed by *Callon et al.* [43], i.e., *I* represents central and visible topics, *II* comprises isolated topics, *III* contains peripheral topics, and *IV* includes unstructured topics. Clusters are represented in the strategic diagram by the term of the cluster which has the highest mean TF-IDF value in a sub-set. For example, in Figure 2.11 the term clusters indicated by *guardcell* and *oocyt* are in the first quadrant and thus are relatively central and well-structured topics for that group and period.

Based on the quadrant in which a cluster or research area is located, it can thus be classified as mainstream research or, on the contrary, as being of secondary importance because being isolated, peripheral, or unstructured in the network [257]. When drawing a series of co-word maps for different periods, dynamic changes of the different sub-fields of a research area become visible.

### Synchronic and diachronic common terms

Research agendas were analyzed synchronically—the relationship among clusters in the same time period—and diachronically—the evolution of clusters over time [53]. How can divergence of research topics between two groups of documents be grasped? Is it plausible to state that difference in underlying structure between both groups of documents means absence of common terms in those two groups of documents? Do synchronic and diachronic common terms tend to introduce relations among co-word clusters? This question suggests that topics located in different strategic diagrams could be cognitively linked to one another despite the fact that, at any given moment in time, these links might not yet be

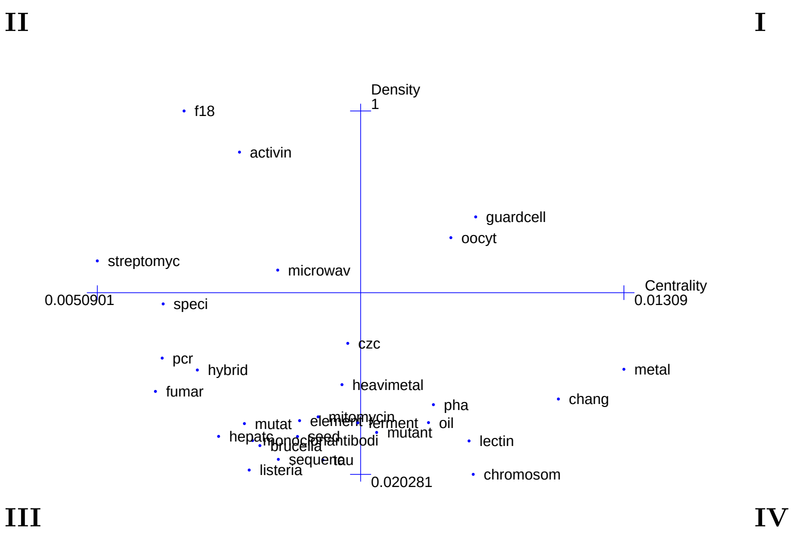


**Figure 2.11:** Strategic diagram for $F_2$ in 1995–1997.

identified. This rather strong assumption might lead to qualitative indications. If we find common terms, does it convey no divergence of research agendas in industry and government? In the affirmative case, how powerful should common terms be to postulate no divergence? Regarding this robustness of common terms, we can use three approaches: the strategic diagram quadrants, mean TF-IDF value and *theoretical ambitiousness.* Firstly, if we split the common words into the four quadrants, then we could consider that central and visible topics are more powerful than the other ones. Secondly, if we order common terms decreasingly according to their mean TF-IDF values, we could consider that higher ones are more powerful than others. Thirdly, if we rank some of the common terms according to the theoretical ambitiousness level of *Rip* and *Courtial* (see Table 2.6) [228], we could consider that those placed at higher levels are more powerful than those placed at lower levels. Here, we only touch upon results, for a more detailed analysis we refer to the manuscript [231].

### 2.4.5 Results and discussion

Synchronically, we looked for common terms suggesting cognitive linkage between $F_1$ and $F_2$ clusters in the three time periods and assessed their robustness using the three different approaches. Newly appearing biotech research themes, mainly measurement and monitoring, were predominant when intersecting term clusters of $F_1$ and $F_2$ in the same periods. Diachronically, when intersecting $F_1$ clusters in a certain period with $F_2$ clusters in a future period, the same observation could be made. Nonetheless, functional explanation and input-output relation were almost as frequent as measurement and monitoring.

**Table 2.6:** Theoretical ambitiousness level of common terms (Source: *Rip* and *Courtial* [228]).

| Level | *Rip & Courtial* | *Weingart & Van den Daele* |
|---|---|---|
| 1 | Screening | Measurement, monitoring |
| 2 | Costs | |
| 3 | Design | Measurement, monitoring<br>Functional explanations, input-output relations |
| 4 | Immobilization | |
| 5 | Product isolation | |
| 6 | Parameter optimization | Functional explanations, input-output relations |
| 7 | Mathematical modeling | Functional explanations, input-output relations<br>Causal explanation, mechanisms |
| 8 | Physical kinetics | |
| 9 | Biokinetics | |
| 10 | Biodynamics | Causal explanation, mechanisms |

As we found some common terms, characterized by high mean TF-IDF, about newly appearing research topics, and usually measurement or monitoring, does it mean absence of deviation of research topics and no effect of MTAs on research agenda setting in industry and government? If MTAs signed in industry and government would have an effect on the research agenda choice in the same sector, would it mean that terms should differ between co-word clusters stemming from documents that used MTAs and those which did not?

Regarding the approach to detect convergence of research agendas, is it sound to decide whether MTAs affect them by just looking at common terms? Before any clustering effort, term selection was performed by implementing a few term filters as described above. These filters do not necessarily pose a problem, but we should keep in mind that they have been applied and that every common term has passed these filters. Some other (common) terms may not have passed them, but then they are not the best descriptors for research in a sub-set.

For validation of results two steps were used. Firstly, practitioners from industry and government were asked to judge the impact of MTAs on choices in their research agendas. Secondly, as they generally did not suffer from MTAs for defining research agendas, we searched for divergence of research agenda between $F_1$ and $F_3$ by using co-word analysis.

Based on the co-word analysis used to detect if the first group of documents overlaps with the third group, we cannot conclude that agreements signed by industry and government affect research agenda setting in academia. The results are in line with the opinion of interviewees from industry and government laboratories who generally do not consider themselves constrained in their choice of research agendas when signing agreements for receiving research materials. On the contrary, MTAs offer important leverage for advancement of their lines of research due to access of materials to carry out the research project.

If MTAs signed in industry and government might have an effect on the research agenda choice in academia, then would it mean that terms should differ between co-word clusters stemming from $F_1$ and $F_3$? Before applying the filters, we compared the vocabulary of industry and government—645 terms—to that of academia—15019 terms—and we obtained 62 non-overlapping terms. So, 90% of $F_1$ terms were included in $F_3$ before filtering. If we look at important terms (after filtering) we still find 31 common terms, or more than 10 percent. Does this mean that there is 'overlap' of research topics between F1 and F3? This modest but existing overlap might perhaps indicate no effect of MTAs signed in industry on research agenda setting in academia.

### 2.4.6 Concluding remarks

The adopted methodology could not detect an effect of MTAs signed in industry and government on research agenda setting, neither in the same sector nor in academia. Nevertheless, strong conclusions can not be drawn. It would be interesting to design another setting in which the purpose is to assess whether the adopted techniques are able to detect known existing divergence. The methodological work undertaken at least indicated that the research themes identified by the co-word technique are relatively stable when using alternative statistical procedures, thereby alleviating the concern that clusters of terms might be nothing more than very unstable statistical artifacts.

## 2.5 Towards mapping library and information science

In this second case study we apply the quantitative linguistic methodology on a set of five journals representing the field of library and information science (LIS), with main focus on IS. Almost 1000 articles and notes published in the period 2002–2004 have been selected for this exercise. The optimum solution for clustering LIS is found for six clusters. The combination of different mapping techniques, applied to the full text of scientific publications, results in a characteristic tripod pattern. Besides two clusters in bibliometrics, one cluster in information retrieval and one containing general issues, webometrics and patent studies are identified as small but emerging clusters within LIS.[8]

### 2.5.1 Introduction

Although bibliometrics has early been applied to study its own field and the field of library and information science [248], relatively few studies have been devoted to general aspects or concept networks of this field. *Bonnevie* has

[8] This study has been published in the journal *Information Processing & Management* [137].

used primary bibliometric indicators to analyze the *Journal of Information Science* [30], while *He* and *Spink* compared the distribution of foreign authors in *Journal of Documentation* and *Journal of the American Society for Information Science and Technology* [121]. Bibliometric trends of the journal *Scientometrics*, another important journal in this field, have been examined by *Schubert* and *Maczelka* [240], *Wouters* and *Leydesdorff* [270], *Schoepflin* and *Glänzel* [237], *Schubert* [239], and *Dutt et al.* [74]. Main journals of the field have also been characterized in terms of journal co-citation and keyword analyses [176, 177]. The co-citation network of highly cited authors active in the field of IR was studied by *Ding*, *Chowdhury*, and *Foo* [68]. Finally, *Persson* analyzed author co-citation networks based on documents published in the journal *Scientometrics* [220, 221]. *Courtial* has studied the dynamics of the field by analyzing co-occurrence of words in titles and abstracts [54]. He described scientometrics as a hybrid field consisting of invisible colleges. The much broader field of LIS is even more heterogeneous and comprises subdisciplines such as traditional library science, IR, scientometrics, informetrics, patent analyses and most recently the emerging specialty of webometrics.

### 2.5.2 Main objectives

The challenge is not the number of articles (almost 1000 full-text articles), but the heterogeneity of this field and the variety of terms and concepts used. According to the observations by *Glänzel* and *Schoepflin* [105], new topics emerged very early in the field and subdisciplines began drifting apart. In order to monitor the situation in the field of LIS about one decade later, we conduct our research along the following questions.

1. Can the heterogeneity be characterized by means of quantitative linguistics?
2. What are the main topics in current research in information science?
3. Have new, emerging topics already developed their own 'terminology'?
4. Can cognitive structure be analyzed using multivariate techniques?
5. How are topics and subdisciplines represented in important journals?

To answer these questions, we elaborate on vocabularies for subdisciplines within LIS, and compare different methods of clustering and mapping in order to reach the optimal presentation of the cognitive structure of the field.

### 2.5.3 Material and methods

We have selected a set of journals with strong focus on scientometrics, informetrics and related specialties. The document set used for our study consists of 938 full-text articles and notes, published between 2002 and 2004 in one of five journals. In particular, Table 2.7 shows the distribution of the 938 documents over selected journals. An overview of the text-based analysis is presented in Figure 2.12. Most of the steps involved have been discussed in Section 2.1. Some additional steps are discussed below.

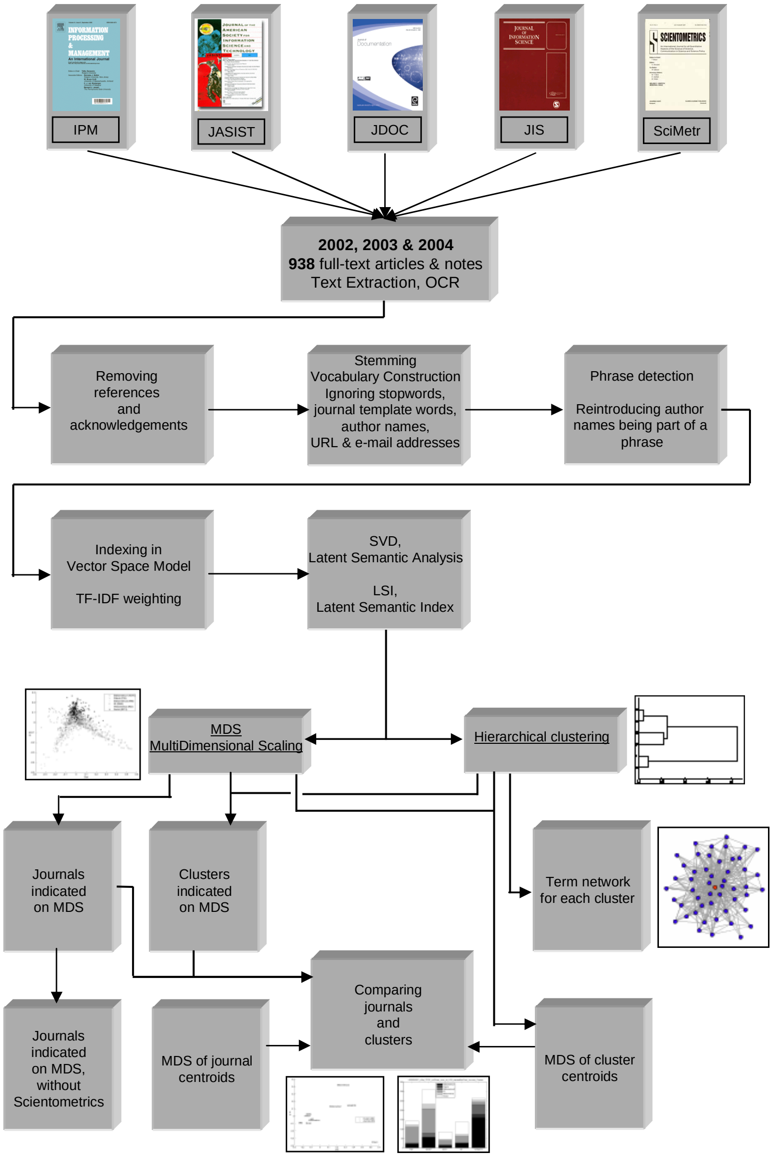

IPM
JASIST
JDOC
JIS
SciMetr
2002, 2003 & 2004
938 full-text articles & notes
Text Extraction, OCR
Removing references and acknowledgements
Stemming
Vocabulary Construction
Ignoring stopwords, journal template words, author names, URL & e-mail addresses
Phrase detection
Reintroducing author names being part of a phrase
Indexing in Vector Space Model
TF-IDF weighting
SVD, Latent Semantic Analysis
LSI, Latent Semantic Index
MDS
MultiDimensional Scaling
Hierarchical clustering
Journals indicated on MDS
Clusters indicated on MDS
Term network for each cluster
Comparing journals and clusters
Journals indicated on MDS, without Scientometrics
MDS of journal centroids
MDS of cluster centroids

**Figure 2.12:** Overall framework of the analysis. The full text of 938 articles and notes from 3 publication years of 5 journals is analyzed in various successive steps that have been discussed in detail throughout this chapter. The arrows indicate the flow of the textual information through different components of the analysis pipeline. Multidimensional scaling and hierarchical clustering are subsequently applied to detect, interpret and visualize different sub-fields of LIS.

**Table 2.7:** Distribution of the 938 articles and notes over 5 selected journals.

| **Journal** | **Number of papers** | **%** |
|---|---|---|
| Information Processing & Management (IPM) | 143 | 15.3 |
| Journal of the American Society for Information Science and Technology (JASIST) | 309 | 32.9 |
| Journal of Documentation (JDoc) | 85 | 9.1 |
| Journal of Information Science (JIS) | 137 | 14.6 |
| Scientometrics (SciMetr) | 264 | 28.1 |
| Total | 938 | 100.0 |

### Text representation and pre-processing

The term-by-document matrix $A$ is transformed into a latent semantic index $A_k$ (LSI). A latent semantic analysis is advisable, especially when dealing with full-text documents in which a lot of noise is observed (for instance, stemming from OCR errors, extracted tables, etc.). Based on the decay of singular values we set the number of factors $k$ to 150. A much lower number of LSI factors (for example, 10) might provide substantially higher Silhouette values (see Section 2.2.3), but results for this case study were obtained prior to experiments concerning the best number of LSI factors. Hence, the number 150 was determined visually and was in accordance with the widely accepted consideration that between 100 and 300 factors is a good choice [62, 24].

**Automatically separating acknowledgements and references from article content** The aim was to analyze only the pure scientific content that is written in the body of a paper and to exclude all bibliographic or other components. Acknowledgements introduce a lot of extra terms relating to institutions, funding agencies, persons, etc. These paragraphs were omitted in order to prevent that similarity of papers could be influenced by, for example, common acknowledged research funding. Article notes and appendices were considered not problematic and no special effort was done to remove them. In practice, they were removed only when they occurred after reference lists.

**Neglecting author names not part of a phrase** Each of an article's references usually has at least one anchor somewhere in the full text. In order not to let cited author names influence text mining and, above all, clustering results, they were semi-automatically removed. However, often an author's name has become eponymic and thus part of a phrase that has much power in describing the content of an article. Some examples are: phrases describing a law (such as *Lotka's law, Bradford's law*), disease (*Alzheimer disease*), model, index (*Price Index*), indicator or method. Such bi-grams were extracted from the texts and added to the phrase list.

### Multidimensional scaling and clustering

In order to get a view of the field, we first applied multidimensional scaling (MDS). Next, agglomerative hierarchical clustering with *Ward*'s method (see Section 2.3) was chosen to subdivide the documents into clusters. A reason for not expecting perfect clustering results is that, for the present study, we only made use of text and discarded all other information. In chapter 4, we asses the performance of different schemes for integration of textual and bibliometric information.

To determine the number of clusters, we used the stability-based method as proposed by *Ben-Hur et al.* [16] and described in Section 2.3.3. A second opinion was offered by observing the plot of mean Silhouette values for 2 up to 25 clusters (as in Figure 2.14). We also observed the dendrogram that resulted from hierarchical clustering of the documents (see Figure 2.15). For ease of interpretation we also made a table of summary statistics, and a term network for each cluster. A term network is mainly intended to provide a qualitative rather than quantitative way of cluster evaluation. It shows the best 50 TF-IDF terms. An edge between two terms indicates that both co-occur in a document of the corresponding cluster, but within a given distance, set to 1 in this case (ignoring stop words). We used Biolayout Java by *Enright* and *Ouzounis* to visualize the term networks [78].

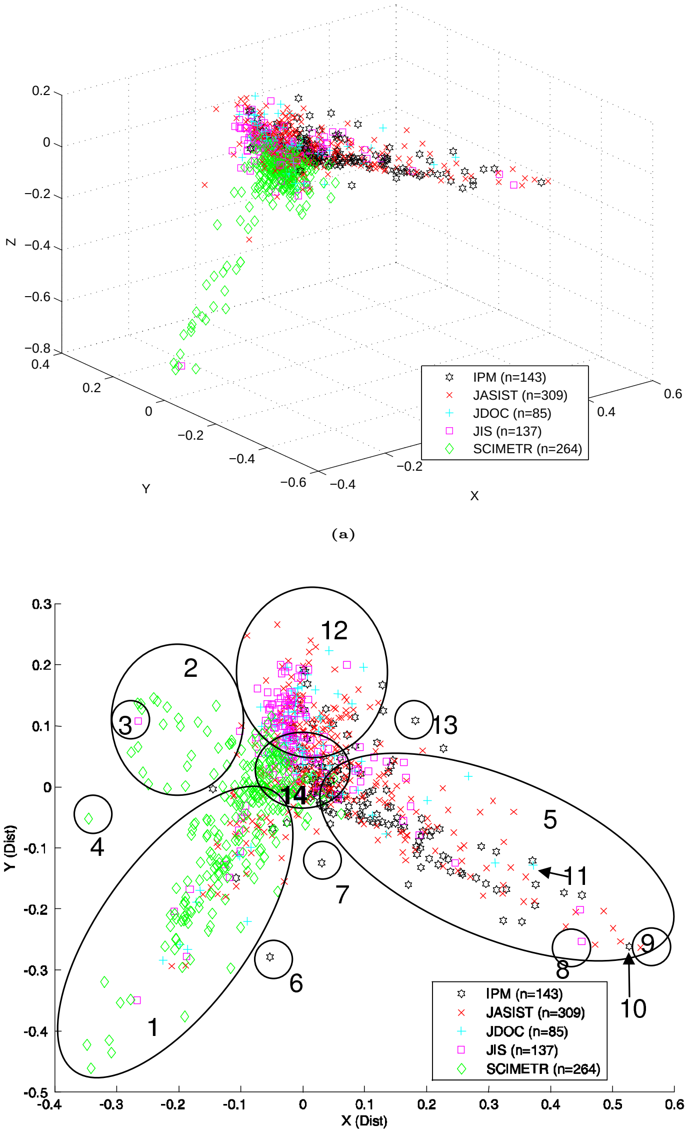
## 2.5.4 Results

### Indexing

The list of detected phrases contained 261 instances and we wrote 58 synonym rules, most of which mapped pairs like coauthor and 'co(-) author', or dealt with acronyms such as wif (web impact factor). An initial indexing phase resulted in a vocabulary of 65 019 terms, but after pre-processing the final index to be used for subsequent analyses only contained 11 151 stemmed terms or phrases. In this index *Zipf*'s curve was cut off to neglect all terms occurring in only one document or in more than 50%. The final term-by-document matrix $A$ was thus of size $11\,151 \times 938$, transformed by LSI to $A_{k150}$ $(150 \times 938)$. Appendix A contains a table with the most important words for each journal and for the whole data set (highest mean TF-IDF scores).

### Visualizing library and information science

Figure 2.13 shows the MDS map of the 938 articles and notes in three and two dimensions. Each of the five journals considered is indicated with a different symbol and color. The journal *Scientometrics* can be largely separated from the other journals (which is also confirmed by different term profiles in Appendix A), and exhibits two different foci (best visible in (b)).



(a)

(b)

**Figure 2.13: (a)**. 3D multidimensional scaling plot of the 938 LIS articles or notes. Each of five journals is indicated with a different symbol and color. **(b)**. Projection of (a) on the X-Y plane. The field of LIS has a tripod shape. The journal *Scientometrics* can be largely separated from other journals and exhibits two different foci. Numbered ellipses indicate (groups of) publications that are discussed in the text.

In what follows, different subsets of papers indicated in Figure 2.13(b) are analyzed in more detail. The first 'leg', indicated by the ellipse with number 1 and by and large containing the first focus of the journal *Scientometrics*, clearly contains papers in bibliometrics. The best 10 TF-IDF terms for 'leg' #1 are: *citat, cite, impact factor, self citat, co citat, scienc citat index, citat rate, isi, countri* and *bibliometr*. The second 'leg of *Scientometrics*', indicated by number 2, is characterized by the best terms *patent, industri, biotechnolog, inventor, invent, compani, firm, thin film, brazilian* and *citat*. The *JIS* paper (#3) embedded in this patent 'leg' might be considered an outlier for that journal, but it was put in the right place since it is concerned with 'The many applications of patent analysis' (Appendix B: Breitzman & Mogee, 2002). One *Scientometrics* paper (#4) seems not to belong to either focus. Indeed, it is about 'Patents cited in the scientific literature: An exploratory study of reverse citation relations' (Appendix B: Glänzel & Meyer, 2003). An important focus of LIS is indicated by ellipse #5 and can be profiled as 'Information Retrieval' (IR) when looking at the highest scoring terms: *queri, search engin, web, node, music, imag, xml, vector* and *weight*. 'Interdisciplinary' *IPM* papers (#6 and #7), between ellipses #1 and #5, are the following: 'Mining a Web citation database for author co-citation analysis' (Appendix B: He & Hui, 2002) and 'Real-time author co-citation mapping for online searching' (Appendix B: Lin et al., 2003). While *Scientometrics* seems to have never published any paper specifically about IR, all other considered journals have (e.g., papers #8, #9, #10 and #11 in Figure 2.13(b)). The fourth distinguishable subpart of LIS (#12) is about *digit, internet, servic, seek, behaviour, health, knowledg manag, organiz, social* and *respond*; and hence encompasses more social aspects. The *IPM* paper bridging the gap between IR and more social oriented research (#13) is entitled 'The SST method: a tool for analyzing Web information search processes' (Appendix B: Pharo & Jarvelin, 2004). The goal of that paper was 'to analyze the searching behaviour, the process, in detail and connect both the searchers' actions (captured in a log) and his/her intentions and goals, which log analysis never captures'. The remaining large subpart is somewhat the central part (#14). It consists of papers leading to a mean profile containing the terms *web, web site, classif, domain, web page, languag, scientist, region, catalog* and *web impact factor*.

### Clustering full-text articles to map LIS

We have experimented with the $k$-means clustering algorithm, but as expected, the hierarchical algorithm seemed to outperform it, even when using a more intelligent version of $k$-means in which the means are initialized by a preliminary clustering on a 10% subsample and in which the best out of ten runs is selected ($k$-means can get stuck in local minima). At first sight, however, $k$-means did seem to do well: the cluster Silhouettes were a bit nicer than for hierarchical clustering. However, upon closer investigation some undesirable effects showed up due to the nature of the algorithm. For instance, one of the clusters was a combination of papers about patents and papers about music information retrieval (MIR). This was definitely a spurious merge of clusters relatively far from other clusters. The reason why they ended up in one cluster is probably the averaging effect of $k$-means. In every step of iteration, each document is assigned to the cluster with closest mean and each mean is updated to become the average document profile in its cluster. The MDS diagram of documents in that cluster indeed showed two very different orientations. The clustering results and MDS diagrams in this section will corroborate that patent papers are closer to bibliometrics papers and MIR papers closer to information retrieval, which complies better with our intuition.

Although the stability plot exhibited no obvious number of clusters, a few observations could be made. Partitioning documents into two groups resulted in the most stable solution. However, we were looking for a somewhat finer-grained clustering solution. Other relatively stable options were 3, 4, 5, 6, and maybe even 7 clusters, but not more. A second opinion was offered by observing the plot of mean Silhouette values, assessing the overall quality of a clustering solution for 2 up to 25 clusters (see Figure 2.14). It is clear that a local maximum was present at six clusters.

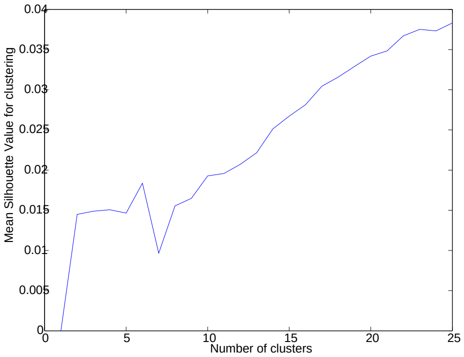


**Figure 2.14:** Mean silhouette coefficient for solutions of 2 up to 25 clusters, with local maximum at 6 clusters.

After the sharp drop at 7 clusters, the mean Silhouette value increases again and from 10 clusters onwards it is larger than the value for 6 clusters, but according to the stability diagram those clustering solutions are less stable. Hence, we chose 6 as the number of clusters. However, the overall mean Silhouette values each seem low, again hinting at groups of documents that are not clearly separable. Nevertheless, a standard $t$-test for the difference in means revealed that the difference between mean Silhouette values for 6 and 7 clusters was statistically significant at the 1% significance level ($p$-value of $2.25 \cdot 10^{-7}$).

Figure 2.15 depicts the dendrogram that resulted from hierarchical clustering, cut off at 25 clusters. The vertical line illustrates the cut-off point for our 6 clusters with best terms *countri, patent, citat, queri, web site* and *seek* for c1 to c6, respectively. For each of 25 clusters, the best mean TF-IDF term is shown.

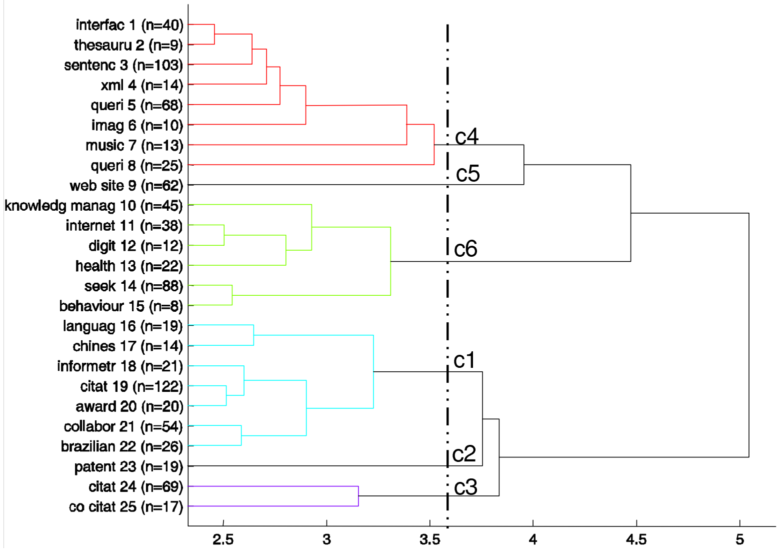


**Figure 2.15:** Dendrogram, cut off at 25 clusters on the left-hand side and at 6 clusters (c1–c6) at the vertical line, for hierarchical clustering of 938 LIS papers. For each of 25 clusters, the best mean TF-IDF term is shown.

The 6 clusters formed two groups according to their size, particularly three large clusters with more than 200 papers each and three small ones with less than 100 articles each. The large clusters are Cluster 1, manually labeled as 'Bibliometrics1', Cluster 4, labeled as 'IR', and Cluster 6, labeled as 'Social'. Cluster names have been chosen based on the terms representing these clusters. Figure 2.16 shows the term network for the IR cluster (c4). We refer to the published manuscript [137] for the networks and for a detailed analysis of other clusters.

The term network of **Cluster 1** indicates that these papers are concerned with domain studies, studies of collaboration in science, citation analyses, national research performance and similar issues. Indeed, analysis of the papers close to the medoid, representing about 20% of all papers in the cluster, confirmed this assumption. The medoid is a paper by *Persson et al.* on 'Inflationary bibliometric values: The role of scientific collaboration and the need for relative indicators in evaluative studies' (Appendix B: Persson et al., 2004). Besides application, this cluster also comprises the sociological approach, technical questions in the context of bibliometrics and IR, as well as database-related aspects.

The smaller bibliometrics cluster (**Cluster 3**, manually labeled as 'Bibliometrics2') is of more methodological/theoretical nature. The medoid document is the state-of-the-art report 'Journal impact measures in bibliometric research' (Appendix B: Glänzel & Moed, 2002).

The term networks for the two bibliometrics clusters just described contain a few overlapping terms (*bibliometr, chemistri, citat, citat rate, cite, cluster, coun-*

*tri, impact factor, isi, physic, rank* and *scienc citat index*). An MDS plot that only considers the two Bibliometrics and the Patent clusters (not shown) confirmed that there is no clear border between Bibliometrics1 and Bibliometrics2, but that there is a gradual transition between methodology and application. One of the papers was clearly an IR paper about collaborative filtering and should even have been put in another cluster. But by application of the *Porter* stemmer [225], the stem for 'collaborative' (*collabor*) is the same as for 'collaboration', which was the second most important term for the Bibliometrics1 cluster. This might just serve as an example for which incorporation of link information in the clustering process might prevent the spurious association with the Bibliometrics1 cluster.

The small **Cluster 2** (19 papers) represents patent analysis. A paper on 'Methods for using patents in cross-country comparisons' forms the medoid of this cluster (Appendix B: Archambault, 2002). This cluster proved to be homogeneous; all papers are concerned with technology studies, linkage between science and technology, and are at least partially relying on patent statistics. The MDS plot that only considers the two Bibliometrics and the Patent clusters (not shown) indicated that the Patent cluster is much closer to Bibliometrics1 than to Bibliometrics2. The dendrogram of Figure 2.15 reveals that Bibliometrics1 ('c1') indeed is combined first with Patent ('c2') before being combined with Bibliometrics2 ('c3').

**Cluster 4**, with 282 papers, is the largest one. We have labeled it 'Information Retrieval'. The medoid paper is entitled 'Querying and ranking XML documents' (Appendix B: Schlieder & Meuss, 2002). The full spectrum of IR related issues can be found here. Both theoretical and applied topics are represented. Music retrieval is also covered by this cluster; among others, all papers of the special issue on music information retrieval (*JASIST* 55 (12), 2004) can be found here.

**Cluster 5**, with 62 papers, belongs to the small clusters. Both terms and papers close to the medoid characterize this cluster as 'Webometrics'. The medoid paper is entitled 'Motivations for academic Web site interlinking: evidence for the Web as a novel source of information on informal scholarly communication' (Appendix B: Wilkinson et al., 2003).

**Cluster 6** (213 papers) proved to be the most heterogeneous cluster. We have labeled it 'Social', however, we could also have called it 'General & miscellaneous issues'. 'Approaches to user-based studies in information seeking and retrieval: a Sheffield perspective' is the title of the medoid paper (Appendix B: Beaulieu, 2003). Knowledge management, social information, evaluation of digital libraries, user feedback, user requirements for information systems, special aspects of IR such as contexts of information seeking, gender issues, use of internet facilities, etc., are among the topics covered by this cluster.

Table 2.8 shows the share of documents in each cluster and the share of terms or phrases from the complete vocabulary that have been used in one or more of the included papers. Next, the percentages of terms that are among the 5% best

TF-IDF terms for a cluster, and which are also present in the list of 5% best terms of another cluster, are indicated. The most frequently common terms are *citat, cluster, web, countri, domain, scientist, search engin, chemistri, queri, score, map, compani, industri, internet, task, bibliometr, collabor* and *china*.

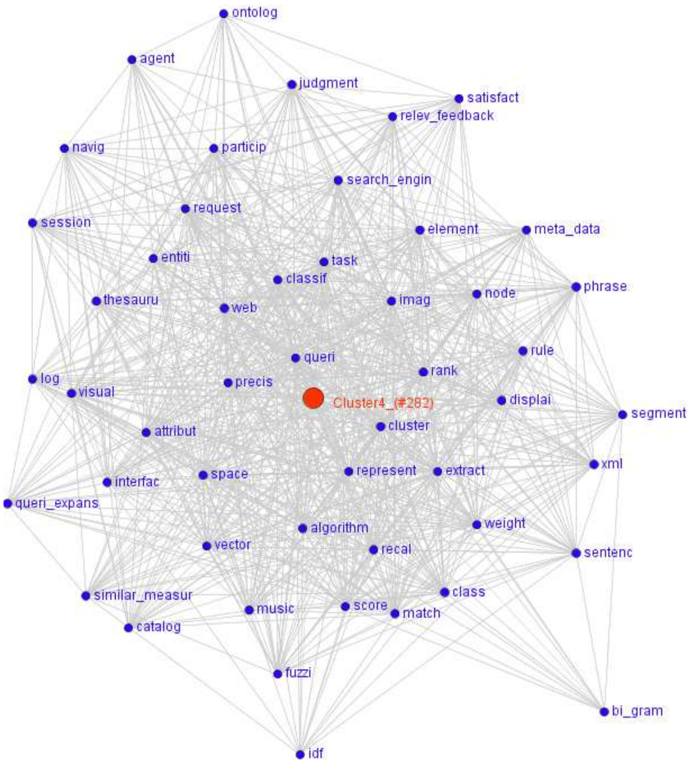
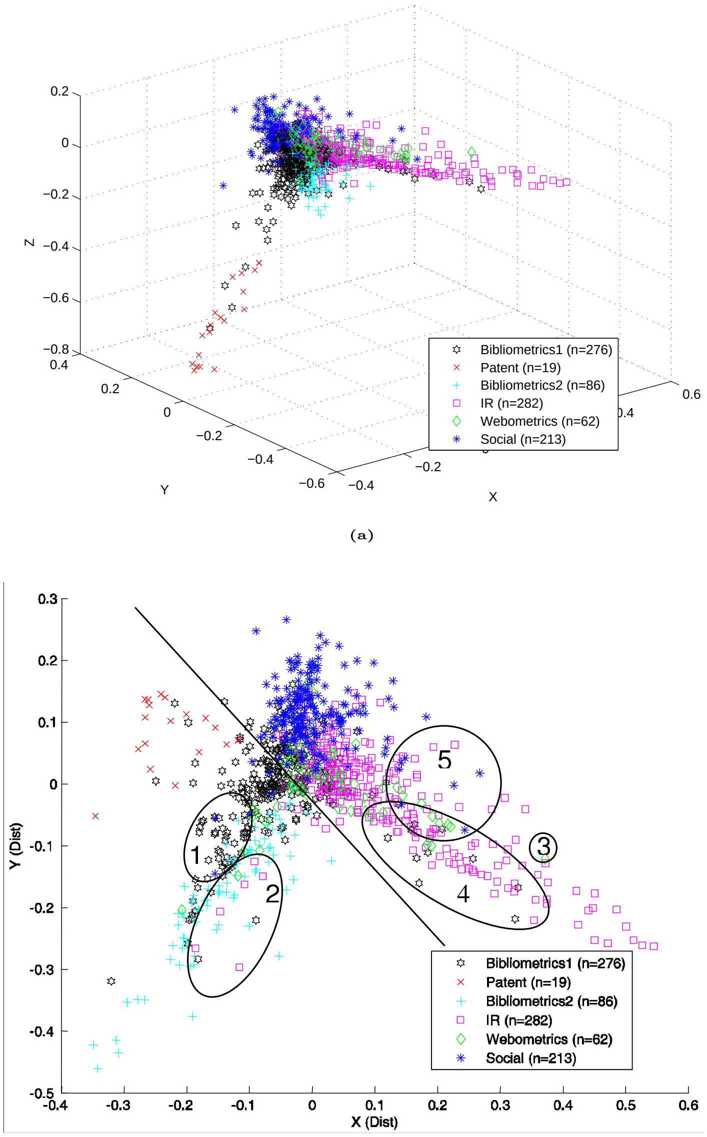


**Figure 2.16:** Term network for Cluster 4 (282 documents), labeled as 'IR' (Information Retrieval). The best 50 TF-IDF terms for the cluster are shown. An edge between two terms indicates that both co-occur next to each other in at least one document of the corresponding cluster (ignoring stop words). The full spectrum of IR related issues can be found in this cluster. Both theoretical and applied topics are represented. Although 'traditional' information retrieval is covered as well, Web search related issues are in the foreground. Music retrieval is covered by this cluster as well.

Figures 2.17(a) and (b) show the same MDS maps as in Figure 2.13, but now clusters instead of journals are indicated. Note that there is no correspondence between journals and clusters with the same symbol or color.

The Patent cluster can be clearly separated from the rest of LIS. The subspace under the line is almost completely occupied by Bilbiometrics1, Bibliometrics2 and Patent. We verified papers that were put in a seemingly suspicious group. Firstly, there are Social papers located in the middle of the bibliometrics



(a)

(b)

**Figure 2.17: (a)**. 3D multidimensional scaling plot of 938 LIS articles or notes. **(b)**. Each of six clusters is indicated with a different symbol and color. Projection of (a) on the $X$-$Y$ plane.

**Table 2.8:** Share of documents and terms in each cluster and share of the 5% best terms in common with other clusters

| Cluster number & name | Share of documents (%) | Number of terms (%) | Share (%) of 5% best TF-IDF terms in common with cluster | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | 1 | 2 | 3 | 4 | 5 | 6 |
| 1. Bibliometrics1 | 29.4 | 71.4 | - | 14 | 29 | 27 | 14 | 32 |
| 2. Patent | 2.0 | 21.0 | 46 | - | 25 | 17 | 12 | 22 |
| 3. Bibliometrics2 | 9.2 | 44.5 | 46 | 12 | - | 27 | 12 | 17 |
| 4. IR | 30.0 | 70.0 | 27 | 5 | 17 | - | 13 | 29 |
| 5. Webometrics | 6.7 | 27.1 | 38 | 9 | 21 | 34 | - | 25 |
| 6. Social | 22.7 | 72.2 | 32 | 6 | 10 | 29 | 9 | - |
| Total | 938 documents | 11 151 distinct terms | | | | | | |

clusters (#1). Were they rightfully added to the Social cluster? Remember that no 'fuzzy clustering' was performed, so a paper could be attributed to only one cluster. The titles gave a clue about the social scope of the papers. Next, based on the titles it could be concluded that the six IR papers that are as well embedded in bibliometrics space (#2) are indeed closely related to IR and thus correct members of that cluster.

The paper that has most deeply infiltrated the IR space though still belongs to Webometrics (#3) is 'Automatic performance evaluation of Web search engines', by *Can et al.* (Appendix B: Can et al., 2004). Again a straightforward choice. However, for most of the 11 Bibliometrics1 papers grafted onto the IR 'leg' (#4) it was not clear why they were put in the Bibliometrics1 instead of the IR Cluster. The use of common words among IR and Bibliometrics1 papers (such as *cluster, english, arab, entropi, sample, poisson*) might have contributed to the high similarity of those papers to the Bibliometrics1 cluster.

By observing the plot of individual Silhouette values for each paper in the Bibliometrics1 cluster (not shown), it was apparent that the Bibliometrics1 cluster, besides being the second largest cluster, contained the highest share of negative Silhouette values. This means that the corresponding papers had better be put in another cluster. The worst score, as low as -0.4, indicated that this paper was definitely put in a wrong cluster. This might be an illustration of an early wrong merger by the agglomerative hierarchical clustering algorithm [146]. Since negative Silhouette values can be detected, the corresponding documents and interpretations can be handled with care.

Finally, the mixed character of about half of the 7 papers of the Social cluster that are also most connected to the IR field (#5), was obvious. The centroid of a cluster is defined as the linear combination of all documents in it and is thus a vector in the same vector space. For each cluster, the centroid was calculated and the MDS of pairwise distances between all centroids is shown in Figure 2.18. As expected, the Patent cluster is the most separated one, and closest to the bibliometrics clusters. The more applied Bibliometrics1 cluster is closer to IR and Social than Bibliometrics2 is. Webometrics is, however, somewhat closer to the more methodological Bibliometrics2 cluster.

### Clustering without LSI

As already mentioned in Section 2.2.3, the number of factors for the latent semantic index is difficult to account for. In order to assess the effect of LSI on the clustering results, Figure 2.19 compares the cluster centers found as described at the end of the previous section (using 150 LSI factors, clustering '*A*' further on) to those of a clustering not using LSI but on the plain term-by-document matrix (clustering '*B*'). In the latter case, there is one extra cluster ('Cluster 7') because the plot of the mean Silhouette coefficients (as in Figure 2.14, but not shown here) revealed a local maximum for 7 clusters. The LSI transformation seems not to have that much influence as most of the *A* clusters correspond to one *B* cluster, except for the new cluster. When analyzing its contents, we observed that Cluster 7 is a dense cluster containing 14 documents about music information retrieval (MIR) with the largest mean Silhouette coefficient of all seven clusters. The terms with highest mean TF-IDF score are: *music, audio, pitch, mir, melodi, song.* Cluster 7 contains the complete special issue of *JASIST* about 'Sound Music Information Retrieval' (*JASIST* 55 (12), 2004), another *JASIST* paper and one paper from *IPM*, *SciMetr* and *JIS*. The *JIS* paper, being the medoid or the closest paper to the centroid and thus the most characteristic for the cluster, is a paper of *Aura Lippincott* about 'Issues in content-based music information retrieval' (Appendix B: Lippincott, 2002). Cluster 7 is closest to Cluster 6 and on the dendrogram (not shown), Cluster 7 is first combined with that Cluster 6, which is very close to the IR cluster from clustering *A*. Moreover, that IR cluster contains all 14 papers of Cluster 7 (MIR) of clustering *B*.

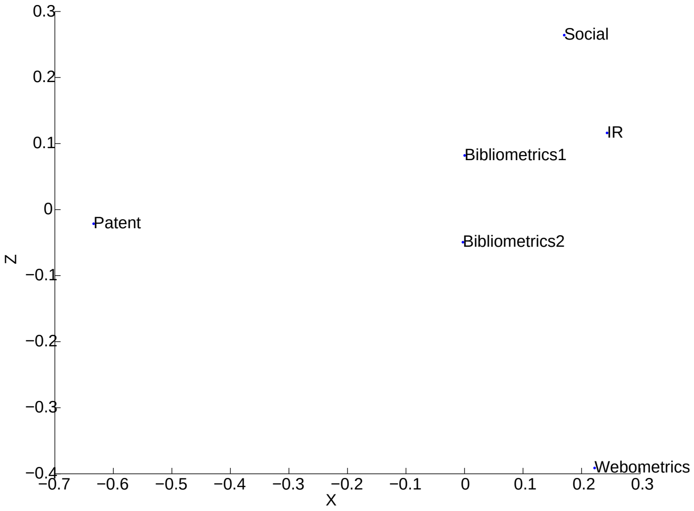


**Figure 2.18:** MDS plot showing distances between the centers (centroids) of the six clusters.

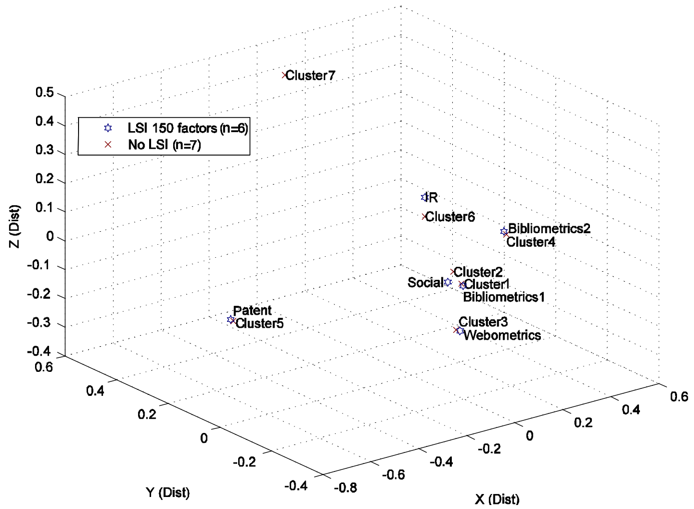


**Figure 2.19:** MDS plot comparing the cluster centroids of the six clusters found in the LSI-transformed concept-by-document matrix (150 factors, clustering '*A*'), with the seven cluster centers when not using LSI (clustering '*B*').

Now, why was the number of clusters higher when LSI was not used? Why was MIR then considered a separate cluster? A possible explanation is that it is an illustration of the power of latent semantic indexing to identify the general concept of information retrieval and the fact that music information retrieval is included as a part of it. Indeed, the most important terms in the MIR cluster are all very specifically about music, but because of the (possibly higher-order) co-occurrences with a lot of general information retrieval terms, they are mapped on the same LSI factors (each is a linear combination of terms). Looking at the dendrogram of Figure 2.15, in the case of LSI, the MIR cluster is only split off when asking for 8 or more clusters (in this case it only contains 13 papers, the paper that was most distant from the centroid here belonging to another cluster). As we were trying to understand the field of LIS and looking for overall patterns, we preferred the solution in which a highly specific and, in this data set at least, temporary cluster like MIR was considered part of the more general concept of IR. Thus, we deem it an advantage of LSI, next to its general noise reduction capabilities.

### 2.5.5 Comparing journals and clusters

The two-dimensional projection of Figure 2.20 provides interesting insight in the journal presentation of LIS. IR and *IPM* collide in this 2D projection. This means that Cluster 4 ('IR') is very close to the scope of this journal. The 'Social' cluster with general and miscellaneous topics, as well as 'Webometrics', are close to *JIS*, *JDoc* and *JASIST*, too. Moreover, the 'Social' cluster is almost

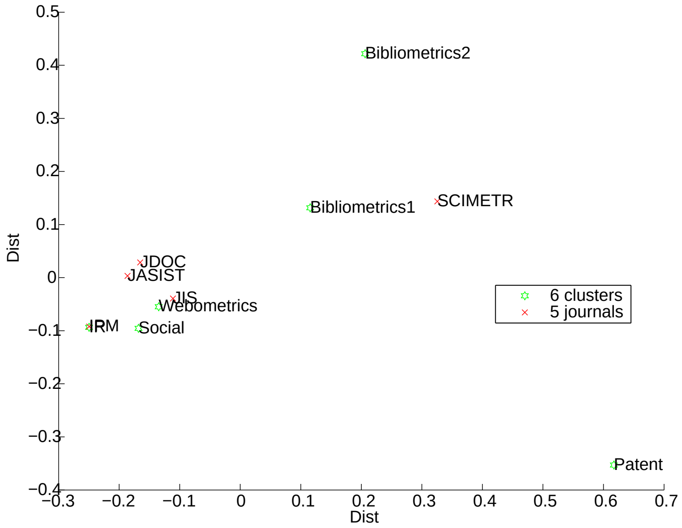


**Figure 2.20:** MDS plot with six cluster centroids and five journal centroids.

equidistant to all traditional journals in information science. Although this is a 2D-projection, we can conclude that those three clusters are mainly represented by the four above-mentioned journals. The remaining three clusters, namely Bibliometrics1, Bibliometrics2 and Patent, form a triangle in the center of which the journal *Scientometrics* is located. The relatively large distances among these clusters and between each cluster and the journal strongly indicate that a quite large spectrum of bibliometric, technometric and informetric research using different vocabularies is covered by the journal *Scientometrics*. This observation is in line with the findings by *Schoepflin* and *Glänzel* [237] that scientometrics consists of several subdisciplines such as informetric theory, empirical studies, indicator engineering, methodological studies, sociological approach, and science policy; and that case studies and methodology became dominant by the late 1990s. At the end of the 1990s, technology related studies based on patent statistics also became an emerging subdiscipline in the field. This trend was confirmed by the size of the bibliometric/technometric clusters (see Section 2.5.4). The patent cluster, still the smallest one, has the largest distance from all other clusters. For a visualization of the share of each journal's papers in the different clusters and the share of the clusters' papers published in the five journals, we refer to the manuscript [137].

### 2.5.6 Discussion and conclusion

We have analyzed the concept structure of five journals representing a broad spectrum of topics in the field of library and information science (LIS). We have focused on the analysis of the 'pure' text corpus, excluding any bibliographic or bibliometric components which might influence the quantitative linguistic analysis of the scientific text. We have excluded author names (except for eponyms), addresses, cited references, journal information and acknowledgements, which might otherwise already have provided cognitive links to other relevant literature. We have applied different techniques of clustering and visualization of the structure of the field and of its journals.

Cluster-stability analysis according to *Ben-Hur* and the mean Silhouette value (see Section 2.3.2) resulted in an optimum of six clusters for the selected journals and for the period 2002–2004. We have found two clusters in bibliometrics, of which a big one in applied bibliometrics/research evaluation and a smaller one in methodological/theoretical issues; we have also found two large clusters in IR and general and miscellaneous issues and, finally, two small emerging clusters in webometrics and patent and technology studies. Within the IR cluster, we have found a small subcluster on music retrieval.

The combination of cluster analysis, MDS, and journal assignment has revealed interesting details about cognitive journal structure and cluster representation by journals. The about 1000 LIS papers form a characteristic 'tripod' in the 3D multidimensional scaling plot. According to expectations, IR, General issues and Webometrics were represented by four of the five journals, namely *JIS*, *IPM*, *JASIST* and *JDoc*, whereas the two bibliometrics and the patent clusters were the domain of the journal *Scientometrics*. The papers published in *Scientometrics* were arranged in two of the three legs forming the tripod. The 'two legs' were formed by Bibliometrics1 and Patent on the one hand, and Bibliometrics1 and Bibliometrics2 on the other hand. The border between the two bibliometrics clusters is fuzzy; there is a gradual transition between methodology and application. From the viewpoint of concept structure, patent analysis can be considered an extension of evaluative bibliometrics. Moreover, the cluster dendrogram has shown that Bibliometrics1 is combined first with Patent, before being combined with Bibliometrics2.

## 2.6 Concluding remarks

In this chapter we have presented the text mining framework that has been developed in the course of this thesis. We have discussed at length the adopted Vector Space Model, including all necessary pre-processing, indexing and weighting steps. A reduction of dimensionality by feature selection, Latent Semantic Indexing (LSI), and Random Indexing (RI) was also described, which is indispensable because of the *curse of dimensionality.*

Our combined semi-automatic strategy for determining the number of clusters is based on a combination of distance-based and stability-based methods. The stability-based method of *Ben-Hur et al.* is used to determine a statistically optimal number of clusters [16]. A second opinion is offered by observing the dendrogram in order to find an appropriate cut-off level. In addition, a local maximum is sought in the curves with mean text-based and citation-based Silhouette values for various numbers of clusters. Finally, the quality of clustering solutions can be verified by a plot with Silhouette values for all objects.

We have contributed to an important open research problem in LSI research, namely the debate about the number of LSI factors. We investigated the relationship between number of factors, number of clusters, and clustering performance. In general, for the bioinformatics data set the clustering performance was significantly higher for a smaller number of factors. It was put forward that a very modest number of factors might deliver local maxima in clustering performance, on condition that there are no fewer LSI factors than the desired number of clusters. However, this should be further assessed using other corpora as well. A limited number of factors has also direct advantages in terms of storage needs and processing time. Our observations are also supported by a recent study of *Kontosthatis* [152]. Interestingly, LSI and RI to some extent model semantics by mere mathematical processing. Besides RI, 'second-order similarities' have been introduced as an alternative to LSI. It does not suffer from the need to determine a number of factors. Unfortunately, the method has bad scaling properties and should thus merely be considered illustrative.

The introduced algorithms have also been demonstrated in two case studies. Firstly, a co-word analysis was performed as a means to elucidate the effect of material transfer agreements (MTAs) on research agenda setting in biotechnology. The analysis of strategic diagrams and their dynamics, and of synchronic and diachronic common terms could not indicate that MTAs signed in industry and government affect research agenda setting, neither in the same sector nor in academia. Nevertheless, strong conclusions could not be drawn.

In the second case study, a spectrum of data mining techniques was used to unravel and visualize the concept structure of the field of library and information science (LIS). We have focused on the analysis of the 'pure' textual content present in 5 journals, excluding any bibliographic or bibliometric components which might influence the quantitative linguistic analysis of the scientific text. The optimum solution for clustering LIS was found for six clusters. However, the goal of Chapter 4 is to integrate text mining and bibliometric techniques in the hope for even better performance. The LIS case study will be revisited in that chapter and a hybrid analysis will point towards 5 instead of the current 6 text-based clusters by merging two bibliometrics clusters.

To conclude, statistical and mathematical techniques prove powerful methods to map knowledge embedded in texts. Nonetheless, the question arises whether indicators of cited references, bibliographic coupling, and cross-citations among subclusters might be appropriate tools to improve the mapping methodology by combining text analysis with bibliometric methods.

# Chapter 3

# Bibliometrics and network analysis

The goal of this chapter is to provide another view on information present in large-scale corpora of scientific publications and patents. Whereas Chapter 2 discussed the use of text mining techniques for mapping knowledge domains, this chapter will focus on analysis of networks that emerge from many individual acts of authors reading and citing other scientific works or collaborating in the same research endeavor. These extremely large networks of science and technology (S&T) often exhibit self-organizing properties and can be analyzed with techniques from bibliometrics and graph theory in order to rank important entities, and for clustering, extraction of communities, collaborative filtering, etc. The science of evolving networks can even contribute to the detection of emerging and converging clusters representing scientific specialties, new technologies, and hot topics.

Sections 3.2 and 3.3 introduce and briefly describe some bibliometric indicators and types of networks that we consider. Next, Section 3.4 discusses the HITS and PageRank algorithms used in information retrieval for identification of important authorities and hubs. Section 3.5 touches upon graph partitioning and detection of communities. Finally, in Section 3.6, we apply a selection of techniques to the interdisciplinary field of bioinformatics. The scope of this chapter is limited to some important aspects of bibliometrics and link analysis and it is certainly not intended to provide an exhaustive survey of these most interesting research topics.

Chapter 4 is devoted to the investigation of possible ways to incorporate this network thinking with text mining algorithms of Chapter 2, in order to come up with a hybrid methodology for information retrieval and for mapping of fields of science. The integration method with the best properties is then deployed in Chapter 5 to provide a hybrid and dynamic clustering of the bioinformatics field. There, textual and bibliometric properties of resulting clusters and *cluster chains* are analyzed in detail.

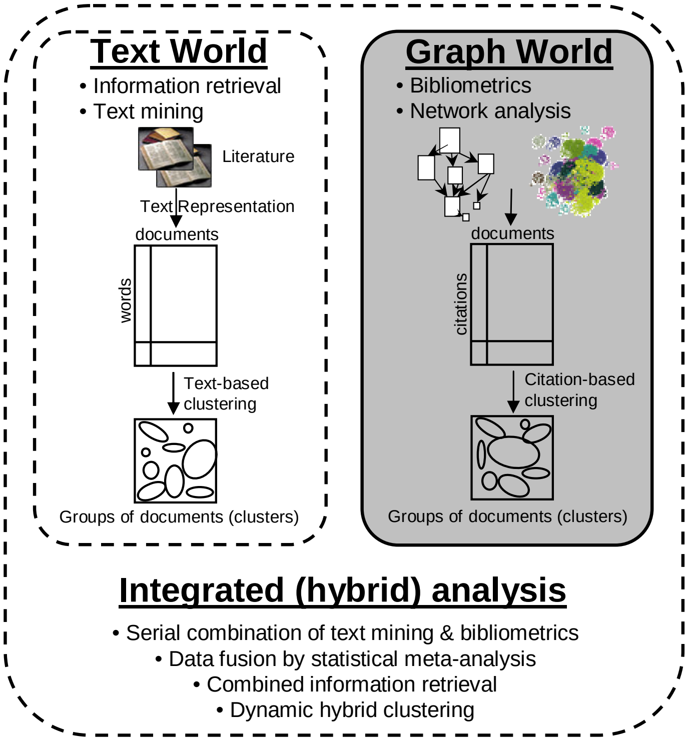

Text World
• Information retrieval
• Text mining
Literature
Text Representation
documents
words
Text-based clustering
Groups of documents (clusters)
Graph World
• Bibliometrics
• Network analysis
documents
citations
Citation-based clustering
Groups of documents (clusters)
Integrated (hybrid) analysis
• Serial combination of text mining & bibliometrics
• Data fusion by statistical meta-analysis
• Combined information retrieval
• Dynamic hybrid clustering

**Figure 3.1:** Whereas Chapter 2 discussed the use of text mining techniques for mapping knowledge domains, this chapter is focused on bibliometrics and the analysis of large citation or collaboration networks. A selection of techniques is applied to the field of bioinformatics.

## 3.1 Introduction

Bibliometrics is an interdisciplinary science in which statistical and mathematical indicators, methods and models are used to study written scientific communication or more general information, mostly collected in large databases containing scientific publications or patents [96, 90, 32, 207, 206, 237]. The purpose of bibliometrics is to measure activity in and structure and evolution of science and technology, as well as the connection between both realms [132, 61, 13, 166]. Although somewhat more general in scope, *bibliometrics* is often used today synonymously with *scientometrics*. The dynamics of this field have been studied by *Courtial* who analyzed the co-occurrence of words in titles and abstracts [54]. He described scientometrics as a hybrid field consisting of invisible colleges. We refer to Section 2.5 for a more detailed computational linguistic analysis of the field.

In bibliometrics a lot of indicators are used pertaining to publication and citation statistics, which provide a means to quantitatively analyze science and technology structure, performance, and evolution. The bibliometric mapping and monitoring offer important tools that give quantitative input to support and supplement science and technology policies and innovation management [207, 171]. The measurement of research performance of authors, institutions and nations is increasingly important for strategic positioning, for the evaluation of publications and scientific journals, and for an optimal use of funding. The micro, meso, and macro levels of aggregation can be distinguished. At the micro level, subjects of analysis are individuals or research groups, whereas the meso level studies journals and institutions, and the macro level addresses regions, countries or even groups of countries.

Because bibliometrics investigates the structure and evolution of various types of citation and collaboration networks, it is much related to *link* or *network analysis* and *graph theory* in general, and much cross-fertilization occurs between both fields. Since the last decade, a lot of research has been conducted regarding analysis of large-scale directed and undirected graphs and their statistical and dynamic properties [250, 6, 71, 198, 35, 178, 213].

Graph analytic algorithms are very popular in data mining, pattern recognition, strategic positioning, trend detection, science and technology policies, fraud detection, analysis of financial networks, epidemiological research, intelligence services, etc. Different algorithms for visualization of evolving networks have been compared by *Chen* and *Morris* [47, 194].

A lot of networks exhibit self-organizing phenomena. They are characterized by absence of regulation in the form of planned global organization, but with structure emerging from an enormous amount of local interactions. Local and global characteristics of networks help to define network topologies such as small worlds [182]. A *small-world* network is a network that is to a large extent locally clustered and in which the average shortest path between two nodes or vertices is small, even when the size of the network grows very large

[268]. *Preferential attachment* to nodes with a high *degree*, i.e., already having a lot of connections, introduces a popularity bias and is the major cause of the small-world phenomenon, leading to a dynamic of *rich-get-richer* in which newcomers mostly attach to well-connected nodes [11]. Various models for evolving networks are based on growth and preferential attachment [12, 140].

The degree distribution $P(k)$ gives the probability that a random node from the network has $k$ connections with other nodes, or, in other words, $P(k)$ is the fraction of vertices having degree $k$. Different types of degree distributions can be distinguished when plotted on a *log-log* scale with the logarithm of degree on the $X$-axis and the logarithm of the number of nodes with this degree on the $Y$-axis. A degree distribution generated by preferential attachment has a fat tail for the relatively smaller number of nodes with unusual high connectivity. In a small-world network the degree distribution $P(k)$ follows a power law, i.e.,

$$P(k) \sim k^{-\gamma}, \tag{3.1}$$

and therefore leads to a straight line in the *log-log* plot. The term *scale-free* network refers to the fact that the degree distribution $P(k)$ remains unchanged up to a multiplicative factor under a rescaling of $k$. Such power law forms are the only solutions to $P(a \cdot k) = b \cdot P(k)$ [198]. Power law degree distributions are present in various kinds of networks in nature and technology [11, 148].

## 3.2 Citation analysis

A workhorse among bibliometric techniques is citation analysis. Most scientific work cites previous research on which it is based or which is considered to be relevant for the subject. These *citations* are collected in the list of *cited references* or the *bibliography* of a publication. Ground-breaking work in the area of citation analysis has been described almost half a century ago by *Price* [58]. Other widely recognized authorities in the field are *Garfield* [86, 87] and *Small* [246, 244, 245]. Other authors have devoted substantial research effort to citation studies as well [36, 94, 153, 165, 186, 219]. *Glänzel* described a statistical model for citation processes in the context of predictions of future citation rates [93].

Individual scientists contribute their findings to the scientific community and in return they can expect to receive various forms of recognition from their peers, for example, in the form of citations [115]. Research on citations has shown highly skewed distributions, with a large majority of publications never cited, while a handful receive exceedingly large numbers of citations [5]. When a paper is highly cited, more people are made aware of it and its visibility increases the chance of getting even more citations [4]. High citation scores result from many researchers' decisions to cite a particular paper. *Price* has shown that in-degree distributions of citation networks follow a power law [59, 60]. Preferential attachment in citation network formation bears strong similarity to the more general phenomenon of cumulative advantage [180], in which those who experience early success capture a larger share of subsequent rewards. This *Matthew effect* can be observed in network configuration [179]. The more connected the nodes are, the more new nodes will be attached to them.

Citation counts are used to gauge the overall impact of research output on the scientific community and are generally used to measure quality and research performance of individuals, research groups or nations [52]. An average citation per paper gives an indication of the aggregate level of influence, whereas highly cited papers reflect more important contributions to the field. Although different reasons for citing prior scientific work are conceivable, in general a citation represents endorsement of the previous work and thus signals quality. *Garfield* studied the work of *Nobel* Prize winners and found that they were among the top 0.1% most cited authors [86]. *Zuckerman* found that publication counts, citation counts, and peer ratings were intercorrelated [281]. However, refutation might just as well be a reason to cite prior work. Ideally, the context surrounding citation anchors in a text should be analyzed by natural language processing techniques for clues about the specific reason why each citation was given.

### 3.2.1 Science Citation Index Expanded

Besides the patent databases *EPO* and *USPTO* used for the study in Section 2.4, we mainly use the *Science Citation Index Expanded* (SCIE). The SCIE is one of the databases of the *Institute for Scientific Information* (*ISI*, Philadelphia, PA, USA), which are widely accepted as basic sources for bibliometric analyses. The *Steunpunt O&O Indicatoren*[1] has access to the underlying data of the complete *Web of Science*, including the SCIE, as well as to the *ISI Proceedings data.* The WoS data are fully available since 1981 and the Proceedings data since 1991.

### 3.2.2 Cited reference characteristics

Bibliographies of publications can be analyzed to characterize the *hardness* of fields and subdisciplines in science and social sciences. The *mean reference age* is the average publication year of references cited in a journal or in a subfield. The *share of serials* in all references is the percentage of references that are given to *serial* literature such as journals or other regularly appearing series, in contrast to books, reports or monographs [237]. These indicators reflect typical differences in communication behavior in the sciences, social sciences and humanities [96].

### 3.2.3 Citation graphs

All citations among a set of scientific articles can be collected in a *citation* or *literature network.* Cliques or communities of related research can be identi-

[1] Steunpunt O&O Indicatoren, Katholieke Universiteit Leuven, Dekenstraat 2, B-3000 Leuven, Belgium.

fied and evolution of different subject areas and emergence of new topics in research or technology can be perceived. Properties of citation networks have been analyzed, among others, by *Newman* [198, 200] and *Redner* [227]. Figure 1.5 on page 6 visualizes a citation network that was built by using bibliographic information from the Web of Science. The literature network was constructed using as seed papers all 138 papers of *Ljung L* [168] known to the WoS, and by extending the network with all cited and all citing publications.

### 3.2.4 Co-citation

In a co-citation network two publications (or authors, cf. [269]) are connected if both are cited by the same third publication. The underlying assumption is that co-citation indicates related subject areas. The symmetric co-citation strength has a value between 0 and 1 and is measured by *Salton*'s cosine similarity (see Figure 3.2). The co-citation strength $CC(x,y)$ between two papers $x$ and $y$ is

$$CC(x,y) = \frac{N_{xy}}{\sqrt{N_x \cdot N_y}}, \tag{3.2}$$

with $N_x$ the total number of citations received by paper $x$, $N_y$ the total number of times paper $y$ has been cited, and $N_{xy}$ the number of publications that have cited both $x$ and $y$ (in other words, the number of bibliographies that contain references to both $x$ and $y$). This formula resembles the measure used in text mining to quantify text-based similarity of a pair of documents (see Section 2.1.2, [236]). It can indeed analogously be used with Boolean input vectors indicating all articles that cite a given article.

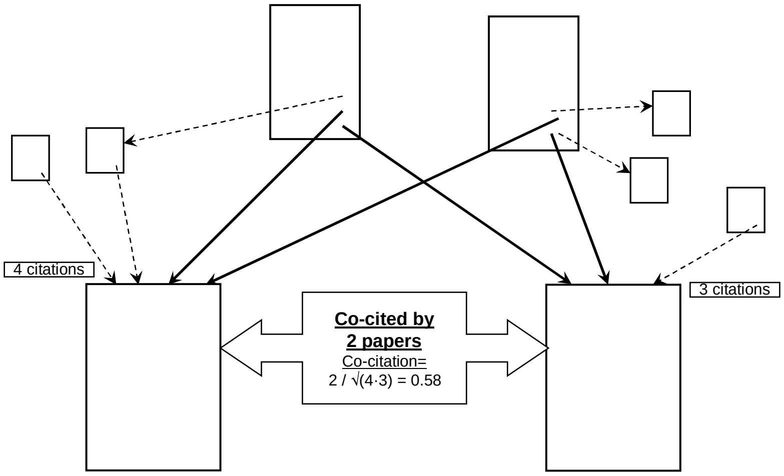


**Figure 3.2:** Co-citation. The lower two papers have received 4 and 3 citations, respectively, and are both cited by the same 2 other papers. Consequently, their co-citation strength is $\frac{2}{\sqrt{4 \cdot 3}} = 0.58$.

Like co-word analysis, co-citation analysis can provide maps of activity based on publications and can be used to monitor dynamics of research themes. *Small* [243] introduced co-citation-based clustering. Progressive visualization of the evolution of co-citation networks has been researched by *Chen* [48].

### 3.2.5 Bibliographic coupling

In a bibliographic coupling (BC) network two nodes (publications) are connected if they have at least one cited reference in common [147] (see Figure 3.3). The strength of coupling, $BC(x, y)$, is also measured by *Salton*'s cosine measure. Hence, the same formula as for co-citation can be used, but with $N_x$ and $N_y$ the number of references in paper $x$ and paper $y$, respectively, and $N_{xy}$ the number of references in common.

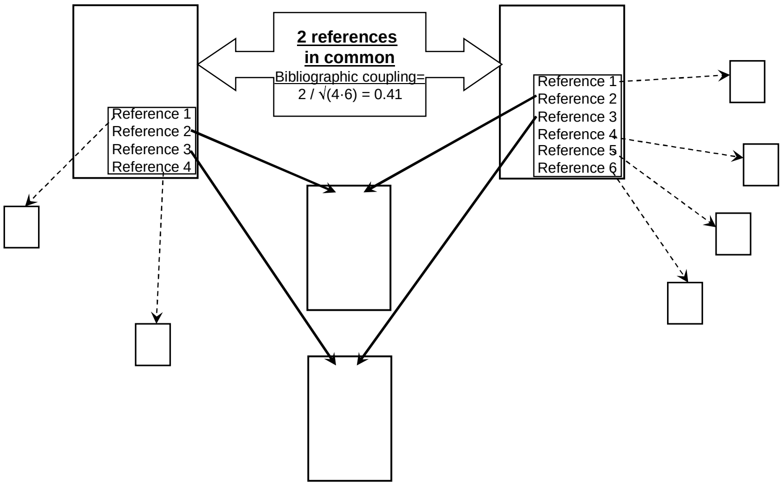


**Figure 3.3:** Bibliographic coupling. The bibliographies of the upper two papers contain 4 and 6 cited references, respectively. Both bibliographies have two cited references in common. Consequently, the bibliographic coupling strength between both papers is $\frac{2}{\sqrt{4 \cdot 6}} = 0.41$.

*Glänzel* and *Czerwon* have used bibliographic coupling to identify core documents that represent 'hot' and research-front topics [99]. *Van Raan* has analyzed network characteristics of a reference-based, bibliographic coupling publication network, in function of the age of references [259].

An advantage of bibliographic coupling over co-citation is that BC does not need time to build up a sufficient amount of citations. All necessary information is available when a paper is published, which is an important advantage for purposes of emerging trend detection (ETD). Recently published papers that are closely related by bibliographic coupling links can provide snapshots of early stages of a specialty's evolution [96].

### 3.2.6 Mean Observed and Mean Expected Citation Rate

The *Mean Observed Citation Rate* (MOCR) is defined as the ratio of citation count to publication count. It reflects the factual citation impact of any unit such as a country, region, institution, or research group. If the underlying paper set is restricted to a single, possibly cross-disciplinary subject, the *subject-standardized Mean Observed Citation Rate* ($MOCR|_f$) can be used as well, which is simply the ratio of the unit's MOCR value and the world standard of the field.

The *Mean Expected Citation Rate* is the average citation rate, measured in any appropriate time frame, of all papers published in the same journal in a specific year. For a set of papers assigned to an institution, country or region in a given field, the indicator is the average of individual expected citation rates over the set [96].

### 3.2.7 Impact Factor

The *ISI* journal Impact Factor was introduced by *Garfield* [85] and is yearly reported in the *Journal Citation Reports*[2] of the *Science Citation Index.* Much care should be taken when using impact factors for research evaluation, especially when the goal is to compare different subject areas or individual performance. For instance, journal coverage in bibliographic databases is a matter of concern, as is the possible bias towards dominating languages. Moreover, authors can attain higher citation rates in larger fields of science. In addition, citation distributions can be very skewed and the chosen citation window is an important factor as well [31, 88].

We use two differently defined impact factors for the evaluation of bioinformatics research. We use the Impact Factor defined by *ISI* as the mean number of citations given in a specific year $X$ to articles published in a journal during the two preceding years $X - 1$ and $X - 2$. Besides, we also count citations in a 3 year citation window.

### 3.2.8 Hirsch-index

For measuring the visibility of an individual scientific author, *Jorge Hirsch* fathered the $h$-index, which is based on the number of citations each of an author's articles receives [124]. Scientists have an $h$-index equal to $h$ if $h$ of their $N_p$ publications have at least $h$ citations each, and the rest $(N_p - h)$ have at most $h$ citations each. The $h$-index is the highest number of papers, published over $n$ years, that have each received at least that number of citations. Thus, an author with an $h$-index of 50 has written at least 50 papers that have each received at least 50 citations [10]. If a scientist has 10 papers, 9 of which are cited 9 times, and the 10th is cited 10 times, then there are $h = 9$ papers having at least $h$

[2] http://scientific.thomson.com/products/jcr/, visited in January, 2007.

citations. Therefore, the scientist's $h$-index equals 9.

Even when researchers are retired, the $h$-index remains useful as a measure of cumulative achievement. Visibility may continue to increase over time, even long after the scientist has stopped publishing [124]. Several advantages and disadvantages of this new measure have been discussed by *Glänzel* [98]. The statistical background and mathematical properties of the $h$-index have among others been analyzed by *Glänzel* [97], *Egghe* and *Rousseau* [77], and *Burrell* [38]. In general, the $h$-index depends on the specific discipline. For example, in biosciences and biotechnologies h-indices tend to be higher than those in physics [124]. *van Raan* has considered research groups rather than individual scientists and he has shown that the $h$-index and several standard bibliometric indicators both correlate with peer judgement [260].

## 3.3 Scientific collaboration

Teamwork is of paramount importance in contemporary science, especially in interdisciplinary research topics. Apart from individual scientists co-authoring publications, interorganizational and international collaboration patterns can be distinguished as well. The absolute number of international papers and their share in the total national publication output serve as basic indicators of international co-authorship relations and scientific collaboration. Such national characteristics have been studied by *Glänzel* [95]. In Section 3.6.4, we map international, interorganizational, and author collaboration for the bioinformatics field.

### 3.3.1 Co-authorship networks

Author collaboration can be represented in a co-authorship network, in which the nodes are individuals that are linked if they have co-authored at least one publication [157].

*Newman* has investigated co-authorship networks to answer questions about collaboration patterns and how they vary between subjects and over time [200]. He has also shown that the diameter of collaboration networks (i.e., the longest shortest path between two vertices, or the *geodesic distance*) is small and the *clustering coefficient* high, indicative for power law degree distributions [197, 198].

*Wagner* and *Leydesdorff* have analyzed the growth of international collaboration in science and tested the hypothesis that international collaboration is a self-organizing network with preferential attachment [264]. *Börner*, *Maru* and *Goldstone* have reviewed models for the structure and dynamics of scientific evolution and have introduced a model for the simultaneous evolution of author and paper networks [33]. *Morris* and *Goldstein* have introduced a qualitative team-based model of research in a specialty and a quantitative growth model based hereupon [193].

### 3.3.2 Interorganizational collaboration

Although Material Transfer Agreements (MTAs) may be useful to exchange research materials between laboratories, academics and policymakers have suggested that the trend towards their standardization might impede the progress of science by constraining research collaboration patterns (see Section 2.4; [230]). The goal of research in progress is to detect discontinuity of interorganizational collaboration in biotechnology that can be attributed to these agreements. The sampled organizations and their collaborations are described with the help of graph theory using technology transfer indicators. *Gay* and *Dousset* have also studied large-scale topology and dynamics of collaboration networks in a major segment of biotechnology industry [91]. They have found accordance with the *fitter-get-richer* hypothesis proposed by *Bianconi* and *Barabasi* [27].

For illustrative purposes, Figure 3.4 presents the interorganizational collaboration network in the Belgian biotechnology sector. We utilized *Pajek* [15], a package for the analysis and visualization of networks. *Pajek* employs two powerful minimum energy or spring-embedded network drawing algorithms to represent network data in two dimensions. These algorithms simulate the network of collaborations as a system of interacting particles, in which organizational nodes repel one another unless network ties act as springs to draw connected nodes closer together. Spring-embedded algorithms iteratively locate a representation of the network that minimizes the overall energy of the system, by reducing the distance between connected nodes and maximizing the distance between unconnected nodes.

The rate at which new organizations appear in the network is partly determined by the success existing nodes have in making progress on a technological frontier. Many of the network participants are multivocal, i.e., they are capable of performing multiple activities with a variety of constituents [39]. But multivocality is not distributed evenly. Those organizations that are more centrally located in industry have access to more sophisticated and diverse collaborators and have developed richer protocols of collaboration [226]. As combinations of collaborators and research agendas unfold, dynamics emerge. Organizational research choices may turn into similar topics, or research trends may cluster and find coherence only in small, densely connected groups. Research agenda choices made early may strongly affect subsequent opportunities, but path dependence might be offset by a constant flow of new arrivals and departures.

## 3.4 Link-based ranking algorithms

Web information retrieval methods based on eigenvectors such as HITS, PageRank, and SALSA [162] , have been surveyed by *Langville* and *Meyer* [161]. Another interesting paper on this subject is by the hand of *Robinson* [229].

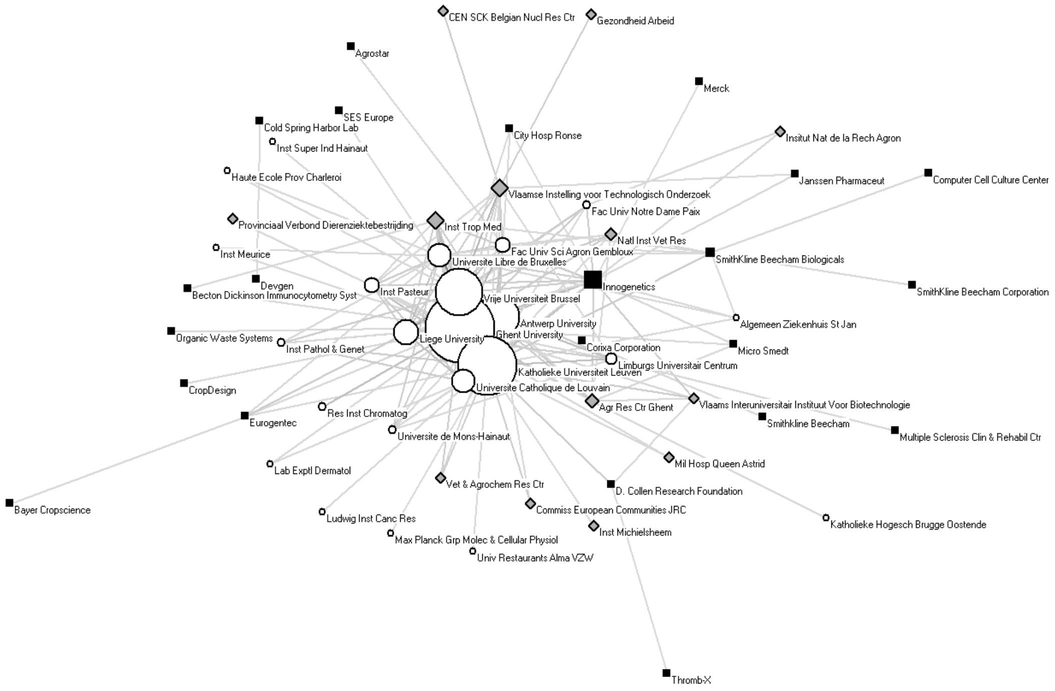

CEN SCK Belgian Nucl Res Ctr
Gezondheid Arbeid
Agrostar
Merck
SES Europe
Cold Spring Harbor Lab
City Hosp Ronse
Inst Super Ind Hainaut
Insitut Nat de la Rech Agron
Haute Ecole Prov Charleroi
Janssen Pharmaceut
Computer Cell Culture Center
Vlaamse Instelling voor Technologisch Onderzoek
Fac Univ Notre Dame Paix
Provinciaal Verbond Dierenziektebestrijding
Inst Trop Med
Natl Inst Vet Res
Fac Univ Sci Agron Gembloux
Inst Meurice
SmithKline Beecham Biologicals
Universite Libre de Bruxelles
Devgen
Becton Dickinson Immunocytometry Syst
Inst Pasteur
Innogenetics
SmithKline Beecham Corporation
Vrije Universiteit Brussel
Antwerp University
Algemeen Ziekenhuis St Jan
Organic Waste Systems
Ghent University
Liege University
Inst Pathol & Genet
Corixa Corporation
Micro Smedt
Limburgs Universitair Centrum
Katholieke Universiteit Leuven
CropDesign
Universite Catholique de Louvain
Vlaams Interuniversitair Instituut Voor Biotechnologie
Agr Res Ctr Ghent
Res Inst Chromatog
Eurogentec
Smithkline Beecham
Multiple Sclerosis Clin & Rehabil Ctr
Universite de Mons-Hainaut
Mil Hosp Queen Astrid
Lab Exptl Dermatol
Vet & Agrochem Res Ctr
D. Collen Research Foundation
Bayer Cropscience
Commiss European Communities JRC
Ludwig Inst Canc Res
Katholieke Hogesch Brugge Oostende
Inst Michielsheem
Max Planck Grp Molec & Cellular Physiol
Univ Restaurants Alma VZW
Thromb-X

**Figure 3.4:** Network of collaborations in Belgian biotechnology between 1992 and 2000. Node size represents number of publications, i.e., the larger the node, the more productive. Link length represents number of collaborations: the closer, the more collaborative. Node shape represents the type of sector: rounded nodes stand for academia, squared nodes symbolize industry, and rhomboidal nodes denote governmental research institutions. A heavily interlinked core is visible with mainly big academic institutions. More peripheral institutions have contributed less to the biotechnology literature and have less links with others.

### 3.4.1 HITS

The original goal of the HITS algorithm, introduced by *Kleinberg* in 1997, was to find the most '*authoritative*' and the best '*hub*' Web pages among an extended result set retrieved from a search engine, including all referring Web sites and all Web sites referred to [149]. These pages together with hyperlinks among them form a directed graph. A hyperlink (directed edge) in the graph is considered to represent recommendation, just as citations in a literature network.

A Web page is considered an *authority* or very relevant for its topic if it is referred to by a lot of other Web pages that are of high quality as well. When a page is a good *hub* it means that it links to many relevant Web sites. Each node is annotated with an authority score and a hub score, which are iteratively updated corresponding to a mutual reinforcement principle between them. This principle states that *a Web page is a good authority if it is pointed to by many good hubs and a Web page is a good hub if it points to many good authorities.* In each iteration the authority scores are replaced by the sum of the hub scores of all referring pages, and hub scores are replaced by the sum of authority scores of all pages referred to. Each iteration is concluded by normalization of hub and authority scores. Mostly, 20 iterations are sufficient for this iterative procedure to converge to stable hub and authority scores for each node. In practice, the same results are obtained by techniques from linear algebra. Let the *adjacency* or *connectivity matrix* $A$ of the graph contain binary values indicating hyperlinks between Web pages. The principal eigenvectors of $A^T \cdot A$ and $A \cdot A^T$ then contain the same authority and hub scores, respectively.

A lot of research has been conducted after this initial introduction of the HITS algorithm. An interesting generalization of HITS was given by *Blondel* and *Van Dooren* to measure similarity of nodes in directed graphs [29]. They have also applied the method to detect synonyms in dictionaries. *Ding et al.* found a relationship between co-citation and authority, and between hubs and bibliographic coupling, and consequently a high correlation between authority scores and in-degree, and between hub scores and out-degree [67]. Unfortunately, HITS has inherent problems such as *topic drift* or simplicity of adversarial information retrieval [25, 161].

The HITS algorithm can also be applied to other directed graphs such as social or citation networks. In the context of literature and patent networks we have adopted HITS to determine representative papers of clusters. An authority might be an important or seminal publication at the origin of a discipline, while a good hub often represents an important survey.

In the context of text mining, we have used a modified version of HITS and have assessed the performance of its mutual reinforcement principle to detect terms and sentences with high saliency scores in documents [138]. Similar goals have been pursued by, among others, *Zha* [275], *Erkan* and *Radev* [80], and *Moens*, *Uyttendaele* and *Dumortier* [188]. In a bipartite graph, all terms of a document were represented by 'term nodes' having outgoing directed edges to 'sentence nodes' for each sentence in which they occured. The mutual reinforcement principle could then be reformulated as: *A salient term is a term that occurs in a lot of salient sentences and a salient sentence is a sentence that contains a lot of salient terms.* After application of the algorithm, the saliency score of a sentence was given by its authority score and the saliency score of a term was equal to its hub score (or vice versa if the directed edges would have been reversed).

In other words, the result of the power method applied to a term-by-document matrix $B$ can be related to the result of the HITS algorithm considering a graph containing a node for each document and a node for each term, and with directed links from each term to every document in which it occurs. Given that HITS only returns authorities and hubs from the largest connected component in a graph, we can conclude that the converged scores will give the most important terms and documents about the most important subject in the text collection. Interestingly, those hub scores for terms are the same as the scores on the dominant left eigenvector resulting from Latent Semantic Indexing of the term-by-document matrix (LSI, see Section 2.2.3). HITS determines hub and authority scores for nodes in a graph by the dominant eigenvectors of $A \cdot A^T$ and $A^T \cdot A$, respectively, with $A$ the adjacency matrix of the graph. LSI uses the left and right singular vectors of a term-by-document matrix $B$, which correspond to the eigenvectors of $B \cdot B^T$ and $B^T \cdot B$. This connection between HITS and LSI was observed by *Ng*, *Zheng* and *Jordan* [203].

### 3.4.2 PageRank

At about the same time that *Kleinberg* introduced the HITS algorithm, *Brin* and *Page* introduced the PageRank algorithm, which is used in the popular search engine *Google* to measure the relative importance of Web pages. PageRank provides an off-line calculated global ranking for every Web page based on the graph of the World Wide Web, while neglecting all textual content [212, 37]. The PageRank of a Web page can be understood as the probability that the page will be visited by a *random surfer* that randomly and with equal probability follows hyperlinks and once in a while 'teleports' to a random page anywhere on the Web. The PageRank is the stationary distribution of a Markov chain representing such an infinite random walk and is computed as the dominant eigenvector of the probability transition matrix. We apply PageRank to citation graphs as a way to characterize clusters of bioinformatics publications by representative papers.

### 3.4.3 Stability

An important issue is stability. Because HITS only considers the largest eigenvectors of $A \cdot A^T$ and $A^T \cdot A$, the results can be very unstable under small perturbations. Hence, if a few nodes would be added to the network—in case of a citation network this would mean a few more publications—very different results might be obtained. *Ng*, *Zheng* and *Jordan* have investigated this instability and found that stability is determined by the 'eigengap', i.e., the difference between the largest and second largest eigenvalues [203]. A large eigengap gives quite stable results, while a small eigengap might even cause the first and second eigenvector to swap places. The PageRank algorithm seemed much less sensitive to small perturbations. *Ng*, *Zheng* and *Jordan* have also proposed two other, more stable variants of HITS, namely *Randomized HITS*, which merges the hubs-and-authorities notion from HITS with a stabilizing 'reset' mechanism from PageRank, and *Subspace HITS*, which provides a principled way of combining multiple eigenvectors from HITS to yield aggregate authority scores [204]. *Cohn* and *Chang* have also described PHITS, a method to probabilistically identify authoritative documents [50]. In PHITS, a probabilistic model replaces the eigenvector analysis of HITS.

## 3.5 Graph partitioning and community structure detection

*Graph partitioning* and the *detection of community structure*, including the mapping of dynamic changes in intellectual communities, are important applications of network analysis. Although graph partitioning and community structure detection are related concepts in the sense that both refer to the clustering of networks, the underlying goals and means of these lines of research differ [202]. Graph partitioning has applications mainly in parallel computing and integrated circuit design, for which the number and size of groups are usually known, whereas the detection of communities is a data analysis technique to detect an unknown number (or the absence) of groups of densely connected vertices with less linkage between those groups. Examples are communities on the WWW of users active in the same area of research or having the same interests, communities of authoritative Web pages or publications linked together by hub pages, etc.

*Newman* has described an effective method to detect community structure by optimization of *modularity* [199, 202, 201]. Modularity is a quality function that is usually maximized over possible divisions of a network. Up to a multiplicative constant, modularity measures the number of intra-cluster edges minus the expected number in an equivalent network with the same community divisions but with edges placed at random. Intuitively, in a good set of communities there are more edges within (and fewer edges between) communities than could be expected from random wiring. The expected number of edges between two nodes is based on their respective degrees and on the total number of edges in the network. *Newman* has shown that modularity can be reformulated in terms of the leading eigenvector of the *modularity matrix* [202]. To subdivide a network into two clusters, vertices are grouped according to the signs of the values in the eigenvector corresponding to the most positive eigenvalue. The absolute values measure how firmly each vertex belongs to the cluster it is assigned to. Modularity is also defined for networks that have multiple edges between two nodes, in which case those edges are replaced by one weighted edge indicating the number of edges it represents.

Some existing clustering algorithms discover groups in networks by explicitly working with the adjacency matrix of the graph. Examples are spectral clustering algorithms that consider the eigenvalue spectrum of the adjacency matrix, as introduced by *Donath* and *Hoffman* [70], and the Markov CLuster algorithm (MCL) of *van Dongen* [258]. MCL is based on random walks in a graph and simulates (alternating) flow expansion and contraction by algebraic operations on stochastic matrices. As a result, the flow between dense, sparsely connected regions will evaporate. MCL converges relatively fast and does not need the number of clusters as an input parameter, but granularity of the clustering outcome depends on other parameters. Other examples of link-based clustering algorithms are described by *Hopcroft* [127, 128] and *Pivovarov* (EqRank) [222].

The HITS algorithm has also been used by *Gibson, Kleinberg* and *Raghavan* to detect multiple Web communities by observing multiple eigenvectors, the dominant eigenvector defining the dominant community [92]. Other authors have addressed the detection of communities and their evolution as well, among which, *Kumar et al.* [159], *Flake*, *Lawrence* and *Giles* [83], *Hopcroft et al.* [127, 128], *Toyoda* and *Kitsuregawa* [255], and *Newman* [201, 202].

Clustering algorithms specifically developed for graphs belong to another paradigm than clustering algorithms that operate in high-dimensional vector spaces, like those that we use for clustering textual data. However, conversions between both worlds are possible. A collection of documents can indeed be modeled as a graph, and a graph can be considered as a set of high-dimensional vectors. We have experimented with MCL, for example, to cluster the citation network in Figure 1.5 on page 6 (each cluster has a different color), but we mainly work with the vector space representation of citation graphs when combining textual and citation information in Chapter 4. The reason for this choice is the equivalence of the citation-based similarity measures *co-citation* and *bibliographic coupling* with cosine-based text vector similarities. *Modha* and *Spangler* have taken the same stance when developing their toric $k$-means algorithm [185]. Moreover, we also incorporate other bibliometric information that is not directly representable in a network. For instance, we investigate the integration of textual content with a combination of *mean reference age* and *share of serials* [136].

## 3.6 Bibliometric analysis of bioinformatics

### 3.6.1 Introduction

A subject-delineation strategy has been developed for retrieval of *core* literature in bioinformatics from the Web of Science and MEDLINE[3] databases [101, 102]. It is a combination of textual components and bibliometric, citation-based techniques, and will therefore be discussed in detail in Chapter 4 (Section 4.6). At this point, it suffices to note that the retrieval strategy resulted in 7401 publications. In this section, we analyze the bioinformatics documents from the bibliometric point of view. In particular, we investigate journal coverage, journal-to-journal citations, growth dynamics and impact of the field. Next, we evaluate publication activity and citation impact of the most active countries, and, finally, international, interorganizational, and author collaboration are mapped.[4]

Bioinformatics is an interdisciplinary field that emerged from the increasing use of computer science and information technology for solving problems in biomedicine, mostly at the molecular level. *Ouzounis* and *Valencia* have provided a review of early stages of the long history of bioinformatics [211]. In recent studies by *Patra* and *Mishra* [216] and *Perez-Iratxeta et al.* [218], evolution and trends in bioinformatics research have been studied. The field has been characterized as an emerging, dynamically evolving discipline with astonishing growth dynamics. The studies were based on the MEDLINE database and partially on NIH-funded project grants. In both cases, bioinformatics was analyzed in a broad biomedical context. *Perez-Iratxeta et al.* quantified the growth of three major informatics topics (*computational methods*, *databases* and *internet*) in research over the past three decades, by calculating the percentage of publications containing various related keywords through time [218]. Recent trends in the use of bioinformatics topics were also contrasted with a more general rise in the use of computers.

In the present study, all bibliometric results are based on raw bibliographic data extracted from the 14-year annual volumes (1991–2004) of the Web of Science Edition of the *Science Citation Index Expanded*$^{TM}$ (SCIE) of *Thomson Scientific* (Philadelphia, PA, USA). Publication data have been matched with MEDLINE. Only papers recorded as *article*, *note*, or *review* in the SCIE were taken into consideration. Papers recorded as *Letter to the Editor* were excluded since this document type tends to cause biases in the application of bibliographic coupling and co-citation analyses [99].

[3] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi

[4] The results presented here are accepted for publication in *Proceedings of the 11th International Conference of the International Society for Scientometrics and Informetrics (ISSI)* [102].

### 3.6.2 Journal coverage of bioinformatics literature in the SCIE

In total 7401 *articles*, *notes*, or *reviews* in bioinformatics were retrieved for the period 1981–2004.

The bibliometric study by *Patra* and *Mishra* was based on MeSH terms and adopted a rather liberal domain delineation strategy that was tailored towards maximal recall. They selected 14 563 journal articles [216], that is, about twice as many as we have found. The main reason is the broad interpretation of bioinformatics resulting from the less restricted search strategy. The other reason is the broader coverage of the underlying database. We aimed at a very strict interpretation of the field, at retrieving the very core of bioinformatics with practically no noise. This was essential for having a solid groundwork for cluster analysis of the retrieved literature. Nonetheless, their ranking of important journals by and large coincides with the list that we have found, and, surprisingly, the number of articles (5387) in the top 20 journals is almost exactly the same as in the present study (5390). Core journals can, of course, be found at the top of the list (see Table 3.1).

**Table 3.1:** The 25 most frequently used journals for publishing bioinformatics literature.

| Rank | Journal | Frequency |
|---:|---|---:|
| 1. | BIOINFORMATICS | 1900 |
| 2. | COMPUTER APPLICATIONS IN THE BIOSCIENCES | 724 |
| 3. | NUCLEIC ACIDS RESEARCH | 594 |
| 4. | JOURNAL OF COMPUTATIONAL BIOLOGY | 397 |
| 5. | JOURNAL OF MOLECULAR BIOLOGY | 241 |
| 6. | BMC BIOINFORMATICS | 239 |
| 7. | GENOME RESEARCH | 203 |
| 8. | PNAS USA | 189 |
| 9. | NATURE | 116 |
| 10. | MOLECULAR BIOLOGY AND EVOLUTION | 107 |
| 11. | SCIENCE | 107 |
| 12. | PROTEIN SCIENCE | 92 |
| 13. | PROTEINS-STRUCTURE FUNCTION AND GENETICS | 88 |
| 14. | PROTEIN ENGINEERING | 84 |
| 15. | MOLECULAR PHYLOGENETICS AND EVOLUTION | 63 |
| 16. | NATURE GENETICS | 56 |
| 17. | JOURNAL OF MOLECULAR EVOLUTION | 54 |
| 18. | CURRENT OPINION IN STRUCTURAL BIOLOGY | 51 |
| 19. | GENOMICS | 46 |
| 20. | PROTEINS-STRUCTURE FUNCTION AND BIOINFORMATICS | 44 |
| 21. | FEBS LETTERS | 37 |
| 22. | GENOME BIOLOGY | 37 |
| 23. | TRENDS IN BIOCHEMICAL SCIENCES | 33 |
| 24. | GENETICS | 30 |
| 25. | TRENDS IN GENETICS | 30 |

Figure 3.5 presents all journals with more than 20 papers in our set. The size of a node represents the square root of number of publications. Arcs are weighted to represent the number of journal-to-journal citations. Arcs corresponding to less than 20 citations were removed. The *Kamada-Kawai* algorithm that was used for layout consequently put journals in close vicinity if their respective sets of papers relatively often cited each other. *Pajek* was used for visualizing the network [15].

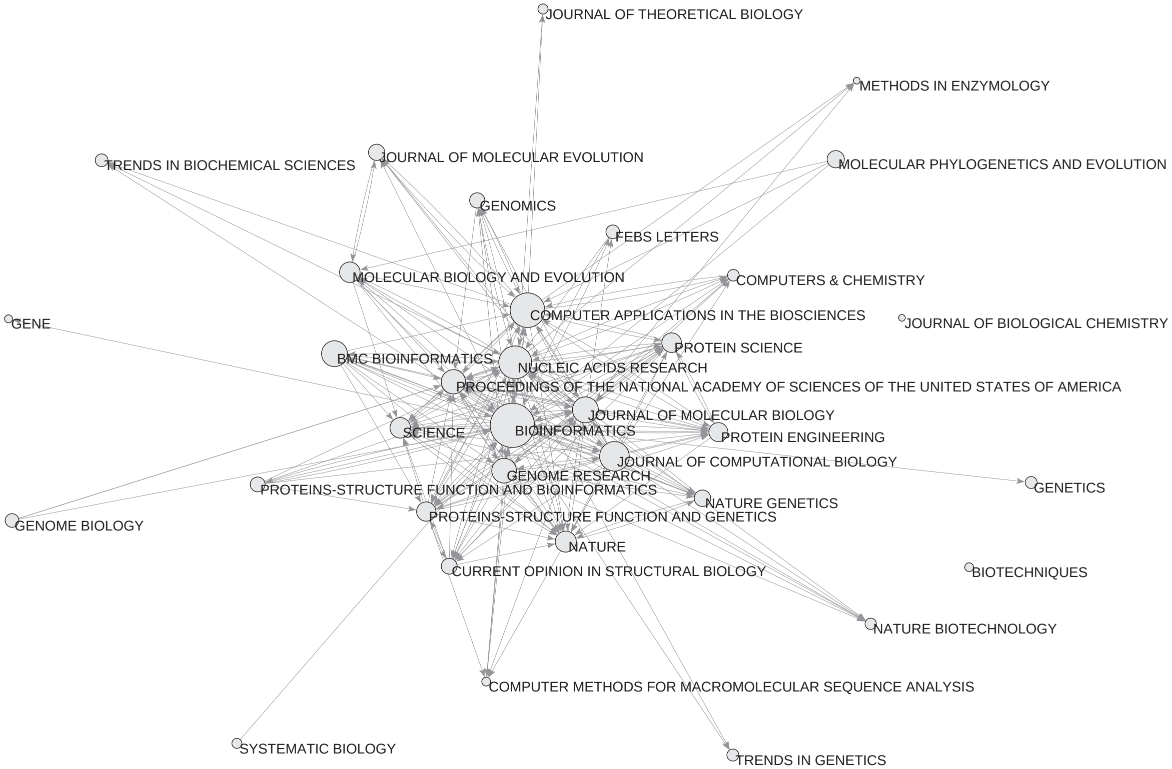

JOURNAL OF THEORETICAL BIOLOGY
METHODS IN ENZYMOLOGY
TRENDS IN BIOCHEMICAL SCIENCES
JOURNAL OF MOLECULAR EVOLUTION
MOLECULAR PHYLOGENETICS AND EVOLUTION
GENOMICS
FEBS LETTERS
MOLECULAR BIOLOGY AND EVOLUTION
COMPUTERS & CHEMISTRY
GENE
COMPUTER APPLICATIONS IN THE BIOSCIENCES
JOURNAL OF BIOLOGICAL CHEMISTRY
PROTEIN SCIENCE
BMC BIOINFORMATICS
NUCLEIC ACIDS RESEARCH
PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA
JOURNAL OF MOLECULAR BIOLOGY
BIOINFORMATICS
SCIENCE
PROTEIN ENGINEERING
JOURNAL OF COMPUTATIONAL BIOLOGY
GENOME RESEARCH
GENETICS
PROTEINS-STRUCTURE FUNCTION AND BIOINFORMATICS
NATURE GENETICS
GENOME BIOLOGY
PROTEINS-STRUCTURE FUNCTION AND GENETICS
NATURE
CURRENT OPINION IN STRUCTURAL BIOLOGY
BIOTECHNIQUES
NATURE BIOTECHNOLOGY
COMPUTER METHODS FOR MACROMOLECULAR SEQUENCE ANALYSIS
SYSTEMATIC BIOLOGY
TRENDS IN GENETICS

**Figure 3.5:** Journal-to-journal citation network showing journals with more than 20 papers in the set. Node size represents square root of number of publications; arc weights represent number of cross-journal citations. Only arcs corresponding to more than 20 citations were retained. By using the *Kamada-Kawai* layout algorithm, journals were put close to each other when their respective sets of papers often cited each other.

Journals in computational and molecular biology as well as the important multidisciplinary journals *PNAS USA*, *Nature* and *Science* are the most important publication channels for bioinformatics research. Although merely 5 core journals were included in the delineation strategy, the subsequent steps of the bibliometric retrieval provided the inclusion of a lot more journals. The huge number of journals in which the papers were scattered according to the paper by *Patra* and *Mishra* could thus be confirmed.

The initial set of 5 core journals was considered an *unconditional criterion* in the delineation strategy, meaning that every paper published in these journals was admitted in the set. Consequently, these journals are most represented in our bioinformatics set. Although the adopted delineation strategy was carefully formulated in order to compose a set of bioinformatics papers as representative as possible, we are aware that the actual choices made might introduce a bias. The analyses to be presented in the next subsections are based on the data set at hand and are not intended to make any quality judgement about journals, authors or papers. Likewise, journals that are not or hardly present in Figure 3.5 might as well publish important bioinformatics papers, but be possibly neglected by the strategy put forward.

### 3.6.3 Evolution of publication output and citation impact

Figure 3.6 gives a picture of the increase in yearly number of bioinformatics publications. The growth of publications lies in between the linear model in the first half of the period and the exponential model for the second half (similarly as observed in nanoscience and -technology, cf. [103]). Literature growth clearly characterizes the field as a young, emerging, and dynamically evolving discipline.

The dynamic growth of literature in bioinformatics is outrun by an even more powerful increase of citations. The patterns are shown in Figure 3.7. For this figure, citations were counted in a three-year window: in the year of publication and the two subsequent years. For instance, if papers published in 1999 were considered, all citations received in the period 1999–2001 have been counted. Because of the use of 3 year citation windows, citations could be counted for papers published up to 2003 (citations received in 2003–2005).

The evolution of the field's Mean Observed Citation Rate (see Section 3.2.6) is presented in Figure 3.8. The strong linear increase of citation impact in the 1990s is followed by a sharp decline in the new millennium. The reason for this phenomenon is not clear. However, a similar decline of citation impact has been observed for nanoscience and -technology [103]. It seems that emerging fields are characterized first by a growth of citation impact exceeding that of the publication output, then by stagnation, and later on by the decrease of impact while the powerful increase of publication activity continues.

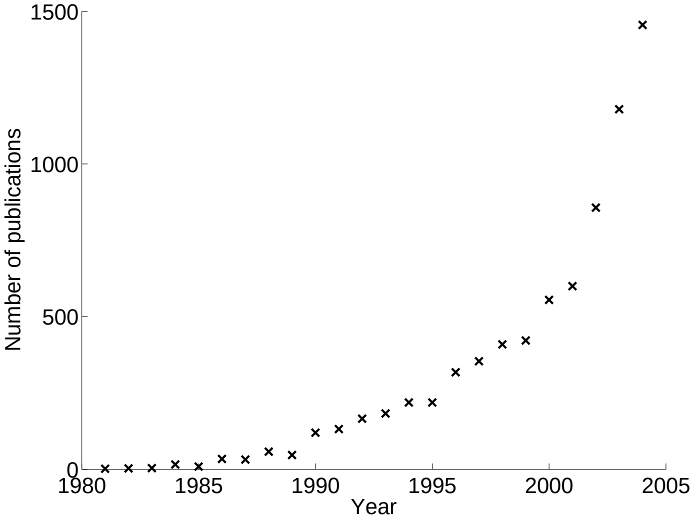


**Figure 3.6:** Evolution of publication output in bioinformatics.

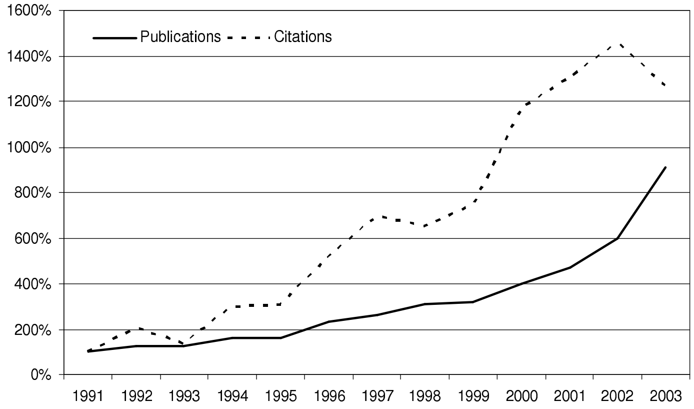


**Figure 3.7:** Annual change of citations compared with that of publications in bioinformatics for 1991–2003 (1991 = 100%).

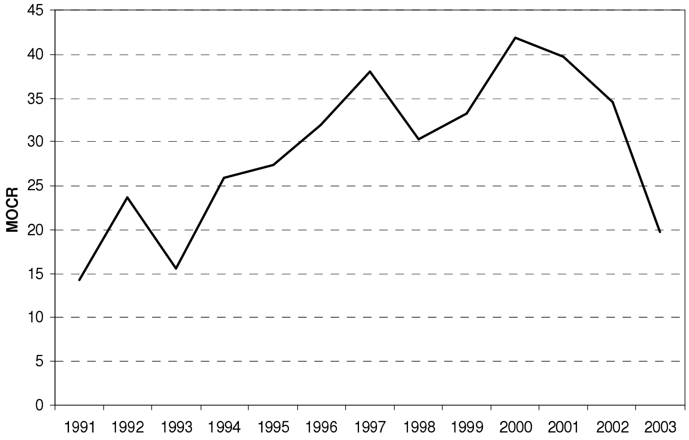


**Figure 3.8:** Evolution of Mean Observed Citation Rate in the period 1991–2003.

### Publication output and citation impact of the 30 most active countries

For the analysis of national publication activity and citation impact, the 30 most active countries in the period 1991–2004 have been selected. Countries with less than 30 papers in the 14-year period were excluded for reasons of statistical reliability.

The publication output of the 30 most active countries in bioinformatics and their share in the world total in this field are presented in Table 3.2. In order to provide information about the evolution of national publication activity in the field, the period 1991–2004 has been split into two sub-periods, particularly, 1991–1997 and 1998–2004. National data in Table 3.2 are ranked in descending order of publication output in the whole 14-year period. If we compare the list with similar lists on national publication output in all fields combined, we can conclude that those countries that are most active in scientific research in all fields combined have top activity in bioinformatics research, too.

However, the three 'leading' countries, USA, UK, and Germany rank distinctly higher in bioinformatics than in all fields combined [106]. The USA have contributed to half of the total publication output of the 30 most active countries. Altogether with UK and Germany, they contributed to three quarters of the total. Although publication counts for the first period are small, we can observe a powerful growth of publication activity in China and other emerging scientific nations such as South Korea, Taiwan and Brazil [100]. National representation also confirms the findings by *Patra* and *Mishra* [216]. The citation impact, in the sub-periods 1991–1997 and 1998–2003, of the 30 most active countries with at least 25 papers in the period 1991–2003 is shown in Table 3.3. We used the share of author self-citations $f_S$ and the subject-standardized Mean Observed Citation Rate ($MOCR|_f$). Both were obtained by scripts made available by the *Steunpunt O&O Indicatoren.*[5]

[5] Steunpunt O&O Indicatoren, Katholieke Universiteit Leuven, Dekenstraat 2, B-3000 Leu-

**Table 3.2:** Publication output of the 30 most active countries in sub-periods 1991–1997 and 1998–2004.

| | 1991–1997 | | 1998–2004 | | 1991–2004 | |
|---|---|---|---|---|---|---|
| **Country** | **Papers** | **Share** | **Papers** | **Share** | **Papers** | **Share** |
| USA | 721 | 46.8% | 2923 | 52.8% | 3644 | 51.5% |
| GBR | 235 | 15.3% | 767 | 13.9% | 1002 | 14.2% |
| DEU | 189 | 12.3% | 594 | 10.7% | 783 | 11.1% |
| FRA | 121 | 7.9% | 331 | 6.0% | 452 | 6.4% |
| JPN | 74 | 4.8% | 232 | 4.2% | 306 | 4.3% |
| CAN | 49 | 3.2% | 223 | 4.0% | 272 | 3.8% |
| ITA | 60 | 3.9% | 150 | 2.7% | 210 | 3.0% |
| ESP | 39 | 2.5% | 146 | 2.6% | 185 | 2.6% |
| ISR | 33 | 2.1% | 144 | 2.6% | 177 | 2.5% |
| SWE | 14 | 0.9% | 161 | 2.9% | 175 | 2.5% |
| RUS | 56 | 3.6% | 118 | 2.1% | 174 | 2.5% |
| AUS | 21 | 1.4% | 134 | 2.4% | 155 | 2.2% |
| CHE | 47 | 3.1% | 100 | 1.8% | 147 | 2.1% |
| CHN | 7 | 0.5% | 139 | 2.5% | 146 | 2.1% |
| BEL | 24 | 1.6% | 108 | 2.0% | 132 | 1.9% |
| DNK | 12 | 0.8% | 83 | 1.5% | 95 | 1.3% |
| NLD | 18 | 1.2% | 77 | 1.4% | 95 | 1.3% |
| IND | 16 | 1.0% | 72 | 1.3% | 88 | 1.2% |
| SGP | 6 | 0.4% | 73 | 1.3% | 79 | 1.1% |
| POL | 5 | 0.3% | 53 | 1.0% | 58 | 0.8% |
| NOR | 6 | 0.4% | 45 | 0.8% | 51 | 0.7% |
| IRE | 7 | 0.5% | 43 | 0.8% | 50 | 0.7% |
| TWN | 1 | 0.1% | 47 | 0.8% | 48 | 0.7% |
| AUT | 5 | 0.3% | 42 | 0.8% | 47 | 0.7% |
| FIN | 5 | 0.3% | 41 | 0.7% | 46 | 0.7% |
| KOR | 1 | 0.1% | 44 | 0.8% | 45 | 0.6% |
| BRA | 0 | 0.0% | 44 | 0.8% | 44 | 0.6% |
| NZL | 6 | 0.4% | 36 | 0.7% | 42 | 0.6% |
| HUN | 11 | 0.7% | 27 | 0.5% | 38 | 0.5% |
| GRC | 5 | 0.3% | 30 | 0.5% | 35 | 0.5% |
| **WORLD** | **1540** | **100.0%** | **5536** | **100.0%** | **7076** | **100.0%** |

**Table 3.3:** Citation impact and self-citation rate $f_S$ of the 30 most active countries in 1991–2003 in the two sub-periods 1991–1997 and 1998–2003.

| | | 1991–1997 | | | 1998–2003 | | | 1991–2003 | |
|---|---|---|---|---|---|---|---|---|---|
| **Country** | **Papers** | $MOCR\|_f$ | $f_S$ | **Papers** | $MOCR\|_f$ | $f_S$ | **Papers** | $MOCR\|_f$ | $f_S$ |
| USA | 721 | 1.28 | 10.1% | 2162 | 1.37 | 9.1% | 2883 | 1.35 | 9.3% |
| GBR | 235 | 1.17 | 12.0% | 594 | 1.47 | 11.0% | 829 | 1.39 | 11.2% |
| DEU | 189 | 1.24 | 13.8% | 429 | 1.48 | 11.2% | 618 | 1.41 | 11.8% |
| FRA | 121 | 2.09 | 12.5% | 247 | 1.66 | 9.6% | 368 | 1.78 | 10.6% |
| JPN | 74 | 1.01 | 17.3% | 157 | 1.97 | 10.9% | 231 | 1.68 | 12.0% |
| CAN | 49 | 2.96 | 11.1% | 140 | 2.15 | 10.1% | 189 | 2.34 | 10.4% |
| ITA | 60 | 0.90 | 19.4% | 103 | 0.73 | 19.9% | 163 | 0.78 | 19.7% |
| RUS | 56 | 0.16 | 26.7% | 94 | 0.52 | 17.6% | 150 | 0.39 | 18.9% |
| ISR | 33 | 0.40 | 21.5% | 112 | 2.06 | 9.1% | 145 | 1.73 | 9.7% |
| ESP | 39 | 1.20 | 17.5% | 99 | 2.18 | 10.3% | 138 | 1.93 | 11.5% |
| SWE | 14 | - | - | 105 | 1.63 | 9.5% | 119 | 1.85 | 8.8% |
| CHE | 47 | 2.24 | 12.0% | 68 | 3.04 | 8.6% | 115 | 2.69 | 9.7% |
| AUS | 21 | - | - | 90 | 2.49 | 8.9% | 111 | 2.13 | 9.2% |
| BEL | 24 | - | - | 71 | 0.88 | 14.2% | 95 | 1.14 | 17.5% |
| CHN | 7 | - | - | 79 | 1.96 | 8.9% | 86 | 1.90 | 9.2% |
| DNK | 12 | - | - | 61 | 1.64 | 8.0% | 73 | 1.78 | 8.5% |
| NLD | 18 | - | - | 47 | 2.56 | 8.5% | 65 | 2.42 | 10.5% |
| IND | 16 | - | - | 42 | 0.29 | 19.4% | 58 | 0.23 | 20.1% |
| SGP | 6 | - | - | 42 | 0.59 | 22.9% | 48 | 0.54 | 23.2% |
| NOR | 6 | - | - | 35 | 1.65 | 10.4% | 41 | 1.50 | 10.9% |
| POL | 5 | - | - | 34 | 0.52 | 27.3% | 39 | 0.69 | 25.2% |
| IRE | 7 | - | - | 30 | 4.31 | 7.2% | 37 | 4.40 | 9.1% |
| FIN | 5 | - | - | 31 | 0.56 | 13.3% | 36 | 0.55 | 13.9% |
| HUN | 11 | - | - | 22 | - | - | 33 | 0.42 | 16.8% |
| NZL | 6 | - | - | 24 | - | - | 30 | 0.83 | 13.3% |
| AUT | 5 | - | - | 24 | - | - | 29 | 0.60 | 17.6% |
| BRA | 0 | - | - | 27 | 0.26 | 32.9% | 27 | 0.26 | 32.9% |
| GRC | 5 | - | - | 21 | - | - | 26 | 2.33 | 10.6% |
| TWN | 1 | - | - | 25 | 0.21 | 26.7% | 26 | 0.21 | 28.0% |
| KOR | 1 | - | - | 20 | - | - | 21 | - | - |
| **WORLD** | **1540** | **1.00** | **11.3%** | **3967** | **1.00** | **10.2%** | **5507** | **1.00** | **10.5%** |

Ireland has the highest $MOCR|_f$, but this based on only 37 papers in the complete period. The high relative citation impact of Canada, Switzerland, Australia and the Netherlands (more than twice the world standard) is worth mentioning. This is contrasted by the relatively low impact of Russia and Italy in all sub-periods, although their publication activity is quite high. In general, the share of author self-citations $f_S$ of about 10% is low in this field; national deviation from this standard follows the patterns observed from other science fields [103]. The overall high impact is partially a consequence of the citation-based component of the retrieval strategy. A study of bibliographic coupling by *Glänzel* and *Czerwon* has shown that retrieval based on strong coupling links results in higher-than-average citation impact [99]. Citation aided tools in information retrieval and data mining necessarily imply a certain bias concerning visibility of literature. The better depiction of the structure of the information space is to the detriment of loosely linked and less visible documents.

## 3.6.4 Global collaboration networks

### Mapping bilateral co-authorship links

In order to measure the strength of bilateral collaboration, an appropriate similarity measure based on country pairs has been used. Multinational collaboration is therefore split up into multiple bilateral relations. Figure 3.9 visualizes the international collaboration network. The weight $w_{a,b}$ on the edge linking two countries $a$ and $b$ is normalized by *Salton*'s cosine measure [236], i.e.,

$$w_{a,b} = \frac{N_{a,b}}{\sqrt{N_a \cdot N_b}}, a \neq b, \qquad (3.3)$$

with $N_{a,b}$ the number of joint publications for which at least countries $a$ and $b$ collaborated, and $N_a$ and $N_b$ the total publication output of countries $a$ and $b$, respectively. In contrast to self-citations in bibliographic coupling or co-citation analyses, we do not consider 'self-collaboration' and therefore define $w_{a,a} = 0$.

Figure 3.9 shows with which other countries a specific country mostly collaborated. For example, Hungary mostly collaborated with New Zealand, the USA, India, and Austria. In general, geographical proximity and common language are important factors, as strong collaborations very often occur between neighboring countries, such as, for example, Sweden and Denmark, or Canada and the USA. Countries with a varied set of cooperating nations can also be distinguished from countries that merely have a few. Since the figures are based on all bioinformatics papers retrieved for 1980–2004, countries such as Czechoslovakia, GDR and FRG still appear in the diagram; however, because of the dynamic growth of the field, their role in the complete set is marginal. In Figure 1.8 on page 9, the same collaboration network was shown with *Kamada-Kawai* layout, but only for countries with more than 20 publications in the set. the *big* countries, USA, UK, Germany, France, and Japan can be found in the center of the

ven, Belgium.

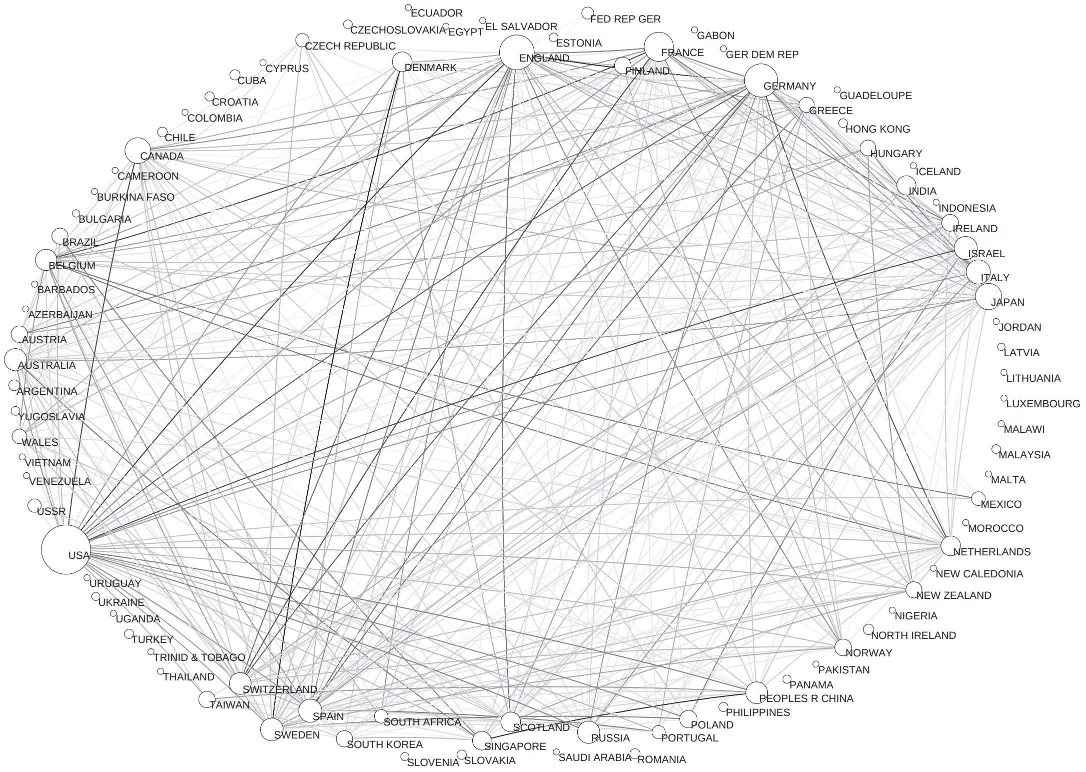


**Figure 3.9:** International collaboration network. Node size represents square root of number of publications and the gray level of an edge represents number of collaborative publications (mutual co-authorship). This figure shows with which other countries a specific country mostly collaborated. For example, Hungary mostly collaborated with New Zealand, the USA, India, and Austria. In general, neighboring countries often collaborate intensively. For example, Sweden and Denmark, or Canada and the USA.

diagram. The appearance of the emerging nations such as China, Singapore, Korea, and Brazil as nodes in the collaboration network is worth mentioning.

Figure 3.11 visualizes the co-authorship network containing all authors that have a within-set $h$-index of at least 12 (see Section 3.2.8). The bioinformatics community proves to be quite homogeneous with a lot of important authors that intensely collaborate. The average distance between two authors in the author collaboration network is 4.66.

Finally, Figure 3.10 illustrates Lotka's law for the bioinformatics set, with power law exponent $\gamma = 2.49$. Lotka's law of scientific productivity asserts that the ditribution of the number of papers written by individual scientists follows a power law. A very large number of publications is produced by only a few authors, whereas most authors only publish once. *The number (of authors) making $n$ contributions is about $1/n$ of those making one; and the proportion of all contributors that makes a single contribution is about 60 per cent* [172].

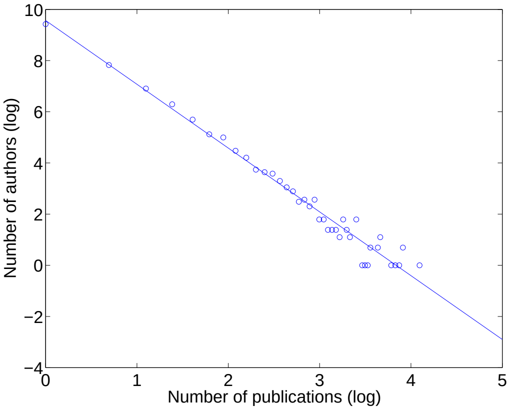


**Figure 3.10:** Lotka's law of scientific productivity [172]. Lotka's law asserts that the ditribution of the number of papers written by individual scientists follows a power law. A very large number of publications is produced by only a few authors, whereas most authors only publish once. The slope of the fitted line is $-2.49$, hence the power law exponent $\gamma = 2.49$.

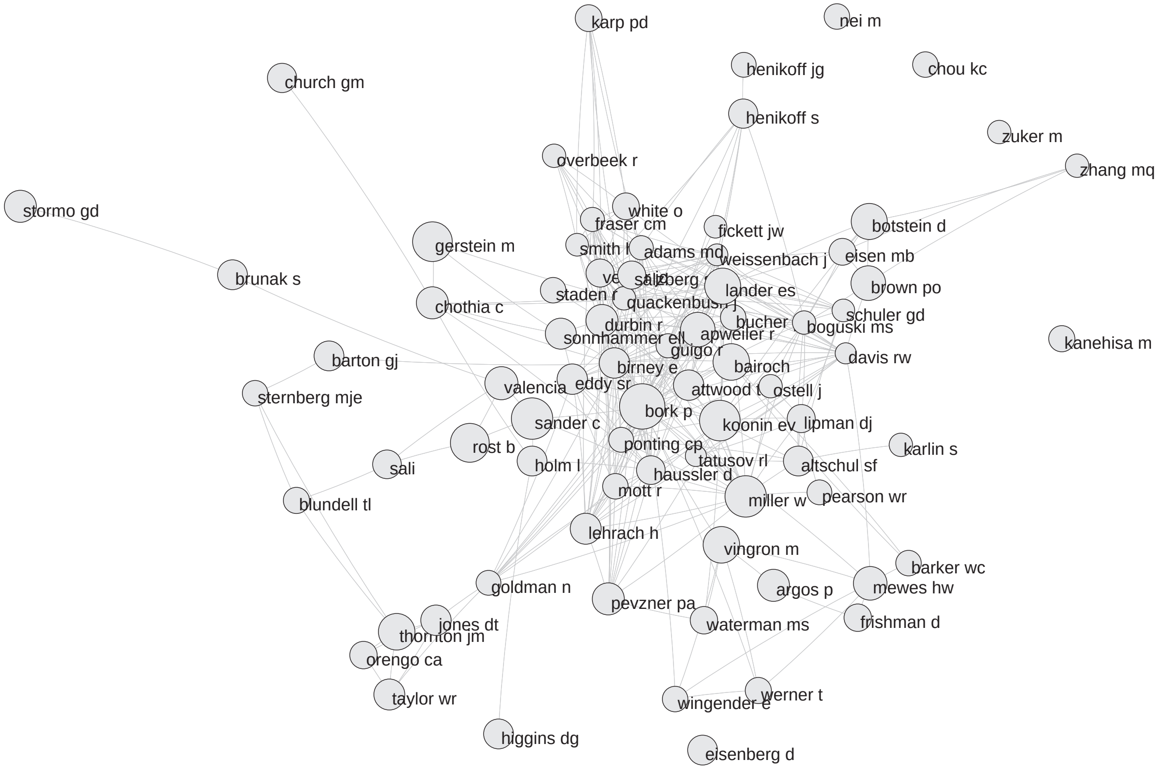


**Figure 3.11:** Author collaboration. Only authors that have a within-set *h*-index larger than 12 are shown. Node size represents number of publications, edge weights represent number of collaborative publications. *Kamada-Kawai* layout with manual adjustment to ensure readability (*Craig Venter*'s name is partly hidden). The bioinformatics community proves to be quite homogeneous with a lot of important authors that intensely collaborate.

### 3.6.5 Discussion

The field of bioinformatics proved a young, emerging field characterized by a powerful, from the late 1990s on almost exponential growth of literature. Beyond several core journals, important periodicals in molecular biology as well as the multidisciplinary journals *Science*, *Nature* and *PNAS USA* proved to be the most important publication channels. Our study has confirmed findings by other recent studies concerning publication patterns. The partially citation-based subject delineation supported the identification of rather visible publications; the citation analysis characterized bioinformatics as a field with very high overall citation scores. According to our expectations, the extent of international collaboration is in keeping with that of other emerging interdisciplinary fields. The *big* countries form the nodes of the global co-publication network.

## 3.7 Concluding remarks

Contrary to the textual approach of Chapter 2, in this chapter we have focused on a selection of bibliometric and graph analytic techniques, which present a different view on information concerning scientific publications and patents contained in massive bibliographic databases. This chapter contained a brief description of some major topics in network analysis such as the emergence of scaling and self-organization in small-world networks. Algorithms for ranking result sets from Web information retrieval, particularly, HITS and PageRank, have been described and will further be adopted in Chapter 5 to analyze literature networks. A succinct section discussed graph partitioning and detection of community structure and dynamics.

We opted not to use specific graph partitioning algorithms such as spectral clustering when integrating text mining and citation-based techniques. Contrary, we mainly work with the vector space representation of citation graphs. This choice was suggested by the resemblance between the bibliometric measures co-citation and bibliographic coupling, and vector space clustering techniques. In addition, vector space methods have a long tradition in bibliometrics. We were also influenced by the toric $k$-means algorithm of *Modha* and *Spangler*, which also utilizes a vector space stance [185]. Moreover, clustering in vector spaces will prove a valuable approach to integrate citation structures or textual information with other bibliometric indicators which do not have a direct analogy with network structure.

The chapter was concluded with a bibliometric analysis of the young, emerging, interdisciplinary field of bioinformatics. Journal coverage, evolution of publication output and citation impact, as well as author and international collaboration networks were described. From the late 1990s on, an almost exponential growth of literature could be observed. Citation analysis characterized bioinformatics as a field with very high overall citation scores. However, the strong linear increase of citation impact in the 1990s was followed by a sharp decline in the new millennium. Countries that are most active in scientific research in general have top activity in bioinformatics, too. USA, UK and Germany rank distinctly higher in bioinformatics than in science in general and have altogether contributed to three quarters of the total publication output of the 30 most active countries. The bioinformatics community is quite homogeneous with a lot of important authors intensely collaborating.

The textual and graph-based approaches provide different perceptions of similarity between documents or groups of documents. We deem it a very interesting research topic to incorporate both viewpoints and we hypothesize that an integrated approach leads to a better comprehension of the structure and dynamic properties of textual corpora. The topic of Chapter 4 is to assess different means of integrating bibliometric and citation-based techniques with text mining. The hybrid methodology that provides the best clustering and classification performance is then demonstrated in Chapter 5 to come up with a hybrid and dynamic clustering of bioinformatics. Each detected cluster is then further profiled by text-based and link-based techniques.

# Chapter 4

# Hybrid analysis combining text mining and bibliometrics

The previous two chapters were devoted to text mining on the one hand, and to bibliometrics and link analysis on the other. Both worlds have proven to provide effective and valuable algorithms for mapping of knowledge, for charting S&T fields, and to monitor scientific processes. Statistical analysis of a textual corpus provides information on included topics, whereas bibliometrics can elucidate other relationships based on various indicators that also convey important clues for mapping purposes.

## 4.1 Introduction

In this chapter we asses the performance of various schemes to integrate text mining and bibliometrics. The long-term goal is an accurate unsupervised clustering of science or technology fields, towards the detection of emerging fields or hot topics.

Sometimes textual information can indicate similarities that are not visible to bibliometric techniques, and vice versa. For example, we encountered two papers in a data set about library and information science with bibliographic coupling similarity equal to 0 (i.e., they have no common references, see Section 3.2.5 for more information), but with more than 95% textual cosine similarity (see Section 2.1.2). Both papers were of *Ding* and *Foo* and were published in *Journal of Information Science* (Appendix B: Ding et al., 2002a; Ding, 2002b). The reason why both papers were not bibliographically coupled is that they mostly cited literature not published in periodicals or serials such as journals. However, both papers were correctly identified as being very similar by the tex-

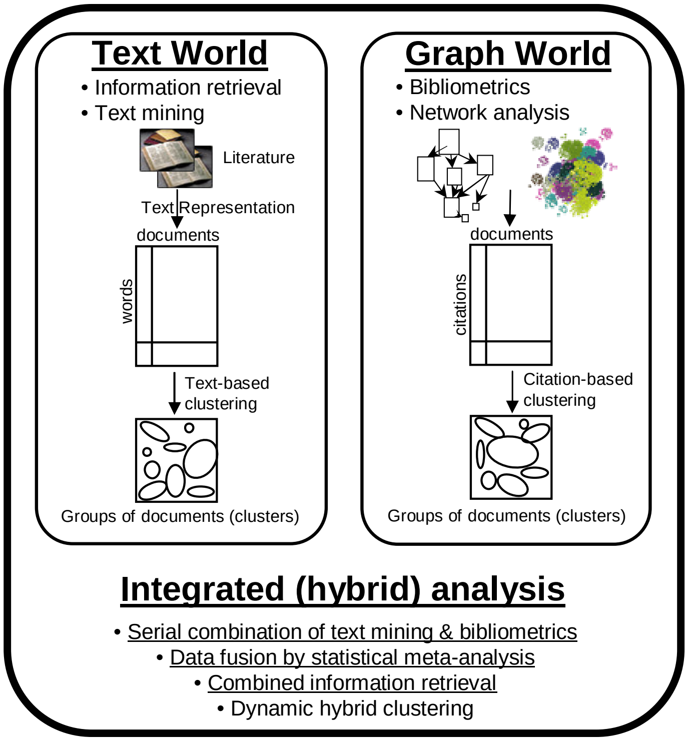

Text World
• Information retrieval
• Text mining
Literature
Text Representation
documents
words
Text-based clustering
Groups of documents (clusters)
Graph World
• Bibliometrics
• Network analysis
documents
citations
Citation-based clustering
Groups of documents (clusters)
Integrated (hybrid) analysis
• Serial combination of text mining & bibliometrics
• Data fusion by statistical meta-analysis
• Combined information retrieval
• Dynamic hybrid clustering

**Figure 4.1:** In this chapter we investigate to what extent text mining and bibliometric methods can supplement each other and whether they can be performed serially to improve individual approaches to science mapping. Next, we devise a methodology for deeply combining text mining and bibliometrics by integrating text-based and bibliometric information before application of a clustering algorithm. We asses the performance of various schemes to integrate textual content and citations and we show that text is more powerful than cited references, but that the best outcome is obtained by integration. Subsequently, we revisit the mapping of library and information science by using hybrid methods. Finally, a hybrid information retrieval strategy consisting of textual and bibliometric components is described and applied to delineate the core literature in bioinformatics.

tual cosine similarity, as they were follow-up papers, namely part I and II of *'Ontology research and development. A review of ontology generation'*. As an aside, there was actually one cited reference common to both papers, but the cited work was published more than 10 years before the papers under investigation, which is a common threshold for bibliographic coupling or co-citation analyses.

On the other hand, based on text alone, true document similarity can be obscured by differences in vocabulary use, or spurious similarities might be introduced as a result of textual pre-processing, or because of polysemous words or words with little semantic value. For instance, documents about music information retrieval might erroneously be linked to patent-related research based on common terms that are used in both contexts, such as *title, record, creative,* and *business*.

### 4.1.1 Related research

The idea of combining bibliometric or citation information with textual content is not new. Bibliometric methods have already been combined with the analysis of indexing terms, subject headings, or keywords extracted from titles and abstracts [43, 208, 279]. Integration has also been pursued to obtain improved performance in information retrieval, bibliometric mapping of science, clustering, and classification.

For **retrieval** purposes, *Bharat* and *Henzinger* augmented the classical HITS algorithm (see Section 3.4.1) with content analysis [25]. The Automatic Resource Compilation algorithm by *Chakrabarti et al.* [46] also extended HITS with analysis of the text surrounding hyperlinks. *Calado et al.* [42] assessed how the use of local link information in the Web compares with the use of global link information, both obtained from the HITS algorithm. They used Bayesian networks to combine link-based and text-based evidence and obtained better retrieval results.

With regard to the **bibliometric mapping of science**, the idea of studying the full text of scientific literature by means of statistics, and combining these tools with bibliometrics, was already present in the work of *Mullins*, *Snizek*, and *Oehler* [196, 247]. *Braam*, *Moed*, and *van Raan* suggested to combine co-citation with word analysis in the context of evaluative bibliometrics in order to improve efficiency of co-citation clustering [34]. The integration of full-text based techniques, above all of text mining into bibliometric standard methodology, has also been advanced by *Kostoff* [156, 155].

*Modha* and *Spangler* [185] introduced the toric k-means algorithm for **clustering** hypertext documents using words, out-links and in-links. The relative importance of these information sources was determined by searching the parameter space for an optimal figure-of-merit. Similarity was calculated as a weighted sum of the inner products between the individual text-based or link-based components. A comparable linear combination of document similarities is described in Section 4.3.1 and is one of the tested methods in Section 4.4. However, in the present work we combine the method with the hierarchical clustering algorithm instead of k-means. *Wang* and *Kitsuregawa* evaluated a contents-link coupled clustering algorithm for retrieved Web pages and studied the effect of out-links, in-links, specific terms, and their combination [266]. Results suggested that both links and contents are important for Web page clustering and that much better results are achieved with appropriate integration weights. *He et al.* [122] discussed Web document clustering by incorporating information from hyperlink structure, co-citation patterns, and textual contents of documents. The hyperlink structure was used as the dominant factor in the combined similarity measure, and the textual content was used to modulate the strength of each hyperlink. The resulting weighted graph was the input to a spectral clustering method.

*Joachims et al.* [141] combined kernel functions for text and co-citation in Support Vector Machine **classification** of hypertext. *Fisher* and *Everson* [82] observed that link information can be useful when the document collection has a sufficiently high link density and if the links are of sufficiently high quality. However, the addition of link information was detrimental for some data sets. For classification they used PLSI and Probabilistic HITS, introduced by *Cohn* and *Hofmann* [51] who described a joint probabilistic model for the contents and interconnectivity of document collections. A mixture model was proposed to define 'topic' factors based on textual content and links. A parameterized stochastic process mimics the generation of documents as part of a larger collection. Their method allowed to identify principal topics of a collection as well as authoritative documents within those topics. Following *Cohn* and *Hofmann*, *Erosheva*, *Fienberg*, and *Lafferty* [81] used a mixed-membership model for both the terms from abstracts and the references in bibliographies, but membership scores were treated as random Dirichlet realizations. By using a Bayesian network model, *Calado et al.* [41, 40] combined link-based similarity measures with text-based classifiers to improve classification results for Web collections. In their experiments on Web pages, the link information alone outperformed the text-only classifier, but the combination could improve results. Finally, *Zhang et al.* applied genetic programming techniques to discover the best fusion framework to integrate citation-based information and structural content in order to improve document classification [276].

### 4.1.2 Overview of the chapter

In the next section we investigate to what extent text mining and 'traditional' bibliometric methods can supplement each other and whether they can be performed serially. It is shown that full-text mining provides reliable results in representing structural aspects of research, whereas bibliometric measures can, in turn, reflect formal characteristics of documented scientific communication that might supplement results obtained from content-based analyses. Bibliometric indicators can, for instance, provide information on how 'theoretical' or ‘applied’ research within the same topic is.

Next, in Section 4.3, we devise a methodology for deeply combining text mining and bibliometrics by integrating text-based and bibliometric information early in the mapping process. More specifically, various information sources are incorporated before an actual clustering algorithm is applied. We mathematically and statistically combine document dissimilarity matrices based on textual information with dissimilarity matrices based on network structure or based on other bibliometric indicators. The integrated document distances can then be passed to a learning algorithm. Weighted linear combination of distance matrices, as well as *Fisher*’s inverse chi-square method (also referred to as *Fisher’s omnibus test*) from statistical meta-analysis, are discussed. Finally, we propose an approach to using Random Indexing for data integration.

Section 4.4 then contrasts clustering and classification performances of sheer text and citation-based methods, of *Fisher*’s inverse chi-square method, of linear combinations and of other data integration schemes. We demonstrate that, in general, text is more powerful than cited references, but that the best outcome is obtained by integration. The introduced integration method based on *Fisher*’s inverse chi-square proves to significantly outperform corresponding text-only and link-only methods, as well as other integration schemes.

Subsequently, in Section 4.5, we revisit the mapping of library and information science (LIS) by using hybrid methods. The added value of an integrated analysis is qualitatively assessed and we investigate whether the clustering outcome is a better representation of the field, compared with text-only clustering as discussed in Section 2.5.

Finally, a hybrid information retrieval strategy consisting of textual and bibliometric components is described and applied to delineate the core literature in bioinformatics.

## 4.2 Mapping by serial combination of text mining and bibliometrics

In this section we investigate to what extent full-text based structural analysis of scientific articles and ‘traditional’ bibliometric methods can supplement each other and whether they can be performed serially to improve on the individual approaches to the mapping of science. The subject of analysis is contemporary bibliometrics and its subdisciplines, as represented in the 2003 volume of the journal *Scientometrics*. We compare text-based clustering results with those of a clustering based on bibliometric indicators, and we assess whether both provide complementary information.

### 4.2.1 Introduction

In a study by *Glenisson*, *Glänzel*, and *Persson* [110], full-text analysis and traditional bibliometric methods were serially combined to improve the efficiency of the individual methods. This methodology was applied to a special issue of *Scientometrics*.[1] The study was based on 19 selected papers that were assigned to five categories. The outcomes have shown that such hybrid methodology can be applied to both research evaluation and information retrieval. The bibliometric part of the pilot study was restricted to simple statistical functions obtained from the papers' reference lists, particularly the *mean reference age* and the *share of references to serial literature* (see Section 3.2.2). Because of the limited number of papers underlying the study, it has to be considered a pilot study that was further extended and confirmed by *Glenisson et al.*[2] Relevant results of this manuscript are discussed in the following sections. The number of papers under study was increased to the complete publication year 2003, i.e., vols. 56–58 of the journal *Scientometrics*, comprising 85 research articles and notes. This data set covered a broader and more heterogeneous spectrum of bibliometrics and related research.

### 4.2.2 Methods

The text representation and pre-processing steps used for this study are comparable to those described in Section 2.1. An overview of the text-based and bibliometric analysis is presented in Figure 4.2: we cluster the documents under consideration with a hierarchical method and compare these results with expert category assignments as well as with a bibliometric analysis. For interpretation purposes, we present top-scoring terms from each cluster in term networks.

Due to the lack of ground truth and the difficulties to define a crisp categorization, we provide an in-depth analysis of how bibliometric, text-based and expert category information provide different views on the thematic structure of the document collection. For a comparison of results from using full-text information with the outcomes of an analysis based on titles and abstracts, and based on terms from the reference lists, we refer to the published manuscript [109]. In short, it has been observed that analysis of full texts provided more pronounced cluster structures than title and abstracts, which, in turn, did better than the information captured in reference titles. Moreover, when manually comparing co-word maps across the three data structures, it was found that the use of full text included more relevant phrases for interpretation.

[1] Scientometrics (2004), vol. 60, issue 3, pp. 273-534.

[2] The study presented here has been published in the journal *Information Processing & Management* [109].

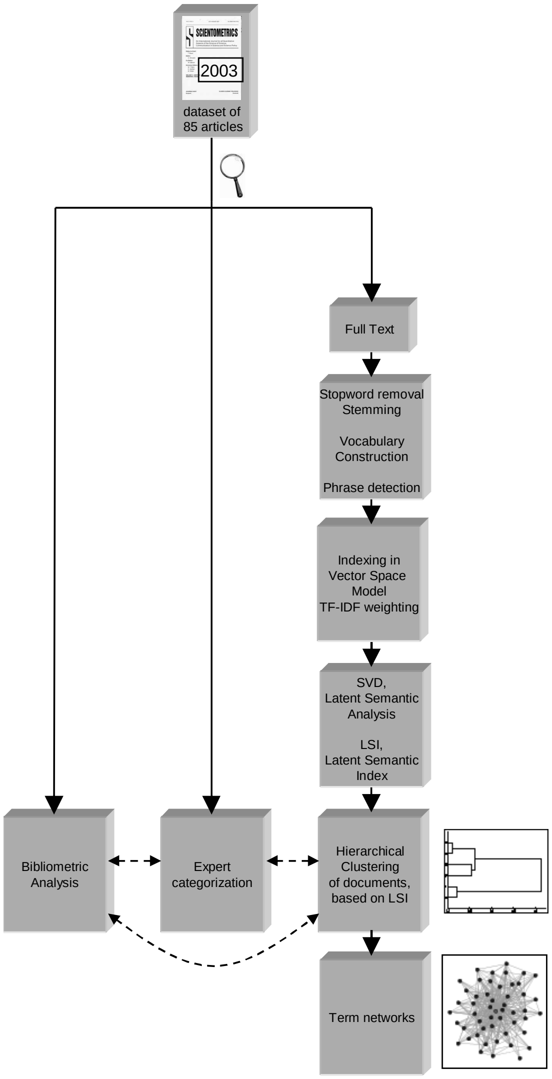


**Figure 4.2:** Overview of the analysis of a set of 85 articles and notes published in *Scientometrics* in 2003. The documents under consideration are pre-processed and clustered with a hierarchical method and the result is compared with expert category assignments as well as with a bibliometric analysis. The manuscript [109] also contrasts the full-text based clustering with clustering based on terms from only titles and abstracts and based on terms from the reference lists. However, this comparison will not be discussed here.

### 4.2.3 Material

The complete publication year 2003 of the journal *Scientometrics* comprises vols. 56–58 with three issues each. Letters to the editor, items on individuals, news items, editorial material and reviews have been omitted from the data set, so that altogether 85 papers were selected. The classification in Table 4.1 was put forward by *Glenisson et al.* [109].

**Table 4.1:** Category scheme for scientometrics papers and their distribution over categories.

| Abbreviation | Description | Share (%) |
|:---:|---|---:|
| A | Advances in scientometrics | 31.8 |
| E | Empirical papers/case studies | 34.1 |
| M | Mathematical models | 2.4 |
| P | Political issues | 17.6 |
| S | Sociological approaches | 3.5 |
| I | Informetrics/Webometrics | 10.6 |

The formerly by and large clear borderlines between indicator research, sociology of science, informetric laws, and science policy have become more and more fuzzy, and are gradually fading away. An increasing number of papers requires double or even triple assignment, in several cases a simultaneous assignment to the categories E, A, and P would be most appropriate. The category assignment thus remains imperfect. A positive effect might be expected from the combination of bibliometric and text-mining methods to monitor, describe, and understand the structure of a field like scientometrics.

### 4.2.4 Clustering of scientometrics in 2003

To get a view on research themes covered by the journal *Scientometrics* in 2003, the title, abstract, and full text of 85 articles were processed and indexed, while ignoring reference lists. Latent Semantic Indexing was used to reduce the rank of the term-by-document matrix to 6 (see Section 2.2.3). The number of clusters was also found to be 6 by observing the dendrogram, the stability diagram of *Ben-Hur et al.* [16], and the Silhouette plot. This combined methodology has been discussed in Section 2.3.2.

The clustering outcome was contrasted with the expert categorization by using the *Rand* index (see Section 2.3.2). A quite low, but still significant value was reported. We again refer to the manuscript [109] for a detailed analysis of this discrepancy by exploration of the confusion table as well as of the lists of documents ranked according to their distance from the corresponding cluster medoid.

The cognitive structure of Scientometrics was visualized with term networks.

For illustrative purposes we show the content structure of cluster 2 in Figure 4.3. It is dominated by empirical papers and case studies and relates above all to national and institutional aspects as well as to science fields. We labeled this cluster as *Case studies and traditional bibliometric applications*.

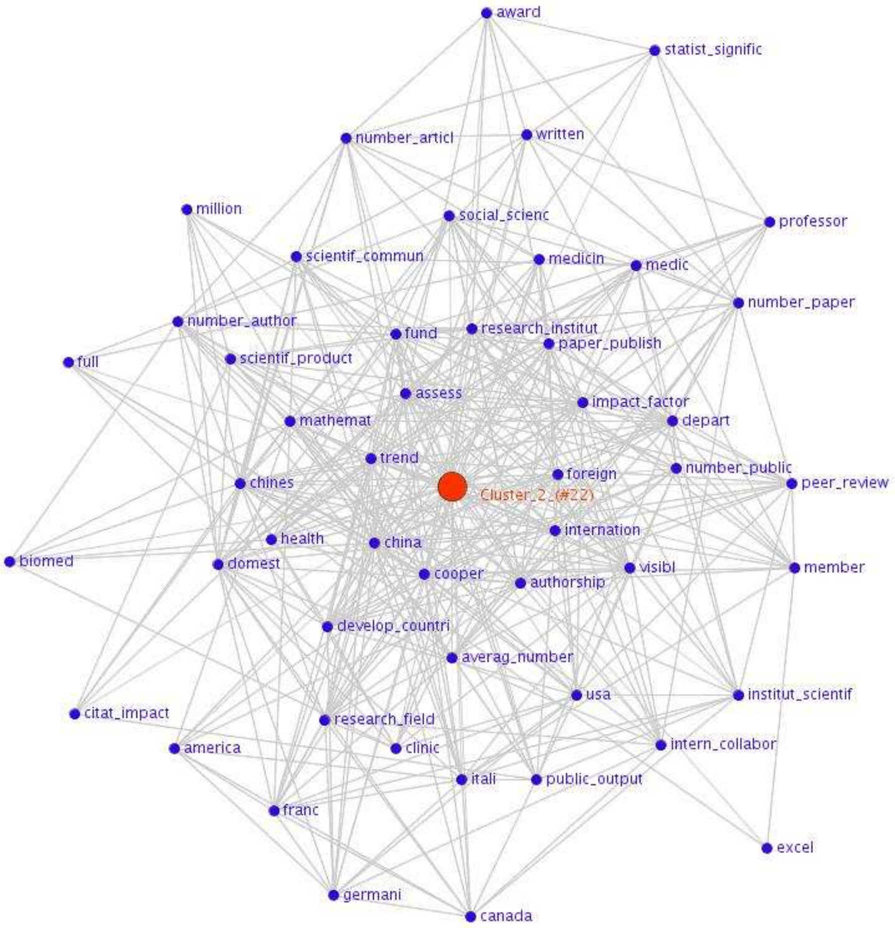


**Figure 4.3:** Term network for cluster 2. The best 50 TF-IDF terms for the cluster are shown. An edge between two terms indicates that both co-occur next to each other in at least one document of the corresponding cluster (ignoring stop words). Cluster 2 is dominated by empirical papers and case studies and relates above all to national and institutional aspects as well as to science fields. We labeled this cluster as *Case studies and traditional bibliometric applications*.

The comparison of the topic structure based on member articles of each cluster with the term networks and with the category assignment in Table 4.1, showed only a partial accordance. The two large categories A and E, covering 65% of all papers, proved heterogeneous. Category A (jointly with category

M) has three sub-clusters, whereas Category E falls apart into three other sub-clusters. Policy relevant issues are also covered by two of these clusters. Only Category I was clearly represented by one corresponding cluster. The full-text analysis substantiates that nowadays both methodological and empirical research each have at least two different main foci. One is based on scientometric standard techniques such as classical indicators, the other are clearly broadening the scope of traditional bibliometrics.

### 4.2.5 Serial combination of text-based clustering and bibliometrics

The statistical analysis of the full texts provided a relational chart of the structure represented by the documents under study. As already used in the pilot study [110], the *mean reference age* (MRA) and the *share of serials* in all references can be used to characterize fields and subdisciplines in the sciences and social sciences. In what follows we check whether these indicators can be used to characterize the six clusters found by the statistical full-text analysis. We first combine the bibliometric approach with the full-text analysis by means of aggregating both results: Figure 4.4 shows the relation between *mean reference age* and *share of serials* with the cluster results as overlay. Clusters are named in the legend by the title of their medoids (i.e., representative elements). Our example, Cluster 2 (indicated by its medoid *Changing trends in publishing behaviour*), is characterized by a medium MRA. The two special issues (*Triple Helix Conference* and *S&T Indicators Conference*) are indicated by ellipses. These issues form surprisingly homogeneous groups, although, in general, there is not much correspondence between text-based cluster membership and common bibliometric characteristics. Papers with similar content might thus have different bibliometric characteristics depending on target readership and field of application. Therefore we deem it an interesting option to integrate these two disparate information sources earlier in the segmentation process. We develop details to such an approach in Sections 4.3 and 4.4.

### 4.2.6 Concluding remarks

A combination of full text and bibliometric information was applied to map 85 papers published in volume 2003 of *Scientometrics*. A fine-grained structure was studied using six clusters and indicators of cited references. A similar polarization of scientometrics literature was found in Section 2.5. Text-based clustering results were compared with those of a clustering based on bibliometric indicators. It was clear that clusters found through application of text mining provided additional information that could be used to extend and explain structures found by bibliometric methods, and vice-versa. Full-text analysis has shown that within categories, such as methodological or empirical research, substantial differences in profile and orientation can occur. The 2003 volume of *Scientometrics* represents almost the complete and heterogeneous spectrum

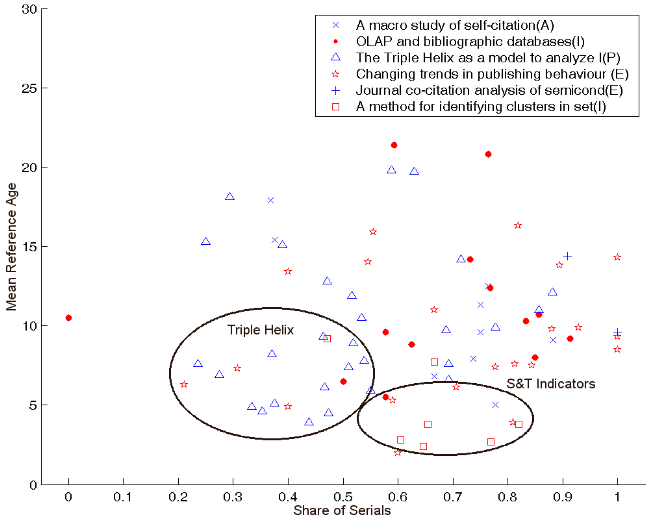


**Figure 4.4:** Plot of *mean reference age* vs. *share of serials* for the documents in different text-based clusters. Clusters are represented in the legend by their medoid documents. In general, there is not much correspondence between text-based cluster membership and common bibliometric characteristics. Papers with similar content might have different bibliometric characteristics depending on target readership and field of application.

of scientometric, informetric and technometric research activity and also covers topics beyond the mainstream in the field. Nevertheless, serial combination of text-mining and bibliometric techniques proved an appropriate tool to unravel the cognitive structure. Hybrid methodologies combining data-mining techniques and bibliometric methods will therefore be developed in subsequent sections and will prove valuable tools to facilitate endeavors in mapping fields of science.

## 4.3 Integrating text and bibliometric information

This section aims at devising a methodology for deeply combining text mining and bibliometrics by integrating text-based and bibliometric information early in the mapping process. More specifically, multiple information sources are incorporated before the clustering algorithm is applied.

The actual integration is achieved by combining various distances between the same pair of documents. Each distance results from possibly different distance measures exploiting different views on the documents. The requisite input for many clustering algorithms indeed includes pairwise (mutual) distances between all objects (documents). These distances can be based on text, on citations or other bibliometric properties, or on a combination of any of these information sources.

We describe weighted linear combination of distance matrices as well as an integration method based on *Fisher*'s inverse chi-square. The quest for even more scalable integration methods leads us to propose an integration scheme based on Random Indexing. It has not been thoroughly assessed in our research so far, but some promising results are shown. Both former methods can be considered *intermediate* integration methods: mutual document similarities are calculated in separate spaces, but integrated before application of the clustering algorithm. Besides RI-based integration, we also use other *early* integration methods that incorporate data even before distance calculation (for example, by appending vectors).

For illustrative purposes, the number of data sources is restricted to two, but straightforward extensions are available to incorporate more. We integrate textual content and citations present in data sets containing bioinformatics and LIS publications, but other bibliometric indicators can be combined as well. Section 4.4 then investigates how clustering and classification performances of linear combinations, of Fisher inverse chi-square method, and of other integration schemes compare with each other and with text-only and link-only methods.

For each data source, such as a normalized *term-by-document* matrix $A$ or a normalized *cited_references-by-document* matrix $B$, square distance matrices

$D_t$ and $D_{bc}$ can be constructed as follows:

$$\begin{aligned} D_t &= O_N - A^T \cdot A \\ D_{bc} &= O_N - B^T \cdot B \end{aligned} \tag{4.1}$$

with $N$ the number of documents and $O_N$ a square matrix of dimensionality $N$ with all ones. *bc* refers to bibliographic coupling. Figure 4.5 provides a visualization of these distance matrix calculations.

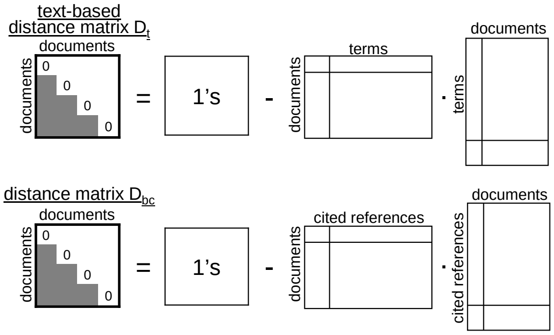


**Figure 4.5:** Visualization of distance matrix calculations.

Figure 4.6 illustrates our motivation for integrating text and citation information. It shows the amount of overlapping and distinctive information present in the text world vs. the citation world for 7401 bioinformatics publications. The figure results from application of the Quotient Singular Value Decomposition (QSVD) on $D_t$ and $D_{bc}$. For general information on QSVD we refer to *Van Loan* [170] and *De Moor et al.* [191, 190, 189]. *Alter et al.* [7] have recently used QSVD to compare two genome-scale yeast and human cell-cycle expression data sets.

The plot shows the sorted antisymmetric 'angular distance' between the data sets, which visualizes the relative significance of patterns in text vs. citations. Values above the horizontal line at $\pi/8 = 0.39$ represent patterns that are highly expressed in the text relative to the citations, whereas negative values below the line at $-\pi/8 = -0.39$ are more significant for the citation data. Patterns in-between are significantly present in both data sources, or in neither of them. The value 0 indicates equal significance. Hence, Figure 4.6 essentially shows that there is definitely information common to both data sources. Stated otherwise, the similarities as conceived by automatic text-based methods correspond in part to those manifested by numerous individual actions of authors citing documents. Nonetheless, there is also quite some information that is only present in one of both data matrices, which is the reason for our endeavors to integrate both worlds.

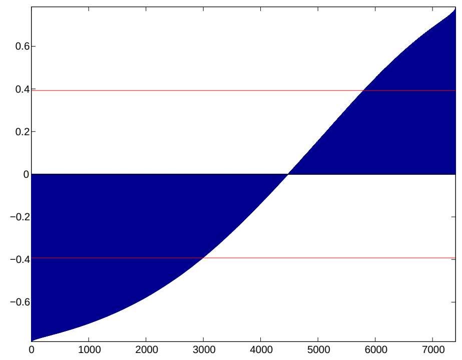


**Figure 4.6:** Sorted *angular distances* between $D_t$ and $D_{bc}$, indicating the relative significance of patterns in the text and citation worlds. Values above the horizontal line at $\pi/8 = 0.39$ go with patterns that are highly expressed in text relative to citations, whereas negative values below the line at $-\pi/8 = -0.39$ indicate patterns that are more significant for the citation data. Patterns with values in-between are significantly present in both data sources (0 indicates equal significance), or in neither of them. There is definitely information common to both data sources, but there is also quite some information only present in one of both data matrices.

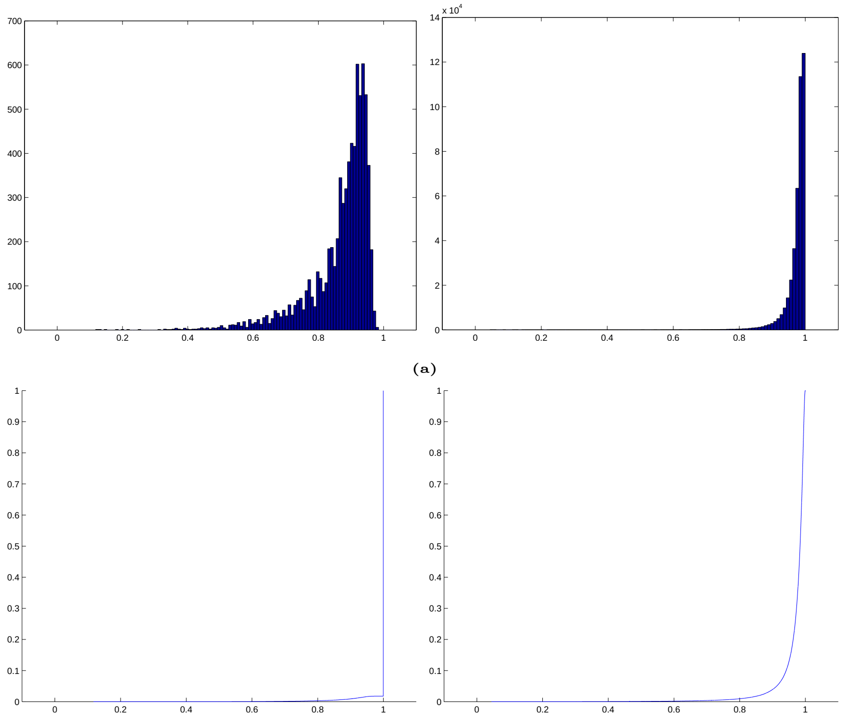
## 4.3.1 Weighted linear combination of distance matrices

The distance matrices $D_t$ and $D_{bc}$ can be combined into an integrated distance matrix $D_i$ by a weighted linear combination (linco) as follows:

$$D_i = \alpha \cdot D_t + (1 - \alpha) \cdot D_{bc} \tag{4.2}$$

The resulting $D_i$ can then be used in clustering or classification algorithms. A comparable methodology was described as the toric $k$-means algorithm by *Modha* and *Spangler* [185], but in the present work it was used with the hierarchical clustering algorithm instead. Although this is a very attractive, easy and scalable integration method, caution should be taken as a linear combination might neglect different distributional characteristics of various data sources. In Figure 4.7(a) we plot the histograms of mutual distances (different from 1) between documents from the LIS data set based on bibliographic coupling (left) and textual information (right). Although in this case the use of *Salton*'s cosine measure in both worlds leads to the same interval (range) of possible distances, the actual distance distributions differ. Also note the different scale on the $Y$-axis in both figures. Figure 4.7(b) shows the empirical cumulative distribution functions of all mutual distances, including those equal to 1. The differences become even more apparent, and the sparsity of bibliographic coupling is noticeable as a large amount of distances are equal to 1 ($> 95\%$).

The discrepancy in distributional characteristics can turn out even more severe when other information sources are considered. For instance, we have combined textual distances with artificial Euclidean distances computed in a



(a)

(b)

**Figure 4.7: (a)**. Histograms with mutual distances smaller than 1, and **(b)** empirical cumulative distribution functions of all mutual distances, between documents based on bibliographic coupling (left) and textual information (right). The distance distributions clearly differ (also note the different scales on the $Y$-axis in (a)). This difference in distributional characteristics is neglected by linear combinations.

two-dimensional space formed by two bibliometric indicators.[3] Different data matrices (such as *term-by-document* and *indicator-by-document*) may indeed require a different choice of distance metric. If the integration weight $\alpha$, which is difficult to determine, would naively be set to 0.5, differences in corresponding distributions might lead to an unequal or unfair contribution of both data sources in the ultimate integrated data, and thus possibly yield suboptimal results by implicitly favoring text over bibliometric information or vice versa. Spurious and strong (dis)similarities might obliterate good relationships established by the other data source. Moreover, different distributional characteristics create additional problems on the transparency of the integration weight.

### 4.3.2 *Fisher*'s inverse chi-square method

As a plain linear combination might not be the best solution for integrating textual and bibliometric information, we developed a methodology based on *Fisher*'s inverse chi-square method. *Fisher*'s inverse chi-square is an omnibus statistic from statistical meta-analysis to combine $p$-values from multiple sources [123]. In contrast to the weighted linear combination procedure, this method can handle distances stemming from different metrics with different distributional characteristics and avoids domination of any specific information source. *Glenisson* [108] has proposed this method as a means to integrate distances stemming from both text and gene expression data. In this section, the method is described in more detail and the rescaling of distances is improved by calculating $p$-values with respect to randomized data sets. This randomization is a necessary condition for having valid $p$-values. We also propose ways to estimate the integration weight $\lambda$ and present a modified formula for bibliographic coupling and a superimposed noise factor in order to tackle the problem of discontinuous test statistics. Finally, the use of the method in combination with SVD is discussed.

Figure 4.8 illustrates the concept of distance integration by using *Fisher*'s inverse chi-square method to combine $p$-values compared to randomized data sets. All text-based and link-based document distances in $D_t$ and $D_{bc}$, as described in the previous section, are transformed to $p$-values with respect to the cumulative distribution function of distances for randomized data. In our setting, a $p$-value means the probability that the similarity of two documents could be at least as high just by chance.

The randomized data sets can be constructed in several ways. The randomization should be as complete as possible, but should obey some rules that apply to the nature of the data. For instance, concerning cited references (bibliographic coupling), all citations in all reference lists of the complete document set are randomly shuffled, while retaining the number of references in each document as well as the total number of times each individual reference is cited (popularity). However, after randomization a document should never cite any

[3] This study has also been presented at the 10th international conference of the International Society for Scientometrics and Informetrics (ISSI) [136].

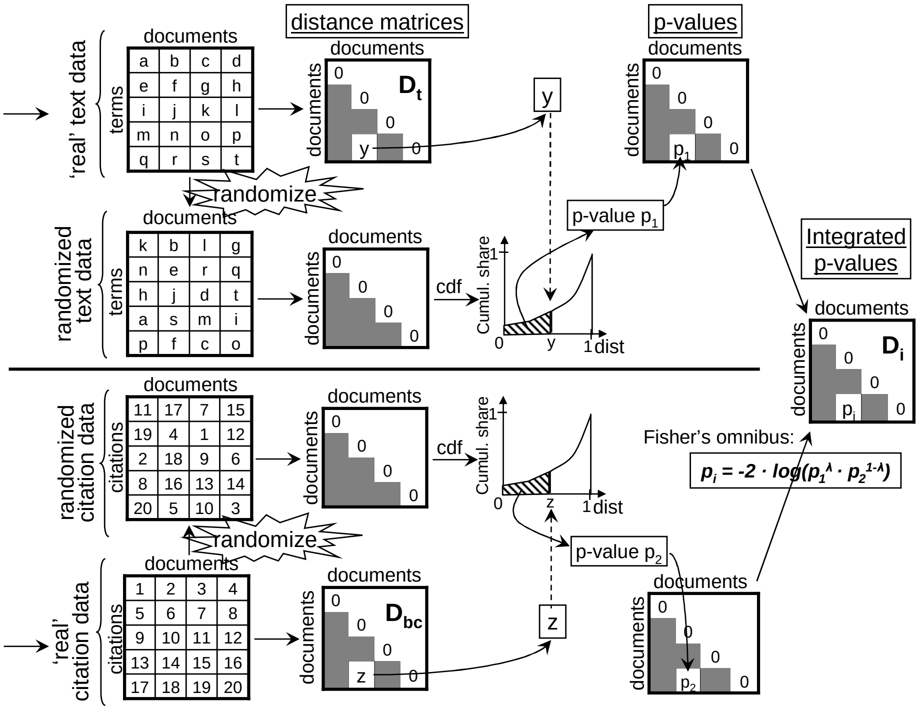
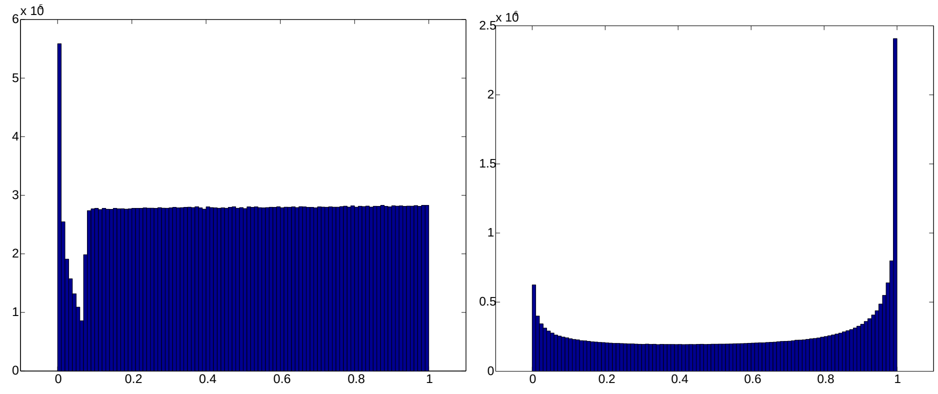


**Figure 4.8:** Distance integration by using *Fisher*'s inverse chi-square method. All text-based and link-based document distances in $D_t$ and $D_{bc}$ are transformed to $p$-values with respect to the cumulative distribution function of distances for randomized data. For randomization, term occurrences (and citations) are randomly shuffled between documents, while maintaining the average characteristic document frequency of each term. This randomization is a necessary condition for having valid $p$-values. In our setting, a $p$-value means the probability that the similarity of two documents could be at least as high just by chance. An integrated statistic $p_i$ can be computed from the $p$-values for the textual data ($p_1$) and for the link data ($p_2$) by application of *Fisher*'s omnibus. The ultimate matrix with integrated $p$-values is the new integrated document distance matrix that can be used in clustering or classification algorithms.



(a)

(b)

**Figure 4.9:** Histogram of $p$-values corresponding to **(a)** bibliographic coupling, and **(b)** textual pairwise document distances, for the real data compared to randomized data. The distributions of $p$-values are not uniform because if they were there would be no structure in the data as the distribution of distances would be the same as for randomized data. The peaks at 0 and 1 indicate that for the real data, compared to random data, more document pairs have a very small or a very large distance. For bibliographic coupling (a), a peak around 1 is missing because the number of document pairs that are not bibliographically coupled is very large for the real data as well as for random data.

other document more than once. In the 'real world', references in one scientific article are also unique. Neglecting this rule would probably not severely influence results, but for the textual data such blind randomizations might destroy important properties of human language. If randomization of textual content would also be implemented by simply permuting all occurrences of all terms in all documents, the fact would be neglected that in human discourse some terms carry much more meaning than others, and in general have a very skewed distribution over documents. Other terms, such as conjunctions, are present in every document.

We considered different randomization schemes and finally opted for the somewhat more conservative randomization which maintains the relative importance between terms by keeping the inverse document frequencies for each term from the real data intact. Hence, term occurrences are randomly shuffled between documents, but the average characteristic document frequencies per term are preserved. These inverse document frequencies are measures for the *a priori* relative importance of terms in the distance calculation between documents. It should also be noted that the choice of text randomization scheme will largely be compensated by an automatic determination of the integration weight $\lambda$ as will be explained further on.

For a correct application of *Fisher*'s inverse chi-square method the input test statistics should be continuous. Indeed, the distribution of $p$-values for randomized data will only be uniform under this essential condition. However, this is not the case for the sparse BC since most pairs of scientific articles do not have any reference in common. Consequently, the bibliographic coupling matrix contains an enormous amount of values equal to zero (cf. the many values equal to 1 in Figure 4.7 (left) which shows the complement of BC). The few non-zero entries will be converted to very small $p$-values ($< 0.05$), while all other $p$-values will equal 1. This is far from a desired uniform distribution of $p$-values under the null hypothesis (i.e., absence of any structure in the data). For this problem, we defined a slight, rank-preserving modification to the original formula for BC by adding a constant 0.01 to the numerator. The new *dense* bibliographic coupling between two papers $x$ and $y$ is then:

$$dBC(x, y) = \frac{N_{xy} + 0.01}{\sqrt{N_x \cdot N_y}}, \tag{4.3}$$

with $N_x$ and $N_y$ the number of references in paper $x$ and paper $y$, respectively, and $N_{xy}$ the number of references in common. The advantage of this formula is that it leads to a much larger set of possible values. Furthermore, all zero bibliographic coupling values will be distributed between 0 and 0.01, depending on the lengths of the reference lists $N_x$ and $N_y$. This is meaningful as the chance for two papers to have no references in common is indeed lower when the reference lists are longer.

In practice, however, the result of the formula is still a finite set of discrete values. Corrections for continuity exist that might be able to counter the persistent problem of discrete input values [265, 160], but instead we opted to superimpose Gaussian noise, a normally distributed error with mean 0 and a standard deviation of 0.0025. Addition of an appropriate random variable was introduced as an alternative to continuity corrections by *Pearson* [217]. The random noise will not deteriorate results since the error to be expected from missing references in the WoS database is much higher. Moreover, this new formula for dense BC, including the noise factor, can lead to comparable clustering performances and even to classification accuracies significantly higher than those of the original formula, as will be shown in Section 4.4.

Actually, the added noise factor even makes the adjusted formula for dBC superfluous, but we opted to keep it. Otherwise, if only the noise factor would be used, the rank order between various $p$-values corresponding to zero BC would just be based on coincidence. With the constant added to the numerator, this ranking depends more on the lengths of the reference lists. Hence, a pair of papers with no common references can get a lower $p$-value if the reference lists are shorter.

With a data set containing full-text articles we observed no similar continuity problem when calculating textual distances, because every single distance value was unique and different from 1, but when dealing with titles and abstracts the chance of having no overlapping terms between two documents is higher. A similar addition of a constant in the numerator and a noise factor can analogously be applied when calculating textual similarities.

If the $p$-values for the textual data ($p_1$) and for link data ($p_2$) are calculated, an integrated statistic $p_i$ can be computed as

$$p_i = -2 \cdot log(p_1^{\lambda} \cdot p_2^{1-\lambda}), \tag{4.4}$$

with $0 < \lambda < 1$ the integration weight determining the relative quality of both data sources and their contribution to the ultimate incorporated data. If the null hypothesis is true (i.e., for randomized data), the distribution of $(p_1^{\lambda} \cdot p_2^{1-\lambda})$ is uniform and the integrated statistic has a chi-square distribution with 4 degrees of freedom [123]. The complement of the integrated $p$-value, $(1 - p_i)$, is the new integrated document similarity measure that can be used in clustering or classification algorithms.

Figure 4.9 shows the histogram of $p$-values corresponding to bibliographic coupling (a) and textual pairwise document distances (b) for the real data compared to randomized data sets. The distribution of $p$-values for the real data is not uniform because if it were there would be no structure in the data as the distribution of distances would be the same as for randomized data. The peaks at 0 and 1 indicate that for the real data, compared to the random data, more document pairs have a very small or a very large distance. For bibliographic coupling (a), a peak around 1 is missing because the number of document pairs that are not bibliographically coupled is very large for the real data as well as for the random data.

The weight $\lambda$ can be used to tune the relative importance or 'quality' of both information sources, which is an important issue. For example, term-based approaches, bibliographic coupling, co-citation information, or bibliometric indicators each have particular strengths and weaknesses on particular types of data. However, choosing a good value for $\lambda$ is not straightforward. We propose to define $\lambda$ by choosing a value $x$ for the *smallest but still significant* bibliographic coupling link (for example, $x = 0.05$) and a value $y$ for the smallest text-based similarity that is also still significant (for example, $y = 0.1$). $x$ And $y$ can be based on visual inspection of the histogram of similarities, in combination with some experience. Next, convert the distances $(1 - x)$ and $(1 - y)$ to $p$-values $p_x$ and $p_y$, respectively, and choose $\lambda$ such that both weakest still significant links have the same contribution in $p_i$, by asking that

$$p_x^{\lambda} = p_y^{1-\lambda} \tag{4.5}$$

Therefore we compensate for the fact that significant similarities are not as numerous in both data sets.

*Fisher*'s inverse chi-square method can also be applied if SVD is used as a pre-processing step for either the textual data (LSI), either for the citation-based component, or for both. The random document vectors should then first be projected in the same space of reduced dimensionality before calculating the distribution of document similarities. After application of SVD, intuitively defining the *smallest but still significant distance* by an expert becomes much more difficult and, moreover, the distribution of document similarities usually no longer has a clear cut-off point. However, a parameter sweep can still be performed and the difficulty of defining $\lambda$ is compensated by the augmented performance after applying SVD, especially on the textual data.

Alternatively, relative classification accuracies of classifiers trained on either data type might help in automatically providing an estimate for $\lambda$. Likewise, fast clustering procedures could help in estimating meaningful structure present in either data source. For instance, by computing a *Silhouette Value per Clustering* (SVC) for each data type [133], we can estimate the relative quality of each data source and use this as an educated guess for $\lambda$:

$$\lambda = \frac{SVC^t}{SVC^{bc} + SVC^t} \tag{4.6}$$

*Fisher*'s inverse chi-square method was also used in a setting where textual data was combined with the two bibliometric features of Figure 4.4, namely *mean reference age* and *share of serials*. Pairwise document distances were computed with the classical Euclidean distance measure in the two-dimensional space formed by both features. We recall that the use of different distance metrics entails difficulties for incorporating information and justifies the use of the complex *Fisher*'s inverse chi-square method.

### 4.3.3 Integrated Random Indexing

Random Indexing (RI), described in Section 2.2.4, can also be used to index citations in bibliographies besides words in texts. Furthermore, the method can even be modified to obtain an *integrated* random index containing both textual and citation information. We propose an approach to using RI for information integration in the remainder of this section. As an aside, a weighted linear combination could also be used to integrate mutual similarities stemming from two distinct random indices, one based on text and one based on citations, but we do not discuss that option in detail here because of analogy with Section 4.3.1. Moreover, it would necessitate storing two random indices in memory.

We have observed that an RI with textual information from the LIS data set provided good clustering performance, and that an RI with citation information even did slightly better than bibliographic coupling (probably due to the incorporation of context). Since these two experiments provided promising results, namely that the *bag-of-concepts* approach of RI seemed to work on the LIS data set, we propose to modify RI for data integration. To the best of our knowledge RI has not been used for data integration before, although *Sahlgren* has put forward the possibility to introduce linguistic properties into the model [233].

Following Section 2.2.4, *context* vectors can also be constructed for all citations in the data set, and an integrated *bag-of-concepts* representation of a document can then be built by adding all (weighted) context vectors of all terms and of all cited references occurring in the document. If the integrated

RI would be constructed without weighting the relative contributions of words and citations, again one of the data sources could dominate the other, especially because a term-by-document matrix $A$ is usually much more dense than the matrix $B$ which indexes cited references in each document. Therefore, we propose the following weight $\alpha$ to boost the contribution of citation context vectors to the final bag-of-concepts representation of a document, relative to word context vectors. It is based on the relative sparseness of both $A$ and $B$.

$$\alpha = \frac{\frac{\sum_i \sum_j A_{i,j}}{t \cdot d}}{\frac{\sum_i \sum_j B_{i,j}}{r \cdot d}} = \frac{r \cdot \sum_i \sum_j A_{i,j}}{t \cdot \sum_i \sum_j B_{i,j}}, \tag{4.7}$$

with $d$ the number of documents, $t$ the term dimension of $A$, and $r$ the reference dimension of $B$. In Section 4.5.4, a small experiment with integrated Random Indexing is discussed.

## 4.4 Assessing various integration schemes for text & link information

### 4.4.1 Introduction

As stated in the outset of this chapter, hybrid clustering methods that exploit both text and citations might achieve better results than pure text-based or link-based methods. The purpose of this section is to assess clustering and classification performances of methods that use just text or only cited references, of *Fisher*'s inverse chi-square method (see Section 4.3.2) and of other schemes for integration of textual contents with the structure of the citation graph.[4] A set of documents published in a list of core bioinformatics journals is extracted from the Web of Science (WoS) and extended with bibliometrically related records. We only consider cited references and evaluate clustering results by the mean Silhouette coefficient. We also assess the performance in a classification setting for which we construct a 'ground truth' based on the Medical Subject Headings[5] (MeSH), annotated by experts. To retrieve the MeSH terms, each WoS document was matched against MEDLINE. Latent semantic analysis and a hierarchical clustering technique were applied to the *MeSH-by-document* matrix to determine document clusters, which were then post-processed by an automatic iterative shrinking technique to only retain well-defined categories. Clustering and classification performances of sheer text and citation-based methods, of *Fisher*'s inverse chi-square method and of other data integration schemes, are compared.

We are interested in unsupervised clustering rather than building an optimal classifier assigning documents to predefined categories, since an accurate

[4] The results presented in this section have been presented at the *International Conference on Multidisciplinary Information Sciences & Technologies* (InSciT2006) [139].

[5] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=mesh

classification of scientific articles is not available and would otherwise indisputably be outdated because of the dynamic nature of contemporary science and technology. However, we also evaluate the various experimented data types in a classification setting as it offers a well-grounded basis for assessing relative performance.

### 4.4.2 Material

Our data set consists of 5188 bioinformatics-related papers which are available in the *ISI* Web of Science database (WoS). For details about the subject delineation strategy we refer to section 4.6. For this experiment only 5188 out of 7401 publications were considered, namely those that could be matched with MEDLINE and that were annotated with at least 5 MeSH terms.

### 4.4.3 Methods

We assessed the performance of one hierarchical clustering and one classification algorithm using 13 different data types, 2 of which consisted of only textual information, 3 only of cited references, whereas 8 were integrated data types. Table 4.2 lists the 13 experimented data types.

For sheer textual data types (#1 & #2) we indexed titles and abstracts in the Vector Space Model, while neglecting stop words, URLs, and e-mail addresses, and cutting off *Zipf*'s curve. Bigrams (phrases composed of two words) were detected in a candidate list of all noun phrases, MeSH phrases, and index terms. The *Porter* stemmer and the TF-IDF weighting scheme were applied, and for data type #2 the dimensionality (8679 terms) was reduced to 50 factors by using Latent Semantic Indexing (LSI) (see Section 2.1).

Except for data types #4, #11, #12 and #13, which have been discussed in Section 4.3.2, similarities between documents were computed by the cosine measure between the normalized textual, citation-based or integrated vectors. For vectors with cited references this corresponds to bibliographic coupling (#3) (see Section 3.2.5). Dimensionality reduction of the *references-by-documents* matrix from 38 660 references to 50 factors was also performed by using a truncated SVD (#5). For data type #6, both the text vector and out-link vector of a document were concatenated, and data type #7 is derived from the application of SVD on the concatenated matrix. Integrated data types #8, #9, and #10 result from a weighted linear combination of document similarities (see Section 4.3.1), possibly with SVD applied on either component. Data type #8 is equivalent to #6 if an integration weight $\alpha = 0.5$ is used.

#### Clustering and classification

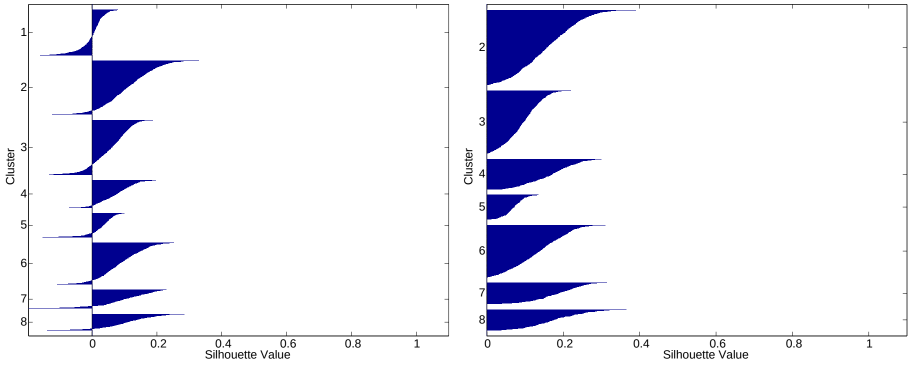
We assessed the quality of clustering results by calculating the *adjusted Rand* index and the mean Silhouette coefficient [146] (see Section 2.3.2). Since no expert-

**Table 4.2:** The 13 experimented data types, indicating (with 'x') whether they contain a text component, a citation-based component, or both, and whether SVD was used for dimensionality reduction.

| Data type number and description | Text component | | Citation-based component | |
|---|---|---|---|---|
| | SVD | No SVD | SVD | No SVD |
| 1. Term-by-document matrix | | X | | |
| 2. Latent Semantic Index (LSI) | X | | | |
| 3. Bibliographic coupling (BC) | | | | X |
| 4. "Dense" bibliographic coupling | | | | X |
| 5. Truncated SVD of references-by-document matrix | | | X | |
| 6. Concatenation of text and reference vectors | | X | | X |
| 7. Concatenation of text and reference vectors, with SVD | X | | X | |
| 8. Linear combination of similarities, without SVD | | X | | X |
| 9. Linear combination of similarities, with LSI | X | | | X |
| 10. Linear combination of similarities, with LSI & SVD | X | | X | |
| 11. *Fisher*'s inverse chi-square method, without SVD | | X | | X |
| 12. *Fisher*'s inverse chi-square method, with LSI | X | | | X |
| 13. *Fisher*'s inverse chi-square method, with LSI & SVD | X | | X | |

made classification of the bioinformatics papers is available, we constructed a 'ground truth' classification based on an optimal clustering of documents, indexed only by their MeSH terms which were never used in further experiments nor in data types. The resulting document clusters, to be considered as classes, were post-processed by an automatic iterative shrinking technique to retain only well-defined categories. One noise cluster was detected and removed. Figure 4.10 shows the quality of the classes (Silhouette plot) before and after iterative shrinking. The ultimate set of 7 clusters was used as the gold standard classification of documents. Besides, to also assess performances at another level of granularity, a coarser-grained classification was used that contained only two iteratively shrunk classes.

For classification experiments we adopted the $k$-Nearest Neighbour classifier (kNN) which classifies a document based on the majority class of the $k$ nearest neighbours. For each data type, 10-fold cross-validation with stratified sampling was used to determine the optimal value for $k$, as well as the optimal integration weight $\alpha$ or $\lambda$ for data types from #8 to #13 (Table 4.2). $k$ was chosen from the set $\{5,10,20,50,0.01,0.05,0.1,0.2,0.5\}$, with $k < 1$ denoting a locally adaptive neighbourhood, whereas $\lambda$ was chosen out of 50 equidistant values between 0 and 1. For each distinct data type, the values for $k$ and $\lambda$ that resulted in maximal cross-validation performance were selected, and final classification performances were assessed on independent test sets by calculating micro-averaged accuracies and AUCs (Area Under the ROC Curve, [118]). All experiments were repeated with 20 different random partitions in training, validation and test sets. Boxplots were drawn for each data type and a Wilcoxon signed rank test ($\alpha = 0.05$) was done on all pairs [56].



(a)

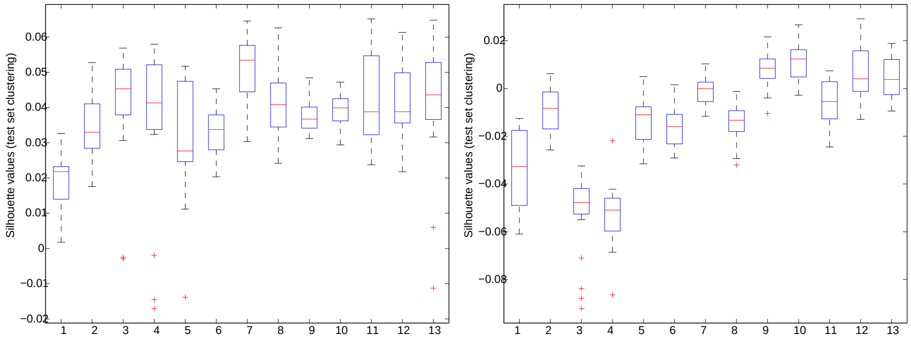
(b)

**Figure 4.10:** Silhouette plots before **(a)** and after **(b)** iterative shrinking of the document clusters based on MeSH. After the shrinking procedure all Silhouette values were positive. Hence, all documents clearly belong to the right cluster and the quality of clusters is improved.

### 4.4.4 Discussion of results

#### Clustering

Figure 4.11 shows clustering performances assessed by overall mean Silhouette coefficient for all data types in Table 4.2, when clustering 20 random test sets into 2 coarse-grained clusters (a) and into 7 finer-grained clusters (b). Silhouette values were calculated on data type independent MeSH-by-document matrices. Figures for *adjusted Rand* index are not included here since relative results were comparable.



(a)

(b)

**Figure 4.11:** Test set clustering performance measured by mean Silhouette value for 2 **(a)** and 7 **(b)** clusters.

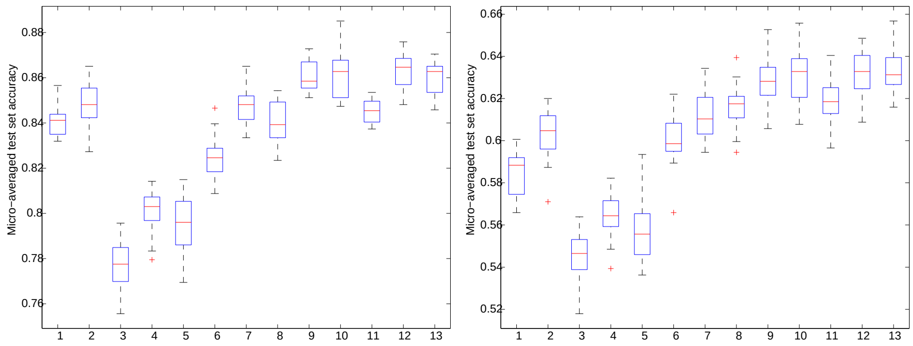
By observing Figure 4.11, we see that: (i) On the coarse level, concatenated matrices subsequently reduced by SVD (#7) provide significantly better results than most other types, although the Wilcoxon signed rank test does not reject equal means when comparing with *Fisher*'s inverse chi-square method (#11 & #13). However, for the more detailed view #7 is outperformed by *Fisher*'s inverse chi-square method and linear combinations with LSI/SVD (#9, #10, #12 & #13), which are the only data types obtaining mainly positive Silhouettes. There is no significant difference between *Fisher*'s inverse chi-square method and corresponding linear combinations.

(ii) For coarse-grained clustering, standard BC (#3) yields quite good results, better than SVD (#5) and text-only (#1), and at first sight even better than LSI (#2) although this difference is not significant. The good performance of BC even degrades when references and textual information are naively concatenated (#6). However, we have observed that for any clustering with more than two clusters, the results for BC were bad.

(iii) On the finer level, pure text or links without SVD (#1, #3 & #4) perform the worst. Application of SVD gives some improvement (#2, #5), but not enough. In this case, mere text is preferred over bare links.

### Classification

Figure 4.12 depicts the classification performance measured by accuracies on 20 random independent test sets, when classifying into 2 categories (a) and into 7 categories (b). Test set AUCs are not shown here because the conclusions were comparable and AUCs are only easily computable for binary classifications.



(a) (b)

**Figure 4.12:** Classification performance measured by test set accuracy for 2 **(a)** and 7 **(b)** classes.

Observing Figure 4.12, we see that: (i) Best performances are obtained by linear combinations and *Fisher*'s inverse chi-square method with SVD applied on the textual component (#9, #10, #12 & #13), but there are no significant differences among these 4 data types.

(ii) Link-only types (#3, #4 & #5) are the worst on both coarse and fine levels. Standard BC (#3) is significantly worse than dense BC (#4), but to a large extent this might be due to papers without references. For these papers, BC provides no information to the kNN classifier, whereas the dense BC between papers with no common references is lower for longer reference lists, i.e., a lower chance to have no reference in common. Surprisingly, dense BC also achieves significantly higher accuracies than SVD on references (#5).

(iii) Text-only methods outperform link-only methods on both levels of detail. Text without SVD (#1) is even better than references with SVD (#5). While on the coarse level plain text (#1) performs as well as linear combination without SVD (#8) and even significantly better than merely concatenated vectors (#6), on the finer level the performances of text-only methods degrade and the outcome of #1 is surpassed by any method that also incorporates out-link information. For binary classification, LSI (#2) is only significantly surpassed by linear combinations and *Fisher*'s inverse chi-square method that also use SVD on textual data (#9, #10, #12 & #13), but linear combination without SVD is worse (#8). On the finer level, LSI is also outperformed by concatenation with SVD (#7) and by linear combination and *Fisher*'s inverse chi-square method without SVD (#8 & #11).

### 4.4.5 Conclusion

The performance of unsupervised clustering and classification of scientific papers can significantly be improved by integrating textual content of titles and abstracts with cited references ('out-links'). In general, for scientific titles and abstracts which are clean pieces of text, text-only information was much more powerful than cited references alone. Dimensionality reduction by SVD can greatly improve results, especially when applied to the textual information.

The introduced integration method based on *Fisher*'s inverse chi-square has shown to significantly outperform corresponding text-only and link-only methods, as well as a mere concatenation of vectors. Only for the coarse-grained clustering (2 clusters) the SVD of concatenated matrices did at least equally well. *Fisher*'s inverse chi-square method, however, did not significantly outperform corresponding linear combinations when SVD had been applied. Given the higher complexity of implementing *Fisher*'s inverse chi-square method and a reduced scalability, a carefully chosen weighted linear combination might be the preferred solution for integrating textual and citation information if LSI is used. However, as discussed in Section 4.3.2, the inverse chi-square method is generic and can be used to incorporate distances with highly dissimilar distributional characteristics, such as textual distances and distances based on the bibliometric features *mean reference age* and *share of serials* [136]. In the next section, we further examine the relative performance of linco and *Fisher*'s inverse chi-square method, quantitatively as well as qualitatively.

## 4.5 Hybrid mapping of library and information science

### 4.5.1 Introduction

In section 2.5 the concept structure of LIS was obtained by using full-text mining of almost 1000 articles and notes published in the period 2002–2004 in 5 representative journals. Only the 'pure' text corpus was analyzed, excluding any bibliographic or bibliometric component. Nevertheless, in previous sections of this chapter it is shown that better results can be obtained by hybrid methods that exploit both text and citations. In this section, we also qualitatively asses the added value of such an integrated analysis and investigate whether the clustering outcome is a better representation of the field.[6] As discussed in the previous section, the integration method based on *Fisher*'s inverse chi-square and another one based on linear combination of distance matrices (linco) were the best methods and significantly outperformed corresponding text-only and link-only methods. However, no significant difference could be observed among both hybrid methods. Reason enough for a more detailed comparison in a different setting. We use these methods to provide a new mapping of LIS by using the full-text as well as citations, and we compare the results with the text-only clustering of Section 2.5. Because the two Bibliometrics clusters are merged by the present hybrid clustering, the number of clusters for the field is 5, one cluster less than the number reported in the text-only setting. Term networks present the updated cognitive structure of the field and are complemented by representative publications. In addition, for data integration with Random Indexing we report on a promising result.

### 4.5.2 Data set

For this study the same data set was used that has been introduced in Section 2.5, except for the exclusion of 24 publications. By matching the articles with the WoS database, we noticed that 22 were actually not indicated as *article* or *note*. There was one *letter* among them, 6 were *reviews*, 11 *editorial materials* were included, as well as 4 *biographical-items*. Finally, two duplicate publications were detected in the original set. The exemption of 22 unique papers (or 2.3%) for this analysis, will presumably not distort results much, particularly because the documents were removed from clusters in a reasonably stratified manner (12 from the largest Bibliometrics cluster, 5 from IR, 3 from the Social cluster, and 1 each from Webometrics and Bibliometrics2). Moreover, the optimal number of clusters for text-only clustering of the remaining 914 publications still amounts to 6 as shown in the following subsections. The text-based clustering has been redone on the smaller data set before comparing with the hybrid clustering.

[6] The results presented here are accepted for publication in *Proceedings of the 11th International Conference of the International Society for Scientometrics and Informetrics (ISSI)* [135].

### 4.5.3 Methodology

For the text representation we refer to Section 2.1 and for the clustering methodology to Section 2.3, but we do mention again that we use a 'hard' clustering algorithm, which means that every publication is assigned to exactly 1 cluster.

### 4.5.4 Results

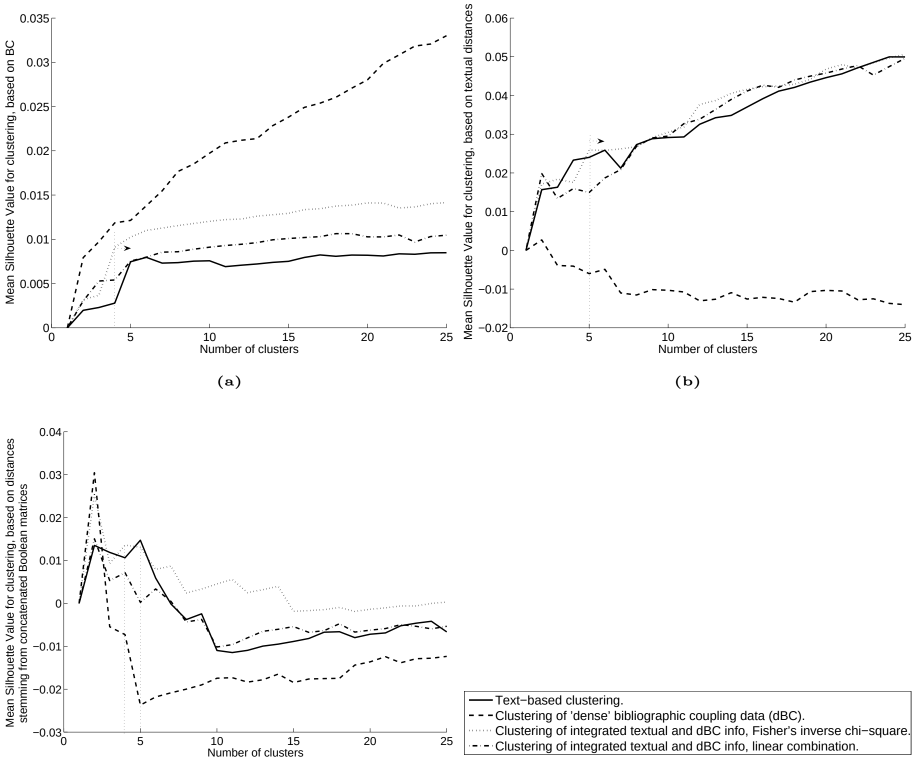
#### Number of clusters

As Silhouette values are intrinsically based on distances [232], depending on the chosen source of distances different Silhouettes can be calculated (see Section 2.3.2). In Figure 4.13 we use the complement of cosine similarity as distance measure, but each time with a different input matrix. In (a), the *cited_references-by-document* matrix was used, whereas the *term-by-document* matrix was the input for (b). Finally, for (c), integrated distances were calculated from both matrices concatenated.

**Table 4.3:** Optimal number of clusters for *Fisher*'s inverse chi-square method as perceived by the stability-based method (Figure 4.14) and by different mean Silhouette curves in Figure 4.13 using link-based (a), text-based (b) and integrated distances (c).

| Evaluation method | Number of clusters |
|---|---|
| Mean Silhouette value based on BC (Figure 4.13(a)) | $\geq 4$ |
| Mean text-based Silhouette value (Figure 4.13(b)) | $\geq 5$ |
| Mean 'hybrid' Silhouette value (Figure 4.13(c)) | 4 or 5 |
| Stability diagram (Figure 4.14) | 3, 4 or 5 |

In the experiments of Figure 4.13, the integration weight was set to 0.5 for both linco and *Fisher*'s inverse chi-square method for simplicity of comparison, but conclusions with regard to number of clusters remain the same (see also Table 4.3). Regarding the number of clusters for hybrid clustering by *Fisher*'s inverse chi-square method, the curve with citation-based Silhouettes (Figure 4.13(a), curve for '*Fisher*'s inverse chi-square') hints towards 4, 5, 6, or more clusters, whereas the text-based Silhouettes show a local maximum for 5 clusters (b). Figure 4.13(c) suggests 4, or maybe 5 clusters, but not more. By observing the stability diagram in Figure 4.14 it can be concluded that a solution with 5 clusters is clearly more stable than 6 clusters, while not differing that much in stability from 3 or 4 clusters. Based on these findings we chose 5 as the number of clusters for the inverse chi-square integrated clustering. On the dendrogram (not shown), five clusters could also be considered as a nice cut-off point.



**(a)**

**(b)**

**(c)**

**Figure 4.13:** Silhouette curves with mean Silhouette coefficient for clustering solutions of 2 up to 25 clusters for text-only clustering, link-only clustering, integrated clustering with *Fisher*'s inverse chi-square method, and integrated clustering by linear combination of document similarities. Silhouette values are based on distances calculated from **(a)** the complement of bibliographic coupling, **(b)** the complement of textual similarities, and **(c)** the complement of cosine similarities calculated on concatenated matrices with text (weighted by IDF) and cited references.

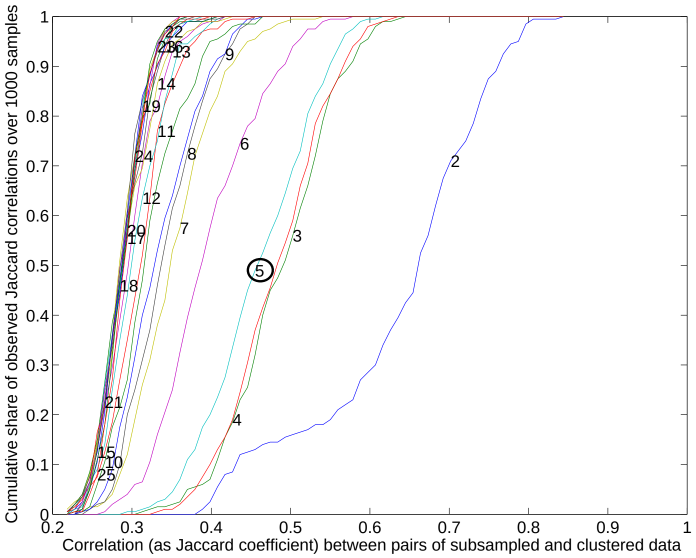


**Figure 4.14:** Stability diagram for determining the number of clusters for hybrid clustering with *Fisher*'s inverse chi-square method as explained in section 2.3.3 [16].

### Comparing *Fisher*'s inverse chi-square method with linear combination, text-only & link-only clustering

When using the same composite procedure for determining the number of clusters with the linco method, two observations could be made. Firstly, when clustering the linearly combined distance matrices, the best number of clusters was 8, compared to 5 for *Fisher*'s inverse chi-square method. Secondly, linco came up with very large, noisy clusters for any solution with less than 8 clusters. For example, when asking for 5 clusters, the largest cluster contained 722 out of 914 documents. Figure 4.13 also presents a more detailed comparison of the performances of linco, *Fisher*'s inverse chi-square method, text-only and link-only clusterings. Main observations are summarized in Table 4.4.

In (a), not surprisingly, the link-based clustering of dBC values performs best. However, when validating with textual (b) or integrated distances (c), this link-only clustering performs very poorly. Integration of text and cited references leads to better Silhouettes than pure text-based methods in (a). From the same figure it is also clear that *Fisher*'s inverse chi-square method does a better job than linco, perhaps an illustration of textual information possibly dominating citations in case of plain linear combinations. Furthermore, linco provided somewhat less stable clustering than *Fisher*'s inverse chi-square method.

Quite favorable but a little counterintuitive is that, when the validation relies on pure text-based Silhouettes (b), *Fisher*'s inverse chi-square method does at

**Table 4.4:** General appreciation of clustering different data types by observing the mean Silhouette curves in Figure 4.13 using link-based (a), text-based (b) and integrated distances (c). A lower value indicates a better appreciation, 1 is best and 4 is worst. Different values are possible for different ranges of cluster numbers, indicated between brackets.

| | Text-based clustering | Clustering of dBC | *Fisher*'s inverse chi-square | Linear combination |
|---|---|---|---|---|
| Mean Silhouette value based on BC (see Figure 4.13(a)) | 4 | 1 | 2 | 3 |
| Mean text-based Silhouette value (see Figure 4.13(b)) | 3 ($c = 2$,$c > 10$)<br>1 or 2 otherwise | 4 | 1 or 2 | 3 ($c = 3..6$)<br>1 or 2 otherwise |
| Mean 'hybrid' Silhouette value (see Figure 4.13(c)) | 1 ($c = 3$ or 5)<br>4 ($c = 2$)<br>2 or 3 otherwise | 1 ($c = 2$)<br>4 otherwise | 2 ($c = 2$, 3 or 5)<br>1 otherwise | 2 or 3 |

least equally well as the pure text-based clustering (which actually *plays a home game* here), except for a four clusters solution. The linco method is the best one on a very coarse level of aggregation with only two clusters, but then goes down. From 7 clusters onwards linco again does as good as or even better than text-based clustering and for more than 10 clusters it competes with the curve for *Fisher*'s inverse chi-square. Thus, based on evaluation with textual Silhouettes, *Fisher*'s inverse chi-square method in general again outperforms linear combination.

In (c), which represents the most natural way of evaluating integrated clusterings, namely by basing the Silhouettes on integrated data, *Fisher*'s inverse chi-square method is again the method of choice. Surprisingly, the clustering of linearly combined data is not better than the text-only clustering, maybe another illustration of textual data dominating citations. Interestingly, the local maximum of the text-based solution at six clusters in figure (b), and as also described in Section 2.5, also decreases to 5 clusters in (c), when evaluated with integrated Silhouette values.

### Computational cost

The computational cost of *Fisher*'s inverse chi-square method is higher than the cost of the linear combination method. First of all, randomization of the data is an extra time consuming step. However, for a data set with 7401 documents, 18 163 terms, and 67 140 cited references, an implementation that was not optimized for speed only took a few minutes on a 2.4 MHz machine with 4 GB of memory. Randomization can also be restricted to a sub-sample of the documents in case of a very large data set.

Another time consuming step next to the actual clustering algorithm is the calculation of mutual document distances. For *Fisher*'s inverse chi-square method 4 different types of pairwise document distances have to be computed: for the real text data, for the real references, and for both randomized variants. For linear combination only two pairwise distance matrices are computed.

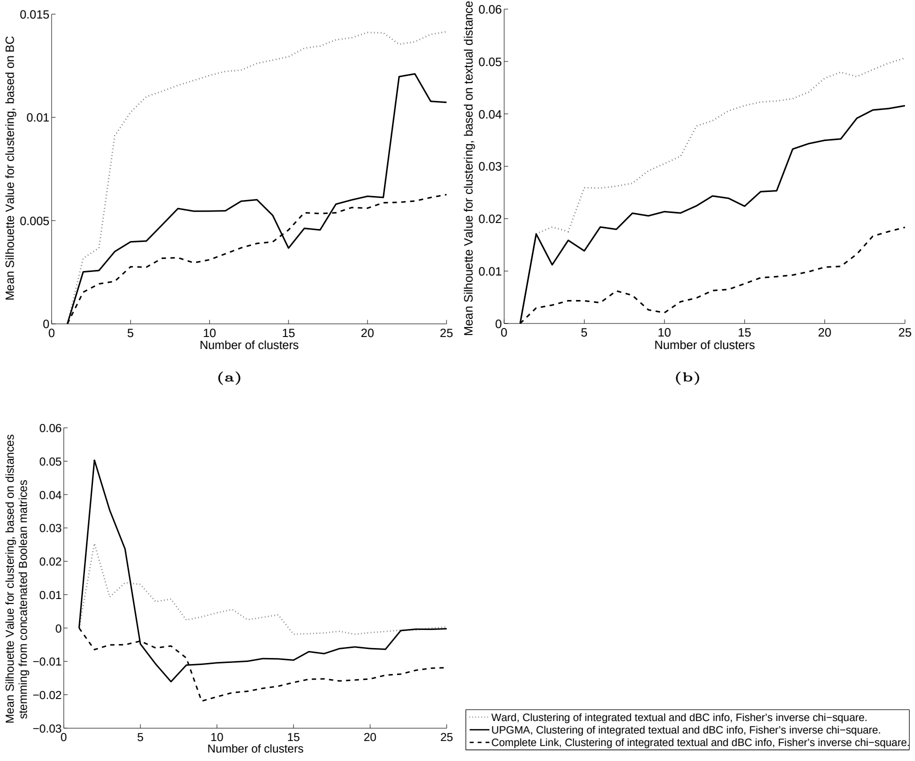
The actual integration formula is also a bit more costly in the case of *Fisher*'s omnibus. The time complexity of the hierarchical clustering algorithm will be of the same order for each of the several integration methods and for text-only and link-only clustering. This amounts to at least $O\big(n^2 \cdot log(n)\big)$ and typically $O(n^2)$ or even worse for a standard algorithm, with $n$ the number of documents. Without using techniques such as parallel programming, memory mapping, and without advanced clustering algorithms that balance the trade-off between space and time complexity, the number of documents that can be clustered by a standard hierarchical clustering procedure in combination with *Fisher*'s inverse chi-square method (dense distance matrix) is about 44 000 on a machine with 16 GB of RAM.

### Linkage method

Agglomerative hierarchical clustering can use various strategies to decide which documents or clusters to merge in each iteration step of the algorithm. As already mentioned in Section 1.3, *single linkage* defines the distance between two clusters as the smallest distance between any two points (one from each cluster). *Complete linkage* considers the maximal distance between any two points from both clusters. The more advanced UPGMA (unweighted pair group method using arithmetic averaging), also referred to as *group average*, calculates the distance between clusters as the weighted average of all mutual distances between objects from both clusters. The result is *unweighted* given the equal contribution of each distance. Other linkage methods exist as well [133], but have not been considered in this dissertation. For single linkage, complete linkage, and UPGMA, in each iteration those documents or clusters with the smallest distance are merged. In *Ward*'s method, those objects are grouped such that the increase in within-cluster error sum of squares over all clusters is minimized [267, 133, 146].

The clustering methods that have been experimented in the previous two sections can hence apply different linkage methods. For *Ward*'s method, the distance matrix is expected to be Euclidean, which means that all distances can be embedded in a Euclidean space. This property can be checked by looking at the eigenvalues of the distance matrix. Negative eigenvalues indicate that the distances can not completely be represented in Euclidean space. The hybrid distance matrix obtained by *Fisher*'s inverse chi-square method does not contain Euclidean distances. Hence, in strict sense, another linkage method should be used. However, in each experiment we have observed that *Ward*'s method yet outperformed other linkage methods. Figure 4.15 contrasts the performance of *Ward*'s method, UPGMA, and complete link for hybrid clustering with *Fisher*'s inverse chi-square method. *Ward*'s method clearly provides the best results, despite the non-Euclidean input matrix. Only in Figure 4.15(c), for 2, 3 or 4 clusters UPGMA does better than *Ward*. Complete link is the worst method. An additional reason why we used *Ward* instead of other methods is comparability with the text-based clustering of LIS which was discussed in Section 2.5.

*Ward*'s method is the merging criterion of choice for text-based and link-based clustering as well (see Figure 4.16). However, distance matrices derived from bibliographic coupling are usually non-Euclidean because of the use of cosine similarity (we have indeed observed negative eigenvalues). For text, we applied the cosine similarity measure because it has proven to be very effective in text mining and information retrieval, but also because of comparability with bibliographic coupling. Other authors have also opted for the combination of cosine similarity or *Pearson* correlation with *Ward*'s method, among others *Morris et al.* [194, 195] and *Leydesdorff* [164] (p. 165).

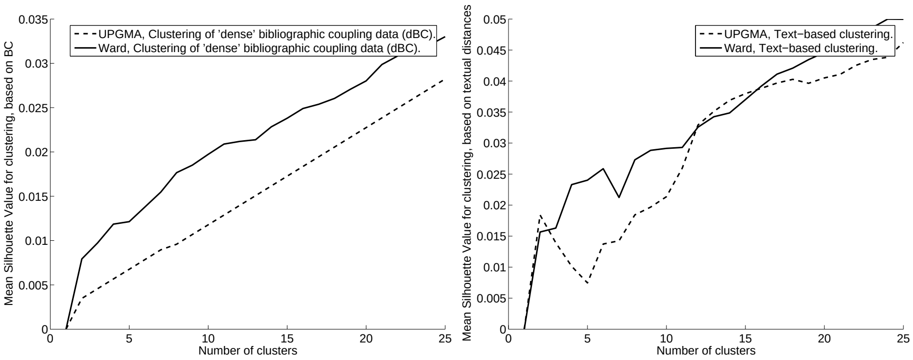


(a)

(b)

(c)

**Figure 4.15:** Silhouette curves with mean Silhouette coefficient for clustering solutions of 2 up to 25 clusters, for integrated clustering with *Fisher*'s inverse chi-square method. As shown in the legend, each plot contains three different curves for three different linkage methods: *Ward*'s method, UPGMA, and complete linkage. Silhouette values are based on distances calculated from **(a)** the complement of bibliographic coupling, **(b)** the complement of textual similarities, and **(c)** the complement of cosine similarities calculated on concatenated matrices with text (weighted by IDF) and cited references. Although the integrated distance matrix is not Euclidean, *Ward*'s method in general obtains the best results.



**(a)** **(b)**

**Figure 4.16:** Silhouette curves with mean Silhouette coefficient for clustering solutions of 2 up to 25 clusters, for bibliographic coupling **(a)** and text-based clustering **(b)**. Each plot contains two different curves for two different linkage methods: UPGMA and *Ward*'s method. Silhouette values are based on distances calculated from **(a)** the complement of bibliographic coupling and **(b)** the complement of textual similarities. Although the cosine similarity measure does not necessarily produce Euclidean distance matrices, *Ward*'s method outperforms UPGMA in both cases.

### Random Indexing results

As an aside, in Figure 4.17 we give a promising, indicative result for RI integration (see Sections 2.2.4 and 4.3.3). Weighted data integration by Random Indexing seems to do better than a linear combination of mutual distances calculated from two distinct RIs. Moreover, for less than 13 clusters the integrated random index often provides better results than *Fisher*'s inverse chi-square method and original linear combination. However, it should be noted that the integration weights for linco and *Fisher*'s inverse chi-square method were naively set to 0.5 in this experiment, whereas for the integrated RI the weight was calculated as explained in Section 4.3.3.

Although RI is a promising methodology, not much literature has been devoted to it yet and more research needs to be conducted, for instance, to determine the performance and a necessary minimal dimensionality in case of huge data sets. For example, in another, larger document set, results proved less favorable for RI integration. The performance in a classification setting should be assessed in detail as well. The outcome of this limited full-text experiment is positive, but because of the limited scope we can not provide conclusive answers. Nevertheless, we do remain cautiously optimistic about RI and data integration via RI.

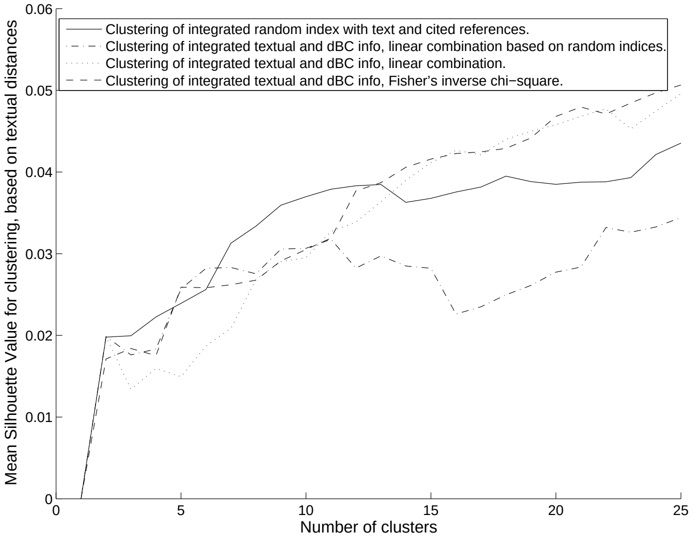


**Figure 4.17:** Silhouette curves with mean Silhouette coefficient for clustering solutions of 2 up to 25 clusters for integrated random index clustering; integrated clustering by linear combination of similarities based on a text-based RI and a citation-based RI; integrated clustering by linear combination of original document similarities; and *Fisher*'s inverse chi-square method. Silhouette values are based on distances calculated as the complement of textual similarities.

### Hybrid mapping by *Fisher*'s inverse chi-square method

Figure 4.18 presents the cognitive structure of LIS as a term network consisting of, for each of 5 clusters, the best 20 stemmed terms or phrases from titles or abstracts according to mean TF-IDF scores. We have labeled the clusters based on their most significant terms and most representative publications. In order to determine these papers, we looked at the largest cosine similarities to the mean cluster profile (centroid). Besides the labels, the two medoid papers closest to the corresponding centroids are presented in Table 4.5. Three large and two smaller clusters can be distinguished. Publications in the three larger classes are concerned with rather traditional subdisciplines of the LIS field, particularly, with *Information Retrieval* (IR), *Bibliometrics* and with what we called *'Social aspects'*. The latter term is probably not the best description but it clearly refers to the fact that many of the papers in this cluster deal with user and community relevant questions, their composition or special demands. The two smaller classes represent relatively new and emerging topics in LIS, namely, *Patent analysis* and *Webometrics*. Hence, the hybrid clustering result contains the same topics as found by the text-based clustering of Section 2.5.4, except for the merger of the two Bibliometrics clusters. The medoid for the IR cluster has not changed and the medoid of the former Bibliometrics1 cluster is also listed for the new merged Bibliometrics cluster. Figure 4.18 also visualizes the interconnections between clusters. Clusters 3 and 5 are connected through the science-technology interface as represented among others by national science and technology indicators, patent citations, industry research, patenting universities and inventor-author coactivity.

Interdisciplinary research in the intersection of science and technology—here represented by the stem *nanotechnolog*—is also one of the bridges between the two paper sets. Citations and their equivalents on the Web (in-links/out-links) form the important connection between the Bibliometrics cluster and the Webometrics cluster, which, in turn, is strongly linked to the general/Social cluster through the Web use. Finally, the stem *queri* connects Webometrics with IR. Here, *search engin*, *crawler* and *algorithm* form a strong interface.

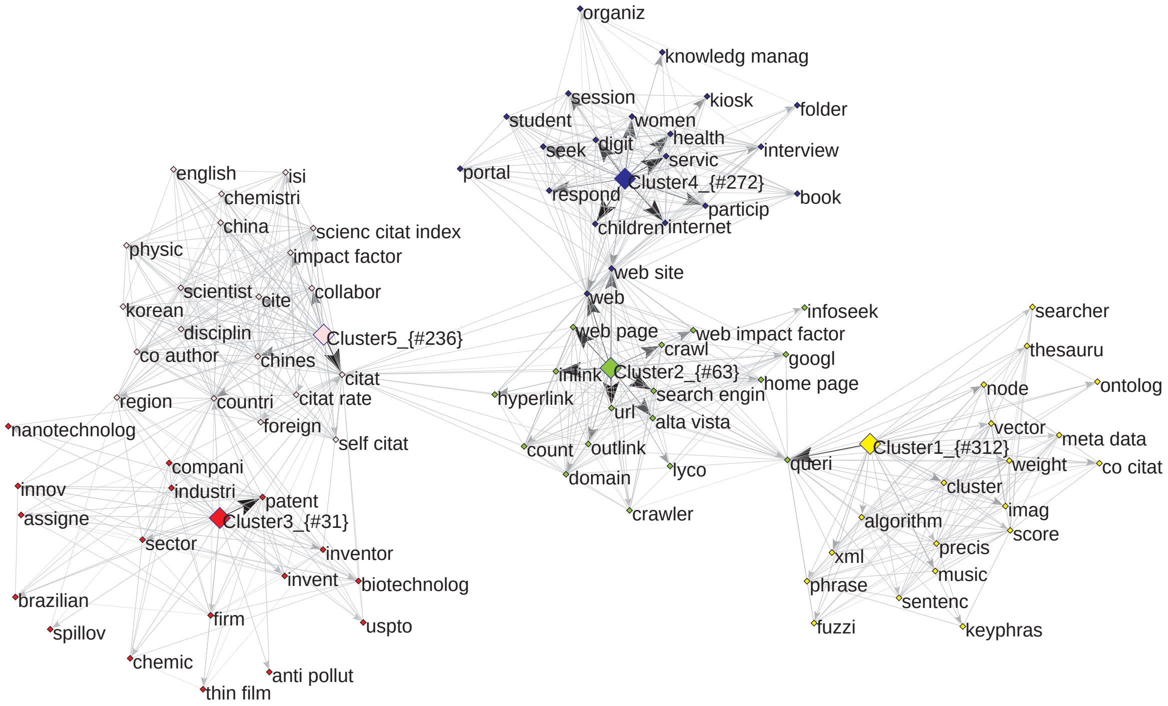


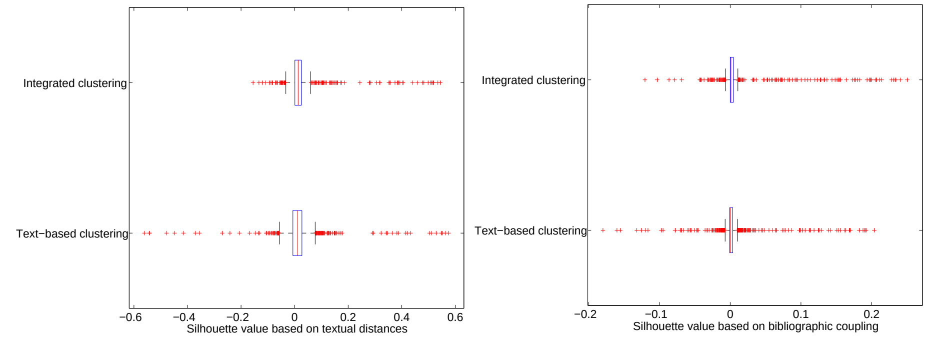
**Figure 4.18:** Term networks with for each of five LIS clusters the best 20 stemmed terms or phrases from titles or abstracts according to mean TF-IDF scores. Each cluster has its own 'central node', represented as a diamond, which also indicates the number of members. Each central node points to the best 20 terms for the cluster. When a term is among the best 20 for more than one cluster, it is only repeated once but connected to all corresponding cluster nodes. The gray level and thickness of an arc reflect the importance of a word for a cluster. Two terms are connected if both occur next to each other in one or more papers of the same cluster (only considering important words); the more co-occurrences, the closer the terms. *Pajek* was used for visualization [15].

**Table 4.5:** For each of 5 clusters the two medoid papers, which are the publications with largest cosine similarity to the mean cluster profile.

| **Cluster 1. Information Retrieval (312 documents)** |
|---|
| Schlieder, T. & Meuss, H. (2002). Querying and ranking XML documents. *JASIST*, 53, 489–503. |
| Huang, C. K., Chien, L. F., & Oyang, Y. J. (2003). Relevant term suggestion in interactive Web search based on contextual information in query session logs. *JASIST*, 54, 638–649. |
| **Cluster 2. Webometrics (63 documents)** |
| Thelwall, M. & Harries, G. (2004). Do the Web sites of higher rated scholars have significantly more online impact? *JASIST*, 55, 149–159. |
| Thelwall, M. & Harries, G. (2003). The connection between the research of a university and counts of links to its web pages: An investigation based upon a classification of the relationships of pages to the research of the host university. *JASIST*, 54, 594–602. |
| **Cluster 3. Patent (31 documents)** |
| Bhattacharya, S. (2004). Mapping inventive activity and technological change through patent analysis: A case study of India and China. *Scientometrics*, 61, 361–381. |
| Meyer, M., Sinilainen, T., & Utecht, J. T. (2003). Towards hybrid Triple Helix indicators: A study of university-related patents and a survey of academic inventors. *Scientometrics*, 58, 321–350. |
| **Cluster 4. Social (272 documents)** |
| Hargittai, E. (2002). Beyond logs and surveys: In-depth measures of people's Web use skills. *JASIST*, 53, 1239–1244. |
| Marchionini, G. (2002). Co-evolution of user and organizational interfaces: A longitudinal case study of WWW dissemination of national statistics. *JASIST*, 53, 1192–1209. |
| **Cluster 5. Bibliometrics (236 documents)** |
| Al Qallaf, C. L. (2003). Citation patterns in the Kuwaiti journal Medical Principles and Practice: The first 12 years, 1989-2000. *Scientometrics*, 56, 369–382. |
| Persson, O., Glänzel, W., & Danell, R. (2004). Inflationary bibliometric values: The role of scientific collaboration and the need for relative indicators in evaluative studies. *Scientometrics*, 60, 421–432. |

The question arises of what the added value is of the combination of the two methods, the text-based and the bibliometrics aided approach. From the technical viewpoint, the appropriate choice of weight $\lambda$, automatically determined equal to 0.43 by the method described in Section 4.3.2, results in a somewhat better evaluation of clustering. If we compare the text-only approach and the hybrid solution, we clearly see a measurable improvement by the combination.

Figure 4.19 presents box and whisker plots of the Silhouette values of all 914 documents for the hybrid clustering solution ('Integrated clustering', on top) and for the text-based clustering (lower part). In (a), the Silhouette values are based on textual distances, whereas the complement of bibliographic coupling was used as the distance measure in (b). Although the Wilcoxon signed rank test does not reject equal means for (a), it is clear that less documents have a highly negative Silhouette value in the hybrid case. As explained above, evaluation with text-based Silhouette values is in favor of text-only clustering, so it is worth mentioning that hybrid clustering is actually not inferior. On the other hand, as expected, the Wilcoxon signed rank test does reject equal means in case of bibliographic coupling, to the advantage of hybrid clustering ($p = 2.52 \cdot 10^{-4}$).

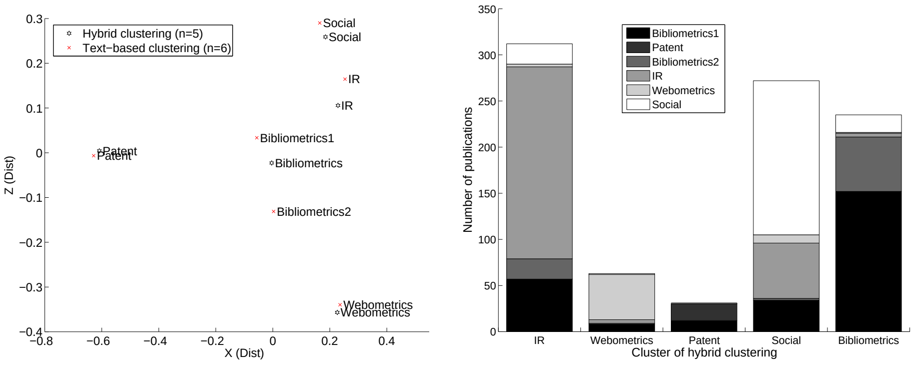


(a) (b)

**Figure 4.19:** Box and whisker plots of the Silhouette values of all 914 documents for the hybrid clustering solution ('Integrated clustering', on top) and for the text-based clustering (lower part). The extent of a box indicates the interquartile range. In **(a)**, the Silhouette values are based on textual distances, whereas the complement of bibliographic coupling was used as distance measure in **(b)**. Although the Wilcoxon signed rank test does not reject equal means for **(a)**, it is clear that less documents have a highly negative Silhouette value in the hybrid case. As expected, the Wilcoxon signed rank test does reject equal means in case of bibliographic coupling **(b)**, to the advantage of hybrid clustering ($p = 2.52 \cdot 10^{-4}$).

In Figure 4.20(a), the centroids of the six clusters of the text-only approach are compared with those of the five clusters of the hybrid method. Certain shifts around the merged Bibliometrics cluster can be observed. This change, however, also concerns other clusters. The centroid of the new merged Bibliometrics cluster is located nicely between the former two centroids. The Patent cluster is still the most distant one and has grown from 19 to 31 papers. Thus, some patent-related publications had been put in one of the bibliometrics clusters by the text-based algorithm, whereas the incorporation of citations has led to a more clear demarcation between patent and bibliometric studies. In the text-only setting, the Patent cluster was closer to Bibliometrics1 than to Bibliometrics2 [137] and Bibliometrics1 was even combined with Patent before being combined with Bibliometrics2 (see Figure 2.15 on page 70). The present hybrid results correspond more to our intuition: there is only one Bibliometrics cluster and the Patent cluster is only merged with bibliometrics in a later stage. In Section 2.5.4 it has already been stipulated that the Patent cluster could be clearly separated from the rest of LIS and that there was no clear border between both former bibliometrics clusters.

This leads immediately to the question of 'migrated' papers, that is, more than a quarter of the papers were assigned to a different cluster according to the hybrid scheme. The *Rand* index comparing the textual and hybrid clustering solution equals 0.75, while the *adjusted Rand* index was 0.337. Figure 4.20(b) visualizes the overlap between hybrid and text-based clusters.

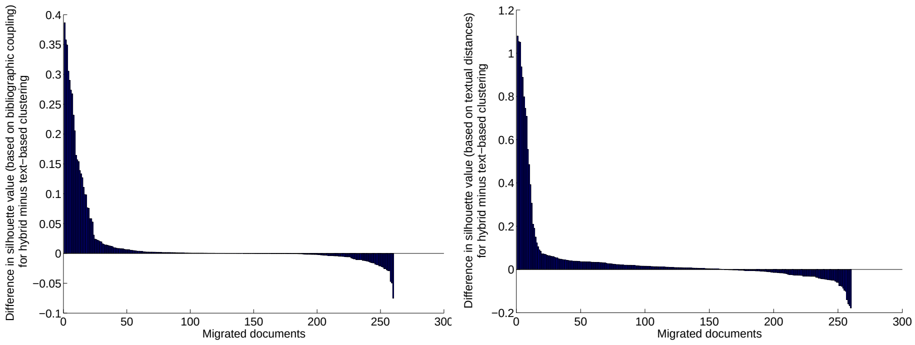


(a)

(b)

**Figure 4.20: (a)**. Multidimensional scaling (MDS) plot comparing the cluster centers (centroids) of the six clusters found by the text-based clustering, with the five cluster centers of the hybrid clustering. The centroid of the new merged Bibliometrics cluster is located nicely between the former two centroids. The Patent cluster is still the most distant one. The present hybrid results correspond more to our intuition: there is only one Bibliometrics cluster. In Section 2.5.4 it has already been stipulated that the Patent cluster could be clearly separated from the rest of LIS and that there was no clear border between both former bibliometrics clusters. **(b)**. The overlap of each of 5 clusters determined by hybrid clustering with *Fisher*'s inverse chi-square method, with the text-based clusters. More than a quarter of the papers were assigned to a different cluster according to the hybrid scheme.

By checking paper assignment to clusters according to the two methods manually, we found that many of these 'migrated' papers were originally misplaced in the text-based approach, like the 'new' patent papers discussed above. Nonetheless, incorrectly assigned papers still occur in the combined classification, too, but this is probably unavoidable when using the agglomerative hierarchical clustering algorithm. One of the disadvantages is that wrong choices (merges) that are made by the algorithm in an early stage can never be repaired [146]. To distinguish the good from the bad migrations, we sorted all migrated documents according to descending difference in text-based Silhouette values for hybrid minus text-based clustering, as visualized in Figure 4.21. The prevalence of positive values indicates that there are more correct than spurious migrations.



(a) (b)

**Figure 4.21:** For all migrated documents the difference in Silhouette value for the hybrid clustering minus the text-only clustering, sorted in descending order. The prevalence of positive values indicates that there are more correct than spurious migrations. **(a)**. Silhouette values are based on the complement of bibliographic coupling. **(b)**. Silhouette values are based on text.

A few of many examples of good migrations are the following. A paper of *Nie* about 'Query expansion and query translation as logical inference' migrated from the text-based Bibliometrics1 cluster to the hybrid IR cluster (Appendix B: Nie, 2003). Next, the paper of *Faba-Perez et al.* about "Sitation' distributions and Bradford's law in a closed Web space' was put in the Webometrics cluster instead of Bibliometrics1 (Appendix B: Faba-Perez et al., 2003). Finally, 'Knowledge integration in virtual teams: The potential role of KMS' by *Alavi* & *Tiwana* changed from Bibliometrics1 to the more Social cluster (Appendix B: Alavi et al., 2002). On the other hand, less obvious migrations could also be observed. For example, 'Empirical evidence of self-organization?' (Appendix B: van den Besselaar, 2003) moved from Bibliometrics1 to IR, and the same goes for a paper by *Leydesdorff*, 'Indicators of structural change in the dynamics of science: Entropy statistics of the SCI Journal Citation Reports' (Appendix B: Leydesdorff, 2002).

Figure 4.22 visualizes the effect of migration after merging the two bibliometrics clusters through combining the text-only with the citation-based method. 24 new papers appear in the very center of the new Bibliometrics cluster as consequence of migration. Other documents of the former Bibliometrics1 and

Bibliometrics2 clusters are not included in the new one, among which the patent-related publications.

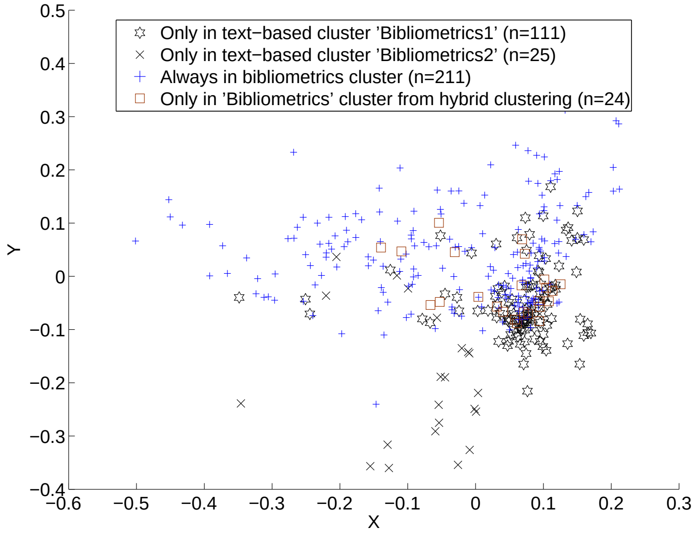


**Figure 4.22:** Multidimensional scaling (MDS) plot only considering documents in the two bibliometrics clusters of the text-based solution and documents in the bibliometrics cluster of the hybrid clustering. A distinction is made between documents that were only in one outcome assigned to a bibliometrics related cluster, and documents that were consistently assigned to bibliometrics. 24 new papers appear in the very center of the new Bibliometrics cluster as consequence of migration. Other documents of the former Bibliometrics1 and Bibliometrics2 clusters are not included in the new one, among which the patent-related publications.

There is still another reason for the 'success' of the hybrid or bibliometrics aided classification beyond any technical considerations. Any lexical (text-based) approach is usually based on rather rich vocabularies and peculiarities of natural language. The result is, according to our observations, a rather 'smooth' or gradual transition between what is related and what is not. The relationship between documents is, therefore, somewhat fuzzy and not always reliable. On the other hand, if strict citation-based criteria are applied, that is, if non-periodical references (i.e., references to non-serial literature such as books or reports) and occasional coupling links are removed, the resulting *citations-by-document* matrix becomes extremely sparse. In this case, rejection of relationship tends to be unreliable. The modification to the original formula for bibliographic coupling by adding a constant 0.01 to the numerator (see Section 4.3.2) helps smoothing the 'singularity', but is not able to overcome it. This might explain the low efficiency of coupling- (and co-citation-) based clustering techniques. The combination of the two worlds helps to improve the reliability of relationship and therefore of the clustering algorithm as well.

### 4.5.5 Concluding remarks

Contrary to the full-text-based clustering for which the optimal number of clusters was 6 (see Section 2.5), the field of library and information science was subdivided into 5 classes by the hybrid clustering method based on *Fisher*'s inverse chi-square. The two bibliometrics clusters from the text-only clustering to a large extent merged into a single cluster. The optimal number of clusters is still a difficult issue and depends on the adopted validation and the applied similarity measures, as well as on the input data, be it mere text, just citations or a combination. For the integrated clustering by linear combination even 8 clusters could be perceived as the best choice. Although this algorithm is a very attractive, easy and scalable integration method that was not inferior to *Fisher*'s inverse chi-square method in the previous section, it was outperformed in the present setting with regard to the Silhouette coefficient and stability. By integrating text and citations, *Fisher*'s inverse chi-square method again performed better than the pure text-based method. We compared the six clusters of the text-only approach with the five clusters of the hybrid method. Quite some papers had 'migrated' to another cluster. Many of these were originally misplaced in the text-based approach, so we clearly observed an improvement by the combination. On the other hand, incorrectly assigned papers still occurred in the combined classification. This is probably unavoidable when adopting the agglomerative hierarchical clustering algorithm. We think that, in order to gain even better performance, a transition should be made towards fuzzy clustering algorithms. A promising result for integration by Random Indexing was also provided, but should be regarded only illustrative as more experiments should be conducted.

## 4.6 Bibliometric retrieval

A combined methodology consisting of textual and bibliometric components can also be applied within the framework of Information Retrieval (IR). Particularly, one of the most important endeavors in most bibliometric domain studies is the delineation of research fields, especially when dealing with emerging or complex interdisciplinary fields such as nanoscience/-technology, biotechnology or bioinformatics.

Although the field of IR has a long history with remarkable accomplishments, often users still have to resort to extensive search strings to filter bibliographic databases or to compose a set of publications representing a scientific domain. For instance, to delineate the nanotechnology domain, a search strategy encompassing a complete page full of query terms was used by *Glänzel et al.* [103] (p. 14 and 17). Furthermore, traditional textual query-based retrieval only works if one knows what to look for and is thus not effective when one wants, for instance, to detect or delineate new fields. Indeed, the term *bioinformatics* was only introduced years after the actual germination of the field.

In the following, we describe the application of a subject-delineation strategy for demarcating core literature in the bioinformatics field[7] [102, 139]. This bibliometric retrieval (BR) strategy should be understood as the extension of traditional information retrieval by adding citation-based components and is geared to the delineation of subject fields. The general BR model has been developed for the delineation of stem-cell research by *Glänzel et al.* [107]. *Zitt* and *Bassecoulard* have used a related strategy [280].

## Delineation of the research field bioinformatics

A bibliometric study of the bioinformatics field by *Patra* and *Mishra* was based on MeSH terms from MEDLINE and adopted a rather liberal delineation strategy that was tailored towards maximal recall. It collected all records containing keywords such as *Bioinformatics* or *Genomics* in any field, including journal information and author address [216]. This rather broad coverage of the field included almost double the number of articles that are retrieved by our bibliometrics-aided retrieval strategy. The latter is more stringent in the sense that the aim is a very strict interpretation of the field by collecting a reliable set with *core* bioinformatics literature, while minimizing the amount of included noise documents. The use of WoS compared to MEDLINE also means a less broad coverage of bioinformatics journals. On the other hand, citation-based components can procure documents that might be overlooked by mere text-based techniques.

Our strategy has a strong bibliometric component and is based on bibliographic coupling ('horizontally' searching at the same time level) as well as on references and citations ('vertically' searching in the past and future, respectively).

The data set is extracted from the Web of Science (WoS) Edition of the *Science Citation Index Expanded*$^{TM}$ (SCIE) of *Thomson Scientific* (Philadelphia, PA, USA), publication years 1981–2004. Another data source has been used, namely the subject headings annotated to MEDLINE records that were matched with the *ISI* WoS data set. These MeSH terms are also used in part for validation and to refine the retrieval made in the SCIE database.

Our *bibliometric retrieval* strategy (BR) logically consists of two parts which, in turn, can have several components each. The first part comprises *unconditional criteria*, which can include, for example, a keyword search strategy ($UC_3$) and *core* journals covered by the Web of Science ($UC_1$) and the MEDLINE database ($UC_2$), i.e., journals that almost solely publish bioinformatics papers.

[7] This bibliometric retrieval strategy was also presented at the *9th International Conference on Science and Technology Indicators* in Leuven, Belgium, September 7-9, 2006.

- $UC_1$: Journal in WoS = Bioinformatics (formerly Computer Applications in the Biosciences), Journal of Computational Biology, Briefings in Bioinformatics, BMC Bioinformatics.[8]
- $UC_2$: Journal in MEDLINE = In Silico Biology, PSB On-line Proceedings, Applied Bioinformatics, PLoS Computational Biology.
- $UC_3$: Keywords in title = bioinformatics, computational biolog*, systems biology.

In other words, all papers meeting at least one of the criteria $UC_1$, $UC_2$ or $UC_3$ are deemed relevant. Furthermore, this set can be extended with result sets obeying so-called *conditional criteria* ($CC_1$ and $CC_2$), each of which results in related but not necessarily core literature. In particular, the conditional criteria comprise conditions for reference ($CC_1$) and citation ($CC_2$) links.

- $CC_1$: Publications cited by $UC1$.
- $CC_2$: Publications citing $UC1$.

All papers meeting at least one of the criteria $CC_1$ and $CC_2$ are considered potentially relevant, but might not directly be concerned with bioinformatics. Only that part of literature which meets further restrictive criteria will be considered truly relevant. In order to reduce or even exclude noise, the conditional criteria are made subordinate to thresholds $T_i$ for relevancy.

The bibliometrics aided retrieval strategy (BR) for identifying relevant papers in bioinformatics can thus be obtained by the following logical combination:

$$BR_{bioi} = \big[UC_1 \vee UC_2 \vee UC_3\big] \vee \big[(CC_1 \wedge T_1) \vee (CC_2 \wedge T_2)\big] \tag{4.8}$$

The BR strategy can be fine-tuned by extending or reducing the sets of criteria and by adjusting the thresholds for bibliometric components such as number or share of references and strength of citation, reference and coupling links. An extra conditional criterion might be, for example, the occurrence of keywords such as *bioinformatics* in the address field or 'reference string'.

Table 4.6 presents the effect of adjusting the strength of citation/reference links on the number of retrieved documents. We used four different thresholds based on the absolute number of citations and references. In addition, results of the first unconditional criterion as well as the disjunctive combination of $UC_1$ with the first conditional criterion are shown. Since the thresholds $T2 = T3 < 3$ still resulted in perceptible noise, we decided to use $T2 = T3 = 3$ for the study. All in all 7655 records were thus retrieved, among which there were 7401 articles, notes and reviews. There were 67 140 citations between these records. If this set would have been extended with all external citing and cited records, it would yield a set of 261 221 records.

Each record retrieved from the Web of Science was also matched against the MEDLINE database in order to retrieve Medical Subject Headings (MeSH). In

[8] Other journals or proceedings that were considered part of the core were not available in WoS at the time of retrieving, among others, *PLoS Computational Biology*, *Applied Bioinformatics*, *In Silico Biology*, *PSB On-Line Proceedings*, *Online Journal of Bioinformatics*, and *IEEE Transactions on Computational Biology and Bioinformatics*.

**Table 4.6:** Number of records retrieved for different combinations of criteria.

| Strategy | Threshold $T_1 = T_2$ | Records retrieved |
|---|---|---|
| $UC_1$ | / | 3386 |
| $UC_1 \vee CC_1$ | / | 9620 |
| BR | 1 | 41 995 |
| BR | 2 | 13 239 |
| BR | 3 | 7655 |
| BR | 4 | 5470 |

short, a search key was generated based on publication year, volume, pagination and first characters of author names and title. The '*Levenshtein* string edit distance' was additionally used on the title string as a second validation measure to detect spurious matches when multiple hits were found or when no direct key match could be identified. In the latter case, when a search for the exact title failed as well, the standard cosine measure in the Vector Space IR Model (see Section 2.1.2) was used to match the title and abstract from the WoS with the entire MEDLINE database. From top ranking documents, the record with lowest *Levenshtein* distance for the title was considered a potential match. In the lack of sufficient evidence, it was still withheld for manual verification. All in all, 6272 direct key matches were found for the 7401 articles, notes, and reviews in our bioinformatics set. Out of 1127 records that could not be matched immediately by their search key, 533 could be matched based on the title.

## 4.7 Concluding remarks

The complex nature of mapping various aspects of knowledge motivates approaches that integrate different viewpoints on the same data. We proposed various schemes to integrate textual and bibliometric methods and we were able to improve on both existing approaches. As a conclusion, we believe that such hybrid methodologies are valuable tools to facilitate endeavors in mapping fields of science and technology. Moreover, the combination of text-based and bibliometric components in *bibliometric retrieval* could also be used to improve the complex delineation of interdisciplinary research fields like bioinformatics, thus opening new perspectives in research evaluation, too.

A serial combination of full-text mining and bibliometric techniques was applied in the mapping of the 2003 volume of *Scientometrics*. It was clear that clusters found through application of text mining provided additional information that could be used to extend, improve, and explain structures found by bibliometric methods, and vice-versa. Reference-based citation measures could help to fine-structure clusters determined by text-based analysis and the combination of text-mining and bibliometric techniques proved an appropriate tool to unravel cognitive structure.

The performance of unsupervised clustering and classification of scientific papers could significantly be improved by profoundly integrating textual content of titles and abstracts with cited references. In general, text-only information was much more powerful than cited references alone and dimensionality reduction by SVD could greatly improve results, especially when applied to textual information. However, the best outcome was obtained by integration.

Next to an approach to integrate data based on Random Indexing, we also devised an integration method in which pairwise distances between documents are converted to $p$-values compared to randomized data sets, and in which *Fisher*'s inverse chi-square method is then used to combine the $p$-values from all information sources. This method can handle distances stemming from metrics with different distributional characteristics and avoids domination of any specific data source. For a correct application of *Fisher*'s inverse chi-square method we introduced a slight, rank-preserving modification to the formula for bibliographic coupling. The integration method has shown to significantly outperform corresponding text-only and link-only methods, as well as a mere concatenation of vectors. In the experiments of Section 4.4, *Fisher*'s inverse chi-square method, however, did not significantly outperform corresponding linear combinations when SVD had been applied. Given the higher complexity of implementing *Fisher*'s inverse chi-square method and a reduced scalability, a carefully chosen weighted linear combination might be the preferred solution for integrating textual and citation information if LSI is used. Nevertheless, this latter algorithm, which offers a very attractive, easy and scalable integration method, was yet outperformed by *Fisher*'s inverse chi-square method with regard to Silhouette coefficient and stability for a hybrid mapping of the LIS field. Furthermore, the inverse chi-square method is generic and can be used to incorporate distances with highly dissimilar distributional characteristics, such as textual distances and distances based on bibliometric features like *mean reference age* and *share of serials*.

We compared six clusters found by text-only clustering of LIS with five clusters of the hybrid method in which two bibliometrics clusters were to a large extent merged into a single cluster. Quite some papers that were originally misplaced in the text-based approach had migrated to another cluster, so we clearly observed an improvement by the combination. On the other hand, incorrectly assigned papers still occurred in the combined classification. This is probably unavoidable when adopting the agglomerative hierarchical clustering algorithm. We think that, in order to gain even better performance, a transition should be made towards fuzzy clustering algorithms. In the meantime, spurious assignments of documents to clusters can be detected by validation measures such as the Silhouette coefficient, and fuzziness might to some extent be mimicked by taking into account document similarities to all cluster centroids.

# Chapter 5

# Dynamic hybrid mapping of bioinformatics

In Section 4.6, we have introduced *bibliometric retrieval*, a subject delineation strategy conceived by Glänzel *et al.* [101], and we have applied it to delineate the bioinformatics field. In Section 3.6, the resulting set of bioinformatics publications was analyzed from a bibliometric point of view, particularly, growth dynamics, international and author collaboration, patterns of national publication activity, and citation impact.

In this chapter, the bioinformatics field is further analyzed, focusing on the cognitive structure as perceived by our hybrid clustering algorithm based on *Fisher*'s inverse chi-square method, that has been introduced in Section 4.3.2. The algorithm provides an integrated analysis of both text and citation worlds. In Section 5.2, for each cluster we provide term and collaboration networks, representative publications, relative importance for the 5 most active countries, as well as citation patterns and *'naive' dynamics* of the cluster. The term *naive* refers to the fact that publication years are not considered during *'static'* clustering, but only afterwards, when clusters are already formed.

Subsequently, in Section 5.3 we introduce *dynamic hybrid clustering* for matching and tracking clusters through time, which is an important research topic in the light of dynamic document sets and emerging trend detection. The resulting *cluster chains*, their structure and evolution, and various statistics are analyzed and compared with clusters found by the *static* hybrid clustering of the whole bioinformatics set. 'Dynamic' networks allow the observation of shifts in collaboration patterns and in terminology within cluster chains. To provide an example, we zoom in on the *Systems Biology & molecular networks* chain. Next, external chain-to-chain citations are visualized for each period, and finally, the *ISI* journal Impact Factor [89] of each cluster chain is plotted through time.

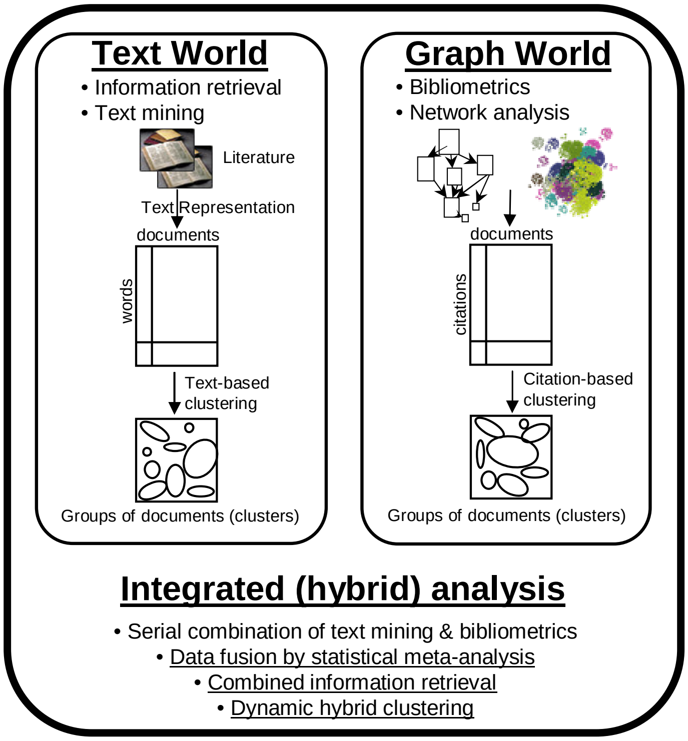

Text World
• Information retrieval
• Text mining
Literature
Text Representation
documents
words
Text-based clustering
Groups of documents (clusters)
Graph World
• Bibliometrics
• Network analysis
documents
citations
Citation-based clustering
Groups of documents (clusters)
Integrated (hybrid) analysis
• Serial combination of text mining & bibliometrics
• Data fusion by statistical meta-analysis
• Combined information retrieval
• Dynamic hybrid clustering

**Figure 5.1:** In this chapter the bioinformatics field is further analyzed, focusing on the cognitive structure as perceived by our hybrid clustering algorithm based on *Fisher*'s inverse chi-square method. The algorithm provides an integrated analysis of both text and citation worlds. Subsequently, we introduce *dynamic hybrid clustering* for matching and tracking clusters through time.

## 5.1 Material and methods

Our data set consists of 7401 bioinformatics-related *articles*, *notes*, and *reviews* (see Section 4.6). From each record we considered author and country information, the textual information present in titles and abstracts, author keywords, and the MeSH[1] fields from MEDLINE (excluding those acknowledging research funding). In addition, we collected all cited references and all citing papers.

### 5.1.1 Text analysis

For a discussion on the analysis of textual data we refer to Sections 2.1, 2.2.2 and 2.2.3. In short, all textual content was indexed and encoded in the Vector Space Model using the TF-IDF weighting scheme, and text-based similarities were calculated as the cosine of the angle between the vector representations of two papers. *Stop words*, URLs, and e-mail addresses were not taken into account during indexing and on all remaining terms from titles and abstracts the *Porter* stemmer was applied. *Dunning*'s *log*-likelihood method for detection of bigrams was followed to detect composite terms from a candidate list of MeSH descriptors, author keywords, and noun phrases identified by LT POS and LT CHUNK. Finally, the dimensionality of the term-by-document matrix was reduced from 18 163 term dimensions to the 10 factors by Latent Semantic Indexing (LSI) (see Section 2.2.3).

### 5.1.2 Citation analysis

Important and highly recognized bioinformatics papers can be identified in each subfield by analyzing the citation graph, a topic which has received ample treatment in Section 3.2. We use the link-based algorithms HITS [149] and PageRank [37] (see Sections 3.4.1 and 3.4.2), and we also consider (average) numbers of citations and the *ISI* Impact Factor [89].

### 5.1.3 Hybrid analysis

To subdivide the bioinformatics papers into clusters we used agglomerative hierarchical clustering (see Section 2.3). We determined the number of clusters to be 9 by observing the dendrogram, by looking for a local maximum in the text-based and citation-based mean Silhouette curves (see Figures 2.6 and 2.7 on pages 50 and 51), and by using the stability-based method of *Ben-Hur et al.* (see Figure 2.9 on page 53). Note that journal or author information was of course never used for clustering.

The requisite input for many clustering algorithms includes pairwise distances between all objects (scientific publications here). These distances can be based on text, on citations, or on a combination of both information sources

[1] http://www.nlm.nih.gov/mesh/, visited in January 2007.

(see Section 4.3). The performance of clustering can significantly be improved by integrating textual content with citations, as has been dilated upon in detail in Chapter 3.

For (*static* and *dynamic*) hybrid clustering of bioinformatics we used *Fisher*'s inverse chi-square method of Section 4.3.2 to integrate both textual similarity and citation information (bibliographic coupling). The dynamic clustering methodology is discussed in Section 5.3.

### 5.1.4 Dynamic term networks

For visualization we again determined for each group (cluster, cluster chain, or period) the best words or phrases according to mean TF-IDF weights (see Figure 5.4 for an example). Each group has its own 'central node', represented as a diamond, which also indicates the number of members. Each central node points to the best keywords. When a keyword is among the best for more than one group, it is only repeated once but connected to all corresponding central nodes. The gray level and thickness of an arc reflect the importance of a word for a group. Two terms are connected if both co-occur in one or more papers of the same group; the more co-occurrences, the closer the terms. *Pajek* was used for visualization [15].

## 5.2 Hybrid clustering results

Figure 5.2 depicts the dendrogram that resulted from hybrid hierarchical clustering of the bioinformatics publications, cut off at 9 clusters on the left-hand side. For each of 9 clusters, the number of publications and the best mean TF-IDF term or phrase are shown. These automatically determined labels already give a quite good impression of the contents of the clusters. As explained in Section 2.3.3, a dendrogram visualizes the (integrated textual and link-based) distances between clusters. For example, the *rna* and the *protein* clusters (#1 & #2) would be merged first when asking for eight instead of nine clusters. The *microarrai* cluster (#9) is most distinct from all other clusters and would only be merged in a final phase to form the *trivial cluster* containing the complete bioinformatics set.

After observing the contents of all clusters in detail we were able to propose a name for each cluster as given in Table 5.1 and Appendix C. In the first table, the cluster size is indicated next to the characterization by most salient author keyword, by best stem or phrase from titles and abstracts, and by best TF-IDF MeSH term. For MeSH terms, the *TF* factor was either 1 or 2 for *minor* and *major* MeSH descriptors, respectively. With 205 publications, Cluster 1 (labeled *RNA structure prediction*) is the smallest one; all other clusters have more than 600 and less than 1200 papers.

We determined representative publications by using five different methods that rank the papers in each cluster according to different criteria of importance.

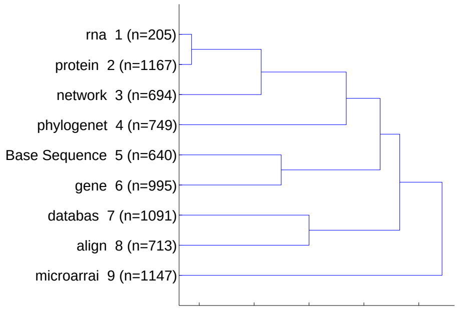


**Figure 5.2:** Dendrogram, cut off at 9 clusters on the left-hand side, for hybrid hierarchical clustering of the 7401 bioinformatics publications. For each of 9 clusters, the number of publications and the best mean TF-IDF term or phrase are shown. Table 5.1 contains the name of each of the clusters, besides the best terms from several vocabularies.

Appendix C lists for each cluster the two papers on top of each ranking: *(1) medoids*, which are the papers most similar to the mean cluster profile (the *centroid*), *(2)* documents that received most citations from within the cluster, *(3, 4)* the best two *authorities* and best *hubs* determined by the HITS algorithm, and *(5)* the papers with highest PageRank.

**Table 5.1:** The 9 clusters obtained from the hybrid clustering algorithm.

| Cluster | Name | Number of papers | Best author keyword | Best term from titles and abstracts | Best MeSH term |
|---|---|---|---|---|---|
| 1 | RNA structure prediction | 205 | rna secondary structure | RNA | Nucleic Acid Conformation |
| 2 | Protein structure prediction | 1167 | protein structure prediction | protein | Proteins/chemistry |
| 3 | Systems biology & molecular networks | 694 | bioinformatics | network | Models, Biological |
| 4 | Phylogeny & evolution | 749 | phylogeny | phylogenet | Phylogeny |
| 5 | Genome sequencing & assembly | 640 | sequencing hybridization | base sequenc | Base Sequence |
| 6 | Gene/promoter/motif prediction | 995 | gene regulation | gene | Sequence Analysis, DNA/methods |
| 7 | Molecular DBs & annotation platforms | 1091 | genome analysis | databas | Databases, Factual |
| 8 | Multiple sequence alignment | 713 | sequence alignment | align | Sequence Alignment/methods |
| 9 | Microarray analysis | 1147 | microarray | microarrai | Oligonucleotide Array Sequence Analysis/methods |
| | Complete bioinformatics set | 7401 | bioinformatics | protein | Algorithms |

Figures 5.3 and 5.4 depict the cognitive structure of bioinformatics. For each cluster, Figure 5.3 indicates the best 10 TF-IDF MeSH terms. Figure 5.4(a) presents a term network consisting of the best 10 terms or phrases from titles and abstracts, according to mean TF-IDF scores, whereas in (b) the best 10 author keywords are shown.

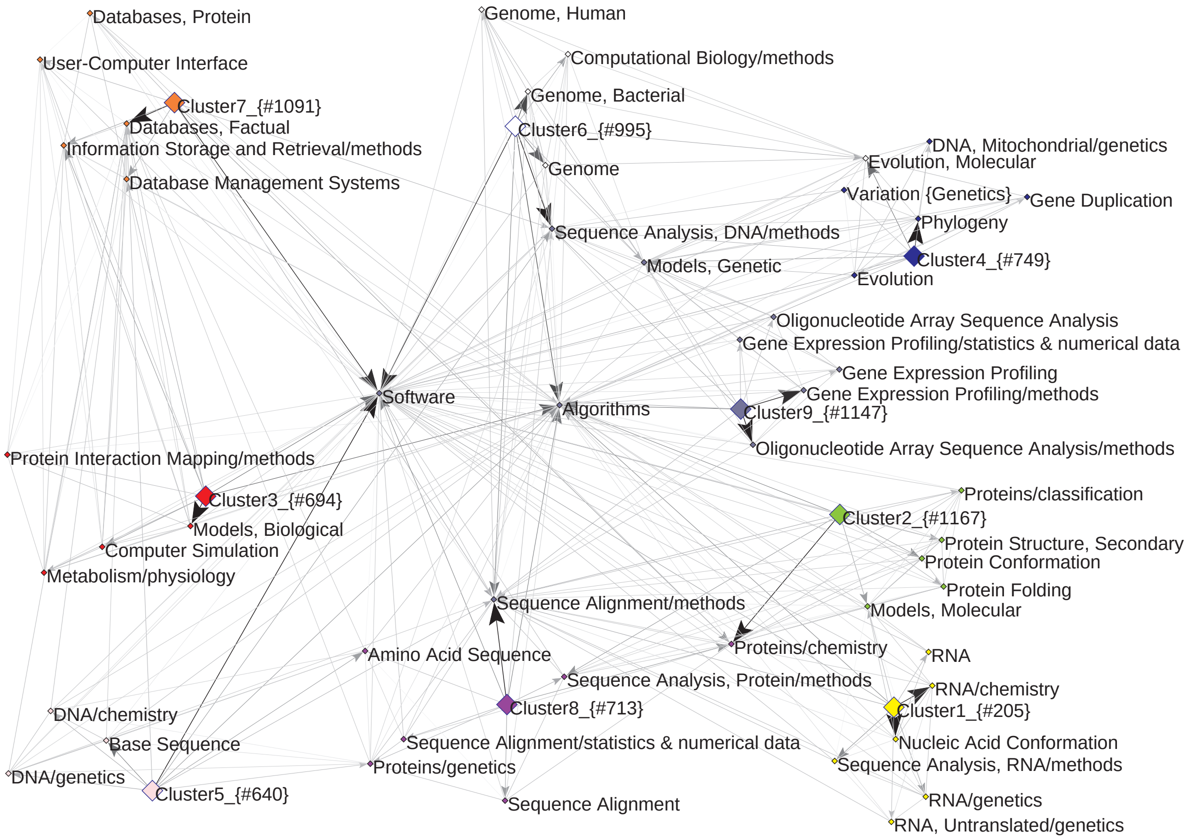
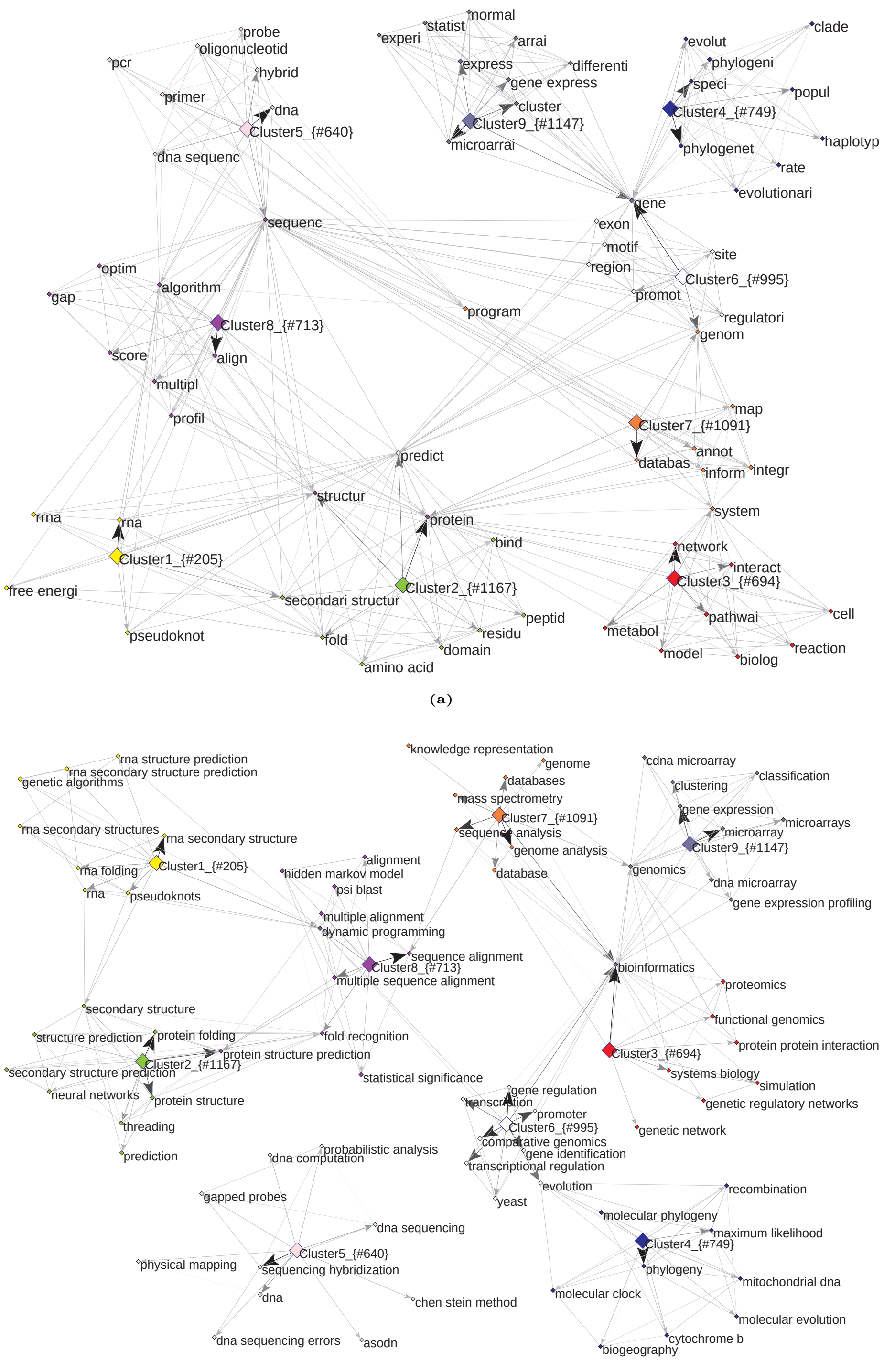


**Figure 5.3:** Term network with for each of nine clusters the best 10 TF-IDF MeSH terms.



(a)

(b)

**Figure 5.4:** Term networks with for each of nine clusters: **(a)** the best 10 terms or phrases from titles and abstracts, and **(b)** the best 10 author keywords, according to mean TF-IDF scores.

### 5.2.1 Cluster representation of the 5 most active countries

The breakdown of national publication output by clusters does not allow any reliable quantitative analysis for most of the 30 selected countries because of the often too small publication sets. We restrict the analysis to the five leading countries, particularly, the USA, UK, Germany, France, and Japan. Their share in the nine individual clusters is shown in Figure 5.5, which was generated by using scripts made available by the *Steunpunt O&O Indicatoren*[2] [102].

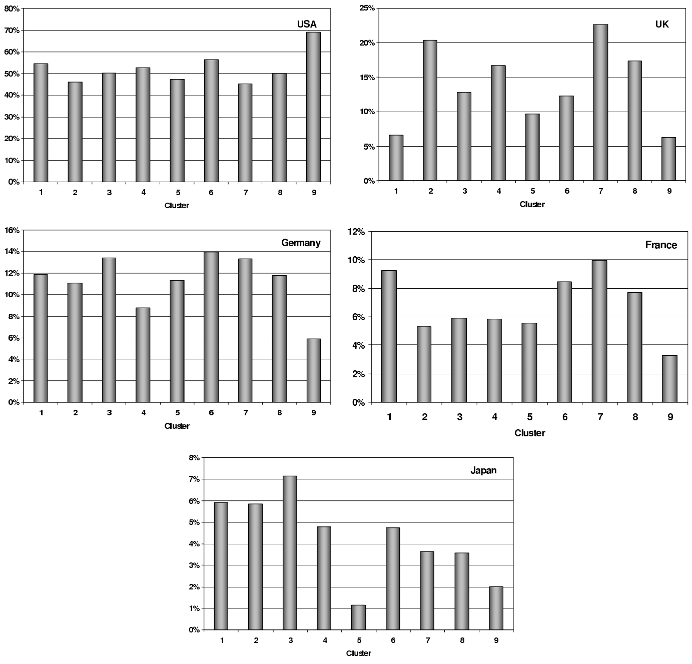


**Figure 5.5:** National representation of the five most active countries by clusters.

The US with a share of about 45% and more are predominant in most subdisciplines. Above all, Cluster 9 (*Microarray analysis*) is dominated by the USA with 70% of all papers. Germany has a well-balanced high share in all clusters as well, except for Cluster 9. The other three countries reflect a rather heterogeneous picture; the British contribution to clusters 2 (*Protein structure prediction*) and 6 (*Gene/promoter/motif prediction*) is worth mentioning, however, the contribution to Cluster 1 (*RNA structure prediction*) and 9 (*Microarray analysis*) is rather small. The situation in France is similar: the strong contribution to Cluster 1 and 7 (*Molecular DBs & annotation platforms*) is contrasted by a low share in Cluster 9. The extremes in the Japanese publication output can be found in Cluster 3 (*Systems Biology & molecular networks*) with 7% of the world total and Cluster 5 (*Genome sequencing & assembly*) with 1%.

[2] Steunpunt O&O Indicatoren, Katholieke Universiteit Leuven, Dekenstraat 2, B-3000 Leuven, Belgium.

### 5.2.2 Author collaboration

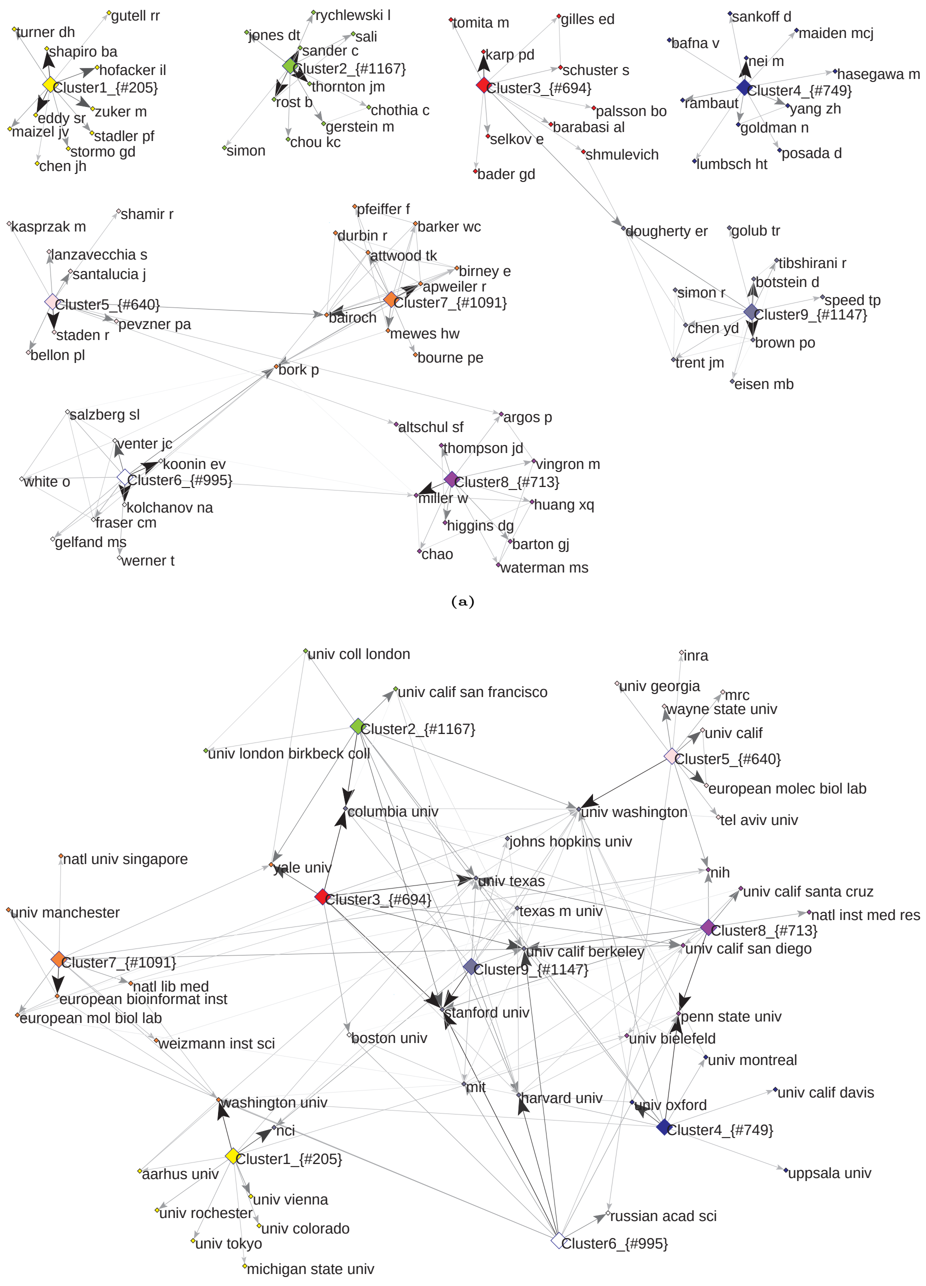
Figure 5.6 presents collaboration networks with for each cluster the 10 most prolific authors in (a), and the 10 most prolific institutions in (b), each in terms of number of publications. Some authors are among the ten with most publications for more than one cluster (among others, *Peer Bork*). Figure (b) is more dense, indicating quite logically that a lot among the best institutions are very active in more than one cluster.

### 5.2.3 'Naive' dynamics

Figure 5.7 provides a view on how much attention the bioinformatics community has devoted to the different subfields through time. In (a), the yearly number of publications in each cluster is plotted. The rise of *Microarray analysis* and of *Phylogeny & evolution* is striking. The former contained by far the most publications in 2004. In (b), for each cluster a box and whisker plot indicates the distribution of publication years of all of its member papers.

Some of the clusters, e.g., *RNA structure prediction* (#1) and *Genome sequencing & assembly* (#5), clearly represent older subfields that are (relatively) almost fading away. On the other hand, the clusters *Systems Biology & molecular networks* (#3) and *Microarray analysis* (#9) are very recent subfields in which a lot of research is conducted today. Cluster 4, *Phylogeny & evolution*, actually represents a relatively old research field, but new developments in bioinformatics made it regain a lot of attention since the start of the new millennium, clearly visible in the figure.

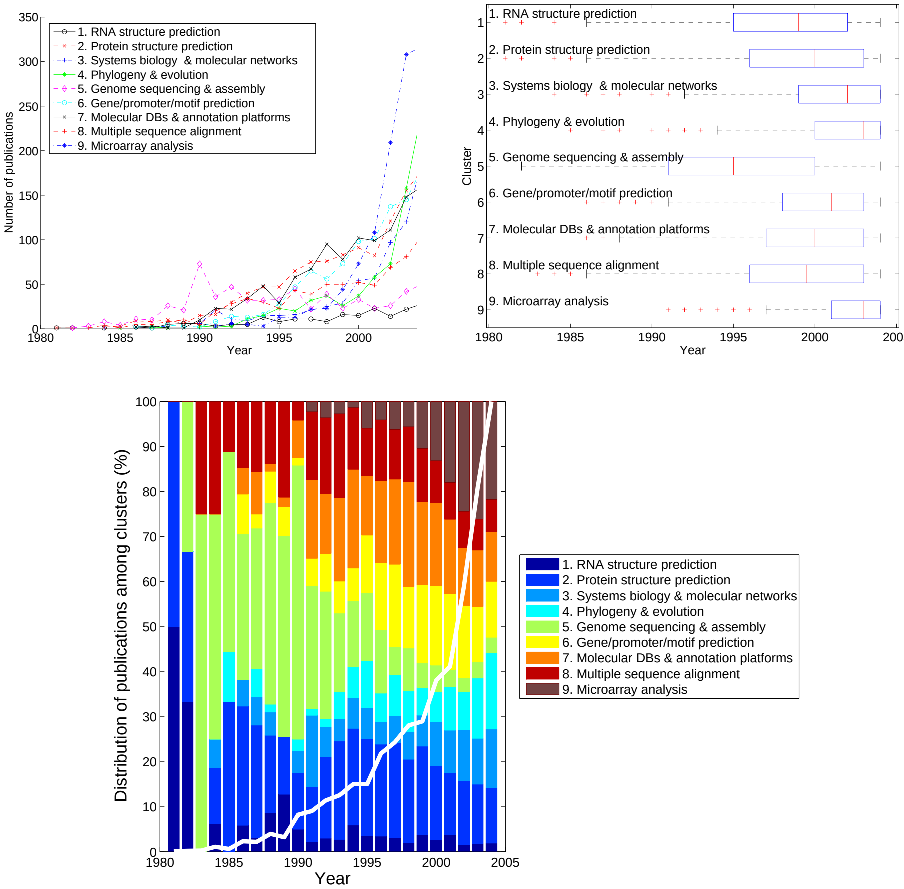
Figure 5.7(c) provides a different view on the same data. Here, the share (in %) of the yearly publication output that belongs to each cluster is shown with a different color. The white line depicts the yearly number of bioinformatics publications relative to the number of publications that were published in the year 2004 (1455). This way of visualizing demonstrates the relative growing and fading of the different topics in bioinformatics. An upward trend in relative number of publications can again definitely be ascribed to the clusters *Microarray analysis* (#9), *Phylogeny & evolution* (#4), and *Systems Biology & molecular networks* (#3).



(a)

(b)

**Figure 5.6:** Collaboration networks with for each of nine clusters: **(a)** the 10 most prolific authors, and **(b)** most prolific institutions, both in terms of number of publications. Some authors are among the ten with most publications for more than one cluster (among others, *Peer Bork*). Figure (b) is more dense, which indicates that a lot of institutions highly contribute to more than one cluster.



(c)

**Figure 5.7:** 'Naive dynamics' of the 9 clusters providing a view on how much attention the bioinformatics community has devoted to the different subfields over time. **(a)**. Yearly number of publications in each cluster. The rise of *Microarray analysis* and of *Phylogeny & evolution* is striking. **(b)**. Box and whisker plots of the publication years in each cluster. The extent of a box indicates the interquartile range, the median publication year is indicated with an internal vertical line. **(c)**. Share (in %) of total yearly publication output in each cluster. The white line indicates the yearly number of publications, relative to the number in 2004 (1455). This way of visualizing demonstrates the relative growing and fading of the different topics in bioinformatics. An upward trend in relative number of publications can be ascribed to the clusters *Microarray analysis* (#9), *Phylogeny & evolution* (#4) and *Systems Biology & molecular networks* (#3).

The citation network is visualized in Figure 5.8 with a different color for each of 9 clusters. Each node represents one publication. Only 1798 publications are shown that cite at least two other papers in the bioinformatics set and that are cited twice or more from within the set and five times in total. In general, member papers of one cluster are localized in close vicinity in the network, although this is not required by the hybrid clustering algorithm that also considers textual similarities besides the structure of the citation network.

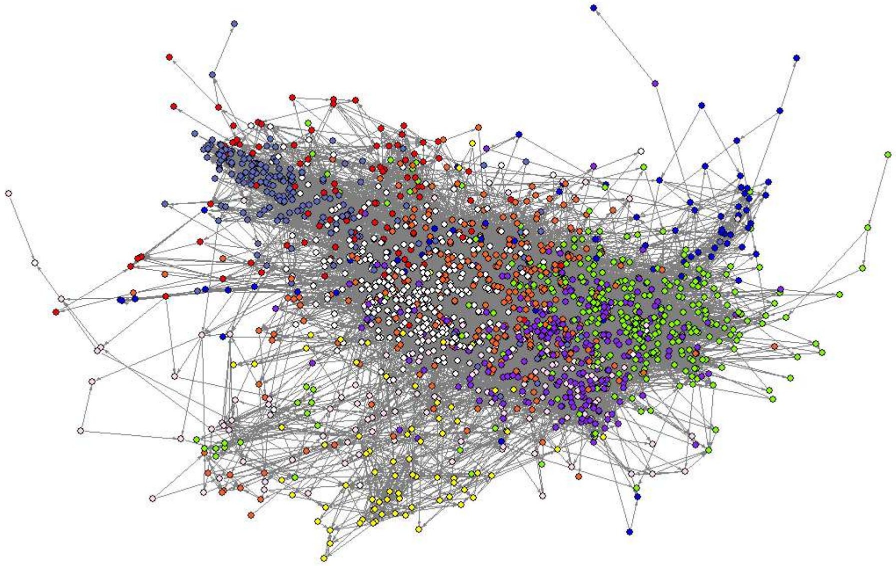
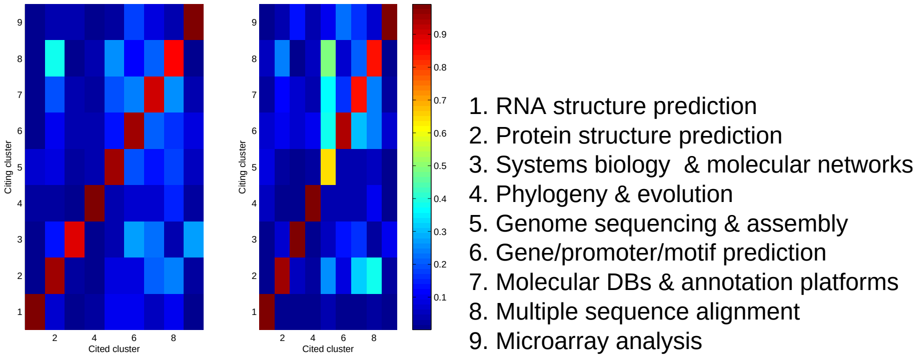


**Figure 5.8:** Citation network with a different color for each of 9 clusters. Each node (dot) represents one publication. Only 1798 publications are shown that cite at least two other papers in the bioinformatics set and that are cited twice or more from within the set and five times in total. *Fruchterman Reingold* layout in *Pajek* [15].

In Figure 5.9, one can observe cross-cluster citation patterns. In (a), each row is normalized, representing the 'citing' pattern of a cluster, whereas in (b) each column or 'cited by' pattern is normalized. Most citations are internal to a cluster, which explains the dark diagonals. For example, the upper line of the matrix in (a) indicates that, next to the obvious great majority of within-cluster citations, most external citations from the *microarrai* cluster (#9) are given to the *gene* cluster (#6). In (b), from the ninth column it is clear that the *microarrai* cluster receives external citations almost solely from clusters 3 and 6.



(a)

(b)

**Figure 5.9:** Cluster-to-cluster citation pattern. **(a)**. The rows of the matrix are normalized ('citing' pattern for each cluster). **(b)**. The columns of the matrix are normalized ('cited by' pattern for each cluster). Most citations are internal to a cluster, which explains the dark diagonals.

## 5.3 Dynamic hybrid clustering

In Figure 5.10 we give the same picture as in Figure 3.6(a) on page 100, but for the dynamic analysis we consider time windows used to subdivide the bioinformatics set into different periods. Seven periods have been defined, while striving for an approximately equal number of publications in each period.

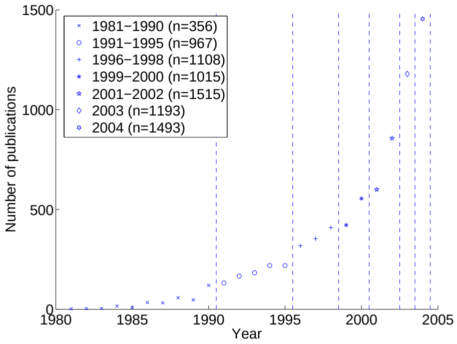


**Figure 5.10:** Evolution of publication output in bioinformatics. Time windows defined for the dynamic analysis are indicated with vertical lines as well as in the legend.

### 5.3.1 Matching and tracking clusters through time

Our strategy for dynamic clustering, namely by matching and tracking clusters through time, is demonstrated in Figure 5.11. Each horizontal level represents one period, indicated by the label of the leftmost circle and with a different gray level. Node size represents number of publications and for each cluster the best TF-IDF term is shown.

In each period, a separate hybrid clustering was performed and the optimal number of clusters was again determined by observing the dendrogram, Silhouette curves, and *Ben-Hur* stability plot. Next, a *complete graph* was constructed with all cluster *centroids* from each period as nodes, and with all mutual cosine similarities as edge weights, calculated in the 10 dimensional latent semantic space. For clarity, we stress that the clustering algorithm used to subdivide all documents in a specific period is the same hard algorithm as used for the static hybrid clustering: documents are assigned to exactly one cluster only.

### 5.3.2 Chains of clusters

Next, a two-step approach was followed in order to form *'cluster chains'*. Firstly, in the complete graph only those edges with weights (i.e., cosine similarities between cluster centroids) larger than $T1 = 0.95$ were retained. All other edges with values below $T1$ were discarded. Somewhat surprisingly, most cluster chains were well established after this single step with stern requirement of

95% similarity. Secondly, clusters that had no similarity larger than $T1$ with any other cluster, and hence were completely detached, were yet allowed to join an existing chain if their similarity to each member of that chain was larger than $T2 = 0.8$. Such clusters are depicted as a diamond instead of a circle. The two thresholds $T1$ and $T2$ were determined by observing Figure 5.12.

Further on, clusters resulting from the *static* hybrid clustering algorithm used in Section 5.2 will be mentioned as *clusters*, whereas *cluster chains* or *chains* shall be used for the *dynamic* hybrid clustering. Although the static clustering algorithm came up with 9 clusters, figure 5.11 suggests that in total 11 cluster chains could be distinguished, 3 of which contain publications from all seven periods between 1981 and 2004. Five chains emerged in 1991 and were still present in 2004. Table 5.2 presents the name of each cluster chain.

**Table 5.2:** The 11 cluster chains obtained by dynamic hybrid clustering.

| Cluster chain number | Name |
|---|---|
| 1 | Sequence DBs & analysis |
| 2 | RNA structure prediction |
| 3 | Gene regulation |
| 4 | Phylogeny & evolution |
| 5 | Multiple sequence alignment |
| 6 | Databases & software |
| 7 | Protein structure prediction |
| 8 | Genome analysis |
| 9 | Systems biology & molecular networks |
| 10 | Microarray analysis |
| 11 | Clustering of gene expression |

The *Microarray analysis* chain (#10) appeared in 1999–2000 and the *Clustering of gene expression* chain (#11) one period later (2001–2002). The chain on the left-hand side in Figure 5.11 (#1) lasted from the first until the third period. Besides these groups of documents that are connected in cluster chains, some others are not connected to any chain. By disregarding these clusters that could not be linked to any other cluster in another period, our dynamic methodology of tracking clusters through time can be considered less *'hard'* than the standard hierarchical clustering algorithm in the sense that not all publications need to be attributed to at least one chain. Subsets of documents that do not clearly belong to any of the chains can be discarded.

### 5.3.3 Comparing clusters with cluster chains

Concept networks and detailed publication lists revealed that most clusters correspond to one cluster chain, except for Cluster 5. This is also illustrated in Figure 5.13. Hence, we have often given the same name to cluster chains. The figure visualizes intersections between clusters and chains with the *Jaccard* sim-

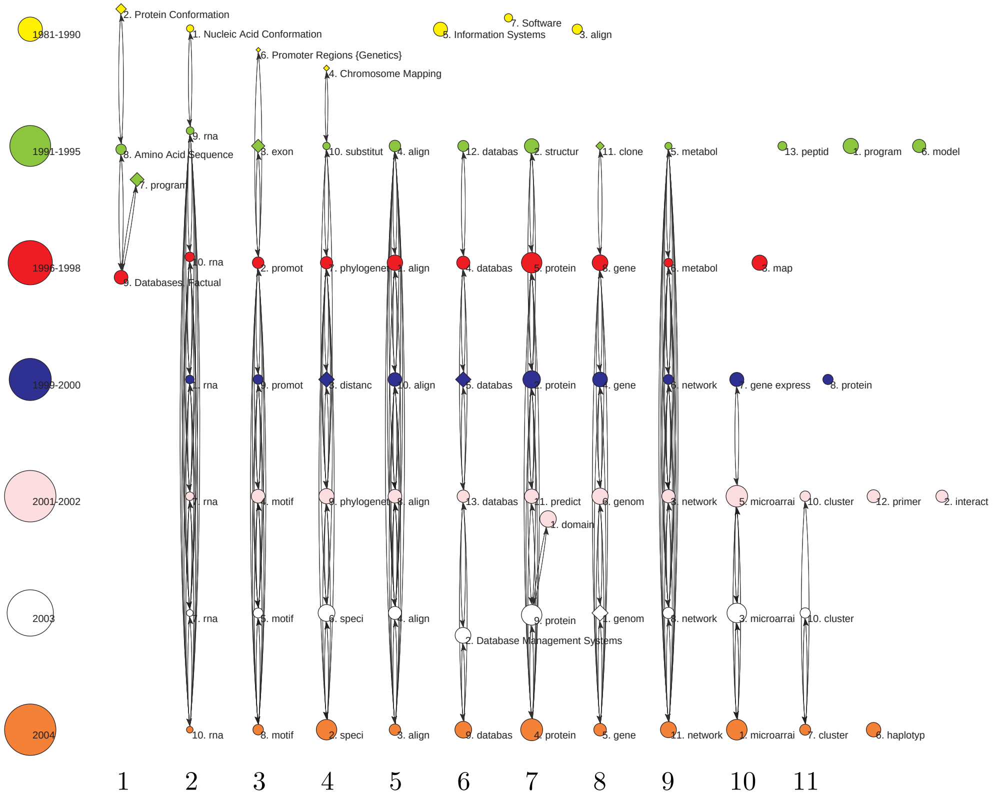


**Figure 5.11:** Dynamic clustering: matching and tracking clusters through time. Each horizontal level represents one period, indicated by the label of the leftmost circle and with a different gray level. Node size represents number of publications and for each cluster the best TF-IDF term is shown. Tiny numbers indicate the number of a cluster in its own period, whereas the numbers in bigger font indicate the *cluster chain* number. Corresponding names for each cluster chain are given in Table 5.2. *Pajek* was used for visualization [15].

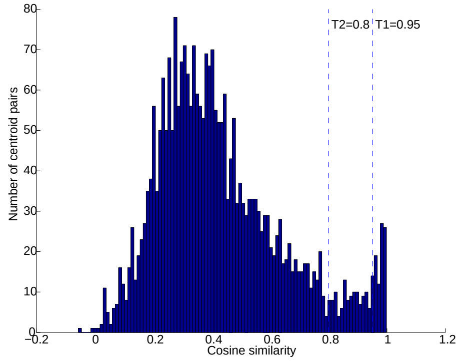
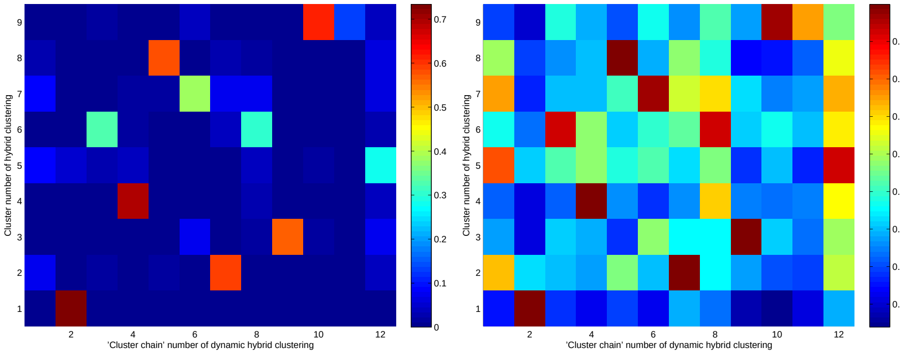


**Figure 5.12:** Histogram of mutual similarities between all cluster centroids of Figure 5.11. A clear demarcation of strong ($T1$) and less strong ($T2$) cluster matches could be defined visually. A few similarities are smaller than 0 because of latent semantic vector calculations.

ilarity coefficient in (a) (see Section 2.3.2), and centroid cosine similarities in (b) (see Section 2.1.2).



**(a)**

**(b)**

**Figure 5.13:** Comparing clusters from the *static* hybrid clustering with *cluster chains* from the dynamic hybrid clustering. Similarities between corresponding centroids are measured as: **(a)** *Jaccard* coefficients and **(b)** cosine similarities between centroids.

Cluster 5, which was called *Genome sequencing & assembly*, contains a lot of publications that are not included in any chain (but actually in #12, called *Not in chain*). Another part is member of the *Sequence DBs & analysis* chain (#1), besides papers from Cluster 2, 7, and 8. Other publications of Cluster 5 are to a lesser extent spread out over clusters 2, 3, 4, 8, & 10.

Cluster 6 (*Gene/promoter/motif prediction*) has largely been split in chain 3 and chain 8, and the *Microarray analysis* cluster (#9) has split off the *Clustering of gene expression* chain (#11). According to the dendrogram at higher resolution for the static clustering (not shown), the *Microarray analysis* cluster would indeed be the first one to split in two parts if more than nine clusters would be desired.

Finally, Cluster 7 overlaps for the larger part with chain 6, but to a lesser extent also with chains 1, 7, and 8.

The chains that have been found by the dynamic clustering procedure might be more accurate than the clusters found with the standard algorithm. If, for a certain period, a non-optimal number of clusters would be chosen, the strategy of tracking and matching of clusters through time can compensate for it by joining more clusters of that period to the same cluster chain. Likewise, a cluster chain might also split up into different branches, when, for example, two centroids of a later period are both linked to the same one of a previous period and both develop further in dissociated chains. In our data set such dissociation is not observable, but the joining of two centroids of the same period into one chain, is. For example, in the period 2001–2002, two clusters (*'11. predict'* and *'1. domain'*) are both attached to Cluster chain #7. If a line of research would be discontinued in a certain period, but be resumed again in a later one, this would also be detected and the resulting chain would just bridge the period with no activity in that area. A drawback, however, is that some clusters can still be overlooked by application of the visually defined, simple similarity thresholds. Improvement for the dynamic methodology might be obtained by using more complex rules for the forming of the chains of clusters.

### 5.3.4 Term networks

For each of the 11 cluster chains, Figure 5.14 presents the term networks with the best 10 keywords given by authors, and Figure 5.15 the best 10 title or abstract terms. We do not describe these networks in detail since most chains correspond to one cluster. The central node of each cluster chain reveals the chain number, the chain name, and the number of publications in the chain. Note that chain numbers do not correspond to cluster numbers of the *static* clustering. Figure 5.16 visualizes collaboration networks for the cluster chains, showing the 10 most prolific authors.

### 5.3.5 Chain properties

Figure 5.17 presents various statistics of the cluster chains of Figure 5.11. 'Chain' *ALL* on the right-hand side refers to the complete set of 7401 publications; 'chain' #12 represents the set of documents that were not a member of any chain. The four sub-figures contain box and whisker plots of number of authors, institutions, references, and citations, for each publication in each cluster chain. The natural logarithm of each statistic was used because of the highly skewed distributions; the number of references and citations were beforehand increased by 1 to prevent minus infinite values for $log(0)$.

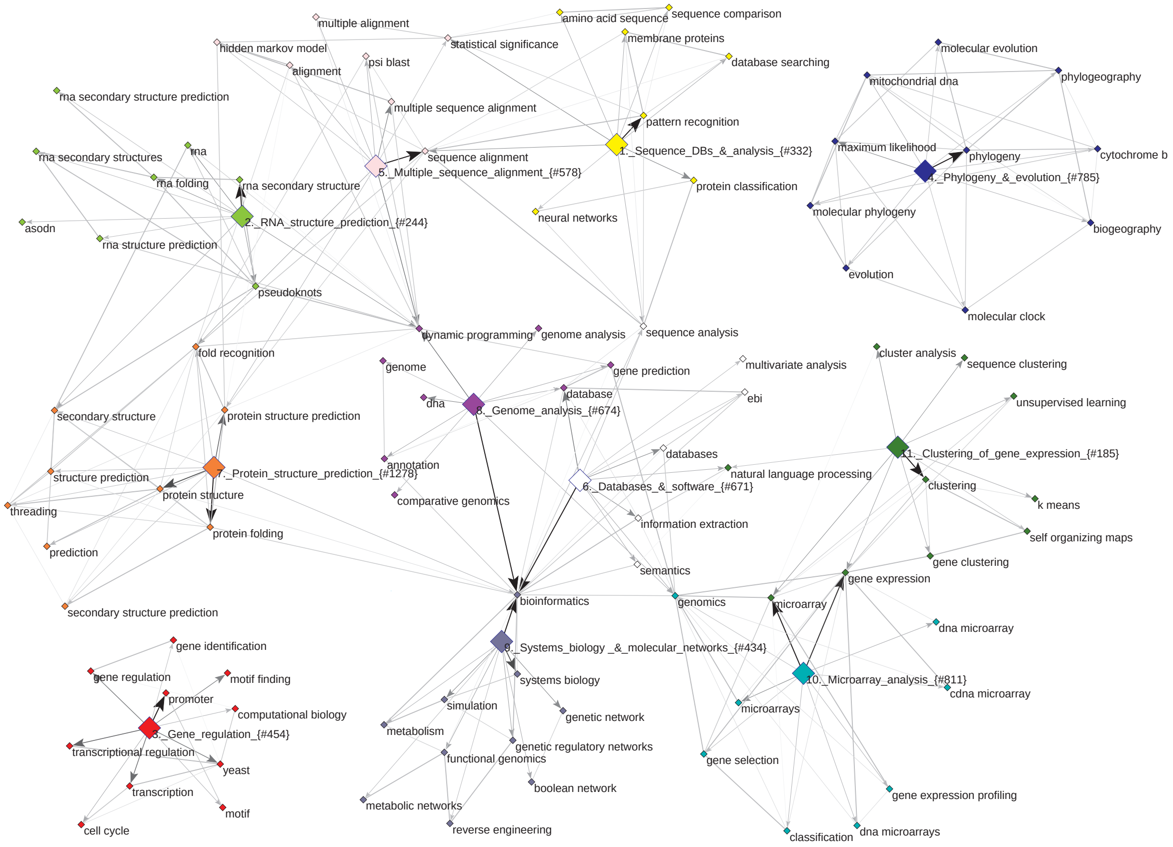


**Figure 5.14:** Term networks for the *cluster chains* with the best 10 author keywords according to mean TF-IDF scores.

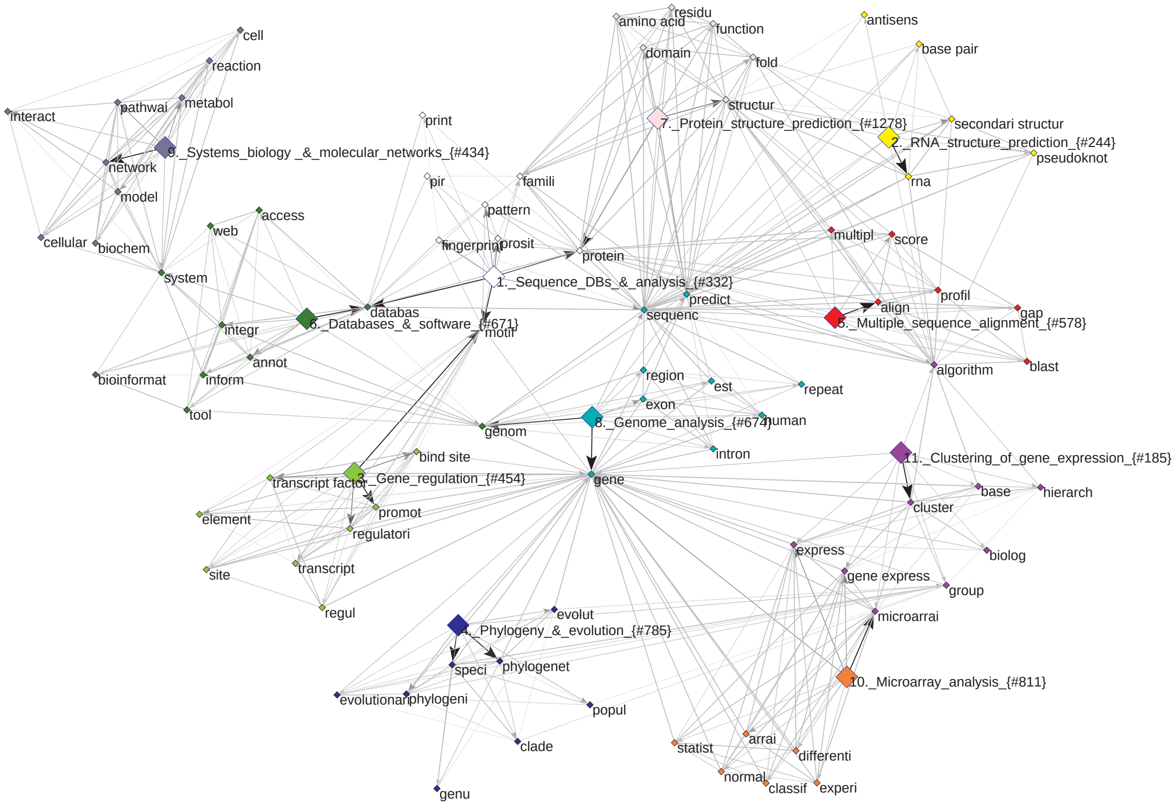


**Figure 5.15:** Term networks for the *cluster chains* with the best 10 terms or phrases from titles and abstracts, according to mean TF-IDF scores.

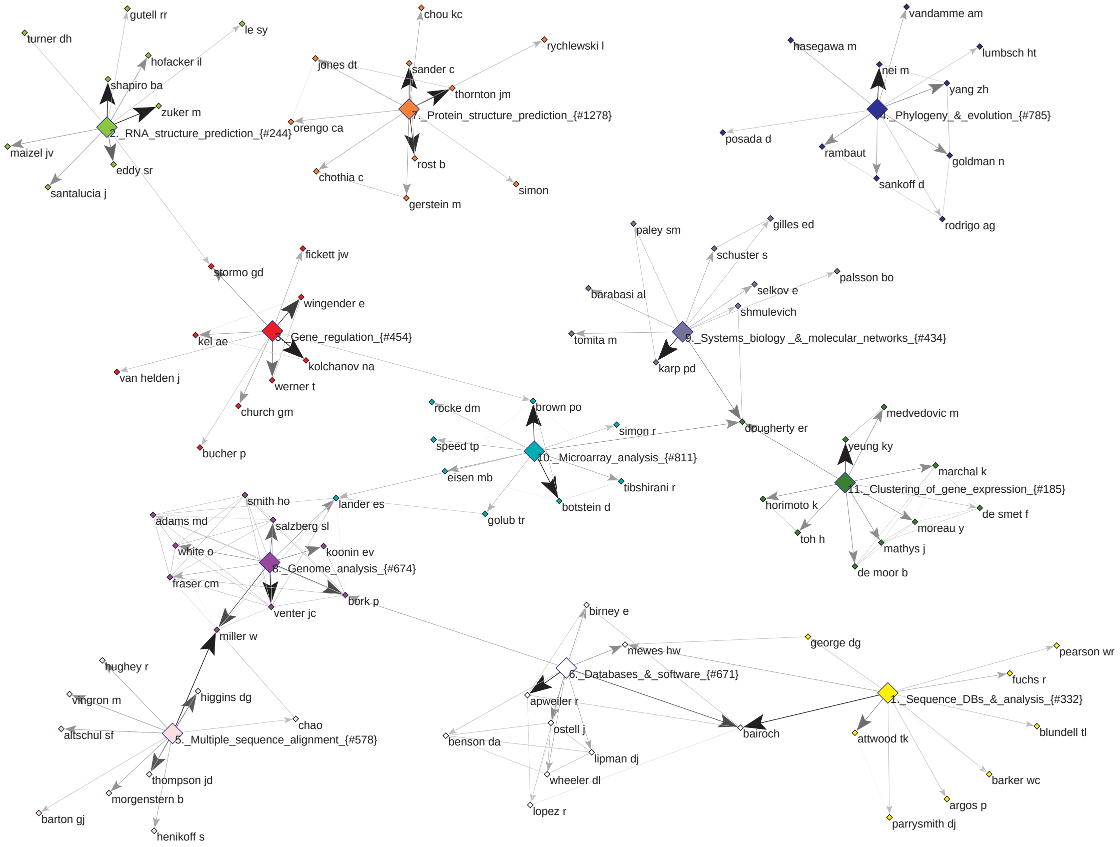


**Figure 5.16:** Collaboration networks with the 10 most prolific authors in each cluster chain.

In the upper plot, the chain with the highest median number of co-authors is *Microarray analysis* (#10), but chain 8 (*Genome analysis*) has the most outliers with peaks in number of co-authors, such as on papers related to the Human Genome Project. On the other hand, the oldest chain (#1) in general has more papers for which only a few people collaborated. From the second, related plot it is (quite logically) clear that chains 8 and 10 also have the highest number of collaborating institutions. Chains 4 and 11 seem to represent subdisciplines which also have a larger than average number of cooperating institutions.

The third sub-plot gives patterns of number of references, but the last one with number of citations is more interesting. Differences between the various chains are visible, but it should be noted that no normalization has been done here with respect to the age of publications. The *Clustering of gene expression* chain (#11) is cited the least, but it is the most recent chain that only germinated in 2001. Chains 4 and 6 also have a relatively lower number of citations, whereas chains 8, 3, 2, 7, & 5 are more frequently cited. Finally, the oldest chain (#1) seems to have gathered most citations up till now.

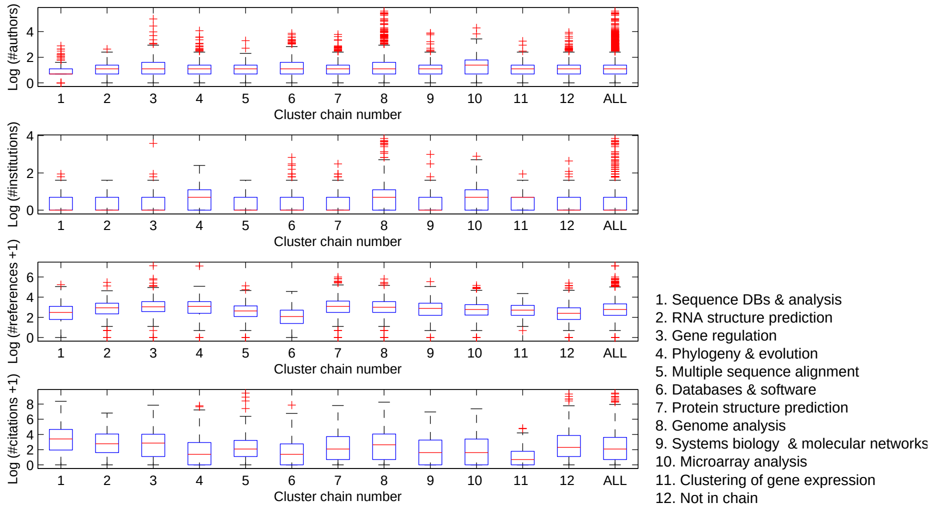


**Figure 5.17:** Box and whisker plots for number of authors, institutions, references, and citations for each publication in each cluster chain of Figure 5.11. 'Chain' 12 represents the set of publications that were *not* allocated to any chain and 'chain' *ALL* is the complete set of 7401 papers.

### 5.3.6 Dynamics

Figure 5.18 visualizes the relative activity in all chains, analogously as Figure 5.7 does for *static* clusters. It is clear that the share of publications *Not in chain* (#12) diminishes mostly with respect to previous years. This is an indication of the bioinformatics field starting to form crisp lines of research, especially after the year 1990.

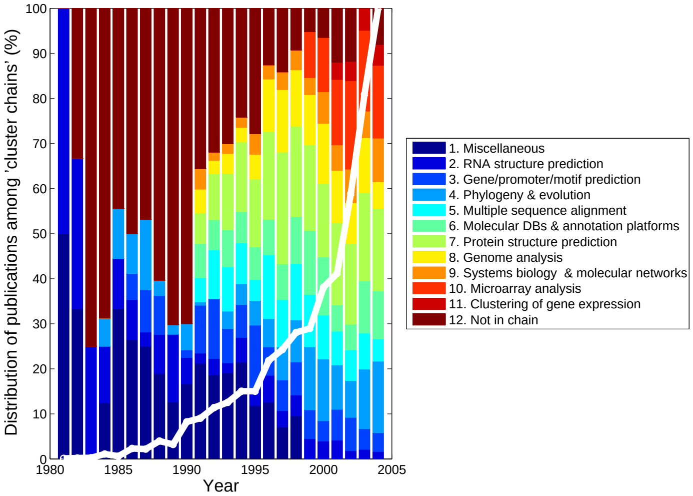


**Figure 5.18:** Distribution of the total yearly publication output among cluster chains. The white line indicates the yearly number of publications, relative to the number in 2004 (1455). ‘Chain’ 12 actually represents all publications that are *not* connected to any cluster chain.

## 5.3.7 The chain *Systems Biology & molecular networks*

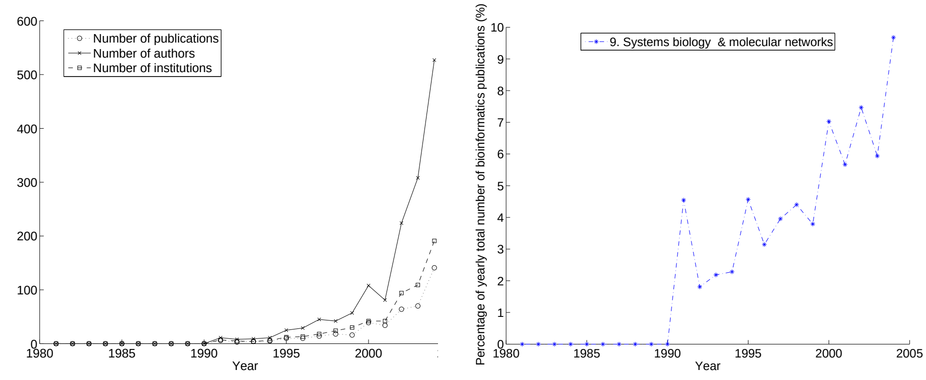
Computational Systems Biology studies biological systems at various scales, their building blocks, and how these form networks of relationships. Dynamic quantitative models are built based on properties of the components, and even allow predictions.

Figure 5.19(a) plots the yearly number of publications in the *Systems Biology & molecular networks* cluster chain (#9), as well as the yearly number of unique authors and institutions. Figure 5.19(b) reveals the sharp rise in attention devoted to systems biology by the bioinformatics community. In 2004, almost 10% of publications were related to this subdiscipline.

### Dynamic term networks

A dynamic term network allows the observation of shifts in vocabulary and focus of a specific (sub-)field of interest. A central node is annotated with an indication of the period (such as ‘1991–1998’), with the period number, and with the number of publications.

Figure 5.20 shows *dynamic* term networks for the *Systems Biology & molecular networks* cluster chain, with for different time periods the best 10 terms



(a)

(b)

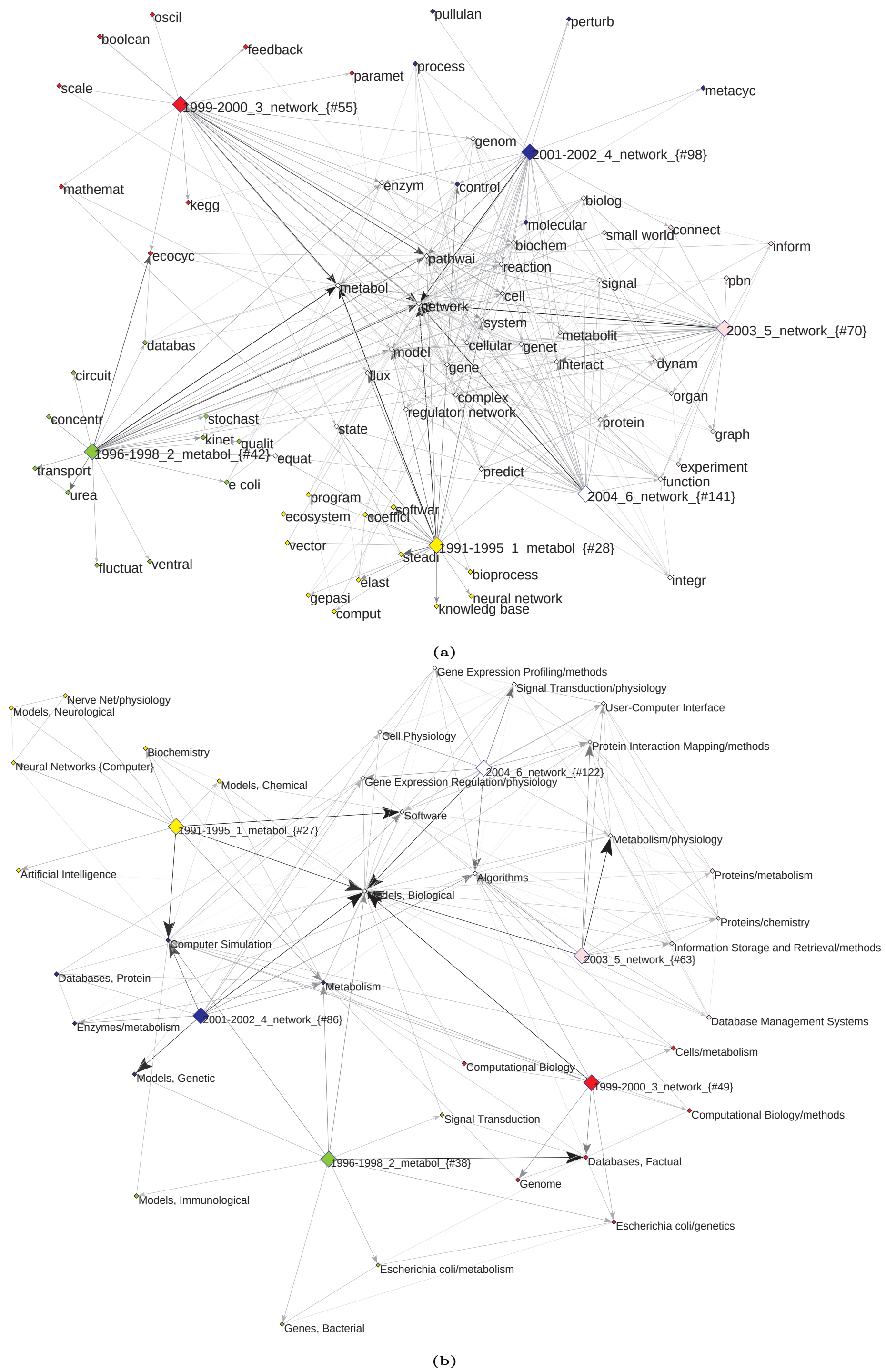
**Figure 5.19: (a)**. Evolution of number of publications, authors, and institutions in the *Systems Biology & molecular networks* chain. **(b).** Share (in %) of total yearly publication output in this chain.

or phrases from titles and abstracts in (a), and the best 10 MeSH terms in (b). Here, the *TF* factor was again either 1 or 2, for *minor* and *major* MeSH descriptors, respectively. Figure 5.21 contains the best TF-IDF author keywords.

The network based on terms from titles and abstracts is denser than the author keyword network because usually just a few author keywords are annotated to a publication and consequently these have less chance to co-occur with others on the same document. The central node in the lower right corner of Figure 5.21 (*1996-1998_2_metabol_{#42}*) corresponds to the period 1996–1998, which accounts for 42 papers (best described by the term *metabol*). It is a bit isolated in the sense that none of its terms are also among the best for another chain, and no term has co-occurred with one of the salient terms of another period. Looking in a clockwise manner, starting in the lower left corner, temporal keywords of successive periods are illustrated.

### Dynamic collaboration networks

Figure 5.22 presents dynamic collaboration networks for the *Systems Biology & molecular networks* cluster chain with for each period the 10 most prolific authors in (a), and the 10 most most prolific institutions in (b). Important authors are present in the networks. Among others, *Bernhard Palsson*, *Peter Karp*, *Leroy Hood*, and *Minoru Kanehisa*.



(a)

(b)

**Figure 5.20:** Dynamic term network for the *Systems Biology & molecular networks* cluster chain (#9) with, for each period and according to mean TF-IDF scores: **(a)** best 10 terms or phrases from titles and abstracts, **(b)** best 10 TF-IDF MeSH terms.

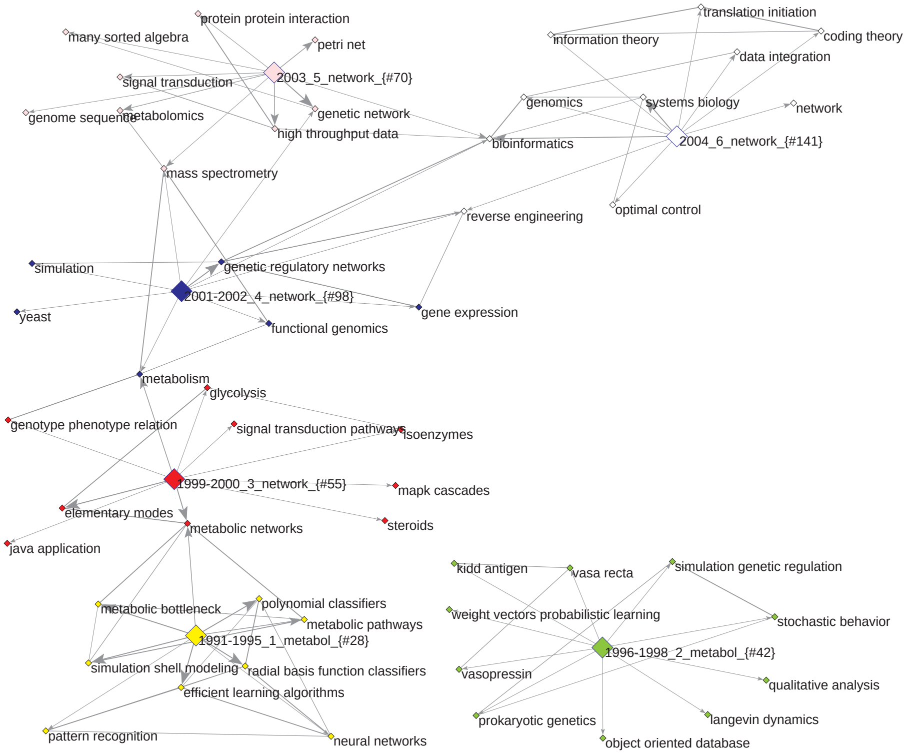
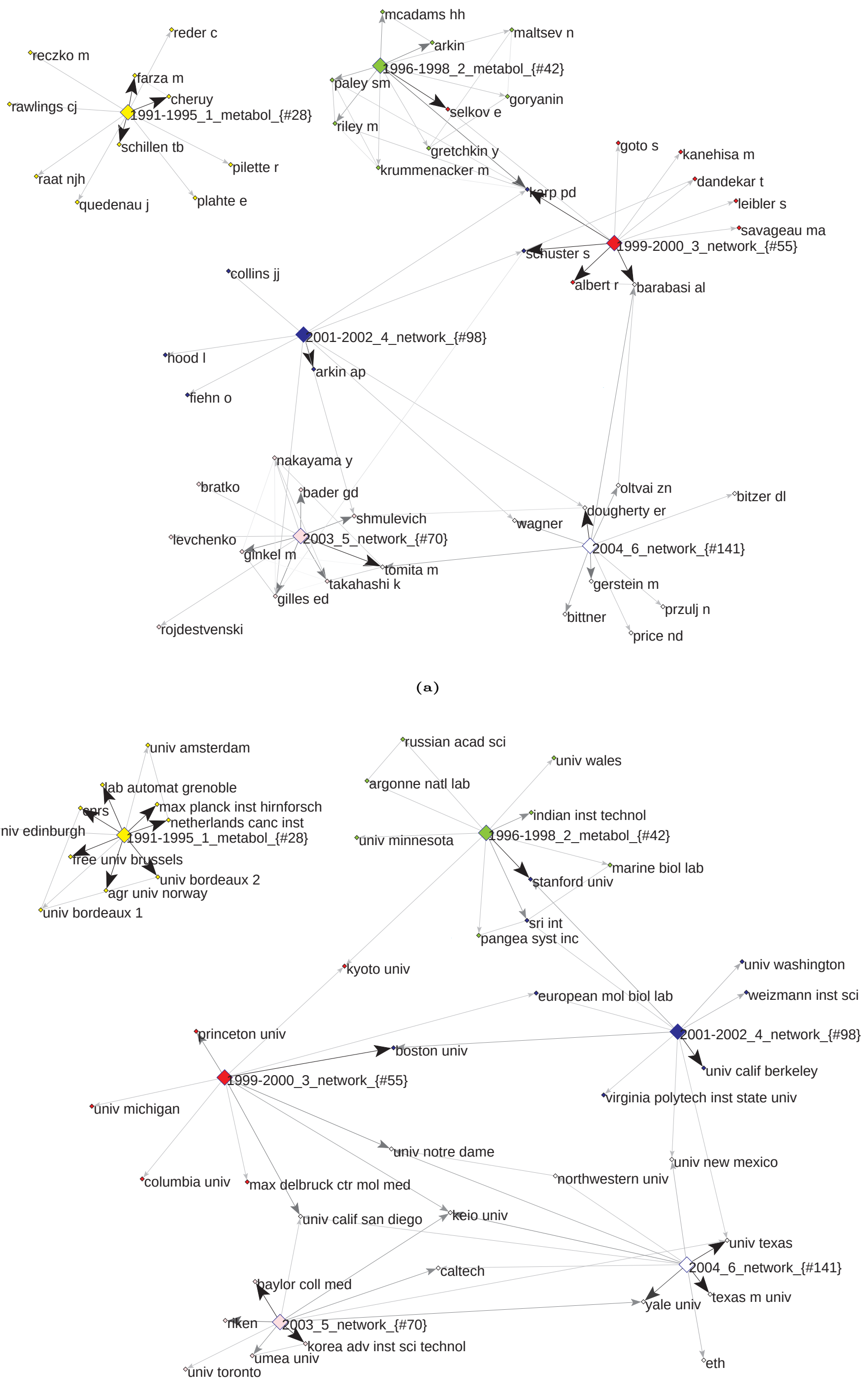


**Figure 5.21:** Dynamic term network for the *Systems Biology & molecular networks* cluster chain with the best 10 author keywords for each period.



(a)

(b)

**Figure 5.22:** Dynamic collaboration networks for the *Systems Biology & molecular networks* cluster chain with, for each period: **(a)** the 10 most prolific authors and **(b)** the 10 most most prolific institutions.

### 5.3.8 Cross-chain citations

Another interesting point of view is given in Figure 5.23. It visualizes the number of cross-chain citations through time. For example, chain 10 (*Microarray analysis*) mostly cited chain 3 (*Gene regulation*) between 1999 and 2003, but this citing pattern became somewhat more diffuse in 2004. On the other hand, the same *Microarray analysis* chain received in 2003 most citations from chain 11 (*Clustering of gene expression*), whereas in 2004 also from *Systems Biology & molecular networks* (#9).

Figure 5.23 actually illustrates the evolution in citation patterns and thus of interdependencies between chains, but the figure does not uncover relative growth or decrease in number of citations to a specific chain through time. In each period, the relative amount of external cross-chain citations is indicated with a color code, but they can thus not be compared across periods. However, such diagrams can yet aid emerging trend detection. For example, if a new chain is brought into existence in a certain period and quickly gets highly cited, it might represent a *hot topic*. Likewise, if two chains that rarely cited each other suddenly have a high share of cross-citations, possibly mediated by a third, more recent chain that cites both other chains, then this might as well be an indication.

### 5.3.9 Impact

The evolution of the field's mean observed citation impact was presented in Figure 3.8 on page 101. A strong linear increase of citation impact was observed in the 1990s, followed by a sharp decline in the new millennium. Figure 5.24(a) shows the yearly *Impact Factor* for each cluster chain, defined as the overall mean number of citations given in a specific year $X$ to articles belonging to the chain that were published during the two preceding years $X-1$ and $X-2$. In Figure 5.24(b), the average Impact Factor over all years in a period is indicated for each chain. The cluster chain with clearly the highest impact is *Genome analysis*, but early *Microarray analysis* had a very high impact, too. During the nineties, the *Gene regulation* chain also achieved very high Impact Factors. A sharp rise in impact, over the years, for the *Systems Biology & molecular networks* is also remarkable. Although in Figure 5.17 chain 1 seemed, in general, to have received the highest number of citations, the mean Impact Factor is less distinctive. The age of the concerned papers was thus indeed an important factor.

The *ISI* Impact Factor for the complete bioinformatics set ('All') was, in general, less than 10 in the early nineties, but increased during later years of the previous millennium towards 15, and even to approximately 17 in 2002. This overall high impact is partially a consequence of the citation-based component of the retrieval strategy (see Section 4.6). An observation that has already been made in Section 3.6.3 is that for most clusters a drop in impact has occurred in the year 2003 or 2004. The overall Impact Factor confirms this observation by dropping back to approximately 10 in 2004.

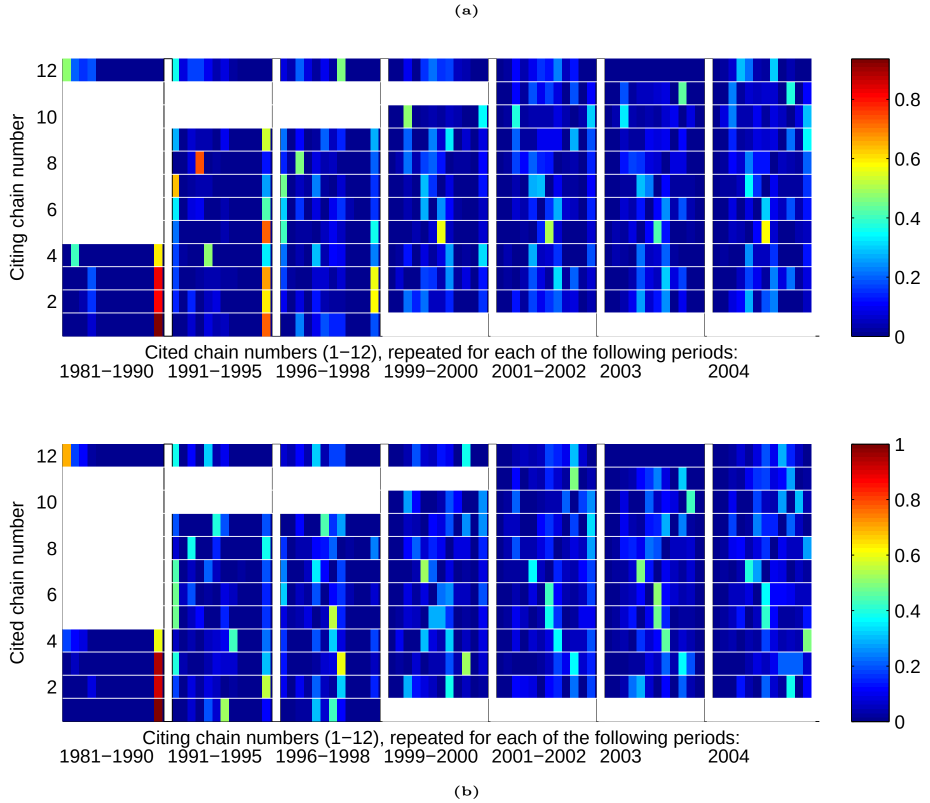
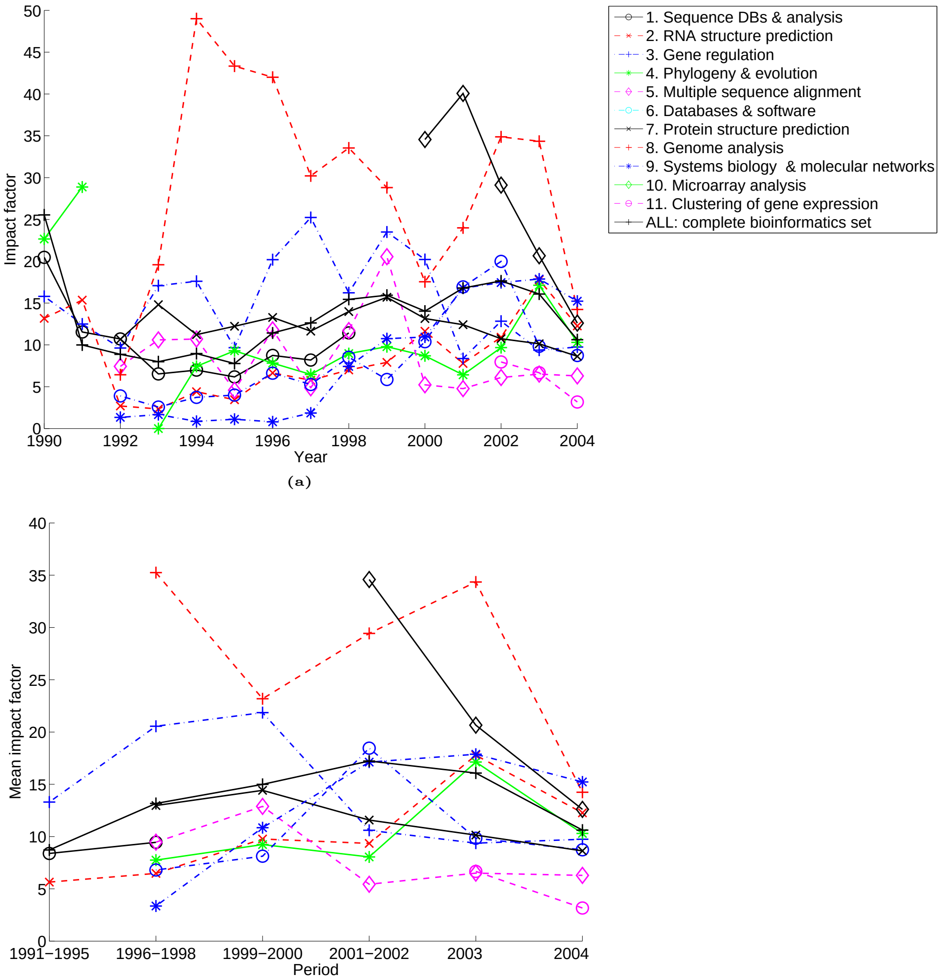


**Figure 5.23:** External chain-to-chain citations. Each row represents one detected cluster chain of Figure 5.11, except for number 12 on the upper row, which represents the set of publications that were *not* allocated to any chain. **(a)**. For each period indicated on the $X$-axis, the proportion of external citations from each cluster chain on the $Y$-axis to all other cluster chains is indicated with a color code. 'Self-citations' to the same cluster chain are not taken into account, neither are citations to papers outside the bioinformatics set. **(b)**. For each cluster chain the proportion of citations received from other chains.



(a)

(b)

**Figure 5.24: (a).** The yearly Impact Factor for each cluster chain, defined as the overall mean number of citations given in a specific year $X$ to articles belonging to a chain that were published during the two preceding years $X-1$ and $X-2$. **(b).** The average Impact Factor per period for each chain. The cluster chain with clearly the highest impact is *Genome analysis*, but early *Microarray analysis* had a very high impact, too. During the nineties, the *Gene regulation* chain also achieved high Impact Factors. A sharp rise in impact, over the years, for the *Systems Biology & molecular networks* is also remarkable. For most clusters a drop in impact has occurred in the year 2003 or 2004. The overall Impact Factor confirms this observation by dropping back to approximately 10 in 2004.

## 5.4 Concluding remarks

In this chapter the subject domain of interest was the bioinformatics field, characterized by an exponential increase in publication output during the last two decades. The demarcation of the field was achieved by the bibliometric retrieval scheme that has been introduced in Section 4.6. Seven consecutive periods containing approximately the same number of publications were defined.

The aim of the first part of this chapter was to demonstrate our hybrid clustering procedure based on *Fisher*'s inverse chi-square method, which was revealed in Chapter 4 as the preferred method for integrating textual content and citation information, at least if implementational complexity is not an issue. Otherwise, a much easier linear combination could certainly also provide satisfactory results that are significantly higher than text-only or link-only clustering algorithms. Our combined strategy for defining the number of clusters suggested nine subdisciplines for bioinformatics. For each cluster we provided term and collaboration networks and the most representative publications according to, for example, HITS and PageRank. Next, cluster representation of the 5 most active countries was analyzed, as well as *'naive' dynamics* to provide a view on how much attention the bioinformatics community has devoted to the different subfields through time. Finally, the citation network and cross-cluster citation patterns were given.

A methodology was developed for dynamic clustering in the second part of the chapter. The same hybrid clustering algorithm was applied multiple times, but each time restricted to publications in one of the defined periods. Eleven *cluster chains* could be identified by matching and tracking clusters through time. Their concept networks, their evolution, and various statistics were analyzed. The chains were compared with the clusters found by the *static* hybrid clustering of the complete bioinformatics set. Most cluster chains corresponded to one cluster. 'Dynamic' networks allowed the observation of shifts in collaboration patterns and in terminology used within cluster chains. The *Systems Biology & molecular networks* chain was analyzed in more detail. Next, external chain-to-chain citations were investigated in each period to visualize the evolution of citation patterns and, hence, of dynamic interdependencies between chains. Finally, the yearly impact of each cluster chain was determined, based on the *ISI* Impact Factor.

Dynamic clustering is an important research topic in the light of dynamic document sets and emerging trend detection. The chains that have been found by the dynamic procedure might be more accurate than clusters resulting from the standard algorithm applied to the complete data set at once. Our dynamic methodology of tracking clusters through time can be considered less *'hard'* than the standard hierarchical clustering algorithm. Indeed, subsets of documents that do not clearly belong to any of the chains can be discarded. A drawback, however, is that some clusters can still be overlooked by application of the visually defined, simple similarity thresholds. Improvement for the dynamic methodology can still be obtained by using more complex rules for the forming of cluster chains. On the other hand, the method can compensate for wrong choices of cluster numbers in a specific period by joining more than one cluster of the period in the chain, or by splitting the chain into multiple dissociated chains. A chain might as well skip some periods in which a research subject has not received attention from the community.

To conclude, the hybrid clustering algorithm exploiting information from both text and citation worlds, possibly complemented with the strategy of tracking clusters through time, provide very powerful tools to unravel the cognitive structure of scientific or technological fields (or of other dynamic document sets), to cast eyes upon the evolution of existing subdisciplines, and to aid detection of emerging or converging clusters and hot topics.

# Chapter 6

# General conclusions and perspectives

## 6.1 Conclusions

### 6.1.1 Hybrid clustering

The complex nature of mapping various aspects of knowledge motivates approaches that incorporate different viewpoints on the same data collection. Textual and bibliometric or graph-analytic techniques can provide different perceptions of similarity between documents or groups of documents and different methods to observe dynamics in massive and evolving bibliographic databases. This complementarity was demonstrated by serial combination of text mining and bibliometric techniques. We proposed various schemes to integrate textual and bibliometric methods and our hypothesis was confirmed that such an integrated approach leads to a better comprehension of the structure and dynamic properties of textual corpora. Such hybrid methodologies are valuable tools to facilitate endeavors in mapping fields of science and technology and in research evaluation.

Performance of unsupervised clustering and classification of scientific publications was significantly improved by profoundly integrating textual content with citations. In general, text-only information proved much more powerful than mere citations and dimensionality reduction by SVD greatly improved results, especially when applied to textual information. However, the best outcome was obtained by integration of the heterogeneous information sources. A combination of text-based and bibliometric components was also used in *bibliometric retrieval* to improve the complex delineation of interdisciplinary research fields such as bioinformatics.

An important remark regarding clustering is that we opted not to use special graph partitioning algorithms when integrating text mining and citation-based techniques. We mainly worked with vector space representations of citation graphs. This choice was suggested by the resemblance between bibliometric similarity measures such as *co-citation* or *bibliographic coupling*, and vector space clustering techniques.

Besides integrating data by using Random Indexing, we also devised an integration method in which pairwise distances between documents were converted to $p$-values compared to randomized data sets, and in which *Fisher*'s inverse chi-square method was then used to combine the $p$-values from all information sources. This method is generic and could be used to incorporate distances with highly dissimilar distributional characteristics, such as textual distances and distances based on bibliometric indicators. For a correct application we introduced a slight, rank-preserving modification to the formula for bibliographic coupling. The integration method was shown to significantly outperform corresponding text-only and link-only methods, as well as a mere concatenation of vectors. However, in experiments of Section 4.4 on page 130, *Fisher*'s inverse chi-square method did not significantly outperform corresponding linear combinations when SVD had been applied. Given the higher complexity of implementing *Fisher*'s inverse chi-square method and a reduced scalability, a carefully chosen weighted linear combination might be the preferred solution for integrating textual and citation information, on condition that LSI is used. Nevertheless, linear combination, which offers a very attractive, easy and scalable integration method, was yet outperformed by *Fisher*'s inverse chi-square method in another experiment regarding the Silhouette coefficient and stability.

Our hybrid clustering procedure based on *Fisher*'s inverse chi-square method was demonstrated in two case studies. The goal of the first case study was to unravel and visualize the concept structure of the field of library and information science based on more than 900 publications from a set of 5 journals. We compared five clusters found by hybrid clustering with six clusters that were found by text-only clustering and we clearly observed an improvement by the hybrid method. On the other hand, incorrectly assigned papers still occurred in the hybrid classification as well. We think that, in order to gain even better performance, a transition should be made towards fuzzy clustering algorithms. In the meantime, spurious assignments of documents to clusters can be detected by validation measures such as the Silhouette coefficient, and fuzziness might to some extent be mimicked by taking into account document similarities to all cluster centroids.

In the second case study the subject domain of interest was the bioinformatics field. The demarcation of the field was achieved by a bibliometric retrieval scheme. Next to a bibliometric analysis of this interdisciplinary field, characterized by an exponential increase in publication output during the last two decades, our combination of strategies for defining the number of clusters suggested nine subdisciplines. For each cluster we provided term and collaboration networks and the most representative publications according to, for example, HITS and PageRank. Next, cluster representation of the 5 most active countries was analyzed, as well as *'naive' dynamics*.

### 6.1.2 Dynamic hybrid clustering

A flexible methodology was developed for dynamic clustering, which is an important research topic in the light of dynamic document sets and emerging trend detection. It provided a view on how much attention the bioinformatics community has devoted to different subfields over time. Eleven *cluster chains* could be identified by matching and tracking clusters through time. Most cluster chains corresponded to one bioinformatics cluster. Cross-citations among chains were analyzed in each period to visualize the evolution of citation patterns and, hence, of dynamic interdependencies between chains. The chains that were found by the dynamic procedure might be more accurate than the clusters resulting from the standard algorithm applied to the complete data set at once. The dynamic methodology of tracking clusters through time can be considered less *'hard'* than the standard hierarchical clustering algorithm. Indeed, subsets of documents that do not clearly belong to any of the chains can be discarded. Improvement of the dynamic methodology can still be obtained by using more complex rules for forming cluster chains.

### 6.1.3 Number of clusters and LSI factors

The combined semi-automatic strategy used throughout this dissertation for determining the optimal number of clusters is a combination of distance-based and stability-based methods. To determine the optimal number of clusters with regard to stability, we used the method proposed by *Ben-Hur et al.* [16]. A second opinion was offered by observing the dendrogram in order to find an appropriate cut-off level. In addition, a local maximum was sought in the curves with mean text-based and citation-based Silhouette coefficients for various numbers of clusters. Finally, quality of the ultimate clustering solution was verified in a plot with Silhouette values for all objects.

To overcome the *curse of dimensionality* we used feature selection, Latent Semantic Indexing (LSI) and Random Indexing (RI). Interestingly, LSI and RI to some extent model semantics by mere mathematical processing. We have contributed to an important open research problem in LSI research, namely the debate about the number of LSI factors. We investigated the relationship between number of factors, number of clusters, and clustering performance. In general, for the bioinformatics data set clustering performance was significantly higher for a smaller number of factors. A very modest number of factors delivered local maxima in clustering performance, on condition that there were no fewer LSI factors than the desired number of clusters. However, this should be further assessed using other corpora as well. A limited number of factors has also direct advantages in terms of storage needs and processing time. Our observations are supported by a recent study of *Kontosthatis* [152].

To conclude, statistical and mathematical techniques from text mining, bibliometrics, and link analysis proved very powerful methods for mapping of knowledge embedded in texts. The proposed hybrid clustering algorithms exploiting information from both text world and graph world, possibly complemented with the strategy of tracking clusters through time, provide even more accurate means to unravel the cognitive structure of scientific or technological fields (or of other dynamic document sets), to cast eyes upon the evolution of existing subdisciplines, and to aid detection of emerging or converging clusters and hot topics.

Applications discussed in this dissertation were mainly focused on the clustering of scientific or technological fields. However, most of the algorithms are generic and can be applied in different contexts as well. A straightforward extension is the analysis of sets of Web pages connected by hyperlinks. Other examples for which link structure can be analyzed in combination with textual content are networks of knowledge such as *Wikipedia*, semantic wiki's, Web logs, newsgroups, and e-mail archives. Web pages or documents often consulted together might also be arranged in a network allowing an integrated analysis. Corporate knowledge management might benefit from a hybrid analysis for the demarcation and categorization of available knowledge within companies. Finally, the document need not be the unit of analysis since the methods can also be used to profile and cluster journals, authors, institutions, etc.

## 6.2 Further research

Although a lot of research has already been conducted in the areas of bibliometrics and text mining, paramount challenges still remain regarding algorithms and numerical methods tailored towards the hybrid and dynamic analysis of massive databases. The immense scale necessitates very fast or parallelizable algorithms. Moreover, the very high dimensionality of the data mining problems involved leads to the inherent curse of dimensionality (see Section 2.2.1), which continuously needs to be tackled with efficient large-scale reduction techniques. In addition, stability and robustness of clustering algorithms and dynamic analysis of textual corpora and of networks remain major challenges.

As already mentioned in the previous section, one experience was that (hybrid) clustering algorithms can deliver very good, yet imperfect results. One reason is the intrinsic 'hard' nature of the applied hierarchical clustering algorithm. In order to gain even better performance, a transition should be made towards fuzzy clustering algorithms [134], and to algorithms that restrict the outcome to stable structures. The advantage of a shift from LSI towards PLSI [126] or LDA[28] is also worth considering. Other algorithms might be proposed as well to analyze and cluster data sets based on an integrated textual and graph analytic stance and should be compared with, for example, (mutual) spectral graph algorithms.

It would be interesting to assess the performance of the hybrid clustering methods by using other evaluation measures as well. For instance, *modularity* could be used (see Section 3.5; [202, 201, 199]). In order to use modularity as a quality measure for an integrated clustering, we could consider a network of documents linked by edges that are weighted with the real-valued pairwise document similarities. The modularity matrix can then be constructed from the weighted adjacency (similarity) matrix minus a matrix of the same size containing expected weights between each pair of nodes. For example, to evaluate an integrated clustering found by using *Fisher*'s inverse chi-square method, the elements $B_{j,k}$ of the symmetric modularity matrix $B$ might be computed as follows:

$$B_{j,k} = S_{j,k} - \frac{\sum\limits_{l \neq j} S_{j,l} \cdot \sum\limits_{l \neq k} S_{l,k}}{\sum\limits_{l} \sum\limits_{m \neq l} S_{l,m}}, \tag{6.1}$$

where $S$ is the similarity matrix obtained by taking the complement of each integrated $p$-value $(1 - p_i)$ from the integrated distance matrix $D_i$ (see Figure 4.8 on page 125). Given that *Fisher*'s inverse chi-square method can be used to integrate several information sources (possibly more than two), modularity can thus measure the quality of a clustering of a network with different types of connections, or the quality of a clustering based on combined pairwise similarities between documents.

Another inquiry to pursue is the use of semi-supervised and active learning techniques. Construction of training sets for supervised learning by annotating data is a daunting task in a lot of applications, especially in very large databases. Semi-supervised or active learning can select the most 'valuable' non-annotated examples, most useful for improvement of the model, to present them for human annotation [238, 14, 205, 158].

For scalability towards massive data sets, such as the world's total scientific publication output, we can investigate efficient decompositions of the large involved low-rank matrices by exploiting sparsity and non-negativity. For example, non-negative matrix factorization (NMF) [22, 242], CUR, CX and QR decompositions. Furthermore, sampling schemes can approximate any large matrix by using only small random subsamples of the data. Existing algorithms for (large-scale) clustering have already been described [18, 134, 272]. By making use of computationally less demanding clustering algorithms with much better scaling properties than standard hierarchical clustering, a parallelized large-scale hierarchical clustering is still feasible for annual volumes of the Web of Science. By iteratively and recursively dividing the set in parts, a top-down solution can be implemented, different divisions being 'coarse-grained parallelizable' and performed by independent machines.

The Singular Value Decomposition (SVD) of a matrix has been extended to simultaneously decompose two or more matrices [169, 170, 192, 191, 190, 189]. The generalized SVD has, for example, already been applied in bioinformatics

[7] and in text mining in the context of text categorization [129, 167]. The applicability of GSVD for data integration should be further investigated, for instance, to find common patterns in text and link data.

Another promising path to follow is the shift towards multilinear (tensor) algebra, which provides a very interesting framework for dimensionality reduction, data integration, community structure detection, and for the incorporation of a time dimension for dynamic analyses. One possible multilinear decomposition, known as PARAFAC (PARAllel FACtors) or Candecomp (Canonical Decomposition) [256, 119, 45] has been applied in text mining to the analysis of data with multiple linkages between objects [150]. In this multi-link or higher-order link analysis, PARAFAC allows the incorporation of mutual document similarities of various origins stored in adjacency tensors [72]. The Tucker decomposition is another generalization of the SVD. A 3-way Tucker decomposition might for example be used to analyze *user* × *query_terms* × *web_page* tensors for personalized Web information retrieval [251]. *Author* × *term* × *time* tensors have also been decomposed to separate different streams of conversations [2]. Another interesting algorithm is DEDICOM (DEcomposition into DIrectional COMponents). DEDICOM summarizes a large matrix into a smaller one containing patterns that can be combined to describe many relationships among the components. By using a 3-way extension of DEDICOM, even analysis through time is within reach [8]. DEDICOM might, for example, also be used for trend analysis of cross-cluster citation patterns that might reveal emerging, growing, stable, and fading themes, domains and communities in dynamic databases. Dynamic tensor analyses can detect evolving patterns in time series of graphs.

## 6.3 Perspectives

Science and technology policies increasingly rely on measurements of scientific, technological and innovative activities using an extended set of indicators. In such studies, an accurate demarcation and categorization of scientific and technological fields is needed. This is a possible application area for retrieval and clustering techniques discussed in this dissertation. Additionally, they might assist in the identification of new, emerging and converging fields in science, social sciences and technology by means of combined text-based, bibliometric and graph analytical approaches. The structural topic analysis (by hybrid clustering to unravel the cognitive structure of S&T fields), together with dynamic analysis of citation graphs, publication activity, and citation impact, might allow future studies of dynamics in the structure of science and technology. Comprehension of the internal and interaction dynamics of emerging technologies might contribute to an informed understanding of the co-evolutionary processes in science and technology. The mutual influence and coherence of scientific and technological activities receives a lot of attention from researchers and policy makers. Also in industry, a lot of interest is observable for methods to detect the emergence of fields with potential industrial applications. Often a new hot topic, technology, or discipline emerges from the combination or the mutual influence of several distinct domains, which manifests itself through shifts in literature and underlying citation networks.

Another potential domain of application is corporate knowledge management. A problem that is often encountered by any innovation driven company is the so-called reinvention of the wheel. Effort and money are invested in the development of solutions and technology that are already available within the same or in other companies. Advanced clustering algorithms applied to massive collections of documents found on a company's intranet might provide a mapping of the knowledge that is available within the company.

The continuous rise of computing power might one day allow a large-scale mapping of the scientific universe explorable at various levels of detail. What's more, application of advanced natural language processing and machine summarization at the scale of large bibliographic corpora might offer some insight into semantics beyond mere statistical processing.

# Bibliography

[1] N. Abe and M. Kudo. Non-parametric classifier-independent feature selection. *Pattern Recognition*, 39(5):737–746, 2006.

[2] E. Acar, S. A. Çamtepe, M. Krishnamoorthy, and B. Yener. Modeling and multiway analysis of chatroom tensors. In P. Kantor, G. Muresan, and F. Roberts et al., editors, *Proceedings of the IEEE International Conference on Intelligence and Security Informatics*, volume 3495 of *Lecture Notes in Computer Science*, pages 256–268, 2005.

[3] C. C. Aggarwal, A. Hinneburg, and D. A. Keim. On the surprising behavior of distance metrics in high dimensional spaces. In *ICDT '01: Proceedings of the 8th International Conference on Database Theory*, pages 420–434, London, UK, 2001. Springer-Verlag.

[4] D. W. Aksnes. Characteristics of highly cited papers. *Research Evaluation*, 12(3):159–170, 2003.

[5] D. W. Aksnes and G. Sivertsen. The effect of highly cited papers on national citation indicators. *Scientometrics*, 59(2):213–224, 2004.

[6] R. Albert and A. L. Barabasi. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47–97, 2002.

[7] O. Alter, P. O. Brown, and D. Botstein. Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. *Proceedings of the National Academy of Sciences of the United States of America*, 100(6):3351–3356, 2003.

[8] B. W. Bader, R. Harshman, and T. G. Kolda. Temporal analysis of social networks using three-way dedicom. Technical report SAND2006-2161, Sandia National Laboratories, Albuquerque, New Mexico and Livermore, California, 2006.

[9] R. A. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., 1999.

[10] P. Ball. Index aims for fair ranking of scientists. *Nature*, 436(7053):900, 2005.

[11] A.-L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[12] A. L. Barabasi, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaborations. *Physica A*, 311(3-4):590–614, 2002.

[13] E. Bassecoulard and M. Zitt. Patents and publications: the lexical connection. In H. F. Moed, W. Glänzel, and U. Schmoch, editors, *Handbook of quantitative science and technology research. The use of publication and patent statistics in studies of S&T systems*, pages 665–694. Kluwer Academic Publishers, Dordrecht, 2004.

[14] S. Basu, A. Banerjee, and R. J. Mooney. Semi-supervised clustering by seeding. In *ICML '02: Proceedings of the Nineteenth International Conference on Machine Learning*, pages 27–34, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc.

[15] V. Batagelj and A. Mrvar. Pajek - analysis and visualization of large networks. *Graph Drawing*, 2265:477–478, 2002.

[16] A. Ben-Hur, A. Elisseeff, and I. Guyon. A stability based method for discovering structure in clustered data. In *Pacific Symposium on Biocomputing*, pages 6–17, 2002.

[17] M. K. Bergman. The deep web: Surfacing hidden value. http://www.press.umich.edu/jep/07-01/bergman.html, visited in March 2007.

[18] P. Berkhin. Survey of clustering data mining techniques. Technical report, Accrue Software, San Jose, CA, 2002.

[19] M. W. Berry, editor. *Computational information retrieval.* Society for Industrial and Applied Mathematics, 2001.

[20] M. W. Berry, editor. *Survey of Text Mining.* Springer-Verlag New York, Inc., 2003.

[21] M. W. Berry and M. Browne. *Understanding search engines: mathematical modeling and text retrieval.* Society for Industrial and Applied Mathematics, 1999.

[22] M. W. Berry, M. Browne, A. M. Langville, P. V. Pauca, and R. J. Plemmons. Algorithms and applications for approximate nonnegative matrix factorization. *Preprint submitted to Elsevier Preprint*, June 2006.

[23] M. W. Berry, S. T. Dumais, and G. W. Obrien. Using linear algebra for intelligent information retrieval. *SIAM Review*, 37(4):573–595, 1995.

[24] M. W. Berry, S. T. Dumais, and A. T. Shippy. A case study of latent semantic indexing. Technical report UT-CS-95-271, Knoxville, TN, USA, 1995.

[25] K. Bharat and M. R. Henzinger. Improved algorithms for topic distillation in a hyperlinked environment. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 104–111, New York, NY, USA, 1998. ACM Press.

[26] S. Bhattacharya and P. K. Basu. Mapping a research area at the micro level using co-word analysis. *Scientometrics*, 43(3):359–372, 1998.

[27] G. Bianconi and A. L. Barabasi. Competition and multiscaling in evolving networks. *Europhysics Letters*, 54(4):436–442, 2001.

[28] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(4-5):993–1022, 2003.

[29] V. D. Blondel, A. Gajardo, M. Heymans, P. Senellart, and P. Van Dooren. A measure of similarity between graph vertices: Applications to synonym extraction and web searching. *SIAM Review*, 46(4):647–666, 2004.

[30] E. Bonnevie. A multifaceted portrait of a library and information science journal: the case of the Journal of Information Science. *Journal of Information Science*, 29(1):11–23, 2003.

[31] M. Bordons, M. T. Fernandez, and I. Gomez. Advantages and limitations in the use of impact factor measures for the assessment of research performance in a peripheral country. *Scientometrics*, 53(2):195–206, 2002.

[32] C. L. Borgman and J. Furner. Scholarly communication and bibliometrics. *Annual Review of Information Science and Technology*, 36:3–72, 2002.

[33] K. Borner, J. T. Maru, and R. L. Goldstone. The simultaneous evolution of author and paper networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101:5266–5273, 2004.

[34] R. R. Braam, H. F. Moed, and A. F. J. van Raan. Mapping of science by combined cocitation and word analysis .2. dynamic aspects. *Journal of the American Society for Information Science*, 42(4):252–266, 1991.

[35] U. Brandes and T. Erlebach, editors. *Network Analysis : Methodological Foundations (Lecture Notes in Computer Science).* Springer, March 2005.

[36] T. Braun and W. Glänzel. Chemistry research in Eastern Central Europe (1992–1997) - facts and figures on publication output and citation impact. *Scientometrics*, 49(2):187–213, 2000.

[37] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117, 1998.

[38] Q. L. Burrell. Hirsch's h-index: A stochastic model. *Journal of Informetrics*, 1(1):16–25, 2007.

[39] R. Burt. *Structural holes.* Harvard University Press, 1992.

[40] P. Calado, M. Cristo, M. A. Gonçalves, E. S. de Moura, B. Ribeiro-Neto, and N. Ziviani. Link-based similarity measures for the classification of web documents. *Journal of the American Society for Information Science and Technology*, 57(2):208–221, 2006.

[41] P. Calado, M. Cristo, E. Moura, N. Ziviani, B. Ribeiro-Neto, and M. A. Goncalves. Combining link-based and content-based methods for web document classification. In *CIKM '03: Proceedings of the twelfth international conference on information and knowledge management*, pages 394–401, New York, NY, USA, 2003. ACM Press.

[42] P. Calado, B. Ribeiro-Neto, N. Ziviani, E. Moura, and I. Silva. Local versus global link information in the web. *ACM Transactions on Information Systems*, 21(1):42–63, 2003.

[43] M. Callon, J. P. Courtial, and F. Laville. Co-word analysis as a tool for describing the network of interactions between basic and technological research - the case of polymer chemistry. *Scientometrics*, 22(1):155–205, 1991.

[44] M. Callon, J. P. Courtial, W. A. Turner, and S. Bauin. From translations to problematic networks - an introduction to co-word analysis. *Social Science Information Sur les Sciences Sociales*, 22(2):191–235, 1983.

[45] J. D. Carroll and J. J. Chang. Analysis of individual differences in multidimensional scaling via an n-way generalization of Eckart-Young decomposition. *Psychometrika*, 35:283–319, 1970.

[46] S. Chakrabarti, B. Dom, P. Raghava, S. Rajagopalan, D. Gibson, and J. Kleinberg. Automatic resource compilation by analyzing hyperlink structure and associated text. *Computer Networks and ISDN Systems*, 30(1-7):65–74, 1998.

[47] C. Chen and S. Morris. Visualizing evolving networks: Minimum spanning trees versus pathfinder networks. In *InfoVis 2003: 9th IEEE Symposium on Information Visualization*, Los Alamitos, CA, USA, 2003. IEEE Computer Society.

[48] C. M. Chen. Searching for intellectual turning points: Progressive knowledge domain visualization. *Proceedings of the National Academy of Sciences of the United States of America*, 101:5303–5310, 2004.

[49] MEDLINE citation counts by year of publication. http://www.nlm.nih.gov/bsd/medline_cit_counts_yr_pub.html, visited in march 2007.

[50] D. Cohn and H. Chang. Learning to probabilistically identify authoritative documents. In *ICML '00: Proceedings of the Seventeenth International Conference on Machine Learning*, pages 167–174, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

[51] D. Cohn and T. Hofmann. The missing link - a probabilistic model of document content and hypertext connectivity. In *Neural Information Processing Systems 13*, 2001.

[52] S. Cole and J. Cole. Scientific output and recognition: A study in the operation of the reward system in science. *American Sociological Review*, 32:377–390, 1967.

[53] J. P. Courtial. Qualitative models, quantitative tools and network analysis. *Scientometrics*, 15(5-6):527–534, 1989.

[54] J. P. Courtial. A coword analysis of Scientometrics. *Scientometrics*, 31(3):251–260, 1994.

[55] W. Daelemans and A. van den Bosch. *Memory-Based Language Processing (Studies in Natural Language Processing)*. Cambridge University Press, 2005.

[56] B. Dawson-Saunders and R.G. Trapp. *Basic & Clinical Biostatistics*. Prentice-Hall International Inc., 1994.

[57] M. A. de Looze and J. Lemarie. Corpus relevance through co-word analysis: An application to plant proteins. *Scientometrics*, 39(3):267–280, 1997.

[58] D. J. de Solla Price. *Little Science, Big Science*. Columbia University Press, 1963.

[59] D. J. de Solla Price. Networks of scientific papers. *Science*, 149(3683):510–515, 1965.

[60] D. J. de Solla Price. A general theory of bibliometrics and other cumulative advantage processes. *Journal of the American Informatics Society*, 27:292–306, 1980.

[61] K. Debackere and R. Veugelers. The role of academic technology transfer organizations in improving industry science links. *Research Policy*, 34(3):321–342, April 2005.

[62] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[63] I. S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge Discovery and Data Mining (KDD '01)*, pages 269–274, New York, NY, USA, 2001. ACM Press.

[64] I. S. Dhillon, S. Mallela, and R. Kumar. A divisive information theoretic feature clustering algorithm for text classification. *Journal of Machine Learning Research*, 3:1265–1287, 2003.

[65] I. S. Dhillon and D. S. Modha. Concept decompositions for large sparse text data using clustering. *Machine Learning*, 42(1/2):143–175, 2001.

[66] C. Ding, X. He, H. Zha, and H. D. Simon. Adaptive dimension reduction for clustering high dimensional data. In *ICDM '02: Proceedings of the 2002 IEEE International Conference on Data Mining*, pages 147–154, Washington, DC, USA, 2002. IEEE Computer Society.

[67] C. H. Q. Ding, H. Y. Zha, X. F. He, P. Husbands, and H. D. Simon. Link analysis: Hubs and authorities on the world wide web. *SIAM Review*, 46(2):256–268, 2004.

[68] Y. Ding, G. Chowdhury, and S. Foo. Mapping the intellectual structure of information retrieval studies: an author co-citation analysis, 1987-1997. *Journal of Information Science*, 25(1):67–78, 1999.

[69] Y. Ding, G. G. Chowdhury, and S. Foo. Bibliometric cartography of information retrieval research by using co-word analysis. *Information Processing & Management*, 37(6):817–842, 2001.

[70] W. E. Donath and A. J. Hoffman. Algorithms for partitioning of graphs and computer logic based on eigenvectors of connection matrices. Technical report, IBM, 1972.

[71] S. N. Dorogovtsev and J. F. F. Mendes. Evolution of networks. *Advances in Physics*, 51(4):1079–1187, 2002.

[72] D. M. Dunlavy, T. G. Kolda, and W. P. Kegelmeyer. Multilinear algebra for analyzing data with multiple linkages. Technical report SAND2006-2079, Sandia National Laboratories, Albuquerque, New Mexico and Livermore, California, 2006.

[73] T. Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74, 1993.

[74] B. Dutt, K. C. Garg, and A. Bali. Scientometrics of the international journal Scientometrics. *Scientometrics*, 56(1):81–93, 2003.

[75] C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1:211–218, 1936.

[76] Miles Efron. Eigenvalue-based model selection during latent semantic indexing. *Journal of the American Society for Information Science and Technology*, 56(9):969–988, 2005.

[77] L. Egghe and R. Rousseau. An informetric model for the hirsch-index. *Scientometrics*, 69(1):121–129, 2006.

[78] A. J. Enright and C. A. Ouzounis. Biolayout - an automatic graph layout algorithm for similarity visualization. *Bioinformatics*, 17(9):853–854, 2001.

[79] European Patent Office (EPO). Annual report 2005, http://annual-report.european-patent-office.org/2005/_pdf/epo_anrep05.pdf, visited in march 2007.

[80] G. Erkan and D. R. Radev. Lexrank: Graph-based centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479, 2004.

[81] E. Erosheva, S. Fienberg, and J. Lafferty. Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences of the United States of America*, 101:5220–5227, 2004.

[82] M. J. Fisher and R. M. Everson. When are links useful? Experiments in text classification. In *ECIR '03: Proceedings of the 25th European Conference on IR Research*, pages 41–56, 2003.

[83] G. W. Flake, S. Lawrence, and C. L. Giles. Efficient identification of web communities. In *KDD '00: Proceedings of the sixth ACM SIGKDD international conference on Knowledge Discovery and Data Mining*, pages 150–160, New York, NY, USA, 2000. ACM Press.

[84] J. H. Friedman. On bias, variance, 0/1-loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery*, 1(1):55–77, 1997.

[85] E. Garfield. Citation indexes to science: A new dimension in documentation through the association of ideas. *Science*, 22:108–111, 1955.

[86] E. Garfield. Citation indexing for studying science. *Nature*, 227:669–671, 1970.

[87] E. Garfield. *Citation indexing: its theory and applications in science, technology and humanities*. Wiley, 1979.

[88] E. Garfield. Long-term vs. short-term journal impact: Does it matter? *Scientist*, 12(3):11–12, 1998.

[89] E. Garfield. The history and meaning of the journal impact factor. *Journal of the American Medical Association*, 295(1):90–93, 2006.

[90] É. Gauthier. Bibliometric analysis of scientific and technological research: A users guide to the methodology. Technical report ST-98-008, Statistics Canada, Science and Innovation Surveys Section, Ottawa, Ontario, Canada, 1998.

[91] B. Gay and B. Dousset. Innovation and network structural dynamics: Study of the alliance network of a major sector of the biotechnology industry. *Research Policy*, 34(10):1457–1475, 2005.

[92] D. Gibson, J. Kleinberg, and P. Raghavan. Inferring web communities from link topology. In *HYPERTEXT '98: Proceedings of the ninth ACM conference on Hypertext and Hypermedia : links, objects, time and space—structure in hypermedia systems*, pages 225–234, New York, NY, USA, 1998. ACM Press.

[93] W. Glänzel. On the possibility and reliability of predictions based on stochastic citation processes. *Scientometrics*, 40(3):481–492, 1997.

[94] W. Glänzel. Science in Scandinavia: A bibliometric approach. *Scientometrics*, 48(2):121–150, 2000.

[95] W. Glänzel. National characteristics in international scientific co-authorship relations. *Scientometrics*, 51(1):69–115, 2001.

[96] W. Glänzel. Bibliometrics as a research field. A course on theory and application of bibliometric indicators. Course script, Katholieke Universiteit Leuven, Leuven, Belgium, 2005.

[97] W. Glänzel. On the h-index - a mathematical approach to a new measure of publication activity and citation impact. *Scientometrics*, 67(2):315–321, 2006.

[98] W. Glänzel. (on the opportunities and limitations of the h-index, in chinese). *Science Focus*, 1(1):10–11, 2006.

[99] W. Glänzel and H. J. Czerwon. A new methodological approach to bibliographic coupling and its application to the national, regional and institutional level. *Scientometrics*, 37(2):195–221, 1996.

[100] W. Glänzel, K. Debackere, and M. Meyer. 'Triad' or 'tetrad'? on global changes in a dynamic world. *To be published in Scientometrics*, 2007.

[101] W. Glänzel, F. Janssens, S. Speybroeck, A. Schubert, and B. Thijs. Towards a bibliometrics-aided data retrieval for scientometric purposes. In *Book of abstracts of the 9th International Conference on Science & Technology Indicators*, Leuven, Belgium, 2006. Katholieke Universiteit Leuven.

[102] W. Glänzel, F. Janssens, and B. Thijs. A comparative analysis of publication activity and citation impact based on the core literature in bioinformatics. 2007.

[103] W. Glänzel, M. Meyer, M. Du Plessis, B. Thijs, T. Magerman, B. Schlemmer, K. Debackere, and R. Veugelers. Nanotechnology - analysis of an emerging domain of scientific and technological endeavour. Technical report, Steunpunt O&O Statistieken, Leuven, Belgium, 2003.

[104] W. Glänzel, M. Meyer, B. Schlemmer, M. Du Plessis, B. Thijs, T. Magerman, K. Debackere, and R. Veugelers. Biotechnology: An analysis based on publications and patents. Technical report, Steunpunt O&O Statistieken, Leuven, Belgium, 2003.

[105] W. Glänzel and U. Schoepflin. Little scientometrics, big scientometrics ... and beyond. *Scientometrics*, 30(2-3):375–384, 1994.

[106] W. Glänzel, A. Schubert, and T. Braun. A relational charting approach to the world of basic research in twelve science fields at the end of the second millennium. *Scientometrics*, 55(3):335–348, 2002.

[107] W. Glänzel, A. Verbeek, M. Du Plessis, B. Van Looy, T. Magerman, B. Thijs, B. Schlemmer, K. Debackere, and R. Veugelers. Stem cells - analysis of an emerging domain of scientific and technological endeavour. Technical report, Steunpunt O&O Statistieken, Leuven, Belgium, 2004.

[108] P. Glenisson. *Integrating scientific literature with large scale gene expression analysis.* Ph.D. thesis, Faculty of Engineering. Katholieke Universiteit Leuven, Belgium, 2004.

[109] P. Glenisson, W. Glänzel, F. Janssens, and B. De Moor. Combining full text and bibliometric information in mapping scientific disciplines. *Information Processing & Management*, 41(6):1548–1572, 2005.

[110] P. Glenisson, W. Glänzel, and O. Persson. Combining full-text analysis and bibliometric indicators. A pilot study. *Scientometrics*, 63(1):163–180, 2005.

[111] G. Golub and C. Van Loan. *Matrix Computations.* John Hopkins University Press, third edition, 1996.

[112] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101:5228–5235, 2004.

[113] B. I. Groenen and J. F. Patrick. *Modern Multidimensional Scaling. Theory and Applications.* Springer, 2005.

[114] P. Grzybek and E. Kelih. Anton S. Budilovic (1846-1908) - a forerunner of quantitative linguistics in Russia? *Glottometrics*, 7:94–97, 2004.

[115] W. Hagstrom. *The scientific community.* Basic Books, 1965.

[116] M. Halkidi, Y. Batistakis, and M. Vazirgiannis. On clustering validation techniques. *Journal of Intelligent Information Systems*, 17(2-3):107–145, 2001.

[117] D. Hand, H. Mannila, and P. Smyth. *Principles of Data Mining.* Cambridge, MA: MIT Press, 2001.

[118] J. A. Hanley and B. J. Mcneil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, 1982.

[119] R. Harshman. Foundations of the PARAFAC procedure: Models and conditions for an 'explanatory' multi-modal factor analysis. *UCLA Working Papers in Phonetics*, 16:1–84, 1970.

[120] E. Hatcher and O. Gospodnetic. *Lucene in Action.* Manning Publications Co., 2004.

[121] S. Y. He and A. Spink. A comparison of foreign authorship distribution in JASIST and the Journal of Documentation. *Journal of the American Society for Information Science and Technology*, 53(11):953–959, 2002.

[122] X. He, H. Zha, C. H. Q. Ding, and H. D. Simon. Web document clustering using hyperlink structures. *Computational Statistics & Data Analysis*, 41(1):19–45, November 2002.

[123] L. V. Hedges and I. Olkin. *Statistical Methods for Meta-Analysis.* Academic Press, 1985.

[124] J. E. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102(46):16569–16572, 2005.

[125] T. Hofmann. Probabilistic latent semantic analysis. In *UAI'99: Proceedings of Uncertainty in Artificial Intelligence*, Stockholm, 1999.

[126] T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57, New York, NY, USA, 1999. ACM Press.

[127] J. Hopcroft, O. Khan, B. Kulis, and B. Selman. Natural communities in large linked networks. In *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge Discovery and Data Mining*, pages 541–546, New York, NY, USA, 2003. ACM Press.

[128] J. Hopcroft, O. Khan, B. Kulis, and B. Selman. Tracking evolving communities in large linked networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101:5249–5253, 2004.

[129] P. Howland, M. Jeon, and H. Park. Structure preserving dimension reduction for clustered text data based on the generalized singular value decomposition. *SIAM Journal on Matrix Analysis and Applications*, 25(1):165–179, 2003.

[130] C.-J. Huang and W.-C. Liao. Application of probabilistic neural networks to the class prediction of leukemia and embryonal tumor of central nervous system. *Neural Processing Letters*, 19(3):211–226, 2004.

[131] L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2(2-3):193–218, 1985.

[132] A. B. Jaffe and M. Trajtenberg. *Patents, Citations & Innovations: A Window on the Knowledge Economy.* MIT Press, 2002.

[133] A. K. Jain and R. C. Dubes. *Algorithms for clustering data.* Prentice-Hall, Inc., 1988.

[134] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: A review. *ACM Computing Surveys*, 31(3):264–323, 1999.

[135] F. Janssens, W. Glänzel, and B. De Moor. A hybrid mapping of information science. 2007.

[136] F. Janssens, P. Glenisson, W. Glänzel, and B. De Moor. Co-clustering approaches to integrate lexical and bibliographical information. In P. Ingwersen and B. Larsen, editors, *Proceedings of the 10th international conference of the International Society for Scientometrics and Informetrics (ISSI)*, volume 1, pages 284–289, Stockholm, Sweden, July 2005. Karolinska University Press.

[137] F. Janssens, J. Leta, W. Glänzel, and B. De Moor. Towards mapping library and information science. *Information Processing & Management*, 42(6):1614–1642, 2006.

[138] F. Janssens and B. De Moor. Application of HITS algorithms to detect terms and sentences with high saliency scores. Technical report 04-29, ESAT-SISTA, K.U.Leuven, Leuven, Belgium, 2003.

[139] F. Janssens, V. Tran Quoc, W. Glänzel, and B. De Moor. Integration of textual content and link information for accurate clustering of science fields. volume I of *Proceedings of the I International Conference on Multidisciplinary Information Sciences & Technologies (InSciT2006). Current Research in Information Sciences and Technologies*, pages 615–619, Mérida, Spain, October 2006.

[140] H. Jeong, Z. Neda, and A. L. Barabasi. Measuring preferential attachment in evolving networks. *Europhysics Letters*, 61(4):567–572, 2003.

[141] T. Joachims, N. Cristianini, and J. Shawe-Taylor. Composite kernels for hypertext categorisation. In *ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning*, pages 250–257, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[142] W. Johnson and J. Lindenstrauss. Extension of Lipshitz mapping to Hilbert space. *Contemporary Math.*, 26:189–206, 1984.

[143] P. Kanerva, J. Kristofersson, and A. Holst. Random Indexing of text samples for latent semantic analysis, 2000.

[144] J. Karlgren and M. Sahlgren. From words to understanding. In Y. Uesaka, P. Kanerva, and H. Asoh, editors, *Foundations of Real-World Intelligence*, pages 294–308. CSLI Publications, 2001.

[145] S. Kaski. Dimensionality reduction by random mapping: Fast similarity computation for clustering. In *IJCNN'98: Proceedings of the International Joint Conference on Neural Networks*, pages 413–418, 1998.

[146] L. Kaufman and P. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis.* John Wiley and Sons Inc., 1990.

[147] M. M. Kessler. Bibliographic coupling between scientific papers. *American Documentation*, 14:10–25, 1963.

[148] J. Kleinberg. The small-world phenomenon: An algorithmic perspective. Technical report, Ithaca, NY, USA, 1999.

[149] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[150] T. G. Kolda, B. W. Bader, and J.P. Kenny. Higher-order web link analysis using multilinear algebra. In *ICDM '05: Proceedings of the Fifth IEEE International Conference on Data Mining*, pages 242–249, Washington, DC, USA, 2005. IEEE Computer Society.

[151] A. Kontostathis and W. M. Pottenger. A framework for understanding Latent Semantic Indexing (LSI) performance. *Information Processing & Management*, 42(1):56–73, 2006.

[152] A. Kontostathis and W. M. Pottenger. Essential Dimensions of Latent Semantic Indexing (EDLSI). In *Proceedings of the 40th Annual Hawaii International Conference on System Sciences (CD-ROM). Computer Society Press. To appear*, 2007.

[153] R. N. Kostoff. The use and misuse of citation analysis in research evaluation - comments on theories of citation? *Scientometrics*, 43(1):27–43, 1998.

[154] R. N. Kostoff and J. A. Block. Factor matrix text filtering and clustering: Research articles. *Journal of the American Society for Information Science and Technology*, 56(9):946–968, 2005.

[155] R. N. Kostoff, H. A. Buchtel, J. Andrews, and K. M. Pfeil. The hidden structure of neuropsychology: Text mining of the journal Cortex: 1991-2001. *Cortex*, 41(2):103–115, 2005.

[156] R. N. Kostoff, D. R. Toothman, H. J. Eberhart, and J. A. Humenik. Text mining using database tomography and bibliometrics: A review. *Technological Forecasting and Social Change*, 68(3):223–253, 2001.

[157] H. Kretschmer. Author productivity and geodesic distance in bibliographic coauthorship networks, and visibility on the web. *Scientometrics*, 60(3):409–420, 2004.

[158] A. Krithara, C. Goutte, M. Amini, and J. Renders. Active, semi-supervised learning for textual information access. In *IIIA '06: International Workshop on Intelligent Information Access*, Helsinki, Finland, July 2006.

[159] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the web for emerging cyber-communities. In *WWW '99: Proceeding of the eighth international conference on World Wide Web*, pages 1481–1493, New York, NY, USA, 1999. Elsevier North-Holland, Inc.

[160] H. O. Lancaster. The combination of probabilities arising from data in discrete distributions. *Biometrika*, 36:370–382, 1949.

[161] A. N. Langville and C. D. Meyer. A survey of eigenvector methods for web information retrieval. *SIAM Review*, 47(1):135–161, 2005.

[162] R. Lempel and S. Moran. SALSA: the stochastic approach for link-structure analysis. *ACM transactions on information systems*, 19(2):131–160, 2001.

[163] E. Leopold, M. May, and G. Paaß. Data mining and text mining for science & technology research. In H. F. Moed, W. Glänzel, and U. Schmoch, editors, *Handbook of quantitative science and technology research. The use of publication and patent statistics in studies of S&T systems*, pages 187–213. Kluwer Academic Publishers, Dordrecht, 2004.

[164] L. Leydesdorff. Why words and co-words cannot map the development of the sciences. *Journal of the American Society for Information Science*, 48(5):418–427, 1997.

[165] L. Leydesdorff. Theories of citation? *Scientometrics*, 43(1):5–25, 1998.

[166] L. Leydesdorff. The university-industry knowledge relationship: Analyzing patents and the science base of technologies. *Journal of the American Society for Information Science and Technology*, 55(11):991–1001, 2004.

[167] T. Li, S. Zhu, and M. Ogihara. Efficient multi-way text categorization via generalized discriminant analysis. In *CIKM '03: Proceedings of the twelfth international Conference on Information and Knowledge Management*, pages 317–324, New York, NY, USA, 2003. ACM Press.

[168] Professor Lennart Ljung. http://www.control.isy.liu.se/∼ljung/, visited in march 2007.

[169] C. F. Van Loan. A general matrix eigenvalue algorithm. *SIAM Journal on Matrix Analysis and Applications*, 12(6):819–834, December 1975.

[170] C. F. Van Loan. Generalizing the singular value decomposition. *SIAM Journal on Numerical Analysis*, 13(1):76–83, 1976.

[171] B. Van Looy, E. Zimmermann, R. Veugelers, A. Verbeek, J. Mello, and K. Debackere. Do science-technology interactions pay off when developing technology? An exploratory investigation of 10 science-intensive technology domains. *Scientometrics*, 57(3):355–367, 2003.

[172] A. J. Lotka. The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences*, 16:317–323, 1926.

[173] P. Lyman, H. R. Varian, K. Swearingen, P. Charles, N. Good, L. L. Jordan, and J. Pal. How much information? 2003. http://www.sims.berkeley.edu/research/projects/how-much-info-2003/printable_report.pdf, visited in March 2007.

[174] C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, 1999.

[175] K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate analysis*. Academic Press, Harcourt Brace & Co, 1979.

[176] I. V. Marshakova. Journal co-citation analysis in the field of information science and library science. In P. Nowak and M. Gorny, editors, *Language, information and communication studies*, pages 87–96. Adam Mieckiewicz University, Poznan, 2003.

[177] I. Marshakova-Shaikevich. Bibliometric maps of field of science. *Information Processing & Management*, 41(6):1534–1547, 2005.

[178] F. Menczer. Evolution of document networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101:5261–5265, 2004.

[179] R. K. Merton. The Matthew Effect in science. *Science*, 57:68–72, 1968.

[180] R. K. Merton. *The Sociology of science: Theoretical and empirical investigations*. University of Chicago Press, 1973.

[181] J. Michel and B. Bettels. Patent citation analysis - a closer look at the basic input data from patent search reports. *Scientometrics*, 51(1):185–201, 2001.

[182] S. Milgram. The small world problem. *Psychology Today*, 67(1), 1967.

[183] G. W. Milligan and M. C. Cooper. A study of the comparability of external criteria for hierarchical cluster-analysis. *Multivariate Behavioral Research*, 21(4):441–458, 1986.

[184] T. M. Mitchell. *Machine Learning*. McGraw-Hill, 1997.

[185] D. S. Modha and W. Scott Spangler. Clustering hypertext with applications to web searching. In *HYPERTEXT '00: Proceedings of the eleventh ACM on Hypertext and Hypermedia*, pages 143–152, New York, NY, USA, 2000. ACM Press.

[186] H. F. Moed, T. N. van Leeuwen, and J. Reedijk. Towards appropriate indicators of journal impact. *Scientometrics*, 46(3):575–589, 1999.

[187] M.-F. Moens. *Automatic Indexing and Abstracting of Document Texts*. Kluwer Academic Publishers, 2000. (The Kluwer International Series on Information Retrieval 6).

[188] M.-F. Moens, C. Uyttendaele, and J. Dumortier. Abstracting of legal cases: the potential of clustering based on the selection of representative objects. *Journal of the American Society for Information Science*, 50(2):151–161, 1999.

[189] B. De Moor. On the structure of generalized singular value and QR decompositions. *SIAM Journal on Matrix Analysis and Applications*, 15(1):347–358, 1994.

[190] B. De Moor and P. Van Dooren. Generalizations of the singular value and QR decompositions. *SIAM Journal on Matrix Analysis and Applications*, 13(4):993–1014, 1992.

[191] B. De Moor and H. Zha. A tree of generalizations of the ordinary singular value decomposition. *Linear Algebra and its Applications*, 147:469–500, 1991.

[192] B. L. R. De Moor and G. H. Golub. The restricted singular value decomposition - properties and applications. *SIAM Journal on Matrix Analysis and Applications*, 12(3):401–425, 1991.

[193] S. A. Morris. Manifestation of emerging specialties in journal literature: A growth model of papers, references, exemplars, bibliographic coupling, cocitation, and clustering coefficient distribution. *Journal of the American Society for Information Science and Technology*, 56(12):1250–1273, 2005.

[194] S. A. Morris, G. Yen, Z. Wu, and B. Asnake. Time line visualization of research fronts. *Journal of the American Society for Information Science and Technology*, 54(5):413–422, 2003.

[195] S. A. Morris and G. G. Yen. Crossmaps: Visualization of overlapping relationships in collections of journal papers. *Proceedings of the National Academy of Sciences of the United States of America*, 101:5291–5296, 2004.

[196] N. Mullins, W. Snizek, and K. Oehler. The structural analysis of a scientific paper. In A. F. J. van Raan, editor, *Handbook of quantitative studies of science and technology*, pages 81–105. Elsevier Science, New York, 1988.

[197] M. E. J. Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences of the United States of America*, 98(2):404–409, 2001.

[198] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003.

[199] M. E. J. Newman. Analysis of weighted networks. *Physical Review E*, 70(5), 2004.

[200] M. E. J. Newman. Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences of the United States of America*, 101:5200–5205, 2004.

[201] M. E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74(3), 2006.

[202] M. E. J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 103(23):8577–8582, 2006.

[203] A. Y. Ng, A. X. Zheng, and M. I. Jordan. Link analysis, eigenvectors and stability. In B. Nebel, editor, *IJCAI 2001: Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, pages 903–910, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[204] A. Y. Ng, A. X. Zheng, and M. I. Jordan. Stable algorithms for link analysis. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 258–266, New York, NY, USA, 2001. ACM Press.

[205] H. T. Nguyen and A. Smeulders. Active learning using pre-clustering. In *ICML '04: Proceedings of the twenty-first international conference on Machine Learning*, page 79, New York, NY, USA, 2004. ACM Press.

[206] E. Noyons. Bibliometric mapping of science in a science policy context. *Scientometrics*, 50(1):83–98, 2001.

[207] E. C. M. Noyons. *Bibliometric mapping as a science policy and research management tool*. DSWO Press, 1999.

[208] E. C. M. Noyons and A. F. J. van Raan. Bibliometric cartography of scientific and technological developments of an research-and-development field - the case of optomechatronics. *Scientometrics*, 30(1):157–173, 1994.

[209] O. B. Onyancha and D. N. Ocholla. An informetric investigation of the relatedness of opportunistic infections to HIV/AIDS. *Information Processing & Management*, 41(6):1573–1588, 2005.

[210] E. J. Otoo, A. Shoshani, and S.-W. Hwang. Clustering high dimensional massive scientific datasets. *Journal of Intelligent Information Systems*, 17(2-3):147–168, 2001.

[211] C. A. Ouzounis and A. Valencia. Early bioinformatics: the birth of a discipline - a personal view. *Bioinformatics*, 19(17):2176–2190, 2003.

[212] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the Web. Technical report, Stanford Digital Library Technologies Project, 1998.

[213] G. Palla, A. L. Barabasi, and T. Vicsek. Quantifying social group evolution. *Nature*, 446(7136):664–667, 2007.

[214] C. H. Papadimitriou, P. Raghavan, H. Tamaki, and S. Vempala. Latent semantic indexing: a probabilistic analysis. *Journal of Computer and System Sciences*, 61(2):217–235, October 2000.

[215] L. Parsons, E. Haque, and H. Liu. Subspace clustering for high dimensional data: a review. *SIGKDD Explorer Newsletter*, 6(1):90–105, 2004.

[216] S. K. Patra and S. Mishra. Bibliometric study of bioinformatics literature. *Scientometrics*, 67(3):477–489, 2006.

[217] E. S. Pearson. On questions raised by the combination of tests based on discontinous distributions. *Biometrika*, 37:383–398, 1950.

[218] C. Perez-Iratxeta, M. A. Andrade-Navarro, and J. D. Wren. Evolving research trends in bioinformatics. *Briefings in Bioinformatics*, 2006.

[219] B. C. Peritz. On the objectives of citation analysis - problems of theory and method. *Journal of the American Society for Information Science*, 43(6):448–451, 1992.

[220] O. Persson. A tribute to Eugene Garfield - discovering the intellectual base of his discipline. *Current Science*, 79(5):590–591, 2000.

[221] O. Persson. All author citations versus first author citations. *Scientometrics*, 50(2):339–344, 2001.

[222] G. Pivovarov and S. T. Seus. EqRank: a self-consistent equivalence relation on graph vertexes. *SIGKDD Explor. Newsl.*, 5(2):185–190, 2003.

[223] N. L. M. M. Pochet, F. A. L. Janssens, F. De Smet, K. Marchal, J. A. K. Suykens, and B. L. R. De Moor. M@CBETH: a microarray classification benchmarking tool. *Bioinformatics*, 21(14):3185–3186, 2005.

[224] A. L. Porter and N. C. Newman. Patent profiling for competitive advantage. In H. F. Moed, W. Glänzel, and U. Schmoch, editors, *Handbook of quantitative science and technology research. The use of publication and patent statistics in studies of S&T systems*, pages 587–612. Kluwer Academic Publishers, Dordrecht, 2004.

[225] M. F. Porter. An algorithm for suffix stripping. *Program-Automated Library and Information Systems*, 14(3):130–137, 1980.

[226] W. W. Powell, K. W. Koput, and L. Smith-Doerr. Interorganizational collaboration and the locus of innovation: Networks of learning in biotechnology. *Administrative Science Quarterly*, 41(1):116–145, 1996.

[227] S. Redner. How popular is your paper? An empirical study of the citation distribution. *European Physical Journal B*, 4(2):131–134, 1998.

[228] A. Rip and J. P. Courtial. Co-word maps of biotechnology - an example of cognitive scientometrics. *Scientometrics*, 6(6):381–400, 1984.

[229] S. Robinson. The ongoing search for efficient web search algorithms. *SIAM News*, 37(9), 2004.

[230] V. Rodriguez. Material transfer agreements: open science vs. proprietary claims. *Nature Biotechnology*, 23(4):489–491, 2005.

[231] V. Rodriguez, F. Janssens, K. Debackere, and B. De Moor. Do material transfer agreements affect the choice of research agendas? The case of biotechnology in Belgium. *Scientometrics*, 71(2):239–269, 2007.

[232] P. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(1):53–65, 1987.

[233] M. Sahlgren. Towards a flexible model of word meaning. In *Proceedings of the AAAI Spring Symposium 2002, Stanford University, Palo Alto, California, USA*. 2002.

[234] M. Sahlgren. An introduction to Random Indexing. In H. Witschel, editor, *Proceedings of Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering (TKE'05)*, 2005.

[235] M. Sahlgren and R. Cöster. Using bag-of-concepts to improve the performance of support vector machines in text categorization. In *COLING'04: Proceedings of the 20th International Conference on Computational Linguistics*, pages 487–493, 2004.

[236] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., 1986.

[237] U. Schoepflin and W. Glänzel. Two decades of 'scientometrics' - an interdisciplinary field represented by its leading journal. *Scientometrics*, 50(2):301–312, 2001.

[238] G. Schohn and D. Cohn. Less is more: Active learning with support vector machines. In *ICML '00: Proceedings of the Seventeenth International Conference on Machine Learning*, pages 839–846, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

[239] A. Schubert. The web of Scientometrics - a statistical overview of the first 50 volumes of the journal. *Scientometrics*, 53(1):3–20, 2002.

[240] A. Schubert and H. Maczelka. Cognitive changes in scientometrics during the 1980s, as reflected by the reference patterns of its core journal. *Social Studies of Science*, 23(3):571–581, 1993.

[241] P. K. Shah, C. Perez-Iratxeta, P. Bork, and M. A. Andrade. Information extraction from full text scientific articles: Where are the keywords? *BMC Bioinformatics*, 4, 2003.

[242] F. Shahnaz, M. W. Berry, V. P. Pauca, and R. J. Plemmons. Document clustering using nonnegative matrix factorization. *Information Processing & Management*, 42(2):373–386, 2006.

[243] H. Small. Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24:265–269, 1973.

[244] H. Small. Co-citation in scientific literature - new measure of relationship between 2 documents. *Current Contents*, (7):7–10, 1974.

[245] H. Small. On the shoulders of Robert Merton: Towards a normative theory of citation. *Scientometrics*, 60(1):71–79, 2004.

[246] H. Small and B. C. Griffith. Structure of scientific literatures .1. identifying and graphing specialties. *Science Studies*, 4(1):17–40, 1974.

[247] W. E. Snizek, K. Oehler, and N. C. Mullins. Textual and nontextual characteristics of scientific papers - neglected science indicators. *Scientometrics*, 20(1):25–35, 1991.

[248] B. Stefaniak. Periodical literature of information-science as reflected in referativnyj-zhurnal, section-59, informatika. *Scientometrics*, 7(3-6):177–194, 1985.

[249] G. Stolovitzky. Gene selection in microarray data: the elephant, the blind men and our algorithms. *Current Opinion in Structural Biology*, 13(3):370–376, June 2003.

[250] S. H. Strogatz. Exploring complex networks. *Nature*, 410(6825):268–276, 2001.

[251] J.-T. Sun, H.-J. Zeng, H. Liu, Y. Lu, and Z. Chen. CubeSVD: a novel approach to personalized web search. In *WWW '05: Proceedings of the 14th international conference on World Wide Web*, pages 382–390, New York, NY, USA, 2005. ACM Press.

[252] R. J. W. Tijssen and A. F. J. van Raan. Mapping co-word structures - a comparison of multidimensional-scaling and Leximappe. *Scientometrics*, 15(3-4):283–295, 1989.

[253] R. Todorov and M. Winterhager. Mapping Australian geophysics - a co-heading analysis. *Scientometrics*, 19(1-2):35–56, 1990.

[254] K. Torkkola. Discriminative features for text document classification. *Pattern Analysis & Applications*, 6(4):301–308, 2003.

[255] M. Toyoda and M. Kitsuregawa. Creating a web community chart for navigating related communities. In *HYPERTEXT '01: Proceedings of the twelfth ACM conference on Hypertext and Hypermedia*, pages 103–112, New York, NY, USA, 2001. ACM Press.

[256] L. R. Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31:279–311, 1966.

[257] W. A. Turner and F. Rojouan. Evaluating input output relationships in a regional research network using co-word analysis. *Scientometrics*, 22(1):139–154, 1991.

[258] S. van Dongen. A cluster algorithm for graphs. Technical report INS-R0010, National Research Institute for Mathematics and Computer Science, Amsterdam, the Netherlands, 2000.

[259] A. F. J. van Raan. Reference-based publication networks with episodic memories. *Scientometrics*, 63(3):549–566, 2005.

[260] A. F. J. van Raan. Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups. *Scientometrics*, 67(3):491–502, 2006.

[261] A. F. J. van Raan and R. J. W. Tijssen. The neural net of neural network research - an exercise in bibliometric mapping. *Scientometrics*, 26(1):169–192, 1993.

[262] J. Vertommen, F. Janssens, J. Duflou, and B. De Moor. Multiple-vector user profiles for knowledge management systems. Technical report 06-22, ESAT-SISTA, K.U.Leuven, Leuven, Belgium, 2006.

[263] J. Vertommen, F. Janssens, B. De Moor, and J. Duflou. Advanced personalization and document retrieval techniques in support of efficient knowledge management. In *Proceedings of the 2nd International Seminar on Digital Enterprise Technology (DET2004)*, Seattle, Washington, USA, Sep. 2004.

[264] C. S. Wagner and L. Leydesdorff. Network structure, self-organization, and the growth of international collaboration in science. *Research Policy*, 34(10):1608–1618, 2005.

[265] W. A. Wallis. Compounding probabilities from independent significance tests. *Econometrica*, 10:229–248, 1942.

[266] Y. Wang and M. Kitsuregawa. Evaluating contents-link coupled web page clustering for web search results. In *CIKM '02: Proceedings of the eleventh international Conference on Information and Knowledge Management*, pages 499–506, New York, NY, USA, 2002. ACM Press.

[267] J. H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58:236–244, 1963.

[268] D. J. Watts. *Small worlds: the dynamics of networks between order and randomness.* Princeton University Press, 1999.

[269] H. D. White. Pathfinder networks and author cocitation analysis: A remapping of paradigmatic information scientists. *Journal of the American Society for Information Science and Technology*, 54(5):423–434, 2003.

[270] P. Wouters and L. Leydesdorff. Has Price's dream come true - is scientometrics a hard science? *Scientometrics*, 31(2):193–222, 1994.

[271] R. E. Wyllys. Measuring scientific prose with rank-frequency (Zipf) curves - new use for an old phenomenon. *Proceedings of the American Society for Information Science*, 12:30–31, 1975.

[272] R. Xu and D. Wunsch. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3):645–678, 2005.

[273] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In *ICML '97: Proceedings of the Fourteenth International Conference on Machine Learning*, pages 412–420, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc.

[274] K. Y. Yeung and W. L. Ruzzo. Principal component analysis for clustering gene expression data. *Bioinformatics*, 17(9):763–774, 2001.

[275] Hongyuan Zha. Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering. In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 113–120, New York, NY, USA, 2002. ACM Press.

[276] B. Zhang, Y. Chen, W. Fan, E. A. Fox, M. A. Goncalves, M. Cristo, and P. Calado. Intelligent fusion of structural and citation-based evidence for text classification. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 667–668, New York, NY, USA, 2005. ACM Press.

[277] G. K. Zipf. *Human behavior and the principle of least-effort: An introduction to human ecology.* Addison-Wesley, 1949.

[278] M. Zitt. A simple method for dynamic scientometrics using lexical analysis. *Scientometrics*, 22(1):229–252, 1991.

[279] M. Zitt and E. Bassecoulard. Development of a method for detection and trend analysis of research fronts built by lexical or cocitation analysis. *Scientometrics*, 30(1):333–351, 1994.

[280] M. Zitt and E. Bassecoulard. Delineating complex scientific fields by an hybrid lexical-citation method: An application to nanosciences. *Information Processing & Management*, 42(6):1513–1531, 2006.

[281] H. Zuckerman. Nobel laureates in science: Patterns of productivity, collaboration, and authorship. *American Sociological Review*, 32:391–403, 1967.

# Curriculum vitae

Frizo Janssens was born in Antwerp, Belgium, on September 19th, 1978. In the year 2000 he obtained a Master of Science in Industrial Engineering, Electronics, Information and Communication Technology (cum laude) from the De Nayer Instituut (Sint-Katelijne-Waver, Belgium). Afterwards he continued his studies at the Katholieke Universiteit Leuven where he received a Master in Artificial Intelligence, Engineering and Computer Science (magna cum laude). The thesis was devoted to the *'Design of an Intelligent Interface: Interfacing a Bibliographic Database'*. In October 2002 he enrolled in a predoctoral program at the Department of Electrical Engineering (ESAT), in the lab of Signals, Identification, System Theory and Automation (SCD/SISTA). Since then he also worked as a researcher in the bioinformatics group, under supervision of prof. dr. ir. Bart De Moor. Since June 2005 he was a visiting research fellow in bibliometrics and data mining at the Steunpunt O&O Indicatoren (Leuven, Belgium; formerly Steunpunt O&O Statistieken).

Contact information:
frizo.janssens@esat.kuleuven.be
frizo.j@gmail.com
+32 486 270537

# Appendix A

# Textual journal profiles

Table A.1: The 50 most important stems or stemmed phrases according to mean TF-IDF score, for each LIS journal (see Section 2.5.3 on page 63) and for the complete data set (938 full-texts articles or notes). Multiple **X**'s in a column mean that the corresponding terms are sorted by decreasing weight for that journal. When a term is also present in the list of a journal more to the left in the table, it is marked with a '*' on the same row, meaning that it is taken out of the ordered list for that journal. Note the last important term for *IPM*, 'speciyc', which might be an illustration of errors that can occur when using OCR or text extraction techniques.

| Term | IPM | JASIST | JDOC | JIS | SciMetr | All journals |
|---|---|---|---|---|---|---|
| queri | **X** | * | * | * | | * |
| imag | **X** | | | * | | * |
| node | **X** | * | | | | * |
| cluster | **X** | | | | * | * |
| vector | **X** | * | | | | * |
| algorithm | **X** | | | | | * |
| fuzzi | **X** | | | | | |
| weight | **X** | | | | | * |
| similar measur | **X** | | | | | |
| paus | **X** | | | | | |
| session | **X** | * | | | | |
| dierent | **X** | | | | | |
| segment | **X** | | | | | |
| web | **X** | * | | | | * |
| sentenc | **X** | | | | | |
| bi gram | **X** | | | | | |
| web page | **X** | | | | | * |
| precis | **X** | | | | | |
| represent | **X** | | | | | |
| speciyc | **X** | | | | | |
| task | | **X** | * | * | | * |
| particip | | **X** | | | | * |
| student | | **X** | | | | * |
| children | | **X** | | | | |
| cognit | | **X** | | | | * |
| seek | | **X** | | * | | * |
| music | | **X** | | | | * |
| behavio(u)r | | **X** | | | | * |
| catalog | | **X** | | | | |
| co citat | | **X** | | | * | * |
| interact | | **X** | | | | * |
| scienc technolog | | **X** | | | | |
| score | | **X** | | | | * |
| digit | | **X** | * | * | | * |
| search engin | | **X** | | | | * |

Continued on next page...

| Term | IPM | JASIST | JDOC | JIS | SciMetr | All journals |
|---|---|---|---|---|---|---|
| book | | | X | | | |
| organis | | | X | | | |
| thesauru | | | X | | | |
| frbr | | | X | | | |
| kiosk | | | X | | | |
| servic | | | X | * | | * |
| loan | | | X | | | |
| jcsm | | | X | | | |
| epistemolog | | | X | | | |
| film | | | X | | | |
| entiti | | | X | | | |
| serendip | | | X | | | |
| women | | | X | | | |
| fiction | | | X | | | |
| borrow | | | X | | | |
| health | | | X | * | | * |
| alzheim diseas | | | | X | | |
| meta data | | | | X | | |
| asset | | | | X | | |
| law | | | | X | | |
| knowledg manag | | | | X | | |
| respond | | | | X | | |
| internet | | | | X | | * |
| topic map | | | | X | | |
| creation | | | | X | | |
| organiz | | | | X | | |
| web site | | | | X | | * |
| regul | | | | X | | |
| sim | | | | X | | |
| legisl | | | | X | | |
| preserv | | | | X | | |
| citat | | | | | X | * |
| patent | | | | | X | * |
| cite | | | | | X | * |
| countri | | | | | X | * |
| collabor | | | | | X | * |
| impact factor | | | | | X | * |
| scienc citat index | | | | | X | |
| scientist | | | | | X | * |
| korean | | | | | X | |
| self citat | | | | | X | |
| chines | | | | | X | * |
| brazilian | | | | | X | |
| physic | | | | | X | * |
| chemistri | | | | | X | |
| citat rate | | | | | X | |
| co author | | | | | X | |
| isi | | | | | X | |
| co authorship | | | | | X | |
| network | | | | | | X |
| rank | | | | | | X |
| domain | | | | | | X |
| languag | | | | | | X |
| social | | | | | | X |
| electron | | | | | | X |
| china | | | | | | X |
| classif | | | | | | X |
| disciplin | | | | | | X |
| resourc | | | | | | X |
| item | | | | | | X |
| industri | | | | | | X |
| interfac | | | | | | X |
| map | | | | | | X |
| titl | | | | | | X |

# Appendix B

# Bibliographic sources of papers subjected to analysis

Bibliographic sources of papers referred to in the text as subject of analysis (in alphabetical order of the first authors):

Alavi *et al.* (2002). JASIST, 53(12):1029.
Aljlayl *et al.* (2002). JASIST, 53(13), 1139.
Archambault (2002). Scientometrics, 54(1), 15.
Beaulieu (2003). Journal of Information Science, 29(4), 239.
Blair (2002). Information Processing & Management, 38(2), 293.
Brajnik *et al.* (2002). JASIST, 53(5), 343.
Breitzman& Mogee (2002). Journal of Information Science, 28(3), 187.
Can *et al.* (2004). Information Processing & Management, 40(3), 495.
Christoffersen (2004). Scientometrics, 61(3), 385.
Ding *et al.* (2002a). Journal of Information Science, 28 (2):123.
Ding (2002b). Journal of Information Science, 28(5):375.
Dominich (2003). Information Processing & Management, 39(2), 167.
Dominich *et al.* (2004). JASIST, 55(7), 613.
Egghe& Rousseau (2002). Scientometrics, 55(3), 349.
Faba-Perez *et al.* (2003). Journal of Documentation, 59(5):558.
Ford *et al.* (2002). Journal of Documentation, 58(1), 30.
Glänzel& Meyer (2003). Scientometrics, 58(2), 415.
Glänzel& Moed (2002). Scientometrics, 53(2), 171.
He *et al.* (2002). Information Processing & Management, 38(5), 727.
He& Hui (2002). Information Processing & Management, 38(4), 491.
Larsen (2002). Scientometrics, 54(2), 155.
Lee *et al.* (2004). Information Processing & Management, 40(1), 145.
Lehtokangas *et al.* (2004). Information Processing & Management, 40(6), 973.
Lewison (2002a). Scientometrics, 54(2), 179.
Lewison (2002b). Scientometrics, 53(2), 229.
Leydesdorff (2002). Scientometrics, 53(1):131.
Lin *et al.* (2003). Information Processing & Management, 39(5), 689.
Lippincott (2002). Journal of Information Science, 28(2), 137.
Muresan *et al.* (2004). JASIST, 55(10), 892.
Nie (2003). JASIST, 54(4) 335.
Niemi& Hirvonen (2003). JASIST, 54(10), 939.
Ozmutlu& Spink (2002). Information Processing & Management, 38(4), 473.

Pennanen& Vakkari (2003). JASIST, 54(8), 759.

Persson *et al.* (2004). Scientometrics, 60(3), 421.

Pharo& Jarvelin (2004). Information Processing & Management, 40(4), 633.

Pirkola *et al.* (2003). Information Processing & Management, 39(3), 391.

Pudovkin& Garfield (2002), JASIST, 53(13), 1113.

Schlieder& Meuss (2002). JASIST, 53(06), 489.

Schneider& Borlund (2004). Journal of Documentation, 60(5), 524.

Spink *et al.* (2004). Information Processing & Management, 40(1), 113.

Tombros *et al.* (2002). Information Processing & Management, 38(4), 559.

Vakkari *et al.* (2003). Information Processing & Management, 39(3), 445.

van den Besselaar (2003). JASIST, 54(1):87.

Wilkinson *et al.* (2003). Journal of Information Science, 29(1), 49.

Wormell (2003). Journal of Information Science, 29(3), 193.

Zhao& Logan (2002). Scientometrics, 54(3), 449.

# Appendix C

# Representative publications for 9 bioinformatics clusters

Table C.1: For each of 9 clusters: the two publications with largest cosine similarity to the mean cluster profile (*medoid* papers); the two papers most cited from within the cluster; the two best authorities and best hubs detected by the HITS algorithm; and the two papers with highest PageRank according to *Google*'s algorithm.

| | |
|---|---|
| **Cluster 1. RNA structure prediction (n=205)** | |
| Medoids | Major *et al.* Computational methods for RNA structure determination. *Current Opinion in Structural Biology* 11 (3):282-286, 2001.<br>Tinoco *et al.* How RNA folds. *Journal of Molecular Biology* 293 (2):271-281, 1999. |
| Most cited | Mathews *et al.* Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *Journal of Molecular Biology* 288 (5):911-940, 1999.<br>Zuker. On Finding All Suboptimal Foldings of An RNA Molecule. *Science* 244 (4900):48-52, 1989. |
| Authorities | Mathews *et al.* Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *Journal of Molecular Biology* 288 (5):911-940, 1999.<br>Rivas *et al.* A dynamic programming algorithm for RNA structure prediction including pseudoknots. *Journal of Molecular Biology* 285 (5):2053-2068, 1999. |
| Hubs | Zuker. Calculating nucleic acid secondary structure. *Current Opinion in Structural Biology* 10 (3):303-310, 2000.<br>Higgs. RNA secondary structure: physical and computational aspects. *Quarterly Reviews of Biophysics* 33 (3):199-253, 2000. |
| PageRank | Zuker *et al.* Optimal Computer Folding of Large RNA Sequences Using Thermodynamics and Auxiliary Information. *Nucleic Acids Research* 9 (1):133-148, 1981.<br>Freier *et al.* Improved Free-Energy Parameters for Predictions of RNA Duplex Stability. *Proceedings of the National Academy of Sciences of the United States of America* 83 (24):9373-9377, 1986. |
| **Cluster 2. Protein structure prediction (n=1167)** | |
| Medoids | Di Francesco *et al.* FORESST: fold recognition from secondary structure predictions of proteins. *Bioinformatics* 15 (2):131-140, 1999.<br>Garnier *et al.* The Protein-Structure Code - What Is Its Present Status. *Computer Applications in the Biosciences* 7 (2):133-142, 1991. |
| Most cited | Murzin *et al.* SCOP - A Structural Classification of Proteins Database for the Investigation of Sequences and Structures. *Journal of Molecular Biology* 247 (4):536-540, 1995.<br>Orengo *et al.* CATH - a hierarchic classification of protein domain structures. *Structure* 5 (8):1093-1108, 1997. |
| Authorities | Murzin *et al.* SCOP - A Structural Classification of Proteins Database for the Investigation of Sequences and Structures. *Journal of Molecular Biology* 247 (4):536-540, 1995.<br>Jones *et al.* A New Approach to Protein Fold Recognition. *Nature* 358 (6381):86-89, 1992. |
| Hubs | Eisenhaber *et al.* Protein-Structure Prediction - Recognition of Primary, Secondary, and Tertiary Structural Features from Amino-Acid-Sequence. *Critical Reviews in Biochemistry and Molecular Biology* 30 (1):1-94, 1995.<br>Bohm. New approaches in molecular structure prediction. *Biophysical Chemistry* 59 (1-2):1-32, 1996. |
| PageRank | Chothia *et al.* The Relation Between the Divergence of Sequence and Structure in Proteins. *Embo Journal* 5 (4):823-826, 1986.<br>Kyte *et al.* A Simple Method for Displaying the Hydropathic Character of a Protein. *Journal of Molecular Biology* 157 (1):105-132, 1982. |
| **Cluster 3. Systems biology & molecular networks (n=694)** | |
| Medoids | Xiong *et al.* Network-based regulatory pathways analysis. *Bioinformatics* 20 (13):2056-2066, 2004. |

Continued on next page...

| | |
|---|---|
| | Lukashin *et al.* Topology of gene expression networks as revealed by data mining and modeling. *Bioinformatics* 19 (15):1909-1916, 2003. |
| Most cited | Jeong *et al.* The large-scale organization of metabolic networks. *Nature* 407 (6804):651-654, 2000.<br>Ito *et al.* A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences of the United States of America* 98 (8):4569-4574, 2001. |
| Authorities | Jeong *et al.* The large-scale organization of metabolic networks. *Nature* 407 (6804):651-654, 2000.<br>Ito *et al.* A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences of the United States of America* 98 (8):4569-4574, 2001. |
| Hubs | Xia *et al.* Analyzing cellular biochemistry in terms of molecular networks. *Annual Review of Biochemistry* 73:1051-1087, 2004.<br>You. Toward computational systems biology. *Cell Biochemistry and Biophysics* 40 (2):167-184, 2004. |
| PageRank | Karp *et al.* EcoCyc: Encyclopedia of Escherichia coli genes and metabolism. *Nucleic Acids Research* 26 (1):50-53, 1998.<br>Bono *et al.* Reconstruction of amino acid biosynthesis pathways from the complete genome sequence. *Genome Research* 8 (3):203-210, 1998. |
| **Cluster 4. Phylogeny & evolution (n=749)** | |
| Medoids | Negrisolo *et al.* Morphological convergence characterizes the evolution of Xanthophyceae (Heterokontophyta): evidence from nuclear SSU rDNA and plastidial rbcL genes. *Molecular Phylogenetics and Evolution* 33 (1):156-170, 2004.<br>Stefanovic *et al.* Testing the phylogenetic position of a parasitic plant (Cuscuta, Convolvulaceae, Asteridae): Bayesian inference and the parametric bootstrap on data drawn from three genomes. *Systematic Biology* 53 (3):384-399, 2004. |
| Most cited | Posada *et al.* MODELTEST: testing the model of DNA substitution. *Bioinformatics* 14 (9):817-818, 1998.<br>Huelsenbeck *et al.* MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17 (8):754-755, 2001. |
| Authorities | Posada *et al.* MODELTEST: testing the model of DNA substitution. *Bioinformatics* 14 (9):817-818, 1998.<br>Huelsenbeck *et al.* MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17 (8):754-755, 2001. |
| Hubs | Delsuc *et al.* Molecular systematics of armadillos (Xenarthra, Dasypodidae): contribution of maximum likelihood and Bayesian analyses of mitochondrial and nuclear genes. *Molecular Phylogenetics and Evolution* 28 (2):261-275, 2003.<br>Douady *et al.* Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. *Molecular Biology and Evolution* 20 (2):248-254, 2003. |
| PageRank | Saitou *et al.* The Neighbor-Joining Method - A New Method for Reconstructing Phylogenetic Trees. *Molecular Biology and Evolution* 4 (4):406-425, 1987.<br>Jin *et al.* Limitations of the Evolutionary Parsimony Method of Phylogenetic Analysis. *Molecular Biology and Evolution* 7 (1):82-102, 1990. |
| **Cluster 5. Genome sequencing & assembly (n=640)** | |
| Medoids | Barber *et al.* SequenceEditingAligner - A Multiple Sequence Editor and Aligner. *Genetic Analysis-Biomolecular Engineering* 7 (2):39-45, 1990.<br>Staden. Searching for Patterns in Protein and Nucleic-Acid Sequences. *Methods in Enzymology* 183:193-211, 1990. |
| Most cited | Devereux *et al.* A Comprehensive Set of Sequence-Analysis Programs for the VAX. *Nucleic Acids Research* 12 (1):387-395, 1984.<br>Pearson *et al.* Improved Tools for Biological Sequence Comparison. *Proceedings of the National Academy of Sciences of the United States of America* 85 (8):2444-2448, 1988. |
| Authorities | SantaLucia. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proceedings of the National Academy of Sciences of the United States of America* 95 (4):1460-1465, 1998.<br>Allawi *et al.* Thermodynamics and NMR of internal GT mismatches in DNA. *Biochemistry* 36 (34):10581-10594, 1997. |
| Hubs | Kaderali *et al.* Selecting signature oligonucleotides to identify organisms using DNA arrays. *Bioinformatics* 18 (10):1340-1349, 2002.<br>Vallone *et al.* AutoDimer: a screening tool for primer-dimer and hairpin structures. *Biotechniques* 37 (2):226-231, 2004. |
| PageRank | Wilbur *et al.* Rapid Similarity Searches of Nucleic-Acid and Protein Data Banks. *Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences* 80 (3):726-730, 1983.<br>Lipman *et al.* Rapid and Sensitive Protein Similarity Searches. *Science* 227 (4693):1435-1441, 1985. |
| **Cluster 6. Gene/promoter/motif prediction (n=995)** | |
| Medoids | Park *et al.* Comparing expression profiles of genes with similar promoter regions. *Bioinformatics* 18 (12):1576-1584, 2002.<br>Kielbasa *et al.* Combining frequency and positional information to predict transcription factor binding sites. *Bioinformatics* 17 (11):1019-1026, 2001. |
| Most cited | Burge *et al.* Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology* 268 (1):78-94, 1997.<br>van Helden *et al.* Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *Journal of Molecular Biology* 281 (5):827-842, 1998. |
| Authorities | Burge *et al.* Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology* 268 (1):78-94, 1997.<br>van Helden *et al.* Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *Journal of Molecular Biology* 281 (5):827-842, 1998. |
| Hubs | Mathe *et al.* Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Research* 30 (19):4103-4117, 2002.<br>Bulyk. Computational prediction of transcription-factor binding site locations. *Genome Biology* 5 (1), 2004. |
| PageRank | Uberbacher *et al.* Locating Protein-Coding Regions in Human DNA-Sequences by a Multiple Sensor Neural Network Approach. *Proceedings of the National Academy of Sciences of the United States of America* 88 (24):11261-11265, 1991.<br>Bucher *et al.* Compilation and Analysis of Eukaryotic Pol-II Promoter Sequences. *Nucleic Acids Research* 14 (24):10009-10026, 1986. |
| **Cluster 7. Molecular DBs & annotation platforms (n=1091)** | |
| Medoids | Andrade *et al.* Automated genome sequence analysis and annotation. *Bioinformatics* 15 (5):391-412, 1999.<br>Wu *et al.* The Protein Information Resource: an integrated public resource of functional annotation of proteins. *Nucleic Acids Research* 30 (1):35-37, 2002. |

Continued on next page...

| | |
|---|---|
| Most cited | Bairoch *et al.* The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research* 28 (1):45-48, 2000.<br>Berman *et al.* The Protein Data Bank. *Nucleic Acids Research* 28 (1):235-242, 2000. |
| Authorities | Bairoch *et al.* The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research* 28 (1):45-48, 2000.<br>Hofmann *et al.* The PROSITE database, its status in 1999. *Nucleic Acids Research* 27 (1):215-219, 1999. |
| Hubs | Kriventseva *et al.* Clustering and analysis of protein families. *Current Opinion in Structural Biology* 11 (3):334-339, 2001.<br>Murvai *et al.* The SBASE protein domain library, release 7.0: a collection of annotated protein sequence segments. *Nucleic Acids Research* 28 (1):260-262, 2000. |
| PageRank | Henikoff *et al.* Automated Assembly of Protein Blocks for Database Searching. *Nucleic Acids Research* 19 (23):6565-6572, 1991.<br>Wallace *et al.* PATMAT - A Searching and Extraction Program for Sequence, Pattern and Block Queries and Databases. *Computer Applications in the Biosciences* 8 (3):249-254, 1992. |

| **Cluster 8. Multiple sequence alignment (n=713)** | |
|---|---|
| Medoids | Jaroszewski *et al.* Improving the quality of twilight-zone alignments. *Protein Science* 9 (8):1487-1496, 2000.<br>Morgenstern *et al.* Multiple DNA and protein sequence alignment based on segment-to-segment comparison. *Proceedings of the National Academy of Sciences of the United States of America* 93 (22):12098-12103, 1996. |
| Most cited | Altschul *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 25 (17):3389-3402, 1997.<br>Thompson *et al.* Clustal-W - Improving the Sensitivity of Progressive Multiple Sequence Alignment Through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice. *Nucleic Acids Research* 22 (22):4673-4680, 1994. |
| Authorities | Altschul *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 25 (17):3389-3402, 1997.<br>Thompson *et al.* Clustal-W - Improving the Sensitivity of Progressive Multiple Sequence Alignment Through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice. *Nucleic Acids Research* 22 (22):4673-4680, 1994. |
| Hubs | Gotoh. Multiple sequence alignment: Algorithms and applications. *Advances in Biophysics* 36:159-206, 1999.<br>Lecompte *et al.* Multiple alignment of complete sequences (MACS) in the post-genomic era. *Gene* 270 (1-2):17-30, 2001. |
| PageRank | Fitch *et al.* Optimal Sequence Alignments. *Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences* 80 (5):1382-1386, 1983.<br>Waterman. General-Methods of Sequence Comparison. *Bulletin of Mathematical Biology* 46 (4):473-500, 1984. |

| **Cluster 9. Microarray analysis (n=1147)** | |
|---|---|
| Medoids | Tsai *et al.* An evolutionary approach for gene expression patterns. *IEEE Transactions on Information Technology in Biomedicine* 8 (2):69-78, 2004.<br>Wang *et al.* A generalized likelihood ratio test to identify differentially expressed genes from microarray data. *Bioinformatics* 20 (1):100-104, 2004. |
| Most cited | Eisen *et al.* Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America* 95 (25):14863-14868, 1998.<br>Golub *et al.* Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* 286 (5439):531-537, 1999. |
| Authorities | Eisen *et al.* Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America* 95 (25):14863-14868, 1998.<br>Golub *et al.* Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* 286 (5439):531-537, 1999. |
| Hubs | Hackl *et al.* Analysis of DNA microarray data. *Current Topics in Medicinal Chemistry* 4 (13):1357-1370, 2004.<br>Sebastiani *et al.* Statistical challenges in functional genomics. *Statistical Science* 18 (1):33-60, 2003. |
| PageRank | Schena *et al.* Quantitative Monitoring of Gene-Expression Patterns with A Complementary-DNA Microarray. *Science* 270 (5235):467-470, 1995.<br>Shalon *et al.* A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Research* 6 (7):639-645, 1996. |