Jointly published by Akadémiai Kiadó, Budapest and Springer, Dordrecht

Scientometrics, Vol. 75, No. 3 (2008) 607–631
DOI: 10.1007/s11192-007-2002-7

A hybrid mapping of information science

FRIZO JANSSENS,[a,b] WOLFGANG GLÄNZEL,[b,c] BART DE MOOR[a]

[a] *K. U. Leuven, Department of Electrical Engineering ESAT-SCD, Leuven (Belgium)*
[b] *K. U. Leuven, Steunpunt O&O Indicatoren, Leuven (Belgium)*
[c] *Hungarian Academy of Sciences, ISPR, Budapest (Hungary)*

Previous studies have shown that hybrid clustering methods that incorporate textual content and bibliometric information can outperform clustering methods that use only one of these components. In this paper we apply a hybrid clustering method based on Fisher's inverse chi-square to integrate full-text with citations and to provide a mapping of the field of information science. We quantitatively and qualitatively asses the added value of such an integrated analysis and we investigate whether the clustering outcome is a better representation of the field by comparing with a text-only clustering and with another hybrid method based on linear combination of distance matrices. Our dataset consists of almost 1000 articles and notes published in the period 2002–2004 in 5 representative journals. The optimal number of clusters for the field is 5, determined by using a combination of distance-based and stability-based methods. Term networks present the cognitive structure of the field and are complemented by the most representative publications. Three large traditional sub-disciplines, particularly, information retrieval, bibliometrics/scientometrics, and more social aspects, and two smaller clusters about patent analysis and webometrics, can be distinguished.

Introduction

The long-term goal of this research is an accurate unsupervised clustering of science or technology fields, towards the detection of new emerging fields or hot topics.

Received September 18, 2007

Address for correspondence:
FRIZO JANSSENS
K.U. Leuven, Department of Electrical Engineering ESAT-SCD
Kasteelpark Arenberg 10, B-3001 Leuven, Belgium
E-mail: Frizo.Janssens@esat.kuleuven.be

0138–9130/US $ 20.00

The idea of combining bibliometric or citation information with textual content is not new for it has already been pursued to obtain improved performance in information retrieval (e.g., [CALADO & AL., 2003; COHN & HOFMANN, 2001]), bibliometric mapping of science [MULLINS & AL., 1988; SNIZEK & AL., 1991; BRAAM & AL., 1991; GLENISSON & AL., 2005; JANSSENS & AL., 2006B], clustering (a.o., [MODHA & SPANGLER, 2000; WANG & KITSUREGAWA, 2002, JANSSENS, 2007A]), and classification (e.g., [JOACHIMS & AL., 2001; CALADO & AL., 2006]).

Sometimes textual information can indeed indicate similarities that are not visible to bibliometric techniques, and vice versa. As an example, we encountered two papers with bibliographic coupling similarity equal to 0 (i.e., they have no common references), but with more than 95% textual cosine similarity. Both papers were by Ding and Foo and were published in *Journal of Information Science* (Appendix: DING & 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. However, both papers were correctly identified as being very similar by the textual 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 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.

In an earlier study, the concept structure of (library and) information science (IS) was obtained by full-text mining of almost 1000 articles and notes published in the period 2002–2004 in 5 representative journals with strong focus on information science [JANSSENS & AL., 2006A]. The optimum solution for this text-based clustering of IS was found for six clusters. Besides two clusters in bibliometrics, one cluster was found in information retrieval, another one containing general and miscellaneous issues, and webometrics and patent studies were identified as small but emerging clusters within IS. In that study only the 'pure' text corpus was analyzed, excluding any bibliographic or bibliometric components. However, the authors have assessed the performance of clustering and classification algorithms using various data integration schemes on a dataset with bioinformatics-related publications [JANSSENS & AL., 2006B]. The best outcome was obtained by hybrid methods that exploited both text and citations. An integration method based on Fisher's inverse chi-square and another one based on linear combination of distance matrices were among the best methods and significantly

outperformed corresponding text-only and link-only methods, as well as a concatenation of vectors. In the present study we use these methods to map IS by using the full-text information as well as citations, and we compare the results with the text-only clustering.

Dataset

The document set used for our study consists of 914 full-text articles or notes, published between 2002 and 2004 in one of five journals. Table 1 shows the distribution of the 914 documents over the journals. We decided to select these five journals by two reasons; firstly we think that these journals representatively cover the main topics of modern information science including information retrieval, bibliometrics and webmetrics; and secondly, the full-text versions of all publications of these journals were available online. Numerous articles were only available as graphic PDF files, and had therefore to be converted into text using OCR techniques. This procedure, which included careful manual corrections, finally resulted in the limitation to a selection of five journals.

This data set is the same as introduced by JANSSENS & AL. [2006A], except for the exclusion of 24 publications. By matching the articles with the Web of Science (WoS) database of Thomson Scientific (Philadelphia, PA, USA), we noticed that 22 were actually not listed 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 Information Retrieval (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. For comparing hybrid and text-based clustering, the latter one has first been redone on the smaller data set.

Table 1. The distribution of the 914 articles or notes over the 5 selected journals

Journal	Number of papers	%
Information Processing & Management	139	15.2
Journal of the American Society for Information Science and Technology (JASIST)	306	33.5
Journal of Documentation	82	9.0
Journal of Information Science	131	14.3
Scientometrics	256	28.0
Total	914	100.0

Methodology

Text representation

All textual content was indexed with the Jakarta Lucene[1] platform [HATCHER & GOSPODNETIĆ, 2004] and encoded in the vector space model using the TF-IDF weighting scheme [BAEZA-YATES & RIBEIRO-NETO, 1999]. Text pre-processing comprised the following steps:

- Text extraction from full-text papers, preceded by Optical Character Recognition if necessary.
- Automatically removing acknowledgements and cited references from article content.
- Neglecting of author names, stop- and template words, URLs and e-mail addresses and stemming all remaining terms with the PORTER [1980] stemmer.
- Phrase and synonym detection and reintroducing author names being part of a phrase (see [DUNNING, 1993; MANNING & SCHÜTZE, 2000]).

The dimensionality of the *term-by-document* matrix was reduced to 150 factors by latent semantic indexing (LSI) [DEERWESTER & AL., 1990; BERRY & AL., 1995]. Text-based similarities were calculated as the cosine of the angle between the vector representations of two papers [SALTON & MCGILL, 1986].

Figure 1 presents an overview of the textual analysis.

Citation analysis

The cosine measure used to quantify the text-based similarity between any two documents can analogously be used with Boolean input vectors indicating the cited references in an article, or indicating all citing articles. This corresponds to bibliographic coupling [KESSLER, 1963] and co-citation, respectively, which are two citation-based measures of similarity. In the present study we use bibliographic coupling and combine it with text-based similarities in order to obtain an integrated measure that can be used by a clustering algorithm.

Clustering

To subdivide the IS papers into clusters we used the agglomerative hierarchical clustering algorithm with Ward's method [WARD, 1963; JAIN & DUBES, 1988]. It is a 'hard' clustering algorithm, which means that every publication is assigned to exactly 1 cluster. We determined the optimal number of clusters by observing the dendrogram, curves with mean Silhouette coefficient for various numbers of clusters [KAUFMAN & ROUSSEEUW, 1990], and by using the stability-based method of BEN-HUR & AL. [2002].

[1] http://lucene.apache.org/, visited in September 2007.

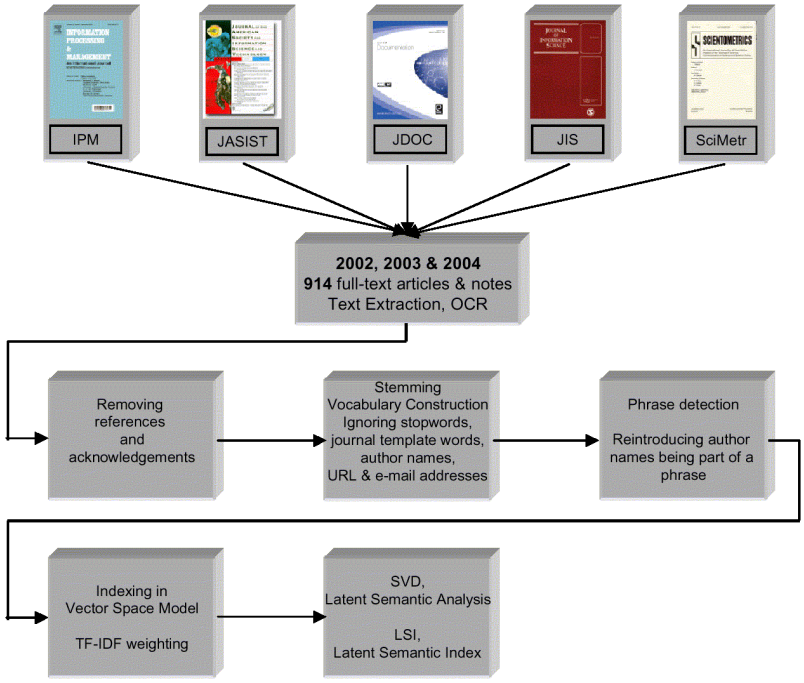

Figure 1. Overview of the textual analysis

The latter method allows to visually and quantitatively determine the most stable number of clusters *k* by inspection of a stability diagram (see Figure 5 for an example).

The main idea is that the perceived structure should remain stable if only a subsample of objects is available, or if noise objects are added to the set. Multiple subsamples (e.g., 200) are randomly drawn from the data 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). Next, the overlap between each pair of clustered subsamples is quantified by using, for example, the Jaccard coefficient (for a specific number of clusters). Each number of clusters 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 the clustered subsamples, and the more stable the clustering solutions with that specific number of clusters. A curve representing a certain number of clusters can be interpreted as how many percent of the subsample pairs (Y-axis) have a Jaccard value lower than or equal to the corresponding values on the X-axis. 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.

The Silhouette value for a document ranges from –1 to +1 and measures how similar it is to documents in its own cluster vs. documents in other clusters [ROUSSEEUW, 1987]. The mean Silhouette value for all documents is a measurement of the overall quality of a clustering solution with a specific number of clusters.

The requisite input for many clustering algorithms includes mutual distances between all objects (scientific publications here). These distances can be based on the text, on citations, or on a combination of both information sources. In the next subsections we describe linear combinations of distance matrices as well as Fisher's inverse chi-square method.

Weighted linear combination of distance matrices

For both data sources, i.e. the normalized *term-by-document* matrix A and the normalized *cited_references-by-document* matrix B, square distance matrices D_T and D_{BC} can be constructed as follows:

$$D_T = O_N - A'.A$$

$$D_{BC} = O_N - B'.B\,,$$

with N the number of documents and O_N a square matrix of dimensionality N with all ones. 'BC' refers to bibliographic coupling. These 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:

$$D_i = \alpha \cdot D_T + (1-\alpha) \cdot D_{BC}$$

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 & SPANGLER [2000]. Although being an attractive, easy, and reasonably scalable integration method, caution should be taken as a linear combination might neglect different distributional characteristics of various data sources [JANSSENS, 2007A : 122–124]. Such differences in distributions might lead to an unequal or unfair contribution of data sources in the ultimate integrated data, and might thus yield suboptimal results by implicitly favoring text over bibliometric information or vice versa.

Fisher's inverse chi-square method

As a plain linear combination might not be optimal for integrating textual and bibliographic coupling (BC) 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 [HEDGES & OLKIN, 1985]. 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 information source.

Figure 2 illustrates the concept of distance integration by Fisher's inverse chi-square method. All text-based and link-based document distances in D_T and D_{BC}, as described above, are transformed to *p*-values with respect to the cumulative distribution function of distances for randomized data. This randomization is a necessary condition for having valid *p*-values. In our setting, a *p*-value means the probability that the similarity between two documents could be at least as high just by chance.

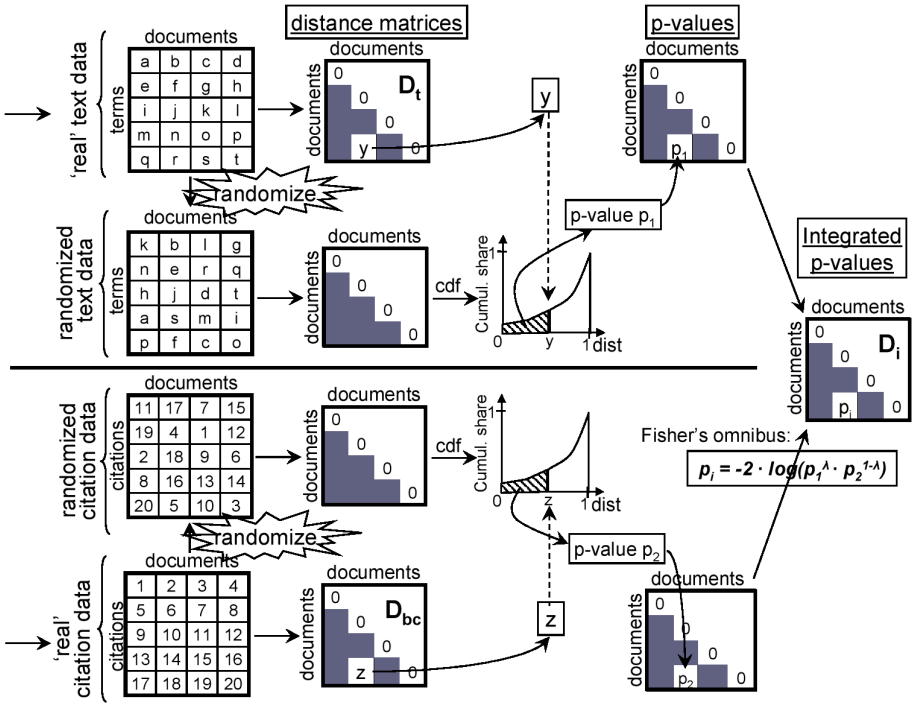

Figure 2. Distance integration by using Fisher's inverse chi-square method.
The ultimate matrix with integrated *p*-values can be used for clustering

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. Blind randomizations might destroy important properties of human

language. We considered different randomization schemes and finally opted for the somewhat conservative randomization which maintains the relative importance between terms by keeping the inverse document frequency 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.

For a correct application of Fisher's inverse chi-square method the input test statistics should be continuous. However, this is not the case for the sparse BC since most pairs of scientific articles do not have any reference in common. 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 BC' between two papers *x* and *y* is then $(Nxy + 0.01)/\sqrt{Nx \cdot Ny}$, 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. In practice, however, it is still a finite set of discrete values, so we also superimposed Gaussian noise (standard deviation of 0.0025). The random noise will not deteriorate results since the error to be expected from for instance missing references in the WoS database is much higher.

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 $p_i = -2 \cdot \log(p_1^{\lambda} \cdot p_2^{(1-\lambda)})$. If the null hypothesis is true (i.e., in the case of 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 [HEDGES & OLKIN, 1985]. 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.

In Figure 3, the distributions of *p*-values for the 'real data' 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.

The weight λ can be used to tune the relative importance or 'quality' of both information sources; however, choosing a good value for λ is not straightforward. We propose to define λ by choosing a value *x* for the *smallest but still significant* BC link (e.g., $x = 0.03$) and a value *y* for the smallest text-based similarity that is also still significant (e.g., $y = 0.1$). *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 λ such that both *weakest still significant links* have the same contribution in p_i, by asking that $p_x^{\lambda} = p_y^{1-\lambda}$. Therefore we compensate for the fact that significant similarities are not as numerous in both datasets.

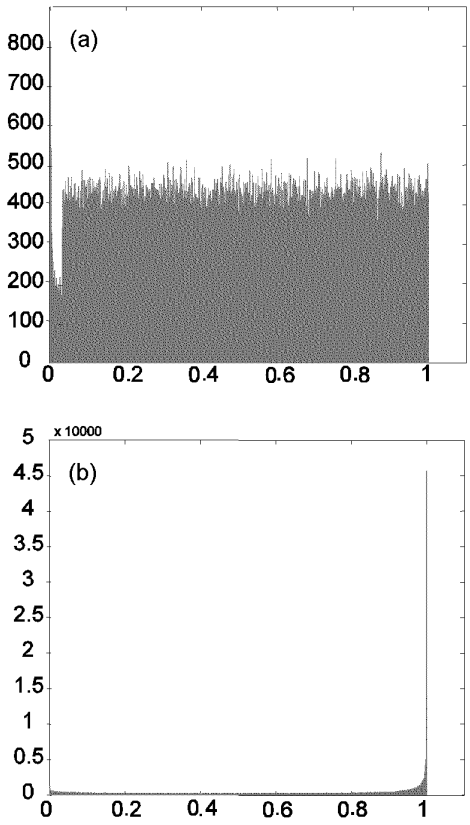

Figure 3. Histogram of *p*-values corresponding to (a) bibliographic coupling and (b) textual pairwise document distances for the real data w.r.t. randomized data

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 more difficult. However, a parameter sweep can still be performed and the difficulty of defining λ is compensated by the augmented performance after applying SVD, especially on the textual data.

Term networks

For visualization of clustering results we determined for each cluster the best words or phrases according to mean TF-IDF weights (see Figure 8 for an example). In a term network, each cluster has its own 'central node', represented by a diamond, which also indicates the number of publications. Each central node points to the best 20 keywords for the cluster. When a keyword 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 [BATAGELJ & MRVAR, 2002].

Results

Optimal number of clusters

Since Silhouette values are based on distances [ROUSSEEUW, 1987], depending on the chosen source of distances different Silhouettes can be calculated. In each case (a), (b) and (c) from Figure 4, we used 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 2. Optimal number of clusters for Fisher's inverse chi-square method as perceived by the stability-based method (Figure 5) and by different mean Silhouette curves in Figure 4 using link-based (a), text-based (b) and integrated distances (c)

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

In the experiments of Figure 4, 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 2). Regarding the optimal number of clusters for hybrid clustering by Fisher's inverse chi-square method, the curve with citation-based Silhouettes (Figure 4(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(c) suggests 4, or maybe 5 clusters, but not more.

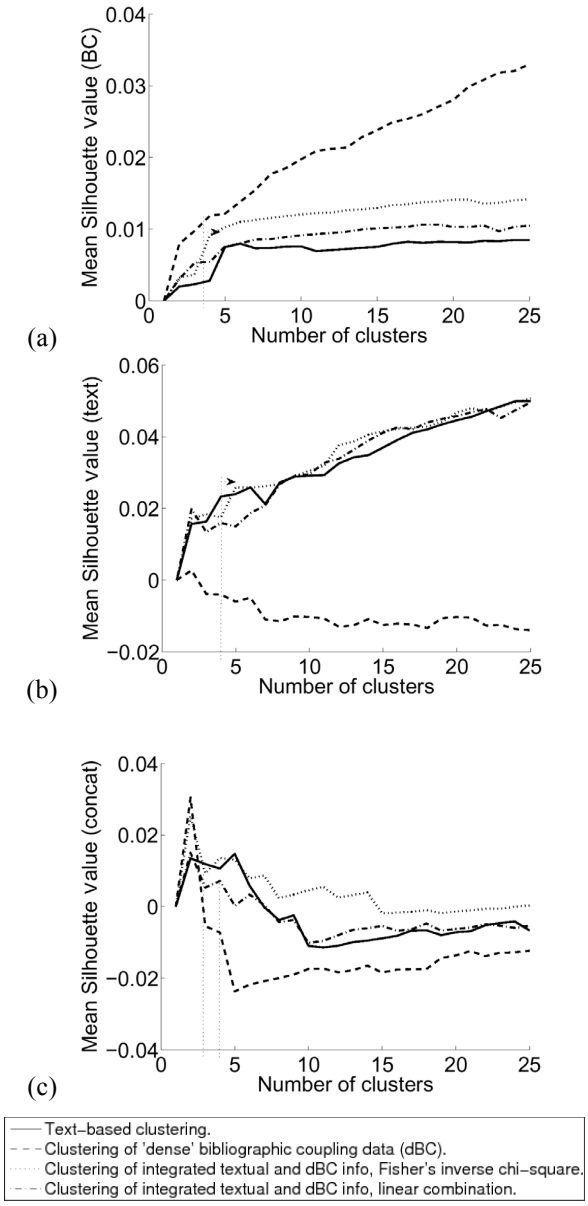

Figure 4. Silhouette curves with mean Silhouette coefficient for clustering solutions of 2 up to 25 clusters for text-only clustering, link-only clustering, integrated clustering with Fisher's inverse chi-square method, and integrated clustering by linear combination of document similarities. Silhouette 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

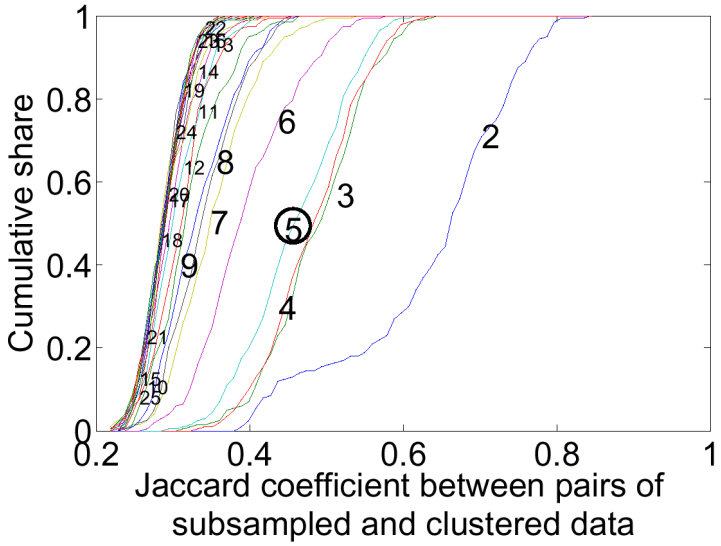

Figure 5. Stability diagram for determining the number of clusters for hybrid clustering with Fisher's inverse chi-square method (BEN-HUR & AL., 2002)

By observing the stability diagram in Figure 5 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 optimal 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.

Comparing Fisher's inverse chi-square method with linco, text-only & link-only clustering

When using the same composite procedure for determining the optimal number of clusters with the linco method, two observations could be made. First, when clustering the linearly combined distance matrices, the optimal 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 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 3. 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 dominating citations in case of plain linear combinations. Furthermore, linco provided somewhat less stable clusterings than Fisher's inverse chi-square method.

Table 3. General appreciation of clustering different data types by observing Silhouette curves in Figure 4. 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 4a)	4	1	2	3
Mean text-based Silhouette value (see Figure 4b)	3 (c=2, c>10) 1 or 2 otherwise	4	1 or 2	3 (c=3..6) 1 or 2 otherwise
Mean 'hybrid' Silhouette value (see Figure 4c)	1 (c=3 or 5) 4 (c=2) 2 or 3 otherwise	1 (c=2) 4 otherwise	2 (c=2, 3 or 5) 1 otherwise	2 or 3

Quite favorable but a little counterintuitive is that, when the validation relies on pure text-based Silhouettes (b), Fisher's inverse chi-square does at 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 Fisher curve. 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 here than the text-only clustering, maybe another illustration of textual data dominating citations. Interestingly, the local maximum of the text-based solution at 6 clusters in (b), and as also described by JANSSENS & AL. [2006A], also decreases to 5 clusters in (c), when evaluated with integrated Silhouette values.

Linkage method

Agglomerative hierarchical clustering can use various strategies to decide which documents or clusters to merge in each iteration step of the algorithm. For example, 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 given 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 error sum of squares over all clusters is minimized [WARD, 1963; KAUFMAN & ROUSSEEUW, 1990]. Other linkage methods exist as well [JAIN & DUBES, 1988], but have not been considered. For single linkage, complete linkage, and UPGMA, in each iteration those documents or clusters with the smallest distance are merged.

The clustering methods that have been experimented 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 6 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 6(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 IS which was discussed in [JANSSENS & AL., 2006A].

Ward's method is the merging criterion of choice for text-based and link-based clustering as well (see Figure 7). 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 & AL. [2003, 2004].

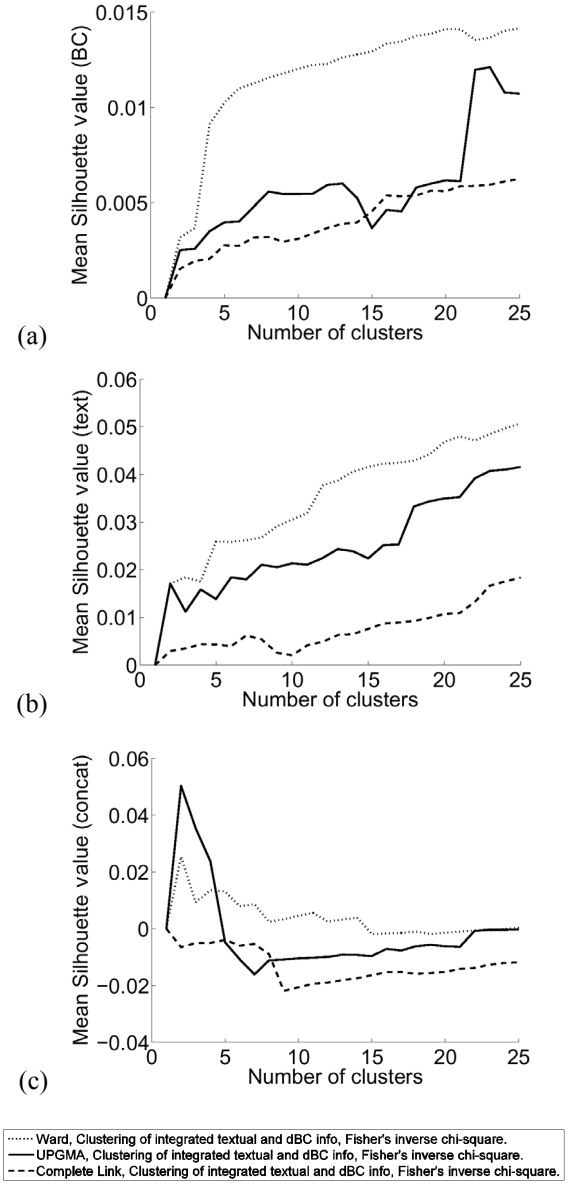

Figure 6. Silhouette curves with mean Silhouette coefficient for clustering solutions of 2 up to 25 clusters, for integrated clustering with Fisher's inverse chi-square method. As shown in the legend, each plot contains three different curves for three different linkage methods: Ward's method, UPGMA, 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 best results

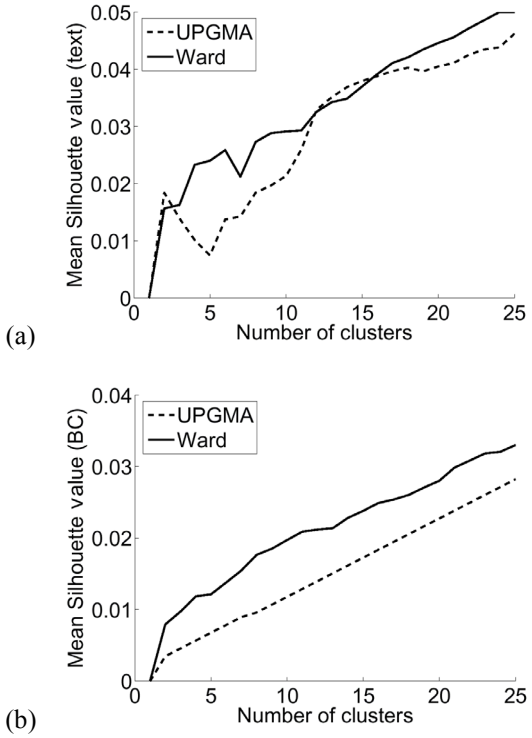

Figure 7. Silhouette curves with mean Silhouette coefficient for clustering solutions of 2 up to 25 clusters, for bibliographic coupling (a) and text-based clustering (b). Each plot contains two different curves for two different linkage methods: UPGMA and Ward's method. Silhouette values are based on 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

Hybrid mapping of the field by using Fisher's inverse chi-square method

Figure 8 presents the cognitive structure of IS 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 (see Table 4). In order to determine these papers, we looked at the largest cosine similarities to the mean cluster profile (centroid). Three large and two smaller clusters can be distinguished. Publications in the three larger classes are concerned with rather traditional sub-disciplines of the IS field, particularly, with Information Retrieval (IR), Bibliometrics/scientometrics 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, etc. The two smaller classes represent relatively new and emerging topics in IS, namely, Patent analysis and Webometrics. Hence, the hybrid clustering result contains the same topics as found by the text-based clustering [JANSSENS & AL., 2006A], except for the merger of the two Bibliometrics clusters. Figure 8 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 liked to the general/Social cluster through the Web use. Finally, the stem *queri* connects Webometrics with IR. Here, *search engin*, web *crawler* and *algorithm* form a strong interface.

Table 4. 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.

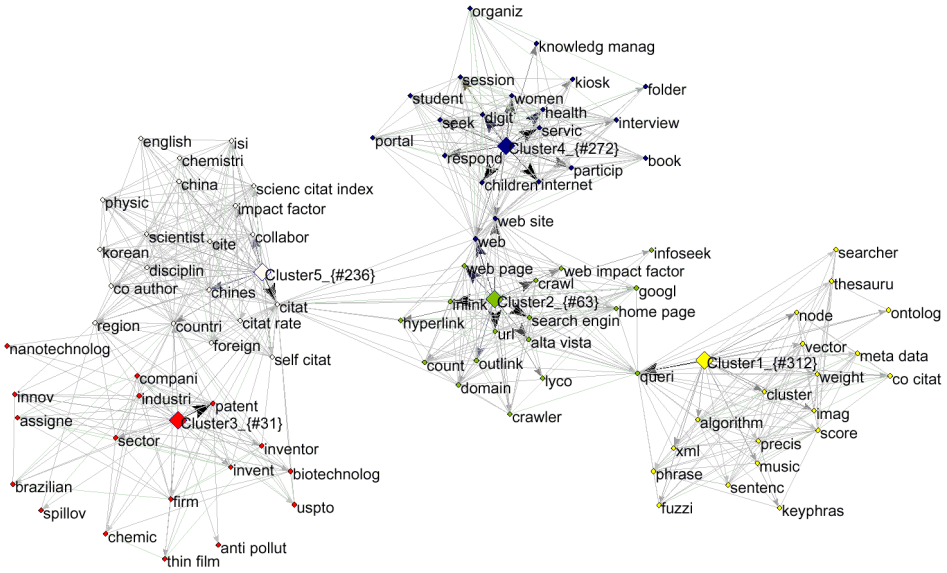

Figure 8. Term networks with for each of five clusters the best 20 stemmed terms or phrases from titles or abstracts according to mean TF-IDF scores

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 λ, automatically determined equal to 0.43 by the method described above, 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.

In Figure 9(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 and Bibliometrics1 was even combined with Patent before being combined with Bibliometrics2 [JANSSENS & AL., 2006A]. 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.

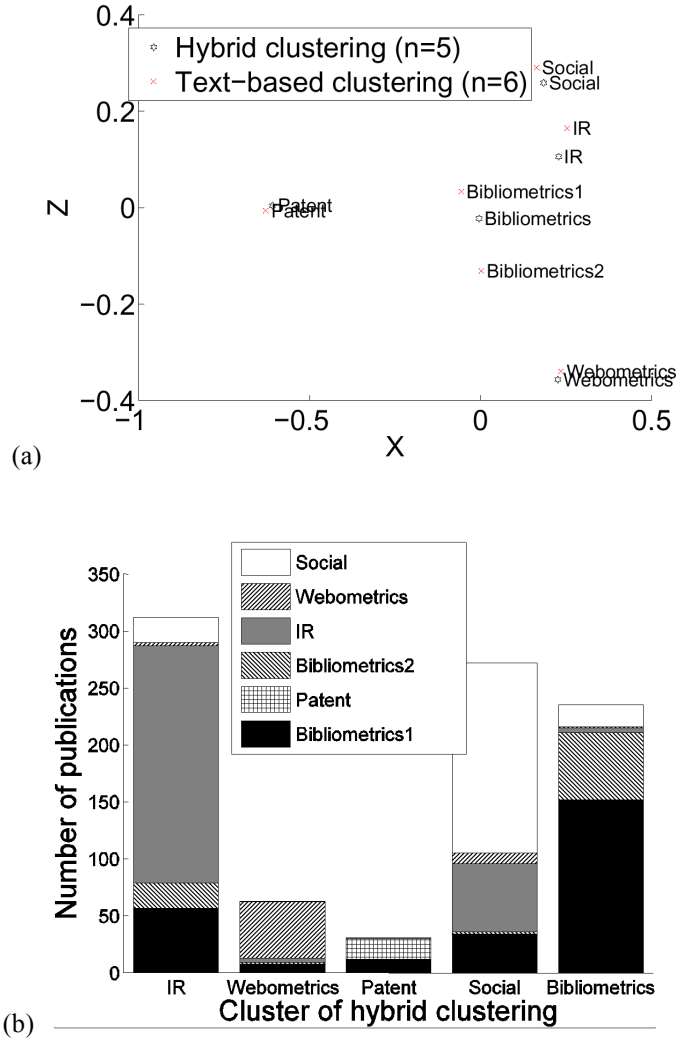

Figure 9. (a) Multidimensional scaling (MDS) plot comparing the cluster centers (centroids) of the six clusters found by the text-based clustering, with the five cluster centers of the hybrid clustering. (b) The overlap of each of 5 clusters determined by hybrid clustering with Fisher's inverse chi-square method, with the text-based clusters

This leads immediately to the question of 'migrated' papers. More than a quarter of the papers were assigned to a different cluster according to the hybrid scheme.

Figure 9(b) visualizes the overlap between hybrid and text-based clusters. 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.

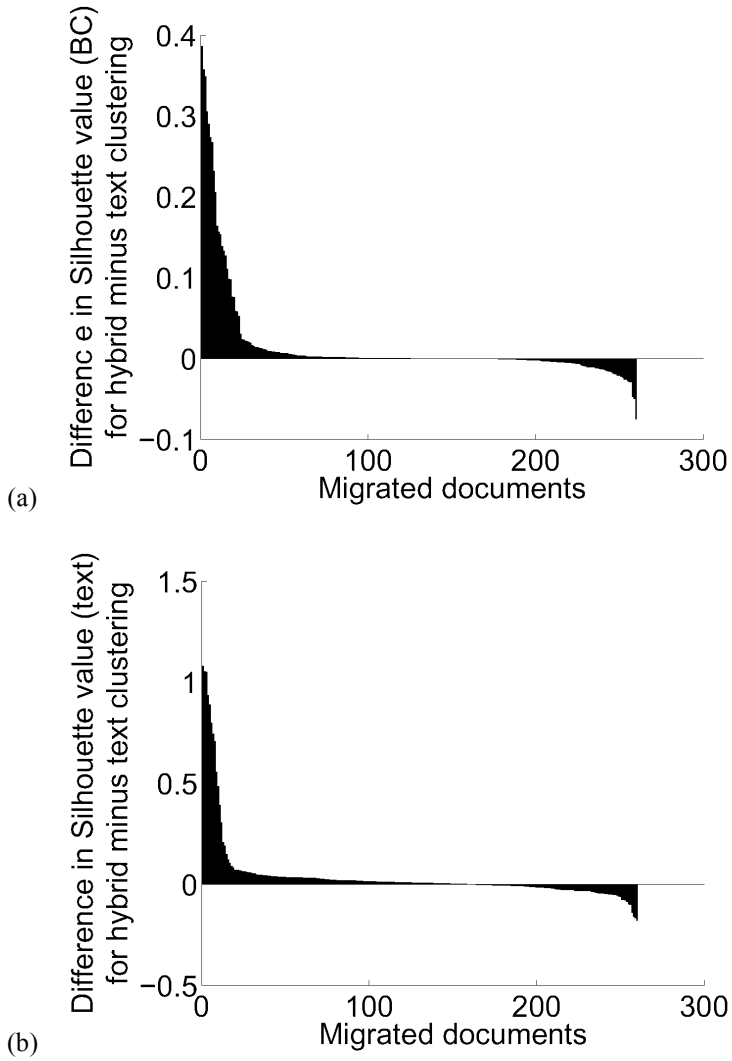

Figure 10. For all migrated documents the difference in Silhouette value for the hybrid clustering minus the text-only clustering, sorted in descending order. The prevalence of positive values indicates that there are more correct than spurious migrations. (a) Silhouette values are based on the complement of bibliographic coupling. (b) Silhouette values are based on text

One of the disadvantages is that wrong choices (merges) that are made by the algorithm in an early stage can never be repaired [KAUFMAN & ROUSSEEUW, 1990]. 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 10. The prevalence of positive values indicated that there are more correct than spurious migrations.

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: 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: FABA-PEREZ & 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: ALAVI & AL., 2002).

On the other hand, less evident migrations could also be observed. For example, "Empirical evidence of self-organization?" (Appendix: 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: LEYDESDORFF, 2002).

Figure 11 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.

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 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 above-mentioned modification to the original formula for bibliographic coupling by adding a constant 0.01 to the numerator 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 techniques helps to improve the reliability of relationship and therefore of the clustering algorithm as well.

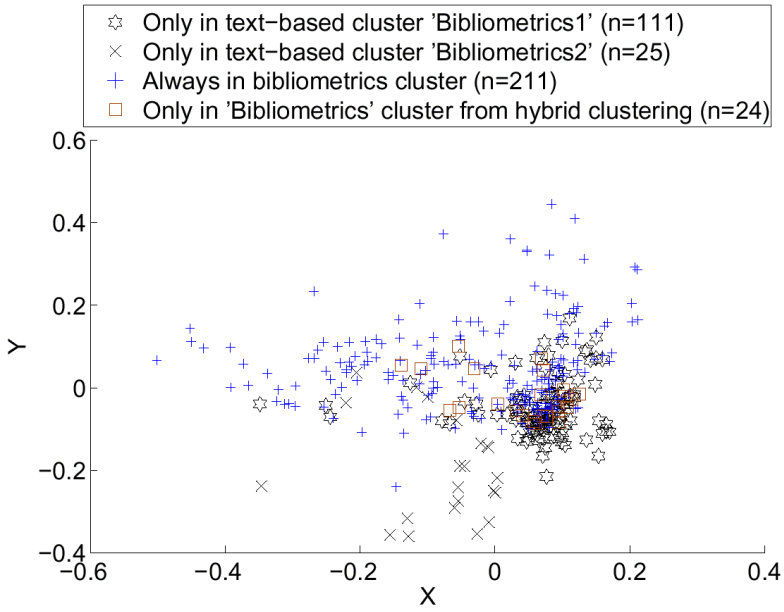

Figure 11. Multidimensional scaling (MDS) plot only considering documents in the two bibliometrics clusters of the text-based solution and documents in the bibliometrics cluster of the hybrid clustering. A distinction is made between documents that were only once assigned to a bibliometrics related cluster, and documents that were consistently assigned to bibliometrics

Conclusion

The field of information science was subdivided into 5 classes by using a hybrid clustering method based on Fisher's inverse chi-square, incorporating the full text of scientific publications and their cited references. Three large clusters could be distinguished: one containing research in information retrieval, another one about bibliometrics/scientometrics, and, finally, a collection of more 'socially' directed topics. The two smaller classes, patent analysis and webometrics, represented relatively new and emerging topics in IS.

In order for the inverse chi-square method to be applicable, we used a slightly modified formula for bibliographic coupling and also superimposed a noise component. A method was proposed to determine the integration weight λ for tuning the relative importance of two information sources.

The number of clusters was semi-automatically determined by applying a combination of distance-based and stability-based methods. However, the optimal number is still a difficult issue and depends on the adopted validation and chosen

similarity measures, as well as on data representation, be it mere text, just citations or a combination. For a text-only clustering, 6 clusters seemed to be optimal, whereas for an integrated clustering by linear combination even 8 could be perceived as the best choice.

This latter algorithm, although being an attractive, easy and reasonably scalable integration method that had previously not been shown to be inferior to Fisher's inverse chi-square method, was in the present setting outperformed with regard to the Silhouette coefficient and stability.

By integrating text and citations, Fisher's inverse chi-square method also did quantitatively and qualitatively 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 as well. However, this is probably unavoidable when using the hard agglomerative hierarchical clustering algorithm. We think that, in order to gain even better performance, a transition should be made towards fuzzy clustering algorithms.

*

An extended version of a paper presented at the *11th International Conference on Scientometrics and Informetrics*, Madrid (Spain), 25-27 June 2007 [JANSSENS & AL., 2007B].

This work was supported by Research Council KUL: GOA AMBioRICS, CoE EF/05/007 SymBioSys, several PhD/postdoc & fellow grants. Flemish Government: Steunpunt O&O Indicatoren; FWO: PhD/postdoc grants, projects G.0407.02 (support vector machines), G.0413.03 (inference in bioi), G.0388.03 (microarrays for clinical use), G.0229.03 (ontologies in bioi), G.0241.04 (Functional Genomics), G.0499.04 (Statistics), G.0232.05 (Cardiovascular), G.0318.05 (subfunctionalization), G.0553.06 (VitamineD), G.0302.07 (SVM/Kernel), research communities (ICCoS, ANMMM, MLDM); IWT: PhD Grants, GBOU-McKnow-E (Knowledge management algorithms), GBOU-SQUAD (quorum sensing), GBOU-ANA (biosensors), TAD-BioScope-IT, Silicos; SBO-BioFrame. Belgian Federal Science Policy Office: IUAP P6/25 (BioMaGNet, Bioinformatics and Modeling: from Genomes to Networks, 2007-2011). EU-RTD: ERNSI: European Research Network on System Identification; FP6-NoE Biopattern; FP6-IP e-Tumours, FP6-MC-EST Bioptrain.

References

BAEZA-YATES, R., RIBEIRO-NETO, B. (1999), *Modern Information Retrieval*. Cambridge: Addison-Wesley.

BRAAM, R. R., MOED, H. F., VAN RAAN, A. F. J. (1991), Mapping of science by combined cocitation and word analysis. 2. Dynamic aspects. *JASIS*, 42 : 252–266.

BATAGELJ, V., MRVAR, A. (2002), Pajek - Analysis and visualization of large networks. *Graph Drawing*, 2265 : 477–478.

BEN-HUR, A., ELISSEEFF, A., GUYON, I. (2002), A stability based method for discovering structure in clustered data. In: *Pacific Symposium on Biocomputing* (vol. 7, pp. 6–17), Retrieved September 9, 2007 from: http://helix-web.stanford.edu/psb02/benhur.pdf.

BERRY, M., DUMAIS, S. T., O'BRIEN, G. W. (1995), Using linear algebra for intelligent information retrieval. *SIAM Review*, 37 (4) : 573–595.

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.

COHN, D., HOFMANN, T. (2001), The missing link – a probabilistic model of document content and hypertext connectivity. *Neural Information Processing Systems*, 13.

DEERWESTER, S., DUMAIS, S. T., FURNAS, G. W., LANDAUER, T. K., HARSHMAN, R. (1990), Indexing by latent semantic analysis. *JASIS*, 41 (6) : 391–407.

DUNNING, T. (1993), Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19 (1) : 61–74.

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.

HEDGES, L. V., OLKIN, I. (1985), *Statistical Methods for Meta-analysis*. San Diego: Academic Press.

JAIN, A., DUBES, R. (1988), *Algorithms for Clustering Data*. New Jersey: Prentice Hall.

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. (2007A), *Clustering of Scientific Fields by Integrating Text Mining and Bibliometrics*. Ph.D. thesis, Faculty of Engineering, Katholieke Universiteit Leuven, Belgium, http://hdl.handle.net/1979/847.

JANSSENS, F., GLÄNZEL, W., DE MOOR, B. (2007B), A hybrid mapping of information science. In: D. TORRES-SALINAS, H. MOED (Eds) *Proc. of the 11th International Conference of the International Society for Scientometrics and Informetrics* (ISSI2007) (pp. 408–420), Madrid, Spain.

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),

KAUFMAN, L., ROUSSEEUW, P. J. (1990), *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: John Wiley and Sons Inc.

KESSLER, M. M. (1963), Bibliographic coupling between scientific papers. *American Documentation*, 14 : 10–25.

MANNING, C. D., SCHÜTZE, H. (2000), *Foundations of Statistical Natural Language Processing*. Cambridge: MIT Press.

MODHA, D. S., SPANGLER, W. S. (2000), Clustering hypertext with applications to web searching. *ACM Conference on Hypertext* (pp. 143–152).

MORRIS, S. A., YEN, G., WU, Z., ASNAKE, B. (2003), Time line visualization of research fronts. *Journal of the American Society for Information Science and Technology*, 54 (5) : 413–422.

MORRIS, S. A., YEN, G. G. (2004), 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.

MULLINS, N., SNIZEK, W., OEHLER, K. (1988), The structural analysis of a scientific paper. In: A. F. J. VAN RAAN (Ed.), *Handbook of Quantitative Studies of Science and Technology* (pp. 81–105), New York: Elsevier Science.

PORTER, M. F. (1980), An algorithm for suffix stripping. *Program*, 14 (3) : 130–137.

ROUSSEEUW, P. J. (1987), Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20 : 53–65.

SALTON, G., MCGILL, M. J. (1986). *Introduction to Modern Information Retrieval*. New York: McGraw-Hill, Inc.

SNIZEK, W., OEHLER, K., MULLINS, N. (1991). Textual and nontextual characteristics of scientific papers: Neglected science indicators. *Scientometrics*, 20 (1) : 25–35.

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).

WARD, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58 : 236–244.

Appendix

Bibliographic sources of papers referred to in the text as subject of analysis:

ALAVI & AL. (2002). *JASIST*, 53 (12) : 1029–1037.
DING & AL. (2002A). *Journal of Information Science*, 28 (2) : 123–136.
DING (2002B). *Journal of Information Science*, 28 (5) : 375–388.
FABA-PEREZ & AL. (2003). *Journal of Documentation*, 59 (5) : 558–580.
LEYDESDORFF (2002). *Scientometrics*, 53 (1) :131–159.
NIE (2003). *JASIST*, 54 (4) : 335–346.
VAN DEN BESSELAAR (2003). *JASIST*, 54 (1) : 87–90.