Volume 5, Issue 10, October-2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Comprehensive Survey on Personality of Authors on Blog Data

Shailendra Saxena
Department of Computer Science and Engineering, IIMT College of Engineering,
Greater Noida, Uttar Pradesh, India

Abstract: ***A weblog or blog is defined as a "frequently update websites consisting of dated entries arranged in reverse chronological order so that the most recent past appears first". Blogs are usually maintained by an individual with regular entries of commentary, descriptions of events or other material such as graphics or video.***
In this paper I have surveyed and analyzed the blog author's personality. In today's scenario Internet is the best way for sharing the individual thoughts. With respect to Internet, Weblogs (Blogs) give the opportunity to individuals for writing and expressing their feeling in front of others. On the basis of written blog it is a great challenge to identify the trait of the authors.
Here, I performed a survey and found the precious constraints viz. High Extrovert, Low Extrovert, High Introvert, Low Introvert which are very useful while we decide the personality of authors. To overcome previous problems, we analyzed the text written in blog. This analysis leads us to identify the personality of the author. As predicted High extrovert generally more emphasize on present tense with adjective which has positive felling. Author belongs to either high introvert or low extrovert use generally negative emotion thought, but more features is available in high introvert e.g. short sentences, maximum first person pronoun. There is one more category known as low introvert, here authors generally discuss about the current scenario as well as the sentences are comparatively large with positive feelings.

Keywords: POS, LIWC, OC, NC

I. INTRODUCTION

Blogs are one of the important forms of user's contribution to the web contents. Blogs are usually maintained by an individual with regular entries of commentary, descriptions of events, or other material such as graphics or video. Entries are commonly displayed in reverse-chronological order. *Blog* can also be used as a verb, meaning *to maintain or add content to a blog*. A weblog is defined as a " frequently update websites consisting of dated entries arranged in reverse chronological order so that the most recent past appears first" (Herman, Jahn and Ryan. 2005). The most common form of weblog is the personal diary or journal blog (Herring, Scheidt, Sabrina Bonus and Wright. 2005).
This modern blog i.e. online diary is used to would keep a running account of personal lives. Most such authors called themselves bloggers or diarists or journalists (Justin Hall. 1994)

II. ADDING AGE TO BLOGS

Blog writing came into existence in year 1980. Various personals were involved in blog writing it was the time when blogs were spreading their arms. Several improvements and analysis were also stated to work on blogs like-contents analysis, personality classification and blogger's trait identification etc.
In the year 2004, the blogs secured more attention as political consultants, news services and can be used as tools for outreach and opinion forming. Blogging was established by politicians and political candidates to express opinions on war and other issues and cemented blogs' role as a news source. (Howard Dean and Wesley Clark.)
In the year 2005, some analysis work performed on blogger's contents. Some motivation for blogging has been related back to psychological profiles of the author: exhibitionist, extrovert, creativity, auto-therapeutic, self-presentation and self promotion were found based on an analysis of polish blogs (Cywinska-Milonas, 2003 cited in Trammeli, Tarkowski, Hofmokl and Sapp. 2006). For similar promotion and presentation reasons businesses have started their own 'corporate weblogs'-in order to better connect with customers (Kelleher and Miller 2006; Puschmann, Forthcoming).
The impact of blogging upon the mainstream media has also been acknowledged by governments. In 2009, the presence of the American journalism industry had declined to the point that several newspaper corporations were filing for bankruptcy, resulting in less direct competition between newspapers within the same circulation area. Discussion emerged as to whether the newspaper industry would benefit from a stimulus package by the federal government.

III. EVOLUTION

Author identification came into existence in the 1980s. Starting from email authorship identification, De Vel analyzed stylistic attributes to discover forensics in email. [3]. He showed a good comparison and provided a relatively better result. However it cannot be applicable to blog data because email text is shorter than diaries text. Email text is usually a topical dialogue between two authors while blog text is from the author to the public, at least the intended group.

Heylighen and Dwaele [4] explored the notion of implicitness in text, by developing a unitary measure of context. They considered parts-of-speech (POS) as they related to Dixie: that is to say POSs that require anchoring with spatial-temporal context of utterance in order to be properly interpreted.

Arjun Mukherjee and Bing Liu [5] worked on gender classification of Blog authors and gave some methods to improve it. They studied various gender preferential features consisting of a set of signals that has been used in an email gender classification task.

They proposed two new techniques to improve the current result. The first technique introduces a new class of features which are variable length POS sequence patterns mined from the training data using a sequence pattern mining algorithm. The second technique is a new feature selection method which is based on an ensemble of several feature selection criteria and approaches. Empirical evaluation using a real life blog data set shows that these two techniques improve the classification accuracy of the current state-of-the-art methods significantly. These results can be embedded in classification of author's traits.

Scott Nowson & Jon Oberlander [6] reported some new classification blog author personality. They took four important personality traits viz. Extraversion, Neuroticism, Agreeableness and conscientiousness. The main aim of their study was to explore the application of language models and classifiers developed using the cleaner and more reliably defined original corpus (OC) data to the more coarse-grained and noisier new corpus (NC) data. They developed a language model named as Model I, Model II, Model III and Model IV working on some training sets and two classifiers named 2-class classification and 3-class classification of the NC data. It was not a big success model but indeed this most finely tuned models and classifiers seem to suffer least in the scaling up procedure.

IV. RECENT SCENARIO

Personal weblogs (blogs) provide individuals with the opportunity to write freely and express themselves online in the presence of others. The blog contents can be interrogated and then used to examine the personality of authors [8]. Using a large blog corpus wedded with LIWC text analysis program [7], they examined the content of blogs to provide insight into the role of personality in motivation for blogging.

According to this paper, personality is decomposed into five numbers of measurable factors or traits i.e. Neuroticism, Extraversion, Openness, Agreeableness and Conscientiousness. This was one of the recent developments in the trait identification. A large blog corpus is used to examine the role of personality in motivation for blogging. Linguistic content of the blogs revealed cathartic and auto-therapeutic tendencies of high Neurotic bloggers; high level life documentation and emotion expression by high Extroverts; commentary and evolution by high Openness scorers and reports of daily life by high Conscientiousness bloggers. Highly Agreeable authors demonstrate similar strategies to those adopt in other contexts.

Haytham Mohtasseb and Amr Ahmed [2] experimented and presented a significant improvement in personal blogs. The improvement in authorship identification was done by utilizing hybrid collection of linguistic features that best capture the style of users in diaries blogs. The work also confirms the best ranges of several parameters that affect the final result of authorship identification. They explained the stages of authorship identification framework that starts with text collection. Next, in features extraction stage, every blog post is converted to a features vector, strong the values in a Relational databases which simplify and increase the speed of all the ongoing experiments.

V. CHALLENGES

In today's environment, where everyone is independent to write and express his views, the role of blogging is very important as a form of expression. Although blogs provide freedom to their authors to express their topic and style, in most cases the blogger's behavior is consistent with the predictions based on their personality type or their linguistic behavior in other media. It is very difficult to identify the personality of the author after analyzing the written blog because writing of text totally depends on the surrounding of author. The thought of author varies according to the surrounds, so for improving the accuracy it is better if we have many text written by a author under various surroundings and we consider all the text while we classifying the text under our category. For classification it is also required that error while classification should be minimum. It is also a important point that what should be the standard so that we can say it is the moderated length of a sentence.

Considering these points, few more areas in which blog corpus data can be used to identify the authors.

(i) Author classifications like Ambivert, High Extrovert or High Introvert etc. may also be distinguished from blog data.
(ii) Two or more classification can be combined due to different and varying author traits.

VI. CONCLUSION

Blog, E-mail and easy writing are the way of sharing the thoughts. Also thoughts reflect the personality of a person. The blog contents can be useful to identify the author's trait. On study of several works done previously, five major personalities can be observed from the posted contents. The matter of discussion is that there is very small difference between features of authors of different categories. Therefore distinguished feature must be scanned and identified. On the first occasion, it might give some approximated result but errors may again be studied. One solution may have some additional features of author's traits so that expected result can be closer.

REFERENCES

[1] S. Nowson and J. Oberlander. The identity of bloggers: Openness and gender in personal weblogs. AAAI Spring Symposium, Computational Approaches to Analyzing Weblogs, Stanford University., 2006.

[2] Haytham Mohtasseb and Amr Ahmed (2009): *More Blogging Features for Authorship Identification*. In Proceedings of International Conference on Knowledge Discovery (ICKD'09), Philippines.

[3] O. de Vel, A. Anderson, M. Corney and G. Mohay. Mining email content for author identification forenscis. ACM SIGMOD record, 30(4):55-64, 2001.

[4] J.M. Dewale and A Furnham Extraversion: The unloved variable in applied linguistic research. Language Learning 49:509-514. 1999.

[5] Arjun Mukharjee, Bing Liu Improving Gender Classification of Blog Authors Proceedings of the 2010 conference on Empirical Methods in natural Language Processing, pages 207-217 MIT. Massachusetts, USA 9-11 October 2010.

[6] Scott Nowson, Jon Oberlander Identifying more bloggers ICWSMH 2007 Boulder Colorado USA.

[7] Yla R. Tausczik and James W. Pennebaker "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods" published in "Language Style Matching Predicts Relationship Initiation and Stability" Psychological Science January 1, 2011 22: 39-44.

[8] Alastair J. Gill, S. Nowson and J. Oberlander "What are they Blogging About? Personality, Topic and Motivation in Blogs" Association for the Advancement Intelligence 2009.

