Mining Interest Navigation Patterns Based on Hybrid Markov Model

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 interest. The interest is associated with his navigation patterns. The interest navigation patterns represent different interest of the users. In this paper, hybrid Markov model is proposed for interest navigation pattern discovery. The novel model is better in prediction overlay rate and prediction correct rate than traditional Markov models. User group interest is also defined in this paper. The probability of user group interest navigation from one page to another is computed by navigation path characteristics and time characteristics. Compared with the previous ones, the results of the experiment show that the performance is improved efficiently by the hybrid Markov model.

l Introduction

When a user accesses a Website, he has some interest and different users have different interest when they access the Website through different paths. The different interest of the users is associated with interest navigation patterns. So, mining navigation pattern can reflect the users' access interest.

There are many efforts toward mining various patterns from Web logs, such as "Footprints" [1], "WUM" [2], etc. "Footprints" takes an optimizing approach. Its idea is that visitors to a Website leave their "Footprints" behind. Over time, "Paths" accumulate in the most heavily traveled areas. New visitors to the site can use these well worn paths as indicators of the most interesting pages to access. "WUM" improves this approach. It defines the g-sequences in order to mine the navigation patterns and gives a mining language MINT. Chen et al[3] map the log data into relational tables and employ the standard data mining approaches to discover the user navigation patterns. Borges and Levene [4] apply hypertext probabilistic grammar to discover the user navigation patterns and propose the use of entropy as an estimator of the statistical properties of the grammar.

All these approaches ignore the facts that users are actually interested in only a certain part of every Web page and the characteristics of every access Web page can reflect interest intensity of Web page content. They only discover the navigation patterns according to the users' access sequence.

In this paper, we define user group and user group interest intensity. Based on these definitions, we present hybrid Markov model for interest navigation patterns.

H. Legind Larsen et al. (Eds.): FQAS 2006, LNAI 4027, pp. 470 – 478, 2006.

The novel model can help the designer of the Website understand the users' interest better and improve the design of the Website. At the same time, the model can help users to understand their access behaviors and access content better.

2 Definitions

2.1 Keyword of User Access Page

When users access Website, there are some basic facts, such as: Every user accesses Website by different paths, every Web page in Website includes one keyword at least and these keywords can be used as representation of pages' contents. The Web pages the user access should be those that users are interested because users have intentions when they access a Website.

The user access keyword represented as *k* set indicates the contents of a Web page and they are a simple description of the contents of a Web page. A Web page may include several keywords. The contents of Web page can be represented by the Web page's keywords. In order to indicate user's access interest, we give the definition as follows:

Definition 1. The transaction of the user access keyword: The user access transaction is composed of Web pages that the user accesses. Every Web page is represented by a group of keywords. As every url comprises of several keyword σ and every keyword is represented by *K* set, the user access keyword transaction is defined as follows:

$$t_k(u_i, url_s) = < k_1, k_2, \ldots\ldots, k_j > \tag{1}$$

Here, $\boldsymbol{u}$ is the user's set, u_i is the user who accesses Web page currently, url_s is currently accessed Web page, k_j is currently accessed keyword.

The support to the user access keyword is the access number of certain keyword k_j in a user access keyword transaction through url_s, defined as $\text{support}(u_i, url_s, k_j)$.

The user access time length is composed of four parts. There are page loading time, page sending time, user view time and user request time. As the log time is recorded by seconds, page loading time and page sending time are close to zero, the time length is composed of user viewing time and page sending time. For a Website, page sending time should be less than time window *C*—a constant, which is mostly several seconds. As for pages that users are interested, time length is longer than *C* value.

Definition 2. The time length of user access keyword: the user will access longer time in the page if he is interested in one concept, suppose a url_s's access time length as $\mathrm{L}(u_i, url_s)$, if the page has *f* keywords $k_1, k_2, \ldots\ldots, k_f$, the time length of the user accesses keyword k_j is :

$$\text{length}(u_i, url_s, k_j) = \begin{cases} \dfrac{\mathrm{L}(u_i, url_s)}{f} & \text{if } k_j \text{ in } url_s; \\ 0 & \text{if } k_j \text{ not in } url_s. \end{cases} \tag{2}$$