[a]Frizo Janssens, [a]Viet Tran Quoc, [b]Wolfgang Glänzel, [a]Bart De Moor
a Katholieke Universiteit Leuven, ESAT-SCD, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium
b Katholieke Universiteit Leuven, Steunpunt O&O Statistieken, Dekenstraat 2, B-3000 Leuven, Belgium

Integration of textual content and link information for accurate clustering of science fields

Pages: 615-619

Current Research in Information Sciences and Technologies

Multidisciplinary Approaches to Global Information Systems

Vicente P. Guerrero-Bote (Editor)

Proceedings of the I International Conference on Multidisciplinary
Information Sciences and Technologies, InSciT2006
Mérida - SPAIN
October, 25th-28th, 2006

ISBN-10 CD-ROM: 84-611-3106-1
ISBN-13 CD-ROM: 978-84-611-3106-8

ISBN-10 Whole Edition: 84-611-3103-7
ISBN-13 Whole Edition: 978-84-611-3103-7

ISBN-10 Volume I: 84-611-3104-5
ISBN-13 Volume I: 978-84-611-3104-4

Open Institute of Knowledge
(Instituto Abierto del Conocimiento)
Antonio Álvarez 6
06005 Badajoz, SPAIN
Phone: +34 924221935
Fax: +34 924221935
http://www.instac.es

Integration of textual content and link information for accurate clustering of science fields

[a]Frizo Janssens, [a]Viet Tran Quoc, [b]Wolfgang Glänzel, [a]Bart De Moor
[a]Katholieke Universiteit Leuven, ESAT-SCD, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium
[b]Katholieke Universiteit Leuven, Steunpunt O&O Statistieken, Dekenstraat 2, B-3000 Leuven, Belgium

Hybrid clustering methods that exploit both text and citations might achieve better results than pure text-based or link-based methods. We experiment with Fisher's inverse chi-square method to derive a hybrid solution to map the bioinformatics field. A set of documents published in a list of core bioinformatics journals is extracted from the Web of Science (WoS) and extended with bibliometrically related records. Pairwise distances between documents are converted to p-values with respect to randomized datasets and the inverse chi-square method is used to combine the p-values from both information sources. This method can handle distances stemming from metrics with different distributional characteristics and avoids domination of any data source. For a correct application of Fisher's method we also introduce a slight, rank-preserving modification to the formula for bibliographic coupling. We evaluate clustering results by the mean Silhouette coefficient and also assess the performance in a classification setting for which we construct a 'ground truth' based on the Medical Subject Headings (MeSH), annotated by experts. To retrieve the MeSH terms, each WoS document is matched against Medline. Latent semantic analysis and a hierarchical clustering technique are applied to the MeSH-by-document matrix to determine document clusters, which are then post-processed by an iterative shrinking technique to only retain well-defined categories. We compare clustering and classification performances of sheer text and citation-based methods, of Fisher's method and of other data integration schemes. In general, text is more powerful than cited references, and dimensionality reduction by SVD further improves results; however, the best outcome is obtained by integration.

Keywords: hybrid clustering, Fisher’s inverse chi-square method, classification, text mining, bibliometrics, bioinformatics.

1 Introduction

The purpose of this paper is to assess clustering and classification performances of methods that use just text or only cited references, of Fisher's inverse chi-square method [5] and of other schemes for integrating textual contents with the structure of the citation graph. 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. We will only consider cited references (also called 'out-links' in webometrics), but straightforward extensions are available to also incorporate received citations.

The idea of combining bibliometric or link information with textual content is not new for it has already been pursued to obtain improved performance in information retrieval [2], bibliometric mapping of science [a.o., 1,4,6], clustering [a.o., 9,10], and classification [e.g., 7,3].

We are interested in unsupervised clustering rather than building an optimal classifier assigning documents to predefined categories, since an accurate classification of scientific articles is not available and would otherwise indisputably be outdated because of the dynamic nature of contemporary science and technology. However, we will also evaluate the various experimented data types in a classification setting as it offers a well-grounded basis for assessing relative performance.

2 Material

Our dataset consists of 5188 bioinformatics-related papers which are available in the ISI Web of Science database (WoS), publication years 1990-2004. For the delineation of the bioinformatics field, we extracted a set of all articles, notes and reviews published in a list of core bioinformatics journals or satisfying a keyword search strategy, namely documents containing ‘*bioinformatics*’, or ‘*computational biolog**’ or ‘*systems biology*’ in the titles. The resulting set was then extended with bibliographically coupled records and with papers, from a set of 50 extra journals, that at least three times cite or are cited by core documents. We used the following core journals from WoS: *Bioinformatics* (formerly *Computer Applications in the Biosciences*), *Journal of Computational Biology*, *Briefings in Bioinformatics* and *BMC Bioinformatics*. Each included paper was matched against the Medline database. For each matched record, associated Medical Subject Headings (MeSH), which are annotated by human experts, were retrieved.

3 Methods

We assessed the performance of one clustering and one classification algorithm using 13 different data types, 2 of which consisted of only textual information, 3 only of cited references, while 8 were integrated data types. Table 1 lists the 13 experimented data types.

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

Except for data types #4, #11, #12 and #13, which will be discussed in section 3.1, similarities between documents were computed by the cosine measure between the normalized textual, citation-based or integrated vectors. For the vectors with cited references this corresponds to bibliographic coupling (#3). Dimensionality reduction of the references-by-documents matrix from 38660 references to 50 factors was also performed by using a truncated SVD (#5). For data type #6, both the text vector and out-link vector of a document were concatenated, and data type #7 is derived from the application of SVD on the concatenated matrix. Integrated data types #8, #9 and #10 result from a weighted linear combination of document similarities, possibly with SVD applied on either component.

Table 1. The 13 experimented data types, indicating (with 'X') whether they contain a text component, citation-based component, or both, and whether SVD was used for dimensionality reduction.

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

3.1 Fisher's inverse chi-square method

The last three types in Table 1 are new and integrate both textual and bibliographic coupling (BC) information by making use of the inverse chi-square method, which is an omnibus statistic from statistical meta-analysis to combine p-values from multiple sources [5]. This method can handle distances stemming from different metrics with different distributional characteristics, and avoids domination of any information source [6]. All text-based and link-based document distances are first transformed to p-values with respect to the cumulative distribution function of distances for randomized data. In our setting, a p-value means the probability that the similarity between two documents could be at least as high just by chance. For a correct application of Fisher's omnibus method the input test statistics should be continuous. However, this is not the case for the sparse BC since most scientific articles do not have any references 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 *A* and *B* is then $(Nab + 0.01)/\sqrt{Na \cdot Nb}$, with *Na* and *Nb* the number of references in paper *A* and paper *B*, respectively, and *Nab* 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 addition of random noise does not deteriorate results since the error to be expected from for instance missing references in the WoS database is much higher. Moreover, as discussed later in section 4, this new formula for dense BC, including the noise factor, can lead to comparable clustering performances and even to classification accuracies significantly higher than those of the original formula. 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. 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. If SVD had been applied on the text or citation-based component, the random document vectors should be projected in the same space of reduced dimensionality. The weight λ can be used to tune the relative importance of both information sources; however, choosing a good value for λ is not straightforward. Especially if SVD is used, a parameter sweep should be performed.

3.2 Clustering and classification

We adopted a hierarchical clustering algorithm using Ward's method and we assessed the quality of clustering results by calculating the adjusted Rand index and the mean Silhouette coefficient [8].

We also evaluated all data types in Table 1 in a classification setting, as it offers a well-grounded basis for comparing relative performances. Since no expert-made classification of the bioinformatics papers is available, we constructed a 'ground truth' classification based on an optimal clustering of documents, indexed only by their MeSH terms which were never used in further experiments nor in data types. The resulting document clusters, also considered as classes, were post-processed by an iterative shrinking technique to retain only well-defined categories. One noise cluster was also detected and removed. Figure 1 shows the quality of the classes (Silhouette plot) before and after iterative shrinking. The ultimate set of 7 clusters was used as the gold standard classification of documents. Besides, to also assess performances at another level of granularity, a coarser-grained classification was used that contained only two iteratively shrunk classes.

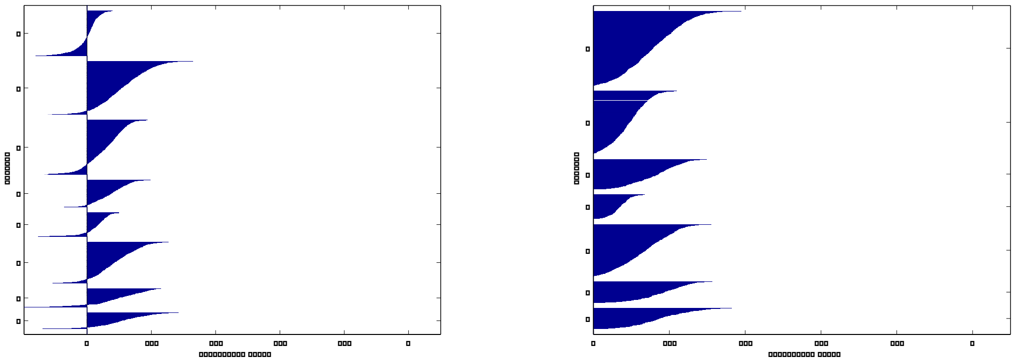

Fig. 1. Silhouette plots before (left) and after (right) iterative shrinking of the document clusters based on MeSH.

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

4 Discussion of results

4.1 Clustering

Figure 2 shows clustering performances assessed by overall mean Silhouette coefficient for all data types in Table 1, when clustering 20 random test sets into 2 coarse-grained clusters (left) and into 7 finer-grained clusters (right). Silhouette values are calculated on data type independent MeSH-by-document matrices. Figures for adjusted Rand index are not included here since relative results were comparable.

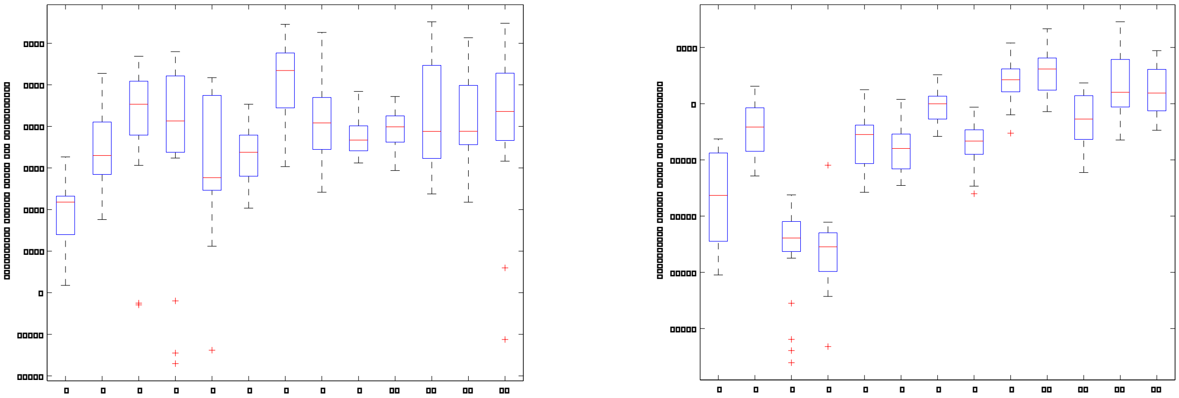

Fig. 2. Test set clustering performance measured by mean Silhouette value for 2 (left) and 7 (right) clusters.

By observing Figure 2, we see that: (i) On the coarse level, concatenated matrices subsequently reduced by SVD (#7) provide significantly better results than most other types, although the Wilcoxon signed rank test does not reject equal means when comparing with Fisher's inverse chi-square method (#11 & #13). However, for the more detailed view #7 is outperformed by Fisher's method and linear combinations with LSI/SVD (#9, #10, #12 & #13), which are the only data types obtaining mainly positive Silhouettes. There is no significant difference between Fisher's method and corresponding linear combinations.

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

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

4.2 Classification

Figure 3 depicts the classification performance measured by accuracies on 20 random independent test sets, when classifying into 2 categories (left) and into 7 categories (right). Test set AUCs are not shown here because the conclusions were comparable and AUCs are only easily computable for binary classifications.

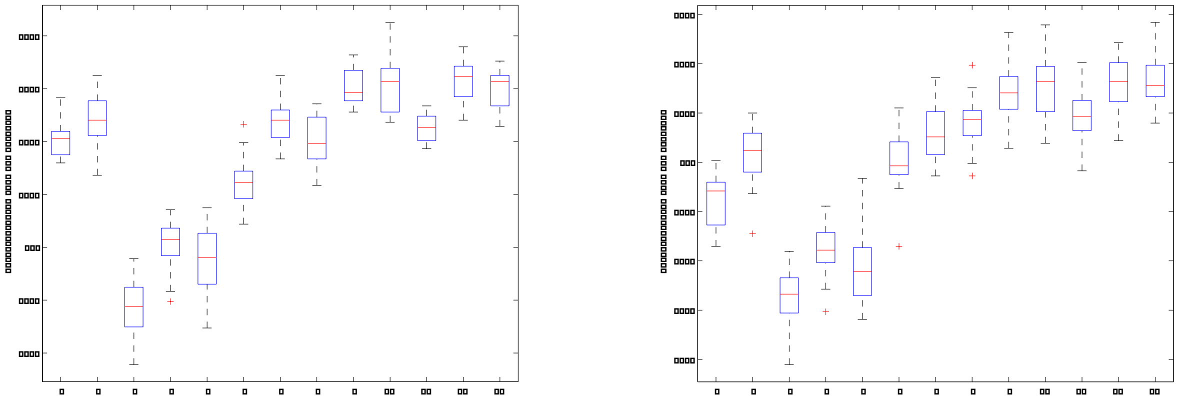

Fig. 3. Classification performance measured by test set accuracy for 2 (left) and 7 (right) classes.

Observing Figure 3, we see that: (i) Best performances are obtained by linear combinations and Fisher's method with SVD applied on the textual component (#9, #10, #12 & #13), but there are no significant differences among these 4 data types.

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

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

5 Conclusion

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

The introduced integration method based on Fisher's inverse chi-square has shown to significantly outperform corresponding text-only and link-only methods, as well as a mere concatenation of vectors. Only for the coarse-grained clustering (2 clusters) the SVD of concatenated matrices did at least equally well. Fisher's omnibus method, however, did not significantly outperform corresponding linear combinations when SVD had been applied. Given the higher complexity for implementing Fisher's method and a reduced scalability, a carefully chosen weighted linear combination might be the preferred solution for integrating textual and citation information if LSI is used. However, the inverse chi-square method is generic and can be used to incorporate distances with highly dissimilar distributional characteristics, like e.g. textual distances and distances based on the bibliometric features Mean Reference Age and Share of Serials [6].

Acknowledgment

Research Council KUL: GOA AMBioRICS, CoE EF/05/007 SymBioSys, several PhD/postdoc & fellow grants; Flemish Government: FWO: PhD/postdoc grants, projects G.0407.02 (support vector machines), G.0413.03 (inference in bioi), G.0229.03 (ontologies in bioi), G.0499.04 (Statistics), research communities (ICCoS, ANMMM, MLDM); IWT: GBOU-McKnow (Knowledge management algorithms), GBOU-SQUAD (quorum sensing), TAD-BioScope, Silicos; Belgian Federal Science Policy Office: IUAP P5/22 (`Dynamical Systems and Control: Computation, Identification and Modelling, 2002-2006); EU-RTD: FP6-NoE Biopattern; FP6-IP e-Tumours, FP6-MC-EST Bioptrain.

References

[1] Braam, R., Moed, H., & Van Raan, A. 1991. Mapping of science by combined co-Citation and word Analysis. 2. Dynamical aspects. Journal of the American Society for Information Science, 42 (4), pp. 252-266.

[2] Calado, P., Ribeiro-Neto, B., Ziviani, N., Moura, E., & Silva, I. 2003. Local versus global link information in the Web. In ACM Transactions on Information Systems (TOIS), 21 (1), pp. 42-63.

[3] Calado, P., Cristo, M., Moura, E.S., Gonçalves, M.A., Ziviani, N., & Ribeiro-Neto, B. 2006. Linkage similarity measures for the classification of web documents. Journal of the American Society for Information Science and Technology (JASIST), 57 (2), pp.208-221.

[4] Glenisson, P., Glänzel, W., Janssens, F., & De Moor, B. 2005. Combining full-text and bibliometric information in mapping scientific disciplines. Information Processing & Management, 41 (6), pp.1548–1572.

[5] Hedges, L.V., & Olkin, I. 1985. Statistical methods for meta-analysis. San Diego: Academic Press.

[6] Janssens, F., Glenisson, P., Glänzel, W., & De Moor, B. 2005. Co-clustering approaches to integrate lexical and bibliographical information. In Ingwersen, P., & Larsen B. (Eds.), Proc. 10th Int'l Conference of the ISSI 1, pp.284-289.

[7] Joachims, T., Cristianini, N., & Shawe-Taylor, J. 2001. Composite kernels for hypertext categorisation. Proc 18th Int'l Conference on Machine Learning (ICML), pp. 250-257.

[8] Kaufman, L., & Rousseeuw, P.J. 1990. Finding groups in data: An introduction to cluster analysis. New York: John Wiley and Sons Inc.

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

[10] Wang, Y., Kitsuregawa, M. 2002. Evaluating contents-link coupled web page clustering for web search results. Proc. 11th int'l conference on Information and knowledge management (CIKM), pp. 499-506.