Knowledge Discovery (Email Harvesting, Gender Identification & Prediction) in Social Network Data (Facebook 100 Million URL)

Nancy.P[1], R. Geetha Ramani[2]

***Abstract*—Online social networks are the one that pave way for various users to contact each other, give and take information and share their views among themselves. In online social networks the members usually announce a profile, which consists of work and education, arts and entertainment and some basic information like gender, e-mail, etc., Such profile information helps in identifying people, know about their interest, and interact with them in need. The objective of this paper is to extract name, email address and gender from a URL and to predict the gender if unspecified. The Dataset used in this work is a list of 100 million Facebook URLs. In this paper, an Algorithm is devised to extract the information from the profile of Facebook users automatically without any human intervention and a methodology is devised to predict the gender of a user from the first name , if the gender information is not specified in the user profile. This research work paves a way to identify the email communities in Facebook. The outcome of this research reveals the fact that most of the email domains of the facebook user's fall into yahoo, hotmail, Gmail and msn. The other domains are with least number of users. The users with Yahoo id are higher when compared to other email domains.**

***Index Terms*—Algorithm, Email address, Extraction, Facebook, Gender, Prediction, Social Networks.**

I. INTRODUCTION

Online social networks are the one which pave way for various users to contact each other, give and take information and share their views among themselves. MySpace (over 275 million users), Facebook has more than 400 million users (Facebook), Twitter has more than 40 million users (Twitter) are examples of wildly popular networks used to share among users. In online social networks the members usually announce a profile, which consists of work and education, arts and entertainment and some basic information like gender, e-mail, etc., .Such profile information helps in identifying people, knowing about their interest, and interacting with them in time of need. However, in practice, not all users provide information about themselves. The profile of such people is said to be private. As per today's practice the members of the Facebook are asked to enter the profile information manually and it depends on the members, whether they wish to enter his/ her details or avoids revealing the details. The profile is said to be public if the information about the member is made public and it is private if the information is not revealed. In this paper we propose a new Algorithm to retrieve the name, e-mail address and gender of a member from a URL. The Data set used in this research is 100 million Facebook URL which was hacked by Ron Bowes, an Internet Security Consultant.

Facebook does not reveal a user's email address to any other user that is not in his friend list. In case the harvester is in the list, the user's email address is presented as a GIF image to prevent automated extraction. Twitter, on the other hand, does not reveal a user's email address in any form. However, the personal information that is revealed includes the user's name, personal web page, location and a short bio description [15].

The Web is an enormous source of information contained in billions of individual pages. Information extraction (IE) tries to process this information and make it available to structured queries. Most often, information extraction systems are targeted towards specific domains of interest and involve either manual or semi-automatic learning of the target examples involved [16]. In contrast, the goal of automatic information extraction is to discover relations between data items of interest and similar data items on a large scale and independently of their domain without any training [2, 14]. The common format used by a web page is HTML. Data extraction from HTML is normally done with the help of wrappers. Existing Wrapper generation have the following features:

First, the wrapper generator works with information provided by the user or by external tool. Second, it is usually assumed that the wrapper works by knowing about the schema of data that is to be extracted. Finally, wrappers are generated by examining one HTML page at a time.

Another problem that prevails in the extraction of web page data includes the dominance of human factor (users) in the extraction process. In several similar applications such as RoboMaker (OpenKapow), YahooPipes, or Karma, that problem may occur because users search and select data

Manuscript revised July 17, 2012.

Nancy is a Full-time PhD research scholar in the Department of Computer Science and Engineering, Rajalakshmi Engineering College (affiliated to Anna University, Chennai), Thandalam, Chennai, India. Phone: 91-9841242291 (e-mail:graciarun@gmail.com)

Dr.R.Geetha Ramani is Associate Professor, Department of Information Science and Technology, College of Engineering, Anna University, Chennai, India (e-mail: rgeetha@yahoo.com).

table from a single web page manually. Since it is time consuming and costly, the process becomes less effective and efficient.

With respect to prediction of gender, the earlier approaches used the information provided by friends of a user based on user's affiliation in various groups. The accuracy of the prediction techniques was also low.

In this research we propose an algorithm for automatic extraction of data (name, email address and gender) from Facebook URLs and also combine the process of prediction of gender if unspecified in the user profile. The number of steps involved in the process of extraction of web information is less when compared to previous approaches. The proposed algorithm does not require Data Cleaning as the extraction process is highly accurate. The techniques used for prediction of gender include usage of first name of the user mentioned in the user profile.

A. *Paper Organization*

The paper is organized in the following manner: Section 2 gives a brief description of the related work. Section 3 narrates the proposed design of the work (overview of the system design, and the steps involved in the process), description of the Dataset. Section 4 explains about the experimental results obtained projects the results obtained. Section 5 concludes the paper.

II. RELATED WORK

The work carried out so far by other researches that are related to retrieval of web information and prediction of gender is concisely presented here.

Gatterbauer [16] employed DOM (Document Object Model) as an approach for extracting web information and determining patterns from HTML tags or code structure in a web page. Gultom [17] used an approach to implement web table extraction and used the concept of mashing from HTML web pages by implementing the application they developed. It also used the concept of DOM generation for the HTML tags of the Web page.

Yanhong Zhai [20] proposed Partial Tree Alignment method which extracts data in two steps (1) Identifying individual data records in a page, and (2) Aligning and extracting data items from the identified data records.

Elena Zheleva [3] showed how an adversary can exploit an online social network with a mixture of public and private user profiles to predict the private attributes of users. Liu [7] used a Bayesian network approach to model the causal relations among people in social networks, and studied the impact of prior probability, influence strength, and society openness to the inference accuracy on a real online social network. Their experimental results revealed that personal attributes can be inferred with high accuracy especially when people are connected with strong relationships. Further, even in a society where most people hide their attributes, it is still possible to infer privacy information.

Hetherly [13] and his team explained how to launch inference attacks using released social networking data to predict undisclosed private information about individuals. They devised three possible sanitization techniques that could be used in various situations and explored the effectiveness of these techniques by implementing them on a dataset obtained from the Dallas/Fort Worth, Texas network.

Polakis [15] demonstrated how names extracted from social networks can be used to harvest email addresses. Cong Tang [4] and team developed a new and powerful technique for inferring gender for users who do not explicitly specify their gender. Having inferred the gender of most users in their Facebook dataset, gender characteristics were learnt and analysis on how males and females behave in Facebook was carried out. Different Gender prediction techniques like Offline Name List Predictor, Facebook Generated Name List Predictor, Local Information Predictor and Friend Information Predictor were designed and implemented individually. This research work has combined Offline Name List and Facebook Generated Name List for predicting the gender of a Facebook user.

III. PROPOSED DESIGN OF THE SYSTEM

This Section gives a brief description about the Dataset used for this research, the design and Architecture of the proposed System. The various steps involved in the algorithm for the process of extraction of information from the web page (Facebook user profile) are discussed in this section. The Techniques used in the Prediction of is also explained.

A. *Dataset Description*

The original dataset considered for this research is a torrent file downloaded from the blog of Skull Security. It was generated around July 15, 2010, by Ron Bowes, an internet security consultant. (Check out http://www.skullsecurity.org for more information). He crawled the Facebook server of United States, and obtained the profiles of more than 100 million Facebook members. The dataset includes the Facebook URLs of various persons, the first name and last name of the users and the coresponding counts of the names as shown in Table 1.

Table I
DESCRITION OF THE DATASET

File Name	Description
Facebook.rb	The script used to generate these files
Facebook.nse	The script that will be used for the second pass
Facebook-URLs	The full URLs to every profile
Facebook-names-original	All names, including duplicates
Facebook-names-unique	All names, no duplicates
Facebook-names-with count	All names, no duplicates but with count
Facebook-firstnames-withcount	All first names (with count)
Facebook-lastnames-withcount	All last names (with count)
Facebook-flast-withcount	All first initial last names(with count)
Facebook-first.l-withcount	All first name last initial (with count)

Of the original torrent file, this research work focuses only Facebook-URLs. A Sample file of Facebook URL is shown in Table 2.

Table II

A SAMPLE FACEBOOK URL FILE

http://en-us.facebook.com/people/-/100000218612110
http://en-us.facebook.com/people/-/100000226945128
http://en-us.facebook.com/people/-/100000233424427
http://en-us.facebook.com/people/-/100000234406002
http://en-us.facebook.com/people/-/100000247916023
http://en-us.facebook.com/people/-/100000249924756
http://en-us.facebook.com/people/-/100000254263318
http://en-us.facebook.com/people/-/100000297669803
http://en-us.facebook.com/people/-/100000317949277
http://en-us.facebook.com/people/-/100000361441792
http://en-us.facebook.com/people/-/100000397174436
http://en-us.facebook.com/people/-/100000399691327
http://en-us.facebook.com/people/-/100000425418926
http://en-us.facebook.com/people/-/100000446621908
http://en-us.facebook.com/people/-/100000513011966
http://en-us.facebook.com/people/-/100000515000987
http://en-us.facebook.com/people/-/100000518880222
http://en-us.facebook.com/people/-/100000605485416
http://en-us.facebook.com/people/-/100000638025816
http://en-us.facebook.com/people/-/100000727219704
http://en-us.facebook.com/people/-/100000750793361
http://en-us.facebook.com/people/-/100000638025816
http://en-us.facebook.com/people/-/100000727219704
http://en-us.facebook.com/people/-/100000750793361
http://en-us.facebook.com/people/-/100000842505349

B. *Overview of the System Design.*

The overview of the System Design is shown in Figure 1.

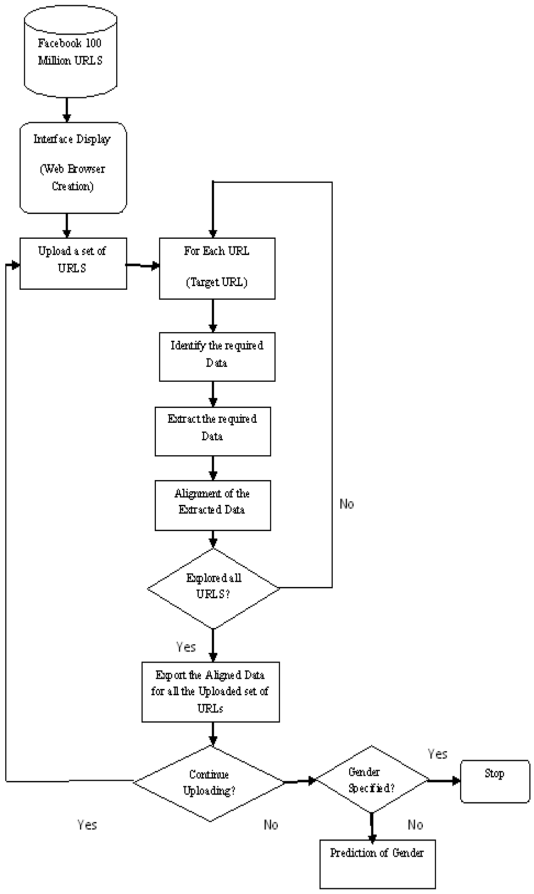

Facebook 100 Million URLS
Interface Display (Web Browser Creation)
Upload a set of URLS
For Each URL (Target URL)
Identify the required Data
Extract the required Data
Alignment of the Extracted Data
Explored all URLS?
No
Yes
Export the Aligned Data for all the Uploaded set of URLs
Continue Uploading?
Yes
No
Gender Specified?
Yes
Stop
No
Prediction of Gender

Fig 1. Overview of the System

C. *Steps involved in the Process*

The algorithm designed is to implement the process of extracting the required data from the web page (Facebook User Profile). The required data which is to be extracted includes the user name, email address, gender. The general flow of the work is briefly outlined as Algorithm and shown in Figure 2.

```
1: Create a Web browser
2: Open a session
3:Upload the URL list
4:    Repeat
5:      Open a single URL
6:      Doc=htmldoc (URL)
7:      If (doc! =data)
8:      The Data is unavailable
9:      Else
10:     For each data € doc
11:          Extractdata (data, doc)
12: Place the extracted data in the Table
13:          Delete the Read URL
14:    Until all URL's are read
15:  Prediction of Gender if unspecified.
```

Fig 2. General Flow of the Algorithm

In the above Algorithm 'data' denotes the required data which is to be extracted (Name, email address and gender). In Step 1, a Web browser is created to get connected to Facebook.com. The Web browser also enables to implement the algorithm for extracting the data from web pages using graphical User interface. The web Browser is designed to enable creation of new session in Facebook.com, upload a set of URLs, to view the data extracted and to export the extracted data into an Excel Worksheet. Step 2 deals with creation of session. The session can be established by logging in a Facebook account. In step 3, the end-user has to select a list of URLs for which data is to be extracted. This is the place where human intervention is required. The original dataset contains about 100 million URLs which are voluminous to be loaded at once. Hence the URLs are divided into subsets to be loaded into the browser. From the set of URLs uploaded, the profile page of each URL is opened one by one to extract the required data. Step 4 to Step 15 is explained in the following subsections in detail.

Data Identification

For each URL loaded, the profile page is displayed in the web browser. In this step, html code of the loaded page is retrieved and stored. The code segment related to the retrieval of html code is as shown in Figure 3.

```
   mshtml.IHTMLDocument2 doc =
(mshtml.IHTMLDocument2)_ActiveWebBrowser.
                                          Docume
                                          nt;
    Page code = ((mshtml.HTMLDocumentClass)(doc)).
   documentElement.outerHTML;
   str = Page code
```

Fig 3. Code segment for retrieving HTML source

The expected data which is to be extracted includes name, email address and gender of every Facebook user from his /

her profile. The Page code shown above contains the entire details of the user in HTML format. The variable k in the following code denotes the total number of tags found in Page code, Variable Gen 1 is used in finding the existence of expected data and length denotes length of the splitted document. The Pseudo code used in identifying the required data is shown in Figure 4.

```
For I= 1 to k
//If the Page code contains data
If (Page code! = null)
//Split the Doc into two based on the
Existence of the data
Gen1=Regex.split (Page code,” data</TH>”)
//If length is 2 expected data exits
Length1=gen1.length
```

Fig 4. Algorithm for Data Identification

Data Extraction

In this step as the existence of the required data is assured the Doc is splitted into two segments. The first segment consists of all the HTML tags and values excluding the expected data. The Second segment consists of the remaining values and tags. The Pseudo code of the extraction process is shown in Figure 5.

```
For I= 1 to k
//If the Doc contains data
If (Doc! = null)
//Split the Doc into two based on the existence of the data
Gen1=Regex.split (Doc,”data</TH>”)
//If length is 2 expected data exits
Length1=gen1.length
If (length1==2)
Tempstring=gen1 [1].to string ();
Data=Read the Substring of Tempstring until the tag ends
```

Fig 5. Algorithm for Data Extraction

Data Alignment

In the Data Alignment stage, the data extracted from each URL is viewed in the developed web browser. It can be used to view the extracted data immediately for the loaded page. The algorithm used for viewing the extracted is given in Figure 6. The extracted data is represented as subitem in the algorithm.

```
Data alignment (URL, subitem)
Assign the Header of each subitem
Row=1;
While (Not EOF ()) For each URL uploaded
Add extracted subitem under its header
Increment Row
```

Fig 6. Algorithm for Data Aligment

Data Export

In the Data Export stage, the listed sub items (extracted data) are exported into an excel sheet for further analysis and visualization. The term subitem in the Algorithm denotes the data extracted. The algorithm is shown in Figure 7.

```
Open a New Excel Application
Open a New Worksheet
Row = 1;
Column =1;
    For each (header in data alignment)
        Column=1;
        For each subitem
        Worksheet cell (Row, Colum) = subitem;
      Increment column;
 Increment Row
```

Fig 7. Algorithm for Data Export

Gender Prediction

The extracted data exported into the excel worksheet in the previous step contains some unspecified values. The extracted data contains the name, URL, Email address and gender. Of all the extracted data, gender of a person can be predicted if unspecified and the prediction uses the algorithm shown in Figure 8.

Popular baby names [18] is a first name list USA baby name list which consists of 1,736 male names and 2,023 female names.

Facebook namelist [19] also list the first name with a count of male and female.

```
Gender Prediction ($name, $gender)
For every unspecified gender for $name
  Repeat
    If $name appears in the worksheet
    Assign Gender with high probability
    Else if $name €popular baby names
    Assign Gender with high probability
    Else if $name € Facebook namelist
    Assign Gender with high probability
  Until the all the unspecified gender is predicted.
```

Fig 8. Gender Prediction Algorithm

Throughout the Gender Prediction Algorithm, the Probability was calculated using Bayes Theorem. The idea

P (A) = proportion of trials producing outcome as Male
P (B) = proportion of trials producing outcome as Female
If we consider only trials in which A occurs, the proportion in which B also occurs is P (B|A). If we consider only trials in which B occurs, the proportion in which A also occurs is P (A|B). In simpler form,

For events A and B, provided that P (B) ≠ 0

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

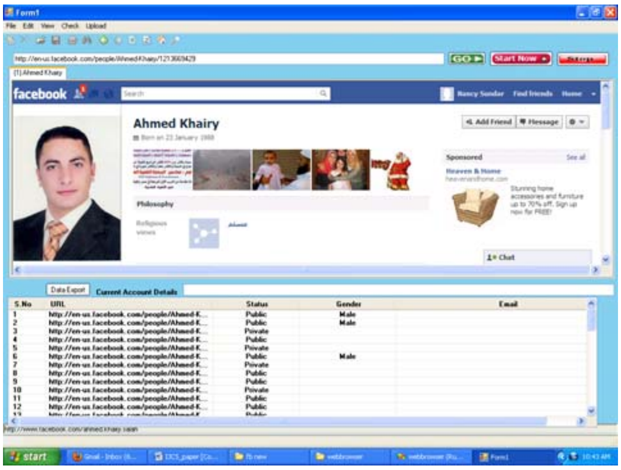
Similarly P (B|A) is found. The higher of the two is selected for predicting the Gender.

A brief description of the various processes involved in the process of extracting the required data from the Web Page (Facebook URL) and Gender prediction is seen in the above section.

IV. EXPERIMENTAL RESULTS

The entire application was developed using .NET as it has many inbuilt features which are useful in developing Web browser creation and in extraction of the web data. The Original Dataset consisted of 100 Million Facebook URLs which is about 1.65 GB. A Software named Text file Splitter is used to Split the original Data (100 Million Facebook URLs) into Small Text files. The size of the Splitted file was about 1.35 MB with about 25000 URLs in each file. The total number of Splitted files was about 1232 text files which contained only 30 million URLs. As first phase of the research, only 30 million URLs were considered for extraction of required data and the results obtained are pertained to 30 million URLs only.

The snapshot of the application with a web Page loaded for a specific URL with some of the extracted data is shown in Figure 9.

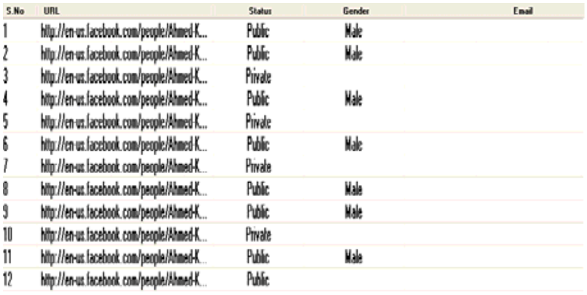


Data Export | Current Account Details

S.No	URL	Status	Gender	Email
1	http://en-us.facebook.com/people/Ahmed-K...	Public	Male	
2	http://en-us.facebook.com/people/Ahmed-K...	Public	Male	
3	http://en-us.facebook.com/people/Ahmed-K...	Private		
4	http://en-us.facebook.com/people/Ahmed-K...	Public		
5	http://en-us.facebook.com/people/Ahmed-K...	Private		
6	http://en-us.facebook.com/people/Ahmed-K...	Public	Male	
7	http://en-us.facebook.com/people/Ahmed-K...	Private		
8	http://en-us.facebook.com/people/Ahmed-K...	Public		
9	http://en-us.facebook.com/people/Ahmed-K...	Public		
10	http://en-us.facebook.com/people/Ahmed-K...	Private		
11	http://en-us.facebook.com/people/Ahmed-K...	Public		
12	http://en-us.facebook.com/people/Ahmed-K...	Public		

Fig 9. A snapshot of the application with a web page loaded for a specific URL.

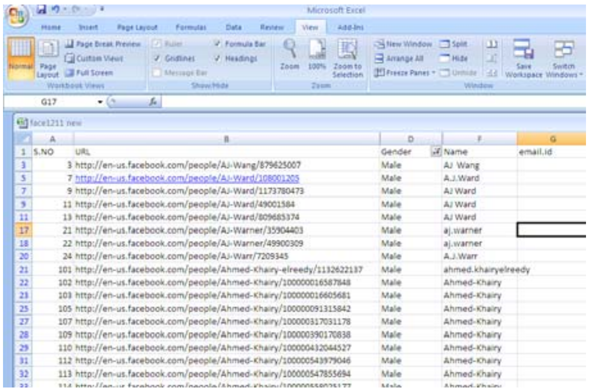
The extracted data is properly aligned under specific headers and the snapshot for the Data alignment is shown in Figure 10. The data has some unspecified values for gender which are to be predicted.



S.No	URL	Status	Gender	Email
1	http://en-us.facebook.com/people/Ahmed-K...	Public	Male	
2	http://en-us.facebook.com/people/Ahmed-K...	Public	Male	
3	http://en-us.facebook.com/people/Ahmed-K...	Private		
4	http://en-us.facebook.com/people/Ahmed-K...	Public	Male	
5	http://en-us.facebook.com/people/Ahmed-K...	Private		
6	http://en-us.facebook.com/people/Ahmed-K...	Public	Male	
7	http://en-us.facebook.com/people/Ahmed-K...	Private		
8	http://en-us.facebook.com/people/Ahmed-K...	Public	Male	
9	http://en-us.facebook.com/people/Ahmed-K...	Public	Male	
10	http://en-us.facebook.com/people/Ahmed-K...	Private		
11	http://en-us.facebook.com/people/Ahmed-K...	Public	Male	
12	http://en-us.facebook.com/people/Ahmed-K...	Public		

Fig 10. Data Alignment in the Web browser

All the extracted data is exported into a excel worksheet for further analysis and exported data is shown in Figure 11.



	A	B	D	F	G
1	S.NO	URL	Gender	Name	email.id
3	3	http://en-us.facebook.com/people/AJ-Wang/879625007	Male	AJ Wang	
5	7	http://en-us.facebook.com/people/AJ-Ward/108001205	Male	A.J.Ward	
7	9	http://en-us.facebook.com/people/AJ-Ward/1173780473	Male	AJ Ward	
9	11	http://en-us.facebook.com/people/AJ-Ward/49001584	Male	AJ Ward	
11	13	http://en-us.facebook.com/people/AJ-Ward/809685374	Male	AJ Ward	
17	21	http://en-us.facebook.com/people/AJ-Warner/35904403	Male	aj.warner	
18	22	http://en-us.facebook.com/people/AJ-Warner/49900309	Male	aj.warner	
20	24	http://en-us.facebook.com/people/AJ-Warr/7209345	Male	A.J.Warr	
21	101	http://en-us.facebook.com/people/Ahmed-Khairy-elreedy/1132622137	Male	ahmed.khairyelreedy	
22	102	http://en-us.facebook.com/people/Ahmed-Khairy/100000016587848	Male	Ahmed-Khairy	
23	103	http://en-us.facebook.com/people/Ahmed-Khairy/100000016605681	Male	Ahmed-Khairy	
25	105	http://en-us.facebook.com/people/Ahmed-Khairy/100000091315842	Male	Ahmed-Khairy	
27	107	http://en-us.facebook.com/people/Ahmed-Khairy/100000317031178	Male	Ahmed-Khairy	
28	109	http://en-us.facebook.com/people/Ahmed-Khairy/100000390170838	Male	Ahmed-Khairy	
29	110	http://en-us.facebook.com/people/Ahmed-Khairy/100000432044527	Male	Ahmed-Khairy	
31	112	http://en-us.facebook.com/people/Ahmed-Khairy/100000543979046	Male	Ahmed-Khairy	
32	113	http://en-us.facebook.com/people/Ahmed-Khairy/100000547855694	Male	Ahmed-Khairy	

Fig 11. Data Exported into Excel Worksheet

Out of 30.82 Million URLs, 7.93 Million URL's were found to be under private category and 22.8 Million URL's were found be under public category. From 22.8 URL's, Gender information was present in 19.45 Million URL's. However, 13311 URLs' contained email address. From the extracted email addresses, it was found that users belong to various domains like yahoo, hotmail, Gmail and msn. 6004 users were found to be under yahoo domain, 2792 in hotmail, 2143 in Gmail and 2372 in msn. The majority of email domain communities which persisted in Facebook ULRs is shown in Figure 12.

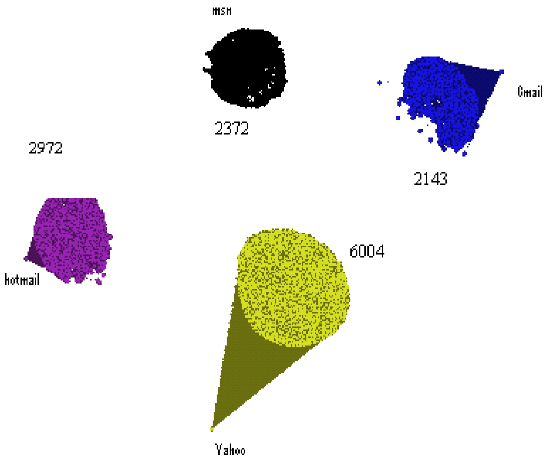


Fig 12. Majority of email domain Communities in Facebook

Out of 19.45 Million URLs in which Gender Information existed, it was found that 11.47 Million users were male and 7.97 users were female. In 3.45 Million URLs, the Gender information was not specified. We used a combination of Gender Prediction techniques (Name centric approach) like popular baby name list, facebook name list to predict the Gender of 3.45 Million URLs. From the predicted Genders, 1.85 Million were predicted to be male and 1.57 Million were predicted to be female.

V. Conclusion

In this paper we have considered a voluminous dataset which contains the URL of more than 100 million Facebook users. Of 100 million URLs, this research work explored only 30 million URLs as a first phase. This research focuses on extracting e-mail address identifying the gender and predicting the gender if unspecified in the user profile. The various steps in terms of algorithmic techniques for extracting the content from user profile information were discussed. From this research it is clear that the domains of majority of email domain community falls into yahoo, hotmail and Gmail and msn of which yahoo have the highest ranking. It can be further concluded that only less than 0.25%of users have specified email address in their profiles. It can be concluded that majority of the email domain of Facebook users fall into yahoo, hotmail, Gmail and msn. The other domains are with least number of users. The users with Yahoo id are higher when compared to other email domains. The paper also uses various offline gender prediction techniques and predicted the gender of user. Further it is clear that any information if provided in the profile can be retrieved automatically and retrieval of required data from all the 100 million URLs is the next step ahead in our research.

References

[1] Facebook statistics, available at: http://www.facebook.com/press/info.php?statistics.

[2] Mislove, A., Viswanath, B., Gummadi, K.P., Druschel, P.: You are who you know: Inferring user profiles in online social networks. In: WSDM (2010)

[3] Zheleva, E., Getoor, L.: To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles. In:WWW(2009)

[4] Cong Tang, Keith Ross†, Nitesh Saxena†, and Ruichuan Chen : What's in a Name: A Study of Names, Gender Inference, and Gender Behavior in Facebook

[5] Krishnamurthy, B., and Wills, C. E. On the leakage of personally Identifiable information via online social networks. In WOSN '09: Proceedings of the 2nd ACM workshop on Online social networks (New York, NY, USA, 2009), ACM, pp. 7–12.

[6] Viswanath, B., Mislove, A., Cha, M., and Gummadi, K. P. On the evolution of user interaction in facebook. In WOSN '09: Proceedings of the 2nd ACM workshop on Online social networks (2009), ACM, pp. 37–42.

[7] Xu, W., Zhou, X., Li, L.: Inferring Privacy Information via Social Relations. In: 24th ICDE Workshop. pp. 154–165 (2008)

[8] Facebook. http://www.facebook.com.

[9] Facebook 100 million user profile : http://www.skullsecurity.org

[10] Lindamood, J., Kantarcioglu, M.: Inferring Private Information Using Social Network Data. Tech. Rep. UTDCS-21-08 (2008)

[11] He, J., Chu, W.W., Liu, Z.: Inferring privacy information from social networks. In: ISI. pp. 154–165 (2006)

[12] Facebook updates privacy settings, available at: http://blog.facebook.com/blog.php?post=197943902130

[13] Heatherly,R.,Kantarcioglu,M.,Thuraisingha B., Lindamood, J.: reventing Private Information Inference Attacks on Social Networks. Tech. Rep. UTDCS-03-09, University of Texas at Dallas (2009)

[14] H. Chun, H. Kwak, Y-H. Eom, Y-Y. Ahn, S. Moon and H. Jeong.Online Social Networks: Sheer Volume vs Social Interaction. In Proc. of IMC, 2008.

[15] I. Polakis, G. Kontaxis, E.Markatos Using Social Networks to harvest e-mail addresses WPES'2010.

[16] Gatterbauer, W., P. Bohunsky, M. Herzog, B. Krupl and B.Pollak,2007. Towards DOMain Independent Information Extraction from Web Tables. Proceeding of the International World Wide Web Conference Committee (IW3C2), May 8-12, ACM, Banff, Alberta, Canada, pp: 71-80.

[17] Rudy A.G.Gultom,Riri Fitri Sari,Bagio, Proposing the new Algorithm and technique development for Integration Web Table Extraction and building a Mashup Journal of Computer Science 7 (2): 129-142, 2011 ISSN 1549-3636.

[18] Popular baby names, available at http://www.ssa.gov.OACT/babynames

[19] Facebook name list, available http;//sites.google.com/site/facebooknamelist/.

[20] Yanhong Zhai, Bing Liu. Web Data Extraction Based on Partial Tree Alignment in *WWW 2005*, May 10-14, 2005, Chiba, Japan. ACM 1-59593-046-9/05/0005.

[21] Gengxin Miao, Junichi Tatemura, Wang-Pin Hsiung, Arsany Sawires2,Louise E. MoserExtracting Data Records from the Web Using Tag Path Clustering in *WWW 2009*, April 20–24, 2009, Madrid, Spain. ACM 978-1-60558-487-4/09/04.ACM 978-1-60558-487-4/09/04.

[22] Valter Crescenzi Giansalvatore Mecca Paolo Merialdo RoadRunner: Towards Automatic Data Extraction from Large Web Sites in Proceedings of the 27th VLDB Conference, Roma, Italy, 2001

AUTHORS PROFILE

Dr.R.Geetha Ramani is Associate Professor in Department of Information Science and Technology, College of Engineering, Anna University, Guidy, Chennai, India.. She has more than 15 years of teaching and research experience. Her areas of specialization include Data mining, Evolutionary Algorithms and Network Security. She has over 50 publications in International Conferences and Journals to her credit. She has also published a couple of books in the field of Data Mining and Evolutionary Algorithms. She has completed an External Agency Project in the field of Robotic Soccer and is currently working on projects in the field of Data Mining. She has served as a Member in the Board of Studies of Pondicherry Central University. She is presently a member in the Editorial Board of various reputed International Journals.

Mrs. P. Nancy completed her M.E. Computer Science & Engineering in Department of Science and Engineering at Bannari Amman Institute of Technology, Sathyamangalam, affiliated to Anna University, Chennai, India. She has more than 5 years of teaching experience. Presently she is pursuing her Ph.D in Computer Science and Engineering at Rajalakshmi Engineering College, affiliated to Anna University of Technology, Chennai. Her areas of interest include Data Mining, Data Structures, Computer Networks and Software Engineering. She has attended and presented papers in National and International Conferencesand journals.