Identification of Human Behavioral Traits Using Social Media

Shruti Kohli, Ankit Gupta
Department of Computer Science and Engineering
Birla Institute of Technology, Mesra, Ranchi, India

Abstract

Today's web has become a primary source of finding information related to human psychology and behavior because of the abundance of user data available on various data repositories which are mostly freely available to the researchers. Social media has become such a repository of an enormous amount of data because of its popularity and number of users using internet in various forms. Web analytics has emerged as one of the web mining paradigm to extract useful information from this large pool of data. In this paper, data available on YouTube is being analyzed to find some human personality traits. The data is collected with the help of 30 undergraduate students selected randomly, to maintain the integrity of the results, and then the data was processed with the assistance of Weka- a data mining tools. MS-Excel spreadsheet was also used to compile some to the results.

1. Introduction

Today's web is a result of a continuous and concurrent research and development of the consortium of technologies proposed by individuals and groups. During 1990's, the concept of *hypertext* and *World Wide Web* changed the face of the then Internet. These concepts lead to the generation of the era of Web 1.0. *Web 1.0* consists of static web pages, and the responsibility of creating information and contents at these pages was on companies/ organizations that were responsible for creation and maintenance of the pages. End-users of the *Web 1.0* were considered as consumers of the information with no role in creation and update of information. The second era of web, *Web 2.0* is considered as both a technology and usage paradigm. Web 2.0 empowers the end-user to control and create the content at the web and participate with the various web related activities so as to relate with other users [34]. "Social Media" is an application of Web 2.0. Recent advancement of Internet in terms of usability, speed and reach to mass has made social media a new way to share one's individual thought anonymously. It can be comfortably said that "Social Media" is responsible for a revolutionary new trend that is of interest to companies, institutions and various organizations for finding emerging and unique patterns in human behavior and Psychology.

One of the earlier example of primitive social media was "Usenet" created by Tom Truscott and Jim Ellis of Duke University[2]. In Last decade, "Open Diary", a social networking site founded by Bruce and Susan Abelson was introduced to the public in large[3]. "Weblog" term was also used for the first time around the same time.

Kaplan et al. [4] describes social media as *"..a group of Internet -based applications that build on the ideological and technical foundations of Web 2.0 , and that allow the creation and exchange of User Generated Content*". The term "Web 2.0" was first used in 2004. It can be considered as a new paradigm where software designer and users started using WWW. Web 2.0 gave the public at large flexibility to modify content and applications in a participatory and collaborative fashion [5]. It ended an era where such modification was done by respective website/database administrator. As the social media is responsible for social presence, and social presence is closely related to media richness that is information hidden and needs to be extracted and transformed into some knowledge base [6]. Social media serve numerous purposes and exist in a variety of forms. It can be broadly classified into following six categories [4]:

1) Collaborative Projects: It can be described as the creation of contents by many of its users jointly and concurrently. e.g. Wikipedia.

2) Blogs: These are the earliest form of the social media. These are date-stamped entries based websites.

3) Content Communities: Sharing of media content is the primary objective of Content Communities. e.g. YouTube, Flickr, etc.

4) Social Networking site: These sites enables various users to connect with each other by inviting friends and colleagues followed by sharing of useful information among themselves, e.g. facebook.

5) Virtual Game world: They replicate a 3-D world environment so the user the games can visualize himself/herself in customized avatars and then interact and communicate with each other. e.g. "World of Warcraft".

6) Virtual Social worlds: It allows its users to live a virtual life similar to real life. e.g. "Second life".

In this paper, data obtained from YouTube, a variant of social media, and is an example of Content Community is being analyzed for finding some aspects of human behavior. This paper further tries to find relationships among various attributes that constitutes the compiled dataset.

Rest of the paper is as follows: Section 2 provides some insight to web mining and web analytics, Section 3 discusses the importance of social media analysis along with some scholarly work, Section 4 gives insight of YouTube statistics, Section 5 explains the various steps involved in this experiment, Section 6 analyzes aspects of the result and Section 7 concludes the paper.

2. Web Mining

Data mining is a term frequently used to mine the appropriate information from a group of structured or unstructured data. This process includes convenient extraction of patterns from a group of data that exhibits some kind of knowledge [30]. Earlier, Data mining was used as one of the many steps of the knowledge discovery process but later many researchers have started using data mining as synonyms of the knowledge process [31]. In the case of web-based data, web mining is referred to as the process of mining/extracting useful information. Web mining is further divided into 3 basic categories, namely [32]:

Web Content Mining: It refers to the study of the content of any website like images, videos, etc.

Web Structure Mining::It refers to the process of deducing information about the overall structure of any website, e.g. hyperlinks of website.

Web Usage Mining: It is the process of inferring knowledge and finding various usage patterns based on the users activities obtained using various means, like web server log.

A server log file contains all the information related to a user who by any means, communicated with the website. Web Analytics deals with the methods for measurement, data collection, data analysis and providing the related feedback on the internet for the motive of understanding behavior of the customer using website [33]. The benefit of studying behavior of the customer leads to optimizing the usage of the site. Web Analytics is the art and science of improving various features and characteristics of websites to increase their efficiency by improving the customer's overall browsing experience. This is a science because it uses statistics, data mining techniques, report summaries, operations and a methodological process to process. It can be said as art because to improve websites there requires a deep level of creativity, imagination, analysis, balancing user-centric design, promotions, content, images and more.

3. Motivation and Some Earlier Work

As per June 2014, approximately 1.31 billion people were using YouTube. Average time spend on browsing this social networking site is 18 minutes. Another social networking, Twitter (more precisely microblogging site) has about 0.645 billion users while flicker has about 0.10 billion active accounts [13], [14], [15], [16]. Many researchers are trying to exploit this vast and flexible collection of information for social and economic research. Raacke et al. [28] indicates the requisite to explore the personal and social needs that any individuals fulfill by using Facebook and MySpace. Nyland [29] have argued that particular indulgence and uses of social network sites (SNSs) may moderate different social outcomes, such as civic and political involvement.

Sekiguchi et al. [24] worked on detection of various topics of blogs based on interest similarities of users. Eysenbach analyzed health-related searches on the World Wide Web and reported that in Google 6.75 million health-related search per day are conducted [25]. Nicholas et al. [26] studied the usage of content from five health-related web pages by analyzing transaction files. They also compared and analyzed the content in terms of thirt seven different health categories or five topics, respectively, and found substantial differences in content between the five topics. The content analysis relies on page headings only and misses, therefore, an in-depth content analysis. This work is extended by [27] while materials provided by different medical social media tool was analyzed.

Xiang et al. [7] investigated the social media in travel domain and analyzed the extent to which social media appear in search engine results in the context of travel-related searches. Chen [8]proposed a web-based tool named SWAB (Social Web Analysis Buddy) that aims to integrate both qualitative analysis and large-scale data mining techniques in Social media domain. The SWAB tool was designed to supports asynchronous collaboration among researcher's conducting inductive content analysis on textural data from user's on-line posts and conversations. It then aggregates the results and calculates the agreement among researchers, and builds modeling algorithms based on the qualitative results to classify large-scale social media text content.The tool was inspired from Atlasti [9] and NVivo [10] which were developed for qualitative analysis of social media data.

Table 1. Songs Summary

Characteristics	Views	Likes	Total Opinion
Minimum Value	290738	100	251
Maximum value	2239545	8523195	8631074
Mean	4720991	81339	85508
Standard deviation	4465374	746845	756203

In a survey of 1,715 college students conducted to examine their satisfaction with social networking (Facebook) along with their civic and political participation in daily life. The study concluded that a user's gratification is highly dependable on their demographic characteristics like gender, hometown and year in school [11].

Gilbert et al. [21] analyzed 340,000 online friendship and 200,000 interpersonal messages with the help of social net- working site to find the difference between rural and urban user and established that the rural people have fewer friends in their social network. The study was extended to analyzed gender wise distribution also that concluded that privacy feature is one the issue where different gender uses their privacy and security issues differently

4. Youtube Statistics

YouTube is among the most successful video sharing site. As per survey conducted by "Alexa", YouTube is the second largest in the world in terms of traffic [19].YouTube uses Sorenson Spark H.263 video codec, and its playback technology is based on Adobe Flash Player. The use of this techniques has made YouTube undisputed leader in online video sharing domain[20].A detailed technical report on YouTube and its functioning can be found at [18].

Some of the statistics related to YouTube are [1]:

- Number of Unique users who visit YouTube per month -more than 1 billion
- Hours of video watched per month-6 billion hours.
- Hours of video uploaded per 60 seconds-100 hours
- YouTube is available in 61 language
- The website is available as a localized version in 61 countries across the globe.
- Its reach is more than the conventional cable network of US.
- Addition of millions of subscribers per day.
- More than 1 million advertisers using Google Ad.
- Scanning of more than 400 years of video every year

5. Experimental Setup

5.1. Data Set Collection

Dataset for the experiment was collected from a popular content community site-www.youtube.com. A group of 30 undergraduate students (of age group 17-21, Native Hindi language speaking people) were randomly divided into 10 groups, each consisting of 3 students. Each group was asked to give the following details for 20 Hindi Language songs of their preferred choice on these four parameters , namely:

- Name of the songs(NAME)
- Total views of the song(VIEWS)
- Number of likes for the song(LIKES)
- Number of dislikes for the songs(DISLIKES)

A total of 200 Song details were collected. Of these, 130 songs were finalized after cleaning the data. Data cleaning process included removal of repeated songs , removal of songs with incorrect entries and then removal of Ten % songs from both the extremes according to the view. Time range of the song was from 1960's to present. Another attribute was added as:

Total Opinion = Likes + Dislikes.

Details of the dataset are provided in Table I.

5.2. Experiment

The analysis is divided into two separate mutually exclusive modules. First module involves use of Machine learning tool, WEKA (Waikato Environment for Knowledge Analysis) [12] to conduct regression analysis on the given attributes for possibly finding some relationship among various attributes while the second module includes finding distinct pattern of anonymous human behavior. Microsoft Excel [17] is used for this purpose.

Some regression outputs from WEKA tool are consolidated in Figure 3, 4, 5 and 6.

Some statistics obtained using Microsoft Excel are as follows:

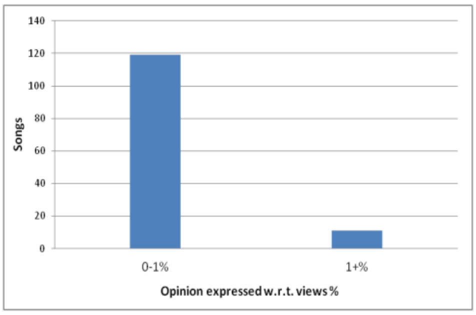

Figure 1. Opinion Percentage of the songs

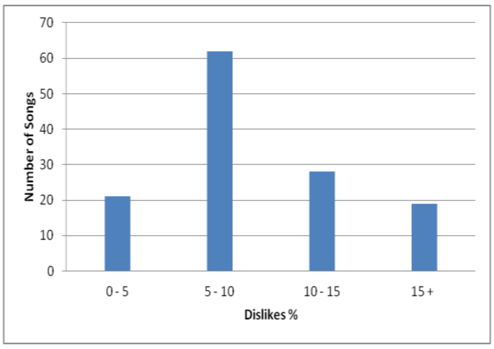

Figure 2. Dislikes percentage of the songs

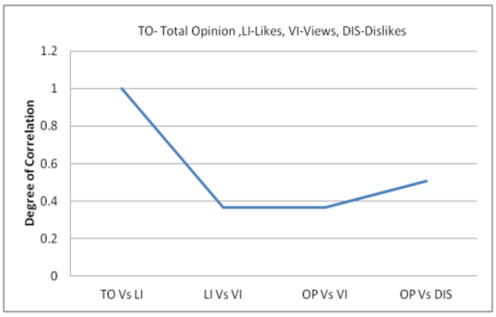

Figure 3. Degree of Correlation of parameters

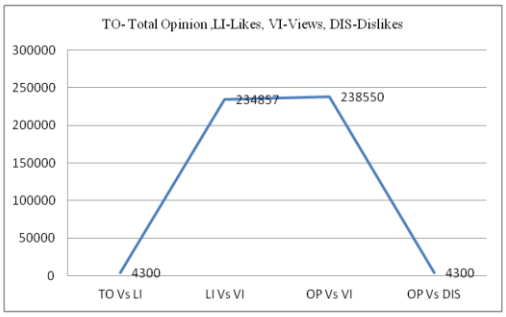

Figure 4. Mean Absolute Error of parameters

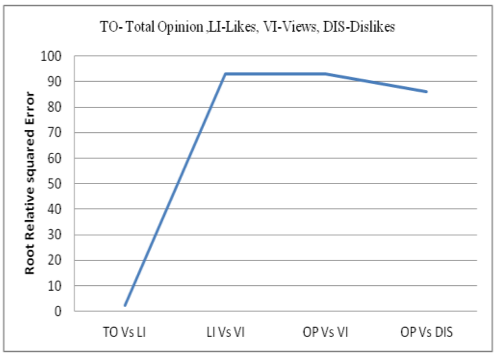

Figure 5. RRSE of parameters

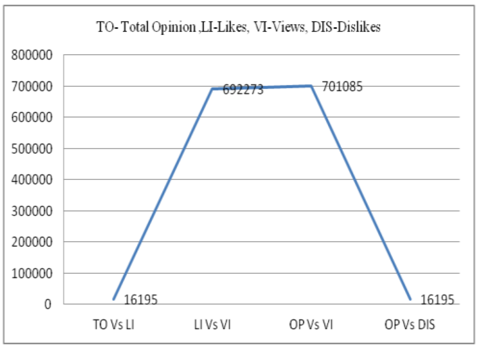

Figure 6. Root mean Squared error of parameters

6. Result analysis

RQ1: Can we depict Leadership, innovative and decision-making ability from this data?
Ans: Figure 1 can be referenced for this question. Figure 1 explains that out of 130 songs in analysis, only 119 songs received some kind of opinion and only 0-1% of total views can be converted into opinion. This fact emphasizes the actual nature of human behavior that very few people represent leadership quality, innovative behavior, and decision-making capability. This statement can be further authenticated by the fact that these opinions were collected in a significant scattered amount of time and mostly voluntarily without any supervision or without fear of exposing of identity.

RQ2: Do people have tendency to resist any popular theme.
Ans: Figure 2 represents the dislike percentage of the songs. The songs were collected and compiled randomly and according to the popularity by the students. So, based on Figure 2, it can be concluded that there will be some person who will always oppose any theme which is liked by other individuals. It can further analyzed any popular topic will be resisted by usually 5-10% of population of discourse.

RQ3: Can any relationship be established between different attributes.
Ans: Figure 3 is representing degree of correlation (DOC) between various attributes. It can be seen that Total Opinion and Likes are showing highest rate of correlation which can further analyzed to state that a relationship can be established among these two attributes. Figure 3, 4 and 5 further claims that various errors will be minimal for this relationship.

Figures 2, 3, 4 and 5 also conclude that the relationship between other attributes, namely “views and likes, Opinion and View” and “opinion and dislikes” cannot be easily established as various error rates are significantly high. “Opinion and Dislikes” is showing lesser rate of error in Figure 4 and Figure6, so it can be said that they represent a marginal tendency to have a relationship with a high degree of error.

7. Conclusion and Future Direction

In this paper, data from a variant of content community website, YouTube was collected and compiled for finding some of the traits of human behavior. As most of the data in this kind of websites are provided by the users, most of the time anonymously or without much fear of any other external sources, the finding of this study are more or less represents some human trait trends.

This work will be extended in the near future in following dimension to make the results obtained in this work more reliable.

1. Person: Future work will include data collection from more number of people of different age group, educational background and if possible, of different ethnic background to collect more number of people opinion.
1. Data: Data will be gathered based on the language, and various genre of music.
2. Time Period: Future result will also be based on the time period of songs for which the data is collected.
3. More number of traits as attributes will be included.
4. Text analysis: This work is based on numerical attributes but future work will comprise of “comments” of the songs also which will be analyzed with the help of some lexical analyzer like Sentiwordnet.

8. References

[1]https://www.youtube.com/yt/press/statistics.html (Access date: 28.08.2014).

[2] M. Hauben, and R. Hauben, ‘Netizens: On the history and impact of Usenet and the Internet,’ First Monday 3.7 (1998).

[3] A. Miura, ”Can weblogs cause the emergence of social intelligence?: causal model of intention to continue publishing weblog in Japan.” Ai, Society 22.2 (2007): 237-251.

[4] A. M. Kaplan, and M. Haenlein, ”Users of the world, unite! The challenges and opportunities of Social Media,” Business Horizons 53.1 (2010): 59-68.

[5] T. M. Harrison, and B. Barthel, ”Wielding new media in Web 2.0: exploring the history of engagement with the collaborative construction of media products.” New Media and Society 11.1-2 (2009): 155-178.

[6] R. L.Daft, and R. H. Lengel, ”Organizational information requirements, media richness and structural design.” Management science 32.5 (1986): 554-571.

[7] Z. Xiang, and U. Gretzel. ”Role of social media in online travel information search.” Tourism management 31.2 (2010): 179-188.

[8] X. Chen, K. Madhavan, and M. Vorvoreanu. ”A Web-Based Tool for Collaborative Social Media Data Analysis.” Cloud and Green Computing (CGC), 2013 Third International Conference on. IEEE, 2013.

[9] ATLAS.ti: The Qualitative Data Analysis and Research Software. [Online]. Available: http://www.atlasti.com/ index.html. [Accessed: 10-May- 2013].

[10] NVivo 10 research software for analysis and insight. [Online].Available: http://www.qsrinternational.com/ products nvivo.aspx. [Access date: 11- May-2013].

[11] N. Park, K. F. Kee, and S. Valenzuela. ”Being immersed in social networking environment: Facebook groups, uses and gratifications, and social outcomes.” Cyber Psychology & Behavior 12.6 (2009): 729-733.

[12] http://www.cs.waikato.ac.nz/ml/weka/ (Access Date: 07.02.2014)

[13]http://www.statisticbrain.com/facebook-statistics/ (Access Date: 29.08.2014)

[14] https://zephoria.com/social-media/top-15-valuable-facebook-statistics/ (Access Date: 29.08.2014)

[15] http://newsroom.fb.com/company-info/

[16]http://socialnetworking.findthebest.com/q/83/358/How-many-peopleuse-Flickr-social-networking-site (Access Date: 29.08.2014).

[17] Microsoft, (2007). Microsoft Excel [computer software]. Redmond, Washington: Microsoft.

[18] X. Cheng, C. Dale, and J. Liu, ”Statistics and social network of youtube videos,” Quality of Service, 2008. IWQoS 2008. 16th International Workshop on. IEEE, 2008.

[19] Alexa, http://www.alexa.com. (Access Date: 29.08.2014)

[20] YouTube: Video Format (Wikipedia), http://en.wikipedia.org/wiki/Youtube#Video format.

[21] E. Gilbert, K. Karahalios, and C. Sandvig, ”The network in the garden: an empirical analysis of social media in rural life,” Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 2008.

[22] J. Leskovec, L. A. Adamic and B. A. Huberman, ”The dynamics of viral marketing,” In Proceedings of the 7th ACM Conference on Electronic Commerce, 2006.

[23] B. Jansen, M. Zhang, K. Sobel, and A. Chowdury, ”Twitter power: Tweets as electronic word of mouth,” Journal of the American Society for Information Science and Technology, 2009.

[24]Y. Sekiguchi, H. Kawahima, H. Okuda, M. Oku. Topic detection from blog documents using user's interest", in proceedings of the seventh international conference on mobile data management.2006.

[25] Eysenbach, G., & Kohler, C. (2003). What is the prevalence of health-related searches on the World Wide Web? Qualitative and quantitative analysis of search engine queries on the internet. In *AMIA Annual Symposium Proceedings* (Vol. 2003, p. 225). American Medical Informatics Association.

[26] D. Nicholas, P. Huntington, J. Homewood. Assessing used content across five digital health information services using transaction log file. Journal of information science 29(6) (2003) 499-515.

[27] Kerstin Denecke, W. Nejdl. How valuable is the social media data? Content analysis of the medical web. Information Sciences 179 (2009) 1870-1880.

[28] Raacke JR, Bonds-Raacke J. MySpace and Facebook: applying the uses and gratifications theory to exploring friend networking sites. Cyber Psychology & Behavior 2008; 11: 169–74.

[29] Nyland R, Marvez R, Beck J., (2007) MySpace: social networking or social isolation? Paper presented at the midwinter conference of the Association for Education in Journalism and Mass Communication, Reno, NV

[30] J. Han, M. Kamber, "Data Mining concepts and techniques," Morgan Kauffman Publisher, 2nd edition, 2006.

[31] G. Mariscal, O. Marban, C. Fernandez, "A Survey of data mining and knowledge discovery process models and methodologies," The Knowledge Engineering Review, vol. 25:2 pp. 137-166, 2010.

[32] S. Kohli , A. Gupta,"An Ordered Weighted Operator Approach towards Web Usage Mining", 5th International Conference on Computer And Communication Technology (ICCCT-14) held at Motilal Nehru National Institute of Technology, Allahabad on 26-28 September 2014.

[33] Kumar, L., Singh, H., & Kaur, R. (2012, August). Web analytics and metrics: a survey. In *Proceedings of the International Conference on Advances in Computing, Communications and Informatics* (pp. 966-971). ACM.

[34] Kohli, S. and Gupta, A. (2014) 'Fuzzy information retrieval in WWW: a survey', *Int. J. Advanced Intelligence Paradigms*, Vol. 6, No. 4, pp.272–311.