Mining Multiple Text Sequence with Key Management

G.V Sam Kumar
Assistant professor
Sathyabama University,
India

A. Angel Princes
M.Tech Student
Sathyabama University,
Chennai-India

R.Karthiga
M.Tech Student
Sathyabama University,
Chennai-India

T.Rajesh
M.Tech Student
Sathyabama Universty
Chennai-India

ABSTRACT
A Text stream is a sequence of chronologically ordered documents, being generated in various forms. Multiple text streams that are correlated to each other by sharing common topics. Our aim is to extract the knowledge of the text stream from the listed documents. In particular, vulnerabilities could include compromise of data security and loss of information which leads to data leakage. To provide a data security and privacy a key management is used. Documents from different sequences about the same topic may have different time stamps termed as asynchronous. Here we first, us e Apriori Algorithm to extract the common topics for the search text from the given data set based on the time stamps using Timestamp-Based Protocols. We also use vormetric encryption algorithm, which combines Encryption and integrated key management to protect and control access to sensitive files on file servers. Second, Ranking is involved in both admin side and user side of mining work which is based on usability of documents.

Keywords
Mining multiple text sequence, Ranking, key management.

1. INTRODUCTION
Data mining is the knowledge discovery from the general data. The gathering and formulating knowledge from data using pattern extraction methods is referred as the data mining. Inevery day lacks of data's are updated in WWW. That data's are stored in huge data base called as the data-warehousing. In that huge data base we need to extract the exact information using some techniques is called data mining.

Text mining is the new born of data mining. In Text Mining, patterns are extracted from natural language text rather than data bases.That means, automatically extracting information from a usually large amount of different unstructured textual resources. Text mining is determined by the computer new, unfamiliar data by automatically extract data from various resources. Text mining is unfamiliar from well-known within web search. The problem is brash onside all the material that currently is not applicable to your needs in order to find the applicable information. The goal of the text mining is to find the unknown data that are not available. Ranking can be used as a scoring function which score the web pages or the document.

For focusing the document and selects a term from the document for which we like to get information that have high quality. Data security is one important aspects in the data mining is data security when the large Amount of data found in the data pool may contain important data .in order to provide security to that type of data need to give some protection like key management and data masking. This data leakage can be prevent by using M-Score method.

2. RELATED WORK
In Time information for the task of establishing and pursue topics in time-stamped text data is the viewing of recent studies.[9]Probalistic topic modeling is the new method that are used to open up and detect the structure of the data.[2] several of topic models that are most closely related to the Entity Topic Model (ETM).Latent Dirichlet Allocation (LDA)is one of the most well-known topic models(solution is)(A Gibbs sampling-based algorithm is proposed to learn the model).[6]Temporal text mining is the task of discovering and summarizing the growth patterns of themes in a text stream. This new text mining problem and present general probabilistic methods for solving this problem through discovering latent themes from text; constructing an evolution graph of themes and analysing life cycles of themes.[11]Bursty events detection in a text stream, where a text stream is a sequence of chronologically ordered documents, and a hot bursty event is a minimal set of bursty features that occur together in certain time windows with strong support of documents in the text stream.

Timestamp uses the timestamp protocol which provide the information about the file uploaded and the user history of the file. For the transaction of the file from admin to the user can be done with the help of apriori algorithm which is the algorithm for the transaction of the file or document from the admin to the user. The solution for this problem is new novel parameter free probabilistic approach, called feature-pivot clustering.[12]In Wikipedia, each and every article describes one concept. So the one concept is usually described in multiple languages, each language corresponding with one article.[6] All documents associated with one concept are similar in their topics. To solve this problem we use topic modelling algorithms to mine multilingual topics from Wikipedia. [8] Assigning a score that represents the sensitivity level of the data that a user is exposed to, the misuse ability weight can determine the extent of damage to the organization if the data is misused.

Using this information, the organization can then take appropriate steps to prevent or minimize the damage.we collected a group of English and Chinese Web pages from Open Directory Project (ODP) website. A major obstacle to fully integrated deployment of statistical learners is the assumption that data sits in a single table, even though most real-world.[7] Databases have complex relational structures. To solve this problem, an integrated approach to building regression models from data stored in relational databases. The solution inductive logic programming (ILP) methods.[]The problem is to preserving the conceptual similarities and eliminating the speed when increasing the accuracy.[7] To improve this a data clustering algorithm were developed with concept of k-means algorithm. In this work solutions to problems such as high dimensionality and scalability associated with existing techniques of mining web documents on the web were provided by proposing an improved data clustering algorithm. [3]The problem is inferring and modelling topics in series of documents with well-known publication dates.

The documents are given in time series are distinguished as the topic. The problem is solved by using the method such as efficient variation Bayesian (VB) inference and topic over time.[5]Rank box were introduced for the complex relationship on the semantic web. Which provide efficient ranking method for the complex relationship in mining.Automatically rank according to the user preferences. Improve the capture of the user preference

3. EXISTING METHOD

Retrieving the particular data from the WWW is possible. In that retrieval users are facing lot of difficulties. In case of retrieving specific content they retrieving the related content but not exact content because of the data replica and newly uploaded data is available. From the report of the researchers of the US says that once in 15 days the amount of data are getting doubled in www. For this problem the scientist were analysis some of the methodologist from topic mining research (year).some of the technologies were introduced .some of the technologies has some best solution like ETM, LDA.some of the method has some drawbacks.

One of the best and popular methodology is top cat in data mining .in his method they focusing the topic based retrieval searched. In a pool of text some of the text need to eliminate the connectivity word because that words should not present in no.of times in that group because of connecting the two words which don't have the meaning that standalone this is called the pre-processing steps. After pre-processing is done topic based retrieval which is nothing but we have scan each and every word in that pool. After the scanning need to count repetition of words and rank accordingly. In that ranking order which is having topmost level order treated as topic for particular pool .If anyone will search related to that topic it will retrieval that particular topic pool data will retrieved .This is one of the best method for the topic mining based retrieval. But it have some drawback that is time-consuming for mining the topic and the performance is done by seeing the old data and the new knowledge cannot be update to the users.

Another drawback is the retrieval data is not protected and there leads to the data leakage that may happen in the organization.

4. OUR WORK

Topic retrieval is the method of retrieving the particular topic from the large amount of topic pool which make the users to retrieve the topic easily To overcome this the proposed system provide the security for the retrieval data and the time consuming is reduced by ranking the document based on the user feedback. Knowledge is becoming essential resources which provides advantage and giving rise to knowledge management. Manyorganization collect large amount of data pool and stored it in the database but it is difficult to find the valuable information that are hidden in the data pool. Similarly the web document that are found in the web contain lot of valuable information which give provide high quality knowledge to the users.

But it is difficult to find the valuable knowledge from the large amount of data pool. The information or the data that are found in the data pool may have some important data that are used in the organization to provide security to those information is very important our work is to overcome this problems of the existing system where the knowledge is taken from the high quality topic which contain the quality information this quality information can be retrieved from the data pool with the help of ranking where the user can rank the quality document so the other user can search the document based on the ranking. The human use to retrieve the knowledge based on the importance of the topic the important of the topic can be find by ranking each topic based on the content of that document .To fetch the relevant information we use the ranking algorithm.The ranking algorithm is used to calculate the frequency usage of documents. In this paper proposing the ranking is RA-SVM ranking.

In the RA- SVM ranking is calculate the user's relevance i.e., who is viewing the document and how many times the document is viewed and download. This is used for calculating the frequency of the usage of the document frequency. In a particular document the list of words are indexed. Based on the iteration of the words and word frequency the document is retrieved and also the topic that contain the document is retrieved based on the time stamp with the help of time stamp protocol.for example here we take the data pool as the text file document.This is done when the admin upload the files to each domain of the organization the files uploaded in the encrypted form in order to maintain privacy of the files. The user of the each domain provided with the unique key to open the files when the user gives the keyword for example when the user gives the keyword java the java related files will be retrieved and the user can find the searched topic based on the ranking according to the user feedback and based on the timestamp. The timestamp will be based on when the files were uploaded and edited according to the ranking to give more quality data to the specific topic and get the knowledge.

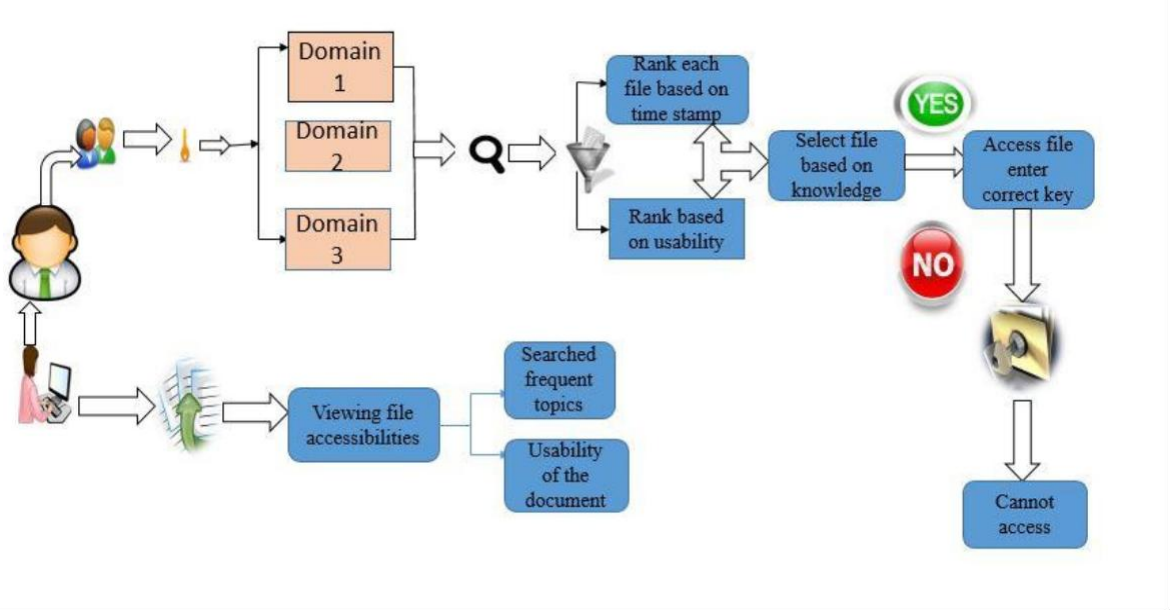

Figure: 1 Flow Diagram of Mining Multiple Text Sequence with Key Management

After retrieval of the quality topic the user must use the key to decrypt the encrypted files and can view the files, download the files accordingly. The admin of the organization can view the usability of the file that are highly ranked the performance is increased. Each and every time the user will get the new knowledge based on the uploaded and edited files. The security to the files or document can be given in the form of encrypted files where it is important for the organization to maintain the security to the document. The method used for the security is the vormetric encryption method where the document of the common topic uploaded in encrypted form. Data security is the one important aspects in the data mining is data security when the large amount of data found in the data pool that contain important data.so it is need to provide security. The ranking is done based on the user feedback and usability of the files both admin and users can view the ranks of the document the admin can view the rankings and can improve the document. Timestamp uses the timestamp protocol which provide the information about the file uploaded and the user history of the file. For the transaction of the file from admin to the user can be done with the help of apriori algorithm which is the algorithm for the transaction of the file or document from the admin to the user. This Apriori algorithm is used to retrieve the document after the document is ranked by the RA-SVM based on the dataset which is in the form of text files collection on various topic.

5. ALGORITHM

$\text{Apriori}(T, \epsilon)$

$L_1 \leftarrow \{\text{large } 1-\text{itemsets}\}$

$k \leftarrow 2$

while $L_{k-1} \neq \textit{emptyset}$

$C_k \leftarrow \{a \cup \{b\} \mid a \in L_{k-1} \wedge b \in \bigcup L_{k-1} \wedge b \notin a\}$

for transactions $t \in T$

$C_t \leftarrow \{c \mid c \in C_k \wedge c \subseteq t\}$

for candidates $c \in C_t$

$count[c] \leftarrow count[c] + 1$

$L_k \leftarrow \{c \mid c \in C_k \wedge \ count[c] \geq \epsilon\}$

$k \leftarrow k + 1$

return $\bigcup_k L_k$

Apriori is one of the classic algorithm for learning association rules. Apriori is designed to operate on databases containing transactions (for example, collections of items bought by customers, or details of a website frequentation).The major goal of the algorithm is to extract useful information from large amounts of data. Association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases. Other algorithms are designed for finding association rules in data having no transactions and timestamp.

6. TERM FREQUENCY

Term frequency is the mathematical calculation for finding no.of words in the document it is used as a weighting factor for the information retrieval and text mining. Here is the mathematical calculation for the term frequency.

Logarithmically scaled frequency,

$$tf(t,d)=\log(t,d+1) \text{ ----------Equation(1)}$$

Augmented frequency.to prevent a bias towards longer document, given by

$$Tf(t,d)= \frac{0.5+0.5*f(t,d)}{Max\{f(w,d):w€d\}} \text{ --------Equation(2)}$$

The inverse frequency of the document can be find by calculating the total no.of document divided by no.of terms in the document, given by

$$Idf(t,d) = \frac{\log|D|}{|\{d€D:t€D\}|} \text{ --------Equation(3)}$$

With d, total no.of documents in the corpus

|{d€D:t€d}|→number of documents where the term t .

Then tf-id is calculated as,

$$t.fidf(t,d,D)=t.f(t,d)*idf(t,D) \text{ ------Equation(4)}$$

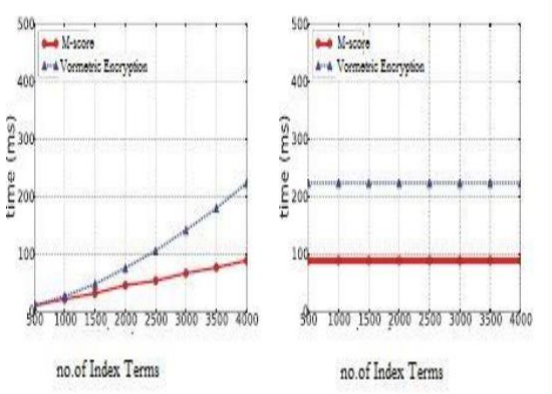

Figure2: Comparison Vormatric Encryption with M-Score

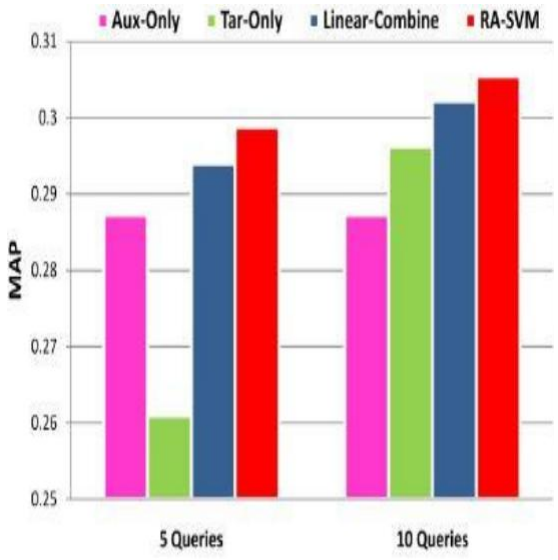

Figure3: Comparison RA-SVM with Other Ranking

7. CONCLUSION

From the graph e consider the ranking and encryption from existing methodologies.For the ranking provide the better result using RA-SVM.with the existing methodologies.Ra-SVM is a technique for getting more relevant data.In existing ranking Aux-only,Tar- only,Linear combine in this they retrieve the relevant data but not very effective.But in RA-Svm we get very relevant data for the users knowledge .In encryption graph he security level is more when compare to the existing.So, this paper conclude the combination of RA-SVM,Apriori and vormetricencryption,These three concept retrieve more relevant or more exact document from the collection of database with secure manner.

8. ACKNOWLEDGE

We would like to thank Sathyabama University for giving us a platform to enhance our knowledge. We would like to express our special thank those who motivated us to prepare this paper and peachy guidance to this paper and also we would like to express our deepest thanks to all those who made us possible to complete this work.

9. REFERENCES

[1] Gabriel Pui Cheong Fung, Jeffrey Xu Yu, Philip S. Yu, Hongjun Lu "Parameter Free Bursty Events Detection in Text Streams" In Proceedings of the 31st international conference on Very large data bases (2005), pp. 181-192 Key: citeulike:8947028}

[2] Hyungsul Kim, Yizhou Sun, Julia Hockenmaier and Jiawei Han"ETM: Entity topic models for mining documents associated with entities".Data Mining(ICDM),2012 IEEE 12th International conference on digital object,2012

[3] IulianPruteanu-Malinici, Lu Ren, John Paisley, Eric Wang and Lawrence Carin"Hierarchical Bayesian Modeling of Topics in Time-Stamped Documents"Pattern Analysis and Machine Intelligence,IEEE Transaction on volume:32,issue:6,2010.

[4] LoulwahAlSumait, Daniel Barbar´a, Carlotta Domeniconi "On-Line LDA: Adaptive Topic Models for Mining Text Streams with Applications to Topic Detection and Tracking"Data Mining,(ICDM)'08.Eigth International conference on Digital Object ,2008.

[5] Na Chen"Rank box: An Adaptive Ranking System for Mining Complex Semantic Relationships Using User Feedback"Information Reuse and Integration(IRI),2012 IEEE 13th International Conference on Digital Object,2012

[6] Qiaozhu Mei "Discovering Evolutionary Theme Patterns from Text An Exploration of Temporal Text Mining"KDD '05 Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, 2005

[7] RamchandraYenape&SharvariGovilkar"New Data Clustering Algorithm for Mining Web Documents"International Journal on Advanced Computer Theory and Engineering (IJACTE), ISSN (Print) : 2319 – 2526, Volume-1, Issue-1, 2012

[8] K.Sundaramoorthy ,Dr.S.Srinivasa Rao Madhane"Efficient Method of Detecting Data Leakage Using Misusability Weight Measure"International Journal of Computational Engineering Research||Vol,03||Issue,4|

[9] Thomas HofmannInternational, Berkeley,"Probabilistic Latent Semantic Indexing" Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval CA &EECS Department, CS Division, UC

[10] Xing Yi and James Allan"Evaluating topic models for information retrieval"CIKM '08 Proceedings of the 17th ACM conference on Information and knowledge management,2008.

[11] Xuanhui Wang, ChengXiangZhai, Xiao Hu, Richard "Mining Correlated Bursty Topic Patterns from Coordinated Text StreamsProceedingKDD '07 Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, 2007.

[12] Xiaochuan Ni, Jian-Tao Sun, Jian Hu, Zheng Chen "Mining Multilingual Topics from Wikipedia", WWW'09 Proceedings of the 18th international conference on World wide web,2009.