Journal of Korean
Data & Information Science Society
2005, Vol. 16, No. 1, pp. 157~163

Designing Summary Tables for Mining Web Log Data

Jeong Yong Ahn[1)]

Abstract

In the Web, the data is generally gathered automatically by Web servers and collected in server or access logs. However, as users access larger and larger amounts of data, query response times to extract information inevitably get slower. A method to resolve this issue is the use of summary tables. In this short note, we design a prototype of summary tables that can efficiently extract information from Web log data. We also present the relative performance of the summary tables against a sampling technique and a method that uses raw data.

Keywords : Association rules, Preprocessing, Summary tables, Web data mining

1. Introduction

With the advances of the Web, it is being used for many commercial purposes. Electronic commerce(EC) is increasing by leaps and bounds and is critical for business. EC enhances the competitiveness of organizations by lowering transaction costs, focusing on differentiating their products and services (Lederer et al., 1998; Lin, 2000), and providing a rich source of information about the customers. According to the strength of EC, the development of new business models supported by information technology, for example CRM(customer relationship management) and BI(business intelligence), has become an important issue in the business community (Wang, 2001).

Web-based organizations often generate and collect large volumes of data such as Web logs in their daily operations. The organizations are looking for easy to use, powerful ad hoc query, reporting and analysis solutions that will empower their staff and customers to access the information they need from corporate data

1)) Associate Professor, Division of Mathematics and Statistical Informatics, Chonbuk National University, Chonbuk, 561-756 Korea.
E-mail : jyahn@chonbuk.ac.kr

warehouses, departmental data marts or online transaction processing systems, in order to make informed business decisions. However, as users access larger and larger amounts of data, query response times inevitably get slower.

A method to resolve this issue is the use of summary tables. A summary table pre-computes the answer to frequently asked queries so that the results can be returned over and over again, avoiding recalculation of the same results. Summary database tables can create not only different results in both query response times and database architecture compared to traditional relational databases but also many advantages for users within a CRM/BI environment, and are able to support a variety of data mining tasks (Hou, 1999).

There are many efforts towards mining information from log data (Cooley et al., 1999; Pei et al., 2000; Lee and Kim, 2003; Koh and Lee, 2004). However, most studies are focused to extract some information, and usually provide little or no design concepts and implementations of summary tables. In this article, we design summary tables that can efficiently extract information from Web log data. With the summary tables, we can get more effective results such as efficiency of time and producing the data information without loss of generality. We also present experimental results of applying the summary tables against a sampling technique and a method that uses raw data.

2. Design of summary tables

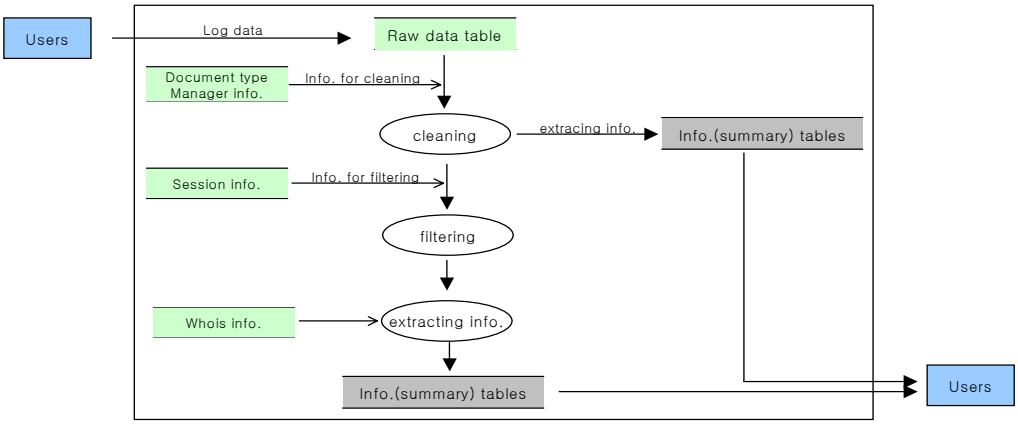
Web log data are usually large volumes of data. Log data more than tens or hundreds of giga-bytes are accumulated per day in cases of big Web site, although there are many differences according to each Web site. Therefore, we must consider the property for the first time when we analyze log data.

One of the methods for the analysis of large volumes of data is to use sampling techniques from a statistical viewpoint. The method is very significant because most statistical information can be replaced by approximate information. Recently, the advance of information technology and sampling algorithms from databases offers an environment that can take advantage of this method more easily, and we expect many studies in this field in the future. However, the method can not be free from large amounts of data.

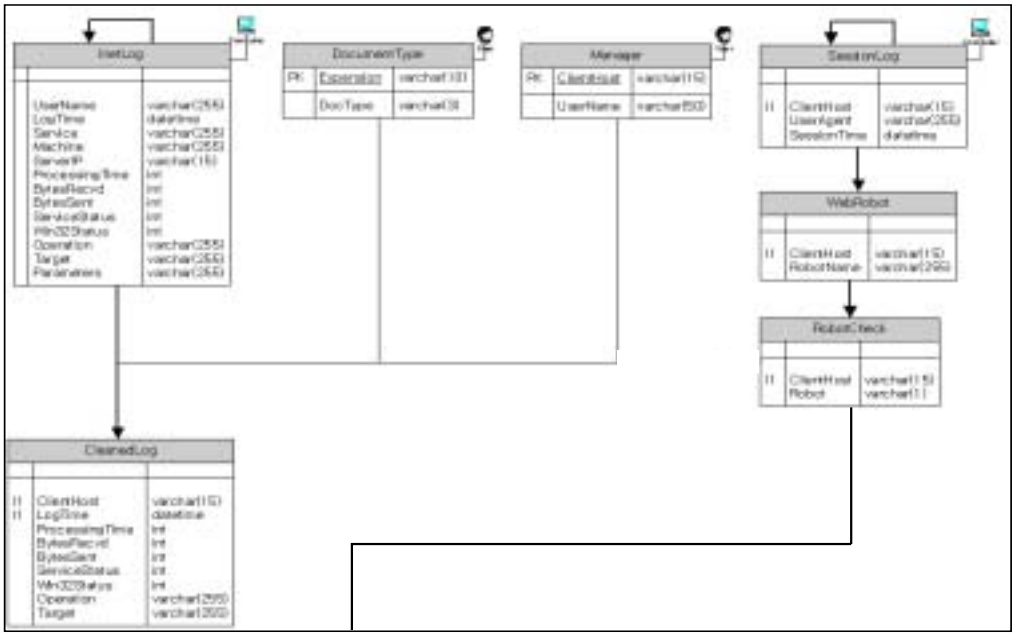
Another method is to use summary tables. Summary tables hold pre-aggregated and pre-joined data. In addition they hold results of frequently run queries that take a long time to complete. Therefore, the method can improve query response times (Chen et al., 1988; Han, 1998). Log data has the features in which include redundant or irrelevant data in addition to large volumes, and additional information is required for analysis.

When we consider these features, the method is more appropriate than the other methods. The summary tables designed in this study have pre-computed/ aggregated/joined summary information for analysis of log data. We remove

redundant data, store necessary information on summary tables, and use it in the analysis. The main concepts to design summary tables are as follows: (a) The summary tables integrate all additional information such as WHOIS information, Web administrator, Web search engine and so on. (b) All processes to summarize and aggregate the data are automatically conducted in step by step. It is an evolution process to update the summary tables. For this work, we use the agent functions for data management of DBMS. (c) The summary information is available to end users. The users obtain information using the SELECT access.



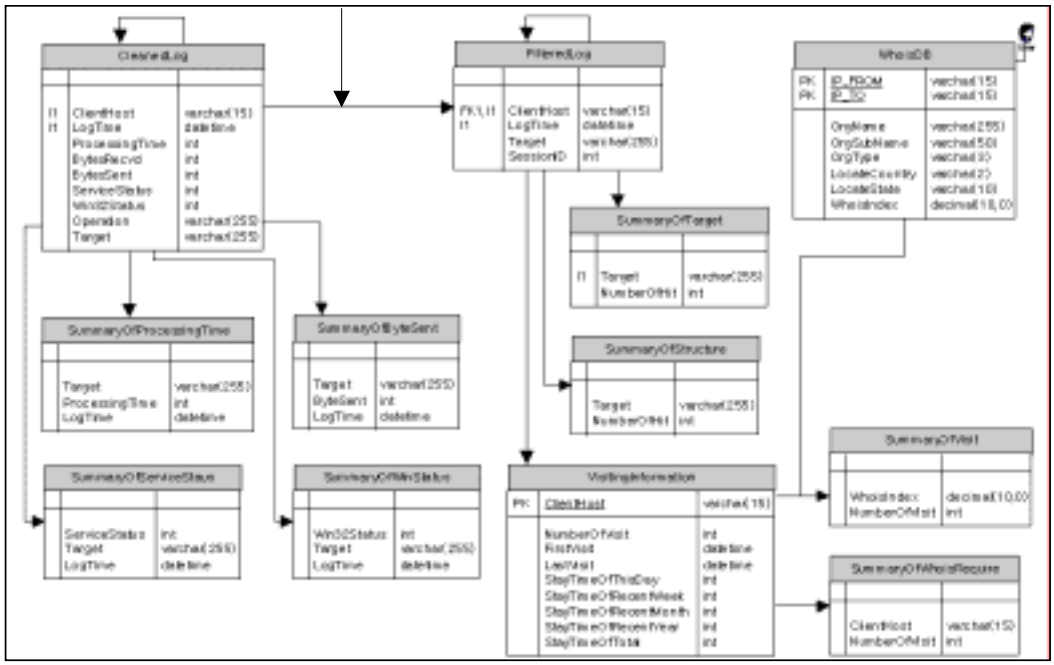
<Fig. 1> Data flow diagram



<Fig. 2> Design of summary tables - I

Fig. 1 is a data flow diagram that present the process to make summary tables from log data. First, log data is stored in a table. The stored data have to pass the process of data cleaning. Data cleaning is performed by using the several information such as document type, manager information, and so on. We extract some information from the cleaned data. After then, cleaned data have the process of filtering. In the process, cleaned data have to be partitioned into logical clusters that represent a single user transaction. We can extract some information from the filtered data.

Fig. 2 and Fig. 3 present the structure of summary tables designed in this study. A quadrangle in the figure is a table. For example, first quadrangle is a table called 'InetLog'. The summary tables are designed as follows: All transactions (log entries, raw data) are stored in InetLog table. In other words, the data of the table are ones that did not undergo any preprocessing. The table, CleanedLog, has the data cleaned from InetLog. Data cleaning is performed by checking the suffix of the URL name and using the information of several tables such as information of document types, managers, Web robot, and so on.



<Fig. 3> Design of summary tables - II

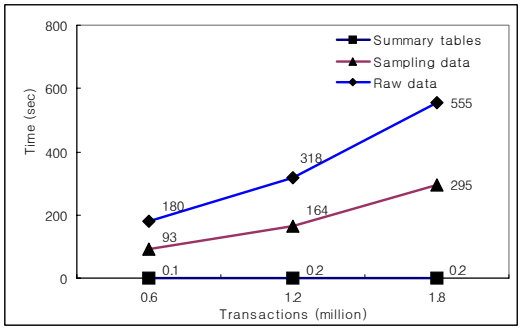
The table, FilteredLog, has the information of sessions (logical clusters) that represent a single user transaction and the visiting time of Web pages. One of the significant factors which distinguish Web data mining from other data mining activities is the process for identifying user log entries (Ahn, 2002; Mobasher et

al., 1996), and this process requires much time. SummaryTarget stores the information of the visiting number of Web pages. SummaryServiceStatus and SummaryWinStatus have the information of traffics, and VisitingInformation has the information of visiting time for a week, a month, and a year recent.

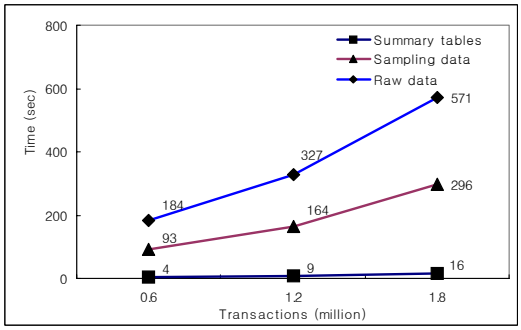
3. Performance

To assess the relative performance of the method that uses summary tables, we performed some experiments on an IBM PC(Pentium IV) with a CPU clock rate of 1.5GHz and 1.5GB of main memory. For experiments, we use 3 data sets. The number of transactions of the data sets is approximately 0.6, 1.2, 1.8 million respectively. And the number of sessions is approximately 15,100, 33,500, 51,700 respectively.

The experiments compare the performance of the three methods that use summary tables, raw data, and sampling technique. Fig. 4 and Fig. 5 show the execution time for hit counts of Web pages from SummaryTarget and discovering association rules from FilteredLog respectively. Association rule is a data mining technique to extract relationships from data collected by Web servers. We use the rules to discover the visiting propensity to Web pages of users. For example, we can discover the following relationships: 30% of users who accessed the Web page with URL /compstat/default.htm, also accessed the page /compstat/resource/link/list.asp.



<Fig. 4> Hit counts



<Fig. 5> Association rules

Execution time of the method that uses raw data includes following process times: data cleaning, identifying sessions, and discovering information. For sampling technique, we draw 5% (sample sizes are approximately 3,000, 5,600, 8,900 respectively) of cleaned data using the rejection method algorithm proposed in Vitter(1984). In Mannila et al.(1994), it is pointed out that a sample size of 3000 log entries gives an extremely good approximation of the large data sets. We

think, therefore, our sample sizes are appropriate. Execution time of the sampling technique includes the process times for data cleaning, sampling, identifying sessions, and discovering information.

Summary tables designed in above section have the pre-joined/pre-computed information. This method requires only the process times for discovering information, hence don't takes much time. As shown in Fig. 5 (The figure is a result in case that minimum support is 0.15), the execution times of summary table method are very efficient.

4. Conclusions

Web data mining is the application of data mining technologies to huge Web data repositories and an important research area in Web-based business model. Analyzing the Web access logs can help understand the customer behaviors and the Web structures and capturing the characteristics of the customers of a business web site is an important task for their marketing department.

In this study, we design a prototype of summary tables to analyze log data. It is physically difficult to create all the summary tables required to support all possible queries. Summary tables, however, are the largest factor influencing performance and scalability of a data warehouse implementation. With using the summary tables designed in this study, we can get more effective results as follows: efficiency of calculation time expended in data analysis, efficient utilization of additional information, and producing the data information without loss of generality.

References

1. Ahn, J. Y. (2002), A study on the mining access patterns from Web log data, *IEICE Transactions on Information and Systems*, E85-D(4), 782-785.
2. Chen, M.C., McNamee, L. P. and Melkanoff, M. (1988), A model of summary data and its applications in statistical databases, *Proc. of the International Conference on Statistical and Scientific Database Managemen*t, 356-372.
3. Cooley, R., Mobasher B. and Srivastava, J. (1999), Data preparation for mining world wide web browsing patterns, *Knowledge and Information Systems*, 1(1), 5-32.
4. Han, J. (1998), Towards on-line analytical mining in large databases, *ACM SIGMOD Record*, 27(1), 97-107.
5. Hou, W. C. (1999), A framework for statistical data mining with

summary tables, *Proc. of the International Conference on Scientific and Statistical Database Management*, 14-23.
6. Koh, B. S. and Lee, G. E. (2004), Web Log Analysis System Using SAS/AF, *Journal of the Korean Data & Information Science Society*, 15(2), 317-329.
7. Lederer, A. L., Mirchandi D. A. and Sims, K. (1998), Using WISs to enhance competitiveness, *Communications of the ACM*, 41(7), 94-95.
8. Lee, S. B. and Kim, M. S. (2003), A study of the reliability of Web services using client sides errors, *Journal of the Korean Data & Information Science Society*, 14(2), 217-221.
9. Lin, B. (2000), Electronic commerce: the emerging technology and its impacts, *Human Systems Management*, 19(4), 225-227.
10. Mannila, H., Toivonen H. and Verkamo, A. I. (1994), *Efficient algorithms for discovering association rules*, Knowledge discovery in databases, 181-192, Washington, AAAI Press.
11. Mobasher, B., Jain, N., Han E. H. and Srivastava, J. (1996), Web mining: pattern discovery from World Wide Web transactions, *Technical Report*, University of Minnesota.
12. Pei, J., Han, J., Mortazavi-asl, B. and Zhu, H. (2000), Mining access patterns efficiently from Web logs, *Proc. of the Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 396-407.
13. Vitter, J. S. (1984), Faster methods for random sampling, *Communications of the ACM*, 27(7), 703-718.
14. Wang, S. (2001), Designing information systems for electronic commerce, *Industrial Management and Data Systems*, 101(6), 304-314.

[received date : Oct. 2004, accepted date : Feb. 2005]