Textual Content, Cited References, Similarity Order, and Clustering: An Experimental Study in the Context of Science Mapping

Per Ahlgren[1] and Cristian Colliander[2]

[1] *per.ahlgren@sub.su.se*
Department of e-Resources, University Library, Stockholm University, SE-106 91 Stockholm (Sweden)

[2] *cristian.colliander@hs.hj.se*
University Library, Jönköping University, SE-551 11 Jönköping (Sweden)

Abstract

This paper deals with document-document similarity approaches, the issue of similarity order, and clustering methods, in the context of science mapping. Using two data sets of bibliographic records, associated with the fields of information retrieval and scientometrics, we investigate how well two document-document similarity approaches, a text-based approach and bibliographic coupling, agree with ground truth classifications (obtained by subject experts), under first-order and second-order similarities, and under four different clustering methods. The clustering methods are average linkage, complete linkage, Ward's method and consensus clustering. The performance of first-order and second-order similarities is compared within the two document-document similarity approaches, and under each clustering method. We also compare the performance of the clustering methods. The results show that the text-based approach consistently outperformed bibliographic coupling with regard to the information retrieval data set, but performed consistently worse than the latter approach regarding the scientometrics data set. For the similarity order issue, second-order similarities performed better than first-order in 12 out of 16 cases. Average linkage had the best overall performance among the clustering methods, followed by consensus clustering. The main conclusion of the study is that second-order similarities seem to be a better choice than first-order in the science mapping context.

Introduction

In many cases, science mapping aims at dividing the units of analysis into groups such that the units in a given group exhibit (1) a high degree of mutual similarity, and (2) a low degree of similarity to the units in the other groups. When the units of analysis are documents, for example journal articles, one is normally interested in semantic similarity. In this paper, the term *similarity* and its morphological variants refer to semantic similarity. When a given scientific field is to be mapped on the basis of a set of documents, the textual content of two documents can be used as an information source for measuring the similarity between them. In the literature, several similarity measures, based on terms that occur in both documents, have been suggested (Boyce, Meadow, & Kraft, 1994; Salton & McGill, 1983). Text-based approaches to document-document similarity have been employed in science mapping by, for instance, Glenisson, Glänzel, Janssens, & De Moor (2005) and Janssens, Leta, Glänzel, & De Moor (2006), where bibliometrics and library and information science were mapped, respectively.

Bibliographic coupling (Kessler, 1963a, 1963b, 1965), another approach to document-document similarity, uses the cited references of two documents as a source for measuring the similarity between them. The (unnormalized) similarity between the documents is taken to be their *coupling strength*: the number of shared references.

The two information sources, textual content and cited references, can be combined (Ahlgren & Colliander, 2009; Janssens, 2007; Janssens, Glänzel, & Moor, 2007; Janssens, Quoc, Glänzel, & Moor, 2006). The combination may be achieved by different methods, for instance by statistical combination of two dissimilarity values associated with the same pair of documents. Regardless of the underlying approach, when the similarity values have been

Proceedings of ISSI 2009, pages 862–873.
Edited by B. Larsen and J. Leta.

obtained, the documents may be automatically grouped into pairwise disjunct sets by the application of clustering methods.

Earlier research has provided insight into the relative performance of document-document similarity approaches with respect to clustering and automatic classification. In several studies, citation-only methods have performed worse than text-only methods (Ahlgren & Colliander, 2009; Ahlgren & Jarneving, 2008; Calado, et al., 2006; Calado, et al., 2003; Zhu, Yu, Chi, & Gong, 2007). However, there are studies that report mixed results concerning the relative performance of the two kinds of methods (Couto, et al., 2006; Janssens, Quoc, et al., 2006). With regard to methods that combine textual data and citation data, Cao & Gao (2005) report a performance gain when text is combined with citations, compared to the best performing text-only method of their study. Zhu, et al. (2007) proposed a combination technique that performed better than the text-only method of their study, and bibliographic coupling and text-only were outperformed by combination methods in Janssens, Quoc, et al. (2006). However, in the study by Ahlgren & Colliander (2009) performance declined, with a few exceptions, when textual data was combined with citation data, compared to textual data only.

The behavior of different document-document similarity approaches can be 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. One may go one step further, though, 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. The second-order strategy takes higher-order co-occurrences of terms into account, a property that the strategy has in common with Latent Semantic Indexing (LSI).

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. Ahlgren & Colliander (2009), who used articles on information retrieval (IR) as test documents, compared experimentally nine document-document similarity methods with respect to the degree of agreement between obtained cluster solutions and a ground truth classification. For clustering, complete linkage was used, and the performance of the nine methods was studied under both first-order and second-order similarities. The tested methods consistently performed better under second-order similarities, an interesting outcome.

In this work, where two data sets are used, we investigate how well two document-document similarity approaches, a text-based approach and bibliographic coupling, agree with ground truth classifications, under first-order and second-order similarities, and under four different clustering methods. The performance of first-order and second-order similarities are compared within the two document-document similarity approaches, for both data sets and under each clustering method. The similarity order issue has not, to our knowledge, been treated much in the bibliometric literature. In this work, then, this issue is further illuminated. We also compare the performance of the clustering methods. One of these methods is consensus clustering, and we have not seen any bibliometric work where this approach has been applied.

The remainder of the paper is organized as follows. In the next section, we describe the data used in the study, as well as applied methods. The data and methods section is followed by the findings section, and the findings are discussed in the subsequent section. In the last section, we put forward conclusions.

Data and methods

Two sets of raw data were used in the study. The first set, which we denote by "InfRet", contains 43 bibliographic records of articles, published in the journal *Information Retrieval* and indexed in the Web of Science during 2004-2006. InfRet, which corresponds to the IR field, has been used in two earlier, and to our study related, works (Ahlgren & Colliander, 2009; Ahlgren & Jarneving, 2008). The second set, denoted by "SciMet", contains 58 bibliographic records of articles, related to the field of scientometrics and published in the four journals *Journal of Documentation*, *Journal of Information Science*, *Journal of the American Society for Information Science and Technology*, and *Scientometrics*. All 58 records occur in the 2001 and 2002 CD-ROM volumes of the SSCI. SciMet was generated by Jarneving (2005), who combined bibliographic coupling and the complete linkage method in science mapping. Each record in InfRet and SciMet contains at least one cited reference, a title and an abstract.

The cosine measure (Baeza-Yates & Ribeiro-Neto, 1999) was used to compute the similarity between two articles, for both document-document similarity approaches, and regardless of the type of similarities. The 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:

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

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 $sim1$ as follows:

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

where n is the number of documents in the collection.

The Silhouette measure (briefly treated later in this section), which was used to obtain a best number of clusters, is defined in terms of dissimilarities. Therefore, we converted the similarity values obtained by Eqs. (1) and (2) to corresponding dissimilarity values by subtracting a given similarity value from 1.

Approach based on a term-by-article matrix of tf-idf values

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 order to counteract the problem of morphological variation of terms, each remaining term was transformed to its stem by the Porter stemmer (Porter, 2001).

For each of the two data sets, a term-by-article matrix $\mathsf{A} = \{a_{mi}\}$ was created, where a given row contained the weights of the corresponding term (stem) across the articles (columns) represented by the records in the data set. We applied 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 articles in the collection, and n_m the number of articles in the collection in which term t_m occurs. Eq. (1) was applied to A, which yielded an article-by-article first-order similarity matrix, which was transformed to a first-order dissimilarity matrix. From the first-order similarity matrix, a second-order similarity matrix was obtained with the aid of Eq. (2). This latter matrix was then transformed to a second-order dissimilarity matrix. In this way, two dissimilarity matrices, one first-order and one second-order, were obtained from InfRet, and the same holds for SciMet. We let "text-tfidf" denote the textual approach described in this section.

The bibliographic coupling approach

For each of the two data sets, after editing a few spelling variants of the cited references, a reference-by-article matrix B was created. A given row m in B, corresponding to the reference r_m, contained the weights of r_m across the 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.

We applied Eq. (1) to B, and in this case the cosine of the angle between the Boolean column vectors of B was measured. This application yielded an article-by-article first-order 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).

The first-order similarity matrix was transformed into a first-order dissimilarity matrix. From the first-order similarity matrix, a second-order similarity matrix was obtained with the aid of Eq. (2). This latter matrix was finally transformed to a second-order dissimilarity matrix. Just as in the text-tfidf case, then, four dissimilarity matrices were obtained, two (one first-order and one second-order) from each data set. We let "bc" denote the bibliographic coupling approach.

Clustering methods

The other studies that have used the data sets InfRet and SciMet, (Ahlgren & Colliander, 2009; Ahlgren & Jarneving, 2008) and Jarneving (2005), respectively, have used exactly one clustering method, complete linkage, in order to group test articles. In this study four clustering methods were applied: average linkage, complete linkage, Ward's method and consensus clustering.

Average linkage defines the dissimilarity between two clusters, C_k and C_i, as the average dissimilarity across all pairs of objects (o_a, o_b) such that $o_a \in C_k$ and $o_b \in C_i$. Complete linkage defines the dissimilarity between C_k and C_i as the maximum dissimilarity between o_a and o_b, where $o_a \in C_k$ and $o_b \in C_i$. (Everitt, Landau, & Leese, 2001) In Ward's method, the dissimilarity between a cluster C_k and a (union) cluster $C_i \cup C_j$ formed by fusion of the clusters C_i and C_j, $d(C_k, C_i \cup C_j)$, is given by (Everitt, et al., 2001):

$$d(C_k, C_i \cup C_j) = \alpha_i d(C_k, C_i) + \alpha_j d(C_k, C_j) + \beta d(C_i, C_j), \quad (4)$$

where

$$\alpha_i = (n_k + n_i)/(n_k + n_i + n_j), \alpha_j = (n_k + n_j)/(n_k + n_i + n_j), \beta = -n_k/(n_k + n_i + n_j),$$

and n_k (n_i, n_j) is the number of objects in cluster C_k (C_i, C_j). (4) is a recurrence formula, where the dissimilarity between two clusters is given in terms of the dissimilarities between one of the involved clusters and the two components of the other cluster, and the dissimilarity between these two components.

Consensus clustering can be obtained by integrating the information contained in two or more dendrograms, where the dendrograms correspond to different clustering methods but are associated with the same underlying dissimilarity matrix. Let d_i and d_j be documents, and let T be a dendrogram (a tree), obtained by applying a given clustering method to a matrix of document-document dissimilarities. A matrix $\{h_{ij}\}$ that corresponds to T is constructed, where h_{ij} is the fusion coefficient associated with the smallest subtree of T such that both d_i and d_j belong to it. The value h_{ij} is the *ultrametric distance* between d_i and d_j, and indicates the difference between d_i and d_j in the classification (Gordon, 1999).

When t $(2 \leq t)$ matrices of ultrametric distances, corresponding to t dendrograms (which in turn correspond to t clustering methods), are obtained, the information in the matrices are combined. Let $\{\{h_{ijr}\} : i, j = 1, \ldots, n; r = 1, \ldots, t\}$ be the set of these matrices. We need to find a matrix $\{u_{ij}\}$ of distances such that (Gordon, 1999) (a) the *ultrametric condition* is satisfied, i.e., each triple (d_i, d_j, d_k) is such that the two largest values in the set $\{u_{ij}, u_{ik}, u_{jk}\}$ are equal, and (b) the distances minimize

$$L_2(\{u_{ij}\}) = \sum_{r=1}^{t} w_r \sum_{1 \leq j < i \leq n} (h_{ijr} - u_{ij})^2. \quad (5)$$

In (5), w_r is a weight assigned to the *r*th matrix (and thereby, implicitly, to the *r*th clustering method) (Hornik, 2005). When a matrix $\{u_{ij}\}$ that fulfills the conditions (a) and (b) has been found, it uniquely determines, due to the fulfillment of the ultrametric condition (a), a new dendrogram, which is based on the t original ones. To achieve a minimization of (5), an iterative function minimization algorithm was applied (De Soete, 1984). The algorithm is not guaranteed to give an optimal solution, i.e., a minimal value. In some cases, an approximation of the minimal value is given as output.

Different ways to weight the matrices in question exist. In this study, we used the *cophenetic correlation coefficient* (Sokal & Rohlf, 1962). Let $\{dis_{ij}\}$ be an original document-document dissimilarity matrix. The cophenetic correlation coefficient between $\{dis_{ij}\}$ and the matrix $\{h_{ijr}\}$ is the Pearson correlation coefficient *r* between the corresponding lower left elements

of the two matrices. For a given data set in the study, a given approach and a given similarity order, three matrices of ultrametric distances were obtained, corresponding to average linkage, complete linkage and Ward's method. The weight w for such a matrix was set to the *r* value between the matrix and the document-document dissimilarity matrix used for the given combination of data set, approach and similarity order. It turned out that matrices corresponding to average and complete linkage were assigned fairly similar weights in the eight cases. These weights are considerably higher than the weights assigned to matrices corresponding to Ward's method. It should be observed that without any weighting, or equivalently, if we set the parameter w_r to 1, Ward's method is implicitly weighted higher than the other two methods. The reason is that the maximal ultrametric distance for both average and complete linkage is 1, whereas this is not the case for Ward's method.

For each of the two approaches text-tfidf and bc, and under both first-order and second-order similarities, we combined information from three dendrograms, corresponding to average linkage, complete linkage and Ward's method. With two data sets, we thereby obtained $2 \times 4 = 8$ consensus dendrograms.

All clustering was handled by R, a free software environment for statistical computing and graphics ("The R project for statistical computing," 2008).

Best cut

The Silhouette measure (Kaufman & Rousseeuw, 1990) was used in order to obtain a best cut, i.e., a best number of clusters. This measure contrasts coherence to separation by comparing within-cluster dissimilarity to between-cluster dissimilarity. For more information regarding the Silhouette technique for obtaining a best number of clusters, the reader is referred to Kaufman & Rousseeuw (1990).

Ground truth classifications and external validation

An IR expert performed a subject classification of the 43 articles corresponding to the data set InfRet. The classification was based on the title and abstract fields from the 43 records in InfRet. The expert was instructed to assign a natural language label to each generated class.

With regard to the data set SciMet, an expert on bibliometrics performed a classification of the corresponding 63 articles. The expert used bibliographic data printed on cards, where each card represented one of the 58 articles. For example, the title and abstract of the corresponding article were printed on the card. The expert was instructed to group the cards according to subject similarity, and to assign a label to each generated class. (Jarneving, 2005)

In Appendix A (Tables 4; 5), the two ground truth classifications are given, together with the labels generated by the subject experts.

For external validation, the agreement between a given cluster solution and a given ground truth 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.

Findings

The outcome of the experiment is reported in this section, where we let "FO" ("SO") denote first-order similarities (second-order similarities). In Table 1, which concerns the data set InfRet, agreement values (Rand index) are given, under FO and SO, and under each clustering

method. For each cluster solution, the table further reports the number of clusters in the solution. Note that the results we obtained under complete linkage were obtained also by Ahlgren & Colliander (2009). Table 2 has the same structure as Table 1, but concerns the data set SciMet.

Table 1. InfRet data set (n = 43). Values on the Rand index under first and second-order similarities, under each clustering method. Number of clusters is given within parentheses.

	Complete		**Average**		**Ward**		**Consensus**	
Approaches	FO	SO	FO	SO	FO	SO	FO	SO
1. text-tfidf	0.4732 (20)	*0.7076* (17)	0.5057 (24)	0.6662 (20)	0.3700 (18)	0.4732 (19)	0.5442 (17)	0.6318 (21)
2. bc	0.1655 (20)	0.1823 (21)	0.3023 (12)	0.2506 (19)	0.1368 (21)	0.1406 (18)	0.1540 (19)	0.1758 (16)

Table 2. SciMet data set (n = 58). Values on the Rand index under first and second-order similarities, under each clustering method. Number of clusters is given within parentheses.

	Complete		**Average**		**Ward**		**Consensus**	
Approaches	FO	SO	FO	SO	FO	SO	FO	SO
1. text-tfidf	0.2237 (31)	0.3597 (23)	0.4073 (25)	0.4547 (25)	0.2288 (26)	0.3919 (22)	0.3919 (23)	0.3882 (24)
2. bc	0.4205 (17)	0.4144 (20)	0.4188 (21)	*0.5359* (14)	0.4315 (16)	0.4294 (16)	0.4294 (16)	0.4477 (16)

With respect to InfRet, the highest value on the index (0.7076) is obtained by text-tfidf, under complete linkage and under SO (Table 1). The value 0.7076 indicates a strong agreement between the cluster solution generated by text-tfidf, together with complete linkage and SO, and the ground truth classification for InfRet. text-tfidf performs consistently, and by far, better than bc: in each of the eight cases, bc is clearly outperformed by text-tfidf. The lowest value on the index is then associated with bc (under Ward's method, FO). For the SciMet data set (Table 2), though, the picture is different. The highest value on the index (0.5359) is obtained by bc, under average linkage and under SO, and indicates a good approximation of the ground truth classification for SciMet. Moreover, bc performs better than text-tfidf in each of the eight cases, even if the differences between the two approaches are less pronounced than the corresponding InfRet differences. The lowest value on the index is then associated with text-tfidf (under complete linkage, FO). We observe, however, that the number of clusters for text-tfidf in all eight cases is considerably higher than nine, the number of classes in the ground truth classification for SciMet (Appendix A, Table 5).

Next we compare the performance of FO and SO. There are $2 \times 8 = 16$ cases such that FO and SO can be compared and such that approach and clustering method are constant in the comparisons (Tables 1; 2). In 12 of these 16 cases, SO have a better performance than FO. With regard to text-tfidf, SO perform (considerably) better than FO in seven of the eight cases. In the remaining case, SciMet in conjunction with consensus, FO performs slightly better than SO (Table 2). In the five SciMet cases where SO perform better than FO, the difference between them is not very large, with one exception (bc, average linkage).

In order to illuminate the performance of the four clustering methods, we computed, for each method, and for both InfRet and SciMet, the mean Rand index value across the two

approaches and across the two similarity orders. We also computed means across the data sets, approaches and similarity orders. Table 3, where the rows are ordered descending according to the values in last column (InfRet+SciMet), reports the outcome of the computations. Overall (InfRet+SciMet), average linkage has the best performance, followed by, in turn, consensus clustering, complete linkage and Ward's method. Also in the InfRet and SciMet cases, average linkage performs best. In the InfRet case, average linkage is followed by, in turn, complete linkage, consensus clustering and Ward's method, which in this case is clearly outperformed by the other three methods. In the SciMet case, average linkage is followed by, in turn, consensus clustering, Ward's method and complete linkage.

Table 3. Mean values on the Rand index for the four clustering methods.

	InfRet	SciMet	InfRet+SciMet
1. Average	0.4312	0.4542	0.4427
2. Consensus	0.3765	0.4143	0.3954
3. Complete	0.3822	0.3546	0.3684
4. Ward	0.2802	0.3704	0.3253

We note, finally, that with respect to the eight possible combinations of data set, approach and similarity order, there is no combination such that consensus clustering has the worst performance (Tables 1; 2).

Detailed comparison of two cluster solutions

We now investigate in detail two cluster solutions obtained by the Silhouette technique. We compare the solution that corresponds to Ward's method, in conjunction with InfRet, text-tfidf and SO (Ward_InfRet_text-tfidf_SO), with the solution corresponding to complete linkage, in conjunction with the same three attribute values (CL_InfRet_text-tfidf_SO). Clearly, then, we compare Ward's method and complete linkage, while holding data set, approach and similarity order constant. The value on the Rand index for the solution Ward_InfRet_text-tfidf_SO is 0.4732, while the corresponding value for CL_InfRet_text-tfidf_SO is 0.7076, the highest observed value on the index (Table 1).

In Figs. 1 and 2, dendrograms indicating the two solutions Ward_InfRet_text-tfidf_SO and CL_InfRet_text-tfidf_SO are given. In these dendrograms, the 43 articles are represented by case labels, like "i39". The ground truth classification for InfRet has 15 classes (Appendix A, Table 4). This classification contains fewer classes compared to the solution Ward_InfRet_text-tfidf_SO (19 clusters; Fig. 1; Table 1). CL_InfRet_text-tfidf_SO comes closer to the ground truth classification in this respect, with its 17 clusters (Fig. 2; Table 1).

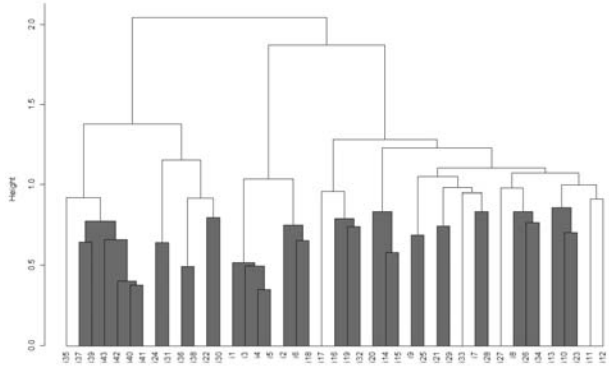

Figure 1. Dendrogram visualizing the cluster solution Ward_InfRet_text-tfidf_SO.

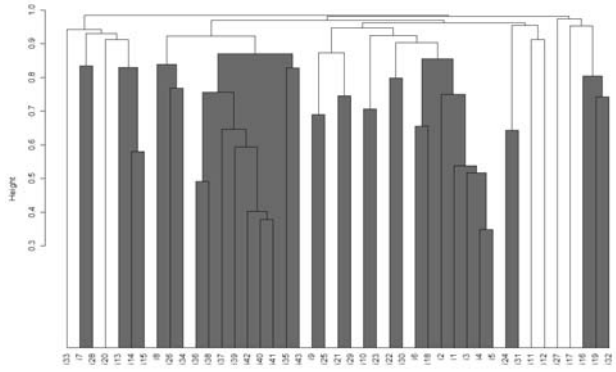

Figure 2. Dendrogram visualizing the cluster solution CL_InfRet_text-tfidf_SO.

To illustrate the better approximation of the classification given by CL_InfRet_text-tfidf_SO, we consider how the articles in the classes "Structured document retrieval" and "CLIR (Cross-language IR)" (Appendix A, Table 4) are grouped in the two cluster solutions. The former class has seven articles: i1, i2, i3, i4, i5, i6 and i18. In the solution Ward_InfRet_text-tfidf_SO, these articles are distributed over two clusters, with four articles in one cluster, three in the other (Fig. 1). By contrast, CL_InfRet_text-tfidf_SO has a cluster that perfectly matches the class "Structured document retrieval", i.e., there is a cluster in the solution that is identical to the class (Fig. 2).

The class "CLIR (Cross-language IR)" is the largest class in the classification with its eight articles (i35, i37, i38, i39, i40, i41, i42, i43). In the solution for Ward_InfRet_text-tfidf_SO, there is a cluster that contains six of the eight articles, and no other article (Fig. 1). The remaining two articles in the class, i35 and i38, are distributed over two clusters, one singleton cluster and one cluster with two articles (Fig. 1). In the CL_InfRet_text-tfidf_SO solution, though, all eight articles in the class belong to the same cluster, and this cluster contains only one other article, i36 (Fig. 2).

Discussion

In this study, we have dealt with two approaches to document-document similarity, the similarity order issue, and clustering methods. Two data sets of bibliographic records over articles were used, corresponding to two different fields of science, and the cosine measure was used to compute the similarity between articles. We compared the performance of two approaches (bc and text-tfidf), under first and second-order similarities, and under four different clustering methods, with respect to how well the approaches agreed with two ground truth classifications, generated by subject experts. We further compared the performance of first and second-order similarities, as well as clustering method performance. The cluster solutions used in the study were obtained by the Silhouette technique, and the agreement between a cluster solution and a ground truth classification was quantified by means of the adjusted Rand index.

For the data set InfRet, the bc approach was consistently and by far outperformed by the text-tfidf approach. For the data set SciMet, we obtained a different outcome: bc performed consistently, and in some cases by far, better than text-tfidf. It is clear, then, that the effect of the approach factor of this work does not depend on the clustering method applied. A partial explanation of the better performance of bc, relative text-tfidf and regarding SciMet, is that text-tfidf consistently gave rise to cluster solutions with a greater number of clusters than the corresponding bc solutions, whereas the number of classes in the ground truth classification

for SciMet is consistently less than the number of clusters in the bc solutions. It is further reasonable to believe that differences between the two fields IR and scientometrics, like degree of vocabulary standardization, underlie the fact that the relative performance of the two approaches varies across the two data sets.

With regard to the similarity order issue, second-order similarities performed better than first-order in 12 out of 16 cases. This outcome is in line with the similarity order results reported by Ahlgren & Colliander (2009), and provides some evidence that second-order similarities are fairly robust with respect to approach and clustering method. For clustering methods, average linkage had the best overall (InfRet+SciMet) performance, with regard to mean values on the Rand index, followed by, in turn, consensus clustering, complete linkage and Ward's method. Average linkage also performed best in this mean respect when InfRet and SciMet were considered separately. Worth noting is that with respect to the eight possible combinations of data set, approach and similarity order, there was no combination such that consensus clustering had the worst performance.

The ground truth classification for InfRet is associated with 15 themes, while the corresponding number for the SciMet classification is eight. This indicates that the field of IR is more heterogeneous regarding research themes than the scientometric field. However, the ground truth classifications reflect the views of two subject experts, and other experts might generate classifications that deviate from the ones used in this work. In the light of this consideration, and the fact that we worked with relatively small data sets, the findings of the study should be interpreted with some caution.

Conclusions

It is possible to achieve good approximations of the two ground truth classifications used in this study by means of automatic grouping of articles. In science mapping, first-order similarities are, to our knowledge, normally used. However, the main conclusion of the study is that second-order similarities seem to be a better choice than first-order in the science mapping context. Consensus clustering behaved in a promising way in the study. For future work, it would be interesting to further investigate the relative performance of consensus clustering, using larger data sets or sets associated with other fields than IR and scientometrics. In a study along these lines, one may also compare the performance of the method we use in this work for weighting matrices of ultrametric distances with the performance of alternative weighting approaches.

Acknowledgments

The authors wish to thank Bo Jarneving for delivering files related to the data set SciMet.

References

Ahlgren, P. & Colliander, C. (2009). Document-document similarity approaches and science mapping: Experimental comparison of five approaches. *Journal of Informetrics*, *3*(1), 49-63.

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.

Baeza-Yates, R. & Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Harlow: Addison-Wesley.

Boyce, B. R., Meadow, C. T. & Kraft, D. H. (1994). *Measurement in Information Science*. San Diego: Academic Press.

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

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).

Cao, M. & Gao, X. (2005). Combining contents and citations for scientific document classification. In *AI 2005: Advances in Artificial Intelligence* (Vol. 3809/2005, pp. 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).

De Soete, G. (1984). A least squares algorithm for fitting an ultrametric tree to a dissimilarity matrix. *Pattern Recognition Letters, 2*(3), 133-137.

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

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), 1548-1572.

Gordon, A. D. (1999). *Classification* (2nd ed.). Boca Raton: Chapman & Hall/CRC.

Hornik, K. (2005). A CLUE for CLUster Ensembles. *Journal of Statistical Software, 14*(12), 1-25.

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.* 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. & De Moor, B. (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).

Jarneving, B. (2005). *The Combined Application of Bibliographic Coupling and the Complete Link Cluster Method in Bibliometric Science Mapping.* Doctoral dissertation. Valfrid, Borås.

Kaufman, L. & Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: Wiley.

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

Kessler, M. M. (1963b). Bibliographic coupling extended in time: 10 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.

Porter, M. (2001). Snowball: A language for stemming algorithms. Retrieved December 13, 2008 from http://snowball.tartarus.org/texts/introduction.html.

Salton, G. & McGill, M. J. (1983). *Introduction to Modern Information Retrieval*. New York: McGraw-Hill.

Sokal, R. R. & Rohlf, F. J. (1962). The comparison of dendrograms by objective methods. *Taxon, 11*(2), 33-40.

Stopword List 1 (2000). Retrieved December 13, 2008 from http://www.lextek.com/manuals/onix/stopwords1.html.

The R project for statistical computing (2008). Retrieved December 13, 2008 from http://www.r-project.org.

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

White, H. D. & McCain, K. (1998). Visualizing a discipline: An author cocitation analysis of information science, 1972–1995. *Journal of the American Society for Information Science*, *49*(4), 327-355.

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).

Appendix A Ground truth classifications

Table 4 gives the ground truth classification for the InfRet data set, while Table 5 gives the corresponding classification for the SciMet data set (note that class 9, the only singleton class in the classification, was not labeled by the subject expert).

Table 4. Ground truth classification of the 43 articles corresponding to the InfRet data set.

Class	Label (generated by the subject expert)	Articles (represented by case labels)
1	CLIR (Cross-language IR)	i35,i37,i38,i39,i40,i41,i42,i43
2	Structured document retrieval	i1,i2,i3,i4,i5,i6,i18
3	Ranking/data fusion	i7,i10,i13
4	Distributed IR	i8
5	Web IR	i9,i14,i15,i21
6	Question Answering	i11
7	IR models	i12,i16,i17,i19
8	Text classification & clustering	i20,i24,i31
9	IR interfaces & interaction	i22,i29
10	Query expansion	i23
11	IR evaluation	i25,i28
12	Filtering & recommendation	i26,i33,i34
13	Compression & efficiency	i27
14	NLP in IR	i30,i36
15	Topic detection and tracking	i32

Table 5. Ground truth classification of the 58 articles corresponding to SciMet data set.

Class	Label (generated by the subject expert)	Articles (represented by case labels)
1	Indicator development; journal impact factor; journal classification; measurement process	s12,s13,s2,s36,s37,s40,s56
2	Mathematical distributions	s16,s17,s35,s44,s53,s58,s6,s7,s9
3	Mapping	s15,s45,s46,s51,s52
4	Collaboration	s10,s21,s22,s24,s41,s42,s43,s48,s54
5	Webometrics	s1,s14,s19,s20,s26,s3,s32,s38,s4,s8
6	Science policy; science & technology; patents	s27,s29,s34,s5,s50,s55
7	Citation behavior	s11,s31,s33
8	Case studies of particular fields	s18,s23,s25,s30,s39,s47,s49,s57
9	-	s28