Sikandar
01-243171-014

A Text Formalization based Approach for Improved Query Response

Masters of Science in Computer Science

Supervisor: Dr.Arif Ur Rahman

Department of Computer Science
Bahria University, Islamabad

September 13, 2019

BAHRIA UNIVERSITY
PAKISTAN

MS-13
Thesis Completion Certificate

It is to certify that the above student's thesis has been completed to my satisfaction and, to my belief, its standard is appropriate for submission for evaluation. I have also conducted plagiarism test of this thesis using HEC prescribed software and found similarity index at ____________ that is within the permissible set by the HEC. for MS/MPhil/PhD.

I have also found the thesis in a format recognized by the BU for MS/MPhil/PhD thesis.

Principle Supervisor's Signature: ____

Principle Supervisor's Name: Dr.Arif Ur Rahman (Assistant Professor)

29 August, 2019

MS-14A
Author's Declaration

I, Sikandar hereby state that my MS thesis titled **"A Text Formalization based Approach for Improved Query Response"** is my own work and has not been submitted previously by me for taking any degree from **"Bahria University, Islamabad"** or anywhere else in the country / world.

At any time if my statement is found to be incorrect even after my Graduate the university has the right to withdraw cancel my MS degree.

SIKANDAR
01-243171-014
September 13, 2019

MS-14B
Plagiarism Undertaking

I, Sikandar solemnly declare that research work presented in the thesis titled

A Text Formalization based Approach for Improved Query Response

is solely my research work with no significant contribution from any other person. Small contribution / help whenever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero tolerance policy of Bahria University and the Higher Education Commission of Pakistan towards plagiarism. Therefore, I as an author of the above titled thesis declare that no portion of my thesis has been plagiarised and any material used is properly referred / cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of MS degree, the university reserves the right to withdraw / revoke my MS degree and HEC and the university has the right to publish my name on HEC / University Website on which name of students who submitted plagiarised thesis are placed.

SIKANDAR
01-243171-014
September 13, 2019

Abstract

Information retrieval (IR) means automatic retrieval of required information into a well-organized form from different sources. Early retrieval methods were ineffectual to beat the scalability and adaptability of natural language text, but with the arrival of internet, the scope and variance of applications that depends on diverse information retrieval and extraction has raised noticeably. As our society is becoming more information dependent, people need a way to easily access the information available for their needs. Applications such as inventory systems, medical decision support systems and institutional databases that keep the records of their day to day activities lead the practitioners to further research. To meet with this need, many approaches have been put into operation to catch the important information from text and make them accessible for decision making but the results are still substandard. In this thesis a new model for rules retrieval is presented. This model is based on Latent Dirichlet Allocation (LDA). The proposed LDA model is applied to a corpus of XML documents yielding topic terms. Using these topic terms as indexing terms brings significant improvement in the results as compared to alternative models.

Acknowledgments

It was an incredible joy working under the Supervision of Dr. Arif Ur Rahman, who gave all that I expected to succeed. His motivation and direction made this Master of Science degree fulfilling. He generally advanced my work and urged me to go out on a limb in my work and to express my thoughts with no confinements. I feel blessed to have been his understudy. I might want to say thanks to Dr. Arif Ur Rahman for the gigantic trust and tolerance he has over me. He always bolstered and guided me in each and every progression.

I might want to thank every one of my companions at Bahria University, Islamabad and my family for the endless help. I might want to extraordinarily thank to my parents for their everlasting affection and support for me to succeed.

SIKANDAR
Bahria University Islamabad, Pakistan

June 2019

*"Listen to your elder's advice, not because they are always right
but because they have more experiences
of being wrong."*

Dr Allama Muhammad Iqbal

Contents

List of Figures

List of Tables

Acronyms and Abbreviations

IE	Information Extraction
IR	Information Retrieval
PR	Passage Retrieval
LDA	Latent Dirichlet Allocation
NER	Name Entity Recognition
TF-IDF	Term Frequency - Inverse Document Frequency
XML	Extensible Markup Language
DTD	Document Type Definition
NLP	Natural Language Processing
DTM	Document Term Matrix
Time ML	Time Markup Language
XSD	XML Schema Definition
QE	Query Engine

Chapter 1

Introduction

Information streams in at a phenomenal pace and should be managed in an auspicious way. The dynamic nature of information making it difficult for human beings to look through and analyze the data manually. Current research in information extraction is aiming to provide appropriate solutions for overcoming these issues by applying advanced techniques to databases and other electronically stored data. As technology is revolutionizing every industry, advance research in information retrieval in social, medical and legal domain also gaining interest.

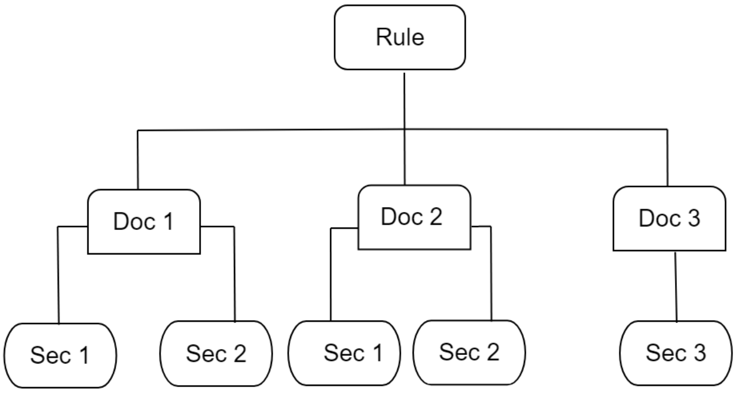


Figure 1.1: Information Retrieval Task

For example the rules of academic institutions can be related to various activities like employee performance, students' performance (grading criteria) and promotion rules etc. These rules are evolve with the passage of time. This evolution may result in various versions of the same rule. The retrieval of relevant (related) rules from various documents is important for analysis for decision-makers. The decision makers may be interested in getting to know the various states of the same rule before proposing any changes. Therefore, the main issue is to retrieve the various versions of the related rules from a rulebook and present to readers in an easy to understand manner. These rules are placed in a rulebook and always pose the challenge of its complex structures and difficult to store it in a more user-friendly way so that it can be manipulated later when required. It is now a challenge for all the related people to compile these rules in such way that it can be referred in no time by going through all the rules regarding a matter. It cannot be done by reading every single line regarding any specific issue. One way of doing it the faster way is to convert the text into XML format and make it available for querying without requiring to learn any specific structure query language that is often the case when documents are encoded in XML format.

1.1 Toy Example

LDA model is based on assumption. Each and every document in a collection share a number of un-known topics. This assumption leads to that documents in a collection carry multiple topics and there may exist multiple latent topics in a single document.

- HS 1: Grading policy of HS is approved by the director.
- HS 2: Director HS called for meeting.
- HS 3: Entrance test for admission to HS to be held on Ist January.
- CS 1: CS program to be launched is subjected to HEC's approval.
- CS 2: Director briefed the faculty in the meeting.
- SCM 1: SCM and CS are to follow the same grading policy.
- SCM 2: Entrance test to be conducted for Admission in SCM at BUIC.
- SCM 3: SCM program at BUIC is approved by HEC.

In this example there are 8 different documents with 27 unique terms (after removing stop words) from three different disciplines: Health Sciences, Computer Science and Supply Chain Management. These documents are labeled by the first two leading characters of related discipline: HS, CS and SCM. For example, HS 2 is the second document in Health Sciences and it is about meeting. Here each document blends different topics in different proportion. The idea is to calculate the probability of words in topics to accurately assign words to topics.

To reflect documents in terms of topic the proposed technique used β matrix generated by LDA. This distinctive representation of documents in the proposed model makes it possible to calculate the query resemblance with documents while retrieving and ranking documents. The aftereffects of our computational analysis demonstrate that the proposed LDA technique that is basically a topic modeling technique is practical in domain of information retrieval. As LDA is often used for topic modeling it tends to be utilized in information systems without naming and labeling data. e.g. web crawlers that crawls the World Wide Web, delivers data at the latest with the list of similar documents to a user query.

1.2 Motivation and Problem Description

We are living in the age, where it is extremely necessary for business firms to keep record of their day to day operations including electronic and paper data. At the time of any kind of controversy, opponents are enforced to go through all the data and retrieved documents concerning the problem. The primitive barrier while retrieving relevant documents is a large volume of data that prevent access to sensible data. Going through this data is time consuming and irritating task. Institutions change their rules and regulations with the passage of time. This modification in rules and regulations may result in various versions of the same rule and stored in different documents. When decision makers want to analyze all the previous versions before proposing any change, it is difficult to go through all the versions in a flash. It cannot be done manually by reading every single line.

1.3 Research Contribution

The existence of schema allows numerous optimization operations for working with XML documents. From the existing literature of XML schema, it is concluded that for many domains there do not exist a schema or schema that is not acceptable. It is very important in the existing scenario to generate XML schema for migration of rules into XML format. The presence of schema not only assists in optimizing and easing the search process, but also facilitate semantic retrieval based on keywords. In this particular scenario defining of the schema was one of the essential tasks to be accomplished. That is why we have defined our own schema for rules in a rulebook to make searching optimized by giving users more accurate results.

Our another contribution is the algorithm. Once the raw text in documents is initially processed and migrated to XML format, our proposed LDA algorithm is run at various levels to extract topics from the rules. To retrieve documents of interest the LDA generated topic terms are used as index terms.

1.4 Thesis Organization

The remaining chapters discuss the proposed model and are sorted as follows: Chapter 2 outlines different Information Extraction, Information Retrieval techniques and related concepts. Chapter 3 presents the applied methodologies and tools to achieve the results of the thesis. Chapter 4 describe the performance matrices used, demonstrating the effectiveness of our proposed algorithm. Chapter 4 is followed by a brief conclusions chapter.

Chapter 2

Literature Review

2.1 Introduction

Automatically and easily accessing rules in a rulebook is a challenge due to increase number of documents in electronic format. The information in rules are important to professionals and computer applications. These applications store the data and help the professionals in retrieving relevant information performing difficult clerical work that requires accuracy and speed. Rules in a rulebook include natural language text with some kind of obligation, permission and prohibition in it. Therefore, identification of these kind of words in document plays a significant role in extracting the relevant material. In this chapter various IE and IR approaches are presented.

2.2 Latent Dirichlet Allocation

As our society is data oriented and large number of devices connected results in billions of terabytes data. It is realized that online information increments with high pace and human can't peruse this at speed [1]. A search engine finds the list of documents related to search query, but consuming all these documents is not possible for humans. A useful way is needed to provide general view about a document. The topic modeling is an approach to extract the theme of a document. The topic modeling approach learns in an un-supervised manner that explore the hidden patterns (topics) of a document. A document contains some topics and topic is composed of group of words, topic model learns these topics using machine learning algorithms [2]. Once the topics are learned then it can be used for information retrieval by finding similarity and relationship between query terms and topic terms.

At present research community acknowledges it to be one the fascinated concern of research and numerous approaches have been presented among which LDA is a popular approach towards information retrieval. Some other topic modeling techniques are also derived from Latent Dirichlet Allocation.

2.2.1 Dynamic Topic Model

It is an extension of LDA. In dynamic topic modeling the documents in corpus are sequentially organized and captures the evolutions of topics over time [3]. This type of topic modeling focus on categorical data as compare to time series data which focuses on continuous data. Dynamic Topic Modeling unlike LDA is sensitive to both grouping of documents observable in corpus and the sequence of words that appears in a document. This order of documents plays a vital role in topic modeling. Similar like LDA each document in dynamic topic modeling is perceived as distribution of hidden concepts and each concept is considered to be a set of words.

2.2.2 Hierarchical Topic Model

Hierarchical topic modeling is another type of un-supervised topic modeling which learns topic hierarchies form the data [4]. Hierarchical model addresses the problem of not depicting relations among topics in LDA. Learning hierarchies of topics are useful for finding relationship among topics. In this model the topic hierarchies are represented by nested Chinese restaurant process. Every node in the hierarchy tree is considered a topic and is explored as cluster of words. Hence hierarchical topic modeling describes the corpus organization more accurately and learning of topic hierarchies helps in enhancing the prediction performance.

2.2.3 Correlated Topic Model

Correlated topic model like Hierarchical topic modeling models the correlation among topics. It makes an independence assumption among topics. CTM unlike LDA provides a graphical representation of relationship between topics. This approach offers increasingly reasonable mechanism of modeling structure of topics where the contents of two different latent topics may corresponds with each other. e.g. a document about “semantic web” is bound to be likewise about “information retrieval” than about the topic “genetics”. It is the logistic normal function in CTM that differs it form LDA [5].

2.2.4 Entity Topic Model

Topic models are un-supervised and sometimes these models do not discover topics that are truly understandable due to lack of coherence. To resolve this specific problem few knowledge-based models have been initiated that make use of existing domain knowledge to enhance the topic coherence. To increase topic coherence vast number of background knowledge represented as ontologies can be associated with topic models [6]. This ontological representation of information leads to interpret the real-world entities and relationship among them. EntLDA is a knowledge-based topic model that has integrated the entity base ontology with Latent Dirichlet Allocation to discover topics that are truly understandable. The knowledge graph and LDA were used to extract the topics. In the model entities relationship in topics were established with the semantic graph

2.3 Information Retrieval Approaches

Information Retrieval is important in each and every field and has given a lot of attention to help the practitioners in accessing high quality information with in no time. There are many libraries for medicine and marketing but law has received less attention compared to other domains. To work with legal and regulatory text a package named as LexNLP providing both tools and data for researchers [7]. The goal of LexNLP was to apply machine learning with NLP to offer productive tools for analysis of statutes, contracts, bids and wills etc. IR and NLP are the two different research domains but still few researchers employing NLP techniques in information retrieval domain. Tokenization, stemming, parts of speech tagging and few others are NLP approaches that are applicable to information retrieval. Low-level approaches like stemming and stop word removal bring improvements while other higher-level approaches like chunking and parsing cause decrease in accuracy while experimenting it on a large collection [8]. NLP techniques need to be optimized for IR to produce fruitful outcome. Following are some of the IR approaches.

2.3.1 NLP Based Approaches

A combine framework for rules extraction, syntax based and logic-based extraction is introduced in [9]. They used Stanford parser for sentence tagging and WordNet to handle the unevenness of natural language text and exploit Boxer Framework for finding dependencies between chunks of text in logic-based rule extraction. After text extracted from source documents it follows two branches upper and lower. Different technique is applied at each branch. In lower branch the extracted rules were analyzed by upper branch using combinatory categorical grammar NLP tool.

To overcome the problem of information extraction one of the solutions is to annotate the text. In case of extraction from collection of documents where the nature of the text is un-strucutred the annotation could be beneficial. System Tag Me that yield structure data of un-structured text like search engine snippets is presented in [10]. This system uses a sequence of terms known as anchor which was used to annotate page and create set of pages that are linked together.

An architecture for medical decision support system MDDS that support sharing of patient health privacy data meeting with medical laws with no or less human intervention is presented in [11]. The spotlight in current scenario was formalization of lawful content into logical rules for extraction. The logical rules in MDDS were further used for decision making for granting or denying access to patient's medical record obeying the medical laws. This system analyzes the request from different entities and read from logical rules to deny or grant access to patient data.

Implicit information in legal text can be made explicit with gaining knowledge of discourse patterns in legal documents [12]. Also describe the different types of legal documents e.g. legal proof (deeds, wills and contracts), judicial proceedings (warrants, court decision) and statute laws that are followed by citizens. All the documents exist different information that need to be identified at the time of information extraction.

Besides of the fact that laws are evolving with great speed and has certain limitations so memorizing and applying these laws for human like a computer is impossible. Laws have mostly complex grammatical structure therefore storing and understanding of such complicated structure text is not an easy task however by representing these laws and regularization in a formalized style the above problem can be solved [13]. An approach is presented that is based on NLU and CTL to achieve formal representation of legal text after which computers could better understand legal text and serve people better than before.

Making machines intelligent is the need of today's world which will reduce the burden of humans, tackling many complex tasks manually. In legal domain there is also a need for such systems that can read and understand natural language text e.g. contracts, licenses and agreements etc. For this one need to translate the documents into a format that is machine readable that allow automated processing and verification of such text but keep in mind these kinds of transformation also require a intense reliability [14]. Active learning may overcome the problem more smoothly rather than conventional machine learning approaches where the most certain examples were hand-tagged in contrast with frequent approach of annotating examples by annotators about which the classifier were uncertain about.

The present information extraction techniques have the potential of effectively extracting entities such as person, company name and location but how ever this is not always the case, there is some content stored of unstructured nature where entities have the relationship between them and include much more information rather than simple entities E.g. news and articles. Current research on developing such systems that could make the information more explicit than before by Time Markup language [15]. TimeML that is based on extension markup language is particularly designed for annotating text with tags making implicit information explicit. TimeML have three basic tags. Following are some basic tag types used by the system. TIMEX labels are utilized for annotation of events and its time of occurrence. To explain occasion articulations, EVENT labels were utilized giving hook to relate them to different occasions and times and TLINK labels show the fleeting relations that hold among times and occasions.

Information extraction systems use different techniques and approach helping practitioners [16]. In traditional IR systems a document is represented by keywords. These keywords are known as index terms, where each index is assigned a certain weight that measures the closeness among the term and document. Knowledge representation systems are more up to date archetype making use of knowledge characterization mechanism, finding similarity between cases. The similarity between cases in obtained through syntactical similarity or semantic attributes.

Previously legal text was annotated partially automatic for easy retrieval but the gigantic amount of documents made it necessary to find a way to annotate the text in legal documents automatically. The automatic annotation of legal text has been an area of interest for researchers specially in document classification and information retrieval. They have applied numerous machine learning strategies but yet these are insufficient. Some advance documents indexing and retrieval require text content to be semantically represented more enriched. Designing an NLP-based system called SALEM with the aim to automatically assign a tag to legal text in law paragraphs by classifying law paragraphs and to extract relevant text fragments is proposed in [17].

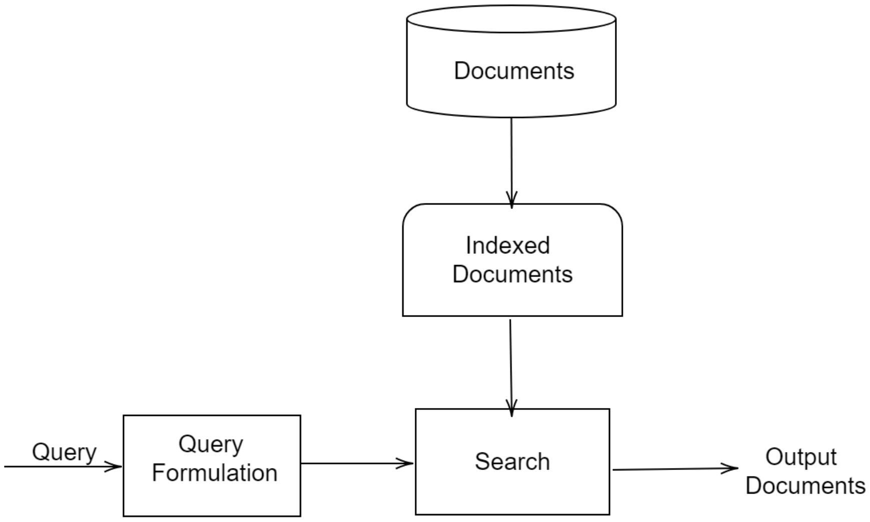


Figure 2.1: NLP based Retrieval

2.3.2 XML Based Approaches

Information retrieval from XML documents where the layout of each document is different from every other document requires some magical skills for finding similarity and ranking. A way to deal with most reasonable solution to client search query is presented in [18]. The problem of retrieving information from XML documents against a user query was reduced to un-ordered tree inclusion problem. The problem was extended to an optimization problem when cost model is applied which was proved to be enough to determine how similar were the XML parts to user query. Their most significant contribution was the algorithm that finds all the matches and ranking was based on its similarity to the query. It is demonstrated that issues in automatic XML markup of documents at micro and macro level that will help the research community to transform the data into XML format [19]. Once all the data is transformed it will be then easy to retrieve and or extract relevant material supporting user query.

XML assumes a significant job as it is a standard language to represent structure data of traditional HTML web. There are several knowledge management repositories that store web-based data and documents in XML format. It is possible to extract knowledge if there exist formal representation of ontology of semantics about data. This could be done when ontology is applied to XML data that is not represented explicitly in the repository [20]. Ontology integration with XML lead us not only to extract knowledge but also sharing of organizational data over the semantic thus it can be used both internal and external to organization inferencing knowledge and supporting data sharing.

The abundance of data is directly related to the quality of relevant data. Large data to be search the results might be affected. In this scenario filtration is needed to include relevant data and exclude irrelevant data based on a define criteria. Filters are the tools that can be utilized to achieve contents that are more suitable to user's need. The filtration process helps in extracting useful information and omit the meaningless information which are not necessary to be retrieved to users. Only extracting relevant information is more productive compared to conventional search and retrieve. XML filtering language for information filtration process with few advantages achieving economical trade-off between simplicity and power of use [21].

In IR system retrieval may not be the primary problem but finding relevant material using any scoring formula or ranking algorithm is of key importance. The indexing of documents accomplishing a prime task in ranking the relevant material. A method of conceptualization, a re-weighting method for XML documents which use the hierarchical ordering of XML elements is presented by [22]. XML documents are comprised of elements known as root elements parent elements and child elements. In the proposed architecture each element is assigned a number as its index number. In the previous approaches the ancestor's weight was not considered in weighting that element.

Many languages such as Document Type Definition, Document Content Description and XML schema have been proposed to generate schemas for XML contents. XML schema of documents offer several services like illustrating the formation of XML elements, XML documents transformation to other types of data and query pruning and rewriting but XML schema is not often necessary like some web documents do not have schema and cannot leverage the schema. Some automatic tools are also in the market but their performance is not up to the mark generating poor quality schema and consuming lot of time. The author presented their own approach for schema extraction from XML documents, designed a content model for conciseness and applied heuristic rules to achieve accuracy. [23].

As earlier mention XML provide basis for data transmission and data integration it can also be applied to many other fields such as databases and web page releasing etc. Increase in quantity of XML documents, XML retrieval is also an area of interest for researchers. To fulfill this need languages like XQeury and Xpath are introduced for XML retrieval. The problem is that these languages have certain limitations as these languages have complex structure require users to be familiar with languages as well as document structure. Experiments focused on XML full text retrieval based on keywords similar to HTML text proposing inverted-file index retrieval method [24]. The proposed method has the position information as well as XML element's route information.

XML sources are explored by the XML data processing system from the database point of view. Fulfilling the users need these query results are boxed as set of XML elements. In XML information system finding relevant documents the results are not sufficient. Some extra effort is required to accomplish this task. In order to encourage the assessment of Boolean Queries inverted file index is integrated with path index [25]. The first processing step was to find-out what terms to be indexed. They have made use of normalization algorithm and stemming algorithm to find-out the index terms and term distribution to generate weight facilitating the ranking process. The proposed method has the potential to support large XML document collection.

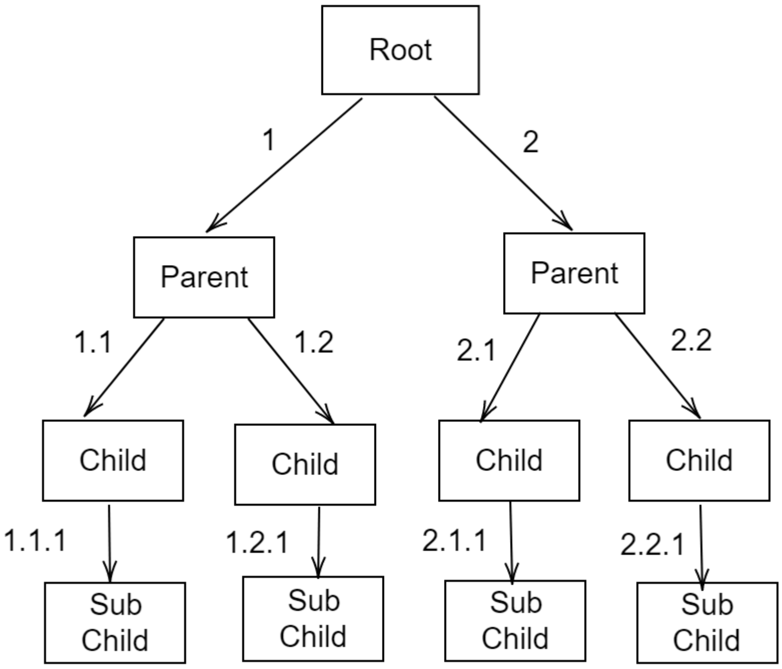


Figure 2.2: XML Retrieval

2.3.3 Passage Retrieval

To generate results against a user query it could be beneficial to divide the documents into different sections, parts and passages. In information retrieval domain where there is a huge corpus the retrieved information may not meet the needs of the user, producing junk and misses important documents that could benefit users. An approach for retrieving only relevant passages rather than whole documents is described in [26]. The text inside documents are assigned score or weighted which are then combined into fragments, answering a question or fulfilling the needs of user query.

To search the whole document against a user query it is now more common to retrieve only relevant passages. The document's length may affect the searching process as there may exist a large and irrelevant text. In these cases, passage retrieval could be beneficial in response to a user query offering efficient and effective retrieval. A method of dividing text into field coherent passages is showed in [27]. To show how topics swell, reduced and shifts sentence after sentence they have used field associated words and phrases concluding that the proposed approach not only extract relevant passages but effectively represent the semantic information of text.

Normally documents are un-structured or semi-structured having different sections and sub-sections exhibiting text of different nature. The retrieval of text from different sections which contains information is essential in fulfilling the users query. The user may be interested in only knowing few things and retrieving text as per the user's interest [28]. The proposed survey narrating and categorizing different extraction method based on input data that may be structured, un-structured or semi-structured.

Human computer interaction is also one of the information retrieval discipline in which human write a query and computer responds accordingly known as short text conversation. To get a reasonable answer from computer is one of the most challenging tasks. Many researches aimed to overcome the problem in recent decades but the progress was unsatisfactory. One of the reasons was absence of real conversation data. A retrieval-based STC to make the system able to generate reasonable response is proposed in [29]. Their experiments with retrieval-based STC have shown significant improvements leveraging huge amount of data available on social media.

Despite of IE and IR Question Answering is also one of the domain researchers have focused on. It is a bit different from IE and IR. IR systems retrieve relevant documents relate to a query and IE aiming at extracting useful information form un-structured text while Question Answering is the integration of both IE and IR answering specific questions that are required in the context. All of the above have some different and some similar modules [30].

Nowadays more and more people are gaining interest towards online shopping including articles and magazine. To buy an article or magazine of your interest one way is to skim through the magazine abstract. Creating abstract by professional is time consuming task therefore there is need for automatic abstract creation that have the potential to attract the reader or buyer [31]. Previous work presented and implemented a system for automatic creation of abstract and is consist of three steps. (1) Text Analysis that is responsible for identifying potential text content. (2) Selection and generalization of text that is relevant for inclusion in the abstract. (3) Text generation that responsible for editing and rewriting of text of abstract.

Information Extraction nowadays also part of financial organizations. Many organizations make use financial data available on internet to have a look at the current status of competitors and to predict the company's strong performance in future [32]. But the problem is that this data is not analyzed. IE make companies and their stakeholders to analyzed the data and answer questions. Information Extraction can be beneficial in media analysis, market segmentation and creating new management policies.

2.3.4 Ranking Techniques

Ranking the results generated against a user query is also a challenging task after finding the relevant documents, passages and sentences that depends on the situation. The researchers are also keen to write algorithms that rank outcomes dependent on its comparability to the query. Only extracting and or retrieving relevant material as a result of a query plays a vital role at the time of analysis once the all the relevant material is retrieved. There are different approaches naming only few e.g. query-term matching, document matching etc. The author has also proposed a framework to extract answer bearing passage, estimating its quality which was later integrated into a ranking model [33]. The quality estimation process was used as an evidence providing a strong indication of document relevance.

In IR-systems to find how close are the query terms to document index terms and how often these words appear in documents are exploited. The issue in this kind of retrieval is that documents having large size have higher chance of retrieval due to increased frequency of terms that match with query terms. To overcome this problem the best solution is to measure the similarity of sentences and passages the makeup the document. In this way document size will not affect the query performance as only relevant documents will be retrieved. A model based on two different modules:(1) Indexation module and (2) document-extracting module is presented in [34]. Indexation module was responsible to assist the document-extracting module by generating dictionaries of terms appear in the archive with its recurrence and position in the archive. Document-Extracting Module aimed at extracting documents based on user's query similarity, queries terms were sorted according to its location and frequency in the document. Query terms found in documents were retrieved and were ranked based on overall similarity of documents to query terms.

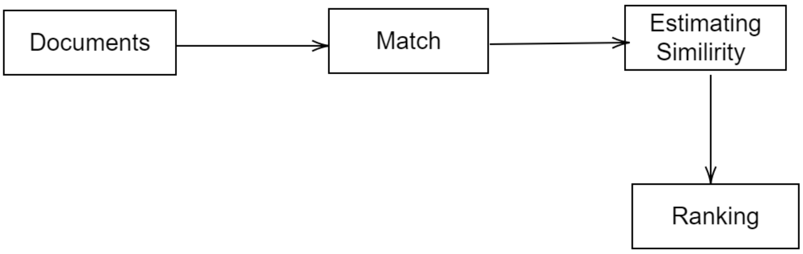
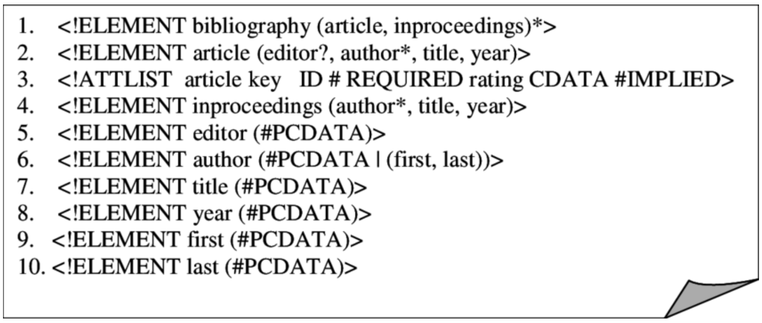


Figure 2.3: Results Ranking

2.4 Existing Standard DTD's

The structure of an XML is defined by Document type Definition. XML contents are organized in specific structure including XML elements and attributes. The grammar of each XML document is defined by this DTD. An XML parser checks the grammar of XML document to know whether it obeys the grammatical rules defined by DTD. Validity and Vocabulary of XML documents is also checked by DTD against grammatical rules possess by XML language. A DTD is used to form a well-structured XML document. DTD is widely used but have limited support for integrity constraints. A single DTD can be used in many XML files. Figure 2.4 present an example of a DTD.



```
<!ELEMENT bibliography (article, inproceedings)*>
<!ELEMENT article (editor?, author*, title, year)>
<!ATTLIST  article key   ID # REQUIRED rating CDATA #IMPLIED>
<!ELEMENT inproceedings (author*, title, year)>
<!ELEMENT editor (#PCDATA)>
<!ELEMENT author (#PCDATA | (first, last))>
<!ELEMENT title (#PCDATA)>
<!ELEMENT year (#PCDATA)>
<!ELEMENT first (#PCDATA)>
<!ELEMENT last (#PCDATA)>
```

Figure 2.4: XML DTD

2.5 Type of XML DTDs

2.5.1 MathML

Mathematical Markup Language (MathML) is consists of XML tags resembling HTML tags. In MathML these tags are used for encoding of math contents in terms of both presentation and semantics [35]. It purpose was to make mathematical equation meaningful to different applications. It marks-up mathematical expressions to support the use of mathematical expressions in HTML or other Web-pages. MathML could be used and beneficial where there is need for to encode mathematical notations. It offers high quality visual display format.

2.5.2 News Industry Text Format

It is an open standard develop by IPTC. News Industry Text Format (NITF) exploit XML to encode the contents and deals with the structure of news stories. NITF documents are far more useful and searchable as compare to HTML pages because meta-data is applied through news content [36]. The publishers can translate these documents into any other format they wish. For moving news articles on the internet NITF opened a gateway by marking-up these articles.

2.5.3 Text Encoding Initiative

Text Encoding Initiative (TEI) purpose is to provide guidelines for creation and management of any type of data in digital form used by researchers in digital humanities [37]. It can be applied to any data such as sound or video but its primary focus is on text data. To present text for online research TEI guidelines are widely followed by individual scholars, museums, publishers and libraries. It uses a computer markup language known as XML to digitally describe a text in humanities.

2.5.4 Chemistry Markup Language

Chemistry Markup Language (CML) is an XML based markup language to manage molecular information. It defines schema for elements and attributes that can appear in CML document [38]. CML tackle the problem of sharing chemical information over internet. It is also similar to HTML tags but with some chemistry tied and it requires to be strictly adhered to SGML. New terms or glossaries can be added without requiring to revise the language and modify the software.

The extensible Markup Language (XML) innovations have been promising to give a typical standard component to interoperable intermixing of various IT forms and have been increasing uncommon acknowledgment from the fundamental to the most confounded business and logical procedures. Moreover, a wide range of areas, content providers and associations have been distributing and exchanging information over the Internet by the use of XML. Starting late, Web administrations dependent on XML innovations have been developing as the accepted system for trading organized data among associations and applications. Due to its adaptable nature and simplicity of execution, XML serves very well as a self-reliant transport format. There are diverse DTDs accessible for various domains but these DTDs are domain specific and cannot be used in any other domain and most importantly none of them is supportive to our work. Well-organized XML schema in different domains is required for efficient implementation of XML. In this perspective structure of XML plays a vital role in development of software and ought to be estimated for effortlessness of maintainability. The maintainability is considered to be one of the most important factors that affect management and quality of software project. For generating XML documents XML schemas and DTDs define by W3C are well-known schema languages. The earlier mentioned DTDs in this chapter are available for different domains e.g. mathematics, medical and for news articles etc, but none of them is supportive to our work. In this study we have tried to develop our own XML schema to make possible efficient retrieval of information from various XML markups.

2.6 Summary

Earlier in this chapter few extraction methods and comparison of several ranking algorithms were discussed to extract relevant data. Data is the collection and representation of facts and figures that is organized in understandable manner and globally accepted. These facts and figures represent an individual's views and perspectives. Data is the basic unit of communication that facilitates us in acquiring and accessing more and more knowledge about a subject that we are interested in. Capturing data of interest is the main challenge encountered by users. It is nowadays in active area of research to capture small part of the data or extracting relevant passages that has led the researchers to introduce searching techniques based on key terms and pattern matching. One of the main issues in IR systems is that they are not aware of the context. For many researchers nowadays it is a center of attention to develop and design IR systems that are fully context aware and to make those systems able to respond differently in different context using the user's behavioral data [39]. All of the aforementioned approaches can be beneficial to design context aware IR systems. The automatic processing of documents plays a significant role in the effectiveness of information retrieval system. Pointing out the existed issues in the process that causes hurdles to meet the required needs other common issue identified was the correct representation of information. The existing approaches have some kind of limitations not using the relationship description of linguistic information and document context. This problem can be unfolded by using two kind of information, linguistic information of documents and semantic information [40]. The structure complexity of databases and finding out semantic relationship between data stored in databases, utilizing knowledge representation techniques and interactive query generation with help of ontology specially stressing on enhancing the interface between data and search queries could help the practitioners to beat these difficulties by bringing the outcomes close to users search requirement [41].

Chapter 3

Methodology

This chapter presents the proposed methodology for achieving our research goals. In subsequent sections some theoretical and practical research activities involved in our research are also illustrated. As discussed earlier the primary focus of our research is to design an approach that can retrieve information on the fly by taking an advantage of IE and IR techniques.

The proposed scheme in this study is composed of the following steps. It includes a pre-processing step that is followed by Query Engine functionalities. In first step documents from corpus are compiled, the compiled documents are then converted into an XML format. In the next step a query engine implements query functionalities against the corpus. In this step Query Engine retrieved all relevant rules from a rulebook including sections, subsections, bullets and enumerations. In the last step the rules extracted from rulebook are dispatched to the end-user through user input/output Interface. Figure 3.1 presents the basic framework.

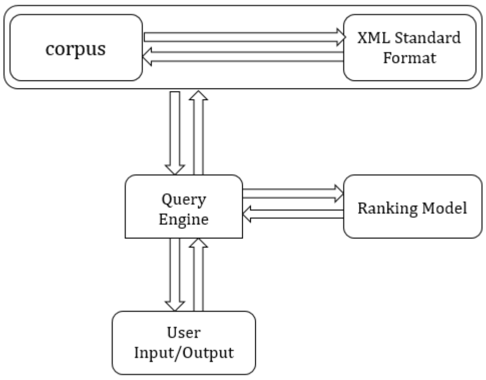


Figure 3.1: Framework

3.1 Pre-processing

Pre-processing is an essential task and to be perform initially. Its purpose is to enhance the text in such a way that retrieval algorithm will be able to retrieve only relevant information from set of documents. Learning algorithms are unable to deal with the raw structure of text in documents, therefore the text is pre-processed and transformed into a form that algorithms are able to practise with. Pre-processing prepares the data makes the analysis feasible and improved the effectiveness of the results. Some of the well-known steps in pre-processing are tokenization, stop words removal, stemming and lemmatization which are briefly discussed in following sections.

3.1.1 Tokenization

It is a critical task that split the text into segments called tokens. Tokens can be words or phrases. First the documents are segmented into sentences later on sentences are split into tokens. In tokenization process some characters are removed e.g. brackets, periods, white spaces and quotation symbols etc. The significance of tokenization is that it helps to reduce complex mathematics involved in IR up to certain degree. Various NLP libraries and tools are used for tokenization defined in python and java language. Some of the well-known libraries are OpenNLP and Stanford CoreNLP [42] and are widely used in NLP and IR.

Input: Hard work is key to success.
Output: “hard”,“work”,“is”,“key”,“to”,“success”.

3.1.2 Stop Word Removal

Stop words removal is an important step in text pre-processing and is widely used in natural language processing, information retrieval and text processing applications. It plays a vital role in text categorization and sentiment analysis which are necessary operation to be performed to find similarity between documents. Stop words are language dependent and its purpose of removal is to filter out useless data. Stop words are commonly used words and appears frequently in text that are often not necessary. Stop words often include preposition and pronouns to which search engine are programmed to ignore both at the time of indexing and retrieving against a user search query [43].

Input: Knowledge is power.
Output: “Knowledge”, “power”.

Removing of stop words from corpus helps in reducing the dimensions that is important to increase both the speed and accuracy. One of the important aspects of stop words are the removal of words that are frequently occur in in each sentence. Another aspect is to identify the words that do not carry some valuable information but occur frequently. There are different libraries available for stop words removal but the procedure is similar across all libraries.

3.1.3 Stemming

In stemming process, the tokens created at the time of tokenization are now reduce to root form. For example, the words "fishing", "fished" and "fisher" are stemmed to root word "fish". Stemming is important in NLP and a part of linguistic studies in morphology in information extraction and information retrieval. More results are returned when more form of words is recognized and searched. Stemming is associated with Query Engines to boost algorithm's accuracy and recall. Modern IR systems use some linguistic tools to increase the productiveness of query search results. Stemming overcomes the problem of complex morphology where single word have hundreds or thousands of variants.

There are two types of errors that occurs during stemming. Over-stemming and Under-stemming. Over-stemming is when two different words are stemmed to same root but should not have been and need to stem to their own root word. Under-stemming is when two words should stem to the same root but actually stemmed to different root words. There many stemming algorithms exist like porter stemmer [44] , Loins Stemmer [45] , and Snowball stemmer [46].

3.1.4 Word Dictionary

It is a collection of unique terms in a corpus. It is also known as vocabulary. Tokens N which is the total number of words in a document come from vocabulary V. E.g. Shakespeare corpus produces tokens N=884,647 from vocabulary of V=29,066.

3.2 Document Term Matrix

It is a matrix representation of documents along with terms. Rows represents documents in a collection and columns corresponds to terms or words in a collection. DTM shows existence of each unique word and its count in each document. A document term matrix represents the documents in a corpus numerically. This numerical representation allows further processing or analytical investigation and is commonly used in text mining and analytics for ranking, categorization and document similarity. There are many ways of representing documents as a matrix among which Tf-idf and bag of words are well-known.

3.2.1 Bag of Words

Bag of words create a table where each document is represented as a row and each column is represented as unique term. BoW only consider the frequency of terms in document irrespective of its grammatical structure. LDA exploit bag of words representation of documents. Let us consider the following example.

Doc-1 I love cats.
Doc-2 I hate cats and dogs.
Doc-3 Dogs love to swim

Table 3.1: Bag of Words

	I	Love	cats	hate	and	dogs	to	swim
Doc-1	1	1	1	0	0	0	0	0
Doc-2	1	0	1	1	1	1	0	0
Doc-3	0	1	0	0	0	1	1	1

In table 3.1 the '1' represents presence of the term in a document and '0' represents absence of the term in that document. This representation of terms in documents can be used to calculate tf-idf.

3.2.2 TF-IDF

BoW is sufficient to convert raw text into numerical form but in order to identify signature words in documents there is need for better transformation. In Tf-idf a word is considered important if it regularly appears in document. In Tf-idf word count appears in a document is replaced by a score that is dot product of Tf and idf.

$$Tf-idf_{t,d} = (1+\log \mathrm{Tf}_{t,d}) \cdot \log \frac{N}{\mathrm{df}_t}$$

3.3 Latent Dirichlet Allocation

Documents are described by identifying the topics in it. This process is known as topic modeling. Topics are emerged from words at the time of topic modeling process. There are many topic modeling techniques like vector-based techniques and probabilistic techniques. Latent Dirichlet allocation (LDA) is popular one. LDA is a generative probabilistic model which assumes topics from set of documents and each topic in a document is represented by set of words [47]. To assign words correctly to topics LDA model is used in the proposed model with certain parameters as input, counting the probability of words in topics and probability of topics in documents with number of iterations until all words are correctly assigned to topics and topics to documents.

In information retrieval we often have collection of documents such as news articles and blogs. These articles and blogs are a mixture of topics containing hundreds or thousands of words. In order to retrieve topics of interest a topic modeling technique is preferred to retrieve topics of interests from a large collection of documents [48]. So, one can say that the intuition behind LDA is to uncover the main theme that is hidden in the data.

Though LDA is topic modeling approach but it could be beneficial in information retrieval. Input to LDA is a document term matrix (DTM) and generate output that is consist of two tables known as topic term matrix and document topic matrix. The first one describes the probability of words in a particular topic and later one describes the probability of topics in a particular document. After LDA is trained and DTM is decomposed into topic term matrix and document term matrix after then it is easy to apply cosine similarity measure to evaluate similarity between query terms and documents.

3.4 Indexing

Indexing is the process of tagging or associating information with a file so that later on it can be search and retrieved. Index information is integrated in a database, record or document management system which serve the purpose for users to locate the documents [49]. Indexing make the information searchable that users will later use to find only relevant information from collection of indexed documents. There are two types of indexing approaches used in the proposed model. Once the documents are pre-processed the word vectors created were used as indexing terms. The goal of this kind of indexing is to recognize the semantics in documents. This indexing is used by Query Engine to access, filter and deliver results to user satisfying his or her need. The other type is the topic terms generated by LDA to recognize document's content e.g. the main topics in documents and the basic units (indexing units) to represent them.

3.5 Query Engine

Whenever a user write a query QE is responsible to process it. QE compares the terms or string in user query with the indexed terms in documents. The query terms may contain in more than one document. The Input/Output module is associated with query engine. Through Input/Output interface the user input a query and a Query Engine implement its query functionalities to disseminate the results to the user. An Input/Output interface is created to interact with QE to present results to user.

3.6 Ranking Model

Ranking model is assigned an important task in information retrieval to retrieve all documents matching a query and put the most relevant results on the top of the list. It merges and sort the results in descending order with most relevant appearing at the top followed by less relevant results from different search nodes [50]. The query terms may exist in different documents and it is also possible that only few of the query terms may contain in a document. In such situation ranking plays an important role as the user may interest in only the most relevant results. Ranking is mechanism that determines how similar are the set of documents to the query.

3.6.1 Vector Space Model

Calculating similarity and dissimilarity is fundamental to IR. Vector Space Model (VSM) calculates the query-documents similarity. Each and every document in a collection is represented as vector of weights. Assume we have V no of documents in a collection then we have |V|-dimensional vector space. Terms are the axis while query and documents are the vectors in space. Once the documents are transformed into vectors then similarity is calculated and documents are ranked according to their similarity to query. Cosine similarity and Euclidean Distance are well-known measures to find similarity in VSM. For finding similarity Euclidean distance is not an appropriate idea under given circumstances as it consider the document's magnitude. Therefore, we have employed cosine similarity to rank documents according to their angles with the query by calculating the cosine similarity.

3.6.2 Cosine Similarity

Cosine similarity is the popular similarity measure often used in text mining, topic modeling and IR. The vector representation of documents in vector space model allow cosine similarity to identify how close or similar are a set of documents to query. Cosine similarity computes the angle between two or more vectors. Cosine similarity has competitive edge over Euclidean distance where documents are apart due to size of the document. Smaller the angle, higher will be the cosine similarity.

$$SC(D,Q) = \sum_{j=1}^{t} w_{qj} \times d_{ij}$$

3.6.3 Euclidean Distance

Euclidean distance is often use for measuring similarity and dissimilarity. Euclidean distance use Pythagorean distance formula to calculate the similarity between. The calculated distance is known as scores that interpret how similar are the documents to the query. Smaller is the distance, higher will be the similarity among objects. The propose model is also tested with Euclidean distance and it came to conclusions that results were not satisfactory. Using Euclidean Distance is not beneficial in IR as it does not take into account the magnitude of documents. This is the reason behind not preferring Euclidean distance as a similarity measure over cosine similarity in the proposed scheme.

3.7 Proposed Model

The proposed model is an LDA based information retrieval model where the documents are first pre-processed and as a result a document-term matrix is generated. An LDA is trained on DTM and it gives output as topic-term matrix calculating probability of word or terms in a topic and document-topic matrix estimating probability of topics in documents. When query is typed the similarity of query terms and topic-term matrix is calculated and those documents are retrieved, ranked and presented to user that have terms similar to query terms. The architecture of the model is proposed in figure 3.2.

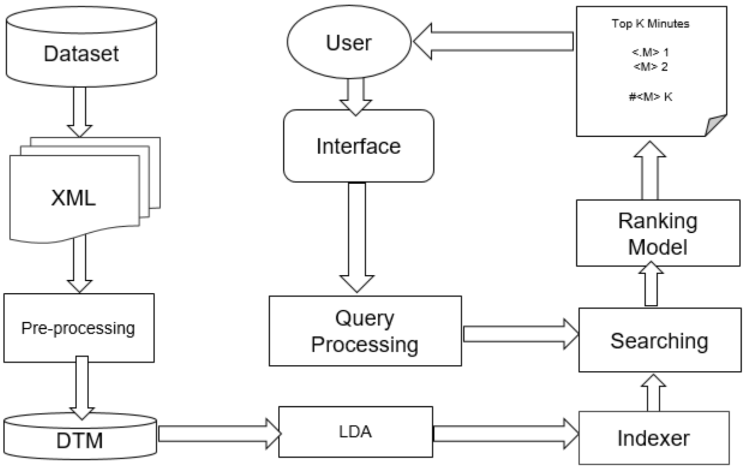


Figure 3.2: Proposed Model

Algorithm 1 Proposed Algorithm

```
Read Dataset
Convert into XML
Pre-process
    a.Tokenization
    b.Stemming
    c.Stop words removal
Document-Term Matrix
    a.Bag of words
Train LDA with XML document-term matrix
    a.Topic-term matrix (words distribution)
    b.Document-topic matrix (topic distribution)
Topic terms indexing
Query processing
Compute Cosine similarity of query terms $q_i$ to topic terms $t_i$. Sim ($q_i$ , $t_i$ )
    a.Retrieve top $k$ topics
    b.Locate and retrieve top $k$ documents of top $k$ topics
Rank top $k$ documents
```

Chapter 4

Experiments and Results

4.1 Experimental Setup

Implementation of proposed model is carried out in java language. This section briefly discusses tools and libraries used in the proposed model.

4.1.1 Spring Tool Suit

Spring tool suit (STS) is an integrated development environment for java programming released in 2019. STS is built on the top of eclipse. It offers built in environment for developers to run, debug and implement application. STS has ready to go features specially intended to give strength to cloud environment and spring projects.

4.1.2 Java

Java is one of the general purpose programming language. Due to availability of large number of open source libraries for XML parsing like JAXB and JAXP the implementation of proposed model is carried out in java. The latest version of java 11.0.1 is exploited to achieve the results of the thesis.

4.1.3 Stanford CoreNLP

Stanford CoreNLP is also open source library in java that provide a set of tools that can be applied to natural language text. The main focus of Stanford CoreNLP library is to make it simple and easy to apply a set of text analysis tools to a piece of text. The motive behind Stanford CoreNLP is to write a few lines of code to process the text. It includes stemming, tokenization, stop word removal and many other text analysis tools. In the proposed model results of both Open NLP and Stanford CoreNLP were combined to avoid any ambiguities at pre-processing step.

4.2 Datasets

Dataset used in this thesis is a corpus consist of documents of semi-structured text holding minutes of the meeting. These minutes are the rules we want to retrieve from the rulebook. Meeting minutes can be characterized as composed or recorded documentation that is utilized to guide individuals regarding what occurred during the meeting. Minutes of the meeting are essential as it give details of the meeting e.g. who were the participants, calendar and dues, defining actions, responsibilities, decisions and discussions. The text in those documents include sections, subsections and enumerations. Each document is different from every other document holding text of different nature some contains table while others don't. This dataset contain 1100 hundred documents. To overcome the problem of diversity of document's structure we convert those documents into XML format and considered each minute of meeting as a document, in order to facilitate us at the time of information retrieval.

In order to evaluate the performance of the proposed model it is also tested with other dataset known as classic4 dataset for comparison purpose. This dataset is known as benchmark dataset in text mining and information retrieval. This dataset is composed of four different collections. For evaluation we have only used two collections CISI and CACM. The first one contain 1460 documents and later one have 3204 documents. These two collections were not in XML format and comparison cannot be drawn, therefore XML tags were ignored with slight adjustment at the pre-processing step to draw an analogy between text of two different natures.

4.3 Evaluation

The output of a query can be categorized into different categories whether relevant or non-relevant. As more and more people are data oriented, they have different thoughts different needs so different systems are required to fulfill their needs. Some people are interested in reading a whole document while some others may be interest in knowing some well-organized statistical information still some others are keen to find answer to a specific question rather than reading a whole document. In information retrieval domain results are also evaluated to know the effectiveness of the information retrieval technique. To evaluation the proposed model following techniques have used to evaluate the results of the thesis.

4.4 Automated

In this type of evaluation, the system is evaluated using recall and mean average precision.

4.4.1 Recall

To find out how much the results of the proposed model were accurate a recall value was calculated for relevant documents retrieved among total number of relevant documents. It was considered to be the critical task because one cannot afford to miss a single relevant document as it notifies how much relevant documents did we miss.

Table 4.1 and 4.2 shows recall values for proposed method and other methods for two different datasets CISI and CACM.

Table 4.1: Recall (CISI)

Method	Recall
LDI	0.26
UniEnM	0.30
EnM	0.33
Proposed Method	**0.41**

Table 4.2: Recall (CACM)

Method	Recall
LDI	0.29
UniEnM	0.39
EnM	0.36
Proposed Method	**0.43**

The proposed model is also trained and tested with our own dataset. To overcome the problem of diversity of document's structure documents in dataset were converted into XML format to improve query performance. This experiment illustrated that the proposed model with XML dataset achieved higher *recall* value of 0.45 that is better than *recall* values of testing the model with other two datasets (CICI and CACM).

4.4.2 Mean Average Precision

Mean average precision is used for multiple or for set of queries. Precision is calculated against each single query and an average precision is calculated for set of queries. Table 4.3 shows MAP of the proposed method.

$$MAP = \frac{\sum_{q=1}^{Q} AveP(q)}{Q}$$

Table 4.3: MAP

Method	CISI	CACM
TFIDF	0.0935	0.1177
LSI	0.1229	0.1094
pLSI	0.1223	0.0973
LDI	0.1429	0.1439
EnM	0.1637	0.1896
Proposed Method	**0.1939**	**0.1937**

4.5 User-Based Evaluation

In user-based evaluation client collaborate with a running framework and note the readings. Values are gathered by asking the user (meetings, surveys) etc. Watching her conduct amid utilize, or naturally recording co-operations and after that subjecting framework logs to different investigations (e.g. click through and conversation rate).

To manually evaluate the proposed model 20 different users were visited and were asked to write 15 different queries of their own choice. The users were then asked to assess the results against each query with a label "satisfied" or "unsatisfied". Results were calculated for each user, as a fraction of satisfied or unsatisfied results among total number of queries. Table 4.4 shows number of satisfied and unsatisfied results for a single user.

Table 4.4: Query Results

	Satisfied	Unsatisfied
Query-1	✓	
Query-2	✓	
Query-3		×
Query-4	✓	
Query-5	✓	
Query-6	✓	
Query-7	✓	
Query-8	✓	
Query-9	✓	
Query-10	✓	
Query-11	✓	
Query-12	✓	
Query-13	✓	
Query-14		×
Query-15	✓	

$$Satisfied = \frac{13}{15}\ Unsatisfied = \frac{2}{15}$$

Table 4.5 shows the ratio of satisfied and unsatisfied results for each person we have visited. There were 20 p and 15 q, So the total number of queries = p x q = 300 among which 245 are with satisfied results and remaining 55 with unsatisfied results. Maximum satisfied result is 14/15 and minimum satisfied result is 9/15. Maximum unsatisfied result is 6/15 and minimum unsatisfied result is 1/15. The percentage of satisfied results is 81.6%.

Table 4.5: Satisfied Unsatisfied Ratio

Person-ID	Satisfied	Unsatisfied
Person-1	13/15	2/15
Person-2	12/15	3/15
Person-3	14/15	1/15
Person-4	11/15	4/15
Person-5	10/15	5/15
Person-6	14/15	1/15
Person-7	14/15	1/15
person-8	12/15	3/15
Person-9	14/15	1/15
Person-10	9/15	6/15
Person-11	12/15	3/15
Person-12	13/15	2/15
Person-13	12/15	3/15
Person-14	14/15	1/15
Person-15	11/15	4/15
Person-16	10/15	5/15
Person-17	14/15	1/15
person-18	12/15	3/15
Person-19	14/15	1/15
Person-20	10/15	5/15

Total	**Satisfied**	**Unsatisfied**
300	245	55

Chapter 5

Conclusions

This section reflects the key points and the knowledge gained while conducting the thesis to achieve the desired results. Furthermore, the impact of our research contributions in relation to previous research will be briefly highlighted.

5.1 Conclusion

In IE and IR systems XML could be a game changer. Once the document is migrated into XML format its benefits could be exploited. Representing natural language text into XML format the text become more easily accessible. This XML format help systems in identifying the start and end of an entities, sentences and paragraphs. Computer don't understand natural language text as good as humans. Without specific XML representation documents are just bag of words and limits the operation that could be perform at the time of retrieval.

Proper indexing of documents as large number of documents are indexed with similar terms but carry different meaning in different contexts and effects the retrieval process. The proposed approach in this thesis known as topic terms indexing of documents improved the query performance. The topic terms generated by LDA are used as indexing terms and allows fast retrieval of relevant material when needed at the time of analysis for decision making.

Because of the fame of XML format few query languages e.g. XPATH and XQUERY have been proposed in last few years to make searching optimized. These languages are rich enough to take into consideration the structure and content of XML document and assure the improved precision of retrieval process. One of the downsides is that this type of search require learning specific query languages. The proposed approach in this thesis have the potential to query XML documents with keyword search and computing the similarity score of XML documents with respect to un-structured query like in search engines without requiring to learn specific query languages. The advantage of proposed method is that it donot require domain knowledge about the structure of XML documents. The proposed method also overcome the problem of HTML data model as it donot capture much semantics. This method of querying XML documents brought up new opportunities to make searching optimized.

Making machines intelligent is the need of today's world which will reduce the burden of humans, tackling many complex tasks manually. In IR domain there is also a need for such systems that can read and understand natural language text e.g, contracts, licenses, agreements, wills, deeds, warrants and court decision etc. For this need to translate the documents into a format that is machine readable that allow automated processing and verification of such text. These kind of transformation also requires an intense reliability.

To extract information from the collection of unseen documents topic modeling is a tool that can leverage the IR process. Due to huge volume of data, topic modeling can be used for knowledge discovery and data mining. These models use statistical assumption for discovering topics based on statistical learning model. Some of these models are based on linear algebra and other are probability base models. LDA is renowned probabilistic model which is being utilized by numerous analysts and having numerous expansions as discussed in chapter 2 of this thesis.

Appendix A

Proposed XML Schema

```
<?xml version="1.0" encoding="UTF-8"?>
<tns:meetings xmlns:tns="http://www.example.org/NewXMLSchema"
 xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
 xsi:schemaLocation="http://www.example.org/NewXMLSchema
    NewXMLSchema.xsd ">
      <meeting>
      <Title>30th Meeting of the Academic Council</Title>
      <Date>2017-10-03</Date>
      <Start>10:00</Start>
      <End>14:00</End>
      <Participants>
            <Name>Vice Adm (R) Tanveer Faiz HI(M)</Name>
            <Designation>Rector</Designation>
            <Role>In Chair</Role>
            <Name>Rear Adm (R) Shahid Saeed HI(M)</Name>
            <Designation>Pro-Rector/DGIC</Designation>
            <Role>Member</Role>
            <Name>Rear Adm Mukhtar Khan HI(M)</Name>
            <Designation>DG IMA</Designation>
            <Role>Member</Role>
      </Participants>
      <Minutes>
            <Minute>
            <Number>1</Number>
            <Itemnumber>1237</Itemnumber>
      <Subject> Draft minutes of the 29th ACM were
         communicated to all members and non-member
```

```
   participants, for comments on 24th May. No comments
   or observations had been received.</Subject>
      <Body></Body>
      </Minute>
      <Minute>
      <Number>2</Number>
      <Itemnumber>1238</Itemnumber>
      <ID></ID>
      <Subject>Consequently, the draft minutes were
         processed on file and the approved minutes
         were then disseminated on 30th May. There had
         been no comments or observations on the
         approved minutes either.</Subject>
      <Body></Body>
<Details>
      <Issue></Issue>
      <Status></Status>
</Details>
</Minute>
<ReportedBy>
      <Name1>BUMDC medical section</Name1>
      <Designation1>Director</Designation1>
</ReportedBy>
<Discussions>
<Discussion>Dean ES briefed on readiness of the BS CS
   and BS IT programmes for accreditation, suggesting
   that the Dept was ready; there were few deficiencies
   which the Dept had been asked to rectify by 9th
   Oct.</Discussion>
<Decision>BULC permitted to start BS professional
   Psychology wef Fall 2018 subject to successful
   accreditations visits in respect of BS CS and BS IT
   and completion of the 4th floor</Decision>
<Discussion>On BS Psychology, concern was raised on
   lack of space. DLC informed that the CU had two
   classrooms and one lab to cater to 25 BS Psychology
   students. Dean PP endorsed, stating 3-4 classrooms
   would serve the purpose. She added that BS
   Psychology had good prospects in Lahore and that
   there was no issue with acquiring the required
   faculty. Dean ES was of the view that whereas it was
   important to start a third department to justify a
   campus, equally important was providing a good
   workplace environment to the faculty, including lab
   engineers.</Discussion>
<Decision>5% admission waiver for BBA and BS IT
   programmes extended to Spring 2018
   intakes.</Decision>
<Discussion>Winding up the discussion, the Chair
   approved BS Psychology, subject to satisfactory
   accreditation visits and completion of the 4th
   floor.</Discussion>
<Decision>Point to remain on agenda and progress
   reported.</Decision>
</Discussions>
<Responsibles>
      <Name2>DG BUMDC</Name2>
      <Designation2>nill</Designation2>
</Responsibles>
<Actions>
      <Action/>
      <Name3>Dean HS</Name3>
      <Designation3>Dean HS</Designation3>
      <Action>Implementation of the Decision c.</Action>
</Actions>
</meeting>
</meetings>
```

References

[1] Mihai Lupu, Katja Mayer, John Tait, and Anthony J Trippe. *Current challenges in patent information retrieval*, volume 29. Springer, 2011. Cited on p. 4.

[2] Yanshan Wang, Jae-Sung Lee, and In-Chan Choi. Indexing by latent dirichlet allocation and an ensemble m odel. *Journal of the Association for Information Science and Technology*, 67(7):1736–1750, 2016. Cited on p. 4.

[3] David M Blei and John D Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120. ACM, 2006. Cited on p. 5.

[4] Thomas L Griffiths, Michael I Jordan, Joshua B Tenenbaum, and David M Blei. Hierarchical topic models and the nested chinese restaurant process. In *Advances in neural information processing systems*, pages 17–24, 2004. Cited on p. 5.

[5] David M Blei, John D Lafferty, et al. A correlated topic model of science. *The Annals of Applied Statistics*, 1(1):17–35, 2007. Cited on p. 5.

[6] Mehdi Allahyari and Krys Kochut. Discovering coherent topics with entity topic models. In *2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 26–33. IEEE, 2016. Cited on p. 6.

[7] II Bommarito, J Michael, Daniel Martin Katz, and Eric M Detterman. Lexnlp: Natural language processing and information extraction for legal and regulatory texts. *arXiv preprint arXiv:1806.03688*, 2018. Cited on p. 6.

[8] Thorsten Brants. Natural language processing in information retrieval. In *CLIN*, 2003. Cited on p. 6.

[9] T. F. Gordon, G. Governatori, and A. Rotolo. Rules and norms: Requirements for rule interchange languages in the legal domain. In *NLP Approaches for Information Retrieval*, volume 5858, page 282, 2009. Cited on p. 6.

[10] Paolo Ferragina and Ugo Scaiella. Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1625–1628. ACM, 2010. Cited on p. 6.

[11] Imran Khan, Muhammad Sher, Javed Khan, Syed Saqlain, Anwar Ghani, Husnain Naqvi, and Muhammad Ashraf. Conversion of legal text to a logical rules set from medical law using the medical relational model and the world rule model for a

medical decision support system. In *Informatics*, volume 3, page 2. Multidisciplinary Digital Publishing Institute, 2016. Cited on p. 7.

[12] Marie-Francine Moens, Caroline Uyttendaele, and Jos Dumortier. Intelligent information extraction from legal texts. *Information & Communications Technology Law*, 9(1):17–26, 2000. Cited on p. 7.

[13] T. Lv, S. Huang, X. Piao, K. Gao, and Y. Jia. A Method of Legal Text Formalization. *Scientia Sinica Physica, Mechanica Astronomica*, 43:26–30, 2012. Cited on p. 7.

[14] Cristian Cardellino, Laura Alonso Alemany, Serena Villata, and Elena Cabrio. Improvements in information extraction in legal text by active learning. In *Proceedings of the 28th Annual Conference on Legal Knowledge and Information Systems*, pages 21–30, 2015. Cited on p. 7.

[15] Frank Schilder, Graham Katz, and James Pustejovsky. Annotating, extracting and reasoning about time and events. In *Annotating, extracting and reasoning about time and events*, pages 1–6. Springer, 2007. Cited on p. 7.

[16] Stefanie Brüninghaus and Kevin D Ashley. Improving the representation of legal case texts with information extraction methods. In *Proceedings of the 8th international conference on Artificial intelligence and law*, pages 42–51. ACM, 2001. Cited on p. 8.

[17] Claudia Soria, Roberto Bartolini, Alessandro Lenci, Simonetta Montemagni, and Vito Pirrelli. Automatic extraction of semantics in law documents. In *Proceedings of the V Legislative XML Workshop*, pages 253–266. European Press Academic Publishing, 2007. Cited on p. 8.

[18] Torsten Schlieder and Felix Naumann. Approximate tree embedding for querying xml data. *IEEE*, 2000. Cited on p. 8.

[19] Mohammad Abolhassani, Norbert Fuhr, and Norbert Govert. Information Extraction and Automatic Markup for XML Documents. In *Intelligent Search on XML Data*, pages 159–174. Springer, 2003. Cited on p. 9.

[20] Henry M Kim and Arijit Sengupta. Extracting knowledge from xml document repository: a semantic web-based approach. *Information Technology and Management*, 8(3):205–221, 2007. Cited on p. 9.

[21] D Ballis and D Romero. Filtering of xml documents. In *2nd International Workshop on Automated Specification and Verification of Web Systems (WWV'06)*, pages 19–28. IEEE, 2006. Cited on p. 9.

[22] Paavo Arvola, Marko Junkkari, and Jaana Kekäläinen. Generalized contextualization method for xml information retrieval. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 20–27. ACM, 2005. Cited on p. 9.

[23] Jun-Ki Min, Jae-Yong Ahn, and Chin-Wan Chung. Efficient extraction of schemas for xml documents. *Information Processing Letters*, 85(1):7–12, 2003. Cited on p. 10.

[24] Ji-xin Zhang and Xia Sun. Keyword retrieval technology research of xml document. In *2011 3rd International Workshop on Intelligent Systems and Applications*, pages 1–3. IEEE, 2011. Cited on p. 10.

[25] Evangelos Kotsakis. Structured information retrieval in xml documents. In *Proceedings of the 2002 ACM symposium on Applied computing*, pages 663–667. ACM, 2002. Cited on p. 10.

[26] Gerard Salton, James Allan, and Chris Buckley. Approaches to passage retrieval in full text information systems. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 49–58. ACM, 1993. Cited on p. 11.

[27] Samuel Sangkon Lee, Masami Shishibori, Toru Sumitomo, and Jun-ichi Aoe. Extraction of field-coherent passages. *Information processing & management*, 38(2):173–207, 2002. Cited on p. 11.

[28] S. Boag, P. Dube, K. El Maghraoui, B. Herta, W. Hummer, K. R. Jayaram, R. Khalaf, V. Muthusamy, M. Kalantar, and A. Verma. A review: Information extraction techniques from research papers. *arXiv e-prints*, pages 56–59, May 2018. Cited on p. 11.

[29] Zongcheng Ji, Zhengdong Lu, and Hang Li. An information retrieval approach to short text conversation. *arXiv preprint arXiv:1408.6988*, August 2014. Cited on p. 11.

[30] Dan Moldovan and Mihai Surdeanu. On the role of information retrieval and information extraction in question answering systems. In *International Summer School on Information Extraction*, pages 129–147. Springer, 2002. Cited on p. 11.

[31] Marie-Francine Moens and Jos Dumortier. Use of a text grammar for generating highlight abstracts of magazine articles. *Journal of documentation*, 56(5):520–539, 2000. Cited on p. 12.

[32] Hamish Cunningham. Information extraction, automatic. *Encyclopedia of language and linguistics,*, pages 665–677, 2005. Cited on p. 12.

[33] Evi Yulianti, Ruey-Cheng Chen, Falk Scholer, W Bruce Croft, and Mark Sanderson. Ranking documents by answer-passage quality. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 335–344. ACM, 2018. Cited on p. 12.

[34] Fernando Llopis, Antonio Ferrández, and José Luis Vicedo. Text segmentation for efficient information retrieval. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 373–380. Springer, 2002. Cited on p. 12.

[35] Robert Miner. The importance of mathml to mathematics communication. *Notices of the AMS*, 52(5):532–538, 2005. Cited on p. 13.

[36] David Michael Parks. News story markup language and system and process for editing and processing documents, March 14 2000. US Patent 6,038,573. Cited on p. 14.

[37] David T Barnard and Nancy M Ide. The text encoding initiative: Flexible and extensible document encoding. *Journal of the American Society for Information Science*, 48(7):622–628, 1997. Cited on p. 14.

[38] Peter Murray-Rust, Henry S Rzepa, and Michael Wright. Development of chemical markup language (cml) as a system for handling complex chemical content. *New journal of chemistry*, 25(4):618–634, 2001. Cited on p. 14.

[39] Ian Ruthven. Information retrieval in context. In *Advanced Topics in Information Retrieval*, pages 187–207. Springer, 2011. Cited on p. 15.

[40] Denis Andrei de ARAUJO, RIGO Sandro José, Carolina Müller, and Rove Chishman. Information extraction from legal documents using linguistic knowledge and ontologies. *ACM*, 2009. Cited on p. 15.

[41] Kamran Munir and M Sheraz Anjum. The use of ontologies for effective knowledge modelling and information retrieval. *Applied Computing and Informatics*, 14(2):116–126, 2018. Cited on p. 15.

[42] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60, 2014. Cited on p. 17.

[43] Kranti Vithal Ghag and Ketan Shah. Comparative analysis of effect of stopwords removal on sentiment classification. In *2015 International Conference on Computer, Communication and Control (IC4)*, pages 1–6. IEEE, 2015. Cited on p. 17.

[44] Martin F Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980. Cited on p. 18.

[45] Wahiba Ben Abdessalem Karaa and Nidhal Gribâa. Information retrieval with porter stemmer: a new version for english. In *Advances in computational science, engineering and information technology*, pages 243–254. Springer, 2013. Cited on p. 18.

[46] Martin F Porter. Snowball: A language for stemming algorithms, 2001. Cited on p. 18.

[47] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003. Cited on p. 20.

[48] Mark Steyvers and Tom Griffiths. Probabilistic topic models. *Handbook of latent semantic analysis*, 427(7):424–440, 2007. Cited on p. 20.

[49] Robert WP Luk, Hong Va Leong, Tharam S Dillon, Alvin TS Chan, W Bruce Croft, and James Allan. A survey in indexing and searching xml documents. *Journal of the American society for Information Science and Technology*, 53(6):415–437, 2002. Cited on p. 20.

[50] Torsten Schlieder and Holger Meuss. Querying and ranking xml documents. *Journal of the American Society for Information Science and Technology*, 53(6):489–503, 2002. Cited on p. 21.

final thesis

ORIGINALITY REPORT

15%	10%	8%	11%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Higher Education Commission Pakistan Student Paper	3%
2	dblp.dagstuhl.de Internet Source	1%
3	Submitted to University of Sheffield Student Paper	<1%
4	docplayer.net Internet Source	<1%
5	oro.open.ac.uk Internet Source	<1%
6	onlinelibrary.wiley.com Internet Source	<1%
7	Submitted to University of Pretoria Student Paper	<1%
8	Lecture Notes in Computer Science, 2015. Publication	<1%
9	hal.inria.fr	