Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 78 (2016) 5 – 12

Procedia Computer Science

International Conference on Information Security and Privacy, 11-12 December 2015, Nagpur, INDIA

User Profiling for University Recommender System using Automatic Information Retrieval

Sumitkumar Kanoje[a], Debajyoti Mukhopadhyay[a], Sheetal Girase[a]

[a]*Dept. of Information Technology,MIT Pune -38, India*

Abstract

User Profiling is the process of Extracting, Integrating and Identifying the keyword based information to generate a structured Profile and then visualizing the knowledge out of these findings. User profiling helps personalizing a system to work according to user. Therefore user profiling or personalization is one of the major concepts used for accessing the user relevant information, which can be used to solve the difficult problems of recommender system like classification and ranking of items in accordance with an individual's interest.

In this paper we focus on the problem of user profiling in which we aim at finding, extracting and integrating keyword based information from various web sources to generate a structured profile. Further we do some experiments on the profiled information to generate knowledge out of it.

Peer-review under responsibility of organizing committee of the ICISP2015

*Keywords:*User Profiling; Information Retrieval; Data Mining

1. Introduction

Education is the basic need of every individual. Now days everyone is bothered about choosing a right institution of his choice and interest. These interests of every individuals change according to their lifestyle. Someone might be more interested in the quality of teaching faculty whereas someone might be interested in the head counts of placements and the packages offered by the companies that come to an institution for recruitment drive. Also there might be chances of someone looking for infrastructure and the social life at the institution. So choosing the right institution makes one's life better at the later stages of life.

While making these decisions, individuals have to go through a large number of physical documents and

* Corresponding author. Tel.: +0-000-000-0000 ; fax: +0-000-000-0000 .
*E-mail address:*author@institute.xxx

Peer-review under responsibility of organizing committee of the ICISP2015
doi:10.1016/j.procs.2016.02.002

prospects of the institutions. Still making the right choice becomes a difficult task for them. For easing this task there should be a common portal which would discuss all the required qualities of an institution that the user is interested in. But there is no provision of getting all these data at one place that could be utilized properly by these people. So our task is to develop a user profiling system which will profile all the information about a user and an Institution. This information can be well utilized for the sake of helping various users, who would be spending more of their time on other issues rather than wasting their time on the decision making of choosing right institute. So the solution for the problems of the various users can be well understood as a better user profiling system that would give a good insight of the institutions to these Students or their parents.

As users are less interested in manually filling the large and hectic forms there is a challenge of knowing user in an implicit way[2]. The task of user profiling gets tougher without having latest information about them. Now day's data is available on various platforms where the users are involved in various social activities. If this information is well utilized there would be great help for profiling users into any system[1, 3]. An innovation is necessary in extracting this information from various sources and converting this unstructured information into a structured keyword based profile.

To help such students who are seeking to take admissions into various universities/institutes/colleges, we are proposing a portal where such students can obtain all such relevant information which will be profiled according to their interest.

2. Problem Formulation

The Problem Statement is to develop a User Profiling System for recommendation of various Universities/Institutions/Colleges by extracting, integrating and identifying the keyword based information to generate a structured Profile and then visualizing the knowledge out of these findings.

When humans come across making choice they find it difficult to obtain the most suitable information that is hidden in the deluge of information[6]. When there is mass of content available with us, important questions is raised over its effective use[2].

Recommender systems provide advice to users about items they might be interested in. Users can navigate through large information spaces of product descriptions, news articles or other items with the help recommendations made by such systems[7]. Recommending such items according to user interest involves processing through these large digitized information spaces; profiling this information properly makes it easier for recommendation system to recommend them to users, User Profiling comes into picture in this scenario[8, 9].

Profiling of a web user is the process of obtaining values of different properties that constitute the user model[13]. User profiling is typically either knowledge-based (already known/factual) or behavior-based[13]. A typical user profiling system is aimed at finding, extracting, and integrating the keyword based user profile from the web[15].

Our most important task will be creating a profile for each individual educational organization, which will contain basic information such as name, establishment, accreditation, etc.; contact information such as address, email, and telephone number; Course information such as course name, Intake, and institution review attributes which will be rated by each user. For each educational institute, profile information will be obtained from web pages introducing them. Also we do use the techniques such social profiling[5] to do user profiling. We will also be implementing a search facility easy for retrieval a particular profile.

3. Proposed Work

3.1. Profile Extraction

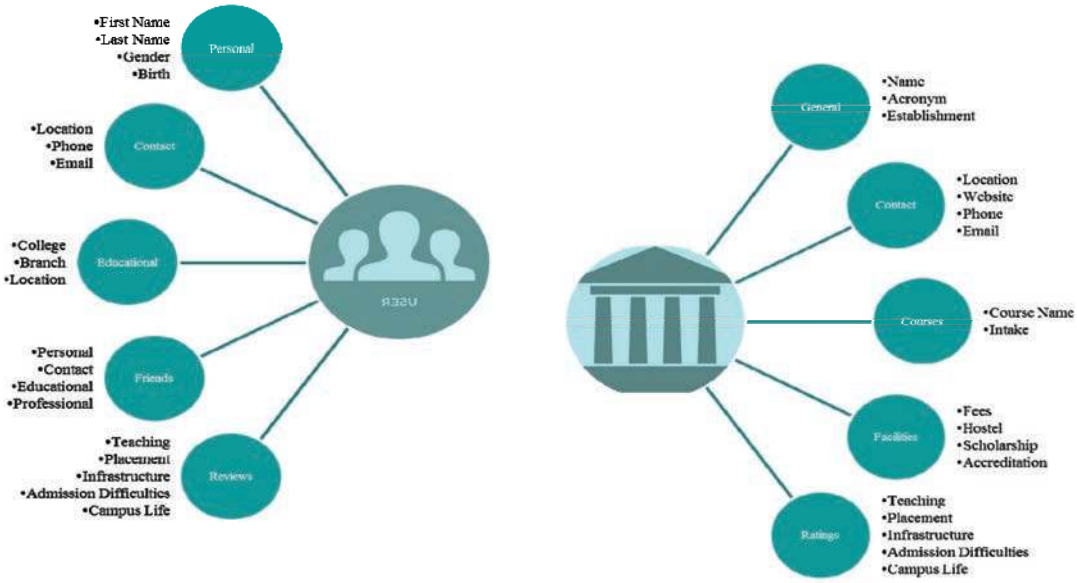
Profile extraction is nothing but extracting the useful information about a user from different sources[13]. In our application context we have to profile all Indian educational institutes and the user. That's why our system has modeled institute profiling and user profiling into two separate parts.

3.1.1. Institute Profile

Identifying different attributes is necessary for institute profiling. Fig.1 below represents the university attributes that will be considered for extracting the university profile. It includes the basic information, contact information, courses information, facilities information and ratings information. This information is used for the knowledge discovery at later stages in the user profiling process.

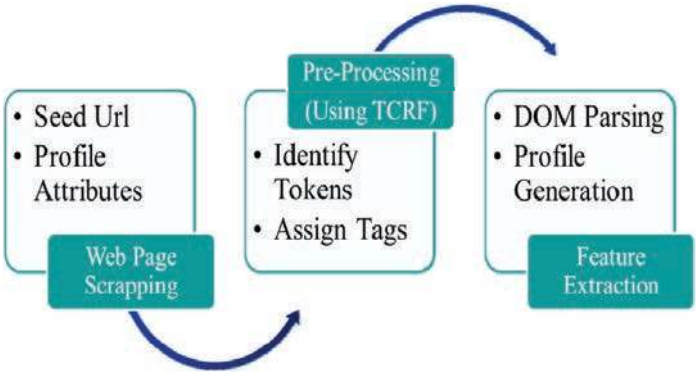
Fig.1. Institute Profile Showing Different attributes of each Institute

Fig.2. User Profile presenting Different attributes of each User/Student



There are three important steps for institute profiling: Web page scrapping, Pre-processing, and Feature Extraction using DOM Parsing. In web page scrapping, given a seed URL, we first get a list of web pages by retrieving all links on this page and then identify the page attributes which are of our interest. We define features such university name, address, pincode, establishment and so on to get these feature parsed from the DOM.

Fig.3. Institution Profile Extraction Process



3.1.1.1. Algorithm for Profile Extraction

In pre-processing, as shown in Fig.3 we identify all URL's which are of our interest on the seed URL and after getting all the URLs from that page, we go on scrapping each page one by one. Now once the individual page is

```
Algorithm: University/College Extraction
INPUT: seed_url
OUTPUT: List of Profile Attributes
Parse(seed_url)
{
  List Attributes;
  urls = select(All urls from
seed_url_page);
  for each url in urls, do
  {
    DOM = Parse(url); List Attribute;
    for each node in DOM
    {
      if (node matches condition)
      {
      Attribute.attr = node.value;
      }
     }
     Attributes.add(attribute);
   }
   return Atrributes;
}
```

scrapped we store this page as DOM i.e. Document Object Model. The DOM stores the page in a tree structured format in which each node of the tree contains some of the data. Now to do feature extraction we use Tree Conditional Random Field method in which each node of the DOM tree is checked by a condition, which finds whether that node contains our required data attribute or not. If the condition is successful it is considered that the node contains our required information and that particular information is then stored into our dataset file.

3.1.2. User Profile

User profile generation is done when we get users complete information while he registers into our system. We have identified different user attributes for profiling him into our system. As Shown above in Fig.2 user profile contains different attributes such as personal information, contact information, educational information and professional information. Each user profile also stores the reviews given by the particular user to an institution or the university. This information is more important as it will be used for the knowledge discovery based on the different user reviews.

The user profiling process is simplified by using the implicit information from user registration process via Facebook[5]. All the information present in the user's social profile is retrieved from his social account and utilized for this purpose.

We make use of social networking website facebook[5] to extract various user attributes implicitly from his facebook account. The process of extracting information from social networking website can be viewed in Fig. 4 System Architecture. We also take reviews about institution from users once they have signed up into our system.

3.2. Profile Integration

We have scraped the complete university & user data from various sources in profile extraction step. For integrating this extracted data we use profile integration. The method inevitably has the many aspects for this problem.

In this phase we combine the data obtained from various sources into meaningful and valuable information. This includes the use of various techniques such as cleansing the data, varying the quality of the data, and also removing some of the data[1]. The removal of some attributes from collected data is done to reduce the complexity of processing such type of data.

3.3. System Architecture

Figure below shows the overall architecture of proposed system.

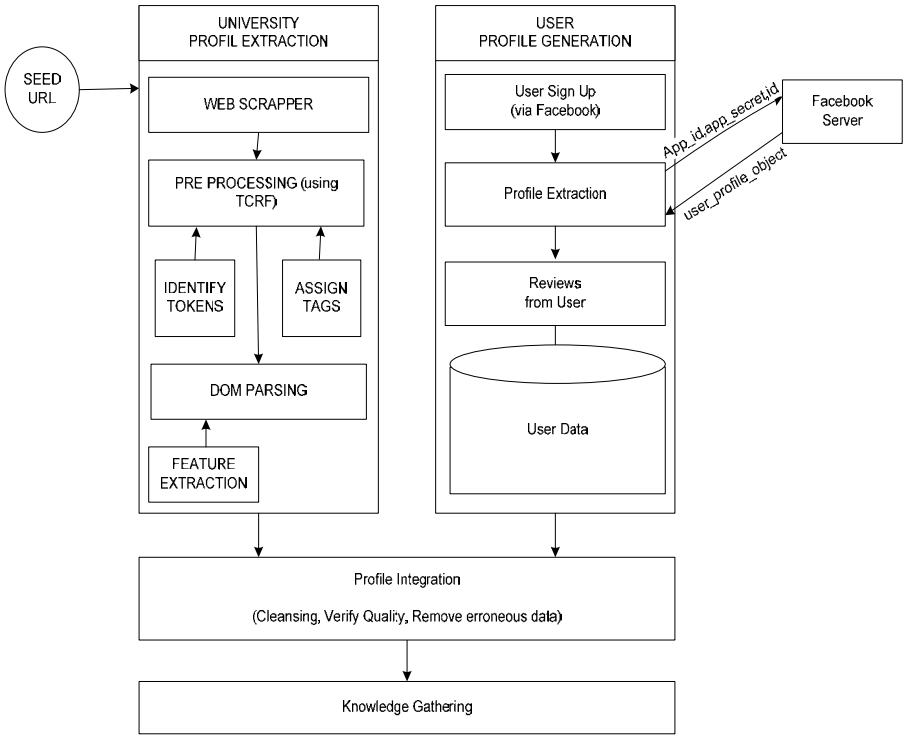


Fig.4. System Architecture

As shown in the figure Institution profiling as well as User Profiling are simultaneous process.

3.4. Knowledge Gathering

After extracting and integrating the information from various data sources, we obtain user profiles which consists a set of profile properties and set of keywords for each profile. Next step is to do user interest analysis based on the profile attributes and the user ratings obtained from users[9].

We already have all the information about universities/institutions/colleges. This information includes its basic information, contact information, courses information and Ratings information. Also we store the admission cut-off scores of each course/institute. We can well utilize this information to find knowledge out of it. This information can be used to find out the user interested institutions based on some information based on the user's profile.

To find out the interested institutes, we use cut off score and find the user interested institutes between the accepted cut off range, but the task doesn't ends here. There might be many results generated based on the cut off score but particularly which institute is more likely to be of users interest is unknown. Finding out this is the most important task.

As far as users concern we can take weights for the different criteria's such as teaching, placements, infrastructure, social life and admission difficulties. When we apply different weights to these criteria's the rank of that institute changes according to the weights assigned to these criteria's. The rank of each institute can be calculated as follows.

$$rank_i = \sum_{k=0}^{n} C_{ki} * W_{ki}$$

Where,

$rank_i$ = Rank of institute *i*

C_{ki} = Criteria k of institute *i*

W_{ki} = Weight for Criteria k of institute *i*

Before calculating the rank of the institute we need to normalise the criteria value of the each criteria that can be achieved by using the following equation

$$normalisation(i) = \frac{(i - old_{min}) * (new_{max} - new_{min})}{(old_{max} - old_{min}) * (new_{min})}$$

Where,

i = old value

old_{max} = old maximum value

old_{min} = old minimum value

new_{max} = new maximum value

new_{min} = new maximum value

4. Experimental Results

After performing the profiling of all the universities, we have successfully collected the datasets of universities/institutions from various web sources. This dataset contains list of universities from India, US, and a list of all Engineering Institutes from Maharashtra State.

4.1. Datasets Collected

The Statistics of the collected datasets is as follows

- **116** US Universities
- **511** India Universities
- Maharashtra Engineering Institutes
  - **255** – Institutes
- All Technical Institutions in Maharashtra

- **173** – Amravati Region
- **232** – Aurangabad Region
- **359** – Mumbai Region
- **529** – Nagpur Region
- **293** – Nashik Region
- **633** – Pune Region

4.2. Experimentations Done on

To find out how the ranking of the colleges changes according to the change in the preferences of the users we have performed experiments on the following datasets

- 116 US Universities
- 511 India Universities
- Maharashtra Engineering College
  - 255–Institutes

4.3. Experimental Results

To analyze interests of the users we have calculated the rank of the each institute with respect to the weights assigned for different criteria's.

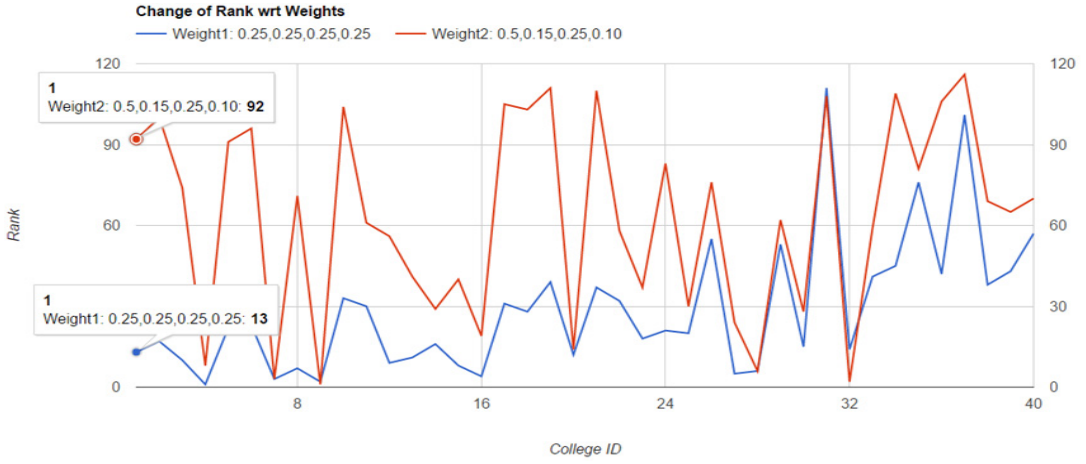


Fig.5. Change in Rank with respect to change in weights of the Criteria's

The above line chart in Fig. 5 shows two lines in blue and red colors, Blue line represents different weights of criteria's given to each of the criteria as 0.25 to Teaching, 0.25 to Placements, 0.25 to Infrastructure and 0.25 to Admission Difficulties whereas Red line presents different weights given to each of the criteria's as 0.5 to Teaching, 0.15 to Placements, 0.25 to Infrastructure and 0.10 to Admission Difficulties

5. Conclusion

There is very less precise and exact data available about Universities/Institutes/Colleges as well as users. This has brought the need of a one stop portal where this information could be placed in a systematic manner and can be accessed by the users for better decision making. So we have used our Profile extraction model to extract data from

various web sources. Also we have profiled users implicitly based on their social networking website. After getting this data we have converted this unstructured data into structured keyword based Profile.

While there was no provision for recommending Universities to users, we have built a User Profiling System for Universities and Users. Once the profile is generated, extracting knowledge out of this data was a big deal. We have managed to find out change in the ranks of the institutes with respect to user interest, which we have analyzed by calculating the ranks of each institute with changing criteria weights.

References

1. Fawcett, Tom, and Foster J. Provost. "Combining Data Mining and Machine Learning for Effective User Profiling." In KDD, pp. 8-13, Portland, Oregon, USA 1996.
2. Rajaraman, Anand, and Jeffrey David Ullman. “Mining of massive datasets”. Cambridge University Press, 2011.
3. McGarry, Ken, Andrew Martin, and Dale Addison. "Data Mining and User Profiling for an E-Commerce System." In Classification and Clustering for Knowledge Discovery, pp. 175-189. Springer Berlin Heidelberg, 2005.
4. Kozmina, Natalija, and Laila Niedrite. "Olap personalization with user-describing profiles." In Perspectives in Business Informatics Research, pp. 188-202. Springer Berlin Heidelberg, 2010.
5. Li, Rui, Shengjie Wang, et.al. "Towards social user profiling: unified and discriminative influence model for inferring home locations." In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 1023-1031. ACM, Beijing, China, 2012
6. Bohté, Sander M., William B. Langdon, and H. L. Poutré. "On Current Technology for Information Filtering and User Profiling in Agent-Based Systems, Part I: A Perspective." CWI, Centre for Mathematics and Computer Science (2000): 1-12..
7. Bedekar, Mangesh, et.al. "Web Search Personalization by User Profiling." Emerging Trends in Engineering and Technology ICETET'08, Nagpur, India 2008.
8. Gatziolis, Kleanthis, and Anthony C. Boucouvalas. "Discovering the impact of user profiling in e-services." In Telecommunications and Multimedia (TEMU), 2014 International Conference on, pp. 208-213. IEEE, Heraklion, Crete, Greece 2014.
9. Gorbunov, Jurij Aleksandrovich, Lev Nikolaevich Krotov, and Elena L'vovna Krotova. "Legitimate User Profiling Based on the Third Order Spline Approximation of the Initial Data Sequence." World Applied Sciences Journal29, no. 12 (2014): 1605-1610.
10. Amato, Giuseppe, and Umberto Straccia. "User profile modeling and applications to digital libraries." In Research and Advanced Technology for Digital Libraries, pp. 184-197. Springer Berlin Heidelberg, 1999..
11. Teixeira, Cláudio, Joaquim Sousa Pinto, and Joaquim Arnaldo Martins. "User profiles in corporate scenarios." In Internet and Web Applications and Services, 2008. ICIW'08. Third International Conference on, pp. 614-619. IEEE, Athens, Greece, 2008.
12. Liang, Huizhi, Jim Hogan, and Yue Xu. "Parallel user profiling based on folksonomy for large scaled recommender systems: An implimentation of cascading mapreduce." In Data Mining Workshops (ICDMW), 2010 IEEE International Conference on, pp. 154-161. Sydney, Australia IEEE, 2010.
13. Sumitkumar Kanoje, Sheetal Girase, Debajyoti Mukhopadhyay, “User Profiling for Recommender System,” 4th Post Graduate Conference for Information Technology (iPGCon-2015), Amrutvahini College of Engineering, Sangamner, 24-25 March 2015.
14. Schafer, J. Ben, Joseph A. Konstan, and John Riedl. "E-commerce recommendation applications." In Applications of Data Mining to Electronic Commerce, pp. 115-153. Springer US, 2001
15. BURKE, R. 2000. Knowledge-based Recommender Systems. In: A. KENT (Ed.) Encyclopaedia of Library and Information Systems, Vol. 69, Supplement 32.