A Text Mining Method to Trace Influential Product Features Based on PageRank

Qin Liu, Zhuchao Yu* and Shucheng He
School of Business Administration, Northeastern University, No. 195 Chuangxin Road, Hunnan District, Shenyang, 110169, China.
*Correspondence: zcyu@mail.neu.edu.cn

Abstract. As a technology of Web2.0, social media has become a popular tool for both enterprises and customers to understand other voices. Therefore, an important project for enterprises to have an outstanding feedback is to use social media to promote and spread products correctly, so it's necessary to mine some product features which customers care mostly. In this paper, we crawled online data from website DouBan to trace those features through their posted contents, such as descriptions, reviews, comments, etc. Firstly, we computed received attention of these contents by self-information algorithm and measured similarity by cross-entropy algorithm. Secondly, using the attention value as node weight and similarity as edge weight, we computed influence of product features in PageRank network by the proposed method. Experimental results show that it is books’ authors, genres and themes that consumers consider of mostly when they choose a book commonly. Furthermore, when it comes to a novel, the feature character is also one of the focuses while other features are not as important as them. This research informs a way for enterprises to choose emphasis in an online promotion, which can be very helpful for recommending products to target customers in a proper way.

1. Introduction
With the advent of Web2.0, every individual shares their opinions, reviews, interests and preferences via social media daily, thus it is useful to mine various data to extract knowledge from it with a sufficient quality [1].

Enterprises gradually notice the benefits, in terms of cost reductions, improvements in customer relation and enhanced accessibility of information [2], of social media to propagate products, and join in the big world of social network but it also can be a challenge [3].

Social media allows users to share their good or bad opinions, which supports other users’ purchasing behaviours [4]. It also provides enterprises with a detailed profile of products. Enterprises, as participants of social media websites, absolutely can construct articles for marketing either. Thus, how to use social media for marketing is an attractive topic for enterprises [3].

The focus of our study is to determine what features of products make their propagation successful and then help enterprises find the emphasis of recommendation. As such, this study takes a step in exploring what kinds of posted contents are considered as popular, finding the most influential factors of products to achieve better product propagation and even more purchase orders.
2. Related Work

2.1. Received Attention Measure
Although social media spreads rapidly, some generated contents don’t get any likes or retweets and cannot be diffused [5]. As a result, enterprises need to have an understanding of the characteristics of products that drive the diffusion on social website [6].

Some research works in the influence of certain tweet elements, such as hashtags, mentions, number of words or links, on their diffusion among the virtual community [7]. They use the number of likes and retweets as measures of popularity and dissemination capabilities [5]. However, only the number of likes or retweets allows potential generated content to get lost and unused [8]. In consequence, Gao [9] and his co-writers attempt to use self-information to measure the popularity, because self-information considers the amount of uncertainty one has.

2.2. Relationship Measure
Relationship often refers to connectedness in networks, which involves two subjects, and is always defined by similarity. As for this research, two subjects are both in form of text. Prior researchers define text similarity by text vector distance [10]. However, different research topics apply different methods.

A piece of text also can be seen as a probability distribution of words. To measure the similarity between two probability distributions, cross entropy [12] is a special method, which is an index for judging optimized value of the optimization model. Specifically, cross entropy performs well and shows efficiency and stability in previous research [13], which this paper chooses to be in line with.

2.3. Influence Measure Considering Networking
Since Larry Page and Sergey Brin [14] developed PageRank algorithm and applied it in a field of web page ranking, PageRank has gotten lots of concern, especially personalized PageRank. For example, Xu [15] and Ding [16] measured the spreading ability of nodes with the function of the weighted PageRank network.

At present, PageRank is normally developed in various fields and becomes mature and easily applied, which offers a perfect reason for employing it in this research.

2.4. Product feature Extraction
Emphasis of feature extraction varies with topics. For example, Christian [17] focuses on latent geographic features of social media while Neumann [18] concentrates on content features of social media behaviours. Since it is the most influential feature that this research aims to trace, it is essential to define product features properly and accurately.

Overviewing the previous researches about product features, they find influential characteristics of products from review extraction results by methods such as PageRank [19]. However, there are already defined product features in many e-commercial websites and it can cool the process. Take books for instance, the features are listed in table 1.
Table 1. Features of books.

| No. | Features       | Word example            |
|-----|----------------|-------------------------|
| 1   | Book title     | The long night          |
| 2   | Author         | Zi Jin, Zi Jinchen      |
| 3   | Publisher      | Publisher               |
| 4   | Content        | Challenges, evidence, testimony |
| 5   | Character      | Jiangyang, the prosecutor, witnesses |
| 6   | Genre          | Novel, crime, detective, reasoning |
| 7   | Year/Location  | China, USA, Korea, Japan |
| 8   | Theme          | Growth, brotherhood, love, justice |
| 9   | Quality        | Special, wonderful, attractive |
| 10  | Others         | Movie, DouBan, adaptation, category |

3. Method

3.1. Data Collection

In this study, the data was collected by web crawling from DouBan (www.douban.com), consisting of 3,305 records in form of 1,554 pieces of reviews and 1,701 pieces of comments generated in ranging from 2016 to 2018, involving 50 books. DouBan offers a popular and open community for sharing ideas about books, movies and music, so it is an ideal platform for our research.

Furthermore, $C$ represents the set of reviews with size of $n$ and $B$ represents the set of books with size of $m$.

3.2. Review Received Attention

Review popularity reflects how other users like the post and their willing to share it. As is said previously, the number of likes and retweets represents received attention of generated content and also self-information algorithm considers uncertainty. So, this study is going to be in line with the former researches, using the number of likes, retweets and other user feedback data and the self-information computation. Thus, the popularity $Pop_i$ of review $i$ is measured by Equation (1).

$$Pop_i = -\ln \frac{1}{\text{like}_i + \text{disagreement}_i + \text{comment}_i}$$ (1)

where $\text{like}_i$ represents its number of likes, $\text{disagreement}_i$ represents its number of disagreements and $\text{comment}_i$ represents its number of comments.

3.3. Relationship between Review and Product Characteristics

To help enterprises have an understanding of features of products that prompt propagation, first thing to do is to have an understanding of the relationship between review and product characteristics. The meaning of generated content is always expressed in form of text, though there are also something like hashtags, links and pictures. Therefore, this relationship is about review text and product description.

Since posted contents are in form of text, word segmentation cannot be avoided. First step, segment these texts into word set $W_r$ and $W_h$, and then illustrate the relationship after having found the similarity or difference. And the similarity measurement the study use is based on cross-entropy, as was noted before, the cross-entropy value can be represented by $CE_{rh}$, which is as Equation (2) shown,
\[ CE_{rb} = -\sum_{w \in W} P_r(w) \log P_b(w) \]  

(2)

in which \( W = W_r \cup W_b \), \( P_r(w) \) represents the probability distribution of one piece of review \( r \) and \( P_b(w) \) represents the probability distribution of one product \( b \). In other words, \( CE_{rb} \) also symbolizes the gap between review \( r \) and product \( b \).

3.4. Influential Features of Product

Since only one type of node is allowed in PageRank network, some translation should be done. Previously, the relationship is between one product and one piece of review. Therefore, there we need to change the subjects of the relationship to one piece of review and another piece of review and eliminate distractions from products, which is achieved by the following Equation (3).

\[ \text{Div}_{r_1 r_2} = \frac{1}{m} \sum_{b \in \mathcal{B}} |CE_{r_1 b} - CE_{r_2 b}| \]  

(3)

where \( \text{Div}_{r_1 r_2} \) represents the divergence between review \( r_1 \) and review \( r_2 \), and its value is the average summarization of every absolute value of the subtraction of \( CE_{rb} \) and \( CE_{r'b} \) that represent the cross-entropy values of different reviews and the same product \( b \). However, it can be noticed that \( \text{Div}_{r_1 r_2} \) is negative to their similarity. Consequently, a mathematical change would better be made. The final edge weight \( W_{r_1 r_2} \), which involves review \( r_1 \) and review \( r_2 \) can be computed by Equation (4).

\[ W_{r_1 r_2} = \max_{i \in \mathcal{C}, j \in \mathcal{C}} CE_{r_1 i} - CE_{r_2 i} + \varepsilon \]  

(4)

In the PageRank network, popularity is as the node weight. The weight of edges between review nodes is also ready. Next, the final PageRank score, which symbolizes influential capabilities of reviews, will be got when a value between two iterations reaches within the setting limitation. The PageRank score \( PR \) is valued by Equation (5) as below.

\[ PR = \alpha p + (1 - \alpha) W \cdot PR \]  

(5)

where \( \alpha \) \((0 < \alpha < 1)\) is the dumping coefficient and set 0.15 here, and that

\[ p = \left( \text{Pop}_1, \text{Pop}_2, \ldots, \text{Pop}_n \right)^T \left( \sum_{i=1}^n \text{Pop}_i \right)^{-1} \]  

and \( W = \left( W_{r_1 r_2} \right)_{\text{max}} \).

As for now, \( PR \) score as the influence of review is achieved. To find the most influential factor in the acceleration of the propagation and diffusion duration, the computed influence has to be translated into the influence of factors. As is said earlier, it can be obtained in the following way. Moreover, \( PR(w) \) represents the PageRank score of each one word satisfying the limitation and its computing equation is Equation (6).

\[ PR(w) = \beta \cdot PR(C_r) \]  

(6)

in which \( \beta \) is frequency of word \( w \) in this review \( r \).

According to the previous setting, product characteristics can be divided into 9 features. Thus, the content of the review must have been all about these aspects. So, one product’s influence of one aspect \( \text{Infl}(\text{Fea}_i) \) can be gained by Equation (7).
\[ \text{Infl}(\text{Fea}_i) = \sum_{r \in C} \sum_{w \in W_r} \text{PR}(w) \] (7)

where \( C \) is the set of reviews, \( W_r \) is the set of word segmented from review \( r \) and \( \text{Fea}_i \) represents the No. \( i \) feature listed in table 1.

Of course, if we want to summarize some common points to discover the aspects that users are commonly focusing on, something further should be made.

3.5. Results and Discussion

The average feature score gotten by the proposed method of 50 books is shown in figure 1, where we can see that genre, theme, author and character rank top 4 evidently. In other words, for this sample, users discuss more about genre, theme, author and character of books on average.

In order to discover whether there are differences existing between different books, we exemplify the difference by listing 4 different types of books, as is listed in table 2 below. As shown in column 'Homo Deus', genre, author and theme rank as the three highest priorities when people are talking about Homo Deus: A Brief History of Tomorrow (Homo Deus), which is well-known that this book is about future from the point of view of Jacques, the author, and is about prediction of changes in the future. As a result, it is reasonable for users to talk more about those three features.

| Ranking | Homo Deus       | I Am A Novelist | The Long Night | L'amica Geniale |
|---------|-----------------|-----------------|----------------|-----------------|
| 1       | Genre           | Location        | Author         | Author          |
| 2       | Author          | Genre           | Genre          | Genre           |
| 3       | Theme           | Theme           | Location       | Theme           |
| 4       | Others          | Author          | Character      | Character       |
| 5       | Location        | character       | Publisher      | Publisher       |
| 6       | Publisher       | Others          | Theme          | Location        |
| 7       | Quality         | Quality         | Quality        | Content         |
| 8       | Content         | Publisher       | Others         | Others          |
| 9       | character       | content         | Content        | Quality         |

The book I Am A Novelist is a biography of Haruki Murakami and it talks about his writing, life and attitudes. So, its ranking result is a bit different from book Homo Deus’ s. Its location ranks at first, which is followed by genre, theme, author.

Even though The Long Night and L'amica Geniale are novels, there is still discrimination, as table 2 shown. But the most popular thing is even about their authors and genres and details, which varies with books themselves.
Having looked through these ranking results, it can be easily to understand that though every individual result looks different, these three features, author, genre and theme, are close to each other magically. Consequently, it is not hard to conclude that author, genre and theme are the three features of books that users care about most.

4. **Conclusion**

In this paper, we have investigated potential features for promoting products. By using content and reviews, some commonly prevalent features of products have been traced. Even though detailed features vary with books, what users and customers discuss always keep steady via social media. Thus, it is very convenient and efficient to present an emphasis in the process of promotion to attract customers’ eyes and recommend proper products attractively to improve the ratio of translating interestingness to purchase behaviours. At present, we have only carried out a partial research. Further investigation is needed to dig out more detailed and subdivided features.

5. **Acknowledgements**

Our thanks to the National Natural Science Foundation of China under projects No. 71171043, and the National Training Programs of Innovation and Entrepreneurship for Undergraduates in Northeastern University, China.

6. **References**

[1] Sánchez D L, Revuelta J, De la Prieta F, Gil-González A B and Dang C 2016 Twitter User Clustering Based on Their Preferences and the Louvain Algorithm *Trends in Practical Applications of Scalable Multi-Agent Systems, the PAAMS Collection. Advances in Intelligent Systems and Computing* vol 473 (Berlin: Springer) pp 349–56

[2] Tajudeen F P, Jaafar N I, Ainin S 2018 Understanding the impact of social media usage among organizations. *Information & Management* vol 55 (Amsterdam: Elsevier) pp 308–21

[3] Gu C and Wang S 2012 Empirical Study on Social Media Marketing Based on Sina Microblog 2012 *Second Int. Conf. on Business Computing and Global Informatization (BCGIN)* (October 12–14, 2012 China: Shanghai) pp 537–40

[4] Schmunk S, Hüpken W, Fuchs M and Lexhagen M 2013 Sentiment Analysis: Extracting Decision-Relevant Knowledge from UGC *Proc. Int. Conf. on Information and Communication Technologies in Tourism 2014* (January 21–24, 2014 Ireland: Dublin) pp 253–65

[5] Alboqami H, Al-Karaghouli W, Baeshen Y, Erkan I, Evans C and Ghoneim A 2015 Electronic word of mouth in social media: the common characteristics of retweeted and favourited marketer-generated content posted on Twitter *Int. Journal of Internet Marketing and Advertising* vol 9 (Geneva: Inderscience) pp 338–58

[6] Lahuerta-Otero E, Cordero-Gutiérrez R and Prieta-Pintado F D 2018 Retweet or like? That is the question *Online Information Review* vol 42 (Bingley: Emerald) pp 562–78

[7] Hao X, Zheng D, Zeng Q and Fan W 2016 How to strengthen the social media interactivity of e-government: Evidence from China *Online Information Review* vol 40 (Bingley: Emerald) pp 79–96

[8] Mackiewicz J and Yeats D 2014 Product Review Users’ Perceptions of Review Quality: The Role of Credibility, Informativeness, and Readability *IEEE Transactions on Professional Communication* vol 57 (New York: IEEE) pp 309–24

[9] Gao Z, Hu Y and Zhang L 2013 A Epidemic Style Super-node Election Method Based on Self-information Theory 2013 15th Int. Conf. on Advanced Communications Technology (ICACT) (Jan. 27 – 30, 2013 South Korea: PyeongChang) pp 63–71

[10] Zhang K, Luo J and Chen X 2013 Text similarity computing based on sememe Vector Space 2013 4th *IEEE Int. Conf. on Software Engineering and Service Science (ICSESS)* (May 23–25, 2013 China: Beijing) pp 208–11

[11] Spark J K 1972 A statistical interpretation of term specificity and its application in retrieval *Journal of Documentation* vol 28 (Bingley: Emerald) pp 11–21
[12] Reuven Y Rubinstein 1997 European Journal of Operational Research vol 99 (Amsterdam: Elsevier) pp 89–112
[13] Farhadinia B and Xu Z 2018 Novel hesitant fuzzy linguistic entropy and cross-entropy measures in multiple criteria decision making Applied Intelligence vol 48 (Berlin: Springer) pp 3915–27
[14] Brin S and Page L 1998 The anatomy of a large-scale hypertextual Web search engine Computer Networks and ISDN Systems vol 30 (Amsterdam: Elsevier) pp 107–17
[15] Xu S, Wang P and Zhang C 2019 Identification of influential spreaders in bipartite networks: A singular value decomposition approach Physica A: Statistical Mechanics and its Applications vol 513 (Amsterdam: Elsevier) pp 297–306
[16] Ding Z, Jia Y, Zhou B, Han Y, He L and Zhang J 2013 Measuring the spreadability of users in microblogs Journal of Zhejiang University SCIENCE C vol 14 (Hangzhou: Zhejiang University Press) pp 701–10
[17] Sengstock C and Gertz M 2012 Latent geographic feature extraction from social media Proc. 20th Int. Conf. on Advances in Geographic Information Systems (SIGSPATIAL ’12) (November 06–09, 2012 USA: California) (New York: ACM) pp 149–58
[18] Neumann S, Li C, Lo C, Lee C, Rajwani S, Sood S, Foster B A, Hadgis T, Savir Y, Michaels F and et al 2017 Content Feature Extraction in the Context of Social Media Behavior Int. Conf. on Augmented Cognition. AC 2017. (July 9–14, 2017 Canada: Vancouver) Augmented Cognition Neurocognition and Machine Learning. Lecture Notes in Computer Science (Berlin: Springer) p 10284
[19] Yan Z, Xing M, Zhang D and Ma B 2015 EXPRS: An extended pagerank method for product feature extraction from online consumer reviews Information & Management vol 52 (Amsterdam: Elsevier) pp 850–58