International Journal of Research in Advent Technology, Vol.5, No.1, January 2017
E-ISSN: 2321-9637
Available online at www.ijrat.org

A Survey on Efficient Data Retrieval for Cloud Computing

Manisha Patel[1], Prof. Umesh Lilhore
CSE, RGPV University/ NIIST, Bhopal, India
Email: manishakhushi.patel@gmaill.com [1] *umeshlilhore@gmail.com,*[2]

Abstract- As the digital data increases on servers different researcher have focused on this field. As a variety of issues are occur on the server such as security, maintenance, data handling, retrieval etc. So fetching from this bulk data is highly required for the effective use of digital world. In this paper text article retrieval is shown by explaining various techniques of fetching with privacy document and user query. Here diverse attributes for the text document fetching is explained in detailed with their necessities as attribute vary as per text study. Paper has concise diverse evaluation parameters for the study and comparison of relevant documents techniques.

Index Terms- Supervised Classification, Text Mining, , Text Feature, Text Ontology, Un-supervised Classification.

1. INTRODUCTION

With evolution of computers the life of people became more and more easily. They where able to keep there data on there 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 there own way during there 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.
The cloud term is of interest not just to the patient clients but to organizations as well. With organization as a consumer the concern of data security becomes multifold. Consider a typical example of small scale business that has different departments like HR, Finance, etc. It is already know that for extracting knowledge from the raw data, mining technique is apply where generation of patterns is very important task. Same thing of pattern generation is also done in text mining where generation of patterns from the text is required for the information extraction. Although work is different but mining task in text and data is much different. In case of data mining implicit information is present in the dataset: here hidden unknown and hard extraction of information is done in the text. In case of text mining information is not hidden in form of any unknown form where extraction of information is very hard. One more issue in txt mining is the understanding of the knowledge in case of different human wordings, as sentence is different in different cases.

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.

2. 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 features 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 belongs 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 a 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. This 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 this 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.

3. 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 .

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.

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 hyperplane is build between the items this hyperplane classify the items into binary or multi class. In order to find the hyperplane 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 is obtained by the training of SVM. So SVM can perfectly classify binary items by using that calculated hyperplane.

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 hyperplane.

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 uncertainity properties.

Limitation: Here deep study is required to develop those stochastic relation, accuracy is depend 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 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.

4. RELATED WORK

Shrilakshmi Prasad, B. S. Mamatha in [2] has makes the penetrating easy, the directory file should be build for each article. The directory file has the keyword and its add up in the exacting article. The normal index file tends to association assail as with the significant words and their add up, the pleased of the leader can be documented. In this work, examiner shows and solve the problem of association assail by hiding the index file using Paillier cryptographic method. So cloud will have to face of penetrating the directory file with the look for query where both will be in an encrypted format. Hence isolation of the text will be potted. Cosine similarity look for is used to get back the top matching documents based on their significance score. The loveliness of the proposed scheme is the user can give manifold keywords in their search query.
Mohinder Singh*, Navjot Kaur in [10] has removed needless information in the document. The DOM Parser Tree Algorithm to filter the web pages from superfluous data and give the dependable output. The article Object Model Parser Tree Algorithm retrieve the HTML links. According to these relations the pages are right of entry. Then the information with is helpful for consumer, is drive to the desk. The DOM Parser Tree Algorithm works upon tree arrangement and have used the table for productions. As the results are exposed in the table, the in sequence shown in the table is accurate and dependable for the user. The user hits the data which he or /she wants to access time by time. The data dynamically fetched from that particular website or link.

Samiksha Chakule, Ashwini Borse, Shalaka Jadhav, Dr. Mrs. Y.V. Haribhakta in [13] has solved an problem of browse from end to end large amount of data on web to get the consequences of his interests. This difficulty has resulted into web page identification which was mapped to different meanings mixed jointly in the consequence list. This increments the load on look for engine and thus reduce its presentation. Here its have planned an movement based on the use of Wikipedia as a knowledge base in information recovery action on web. Our purpose is to solve uncertainty in query by means of semantic information of Wikipedia. This unit search supporting words for a known query contained ambiguous conditions. Wikipedia helps in proving large amount of knowledge that is based on structured fashion to compute semantic relatedness between texts. Also it is a large source of data which is modernized constantly.

5. TEXT PREPROCESSING

As document is collection of paragraphs. Paragraphs are collection of sentences. While sentences are collection of words. So whole preprocessing focus on

Author	***Technique***	***Merit***	***De-merit***
Jian Ma, Wei Xu, Yonghong Sun, Efraim Turban, Shouyang Wang, and Ou Liu. 2012. [8]	*A novel ontology-based text-mining approach to cluster research proposals based on their similarities in research areas.*	*Classify on the basis of keywords present in research papers.*	*Less efficient as patter n based classification work well.*
Souneil Park, Jungil Kim, Kyung Soon Lee, and Junehwa Song. 2013. [9]	*A modified version of HITS algorithm and an SVM classifier trained with pseudorelevant data for article analysis.*	*Efficientl y classify documen ts on the basis of disputant s.*	*Need prior knowl edge for disput ant identi ficati on.*
Esra Saraç, Selma Ayşe Özel. 2013. [1]	*In this study, used FA to select a subset of features, and to evaluate the fitness of the selected features classifier of the Weka data mining tool is employed.*	*It require less execution time.*	*Classi ficati on accur acy is quite less.*
Yuefeng Li, Abdulmohse n Algarni, Mubarak Albathan, 2015. [7]	*This work adopt term as well pattern feature for classifying document in two category.*	*Utilizatio n of both feature increase the classifica tion accuracy .*	*Proce ss requir ed high execu tion time.*

word in the document without any punctuations. So in pre-processing of document there are two common steps first is stop word removal, and second is stem word removal. [8]

Stop List Removals: As sentence is frame with number of words but some of those words are just use to contruct a proper sentence although it does not make any information in the sentence. So identification of those words then removing is term as Stop word removal. So a list of words is store by the researcher which help in identifying of stop words. This removal of stop words help in reduce the execution time of the algorithm, at the same time noisy words which not give any fruitful information is also removed. Stop words are like {a, the, for, an, of, and, etc.}. So text document is transform into collection of words which is then compare with these words and then each match word is removed from the document.
Inorder to understand this assume an sentence {India is a great country in the world} then after pre-processing it become {India, great, country, world} while stop words {is, a, in, the} in the sentence are removed. Let

Stem Word Removal In this words which are almost similar in prefix are replace by one word. This can be said collection of words share same word is term as stem. So there occurrence in the document make same effect but while processing in text mining algorithm it make different so update each word from the collection into single word is done in this stem word removal pre-processing step. Let us assume an collection of words for better understanding of this work. Collection of word is {play, plays, playing} then replace each with word {play}.

6. PRIVACY PRESERVING ALGORITHM

AES is obviously a substitute for DES was needed have hypothetical attacks that can smash it have demonstrated comprehensive key seek attacks can use Triple-DES – but time-consuming. Here four steps of 11 / 13/ 17 round is taken.

rounds in which state undergoes:

- Byte substitution (1 S-box used on every byte)
- Shift rows (permute bytes between groups/columns)
- Mix columns (subs using matrix multiply of groups)
- Add round key (XOR state with key material)

view as alternating XOR key & scramble data bytes initial XOR key material & incomplete last round with fast XOR & table lookup implementation.

Pailler cryptosystem: This cryptosystem is base on the public and private key concept.

Here input vector D[n], will be encrypt by this algorithm.

- Choose two large prime numbers p and q randomly and independently of each other such that gcd(pq,$(p-1)(q-1)$)=1.
- Compute RSA modulus $n = pq$ and Carmichael's function $\lambda = lcm(p-1, q-1)$
- Select generator g, Select α and β randomly from a set $\mathbb{Z}n^2$ then calculate $g=(\alpha n+1)\beta * \beta \bmod(n^2)$
- Calculate the following modular multiplicative inverse $\mu = \text{mod(n)} / (L(g\lambda \bmod(n*n)^{-1}$ Where the function L is defined as $(u)=(u-1)/n$.

So The public key is (n, g), private key is (λ, μ).

7. 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. 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 an strong ontology for same.

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, "Semisupervised 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] Ning Zhong, Yuefeng Li, And Sheng-Tang Wu "Effective Pattern Discovery For Text Mining". Ieee Transactions On Knowledge And Data Engineering, Vol. 24, No. 1, January 2012.

[8] Jian Ma, Wei Xu, Yong-Hong Sun, Efraim Turban, Shouyang Wang, And Ou Liu. "An Ontology-Based Text-Mining Method To Cluster Proposals For Research Project Selection". Ieee Transactions On Systems, Man, And Cybernetics—Part A: Systems And Humans, Vol. 42, No. 3, May 2012

[9] Disputant Relation-Based Classification For Contrasting Opposing Views Of Contentious News Issues Souneil Park, Jungil Kim, Kyung Soon Lee, And Junehwa Song. 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] Samiksha Chakule, Ashwini Borse, Shalaka Jadhav, Dr. Mrs. Y.V. Haribhakta. "Literature Survey on Knowledge Based Information Retrieval on Web". International Journal of Engineering Science and Innovative Technology (IJESIT) Volume 3, Issue 3, May 2014 352.

[14] Wenhai Sun, Bing Wang, Ning Cao, Ming Li, Wenjing Lou, Y. Thomas Hou, and Hui Li. "Verifiable Privacy-Preserving Multi-Keyword Text Search in the Cloud SupportingSimilarity-Based ranking ieee transactions on parallel and distributed systems, vol. 25, no. 11, november 2014.