International Journal of Latest Trends in Engineering and Technology
Special Issue SACAIM 2016, pp. 422-425
e-ISSN:2278-621X

PATTERN DISCOVERY FOR TEXT MINING USING PATTERN TAXONOMY

Sandesh Shetty[1]and Shaikh Obaid Ahmed[2], Hemalatha N[3]

Abstract-This Paper presents an innovative and effective pattern discovery technique which includes the process of pattern deploying and pattern evolving, to improve the effectiveness of using and updating discovered patterns for finding relevant and interesting information. Many other techniques have been developed for mining useful patterns in text documents. This paper represents an effective way in finding patterns with minimum support and confidence.

Keywords –Text mining, Knowledge Discovery, Data mining, Sequential Pattern, Pattern Taxonomy Model.

I. INTRODUCTION

Text mining is the discovery of interesting knowledge in text documents. It is challenging issue to find accurate knowledge in text documents to help users to find what they want. Many applications, such as market analysis and business management, can benefit by the use of the information and knowledge extracted from a large amount of data. Knowledge discovery can be effectively use and update discovered patterns and apply it to field of text mining.

Knowledge discovery can be viewed as the process of nontrivial extraction of information from large databases, information that is implicitly presented in the data, previously unknown and potentially useful for users.Therefore Data mining is an important step in the process of knowledge discovery in databases. 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[1][2][3].

This paper represents an effective way in finding patterns with minimum support and confidence. Algorithms used sequential pattern, closed sequential pattern and vertical mining of sequential generation of patterns. Algorithms uses a particular large datasets. Datasets which consists of different data can be used to retrieve information through it.

II.METHOD FOR FINDING PATTERNS

A. Sequential Pattern Mining Using Co-occurrenceInformation(SPAM) algorithm

The Sequential Pattern Mining (SPAM) algorithm using a vertical representation are the most efficient for mining sequential patterns in dense or long sequences, and have excellent overall performance. The vertical representation allows generating patterns and calculating their supports without performing costly database scans[4].

[1] *Aloysius Institute of Management And Information Technology(AIMIT),Mangalore, Karnataka, India.*
[2] *Aloysius Institute of Management And Information Technology(AIMIT),Mangalore, Karnataka, India.*
[3]*Aloysius Institute of Management And Information Technology(AIMIT),Mangalore, Karnataka, India.*

Algorithm:
SPAM(SDB, minsup)
1. Scan SDB to create V(SDB) and identify F1, the list of frequent items.
2. FOR each item s € F1,
3. SEARCH(s, F1, {e € F1 | e >lex s}, minsup).
SEARCH(pat, Sn, In, minsup)
1. Output pattern pat.
2. Stemp :=Itemp := Ø
3. FOR each item j € Sn,
4. IF the s-extension of pat is frequent THEN Stemp :=StempÙ{i}.
5. FOR each item j €Stemp,
6. SEARCH(the s-extension of pat with j, Stemp , {e €Stemp | e >lex j}, minsup).
7. FOR each item j €In,
8. IF the i-extension of pat is frequent THEN Itemp :=ItempÙ{i}.
9. FOR each item j €Itemp,
10. SEARCH(i-extension of pat with j, Stemp , {e €Itemp | e >lex j}, minsup).

B. VGEN(Vertical Mining of Sequential Generator Patterns)

VGEN is fast Vertical Mining of SequentialGenerator Patterns. Sequential pattern mining is a popular data mining task with wide applications. However, the set of all sequential patterns can bevery large. To discover fewer but more representative patterns, several compact representations of sequential patterns have been studied. Theset of sequential generators is one the most popular representations. It was shown to provide higher accuracy for classification than using allor only closed sequential patterns[5]. Furthermore, mining generators is a key step in several other data mining tasks such as sequential rulegeneration. However, mining generators is computationally expensive. To address this issue, they propose a novel mining algorithm named VGEN which we have used to generate results.
(Vertical sequential GENerator miner).

Algorithm:
PATTERN-ENUMERATION(SDB, minsup)
1. Scan SDB to create V(SDB) and identify Sinit, the list of frequent items.
2. FOR each item s ∈ Sinit,
3. SEARCH((s), Sinit, the set of items from Sinit that are lexically larger than s, minsup).
SEARCH(pat, Sn, In, minsup)
1. Output pattern pat.
2. Stemp :=Itemp :=Ø
3. FOR each item j ∈ Sn,
4. IF the s-extension of pat is frequent THEN Stemp :=Stemp ∪ {i}.
5. FOR each item j ∈ Stemp,
6. SEARCH(the s-extension of pat with j, Stemp , elements in Stemp greater than j, minsup).
7. FOR each item j ∈ In,
8. IF the i-extension of pat is frequent THEN Itemp :=Itemp ∪ {i}.
9. FOR each item j ∈ Itemp,
10. SEARCH(i-extension of pat with j, Stemp , all elements in Itemp greater than j, minsup).

C. Clospan Algorithm(Mining Closed Sequential Pattern)

Instead of mining the complete set of frequent subsequences, we mine frequent closed sub sequences, only i.e..,
those containing no super sequence with the same support (i.e., occurence frequency). By exploring novel global optimization techniques, an efficient algorithm called **CloSpan**(Closed Sequential pattern mining)is used in this paper[6 7 8]. Moreover **CloSpan**can mine really long sequences, which to the best of our knowledge, is un-minable by other discussed algorithms. Finally **CloSpan**produces a significantly less number of discovered sequences than the traditional(i.e..full set)methods while preserving the same expressive power since the whole set of frequent subsequences, together with their supports, can be derived easily from our mining results.

Algorithm:
Clospan(s, Ds, min_sup, L)
Input: A sequence s, a projected DB Ds, and min_sup.
Output: The prefix search lattice L.
1. Check whether the discovered sequence s' exists s.t.
either s $\underline{C}$s' or s' $\underline{C}$s and I(Ds)=I(Ds');
2. If such super pattern or sub pattern exists then
3. Modify the link in L;
4. else insert s into L;
5. Scan Ds once, find every frequent item α such that
    (a) s can be extended to (s Δi α), or
    (b) s can be extended to (s Δs α);
6. If no valid α available then
7. return;
8. FOR EACH valid α do
9. callClospan(s Δi α, Ds Δi α, min_sup, L);
10. for each valid α do
11. callClospan(s Δs α, Ds Δs α, min_sup, L);
12. return;

III.RESULTS

Algorithms which we discussed above provides an effective way for mining text. Algorithms where performed on a particular dataset to generate results. Generated Results show support count, maximum memory used by algorithms, Number of patterns generated and support count in Table 1. From the Table we have SPAM Algorithm having the minimum memory compared to other algorithms but has only 15 frequent patterns sets generated. On the other hand, though VGEN has memory little more than SPAM but has 16 frequent pattern sets generated with support count 4. This experiment shows that VGEN may be the better Algorithm for mining text compared to SPAM and Clospan. This values may differ according to different datasets and its features.

Table 1 Experimental Results

METHOD	Frequent Pattern Set	Minimum Support	Support Count	Max Memory(mb)
SPAM	15	0.5	2	1.28013
VGEN	16	0.5	4	1.31114
Clospan	14	1	3	1.88812

IV.CONCLUSION

In this research work, an effective pattern discovery technique has been used to overcome the low-frequency and misinterpretation problems for text mining[8]. VGEN shows a better result in getting high number of patterns, with minimum support, and using less memory in the system. In this paper using this algorithms we discussed that VGEN shows high efficiency in Text Mining using Pattern Taxonomy.

REFERENCES

[1] K. Aas and L. Eikvil, "Text Categorisation: A Survey," TechnicalReport Raport NR 941, Norwegian Computing Center, 1999.

[2] R. Agrawal and R. Srikant, "Fast Algorithms for MiningAssociation Rules in Large Databases," Proc. 20th Int'l Conf. VeryLarge Data Bases (VLDB '94), pp. 478-499, 1994.

[3] H. Ahonen, O. Heinonen, M. Klemettinen, and A.I. Verkamo,"Applying Data Mining Techniques for Descriptive PhraseExtraction in Digital Document Collections," Proc. IEEE Int'lForum on Research and Technology Advances in Digital Libraries(ADL '98), pp. 2-11, 1998.

[4] Ayres, J., Flannick, J., Gehrke, J., Yiu, T.: Sequential pattern mining using a bitmap representation. In: Proc. 8th ACM SIGKDD Intern. Conf. Knowledge Discovery and Data Mining, pp. 429–435. ACM (2002)

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

[6] Agrawal, R., Ramakrishnan, S.: Mining sequential patterns. In: Proc. 11th Intern.Conf. Data Engineering, pp. 3{14. IEEE (1995)

[7] S.-T. Wu, Y. Li, and Y. Xu, "Deploying Approaches for PatternRefinement in Text Mining," Proc. IEEE Sixth Int'l Conf. Data Mining (ICDM '06), pp. 1157-1161, 2006.

[8] S.-T. Wu, Y. Li, Y. Xu, B. Pham, and P. Chen, "Automatic Pattern-Taxonomy Extraction for Web Mining," Proc. IEEE/WIC/ACM Int'l Conf. Web Intelligence (WI '04), pp. 242-248, 2004.