Asian Journal of Computer Science and Technology
ISSN: 2249-0701 (P) Vol.8 No.S1, 2019, pp.41-45
© The Research Publication, www.trp.org.in
DOI: https://doi.org/10.51983/ajcst-2019.8.S1.1961

Mining Sequential Pattern of Data in Textual Document Using Data Mining Classification Technique

J. Jayasudha[1] and A. Christina Esther[2]
[1&2]Assistant Professor, Department of Computer Science,
Kovai Kalaimagal College of Arts and Science, Coimbatore, Tamil Nadu, India
E-Mail: jayasudhayuvaraaj@gmail.com,christinaesther1920695@gmail.com

(Received 11 December 2018; Revised 21 December 2018; Accepted 31 January 2019; Available online 8 February 2019)

***Abstract* - Text document were transmitted over the internet for the text communication. So they were occurred many problems like repeated text occurred because of same data were provided in the internet. To characterize and extracting that is a most critical task for the researchers. Many researchers were characterized and applied in many fields like real-life scenarios, such as real-time monitoring on abnormal user behaviors, etc. In this case to detect and characterize the personalized behavior of the user were provide some drawbacks. To solve this problem, this paper analyzing the sequential data and characterize the user behavior with the help of the data mining sequential pattern matching algorithm.**

***Keywords:* Text Mining, Textual Document, Sequential Analyses, Personalize and Abnormal Behavior**

I. INTRODUCTION

Today data mining is the major research area for every researcher especially in textual mining field. Extracting of the sequential information in a huge database is an indispensable for every person who is published their document in internet. Data Mining (DM) is the skillful of resolve expensive information from the large database. The purpose of the DM is to discover information and present it in an understanding that is simply understandable to the people. Knowledge detection in database is detailed method which presenting a number of functional, relevant information. Data mining, or information detection, is the procedure of excavation and investigate enormous sets of data and then eradicate the connotation of the data Jain, Nikita, and Vishal Srivastava[1]. In DM they were many perform are used like sequential analyses, association rule, prototype matching, feature extraction, predictive analyses etc. With the support of the DM analytical analyses it mines the possible result in every field like web mining, text mining, spatial mining, etc. for our research here used the textual document to mine the related and sequential features. With the speedy growth of text communication accessible in the internet, user desires to share the information in both formatted and unformatted manner. In the practice of formatted text, it has the extra features like text style, size, color, etc. but in the form of unformatted text it is like a plain text. To extort the relevant feature of text is called a text mining apparatus, Text mining is the discovery of attractive acquaintance in text documents. It is a difficult issue to discover amazing knowledge (or features) in text documents. Text mining is an information discovery method that manages to pay for computational astuteness. Whereas mining the text the user have to scrutinize the text content in the text document. The text analysis is one of the vital factors in this real world to scrutinize the associated features. Text Analytics coherent a set of languages, mathematics, and machine learning methods that shape and form the information rewarded of textual cause for commerce intelligence Padhy, Neelamadhab, Dr Mishra, and RasmitaPanigrahi [2]. Text mining is the discovery of amazing knowledge in text documents. It's intricate matter to find out suitable data in text documents to help users to search for out what they want. It is a demanding work to use those sketches and also obtain them up to date. Earlier appearance based methods are make available by Information Retrieval (IR) techniques. All appearance based process tolerates from troubles such as polysemy and meanings. When a word has a diversity of meanings, it is known as polysemy. When diverse words have the correspondent meaning, it is called synonymy Zhu, Jiaqi, KaijunWang[3]. Therefore the semantic meaning of a collection of discovered terms are unpredictable for act in response what users crave. The foundation of text mining is to development formless data, mine consequential information in text collection. Information can be obtains by threat to gain rundown for the vocabulary restricted in the documents. Hence, we can inspect documents and conclude comparison between them. Text mining also referred as text data mining, roughly corresponding to text analytics, it refers to perform of obtain high superiority of information from text and towering quality of information is consequential from end to end devising of patterns. Text analysis occupies information recovery, lexical scrutiny, word happening distributions, pattern recognition, information mining and significant feature extraction G.Chandrashekar and F. Sahin[4]. Also recognized as knowledge discovery, data mining is to execute of penetrating for patterns in necessities of data. Data Mining, on the other hand, is the extraction of tricky to comprehend or hidden sequential information from huge databases or data warehouses Han, Jiawei, Jian Pei [5],To this end, data mining uses computational techniques from information and sequence analyses. Seeming for prototype in data thus defines the scenery of data mining. A sequence analyses replica is built by data mining tools and techniques. Sequential analytics is used to conclude the possible outlook result of an experience or the likelihood of a conditions happening in textual document. It is the separation of data mining come inside reach of with the approximation of future probability and trends.

II. LITERATURE REVIEW

Yuefeng Li *et.al* [6] analyze the existing popular text mining and classification processes have assumed term based approaches. Nevertheless, they have all experience from the problems of polysemy and synonymy. Over the years, people have recurrently held the submission that pattern-based process should execute higher than term-based ones in unfolding user partiality but many carrying out tests do not sustain this hypothesis. The revolutionary technique obtainable in paper makes go through for this difficulty. This method establish both positive and negative patterns in text documents as superior level features in classify to candidly weight low-level features (terms) based on their specificity and their distributions in the greater level features. Extensive carrying out tests using this technique on Reuters Corpus Volume 1 and TREC topics show that the proposed approach severely outperforms both the state-of-the-art term-based methods underpinned.

NingZhong*et.al* [7] examine many data mining system have been anticipated for mining useful prototype in text documents. Nevertheless, how to effectively use and notify exposed prototype is still an open research issue, predominantly in the domain of text mining. While most presented text mining technique implicit term-based approaches, they all experience from the troubles of polysemy and synonymy. Over the years, people have frequently held the proposition that outline (or phrase)-based approach should implement superior than the term-based ones, but many carrying out tests do not hold this hypothesis. This paper shows an ingenious and successful pattern detection technique which includes the progression of example deploying and pattern growing, to advancement the success of using and modernize discovered patterns for judgment relevant and motivating information. Noteworthy experiments on RCV1 data compilation and TREC topics make obvious that the projected enlightenment achieves heartening performance.

C.Kanakalakshmi and Dr. R.Manickachezian [8] examine the pulling out of practical information from shapeless textual data through the gratitude and penetrating of motivating patterns. The detection of applicable features in real-world data for connecting user information requirements or partiality is a new confront in text mining. Significance of a feature designate that the features is at all times essential for an most constructive subset, it cannot be undisturbed without offensive the pioneering provisional class sharing. They proposed an adaptive method for relevance feature recognition is conversed to find useful features accessible in a feedback set, as well as both positive and negative documents, for performance what users need. Thus, this paper talk about the methods for importance feature discovery using the simulated annealing rough calculation and genetic algorithm, a inhabitants of applicant solutions to an optimization problem on the way to better solutions.

HarpreetKaur, Rupinder Kaur *et .al* [9] observed that the text mining using the pattern sighting usually uses only the text material in typical fonts i.e. it does not believe the bold, underline or italic or smooth the larger fonts as the key text sample for text mining. This generates difficulty numerous a times when the key words are eliminate from the commentary by the algorithm itself. In that case, momentous keywords are left from the most significant stream of text patterns. In their projected work, patterns are exhumes in both positive and negative reaction. It then unconsciously classifies the patterns into clusters to find applicable patterns as well as destroy noisy patterns for a given topic. A novel prototype organizing approach is proposed to remove alternative features of text documents and use them for improving the retrieval performance. The projected move toward is appraised by remove features from RF to advancement the presentation of information filtering (IF).

Muthuvalli.A.R and Manikandan.M [10] study the feature clustering method to instinctively group terms into the three group positive detailed features, general features, and negative detailed features. The first issue in using irrelevant documents is how to choose a apposite set of irrelevant documents since a very large set of negative example is typically get hold of For example, a Google Search can revisit millions of documents; though, only a few of those documents may be of concentration to a Web user. Perceptibly, it is not well-organized to use all of the irrelevant documents. This demonstration is a supervised advance that needs a training set including both relevant documents and irrelevant documents. It also makes available suggestion for wrongdoer (irrelevant) compilation and the use of detailed stipulations and all-purpose terms for presentation user information needs. This model finds both positive and negative advice and the RFD used unimportant documents in the preparation set in order to eradicate the noises and also it can achieve the reasonable performance.

III. SEQUENTIAL ANALYSES FOR TEXTUAL DOCUMENT

Sequential pattern mining is the mining of co- occurring of text occurred in the text document. This mining task is used to mine the repeated text occurred in the document. It is the selection of a subset of feature used to signify the data. In text categorization it spotlight on recognizing sequential information lacking stirring the correctness of the classifier. It is used to find the text with the help of the feature selection methods. In text documents feature can be term, pattern, and disapproval. However, the traditional feature assortment methods are not successful for choosing text features for responding the relevance subject because significance is an introverted class problem Tobji, MA Bach, [11]. For solving the problem this paper provides a sequential analyzer algorithm to find co incidence of data for this analyzing we used some mechanism they are.

1. Data Preprocessing

The data processing task is also one of the criteria which must be taken care in the process of data mining. The data input to a data mining algorithm need not be in proper format and is hence not suitable for processing efficiently. In such a case, we need to see the data is in proper format so that it is suitable for processing. This case generally arrives when we try to mine the data using the existing data mining tools or algorithmsR. Agrawal and R. Srikant [12]. Different Data mining tools available in the market have different formats for input which makes the user forced to transform the existing input dataset into the new format. This itself is very time consuming, laborious and has a chance of data loss as the data is to be entered manually into a new format to be supported by the tool. For this preprocessing the Apriori algorithms were used to discover intra-transaction of associations. However the sequence mining task is defined as discovering inter-transaction associations – sequential patterns – across the same or similar data.

```
INPUT: A dataset D and a support threshold s
OUTPUT: All sets that appear in at least s transactions of D F is set of frequent item sets
C is set of candidates        C ← U
Scan database to count support of each item in C
Add frequent items to F
Sort F least-frequent-first (LFF) by support (using quick sort)
Output F
for all f ∈F, sorted LFF do
for all g ∈F, supp(g) ≥ supp(f), sorted LFF do
Add {f, g} to C
end for
Update index for item f
end for
while |C| > 0 do
{Count support}
for all t ∈D do
for all i ∈t do
Relevant Cans ← using index, compressed cans from file that start with i
for all Compressed Cans∈Relevant Cans do
if First k − 2 elements of Compressed Cans are in t then
Use compressed candidate support counting technique to update appropriate
support counts
end if; end for; end for; end for
Add frequent candidates to F
Output F
Clear C
{Generate candidates}
Start ← 0
for 1 ≤ i ≤ |F| do
if i == |F| OR fi is not near-equal to fi−1 then
Create super candidate from fstart to fi−1 and update index as necessary
Start ← i  end if;        end for
{Candidate pruning—not needed!}
Clear F; Reset hash; end while
```

Algorithm 1 Frequent pattern growth with Apriori algorithm

2. Text Categorization

Text categorization (or text classification) is the assignment of natural language documents to predefined categories according to their content. In this phase have to gather the text and to categorize it according to the attributes which is to be set in the phase of the given element which is occurred in the sequential onX. Li and B. Liu Srikant, Ramakrishnan[13, 14]. Automatic text categorization has many practical applications, including indexing for document retrieval, automatically extracting metadata, word sense disambiguation by detecting the topics a document covers, and organizing and maintaining large catalogues of Web resources.

3. Sequential Discovery

Sequential patterns are the sequences whose support exceeds the minimal support which is defined by the user. Sequences of events, items or tokens occurring in an ordered metric space appear often in data and the requirement to detect and analyze frequent subsequences is a common problem. Sequential Pattern Mining arose as a sub-field of data mining to focus on this fieldLodhi, Sanjaydeep Singh, PremnarayanArya, and DilipVishwakarma[15]. For these analyses, this paper used the Frequent pattern growth (FP-Growth) type algorithms are often regarded as the fastest item enumeration algorithms.

```
Sequential node selection (node n = (s1, ....,sk), Sn , In)
Begin
(1)  Stemp = φ.
(2)  Itemp = φ.
(3)  For each (i Sn)
(4)  if ((s1, ....., sk , {i}) is frequent)
(5)  Stemp = Stemp {i}
(6)  For each (i Stemp)
(7)  DFS-Pruning((s1,…………...,sk,{i}),Stemp,
             all elements in Stemp greater than i )
(8)  For each (i In)
(9)  if ((s1, ..... , sk » {i}) is frequent)
(10)  Itemp = Itemp {i}
(11)  For each (i Itemp)
(12)  DFS-Pruning ((s1, ......,sk {i}), Stemp,
             all elements in Itemp greater than i)
End
```

Algorithm 2 Sequential pattern

FP-Growth generates a compressed summary of the data set using two passes in a cross referenced tree, the FP-tree, before mining item sets by traversing the tree and recursively projecting the database into sub-databases using conditional FP-Trees. In the first database pass, infrequent items are discarded, allowing some reduction in the database size. In the second pass, the complete FP-Tree is built by collecting common prefixes of transactions and incrementing support counts at the nodes. at the same time, a header table and node links are maintained, which allow easy access to all nodes given an item. The mining step operates by following these node links and creating (projected) FPTrees that are conditional on the presence of an item (set), from which the support can be obtained by a traversal of the leaves. This technique will provide better result when compared to others.

Input: A transaction database DB and a minimum support threshold ξ .
Output: FP-tree, the frequent-pattern tree of DB.**Step 1:** Scan the transaction database DB once.
Step 2: Collect F, the set of frequent items, and the support of each frequent item.
Step 3: Sort F in support-descending order as FList, the list of frequent items.
Step 4: Create the root of an FP-tree, T , and label it as "null". For each transaction Trans in DB do the following.
Step 5: Select the frequent items in Trans and sort them according to the order of FList.
Step 6: Let the sorted frequent-item list in Trans be [p | P], where p is the first element and P is the remaining list. Call insert tree ([p | P], T). The function insert tree ([p | P], T) is performed as follows. If T has a child N such that N.item-name = p.item-name, then increment N's count by 1; else create a new node N, with its count initialized to 1, its parent link linked to T , and its node-link linked to the nodes with the same item-name via the node-link structure. If P is nonempty, call insert tree(P, N) recursively.

Algorithm 3 FP-tree construction

IV. EXPERIMENTAL RESULTS

In the experimental analysis is used to analyze the sequential pattern extraction in the textual document. The proposed method where implemented in the field of internet area to analyze the frequent data over the internet. Our proposed system is very helpful to find out the perfect result in the case of time, accuracy, frequent data, etc,.Ourproposed methodology improvedthe accuracy, time etc and is used to create the experimental setup. The performance of the proposed method is evaluated in terms of,

1. Accuracy
2. Time and cost management
3. Packet Delivery Ratio
4. Energy consumption

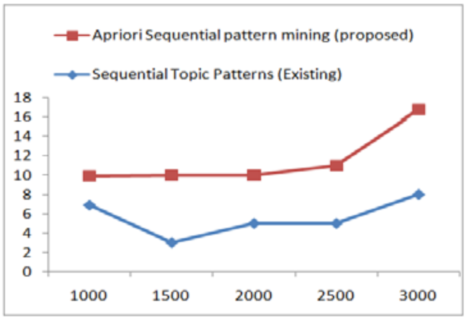

Fig. 1 Accuracy Improvement

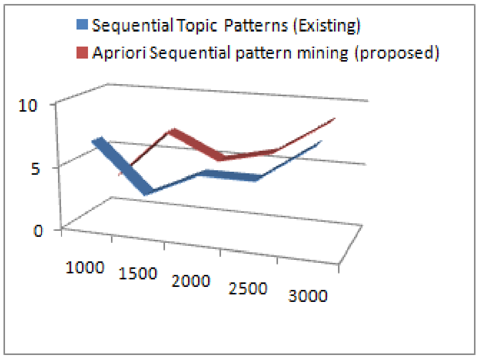

Fig.2 Time and cost analyses

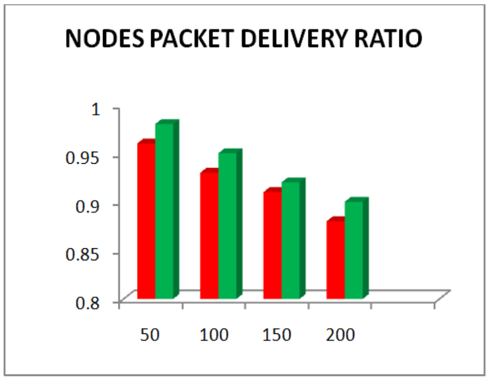

Fig.3 Packet Delivery Ratio

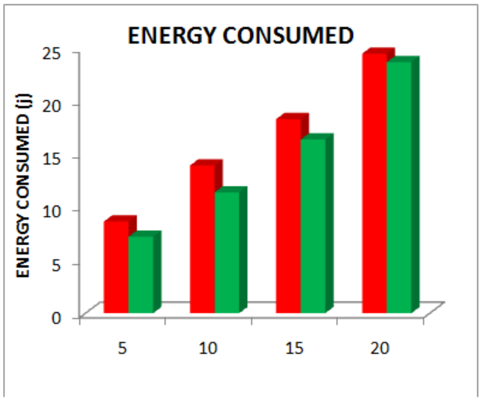

Fig.4Energy consumption

V. CONCLUSION

Text mining is an important research area in every field. Mining the text is useful over the internet. While extracting the text sequential occurrence of data where occurred. So they were many problem arise due to this issues. They were many methods were found to solve this issues. Many researchers were did many thing and some provide better and some provide failure in some of the case. So solve this issues here this paper provide analyzing the sequential data and characterize the user behavior with the help of the data mining sequential pattern matching algorithm used with the help of the Frequent pattern growth algorithm to show the effective result for sequential analyses of the data.

REFERENCES

[1] Jain, Nikita, and Vishal Srivastava, "Data Mining techniques: A survey paper", *IJRET: International Journal of Research in Engineering and Technology*,Vol. 2, No. 11, pp. 2319-1163,2013.

[2] Padhy, Neelamadhab, Dr. Mishra, and RasmitaPanigrahi, "The survey of data mining applications and feature scope", *arXiv preprint arXiv*, pp. 1211.5723, 2012.

[3] Zhu, Jiaqi, Kaijun Wang, Yunkun Wu, Zhongyi Hu, and Hongan Wang, "Mining User-Aware Rare Sequential Topic Patterns in Document Streams", *IEEE Trans. Knowl. Data Eng,* Vol. 28, No. 7, pp. 1790-1804, 2016

[4] G. Chandrashekar and F. Sahin, "Asurvey on feature selection methods," in Comput. Electr. Eng., vol. 40, pp. 16–28, 2014.

[5] Han, Jiawei, Jian Pei, and MichelineKamber, "Data mining: concepts and techniques", *Elsevier*, 2011.

[6] Li, Yuefeng, AbdulmohsenAlgarni, and NingZhong, "Mining positive and negative patterns for relevance feature discovery", In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 753-762. ACM, 2010.

[7] Zhong, Ning, Yuefeng Li, and Sheng-Tang Wu, "Effective pattern discovery for text mining", *IEEE transactions on knowledge and data engineering,* Vol. 24, No. 1, pp. 30-44, 2012.

[8] Kostoff, and N.Ronald, "Method for data and text mining and literature-based discovery", U.S. Patent 6,886,010, issued April 26, 2005.

[9] T. A. Pawar,., and N. D. Karande, "Effective Pattern Discovery For Text Mining Using Pattern Based Approach", *International Journal of Advance Research in Computer Science and Management Studies, ISSN*, pp. 2321-7782,2014.

[10] Loh, Stanley, Leandro Krug Wives, and José Palazzo M. de Oliveira, "Concept-based knowledge discovery in texts extracted from the web", *ACM SIGKDD Explorations Newsletter*, Vol. 2, No. 1, pp. 29-39, 2000.

[11] Tobji, MA Bach, B. Ben Yaghlane, and KhaledMellouli, "A new algorithm for mining frequent itemsets from evidential databases", In *Proceedings of IPMU*, Vol. 8, pp. 1535-1542. 2008.

[12] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In Proc. ACM SIGMOD Int. Conf. on Management of Data, Minneapolis, MN, 1994.

[13] X. Li and B. Liu, "Learning to classify texts using positive and unlabeled data,"*in Proc. 18th Int. Joint Conf. Artif. Intell,* pp. 587–592, 2003.

[14] Srikant, Ramakrishnan, and RakeshAgrawal, "Mining sequential patterns: Generalizations and performance improvements", In *International Conference on Extending Database Technology*, pp. 1-17. Springer, Berlin, Heidelberg, 1996.

[15] Lodhi, Sanjaydeep Singh, PremnarayanArya, and Dilip Vishwakarma, "Frequent Itemset Mining Technique in Data Mining", *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)* Vol. 1, No. 5, pp.395-402, 2012.