High Level Implementation and Framework of a Novel Algorithm for Prescriptive, Predictive and Preventive Analysis of Personality Data

Muhammad Uddin[1], Jeongkyu Lee[2]
Department of Computer Science and Engineering
University of Bridgeport, CT. USA

Big data mining and analytics have grown tremendously in the last 2 decades. It has positively affected many industries and research communities in utilizing the knowledge and valuable information extracted from unstructured or semi-structured big data, that we produce every second. With ease of use of social networking such as Facebook, Twitter and others, we produce tremendous amount of data that has tendency to understand people's personality. The personality if understood with more accuracy can play vital role in prescriptive, predictive and preventive analysis and decision making for many individuals. In this paper, we provide the high level framework of proposed model and one algorithm for PERFECT (Predicting Educational Relevance for an Efficient Classification of Talent) Engine, as our research work in progress. We emphasize on 3 Ps of analytics as main categories for our direction by concluding the future direction and real world applications.

Corresponding Author:
1. *Muhammad Fahim Uddin. - PhD candidate at the University of Bridgeport.*
2. *Dr. Jeongkyu Lee. – Advisor and Professor in the department of computer science and engineering*

Introduction

Our main motivation comes from the fact that data generation is real and is happening every second. We all are data points, data generators, data consumer and data collectors. With so much of raw data in hand, we need to improve the way we analyze and mine data and extract new knowledge and insights. Prediction in big data is a key feature that industry depends greatly upon. In our research, we take the personality prediction correlated with their role in academic and real world to provide useful insights for future.

The personal data in industry is rapidly growing into a multibillion business[1]. More firms are increasingly venturing into the enterprise in a move to expand their client bases. In business perspectives, entities view the strategic use of personal data as absolute measures of aligning products to consumer demands. This step is critical to creating a business niche. Similarly, hackers strategically use malware to steal personal data from various institutions[2]. Today's big data and social networking platform have revolutionized the way we exhibit ourselves and our personality attributes. The Five Factor Model (FFM)[1][2] of personality uses personal data to classify an individual into one of the OCEAN personality traits, classifications that can subsequently be exploited by business entities. The fascination of firms with personal data has seen a surge in the number of blog sites and news posts from various organizations[3]. However, further understanding of such data to extract specific knowledge and understanding of the industry is still an open research area.

The OCEAN model[4] analyzes the personality of technology users based on the five basic foundations which include *openness, conscientiousness, extraversion, agreeableness, and neuroticism*. While the five factor model is

[1] S. Kroft, "The Data Brokers: Selling your personal information", Cbsnews, 2014. [Online]. Available: http://www.cbsnews.com/news/the-data-brokers-selling-your-personal-information/. [Accessed: 05- Feb- 2016].

[2] J. Pagliery, "Criminals use IRS website to steal data on 104,000 people", CNNMoney, 2015. [Online]. Available: http://money.cnn.com/2015/05/26/pf/taxes/irs-website-data-hack/. [Accessed: 05- Feb- 2016].

used widely, it exhibits some challenges. For instance, the elements lack independence from each other. Some facets show a negative correlation as noted between neuroticism and extroversion. Also, the model fails to explain all human personalities. Some of the features that the OCEAN model fails to take into account include religiosity, sexiness, conservativeness and honesty[3]. Therefore, the sole reliance on data companies on the FFM traits model to understand user personalities is inaccurate. However, an integration of the FFM and personality features can provide an accurate reflection of the aspirations and objectives of an individual. Such proper understanding can be applied in guidance and counseling to enable learners to make good career choices. Besides, the integrated data model can be exploited to facilitate growth at personal and institutional levels.

7 V's of Big Data

In our previous publication[5], we extended the general model of 3 V's to 7 V's in one paper and context to elaborate the importance of it any and every data analytics challenges. We provide the snapshot in below table

Table 1 – 7 Vs of Big Data

Type	Description
Volume	The size it is being generated in.
Velocity	The speed it is being generated with
Variety	Various forms of data
Veracity	Truthfulness of data
Validity	Correctness and Accuracy of Data
Volatility	Retention relevance and importance
Value	Desired outcome

Related Study

FFM (OCEAN) [6][7] is widely used in research to predict and classify the various individual based on their behavior. *A study in* [1] discusses the measurement and Big 6 and Big 2. One of the challenges in FFM is different interpretation of its measurements that introduces some complexity.

A Study[2] prepared on Big 5 Inventory Report (BFI) for better assessment of personality traits. One of their focuses is on language translation for individuals from many other nations. They include understand and verify if FFM model is standing well across various cultures and validity of individual's profiles in such various nations.

Social Media including but not limited to Facebook, Twitter, LinkedIn, flicker, Pinit, Google+ and various blogs have great tendency to capture and store the personality features that are indirectly hidden the post user generates on daily basis, in various context of industry and discussion threads.

A *study done in* [8] performances a examination of behavior traits in predicting personality and analyze to what extent is the relevance. They utilize Facebook and Twitter data as test data. They conclude the same success in prediction personality from behavior analysis as compared to results that can be achieved using text analysis. They also take the friends and follower behavior into consideration.

Micro-blogging and predicting the personality is hot research area. *A study done in* [4] includes 444 users for their test. They utilize real time online individuals' behaviors to come up with algorithm using regression technique in data mining. They showed improved accuracy in relevance of related study and provide great reflection of online attitudes. They suggest continuing work in direction of other parameters like mental state and social networking behaviors.

A numerous studies [9][10][11] have shown the characteristics of data mining and personality traits such as FFM with Social networking data, to be very useful to categorize personalities accordingly.

Machine learning and its prediction power[12][13] have emerged and evolved into many sophisticated algorithms.

A study done in [14]machine learning technique based on Rough Sets to extract rules for their prediction work. They claim to have better efficiencies as compared to SVM (Support Vector Machine) methodologies[15][16].

Various comprehensive reviews[17][18] on Search of individual personality types for constructive real world application has been conducted. They support the future mining of data to address challenges and issues, at hand. Similar to data on social networking that contributes to Personality Prediction Opportunities and research, Educational data mining Big data Education[19]

[3] "Bipolar subtypes have differing personality traits", Springer Healthcare News, vol. 1, no. 1, 2012.

contributes to prediction and classification of factors in huge data set for success and failures of students and educational system in context.

A stream of research work[20] [21] has supported and advanced the techniques and data mining[22][23] algorithm to improve prediction accuracies. Various case studies[24][25] support to predict drop outs and improve student performance. Student behavior[26][27] prediction plays role in such type of data mining and conclude predictive metrics and measures. Student Recommender system[28] has matured through various researches in student data mining[29][30].

Predictive, Preventive, Descriptive and Prescriptive Analytics in Context

In this section, we discuss four types of analytics[31]. Predictive Analytics is the branch of big data Analytics that utilizes data and knowledge from past and correlate it with current data to predict the likeliness of future events to some accuracy. It combines the processes of data collection, data modeling, statistical techniques, data mining methods and deployment. It tells us, *what will happen or what can happen?*

Preventive Analytics is relatively newer branch of analytics and deals with facts extracted with Predictions and put in perspective to take measure to prevent undesired things from happenings. For example, after prediction of epidemics, the next goal is to do data analysis to find efficient measures to be taken accordingly. It essence, Preventive Analytics tells us *what should be avoided?*

Descriptive Analytics deals with description of past events. In other words, it tells us *what has happened*?

Prescriptive Analytics generally is known to include all three descriptive, predictive and prescriptive analyses. It further provides advices on particular situation and cases that data mining project is trying to accomplish. Analogy is pretty much when doctor prescribes medicine on certain disease, patient is having. It tells us *what we should do to make it happen the way want it to be.*

PERFECT Algorithm

Based on our review of literature of Educational Data Mining and Personality Prediction, we correlate the gaps and introduce our family of algorithms known as PERFECT algorithm. This consists of several algorithms in conjunction with mathematical modeling and development is in progress at the time of this writings.

Figure 1 shows our concept of all three types of Analytics through lens of our PERFECT algorithm

This figure shows our study about Predictive, Preventive and Prescriptive Analytics for PERFECT Algorithm.

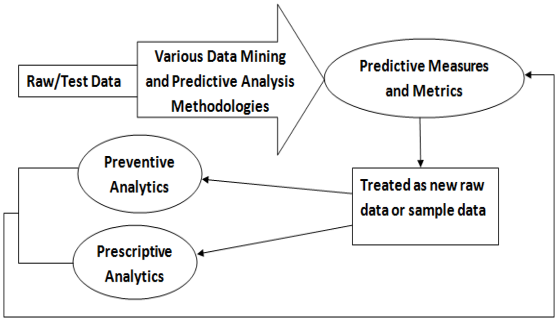
Raw/Test Data
Various Data Mining and Predictive Analysis Methodologies
Predictive Measures and Metrics
Preventive Analytics
Treated as new raw data or sample data
Prescriptive Analytics

Figure 1 – PERFECT Oriented Vision for 3P's

High Level of Our Algorithms

In this paper, we present the high level of the proposed algorithms that are currently working in progress.

Our Family of PERFECT Algorithm consists of several sub-algorithms/tasks. The diagram below shows the black box and vision of our final research work. We include one algorithms and high level of model and framework/work flow in figure 2.

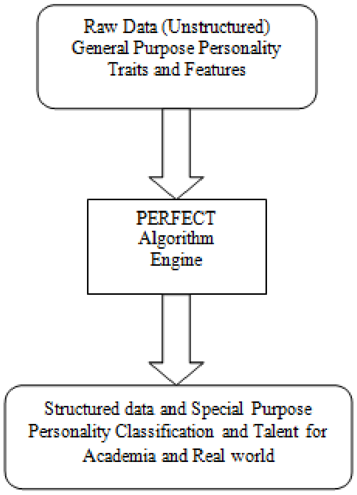
Raw Data (Unstructured)
General Purpose Personality
Traits and Features
PERFECT
Algorithm
Engine
Structured data and Special Purpose
Personality Classification and Talent for
Academia and Real world

Figure 2 – PERFECT Algorithm Engine Process Diagram

In the following paragraph, we introduce the high level proposed algorithm at preliminary level of our research.

Separating Signal from Noise:

Stream of data contains irrelevant or redundant data, we call it *Noise*. Relevant data is Signal that has tendency to provide useful information in context. Structures and relationship that we identify out of Noisy data contributes also the signal or useful information/data that we further utilize in our algorithm family/Engine.

In this algorithm, we learn to detect Noise and remove it before further processing of data to reduce cost and time overhead.

Definition: *To our context, Noise is defined as the following:*

Redundant data
Irrelevant Data, and
Unstructured or Non-relational that has no relation with other data elements or entities.

Algorithm # 1 – *Noise Removal and Feature Extraction – High Level*

1. Initialize Big Data Set, Valid Data set and Invalid Data set.
2. Define 3 Vs of Big Data as (Variability, Variety and Veracity)
3. Initialize Inconsistency of dataset and Noisy data set.
4. Define Threshold.
5. Input 3 V's and Threshold value
6. Output Dataset for Noise and Probability Function
7. Big Data set Processing to separate valid and invalid set
8. Using Threshold, find out Probability of valid data set for 3 V's and then do the division accordingly.
9. Extract Signal (valid) and Noise (invalid) accordingly.
10. Detect the inconsistencies in data set using 3 Vs.
11. Based on threshold, increment signal or noisy data set.
12. Repeat Process unless all checked.

Results, Discussions and Conclusion

To keep the algorithm output scalable and effective, we only use 3 Vs for this algorithm to separate valuable data from noisy data. We test our model 135 MB sample size and Noise data extraction accuracy comes out to be about 93 %. For size of 360 MB set, our accuracy was about 89%. This Algorithm introduced in this paper, serves as initial step towards PERFECT Algorithm Engine, where Algorithm # 2 and Algorithm # 3 will utilize the Noise separation in test data set to extract relevant features. Such relevant features, are used to do the Educational Data mining and career prediction to improve academic performances, reduce drop outs, maintain good retention rates and help avoid getting caught in wrong careers using predictive, prescriptive and preventive analysis. Relevant features are structured data points that have impact on academic success of an individual. Such features also can be used to build retention prediction models. Personality features of an individual have great impact on educational choices they should make. It is a challenge to predict high success for a particular education and career line. More often, decisions are made solely on available academic choices and other supportive trends, such popularity, affordability and family inclination. While FFM uses questionnaire, and have reported limitations reported in literature[32], we therefore, suggest using an individual data, in regard to their personalities and talent; they exhibit in social networking sites to predict the educational and career directions having high success rate for them. We include only 3 V's for the work presented in this paper. In future, we may utilize rest of V's that may have an impact on results.

Future Works

In continuation of above Algorithms, our future release of publication will add about 4 more algorithms to show prediction, prevention and prescription strategies for a student in college to choose right line of education that finally results in right career. This work will show improved accuracy and prediction rates for success in light of the data that is related to their Academics and career.

References

[1] J. A. N. Cieciuch, "the Big Five and Belbin," 2014.

[2] D. P. Schmitt, J. Allik, R. R. McCrae, and V. Benet-Martinez, "The Geographic Distribution of Big Five Personality Traits: Patterns and Profiles of Human Self-Description Across 56 Nations," *J. Cross. Cult. Psychol.*, vol. 38, no. 2, pp. 173–212, 2007.

[3] T. A. Byington, "Communities of Practice: Using Blogs to Increase Collaboration," *Interv. Sch. Clin.*, vol. 46, no. 5, pp. 280–291, 2011.

[4] S. Bai, B. Hao, A. Li, S. Yuan, R. Gao, and T. Zhu, "Predicting big five personality traits of microblog users," *Proc. - 2013 IEEE/WIC/ACM Int. Conf. Web Intell. WI 2013*, vol. 1, pp. 501–508, 2013.

[5] M. A. Khan, M. F. Uddin, and N. Gupta, "Seven V's of Big Data understanding Big Data to extract value," *Proc. 2014 Zo. 1 Conf. Am. Soc. Eng. Educ.*, pp. 1–5, Apr. 2014.

[6] G. Saucier and L. R. Goldberg, "The Language af Personality : Lexical Perspectives on the Five-Factor Model," *Five-Factor Model Personal. Theor. Perspect.*, pp. 21–50, 1996.

[7] L. Goldberg, "From Ace to Zombie: Some explorations in the language of personality," *Advances in personality assessment*, vol. 1. pp. 203–234, 1982.

[8] S. Adali and J. Golbeck, "Predicting Personality with Social Behavior," *2012 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Min.*, pp. 302–309, 2012.

[9] O. Celiktutan, E. Sariyanidi, and H. Gunes, "Let me tell you about your personality!†: Real-time personality prediction from nonverbal behavioural cues," *2015 11th IEEE Int. Conf. Work. Autom. Face Gesture Recognition, FG 2015*, p. 6026, 2015.

[10] O. Celiktutan and H. Gunes, "Automatic Prediction of Impressions in Time and across Varying Context: Personality, Attractiveness and Likeability," *IEEE Trans. Affect. Comput.*, vol. 3045, no. January 2016, pp. 1–1, 2016.

[11] G. Chittaranjan, B. Jan, and D. Gatica-Perez, "Who's who with big-five: Analyzing and classifying personality traits with smartphones," *Proc. - Int. Symp. Wearable Comput. ISWC*, pp. 29–36, 2011.

[12] R. Wald, T. Khoshgoftaar, and C. Sumner, "Machine prediction of personality from Facebook profiles," *Proc. 2012 IEEE 13th Int. Conf. Inf. Reuse Integr. IRI 2012*, pp. 109–115, 2012.

[13] C. M. Bishop, *Pattern Recognition and Machine Learning*, vol. 4, no. 4. 2006.

[14] U. Gupta and N. Chatterjee, "Personality Traits Identification Using Rough Sets Based Machine Learning," *2013 Int. Symp. Comput. Bus. Intell.*, pp. 182–185, 2013.

[15] and C.-J. L. Chih-Wei Hsu, Chih-Chung Chang, "A Practical Guide to Support Vector Classification," *BJU Int.*, vol. 101, no. 1, pp. 1396–400, 2008.

[16] C. C. J. C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Min. Knowl. Discov.*, vol. 2, no. 2, pp. 121–167, 1998.

[17] F. Pianesi, "Searching for personality,"

IEEE Signal Process. Mag., vol. 30, no. 1, pp. 146–158, 2013.

[18] A. Vinciarelli and G. Mohammadi, “A survey of personality computing,” *IEEE Trans. Affect. Comput.*, vol. 5, no. 3, pp. 273–291, 2014.

[19] L. Cen, D. Ruta, and J. Ng, “Big Education : Opportunities for Big Data Analytics,” pp. 502–506, 2015.

[20] R. Shaun, J. De Baker, and P. S. Inventado, “Chapter 4: Educational Data Mining and Learning Analytics,” *Springer*, vol. Chapter 4, pp. 61–75, 2014.

[21] N. T. N. Hien and P. Haddawy, “A decision support system for evaluating international student applications,” *Proc. - Front. Educ. Conf. FIE*, pp. 1–6, 2007.

[22] R. Jindal and M. D. Borah, “A Survey on Educational Data Mining and Research Trends,” *Int. J. Database Manag. Syst.*, vol. 5, no. 3, pp. 53–73, 2013.

[23] R. S. J. D. Baker, “Data mining for education,” *Int. Encycl. Educ.*, vol. 7, pp. 112–118, 2010.

[24] A. Merceron and K. Yacef, “Educational data mining: A case study,” *Artif. Intell. Educ. Support. Learn. through Intell. Soc. Inf. Technol.*, pp. 467–474, 2005.

[25] G. W. . Dekker, M. . Pechenizkiy, and J. M. . Vleeshouwers, “Predicting students drop out: A case study,” *EDM’09 - Educ. Data Min. 2009 2nd Int. Conf. Educ. Data Min.*, pp. 41–50, 2009.

[26] J. Sheard, J. Ceddia, J. Hurst, and J. Tuovinen, “Inferring student learning behaviour from website interactions: A usage analysis,” *Educ. Inf. Technol.*, vol. 8, no. 2002, pp. 245–266, 2003.

[27] A. El-Halees, “Mining Students Data To Analyze Learning Behavior : a Case Study Educational Systems,” *Work*, 2008.

[28] M. Goga, S. Kuyoro, and N. Goga, “A Recommender for Improving the Student Academic Performance,” *Procedia - Soc. Behav. Sci.*, vol. 180, no. November 2014, pp. 1481–1488, 2015.

[29] J. K. J. Kalpana and K. Venkatalakshmi, “Intellectual Performance Analysis of Students’ by using Data Mining Techniques,” vol. 3, no. 3, pp. 1922–1929, 2014.

[30] A. A. Al-shargabi and A. N. Nusari, “Discovering vital patterns from UST students data by applying data mining techniques,” *2010 2nd Int. Conf. Comput. Autom. Eng. ICCAE 2010*, vol. 2, no. 2, pp. 547–551, 2010.

[31] IBM Software, “Descriptive, predictive, prescriptive: Transforming asset and facilities management with analytics,” *Thought Leadersh. White Pap.*, no. October, 2013.

[32] G. J. Boyle, “Critique of the five-factor model of personality,” 2008.