IJARIIT INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume 4, Issue 3)
Available online at: www.ijariit.com

Age detection system based on web browsing pattern

Shivam B. Rathod
rathod.shivam289@gmail.com
Shram Sadhana Bombay Trust's College of Engineering and Technology, Jalgaon, Maharashtra

Priyanshu Arora
priyanshu11arora@gmail.com
Shram Sadhana Bombay Trust's College of Engineering and Technology, Jalgaon, Maharashtra

Reshma Patil
rkpatil3197@gmail.com
Shram Sadhana Bombay Trust's College of Engineering and Technology, Jalgaon, Maharashtra

Harshal R. Patil
harshal612patil@gmail.com
Shram Sadhana Bombay Trust's College of Engineering and Technology, Jalgaon, Maharashtra

ABSTRACT

Today almost everyone around the globe uses the internet. These internet users have different attributes like age, browsing behavior, gender, location, etc. that plays an important role for different business enterprises to target their users and also helps in providing the better user experience. Earlier researches demonstrate that the user browsing behavior is related to their basic attributes. In this project, an approach has been used to detect the age group of users based on their browsing patterns using a data mining algorithm. Age information of different users is inferred using the web browsing behavior which varies for different users.

Keywords: *Browsing behavior, FP growth, Age detection technique.*

1. INTRODUCTION

Many of the network services like searching the websites, social media, etc., now give an improved user experience by giving more consideration to different services. Like My Yahoo and Google allow users to personalize explicitly their choices by only showing the interesting areas and information. This will help the different organizations to acquire the browsing pattern data so as to target their users. Generally, it is not easy to find demographic information. The diversity of online browsing behavior of the user helps in guessing the unrecognized user demographic attribute such as age.

To find the demographic information is not so simple. Some users of the internet are afraid to disclose this type of confidential information to free to all. In commerce as well as in academic circles there is a large interest to assume the demographic information of the users. The earlier study is on demographic prediction in which the more attention gives to designing the variation of linguistic writing and also the style of speaking which is related to the demographic attributes that generally with the user gender. There is a major difference between both, the style of writing and also the content among the authors having different age is analyzed by the Koppel.

The Internet is the basis for business for most of the enterprises. As a large number of user’s use of the internet, their browsing pattern is required by these organizations and enterprises. They help to enhance the experience of various users. Along with this it also provides new and improved services in order to support the pre-existing customer as well as attracting new ones simultaneously. Also, behavior addressing provides the facility to the organizations to target the right users. Also, there are many websites which are not appropriate for the children below 18 years of age. Thus it helps in examining the complications in guessing the age of the user of the websites dependent on their browsing activities.

The main issue is to detect the age of the users according to their search patterns. Earlier the problem of detecting the age is done using a regression model. The main focus was on designing the different linguistics writing and way of speaking is associated with the demographic feature. The solution is based on dividing the users into two different categories depending on their browsing behavior as well as profile.

In this paper, the website users’ age group is detected based on their browsing behavior, where the website viewed serves as the basis to predict the age of users. In the proposed solution, the information is taken from two different user groups, adults, and youngsters. Then it is analyzed to predict the age group of the user, they are divided into sessions then FP-growth algorithm is applied to it. It uses a pre-computed comparison of internet activities of the internet users. Based on the above method, the given solution results in giving a good accuracy.

2. RELATED WORK

All Business Enterprise want to focus on their user which provides them profit. As all business nowadays is mostly processed online so the target users should be done on online resources, i.e. through their web behavior. Basic attributes of web behavior are age, gender, location etc. but this system focuses on the age of the consumer. In order to provide various organizations information of user's age the project focuses on a system which will provide an age for the different users over the internet. According to the usage history of the user the age of the user is determined. The System allows the user to browse and view their history. The proposed software uses this history to categories the URL according to their type and the URL which are used by almost all age group is filtered out.

Using Regression model Smith, Rose frames and Nguzen predict the age of the text writer. Telephone conversations, online forum data, and blog data are used to create a new data set. Data from all sources are combined in a model and that model is trained by the domain adaptation technique. It also works separately on every data source. Previous Studies of demographic prediction basically grabs attention on the variety of the linguistics writing and speaking styles related to the demographic characteristic that also precisely involve age of the user. According to Koppel there are large deviations in both typing style and content between authors of varying ages. On the basis of these differences on blogs content and style, Multi-Class Real Winnow algorithm was used by them to study models that group blogs according to the writer gender and age. Hu, Zeng, Niu, and Chen researched the trouble of predicting internet user's demographic attributes like age based on their browsing pattern, in which the browsed websites information is treated as a cached variable to predict the age of any users.

3. METHODOLOGY

Browser History is collected and stored in the log file. Browsing history consists of URL of Web-site, time and date of accessing the web-site, and IP address of the user. The log file is separated into different session files which are created by differentiating the IP addresses or by the division of websites accessed in 30 minutes. The Preprocessing Database is used to store the data of the different sessions or users.

3.1. Session Algorithm

The system starts with an unsorted log file which serves as an input file. Then a session algorithm is applied to the given input file for log cleaning and conversion:

```
Input: Weblog file
Output: Relevant records saved in DB log
Method:
For each Record in Weblog file
  Read fields
    If fields= {*.gif,*.jpg, *.css} OR {404,500} then
          >>Remove Records
    Else
        >>Save Records in the log table
    End if
Until no more Records
```

Then an algorithm is applied for identifying the user and computing the sessions.

```
Input: Relevant records saved in DB log
Output: Set of sessions
Method:
    For each Record in DB
      Repeat steps
        Compare ip-address of first entry with ip-address of the second entry
        If both are same, identify both entries are from the same user
     Until last entry
     For each user
        Order records by the time
        Identify <=30 minutes entry from the first entry of web page && minimum 5 page views in a session
     Until last entry
```

The sessions files are then combined into single data set in comma separated format. Then this data set is provided as input to the FP growth algorithm to predict age. This algorithm uses a basis for predicting the age group of the users which is shown in a graph. Figure 1 shows the comparison of internet activities among various users of both the categories [1].

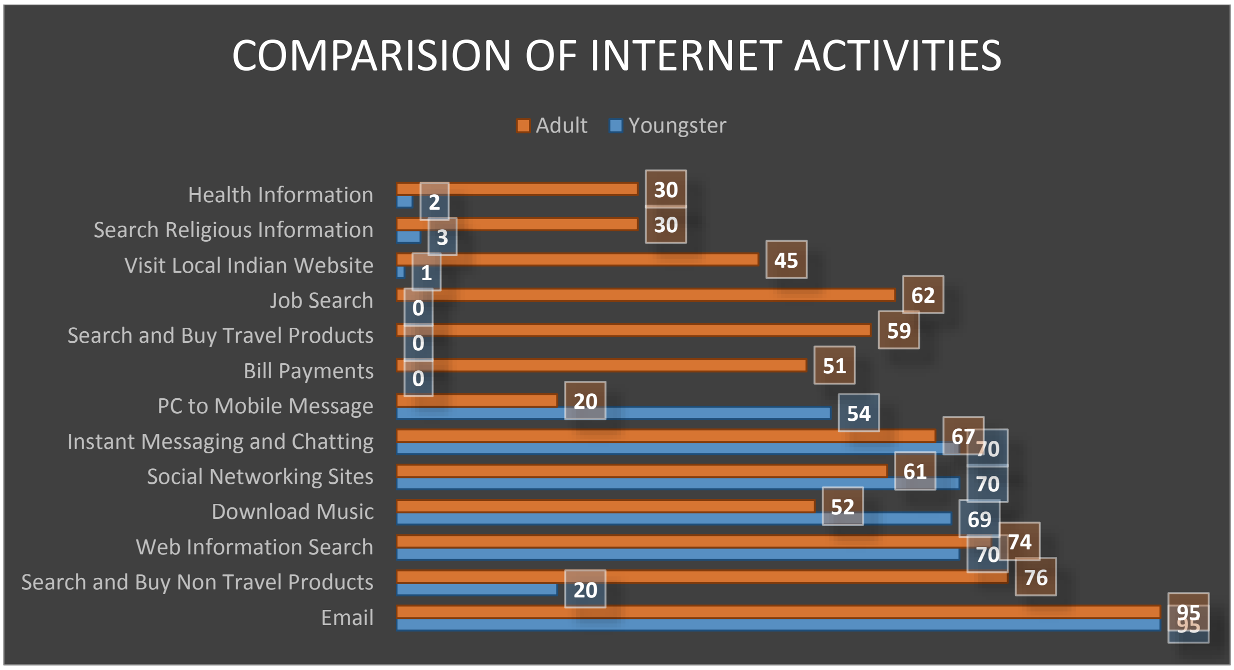


Figure 1: Comparison of internet Activities

3.2. FP Growth algorithm

Frequent pattern growth algorithm uses different techniques for mining frequent pattern from a database. It is an improvement to Apriori algorithm. Unlike Apriori it does not involve candidate item set generation. This process includes two steps in which first step builds an FP tree using two passes while the second step involves the extraction of frequent patterns from FP tree [4].

```
Input: constructed FP-tree
Output: complete set of frequent pattern
Method: call FP-growth (FP-tree, null)
Procedure FP-growth (Tree, α)
{
    1) If Tree contains a single path P then
     2) For each combination do generate pattern β
α with support = minimum support of nodes in β.
     3) Else for each header ai in the header of Tree
Do {
     4) Generate pattern β = ai α with support = ai.support;
    5) Construct β.s conditional pattern base and then β.s conditional FP-tree Tree β
    6) If Tree β= null
     7) Then call FP-growth (Tree β, β)}
}
```

3.3. Comparison of FP tree with Apriori algorithm

In this paper FP tree is used for data mining over Apriori because it is better than the latter in many aspects. The storage structure of the FP growth is tree based which is better than the array-based storage structure used in Apriori. There are only two scans in FP growth. It uses less memory space as compared to the latter. Even the run time is also less [3].

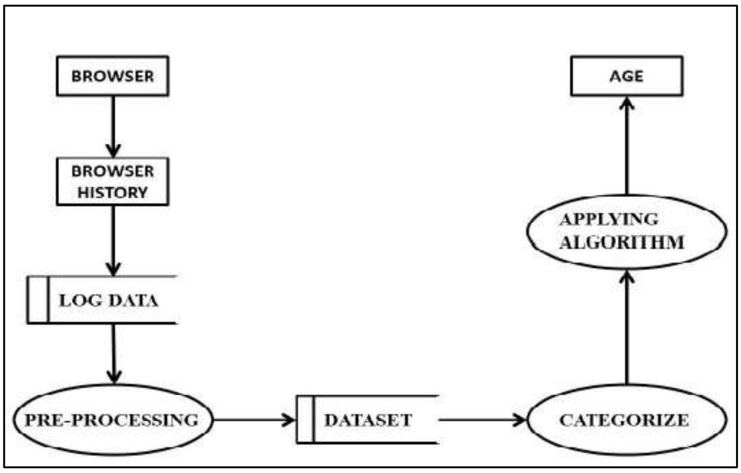
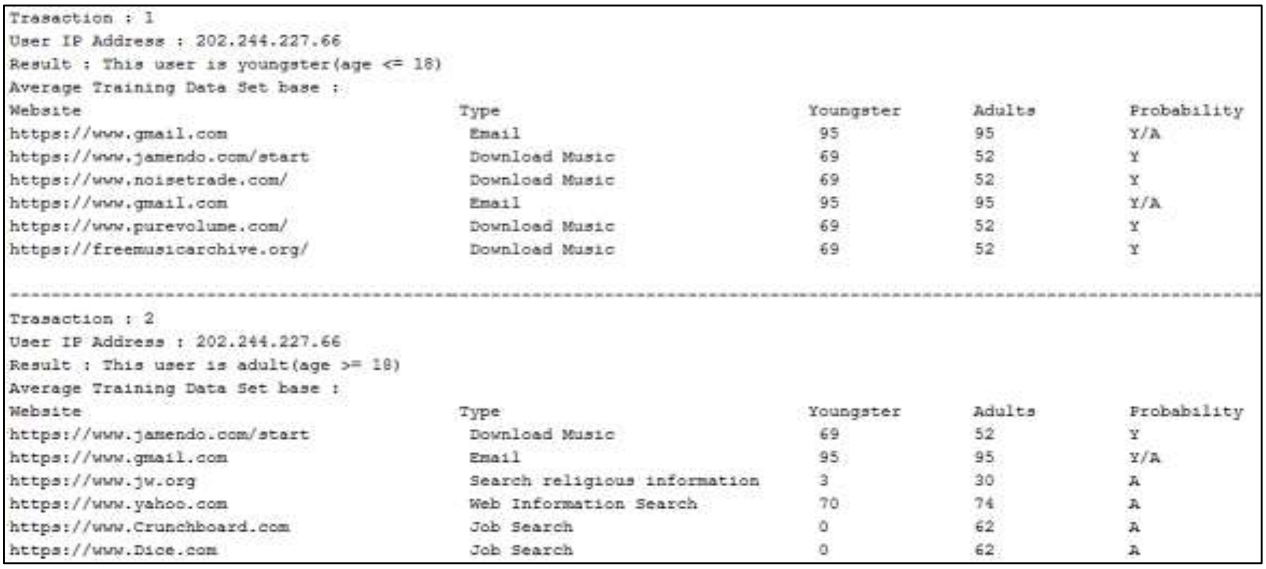


Figure 2: Process Flow

4. RESULTS AND DISCUSSION

4.1. Results



```
Trasaction : 1
User IP Address : 202.244.227.66
Result : This user is youngster(age <= 18)
Average Training Data Set base :
Website                                 Type                              Youngster      Adults        Probability
https://www.gmail.com                    Email                             95            95            Y/A
https://www.jamendo.com/start            Download Music                    69            52            Y
https://www.noisetrade.com/              Download Music                    69            52            Y
https://www.gmail.com                    Email                             95            95            Y/A
https://www.purevolume.com/              Download Music                    69            52            Y
https://freemusicarchive.org/            Download Music                    69            52            Y

----------------------------------------------------------------------------------------------------------------------
Trasaction : 2
User IP Address : 202.244.227.66
Result : This user is adult(age >= 18)
Average Training Data Set base :
Website                                 Type                              Youngster      Adults        Probability
https://www.jamendo.com/start            Download Music                    69            52            Y
https://www.gmail.com                    Email                             95            95            Y/A
https://www.jw.org                       Search religious information      3             30            A
https://www.yahoo.com                    Web Information Search            70            74            A
https://www.Crunchboard.com              Job Search                        0             62            A
https://www.Dice.com                     Job Search                        0             62            A
```

Figure 3: Result

Figure 3 displays the result of the predicted age group of the user. The target users initially were grouped according to the sessions and then data mining techniques are applied on it. So, the URL’s are categorized according to their type and whether they are used by adults or youngsters. Then the count is analyzed for youngsters and adults and whichever is greater that session’s user is declared as youngster or adult accordingly.

Through this system, a method has been proposed for predicting the user's age group based on their web browsing patterns using FP growth algorithm. It includes the data mining techniques which has minimum support and minimum confidence concepts used. If implemented as a web service, this system can be used by online advertisement agencies for improving the efficiency of their advertisements.

5. REFERENCES

[1] D. U. Misha Kakkar 2013.Web browsing behaviors based age detection," International Journal of Soft Computing and Engineering, vol. 3, p. 3.

[2] P. B. T. S. M. R. S. K. R. Chinmay Prakash Swami, Nuzhat Faiz Shaikh 2015. Detecting the age of a person through web browsing patterns," International Journal of Computer Applications, vol. 118, no. 12, p. 5.

[3] M. S. Mrs. M.Kavitha2016. Comparative study on apriori algorithm and fp growth algorithm with pros and cons," International Journal of Computer Science Trends and Technology, vol. 4.

[4] A. S. A. Alghamdi 2011.E_cient implementation of fp growth algorithm-data mining on medical data," International Journal of Computer Science and Network Security, vol. 11.