Improving the Accuracy of Co-citation Clustering Using Full Text[1]

Kevin W. Boyack,[*] Henry Small[**] and Richard Klavans[**]

[*] *kboyack@mapofscience.com*
SciTech Strategies, Inc., Albuquerque, NM 87122 (USA)

[**] *hsmall@mapofscience.com; rklavans@mapofscience.com*
SciTech Strategies, Inc., Berwyn, PA 19312 (USA)

Abstract

Historically, co-citation models have been based only on bibliographic information. Full text analysis offers the opportunity to significantly improve the quality of the signals upon which these co-citation models are based. In this work we study the effect of citation proximity on the accuracy of co-citation clusters. Using a corpus of 270,521 full text documents from 2007, we compare the results of traditional co-citation clustering using only the bibliographic information to results from co-citation clustering where proximity between reference pairs is factored into the pairwise relationships. We find that accounting for reference proximity from full text can increase the textual coherence (a measure of accuracy) of a co-citation cluster solution by 9-20% over the traditional approach based on bibliographic information.

Introduction

For the past 45 years, models of the scientific literature have relied solely on bibliographic information. In general, researchers have utilized the very large bibliographic databases, such as the Web of Science, Scopus and Medline, to create many different types of document-level models of the scientific literature. These models have been developed using a variety of citation-based (co-citation, bibliographic coupling, and direct citation) and text-based (e.g., co-word, LSA, topic modelling) methodologies. Hybrid methods (using both citation and textual data) have also been recently been proposed and tested.

The fact that these models have been largely limited to the use of bibliographic data is historical—it goes back to the days when very little information about the published literature was available in digital form and computing capabilities were meager. Even as more information has become available electronically and computing capabilities have greatly increased, most studies have continued to be based solely on bibliographic data. We suppose that this is due to a variety of factors—an existing comfort level gained by decades of working with bibliographic data, the existence of methods to work with bibliographic data, and the lack of access to full text of scientific articles being primary among them.

[1] This research is supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center (DoI/NBC) contract number D11PC20152. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government. We also gratefully acknowledge Elsevier BV for providing the full text data used in this research.

The reasons to limit citation analysis and science modeling to bibliographic data are declining with each succeeding year. Full text is becoming more available. For example, several hundreds of thousands of full text documents are available from PubMed Central and Citeseer. IARPA has recently initiated its FUSE program to investigate how full text can improve characterization of scientific and technical emergence. Publishers are starting to explore how information from their full text holdings can be converted into viable products and services. Major advances can be expected as full text document collections become more available and research proceeds.

Full text analysis has the potential to fundamentally change the theory and practice of citation analysis and the modelling and measuring of science. Full text contains additional information that has not been available in bibliographic data. At a minimum this includes reference position, proximity of cited references within the text, multiple references at the same reference point, multiple mentions of references (so-called *op. cit.*), section information, and words indicating how an author feels about a reference.

Recent work suggests that the actual similarity between two references in a document is related to their proximity in the text (Callahan, Hockema & Eysenbach, 2010; Elkiss & *al.*, 2008; Gipp & Beel, 2009; Liu & Chen, 2012). We build upon that work by examining if use of reference proximity will result in a more accurate co-citation model of science. In this study we use a relatively large full text document corpus to compare the results of traditional co-citation clustering using only bibliographic information to results from co-citation clustering where proximity between reference pairs is factored into the pairwise relationships. We purposely focus on the clustering of references in this study. Of the three citation-based methods that are commonly in use, co-citation analysis is one that can be directly improved by knowing where the citation appears in the document. The remainder of this article will provide background information relative to the study, and will then describe the data, methods, and results of the study.

Background

Full text sources have been used in studies of citation theory and behavior for decades (Bornmann & Daniel, 2008). These studies have typically been done manually using small (tens to hundreds) sets of documents. Although citation indices have been used to identify citing documents, analysis has largely been done using printed versions of articles.

Researchers have recently developed methods to process full text documents electronically in ways that enable further analysis of references. For example, Citeseer (Giles, Bollacker & Lawrence, 1998; Lawrence, Giles & Bollacker, 1999) displays citing sentences, also known as citances (Nakov, Schwartz & Hearst, 2004), allowing users to understand the reasons for and the context in which others have cited a particular article. Sentiment analysis, an area of active research (Agarwal, Choubey & Yu, 2010; Kilicoglu & Bergler, 2008; Ritchie, Teufel & Robertson, 2008; Small, 2011; Teufel, 2010; Teufel, Siddharthan & Tidhar, 2006), attempts to correlate words and patterns from these citances to citation types and roles. Small and Klavans (2011) published the first study combining co-citation analysis with citation context analysis using a combination of Elsevier and Scopus data.

More recently, several researchers have explored the hypothesis that references that are closer together in the text are more similar than references that are further apart in the text. Elkiss & *al.* (2008) examine reference proximity for 2,497 cited papers extracted from PubMed Central in the

context of creating citation summaries, and found that textual cohesion between citance-based summaries associated with co-cited references follows the pattern [sentence > paragraph > section > article]. Gipp & Beel (2009) were the first to propose using modified co-citation weights based on proximity. Their citation proximity index gives full weight to references in the same sentence, ½ weight to those in the same paragraph, ¼ weight to those in the same chapter, etc. They found that citation-proximity weighted relatedness measures performed twice as well at retrieving relevant documents (were used) than the traditional unweighted approach over 21 paired sets of three documents. Callahan & *al*. (2010) provide a very detailed discussion of the issues, practicalities and implications around weighting based on co-citation proximity. They present case examples showing that reference similarity is correlated with proximity, but do not provide any statistical data.

The most comprehensive study to date, by Liu & Chen (2012), correlates co-citation proximities from the full text of articles in 22 open access journals (22,885,839 co-citations) from BioMed Central (BMC) with section locations and co-citation frequencies. They also showed overlays of sentence, paragraph, section, and article level co-citations on a map of the co-citation network. They found that, in general, closer co-citation proximity is correlated with higher co-citation frequencies. Given that higher frequencies are the basis for nearly everything associated with citation or co-citation analysis, this study provides strong evidence that proximity information from full text results in more accurate co-citation similarity values. Liu & Chen also found that sentence-level co-citations preserved the essential structure of the full (traditional) co-citation network.

Data

In this study, we take the next step in exploring the hypothesis that citation proximity from full text can improve co-citation analysis. We performed multiple co-citation clusterings on the cited references from a set of 270,521 full text documents to determine if co-citation similarities based on reference proximity would give a more accurate set of clusters than similarities based on the traditional approach.

The full text dataset used for this analysis is Elsevier's full text for publication year 2007. The data were obtained in a standardized XML format that allows for consistent extraction of source article metadata, reference tags, reference positions, citances, and reference metadata. Reference metadata were easily matched to reference tags because the reference information included the tag numbers. Source article information (DOI, author, journal, volume, page) and reference information from the full text were matched to metadata from Scopus to join Scopus article IDs to both the source articles and cited references. These integer IDs facilitate efficient calculation of co-citation frequencies and weights. Once the source article had been matched, matching of full text references to the reference information for the article listed in Scopus was relatively easy using the assumption that no two references in a single article would have identical page numbers. In cases with multiple page number matches, matches based on volume and/or journal were used.

Only those data that could be positively matched with Scopus data in terms of source and reference data were used in the analysis. This comprised 270,521 full text articles from 1,606 journals, containing 4,484,815 unique references and 12,569,686 individual reference mentions where the Scopus article ID was identified for the reference. Although Elsevier full text content is not evenly distributed across all areas of science, all major areas of science do have coverage, as shown in Figure 1. Coverage is lowest in computer science, where conference proceedings dominate, and in the social sciences.

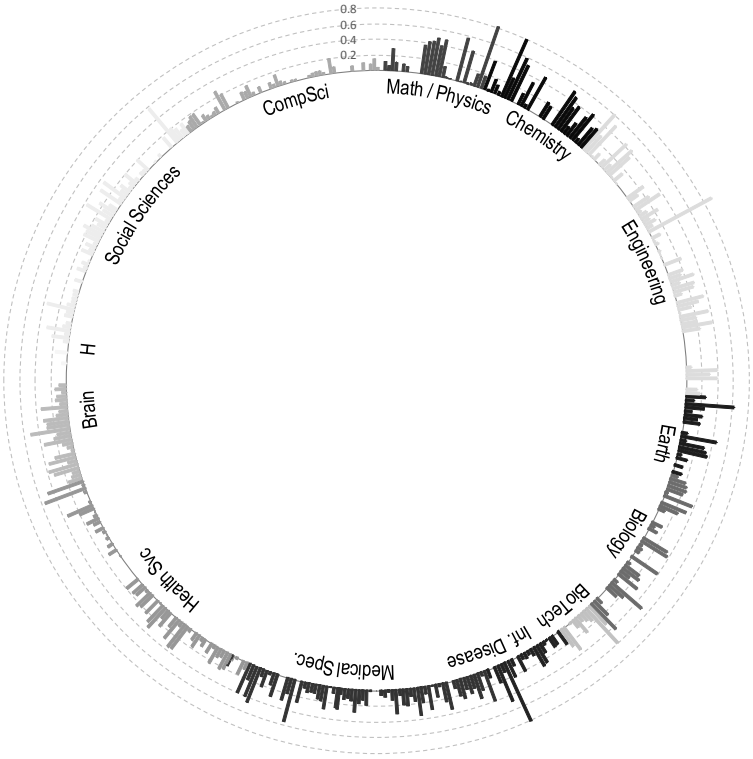

Figure 1. Elsevier full text by discipline as a fraction of Scopus content.

Methodology

As mentioned above, we performed multiple co-citation clusterings of the cited references from a large set of full text documents. Before describing the full process in detail, it is useful to discuss co-citation weighting schemes.

Co-citation weighting

Elkiss & *al*. (2008), Gipp & Beel (2009), and Liu & Chen (2012) all subdivided full text documents in terms of sentence, paragraph, section, and article, and performed analyses at these levels. This is a logical and very defensible choice, especially for small studies or for studies using journals from a single discipline and source. The Elkiss and Liu studies both used BMC journals, which can be expected to have relatively standardized section types and naming. Although the starting and ending points of sections are easily distinguished in the Elsevier full text XML, the naming conventions and orderings of sections and subsections from 1,600+ journals spread over all scientific disciplines are anything but standard. Since we are using the same data extractions as source materials for sentiment analysis (Small, 2011) over this corpus, we decided to avoid using sections.

Specification of reference position in full text documents need not rely on sentence, paragraph, and section demarcations. Relative positions can be calculated using byte (or character) offsets. These offsets can be used with or without normalization. We choose to normalize by the length of the body of the article, and correspondingly convert each reference position into a centile position within the body of the article. Teufel (2010, p. 290) uses a similar methodology, dividing texts into 20 equal parts by byte (equivalent to 5 centiles each) before then recombining segments of different lengths for analysis.

The matrices in Figure 2 show the co-citation weights for the three weighting schemes based on the reference positions shown in the example article at the top of the figure. In the two proximity-based weighting schemes, the minimum distance between any two cited references is used. For scheme P1, references that are in the same bracket are given a weight of 4, while those within 5, 15, and 25 centiles are given weights of 3, 2, and 1, respectively. Scheme P2 is similar, but uses centile proximity bases of 5, 10 and 15 rather than 5, 15 and 25. As more stringent proximity criteria are applied, the numbers of zeros in the matrix increases. We note that although these choices of proximity criteria may appear to be somewhat arbitrary, they are meant to simulate, using centiles, distinctions between sentence, paragraph, and section.

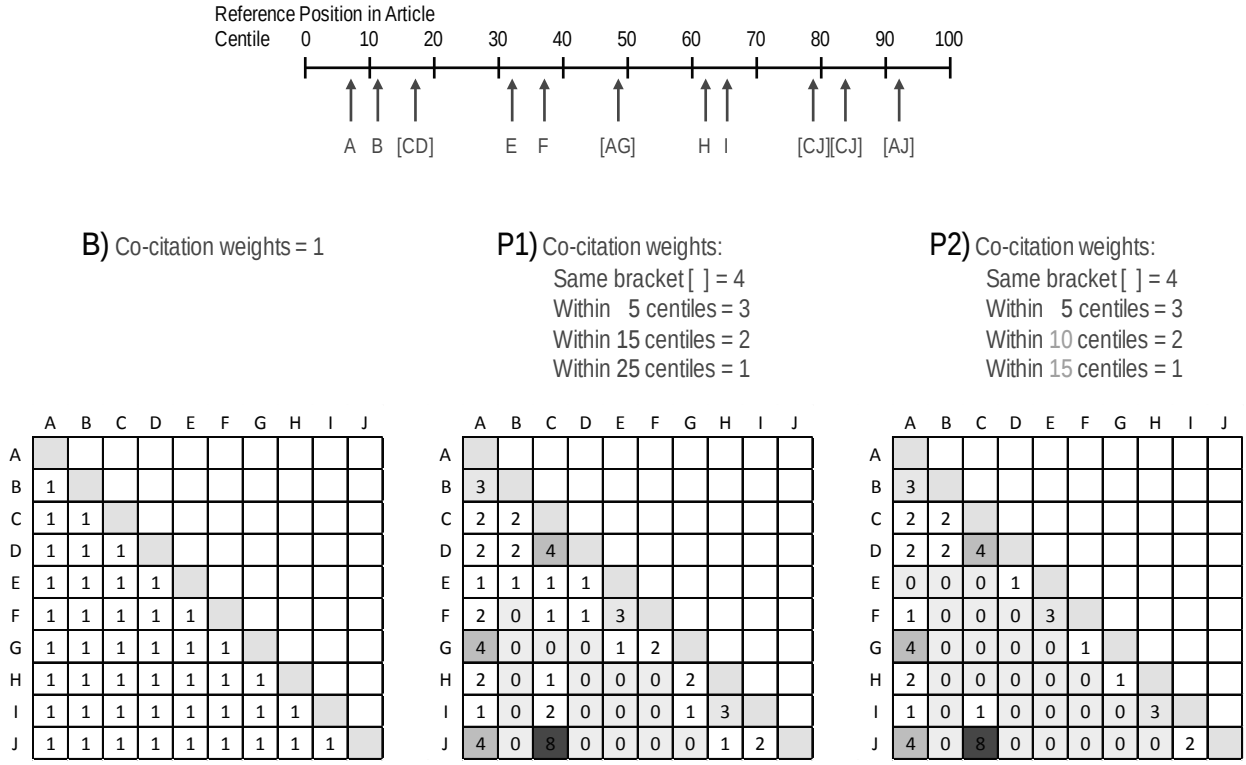

Figure 2. Examples of co-citation weighting schemes based on reference positions in full text: bibliographic (B), and two proximity approaches (P1 and P2) based on centile positions.

Co-citation clustering process

Each co-citation clustering was done using a modified version of our standard co-citation methodology (Boyack & Klavans, 2010; Klavans & Boyack, 2011) as follows:

- Citing thresholds are applied to obtain a subset of the cited references that can then be clustered. In this study we ran two sets of calculations using two different thresholds:
  - Trial 1 (T1)—References that were cited $2 \leq nc \leq 30$ times within the full text set were retained. This resulted in 1,405,800 unique references.
  - Trial 2 (T2)—References that were cited $4 \leq nc \leq 30$ times within the full text set were retained. In addition, references from 2006 or 2007 (≤1 year old) that were cited 2 or 3 times were also kept. This resulted in 435,791 unique references.

- For each trial, three different co-citation weighting schemes were applied as shown in Figure 2. For each trial and scheme, weighted co-citations were calculated on a per paper basis, and converted into modified frequencies where $f = wt/log(n*(n+1)/2)$, *wt* is the weighting factor from Figure 2, and *n* is the number of cited references for the article.
- Modified frequencies *f* were summed by co-citation pair over all citing papers. Cosine index similarity values were than calculated for each pair of citing papers as $S_{ij} = f_{ij} / sqrt(\sum(f_i)* \sum(f_j))$.
- We then filter this similarity matrix to include only the top-n S_{ij} values for each node *i*, where n varies from 5 to 15 based on the log of column sums $\sum(f_i)$.
- For each trial, scheme combination, the DrL/OpenOrd graph layout routine (Martin, Brown, Klavans & Boyack, 2011) was run using the filtered similarity as input, and using a cutting parameter of 0.975 (maximum cutting). DrL uses a random walk routine and prunes edges based on degree and edge distance; long edges between nodes of high degree are preferentially cut. At the end of the layout calculation, each article has an x,y position, and roughly 40% of the original edges remain. Articles are then assigned to clusters using an average-linkage clustering algorithm that uses article positions and uncut edges as input.

Evaluation

Each co-citation cluster solution is evaluated by calculating the textual coherence for that solution using methods we established in other large-scale studies (Boyack & Klavans, 2010; Boyack & *al*., 2011) to compare the accuracies of multiple citation-based and text-based cluster solutions. We first calculate the Jensen-Shannon divergence (JSD) for each cluster. JSD quantifies the divergence between two probability distributions—in this case between the word distribution for a document and the word distribution for the cluster in which the document resides. The JSD value for a cluster is the average of the JSD values of its constituent documents. Since JSD is cluster-size dependent, coherence is calculated as the difference between the cluster JSD and the JSD value for a random cluster of the same size. The average coherence value for the entire solution is calculated as the size-weighted average of the cluster coherence values.

Results

As mentioned above, six separate co-citation calculations were run, with two trials based on different reference thresholds, and three weighting schemes for each trial. The first trial was based on 1,405,800 references (31.3% of the available references), while the second trial was based on 435,791 (9.7%) references. Table 1 shows the distribution of co-citation weights for each trial and scheme.

For weighting scheme P1, co-cited references that were more than 25 centiles apart in the text (and given a 0 weight) comprised roughly 37% of all pairs. For weighting scheme P2, co-cited references that were more than 15 centiles apart in the text (and given a 0 weight) comprised roughly 49% of all pairs. These fractions were only minimally affected by the different thresholds used to create the reference sets T1 and T2. Co-citations in the same bracket (weights 4 and 8) were only 4.6% and 5.0% of the total available co-citations for the two trials. Co-citations with a weight of 3 (those within 5 centiles, or roughly paragraph level assuming 20 paragraphs in an article) comprised 24-25% of the available co-citations.

Table 1. Properties of the co-citation trials and weighting schemes.

	T1 – B	T1 – P1	T1 – P2	T2 – B	T2 – P1	T2 – P2
#Refs	1,405,800	1,405,800	1,405,800	435,791	435,791	435,791
#Co-cites	76,450,677	76,450,677	76,450,677	24,478,742	24,478,742	24,478,742
%Wt=0		37.14%	49.50%		36.43%	48.57%
%Wt=1	100%	12.36%	8.83%	100%	12.14%	8.71%
%Wt=2		21.88%	13.06%		21.70%	12.99%
%Wt=3		24.03%	24.03%		24.77%	24.77%
%Wt=4		3.88%	3.88%		4.22%	4.22%
%Wt=8		0.70%	0.70%		0.81%	0.81%
Top-n Sims	10,150,885	9,426,229	9,089,754	2,986,253	2,862,835	2,784,562
#Clusters	102,357	151,655	162,343	39,740	58,613	60,915
#Refs/Clust	13.7	9.3	8.7	11.0	7.4	7.2

Use of a non-uniform weighting scheme has a large effect on the top-n similarity values that are used as input to the clustering process. Of the 10,150,885 similarity pairs used for the T1-B calculation, only 52% of those pairs were present in the T1-P1 similarity file. Roughly half of them had been replaced by pairs whose similarity was enhanced by close proximity in full text. Those that remained had significantly higher similarity values in the T1-P1 set. Figure 3 shows that the similarity density functions for trial 1 (T1) calculations shift significantly to higher values as a result of using proximity weighting. The curves for T2, although not shown here, exhibit a nearly identical shift.

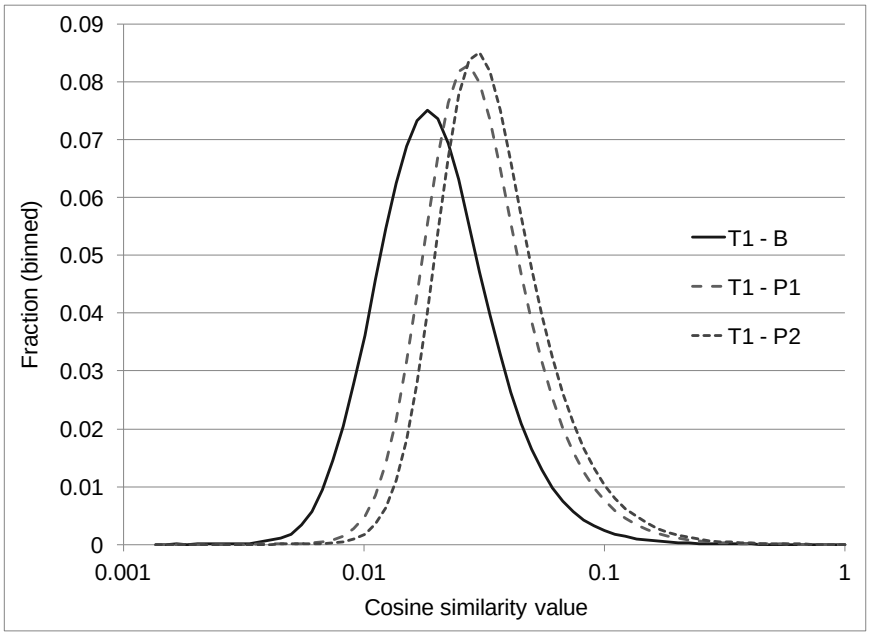

Figure 3. Cosine similarity probability density functions for trial 1 calculations.

The clustering process that we use tends to create very small clusters; these clusters can be thought of as individual research problems. Numbers of clusters and average clusters sizes are reported in Table 1, while cluster distributions are shown for each of the solutions in Figure 4.

The lack of smoothness in all six curves at a cluster size of 15 is an artefact due to lack of uniformity in the size bins used to plot the data. The largest clusters for each solution are between 100 and 200 reference papers in size. There are many clusters with only two reference papers in each solution. Comparison of the distributions shows that as the proximity criteria are tightened, the number of clusters increases and the average cluster size decreases. One possible (and we think, likely) explanation for this effect is that the addition of zeroes into the solution space effectively removes weak links and thus disaggregates clusters that would be larger if those weak links were included.

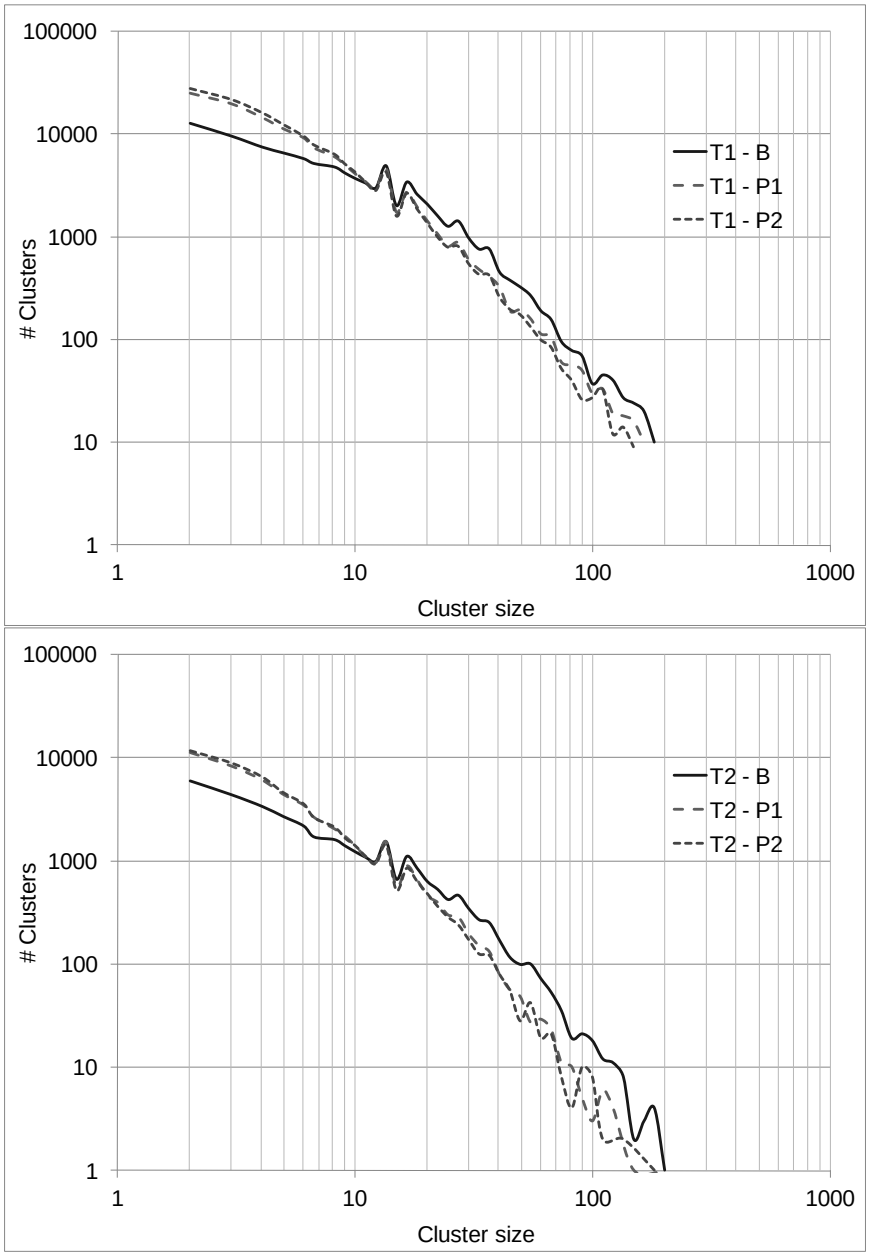

Figure 4. Cluster size distributions for the co-citation cluster solutions.

Coherence curves for each of the co-citation cluster solutions are shown in Figure 5 as a function of cluster size. For each of the two trials, incorporation of proximity information into the similarity between references significantly increased the accuracy of the solution as measured by coherence. The gain in coherence is more pronounced for large clusters than for small clusters. Table 2 shows that the proximity weighted solutions P1 and P2 had 10.6% and 11.4% greater coherence, respectively, than the unweighted bibliographic solution for Trial 1. Proximity weighted solutions had 9% higher coherence than the bibliographic solution for Trial 2 as well. If one concentrates on the larger clusters, Table 2 shows that the coherence of clusters with 10 or more reference papers is significantly enhanced (from 16-20%) by weighting of reference papers similarities by proximity.

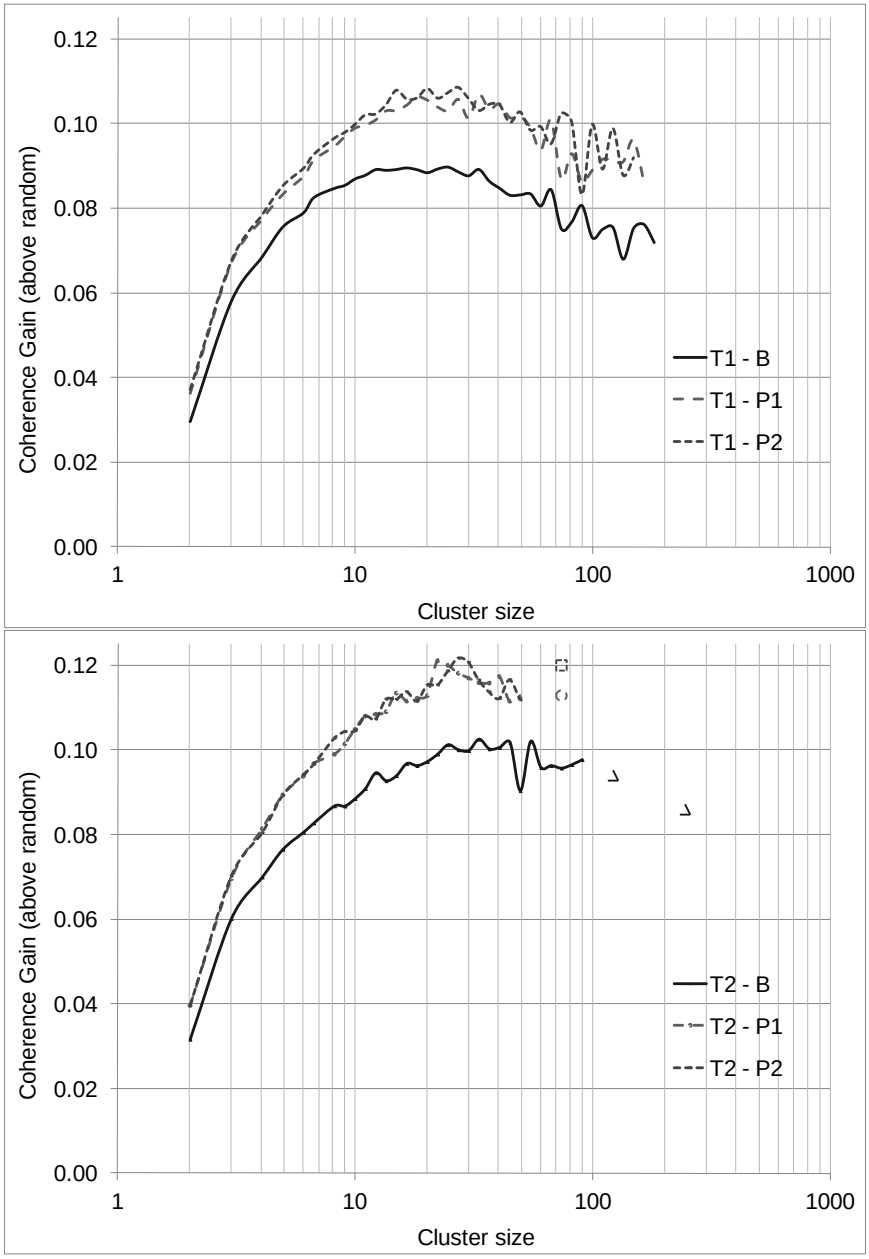

Figure 5. Coherence curves as a function of cluster size for the co-citation cluster solutions.

Table 2. Coherence values for the co-citation cluster solutions for different cluster size ranges. Percentage gain numbers in parentheses for the proximity enhanced solutions are compared to the corresponding bibliographic (B) solutions.

	T1 – B	T1 – P1	T1 – P2	T2 – B	T2 – P1	T2 – P2
Coh (all)	0.08277	0.09153 (+10.6%)	0.09219 (+11.4%)	0.08869	0.09680 (+9.1%)	0.09656 (+8.9%)
Coh (Cl≥5)	0.08565	0.09794 (+14.3%)	0.09950 (+16.2%)	0.09327	0.10604 (+13.7%)	0.10657 (+14.3%)
Coh (Cl≥10)	0.08671	0.10160 (+17.2%)	0.10373 (+19.6%)	0.09637	0.11222 (+16.4%)	0.11281 (+17.1%)

The results also show that, in general, the weighting scheme based on closer proximities, P2, had higher coherence values than the weighting scheme with slightly more distant proximities. However, this effect was not strong compared to the effect of using weighting vs. not using weighting.

Discussion

The results of this study provide strong evidence that the use of reference proximity information from full text significantly increases the accuracy of co-citation clustering. This is important because decisions in the areas of planning, evaluation, and policy can often be dependent upon an accurate understanding of the structure and dynamics of science and technology.

Although this study shows significant gains in accuracy when proximity information is incorporated in co-citation clustering, there is still much work to be done. As mentioned above, proximity can be represented in different ways. One can use sentence, paragraph, and section structures as exemplified in previous studies (Elkiss & *al*., 2008; Gipp & Beel, 2009; Liu & Chen, 2012), or one can use byte (character) counts and offsets as we have done. The effect of different weighting schemes also need to be compared. Gipp & Beel used a weighting scheme based on factors of 2 (1, ½, ¼, $^1/_8$, etc.), while we primarily used increments of one (1,2,3,4,8). There is likely an optimal weighting scheme that further investigation will uncover. Recent work showing that hybrid (text + citation) approaches can be more accurate than citation-only approaches suggests that the addition of textual features from full text might improve upon these proximity results. In addition, it is possible that use of full text information can lead to more accurate journal co-citation and author co-citation analysis as well.

References

Agarwal, S., Choubey, L. & Yu, H. (2010). Automatically Classifying the Role of Citations in Biomedical Articles. *AMIA 2010 Symposium Proceedings*, 11-15.

Bornmann, L. & Daniel, H.-D. (2008). What Do Citation Counts Measure? A Review of Studies on Citing Behavior. *Journal of Documentation*, 64 (1), 45-80.

Boyack, K.W. & Klavans, R. (2010). Co-citation Analysis, Bibliographic Coupling, and Direct Citation: Which Citation Approach Represents the Research Front Most Accurately? *Journal of the American Society for Information Science and Technology*, 61 (12), 2389-2404.

Boyack, K.W., Newman, D., Duhon, R.J., Klavans, R., Patek, M., Biberstine, J.R. & *al*. (2011). Clustering more than Two Million Biomedical Publications: Comparing the Accuracies of Nine Text-Based Similarity Approaches. *PLoS One*, 6 (3), e18029.

Callahan, A., Hockema, S. & Eysenbach, G. (2010). Contextual Cocitation: Augmenting Cocitation Analysis and its Applications. *Journal of the American Society for Information Science and Technology*, 61 (6), 1130-1143.

Elkiss, A., Shen, S., Fader, A., Erkan, G., States, D. & Radev, D. (2008). Blind Men and Elephants: What Do Citation Summaries Tell Us about a Research Article? *Journal of the American Society for Information Science and Technology*, 59 (1), 51-62.

Giles, C.L., Bollacker, K. & Lawrence, S. (1998). Citeseer: An Automatic Citation Indexing System. Paper presented at the Proceedings of the Third ACM Conference on Digital Libraries (DL '98).

Gipp, B. & Beel, J. (2009). Citation Proximity Analysis (CPA)—A New Approach for Identifying Related Work based on Co-Citation Analysis. *Proceedings of ISSI 2009*, 2, p. 571-575.

Kilicoglu, H. & Bergler, S. (2008). Recognizing Speculative Language in Biomedical Research Articles: A Linguistically Motivated Perspective. *BMC Bioinformatics*, 9 (Suppl 11), S10.

Klavans, R. & Boyack, K.W. (2011). Using Global Mapping to Create More Accurate Document-Level Maps of Research Fields. *Journal of the American Society for Information Science and Technology*, 62 (1), 1-18.

Lawrence, S., Giles, C.L. & Bollacker, K. (1999). Digital Libraries and Autonomous Citation Indexing. *IEEE Computer*, 32 (6), 67-71.

Liu, S. & Chen, C. (2012). The Proximity of Co-Citation. *Scientometrics, early online*.

Nakov, P.I., Schwartz, A.S. & Hearst, M.A. (2004). Citances: Citation Sentences for Semantic Analysis of Bioscience Text. Paper presented at the SIGIR 2004 Workshop on Search and Discover in Bioinformatics.

Ritchie, A., Teufel, S. & Robertson, S. (2008). Using Terms from Citations for IR: Some First Results. *Proceedings of the European Conference on Information Retrieval (ECIR)*, 211-221.

Small, H. (2011). Interpreting Maps of Science Using Citation Context Sentiments: A Preliminary Investigation. *Scientometrics*, 87 (2), 373-388.

Small, H. & Klavans, R. (2011). Identifying Scientific Breakthroughs by Combining Co-Citation Analysis and Citation Context. *13th International Conference of the International Society for Scientometrics and Informetrics*, 783-793.

Teufel, S. (2010). *The Structure of Scientific Articles: Applications to Citation Indexing and Summarization*. Stanford: CSLI Publications.

Teufel, S., Siddharthan, A. & Tidhar, D. (2006). Automatic Classification of Citation Function. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)*, 103-110.