International Journal of Latest Trends in Engineering and Technology
Special Issue SACAIM 2016, pp. 544-548
e-ISSN:2278-621X

MODIFIED PATTERN-BASED WEB MINING USING DATA MINING TECHNIQUES

Prajwal Poonja[1], M Madhana Kumar[2], Abhishek Kelkar[3] and Mrs. Annapoorna Shetty[4]

Abstract-Many data mining techniques have been proposed for fulfilling various knowledge discovery tasks in order to achieve the goal of retrieving useful information for users. Data mining techniques include association rule mining, frequent itemset mining, sequential pattern mining, maximum pattern mining and closed pattern mining. However, how to effectively exploit the discovered patterns is still an open research issue, especially in the domain of Web mining. We compare these data mining methods based on the use of several types of discovered patterns.
Keywords –Web mining, Knowledge Discovery, Data mining, Pattern Taxonomy Model.

I. INTRODUCTION

Web Mining can be broadly defined as the discovery and analysis of useful information from the World Wide Web. Web Mining is the process of using data mining technique and algorithm to extract information directly from the web documents and services, web content, server logs. Web mining and Data mining helped discovery of useful Information and Knowledge.Market Analysis and business management can be benefited by information extracted from large amount of data.

Knowledge discovery is the process of discovering useful knowledge from a collection of data. It includes data preparation and selection, data cleansing, incorporating prior knowledge on data sets and interpreting accurate solutions from the observed results.The significant number of data mining techniques have been presented in order to perform different knowledge tasks. These techniques include association rulemining, frequent item set mining, sequential pattern mining, maximum pattern mining and closed pattern mining. The World Wide Web provides rich information on an extremely large amount of linked Web pages. Such a repository contains not only text data but also multimedia objects, such as images, audio and video clips. Data mining on the World Wide Web can be referred to as Web mining which has gained much attention with the rapid growth in the amount of information available on the internet. Web mining is classified into several categories, including Web content mining, Web usage mining and Web structure mining.Most Web text mining methods use the keyword-based approaches, whereas others choose the phrase technique to construct a text representation for a set of documents. It is believed that the phrase-based approaches should perform better than the keyword-based ones as it is considered that more information is carried by a phrase than by a single term.Although phrases carry less ambiguous and more succinct meanings than individual words, the likely reasons for the discouraging performance from the use of phrases are: (1) phrases have inferior

[1] *Aloysius Institute of Management and Information Technology (AIMIT) St Aloysius College (Autonomous) Mangaluru, Karnataka, India*
[2] *Aloysius Institute of Management and Information Technology (AIMIT) St Aloysius College (Autonomous) Mangaluru, Karnataka, India*
[3] *Aloysius Institute of Management and Information Technology (AIMIT) St Aloysius College (Autonomous) Mangaluru, Karnataka, India*
[4] *Aloysius Institute of Management and Information Technology (AIMIT) St Aloysius College (Autonomous) Mangaluru, Karnataka, India*

statistical properties to words, (2) they have a low frequency of occurrence, and (3) there are a large number of redundant and noisy phrases among them [4], [5]. In order to solve the above mentioned problem, new studies have been focusing on finding better text representatives from a textual data collection. One solution is to use the data mining techniques, such as sequential pattern mining, for building up a representation with the new type of features [6]. Such data mining-based methods adopted the concept of closed sequential patterns and pruned non-closed patterns from the representation with an attempt to reduce the size of the feature set by removing noisy patterns. However, treating each multi-terms pattern as an atom in the representation seems likely to encounter the low-frequency problem while dealing with the long patterns [7]. Another challenge for the data mining-based methods is that more time is spent on uncovering knowledge from the data; consequently less significant improvements are made compared with information retrieval methods [8].

II.PROPOSED ALGORITHM

A. SEQUENTIAL PATTERN MINING (SPM) ALGORITHM

The Sequential Pattern Mining (SPM) algorithm's main purpose is to eliminate the non-closed patterns during the process of sequential patterns discovery. This is done by applying the pruning scheme. Recursive algorithm plays role in the first line of algorithm, which describe the pruning procedure. In this algorithm, all (*n-1*) *Terms* of length patterns are diagnosed to determine whether or not they are closed patterns after all *nTerms* of length patterns are generated from the previous recursion.Recursively algorithm repeats until there is no more pattern discovered**.** As a result, the output of algorithm Sequential Pattern Mining is a set of closed sequential patterns with relative supports greater than or equal to a specified minimum support.

Method:

1: $SP \leftarrow SP - \{Pa \in SP \mid \exists Pb \in PL$ Such That $len(Pa) = len(Pb) - 1$ And $Pa < Pb$ And $supp_d(Pa) = supp_d(Pb)\}$
   //Pattern Pruning
2: $SP \leftarrow SP \cup$ PL *//nTerms pattern set*
3: $PL' \leftarrow \phi$
4: **For Each** *pattern p* In *PL* **Do Begin**
5: generating *p-projected* database *PD*
6: **For Each** frequent *term t* In *PD* **Do Begin**
7: $P' = p \oplus t$ //sequence extension
8: **If** $supp_r(P') \geq min\_sup$ **Then**
9: $PL' \leftarrow PL' \cup P'$
10: **End If**
11: **End For**
12: **End For**
13: **If** $|PL'| = 0$ **Then**
14: Return //no more pattern
15: **Else**
16: Call SPMining(PL', *min_sup*)
17: **End If**
18: Output *SP*

B.GSP ALGORITHM

GSP algorithm (*Generalized Sequential Pattern* algorithm) is an algorithm used for sequence mining. It works by counting the occurrences of all singleton elements in the database. Then, the transactions helps in removing the non-frequent items. At the end, each transaction consists of only the frequent elements it originally contained. This modified database becomes an input to the GSP algorithm. Process requires one pass over the entire database. Multiple database passes are also made by GSP algorithm.

```
Method:
F1 = the set of frequent 1-sequence
  k=2,
  do while F(k-1)!= Null;
     Generate candidate sets Ck (set of candidate k-sequences);
        For all input sequences s in the database D
           do
        Increment count of all a in Ck if s supports a
        Fk = {a ∈ Ck such that its frequency exceeds the threshold}
           k = k+1;
        Result = Set of all frequent sequences is the union of all Fks
           End do
  End do
```

C.NON SEQUENTIAL PATTERN MINING ALOGRITHM

Non Sequential Patterns from a set of textual documents is another application of the data mining mechanism. It can be treated as frequent itemsets extracted from a transactional database.

Algorithm: Non_SPMining(*NP*, *FT*, *mini_sup*)
Input: NP - To Non Sequential patterns a list of NTerms;
FT - Frequent Patterns list of 1Term,
mini_sup - Minimum Support.

Output: FP - Frequent Non Sequential Patterns set.

Algorithm: Non_SPMining(*NP*, *FT*, *mini_sup*)

Input: NP - To Non Sequential patterns a list of NTerms;

FT - Frequent Patterns list of 1Term,

mini_sup - Minimum Support.

Output: FP - Frequent Non Sequential Patterns set.

Method:

1: *FP*←*FP* È NP

2: *NP* ←NULL

3: **For Each** pattern *p* in *NP* **Do Begin**

4: **For Each** frequent term *t* in *FT* **Do Begin**

5: *P'*= *p* È {t} //pattern growing

6: **If** *supor*(*P*) ³ *mini_sup* **Then**

7: *NP* ←*NP* È*P*

End If

End For

8: **End For**

9: **If** | *NP*| = 0 **Then**

10:Return

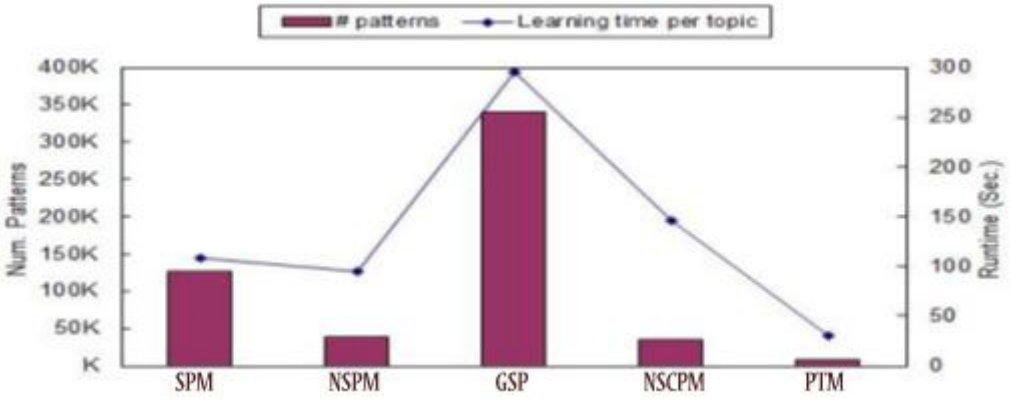
11:**Else**

12: Call Non_SPMining(*NP'*, *FT*, *mini_sup*)

15: **End If**

16: Output *FP*

III.RESULTS



IV.CONCLUSION

In general, a significant amount of patterns can be retrieved by using the data mining techniques to extract information from Web data. Many data mining techniques have been proposed in the last decade. Our progress in this study is just by comparing GSP with SCPM and NSCPM. A comprehensive comparison of data mining methods applied for Web mining task is performed in this study. The experimental results show that closed pattern methods, such as GSP,SCPM and

NSCPM, have better performance due to the use of pruning mechanism in the pattern discovery stage.

REFERENCES

[1] V. Devedzic, "Knowledge discovery and data mining in databases," in *Handbook of Software Engineering and Knowledge Engineering*, vol. 1, 2001, pp. 615-637.

[2] W. J. Frawley, G. Piatetsky-Shapiro, and C. J. Matheus, "Knowledge discovery in databases: an overview," *AI Magazine*, vol. 13, pp. 57-70,1992.

[3] D. D. Lewis, "An evaluation of phrasal and clustered representations on a text categorization task," *SIGIR*, 1992, pp. 37–50.

[4] F. Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys*, vol. 34, no. 1, pp. 1-47, 2002.

[5] N. Zhong, Y. Li, and S. T. Wu, "Effective Pattern Discovery for Text Mining," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 1, pp. 30-44, 2012.

[6] S. T. Wu, Y. Li, Y. Xu, B. Pham, and P. Chen, "Automatic pattern-taxonomy extraction for web mining," in *Proc. IEEE/WIC/ACMInternational Conference on Web Intelligence*, 2004, pp. 242-248.

[7] S. T. Wu, Y. Li, and Y. Xu, "An effective deploying algorithm for using pattern-taxonomy," in *Proc. iiWAS05*, 2005, pp. 1013-1022.

[8] S. T. Wu, Y. Li, and Y. Xu, "Deploying approaches for pattern refinement in text mining," in *Proc. ICDM*, 2006, pp. 1157-1161.