RESEARCH ARTICLE OPEN ACCESS

# A Supervised Learning Method for Avoiding the Ambiguity Problem in User Profile Extraction in Web Pages

S.Nagarajan [1], Dr.K.Perumal [2]
Department of Computer Applications
School of Information Technology
Madurai Kamaraj University
Madurai - India

**ABSTRACT**
In this paper focus on user profiling, which is aimed at identifying, extracting and merging the semantic based user profile from the web. Web user profiling was often done by creating a list of keywords for the user, which is not sufficient for main applications. In this paper the profiling problem are formalized as several subtasks: profile extraction, integration, and user's interest discovery. Introduced an approach to work with the task of profiling. To deal with the name ambiguity problem while integrating profile information extracted user profiles, and constructing user interest model. The implementation results on the online system reflect that the combined approach. The joint approach to various profiling tasks outperforms several methods.

***Keywords:-*** Information Extraction,Imformation Retrieval,Text Mining,Data Mining

## I. INTRODUCTION

Web user profiling is the process of acquiring values of different user properties that makes the user model. To mine the user's interest from their historical data, considerable efforts are being made. User's interest can be represented by creating a list of relevant keywords but this is not sufficient for modeling and understanding user's behaviors. To provide high quality web service a complete user profile along with the educational background, experience and contact information is very important.

This user profiling helps the online advertising firms to target more customers based in the current position rather than focusing only on their interest. Formerly, user profiling is considered as an engineering issue and was done manually or conducted separately in an ad-hoc manner. In web based social networking site such as Facebook and MySpace, certain user profiles would be incomplete, just because they do not wish to fill those details.

There may also exists some profiles with inconsistent and irrelevant details. User profiling can also be done by using the list of keyword generated using statistical methods. For example, discovering most often used words from the user entered information. However, in this composition few semantic information like affiliation and location are ignored. Works are been conducted to automate the process of building the used profile by data extraction technology. The proposed work use predefined rules or specific machine learning models to extract the various types of profile data in a distributed pattern. The profile information in user's related documents are retrieved from web page. In rest of the paper Section 2 formalizes the problem of the Web User Profiling. The Overview of the approach in Section 3, process of profile extraction in Section 4 and Section 5 deals with name ambiguity while bringing together the extracted profiles. In Section 6 the experiment results are presented and Section 7 gives the demonstration system.

## II. RELATED WORKS

Based on the techniques, concepts and methods, information filtering is done.

### *A. User profiling with online profiles*

User profiling is a field of AI that is focused to gather information about the user and then using the information to adapt a system to the user. The aim of user profiling is to filter information based on its relevance for the user. In this paper , we focus on the extraction of user models in the form of terms that represents the user's online content.

Lops et al. [8] introduce a paper recommendation system based on the entities provided by the LinkedIn profiles such as "Interest", "Groups", "Associations', etc..In the network, vectors of adjacent users are added to the user's vector. The similarity between the paper's vector and user's vector is then calculated to recommend the appropriate users using recommendation engine.

Abel et al. [6] introduces almost similar approach where Twitter data along with the URLs content from the tweet of the user is used instead of LinkedIn data. Additionally, entity recognition is used to enrich the user model. In vector space, the user model is again represented as the articles are recommended to the user. Burke [3] finds the interests of researchers using probabilistic topic modeling. To analyze the terms in large bodies of text and how they are interconnected, the above method relies on statistical methods.

***B Collaborative filtering***

Collaborative filtering is the one possible approach to enrich the sparse datasets. In order to generate the user specific recommendations, this technique makes use of the activity of other users. On the basis of common interests, user relations are formed which makes it possible to enrich the user profile. Kostoff et al. [6] extracted terms from cited papers to describe the topics of the cited paper. To detect the general theme of an article, this kind of trans-citation is proved to be very successful.

## III. PROBLEM DEFINITION

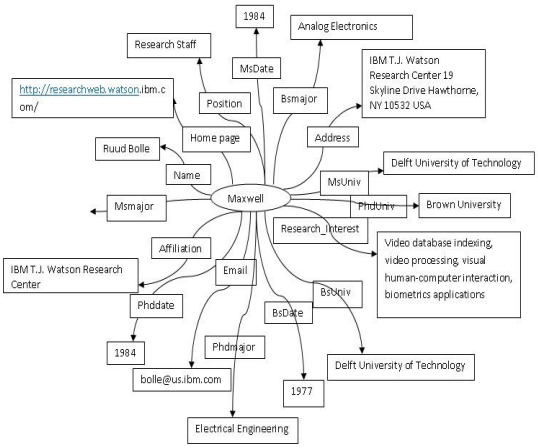
The profile schemas may vary for different applications. this paper uses the researcher's social networking site profile for explanation. It describe three key issues: profile extraction, user interest and name disambiguation.

In this paper uses data from LinkedIn system for study. This system provides a social networking platform for students and professional. More than 300 million user profiles exist in LinkedIn. For name ambiguity problem (different profile with same name), an examination on 200 randomly selected profile name was made and found that more the 40% of the name have the ambiguity problem.

## IV. OVERVIEW OF THE APPROACH

In this paper, focus on user's information based on its relevance for the user. Using collaborative filtering enrich the sparse dataset, which is extracted profile information.

| Network | Subject | Data Field |
|---|---|---|
| LinkedIn | User | profile.{industry, headline, summary, specialties, interests, skills, educations, positions } |
| LinkedIn | Connections of user | profile.{industry, headline, summary, specialties, positions} |




### Schema of User Profile

To retrieve the information required for composing APIs provided by LinkedIn were used. Apart from the required profile, the users connected the profile are also extracted.

User Profiling for Semantic Web Semantic web paves a way for the machines to interact by explicitly providing semantic data, like higher level information about the page, for example functionality of a web service or topic category of the page content. User profiling is required when enhance the interaction between the machines. Semantic web is about data that are self-explanatory and about the data that are annotated. These data about data enables efficient communication among computers for the purpose of building better services for the end user. Since data, that cannot be explicitly annotated, in most cases can be understood in more than a way, the possible source of annotation may come from the information about the user, which is represented in different ways. User profiling generalizes this collected abstract and aggregate information about the user behavior. The user profile is further used for annotating the data in the way that the web services are able to provide personalized information which increases the user efficiency when communicating with computer.

Therefore, user profiling is a significant source of meta-data about user perspective in relation with the data understood. The focus of this paper is to increase the efficiency of user activity by delivering more personalized information.

***A.Profile Extraction:***

An automated web crawler is developed. The crawler would visit profile pages based on a randomly generated list of id numbers using the Rand function. To collect relevant bit of data regular expressions are used. The collected data are analyzed for further finding .The data are organized hierarchically. It is categorized as basic information containing name, photo, email id and address, educational history and the professional information.

***B.User interest***

Rather than directly extracting the user's interest the extracted profile information is analyzed to discover the user's interest. Each user may have a set of interest which is represented by a mixture of words and their probability may differ. The user interest is defined on the basis of hobbies. Each hobby is defines as $c=\{(a_1,p(a_1|c)),\ldots,(a_{N1},p(a_{N1}|c))\}$. The definition means that a hobby is represented by a mixture of words. Therefore, the interest of the user x is defined as a set of hobby distribution $\{P(c|x)\}_c$.

Name disambiguation:

The process of extraction does not happen in a homepage but the data from the existing online source are also integrated.

A formal definition of the name disambiguation task is, given a person name x, all profile name having x is denoted by Q = $\{q_1,q_2,\ldots, q_n\}$. For each y profile names $\{b_i^{(0)},b_i^{(1)},\ldots b_i^{(y)}\}$, the profile name that are going to disambiguate as the principle profile. If there are ju actual profiles having the name x, our task is to identify the actual person of the profile $y_h$ ,h Є[1,k].

## V. WEB DATA PROCESSING

***A. Data Sources***

There are different types of data that are available in the web pages. The types are A) Data taken from web logs B) Content data .

***B. Data taken from web logs***

Web servers usually maintain web logs which comprises of information about the users accessing sites. Logs are as simple as text files, each line corresponds to one access. Common Log File Format(CLF) and the Extended Common Log File Format(ECLF) are the two formats which are mostly used in log files. The information in the web log contains 1) IP address of the user 2) Authentication name of the user 3) Date and time stamp of the access 4) HTTP request 5) Response status 6) Requested resource size 7) Referrer URL 8) Browser identification of the user. Some sites does not require authentication, in that case the authentication name of the user will not be available in the log file.

***C. Content data***

All the contents that are accessed by the user refers to content data.

Here , the contents refers not only to the text data but also to other multimedia elements and images.

***D. Data Preprocessing***

Data preprocessing is the first step in profile extraction process. Data preparation phase is done before preprocessing. Thus data preparation phase is divided in to two categories. They are 1) Web access data preparation 2) Content data preparation.

***E. Web access data preparation***

The outcome of the data preparation of web access is huge and it goes well with the collaborative filtering methods.

***F. Data cleaning***

All access to the content need not be taken in to consideration. Data cleaning is the process of removing access to irrelevant items , access made by web crawlers , and failed requests.

***E. Efficient User Identification***

Because of the reasons, many users can have same IP address and one user can have different IP addresses, user identification using IP address is a poor idea.

The first problem is due to the side effect of intermediary proxy devices and local network gateways. When ISP is performing load balancing over several networks, second problem arises. "Browser" and "Referrer" fields contain information which are used to distinguish between users having same IP . But this does not achieve complete distinction. For better identification, cookies can be used. Assigning usernames and passwords to user is good user identification.

## VI. PROFILE EXTRACTION

The process of profile extraction comprises of three steps namely relevant page finding, preprocessing and tagging. In the first step say relevant page finding, once given a user name, will get a list of pages displayed by the search engine. Using binary classifier, home page or introducing

page will be identified. Defining features such as whether the URL address contains the user name or whether the page title contains the user name is done using Support Vector Machines (SVM) as the classification model.

In the second step say preprocessing, it will first segment the text in to tokens. Then, possible tags will be assigned to each token. The tokens in turn form the basic units and the pages form the tree structure of units or sequence of units in the tagging problem. In tagging, a sequence of units or tree structure of units will be given, and then it determine the most likely corresponding tags using a trained tagging model. Each tag corresponds to a property.

In the first step, the tokens in the web page are identified heuristically. There are five types of tokens defined here. They are 'standard word', 'special word', '<image>'token, term and punctuation mark. Standard words represent unigram words in natural language. Special words include IP address, date, number, unnecessary tokens, email address, percentage, URL, words containing special symbols, etc. Regular expressions are used to identify special words. `< image >' tokens are `< image >' tags in the HTML file. They are identified by parsing the HTML file. Terms are base noun phrases extracted from the web pages. Punctuation marks include exclamation mark, question and period.

In the second step, based on the type of each token, tags will be assigned. In case of standard word, assign all possible tags. In case of special word, the tags assigned will be Position, Affiliation, Email, Address, Phone, Fax, and Bsdate, Msdate , and Phddate. For `< image >' token, Photo and Email tags will be assigned. Position, Affiliation, Address, Bsmajor, Msmajor, Phdmajor, Bsuniv, Msuniv, and Phduniv tags will be assigned to the term token. By this way, each token can be assigned several possible tags. Using the tags, it can perform extraction of 16 profile properties, which cover 95.71% of the property values on the Web pages.

### *A. Extraction Model using Conditional Random Fields*

The tagging model used here is Conditional Random Fields (CRF). It is a special case of MRF. Given a sequence of observations tokens, CRF is a conditional probability of a sequence of labels y. However , the previous linear chain CRF is not able to model the hierarchial dependencies . It can only model the hierarchically laid-out information.

#### *1) Linear chain CRFs*

Conditional Random Fields are undirected graphical models. Here, X is a random variable over sequences of data to be labeled and Y is a random variable over corresponding label sequences. All components $Y_i$ of Y, are assumed to range over a finite label alphabet Y. CRFs construct a conditional model P(Y|X) with a given set of features from paired observations and label sequences. A CRF is a random field globally conditioned on the observation X.

By the fundamental theorem of random fields , the conditional distribution of the labels y given the observation of data x has the form

$$P(y|x) = \frac{1}{Z(x)}\left(\sum_{e\in E,j} \lambda_j t_j(e, y|_e, x) + \sum_{v\in V,k} \mu_k s_k(v, y|_v, x)\right)$$

*2) Tree -structured Conditional Random Fields(TCRFs)*

In hierarchically laid- out information, Linear-chain CRFs cannot model dependencies.Here, Parent vertex represents upper level vertex and child vertex represents lower level vertex.Tree-structured Conditional Random Fields (TCRFs) can model the parent child dependencies. In TCRFs, vertices which are in the same level are represented as sibling dependencies.

Here, X is used to denote the random variable over observations; Y is used to denote the corresponding labels. $Y_i$ is a component of Y at the vertex i. In TCRFs, considering one or two vertices as a clique. It can be viewed as a finite-state model.

Each variable $Y_i$ has a finite set of values. One-to-one mapping is assumed between the states and labels. Thus, dependencies between components $Y_i$ can be viewed as a transition between states.

Both the linear chain CRFs and TCRFs have the same form except one difference. In TCRFs the edges include parent-child edges, child-parent edges and sibling-vertices edges, whereas, in linear CRFs, the edges represent the transition from the previous state to current state.

$$P(y|x) = \frac{1}{Z(x)} exp\left(\sum_{c\in C,j} \lambda_j t_j(C, y|_c, x) + \sum_{v\in V,k} \mu_k s_k(v, y|_v, x)\right)$$

For Example, in user profile extraction, the observation x in TCRFs to the identified homepage/introduction page. By converting the web page in to a DOM tree, the tree form will be obtained. In the tree structure, the root node represents the web page, leaf node represents the word token and the inner node represents the

coarse information block. Coarse information block is the block containing the contact information. The label y of the inner node corresponds to one type of coarse information whereas the label y of the leaf node corresponds one of the profile properties.

***B. Name disambiguation:***

After analyzing the extracted user data from the online sources, for integration using the user's name. The method has name ambiguity problem.

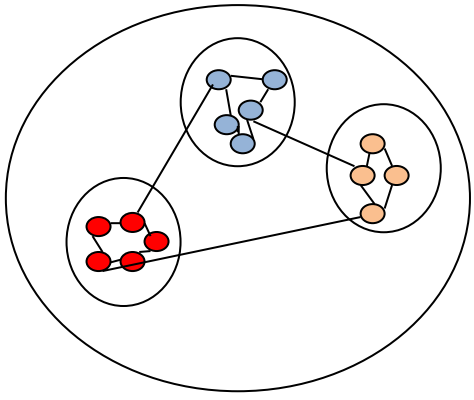
The objective of name disambiguation is to disambiguate n profiles $P=\{p_1,p_2,p_3,\ldots,p_n\}$ that contains the name j to k actual names $\{y_1,y_2,\ldots,y_k\}$ with respect the name x. To deal with this problem a probabilistic model is proposed. Profiles with same email id have same name is our intuition.

Using Hidden Markov Random Field (HMRF) method to determine the dependencies between the extracted data. The main function in the HMRF model is a posterior probability of variables that are hidden in the observation, which are used as criteria for selection

***C. User interest***

So far the basic details in a profile are extracted. User interest is predicted by studying the extracted details. For example if the user's hobby is related to book and he is a member of a community or group named 'Book worms' it could predict that the user's interest may be books.

Certain times the user's hobby mentioned in the social network site may be a combination of many terms. Therefore the most probable result is mined through and suggested. In this paper identifying the user interest and suggest a related community for the user.



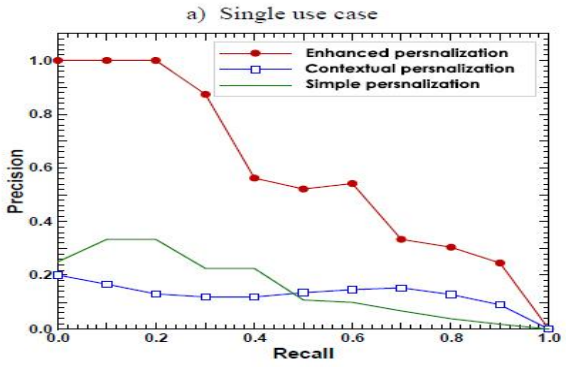
Structure of Social Network Community

A smaller compressed group within a social network is a Community. In a social network, users with similar interest form a community. Applying the apt tools to identify and understand the behavior if the community is crucial. Diverse clustering techniques are used to detect community on a network. The most commonly used technique is hierarchical clustering. This technique is a combined approach of many techniques that are used to group nodes to reveal the strength of individual group. Structural equivalence measures of clustering focus on common connections shared by two nodes. Two users on a network with many mutual friends are closer than two users with few mutual friends on the network.

## VII. EXPERIMENTAL RESULTS

This method was tested with social networking sites. The proposed system was uncomplicated to understand and it is implemented with the help of Dotnet and run on desktop PC with 3.09 GHz Intel and 1.09 GB RAM. The output is more accurate and thus it can significantly improve the performance of inferring user search goal.

To analyze the algorithms in terms of satisfying the user, they design a social networking site. The data extraction happens in both the social network site . The user's personal details are retrieved from Social Networking Sites. Using this API only the basic user details are extracted. The profile details from the social network site designed is also extracted. These are combined to make a user profile. For integrating the profile the user name is checked for disambiguation. User interest is one of the important data for profiling which is determined by analyzing the extracted data and looking for the community the user is in. The community based on the determined result is suggested for the user to join.



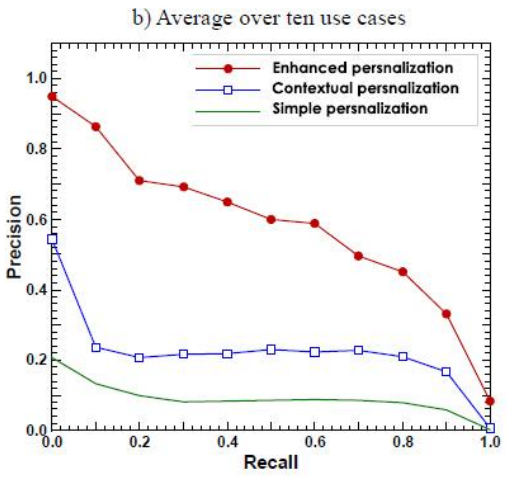



Fig1: Comparative performance of personalized search with various contextualization, showing the precision vs. recall curve for a) one of the scenarios, and b) the average over 10 scenarios.

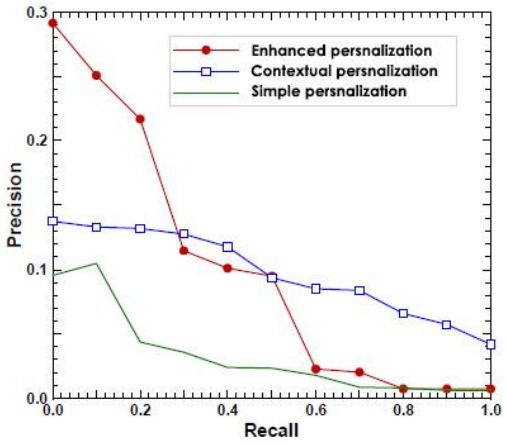



Fig2a: The precision vs. recall curve

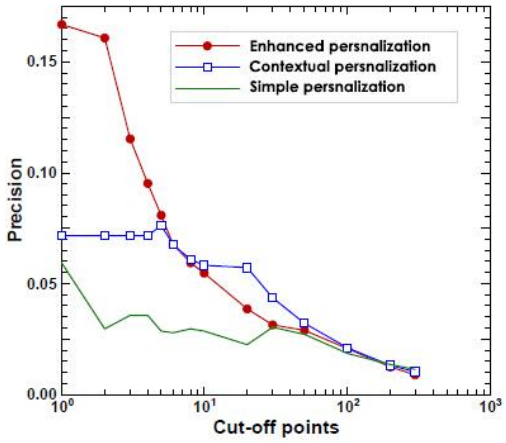



Fig 2b: The precision at cut-off points

Comparative performance of personalized search with various contextualization tested with 18 subjects on three proposed tasks. The graphics show in fig 2a, the precision vs. recall curve, and in fig 2b the precision at cut-off points. The results are averaged over the set of all users and tasks.

## VIII. CONCLUSION AND FUTURE WORK

Semantic annotations and meta-data for the data that are used for different application are provided by semantic web .User profile serves as a source of annotation where semantic is not sensible. User profiles are built by data mining methods. Therefore these user profiles can be used in situations like auto document filling, text tagging and information extraction which will be time consuming.

Recent relevant work includes mining the extracted data, where information extraction is implemented to automatically to collect details about companies Then data mining methods are used to extract data. Since web documents contain hyperlinks arranged in a graph structure, research efforts are made using the graph structure to enhance document categorization. Implementation of these methods in web environment is a matter of future research.

## REFERENCES

[1] Aggarwal, C.: *An introduction to social network data analytics.*Springer US, 2011.

[2] Bakshy, E., Hofman, J. M., Mason, W. A., Watts, D. J.:Identifying influencers on twitter. In *Fourth ACM International Conference on Web Seach and Data Mining (WSDM)*, 2011.

[3] Burke, R.: Hybrid recommender systems: Survey and experiments. *User Modelling and User-Adapted Interaction*,12(4):331–370, 2002

[4] Papadopoulos, S., Kompatsiaris, Y., Vakali, A., Spyridonos, P.:Community detection in social media. *Data Mining and Knowledge Discovery*, *24*(3), 515-554, 2012.

[5] Ahmad Abdel-Hafez and Yue Xu. A survey of user modelling in social media websites. Computer and Information Science, 6(4):p59, 2013.

[6] Fabian Abel, Qi Gao, Geert-Jan Houben, and Ke Tao. Analyzing user modeling on twitter for personalized news recommendations. User Modeling, Adaption and Personalization, pages 1–12, 2011.

[7] Fabian Abel, Qi Gao, Geert-Jan Houben, and Ke Tao. Semantic enrichment of twitter posts for user profile construction on the social web. In Proceedings of the 8th extended semantic web conference on The semanic web: research and applications - Volume Part II, ESWC'11, pages 375–389, Berlin,Heidelberg, 2011. Springer-Verlag.

[8] Pasquale Lops, Marco de Gemmis, Giovanni Semeraro, Fedelucio Narducci, and Cataldo Musto. Leveraging the linkedin social network data for extracting content-based user profiles. In Proceedings of the fifth ACM conference on Recommender systems, pages 293–296. ACM, 2011.

[9] Ackerman, M., Billsus, D., Gaffney, S., Hettich, S., Khoo, G., Kim, D.J., Klefstad,R., Lowe, C., Ludeman, A., Muramatsu, J., Omori, K., Pazzani, M.J., Semler, D.,Starr, B., Yap, P., 1997. Learning Probabilistic User Profiles, AI magazine, 47-56, Vol. 18, no. 2.

[10] Armstrong, R., Freitag, D., Joachims, T., Mitchell, T., 1995. WebWatcher: A Learning Apprentice for the World Wide Web, AAAI 1995 Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments, Stanford.

[11] Balabanovic, M., Shoham, Y., 1997. FAB: Content-based collaborative recomender. Communic. ACM 40, 3, 66-72.

[12] Y.H. Wu, Y.C. Chen and Arbee L. P. Chen, Index Structures of User Profiles for Efficient Web Page Filtering Services, *Proceedings of IEEE*

[13] *Conference on Distributed Computing Systems*, p. 644-651, 2000.

[14] [13] F.A. Aniscar and C. Tasso. "ifWeb: a Prototype of User-Model-Base Intelligent Agent for Document Filtering and Navigation in the World Wide Web", *Proc. of the workshop Adaptive Systems and User Modelling on the World Wide Web,6th International Conference on User Modelling UM97*, 1997.

[15] publisher over 7 technical papers in conferences and journals. His area of research includes Data mining and analytics and information retrieval.