Review Paper

Approach to Recover CSGM Method with Higher Accuracy and Less Memory Consumption using Web Log Mining

Shrivastva Neeraj and Lodhi Singh Swati
IES IPS Academy, Indore, MP, INDIA

Available online at: www.isca.in
Received 14th June 2012, revised 23rd June 2012, accepted 13th July 2012

Abstract

Sequential pattern mining is an important mining technique which discovers closed frequent sub sequence from a sequence database. However it is very difficult as it generates explosive number of sub sequence in candidate generator and test approach. Previous sequential pattern mining algorithm closed sequence-sequence generator mining (CSGM) mine full set of frequent sub sequence satisfying a min_sup and max_sup threshold in sequence database. This algorithm is not suitable for datasets that are too dense or too sparse, which is prohibitively expensive in both time and space. In this paper we analyze the existing methods of sequential pattern mining and after analysis we propose an enhance algorithm for sequential pattern mining. Thus the main purpose this method is aiming to solve is to develop new techniques based on the closure concept for effectively and efficiently discovering non-redundant sequential association rules from sequential datasets with higher accuracy, less memory and time consumption. After performance analysis we use modified algorithm for mining useful data from web log.

Keyword: Sequential pattern mining, CSGM, web mining.

Introduction

Sequential pattern mining[1] identifies sequential patterns appearing with enough support It has potential application in many areas such as analysis of market data, purchase histories, web logs, etc. Sequential rules express temporal relationships among patterns. It can be considered as a natural extension to many spurious patterns by introducing the notion of confidence to the set of patterns. Only rules satisfying both support and confidence thresholds are mined. Sequential rules extend the usability of patterns beyond the understanding of sequential data. Some examples of useful rules include: market data. If a customer buys a car, he/she will eventually buy car insurance. This is potentially useful in designing personalized marketing strategy that the patient will need a treatment for dengue fever.

These studies include those mining a compact representation of patterns, referred to as closed patterns and generators. These compact representative patterns can be mined with much more efficiency than the full set of frequent patterns.

Sequential pattern mining is the procedure to extract frequent sequences from sequential transaction databases. Many algorithms have been introduced to mine frequent sequences, closed sequences and sequence generators like CSGM (closed sequence-sequence generator mining). The CSGM algorithm is an extension of the closed sequence mining algorithm CloSpan[2].

CloSpan uses a similar projection database concept and it is an extension of the Prefix Span[3] algorithm. Instead of mining all frequent sequences, CloSpan mines only closed sequences. The Clospan algorithm consists of two stages for the first stage, the candidate set is generated using the same pruning technique as Prefix Span while incorporating an early termination condition to eliminate sequences which are unlikely closed sequences. The purpose of CSGM is to generate both sequential generators and closed sequential patterns together by scanning sequential database only once, with the hope of reducing time cost compared to generating closed sequences and generators separately. The CSGM algorithm uses a similar prefix-search-lattice data structure and the projected database concept as for CloSpan. The detailed procedure of this algorithm consists of two major steps: scanning database and mining the entire candidate set of closed. CSGM has a similar complexity to the closed sequence mining algorithm CloSpan but it is considerably faster than conducting sequential generator mining and closed sequential pattern mining separately

This paper focuses on the new way of developing non redundant association rule for numerical and nominal data set. We Propose a method that efficiently discovering non-redundant sequential association rules from sequential datasets with higher accuracy, less memory and time consumption. To simulate our complete task performances study of both algorithm we proposed web log Mining tool.

The main role of this paper is: i. Introduction of sequential pattern mining that identifies sequential patterns appearing with enough support. ii. Introduction to CSGM method that generate both sequential generators and closed sequential patterns together with the hope of reducing time cost compared to generating closed sequences and generators separately. iii. Introduction of our proposed method that is useful to mine the compressed set of non-redundant rules and show that it performs much faster and require less memory space than closed sequence-sequence generator mining. iv. Introduction of web log mining that is use to To simulate our complete task performances study of both algorithm.

Web Log Mining

Web usage mining[4] is the type of web mining activity that involves the automatic discovery of user access patterns from one or more web servers. Web usage mining is also known as web log mining. As more organizations rely on the internet and the world wide web to conduct business, the traditional strategies and techniques for market analysis need to be revisited in this context. Organizations often generate and collect large volumes of data in their daily operations. Most of this information is usually generated automatically by Web servers and collected in server access logs. Other sources of user information include referrer logs which contains information about the referring pages for each page reference, and user registration or survey data gathered via tools such as CGI scripts.

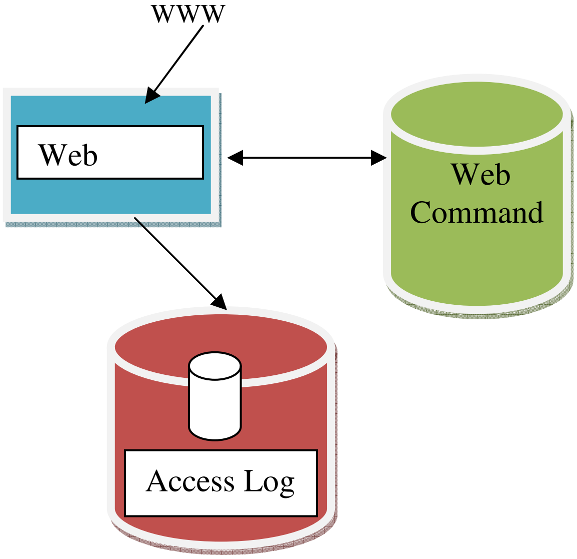

Figure-1
Web Log mining

In figure 1, web server register a log entry for every single access they get. A hug number of access are registered and collected in an ever growing web log. Here web log provide rich information about web dynamic.

Web server record access information as a click stream data into log. Whenever a user hits a page the log data is collected automatically in Web servers. It represents the accurate navigational behavior of visitors. There are valuable information in profile such as access pattern of user, the type of explore.

Introduction of Our Proposed Work

In our work we recover the problem of CSGM method. We increase accuracy of generating non redundant association rule for both nominal and numerical data with less time complexity and memory space and in this method we use N-fold cross validation technique for performance evaluation and for classification of data set we are use ID3 decision learning algorithm with some modification in entropy calculation. In our proposed method we Use $\log_n$ in the place of $\log_2$.

System Architecture

To simulate our complete task performances study of both algorithms we proposed web log mining tool that architecture are given below

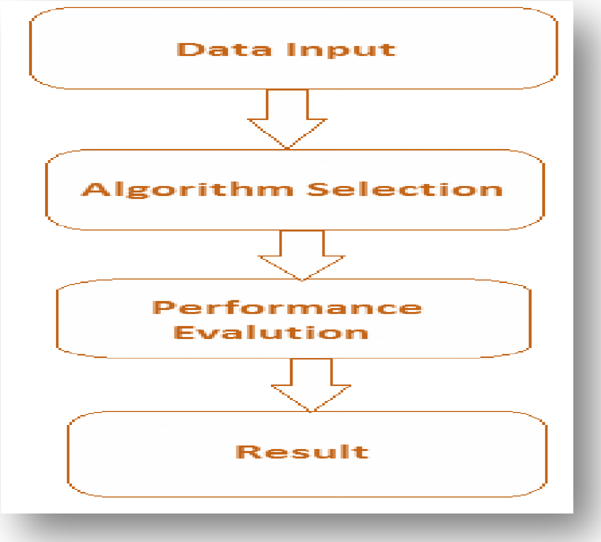

Figure-2
System Architecture

Data Input: Whenever a user hits a page the log data is collected automatically in web servers. It represents the accurate navigational behavior of visitors. It is the primary source of data in web usage mining. Each hit against the server, corresponding to an HTTP request, generates a single entry in the server access logs. There are different forms of log files like Apache, IIS etc.

Each log entry may contain fields such as date time s-ip cs-method cs-uri-stem cs-uri-query s-port cs-username c-ip cs (user-agent) sc-status sc-substatus sc-win32-status sc-bytes cs-bytes. Different servers have different log formats nevertheless the data in this log fragment is pretty typical of the information available. Let's look at one line from the above fragment

```
    ppp931.on.bellglobal.com
      - -
   [26/Apr/2000:00:16:12 -0400]
   "GET /download/windows/asctab31.zipHTTP/1.0"
    200
   1540096
"http://www.htmlgoodies.com/downloads/freeware/webdevelopment/15.html"
      "Mozilla/4.7 [en]C-SYMPA  (Win95; U)"
```

Here
IP address: "ppp931.on.bellglobal.com"
Username etc: "- -"
Username etc. Only relevant when accessing password-protected content.
Timestamp: "[26/Apr/2000:00:16:12 -0400]"
Accessrequest:"GET/download/windows/asctab31.zip HTTP/1.0"
Result status code : "200"
Bytes transferred : "1540096"
User Agent : "Mozilla/4.7 [en]C-SYMPA (Win95; U)"

Algorithm Selection: Previous algorithm CSGM[5] (closed sequential sequential generator mining) is having some aspects that need improvement. In some datasets which has few distinct events/items, the size of the generated frequent sequences sets and closed sequences sets are similar at certain minimum support settings. In this case, the non-redundant rule mining method does not have a large reduction ratio compared to the full sequential rules. Previous algorithm is not suitable for datasets that are too dense or too sparse. Previous algorithm CSGM (closed sequential sequential generator mining) is have some aspects that need improvement. In some datasets which has few distinct events/items, the size of the generated frequent sequences sets and closed sequences sets are similar at certain minimum support settings. In this case, the non-redundant rule mining method does not have a large reduction ratio compared to the full sequential rules. Previous algorithm is not suitable for datasets that are too dense or too sparse. So we proposed a method that recovers CSGM method with high accuracy and less memory and time consumption. In this method we are used decision tree learning algorithm for decision making for capturing knowledge in the system

Definition 1 (Decision tree): Decision tree are commonly used for gaining information for the purpose of decision -making. Decision tree starts with a root node on which it is for users to take actions. From this node, users split each node recursively according to decision tree learning algorithm. The final result is a decision tree in which each branch represents a possible scenario of decision and its outcome.

Definition 2 (Information Gaining): Information gain as the expected reduction of entropy related to specified attribute when splitting a decision tree node.

Algorithm of proposed method

```
1.  Import web log fie
2.  Filter data in row column format
3.  Find user sessions
4.  User sessions defined as a class
5.  Get all unique attribute values
6.  Calculate the threshold according to class values
    using formula
7.  n= no class in dataset.
8.  $threshold = - \frac{class\ (a)}{No.of\ rows}\ \log n\ \frac{class\ (a)}{No.\ of\ rows}$
9.  Calculate info gain for all attributes using formula
10. n= no of attribute in a column.
     $Gain = threshold - \frac{attribute}{total}\ \log n\ \frac{attribute}{total}$

11. Sort all attribute value accordingly to best
    attribute values.
12. Create Sub. Sets of all sorted data set.
13. Repeat till all attribute get a unique value.
```

Performance Evolution: To compare the performance of CSGM method and our proposed method we use N-fold Cross validation technique[6]. In this method the data set is divided into *k* subsets, and the holdout method is repeated *k* times. Each time, one of the *k* subsets is used as the test set and the other *k-1* subsets are put together to form a training set. Then the average error across all *k* trials is computed. The advantage of this method is that it matters less how the data gets divided. Every data point gets to be in a test set exactly once, and gets to be in a training set *k-1* times. The variance of the resulting estimate is reduced as *k* is increased. A variant of this method is to randomly divide the data into a test and training set *d*ifferent times. The advantage of doing this is that you can independently choose how large each test set is and how many trials you average over.

To find an optimal way to classify learning set in this method we used ID3 decision learning tree algorithm. The basic idea of ID3 algorithm is to construct the decision tree by employing a top-down, greedy search through the given sets to test each attribute at every tree node. In order to select the attribute that is most useful for classifying a given sets.

The sample data used by ID3 has certain requirements, which are: i. Attribute-value description - the same attributes must describe each example and have a fixed number of values. ii. Predefined classes - an example's attributes must already be defined, that is, they are not learned by ID3. iii. Discrete classes - classes must be sharply delineated. Continuous classes broken up into vague categories such as a metal being "hard, quite hard, flexible, soft, quite soft" are suspect. iv. Sufficient examples - since inductive.

Generalization is used (i.e. not provable) there must be enough test cases to distinguish valid patterns from chance occurrences. The entropy obtains from ID3 is not able to measure the proper impurities of data set as the accuracy varies along with the data sets. For resolving this problem we are using $\log_n$ to eliminate the impurities along with the variation in data sets to find the accuracy.

Related Work

Sequential pattern mining was first introduced by Agrawal and Srikant[1]. It is the process of extracting certain sequential patterns whose support exceeds a predefined minimal support threshold. Since the number of sequences can be very large, and users have different interests and requirements, to get the most interesting sequential patterns usually a minimum support is predefined by the users. By using the minimum support we can prune out those sequential patterns of no interest, consequently making the mining process more efficient. Later, a series of data projection based algorithms were proposed, which included FreeSpan[7] and Prefix span[3] several closed sequence mining algorithm were also introduced, such as CloSpan[2], TSP[8], CSGM[5].

Closed Sequential pattern mining: A closed sequential pattern is a sequential pattern included in no other sequential pattern having exactly the same support. The first algorithm designed to extract closed sequential patterns is CloSpan[2] with a detection of non-closed sequential patterns avoiding a large number of recursive calls.

CloSpan is based on the detection of frequent sequences of length 2 such that “*A* always occurs before/after *B*”. First, it adopts a novel sequence extension, called BI-Directional Extension, which is used both to grow the prefix pattern and to check the closure property. Second, in order to prune the search space more deeply than previous approaches, it proposes a Back Scan pruning method. The main idea of this method is to avoid extending a sequence by detecting in advance that the extension is already included in a sequence. One major difference between CloSpan and Prefix span is that CloSpan implements an early termination mechanism to avoiding unnecessary traversing of search space by using both backward sub-pattern and backward super-pattern methods, some patterns will be absorbed or merged, and the search space growth can be reduced.

Sequential Generator Mining: In a sequential database, the sequential generator[9] refers to patterns without any subsequence with the same support. Sequential generators used together with closed sequential patterns can bring additional information, which closed sequential patterns alone are not able to provide. According to the Minimum Description Length[10] generator are the minimal member and preferable over closed patterns in terms of association rule induction and classification. The gen miner method[10] is the first sequential generator mining algorithm; it fills the research gap in sequential generator mining. The performance study shows that Gen Miner can run a lot faster than a full frequent sequential pattern mining method, such as Prefix Span, and its speed can be on par with or at times faster than that of closed sequential pattern mining algorithms, such as CloSpan.

CSGM method: CSGM algorithm[5] uses a similar prefix-search-lattice data structure and the projected database concept as for CloSpan. The detailed procedure of this algorithm consists of two major steps: scanning database and mining the entire candidate set of closed sequential patterns and their corresponding generators; then eliminating all non-closed sequential patterns.

The CSGM algorithm[5] first scans the sequential database once, and finds all frequent length-1 sequences. These length-1 sequences are those patterns containing only one item. Since the generators of length-1 sequences are themselves, we put these sequences and a set of their corresponding generators together as sequence-generator pairs, and we also find the corresponding project databases for these sequences

CSGM methods have some aspects that need improvement. In some datasets, e.g. the MSNBC dataset, which has few distinct events/items, the size of the generated frequent sequences sets and closed sequences sets are similar at certain minimum support settings. In this case, the non-redundant rule mining method does not have a large reduction ratio compared to the full sequential rules CSGM algorithm is not suitable for datasets that are too dense or too sparse (i.e. those datasets either with only a few events or that have too many distinct events). CSGM has similar complexity to closed sequence mining and sequential generator mining like require high memory space, less accuracy, but it is considerably faster than conducting sequential generator mining and closed sequential pattern mining separately.

Conclusion

We have described N fold crossvalidation techinque for perfomance evaluation of our proposed algorithm and give brief description of ID3 learning algorithm which we are used for classification. So we classify the data set in our web log file. In this paper, we highlight the problem of CSGM sequential pattern mining method and give the solution to overcome the problem of CSGM method and generate non redundant association rule with less memory and time consumption. We plan to improve the efficiency of searching through classifying documents and suggest similar documents by ID3 learning decision tree algorithm.

Researcher, who is wanted to work in sequential pattern mining this paper, is very useful. In this paper we give only overview of our proposed work.hum next paper me iska implementation denge ye future work hai.

References

1. Agrawal R. and Srikant R., Mining sequentialpatterns, Proceedings of the Eleventh International Conference on Data Engineering **(1995)**

2. Yan X., Han J. and Afshar R., CloSpan: Mining Closed Sequential Patterns in Large Datasets **(2003)**

3. Pei J., Han J., Mortazavi-Asl B., Pinto H., Chen Q., Dayal U. et al., PrefixSpan: Mining Sequential Patterns by Prefix-Projected Growth **(2001)**

4. Han J., Pei J. and Yin Y., Mining frequent patterns without candidate **(2000)**

5. Hao Zang and Yue Xu Yuefeng Li, Non-Redundant Sequential Association Rule Mining and Application in Recommender Systems Proceedings of IEEE International Conference on Data Mining **(2010)**

6. Jon Starkweather may, Cross validation technique **(2010)**

7. Han J., Pei J., Mortazavi-Asl B., Chen Q., Dayal U. and Hsu M.C., Free Span: frequent pattern-projected sequential pattern mining **(2000)**

8. Tzvetkov P., Yan X. and Han J., TSP: Mining top-k closed sequential patterns, Knowledge and Information Systems, **7(4),** 438-457 **(2005)**

9. Xu Y. and Li Y., Generating concise association rules **(2007)**

10. Li J., Li H., Wong L., Pei J. and Dong G., Minimum description length principle: generators are preferable to closed patterns **(2006)**

11. Gaul W. and Schmidt-Thieme L., Mining Generalized Association Rules for Sequential and Path Data. Proceedings of the 2001 IEEE International Conference on Data Mining **(2001)**

12. Xu Y. and Li Y., Concise representations for approximate association rules **(2008)**