Are Cited References Meaningful? Measuring Semantic Relatedness in Citation Analysis

Hassan Alam[1], Aman Kumar[2], Tina Werner[3], Manan Vyas[4]

BCL Technologies
(hassana[1], amank[2], twerner[3], mvyas[4])@bcltechnologieshttp

Abstract. In this proof-of-concept study we use standard cosine similarity measure to calculate the semantic similarity between two pieces of text – the citing document and the cited text. Three subject matter experts then evaluate the citing and the cited text based on the cosine score to give their judgement on the semantic similarity between the two pieces of text.

Keywords: Bibliometrics, citation analysis and network analysis for IR

1 Introduction

Researchers and scientists in both academia and industry present and publish their research work in a variety of places and platforms. Because of career pressure and other factors, they are encouraged to publish and present more and more. The large and rapidly increasing amount of scientific literature online and otherwise (book, journals) has triggered intensified-research into understanding the effectiveness and quality of this research work.

When reading a research paper we often glance at the bibliography or the list of references for additional information. An author cites the references when they look up for information while preparing for their research paper and they want to acknowledge all the sources they have used in the process of writing that paper. Ideally, authors are expected to report the sources even though they do not quote directly from that source. Readers can use the referenced list to check for the accuracy of the published material and that establishes credibility for the author. But as a reader, we may not have time to do further consultation because of the sheer size of the cited material.

In this study we are building a proof-of-concept system that looks only at the relevant parts of the cited material that is appropriate for evaluating the claims made by the original author in a given paragraph. This system analyzes the text around the cited sentences or text in the original article and tries to find the cited material from the referenced articles to check if the cited text is semantically related to the citing text in the original document.

Compared to humans, this tool cuts down the time considerably in reading and analyzing the cited material.

Our goal in this study is do a proof-of-concept study 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 the cited documents discuss the same topics, we assume that the concepts 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 text. Once we establish that this similarity metrics for this specific task gives satisfactory results, we will implement other semantic similarity measures such as Latent Semantic Indexing and evaluate the results.

2 Related Work

Author in [1] explores reasons why citing and cited works may be related. The analysis indicates that factors such as sources of the cited document, citing work, frequency of a work cited, and type of citing articles predict closer relatedness between citing and cited works. Authors in [4], [6] and references there-in discuss several measures of similarity and relatedness, such as the Pearson correlation and conclude that the cosine index performs the best.

In this preliminary work, first, we use standard similarity index – cosine similarity score to establish similarity between two pieces of text, and then we use manual judgment to understand what type of citing and cited texts closely match semantically with each other.

This paper is organized as follows. In section 2 we describe the methodology we adopt to do the empirical analysis to establish semantic relatedness. In section 3 we describe the experiment set up for this task, followed by a discussion of the evaluations, and conclusions.

3 Methodology

In this study we want to investigate the degree to which automated methods can reflect, match, or even predict human judgments and to understand the semantic relationship between the original article and the cited text.

The automated system calculates the cosine similarity between all sentence pairs, which is then compared with the Subject Matter Expert's (SME) relevancy judgment. The idea is that can we correlate the semantic similarity of two sentences and ascertain the relationship of relevance between the citing and the cited text.

For Data Preprocessing, we used the stop word list [7] to get rid of the stop words for further processing.

For stemming**,** we used Porter stemming algorithm [8] which is a Java implementation. The motivation for stemming is that if we do not do stemming, the *tfidf* counts will yield false results. *tf* as such is not sufficient for our goals of predicting an article's relevancy or establishing similarity between two pieces of texts. Using the *inverse document* frequency lowers the weight of common terms. A weight is created by the tfidf for each term. This establishes a balance between how often a term appears in an individual document and with how many documents use the term. In this model, a common, more frequent term is weighted lightly and an unfamiliar or rare term is weighted more heavily. This results in identifying discriminative terms. Mathematically, tfidf weight is calculated using the standard formula:

$$weight\ (i,j) = tf(j) * \log(\frac{i}{idf(j)})$$

Where, i is the term, and j is the document.

Normalization

The term frequencies can be influenced by difference in the length of the article. A more frequent term in a long article will skew the results. Also, it's likely that in short article a term gets repeated a number of times which may lead to misleading results. In order to mitigate the effect due to article length and term frequencies we need to normalize the term weight for each article. The normalization of term weight is expressed mathematically as:

$$norm(D) = \sqrt{}(\sum w(j)^2)$$

Here j is the document

Cosine Similarity Score

In order to compute the similarity of each pair of the compared items, the cosine similarity gives a numerical value that describes by how much the two compared items are close to each other. A group of cosine similarity score creates a natural ordering of comparisons in which the highest values are the most similar and the lowest values are the least.

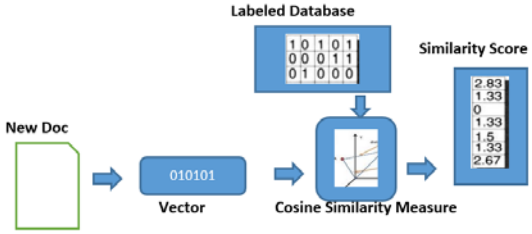


Fig. 1. Computing Similarity score

This cosine similarity score computes a value, adjusted for article length, to depict the similarity for each sentence pair, based on the values of shared terms. Mathematically, it is represented as follows.

$$\boldsymbol{Cosine\ (D_1, D_2)} = \sum(wD_1(j)*wD_2(j)/norm(D_1)*norm(D_2))$$

To summarize, the algorithm we implemented for this proof-of-concept system is as follows.

Step 1: Term frequencies and inverse document frequencies is calculated for each individual stemmed term.

Step 2: The term frequencies are combined to create a TF*IDF score.

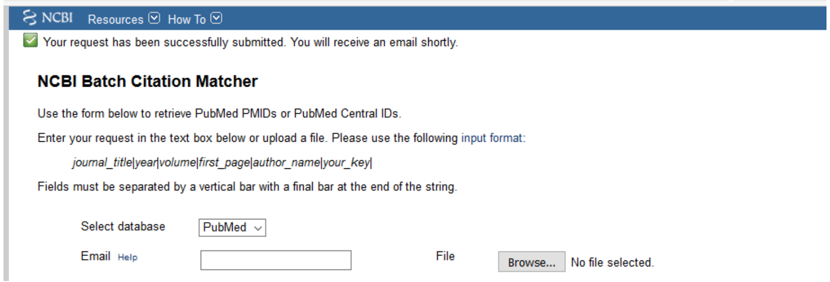
Step 3: The TF*IDF score is then normalized to account for varying lengths between sentences.

Step 4: This normalization is then be used to calculate cosine similarity between each citing sentence and every sentence in the cited article.

Step 5: The similarity score is compared with manual assessments of whether the paired sentences from the citing and citing articles cite or support one another.

3.1 Data

We wrote a tool to extract data from NLM/NCBI [9]. 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 is indexed, and 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. Here's a snapshot of the NCBI Batch Citation Matcher.



✓ Your request has been successfully submitted. You will receive an email shortly.

NCBI Batch Citation Matcher

Use the form below to retrieve PubMed PMIDs or PubMed Central IDs.

Enter your request in the text box below or upload a file. Please use the following input format:

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

Fields must be separated by a vertical bar with a final bar at the end of the string.

Select database: PubMed

Email Help

File: No file selected.

Fig. 2. NCBI Batch Citation Matcher (https://www.ncbi.nlm.nih.gov/pubmed/batchcitmatch)

Using this interface at NCBI we can submit extracted citations in batches that could range from fifty thousand to one hundred thousand at a time, and load the responses from the NLM back into our database. This allows us to link the articles to their PMID using the title, date, journal, etc. from each citation.

The assumption is that the full text of each article includes 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.

4 Experiment

We extracted 50 journal articles from PubMed. For each citation in an article the tool extracted the corresponding paper. The tool then extracted two sentences before and two sentences after the citation in the original document and tried to match the words in those sentences with the target document using the cosine similarity metric. This process generated a cosine similarity index for each citation in the original document.

Once we have the cosine similarity measurements, we picked up the pairs (citation in the original sentence and relevant parts in the cited document) that scored higher than 0.90. Three subject matter experts (SME - clinical experts in this case) then manually evaluated the citations sentences and the cited documents, and decided which of the correlated documents matched the most. The human experts based their judgement mainly on semantic matching of the sentences in the two documents and not just on matched strings.

4.1 Results

For manual evaluations the three SMEs looked at 100 matched set that scored higher than 0.90 cosine similarity score. SMEs rated their assessment on a scale of 1-100, 100 being the best match. For example, SME-1 found that out of the 100 paired texts, 62 talked about the same concept. In this preliminary study we did not analyze the disagreements between the SMEs. Table 1 gives a summary of this evaluation process.

Table 1. SME's evaluations of 100 paired texts

SME	Semantic Relatedness (> 0.90 cosine score)
SME-1	62
SME-2	67
SME-3	64

5 Conclusions

In this proof-of-concept study we analyzed the textual similarity between citation text in an original research paper from PubMed and the corresponding text in the cited document. We tried to understand how close the author was in citing the cited paper. We first used cosine similarity measure to come up with a paired list of citation text and cited text. We then looked at 100 such pairs with a cosine similarity score of over 0.90. The system recorded an average accuracy of 64.33% based on the evaluations of the three SMEs. For future work, we plan to extend the similarity metrics using the WordNet synset hierarchy and distributional similarity and Latent Semantic Analysis (LSA) index.

References

1. Bonzi, S.: Characteristics of a literature as predictors of relatedness between cited and citing works. Journal of the American Society for Information Science, 33(4):208-216 (1982).
2. Boyack, K. W., Small, H., and Klavans, R.: Improving the Accuracy of Co-citation Clustering Using Full Text. In Proceedings of 17th International Conference on Science and Technology Indicators. (2012)
3. Corley, C., and Mihalcea, R.: Measuring the Semantic Similarity of Text, in Proceedings of the ACL workshop on Empirical Modeling of Semantic Equivalence and Entailment, pp. 13−18. (2005)
4. Klavans, R., Boyack, K.W.: Identifying a Better Measure of Relatedness for Mapping Science. Journal of the American Society for Information Science and Technology. 57 (2) pp. 251-263 (2006)
5. Madylova, A., and Oguducu, S.G.: A taxonomy based semantic similarity of documents using the cosine measure, in Proceeding of International Symposium on Computer and Information Sciences, 2009, pp. 129−134 (2009)
6. van Eck, N. J., Waltman.: Appropriate Similarity Measures for Author Co-citation Analysis. Journal for the American Society for Information Science and Technology. 59 (10) pp. 1653-1661. (2008)
7. https://www.ncbi.nlm.nih.gov/books/NBK3827/table/pubmedhelp.T.stopwords/
8. http://www.nltk.org/howto/stem.html
9. https://www.ncbi.nlm.nih.gov/pubmed/batchcitmatch)