A SURVEY ON: A DOCUMENT PATTERN DETECTION APPROACH

[1]Pooja Singh and [2]Nitesh Gupta
[1]Research Scholar, [2]Assistant Professor
[1,2]RGPV NRI, Bhopal, India

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.

IndexTerms **- Document annotation, sequential pattern, topical pattern, data modeling, analysis approach.**

I. INTRODUCTION AND MOTIVATION

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 labeling 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 modeling analysis
4. Content classification or mapping component
5. Labeling component
6. Sequential analysis

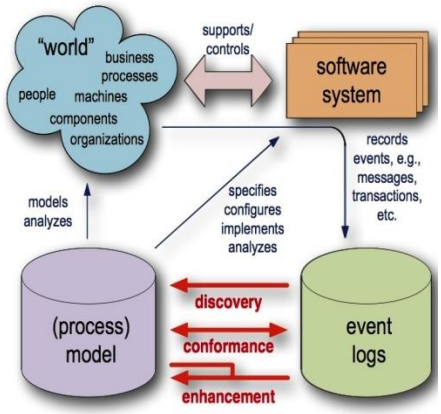

Figure 1: Process flow diagram of mining and document handling scenario

In the figure 1 above, a complete processing scenario of process mining, document sharing, storage and accessing system is given. This figure also shows the flow performed by the data mining, document process in system.

II. LITERATURE REVIEW

A. 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 extending for large data. A research area specified in sequential pattern detection which is rare and extended work over topical research.

B. 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.

C. 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].

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

In this paper [5] a study is performed with spatial 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].

E. 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.

- They have either taken a small dataset for the research through which a proper result cannot be considered as true [11].
- 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.

F. MadhurAggarwal, Anuj Bhatia

In this paper [13] author discuss about the web pattern and discovery of the content based on web data. There is different approach such that they work with pattern extraction and working with data dictionary. There is 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 approaches with advantage and disadvantage of pattern analysis approach.

G. 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 approaches which is SP Mining and NSP Mining approach are processed. A closed pattern approach is performed which generates the low recall situation [15].

Table 2.1: Comparison analysis of previously given technique.

Sr. No.	Author	Algorithm	Advantage	Remark
1.	Jiaqi Zhu, Member, Kaijun Wang, Yunkun Wu, Zhongyi Hu, and Hongan Wang	STP Candidate based FP growth approach	Fast execution and rapid mining technique	Execution is fast but long iterations were used.
2.	Philippe Fournier-Viger, Jerry Chun-Wei Lin	Sequential pattern analysis	Sequential data is considered than topical relation.	Topical data and sequential both can combine with the approach.
3.	Zhou Zhao, Da Yan and Wilfred Ng	UPrefix span technique	A Prefix data analysis detection is performed as quick process.	Data outcome may get wrong values.
4	Y. Li, J. Bailey, L. Kulik, and J. Pei	Spatial temporal analysis	Enable to process large dataset.	Quick processing.

In the table 1 above, a comparison analysis and technique overview performed by existing author is discussed.

III. Conclusion

Social data and different document arise in different industry. Various documents get a writing style of topical analysis. Document over a particular topic gives the understanding of word and relation in between data. Topical relation understanding is common and performed by multiple techniques in previous research. A rare sequence which is sequential pattern detection is performed by less research organization.In recent base paper work URSTPs approach is discussed which works with rare sequential approach pattern analysis. Previous technique such as pattern FP growth, rank analysis, and LDA and CTM approach is discussed here to understand document pattern analysis. In this paper multiple paper survey is discussed which use in analysis of topical and sequential search, document analysis. Recent search shows the effectiveness of URSTPs computes high precision and accuracy parameters. Further work is going to enhance CTM with large twitter dataset.

References

[1] Jiaqi Zhu, Member, IEEE, Kaijun Wang, Yunkun Wu, Zhongyi Hu, and Hongan Wang, "Mining User-Aware Rare Sequential Topic Patterns in Document Streams", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 28, NO. 7, JULY 2016.

[2] Philippe Fournier-Viger, Jerry Chun-Wei Lin, Rage UdayKiran, "A Survey of Sequential Pattern Mining", Data Science and Pattern Recognition c 2017 ISSN XXXX-XXXX Ubiquitous International Volume 1, Number 1, February 2017.

[3] Z. Zhao, D. Yan, and W. Ng, "Mining probabilistically frequent sequential patterns in large uncertain databases," IEEE Trans. Knowl. Data Eng., vol. 26, no. 5, pp. 1171–1184, May 2014.

[4] X. Yan, J. Guo, Y. Lan, and X. Cheng, "A biterm topic model for short texts," in Proc. 22nd Int. Conf. World Wide Web, 2013, pp. 1445–1456.

[5] Y. Li, J. Bailey, L. Kulik, and J. Pei, "Mining probabilistic frequent spatio-temporal sequential patterns with gap constraints from uncertain databases," in Proc. IEEE 11th Int. Conf. Data Mining, 2013, pp. 448–457.

[6] Yuening Hu, Jordan Boyd-Graber, Brianna Satinoff, and Alison Smith. Interactive topic modeling. Machine Learning, 2013.

[7] SanjeevArora, RongGe, Yoni Halpern, David M. Mimno, AnkurMoitra, David Sontag, Yichen Wu, and Michael Zhu. A practical algorithm for topic modeling with provable guarantees. In ICML, 2013.

[8] J. Chae, D. Thom, H. Bosch, Y. Jang, R. Maciejewski, D. S. Ebert, and T. Ertl, "Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition," in Proc. IEEE Conf. Vis. Anal. Sci. Technol., 2012, pp. 143– 152.

[9] Z. Zhang, Q. Li, and D. Zeng, "Mining evolutionary topic patterns in community question answering systems," IEEE Trans. Syst., Man, Cybern. A, vol. 41, no. 5, pp. 828–833, Sep. 2011.

[10] M. Muzammal, "Mining sequential patterns from probabilistic databases," in Proc. 5th Pacific-Asia Conf. Adv. Knowl. Discovery Data Mining, 2011, pp. 210–221.

[11] T. Bernecker, H.-P. Kriegel, M. Renz, F. Verhein, and A. Zuefle, "Probabilistic frequent itemset mining in uncertain databases," in Proc. ACM SIGKDD, 2009, pp. 119–128.

[12] N. R. Mabroukeh and C. I. Ezeife, "A taxonomy of sequential pattern mining algorithms," ACM Comput. Surv., vol. 43, no. 1, pp. 3:1–3:41, 2010.

[13] Jason Chuang, Sonal Gupta, Christopher D. Manning, and Jeffrey Heer. Topic model diagnostics: Assessing domain relevance via topical alignment. In ICML, 2013.

[14] MadhurAggarwal, Anuj Bhatia," Pattern Discovery Techniques in Online Data Mining", International Journal of Engineering and Technical Research (IJETR) ISSN: 2321-0869 (O) 2454-4698 (P), Volume-3, Issue-7, July 2015.

[15] Sheng-Tang Wu and Yuefeng Li, "Pattern-Based Web Mining Using Data Mining Techniques", International Journal of e-Education, e-Business, e-Management and e-Learning, Vol. 3, No. 2, April 2013.

[16] https://alchetron.com/Process-mining