# The Use of Persona in Recommendation System and Privacy Protection

Suduo Li[a], Kaiying Deng[b] , Jingwei Deng[c], Yingxing Li[d]

College of Mathematics & Computer Science, North-West University for Nationalities, Lanzhou 730124, China

[a]email: 435752897@qq.com, [b]email: 470601995@qq.com,
[c]email: 494521241@qq.com, [d]email: liyx555@126.com,

**Keywords:** Persona, Digitalized User Model, Personalized Recommendation, Privacy Protection

**Abstract.** Aimed at the contradiction between personalized recommendation and privacy protection, this paper puts forward the basic idea of persona, a digitalized user model. The method uses browser to analyze user's access behavior, gets a comprehensive and accurate user model, so as to help realize the personalized recommendation. In consideration of privacy, let user model saved on the client side, and user can decide to what extent browser will offer his/her own user characteristics to the target website, so user's privacy is fully protected. Thus it solved the contradiction between the personalized recommendation and privacy protection successfully.

## Introduction

As the exponential growth in international information, the traditional search engine no longer satisfies users' needs. For it can only offer the same sorted results to all users, and can't give suitable services to distinct users according to their preferences. It's very hard to make choice from the large amounts of search results. This is the question of disorientation and information overload. Recommendation system is used to solve such problem. It's an application which establishes user's interest model according to analyzing his browsing behavior so as to push the needed information to him. Since promoted by Marko Balabanovic et al in 1995, recommendation system has got a rapid development and made enormous wealth in e-commerce. Now it has become a standard web design in social networking sites and e-commerce sites, etc. Its importance is beyond doubt. There is a famous case that the world's biggest films online renting website, Netflix, declared in Oct 2006 that if anyone can promote the prediction ability of their recommendation system by 10%, he or she will get one million dollars as a reward, indicating the great value of recommendation system to Netflix[1].

However, there is a contradiction between recommendation system and privacy protection. Recommendation system needs to collect user's personalized information. The more accurate and comprehensive of user's personalized information it masters, the more effective the recommendation system is. But the process inevitably involves with user's privacy, such as collecting user's basic information, preferences, browsing behavior and contents, storing, processing, transmitting and computing user's personal information without permission. So many people refuse to give their own data because of privacy considerations. Thus the accuracy of recommendation system is significantly reduced and impeded its application. How to deal with privacy protection is an important issue that recommendation system must address.

## Problems in Privacy Protection

Prof. Alan Westin in Columbia University pointed out that "Privacy is the claim of individuals, groups, or institutions to determine for themselves when, how and to what extent information is communicated to others" [2]. Privacy protection is aim to take some security methods to prevent someone leak and abuse user's privacy. There are three kind of privacy protections[3]: Protect the information which can uniquely identify a person; Protect the right of a person to seclusion oneself; Protect the right to control one's own information data. Many countries have made some

 

corresponding laws to protect people's right of privacy. Many websites expressed their privacy protection strategies to visitors. But all these need certain technologies to support the work.

Considered from the aspect of software architecture, user's personalized information can be stored in server-side or in client-side[4]. At present most applications use the server-side mode because after years of running, many large websites have accumulated rich user information to use. But this caused great privacy threaten to users, for they have no way to see or control their own data. Either the administrator or an invader of the website can get user's information easily. In addition this mode can't get a comprehensive user profile. For a server has user's few data only when user visits its website. It doesn't know user's preference from other websites. When a user comes to a website for the first time, recommendation will face the problem of cold start. Because of knowing nothing about the user, it can only give a generalized recommendation which probably may not suitable for the user.

For these reasons many scholars proposed to use distributed architecture, make recommendation via client, P2P or agent, etc. The common idea is to store user's information in client-side. Obviously the client-side is more safety because in this way user can fully control his data. But there are still many problems to be solved. If user profile modeling and result resorting are all done by client, it will increase client's work and make unnecessary flow in network. The other drawback is that due to lack of user's personalized information, the recommendation algorithm can't use some knowledge that is only available on the server side (e.g., PageRank score of a result document.). So the reasonable way is to use client-side mode while let user provide his partial information on his willingness. Paper [5] applied this way but it offered two parameters for users to determine the content and the amount of privacy exposed. In fact it's rather difficult to make decision through such abstract numbers. We need a way to do it easily and automatically.

The other problem is the real-time demand of user's modeling. Users' interests and preferences and concern of privacy often change with time, occasion, feelings, etc. Privacy protection technologies haven't think of this at present[6]. In our paper we introduced a forgetting factor in user modeling , let user's recent behavior make more contribution to his profile.

Besides the sensitivity of different privacy data is also an important issue. Research have shown that user's sensitivity of personal data differ with data's type. For example , most people are much sensitive about their information of identity, savings, credit, etc., but are less sensitive about their occupation, interests, education, and so on. Paper[5] used the parameter "minDetail" to determine the exposed privacy, it means the same target is used in all attributes of user's features. It is evidently not in conformity with the reality. Actually there is no different treatment between sensitive and non-sensitive attributes in current privacy protection technologies. We proposed using a visualized mask to hide user's sensitive data according to user's willingness. Thus user's profile can be very concrete in some aspects while very vague in others.

## User Model: Persona

From above analysis we know that in order to consider both accurate recommendation and privacy protection, we should analyze and store user's profile at client-side, confuse the real profile with a mask and get a pseudo profile, which only contains partial information of user's. When visiting a website, browser will send the web request along with user's pseudo profile so as to make recommendation by website. We designed a mechanism to analyze user's behavior also alter the mask easily and automatically. For this assumption we introduced the conception of persona.

### A. Digitalized User Model

Persona, also known as user model, an abstract of a user, is used to represent a user group. Here we discussed is the web persona, which is the outline of the real characteristics of the target user group of a site, also is the prototype of real users. As for present research, persona is always described in narrative form such as words and photos. In this paper, we propose to describe persona with digitalized attribute value. The idea is to make user's character easy to compute.

A representative persona can be described by gender, age, region, occupation, hobbies and other attributes. Assume that each attribute were expressed at a certain digits, all these digits stitching

together constitute the user's persona value. Our aim is to study how to assign each attribute with a number through the theory of user behavior analysis.

In the all properties mentioned above, region is the most recognizable feature. Once obtaining the user's IP address it can be mapped to a region, accurate to a city and even to a work unit. Suppose using area code, region property can be expressed by certain digits.

Judgment on gender, there are a variety of different algorithms. According to the results of the statistical analysis, gender may have a different interest in the different vocabularies to some extent. Such as the research of Vdoing (a kind of software about website traffic statistics and analysis system) shows that there are only 1% of female are sensitive to the word "software", while for male this number is up to 99%. At present professional research has been able to give any keyword its distribution of gender, and the distribution of men and women on any URL. In addition, Twitter's algorithm pointed out that men and women have differences in the habit of using language, women are more likely to use emoticons, abbreviations, repeated letters to express their emotions[7]. Therefore, the probability of gender can be obtained by comprehensively evaluating the information such as the URL each time the user visited, the search terms he used, the comments he posted and so on. Thus the gender property in persona can be expressed by a number from 0 to 9 , digital approximates to 0 indicates the user is more likely to be a man, approximates to 9 indicates the user is more likely to be a women, the middle value of 5 indicates that the gender is unknown.

Similarly, other properties can also be obtained through the user's visiting behavior. Some properties have exact value and can be represented by coding, such as region; while some properties have to determine the value using a tendentious judgment through the probability distribution, such as gender, age, occupation, hobbies, etc. All these attribute values stitching together constitutes the persona value of the user.

### B. Hierarchical and Paralleling User Model

User characteristics are hierarchical inherently, it describes the granularity of an attribute. For example, when mentioned "region", it can be layered as country, province, city, district, road, etc. Obviously occupation，hobby and many other attributes are the same satiations. While gender only has one layer.

So we can describe user's characteristics by a layered strings, such as O84.4.1P86.62，means occupation is regular higher education, position is in Gansu,China. (coding method referenced from http://www.360doc.com/content/14/1023/11/5052258_419170242.shtml). Letter indicates a kind of attribute, such as "O" means occupation and "P" means position. Number is corresponding attribute values, ". "is a separator means the attribute value also has a lower level. The more separators, the more specific of the attribute values. According to user's demand for privacy protection., browser may sent the user characteristics like OC86, that is Chinese nationality, occupation is unknown. For those frequent visited reliable sites, it may send out more specific characteristics.

Besides hierarchical relationship, some attributes of user's characteristics are paralleling relationships, such as hobby. A man may have many hobbies, such as sports, music, etc. Sports can be specific to fencing, swimming and so on. Assumes that the degree of be fond divided from 0 (dislike) to 9 (greatly enjoy), below tied paralleling attributes expressed in parentheses , separated by commas. So the attribute "hobby" might be expressed as: H (Sport7 (fencing6, swimming9), music8), which means the user's preference of sport is level 7, including fencing level 6, swimming level 9. And the preference of music is level 8.

### C. Standard Structure of User Model

In this method, user's characteristics can be described by tree data structure. Sibling nodes represent paralleling relationship, and child nodes represent hierarchical relationship. The division of the attribute value must have a certain standard. Some attribute values already have some standard for reference, such as occupation, position, etc. But there are also some attributes have no standard, such as hobby. Thus needs to create some rules to evaluate by percentage rate.

For example, mentioned as Occupation, the United Nations Economic and Social Affairs Bureau has made International Standard Industrial Classification of All Economic Activities, recommends

global adoption. It divides national economy into 10 categories, each category was further divided into main class, middle class, and small class. In 1990 it was revised for the third time. In which all economic activities were divided into 17 categories, 60 main classes, 173 middle classes, 306 small classes. For example, the code of software development is 6510. Of cause there are other standard for occupation to reference. Our previous example applies the class categories in the new Classification of National Economic Industries in China (GB-T4754-2002).

For position, it's very easy to get user's IP address and learn his/her rough place. If it is a mobile application, user's location can be accurate to a street, a small residential area by mobile device with GPS function. By means of big data analysis, we can recognize the same user with a variety of equipment, making it easier and more precise to determine the user's location. On location coding, we can consult states, provinces and cities coding region, and the national administrative divisions code, etc. More accurate location code can use GPS coordinates.

For those attributes that have no ready coding standard to reference, such as gender, age, hobbies, and so on, we can figure out a probability distribution of user's attributes according to amount of old user information amassed by websites and the content and frequency of user's visiting behavior, so as to get a tendentious judgment, and quantify it to level of 0 ~ 9.The concrete calculation method is given below.

**D. Algorithm of Computing the Attribute Value of User Model**

Browser is the place that user must pass by when visiting the Internet, so it's the most suitable place to collect user's browsing behavior. Let browser add a function of recording persona. It calculates the user's persona value according to characteristics corresponding to the URL of webpage which user is opening. When Browser sends a request to the web server which provides personalized services, it also sends user's persona value. So the website could give personalized information recommendation based on user persona to improve user's browsing experience. At the same time web server updates characteristics of the URL regularly according to visitors' persona. The specific algorithm is described as follows:

1. According to the known sample data, web server computes the probability distribution of attribute values on each URL, sets the initial value for $URL_1(X)$, $X$ is a vector, on behalf of all the attributes that need to be described by probability distribution ;

2. Once user opens a webpage, browser then calculates user's persona value. If it's the first time it analyzes user's behavior, the value will be:

$$Persona_1(x) = URL_m(x) \quad (1)$$

Subscript $m$ represents the number of times that web server updates characteristic of the URL when the user visiting the URL;

3. Browser modifying user's persona value if it analyzes user's behavior for the times $i$:

$$Persona_i(x) = \frac{1}{i}[(i-1)Persona_{i-1}(x)\rho^T + URL_m(x)] \quad (2)$$

Where $\rho$ represents forgetting factor, $0 <= \rho <= 1$, used to weaken the influence of old data to the user's characteristics. $T$ represents the interval cycles of the nearest visits. $T$ should be set a reasonable interval according to user's average access frequency, such as 3 days, or a week, and so on.

4. $Persona_i(x)$ represents the multiple attributes of a user's latest probability distribution, quantitative to the range of 0 ~ 9, the attribute values of the latest scores are calculated as follows:

$$Score_i(x) = round(10\ Persona_i(x), 0) \quad (3)$$

5. When browser sends user's requests of the URL information to the web server, it also sends the encrypted persona of the user together. Web server statistics all user's persona on same URL in a period and gets the average persona value noted by $E(Persona(X))$;

6. Web site updates characteristics of each URL regularly:

$$URL_m(X) = \frac{1}{m}[(m-1)URL_{(m-1)}(X) + E(Persona(X))] \quad (4)$$

The superiority of the algorithm is to realize the data sharing. It enables user model derived from different websites to be taken together, and form a more comprehensive, complete user model which can be used for all sites provide personalized service.

### E. Pseudo the User Model

From above we know that through information exchange between the browser and the server, browser could get rich user visiting behavior, to form a full and accurate user model. It can even know users better than user themselves, so as to get a better understand about user's real demand, and benefit users to enjoy personalized recommendation service provided by websites.

But in fact, for reasons of privacy, we needn't completely expose out the comprehensive and accurate user model. Browser will provide a visual interface that allows user to view their own persona, and modify the degree of exposure of those hierarchical attributes, such as location, occupation, interest ,etc. If user feel too much trouble and don't want to set, browser will use the default algorithm to make decision. At last it send the target site a pseudo user model.

For website, it only needs necessary user information. Such as shopping site, it only interested in the user's purchase behavior, and care little about occupation, age, etc. Music site is only interested in user's playlist, but has nothing to do with user's occupation, education, shopping behavior, etc. So the website should state which attributes it concerns about. Before sends out user's requests, browser should get the needed attribute list of user's from the website, together with the site's credibility, intercept partial characteristics from the whole persona and calculate the content with certain fuzzy factor , then send it out to the target website.

## Discussions

The innovation points of this paper is embodied in the following aspects:

About the user modeling, this paper puts forward the basic idea of digital persona. This not only makes the description of user's characteristics more concise and normative, but also makes the user clustering easier. For just comparing the number of corresponding attributes, it can know the differences between users.

When building user model, we used the forgetting factor in the algorithm to meet the real-time demands of modeling, let user's recent behaviors make more contribution to his profile.

Different with existing user modeling algorithm that can only rely on the server of visiting website to collect user's behavior, this method solved the problem of data acquisition comprehensively through the interaction of browser and web server. By this way, user's every browsing is helpful to create his user model.

In order to respect user's privacy, the algorithm allows the user to take the initiative to modify or close his persona values. Browser will ask the user whether to submit his persona value to the visiting website, or determine it automatically via safety rules set in advance. So different from the situation of collecting user's visiting information by current browser, in this method the browser doesn't record user's all detailed browsing process, but just converts user's concrete visiting behavior to a number which represents user's characteristics. Whether to use and how to usethe number is entirely depended on the user.

## Conclusion

Through the attribute division of user's persona combined with user behavior analysis, this paper proposed the concept of describing persona with digital, which will make the user identification and user clustering easier to implement. It states the basic idea of constructing digital user model through the interaction between web server and browser. The future work is to further study the influences of user behavior to user modeling, and give more accurate method of persona digitized algorithm.

## Acknowledgement

This work is supported by the Fundamental Research Funds for the Central Universities of China (31920150080, 31920150039); Humanities and Social Sciences Youth Projects of the Ministry of Education of China(12YJCZH027, 13YJCZH029).

## References

[1] Haralambos Marmanis, Dmitry Babenko, Algorithms of the intelligence web (Manning Publications, 2009).

[2] Westin, A.F., Privacy & Freedom( The Bodley Head, London ,1967)

[3] WANG Yang, KOBSA A, Privacy enhancing technologies (GUPTAM, SHARMANR, Handbook of Research on Social and Organizational Liabilities in Information Security.Hershey:IGIGlobal,2009:203-227).

[4] Xuehua Shen, Bin Tan, ChengXiang Zhai, Privacy Protection in Personalized Search, ACM SIGIR Forum, Vol.41 No.1 June 2007

[5] Y Xu，K Wang，B Zhang，Z Chen, Privacy-enhancing personalized web search, International Conference on World Wide Web, 2007:591-600)

[6] Guoxia WANG, Lijun WANG, Heping LIU, Study progress of privacy protection techniques used in Personalized recommendation system, Application Research of Computer, Vol29 ,No6 ,Jun 2012,

[7] New Twitter Algorithm Could Out Dudes Pretending to Be Lesbians. http://news.yahoo.com/twitter-alogrithm-could-dudes-pretending-lesbians-133105472.html