



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

## 1. Introduction

1.1 Overview

1.2 Motivation

## 2. Problem Formulation

2.1 Problem Definition

2.2 Scope

## 3. Literature Survey

3.1 Personality Test

3.2 Previous work

3.3 Classification Methods

3.4 Interaction Data

## 4. Proposed Methodology

4.1 Constraints

4.2 General representation of model

4.3 Components

4.4 Data

4.5 Feature Set

4.6 Feature Extraction

4.7 Data Parsing

4.8 Fuzzy multi-label class classifier

## 5. Hardware & Software requirements

## 6. Result and Analysis

7. Conclusion

8. References

## **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 several 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 change 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 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](https://www.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

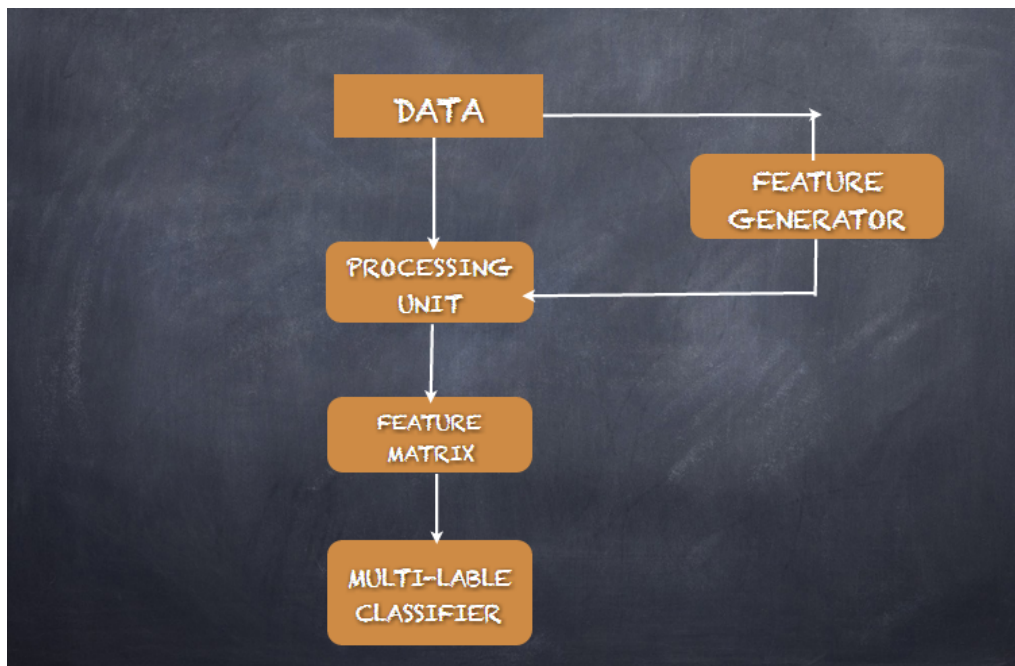


Fig 1:Diagram depicting the model of the framework

### 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 are 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.

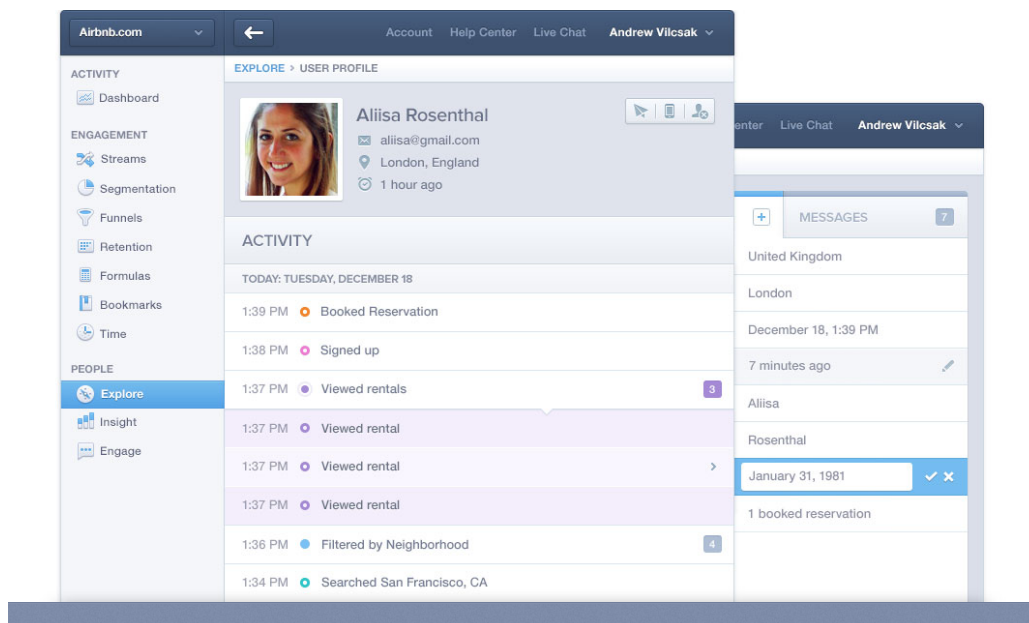


Fig 2 User activity feed in Mix-Panel Analytics

## 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](http://SKBMART.com) to track user interaction data.

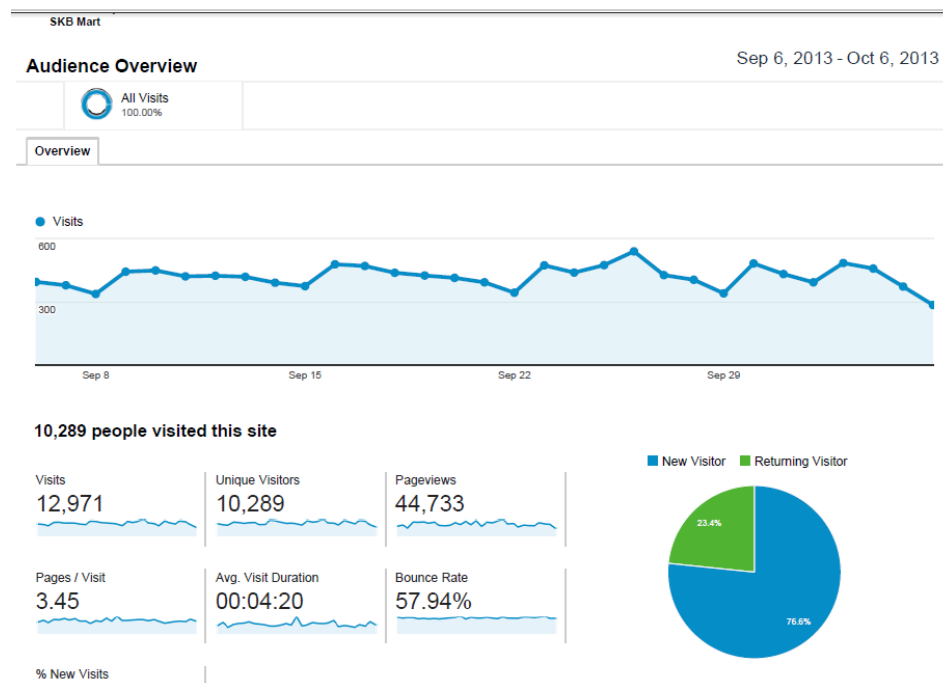


Fig 2 Google Analytics data of [Skbmart.com](http://Skbmart.com)

## 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.

```
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025
1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079
1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187
1188
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241
1242
1243
1244
1245
1246
1247
1248
1249
1250
1251
1252
1253
1254
1255
1256
1257
1258
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1289
1290
1291
1292
1293
1294
1295
1296
1297
1298
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349
1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403
1404
1405
1406
1407
1408
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457
1458
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511
1512
1513
1514
1515
1516
1517
1518
1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1560
1561
1562
1563
1564
1565
1566
1567
1568
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1579
1580
1581
1582
1583
1584
1585
1586
1587
1588
1589
1590
1591
1592
1593
1594
1595
1596
1597
1598
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1610
1611
1612
1613
1614
1615
1616
1617
1618
1619
1620
1621
1622
1623
1624
1625
1626
1627
1628
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1670
1671
1672
1673
1674
1675
1676
1677
1678
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727
1728
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1780
1781
1782
1783
1784
1785
1786
1787
1788
1789
1790
1791
1792
1793
1794
1795
1796
1797
1798
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1830
1831
1832
1833
1834
1835
1836
1837
1838
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889
1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1940
1941
1942
1943
1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2049
2050
2051
2052
2053
2054
2055
2056
2057
2058
2059
2060
2061
2062
2063
2064
2065
2066
2067
2068
2069
2070
2071
2072
2073
2074
2075
2076
2077
2078
2079
2080
2081
2082
2083
2084
2085
2086
2087
2088
2089
2090
2091
2092
2093
2094
2095
2096
2097
2098
2099
2100
2101
2102
2103
2104
2105
2106
2107
2108
2109
2110
2111
2112
2113
2114
2115
2116
2117
2118
2119
2120
2121
2122
2123
2124
2125
2126
2127
2128
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159
2160
2161
2162
2163
2164
2165
2166
2167
2168
2169
2170
2171
2172
2173
2174
2175
2176
2177
2178
2179
2180
2181
2182
2183
2184
2185
2186
2187
2188
2189
2190
2191
2192
2193
2194
2195
2196
2197
2198
2199
2200
2201
2202
2203
2204
2205
2206
2207
2208
2209
2210
2211
2212
2213
2214
2215
2216
2217
2218
2219
2220
2221
2222
2223
2224
2225
2226
2227
2228
2229
2230
2231
2232
2233
2234
2235
2236
2237
2238
2239
2240
2241
2242
2243
2244
2245
2246
2247
2248
2249
2250
2251
2252
2253
2254
2255
2256
2257
2258
2259
2260
2261
2262
2263
2264
2265
2266
2267
2268
2269
2270
2271
2272
2273
2274
2275
2276
2277
2278
2279
2280
2281
2282
2283
2284
2285
2286
2287
2288
2289
2290
2291
2292
2293
2294
2295
2296
2297
2298
2299
2300
2301
2302
2303
2304
2305
2306
2307
2308
2309
2310
2311
2312
2313
2314
2315
2316
2317
2318
2319
2320
2321
2322
2323
2324
2325
2326
2327
2328
2329
2330
2331
2332
2333
2334
2335
2336
2337
2338
2339
2340
2341
2342
2343
2344
2345
2346
2347
2348
2349
2350
2351
2352
2353
2354
2355
2356
2357
2358
2359
2360
2361
2362
2363
2364
2365
2366
2367
2368
2369
2370
2371
2372
2373
2374
2375
2376
2377
2378
2379
2380
2381
2382
2383
2384
2385
2386
2387
2388
2389
2390
2391
2392
2393
2394
2395
2396
2397
2398
2399
2400
2401
2402
2403
2404
2405
2406
2407
2408
2409
2410
2411
2412
2413
2414
2415
2416
2417
2418
2419
2420
2421
2422
2423
2424
2425
2426
2427
2428
2429
2430
2431
2432
2433
2434
2435
2436
2437
2438
2439
2440
2441
2442
2443
2444
2445
2446
2447
2448
2449
2450
2451
2452
2453
2454
2455
2456
2457
2458
2459
2460
2461
2462
2463
2464
2465
2466
2467
2468
2469
2470
2471
2472
2473
2474
2475
2476
2477
2478
2479
2480
2481
2482
2483
2484
2485
2486
2487
2488
2489
2490
2491
2492
2493
2494
2495
2496
2497
2498
2499
2500
2501
2502
2503
2504
2505
2506
2507
2508
2509
2510
2511
2512
2513
2514
2515
2516
2517
2518
2519
2520
2521
2522
2523
2524
2525
2526
2527
2528
2529
2530
2531
2532
2533
2534
2535
2536
2537
2538
2539
2540
2541
2542
2543
2544
2545
2546
2547
2548
2549
2550
2551
2552
2553
2554
2555
2556
2557
2558
2559
2560
2561
2562
2563
2564
2565
2566
2567
2568
2569
2570
2571
2572
2573
2574
2575
2576
2577
2578
2579
2580
2581
2582
2583
2584
2585
2586
2587
2588
2589
2590
2591
2592
2593
2594
2595
2596
2597
2598
2599
2600
2601
2602
2603
2604
2605
2606
2607
2608
2609
2610
2611
2612
2613
2614
2615
2616
2617
2618
2619
2620
2621
2622
2623
2624
2625
2626
2627
2628
2629
2630
2631
2632
2633
2634
2635
2636
2637
2638
2639
2640
2641
2642
2643
2644
2645
2646
2
```

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

SSSVSVSVWSVSVVCVCBSVSVSVSSSS

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

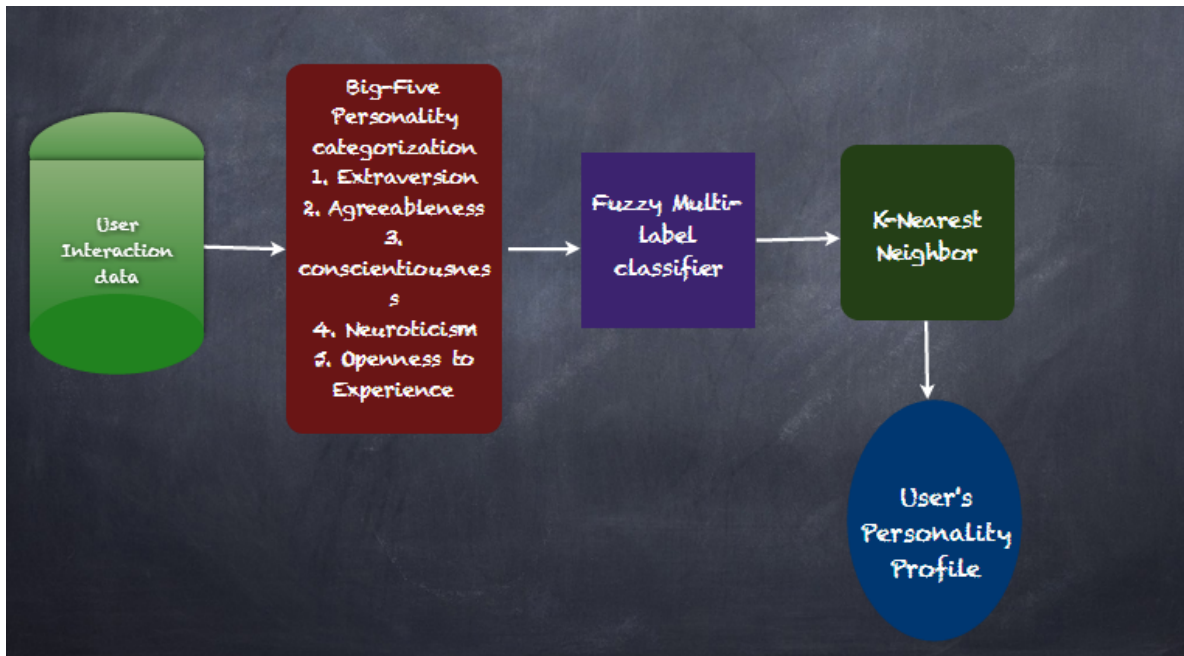


Fig 4 Personality Profile classification structure

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

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

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

### ***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.

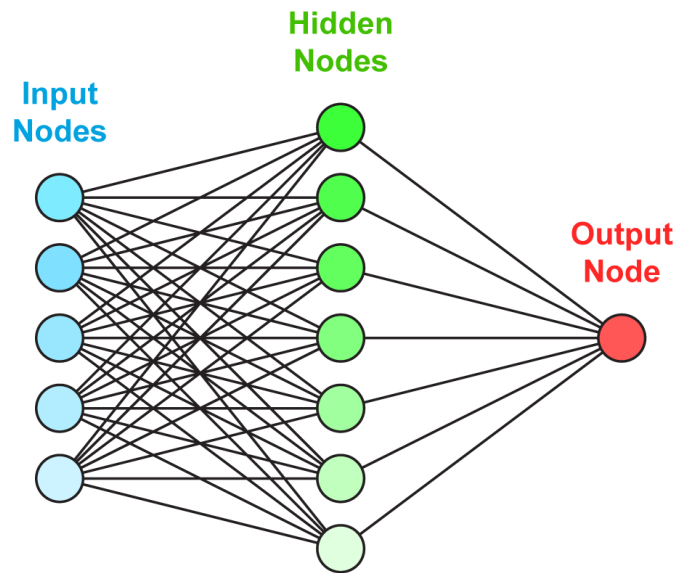


Fig 4 Diagram of a simple Neural Network

### ***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.

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

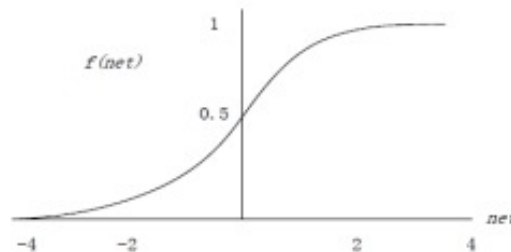


Figure 2 The Sigmoid Function [Source: Callan (1999)]

“Closeness” is defined in terms of a distance metric, such as Euclidean distance. The Euclidean distance between two points or tuples, say,  $a = a_1, a_2, \dots, a_n$  and  $b = b_1, b_2, \dots, b_n$  is



$$D(a, b) = \text{EuclidianDistance}(a, b)$$

Nearest-neighbor classifiers use distance-based comparisons that intrinsically assign equal

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

weight to each attribute. They therefore can suffer from poor accuracy when given noisy or irrelevant attributes.

### **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.

$$\text{Mean} = 25$$

$$\text{Standard deviation} = 5$$

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  $P_i$  and  $A_i$  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:

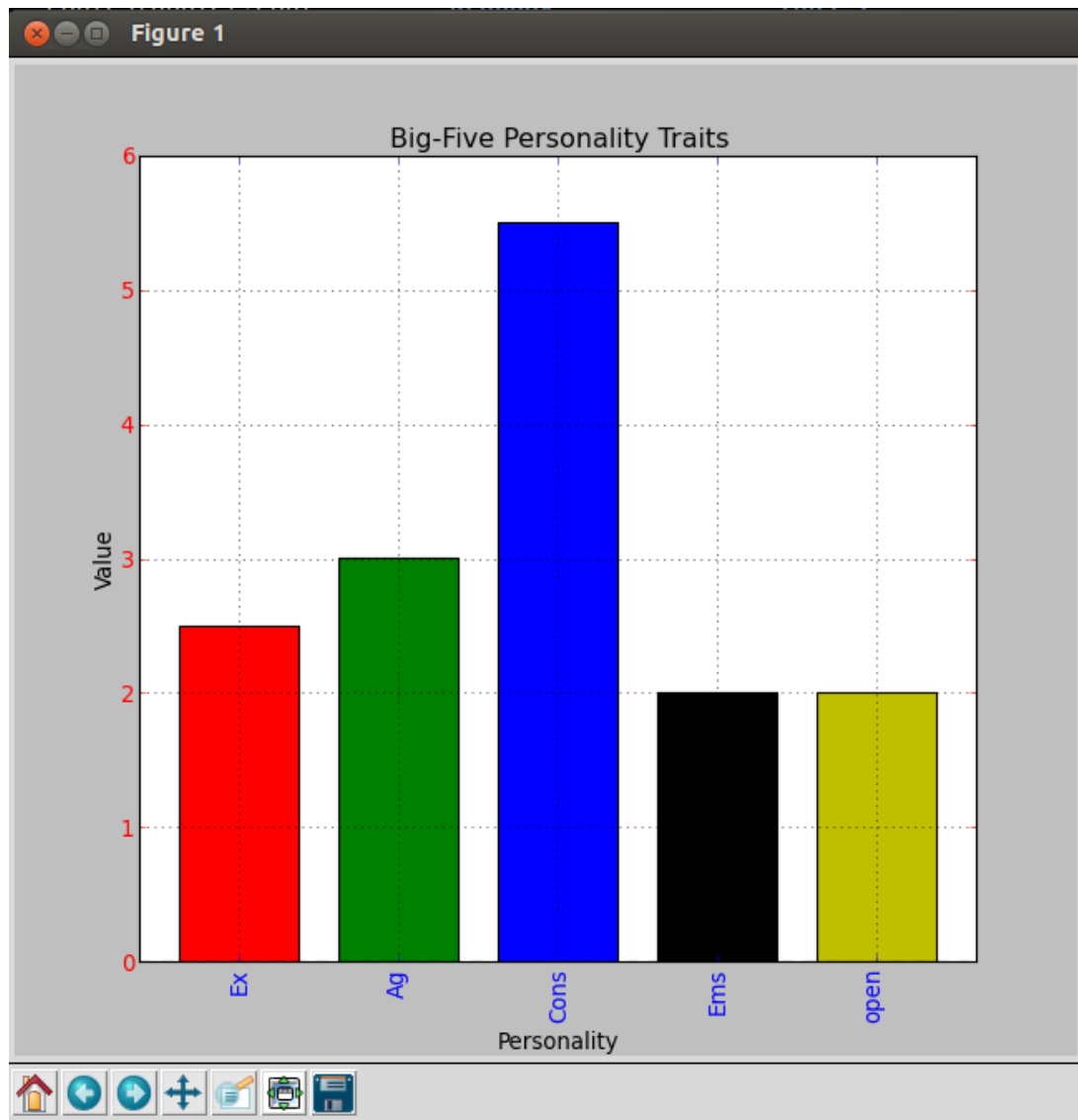
$$MAPE \equiv \left( \sum_{i=1}^n \left| \frac{(P_i - A_i)}{A_i} * 100 \right| \right) \div n,$$

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) and neo 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: Mechanisms 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.



