2003－GN－48 (6)
2003／5／23

Know-when knowledge discovery: an empirical methodology to identify the lost users in mobile Internet services

Toshihiko Yamakami†

The mobile Internet is one of the most promising application domains in the computer communications. The mobile handsets close the gap between end users and the computer communications using the 24-hour 365-day availability. The emerging success in Japan has demonstrated the potential capabilities for the next generation Internet platform. The close relationship to the end user reveals a new research area of the user behavior analysis. Using the user transaction logs for over a year, the author tries to build up a new knowledge domain: know-when knowledge. A methodology to capture the lost users is important to manage the subscription-based mobile Internet services. Using the know-when knowledge acquisition, the user performs the long-term analysis of user behavior patterns and prediction of the lost user probability.

1. Introduction

The mobile Internet is rapidly penetrating the every day life. Especially, the micro-browsers embedded in the mobile handsets have become the every-day gadget since the launch in 1999. After three years, we witness more than 59 million users who access the Internet services from the micro browsers in mobile handsets at the end of 2002. In this paper, the author explores a methodology to capture the user behavior characteristics in the Internet service use using micro browsers in the mobile handsets. In this sense, the term mobile Internet denotes the Internet use using micro browsers in the mobile handsets in this paper. The motivation of the research is based on the assumption that the mobile Internet access with the micro browsers will outnumber the wired web access in the near future. With the 24-hour and 365-day availability, the mobile Internet user behavior reflects the personal behavior patterns. In addition, the restrictions in the mobile Internet may lead to the different user behavior patterns. The author performed field studies in the commercial mobile Internet services since 2000 in order to identify the mobile Internet user behavior. The rapidly growing the mobile Internet provides a series of challenges. Especially, the lack of the stable analyzing methodology makes the inter-service comparison difficult. In this paper, the author presents the issues in the long-term mobile Internet user behavior observation and proposes a methodology to identify the use patterns.

† ACCESS, 2-8-16 Sarugaku-cho, Chiyoda-ku, Tokyo, JAPAN, e-mail:yam@access.co.jp

2. Purpose of the Research

It is known that the mobile Internet has increased uncertainty about the traffic variations. In the past, the integration of the mobile Internet was underestimated due to the relatively small amount of data size per each transaction. The data size is small because the display size is limited on the mobile handsets. However, some of the mobile Internet has millions of paid users, which gives a significant challenge on the database transaction capabilities. Understanding the user behaviors is important with the following two reasons.

- To provide the base technology for service deployment, transition, and integration
- To provide the base technology for user marketing

2.1 The Past Research

The transaction log analysis dated back to 1980's for usability, traffic and system analysis. In 1990's there were significant study in the wired web analysis. Rosenstein examined the accuracy and pitfalls in web server log[7]. Burgess used the transaction log to identify the system anomaly detection to set criteria for statistical state for hosts[2]. In addition, service level logs are analyzed for the emerging e-

commerce usage analysis. Padmanabhan made a popular web site log study for the server side to identify the content dynamics and temporal stability for user access for content management purpose[5]. Jansen made a search engine study to identify the sessions and query sequence analysis[3]. The literature about the success of the mobile Internet is rarely available at the moment. The user interface study illustrates the user interface constraints on the mobile handsets[1], however, it did not provide any cue for the rapid mobile Internet growth mainly in Asia. In the CSCW research, the long-term time factors in the social rhythms were studied in the medical environment[6]. The social approach like[9] is still in the early stage to identify the culture factor in the mobile Internet. The author performed the initial study on the mobile Internet usage pattern[8] with the approach unique identifier tracking analysis to make use of the unique mobile Internet characteristics. However, this research was done on relatively short-term log (6 months). The longer-term research case studies[10] illustrated the difficulties in the long-term field studies in the commercial services. In order to understand the transaction log, the user model is necessary. In e-commerce, the Customer Behavior Model Graph approach was proposed by Menasce[4] for workload characterization. This study inspired the author to coin the aging model to make clustering criteria.

2.2 Challenges

The dynamism in the mobile Internet makes the stable analysis very difficult. It comes from the two aspects of the dynamism in the mobile Internet:

- the total usage is dominated by the heavy users, and
- the volatility of the users makes the stable comparison difficult

The mobile e-commerce is expanding, however, the systematic research on the mobile user behavior is still in the early stage. Especially, the biggest challenge is how the volatile user behavior can be dynamically captured on the web. It is common to witness the "easy come and easy go" users on the mobile Internet services. A systematic methodology to capture the user dynamism is necessary for the mobile e-commerce. It is important to capture the lost users in a timely manner because it is difficult to recover the lost users in the mobile e-commerce. Especially, the effective reliable clustering of the users is crucial for the exploration in this emerging domain.

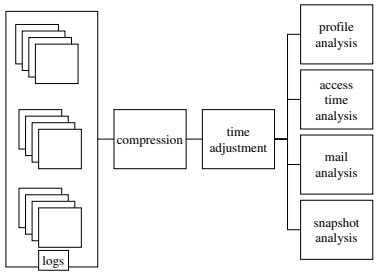

Fig. 1 A command log analysis tool configuration

3. Unique Identifier Analysis

One of the key factors in the mobile Internet is the user track ability using the unique user identifier from the mobile carriers. The user identifier is usually 16 or more alphanumeric character long, e.g. " 310SzyZjaaerYlb2 ". It is distinguished that the mobile Internet is aimed at the paid services. It is unique in the mobile Internet service domain. **Fig. 1** illustrates the command log analysis tool configuration. The time adjustment is used to coordinate the differences on the multi-server site configuration. It runs on PHP 4.1.2.

This paper focuses on the traffic analysis on the mobile Internet. The questionnaires and interviews to identify the causes and user perceptions will be for further studies. The command transaction logs for the mobile Internet service is used to make usage pattern analysis. The transaction logs are obtained from the subscription-based paid services in the mobile Internet in Japan. It is common that the subscription fee range is 1 to 3 US dollars per month. In addition, many services consist of a paid subscription-based service part and a free-of-charge service part.

4. Know-when knowledge discovery

One of the heavily discussed topics in the web knowledge discovery is the performance of the discovery systems. The web logs are great sources for data mining today. It gives a chal-

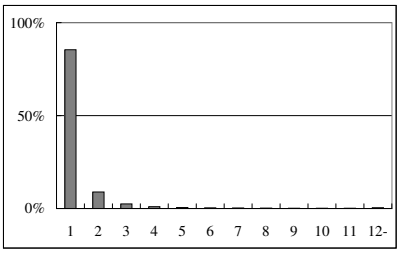

Fig. 2 The number of user visit months during one and a half year commercial service field study

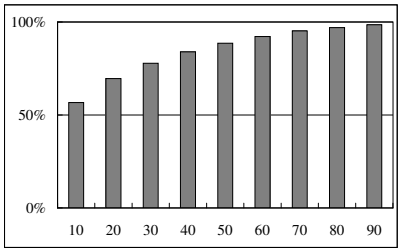

Fig. 3 Top n% users' ratio in the total access in an accumulated manner

lenge for how the discovery system can perform the discovery function in an efficient manner. However, the author focuses on the different side of the time factor in the knowledge discovery. It is the know-when knowledge discovery. In the real world, the human user or a group of them change the behavior in a long-term. It is a challenge for the computer technology to detect the changes over a long span of time. For example, a user gradually loses the interest to some sites, it is important to detect such a behavior change. However, this type of discovery of the time-related aspects of the user behavior is still a challenge. The unique characteristics of the user identifier uniqueness of the mobile Internet should be effectively utilized.

5. Aging Analysis

5.1 Need for Segmentation Analysis

The author performed field studies in the mobile Internet user transaction logs since 2000. **Fig. 2** illustrates the typical visit month statistics in one of the field studies during January 2001 to July 2002. The log comes from the subscription-based transaction logs for business use news service. It should be noted that the 85% of the user showed only one of the months during the observation period. More than 90% users showed only less than 3 months.

Fig. 3 illustrates the ratio of the top n% users' ratio in the total access. It comes from the same service as depicted in Fig. 2 in January 2001. The 10% users occupy the 60% of the total mobile web access. 20% users occupied the more than 70% of the total access.

One of the challenging issues in the mobile Internet is the dynamism in the usage. It varies from time to time and easy to change. A large number of users come and cannot find the usefulness on the top page of the mobile web service and are gone. On the contrary, some of the addicted users account for the large portion of the database access for subscription check and for latest content. The network workload and database workload heavily depends on the ratio of the heavy-use and the super light use users.

5.2 Aging Analysis

In the preliminary study, it is shown that the significant number of users just visit the mobile Internet sites just a small number of times. The author calls it as " easy-come and easy-go " pattern. In the mobile Internet, the information for the sites is available in a very limited manner. They click the hyperlink without having sufficient knowledge about the next page. In the PC Internet, they can check the URL on the browser without doing any special operation, however, the space to show the next link URI is just not available on the mobile handsets. During the past research, the author found that the use pattern is heavily affected by the user experience. The factor to indicate user behavior pattern dynamics caused by the length of the user experience is called " aging " here. Especially, the number of months since the user started to use the service is measured for comparison. **Fig. 4** illustrates an aging transition model for the aging observation.

6. A Case Study

6.1 Setting

A case study is performed to identify the user dynamism and user segmentation. A commercial service is used to record user command

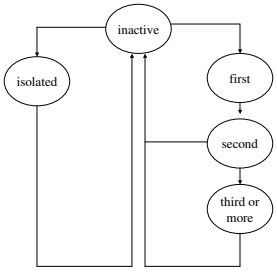

Fig. 4 An Aging Transition Model

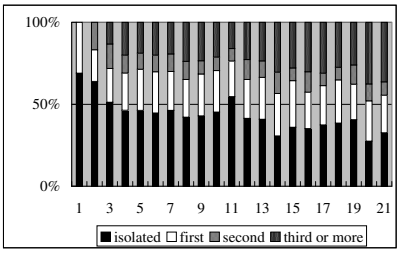

Fig. 5 The user ratio during the long-term monthly analysis in a service in the carrier A

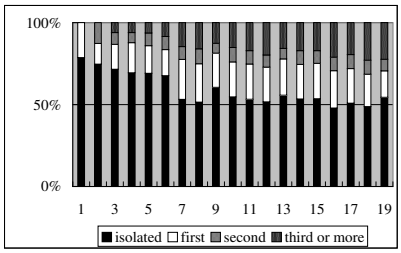

Fig. 6 The user ratio during the long-term monthly analysis in the same service in the carrier B

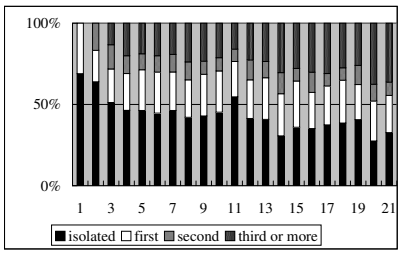

Fig. 7 The user ratio during the long-term monthly analysis in the same service in the carrier C

logs. The service log is from September 2000 to June 2002 is analyzed. The service is an official site for the three different carriers. The service launch was from the July 2000 to January 2001 depending on the carrier. It is intended for the business people information service, partly charged and some of the content is provided for free. The service content is slightly different from carrier to carrier, however, the major information content is same and updated on a weekday daily base. The main content is converted into each carrier specific content format in the respective markup language.

6.2 Observation

Depending on the launch date, the length of the long-term observation differs from service to service. **Fig. 5** depicts the aging based user segments on the target service in the carrier A. **Fig. 6** depicts the same segment data for the service in the carrier B. Then, **Fig. 7** illustrates the same segment transition from the log data on the same service in the carrier C.

The ratio of isolated users, called easy-come easy go users, differs from service to service depending on the carrier. It could come from each user's user segment or differences of marketing. The cause analysis of the differences is for further studies. During the October 2001 to December 2002, the three carriers launched the IMT-2000 services, which may impact the user expectations and user interface structures. The impact on these infrastructure changes is for further studies.

6.3 Simulation

Based on the clustering data on the aging analysis, the user behavior from the first month to the next month is evaluated using the threshold value. The threshold value is set to the middle value between the average isolated users and the average first users. It is assumed that the difference of the average command access numbers between these two groups can be a good assessment of the clustering measure. It is a preliminary study to make use of the clustering data for the user behavior prediction. The simulation part is also implemented in PHP 4.1.2. The data from 2000 to July 2001 is used for the

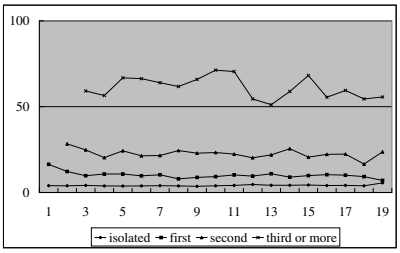

Fig. 8 Average number of access per month based on the aging analysis from January 2001 to July 2002.

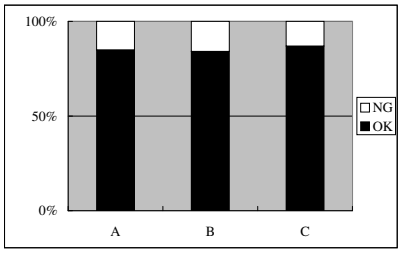

Fig. 9 Preliminary Result to predict the user presence in the next month from the first month in the carrier A, B, and C.

simulation.

7. Experience

From the aging analysis, each cluster's average web access count per month is depicted in **Fig. 8** for a carrier B service from January 2001 to July 2002.

With the long-term history data, this gives a good estimation how a threshold value for super heavy user and super light user identification. The technique is general and applicable to a wide range of applications with user identifiers.

The threshold value predicts the user presence in the next month in the accuracy ranging from 84% to 87 % in the three cases. The simulation is performed based on the first month presence and the second month presence data. The accuracy about the prediction on the simulation data is shown in **Fig. 9**. The accuracy is defined by the ratio of the accurate next month presence based on the threshold value against the first month user access numbers. The accuracy needed for the mobile e-commerce marketing is still for further studies.

The dynamism in the mobile Internet comes from the three factors:

- Market Emergence Speed,
- User Behavior Uncertainty Due to the Limited Marketing Information, and
- Market Imbalance.

The fast penetration of the mobile Internet put the use scenes dynamic and transitive. In three years since the launch in 1999, the mobile Internet using the handset micro browsers penetrated 40The user behavior is uncertain because the user does not have sufficient knowledge about sites. The information available on the sites is limited due to the limitation of the total display size. The information on the mobile Internet is fragmented due to this limitation. Many links are terse for conciseness. The mobile Internet has more volatile traffic structure compared with the PC Internet. Some of the cites have a drastic traffic increase pattern. The initial traffic in such a site is unpredictable. Especially, the information about the newly created cites are transferred by the very limited marketing channel, like the carrier top notice pages. It causes intensive heavy access patterns on the site. To cope with the dynamism, there are two issues:

- How to segment the heavy traffic, and
- How to identify the causes of the dynamism.

This research needs the following criteria:

- General applicability,
- Sensitivity to the usage pattern dynamism,
- Ability to segment the usage patterns, and
- Stability on the long-term analysis

About the general applicability, the usage log analysis and aging analysis are applicable to a wide range of applications using the simple time-stamped command logs. From this point, this methodology is promising for a base technique for mobile Internet use analysis. In addition, this method can be applicable for a wide variety of span analysis, like days, weeks, months, 3 months, years, and so on. A careful design of the log format and log collection will provide a good technique for the long-term use analysis. About the sensitivity to the usage pattern dynamism, it is usable to identify the differences among these three services. The sensitivity is not verified for the cause analysis. Therefore, the effective of the sensitivity

is for further studies. About the ability to segment the usage patterns, the number of the segments is the important. The ability to identify the best segment number is not covered by this research. The basic assumption is that there are two or three different user segments. The most obvious example of the two-segment pattern is heavy users and light users. An example of three-segment pattern is heavy, common and light users. These hypotheses are not verified in this research. It includes the in-depth cause analysis and it is for further studies. About the stability over the long-term analysis, it gives a stable result over time. It is a macro analysis. Therefore, it needs supplementary micro-analysis to identify the real dynamics in the usage transitions. Considering the wide variety between the heavy users and easily vanishing one-time users, the aging analysis is simple and generally applicable that is an advantage of this method.

8. Conclusions

With advances in computer communication technologies, we are approaching the ubiquitous computing environment where anytime and anyplace computing services will be available. Such an advance will lead to the demands of identifying 24-hour user behavior. With the emerging experience with the mobile Internet, it is promising to get the user behavior study using the long-term accumulation of user behavior data. The author gives the first trial of mobile Internet usage pattern analysis in order to estimate the future user behavior using the available log data.. In this research, the unique identifier tracking analysis is extended for a general framework of aging analysis over a span of time. A general aging analysis with user identifiers in the mobile Internet is used to predict the next month user presence. From the preliminary study on the cluster based on the aging analysis, a lost-user simulation is performed to explore the efficient detection of the user behavior. The result is approximately 85 % accuracy for the first month user data. The dynamic clustering is a hot topic in the mobile e-commerce because it is crucial for business to capture the user behavior and make effective marketing. The quantitative analysis on the long-term mobile Internet usage transitions is useful for the effective mobile e-commerce.

References

1) G. Buchanan, S. Farrant, M. Jones, H. Thimbleby, G. Marsden, and M. Pazzani, “ Improving mobile internet usability ”, The tenth international World Wide Web conference on World Wide Web, ACM Press, May 2001.
2) M. Burgess, H. Haugerud, S. Straumsnes, and T. Reitan, “ Measuring system normality ”, ACM Transactions on Computer Systems, Vol. 20 No. 2, pp. 125-160, May 2002.
3) B. Jansen, A. Spink, J. Bateman, and T. Saracevic, “ Real life information retrieval: a study of user queries on the Web “ , ACM SIGIR Forum , Vol. 32 No. 1, pp. 5-17, Apr. 1998.
4) D. Menasce, V. Almeida, R. Fonseca, and M. Mendes, “ A methodology for workload characterization of E-commerce sites ”, Proceedings of the first ACM conference on Electronic commerce, pp. 119-128, Nov. 1999.
5) V. Padmanabhan and L. Qiu, “ The content and access dynamics of a busy Web site: findings and implications ”, ACM SIGCOMM Computer Communication Review, Proceedings of the conference on Applications, Technologies, Architectures, and Protocols for Computer Communication, Vol. 30 No. 4, pp. 111-123, Aug. 2000.
6) M. Reddy and P.l Dourish, “ A social sense of time: A finger on the pulse: temporal rhythms and information seeking in medical work ”, Proceedings of the 2002 ACM conference on Computer supported cooperative work , pp. 344 - 353, Nov. 2002.
7) M. Rosenstein,“ What is actually taking place on web sites: e-commerce lessons from web server logs ”, Proceedings of the 2nd ACM conference on Electronic commerce, pp. 38-43, Oct. 2000.
8) T. Yamakami, “ Unique Identifier Tracking Analysis: A Methodology To Capture Wireless Internet User Behavior ”, IEEE ICOIN-15, Beppu, Japan, Feb. 2001.
9) T. Yamakami, “ An AIMS model-based approach: Toward Understanding Wireless Minds ”, IPSJ DiCoMo2002, pp.277-280, Toi, Japan, July 2002.
10) T. Yamakami,“ A Case Study on Mobile Internet Use Analysis：Implications from Long-term Mobile Internet Observation ”, IPSJ Technical Reports IPSJ-GN-45-20, pp. 113-118, Oct. 2002.