USERS' TWO-DIMENSIONAL INTEREST AND ITS STORAGE METHOD*

Xu Baowen* ** Zhang Weifeng*

* (Department of Computer Science and Engineering, Southeast Univ., Nanjing 210096)
** (State Key Laboratory of Software Engineering, Wuhan University, Wuhan 430072)

Abstract The users' interest can be mined from the web cache and can be used widely. The interest can be specialized by the two-tuple (term, weight) in the simple interest model, in which the association relations are not mined, and then the interest cannot be associated in expressing the users' interest. Based on analyzing the WWW cache model, this letter brings forward a two-dimensional interest model and gives the interrelated methods on how to store the two-dimensional interest model effectively.

Key words WWW; Internet; Interest model; Data mining; Cache

I. Introduction

The WWW cache technology is an important measure to improve the performance of WWW in the definite bandwidth. In the WWW cache technology, the popular web pages are stored in the places near the users, and then the speed of accessing these web pages can be improved. At present, this technology is generally used in the Internet explorers and WWW proxies. There are historical informations recently visited by the users in the WWW cache. The traditional scheduling method of WWW cache such as LRU, LFU[1], and so on, stores the web pages, which are recently visited and frequently used, into the WWW cache[1]. In this view, the information in the cache shows the users' recent interest states. The users' interest information can be widely used in customizing the web page, filtering the information, pre-fetching the information, and so on[2–4]. There have been a lot of researches on how to mine the characteristic information and how to produce the users' interest. The interest is usually described by the pair (term, weight) in these interest models, which are named as the simple interest models. The simple interest models do not mine the association rules between the interests, so the associations cannot be realized when expressing the users' interest. As well known, the knowledge should be associated with each other based on the composing of the knowledge. So the users' interest cannot be denoted relevantly.

To solve the problems of the simple interest model in expressing the users' interest, in this letter, we bring forward a two-dimensional interest model. This model does not simply extend the simple interest model, but fully improves the model and corresponding arithmetic. The reasoning relations are considered well by importing the concepts of two dimensions. Compared with the simple interest model, the two-dimensional interest model need to concern how to store the two-dimensional model effectively.

In this letter, we will first introduce the simple interest model and its insufficiency. Secondly, we will present a two-dimensional interest model and analyze its storage problem.

*Supported in part by the National Natural Science Foundation of China (No.60073012), Natural Science Foundation of Jiangsu(BK2001004), Visiting Scholar Foundation of Key Lab. in University.
Finally, aiming at the storage problem, we give the rough set based storage method.

II. Simple Interest Model and Its Insufficiency

The simple interest model is a simple method describing the users’ interest. In this model, the users’ interest is represented as a two-tuple, (interest term, interest weight). Some interest can construct a interest set. The set of all interest makes up of the full set (the dictionary). The full set of interests is denoted as \( T = \{t_1, t_2, \cdots, t_m\} \), where \( t_1, t_2, \cdots, t_m \) represent the interest lemmas, respectively, \( m \) is the size of the dictionary \( T \). Based on the interest set, some basic concepts are given first for the easy discussion.

The web pages in the WWW cache show the users’ recent interest states. The interest can be transformed into interest node according to some style. Before we introduce how to transform the users’ interest information into the interest nodes, the web pages should be described in some styles. The text set in the WWW cache is denoted as \( D \), \( D == \{d_1, d_2, \cdots, d_n\} \). For the lemma \( t_i \) in the dictionary, its frequency appearing in the document \( d_j \) (term frequency) is denoted as \( tf_{ij} \); and its all frequency (no matter how many times it appears, it is recorded as 1) is denoted as \( df_i \) (the document frequency). The reciprocal of \( df_i \) is named as the invert document frequency, and is denoted as \( idf_i \).

For the simple interest model, it is an important part to compute the weights of the interest nodes. The data mining of the simple interest model is mainly the text based data mining\(^2\). In its interest computing method, all the documents in \( D \) are looked as one hypertext. The term frequency of the lemma in this hypertext document is equal to the interest weight of the interest lemma. This method is simple, but there exist some problems inescapably. For example, if one lemma’s frequency is too big, the lemma can not distinguish the users’ interest definitely.

With the improvement of computer performance, the storage space and run speed, the size of the WWW cache is becoming larger and larger, which ensures that the users’ interest can be reviewed in another point of view. How to mine the users’ potential interest from the WWW cache has become the emphases of research currently. Based on fully analyzing the WWW model and the WWW cache structure model, we regard that the WWW content mining, WWW structure mining and WWW log mining should be used synthetically to mine the users’ interest as sufficiently as possible. Therefore, we specially bring forward the two-dimensional interest model, which can be used to describe the users’ interest mined from the WWW cache.

III. Two-dimensional Interest Model

In the two-dimensional interest model, the users’ interest model is extended to two dimensions, i.e., not only the interest weight is considered, but also the reasoning relations are considered. This model extends greatly the connotation of the interest, makes the interest not be self-governed, and leads to the expanding of information to store the model.

Definition 1 The interest association rule is a triple, Rule \((Node(t_s), Node(t_t))\) \(= (Node(t_s), \text{weight}, Node(t_t))\), where weight is the probability from the interest node \( Node(t_s) \cdot t_s \) to the interest node \( Node(t_t) \cdot t_t \), \( 0 < \text{weight} \leq 1 \).

Definition 2 The interest association repository is a set of interest association rules, where

\[
\text{RULE} = \{ \text{Rule} \,(\text{Node} \,(t_s), \text{Node} \,(t_t)) \mid \text{Rule}(\text{Node} \,(t_s), \text{Node}(t_t)) \, \text{is a rule} \}
\]

The interest association rules in the interest association repository must satisfy:

\[
\sum_{P(Node(t_s))} \text{Rule} \,(\text{Node} \,(t_s), \text{Node}(t_t)) \cdot \text{weight} = 1,
\]