Indian Statistical Institute, Kolkata

M. Tech. (Computer Science) Dissertation

The Bibliographic Citation Recommendation Problem

A dissertation submitted in partial fulfillment of the requirements
for the award of Master of Technology
in
Computer Science

Author:
Kunal Ray
Roll No: MTCS-1326

Supervisor:
Dr. Mandar Mitra
CVPR Unit, ISI

M.Tech(CS) DISSERTATION THESIS COMPLETION CERTIFICATE
Student: Kunal Ray (MTCS1326)
Topic: The Bibliographic Citations Recommendation Problem
Supervisor: Dr. Mandar Mitra

This is to certify that the thesis titled **"The Bibliographic Citations Recommendation Problem"** submitted by Kunal Ray in partial fulfillment for the award of the degree of Master of Technology is a bona-fide record of work carried out by him under my supervision. The thesis has fulfilled all the requirements as per the regulations of this Institute and, in my opinion, has reached the standard needed for submission. The results contained in this thesis have not been submitted to any other university for the award of any degree or diploma.

Date : *MandarMitra*

Dedication

To my wife, without whose, occasionally annoying proddings, genuine help and wholehearted encouragement, it would not have been possible for me to be where I am today.

Acknowledgements

I would like to thank my dissertation supervisor Dr. Mandar Mitra for agreeing to guide me and for helping me to undertake work in the topic.

I would also like to thank Mr. Dwaipayan Roy, currently pursuing his Ph.D. under the guidance of Dr. Mitra for helping me in many ways for understanding the problem and doing this work. Thanks are also due to Dipasree Pal, Ayan Bandopadhyay and Kripabandhu Ghosh for help and cooperation.

Last but not the least I am grateful to my organization, Proof & Experimental Establishment, a laboratory under the Defence Research and Development Organization (DRDO), Government of India, for allowing me to pursue M.Tech in Computer Science from the Indian Statistical Institute, Kolkata, under study leave.

Abstract

An essential step in authoring a research paper is inclusion of appropriate references or citations. Incorporating relevant references increase academic weight of the paper by presenting links to similar contributions and highlighting the novelty of the work under discussion. This work is becoming increasingly more demanding with increase in volume of published works. Recommender systems for bibliographic citations aim to ease the burden on the author by suggesting possible references globally as well as contextually. More than 200 papers have been published in last two decades exploring various approaches to the problem. In spite of this, no definitive results are available about what approaches work best. Conflicting reports have been published regarding the relative effectiveness of content-based and collaborative filtering based techniques. Arguably the most important reason for this lack of consensus is the dearth of standardized test collections and evaluation protocols, such as those provided by TREC-like forums; forcing research workers to use their own data sets for experiments. A practice that makes objective comparison of techniques a near impossibility. Recent publication of "CiteseerX: A scholarly big data set" makes available raw material for addressing the problem, pending making it into a standard test-evaluation framework. We discuss in this report our efforts in designing a test collection with a well defined evaluation protocol by solving problems with the data set, supplementing the data set with standard queries and their relevance judgments. We also report performances of some standard proposed recommendation approached on our test setup.

Contents

Chapter 1

Introduction

An essential but typically quite daunting step in authoring a research paper is inclusion of proper, important and relevant references/citations in proper places in the document. Appropriate citations indicate relevance and importance of the area of work in general and the specific work under discussion in particular, while highlighting novelty and comparative progress envisioned in the article. Properly placed relevant references provide, to any reader of the article, pointers to relevant as well as contrasting areas of work. Nowadays online access of journal editions and conference proceedings has reduced the problem of access to almost non-existence, but has created the problem of information overload. There are simply too many research papers which can be potentially valuable references. In this scenario it is not difficult to overlook highly relevant references if they have been relatively sparsely cited; this often reduces the impact of the research article. With the increasing volume of published research works, this problem is aggravating at an alarming rate.

1.1 Recommender Systems to the rescue

Recommender Systems constitute a technology that has traditionally been most effective in such a desperate situation of information overload. Recommender systems (RS) are typically semi or fully automated interactive systems that assist human beings in making the so called correct decisions when faced with numerous, closely *matched* options. They are most commonly used in e-retail websites nowadays. An effective recommender system is supposed to aid in making correct decisions. Its effectiveness may be manifested in the system being able to identify and highlight the important characteristics that distinguish the close choices, or in simply producing a pruned choice list. The primary approaches used by RS can be broadly categorized into *content based approaches* and *collaborative filtering* (CF) based approaches. In the first case, the system knows the deciding factors behind the making of choices. These are the content based on which similar/ dissimilar options are presented/ pruned. In the second approach, the system has no idea about the options. It only models user behavioral patterns in making choices, assuming at least a majority of already made choices to be correct. There are also hybrid approaches that attempt to combine the best of

both approaches.

1.2 The bibliographic citation recommendation problem

In the context of the bibliographic citation recommendation problem, the input is a partially or fully written paper with or without the probable citation locations marked. The system returns a ranked list of possible citations for the local or global context as required. A much more advanced system may be able to suggest the optimal locations of the citations also along with their ranked list, given a paper. The typical requirement is a fairly large corpus of preprocessed citation candidates from which the ranked proposals come. Reported approaches for tackling the problem range from content based to collaborative filtering based to hybrid approaches and some novel ideas that are difficult to categorize.

1.3 Motivation

The bibliographic citation recommendation problem is an actively researched area, with more than 200 articles published in the last two decades [1]. However, the absence of a standard evaluation framework, that can be used to quantitatively assess proposed methods has remained a major difficulty. This makes comparison of published works difficult if not impossible. The recent publication of CiteseerX [2] makes available a moderately large sized document corpus in the public domain, and is a first step towards plugging this gap. However, this corpus cannot be used as such for evaluation due to the presence of noise. Further in the absence of standard queries and their relevance judgments, this is nothing more than a fairly large document collection.

1.4 Our work

We have aimed to contribute towards remedying the situation by removing noise from the CiteseerX, collection and generating sets of standard queries with relevant judgments. This converts the document collection into a usable evaluation framework. We subsequently tried a number of baseline approaches on the evaluation framework and have presented the results.

The contributions of this dissertation are:

- Creation of a usable *test collection* for the bibliographic citation recommendation system from the raw data provided by CiteseerX
- Comparison of the following three major approaches to the bibliographic citation recommendation problem using the above as a common test-bed

– Content based search (TF-IDF, BM25, LM) along with query expansion using WordNet.
– Reference Directed Indexing along with a robustness study.
– Statistical Machine Translation based approach.

Chapter 2

Related work

In the group of technologies that aims to reduce human effort in unsupervised or semi-supervised ways, one class that has been very effective comprises recommender systems. It is no surprise that significant research effort has been directed towards utilizing their services to reduce difficulties in writing research papers. A major effort in writing a research article is the inclusion of proper and relevant references in proper and relevant places. Considering the information overload that results from the online availability of almost all journals, and the proliferation of new and upcoming journals and conferences, this involves quite a bit of work. According to a recent survey [1] more than 200 research articles have been published on recommender systems for bibliographic citation, starting with the first publication on recommender systems for research citations way back in 1998 [3]. Principally, three approaches and their combinations have been tried out. They are, in decreasing order,

1. content based approaches,
2. collaborative filtering based approaches,
3. graph based recommendations treating papers as nodes connected by citations as edges.

Here we are primarily concerned with the content based approaches.

Content based approaches

In content based approaches, one typically indexes a corpus of research papers. Subsequently, the client paper (i.e., the paper that is being written) or parts thereof is/are submitted to a searching routine that uses some similarity measure to find and rank search results and displays them. This basic approach has two variants.

Local Context: In this case, the author specifies a citation location, and results are specific for this location. This introduces an associated requirement of choosing the right sized context to form the query. If the chosen context is appropriate, keywords that are most relevant to the search query are included, making the subject matter of the query more focused. On the other hand too large a context will reduce the focus of the query by

incorporating non-relevant keywords. Too short a context, contrarily, will eliminate relevant keywords from the query reducing its worth.

Global Context: Here the full paper is treated as the context. The system in this case has to contend with the highly non-trivial problem of unsupervised identification of citation locations. Also here, the paper as a whole forms the query and since a paper typically talks of multiple, related topics the subject matter of the query is far less well-defined.

Reference directed indexing (RDI)

One interesting approach uses terms from citing documents to index a cited document, instead of the cited document's own terms [4]. This approach, termed reference directed indexing (RDI) [5], yields improved results. The most common approach is to use a variable sized citation context window around the point of citation, to pick up the terms to index. The size can be character based, word based or even sentence based. In this approach, the size of the windows, vis-a-vis, the choice of the terms to index assumes critical importance [6]. This type of concept modeling with a collaborative filtering flavor (the training papers collaborate in their description of a common cited paper) has been used by use of co-citation index and various graph similarity measures [7]. Taking the CF ideas further, some publications have attempted to introduce a personalization flavor by trying to model the users of the system and their research areas [8, 9].

Machine translation based approach

One recently introduced approach, treats the language used in citing paper to be different from the language used in the cited papers. Two variations of this approach has been published. In one case [18] statistical machine translation tools are taught the two languages. A submitted query is translated into the language of cited papers. Subsequently, usual IR techniques are used to find most papers most similar to the translated query. In another approach [19], the cited documents are identified by unique identifiers. All citation contexts of a paper are concatenated into a sentence in source language. The concatenated list of all cited documents, encoded in unique ids, form the paired sentence in target language. These pairs are used to train machine translation tools about the mapping between the languages. A given query is tokenized, and for each term, the transition probabilities to the unique ids are taken. They are boosted by Inverse Document Frequency, an IDF like factor found by treating each citation context as a document, and accumulated. The final list of unique ids are sorted to generate the search output. The second approach has been reported to perform much better compared to the first.

In spite of such a large volume of work the absence of any standard evaluation framework has remained a problem. Most of the published work has used a proprietary corpus of documents, for example Microsoft Academic [10], Rexa Database [7], etc. As a result, it

is difficult to compare the techniques objectively. Some researchers, in the absence of a standard data set, have worked towards building a test collection. Ritchie et al have used in one occasion all accepted papers in a conference [11] to create their data set. In another occasion they have used all papers published up to a period in one field in a journal as the document corpus[12] to build their evaluation data set. To construct the queries and relevance judgments they contacted a number of authors from their corpus and asked for the research question that prompted the papers. These questions formed their queries, and the authors provided relevance judgments on a four point scale. Sugiyama and Kan selected fifty prolific research workers, and used their published works as a document corpus [8]. These are commendable efforts to create test collections, but the collections are very small in size, being of the order of a few thousand to tens of thousands of documents.

As pointed out in the survey referred to earlier, [1] the absence of standard test collections makes comparison of various proposed techniques difficult if not impossible. This is a research necessity as conflicting reports about the effectiveness of various techniques have been published. The recent publication of 'CiteSeerX: a scholarly big dataset' [2] is an effort towards plugging this void. We describe this corpus in the next chapter.

Chapter 3

Creating a test collection

The field of information retrieval being a semi-empirical area of research, the necessity of a standard test collection cannot be overemphasized. To build a test collection, we first need a large document collection. The document collection turns into a test collection once it is supplemented with a set of search queries and relevance judgments. We have used the CiteseerX document collection [2] to create a test collection. The major steps involved are:

- Correcting the noises in the corpus due to improper handling of Unicode characters,
- Supplementing the collection with more than 2500 queries in the form of citation contexts, and relevance judgments.

Subsequently, we have performed test runs on the set following content based frameworks, reference directed indexing and a statistical machine translation based framework. For the first two runs, we have used Apache Lucene for indexing and retrieval. For the third one we have used the "MGiza++" alignment tool for training and prediction. The "trec_eval" software has been used for evaluation of the results.

3.1 Collection Overview

We first describe CiteseerX as a document collection. The collection consists of 630,351 XML files. Each file corresponds to one article and is identified by its DOI as given during generation of the collection. The articles are drawn from various sub disciplines of Computer Science, Communication, Mathematics, Statistics, etc. The files contain appropriately tagged article metadata like venue/ journal, author(s), year of publication, etc. The content of the paper is generally not fully present.

The title and abstract are present in their entirety. The actual content of the paper around each point of citation in the paper body, form a *citation context*, for that citation. The paper body is only present as a compendium of these citation contexts of 2l character window (-l,+l) centered at the point of citation. For this collection l is taken as 200. The citation contexts (of the citing document) are present along with the title, author(s), venue and year of publication of the cited document, tagged together within a single `<raw>` tag, and

a cluster identifier number which is an unique identification number of the cited document. For cited papers which are not part of the collection, a cluster identifier number of 0 is assigned. So the paper body is represented as a collection/ compendium of the *citation contexts* of size 400 characters. Since repeated citations over same context are common, the actual anchoring citation (around which the citation context window is computed) within the context body (in numerical, author/year, and other formats) is marked by a special set of delimiters: "=-=" and "-=-". The full text of the paper is, however, available separately via http://csxstatic.ist.psu.edu/about/data. This data is however not in XML tagged format but is just a text dump of whole paper without the presence of any markup and so is not usable directly.

There are a number of problems with the collection which precludes its use as such. They are:

- Noisy data due to probable improper handling of Unicode characters. For example: `< of the memories: adi = (qi, x) =-=(1) w-=-here (.,.) denotes>`
- Presence of multiple versions of same paper with different DOI but same cluster id, which differs from each other by in most case only a few characters. This has happened as during collection buildup multiple versions of same papers have been collected by crawlers from different source locations. These versions differing at most by a few characters among each other, are identified by different unique DOIs but are the same paper and so share the same unique cluster id.
- Presence of citations which have a non-zero cluster id but are absent from the collection.

In the following sections we discuss ways and means of overcoming these problems.

3.2 Noise removal

Unicode mishandling: As discussed earlier, the citation contexts are four hundred character windows around the citation. A problem has been encountered which probably originated due to the improper handling of Unicode characters during selection of the context window. In a number of cases, the special citation delimiter symbols have been placed within a word, creating two typically meaningless sequences of letters, separated by the delimiter. Thus, during tokenization, a legitimate word is separated into two pieces of gibberish. Many Unicode characters are multi-byte; they appear to have been counted more than once when the delimiter is inserted in the middle of the window. Thus, the delimiters have often been placed a few byte positions away from their intended location. We have rectified this problem by parsing the context, searching for Unicode characters in it, and re-adjusting the positions of the tag symbols appropriately.

Partial words: We often have citation contexts beginning and ending with meaningless partial words. This has happened due to strict adherence to a fixed character window

around the citation context. We have largely ignored this problem since it introduces at most 2 "noise" words in any context.

Multiple citations: Two other possible design decisions taken during the preparation of the data set can create problems for content based search. First, if a reference is cited multiple times in a paper, the 400 character windows around the different citations are simply concatenated together *without any delimiters* to create a single context for that reference. Secondly, if a number of references are cited together, the same context is repeated that many times, once corresponding to each reference. In situations where a reference was not matched with any context, the context simply reads "None".

3.3 Query Generation

As mentioned earlier, to convert a document collection into an evaluation data set, we need to supply a set of standard queries along with relevance judgments. For simplicity we have modeled the bibliographic recommendation as an ad-hoc search problem. The queries are taken to be of four forms:

- each *citation context* is a query,
- title of paper + one *citation context* forms a query (title is common for each query from a paper),
- abstract of Paper + one *citation context* is a query (abstract is common for each query from a paper),
- title of paper + abstract of paper + one *citation context* is a query (title + abstract is common for each query from a paper)

The reference(s) cited in the centre of the context are taken to be the relevant documents for the corresponding query. For example in

```
ralised to MDPs [13], abstract interpretation has been applied to
MDPs [20], and various bisimulation equivalences and simulation
preorders allow model aggregation prior to model checking, see e. g.
, =-=[4, 23]-=-. Recent techniques that have been proposed include
abstraction of MDPs by two-player stochastic games [18], and symmetry
reduction [19]. To our knowledge, threevalued abstraction of continuous
-time s
```

the relevant documents correspond to the citations are identified by the numerals 4 and 23. One (possibly major) drawback of this approach is that, we regard as non-relevant all papers that are cited in the context, but are not part of the central citation. For example the citations identified by the numerals 13, 20, 18 and 19.

To choose query papers, we arranged the papers in the collection in increasing number of citations, rejected the top and bottom 30%, and randomly sampled the rest. This resulted in a set of 226 papers that were used to form queries. For each of these papers, we picked the distinct contexts that also satisfy the following conditions:,

- firstly the context is not "None",
- secondly the relevant gold document belongs to the collection.

This has finally resulted in 2,826 distinct contexts. Each context, together with the title and abstract of the containing paper constitutes a query. Most queries have a single relevant document but some have more.

The decision of considering the actually cited papers to be the gold is questionable. Given the title, abstract and a particular context, the decision of which paper(s) to cite should ideally be objective and consistent across authors. Unfortunately this is generally not true. However, despite shortcomings, this assumption has two advantages:

- firstly it corresponds to a real life situation,
- secondly it eliminates human assessment effort.

Chapter 4

Recommendation approaches

4.1 Content Based Search

We used standard information retrieval paradigms as implemented in Lucene to study elementary content based approaches on the prepared data set and chosen queries. We parsed the documents and indexed the title, abstract and citation contexts using a Lucene analyzer that embodies Porter Stemmer [13] and stop-word removal using the built-in stop-word list for English.

The queries were formulated from the chosen query documents in 3 ways. For the first approach, a query was formed by concatenating, for each query, the title with the abstract of the document, and a citation context. Thus for all the queries from a chosen document, the title and abstract are common. In the second approach we take only the title and a citation context. In the third each citation context forms the query; the title and abstract of the document are not included. For each query, the top 100 documents are retrieved.

The ranking models we tried are:

1. Lucene's built-in TF-IDF vector based ranking scheme,
2. Lucene's implementation of the probabilistic model BM25 [14],
3. language modeling (LM) [15, 16] with Lucene's implementation of both Jelinek-Mercer (JM) and Dirichlet (D) smoothing [17].

A range of parameter values were tried for models in (ii) and (iii). The results as presented later, show that across all ranking models, the third approach to query formation produced significantly superior results as compared to the first. A possible reason can be dilution of importance for common tokens between query and gold, by expansion of the query with title and abstract which principally talk about the citing paper and not the cited ones except in fairly general terms. Also the BM25 model (with proper parameter values) produced superior results compared to the other ranking models for all three query variants, though the differences were not always great.

4.1.1 Query Expansion

A possible problem for any content based approach is vocabulary mismatch. This happens because of different styles of writing adopted by different authors. This is fairly obvious and has been reported in earlier work as well. A well-known remedy for the situation is to expand the query by adding semantically related terms in the hope of achieving increased vocabulary overlap. This is called *query expansion*, it comes with the possibility that expanded queries might mean something different from the original, a phenomenon called *query drift*.

Wordnet based approach

The Wordnet is a standard lexical resource for natural language processing. We tried expanding the queries by using synonyms from Wordnet. The synonyms generated from WordNet were also included along with the original query terms in the expanded query. To evaluate the performance of query expansion we tried the third approach of query formation with the default ranker for Lucene. Since Wordnet based synonym expansion blows up the query size to a great extent, the third approach was chosen because it produces shorter original queries. The performance was poor as compared to unexpanded queries. This is most likely due to query drift caused by the presence of all possible synonyms of query terms.

It appears that query expansion, using a general-purpose lexical resource like WordNet does more harm than good for content based approaches. However, we could not undertake a really thorough and detailed study involving various techniques for query expansion, which is undoubtedly required before one can draw reliable conclusions in this regard.

4.2 Reference Directed Indexing

We have seen that content based approaches are only moderately effective. A possible reason for this is the lack of word overlap between source and query. We hypothesize that the mismatch occurs because, a paper discusses many things in a language local to the paper only, while the papers citing this paper use language local to the citing paper only. This is because individual authors may have their distinctive ways of describing a concept which do not necessarily overlap by a large extent. Whenever they do, we get good results from content based searches, but not otherwise. This problem was investigated by Bradshaw [5] and later Ritchie et al [4]. They suggested indexing documents with the terms used by the citing papers. This ensures that we always identify papers with one particular language; that of the citing documents. Of course, this will vary across the citing documents, *but hopefully not by a great extent*. This is the idea behind *reference directed indexing*.

To generate data for this approach we take following steps:

for each file in collection

parse the file and collect the citation contexts and their gold cluster ids.

for each citation context

if the citation context is not *None* and the cluster id is not *0* and there exists *DOI* for this cluster id in the collection,

then

remove single letters and numerals from the citation context

open a *text file* in the name of the cluster id

append the modified citation context to the file and add a newline

close the file

We call the resulting data reference directed (RD) data. In the RD data each cluster id is represented by a text file containing the *citation contexts* arranged as one context per line. Once this was done, we removed all citation contexts that were present in the queries. We call the resulting data RD pruned (RDP) data. Subsequently, this data was indexed using Lucene. We did not drop stop-words; we did not perform stemming. We used the BM25 similarity measure with the set of parameter values that gave good results for content based search. The results, presented in Chapter - 5, show a massive improvement.

Robustness study for RDI: To study the robustness of the approach, we next undertook the following study. We randomly dropped a percentage of the citation contexts from the RDP data and repeated the indexing and searching. This was done from 10% drop (90% retention) to 80% drop (20% retention) in steps of ten. Each study was repeated ten times. This means cluster id files with 10 or fewer lines of context information were untouched. The files with say 15 lines were reduced to 10 lines for all drops of more than 30%. The results of this study show that the performance degrades with increasing percentage of drop as expected, but slowly; and even with only 20% retained data, the performance is superior to the best obtained using content based search.

This indicates that reference directed indexing is a very effective idea. However, as discussed by Ritchie et al [4, 6], it is a highly non-trivial problem to correctly identify important terms for indexing from the citing document. The CiteseerX data originally comes with the paper body being a compendium of 400 character context windows of citations only. So no experiments could be performed to determine whether this size is optimal or not. We also did not study whether all terms of the window are equally important or not, Ritchie et al defined the context windows in terms of word and sentence count. They reported improvements with increase in windows size up to a point. With *vary* large windows, the performance fell, probably due to query drift. They further observed that both word based and sentence based context windows work similarly as long as they are long enough. However, as their collection size was small (9800 documents), more work is needed before arriving at a concrete conclusion.

A major criticism of the approach of RDI is that, to work properly it requires papers to be cited at least once, which may not hold for recent papers. This is a valid criticism as in such a case the paper will be in the collection but there will be no entry for it in the index.

4.3 Machine Translation Based Approach

A novel idea tries to address the vocabulary mismatch problem between a citing paper and a cited paper by treating the two to be in different languages. Then one uses statistical machine translation tools to learn the mapping between the two languages[18]. A search query then becomes a translation client, and the translated output produces the query in the cited paper's vocabulary. Subsequently, one searches for that query using usual IR techniques. So this is a context aware technique with translation replacing query expansion.

A recent paper [19] uses similar ideas but uses unique identifiers (cluster ids in our case) of cited papers as the target language. The primary problem of vocabulary mismatch between citing and cited papers is again considered, but while one assumes that the citation contexts are written in natural language, the citations are encoded in terms of the unique identifiers like cluster id, and are in the so called reference language. The statistical machine translation tool then learns to translate between the two. As training data, all citation contexts of a paper are concatenated to form one sentence of the source language. All the corresponding citations (encoded in unique id) are then concatenated forming the corresponding line of the target language. Thus, each paper in the training data contributes one sentence pair. After the learning is over a transition table is generated. This table provides translation probabilities of words in contexts to unique ids in reference list.

Given a query:

the query is tokenized

for each term in the tokenized query

from the transition table pick up all cluster ids with non-zero transition probability

boost all the probabilities by the Inverse Context Frequency (ICF) (which is nothing but IDF treating each context as a document)

for each cluster id picked in previous step accumulate the boosted probability

perform descending sort of all the picked up cluster ids for the accumulated probability values

return top n cluster ids of the sorted list as search response

As per the recent paper [19], the later technique performs better than the former one of query translation and searching. So for testing the performance of statistical machine translation approach to bibliographic citation recommendation problem, we used the later approach for citation recommendation on our evaluation framework. We have used, as in the paper, Giza++ to generate the transition table. Subsequently we have ran our standard queries with ICF boosting. The results are presented in Chapter - 5.

Chapter 5

Results

Table 5.1 presents the results obtained using various approaches on the test collection, and described in Chapter 4 using the developed evaluation framework. For all cases of content-based methods, the corpus data have been indexed in three fields: tile, abstract and body (consisting of citation contexts); the searching has been done taking all three fields weighted equally. For the reference directed indexing, the indexed data is only the citation contexts of citing papers and so only one field was used.

From the table we see that among the content based approaches, there is an order of magnitude difference between queries formed by taking only contexts, and queries formed by taking title and abstract along with the contexts. The case of query formed by taking title and context is only slightly worse than that of only the context. The most likely reason behind this, is the dilution of focus in the query by terms relevant to whole document, and not only to the query concerned. This is further confirmed since the inclusion of title in query worsens the result slightly due to much smaller number of terms coming from it. From the table we further see that the RDI approach is far superior compared to the other approaches for all similarity measures. Notably, the simpler TF-IDF similarity measure provided by Lucene outperforms the sophisticated approaches of both BM25 and Language model. This indicates that the basic idea of RDI, of citing documents agreeing in their description of the cited document, is valid and quite useful. The reported efficiency of machine translation based approach is not reflected in the results generated by us, with the performance close to but poorer than the context only query formation approach of content based search. A possible reason behind this apparent anomaly might be that, the so called reference language is just a sequence of unique numbers, and not sufficiently like a natural language to exploit the power of statistical machine translation tools.

We next present the data for the robustness study of the RDI approach. The data is presented in tabular as well as graphical format. The table 5.2 and the figures, show that the RDI approach appears to be remarkably robust as the performance deterioration with reduction in data slow and steady. A point to note is that even when only 20% of the corpus is used to create the reference directed index, retrieval effectiveness remains significantly superior to that obtained using other approaches.

Table 5.1: Baseline Results

Content-based methods

Model	Query From	Relevant Returned	MAP	Reciprocal Rank
BM25 (1.2,0.4)	Title + Abstract + Context	1221	0.0868	0.0886
BM25 (1.2,0.4)	Title + Context	1573	0.1546	0.1583
BM25 (1.2,0.4)	Context only	1549	0.1621	0.1660
Dirichlet (100)	Context only	1547	0.1582	0.1621
Jelinek-Mercer (0.5)	Context only	1573	0.1642	0.1681
Lucene TF-IDF	Context only	1445	0.1488	0.1527
BM25 (1.2,0.4) + WordNet	Context only	644	0.053	0.0545

Reference Directed Indexing

Model	Query From	Relevant Returned	MAP	Reciprocal Rank
BM25(1.2,0.4)	Context only	2550	0.4978	0.5024
Lucene TF-IDF	**Context only**	2578	**0.5022**	**0.5070**
Dirichlet(100)	Context only	**2584**	0.4713	0.4765
Jelinek-Mercer(0.5)	Context only	2617	0.4921	0.4966

Translation Based Approach

Model	Query From	Relevant Returned	MAP	Reciprocal Rank
Translation Approach	Context only	1308	0.0934	0.0958

Table 5.2: Robustness Study for RDI. The BM25 (1.2,0.4) model has been used for all experiments reported in this table.

% corpus used for RDI	Relevant Returned		MAP		Reciprocal Rank	
	Mean	SD	Mean	SD	Mean	SD
90%	2550.1	6.35	0.49	0.0018	0.50	0.0017
80%	2542.8	8.39	0.49	0.0026	0.49	0.0028
70%	2520.8	7.28	0.47	0.0019	0.48	0.0018
60%	2497.9	7.45	0.47	0.0032	0.47	0.0034
50%	2496.7	5.1	0.46	0.0029	0.46	0.0029
40%	2469.7	9.80	0.44	0.0019	0.45	0.0018
30%	2451.1	12.81	0.42	0.0037	0.43	0.0040
20%	2435	8.39	0.41	0.0037	0.42	0.0037

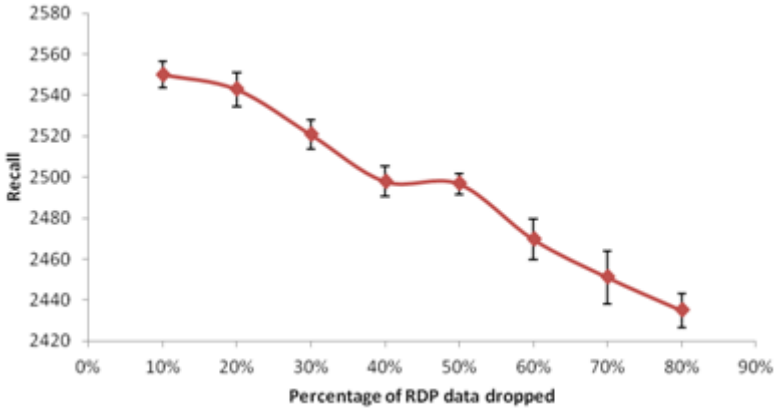

Figure 5.1: Recall

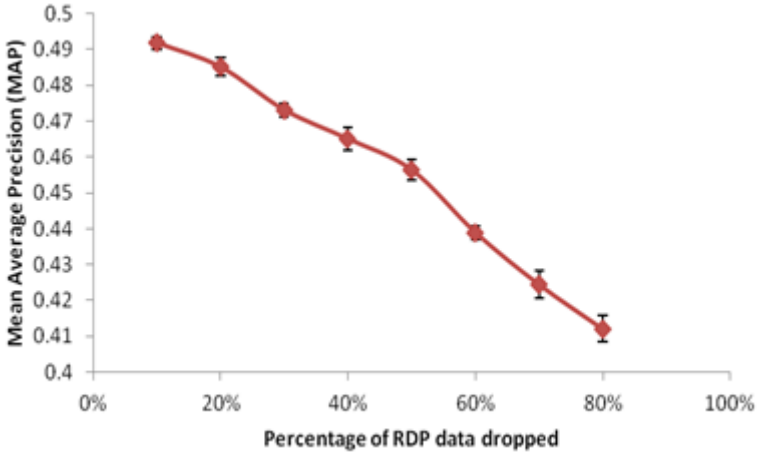

Figure 5.2: Mean Average Precision

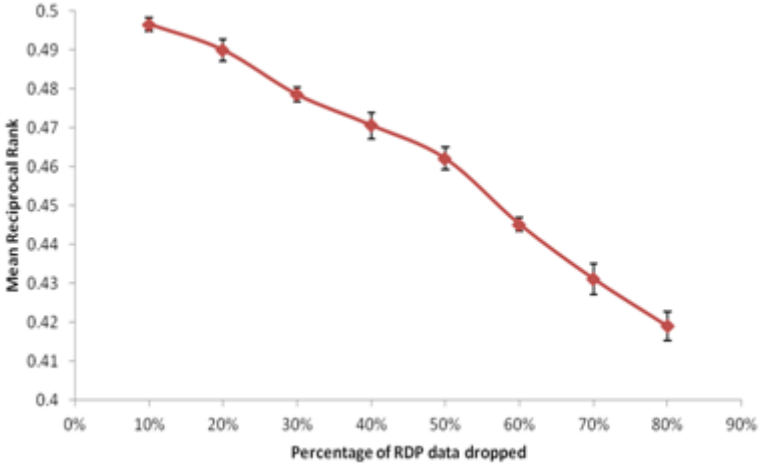

Figure 5.3: Reciprocal Rank

Chapter 6

Conclusion

The problem of bibliographic citation recommendation is one of the actively researched fields today. In spite of a significant volume of published work, a glaring void has remained in the form of the absence of publicly available standard evaluation frameworks. This is a major obstacle in objectively comparing various proposals. CiteseerX, a moderately large collection of documents, has been published recently in an attempt to plug this void. However, in the absence of queries and relevance judgments, this remains just a document collection. Moreover, the collection has quite a few problems which prevent its use as such for evaluation.

We have:

- tried to remove noise due to improper handling of Unicode characters,
- have constructed a set of standard queries along with relevance judgments,

to convert the document collection into a standard evaluation framework.

We have used the test collection to compare a number of baseline as well as recently proposed approaches, which include content based search approaches including query expansion using WordNet, reference directed indexing, and a statistical machine translation based approach. The observations can be summarized as follows:

- We have seen by far the best performance has been achieved with reference directed indexing, for which we have demonstrated robustness to random deletion of data as well.
- It appears that, in spite of its sophistication, the translation based approach does not yield very good results. This is contrary to expectation as it has been reported to work very well [18, 19].
- Content-based search has also been reported to be ineffective, but we find that the results are not as bad as are usually reported if only the citation context is considered as query.

However, much work needs to be done. A more careful study of various approaches is required before we can arrive at concrete judgments about the merits and demerits of the

various techniques. We have also not undertaken a detailed study about the query expansion problem. We have also been unable to undertake studies to judge whether the 400 character citation context window is optimal or not. These remain as future work which we hope will be undertaken later.

Bibliography

[1] J. Beel, B. Gipp, and C. Breitinger, "Research paper recommender systems – a literature survey," *International Journal on Digital Libraries*, 2015.

[2] C. Caragea, J. Wu, A. Ciobanu, K. Williams, J. Fernández-Ramírez, H.-H. Chen, Z. Wu, and L. Giles, "CiteSeer x: A Scholarly Big Dataset," in *Advances in Information Retrieval*, pp. 311–322, Springer, 2014.

[3] C. L. Giles, K. D. Bollacker, and S. Lawrence, "CiteSeer: An automatic citation indexing system," in *Proceedings of the third ACM conference on Digital libraries*, pp. 89–98, ACM, 1998.

[4] A. Ritchie, S. Teufel, and S. Robertson, "Using terms from citations for IR: some first results," in *Advances in Information Retrieval*, pp. 211–221, Springer, 2008.

[5] S. Bradshaw, "Reference directed indexing: Redeeming relevance for subject search in citation indexes," in *Proceedings of the European Conference on Research and Advanced Technology for Digital Libraries (ECDL)*, pp. 499—-510, 2003.

[6] A. Ritchie, S. Robertson, and S. Teufel, "Comparing citation contexts for information retrieval," in *Proceedings of the 17th ACM conference on Information and knowledge management*, pp. 213–222, ACM, 2008.

[7] T. Strohman, W. B. Croft, and D. Jensen, "Recommending citations for academic papers," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 705–706, ACM, 2007.

[8] K. Sugiyama and M.-Y. 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, 2013.

[9] J. Lee, K. Lee, and J. G. Kim, "Personalized academic research paper recommendation system," *ArXiv Preprint*, vol. http://arxiv.org/abs/1304.5457, pp. 1—8, 2013.

[10] A. Livne, V. Gokuladas, J. Teevan, S. T. Dumais, and E. Adar, "CiteSight: supporting contextual citation recommendation using differential search," pp. 807–816, ACM Press, 2014.

[11] A. Ritchie, S. Teufel, and S. Robertson, "Creating a test collection for citation-based IR experiments," in *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pp. 391–398, Association for Computational Linguistics, 2006.

[12] A. Ritchie, S. Robertson, and S. Teufel, "Creating a test collection: Relevance judgements of cited & non-cited papers," in *Large Scale Semantic Access to Content (Text, Image, Video, and Sound)*, pp. 251–261, LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, 2007.

[13] M. F. Porter, "Readings in information retrieval," ch. An Algorithm for Suffix Stripping, pp. 313–316, 1997.

[14] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: Bm25 and beyond," *Information Retrieval*, vol. 3, no. 4, pp. 333–389, 2009.

[15] D. Hiemstra, *Using language models for information retrieval*. PhD thesis, Ph.D. thesis, University of Twente, 2001.

[16] J. M. Ponte and W. B. Croft, "A language modeling approach to information retrieval," in *Proceedings of the 21st ACM SIGIR Conference on Research and Development in IR*, SIGIR '98, pp. 275–281, 1998.

[17] C. Zhai and J. Lafferty, "A study of smoothing methods for language models applied to ad hoc information retrieval," in *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 334–342, ACM, 2001.

[18] Y. Lu, J. He, D. Shan, and H. Yan, "Recommending citations with translation model," in *Proceedings of the 20th ACM international conference on Information and knowledge management*, pp. 2017–2020, ACM, 2011.

[19] W. Huang, S. Kataria, C. Caragea, P. Mitra, C. L. Giles, and L. Rokach, "Recommending citations: translating papers into references," in *Proceedings of the 21st ACM international conference on Information and knowledge management*, pp. 1910–1914, ACM, 2012.