International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Survey of User Modeling Techniques with Specific Emphasis on Considering Demographic Attributes

Prajakta Shitole[1], Mrs. M. A. Potey[2]

[1]Computer department, Savitribai Phule Pune University, Pune, India

[2]D. Y. Patil College of Engineering, Akurdi, Pune, India

Abstract: ***Users demographic attributes like age, gender location etc. plays an essential role in today's web applications. Previous research shows that there is relationship between users' browsing behavior and their basic characteristics. In this paper we discuss various Data Mining and machine learning approaches for modeling and prediction of demographic characteristics. Data mining and machine learning techniques have the ability to handle large amounts of data and to process uncertainty. These characteristics make these techniques suitable for automatic generation of user models that simulate human decision making. The paper also represents guidelines that show which techniques may be used more efficiently according to the task implemented by the application.***

Keywords: Demographic Attributes, Data Mining, Machine Leaning, user model, uncertainty, human decision making.

1. Introduction

Demographic information plays an important role in personalized web. However, it is usually not easy to obtain this kind of personal data such as age and gender. The demographic attributes of the web users can be predicted from their Web browsing behaviors, in which the Webpage view information is treated as a hidden variable to propagate demographic information between different users [5]. To date, investigations of Web activity have typically relied on three types of data sources: surveys, toolbar tracking, and aggregated site-level analytics. Surveys are perhaps the most popular methodology for investigating Internet use, for measuring attitudes and general usage trends, surveys typically involve small, non-representative samples of the population, and rely on users' in complete and sometimes in accurate statements of their own online behavior. In contrast, toolbar tracking data (e.g., as collected via user-installed toolbars distributed by Google, Microsoft, and Yahoo!) provide complete browsing histories for millions of users. Though quite large in scale, these data often suffer from significant sample bias, and perhaps more importantly, it is difficult to disentangle an individual's browsing history from that of an entire household.

Internet usage nowadays knows no real age limits. There is a difference between children and adults in their search behavior. Many children have access to the Internet and explore the web from a young age. Unfortunately, the web adapts slowly to the needs of children. There are many good techniques in IR for adults too, but still not much scientific insights on how to design search engines for age specific users in terms of both user interfaces and underlying algorithms. To characterize the search behavior of users with varying ages and the browsing activities that lead to search. In terms of search behavior, there is a need to focus on identifying the difficulties that different aged based users encounter on the internet when they search for information with a state-of-the-art search engine. Given that current search engines are not designed for age specific user requirements, it is hypothesized that they struggle in their way to find information on the Web and that these difficulties are quantifiable from their interactions with the search engine.

Interactive software systems are employed by users with varied goals, interests, levels of expertise, abilities and preferences. In order to cater to a wide range of web users, many programs already allow user (or system administrators) to tailor the program behavior to individual needs (usually by completing preference menus or editing profile files). Since User-initiated and user-selected adaption is often not possible due to the amount and nature of the required adaptive behavior [2]. For this purpose, systems must make assumptions about the user. A user model contains the system's assumptions about all aspects of the user which are deemed relevant for tailoring the dialog behavior of the system to the user. A user modeling component in an interactive system draws assumptions about the user based on the interaction with the user, stores them in an appropriate representation system, infers additional assumptions from initial ones, maintains the consistency in the set of current assumptions and supplies other system components with assumption about the user.

2. Related Work

The most relevant literature on user modeling, capturing search behavior and information retrieval based on demographic attributes are described in the following paragraphs.

A. Information Seeking

The first studies attempting to characterize the search behavior of children have been carried out using systems with no Internet, such as electronic libraries, CD-ROMs, and OPACs (Online Public Access Catalogs). The first studies attempting to characterize the search behavior of children have been carried out using systems with no internet, such as electronic libraries, CD-ROMs, and OPACs (Online Public Access Catalogs). Solomon [15] explored the search success of elementary school children when using an OPAC. The author found that children were able to use the system effectively when engaging in simple searches. However, they

found that complex searches were hampered by the lack of mechanical skills of children. They pointed out that factors such as typing on the keyboard, spelling, limited vocabulary, and reading expertise are skills that are not developed enough in children in order to use the OPAC system studied [16].

Sergio Daurate Torres et al.[6] employed toolbar logs from a commercial search engine to characterize the browsing behavior of young aged users, particularly for understanding the activities on the Internet that trigger search. They quantified the proportion of browsing and search activity in the toolbar sessions and we estimated the likelihood of a user to carry out search on the Web vertical and multimedia verticals (i.e., videos and images) given that the previous event is another search event or a browsing event. It was even observed that these metrics clearly demonstrate an increased level of confusion and unsuccessful search sessions among children. They also found a clear relation between the reading level of the clicked pages and characteristics of the users such as age and educational attainment.

Bernard J. Jansen et al. [9] evaluated the effect of gender targeted advertising on the performance of sponsored search advertising. They used the Microsoft ad Center Labs Demographics Prediction Tool (http://adlab.microsoft.com/ Demographics-Prediction/DPUI.aspx). This application takes a given search phrase and provides the probability that the query is male or female-oriented (i.e., provides a probability for both within range of 0–1 inclusive).

B. Related Query Log Analysis

Hongning Wang [8] studied the problem of user modeling in the search log data and propose a generative model, dp Rank, within a non-parametric Bayesian framework. The dpRank model identifies each individual user's latent search interests and his/her distinct result preferences in a joint manner by postulating generative assumptions about a user's search behaviors. The experimental results on a large-scale news search log data set validate the effectiveness of the proposed approach, not only provides in-depth understanding of a user's search intents but also benefits a variety of personalized applications.

Duarte Torres et al. [17] constructed two sets of search sessions from the AOL search logs: the first with users accessing information aimed at the general public (i.e., content targeted at non child users), and the second with users accessing information that was aimed at children. The set of search sessions in the next step were constructed by using a set of carefully selected urls aimed at children from the Kids and Teens section of the Dmoz Open Directory.

J Hu. et al.[5] made a first approach to predict users' gender and age from their Web browsing behaviors, where the Webpage view information is treated as a hidden variable to propagate demographic information between different users.

Weber and Castillo [18] presented a query logs study on how search differs in users with different demographics. They used demographic information that was derived from the US-census and user profile information to describe search patterns and behaviors for population segments with different demographic characteristics. They also employed an analogous methodology to show that the reading level of the urls clicked by children also varies across demographic features.

C. Browsing Behavior

Liat Antwarg et al.[7] introduced a novel approach to generate an intention prediction model of user interactions with systems. As part of this new approach, they included personal aspects, like user characteristics, that can increase prediction accuracy. The model is automatically trained according to the fixed attributes of user(e.g., demographic data such as age and gender) and the user's sequences of actions in the system.

3. Demographics and User Modeling

A user model should capture the behavior (patterns, goals, interesting topics, etc.) a user shows when interacting with the Web. A user model is defined as a set of information structures designed to represent one or more of the following elements [1]: 1) representation of plans, preferences, goals, tasks, and/or abilities about one or more types of users; 2) representation of relevant common characteristics of users pertaining to specific user subgroups or stereotypes; 3) the classification of a user in one or more of these subgroups or stereotypes; 4) the recording of user behavior; 5) the formation of assumptions about the user based on the interaction history; and/or 6) the generalization of the interaction histories of many users into groups.

User models can be classified according to two main elements: (1) the granularity of the model, where a model can be created for each individual user (content-based modeling) or for clusters of users (collaborative modeling); and (2) the type of task for which the model is going to be used. Frias-Martinez [4] has defined four basic types of tasks: (i) Prediction (P), (ii) Recommendation (R), (iii) Classification (C) and (iv) Filtering (F). Prediction is the capability of anticipating user needs using past user behavior.

There are many ways to create user model [2], One of the processes presented in Fig. 1 is the automatic generation of user models from the interaction data between the users and the system (done by the UM Generation module). Fig. 2 presents the basic steps of this module: 1) data collection; 2) preprocessing; 3) pattern discovery; and 4) validation and interpretation.

There is a tremendous change in the behavior of different demographic attributes. Demographic user modeling is done in according to understand and adapt these behavioral differences. All demographic groups spend the majority of their time on the same popular activities (e.g., social media and e-mail), there are pronounced disparities in how frequently different groups access the particularly relevant categories of health, news, and reference.

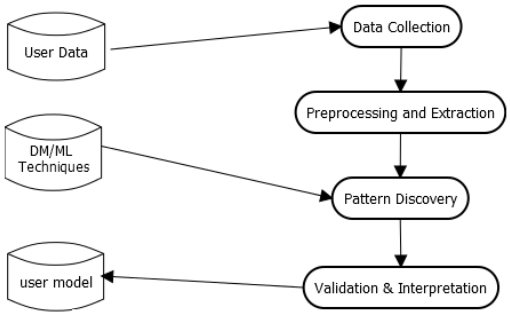

Figure 1: Automatic user modeling [2].

User model heuristics are used in each step of the process to extract, adapt, and present the knowledge in a relevant way. The process of generation of user models using data mining/machine learning techniques can be seen as a standard process of extracting knowledge from data where UM is used as a wrapper for the entire process.

4. Study of Browsing Behavior

This section focuses on how users distribute their online time across different categories of websites, focusing on how these usage patterns relate to an individual's overall Internet activity level [3]. For example, do the heaviest Web users behave qualitatively differently than those who use the Web infrequently, or do they simply visit the same types of sites more often. The following figure depicts the Variation in time spent on popular categories with overall activity.

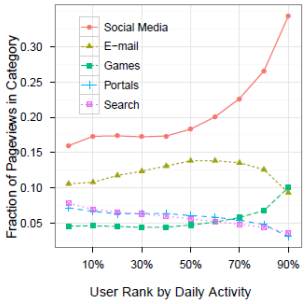

Figure 2: Web Usage by different users [3].

For example, while women spend nearly 30% of their time on social media sites, men spend approximately 20%. Social media usage, in fact, constitutes the most noticeable difference across demographic groups, with older, more educated, male, wealthier, and Asian Internet users spending a smaller fraction of their time on this category. Moreover, lower social media use by these groups is often accompanied by higher e-mail volume, similar to the overall trend noted earlier.

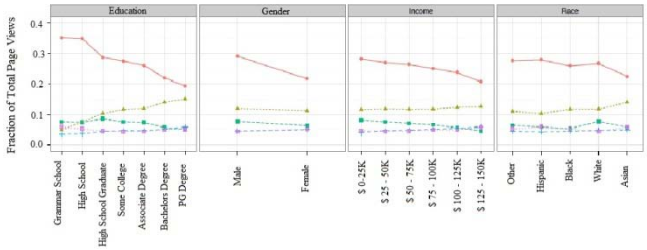

Figure 3: Fraction of time spent on the five most popular categories, split by demographic group [3].

The percentages shown in Figure 3 are normalized within group to adjust for variation in the total amount of time different groups spend online. While Figure 4 focuses on the age based interests of the users and depicts the distribution of topics for informational queries over different age groups.

Study of browsing behavior also emphasizes on the identifying and measuring the search difficulties. User may face difficulties in search for difficulties in meeting the expectations due to reasons like: (1) Short query length, (2) natural language in the query, (3) Click position bias, (4) Click Duration, (5) Click on ads, (6) Accidental Clicks.
To reduce the effects of these factors extended the switch patterns. The definitions are described as follows:

Definition 1 *(Start of the Session → Web Query)*. This pattern occurs when the search query is the first event of the session.
Definition 2 *(Web Query → Web Query)*. This pattern occurs when the event before the search query is a search query.

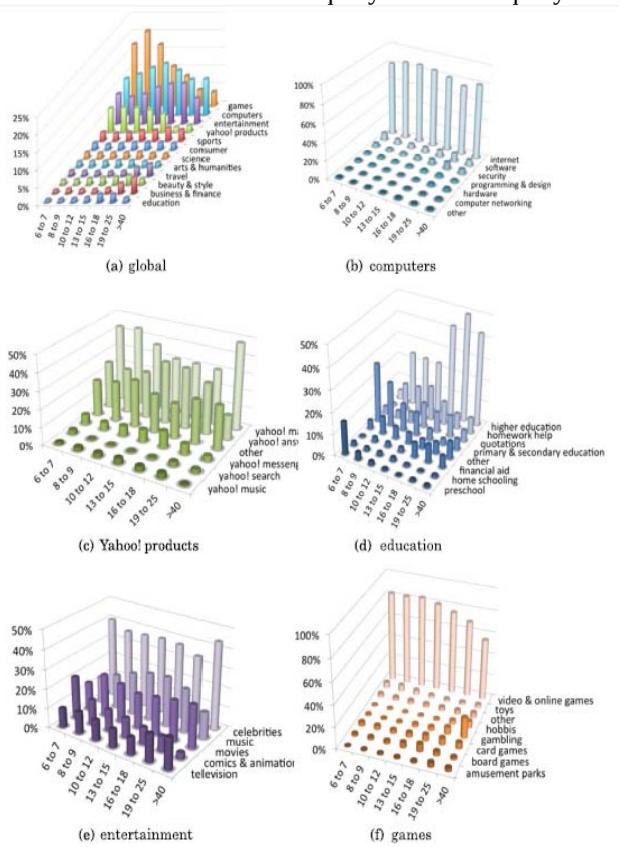

Figure 4: Distribution of topics for informational queries over different age groups [6].

Definition 3 *(Web Result Event → Web Query)*. This pattern occurs when a search result event leads to a new search query event.

Definition 4 *(Browsing Event → Web Query)*. This pattern occurs when a browsing event leads to a query search event.

5. Approaches to User Modeling

In this section we discuss the approaches that are used for effective user modeling. When we deal with the demographic attributes, we need to model uncertainties,

probabilistic estimation etc. An appropriate solution is to use the data mining and machine learning techniques.

As presented in Figure 1, the phase of pattern discovery finds out relevant information about the behavior of a user (or set of users) when interacting with the Web. Data mining and machine learning techniques are ideal for that process because they are designed to represent what have been learned from the input data with a structural representation. This representation stores the knowledge needed to implement the two types of tasks previously described in section II.
Table 1 discusses few applications of user models based on data mining approaches.

Table 1: Examples of some Data Mining based User Models

	Application	**Input Data**	**Results**
Mobasher et. al [10]	Capture of web usage interests by using K-means clustering.	User logs from the university of Minnesota Comp. department server	Example of the implementation of User model in a commercial site
Lin et al. [11]	Learning user preference model of multimedia internet files.	200 records of user email	Average correct acceptance and average correct refusal of each user preference modeling.
Beck et al. [12]	Construction of an user model for an adaptive tutor with J4.8 & Naïve Bayes Classifier.	Data Collected from the interaction of students with the tutor	Naïve Bayes Classifier outperforms J4.8 for individual modeling.
Ruvini [13]	A front – end to Google search engine that uses SVM to infer the user's goal.	Recorded user interactions with Google.	The problem of small screen sizes satisfactorily solved.
Biedal et. Al [14]	Classification & tracking of user navigation	Data generated from online Encyclopedia	A labeled approach to the problem produces better accuracy
Liat Antwarg [7]	Attribute-Driven Hidden Markov Model	obtained from a web application used at Ben-Gurion University (BGU	predicting user search intentions.

Table 2: Comparative Analysis of Data Mining Approaches

	Complexity	Dynamic Modelig	Labeled/ Unlabelled	Noisy Data
K-means	*O(k,m,n,i)*	No	Unlabeled	No
Associat-on Rules	*O(n)*	No	N/A	Yes
Decision Trees	$O(A^2 N)$	Yes	Labeled	Yes
SVM	$O(N^3)$	Yes	Labeled	No
NN	NP complete for 3 & more Layer N/w	Yes	Both	No

Comparative Analysis of data mining approaches is shown in Table 2. The parameters of the comparison are: (1) Complexity, i.e. the computational complexity in offline processing. (2) Dynamic Modeling, which indicates the suitability of the technique to change a user model on-the-fly, (3) Labeled/ Unlabeled, which reflects the need of labeled data, (4) Noisy data, reflects the ability of the technique to handle noisy data.

The complexity for user model using K-means technique is *O(k, m, n, i)* where *n* is number of instances to cluster, *m* is number of attributes, *k* number of clusters, *i* number of iterations, with $i = O(n)$ [10]. Association rule approach has the complexity of *O(n)* [11] where *n* is the number of feature vectors. Complexity using decision trees for single attribute, for multi-way splits on *A* discrete variables and data size of *N* is $O(A^2 N)$ [12]. Whereas the complexity of solving a Quadratic Optimization problem at each iteration using Support Vector machine (SVM) is $O(N^3)$ with *N* as total number of iterations[13]. The problem of User modeling with Neural Networks (NNs) is NP – Complete for 3 or more layered network [14].

Table 3: Techniques Recommended for each combination of decision variables

Task	*Interpretability*	
	Needed	*Not Needed*
Prediction	Decision Trees, Association Rules	Neural N/w, K-means clustering
Recommendation	Decision Trees, Association Rules	Neural N/w,
Classification	Decision Trees.	Neural N/w, SVM, K-means clustering
Filtering	Decision Trees, Association Rules	Neural N/w, K-means clustering

6. Conclusions

This paper has presented a brief review of the state of the art techniques used within the area of user modeling systems. It has also stated the importance of considering demographic attributes in user models. The review demonstrates that one of the main problems that the development of UM faces is the lack of any kind of standardization when the demographic attributes are considered for the design of user models. In order to improve this situation this paper has tried to give a set of guidelines that formalize the design of user models using Data mining and machine learning approaches.

7. Acknowledgment

We express our thanks to publishers, researchers for making their resource available & teachers for their guidance. We also thank the college authority for providing the required infrastructure and support. Last but not the least we would like to extend a heartfelt gratitude to friends and family members for their support.

References

[1] Albrecht D.W. Nicholson A.E. Zukerman, I. predicting users request on the *www.Proceedings of the 7th International Conference on User Modeling, UM99, 275-284.*
[2] Enrique Frias-Martinez, Sherry Y. Chen, and Xiaohui, "Survey of Data Mining Approaches to User modeling for Adaptive Hypermedia" *IEEE Transactions on*

Systems, Man, and Cybernetics—part c: applications and reviews, vol. 36, no. 6, November 2006.

[3] Sharad Goel, Jake M. Hofman, M. Irmak Sirer, "Who Does What on the Web: A Large-scale Study of Browsing Behavior" Copyright 2012, *Association for the Advancement of Artificial Intelligence (www.aaai.org)*

[4] E. Frias-Martinez, G. Magoulas, S. Chen, and R. Macredie. Modeling human behavior in user-adaptive systems: Recent advances using soft computing techniques. *Expert Syst. Appl.*, 29(2):320{329, aug 2005.

[5] Jian Hu, Hua-Jun Zeng, Hua Li, Cheng Niu, and Zheng Chen. Demographic prediction based on user's browsing behavior. In *Proceedings of the 16th International Conference on World Wide Web, WWW '07*, pages 151{160, New York, NY, USA, 2007. ACM.

[6] Sergio Duarte Torres, Ingmar Weber, and Djoerd Hiemstra. Analysis of search and browsing behavior of young users on the web. *ACM Trans. Web*, 8(2):7:1{7:54, March 2014.

[7] L. Antwarg, L. Rokach, and B. Shapira. Attribute-driven hidden markov model trees for intention prediction. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 42(6):1103{1119, Nov 2012.

[8] Hongning Wang, Cheng Xiang Zhai, Feng Liang, Anlei Dong, and Yi Chang. User modeling in search logs via a nonparametric bayesian approach. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, WSDM '14, pages 203{212, New York, NY, USA, 2014. ACM.

[9] Bernard J. Jansen, Kathleen Moore, Stephen Carman, "Evaluating the performance of demographic targeting using gender in sponsored search" Information Processing and Management 49 (2013) 286–302.

[10] B. Mobasher and R. Cooley, "Automatic personalization based on *Web usage mining," Common. ACM*, vol. 43, no. 8, pp. 142–151, Aug. 2000.

[11] F. Lin, L. Wenyin, Z. Chen, H. Zhang, and T. Long, "User modeling for efficient use of multimedia files," in *Proc. Advances in Multimedia Information Processing PCM 2001: 2nd IEEE Pacific Rim Conf. Multimedia*, 2001, pp. 24–26.

[12] J. Beck, P. Jia, J. Sison, and J. Mostow, "Predicting student help-request behavior in an intelligent tutor for reading," *in Proc. 9th Int. Conf. User Model., Lecture Notes in Artificial Intelligence*, vol. 2702, Berlin, Germany: Springer-Verlag, 2003, pp. 303–312.

[13] J. Ruvini, "Adapting to the user's Internet search strategy," presented at the *9th Int. Conf. User Modeling, Lecture Notes in Artificial Intelligence*, vol. 2702, Berlin, Germany: Springer-Verlag, 2003.

[14] S. Bidel, L. Lemoine, F. Piat, T. Artieres, and P. Gallinari, "Statistical machine learning for tracking hypermedia user behavior," presented at the 2nd Workshop Machine Learning, Information Retrieval, and User Modeling, *9th Int. Conf. User Modeling, Pittsburgh, PA*, 2003.

[15] Solomon, P. 1993. Children's information retrieval behavior: *A case analysis of an opac. J. Amer. Soc. Inf. Sci*. 44, 5, 245–264.

[16] Broch, E. 2000. Children's search engines from an information search process perspective. *School Libr. MediaRes.3.* http://www.ala.org/aasl/aaslpubsandjournals/slmrb/slmrcontents/volume32000/childrens.

[17] Duarte Torres, S., Hiemstra, D., and Serdyukov, P. 2010. An analysis of queries intended to search information for children. In *Proceedings of the 3rd Symposium on Information Interaction in Context (IIiX'10). ACM Press*, New York, 235–244.

[18] Weber, I. and Castillo, C. 2010. The demographics of web search. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'10).* ACM Press,