Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 112 (2017) 1964–1973

Procedia Computer Science

www.elsevier.com/locate/procedia

21st International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES2017, 6-8 September 2017, Marseille, France

Personality Assessment using Twitter Tweets

Nadeem Ahmad [a*], Jawaid Siddique[a]

[a]*CS & IT Department, The University of Lahore, 1-Km Defence Road, Lahore Pakistan.*

Abstract

Social media establishes uninterrupted connectivity, between its users and external world through revealing personal details and their viewpoints in every aspect of life. The focal aim of this study is to analyze how twitter (dataset) can be used to improve the user experience with personality assessment. The article, in its primary context, exhibits the psychological profiling of users on the basis of the dataset. This profiling can be a very useful tool for career progression, job satisfaction and setting preferences in different interfaces. We propose a way in which the user's personality can be predicted through information mapping available to the public on their personal Twitter using DISC (Dominance, Influence, Compliance, Steadiness) assessment. The outcomes of this study can be useful for information retrieval (search engine), content selection mechanism and positioning product & services.

Peer-review under responsibility of KES International

Keywords: Social Networks, Pattern Recognition, Personality Assessment, DISC, Algorithms.

* Corresponding author. Tel.: +92 322 644 3536
E-mail address: nadeem.ahmad@cs.uol.edu.pk

1877-0509
Peer-review under responsibility of KES International
10.1016/j.procs.2017.08.067

1. Introduction

Personality is made up of the characteristics, patterns of thoughts, feelings, and behaviors, which makes a person unique. Behavioral modification and modulation are the core areas that provide rationale towards human interaction and socially balanced relationships. Personality assessment of social media users can be performed by keeping in view their social background, demographics and ethnic origins. Personality Assessments (PA) can be used for Information Retrieval (Search Engine), Content Selection Mechanism, Positioning Product & Services and many other areas.

There are many models available for personality assessment, like Myers Briggs Type Indicator (MBTI) [16], DISC Assessment [20], Strength Finder [18] and Big Five Personality Traits [15] etc.

Myers Briggs Type Indicator (MBTI): It has been a widely known and accepted personality assessment test around the globe since last 55 years. This test is based upon Carl Jung Typology theory and describes human personality in four fundamental dimensions as extroversion/introversion, sensing/intuition, thinking/feeling, and judging/perceiving [16]. The test categorizes the human personality type in one of the popular sixteen categories that further exemplify human personality with reference to its basic nature and preferences. Since the test is descriptive and analytic in nature that is why one needs to be very good at analytical interpretation and psycho-analytic profiling.

DISC: It is a simple, applied and more intuitive personality assessment test, introduced way back in 1928. DISC performs behavioral assessment keeping four principal and key behaviors as benchmark which includes: Dominance, Influence, Stability and Compatibility. DISC explicitely concentrates on behavioral preferences that's why it's more applied, explanatory and comprehensible as compared to MBTI [19].

Strength Finder: Gallup in 2001 introduced this new personality assessment technique [17]. Strength finder technique revolves around 34 basic talents (strength) themes that encompass a human's personality and invites people to reveal their strength and use it in order to be more successful and productive in life. The test is more related to Positive psychology undertaken for personality assessment. The strength finder test is more restrictive on a proactive strategy, in contrast with MBTI and DISC [19]. It lacks the intuitive model that team members can be moved [17,19]. It seems difficult to remember 34 strengths where relationship among them is not clearly described. It also does not explicitly specify top weaknesses, either individuals or teams.

Multidisciplinary collaborations in fields such as computational linguistics has been undertaken to deduce sense out of large datasets that describes the similarities and differences between groups on their differential language use. In contrast, in this article a data-driven collection of words, phrases, and topics is extracted, in which the lexicon is based on the tweets of the text being analyzed. It appears beneficial while describing logical differences between the two groups in any given cell, and allows one to experience unexpected results. We call approaches like ours, which do not rely on a priori word or category judgments, open-vocabulary analysis.

Mostly the data accessible on the web and intranets is by and large displayed in the form of text documents. Usually these data sources are organized and sorted in a specific way so that extraction of information from these sources could be easy and instant. The analysts and professionals may search for classifications of text sources by some automatic technique in areas like information retrieval and content mining [7], [8].

Data clustering is a methodology for clustering data to extract information from text documents. Data clustering and text mining can be further subdivided into document clustering and text clustering. The base and source of text clustering is upon Information Retrieval (IR), Natural Language Processing (NLP) and Machine Learning (ML) [8].

Many techniques of document clustering are, in practice, for example a novel technique for clustering of documents, which have been written in a language evolving during different historical periods [28]. Based on two

phases 1) the text is converted into a string of four numerical codes, and then 2) in the next phase, texture features are pull out from the obtained image in order to create document feature vectors. Afterward, a clustering algorithm is engaged on the feature vectors to classify documents from different historical periods of the language.

Another innovative technique is consisting on map each letter of the text with certain script type [27]. It is made according to characteristics concerning the position of the letter in the baseline area. In order to extract features, the co-occurrence matrix is computed. Then, the texture features are calculated. Extracted measures show meaningful differences due to dissimilarities in the script and language characteristics.

In our study, the user personality is predicted by analyzing the twitter tweets by using DISC framework. The text mining technique (clustering) and sentiment analysis are performed on more than 1 million tweets. Some pre-processing steps were performed to clean the noisy data and finally user personality is mapped on DISC to find out the personality characteristics.

In section 2, related work regarding text mining and classification of user based on different attributes are described. Section 3, describe our methodology while Section 4 covers Social media tools and analytics. Section 5 describes experiments and exhibits results in detail. Section 6 summarizes the article and provide insight for future extension of work.

2. Related Work

Social networking websites can help to eradicate the coordination issues among individuals that are at a considerable physical distance [21]. It can build the feasibility of social campaigns [22] by distributing the necessary data at any place and time. In any case, in social networking websites, individuals for the most part use unstructured or semi-structured language for correspondence.

Individuals are less interested in correct spellings and the precise etymological sense revleaed in sentences during everyday communication. This may incite different sorts of ambiguities, for instance, lexical, syntactic, and semantic understanding of sentences may be compromised [23]. Therefore, mining logical patterns with exact data from such unstructured types of data is a complex job.

A review by Berry and Castellanos in 2004, expressed that text mining procedures are utilized to extract information from different documents [24]. The text mining techniques are helpful in forming new knowledge by expoiting exisiting relationships in extracted information or may start a new era of more investigation to get appropriate patterns for data classification. The text mining techniques are different from usual search on web. In web searching the user gets information which others posted on web without caring the required structure while in text mining only required data is provided by putting aside all irrelevant data.

It is widely believed that powerful information is locked within the vast amounts of data. The use of large sets of user data for determining the personality on the social media is common. Recent studies have been able to successfully build models to predict a wide range of these attributes of the user such as age [1], gender [2], occupation [3], personality [4, 9], assessing job candidates [10], political orientation [5], deciding whether to join (Dating) someone [11], assessment by emails [14] and location [6].

A novel framework for community discovery is proposed by Amelio & Pizzuti [25] in multidimensional networks, which is based on multi-objective optimization and local search in context of social media. In another study, a new methodology is proposed by Amelio & Pizzuti which is known as SNMOGA (Signed Networks with Multi Objective Genetic Algorithms) that optimizes the concept of modularity and frustration by applying genetic algorithms [26]. Detecting community structure in Signed networks based on positive (friendly) or negative (antagonistic) relations is an interesting research phenomena which enables us to determine the instability and predicts the possible changes in group organization. Existing studies have used varied category of information,

collected from social networking sites by using hypothesis homophily (i.e., "love of the same", is the tendency of individuals to associate and bond with similar others, as in the proverb "birds of a feather flock together") [7]. Furthermore, people can make personality assessment from Facebook (social network sites, SNSs) profiles [8].

3. Methodology and Analysis Requirements

This study aims for a broader, larger and more relevant personality assessment than the media and polls have. The twitter sentiment/texts are expanded by classifying tweets into four categories Dominance, Influence, Submission, and Compliance (DISC). DISC has proved predictive validity and compatibility with earlier knowledge (Social Sciences & Marketing) with understandable dimensions.

Text mining and sentiment analysis were performed for each user based on his/her recent tweets. We have downloaded over one million tweets using keywords. Our keywords for the search are summarized in Figure 1.

Dominance

Influence

Steadiness

Compliance

Figure 1: Keywords for search, Word Cloud

The main reasons for the use of social networking are, first, the easy availability of and access to a set of generalized data, and secondly, access to tools that allow for the "profound" analysis of data sets [4]. It is significant

to mention that researchers have access to the data sets like Twitter for experimental purpose. The focused methodology relevant to this study can be classified in three staple ways i.e. data, data analysis and data visualization as summarized in Figure 2.

3.1 Public Data and Data Analysis

As discussed earlier a dataset of one million tweets, freely available on Internet for research purposes, was extracted by using Rapidminer on which text mining is applied through R language. R is an open source language and environment for statistical computation and graphics. Its various packages are used to carry out text processing [12].

RapidMiner is a data science software platform which provides an integrated environment for machine learning, deep learning, text mining, and predictive analytics. It is used for business and commercial applications and also for research, education, training, rapid prototyping, and application development. The tool supports all steps of the machine learning process, including, data preparation, validation, results visualization and optimization [13].

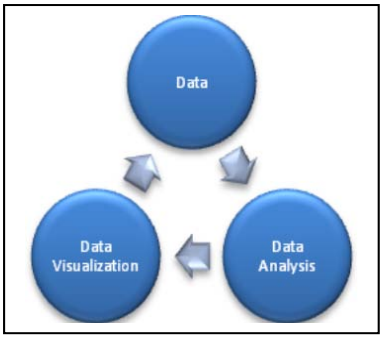


Figure 2: Methodology used in data analysis

3.2 Data Visualization

Visualization tools for researcher are evidently required, so that they can visualize data in some graphic form and fulfill the goal of communicating information clearly and effectively. Figure 2 depicts the overall process of the data analytics process. The process begins with analytics dashboard, followed by data analysis and ends with data visualization.

4. Social Media Analytics Tools and Techniques

One of the main features of the massive amount of data generated by the user is an online texts disorder with high diversity. In order to analyze such data-sets; natural language processing (NLP), text analytics and computational linguistics methods are used to classify and find out subjective information from source text [5]. The overall purpose is to find out the viewpoint of a writer with respect to some subject matter or the overall contextual divergence of a document.

The techniques exist includes: sentiment analysis, machine learning (ML), computational statistics, supervised learning methods etc., [6, 7]. The techniques of computational science for the analysis of social media data, contains bag-of-words model semantic orientation, machine learning, etc. [5]. Figure 3 illustrates the existing computational science techniques that can be used for analyzing (text mining) social media data.

Figure 3: Pre-processing Steps

5. Experiments and Results

Figure 1, presents words and phrases that most distinguished in each group. Even though dominant words in each word cloud generally reflect what might be expected based on decades of questionnaire-based personality research while the surrounding words suggest processes underlying each group.

Words were divided into four groups, as mentioned earlier:

1. $D = \{D_1, D_2, D_3, \ldots D_{27}\}$
2. $I = \{I_1, I_2, I_3, \ldots I_{27}\}$
3. $S = \{S_1, S_2, S_3, \ldots S_{27}\}$
4. $C = \{C_1, C_2, C_3, \ldots C_{27}\}$

and $X = \{D, I, S, C\}$ (whole corpus).

First step was pre-processing (Figure 3), in which the extracted text is cleaned by removing punctuations, numbers, common words, etc. in order to increase robustness and efficiency of the results.

On execution of two R scripts to 1) clean & prepared data set and 2) calculate the word frequency of each group of data set; we got the Word Cloud Weight (numbers) for each group.

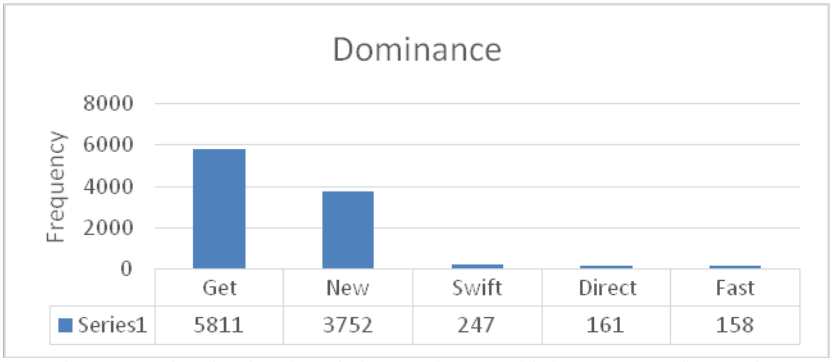


Figure 4: Depicts the plot of words for 'Dominance' with frequency (top five words).

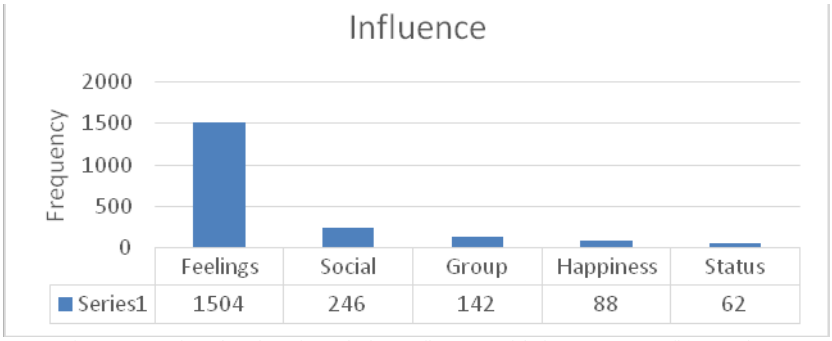


Figure 5: Depicts the plot of words for 'Influence' with frequency (top five words).

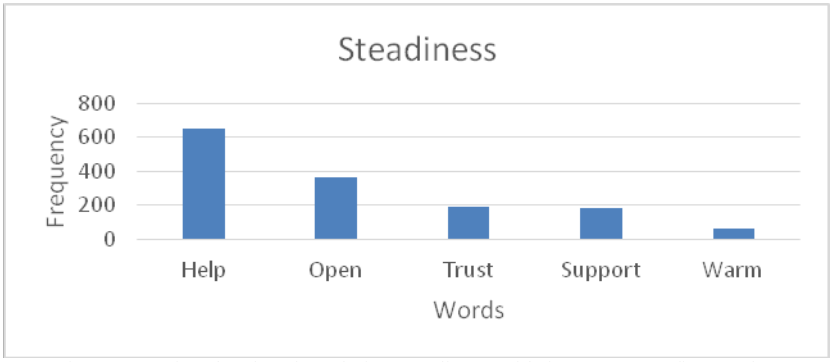


Figure 6: Depicts the plot of words for 'Steadiness' with frequency (top five words).

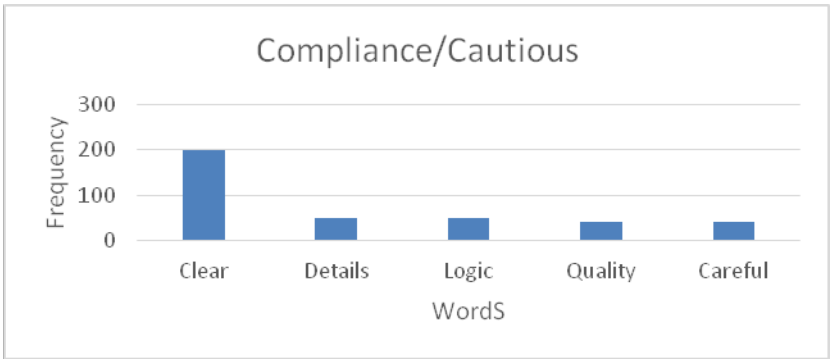
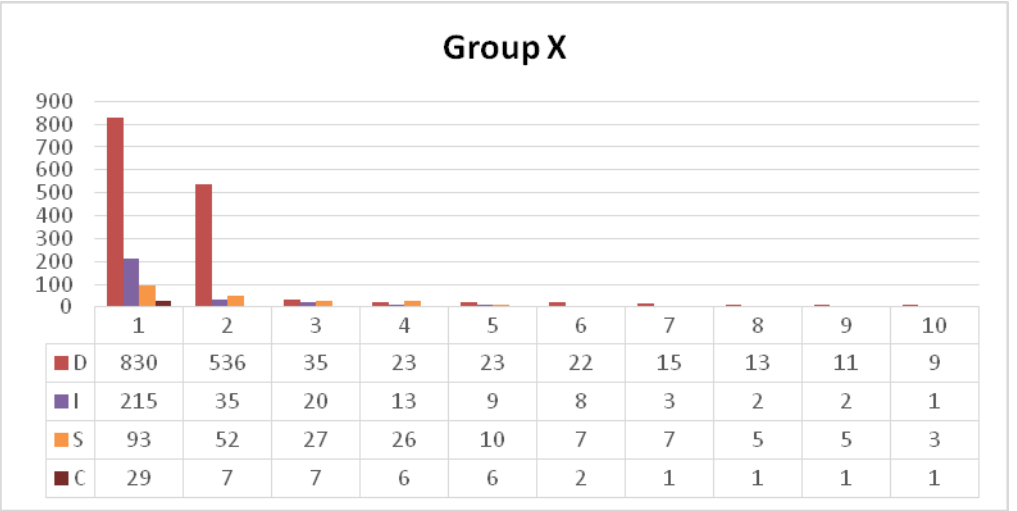


Figure 7: Depicts the plot of words for 'Compliance' with frequency (top five words).

Figure 4, 5, 6, & 7 shows top five results of individual groups. Whereas figure 8 presents holistic corpus (group X) with top 10 words from each group.



	1	2	3	4	5	6	7	8	9	10
D	830	536	35	23	23	22	15	13	11	9
I	215	35	20	13	9	8	3	2	2	1
S	93	52	27	26	10	7	7	5	5	3
C	29	7	7	6	6	2	1	1	1	1

Figure 8: What is the heading of this graph? For understanding and clarity, above graph has been rescaled.

Considering words that occur at least 20 times (word frequency, can be increased or decreased according to the requirements); results shows there are 6 lexemes with word count greater than 20 in group "D", 3 lexemes in group "I", 4 lexemes in group "S" and 1 lexeme in group "C", assuming all the words have same priority; the calculations show that the user primarily belongs to "Dominance" category with some additional qualities from "Influence" and "Steadiness", by combing these two means "People Oriented" as exhibited in figure 9 showing DISC framework. So, the person psyche is Dominant with People Oriented approach.

6. Conclusion & Future Work

Effects of the growth of social networks sites (SNS) are very heavy on the techniques developed for text mining in social networks. Text mining provides a proficient way to execute and make use of datasets. In this paper, we have reviewed the text mining technique (clustering) which is used for social network analysis and mapped its results to a framework called DISC. The practical implementation of the results can be in the field of information retrieval, content selection, product positioning and psychological assessment of the user.

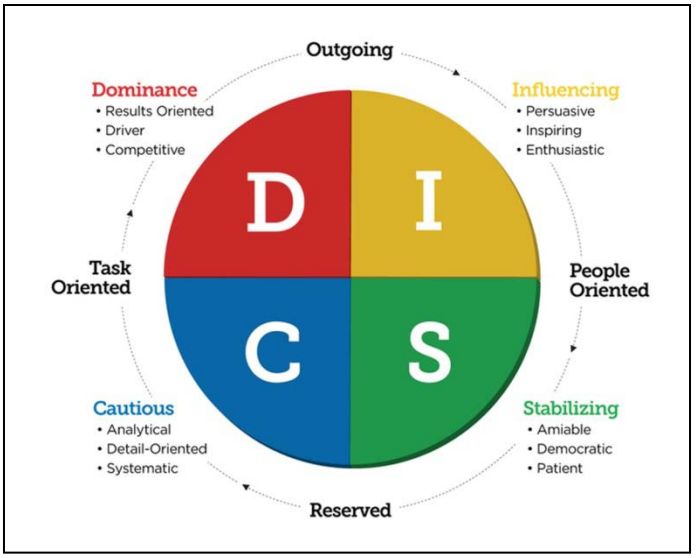


Figure 9: DISC (Dominance, Influence, Compliance, Steadiness) Framework

Psychoanalytic profiling has been a well received and proven method with scores of techniques that are undertaken to study multi-dimensionality in an individual's personality. Social media has changed the linguistic syntax across the globe and now people seem more comfortable to express themselves with tag words instead of using a statement or complete sentence. These tag words are helpful in generating themes which assists researchers in drawing out respective conclusion with maximum precision. This study is a very initial attempt towards 'people profiling' with the help of tag words that can be a very useful tool in number of other areas as marketing, promotions, advertising, sales, IT applications and anthropological studies etc.

The peculiarity of this research study is that it provides a framework for using text mining technique in altogether a novel area. Keeping the same track, parameters of geolocation (latitude & longitude) can be added to regionalize the assessment. Parameters of trend and taste can further be added to gain a comprehensive personality profile that can provide a fundamental ground for extending findings of this research to other disciplines. We invite other researchers to please come forward and take initiatives to enrich the body of knowledge in the very right perspective of this new area and take the matter further through exploring new-fangled dimensions.

References

1. Rao, D.; Yarowsky, D.; Shreevats, A.; and Gupta, M. 2010. Classifying Latent User Attributes in Twitter. In Proceedings of the 2nd International Workshop on Search and Mining User-generated Contents, SMUC, 37–44
2. Burger, D. J.; Henderson, J.; Kim, G.; and Zarrella, G. 2011. Discriminating Gender on Twitter. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP, 1301–1309
3. Preot¸ iuc-Pietro, D.; Lampos, V.; and Aletras, N. 2015. An Analysis of the User Occupational Class through Twitter Content. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, ACL, 1754–1764.
4. Schwartz, H. A.; Eichstaedt, J. C.; Kern, M. L.; Dziurzynski, L.; Ramones, S. M.; Agrawal, M.; Shah, A.; Kosinski, M.; Stillwell, D.; Seligman, M. E.; et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. PloS ONE 8(9).
5. Pennacchiotti, M., and Popescu, A.-M. 2011. A machine learning approach to twitter user classification. In Proceedings of the 5th International AAAI Conference on Weblogs and Social Media, ICWSM, 281–288.
6. Cheng, Z.; Caverlee, J.; and Lee, K. 2010. You are where you Tweet: A Content-Based Aproach to Geo-Locating Twitter Users. In Proceedings of the 19th ACM Conference on Information and Knowledge Management, CIKM, 759–768.
7. Rout, D.; Preot¸iuc-Pietro, D.; Kalina, B.; and Cohn, T. 2013. Where's @wally: A Classification Approach to Geolocating Users based on their Social Ties. HT, 11–20.
8. Evans et al. 2008, Gosling et al. 2007 & 2011
9. Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., ... & Seligman, M. E. (2015). Automatic personality assessment through social media language. Journal of personality and social psychology, 108(6), 934.
10. Kluemper, D. H., & Rosen, P. A. (2009). Future employment selection methods: evaluating social networking web sites. Journal of managerial Psychology, 24(6), 567-580.
11. Zhao, S., Grasmuck, S., & Martin, J. (2008). Identity construction on Facebook: Digital empowerment in anchored relationships. Computers in human behavior, 24(5), 1816-1836.
12. Ripley, B. D. (2001). The R project in statistical computing. MSOR Connections. The newsletter of the LTSN Maths, Stats & OR Network, 1(1), 23-25.
13. Klinkenberg, R. (Ed.). (2013). RapidMiner: Data mining use cases and business analytics applications. Chapman and Hall/CRC.
14. Shen J, Brdiczka O, Liu J. Understanding email writers: personality prediction from email messages. In: Carberry S, Weibelzahl S, Micarelli A, Semeraro G, editors. User modeling, adaptation, and personalization. Lecture notes in computer science, vol. 7899. Berlin, Heidelberg: Springer; 2013. p. 318–30

15. Judge, T. A., Higgins, C. A., Thoresen, C. J. and Barrick, M. R. (1999), THE BIG FIVE PERSONALITY TRAITS, GENERAL MENTAL ABILITY, AND CAREER SUCCESS ACROSS THE LIFE SPAN. Personnel Psychology, 52: 621–652.
16. Myers, I. B., McCaulley, M. H., & Most, R. (1985). Manual: A guide to the development and use of the Myers-Briggs Type Indicator (Vol. 1985). Palo Alto, CA: Consulting Psychologists Press.
17. Buckingham, M., & Clifton, D. O. (2001). Now, discover your strengths. Simon and Schuster.
18. Rath, T. (2007). StrengthsFinder 2.0. Simon and Schuster.
19. Reynierse, J. H., Ackerman, D., Fink, A. A., & Harker, J. B. (2000). The effects of personality and management role on perceived values in business settings. International Journal of Value-Based Management, 13(1), 1-13.
20. Shaffer, D., Schwab-Stone, M., Fisher, P., Cohen, P., Placentini, J., Davies, M., ... & Regier, D. (1993). The diagnostic interview schedule for children-revised version (DISC-R): I. Preparation, field testing, interrater reliability, and acceptability. Journal of the American Academy of Child & Adolescent Psychiatry, 32(3), 643-650.
21. Evans, B. M., Kairam, S., & Pirolli, P. (2010). Do your friends make you smarter?: An analysis of social strategies in online information seeking. Information Processing & Management, 46(6), 679-692.
22. Li, J., & Khan, S. U. (2009, November). MobiSN: Semantics-based mobile ad hoc social network framework. In Global Telecommunications Conference, 2009. GLOBECOM 2009. IEEE (pp. 1-6). Ieee.
23. Sorensen, L. (2009, May). User managed trust in social networking-Comparing Facebook, MySpace and Linkedin. In Wireless Communication, Vehicular Technology, Information Theory and Aerospace & Electronic Systems Technology, 2009. Wireless VITAE 2009. 1st International Conference on (pp. 427-431). IEEE.
24. Berry Michael, W. (2004). Automatic Discovery of Similar Words. Survey of Text Mining: Clustering, Classification and Retrieval", Springer Verlag, New York, 200, 24-43.
25. Amelio, A., & Pizzuti, C. (2016). Evolutionary clustering for mining and tracking dynamic multilayer networks. Computational Intelligence.
26. Amelio, A., & Pizzuti, C. (2016). An Evolutionary and Local Refinement Approach for Community Detection in Signed Networks. International Journal on Artificial Intelligence Tools, 25(04), 1650021.
27. Brodić, D., Amelio, A., & Milivojević, Z. N. (2016). Language discrimination by texture analysis of the image corresponding to the text. Neural Computing and Applications, 1-22.
28. Brodić, D., Amelio, A., & Milivojević, Z. N. Clustering documents in evolving languages by image texture analysis. Applied Intelligence, 1-18.