Detecting the Age of a Person through Web Browsing Patterns

Chinmay Prakash Swami
Department of Computer Engineering, MESCOE,
Pune – 411001 Maharashtra,
India

Sumit Maroti Raut
Department of Computer Engineering, MESCOE,
Pune – 411001 Maharashtra,
India

Prasad Bhalchandra Tarte
Department of Computer Engineering, MESCOE,
Pune – 411001 Maharashtra,
India

Nuzhat Faiz Shaikh
Department of Computer Engineering, MESCOE,
Pune – 411001 Maharashtra,
India

Sagar Kisan Rakshe
Department of Computer Engineering, MESCOE,
Pune – 411001 Maharashtra,
India

ABSTRACT

As more and more people are using the internet, the demographics of these people such as their age, their gender, their location etc. are highly valued by several business enterprises. We are cognizant about the fact that each user has his/her preferences when it comes to browsing over the internet. This paper designed a system that predicts the age of the user based on his/her web browsing patterns using the A-priori algorithm.

General Terms

Data Mining, Age Prediction

Keywords

A-Priori Algorithm, User demographic Prediction, Association Rule Learning

1. INTRODUCTION

There are wide array of enterprises whose business is depended on the internet. Figure 1 shows market values of such enterprises which was published by 'Statista' one of the world's largest statistics portal [10].

Due to this these organizations are concentrating more towards improving the user experience as well as providing modified services to sustain the pre-existing customers as well as attract new ones. For example various shopping sites suggest various items the user might interest in based on what he has previously searched. Most of the internet based industries generate huge income via showing advertisements on their sites. Figure 2 illustrates this fact. One of the most popular techniques used by these industries is known as "Behavior targeting".

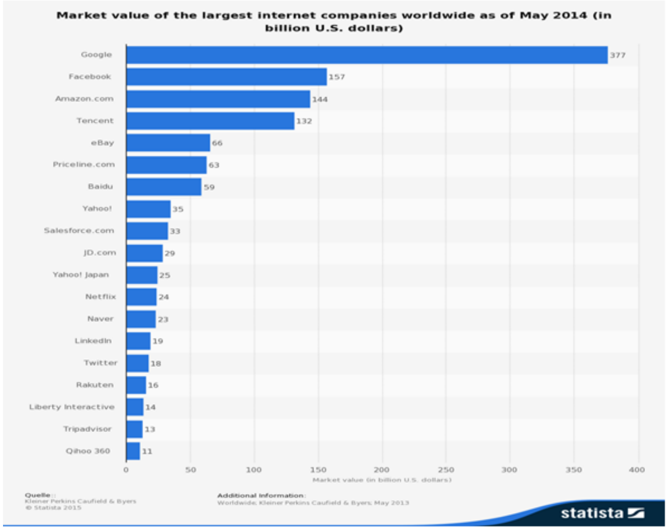

Fig 1 the statistic depicts the market value of the largest internet companies worldwide in 2014.

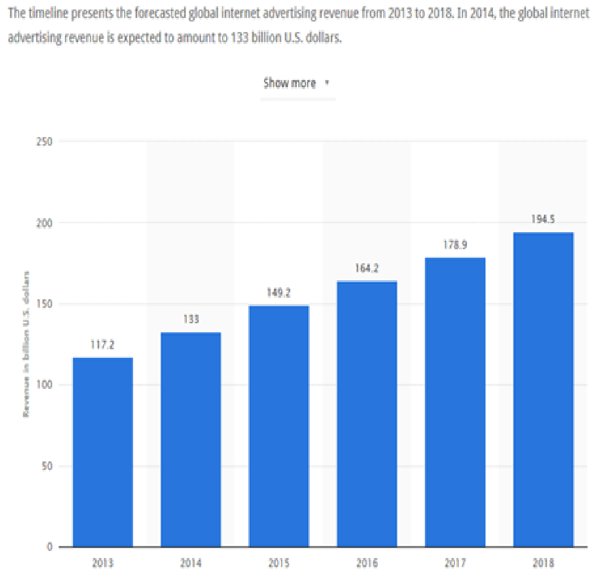
The timeline presents the forecasted global internet advertising revenue from 2013 to 2018. In 2014, the global internet advertising revenue is expected to amount to 133 billion U.S. dollars.

Fig 2 Global internet advertising revenue from 2013 to 2018('Statista')

Behavior targeting consists of multiple technologies and techniques used by online website publishers and advertisers whose primary objective is to increase the effectiveness of advertising using user web browsing behavior information. In other words behavior targeting uses information gathered from users web browsing behavior to select advertisements that should be displayed such that it interests the user [9]. Hence we can say that web browsing behavior of the user is of great importance to such organizations. Table 1 represents the internet usage in terms of user's age and was published after survey conducted by "statista" in 2014

Table 1 Internet usage statistics with respect to user's age

Age Group	Internet Usage
15-24	26.7
25-34	26.6
35-44	20.3
45-54	13.7
55+	12.7

Internet usage is in terms of Percentages

Hence having insight of users age is bound to affect the effectiveness of the advertisements that are displayed. For example if the user's age is 55+ then he's most likely to be interested in advertisements regarding some pilgrimage packages or some "Yoga" guru's campaigns or retirement policy providing sites etc. But if such advertisements are displayed to a user of age say 26 then he's most probably going to ignore it hence reducing the effectiveness of advertisements.

But asking for such sort of information from the user is not always rewarding as we are aware of the fact that cybercrime has grown since past couple of decades and may keep growing. Hence internet users tend to be discreet about their online activities and avoid revealing it. There are various techniques used for predicting the age the user [1] one of which is described by Misha Kakkar, and Divya Upadhyay which uses a neural network technique known as Multi-Layer Perception using Back propagation Algorithm (BPNN) [3]. Santosh Kabbu and his colleague's work described various machine learning techniques used to predict the demographic information of the website using only the information obtained from various aspects of website and not relying on human input [4]. Also various natural language processing techniques can be used to predict the user's age and gender [5]. Along with these various other techniques exist for predicting the user demographic information's [1] [6] [7].

In this paper we have discussed a method for predicting the age of the user with the help of a data mining and machine learning algorithm known as A-Priori [1] [7].

In this paper we have discussed a method for predicting the age of the user with the help of a data mining and machine learning algorithm known as A-Priori [1] [7]. The remaining paper is divided into 6 sections. Section 1 contains introduction, section 2 contains high level system architecture. Section 3 contains basic explanation of A-priori algorithm. And section 4 and 5 contains results and conclusion respectively and section 6 contains Future work.

2. HIGH LEVEL SYSTEM ARCHITECTURE

The system consists of following main modules [2]:

2.1 Browser extension:

Browser extension is used for gathering and storing browsing history in the local database on local machine. Browsing history consists of URL of website, time requires for accessing particular website.

2.2 Server database:

Server database is a place where the browsing data from various clients is collected and stored. All the data collected on client side is sent to server through network.

2.3 Client application:

Using client application client perform following tasks:

1. View their browsing data.

2. Filter the data and upload data to the server.

3. Enter their age (Only during training phase)

2.4 Admin application:

Admin application is where admin can apply data mining and view the results.

Using admin application admin perform following tasks:

1. Login into the system with valid credentials.

2. View entire history of various clients

3. Filter the URLs such as search engines or the sites which are used by almost all age groups.

4. Apply data mining on the gathered data to get the final output.

The system architecture is explained in paper reference number [2] [1]. The detail working of the system is explained in the Figure 4 which is High-level System Architecture.

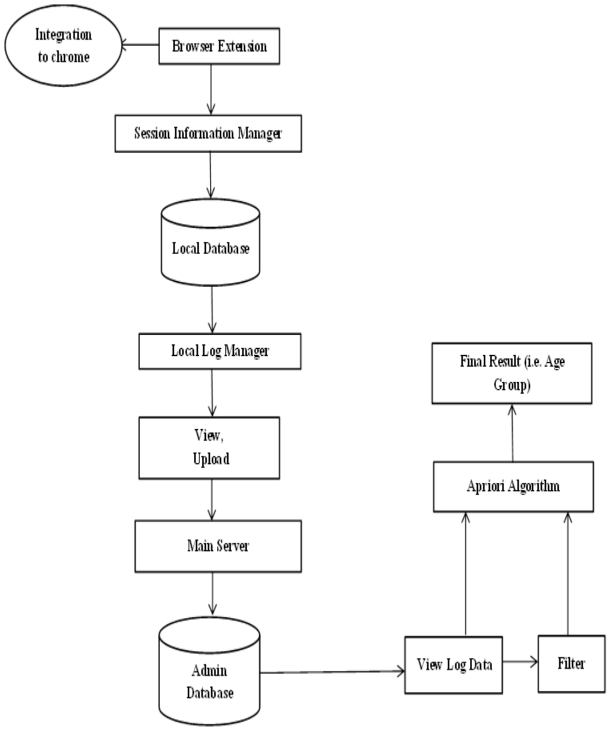

Fig 4 High-Level System Architecture

In the above Figure 4 the flow of the working system is explained such as the browser extension which is written in

JavaScript is integrated into the Chrome browser (Currently built for Chrome only).The client will access the websites and all the history of URLs of clients will be collected with the help of extension into the local database in the local machine. Session Information Manager is the web service which is used to collect and send all the history from extension to the local database. Local log manager is the client side module which can be accessed for viewing the history of self .Then Client can view the data and along with the age of client the data can be uploaded to the main server. The main server collects the whole log data of various clients and stores it into global database.

The admin need to log in into the system with valid credentials to perform the operations on the database to get the output. The admin can view entire log history of different clients altogether and also admin can filter the websites such as search engines (i.e. Google, Yahoo) and the websites that are mostly accessed by the people of almost all age groups(i.e. Online shopping sites etc.).The age groups here are divided in the scale of 10 such as 10-20, 20-30 and so on. After filtering the task of admin is to apply mining on the collected data. The Apriori algorithm is used for data mining and forming strong association rules. The min_support and min_conf values are set before going to calculate the actual output and on comparing with these values some of the rules are neglected and remaining which are strong are selected. Finally from the strong association rules we get the output.

3. APRIORI ALGORITHM

The Apriori algorithm is used for mining of frequent item sets and also for association rule learning on the transactional databases. The Apriori Algorithm is an influential algorithm for the mining of item sets which are used frequently for Boolean association rules. The Apriori algorithm uses the most frequent item sets to generate the strong association rules.[11] Apriori algorithm is basically designed to operate on the databases containing transactions such as items bought by the customers, market basket analysis etc. where the collection of items are made. Apriori uses a "bottom up" approach, where frequent subsets of item sets in a transaction are extended one item at a time which is known as candidate generation step and groups of candidates are tested against the data. The algorithm will terminate at a point when no further successful extensions are found [8].

We use the minimum support and minimum confidence concepts here.

The support SUPP(X) of an item set X is defined as the proportion of transactions in the data set which contain the item set. In the database containing five transactions, the item set {mouse, keyboard, usb} has a support of 1/5=0.2 since it occurs in 20% of all transactions (1 out of 5 transactions). [11]

The confidence of a rule is defined conf(X=>Y) =SUPP(X U Y)/SUPP(X). For example, the rule {mouse, keyboard}=>{usb} has a confidence of 0.2/0.2=1.0 in the database, which means that for 100% of the transactions containing mouse and keyboard the rule is correct (100% of the times a customer buys mouse and keyboard, usb is bought as well).[12]

Pseudo-code:

```
Ck: Candidate item set of size k
Lk: frequent item set of size k
L1= {frequent items};
for (k= 1; Lk! =NULL; k++) do begin
Ck+1= candidates generated from Lk;
for each transaction tin database do
Increment the count of all candidates in Ck+1that are contained in t
Lk+1= candidates in Ck+1with min_support
end
return UkLk;
```

In Figure 5, the working of Apriori algorithm is shown, in the first step there is an input of list of item sets hold by the transactions. By using Apriori algorithm we have to find frequent items (the items which are appearing frequently in the dataset).

In second step we set the minimum support and confidence values as threshold value. We then generate the pattern of frequent item sets such as (1-frequent item set, 2-frequent item sets and so on.).We have to calculate the support count of each item set in the above pattern. We now compare the supp count of each item set with minimum support value and which are less than that are to be pruned.

After finding the support count of all generated frequent items and pruning is done, we will get the set of items which are used to form the strong association rules. By comparing the minimum confidence value of item sets with the threshold value we generate the final strong association rules.

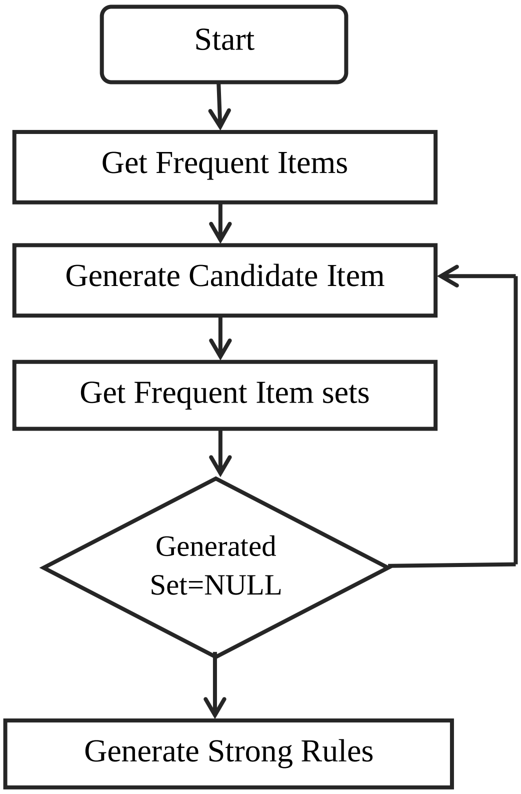

Fig 5 Apriori Algorithm Flow

4. RESULTS

As discussed earlier the Apriori algorithm is used to form strong association rules with the help of minimum support and minimum confidence concepts. With the threshold values of minimum support and minimum confidence the rules which were formed were compared and pruned. We got the final strong rules to be considered to calculate the final output. The final output is the users' age group and confidence value that should be greater than the threshold value of minimum confidence.

As we have used Apriori algorithm we could relate many more things with the age parameter such as location, time etc. Using these parameters we could associate the age with it such as Age-Location, Age-URL & Location- URL. We have used www.whatismyip.com website to get the users IP address to know the location of user. The browser extension can fetch the URLs request and response time as well.

As shown in the following figure 5, the analysis is done and it is shown in the tabular format as final confidence value of particular user for particular websites. At first we tested our system taking five users and their web browsing behavior and calculated the precision of our system, we found that for ten different combinations, seven combinations were correct. At present we have smaller database for ten users, hence there was less accuracy, but as soon as time passes the database will grow in size and we will have more URLs of different users hence accuracy goes on increasing. The approximate accuracy of this system is about 90-93%.

As shown in table 1 there can be any combination of users with their URLs history hence we get different values for different users. The values which are greater than minimum confidence threshold are considered to be strong rules, for our system seven out of ten rules formed correctly and for the common sites for any age group it's showing the confidence value less than minimum value that is calculated before.

From table 2, the graphical representation of data is shown below in figure 7. We have calculated the final confidence value for the various combinations of user and the web browsing patterns. The final confidence value lies between 0 and 100.The minimum confidence value is set as threshold value which on comparing with the final confidence value gives the correct or incorrect result.

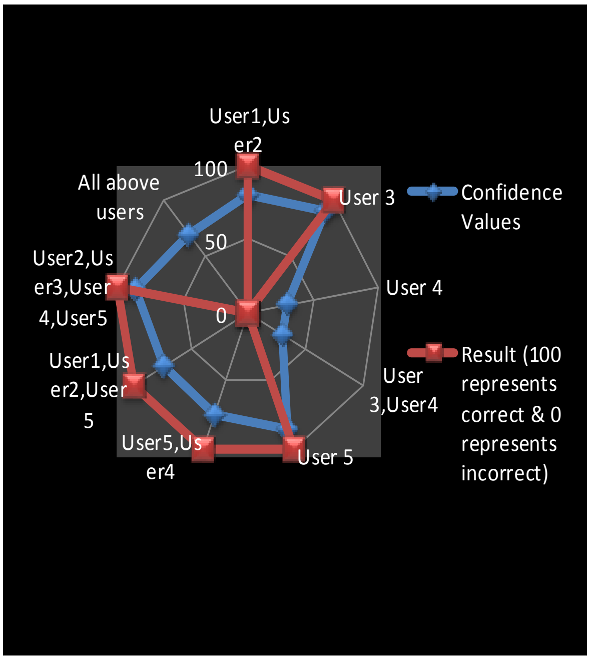

Fig 7 Graphical Representation of Result Analysis

5. CONCLUSION

Through this paper, a method has been proposed for predicting the user's age based on their web browsing patterns using Apriori algorithm. It includes the association rule and has minimum support and minimum confidence concepts which helps us to form strong rules. Final result gives an accuracy of 90%. If implemented as a web service, this system can be used to stop underage children or teens from accessing censored contents online (i.e. parental control can be achieved). Also it can be used by online advertisement agencies for improving their efficiency of their advertisements.

6. FUTURE WORK

The range of age group can also be reduced to 5 or 4. The underlying algorithm can changed to CLA's (Cortical Learning Algorithms) [13]. It can also be implemented on Android and Windows operating systems (mobile) to reach wide array of people as use of mobiles is increasing.

We plan on extending our work on other attributes such as Gender, Occupation etc. to increase the usability of this system and to reach out to much wider array of people.

7. REFERENCES

[1] Chinmay Swami, Prasad Tarte, Sagar Rakshe, Sumit Raut , Nuzhat F. Shaikh, "Detecting the age of a person through web browsing patterns: A Review " International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307,Volume-4, Issue-5,September 2014.

[2] Chinmay Prakash Swami, Prasad Bhalchandra Tarte, Sagar Kisan Rakshe, Sumit Maroti Raut , Nuzhat Faiz Shaikh, "Detecting the age of a person through web browsing patterns" National Conference on, Modeling, Optimization and Control, 4th -6th March 2015, NCMOC – 2015. (Article in a conference proceedings)

[3] Misha Kakkar, DivyaUpadhyay, "Web Browsing Behaviors based age detection", International Journal of Soft Computing and Engineering (IJSCE), ISSN: 2231-2307, Volume-3, Issue-1, March 2013

[4] K. Santosh Kabbu, Eui-Hong Han, George Karypis, "Content-Based Methods for Predicting Web-Site Demographic Attributes". NSF (IIS-0905220, OCI-1048018,IOS-0820730), NIH (RLM008713A), and the Digital Technology Centreat the University of Minnesota

[5] Claudia Peersman, Walter Daelemans, Leona Van Vaerenbergh,, "Predicting Age and Gender in Online Social Networks", Conference'10, Month 1–2, 2010, City, State, Country, Copyright 2010 ACM 1-58113-000-0/00/0010.

[6] Y. Jonathan Schler, Moshe Koppel, Shlomo Argamon,James Pennebaker "Effects of Age and Gender on Blogging" Copyright © 2005, American Association for Artificial Intelligence(www.aaai.org).

[7] M. Reyhaneh Tamimi, Prof. Dr.Mohammad pourzarandi, " The Application of Web Usage Mining In E-commerce Security", 978-1-4799-0393-1/13/$31.00 ©2013 IEEE

[8] K. Geetha, Sk. Mohiddin, "An Efficient Data Mining Technique for Generating Frequent Item Sets", In: Proceeding of IJARCSSE, ISSN 2277-128X, Vol. 3, Issue 4, April 2013.

[9] Chen, Jianqing; Jan Stallaert (2014). "An Economic Analysis of Online Advertising Using Behavioral Targeting". MIS Quarterly 38 (2): 429–449.

[10] http://www.statista.com.

[11] http://en.wikipedia.org/wiki/Apriori_algorithm#cite_note-apriori-1

[12] Hierarchical Temporal Memory including HTM cortical Learning Algorithms version 0.2.1, September 12, 2011.Numenta,Inc.2011. http://numenta.org/htm-white-paper.html

8. ACKNOWLEDGMENTS

It gives us great pleasure and satisfaction in presenting this project on “Detecting the age of a person through web browsing patterns”.

We are taking this opportunity to express our profound gratitude and deep regards to our guide for her exemplary guidance, monitoring and constant encouragement throughout the course of this thesis.

We would like to thank all those, who have directly or indirectly helped us for the completion of the work during this project.