Predicting Application Review Rating with SAS® Text Miner

Tianxi Dong, Jonghyun Kim, Zhangxi Lin,
Rawls College of Business Administration, Texas Tech University, Lubbock, Texas 79409
Rawls College of Business Administration, Texas Tech University, Lubbock, Texas 79409
Rawls College of Business Administration, Texas Tech University, Lubbock, Texas 79409

ABSTRACT

With the wide proliferation of text-based data on the Internet, there is a need for dealing with the information overload. The large amount of online user reviews may present an obstacle to developers who want to know users' feedback and potential customers who are interested in applications. Here we employ text analysis provided in SAS® Text Miner to predict the overall and feature-based ratings for online application reviews. We use examples from Android Market and Apple Store, the real world of online application stores. The findings may aid in promoting the sales of applications by better satisfying customer demands.

Keywords: online reviews, application store, text mining, Android Market and Apple Store

INTRODUCTION

Recently with the popularity of smart phones and tablet personal computers, thousands of application developers have flourished. The sale of applications is important for both developers and application. The rating information, however, is not informative enough for developers to improve the applications. Developers must also understand the textual content of user reviews. Many users provide the reasons why they do not like the application, for example, "Game is awesome! However, needs more levels and lags quite often." The developers can improve the application by increasing levels and reducing the lags. If we only focus on the star rating, no one will know how to improve this application or why this application is so popular. In order to assist the developers to more easily find helpful review information to improve the application, a text mining based method of summarizing online reviews was investigated.

RESEARCH METHODS

Step 1: Latent features identification in the review data
Step 2: Manual Coding
Step 3: Prediction with SAS® Enterprise Miner

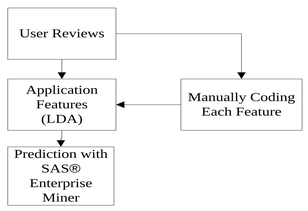

Figure 1. Research Conceptual Model

DATA COLLECTION

The *Android Market and Apple Store* and *Apple Store* offer users a friendly feedback system to share their opinions and experiences about the applications. The user review system includes detailed comments and a five-star user rating system.

To collect review data, we chose Angry Birds, one of the most popular applications in the Android Market and Apple Store, as our target application. The total number of user reviews was 953,619 (from Android Market) and 817,913 form Apple Store.

Figure 2a. Rating distribution of Android Figure 2b. Rating distribution of iOS

TEXT MINING WITH SAS® ENTERPRISE MINER

Manual Coding

Two people coded scores of each feature independently based on reviews. The final result of the scoring is the average of two independent scoring. Score 1 and 2 indicate negative scores, 4 and 5 indicate positive scores for each feature. Score 3 indicates that review does not mention the feature. In addition, for increasing accuracy of scoring, the two coders discussed and got consensus on records that have larger difference than the threshold.

	ID	OS	Ver	Entertainment	Quality	Performance	Interface	Rating	Customer_Reviews
2	1	iOS	2.3.0	3	3	1	3	2	Why would you bother re leasing an update that doesn't support the new phone?
3	2	iOS	2.3.0	3.5	3	1	3	1	What exactly are the developers doing? The biggest game on the iPhone and still no support for the iPhone 5.
4	3	iOS	2.3.0	3	1	2	3	3	I don't have the iPhone 5 so ... But I just downloaded the new update and it made the whole game very low quality so I don't s
5	4	iOS	2.3.0	4	1.5	3	3	3	The new bad piggies update just came outfor Angry Birds! I was exited! Until I updated it and it toteld my graphics! Fix it!
6	5	iOS	2.3.0	4	2	2	2	3	C'mon guys ... one of the most popular apps on the App Store and several weeks after the iPhone 5 is released Angry Birds st
7	6	iOS	2.3.0	3.5	3.5	1	3.5	1	Will give five stars when updated
8	7	iOS	2.3.0	3.5	3.5	1.5	3.5	4	5 stars when you can update for new iPhone 5 screen.

Figure 3. Examples of Scored Dataset

Prediction Model

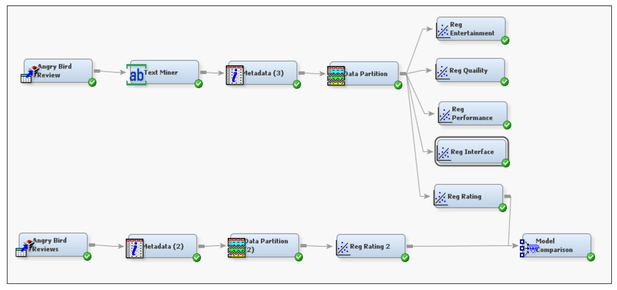

Figure 4. Prediction Model

SAS® Enterprise Miner is used to set up a text analysis project. The software component gives flexibility in terms of setup and storage so that you can identify the location of the project and the corresponding data sets.

RESULTS AND DISCUSSION

Application Features	Related words	Positive Frequency	Related words	Negative Frequency
Entertainment	love	49	awful	4
	great	16		
Quality	mighty	5	bad	31
			terrible	21
Performance			late	3
			sad	4
Interface	amazing	2	annoying	4
			difficult	2

Table 1. Results Summarization of Text Mining Analysis

Figure 5a to 5d separately show the terms related to Game (Entertainment Feature), Graphics (Quality Feature), Update (Performance Feature) and Advertisement (Interface Feature). Figure 5e is the terms related to the game *Angry Birds*.

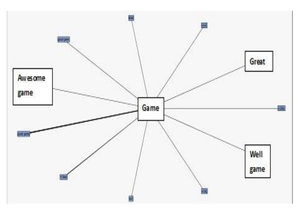

Figure 5a. Terms Related to Game

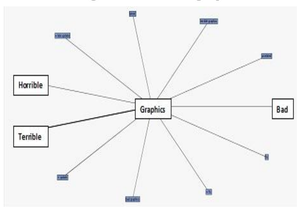

Figure 5b. Terms Related to Graphics

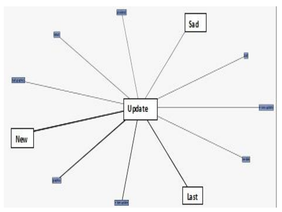

Figure 5c. Terms Related to Update

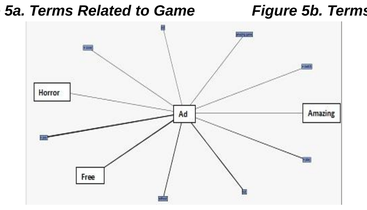

Figure 5d. Terms Related to Advertisement

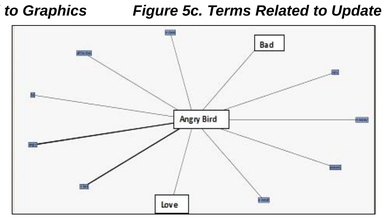

Figure 5e. Terms Related to Angry Bird

Figure 6 shows the result of prediction model for overall rating using a linear regression model. SAS® Enterprise Miner analyzed reviews using a text mining method. Using the result of the text mining, our linear regression model predicts the overall rating of the review.

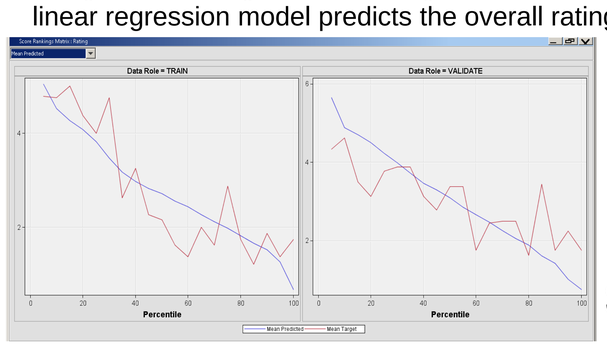

Figure 6. Performance of Prediction Model

CONCLUSION

In this paper, we examined a previously ignored yet important research question concerning the online user reviews: How can we enhance the communication between application developers and users? We addressed this question by investigating a text mining based method for mining and summarizing the users' opinions about the applications, which are contained in the online reviews. This study examined how to identify the critical information contained in the online reviews by employing text mining method. The major contribution of this study is to add to understanding on how to identify the features users most frequently mentioned in the reviews and the corresponding sentiment towards these features. This study also has significant implications for website designers in that it can guide them in designing multi-dimensional rating mechanisms that may satisfy users' diversified demands.

Reference

1. Blei, D.M., Ng, A.Y., & Jordan, M., I (2003) Latent dirichlet allocation. Journal of Machine Learning Research, 3, 993–1022.
2. Blei, D. M. & Lafferty, J. D.(2006). Dynamic topic models. The Proceedings of the 23rd international conference on machine learning (ICML2006), page 113-120, Pittsburgh, Pennsylvania, USA.
3. Cerrito, P. (2011). Sentiment Mining Using SAS® Text Miner®. Paper presented at the SAS Global Forum 2011, Las Vegas.
4. Clauset, A., Newman, M. E. J., & Moore, C. (2004). Finding community structure in very large network. Physical Review E, 70, 066111
5. Duan, W., Gu, B., & Whinston, A. B. (2008). Do online reviews matter? — An empirical investigation of panel data. Decision Support Systems, 45(4), 1007-1016. doi: 10.1016/j.dss.2008.04.001
6. Erdos, P., & Renyi, A (1959) On random graphs. I, *Publicationes Mathematicae, 6*, 290-291.
7. Ghose, A., & Ipeirotis, P. G. (2011). Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics. Knowledge and Data Engineering, IEEE Transactions on, 23(10), 1498-1512.
8. Girvan, M., & Newman, M. E. J (2002) Community structure in social and biological networks. Proceedings of the National Academy of Sciences of the United States of America, 9912, 7821–7826.
9. Hu M.Q. and Liu B. (2004), Mining and Summarizing Customer Reviews, Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 168-177.
10. Iwata, T., Yamada, T, Sakurai Y, & Ueda, N. (2010). Online Multiscale Dynamic Topic Models. The 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 7.
11. Li, D., He, B., Ding, Y., Tang, J., Cassidy, S., Qin, Z., Yan, E., Li, J., and Dong, T. Community-based Topic Modeling for Social Tagging. In Proceedings of the Nineteenth Conference on Information and Knowledge Management (CIKM'2010). (short paper).
12. Li, D., Li, J., & Lin, Z. (2008). Online consumer-to-consumer market in China – A comparative study of Taobao and eBay. Electronic Commerce Research and Applications, 7(1), 55-67. doi: 10.1016/j.elerap.2007.02.010
13. Milgram, S (1967) The small world problem, Psychology Today, 2, 60–67.
14. Newman, M. E. J (2001) The structure of scientific collaboration networks. Proceedings of the National Academy of Sciences of the United States of America, 982, 404-409.
15. Newman, M. E. J (2003) The structure and function of complex networks. Siam Review 45, 167–256 .
16. Nikolay Archak, Anindya Ghose, Panagiotis G. Ipeirotis, "Deriving the Pricing Power of Product Features by Mining Consumer Reviews," *Management Sciences*, Vol. 57, No. 8, August 2011, pp. 1485–1509.
17. Shirky, C. (2005). Ontology is overrated: Categories, links, and tags. Retrieved May 20, 2010 from http://shirky.com/writings/ontology overrated.html.
18. Tan, C., Tang, J., Sun, J., Lin, Q and Wang, F. Social Action Tracking via Noise Tolerant Time-varying Factor Graphs. In Proceedings of the Sixteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD'2010). pp. 1049-1058.
19. Toffler, A. (1980): The Third Wave, London et al. (Collins).
20. Torunski, L. (2009). Smart and simple webcrawler. Retrieved May 20, 2010 from https://crawler.dev.java.net.

Dong,
Kim,
Lin

SAS® GLOBAL FORUM