International Journal of Current Trends in Engineering & Technology
ISSN: 2395-3152
Volume: 03, Issue: 04 (JULY– AUGUST, 2017)

A Survey on Various Techniques and Characteristics of Text Document Fetching

Shruti Pathak, Prof. Manish Misra
Department of Computer Science & Engineering
Vaishnavi Institute of Science & Technology, Bhopal
shrutirpathak@gmail.com

Abstract:— As the digital data increases on server's different researcher have focused on this field. As various issues are arise on the server such as data handling, security, maintenance, etc. In this paper text document retrieval study is done with various techniques of fetching with their implementations. Here different features for the text document retrieval is explained in detailed with their 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

With evolution of computers the life of people became more and more easily. They were able to keep their data on their devices, and started finding ways to make them accessible to others, for example say by using poppy, writable disks, which was followed by portable hard-disk, all these where expensive in their own way during their time. The data was very much private on personal devices like PC, laptops, mobile phones etc, therefore sharing data with others was considered to be expensive. As the world of computing got more advanced the ways for sharing data started becoming cheaper and cheaper. In recent years a new term has evolved call "Cloud" which is provided by different provides, and which is nothing but facility or service of different resources or apparatus like platform, hardware, software, storage's etc, and this make user free from maintenance which has increase the importance of the work as all these are the cloud service provider responsibility. Now to provide such service to the client, naturally the provider's must have and rather can have access to resources which are used by the people/clients. Among the reasons these access are greatly required are for maintenance perspective. As thousands of client are using those service, so infrastructure tends to be capable for making support of this work. In cloud 24x7 Service availability, data maintenance between various devices, then availability of data via any devices, web browser based connectivity. So problem of information fetching is not so amenable to automatic processing of text document. But with the use of text mining approach document is converting into appropriate format so that computer can easily digest whole document. This can be understand as by introducing the text mining approach document is convert into computer readable and under stable format so without any manual interruption system can treat whole data for information interpretation. As text mining involves applying very computationally intensive algorithm to large document collection, IR can speed up the analysis considerably by reducing documents for analysis. For example if interested in mining information only about protein interaction, might restrict our analysis to documents that contains the name of a protein or some form of the web 'to interact' or one of its synonymous.

II. FEATURES OF DOCUMENT MINING

1) Title feature:- The word in sentence that also occurs in title gives high score. This is determined by counting the number of matches between the content word in a sentence and word in the title. In [4] calculate the score for this feature which is the ratio of number of words in the sentence that occur in the title over the number of words in the title.

2) Sentence Length:- This feature is useful to filter out short sentence such as datelines and author names commonly found in the news articles the short sentences are not expected to belong to the summary. In [5] use the length of sentence, which is the ratio of the number of words occurring in the sentence over the words occurring in the longest sentence of the documents.

3) Term Weight:- The frequency of the term occurrence with documents has been used for calculating the importance of sentence. The score of a sentence can be calculated as the sum of the score of words the sentences. The score of important score wi of word i can be calculated by traditional tf.idf method.

4) Sentence position:- Whether it is the first 5 sentence in the paragraph, sentence position in text gives the

importance of the sentences. These features can involve several items such as the position of the sentence in the documents, section and the paragraph, etc, proposed the first sentence of highest ranking. The score for these features in [6] consider the first 5 sentence in the paragraph.

5) Sentence to sentence similarity:- This feature is a similarity between sentences for each sentence S , the similarity between S and each other sentence is computed by the cosine similarity measure with a resulting value between 0 and 1 [6]. The term Weight wi and wj of term t to n term in sentences Si and Sj are represented as the vector. The similarity of each sentence pair is calculated based on similarity.

6) Proper Noun:- The sentence that contains more proper noun (name entity is an important and is most probably include in the document summary. The score for this feature is calculate as the ratio of the number of proper noun that occur in the sentence, over the sentence length. S_f(6)S = No. Proper noun in S/Sentence Length (S).

7) Thematic Word:- The number of thematic word in the sentence, this feature is important because term that occurred frequently in a document are probably related to the topic. The number of thematic word indicates the word with maximum possible relativity. We used the top 10 most frequent content word for consideration as thematic. the score for these features is calculated as the ratio of the number of thematic words that occur in the sentence over the maximum summary of thematic word in the sentence. S_f7(S) = No. Thematic word in S/Max (No. Thematic word).

III. RELATED WORK

Lin, Yung-Shen, Jung-Yi Jiang, and Shie-Jue Lee [7] proposed a new similarity measure algorithm for text classification and clustering. This takes many cases for similarity calculation, which are features from both documents, features from a single document and features are not in the given documents. The authors generated the awareness of detecting presence and absence of features, features have non-zero values. This has been applied in hierarchical clustering and KNN clustering algorithms. But the proposed work has been investigated only few clustering algorithm and doesn't provide accuracy in similarity finding.

Li, Zechao, et al[8] developed a novel unsupervised feature selection algorithm, named as clustering guided sparse structural learning (CGSSL). This integrates the cluster analysis and sparse structural analysis as a joint framework. The authors used the nonnegative spectral clustering for accurate cluster label detection. The cluster labels are predicted using non-negative analysis.

Massimo Melucci [13] present a class of RF algorithms inspired by quantum detection to re-weight the query terms and to re-rank the document retrieved by an IR system. Focuses on explicit RF and on pseudo RF. Implicit RF is based on observations (e.g., click-through data) that are proxies of relevance. The main problem with proxies is that they are not necessarily reliable indicators of relevance and thus should be considered noisy. How quantum detection can help "absorb" noise can also be investigated in the future work.

Deepali D. Rane et.al, [13] proposed implementation of the encryption and decryption, Secure index construction is successfully completed with desirable performance. After index construction it will get compressed and will be stored in .cfs file format. After firing single-keyword query, user will get all documents that contain the specified keyword. The advantages are protects data privacy by encrypting documents before outsourcing, rank based retrieval of the documents, To easily access the encrypted data by multi keyword rank search using keyword index. The Disadvantages of the proposed system are single-keyword search without ranking, Boolean keyword searching without ranking, single-keyword search with ranking, Rarely sorting of the results i.e. no index creation and ranking, Single User search.

Bing Wang et.al, [14] proposed a novel construction of a public key searchable encryption scheme based on inverted index. This scheme overcomes the one-time-only search limitation in the previous schemes. The disadvantages of the proposed system are first of all, the keyword privacy is compromised once a keyword is searched. As a result, the index must be rebuilt for the keyword once it has been searched. Such solution is counterproductive due to the high overhead suffered. Secondly, the existing inverted index based searchable schemes do not support conjunctive multi-keyword search, which is the most common form of queries now a days. The advantages are exploring the problem of building a searchable encryption scheme based on the inverted index, Achieve secure and private matching between the query trapdoor and the secure index.

IV. TECHNIQUES OF DOCUMENT RETRIEVAL

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 . 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. 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.

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 a 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 ia an item to be classify then W is vector while B is constant. Here W and B are obtained by the training of SVM. So SVM can perfectly classify binary items by using that calculated hyper plane. 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. In this classification multiclass items are not perfectly classify as number of items reduce gap of hyper plane. 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 depending 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. Here deep study is required to develop that stochastic relation, accuracy is depending on prior knowledge. Sagayam, Srinivasan, Roshni, 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.

Ghosh, Roy, Bandyopadhyay 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 un-trusted 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. 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 their required features are discussed in detailed. Here paper related work of researchers done in this field. So it can be concluded that one strong algorithm is required that can effectively classify and retrieve document while it need a strong ontology for same.

V. REFERENCES

[1]. Selma Ayşe Özel. Esra Saraç " Web Page Classification Using Firefly Optimization ", 978-1-4799-0661-1 /13 /$31.00 ©2013 IEEE.

[2]. Shrilakshmi Prasad, B. S. Mamatha." Retrieving documents from encrypted cloud data in a secured way using cosine similarity search with multiple keyword search support. " International Journal of Advance Research in Computer Science and Management Studies. Volume 4, Issue 5, May 2016

[3]. G. Salton, C. Buckley, "Term-Weighting Approaches In Automatic Text Retrieval" Information Processing And Management 24, 2008. 513-523.

[4]. L. Suanmali, N. Salim, M.S. Binwahlan, "Srl-Gsm: A Hybrid Approach Based On Semantic Role Labeling And General Statistic Method For Text Summarization", Research Article- Journal Of Applied Science, 2010.

[5]. M. K. Dalal, M. A. Zaveri, "Semi supervised Learning Based Opinion Summarization And Classification For Online Product Reviews", Hindawi Publishing Corporation Applied Computational Intelligence And Soft Computing, Volume 2013.

[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]. Mohinder Singh*, Navjot Kaur . "Retrieve Information Using Improved Document Object Model Parser Tree Algorithm". Mohinder Singh, Navjot Kaur / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com Vol. 3, Issue 4, Jul-Aug 2013, pp.2671-2675.

[11]. Sagayam R, Srinivasan S, and Roshni S, (2012), A Survey of Text Mining: Retrieval, Extraction and Indexing Techniques, International Journal Of Computational Engineering Research, 2(5).

[12]. 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)..

[13]. Massimo Melucci, "Relevance Feedback Algorithms Inspired By Quantum Detection", IEEE Transactions On Knowledge And Data Engineering, Vol. 28, NO. 4, APRIL 2016.

[14]. 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.

[15]. 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.