Master of Science in Computer Science
March 2017

An approach to Language Modelling for Intelligent Document Retrieval System

Aditya Kamma

Faculty of Computing
Blekinge Institute of Technology
SE-371 79 Karlskrona Sweden

This thesis is submitted to the Faculty of Computing at Blekinge Institute of Technology in partial fulfillment of the requirements for the degree of Masters of Science in Computer Science. The thesis is equivalent to 20 weeks of full time studies.

Contact Information:
Author:
Aditya Kamma
E-mail: adka14@student.bth.se

University advisor:
Lawrence Henesey
Department of Computer Science and Computer Systems Engineering
Email: larry.henesey@bth.se

Faculty of Computing	Internet : www.bth.se
Blekinge Institute of Technology	Phone : +46 455 38 50 00
SE-371 79 Karlskrona, Sweden	Fax : +46 455 38 50 57

ABSTRACT

The integration of document indexing and document retrieval model has been extensively received by many researchers, but due to lack of adequate indexing model it is difficult problem. A better indexing model would be helpful to solve the problem. In this study, different techniques for document retrieval models using snowball sampling are identified I the Literature Review. Based on the results of the literature review an indexing model was identified from the retrieval models and retrieval algorithms. Calculate the precision/recall values for the document and compare the results of the proposed model with ordinary language model. Experiment was conducted on two different query sets of TREC 202-250 and 51-100 based on term weighting and structured queries.

Conclusions: Like any retrieval model, the proposed model performs an outstanding performance than the ordinary language model which was performed on two different TREC query sets. We can conclude that a better indexing model with use of term weighting and ranked retrieval provides fast and easy access for the documents in the retrieval systems.

Keywords: Retrieval System, Information Retrieval, Language based Modelling.

ACKNOWLEDGEMENT

I thank my supervisor, Asst. Prof Lawrence Henesey for his magnificent supervision. I thank him for his patience and encouragement. This work would not have been possible without his immense knowledge, exceptional guidance and extraordinary support.

I am grateful to my family- Daddy, Mummy for believing in me till date and possibly forever. Special thanks to my dear friend who stood by my side in thick and thin. I dedicate this fulfillment to everyone who is special to me with a promise that the best is yet to come!

Finally, to all the friends and foes who made me laugh and cry, thank you for making my life at Karlskrona the most memorable and delightful.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

IR	**Information Retrieval**
ED	**Electronic Documentation**
PDA	**Personal Data Assistant**
RA	**Retrieval Algorithms**
RM	**Retrieval Models**
TREC	**Text REtrieval Conference**

1 INTRODUCTION

A universal storehouse of Knowledge and culture in the Digital world has allowed a direct sharing of ideas and information in unpredictable rate. So, there is a need of accessing the data in the digital world in the form of documents. These documents are useful for sharing the information to every user which is considered as Information Retrieval. Document retrieval is most commonly referred as Information Retrieval by researchers in the field. It is a computerized process of producing relevance ranked list documents based on inquirer's request. Does Information Retrieval require any introduction to present day technologies? Because most of the surveys shows that 85% users of internet use search engines (e.g. Google, Yahoo, Alta Vista) to find information[1] which is in the form of documents. In the past, traditionally information is stored in the form of a conventional book format but not taking the advantage of present technology offered by modern computer systems and networks. However, the use of paper book is not only inconvenient but counter-productive in many industries e.g. In maintenance industry, field engineers carry their manuals daily to work, as these manuals are bulky and often not up to date[2]. So, the need of Electronic documents that provide quick access to right information for the right user.

The delivery, storage and retrieval of electronic documents is called Document Retrieval System (DRS) which is being currently undertaken. The technology provided is to make sure the information is provided reliably and delivery to right person, right place, at the right time and in the right format. The main purpose of this system is to select from a large quantity of documents to a manageable number that satisfy a need for information. The primary problem in Document retrieval system is representation, representing documents by storing them in some form of description in a database. In order, to accomplish this task the system needs a representation of each document in the database. Many of the modern information retrieval systems, like search engines are designed especially for the users who aren't familiar with representation, collection of documents and use of Boolean operators. The main requirements of these systems are, firstly the user should be able to enter the natural language like words, phrases and sentences etc., without the use of operators. This implies a full text information retrieval system in which every word in the document is indexed in the system. Secondly the system should rank the retrieved documents by their degree of probability. Thirdly, the system should support the search formulation from the user feedback. These three requirements form the base for the research presented in this thesis.

The following sections introduces the technical vocabulary and disciplines of Information retrieval.

1.1 A Definition

According to Savino and Sebastian[1], Information retrieval is the name of the process or method where by an user of information is able to convert his need into actual list of citations to documents in storage containing information which is useful to the user. It also stores and manages the information on documents. These systems assist the users in finding the useful information they need and they explicitly don't return the information, they instead return the locations and existence of the document that information might contain. The documents which satisfies the user information are relevant documents and doesn't satisfies are irrelevant documents. A perfect retrieval system would be that retrieves relevant documents but no irrelevant documents, but there will not be any perfect retrieval system as search statements are necessarily incomplete and relevance depends on subjective opinion of the user.

1.2 Background

In the recent past, Computer Hardware has become increasingly cheaper. So people trend to use them not even at work but also at home. The trend of using Laptops, and PDA made PC's to lose their dominance. Handling of documents from more available devices will increase the trend more and more editing and storing of documents. The rest of the section provides the reader a brief introduction to different fields of DRS and its State – of –the Art techniques in this field of Information retrieval.

1.2.1 Electronic Documents

Electronic Document is the soft copy of a paper document can contain a non-sequential information stored in a manner that requires an electronic device to display, interpret and process it. This contains the documents stored on a magnetic disk, optical media as well as electronic mail. It is becoming increasingly prevalent in today's life. Scientific papers are made available in the internet in electronic form are widely used every day. Electronic documents provide competitive advantages to financial institutions where the documents are intensive to business. It also provides massive amounts of time and space to transportation industries by going to paperless documents, as transportation industries have tons of papers like invoices, tracking records and more.

1.2.2 Information Retrieval

There are three main basic process an Information retrieval system has to support: representation of documents, representation of user information and comparison of both representations.

The documents are usually represented using indexing process. This indexing process results in formal representation of the documents, so this turns to be a document representation or index representation. An algorithm that identifies words in a text and put them into lower case to derive the index representation using trivial algorithms in some text retrieval systems. In the indexing process the actual storage of documents contains partial information for instance like title, abstract and information.

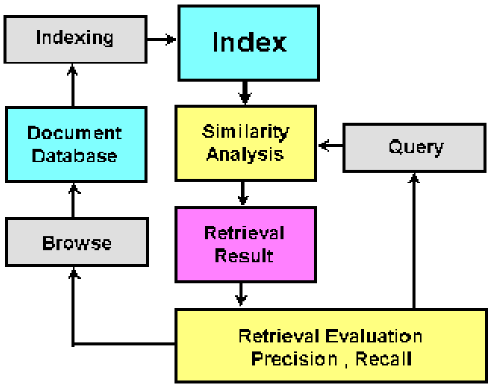

Figure 1.1: An over view of Information Retrieval.

Query formulation is defined as process of representing the information problem or need, it results in formal representation of query. In broad sense query formulation is defined as complete interaction between the user and system, leading not only a query and but also a better understanding of user information need. Query formulation is generally denoted as automatic formulation of query where there are no previous retrieved documents, which is the formulation of initial query and formulation of successive queries is called relevance feedback.

The user and system communicate with each other using respective queries and retrieves the set of documents. The most natural form of communication is used to communicate with each other for information needed, such a natural communication method is called request. In the automatic query formulation, it takes the input as request and gives the output as initial query. Based on the initial query some or all words in the request are converted to query terms by a trivial algorithm. Relevance feedback inputs the initial query to some retrieved relevant or irrelevant documents to output a successive query. There are different approaches used to retrieve the information e.g. Boolean retrieval, Genetic Algorithms, Vector-space model. The use of Information Retrieval is ease of access to information in a short period. It is used in almost every field of computer science to access the information quick and fast.

1.2.3 Mathematical models of Information Retrieval

The models used for Information Retrieval guides the Information retrieval systems. Generally Professional searchers use traditional information retrieval systems in which matching process is automated and indexing, query formulation are processed manually. For this kind of traditional information retrieval systems matching process is only modelled. Boolean model is perfect example for matching model which either retrieves the documents or not without ranking them. In the present-day Information retrieval systems like search engines which are usually used by non-professional users, query formulation is automated. But mathematical models for these systems still model the matching process. There are many models used by Information retrieval systems for matching process of ranked retrieval systems, which are called as approximate matching models. This model use frequency distribution of terms to compute the ranking of retrieved sets. From the practical view, all the classical retrieval model like

Boolean model, vector space model and probabilistic model represents problems of information retrieval representing structured queries, initial term weighting and relevance feedback.

1.3 Aims & Objectives

Aim:

The project is primarily designed to identify the different mathematical models of information retrieval and calculate retrieval time of each document using Language Modelling and provides the user a better Information retrieval system.

Objectives:

- Identify different Information retrieval systems that are existing,
- Identify different mathematical models and algorithms used in the retrieval systems.
- Calculate the retrieval times of the documents using the identified model.
- Compare the values of the proposed model with other statistical models used in the retrieval systems.

1.4 Research Questions

RQ1: Which mathematical models and algorithms that are used in the Information retrieval systems?

RQ2: What are the precision and recall values of documents using Language Modelling?

RQ3: Does the proposed model provides the better results than other Statistical models?

1.5 Research Gap

DRS is the major expansion in the domain of information retrieval, most organizations at enterprise level have a little understandings and commitment to high quality documents access and management level. So there is a need for building a repository of documents with research and development in Information Retrieval. Much research has been done on the retrieval of documents and implemented different retrieval strategies but there are existing models of Information retrieval that tops in the performance of term weighting algorithms and none of the existing model accounts for indexed modelling using both term weighting and ranked retrieval. The main intension of the research is to improve the performance of Information retrieval systems by implementing Language based Modeling.

1.6 Thesis Outline

The following this is outlined as follows: The “Introduction” provide the reader the short description of about the content of study that will be dealt with in the thesis work. The “Related work” section of the documents describes in short, other studies that have been conducted in latest tools and retrieval techniques used for fast retrieval of documents. The next section describes the different models used for Information retrieval. The next section explains about the proposed model used for Information retrieval systems. The “Methodology” section then explains about the how the thesis has been conducted followed by “Results” and “Analysis” section that demonstrates the findings of the study and analyses the findings using statistics. The “Discussion” section validates the study and finally the “Conclusion” sections demonstrates the knowledge gained through the studies.

2 RELATED WORK

The need for improving the communication between the business partners and supporting the data exchange between the business partners in the form of electronic documents has led to the trend of Electronic data interchange[3] and managing the documents by quick and easy retrieval techniques is Document Retrieval System[4].

As organizations are moving towards the paperless environment due to increase in number of paper documents and storage space made the Document Retrieval System (DRS) must be in place to cope the increase in volumes of documents and made documents available for future use. This is proven to be an appropriate tool for handling the documents in transportation industries.

Owen stated that access and retrieval of documents is more efficient and reported with no missing and loss of files and efficiently managing documents and records through its life cycle[5]. [6]explains 90% of Garten Group information is contained in the documents and 40% of their time in dealing with their documents. So document retrieval has been taken a prominent role in Information Retrieval System[7]. Information Retrieval System is to acquire desired information in response to user queries will help the users to efficiently acquire desired information[8] in order to build that system J.Zobel at al proposed inverted schema for fast access file system[9] and Doszkocs et al provided a document retrieval overview with the use of connectionist models[10]. FIRE (Friendly Information Retrieval Engine) a new multimodal interface situated in an Intelligent Environment for retrieving information from the web[11]. A retrieval tool called Ontalk provides a semi-automatic metadata generator and an ontology based search engine for electronic documents [12].

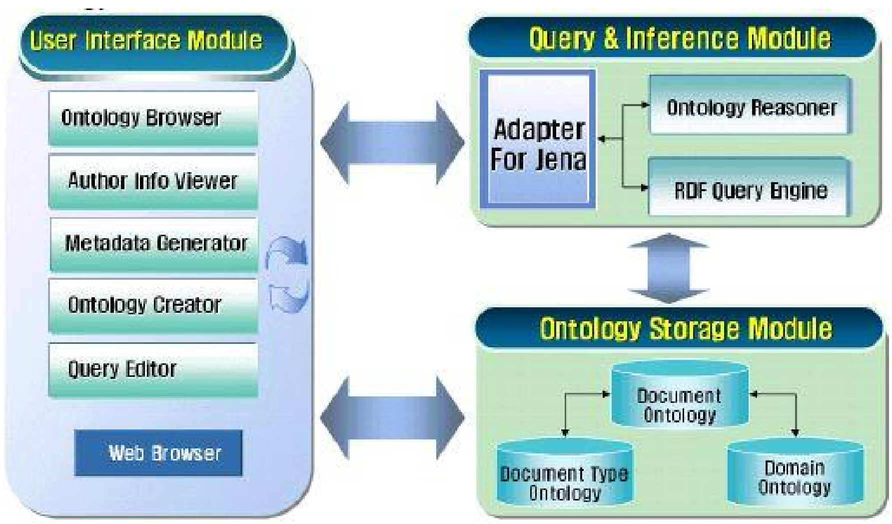

Figure 2.1: Architecture of Ontalk Retrieval System [12]

In a document retrieval system the document is said to be relevant when the user want it and relative to his search for information and if doesn't want the document is non-relevant[12]. Different researches proposed many solutions techniques for retrieval of relevant documents based on user queries[8] Kim and Zhang proposed several factors to rank the retrieved documents and Persin et al proposed an evaluation technique that uses early recognition of documents that are likely to be ranked high. A retrieval model proposed by Robertson Spark- Jones use of ontology information into the query expansion process[13]. Horng, Yeh and Gordon suggested Genetic Algorithms to information retrieval for crossover and mutation operation[14]. Vector space model used as a standard model in information retrieval for automating indexing of documents[15] and wong improved the model by computing term correlations directly from automatic indexing scheme. [14] [15] the text is tokenized into words and vector space model is created. QR factorization and LSI [15] are applied to retrieve the documents which are most relevant to the query. Vector space model, Probabilistic model and Language model improved the efficiency of searching have been attempted by many authors[16].

3 INFORMATION RETRIEVAL MODELLING

3.1 Introduction

The two good reasons for having models of Information retrieval, it provides the means for academic research and serves as a blue print to implement the retrieval system. In the present world, mathematical models are used to in many scientific areas to understand the objectives and reason the behavior of the real world. According to Webster's[1] model is the presentation of the mathematical description of data and inference. Information retrieval model predicts and explains the user will find the relevant information upon the user query. A naive approach of the information retrieval would be scanning the collection of documents in search for the needed information. Linear scanning is appropriate for small scale collection of documents but for large scale collection of documents data structures are built for the speedup of the retrieval process. The three problems that Information retrieval must solve is term weighting, relevance feedback and structured queries. Term weighting is weight of a term which is the value of the importance of the term. The greater the weight, the more value to the term, but term weighting is not a trivial problem. The second problem Information retrieval must solve is relevance feedback, which uses the example of the relevant documents to improve the retrieval of relevant documents. The last problem for the Information retrieval is structured queries in which it defines the relation between the query terms.

The following section will describe the different models for the Information retrieval. The next section presents the Boolean model, strictly speaking the model is a more of model of data retrieval than a model of Information retrieval and it is served as role model for many approaches. The next section presents about different ranked retrieval models and last section explains about term weighting of the user queries.

3.2 Boolean Model

Boolean model is the first model of the Information retrieval and probably the most criticized model of all the models. It was developer around 1960's and it was used as the leading model by many commercial retrieval systems until 1990's. Firstly the model gives the user complete control of the system. If the retrieved document set is too big or too small, it is clearly shows that which operator is used to produce the smaller or bigger sets. Use of wildcard operators and proximity operators makes the Boolean model a powerful full text retrieval model. Besides advantages the Boolean model has many clear disadvantages, this model doesn't provide the ranking of a retrieved documents, which might lead the system to make frustrating decisions.

The main failure for the Boolean model is its inability to rank the retrieved documents, so many of the modern full text retrieval systems use ranking as utmost importance. Boolean model doesn't fit the need for the modern retrieval systems. But many of the ranked retrieval models takes their base from the Boolean model, so this model serves as role model for the ranked retrieval model.

3.3 Ranked Retrieval Model

The Boolean model inability to rank documents is addressed by these models. These models use statistics of number of occurrence of term in the document or index to compute ranking. Automatic query formulation is other key for the ranked retrieval models. This address the difficulties of non – expert users with Boolean operators.

Luhn's was the first person to approach statistically for searching information[1]. He also said that degree of similarity between representations of the prepared document and document collection are used to search the collection.

3.3.1 Vector Space Model

Based on Luhn's approach Salton and McGill suggested a model which has strong theoretical motivation. He considered the query and index representation as vectors in high dimension Euclidean space, where each term is assigned a separate dimension.

The main disadvantage of the vector space model is it doesn't clearly mention the values of the vector component. There are many other problems vector space model must deal with term dependences are not possible to include in the model and calculation of the cosine needs all the vector space values which are not available. In real world, normalized vales and vector product algorithms are used to calculate the vector space model but normalized weights and normalized values must store in separate file which would require a more storage space.

3.3.2 Probabilistic Model

This model is completely against vector space model to use the degree of similarity between query and index representation. This model argues that retrieval system should rank the documents in the collection in order of their probability of relevance.

Probabilistic model is one of the rare retrieval models need not to implement an additional term weighting algorithms. For this reason, it is one of the most influential retrieval models. But unfortunately, the distribution of terms over relevant and non-relevant documents is unavailable in many applications. But there also disadvantages using probabilistic model it defines a partial ranking of documents.

3.4 Term Weighting

All the models of ranked retrieval model require term weighting algorithm before they are implemented. The most important factor of the search retrieval systems is weighting of search terms. The development of weighting terms is implemented from the past 25 years and were experimented in many smart retrieval systems.

3.4.1 Inverse Document Frequency(idf) Weighting

The term df is defined as document frequency which is the number of documents a term occurs in. The terms having a low document frequency are more specific that terms having high frequency. This system will match non-frequent terms as more valuable than one with frequent terms. The weight $log(N/df)$ is the inverse document frequency idf .The ranking algorithm using term weighting is $idf(t) = log\_e(Total\ number\ of\ documents\ /\ Number\ of\ terms\ in\ a\ document\ t)$.

3.4.2 Probabilistic weighting

Probabilistic weighting is used in probabilistic model which suggest a simple term weighting algorithm that suggest binary document weights dk and relevance weights $qk = \log(N/df)$ of the query terms where df is document frequency and N is the total number of documents in the collection. This weighting algorithm turns to be a different term weighting algorithm.

3.4.3 $tf.idf$ Weighting

It is the combination of term frequency tf and inverse document frequencyidf. This algorithm is a breakthrough in term weighting of information retrieval systems. Most of the major smart systems uses this algorithm.

4 METHODOLOGY

This section of study describes the steps involved during the thesis. Initially basic Literature Review was conducted in the field of Information retrieval systems to gain the knowledge of different retrieval systems running in the market. During this study a limited research was identified in this field and there is a scope for lot of research to be done.

Many Challenges were identified for retrieval systems, but there exist many models of Information retrieval that tops in the performance of term weighting and ranked retrieval. But none of the existing models accounts for indexing modelling using both term weighting and ranked retrieval was observed and this research gap was observed to formulate the research questions that were focused as a part of thesis.

After formulating the research questions, there is a need to observe the existing Literature for deep and detailed knowledge of retrieval models. The findings of Literature review is presented in Section 2 as “Related Work”. To gain the deep knowledge of different existing Literatures are considered in different fields like E-Health Records etc.

The results from the Literature Review were used as inputs for conducting the experiment to compare the performance of ordinary language model and proposed model. Based on the results obtained from the proposed model, conclusions were drawn and described in the section 5 of this report. Figure 4.1 shown below, gives the brief overview of the methodology during the study.

4.1 Literature Review

During the Literature Review, studies related to DRS, retrieval models and retrieval algorithms used for DRS were considered. The research papers pertaining to different areas were obtained from different sources like IEEE, Inspec and ScienceDirect Databases.

The search results are performed in 8 different stages using a different set of keywords each time. The keywords are chosen in such a way that to narrow down the field of DRS. Later the keywords are made broader to search for different studies that reported different retrieval models and retrieval algorithms for DRS, how these tools and techniques used for improving the performance of retrieval document. The list of keywords and search results obtained from different database of each search are listed in the Table 4.1

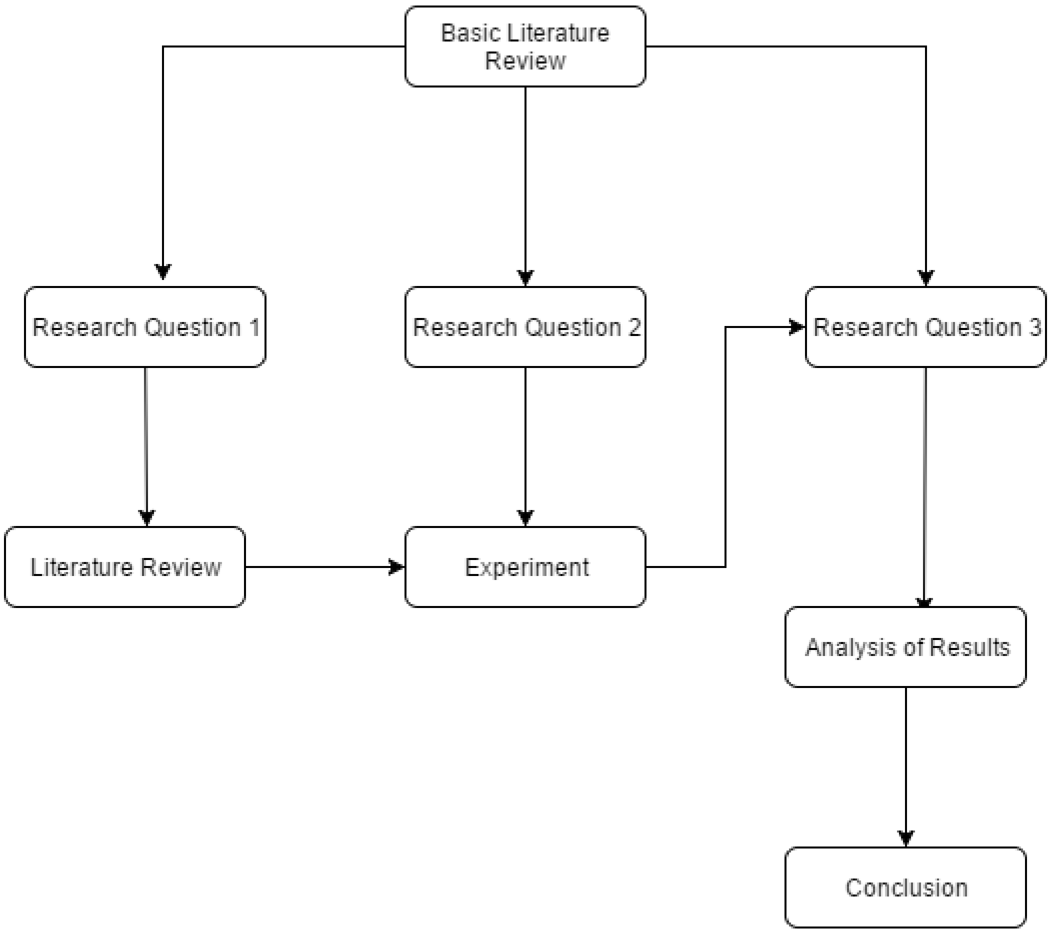

Figure 4.1 Overview of methodology

A total of 1372 Journal and Conference papers were obtained from the Literature Review conducted. These results obtained from the Literature Review were then filtered to only articles with full text available from the database, this exclusion criteria reduced the results to 731 articles. Later these articles are limited to articles that are presented from last 15 years, these causes the search results to 30 articles among which 11 articles were selected which seems to be relevant and interesting to the topic. Finally, 11 articles were selected from the search results, despite obtaining many papers from the literature review, which discussed mostly on E-health records retrieval. Furthermore, the Literature Review was enlarged using Snowballing technique[17] by looking forward and backward into the references of selected papers.

Keywords	Results Obtained		
	Inspec	**IEEE**	**ScienceDirect**
"Electronic Document" AND "Retrieval System"	1195	386	467
"Electronic Document Retrieval System" AND "Information Retrieval Models"	99	30	244
"Electronic Document Retrieval System" AND "Document Retrieval Models"	22	5	138
"Electronic Document Retrieval System" AND "Retrieval Models"	205	56	327
"Electronic Document Retrieval System" AND "Retrieval Algorithms"	214	78	286
"Electronic Document Retrieval System" AND "Information Retrieval Model" AND "Document Retrieval Model"	6	1	102
"Electronic Document Retrieval System" AND "Retrieval Model" AND "Retrieval algorithms"	33	13	233
"Electronic Document Retrieval System" AND "Information retrieval model" AND "document retrieval model" AND "retrieval model" AND "Retrieval algorithms"	1	0	64

Table 4.1 Search Results of Literature Review

4.1.1 Inclusion and Exclusion Criteria

Inclusion Criteria:

- Studies that discussed about DRS, different retrieval models and retrieval algorithms.
- Journal and Conference Articles were considered that were published in the last 15 Years (2000 - 2015).

Exclusion Criteria:

- Articles which were written in English were considered.
- Articles other than Journal and Conference Articles are excluded based on the quality constraint.
- Articles that were not considered, which doesn't have "Full Text"
- of article in the search database. These articles need to be purchased to get full text access to the article, so these articles were excluded.

4.2 Model description

We would like propose a new model for Information retrieval system using combination of term weighting and ranked retrieval. From the proposed model, we infer for each document and rank each document per estimate of producing the query. We estimate the probability of the query using the equation $P(Q|Md)$, given the language model of document d. The probability of term t under the term distribution of the document d, $Pml(t|Md) = t\frac{f(t,d)}{dld}$ where $tf(t,d)$ is the raw frequency of term t in the document d and dld is the total no of the tokens in the document d. We assume that a query terms occur independently on a language model, which leads to a ranking formula $\pi t \epsilon Q \; Pml(t,d)$ for each document. The most practical problem of this model is that probability of zero is not assigned to a document, which means there is a possibility of missing one or more query terms. In addition to this practical consideration we assume$P(t|Md) = 0$, instead we make the assumptions that non-occurring term is possible. This provides practical problems with most reasonable distribution. The other problem with the estimators is, if we get an arbitrary sized sample of data it will be confident to the estimator. But however, we have only a document sized sample of data for the distribution. To solve this problem, we need large amounts of data.

In order to benefit the robustness and minimize the risk we have taken the risk of considering term t in a document d using geometric distribution[18]. Now we use this parameter to calculate $P(Q|Md)$ to estimate the probability of producing the query for a given document.

5 Empirical Results

5.1 Data

The experiment of precision/recall was performed on two different sets of data; the first set was TREC topics 202-250 on TREC disks 2 & 3 and second set on TREC topics 51-100 on TREC disk 1 & 2. We choose these data sets as they are different from each other, the set 51-100 are a list of good terms and set 202-250 having natural language queries consists of one sentence each.

5.2 Implementation

A research prototype retrieval engine was implemented known as Labrador to test the approach. In High throughput retrieval systems, this engine was implemented using topic segmentation work [18]. To achieve the high throughput by keeping much of the index in memory which reduces the waiting time of the device. By compressing, it will store large portions of memory in the index by using a bit level index compression. This approach uses stemming, tokenization and standard term weighting $tf.idf$ in our language modelling approach. Labrador is considered as best research prototype as it works very well but it lacks in error checking and verification which is a failure to meet certain standards. This prototype is not suitable for wide research, but it is performed for these experiments.

5.3 Recall/Precision Experiment

In the table 4.1 and 4.2 we see recall/precision results as well as the average precision values conducted on TREC topics 202-250 and TREC topics 51-100. The first two columns in the table were used to compare for the baseline of the new approach. The baseline result was obtained using term weighting$tf.idf$. The third column represents the percentage change on first and second column.

	tf.idf	LM	%change
Rel:	6550	6550	
Rret:	3145	3479	+ 10.6
Prec			
0.00	0.6349	0.6950	+9.46
0.10	0.5241	0.5910	+12.76
0.20	0.4514	0.5110	+13.20
0.30	0.3761	0.4249	+12.97
0.40	0.2987	0.3387	+13.33
0.50	0.2093	0.2579	+23.22
0.60	0.1357	0.1558	+14.81
0.70	0.0894	0.1197	+33.89

0.80	0.0450	0.0763	+69.55
0.90	0.0089	0.0160	+79.77
1.00	0.0019	0.0057	+200
Avg:	0.2523	0.2901	+14.98
Prec:			
5	0.5367	0.5681	+5.85
10	0.4932	0.5109	+3.58
20	0.4126	0.4623	+12.04
30	0.3810	0.4097	+7.53
40	0.3359	0.3798	+13.06
50	0.2989	0.3187	+6.62
100	0.2086	0.2876	+37.87
200	0.1658	0.2014	+21.47
500	0.1387	0.1632	+17.66
700	0.1089	0.1132	+39.48
1000	0.0635	0.9007	+13.84
RPr:	0.2858	0.3923	+37.26

Table 5.1: Comparison of term weighting $tf.idf$ and Language modelling on TREC topics 202-250

The Language modelling approach achieves a better precision at all levels of recall, mostly at all significant levels. Also, there is an improvement on recall, average precision and R – precision, where precision after R documents where R is equal to number of relevant documents for each query. On the second part of the table there is a significant improvement of recall. The results on the table 5.1 also shows the improvement on the levels of precision at all levels of recall.

	tf.idf	LM	%change
Rel:	10426	11426	
Rret:	7992	8724	+9.15
Prec			
0.00	0.8763	0.9632	+9.91
0.10	0.7156	0.8045	+12.42
0.20	0.5624	0.6456	+21.62
0.30	0.4427	0.5408	+22.15
0.40	0.3439	0.4357	+27.07
0.50	0.2451	0.3681	+50.18
0.60	0.1637	0.2182	+33.29
0.70	0.1169	0.1432	+22.49
0.80	0.0324	0.0301	-7.09
0.90	0.0167	0.0057	-6.58
1.00	0.0008	0.0004	-5.00
Avg:	0.3195	0.377	+18.21
Prec:			
5	0.5367	0.5681	+5.85
10	0.4932	0.5109	+3.58

20	0.4126	0.4623	+12.04
30	0.3810	0.4097	+7.53
40	0.3359	0.3798	+13.06
50	0.2989	0.3187	+6.62
100	0.2086	0.2876	+37.87
200	0.1658	0.2014	+21.47
500	0.1387	0.1632	+17.66
700	0.1089	0.1132	+39.48
1000	0.0635	0.9007	+13.84
RPr:	0.2858	0.3923	+37.26

Table 5.2: Comparison of term $tf.idf$ weighting and Language modelling on TREC topics 51-100.

6 ANALYSIS & DISCUSSION

6.1 Identified Techniques

During the Literature review, different techniques were identified for enhancing the retrieval speed of relevant document are described below.

- Vector Space Model: Documents and query are represented as vector in the term space in which similarity between the two vectors are computed.
- Boolean Model: It is the most adopted model which is a classical information retrieval used by many retrieval systems.
- Probabilistic Retrieval: In this retrieval model the term appeared in the relevant document is computed for each term in the collection. The similarity measure is computed as combination of probabilities of each of the matching term.
- Latent Semantic Indexing: The terms occurred in the document is represented with a term- document matrix. The documents which have same semantics are located close to one another in a multi- dimensional space.
- Language based Modelling: In this approach a document is a good match to a query, if the document model is likely to generate the query which in turn will happen if the document contains query terms.
- Genetic Algorithms: The relevant documents can be find by generating an optimal query. An initial query used with random or estimated weights and new queries are generated based on the modifying the new weights.

Based on the identified techniques the new retrieval model is evaluated by comparing it with some existing models in a controlled environment. Traditionally two distinct problems must be solved by these models are the following:

1. Term weighting
2. Ranking Algorithm

Based on the following problems, two retrieval tasks were designed, these tasks serve to illustrate that language based modeling system perfectly suitable for, to rank the documents from the relevance information, the ability to perform structured queries and ability to show the probability of relevance estimation from the relevant documents. For each of the retrieval tasks, the language model is compared with other traditional models. Ideally we take these traditional models of information retrieval for comparison, Boolean model, vector model and probabilistic model. The Boolean retrieval model does not provide ranking of documents. Of the two-vector space model and probabilistic model, vector space model was introduced in late 1970's and probabilistic model was used to perform TREC experiments in 1990's.

6.2 Improving the Language Model

According to proposed model we should yield a better retrieval performance, simple improving of estimates to the language model will smoothen the estimates of the average probability of low document frequency. In order to do this, we binned the low frequency data by document frequency. This new estimate is incorporated in to our new model and is run on the TREC topics 202-250 and TREC 51-100 and these results are shown in the table 6.1 and 6.2.

	LM	LM2	%change
Rel:	6550	6550	
Rret:	3479	3456	
Prec			
0.00	0.6950	0.7095	+2.08
0.10	0.5910	0.6137	+3.84
0.20	0.5110	0.5394	+5.55
0.30	0.4249	0.4324	+1.76
0.40	0.3387	0.3426	+1.15
0.50	0.2579	0.2590	+0.42
0.60	0.1558	0.1558	+0.00
0.70	0.1197	0.1129	-5.68
0.80	0.0763	0.0937	+22.8
0.90	0.0160	0.0197	+23.12
1.00	0.0057	0.0063	+10.52
Avg:	0.2901	0.2986	+2.93
Prec:			
5	0.5681	0.5926	+4.31
10	0.5109	0.5537	+8.37
20	0.4623	0.5092	+10.14
30	0.4097	0.4563	+11.37
40	0.3798	0.3921	+3.23
50	0.3187	0.3346	+4.98
100	0.2876	0.3019	+4.97
200	0.2014	0.2246	+11.51
500	0.1632	0.1819	+11.45
700	0.1132	0.1186	+4.77
1000	0.9007	0.9009	+22.2
RPr:	0.3923	0.415	+5.78

Table 6.1: Comparison of Ordinary Language model and proposed model tested on TREC queries 202-250.

	LM	LM2	%change
Rel:	11426	11426	
Rret:	8724	8935	+2.41
Prec			
0.00	0.9632	0.9638	+0.06
0.10	0.8045	0.8050	+0.05
0.20	0.6456	0.6490	+0.34
0.30	0.5408	0.5438	+0.30
0.40	0.4357	0.4400	+0.43
0.50	0.3681	0.3680	-0.01
0.60	0.2182	0.2178	-0.04
0.70	0.1432	0.1460	-0.72
0.80	0.0301	0.0319	+0.01
0.90	0.0057	0.0080	+0.02
1.00	0.0004	0.0010	+0.006
Avg:	0.377	0.379	+0.53
Prec:			
5	0.5681	0.5700	+0.019
10	0.5109	0.5219	+0.011
20	0.4623	0.4690	+0.006
30	0.4097	0.4127	+0.003
40	0.3798	0.3850	+0.002
50	0.3187	0.3200	+0.001
100	0.2876	0.2890	+0.001
200	0.2014	0.2023	-0.085
500	0.1632	0.1637	+0.000
700	0.1132	0.1139	+0.000
1000	0.9007	0.9005	-0.000
RPr:	0.3923	0.3957	+0.003

Table 6.2: Comparison of Ordinary Language Model & Proposed model tested on TREC queries 51-100.

These results show a significant change in improvement of precision at all levels of recall and significant improvement at all levels of recall and an average. Our assumption is smaller improvement in the query set is due to longer average length of the query when compared to another query set. It shows that low frequency terms effect the average.

7 CONCLUSION AND FUTURE WORK

It can be concluded that the proposed retrieval model significantly improves the performance than the ordinary language model. The proposed model will provide the effective retrieval system when the following conditions are met language models are the accurate representation of the data, users should have a knowledge of term distribution and understanding of retrieval approach.

Performance of proposed language model was better than the ordinary model performed on two different query sets. These experiments show that simple smoothing techniques provides the better results than the baseline results on both the query sets. The retrieval model was evaluated on two large scale retrieval collections from TREC. The results show that the performance of the proposed language model is better than the ordinary language model.

It is also possible to elaborate that other smoothing techniques such as data transformation can yield better results which is planned to investigate in the future. Query expansion techniques such as local feedback or relevance feedback are used as intuitive ways to generate language model.

REFERENCES

[1] D. Hiemstra, *Using language models for information retrieval*. Taaluitgeverij Neslia Paniculata, 2001.

[2] V. Konstantinou and P. Morse, “Electronic documentation system: using automated hypertext techniques for technical support services,” in *Proceedings of the 10th annual international conference on Systems documentation*, 1992, pp. 1–6.

[3] W. Elgarah, N. Falaleeva, C. C. Saunders, V. Ilie, J. T. Shim, and J. F. Courtney, “Data exchange in interorganizational relationships: review through multiple conceptual lenses,” *ACM SIGMIS Database*, vol. 36, no. 1, pp. 8–29, 2005.

[4] R. H. Sprague Jr., “Electronic document management: challenges and opportunities for information systems managers,” *Manag. Inf. Syst. Q.*, vol. 19, no. 1, pp. 29–49, Mar. 1995.

[5] K. G. Alberto, C. M. Abella, M. G. C. E. Sicat, J. D. Niguidula, and J. M. Caballero, “Compiling Remote Files: Redefining Electronic Document Management System Infrastructure (CReED),” in *Information and Multimedia Technology, 2009. ICIMT’09. International Conference on*, 2009, pp. 347–350.

[6] D. C. Blair, “The challenge of commercial document retrieval, Part I: Major issues, and a framework based on search exhaustivity, determinacy of representation and document collection size,” *Inf. Process. Manag.*, vol. 38, no. 2, pp. 273–291, Mar. 2002.

[7] L. Azzopardi and V. Vinay, “Accessibility in information retrieval,” in *Advances in Information Retrieval*, Springer, 2008, pp. 482–489.

[8] A. Al-Dallal and R. S. Abdulwahab, “Achieving high recall and precision with HTLM documents: An innovation approach in information retrieval,” in *World Congress on Engineering 2011 (WCE 2011), 6-8 July 2011*, 2011, vol. vol.3, pp. 1883–8.

[9] Moon Soo Cha, So Yeon Kim, Jae Hee Ha, Min-June Lee, Young-June Choi, and Kyung-Ah Sohn, “CBDIR: fast and effective content based document information retrieval system,” in *2015 IEEE/ACIS 14th International Conference on Computer and Information Science (ICIS), 28 June-1 July 2015*, 2015, pp. 203–8.

[10] H. Chen, “A machine learning approach to document retrieval: an overview and an experiment,” in *Proceedings of the Twenty-Seventh Hawaii International Conference on System Sciences, 1994*, 1994, vol. 3, pp. 631–640.

[11] R. S. Khokale and M. Atique, “Intelligent Interface for Web Information Retrieval with Document Understanding,” in *Human-Computer Interaction. Applications and Services. 16th International Conference, HCI International 2014, 22-27 June 2014*, 2014, vol. pt. III, pp. 21–31.

[12] B. Mianowska and Ngoc Thanh Nguyen, “A Method of User Modeling and Relevance Simulation in Document Retrieval Systems,” in *Agent and Multi-Agent Systems: Technologies and Applications. 5th KES International Conference, KES-AMSTA 2011, 29 June-1 July 2011*, 2011, pp. 138–47.

[13] J. Bhogal and A. Macfarlane, “Ontology based query expansion with a probabilistic retrieval model,” in *Multidisciplinary Information Retrieval. 6th Information Retrieval Facility Conference, IRFC 2013, 7-9 Oct. 2013*, 2013, pp. 5–16.

[14] A. A. A. Radwan, B. A. A. Latef, A. Mgeid, A. Ali, and O. A. Sadek, *Using Genetic Algorithm to Improve Information Retrieval Systems*. .

[15] N. Jamil, N. A. Jamaludin, N. A. Rahman, and N. Sabari, “Implementation of vector-space online document retrieval system using open source technology,” in *2011 IEEE Conference on Open Systems, 25-28 Sept. 2011*, 2011, pp. 395–9.

[16] VanNhon Do, T. T. Huynh, and TruongAn PhamNguyen, “Semantic representation and search techniques for document retrieval systems,” in *Intelligent Information and Database Systems. 5th Asian Conference, ACIIDS 2013, 18-20 March 2013*, 2013, vol. pt.I, pp. 476–86.

[17] C. Wohlin, “Guidelines for snowballing in systematic literature studies and a replication in software engineering,” in *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering*, 2014, p. 38.

[18] J. M. Ponte and W. B. Croft, “A language modeling approach to information retrieval,” in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, 1998, pp. 275–281.