A hierarchical and optimal clustering of WoS journal database by hybrid information

Xinhai Liu[1], Wolfgang Glänzel[2,3], Bart De Moor[4]

[1]*Xinhai.liu@esat.kuleuven.be*
K.U. Leuven, Dept. of Electrical Engineering ESAT-SCD, Leuven (Belgium)

[2]*Wolfgang.Glanzel@econ.kuleuven.be*
K.U. Leuven, Steunpunt O&O Indicatoren (ECOOM) and Dept. MSI, Leuven (Belgium)

[3]*glanzw@iif.hu*
Hungarian Academy of Sciences, ISPR, Budapest (Hungary)

[4]*Bart.DeMoor@esat.kuleuven.be*
K.U. Leuven, Dept. of Electrical Engineering ESAT-SCD, Leuven (Belgium)

Abstract

Previous studies have shown that hybrid clustering methods based on textual and citation information outperforms clustering methods that use only one of these components. However, former methods focus on the vector space model. In this paper we apply a hybrid clustering method which is based on the graph model to map the Web of Science database in the mirror of the journals covered by the database.

Compared with former hybrid clustering strategies, our method is very fast and even achieves better clustering accuracy. In addition, it detects the number of clusters automatically and provides a top-down hierarchical analysis, which fits in the practical application.

We quantitatively and qualitatively asses the added value of such an integrated analysis and we investigate whether the clustering outcome provides an appropriate representation of the field structure by comparing with a text-only clustering and with another hybrid method based on linear combination of distance matrices. Our dataset consists of about 8000 journals published in the period 2002–2006. The cognitive analysis, including the ranked journals, term annotation and the visualization of cluster structure demonstrates the efficiency of our strategy.

Introduction

The objective of this research is an accurate unsupervised clustering of science or technology fields, towards the detection of new emerging fields. The idea of combining citation information with textual content is not new for it has already been pursued to obtain improved performance in information retrieval (e.g., Calado et al., 2003), bibliometric mapping of science (Mullins et al., 1988; Snizek et al., 1991; Braam et al., 1991a,b; Glenisson et al., 2005; Janssens et al., 2006b), clustering (e.g., Modha & Spangler, 2000; Wang & Kitsuregawa, 2002), and classification issues (e.g., Joachims et al., 2001; Calado et al., 2006).

Sometimes textual information can indeed indicate similarities that are not visible through citation links, and vice versa. On the other hand, based on text alone, true document similarity might be obscured by differences in vocabulary use, or spurious similarities might be introduced as a result of textual pre-processing like stemming, 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, business*, etc. Consequently, the combination of textual data and citation data is thought as a promising method to deal with scientific publication. Some hybrid clustering has been carried out, such as Janssens et al. (2006a,b) putting forward a hybrid clustering strategy based on a convex combination of distance matrices method (WLCDM) and Fisher's inverse χ^2 method. Most of the applied hybrid methods are based on vector space model.

Due to the growth of information and the availability of huge databases during the last decades, handling the amount of data has become a real challenge to information science. The vector-space based hybrid clustering is usually limited to tackle scalable data. Furthermore, the number of clusters is often considered to be known or is estimated based on variety of measures. Therefore a non-parameter or no-hypothesis clustering is needed. In addition, clustering methods often returns a one level cluster structure which does not always reflect the nature of data structure correctly because the real data is pretty complicated. Therefore, multi-level of cluster structure or hierarchical cluster structure should be preferred.
However, in the last few years, there has been a concerted interdisciplinary effort to develop mathematical tools and computer algorithms to detect community structure in large networks. Such a problem is often computationally intractable and therefore requires approximation methods in order to find reasonably good partitions in a reasonably fast way. The rapidity of the algorithm has become a crucial factor due to the increasing size of the networks to be analyzed. A large variety of methods have been developed in order to address this problem (Fortunato, 2010). In particular, the recent method called "Louvain method", which is based on approximate modularity optimization, outperforms the alternative methods in terms of computation time, while having an excellent accuracy (Blondel et al., 2008). The Louvain method has been employed in the analysis of scientific knowledge. Lambiotte and Panzarasa (2009) discussed the community detection of scientific collaboration network by Louvain method. Rafols and Leydesdorff (2009) investigated the clustering of Louvain method of the 2006 edition of the Journal Citation Report (JCR) and compared the results with those of other three classification schemes. In former research, we have used the method to cluster the subjects structure of the Web of Science (WoS) based on ISI classification based on citation link data (Zhang et al., 2010). However, graph partition methods usually focus on link structure and ignore attribute similarities. Therefore, we put forward a hybrid strategy based on network (graph) model to deal with the clustering problems mentioned before. Our strategy is able to facilitate the clustering task by several ways: combining citation links and textual information, is the self optimizing and provides a hierarchical analysis.
In a related approach, He et al. (2001) implemented the combination of hyberlink structure and textural similarity to cluster the Webpages. Here we use the cross-citation link and the k-nearest neighbour relationship and modularity optimization based on the Louvain method.

Data sources and methodology

The raw dataset contains more than 6,000,000 publications (articles, letters, notes, and reviews) indexed by the WoS database of Thomson Reuters for the period 2002–2006. In pre-processing, the ambiguities of journal names, author names and bibliographic data are resolved. We only keep the journals have at least 50 papers and have more than 30 citations. After pre-processing, we obtain 8,305 journals as the journal data set adopted in this paper.

Text Mining Analysis

In a first step, we have retrieve lexical and citation information from the selected journals. The titles, abstracts and keywords of journal publications are indexed by a Jakarta Lucene[1] platform (Hatcher & Gospodnetić, 2004) based on a text mining program without controlled vocabulary. The index result contains 9,473,061 terms and we cut the Zipf curve of terms at the head and the tail to remove the rare terms, stopwords and common words. After Zipf cut, 669,860 meaningful terms are kept as the attribute representations in vector space model. All textual content was encoded in the vector space model using the TF-IDF weighting scheme

[1] http://lucene.apache.org/, visited in November 2006.

(Baeza-Yates and Ribeiro-Neto, 1999). The paper-by-term vectors are then aggregated to journal-by-term vectors as the representations of the lexical data. Text-based similarities were calculated as the cosine of the angle between the vector representations of two papers (Salton and McGill, 1986).

Citation analysis

In a second step, we analysed the cross-citations links among the selected journals. Citations among individual papers were aggregated to the journal level. We ignored the direction of citations links by symmetrizing the cross-citation matrix. Each row of this cross-citation matrix can be taken as a (citation-) link vector of the corresponding journal. Though the representation of citations actually forms a sparse graph, we can also regard it as journal-by-citation vectors, where the similarities of journals are measured by constructing the empirical kernel, which is equivalent to the cosine value of journal-by-citation vectors.

Clustering strategy

1. Modularity

Modularity is a benefit function used in the analysis of networks or graphs (Newman, 2004a):

$$Q = \frac{1}{2m}\sum_{ij}[A_{ij} - \frac{k_i k_j}{2m}]\delta(c_i, c_j) \qquad (1)$$

where A_{ij} represents the weight of the edge between vertex i and vertex j; $k_i = \sum_j A_{ij}$ is the sum of the weights of the edges attached to vertex i; c_i is the community to which vertex i belongs; the δ function $\delta(u,v)$ is 1 if $u = v$ and 0 otherwise and $m = 0.5\sum_{ij} A_{ij}$. The exact optimization of modularity is a problem that is computationally hard. A number of algorithms have been recently introduced in order to deal with this problem. The basic idea is to recurrently merging community that optimize the production of modularity between two vertices:

$$\Delta Q = \begin{cases} A_{ij} - \frac{k_i k_j}{2m}, & \text{if } (i, j) \text{ are connected} \\ 0, & \text{otherwise} \end{cases}$$

2. Louvain method

Based on modularity optimization, the Louvain method incorporates a multi-level organization and consists of two phases that are repeated iteratively. First, the algorithm looks for "small" communities by optimizing modularity in a greedy, local way. Second, the algorithm aggregates nodes of the same community and builds a new network whose nodes are the communities. These phases are repeated iteratively until a maximum of modularity is attained and optimal partition of the network into communities is found. The choice of this method for community detection is motivated by its excellent accuracy and its rapidity which allows us to study networks of unprecedented size (e.g. the analysis of a typical network of 2 million nodes only takes 2 minutes). The Louvain method has also been to shown to be very accurate by focusing on ad-hoc networks with known community structure. Moreover, due to its hierarchical structure, which is reminiscent of renormalization methods, it allows to look at communities at different resolutions.

The output of the program therefore gives several partitions. The partition found after the first step typically consists of many communities of small sizes. At subsequent steps, larger and larger communities are found due to the aggregation mechanism. This process naturally leads to hierarchical decomposition of the network. The Louvain method can be regarded as a hierarchical partition method from graph view.

3. Hybrid clustering

Thank to the merits of Louvain method, taking each journal as a vertex, we can directly carry out the clustering analysis of the database by cross-citation data. Since the text data of WoS is available, we intend to utilize the textual information which is supposed to complement the citation data. By combining these two information, it is expected to obtain a robust cluster structure. However, the question remains of how to deal with these two heterogeneous data in graph space arises?

Inspired by the research work in webpage clustering (He et al, 2001), we are able to integrate the cross-citation with text in the following way. The link structure is determined by cross-citation, that is, if there is cross-citation relationship between two journals, there will be a link, while the edge strength determined by the textual similarity. Consequently, a fused graph is generated and the Louvain method can be implemented on it to obtain a final partition result. In the empirical test on the fused graph, we found that some edges with weak strength (weak textual similarity) have negative impact on the final partition. This kind of edge can be understood as although two journals are cross-cited (relevant) but they share less textual terms (not similar). In common sense, if two journals are cross-cited each other but share less textual terms, they should not be clustered into the same category. Therefore, to neglect their negative impact in the partition, we even go further to use the k-nearest neighbours constraint to filter out these edges. By this means, we strengthen the effect of those journal vertices which are cross-cited and share more textual similarity.

Compared with hybrid clustering in vector space, our hybrid strategy is completely distinct, in particular.

- Out strategy does not require any previous setting, such as, the number of clusters,
- Integration schemes are different; we use the cross-citation link structure to couple the textual similarity, plus the KNN constraint to neglect the un-useful edges,
- Partition scheme are different; in the vector space model, we can use kernel k-means, ward-linkage method and spectral clustering while in graph space, the modularity optimization based on the Louvain method is employed.
- Final cluster structures are different, in the vector space, usually one level clustering result is provided while the optimal hierarchical structure is offered which would more fit in with the practical tasks.

In addition, we grasp the core information of the textual and citation data while neglecting large data without sufficient information, so that our strategy is applicable to the large-scale application.

Results

Fixing the number of clusters as 22 (the number of ESI category)

In order to compare our hybrid strategy in the graph model with single-data clustering solution in the graph model as well as with the hybrid strategy in the vector space, we fix the cluster number the same as the number of standard ESI fields (used by Thomson Reuters' Essential Science Indicators [2] (ESI)). Otherwise, it is impossible to compare different clustering results with different cluster numbers. Three vector space based clustering solutions are compared: clustering on TF-IDF data, clustering on cross-citation data and hybrid strategy of WLCDM (Janssens et al., 2008). The final partition is implemented by Ward's linkage method (Jain and Dubes, 1988). The three counterpart solutions in the graph model are

[2] http://www.esi-topics.com/fields/index.html.

compared: clustering on TF-IDF data (we apply the top 100 nearest neighbour constraint to sparse the full textual similarity matrix to obtain a graph), clustering on cross-citation data and our hybrid strategy. The final partition is implemented by the Louvain method.
Two external evaluation measures are adopted to gauging the clustering performance against ESI fields as benchmark. One is Adjusted Rand Index (ARI) (cf. Hubert and Arabie, 1985), the other one is the Normalized Mutual Information (NMI) (Strehl et al., 2002). Both are trying to measure the overlap between clustering result and benchmark category. The larger the evaluation values, the better clustering performance. The results are shown in Table 1. Our hybrid strategy proved to be the best clustering method according to both NMI and ARI. The corresponding solutions in the graph model seem to be better than their counterparts in vector space. Due to the final partition algorithms in both the vector space and the graph model, the clustering result of each strategy is unique.

Table 1 the clustering evaluation with fixed cluster (22)

Models	*Methods*	*NMI*	*ARI*
Vector space	TFIDF	0.5080	0.2676
	CRC	0.4532	0.1604
	WLCDM	0.5161	0.2885
Graph space	TFIDF (KNN100)	0.5309	0.29
	CRC (Salton)	0.5640	0.3209:
	Fused graph (Hybrid)	**0.5768**	**0.3407** :

In addition, we compared the computational time of the related algorithms. The experiment was carried out on a CentOS 5.2 Linux system with a 2.4G Hz CPU and 16 G Bytes memory. As shown in Table 2, the clustering solutions in the graph model are distinctly (by almost two orders of magnitude) faster than their counterparts in the vector space.

Table2: the comparison of running time by different clustering schemes

Models	*Methods*	*Time (seconds)*
Vector space model	TFIDF	600
	CRC	600
	WLCDM	700
Graph model	TFIDF(KNN)	8
	CRC(Salton)	9
	Fused graph	12

The excellent performance of the new hybrid strategy is based on two facts: (1) the information fusion which usually leads to a robust and better clustering structure and (2) the partition efficiency of the Louvain method.

The hierarchical clustering structure optimized by the Louvain method

Based on the above results, we decided to implement the optimal clustering, that is, the optimal cluster number and the optimal hierarchical structure are found automatically during the partition process. No input parameters are thus needed for this optimum partition, only the adjacent matrix of this fused graph, which is generated by combining the cross-citation link structure with textual similarity. The partition strategy of the Louvain method is able to find the optimal cluster number by maximizing the modularity. When a local maximum modularity is reached the number of clusters and its related clustering structure are recorded. Because of the aggregating optimal mechanism, the different cluster structure has hierarchical relationship. Thus we can obtain a graph structure with various resolutions.

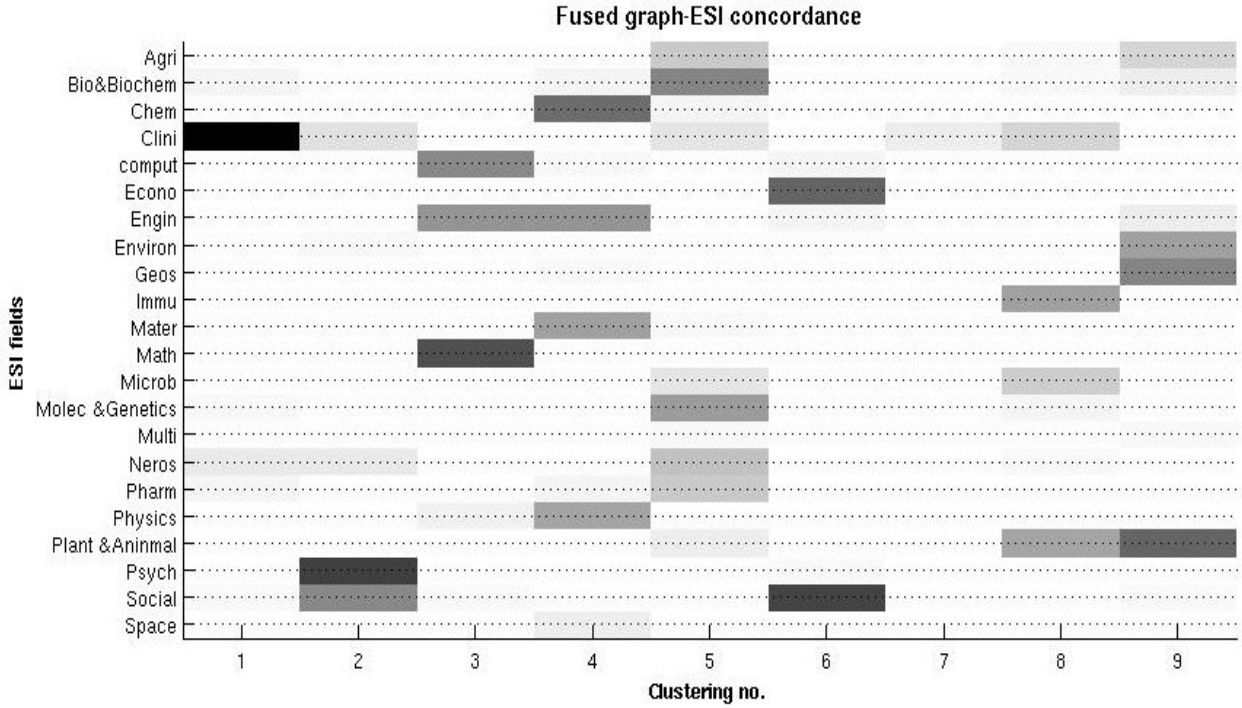

Figure 1 the concordance between cluster obtained the fused graph and ESI category

We have stopped at two levels of partitions obtained by the hierarchical strategy. At the higher level, we got 9 clusters for the fused graph while in the lower level, we obtained 45 clusters. The cognitive analysis of the second level is shown below.

Since the optimal cluster number (9 or 45) differs from the number of ESI fields (22), we can not gauge the clustering performance against the two evaluations of the previous subsection. But we still can take the ESI as a reference standard to determine if our clustering results are meaningful by finding any concordance between them. As shown in Figure 1, concordance between our clustering solution (9 clusters) and the ESI scheme are visualized by gray-scaled cells representing the Jaccard index for each cluster and field pair. The darkest cells represent the best-matching pairs of fields and clusters. It is clear that each of our 9 clusters corresponds to one ESI field or several relevant ESI fields. For instance, cluster #7 corresponds to the Clinical Medicine, #6 corresponds to the fields of Economic & Business and Social Sciences, and #3 corresponds to Computer Science, Engineering and Mathematics.

Table 3: the five most important journals of each cluster according to a modified version of Google PageRank algorithm,

Cluster1	*Cluster 2*	*Cluster 3*
1.Nat Rev cancer	1. Annu Rev Psycho	1. J Roy Stat Soc S.B
2. Caner cell	2. Psycho Meth	2. Fund comp math
3. CA-cancer J Clin	3. Psych B.	3. Biostat
4. Annu Rev Med	4. Psych Rev	4. J Amer Math Soc
5. B. B. Rev cancer	5. Behav Brai Sci	5. Anna Math
Cluster 4	*Cluster 5*	*Cluster 6*
1. Rev Mod Phys	1. Nat Rev molec cell bio	1. Quart J econ
2. Nat material	2. Nat Rev genetics	2. J econ liter
3. Chem Rev	3. Deve cell	3. J finance
4. Annu Rev Astron & Astrop	4. Nat Rev neruos	4. J finance econ
5. Mate sci & eng Rep	5. Annu Rev Bioche	5. J poli econ
Cluster 7	*Cluster 8*	*Cluster 9*
1. Prog retin & eye res	1. Nat Rev immu	1. Annu Rev Ecolog evo & Sys
2. Invest ophth & visua sci	2. Annu Rev Immu	2. Ocean & Marin Bio
3. Surv ophth	3. Nat immu	3. Syst Bio
4. Molec vision	4. Nat Medi	4. A Muse Novi
5. Archi ophth	5. J Exp Med	5. Annu Rev Entom

Table 4: the 30 best TF-IDF terms describing the 9 hybrid citation-textual clusters

Cluster	*Best 30 terms*
1	patient tumor cancer clinic cell arteri diseas therapi surgeri carcinoma renal diabet coronari lesion transplant pain postop surgic gene blood dose bone breast prostat women liver resect hospit rat protein lung tumour tissu stent receptor serum cardiac graft children hypertens ventricular acut malign biopsi syndrom mice pulmonari myocardi infect ltd
2	patient children schizophrenia student health adolesc nurs women disord depress symptom clinic cognit school teacher mental psychiatr ltd social educ anxieti hospit smoke emot suicid psycholog interview child questionnair abus sleep physician ptsd medic score adhd intervent care alcohol men infant particip sexual adult violenc inc drug risk wilei antipsychot
3	algorithm algebra graph fuzzi finit ltd wireless theorem antenna wilei queri polynomi semant inc nonlinear robot asymptot qo equat packet infin bandwidth xml network user scheme multicast manifold fault server nois bit simul web circuit cdma integ let cach paper filter traffic servic machin architectur topolog watermark wavelet queue bound
4	film temperatur ltd alloi crystal atom ion polym quantum catalyst galaxi dope magnet metal oxid hydrogen diffract optic particl thermal wilei bond beam spin spectroscopi rai angstrom electron si spectra acid nm laser surfac adsorpt energi nanoparticl carbon cu poli dielectr nmr layer ligand nanotub solvent anneal molecul electrod fe
5	protein gene cell receptor rat neuron mice kinas bind mutant transcript acid mrna dna ca2 phosphoryl ltd mutat enzym inhibit peptid inhibitor inc apoptosi membran beta genom brain mous insulin muscl rna patient subunit oxid vitro amino chromosom tissu alpha pathwai wilei vivo induc assai tumor metabol coli liss regul
6	polit firm polici market price social busi tax wage ltd economi capit organiz war trade welfar reform court parti democraci labour corpor invest women discours democrat countri employe pavement econom compani financi employ crime servic household moral urban argu labor religi sector incom monetari earn vote job privat investor ethic
7	corneal retin ey patient glaucoma acuiti iop iol macular cataract intraocular lasik ocular surgeri cornea retina len choroid postop myopia vitrectomi astigmat refract phacoemulsif rpe ophthalmolog cnv retinopathi vitreou keratoplasti conjunctiv intravitr neovascular uveiti cell epitheli latanoprost keratomileusi anterior lens kerat visual myopic scleral preoper clinic glaucomat amd optic eyelid
8	infect cell patient il hiv viru vaccin mice protein gene antibodi immun antigen cd4 ifn cow diseas dog clinic cytokin receptor cd8 viral milk pcr lymphocyt calv serum hla hcv macrophag assai tumor hors tnf therapi strain blood pig diet allergen dna asthma cattl vitro broiler rna arthriti inflammatori tuberculosi
9	soil speci plant forest sediment habitat water ltd lake basin ocean river biomass season sea fish crop leaf cultivar seedl rock seismic seed predat temperatur fault climat larva veget isotop ecosystem rainfal earthquak nutrient prei carbon assemblag mantl shoot egg tecton wind magma tree winter ha groundwat taxa flower wheat

To better understand the structure of clustering, we applied a modified Google PageRank algorithm (Janssens et al., 2009) to analyze the journals within each cluster. Using the algorithm, we investigated the five most highly ranked journals in each cluster and presented them in Table3. Moreover, for the journals presented in Table 3, we re-investigated the titles, abstracts and keywords that have been indexed in the text mining process, the indexed terms were sorted by their frequencies and for each cluster, the thirty most frequent terms were used to label the obtained clusters. The best TF-IDF terms of each journal cluster are shown in Table 4.

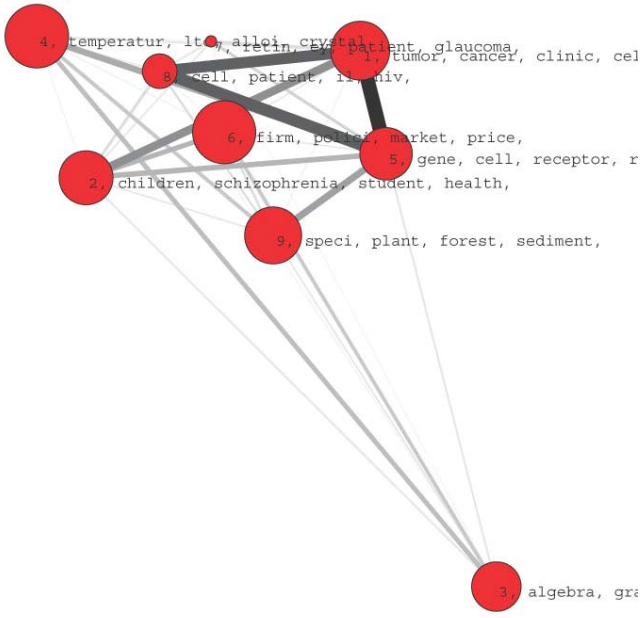

Figure 2 Network structure of the 9 journal clusters (the 2rd level).

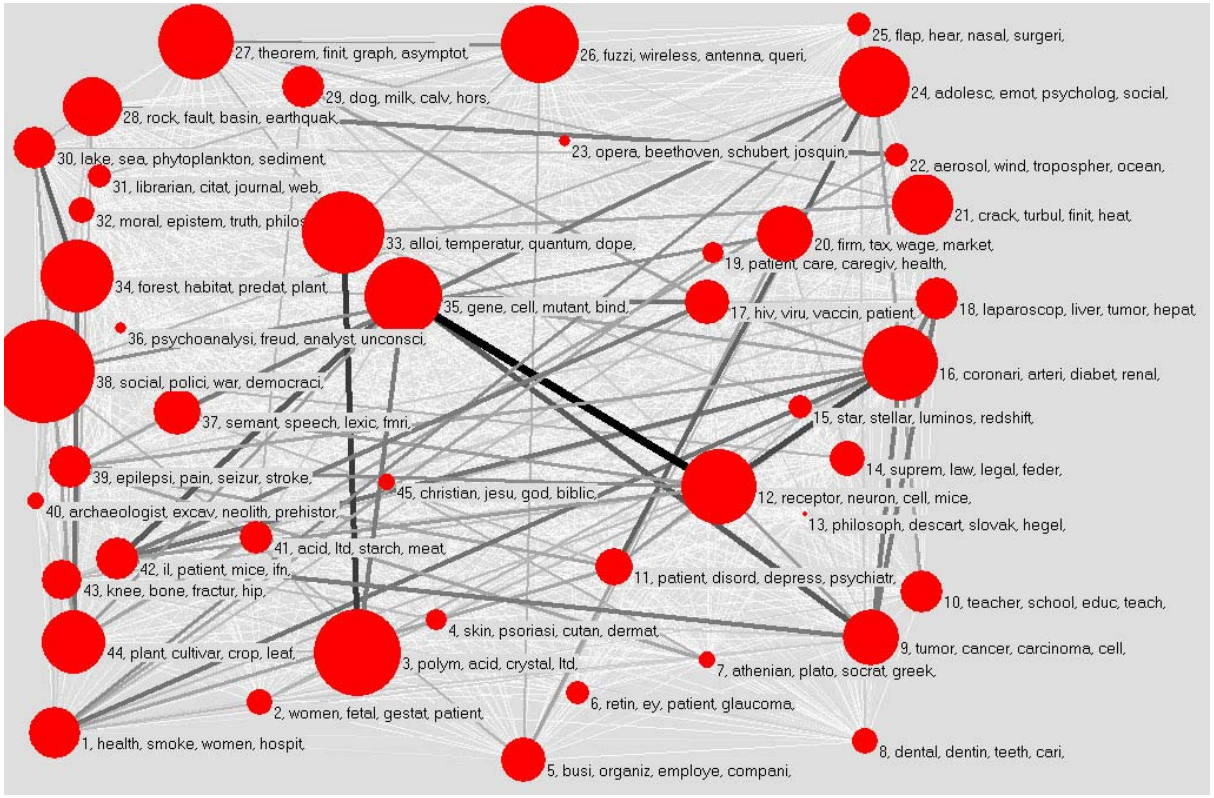

Figure 3 Network structure of the 45 journal clusters (the first level)

Table 4: the three most important journals of each cluster according to a modified version of Google PageRank algorithm

Cluster 1	*Cluster 2*	*Cluster 3*	*Cluster 4*	*Cluster 5*
1. Milkb Quar 2. Annu Rev Pub Healt 3. A J Epide	1. Twin Res 2. I J Obst &Gyna 3. Hum Repr	1. Chem Rev 2. Prog Ploy Sci 3. Acc Chem Res	1.J Inv Derm S P 2. A J Clin Derm 3.J Inve Derm	1. Admi Sci Quar 2. Mis Quar 3. Aca Mana J
Cluster 6	*Cluster 7*	*Cluster 8*	*Cluster 9*	*Cluster 10*
1. Prog Ret eye Res 2. Inv Ophth & Vis Sci 3. Sur Ophth	1. Class Antiq 2. T A Philo Asso 3. A J Philo	1. Crit Rev Ora Bio Med 2. J Dent Res 3. Dent Mater	1. Nat Rev cancer 2. Cancer cell 3. Ca-cancer J Clin	1. Rev Educ Res 2. A Educ Res J 3. Educa Eval Poly Anal
Cluster 11	*Cluster 12*	*Cluster 13*	*Cluster 14*	*Cluster 15*
1. Arch Gener Pysychi 2. Molec Psychi 3. Bio Psychi	1. Nat Rev Neros 2. Physi Rev 3. Annu Rev Neros	1. Filoso Casopis 2. Filozo	1. Yale Law J 2. Univ Chica Law Rev 3. Stanf Law REv	1. Annu Rev Astron & Astrop 2. Astrop J Supp S 3. Astroph J
Cluster 16	*Cluster 17*	*Cluster 18*	*Cluster 19*	*Cluster 20*
1. Annu Rev Med 2. N Eng J Med 3. Circul	1. Nat Rev Microb 2. Clin Microb Rev 3. Lanc Infec Disea	1. Gastroe 2. Anna Surg 3. Hepato	1. Geronto 2. A J Criti Care 3. Nurs Res	1. Quart J Econ 2. J Econ Liter 3. J Finace
Cluster 21	*Cluster 22*	*Cluster 23*	*Cluster 24*	*Cluster 25*
1. Annu Rev flu Mecha 2. Prog Energ Comb SCi 3. J Mecha Phys Sol	1. J Hydrom 2. Clima Dynam 3. J Clima	1. J A Musico Soci 2. Musi Theo Spectr 3. Music Anal	1. Annu Rev Psych 2. Psych Meth 3. Psycho Bull	1. Audio Neuro-oto 2. Ear & Hear 3. Laryngp
Cluster 26	*Cluster 27*	*Cluster 28*	*Cluster 29*	*Cluster 30*
1. Acm Comp Surv 2. J Acm 3. J Machi learn Res	1. J Roy Stat Soc S B 2. BioStat 3. J A Mat Sco S B	1. Rev Minerl & Geoche 2. Earth Sci Rev 3. Annu Rev Earth & Plane Sci	1. Veter Res 2. J Feli Med & Surg 3. J Dairy SCi	1. Oceanog & Marine Bio 2. Fish & fisher 3. Prog Oceanog
Cluster 31	*Cluster 32*	*Cluster 33*	*Cluster 34*	*Cluster 35*
1. Libra & Infor Sci Res 2. P Libra & Acad 1. 3 Colle & Res Libra	1. Ethics 2. Philos & Pub Affai 2. 3. J Philos	1. Rev Mode Physi 2. Nat Mater 3. 3. Mater Sci & Eng R Rep	1. Annu Rev Ecolo Evol & Syst 2. syst Bio 4. 3. A Muse Novi	1. Nat Rev Molec cell Bio 2. Nat Rev genetics 5. 3. Devel cell
Cluster 36	*Cluster 37*	*Cluster 38*	*Cluster 39*	*Cluster 40*
1. Psychoa Dial 2. Psychoa quart 3. J A Psycho Asso	1. Psycho Rev 2. Behav & Bra Sci 3. Tre Cogn Sci	1. A Polit Sci Rev 2. Annu Rev Soc 3. A Soci Rev	1. Lanc Neuro 2. Brain 3. Anna Neuro	1. J Anthr Archae 2. A Antiq 3. J Archa Met & Theo
Cluster 41	*Cluster 42*	*Cluster 43*	*Cluster 44*	*Cluster 45*
1. Crit Rev food Sci & Nutri 2. I J food Microb 3. A J grape & winde Res	1. Nat Rev Immu 2. Annu Rev Immu 3. Nat Immu	1. Exerc & Sport Sci Rev 2. J bone Min Res 3. Bone	1. Global Chan Bio 2. Criti Rev plant Sci 3. Adva Envir Res	1. J Ear Chris Stud 2. J Bib Liter 3. N Testa Stud

To visualize the clustering result, the structural mapping of the 9 categorizations obtained from the hybrid strategy is presented in Figure 2. For each cluster, the three most important terms are shown. The network is visualized by Pajek (Batagelj and Mrvar, 2003). The edges represent the strength of cross-citation links and darker colour represents more links between the clusters pairs. The circle size is proportional to the number of journals in each cluster.

Table 5 the 20 best TF-IDF terms describing the 9 hybrid citation-textual clusters

Cluster	*Best 10 terms*
1	health;smoke;women;hospit;children;physician;alcohol;cancer;care;clinic;risk;medic;adolesc;ltd;ci;drug;smoker;mortal;infant;intervent;tobacco
2	pregnanc;women;fetal;gestat;patient;ivf;preterm;vagin;matern;uterin;cesarean;endometriosi;ovarian;embryo;oocyt;endometri;icsi;sperm;fetus;infant;
3	catalyst;polym;acid;crystal;ltd;ligand;wilei;nmr;ion;angstrom;bond;adsorpt;hydrogen;solvent;atom;copolym;poli;temperatur;oxid;molecul;polymer;
4	skin;psoriasi;cutan;dermat;lesion;keratinocyt;dermatolog;melanoma;hair;clinic;acn;wound;cell;atop;uvb;diseas;melanocyt;dermatologist;therapi;epiderm;
5	firm;busi;organiz;employe;compani;market;custom;brand;retail;supplier;corpor;ltd;advertis;strateg;manageri;manag;leadership;price;ethic;job;wilei
6	corneal;retin;ey;patient;glaucoma;acuiti;iop;iol;macular;cataract;intraocular;lasik;ocular;surgeri;cornea;retina;len;choroid;postop;myopia;vitrectomi
7	roman;athenian;plato;socrat;greek;ovid;cicero;homer;poem;aristotl;poet;horac;herodotu;catullu;euripid;iliad;poetri;literari;aeneid;plautu;aristophan
8	periodont;dental;dentin;teeth;cari;patient;implant;mandibular;enamel;gingiv;orthodont;tooth;bone;dentur;maxillari;resin;oral;incisor;endodont
9	tumor;cancer;carcinoma;cell;prostat;breast;tumour;gene;chemotherapi;p53;malign;apoptosi;therapi;bladder;clinic;radiotherapi;metastat;protein
10	student;teacher;school;educ;teach;classroom;curriculum;learn;learner;faculti;instruct;skill;literaci;academ;undergradu;children;pedagog;colleg;medic
11	schizophrenia;patient;disord;depress;psychiatr;suicid;antipsychot;symptom;sleep;bipolar;mental;antidepress;ptsd;anxieti;clinic;psychosi;schizophren;psychot
12	rat;receptor;neuron;cell;mice;protein;mrna;ca2;brain;gene;insulin;kinas;ltd;muscl;inhibit;patient;inc;synapt;acid;gaba;cortex;dose;hippocamp
13	philosophi;philosoph;descart;slovak;hegel;ethic;moral;masaryk;kant;husserl;cogito;frege;kierkegaard;patocka;heidegg;sartr;metaphys;stoic;ontolog;plato
14	court;suprem;law;legal;feder;doctrin;litig;wto;crimin;judici;justic;lawyer;statut;claus;amend;corpor;liabil;professor;tax;jurisdict;tort;plaintiff;attornei
15	galaxi;star;stellar;luminos;redshift;galact;ngc;solar;telescop;dwarf;supernova;accret;quasar;cospar;pulsar;rai;emiss;nebula;disk;radio;agn;halo;cloud;interstellar
16	coronari;arteri;diabet;renal;transplant;clinic;ventricular;diseas;hypertens;cardiac;therapi;myocardi;blood;stent;pulmonari;aortic;insulin;hospit;graft
17	infect;hiv;viru;vaccin;patient;protein;viral;cell;gene;hcv;antibodi;mice;strain;pcr;malaria;immun;antigen;rna;tuberculosi;cd4;clinic;parasit
18	laparoscop;liver;tumor;hepat;resect;pancreat;gastric;surgeri;cancer;pylori;endoscop;postop;surgic;ct;carcinoma;bowel;diseas;esophag;lesion;mri
19	nurs;patient;care;caregiv;health;student;hospit;educ;clinic;women;staff;midwiv;midwiferi;interview;pain;profession;intervent;satisfact;home
20	price;firm;tax;wage;market;pavement;polici;trade;economi;monetari;capit;earn;invest;forecast;traffic;incom;investor;welfar;asset;stock
21	crack;turbul;finit;heat;flame;shear;vibrat;concret;beam;reynold;veloc;acoust;elast;vortex;temperatur;combust;equat;steel;convect;flow;wave
22	cloud;aerosol;wind;troposphere;ocean;atmospher;stratospher;radar;convect;ozon;rainfal;sst;ionospher;forecast;climat;tropic;flux;cyclon;monsoon;sea
23	music;opera;beethoven;schubert;josquin;symphoni;bach;tonal;song;motet;handel;brahm;sonata;adorno;debussi;schoenberg;piano;minuet;schenker
24	children;adolesc;emot;psycholog;social;child;anxieti;student;women;cognit;school;sexual;violenc;patient;parent;depress;disord;symptom;abus
25	flap;hear;nasal;surgeri;cochlear;postop;ear;surgic;nerv;implant;cleft;laryng;endoscop;neck;sinu;facial;vestibular;tumor;children;vocal
26	algorithm;fuzzi;wireless;antenna;queri;semant;robot;qo;packet;graph;xml;user;bandwidth;multicast;network;wilei;ltd;web;fault;cdma;server;bit
27	algebra;theorem;finit;graph;asymptot;infin;equat;polynomi;manifold;ltd;inc;let;nonlinear;banach;inequ;semigroup;singular;wilei;convex;cohomolog;conjectur
28	seismic;rock;fault;basin;earthquak;magma;sediment;tecton;mantl;crustal;volcan;subduct;magmat;isotop;metamorph;crust;ocean;basalt;lithospher;sedimentari
29	cow;dog;milk;calv;hors;diet;broiler;cattl;herd;pig;dairi;breed;carcass;heifer;lamb;goat;silag;rumen;cat;fed;sheep;rumin;anim;infect
30	fish;lake;sea;phytoplankton;sediment;speci;fisheri;habitat;river;ocean;spawn;estuari;benthic;larva;trout;biomass;reef;zooplankton;salmon;water
31	librari;librarian;citat;journal;web;metadata;catalog;bibliometr;literaci;academ;book;user;librarianship;scholarli;onlin;servic;internet;bibliograph;digit
32	philosoph;moral;epistem;truth;philosophi;metaphys;kant;argument;semant;argu;wittgenstein;epistemolog;realism;frege;god;logic;ontolog;heidegg;aquina;
33	film;alloi;temperatur;quantum;dope;magnet;crystal;optic;si;beam;atom;laser;spin;anneal;ion;diffract;dielectr;silicon;microstructur;electron;gan
34	speci;forest;habitat;predat;plant;soil;seedl;prei;bird;tree;larva;egg;genu;femal;forag;male;parasitoid;taxa;veget;breed;seed;nest;beetl;season
35	protein;gene;cell;mutant;bind;dna;transcript;kinas;receptor;mutat;enzym;genom;phosphoryl;rna;acid;chromosom;peptid;mrna;membran;coli;inhibitor;mice
36	psychoanalyt;psychoanalysi;freud;analyst;unconsci;countertransfer;psychoanalyst;psychic;dream;analysand;patient;jung;fantasi;ego;psychotherapi;narcissist
37	phonolog;semant;speech;lexic;fmri;word;task;verb;sentenc;languag;children;cognit;memori;perceptu;stimuli;stimulu;auditori;cortex;cue;speaker;brain
38	polit;social;polici;war;democraci;democrat;parti;women;discours;religi;reform;crime;sociolog;urban;polic;elector;geographi;labour;ideolog;essai;vote
39	epilepsi;pain;seizur;stroke;aneurysm;clinic;cerebr;migrain;headach;brain;spinal;lesion;arteri;neurolog;mri;diseas;eeg;nerv;parkinson
40	archaeolog;archaeologist;excav;neolith;prehistor;potteri;settlement;maya;ritual;palaeolith;burial;bronz;assemblag;monument;roman;lithic;archaic
41	chees;acid;ltd;starch;meat;milk;flour;ferment;wine;antioxid;protein;juic;cook;food;monocytogen;pulp;wheat;cultivar;aroma;dough;temperatur
42	cell;il;patient;mice;ifn;cytokin;cd4;immun;receptor;cd8;antigen;gene;hla;antibodi;protein;allergen;asthma;arthriti;tnf;lymphocyt;diseas;tumor
43	knee;bone;fractur;hip;arthroplasti;tendon;femor;ligament;injuri;pain;muscl;bmd;athlet;flexion;ankl;cartilag;radiograph;spine;tibial;screw
44	soil;plant;cultivar;crop;leaf;water;ltd;wheat;sludg;shoot;seedl;irrig;seed;biomass;sediment;ha;weed;wastewat;tillag;groundwat;rice;manur
45	gospel;christian;jesu;god;biblic;hebrew;psalm;testament;theologi;luke;paul;church;bibl;divin;sermon;text;jeremiah;theolog;prophet;scriptur;jewish;christ

For the lower level partition with 45 clusters, the best 20 representative terms in each cluster is shown in Table 5. The top three journals of each cluster are listed in Table 6 (Cluster #13 only has two journals). Its cluster structure is visualized in Figure 3 analogously to Figure 2. We illustrate the hierarchical structure between the two partitions in Figure 4, which provides different resolutions of the science mapping. The clusters of different level partition are

annotated by the number of journals within this cluster and it related main subjects. For instance, cluster #1 stands for clinical medicine and neuroscience, and has the following substructure: 9 subfields including obstetrics, dental dermatology, cancer, medicine, surgery, audio, neurology and bone.

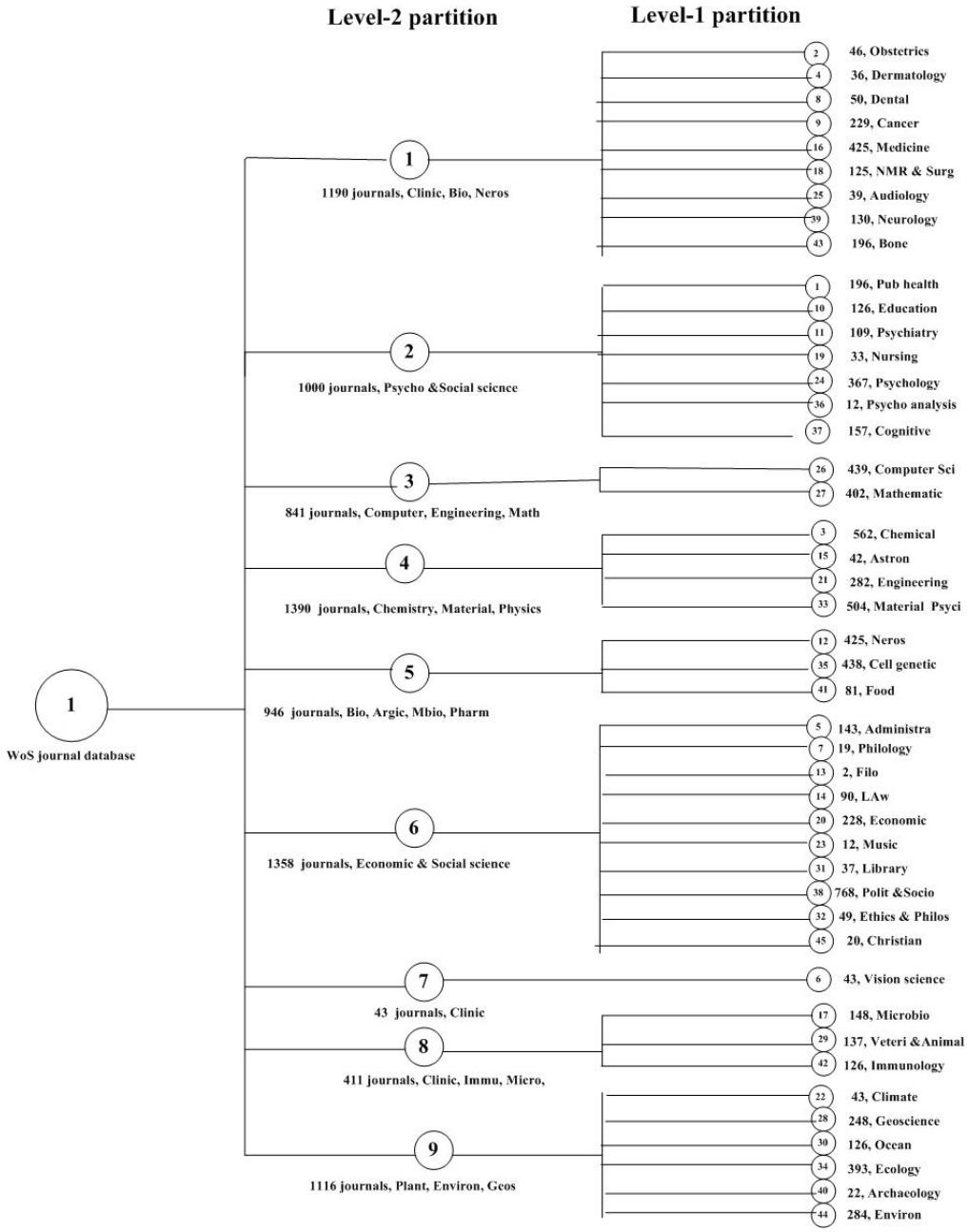

Figure 4. the hierarchical structure of the clustering results

Conclusion

In this study we have presented a new hybrid clustering strategy based on graph model, which proved efficient and extremely fast. It is able to automatically provide optimum partitions at several hierarchical levels without any previous input. By combining textual and citation information, the strategy provided more robust cluster structures than hybrid clustering strategies based on vector space model.

Even for given number of clusters the new method outperformed analogous cluster algorithms based on the vector space model. The self-optimization scheme of the Louvain method provided an optimum two-level hierarchical cluster structure. The cognitive analysis based on

the textual component provided information for labelling and term annotation, the ranked journals and the visualization of the cluster structure also verified the validity of the new strategy.

The hybrid strategy is expect to provide a powerful tool to scientometrics and informetrics, as it can handle large-scale data, carry out the immediate partition, automatically optimize the cluster and provide a hierarchical system in practically one process and in an very short time.

Acknowledgement

Bart De Moor is a full professor at the Katholieke Universiteit Leuven, Belgium. Research supported by (1) China Scholarship Council (CSC, No. 2006153005); (2)Research Council KUL: GOA Ambiorics, GOA MaNet, CoE EF/05/006 Optimization in Engineering(OPTEC), IOF-SCORES4CHEM, several PhD/postdoc & fellow grants; (3)FWO: PhD/postdoc grants, projects: G0226.06 (cooperative systems and optimization), G0321.06 (Tensors), G.0302.07 (SVM/Kernel), G.0320.08 (convex MPC), G.0558.08 (Robust MHE), G.0557.08 (Glycemia2), G.0588.09 (Brain-machine) research communities (ICCoS, ANMMM, MLDM); G.0377.09 (Mechatronics MPC); (4)IWT: PhD Grants, Eureka-Flite+, SBO LeCoPro, SBO Climaqs, SBO POM, O&O-Dsquare; (5)Belgian Federal Science Policy Office: IUAP P6/04 (DYSCO, Dynamical systems, control and optimization, 2007-2011); (6)EU: ERNSI; FP7-HD-MPC (INFSO-ICT-223854), COST intelliCIS, FP7-EMBOCON (ICT-248940); (7)Contract Research: AMINAL; Other: Helmholtz: viCERP; ACCM; Bauknecht; Hoerbiger; (8) Flemish Government: Center for R&D Monitoring (ECOOM).

References

Baeza-Yates, R. & Ribeiro-Neto, B. (1999). *Modern information retrieval*. Cambridge: Addison-Wesley.

Batagelj, V. & Mrvar, A. (2002). Pajek - Analysis and visualization of large networks. *Graph Drawing*, 2265, 477-478.

Blondel, V. D., Guillaume ,J.-L., Lambiotte, R. and E. Lefebvre. (2008) Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment.* 10, 10008.

Braam, R.R., Moed, H.F. & van Raan, A.F.J. (1991a). Mapping of science by combined cocitation and word analysis, Part 1: Structural aspects. *Journal of the American Society for Information Sci*ence, 42 (4), 233–251.

Braam, R. R., Moed, H. F. & van Raan, A. F. J. (1991b). Mapping of Science by Combined Cocitation and Word Analysis. 2. Dynamic Aspects. *JASIS*, 42, 252-266.

Calado, P., Ribeiro-Neto, B., Ziviani, N., Moura, E. & Silva, I. (2003). Local versus global link information in the Web. *ACM Transactions on Information Systems*, 21, 42-63.

Calado, P., Cristo, M., Goncalves, M. A., de Moura, E. S., Ribeiro-Neto, B. & Ziviani, N. (2006). Link-based similarity measures for the classification of Web documents. *JASIST*, 57, 208-221.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486, 75-174.

Glenisson, P., Glänzel, W., Janssens, F. & De Moor, B. (2005). Combining full text and bibliometric information in mapping scientific disciplines. *Information Processing & Management*, 41, 1548-1572.

Hatcher, E. & Gospodnetić, O. (2004). *Lucene in Action*. New York: Manning Publications Co.

He, X.; Ding, C. H. Q.; Zha, H. & Simon, H. D.(2001) Automatic Topic Identification Using Webpage Clustering *Proceedings of the 2001 IEEE International Conference on Data Mining.*

Hedges, L.V. & Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego: Academic Press.

Hubert, L & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2 (1): 193–218.

Jain, A. & Dubes, R. (1988). *Algorithms for clustering data*. New Jersey: Prentice Hall.

Jain, A. K. (2001). Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8), 651-666.

Janssens, F., Leta, J., Glänzel, W. & De Moor, B. (2006a). Towards mapping library and information science. Information Processing & Management, 42(6), 1614–1642.

Janssens, F., Tran Quoc, V., Glänzel, W. & De Moor, B. (2006b). Integration of textual content and link information for accurate clustering of science fields. In V. P. Guerrero-Bote (Ed.), Proc. of the I Intl. Conf. on Multidisciplinary Information Sciences and Technologies (InSciT2006) (pp. 615-619), Mérida, Spain.

Janssens, F., Glänzel, W., & De Moor, B. (2008). A hybrid mapping of information science. *Scientometrics*, 75(3), 607-631.

Joachims, T., Cristianini, N. & Shawe-Taylor, J. (2001). Composite kernels for hypertext categorisation. In *Proceedings of the 18th International Conference on Machine Learning* (ICML) (pp. 250-257).

Lambiotte, R. & Panzarasa, P. Communities (2009). Knowledge Creation, and Information Diffusion. *Journal of Informetrics*, 3(3), 180-190

Modha, D. S. & Spangler, W. S. (2000). Clustering hypertext with applications to web searching. *ACM Conference on Hypertext* (pp. 143-152).

Mullins, N., Snizek, W. & Oehler, K. (1988). *The structural analysis of a scientific paper*. In A. F. J. vanRaan (Ed.), Handbook of quantitative studies of science and technology (pp. 81–105). New York: Elsevier Science.

Newman, M. (2004). Detecting community structure in networks. Eur. Phys. J. B, 38, 321–330.

Rafols, I. & Leydesdorff, L. (2009). Content-based and Algorithmic Classifications of Journals: Perspectives on the Dynamics of Scientific Communication and Indexer Effects. *J. Am. Soc. Inf. Sci. Technol*., 60(9), 1532-2882

Salton, G. & McGill, M. J. (1986). *Introduction to modern information retrieval*. New York: McGraw-Hill, Inc.

Strehl, A., & Ghosh, J. (2002). Cluster Ensembles-A Knowledge Reuse Framework for Combining Multiple Partitions. *Journal of machine learning research*, 3, 583—617.

Snizek, W., Oehler, K. & Mullins, N. (1991). Textual and nontextual characteristics of scientific papers: Neglected science indicators. *Scientometrics*, 20(1), 25–35.

Tan Pang-Ning, Steinbach Michael and Kumar Vipin (2005). Introduction to Data Mining.

Wang, Y. & Kitsuregawa, M. (2002). Evaluating contents-link coupled web page clustering for web search results. In *Proc. of the 11th intl. conf. on Information and knowledge management* (CIKM) (pp. 499-506).

Zhang L., Liu X.H., Janssens F., Linag L & Glanzel W. (2010). Subject clustering analysis based on ISI category classification. *Journal of Informetrics*, 4(2), 185-193.