Blind men and elephants: What do citation summaries tell us about a research article?

Aaron Elkiss, Siwei Shen, Anthony Fader, Güneş Erkan, David States, and Dragomir Radev

University of Michigan, Ann Arbor MI 48109

Abstract

The old Asian legend about the blind men and the elephant comes to mind when looking at how different authors of scientific papers describe a piece of related prior work. It turns out that different citations to the same paper often focus on different aspects of that paper and that neither provides a full description of its full set of contributions. In this paper we will describe our investigation of this phenomenon.

We studied *citation summaries* in the context of research papers in the biomedical domain. A citation summary is the set of citing sentences for a given article and can be used as a surrogate for the actual article in a variety of scenarios. It contains information that was deemed by peers to be important. Our study shows that citation summaries overlap to some extent with the abstracts of the papers and that they also differ from them in that they focus on different aspects of these papers than the abstracts do. In addition to this, co-cited articles (which are pairs of articles cited by another article) tend to be similar. We show results based on a lexical similarity metric called *cohesion* to justify our claims.

1 Introduction

Demand for automatic curation of scientific articles (e.g., biomedical publications) has increased recently as a result of the large volume of existing literature and the accelerating rate at which new papers are published (Cohen and Hersh, 2005). Scientific journal articles can be characterized by their dense and varied content and a large number of citations. The network of citations of these articles is an important component in automatic analysis of articles; it has been heavily studied by researchers in natural language processing, bibliometrics, complex systems, social networks, etc. (Garfield, 1955; Menczer, 2004; Newman, 2001). The text of sentences containing citations is of particular interest. Some of the initial research done on using citations to determine the content of articles was done in (Bradshaw, 2002, 2003), where the author improves search engine results with a method called Reference Directed Indexing. Recently, these *citing sentences* have been used to support automatic paraphrasing (Nakov et al., 2004) and automatic survey paper generation (Nanba et al., 2004a,b; Nanba and Okumura, 2005).

In this paper we provide a quantitative analysis of textual relationships induced by citing sentences with a view towards potential applications in summarization and information retrieval. We describe a new similarity metric, *cohesion*, and use it to analyze a corpus of biomedical journal articles from PubMed Central Open Access (PMCOA). We examine the textual relationship between the abstract of an article and the set of all sentences that cite it , also known as *citation summaries* (Figure 4) as well as the textual relationship between pairs of articles cited in the same citing sentence (Figure 5). The most salient finding is that co-citation implies textual similarity. Further, the similarity of the co-cited papers is proportional to the proximity of their citations in the citing article. For example, papers co-cited in the same sentence tended to be more similar than

papers co-cited in the same paragraph (Kessler, 1963; Small, 1973; Nanba and Okumura, 1999; Nanba et al., 2004b).

BACKGROUND: The requirement of a large amount of high-quality RNA is a major limiting factor for microarray experiments using biopsies. An average microarray experiment requires 10-100 microg of RNA. However, due to their small size, most biopsies do not yield this amount. Several different approaches for RNA amplification in vitro have been described and applied for microarray studies. In most of these, systematic analyses of the potential bias introduced by the enzymatic modifications are lacking. RESULTS: We examined the sources of error introduced by the T7 RNA polymerase based RNA amplification method through hybridisation studies on microarrays and performed statistical analysis of the parameters that need to be evaluated prior to routine laboratory use. The results demonstrate that amplification of the RNA has no systematic influence on the outcome of the microarray experiment. Although variations in differential expression between amplified and total RNA hybridisations can be observed, RNA amplification is reproducible, and there is no evidence that it introduces a large systematic bias. CONCLUSIONS: Our results underline the utility of the T7 based RNA amplification for use in microarray experiments provided that all samples under study are equally treated.

Figure 1: Abstract of PubMed article 12445333, "Optimization and evaluation of T7 based RNA linear amplification protocols for cDNA microarray analysis."

Many authors have made efforts to pinpoint the sensitive steps within this technique [10-13], optimizing the labeling, purification and variation of enzymatic and non-enzymatic components.

A more recent paper also describes an observed reduction of aRNA yield after 5 hours of amplification [11].

Although T7-based approaches for amplification of mRNA have been described [14-17], these rely on the 3' polyA tails for priming and incorporation of the T7 promoter.

This is as expected for mammalian full-length cDNA and is in agreement with previously published observations (13,17,20).

Several other groups [8,10,14,15] have applied Pearson correlation coefficients between log ratios in order to show the reproducibility of the RNA amplifications.

Figure 2: Citing sentences of PubMed article 12445333, "Optimization and evaluation of T7 based RNA linear amplification protocols for cDNA microarray analysis."

1.1   Citing Sentences and Abstracts

We define the *citing sentences* of an article A to be the collection of sentences that contain citations to A. Both the abstract and the citing sentences of an article can be considered as a kind of summary of the article. The abstract is produced by the authors of an article and conveys the central ideas of the article from the authors' perspective. In contrast, the citing sentences are a collaborative summary that indicates what other researchers found relevant, interesting or novel about the article. Thus, the citing sentences of an

article (Figure 4) can be used to produce a different kind of summary from the traditional abstract.

Bradshaw in (Bradshaw, 2003) showed that citing sentences do indeed provide many different perspectives on the same article. Recent work by Nakov and his colleagues (Nakov et al., 2004) has already shown the utility of text in articles near citations, which they neologized as "citances", which they use to automatically learn paraphrases from biomedical papers.

1.2   Example Abstract and Citing Sentences

For context we provide and analyze an example abstract and citation summary sentences for an article (Kenny et al., 2002) randomly selected from one of the 2,497 used in the study (see Section 2). This article's PubMed ID is 12392602.

The abstract and citing sentences are displayed in Figure 3 and Figure 6 respectively. Sentences in the abstract and citing sentences relate several types of information: background or context information, intermediary information such as experimental methods, and results. Virtually all abstracts contain all these types of information (Nanba et al., 2004c), but the set of citing sentences may refer to only one or two of them. Also, a citing sentence may itself be background or intermediary information in the context of its own article even though it refers to results from the article it cites. In this case, there are five citing sentences: three from (Kenny et al., 2005) (PubMed ID 15642117) and two from (Goverdhana et al., 2005) (PubMed ID 15946903). One of the citing sentences in (Kenny et al., 2005) refers to results and the other two to experimental methods (intermediary information) while both of the citing sentences in (Goverdhana et al., 2005) cite results.

It is known that researchers cite other papers for a variety of reasons. In (Nanba and

Okumura, 1999) and (Nanba et al., 2004a) the authors define three classes of citations: citations that base current work on the cited paper (type B), citations that compare current work to related papers or point out problems (type C), and citations that do not fall into either of the previous two classes (type O). The variability of citation types may help produce a more comprehensive summary by describing different aspects of the same article.

Tetracycline-regulated systems have been used to control the expression of heterologous genes in such diverse organisms as yeast, plants, flies and mice. Adaptation of this prokaryotic regulatory system avoids many of the problems inherent in other inducible systems. There have, however, been many reports of difficulties in establishing functioning stable cell lines due to the cytotoxic effects of expressing high levels of the tetracycline transactivator, tTA, from a strong viral promoter. Here we report the successful incorporation of tetracycline-mediated gene expression in a mouse mammary epithelial cell line, HC11, in which conventional approaches failed. We generated retroviruses in which tTA expression was controlled by one of three promoters: a synthetic tetracycline responsive promoter (TRE), the elongation factor 1-alpha promoter (EF1 α) or the phosphoglycerate kinase-1 promoter (PGK), and compared the resulting cell lines to one generated using a cytomegalovirus immediate early gene promoter (CMV). In contrast to cells produced using the CMV and PGK promoters, those produced using the EF1 α and TRE promoters expressed high levels of β-galactosidase in a tetracycline-dependent manner. These novel retroviral vectors performed better than the commercially available system and may have a more general utility in similarly recalcitrant cell lines.

Figure 3: Abstract of PubMed article 12392602, "Retroviral vectors for establishing tetracycline-regulated gene expression in an otherwise recalcitrant cell line."
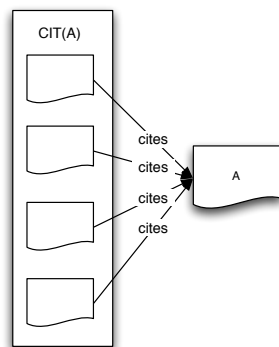

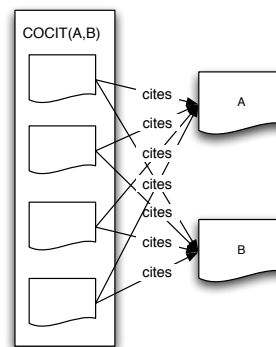Figure 4: Citation topology 1: All papers citing a given paper

Figure 5: Citation topology 2: All papers citing a given pair of papers

> Consequently, tTA expression is minimal in the presence of tetracycline and, upon tetracycline withdrawal, tTA activates its own transcription in an autoregulatory manner [24].
>
> For the tetracycline dose response curve, 5000 HC11-lacZ cells for each condition, were cultured in triplicate in 96 well plates for 72 hours, and beta-galactosidase activity was determined as previously described [24].
>
> These cell lines were established using a novel autoregulatory system in which the expression level of the tetracycline transactivator (tTA) protein is minimised during routine culture and is induced upon withdrawal of tetracycline with concomitant upregulation of the transgene-of-interest [24].
>
> Studies by Kenny and co-workers [191] established successful Tet-OFF-based regulation from retroviral vectors and demonstrated the effectiveness of the TRE promoter in achieving stringent regulation of gene expression [191].
>
> Upon evaluation of different promoters to drive tTA expression, such as CMV, elongation factor 1α, and phosphoglycerate kinase-1, in combination with the TRE they observed that only the CMV promoter in combination with the TRE promoter produced successful regulatable β-galactosidase expression when controlled by the Tet-OFF regulatory switch in HC11 mouse mammary epithelial cell lines [191].

Figure 6: Citing sentences of PubMed article 12392602, "Retroviral vectors for establishing tetracycline-regulated gene expression in an otherwise recalcitrant cell line."

1.3   Co-citations

The relationship between abstracts and citing sentences is not the only relevant data that we can extract from citing sentences. Another interesting feature of citing sentences is that of *co-citation*. A citing sentence can contain references to two or more other articles; these articles are said to be co-cited by the citing sentence (Figure 5). Co-citations can occur at various granularities: sentence level, paragraph level, section level, and article level. Articles can be co-cited by only one paper or by many different papers; this may also have a relationship to the similarity of the co-cited papers.

1.4   Research Hypotheses

Intuitively, citing sentences should be a valuable source for mining the knowledge in the cited publications. If the information contained in the citing sentences is more focused than the information contained in the abstract, extractive summaries (summaries containing the most salient sentences from the article) based on the citing sentences may be useful. They would provide a more concise summary of the abstract and contain specif-

ically the information from the article that others found useful. In addition, examining co-cited articles might provide a fast and useful way to find articles similar to one under consideration.

More specifically, we attempt to confirm the following hypotheses:

- The citation summary of an article is similar to that article's abstract.

- Citing sentences contain more focused information than the abstract.

- The amount (or diversity) of information contained in the citing sentences converges as the number of citing sentences grows.

- Co-citation is highly correlated with textual similarity; as the focus of co-citation (sentence, paragraph, section or article) becomes smaller and as the number of co-citations increases, textual similarity will increase.

1.5   Comparing Texts: Cohesion

To test these hypotheses we will need a quantifiable notion of focused information. If sentences in a text tend to be similar to each other, then the information is focused because textually similar sentences are likely to be on the same topic. In particular, textually similar citing sentences probably cite the same aspect of the cited article. In addition to this, we would like to quantify the similarity between two texts (abstracts and citing sentences) in such a way that the self-similarity of a text can be compared to its similarity with another text. In other words, we would like to compare the self-similarity of the abstract to its similarity to the set of citing sentences. Both quantities will be based on a sentence-oriented version of the standard *tf·idf* approach (Salton and Buckley, 1988). For comparing co-cited texts, normal document-oriented *tf·idf* suffices.

To calculate the similarity between a pair of texts we use the average weighted cosine similarity over all sentence pairs. Each sentence is represented as a vector $S \in \mathbb{R}^{\infty}$ where each nonzero element of S is a weighted count of the number of occurrences of a word in the sentence. Each element is defined as $-\log(\frac{s(w)}{|\mathbf{S}|})c_S(w)$ where $s(w)$ is the number of sentences containing word w, $|\mathbf{S}|$ is the total number of sentences in the corpus and $c_S(w)$ is the number of times word w appears in sentence S. Thus, less frequently occurring terms are more heavily weighted, since they carry more information, while words that occur in nearly every sentence (e.g. "the", "of", etc.) have a very small weight since they carry little information. The *cross cohesion* between two distinct texts comprising sentences $S_1 \ldots S_m$ and $T_1 \ldots T_n$ is calculated as

$$C_c(S,T) = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{\langle S_i, T_j \rangle}{||S_i|| \; ||T_j||}$$

where $\langle \cdot, \cdot \rangle$ is the Euclidean dot product and $||S||$ is the magnitude of vector $|S|$ using the Euclidean norm. $\frac{\langle S_i, T_j \rangle}{||S_i|| \; ||T_j||}$ is the cosine of the smaller angle between the two vectors. Its value lies between 0 and 1, where 0 indicates the two sentences have no words in common and 1 indicates they have exactly the same words, although possibly in a different order.

We can also compute the *self cohesion* of a single text $S_1 \ldots S_n$ by comparing each sentence with each other sentence:

$$C_s(S) = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} \frac{\langle S_i, S_j \rangle}{||S_i|| \; ||S_j||}$$

This avoids comparing each sentence to itself (such comparisons always have a cosine of 1) as well as redundant comparisons (since the inner product is symmetric). C_s reflects the homogeneity of information in a piece of text. For example, the abstract displayed in Figure 7 has a relatively high self cohesion of 0.32 while the abstract displayed in Figure 8 has a much lower self cohesion of 0.013. We see the first abstract mentioned 'SCit' in all 4 sentences and 'side chain conformations' in 3 of the sentences, while the

second abstract has much less lexical overlap between sentences - other than 'PDB' in two sentences, only extremely common words ('is','the','and') are shared.

> SCit is a web server providing services for protein side chain conformation analysis and side chain positioning. Specific services use the dependence of the side chain conformations on the local backbone conformation, which is described using a structural alphabet that describes the conformation of fragments of four-residue length in a limited library of structural prototypes. Based on this concept, SCit uses sets of rotameric conformations dependent on the local backbone conformation of each protein for side chain positioning and the identification of side chains with unlikely conformations. The SCit web server is accessible at http://bioserv.rpbs.jussieu.fr/SCit.

Figure 7: Abstract of (Gautier et al., 2004); $C_s = 0.32$

> The Protein Data Bank (PDB; http://www.pdb.org) is the primary source of information on the 3D structure of biological macromolecules. The PDB's mandate is to disseminate this information in the most usable form and as widely as possible. The current query and distribution system is described and an alpha version of the future re-engineered system introduced.

Figure 8: Abstract of (Bourne et al., 2004); $C_s = 0.013$

It is important to note that high C_c is not in fact a necessary condition for information similarity (or information homogeneity in the case of C_s). If two sentences have a large cosine similarity there is a large degree of lexical overlap among information-bearing words in its constituent sentences and therefore a high chance of information overlap, but the inverse is not true: lack of lexical overlap does not necessarily imply two sentences do not carry the same information. As an example, "Androgen receptor was found to bind to RAN in *Homo sapiens*" contains all the information that "Human DHTR interacts with RASL2-8" contains (DHTR is an abbreviation for dihydrotestosterone receptor, dihydrotestosterone is a synonym for androgen; likewise RAN and RASL2-8 refer to the same protein), but have zero cosine since they share no lexical tokens. This problem occurs frequently in biomedical literature in particular because many concepts such as genes or proteins have multiple aliases that depend on context. Ontologies, synonym dictionaries, etc., would help with this problem, but introduce a large degree of additional

X	Y	$\rho_{\mathbf{X},\mathbf{Y}}$
$\cos_s(ABS(A), ABS(B))$	$\cos_s(BODY(A), BODY(B))$	0.723
$\cos_d(ABS(A), ABS(B))$	$\cos_d(BODY(A), BODY(B))$	0.810
$\cos_d(ABS(A), ABS(B))$	$\cos_s(ABS(A), ABS(B))$	0.922
$\cos_d(BODY(A), BODY(B))$	$\cos_s(BODY(A), BODY(B))$	0.810
$\cos_s(ABS(A), ABS(B))$	$\cos_{s_s}(ABS(A), ABS(B))$	0.898
$\cos_d(ABS(A), ABS(B))$	$\cos_{s_s}(ABS(A), ABS(B))$	0.877
$\cos_{d_s}(ABS(A), ABS(B))$	$\cos_{s_s}(ABS(A), ABS(B))$	0.874
$\cos_d(ABS(A), ABS(B))$	$\cos_{d_s}(ABS(A), ABS(B))$	0.832
$\cos_{d_s}(ABS(A), ABS(B))$	$\cos_s(ABS(A), ABS(B))$	0.688
$C_{c_d}(ABS(A), ABS(B))$	$\cos_d(ABS(A), ABS(B))$	0.866
$C_{c_{d_s}}(ABS(A), ABS(B))$	$\cos_d(ABS(A), ABS(B))$	0.860
$C_{c_s}(ABS(A), ABS(B))$	$\cos_s(ABS(A), ABS(B))$	0.857
$C_{c_{s_s}}(ABS(A), ABS(B))$	$\cos_s(ABS(A), ABS(B))$	0.846

Table 1: Correlation $\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$ for various related measures of textual similarity. *cos* and cross cohesion C_c are parameterized on the vector space representation used: d for the usual *tf·idf*, s for *tf* times inverse sentence frequency, and d_s and s_s for the corresponding versions with each token stemmed with Porter's stemmer.

complexity to a simple lexical measure. The author of (Bradshaw, 2002) showed that this property is actually useful in a Reference Directed Indexing system, where a diverse group of keywords provides a larger target for searchers.

To validate cohesion as a measure of textual similarity, we compared it with various other cosine-based measures of textual similarity on the experimental corpus (see Section 2). Table 1 shows how the cohesion metric correlates with raw cosine similarity, how cosines computed with "inverse sentence frequency" correlate with cosines computed with "inverse document frequency", and how cosine similarity of the abstract correlates with cosine similarity of the body. We also compare *tf·idf* using stemming with Porter's algorithm (Porter, 1997) to unstemmed versions. Correlations were computed using the cosines between all co-cited papers in the corpus (see Section 2) as a sample. The correlation is generally quite strong, as would be expected. The correlation between stemmed and unstemmed versions give some credence to the supposition that simple lexical similarity is a good proxy for underlying semantic similarity.

1.6   Experiments

Having defined self-cohesion and cross cohesion, we can quantify the notions suggested in the hypotheses. To determine the similarity between abstract and citing sentences we compute two quantities. The first is the cross cohesion between citing sentences and the corresponding abstract, $C_c(CIT(A), ABS(A))$. The second is the information composition of the abstract and citing sentences – that is, where in the article the information in the abstract and information in the citing sentences come from. To determine this we split the article into paragraphs $P_1, \ldots, P_m$ and sections $S_1, \ldots, S_n$ and compute the similarity between each section and paragraph and the citing sentences and abstract. We expect a correlation between the information-source distributions for abstracts and citing sentences.

To test whether citing sentences in general contain more focused information than the abstract, we check if the self cohesion of the citing sentences $C_s(CIT(A))$ exceeds the self cohesion of the abstract $C_s(ABS(A))$.

To test whether articles having many citations are likely to be cited for a few things repeatedly, we compute the self-cohesion of citing sentences $C_s(CIT(A))$ and see if there is a correlation with respect to the number of citing sentences $|A|$.

To test the relationship between co-citation and textual similarity, we computed the number of times each pair of citations were cited together in the same article, at the sentence, paragraph, section, or complete article level. In addition, we counted the number of distinct papers citing each pair at the article, section, paragraph or sentence level. These values were compared to the cosine similarity between the bodies and abstracts of the two articles. This comparison uses the traditional document-oriented *tf·idf* rather than the sentence-oriented cohesion.

2   Experimental Data

Given the proposed experiments, we require a corpus of articles to analyze. Since we are interested in information retrieval and summarization particularly in a biomedical context, we use biomedical journal articles for our experiments.

2.1   Data Collection

The primary set of articles we analyzed comes from the free PubMed Central (PMC) repository at `http://www.pubmedcentral.gov`. We downloaded all 13,520 open access articles available as of October 1, 2005 from the NCBI FTP site at `ftp://ftp.ncbi.nlm.nih.gov/pub/pmc`. Of these, 2,497 were cited by at least one other paper in PubMed Central. In addition we retrieved all papers in PubMed Central citing the open access subset and extracted the citing sentences. Figure 10 gives the distribution of number of citations per article.

2.2   Data Preprocessing

The articles come in an XML format with references and citations marked so that they can be extracted unambiguously. We also extracted the abstract and body from each article and segmented into sentences using Adwait Ratnaparkhi's MXTERMINATOR, a maximum entropy based sentence-boundary recognition tool (Reynar and Ratnaparkhi, 1997). Rather than using the default model, trained on Wall Street Journal news articles, we used a biomedical article specific model trained on 50 randomly selected articles comprising approximately 100,000 words.

2.3   Data Statistics

We present some statistics of the 2,497 articles investigated. The articles came from most of the open access journals available through PubMed Central. Table 2 lists those journals with more than 20 articles cited by other articles in PubMed Central. Over 1,000 of the articles are from the BioMed Central (BMC) family of journals. Genome Biology had 307 articles; Nucleic Acids Research, Breast Cancer Research and Critical Care each had over 100. The complete list of journals with open access articles is available at `http://www.pubmedcentral.nih.gov/about/openftlist.html`.

Figure 9 shows the distribution of the number of sentences in the abstract, which is approximately normal with $\bar{x} = 9.67$ and $s = 4.82$. The distribution of the number of retrieved citing sentences is quite close to a power law (see Figure 10) with $k = -1.9663$ and $r^2 = 0.958$. The distribution of the number of papers cited by each of the 13,520 open access articles is approximately normal with $\bar{x} = 41.66$ and $s = 37.31$. The distribution of the number of retrieved citing sentences is quite close to a power law (see Figure 10) with $k = -1.9663$ and $r^2 = 0.958$. The distribution of the number of papers cited by each of the 13,520 open access articles is also approximately normal with $\bar{x} = 41.66$ and $s = 37.31$.

2.4   Data Analysis

Given an article A, we retrieve its abstract, $ABS(A)$, and a set of sentences from other PubMed Central papers that cite article A, $CIT(A)$. Because PubMed Central does not contain *all* biomedical journal articles, $CIT(A)$ is unlikely to contain every citation for A. However, we do assume that it contains a representative subset. Additionally, we compute all pairs of co-cited papers (A, B) such that there exists some C such that C

No. of Articles	Journal Title
379	Nucleic Acids Research
361	Genome Biology
204	BMC Bioinformatics
193	Critical Care
170	Breast Cancer Research
169	PLoS Biology
119	BMC Genomics
115	Arthritis Research
77	BMC Microbiology
68	Arthritis Research & Therapy
67	Health and Quality of Life Outcomes
65	PLoS Medicine
62	Respiratory Research
56	Reproductive Biology and Endocrinology
53	BMC Cancer
51	BMC Evolutionary Biology
48	BMC Cell Biology
48	BMC Infectious Diseases
47	BMC Public Health
43	Malaria Journal
42	Evidence-based Complementary and Alternative Medicine
40	BMC Neuroscience
36	BMC Molecular Biology
35	Molecular Cancer
35	BMC Biotechnology
34	BMC Genetics
33	BMC Medical Research Methodology
31	BMC Biochemistry
30	Retrovirology
28	Journal of Biology
25	Current Controlled Trials in Cardiovascular Medicine
23	International Journal of Health Geographics
23	BMC Health Services Research
23	Journal of Biomedicine and Biotechnology
22	BMC Developmental Biology
22	Journal of Translational MEDICINE

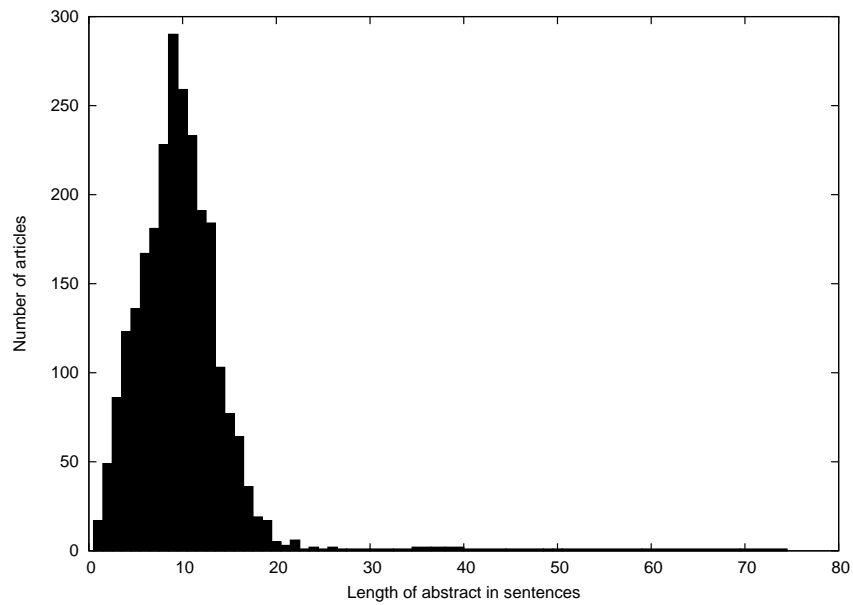Table 2: Journals with more than 20 open access articles cited by other articles in PubMed Central

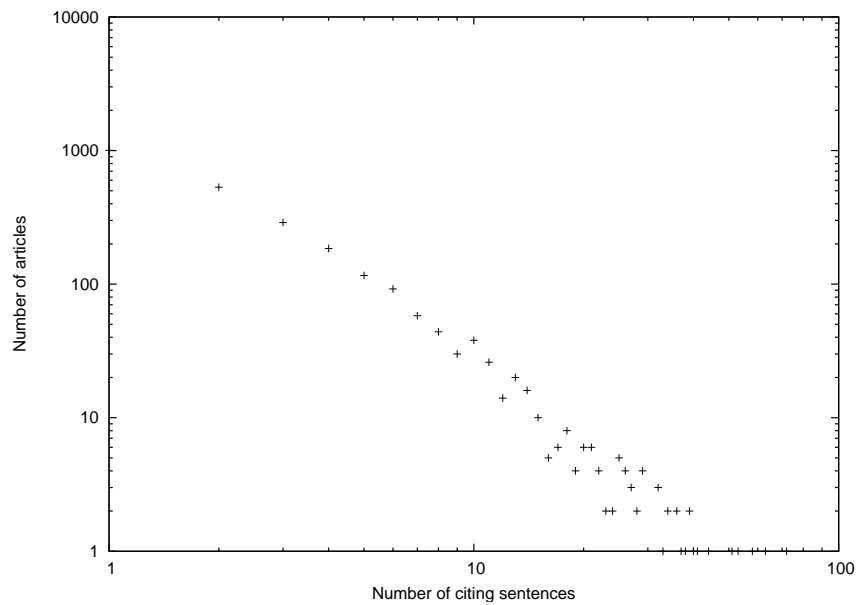Figure 9: Distribution of the number of sentences in the article abstracts.

Figure 10: Distribution of the number of retrieved citing sentences of the articles, on a log-log scale. The regression is $\log(y) = -1.9663\log(x) + 3.3173$; $r^2 = 0.958$

cites A and B.

3   Experiments and Results

Experiments testing the hypotheses from section 1.4 are described in this section along with results for the 2,497 articles from PubMed Central.

3.1   Abstracts vs. Citing Sentences

Consider the following text pairs for which we compute cohesion scores.

- $C_s(CIT(A))$: Do all the citing sentences of A cite A for the same reason, or do different papers cite different aspects of A?

- $C_s(ABS(A))$: Is the abstract tightly focused or does it give a broader overview of various aspects of A?

- $C_c(CIT(A), ABS(A))$: Is all the information in the citing sentences contained in the abstract, or is there some divergence? How much information is shared between the two?

- $C_c(CIT(A), CIT(B))$, $C_c(ABS(A), CIT(B))$ where B is a randomly chosen article not identical to A: We need negative controls to ensure that the above cohesions are higher than expected by chance.

Our goal is to show the cross cohesion between $ABS(A)$ and $CIT(A)$ is lower than the two self cohesion scores but significantly higher than the two cross cohesion scores involving randomly chosen $CIT(B)$. This might suggest that there is information contained in the citing sentences of A that is not contained in the abstract of A. It follows

	n	$\bar{x}$	s
$C_s(CIT(A))$	1527	0.1321	0.1131
$C_s(ABS(A))$	2480	0.1176	0.0579
$C_c(ABS(A), CIT(A))$	2497	0.0820	0.0545
$C_c(CIT(A), CIT(B))$	2497	0.0110	0.0104
$C_c(ABS(A), CIT(B))$	2497	0.0090	0.0068

Table 3: Statistics of cohesion between various texts.

from this that citing sentences might be a potentially useful resource for summarization and information retrieval, giving an alternative view of the salient parts of the cited article.

Table 3 lists the sample mean and standard deviation of the various cohesion scores. The average $C_c(ABS(A), CIT(A))$ is less than the average $C_s(ABS(A))$ and $C_s(CIT(A))$ and greater than the two negative controls. C_s is only defined on sets of more than one sentence and was computed on only 2,480 abstracts and 1,527 sets of citing sentences. The difference between $C_s(ABS(A))$ and $C_s(CIT(A))$ is significant at the 99.9999% confidence level as reported by a paired t-test computed for the 1,521 articles for which both self cohesion scores were defined. All the other pairwise differences are that significant or even more so. Figures 11-15 display the distribution of each cohesion score. $C_s(CIT(A))$ is roughly normal but less tightly peaked than $C_s(ABS(A))$ although the means are close; $C_c(CIT(A), ABS(A))$ has a significantly lower mean, and the negative control C_c values are very low.

Results from this experiment confirm that the $C_s(CIT(A))$ is consistently higher than the self cohesion of $C_s(ABS(A))$ for the same A. That is, the contents of citing sentences exhibit a greater uniformity than the contents of the corresponding abstract. This confirms the common sense notion that the abstract serves as a synopsis of the entire article while citations of the article focus on notable aspects of what is presented in the paper.
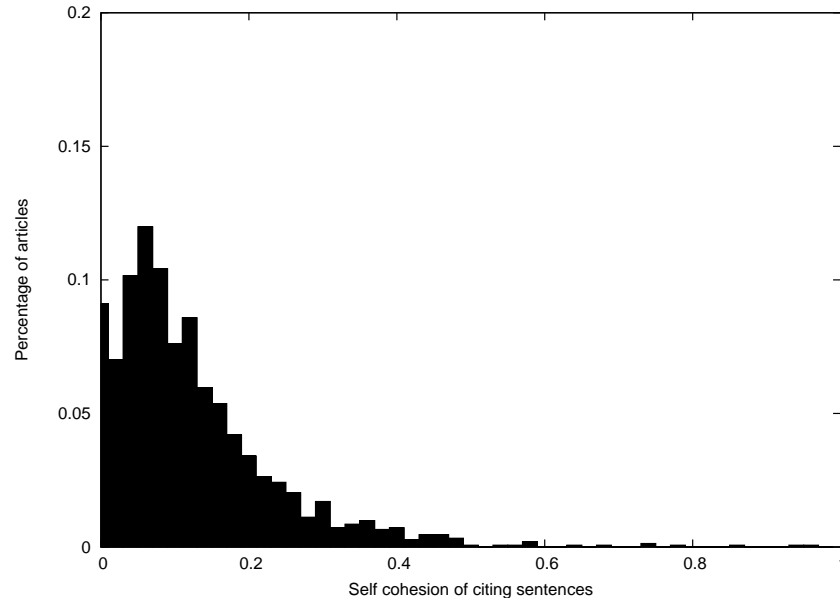
Figure 11: Distribution of self cohesion of $CIT(A)$ - $n = 1527$, $\bar{x} = 0.1321$, $s = 0.1131$
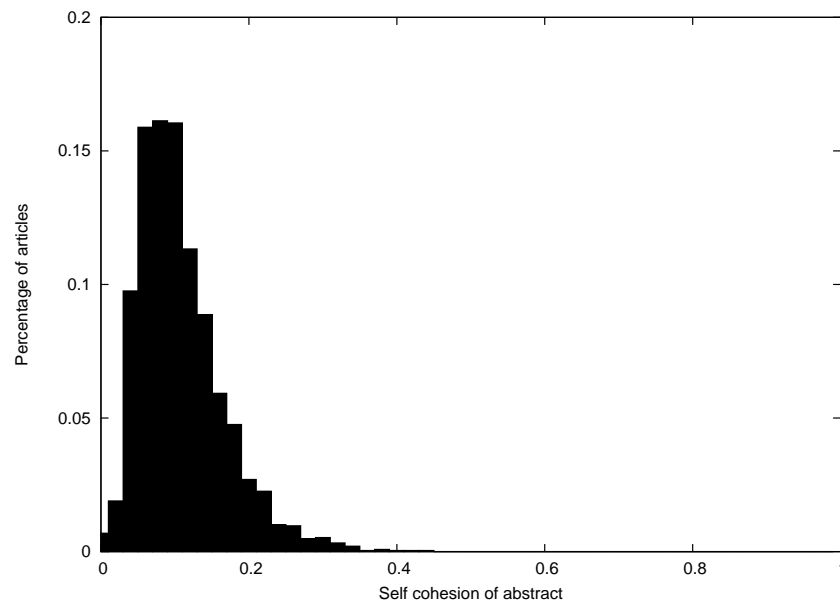

Figure 12: Distribution of self cohesion of $ABS(A)$ - $n = 2480$, $\bar{x} = 0.1176$, $s = 0.0579$
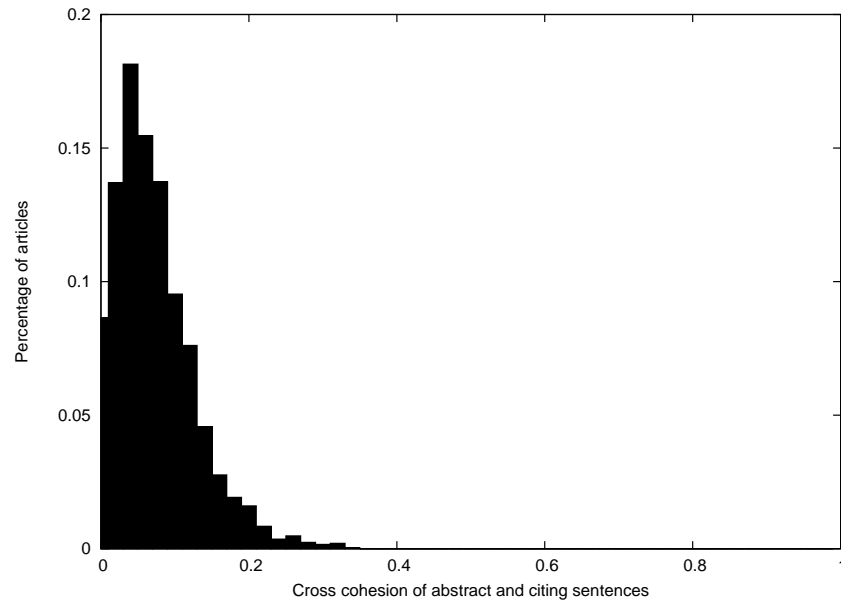
Figure 13: Distribution of cross cohesion of $CIT(A) \times ABS(A)$ - $n = 2497$, $\bar{x} = 0.0820$, $s = 0.0545$
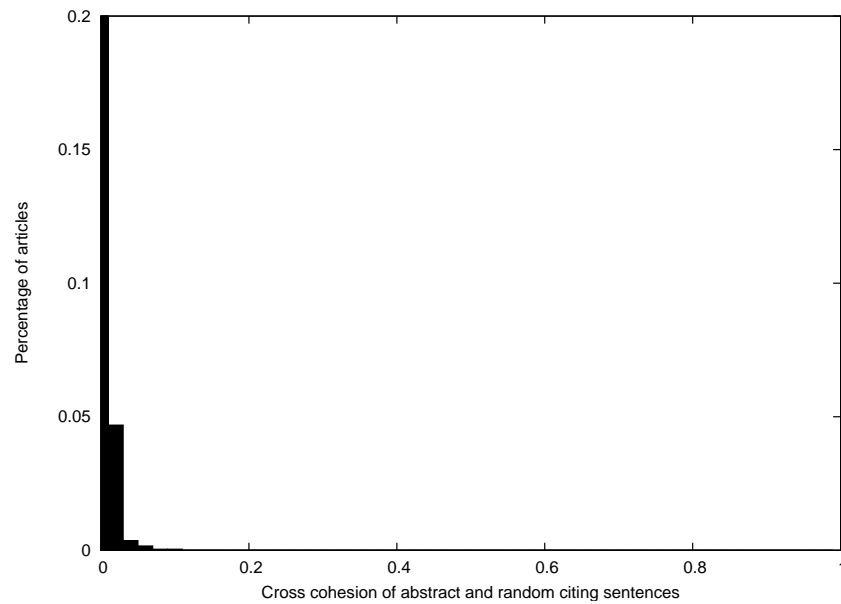

Figure 14: Distribution of cross cohesion of $ABS(A) \times CIT(B)$ - $n = 2497$, $\bar{x} = 0.0090$, $s = 0.0068$
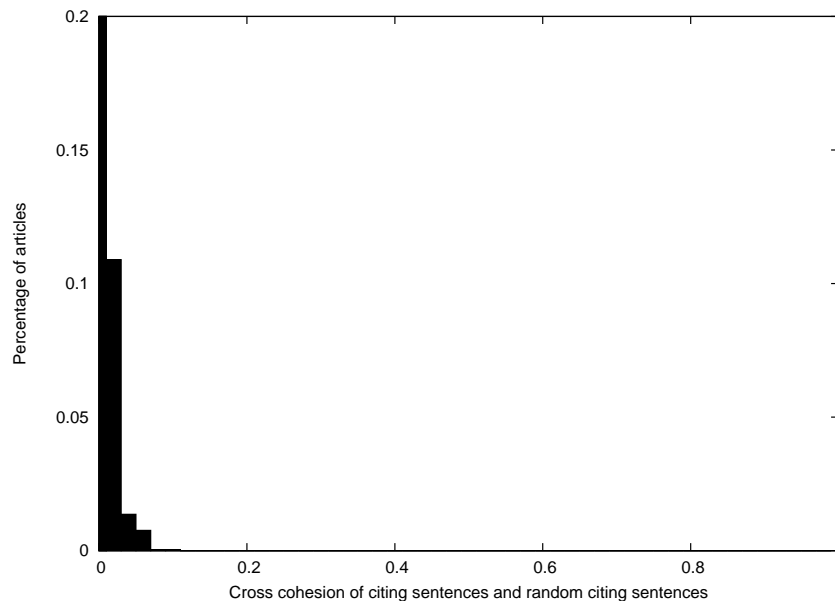
Figure 15: Distribution of cross cohesion of $CIT(A) \times CIT(B)$ - $n = 2497$, $\bar{x} = 0.0110$, $s = 0.0104$

> Heritable, but reversible, changes in transposable element activity were first observed in maize by Barbara McClintock in the 1950s. More recently, transposon silencing has been associated with DNA methylation, histone H3 lysine-9 methylation (H3mK9), and RNA interference (RNAi). Using a genetic approach, we have investigated the role of these modifications in the epigenetic regulation and inheritance of six Arabidopsis transposons. Silencing of most of the transposons is relieved in DNA methyltransferase (met1), chromatin remodeling ATPase (ddm1), and histone modification (sil1) mutants. In contrast, only a small subset of the transposons require the H3mK9 methyltransferase KRYPTONITE, the RNAi gene ARGONAUTE1, and the CXG methyltransferase CHROMOMETHYLASE3. In crosses to wild-type plants, epigenetic inheritance of active transposons varied from mutant to mutant, indicating these genes differ in their ability to silence transposons. According to their pattern of transposon regulation, the mutants can be divided into two groups, which suggests that there are distinct, but interacting, complexes or pathways involved in transposon silencing. Furthermore, different transposons tend to be susceptible to different forms of epigenetic regulation.

> As noted by McClintock, the rate of transposition in a given germ line can change over time, with cycles of activation and silencing lasting several generations (cited in reference 35).
> In Arabidopsis and maize, a low level of transcription is maintained despite the production of small interfering RNAs (35, 46).

Figure 16: The abstract (top) and citing sentences (bottom) of PMC article 300680. $C_c = 0.0408$

	n	$\bar{x}$	s
$C_s(ABS(A))$	66	0.1287	0.0856
$C_s(CIT(A))$	57	0.0894	0.0790
$C_c(ABS(A), CIT(A))$	66	0.0708	0.0553
$C_c(CIT(A), CIT(B))$	66	0.0209	0.0178
$C_c(ABS(A), CIT(B))$	66	0.0176	0.0214

Table 4: Statistics of cohesion between various texts for the small collection of 66 articles from the ACM Digital Library.

3.1.1   ACM Digital Library

For additional validation, we also examined a small collection of 66 articles from the 11th and 12th international conference on the WWW, obtained from the ACM Digital Library (`http://portal.acm.org/dl.cfm`) to see if the trends observed for the biomedical articles held for another domain. Figure 17 shows the distribution of abstract length with $\bar{x} = 7.38$ and $s = 2.85$. The collection contained 305 citations. The results of the experiment are summarized in Table 4. Here the self cohesion of the abstracts is significantly higher than the self cohesion of the citations. One explanation for this might be that in this domain each citing article contains fewer citing sentences than in the biomedical domain, and citing sentences in different articles tend to be less similar than citing sentences in the same article. $C_s(ABS(A))$ and $C_s(CIT(A))$ are still both greater than $C_c(ABS(A), CIT(A))$, which is again greater than the negative controls. Thus we can conclude that the abstracts and citing sentences are still significantly similar in this domain.

3.2   Information Source

Another experiment we carried out was to analyze $ABS(A)$ and $CIT(A)$ to locate where in the original article information came from. This is another way to test how the information content of $ABS(A)$ differs from the information content of $CIT(A)$. A very strong correlation would imply that the abstract and citing sentences both focus on
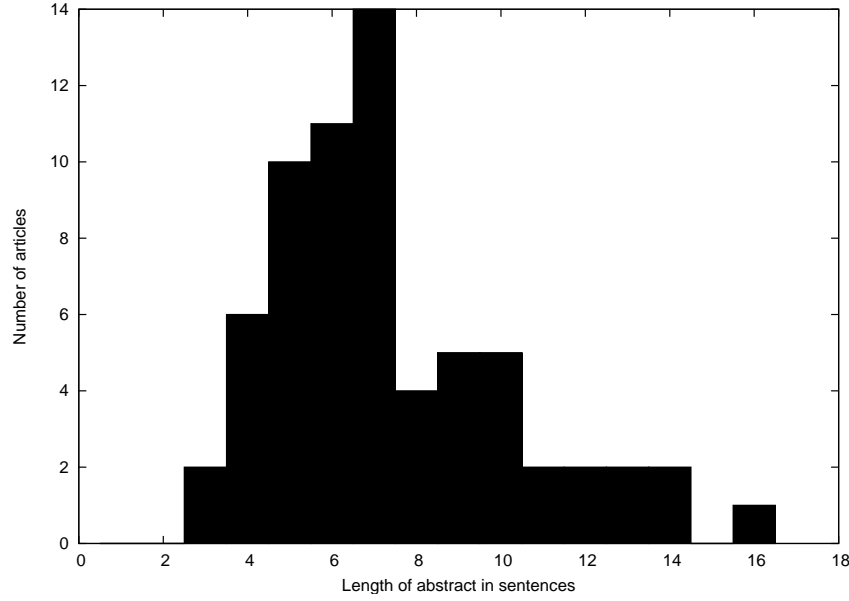
Figure 17: Distribution of the number of sentences in the article abstracts for the ACM Digital Library articles.

information from the same parts of the article, while a lower correlation implies they have differing focuses.

An article A can be divided into paragraphs $P_1, \ldots, P_m$ or sections $S_1, \ldots, S_n$ (where a section is a natural division of the article made by the author). We can then look at the cross cohesion of each piece of the article (either P_i or S_i) with $ABS(A)$ and $CIT(A)$. If the abstract and citing sentences highlight information from the same part of the article, we would expect to see a correlation between $C_c(P_i, ABS(A))$ and $C_c(P_i, CIT(A)$ (and similarly for S_i in place of P_i). To test this, we collected two sets, each containing pairs of cross cohesion values:

$X_{par} = \{(C_c(P_i, ABS(A)), C_c(P_i, CIT(A))) \; : \; \text{for each paragraph } P_i \text{ in each article } A\}$

$X_{sec} = \{(C_c(S_i, ABS(A)), C_c(S_i, CIT(A))) \; : \; \text{for each section } S_i \text{ in each article } A\}.$

Table 5 gives some information about these quantities. We then computed the correlation

	n	$\bar{x}$	s
$C_c(P_i, ABS(A))$	116794	0.0575	0.0588
$C_c(P_i, CIT(A))$	116794	0.0399	0.0540
$C_c(S_i, ABS(A))$	31047	0.0624	0.0482
$C_c(S_i, CIT(A))$	31047	0.0438	0.0427

Table 5: Statistics for the cross cohesion of paragraphs and sections with abstracts and citing sentences.

coefficients $\rho_{par} = 0.565$ for X_{par} and $\rho_{sec} = 0.564$ for X_{sec}. These values suggest that the citing sentences and abstract tend to be similar to the same parts of the article, but there are also regions of the article that are similar to just the abstract or just the citing sentences.

3.3 Self Cohesion of Citing Sentences

We would also like to see if information content converges as the number of citing sentences increases. Information convergence would imply that as the number of citations for an article grows, most citing sentences tend to cite the article for a very small number of reasons. If the citing sentences behave in this fashion, they could provide very tightly focused summaries of an article. In the context of search engines, this supports the work of (Bradshaw, 2002, 2003), showing that terms contained frequently in citations can be used to boost the relevance of an article. To test whether information in the citing sentences converges as the number of citing sentences grows, we compute the self cohesion of $CIT(A)$ for each article A and observe the correlation with the number of citing sentences.

However, as seen in Figure 18 the average self cohesion actually decreases somewhat as the number of citing sentences increases up to 20. Beyond 20 citations there are only a few articles with each number of citations (see figure 10), thus discerning any trend is difficult. Therefore we cannot use the self cohesion of the citations to confirm the hypothesis that information in the citing sentences converges as the number of citing sentences increases.

As mentioned previously, low self cohesion does not necessarily imply lack of information convergence, hence we cannot conclusively reject the hypothesis either.

One explanation for the trend seen could be that up to a point, the number of things papers are cited for increases, therefore decreasing the self cohesion, but beyond that point papers tend to be cited for the same few things repeatedly. This seems to be the case for the articles in the study with more than 50 citing sentences (the citing sentences are not reproduced here for space constraints; the relevant PubMed IDs are 12144710, 12537568, 12182760, 11734060, 12537572), but since only a few papers are cited frequently a much larger collection of articles than the 2,497 used in the study would be needed to examine this more rigorously. Another likely reason cohesion might decrease is synonymy and related phenomena, e.g. "flies" vs. "*Drosophila melanogaster*"; different authors can paraphrase articles in different ways and might not use exactly the same lexical items. This type of behavior was described by Bradshaw in his dissertation (Bradshaw, 2002).
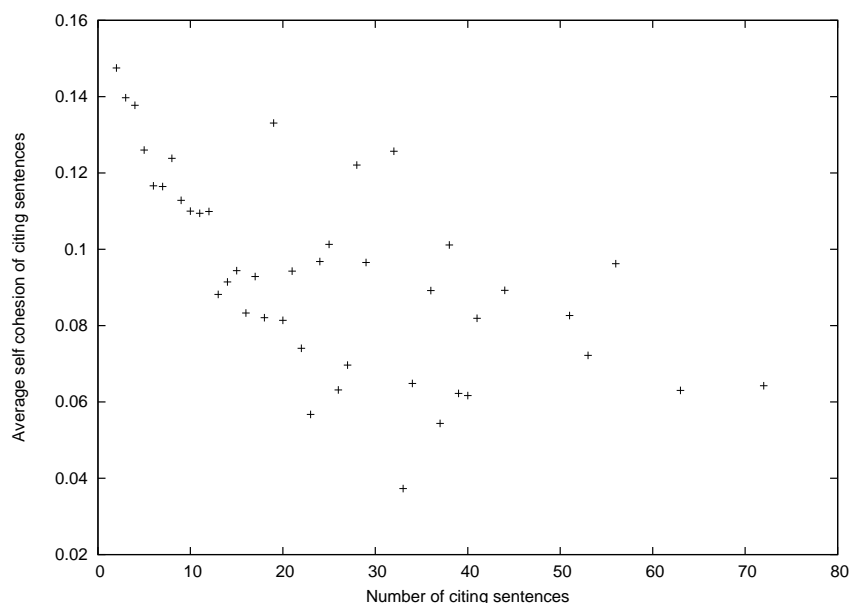

Figure 18: Average self cohesion of CIT(A) by number of citing sentences.

3.4 Co-citations

Citing sentences contain additional information beyond their plain text content – each citing sentence can potentially cite many different articles. Obviously, co-citation is generalizable to granularities larger than a single sentence. Natural divisions are sentence, paragraph, section, and article. We computed the cosine similarity between each pair of co-cited papers as well as the cosine similarity between an equal number of randomly chosen non-co-cited papers. There is a very large and significant difference ($\rho < 0.001$) in the average cosine similarity for papers that are not co-cited compared to those that are co-cited at any granularity (Figure 21). In addition, there is a modest but often significant increase in average similarity when papers are co-cited additional times (Figures 19 and 20). In particular the difference between 1 and 2 co-citations is significant at or below the $\rho = 0.001$ level for each co-citation granularity. There is also a large and significant ($\rho < 0.001$) difference in similarity between papers co-cited only in the same paper vs. papers co-cited at smaller granularities (Figure 21).

However, there is no strong correlation in general between number of co-citations and similarity either between abstracts (Figures 19) or body text (Figure 20). More than one sentence, paragraph or section can co-cite a pair of articles, and one can ask if limiting the count of co-citations at these granularities to distinct articles produces a correlation; however, it does not appear to make a significant difference. The fact that articles are co-cited is a strong indication of similarity, but additional co-citations do not imply the articles are more similar. This finding implies that the ability to navigate from a given article directly to co-cited articles would be a useful way to find related work, but ranking co-cited articles by the number of co-citations would probably not be useful.
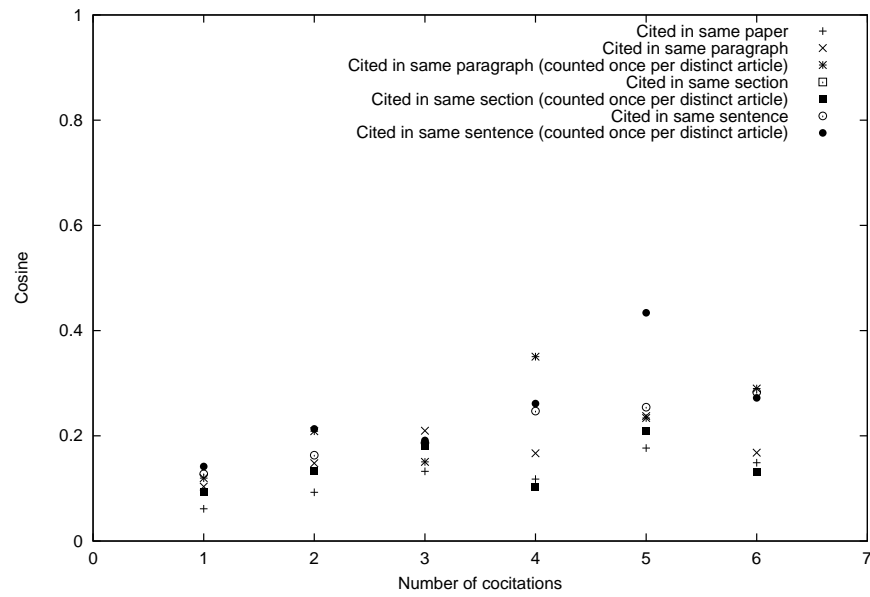
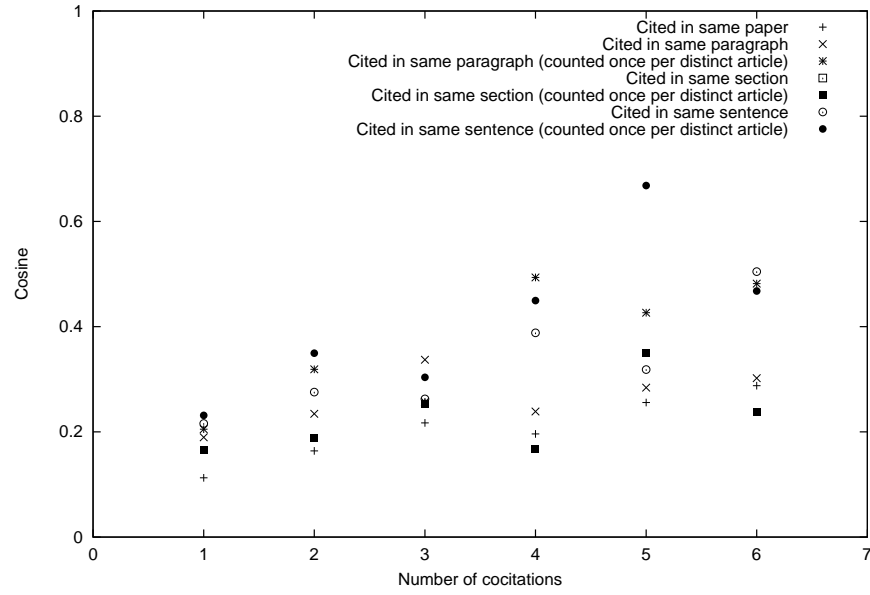Figure 19: Number of co-citations of A,B vs. $cos_d(ABS(A), ABS(B))$

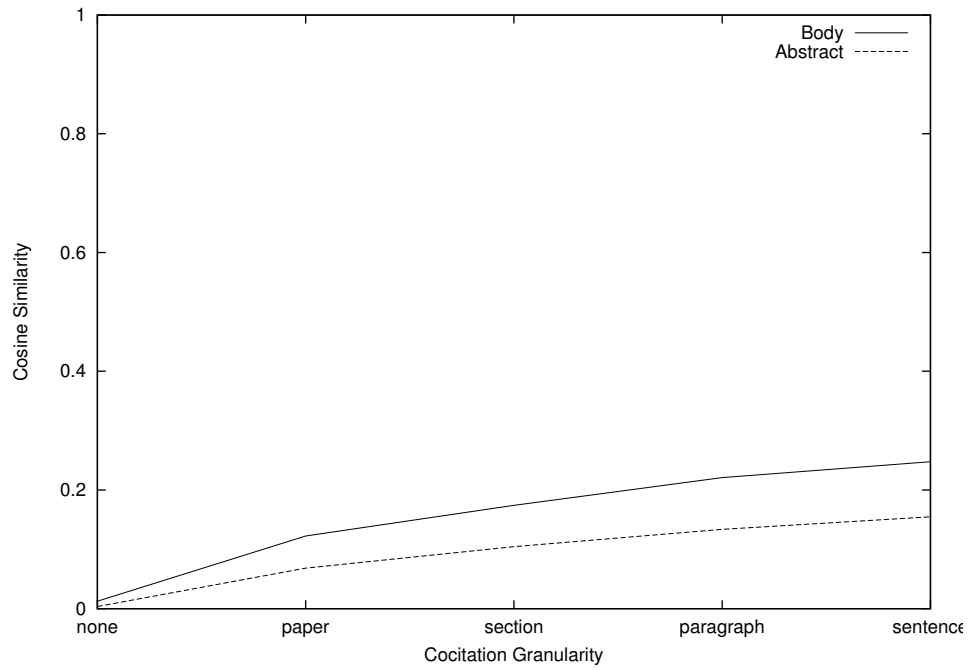Figure 20: Number of co-citations of A,B vs. average $cos_d(BODY(A), BODY(B))$

Figure 21: Granularity of co-citation vs. average cos_d

4   Conclusion

We introduced a sentence-oriented cosine similarity metric called *cohesion* which is highly correlated ($\rho \geq 0.84$) with the document-oriented *tf·idf*-weighted cosine similarity which it is based on. We defined two variants of cohesion, self-cohesion and cross cohesion, which allow comparison of the self-similarity of a document to its similarity with another document.

We used cohesion to analyze a corpus of biomedical journal articles. Mean self-cohesion of the citing sentences is somewhat higher than that of citations, however variance of self cohesion of citing sentences is higher; self-cohesion of either is higher than cross cohesion of citing sentences and abstract of the same paper, which is much higher than cross cohesion with citing sentences of a random paper. Additionally, the cross cohesion of abstract with some block of article text and cross cohesion of citing sentences with the same block is moderately ($\rho = 0.493$) but significantly correlated. This suggests that abstracts and citing sentences share some, but certainly not all, content in common.

However, the self cohesion of citing sentences decreases up to a point as the number of citing sentences increases. For articles with more than 20 or so citing sentences there is no observable trend, but the number of samples available is small. This is contrary to the expectation that self-cohesion would increase or remain relatively constant as the number of citing sentences increased since one would expect a paper to be cited for a small number of different things. This premise is not necessarily false because there are a number of confounding factors, such as synonyms and citing sentences referring to other papers or other topics in addition to the actual cited paper.

Since citing sentences appear to be somewhat more focused than the abstract and contain additional information not in the abstract, they could be useful as a supplement.

In the absence of an abstract, the citing sentences may provide a good substitute, especially in the context of automatic summarization. There has been ongoing research to make machines produce automatic summaries of an article, especially when the abstract of an article is not provided (news articles for example) or is not freely available (Luhn, 1958; Kupiec et al., 1995; Radev et al., 2002; Teufel and Moens, 2002). Automatically produced summaries are normally *extractive*, that is, they consist of a set of sentences from the article that provides an overview of the information in the article. This is much more tractable than the general problem of free-text summarization, but it is still quite challenging. Since there does seem to be a small but quantifiable difference in the information content of citing sentences as compared to abstracts, using the citing sentences as a guide to the salient aspects of an article in conjunction with other methods may assist in creating more useful extractive summaries.

We also examined the relationship between co-cited papers. They are significantly more cosine-similar than two random papers. Papers co-cited at a smaller granularity (in same paper vs. in same section, paragraph, sentence) are more cosine-similar than papers co-cited at a larger granularity. Papers co-cited twice are significantly more similar than papers co-cited only once; there are some additional significant differences as the number of co-citations goes up. However, the number of co-citations is not directly correlated to the cosine similarity. This suggests that the ability to browse co-cited articles would be useful in finding related work.

We want to conclude using a reprise of the title and first paragraph of this paper by repeating the observation that the citing sentences of an article are similar to the observations in the story of the blind men and the elephant: each sentence gives a focused perspective of the cited article and not necessarily a complete summary.

5   Acknowledgements

References

Bourne, P. E., Addess, K. J., Bluhm, W. F., Chen, L., Deshpande, N., Feng, Z., Fleri, W., Green, R., Merino-Ott, J. C., Townsend-Merino, W., Weissig, H., Westbrook, J., and Berman, H. M. (2004). The distribution and query systems of the RCSB Protein Data Bank. *Nucleic Acids Res*, 32(Database issue):223–225.

Bradshaw, S. (2002). *Reference Directed Indexing: Indexing Scientific Literature in the Context of Its Use*. PhD thesis, Northwestern University.

Bradshaw, S. (2003). Reference directed indexing: Redeeming relevance for subject search in citation indexes. In *Proceedings of the 7th European Conference on Research and Advanced Technology for Digital Libraries*.

Cohen, A. M. and Hersh, W. R. (2005). A survey of current work in biomedical text mining. *Briefings in Bioinformatics*, 6(1):57–71.

Garfield, E. (1955). Citation indexes for science: a new dimension in documentation through association of ideas. *Science*, 122(3159):108–111.

Gautier, R., Camproux, A.-C., and Tuffery, P. (2004). SCit: web tools for protein side chain conformation analysis. *Nucleic Acids Res*, 32(Web Server issue):508–511.

Goverdhana, S., Puntel, M., Xiong, W., Zirger, J. M., Barcia, C., Curtin, J. F., Soffer, E. B., Mondkar, S., King, G. D., Hu, J., Sciascia, S. A., Candolfi, M., Greengold, D. S., Lowenstein, P. R., and Castro, M. G. (2005). Regulatable gene expression systems for gene therapy applications: Progress and future challenges. *Mol Ther*, 12(2):189–211.

Kenny, P. A., Enver, T., , and Ashworth, A. (2005). Receptor and secreted targets of wnt-1 β-catenin signalling in mouse mammary epithelial cells. *BMC Cancer*, 5(3).

Kenny, P. A., Enver, T., and Ashworth, A. (2002). Retroviral vectors for establishing tetracycline-regulated gene expression in an otherwise recalcitrant cell line. *BMC Mol Biol*, 3(13).

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

Kupiec, J., Pedersen, J. O., and Chen, F. (1995). A trainable document summarizer. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 68–73.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research Development*, 2(2):159–165.

Menczer, F. (2004). Correlated topologies in citation networks and the Web. *The European Physical Journal B - Condensed Matter and Complex Systems*, 38(2):211–221.

Nakov, P. I., Schwartz, A. S., and Hearst, M. A. (2004). Citances: Citation sentences for semantic analysis of bioscience text. *Workshop on Search and Discovery in Bioinformatics at SIGIR 2004*.

Nanba, H., Abekawa, T., Okumura, M., and Saito, S. (2004a). Bilingual presri: Integration of multiple research paper databases. In *Proceedings of RIAO 2004*, pages 195–211, Avignon, France.

Nanba, H., Kando, N., and Okumura, M. (2004b). Classification of research papers using citation links and citation types: Towards automatic review article generation. In *Proceedings of the American Society for Information Science (ASIS) / the 11th SIG Classification Research Workshop, Classification for User Support and Learning*, pages 117–134, Chicago, USA.

Nanba, H., Kando, N., and Okumura, M. (2004c). Classification of research papers using citation links and citation types: Towards automatic review article generation. In *Proceedings of the American Society for Information Science (ASIS) / the 11th SIG Classification Research Workshop, Classification for User Support and Learning*, pages 117–134, Chicago, USA.

Nanba, H. and Okumura, M. (1999). Towards multi-paper summarization using reference information. pages 926–931.

Nanba, H. and Okumura, M. (2005). Automatic detection of survey articles. In *Research and Advanced Technology for Digital Libraries, 9th European Conference, ECDL 2005*, pages 391–401, Vienna, Austria.

Newman, M. E. J. (2001). The structure of scientific collaboration networks. 98(2):404–409.

Porter, M. F. (1997). An algorithm for suffix stripping. In *Readings in information retrieval*, pages 313–316. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Radev, D., Hovy, E. H., and McKeown, K. (2002). Introduction to the special issue on summarization. *Computational Linguistics*, 28(4):399–408.

Reynar, J. C. and Ratnaparkhi, A. (1997). A maximum entropy approach to identifying sentence boundaries. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pages 16–19, Washington DC.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.*, 24(5):513–523.

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. 24:265–269.

Teufel, S. and Moens, M. (2002). Summarising scientific articles - experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4).