Discovering Semantic Relatedness between Scientific Articles through Citation Frequency

Abdul Shahid, Muhammad Tanvir Afzal, Muhammad Abdul Qadir,

1Institute of information technology, KUST, Kohat
2,3Department of Computer Science, Muhammad Ali Jinnah University, Islamabad Campus

Abstract: Finding relationship between scientific articles is dire need of scholarly community. Current systems support the task to some extent, for example showing cited-by articles, related-articles etc. However, in many cases, these systems show a long list of related documents. For example, if a paper has been cited-by 500 papers, the Google Scholar shows the list of those 500 papers. The system itself is unable to recommend the most relevant cited-by articles for a cited article. This hinders the researchers to skim all of the cited paper to search the most suitable papers because of the fact that many researchers just cite the paper to give background study of the topic etc. In this way there is no strong relationship between the both papers. Therefore, there is a need for a system that can discover semantic relatedness between these documents and could recommend a few most relevant cited-by articles. The semantic relatedness could mean different for different people. In this paper, we consider the cited-by paper more relevant to cited paper if the cited paper has been referred within the cited-by paper’s text more than a threshold value. We have empirically proved that Citation Frequency could be another measure to find semantically related set of articles. We have tested our system for a dataset of online journal such as: Journal of Universal Computer Science (J. UCS), we found that if cited-by paper refers the cited paper more than five times, both are semantically related and shows a strong relationship.

Key words: citation frequency, citation relationship, document relatedness, semantic relationship

INTRODUCTION

The growth of digital publications is exponential (Afzal et al., 2007) and to find the relevant information is crucial task. If we look on the stats then we came to know that the world information is doubling every eighteen months and in scientific domain every 5 years (http://www.activated.org/books/chapter.php?id=17.10). The scientifically acknowledged systems such as ISI(http://www.isiknowledge.com/), and Citeseer (http://citeseer.ist.psu.edu/) index large set of information. For example only ISI has indexed about 9,000 international and regional selected journals and book series. Similarly if we search on Google Scholar (http://scholar.google.com.pk.) for ontology papers, we find one millions papers and the first entry titled “Gene ontology: tool for the unification of biology” has more than 6000 citations.

The researchers (Simon Buckingham Shum, 1998) are always having an interest in knowning relationship between the articles. They need to know semantic relationship between the publications. The semantic relationship is hard to find between two articles with current approaches. The existing approaches cannot be easily extended to find semantic relationship between two papers.

There are various approaches that are being used in the literature to find the relationship or related papers for a focused paper such as: Co-citation analysis, bibliography analysis (Giles et al., 1998; Ratprasartporn and Ozsoyoglu, 2007). Text-based relatedness (Giles et al., 1998), Context based relatedness (Ratprasartporn and Ozsoyoglu, 2007) and the technique developed (Afzal et al., 2007) for creating links into the future. The detail of these techniques can be found in related work. These techniques do not have the capability to show any kind of semantic relationship between papers.

In scientific domain, Garfield, (1955) described that a strong relationship can be seen through the citations. However, the cited-by paper cites a paper for some reasons, for example, some times to just give a background study, sometimes the cited-by paper is actually extending the work mentioned in cited paper etc. These reasons are hard to identify and need strong NLP, Machine Learning, and Artificial Intelligence techniques. However, the problem can be indirectly solved such that if we analyze the text of the cited-by paper, we can have some
patterns. For example 1) a paper is only cited once in the text of the cited-by paper or 2) a paper is cited more than once in the text of the cited paper. If we could define a threshold value on number of citations for which both of the papers have strong relationship, it would be great. This is where the system contributes.

We investigated and proposed that Citation frequency could be another important filter in order to find semantic relatedness between scientific works. This cited-by article lists set of references at the end of the paper. The Citation Tag is referred to the information in the reference that is used to cite that article in the body of the document e.g. “(http://www.activated.org/books/chapter.php?id=17.10.)”, “(Afzal, 2008)” etc. For example if we want to find semantic relatedness for a cited paper which has been referred at number 10 in the cited-by article. The citation tag is searched within the text of the cited-by article. The system automatically calculates the number of times; the cited paper has been referred in the text of the cited-by article. In this way, the system build a relation {cited paper, Cited-by paper, citation frequency}. Subsequently, the relation is arranged in descending orders with respect to citation frequency. Sometimes, the cited paper was referred more than 20 times in the cited-by paper’s text. We manually explored the cited paper and all of its citations to see the semantic relationship between cited and cited-by paper. The results shows that if the cited-by paper refers the cited paper more than the threshold value, then both of the papers have a strong semantic relation.

Our propose technique is supported by empirical results that we achieved after performing experiment for an online Journal such as: Journal of Universal Computer Science (J. UCS) dataset. Based on our results we argue that Citation Tag frequency could be another filter which can be easily extended to finds the semantic relationship between the articles. We used J. UCS dataset containing about 1200 documents. We found Citation Tag frequency of their references in each document and we listed about 70 papers in descending order based on Citation Tag Frequency. We found that when Citation Tag frequency is higher, we are almost certain that these two documents are highly related. However, when the Citation tag frequency is not too high, we can make decision that these two documents are not strongly related.

In section 2 of this document the related work has been discussed, section 3 describes in detail proposed technique and in section 4 shows the results that provides proof of the efficiency of our proposed technique. Section 5 contains future directions and conclusion.

Related work:

Various approaches in literature have been discussed for finding document relatedness like Text-Based (Ratprasartporn and Ozsoyoglu, 2007; Giles et al., 1998) relatedness. In this approach text of the two papers are extracted and similarity between texts is calculated. The more the text is similar the more the documents are related.

Context based technique (Ratprasartporn and Ozsoyoglu, 2007) have proposed and described to find related papers. The context is ontological term which characterizes one more topic. The papers are grouped in contexts and the relatedness among the contexts is determined using various approaches like Count-Based Context Relatedness, Text-Similarity-Based Context, relatedness, Structure-Based Context Relatedness. Once the relatedness among the context is determined the paper to context relationship is explored by finding top-k relevant context to the paper. Afterward representative context is determined to find out relationship between the papers. The authors conclude that by using the measures of context-to-context and paper-to-context relatedness we can use and define the relatedness between papers is basically equal to the relatedness between two sets of selected contexts that represent the papers.

Citation based (co-citation (Small, 1973) and bibliographic coupling (Kessler, 1963)) techniques have also been studied in literature to find out papers relatedness. In co-citation the papers P1 and P2 is highly related if P1 and P2 are co-cited by larger numbers of documents. Similarly in bibliographic coupling two papers P1 and P2 is highly relevant when P1 and P2 have larger numbers of common citations.

Citeseer is one of the most widely know citation mining engine. CiteSeer uses different approaches to find related papers. Before applying any technique CiteSeer normalize certain aspects to improve the results like Conversion to lowercase, Removal of hyphens, Removal of citation tags, Expansion of common abbreviations e.g. conf.–conference, etc. The CiteSeer uses three approaches (Giles et al., 1998) in order find relatedness between the papers 1) word vectors: which is use to find the term similarity between documents. 2) String distance: CiteSeer uses Likeit string algorithm to find similarity between the headers of the documents. Header of the documents is simply the text before the abstract section 3) and citations: in which common citations between the documents is found to find the relatedness between the documents.

(Afzal et al., 2007) In his paper the authors are focusing on finding the related papers using his proposed technique. The idea published with name creating links into the future. The Author focuses on extracting Authors, Title and Venue information from citation. After that venue is disambiguated in and focus is on

1600
specific venue papers. Then direct matching over the current paper and the paper from venue is performed and if the match is successful he creates a link. When the direct match fails then he applies partial matching over the title and then using some threshold value to verify the comparison results. If the matching value greater then threshold value he finds the paper are related and in case of fail comparison he once check the value of authors.

All these above approaches are kind of unintelligent techniques. They only find the relatedness between the papers based on topic, references, metadata and text similarity etc. These techniques do not have the capability to show any kind of semantic relationship between papers. While our approach is different in sense that objective is not only to find related document. We are also focusing to find semantic relationship (the cited work is extended, the cited work is mentioned just background study, the cited paper work is negated etc) between the papers. We are extracting the referencing and extracting the citation Tag out of them. The references are handpicked and therefore it has high probability that those references will have high relevance with the paper itself. Furthermore when we find the Citation Tag frequency for each references it gives us a clear picture “How one paper is related to another one?”

Propose Technique:

The overall architecture of Citation Tag Frequency Technique is shown in Fig 1. The objective of our technique is to find is to find the most relevant papers from the citations of the focused paper. Our system consists different modules 1) Document Fetcher 2) Document Parser. The Document Parser itself is divided into three sub modules 2.1) Citation Tag Identifier 2) Section Identifier and 3) Citation Tag Frequency Calculator. Document is responsible for retrieving documents detail (document metadata & document text). The main module is Document Parser. Citation Tag Identifier in Document parser identifies Citation Tag from document reference list. The Section Identifier in Document parser divides the document in section and Citation Tag identifier result and Section identifier result are passed to Citation Tag Frequency Calculator. The Citation Tag Frequency Calculator calculates citation frequency and the results are persisted for further analysis. The details of each module are explained below:

**Fig. 1:** System Architecture

**Document Fetcher:**

Document fetcher retrieves document from the dataset. The fetched information consists of references from the database and its corresponding .txt file and then the document is passed over to the parser.

For this experiment, we selected the dataset of Journal of Universal Computer Science (J. UCS). The reason for this selection is based on the fact that J. UCS publishes articles in all domains of computer science. In this way, we will not only be able to find relevant papers within the focused domain, instead we will be able to find relevant paper across different domains. In our previous experiments (Afzal et al., 2007) we devised a technique known as TIERL to discover citations for articles published within JUCS. The discovered citations are further mined to find most relevant cited-by papers for the focused paper. This dataset is based on database and document files. The database contains metadata for each publication such as: publication title, Publication ID and Publication references list. In our previous work (Afzal et al., 2007), we converted all PDFs
to text documents files. These text document files are further mined to discover the semantic relationship between documents.

**Document Parser:**

The document fetcher send document and reference list to the Document parser. The document parser is sole responsible for mining the document text and discovering semantic relationship between documents. The document parser is further subdivided into three sub-processes: 1) citation tag identifier 2) Section Identifier and 3) Citation Tag Frequency Calculator. The citation tag identifier recognizes citation tag in the reference list of the paper such as (Afzal et al, 2010) or (Ratprasartporn and Ozsoyoglu, 2007). The Section Identifier identifies the sections in the document such as: introduction, related work, conclusion etc. The Citation Tag Frequency Calculator calculates the citation frequency for a Citation Tag both in the whole document and within each section identified by section identifier. These sections are explained below in details.

**Citation Tag Identifier:**

The Citation Tag is the information in the reference that is used to cite that article in the body of the document e.g. “(http://www.activated.org/books/chapter.php?id=17.10.)”, “(Afzal 2008)”. This module parse references list of document one by one and find out the citation tag pattern. After manual inspection of the dataset, we found that different authors use different citation tags such as: 20, (Afzal 2009) and (http://www.activated.org/books/chapter.php?id=17.10.). We considered all of these patterns in our implementation. We faced some problems, sometimes citation tag was written as “[Afzal 2010” instead of “[Afzal et al 2010]” and some time “[Afzal, 2010]” instead of “[Afzal 2010]”. As we verified all the selected citation manually and thus we are sure that our citation frequency result are error free. Once Citation is ready then it is passed to Citation Tag Frequency Calculator.

**Section Identifier:**

The section Identifier module identifies the sections in the text of the document. The sections of documents could be e.g. Introduction, Related work, Conclusion etc. We identify the section with help of regular expressions. We analyzed the documents and we found that there is usually double empty when new paragraph starts. Therefore, we used “\n\n” regular expression to find paragraph. Then to find the sections we observed that the sections are generally get starts with number and then space and there might be period (.) sign and then section label. And thus we used the “^[0-9]\s\w +.*” regular expression to find the section labels. We found that most of the time it identifies properly but in some cases it fails which is mainly because the text file of document is not properly built. However, in our case we manually checked all the results so that we sure about our correctness of computed data.

**Citation Tag Frequency Calculator:**

The Citation Tag Frequency Calculator receives citation tags from Citation Tag Identifier and section information from Section Identifier. Subsequently, it finds citation Tag in the running text of various sections. At the end, this module list citation frequency in the whole text and within different sections.

**Citation Tag Frequency Data:**

The Parser then populates the data in database for further usage. The populated consists of (Current Document, Citation Tag, Total Frequency, Section, Section Citation Tag Frequency). Currently we manually analyzed the populated data to deduce the reason for extraordinarily citation Tag frequency in document.

**Document parser Algorithm:**

The process of building of citation tag frequency to build summary data is summarized Figure 2.

Algorithm ComputeCitationTagFrequency(Documents)
1. for each ‘document in DocumentDataSet’ do
2. get ‘Text’ Current Document
3. get ‘References’ For Current Document
4. get ‘All Sections’ in Document Text
5. for each ‘Reference’ do
6. set Current_Citation_Tag_Section_Frequency = 0
7. set Current_Citation_Tag_Document_Frequency = 0
8. get ‘Citation Tag’ From Current Reference
9. get Current_Citation_Tag_Document_Frequency.
10. for each ‘Section in Text’ do
11. get Current_Citation_Tag_Section_Frequency
12. get ‘Section Label’
13. Persist Citation Information
   (Current document ID,
    Citation Tag,
    Current_Citation_Tag_Document_Frequency,
    Section Label,
    Current_Citation_Tag_Section_Frequency)
14. end
15. end
16. end

Fig. 2: Algorithm

The data is computed by algorithm and we have crossed verify it manually. We then select the paper and read to identify the reasons “Why a particular reference has got high citation frequency?” The results are shown in the result section of the paper.

RESULTS

We used J. UCS dataset containing about Total Documents Metadata Available of 1460 documents. The Text Document we used (pdf & txt) vol. 14, Issue. 6 (2008). The total references we parsed were 16404. We found Citation Frequency in each document. We selected and listed about 70 papers in descending order based on citation Tag Frequency. The maximum a Citation Tag frequency in a document was 20. The next was 17 and then some were having 16, 15 &15 and so on. We manually analyzed that why an article refers so many times? We found various citation reasons shown in Table 1. Once analysis was performed we found a threshold value 5 than can be used to determine the semantic relationship between the articles. The summarized results are shown in Table 2 & Table 3. The Table 2 summarized results consisting of all those Citation Frequency in Cited-By Article whose value is greater than 5. Similarly the Table 3 shows all those results whose Citation Frequency in Cited-By Article is less than or equal to 5.

The first column in Table 1 & 2 shows the Total Number of papers and the Second column shows the Reasons. The results should be interpreted like R-1 found in Total Number of 14 papers having Citation Frequency more then 5. The above results show that cited-by document & cited document has strong semantic relationship for Citation frequency more than 5. And the cited-by & cited document are not strongly related if the Citation Frequency is less than or equal to 5.

Table 1: Citation Reasons

<table>
<thead>
<tr>
<th>Citation Reason ID</th>
<th>Citation Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>R-1</td>
<td>Definition &amp; Theories usage</td>
</tr>
<tr>
<td>R-2</td>
<td>Extension, Variant</td>
</tr>
<tr>
<td>R-3</td>
<td>Comparison</td>
</tr>
<tr>
<td>R-4</td>
<td>Background, Related Work or Homepage for Pioneer</td>
</tr>
</tbody>
</table>

Table 2: Citation Frequency in Cited-By Article > 5

<table>
<thead>
<tr>
<th>Total Number Papers</th>
<th>Reasons</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>R-1 (definition used)</td>
</tr>
<tr>
<td>31</td>
<td>R-2 (extension/variant)</td>
</tr>
<tr>
<td>4</td>
<td>R-3 (Comparison)</td>
</tr>
<tr>
<td>0</td>
<td>R-4 (Background/Related)</td>
</tr>
</tbody>
</table>

Table 3: Citation Frequency in Cited-By Article <= 5

<table>
<thead>
<tr>
<th>Total Number Papers</th>
<th>Reason</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>R-1 (definition used)</td>
</tr>
<tr>
<td>0</td>
<td>R-2 (extension/variant)</td>
</tr>
<tr>
<td>0</td>
<td>R-3 (Comparison)</td>
</tr>
<tr>
<td>16</td>
<td>R-4 (Introduction/Related)</td>
</tr>
</tbody>
</table>
Conclusion & Future work:
Finding relatedness between scholarly knowledge is a crucial task. Currently numbers of systems are available to find related papers based on number of analysis such as: co-citation, bibliographic, content based, context based etc. However, less attention has been paid to find semantic relationship between the articles. We have analyzed the text of cited-by paper and found the citation frequency for the cited paper. We exercised our approach on J. UCS corpus and our result shows that for all papers that have been cited more than five times in the cited-by paper text has strong semantic relationship.

We hindered by the status of “Citation Frequency” results from higher to lower in taking decision. We were not sure that what decision at which level of Citation frequency can be taken? Then we decided to find sections in whole document which will help us in decision making. We found that keeping track of citation tag frequency in section can help us in deciding that “How two documents are related to each other?”

In future, we will extend the dataset and will focus to assign some weights to different sections of the cited-by paper. For example, in the text of the cited-by paper, the citation in the result section will have more value than the citation in the related work section.

ACKNOWLEDGEMENT
This is the extended version of our own paper presented and published as Conference proceedings in “International Conference on Computers & Emerging Technologies” (ICCET 2011) held on 22-23 April 2011 at Shah Abdul Latif University, Khairpur, Sindh, Pakistan.

REFERENCES
CiteSeer, http://citeseer.ist.psu.edu/
ISI Web of Knowledge, http://www.isiknowledge.com/