RECOMMENDING RELEVANT PAPERS USING IN-TEXT CITATION FREQUENCIES AND PATTERNS

(PhD Thesis)

by

ABDUL SHAHID
PC103009
PhD Candidate
ashahid@kust.edu.pk

Thesis Supervisor
Dr. Muhammad Tanvir Afzal
Associate Professor (Computer Science)
mafzal@jinnah.edu.pk

Faculty of Computing
Mohammad Ali Jinnah University,
Islamabad Campus

ABSTRACT

Scientific publications are growing exponentially. For example, more than 50 million journal papers have been published till now, and more than 2 million journal papers are added to the scientific knowledge every year. The published conference papers are in billions, and millions others are added every year. The world famous scientific databases such as Web of Science, Scopus, and PubMed etc index millions of such scientific papers, and that also despite the fact that their index either belongs to specialized domain or it is selective. There is another comprehensive index known as Google Scholar, indexes huge scientific knowledge from different domains. These systems make available the scientific knowledge to researchers. The advancement in research is always possible by standing on the shoulders of others. However, when users attempt to identify relevant papers from the mentioned systems or other similar systems, they are given millions of papers and are asked to select the most relevant papers manually by skimming those millions of papers. This creates frustration, and generally all of the selected papers do not belong to the list of papers which the users must read. In this task, many important papers are overlooked by the users as well.

The identification of relevant papers from such a big data has attracted a number of researchers across the globe to find solutions to this problem. The contemporary approaches use a variety of techniques for the identification of the relevant documents such as content based approaches, metadata based approaches, collaborative filtering based approaches, co-citation analysis, and bibliographic analysis etc. However, the state-of-the-art research lacks in many directions such as its inability to find the nature of relationship between scientific documents and its failure to find how strongly two scientific documents are linked up, based on their relationship strength.

To address these issues, this thesis designs, implements, and evaluates a novel approach that facilitates researchers to identify the most relevant papers in their domains. The proposed approach identifies the most relevant papers from the list of cited-by papers for the cited paper. This thesis works on the in-text citation frequencies and in-text citation patterns to identify the most relevant papers. In-text citation frequency is the number of occurrences of citations of one paper in the text of the other paper. In-text citation frequency patterns are the in-text citation evidences in different sections of the paper. The system has been implemented as a prototype for

CiteSeer. The proposed system has been evaluated using a number of user studies. The proposed approach shows encouraging results and assists the scientific community to identify the most relevant papers from a huge list of papers.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

This research focuses on enhancing the identification of relevant documents in the scientific domain. This document presents and reviews state-of-the-art scholarly work and systems for ranking relevant papers for a cited paper from a citation list of the cited-by papers. Furthermore, it presents the developed technique and its evaluations through experiments.

Section 1.1 of the endeavor gives the background and importance of the tasks related to improving the identification of relevant documents. In section 1.2, motivations for this work are explained. Section 1.3 caters for the problem statement of this study.

1.1 Background

The growth of digital publications is exponential [Afzal et al., 2007], and finding the relevant information is a crucial task. In 1950, there were 60,000 Journals, and the estimate for the year 2000 was 1 million [Larsen and Ins, 2010]. Furthermore, British Library Lending Division (BLLD) indexed 43,000 journals in 1982 [Larsen and Ins, 2010]. According to Ulrich's International Database, about 250,000 journals were published in 2004 [Dalen and Arjo, 2005]. PLOS (Public Library of Science) was started in 2006 and has published 10,000 articles in just 4 years [PLOS, 2010]. Furthermore, it has been reported that the world information is doubling every eighteen months, and in the scientific domain, the information is doubling in every 5 years [Larsen and Ins, 2010]. The scientifically acknowledged systems such as ISI Web of Knowledge[1], Google Scholar[2] and CiteSeer[3] have also indexed large sets of information. For example, at the time of writing this document, only ISI has indexed about 17,581 international and regional selected journals [ISI, 2013]. Searching a term such as "ontology papers" on Google Scholar results into millions of papers. For instance, the first entry titled "Gene ontology: tool for the unification of biology" has more than 12,400 citations. Considering further, if a person reads 10 papers daily, in that case, he would have to spend three and a half years to read all relevant papers.

[1] http://thomsonreuters.com/web-of-knowledge/
[2] http://scholar.google.com.pk/
[3] http://citeseer.ist.psu.edu/

We also should not forget that 2 million papers are published every year and are made part of the big data [Jinha, 2010].

The current state-of-the-art citation indexes (CiteSeer and GoogleScholar) offer the ranking of citation services. They provide various options for exploiting the relevant citations list, such as ranking citation on the basis of citations count, date-wise (ascending, descending), or listing of citations from a particular date onward etc. The algorithm of Google Scholar for ranking citations is unknown. However, it has been reported that citations count play a vital role in the ranking of the relevant citations [Beel and Gipp, 2009c]. Ranking relevant citations in ascending or descending order or on the basis of citations count cannot ensure that the relevant citations will be ranked at the top of the list. A detailed case study has been explained in the motivation section of this chapter.

Furthermore, researchers [Shum, 1998][Kaplan et al., 2009] are continuously exploring the semantic relationship between publications. By semantic relationship they mean to find how two papers are related to each other. For example, discovering documents that are using or extending a particular document, documents analyzing the cited paper's problem, or documents problematizing a methodology are closely associated the particular document etc. However, with the current approaches, the semantic relationship is hard to find between two articles.

Identification of the relevant document is an active area of research [Beel et al., 2013]. There are various approaches proposed in literature to identify the relevant papers. For example, Justin et al., proposed the visualization of citations network to find important papers [Justin et al., 2012]. Similarly, Benjamin and Schafer have proposed the visualization of citations along with the relationship of citations for the focused paper to identify relevant papers [Benjamin and Schafer, 2010]. Citations between the papers mimic links between the web pages; therefore, variant of PageRank has been proposed to identify relevant documents [Pruitikanee et al., 2013][Haddadene et al., 2012]. In the previous decade, techniques such as text based relatedness [Nattakarn and Ozsoyoglu, 2007] and the identification of future relevant papers [Afzal et al., 2007] have been proposed. Similarly, in the previous century, the state-of-the-art techniques such as co-citation analysis [Small, 1973] and bibliography

analysis [Kessler, 1963] had been proposed. The details of these techniques can be found in related work section. However, these techniques do not have the capability to show any kind of semantic relationship between papers.

In scientific domain, Garfield [Garfield, 1964] described that the relationship can be seen through the citations. The cited-by paper may cites a paper for many reasons, for example, to cover a background study, or extending the work mentioned in the cited paper. These reasons are hard to identify from the text and need NLP, Machine Learning, and Artificial Intelligence techniques. However, the problem can be indirectly solved by analyzing the text of the cited-by paper, based on the citation patterns. For example, 1) the 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, and 3) the identification of citation patterns in the paper's sections (introduction, related work, methodology etc). It may be possible to identify the relationship strength and relationship nature by observing the in-text citation frequency and in-text citation patterns, respectively. The thesis focuses on such identifications for the ranking of relevant citations.

We investigated and proposed that in-text citation frequency and in-text citation patterns could be other important filters in the Relevant Citation Rankings. The in-text citation frequency and in-text citation patterns have been explained in Figure 1.1, which explains various concepts such as cited-by article, cited article, in-text citation frequency and in-text citation patterns.

When a paper "A" cites paper "B", the "A" is called cited-by paper and the "B" is called cited paper. Every article contains a specific section called "References" section which lists all references. Furthermore, these referenced articles (cited articles) are referred in the body text of the cited-by article. The in-text citation frequency refers to the number of occurrences in body text of the cited-by article for a cited article. For example, in Figure 1.1, the first cited article in the reference list has been referred twice in the body text of the cited-by article. Similarly, the third article has been referred four times in the body text of the cited-by article. In-text citation patterns refer to the evidence of in-text citations of cited article in different section of the paper. For example, in the shown Figure 1.1, the third cited article "[3]" has been referred in "Introduction" and "Discussion" sections.

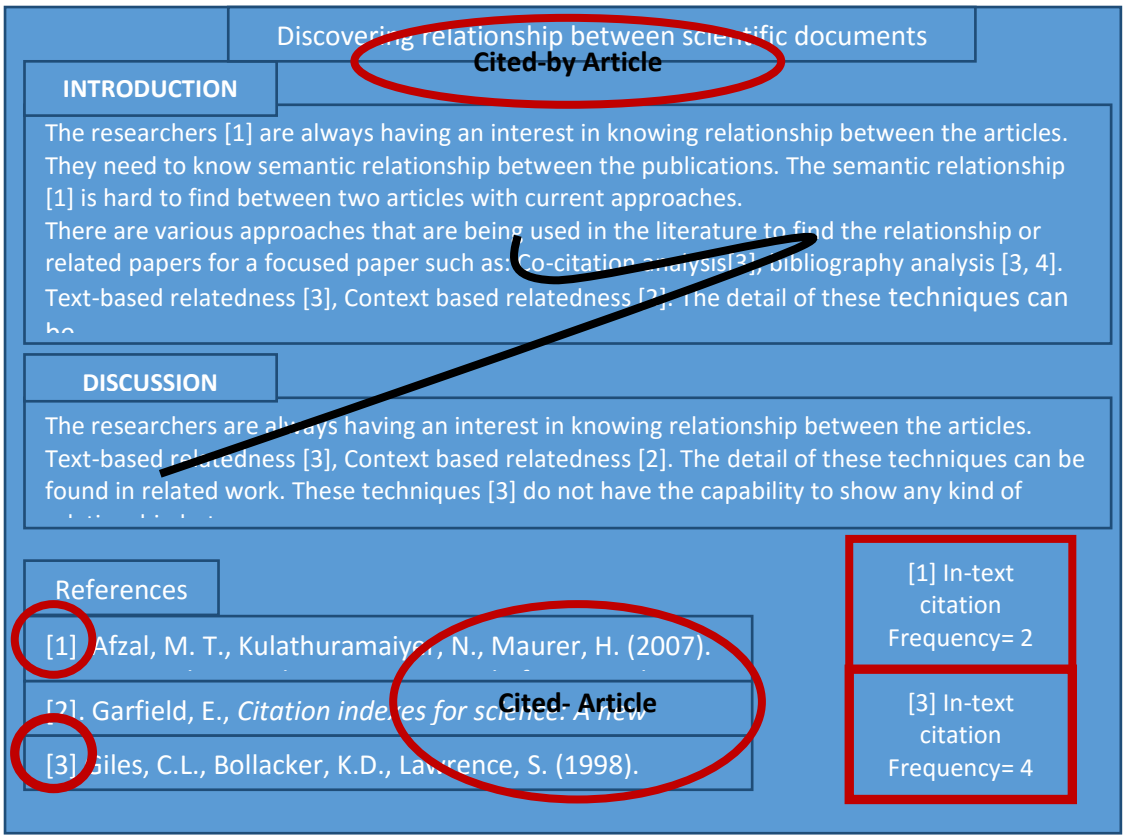


Figure 1.1: In-text Citation Frequency and In-text Citation Patterns

The cited articles are referred with the help of citation tags. The citation tag is referred to the information in the reference that is used to cite that article in the body of the cited-by document e.g. in Figure 1.1, "[1]", "[2]", and "[3]" are referred as citation tags. Sometimes, these tags are alphanumeric such as: "[Afzal et al 2008]". Similarly, sometimes, different variations are used for numeric one like: "1" and "[1]".

1.2 Motivation

Information is considered as the most important valuable asset. The World Economic Forum has recently declared the information as a new class of economic asset like currency or gold [WEF, 2011]. The importance of the current Relevant Citation Ranking can be understood using a case study scenario. For example, one wants to identify the relevant paper on the topic of "PageRank". Searching on Google Scholar provides more than 37,200 research documents where the first entry, titled "The PageRank citation ranking: Bringing order to the Web," has more than 6,500 citations. On clicking the cited-by articles, a ranked list of relevant papers is shown. The top 10 results of cited-by articles are shown in Table 1.1. Google Scholar algorithm for ranking the cited-by documents is unknown. It has been reported in

literature that *Citation count* plays a vital role in ranking the citations [Beel and Gipp, 2009c], and the same was observed in our study too. However, the quality of results is not very good and needs improvement in terms of ranking relevant citations. For example, the papers at 4th and 6th positions are from network domain which is not the focus of the cited paper. Similarly, the results at 2nd, 3rd and 5th positions are books that are of generic nature which might not be too relevant.

Table 1.1: Top 10 Results of Total 6,521

1	The Anatomy of a Large-Scale Hypertextual Web Search Engine
2	Modern information retrieval
3	Linear Methods for Regression
4	The Structure and Function of Complex Networks
5	Introduction to Information Retrieval
6	The EigenTrust Algorithm for Reputation Management in P2P Networks
7	A Survey of Trust and Reputation Systems for Online Service Provision
8	What is Twitter, a Social Network or a News Media?
9	Measurement and Analysis of Online Social Networks
10	Folksonomies - Cooperative Classification and Communication Through Shared Metadata

Furthermore, in this study, top 200 documents, returned by the Google Scholar, were investigated to identify any relevant papers which are not part of the top 10 results; and to our surprise, it was discovered that at 64th position, there exists a citation with the title "Inside PageRank". Similarly, at 145th position, another very relevant citation titled "Weighted PageRank Algorithm" was discovered. Similarly, at 199th position, another important citation was found with the title "Using PageRank to Characterize Web Structure". These mentioned papers should have been ranked higher than their mentioned positions if not part of the top 10 results. This case study clearly explains

the need of devising a technique through which the relevant citations ranking can be improved.

There are certain techniques available to extract relevant information which have their own shortcomings and often present a long list of irrelevant information. Therefore, it is needed to make a critical analysis of state-of-the-art systems and to propose an innovative technique, which can extract and present the most relevant ranked list of papers.

1.3 Problem Statement

The identification of relevant papers is a dire need of the scientific community. The task of extracting relevant papers is challenging due to the large amount of availability of scientific publications. Therefore, research community is engaged in developing state-of-the-art techniques and approaches for extracting relevant information. The existing systems for extracting relevant papers use different techniques such as 1) Content based, 2) Metadata based, 3) Collaborative filtering based, and 4) Citations based analysis. The existing approaches, however, do not consider the nature of the relationship between articles, and are unable to provide ranking based on the relationship strength. With support from comprehensive critical analysis of the domain, the problems are listed below:

1. The contemporary techniques and systems do not consider nature of relationship between cited and cited-by documents.
2. The state-of-the-art techniques and systems do not consider relationship-based strength for ranking cited and cited-by documents.
3. The content based approaches have vocabulary issues, as they compute relatedness just by considering the content of two documents irrespective of the concepts used in the papers.
4. The metadata based systems are based on just few terms to identify relationship and thus less number of recommendations are made.
5. The citation based approaches mostly consider citation network information instead of considering the concepts discussed in the papers.
6. Finally the collaborative filtering based techniques suffer from many issues like cold start problem, gray sheep, black sheep, and data sparsity issue.

1.4 Purpose of the Research

Relevant document extraction is an important task. The focus of the current thesis is to develop a state-of-the-art technique for discovering relevant information. Before explaining the purpose of the current thesis, we briefly explain the current state-of-the-art systems followed by the purpose of this research.

1.4.1 Current State-of-the-art Systems

Many techniques are being used in literature to extract relevant research papers for a focused paper. The existing systems have a number of limitations such as 1) vocabulary issues, where similar terms do not ensure that the papers are similar, 2) inability to identify the relationship nature between the research documents, and 3) failure to identify the relationships strength between research documents. A lot of discussion has been provided on this issue in the second chapter of the thesis.

Apart from the research efforts, there are certain large scale applications available which are being used by the scientific users to gather relevant documents. These applications can be categorized as Search Engines (Google, Yahoo, and Bing), Citation Indexes (Citeseer, GoogleScholar), digital libraries (IEEE, ACM) and socially maintained databases (CiteULike, Bibsonomy). Search Engines are the most generic category and its results are web pages from blogs, discussion forums, and resumes etc. Mostly, the retrieved items contain generic results and thus a large number of non-relevant documents are retrieved. The second category, Citation Indexes, like CiteSeer and Google Scholar, are specialized services for the scientific community. The CiteSeer scope is confined to scholarly documents related to the domain of Computer Science, while GoogleScholar covers other disciplines as well. However, it also results thousands of documents, some of which are relevant, while others are irrelevant. Broadly speaking, the current research and applications lacks in the identification of the relationship nature, identification of relationship strength, and availability of innovative visualization.

1.4.2 Objective of the Research

The objective of this research is to develop a technique that assists the scientific community to find specific research papers of their interest, relevant to particular

papers. More precisely, this thesis aims to help researchers to identify the most important papers from thousands of papers that are methodologically relevant. For example, the papers which have extended the work of the focused paper, or they have compared their results with the focused paper etc.

1.5 Applications of the Research

The developed techniques in this research can be adapted for use in a number of application domains by adapting it for other domains. Few of them are listed below:

a. Extending the current state-of-the-art techniques for extracting relevant documents such as

   1. Co-citation and
   2. Bibliographic coupling

b. Revising the quality measure, defined for research paper such as

   1. Impact Factor
   2. H-index

c. Extending the current citation indexes

   1. ISI Web of Knowledge
   2. Google Scholar
   3. CiteSeer

d. Identification of relevant documents for:

   1. A researcher when one wants to gather the most seminal papers relevant to a topic.
   2. A researcher when one wants to gather introductory papers on a topic.

Chapter 2

Literature Review

Since the advent of World Wide Web (WWW), the scientific literature dissemination has become easy, and thus, users can have access to huge scientific knowledge repositories [Bollacker, 2000]. The publishing venues such as journals, conferences, and open archives are increasing. This results in a tremendous growth in the quantity and diversity of research publications. Researchers are always interested to find relevant research, specific to their areas of interest. However, information overload makes this task complicated, and thus, extraction of relevant document from huge document repositories is challenging. Just to get an idea that how much research documents are available on particular topic, some queries were executed on a well-known citation index: Google Scholar. The total number of results (research papers) retrieved against each query is shown in Table 2.1. For the queries presented in the table, the average numbers of results returned are more than 3 million. This becomes difficult for users to focus the most relevant and interesting papers which they should read from these millions of results. These results give us an idea about the size of digital publications. Therefore, different researchers have proposed and developed number of research techniques and systems for identification of relevant documents.

Table 2.1: Query Results using Google Scholar

S.No	Query Term	Total Results
1	Information Retrieval	2,910,000
2	Knowledge Management	3,370,000
3	Citation Analysis	3,960,000
4	Text Similarity	2,660,000

However, evaluating the state-of-the-art techniques without a benchmark is a challenging task. Therefore, this chapter presents the evaluation criteria for evaluating state-of-the-art system. Furthermore, different types of data sources and contemporary techniques have been described and evaluated based on the defined evaluation criteria. Subsequently, existing applications have also been discussed and evaluated.

The previous research based solutions exploit different data sources such as: content, metadata, collaborations and citations as shown in the Figure 2.2. The studied systems have been placed into different categories based on the techniques being used by them. The categories based on techniques have been shown in the Figure 2.2 are: syntactic, semantic, collaborative filtering, bibliographic analysis, and hybrid. These approaches are not rivals of each other, rather complementary in nature. Therefore, there are different studies reported in literature which have used these techniques in merger to identify most relevant research papers. In this study, a section has been included which discusses all those research techniques in this chapter.

There are some applications as well that have been deployed to support users to identify relevant research papers. Such applications include: search engines, citation indexes, digital libraries and socially maintained databases. All of these systems have been discussed in section 2.5 of this chapter.

In this research, the vision is to identify all relevant papers for a paper "A" from a huge corpus of research papers $\sum_{k=1}^{n} RP_k$.The relevance is measured based on the following criteria: 1) whether the relationship nature between paper "A" and the identified relevant papers have been established [Garfield, 1964] [Teufel, 2006], 2) whether the identified relevant research papers have relationship based strength associated with the paper "A", and 3) whether the ranking of the evaluated technique can be altered by some external means.

Finally, this chapter briefly describes the data-sources, and highlights the strengths and limitations of the systems that use a particular dataset for the identification of relevant documents. Afterwards, the techniques (that operate on various data sources) have been critically discussed.

2.1 Data Sources

Different data sources are used for the identification of relevant papers for a source paper such as: content of papers, metadata of papers, collaborations (e.g. same user is co-downloading different papers etc.), and citations data set. It has been shown in the Figure 2.1 of this chapter. In the following sections, the mentioned data sources have been discussed in details.

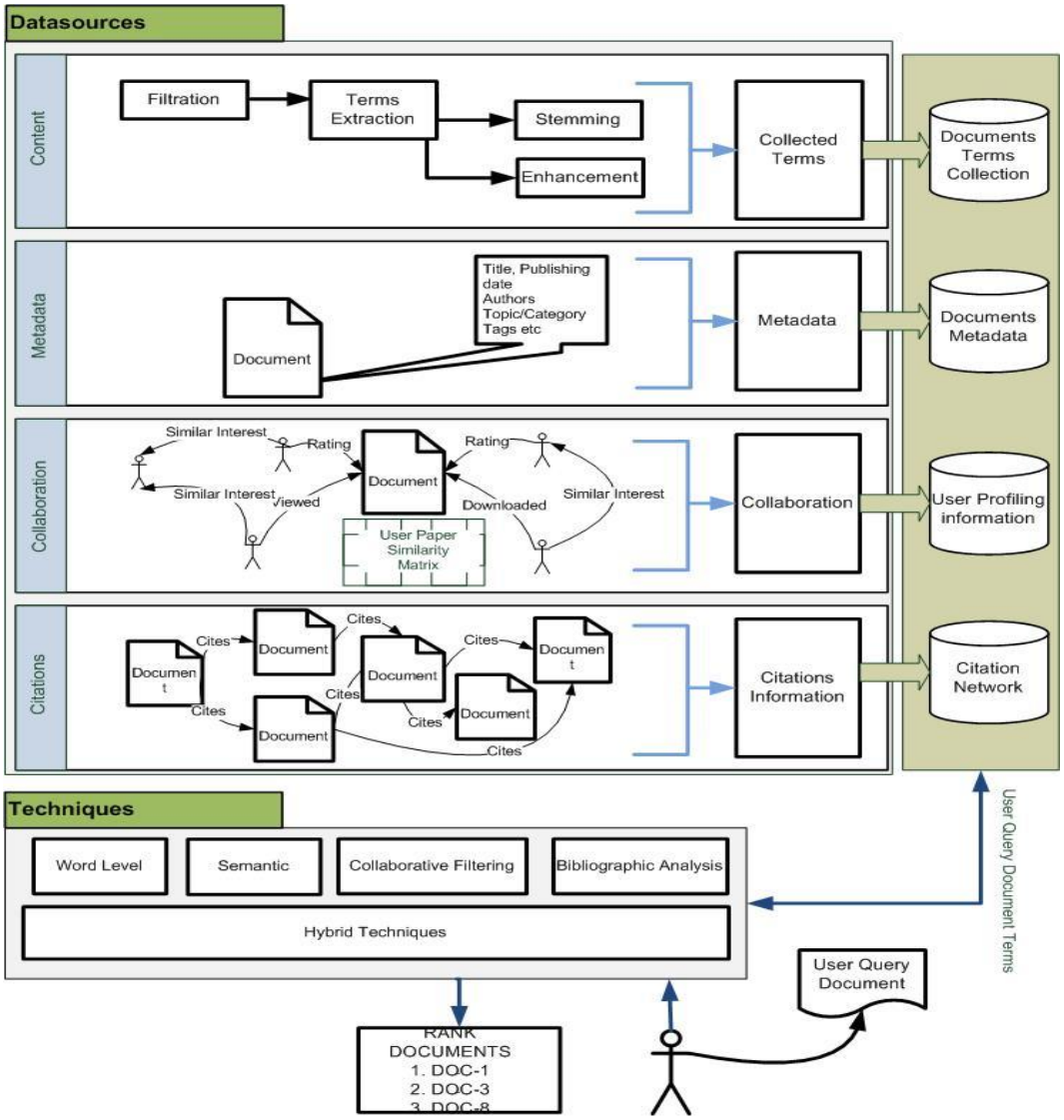


Figure 2.1: Data Sources and Techniques Used to Identify Relevant Documents

In the Figure 2.1: state-of-the-art techniques and data sources which are used to identify relevant research papers have been presented in an organized way.

2.1.1 Content

The content of research paper normally means the whole text of the paper. The techniques that operate on content of a research paper, assume that relevant papers would be those papers which are similar in content to the source paper. These techniques compute the similarity between the content of two research articles. Thus, more the text is similar, the more the documents are relevant.

There are certain strengths and weaknesses associated with content based techniques. The strengths are: content based approaches are generic in nature and thus can be used for any kind of documents/reviews/news etc. Furthermore, well-defined distance measures are available for calculating the similarity between two texts. However, there are certain weaknesses which need proper attention while considering it for ranking of relevant papers. The weaknesses are: for instance vocabulary issues e.g. United States of America and USA is different by its vocabulary but they are topically same, therefore, it can be overlooked if not properly considered. Similarly, context information can also mislead the results e.g. “Apple computers” and “apple pie” will be matched just due to the fact that the term apple exist in both of the text segments, which are contextually different. Finally, for large documents, the content based approaches will need more time to find relevancy score between documents.

Furthermore, there are certain other important concerns which are specifically associated with effectiveness of the retrieved results for scientific domain. The distance measures such as Levenstein, jaccardetc, rank the retrieved results based on number of terms matched within two documents. However, more terms may be matched into two different documents just because of similar vocabulary used which don’t ensue that they are belonging to the same topic. The strengths and limitation of content based approaches are summarized below:

Strengths:

- Generic- can be used for any kind of documents
- Well-defined measures are available for computing similarity between documents/text
- Able to identify the relationship strength between documents.

Weaknesses:

- Resource hungry;
- Synonyms/Vocabulary issue;
- Context ambiguity: example “Apple Computers” and “apple pie” are matched;
- Two documents “A” and “B” will be considered as very relevant if both have similar terms and talking about different things;

- Two documents “A” and “B” will be considered as irrelevant even if both are talking about similar thing and using different set of vocabularies; and
- High chance of more similarity between documents by the same authors even if the documents belong to different areas but the vocabulary belongs to the same author.

2.1.2 Metadata

Another data source being used by the researchers is the metadata of research papers. Metadata can be defined as data about the data. In the context of research articles, the metadata could be “Title of the paper”, “author(s) of the paper”, “keywords”, “ACM topics (if any)” etc. Technique that discovers the relevant papers based on metadata of research articles are categorized as metadata based techniques. With the recent advancement in the web (Web 2.0), the resources (in our case, research articles) no longer remain passive items. The users can add extra information or annotate them for future use. Different online services specifically have been built for serving scientific communities such as CiteULike[4] etc. The user of these services can annotate the research articles, bookmark the references etc. We believe this is another kind of metadata. Thus, we have divided metadata into two main categories such as traditional metadata and metadata acquired from social tagging and bookmarking.

Traditional metadata refers to the data about research articles such as “Title of the paper”, “author(s) of the paper”, “keywords”, “ACM topic (if any)” etc. Different authors have tried to discover relevant papers with the help of metadata exploitation of research articles. Generally speaking, metadata has been remained less focused area for the identification of relevant papers. One such system was developed byAfzal et al [Afzal et al., 2007]. They proposed to use authors, ACM topic information and published date information for the identification of relevant papers. The idea is that when a reader opens a paper “A”, then all of its (A’s) relevant papers are displayed to the users that are published later in time, by the same team of authors in the same topic. The idea has been implemented and verified in an online Journal, Journal of Universal of Computer Sciences (J.UCS). However, there are some limitations, such

[4] http://www.citeulike.org

as, when a paper does not have the topic information then it will not be possible to find its relevant paper. Similarly, when author(s) of paper has published their papers in different topics/venues, then the proposed approach may retrieve irrelevant papers or it becomes hard to identify such metadata from research papers.

Normally, citations of the papers are openly available and therefore references are sometime considered as metadata. Authors such as Sajid et al. have used paper's references to discover the paper topic [Sajid et al., 2011]. They extract the references from the current paper and its ACM topics. Each paper's reference and its ACM topic are paired. When new document arrives, the references of new document are searched within the established pairs and wherever the reference is matched, the corresponding associated topic is retrieved. This technique has also been tested on J.UCS. However, this kind of systems needs large number of papers whose references and ACM topics are known to make topic-reference pair. Furthermore, if a reference does not find its match in the established topic-reference pair, it becomes difficult to classify papers. There are large number of articles that doesn't follow ACM classification. Therefore, generalizations of such techniques are required to be implemented and evaluated.

On the other hand, in the context of social web, the users and item are no more passive entities. Thus, it has created new possibilities to find relevant papers by offering different services like CiteULike, Bibsonomy etc. These services are based on usage of social tagging and bookmarking of references and web pages (which we call as metadata - data about research articles). These services allow the users to save their references or a web page with their own defined tags or keywords. There are some popular social tagging and bookmarking services designed specifically for scientific community such as CiteULike, Bibsonomy[5]& Delicious[6] etc. Tags from the CiteULike have been exploited in different research studies such as [Khan et al., 2012] [Bogers and Bosch, 2008].

Techniques that operate on metadata have definitely an edge over techniques that require content of the papers. For example, metadata is freely available, while on the other hand, there are large numbers of digital repositories which have restricted access to the content (i.e. content are not freely available). Furthermore, the metadata based techniques can quickly determine whether two papers are relevant or not as compared

[5] http://www.bibsonomy.org/
[6] https://delicious.com/

to content based techniques. However, major limitation of the metadata based techniques is that they are normally dataset dependent. They cannot be generalized, for example, one type of metadata available in one dataset may not be available in other dataset. Normally, author, title and published information are available in every dataset. However, these information are too less to compute any type of relatedness between the research articles. Sometimes, the metadata is not freely available. In this case the metadata is extracted automatically. For automatic extraction of metadata from research papers, there are certain efforts such as rule based approaches [Klink and Kieninger, 2001], and statistical based approaches [Andrew et al., 2000] etc.

We have summarized the strengths and weaknesses of metadata based techniques as follow:

Strengths:

- Rely on just few key terms (title, author, topic etc.); and
- Need less effort to compute the relatedness.

Weaknesses:

- No comprehensive study available to evaluate its performance,
- Specialized technique related to some specific dataset,
- Two papers “A” and “B” will be considered as very relevant when both have common terms in title, irrespective of the area etc.; and
- Two papers “A” and “B” will be considered as irrelevant when both do not have common terms in e.g. title, whether both are talking about the same topic and problem.

2.1.3 Collaboration

Apart from the metadata and content, there is another important data source being used for the identification of relevant papers. This data source is acquired from the collaborative profiles of users by applying collaborative filtering. Collaborative filtering is the process of filtering information based on collaboration of community members. Collaborative filtering have been used by number of applications such as: retrieving best seller list [Linden et al., 2003], retrieving relevant music [Cohan and Fan, 2000], and movies [Koren et al., 2009] etc. Collaborative filtering, can be seen as “(CF) is the prediction of a small subset of items (filtering) for a specific user, that is

derived out of the taste information of many other users (collaboration)" [Pohl et al., 2007].

Collaborative filtering is performed based on user profiles. The user profile is the digital representation of the persons or user profile is the computer representation of a user model. Furthermore, user profiles are enriched by monitoring user's activities. Different systems perform profiling either one way or another in order to suggest relevant items. In presenting the related information, it is assumed that the users of similar interest (group) would need similar objects. Furthermore, the user profiles exhibit the behavior of the users in a system. Therefore, first it is necessary to build user profile. There are two main types of processes followed to build user profile such as 1) explicit rating, and 2) implicit ratings. In explicit rating, the user is asked to rate the presented information while in implicit rating the click stream, co-downloads and other type of actions of the users are analyzed.

There are certain limitations associated with collaborative filtering based techniques as well. The worth mentioning are some general limitations that are summarized by Su and Khoshgoftaar [Su and Khoshgoftaar, 2009]. These limitations are described in below section.

a) Data Sparsity Issue:

Cold Start Problem: When user or new Item is added, then finding relevant information is very difficult as there is not enough information: Sometimes, this issue is called new user or new item problem. New items cannot be recommended until some user rates it.

b) Reduce Coverage Problem:

Coverage can be defined as the percentage of items that the algorithm could provide recommendation for. The reduce coverage problem occurs when the number of users' rating is very small as compared to large number of items in the system, and the recommender system may be unable to generate recommendation for them.

c) **Synonymy**:

This parameter refers to the tendency of a number of same or very similar items to have different names. Most of the systems are unable to find this latent association

and thus deals the same items as different. Like Children Movies and Children Films are considered differently.

The attempts are made in direction of automatic expansion of word or construction of thesaurus [Chen and Chen, 2007]. The problem in this approach is that some added terms can have different meaning.

d) Gray Sheep:

This refers to the people whose opinion do not consistently agree or disagree with any other group of people and thus do not benefit from Collaborative filtering.

e) Black sheep

Black sheep are the opposite group whose unique taste makes the process of recommendation impossible. The manual recommendation faces similar problems in such cases. So, Black sheep are acceptable failure.

f) Shilling attacks:

In cases where anyone can provide recommendations, people can provide tons of positive recommendations to own material and negative recommendation to others.

g) Privacy Issues:

People may not want that their habits are widely known.

The overall strengths and weaknesses of collaborative filtering for recommending relevant documents have been summarized below:

Strengths:

- Generic- Can be used for any kind of items (documents)

Weaknesses:

- Hard to have real setup access for researchers, might be meaningful for a dedicated digital library;
- It can only recommend the article to the user as per user behavior which matches with other users,
- Cold Start problem is very common to scientific domains as users are less and items are too many,
- Recommendation heavily depends on user input and thus Graysheep, Blacksheep and Shilling attacks need careful attention; and

- Unable to identify relationship nature between related documents.

2.1.4 Citations

For identification and recommendation of relevant papers, another important dataset being used is the citation. Citation is naturally systematic process in which researchers cite others work and thus create a citations network. This citation network information play very important role in scientific community. It has been used in a number of applications such as calculating impact factor of Journal [Garfield, 2006], ranking of universities, calculating importance of a researcher (H-index) [Hirsch, 2005] etc. Citation is also important for many other reasons such as ensuring novelty of authors work, validating author argument in the light of mentioned research work etc. Researchers refer to others' research works, that way a relationship between the documents is explicitly mentioned by author. However, various state-of-the-art techniques have been proposed in the literature for identifying relevant papers from the citation network.

The most widely known primitive techniques that require citation network information are Bibliographic coupling [Kessler, 1963] and co-citations [Small, 1973]. Details of these techniques will be discussed in citations network techniques.

Similar to the predecessors, citations network based techniques have also some strengths and limitations which have been summarized below:

Strengths:

- Specifically deals with scientific documents;
- Can be automatic as information about citations can be extracted;
- It is not depended on too many inputs;
- Important measures such as impact factor, h-index are based on citations; and
- Researchers normally cite relevant papers only; therefore, citation network is already a connection of related papers.

Weaknesses:

- Network of papers are required to compute related documents;
- Unable to discover the nature of relationship between the documents;
- Two documents "A" and "B" are considered relevant if either "A" has cited "B" or "B" has cited "A" irrespective of the area; and

- Citations are made by authors and thus sometimes ceremonial citation can be given just for getting undue popularity.

In summary, the state-of-the-art techniques for identifying relevant papers make use of the above discussed data sources. This section mentioned only the generic characteristics of the systems associated with a particular data sources.
In rest of the chapter, first, we will explain the generic methodologies for each type of techniques that work based on the aforementioned data sources. Afterwards, all of the individual research work has been critically evaluated.

2.2 Methodologies to Identify Relevant Research Papers

For literature review, more than 100 recent papers were considered and among them about 60 relevant papers were selected. We have categorized the techniques that identify and recommend relevant papers into four categories such as: Word level similarity, semantic similarity, collaborative filtering, and bibliographic analysis. It has also been shown in the Figure 2.1. This section discusses generic methodology used by each type of technique to identify and recommend relevant research papers.

2.2.1 Generic Methodology for Computing Word Level Similarity

Summarizing the methodology, in the first step, the document text is filtered. The filtration process normally contains the stop word removal process. Then filtered text is processed to compute the representative word/terms for each document. Finally, to increase the recall of the system, stemming process is performed. Below is the detail for each step:

a) Data Filtration Phase

In the data filtration phase, the stop words are removed from the content so that important terms can be retrieved from the text. Normally, a stop words list is used for this purpose. However, sophisticated approaches can also be used to find frequently occurring terms in a dataset such as: term-based random sampling [Lo et al., 2005], or building domain specific stop words list [Masoud and Kamel, 2008]. At this point, the document text is just filtered version. The representative terms yet not computed. Afterwards, a term extraction process is performed on this filtered version of text.

b) Terms Extraction

There are number of options available for term extraction from the documents. We only list few of them that have been referred frequently in the literature for identification of relevant papers.

i) Yahoo! Term Extractor

Yahoo! Terms Extractor is a tool developed by Yahoo[7] for key term extraction. Its implementation is available via open API. It takes a text of the document as input and returns a list of significant words or phrases extracted from that document. Yahoo term extraction implementations have been enhanced in 2011 and now it is called Yahoo Content Analysis API. According to the Yahoo, "The Content Analysis Web Service detects entities/concepts, categories, and relationships within unstructured content. It ranks those detected entities/concepts by their overall relevance, resolves those if possible into Wikipedia pages, and annotates tags with relevant metadata"[8]. However, its earlier version have been used by number of past studies for extracting Key-terms such as [Afzal, 2009][Wisamet al., 2006] etc.

ii) TF-IDF Based Terms Extraction

TF-IDF (term frequency – inverse document frequency) computes the important Key-terms for a document. It is often used in information retrieval and text mining tasks. This scheme is based on two parts TF- term frequency (the frequency of the term in the document itself, the more the term found the better it is) and IDF, it is based on counting the number of documents in the collection in which term in question contained. It means that y term which occurs in many documents is not a good to be retrieved. The IDF was first published as term specificity [Sparck, 1972] and later on became popular as inverse document frequency (IDF). The TF-IDF computes the relative word frequency in a document compared to the inverse frequency of that word in all documents in the corpus. The words that are common in a particular document or small set of document gets the higher TF-IDF value as compared to the common words across the corpus. The generic form of TF-IDF is shown in Equation 2.1. A well-known indexing and searching service, Lucene[9] make use of TF-IDF score to compute text similarity.

[7] https://www.yahoo.com/
[8] http://developer.yahoo.com/contentanalysis/
[9] https://lucene.apache.org/

$$tf - idf_{t,d} = tf_{t,d} \times idf_t$$

The t *represents the Term and* d *means the document.* *(Eq 2.1)*

$$idf(t, D) = \log \frac{|D|}{|\{d \in D : t \in d\}|}$$

iii) KEA

KEA [Ian H. et al., 1999] is a Key-phrase extraction algorithm that makes use of Machine learning techniques for extraction of key phrases from the text. It was developed by New Zealand Digital Library (NZDL). The KEA generates key phrases for a document which can be used for computing documents similarity. The KEA was used for computing similarities between research articles [Jones &Paynter, 1999].

c) Stemming

Once the terms are extracted, the process of stemming over terms is initiated. Stemming is the process of reducing a word to its stem or root form. The words are stemmed so that morphological variants of a word (i.e. equivalent in nature) can be matched. For example “Production”, “Producing”, and “Produces” belong to the same stem i.e. “produc”. Different stemming algorithms have been proposed in literature such as Porter [Porter, 1980], Paice/Husk [Paice, 1990], Lovins [Lovins, 1968], and Krovetz stemming algorithm [Krovetz, 1993].

Once the terms are extracted and refined, different distance measures are used to compute similarity between the terms representation of the documents. Those similarity measures are usually referred as distance measures. Few of the widely used distance measure are discussed below:

1) Levenshtein Edit Distance:

The Levenshtein distance from one document (A) to another document (B) is the minimum number of character edits needed to transform A-document equivalent to B-document.

Mathematically, the Levenshtein distance between two strings a, b is given in equation 2.2.

$$lev_{a,b}(i,j) = \begin{cases} \max(i,j) & if\ \min(i,j) = 0 \\ \min \begin{cases} lev_{a,b}(i-1,j)+1 \\ lev_{a,b}(i,j-1)+1 \\ lev_{a,b}(i-1,j-1)+[a_i \neq b_j] \end{cases} & otherwise \end{cases} \qquad (Eq\ 2.2)$$

Note that the first element in the minimum corresponds to deletion (from a to b), the second to insertion and the third to match or mismatch, depending on whether the respective symbols are the same. A well-know citation index such as CiteSeer [Giles et al., 1998] have used Levenshtein edit distance modified version titled as LikeIT.

2) Jaccard Similarity:

Jaccard similarity, measures the similarity between two sets. It can be computed by dividing the intersection of common terms by total number of terms in documents. [HADDADENE et al., 2012] have proposed to use Jaccard co-efficient for computing document similarity. The generic form Jaccard similarity equation is shown in 2.3.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \qquad (Eq\ 2.3)$$

3) Cosine Similarity:

Cosine similarity is a measure of similarity between two vectors. The term vectors are given as input and the similarity between the vectors are then computed using the equation 2.4.

$$Similarity = \cos(\theta) = \frac{A \bullet B}{\|A\|\ \|B\|} \qquad (Eq\ 2.4)$$

The cosine similarity is commonly used for document clustering [Sandhya et al., 2008]. When the cosine value is 1, the two documents are identical and 0 when there is nothing in common between them.

Finally, the distance measure can also be useful in number of other areas such as Spell checking [Mu et al., 2006], DNA analysis [Sölkner et al.,1998] and Plagiarism detection [Mario et al., 2009]etc.

2.2.2 Generic Methodology for Computing Semantic Similarity

In the previous section, generic methodology of word level similarities between research papers has been discussed. This section presents the generic methodology normally used in computing semantic similarity for identifying relevant documents.

The semantic similarity methodology share few methodological steps of word level similarity. For example, the term extraction and the comparison part of afore-explained technique for identifying relevant papers. However, for determining the semantics of the documents, the extracted terms are further enriched by using different kind of pre-compiled data sources such as WordNet (discussed below).

a) WordNet

WordNet is a large lexical database for English language developed at the University of Princeton [Miller et al., 1990]. WordNet groups the words together based on their meanings and senses. Furthermore, WordNet specifies semantic relationship between the words. The early approaches can be considered as bag of words (Terms), which ignores the relationship between the terms. Whereas the WordNet based approaches extracts generic concepts for all the terms to form more knowledge-wise enriched representation of the document. In literature, the use of WordNet ontology for retrieval of hypernym/hyponymy or senses of words/keywords has been reported [Khan et al., 2012].

b) Wikipedia

Wikipedia is the world's largest collaboratively edited source of encyclopedic knowledge. Therefore, some techniques have made use of Wikipedia to find semantic similarity between the words. Wikipedia have been used to compute semantic similarity between the words [Zhiqiang et al., 2009].

2.2.3 Collaborative Filtering

The collaborative filtering based approaches follow a generic methodology for recommending relevant documents.

a) User Rating Matrix

In the collaborative filtering methodology, a user rating matrix is built. Once user rating matrix is created, then different algorithm of collaborative filtering is applied on user rating matrix. A generic user rating matrix is shown in Table 2.2. Different algorithm computes values for user-rating for example Citations, Downloads etc. Then neighbors are computed and thus new items are recommended to the user.

Table 2.2: User - Item Rating Matrix

	Item1	Item-2	Item-3	….	…	Item-n
User-1	√	√	√			√
User-2				√	√	√
…						√
…						√
User-m	√	√	√	√		√

The success of collaborative filtering in other domains (e.g. retrieving Best seller list, retrieving relevant music, movies etc.) has attracted different researchers to exploit its efficiency in the relevant citations ranking domain. In the context of paper recommendation systems, there are number of research studies which provide evidence of the usage of collaborative filtering techniques. The detail of individual techniques has been explained in section 2.4 of this chapter.

2.2.4 Bibliographic Analysis

Similarly, to its predecessor's techniques, the citation based approaches follow a specific methodology for identifying relevant research papers. In order to find relevant citation, the references of the papers are automatically extracted and citations to papers are marked. Due to the large number of publications, this task involves a great amount of human effort if done manually. Alternatively, an approach for autonomous citation discovery can be applied. On the other hand, autonomous citation indexes services have been built such as Google Scholar and CiteSeer are well known to scientific community. These services automatically extract the

references of the papers and mark the citations. The methodological steps are explained below in detail:

a) References Extraction:

The research papers establish a link to previously published research papers by referring them in their references list. Therefore, at first step these references are extracted. The extracted references are then parsed to identify the citation for a particular paper.

b) Reference Tokenization:

Afterwards, each reference is tokenized to its components. A reference normally has multiple components such as: authors' names, title of the publication, venue, date of publication etc. Different techniques have been proposed in literature for the extraction of these tokens such as FLUX-CIM [Cortez et al., 2007], ParsCit [Isaac et al., 2008] and heuristic based parser [Giles et al., 1998]

c) Link Establishment Between Cited and Cited-by Papers:

Based on extracted component from a reference, the cited and cited-by paper link is established. The well-known citation indexes such as: CiteSeer performs the heuristic based process for establishing the link between cited and cited-by documents [Giles et al., 1998]. There are certain other autonomous citation techniques that have been proposed in literature for the same task such as: Template based Information Extraction using Rule based Learning (TIERL) [Afzal et al., 2010].

2.3 Techniques for Identifying Relevant Papers

In this section, we have critically evaluated all major techniques found in the literature for identifying relevant papers. The techniques found in literature have been organized date-wise i.e. recent on top. Finally, they have been summarized in Table 2.3 of this section.

Relevant paper identification is an active area of research in the recent past. Different techniques have been proposed in literature; recently Pruitikanee et al [Pruitikanee et al., 2013] have proposed a technique for identification of relevant papers consisting of various steps. First of all those papers should be selected which contains at least one keyword of user query. In second step, they have proposed to form fuzzy clusters based on similarity between the documents. In third step, representative papers are

produced by extracting text from the papers satisfying user query. Finally, ranking is performed based on PageRank algorithm. However, they have not shown that how PageRank will be used on documents in which links are not considered. Furthermore, the system has not been evaluated on any dataset. The proposed approach is based on text similarity and thus it will not be able to find nature of relationship between two documents.

PageRank is well-known Google ranking algorithm for web pages[Page and Brin, 1998]. Different researchers have proposed their techniques similar in nature to Google PageRank for identification of relevant papers. For example, Haddadene et al., have proposed a modified version of well-known algorithm i.e. PageRank for academic paper ranking [Haddadene et al., 2012]. In PageRank algorithm, higher authority scores are assigned to those web pages that have many in-links from authoritative pages with relatively few out-links. They have used similarity score between the citing and cited articles. They first compute similarity between the documents' text and then use PageRank for ranking the documents. They have used Jaccard coefficient for document content similarity. The PageRank algorithm, the rank score of a page, p, is equally divided among its outgoing links. It gives the same score to its outgoing links nodes; whereas, the proposed scheme give higher weight to those cited-by papers which has higher text similarity with cited article. However, they have not validated their system and have not provided any results.

A task focused strategy has been adopted to make recommendation in digital library [Yang and Lin, 2012]. A task focused strategy employ task profiles of users (i.e. sets of recently accessed articles), instead of long-term interest profiles. The proposed system has combined common citation analysis, co-author relationship analysis, and citation network analysis technique (i.e. CiteRank algorithm). The proposed system has been evaluated using CiteSeerX dataset. It was found that content-citation approach gives the highest-quality article recommendations. However, fine grained level recommendations are being considered.

CiteSeer algorithm for identification of relevant (CCIDF) has been extended by Huynh et al. [Huynh et al., 2012]. Huynh et al. named their technique as CCIDF+. They have proposed to use the weight of co-citation along-with co-reference in CCIDF. They have conducted a user study to evaluate CCIDF+ based recommendations. Their results indicate improvement in recommendation of relevant

documents. However, evaluation has been performed on very limited scale. Furthermore, the proposed system is not looking into discovering any kind of relationship between the research articles.

In-text citation frequency importance has been recognized by authors such as Hou et al [Hou et al., 2011]. Their experiment is based on biomedical domain. At present, the related papers are considered as those papers that have bibliographically coupled. In their experiment, those papers were taken that were coupled more than 10 times and it was found that those references (i.e. references for couples papers) were highly cited in the body text of the target paper. Their experiments support our hypothesis (will be discussed in next chapters). They have focused on redefining quality measure. However, they have not talked about discovering nature of relationship between the papers with the help of in-text citation frequency. We have exploited in-text citation frequencies and found that in-citation frequency and patterns can be used for discovering nature of relationship between cited and cited-by documents [Shahid et al., 2011].

Citation Authority Diffusion (CAD) technique was proposed to identify and analyze important papers from survey articles [Chen et al., 2011]. This improves the novelty of literature survey. Their system is called "Survey Importance Measurement (SIM)" which is available online as a web service. For experimentation, they selected all the academic papers published before year 2008 from CiteseerX dataset. There were a total of 456,787 unique papers and 1,612 papers with quality references were selected as the testing dataset. Apart from CiteSeer dataset, their system extracts survey papers from Google Scholar and builds a citation graphs for the papers which have been referred in the references. Furthermore, the novel papers are selected using number of procedures such as: calculating author (cites), author (cited), and common references in the cited papers. The system achieves a reasonable accuracy; however, there are some concerns that need to be addressed such as: 1) there is still a lot of articles not referenced. Therefore, it would lose the novelty assessment, 2) the system has been tested for CiteSeer limited dataset, therefore, a generalization is required for huge datasets, 3) the system is only tested for computer science literature, however, this may or may not get reasonable accuracy for other domains. Furthermore, the system does not exploit paper semantics e.g. thus the system will not be able to determine how strongly two papers are connected to each other.

The relevant paper identification task has also been exploited in literature based on common authors, common references and citations analysis [Taheriyan, 2011]. The proposed technique performs well when the network of paper is dense. They have used ACM (255) papers for verification of their system. The overall precision and recall of the system is good. However, the performance is based on the graph density (network of the paper share more information). For example, the selected set of paper contains 115 common authors, and similarly, 205 are direct relationship between the documents. Furthermore, as it considers only surface level detail of the paper (authors, references and citations) and thus would not be able to identify relationship nature between the cited and cited-by documents.

Citation context should be determined in order to recommend citations [He et al., 2010]. For that purpose, placeholders are provided in text of the paper where citation is needed. In order to recommend citations, the paper content (Global context) and placeholder citation (local context) is considered. For each citation placeholder, the words surrounding the placeholder are collected as the context of the placeholder. Total 100 words for citation context are considered, 50 words before the placeholder and 50 words after the placeholders. Then context obvious (Top N paper whose abstract and title are written by the same team of author's etc.) and context aware methods (Top N papers whose in-links are similar to the target paper and Top N papers whose out-link are similar to the target paper) have been used to retrieve the relevant documents. At the end the Context-aware relevance model is compared with other baseline methods in global recommendations and proved that their system outperformed others. However, the results have been verified on around 1500 papers of CiteSeer and thus require to be validated on larger dataset. Furthermore, the result on local context needs to be improved as only 34% relevant papers were found in top 5 papers.

The system presented by Zhang and Li [Zhang and Li, 2010] creates user profiles, based on the documents viewed by the user. Concept tree is built for each user based on the documents that he/she has viewed. This is different from the earlier approaches in a sense that predecessor creates the keyword based vectors. It will overcome the sparseness and semantic ambiguity issues of keywords based vectors approaches. Afterwards, the correlation strength between users is computed using tree-edit distance. The users of similar interest are discovered by exploiting the relationship

between the users. They have used the dataset of National Science and Technology Library. Furthermore, they have proposed to build a concept tree instead of word vectors so they should have provided a comparison of performance gain. Finally, as this system can be classified as collaborative filtering based techniques so it can only retrieve relevant papers while ignoring the nature of relationship between the documents and will also have general issues listed in section 2.1.3.

Scienstein, is a paper recommender system proposed by Gipp et al [Gipp et al., 2009c]. In this recommender system the authors have adopted a hybrid approach such as a) content analysis, b) Citations, c) Authors, d) Sources: Further papers of that source, and e) rating scores to discover relevant papers. In content analysis, they have proposed to use Naive Bayesian and Support Vector machine to calculate similarity between the documents. In citations part, they have proposed to use co-citation, bibliographic coupling, cited by approaches to find relevant documents. They have also proposed an extended version of co-citation [Gipp et al., 2009a]. The Citation Proximity Index (CPI) is calculated i.e. if two in-text citations occur in the same sentence then its CPI will be 1 and CPI of in-text citation at paragraph level will be 1/2 and so on. Thus, if two co-cited paper will have higher CPI, those will be highly related. In [Gipp et al., 2009b] they have also proposed Citation Order Analysis (COA) in which the order of in-text citation is considered to find similar documents. In the metadata part, they have proposed to use the authors and source information to extract relevant research. In collaboration filtering part, explicit and implicit rating is proposed to be used to find similar users. It is argued that, through explicit rating mechanism, number of objective can be achieved such as: author's own recommendations improvements. They have also proposed the usage of implicit rating. They defined 22 different actions to monitor the users such as downloading, printing, viewing and editing detail of document etc. In summary, authors have combined approximately all of the known techniques. However, the system has not been verified on any dataset. It is just hypothetical system. Moreover, as this system has proposed the usage of all well-known techniques for relevant document retrieval, then different questions arise such as: which techniques weight should be given more preference, what should be the order of execution of different techniques etc. Furthermore, the limitation associated (as discussed in the above sections) with those techniques will also become part of the system.

Co-citation technique has been extended recently by different authors such as: Gipp et al [Gipp et al., 2009a], and Liu and Chen [Liu and Chen, 2011]. The authors of these scholarly works have analyzed the distributions of co-citations at four levels of proximity (such as journal articles, organizational sections in articles, co-citation frequency groups, and roles of co-citations) with reference to corresponding traditional co-citation network [Liu and Chen, 2011]. They found that sentence-level co-citations play a predominant role in forming the overall co-citation network. Their results indicated that sentence-level co-citations are potentially more efficient candidates for co-citation analysis because they tend to preserve the essential structural components of the corresponding traditional co-citation network and they tend to appear much infrequent in comparison to loosely coupled article-level only co-citations. However, the proposed system cannot identify the nature of relationship between documents because, their proposed system does not consider cited and cited-by papers semantics.

In collaborative filtering, the rating matrix (paper-citation) is used to mimic the user-item rating matrix. However, the rating values that are used between the paper and citation are Boolean (either cited or not). This problem was investigated by [Vellino, 2009] and proposed to use a well-known PageRank algorithm to replace the Boolean values by PageRank (citations) values. Their results show that using the PageRank values in rating matrix decreases the quality of recommendations. However, it is recommended to apply more experiments to get more insights and it will not be wise to conclude that PageRank algorithm is useless in recommending research articles in the context of collaborative filtering environment.

Papyres [Naak et al., 2008] is a paper management system that offers services such as management of scientific resources in efficient way, retrieval of relevant documents. Papyres has proposed hybrid cascade approach (one's output is used as input for others). The technique proposed in this system uses the features of items, and furthermore the user's interest features is used to make the recommendations. The Papyres accumulates user ratings of items, identifies users with common ratings, and offers recommendations based on inter-user comparison. The inter-user correlation is then evaluated based on Pearson correlation to find the suitable researchers. Papyres takes 30 researchers as neighbors. Furthermore, Papyres made use of links and RSS feeds for recommendation that researchers have stored or subscribed. However, the

evaluations of the system results are not provided and authors themselves mentioned this is as future work. Furthermore, the recommendations are based on user's interest and thus limitation of collaborative filtering will entails in their system.

The evaluation of the web also opened new venues for identification of relevant documents. In web 2.0, the users are able to annotate, bookmark and comment on different objects of websites. CiteULike is one such service specially designed for research community. CiteULike dataset have been used by researchers such as Bogers and Bosch [Bogers and Bosch, 2008], they applied various collaborative filtering mechanisms to discover relevant documents. They have collected five different type of metadata from CiteULike Topic-related (e.g. article topic, title etc.), Person-related (authors, publisher etc.), Temporal (e.g. publication info), Miscellaneous (e.g. publisher details, volume and number information etc.), and User-specific (tags, comments and reading priorities). The user based and item based recommendation algorithms have been used. It was found in the experiment that user based collaborative filtering works good in order to discover relevant information. However, as mentioned in the limitation of collaborative filtering based approaches that they do not consider the internal features of the document and thus they are unable to find either the relationships strength or the relationship nature between the documents.

Digital Library should not work as only information provider; rather it should be a common place where user can collaborate and share and organize their knowledge [Avancini and Candela, 2007]. Based on collaborative information, the system should recommend communities along-with relevant documents to the users/communities of the similar interest. Based on this idea, they have prototyped an application called CYCLADES. The system functionality can be summarized as following a) user can search (ad-hoc search, filtered search, and what's new, on-demand etc.) for information, b) users can organize the information space in folder paradigm based approach, c) collaborate with others having similar interests, and d) get recommendations. As mentioned earlier, the recommendation is not limited only to relevant documents rather user, communities and collections for searches are also recommended. The similarity is computed between the user's actions and folders (containing different documents). Cosine similarity measure is used to compute similarity and recommendations are made by the system. The documents contents are extracted and thus only those documents will be retrieved that are content-wise

similar. Therefore, all those limitations that exist in content based approaches will also be part of this system. Furthermore, the documents are selected based on similar users (i.e. user having similar interest), the system will be unable to determine by what relationship two documents are connected.

Citation recommendation for a paper was explored by Strohman et al [Strohman et al., 2007]. They have used Rexa dataset that contains about one million documents. There model consists of two stages, in first stage 100 relevant documents are retrieved and then citations of those documents are added to enrich the relevant document result. Furthermore, the documents are ranked on different features such as text similarity, publication year, citation count, co-citation etc. and finally their weights are combined to compute the relevant documents. It was noted that surprisingly all other measures outperform the text similarity feature. The publication date, citation count and title of the paper add very little to calculate overall relevancy of the document. Therefore, this study suggests exploring of new techniques that could extract relevant documents more precisely.

Pohl et al [Pohl et al., 2007] used http-server logs to check and compare with co-citation. Co-citation needs more time to be implemented; because citations data are not available on time as compared to digital access data. Based on digital access records, co-downloads were calculated and demonstrated that co-downloads out performed co-citation for recommendation on recently published articles in arXiv system. However, there are some concerns about the effectiveness of the adapted procedure because simply co-downloading two papers does not mean they are related. Therefore, it is needed to have more statistics on quality of related papers, recommended based on co-downloads. Furthermore, co-downloading feature doesn't ensure that when two papers are co-downloaded then there exist any correlations between them.

As already discussed, Google PageRank algorithm inspired many researchers in the field of identification of relevant documents [Gori and Pucci, 2006]. This work has proposed a biased version of a well-known PageRank algorithm, titled as PaperRank algorithm. First, they build a symmetric matrix for a citation graph, then that matrix is normalized to build stochastic matrix and finally the correlation matrix between the papers is calculated by using biased version of PageRank algorithm. In order to test PaperRank algorithm, ACM Portal Digital Library dataset was built by crawling the

said ACM website. The initial results presented in their paper show that their system recommends related paper with in top 20's list. However, the proposed technique is unable to determine refined relationship (discussing same problem, working on same idea etc.) between two documents.

User profile based recommendations of scientific papers have been seen by different researchers such as [Sugiyama and Kan, 2010]. In their work, the authors' profiles are made enriched with the help of terms extracted from their past papers, reference of their papers, and citations of the papers. The authors have been divided in two categories i.e. junior researchers and senior researchers. The junior researchers are those who have written only single paper, whereas, senior are those who have written more than one paper. They have evaluated their proposed system based on well-known measure of NDCG (Normalized Discounted Cumulative Gain) and MRR (Mean Reciprocal Rank). They asked the 15 junior and 13 senior researchers to mark the relevancy of the recommendations. It was found that recommendation accuracy was increased when context of the paper in form of references and citations were considered. Furthermore, the results improved when un-important references were pruned. They have also improved their technique and results [Sugiyama and Kan, 2013] by incorporating collaborative filtering feature and allocating weights to different terms occurring in various logical sections (abstract, introduction, conclusion etc) based on their importance. However, there are certain concerns; for example, the proposed system will not be helpful for novice researchers who have not yet published any paper. Furthermore, relevant documents are considered as binary i.e. either relevant or not relevant, refined recommendations are not provided to the authors. It can be said that only topically relevant papers are recommended to the users.

Links in to the future, the idea was proposed by Maurer, see [Maurer, 2001] and realized by Afzal et al [Afzal et al., 2007]. They have proposed to use topic, author, publication year and citations information for finding relevant documents of focused paper. The idea is that when a reader opens a paper "A" then all of its (A's) relevant papers are displayed to the users that are published later in time, by the same team of authors in the same topic. They idea was initially applied and realized in online Journal: Journal of Universal Computer Science (J.UCS). This idea was also applied on extracting documents from the web [Afzal, 2009]. However, citation of paper does

not mean that there exists a strong relationship between the documents. Thus, recommending documents based on citations may retrieve too many irrelevant documents. Therefore, only those documents should be considered that are strongly connected.

The citation functions identification has been studied by different researchers for long time since 1970's. Various manual citation annotation schemes for motivation have been proposed in literature such as [Garfield, 1964] [Small, 1982]. There are schemes proposed by authors to exploit the citation function such as [Spiegel-Rusing, 1977] [O'Connor, 1982] [Swales, 1990]. The main idea is to exploit the content around the citation. Automatic citation classification function is a complex and challenging task. For automatic classification of citation, few authors have suggested to use cue phrases [Tuefel et al., 2006] to categorize the citation function. Whereas, co-reference chain analysis have also been proposed by some authors such as [Kaplan et al., 2009]. These techniques use the citation content and manual annotation is performed. In order to predict citation function classes, machine learning techniques are applied. However, the proposed approaches are evaluated on small size of dataset. For example [Tuefel et al., 2006] has evaluated their approach on 116 research articles while [Kaplan et al., 2009] have taken 4 cited papers and their citation from CiteSeer and thus have collected 38 papers. These systems classify the relationship between the documents based on citation content. Thus, based on single author's sentence, a citation would be classified which does not mean that cited and cited-by documents are strongly connected to each other. Therefore, it is important to have such a system that first consider all of the in-text citation occurrences of a citation and then decide about citation function. Furthermore, [Tuefel et al., 2006] results also report that citation classification at abstract level has got accurate results as compare to deep level.

Papit is paper management and sharing system [Watanabe et al., 2005]. It allows the users to share papers, and system recommends papers, classify and retrieve papers as per demand. For recommender task, the user's models are constructed based on their research paper's viewing history. Scale free network idea has been used to compute user's interest. The words are considered as vertices and the edges are computed based on their co-occurrence score in documents. When user views another paper, new vertices and edges are created and thus network is further scaled. The system

computes user's interest. However, the system doesn't consider the paper semantics, thus it is not possible to compute relationship between the papers. Furthermore, the proposed system works on basic assumption that every paper viewed by the users will be considered as relevant which might not be true every time. Therefore, the user's computed interest may lead the system to recommend irrelevant papers.

McNee et al proposed a system which takes a target paper and its references as input, the system then recommends more citations [McNee et al., 2002]. They have used six different algorithms (four collaborative and two non-collaborative) for recommending citation for a target paper. The four collaborative algorithms are: Co-Citation matching, User-Items CF, item-item CF and Naïve Bayesian classifier. The two non-collaborative filtering algorithms are: Localized Citation Graph Search and Keyword search (Google). For collaborative algorithms, the rating matrix between the users (papers) and items (citations) has been used so that collaborative filtering startup problem may not occur. ResearchIndex (NEC Research Institute (CiteSeer)) dataset was used for overall experiment, two experiments (offline and online) were performed. Three metrics such as: a) rank (precision), coverage (recall), and c) effective coverage (f-measure) were used to evaluate the overall results. In offline experiment, it was found that Item-Item, User-Item, and Graph Search all perform substantially better than the co-citation or Bayesian classifier; whereas, in their online experiment, the questionnaire was designed and an online survey was conducted to evaluate all of the six recommender algorithms. They found that Google produce least novel and most familiar results while the item-item and user-item produces less relevant but more novel items. It was found that novelty and relevancy is inversely related to each other. At the end, they were unable to nominate any algorithm that best suites in all of the cases. Their result shows that choice of algorithm affects the overall recommendation results. Furthermore, they have also presented some pitfalls [McNee et al., 2006] which should be avoided when recommending research articles. They have used HRI (Human-Recommender Information) theory to evaluate the overall recommendations. The pitfall are: a) not building user confidence (trust failure), b) not generating any recommendations (knowledge failure), c) generating incorrect results, d) recommendations (personalization failure), and e) generating recommendations to meet the wrong need (context failure).

The RAAP (Research Assistant Agent Project)uses content and collaborative filtering techniques in order to achieve best results[Joaquin et al., 1998]. They have used new text classification algorithm for classifying documents in specific domain. Apart from the document classification, user actions are monitored to support collaborative filtering. The user bookmarks details are used to actively learn their profiles. The system has been tested on limited number of users and the system accuracy is satisfactory. Therefore, it is needed to verify the system for large number of users and check the performance and accuracy of the system. Furthermore, as mentioned earlier, due to collaborative filtering, their system will be unable to find relationship between the scientific documents.

The widely known techniques based on citations are: bibliographic coupling [Kessler, 1963] and Citation based (co-citation [Small, 1973]. Bibliographic coupling is state-of-the-art technique which describes that two papers P1 and P2 is highly relevant when P1 and P2 have larger numbers of common references. Thus, two papers will be strongly related to each other if they have large number of common references. In co-citation the papers P1 and P2 are related to each other if P1 and P2 are co-cited by other research papers. Thus, when two papers p1 and p2 are co-cited in too many other papers, then they are highly related to each other.

These evaluated approaches have been summarized in Table 2.3. The proposed techniques have been listed in second column under the heading of "*Systems*". The third column lists the data sources used by the proposed system and it is labelled as "*Data Sources*". The next column is "*Methodology*" that explains the methodology adopted by the proposed system for identification of relevant documents. Finally, the last two columns of Table 2.3 list the strengths and limitations of each proposed technique.

The critical survey of the literature highlights the following limitations: 1) they do not identify the nature of relationship between scientific documents, 2) the existing systems do not provide ranking in refined relationship between scientific documents, 3) the existing techniques do not provide innovative visualization through which important scientific document can be easily discovered, and 4) finally the ranking provided by the existing techniques are sometime vulnerable and can be manipulated. Before going into summarized results, these limitations have been discussed in detail:

a) Nature of Relationship

When millions of papers are returned as relevant for a focused paper, then user needs filtration mechanism to select a few most interesting and most important papers for reading. For this case, one of the obvious filtration mechanisms would be to identify the nature of relevance (relationship) between papers. There could be different nature of relationships between scientific papers such as: both worked in the same area, both solved the same problem, one paper extends the algorithm/technique of other paper etc. Such kind of relationship types have already been pointed out in the literature [Garfield, 1964] [Teufel, 2006].

b) Relationship Based Strength

When multiple papers of same nature (i.e. Background study etc.) are found in millions of papers, then how those multiple papers should be ranked? It means that there should be some kind of relationship based strength mechanism through which relevant papers can be ranked. The identification of relationship based strength is very important because we have thousands of relevant papers which are talking about same/similar topics, dealing with same/similar problem etc. Furthermore, based on the relationship based strength, it would become meaningful to acquire a ranked list of relevant papers for a source paper within a particular relationship nature category. .

c) Relevant documents Results Manipulation

Until now, we have discussed the relationship based strength, relationship nature and visualization for the retrieved list of relevant papers for a source paper. However, in the area of limitation we are interested to know whether it is possible to manipulate the ranking of relevant papers based on some criteria. For example, on one hand for a scheme which is ranking relevant papers based on citation counts, it is possible to manipulate the results by self-citations or asking for gift citations, however, this is hard to do. On the other hand, a scheme which is ranking the relevant papers based on number of tags received in social bookmarking, it is easier to manipulate the results by creating more accounts and adding more tags to get good ranking.

Table 2.3: Evaluation of Existing Research based Techniques

S.No	Systems	Data sources	Methodology	Strengths	Limitations
1	[Pruitikanee et al., 2013]	Content + Metadata	Word Level Similarity	• Identify relevant documents • Relevant documents Results cannot be manipulated	• Unable to identify nature of relationship • Relationship based ranking is not considered • Different vocabulary issues • Lack of evaluation
2	[Justin et al., 2012]	Citations	Bibliographic Analysis	• Provide visualization for navigation of citation network	• Unable to identify nature of relationship • Relationship based ranking is not considered • Only citation navigation system
3	[Yang and Lin, 2012]	Citations + Content	Word Level Similarity + Bibliographic Analysis	• Identify relevant documents • Relevant documents Results cannot be manipulated • Combination of multiple technique such as CiteRank, content etc	• Unable to identify nature of relationship • Relationship based ranking is not considered • Accessing an article don't quality it to be included in user task profile. Thus, it may lead to get irrelevant results.
4	[Huynh et al., 2012]	Citations + Content	Word Level Similarity + Bibliographic Analysis	• Identify relevant documents • Relevant documents Results cannot be manipulated	• Unable to identify nature of relationship • Relationship based ranking is not considered • Evaluation on limited dataset
5	[HADDADENE et al., 2012]	Citations + Content	Word Level Similarity + Bibliographic Analysis	• Identify relevant documents • Relevant documents Results cannot be manipulated • Have combined content and pagerank algorithm	• Unable to identify nature of relationship • Relationship based ranking is not considered • Lack of evaluation • Vocabulary issues may corrupt the results
6	[Khan et al., 2012]	Metadata	Word Level Similarity	• Identify relevant documents • Relevant documents Results cannot be manipulated • Less amount of resources are required for implementation of the proposed system	• Unable to identify nature of relationship • Relationship based ranking is not considered • La ck of evaluation • Dependency on third party tool
7	[Hou et al., 2011]	Citations	Bibliographic Analysis	• Identify relevant documents • Relevant documents Results cannot be manipulated • Static in nature	• Unable to identify nature of relationship • Relationship based ranking is not considered • Considered in-text citation frequencies in bibliographic coupling
8	[Chen et al., 2011]	Citations + Metadata	Bibliographic Analysis + Word Level Similarity	• Identify relevant documents • Relevant documents Results cannot be manipulated • improves the novelty of literature by filtering out well-known research documents	• Unable to identify nature of relationship • Relationship based ranking is not considered • Evaluation on limited dataset • Paper semantics are not considered
9	[Taheriyan, 2011]	Citations + Metadata	Bibliographic Analysis + Word Level Similarity	• Identify relevant documents • Relevant documents Results cannot be manipulated	• Unable to identify nature of relationship • Relationship based ranking is not considered • Evaluation on limited dataset • Results are missing for a sparse graph
10	[Liu and Chen, 2011]	Citations + Metadata	Bibliographic Analysis + Word	• Identify relevant documents	• Unable to identify nature of relationship

			Level Similarity	• Relevant documents Results cannot be manipulated	• Relationship based ranking is not considered • Long List of results are produced • Evaluation on limited dataset
11	[Sajid et al., 2011]	Metadata	Word Level Similarity	• Identify relevant documents • Relevant documents Results cannot be manipulated • Less resources are required to implement this technique	• Unable to identify nature of relationship • Relationship based ranking is not considered • Lack of proper evaluation • Only evaluated on small dataset
13	[He et al., 2010]	Citations	Bibliographic Analysis	• Identify relevant documents • Relevant documents Results cannot be manipulated	• Unable to identify nature of relationship • Relationship based ranking is not considered • Window size determination is important to discover citation context • Accuracy of results are very low for local citation context i.e. 34%
14	[Zhang and Li, 2010]	Content + Collaboration	Collaborative filtering + Semantic Similarity	• Identify relevant documents • Relevant documents Results cannot be manipulated • Overcomes the sparseness and semantic ambiguity issue of Keywords based vectors	• Unable to identify nature of relationship • Relationship based ranking is not considered • Lack of evaluation detail
15	[Kaplan et al., 2009]	Citations + Metadata	Bibliographic Analysis + Word Level Similarity	• Identify relevant documents • Nature of relation • Relevant documents Results cannot be manipulated	• Relationship based ranking is not considered • Evaluation on small dataset • Only noun phrases are considered
16	[Gipp et al., 2009a]	Citations + Content	Word Level Similarity + Bibliographic Analysis	• Identify relevant documents • Relevant documents Results cannot be manipulated • Consider in-text citations thus incorporating only important citations	• Unable to identify nature of relationship • Relationship based ranking is not considered • More resources are required to compute CPA. • Accuracy detail of accurate identification of in-text is not given •
17	[Gipp et al., 2009b]	Citations + Content	Word Level Similarity + Bibliographic Analysis	• Identify relevant documents • Relevant documents Results cannot be manipulated • System performance were enhanced by incorporating the order of in-text citations	• Unable to identify nature of relationship • Relationship based ranking is not considered • Evaluated on small dataset
18	[Gipp et al., 2009c]	Content + Metadata+ Collaboration + Citations	Word Level Similarity + Collaborative filtering + Bibliographic Analysis	• Identify relevant documents • Consider user similarity for identification of relevant documents • Also considers the state-of-the-art techniques co-citation and bibliographic coupling • Relevant documents Results cannot be manipulated	• Unable to identify nature of relationship • Relationship based ranking is not considered • System evaluation results have not been provided

19	[Vellino, 2009]	Collaboration + Citations	Collaborative Filtering + Bibliographic Analysis	• Identify relevant documents • Relevant documents Results cannot be manipulated	• Unable to identify nature of relationship • Relationship based ranking is not considered • Low quality results • Recommended for further experiments
20	[Naak et al., 2008]	Metadata+ Collaboration	Collaborative Filtering + Word Level Similarity	• Identify relevant documents • Relevant documents Results cannot be manipulated • Make use of RSS feeds and Subscription of user for recommendations	• Unable to identify nature of relationship • Relationship based ranking is not considered • Lack of proper evaluation • Collaborative filtering issues have been ignored like cold start problem etc.
21	[Bogers and Bosch, 2008]	Metadata + Collaboration	Collaborative Filtering + Word Level Similarity	• Identify relevant documents • Easy to setup and less resources are required to implement the proposed technique	• Unable to identify nature of relationship • Relationship based ranking is not considered • Relevant documents Results can be manipulated • Collaborative filtering issues are not considered and thus the proposed system may provide irrelevant documents
23	[Avancini and Candela, 2007]	Content + Collaboration	Word level Similarity + Collaborative Filtering	• Identify relevant documents	• Unable to identify nature of relationship • Relationship based ranking is not considered • Relevant documents Results can be manipulated • Vocabulary issues • Collaborative filtering issues like cold start
24	[Strohman et al., 2007]	Content + Metadata + Citation	Word level Similarity + Bibliographic Analysis	• Identify relevant documents • Relevant documents Results cannot be manipulated	• Unable to identify nature of relationship • Relationship based ranking is not considered • Katz graph ignores the weights on the edges between the cited and cited-by papers
25	[Pohl et al., 2007]	Citations + Collaboration	Collaborative Filtering + Bibliographic Analysis	• Identify relevant documents	• Unable to identify nature of relationship • Relationship based ranking is not considered • Relevant documents Results can be manipulated • Co-downloading two papers do not mean they are related
26	[Afzal et al., 2007]	Metadata + Citations	Word Level Similarity + Bibliographic Analysis	• Identify relevant documents • Relevant documents Results cannot be manipulated	• Unable to identify nature of relationship • Relationship based ranking is not considered • All of the citations are not strongly relevant
27	[Sugiyama and Kan, 2006] [Sugiyama and Kan, 2007].	Citations + Content	Word Level Similarity + Bibliographic Analysis	• Identify relevant documents • Relevant documents Results cannot be manipulated	• Unable to identify nature of relationship • Relationship based ranking is not considered • The proposed system will not be help for novice researchers who have not yet published any paper
29	[Gori and Pucci, 2006]	Citations	Bibliographic Analysis	• Identify relevant documents • Relevant documents Results cannot be manipulated • Crawled ACM digital library for experiment	• Unable to identify nature of relationship • Relationship based ranking is not considered • Recommend relevant papers in top 20's list. The results should be given for article recommended in top 10 results.
30	[Tuefel et al., 2006]	Citations + Content	Word Level Similarity +	• Identify relevant documents • Relevant documents Results cannot be	• Unable to identify nature of relationship • Relationship based ranking is not considered

			Bibliographic Analysis	manipulated	• Low quality results on refined classifications • Semi automatic approach
31	[Watanabe et al., 2005]	Content + Collaboration	Word level Similarity + Collaborative Filtering	• Identify relevant documents	• Unable to identify nature of relationship • Relationship based ranking is not considered • Every paper viewed by the users will be considered as relevant which might not be true every time • Relevant documents Results can be manipulated
32	[McNee et al., 2002]	Content + Metadata+ Collaboration + Citations	Word Level Similarity + Collaborative filtering + Bibliographic Analysis	• Identify relevant documents • Relevant documents Results cannot be manipulated	• Unable to identify nature of relationship • Relationship based ranking is not considered • Evaluated different techniques
33	[Joaquin et al., 1998]	Content + Collaboration	Word level Similarity + Collaborative Filtering	• Identify relevant documents	• Unable to identify nature of relationship • Relationship based ranking is not considered • Relevant documents Results can be manipulated • Evaluation on small dataset • Did not consider collaborative filtering issues
34	[David et al., 1996]	Citations	Bibliographic Analysis	• Identify relevant documents • Relevant documents Results cannot be manipulated	• Unable to identify nature of relationship • Relationship based ranking is not considered • Citation navigation facility
35	[Small, 1973] - Co-Citation	Citations	Bibliographic Analysis	• Identify relevant documents • Relevant documents Results cannot be manipulated • Dynamic in nature	• Unable to identify nature of relationship • Relationship based ranking is not considered • Don't consider the importance of citation in cited papers
36	[Kessler, 1963] - BCoupling	Citations	Bibliographic Analysis	• Identify relevant documents • Relevant documents Results cannot be manipulated	• Unable to identify nature of relationship • Relationship based ranking is not considered • Static in nature • Bibliographic coupling is considered only on the basis of common references instead important references

2.4 System Applications

In the previous section, the research work has been evaluated. In this section, we are going to evaluate the system application used for relevant research documents identification.

2.4.1 Search Engines

Search engines are the most widely known for searching relevant information. The most widely known search engines are Google, Bing (from MSN) and Yahoo.

The search engines retrieved millions of generic results. As we already know that search engines scope is not limited to scientific domain only, so the user has to filter out the retrieved results in order to find relevant documents. Therefore, user will need a reasonable amount of time to find such relevant information (as mentioned at start of the section). The next option (which is vastly used by the scientific community) is to search relevant documents using Citation Indexes.

2.4.2 Citation Indexes

Citation Index is bibliographic database which indexes publications and their citations. The citation indexes which have been used in this experiment are Google Scholar and CiteSeer. The Google Scholar indexes publication for all disciplines of scientific domains while the CiteSeer indexes only publications of Computer Science domain.

The first widely known research in this category is CiteSeer scholarly work. They have also developed a research based product which will be discussed in the next section. CiteSeer [Giles et al., 1998] is well known citation index; its scope is only limited to computer science domain. CiteSeer extracts documents from the web that are either in PDF or PS format. Then the terms are extracted from the downloaded documents and three different techniques are used to find relevant documents such as 1) word vectors: TFIDF scheme is implemented where stem words are used to represent a document, 2) String distance measure (LikeIT) is used to calculate similarity between headers of the documents and 3) in last common citations: Common Citation x Inverse Document Frequency (CCIDF) have been used to determine related documents. Besides that, [Bollacker, 2000] have also adopted Information Filtering (IF) mechanism to recommend relevant and interesting research to the user. The user's profile is adaptive which is built by manual adjustment and

machine learning. The user's interest profiles (set of keywords, URLs, word vectors citation vectors) are built either by manual adjustment or by observing user behavior (browsing or responding to recommendations). Thus, in this way any document features matching user profile feature is recommended to the user. The recommendation either sent via email or through web interface.

Google Scholar is another major citation index service that is not only specific to computer science as CiteSeer is. Google Scholar ranking algorithm is unknown. [Beel and Gipp, 2009a] in their study reversed engineered the Google Scholar algorithm, and found that Google considers various aspects of articles while ranking documents such as: considering search term in text of the articles and title of the article. However, its ranking weight is lower as compared to other criteria such as citation count. Google Scholar considers citation count as major measure in ranking the documents [Beel and Gipp, 2009c]. They also experimented about the article age i.e. old/new article are ranked differently or not. It was found that article age has no significant role in ranking of research articles [Beel and Gipp, 2009b].

2.4.3 Digital Libraries

There are number of digital libraries which provide different services such as organizing online conferencing, managing journals etc. Digital libraries also provide search facilities over the web so that scientific knowledge can be explored. Digital libraries are often considered as deep web i.e. resources have restricted access. The most widely known digital libraries are: IEEE, ACM, Springer and Elsevier etc. The Digital libraries facilitate users by providing search interfaces. User can perform keyword based queries to retrieve results. Digital libraries results are normally displayed in textual format with 10 to 20 documents per page.

2.4.4 Socially Maintained Databases

Socially maintained databases emerged in the era of Web 2.0. In web 2.0 the users are not considered just consumer rather they also produce information. Different applications emerged such as Facebook, twitter etc. Similarly, specialized applications in scientific domain were built where user can bookmark their references, tag and share with others. These are also used for discovering relevant documents. The most widely known scientific applications are: CiteULike, Bibsonomy etc.

These system applications have been evaluated and summarized based on defined evaluation criteria. The evaluation summary is shown in Table 2.4.

Table 2.4: Evaluation of Research based Products

	Strengths	Limitations
Search Engines		
Google Yahoo	• Generic documents • Widely known • Large number of users	• Ranking of documents can be manipulated • Scope not limited to scientific domain • Millions of generic hits • Long list of irrelevant results • Relevant documents results can be manipulated
Citation indexes		
Google Scholar CiteSeer	• Scientific Community • Autonomous Services • broad coverage	• Unable to identify nature of relationship • Relationship based ranking is not considered • Lack of visualization • Long list of irrelevant results
Digital libraries		
ACM IEEE	• Limited converge • Restricted access	• Unable to identify nature of relationship • Relationship based ranking is not considered • Lack of visualization • Long list of irrelevant results
Socially maintained DB's		
CiteULike	• Enable user to tag documents • Create Reference library	• Unable to identify nature of relationship • Relationship based ranking is not considered • Lack of visualization • Long list of irrelevant results • Relevant documents results can be manipulated

Chapter 3

Proposed Work

In the previous chapter, we have reviewed the state-of-the-art techniques and approaches for discovering relevant documents in detail and critically evaluated the existing systems. One of the approaches for discovering relevant documents is citation analysis. Based on citations, various techniques have been proposed such as bibliographic coupling, co-citations. These techniques are unable to find the nature of relationship between the documents. Furthermore, for exploiting citations reasons, content analysis of citation is performed.

3.1 Hypotheses

Research documents are linked by means of citations. Sometimes, the cited-by papers are in hundreds or even in thousands. All of the cited works are not equally relevant with the cited-by paper. Some citations may be considered very strong, some may be moderate, and some may be weak citations in the context of the citing paper. It is intuitive that cited article with high in-text citation frequency in citing article should be considered as strong citations. But this intuition requires verification/proof. For that matter, this thesis makes the following hypothesis:

Hypothesis 1: In-text citation frequencies have potential to identify the relationship strength (Strong, Medium, Weak) between scientific papers.

As explained above that research documents are linked by means of citations. The authors make citation due to different reason e.g. a paper is cited because cited-by paper works on the same topic, builds its technique based on the techniques mentioned in the cited paper, or sometimes, the cited-by document just cites a particular paper to cover background study. Thus, ranking, based on the nature of relationship between cited and cited-by documents, may help scientific community to find the most relevant documents effectively. Such a discourse analysis has attracted researchers for decades e.g. see [Garfield, 1979]. However, identification of such nature of relationship between cited and cited-by document requires extensive analysis of content [Tuefel et al., 2006]. Broadly speaking, such relationships [Garfield, 1964] between cited and cited-by documents can be classified into two major categories as: 1) methodological relation and 2) non-methodological relation.

The cited papers have the 'methodological relations' when cited-by documents:

a) have worked on the same problem as the cited work has done;
b) have extended/compared their work with cited work; and
c) have used concepts, definitions of the cited work.

The cited papers have the 'non-methodological relations' when cited-by documents:

a) refer cited document only to give background study or highlight the importance of the problem or belong to same/similar area as of cited document; and
b) use the cited document partially (cited work used to complete cited-by paper's methodology).

Based on this discussion, this thesis makes a followup hypothesis:

Hypothesis 2: For a given cited paper, In-text citation frequencies and In-text citation patterns can be used to categorize cited-by papers into two classes such as: 'methodologically related' and 'non-methodologically related' papers.

There are different state-of-the-art techniques which are used to recommend relevant papers. These techniques have been described in detail in Chapter 2. In comparison to rest of the techniques, in-text citations deeply analyze the author intention in citing paper about the cited paper. Therefore, this thesis makes another hypothesis that deals with comparison of results between proposed approach and state-of-the-art techniques:

Hypothesis 3: In-text citation frequencies based approach analyzes cited paper evidences deeply in citing paper text, thus it has potential to produce higher quality results in comparison to other state-of-the-art approaches.

3.2 Methodology to Evaluate Hypotheses

To evaluate the mentioned hypotheses, the following methodology is adopted. Some of the steps of the methodology are required to evaluate multiple hypotheses while some of the steps of the methodology are necessary to evaluate a particular

hypothesis. This detail has been provided with each step to illustrate its requirement for the evaluation of the mentioned hypotheses

Step 1: Comprehensive dataset selection (**this step is required to evaluate all hypotheses**)

Step 2: Pre-processing – Crawling CiteSeer database to download research articles, converting downloaded PDFs to XML format, and extracting references (**this step is required to evaluate all hypotheses**).

Step 3: Identification of sections of citing papers (**this step is required to evaluate hypothesis 2**).

Step 4: Identification of In-text citation frequencies section wise (**this step is required to evaluate all hypotheses**).

Step 5: Evaluation of recommending most relevant papers based on in-text citation frequencies using user study **(this step is only required to evaluate hypothesis 1)**

Step 6: Constructing rules based on in-text citation frequency patterns within different sections (**this step is required to evaluate hypothesis 2.**).

Step 7: In-text citation patterns rules**(this step is required to evaluate hypothesis2)**.

Step 8: Comparisons of proposed approach with state-of-the-art approaches**(this step is only required to evaluate hypothesis 3)**

This chapter explains the methodological steps required to evaluate the framed hypotheses. Therefore, in this chapter we will explain step 1 to 5, and step 7. The rest of the steps 6, 8, and 9 are explained in the next chapter. To perform the steps 1-5, and step 7 this thesis develops a prototype. The proposed system consists of different components such as 1) Pre-Processing-CiteSeer database crawling, 2) Document Parser, and 3) Section Mapper etc. The *Pre-Processing* component involves activities such as document conversion, reference extraction and citation tag identification. The *Pre-Processing* module sends citation tags and documents text to document parser module so that in-text citation of that specific tag in particular document can be determined. The *Document Parser* itself is divided into two sub modules: a) Section Identifier and b) In-text citation frequency calculator. The *Section Identifier* discovers

the document sections. Furthermore, the results of *Citation Tag identifier* module and *Section identifier* module are sent to *in-text citation frequency calculator* module. The *in-text citation frequency calculator* calculates citation frequencies across the sections of the document and the results are persisted in database for further analysis. Different modules of the developed system are explained in the following sections.

3.3 Details of Steps for the Evaluation of Hypotheses

This section explains the methodology steps 1-5in details.

3.3.1 Comprehensive Dataset Selection

In the field of relevant documents identification, there is no standard dataset available for evaluating proposed approaches [Beel and Langer, 2015]. Therefore, similar to predecessors, we had no choice other than creating own dataset. For that matter, we developed a system for acquiring and preparing dataset from CiteSeer [Giles et al., 1998]. The CiteSeer is an open access indexing service that has indexed research paper from computer science domain. Furthermore, it covers all topic of computer science domain and indexes large number of journals and conferences. Therefore, it is suitable resource for preparing dataset and conducting research.

The overall system architecture of data acquiring is shown in Figure 3.1. The system consists of two main parts i.e. CiteSeer crawler module and XML Module. These modules are further divided in sub-modules. The prepared dataset is then shown in the pre-processed section of Figure 3.2. The details of these systems are explained below:

3.3.2 Pre-Processing

The pre-processing is divided into further steps such as 1) CiteSeer Crawler Module, 2) XML Module.

A) CiteSeer Crawler Module

The purpose of this module is to crawl CiteSeer database for given terms and then download those research papers for further analysis. This module works as follow:

i. Topic based Paper's Metadata Extractor

This module starts working by loading extracting topic/searched terms from database. The terms are manually persisted. The topics or searched terms (e.g. ontology engineering, digital libraries etc.) were persisted in database. This module load a particular term and then poses query on CiteSeer database. CiteSeer provides 500 links to the retrieved results. Furthermore, pagination has been applied on retrieved results so that 10 records per page are displayed to end user. This module traverses those records page by page and extracts metadata of each paper and save them in MySQL database.

ii. Citing Papers Extractor

CiteSeer provide citing papers list for a particular paper. Every paper is given a unique identification number called "cid", based on this number all of its citing papers can be extracted. One can access those papers list by clicking cited-by link. All of the citing paper for the current paper is then shown to end user. Again CiteSeer only provide access to top 500 citing papers. Thus, when a paper has more than 500 citations then only 500 are retrievable. Similar to the crawler of previous section, this crawler crawl the citing papers and downloads its metadata. The overall statistics of download papers are shown in Table 3.1

Table 3.1: Term-wise downloaded paper statistics

Terms	Total Papers (Cited and Citing papers
Ontology	1320
Semantic computing	1224
Digital libraries	1336
Semantic web	1120

iii. Papers Downloader

Papers downloader is a separate script that iterates over the metadata of the crawled and downloads the paper. This script updates paper's status when a paper is downloaded. As downloading require lots of time, therefore, this status value help us

to keep track of downloaded and un-downloaded papers. The papers are downloaded in pdf format. So, at this point, we have only research papers in pdf formats.

B) XML Module

The purpose this module is to convert research paper into xml files, so that they can be processed later-on for identification of fine grained information such as in-text citations frequencies of individual reference etc.

i. PDF to XML Convertor

There are different tools available for converting pdf to xml, however, we don't require simple conversion that converts any .pdf document to .xml. A research paper contains structured/semi-structured information that's need to be extracted. Therefore, it was required to either directly acquire information from pdf documents or convert them in more programmable friendly version such as xml. The main intent was to extracts all of its components (especially in-text citations) correctly. Manchester University has developed a tool namely pdfx [Constantin et al., 2013], this tool takes research paper as input and convert in to xml using multiple ontologies like DoCO[10], DEo[11] etc., and also identifies in-text citations in research paper with sufficient accuracy. Thus, this module send research paper in pdf format to pdfx tool using CURL to get it converted in xml format.

ii. XQuery/XPath Solution

In previous section, it was explained how papers are converted into xml format. Therefore, now it was required to extract detailed information about in-text citations and their frequencies. Thus, we developed XQuery and XPath expression based solutions that extract all citations of a research article and then calculates in-text citation frequencies for each reference found in research paper. Along with in-text citation frequencies, this module also extracts section information of in-text citations for exploring in-text citation pattern role in identification of relevant paper finding task. The section extractor/mapper is explained with detail in next section.

[10] http://www.essepuntato.it/lode/http://purl.org/spar/DoCO

[11] http://www.essepuntato.it/lode/http://purl.org/spar/Deo

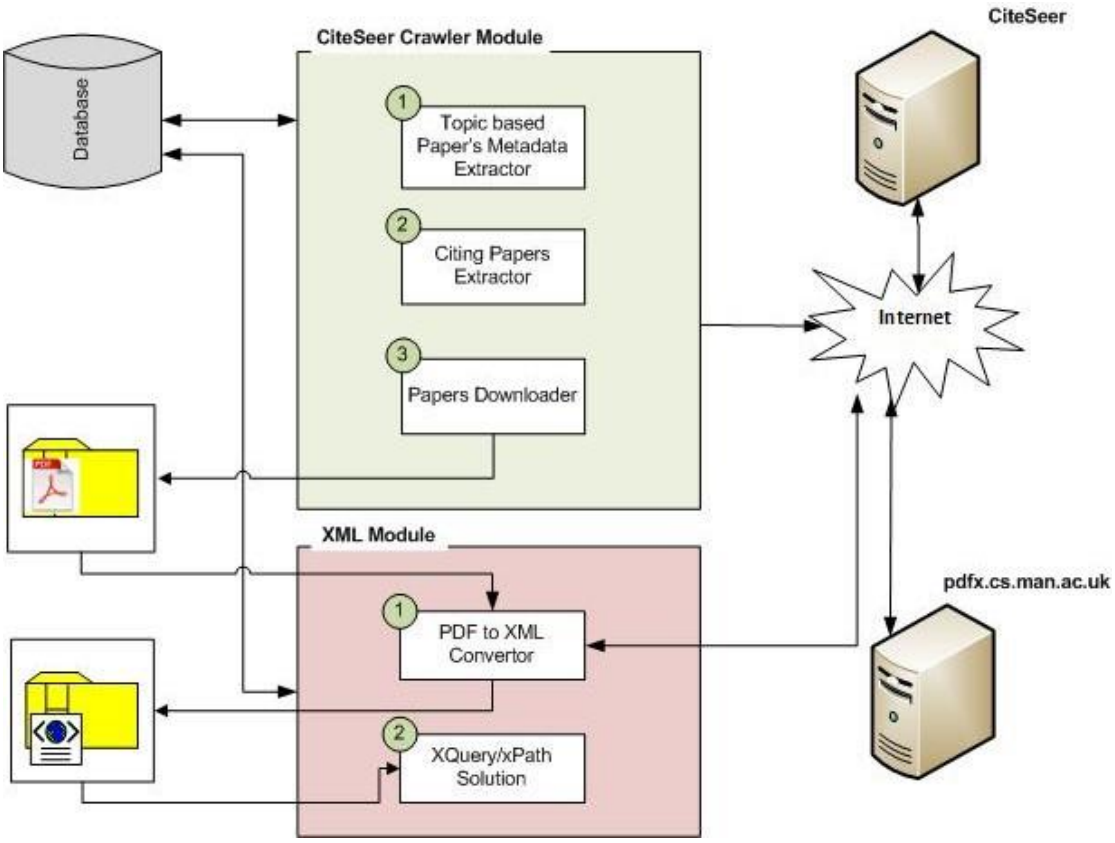


Figure 3.1: System Architecture for Data Preparation

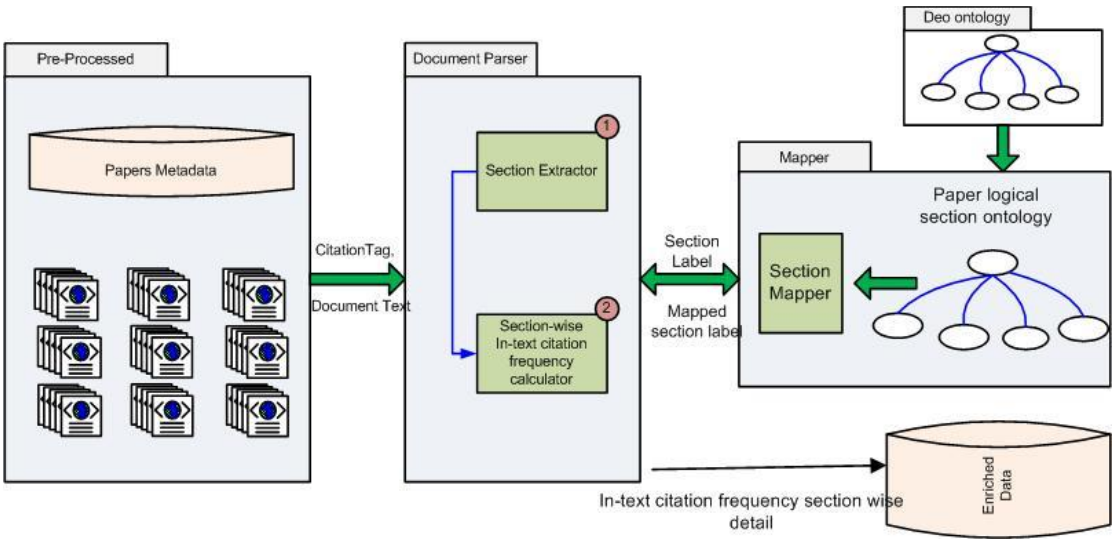


Figure 3.2: Proposed System Architecture

3.3.3 Identification of Sections of Cited-by Papers

The section extractor module identifies the sections in the text of the retrieved document from pre-processor module. The .xml files associated with research document were used for this part. xQuery[12] and Xpath[13] based component was built to extract the paper's sections. The sections of documents could be e.g. "Introduction", "Related work", "Methodology", "Results", "Discussion", and "Conclusion". The extracted sections from some of the papers were manually analyzed and it was found that an author uses different names for the same section. For example, in our data, both section headings "Preliminary Findings" and "Experimental Results and Analysis" represents the section "Results". To understand the problem, a detailed study was performed. The objective of this study was to determine: How often an author uses standard labels for sections in a scientific document? From the standard label, we mean that the names of the sections appearing in the full text of scientific document belong to one/all of the following section labels: "Introduction", "Related Work", "Methodology", "Results", "Evaluation" and "Conclusion". The reason for considering these six sections as standard was that there is scientific agreement of considering them as sections of the scientific documents such as: Kansas State University-Research Paper template [Kansas-template, 2013], Rice University Research-Paper template [Rice-template, 2013] and Boston College University libraries-Research Paper template [Boston-template, 2013].

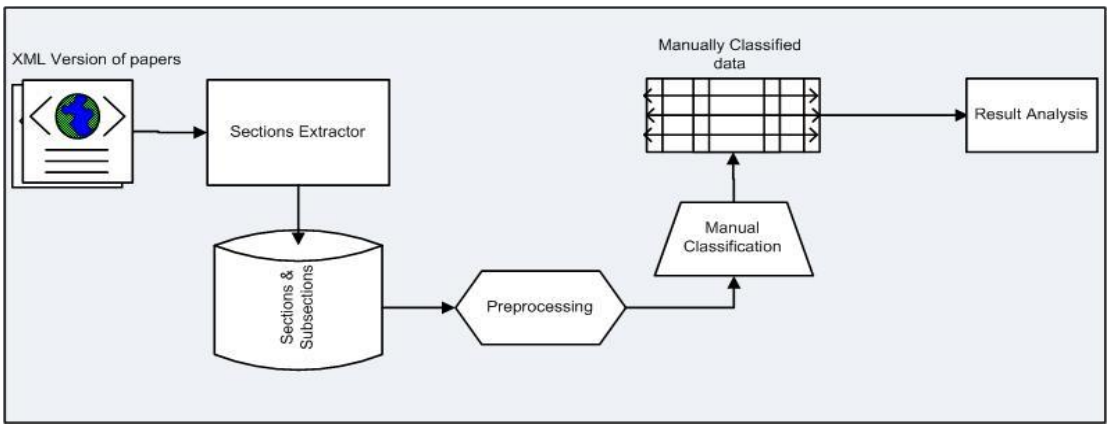


Figure 3.3: Motivational Case Study for Understanding Papers' Sections

[12]http://www.w3.org/XML/Query/
[13]http://www.w3.org/TR/xpath20/

It is important to know the answer to this question, because if authors are already using the same/similar section labels as mentioned above then section-wise classification of scientific documents becomes a trivial task. Therefore, there is a need to perform comprehensive analysis of section names occurring in the scientific documents to get further insights. Therefore, a motivational case study was conducted.

i) Motivational Case Study

The overall architecture of this study is shown in the Figure 3.3. For this study, we performed the following steps: 1) pre-processing of extracted sections 2) manual classification and 3) preliminary results analysis. For this case study, we randomly selected 330 research documents from a total of 1,200 documents. The sections for the 329 papers were 1,833. The detailed procedure of this motivational case study is listed below.

a) Sections filtration

There was a lot of noise in the retrieved sections. For example, some of the retrieved entries were not actually the sections in the real document. Furthermore, one of the major problems was the character encoding. For instance the character "π" was encoded as "Ãƒ Â • Ã¢â€šÂ¬". The text in appendix sections was marked as a separate section. We analyzed text of the extracted sections and used set of general heuristics to remove common discrepancies.

b) Manual Classification

In this step, we manually mapped the retrieved sections instances over the defined standard label of the sections (Introduction, Related Work, Methodology, Results, Discussion and Conclusion). Two human experts (i.e. students that were actively involved in research) classified the sections manually by reading the section titles and looking into the real document. It was made sure to annotate all sections properly as per rhetorical sections labels.

c) Preliminary Results Analysis

The results of the manual inspection are presented in the Table 3.1. The column "Class Name" represents the rhetorical sections which we intend to find in research papers. The value of "Total Papers" column represents the frequency of "Class Name" found in the papers; and "Entries" columns represent the frequency of distinct

section-names found in the papers for each “class name”. The forth column presents the percentages of section-label occurrences which were found from the content of research papers that are same as of the column 1. The fifth column presents the percentages of section-label occurrences which were found from the content of research papers that are different as of the column 1.

One important thing that can be noted easily is that how it is possible that the total number of papers is different in case of “Introduction” and “Conclusion” which is 329 and 263 respectively. The reason for this is that J.UCS contains article of special issues which contain “Introduction” and doesn’t have any “Conclusion” section. There is one other reason which is typically related with conversions from PDF to xml/text which results in some errors during the conversion.

Table 3.1 presents comprehensive statistical insights of 1,833 sections from the 329 research papers. The section “Introduction” was noted as the most compliant section i.e. in 78% of the documents, the section “Introduction” was referred with the same names. However, the section “Methodology” was not referred even a single time with the term “Methodology”. The section “Related Work” was referred with the same/similar terms as “Related Work” in only 30% of the documents. The section “Results” was mentioned with the term “Results” only by 1% of the documents. One important point is that we did not consider only the exact names in the column 1 for comparisons.

Table 3.2: Manual Classification of Sections Label

Class Name	Total Papers	Entries	Section label same as standard sections labels	Section label different From standard sections labels
Introduction	329	378	78%	22%
Related Work	158	184	30%	70%
Methodology	322	829	0%	100%
Results	59	62	1%	99%
Discussion	95	110	20%	80%
Conclusion	263	270	60%	40%

For example, the section “Related Work” can be referred as “Related Work”, “Literature Review”, “State-of-the-art” etc. If an author has mentioned such similar terms for the section, we have considered them as accurate names in the column 4 of

the Table 3.2; however, if a totally different name has been mentioned for the section, then it has been noted in the column 5. For getting more clarity, we have listed some uncommon names for sections ("Introduction" and "Methodology") that have appeared in the text of our sampled data in the Table 3.3. It is obvious from the Table 3.3 that mapping such diversified section names appearing in research papers to standard label of sections is not a trivial task. For example, it is a challenging task to automatically map "historical remarks", "Operator Algebra" and "Preamble" to the standard section "Introduction".

Table 3.3: List of Section Labels in Different article for "Introduction" and "Methodology" Section

Introduction Section–Actual Instances	Methodology Section-Actual Instances
Learning Scenarios in E-Learning	Connection of Process Structures and Communities
Compactness and uniform continuity	Cooperative Knowledge Generation with the Wiki Approach
Historical remarks	Integration of Knowledge Networks into Process-Oriented Structures
Operator algebras	Realisation of the Prototype
From Contents to Activities	Relating Knowledge Processes to Enterprise Business Processes
Definition and Notations	A Survey of Knowledge Management Solutions Source Description
Main Definitions and Problem Statement	Response of academic institutions
Background and Context of Research	Detecting plagiarism
Preamble	Itai-Rodeh Leader Election
Basic Functions of Digital Libraries	Leader Election without Round Numbers
Prologue	Leader Election without Bits
Knowledge Integration	
Preliminaries	
Observation, Computer Systems, and Agents	

ii) Paper's sections mapping over standard sections labels

This has been highlighted in the previous section that the mapping of diversified sections names, appearing in research papers, to the standard section labels is a tough task. In this section, we addressed this important issue by employing a workable methodology.

a) Semantic Structure of Scientific Documents

We evaluated the state-of-the-art systems that have formally defined the structure of scientific documents. The contemporary system such as Semantic Publishing has formally defined the structure of research papers. The Semantic Publishing notion was coined by Seringhaus et al [Seringhaus et al., 2007] [Gerstein et al. 2007]. They suggested the creation of Structured Digital Abstracts (SDAs), which are machine readable documents.

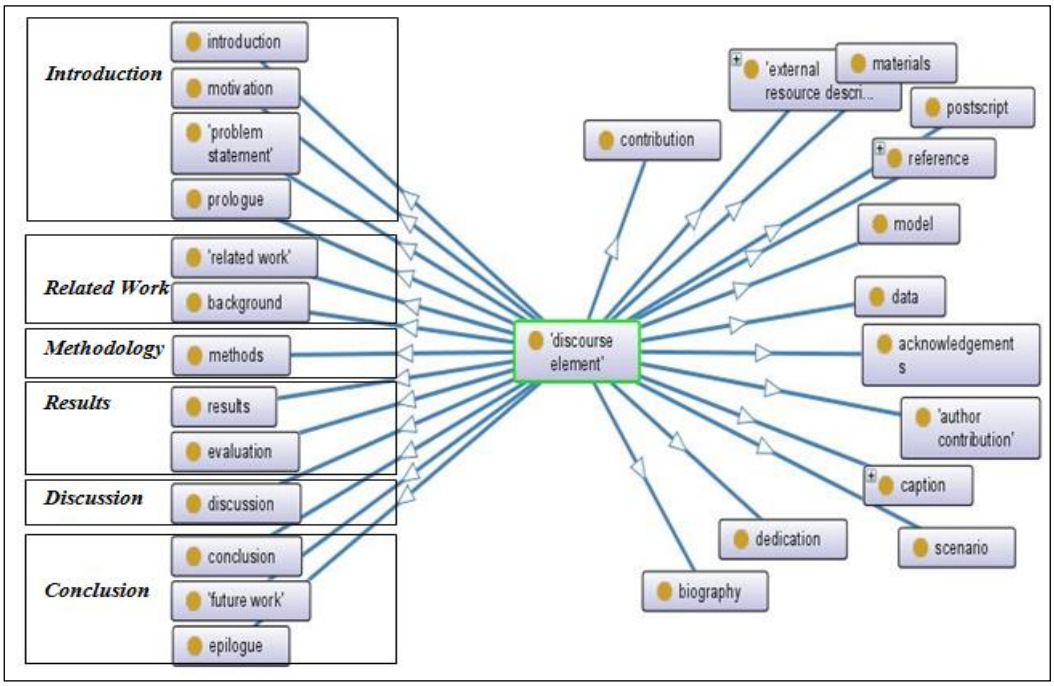


Figure 3.4: Standard Sections Mapping of Research Article over DEo Concepts

Furthermore, Shotton et al. [Shottonet al., 2009] hollow out in this direction and suggested an ontological structure for the purpose of semantic publishing. They proposed an ontology SPAR for structuring scientific documents into standardized format. The SPAR ontologies contain a set of specialized ontologies including DEo.

In DEo ontology, there are different concepts depicting various aspects of research article. The DEo ontology is shown in the Figure 3.4. This ontology presents a comprehensive structure of scientific documents. We have used this ontology to define and populate rhetorical structure of scientific documents. We have grouped up related concepts of this ontology according to the rhetorical sections of research papers on the left part of the Figure 3.4. For example the concept "related work" and "background" of this ontology are grouped under the section "Related Work". The right part of the ontology is of no use for our current task of mapping occurring section names to rhetorical section names. The population of such ontology from unstructured text has not been addressed by the previous systems. This has been addressed in the following section.

Table 3.4: Key-terms used for Mapping

	Standard Sections	**Key Terms**	**Stemmed Terms**
1	Literature Review (LITR)	Background, History, Motivation, Previous, Related, Literature	Background, Histori, Motiv, Previou, Relat, Literatur
2	Results (RES)	Results, Findings, Implementation, Simulations, Experiments	Result, Find, Implement, Simul, Experiment
3	Discussion (DISS)	Discussion, Analyzing, Analysis, Evaluation, Implications, Verification, Comparative, Comparison	Discuss, Analys, Analysi, Evaluat, Implicat, Verif, Compar, Comparison
4	Conclusion (CON)	Epilogue, Future work, Concluding remarks, Conclusion, Final Remarks, outlook	Epilogu, Futur, Conclud, Conclus, Final, outlook

b) Papers' Section Mapping

This section discusses the methodology of mapping papers' section names to standard section names. We have used a layered approach to solve this problem. Firstly, we have built a dictionary of concepts referring to standard sections. This dictionary was built using the basic concepts defined in the DEo ontology and enriching this dictionary by using common labels based on human perception during the analysis of the results presented in motivation section. The standard sections and its corresponding general terms and stemmed version of the terms are referred in the Table 3.4.

Porter's stripping stemming [Porter, 1980] algorithm is used for extracting stemmed key terms so that the words with different endings can be mapped on a single word, for example the words "evaluating", "evaluation", "evaluate" will be mapped to the stem term: "evaluat". Secondly, the paper's template is used for the task of mapping. The template layout gives important information to map especially unknown sections to standard sections labels. For example, introduction would be followed by the related work. Therefore, the first section would be considered as introduction. However, in the case of section "Related Work", as some authors write this section as the second section of the document whereas sometimes, this section appears at the end of the research paper. In this scenario, the dictionary will help to identify the logical corresponding section. The template of a research paper is shown in Figure 3.5. This template is adapted from previous research work [Afzal et al., 2010].

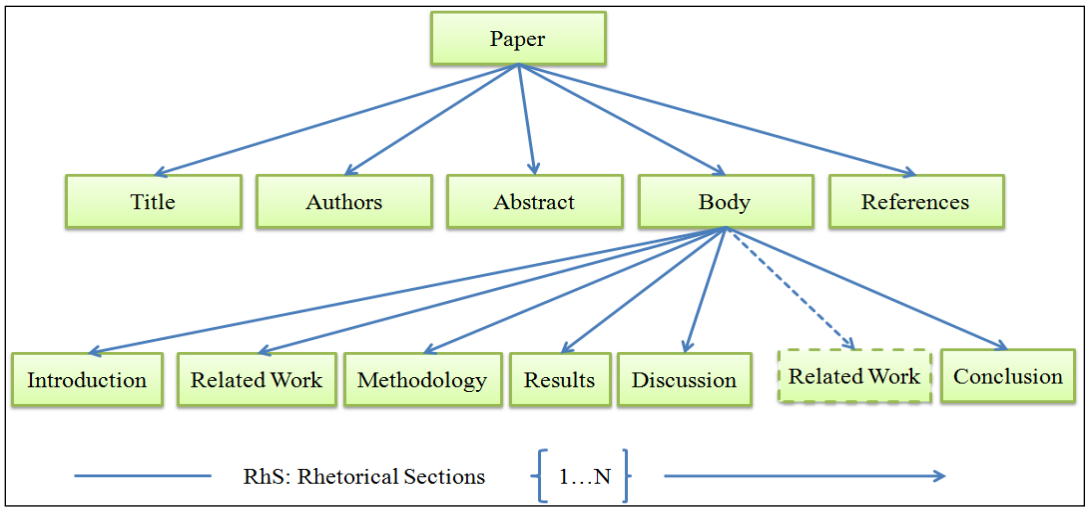


Figure 3.5. Paper Template with Standard Sections

iii) Formal Representation of Sections Mapping Over Standard Sections Procedure

A research article contains various sections which can be represented as below:

$RS_1 = \{INT\}$
$RS_2 = \{LITR\}$
$RS_3 = \{MET\}$
$RS_4 = \{RES\}$
$RS_5 = \{DISS\}$
$RS_6 = \{CON\}$

NOTE: INT = Introduction, LITR = Literature Review, MET= Methodology, RES= Results, DISS= Discussion, CON = Conclusion

Thus a research article sections can be modeled as:

$$\text{Research Article - Sections} = \bigcup_{i=1}^{6} RS_i \qquad (Eq.\ \ 3.1)$$

The upper limit is six which means the research article sections are union of these sections. However, a researcher may or may not use these standard names of sections in his/her research paper. This was also concluded in motivational case study. The sections of document at hand are defined as:

$$\text{Current - Research Article - Sections} = \{S_1, S_2, ...S_n\} \qquad (Eq.\ \ 3.2)$$

Where Sj may or may not be the same as RSi

for each $j = 1, ..., n,\ i =$ 1,2,3,4,5,6

id	section_label
2764	1 Introduction
2765	2 Applying social network analysis in knowledge mana
2766	2.1 Basics of graph theory
2767	2.2 Classifying network analysis metrics
2768	2.3 Specifying wiki networks
2769	3 SONIVIS:Tool
2770	4 Analyzing and evaluating wiki networks to improve
2771	5 Outlook

Figure 3.5. Sections of the Research Paper: "Analyzing Wiki-based Networks to Improve Knowledge Processes in Organizations"

As concluded above that the technique is comprised of two components: 1) Key terms defined for standard sections (See Table 3.3), and 2) Layout information of research paper (see Figure 3.4). The key terms can be modeled for each rhetorical section as below:

$$(\text{KT\_RS})_i = \{T_{i1}, T_{i2}, ...T_{in}\} \text{ Where } i = 2, 4, 5, 6 \qquad \textit{(Eq. 3.3)}$$

Note: $i = 1 = \text{INT}$ and $i = 3 = \text{MET}$

As per assumption, first section will be "Introduction" section; therefore, support from keyword dictionary is not required there. First section can be detected by using the layout information. Furthermore, the terms of "Methodology" section are really diversified, and always evolving over time. Therefore, once all sections have been identified except "Methodology" section, the remaining sections will be considered as "Methodology" section. So, there is no need to use terms for both "Introduction" and "Methodology" section.

For mapping current section of a document over standard sections, Key Terms existence is verified as shown below:

$$\left| \text{KT\_RS}_i \right| = \begin{cases} 1 & \textit{if current section label contains key terms of } KT\_RS_i \\ 0 & \textit{if current section label does not contains key terms of } KT\_RS_i \end{cases}$$

Where $i = 2, 4, 5, 6$ *(Eq. 3.4)*

Thus, with the help of KT-RS and paper layout template, a function for mapping sections of a paper (current research article) over standard sections has been modelled.

$$Map(S_j, RS_i) \equiv$$

$$\begin{cases} \text{INT} & \text{if } j = 1 \\ \text{LITR} & \text{if } j \in \{1, 2, n\text{-}1, n-2\} \textit{ and } KT\_RS_2 = 1 \\ \text{RES} & \text{if } 2 < j < n \textit{ and } KT\_RS_4 = 1 \textit{ and } RES_j > MET_j \\ \text{DISS} & \text{if } 2 < j < n \textit{ and } KT\_RS_5 = 1 \textit{ and } DISS_j > MET_j \\ \text{CON} & \text{if } j = n \textit{ or if } i = n\text{-}1 \textit{ and } KT\_RS_6 = 1 \\ \text{MET} & \text{if } 1 < j < n \textit{ and } S_j \notin \{INT, CON, LITR, RES, DISS\} \end{cases} \qquad \textit{(Eq. 3.5)}$$

Where S_j represents the j^{th} section of the current document and RS represent standard section.

iv) Explanation with Example

In this section the working of algorithm is explained with the help of example. The Figure 3.5 shows the sections extracted from a paper titled, “Analyzing Wiki-based Networks to Improve Knowledge Processes in Organizations”, published in the vol. 14, no. 4 in the Journal of Universal Computer Science (J.UCS). The manual inspection shows that the section with id 2764 can be mapped to the “Introduction” section. The section with id 2765 can be mapped to the “Literature Review” section. The section with id 2769 can be mapped to the “Methodology” section. The section with id 2770 can be mapped to the “Discussion” section, and the section with id 2771 can be mapped to the “Conclusion” section. In the following paragraphs, the importance of proposed approach in using both the key-term dictionary, and the layout information of the scientific document has been shown.

a) Key terms Scenario

The key terms of the current sections are compared with the key terms of standard sections as are shown in the Table 3.4. Stemming is performed using [Porter, 1980] before matching. The key-terms are very important to identify mapped rhetorical sections. For example, with only layout information, it cannot be determined whether the fourth section with id 2770 would be mapped to “Discussion” or to “Methodology” section. This is due to the fact “Discussion” section can lie anywhere in range of section $2 < j < n$. Therefore, in the first step, the key-terms are used to map the sections. In the matching process, the section with id 2764 is mapped to “Introduction” section. The section with id 2765 is mapped over the “Discussion” section because the key term “analysis” is matched. The section with id=2770 was identified as “Discussion” section, again because the key terms “analysis” and “evaluating” is matched. The section with id=2771 was identified as “Conclusion”. In the end, the remaining section with id=2769 was mapped to “Methodology” section. This step enabled to map the section number 1, 3, 4, and 5 correctly. However, the key-term dictionary alone was not able to properly mark the section with id id=2765 and have identified as “Discussion” section.

b) Paper Layout Template Scenario

Paper layout information makes sure that paper sections are arranged in proper order. Thus, by exploiting paper layout information, incorrect marking of sections can be

adjusted. For example, in the Figure 3.5, section 2 with id = 2765 was wrongly marked as “Discussion” section with the help of key-terms dictionary alone; whereas, paper layout information describes that “Discussion” section will come after the “Methodology” section i.e. DISSj>METj. Hence, with the paper layout information, the system is able to correct the wrongly identified section with id=2765. Consequently, the decision of mapping section 2 with id = 2765 over “Discussion” section will be reverted. However, the decision of mapping section with id = 2770 over “Discussion” will remain unchanged. Moreover, the paper layout will further identify that the section with id=2765 is un-mapped, and is between the section “Introduction” and “Methodology”, and the “Discussion” section has already been mapped; therefore, the section with id=2765 will be mapped over “Literature Review” section.

Therefore, it is important to have these two types of information for mapping a research paper sections over standard sections labels.

v) Sections Mapping Results

In this section, we explain the results after applying algorithm as defined in above section over the already manually classified data discussed in motivational case study section. We built the confusion matrix [Perry et al., 1997] which is widely known in scientific community to evaluate classifier results. Confusion matrix helps us in determining how well a classifier can recognize the different classes of the data.

Table 3.5: Confusion Matrix for Defined Classes

Classified as→	Introduction	Methodology	Conclusion	Related Work	Discussion	Results
Introduction	326	46	0	5	1	0
Methodology	2	752	35	3	34	45
Conclusion	0	2	268	0	0	0
Related Work	0	102	2	78	0	2
Discussion	0	37	5	1	61	2
Results	0	8	1	0	5	48

The confusion matrix represents true positive, false negative, false positive and true negative values. We defined different sets for each section that contains the measured values (true positive, false positive) from the proposed classification. These sets contain the recorded values obtained during automatic classifications. The recorded

values are represented with a Count function. For example for “Introduction” class, the recorded value of true positive is shown as below

$$\text{Count (Introduction)} = 326$$

Furthermore, we have calculated the Recall and Precision for each individual class. We have shown below that how precision and recall for the section “Introduction” is calculated. The values of precision and recall for other sections are calculated in the similar way.

Introduction Section

$$Introduction - TP = \{\text{Count(Intrduction)}\}$$

$$Introduction - FN = \begin{Bmatrix} \text{Count(Methodology)}, \\ \text{Count(Conclusion)}, \\ \text{Count(Related Work)}, \\ \text{Count(Discussion)}, \\ \text{Count(Results)} \end{Bmatrix}$$

NOTE: *False Negative values are on horizontal axis in the Table 3.5*

$$Introduction - FP = \begin{Bmatrix} \text{Count(Methodology)}, \\ \text{Count(Conclusion)}, \\ \text{Count(Related Work)}, \\ \text{Count(Discussion)}, \\ \text{Count(Results)} \end{Bmatrix}$$

NOTE: *False Positive values are on vertical axis in the Table 3.5*

Thus by having the values for the true positives and false positives, we can calculate the overall precision and recall as shown below.

$$\text{Recall} = \frac{TruePositive}{TruePositive + FalseNegative}$$

$$\Rightarrow \quad \text{Recall}^{Introduction} = \frac{\text{Count(Introduction}-\text{TP)}}{Count(Introduction-TP) + \sum_{i=0}^{n} \text{Introduction}-\text{FN count(i)}}$$

$$\Rightarrow \quad \text{Recall}^{Introduction} = \frac{326}{326 + 46 + 0 + 5 + 1 + 0}$$

$$\Rightarrow \quad \text{Recall}^{Introduction} = 86\%$$

Now we calculate the precision for the “Introduction” section

$$\text{Precision} = \frac{TruePositive}{TruePositive + FalsePositive}$$

$$\Rightarrow \text{Precision}^{Introduction} = \frac{\text{Count(Introduction}-\text{TP)}}{Count(Introduction-TP) + \sum_{i=0}^{n} \text{Introduction}-\text{FP count(i)}}$$

$$\Rightarrow \text{Precision}^{Introduction} = \frac{326}{326 + 2 + 0 + 0 + 0}$$

$$\Rightarrow \text{Prcision}^{Introduction} = 99\%$$

Finally F1 measure was calculated as below:

$$\text{F1} = 2 \text{ x } \frac{\text{Prcision}^{Introduction} \text{x Recall}^{Introduction}}{\text{Prcision}^{Introduction} + \text{Recall}^{Introduction}}$$

$$\text{F1} = 2 \text{ x } \frac{99 \text{ X } 86}{99 + 86} \Rightarrow 92\%$$

The overall results (precision and recall) are shown in the Figure 3.6 and Figure 3.7. The results are very promising for "Introduction" and "Conclusion" sections. However, for sections "Methodology" and "Discussions", the results are comparatively low. There are two reasons as discussed below: 1) sometimes these sections are not even part of the document; therefore, the algorithm should identify whether these sections are part of the document or not before mapping them, 2) authors use very diversified names for mentioning the names of these sections as was shown motivational case study results. However, the proposed methodology has achieved reasonable results considering the diversified section names as shown in motivational case study results. For example, no one used the term Methodology in the research papers for the section "Methodology", however, our algorithm has achieved 79% precision and 86% recall for mapping unknown names to the section "Methodology". The overall results can be improved by a number of different ways such as a) building a comprehensive list of candidate key-terms (dictionary) for representing a section, and b) exploiting section content.

Finally we can calculate the Macro-Precision and Macro-recall and Macro-F1 values from the values shown in Figure 3.6, Figure 3.7, and Figure 3.8 with the following steps

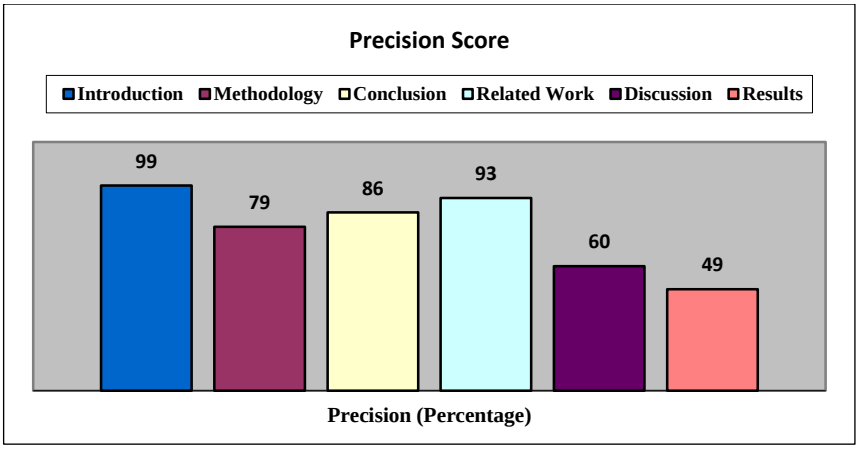


Figure 3.6: Precision Score Received by each Class

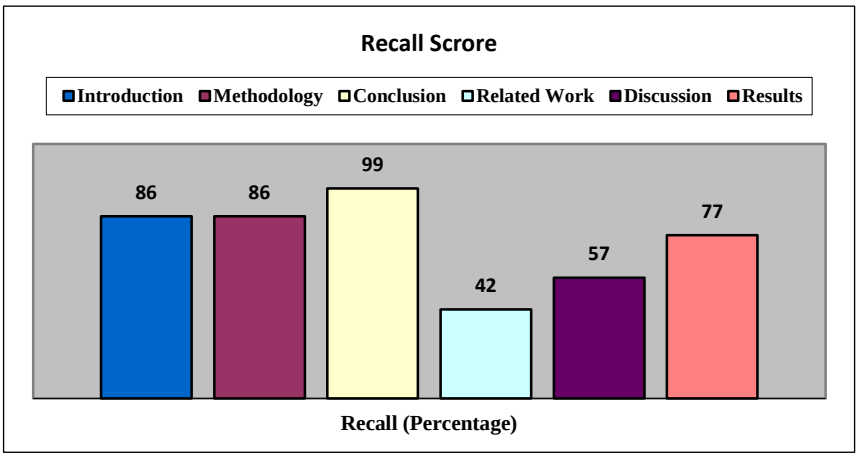


Figure 3.7: Recall Score Received by each Class

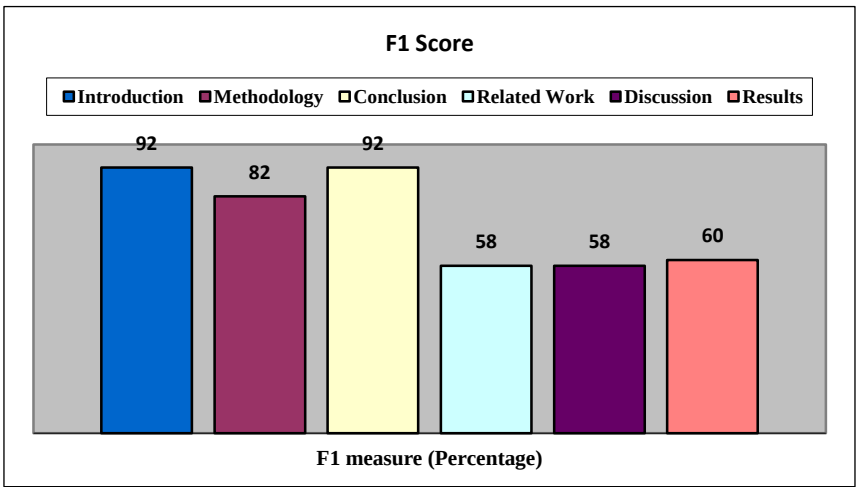


Figure 3.8: F1 Score Received by each Class

Macro Precision

$$\text{Macro} - \text{Precsion} = \frac{1}{\text{n}}\sum_{i=0}^{n}\text{Pi}$$

$$\text{Macro} - \text{Precsion} = 77.6\%$$

Macro Recall

$$\text{Macro} - \text{Recall} = \frac{1}{\text{n}}\sum_{i=0}^{n} \text{Ri}$$

$$\text{Macro} - \text{Recall} = 74.5\%$$

Macro F1 Measure

$$\text{Macro} - \text{F1 measure} = \frac{1}{\text{n}}\sum_{i=0}^{n} \text{F1i}$$

$$\text{Macro} - \text{F1 Measure} = 73.6\%$$

Summarizing the section mapping procedure, the extracted sections from the document were mapped over standard sections labels. In this experiment, 329 papers were randomly selected from the total of 1,200 documents. The system was evaluated based on well-known measure of precision and recall. Precision and recall values were computed for each standard section i.e. “Introduction”, “Related Work”, “Methodology”, “Results”, “Discussion” and “Conclusion”. Finally the macro-precision and macro-recall were calculated recorded 77.6% and 74.5% respectively. The overall F1 measure score received is 73%.It means that about 27% of error may occur in identification of relevant documents based on in-text citation patterns. The next step of methodology is to find the in-text citation frequencies.

3.3.4 Identification of Section-wise In-text citation Frequencies

After the identification of the sections in a document, The *Citation Tag Frequency Calculator* calculates in-text citations in the running text of various sections of that particular document. At the end, this module lists citation frequencies in the whole text and within different sections. The proposed algorithm has been shown in Figure 3.9. This algorithm computes the total in-text citation frequencies for each citation. Furthermore, in-text citations are also computed in each standardized sections (extracted in the previous step). Finally, the computed data is persisted in MySQL database for further analysis.

The proposed system computes the in-text frequencies for each citation. The computed data consists of (Current Document, Citation Tag, Total Frequency, Section, and Section Citation Tag Frequency). The snapshot of the computed data is shown in Table 3.6 and Table 3.7.

AlgorithmComputeCitationTagFrequency(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 using xPath and xQuery*
12. get '*Section Label' using xPath and xQuery*
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**

Figure 3.9: Algorithm for Computing In-text Citation Frequencies

Table 3.6: Paper Reference In-text Citation Frequencies Detail

RefID	Paper	Reference	In-text CF
1	1	[Brown, 89] J. S. Brown, A. Collins, P. Duguid, Situated Cognition and the Culture of Learning. In: Education Researcher 18, 1, 1989, 32-42	1
2	1	[Croft, 89] W. Croft, H. Turtle, A Retrieval Model Incorporating Hypertext Links, In: Hypertext'89, Proceedings, November 5-8 1989, Pittsburgh, USA, ACM, 213â€“224	3

During the analysis of in-text citation frequencies, it was found that in-text citation can be given differently based on its citation tag. There are a number of scientific writing templates for referencing such as IEEE, Harvard, Vancouver and APA etc. Each referencing template has its own way to cite other work. It was observed that there are certain complex scenarios in which identification of in-text citation becomes difficult. In the following section we will explain the reasons for incorrect identification of in-text citations.

Table 3.7: Section-wise In-text Citation Frequencies Detail

Fk_REF-ID	Section	Mapped-Section	In-text CF
1	1. Background	Introduction	1
2	1-Introduction	Introduction	1
2	2- Literature Review	Related Work	1
2	5-Knowledge Production	Methodology	1

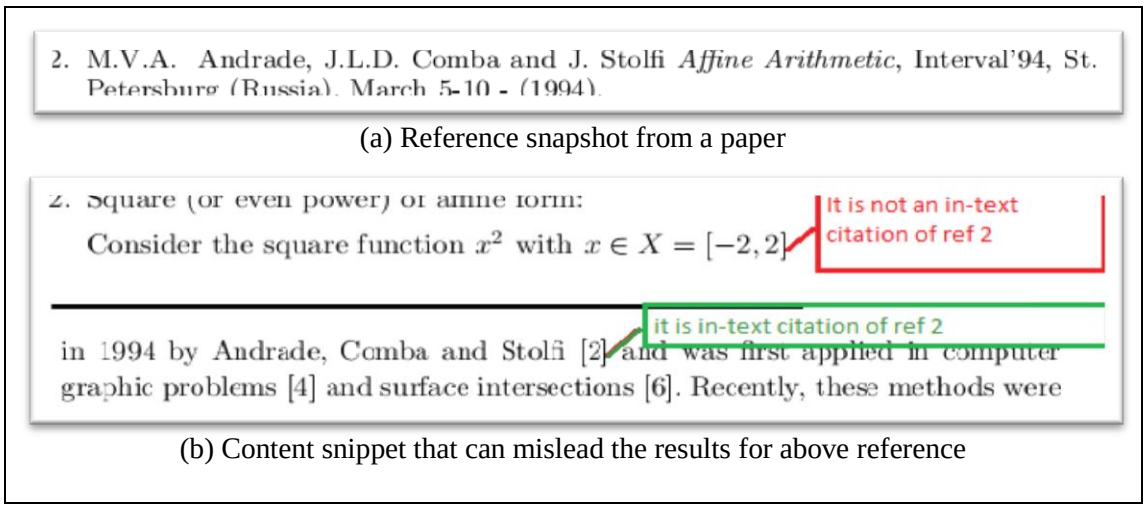


Figure 3.10: (a) Reference Snapshot from a Paper and (b) Content Snippet that can Mislead the Results

i. Real Scenarios from Scientific Documents

Based on manual inspection and analysis of the incorrect results, we are presenting interesting real scenarios from the documents where in-text citation has been identified incorrectly. The following scenarios demonstrate real issues where accurate identification of in-text citations is problematic. These scenarios highlight the ambiguity of identification of citation tags in a typical part of paper's content. Below is the detail of each scenario.

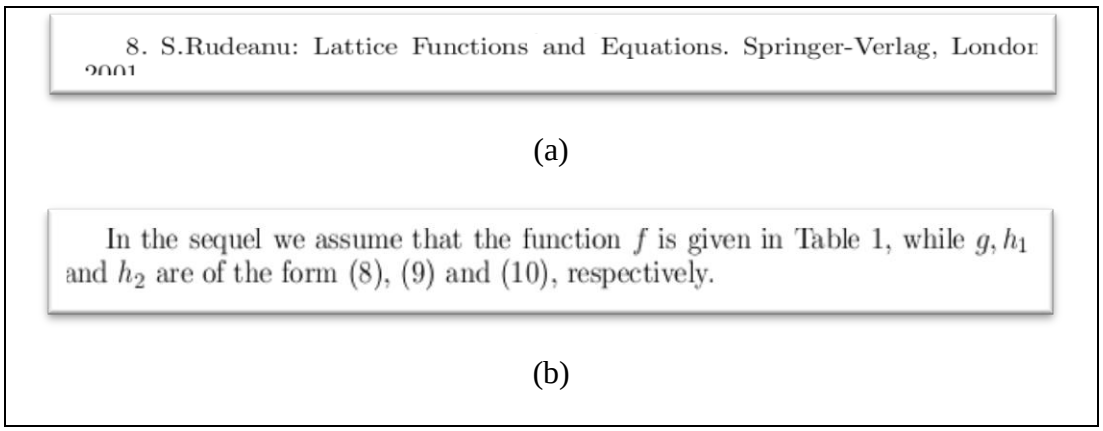


Figure 3.11: Reference Snapshot from a Paper and (b) Content Snippet that can Mislead the Results

a) Scenario 1 – Mathematical Ambiguity

A reference is shown in Figure 3.10 (a) extracted from reference sections of an article. In this case, the citation tag is "2". The citation in the running text of the document could be made using the following citation tags: "[2]", "[2,", ",2]", "[2", "2]". "[,2,]" or it can be hidden in the following citation tag "[1-5]" which is referring all references from 1 to 5. However, Figure 3.10 (b) presents another snippet from the same document where "[-2, 2]" is part of the paper text and does not belong to a citation tag. The tag "[-2,2]" is being used in a mathematical formula for donating an interval. The system may identify incorrectly due this interval values as in-text citation of reference "2". For tackling this type of problems, the automated tool needs to discover the context of the citation and needs to disambiguate between actual citation tag and content of the paper.

b) Scenario 2 – Mathematical Ambiguity

This scenario is an extension of the scenario number 1. A reference is shown in Figure 3.11 (a) from the reference section of an article where its citation tag is "8". In the body text of that article, "(8)" could be the one possible citation tag. However, Figure 3.11 (b) demonstrates a text from the same document where the "(8)" is being referred

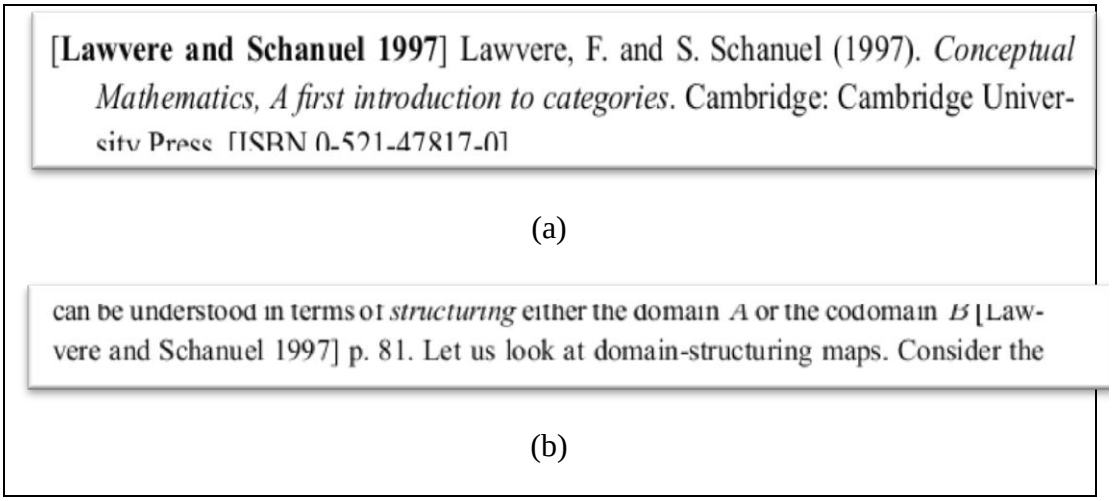


Figure 3.12: Reference snapshot from a paper and (b) Content snippet that can mislead the results

for some mathematical equation defined in that article. Thus, it will again become ambiguous for an automated tool to identify in-text citation accurately.

The equation number and intervals were found the two important misleading content for the accurate identification of in-text citation frequencies.

These kinds of problem may be addressed by disambiguating in-text citation and context of usage of such citation tag in article.

c) Scenario 3 – String Variations

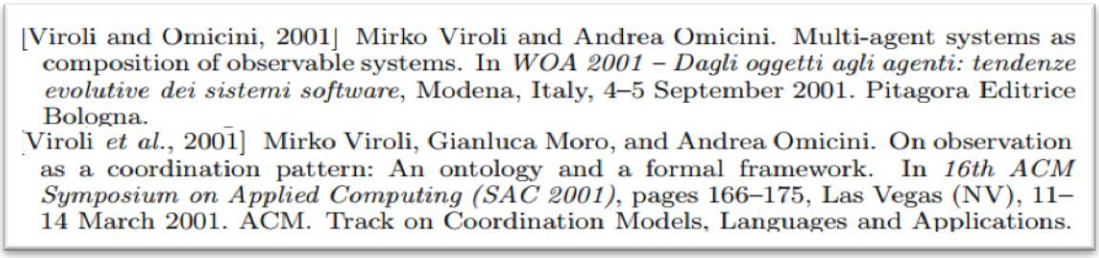
In this scenario, we have shown that hyphen can be used within the citation tag while referring to a particular reference in body text of the document. For example in Figure3.12 (a), the citation tag is "[Lawvere and Schanuel 1997]", however, Figure 3.12 (b) represents a snippet from the same document where the citation tag [Lawvere and Schanuel 1997] is used to refer to that reference. The inclusion of additional characters such as hypen (-) in the in-text citation was another reason.

These types of problems can be resolved using some string comparisons such as edit distance and Levenshtein distance etc.

d) Scenario 4 – Wrong Allotment

In J.UCS dataset we found that some articles have used authors and year information for citation tag. Multiple papers of an author with different team in the same year are referred as shown in Figure 3.13 (a). There are two separate tags for each citation i.e. [Viroli and Omicine, 2001] and [Viroli et al., 2001].

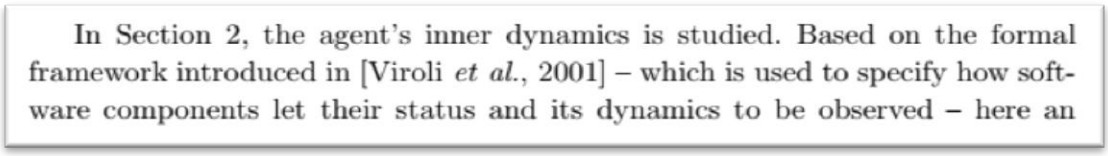
Automated solutions such as pdfx wrongly build a regular expression for citation tag based on only first author and year information.



[Viroli and Omicini, 2001] Mirko Viroli and Andrea Omicini. Multi-agent systems as composition of observable systems. In *WOA 2001 – Dagli oggetti agli agenti: tendenze evolutive dei sistemi software*, Modena, Italy, 4–5 September 2001. Pitagora Editrice Bologna.

Viroli *et al.*, 2001] Mirko Viroli, Gianluca Moro, and Andrea Omicini. On observation as a coordination pattern: An ontology and a formal framework. In *16th ACM Symposium on Applied Computing (SAC 2001)*, pages 166–175, Las Vegas (NV), 11–14 March 2001. ACM. Track on Coordination Models, Languages and Applications.

(a)



In Section 2, the agent's inner dynamics is studied. Based on the formal framework introduced in [Viroli *et al.*, 2001] – which is used to specify how software components let their status and its dynamics to be observed – here an

(b)

Figure 3.13: Reference Snapshot from a Paper and (b) Content Snippet that can Mislead the Results

Therefore, a regular expression, designed to calculate in-text citation of "Viroli, 2001" would mislead the results. Improper building of regular expression was one of the reasons that took part in the overall improper marking of in-text citation. To solve such problems, we should design a regular expression carefully such as in the above case, two separate regular expressions should be designed: [Viroli and Omicine, 2001] and [Viroli et al., 2001].

e) Scenario 5 - Commonality in Content

We found that some authors have used very common citation tags. For example, in the reference entry shown in the Figure 3.14 represents a citation-tag "[p]". Here, the contemporary systems will only use the character "P" as a reference tag, as shown in Figure 3.14. These kinds of citation tags are very sensitive as "P" is common character which may occur many times in the full text of the paper and will mislead the calculation of in-text citation frequencies. These types of problems may be handled by designing proper regular expressions. For example, in the above scenario, the extensive list of regular expression would be as follows:

"[P]", "[P,", ",P]", "[P", "P]". "[,P,]".

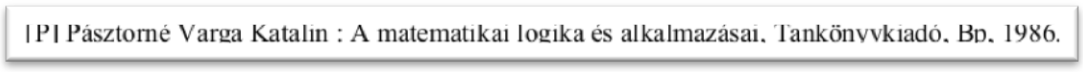


Figure 3.14: Reference Snapshot from a Paper and (b) Content snippet that can Mislead the Results

Finally, the in-text citations frequencies and their distribution across the sections were computed for all selected 1,200 documents were persisted in database for further analysis. During the system evaluation phase (will be discussed in next chapter), it was found that some interesting and complex scenarios may occur during automatic discovery of in-text citations. The identified solution in handling these complex scenarios will be incorporated when the experiments are performed on CiteSeer dataset.

3.3.5 Constructing Rules based on In-text Citation Frequencies and Patterns

The objective of this step (step 6 of overall methodology) is to construct rules based on in-text citation frequencies and in-text citation patterns to identify relationships between cited and cited-by papers. For the proof of concept, the proposed approach was evaluated on the dataset of an online Journal (Journal of Universal Computer Sciences). From the dataset of J.UCS which had 1,200 documents and 15,000 citations pairs, considering all of these pairs were impossible, therefore, for initial experiments, 100 pairs were selected from this dataset. These 100 pairs were selected systematically that they could represent the overall dataset. It was noted that the citation frequencies varies between 1 and 22 in our dataset for different pairs. The selected 100 pairs belonged to different in-text citation frequencies. For this purpose, different groups were made, one group represents all those pairs whose in-text citation frequencies varies between 1-5, the second group represents those pairs whose in-text frequencies varies between 6-10, the third group represents those pairs whose in-text citation frequencies varies between 11-15, whereas the fourth group represents those pairs whose in-text citation frequencies varies between 16-22. From each of these four groups, 25 papers were selected randomly. From these 100 selected papers, rules were constructed from 70 pairs and tested for 30 pairs. The constructed rules have been explained below. However, the testing and its results will be explained in the next chapter. These 70 pairs were given to three domain experts to manually read the citation context in the real papers and to identify nature of relationship.

Based on the analysis of in-text citation frequencies and in-text citation patterns, we were able to construct the following rules for identifying the nature of relationships between cited and cited-by papers.

Remember that following were two main categories of relationships between cited and cited-by papers "Methodological Relations" and "Non-methodological Relations".

The cited papers have the 'methodological relations' when cited-by documents:

1. have worked on the same problem as the cited work has done;
2. have extended/compared their work with cited work; and
3. have used some concepts, definitions of the cited work.

The cited papers have the 'non-methodological relations' when cited-by documents:

1. refer cited document only to give background study; and
2. use the cited document partially(cited work used to complete cited-by methodology).

Rule 1 and **Rule 2** is to determine the methodological and non-methodological relationship between cited and cited-by papers respectively.

Rule 1: The cited papers have the 'methodological relations' when:

```
Citation TypeIS "Paper" and
In-text citation BelongsTo two different Mapped Sections
and
One of the mapped section is other-than
{Introduction or related work} and
In-text citation occurrences in total number of sections
< 4

Or

Citation TypeIS "Paper" and
In-text citation >= 5 and
In-text citation BelongsTo more than two Mapped Sections
and
(
     One of the Mapped Section is {Result, Discussion,
     Conclusion} Or Self Citation Or
     In-text citation occurrences in total number of
     sections>= 4
)
```

Or

In-text citation > 1 and Citation TypeIS "Book"

Rule 2: The cited papers have the 'non-methodological relations' when:

In-text citation only BelongsTo {Introduction, Related Work}

or

In-text citation BelongsTo Sections other than {Introduction, Related Work} and

In-text citation occurrences in total number of sections = 1

Finally, these rules were evaluated on dataset of 30 papers. The system, based on these rules, identifies the methodological and non-methodological relationships with accuracy of more than 80% and the refined categories are found with average accuracy of 76%. The results have been discussed in next chapter.

Alongside this small scale study, we also evaluated these rules with the help of user studied dataset (Benchmark dataset). It was found that these rules improve overall results. These results are discussed in next chapter.

Chapter 4

Result Analysis

In the previous chapter detailed methodological steps for the evaluation of formulated hypotheses were described. Certain methodological steps were left as they were associated with results and evaluation of hypotheses. Below are the overall methodological steps; furthermore, ***bold and italic*** steps are discussed in detail in this chapter.

4.1 Methodology to Evaluate Hypotheses

To evaluate the formulized hypotheses, the following methodology was adopted. In previous chapter step1 to 4 and step 6 were discussed in detail (they are shown in gray color here). However, some of the steps such as step 5, 7 and 8 were associated with results and evaluation of hypotheses and hence will be presented in detail in this chapter. Those steps are shown below as bold and italic for easy comprehension.

Step 1: Comprehensive dataset selection (This step is required to evaluate all hypotheses).

Step 2: Pre-processing – Crawling CiteSeer database to download research articles, converting downloaded PDFs to XML format, and extracting references (**This step is required to evaluate all hypotheses**).

Step 3: Identification of sections of citing papers (**This step is required to evaluate hy**).

Step 4: Identification of In-text citation frequencies section wise (This step is required to evaluate all hypotheses).

Step 5: ***Evaluation of recommending most relevant papers based on in-text citation frequencies using user study (This step is only required to evaluate hypothesis 1).***

Step 6: Constructing rules based on in-text citation frequency patterns within different sections (This step is required to evaluate hypothesis 2).

Step 7:***In-text citation patterns rules (This step is required to evaluate hypothesis 2).***

Step 8:*Comparisons of proposed approach with state-of-the-art approaches (This step is only required to evaluate hypothesis 3).*

4.2 Evaluation of Recommending most Relevant Papers based on In-text Citation Frequencies using User Study

The developed crawler (as explained in the previous chapter) crawled and downloaded research articles found on CiteSeer for specific topics. Total 5,000 documents were downloaded; later on, those documents were converted to xml using pdfx. Thereupon, using Xpath and XQuery based solutions citations and their in-text citation frequencies were extracted. Total of 105,000 references were extracted from those papers.

In Figure 4.1, overall contribution of in-text citation frequencies in whole dataset is shown. It was found that major contributions is of in-text citation frequencies = 1 that is about 60% of overall data. The results follow an intuitive pattern that less number of portion is covered by higher values of in-text citation frequencies. For example, in-text citation frequencies=2 has 16% percent contribution in the dataset, in-text citation frequencies = 3 has covered 7% portion of the data and so on. Another interesting part of this result is that in-text citation frequencies=0 have a significant value. It shows that some of the references given in paper were not referred even single time in the body text of the citing paper. It validates the old results that some references are given in paper to pay un-due credit to some authors [Shahid et al., 2015]. These results clearly indicate that current quality measure should exercise the role of in-text citation frequencies in their overall calculation.

4.2.1 Gold Standard Dataset

Benchmark (Gold Standard) dataset is required for conducting experiments and evaluation of hypotheses. In the field of relevant research papers identification, gold standard dataset does not exist [Beel et al., 2013]. Therefore, the researchers developed their own benchmark dataset for evaluating result of their proposed approach. In spite of number of proposed approaches in this field, there is no such dataset freely available for performing experiments.

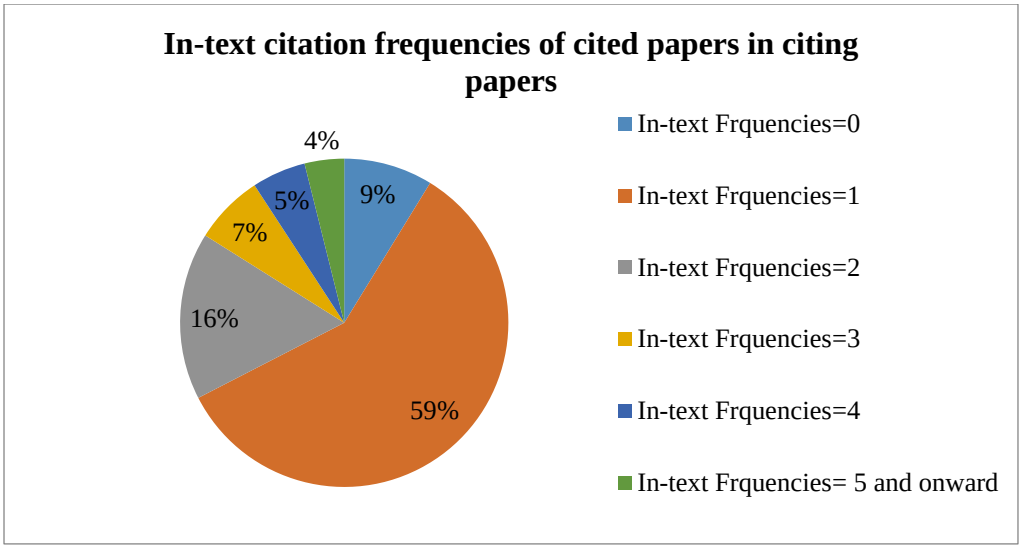


Figure 4.1: Contribution of Citations having Various In-text Citation Frequencies

Therefore, user study was conducted to develop benchmark dataset. Later on, the developed gold standard dataset was used to evaluate the hypotheses established in the previous chapter.

In our case, conducting user study on the whole dataset (as explained in section 4.2) was not feasible. Therefore, as a first step randomly citations pairs that have coverage from all groups mentioned in the Figure 4.1 were selected. Thus, the total records filtered were 12,000 citations pairs. From these 12,000 pairs, 400 citations pairs were randomly selected for user study. The overall in-text citation frequencies distribution among these 400 pairs is given in the Table 4.1. Citation pairs means citing and cited paper e.g. CitationPair (p100, p34) means that the first entry represents the citing paper and second entry represents the cited paper. It means that paper number "p34" is cited-by paper number "p100". There are two most widely known methods for evaluating research paper recommender systems which are online evaluation and offline evaluation [Jannach et al., 2013] [Rashid et al., 2002][Knijnenburg et al., 2012].

Table 4.1: In-text Citation Frequencies Representation in Selected Sample Dataset

In-text frequencies	Total Instances
1	94
2	64
3	74
4	56
>= 5	112

In offline evaluation, promising approach is identified at first step and then that approach is evaluated with detailed user studies. However, it has been criticized for not finding effective approach. Furthermore, it has been observed that offline evaluation results do not necessarily correlate with results from user studies [Beel et al., 2013][Cremonesi et al., 2012][McNee et al., 2006] [Hersh et al., 2000]. McNee et al has criticized offline evaluation in the following words:

"*the research community's dependence on offline experiments [has] created a disconnect between algorithms that score well on accuracy metrics and algorithms that users will find useful*"

Similarly, Jannach et al [Jannach et al., 2013] have stated about offline evaluation that:

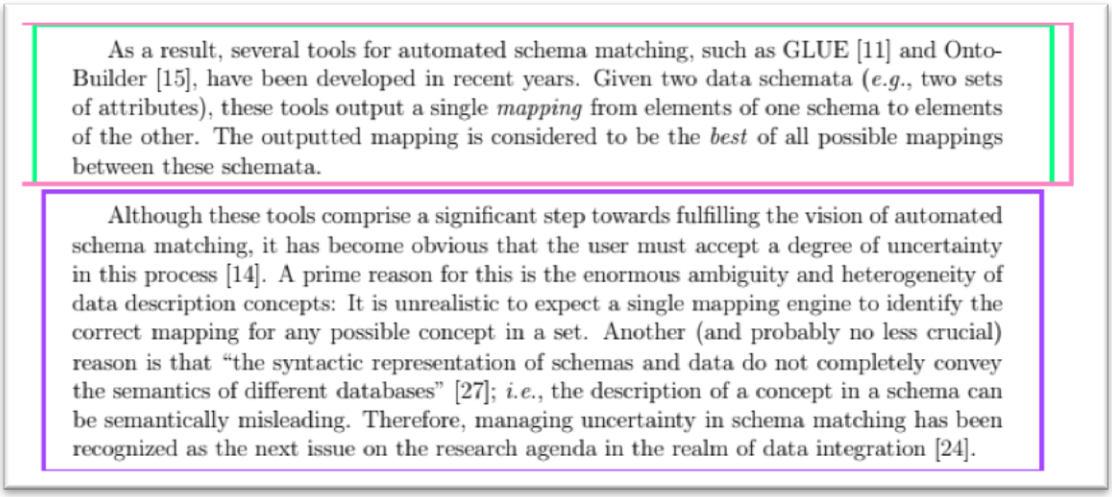
"*the results of offline [evaluations] may remain inconclusive or even misleading" and "real-world evaluations and, to some extent, lab studies represent probably the best methods to evaluate systems*"

Thus, researchers in this field have recommended that identification of relevant document task should be evaluated using online evaluation instead of offline evaluations [Jannach et al., 2013] [Rashid et al., 2002][Knijnenburg et al., 2012].

Therefore, contrary to offline evaluation, first a detailed user study was conducted to evaluate the proposed approach. However, which paper is most relevant and which is not? Such type of questions can only be asked from those who are experienced, or at least actively involved in research. The point is that one cannot invite everyone to conduct user study when working in this field. Thus for user study, those MS and PhD students from two different university were invited who were actively involved in their research work. Total participants in this study were 80.

As already discussed in Chapter 2, authors cite other papers for some reasons. Therefore, authors' sentiment for cited paper can always be found around their in-text citations in the paper. That is called citation context. Different researchers have exploited citation context for discovering sentiment of the authors for cited paper [Teufel, 2006][Kaplan et al., 2009]. Therefore, selected 400 citation pairs context were marked for user for their ease and quick decision about relationship between citing and cited papers. In Figure 4.2(a), it is shown that how citation context were

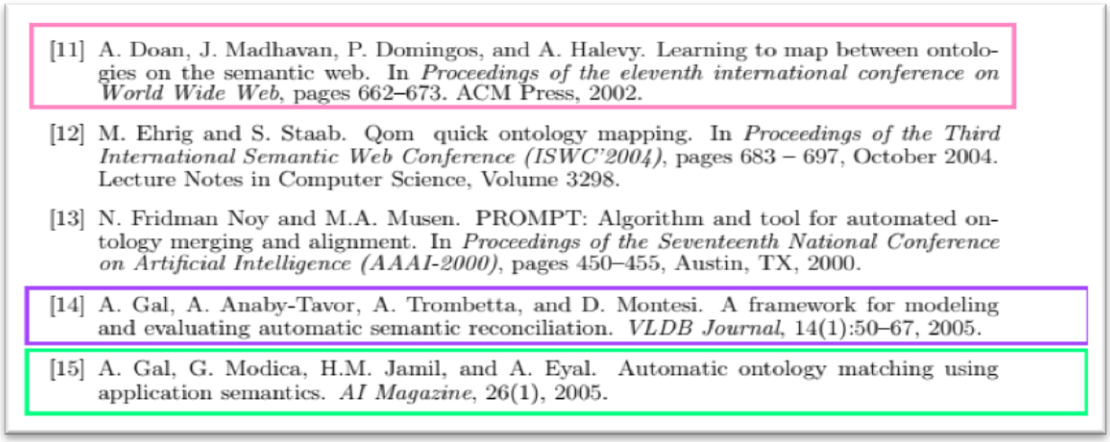
marked for the users. In Figure 4.2(a), in-text citations marked in a paper titled "*Managing Uncertainty in Schema Matching with Top-K Schema Mappings*" are shown, while in Figure 4.2(b), their references are shown. In the similar fashion, wherever in-text citations were found, they were marked for deeper analysis of the users.



As a result, several tools for automated schema matching, such as GLUE [11] and OntoBuilder [15], have been developed in recent years. Given two data schemata (*e.g.*, two sets of attributes), these tools output a single *mapping* from elements of one schema to elements of the other. The outputted mapping is considered to be the *best* of all possible mappings between these schemata.

Although these tools comprise a significant step towards fulfilling the vision of automated schema matching, it has become obvious that the user must accept a degree of uncertainty in this process [14]. A prime reason for this is the enormous ambiguity and heterogeneity of data description concepts: It is unrealistic to expect a single mapping engine to identify the correct mapping for any possible concept in a set. Another (and probably no less crucial) reason is that "the syntactic representation of schemas and data do not completely convey the semantics of different databases" [27]; *i.e.*, the description of a concept in a schema can be semantically misleading. Therefore, managing uncertainty in schema matching has been recognized as the next issue on the research agenda in the realm of data integration [24].

(a) In-text Citation Marking in Body-text of the Paper



[11] A. Doan, J. Madhavan, P. Domingos, and A. Halevy. Learning to map between ontologies on the semantic web. In *Proceedings of the eleventh international conference on World Wide Web*, pages 662–673. ACM Press, 2002.

[12] M. Ehrig and S. Staab. Qom quick ontology mapping. In *Proceedings of the Third International Semantic Web Conference (ISWC'2004)*, pages 683 – 697, October 2004. Lecture Notes in Computer Science, Volume 3298.

[13] N. Fridman Noy and M.A. Musen. PROMPT: Algorithm and tool for automated ontology merging and alignment. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence (AAAI-2000)*, pages 450–455, Austin, TX, 2000.

[14] A. Gal, A. Anaby-Tavor, A. Trombetta, and D. Montesi. A framework for modeling and evaluating automatic semantic reconciliation. *VLDB Journal*, 14(1):50–67, 2005.

[15] A. Gal, G. Modica, H.M. Jamil, and A. Eyal. Automatic ontology matching using application semantics. *AI Magazine*, 26(1), 2005.

(b): References Marking in Citing Article

Figure 4.2: Research Articles Body-text and References Annotation for Participants of the Study

Table 4.2: Total Citations Context in Citing Papers for Selected Dataset

In-text frequencies	Total Instances	Citation Context
1	94	94
2	64	124
3	74	229
4	56	224
5 or greater than 5	112	560
Total Citation context marked for analysis		1,231

Thus we exploited more than 1230 citation contexts in our user study. However, the full text of the papers was also available for users to further explore the context of cited papers. The overall details are shown in Table 4.2.

For annotation purposes, users were given three things: first was selected paper where references were marked along with their in-text citations in the citing paper (as shown in Figure 4.2(a),; second was "Citation reasons form" as shown in Table 4.3; third was the list of selected citations list from the paper.

Users were asked to fill the citation reason code after analysis of citation reason for a particular citation. The selected citation table for paper "*Managing Uncertainty in Schema Matching with Top-K Schema Mappings*" is shown in Table 4.4. To get multiple judgments on same citations pair, it was made sure to assign a citation pair to two different users.

Table 4.3: Citation Reasons Form

Citation Reason	Code
Citing paper has worked on the same problem as the cited work has done.	SPRB
Citing paper has extended/compared their work with cited work	ECW
Citing paper has used some concepts, definitions of the cited work	UCD
Citing paper has used the cited document partially(cited work is used to complete citing paper methodology)	UP
Citing paper has referred cited document only as background study, or highlighting the importance of the research	UBI

4.2.2 Citation Reasons Mapping

For the last fifty years, researchers have been trying to find an answer to the question that why one author has referred to the previous research? Garfield, the pioneer in the citation analysis, has earlier described 15 reasons for citations to answer this question [Garfield, 1979]. It is important to find author's motivation about any citation, because if done successfully, this could revolutionize the whole information sciences

Table 4.4: Selected Sample Citations for User Study

Citing paper Citations	Your Code
[14] A. Gal, A. Anaby-Tavor, A. Trombetta, and D. Montesi…	
[15] A. Gal, G. Modica, H.M. Jamil, and A……	
[26] S. Melnik, E. Rahm, and P.A. Bernstein…..	
[1] J. Aitchison, A. Gilchrist, and D. Bawden…..	
[7] C.R. Chegireddy and H.W. Hamacher. Algorithms…..	
[11] A. Doan, J. Madhavan, P. Domingos, and….	

practices. However, identification of such nature of relationship between cited and cited-by document requires extensive analysis of content.

Table 4.5: Type of Relevancy Grouping

Citation Reason	Code	Mapping on TR
Citing paper has worked on the same problem as the cited work has done.	SPRB	**Methodologically Related**
Citing has extended/compared their work with cited work	ECW	
Citing paper has used some concepts, definitions of the cited work	UCD	
Citing paper has used the cited document partially(cited work is used to complete citing paper methodology)	UP	**Non-Methodologically Related**
Citing paper has referred cited document only as background study, or highlighting importance of the research	UBI	

In the light of previous research, we grouped various citation reasons. One aspect of this grouping deals with the types of relationship between citing and cited papers and the other aspect deals with the strength of the relationship between citing and cited

papers. The same two aspects were later used for the evaluation of the proposed system. The user annotated data was categorized using these two types of classifications. The detail of each classification is given below.

A) Type of Relevancy

Broadly speaking, the relationship between citing and cited paper can be classified into two major categories: methodological relation and non-methodological relation. We mapped citation reasons over these types of relationships. It has been summarized in Table 4.5. Later on, these definitions were used for the evaluation and conducting user studies.

B) Degree of Relevancy

The focus of this evaluation was to discover the strength of relationship between citing and cited papers. Instead of asking the users to give their opinion about citations, citation reasons were categorized. We used three stars scale method such as Strong, Medium, and Weak citations.

Categorization (Strong, Medium, and Weak) of citations are shown in Table 4.6.

Table 4.6: Strength of Relationship Grouping

Citation Reason	Code	Mapping on SR
Citing paper has worked on the same problem as the cited work has done.	SPRB	**Strong**
Citing has extended/compared their work with cited work	ECW	
Citing paper has used some concepts, definitions of the cited work	UCD	**Medium**
Citing paper has used the cited document partially(cited work is used to complete citing paper methodology)	UP	
Citing paper has referred cited document only as background study, or highlighting the importance of the research	UBI	Weak

Based on the above defined criteria, the aforementioned 400 citations were classified. It means that now every annotated citation had a specific category i.e. either methodologically related or non-methodologically related in type of relevancy-based classification. Similarly, each citation was classified into either "Strong", "Medium", or "Weak" relationship with citing paper in degree of relationship-based classification. In the total of 400 citation pairs, there was difference of opinion among annotator on 82 citation pairs in degree of relationship classification, whereas difference of opinion on 49 pairs was recorded in type of relevancy based classified instances. These disputed citation pairs were not considered in the results of final experiments.

In the below sections, this annotated data has been used for evaluating results that support the established hypothesis in the previous chapter.

4.2.3 Experimental Results

This section highlights important results obtained from the proposed technique through the users' judgment. First the correlation of in-text citation frequencies with degree of relevancy has been explained. Secondly, type of relevancy and in-text citations correlation has been described.

The analysis of the following results helped in validating the developed hypotheses. The first hypothesis was associated with degree for relevancy between citing and cited paper:

Hypothesis 1: In-text citation frequencies have potential to identify the relationship strength (Strong, Medium, Weak) between scientific papers.

The second hypothesis was related to the type of relevancy between citing and cited papers.

Hypothesis 2: For a given cited paper, In-text citation frequencies and In-text citation patterns can categorize cited-by papers into two categories such as: 'methodologically related' and 'non-methodologically related' papers.

The third hypothesis was about comparison between different state-of-the-art approaches and in-text citation frequencies based approach. It is given below

Hypothesis 3: In-text citation frequencies based approach analyzes cited paper evidences deeply in citing paper text; thus, it has more potential to produce quality results in comparison to other state-of-the-art approaches.

In the sections below, different experiments results have been presented. The results have been discussed in details.

A) In-text Citation Frequencies Correlation with Degree of Relevancy

In this section the correlation between degree-of-relevancy/strength-of-relationship with in-text citation frequencies was explored. The results are shown in the Table 4.5. As explained in the previous sections, the users were asked to provide annotation for selected citations of their assigned papers. The annotated citations were then categorized based on citation reasons mapping procedure explained in the previous sections. In the computed results in Table 4.7, the first column represents the in-text citation frequencies, the next columns with labels “Strong”, “Medium” and “Weak” represent the number of agreed upon instances that were classified in respective classes. The last column represents the number of cases where inter annotators agreement was not identical. Those disputed cases were not considered in the result.

These results indicate that lower in-text citation frequencies normally represent weaker relationship between citing and cited paper. For example, in total 85 citation instances having in-text citation frequencies = 1, weak relationship was found for 64 instances. Similarly, this pattern holds for in-text citation frequencies= 2. For in-text citation frequencies = 3 and 4, moderate relationship between citing and cited paper was recorded most of the times. However, a good number of Strong relationships was also recorded. Lastly, for in-text citation frequencies = 5 and greater, Strong relationships were found most of the time. However, in that case some medium relationship was also recorded, i.e. 16%.

The consolidated results are also shown for better comprehension in the Figure 4.3 with the help of line graph. The results can be summarized: the lower in-text citation frequencies represent weaker relationship; on the other hand, in-text citation frequencies- 3 and 4- represent medium relationship between citing and cited

documents; finally, for in-text frequencies>= 5, there exists a strong relationship between cited and citing papers.

Table 4.7: Mapping of In-text Citation Frequencies over Degree of Relevancy

Frequencies	Strong	Medium	Weak	Un-decided
In-text citation frequencies = 1	2	9	74	9
In-text citation frequencies = 2	7	13	31	13
In-text citation frequencies = 3	14	29	17	14
In-text citation frequencies = 4	13	21	4	18
In-text citation frequencies = 5 and onward	65	14	5	28

In Figure 4.3, normalized values (i.e. 0 to 1) percentages of number of instances are shown on Y-axis, whereas on X-axis in-text citations frequencies ranges are shown.

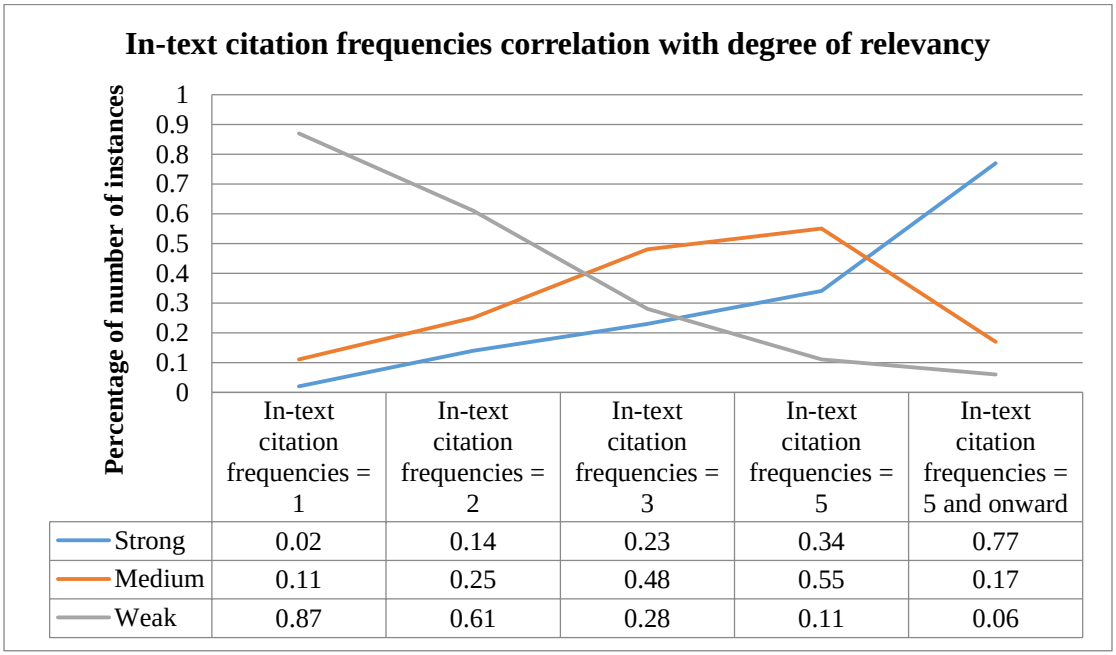


Figure 4.3: In-text Citation Frequencies Mapping over Degree of Relevancy

In the next experiment, the correlation of in-text citation frequencies with type of relevancy has also been exploited. This experiment, was also performed on the same gold standard dataset.

B) In-text Citation Frequencies Correlation with Type of Relevancy

In this experiment, the correlation between in-text citation frequencies and type of relevancy between citing and cited papers was analysed. It was found that in-text citation frequencies have the capability to determine type of relevancy between cited and citing papers. The overall results are shown in the Table 4.8. The first column of

Table 4.8 represents in-text citation frequencies. The second and third columns represent type of relevancy i.e. (M= "Methodologically related" and NM = "Non-Methodologically related"). The values in these columns are the number of instances where inter-annotator agreement was found on mapping them in the same category. The last column "Un-Decided" represents the number of instances where the annotators did not agree on their decision. For example, one annotator placed a citation in "M" category while the other placed the same citation in "NM" class.

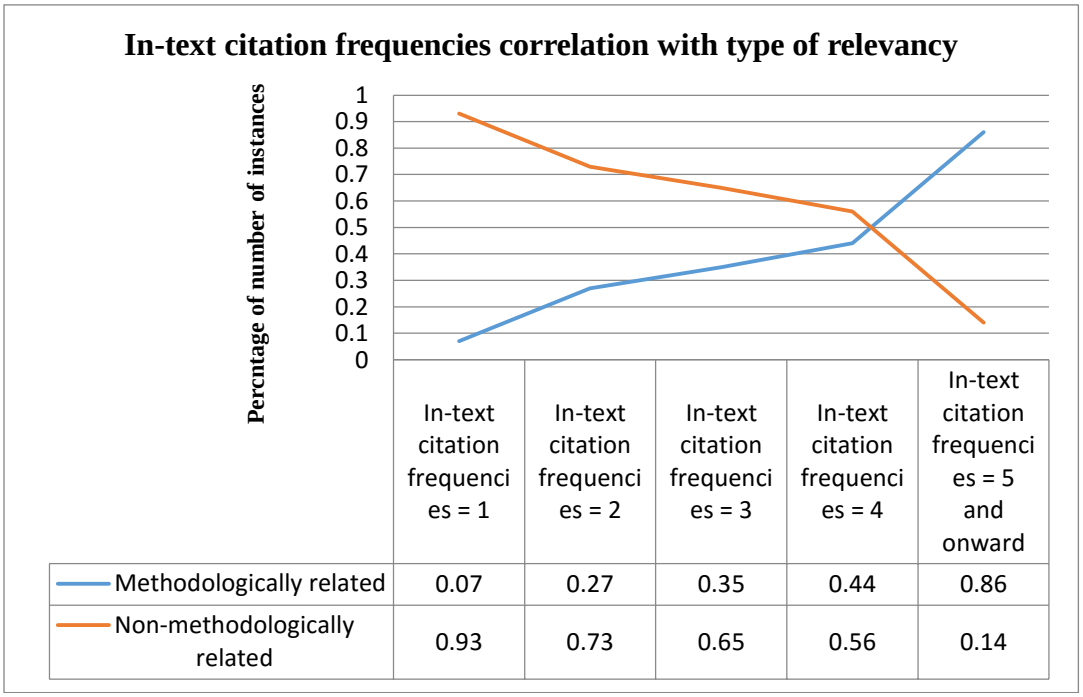
The results show that in-text citation frequencies = 1 means that there is non-methodological relationship between cited and citing paper. Furthermore, there is a strong correlation between annotators on their decision. The similar trend was recorded for those citations whose in-text citation frequencies were equal to two whereas for citations having in-text citation frequencies 3 or 4, the inter-annotator data has too many un-decided values. However, it was noticed that for in-text citation frequency equal to 3 and 4, majority of the time the relationship was non-methodological on the cases where annotators agreed. Furthermore, it was found that citations having higher in-text citation frequencies (in-text citation frequencies >=5) were majority of the time methodologically related.

Table 4.8: Mapping of In-text Citation Frequencies over Type of Relevancy

Frequencies	M	NM	Un-Decided
In-text citation frequencies = 1	6	80	8
In-text citation frequencies = 2	15	41	8
In-text citation frequencies = 3	20	37	17
In-text citation frequencies = 4	22	28	6
In-text citation frequencies = 5 and onward	88	14	10

The in-text citation frequencies correlations with type of relevancy are shown in the Figure 4.4 with line graph. The Methodologically related citations are shown with blue line and Non-Methodologically related citations are represented with brown line. The results indicate that the certainty levels of decision about type of relevancy are related with in-text citation frequencies. Furthermore, the lower the in-text citation frequencies (i.e. 1, 2), the higher the probability of its being "non-methodologically related". Similarly for higher in-text citation frequencies e.g. >=5, the probability being "methodologically related" increases.

The line graph shown in Figure 4.4 reveals an interesting pattern that lower in-text citation frequencies correspond to highly non-methodologically related papers and low number of methodologically related papers. On the other hand, higher in-text citation frequencies map over low non-methodologically related papers and high number of methodologically related papers.



	In-text citation frequencies = 1	In-text citation frequencies = 2	In-text citation frequencies = 3	In-text citation frequencies = 4	In-text citation frequencies = 5 and onward
Methodologically related	0.07	0.27	0.35	0.44	0.86
Non-methodologically related	0.93	0.73	0.65	0.56	0.14

Figure 4.4: In-text Citation Frequencies Mapping over Type of Relevancy

These results indicate that in-text citation frequencies play significant role in identification of type of relationship and strength of relations between citing and cited papers. However, there are certain cases where even higher in-text citation frequencies correspond to non-methodological relationships. Such cases were investigated and it was found that in-text citation frequencies also depend on authors' citation styles. For example, consider the snapshots shown in the Figure 4.5 and Figure 4.6. In the Figure 4.5 in-text citations for reference "19" have been mentioned twice in single line. Similarly, in Figure 4.6, the reference no. 22 from the paper has been referred thrice in just 7 line of paragraph.

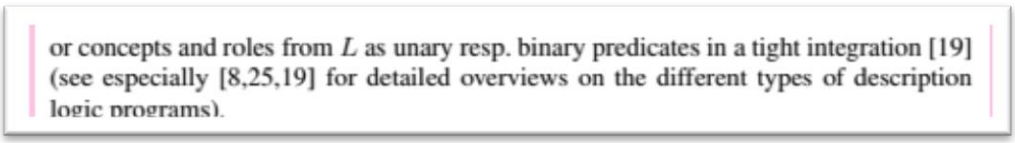

or concepts and roles from L as unary resp. binary predicates in a tight integration [19] (see especially [8,25,19] for detailed overviews on the different types of description logic programs).

Figure 4.5 In-text Citation Occurrences in Single Line

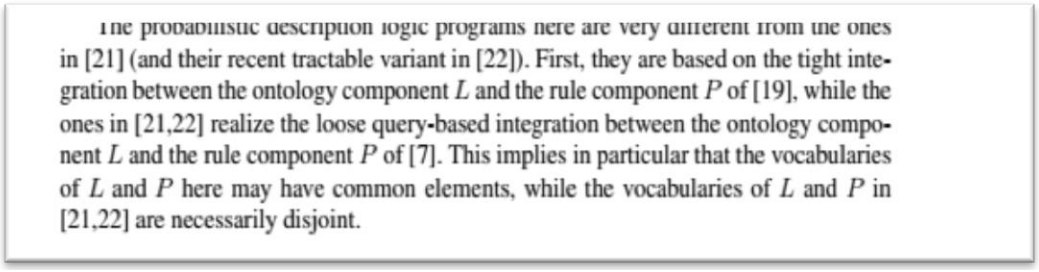

The probabilistic description logic programs here are very different from the ones in [21] (and their recent tractable variant in [22]). First, they are based on the tight integration between the ontology component L and the rule component P of [19], while the ones in [21,22] realize the loose query-based integration between the ontology component L and the rule component P of [7]. This implies in particular that the vocabularies of L and P here may have common elements, while the vocabularies of L and P in [21,22] are necessarily disjoint.

Figure 4.6 Multiple Occurrences of In-text Citation in a Small Paragraph

Summarizing the above experiment, we explored the role of in-text citation frequencies for identification of relevant documents. It was found that higher values of in-text citation frequencies map over methodological relationship between citing and cited papers. Moreover, the lower in-text citation frequencies represent Non-Methodological relationship between citing and cited papers.

The above experiments show that in-text citation correlates with degree of relevancy and type of relevancy. Reasonably high accuracy was determined for lower range (in-text citation frequencies = 1) and high range in-text citation frequencies (in-text citation frequencies >= 5). For example, in case of degree of relevancy, 87% of the time the citations pairs were accurately classified as "Weak" citations for in-text citation frequencies = 1. Similarly, 77% of the time citations were accurately classified as "Strong" citations for in-text citation frequencies >= 5. These results indicated that lower citations having in-text citation frequencies are normally corresponds to "Weak" citations and citations having higher in-text citation frequencies correspond to "Strong" citations.

In the second type of experiment where in-text citation frequencies correlation with type of relevancy was explored, it was found that the important category "methodologically related" citations are usually mapped over high in-text citation frequencies. For example, 86% of the time, "methodologically related" citations were classified for in-text citation frequencies >=5. It means that 86% of citation pairs in total of about 4% of citation pairs having in-text citation frequencies = 5 in Figure 4.1 correspond to "methodologically related" papers. On the other hand, there are high chances for a citation being classified as "non-methodologically related" for in-text citation frequencies = 1. For example, as you can see in Figure 4.4, 93% of citations were classified as "non-methodologically related" for in-text citation frequencies = 1. It means that 93% of the citations pairs in total of about 59% of citations having in-

text citation frequencies = 1 in Figure 4.1 corresponds to "non-methodologically related" papers.

These results clearly demonstrate that in-text citation frequencies help in discovering most relevant citations. On the basis of these findings, **the proof of hypothesis 1was supported with enough and comprehensive experiments**.

However, there are cases where in-text citation frequencies alone can mislead the results as demonstrated in Figure 4.5 and 4.6. Therefore, there is need to consider some more features such as in-text citation patterns.

4.3 In-text Citation Patterns Rules(Step 7 of the Methodology)

In a research article, different sections could be, for instance, "Introduction", "Related work", "Methodology", "Results", "Discussion", and "Conclusion". There is scientific agreement of considering them as sections of the scientific documents such as: Kansas State University-Research Paper template [Kansas-template, 2013], Rice University Research-Paper template [Rice-template, 2013] and Boston College University libraries-Research Paper template [Boston-template, 2013]. The in-text citation occurrences in different sections of papers are referred as patterns of citations. It is important to exploit section information as different researchers have also suggested that in-text citations in different sections may provide a clue about special type of relationship with cited paper [Tuefel,2006].

4.3.1 Testing the Pattern's Rules

The rules were constructed based on the in-text citations occurrences in the logical sections (i.e. "Introduction", "Related work", "Methodology", "Results", "Discussion", and "Conclusion") of a paper. The rule construction mechanism was explained in the previous chapter in detail. In this chapter, our focus is on testing and evaluation of those rules.

For testing and evaluation of rules, 54 citation pairs were selected (24 pairs were added later in initial data set of 70% training pairs and 30% testing pairs as explained in previous chapter). The pairs were selected covering different in-text citation frequencies ranges. For this purpose, different groups were made, one group represents all those pairs whose in-text citation frequencies varies between 1-5, the

second group represents those pairs whose in-text frequencies varies between 6-9, the third group represents those pairs whose in-text citation frequencies varies between 10-22. These selected 54 pairs were given to three domain experts (18 pairs to each expert) to manually read the citation context in the real papers and to identify nature of relationship between cited and cited-by paper. The domain experts were PhD students who had been actively involved in their research for last more than one year in the same domain. The same sets of selected pairs (i.e. 54 pairs) were automatically processed to mark the citation reasons with the help of constructed rules (for nature of relationship identification). Subsequently, the identified citation reasons by domain experts were compared with citation reasons marked with the help of rules.

The results for methodologically related papers and non-methodologically related papers are shown in the Table 4.9. The first column of the table represents the citation reasons. The second column represents the rule number used to identify that particular citation reason. The rule numbers along-with the rules were explained in the section 3.2.5 in chapter 3. The third column "total number of papers" shows the total number of pairs (cited and cited-by) having that particular reason as per the domain expert assessment, and the second last column shows the total number of pairs in which the citation reason was correctly marked, and the last column displays the percentage of accuracy.

Table 4.9: Methodologically and Non-Methodologically Related Pairs Accuracy

Citation Reason	Rules	Total Number of papers	Accurately marked	Accuracy Percentage
Methodologically related pairs	Rule1	42	33	79 %
Non-methodologically related pairs	Rule2	13	11	84%

The overall results are quite encouraging and these initial results support our hypothesis 2.

However, the selected dataset was too small to conclude something concrete. Therefore, it was required to validate the constructed rules on large dataset. Those

rules were evaluated on CiteSeer dataset and the improvements in the results (results of experiments conducted in section 4.2.2) were recorded. These improvements have been discussed in below section.

4.3.2 Improvements in Results with the help of In-text Citation Patterns

Careful analysis of the results discussed in previous section (Section 4.2.2) indicates that in-text citation frequency feature alone is not sufficient because for in-text citations frequencies =3, 35% methodological and 65% non-methodological relationship between citing and cited papers were identified. Similarly, for in-text citation frequencies = 4, the values for the two type of relationship are more close. These results indicated that in-text citation frequencies alone are not sufficient. To cope with this deficiency, in-text citation frequencies were analyzed across all sections of the papers. Research papers were divided into six different logical sections such as: “Introduction”, “Related Work”, “Proposed Work”, “Results”, “Discussion”, and “Conclusion” as explained in detail in chapter 3. The citations given in these sections are normally of similar nature and thus could help in providing more insights into the problem.

The type of relevancy between citing and cited papers was explored again with the help of manually crafted rules (as explained in the previous chapter). Overall improvement was recorded in achieved results. The overall results are shown in Table 4.10. The first column of Table 4.10 represents in-text citation frequencies; second column represents the total number of un-disputed instances. The third column “PM” represents the total number of previous methodological relationships found when patterns were not considered. The fourth column “AM” represents the total number of methodological relationships found after considering patterns. Similarly, the second last and the last column represent the previous and after status of non-methodological relationships found.

Analyzing the results, It was found that overall quality of the results has been improved.

Table 4.10: Mapping of In-text Citation Frequencies over Type of Relevancy

In-text citation Frequencies	Total	Pattern	PM	AM	PNM	ANM	P-I
1	86	Yes	6	6	80	80	0%
2	56	Yes	15	18	41	38	5%
3	57	Yes	20	27	37	30	12%
4	50	Yes	22	38	28	12	32%
= 5 and onward	102	Yes	88	94	14	8	5%

For example, for in-text citation frequencies= 3, the 7 pairs that were marked as "Non-Methodologically Related" pairs were identified as "Methodologically Related" pairs. It means that 7/57 results were improved by 12%.

Similarly for in-text citation frequencies = 4, 16 citation pairs were properly classified, thus results improved 32%. The overall Percent Improved (P-I) is shown in the last column of the Table.4.10.

The overall results indicate that in-text citation frequencies are inversely proportional to non-methodological relationship between citing and cited papers and directly proportional to methodological relationship. It means that the lower in-text citation, the higher its chances of being non-methodological relationship and vice versa. Similarly, the lower the in-text citation frequencies, the lesser the chances are for two papers to be in methodological relationship and vice versa.

By incorporating the patterns, the overall results shown in Table 4.10 have been improved. These results indicate that patterns help in refining the results and thus I argue that **the achieved results are in support of hypothesis 2.**

4.4 Comparisons of proposed approach with state-of-the-art techniques (Step 8 of the Methodology)

The methodology of in-text citation frequencies calculation was explained in the previous chapter. In this section our focus is on the evaluation and testing of in-text citation frequencies based on most relevant papers recommendations and its comparison with state-of-the-art approaches.

The state-of-the-art approaches have been discussed in detail in chapter 2. The approaches for recommending relevant papers can be mainly classified in four major categories such as content, metadata, citations, and collaboration based approaches.

In this research, In-text citation frequencies based approach have been compared with content, metadata, and citations based (bibliographic coupling) approaches. The collaboration based technique was not feasible for comparison as it can only be evaluated on full fledge live system where one can discover taste information of different users in that system.

For the evaluation purposes, the same gold standard dataset was used as explained in section 4.2.1. It is important to note that we already computed in-text citation frequencies for all of the citation pairs. The rest of the techniques were applied on the gold standard dataset and thus result was acquired and compared with proposed approach.

Table 4.11: Sample Paper and their Selected References

Selected References	**Paper: Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network**			
	In-text citation frequencies	**Content similarity**	**Citations (Bibliographic coupling) Based**	**Metadata (Title Terms) Based**
R1	10	0.27	2	1
R2	9	0.30	0	1
…	…	…	…	…
…	…	…	…	…
Rn (in this case n= 7)	2	0.20	0	1

Different state-of-the-art techniques were applied on standard dataset; the summarized table of a sample paper is shown in Table 4.11. In this table, the left most column represents the selected references for the paper “Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network”. In rest of the columns, computed values for different techniques have been shown. For example, in the second column, the in-text citation frequencies are shown. Their values represent that “R1” reference has been referred 10 times in body text of the focused paper. Similarly, the next column represents the content similarity values between selected reference and the titled paper. For example, 0.27 of the abstract terms of the selected reference paper and titled paper were matched. The fourth column represents the bibliographic coupling unit between selected reference paper and titled paper. Finally, in the last column, the title terms matched values between selected references and titled papers were shown.

As already explained in formulation of gold standard dataset, references were classified as strong, moderate and weak citations; therefore, when providing these references to end users, they were ranked e.g. strong citations are shown on top; moderate citations are shown next, and weak citations are displayed at the end of list. In Information retrieval, Discounted Cumulative Gain (DCG) is mainly used to measure the usefulness, or gain, of a proposed system having such graded relevance [Jarvelin and Kekalainen, 2000]. For DCG, each result in the query result list must have a graded relevance score, which I already have in my gold standard dataset. The results with higher relevancy are more useful when appearing on the top positions of a result list. DCG penalizes the highly relevant documents appearing at the bottom positions of a result set. The formula of DCG is given in Equation 4.1.

$$DCG_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{\log_2(i+1)} \qquad \textit{(Eq 4.1)}$$

Where rel_i is the relevance of the i the result in the result list of a query, i is the order of the document in the result list, and p is the order of the last document included in the DCG calculation.

For different research papers, the selected references vary in length. Therefore, the DCG alone was not capable for evaluating the system performance. For this reason, the DCG was normalized for different papers. The normalized cumulative gain, nDCG is formulated as shown in Equation 4.2.

$$nDCG_p = \frac{DCG_p}{IDCG_p} \textit{ (Eq 4.2)}$$

Where $IDCG_p$ is the ideal
that is computed by sorting the documents of a result list by relevance score with the help of user's annotation (i.e. gold standard preparation).

Finally, all the normalized values were averaged to get a single value for each technique for different size of disjoint set of queries. In the sections below, details of these techniques have been discussed.

4.4.1 Proposed Approach based Recommendations

It has already been explained as to how in-text citation frequencies were gathered and computed. For this experiment, selected references were ranked based on in-text citation frequencies. Thus, citations having higher in-text citation frequencies were ranked at top, and references with lower in-text citation frequencies were placed at bottom of the list. As you have already observed that in-text citation frequencies are discrete values; therefore, ranking based on in-text citation frequencies was not an issue.

One important thing which should be noted that is the a situation of tie which may occur while mapping two citation pairs of same in-text citation frequency over different classes i.e. strong, medium and weak citations. In total 400 citation pairs, such situation were carefully observed and total of 44 instances were found where in-text citation between cited and citing papers were same. Furthermore, out of these 44 citation pairs, 35 instances were classified by the users in same category whereas disagreement was recorded on 7 instances between the users. So, only two instances were found in which they were classified in different categories, so we chose to rank the lowest category on top so that we don't give favor to the proposed technique.

Once this ranking was computed for each query document, the ranking was normalized with the help of Ideal Discount Cumulative Gain (IDCG). Finally, all of the results were averaged to get a single value. The overall results are shown in Figure 4.11. The results indicate that an nDCG value of 0.89 was received, which is very good with respect to other techniques such as content similarity, citations based and metadata based techniques. Different disjoint set of queries such as nDCG @5, nDCG @10, nDCG @15, nDCG @20, and nDCG @25 were prepared and executed. The nDCG values for different sets of query documents are shown in Figure 4.7. The overall values are stable and do not vary a lot.

4.4.2 Content based Recommendations

The content based recommendation also sometimes refers to as word level similarity in literature. The word level similarity was used by more than 53% of the researchers who worked in the area of research paper recommendations [Beel et al., 2015].

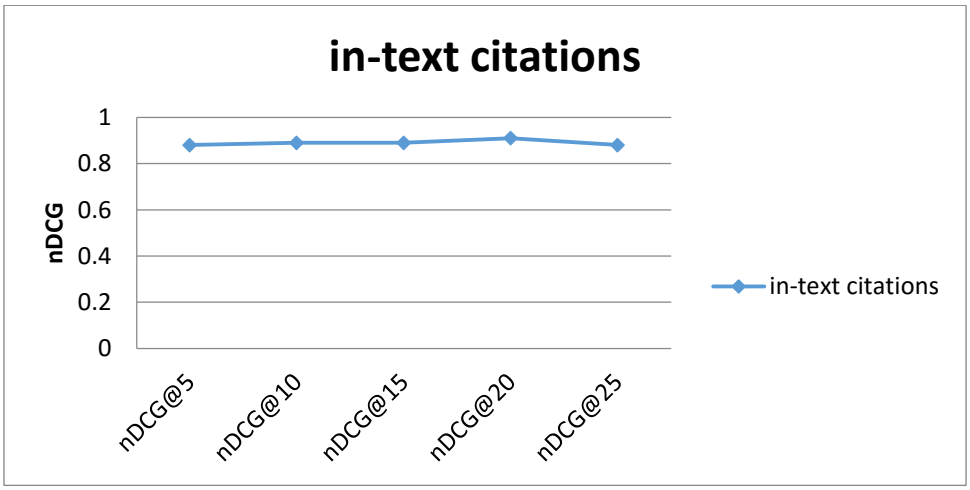


Figure 4.7: nDCG Values of Proposed Approach based Recommendations

A) Paper's Abstracts

For computing word level similarity, the abstracts of cited and citing papers were retrieved. The abstracts of the selected dataset were then indexed using apache Lucene[14] platform. The reason for selecting the Lucene was that it provides a proven, robust, and scalable indexing and retrieval functionality and has been used by many others working in this domain [Liang et al., 2015]. The Lucene accepts documents as a basic unit of information that is used for indexing, storage and retrieval. We have used latest version available to the community i.e. Lucene 4.8.1. The TF-IDF terms vectors were acquired for all of the papers in the selected dataset (i.e. 400 annotated pairs of citation). Finally, cosine similarity was applied to compute document similarity. The overall cosine similarities rounded results up to 2 decimal digits have been shown in Figure4.11. The Lucene provides support for extracting terms from the indexed documents. By default Lucene excludes stop words such as "the", "is", and "and" etc while retrieving terms. The content based technique produced large number of recommendations for each of the source paper (higher recall). It is considered the strength of content based system that they require only two documents to compute relevancy between them. Sample values are shown in Table 4.11 between paper and its selected references. Once these values were computed then it was easy to produce

[14]http://lucene.apache.org/

the ranked list. This rank list was normalized with the help of gold standard ranking. The nDCG values for different sets of queries are shown in Figure 4.8.

A) Paper's whole content

In this second experiment, papers contents were acquired using different sources over the internet. It was found most of the time papers content was not available due to the reason such restricted access or scanned format of documents. The same settings of experiment were applied on the available dataset as explained in previous experiment. It was found that very low similarity was detected because of huge number of noisy terms produced. Few of the terms are listed in Table 4.12 from two indexed papers. The titles of the papers are "Google Scholar's Ranking Algorithm: The Impact of Articles' Age (An Empirical Study)" and "Google Scholar's Ranking Algorithm: The Impact of Citation Counts (An Empirical Study)". As it can be the indexed terms are not properly representing the titled papers. Therefore, due to such issues papers abstracts based document similarity computation was adapted.

Table 4.12: Some of the terms extracted from selected papers

Terms
Lesser
graphic
sharing
now
happy
utilized
Introductive
Elated
Meho
Jeoran
2011
2004
….
….

4.4.3 Bibliographic Analysis

The widely known citation techniques for identification of relevant documents are: bibliographic coupling [Kessler, 1963] and co-citation [Small, 1973]. The fundamental difference between co-citation and bibliographic coupling is that the latter is static is nature whereas the former one is dynamic. In bibliographic coupling,

two papers are considered relevant when they have common references. However, to compute relevant documents based on co-citation, we are required to find those papers that have co-cited targeted papers. Thus, co-citation based recommendations can vary with the passage of time as in future other documents can co-cite those papers. In our current setup, the possible choice was bibliographic coupling. Therefore, common references between citing and cited papers were automatically computed using edit distance algorithm and cross verified manually.

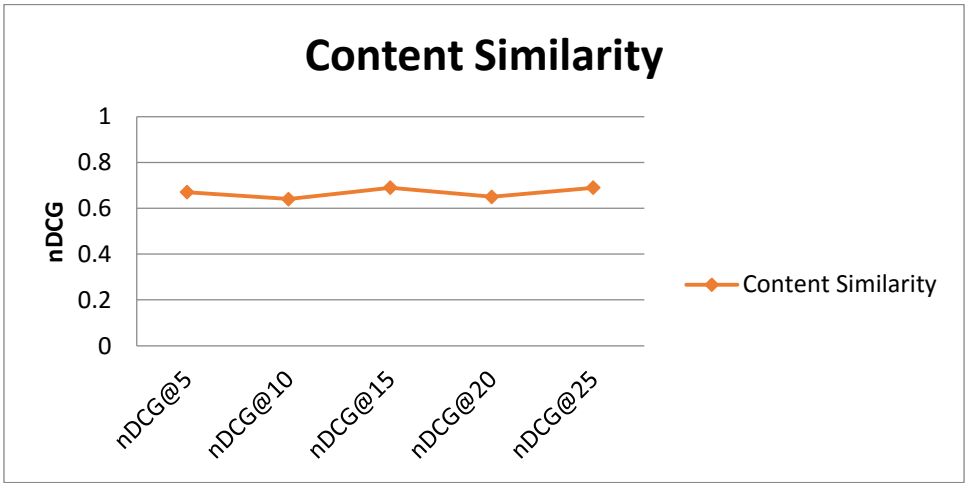


Figure 4.8: nDCG Values of Content Similarity based Recommendations

Afterwards, strength of relationships categories were mapped over different bibliographic coupling units such as 1, 2, 3 etc.

Afterwards, relevant documents were ranked based on a number of common references between citing and cited papers. Relatively low nDCG value was recorded for this technique and that was 0.54. Furthermore, the nDCG values for different set of queries are shown in Figure 4.9.

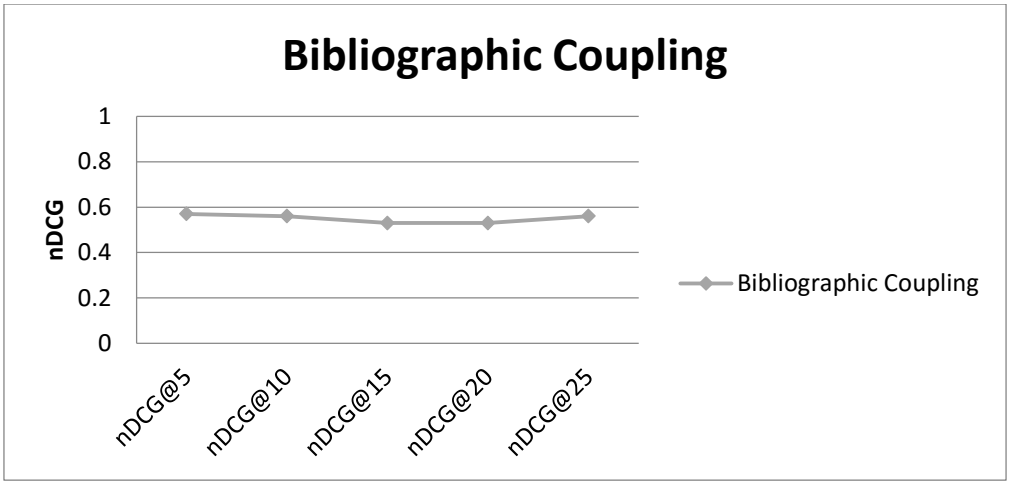


Figure 4.9: nDCG Values of Bibliographic Coupling based Recommendations

Papers may not have common references and thus bibliographic coupling based recommendations will fail to provide any recommendations. This will affect the overall recall of the system. For example, in our case 35% of the time we did not find any bibliographically coupled paper. The overall recommendations made by different techniques are summarized in Figure 4.12.

4.4.4 Metadata based Relevant Documents

Metadata can be defined as data about the data. In the context of research articles, the metadata could be "Title of the paper", "author(s) of the paper", "keywords", "ACM topics (if any)" etc. Technique that discovers the relevant papers based on metadata of research articles are categorized as metadata based techniques.

We also experimented to identify most relevant papers based on different metadata such as papers' title terms matching, keywords matching, and papers' authors matching. Below each one is discussed separately.

A) Paper's Title

An automatic solution was built for this task. The titles of the citing and cited papers were extracted and tokenized based on white spaces. Afterwards, stop words (i.e. "for", "a", "an" etc) were removed and furthermore the filtered terms were stemmed using porter stemming algorithm (Porter 1980). Along-with titles of the paper, some other metadata was also extracted such as authors of the paper, and papers keywords. The purpose of using multiple type metadata was to increase the total number of recommendations produced by this approach.

B) Paper's Authors

Similar to title of the paper, an automatic solution was designed to match the authors of the papers. The authors of the papers were manually extracted and were persisted in database in comma separated manner. For matching authors, edit distance algorithm was applied to find candidate results and then those were manually verified for guarantying 100% accurate results.

The results of experiments on selected dataset revealed that authors of the papers have less capability to recommend relevant papers. However, if authors were found same, then the chance of relevancy was increased because of self- citations.

C) Paper's Keywords

The keywords of the papers were also manually extracted and persisted in database in comma separated manner. Afterwards, using edit distance algorithm they were compared to compute the final results. Papers keywords are very important and can be used to find relevant documents. However, in our case, it was found that most of selected citation pairs were not having the keywords at all.

Once the aforementioned metadata of the papers was ready, then the results were produced. So, based on successful comparison of any of the metadata e.g. paper's title, author, or keywords would result a recommendation. Furthermore, papers were ranked in descending order i.e. maximum matching terms in paper was ranked at the top. In this experiment no any further weight mechanism was used. The achieved results were then compared against gold standard dataset (as explained in section 4.2.1).

Finally the nDCG's values were averaged to compare it with rest of the techniques. It was found that the gain of title+author+keywords based recommendation was around 0.51. The overall nDCG values for different sets of queries are shown in Figure 4.10.

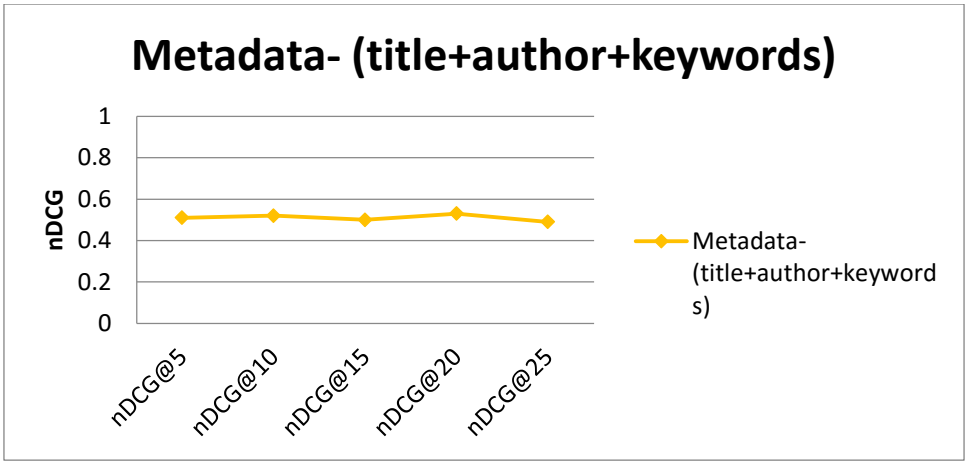


Figure 4.10: nDCG Values of Metadata based Recommendation

All of the research papers have titles, thus, title based recommendations have the capability to provide recommendations for all cases. However, my results indicate that when terms are not matched then metadata based technique does not recommend anything at all. This was reflected in the overall recommendations made by different technique as shown in Figure 4.12. A total of 60% recommendation could be made using title matching technique.

The nDCG values of different techniques are consolidated in Figure 4.11. This Figure represents that in-text citation frequencies based approaches have higher gain as compared to the rest of techniques. The state-of-the-art techniques were tested against different sets of disjoint queries and the results for those sets were uniformed. There was not much change in overall results across those sets of queries.

Apart from the nDCG values, the total recommendations by different approaches in this experiment are also shown in Figure 4.12. The in-text citation frequencies and content has higher recall by providing recommendation for all possible instances. On the contrary, other techniques such bibliographic coupling, title terms matching, authors matching provide less number of recommendations i.e. 65%, 60% and 24% respectively.

My results indicate that in-text citations frequency help in further refining the quality of overall results. In this experiment, in-text citation frequency based identification of relevant documents outperforms the keyword based recommendations.

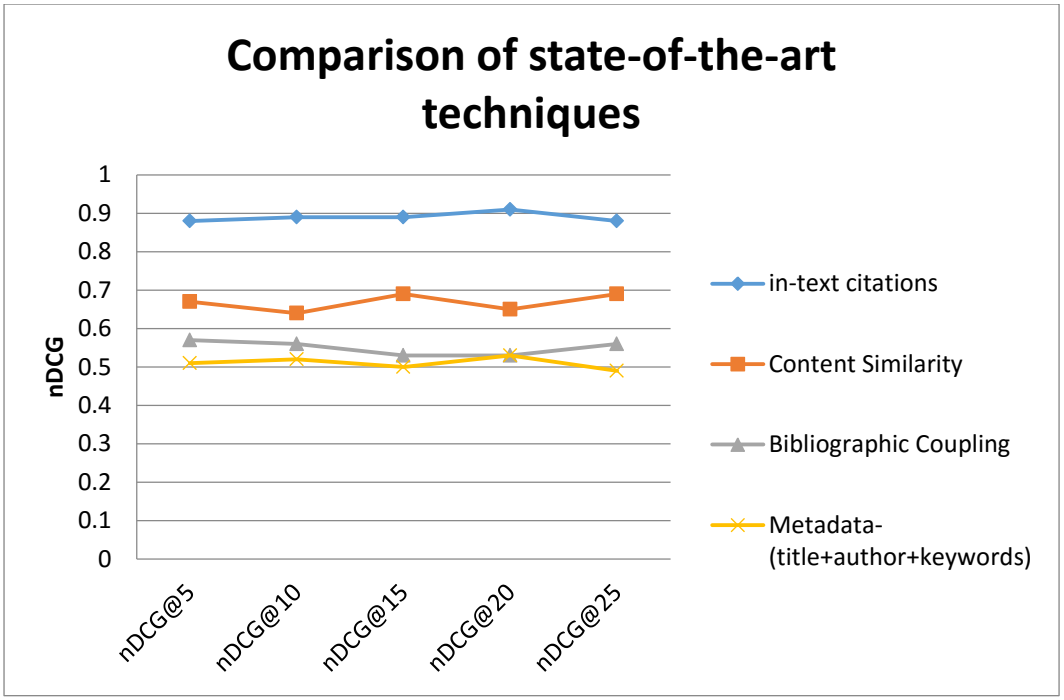


Figure 4.11: nDCG Values of Different Techniques for Different Set of Queries

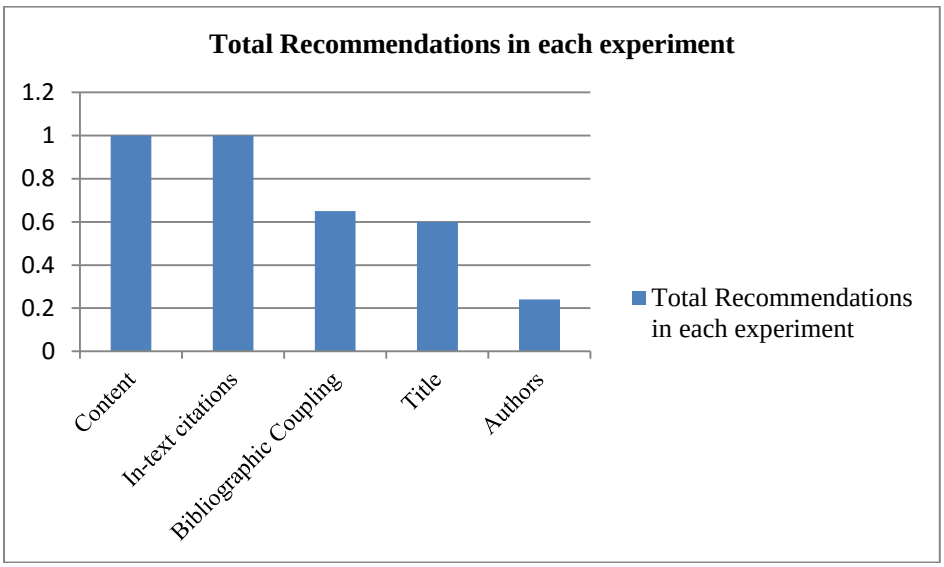


Figure 4.12: Total Recommendations by each Technique

However, there are certain considerations that still need to be explored. I have summarized them as below:

- TF-IDF based scheme was used for terms extraction; other techniques need to be tested in future, for example: Yahoo term extractor and KEA.

- The metadata based approached may improved for integrating some other techniques. For example, keywords can be extracted using some techniques if they are not mentioned explicitly.
- Despite the benefits of in-text citation frequencies based recommendations, it may also add overhead for computation of accurate identification of in-text citation in body-text of the paper.

In summary, despite these considerations, the in-text citation frequencies, content based, metadata, and bibliography based techniques are complementary approaches. The strength of the content based retrieval is to retrieve documents that are not linked up with each other using, for example, citation-link, whereas the in-text citation may help (as we can see in our results) in better recommendations of relevant paper when there exists citation relationship between the documents. ***Thus these results support hypothesis 3.***

Chapter 5

Conclusions and Future Work

Identification of relevant documents is an important task and it is a dire need of scientific community. In literature, different approaches have been proposed to tackle this problem. This thesis presents an evaluation criteria for evaluating these techniques/approaches and systems. The reviewed literature was categorized into four different categories such as identification of relevant documents based on: 1) Content, 2) Metadata, 3) Collaborative filtering, and 4) Citations. Furthermore, the literature was critically reviewed and evaluated based on the defined evaluation criteria.

Based on critical evaluations, it was found that existing techniques and systems have many limitations such as: firstly, they are unable to identify nature of relationships between the research papers; secondly, they are unable to identify relationship strength between scientific papers. Thirdly, content based systems have mainly vocabulary issues, as they compute relatedness just by considering the content of two documents irrespective of the concepts used in the papers. Fourthly, metadata based systems are based on just few terms to identify relationship. Next, citation based approaches consider the citation network information instead of considering the concepts discussed in two papers. Moreover, collaborative filtering suffers from many issues like cold start problem, gray sheep, black sheep, and data sparsity issue etc. Lastly, the results of the contemporary systems are sometimes very easy to manipulate.

To address these issues, the thesis makes three hypotheses:

For these three hypotheses, dataset of around 5,000 papers was extracted and processed. A total of about 105,000 citation pairs along with in-text citation frequencies were computed. Thereupon, 12,000 represented citation pairs were selected in such manner that it has reasonable representation of pairs having in-text citation frequencies in different ranges. In-text citation frequencies ranges refer to, for instance, in-text citation frequencies =1, 2, 3, 4 and finally in-text citation frequencies >=5.

In our domain of research, there exists no gold standard dataset. Therefore, certain citation pairs (i.e. 400 citation pairs) were given to experts of the area for annotation. For annotation purposes, I relied on previous research where renowned researchers have explained different citations reasons an author uses for making a citations. Total 80 experts participated in this study, and I found strong correlation among annotators and thus developed my own Gold Standard dataset.

Based on this developed gold standard dataset, different experimental results were evaluated.

For the first hypothesis, we found that in-text citation frequencies play vital role in identification of degree of relevancy between citing and cited papers. The overall finding can be summarized as following:

Citing and cited papers are strongly related when cited papers are frequently referred in body text of the citing paper and vice versa. It was found that 77% of the times strong relationship was identified where in-text citation frequencies were greater than or equal to 5. Similarly, for lower in-text citation frequencies, weak relationship was marked. For example, 87% of the times weaker relationship was identified where in-text citation frequencies were equal to 1.

The second hypothesis was associated with the nature of relationship between citing and cited papers.

To identify nature of relationship between cited and cited-by papers, rules were constructed based on in-text citation frequencies and in-text citation patterns (in-text citation in different sections of the paper). The results were encouraging. The accuracy of rules for the identification of different nature of relationships, for example, methodologically related, and non-methodologically related were 79%, and 84% respectively.

The results were validated and evaluated with the help of number of detailed user studies. The user studies are a common way for evaluating the proposed approaches especially for evaluating the task of identification of relevant papers. These user studies suggest that in-text citation frequencies have the potential to find most relevant documents in a better way in comparison to contemporary approaches. The proposed system has been implemented as a prototype system for larger dataset

acquired from CiteSeer. The produced results are encouraging, and in support of the argument that in-text citation frequencies and patterns have the potential to discover most relevant documents with high accuracy.

Finally, the proposed approach was compared with word level similarity, metadata, and citation based techniques. All of these techniques were implemented and evaluated on developed gold standard dataset. The qualities of the recommendations from the proposed approach were much higher than the rest. For example, 0.89 nDCG was recorded for proposed approaches whereas 0.66, 0.54, and 0.51 nDCG values were recorded for content similarity, bibliographic coupling, and metadata based techniques respectively. In terms of recall, the word level similarity technique has better recall than our proposed technique.

In this thesis, numbers of contributions were made:
Paper sections mapping technique was designed, implemented, and evaluated that is able to tag each paper's content with standard sections appearing in the scientific document. The overall correctness and completeness of the proposed technique is 77.6% and 74.5% respectively. Identification of the sections in paper was important task as it was required to identify in-text citation patterns.

A novel technique based on in-text citation frequencies and patterns was proposed, implemented and evaluated.

Based on detailed user studies, it was found that the proposed approach has the potential to identify degree of relevancy between citing and cited papers. Furthermore, the proposed approach is helpful in identifying nature of relationship between citing and cited papers.

Lastly, the proposed approach has also been compared with state-of-art (such as content, metadata, and citations based) techniques. In the comparison, the proposed approach outperformed the rest of the techniques.

5.1 Future Work

In general, the in-text citation frequencies and in-text citation patterns based techniques have the strength to apply in many different directions. Few of them are listed below:

- The proposed technique has potential for pursuing future research in different directions. For example, the impact factor can be refined which considers the impact of the cited-by papers instead of only number of citations. The impact of cited-by papers can be found by analyzing in-text citation frequencies of the cited-by paper in citing paper.
- The proposed technique can also be augmented with the rest of approaches; for example, in case of collaborative filtering where papers and citations matrix values can be replaced with citation frequencies values.
- The proposed technique can be adapted to design and formulate innovative visualization that may be helpful in identification of relevant documents in easy and convenient manner.
- The proposed technique can also be integrated with state-of-the-art (i.e. metadata, content, and citations based) techniques to produce more robust hybrid technique for identification of relevant documents.

References

[Afzal et al., 2007] Muhammad Tanvir Afzal et al., "Creating Links into the Future", Journal of Universal Computer Science, vol. 13, no. 9 (2007), 1234-1245

[Afzal et al., 2010] Muhammad Tanvir Afzal et al., "Rule based Autonomous Citation Mining with TIERL," Journal of Digital Information Management, vol. 8, no. 3, pp. 96-204, 2010.

[Afzal, 2009] Muhammad Tanvir Afzal, "Applying Ontological Framework for Finding Links into the Future from Web", In: Proceedings of International Conference on Semantic Systems, pp. 656-662, Graz, Austria, 2-4, Sep. 2009.

[Andrew et al., 2000] McCallum Andrew et al., "Maximum Entropy Markov Models for Information Extraction and Segmentation." In ICML, pp. 591-598. 2000.

[Avancini and Candela, 2007] Henri Avancini and Leonardo Candela, "Recommenders in a personalized, collaborative digital library environment", Journal of Intelligent Information Systems 28.3 (2007): 253-283.

[Beel et al., 2015] Joeran Beel et al., “A Comparison of Offline Evaluations, Online Evaluations, and User Studies in the Context of Research Paper Recommender Systems”, ACM Transactions [yet to be published]

[Beel et al., 2013] Joeran Beel et al., "Research paper recommender system evaluation: a quantitative literature survey.", In Proceedings of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation,

pp. 15-22. ACM, 2013.

[Beel and Gipp, 2009a] Jöran Beel, Bela Gipp, "Google Scholar‘s Ranking Algorithm: An Introductory Overview", In Proceedings of the 12th International Conference on Scientometrics and Informetrics (ISSI’09). Vol. 1. 2009.

[Beel and Gipp, 2009b] Jöran Beel, Bela Gipp, "Google Scholar’s Ranking Algorithm: The Impact of Articles’ Age (An Empirical Study)", Information

Technology: New Generations, 2009. ITNG'09. Sixth International Conference on. IEEE, 2009.

[Beel and Gipp, 2009c] Jöran Beel, Bela Gipp, "Google Scholar's Ranking Algorithm: The Impact of Citation Counts (An Empirical Study)", Research Challenges in Information Science, 2009. RCIS 2009. Third International Conference on. IEEE, 2009.

[Benjamin and Schäfer, 2012] Weitz, Benjamin and Ulrich Schäfer. "A Graphical Citation Browser for the ACL Anthology." In LREC, pp. 1718-1722. 2012.

[Bollacker, 2000] Kurt D. Bollacker, Steve Lawrence, and C. Lee Giles, "Discovering Relevant Scientific Literature on the Web", IEEE Intelligent Systems, Volume 15, Number 2, 2000, pp. 42–47.

[Chen et al., 2011] Chen et al. , "Novelty Paper Recommendation Using Citation Authority Diffusion',Technologies and Applications of Artificial Intelligence (TAAI), 2011 International Conference on. IEEE, 2011.

[Constantin et al., 2013] Alexandru Constantin et al., "PDFX: Fully-automated PDF-to-XML Conversion of Scientific Literature", DocEng'13, Florence, Italy, 2013.

[Cortez et al 2007] Eli Cortez et al., "FLUX-CIM:Flexible Unsupervised Extraction of Citation Metadata", In: Proceedings of Joint Conference on Digital Libraries,pp. 215-224, Vancouver, British Columbia, Canada, 18-23, June. 2007.

[Cremonesi et al., 2012] Cremonesi et al., "Investigating the persuasion potential of recommender systems from a quality perspective: An empirical study", ACM Transactions on Interactive Intelligent Systems (TiiS), 2(2), 11-26, (2012)

[Dalen and Arjo, 2005] Hendrik Van Dalen, and KlamerArjo. "Is science a case of wasteful competition?." Kyklos 58, no. 3 pp: 395-414. 2005.

[Gerstein et al. 2007] Mark Gerstein, Michael Seringhaus and Stanley Fields, "Structured digital abstract makes text mining easy", Nature 2007; 447.doi:10.1038/447142a.

[Giles et al., 1998] C. Lee Giles et al., "CiteSeer: An Automatic Citation Indexing System", In: Proceedings of Third ACM Conference on Digital Libraries, pp. 89-98, Pittsburgh, Pennsylvania, United States. 23-26, 1998.

[Gipp et al., 2009a] Gipp et al., "Citation Proximity Analysis (CPA) – A new approach for identifying related work based on Co-Citation Analysis", in Proceedings of the 12th International Conference on Scientometrics and Informetrics (ISSI'09), Rio de Janeiro, Brazil, 2009.

[Gipp et al., 2009b] Gipp et al., "Identifying Related Documents For Research Paper Recommender By CPA and COA", In Proceedings of The World Congress on Engineering and Computer Science 2009, Berkeley, USA, 2009.

[Gipp et al., 2009c] Gipp et al.,, "Scienstein: A Research Paper Recommender System", In Proceedings of the International Conference on Emerging Trends in Computing (ICETiC'09), pp. 309–315.

[Gori and Pucci, 2006] Marco Gori, Augusto Pucci, "Research Paper Recommender Systems: A Random-Walk Based Approach", Web Intelligence, 2006. WI 2006. IEEE/WIC/ACM International Conference on. IEEE, 2006.

[He et al., 2010] Qi He et al., "Context-aware Citation Recommendation", Proceedings of the 19th international conference on World wide web. ACM, 2010.

[Hersh et al., 2000] William Hersh et al., "Do batch and user evaluations give the same results?", In Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 17–24, 2000.

[Hirsch, 2005] Jorge E Hirsch, "An Index to quantify an individual's scientific research output", PNAS 102 (46), pp. 16569-16572.2005.

[Hou et al., 2011] Wen-RuHou, Ming Li, and Deng-KeNiu. "Counting citations in texts rather than reference lists to improve the accuracy of assessing scientific contribution.",BioEssays 33, no. 10, pp: 724-727. 2011.

[Isaac et al., 2008] Councill, Isaac et al., "ParsCit: an Open-source CRF Reference String Parsing Package." In LREC. 2008.

[ISI, 2013] http://wokinfo.com/ [02Oct 2013]

[Jarvelin and Kekalainen, 2000] Kalervo Järvelinand Jaana Kekäläinen, "IR evaluation methods for retrieving highly relevant documents", In SIGIR 23, pages 41–48, 2000.

[Jannach et al., 2013] Jannach et al., "What Recommenders Recommend–An Analysis of Accuracy, Popularity, and Sales Diversity Effects", In User Modeling, Adaptation, and Personalization. Springer, pp. 25–37, 2013

[Jinha, 2010] Arif Jinha, "Article 50 million: an estimate of the number of scholarly articles in existence." Learned Publishing 23, no. 3 pp:258-263. 2010.

[Joaquin et al., 1998] DELGADO Joaquini et al., "Content-based Collaborative Information Filtering: Actively Learning to Classify and Recommend Documents", Cooperative Information Agents II Learning, Mobility and Electronic Commerce for Information Discovery on the Internet. Springer Berlin Heidelberg, pp:206-215, 1998.

[Jones &Paynter, 1999] Steve Jones, and Gordon W. Paynter. "Automatic extraction of document keyphrases for use in digital libraries: evaluation and applications." Journal of the American Society for Information Science and Technology 53, no. 8, 2002.

[Justin et al., 2012] Mike Justin et al., "Citeology: visualizing paper genealogy". In CHI'12 Extended Abstracts on Human Factors in Computing Systems, pp. 181-190. ACM, 2012.

[Kansas-template, 2013] kansas State University, Electrical and Computer Engineering Department, USA, Research Paper template, Available: http://eece.ksu.edu/~starret/684/paper.html [28 June 2013]

[Kaplan et al., 2009] Dain Kaplan, Ryu Iida and Takenobu Tokunaga, "Automatic Extraction of Citation Contexts for Research Paper Summarization: A Coreference-

chain based Approach",In Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries, ACL-IJCNLP, pp: 88–95, 2009.

[Knijnenburg et al., 2012] Bart P. Knijnenburg,et al., "Explaining the user experience of recommender systems". User Modeling and User-Adapted Interaction, Vol. 22, no. 4, pp. 441–504, 2012.

[Sugiyama and Kan, 2010] Kazunari Sugiyama and Min-Yen Kan. "Scholarly paper recommendation via user's recent research interests." In Proceedings of the 10th annual joint conference on Digital libraries, pp. 29-38. ACM, 2010.

[Sugiyama and Kan, 2013] Kazunari Sugiyama, and Min-Yen Kan. "Exploiting potential citation papers in scholarly paper recommendation." In Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries, pp. 153-162. ACM, 2013.

[Kessler, 1963] Maxwell Mirton Kessler, "Bibliographic Coupling between Scientific Papers". American Documentation 14, 10–25, 1963.

[Khan et al., 2012] Sherafgan Khan et al., "Metadata Based Classification of Scientific Documents," International Journal of Information Studies, vol. 4, no. 4, pp. 184-193, 2012.

[Klink and Kieninger et al., 2001] Stefan Klink and Thomas Kieninger, “Rule-based document structure understanding with a fuzzy combination of layout and textual features”, International Journal on Document Analysis and Recognition, 4(1), 2001, 18-26

[Krovetz, 1993] Robert Krovetz,. "Viewing morphology as an inference process." In Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 191-202. ACM, 1993.

[Lamping and Rao 1994], Jorge Lamping, Rao, R., “Laying out and Visual-izing Large Trees Using a Hyperbolic Space”, In ACM Symposium on User In-terface Software and Technology, pp13-14, Marina del Rey, California, USA, 2-4, Nov. 1994.

[Larsen and Ins, 2010] PederOlesen Larsen and Markus von Ins, "The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index", Scientometrics, pp: 575-603, 2010.

[Liang et al., 2015] Yuan Liang et al., "Exploration and Fulfillment of Search Engine in Economic Model Resource Platform Subsystem Based on Lucene Search Engine." In LISS 2013, pp. 1017-1022. Springer Berlin Heidelberg, 2015.

[Liu and Chen, 2011] Shengbo Liu and Chaomei Chen, The Effects of Co-citation Proximity on Co-citation Analysis", In Proceeding of the International Society for Scientometrics and Informetrics July 4-7, 2011.

[Lo et al., 2005] Rachel Tsz-Wai Lo et al., "Automatically building a stopword list for an information retrieval system." In Journal on Digital Information Management: Special Issue on the 5th Dutch-Belgian Information Retrieval Workshop (DIR), vol. 5, pp. 17-24. 2005.

[Lovins, 1968] Julie Lovins, "Development of a stemming algorithm", MIT Information Processing Group, Electronic Systems Laboratory, 1968.

[Mario et al., 2009] Zechner Mario et al., "External and intrinsic plagiarism detection using vector space models." In Proc. of 3rd Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse and 1st International Competition on Plagiarism Detection, pp. 47-55. 2009.

[Masoud and Kamel, 2008] Makrehchi Masoud, and Mohamed S. Kamel. "Automatic extraction of domain-specific stopwords from labeled documents." In Advances in information retrieval, pp. 222-233. Springer Berlin Heidelberg, 2008.

[Maurer, 2001] Harman Maurer, "Beyond Digital Libraries. Global Digital Library Development in the New Millennium", In: Proceedings of NIT Conference, Beijing, 2001, pp.165-173.

[McNee et al., 2002] Sean M. McNee, Istvan Albert, Dan Cosley, PrateepGopalkrishnan, Shyong K. Lam, Al Mamunur Rashid, Joseph A. Konstan, John Riedl, "On the Recommending of Citations for Research Papers", In:

Proceedings of the 2002 ACM conference on Computer supported cooperative work. ACM, 2002.

[McNee et al., 2006] Sean M. McNee et al., "Don't Look Stupid: Avoiding Pitfalls when Recommending Research Papers", In: CSCW '06: Proceedings 20th anniversary conference on Computer supported cooperative work, New York, NY, USA, pp: 171-180. 2006.

[Miller et al., 1990] George A Miller et al., "Introduction to wordnet: An on-line lexical database*", International journal of lexicography 3, no. 4, pp: 235-244 1990.

[Mu et al., 2006] Li Mu et al., "Exploring distributional similarity based models for query spelling correction." In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, pp. 1025-1032., 2006.

[Naak et al., 2008] Amine Naak et al., "Papyres: a Research Paper Management System", In 5th IEEE Conference on Enterprise Computing, E-Commerce and E-Services, Washington, 2008.

[Nattakarn and Ozsoyoglu, 2007] Ratprasartporn Nattakarn, and Gultekin Ozsoyoglu, "Finding Related Papers in Literature Digital Libraries.", Lecture Notes In Computer Science 4675, pp. 271-284, 200. 2007.

[O'Connor, 1982] John O'Connor, "Citing statements: Computer recognition and use to improve retrieval". Information Processing and Management, 18(3), 1982.

[Page and Brin, 1998] Sergey Brin, and Lawrence Page. "The anatomy of a large-scale hypertextual Web search engine." Computer networks and ISDN systems 30, no. 1. 1998.

[Paice, 1990] Chris D. Paice, "Another stemmer", SIGIR forum, 24,56-61, 1990.

[Perry et al., 1997] Hardin Perry et al., "Statistical Significance and Normalized Confusion Matrices", PhotogrammEng Rem S; 63:735-740, 1997.

[PLOS, 2010] http://blogs.plos.org/everyone/2010/04/02/plos-one-publishes-

10000th-article/

[Pohl et al., 2007] Stefan Pohl et al., "Recommending Related Papers Based on Digital Library Access Records", In: JCDL '07: Proceedings of the 7th ACM/IEEECS joint conference on Digital libraries, New York, NY, USA, ACM (2007) 417-418

[Porter, 1980] Martin F. Porter, "An algorithm for suffix stripping." Program: electronic library and information systems 14, no. 3. 1980.

[Pruitikanee et al., 2013] Siwipa Pruitikanee et al., "Paper Recommendation System: A Global and Soft Approach.", In: Fourth International Conference on Future Computational Technologies and Applications, pp. 21-27. 2013.

[Rashid et al., 2002] RASHID et al., “Getting to know you: learning new user preferences in recommender systems”, In Proceedings of the 7th international conference on Intelligent user interfaces. ACM, pp. 127–134, 2002.

[Rice-template, 2013] Rice University, experimental biosciences, USA, Research Paper template, Available: http://www.ruf.rice.edu/~bioslabs/tools/report/reportform.html , [28 June 2013].

[Sajid el al., 2011] Nasir Ahmad Sajid et al., "Exploiting Reference Section to classify paper’s Topics," in MEDES '11 Proceedings of the International Conference on Management of Emergent Digital EcoSystems, San-Francisco, USA , 2011.

[Schäfer and Kasterka, 2010] Ulrich Schäfer, and UweKasterka. "Scientific authoring support: A tool to navigate in typed citation graphs.", In Proceedings of the NAACL HLT Workshop on Computational Linguistics and Writing: Writing Processes and Authoring Aids, pp. 7-14. 2010.

[Shotton et al., 2009] David Shotton et al., "Adventures in Semantic Publishing: Exemplar Semantic Enhancements of a Research Article.", PLosComputBiol 2009; 5: e1000361. doi:10.1371/journal.pcbi.1000361.

[Shum, 1998] Simon Buckingham Shum, "Evolving the web for scientific knowledge: First steps towards an" HCI knowledge web." In Interfaces, British HCI

Group Magazine. 1998.

[Small, 1973] Henry Small, "Co-citation in the Scientific Literature: A New Measure of the Relationship between Two Documents", Journal of the American Society for Information Science 24(4), pp:28–31, 1973.

[Sölkner et al.,1998] Jhon Sölkner et al., "Genetic variability of populations and similarity of subpopulations in Austrian cattle breeds determined by analysis of pedigrees.", Animal Science 67, no. 02 pp: 249-256, 1998.

[Spiegel-Rusing, 1977] Ina Spiegel-Rusing., "Bibliometric and content analysis" Social Studies of Science, 7:97, 1977.

[Strohman et al., 2007] Trevor Strohman et al., "Recommending Citations for Academic Papers", SIGIR'07, Amsterdam, The Netherlands, 2007.

[Su and Khoshgoftaar, 2009] Xiaoyuan Su and Taghi M. Khoshgoftaar, "A Survey of Collaborative Filtering Techniques," Advances in Artificial Intelligence, vol. 2009, 2009. doi:10.1155/2009/421425

[Swales, 1990] John Swales, "Citation analysis and discourse analysis", Applied Linguistics, 7(1):39–56. 1986.

[Taheriyan, 2011] Mohsen Taheriyan, "Subject Classification of Research Papers Based on Interrelationships Analysis", KDMS'11, San Diego, California, USA. 2011.

[Teufel et al., 2006] Simone Teufel et al., "Automatic classification of citation function", In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2006.

[Huynh et al., 2012] Tin Huynh et al., "Scientific publication recommendations based on collaborative citation networks." In Collaboration Technologies and Systems (CTS), 2012 International Conference on, pp. 316-321. IEEE, 2012.

[Vellino, 2009] Andr´eVellino, "Recommending Journal Articles with PageRank Ratings", Recommender Systems 2009.

[Watanabe et al., 2005] Satoshi Watanabe et al., "A Paper Recommendation Mechanism for the Research Support System Papits", Proceedings of the 2005 International Workshop on Data Engineering Issues in E-Commerce (DEEC'05)

[WEF, 2011] World Economic Forum,http://www.weforum.org/events/world-economic-forum-annual-meeting-2011.

[Wisam et al., 2006] Dakka Wisam et al., "Automatic discovery of useful facet terms." In SIGIR Faceted Search Workshop, pp. 18-22. 2006.

[Yang and Lin, 2013] Wan-Shiou Yang, and Yi-Rong Lin. "A task-focused literature recommender system for digital libraries." Online Information Review 37, no. 4 (2013): 581-601.

[Zhang and Li, 2010] Zhiping Zhang, and Linna Li. "A research paper recommender system based on spreading activation model." In Information Science and Engineering (ICISE), 2010 2nd International Conference on, pp. 928-931. IEEE, 2010.

[Zhiqiang et al., 2009] Lu Zhiqiang et al., "Measuring semantic similarity between words using wikipedia." In Web Information Systems and Mining, 2009. WISM 2009. International Conference on, pp. 251-255. IEEE, 2009.