Expert Systems with Applications 41 (2014) 1680–1689

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Box office prediction based on microblog

Jingfei Du [a,b], Hua Xu [a,*], Xiaoqiu Huang [a,b]

[a] *State Key Laboratory of Intelligent Technology and Systems, Tsinghua National Laboratory for Information Science and Technology, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China*
[b] *Beijing University of Posts and Telecommunications, Beijing 100876, China*

ARTICLE INFO

Keywords:
Box office
Microblog
Social media
Prediction model

ABSTRACT

As the importance and popularity of online social media has become more obvious, there are more researches aiming at making use of information from them. One important topic of this is predicting the future with social media. This paper focuses on predicting box offices using microblog. Compared with previous work which makes use of the count of related microblogs simply, the information from social media has been utilized more deeply in this paper. Two sets of features have been extracted: count based features and content based features. For the former, the information in the aspect of users, which decrease the influence of garbage microblogs, has been exploited. For content based features, a new box office oriented semantic classification method has been provided to make the features more relative with box offices. Meanwhile, more complex machine learning models such as SVM and neutral network have been applied to the prediction method. Our prediction model is more accurate and reliable. With our prediction method, the data in Tencent microblog has been utilized to predict box offices of certain movies in China. With the results, the strength of our method and predictive power of online social media can be completely demonstrated.

1. Introduction

With the development of World Wide Web, people would like to spend more time on the Internet as well as the social media online. Because of their convenience and ease of use, social media, such as microblog, twitter and facebook, have become platforms for recording experiences, sharing opinions and communications. Due to social media's character of high popularity, the opinions that people express on social media can, to some extent, represent the opinions of most people in the real world. From this respect, the social media can be precious resources of information for researchers to observe, understand, and even explore the events and laws of the real world.

This paper studies the power of microblog to predict events and outcomes in real world. More specifically, our goal is to predict the box office of movies using the information we extract from microblog. Moreover, our work demonstrates the predictive power of microblog as well as online social media. Besides the box office, our method can also be used to predict other outcomes in the real world such as sales volume and stock.

Making good use of information from online social media is an important issue in this age. One aspect of this is predicting outcomes and data in real world using social media. There are many researches on it. However, no one can be sure to say the prediction methods based on social media are feasible and accurate to predict all the outcomes. Around this problem, researchers can be divided into two categories. Some researchers believe that information from social media can be used to predict data from real world and provide some practical prediction methods. On the contrary, through their experiments and comparison with traditional methods, other researchers demonstrate that prediction methods based on social media are not appropriate to predict some outcomes and events.

People try various methods to extract information from social media and construct different prediction models to predict various things. For example, Guille and others built a predictive model for the information diffusion in online social networks (Guille & Hacid, 2012). Nguyen and others presented a strategy of building statistical models from the social media dynamics to predict collective sentiment dynamics (Nguyen, Wu, Chan, Peng, & Zhang, 2012). Gupta and others predicted the credibility of information in a tweet (Gupta & Kumaraguru, 2012). Besides these, there are lots work about similar problems (De Choudhury, 2009; Gilbert & Karahalios, 2009; Hong, Dan, & Davison, 2011; Huang, Chen, Luo, & Lee, 2012; Lerman & Hogg, 2010, 2012; Lerman, Intagorn, Kang, & Ghosh, 2012; Qi, Qu, & Tan, 2012; Rosenthal & McKeown, 2011; Tang & Liu, 2009). However, above work is predicting information

[*] Corresponding author at: State Key Laboratory of Intelligent Technology and Systems, Tsinghua National Laboratory for Information Science and Technology, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China. Tel.: +86 15117942752.
E-mail address: du1342157416@gmail.com (J. Du).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.08.065

of social media themselves but not very relative with the real world. There is no direct comparison between our work and their methods. On the other hand, there is also much work on predicting the real outcomes and events with social media. Asur and others used data from twitter and built a model based on the rate at which tweets are created about particular movies as well as the sentiment polarity of these tweets to predict the box offices of the movies (Asur & Huberman, 2010). In their paper, they provided a general prediction model for social media. However, they did not demonstrate the effectiveness of this model through experiments and their model is so simple that can be influenced by various factors. Liu and others focused on content information of blog (Liu, Huang, An, & Yu, 2007). They used sentiment PLSA to find the hidden sentiment factors in blogs and built an autoregressive sentiment-aware model to predict product sales performance. They concentrated on the opinions and sentiments in the blogs. However, because their dataset is from blog, which are long-text social media, they could not make use of features based blogs' count. And Bar-Haim and others predicted stock price movement using stock microblogs (Bar-Haim, Dinur, Feldman, Fresko, & Goldstein, 2011). Their prediction method is mainly finding expert investors and collecting experts' opinions. Besides, there are other researchers who predicted movie ratings and stock market level with twitter data (Mao, Wang, Wei, & Liu, 2012; Oghina, Breuss, Tsagkias, & de Rijke, 2012). Some of methods above are not general enough while others are not accurate enough. Because of different characters of different social media, it is not appropriate to apply the same prediction method to different social media.

Contrarily, other people demonstrated that prediction methods using social media cannot be applied to some problems. For instance, Skoric and others tried to predict election result in Singapore with twitter (Skoric, Poor, Achananuparp, Lim, & Jiang, 2012). However, they found that the predictive power is so weak in this problem because there are many other complex factors. Moreover, Gayo-Avello suggested that the power to predict outcomes based on twitter data is greatly exaggerated, especially for political elections (Gayo-Avello, 2011). Actually, prediction methods based on social media are not so good at predicting outcomes which are influenced by lots of factors and not so relative with social media.

Compared with previous work above, the information has been exploited more deeply and variously in this paper. The improvement of methods in the respects of amount and content of microblog can immensely optimize the prediction result. Different from previous researches, our method mix the count based information and the content based information. Furthermore, instead of linear prediction model, the machine learning models like neutral network and SVM which are more complex to predict the box offices have been chosen. The results of the experiments show that the outcomes can be predicted more precisely and reliably after applying our method.

The rest of this paper is arranged as follow. The problem we would like to solve will be defined in Section 2. Besides this, the framework of our method will be provided in this section. Section 3 describes the feature extraction methods and prediction model specifically. Section 4 presents our experiments to demonstrate the predictive power of the methods and discussion about them. At last, the result of the paper will be concluded and the future work will be presented in Section 5.

2. Problem definition and framework

2.1. Problem definition

As we said in Section 1, our goal is to demonstrate the predictive power of microblog. To achieve this, a more specific problem should be defined. Therefore, our focus turns to predicting the box offices of movies.

Assume x_i to be the information, which is extracted from microblog which is published during the days from t_{i-1} to t_i.Moreover, it is assumed that y_i is the box office of certain movie during days from t_i to t_{i+1}.Without loss of generality, it is assumed that t_i - $- t_{i-1} = 7$. That is to say, the information and box office is split into some periods whose lengths are 7 days. Besides, it is assumed that t_1 is the day when the movie was released. So the problem is to predict y_i using $x_1,x_2,\ldots,x_i$ as well as $y_1,y_2,\ldots,y_{i-1}$. In other word, the prediction of

$$y_i = g(x_1,\ldots,x_i,y_1,\ldots,y_{i-1}) \tag{1}$$

will be our major work. Here, it is confirmed that $i = 3$, so our problem is to predict

$$y_i = g(x_1,x_2,x_3,y_1,y_2) \tag{2}$$

In other word, the box office of the movie on the 3rd week will be predicted using the microblog information from 1 week before it was released to 2 weeks after it was released as well as the box offices of first 2 weeks.

This problem can be divided into 2 sections. One of them is to choose a proper prediction model. The other is to find features to accurately and entirely represent the information of microblog.

2.2. Framework

To solve the problem above, our prediction method can be summarized as Fig. 1.

The microblog website is crawled with the API of it at first to get the microblogs which are relative with the movies. After this, the information such as users, contents and comments is extracted from the microblogs. With the information, our content based method and count based method can be used to get the features and these features are mixed. The mixed features are applied to the machine learning models to build prediction models. The prediction models can be trained by real data. To predict box office of a new movie, we can extract the features as the input of the trained prediction model. The output of the model is the box office of the movie which we predict.

In following sections, the feature extraction methods are going to be described. After that, the prediction models will be provided to solve the problem.

3. Proposed method

3.1. Feature extraction method

3.1.1. Overview

As we can see in the last section, one of the most important parts of our work is to represent the information we can get from microblog as an input of our prediction model. Lots of previous work about prediction using microblog divides this task into two separate sections. One section is extracting the information based on the amount of microblogs, comments and retweets. These features, which named as count based features, are independent on the contents of the microblogs. The other section is making use of the information about the contents of the microblogs such as the amount of microblogs with positive sentiment and the ratio of microblogs with positive sentiment and negative sentiment. Previous work shows these features can be used to predict the box office with good performance. However, there may be interaction effect between the two kinds of features so it may be better to mix the two. In our work, not only count based features but also content based features are improved to make use of the microblog