ARTICLE RANKING WITH CITATION CONTEXT ANALYSIS

by

Metin Döşlü

B.S., Computer Engineering, Boğaziçi University, 2009

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Computer Engineering
Boğaziçi University

2013

ACKNOWLEDGEMENTS

I would like to thank my thesis supervisor Assoc. Prof. Haluk Bingöl for his invaluable guidance, helping during the preparation of this dissertation and especially for his patience. I would also like to thank my committee members Assist. Prof. Arzucan Özgür and Assoc. Prof. Şule Gündüz Öğüdücü for their contributions to this work.

I want to deeply thank all my fellow friends, especially Mahmut Cengiz Uzuntaş, Özgün Özer, Fatma Başak Aydemir, Ezgi Çebi, Fırat Çorapçılar, Yusuf Reyhani, Abdulkadir Yazıcı, İlkay Ozan Kaya, Elif Dede and Gözde Dinç for bearing my thesis talks and their support.

I would like to express my sincere gratitude to Gözde Kaymaz for sharing most of this journey, course projects to eventually this thesis. Your friendship was invaluable.

I would like to express my gratitude to my mother Fethiye Döşlü and my father Mehmet Döşlü for their understanding and support during my long working hours. Last but not least, I also would like to express my sincere gratitude to my sister, my first teacher Semra Döşlü, everything started with the arithmetical operations you taught me.

ABSTRACT

ARTICLE RANKING WITH CITATION CONTEXT ANALYSIS

It is hard to detect important articles in a specific context. Information retrieval techniques based on full text search can be inaccurate to identify main topics and they are not able to provide an indication about the importance of the article. Generating a citation network is a good way to find most popular articles but this approach is not context aware. The text around a citation mark is generally a good summary of referred article. So citation context analysis presents an opportunity to use the wisdom of crowd for detecting important articles in a context sensitive way. In this work, we analyze citation contexts to rank articles properly for a given topic. The model proposed here uses citation contexts in order to create a directed and weighted citation network based on the target topic. We create a directed and weighted edge between two articles if citation context contains terms from the term set we created for the target topic. Then we apply common ranking algorithms for the vertices of network. We showed that this method successfully detects the most prominent articles in a given topic. The biggest contribution of this approach is that we are able to identify important articles in the target topic even though they don't contain the term represents the interested context.

ÖZET

ATIF METNİ ANALİZİ İLE MAKALE SIRALAMA

Belirli bir konuda makaleleri önem sırasına göre sıralamak zordur. Tam metni tarama üzerine kurulu olan bilgi çıkarım teknikleri makalelerin ana konularını tespit etmekte çok başarılı değillerdir. Ayrıca bu bilgi çıkarım teknikleri makalenin önemi konusunda bir çıkarımda bulunamazlar. Bir atıf ağı oluşturmak, en çok önemli olan makaleleri bulmak için iyi bir yöntem olabilir, ancak bu yöntem de makalelerin konusu ile bir ilişki kuramamaktadır. Bir atıf işaretçiğinin çevresindeki metin genellikle atıfta bulunulan makalenin iyi bir özetidir. Bu nedenle, atıf metni analizi konu bazında makaleleri önem sırasına göre sıralamak için araştırmacıların konu üzerindeki genel kanısını kullanmak için bir fırsat sunar. Bu çalışmada verilen bir konu içindeki makaleleri önem sırasına göre sıralamak için atıf metinlerini analiz ediyoruz. Burada sunduğumuz model, hedef konu üzerine kurulu olan yönlü ve ağırlıklı bir atıf ağı kurmak amacıyla atıf metinlerini kullanıyor. Eğer atıf metni hedef konu için oluşturduğumuz terimler grubundan herhangi bir terimi kapsarsa, bu iki makale arasında atıf verenden atıf alana doğru yönlü ve ağırlıklı bir çizgi oluşturuyoruz. Bundan sonra bu ağın çizgeleri sıralamak için genel olarak kullanılan ağ üzerindeki çizge sıralama algoritmalarını kullanıyoruz. Deneylerimizin sonucunda, bu metodun verilen bir konuda en önemli makaleleri üst sıralarda sıraladığını gösterdik. Önerdiğimiz yaklaşımın en büyük katkısı ise verilen bir terim için ilgili makaleler bu terimi içermese dahi sistemimizin bu makaleleri de sırayabilmesidir.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS . . . iii
ABSTRACT . . . iv
ÖZET . . . v
LIST OF FIGURES . . . vii
LIST OF TABLES . . . viii
LIST OF SYMBOLS . . . ix
LIST OF ACRONYMS/ABBREVIATIONS . . . x
1. INTRODUCTION . . . 1
2. RELATED WORK . . . 4
2.1. Citation Context . . . 4
2.2. Network Generation . . . 6
3. METHODOLOGY . . . 7
3.1. Theory . . . 7
3.1.1. Citation Network . . . 8
3.1.2. Term-$\boldsymbol{\alpha}$ Specific Citation Network . . . 10
3.1.3. Term Similarity . . . 12
3.1.4. Similar Term Set Citation Network . . . 15
4. EXPERIMENTS AND RESULTS . . . 17
4.1. Dataset . . . 17
4.2. Citation Network for the Term “power law” . . . 18
4.3. Citation Network for the Term “hadoop” . . . 20
5. CONCLUSION . . . 23
6. FUTURE WORK . . . 25
APPENDIX A: LINK ANALYSIS METHODS . . . 26
A.1. PageRank . . . 26
A.2. HITS Algorithm . . . 27
APPENDIX B: IMPLEMENTATION . . . 29
REFERENCES . . . 33

LIST OF FIGURES

Figure 1.1. Part of an Article About Information Retrieval. 2

Figure 3.1. Citation Network of Example Fake Articles. 9

Figure 3.2. Citation Networks. 11

Figure 3.3. Overlapping Terms. 12

Figure 4.1. Results for “power law” on Google Scholar. 19

Figure 4.2. Results for “power law” on CiteSeerX. 20

Figure 4.3. Results for “hadoop” on Google Scholar. 21

Figure 4.4. Results for “hadoop” on CiteSeerX. 22

Figure B.1. Get Unique Keywords. 31

Figure B.2. Get Citation Related with a Given Bigram. 31

Figure B.3. Get Target Cluster and Source Paper for a Given Citation. 32

Figure B.4. Get Source Cluster. 32

LIST OF TABLES

Table 4.1. Similar Terms for “power law”. 18

Table 4.2. Article Ranking of Citation Network for “power law” term. 18

Table 4.3. Statistics of Similar Term Set Citation Network for “power law”. . 19

Table 4.4. Bigram Rankings in the “power law” Citation Network. 20

Table 4.5. Article Ranking of Citation Network for “hadoop” term. 21

Table B.1. Papers table: Stores metadata and access information for each paper. 29

Table B.2. Citations table: Citations found in a paper. 30

Table B.3. CitationContexts table: Contexts which contain citation. 30

Table B.4. Keywords table: Keywords found in a paper. 30

LIST OF SYMBOLS

a_i	Article with index i
A	Set of all articles
C	Set of edges in citation network
C_α	Set of edges in term specific citation network for the term α
G	Citation network
G_α	Term specific citation network for the term α
G_{S_α}	Similar term set citation network for the term α
S_α	Similar terms set for the term α
T	Set of all terms
T_{ij}	Term set extracted from the citation context in article i which has a citation to article j
α	A term in set T
δ	Threshold value for similarity scores

LIST OF ACRONYMS/ABBREVIATIONS

CPA	Co-citation Proximity Analysis
GFS	Google File System
HITS	Hyperlink-Induced Topic Search
RDI	Reference Directed Indexing

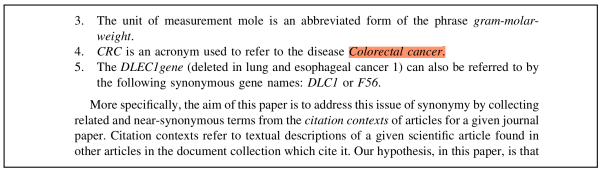
1. INTRODUCTION

A researcher needs to know about related work before starting to study on a topic. In this context, citation indexes such as CiteSeerX[1] are very useful to navigate through related research articles. Some of the citation indexes provide a medium to search over full text of articles. Citation indexes are also able to index articles without access to full text with the help of articles cite them. They also provide a way to evaluate importance of an article because they also report the number of times the article is cited. However still it is an exhaustive work to scan scientific literature to find important articles in the interested topic.

Citations provide a way to measure the relative impact of articles in a collection of scientific literature. A *citation network* is formed by arcs between articles where there is an arc from article i to article j if and only if article i contains a citation to article j. Forming a citation network and ranking articles according to their incoming degree on this network helps to identify important articles over all articles. However it is not an easy task to determine prominent articles in a specific context. Text of an article would contain lots of words not related with its main topic. These words would be used in examples, controversy arguments etc. Search methods which use indexing techniques on full text suffer from these problems. Figure 1.1 shows a part from an article which is about information retrieval [1] and this text part contains a term about "cancer". So any indexing technique which uses full text will index this article also for the term "cancer" although this term is not related with the main argument of the article. Indexing techniques also do not have any information about the importance of the target articles.

Citation networks provide a way to detect the most outstanding articles in the scientific literature. However ranking obtained from a citation network does not contain context information, so one still needs to differentiate them according to the context. Combining indexing techniques with a citation network would be a good approach to

[1] http://citeseerx.ist.psu.edu



3. The unit of measurement mole is an abbreviated form of the phrase *gram-molar-weight*.
4. *CRC* is an acronym used to refer to the disease *Colorectal cancer*.
5. The *DLEC1 gene* (deleted in lung and esophageal cancer 1) can also be referred to by the following synonymous gene names: *DLC1* or *F56*.

More specifically, the aim of this paper is to address this issue of synonymy by collecting related and near-synonymous terms from the *citation contexts* of articles for a given journal paper. Citation contexts refer to textual descriptions of a given scientific article found in other articles in the document collection which cite it. Our hypothesis, in this paper, is that

Figure 1.1. Part of an Article About Information Retrieval.

try, but this hybrid method still suffers from terms not related with the main topic of the article. According to our observation on academic literature search engines, we can say that they use such hybrid systems which consist of full text indexing and citation network as a part of their systems.

A *citation context* is essentially the text surrounding the reference markers used to refer to other scientific works [1]. Length of the citation context can be defined as a predefined number of sentences, words or characters around the citation marker. Citation context provides an useful way to identify the main contributions of an academic publication because authors refer articles by briefly presenting main points of the cited article in citation context. To be cited in an article with specific terms is a significant indication of importance in that topic. More the article is cited with same terms means that this article is more important in the topic which these terms represent.

Citation context is generally formed by the explicit and definitive words that the citing author uses to describe the cited work. Most of the time, the citation context is a very good summary of the cited article or points to some important highlights about it. In other words, citation context contains representative keywords for the cited work. Citation context analysis provides us an opportunity to reason about main topics of the cited articles even though we do not have the contents of articles. Schwartz *et al.* stated that it is hypothesized that through time, the citation context can more accurately describe the most important contributions of an article than its original abstract [2].

It is observed that the link structure formed by citations is analogous to that of

the web, where the links are hyperlinks between web pages. This motivated us to use widely used network analysis methods for web link analyse such as HITS and PageRank to apply on our dataset.

In this work, we present a method to find important articles for a given topic. For this purpose we use terms extracted from citation contexts to create context aware citation networks. Then on these networks, we find the most outstanding articles by using common network analysis methods.

Utilizing citation contexts helps us to find articles especially in the following situations where there is no way to pinpoint them using full text indexing methods:

- Suppose that an article proposed a concept, and later someone else build another concept on the top of this concept. Such as Hadoop was derived from Google File System (GFS) [3] and Google's MapReduce [4] articles. In such situations, if you are looking for important articles for the second concept such as Hadoop, you also would to see articles related with first concept such as MapReduce.
- Suppose that there are closely related concepts such as "power law" and "small world". There can be some articles which talk about "small world" but never mention "power law". In such situations, if you are looking for important articles for "power law", you also would to see articles related with "small world".

2. RELATED WORK

2.1. Citation Context

Bradshaw used citation contexts to index cited papers which they call their method as Reference Directed Indexing (RDI) [5]. The main motivation behind this article is that in a citation context authors describe an article with similar terms to a search query used to search an article for that context. For the citation context, they used approximately 100 words long with 50 words on either side of the point of citation. Then they created a list of index terms for the cited article from citation contexts it is cited. As the number of citation contexts increases which refers to the given article with a specific term then score of this index term for the given article increases. After term indexes for all articles are created in the dataset, for a given query they first check articles which contain all the terms of query in their index list, and then rank them according to their index scores. They tested their results by checking how many relevant documents returned by their search engine based on RDI in the top ten and compared their results with a full text similarity based index search method.

Research of Bradshaw is the closest study to ours. We also used citation context to rank papers for specific topics. Differently, we created a network of articles instead of indexing citation context. Besides, we used bigrams instead of single terms with the motivation that bigrams are more descriptive to define specific research areas. We cross checked bigram list we get from citation context with the keyword list of all articles, so we get rid of meaningless bigrams in citation contexts. Another contribution of ours is that instead of just using search term we also search for similar terms in citation contexts while forming our citation network. This increases accuracy and robustness of our system.

Ritchie *et al.* discussed similarities between the web and scientific literature, making an analogy as hyperlinks between web pages alongside citation links between articles [6]. They mentioned that there are fundamental differences like greater vari-

ability of web pages and the independent quality control of academic texts through the peer review process. They stated that the analogy between hyperlinks and citations is not perfect because the number of hyperlinks varies from web page to web page where the number of citations in papers is somehow restricted. Aljaber *et al.* also makes an analogy between citation context in scientific articles and anchor text in web pages [1].

It is shown that citation contexts can be used to cluster documents [1]. They used citation term representation for each article which is generated from all its citation contexts found in the dataset. For every article, they automatically extracted all of the citation sentences that other articles used to cite to it. Then, they used this citation term representation in order to cluster articles.

It is also shown that citation contexts can be used to summarize articles [7]. They extracted significant keyphrases from the set of citation contexts where keyphrases are expressed using n-grams. Then, they used these key phrases to build the summary.

In a very previous work, citations of scientific articles are classified according to whether they are conceptual or operational, organic or perfunctory, evolutionary or juxtapositional, confirmatory or negational [8].

Although there are variety of academic works that focus on citation contexts, these efforts were relatively on small datasets. For example Bradshaw used 10,000 articles [5] and Ritchie *et al.* used 9,000 articles [6]. Most of the bigger datasets are not well structured and require lots of preprocessing and manual work. Problems coined are generally unsupervised and evaluation of results requires manual work. This makes infeasible to evaluate large result sets. Differently, we used relatively larger dataset that contains around 1.8 million articles, but we still have the problem of evaluating large result sets due to unsupervised nature of our quest. For example, in order to evaluate results Bradshaw listed top 10 articles for every test query they run on both their system and comparison systems. Then, they mixed results and manually checked the relevance of articles in result set without knowing which system found which articles [5].

According to Aljaber *et al.* using terms around citation references with a predefined window size is a simple but effective way to determine useful terms [1]. They tried different window sizes and found that 50 words before and after the citation reference is optimal citation context size for document clustering on their datasets. Similarly Bradshaw also used citation contexts of 100 words length extracted from articles with 50 words on both sides of the citation mark, in this case in order to index the cited articles. Citation contexts in our dataset consist of around 400 characters which are equally divided to both sides of citation marker.

2.2. Network Generation

One can use different methods to create a network over an article dataset.

Bibliographic coupling occurs when two articles have a common citation to same third article. Number of common citations can be used to create an undirected weighted edge between these two articles. The main assumption behind bibliographic coupling, introduced by Kessler, is that similar articles have similar references [9]. Two articles are *bibliographicly coupled* if and only if they cite the same article. An undirected weighted network is obtained by the number of common such articles.

Another way to create a network from papers is to use co-citation analysis. The co-citation count for two articles A and B is the count of articles where articles A and B cite together [10]. We can generate an undirected and weighted network by creating edges between articles using co-citation counts. The main assumption behind co-citation analysis is that similar articles are cited together more frequently. Gipp *et al.* introduced an extended approach termed *Co-citation Proximity Analysis (CPA)* on the top of co-citation analysis [11]. CPA considers the proximity of citations within an article with basic assumption that two articles are more similar if citations to them appear closely. Then, we can calculate weight of an edge between two articles with a function of proximity of citations.

3. METHODOLOGY

3.1. Theory

We need a methodology which is good both for relevance and significance. An article identifies the main contributions of the cited article and uses related terms when citing this article. This gives us invaluable information about the relevance of cited articles with the interested topic. Heavily cited articles with related terms generally mean important contributions in the topic of interest, so more the citation count means more significant the cited article.

The citation context of a citing article may have many possible meanings: it may be off topic or it may convey criticism rather than approval. It is hard to determine the intent of the citation context automatically. But in aggregate, if an article is cited by many articles with the same terms, then it is receiving a kind of collective confirmation in the area of this term represents. We can extract cumulative understanding of the crowd for the cited article from cumulative citation contexts of citing articles.

Best articles are cited by many articles so we decided to follow a simple scheme for determining defining citation terms in citation context. The words that are used to describe the cited paper will stand close to the citation marker, so we used a window of fixed size for citation context like previous studies [5], in our case which is 400 characters long.

We used every bigram in citation context as a defining term for all articles in this citation context. For example, if three articles are cited in a citation context, we pick every bigram in this context as defining terms for all these three articles. We can trivially extend this system to use also other n-gram types but we think that bigrams are the mostly used n-gram types to describe specific terms such as “scale free”, “map reduce”, “preferential attachment” etc. This helped us easily to extract meaningful terms from citation context, where other techniques which only use single words suffer

from problems like synonyms.

We created a simple query engine which takes a bigram and returns an ordered list of scientific articles. This list is ranked according to impact of articles on the scientific community in the topic related with a given bigram where the details are given shortly.

Here we show step by step how we form a context sensitive network for a given term:

3.1.1. Citation Network

Citation context is the text around the citation marker. The size of this text can be defined as a specific number of sentences, words or characters around the citation marker. We can form a citation network, a directed graph, from citation information of articles by creating directed edges from citing article to cited articles. Actually, an edge in a citation network carries more information than just a single binary relation. We can extract terms from citation context which author used in order to explain the cited document.

Definition 3.1. *A* term *is a word or group of words describing a thing or expressing a concept, especially in a particular field.*

Example 3.1. *A small example of citation network formed by six fake articles is shown in Figure 3.1. In the figure, citation contexts are underlined and extracted terms with cited articles are highlighted.*

Let $A = \{a_1, a_2, ..., a_{|A|}\}$ be the set of all articles. We use also lower case Latin letters for the elements of A such as $i, j \in A$.

Let $T = \{\alpha_1, \alpha_2, ..., \alpha_{|T|}\}$ be the set of all terms used in all articles inside A. We use also lower case Greek letters for the elements of T such as $\alpha, \beta \in T$. We selected elements of T for our experiment from article keywords.

Figure 3.1. Citation Network of Example Fake Articles.

Definition 3.2. *$T_{ij} \subset T$ is the set of all terms that appear in at least one citation context in article i with a reference to article j.*

Note that if there is no citation from article i to article j, then $T_{ij} = \emptyset$. It is possible that article i cites article j but the citation context has no term in it, then $T_{ij} \neq \emptyset$. It is also possible that article i cites article j more than once. A term can be presented in all of these citations. It is also possible that it occurs only one of them. In either case the term is in T_{ij}.

Definition 3.3. *A* term labelled citation network *is a directed graph $G(A, C)$ with $C = A \times A$ where $(i, j) \in C$ if and only if article i has at least one citation that refers to article j. The edge $(i, j) \in C$ is labelled with all terms in T_{ij}.*

Example 3.2. *An example of term labelled citation network with $A = \{a_1, a_2, ..., a_6\}$ and $T = \{\alpha_1, \alpha_2, ..., \alpha_5\}$ is given in Figure 3.2(a). Arcs are also labelled, for example $T_{1,3} = \{\alpha_1, \alpha_4, \alpha_5\}$ and $T_{1,2} = \emptyset$. This term labelled citation network is formed from example citation network at Figure 3.1.*

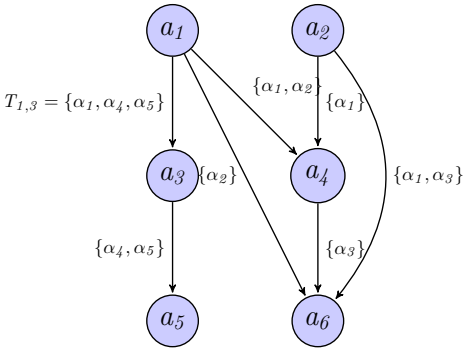
3.1.2. Term-α Specific Citation Network

We can define a term specific citation network by using terms describing edges on the citation network. Then, we can run standard ranking algorithms on this network and find important articles for this topic. Here is the definition of term specific citation network:

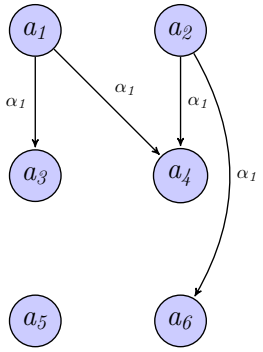
Definition 3.4. *Let $\alpha \in T$ be a term. The subgraph $G_\alpha(A, C_\alpha)$ of $G(A, C)$ is called* term-α specific citation network *where $C_\alpha \subset C$ and $(i, j) \in C_\alpha$ if and only if $\alpha \in T_{ij}$.*

Note that the citation network $G(A, C)$ is the union of all term specific citation networks $G_\alpha(A, C_\alpha)$ where $\alpha \in T$. That is,

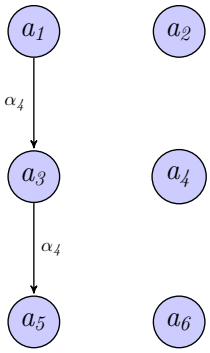
$$G(A, C) = \bigcup_{\alpha \in T} G_\alpha(A, C_\alpha).$$



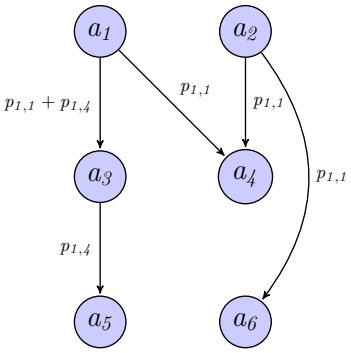
(a) $G(A, C)$.



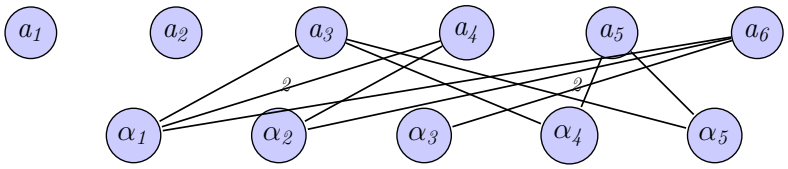
(b) $G_{\alpha_1}(A, C_{\alpha_1})$.



(c) $G_{\alpha_4}(A, C_{\alpha_4})$.



(d) $G_{S_{\alpha_1}}(A, C_{S_{\alpha_1}})$.



(e) Article-Term Citation Bipartite Graph (Weight of unlabelled edges are equal to *1*).

$$F = \begin{array}{c} \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \end{array} \begin{array}{c} \begin{array}{cccccc} a_1 & a_2 & a_3 & a_4 & a_5 & a_6 \end{array} \\ \begin{pmatrix} 0 & 0 & 1 & 2 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 2 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \end{pmatrix} \end{array}$$

(f) Term Document Matrix F.

Figure 3.2. Citation Networks. (a) Term labelled citation network $G(A, C)$. (b) Term-α_1 specific citation network $G_{\alpha_1}(A, C_{\alpha_1})$. (c) Term-$\alpha_4$ specific citation network $G_{\alpha_4}(A, C_{\alpha_4})$. (d) Similar term set citation network $G_{S_{\alpha_1}}(A, C_{S_{\alpha_1}})$ for the similar term set $S_{\alpha_1} = \{\alpha_1, \alpha_4\}$. (e) Article-term citation bipartite graph. (f) Term document matrix F.

Example 3.3. *Figure 3.2(b) and Figure 3.2(c) are term specific citation networks for terms α_1 and α_4, respectively.*

3.1.3. Term Similarity

A term is not generally enough to describe fully a topic in scientific literature by itself and just using a single term is open to noises because of natural language usage such as synonyms etc. For every term, there is a set of articles covered by it. These sets of articles hugely overlap for some of the terms as it is seen in Figure 3.3 for the terms "power law", "scale free" and "preferential attachment".

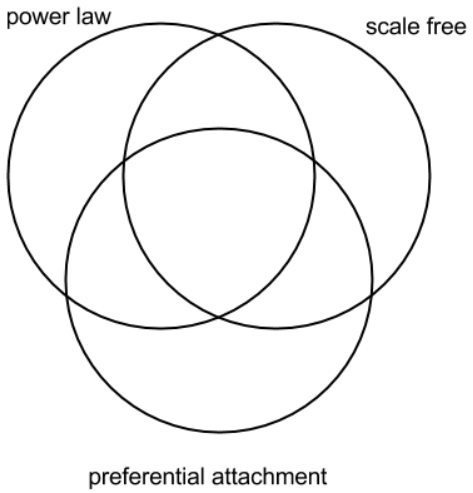


Figure 3.3. Overlapping Terms.

Definition 3.5. *For a given term α,* similar terms *are the set of terms which have a high level of overlapping article coverage with the article coverage of the given term.*

One of the key approaches of this work is that we use similar terms in the network creation process. This helps us to broaden the set of citation contexts we evaluate in the interested topic. Every one of these citation contexts was written by a researcher, so the size of crowd source increases which we utilize to rank articles.

First, we create a term-article frequency matrix, then shortly we will represent the steps we use to find similar terms for the target term.

Definition 3.6. *Term frequencies are related to articles by means of* term document matrix $\mathbf{F} = [f_\alpha j]$ *of size* $|T| \times |A|$*, where the entry* $f_{\alpha j}$ *is the count of how many different articles use the term* α *in related citation context to cite article* j*.*

$$f_{\alpha j} = \textit{indegree of article } j \textit{ in the graph } G_\alpha.$$

F is actually extracted from an undirected weighted bipartite graph between article nodes and term nodes.

Example 3.4. *An example of this undirected weighted bipartite graph for the term labelled citation network in Figure 3.2(a) is shown in Figure 3.2(e). Related term document matrix* **F** *is shown in Figure 3.2(f).*

We want to find similar terms, but especially the discriminative ones which are used to describe smaller set of articles. Simple term frequency has a problem that all terms are considered equally important, but certain terms have little or no discriminating power. For example, a collection of articles on the "cancer" is likely to have the term "cancer" in nearly all citation contexts. So we decided to scale down the weights of terms which occur in lots of citation contexts. In principle, the idea is reducing term frequency weight of a term by a factor that grows with its citation context frequency it appears and *term frequency-inverse document frequency* is a technique which is based on this idea [12]. This method is widely used in information retrieval and text mining and it reflects how important a word is to a document in a collection. So we decided to use it to weight term frequencies we have.

Inverse document frequency for the term α is defined by $g(\alpha)$

$$g(\alpha) = \log \frac{|A|}{\sum_{j=1}^{|A|} \operatorname{sgn}(f_{\alpha j})}$$

where $\text{sgn}(x)$ is sign function defined as

$$\text{sgn}(x) = \begin{cases} 1, & x > 0, \\ 0, & x = 0, \\ -1, & x < 0. \end{cases}$$

Then, let $\mathbf{D} = [d_{\alpha\beta}]$ be a $|T| \times |T|$ diagonal matrix defined by

$$d_{\alpha\beta} = \begin{cases} g(\alpha), & \alpha = \beta, \\ 0, & \alpha \neq \beta, \end{cases}$$

and we define *weighted term document matrix* $\mathbf{N} = [n_{\alpha\beta}]$ of size $|T| \times |A|$ by

$$\mathbf{N} = \mathbf{D} \times \mathbf{F}.$$

Then we need to build a relation between terms. Let $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ be the α^{th} and β^{th} row vectors of $\boldsymbol{N}$, respectively. Elements of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ show the respective weighted term frequencies of terms α and β for the articles in the dataset. If somebody wants to find out how much article coverages of these terms overlap, then he needs to compare corresponding row vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. For this purpose we decided to use sample Pearson correlation coefficient which is widely used in the sciences as a measure of the strength of linear dependence between two samples.

Then we define *sample Pearson correlation matrix* $\mathbf{P} = [p_{\alpha\beta}]$ of size $|T| \times |T|$ and $p_{\alpha\beta}$ is the sample Pearson correlation between term α and β

$$p_{\alpha\beta} = \frac{\sum_{i=1}^{|A|}(\boldsymbol{\alpha}_i - \bar{\boldsymbol{\alpha}})(\boldsymbol{\beta}_i - \bar{\boldsymbol{\beta}})}{\sqrt{\sum_{i=1}^{|A|}(\boldsymbol{\alpha}_i - \bar{\boldsymbol{\alpha}})^2}\sqrt{\sum_{i=1}^{|A|}(\boldsymbol{\beta}_i - \bar{\boldsymbol{\beta}})^2}}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the α^{th} and β^{th} row vectors of $\boldsymbol{N}$, respectively. $\bar{\boldsymbol{\alpha}}$ is the average of entries of vector $\boldsymbol{\alpha}$.

The *sample Pearson correlation coefficient* is a measure of the linear correlation between two samples X and Y, giving a value between *-1* and *1* inclusive. A value of *1* means that a linear equation describes the relationship between X and Y, with all data points lying on a line where Y increases as X increases. A value of *-1* means that all data points lie on a line for which Y decreases as X increases. This case is irrelevant for our dataset, because to get a value of *-1* for two terms α and β, they have to be complement of each other. This is not probable on a large collection of articles. A value of *0* means that there is no linear correlation between the samples.

Definition 3.7. *For a given term* α, similar term set S_α *is defined by*

$$S_\alpha = \{\beta \in T \quad | \quad p_{\alpha\beta} > \delta\} \text{ for some } 0 < \delta < 1.$$

Note that δ which is a cross validation parameter and this value changes among topics. In our experiments, we found the optimum value by trial and error.

Also note that similarity score $p_{\alpha\alpha}$ for the term α itself is equal to *1*. Therefore, $S_\alpha \neq \emptyset$ for all α, since $\alpha \in S_\alpha$.

3.1.4. Similar Term Set Citation Network

Now, we can define a directed and weighted citation network from similar term set S_α based on the term α as:

Definition 3.8. *The subgraph* $G_{S_\alpha}(A, C_{S_\alpha})$ *of* $G(A, C)$ *is called* similar term set citation network *where*

(i) $C_{S_\alpha} = \bigcup_{\beta \in S_\alpha} C_\beta$

(ii) weight w_{ij} *for the edge* $(i, j) \in C_{S_\alpha}$ *equals to similarity score weighted sum* $w_{ij} = \sum_{(i,j) \in T_{ij} \bigcap S_\alpha} p_{\alpha\beta}$.

Example 3.5. *Lets assume that we calculated similarity set as* $S_{\alpha_1} = \{\alpha_1, \alpha_4\}$ *for the term* α_1 *with a given* δ *from Figure 3.2(a). Then we draw similar term set citation*

network $G_{S_{\alpha_1}}(A, C_{S_{\alpha_1}})$ *for the term* α_1 *and it is shown in Figure 3.2(d).*

After forming a similar term set citation network for a given term α, we can run common ranking algorithms on this network and find the most important articles for the topic represented by the term α.

4. EXPERIMENTS AND RESULTS

We run our experiments over our dataset for two different terms “power law” and “hadoop” and here we reported the results.

4.1. Dataset

“SeerSuite is a framework for scientific and academic digital libraries and search engines built by crawling scientific and academic documents from the web with a focus on providing reliable, robust services. In addition to full text indexing, SeerSuite supports autonomous citation indexing and automatically links references in research articles to facilitate navigation, analysis and evaluation” [13]. CiteSeerX where we get our dataset is an instance of SeerSuite. This dataset is a snapshot of June 2012 and it contains nearly 1.8 million scientific articles and 41.5 million citation contexts extracted from these articles.

In our dataset, citation contexts are marked with their citations. On the other hand, we still need to identify which terms in a citing paper refer to which of its citations. This is not an easy problem due to nature of different citation techniques. Ritchie *et al.* stated some of the problems with matching terms in citation context with correct citations [6]:

- Length of text which refers to citations differentiate from article to article.
- If citation context blocks are physically very close to each other, then citation terms for different citations would be overlapped.
- Citation marks may appear in different places such as some citation texts start with citation mark while others end with citation mark.
- Some citation contexts may contain contradictory terms for cited articles.

4.2. Citation Network for the Term "power law"

First we get similar terms with the threshold value $\delta = 0.35$ for "power law" term and their similarity scores are shown in Table 4.1. We considered most popular 3,469 articles and 10,858 terms in similar term detection algorithm, because larger experiments were not computationally feasible. By using these terms, we created a similar term set citation network for the topic represented by "power law" term.

Table 4.1. Similar Terms for "power law".

Term	**Similarity Score**
power law	1.00
degree distribution	0.83
web graph	0.56
preferential attachment	0.45
scale free	0.38

Then we run widely used link analysis methodologies weighted in-degree, HITS [14] and PageRank [15] on this network in order to rank articles. Ranking results are reported at Table 4.2 for the articles which take place in top ten for all methodologies. As it is shown on the table, we are able to identify most prominent articles in "power law" related topic.

Table 4.2. Article Ranking of Citation Network for "power law" term.

Paper \Ranking Method		**In-Degree Weight**	**HITS-Authority**	**PageRank**
Reference	**Title**			
ref [16]	Emergence of Scaling in Random Networks	1	1	1
ref [17]	On Power-Law Relationships of the Internet Topology	2	2	2
ref [18]	Statistical Mechanics of Complex Networks	3	3	6
ref [19]	Collective Dynamics of 'small-world' Networks	5	4	9
ref [20]	The Structure and Function of Complex Networks	6	5	10
ref [21]	Alpha-Power Law MOSFET Model and its Applications to CMOS Inverter Delay and Other Formulas	4	6	10

Table 4.3 shows statistics about similar term set citation network the term "power law."

Table 4.3. Statistics of Similar Term Set Citation Network for “power law”.

Node Count	16,487
Edge Count	10,538
Average Weighted Degree	0.574
Network Diameter	5
Average Path Length	1.24

We also searched “power law” term on academic literature search engines and the top results are shown in Figure 4.1 for Google Scholar[2] and in Figure 4.2 for CiteSeerX. We can comfortably say that our results consist of more related and important articles for the topic represented by “power law”.

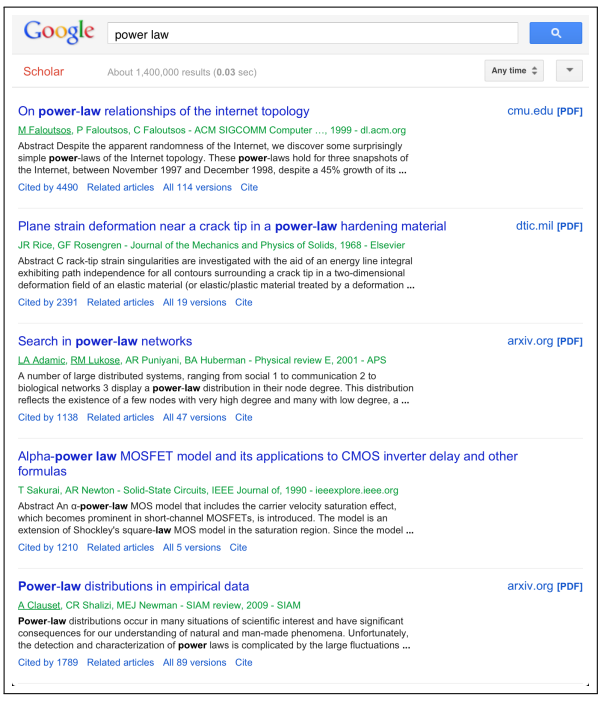


Figure 4.1. Results for “power law” on Google Scholar.

We observe the most obvious benefit of using citation context with the appearance of article “Collective Dynamics of ‘small-world’ Networks” [19] in our list. This article

[2]http://scholar.google.com.tr

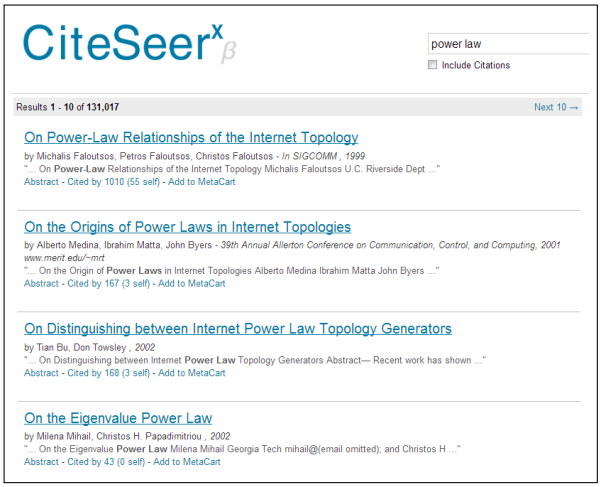


Figure 4.2. Results for "power law" on CiteSeerX.

does not contain the term "power law" and other terms in the term set. However, this is an important article for this topic and other methods miss it while we find it.

We can see the order of most descriptive terms for the top articles in "power law" citation network in Table 4.4.

Table 4.4. Bigram Rankings in the "power law" Citation Network.

Term	**ref [16]**	**ref [17]**	**ref [18]**	**ref [19]**
power law	1	1	2	6
scale free	2	4	1	5
preferential attachment	3	16	8	9
degree distribution	4	2	3	7
random graphs	5	5	6	4
small world	6	7	5	1
social networks	7	6	7	3
complex networks	8	10	4	8
web graph	9	8	26	34
internet topology	10	3	14	26
clustering coefficient	16	25	8	2

4.3. Citation Network for the Term "hadoop"

"Hadoop" itself is a very descriptive term, so we didn't include similar terms while generating network. We used single term citation network which equals to similar term

set citation network with similar term set size is equal to one ($n = 1$).

Hadoop was derived from Google File System (GFS) [3] and Google's MapReduce [4] papers. These papers published respectively in 2003 and 2004 while Hadoop term is coined in 2005. Our method detected these articles as the highest ranking papers as it is seen in Table 4.5 while existing search engines even do not list them.

Table 4.5. Article Ranking of Citation Network for "hadoop" term.

Paper \Ranking Method		**In-Degree Weight**	**HITS-Authority**	**PageRank**
Reference	**Title**			
ref [4]	MapReduce: Simplified Data Processing on Large Clusters	1	1	1
ref [3]	The Google File System	2	2	2
ref [22]	Evaluating MapReduce for Multi-Core and Multiprocessor Systems	3	3	3
ref [23]	Pig Latin: a Not-So-Foreign Language for Data Processing	4	4	4

We searched also for the term "hadoop" on academic literature search engines and the top results are shown in Figure 4.3 for Google Scholar and in Figure 4.4 for CiteSeerX.

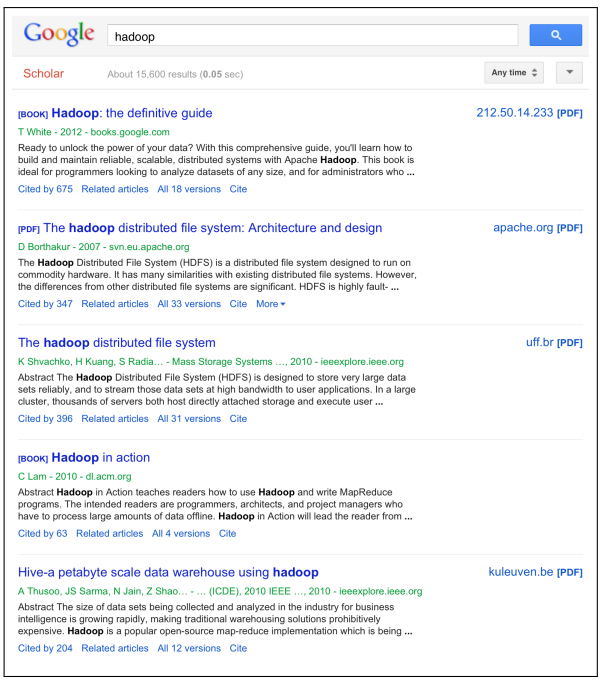


Figure 4.3. Results for "hadoop" on Google Scholar.

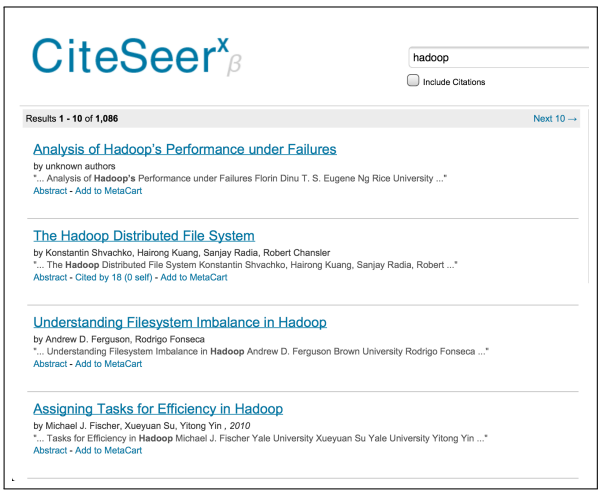


Figure 4.4. Results for "hadoop" on CiteSeerX.

5. CONCLUSION

Citation indexes are generally based on Boolean retrieval, so every article using a set of query terms is equally likely to be listed for the given query. The author of an article uses many words while explaining her research that may contain words not related with main contributions of the article. Hence, unrelated articles may rank highly in search results for a query, simply because they are important articles in another area and contain the query terms. So there is a need for a system which is able to measure both relevance and impact.

We are interested in finding fundamental and important documents in a context sensitive way. We especially target scientific literature because of the existing potential of citation structure. The text around citation marks represents very concise information about cited documents. Probably, the most prominent criticism is that citation analysis based on raw citation counts ignores the underlying reasons for the citation. So, we come up with a solution for this problem.

In this work, we presented a method to utilize citation contexts in order to rank important articles in a topic specific way. For a given term which represents the interested topic, first we formed a set of similar terms. Then we detected citation contexts which contain terms from this set. Only by using detected citations we created topic specific networks. Finally, we applied common link analysis methods in order to find most prominent articles in these topic specific citation networks.

It would draw attention of someone that we did not report the ranking for hub scores while we reported authority scores for the HITS algorithm. This is because we did not find meaningful results for hub scores where most of the nodes had a score of zero. According to our observations in these experiments, the reason for this is that there are no articles which list most of the prominent articles more than others.

This is an unsupervised problem and evaluating results require knowledge in

target context. So we kept our test cases limited in order to be sure that results we found are meaningful. As a future work, we can work with academicians from different research topics in order to evaluate our system broadly.

6. FUTURE WORK

This work utilizes lots of different research areas. So there are lots of different areas where improvements would increase quality of results.

First of all better automatic citation context extraction techniques would help to restrict unnecessary text processed. Given the citation context, it is also very hard to extract meaningful terms from it and this area is open to broad improvements.

Some can try different methods to find similar terms and similarity scores for a given term. Instead of using similarity scores directly in edge weight creation, a function of similarity score can be used to create a different weighting method.

Besides, one can also use full text of articles in order to improve overall system by using citation contexts. It would be an interesting research to compare terms extracted from citation contexts of the citing articles and terms extracted from the cited article.

APPENDIX A: LINK ANALYSIS METHODS

PageRank and HITS are the link analysis methods which are used to rank web pages. We utilized these methods to rank articles in topic specific citation networks we created.

A.1. PageRank

PageRank is a method for calculating a ranking of web pages on the directed unweighed graph of the web where there is an arc from page A to page B if and only if there is a hyperlink to B in A [15]. PageRank is based on incoming links, but not just on the number of them, also the PageRanks of the incoming links are important. Every web page starts with a predefined PageRank value and then PageRank algorithm is run iteratively until the PageRank of all web pages converge.

PageRank can be explained with a random surfer model of the web. Random surfer starts from a random web page, and then selects one of the outgoing links from the web page in a random way or with a small probability jumps to another random web page. The probability that the random surfer visits a web page is its PageRank. And, the d *damping factor* is the probability at each page the random surfer will get bored and visit another random page.

When a web page has no outgoing links, it is assumed that it links out to all other pages in the collection. Its PageRank score is therefore divided evenly among all other pages. These random transitions are added to all nodes in the Web, with a residual probability d in order to be fair with pages that are not sinks.

Here is the equation for the calculation of PageRank at every iteration

$$r_{i+1}(u) = \frac{(1-d)}{N} + d \sum_{v \in N_u^+} \frac{r_i(v)}{|N_v^-|}$$

where $r_i(u)$ is the PageRank score of web page u at the end of iteration i, d is the damping factor, N_v^- is the set of web pages v points to and N_v^+ is the set of web pages that point to v.

Iterations stop when PageRank scores do not change over the iterations more then a tolerance value, so we stop when the following equation is satisfied

$$\forall v, \quad r_{i+1}(v) \cong r_i(v).$$

A.2. HITS Algorithm

Hyperlink-Induced Topic Search (HITS) is a link analysis algorithm that ranks web pages. This algorithm is also known as hubs and authorities [14]. Main idea behind the algorithm is that an important authority represents a web page that is referred by many important hub web pages, and an important hub represents a web page that referred to many other important authoritative web pages. So every web page has two scores, one for hub and one for authority.

The algorithm is a series of iterations of two steps and stops when the scores for hubs and authorities stabilize. In the authority update step, new *authority score* for a web page is calculated by summing hub scores of each web page that points to it. Similarly in the hub update step, new *hub score* for a web page is calculated by summing authority scores of each web page that it points to. Before starting a new iteration there is also a normalization step to prevent score inflation.

Here are the equations for the calculation of authority and hub scores at every iteration

$$a_{i+1}(u) = \sum_{v \in N_u^+} h_i(v),$$

$$h_{i+1}(u) = \sum_{v \in N_u^-} a_i(v)$$

where $a_i(u)$ is the authority score of web page u at the end of iteration i, $h_i(u)$ is the hub score of web page u at the end of iteration i, N_v^- is the set of web pages v points to and N_v^+ is the set of web pages that point to v.

APPENDIX B: IMPLEMENTATION

Source code and documentation about CiteSeerX project are published freely.[3]

We utilized four of the database tables in our dataset. We listed important fields of them and their descriptions in Table B.1 for papers table, Table B.2 for citations table, Table B.3 for citationcontext table and Table B.4 for keywords table.

Table B.1. Papers table: Stores metadata and access information for each paper.

Field	**Description**
id	Unique id identifying each paper. System assigned using the DOI server.
version	Last valid version of paper metadata.
cluster	Identifier of the collection of similar papers/citations. Default value 0, after the inference process is run.
public	Indicates if the paper is available or not.
ncites	Number of citations found within the corpus for the given paper. (It's calculated by the inference process)
versionName	Name of the last valid version of paper metadata.
crawlDate	When the paper was obtained.
repositoryID	Repository identifier where all the files associated to the paper are located. The Repository identifier maps to a physical location in a file system.
ConversionTrace	All the tools (in order) used to get the text version of the paper.
selfCites	Number of self citations. (This data is calculated by the inference process)
versionTime	When the last version of the paper metadata was created/updated.

[3]https://citeseerx.svn.sourceforge.net/svnroot/citeseerx

Table B.2. Citations table: Citations found in a paper.

Field	**Description**
id	Unique id identifying for each acknowledgement. System assigned auto increment.
cluster	Unique identifier that matches the same citation in different papers to a canonical one; including the paper if present within the corpus. The id is assigned by the inference process.
raw	Citation text as found in the paper.
self	Indicates if this citation is a self citation.

Table B.3. CitationContexts table: Contexts which contain citation.

Field	**Description**
id	Unique id identifying for each acknowledgement. System assigned auto increment.
context	Raw text where the citation is mentioned in the paper.

Table B.4. Keywords table: Keywords found in a paper.

Field	**Description**
id	Unique id identifying for each acknowledgement. System assigned auto increment.
keyword	The keyword

From unique keywords we get from "keywords" table we created our term set T. The SQL query for getting unique keywords is shown in Figure B.1.

```
SELECT DISTINCT keyword FROM citeseerx.keywords;
```

Figure B.1. Get Unique Keywords.

We find citations related with a given bigram with the query shown in Figure B.2. In order to solve some simple problems in bigram usage, besides bigram itself we also check cases like plural version of bigram and bigram without any space between two words. We considered most popular 3,469 articles and 10,858 terms in similar term detection algorithm, because larger experiments were not computationally feasible. We used text mining package[4] in R language[5] in order to calculate similar terms and their similarity scores from given bigram.

```
SELECT citationid FROM citeseerx.citationcontexts WHERE
                    context LIKE '%bigram%' OR
                    context LIKE '%bigramPlural%' OR
                    context LIKE '%bigramAdjacent%';
```

Figure B.2. Get Citation Related with a Given Bigram.

Same paper may be crawled from different web sites with different paper ids, so CiteSeerX uses a concept of "cluster" where they group same paper with different paper ids in a single "cluster".

We get target "cluster" and source "paperid" with the query shown in Figure B.3.

Then we find the source "cluster" id from source "paperid" with the query shown in Figure B.4. "cluster" id helps us to group same article which is provided by different

[4]http://cran.r-project.org/web/packages/tm/index.html
[5]http://www.r-project.org

```
SELECT cluster, paperid FROM citeseerx.citations
                        WHERE id = citationid;
```

Figure B.3. Get Target Cluster and Source Paper for a Given Citation.

web pages. We choose a "paperid" to represent every "cluster".

```
SELECT cluster FROM citeseerx.papers WHERE id = paperid;
```

Figure B.4. Get Source Cluster.

We implemented a Java program which goes over whole dataset for a given bigram set and their similarity scores and outputs a file of similar term set citation network in .gdf format[6] . Then we used this file in Gephi[7] and calculated network statistics and article rankings.

[6]http://guess.wikispot.org
[7]https://gephi.org

REFERENCES

1. Aljaber, B., N. Stokes, J. Bailey and J. Pei, "Document Clustering of Scientific Texts Using Citation Contexts", *Information Retrieval*, Vol. 13, No. 2, pp. 101–131, 2009.

2. Schwartz, A. S. and M. Hearst, "Summarizing Key Concepts Using Citation Sentences", *Proceedings of the Workshop on Linking Natural Language Processing and Biology: Towards Deeper Biological Literature Analysis*, BioNLP '06, pp. 134–135, Association for Computational Linguistics, Stroudsburg, PA, USA, 2006.

3. Ghemawat, S., H. Gobioff and S.-T. Leung, "The Google File System", *ACM SIGOPS Operating Systems Review*, Vol. 37, No. 5, p. 29, 2003.

4. Dean, J. and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters", *Proceedings of the 6th Conference on Symposium on Opearting Systems Design & Implementation - Volume 6*, OSDI'04, p. 10, USENIX Association, Berkeley, CA, USA, 2004.

5. Bradshaw, S., "Reference Directed Indexing: Redeeming Relevance for Subject Search in Citation Indexes", *Research and Advanced Technology for Digital Libraries*, Vol. 2769, pp. 499–510, 2003.

6. Ritchie, A., S. Teufel and S. Robertson, "How to Find Better Index Terms Through Citations", *Proceedings of the Workshop on How Can Computational Linguistics Improve Information Retrieval?*, CLIIR '06, pp. 25–32, Association for Computational Linguistics, Stroudsburg, PA, USA, 2006.

7. Qazvinian, V., D. R. Radev and A. Özgür, "Citation Summarization Through Keyphrase Extraction", *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pp. 895–903, Association for Computational Linguistics, Stroudsburg, PA, USA, 2010.

8. Moravcsik, M. J. and P. Murugesan, "Some Results on the Function and Quality of Citations", *Social Studies of Science*, Vol. 5, No. 1, p. 86, 1975.

9. Kessler, M. M., "Bibliographic Coupling Between Scientific Papers", *American Documentation*, Vol. 14, No. 1, pp. 10–25, 1963.

10. Small, H., "Co-citation in the Scientific Literature: A New Measure of the Relationship Between Two Documents", *Journal of the American Society for Information Science*, Vol. 24, No. 4, pp. 265–269, 1973.

11. Gipp, B. and J. Beel, "Citation Proximity Analysis (CPA) – A New Approach for Identifying Related Work Based on Co-citation Analysis", *Scientometrics*, Vol. 2, No. July, pp. 571–575, 2009.

12. Manning, C. D., P. Raghavan and H. Schütze, *Introduction to Information Retrieval*, Vol. 1, Cambridge University Press, 2008.

13. Teregowda, P. B., I. G. Councill, J. P. Fern and C. L. Giles, "SeerSuite: Developing a Scalable and Reliable Application Framework for Building Digital Libraries by Crawling the Web", *Architecture*, p. 14, 2010.

14. Kleinberg, J. M., "Authoritative Sources in a Hyperlinked Environment", *Journal of the ACM*, Vol. 46, No. 5, pp. 604–632, 1999.

15. Brin, S. and L. Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine", *Computer Networks and ISDN Systems*, Vol. 30, No. 1-7, pp. 107–117, 1998.

16. Barabasi, A.-L. and R. Albert, "Emergence of Scaling in Random Networks", *Science*, Vol. 286, No. 5439, p. 11, 1999.

17. Faloutsos, M., P. Faloutsos and C. Faloutsos, "On Power-Law Relationships of the Internet Topology", *ACM SIGCOMM Computer Communication Review*, Vol. 29,

No. 4, pp. 251–262, 1999.

18. Albert, R. and A. L. Barabasi, "Statistical Mechanics of Complex Networks", *Reviews of Modern Physics*, Vol. 74, No. 1, pp. 47–97, 2002.

19. Watts, D. J. and S. H. Strogatz, "Collective Dynamics of 'small-world' Networks", *Nature*, Vol. 393, No. 6684, pp. 440–442, 1998.

20. Newman, M. E. J., "The Structure and Function of Complex Networks", *SIAM Review*, Vol. 45, No. 2, p. 58, 2003.

21. Sakurai, T. and A. R. Newton, "Alpha-Power Law MOSFET Model and Its Applications to CMOS Inverter Delay and Other Formulas", *IEEE Journal of Solid-State Circuits*, Vol. 25, No. 2, pp. 584–594, 1990.

22. Ranger, C., R. Raghuraman, A. Penmetsa, G. Bradski and C. Kozyrakis, "Evaluating MapReduce for Multi-Core and Multiprocessor Systems", *2007 IEEE 13th International Symposium on High Performance Computer Architecture*, Vol. 0, No. October, pp. 13–24, 2007.

23. Olston, C., B. Reed, U. Srivastava, R. Kumar and A. Tomkins, "Pig Latin: A Not-So-Foreign Language for Data Processing", *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, pp. 1099–1110, 2008.