International Journal of Scientific Research & Engineering Trends
Volume 4, Issue I, Mar-apr-2018, ISSN (Online): 2395-566X

A Survey on Various Techniques and Characteristics of Text Document Fetching

Rajeev Kumar
Dept. of Computer Science and Engineering
Mittal Institute of Technology, Bhopal, India
rajeevniranjan999@gmail.com

Prof. Durgesh Wadbude
Dept. of Computer Science and Engineering
Mittal Institute of Technology, Bhopal, India

A.P. Jayshree Boaddh
Dept. of Computer Science and Engineering
Mittal Institute of Technology, Bhopal, India
jayshree.boaddh@gmail.com

Abstract – **Search engines are the major breakthrough on the web for retrieving the information. But List of retrieved documents contains a high percentage of duplicated and near document result. So there is the need to improve the performance of search results. In this paper text document retrieval study is done with various techniques of fetching with there implementations. Here different features for the text document retrieval is explained in detailed with there requirements as feature vary as per text analysis. Paper has brief different evaluation parameters for the study and comparison of relevant documents techniques**.

Keywords- **Classification analysis, Ontology , Supervised classification, Un-supervised Classification, Text Mining,**

I INTRODUCTION

The basic function of communication is transferring the data from one corner to another corner of the world. The data is basically stored in the form of documents, files and these files are arranged under folder or subfolder. The random creation and storage makes them unstructured in nature which results in inefficient data retrieval and modification as well as updation. E-commerce and corporate intranets has led to the growth of organizational repositories containing large and unstructured document collection. So, efficient storage and transmission of documents as well as archiving and information retrieval for document databases have become important research issues. Structured documents must maintain the structure where in addition to pure textual information, the meaning of different sections likes author, title, abstract, heading of section or subsection, paragraph, etc. are also stored within same document whereas unstructured document does not have a predefined manner. Unstructured information is basically text heavy, also contain data such as dates, numbers, facts [1]. As the size of unstructured data in our world continues to increase, text mining tools that allow sifting through this information with ease will become more and more valuable.

Text mining tools are beginning to be readily applied in the biomedical field, where the volume of information on a particular topic makes it impossible for a researcher to cover all the material, much less explore related texts[2]. Text mining methods can also be used by the government's intelligence and security agencies to try to piece together terrorist warnings and other security threats before they occur. Another area that is already benefiting from text mining tools is education. Students and educators can find more information relating to their topics at faster speeds than they can use traditional adhoc searching. The new developments in text mining technology that go beyond simple searching methods are the key to information discovery and have a promising outlook for application in all areas of work[2][3].

Only relevant to user in order to analyze and extract useful information from data. Thus, Text Mining has become more and more popular and essential topic in DM. Text Mining, also known as knowledge discovery (KD) from text, and document information mining (IM), refers to the procedure of extracting fascinating information from very large text quantity for the purposes of determining knowledge[7, 8]. It is an interdisciplinary field involving IR, understanding text, extraction of information, clustering, classification, linkage of concept, visualization, database knowledge, machine learning (ML), and DM. Search engine is the most well known Information Retrieval tool. Application of Text Mining techniques to Information Retrieval can improve the precision of retrieval systems by filtering relevant documents for the given search query[4].

II. REQUIREMENT OF DOCUMENT MINING

1. Information Retrieval

Information retrieval (IR) concept has been developed in relation with database systems for many years. Information retrieval is the association and retrieval of information from a large number of text based documents. The information retrieval and database systems, each

handle various kinds of data; some database system problems are usually not present in information retrieval systems, such as concurrency control, recovery, transaction management, and update. Also, some common information retrieval problems are usually not encountered in conventional database systems, such as unstructured documents, estimated search based on keywords, and the concept of relevance. Due to the huge quantity of text information, information retrieval has found many applications. There exist many information retrieval systems, such as on-line library catalog systems, on-line document management systems, and the more recently developed Web search engines [11].

Predefined sequences in the text, a process called pattern matching. The software infers the relationships between all the identified places, people, and time to give the user with meaningful information. This technology is very useful when dealing with large volumes of text. Traditional data mining assumes that the information being "mined" is already in the form of a relational database. Unfortunately, for many applications, electronic information is only available in the form of free natural language documents rather than structured databases [10].

Categorization involves identifying the main themes of a document by inserting the document into a predefined set of topics. When categorizing a document, a computer program will often treat the document as a "bag of words." It does not try to process the actual information as information extraction does. Rather, the categorization only counts words that appear and, from the counts, identifies the main topics that the document covers. Categorization often relies on a glossary for which topics are predefined, and relationships are identified by looking for large terms, narrower terms, synonyms, and related terms [9].

2. Natural Language Processing

Natural Language Processing (NLP) is an area of research and application that explores how computers can be used to understand and manipulate natural language text. NLP researchers aim to collect knowledge on how human beings understand and use language so that fitting tools and techniques can be developed to make computer systems understand and manipulate natural languages to perform the preferred tasks [3]. The basics of NLP lie in a number of disciplines, viz. computer and information sciences, linguistics, mathematics, electrical and electronic engineering, artificial intelligence and robotics, psychology, etc. Applications of NLP include a number of fields of studies, such as machine translation, natural language text processing and summarization, user interfaces, multilingual and cross language information retrieval (CLIR), speech recognition, artificial intelligence and expert systems and so on[13].

III. Related Work

In [2] says that paper investigate four different methods for document classification: the naive Bayes classifier, the nearest neighbor classifier, decision trees and a subspace method. These were applied to seven-class Yahoo news groups (business, entertainment, health, international, politics, sports and technology) individually and in combination. Study of three classifier combination approaches: simple voting, dynamic classifier selection and adaptive classifier combination. Experimental results indicate that the naive Bayes classifier and the subspace method outperform the other two classifiers on data sets. Combinations of multiple classifiers did not always improve the classification accuracy compared to the best individual classifier. Among the three different combination approaches, adaptive classifier combination method introduced performed the best.

In [3] explains Automatic Text Classification is a semi-supervised machine learning task that automatically assigns a given document to a set of pre-defined categories based on its textual content and extracted features which is also a close related paper to my research. This paper explains the generic strategy for automatic text classification which includes steps such as pre-processing, feature selection using various statistical or semantic approaches, and modeling using appropriate machine learning techniques (Naïve Bayes, Decision Tree, Neural Network, Support Vector Machines, Hybrid techniques). This paper also discusses some of the major issues involved in automatic text classification such as dealing with unstructured text, handling large number of attributes, examining success of purely statistical pre-processing techniques for text classification v/s semantic and natural language processing based techniques, dealing with missing metadata and choice of a suitable machine learning technique for training a text classifier.

In [4] paper discussed about the text mining and its preprocessing techniques. Text mining is the process of mining the useful information from the text documents. It is also called knowledge discovery in text (KDT) or knowledge of intelligent text analysis. Text mining is a technique which extracts information from both structured and unstructured data and also finding patterns. Text mining techniques are used in various types of research domains like natural language processing, information retrieval, text classification and text clustering.

In [5] paper, a hierarchical clustering method is proposed to support more search semantics and also to meet the demand for fast cipher text search within a big data environment. The proposed hierarchical approach clusters the documents based on the minimum relevance threshold, and then partitions the resulting clusters into sub-clusters until the constraint on the maximum size of cluster is

reached. In the search phase, this approach can reach a linear computational complexity against an exponential size increase of document collection. In order to verify the authenticity of search results, a structure called minimum hash sub-tree is designed in this paper. The results show that with a sharp increase of documents in the dataset the search time of the proposed method increases linearly whereas the search time of the traditional method increases exponentially. Furthermore, the proposed method has an advantage over the traditional method in the rank privacy and relevance of retrieved documents.
In [11] has developed a system which can learn from text query examples to improve retrieval performance. This is called relevance feedback and has proven to be effective in improving retrieval performance. When we do not have such relevant examples, a system can assume the top few retrieved documents in some initial retrieval results to be relevant and extract more related keywords to expand a query. Such feedback is called pseudo-feedback or blind feedback and is essentially a process of mining useful keywords from the top retrieved documents. Pseudo-feedback also often leads to improved retrieval performance. One major limitation of many
existing retrieval methods is that they are based on exact keyword matching. However, due to the complexity of natural languages, keyword based retrieval can encounter two major difficulties.
In [12] can perform some types of analysis with a high degree of success. Shallow parsers identify only the main grammatical elements in a sentence, such as noun phrases and verb phrases, whereas deep parsers generate a complete representation of the grammatical structure of a sentence. The role of NLP in text mining is to provide the systems in the information extraction phase (see below) with linguistic data that they need to perform their task. Often this is done by annotating documents with information like sentence boundaries, part-of-speech tags, parsing results, which can then be read by the information extraction tools.

Public Encryption with Keyword search [6] can help to test the given keyword present in the document without learning anything else from the document. Data stored in untrusted server can be encrypted. Search the data by using keyword. By using PEKS reduce the processing time by retrieve only the selected files. By its disadvantage by using the application such as patient record and investigations, a small mistake on spelling on keyword cannot produce any result. Thus by going Fuzzy Keyword Searching.

IV. Techniques of Document Retrieval

1. KNN (K nearest Neighbors algorithm) in [4] is used which utilize nearest neighbor property among the items. This algorithm is easy to implement with high validity and required no prior training parameters. Although K nearest neighbor is also identified as instance based learning in other words classification of items is quite slow. In this classification techniques distance between the K cluster center and classifying item is calculated then assign item to cluster having minimum distance from the cluster center. In case of text mining features from the document is extracted then k labeled node is select randomly which are suppose to be cluster center and rest of nodes or document are unlabeled nodes. Finally distance between labeled and unlabeled node is calculate on the base of feature vector similarity. In this algorithm distance between nodes are estimate in log(k) time .

Advantages: Main significance of this algorithm is that this is robust against raw data which contain noise. In this algorithm prior training is not required as done in most of the neural network for classification. One more flexibility of this algorithm is that this work well in two or multiclass partition.

Limitations: In this work selection of appropriate neighbor is quite high if population of item is large in number. One more issue is that it required much time for finding the similarity between the document features. Because of these limitations this algorithm is not practical with large number of items. So cost of classification increases with increase in number of items.

2. Support Vector Machine (SVM) in [3] is quite famous soft computing technique for item classification which is based on the input feature vector quality and training of the support vector machine. In this technique an hyper plane is build between the items this hyper plane classify the items into binary or multi class. In order to find the hyper plane equation is written as P = B+XxW where X Is a an item to be classify then W is vector while B is constant. Here W and B is obtained by the training of SVM. So SVM can perfectly classify binary items by using that calculated hyper plane.

Advantages: Main significance of the Support Vector Machines is that it is less susceptible for over fitting of the feature input from the input items, this is because SVM is independent of feature space. Here classification accuracy with SVM is quite impressive or high. SVM is fast accurate while training as well as during testing.
Limitations: In this classification multiclass items are not perfectly classify as number of items reduce gap of hyper plane.

3. Fuzzy classification in [5], has classify image data which is highly complex and required stochastic relations for the creation of feature vector from images. Here different types of relations are combined where members of the feature vector is fuzzy in nature. So this relation based image classification is highly depend on the type of image format as well as on the threshold selection.

Advantages: This algorithm is easy to handle, while stochastic relation help in identifying the different uncertainty properties.

Limitation: Here deep study is required to develop those stochastic relations; accuracy is depend on prior knowledge.

IV. Evaluation Parameters

Once pattern are discover then the compared results of the system are need to be evaluate that either results obtain are correct or not. If the obtained results are not valid then training parameter need to be change, if results are still vary then the pattern discovered are also need to be reshuffled or change as per requirement.

Precision = true positives / (true positives+ false positives)
Recall = true positives / (true positives +false negatives)
F-score = 2 * Precision * Recall / (Precision + Recall)

In order to evaluate results there are many parameter such as accuracy, precision, recall, F-score, etc. Obtaining values can be put in the mention parameter formula to get better results.

v. Conclusion

As the writing work of different articles from laboratory, organization, press media, institutes are increasing day by day. Then publishing their work is also increase which is done by most of the journals, news paper, organizations. Here paper has cover an important issue of document retrieval. Various techniques with there required features are discussed in detailed. This paper describes that Text data mining is always used with various fields like Natural Language Processing, Data Mining, Information Extraction, and Information Retrieval. Various techniques for Text mining has been also explained briefly which includes Information Extraction, Topic Tracking, Summarization, Clustering, Categorization and Association Rule Mining

References

[1]. Selma Ayşe Özel. Esra Saraç " Web Page Classification Using Firefly Optimization ", 978-1-4799-0661-1/13/$31.00 ©2013 Ieee.
[2]. Vandana Korde, C Namrata Mahender "TextClassification and classifiers: A survey".ijaia,2012.
[3]. Bhumika, Prof Sukhjit Singh Sehra, Prof Anand Nayyar "A review paper on algorithms used for Text Classification" International Journal of Application or Innovation in Engineering & Management (IJAIEM) 2013.
[4]. Dr. S. Vijayarani , Ms. J. Ilamathi, Ms. Nithya. "Preprocessing Techniques for Text Mining - An Overview". International Journal of Computer Science & Communication Networks,Vol 5(1),7-16 7 ISSN:2249-5789
[5]. Chi Chen, Xiaojie Zhu, Peisong Shen, Jiankun Hu, Member, IEEE, Song Guo, Zahir Tari, and Albert Y. Zomaya. "An Efficient Privacy-Preserving Ranked Keyword Search Method". IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, VOL. 27, NO. 4, APRIL 2016
[6]. Peng Xu and Hai Jin. Public-key encryption with fuzzy keyword search: A provably secure scheme under keyword guessing attack. Cryptology ePrint Archive, Report 2010/626, 2010.
[7]. Lin, Yung-Shen, Jung-Yi Jiang, and Shie-Jue Lee. "A similarity measure for text classification and clustering." IEEE transactions on knowledge and data engineering 26.7 (2014): 1575-1590.
[8]. Li, Zechao, et al. "Clustering-guided sparse structural learning for unsupervised feature selection." IEEE Transactions on Knowledge and Data Engineering 26.9 (2014): 2138-2150.
[9]. Souneil Park, Jungil Kim, Kyung Soon Lee, And Junehwa Song "Disputant Relation-Based Classification For Contrasting Opposing Views Of Contentious News Issues". Ieee Transactions On Knowledge And Data Engineering, Vol. 25, No. 12, December 2013.
[10]. Ghosh S, Roy S, and Bandyopadhyay S K, (2012), A tutorial review on Text Mining Algorithms, International Journal of Advanced Research in Computer and Communication Engineering,1(4)..
[11]. Massimo Melucci, "Relevance Feedback Algorithms Inspired By Quantum Detection",ieee transactions on knowledge and data engineering, vol. 28, no. 4, april 2016.
[12]. Deepali D. Rane and Dr.V.R.Ghorpade " Multi-User Multi-Keyword Privacy Preserving Ranked Based Search Over Encrypted Cloud Data" International Conference on Pervasive Computing (ICPC), 2015.
[13]. Bing Wang, Wei Song, Wenjing Lou, and Y. Thomas Hou "Inverted Index Based Multi-Keyword Public-key Searchable Encryption with Strong Privacy Guarantee" IEEE Conference on Computer Communications (INFOCOM), 2015.