Web User Profiling: Strategies and Challenges

Savita[a] and G Rani[b]
[a]*G D Goenka University, Gurugram, India*
[b]*Manipal University Jaipur, Jaipur, India*
meenu.sahrawat@gmail.com, geetachhikara@gmail.com

Received: *30.04.2019,* ***Accepted:*** *25.05.2019*

Abstract

User profiling plays an important role in the field of recommendation system. User profile identify user's needs and work according to that. User profile is defined as collection of data of user's interest domains. Users are represented through user profiles. This paper gives an overall idea of web user profiling, its methods, challenges, and ways to overcome that challenges, techniques and applications in area of web user profiling.

Keywords- Web store, MagicFG, *CASPER,* Term Frequency

Introduction

User profiling represents every user uniquely. As we know that lots of data is available on internet, it is very difficult for any person to find the information of his relevant use, this create confusing environment for users. So the various machine learning techniques are applied on websites for users in background. To provide the relevant information to the users the process of suggestions or recommendation is the better way. Various machine learning or data mining techniques are used in background like *Bayesian networks,* decision trees, case-based reasoning, association rules and neural networks (Fleuren *et al.*, 2012).

Various applications of user profiling are Search Personalization (Mangest *et al*., 2008), Adaptive Websites, Adaptive Web stores and Customer Relationship Management systems. Adaptive Websites, Adaptive Web stores and Customer Relationship Management systems, for recommending research papers for researchers, e-Tourism (Mariam *et al.*, 2010) based websites, energy management, recommending job according to the qualification and experience by CASPER (Case-Based Profiling for Electronic Recruitment) (*Bradley et al.*, 2010).

Various information of the user is gathered for the purpose of obtaining the profile and important information is taken out from the data collection. Various techniques are used for making profiles of the users from raw data. In the background section we present the background for user profiles or web user profiling where we describe different methods, techniques used previously in this area, applications and, In section material and methods, we describe user profiling methods, techniques and challenges faced in user profiling and solution to overcome from that challenge. In section related work, we describe related work of user profiling. In last section application and conclusion of web user profiling have been described.

Background

A user profile is defined as information which tells about user via user related rules, settings, needs, interests, behaviors and preferences (Araniti *et al.,* 2003; Kuflik and Shoval, 2000; Martin-Bautista *et al.,* 2002; ETSI., 2005; Henczel, 2004). User profile is defined as collection of data of user's interest domains. Users are represented through user profiles uniquely. First step of the user profiling is to collect the information of users. Various types of information can be stored in user profile. An example of e-commerce 1) Personal information- which consists information like city, country, age etc. such type of information can be taken during sign up page of any shopping website. 2) Interests -It consists hobbies, or news related topics. Such types of information can be taken through purchasing or browsing history of the user. 3) Behavior-It can be received implicitly or dynamically; patterns are detected in it.

Information can be gathered either by explicitly (static method or factual method) or implicit method. Explicit method is the method, in which profiles of the users are created manually in which information of user is directly taken by the system through questionnaires or rating methods whereas in implicit method, information is taken by the user browsing behavior dynamically, according to user's need. There are three types of user's profiles 1) Explicit User Profiling 2) Implicit User Profiling 3) Hybrid User Profiling. In explicit user profiling, profiles of the users are created manually in which information of user is directly taken by the system through questionnaires or rating methods. In implicit user profiling method, information is taken by the user through browsing behavior dynamically, according to user requirement. Machine learning techniques are used in it. In hybrid user profiling that is combination of both implicit user profiling and explicit user profiling. It takes static as well as behavioral.

Methods for User Profiling

Through survey two main challenges were found in user profiling which were 1) user profiles of the new user 2) Updating the information of the profile according to the need of the user. To handle these challenges two methods of user profiles have been described below which are content-based method and the collaborative based methods (Fleuren, 2012; Godoy and Amandi, 2005). In collaboration method if you are buying any item, other user's action will be seen for the same like when someone buy bread then the system starts recommending him to buy a butter (reason behind it is, many people who buy bread are also buying butter and not the reason that they both items are related.

Two types of techniques are used in collaborative based filtering which are 1) Memory-based Technique and 2) Model based techniques are used. Problem with collaborative based method 1) Sparsity: When new item enters in database, that item does not have rating, so prediction of such type of item is considered to be poor 2) First-rater: For new users poor recommendations are made, until you have more ratings in their profiles to make comparisons (Khosrowpour, 2005).Whereas in content based filtering, the pre-defined attributes of the products are matched and similar types of products will be recommended. For Example- If a user purchase a Camera, immediately system will starts recommending lenses and other similar model of camera. Various techniques like vector space model, latent semantic indexing, learning, information agents, neural network agents techniques are used in content based filtering.

Related Works

1. User Profiling for University Recommender System Using Automatic Information Retrieval (Kanojea et al., 2015).

In this paper author focus to develop a user profiling system for recommendation of various colleges by focusing on finding, extracting, integrating the information from the web and recommendation system, suggests the colleges of user's choice. The proposed work has been done by using three step 1) Profile extractions which consists institute profile and user profile –It means extracting the useful information from different sources.

In the institute profiling three steps occurs which are web page scrapping, pre-processing and feature extraction using DOM parsing. In web page scrapping, from the seed URL author gets web pages list, from this list author find out the page attributes of his interest. In pre-processing after getting URLs author scrap each page and this page stored as DOM (Document Page Model) which store pages in tree structure format. For the feature extraction, author use condition random filed method, if condition is true information will be stored in dataset. In user profile, the complete information of user is finding out during he registers into the system, that information used for the purpose of knowledge discovery and author make use of social networking website to extract the required attributes. 2) Profile Integration- Data is merged and some attributed removed so that complexity of processing can be reduced. 3) Knowledge gathering- Based on profile attributes user's interest analysis is done. Author may use weights for different criteria such as placement, infrastructure etc. through which rank of the college changes according to the weight assigned, before calculating the rank author must normalize the criteria value.

After performing the profiling, datasets have been collected of various colleges. Experimental work has been done on 116 US Universities, 511 India Universities and 255 Institute of Maharashtra Engineering College. User's interest is analyzed and calculated the rank of the college with respect to weights assigned for different criteria's.

2. Web User Profiling Using Data Redundancy (Xiaotao et al., 2016).

In this paper author has designed an approach MagicFG, for the purpose of extracting the profile attributes from web by making use of big data. To remove errors, approach processes the entire subtask in one modified model. To get rid of redundancy, approach incorporates human knowledge as first order logic and combines logic into extraction model. Main aim of the proposed approach is to design a method which automatically extracts the profile attributes from web. In traditional methods author noticed the problem of finding relevant pages from web and applying model like SVM to extract profile attributes from web pages. Due to this reason, required profile attributes may distribute in web pages which lead to reason of extraction from distributed pages and extraction with data redundancy. To solve this problem, query is constructed and by making use of search engine, to get relevant snippets with query. To solve this problem of ranking the identified information, author proposed MAkov logic graph to get rid of redundancy problem. Profile attributes are of two types categorical- like gender and non categorical like email id and age. Author constructs the query for both types of attributes.

For the categorical attribute query is generated by identifying representative keyword in each candidate keyword and combine them together as the query. Top 10 snippets are obtained, author identify the most representative keywords with highest TF-IDF score. For non-categorical directly keyword used in the attributes name to generate the query. For the extraction of profile attributes two baseline models are there 1) Rule based-A rule-based approach uses rules of thumb or heuristics to determine sentiments. For decision trees, for example It make use of 1) if 2) than and 3) and. 2) Classification based-LR (Logistic Regression) is used and it consists learning and extraction stages. In learning author find the optimal

weight configuration to maximize the log likelihood function of observed instances. In extraction author we see which information we need to extract. It can adjust the weight of different features and combining them to achieve better performance. In MagicFG model, author model the correlation as the first order logic and to leverage logics to improve extraction performance which remove redundancy problem but it was not there in rule and classification-based model which ignores correlation among candidate instance which were seen in returned snippets. Now author introduce how to model data redundancy for both types of attributes.

For modeling non categorical attributes two types of functions are used. 1) Attribute factor function which catches character tics of email-person pair. 2) Complete Consistency-It catches correlation between latent variables. There are three types of first order relationship between latent variable 1) Complete consistency-two latent variables value in any of the condition like should be consistent Like Equals (ei, ej)= Equals (yi, yj). 2) Partial Consistency-Value of the two latent variables is partially consistent in any condition. 3) Prior Knowledge-Describe prior knowledge which can be formalized into first order logic for a specific task. Modelling categorical attributes-one factor graph is built with each node, which represents a query person for example -query is there which add person name as well as gender. Query look like "name his! Her" then approach is used on return snippets. For non categorical attributes, multiple graph is built, each graph is built for each person whereas in categorical attribute only one graph is built.

The effect of proposed approach is seen in both categorical and non categorical attributes by an example of gender and email. To construct a ground –truth dataset, authors take 2,000 researchers from AMiner.org. After extracting author found 34 % of researchers are female and 40% emails are correct. To evaluate the model dataset divided into training and test dataset. Author compare the MagicFG with following methods in terms of precision, recall and F1-score 1) Rule 2) SVM 3) RF 4) LR for extracting gender and emails. Proposed approach shows best extraction model namely FR (+2.12% in terms of F1 score) and in gender inference model perform best method LR (+2.12% in terms of F1 score). In all experiments we set L=10 and search top 10 results by google search. MagicFG approach is compared with various several state-of-art methods TCRF and FGNL. TCRF method make use of only two steps which are 1) Find out user's home page and 2) Extract email from home page with high precision using model TCRF. Results are much better because TCRF choose only home page as data source which ignore useful information on web and in proposed approach query construction will not ignore useful information. So, proposed approach much better precision for email extraction. For gender inference, FGNL (Facebook Generated Name List Predictor) is used as baseline. Most states of the art methods depend on the list of common names of male and females. In an approach high quality name list data can be found from Facebook. They match the user name with the list, it gets match with male or female name, will be treated as the same. Proposed approach performs much better than FGNL in recall because FGNL totally dependent on name list but FGNL performs slightly better in precision which tells advantage of using name list. Proposed approach tested on two real datasets and found improvement in profiling accuracy as compare to others methods.

3. Web User Profiling Based on Browsing Behavior Analysis (Xiao-Xi Fan, 2014).

In this paper author describes a model for web user profiling and identifies a user from web browsing history, by taking two features of browsing behavior that are page view number and page view time of a user for each domain. Four models obtained and compared that which are based on term frequency and

term frequency-inverse document frequency. Methodology steps involve three steps that are 1) Data extraction 2) Data pre-processing and 3) Vector representation. In data extraction, History records are extracted from all the browsers of each computer and combined. Every record consists information like URL, access time etc. If more than one user use, then different profiles are made for each user.

After that data is pre-processed, duplicate records removed with the same URLs and access time, pop up desktop news notice was also removed. Top level and second level domain were extracted from the records. After pre-processing PVN and PVT are calculated. Finally, cosine similarity is applied to calculate similarity score between target and candidate computers. 40 days history is extracted. In vector representation weighted was calculated by using TF and TFIDF over PVN and PVT. Top N domains were chosen and ranked according to their weighting values. Four web user profiles models are TF-PVN model, TFIDF-PVN, TF-PVT and TFIDF-PVT model. 1) In TF (Term Frequency)-PVN (Number of page views) model. This model assign weight to the page view frequency TF-PVN is calculated as number of page views of domain to the total number of page views from computer during observation period. If the domain is frequently used it means that has high TF-PVN. 2) TFIDF-PVN Model-TFIDF-PVN of a domain for the computer consists two sub calculations which are a) How many times domain is visited by the computer and b) Number of computers in the collection visits the domain. IDF is calculated as logarithm of the quotient of total number of computers and number of computers containing the domain. TFIDF-PVN assign high value to domain when it occurs many times in small number of computers and assign value low when domain occurs rarely on many computers or the domain appears on virtually all the computers 3) TF-PVT Model- PVT can be calculated as a difference in the access times between two consecutive pages TF-PVT is the time spent on domain d based on browser history and it can be calculated as the total page view time for all the pages to domain d for a computer to the total page view time for the computer. There is the difference in the weights of TF-PVN and TF-PVT because videos are watched and online shopping is done. TF-PVN has high weights due to this reason. 4) TFIDF-PVT Model- IDF is calculated as logarithm of the quotient of total number of computers and number of computers containing the domain. It enhances the domain weight due to high IDF and domain will get the high weight when it spent most of the time on few computers and low weight is assigned when page view time was short or domain browsed by many computers or domain browsed by many computers. Cosine similarity is measured by checking similarity of two browsing history. Top N domains in target computers have the high similar score and high probability as it used by the same web users as target computer.

Model was tested by taking 51 computers 34 participants. 34 computers for 34 participants one for each participant was assigned to Group 1 and remaining 17 computers were assigned to Group 2. Browsing history of 40 days was extracted from computers and performance of model is calculated as the evaluation matrix which is the proportion of correctly identified examples out of all examples. Target and candidate computers were selected from 51 computers. Three sampling were used (M=15, 33, 51) to check average accuracy as well as weight of domains are computed for target computer. When user identification performed on small group N (Number of domains) =15 from both group and flattens as N increases when N=26 best user identification result occurred 67% and when N=32, For TFIDF-PVN best user identification result was 92% TFIDF is more effective than TF. Each feature remains consistent as the value of M increases. For all feature web user identification accuracy decrease as the size of computer profile increases if high accuracy is needed TF-IDF should be used otherwise TFIDF-PVT is also another option. Author future research plans is to combine PVN and PVT model to get more accuracy in identification.

4. An investigation on Social Network Recommender Systems and Collaborative Filtering Techniques.

In this paper author has implemented and compared two approaches collaborative filtering (CF) and social network recommendation system (SNRS) by making use of mean absolute error (MAE) and accuracy. Collaborative filtering techniques are used to make the decision about the interests of a user by collecting preference of many users. These are of three types 1) Memory based filtering (Determine similarity between item and user) 2) Model based filtering (All users are not used in prediction, instead of that nearest neighbors are used 3) Hybrid based filtering (Combination of model and memory-based model). Probabilistic approach is a new technique in which three factors are taken into the consideration which are user preference, item acceptance, friend inference to predict that a user taste of liking or disliking any item. Main requirement of dataset is 1) Social relationship bond must be there between different individuals 2) Individual ratings given by individuals to different items. 3) Categories to which different items belong to. To generate dataset author, make use of MS excel. Three tables are used to store the required data 1) Relationship Table-It represent the bond between 2 users (x and y) with 100 ×100 entries. The value of x and y lies between 0 to 5. 2) Rating Table-Store the rating between user x to item y with 100 ×10 entries 3) Attribute/category table- It store the attributes each item has with 10 ×10 entries. Disadvantage of collaborative approach are cold start problem and data sparsity.

In implementation of collaborating filtering approach author has used memory-based approach to implement our model where' u' is the active user, 'I' is the item and 'n' is the neighbor of active user. In basic formula of memory-based model, some problems were noticed. If the similarity of the neighbors does not sum up with one then prediction was miss-scaled. Author used new formula to solve the problem of wrong scale, due to usersim coefficient; the result was divided by the sum of all the coefficient. To get exact estimation, average value of rating of users is added because it was subtracted during calculation. Pearson correlation used to measure the correlation between the rating of active user and its neighbor. In implementation of probabilistic approach three factors are taken in consideration which are user preference, item acceptance and friend interference.1) User preference is the probability that a user U will give a rating K to any item I. 2) Item acceptance is the probability that a given item I will get rating K from users providing that item has attributes. 3) Friend interference- This probability is achieved through estimating user similarities either based on user profiles or user rating. To get final probability user U will give rating K to an item I. We multiply three of them and divide by normalizing factor Z can be calculated from training dataset.

We have used two measures to compare two approaches MAE and accuracy. MAE- MAE is the mean of absolute error of the number of observations. Accuracy is the percentage of total number of observations in which the recommended made was exactly same as the actual values. MAE and accuracy values in CF are 0.876 and 35.2% and in SNRS is 0.930 and 33.6%. SNRS has the better efficiency due to the use of user's own preference; item's general acceptance and influence from friends have been taken into consideration.

5. A Web Personalization System Based on User's Interested Domains

In this paper, authors present web recommendation system. The system uses collaborative filtering techniques. The system recommends a list of domains. A user is free to choose web page(s) from any domain. Methods used in recommendation system are K- Mean's algorithm and K-Nearest Neighbor, KNN. System need to collect the training data like interested domains list of different users. Data will be

narrow down into the several clusters and will be put in the knowledge base. Two main steps are 1) cluster the users by using k means algorithms. In k means algorithm first we take mean value, and find nearest number of mean value and put it into the cluster and finally we keep on repeating step 1 and step 2 until we get the same value 2) Find out the new user similar to which cluster (by using k-Nearest Neighbor(k-NN) algorithm.

In the experimental description we have I'= {set of interested domains},P=is a set of web pages and P'=set of recommended ranked list of web pages, which depends on user's selection in I. Knowledge based is prepared after collecting set of I.U= {I1,I2,I3…In} is a two dimensional vector and n=60. k means algorithm is applied on U and select the centroid of a cluster as representative and stored in knowledge base. It saves server time and reduces server traffic. The recommendation system procedure has two steps 1) user's current interested domain is detected by providing list of interested domains 2) Based on user current interested domain system will give to user set of web pages with ranking score P'. If the user is new system will identify user's interest by click history on interested domain list and will update the web pages related to user's click on the same page. Comparison of interested domains recommendation 's accuracy based on random selection and recommendation methods is (73.9% and 62.3%) and web page recommendation accuracy based on interest domain selection and without interested domain is (68.9% and 42.9%). Our system just learns user's interested domains based on click history. In future we need to consider the time spent on each interested domain.

Table 1: Comparison of Literature Survey

Paper Title	Author Name	Purpose	Technique Used	Dataset Used	Tool used Used	Future scope
User Profiling for University Recommender System Using Automatic Information Retrieval (Kanojea *et al.*, 2014)	Sumit Kumar Kanojea, Debajyoti Mukhopadhyaya, Sheetal Girasea	Recommending the colleges to the students of their choice by using user profiling system	Web page scrapping, Pre-processing and Feature Extraction using DOM Parsing	Data collected from various web sources [116 US Universities, 511 India Universities 255 Maharashtra Institute.	Weka	To develop user profiling system for all the Indian Universities or Colleges.
Web User Profiling Using Data Redundancy (Xiaotao *et al.*, 2016)	Xiaotao Gu, Hong Yang, Jie Tangy, Jing Zhang.	Design a method which automatically extract the profile attributes from web.	Approach MagicFG (Markov logic factor graph)	Ground truth dataset was constructed by choosing 2000 researchers randomly.	Imple-mentation done in c++	Author would try to find more better results
Web User Profiling Based on Browsing Behavior Analysis (Xiao-Xi *et al.*, 2015)	Xiao-Xi Fan, Kam-Pui Chow, Fei Xu	Web user identification model which creates a profile of the user based on web browsing activities.	Term Frequency(TF) and Term Frequency – Inverse Document Frequency (TFIDF) and Cosine Similarity Measure.	Browsing history records from July 1, 2013 through August 9, 2013 (40 days) were extracted from each computer from Hongkong university.	Chrome History View and Mozilla History View and IE History Viewv1.70 to extract internet history records from index.dat.	To combine PVN and PVT for more accurate identification model.

Paper Title	Author Name	Purpose	Technique Used	Dataset Used	Tool used Used	Future scope
An investigation on Social Network Recommender Systems and Collaborative Filtering Techniques (Nayebzadeh *et al.*, 2017)	Maryam Nayebzadeh, Akbar Moazzam, Amir Mohammad Saba1, Hadi Abdolrahimpour Elham Shahab	SNRS (Social Network Recommendation System) approach is better than other traditional Collaborative filtering techniques	Collaborative Filtering (CF) and Social Network Recommendation system (SNRS)	Author own dataset with help from yelp.com with requirement of showing Social relationship, individuals rating and category to which different items belong.	MS Excel.	Author will try to overcome the problem of dataset with better results.
A Web Personalization System Based on User's Interested Domains	Minxiao Lie, Lisa Fan.	Helps users in finding out relevant information based on their selection from domain list.	K- Means algorithm and K-Nearest Neighbor.	14 interested domains based on Yahoo directory has been taken with 60 users.	MS excel	Find out the time spent on each interested domain.

Applications of Web User Profiling

User profiling is one of the important parts of recommendation system. It is applicable in the various fields of Search Personalization, Adaptive Websites, Adaptive Web stores and Customer Relationship Management systems, for recommending research papers for researchers, e-Tourism (Kanojea *et al.,* 2015) based websites, energy management, recommending job according to the qualification and experience by CASPER (Case-Based Profiling for Electronic Recruitment) (Xiaotao *et al.,* 2016) and e-governance systems (Rani and Chakraverty, 2012).

Conclusion

User profiles represent each user uniquely according to his behavior and interests. This paper tells about user profiling concepts, techniques, methods pro and cons of web user profiling methods along with its applications. This paper gives an idea of web user profiling with recommendation system. In future author would like to implement various classification and clustering techniques on real world user profiling dataset, so that comparison of various techniques can be done.

References

Araniti, G, Meo, P.D., Iera, A., and Ursino, D. 2003. Adaptive controlling the QoS of multimedia wireless applications through user profiling techniques, *IEEE Journal on selected areas in communication, 21(10), 1546-1556.*

Bradley, et al. 2000. Case-based user profiling for content personalization," In Adaptive Hypermedia and Adaptive Web-Based Systems. *Springer Berlin Heidelberg, 62-72.*

Bedekar, M. et al. 20018. Web Search Personalization by User Profiling. *Emerging Trends in Engineering and Technology ICETET'08, Nagpur India.*

European Telecommunications Standards Institute. 2005. Human Factors (HF); *User Profile Management*, 1- 100, Available: http://www.etsi.org/

Godoy, D., Amandi, A. 2005. User profiling in personal information agents: a survey, *The Knowledge Engineering Review Journal,* 20(4), 329-361.

Henczel, S. 2004. Creating user profiles to improve information quality. *Factiva*, 28(3), 30.

Kanoje, S.K., Girase, S., Mukhopadhyay, D. 2014. User Profiling Trends, Techniques and Applications. *International Journal of Advance Foundation and Research in Computer, 1 (1).* Article retrieved from https://arxiv.org/ftp/arxiv/papers/1503/1503.07474.pdf.

Kanoje, S.K., Mukhopadhyay, D., Girase, S. 2016. User Profiling for University Recommender System using Automatic Information Retrieval. *Procedia Computer Science*, 78: 5-12.

Khosrowpour, M. 2005. Encyclopaedia of information science and technology. *Electron, Book, Hershey, PA Idea Group Reference, 2063-2067.*

Kuflik, T., and Shoval, P. 2000. Generation of user profiles for information filtering-research agenda, *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 313-315.*

Maaike, F. 2012. User Profiling Techniques: A comparative study in the context of e-commerce websites, Nagpur, India.

Martin-Bautista, M.J., Kraft, D.H., Vila, M.A., Chen, J., Cruz, J. 2002. User profiles and fuzzy logic for web retrieval issues. *Soft Computing (Focus), 15(3-4), 365-372.*

Nayebzadeh, M., Moazzam, A., Mohammad, A., Hadi, S., Elham Shahab, A. 2017. Investigation on Social Network Recommender Systems and Collaborative Filtering Techniques. Article retrieved from https://arxiv.org/ftp/arxiv/papers/1708/1708.00417.pdf.

Ouanaim, M. et al. 2010. Dynamic user profiling approach for services discovery in mobile environments. *Proceedings of the 6th IWC MC Conference, 550-554.*

Peterson, G., Shenoi, S. 2014. 10th IFIP International Conference on Digital Forensics (DF). *Springer Heildelberg New York Dordrecht London*.

Rani, G. Chakraverty, S. 2012. Survey of E-Governance Systems with focus on Development Approaches & Interface Quality. *International Journal of Interscience Management Review*, 2(2): 34-42.

Xiaotao, G., Hong, Yang., Jie, Tang*y*., Jing, Zhang. 2016. Web User Profiling using Data Redundancy. *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM),* http://keg.cs.tsinghua.edu.cn/jietang/publications/ASONAM16-Gu-et-al-web-user-profiling.pdf.

Xiao-Xi, F., Kam-Pui, C., Fei, X. 2014. Web User Profiling Based on Browsing Behavior Analysis. 10th IFIP International Conference on Digital Forensics (DF), Vienna, Austria. pp.57-71, ff10.1007/978-3-662-44952-3_5ff. ffhal-01393760.

