Ryan M. Jones. Evaluation of the Effectiveness of Cosine Similarity in Predicting Relevance between Paired Citing and Cited Sentences. A Master's Paper for the M.S. in I.S. degree. May, 2009. 47 pages. Advisor: Catherine Blake.

Citation analysis has a long history in Information Science. We examined the potential of cosine similarity to predict relevance between citing sentences and the articles they cite. An expert evaluated 22,697 pairs of cited and citing sentences, and marked 544 as relevant to one another. Cosine similarity gave 8386 of these pairs a similarity score over zero, which included 339 relevant pairs. (4% precision, 65% recall). Under 0.01% of each cited article was relevant to the citing sentence, making precise retrieval challenging. We performed a detailed error analysis. Cosine similarity performance was reduced by insufficient window size, affixes, hyphenation, acronyms and abbreviations. The following preprocessing steps would improve retrieval performance: using a stemming algorithm that accounts for prefixes, expanding the window of comparison from sentences to paragraphs, identifying synonyms and expanding abbreviations. Further investigation of the possibilities of cosine similarity is necessary, but such investigation is worth pursuit.

Headings:

Text Mining

Citation Analysis

Cosine Similarity

Bibliometrics

EVALUATION OF THE EFFECTIVENESS OF COSINE SIMILARITY IN PREDICTING RELEVANCE BETWEEN PAIRED CITING AND CITED SENTENCES

by
Ryan M. Jones

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Information Science.

Chapel Hill, North Carolina

May 2009

Approved by

____

Catherine Blake

Table of Contents

Table of Figures

1. Motivation and context

There is a great deal of information contained in scholarly articles- not only within the text, but also in the citations between two texts. The article might provide new insight on the article's topic, some new procedure, or discovery, or new way of looking at old ideas. Citations also provide insight into how ideas grow and how knowledge within a field of becomes specialized.

One of the ways to examine the relationship between two scholarly articles is through their citations. A citation is a statement made in an article that cites, or references, a source outside the article for support- usually another article (see Figure 1). Authors often cite other scholars in the same field. When one scholar cites the work of another, they are attesting to a meaningful link between their article and the article they cite. Large numbers of citation links can reveal how ideas transmit between academic specializations.

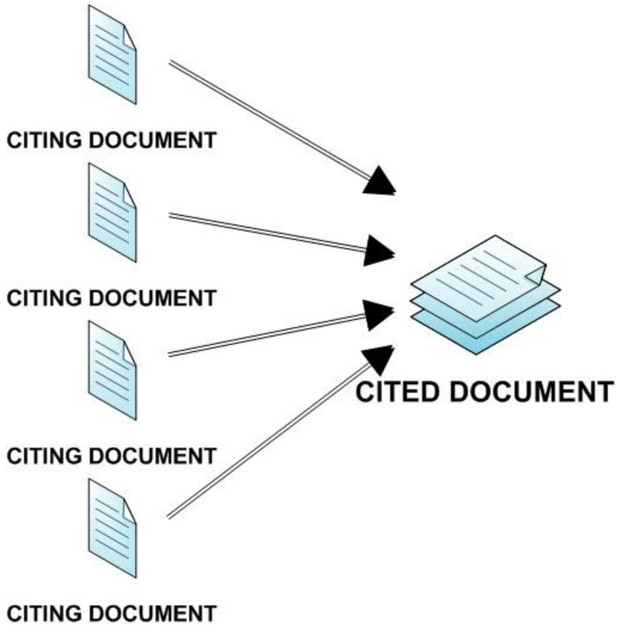


Figure 1: Citing documents versus cited document

More can be learned, not only by who cites who, but by the text an author uses in the citing sentence. The citing sentence contains terms that describe the content of the cited article. For example, these terms can be used for automatic generation of summaries, a shorter statement of the overall content in a particular document. Many systems create extractive summaries, limited to the vocabulary of the article they summarize. The additional vocabulary in the citing sentences may include new terms to describe the content, allowing the summary to go beyond the vocabulary of the summarized article. Each citation also focuses on the aspects of the article that the citing author found most important, so the terms in each citing sentence are more likely to be essential components of a summary.

Since scholars find themselves navigating an ever deeper sea of material relevant to their fields, tools that assist in their analysis of these large amounts of material would be very useful. I began to wonder if the citation text could be used to generate a summary of the content of the cited article, or to identify which parts of the cited article are relevant to the citation. If feasible, this would allow a researcher who is assimilating an article to be directed not only to the articles it cites, but the actual text of those articles that are relevant to the citation. I will investigate the relationship between citations and the cited article, and examine the text in the cited article relevant to the citation. It is possible that a single citing sentence may already be a one sentence summary of the cited article.

2. Previous research

Citation analysis, also called bibliometrics, a major part of my project, has been used to obtain many different kinds of information. Examination of citation practices creates a large, complex web of interaction which can reveal otherwise undetectable trends. For example, co-citations have been used to map out specializations of scientific fields. Co-citation analysis is a specific kind of citation analysis, in which the similarity between two articles A and B is measured by the number of articles have cited both A and B. Since both A and B are referenced by the same article, their content is linked in some way. This is in contrast to another form of citation analysis, bibliometric coupling, which the similarity between two articles A and B is measured by the number of references shared by A and B.

An important investigation of co-citation networks was described by Henry Small in his 1999 paper *Crossing Disciplinary Boundaries.* Although Small acknowledged past research on linking papers by shared vocabulary or index terms, he found that citations "represent(ed) a more direct author-selected dependency," and therefore made a strong foundation for study of inter-textual relationships. This observation motivated my own decision to focus on citation data.

Small had previously observed that citations tended to concentrate "in narrowly defined pockets that correspond roughly to specialties or invisible colleges of researchers." (1974). Researchers within a concise specialty would tend to cite the works of others within that specialty, creating small, interconnected groups of articles that were identifiable as a specialty. In his 1999 paper, Small examined articles that through their citations created a link between one group and another, linking the specialization groups. He observed that the articles drawing on citations from outside their specialty may be introducing an innovation. This could allow us to pinpoint where ideas cross from one field to another. In this and many other ways, examination of citations can reveal new and useful information about the relationships between articles and the ideas within them.

Braam (1991) examined the combination of citation maps with word profiles from a collection of articles and abstracts. This dataset consisted of abstracts from 3400 agricultural publications in *Chemical Abstracts,* and an additional 1384 publications on chemoreception from *BIOSIS.* These articles were combined with citation data from ISI's *Science Citation Index.* The citation maps and word profiles of these documents gave two images of the same dataset. Braam compared term frequencies from the abstracts with clustered co-citation analysis, examining how these profiles changed each year. This revealed new details, such as the development and adoption of specialized terminology within that discipline. Braam notes that when a specialization is unstable, developing quickly, the articles that get cited will vary more widely. In this situation, world profiles may be preferable to citations in establishing specialization groups. Contrasting word profiles with citation analysis gave a new way to track the development of ideas within scientific specialization.

Chen used citation analysis to trace the diffusion of knowledge through fields of science. (Chen, 2004). In this case, knowledge refers to the adoption of new concepts or processes by later writers who cite the earlier writers. Chen coupled citation analysis with other techniques- in this case, network theory and network visualization were applied to the network of citations. Visualization has turned out to be a very useful tool for assessing clusters in the network, as the results are far easier to understand visually. Chen's visualizations show tight clusters, interlinked by points of diffusion where articles (or in this case, patents) cite outside the clustered group.

Mapping science is not the only goal of citation analysis. Citations have also been used as measures of similarity, in various ways. Giles and his colleagues (1999) developed a measure of similarity that depended on common citations between articles, without reference to the text in the article. If two articles cite the many of the same sources, it indicated a degree of similarity in content between those two articles. He thereby proposed an alternative to TFIDF scores. However, this method can only be applied on the level of the whole article, so TFIDF remains a viable choice for determining similarity within the article.

A recent article (Elkis, 2008) entitled *Blind Men and Elephants* discussed the kinds of information can be discerned about an article by considering citation summaries. They concluded that when two articles were co-cited by the same article, those two articles would tend to be similar, and that this similarity increased with the proximity of the two citations within the citing article. This new measure of similarity was compared with tf-idf cosine similarity scores for the same text, and the two were found to perform very much alike, ranking items in close to the same order. Consequently, Elkis proposes, co-citation may be used as a measure of similarity. This article also examined self-cohesion, a measure of similarity between the sentences within a article, and cross-cohesion, similarity between that article and some other entity- in this case, the article’s abstract, and the collection of sentences citing that article. The sentences citing an article are generally more similar to the article’s main text than the article’s own abstract. Abstracts and citing sentences have different characteristics- for instance, the abstract covers the overall content in an article, while the citing sentences often focus on portions of the same article, and may not cover all of the content within. However, Elkis suggests that in the absence of an abstract, citing sentences may profitably replace it, through a process of automatic summarization.

Wangzhong (2006) proposed a measure of similarity based on citation linking through a graph, using two algorithms- the maximum flow metric and the authority vector metric.

Wangzhong concluded that citation and text based analysis are useful complements, confirming Braam's prior observations.

Klavans (2006) discusses several measures of similarity (or relatedness), such as the Pearson correlation. While he notes that little research has been done previously to evaluate the accuracy of relatedness measures, Klavans concludes that the cosine index performs the best. Van Eck (2009) recently reached a similar conclusion after comparing several techniques for measurement of similarity. Their observations helped to inform my use of cosine similarity measures in this project.

Ritchie (2008) performed a series of experiments in which the retrieval performance of the terms from within a document was compared with the terms used to describe that document in citations combined with the terms from within the document. It was discovered that adding the terms from the citations improved overall information retrieval performance. While these experiments were just the beginning of a longer term research project, they already support the idea that citation text has many more possibilities that we've not yet tapped.

3. Research question and approach

My intent is to evaluate the relationship between citing and cited documents, by examining measures of cosine similarity between the citing sentences and the text of the cited scientific articles. Since both the citing and cited documents discuss the same topics, I anticipate that the sentence pairs that are relevant to one another will be more similar than those that are not. If effective, this will allow identification of the material in the cited article that is relevant to the citing sentence.

Most of the studies mentioned above, while affirming the use of cosine similarity and tf-idf as the benchmark for measurement of textual similarity, consider the link of citing document to cited document in aggregate- between citing sentences and the cited article as a whole. The approach used in this work compares sentence to sentence rather than whole article to whole article. Similarity measures may prove useful within the article, perhaps even identifying the phrase within an article to which a citing article refers.

If using similarity to predict relevance within an article proves possible and reliable, it would have potential to develop into time saving research tools, allowing a scholar not only to know the source of citation but to immediately see the actual cited claim. Is it possible to predict which parts of a cited article are relevant to the citing sentence, using only a measure of textual similarity? Can existing metrics like similarity be used to predict sentence with the most information?

If cosine similarity were an ideal predictor for relevance, we could expect that using cosine similarity would result in 100% precision of results, and 100% recall of all relevant material. Since the relevant items would be highly similar to one another, they would dominate the initial results on the list when it is ordered by similarity. This platonic ideal of results is not likely to occur in actual practice, during my experiments. However, knowing the perfect result does allow evaluation of the performance of imperfect results by how closely actual results resemble, or fail to resemble, ideal performance.

Throughout this paper, I will frequently refer to *cited articles* and to *citing articles*. Each of the articles I've selected has, since the date of its publication, been cited a number of times by other scholars writing their own articles. For the purposes of this study, the term *cited* will refer to the original article, and *citing* to those later articles that refer to the original article, for support of a claim or for some other reason.

4. Materials and Methods

This section describes the materials and methods used to explore the degree to which the text from a citing sentence can be mapped to the original cited article automatically.

4.1. Materials

The texts on which I ran my tests were from a collection of chemistry and biomedical journals, which had previously been parsed and assimilated into a database for a previous project (Blake, 2006). The medical journals were from the genomics track of a dataset created by the Text Retrieval Conference to serve as a common test bed for Information

Retrieval research (Vorhees, 2006). The TREC data and the chemistry articles constituted two discrete collections. I assisted Blake with several experiments and some of the data processing they required, which acquainted me with the collection. Consequently, the earliest stages of data preprocessing- term frequency calculation, stemming, and parsing of the original text- had already been performed before I began work on these experiments.

The chemistry journals were a collection of 103,262 full text articles provided by the American Chemistry Society from 27 different journals, all published between 2000 and 2004. These articles had been processed previously for a different set of experiments. (Blake 2006)

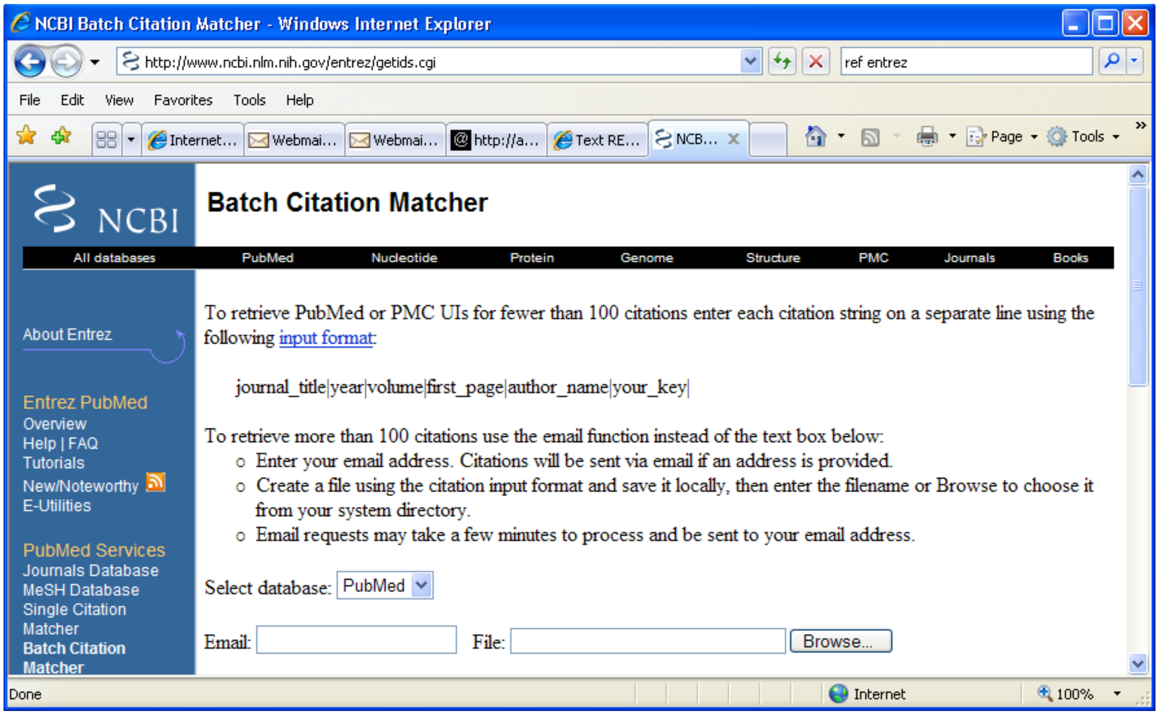
4.1.1. Data Preprocessing

The full text of each article included the list of citations from the end of each article, and the tags within the text of the article that linked each citation to the citing sentence. A citation in the text of the article would be marked with a number, and the corresponding number in the reference section contained the full details of the citation.

For example, a sentence might read "Other researchers thought so too.$_{10}$"and the reference at the end read "10. Jones, R. 2009. *I think so too,"* and so forth. This made it possible to link each sentence making a claim supported by outside material with the identifying information of that outside material- for instance, the title, author, year and journal of a particular journal article, using the reference number. One might then learn how many times an article was cited, but for one problem. The articles proved inconsistent in their citation styles- one might use the full name of a journal, while another might use one of several acceptable abbreviations.

At this point in the processing, it was possible only to tell how many times each article was cited by the same form of the title. The same article might have one count under the name *Journal of the National Academy of Science,* and another count under the abbreviated name *Natl Acad Sci,* when the correct total was the sum of the two enumerations. Fortunately, the National Library of Medicine indexes the 60 journals in the TREC collection. The NLM index includes the full title for each journal, as well as each journal's accepted abbreviations, making it possible to disambiguate and group the varied forms of each journal title under the same identifier. Each article so indexed has a unique identifying number, the Pubmed ID, or PMID. The NLM offers a Batch Citation Matcher at www.ncbi.nlm.nih.gov/entrez/getids.cgi. This citation matcher provides the PMID for each known citation. I uploaded extracted citations in batches that ranged between fifty and one hundred thousand at a time, and loaded the responses from the NLM back into the database. This allowed me to link the articles to their PMID using the title, date, journal, etc from each citation.

Importantly, because each TREC article was already identified by a PMID, this also made it possible to know which other articles cited other articles already in our collection.


NCBI Batch Citation Matcher - Windows Internet Explorer
http://www.ncbi.nlm.nih.gov/entrez/getids.cgi
ref entrez
File Edit View Favorites Tools Help
Internet… Webmai… Webmai… http://a… Text RE… NCB…
Page Tools
NCBI
All databases PubMed Nucleotide Protein Genome Structure PMC Journals Books
About Entrez
Entrez PubMed
Overview
Help | FAQ
Tutorials
New/Noteworthy
E-Utilities
PubMed Services
Journals Database
MeSH Database
Single Citation Matcher
Batch Citation Matcher
Browse...
Done
Internet
100%

Batch Citation Matcher

To retrieve PubMed or PMC UIs for fewer than 100 citations enter each citation string on a separate line using the following input format:

journal_title|year|volume|first_page|author_name|your_key|

To retrieve more than 100 citations use the email function instead of the text box below:

- Enter your email address. Citations will be sent via email if an address is provided.
- Create a file using the citation input format and save it locally, then enter the filename or Browse to choose it from your system directory.
- Email requests may take a few minutes to process and be sent to your email address.

Select database: PubMed

Email: File:

Figure 2: NCBI batch citation matcher

One article might cite forty or fifty others, which included articles both within and without the available collection of full texts. I limited my analysis to those citing articles already in collection, so I would have the text of both the citing article and the cited article. In the case of the TREC collection, the citation information and PMID from each article was used to identify which pairing of citer and cited were both part of the collection. In the case of the chemistry journal collection, this work had been done previously as part of Blake's prior research.

Of the subset of cited articles for which at least some of the citing articles were available, an individual article might have been cited multiple times by other researchers. The number of citations made to each article varied widely. Some articles were cited thousands of times, while many others were cited only once and most were not cited at all. Figure 3 shows the average number of times an article is cited from other articles with the chemistry collection.

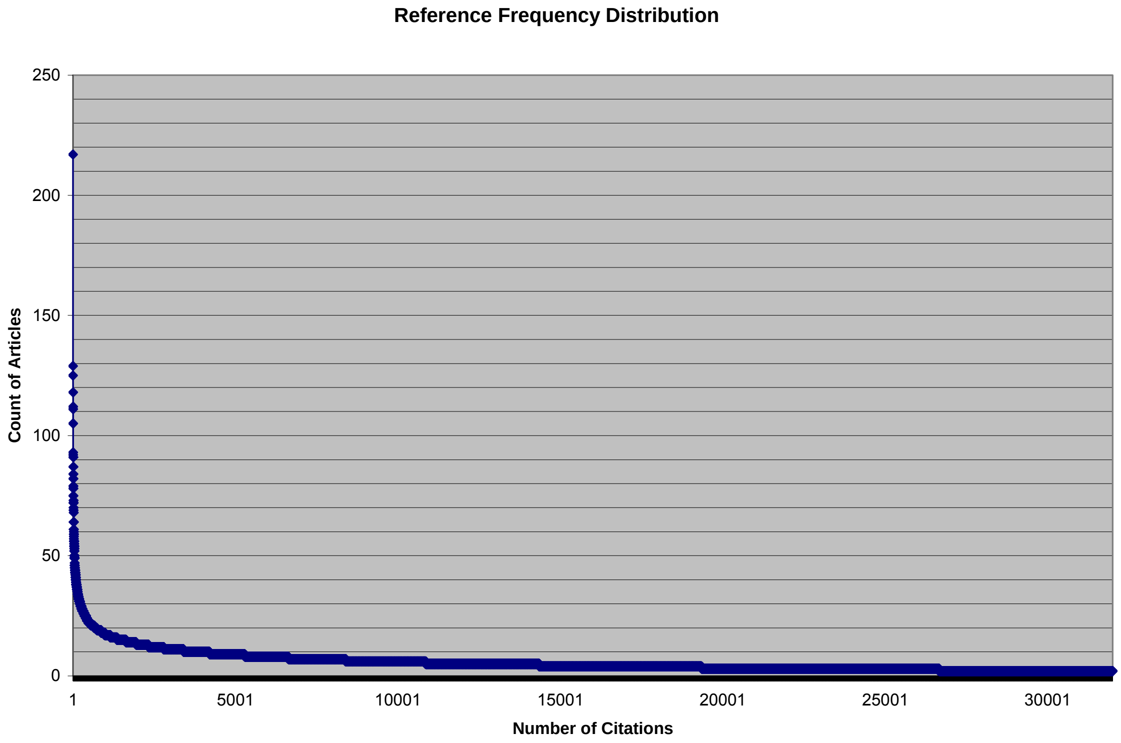


Figure 3: The number of times each cited article is cited within the chemistry collection

Since the median of the curve fell around 20 articles, I selected those articles cited close to this number of times. Of these 96 articles, I selected a smaller experimental set randomly, consisting of nine articles, each cited 20 times by other articles.

The cross product of every sentence in each article with each sentence that cites that article created a set of 22,697 sentence pairs.

4.2. Methods

This section describes the methods employed to prepare data and interpret the trends.

4.2.1. Overview

My intent is to explore the degree to which automated methods can reflect, match, or even anticipate human judgments of relevancy at a sentence level, and to explore the relationship between cited text and the original article. The system measured the cosine similarity between all sentence pairs, which was compared with the expert's relevancy judgment. This will allow an evaluation of the relationship between the two measures, demonstrating whether similarity of text on the sentence level can predict relevance of cited text to citing text.

4.2.2. Stop Words

Before calculating the similarity of a pair of sentences, the system removed stop words from consideration. Stop words are words that have little informational value, such as 'the', 'and', 'a' or 'of'. In these experiments, the system used the stop words provided by the National Library of Medicine (last updated in 2000) which consists of 364 terms. The list was obtained at the following URL:
http://www.ncbi.nlm.nih.gov/entrez/query/static/help/pmhelp.html#Stopwords.
This list of stop words does not include numbers. I did not add numbers to the stop word list, knowing that in many cases, specific numbers, such as measurements, are a distinguishing part of the statements made in these articles.

4.2.3. Stemming

A vital step in the process of determining term frequencies is to reduce the terms in question to their basic stems. The same term might appear in singular form, plural, or in various tenses. Some terms will appear without variation in all articles, and others are altered by prefixes or suffixes. Terms like ‘exposure’, ‘exposing’ and ‘exposed’ are formed off the same basic stem of ‘expose’. Without stemming, the term frequencies will give deceptive results. If all of these are varied forms are reduced to a common stem, which need not be the correct grammatical stem as long as the computer can process it, the result will be a single, more accurate count of related terms, instead of a larger number of separate low counts for terms that are just different forms of the same term. The stemmer used in this case was a Java adaptation of the Porter stemmer (Porter, 1980).

4.2.4. Term Frequency

Term frequency is the number of times each stemmed term appears in a given article. Intuitively, frequency correlates with relevance. For example, if you are searching for a given term, and the term appears twenty times in the first article, but only five in the second, a simple retrieval system should offer the first article as more relevant than the second.

High term frequency does not always indicate relevance. For instance, the size of an article can significantly influence the term frequency. Consider two articles, the first comprising a hundred pages, and the second comprising one page. Now consider that both articles have a term frequency of twenty. Although the search term frequency is the same, the first article's subject has little to do with the term, while the second article is highly relevant. Term frequency is a useful starting point, but is by itself insufficient for predictions of article relevance or similarity.

Term Frequency (TF(*j*)) = (# of stemmed terms) – (stop words in document *j*)

4.2.5. Inverse Document Frequency

The Document Frequency is the number of articles in which a term appears. Common terms in a collection have low discriminative power. Two articles about chemistry should not be judged similar or relevant to one another purely on the basis of the term 'chemical' appearing in both, while in a different field the same term might be more informative. Using the inverse document frequency lowers the weight of common terms.

Inverse Document Frequency (IDF(*i*)) = (# of documents that contain stemmed term *i*)

4.2.6. Term Frequency * Inverse Document Frequency

Salton and Buckley (1987) suggested that terms should be assigned a weight based on the term frequency and the inverse of their document frequency. When calculated as shown below, TF-IDF creates a weight for each term that balances how often a term appears in an individual document with how many documents use the term. If a term appears many times in an article, but is common to all other articles as well, the frequency of that term is less informative than a term that is uncommon in the article collection as a whole. Rare terms are weighted more heavily, common terms more lightly, making it easier to identify discriminative terms.

TFIDF: $\text{Weight}(i,j) = \text{TF(j)} * \text{log2}(i / \text{IDF}(j))$

Where i is the term and j is the document

4.2.7. Normalization of Term Weights

Differences in article length can influence term frequencies. A large number of occurrences of a term makes the document appear relevant, but if the article is very long each of those occurrences constitutes a very small portion of the full document. A short document may mention the same term only a few times, yet those mentions make up a larger portion of the overall article. This will lead to misleading results from term frequencies. The effect of varying article lengths can be mitigated by normalizing term weights for each article.

Normalization of Term Weights: $\text{norm(D)} = \sqrt{(\sum \text{w}(j)2)}$
Where j is the document

4.2.8. Cosine Similarity

Cosine similarity incorporates the information provided by the calculations above to create a numerical value to describe the similarity of each pair of compared items. The value itself means little, but a group of such values creates a natural ordering of comparisons in which the highest values are the most similar and the lowest values are the least.

Cosine(D1,D2) = ∑(wD1(j)*wD2(j) / norm(D1) * norm(D2))

Generate a value of similarity for each sentence pair, based on the values of shared terms adjusted for article length

4.2.9. Calculating Similarity

Term frequencies and inverse document frequencies were calculated for each individual stemmed term. These are combined to create a TF*IDF score, which is then normalized to account for varying lengths between sentences. This normalization is used to calculate cosine similarity between each citing sentence and every sentence in the cited article. These calculations of similarity will be compared with manual assessments of whether the paired sentences from the citing and citing articles cite or support one another. The citing sentences will be compared one with another in the same way. Similarity was calculated using the equations presented in *Text Mining: Predictive Methods for Analyzing Unstructured Information (Weiss, 2005, pp 91-92)*. Sentences containing more than two citations were excluded from this comparison, since the citation text tends to become less specific the more citations it includes.

4.3. Preliminary investigations

4.3.1. Initial test experiment

I selected an article that 158 TREC articles cited, *Efficient presentation of soluble antigen by cultured human dendritic cells is maintained by granulocyte/macrophage colony-stimulating factor plus interleukin 4 and downregulated by tumor necrosis factor alpha,* by Sallusto and Lanzavecchia, a 1994 article from the Journal of Experimental Medicine. I collected the text from all the sentences in our collection that cited this article. It quickly became apparent that all of the sentences referred to one of a few key concepts from the cited article, and that the citing sentences were very similar to one another and used the same distinctive terminology.

The citing articles were written by experts, thus each citation sentence was in effect a set of terms compiled by a field expert and asserted to be relevant to the content of the cited article.

PMID 12649135:Monocytes can differentiate into DCs in vitro when cultured in the presence of granulocyte macrophage-colony-stimulating factor (GM-CSF) and IL-4 or IL-13 for 5 to 7 days.
PMID 12406905: Monocyte-derived dendritic cells (DCs) were generated by culturing PB monocytes from healthy donors in cRPMI supplemented with 50 ng/mL granulocyte-macrophage colony-stimulating factor (GM-CSF) and 100 U/mL IL-4 for 7 days.
PMID 12384430: DCs were generated in vitro from monocytes (MDDCs) in the presence of granulocyte macrophage-colony-stimulating factor (GM-CSF) and IL-4 or from cord blood CD34 progenitors in the presence of GM-CSF and tumor necrosis factor (TNF).
PMID 12149218: These unique features of DCs are increasingly exploited for the design of DC-based vaccines in immunotherapy since sufficient numbers of monocyte-derived DCs (MoDCs) can be obtained through in vitro differentiation of monocytes in the presence of granuloctye-macrophage colony-stimulating factor (GM-CSF) and IL-4.

Figure 4: Four similar sentences from initial sample

If such a high degree of similarity exists between sentences citing the same article, then measures of similarity might be a useful guide for predicting related concepts. We know these sentences shown in Figure 4 are related because they all cite the same article, but had the citation been left out, the content of each phrase is similar enough to conclude that they are related. While it is possible to use different terminology to discuss the same concept, or to make the same claim, the technical terminology of science and chemistry works to limit this- by design, technical terms have few synonyms.

4.4. Creating the Gold Standard

The validity of this study depends on the quality of the annotations. We recruited an annotator who had completed post-doctoral research in chemical engineering to provide relevance judgments, which would be contrasted with cosine similarity scores.

I built a PHP script that generated a HTML form dynamically, which enabled the annotator to evaluate each sentence pair. The annotator was provided with two sentences, the first was the citing sentence, and the second was each sentence from with the article that was cited. The annotator marked each sentence pair as relevant, somewhat relevant, or not relevant. The interface displayed the sentences in the cited article together, allowing the annotator to see the whole article's text and observe how the sentences related to one another. Each article was be annotated with respect to each citing sentence. The process took on average 3 to 5 hours per article.

At least one sentence in each cited article should be relevant to support the link between the two articles made by an expert in the field,. Not all citations are made for supporting material- scientist A might write "Unlike scientist B, who makes claim X, (in cited article Y) I make claim Z". Even if the scientist A does not agree with scientist B's claim, the words used saying so will be relevant.

Citers may cite several articles at once, making a broad statement that is supported collectively by all the cited articles. The citing sentence will become more general, and will tend to have fewer distinctive terms in multi-article citations than when one article is cited. An example of this effect is the following sentence, taken from *Using Raman*

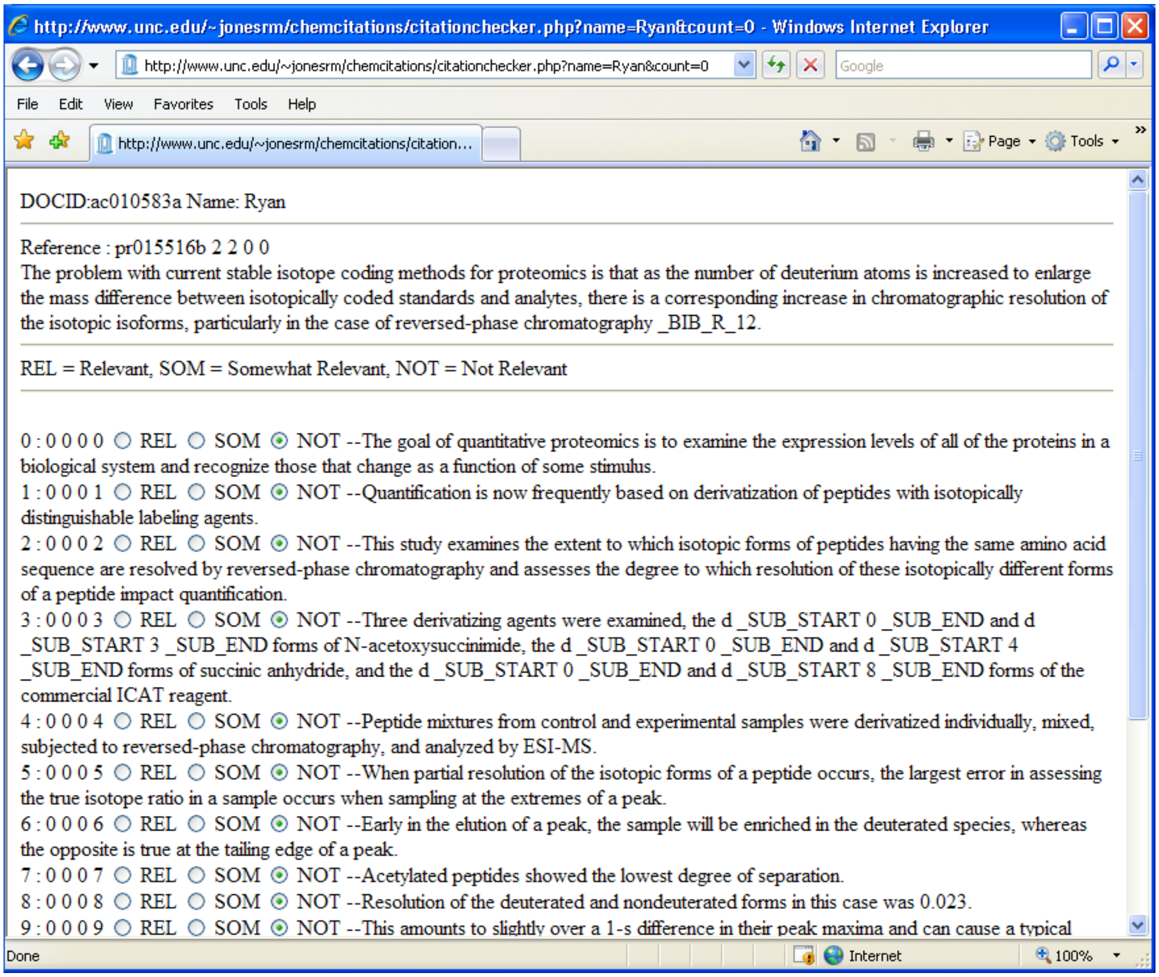
Spectroscopy to Elucidate the Aggregation State of Single-Walled Carbon Nanotubes, Journal of Physical Chemistry B, volume 108 issue 22, 2004, part of our chemistry corpus:

> *Only recently have researchers begun to seriously address this problem.*

This sentence contained five citations to other articles, and is certainly relevant to their content; but, the sentence is so general, that it could equally apply to hundreds of thousands of articles.

This generalization effect, while not universal, waters down the uniqueness of terminology sufficiently that I am disregarding citing sentences that include three or more simultaneous citations, focusing instead on comparing sentences that refer to only one article, where it should be easier to identify a relationships between the citing sentence and cited document.

Using a Java program to execute the queries across all the SQL tables containing article text, the term frequencies and inverse document frequencies for each term were normalized. These results were stored in a single table on an Oracle database. Then, using SQL queries (included below) similarity scores were calculated between all articles using cosine similarity. For the purposes of the equations below, each individual sentence is treated as a separate article, since our interest is within the article, rather than comparison of entire articles. Cosine similarity gave a numerical value measuring the similarity between each pair of phrases, each citing phrase paired to each sentence in the citing article. Stop words were excluded from this comparison of similarity.


http://www.unc.edu/~jonesrm/chemcitations/citationchecker.php?name=Ryan&count=0 - Windows Internet Explorer
http://www.unc.edu/~jonesrm/chemcitations/citationchecker.php?name=Ryan&count=0
Google
File
Edit
View
Favorites
Tools
Help
http://www.unc.edu/~jonesrm/chemcitations/citation...
Page
Tools
Done
Internet
100%

DOCID:ac010583a Name: Ryan

Reference : pr015516b 2 2 0 0
The problem with current stable isotope coding methods for proteomics is that as the number of deuterium atoms is increased to enlarge the mass difference between isotopically coded standards and analytes, there is a corresponding increase in chromatographic resolution of the isotopic isoforms, particularly in the case of reversed-phase chromatography _BIB_R_12.

REL = Relevant, SOM = Somewhat Relevant, NOT = Not Relevant

0 : 0 0 0 0 ○ REL ○ SOM ● NOT --The goal of quantitative proteomics is to examine the expression levels of all of the proteins in a biological system and recognize those that change as a function of some stimulus.
1 : 0 0 0 1 ○ REL ○ SOM ● NOT --Quantification is now frequently based on derivatization of peptides with isotopically distinguishable labeling agents.
2 : 0 0 0 2 ○ REL ○ SOM ● NOT --This study examines the extent to which isotopic forms of peptides having the same amino acid sequence are resolved by reversed-phase chromatography and assesses the degree to which resolution of these isotopically different forms of a peptide impact quantification.
3 : 0 0 0 3 ○ REL ○ SOM ● NOT --Three derivatizing agents were examined, the d _SUB_START 0 _SUB_END and d _SUB_START 3 _SUB_END forms of N-acetoxysuccinimide, the d _SUB_START 0 _SUB_END and d _SUB_START 4 _SUB_END forms of succinic anhydride, and the d _SUB_START 0 _SUB_END and d _SUB_START 8 _SUB_END forms of the commercial ICAT reagent.
4 : 0 0 0 4 ○ REL ○ SOM ● NOT --Peptide mixtures from control and experimental samples were derivatized individually, mixed, subjected to reversed-phase chromatography, and analyzed by ESI-MS.
5 : 0 0 0 5 ○ REL ○ SOM ● NOT --When partial resolution of the isotopic forms of a peptide occurs, the largest error in assessing the true isotope ratio in a sample occurs when sampling at the extremes of a peak.
6 : 0 0 0 6 ○ REL ○ SOM ● NOT --Early in the elution of a peak, the sample will be enriched in the deuterated species, whereas the opposite is true at the tailing edge of a peak.
7 : 0 0 0 7 ○ REL ○ SOM ● NOT --Acetylated peptides showed the lowest degree of separation.
8 : 0 0 0 8 ○ REL ○ SOM ● NOT --Resolution of the deuterated and nondeuterated forms in this case was 0.023.
9 : 0 0 0 9 ○ REL ○ SOM ● NOT --This amounts to slightly over a 1-s difference in their peak maxima and can cause a typical

Figure 5: Interface used by the domain expert to establish the gold standard

The manual annotation allowed comparison with the similarity scores for each sentence pair. By manually annotating which sentences in the cited article are relevant, we create a set of articles wherein we know which sentences are directly relevant to a sentence that cites it, and which are at least partly relevant. This allows us to examine the qualities of the sentences, such as similarity, that might contribute to the semantic relationship between them.

5. Results

The comparison of cosine similarity scores with manual evaluations of relevance revealed that cosine similarity alone, applied to individual sentences, is an insufficient predictor of relevance. However, the process of making the comparison has revealed several adjustments to the technique that may produce improve retrieval performance

5.1. Summary of Human Judgments

Each article was annotated repeatedly, each time with respect to a different citing sentence. The amount of time required to annotate each document varied from two to five hours, depending on the complexity of the comparison and the length of the article involved. The nine cited articles selected for the evaluation had an average of 60 relevant sentences each between all citing sentences, ranging from 33 to 117. Of the 22,697 marked pairs 526 (2.3%) were marked as relevant and an additional 1.2% were marked as at least partly relevant. For each citing sentence there was an average of 7 relevant sentences (ranging from 1 to 30) in the cited document. The small percentage of relevant sentences suggests that automated retrieval will be difficult, because the target is very small, comprising about 0.1% of the article.

The sentences from the cited article marked relevant to the citing sentence occurred the most often (51%) in the introduction section of the cited article. The next most frequent location was the article’s discussion of results (23%). The remainder appeared in the abstract (8%), the conclusion (3%), and a few in the methods section (1%). The remaining portion (14%) was uncategorized.

5.2. Overall results

In order to evaluate the value of cosine similarity as a predictor of relevance, we first need to know what perfect results would look like, so we can see how far the real world results depart from the ideal. Knowing how they differ from ideal results will show us how to improve the metric, or, if necessary, show us that cosine similarity is not useful for this purpose.

Precision and recall are the typical measurement of performance for information retrieval. Precision is the number of relevant documents retrieved divided by the total number of document retrieved. In this experiment we compare sentences, so if ten sentences are retrieved and five are relevant, the search had 50% precision. Recall is the number of relevant documents retrieved divided by the total number of relevant documents that should have been retrieved. Again we use sentences rather than documents, such that if five relevant sentences are retrieved, but another five are marked as relevant, but not retrieved, the search had 50% recall.

The cosine similarity score assigns a numerical value to each pair of sentences where the highest number corresponds to the most similar sentence pair. If a pair of sentences has no words in common, the sentence pair has a similarity score of zero. In these experiments we consider sentence pairs with a cosine similarity score of zero as not retrieved.

Since many of the relevant sentences had none of the terms of the citing sentence, the system rarely achieved 100% recall. The full retrieved set of items with similarity scores greater than zero resulted in 65% recall. At this point of recall, the precision of the results was 4%. Precision was highest at 7.76%, when recall was at 6.84%.

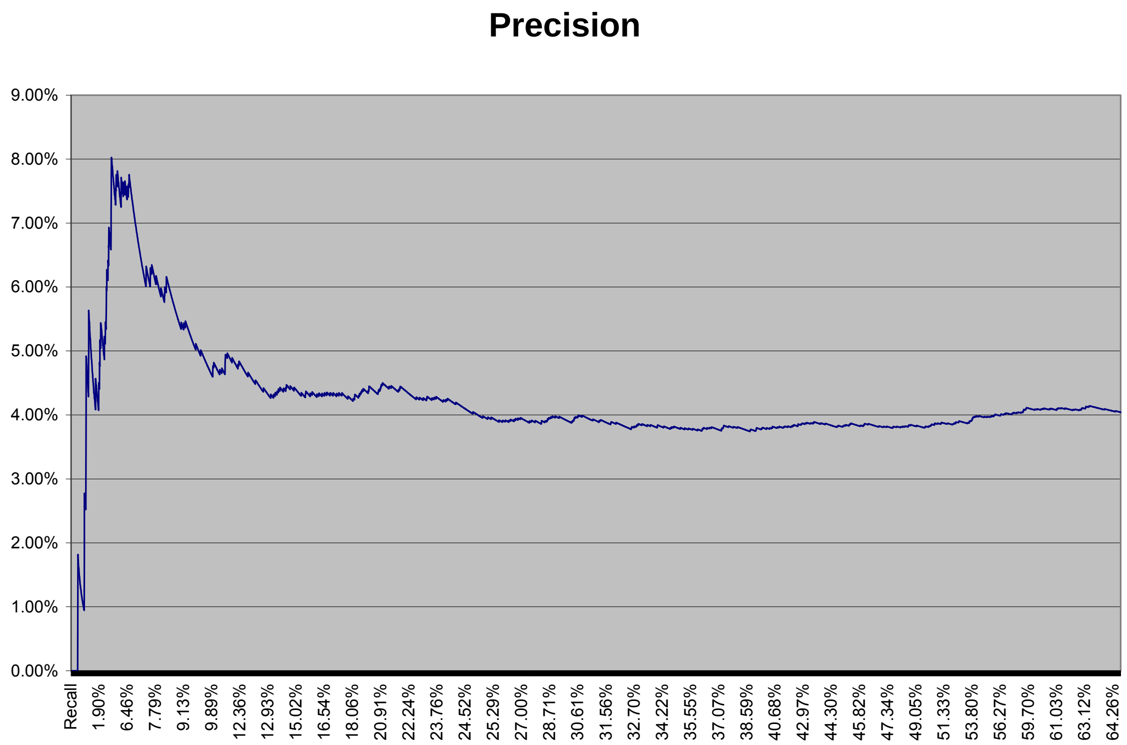


Figure 6- Precision and recall for all cited documents

Another way to look at the performance of the similarity metric is to consider the average position of the relevant sentences in the ranked list. If sentence similarity has no correlation to human relevance, we would expect the sentences marked relevant to range all over the list. Those sentence pairings marked as relevant to one another were, on average, within 34.4% of the top of the list. Considering each marked cited article individually, the average range of results varied from within 44% of the top of the list to within 14% of the top of the list. These results suggest that the majority of relevant sentences were concentrated early in the ordered list of similarity. Each article comprised 1000 to 3300 sentences.

Figure 7 shows examples of sentence pairs, the four highest ranked items, as ordered by cosine similarity. The first item marked as relevant appears fourth in the list.

Citing Sentence	Candidate Cited Sentence	Relevance	Similarity
Previously, we have shown that dimerization of quadruply hydrogen bonding 2-ureido-4 [1H]-pyrimidinone (UPy) derivatives is very strong and has an association constant of 6 x 10 7 M -1 in CDCl .	The dimerization constant of 1 was previously estimated as exceeding 2.2 x 10 6 M -1 (293 K in chloroform) and Zimmerman has shown that quadruply hydrogen bonded DDAA dimers of 2 have a dimerization constant of at least 3 x 10 7 M -1 (CDCl 3).	Not	1.712783
Because of the moderate (2 x 10 4 M -1 , UTr) to high (6 x 10 7 M -1 , UPy) association constants between the units, reversible polymers with a high degree of polymerization were obtained.	From these data, K dim was determined to be (5.7 + - 0.6) x 10 7 M -1 (r = 0.992) for chloroform, (1.0 +- 0.1) x 10 M -1 (r = 0.995) for wet chloroform, and (5.9 +- 0.7) x 10 8 M -1 (r = 0.993) for toluene.	Not	1.647787
Because of the moderate (2 x 10 4 M -1 , UTr) to high (6 x 10 7 M -1 , UPy) association constants between the units, reversible polymers with a high degree of polymerization were obtained.	Figure 1 Plot of the normalized excimer fluorescence of 1b vs concentration, measured in chloroform, chloroform saturated with water, and toluene, curves are derived from the nonlinear curve fit.	Not	1.647787
The high dimerization constant (6 x 10 7 M -1 in chloroform) makes it possible to obtain materials with a high degree of polymerization.	From these data, K dim was determined to be (5.7 + - 0.6) x 10 7 M -1 (r = 0.992) for chloroform, (1.0 +- 0.1) x 10 M -1 (r = 0.995) for wet chloroform, and (5.9 +- 0.7) x 10 8 M -1 (r = 0.993) for toluene.	**relevant**	1.627649
The high dimerization constant (6 x 10 7 M -1 in chloroform) makes it possible to obtain materials with a high degree of polymerization.	Figure 1 Plot of the normalized excimer fluorescence of 1b vs concentration, measured in chloroform, chloroform saturated with water, and toluene, curves are derived from the nonlinear curve fit.	Not	1.627649

Figure 7: Items Ranked by Cosine Similarity

Figure 7 shows that several of the cosine similarity scores are the same. This result suggests that there is insufficient data for comparison since there are not enough terms to distinguish between the similarity of different sentences and thus the sentences cannot be properly ordered. Expanding the window so that more than a single sentence of the cited document is compared, would add more terms to the comparison and thus may improve the cosine similarity metric.

5.3. Error Analysis

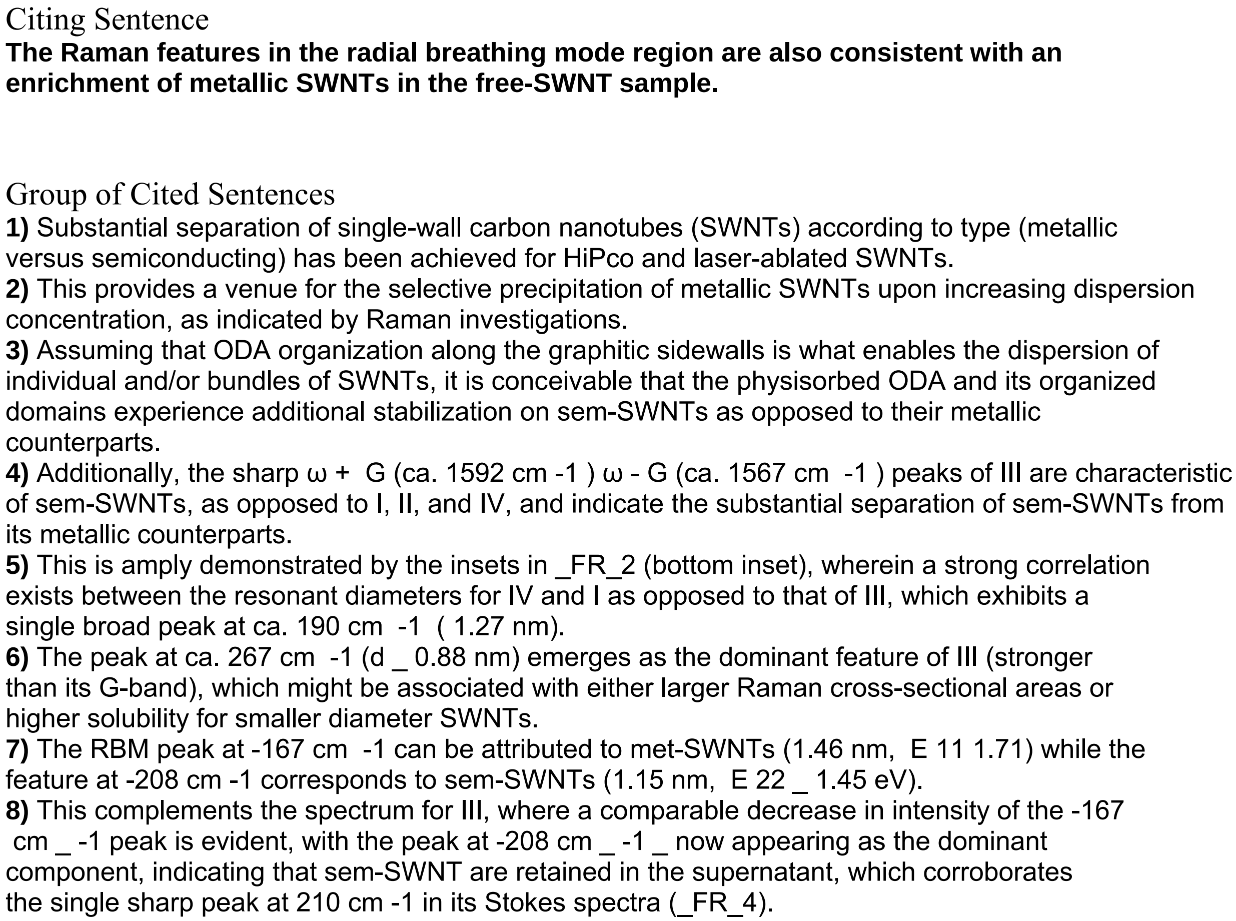
Of the sentence pairs manually marked as relevant, we selected 50 at random and examined each sentence in detail, to identify the factors that lead to low similarity scores. We describe these factors in the following sub-sections.

5.3.1. Adjustments to Window Size

Cosine similarity is typically employed over larger segments of text- most often, whole articles. I compared similarity scores on a sentence to sentence basis, a citing and a cited sentence from two articles, rather than on a full article to full article basis. In practice, the window size of a single sentence appears to have been too small.

About 75% of the time, the cited article contained the information relevant to the citing sentence, in several sentences rather than in a single sentence. Consequently, relevant sentences tended to be marked in clusters. This finding makes sense- complex ideas can take many sentences to discuss. Thus, instead of a high similarity score between two sentences sharing multiple key terms, we observe low similarity scores between the citing sentence and a number of cited sentences, with key terms spread out among them. Thus, the cosine similarity ranking is particularly sensitive to the effects caused by short selections.

Figure 8 provides an example of the sentence spanning phenomena. The number of shared words between the citing sentence and any one of the relevant marked sentences from the cited article low, but when we consider the whole paragraph together, the relationship is more apparent. Specifically the key terms in the citing sentence such as SWNT, Raman, and metallic are distributed throughout the paragraph of the cited article, preventing any single sentence from having a high cosine similarity.



Citing Sentence
The Raman features in the radial breathing mode region are also consistent with an enrichment of metallic SWNTs in the free-SWNT sample.

Group of Cited Sentences
1) Substantial separation of single-wall carbon nanotubes (SWNTs) according to type (metallic versus semiconducting) has been achieved for HiPco and laser-ablated SWNTs.
2) This provides a venue for the selective precipitation of metallic SWNTs upon increasing dispersion concentration, as indicated by Raman investigations.
3) Assuming that ODA organization along the graphitic sidewalls is what enables the dispersion of individual and/or bundles of SWNTs, it is conceivable that the physisorbed ODA and its organized domains experience additional stabilization on sem-SWNTs as opposed to their metallic counterparts.
4) Additionally, the sharp ω + G (ca. 1592 cm -1) ω - G (ca. 1567 cm -1) peaks of **III** are characteristic of sem-SWNTs, as opposed to **I**, **II**, and **IV**, and indicate the substantial separation of sem-SWNTs from its metallic counterparts.
5) This is amply demonstrated by the insets in _FR_2 (bottom inset), wherein a strong correlation exists between the resonant diameters for **IV** and **I** as opposed to that of **III**, which exhibits a single broad peak at ca. 190 cm -1 (1.27 nm).
6) The peak at ca. 267 cm -1 (d _ 0.88 nm) emerges as the dominant feature of **III** (stronger than its G-band), which might be associated with either larger Raman cross-sectional areas or higher solubility for smaller diameter SWNTs.
7) The RBM peak at -167 cm -1 can be attributed to met-SWNTs (1.46 nm, E 11 1.71) while the feature at -208 cm -1 corresponds to sem-SWNTs (1.15 nm, E 22 _ 1.45 eV).
8) This complements the spectrum for **III**, where a comparable decrease in intensity of the -167 cm _ -1 peak is evident, with the peak at -208 cm _ -1 _ now appearing as the dominant component, indicating that sem-SWNT are retained in the supernatant, which corroborates the single sharp peak at 210 cm -1 in its Stokes spectra (_FR_4).

Figure 8: Dispersal of relevant terms over several sentences

The result is that individual sentences often each contain a piece of the relevant idea, but nevertheless rank low when ordered by cosine similarity. This suggests that identifying a single sentence is too focused and that a more appropriate comparison would be between the citing sentence and a paragraph of the cited document. This would still identify relevant portions of the cited article, and would do so more reliably.

5.3.2. Anaphoric References

Another form of interference in the evaluation of similarity comes through anaphors. Of the sample set of 50 sentences, 8, or 16%, continued to cite a key concept anaphorically, using pronouns like ‘the’ or ‘it’ to link the current sentence with the subject that had been introduced in a prior sentence. This reduced their similarity score.

Sentence containing subject: Furthermore, the reported affinity of amine groups for semiconducting SWNTs, as opposed to their metallic counterparts, contributes **additional stability to the physisorbed ODA**.
Relevant sentence: **This** provides a venue for the selective precipitation of metallic SWNTs upon increasing dispersion concentration, as indicated by Raman investigations.

Figure 9: Cited sentence expressing subject anaphorically (emphasis added)

Without resolution, pronouns are of little value, which is why pronouns such as ‘this’ and ‘it’ are in the stopword list. In the example above, something is providing a venue for selective precipitation of single walled nanotubes. The author provides the subject outside of this sentence, and therefore reduces the similarity score erroneously.

The structure of the paragraph offers a solution here as well. In every observed case, the subject of the sentence was defined earlier in the paragraph. Expanding the window to include the full paragraph should resolve this, since concepts cited anaphorically will now include the full identification of the subject. Actively resolving anaphoric references may also improve performance.

5.3.3. Acronyms and Abbreviations

Of the 50 analyzed sentences, 21 of 50 sentences (41%) used acronyms and abbreviations. Further, a third of the sentences with acronyms and abbreviations used more than one in the sentence. This does not include extremely common abbreviations for units of measurement, like cm for centimeter, or C for Celsius. When one sentence used an acronym or abbreviation, and the other used the full term, cosine similarity would not consider them similar even though they cite the same concepts. In some cases, expanding the window size will resolve abbreviations. Many articles give a full explanation of an acronym the first time it is used, and if this falls within the larger window, it will register as more similar. Unfortunately, such initial explanations are usually given only once, usually at the beginning of the document. Even when comparing citing sentence to cited paragraph, it is likely that such clarification will not be included in the comparison. Opening the window even further means that we are no longer pinpointing the source concept in the cited article.

To further muddy the waters, many abbreviations are commonplace, the author is unlikely to offer an explanation. It is unlikely that a reader of a professional scientific article will need a definition of MRI, or what element *Bi* represents. Yet if cited and citer refer to these concepts differently, they will confound the effectiveness of cosine similarity. This did in fact occur- both cited sentences below had a lower score of similarity than they should have. In the first case, the citing article mentions Bismuth and the cited used the abbreviation *Bi.* In the second, the citing article uses the term *metallic*, while the cited article uses *met*, a common sense abbreviation that the authors do not define in the article.

> Nonetheless, our XRD, TEM, and composition analysis have unambiguously demonstrated that the tubular structures in our sample are metal Bi nanotubes.

> The RBM peak at -167 cm -1 can be attributed to met-SWNTs (1.46 nm, E 11 1.71) while the feature at -208 cm -1 corresponds to sem-SWNTs (1.15 nm, E 22 _ 1.45 eV).

Figure 10: Cited sentences featuring abbreviations and acronyms

Expanding acronyms and abbreviations before or in place of stemming may improve the retrieval performance using the cosine similarity metric. Devising a way to perform such expansion is a new project, however, and lies outside of the auspices of this project. However, other researchers have explored this challenge (Bapat, 2009; Schwarz 2003)

5.3.4. Adjustments to the Stemming Algorithm

All terms were stemmed using a java implementation of the Porter stemming algorithm (Porter, 1980) prior to evaluating similarity between the two sentences. The intent was to improve the cosine similarity score by grouping terms with a similar stem, such as 'dimer' and 'dimerize' which would be reduced to 'dim'.

The Porter stemming algorithm did improve the retrieval performance of cosine similarity, but rather exhibited a noteworthy deficiency that is particularly relevant to scientific terminology: the Porter algorithm does not consider remove prefixes. This can be more complicated than removing suffixes, but for the scientific domain, it is essential. My initial expectation was that scientific terminology would, once stemmed by suffix, prove highly consistent and give reliable similarity scores. Scientific terminology deliberately limits synonymy, so that while a poet might describe the same Albertan rose as ‘crimson’, ‘red’, or ‘blushing’, the formal scientific name is consistently *Rosa acicularis*. Many scientific terms and processes are formed by compounding base terms with prefixes. However, despite the consistency and clarity that scientific terminology affords, I found that the same concept could appear with a prefix in the citing document, and without in the cited document, and vice versa.

Figure 11 provides an example of two sentences that are annotated as relevant to one another, but had a similarity score of zero as there are no identical stems shared between the sentences. Although both sentences used the stem term “tube,” differing uses of prefixes between the two prevented the similarity score from reflecting the similarity in concepts.

Recently, we have developed a low-temperature hydrothermal reduction method and successfully synthesized Bi **nanotubes**.	A significant portion (about 30%) of the sample dispersed on the TEM grids shows **tubular** structures, although other nano-sheets were also observed.

Figure 11: Dissimilarity caused by prefixes- emphasis added

Despite a low similarity score, both sentences in Figure 11 share a key concept that is obscured by inconsistent use of prefixes. One mentions ‘nanotubes’, the other ‘tubular’. Although not identical concepts, they are similar, particularly since we already know that the content between the two articles is related. If both articles discuss the concept of tubes, it is likely to be relevant. Stemming of both prefixes and suffixes would have reduced these to the same stem, and thus these terms would contribute to the cosine similarity score. This inconsistency between the two authors' use of prefixes created a false negative, in which the same core concept was treated as two separate terms.

Most stemming algorithms are deliberately cautious with affix removal, since trimming a term excessively can increase the frequency of false positives. In the case of general information retrieval, removing affixes might falsely rank two articles as highly similar to one another. However, in our collection, the risk of false positives between two unrelated articles is mitigated because the two articles are related by virtue of the citation link between them. The terminology within the two articles constitutes a much smaller vocabulary, such that if two terms are reduced to the same stem, they may be less likely to be false positives.

Kantrowitz compared the effects of various stemmers on TFIDF rankings. (Kantrowitz, 2000). The algorithm that included prefix stemming was shown to significantly outperform (by about 30%) the suffix-only Porter stemmer. We anticipate that such an algorithm would have had similar benefits in this case. Bacchin (2002) describes a graph-based algorithm that focused on prefix stemming, which performed as well on a set of Italian articles as other algorithms that had been optimized for Italian. Paice's Lancaster stemmer is another alternative that addresses prefixes. Either of these approaches would produce more consistent stems and thereby produce more reliable cosine similarity scores.

5.3.5. Hyphens

Hyphens were employed inconsistently between authors. One article might refer to a nano-tube and another to a nanotube. Terms with inconsistent use of hyphens will not be included in the cosine similarity unless the punctuation is removed during the preprocessing, or the stemming algorithm accounts for such hyphens.

5.4. Establishing Sentence to Sentence Relevance

This study relies on expert evaluation of relevance to create the gold standard, against which the system retrieval performance is evaluated. The manner in which this expert evaluation is conducted is therefore of utmost importance. In the case of this study, the annotator was provided with each article intact, with the sentences in the same order they were originally written. This allowed the annotator to compare each citing sentence with concepts in the cited document that were expressed in more than one sentence, and to follow the path of each concept through the article on a higher level. This approach helped to demonstrate the need to open the window of comparison, so that the algorithm would behave more like the expert annotator, and make comparisons on a higher level than sentence to sentence.

It would be both useful and interesting, however, to break the ordering of the sentences in each article so that the progression of concepts from sentence to sentence is obscured. This would force the annotator to consider relevance purely on a sentence to sentence basis- to behave as the algorithm employed in this paper. The comparison would become a test of whether cosine similarity can predict human relevance, when the human judges relevance with no knowledge of the sentence context. Anaphors would certainly have an impact in obscuring the accuracy of human judgment.

To explore the degree to which context influenced the manual relevance judgments, I selected the top fifty sentences, as ranked by cosine similarity, from five articles. The sentences were ordered by cosine similarity, and were thus removed from individual context. Interestingly, when the context was removed, the annotator knowledge of the context of the sentence, using only the sentences themselves as the basis of comparison, he was more likely to mark individual pairs as relevant. Pairs that were not previously marked as relevant were marked this time, with the smaller set. This suggests that the performance of cosine similarity may in fact be better than the gold standard would suggest, and that future investigation should take into account the effects of cognitive overload.

The precision of the results was far higher than in the prior experiment. Performance peaked around 75% precision, and dropped to 40% by the time recall within the test set reached 100%.

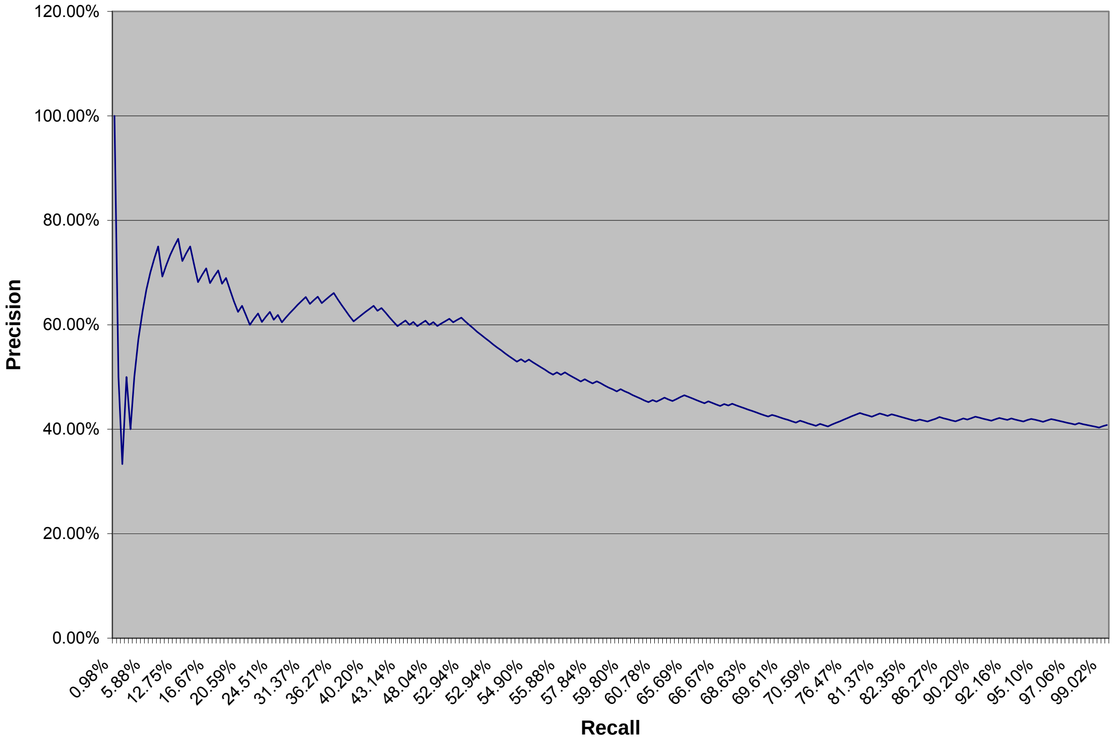


Figure 12- Precision and recall for contextless annotation

It is possible that, divorced from the context of the sentence, the annotator is himself more reliant on textual similarity to make his evaluation, since the meaning of the sentence is more difficult to judge. If this is the case, the vast improvement in cosine similarity's predictive performance is irrelevant- all that we did was force the annotator to rely on human judgments of similarity rather than on human judgments of relevance. The test experiment performed here is insufficient to demonstrate whether this is the case, so the matter demands further investigation.

6. Conclusion and Recommendations

The goal of this study was to explore the relationship between the text used in a citing sentence and the text in the cited article. We asked an expert to annotate sentences in the cited article that he considered relevant. We targeted nine documents that were cited by between 13 and 20 other articles in our collection. On average, only 0.1% of sentences in an article were relevant to an individual citing sentence. The number of relevant sentences ranged between one and thirty-three, with an average of 7.04. When considering all citing sentences, the expert annotator considered only 2.3% of the sentences relevant in the cited article, and an additional 1.3% as partially relevant. The small number of relevant sentences per cited article reflects the difficulty of this retrieval task.

Using cosine similarity between the citing sentence and sentences in the cited article resulted in 8386 sentences with a value greater than zero. Of the returned sentences 339 out of the 544 relevant sentences were identified, giving a recall performance of 65% and precision performance of 4%The retrieval performance of cosine similarity varies between cited articles, but the average performance ranked relevant sentences in the top 35% of sentences.

We conducted an error analysis on fifty sentence pairs selected at random that revealed the following five main issues:

1) Although the citing sentence summarizes the key points made in an article, the cited text will often develop the same idea over multiple sentences, or a paragraph. This spreads the relevant terms from the citing sentence over several

sentences in the cited article. The relevant sentences are individually less similar to the citing sentence than the paragraph as a whole.

2) Authors refer to key concepts anaphorically, so that a sentence that discusses a relevant term does not actually contain that relevant term.
3) Authors refer to key concepts using acronyms and abbreviations. If the citing sentence and cited article do not use the same uncontracted form, acronym or abbreviation, the sentence similarity score will be low, despite being relevant.
4) Authors employ affixes inconsistently between citing sentences and cited articles, which reduces the similarity score, despite being relevant.
5) Authors use hyphens inconsistently, which reduces the similarity score. .

Based on our error analysis, we make the following text transformations recommendations to improve retrieval performance:

1) Expand the text window from a sentence to a paragraph to resolve anaphoric references and capture ideas developed over multiple sentences.
2) Use a stemming algorithm that removes prefixes and hyphens.
3) Use thesauri to resolve abbreviations, acronyms and synonymy.

Given the ever deepening morass of information that scholars must navigate, we must explore tools such as these will become invaluable. Although cosine similarity has been well explored for information retrieval, knowing exactly how and why it fails is critical if we are to improve retrieval performance Further investigation of our recommendations is required to measure the change in effectiveness of the cosine similarity metric, but this study shows that such investigation is worth pursuit.

7. Acknowledgments

I’d like to thank Catherine Blake, for her advice and her confidence in me, and the rest of the faculty at SILS, who taught and inspired me, as well as my parents, who supported me from afar with wise guidance.

Above all else, I am grateful to my wife and children, who have been patient and loving throughout these years, whose faith in me never failed, and who were ultimately my reason for persisting through the challenges graduate school presents. I did not know my life was incomplete until you made it whole.

8. References

Bacchin, M., Ferro, N., Melucci, M. (2002) The Effectiveness of a Graph-Based Algorithm for Stemming. pp. 117-128.

Blake, C. (2006) A Comparison of Document, Sentence and Term Event Spaces. *The 44th Annual Meeting of the Association for Computational Linguistics (ACL), Sydney Australia.* pp. 601-608

Braam, R., Moed, H.F., van Raan A.F.J. (1991) Mapping of Science by Combined Co-Citation and Word Analysis. I. Structural Aspects. *Journal of the American Society for Information Science.* **42 (**4) pp. 233-251.

Braam, R., Moed, H.F., van Raan A.F.J. (1991) Mapping of Science by Combined Co-Citation and Word Analysis. I. Dynamical Aspects. *Journal of the American Society for Information Science.* **42 (**4) pp. 252-266

Chen, C., Hicks, D. (2004) Tracing Knowledge Diffusion. *Scientometrics.* **59 (**2) pp. 199-211.

Elkis, 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) pp. 51-62.

Kantrowicz, M., Behrang, M., Mittal, V. (2000) Stemming and its effects on TFIDF Ranking. *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* pp. 357-359.

Klavans, R., Boyack, K.W. (2006) Identifying a Better Measure of Relatedness for Mapping Science. *Journal of the American Society for Information Science and Technology*. **57 (**2) pp. 251-263.

Paice, C.D. (1990) Another Stemmer. *ACM SIGIR Forum* **24** (3) pp. 56-61.

Porter, M.F. (1980) An algorithm for suffix stripping. *Program* **14** (3) pp. 130−137.

Ritchie, A., Teufel, S., Robertson, S. Using Terms from Citations for IR: Some First Results. *Advances in Information Retrieval.* Springer Berlin/Heidelberg. pp. 211-221.

Salton, G., Buckley, C. (1987) Term Weighting Approaches in Automatic Text Retrieval. *Information Processing and Management* **24 (**5), pp. 513-523.

Schwartz, Hearst. (2003) A simple algorithm for identifying abbreviation definitions in biomedical text. *Pacific Symposium on Biocomputing*. pp 451-62.

Small, H., Griffith, G.C. (1974) The Structures of Scientific Literatures; Identifying and Graphing Specialties. *Social Studies.* **4** (1) pp.17-40

Small, H. (1999) Crossing Disciplinary Boundaries. *Library Trends.* **48** (1) pp. 72-108.

van Eck, N. J., Waltman, (2008) Appropriate Similarity Measures for Author Co-citation Analysis. *Journal for the American Society for Information Science and Technology*. **59** (10) pp. 1653-1661.

Vorhees, E.M., (2006) Overview of the TREC 2006. *Text REtrieval Conference (TREC) 2006 Proceedings.* pp 1-16.

Wangzhong, L., Janssen, J., Milios, E., Japkowicz, N., Zhang, Y. (2006) Node Similarity in the Citation Graph. *Knowledge and Information Systems.* **11 (**1) 105-129.

Weiss, S.M., Indurkhya, N., Zhang, T., Damerau. F.J. (2005) *Text Mining: Predictive Methods for Analyzing Unstructured Information.*Springer Science+Business Media Inc. New York. pp 91-92.