A STUDY ON SEQUENTIAL TOPIC PATTERNS MINING ALGORITHMS AND APPROACHES IN WEB LOGS

[1]S. TAMILARASI, [2]DR.V.VIDYA PRIYA,
[1]M.PHIL RESEARCH SCHOLAR, [2]ASSOCIATE PROFESSOR,
PG and Research Department of Computer Science,
[1]**QUAID-E-MILLATH ARTS & SCIENCE COLLEGE FOR WOMEN (AUTONOMOUS)**
CHENNAI, INDIA

ABSTRACT: SEQUENTIAL TOPIC PATTERNS (STP) is an important topic of data mining task in wide range of applications from content analysis of the web and document stream. It is usually consists of discovering interesting sub sequences in *a set of sequences.* ***The resultant Sequences can be measured in terms of various criteria such as its occurrence frequency, length, and profit. STP has various real-life applications because the fact of that data is naturally programmed as*** *sequences of codes* ***in many fields like bioinformatics, e-shopping, document streams, and social media. In this paper we study both the accuracy and the efficiency of STP mining algorithms and approaches.***

Keywords: Sequential Topic Patterns (STP), User- Aware Rare Sequential Topic Patterns (URSTP), Web Mining

---

Introduction

The World Wide Web (WWW) is the most graphically inviting and easily navigable part of the internet . It gets number of requests from the web users through the world. So it becomes necessity for administrator's to improve the quality of the Web services. web mining is the technique of extracting the required information from the World Wide Web documents and web services [5].

A. Taxonomy of Web Mining

1. Web Content Mining

Web Content Mining is the mining, extraction and integration of useful data, information and knowledge from web page content [6]. Web crawlers, Meta crawlers provide some comfort to users to categorize, filter and interpret the documents.. It mainly focuses on- Web Text mining and Web Multimedia Mining [5].

2. Web Structure Mining

Web structure mining is the process of analyzing or describing the structure of the contents of the website. using graph theory, where web pages are its nodes and its hyperlinks are its edges [7]. Thus the web links from one web page to another shows the relationship among the web and users and it focus only on the structural summary of web pages and web sites. Web structure mining mainly works on- Link mining, Internal structure mining and URL mining [5].It is further divided into 2 types : 1. Extracting patterns and 2. Mining the structure of the web document.

3. Web Usage Mining:

Web Usage Mining is used to define the mode through the users can interact with the servers or to access the available web pages. It also includes information generated by client side transaction from one or more web localities [7]. Its main objective is to find the usage patterns from web applications. It consists of three phases: preprocessing, discovery of usage pattern and analysis of the pattern. This mining method is used by server logs and its aim is getting useful users who can access information in the form of web logs[5].

ALGORITHMS AND APPROACHES OF STP:

Apriori Algorithm:

Apriori works an iterative method known as a level wise search, where n -item sets are used to explore (n+1) item sets. When Scanning first the frequent -1 item sets is found in the database to accumulate the count for each item and collecting those items that satisfy minimum support. The resulting set is denoted as L1. Next L1 is used to find L2, the frequent 2 item sets which further are used to find l3 and so on, until no more frequent n- item sets can be found. The finding of every Ln requires one full scan of the database.

SPADE (Sequential PAttern Discovery using Equivalence classes),

It is uses the equivalence classes for discovering the set of all frequent Items in the data set. The main key features of this approach are as follows:

1. It uses a vertical id-list database format, where it subordinate with each sequence of a list of objects in which it occurs with the time-stamps and it shows all frequent items can be itemized through simple temporal joins or intersections on the vertical id-lists.
2. It also uses the lattice-theoretic approach[2] which splits the original search space (lattice) into smaller pieces (sub-lattices) that can be processed separately in the core-memory. This approach usually needs three database scans, or only a free scan with the some pre-processed news, therefore it minimizing the I/O costs.
3. It again breaks the problem from the pattern search. We propose two different search strategies for enumerating the frequent sequences within each sub lattice: breadth - first and depth - first search.
4. SPADE not only minimizes I/O costs by reducing database scans, but also minimizes computational costs by using efficient search schemes.

cSPADE

Mining frequent sequential patterns with the cSPADE algorithm. This algorithm operates temporal joins along with efficient lattice search methods and keeps for timing constraints.

FREE SPAN-

The Free Span algorithm reduces the cost require to candidate generation and testing of apriori, with satisfying its basic feature [8]. In short, it uses the frequent items to iteratively plan the sequence database into task database while developing subsequence's frequently in each task dataset. Every task divides the database and limits further testing to gradually smaller and manageable units. The important issue is to sizable amount of sequences can appear in more than single task database and the size of database decreases with each iteration.

PREFIX SPAN :

An algorithm for finding sequential patterns in sequence databasesis the only projection based algorithms from all the sequencing pattern mining algorithms. It performs better than the other algorithms like apriori, Free Span, SPADE. It input is a sequence database and a user-specified threshold named *minsup*. It output is finds all frequent sequential patterns occurring in a sequence . The main concept behind it is to successfully discovered patterns is employing the divide and-conquer approach. This algorithm requires huge memory space as compare to the other algorithms in the sense that it requires creation and processing of huge number of projected sub-databases. To explain the sequential pattern, it is necessary to review some definition.

A sequential pattern is a sequence. A sequenceSA = M1, M2, ... Mi, where M1, M2, ... Mi are itemsets is said to occur in another sequence SB = N1, N2, ... Nj, where N1, N2, ... Nj are itemsets, if and only if there exists integers k <= k1 < k2... < km <= m such that M1 ⊆ Nk1, M2 ⊆ Nk2, ... Xi ⊆ Ykj.

The support value of a sequential pattern is the number of sequences that the pattern occurs divided by the total number of sequences in the database.

A frequent sequential pattern is a sequential pattern having a support value should not less than the *minsup* parameter value provided by the user.

For example, if we run Prefix Span with *minsup*= 50 % and with a maximum pattern length of 100 items, 53 sequential patterns are found. The list is too big to show here. An example of pattern found is "(1,2),(6)" which appears in the first and the third sequences (it has therefore a support value of 50%). This pattern has a length of 3 because that contains 3 items. Another pattern is "(4), (3), (2)". It appears in the second and third sequence (it has thus a support value of 50 %). It also has a length of 3 because that contains 3 items.

WAPMINE:

Web access pattern mining from sequential database set which contains web access click in the sequential format by timestamps.

Sequential linear programming (SLP)

SLP algorithms are used successfully in structural organization design .This class of methods is well suited for solving large nonlinear problems due to the fact that it does not require the computation of second derivatives, so the iterations are cheap. SLP algorithms that are proved to be globally convergent are seldom adopted in practice than many other algorithms.

Nonnegative Matrix Factorization (NMF) Approach

Sometimes **NMF** approach isalso called Non Matrix Approximation (NMA)[8] which is a group of algorithms in multivariate analysis and linear algebra where a matrix X is factorized into (usually) two matrices Y and Z, with the property that all three matrices have no negative elements. This non-negativity makes the resulting matrices easier to inspect. Also, in applications such as processing of audio spectrograms or powerful activity, non-negativity is essential to the data being considered.

Since this NMF approach analyses all the data together; i.e., the whole matrix is available from the start. This may be unacceptable in applications where there are too many data to fit into memory or where the data are provided in streaming fashion. One such use is for common filtering in recommendation systems, where there may be many users and many items to recommend, and it would be inefficient to recalculate everything when one user or one item is added to the system. The cost function for optimization in these cases may or may not be the same as for standard NMF, but the algorithms need to be rather different

NMF finds applications in such fields as astronomy, computer vision, document clustering, chemo metrics, audio signal processing, recommender systems and bioinformatics

Smooth Non-Negative Matrix Factorization (SNMF)

- Smooth Nonnegative Matrix Factorization (SNMF)[9] for event detection, by fully leveraging information from query semantics, temporal correlations, and search log records.
- We use the SNMF method rather than the normal NMF method or other MF method to guarantee that the weights for each topic are non-negative and consider the time factor for event development at the same time.
- The basic idea is two-fold: 1) promote event queries through by strengthening their connections based on all available features; 2) differentiate events from popular queries according to their temporal characteristics.

Conclusion

Sequential Topic Pattern mining in document stream and web logs has various techniques. In this paper we discussed overview of the STP and URSTP and which algorithms mostly used for STP Mining . Best Suitable Algorithms are determined for web logs and Data Streams

REFERENCES

1 C. H. Mooney and J. F. Roddick, "Sequential pattern mining approaches and algorithms," ACM Comput. Surv., vol. 45, no. 2, pp. 19:1–19:39, 2013

2. 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

3. Yunkun Wu, Zhongyi Hu, and Hongan Wang, Member, IEEE "Mining User-Aware Rare Sequential Topic Patterns in Document Streams" 2016.
4. C Ding, T Li, MI Jordan, Convex and semi-nonnegative matrix factorizations, IEEE Transactions on Pattern Analysis and Machine Intelligence, 32, 45-55, 2010
5. Brijendra Singh, Hemant Kumar Singh,"Web Data Mining Research: A Survey", IEEE, 2010
6. Arvind Kumar Sharma, P.C. Gupta,"Study & Analysis of Web Content Mining Tools to Improve Techniques of Web Data Mining", International Journal of Advanced Researchin Computer Engineering & Technology, Vol. 1, Issue 8, October 2012.
7. K. Mohammad Mujahid, et al. , "Web Mining: Day-Today",International Journal of Emerging Trends and Technology in Computer Science, Vol. 3, Issue 5, Sept-Oct, 2014
8. Lee, D., Seung, H., et al.: Learning the parts of objects by non-negative matrix factorization. Nature 401 (1999) 788–791 9.
9. Ding, C., Li, T., Jordan, M.: Convex and semi-nonnegative matrix factorizations. Pattern Analysis and Machine Intelligence, IEEE Transactions on 32 (2010) 45–55
10. CM. Chin and R. Fletcher, On *the global convergence of an SLP-filter algorithm that takes EQP steps.* Math. Program., **96** (2003), 161-177.
11. Viger, P., Lin, J. C.-W., Kiran, R. U., Koh, Y. S., Thomas, R. (2017). **A Surveyof Sequential Pattern Mining**. Data Science and Pattern Recognition, vol. 1(1), pp. 54-77.