International Journal of Advanced and Innovative Research (2278-7844) /

Volume 7 Issue 4

A Document Pattern Analysis over the document stream using Enhance NN Architecture

Pooja Singh[#1], Asst. Prof. Nitesh Gupta

[#]*NRI College, CS Department, RGPV University*
India

[1]poojaitech92@gmail.com

***Abstract*-- Document annotation and its availability for the user is important in many application area. Application such as research labs, student labs where continuous document availability and release perform need document investigation. Document writing style gives different pattern with it. Pattern analysis gives understanding of document and its knowledge extraction. There are two type of patterns are available with document which is topical and sequential. Majorly document follows topical pattern which identify in easy manner. Rare document do follows sequential pattern, thus the detection techniques are not available. A research field to detect sequential pattern document is need to investigate. In this paper a survey related to document pattern detection is discussed. Various techniques and pattern analysis over the document stream performed is discussed. Existing author performed user aware rare sequential pattern approach. Related work performed research over twitter real time & synthetic dataset and thus shows the efficiency of their approach. Our further work is to find optimal technique for sequential pattern analysis for document stream in real time dataset.**

***Keywords*-- Document annotation, sequential pattern, topical pattern, data modelling, and analysis approach.**

I. INTRODUCTION

Due to explosive growth of data representation in various formats, a demand is increasing of storing, searching and retrieving document stream data day by day. Large amounts of research have been carried out in document stream and analysis from long time. Document stream gives two type of annotation which is annotation with topical analysis and sequential analysis. Topical analysis gives the relation between the document and its writing style. Different topic and relation in words, their suggested terms is defined by researcher. The traditional approaches for image retrieval can be topical related mining and analysis. In this techniques set of documents are indexed and retrieved by document topic level features like word, its segment and given approach etc. Next technique which is investigate sequential pattern which is pattern relevant to sequence analysis. Here user provides the annotation to the document and then these documents were searched and retrieved later on by specifying text. After manual document annotation topical relation can be retrieved as text documents.

Document Annotation refers to the process of automatically labelling the image by predefined keywords which represent the semantic of data. Word Annotation is done using a Dataset which is primarily known as Twitter dataset etc. Annotated document have an advantage of text based searching. Thus document annotation aims invest large amount of pre-efforts to annotate the data search as accurately as possible to support the image search [1].

Pattern analysis may have the following component.

1. Document Segmentation Component
2. Document level component mining
3. Topical data modelling analysis
4. Content classification or mapping component
5. Labelling component
6. Sequential analysis

II. RELATED WORK

1. Jiaqi Zhu, Member, Kaijun Wang, Yunkun Wu, Zhongyi Hu, and Hongan Wang

In this paper [1] an algorithm rare sequential modelling and pattern analysis is performed by the proposed algorithm. STP candidate based pattern growth algorithm over the document is performed along with Dynamic programming based approach is performed. An exact probability of pattern is determined by the approximation algorithm. A research over the document pattern analysis is performed with Twitter synthetic and real time analysis. 2000 user's data is collected and other dataset with approx. 955 users and approx. 1 lac tweets are collected and observed. An Ubuntu 12.04 Operating system with java programming language is performed. Result performed with parameter as precision and accuracy parameter is performed. The limitation in this work is a limited source of data and short dataset is used for experiment which can be extend for large data. A research area specified in sequential pattern detection which is rare and extended work over topical research.

2. Philippe Fournier-Viger, Jerry Chun-Wei Lin

In this paper [2] a survey over the different sequential pattern analysis is performed. They have specified the different two sequences and time series sequential mining. The value computed either be value of ordered normal series or list of nominal values. They have also mentioned that SPMF data mining library tool is used for open source implementation. The application area of research is in analysis of social media data, developing more enhanced algorithm which are based on GPU, Depth firth search algorithm. A complex data handling strategy and finding meaningful pattern over the given document is discussed by the paper. This paper also discussed about all the previous type of rule mining which is easy to understand and to study more about the sequential pattern approach.

3. Zhou Zhao, Da Yan and Wilfred Ng

In this paper [3] a technique for pattern analysis, measure pattern frequentness based on the possible world semantics is used. An algorithm U-PrefixSpan is performed which speed up the document, which is inspired by PrefixSpan algorithm. They have also discussed traditional sequential pattern extraction which is PrefixSpan works with random variable and other support variable. Different approach such as segmentation, pruning is used for pattern analysis is performed. An issue which is sequence-level uncertain model is addressed and pattern based approach is appended with given approach. A word level, element level and document level sequence extraction and applicability over data are performed. A fast validating method with data processing from its element is used over dataset. Proposed algorithm exhibit low computation time, high number of patterns and high precision and ideal recall value. The proposed work shows the pattern detection is efficient and can be extended for the further research [3] [4].

4. Y. Li, J. Bailey, L. Kulik, and J. Pei

In this paper [5] a study is performed with spatio-temporal dataset which contains the information relation to geo information which contains a different field of research related to space and coordinates. They have focused on the uncertain sequences which need to study and grab the knowledge from them. The pattern with gap constraints is discussed and analyzed over trajectory dataset. The algorithm work with linear transform capacity and pattern detection technique over it. Pattern detection technique such as breadth first search and depth first search is used. This is efficient in pattern detection from the large data [6]. They have also worked with the synthetic and real world dataset to outperform research and show the efficiency of their approach with precision, recall and accuracy parameter using confusion matrix[7,8].

5. Z. Zhang, Q. Li, and D. Zeng

In this paper [9] topic pattern discovery over the large dataset and continues dataset which is related to communication community question and different answers. Thus a popular application web 2.0 is investigated by the paper. Approach work with topic pattern mining which extract the topic analysis from the different topic temporal data. Discovery of different topic data is extracted, extracted topic graph is analyzed by the approach. They have also discussed the life cycle of the extracted topic which helps in exact pattern detection duration and its availability for enhancement. They have proved their work efficiency over the large and real time dataset but it is limited to only topical analysis approach [10].

In all the previous approaches discussed in literature, they have worked with the previous approach either with synthetic or real time data but again worked with short text with following limitation.

1. They have either taken a small dataset for the research through which a proper result cannot be considered as true [11].
2. Large data prediction is not investigated with different document format and pattern detection technique [12].

Thus a further work to working with large dataset and its processing with different format of data is remaining.

6. Madhur Aggarwal, Anuj Bhatia

In this paper [13] author discuss about the web pattern and discovery of the content based on web data. There are different approach such that they work with pattern extraction and working with data dictionary. There are various approach which deals with the pattern extraction, algorithm such as Apriori algorithm, FP- Tree approach and further categorical fuzzy logic based approach is compared by them. They are providing various approach with advantage and disadvantage of pattern analysis approach.

7. Sheng-Tang Wu and Yuefeng Li

In this paper [14] author discussed about the various SCPM and NSCPM approach which deals with the pattern extraction and pattern detection approach. They have discussed pattern taxonomy model which is the improvement of PTM based model for the data analysis and extraction of data. They have taken RCV1 includes 806,791 document letters which is process in order to have the pattern recognition. Two approach which is SP Mining and

NSP Mining approach is processed. A closed pattern approach is performed which generates the low recall situation [15].

III. PREVIOUS APPROACH

In this chapter, previous approach and proposed approach result comparison is performed, as per the monitored results from implementation which is obtained is compared. The proposed algorithm is presented and compared with existing solution. This chapter gives a comparison graph and statically analysis. As per observed, finally it shows the proposed approach is efficient in terms of total net flows,malicious net flows, accuracy, detection rate as well in the implementation analysis.

IV. PROPOSED METHODOLOGY

The Solving set algorithm for finding the Pattern attributes which finding the desired attributes for the ranking optimization thoughtfulness. So in order to find disjoined non considered attributes going to apply the algorithm for finding the Patterns and then we will store the Patterns in a separate dataset for considering next algorithm usage.

The Lineup algorithm helps us to optimize the ranking on providing the attribute of our choice for the consideration, but again line up gives us ranking of the particular provided attributes only, so here to we perform HYBRID DSS. First, find all the relevant candidate data set in our data set, and then we will do HYBRID DSS. Visualization of the ranking will be executed and random choice option for selecting another attribute can be provided to the user for further ranking optimization.

Basic steps of hybrid DSS are as follows:

Step 1: Load data sets from the saved Database.

Step 2: Load Candidate data sets.

Step 3: Compare Candidate data set with the exiting original data set.

Step 4: Employ Hybrid algo to find Maximum no. of Pattern this will take minimum computation time as compared to Lazy Dss and DSS.

Step 5: Employ Pattern dataset over line-up algo.

Step 6: Now with the aid of multiple attribute over step 5 we best rank over the data set.

Table 1: Comparison analysis between the algorithm executions.

The table 1 below, shows the comparison between the approaches presented.

Parameters	**DSS**	**LDS**	**HYBDRIB**
Maximum Computation time	1369	789	174
Minimum Computation time	1090	452	106
Average Computation time	1106	563	129

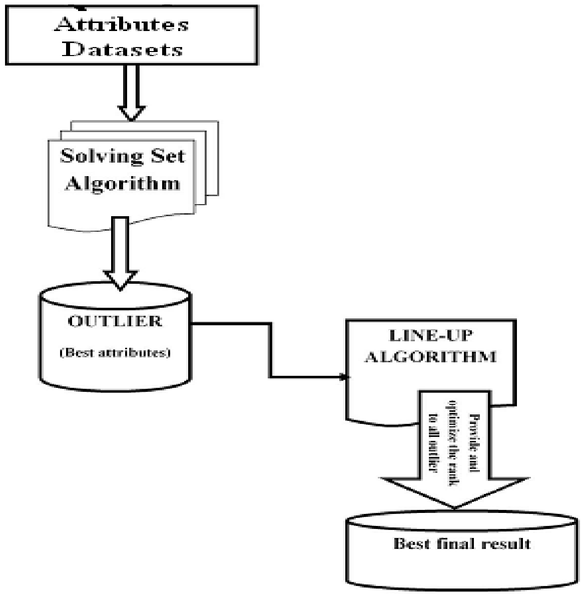

Figure 1.1: flow chart DSS with hybrid.

V. RESULT ANALYSIS

In this section, different observed result which is performed using apache framework is presented. A statically analysis and graphical analysis using the existing as well as proposed technique is presented.

Computing Parameter:

There are mainly three parameter, which is taken for the comparison analysis is taken. Computing parameter such as computation time, computation cost and bandwidth consumption is observed.

Computation Time:
Computing time is the time difference which is observed by subtracting final executing time to initial loading time. A time difference between both the times is observed and call as computation time.
Computing time = final execution time – initial time;
Ct=fet-it;

Computation Cost:
Computing cost is the total cost which can be observed by monitoring different usage resources and aspects such as bandwidth, data consumption, resources etc.
Computing cost = bandwidth consumption cost + Resources consumption cost + cost per second;
Cc= bcc + rcc+cps;

Bandwidth consumption:
It is the total data consumption per unit of time which is taken by the token and complete access monitoring.
Bandwidth consumption = total data consumption/ unit time;
Bc = tdc/ut;

Parameters	**DSS**	**LDS**	**HYBDRIB**
Precision	87.6	88	92.3
Accuracy	65.34	78	93.2
Recall	76.65	81.9	89.56

STATISTICAL ANALYSIS
In this section we will explain about the several calculations performed over different algorithms.

Table 1.1: Computation time comparison.

In the above table the computation cost has been calculated.

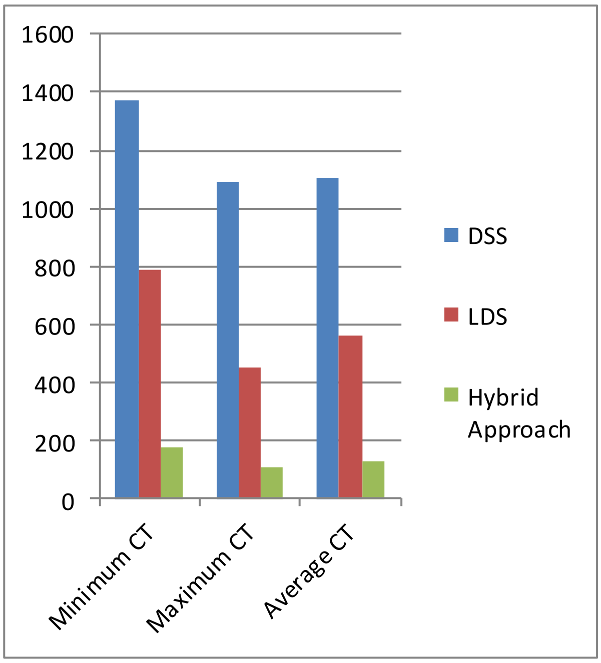

Figure 1.2: Graph Computation time Obtained.

The above graph explains about the computation cost analysis over different modes.

Table 1.2: Computation Accuracy comparison

In the above table the precision, accuracy, recall are calculated by applying different algorithms in them.

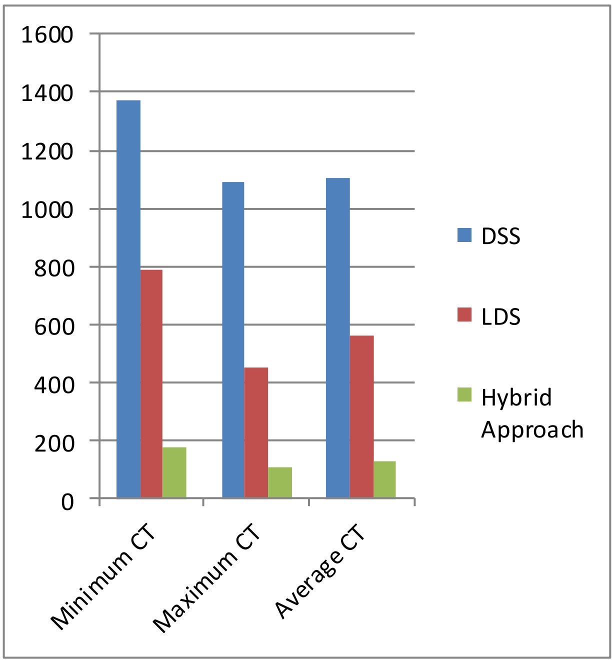

Figure 1.3: Graph Computation parameters Obtained.

In the above figure the parameters obtained by using different algorithms has been shown.

VI. CONCLUSION

By considering all features in dataset for Pattern detection. We have analyzed the result from multiple algorithms that select relevant features for the proposed frameworks. A forbes dataset of top ranking companies related dataset was used for evaluating the performance of system.

we have conducted various experiment of different algorithm and observed results, by considering all features in dataset for Pattern detection. We have analysed the result from multiple algorithms that select relevant features for the proposed frameworks. In this research work have proposed a powerful Hybrid technique in order to obtain the Pattern and we have also implemented existing Distributed solving set and Lazy distributed solving set algorithm and compare them with the hybrid technique on the basis of their computation time. We have found it best as the Hybrid perform a best less time compare with other two algorithm used for the Pattern detection.

REFERENCES

[1] E. Hung and D.W. Cheung, “Parallel Mining of Patterns in Large Database,” Distributed and Parallel Databases, vol. 12, 2002.

[2] E. Knorr and R. Ng, “Algorithms for Mining Distance-Based Patterns in Large Datasets,” Proc. 24rd Int’l Conf. Very Large Data Bases (VLDB), pp. 392-403, 1997.

[3] E. Lozano and E. Acun ˜a, “Parallel Algorithms for Distance-Based and Density-Based Patterns,” Proc. Fifth IEEE Int’l Conf. Data Mining (ICDM), pp. 729-732, 2005

[4] S.D. Bay and M. Schwabacher, “Mining Distance-Based Patterns in Near Linear Time with Randomization and a Simple Pruning Rule,” Proc. Ninth ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining (KDD), 2003.

[5] M.E. Otey, A. Ghoting, and S. Parthasarathy, “Fast Distributed Pattern Detection in Mixed-Attribute Data Sets,” Data Mining Knowledge Discovery, vol. 12, nos. 2/3, pp. 203-227, 2004.

[6] A. Koufakou and M. Georgiopoulos, “A Fast Pattern Detection Strategy for Distributed High-Dimensional Data Sets with Mixed Attributes,” Data Mining Knowledge Discovery, vol. 20, pp. 259-279, 2009

[7] H. Dutta, C. Giannella, K.D. Borne, and H. Kargupta, “Distributed Top-K Pattern Detection from Astronomy Catalogs Using the DEMAC System,” Proc. SIAM Int’l Conf. Data Mining (SDM), 2007.

[8] S. Ramaswamy, R. Rastogi, and K. Shim, “Efficient Algorithms for Mining Patterns from Large Data Sets,” Proc. Int’l Conf. Managment of Data (SIGMOD ’00), pp. 427-438, 2000.

[9] S.D. Bay and M. Schwabacher, “Mining Distance-Based Patterns in Near Linear Time with Randomization and a

Simple Pruning Rule," Proc. Int'l Conf. Knowledge Discovery and Data Mining (KDD '03), 2003

[10] E. Eskin, A. Arnold, M. Prerau, L. Portnoy, and S. Stolfo, "A Geometric Framework for Unsupervised Anomaly Detection: Detecting Intrusions in Unlabeled Data," Applications of Data Mining in Computer Security, Kluwer, 2002.

[11] Chakrabarti, S. et. al.,; "Mining the link structure of the World Wide Web;" IEEE Computer, 32(8), August 1999.

[12] Baeza-Yates,Ricardo; Davis, Emilio; "Web page ranking using link attributes," Proceedings of the 13th internationalWorld Wide Web conference on Alternate track papers & posters, May 2004.