# Web Log Mining and Parallel SQL Based Execution

Masaru Kitsuregawa, Takahiko Shintani,
Takeshi Yoshizawa*, and Iko Pramudiono

Institute of Industrial Science, The University of Tokyo
7-22-1 Roppongi, Minato-ku, Tokyo 106, Japan
{kitsure,shintani,yoshi,iko}@tkl.iis.u-tokyo.ac.jp

**Abstract.** We performed association rule mining and sequence pattern mining against the access log which was accumulated at NTT Software Mobile Info Search portal site. Detail web log mining process and the rules we derived are reported in this paper. The integration of web data and relational database enables better management of web data. Some researches have even tried to implement applications such as web mining with SQL. Commercial RDBMSs support parallel execution of SQL. Parallelism is key to improve the performance. We showed that commercial RDBMS can achieve substantial speed up for web mining.

## 1 Introduction

The analysis of web log to understand the characteristics of web users has been one of the major topics in web mining. The goal is to provide personalized information retrieval mechanism that match the need of each individual user. "One to one marketing" on the web also has similar objectives. The development of the web personalization technologies will certainly benefit e-commerce too.

In this paper, we focused on the mining access log using association rule discovery techniques. We show some mining results from web log of Mobile Info Search(MIS), a location-aware search engine [18]. Usage mining of this unique site could give some interesting insights into the behavior of mobile device users which are the targets of this site.

Here we report some of the web mining techniques based on association rule that can be accomplished by some modified SQL queries on relational database. The integration of web with database techniques has drawn attention from researchers. Some have proposed query languages for the web that is similar with SQL such as Squeal[9] and WebSQL[5]. They emphasize better organization of web data managed in relation database way. We extend this concept for real applications of web mining. We also address the performance problem by paralleling the execution of SQL queries.

Although the amount of log at MIS is not so large, generally at large portal site it tends to be very large. The log can reach several tens of GB per day. Just

* IBM Japan Co.,Ltd. 1-1, Nakase, Mihama-ku, Chiba-shi, Chiba 261-8522, Japan

S. Bhalla (Ed.): DNIS 2000, LNCS 1966, pp. 20–32, 2000.

one day log is not enough for mining. If we are going to use several weeks log, then we have to handle more than one terabyte of data. Single PC server cannot process such huge amount of data with reasonable amount of time.

On the other hand recently most major commercial database systems have included capabilities to support parallelization although no report available about how the parallelization affects the performance of complex query required by association rule mining. This fact motivated us to examine how efficiently SQL based association rule mining can be parallelized and speeded up using commercial parallel database system (IBM DB2 UDB EEE). We propose two techniques to enhance association rule mining query based on SETM [10]. And we have also compared the performance with commercial mining tool (IBM Intelligent Miner). Our performance evaluation shows that we can achieve comparable performance with commercial mining tool using only 4 nodes. Some considerable works on effective SQL queries to mine association rule such didn't examine the effect of parallelization [15][16]. Some of authors have reported a performance evaluation on PC cluster as parallel platform [13][14]. Comparison with natively coded programs is also reported. However we use currently available commercial products for the evaluation here.

## 2 Web Usage Mining for Portal Site

### 2.1 Web Access Log Mining

Access log of a web site records every user requests sent to the web server. From the access log we can know which pages were visited by the user, what kind of CGI request he submitted, when was the access and it also tells to some extent where the user come from. Using those information, we can modify the web site to satisfy the need of users better by providing better site map, change layout of the pages and the placement of links etc. [12] has proposed the concept of adaptive web site that dynamically optimize the structure of web pages based on the users access pattern. Some data mining techniques has been applicated on web logs to predict future user behavior and to derive marketing intelligence[21][7][8]. Currently many e-commerce applications also provides limited personalization based on access log analysis. Some pioneers such as Amazon.com have achieved considerable success.

Here we will show the mining process of real web site. We have collaborated with NTT Software to analyze the usage of a unique search engine called Mobile Info Search.

### 2.2 Mobile Info Search(MIS)

Mobile Info Search (MIS) is a research project conducted by NTT Software Laboratory whose goal to provide location aware information from the internet by collecting, structuring, organizing, and filtering in a practicable form[17]. MIS employs a mediator architecture. Between users and information sources,

MIS mediates database-type resources such as online maps, internet "yellow-pages" etc. using *Location-Oriented Meta Search* and static files using *Location Oriented Robot-based Search*. Users input their location using address, nearest station, latitude-longitude or postal number. If the user has a Personal Handy Phone(PHS) or Geo Positioning System(GPS) unit, the user location is automatically obtained.
The site is available to the public since 1997. Its URL is *http://www.kokono.net*. In average 500 searches are performed on the site daily. A snapshot of this site is shown in Figure 1. [1]

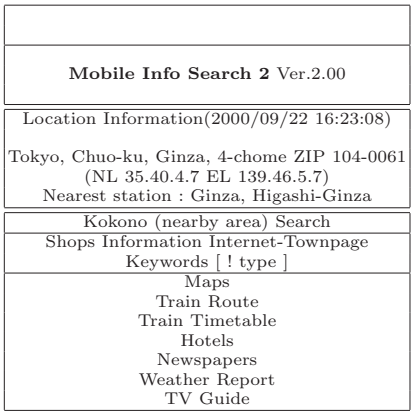


**Fig. 1.** Index page of Mobile Info Search

MIS has two main functionalities :

1. Location Oriented Meta Search
   Many local information on the web are database-type, that is the information is stored in backbone database. In contrast to static pages, they are accessed through CGI program of WWW server. MIS provides a simple interface for local information services which have various search interfaces. It converts the location information and picks the suitable *wrapper* for the requested service. Example of database-type resources provided are shown in table 1.
2. Location-Oriented Robot-Based Search "kokono Search"
   *kokono Search* provides the spatial search that searches the document close to a location. "kokono" is a Japanese word means *here*. *kokono Search* also employs "robot" to collects static documents from internet. While other search engines provide a keyword-based search, *kokono Search* do a location-based spatial search. It displays documents in the order of the distance between the location written in the document and the user's location.

[1] The page is shown in Japanese at *http://www.kokono.net*

**Table 1.** Database-type resources on the Internet

| Service | Location information used for the search |
|---|---|
| Maps | longitude-latitude |
| Yellow Pages | address ( and categories ... etc) |
| Train Time Tables | station |
| Weather Reports | address or region |
| Hotel Guides | nearest station |

## 3 Mining MIS Access Log and Its Derived Rules

### 3.1 Preprocessing

We analyzed the users' searches from the access log recorded on the server between January and May 1999. There are 1035532 accesses on the log, but the log also consists image retrieval, searches without cookie and pages that do not have relation with search. Those logs were removed. Finally we had 25731 search logs to be mined.

- Access Log Format
  Each search log consists CGI parameters such as location information (*address*, *station*, *zip*), location acquisition method (*from*), resource type (*submit*), the name of resource (*shop_web*, *map_web*, *rail_web*, *station_web*, *tv_web*), the condition of search (*keyword*, *shop_cond*). We treat those parameters the same way as items in transaction data of retail sales. In addition, we generate some items explaining the time of access (*access_week*, *access_hour*).
  Example of a search log is shown in Figure 2.
- Taxonomy of Location
  Since names of places follow some kind of hierarchy, such as "city is a part of prefecture" or "a town is a part of a city", we introduce taxonomy between them. We do this by adding items on part of CGI parameter *address*. For example, if we have an entry in CGI parameters entry [address=Yamanashi-ken, Koufu-shi, Oo-satomachi], we can add 2 items as ancestors : [address= Yamanashi-ken, Koufu-shi] at city level and [address=Yamanashi-ken] at prefecture level. In Japanese, "ken" means prefecture and "shi" means city.
- Transformation to Transaction Table
  Finally we have the access log is transformed into transaction table ready for association rule mining. Part of transaction table that corresponds to log entry in Figure 2 is shown in Table 2

### 3.2 Association Rule Mining

Agrawal et. al.[1] first suggested the problem of finding association rule from large database. An example of association rule mining is finding "if a customer

```
0000000003 - - [01/Jan/1999:00:30:46  0900] "GET /index.cgi?
sel_st=0&NL=35.37.4.289&EL=138.33.45.315&address=Yamanashi-ken,
Koufu-shi,Oosato-machi&station=Kokubo:Kaisumiyoshi:Minami-koufu:
Jouei&zip=400-0053&from=address&shop_web=townpage&keyword=
&shop_cond=blank&submit_map=Map&map_web=townpage&rail_web
=s_tranavi&station_web=ekimae&tv_web=tvguideHTTP/1.1" 200 1389
"http://www.kokono.net/mis2/mis2-header?date=1999/01/01.00:27:59
&address=Yamanashi-ken,Koufu-shi,Oosato-machi&NL=35.37.4.289&EL
=138.33.45.315&station=Kokubo:Kaisumiyoshi:Minami-koufu:Jouei&zip
=400-0053&from=address&keyword=&shop_web=townpage&shop_cond=blank
&map_web=townpage&station_web=&tv_web=tvguide""Mozilla/4.0
(compatible; MSIE 4.01; Windows 98)$B!I(B"LastPoint=NL=35.37.4.289
&EL=138.33.45.315&address=Yamanashi-ken,Koufu-shi,Oosato-machi&station
=Kokubo:Kaisumiyoshi:Minami-koufu:Jouei&zip=400-0053;  LastSelect
=shop_web=townpage&shop_cond=blank&keyword=&map_web=townpage&rail_web=
s_tranavi&station_web=ekimae&tv_web=tvguide;Apache=1; MIS=1" "-"
```

**Fig. 2.** Example of an access log

**Table 2.** Representation of access log in relational database

Relation *LOG*

| Log ID | User ID | Item |
|---|---|---|
| 001 | 003 | address=Yamanashi-ken ,Koufu-shi,Oosato-machi |
| 001 | 003 | address=Yamanashi-ken,Koufu-shi, |
| 001 | 003 | address=Yamanashi-ken, |
| 001 | 003 | station=Kokubo: Kaisumiyoshi:Minami-koufu:Jouei |
| 001 | 003 | zip=400-0053 |
| 001 | 003 | from=address |
| 001 | 003 | submit_map=Map |
| 001 | 003 | map_web=townpage |

buys A and B then 90% of them buy also C" in transaction databases of large retail organizations. This 90% value is called confidence of the rule. Another important parameter is support of an itemset, such as {A,B,C}, which is defined as the percentage of the itemset contained in the entire transactions. For above example, confidence can also be measured as support({A,B,C}) divided by support({A,B}).

We show some results in Table 3 and 4. Beside common parameters such as *confidence* and *support*, we also use *user* that indicate the percentage of users logs that contain the rule.

Those rules can be used to improve the value of web site. We can identify from the rules some access patterns of users that access this web site. For example, from the first rule we know that though Akihabara is a well known place in Tokyo

DOCKET ALARM

# Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

## Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

## Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

## Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

## API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

**LAW FIRMS**
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

**FINANCIAL INSTITUTIONS**
Litigation and bankruptcy checks for companies and debtors.

**E-DISCOVERY AND LEGAL VENDORS**
Sync your system to PACER to automate legal marketing.