2009 Eighth International Symposium on Natural Language Processing

# Client-side Mobile User Profile for Content Management Using Data Mining Techniques

W. Paireekreng and K.W. Wong

*Abstract*—**Mobile device can be used as a medium to send and receive the mobile internet content. However, there are several limitations using mobile internet. Content personalisation has been viewed as an important area when using mobile internet. In order for personalisation to be successful, understanding the user is important. In this paper, we explore the implementation of the user profile at client-side, which may be used whenever user connect to the mobile content provider. The client-side user profile can help to free the provider in performing analysis by using data mining technique at the mobile device. This research investigates the conceptual idea of using clustering and classification of user profile at the client-site mobile. In this paper, we applied K-means and compared several other classification algorithms like TwoStep, Kohenen and Anomaly to determine the boundaries of the important factors using information ranking separation.**

## I. Introduction

THE mobile phone has becoming an important device for providing information anytime anywhere. However, due to the limitation of its hardware such as small display screen and input capacities, it is not as easy as using a personal computer. Using information services and mobile internet on mobile devices has several difficulties such as information overload, small screen and input capacities limitation. There is a huge amount of information available online today. It is sometimes common that the information accessed by the desktop PC users could also be accessed by the mobile internet users. One possible way to make the accessing of online information easier is personalisation [13]. There have been several attempts to help mobile users retrieve information and services efficiently. They are content visualisation or content personalisation [16]. This paper focuses on content personalisation only. The primary process to do content personalisation is to know who the users are and how to customise the contents to be delivered to the users at the right time. The user profile is one of the important factors for successful content personalisation. Although some providers need users to apply for the membership, fill in user information form, or login for information services, it is inconvenience for users to submit their information to the providers due to many reasons. Reasons such as privacy issues, time required filling in all the requested information or the hardware's input capacities could cause the inconvenience.

Manuscript received July 15, 2009.

W. Paireekreng is with the School of Information Technology, Murdoch University, Australia, WA, 6150 (e-mail: w.paireekreng@murdoch.edu.au)

K.W. Wong is with the School of Information Technology, Murdoch University, Australia, WA, 6150 (e-mail: k.wong@murdoch.edu.au)

Hence, storing a user profile at the client-side can be viewed as prefer solution. Users can update their profile as they desire. In addition, the profile can be updated automatically using intelligent data mining techniques such as clustering and classification. This can provide appropriate process between the client-side and the server-side to perform the content personalisation.

This paper proposes the client-side user profile with a framework to perform the content personalisation with data mining techniques to cluster and classify the user profile type. This can help content providers facilitate client-side processing on personalisation to increase customer's satisfaction. It also allows content provider to provide appropriate information at the right time at the right place. This will ensure that useful information be delivered to satisfy the users based on demographic factors using simplified techniques computed on the mobile devices.

## II. Literature Review

In order to provide content personalisation or content recommendation, the system needs user profile information. If users can provide more useful information, the chance of getting the content they desire should be higher. This may also lead to better performance of the personalisation system.

### *A. Personalisation*

Personalisation was defined by Ivar Jorstad et.al. [6] as "mechanisms exist to allow a user *U* to adapt, or produce, a service *A* to fit user U's particular needs, and that after such personalisation, all subsequent service rendering by service *A* towards user *U* is changed accordingly". They also presented the concept of personalisation in terms of an customise something to satisfy a person's need.

According to [5], it can be observed that more than 66% of the users from the survey want to personalise their mobile phone content. The research also presented that context-awareness and the situation users are currently in, seem to be important for user personalisation. It is evidence that personalisation can be used as a motivation for many tasks and services. Additionally, it might be better if the user can be provided with an alternative forethought [8].

Consequently, the success of personalisation depends on a few important factors. They are user profile classification which is the separation of user into a small group for serving customised services, and context customisation which is to

978-1-4244-4139-6/09/$25.00 ©2009 IEEE

determine the different context in terms of location and time. It is therefore important to classify the user's group and internet usage behaviour based on their current context information in order to provide the specific content that the user requires. This can be done by content filtering.

The primary data for classifying the group of users by interest includes demographic variables such as gender, age, occupation, and income. The user profile may come from personal data, user's interest or user's preferences through both implicit and explicit findings. There is some example works related to user profile like the one in [3], which they applied symbolic machine learning to discover user's interests and preferences. The research also proposed the ALKEMY algorithm for the decision-list learner online. This work was implemented on the infotainment TV show recommender application. It can be used to predict the class of an individual based on its current decision list. As can be observed from this research, the user profile is an important aspect of personalisation.

### B. User profile on personalisation systems

*1) Background:* The concept of user profile was introduced in the research of Wagner et.al in 2002[13]. The research proposed framework for advanced personalisation of mobile services using profiling technique with the semantic enrich service. The main concept assumes that the user can belong to a specific group. Thus a generalise usage pattern can be applied for that group. In addition, important standardisation efforts can facilitate online security for identity purposes. The profiling technique is device dependence, which is compatible to mobile terminal such as CC/CP and UAProf. Open mobile network using XML interfaces is a representative tool for connecting with other systems to facilitate this as well.

*2) Storing User Profile:* To show the concept of storing the user's personal profile, as can be observed from 'MyLifeBits' [4], that they stored everything related to user's life in the user's PC. They even stored real-time data collected during interaction including mouse clicks and web pages visited. The research also suggested that mobile devices should also store personal data in the same manner. Besides, it should also store the internet usage behavior. In addition, the data should be organised according to the life style and the time of the activities recorded.

The example real world application using user's profile is 'FLAME2008' [14]. The main concept for this work is to use ontology for creating the relationships between the user profile and context awareness based on location and time, in order to obtain the personalised service on the mobile application. The user profile consists of personal data, user interest, and user preferences. It also stores the information relating to context.

*3) User Modeling Construction:* Kobsa in 2001[7] presented the development of the genetic user modeling systems. It described the characteristics of generic user model which mainly consisted of generality including domain independence. In addition, the user profile should be universal. Although the generic profile is created for one application, it can normally be used generally for other applications as well. The research also suggested that the generic user model can be applied for mobile devices in future.

The construction of the user model was expanded in Mobile User Behavior Modeling [12] which was based on task-oriented model using ontology. This approach implemented generality concept in task organisation such as buying a book, entering to a park and scalability towards several providers. After that, it combined the hierarchical task concept with the domain concept.

As can be seen, if we combine the user modeling by relevant tasks and domains, it can provide better personalisation for the specific new mobile content services. Content providers can serve various types of customer's needs according to their profile. In this case, more customised services can then be launched.

### C. Information Ranking

It is important to separate the user's view towards each content item. If it can be defined by how users rate each item in different aspects, it will make the content personalisation process easier. The use of clustering and classification of the user profile could facilitate this.

In the survey of user content personalisation for the mobile users [9,10]. It can be observed that users assigned the content items in various aspects which are importance, up-to-date and time. The research shows that users may rate different items which reflect his/her preferences. It can be separated by demographic factors such as gender, age or income and mobile usage factors such as the network provider and mobile internet capabilities. The other related work is specific to the educational information system [11]. The research also presented that students pay attention in browsing important information differently identified by their profile, such as enrolled year, GPA or income. Moreover, the importance, and up-to-date information also affect the behavior of browsing each content item on the mobile device.

The first algorithm that mentioned the 'importance' weight is PageRank [2]. It a favored algorithm to analyze the World Wide Web and link structure by rating the links using importance of the page. It is ranked by the score of a summary of page that point to the calculated PageRank page.

As can be observed here that separating the aspects of the presented information tends to help the personalisation process, it will be the same when applied to mobile applications.

### D. Machine Learning in Clustering and Classification problems

It seems that data mining is important for businesses especially in finding the customer's needs. This area normally implements Machine Learning Techniques in order to analyze the data correctly. Machine Learning has an important role towards business data processing especially in data mining or knowledge discovery. Data mining process includes data set selection, preprocessing, data analysis and data interpretation [1]. Data Mining can be used to understand the problem context and provide solutions, techniques such as Classification, Prediction, Association and Detection can be used. Many applications also implemented Rule Induction in Machine Learning for data mining. Wu et.al. [15] have shown that some commonly used algorithms in data mining are k-means, SVM, Apriori, PageRank including Naïve Bayes. They also described k-means as a simple iterative clustering method. As this is a simple algorithm, and due to the fact that mobile devices have limited resources, it could suggest that it is appropriate to be implemented at the client side.

As can be seen, personalisation is very important for mobile internet usage especially for content delivery. However, for most providers, users have to subscribe in order to access all websites or services. It can sometime be difficult to manage their username and password and link them to the user's profile. Therefore, we suggest a framework in this paper to save the user profile within the mobile device instead of using the subscription model.

The first step to do the general personalisation is to the user to a user group. Clustering and classification model can be implemented on the mobile phone to perform this task. There are some limitations on the device's resource such as memory and computation. The algorithms to be used on the mobile device should be simple and efficient. We propose user type classification on favorite items based on ranking information.

## III. Research Methodology

### A. Data source and Pre-processing

The data source used do the experiment was obtained from the published research [10]. This data is the survey results on the preferred content of mobile users in Bangkok. This set of data consists the user's preference of contents such as multimedia, news or information services on mobile internet. This data also includes information to separate user's context information such as time-of-day, importance or up-to-date information.

A sample size of 200 data is used. The data cleaning process is done first before using in the experiment. The data was divided into training and testing sets. The training set is used to establish the prediction model and the testing set is used to test the performance. From the training set, the 2 context attributes which would be used in this experiment were chosen. They are the *Importance* and *Up-to-date* information of the mobile content item. From these, there were 2 sets of data:

- Training data set of importance mobile item
- Training data set of up-to-date mobile item

Finally, before the data were used to establish the model, the demographic factors were chosen. They are gender, age, occupation and income.

### B. Proposed clustering and classification framework

The first step to establish the client-side user profile starts from constructing the clustering model. We are assuming that the prediction model does not have any prior knowledge of the nature of the clusters that exist. After the clustering information has been established, classification model will be constructed to predict the future user's preferences and interests. After that, the user profile type and the influenced factors would be submitted to content provider for the next step of processing to obtain the content personalisation. Figure 1 summarise the process.

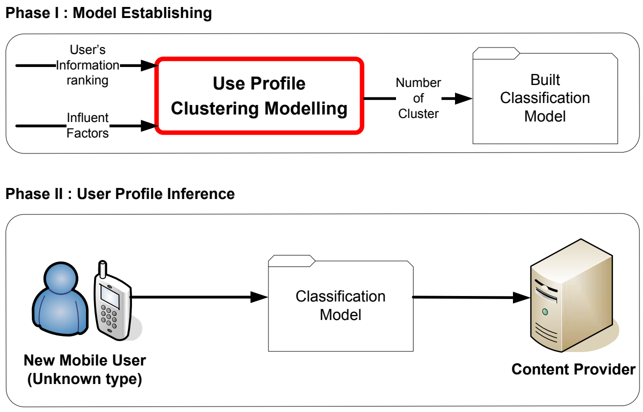


Fig. 1. The client-side user profile processing framework.

## IV. Results

### A. Clustering

The first part of the result is concerned with the clustering model because there is no cluster information related to user profile. Firstly, the importance and up-to-date information data sets with all mobile internet items were used to create the clusters using different algorithm as follows: k-means, TwoStep, Anomaly and Kohenen.

TABLE I
CLUSTERING WITH ALL FACTORS AND ITEMS

| Algorithm and information type | Number of Cluster | Number of small clusters (number < 5% | Iteration |
|---|---|---|---|
| k-means I | 10 | 3 | 6 |
| k-means U | 10 | 6 | 9 |
| TwoStep I | 10 | 0 | - |
| TwoStep U | 10 | 0 | - |

I = Importance of Information, U = Up-to-date of information

The results show that Anomaly and Kohenen may not provide the clusters information appropriately for this data set because the number of clusters is not distributed appropriately. Anomaly separated the cluster into only 2 groups while Kohenen divided into 12 clusters with much gap among the groups. Having said these, optimisation could help to realise the clusters better, but this is out of the scope of this paper. We obtained better results from TwoStep clustering technique but the processing time may take longer. K-means' result is acceptable when performed on the importance data set with the number of iteration as shown in table 1.

The next experiment worked on the selection of the important factors, which were gender, age, occupation and income with the top 7 items ranked from the user's rating of data. The experiment implemented k-means algorithm as it is a simple algorithm which consumes less computational time. This could suggest that it is appropriate to be implemented on the mobile device for user profile classification.

TABLE II
CLUSTERING WITH SEPARATED FACTORS AND 7 SIGNIFICANCE ITEMS

| Factors | Number of Cluster | Number of small clusters (number < 5% | Iteration |
|---|---|---|---|
| Gender – I | 10 | 3 | 14 |
| | 5 | 0 | 8 |
| Age – I | 10 | 1 | 8 |
| | 5 | 0 | 9 |
| Occupation – I | 10 | 0 | 6 |
| | 5 | 0 | 13 |
| Income – I | 10 | 1 | 17 |
| | 5 | 0 | 11 |
| Gender – U | 10 | 3 | 7 |
| | 5 | 0 | 9 |
| Age – U | 10 | 3 | 11 |
| | 5 | 0 | 20 |
| Occupation – U | 10 | 1 | 7 |
| | 5 | 0 | 9 |
| Income – U | 10 | 4 | 16 |
| | 5 | 0 | 7 |

I = Importance of Information, U = Up-to-date of information

It can be observed that if the importance of information was clustered with 7 significance items and demographic factors, it can help to reduce the number of small clusters. The small clusters refer to clusters that consist of small number of data. In our case, it is set at 5%

## B. *Classification*

After the cluster characteristics have been determined, the next step is to build the classification model for future prediction. We use the clustering information as class label for supervised training. There were 4 different types of algorithms implemented for classification, specifically LWL, RepTree, Decision Table and SVMReg. In this paper, the importance and up-to-date of information were also used. In addition, the factors used to consider the performance of the model established are correlation, Mean Absolute Error and Relative Absolute Error. It can be observed from the table 3 that the better performance algorithm is Decision Table followed by RepTree while LWL and SVMReg presented the lower performance on this experimental data set.

TABLE III
CLASSIFICATION WITH ALL FACTORS AND ITEMS

| Factors | Measure | LWL | RepTree | Decision Table | SVMReg |
|---|---|---|---|---|---|
| Importance | CC | 0.5589 | 0.7685 | 0.7751 | 0.6599 |
| | MAE | 2.2857 | 1.4035 | 1.4462 | 1.8023 |
| | RMAE | 2.6099 | 2.0004 | 1.9749 | 2.3918 |
| | RAE | 82.3423% | 50.5622% | 52.1004% | 64.9278% |
| | RRSE | 83.4955% | 63.9983% | 63.1803% | 76.5201% |
| Up-to-date | CC | 0.6599 | 0.7968 | 0.8005 | 0.6276 |
| | MAE | 1.7629 | 1.0906 | 0.1171 | 1.4613 |
| | RMAE | 1.9749 | 1.5641 | 0.1628 | 2.0292 |
| | RAE | 80.4840% | 49.7929% | 58.1753% | 66.7144% |
| | RRSE | 76.2934% | 60.4227% | 59.9396% | 78.3899% |

CC = Correlation Coefficient, MAE = Mean Absolute Error, RMAE = Root Mean Absolute Error, RAE = Relative Absolute Error, RRAE = Root Relative Absolute Error

Secondly, the demographic factors and top 7 significant items show better results overall. Table 4 has shown the correlation coefficient results in terms of the importance of information with different number of classes

TABLE IV
CLASSIFICATION WITH SEPARATED FACTORS AND 7 SIGNIFICANCE ITEMS OF IMPORTANCE OF INFORMATION

| Factors | Number of Classes | LWL | RepTree | Decision Table | SVMReg |
|---|---|---|---|---|---|
| Gender | 10 | 0.7844 | 0.9524 | 0.9246 | 0.7635 |
| | 5 | 0.8510 | 0.9218 | 0.9371 | 0.2033 |
| Age | 10 | 0.6058 | 0.9488 | 0.9487 | 0.1454 |
| | 5 | 0.8743 | 0.9503 | 0.9955 | 0.7102 |
| Occupation | 10 | 0.7629 | 0.9148 | 0.9533 | 0.5936 |
| | 5 | 0.8186 | 0.9436 | 0.9754 | 0.6473 |
| Income | 10 | 0.6177 | 0.8338 | 0.9343 | 0.5609 |
| | 5 | 0.8541 | 0.9489 | 0.9014 | 0.6471 |

However, the up-to-date information seems to be influenced by the number of classes selected. When comparing to 10 classes, the 5 classes results seem to present better correlation coefficient, as shown in figure 2. In addition to the above classification algorithm, results generated from RepTree and Decision Table using 5 classes does not have significance difference in gender, occupation and income factors. The age factor shows more variation in the correlation coefficient. The Decision Table consumed more time to generate the classification results when compared to RepTree.

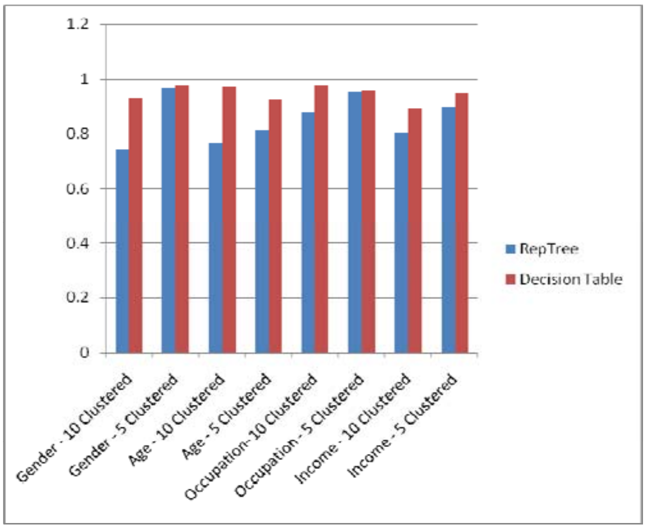

Fig. 2. RepTree and Decision Table were applied on the classification of up-to-date information with demographic factors and different clustering based on correlation coefficient

## V. Discussions And Conclusions

The purpose of this paper is to introduce the concept of constructing of the user profile on the mobile device for content personalisation using data mining techniques. This paper implemented the construction of user profile data from clustering algorithm. The cluster information is then used as class information to build the classification model. It can be observed that there were a lower number of small groups in the results. Although the number of classes used was 5 or 10, there is no significant difference between them. When examine each demographic factor, the occupation factor seems to be provide better separation in the cluster. The Importance of information data set showed better clustering results when compared to the up-to-date information data set. However, gender seems to be important as there are more small clusters compared to other factors. These also implied that gender and preferences should be considered in more detail when considering menu by menu or item by item solution for the purpose of personalisation. For implementing clustering at the mobile devices, K-means algorithm seems to be appropriate from the test results.

For classification, RepTree provide good results, and consumes less time. Thus, it can be used as an algorithm on for classifying user profile on the mobile device. Besides, the RepTree algorithm also worked well in demographic factors.

In conclusion, the paper not only show the algorithms that can be implemented at the mobile devices, but also discover the factors of providing content personalisation.

In future study, user's interest and preference can be incorporated in order to adjust the model to improve the prediction accuracy.

## References


[1] I. Bose and R. K. Mahapatra, "Business Data Mining - A Machine Learning Perspective," Information and Management, vol. 39, pp. 211-225, 2001.

[2] S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine," Computer Networks and ISDN Systems, vol. 30, pp. 107-117, 1998.

[3] J. J. Cole, M. J. Gray, J. W. Lloyd, and K. S. Ng, "Personalisation for User Agents," presented at Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems (AAMAS'05), Utrecht, Netherlands, 2005.

[4] J. Gemmell, G. Bell and R. Lueder, "MyLifeBits: a personal database for everything," Communications of the ACM, vol. 49, pp. 88-95, 2006.

[5] J. Hakkila and C. Chatfield, "Personal Customisation of Mobile Phones - A Case Study," in NordiCHI 2006:Changing Roles. Oslo, Norway, 2006, pp. 409-412.

[6] I. Jorstad, D. V. Thanh, and S. Dustdar, "Personalisation of Future Mobile Services," in 9th International conference on intelligence in service, 2004.

[7] A. Kobsa, "Generic User Modeling Systems," User Modeling and User-Adapted Interaction, vol. 11, pp. 49-63, 2001.

[8] A. Oulasvirta and J. Blom, "Motivations in Personalisation Behaviour," Interacting with Computers, vol. 20, pp. 1-16, 2007.

[9] W. Paireekreng and K. W. Wong, "The Empirical Study of the Factors Relating to Mobile Content Personalization," International Journal of Computer Science and System Analysis, vol. 2, pp. 173-178, 2008.

[10] W. Paireekreng, "Influence Factors of Mobile Content Personalization on Mobile Device User in Bangkok," Research report 2007.

[11] W. Paireekreng, "Influence Factors of Student Mobile Content Personalization on Mobile Device User a Case Study of Students at Dhurakij Pundit University," Research report 2007.

[12] M. Sasajima, Y. Kitamura, T. Naganuma, S. Kurakake, and R. Mizoguchi, "Toward Task Ontology-based Modeling for Mobile Phone Users' Activity," in Proc. of Posters and Demos of the 4th International Semantic Web Conference (ISWC2005), PID-29, Galway, Ireland, Nov. 6-10th, 2005.

[13] M. Wagner, W.-T. Balke, R. Hirschfeld, and W. Kellerer, "A Roadmap to Advanced Personalization of Mobile Services," in ODBASE, CoopIS 2002, Irvine,CA,USA, 2002.

[14] N. WeiBenberg, A. Voisard, and R. Gartmann, "Using Ontologies in Personalized Mobile Applications," in GIS'04. Washington, DC, USA, 2004, pp. 2-11.

[15] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg, "Top 10 algorithms in data mining," Knowledge and Information Systems, vol. 14, pp. 1-37, 2007.

[16] D. Zhang, G. Karabatis, Z. Chen, B. Adipat, L. Dai, Z. Zhang, and Y. Wang, "Personalization and Visualization on Handheld Devices," in SAC'06 Dijon, France, 2006, pp. 1008-1012.