The Web Query Model Based on Statistical Order

Sha Huiping
GuiLin Statistics Bureau
6 Diecai Road
Guilin 541001, P.R.C
E-mail: swppp@163.net

Zhu Guohun
GuiLin Institute of Electronic Technology
1 Jinji Road
Guilin 541004, P.R.C
E-mail: zhuguohun@8848.net

1. Introduction

With the development and application of Internet, and its resources a nd data rapidly growing, the querying problem appears. There are two methods to solve the problem. One method is creating visible navigation image, but only used in the current or other several networks. The other is using Robot resource discovering process, which creates index database by researching relative web page, and provides query patterns such as text retrieval or others. The query results are graded or ordered by certain algorithms or rules, but this technology still has its inadequacy. We propose a model method according users' visiting and relevant degree of information. Because people in same area receive almost same education and culture background, on WWW, most of people use most of time visiting few web sites. The train of thought imitates principles of ATM exchanger based on statistical multiplexing and program locality.

2. Net Searching and Bookmark Manual Disposing

For searching special information manually, we should find series of relative network and put them to my favorite as bookmarks in order to carry the scanning much better. We could possible get rid of web sites which we didn' t visit for long from my favorite.

3. Data Model

There are three elements in data model: data definition, data manipulation and data constraint.
(1) Data definition
First, we define three main data set of model:
<Id, Key, Source_SubIp, Access_Times>
  <Id, Key.Denst_SubIp, Weight>
  <Id, Key, Source_SubIp, Denst_SubIp, Access_Times>
Id means the only identification character. Key is the key query word . Source_SubIp is Ip subnet that we often visit and search for the Key word in this web site. Denst_SubIp is Ip subnet that often contains the keyword. For example, the web sites of GuiLin have "GuiLin Landscape" as Key, so its Ip 202.103.224.*~ ~202.103.268.* is the Denst_SubIp. Access_Times is the number of visit and Weight is visiting degree, both dynamic and variable.

According to key, system could find correlative kind, and find visitor's Ip. According to the relationship between Denst_SubIp and Source_SubIp, system decides its class. Also, using Access_Times help sorting, ordering for visiting.

(2) Data manipulation

Primary operation is mainly according to the above three sets. Its form shows a binary table relationship. As a result, we can management it with the existing database knowledge. Main operations are:

--Selection Operation. Its essence is selection operation in relational algebra Ó (set, term). You can complete it by specifying requested sets and filtrating

it at requirement.

--Join. Its essence is join operation in relational algebra Ë (A X B, connection term).

--Knowledge acquisition. Its essence is adjusting with amount of visitors and times of key word.

--Set Operation. It refers to the formal set operations of point set, such as union, intersection.

--Times Operation. Once time is over, but still nobody visit. All sorts of weight will bring down automatically.

(3) Data constraint

This model mainly has constraints as follows: Id is unique in three sets. Source_SubIp and Denst_SubIp are not void.

4. Conclusion

Practice verifies that the model is authentic and available. To the full text retrieval system that becomes mature day by day, it is primary to reduce retrieval time. That is to say that the purpose is to provide specialty query link for users, and draw useful information from users.