Exploiting User Behavior Changes in Privacy Disclosure by Modified Clustering Technique

Hongchen Wu, Xinjun Wang∗, Zhaohui Peng and Qingzhong Li

School of Computer Science and Technology
Shandong University, Jinan 250101, China
hc_wu@mail.sdu.edu.cn, {wxj, pzh, lqz}@sdu.edu.cn

Abstract

The analysis of user behaviors has been an important subject in recommending research recently. This paper proposes a modified clustering technique, showing that users privacy disclosure may change when they are answering the information requests, and we argues that their attitudes, including risk, useful, appropriate, played an important role behind those changes. We presented the new data structure in our dataset that would be loaded to experiment, e.g. personal information requests, users' answers to those requests, and most importantly, users cluster and attitude for later analysis. Our modified clustering technique would not only locate users privacy disclosure change by comparing the results from learning their past disclosure behaviors and from learning their current disclosures, but also exploit the relationship between the inconsistence in those two results and their attitudes. The data containing users' answers to a questionnaire with personal information requests was integrated to analyze their disclosure behaviors and attitude with the proposed clustering technique. We indeed find some interesting connections between their privacy disclosure change and attitudes, and the exploration of this paper could benefit to any researchers and online community owners who focusing on user-centered strategies and personal-information-requesting issues.

Keywords: Recommender system, User behavior, Privacy disclosure, Clustering technique

1. Introduction

Electronic communities need to collect huge amount of data to capture the interests of users, so that they could apply beneficial strategies such as recommender system to help users find information they need accurately and efficiently among the Big Data [1-3], and predicts what suits users' interests according to their personal information [4, 5]. However, in order to give recommendations to users precisely, it is needed to be familiar with users as much as possible so that recommender system would understand what kind of product they want to buy [6], what movie they want to watch [7, 8], or what music they want to listen to [9,10], *etc*. This has raised great conflicts with users' privacy concern [11-13], since recently came out some social affairs, such as invalidate using of customers' data, disclosing patients' disease information, and even illegally selling account information from bank customers. On one hand, some users are likely to disclose their information for benefits, such as filling some forms in plaza in exchange of discounts when check out, or free membership if reply the email from commercial companies; On the other hand, users are hesitated to disclosing personal information due to risk of unexpected outcomes. This privacy contradiction is still an interesting topic in


∗Corresponding author

privacy-related recommendation problem. There are many famous researchers focus on solving this problem, and one of them is A. Kobsa, who had proposed the strategy of privacy protection, balancing the best personalization while using lowest users' personal data criteria under their privacy disclosing tolerance. Based on this strategy, he also set up the field of user modeling in artificial intelligence dealing with users' beliefs and goals [14], and User-adaptive applications cater to the needs of each individual computer user, taking for example users' interests, level of expertise, preferences, perceptual and motoric abilities, and the usage environment into account [15]. B. Knijnenburg had forwarded a new approach by analyzing users' satisfaction, disclosure tendency, *etc.,* together with user modeling method in helping them make decision with justifications and increasing users' willingness to divulge demographic and contextual information [16, 17], and made analysis in a number of distinct factors to approve that there is no one-for-all strategy in user disclosure personalization.

This paper, we mainly explored that there were some users' behavior changes occurred during they answering the personal information requests, and further analyze their attitude connections behind that phenomenon. Risk, useful, and appropriate, which could be the possible reasons to "persuade" users to change their privacy disclosure behaviors, are three attitudes that we were mainly looking at. A modified clustering technique was applied, which analyzes users' disclosure behaviors by learning from their previous behaviors or from their current behaviors. The comparison of the two results will reveal those users who had changed their behavior, and we could also focus on them to see if the changes were related to their attitudes, like risk, useful, appropriate. This research could be very useful to any researchers and websites designers who want to apply user-centered strategies and want to release their users' anxiety from requesting their personal information, and we indeed find some interesting connections between users disclosure behaviors changes and their attitudes.

This paper is organized as follows. In Section 2, the study background is provided. Section 3 introduces the new data structure which would be loaded to the later experiment, a modified clustering technique that analyzing the users' past disclosures and current behaviors separately, and possible hypotheses about the relationship between users disclosure behavior change and their attitudes. Section 4 presents our experiment applying the modified clustering technique, and also forward the results and discussion. Conclusion and future works are proposed in Section 5.

2. Background

Our recent study had showed that users' behaviors toward personal information request could change due to unknown reason. It was carried by a mobile app "Div" [18] gathering knowledge from the users and distributing that to those who need it, and it is, in other words, a crowdsoucing platform bridging the online users to brainstorm together. When users carry their phone walking along the street and don't know the name of a really beautiful building, this app could help. The answers are mainly come from two ways: coming from the answer acknowledged by most of the public users, or from the answer of an expert, like someone living nearby for 20 years.
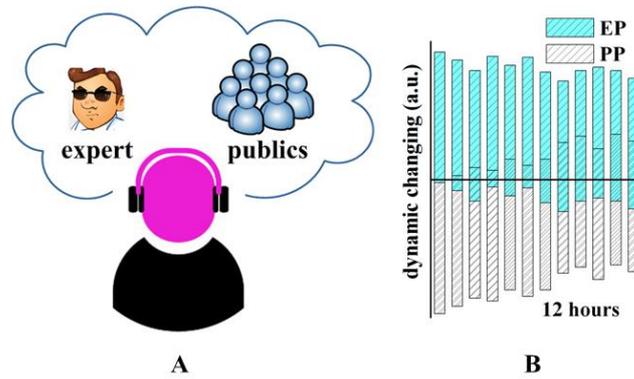
Figure 1. Some users' Trust Changed Overtime

Users could choose the answer either from most public users' opinion, or from the expert's suggestion. The click to each one is counted in the user's own selection record. For example, Div received one request from user X with loading the photo that he saw on the street, and soon Div would post two answers (Figure 1.A). Choosing large group of users will make the count A of this user +1, or count B +1 if he/she selected expert's answer; both +1 when the two answers are equal. If X.count $A > X$. count B, we would say the user X is public preferred user (PP user); or expert preferred user (EP user). This study is used to find out which answer that most users would prefer to, and the results showed that it is almost half-and-half at first 12 hours. However, there is an interesting fact occurred in both group of users: about 1/3 of PP user had changed into EP user while 1/5 of EP users had become PP users in the second 12 hours, no new users joined (Figure 1.B). The reason why the users' trust had changed was fuzzy and vague.

Privacy is recently a big issue related to users who browsing the Internet and receive benefits from "selling" their personal information. Recommender system could be applied for mutual benefits: users would receive good suggestions on what to buy, which suits their need and save time in finding them, because recommender system could make better predictions according to the disclosed personal information, also receive discounts when check out sometimes; managers of the market are hungry for the customers' feedback so they can receive more daily profit according to their personal information, such as placing the most welcome products in an obvious place. However, recent year comes out some terrible issues of companies invading users' privacy, which raise users' concern on privacy protection. Also, in research field, more papers are focusing on privacy related problems, such as Acquisti [19] and Debatin [20] present that privacy and rationality in individual decision making. Traditional theory suggests consumers should be able to manage their privacy. Yet, empirical and theoretical research suggests that consumers often lack enough information to make privacy-sensitive decisions and, even with sufficient information, are likely to trade off long-term privacy for short-term benefits. Their researches mainly discover that users have their privacy disclosure tendencies, and there is no one-fit-all strategy for all users, in other words, each person has his/her privacy tendency on specific items. Furthermore, machine learning techniques, such as Clustering, could be a very useful method in helping us find the knowledge among the Big Data [21]. Our recent research works had taken one step beyond those previous works, and proposed that users' disclosure behaviors may change over time due to some reason. If we could know the reason causing users to release the discomfort and disclose more information to us, the time of cold start will surely shorted and gain more trust from our customers from the electronic communities.

3. Model Implementation

3.1. New Model Elements

This section presents the model elements, which would be used for detecting users' disclosure change in the later experiment. Since our analysis lies in users answering a sequence of requested questions, it is needed to show the data structure that used as input of our experiment. The information we request from users mainly varies in two types: context category or demographic category. Each request has an indicator "*Coin*" showing its sensitiveness is high or low. The more the value of *Coin* close to 1 indicates this is a high sensitive request; otherwise this is a low sensitive request. Users just need to answer the requests with "YES" or "NO", which will update the *Coin* value in return. Beyond that, two core data structure of this paper are shown below:

UserCluster (int Cluster$_1$, int Cluster$_2$, …, int Cluster$_n$)

This data structure shows which cluster each user was belonging to when (s)he finished one request. The clustering technique are applied in later section, and the clustering algorithms are operated on two different ways: clustering users based on each personal information request, which means the cluster of users is point-to-time related and does not consider the connections between different disclosure behaviors, let alone the previous disclosed answers; or clustering users based their past behaviors, which using their past disclosures to predict the current disclosures. These two ways of computing which cluster one user would belongs to are different from whether those previous disclosures were considered in exploiting the knowledge.

We would mainly compare the two clustering results so that we can figure out if some users have inconsistent belongings to the clusters: if the one person's cluster, which learnt by the disclosures 1 to x-1, is different from the other result of clustering algorithm, which learnt by the disclosure x, we could possibly announce that (s)he had changed the disclosure behavior after (s)he answered the No.x request. There are 12 requests related with personal information were asked in experiment section, and we call them r1 to r12 if the clusters were learnt by previous disclosures, while R1 to R12 if the clusters were learnt by current disclosures.

UserAttitude (int Risk[], int Useful[], int Appropriate[])

We will further pick up those users in inconsistent results, and try to find out the reason behind the differences by looking at their answers of risk, useful and appropriate of each request on timeline. Users' attitudes are represented in these three facts. When (s)he finished one request, we would also ask three questions: Do you think this request is risky/useful/appropriate? Users could answer each question with a 7-scale-value selection, varying from 3 (I think disclosing my personal information to this request would be very risky/useful/appropriate) to -3 (disclosing my information to this request would not be risky/useful/appropriate). In the experiment section, we will mainly look at if there were any connections between users' disclosure behaviors and these three attitudes.

3.2. Clustering Method Implementation and Proposed Hypotheses

If one user's cluster changed due to some reason, it is needed to compare his/her past behavior with similar one's behavior whose cluster did not changed. For each participant X who attends our study labeled as Participant X ($a_1(X)$, $a_2(X)$,…, $a_n(X)$), where $a_r(X)$ stands for participant X's answer towards No.r question (r = 1, 2, 3…n), and the distance between two participants x_i and x_j can be defined as follows:

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{n} \left(a_r(x_i) - a_r(x_j) \right)^2}$$

Based on the definition of distance, the clustering algorithm implementation is given below:

```
01  Input:
02      Participants dataset X {x< a₁(x), a₂(x),…, aₙ(x)>}
03  Output:
04      4 Groups of Participants labeled as G₀₀, G₀₁, G₁₀, G₁₁
05  Begin
06  For all participants x<a₁(x), a₂(x),…, aₙ(x)> in dataset X
07  add into table training_samples
08  Initially define the core participant in each cluster randomly or
09  manually as P₀₀, P₀₁, P₁₀, P₁₁
10  Given a sample xq that needed to be classified in one Group
11  Pick up nearest core participant
12  by computing the minimum d(xq, Pxx), where xx = {(0,0), (0,1), (1,0), (1,1)}
13  Groupxx = Groupxx + xq
14  X = X - xq
15  Recalculate the core Pxx of Groupxx
16  Until X = ∅
18  End
```

This algorithm is used to learn users' cluster by previous answers or by current answers, detecting if their behaviors are detected like "first same then different", and the overflow of the algorithm can be viewed in Figure 2, including imported data, visualized cluster assignments from WEKA [22], and the saved clustering results. We also give three possible hypotheses that could relate with users' disclosure behavior change and their attitudes.

H1, risk will cause users to change their disclosure change. We will support this hypothesis if we find the value difference in risk among users who were clustered inconsistently.

H2, useful will cause users to change their mind in disclosure behaviors, and we could accept this argument if there is an obvious gap in useful value between the users who owned inconsistent grouping results.

H3, appropriate will lead to users to change their behaviors in personal information disclosure. We could accept this hypothesis if we discover that users had different perspectives on appropriate when they changed their disclosure behaviors

Now with all data structure and algorithm ready, we can run our experiment.
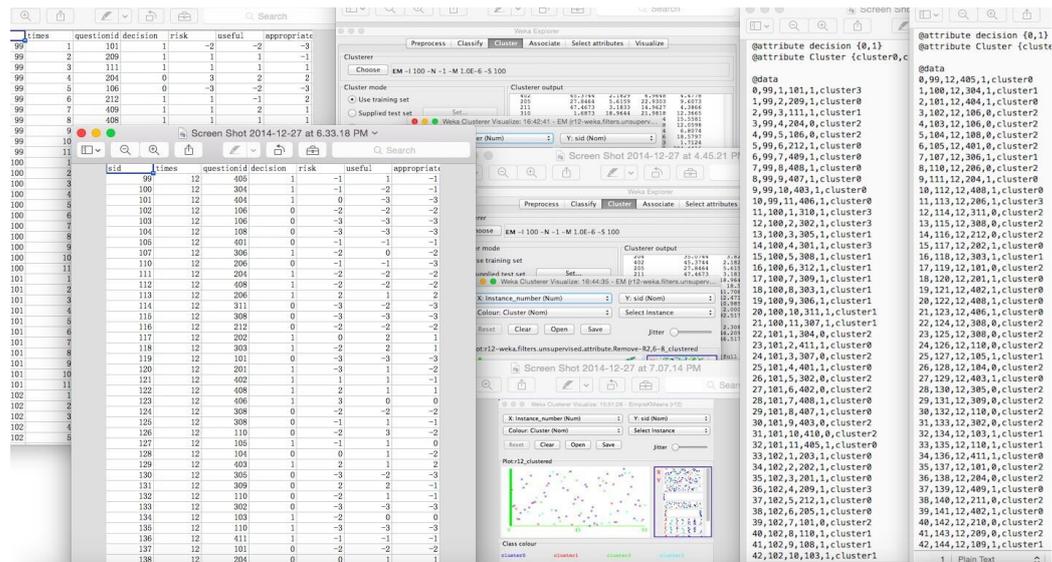
**Figure 2. Overflow of the Modified Clustering Technique, Learning users'
Clusters from Previous Disclosures or from Current Disclosures, then We
Check the Inconsistent Clustering Results by Looking at the Values of Risk,
useful, and Appropriate**

4. Experiment

4.1. Pre-study and Dataset Description

We do the experiment with the data integrated from a recent research [23], which
containing users' answers to a list of 12 personal information requests, and we
mainly analyze their disclosure behaviors the proposed clustering technique. The
dataset is mainly the 12-disclosures behavior dataset, and we cluster our users into
groups based on their disclosures on time-scale rage, where are 12 requests in total.
There were 376 users invited to join this study and we indeed find some of them
who has inconsistent clusters.

4.2. Clustering Results and Discussion

Interestingly, there were some users cannot be grouped to any clusters due to
their up-to-time disclosures were so less related with others, and this phenomenon
were more evident in the clustering algorithm importing users' previous disclosure
behaviors. Thus, some of the users were removed from the cluster results
comparison due to no cluster record. However, even though we removed those
users, we still discovered that there were 91 users can be used for comparing the
grouping results in Figure 3.
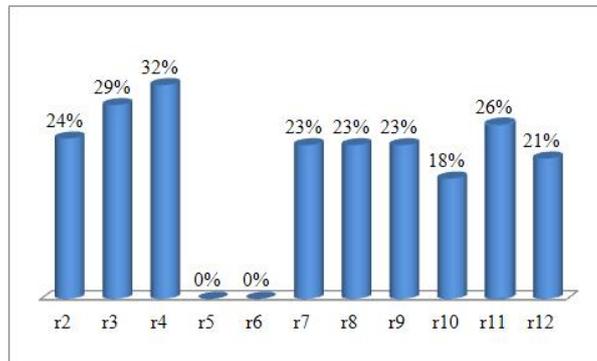
Figure 3. The inconsistence rate of users's clusters in each request

Here are inconsistent rate of these two clustering algorithms, shown in Figure 3. Even though most of users do not change their cluster, there are still some of them changed their disclosure behaviors. To find out what request had caused the grouping inconsistence, we need to further look into the summary values of risk, useful and appropriate in those cases.

For those users do not change their clusters, the percentage of negative value with risk is 44.71% and the percentage of positive value is 43.18%, shown in Figure 4.a. For other users changed their cluster, the percentage of negative value with risk is 38.96% and the percentage of positive value is 45.89%. The percentage of positive value has increased considerably, and thus we would say users changed their disclosure behaviors had considered more risks, so we could say the hypothesis of risk caused users to change their mind is supported and it is the risk that played an important role in disclosure change consideration.

For those who do not change their cluster, the percentage of negative value with useful is 43.76% and the percentage of positive value is 42.59%, shown in Figure 4.b. In contrast, for those users changed their cluster, the percentage of negative value with useful is 36.36%, while the percentage of positive value with useful is 47.62%. We see that there are also obvious differences in this comparison, so we shall say the hypothesis of useful playing an important role in users' disclosure behaviors changing is supported, too.

For users who do not change their cluster, the percentage of negative value with appropriate is 46.94% and the percentage of positive value is 36.71%, shown in Figure 4.c. For comparison, the rest of users who changed their disclosure attitude own the percentage of negative value with appropriate is 42.86% while the positive values get the percentage of 36.71%. As a result, we shall say there is no obvious difference among the percentage division of appropriate behind the users' disclosure behaviors change, so we shall say the hypothesis of users change their mind due to appropriate is not supported.
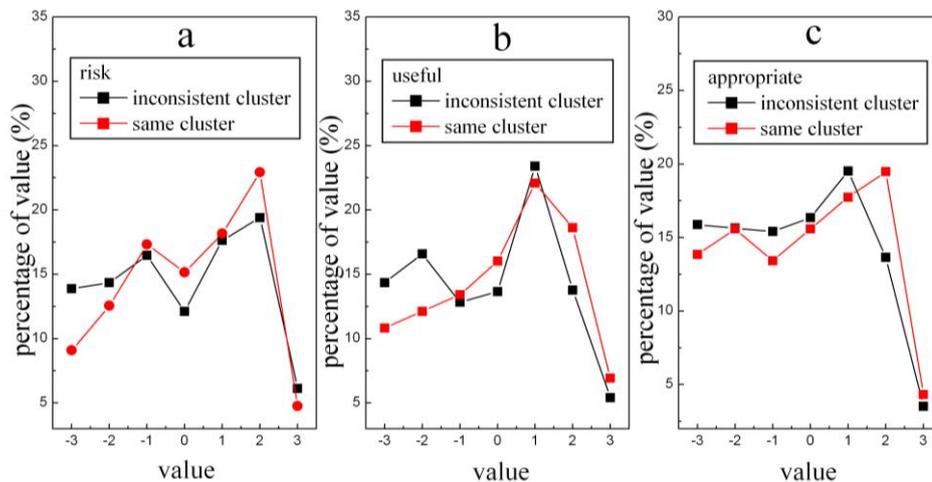
Figure 4. The Comparison of the Values of Risk, useful, Appropriate between the users who were clustered in Inconsistent Groups and the users who were clustered in same Groups

There should be some possible reason why users consider risk and useful as the main attitude evaluating their disclosure behaviors.

Firstly, users mainly evaluate risk and benefit when they are asked to disclose their personal information. For example, when customers are checking out, they probably agree to disclose their name, age, gender, home address in exchange of discounts in bought properties. The plaza is the place where customers are quite familiar with and there are many other customers also agree to disclose their information, so there is more benefit with low risk to disclose the personal information to the trusted plaza. In contrast, there are more risks if users are asked to disclose the personal information to those companies that they don't know. Disclosing information to those less well-known companies cannot guarantee the proper use of their information, neither avoiding negative consequences that caused by illegal use of personal information. Risk should be a major attitude that makes sense in users' decision on information disclosure. Secondly, useful is another important fact in users' decision of whether to disclose personal information or not. We could possibly disclose our personal information to online purchase community so that they can know us better and give us suggestions on what to buy to embrace our Christmas plan with my children's favorite food and toys. We could also know some coming event hosted by my faculty and buddies on Facebook if I disclose which university I am now studying in. Google can also suggest me a nearest restaurant where I can have diner in by disclosing my current location, which could be very useful and saving my time on finding that. One possible fact that could persuade users to disclose more personal information is definitely more useful than before. Thirdly, appropriate could be a fact in users' decision making of disclosing their personal information, but we would say users should have different understanding toward this concept. This paper does not consider the parameter of users in influencing their disclosure decisions. Seniors could possibly disclose less information than junior on their home address and family information. Females could be more sensitive than males to be requested of age-related issues like birthday, children's age, marriage date, etc. Computer-majored users could cease their tongue when being asked to give their online information than others who do not working with computers, because they could possibly know where the information would be used to. However, in this paper, the dataset told us that the appropriate does not make a major role in "persuading" users to change their

disclosure actions, and it is the risk and useful that could lead to change their mind in decision-making.

5. Conclusion and Future Work

This paper had proved that there were indeed some behavior changes took place, and we argue that risk and useful played an important role in that phenomenon. We achieved it by adopting a modified clustering technique which analyzing users' past disclosures and current disclosures separately. As a result, we recommend researchers and electronic website owners to push forward to carry more beneficial and useful policy in information-requesting strategy while less risk voluntary adoptions if they want to know their users better.

In the future work, we could establish some more complicated experiments that combining users' parameters and their attitude, to further exploit the connections of users' lively mode with their privacy disclosing preference. A cross-community recommender system will be also constructed to provide the platform of discovering users' privacy related changes and requesting issues.

Acknowledgements

References

[1]   B. Sarwar, G. Karypis, J. Konstan and J. Riedl, "Item-based collaborative filtering recommendation algorithms", Proceedings of the 10th international conference on World Wide Web, (2001) May 1-5, Hong Kong, China.

[2]   J. G. Boticario, "Content-free collaborative learning modeling using data mining", User Modeling and User-Adapted Interaction, vol. 21, no. 181, (2011).

[3]   B. P. Knijnenburg and A. Kobsa, "Making decisions about privacy: information disclosure in context-aware recommender systems", ACM Transactions on Interactive Intelligent, vol. 3, no. 20, (2013).

[4]   B. P. Knijnenburg, M. C. Willemsen and A. Kobsa, "A pragmatic procedure to support the user-centric evaluation of recommender systems", Proceedings of the 5th ACM International Conference on Recommender Systems, (2011) October 23-27, Chicago, IL, USA.

[5]   Y. Ge, Q. Liu, H. Xiong, A. Tuzhilin and J. Chen, "Cost-aware travel tour recommendation", Proceedings of the 17th ACM SIGKDD conference on knowledge discovery and data mining, (2011) August 21-24, San Diego, California, USA.

[6]   B. Xiao and I. Benbasat, "E-commerce product recommendation agents: use, characteristics, and impact", Mis Quarterly, vol. 31, no. 137, (2007).

[7]   M. Szomszor, C. Cattuto, H. Alani and K. O'Hara, "Folksonomies, the semantic web, and movie recommendation", Proceedings of the 4th European Semantic Web Conference, Bridging the Gap between Semantic Web and Web 2.0, (2007) Jun 3-7, Innsbruck, Austria.

[8]   C. Ono, M. Kurokawa, Y. Motomura and H. Asoh, "A context-aware movie preference model using a Bayesian network for recommendation and promotion", User Modeling, vol. 4511, no. 257, (2007).

[9]   H. C. Chen and A. L. P. Chen, "A music recommendation system based on music data grouping and user interests", Proceedings of the tenth international conference on Information and knowledge management, (2001) November 5-10, Atlanta, Georgia, USA.

[10] Y. C. Zhang, D. Ó. Séaghdha and D. Quercia, "Auralist: introducing serendipity into music recommendation", Proceedings of the fifth ACM international conference on Web search and data mining, (2012) February 8-12, Seattle, Washington, USA.

[11] L. Li, Q. Z. Li, L. J. Kong and Y. L. Shi, "Efficient Query Integrity Protection for Multi-tenant Database", International Journal of Database Theory and Application, vol. 7, no. 31, (2014).

[12] K. Shyong, D. Frankowski and J. Riedl, "Do you trust your recommendations? An exploration of security and privacy issues in recommender systems", Lecture Notes in Computer Science, vol. 3995, no. 14, (2006).

[13] A. Kobsa, "Privacy-enhanced web personalization", Lecture Notes in Computer Science, vol. 4321, no. 628, (2007).

[14] A. Kobsa and W. Wahlster, "User models in dialog systems", Springer-Verlag, New York, **(1989)**.

[15] A. Kobsa and J. Schreck, "Privacy through pseudonymity in user-adaptive systems", ACM Transactions on Internet Technology, vol. 3, no. 149, **(2003)**.

[16] B. P. Knijnenburg and A. Kobsa, "Helping users with information disclosure decisions: potential for adaptation", Proceedings of the 2013 international conference on Intelligent user interface, **(2013)** May 19-22, Santa Monica, California, USA.

[17] B. P. Knijnenburg, A. Kobsa and G. Saldamli, "Privacy in Mobile Personalized Systems: The Effect of Disclosure Justifications", SOUPS Workshop on Usable Privacy & Security in Mobile Devices, **(2012)** July 11-13, Washington, DC, USA.

[18] H. C. Wu, X. J. Wang, Z. H. Peng and Q. Z. Li, "Div-clustering: exploring active users for social collaborative recommendation", Journal of Network and Computer, vol. 36, no. 1642, **(2013)**.

[19] R. Gross and A. Acquisti, "Information revelation and privacy in online social networks", Proceedings of the 2005 ACM workshop on Privacy, **(2005)** November 3-7, Alexandria, VA, USA.

[20] R. Baden, A. Bender, N. Spring and B. Bhattacharjee, "Persona: an online social network with user-defined privacy", Proceedings of the ACM SIGCOMM 2009 conference on Data communication, **(2009)** August 17-21, Barcelona, Spain.

[21] W. Hu and Q. H. Pan, "Data clustering and analyzing techniques using hierarchical clustering method", International Journal of Database Theory and Application, vol. 6, no. 109, **(2013)**.

[22] http://www.cs.waikato.ac.nz/ml/weka/.

[23] H. C. Wu, B. P. Knijnenburg and A. Kobsa, "Improving the prediction of users' disclosure behavior…by making them disclose more predictably?", Proceedings of the tenth Symposium on Usable Privacy and Security, **(2014)** July 9-11, Menlo Park, California, USA.