User Behavior Analysis of Location Aware Search Engine

Iko Pramudiono, Takahiko Shintani,* Katsumi Takahashi,† Masaru Kitsuregawa
Institute of Industrial Science, The University of Tokyo
4-6-1 Komaba, Meguro-ku, Tokyo 153-8505, Japan
E-mail: {iko,shintani,katsumi,kitsure}@tkl.iis.u-tokyo.ac.jp

Abstract

Rapid growth of internet access from mobile users puts much importance on location specific information on the web. An unique web service called Mobile Info Search (MIS) from NTT Laboratories gathers the information and provide location aware search facilities. We performed association rule mining and sequence pattern mining against the access log which was accumulated at the MIS site in order to get some insight into the behavior of mobile users regarding the spatial information on the web. Detail web log mining process and the rules we derived are reported in this paper.

1. Introduction

The emergence of internet has been touted also as the emergence of "borderless world" where users of internet are free from any location restrictions.

However the rapid growth of internet access from mobile users might change the perception somehow. Mobile users use a wide range of devices, among others mobile phones, PDAs, car navigation systems etc. The loose definition of mobile internet users are people that are not fixed to any particular place. It seems that they might become the mainstream of the borderless internet. However since they are "mobile", they concern more about location. Particularly where they are, and the information about places or services they can reach nearby.

On the other hand, there are abundant pages on the Web that contain some kind of address or other form of location information. An experimental location-aware search engine called *Mobile Info Search* (MIS) from NTT Laboratories [10], collects those spatial information and provides a portal to access them. It targets a variety of mobile devices and also supports some position acquiring means. In this paper, we report the mining process of access log from MIS.

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.

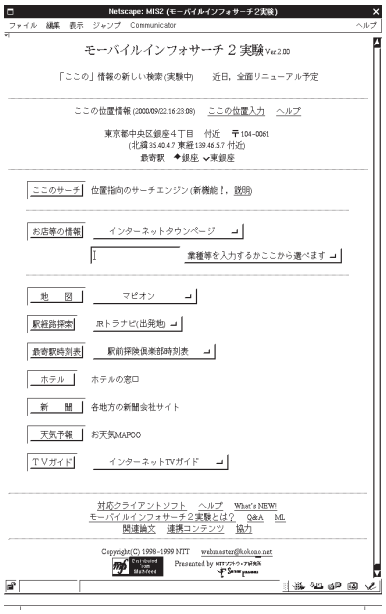
Using those information, we can also modify the web site to satisfy the need of users better by providing better site map, change layout of the pages and the placement of links etc. Perkowitz and Etzioni have proposed the concept of adaptive web site that dynamically optimize the structure of web pages based on the users access pattern[8]. The analysis of web log to understand the characteristics of web users has been one of the major topics in web mining. Some data mining techniques has been applied on web logs to predict future user behavior and to derive marketing intelligence[11][6][7]. Currently many e-commerce applications also provide limited personalization based on access log analysis. Some pioneers such as Amazon.com have achieved considerable success.

The mobility of users certainly will affect what information they want, and how they do shopping. Some e-services have tried to adapt their contents to the location of their visitors. However most of them based on manually defined rules and there are no study on its effectiveness yet.

Here we focus on mining the behavior of user with regard to his/her location. The unique features of MIS allow us to mine those knowledge from MIS access logs. Usage mining of this unique site could also give some interesting insights into the behavior of mobile device users which are the targets of this site.

We mainly use association rule mining and sequential pattern mining techniques. We also examine the addition of location hierarchy and methods to select interesting rules.

*Currently at Hitachi, Ltd., CentralResearch Laboratory 1-280, Higashi-koigakubo, Kokubunji-shi, Tokyo 185-8601, Japan
†NTT Information Sharing Platform Laboratories Midori-cho 3-9-11, Musashino-shi, Tokyo 180-8585, Japan



Mobile Info Search 2 Ver.2.00

Location Information(2000/09/22 16:23:08)

Tokyo, Chuo-ku, Ginza, 4-chome ZIP 104-0061
(NL 35.40.4.7 EL 139.46.5.7)
Nearest station : Ginza, Higashi-Ginza

Kokono (nearby area) Search

Shops Information Internet-Townpage
Keywords [! type]

Maps
Train Route
Train Timetable
Hotels
Newspapers
Weather Report
TV Guide

Fig.1. Index page of Mobile Info Search

2. Mobile Info Search(MIS)

Mobile Info Search (MIS) is a research project conducted by NTT Laboratories whose goal to provide location aware information from the internet by collecting, structuring, organizing, and filtering in a practicable form[9]. 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*.

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]

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

Service	Primary MIS CGI parameter	Location information used for the search
Maps	submit_map	longitude-latitude
Yellow Pages	submit_shop	address,categories
Train Time Tables	submit_station	station
Train Routes	submit_rail	station
Hotel Guides	submit_hotel	nearest station
Weather Reports	submit_weather	address or region
Local Newspaper	submit_newspaper	address or region
Local TV Guide	submit_tv	address or region

Table 1. Database-type resources on the Internet

2.1. User Location Acquisition

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 PHS service was launched in Japan in July 1995. Unlike conventional cellular telephone systems, PHS use many small base stations. Each base station serves a cell whose area is few hundred meters in diameter. Together they form a grid of cells. The base stations are placed in almost every stations, buildings and street crossings. The location acquired from PHS is not the exact location of the user but is actually the position of nearest base station.

Users can also input their location using some map softwares such as “ProAtlas”, or softwares to search train routes and schedule such as “Japan Railway(JR) Travel Navigator”.

2.2. MIS Functionalities

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

	Transactions	Users	Repeat
[from=address]	6749 (39.63%)	1386 (34.87%)	4.87
[from=zip]	3832 (22.50%)	982 (24.70%)	3.90
[from=station]	3418 (20.07%)	790 (19.87%)	4.33
[from=idokeido]	887 (5.21%)	159 (4.00%)	5.58
[from=tranavi95]	641 (3.76%)	209 (5.26%)	3.07
[from=pocke]	586 (3.44%)	216 (5.43%)	2.71
[from=proatlas]	247 (1.45%)	38 (0.96%)	6.50
[from=lkokonavi]	217 (1.27%)	54 (1.36%)	4.02
[from=demone]	206 (1.21%)	89 (2.24%)	2.31
[from=kokonogpslink]	132 (0.78%)	17 (0.43%)	7.76
[from=netscape]	44 (0.26%)	1 (0.03%)	44.00
[from=tranavice]	25 (0.15%)	12 (0.30%)	2.08
[from=lkokonavidemo1]	15 (0.09%)	9 (0.23%)	1.67

Table 2. Location acquisition methods

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. For example, since Institute of Industrial Science (IIS), The University of Tokyo is located in Komaba, when a user's location is at the IIS, *kokono Search* will return home pages that contain the word "Komaba" and other addresses in IIS vicinity.

3. Mining MIS Access Log and its Derived Rules

Here we will report some statistics of MIS site, the process of association rule mining and sequential rule mining and their results.

3.1. Site Statistics

From Table 2, most users still choose to manually input their position. However, there are 80 users with automatic location acquisition devices such as *kokonogpslink* (GPS) and *lkokonavi* (PHS), they performed 364 searches. The "repeat rate" here is simply the number of transactions divided by users.

The distribution of search types is shown in Table 3. Map accounts for almost half of the searches, followed by the *kokono Search* and search on Yellow Pages. Actually more users use the *kokono Search*, although the repeat rate is much lower. We will discuss the problems with *kokono Search* later.

	Transactions	Users	Repeat
map	8572 (45.90%)	1910 (28.34%)	4.49
kokono	4109 (22.00%)	2206 (32.73%)	1.86
shop	3128 (16.75%)	1028 (15.25%)	3.04
rail	1289 (6.90%)	539 (8.00%)	2.39
hotel	665 (3.56%)	397 (5.89%)	1.68
weather	375 (2.01%)	295 (4.38%)	1.27
newspaper	279 (1.49%)	183 (2.72%)	1.52
tv	260 (1.39%)	181 (2.69%)	1.44

Table 3. Search type

3.2. 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. We have also removed logs whose default location since our examination indicated that most of those logs are only trial accesses and do not represent actual mobile users behavior. Finally we had 20750 accesses to be mined.

We use cookie to identify unique users. We use typical time-out threshold for session identification. Our sessionizer ends a session if view time of a page is longer than 30 minutes. Thus we had 5576 sessions with 3642 unique users.

3.3. Access Log Format

Each search log consists CGI parameters such as location information ($address$, $station$, NL, EL, zip), location acquisition method ($from$), resource type ($submit$), the name of resource to search from ($shop\_web$, map_web, $rail\_web$, $station\_web$, tv_web), the condition of search ($keyword$, $shop\_cond$). Summary of the parameters are given in Table 4.

The latitude NL and longitude EL parameters can be set manually or automatically computed from $address$, zip or $station$ parameters. On the other hand, automatic position acquisition such as GPS fetches the latitude and longitude as CGI parameters, and MIS converts them into corresponding address, postal code and nearest station.

The type of web database used for search is specified by CGI primary parameters that begin with $submit\_$. The list of primary parameters is given in Table 1. Some CGI parameters are dependent to their primary parameter. For example parameters $shop\_cond$, $shop\_web$ and $keyword$ are not empty only when searching with Yellow Pages, i.e. when primary parameter $submit\_shop$ is specified. Thus when other search service is used those parameters are redundant, so we eliminated them from further processing of the log.

Parameter	Usage
NL	latitude
EL	longitude
address	address(es)
station	name of station(s)
zip	postal code
from	location acquisition method
keyword	search keyword
shop_cond	search category
submit_(*)	resource type for search
shop_web	resource name for Yellow Pages
map_web	resource name for map
rail_web	resource name for train route
station_web	resource name for train time table
tv_web	resource name for TV guide

Table 4. CGI parameters

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

Table 5. Representation of access log in relational database

```
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=tvguide
HTTP/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)"
"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" "-"
```

Figure 2. Example of an access log

We treat those parameters the same way as items in transaction data of retail sales. In addition, we generate some items describing the time of access ($access\_week$, $access\_hour$).

Example of a search log is shown in Figure 2.

3.4. 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.

The introduction of the hierarchy allows us to find not only rules specific to a location but also wider area that covers that location. This is useful since in many case the locations specified as search condition are sparsely distributed. That is, only few people choose “Oo-satomachi, Koufu-shi, Yamanashi-ken” since only few people have interest in Oo-satomachi that has few population. Thus we may not obtain any rule containing “Oo-satomachi” since the support is low. However the upper hierarchy levels of the search condition, i.e. “Koufu-shi, Yamanashi-ken” and “Yamanashi-ken”, are now counted as separate items. Their support will be higher since they also contain the support of locations below their hierarchy levels. We can expect to have some rules contain “Yamanashi-ken” with sufficient support.

However the hierarchy also adds computational burden during mining proses since the number of items increases significantly. We employ the optimization for generalized association rule mining with taxonomy to effectively prune the candidate itemset generation. [4]

3.5. Transformation to Transaction Table

Finally, we have the access log being transformed into transaction table ready for mining. Part of transaction table that corresponds to log entry in Figure 2 is shown in Table 5.

3.6. Association Rule Mining

Agrawal et. al.[1, 2] first suggested the problem of finding association rule from large database. An example of association rule mining is finding ”if a customer 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}).

Not so many good restaurants in Akihabara ?
[keyword=][address=Tokyo,][station=Akihabara] ⇒ [shop_cond=restaurant]
In Hokkaido, people looks for gasoline stand at night from its address
[access_hour=20][address=Hokkaido,][from=address] [shop_web=townpage] ⇒ [shop_cond=gasoline]
People from Gifu-ken quite often searches for restaurants
[address=Gifu-ken,][shop_web=townpage] ⇒ [shop_cond=restaurant]
However people from Gifu-ken search for hotels on Saturday
[access_week=Sat][address=Gifu-ken,] [shop_web=townpage] ⇒ [shop_cond=hotel]
People from Gifu-ken must search for hotel around stations
[address=Gifu-ken,][shop_web=townpage] [station=Kouyama] ⇒ [shop_cond=hotel]

Table 6. Some results of MIS log mining regarding search condition

Most frequent searches for restaurants around 16:00 if they start from address on Friday
[access_week=Fri][from=address][shop_cond=restaurant] ⇒ [access_hour=16]
Most frequent searches for department store stand at 20:00 if start from address.
[from=address][shop_cond=department] ⇒ [access_hour=20]
Looking for gasoline stand on Sunday ?
[from=address][shop_cond=gasoline][shop_web=townpage] ⇒ [access_week=Sun]
Search for hotels often from station if at Kanagawa-ken
[address=Kanagawa-ken,][shop_cond=hotel] ⇒ [from=station]
People at Osaka start searching convenience stores from ZIP number !
[address=Osaka,][shop_cond=conveni] ⇒ [from=zip]
People at Hokkaido always search convenience stores from address
[address=Hokkaido,][shop_cond=conveni] [shop_web=townpage] ⇒ [from=address]

Table 7. Some results of MIS log mining regarding time and location acquisition method

We show some results in Table 6 and 7. Beside common parameters such as *confidence* and *support*, we also use *user* that indicate the percentage of users that contain the rule. We also tried some other interestingness measures such as *lift* and *chi-square*, but we found that those measures are not so helpful to select useful rules. Instead we use the following method :

- Set the minimum confidence higher when the support of the rule is lower
- Remove items whose low contribution to the confidence of the rule
  For example, suppose we have a rule such as : [access_hour=16][address=Tokyo, Minato-ku][from=station][shop_web=townpage] ⇒ [shop_cond=restaurant] with 44% confidence and then when an item [shop_web=townpage] is removed the confidence becomes 41%. Thus the contribution of [shop_web=townpage] is only 3%, so we prefer the rule without the item.

After finding a shop, check how to go there and the weather
[submit_shop=Shop Info] → [submit_rail=Search Train] → [submit_newspaper=Newspaper] ⇒ [submit_weather=Weather Forecast]
Or decide the plan after checking the weather first
[submit_weather=Weather Forecast] → [submit_shop=Shop Info] [shop_web=townpage] → [submit_kokono=Kokono Search] ⇒ [submit_map=Map]
Looking for shops after closing time
[submit_shop=Shop Info] [access_hour=22] [access_week=Fri] ⇒ [submit_map=Map] [access_hour=22] [access_week=Fri]

Table 8. Some results of sequential pattern mining

Derived association 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 for electronic appliances/parts shopping, user that searches around Akihabara station will probably looks for restaurant. From this unexpected result, we can prefetch information of restaurant around Akihabara station to reduce access time, we can also provide links to this kind of user to make his navigation easier or offer proper advertisement banner. In addition, learning users behavior provides hint for business chance for example the first rule tell us the shortcoming of restaurants in Akihabara area.

Other search results show how location affects the search conditions. The second rule in Table 6 shows that people in Hokkaido, the largest and the most sparse prefecture in Japan, has particular problem to find gasoline stand at night. The rest of the rules show how people at Gifu-ken, a modest prefecture in the middle of Japan, often looks for restaurants. However more people, some of them might be travellers, looks for hotel around the station in the weekend.

Some results in Table 7 show that in addition to the location, time and location acquisition method might affect search conditions. For example, the third rule indicates that hotels in Kanagawa-ken are more likely searched from their nearest station since Kanagawa-ken, being the suburb of Tokyo, has extensive railway. In contrast, the last rule shows that people at Hokkaido are more comfortable to find convenience stores from the address.

3.7. Sequential Rule Mining

The problem of mining sequential patterns in a large database of customer transactions was also introduced by Agrawal et. al.[3, 5]. The transactions are ordered by the transaction time. A sequential pattern is an ordered list (sequence) of itemsets such as "5% of customers who buy both A and B also buy C in the next transaction".

We show some sequential patterns that might be interesting in Table 8. Some patterns indicate the behavior of users that might be planning to do shopping. We can derive from second pattern that significant part of users check the weather forecast first, then they look for the shops in the yellow-pages service called "Townpage" then look again for additional information in the vicinity with *kokono Search* and finally they confirm the exact location in the map.

4. Discussion and Summary

In this paper, we reported the result of mining web access log of a portal site for mobile users called Mobile Info Search. Two techniques are used : the association rule mining and sequential pattern mining.

Although the size of access logs of MIS is comparatively small, using those techniques we can figure out how the behavior of MIS users, which many of them are mobile users, and also the kind of services they use are affected by their location.

Unfortunately the number of mobile users with automatic location acquisition devices is too small so that there are no significant rule derived. A closer examination is needed to know whether their behavior is different than ordinary users.

We found that the spatial information is highly valuable to derive users' preferences, in particular mobile users. Our rules show that items with location information such as $address$ and $station$ increase the confidence of the rules significantly.

The current implementation of *kokono Search* only lists the pages in the vicinity of user's location. *kokono Search* automatically adjust the range of vicinity according to the number of pages. Unfortunately there are many case when the user is overwhelmed by so many pages, such as when the user is at shopping area. Clustering the search results on their contents will help the user to obtain the desired information. Such improved presentation of search results are expected to improve the low repeat rate of *kokono Search*.

We are also considering to perform clustering based on latitude-longitude information instead of address as our future work.

Acknowledgements

We would like to thank people from NTT Laboratories, in particular Dr. Atsuhiro Goto for providing the log file of MIS and helpful discussions.

References

[1] R. Agrawal, T. Imielinski, A. Swami. "Mining Association Rules between Sets of Items in Large Databases". In *Proc. of the ACM SIGMOD Conference on Management of Data*, May 1993.

[2] R. Agrawal, R. Srikant. "Fast Algorithms for Mining Association Rules". In *Proc. of the VLDB Conference*, June 1994.

[3] R. Agrawal, R. Srikant "Mining Sequential Patterns". 'In *Proc. of Int. Conf. on Data Engineering*, March 1995.

[4] R. Srikant, R. Agrawal "Mining Generalized Association Rules" 'In *Proc. of the VLDB Conference*, September 1995.

[5] R. Srikant, R. Agrawal "Mining Sequential Patterns: Generalizations and performance improvements" 'In *Proc. of 5th Int. Conf. on Extending Database Technology*, March 1996.

[6] A. Buchner, M. D. Mulvenna. "Discovering internet marketing intelligence through online analytical Web usage mining" In *SIGMOD Record* (4)27, 1999.

[7] R. Cooley, B. Mobasher, J. Srivistava. "Data preparation for mining World Wide Web browsing patterns" In *Journal of Knowledge and Information Systems* (1)1, 1999.

[8] M. Perkowitz, O. Etzioni. "Towards Adaptive Web Sites: Conceptual Framework and Case Study", In *Proc. of WWW8*, May 1999.

[9] Katsumi Takahashi, Seiji Yokoji, Nobuyuki Miura "Location Oriented Integration of Internet Information - Mobile Info Search". In *Designing the Digital City*, Springer-Verlag, March 2000.

[10] *http://www.kokono.net*. English version is *http://www.kokono.net/english/*.

[11] T. Yan, M. Jacobsen, H. Garcia-Molina, U. Dayal. "From user access patterns to dynamic hypertext linking" In *Proc. of WWW5*, May 1996.