Personalized Web Recommendation Based on Path Clustering

Yijun Yu, Huaizhong Lin, Yimin Yu, and Chun Chen

Computer Institute, Zhejiang University, 310027 Hangzhou, China
yijunyu@mail.hz.zj.cn, linhz@zju.edu.cn,
billyuym@163.com, chenc@cs.zju.edu.cn

Abstract. Each user accesses a Website with certain interests. The interest can be manifested by the sequence of each Web user access. The access paths of all Web users can be clustered. The effectiveness and efficiency are two problems in clustering algorithms. This paper provides a clustering algorithm for personalized Web recommendation. It is path clustering based on competitive agglomeration (PCCA). The path similarity and the center of a cluster are defined for the proposed algorithm. The algorithm relies on competitive agglomeration to get best cluster numbers automatically. Recommending based on the algorithm doesn't disturb users and needn't any registration information. Experiments are performed to compare the proposed algorithm with two other algorithms and the results show that the improvement of recommending performance is significant.

1 Introduction

Currently World Wide Web is developing rapidly. The managers of Websites need good auto-subsidiary tool, which can adjust the configuration of Web page dynamically according to the user's interest. Thus they can improve service and develop peculiar electronic business to satisfy the demands of visitors much better. For the visitors, what they want is a characteristic Web page with much better service that can satisfy various demands and they also expect to receive an illumination from other users who have a similar access interest. From some points of view, visitors are also not very clear about these demands. Moreover, an important method to solve these two demands is to use Web Usage mining to recommend Web personalized Web page.

The term 'Web Usage Mining'[1] was introduced by Cooley et al., in 1997, in which they define Web usage mining as the 'automatic discovery of user access patterns from Web Servers'. Web Usage mining has gained much attention in the literature as a potential approach to fulfill the requirement of Web personalization[2]. Personalized Web recommendation is one form of Web personalization that could find important applications in e-business (such as google.com and Amazon.com) and e-learning sectors. Web pages are recommended to a user according to user interest that anticipates the user's needs. In this paper, we focus on the personalized Web recommendation of Web pages that are adapted according to user interest.

H. Legind Larsen et al. (Eds.): FQAS 2006, LNAI 4027, pp. 368 – 377, 2006.

User access pattern is important in personalized Web recommendation system. The drawbacks of previous methods[3,4] are that not it ignores the facts that the available element will be close to zero with the increase of the user's access time and the number of clustering is given by people, which can not dynamically adjust according to personal access. Most of the research in personalized Web recommendation recently is collaborative filtering[5,6], which needs extra information to distinguish personal activities. However, this input may have some errors and may be outdated. Authors in Paper [7,8] use K-means clustering to recommend Web personalized page, which does not consider the relationship between the user's personality and the access path. The algorithm presented in this paper can overcome these drawbacks.

2 Path Clustering Based on Competitive Agglomeration

Path clustering based on competitive agglomeration (PCCA) is a partition algorithm, not a hierarchical clustering algorithm. The algorithm clusters according to path similarity, improves K-paths algorithm [9], and can get best cluster numbers automatically by competitive agglomeration method.

2.1 Definitions

We should consider that user access is sequence for Web page in clustering. Web users access Web page according to their interest, of course, their interest is different. So, the sequence can present user access interest.

Definition 1. The user access transaction t is defined as:

$$t=<l_1^t.url\ ,\ldots,\ l_r^t.url>, \tag{1}$$

Here, L is the user's access log set, $l \in L$, $l_i^t.url$ is hyperlink address of the Web page.

For urls in current accessing Web page, people always have been more interested in url accessed earlier than url accessed later. It is obviously that there is relation between user interest and user access path. So, we define user interest as follow.

Definition 2. The user's interest is defined as:

$$\mathrm{I}(l_i^t.url)=m-i+1,\quad 1 \le i \le m \tag{2}$$

Here, $l_i^t.url$ is hyperlink address of the Web page, m is the number of Web page in Website. It is obviously that there is relation as below.

$$\mathrm{I}(l_1^t.url)>\mathrm{I}(l_2^t.url)>\ldots>\mathrm{I}(l_i^t.url),\ 1 \le i \le m$$

Definition 3. Given two user access transaction t and s, the similarity between them sim(t,s) is defined as: