A Graph-Based Optimization Algorithm for Website Topology Using Interesting Association Rules

Edmond H. Wu and Michael K. Ng

Department of Mathematics, The University of Hong Kong
Pokfulam Road, Hong Kong
hcwu@hkusua.hku.hk, mng@maths.hku.hk

Abstract. The Web serves as a global information service center that contains vast amounts of data. The Website structure should be designed effectively so that users can efficiently find their information. The main contribution of this paper is to propose a graph-based optimization algorithm to modify Website topology using interesting association rules. The interestingness of an association rule $A \Rightarrow B$ is defined based on the probability measure between two sets of Web pages $A$ and $B$ in the Website. If the probability measure between $A$ and $B$ is low (high), then the association rule $A \Rightarrow B$ has high (low) interest. The hyperlinks in the Website can be modified to adapt user access patterns according to association rules with high interest. We present experimental results and demonstrate that our method is effective.

1 Introduction

Web usage mining refers to mine Web logs to discover user access patterns of Web pages [7]. The effective management of a Website is dependent on awakening the needs of potential and profitable customers, analyzing their behaviors and then improving Web service. Therefore, there is a great demand in developing efficient solutions for Web service, no matter in commerce or industry. A major topic in Web usage mining is mining association rules [11]. An association rule is described by the dependence among Web pages in a Website. The support-confidence framework is well known to determine statistically significant association rules [3,4]. However, there are some drawbacks of this framework while mining association patterns in a Website. For example, when two sets of Web pages $A$ and $B$ are very likely to be visited together, we obtain the association rule $A \Rightarrow B$ or $B \Rightarrow A$. However, if $A$ and $B$ have already been connected by hyperlinks, then such association rule $A \Rightarrow B$ or $B \Rightarrow A$ does not have a high interest. Moreover, an association rule with a very high support rarely exists in Web log mining. This is because in a complex Website with variety of Web pages, and many paths and hyperlinks, one should not expect that in a given time period, a large number of visitors follow only a few paths or Web pages. Otherwise, it would mean that the structure and content of the Website have a
serious problem. Therefore, in general, there are many association rules of about the same support obtained from the Web logs. It is necessary for us to find an additional interestingness measure to identify really interesting association rules. In this paper, we propose a novel interestingness measure called rule interest that is based on Website topology to analyze association rules.

One of the objectives of Web usage mining is to improve the organization of a Website by learning from Web logs. Previous works [6,8,9,10] have focused on the optimization problem of Website topology from different aspects. For instance, Perkowitz and Etzioni [8,9,10] have used clustering algorithms for indexing page synthesis to create an adaptive Website. Garofalakis et al. [6] have used page popularity to rearrange Website structure to make it more accessible and more effective. In this paper, we employ association rules with high interest to develop a graph-based optimization algorithm to modify the current Website topology that can cater to a large population of visitors. Comparing to the other Website optimization solutions, the most virtue of our method is that we not only use on Web logs to analyze user access patterns, but also use the original Website topology together to determine an effective Website structure.

The rest of this paper is organized as follows. In Section 2, we introduce the concept and the calculation of rule interest. In Section 3, we propose a graph-based optimization method to modify a Website topology according to interesting association rules. Section 4 describes the complexity analysis of our algorithm. Experimental results are presented in Section 5. Finally, some concluding remarks are given in Section 6.

2 Discovery of Interesting Association Rules in Website

In order to discover interesting usage patterns from Web logs, we define a new interestingness measure called rule interest in association rules. By using rule interest, together with rule support and confidence, we can identify association rules which are more meaningful to our analysis of visitors’ behaviors.

2.1 Website Topology

First, we regard Website topology as the structure of a Website. The nodes in a Website topology represent the Web pages and the edges among nodes represent the hyperlinks among Web pages. We assume that there is at least one edge to connect every node, that is, every Web page in the Website can be visited through at least one path. Figure 1 shows an example of Website topology. All Web pages are assigned with unique labels.

Table 1. An example of transaction itemsets

<table>
<thead>
<tr>
<th>TID</th>
<th>Itemset</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>ADCE</td>
</tr>
<tr>
<td>200</td>
<td>ABG</td>
</tr>
<tr>
<td>300</td>
<td>BFE</td>
</tr>
<tr>
<td>400</td>
<td>CH</td>
</tr>
</tbody>
</table>

**Fig 1.** A website topology example