Performance Analysis of Sequential Pattern Mining Algorithms on Large Dense Datasets

Dr. Sunita Mahajan 1, Prajakta Pawar 2 and Alpa Reshamwala 3

1 Principal, 2 M.Tech Student, 3 Assistant Professor,
1 Institute of Computer Science, 2, 3 Computer Engineering Department.
1 MET, Mumbai University, Bandra, 2, 3 MPSTME, SVKM’s NMIMS University Mumbai, India

ABSTRACT
Sequential Pattern Mining involves applying data mining methods to large web data repositories to extract usage patterns. With the proliferation of Internet, discovery and analysis of useful information from the World Wide Web becomes a practical necessity. It has become much more difficult to access relevant information from the Web with the explosive growth of information available on the World Wide Web. Web usage mining has become a fertile field of research for improving designs of web sites, analyzing system performance as well as network communications, understanding user reaction, motivation and building adaptive Web sites. Owing to important applications such as mining web page traversal sequences, many algorithms have been introduced in the area of sequential pattern mining over the last decade. Sequential pattern mining algorithms can be classified into (a) Apriori-based such as GSP (Generalised Sequential Pattern) , SPADE (Sequential PAttern Discovery using Equivalent Class) and SPAM (Sequential Pattern mining), (b) Pattern-growth such as Prefixspan and (c) Early-pruning methods such as AprioriAll_Set algorithm. This paper implements all the mentioned sequential pattern mining algorithms on the Web sequential Datasets KDD CUP 2000 and MSNBC. As shown by the simulation results, Prefixspan gives best performance for dense dataset such as KDD cup 2000 and AprioriAll_Set an early pruning algorithm gives best performance for heavily dense dataset. Among the apriori based and pattern growth based algorithms SPADE performs best for heavily dense dataset but requires the maximum memory. Hence for heavily dense dataset, among the apriori based and pattern growth based algorithms Prefixspan gives best performance both in space and time.

Keywords: Generalized Sequential Pattern Mining, Web Usage Mining, Sequential Pattern Mining, AprioriAll, Set Theory, Sequential PAttern Mining, Sequential PAttern Discovery using Equivalence classes

1. INTRODUCTION
Data Mining is the process of extracting useful information which is hidden in large databases. The knowledge or pattern mined could be used to make decisions. Sequential pattern mining is one of the major areas of research in the field of data mining as patterns from a sequence database. With massive amounts of data continuously being collected and stored, many industries are becoming interested in mining sequential patterns from their database. Sequential pattern mining is one of the most well-known methods and has broad applications including web-log analysis, customer purchase behavior analysis and medical record analysis. In the retailing business, sequential patterns can be mined from the transaction records of customers. For example, having bought a notebook, a customer comes back to buy a PDA and a WLAN card next time. The retailer can use such information for analyzing the behavior of the customers, to understand their interests, to satisfy their demands, and above all, to predict their needs. In the medical field, sequential patterns of symptoms and diseases exhibited by patients identify strong symptom/disease correlations that can be a valuable source of information for medical diagnosis and preventive medicine. In web log analysis, the exploring behavior of a user can be extracted from member records or log files. For example, having viewed a web page on “Data Mining”, user will return to explore “Business Intelligence” for new information next time. These sequential patterns yield huge benefits, when acted upon, increases customer royalty.

For example, retail stores often collect customer purchase records in sequence databases in which a sequential pattern would indicate a customer’s buying habit. In such a database, each purchase would be represented as a set of items purchased and a customer sequence would be a sequence of such itemsets. More formally, given a sequence database and a user-specified minimum support threshold, sequential pattern mining is defined as finding all frequent subsequences that meet the given minimum support threshold [2]. GSP [17], PrefixSpan [13], SPADE [24], and SPAM [1] are some well known algorithms to locate such patterns efficiently.

Knowledge extraction from the World Wide Web has become an important and challenging task as enormous amount of data in form of semi-structured nature is available. Web mining is the application of data mining techniques to discover patterns from the Web [1]. In Web Mining, data can be collected at the server-side, client-side, proxy servers, or obtained from an organization’s database, which contains business data or consolidated Web data. The information gathered through Web mining is evaluated by using traditional data mining parameters such as clustering and classification,
association, and examination of sequential patterns [2]. According to analysis targets, web mining can be divided into three different types, which are Web usage mining, Web content mining and Web structure mining.

Web Usage mining has a lot of application in real life such as Improving designs of web sites, analyzing system performance as well as network Communications, understanding user reaction, motivation and Building adaptive Web sites; it is now a very important and useful subject. Web usage mining is concerned with finding user navigational patterns on the Wide Web by extracting knowledge from web logs, where ordered sequences of events in the sequence database are composed of single items and not sets of items, with the assumption that a web user can physically access only one web page at any given point in time. The pattern mining and researches in data mining, machine learning as well as statistics are mainly focused on analysis of the web pattern discovery. As for pattern mining, it could be: - Statistical analysis, used to obtain useful statistical information such as the most frequently accessed pages; Association rule mining [2], used to find references to a set of pages that are accessed together with a support value exceeding some specified threshold; Sequential pattern mining [3], used to discover frequent sequential patterns which are lists of Web pages ordered by viewing time for predicting visit patterns; Clustering, used to group together users with similar characteristics; Classification, used to group together users into predefined classes based on their characteristics.

Currently, most web usage-mining solutions consider web access by a user as one page at a time, giving rise to special sequence database with only one item in each sequence’s ordered event list. Thus, given a set of events $E = \{a, b, c, d, e, f\}$, which may represent product web pages accessed by users in an e-commerce application, a web access sequence database for four users may have four records: [T1, $<abdac>\}$; [T2, $<eaebac>\}$; [T3, $<babfaec>\}$; [T4, $<abfca>\}$. A web log pattern mining on this web sequence database can find a frequent sequence, $abac$, indicating that over 90% of users who visit product $a$’s web page also immediately visit product $b$’s web page and then revisit product $a$’s page, before visiting product $c$’s page. Store managers may then place promotional prices on product $a$’s web page, which is visited a number of times in sequence, to increase the sale of other products. The web log could be on the server-side, client-side, or on a proxy server, each with its own benefits and drawbacks in finding the users’ relevant patterns and navigational sessions.

### 2. RELATED WORK

Mining frequent web access patterns from very large databases (e.g. using click-stream analysis) has been studied intensively and there are a variety of approaches. Most of the previous studies have adopted a sequential patterns mining technique – which aims to find sub-sequences that appear frequently in a sequence database – on a web log access sequence. In web server logs, a visit by a client is recorded over a period of time and the discovery of sequential patterns allows web-based organizations to predict user visit patterns, which helps in targeting advertising aimed at groups of users based on these patterns.

Sequential pattern mining was proposed in [3], using the main idea of association rule mining presented in Apriori algorithm of [2]. Later, three algorithms (Apriori, AprioriAll, and AprioriSome) to handle sequential mining problem were proposed in [3]. Following this, the GSP (Generalized Sequential Patterns) [4] algorithm, which is 20 times faster than the Apriori algorithm in [3] was proposed. The PSP (Prefix Tree for Sequential Patterns) [5] approach is much similar to the GSP algorithm [4]. The main idea of Graph Traversal mining which is proposed by [6][7], is using a simple unweighted graph to reflect the relationship between the pages of Web sites. The Web Utilization Miner (WUM) [8] tool aims to discover sequential patterns that are considered as interesting from a statistical point of view. The WAP-mine, described in [1], is a method that allows the extraction of frequent patterns from the user sessions. The authors of [9] were interested in discovering contiguous sequence patterns in a Web log file; The FS-Miner algorithm [10] is based on the FS-Tree that is a compressed tree used to represent sequences. The ApproxMAP [11] combines clustering and sequential patterns for extraction of multiple alignment sequential pattern mining. Pre-Order Linked WAP-Tree Mining (PLWAP) algorithm has been presented by [12] for efficiently mining of sequential patterns from the Web log. Automatic Log mining via Genetic algorithm to mine sequential accesses from Web log files has been proposed by [13]. An intelligent recommender system known as SWARS (Sequential Web Access based Recommender System) that uses sequential access pattern mining is proposed in [14].

Traditional sequential patterns mining approaches such as Apriori-based algorithms [3, 4] encounter the problem that multiple scans of the database are required in order to determine which candidates are actually frequent. Most of the solutions provided so far for reducing the computational cost resulting from the apriori property use a bitmap vertical representation of the access sequence database [15][16][17][18] and employ bitwise operations to calculate support at each iteration. The transformed vertical databases, in their turn, introduce overheads that lower the performance of the proposed algorithm, but not necessarily worse than that of pattern-growth algorithms. Chiu et al. [19] propose the DISCall algorithm along with the Direct Sequence Comparison DISC technique, to avoid support for counting by pruning nonfrequent sequences according to other sequences of the same length. There is still no variation of the DISC-all for web log mining. Breath-first search, generate-and-test, and multiple scans of the database, which are discussed below, are all key features of apriori-based methods that pose challenging problems, hinder the performance of the algorithms. Pei et al. introduced a compressed data structure called Web Access Pattern tree (or WAP-tree), which facilitates the development
of algorithms for mining access patterns from pieces of web logs [1]. Since then, many modifications were proposed in order to further improve efficiency, by eliminating the need to perform any re-construction of intermediate WAP-trees during mining; for example the Position Coded Pre-order Linked Web Access Pattern mining algorithm [20][21], Conditional Sequence mining algorithm [22] and the modified Web Access Pattern (mWAP) algorithm [23].

Sequential pattern mining algorithms can be classified into apriori-based, pattern-growth, early-pruning, and hybrids of these three techniques. Breadth-first search, generate-and-test, and multiple scans of the database, are all key features of apriori-based methods that pose challenging problems and hinder the performance of the algorithms. The apriori-based algorithms are found to be too slow and have a large search space, while pattern-growth algorithms have been tested extensively on mining the web log and found to be fast, early-pruning algorithms have had success with protein sequences stored in dense databases. Shang Gao et al. approach in [24], relaxes the constraint described in AprioriAll/Some and improves the performance by user oriented and self adaptive approach than the probabilistic knowledge representation. The key features of pattern growth based sequential pattern mining algorithms are: Sampling and/ compression, Candidate Sequence pruning, search space partitioning, tree projections, Depth first traversal, suffix/prefix growth and memory only. For example the pattern growth based PrefixSpan [27] algorithm is implemented in this paper. In this paper, Traditional set based apriori-based algorithm proposed by A. Reshamwala and S. Mahajan in [25], is implemented as the algorithm has acceptable performance measures such as low CPU execution time and low memory utilization when mined with low minimum support values and maximum confidence among the patterns generated. The algorithm performs a Breadth-first search, with the implementation of Hash Map data structure in Java, the support counting is avoided leading to few scans of database and helps in projecting the database in vertical layout as well as positions of the itemsets are coded. The algorithm by applying the Set operations results in database Shrinking. Thus the algorithm handles Candidate sequence pruning by applying the intersection operation that allows them to prune candidate sequences early in the mining process. With the database shrinking the corelations among the patterns generated is highly increased.

3. EXPERIMENTAL RESULTS

A simulation study is done to compare the performances of the Apriori based algorithms: GSP [4], SPADE [29] and SPAM [28], Pattern-growth approach based: PrefixSpan [27] and Early pruning algorithm: AprioriAll_Set[25] to discover sequential patterns from large sequences. 

These algorithms are executed on sequential Datasets KDD CUP 2000, Kosarak10k, LEVIATHAN and MSNBC. These dataset are downloaded from SPMF (Sequential Pattern Mining Framework) which is implemented by Phillipe Fournier-Viguera [26] and available from http://www.philippe-fournier-viger.com/spmf/. SPMF tool is used to analyze and compare dataset statistical parameters.

Dense Data Sets are characterized by having a small number of unique items and a large number of sequences. Probability of frequent sequences is HIGH in dense dataset. Sparse data sets are characterized by having a large number of unique items and a small number of sequences. Probability of frequent sequences is LOW in sparse dataset. Table 1 shows the comparison of the different web datasets. KDD CUP 2000 dataset contains 59,601 sequences of click stream data from an e-commerce. It contains 497 distinct items. The average length of sequences is 2.42 items with a standard deviation of 3.22. In this dataset, there are some long sequences [30]. For example, 318 sequences contain more than 20 items. MSNBC is a dataset of click-stream data. The original dataset contains 989,818 sequences obtained from the UCI repository. Here the shortest sequences have been removed to keep only 31,790 sequences. The number of distinct items in this dataset is 17 (an item is a webpage category). The average number of item sets per sequence is 13.33. The average number of distinct items per sequence is 5.33. Both the datasets are dense dataset, that is, usually there are very less unique items with many repetitions in the sequences of one user.

<table>
<thead>
<tr>
<th>Sr. No.</th>
<th>Statistical Parameters</th>
<th>KDDCUP 2000</th>
<th>MSNBC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Less Dense</td>
<td>More Dense</td>
</tr>
<tr>
<td>1</td>
<td>Number of sequences</td>
<td>59601</td>
<td>31790</td>
</tr>
<tr>
<td>2</td>
<td>Number of distinct items</td>
<td>497</td>
<td>17</td>
</tr>
<tr>
<td>3</td>
<td>Average number of item sets per sequence</td>
<td>2.51</td>
<td>13.33</td>
</tr>
<tr>
<td>4</td>
<td>Average number of distinct items per sequence</td>
<td>2.51</td>
<td>5.33</td>
</tr>
</tbody>
</table>

As shown in the Table 1, the KDD CUP 2000 dataset is a moderately dense whereas the MSNBC is a highly dense dataset.
The experiments were performed on a system having Java SE 1.6.0_26 with NetBeans 7.0 on Windows 7 Professional, Intel Core i5-2400 processor 3.10 GHz with 4 GB RAM. Performance of KDD CUP 2000 dataset can be seen in Figure 2, 3 and 4, where the minimum support ranges from 1 % to 7 %. The Figure 1 shows the number of patterns generated using KDD CUP 2000 dataset.

Figure 1: No. of patterns of KDD cup 2000 Dataset

Figure 2: Performance of KDD cup 2000 Dataset

Figure 2 shows performance of KDD CUP 2000 dataset. The minimum support varies from 3% to 6 % in the figure 2, 3, 4, 5, 6 and 7. The SPADE algorithm gives the best performance followed by SPAM and Prefixspan. GSP give the worst performance and AprioriAll_Set gives approximately uniform performance for varying support values. Memory utilization of KDD CUP 2000 dataset is shown in Figure 3. The SPADE algorithm performs best but requires the maximum memory. Prefixspan requires less memory as compared to SPAM and AprioriAll_Set as it uses the pseudo projection for projected database. GSP requires least memory with the minimum support of 3% and for the rest requires approximately same memory as Prefixspan algorithm.

The Figure 4 shows the number of patterns generated using MSNBC dataset.

Figure 3: Memory Utilization of KDD Cup 2000 Dataset

Figure 4: Performance of MSNBC Dataset

Figure 5: Performance of MSNBC Dataset
Figure 5 shows performance of MSNBC dataset. The GSP algorithm gives the worst performance for heavily dense dataset such as MSNBC. Figure 6, shows the performance of all the algorithms except GSP. AprioriAll_Set outperforms all the algorithms followed by SPADE, SPAM and Prefixspan algorithm. Memory utilization if MSNBC dataset is shown in Figure 7. AprioriAll_Set not only outperforms but utilizes the least memory as compared to all the algorithms. AprioriAll_Set approximately uses uniform memory for varying support values. It is followed by SPAM. Figure 8 below shows the number of discovered patterns by both the datatset with minimum support of 1%. MSNBC a heavily dense dataset discovers 45622 sequential patterns where as KDD Cup 2000 a less dense dataset discovers 77 sequential patterns. Among the apriori based and pattern growth based algorithms SPADE performs best for heavily dense dataset. There is a negligible difference in the performance of Prefixspan and SPAM as shown in Figure 9.

4. CONCLUSION AND FUTURE WORK

Web Usage Mining is the application of pattern mining techniques to usage logs of large Web data repositories in order to produce results that can be used in the design tasks. An important input to these design tasks is the analysis of how a Web site is being used. Usage analysis includes straightforward statistics, such as page access frequency, as well as more sophisticated forms of analysis, such as finding the common traversal paths through a Website. The performances of the Sequential pattern mining algorithms on Heavily dense and dense dataset such as MSNBC and KDD Cup 2000 shows that for heavily dense dataset AprioriAll_Set not only outperforms but utilizes the least memory as compared to all the algorithms. AprioriAll_Set approximately uses uniform memory for varying support values. This is because of its database shrinking property. It is followed by SPAM in memory utilization and SPADE in performance. For dense dataset such as KDD Cup 2000; the SPADE algorithm gives the best performance followed by SPAM and Prefixspan. SPADE gives best performance due to its vertical bitmap layout which helps in faster support counting. GSP gives the worst performance due to its huge candidate generation property and AprioriAll_Set gives approximately uniform performance for varying support values. The SPADE algorithm outperforms other algorithms but requires the maximum memory. Prefixspan requires less memory as compared to SPAM and AprioriAll_Set as it uses the pseudo projection for projected database. GSP requires approximately same memory as Prefixspan algorithm. As shown by the simulation results, Prefixspan gives best performance for dense dataset such as KDD cup 2000 and AprioriAll_Set an early pruning algorithm gives best performance for heavily dense dataset. Among the apriori based and...
pattern growth based algorithms SPADE performs best for heavily dense dataset but requires the maximum memory. There is a negligible difference in the performance of Prefixspan and SPAM. Hence for heavily dense dataset, among the apriori based and pattern growth based algorithms Prefixspan gives best performance.

REFERENCES


[26] SPMF: —Sequential Pattern Mining Framework.


AUTHOR

Ms. Alpa Reshamwala is currently working as an Assistant Professor in the Department of Computer Engineering at MPSTME, NMIMS University. She received her B.E degree in Computer Engineering from Fr. CRCE, Bandra, Mumbai University in 2000 and M.E degree in Computer Engineering from TSEC, Mumbai University in 2008. Her area of Interest includes Artificial Intelligence, Data Mining, Soft Computing – Fuzzy Logic, Neural Network and Genetic Algorithm. She has more than 25 papers in National/International Conferences/Journal to her credit.

Dr Sunita M. Mahajan is currently working as the Principal, Mumbai Educational Trust’s Institute of Computer Science. She has done her Doctorate from S.N.D.T. Women’s University in 1997. She has worked as senior scientist at Bhabha Atomic Research Centre for 31 years and entered educational field after her retirement. She has done extensive work in parallel processing. She has more than 45 papers in National and International conferences and journals to her credit. She has guided many PhD students in distributed computing, data mining, natural language processing etc. Her current field of interest is parallel processing, distributed computing, cloud computing, data mining. She has also written a text book on “Distributed Computing”(New Delhi, Oxford University Press, 2010)

Prajakta Pawar is currently pursuing M.Tech in the Department of Computer Engineering at MPSTME, NMIMS University. She has received her B.E degree from SSJCOE, Dombivli, Mumbai university in 2011. Her area of Interest includes Artificial Intelligence, Data Mining. She has published 2 papers. She has attended one International Conference and received award for the “Excellent Paper”.

AUTHOR

Ms. Alpa Reshamwala is currently working as an Assistant Professor in the Department of Computer Engineering at MPSTME, NMIMS University. She received her B.E degree in Computer Engineering from Fr. CRCE, Bandra, Mumbai University in 2000 and M.E degree in Computer Engineering from TSEC, Mumbai University in 2008. Her area of Interest includes Artificial Intelligence, Data Mining, Soft Computing – Fuzzy Logic, Neural Network and Genetic Algorithm. She has more than 25 papers in National/International Conferences/Journal to her credit.

Dr Sunita M. Mahajan is currently working as the Principal, Mumbai Educational Trust’s Institute of Computer Science. She has done her Doctorate from S.N.D.T. Women’s University in 1997. She has worked as senior scientist at Bhabha Atomic Research Centre for 31 years and entered educational field after her retirement. She has done extensive work in parallel processing. She has more than 45 papers in National and International conferences and journals to her credit. She has guided many PhD students in distributed computing, data mining, natural language processing etc. Her current field of interest is parallel processing, distributed computing, cloud computing, data mining. She has also written a text book on “Distributed Computing”(New Delhi, Oxford University Press, 2010)

Prajakta Pawar is currently pursuing M.Tech in the Department of Computer Engineering at MPSTME, NMIMS University. She has received her B.E degree from SSJCOE, Dombivli, Mumbai university in 2011. Her area of Interest includes Artificial Intelligence, Data Mining. She has published 2 papers. She has attended one International Conference and received award for the “Excellent Paper”.