Study on Sentimental Analysis for Intermedia-UGC

Yong Xu, Dongqin Li, Fengshan Si, Shuqin Huang

Department of Computer Science & Technology, Anhui University of Finance & Economics, Bengbu, China 233030;

Keywords: UGC; Sentimental tendency; Intermedia-Mining

Abstract. Intermedia user generated content is an important media which can be used to express personal sentimental emotion in the Internet. The paper defined the meaning of user's generated content; analyzed the essential characteristics in the view of subject, content, life cycle and media form. Finally, two key issues in the field were proposed.

1. Introduction

With the continuous development of information technology, more and more users like to publish text, pictures, audio, video and other information on network platform. This information can be product reviews, criticism and evaluation, tourism experience, entertainment, education, malicious attacks, news reports, photo sharing, campaign, social network, protest, which is defined user generated content of Web2.0 (UGC User, Generated Content). In recent years, UGC has changed people's way of producing, distributing, using information because of its brief contents, various media format, dynamic aggregation, independent terminal.

Emergence of a large number of UGC not only changes organization's working manner, but also affects network user's thinking habit and behavior. These changes provide a new challenge for business activities and information expression in Web 2.0[1].

It has great commercial and social benefits for people who make a decision based on sufficient information published in various Internet plat form (blog, forum) by more and more Internet users. The information includes personal opinion, individual unique experience of life, people's emotion (such as approval, objection, neutrality), because people could get more knowledge of the product and topic, and then make a scientific decision. As can be shown by a large number of facts that Internet users are more like to agree with the information provided by Internet users than organizations. But on the other hand, the number of UGC has increased, forms of UGC are various, because the UGC is open, innovative which can produced by ordinary users. These are disadvantages for users to understand it.

Sentimental Analysis (SA) on cross media UGC is an effective method to analyze such UGC based on data mining, text understanding, image processing and semantic analysis techniques. SA can mine various forms of media UGC information, identify its emotional tendency (happy, sad, or agree with etc.), and evolution of emotional with time changed.

2. User Generated Content

2.1 What's User Generated Content

Organization and specialist are not the only sources of Internet information because of the emergence of social network under Web2.0. Internet users can also produce and transmit information by generating information on Web which is named User Generated Content (UGC).

User Generated Content (UGC) is new model for people sharing their feeling, experience, opinion by publishing text, image, video. UGC could be used to review item and service, also to response a question in Web platform [2]. World Organization for Economic Cooperation and Development summarized three characteristics of UGC: 1) UGC could be published on Web; 2) UGC should have some novelty; 3) UGC is created by Internet users in common sense [3].

2.2. Characters of UGC

It is generally believed that UGC is often published by amateur audience to express opinion, feeling in the process of accepting some service. UGC can be shared by other audiences free of charge [4]. In theory, the holder of UGC should be all users of the Internet platform. UGC can be original created by Internet user, or can be copied from other sources legally [5]. In practice, the platform based on UGC model develops rapidly, its influence is increasing, such as Wikipedia, YouTube, FaceBook. The platform would be forum for publishing comments and user's reviews, even personal website [6].

UGC is mainly generated by the Internet users spontaneously, which generally do not have any profit purposes. In the case, the Internet users become the role of both producers and consumers from the simple identity of the consumer. Internet users may become specialist at creating such UGC from a random UGC author in the last stage. Such kind of creative groups may hold some purpose, but they would never consume similar works of the groups.

For content, UGC is discrete, de centralized, self-organization and so on. It has the characteristics of openness, community, connectivity. It also has the dual author identity of users and producers. Therefore, the content of UGC is great different with traditional media. As a new media form under the Web 2.0 environment, UGC shows the characteristics of fragmentation, collaborative creation, which emerged in new media, social networks, blogs, etc.

In the aspect of UGC life cycle organization form, it can achieve self-organization state when UGC satisfies certain conditions. Li has studied the UGC self-organization model from the three aspects of UGC self-organization level, self-organization content and self-organization evolution process mechanism[7].

In terms of information representation, UGC shows diversity, including text, images, videos, and external disk files. Most well-known user generated content platform have YouTube, social networking site MySpace, image sharing site Flickr, etc.

We can find that the early UGC mainly show the "content is king" concept by analyzing the UGC development process. Then UGC develops a new stage, which omits the boundary of individuals and organizations in the form of Internet[8].

3. Analyzing on Sentiment

With the increasing amount of UGC information on the Internet platform, the content of UGC is more and more complex, so the traditional analysis method is difficult to meet the needs of the analysis. Sentiment mining (Sentimental Analysis SA) is studied in this situation, which mines user sentiment information contained in a variety of media UGC by using data mining technology, also known as opinion mining etc.

Attitudes, opinions and views of the Internet users can be mined through the analysis on emotional information, which is expressed and transmitted in a variety of media UGC by the network users. The specific emotion analysis method can be studied from user rating view, also can be directly carried out by UGC mining, extraction of user sentiment. Sentiment mining process generally includes three stages of data acquisition and preprocessing, cross media information mining and application.

3.1 UGC Sentiment Analyzing

Most of the domestic and foreign researches on UGC analysis focus on the analysis of product reviews of text, especially on e-commerce products and tourism text format of UGC in first stage. In recent years researchers began to pay attention on UGC published in all kinds of new media platform include product reviews, discussions, experience sharing on sentiment mining problems.

Web 2.0 is thought as the technical foundation of UGC information ecology in academic view. Jenkins compared the characteristics of the new and old media, studied the integration of the new and old media, and pointed out that the new media has an important role in promoting the growth of UGC[9,10]. Baya et al studied review and browse number of video UGC, and made a conclusion that the higher the number of browsing is, the higher the popularity of the UGC is, and the more the number of comments UGC is, the higher the degree of concern is[11].

In recently we can find that the current trend of sentiment analysis is mainly focused on the text type UGC analysis of business or movie review. In the research method, the quantitative analysis methods have been used to model the user's sentiment, but the description of the sentiment category is still using the simple and discrete class representation.

3.2 Classis Method for Sentiment Analyzing

Most scholars adopt two kinds of models: supervised learning and unsupervised learning. Ni thought that sentiment classification is one of the two kinds of sentiment division, based on feature selecting by Bayes, SVM and Rocchio algorithm and information gain [10]. Tumey used the maximum entropy method for feature selection to classify movie reviews [12]. Das used the emotion lexicon to recognize emotion words, quantitative calculate emotional expressions of emotion information, at last the whole text sentiment was classified [13].

In recent years, studying on UGC has attracted many scholars. Most work on UGC analyzing of the Web2.0 service platform is still in the primary stage, for example, sort the reviews according to published time. This will cause the user to miss a lot of high quality UGC. Sometime Web2.0 service platform on UGC classifies UGC to positive and negative, while users need more delicate classifications.

3.3 Meaning of UGC Sentiment Analyzing

In practice, in the situation of a large number of UGC appearing in various forums, bulletin boards, portals and other network platform, users often want to obtain product performance characteristics with the help of UGC to aid decision making; on the other hand, companies want to accurately catch consumer’s response on business experience. It is impossible to get the important decision knowledge in time by analyzing such large number of UGC manually, so it is necessary to study the information technology which could mine the interesting knowledge from UGC effectively.

In a word, UGC sentiment analysis can reveal the user's internal driving mechanism, monitoring network public opinion, guiding micro blog marketing, and can help service providers to provide better personalized products and services. It also can be used to guide the virtual community, SNS community, blog, social quiz website, provide a theoretical basis and reference pictures, video sharing websites and other types of Web2.0 to develop well.

4. Technique of Sentiment Analyzing

It is the basis of realizing the cross media mining that how to describe the features, emerge multi-media UGC, classify UGC effectively. There are variety of models and methods developed in this field.

Xavier Ochoa et al pointed out that UGC has a long tail feature based on efficient empirical analysis on a large number of UGC data [14]. Li studied the advantages of supervised learning and designed an unsupervised feature extraction algorithm, which can be used to mine the sentiment orientation of text reviews [15]. Text mining method was applied to song emotion analysis by Xia, and an emotion vector space model (s-VSM) is designed, simulation experiment is carried out on [16]. Li et al analyzed hotel users’ reviews, selected hotel price, location, service and other objective data, and the user score, user comments, readability. At last, they designed a supervised hotel search ranking model based on the reviews [17]. Yang collected relevant data quantitatively, study the key factors and its mechanism affecting credibility of micro-blog by means of factor analysis and structural equation modeling method after he studied credibility index which can impact credit of micro-blog [1].

So the domestic and foreign scholars also are attracted on cross media UGC sentiment tendency mining technology. Because of the diversity of cross media UGC forms, extraction of UGC characteristics with different media forms, merged model affect the result of UGC sentiment analyzing. It could be concluded that data analysis model of cross media UGC sentiment analysis is a big problem.

5. Future Work

With the rapid development of information technology and e-commerce industry, various forms of media UGC will continue to emerge. Demand of the cross media UGC effective analysis methods and tools will be more obvious. Users are care about effectively using UGC knowledge obtained from sentiment analyzing, and then can make decision effectively.

We will study UGC feature extraction method, merging model of the different forms of media combine the basic theory of cross media UGC in the future. We will research cross media UGC sentiment classification problem, construct cross media UGC emotion influence mechanism on audience behavior for detail.

(1) Cross Media UGC Feature Extraction, uniform representation

Due to the cross media UGC comes from different network platform (including social networks, forums, blogs, etc.) UGC contains rich natural attributes (such as text, visual space, etc.) and social attributes (such as evaluation, heat and preferences). UGC also expresses a topic from various view by using text, image and video and location et al, therefore, unified representation and modeling method of cross media UGC data features will be an interesting research issue.

(2) Formal representation of UGC Sentiment

Due to the human emotion are various, user's sentiment should not be divided into positive and negative simply. According to the social attributes and its massive, heterogeneous, high dimensional , multi-dimension characteristics of cross media UGC data, feature distraction of UGC and representation of sentiment are important topics.

Acknowledgments

Supported by National social science foundation in China under Grant No. 15BTQ043.

References

[1] Xuecheng Yang, Tingting Ge, Bing Lan. Research on influential factors of the brand microblog credibility. Journal of Shanxi Finance and Economics University, 2013, 35(10): 68-80.

[2] Ping Wang, Qijie Chen. A study on the relationship between consumers' status of community and effectiveness of consumer-generated content. Library and Information Service, 2011, 3: 111-120.

[3] Graham Vickery, Sacha Wunsch-Vincent. Organization for Economic Cooperation and Development. Participative Web and User-Created Content: Web 2.0, Wikis and social networking edition complete[M]. Organization for Economic Co-operation and Development, 2007.

[4] Johan Ostman. Information, Expression, Participation: How Involvement in User-Generated Content Relates to Democratic Engagement amont Young People [J]. New Media & Society, 2012(6): 1004-1021.

[5] S Shim, B Lee. Internet Portals Strategic Utilization of UGC and Web 2.0 Ecology [J]. Decision Support Systems, 2009(47): 415-423.

[6] Consumer-Generated Media (CGM) 101: Word-of-Mouth in the age of the Web-Fortified Consumer [EB/OL]. [2012-12-28]. http://www.nielsenbuzzmetrics.com/ whitepapers.

[7] Peng Li. Self-organization of user generated content in Web 2.0 environment. Library and Information Service, 2012, 56(16): 119-126.

[8] Yuxiang Zhao, Zhe Fan, Qinghua Zhu. Conceptualization and research progress on user-generated content. Journal of Library Science in China, 2012, 38(9): 68-81.

[9] H Jenkins. Convergence Culture: Where Old and New Media Collide. 2006.

[10] X Ni, G Xue, X Ling, et al. Exploring in the Weblog space by detecting informative and affective articles. Proc. of the 16th Int. Conf. on World Wide Web, 2007, 281-290.

[11] Herbert Baya, Andreas Essa, Tinne Tuytelaars, et al. Speeded-up robust features (SURF) [J]. Computer vision and image understanding, 2008, 110(3): 346-359.

[12] P D Tumey, L M Littman. Measuring praise and criticism; Inference of semantic orientation from association. ACM Transactions on Information Systems, 2003, 21(4): 315-346.

[13] S R Das, M Chen. Yahoo! for Amazon: sentiment extraction from small talk on the web. Proc. of the 8th Asia Pacific Finance Association Annual Conf., 2001.

[14] Xavier Ochoa, Erik Duval. Quantitative Analysis of User-Generated Content on the Web [C]. Beijing: Web Evolve 2008: Web Science Workshop at WWW2008, 2008.

[15] Shi Li, Qiang Ye, Yijun Li et al. Mining product features and sentiment orientation from Chinese customer reviews [J]. Application Research of Computer, 2010(8): 3016-3019.

[16] Yunqing Xia, Ying Yang, Pengzhou Zhang. Lyric-Based Song Sentiment Analysis by Sentiment Vector Space Model [J]. Journal of Chinese Information Processing, 24(1): 99-103

[17] B Li, A Ghose, P G Ipeirotis. Towards a theory model for product search[C]//Proceedings of WWW 2011, March 28-April 1, 2011.