Co-clustering Approaches to Integrate Lexical and Bibliographical Information

Frizo Janssens[*], Patrick Glenisson[*,**], Wolfgang Glänzel[**,***] and Bart De Moor[*]

frizo.janssens@esat.kuleuven.be
[*]Katholieke Universiteit Leuven, ESAT-SCD, Kasteelpark Arenberg 10,B-3001 Leuven (Belgium)

patrick.glenisson@econ.kuleuven.be
**Katholieke Universiteit Leuven, Steunpunt O&O Statistieken, Dekenstraat 2, B-3000 Leuven (Belgium)

wolfgang.glanzel@econ.kuleuven.be
[***]Hungarian Academy of Sciences, Institute for Research Organisation, Nádor u. 18,
H-1051 Budapest (Hungary)

Abstract
Terms are the building blocks to organize and access information, and hold a key position in information retrieval. In forthcoming work we have shown how a methodology of indexing full-text scientific articles combined with an exploratory statistical analysis can improve on bibliometric approaches to mapping science. Textual documents are indexed and further characterized using data mining techniques and co-word analysis. We start this paper by briefly demonstrating the text mining approach. Whereas statistical processing based on full-text documents provides a relational view based on the topicality represented by these documents, bibliometric components can include other characteristics that describe their position in the set. Therefore we extend on previous work and explore how *hybrid methodologies* that deeply combine text analysis and bibliometric methods can improve the mapping of science and technology. In particular, we propose a method to mathematically combine document similarity matrices resulting from vector-based indices on the one hand, and from selected bibliometric indicators on the other hand. Weighted linear combinations as well as approaches inspired on statistical meta-analysis are presented. Both pitfalls and possible solutions are discussed. The resulting combined similarity matrix offers an attractive way to 'co-cluster' documents based on both lexical and bibliographic information.

Introduction

Bibliometric methods proved valuable tools to monitor and chart scientific processes. When considering publications as atomic entities in scientometric studies, one can readily describe and analyze the relationship between elements of a given set of scientific publications using bibliometric tools. However, lexical information may also convey important clues for such mapping purposes. Therefore, using both sources of information in a supplementary way provides interesting perspectives. The idea of combining bibliometric methods with the analysis of indexing terms, subject headings or keywords extracted from titles and/or abstracts, is not new (Callon et al., 1991; Noyons and van Raan, 1994; Zitt and Bassecoulard, 1994; Kostoff et al, 2001). In a forthcoming publication (Glenisson et al., 2005), we examine how full-text analysis by casting terms of a scientific publication in a vector space can complement more traditional bibliometric indicators for the purpose of science mapping. In this paper we present and extend the main conclusions of the previous work and propose an integrated approach to jointly mine lexical and bibliometric information. Our goal is to improve on both existing lexical and bibliometric approaches to science (or technology) mapping through a hybrid methodology that enables a 'best of both worlds' approach.

Methodology

Preprocessing

As in most data mining endeavors, data acquisition, preprocessing, and cleansing jointly represent up to 80 percent of the overall effort distribution. Preprocessing steps include the removal of stopwords and author names, stemming and the detection of phrases. Stemming involves the removal of word suffixes such as plurals, verb tenses and deflections, and the replacement by their canonized equivalent (Porter, 1980). Bigrams, or phrases composed of two words, were detected using the Dunning likelihood ratio test (Dunning, 1993). After elimination of words that only occur in a single document,

Proceedings of ISSI 2005, pages 284–289.
Edited by P. Ingwersen and B. Larsen.

we finally combined the withheld words and phrases into a final thesaurus by means of which all documents were indexed.

Text representation
We adopted the common vector space model to encode a document in a *k*-dimensional term space where each component w_{ij} represents the weight of term t_j in document d_i. The grammatical structure of the text is hereby neglected and therefore it is also referred to as a 'bag-of-words' representation. The set of all terms t_j is called the vocabulary or thesaurus.
The TF-IDF term weighting scheme is defined as follows:

$$w_{ij} = f_{ij} \log \frac{N}{n_j}$$

where f_{ij} is the number of occurrences of t_j in d_i and is referred to as term frequency (TF). N represents the total number of documents and n_j is the number of documents in the collection that contain term t_j. The logarithm is called inverse document frequency (IDF).

To find the major associative patterns of word usage in the document collection and to get rid of the 'noise variability' in it, we used a transformation based on Latent Semantic Indexing (Deerwester, 1990).

We express similarity between pairs of documents d_{i1} and d_{i2} as the cosine of the angle between the corresponding vector representations:

$$sim(d_{i_1}, d_{i_2}) = \frac{d_{i_1} \cdot d_{i_2}}{\left\| d_{i_1} \right\| \cdot \left\| d_{i_2} \right\|},$$

hence obtaining a similarity matrix S, which is passed as a distance matrix $D = 1 - S$ to Ward's hierarchical clustering procedure.

Scientometrics in 2003 through the eyes of text mining
The complete publication year 2003 of the journal *Scientometrics* has been selected and weeded down to 85 full-text research papers. The results of the text cluster analysis are extensively discussed in (Glenisson, 2005). For illustrative purpose we show the content structure of cluster 2, which is dominated by empirical papers and case studies. The terms in this cluster are presented in Figure 1 and relate above all to national and institutional aspects as well as to science fields. We labeled this cluster as pertaining to case studies and traditional bibliometric applications.

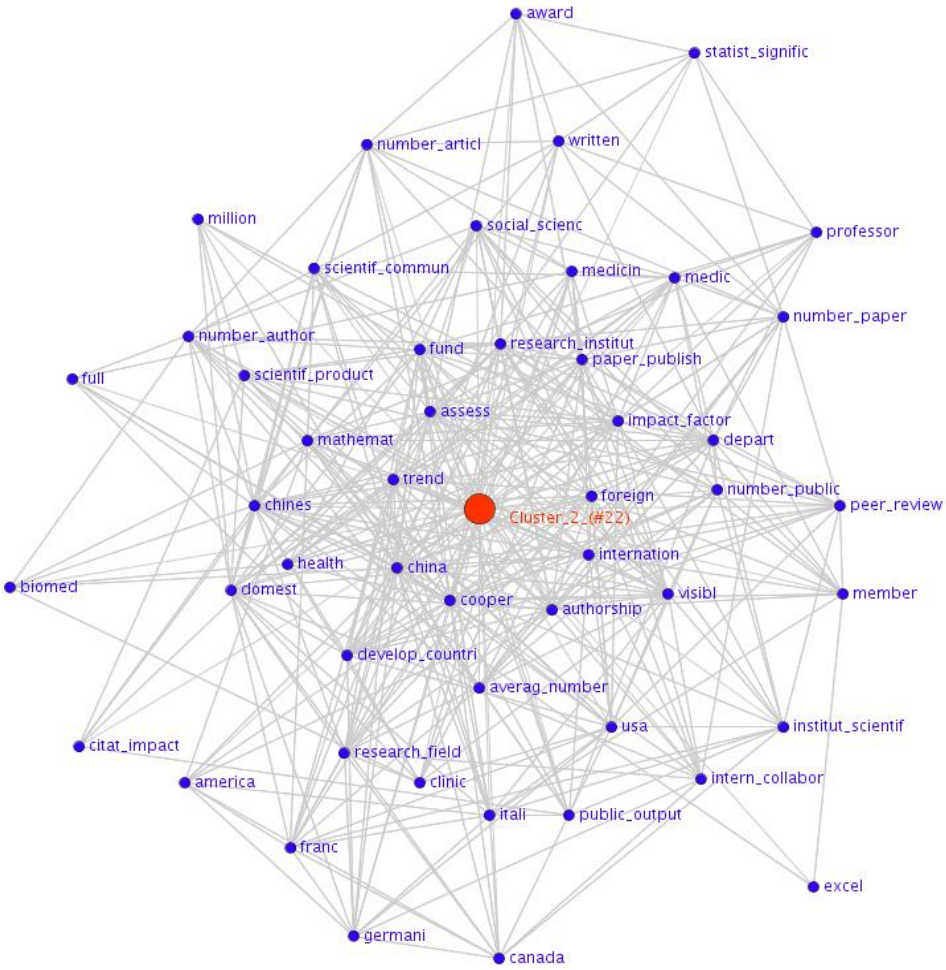

Figure 1: Illustration of a term network of cluster 2 through subsequent co-word analysis.

Combining text mining and bibliometric information

Serialization of text clustering and bibliometrics

We first combine the bibliometric approach with results of the full-text analysis by means of aggregating both results: Figure 2 shows the relation between Mean Reference Age and Share of Serials with the cluster results 'overlaid'. We have indicated the two special issues (Triple Helix Conference and S&T Indicators Conference) by ellipses. These issues form surprisingly homogeneous groups. Our example, cluster 2 (indicated by its medoid 'Changing trends in publishing behaviour'), is characterized by medium Mean Reference Age (MRA).

Papers with similar 'content' might thus have different bibliometric characteristics depending on the target readership and the field of application. Therefore we deem it an interesting option to integrate these two disparate information sources *earlier* in the segmentation process. We develop details to such an approach in what follows.

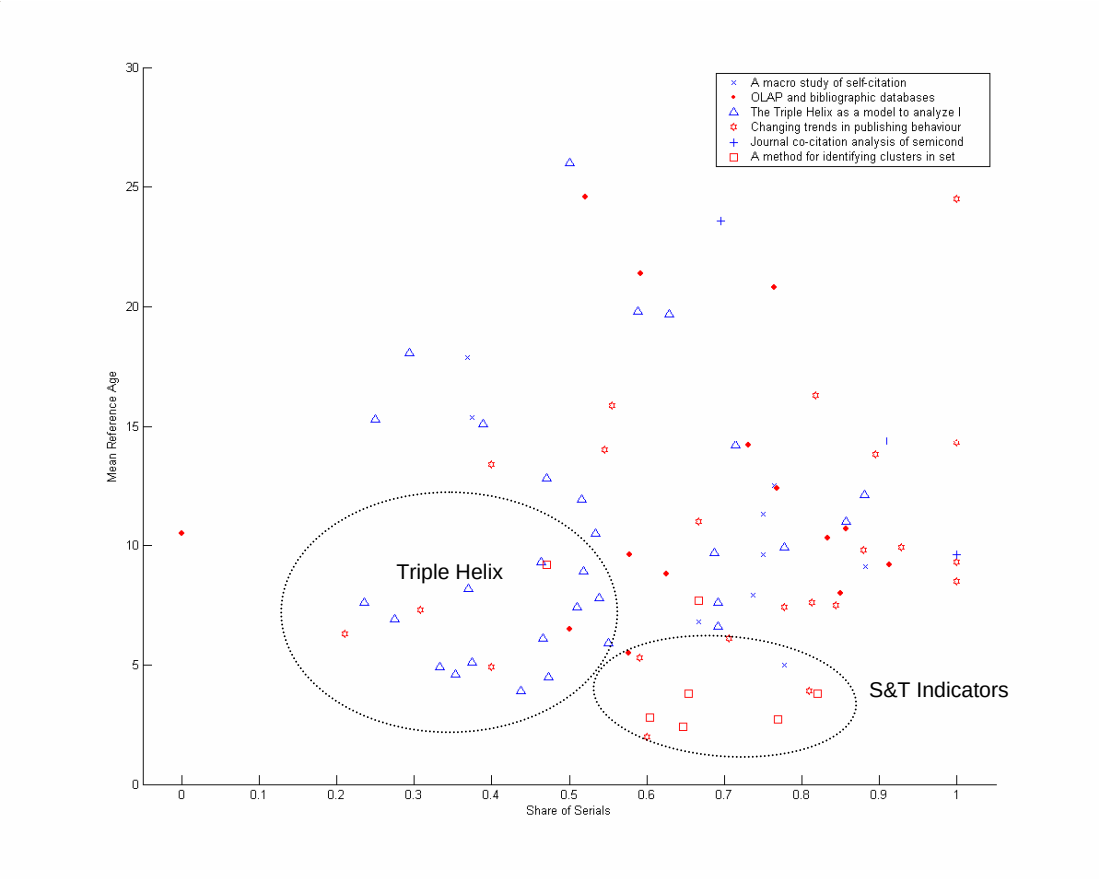

Figure 2: Plot of Mean Reference Age vs. Share of Serials on basis of the co-word clusters. Clusters are represented in the legend by their medoid documents.

Co-clustering of lexical and bibliometric information

We can use the two bibliometric features in Figure 2, Mean Reference Age and Share of Serials, to compute pairwise document distances with the classical Euclidean distance measure. The resulting distance matrix D^{BIBL} can be linearly combined with the previously obtained text-based distance matrix D^{TEXT} using the following scheme:

$$D^{INTEGR} = \lambda D^{TEXT} + (1-\lambda) D^{BIBL},$$

with λ controlling the relative importance of both data sources. The resulting D^{INTEGR} can then be passed to any distance-based clustering algorithm such as hierarchical clustering.

Despite being attractively simple, several problems need to be solved:

- heterogeneous data matrices (e.g., document-term or document-indicator) require a different choice of distance metric;
- spurious and strong (dis)similarities can obliterate good relationships established by the other data source;
- even if distances from both sources span the same range (e.g., [0 1]), they exhibit different distributional characteristics;
- manually balancing confidence between data sources through the parameter λ can be rather arbitrary.

At the top of Figure 3 we show the different distributional characteristics of text-based and bibliometric distances. This implies that for a setting $\lambda = 0.5$ there is no equal contribution of both sources in the mixed representation. Rather, such setting will implicitly favor text over bibliometric information or vice versa depending on the distributional characteristics. Although not necessarily detrimental, it creates additional problems on the transparency of λ.

Therefore, to accommodate for this problem and better equalize the distributions of both sources, we first propose to transform all entries in the distance matrices to *p*-values by computing one-sided cumulative distribution function (cdf) values for each distance value in both representations. The results for D^{TEXT} and D^{BIBL} are displayed at the bottom of Figure 3. A first question that arises is to which extent this transformation conserves the structure of the original distance matrices. Linear matrix correlations are inappropriate in this case as a non-linear transformation was applied. Therefore, we measure the overlap (measured by means of the *Rand* index; (Jain ,1988)) between the results of the same clustering algorithm applied before and after the *p*-value transformation. Preliminary experiments report acceptable *Rand* 'correlations' of 0.85 and 0.95 for D^{TEXT} and D^{BIBL} respectively.

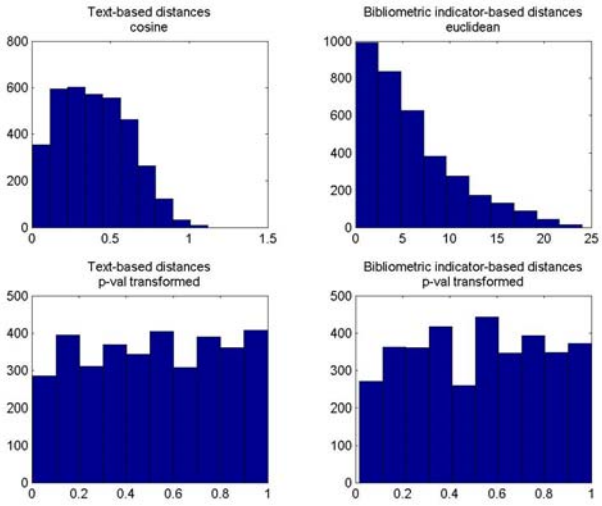

Figure 3: Different distributional characteristics of distances in term vector space (left) and a selected bibliometric space (right). Note that the cosine based distance, $1-\cos(.,.)$, can be larger than one due to negative values in the LSI transformed matrix.

Now that *p*-values are available, we can apply Fisher's omnibus statistic to combine *p*-values from multiple sources:

$$F = -2\log(p_{TEXT}^{\lambda} p_{BIBL}^{1-\lambda}).$$

The resulting cdf-values from the empirical distribution are used as new *p*-values in the combined distance matrix. In contrast to the weighted linear combination procedure, this method can handle distances stemming from different metrics, and provides a smoother way to combine data by avoiding overcompensation by either data source. We are currently validating the results of clustering such integrated similarities in a setting $\lambda = 0.5$. However, it is clear that weighting the data sources according to their 'quality' is an important issue. For example, term-based approaches, bibliographic coupling, co-citation information, or bibliometric indicators each have particular strengths and weaknesses on particular types of data. Although λ can be set manually by researchers with a lot of expertise on the data set at hand, we are investigating how fast clustering procedures can help in *automatically* providing an estimate of 'meaningful' structure present in either data source.

By computing a *Silhouette Value per Clustering (SVC)* (Jain, 1988) for each data type we can estimate the relative quality of each data source and use this as an educated guess for λ:

$$\lambda = \frac{SVC^{TEXT}}{SVC^{BIBL} + SVC^{TEXT}}.$$

For the two data types discussed in this paper (textual and bibliometric) such an exercise resulted in $\lambda = 0.47$. When contrasting text and co-citation analysis on this restricted data set we got $\lambda = 0.61$.

Although these values are *indicative,* they match our intuition about this particular data set and the information types used.

Conclusion

The complex nature of mapping various aspects of knowledge motivates approaches that integrate different viewpoints on the same samples. We proposed a flexible framework that allows co-clustering of lexical and bibliographic information (or any other data source that can be cast in vector space). Although not yet finished, we believe the approach can offer interesting perspectives in several applications.

Acknowledgement

The authors acknowledge support from the Flemish Government (Steunpunt O&O Statistieken), Research Council K.U. Leuven (GOA-Mefisto-666, GOA-Ambiorics, IDO), the Fonds voor Wetenschappelijk Onderzoek - Vlaanderen (G.0115.01, G.0240.99, G.0407.02, G.0413.03, G.0388.03, G.0229.03, G.0241.04), the Instituut voor de aanmoediging van Innovatie door Wetenschap en Technologie Vlaanderen (STWW-Genprom, GBOU-McKnow, GBOU-SQUAD, GBOU-ANA), the Belgian Federal Science Policy Office (IUAP V-22), and the European Union (FP5 CAGE, ERNSI, FP6 NoE Biopattern, NoE Etumours).

References

Callon, M., Courtial, J.P., and Laville, F. (1991). Co-word Analysis as a Tool for Describing the Network of Interactions between Basic and Technological Research: The Case of Polymer Chemistry. *Scientometrics*, **22 (**1), 155-205.

Deerwester S., Dumais S.T., Furnas G.W., Landauer T.K., and Harshman R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41, 391-407.

Dunning T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19, 61-74.

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

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

Kostoff R.N., Toothman D.R., Eberhart H.J., Humenik J.A. (2001). Text mining using database tomography and bibliometrics: A review. *Technological Forecasting and Social Change*, **68** (3), 223-253.

Noyons, E.C.M., Van Raan A.F.J. (1994). Bibliometric cartography of scientific and technological developments of an R&D field. The case of Optomechatronics. *Scientometrics*, 30, 157-173.

Porter M.F. (1980) An algorithm for suffix stripping, *Program*, 14(3), 130-137.

Zitt M., Bassecoulard E. (1994). Development of a method for detection and trend analysis of research fronts built by lexical or co-citation analysis. *Scientometrics* **30** (1), 333-351.