INDIAN INSTITUTE OF
INFORMATION TECHNOLOGY,

Project Report

Building User Personality Profile Using Web Interaction Data

Supervisor: Prof. Dr. Om Prakash Vyas

Prepared By:

Charul (IIT2011141)
Rahul Ranjan (IIT2011136)
Aditya Chaturvedi (IIT2011102)

CANDIDATE'S DECLARATION

Date: 07/10/2013

We hereby declare that the work presented in the project report entitled "**Building User Personality Profile using Web Interaction Data**", submitted for 5th semester mini-project mid-term evaluations of B.Tech(IT) at the Indian Institute of Information Technology, Allahabad, is our original work and will be carried out for a semester until 27th November, 2013 under the guidance of **Prof. Dr. O.P.Vyas.**

Further, due acknowledgements has been made in the report to all concerned and other material used from various resources. The project was done in full compliance with the requirements and constraints of the prescribed curriculum.

Charul (IIT2011141)
Rahul Ranjan (IIT2011136)
Aditya Chaturvedi (IIT2011104)

Place: Allahabad.

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY ALLAHABAD

(Deemed University)

(A centre of excellence in IT, established by Ministry of HRD, Govt. of India)

Date: ______________

This is to certify that this project work entitled “**Building User Personality Profile using Web Interaction Data**” which is submitted by Charul, Rahul Ranjan and Aditya Chaturvedi as their mid-semester report to the Indian Institute of Information Technology, Allahabad is a bonafide record of research work carried out by them under my supervision.

____

Prof. Dr. O.P.Vyas

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY,

ALLAHABAD

(Deemed University)

(A centre of excellence in IT, established by Ministry of HRD, Govt. of India)

Date: ______________

This is to certify that project Report entitled "**Building Personality Profile using Web Interaction Data**" which is submitted by us is the fulfillment of the requirement for the Mini-Project to Indian Institute of Information Technology, Allahabad comprises only our original work and due acknowledgement has been made in the text to all other material used.

Charul	IIT2011141
Rahul Ranjan	IIT2011136
Aditya Chaturvedi	IIT2011102

Table of Content

Abstract

Today web is a place where people spend considerable amount of time. Subsequently, they reveal a lot of their personal details and insights of their lives. We are beginning to understand how some of this interaction data can help us to improve user experience on web and in a way to achieve that user's personality determination is the key aspect of making a behavior model out of it.

Personality has been shown to be relevant to many types of interactions; it has been shown to be useful in predicting job satisfaction, professional and romantic relationship success, and even preference or different interfaces. Until now, to accurately gauge users' personalities, they were required to take a personality test. This made it impractical to use personality analysis in many web-oriented service domains.

So in this Project we are developing a completely new methodology to determine the personality of a Person using his/her interaction data on the web.

1. Introduction

The rapid global growth of internet usage has reinforced the need to study the psychological, social, and economic implications of web services on users. Internet world provide a new outlook for investigating user's need to provide personalized services. Such platforms are programmable, allowing the development of data collection tools to record various behavioral aspects of the user, ranging from how the user react across different contexts of stimulation to analyzing spatial and social dimensions of the everyday life of the user. From the point of view of designing communication features and applications that are tailored to the individual needs and preferences of a user, this data intensive framework provides a wealth of new opportunities, as it allows us to understand the impact of context on user behavior as well as to study individual differences, such as personality of users.

1.1 Overview

A person's value and preference are often reflected in his personality traits. In personality psychology, personality traits play a central role in describing a person. Personality is relatively stable and predictable. However, Personality is not rigid and unchanging; it is normally kept stable over a 45-year period which begins in early adulthood.

Many information systems are trying to determine the personality of the users in-order to customize their website that attracts customer spending and provide individualized products and service information. In this process, they are using traditional ways of determining personality through explicit methods in the form of questionnaire. The questionnaires used in many Big-Five personality studies are typically lengthy. This can be a limitation when a large number of participants at geographically spread areas have to complete questionnaires online. So we need implicit ways to predict the personality of the user.

We are determining the personality of the users through the activities of the user on an e-commerce website. These activities can be viewing a product, search a product, account creation, buying a product, add product to the cart. The methodology and framework proposed in this project can be further extended to any information system.

1.2 Motivation

The lack of cognitive attributes in the current User Models has motivated this project. Nowadays problems with personal psychological details are not the main concern for web system designers and programmers. Some research has been made by affective Computer scientists focusing mainly on the identification and modeling of user's Emotions [21, 22, 23].

Recently, studies from [22, 24, 25] have demonstrated how important psychological aspects of people such as Personality Traits and Emotions are during the human decision-making process. Human Emotion and their models have already been largely implemented in computers, much more than personality.

In personality psychology, personality traits play a central role in describing a person [26]. This topic has also been found to be of vital importance in computing. Several studies have been recently conducted on personality traits and their relationship to the use of Internet and forms of social media such as Youtube, blogs, Facebook and other social networks [27, 28, 29].

2. Problem Formulation

Determining personality of a person using conventional method of questionnaire, for instance TIPI, NEO-PI-R, has already been researched extensively by psychologist and henceforth many models have been proposed which consist different number of questions for calculating the same. In recent years Computer scientist have developed severals methods for Personality determination using "Frequency of key-strokes" and "Textual Analysis". All these are explicit methods and requires user to take up an additional assessment to model his personality profile. But implicit method such as "web-interaction data for assessing personality", is just beginning . Thus, our research question is:

2.1 Problem Definition

"How could we accurately build user personality profile using interaction data in order to integrate that in web-services to build large scale applications using reinforcement learning so that we can provide more personalized information, products or services for people?"

2.2 Scope

The framework developed has a wide scope in User personalization softwares like context-aware search engine, product recommendation, targeted online marketing, social matching and website content personalization. Also, above implementation will provide new horizon for data scientist to determine personality of user accurately and implicitly. Taking this as a basis, models for new platforms can be developed for various personality researches. The future scope of this project aims at providing a cognitive model to unify the various behavior attributes reflected by a user hence consolidating the methods of web personalization.

3. Literature Survey

Personality Schultz described Personality as "an enduring and unique set of characteristics that does not have any chance in response to different situations"[3].Personality accounts for consistently chosen pattern of mental reaction including behavior, emotions and thoughts over situations and time. Personality is more than just superficial physical appearance. Personality is relatively stable and predictable and is unchanged over a 45 year period which begins in young adulthood[2].Earlier more than 18 categories of Personality were described by researches.But each one describes alternative ways to present and differentiate human Personality. Personality can be considered an important aspect of human decision-making process.

Big-Five is currently the most widespread and generally accepted model of personality[5][6].It was developed using factor analytic techniques on adjectives and descriptive phrases of people culled from an English corpus. Big-five represents Personality at the broadest level of abstraction, and each dimension summarizes a large number of distinct, more specific Personality characteristics.The five personality dimensions according to Big-five are

- **Openness to experience** describes the breadth, depth, originality, and complexity of an individual’s mental life.
- **Neuroticism** relates to emotional stability.
- **Conscientiousness** describes socially prescribed impulse control that facilitates task and goal-directed behavior.
- **Agreeableness** is a tendency to be compassionate and cooperative rather than suspicious and antagonistic towards others.
- **Extraversion** implies an energetic approach to the social world and includes traits such as sociability and positive emotionality.

3.1 Personality Test

Questionnaire based: Researches proposed a wide range range of instruments to assess human Personality. Traits; among them most common ones were:

- 240-items NEO-PI-R (Revised NEO (Neuroticism-Extraversion-Openness) Personality Inventory)[7]
- 60-items NEO-FFI (NEO Five-Factor Inventory)[8]
- 44-items BFI (Big Five Inventory)[9]
- 10-items TIPI (Ten-Item Personality Inventory)[10]

We are using TIPI questionnaire to assess the personality of the users because it provides extremely brief measures of Big-Five Personality dimensions.

3.2 Related Work

Facebook Profile: There have been past work on determining personality using the properties of user's Facebook profile such size and density of their friendship network, number of uploaded photos, number of events attended, number of group membership and number of times user has been tagged in photos[11].

Smartphone Usage: Derived the relationship between automatically extracted behavioral characteristics derived from rich smartphone data and self-reported Big-Five personality traits. The features selected were call logs, SMS logs, apps logs, bluetooth logs, Profile logs(Normal,Silent, Beep, Ascending and Ring Once)[12].

Digital Records of Human Behavior: Determined people’s personal attributes such as sexual orientation, ethnicity, religious, happiness, relationship status, use of addictive substances, age and gender ranging from sexual orientation to intelligence using their Facebook Likes.[13]

Textual Data: The ways individuals use words can reflect basic psychological processes, including clues to their thoughts, feelings, perceptions, and personality.The paper focused on determining "personality" of the author through casual written text.They have used machine learning approaches [14].

3.3 Classification Methods

Fuzzy Multi-label Classifier: In multi-label classification, each object may belong simultaneously to many classes. For ML-RBF, the first layer of the corresponding neural layer is constructed by clustering instances of each possible class. the weights of the second layer are then optimized through minimizing function error function[15]

K- nearest neighbor is a widely used method for classifying objects based on closest training examples in the feature space. k-NN is a type of instance-based learning, where the function is only approximated locally and all computation is deferred until classification. This paper evaluates a novel k-NN classifier with linear growth and faster run-time built from binary neural networks[16].

3.4 Interaction Data:

Interaction Data is all the activities performed by a user on an a website. This can be recorded as order of occurrences of the events. For e.g. in an e-commerce website, this may include searching, buying, adding a product to the cart, buying a product, adding a product to wish-list, creating an account, sharing on Facebook, Twitter etc.

4. Proposed Methodology

4.1 Constrains

In this project, we aim to demonstrate the proposed methodology for an E-commerce platform. Hence we track the user interaction data which will comprise of user's activity on an e-commerce website which will include his browsing and buying history.

For this we contacted **SKBMART.com** who have agreed to provide us with the data of their users anonymized properly before delivery. (Privacy issues will be later discussed in detail).

4.2 General Representation of Model

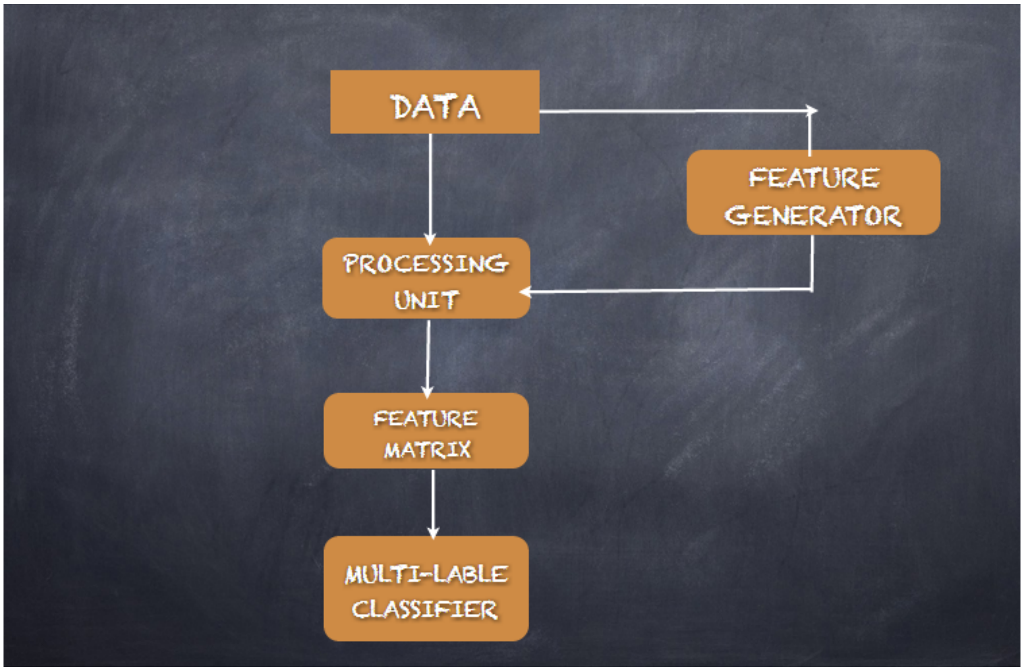


Fig 1:Diagram depicting the model of the framework

4.3 Components

Implementation of the proposed model comprises of the following:-

1. Data Collection System

2. Data Parser

3. Feature Set Generator.

4. Feature Matrix Generator.

5. Questionnaire processing engine.

6. Multi Label Fuzzy Classifier.

4.4 Data

- *Nature:*

The interaction data of user recorded per session is real time and continuous in nature. It tracks up various events that occur along the browsing session of user on our website.

To simplify things up, we have restricted our model to track the most fundamental events that can occurs in our e-commerce website.

Some of which may include-

- Browse a product
- Search a product
- Wish-list a product
- Add to cart
- Buy a product
- Share on Social network

- Write a review

- *Collection*

A real time client side analytics is required to be deployed to track the interaction data of a user on the web. This analytics engine should be able to identify the user and store all the data on cloud.

To generate training data, we are asking some distinct users to participate in the standard personality test which uses TIPI Inventory. A questionnaire of 10 questions is asked from the user as a part of the survey and his explicit personality ratings and calculated and stored along with his interaction data on the website.

4.5 Feature Set

The feature set for this model are devised using two approaches:

1. Intuitive Approach:

Any simple e-commerce system has a fundamental set of activities a user can perform. The entire session log of a user will comprise of a permutation of such activities. Thus, a sequence string can be generated which will determine the entire activity history of a user in a definite session.

Thus, by intuition, we can state that the a user's personality should be dependent on the activities he can perform on the web application. Hence, a naive approach will be to explicitly hardcoded the activity set and the feature set.

2. Enhanced approach:

In this approach, we decline the hypothesis that the personality depends on the individual activities of user on the web and instead look upon it as a contextual parameter, i.e not only the activities as an individual entity is relevant but also, the sequence of activities performed by a user is significant in determining his personality values.

Here we reject the possibility of activities being an independent entity in determining the personality of a user instead consider them as dependent attributes.

Thus, we follow a process to extract features from the data itself.

4.6 Feature extraction

To extract features from the data, a number of standard data mining techniques can be used. The method we propose here, is frequent pattern mining.

Frequent Pattern Mining:

To generate frequent occurring patterns of activities of users on e-commerce website. Using Apriori algorithm for frequent pattern mining.

4.7 Data Parsing

This is used to generate a Feature matrix from the raw data for each object set. The parse engine uses the Feature set and interaction string as input and returns the feature matrix as output. The raw data will be converted to input data for each object (in this case, a user) using this method.

4.8 Fuzzy multi label classifier

Personality of each individual may belong to each of the five classes of Big Five personality. We will process the data and implies the degree of belongingness to each class.

5. Hardware and Software requirements

Programming Languages:

Python, JSON and Javascript.

Python Libraries:

NumPy: It is the fundamental package for scientific computing with Python. It contains among other things: a powerful N-dimensional array object, sophisticated (broadcasting) functions, tools for integrating C/C++ and Fortran code useful linear algebra, Fourier transform, and random number capabilities. Besides its obvious scientific uses, NumPy can also be used as an efficient multi-dimensional container of generic data. Arbitrary data-types can be defined. This allows NumPy to seamlessly and speedily integrate with a wide variety of databases.

SciPy: It is stack, a collection of open source software for scientific computing in Python, and particularly a specified set of core packages.

JavaScript Libraries:

jQuery is a fast, small, and feature-rich JavaScript library. It makes things like HTML document traversal and manipulation, event handling, animation, and Ajax much simpler with an easy-to-use API that works across a multitude of browsers. With a combination of versatility and extensibility, jQuery has changed the way that millions of people write JavaScript.

IDE:

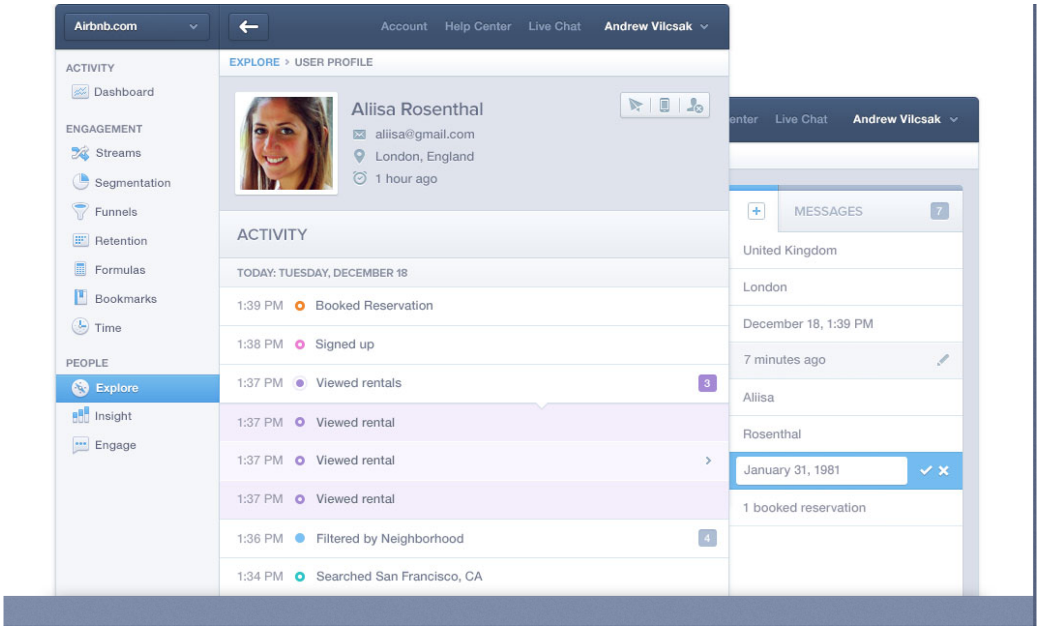
Eclipse, Aptana

Web-Analytics:

Google Analytics: It is a service offered by Google that generates detailed statistics about a website's traffic and traffic sources and measures conversions and sales. The product is aimed at marketers as opposed to web-masters and technologists from which the industry of web analytics originally grew. It is the most widely used website statistics service.

MixPanel Analytics:

The MixPanel Platform essentially gives websites a plug and play analytics offering for their users. Developers insert a few lines of code, and then users can access MixPanel's variety of realtime analytics for their services. The platform offers the ability to track how many comments, subscribers, likes, shares, and page views users are getting. MixPanel will place all of this data on a dashboard for a business' users to check and monitor.



Aliisa Rosenthal
aliisa@gmail.com
London, England
1 hour ago

ACTIVITY

TODAY: TUESDAY, DECEMBER 18

1:39 PM Booked Reservation
1:38 PM Signed up
1:37 PM Viewed rentals 3
1:37 PM Viewed rental
1:37 PM Viewed rental
1:37 PM Viewed rental
1:36 PM Filtered by Neighborhood 4
1:34 PM Searched San Francisco, CA

United Kingdom
London
December 18, 1:39 PM
7 minutes ago
Aliisa
Rosenthal
January 31, 1981
1 booked reservation

Fig 2 User activity feed in Mix-Panel Analytics

6. Analysis and Result

Modeling the Framework:

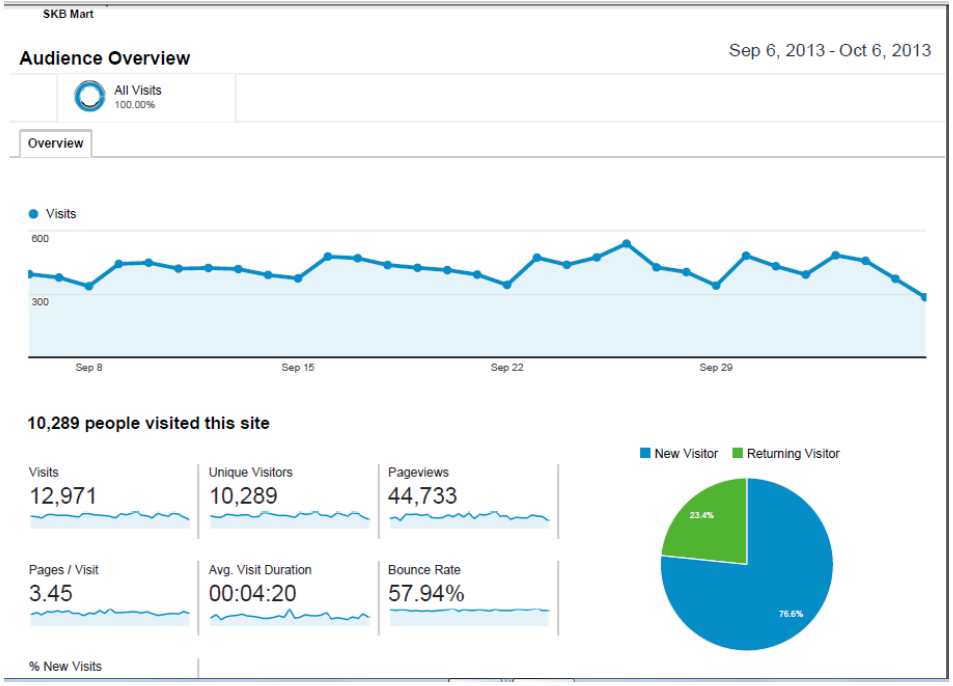
We performed a detailed literature survey on the previous work done in the field of personality study and extraction methods. These helped us in devising a model of our own which we consider will be applicable in determining the personality profile of a person implicitly in any web service.

Formulating Activity Set:

We formulated a set of activities which a user can perform on our e-commerce web platform. Further, we setup event tracking mechanism in analytic engines for the same.

Deployment of Analytics:

Initially we deployed Google analytics engine on our website SKBMART.com to track user interaction data.



SKB Mart

Audience Overview

Sep 6, 2013 - Oct 6, 2013

All Visits 100.00%

Overview

10,289 people visited this site

Visits 12,971

Unique Visitors 10,289

Pageviews 44,733

Pages / Visit 3.45

Avg. Visit Duration 00:04:20

Bounce Rate 57.94%

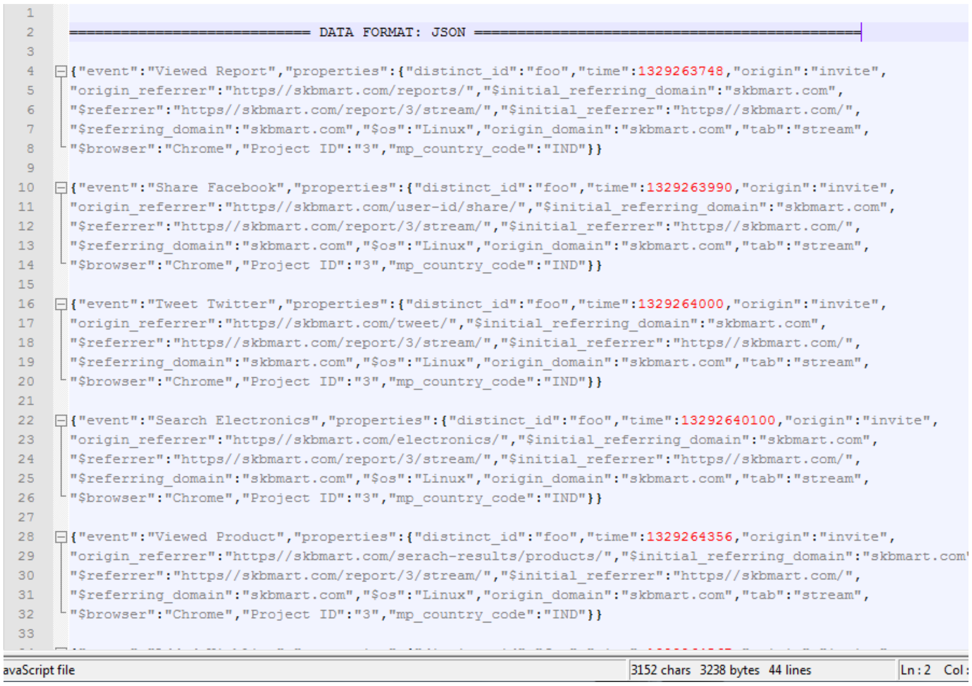
% New Visits

Fig 2 Google Analytics data of Skbmart.com

The nature of data collected was as below

We collected data of about 60 users. The problem with google analytics was that it failed to provide us with session based data per user. The free version of google analytics only provided with summary of the total interaction data of all users. This include page hits, custom funnels, event triggers, conversion and demographic details.

Later, we switched on to a more accurate analytics system called the 'MixPanel'. It is one of the most advanced analytics platform ever for mobile and the web. It helps in tracking various metrics of engagement and not just page views, thus generating an entire activity stream of each user accessing the website.



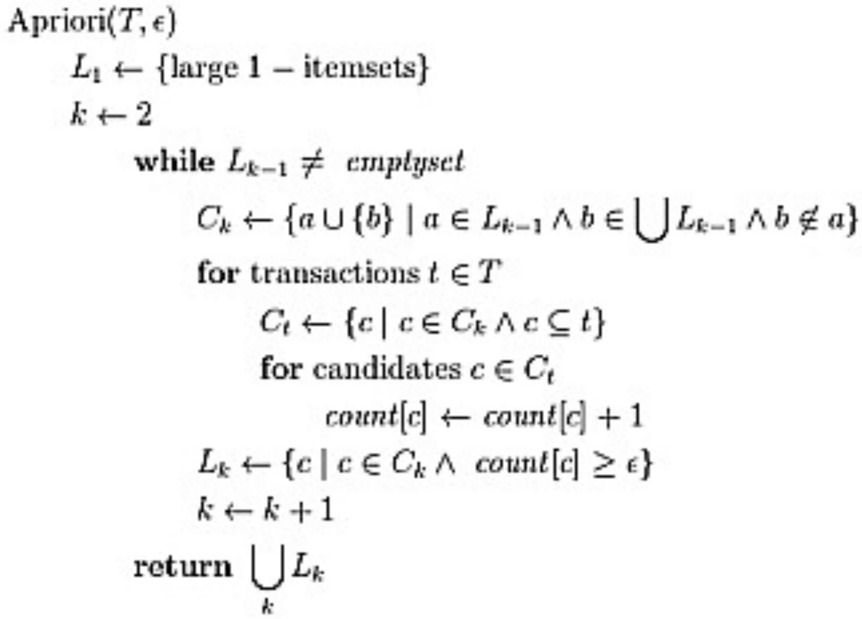
```

============================ DATA FORMAT: JSON =============================================

{"event":"Viewed Report","properties":{"distinct_id":"foo","time":1329263748,"origin":"invite",
"origin_referrer":"https//skbmart.com/reports/","$initial_referring_domain":"skbmart.com",
"$referrer":"https//skbmart.com/report/3/stream/","$initial_referrer":"https//skbmart.com/",
"$referring_domain":"skbmart.com","$os":"Linux","origin_domain":"skbmart.com","tab":"stream",
"$browser":"Chrome","Project ID":"3","mp_country_code":"IND"}}

{"event":"Share Facebook","properties":{"distinct_id":"foo","time":1329263990,"origin":"invite",
"origin_referrer":"https//skbmart.com/user-id/share/","$initial_referring_domain":"skbmart.com",
"$referrer":"https//skbmart.com/report/3/stream/","$initial_referrer":"https//skbmart.com/",
"$referring_domain":"skbmart.com","$os":"Linux","origin_domain":"skbmart.com","tab":"stream",
"$browser":"Chrome","Project ID":"3","mp_country_code":"IND"}}

{"event":"Tweet Twitter","properties":{"distinct_id":"foo","time":1329264000,"origin":"invite",
"origin_referrer":"https//skbmart.com/tweet/","$initial_referring_domain":"skbmart.com",
"$referrer":"https//skbmart.com/report/3/stream/","$initial_referrer":"https//skbmart.com/",
"$referring_domain":"skbmart.com","$os":"Linux","origin_domain":"skbmart.com","tab":"stream",
"$browser":"Chrome","Project ID":"3","mp_country_code":"IND"}}

{"event":"Search Electronics","properties":{"distinct_id":"foo","time":13292640100,"origin":"invite",
"origin_referrer":"https//skbmart.com/electronics/","$initial_referring_domain":"skbmart.com",
"$referrer":"https//skbmart.com/report/3/stream/","$initial_referrer":"https//skbmart.com/",
"$referring_domain":"skbmart.com","$os":"Linux","origin_domain":"skbmart.com","tab":"stream",
"$browser":"Chrome","Project ID":"3","mp_country_code":"IND"}}

{"event":"Viewed Product","properties":{"distinct_id":"foo","time":1329264356,"origin":"invite",
"origin_referrer":"https//skbmart.com/serach-results/products/","$initial_referring_domain":"skbmart.com'
"$referrer":"https//skbmart.com/report/3/stream/","$initial_referrer":"https//skbmart.com/",
"$referring_domain":"skbmart.com","$os":"Linux","origin_domain":"skbmart.com","tab":"stream",
"$browser":"Chrome","Project ID":"3","mp_country_code":"IND"}}

```

Fig 3 Data collected in JSON format

Implementing Feature Extraction:

We used the most standard Apriori algorithm for feature extraction. Apriori is a classic algorithm for frequent item set mining and association rule learning over transactional databases.

It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database.



```
Apriori($T, \epsilon$)
    $L_1 \leftarrow \{\text{large } 1 - \text{itemsets}\}$
    $k \leftarrow 2$
        while $L_{k-1} \neq$ emptyset
            $C_k \leftarrow \{a \cup \{b\} \mid a \in L_{k-1} \wedge b \in \bigcup L_{k-1} \wedge b \notin a\}$
            for transactions $t \in T$
                $C_t \leftarrow \{c \mid c \in C_k \wedge c \subseteq t\}$
                for candidates $c \in C_t$
                    $count[c] \leftarrow count[c] + 1$
            $L_k \leftarrow \{c \mid c \in C_k \wedge \ count[c] \geq \epsilon\}$
            $k \leftarrow k + 1$
        return $\bigcup_k L_k$
```

Output: Top Six are as follows:

{Search, Wishlist } = Support {0.32}
{Add to cart, Buy} = Support{0.63}
{Wishlist, Add to Cart} = Support{0.56}
{Search, share} = Support{0.37}
{Search, Search, comment} = Support{0.29}
{Search, View} = Support{0.79}

Implementing Data parsing:

We created a python engine to parse the data object. It connects to the mysql database which stores the data object and the feature set. The parser then generates a feature matrix using this input set and stores it back to the database.

The sample data set after parsing was as below:

For User_id = 46 ==>

SSSVSVSVWSVSVCVCBSVSVSVSSSS

S->Search , V- View, W-Wishlist, B-Buy, C-Add to Cart

Study of Classification Techniques

Classification is a form of data analysis that extracts models describing important data classes.Such models, called classifiers, predict categorical (discrete, unordered) class labels. For example, we can build a classification model to categorize bank loan applications as either safe or risky. Such analysis can help provide us with a better understanding of the data at large.Many classification methods have been proposed by researchers in machine learning,pattern recognition, and statistics. Most algorithms are memory resident, typically assuming a small data size.

Two Step Process:

Data classification is a two-step process, consisting of a learning step (where a classification model is constructed) and a classification step (where the model is used to predict class labels for given data).

Techniques:

Naive Bayesian classifiers

It assume that the effect of an attribute value on a given class is independent of the values of the

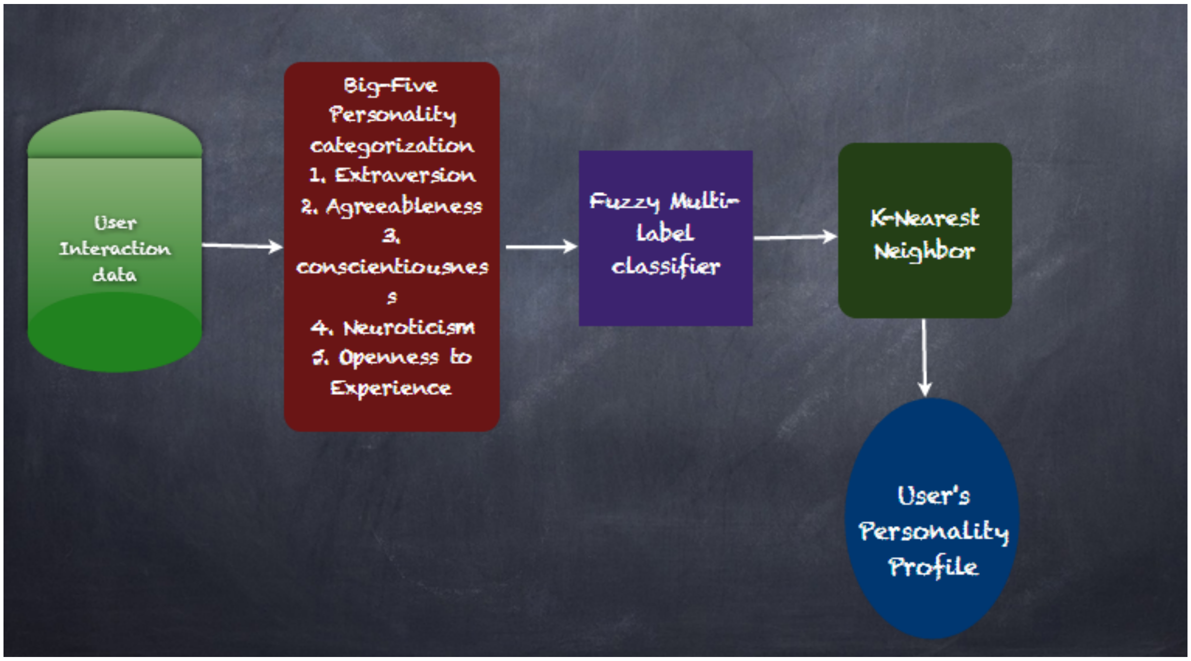

User Interaction data
Big-Five Personality categorization
1. Extraversion
2. Agreeableness
3. conscientiousness
4. Neuroticism
5. Openness to Experience
Fuzzy Multi-label classifier
K-Nearest Neighbor
User's Personality Profile

Fig 4 Personality Profile classification structure

other attributes. This assumption is called class conditional independence. It is made to simplify the computations involved and, in this sense, is considered “naıve” Bayes Theorem:

P(H | X) is the posterior probability, or a posteriori probability, of H conditioned on X.

$$P(H \mid X) = P(X|H) * P(H) / P(X)$$

Multilayer Feed-Forward Neural Network

A multilayer feed-forward neural network consists of an input layer, one or more hidden layers, and an output layer.Each layer is made up of units. The inputs to the network correspond to the attributes measured for each training tuple. The inputs are fed simultaneously into the units making up the input layer. These inputs pass through the input layer and are then weighted and fed simultaneously to a second layer of “neuron like” units, known as a hidden layer. The outputs of the hidden layer units can be input to another hidden layer, and so on. The number of hidden layers is arbitrary, although in practice, usually only one is used. The weighted outputs of the last hidden layer are input to units making up the output layer, which emits the network’s prediction for given tuple.

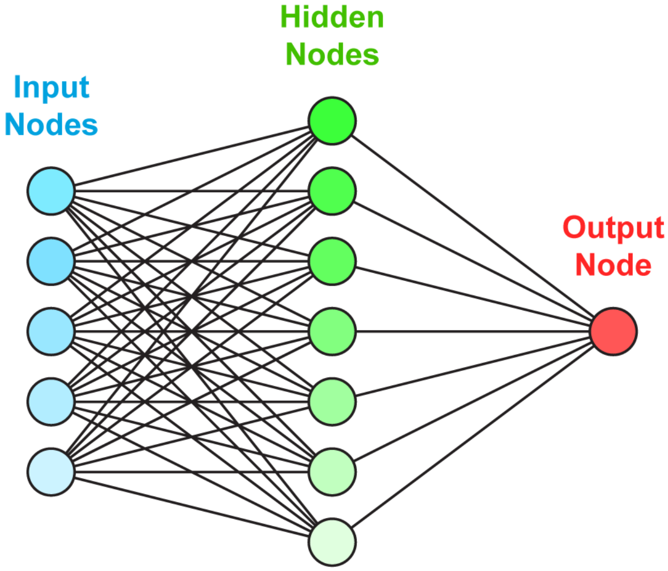


Fig 4 Diagram of a simple Neural Network

k-Nearest-Neighbor Classifiers

Nearest-neighbor classifiers are based on learning by analogy, that is, by comparing a given test tuple with training tuples that are similar to it. The training tuples are described by n attributes. Each tuple represents a point in an n-dimensional space. In this way, all the training tuples are stored in an n-dimensional pattern space.When given an unknown tuple, a k-nearest-neighbor classifier searches the pattern space for the k training tuples that are closest to the unknown tuple. These k training tuples are the k “nearest neighbors” of the unknown tuple.

$$O_k^T = f(net) = \frac{1}{1+\exp^{(-\beta net)}} \quad (\beta > 0) \qquad (7)$$

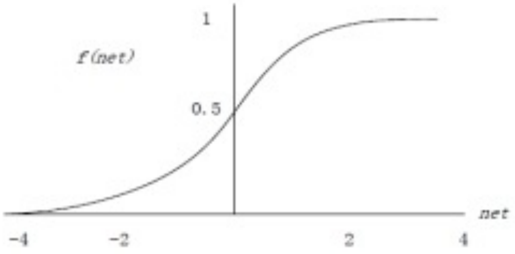


Figure 2 The Sigmoid Function [Source: Callan (1999)]

“Closeness” is defined in terms of a distance metric, such as Euclidean distance. The Euclidean distance between two points or tuples, say, a = a1, a2,...... , an and b = b1, b2, ... , bn is

D (a, b) = EuclidianDiastance(a, b)

Nearest-neighbor classifiers use distance-based comparisons that intrinsically assign equal weight to each attribute. They therefore can suffer from poor accuracy when given noisy or irrelevant attributes.

$$X_{n+1} = (aX_n + b) \bmod m$$

Sampling:

Object Sampling-

We have used 50 object set to train our Classifier network. These objects have been chosen randomly using the pseudo-random number generator algorithm.

Data Sampling-

Different users spend different amount of time on the website. Hence the number of activity count for each of them varied. To train our network, we need to normalize this value. For this, we took out the mean number of activities performed by a user according to our data.

Mean = 25 Standard deviation = 5

Now for each training set, a normalization is performed to pick 25 continuous activities. This is done by considering the median data and trimming the edge vectors.

Training:

Training is performed using Multi-layer Feed forward fuzzy classifier as discussed earlier.

Error Analysis:

Two error measuring criteria are used to evaluate the ANN models for the validation data sets. The first is the Mean Absolute Percentage Error (MAPE), and the second is the forecasting error.

The MAPE is defined as:

where Pi and Ai are the predicted personality rating and the actual personality rating of object i in the set of n objects.

The data model with a smaller MAPE is deemed superior. This error measurement attempts to produce a single number that represents the total error for all objects. This error measurement fails, however, to provide information as to how the error deviates between the objects.

Hence we use forecasting error for the object i which is defined as:

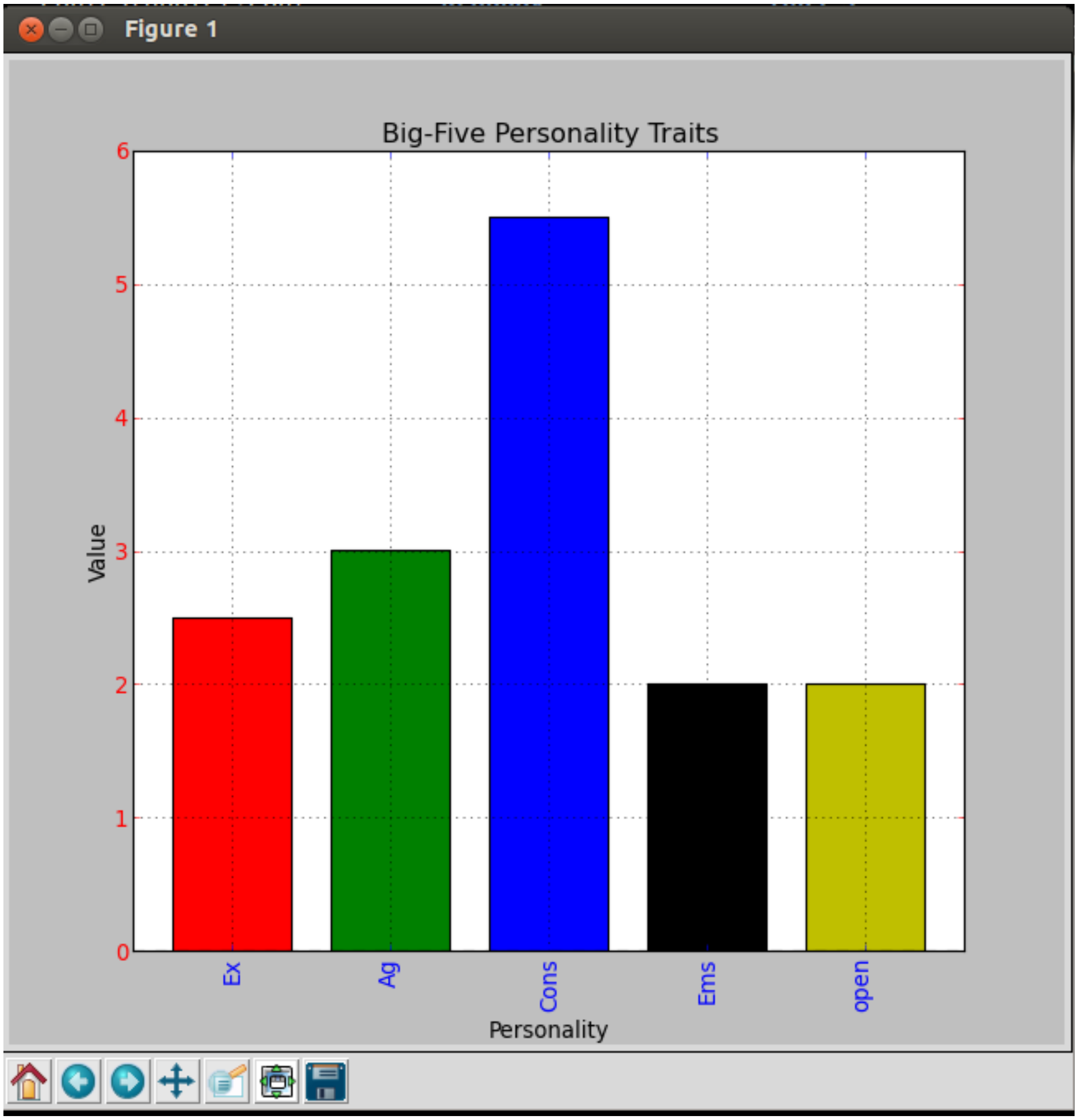
$$MAPE \equiv \left(\sum_{i=1}^{n} \left| \frac{(P_i - A_i)}{A_i} * 100 \right| \right) \div n,$$

Output for our data model is as follows:

MAPE = 20.83% which is more than the acceptable value for perfect data model (15%)

$$FE \equiv \left| \frac{(P_i - A_i)}{A_i} * 100 \right|.$$

User Personality profile for user_id = 46



Ex = Extraversion

Ag =Agreeableness

Cons = Conscientiousness

Ems = Emotional Stability(Neuroticism)

open = Openness to Experience

7. Conclusion

This study lays the basis for research in the prediction and usage of implicit techniques to determine personality traits of a user on web platforms. Our study presents a detailed analysis of the relationship between the personality inventories and aggregated user interaction data in predicting the the Big-Five personality traits. The methodology presented in this paper offers two main benefits. Firstly, the methods are easily scalable to major web service which rely on heavy user customization. Further, the features used are by nature privacy sensitive, which is of paramount importance in this area of research.

The results clearly show that several aggregated web interaction data could be predictive of the Big-Five personality traits. The analysis of this data also highlighted several interesting trends. Many of these trends conform with past work in psychology literature. It was found that extraverts, who are characterized by their outgoing nature, tend to search more and are less likely a potential buyer. On the other hand, users with high value of Conscientiousness are more likely to spend less time and use the service with a buying intent.

Subsequently, it was shown that a machine learning framework based on a supervised learning method can effectively classify an unknown user's Big-Five trait measures without relying on the personality inventories questionnaire as belonging to either the higher half or lower half of the population.

Regarding future work, in our opinion, this work shows the potential for further research into how personality traits can be predicted from web interaction data. Such a system can be extended for not only E-commerce, but other web domains like social networking, dating, tourism etc, where personal choice greatly influences user experience. Moreover, the study of a user's behavior on the web and his cognitive traits needs to be carefully addressed. While quantitative data analysis methods used in this study are suitable for highlighting statistical regularities, qualitative techniques are likely to be needed in order to obtain more insights on the reasons for individuals with a certain personality profile behaving in a given way.

8. References

[1] McCrae, R.R., John, O.P.: An introduction to the five-factor model and its applications.

[2] Stephen Soldz and George E.Vaillant. The Big Five Personality Traits and the Life Course: A 45-Year Longitudinal Study

[3] Duane Schultz. Theories of Personality

[4] Christine L. Lisetti. Personality, Affect and Emotion Taxonomy for Socially Intelligent Agents

[5] L.R. Goldberg. The structure of phenotypic personality traits

[6] O.P. John and S. Srivastava. The big five trait taxonomy: History, measurement and theoretical perspectives.

[7] P. T. Costa and R. R. McCrae. Revised neo personality inventory (neo-pi-r) and neo five-factor inventory.

[8] P. T. Costa and R. R. McCrae. Revised neo personality inventory (neo-pi-r) andneo five-factor inventory (neo-ffi)

[9] S. Srivastava. Measuring the big five personality factors., 2006. (Available in

http://www.uoregon.edu/ sanjay/bigve.html).

[10] Samuel D. Gosling, Peter J. Rentfrow, and William B. Swann Jr. A very brief measure of the Big-Five personality domains.

[11] Yoram Bachrach, Michal Kosinski, Thore Graepel. Personality and Patterns of Facebook Usage

[12] Gokul, Jan Blom, Daniel Gatica-Perez. Mining large-scale smartphone data for personality studies

[13] Michal Kosinski, David Stillwell, and Thore Graepel.Private traits and attributes are predictable from digital records of human behavior.

[14] Shlomo Argamon, Sushant Dhawle. Lexical Predictors of Personality Type.

[15] Zoulficar younes, Fahed Abdallah and Thierry Denoeux. Fuzzy Multi-label Learning Under Veristic Variables.

[16] Victoria J. Hodge and Jim Austin. A Binary Neural k-Nearest Neighbour Technique

[21] Daniel Rousseau and Barbara Hayes-Roth. A social-psychological model for synthetic actors. In AGENTS '98: Proceedings of the second international conference on Autonomous agents, pages 165{172, New York, NY, USA, 1998. ACM Press.

[22] Rosalind W. Picard. What does it mean for a computer to 'have' Emotions? In R. Trappl, P. Petta, and S. Payr, editors, Emotions in humans and artefacts, chapter 7, pages 213{235. A Bradford Book - MIT Press, Cambridge, Massachusetts, 2002

[23] Andrew Ortony. On Making Believable Agents Believable. In R. Trappl, P. Petta,and S. Payr, editors, Emotions in humans and artefacts, chapter 6, pages 189-211. A Bradford Book - MIT Press, Cambridge, Massachusetts, 2002.

[24] Robert Trappl, Sabine Payr, and Paolo Petta, editors. Emotions in Humans and Artifacts. MIT Press, Cambridge, MA, USA, 2003.

[25] Paul Thagard. Hot Thought: Machanisms and Applications of Emotional Cognition. A Bradford Book- MIT Press, Cambridge, MA, USA, 2006.

[26] R. R. McCrae and O. P. John. An introduction to the five-factor model and its applications. Journal.of Personality, 60:175-215, 1992.

[27] M. D. Back, J. M. Stopfer, S. Vazire, S. Gaddis, S. C. Schmukle, B. Eglo and S. D. Gosling. Facebook profiles reflect actual personality, not self-idealization. Psychological Science, 21:372-374, 2010.

[28] S. Counts and K. Stecher. Self-presentation of personality during online profile creation. In Proc. AAAI Conf. on Weblogs and Social Media (ICWSM), 2009.

[29] T. Yeo. Modeling personality influences on youtube usage. In Proc. Int. AAAI Conference on Weblogs and Social Media (ICWSM), 2010.