Data Preparation for Mining Location-Specific Web Access Patterns in Mobile Environment
이동 환경하에서의 지역적 웹접근 형태 분석을 위한 자료 준비

Namshik-Ahn(Dept. of Multimedia)
안남식(멀티미디어과)

Key words : Mining, Mobile environment, GPS(Global Positioning System), MH(Mobile Host), MSS(Mobile Support Station)

ABSTRACT : Recent technological progress in mobile environment has made it possible that the users use the mobile internet massively in their daily lives. In this sense, this paper proposes the data preparation process for gathering location-specific web access patterns in mobile environment. The proposed system architecture, which includes the mobile computing environment and GPS facility, gathers the data on the location and web access patterns of the mobile devices (MH) such as PDA and hand-held PC. Additionally, the data structure for the gathered data is given to facilitate later information discovery process for finding the relationship between the location and web access pattern. Because of the practical difficulties, the data preparation process is simulated with the rational assumptions. Finally, providing the visualized scatter plot in many perspectives, this paper shows the linkage between the location and web access pattern in mobile environment.

1. Introduction

Recent progress in web technologies has changed the way of information access compared with the traditional way, which means visiting libraries or referring to the encylopedia. As more people perform their activities through web interface, the volume of web usage has increased drastically. Therefore, in this sense, the need of web mining, which provides better chances of making higher profit, has been recognized, e.g., especially in business area. Simultaneously, as the mobile technologies advance in increasing pace, many devices to support mobile communication including wireless PDA and hand-held PCs are commercially available. Keeping the pace with this environmental change, much more people have been accessing the web with the mobile devices continually changing their locations. In other words, mobile technology enables the users to continually access the web in different locations, not in a fixed location. Coherent with this trend, the need of analyzing the web usage in mobile environment has been recognized.

Web mining[4] in static environment, which means the user accesses the web from one fixed location, has the defect that it cannot analyze the web usage in mobile environment. In mobile environment, the user continually changes his/her position accessing the web. Therefore, existing web mining only dealing with static environment cannot properly provide the combined information with the locations and access patterns. For example, let us assume that there is a place full of theaters. In this case, many users with mobile devices are inclined to access the web for the information on movies. Therefore, if we find out the characteristics of a location determined by mobile internet access patterns without the prior knowledge, we can use this information for higher profit in business area.

In a specific location, the users are inclined to access the web sites which belong to a certain semantically pre-classified category. As a consequence, we can assess the characteristics inherent in a location through mobile internet access to a certain category of web sites. To put it another way, mobile internet access pattern plays a proxy role in understanding a location. Therefore, in this thesis we propose the mobile architecture combined with the Global Positioning System (GPS) to gather the data to be used in our later mining process, which discovers the information on location-specific access patterns in mobile environment. And using the proposed architecture, we gather the data on location-specific access pattern. Because of the practical difficulties, we simulate the data collection process and generate the data with the rational assumptions to properly depict the real world.

The following sections will give the explanations on system architecture and its simulation process. In section 2, the general mobile system architecture of MSS-MH-GPS will be explained in detail. Next, in section 3, the simulation process for gathering data of web access pattern in mobile environment and its interpretation will be given. Finally, in section 4, we summarize our studies and issue their related future works.

2. Mobile Computing System Architecture with GPS

In this section, we present the mobile computing system model used in our discussion. Figure 2.1 displays the architecture of the general mobile environment. It consists of wired networks of fixed hosts and low-bandwidth wireless cells, each of which consists of a fixed host called Mobile Support Station (MSS) and Mobile Hosts (MHs)[1,2].

Mobile host (MH) can unrestrictedly move either within a cell which

implies a radio coverage area of MSS or between two cells while retaining its network connection. The hosts, other than mobile ones, are steadily connected with a wired network and some of them, called mobile support stations (MSSs), have a wireless interface to enable the communication with MHs. Next, in our architecture, a certain MSS takes the responsibility of supporting MHs in its own geographic area bounded by effective electro-wave arrival distance. The bounded area covered by a MSS is called a cell. Definitely, communication distance between MHs and the MSS concerned is constrained because the MSS supports MHs only within its own cell. In addition, when a MH leaves the cell supervised by a MSS, a hand-off protocol[5] is used to transfer the authority over the MH to the other MSS whose cell the MH moved into. The hand-off during the inter-cell movement is done through the establishment of a new communication link between the MH and its newly related MSS.

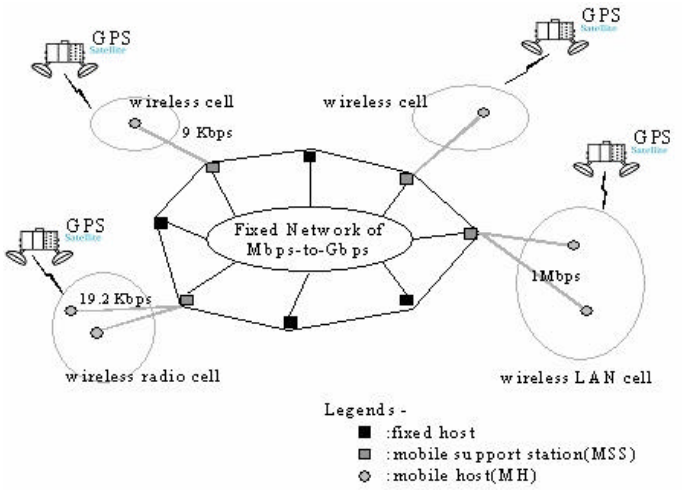


Figure 2.1 Mobile Computing System Architecture with GPS

In traditional mobile environment, roughly speaking, the location of a MH is assumed to belong to the area covered by its connected MSS. Therefore, we cannot identify the exact location of MH at a certain time. To overcome this defect in the traditional environment, i.e., to point out exactly where the MH is within a cell, GPS[5] in our model is introduced to increase the positioning accuracy. In addition, other than the traditional mobile environment, we assume that one MH communicates with only one supervising MSS . This means that we don't consider the situation that the areas covered by MSS overlapped with one another. As a result, in our model, the location data is transferred from GPS to a MH and the web

access is done through the local web browser equipped in the MH. In following sections, more detailed explanations on the roles of each component in our model will be given.

2.1 Roles of Mobile Host (MH)

Technological advances in recent years have made MHs available, smaller in size, larger in capacity and higher in speed. This has enabled the users to access the web in motion, which was impossible with previous terminal-like mobile devices. Moreover, GPS attached to mobile devices, a.k.a., MH can locate the user more accurately than MSS-based locating services. As a consequence, a MH with GPS functionality can report the data on location and web access in a merged form. In other words, MH with GPS provides MSS with the data on when, where and to which URL the user accessed.

2.2 Roles of Mobile Support Station (MSS)

MSS is a network facility which supervises the pre-defined cell with some area and receives the data from MHs in its own cell. After receiving a message from MH, MSS first analyzes the received URL data and finds the category of the URL. After that, MSS produces a new message that tells current MSS, the user of MH, the category of URL, when a message is received and the data items from the MH. This message is stored in a database system and used for later mining process. In this model, the database management system (DBMS) might be a centralized or distributed DBMS which consists of the servers located in a fixed network in mobile computing environment. Furthermore, the inner structure of DBMS is transparent. Therefore, each MSS or server in a fixed network requests the query and receives the query results through the transparent DBMS interface. As a result, we can consult the database freely with ad hoc purposes.

3. Simulation for Data Collection

In our simulation model for data collection, we primarily focus on building the model to properly depict the real world where mobile hosts such as cellular phone and PDA move freely accessing the web without any geographical constraints. To do so, we introduce many assumptions with the rationale based upon statistics.

3.1 Simulation Model

Let us assume that there is a squared region with the area of 2000 \ 2000 where the equipments of MSSs are prepared for mobile communication. In this area, each MSS (Mobile Support Station) covers the squared area of 200 \ 200. Therefore, there exist one hundred MSSs (3.1). In this situation, mobile hosts move continually accessing the web. The number of mobile hosts, k can be the value between zero and positive infinite value (3.2). The location of mobile host at time t is denoted by 2 \ 1 vector with x position and y position values and time index t (3.3).

$$n(MSS) = number\ of\ MSS's = 10 \text{ X } 10 \quad (3.1)$$

$$n(MH) = number\ of\ MH's = k \ where\ 0 \leq k < \infty \quad (3.2)$$

$$Position\ of\ MH_i\ at\ time\ t = P_{i,j} = \begin{bmatrix} x_{t,i} \\ y_{t,i} \end{bmatrix} \ where\ -\infty < x_{t,i} < \infty \text{ and } -\infty < y_{t,i} < \infty \quad (3.3)$$

The above assumption is derived from the process diagram in our model as shown in figure 3.1. In our model, five processes of *DB Server*, *MSS*, *MH*, *User* and *GPS* play critical roles. *GPS* monitors the movement and location of *MH*. User process monitors the user URL requests. *MH* sends the messages *URL*, *Location and MHid* to *MSS* concerned whenever URL requests are done. *MSS* sends the integrated data into database server in distributed database environment.

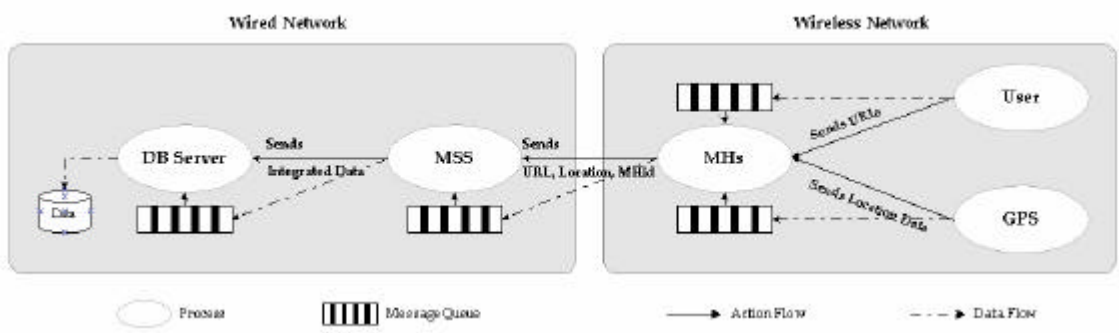


Figure 3.1 Process Diagram for Mobile Communication

User moves and access the web with some time interval. This time interval follows exponential distribution[3]. Statistics has it that the time interval among events follows exponential distribution. Therefore, *Time Between URL Requests of MH*$_i$ and *Time Between Movements of MHi* follows

exponential distribution identically and independently(3.4, 3.5).

(Time Between URL Requests of MH_i) =

$$(URL\ Request_{t+1,i} - \text{URL Request}_{t,i}) \sim \text{i.i.d}\ \exp(\beta_u) \quad (3.4)$$

(Time Between Movements of MH_i) =

$$(\text{Movement}_{t+1,i} - \text{Movement}_{t,i}) \sim \text{i.i.d.}\ \exp(\beta_m) \quad (3.5)$$

The size of movement from one point to another, a.k.a, distance follows log normal distribution because the chance that a man moves to the place near you is much higher. This is not supported by any empirical studies but is drawn from the intuition based on the shape of log normal probability density function. The distance of movement in MH_i follows log normal distribution as shown in (3.6) and figure 3.2.

$$d_i(P_{t+1,i}, P_{t,i}) = \sqrt{(x_{t+1,i} - x_{t,i})^2 + (y_{t+1,i} - y_{t,i})^2} \sim \text{i.i.d.}\ LN(\mu_d, \mu_d^2) \quad (3.6)$$

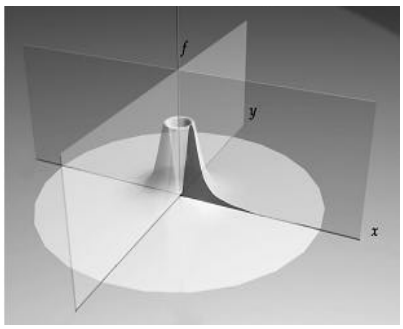


Figure 3.2 Log Normal Distribution of Distance

Direction of movement in MH_i follows uniform distribution because the movement direction is selected randomly with equal chance. Also, *initial point of* MH_i is selected randomly with equal chance in hypothetical area. In addition, *Selected URL* is chosen randomly with equal chance from given URL set. This means that we don't know the personal preference in terms of *URL Request* because we don't have any prior knowledge. It results in the assumption of the *Selected URL* which is randomly selected with equal chance. In other words, *Direction of Movement in* MH_i, *Initial Point of* MH_i and *Requested URL* follows uniform distribution(3.7, 3.8, 3.9) [3]

(Direction of Movement in MH_i at time t)
= $Direction_{t,i}$ ~ i.i.d. $U(P_d)$ where P_d means the probability. (3.7)

(Initial Point of MH_i) = Point of MH_i at time 0

$$P_{0,i}=\begin{bmatrix} x_{0,i} \\ y_{0,i} \end{bmatrix} \sim i.i.d. \begin{bmatrix} U(p_x) \\ U(p_y) \end{bmatrix} where \begin{pmatrix} p_x=\frac{1}{width\ of\ area}, & p_y=\frac{1}{height\ of\ area} \\ -1000 \le x_{0,i} \le 1000, & -1000 \le y_{0,i} \le 1000 \end{pmatrix} \quad (3.8)$$

(Requested URL of MHi at time t) = URLt,i ~ i.i.d. $U(p_{ru})$

where $p_{ru}=\frac{1}{sizeof\ URL\ set}$ (3.9)

The above model is based on statistical probability. Although the model may not fit exactly into our daily lives, it is built upon the rational assumptions. We can obtain as many data as possible varying the parameters of assumed distributions and the exogenous variables such as the area size. The following section will gather data through the experiments and show the result definitely via the visualized dimensional scatter plot.

3.2 Experiments and Results

This section gives the explanations on the environment of the simulation experiments and from the data gathered during the simulation, simple analysis will be done briefly.

3.2.1 Environment for Experiments

The mobile hosts, MHs continually move accessing the web information or services in mobile environment. As shown in figure 3.3, a MH changes the location from point 1 to point 8. Sometimes, it accesses the web and even moves out the region. Our focus has been given to the location in the region where the web accesses happen. The location is surrounded by the circle which implies the related data is stored in database.

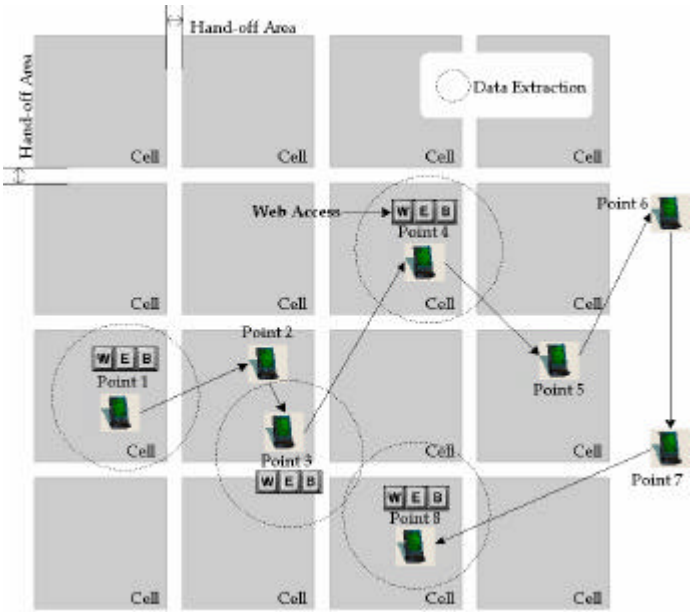


Figure 3.3 Hypothetical Path for a Mobile Host in Mobile Environment

Table 3.1 Simulation Parameters

Name	Value	Description
Area_Size	2000 X 2000	The size of area concerned.(Unit Scale = 1)
ASCBM	200 X 200	Area size covered by MSS (Unit Scale = 1)
n(MH)	1000	Number of mobile hosts
n(URL)	14 X 10	Number of URL list
n(URL_Cat)	14	Number of URL category
Time Span	24 X 60	Time span when data generation is done. (Unit Scale = 1 min.)
b_u	10	Average time between two URL requests. (Unit Scale = 1 min.)
b_m	1	Average time taken during the movement. (Unit Scale = 1 min.)
m_d	$e^{\mu+\sigma^2/2}$	Average distance during the movement. (Unit Scale = 1) (m = ln10, s-square = ln4)
s_d	$e^{2\mu+\sigma^2}(e^{\sigma^2}-1)$	Measure for the spread of Log Normal distribution. (m = ln10, s-square = ln4)
p_d	1/4	The probability of the movement direction in a uniform distribution.
p_x	1/2000	The probability of the initial position in terms of x-coordinate in a uniform distribution.
p_y	1/2000	The probability of the initial position in terms of y-coordinate in a uniform distribution.
p_{ru}	1/140	The probability of the Requested URL in a uniform distribution.

Following this concept, our simulation experiments are done. The parameters in our simulation experiments is classified into two categories: Statistical Parameter and Exgogenous Deterministic Variable. The former is used for the phenomenon on a statistical basis and the latter for the simulation environment. The items of Area_Size, ASCBM, n(MH), n(URL), n(URL_Cat) and Time Span in table 3.1 represent the simulation environment and the rest are for statistical distributions.

3.2.2 Results and Interpretations

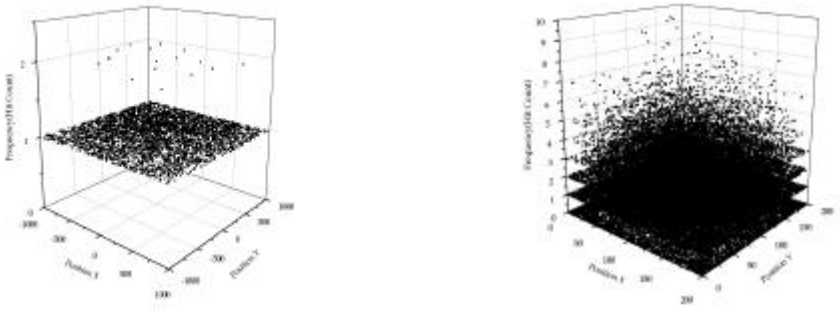
From our simulation, the following data such as *Time*, *MHid*, *MSSid*, *URL*, *Category*, *XPos* and *YPos* have been gathered in the database (table 3.2). *Time* represents when the web access is done, *MHid* denotes the specific MH, *MSSid* specifies the specific MSS through which mobile communication is done, *URL* means the access URL, *Category* denotes the category of accessed URL and the *Position in terms of X and Y* locates where the access is done. In addition, the number of data records gathered in the simulation is equal to 57,225.

Table 3.2 Data Instance of Database in Mobile Environment

Time	MHid	MSSid	URL	Category	Xpos	Ypos
00:00	14	0	www.biblopile.com	1	426	-277
00:00	42	0	www.columbiacountyrealtors.com	11	389	-149
00:00	45	0	www.healthcarey2kguid.com	7	127	249
00:00	50	0	www.angelfire.com	14	22	-756
00:00	57	0	www.hsr.org	7	-424	-411
00:00	95	0	www.cete.org	4	79	-789
00:00	113	0	www.enciclopedia-catalana.com	10	253	596
00:00	119	0	www.nccte.com	4	978	642

By processing the data in table 3.2, we can show the data in a visualized way. The following figure 3.4 shows the 3D scatter plot for location-specific access frequency when the category URL is 1. As told before, URL is classified into 14 categories: *Arts & Humanities*, *Business & Economy*, *Computer & Internet*, *Education*, *Entertainment*, *Government*, *Health*, *News & Media*, *Recreation & Sports*, *Reference*, *Regional*, *Science*, *Social Science and Society & Culture*. For our later analysis, we restrict the category of

URL to *Arts & Humanities* (Category=1). The axis of frequency in panel (a) and (b) of figure 3.4 denotes the web access count under unit scale of one and unit scale of ten respectively. In terms of the scale, zero in scaled-up panel (b) is equal to -1000 in panel (a).

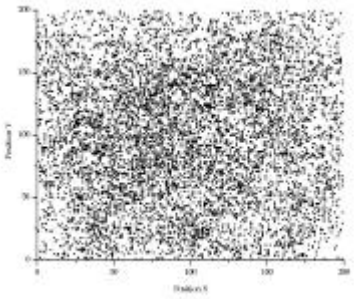


(a) 3D Scatter Plot(Unit Scale=1)

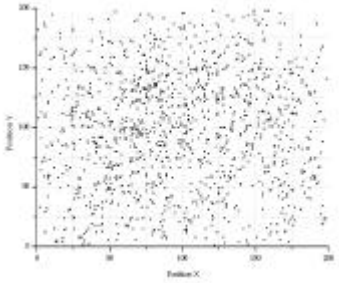
(b) 3D Scatter Plot(Unit Scale=10)

Figure 3.4 3D Scatter Plot for Location Specific Access Frequency

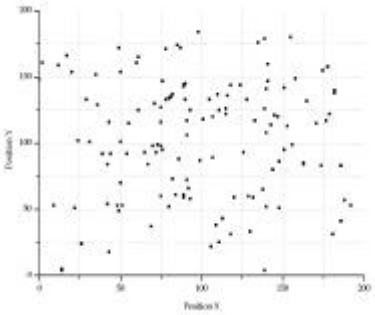
The following figure 3.5 shows the XY projection in terms of access count under the unit scale of 10. The projection is done with the access count level of 3, 5, 7 and 9. As the access count level increases, we can recognize the region which is strongly related to the mobile web access of the URL which belongs to category 1. Although the level of web access count acceptable to prove the relationship between the location and web access cannot be set up determinately, the relationship is prevailing in our analysis. The acceptable level of access count can be set up out of the model. This process of setting up the value can be iterative and the analysis must be done back and forth. Therefore, when the acceptable level of access count is determined during the implementation and information discovery, we can proceed our information discovery process to find out the cluster of the specific locations characterized the mobile web access. Although we characterized the data through simple 2D or 3D scatter plot, we can get the clue to imply the relationship between a location and the mobile web access pattern.



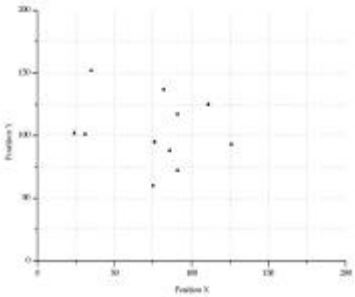
(a) Scatter Plot(Count >= 3)



(b) Scatter Plot(Count >= 5)



(c) Scatter Plot(Count >= 7)



(d) Scatter Plot(Count >= 9)

Figure 3.5 Cross-section Scatter Plot in Terms of Access Count

4. Conclusions and Future Work

In this paper, we propose the data preparation method to find location-specific web access patterns in mobile environment. Previous web usage mining, which deals with the data only from one single server, has defects that the mining cannot extract the information on the web access in mobile environment. This means that the existing web mining does not focus on the origin from which the web access is done in mobile environment. Therefore, to get the information on location, we adopt the MSS-MH-GPS architecture which is composed of MSS, MH and GPS. Using the propose architecture, we specified the needed data items and their formats for discovering the information on location-specific access pattern in mobile environment. Moreover, we prepared the data through the simulated process

which depicts how to use the mobile devices in real world. In addition, we found out the fact that there might be the relationship between location and web access pattern by looking over the data through the visualized 2D or 3D scatter plots.

As a future work, we primarily focus on finding the clusters of the same kinds of web access pattern in terms of geographical location. The clustered area in a category of URLs might be useful especially in on-line or off-line marketing campaign. Our interests and efforts will be given in extracting the information of the clusters.

REFERENCES

(1) D. Barbara, Mobile Computing and Databases-A Survey, *IEEE Transactions On Knowledge and Data Engineering, Vol. 11, No 1*, Jan/Feb 1999.

(2) J. Ioannidis , D. Duchamp and G. Q. Maguire, Ip-Based protocols for mobile internetworking, *Proc. of ACM Symposium on Communication, Architectures and Protocols*, 1991.

(3) A. M. Law and W. D. Kelton, *Simulation Modeling & Analysis*, 2nd Edition, McGraw Hill[LK91]

(4) J. Srivastava, R. Cooley, M. Deshpande, P. Tan, Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, *SIGKDD Explorations*, ACM SIGKDD, Jan. 2000

(5) M. Stemm and R.H. Katz, Vertical handoffs in wireless overlay networks, *ACM Mobile Networking(MONET), Special Issue on Mobile Networking in the Internet*, Fall. 1997.