Journal of Informetrics 3 (2009) 49–63

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Document–document similarity approaches and science mapping: Experimental comparison of five approaches

Per Ahlgren[a,*], Cristian Colliander[b]

[a] Department of e-Resources, University Library, Stockholm University, SE-106 91 Stockholm, Sweden
[b] University Library, Jönköping University, SE-551 11 Jönköping, Sweden

ARTICLE INFO

Article history:
Received 8 October 2008
Received in revised form
16 November 2008
Accepted 17 November 2008

Keywords:
Citation data
Textual data
Data source combination
Cluster analysis
Science mapping

ABSTRACT

This paper treats document–document similarity approaches in the context of science mapping. Five approaches, involving nine methods, are compared experimentally. We compare text-based approaches, the citation-based bibliographic coupling approach, and approaches that combine text-based approaches and bibliographic coupling. Forty-three articles, published in the journal *Information Retrieval*, are used as test documents. We investigate how well the approaches agree with a ground truth subject classification of the test documents, when the complete linkage method is used, and under two types of similarities, first-order and second-order. The results show that it is possible to achieve a very good approximation of the classification by means of automatic grouping of articles. One text-only method and one combination method, under second-order similarities in both cases, give rise to cluster solutions that to a large extent agree with the classification.

1. Introduction

In science mapping based on a set of documents, document–document similarity can be measured in different ways. One approach is to measure the similarity between the texts associated with the documents. Several similarity measures, based on terms that occur in both texts, have been suggested in the literature (Boyce, Meadow, & Kraft, 1994; Salton, & McGill, 1983). Recent examples of works that make use of this text-based approach in science mapping are Glenisson, Glänzel, Janssens, and Moor (2005) and Janssens, Leta, Glänzel, and Moor (2006), where bibliometrics and library and information science were mapped, respectively.

A second approach is bibliographic coupling, introduced to the scientific society through a number of articles in the 1960s (Kessler, 1963a, b, 1965). The point of departure for the bibliographic coupling approach is to measure the similarity of the reference lists of two documents. A single reference shared by two documents was considered to be a unit of coupling between them, and the *coupling strength* between two documents was defined as the number of coupling units (Kessler, 1963a). Clearly, a unit of coupling between two documents is cited by both of them, and bibliographic coupling is a citation-based approach to document–document similarity. Here, we say that two documents are *bibliographically coupled* if their coupling strength is greater than 0. The coupling strength is taken as an indicator of document–document similarity.

A text-based approach can be combined, or integrated, with bibliographic coupling (Janssens, Quoc, Glänzel, & Moor, 2006; Janssens, 2007). Thus, approaches of this kind use two information sources, textual content and cited references. The integration may be achieved by different methods, for instance by statistical combination of two dissimilarity values associated with the same pair of documents.

[*] Corresponding author.
E-mail address: per.ahlgren@sub.su.se (P. Ahlgren).

1751-1577/$ – see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.joi.2008.11.003

When the similarity values have been obtained, regardless of the underlying approach, the documents may be automatically grouped into pairwise disjunct sets by the application of clustering techniques.

In this work, where the research reported in an earlier article (Ahlgren & Jarneving, 2008) is developed, we compare experimentally five approaches to document–document similarity in the context of science mapping. We compare (1) text-based approaches, (2) the citation-based bibliographic coupling approach, and (3) approaches that combine text-based approaches and bibliographic coupling. We investigate how well the approaches agree with a ground truth subject classification of our test documents, when the complete linkage method is used, and under two types of similarities, first-order and second-order.

The remainder of this paper is organized as follows. Section 2 reviews related research. Data and methods are described in Section 3, whereas Section 4 gives the results. The concluding section, Section 5, includes a discussion, as well as conclusions.

2. Related research

In Small and Koenig (1977), the combination of bibliographic coupling (of journals) and clustering (single linkage method) was used in order to automatically group scientific journals. The cluster solution obtained was compared to a manually constructed classification of the journals, and a generally good agreement is reported. Vladutz and Cook (1984) studied the validity of bibliographic coupling as an indicator of semantic similarity between scientific papers, utilizing the Science Citation Index database. According to expert assessments, a large proportion of the papers with the highest coupling strength, with respect to a set of input papers, were closely related by subject to these input papers. Peters, Braam, and van Raan (1995) investigated cognitive resemblance, operationalized as word-profile similarity, and citation relations in chemical engineering publications. It was found that publications bibliographically coupled with a highly cited publication were related by content: the average word-profile similarity (to the word-profile for the totality of the coupled publications) within such a set of coupled publications was significantly higher than the average word-profile similarity (to the word-profile for the totality of the coupled publications) across publications not coupled with the highly cited publication.

Calado, Cristo, Moura, de Ziviani, Ribeiro-Neto, and Gonçalves (2003) used a set of pre-classified Web pages and tested the effectiveness of several link-based similarity measures and text-based methods with respect to automatic classification. In a Web context, two pages are bibliographically coupled if they have at least one common out-link. The classification results obtained by coupling were less accurate than the results obtained by text-based methods, with respect to the measure F_1, which combines precision and recall. This can be explained by the fact that coupling relies on out-links, corresponding to cited references in a journal article, while most pages in the used test collection have no links of this kind. Link-based measures and text-based methods were not only studied in isolation, but also combined. The combinations gave rise to mixed results, though. The work reported by the authors was continued in Calado, Cristo, Gonçalves, Moura, de Ribeiro-Neto, and Ziviani (2006), where five link-based similarity measures were compared to, and combined with, text-based methods. The text-based methods were clearly outperformed, with respect to F_1, by three of the five link-based similarity measures. When these three link-based measures were combined with the text-based methods, irrespective of which, a slightly better performance was achieved, compared to the link-only cases. Bibliographic coupling performed poorly, for the same reason that was given earlier in this paragraph.

Text and citations were combined by Cao and Gao (2005), who dealt with classification of scientific documents. First, textual data were used to compute the weights of the subject categories with respect to a given unknown document. Then these weights were iteratively updated by using category data from documents that are cited by, or that cite, the document. After the last iteration, the document was assigned to the category with the largest weight. In comparison to the best performing text-based method of the study, the combination method improved classification accuracy. In Couto et al. (2006), text-based classifiers were compared to citation-based classifiers. Two document collections were used in the experiments: a Web directory, covering a broad area of topics, and a collection of computer science papers. For the former collection, the text-based classifiers yielded better classification accuracy (measured by F_1) than the coupling classifier. However, for the collection of computer science papers, where the text contained in the title and abstract was used for indexing the papers, the coupling classifier outperformed the text-based classifiers. The authors further combined text-based and citation-based classifiers. In one of the used combination methods, the classification of a document was accomplished by selecting the more appropriate classifier, based on an estimation of its reliability. This method yielded an accuracy gain of about 14%, with respect to the Web directory.

Zhu, Yu, Chi, and Gong (2007) proposed a technique, involving factor analysis, for seamlessly combining textual and link data in classification tasks. The technique involves a joint factorization on both the linkage adjacency matrix and the document-by-term matrix, and a new representation of the documents in a low-dimensional factor space is derived. In the reported experiments, the relative classification performance of the proposed technique, two other text-link combination methods, one text-only and two link-only methods was investigated. Two data sets were used in the experiments: a set of Web pages from computer science departments, and a set of records (abstracts and references) of computer science papers. The computer science papers, but not the Web pages, were *subject* classified by the tested methods, and we therefore concentrate on these papers. Each of the utilized computer science papers belongs to exactly one of four main computer science fields. Moreover, each such field is divided into a number of non-overlapping subfields, so each paper belongs to exactly one computer science subfield. A variant of the proposed technique had the best performance, for each of the four main fields, with respect to the percentage of correct classified papers. The two link-only methods performed poorly: for three of the four main fields, these methods had the two lowest performance values.

Janssens, Quoc, et al. (2006) assessed clustering performance of several document–document similarity approaches. A set of bioinformatics-related papers was used in the experiments. Some approaches used only text, some used only cited references, while a number of approaches integrated textual content with citation-based information. Silhouette values were computed on approach type independent medical subject headings (MeSH)-by-document matrices. It turned out that (normalized) bibliographic coupling performed better than text-only when 2-cluster solutions were considered, while text-only performed better than coupling when 7-cluster solutions were considered. However, both these approaches were outperformed by integrated approaches. Janssens, Glänzel, and Moor (2007) studied the cognitive structure and dynamics of bioinformatics, using a large set of records from the Web of Science. Based on earlier findings, an integrated (statistical combination) approach to document–document similarity was used. The authors conclude that the employed approach is a powerful tool to decipher the cognitive structure of scientific or technological fields.

Ahlgren and Jarneving (2008) worked with a collection of articles on information retrieval (IR). Two document–document similarity approaches were compared in the context of science mapping: bibliographic coupling and a text approach based on the number of common abstract stems. An information retrieval expert performed a classification of the test articles. The cosine measure was used for normalization, and the complete linkage method was used for clustering the articles. The agreement between the two cluster solutions, one for each approach, was fairly low, according to the adjusted Rand index. The classification generated by the expert contained larger groups compared to the coupling and stems solutions, and the agreement between the two solutions and the classification was not high. According to the adjusted Rand index, though, the stems solution was a better approximation of the classification than the coupling solution. With respect to cluster quality, the overall Silhouette value was slightly higher for the stems solution.

3. Data and methods

We used the same raw data as was used by Ahlgren and Jarneving (2008): 43 bibliographic records of genuine articles, published in the journal *Information Retrieval* within the accession period 2004–2006. These records were downloaded from the Web of Science. Each record contained at least one cited reference, a title and an abstract.

The behavior of the document–document similarity approaches was investigated under two types of similarities, first-order and second-order. First-order similarities are obtained by measuring the similarity between columns in a term/reference-by-document matrix, an operation that yields a document-by-document similarity matrix. However, one may go one step further and obtain the similarities by measuring the similarity between columns (similarity *profiles*) in this latter matrix. This operation yields a new document-by-document similarity matrix, populated with second-order similarities. In the first-order strategy, one focuses on the direct similarity between two documents, in the second-order strategy on the way these documents relate to other documents in the data set.[1]

One advantage of the second-order strategy is that it is able to detect that two documents are similar by detecting that there are other documents such that the two documents are both (directly) similar to each of these other documents. Moreover, Janssens (2007) observed good performance of the strategy.

For each of the five document–document similarity approaches, regardless of the type of similarities, the cosine measure (Baeza-Yates, & Ribeiro-Neto, 1999) was used to compute the similarity between two articles. This measure gives the cosine of the angle between the two vectors, which represent the documents d_i and d_j. With respect to first-order similarities, the cosine measure can be formulated as

$$sim1(d_i, d_j) = \frac{\sum_{m=1}^{k} w_{m,i} \times w_{m,j}}{\sqrt{\sum_{m=1}^{k} (w_{m,i})^2} \times \sqrt{\sum_{m=1}^{k} (w_{m,j})^2}} \tag{1}$$

where $w_{m,i}$ ($w_{m,j}$) is the weight of object o_m (a term or a reference) in d_i (d_j). If o_m is a reference, the weights are binary, i.e., a given weight is either 0 (the corresponding reference is absent in the document) or 1 (the corresponding reference is present in the document).

For second-order similarities, we reformulate the cosine measure in terms of *sim*1 as follows:

$$sim2(d_i, d_j) = \frac{\sum_{m=1}^{N} sim1(d_m, d_i) \times sim1(d_m, d_j)}{\sqrt{\sum_{m=1}^{N} (sim1(d_m, d_i))^2} \times \sqrt{\sum_{m=1}^{N} (sim1(d_m, d_j))^2}} \tag{2}$$

where N is the number of documents in the collection, in our case 43.

[1] The two similarity strategies are discussed in Ahlgren, Rousseau, and Jarneving (2003), but in the context of author–author similarity.

For each approach, we converted the similarity values obtained by Eqs. (1) and (2) to corresponding dissimilarity values by subtracting a given similarity value from 1. The reason for using dissimilarities, rather than similarities, is that the Silhouette measure (Section 3.5) is defined in terms of dissimilarities.

Below (Sections 3.1–3.3), we describe the approaches in terms of their associated matrices. The description is given in relation to the first-order strategy, but we occasionally refer to the second-order strategy. With respect to the latter, second-order similarity matrices were obtained from first-order similarity matrices, with the aid of Eq. (2), and the former matrices were then transformed to second-order dissimilarity matrices.

3.1. Text-based approaches

Terms were extracted from the abstract and title of each bibliographical record, neglecting stop words appearing in a freely available stop word list for English ("Stopword List 1", 2000). In an effort to counteract the problem of morphological variation of terms, each remaining term was transformed to its stem by the Porter stemmer (Porter, 2001). In that way, a record with *query* in its abstract or title would have at least one term in common with a record containing *queries* in its abstract or title, since both terms are transformed to *queri* by the Porter stemmer.

3.1.1. Term-by-article matrix of tf-idf values

A term-by-article matrix $\mathcal{A} = \{a_{mi}\}$ was created, where a given row contained the weights of the corresponding term (stem) across the 43 articles (columns). We used the well-known *term frequency-inverse document frequency* (tf-idf) scheme for generating term weights (Baeza-Yates, & Ribeiro-Neto, 1999). $a_{mi} = w_{m,i}$ is then defined as

$$w_{m,i} = freq_{m,i} \times \log\left(\frac{N}{n_m}\right) \tag{3}$$

where $freq_{m,i}$ is the frequency of term t_m in article d_i, i.e., the number of occurrences of term t_m in d_i, N the number of documents in the collection, and n_m the number of documents in the collection in which term t_m occurs. Eq. (1) was applied to $\mathcal{A}$, which yielded an article-by-article similarity matrix, which was transformed to a dissimilarity matrix, $\mathcal{D}_1$.

3.1.2. SVD of the term-by-article matrix of tf-idf values

For the second text-based approach, *singular value decomposition* (SVD) was applied to the term-by-article matrix $\mathcal{A}$, described in the preceding section. The rationale for SVD is that it is reasonable to assume that there is, in a given term-by-document matrix, a latent semantic structure that is somewhat hidden due to the document authors' usage of different words for the same concept. SVD can be used to construct a semantic space (or concept space) that reflects the major associative patterns in the data, and ignores the less important ones (Berry, 1999; Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990).

In order to construct a low-rank approximation of $\mathcal{A}$, only the s largest singular values are kept and the remaining ones are set to zero. This results in a matrix $\hat{\mathcal{A}}$ with rank s, which is the best s-rank approximation to the original term-by-article matrix $\mathcal{A}$, in the least square sense. However, a major difficulty lies in the selection of a value for s. The value should be large enough to fit all the relevant association patterns but small enough so that the noise in the original matrix in not modeled. In some cases an heuristic is be used to obtain the "optimal" value for s. In other cases empirical evidence is used to guide the choice. In this study, we use an heuristic based on the share of cumulated singular values (Wild, Stahl, Stermsek, & Neumann, 2005). Let $s_1, s_2, \ldots, s_p$ be the sequence of singular values, in descending order. Then the value of s is equated to the smallest $i\,(1 \leq i \leq p)$ such that $(s_1 + \cdots + s_i)/(s_1 + \cdots + s_p) \geq 0.9$. Informally, the value of s is equated to the first position in the descending sequence of singular values, where their sum divided by the sum of all the singular values meets or exceeds 0.9. The used heuristic yielded 40 as a value on s.

We applied Eq. (1) to the matrix $\hat{\mathcal{A}}$, which gave rise to an article-by-article similarity matrix. The similarity matrix was then transformed to a dissimilarity matrix, $\mathcal{D}_2$.

3.2. The bibliographic coupling approach

After editing a few spelling variants of the cited references, a reference-by-article matrix $\mathcal{B}$ was created. A given row m in $\mathcal{B}$, corresponding to the reference r_m, contained the weights of r_m across the 43 articles. Here $b_{mi} = w_{m,i}$ is 0 or 1, depending on if r_m is absent or present in article d_i, respectively.

Eq. (1) was applied to $\mathcal{B}$, and in this case the cosine of the angle between the Boolean column vectors of $\mathcal{B}$ was measured. This application yielded an article-by-article similarity matrix. The numerator in Eq. (1) now gives the coupling strength between articles d_i and d_j. In the denominator, the square roots of the lengths of the reference lists of d_i and d_j are multiplied. The coupling strength between two articles was thus normalized with respect to the length of the reference lists (Vladutz & Cook, 1984). Finally, the similarity matrix was tranformed into a dissimilarity matrix, $\mathcal{D}_3$.

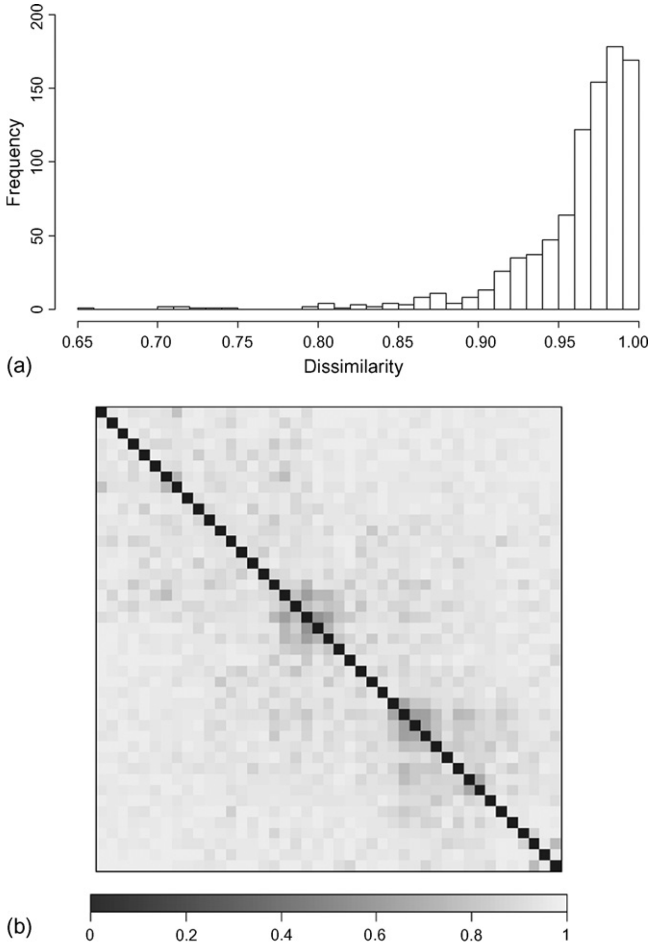

Fig. 1. tf-idf approach. Distribution and underlying dissimilarity matrix. (a) Distribution of dissimilarity values for the tf-idf approach. (b) Visualization of the dissimilarity matrix $\mathcal{D}_1$.

3.3. *Combination approaches*

The information from text-based and citation-based approaches could be combined to potentially improve the clustering of the articles (Janssens et al., 2006; Janssens, 2007). In Figs. 1 and 2, some properties of the two data sources are shown. Fig. 1 concerns the text-based tf-idf approach described in Section 3.1.1, Fig. 2 the bibliographic coupling approach described in the preceding section. Fig. 1(a) shows the distribution of dissimilarity values for the textual approach at stake, while Fig. 2(a) shows the corresponding distribution for the coupling approach. It is clear that the two distributions deviate from each other. In the coupling case, almost all of the observations are found at dissimilarity values greater than or equal to 0.9, with the maximal value 1 having a large share of the observations. In the textual case, the observations are distributed over a longer interval, compared to the coupling case, and the observations at the maximal value 1 constitute, in relation to the coupling case, a small share of the observations. The diverse distribution properties are partly caused by the fact that approximately 80% (1%) of the article pairs with respect to the coupling case (textual case) are such that the two involved articles are not bibliographically coupled (have no common terms). This yields that approximately 80% of the article pairs with respect to the coupling case have the dissimilarity value 1, compared to 1% in the textual case. With a considerably extended sample, one can expect that the observations, in the coupling case, would be distributed over a longer interval, compared to present interval (Fig. 2(a)).

Fig. 1(b) (2(b)) visualizes the dissimilarity matrix $\mathcal{D}_1$ ($\mathcal{D}_3$), from which the distribution in Fig. 1(a) (2(a)) is generated. Matrix shading is used, and entries with lower dissimilarities are plotted darker. It can be seen, for example, that there are article pairs with low dissimilarities (high similarities) based on textual data but with high dissimilarities (low similarities) according to citation data.

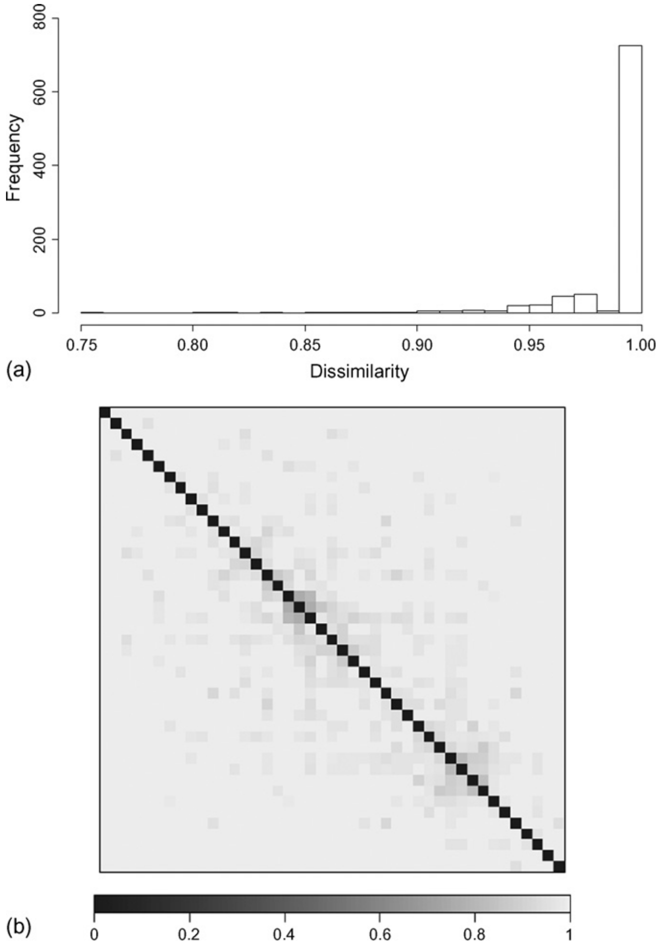

Fig. 2. Bibliographic coupling. Distribution and underlying dissimilarity matrix. (a) Distribution of dissimilarity values for the bibliographic coupling approach. (b) Visualization of the dissimilarity matrix $\mathcal{D}_3$.

3.3.1. Linear combination of dissimilarity matrices

The intuitively easiest way to utilize information from both data sources is probably a weighted linear combination of the two corresponding dissimilarity matrices. An integrated dissimilarity matrix, $\mathcal{D}_{comb}$, is then defined as

$$\mathcal{D}_{comb} = w \times \mathcal{D}_{text} + (1 - w) \times \mathcal{D}_{cit} \tag{4}$$

where $w(0 < w < 1)$ is a weight that can be used to give more importance to a particular source. We combined each of the two text-based approaches with the citation-based approach, and we thereby obtained two integrated dissimilarity matrices. Let $\mathcal{D}_{13}$ ($\mathcal{D}_1$ combined with $\mathcal{D}_3$) and $\mathcal{D}_{23}$ ($\mathcal{D}_2$ combined with $\mathcal{D}_3$) be these two matrices. As for the weight w, we used the value 0.5.

However, a linear combination of this kind might be somewhat problematic as it ignores differences between the involved dissimilarity distributions. So even if the weight w is set to 0.5, the contribution of the data sources might be unequal in the resulting clustering process. In this study, there are differences between dissimilarity distributions (cf. Figs. 1(a) and 2(a)). In the combination approach described in the next section, we use a statistical procedure based on p-values to tackle the indicated problem.

3.3.2. Distribution free combination of transformed dissimilarity matrices

To deal with diverse distribution properties when dissimilarity matrices are to be combined, Janssens et al. (2006) and Janssens (2007) suggest that the dissimilarity values are transformed to p-values under the null hypothesis that the observed similarity between two documents is due to chance alone, i.e., absence of any structure in the data. The p-values are then combined using an statistic from statistical meta-analysis. Our procedure for tackling the problem of diverse distribution properties is heavily inspired by the method developed by these authors.

The omnibus statistical methods for testing the statistical significance of combined results do not depend on the form of the underlying data, but only on the p-values themselves. This is so because observed p-values derived from a continuous test statistic have a uniform distribution under the null hypothesis, regardless of the test statistic or distribution from which they arise (Hedges & Olkin, 1985). This yields, though, that it is necessary to make a small adjustment to Eq. (1) in the case where the dissimilarities are to be calculated from citation data. Since the dissimilarity matrix ($\mathcal{D}_3$) in this case is populated by a vast amount of values equal to maximum dissimilarity (most article pairs do not share any references) these dissimilarities would all be transformed to p-values equal to 1 (and all other transformed to very small p-values). Hence the distribution of p-values under the null hypothesis would be far from uniform. One solution is to make a modification to Eq. (1) by adding a small constant in the numerator, then superposition Gaussian noise (mean = 0, standard deviation = 0.0025) in the numerator, and finally dividing the result by the denominator of Eq. (1). We adopted this solution. 0.01 was used as the constant, and the resulting similarity matrix was transformed to a dissimilarity matrix, $\mathcal{D}_4$. The constant added in the numerator in Eq. (1) has the effect that the ranking between pairs of articles not sharing any references depends more on the lengths of the reference lists, compared to the case where only the noise factor is used.[2] With respect to the second-order strategy, there was no need to use Gaussian noise, since the addition of the constant 0.01 in the numerator of Eq. (1) was sufficient for obtaining, via one first-order and one second-order similarity matrix, a second-order dissimilarity matrix such that its lower left part contained unique dissimilarity values. This matrix is the second-order counterpart to $\mathcal{D}_4$.

We tested the null hypothesis by means of Monte Carlo simulation (Moore, & McCabe, 2005). Let $\mathcal{D}$ be a dissimilarity matrix such that it should be combined with another dissimilarity matrix. Each dissimilarity value in $\mathcal{D}$ was compared to a randomly generated distribution of dissimilarity values, in order to map the value to a p-value. This distribution, based on the matrix $\mathcal{M}$ of weights from which $\mathcal{D}$ was obtained, was generated by applying the following procedure a large number of times:

1. Permute randomly each row in $\mathcal{M}$, independently of each other row in $\mathcal{M}$. Let $\mathcal{M}'$ be the resulting matrix. However, in case of the approach based on SVD of the term-by-article matrix of tf-idf values (Section 3.1.2), we let $\mathcal{M}'$ stand for, not the resulting matrix, but the matrix obtained by SVD of the resulting matrix.
2. For each pair of distinct columns in $\mathcal{M}'$, measure (cosine, or the cosine modification in case of citation data) the similarity between the columns. This gives rise to a similarity matrix, $\mathcal{S}$.
3. Transform $\mathcal{S}$ into a dissimilarity matrix, $\mathcal{D}'$.
4. Record each dissimilarity value that occurs in (the lower left part of) $\mathcal{D}'$.

(For the second-order strategy, the similarity matrix $\mathcal{S}$ in step 2 was transformed to a second-order similarity matrix (Eq. (2)), and this latter matrix was in turn transformed to a dissimilarity matrix.) The procedure described in the list above was iterated 10,000 times. In each iteration, 903 ($= 43 \times 42/2$) simulated dissimilarity values were obtained. A distribution of 9,030,000 ($= 10,000 \times 903$) observations was thereby randomly generated. This distribution was used to transform the dissimilarity values in $\mathcal{D}$ to estimated one-tailed p-values: for a given dissimilarity value d_{ij}, corresponding to the article pair (d_i, d_j), its p-value was computed as the proportion of the observations, which are less than or equal to d_{ij}. Thus, by computing the proportion of the randomly generated dissimilarity values that are less than or equal to the observed value for the pair, we arrived at a p-value for the pair. The p-value indicates the probability to obtain, in the randomized world, a dissimilarity value that is at least as low as the observed dissimilarity value for the two articles. The transformation of the dissimilarity values in $\mathcal{D}$ to p-values gave rise to a dissimilarity matrix of p-values, corresponding to $\mathcal{D}$.

Janssens (2007) used a slightly different simulation approach. A matrix $\mathcal{M}$ of weights was randomized exactly one time. On the basis of the resulting similarity matrix, a distribution of $n(n-1)/2$, where n is the number of documents represented in $\mathcal{M}$, simulated dissimilarity values was generated, and used for transforming original dissimilarity values to p-values. However, we consider our multi-trial approach to be a better choice than the one-trial approach. The multi-trial approach gives a higher precision, i.e., a shorter confidence interval, with respect to the estimation of p-values, compared to the one-trial approach. Further, it is clearly desirable that a given article pair has the same rank in an original dissimilarity matrix and in the corresponding matrix of p-values. We performed an experiment, where the two approaches were compared as to the number of article pairs with the same rank in two such matrices. We used the term-by-article matrix $\mathcal{A}$ and its corresponding dissimilarity matrix, $\mathcal{D}_1$ (Section 3.1.1). Both the one-trial and the multi-trial approach were applied to $\mathcal{A}$, and two distributions of simulated dissimilarity values were thereby obtained. The original dissimilarity matrix $\mathcal{D}_1$ was then transformed to two matrices of p-values, $\mathcal{D}^{one}$ and $\mathcal{D}^{multi}$, where the one-trial distribution was used for generating $\mathcal{D}^{one}$, the multi-trial distribution for generating $\mathcal{D}^{multi}$. The outcome of the experiment was that 435 article pairs had the same rank in $\mathcal{D}_1$ and $\mathcal{D}^{one}$, while all 903 article pairs had the same rank in $\mathcal{D}_1$ and $\mathcal{D}^{multi}$. Behind the relatively low number for $\mathcal{D}_1$ and $\mathcal{D}^{one}$ lies the fact that distinct dissimilarity values in $\mathcal{D}_1$ not seldom were transformed to the same p-value.

We decided to combine each of the two text-based dissimilarity matrices, $\mathcal{D}_1$ and $\mathcal{D}_2$, with $\mathcal{D}_4$. Monte Carlo simulation, and the following transformation of dissimilarity values to p-values, then gave rise to two pairs of matrices of p-values (involving three matrices), corresponding to the matrix pairs $(\mathcal{D}_1, \mathcal{D}_4)$ and $(\mathcal{D}_2, \mathcal{D}_4)$. Now, two p-values are associated with a given article pair and a given pair of original matrices. One value, p_1, originates from textual data and corresponds to (a dissimilarity value in) the first component in the matrix pair, while the other, p_2, originates from citation data and corresponds to (a dissimilarity

[2] The approach to add a small constant in the numerator is called “dense bibliographic coupling” by Janssens (2007). In the same work, it is shown that the approach can lead to comparable clustering performance when compared to the standard approach.

value in) the second component in the matrix pair. In order to combine these two *p*-values, we calculated an integrated statistic, Z_{int}, using the *inverse normal method* (Hedges & Olkin, 1985). We did not apply *Fisher's omnibus test* (Hedges & Olkin, 1985), used by Janssens (2007), since this test has a, relative to our context, drawback. The test is asymmetrically sensitive to small *p*-values compared to large *p*-values (Rice, 1990; Whitlock, 2005). The inverse normal method avoids this drawback.

According to the inverse normal method, p_i $(i = 1, 2)$ is first transformed to a value from a normal distribution with mean 0 and standard deviation 1: the value Z_i such that the probability to obtain a value less than or equal to Z_i is equal to p_i. More formally, Z_i is defined by $p_i = \Phi(Z_i)$, where $\Phi(x)$ is the cumulative distribution function (note that the inverse of the cumulative distribution function, $\Phi^{-1}(x)$, maps p_i to Z_i).

Then, to get the integrated statistic Z_{int} for the article pair, the sum of Z_1 and Z_2 is divided by the square root of the number of *p*-values that should be combined (Hedges & Olkin, 1985):

$$Z_{int} = \frac{\sum_{i=1}^{2} Z_i}{\sqrt{2}} \tag{5}$$

Under the assumption that the null hypothesis is true, Z_{int} has the standard normal distribution. To obtain the combined *p*-value, p_{12}, for the given article pair, Z_{int} for the pair is compared to the standard normal distribution. For example, if Z_{int} is equal to −1.96 (1.96), then p_{12} is equal to 0.025 (0.975). With the aid of Monte Carlo simulation and by using the inverse normal method in conjunction with the standard normal distribution, we obtained two dissimilarity matrices of combined *p*-values. Let $\mathcal{D}_{14}$ and $\mathcal{D}_{24}$ be these matrices, which correspond to the matrix pairs $(\mathcal{D}_1, \mathcal{D}_4)$ and $(\mathcal{D}_2, \mathcal{D}_4)$, respectively.

It might be fruitful to use weights for the individual transformed *p*-values and thereby give more importance to a particular data source (Hedges & Olkin, 1985):

$$ZW_{int} = \frac{\sum_{i=1}^{2} w_i Z_i}{\sqrt{\sum_{i=1}^{2} w_i^2}} \tag{6}$$

where w_i is a non-negative weight. How to select the weights is not obvious. One possibility, though, is to base the selection on the relative structure present in the different data sources. We used the agglomerative coefficient (AC) to measure the clustering structure of a data source (Kaufman & Rousseeuw, 1990). Let *S* be the set objects to be clustered. For a given object *i* in *S*, let $d(i)$ denote the dissimilarity value of *i* to the first cluster it is merged with, divided by the dissimilarity value of the merger in the last step of the clustering process. Then AC is defined as

$$\mathrm{AC} = \frac{1}{|S|} \sum_{i \in S} (1 - d(i)) \tag{7}$$

It is clear from this definition that AC is the mean of all $(1 - d(i))$. The measure takes values on the interval [0, 1]. A value close to 1 indicates that a very clear structure has been found. A value close to 0 indicates that no natural structure is present in the data source.

The weight w_1 of Z_1, which originates from textual data, can now be defined as

$$w_1 = \frac{\mathrm{AC}_{text}}{\mathrm{AC}_{text} + \mathrm{AC}_{cit}} \tag{8}$$

where AC_{text} is the AC associated with the matrix $\mathcal{D}_1$ (Section 3.1.1), AC_{cit} the AC associated with the matrix $\mathcal{D}_3$ (Section 3.2). We thus used exactly one weight for each of the two data sources (also with respect to the second-order strategy). The weight w_2 of Z_2, which originates from citation data, is equal to $1 - w_1$. If the structure in the different data sources is equally good (or bad) according to AC, then $w_1 = w_2 = 0.5$, and Eq. (6) is reduced to Eq. (5). w_1 turned out to be, rounded to four decimals, 0.6423 (w_2 equal to 0.3577).

The relative structure approach for generating weights to transformed *p*-values gave rise to two dissimilarity matrices of combined weighted *p*-values. Let $\mathcal{D}_{(14)_w}$ and $\mathcal{D}_{(24)_w}$ be these matrices, which correspond to the matrix pairs $(\mathcal{D}_1, \mathcal{D}_4)$ and $(\mathcal{D}_2, \mathcal{D}_4)$, respectively. Note that the weight w_1, associated with textual data, is used both when $\mathcal{D}_1$ is combined with $\mathcal{D}_4$, and when $\mathcal{D}_2$ is combined with $\mathcal{D}_4$.

Regarding the second-order strategy, AC_{text} is the AC associated with the second-order counterpart to $\mathcal{D}_1$, whereas AC_{cit} is the AC associated with the second-order counterpart to $\mathcal{D}_3$. For this strategy, $w_1 = 0.6548$, and $w_2 = 0.3452$, values close to the corresponding values in the first-order strategy case.

3.4. Summary of the different approaches

In the list below, we summarize the five different approaches to document–document similarity that we experimentally compare in this work. Note that we actually compare nine document–document similarity methods, since the approaches 4 and 5 involve two and four methods, respectively (we consider the approaches 1–3 as methods). Labels for the nine methods are given within parentheses, and the first-order dissimilarity matrices associated with the methods, and used for clustering, are indicated within parentheses. Each label is separated from its corresponding matrix symbol by a comma (","). "dbc" stands for "dense bibliographic coupling" (Janssens, 2007). We underline that each first-order dissimilarity matrix indicated has a second-order counterpart. Since each method was tested under two types of similarities, first-order and second-order, $9 + 9 = 18$ runs with respect to clustering were done.

1. Approach based on a term-by-article matrix of tf-idf values (text-tfidf, $\mathcal{D}_1$).
2. Approach based on SVD of the term-by-article matrix of tf-idf values (text-tfidf-svd, $\mathcal{D}_2$).
3. The bibliographic coupling approach (bc, $\mathcal{D}_3$).
4. Linear combination of dissimilarity matrices originating from different data sources (text-tfidf_bc, $\mathcal{D}_{13}$; text-tfidf-svd_bc, $\mathcal{D}_{23}$).
5. Distribution free combination of transformed dissimilarity matrices originating from different data sources. Two versions[3]:
   a. Non-weighted version (text-tfidf_dbc, $\mathcal{D}_{14}$; text-tfidf-svd_dbc, $\mathcal{D}_{24}$),
   b. Weighted version (text-tfidf_dbc_weight, $\mathcal{D}_{(14)_w}$; text-tfidf-svd_dbc_weight, $\mathcal{D}_{(24)_w}$).

3.5. Cluster analysis

Cluster analysis was applied in order to group the 43 test articles. For each considered approach, we used the complete linkage method (Everitt, Landau, & Leese, 2001). This method defines the dissimilarity between two clusters, C_1 and C_2, as the maximum dissimilarity between objects o_1 and o_2, where $o_1 \in C_1$ and $o_2 \in C_2$. The clustering, based on the generated dissimilarity matrices, was handled by R, a free software environment for statistical computing and graphics ("The R project for statistical computing", 2008).

To obtain a best cut, i.e., a best number of clusters, we applied the Silhouette measure (Kaufman & Rousseeuw, 1990). This measure contrasts coherence to separation by comparing within-cluster dissimilarity to between-cluster dissimilarity. For a given object i in the data set, let A be the cluster to which i has been assigned, and let $d(i, C)$ be the average dissimilarity of i to all objects of cluster C, where $C \neq A$. The Silhouette value for i, $s(i)$, is then defined as follows:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{9}$$

where $a(i)$ is the average dissimilarity of i to all other objects of A, and $b(i) = \min_{C \neq A} d(i, C)$. The cluster for which the minimum is attained is called the *neighbor* of i. When A is a singleton cluster, $s(i)$ is set to 0. The Silhouette measure takes values on the interval $[-1, 1]$.

In the Silhouette technique for obtaining a best number of clusters (Kaufman & Rousseeuw, 1990), *overall average silhouette values* are used. Such a value, $\bar{s}(k)$, is defined as

$$\bar{s}(k) = \frac{1}{c}\sum_{i=1}^{c} s(i) \tag{10}$$

where k is the number of clusters for a given cluster solution, and c the number of objects in the data set. In general, each value of k gives rise to a different $\bar{s}(k)$. The Silhouette technique selects that value of k for which $\bar{s}(k)$ is maximal.

3.6. External validation

The title and abstract fields from the 43 bibliographical records were extracted and submitted to an IR expert. The expert then performed a subject classification of the 43 corresponding test articles, on the basis of titles and abstracts. In that way, a ground truth classification of the material was obtained. The expert was instructed to assign a natural language label to each class, a label that he thought was indicative of the semantic content of the articles in the class. For external validation, the agreement between a given cluster solution and the classification was quantified by means of the *adjusted* Rand index (Hubert, & Arabie, 1985). This index is a measure of the degree of agreement between two partitions of the same set of objects, and the upper bound of the measure is 1.

[3] Note that the subscript 4 refers to the dissimilarity matrix $\mathcal{D}_4$ (Section 3.3.2).

4. Results

In this section, we report the results of the experiment. For each of the nine document–document similarity methods tested, and under a given type of similarities (first-order or second-order), the agreement between the corresponding cluster solution and the ground truth classification was measured. Table 1 reports the agreement values (Rand index), as well as the number of clusters for each cluster solution. For example, consider the third row from the bottom of the table. The method text-tfidf-svd_dbc, which falls under the distribution free combination approach, has the values 0.2847 (first-order) and 0.4891 (second-order) on the Rand index, and its cluster solution for first-order similarities has 16 clusters, while its solution for second-order similarities has 20 clusters.

First we consider the performance of the methods under first-order similarities. The best performance, according to the Rand index, is obtained by text-tfidf-svd_bc, a method that combines textual data with citation data and that belongs to the linear combination approach. For textual data, this method makes use of the tf-idf term weighting scheme and SVD. For citation data, bibliographic coupling is used. A value of 0.5642 on the index is obtained by the method, which performs slightly better than text-tfidf-svd (0.5617), a text-only approach/method (tf-idf scheme, SVD). These two methods considerably outperform the other seven methods. The best performing method under the distribution free combination approach is text-tfidf-svd_dbc_weight (tf-idf scheme, SVD, dense bibliographic coupling; transformed *p*-values weighted according to the AC), which has the value 0.3739. The citation-only approach/method bc (bibliographic coupling) exhibits the worst performance, with the value 0.1655 (the lowest value across all tested methods and across the two orders of similarities). The next lowest value is 0.2521, obtained by the linear combination method text-tfidf_bc, so bc is clearly outperformed by each other method. text-tfidf_bc performs considerably worse than text-tfidf-svd_bc, the best performing method. However, the cluster solution for text-tfidf_bc has as many as 32 clusters, while the solution for text-tfidf-svd_bc has 16. This difference in number of clusters, together with the fact that the ground truth classification has 15 classes, partly explains the observed performance difference.

Next we consider performance under second-order similarities. The best performance is obtained by text-tfidf, a text-only method (tfidf scheme, not SVD) with the value 0.7076 on the Rand index. This value is greater than 0.5642 (highest value under first-order similarities) and thereby the highest value across all tested methods and across the two orders of similarities. The value 0.7076 indicates a strong agreement between the cluster solution generated by text-tfidf, in conjunction with second-order similarities, and the ground truth classification of the test articles. The distribution free combination method text-tfidf_dbc_weight (tf-idf scheme, not SVD, dense bibliographic coupling; transformed *p*-values weighted according to the AC) also achieves a high value (0.6919). The best performing method under the linear combination approach is text-tfidf-svd_bc (0.5748). As stated above, this method has the best performance under first-order similarities, but it has only the fourth best performance under second-order similarities. Therefore, the results of the experiment show an interaction between the factors “method” and “order”, i.e., the effect of “method” depends on whether first-order or second-order similarities are considered. Again, the citation-only method bc has a poor performance (0.1823), and it is clearly outperformed by the other methods.

In the following list, we put forward some further observations with respect to performance according to the Rand index:

- The methods perform consistently better under second-order similarities than under first-order similarities.
- With regard to the four distribution free combination methods, the two weighted versions perform better than the corresponding unweighted versions, regardless of whether first-order or second-order similarities are considered.
- With a few exceptions, the performance deteriorates when textual data is combined with citation data, compared to textual data only.
- Under first-order similarities, the methods that involve SVD perform better than the corresponding non-SVD methods. Under second-order similarities, though, the effect of SVD is mixed.

Table 1
Values on the Rand index and number of clusters, under first-order and second-order similarities. The highest value at each order (columns 3 and 4) is in *italics*.

Approaches	Order		# clusters (first/second)
	First	Second	
1. text-tfidf	0.4732	*0.7076*	20/17
2. text-tfidf-svd	0.5617	0.6312	19/20
3. bc	0.1655	0.1823	20/21
4. Linear combination			
text-tfidf_bc	0.2521	0.4498	32/20
text-tfidf-svd_bc	*0.5642*	0.5748	16/18
5. Distribution free combination			
text-tfidf_dbc	0.2813	0.4311	16/19
text-tfidf-svd_dbc	0.2847	0.4891	16/20
text-tfidf_dbc_weight	0.3397	0.6919	20/14
text-tfidf-svd_dbc_weight	0.3739	0.4984	16/18

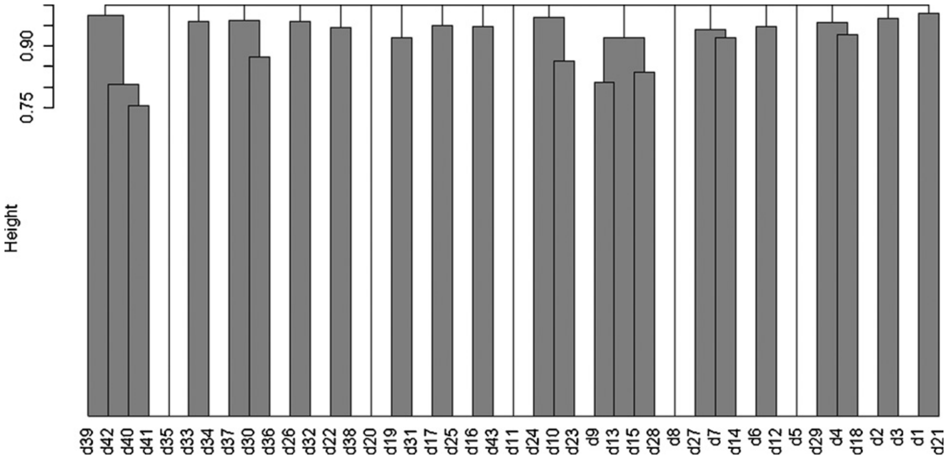

Fig. 3. Dendrogram for the combination bc_FO. The generated cluster solution is visualized.

We now compare in more detail two combinations of method and similarity order: the combination with the worst performance, bc under first-order similarities (bc_FO), and the combination with the best performance, text-tfidf under second-order similarities (text-tfidf_SO). In Figs. 3 (bc_FO) and 4 (text-tfidf_SO), dendrograms corresponding to the two combinations are given. In these dendrograms, the 43 articles are represented by case labels, like “d39”, and the two cluster solutions obtained by the Silhouette technique are indicated. Article titles corresponding to the case labels are given in Appendix A, together with the case labels.

The ground truth classification has, as mentioned above, 15 classes. Five of these contain exactly one article. Table 2 puts forward the classification with, for each class, the label generated by the subject expert. The classification contains considerably fewer classes compared to the bc_FO solution, which has 20 clusters (Fig. 3; Table 1). The text-tfidf_SO solution has 17 clusters (Fig. 4; Table 1).

To illustrate the by far better approximation of the classification achieved by text-tfidf_SO, we consider how the articles in the two classes “Structured document retrieval” and “CLIR (Cross-language IR)” (Table 2) are grouped in the two cluster solutions at stake. The class “Structured document retrieval” has seven articles: d1, d2, d3, d4, d5, d6 and d18. With regard to the bc_FO solution, these articles are distributed over five clusters (Fig. 3). Two of the five clusters have two of the seven articles as members, whereas three of the five clusters have at least one article distinct from each “Structured document retrieval” article in the cluster. By contrast, the text-tfidf_SO solution has a cluster that perfectly matches the class in question, i.e., there is a cluster in the solution that is identical to the class (Fig. 4).

The class “CLIR (Cross-language IR)” has eight articles. In the bc_FO solution, four of these eight articles (d39, d40, d41 and d42) belong to the same cluster (Fig. 3). The remaining four articles (d35, d37, d38 and d43) are distributed over four clusters. Moreover, three of these four clusters contain at least one other article. For example, d38 belongs to a cluster with

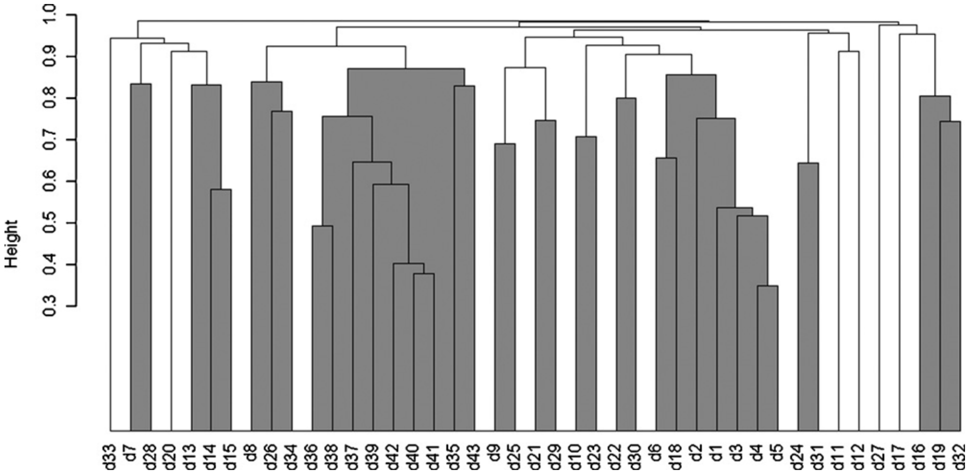

Fig. 4. Dendrogram for the combination text-tfidf_SO. The generated cluster solution is visualized.

Table 2
The ground truth classification of the 43 articles.

Label (generated by the subject expert)	Articles (represented by case labels)
CLIR (cross-language IR)	d35, d37, d38, d39, d40, d41, d42, d43
Structured document retrieval	d1, d2, d3, d4, d5, d6, d18
Ranking/data fusion	d7, d10, d13
Distributed IR	d8
Web IR	d9, d14, d15, d21
Question answering	d11
IR models	d12, d16, d17, d19
Text classification & clustering	d20, d24, d31
IR interfaces & interaction	d22, d29
Query expansion	d23
IR evaluation	d25, d28
Filtering & recommendation	d26, d33, d34
Compression & efficiency	d27
NLP in IR	d30, d36
Topic detection and tracking	d32

two articles, and the other article (d22) has been assigned to the class "IR interfaces & interaction" by the subject expert (Table 2). In the text-tfidf_SO solution, all eight articles in the class belong to the same cluster (Fig. 4), and this cluster contains only one other article, d36, an article that has been assigned to the class "NLP in IR" ("NLP" stands for "Natural Language Processing") by the subject expert (Table 2).

5. Discussion and conclusions

In this study we have considered five approaches to document–document similarity in the context of science mapping. Two of the approaches are text-based, one approach, bibliographic coupling, is citation-based, while the remaining two approaches combine text-based approaches and bibliographic coupling. We experimentally compared nine methods, where each method is either identified with one of the five approaches or falls under one of these, in the context of science mapping. Each method was investigated under two types of similarities, first-order and second-order.

We used 43 articles, published in the journal *Information Retrieval*, as test documents. A subject classification of these articles was performed by an IR expert, in order to obtain a ground truth classification of the material. The cosine measure was used to compute the similarity between articles, and the complete linkage method was used for clustering the articles, irrespective of the underlying combination of method and similarity order. For each such combination, the cluster solution associated with the combination was compared to the ground truth classification with respect to agreement, which was quantified by means of the adjusted Rand index.

For first-order similarities, the method text-tfidf-svd_bc (tf-idf scheme, SVD, bibliographic coupling), which belongs to the linear combination approach, had the best performance (Rand index = 0.5642). Regarding second-order similarities, the best performance was obtained by text-tfidf, a text-only method (tfidf scheme, not SVD), with the value 0.7076 on the Rand index. The citation-only method bc (bibliographic coupling) exhibited the by far worst performance, regardless of similarity order. All nine methods performed consistently better under second-order similarities than under first-order similarities, a result in agreement with observations made by Janssens (2007).

Not only in our study, but also in several earlier studies, where clustering or classification has been employed, citation-only methods have performed worse than text-only methods (Ahlgren & Jarneving, 2008; Calado et al., 2003, 2006; Zhu et al., 2007). However, there are studies that report mixed results concerning the relative performance of the two kinds of methods (Couto et al., 2006; Janssens et al., 2006). With regard to methods that integrate, or combine, textual data and citation data, Cao and Gao (2005) report a performance gain when text is combined with citations, compared to the best performing text-only method of their study, whereas Zhu et al. (2007) proposed a combination technique that performed better than the text-only method of their study. Further, bibliographic coupling and text-only were outperformed by combination methods in Janssens et al. (2006). In our study, with a few exceptions, performance declines when textual data is combined with citation data, compared to textual data only. However, the best performing method under first-order similarities is a (linear) combination method.

This study develops the research reported by Ahlgren and Jarneving (2008), who compared two approahes, bibliographic coupling and a text-only approach (which did not use the tf-idf scheme or SVD, and which was not tested in this study), under first-order similarities. These authors and the authors of the present study used the same set of test documents, the same stemmer (Porter), the same similarity measure (cosine) and the same clustering method (complete linkage). Each of the nine methods tested in the present study, irrespective of similarity order, performed better than the best-performing approach (text-only) in the earlier study. That approach had a value of 0.1162 on the adjusted Rand index (Ahlgren & Jarneving, 2008), while four of the nine methods have, under first-order similarities, a value on the index that exceeds 0.3500 (nine of the combinations of method and similarity order have a value on the index that exceeds 0.4500). Note, though, that the earlier study used the upper tail rule (Mojena, 1977; Wishart, 2005) to obtain a best number of clusters, and not the

Silhouette technique. It is possible that this best cut difference between the two studies to some extent explains the observed performance differences.[4]

The ground truth classification is associated with 15 themes. This indicates that the field of IR, as represented by one volume of articles published in *Information Retrieval*, is fairly heterogeneous regarding research themes. A considerably extended sample of IR articles might result in even more themes. We underscore, though, that the ground truth classification reflects the view of a single subject expert. Other experts might generate classifications that deviate from the one used in this work. For example, a less fine-grained classification might be generated by another expert.

We have demonstrated that it is possible to achieve a very good approximation of the classification by means of automatic grouping of articles. The text-only method text-tfidf and the distribution free combination method text-tfidf_dbc_weight, under second order similarities, and in combination with the complete linkage method and the Silhouette technique, give rise to cluster solutions that to a large extent agree with the classification. However, we worked with a relatively small dataset, and the obtained results, although promising, should be interpreted with some caution.

It might be reasonable to believe that the relative performance of the tested methods would be approximately the same if the sample is extended, given that the additional articles are on IR. However, the relative performance of the methods might change if a sample of articles from another field than IR is used. Field differences may give rise to relative performance of the methods that varies across fields.

One may expect that the term set of the study covers a larger proportion of the IR vocabulary that would be generated from an extended and much larger sample of IR articles, compared to the reference set of the study in relation to the reference set that would be generated from such a sample. One might therefore ask if the bibliographic coupling approach would perform substantially better with an extended and much larger sample. Janssens et al. (2006) used a larger set of publications than we did (bioinformatics-related, though). However, with regard to the issue in question, the empirical evidence from that study is mixed. Bibliographic coupling performed better than text-only when 2-cluster solutions were considered, while text-only performed better than coupling when 7-cluster solutions were considered.

Future research may compare a full text approach to (some of) the approaches treated in this work. It would be interesting to use data sets larger than the one used in the study, sets representing IR as well as other fields. One might also bring in other clustering methods than complete linkage, since it is possible that the effect of the method factor of this work depends on the clustering method applied. Some empirical evidence for the existence of such an interaction is given in Glenisson et al. (2005), where a full text method was compared to text-only methods that use only a part of the text. Finally, the result that the tested methods consistently perform better under second-order similarities is notable, but more studies are needed on the similarity order issue.

Acknowledgements

The authors wish to thank Kalervo Järvelin for performing the subject classification of the test articles, and Daniel Thorburn for giving valuable remarks on Monte Carlo simulation.

Appendix A. Test articles

The titles and case labels of the test articles used in the study are given in Table 3.

Table 3
Article titles and case labels associated with the 43 test articles.

Case label	Article titles
d6	A Bayesian framework for XML information retrieval: searching and learning with the INEX collection
d4	A fusion approach to XML structured document retrieval
d9	A general evaluation framework for topical crawlers
d11	Analysis of statistical question classification for fact-based questions
d31	Augmenting naive Bayes classifiers with statistical language models
d22	Automatic alphabet recognition
d38	Character N-gram tokenization for European language text retrieval
d42	Combination approaches for multilingual text retrieval
d40	Combining multiple strategies for effective monolingual and cross-language retrieval
d10	Comparing rank and score combination methods for data fusion in information retrieval
d17	Complexity reduction in lattice-based information retrieval
d35	Cross-language evaluation forum: objectives results, achievements
d20	Data driven similarity measures for k-means like clustering algorithms
d13	Dempster-Shafer theory for a query-biased combination of evidence on the Web
d39	Dictionary-based cross-language information retrieval: learning experiences from CLEF 2000–2002
d30	How effective is stemming and decompounding for German text retrieval?
d3	Hybrid XML retrieval: combining information retrieval and a native XML database

[4] In case of bibliographic coupling under first-order similarities, tested in both studies, the best cut difference explains the performance difference.

Table 3 (*Continued*)

Case label	Article titles
d21	Index-based persistent document identifiers
d23	Information retrieval with a hybrid automatic query expansion and data fusion procedure
d43	Interactive cross-language document selection
d27	Inverted index compression using word-aligned binary codes
d7	Learning to rank
d34	Learning user similarity and rating style for collaborative recommendation
d8	Mobile agents for distributed and heterogeneous information retrieval
d36	Monolingual document retrieval for European languages
d41	Multilingual information retrieval using machine translation relevance feedback and decompounding
d19	On event spaces and probabilistic models in information retrieval
d33	OSGS—a personalized online store for e-commerce environments
d18	Personalised indexing and retrieval of heterogeneous structured documents
d15	Rank-stability and rank-similarity of link-based Web ranking algorithms in authority-connected graphs
d28	Replicating web structure in small-scale test collections
d26	Scale and translation invariant collaborative filtering systems
d1	Semantic similarity search on semistructured data with the XXL search engine
d32	Simple semantics in topic detection and tracking
d37	Statistical models for monolingual and bilingual information retrieval
d25	System performance and natural language expression of information needs
d12	Test data likelihood for PLSA models
d24	The combination of text classifiers using reliability indicators
d5	The importance of length normalization for XML retrieval
d14	Theoretical study of a generalized version of Kleinberg's HITS algorithm
d2	TIJAH: embracing IR methods in XML databases
d16	Varying retrieval categoricity using hyperbolic geometry
d29	Within-document retrieval: a user-centred evaluation of relevance profiling

References

Ahlgren, P., & Jarneving, B. (2008). Bibliographic coupling, common abstract stems and clustering: A comparison of two document–document similarity approaches in the context of science mapping. *Scientometrics*, *76*(2), 273–290.

Ahlgren, P., Rousseau, R., & Jarneving, B. (2003). Requirements for a cocitation similarity measure, with special reference to Pearson's correlation coefficient. *Journal of the American Society for Information Science and Technology*, *54*(6), 550–560.

Baeza-Yates, R. A., & Ribeiro-Neto, B. A. (1999). *Modern information retrieval*. Harlow: Addison-Wesley. Modeling (Chapter 2).

Berry, M. W. (1999). Matrices, vector spaces, and information retrieval. *SIAM Review*, *41*(2), 335–362.

Boyce, B. R., Meadow, C. T., & Kraft, D. H. (1994). *Measurement in information science*. San Diego: Academic Press. Clustering, Similarity, and Set Membership Measures (Chapter 7).

Calado, P., Cristo, M., de Moura, E. S., Ziviani, N., Ribeiro-Neto, B., & Gonçalves, M. (2003). Combining link-based and content-based methods for web document classification. In *Proceedings of the 12th ACM international conference on information and knowledge management* (pp. 394–401).

Calado, P., Cristo, M., Gonçalves, M. A., de Moura, E. S., Ribeiro-Neto, B., & Ziviani, N. (2006). Link-based similarity measures for the classification of web documents. *Journal of the American Society for Information Science and Technology*, *57*(2), 208–221.

Cao, M., & Gao, X. (2005). Combining contents and citations for scientific document classification. In *AI 2005: Advances in artificial intelligence* (143–152). Berlin/Heidelberg: Springer.

Couto, T., Cristo, M., Gonçalves, M., Calado, P., Ziviani, N., de Moura, E. S., et al. (2006). A comparative study of citations and links in document classification. In *Proceedings of the 6th ACM/IEEE-CS joint conference on digital libraries* (pp. 75–84).

Deerwester, S., Dumais, S. T., Furnas, G., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391–407.

Everitt, B., Landau, S., & Leese, M. (2001). *Cluster analysis* (4th ed.). London: Arnold. Hierarchical Clustering (Chapter4).

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

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando: Academic Press. Tests of Statistical Significance of Combined Results (Chapter 3).

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

Janssens, F. (2007). *Clustering of scientific fields by integrating text mining and bibliometrics*. Unpublished doctoral dissertation. Katholieke Universiteit, Leuven.

Janssens, F., Glänzel, W., & Moor, B. D. (2007). Dynamic hybrid clustering of bioinformatics by incorporating text mining and citation analysis. In *Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 360–369).

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

Janssens, F., Quoc, V. T., Glänzel, W., & Moor, B. D. (2006). Integration of textual content and link information for accurate clustering of science fields. In *InSCit2006, Current research in information sciences and technologies: Multidisciplinary approaches to global information systems* (Vol. I, pp. 615–619).

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*. New York: Wiley. Partitioning Around Medoids (Chapter 2); Agglomerative Nesting (Chapter 5).

Kessler, M. M. (1963a). Bibliographic coupling between scientific papers. *American Documentation*, *14*(1), 10–25.

Kessler, M. M. (1963b). Bibliographic coupling extended in time: Ten case histories. *Information Storage and Retrieval*, *1*(4), 169–187.

Kessler, M. M. (1965). Comparison of the results of bibliographic coupling and analytic subject indexing. *American Documentation*, *16*(3), 223–233.

Mojena, R. (1977). Hierarchical grouping methods and stopping rules: An evaluation. *Computer Journal*, *20*(4), 359–363.

Moore, D. S., & McCabe, G. P. (2005). *Introduction to the practice of statistics* (5th ed.). New York: W.H. Freeman. Bootstrap Methods and Permutation Tests (Chapter 14).

Peters, H., Braam, R., & van Raan, A. (1995). Cognitive resemblance and citation relations in chemical engineering publications. *Journal of the American Society for Information Science*, *46*(1), 9–21.

Porter, M. (2001). Snowball: A language for stemming algorithms. Available from http://snowball.tartarus.org/texts/introduction.html (visited October 1, 2008).

Rice, W. R. (1990). A consensus combined p-value test and the family-wide significance of component tests. *Biometrics*, *46*(2), 303–308.

Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval*. New York: McGraw-Hill. Retrieval Refinements (Chapter 6)

Small, H., & Koenig, M. (1977). Journal clustering using a bibliographic coupling method. *Information Processing & Management*, *13*(5), 277–288.

Stopword List 1. (2000). Available from http://www.lextek.com/manuals/onix/stopwords1.html (visited October 1, 2008).

The R project for statistical computing. (2008). Available from http://www.r-project.org (visited October 1, 2008).

Vladutz, G., & Cook, J. (1984). Bibliographic coupling and subject relatedness. In *Proceedings of the 47th ASIS annual meeting* (pp. 204–207).

Whitlock, M. C. (2005). Combining probability from independent tests: The weighted *z*-method is superior to Fisher's approach. *Journal of Evolutionary Biology*, *18*(5), 1368–1373.

Wild, F., Stahl, C., Stermsek, G., & Neumann, G. (2005). Parameters driving effectiveness of automated essay scoring with LSA. In *Proceedings of the 9th international computer assisted assessment (CAA) conference* (pp. 485–494).

Wishart, D. (2005). Number of clusters. In B. Everitt, & D. D. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (pp. 1442–1446). Chichester: John Wiley & Sons.

Zhu, S., Yu, K., Chi, Y., & Gong, Y. (2007). Combining content and link for classification using matrix factorization. In *Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 487–494).