High Utility Web Access Patterns Mining from Distributed Databases

Md.Azam Hosssain¹, Md.Mamunur Rashid¹, Byeong-Soo Jeong¹, Ho-Jin Choi²

¹Database Lab, Department of Computer Engineering, Kyung Hee University
1 Seochun-dong, Kihung-gu, Youngin-si, Kyunggi-do, 446-701, Republic of Korea

²Computer Science Dept., Korea Advanced Institute of Science and Technology
335 Guseong-dong, Yuseong-gu, Daejeon 305-701, Republic of Korea

{azam, mamun, jeong}@khu.ac.kr, hojinc@kaist.ac.kr

Abstract. Utility based web access pattern mining and knowledge discovery from database has become an interesting research domain in recent time. Traditional pattern mining algorithms deal with only binary occurrence of a web page and do not consider the weight or profit of the web page. Hence, utility-based web path traversal pattern mining technique has evolved and got much interest in recent time. However, the methods of mining high utility web access pattern from distributed databases have not explored yet. All of the existing utility pattern mining algorithms are based on centralized database. In this paper, we have been proposed a parallel and distributed method for mining high utility web access patterns. We also proved that our algorithm maintains downward closure property. Extensive experimental analyses reflect that our proposed technique is highly efficient on large databases in distributed environment

Keywords: Web Access, High Utility Pattern, Distributed Database.

1 Introduction

The World Wide Web offers huge, widely distributed global information system. It contains a rich and dynamic collection of web link, web page access and usage information [1]. Web access pattern mining algorithm discovers the usage patterns of user’s frequent web page access. Therefore, mining frequent web access can be used to improve the webpage design, advertisement placement in most popular page and so on.
Moreover, today’s internet era databases are inherently distributed. Most of the organizations operate business in global markets require to perform data mining on distributed data sources to turn them into realistic and meaningful knowledge for their future use and the volume of data available for usage is very high. This inherent distribution source of data and the voluminous in size emerges to develop scalable parallel and distributed algorithm for data mining.

In recent time, the concept of utility-based web path traversal mining has been developed in [7]. The concept of utility for web path traversal first time was introduces in [5]. They adopted the browsing time of a user at particular web site as a utility of that web site. They used the same definitions of utility from the high utility pattern mining algorithm [2-6]. Farhan et al. [8] proposed the EUWPTM algorithm for utility based web path traversal pattern mining that exploit the pattern growth sequential pattern mining approach.

However, the technique of mining high utility web access pattern from distributed databases have not explored yet. All of the existing utility-based web path traversal pattern mining algorithms are considered the centralized database. In real scenario, web access data mining techniques need to handle the large database. Therefore, it is necessity to focus on large-scale parallel and distributed high utility-based web site traversal patterns mining.

In summary, our main contributions in this work are

- We define the problem for high utility-based web access patterns mining in the parallel and distributed environment
- We have been proposed a distributed method for mining the high utility web path traversal pattern from large databases.
- We also proved that our proposed technique maintain the downward closure property.

The rest of the paper is organized as follows. Section 2 presents the review of some recent related research works. Section 3 describes problem definition and Section 4 presents the proposed method in details. Section 5 elaborates the experimental results and finally Section 6 concludes the paper.

2 Related Work

WEBMINER [1] was proposed to discover the information from the data of World Wide Web transactions. They focused to analyze the interesting and meaningful relationships from a large pole of web database. A web utilization miner (WUM) [9] finds the interesting web navigation pattern. They primarily focused on the MINT query language and its execution mechanism.

The theoretical model and definitions of high utility pattern mining were given in [2]. This method, called MEU (mining with expected utility), cannot maintain the downward closure property. In this approach, they used heuristics to determine whether an itemset should be considered as a candidate itemset. This is impractical whenever the number of distinct items is large and the utility threshold is low. Later, the same authors proposed two new algorithms, UMining and UMining_H, to calculate high utility patterns [3]. In UMining, a pruning strategy based on the utility
upper bound property is used. Furthermore, these methods do not satisfy the downward closure property and may overestimate too many patterns. Based on the definitions of [2], the Two-Phase [4, 5] algorithm was developed to find high utility itemsets. The authors have defined the transaction weighted utilization (twu) [6] and using it they proved that it is possible to maintain the downward closure property.

In recent time, the concept of utility-based web path traversal mining has been developed in [8]. The concept of utility for web path traversal first time was introduces in [7]. They adopted the browsing time of a user at particular web site as a utility of that web site. They used the same definitions of utility from the high utility pattern mining algorithm [2-6]. Farhan et al. [8] proposed the EUWPTM algorithm for utility based web path traversal pattern mining that exploit the pattern growth sequential pattern mining approach. By using the pattern growth mining approach, their technique prunes a huge number of candidate items at first database scan.

3 Problem Definition

Let \( S = \{s_1, s_2, s_3, \ldots, s_j\} \) be a \( j \)-sequence of traversal path and \( D = \{T_1, T_2, T_3, \ldots, T_m\} \) be web-log, where \( T_i \) is traversal path. A traversal path database is presented in the table 1. Transaction \( T_1 \) contains the user browsing sequence B, C, and D websites. The quantity inside the parenthesis is the internal utility that indicates the time spent by the user at that website. Table 2 shows the external importance of a website. As for example, internal utility for website B is 3 where external utility is 10.

<table>
<thead>
<tr>
<th>Table 1 A web traversal Database</th>
<th>Table 2 External Utility</th>
</tr>
</thead>
<tbody>
<tr>
<td>TID</td>
<td>Traversed Path</td>
</tr>
<tr>
<td>-----</td>
<td>----------------</td>
</tr>
<tr>
<td>( T_1 )</td>
<td>B(3), C(2), D(3)</td>
</tr>
<tr>
<td>( T_2 )</td>
<td>A(3), D(2), E(2)</td>
</tr>
<tr>
<td>( T_3 )</td>
<td>B(3), E(2)</td>
</tr>
<tr>
<td>( T_4 )</td>
<td>A(1), B(1), C(1)</td>
</tr>
<tr>
<td>( T_5 )</td>
<td>A(2), B(3), D(5)</td>
</tr>
<tr>
<td>( T_6 )</td>
<td>A(3), B(4)</td>
</tr>
<tr>
<td>( T_7 )</td>
<td>E(1)</td>
</tr>
<tr>
<td>( T_8 )</td>
<td>B(2), D(2)</td>
</tr>
</tbody>
</table>

The formal definition for high utility mining model for web path traversal pattern is as follows:

Definition 1: Utility of a website \( ij \) in transaction \( T_k \) is the quantitative measure of the website which is multiplication of internal and external utility of a web page.

\[
u(i_j, T_k) = I(i_j, T_k) \times e(i_j)
\]

For example, in table 1 \( u(D, T_1) = 3 \times 7 = 21 \).
Definition 2: The utility of a website sequence $S$ in transaction $T_k$ is the sum of utility of the entire website that belongs to that sequence and it is defined by,

$$u(S, T_k) = \sum_{i \in S} u(i, T_k)$$

Definition 3: Transaction Utility ($tu$) of a transaction is the total utility of that transaction and it is defined by,

$$tu(T_k) = \sum_{i \in T_k} u(i, T_k)$$

For example, in Table 1, $tu(T_4) = u(A, T_4) + u(B, T_4) + u(C, T_4) = 6 + 10 + 4 = 20$

Definition 4: Local transaction utility utilization of an itemset $S$, denoted by $ltwu(S)$, is the sum of the transaction utilities of all transactions containing $S$ of a particular site $P_i$ and defined by,

$$ltwu(S) = \sum_{T_k \subseteq D} tu(T_k)$$

For example, $ltwu(BD) = tu(T_1) = 59$ is since BD appears only in $T_1$ in the local site $P_0$.

Definition 5: Global transaction utility utilization of an itemset $S$, denoted by $gtwu(S)$, is the sum of the transaction utilities of all transactions of all the sites that contain $S$ and defined by,

$$gtwu(S) = \sum_{i=1}^{p} \sum_{T_k \subseteq D} tu(T_k)$$

As for example $gtwu(BD) = tu(T_1) + tu(T_5) = 59 + 77 = 136$.

Theorem 1: If $S = \{S_1 \cup S_2 \cup S_3 \cup \ldots \cup S_q\}$ is a pattern where $q$ is a positive integer and $S_j$ is a sub-pattern of $S$ and $j \in [1, q]$, then global transaction utility utilization $gtwu(S) \leq \sum_{i=1}^{p} \min \{ltwu(S_i), ltwu(S_2), ltwu(S_3), \ldots, ltwu(S_q)\}$.

4 Proposed Method

In this section, we present the proposed method for parallel and distributed high utility pattern mining. Our proposed algorithm works in the following steps. 

**Master Node:**
1. Broadcast the information of the master site to all the Slave sites.
2. Wait for local transaction weighted utilization ($ltwu$) from the Slave node.
3. Compute global transaction utility utilization ($gtwu$).
4. Broadcast the $gtwu$.
5. Receive the global potential high utility pattern.
6. Calculate the high utility pattern from potential high utility pattern.

**Slave Node:**
7. Scan the local database.
8. Calculate the local transaction utility utilization (ltwu) of each website.
9. Send (ltwu) to the Master node.
10. Wait for global transaction weighted utilization (gtwu).
11. Build the PHUP tree.
12. Find the potential global high utility patterns and send to Master Node.

In the first database scan each node calculates the local transaction-weighted-utilization (ltwu) as the definition 4 of each website. Global transaction weighted utilization is calculated by using the definition 5 and is broadcasted by Master node. Then each local site builds their global transaction utility table using global transaction-utility-utilization and pruned the items that do not satisfied the given threshold 

\[ \text{min}_\text{util}(\bar{\delta}) \].

As for example, total transaction utility value is 335. If \(\bar{\delta}\) is 25% than the minimum utility value will be 

\[ \text{min}_\text{util} = 0.25 \times 335 = 83.75. \]

\(\text{gtwu}(C) < \text{min}_\text{util}(\bar{\delta})\), so “C” is pruned.

After that local site builds PHUP tree by scanning the local database and find the potential global high utility patterns that satisfied the given threshold 

\[ \text{min}_\text{util}(\bar{\delta}) \].

Each local site calculates the actual utility \(u(S)\) from potential high utility patterns by scanning the database and sends to Master node. Finally, global high utility pattern are accumulated by the Master node.

5 Experimental Results

We experiment our algorithm in the system where all sites are identical. Each site consists of a 2.66 GHz CPU with 2 GB memory and running on Linux. Communications among site are established using Message Passing Interface and programs are implemented in C++ language. Experiments have been carried out on real BMS-Web View-1 datasets which contains several months world of click stream of an e-commerce web site. It has 59602 transactions and 497 distinct items.

It is observed from the figure 1 that as the number of process increases, the execution time decreases. We also see that if the minimum threshold is higher than number of candidates is fewer. Therefore, our approach is highly scalable and efficient.
Figure 1 (b) demonstrates that our proposed algorithm generates less number of candidates than existing algorithm “Utility” [7] since our method maintains the downward closure property.

6 Conclusions

The main contribution of this paper is to provide first time a distributed method for utility-based web sequence access mining. Our proposed maintain the downward closure property that helps to prune the website with low utility at the beginning. So, the proposed method is effectively suited for large databases which provide the high scalability and performance gain and require minimum communication among the process.

Acknowledgement

This work was supported by a grant from the NIPA (National IT Industry Promotion Agency, Korea) in 2012 (Global IT Talents Program) and partly supported by the National Research Foundation (NRF) grant (No. 2011-0018264) of Ministry of Education, Science and Technology (MEST) of Korea.

References