A Quality-Aware Model for Sales Prediction Using Reviews

Xiaohui Yu[†,§], Yang Liu[†,§], Xiangji Huang[†], and Aijun An[†]
[†]York University
Toronto, ON, Canada, M3J 1P3
[§]School of Computer Science and Technology, Shandong University
Jinan, Shandong, China, 250101
xhyu@yorku.ca,yliu@sdu.edu.cn,jhuang@yorku.ca,aan@cse.yorku.ca

ABSTRACT

Writing and publishing reviews online has become an increasingly popular way for people to express opinions and sentiments. Analyzing the large volume of online reviews available can produce useful knowledge that are of interest to vendors and other parties. Prior studies in the literature have shown that online reviews have a significant correlation with the sales of products, and therefore mining the reviews could help predict the sales performance of relevant products. However, those studies fail to consider one important factor that may significantly affect the accuracy of the prediction, i.e., the quality of the reviews. In this paper, we propose a regression model that explicitly takes into account the quality factor, and discusses how this quality information can be predicted when it is not readily available. Experimental results on a movie review dataset confirm the effectiveness of the proposed model.

Categories and Subject Descriptors

H.4.0 [**Information Systems Applications**]: General

General Terms

Algorithm, Experiment

Keywords

Review mining, Review quality mining

1. INTRODUCTION

With the advent of Web 2.0 that centers around user participation, writing and publishing online reviews has become an increasingly popular way for people to share with other users their opinions and sentiments toward products and services. Those online reviews present a wealth of information on the products and services, and if properly utilized, can provide vendors highly valuable intelligence to facilitate the improvement of their business.

A growing number of recent studies have focused on the economic values of reviews, exploring the relationship between the sales performance of products and their reviews [1, 2]. One important aspect that has not been looked at in those prior studies is the effect of the reviews' quality on their predictive power. Quality-wise, not all reviews are created equal. Especially in an online setting where anybody can post anything, the quality of reviews can vary to a great extent. Reviews poorly written, reviews containing no subjective judgment, or even spam reviews, may actually negatively affect the accuracy of the prediction, if they are not properly taken care of.

In this paper, we propose a new model for predicting sales performance using online reviews, with the quality factor explicitly built in. Although there have been some previous studies [3] on the quality (or utility, or helpfulness) of online reviews, they focus on the automatic prediction of the review quality itself. To the best of our knowledge, our work represents the first step towards exploring the economic impact of review quality.

The rest of the paper is organized as follows. In Section 2, we start with presenting the general quality-aware framework for sales performance prediction based on reviews. Then we describe a model with some specific choices for quality and sentiment modeling. In Section 3, we present the experimental results on a movie review dataset, and we conclude this paper in Section 4.

2. THE QUALITY-AWARE MODEL

In order to develop an effective model for sales performance prediction, we must first identify the important factors involved. To this end, we examined an IMDB movie review dataset along with the corresponding daily box office revenues, and conducted several preliminary experiments to evaluate the correlation between the candidate factors and the sales performance. Our efforts reveal that for a particular product, the following factors are among the most indicative: (1) the sales performance of the immediately preceding periods, (2) the sentiments expressed in the reviews, and last but not least (3) the quality of the reviews.

Based on the above observations, we propose the following general model for predicting sales performance. Here we assume that the time unit used is day, but the model is general enough to be extended to finer or coarser time units. Let x_t be the sales figure of the product on day t ($t = 1, \ldots, N$ where $t = 1$ and $t = N$ correspond to the first and last day of the period of interest). We assume that the resulting time series $\{x_t\}_{t=1}^N$ has been properly treated to remove trend and seasonality etc. This is to ensure that the time series $\{x_t\}$ can be considered as approximately stationary so that autoregressive models can be applied. Let v_t be the number of reviews posted at day t. Also, let $\omega_{t,j,k}$ be the quantitative representation of the k-th sentiment factor in the j-th review at time t, which we assume can be obtained based on a model for sentiment analysis (a specific choice of which will be introduced in the sequel). Denote by $\mu_{t,j}$ the quality of the j-th review (either readily-available or predicted by some model) on day t. Then the prediction model can be formulated as follows:

$$x_t = \sum_{i=1}^{p} \phi_i x_{t-i} + \sum_{i=1}^{q} \frac{1}{v_{t-i}} \sum_{j=1}^{v_{t-i}} \rho_{i,k} \sum_{k=1}^{r} \mu_{t-i,j} \omega_{t-i,j,k} + \epsilon_t \quad (1)$$

where p, q, and r are user-defined parameters, ϵ_t is an error term (white noise with zero mean), and ϕ_i, $\mu_{t-i,j}$, and $\rho_{i,k}$ are parameters to be estimated from the training data. p and q specify how far the model "looks back" into the history, whereas r specifies how many sentiment factors we would like to consider. Note that in Equation (1), the sentiment factors are weighted by the quality of the reviews, which reflects the fact that reviews of different levels of quality have different degrees of influence in the prediction.

Modeling sentiments: Note that Equation (1) is a general framework, as it does not limit the methods used for sentiment modeling and quality modeling. Our particular choice for sentiment modeling is the S-PLSA model [2], which has been shown to be effective in sales performance prediction. In essence, it assumes that there are a number of hidden factors or aspects in the documents, and models using a probabilistic framework the relationship among those factors, the documents, and the words appearing in the documents. More specifically, for a given pair of word and document (w, d), S-PLSA models the joint probability $p(w, d)$ as a mixture of conditionally independent multinomial distributions: $p(w, d) = \sum_z p(z)p(w|z)p(d|z)$. What differentiates S-PLSA from conventional PLSA is that the words used in modeling are limited to "appraisal words"[4] that better reflect the sentiments. This model can be trained using the algorithm proposed in [2], and the result are plugged into Equation (1) by setting $\omega_{t-i,j,k} = p(z = k|d_{t-i,j})$ where d denotes the j-th review at time $t - i$.

Modeling quality: Quality indicators for reviews are sometimes readily available. For example, many websites provide summary information about a review in the form of "*x out of y people found the following review helpful*". In such cases, $h = \frac{x}{y}$ can be deemed as the true quality rating of a review, $\mu_{t-i,j}$ [3, 5]. When such information is absent, we can obtain it through the use of a trained prediction model. In particular, we use the *writing style* of a review to help predict its quality. Online reviews often demonstrate dramatically different qualities in terms of writing. Some reviews are highly readable and therefore tend to have better quality, whereas some reviews are either lengthy but with few sentences containing author's opinion or snappy but filled with insulting remarks. Therefore, the writing style provides important indication on the quality of a review, and has been shown to be among the most influential factors on the helpfulness of a review[3, 5].

Other factors are also considered in previous studies on the related problem of helpfulness prediction, such as reviewer expertise and timeliness of reviews [3]. But we note that they are either not able to be applied to general settings due to the lack of user information (in the case of review expertise), or not directly related to quality itself (in the case of timeliness of reviews). Therefore, we do not take them into consideration in developing our model.

Our modeling of the writing style is based on a previous study showing that shallow syntactically features like part-of-speech can be a good indicator of the utility of the review[5]. For our purpose, we consider the part-of-speech that can potentially contribute to the differentiation of writing style due to their implication of the subjectivity/objectivity of a review. The tags chosen include: *qualifiers* (e.g., quite, rather), *modal auxiliaries* (e.g., can, should), *nominal pronouns* (e.g., everybody, nothing), *comparative and superlative adverbs* (e.g., harder, faster), *comparative and superlative adjectives* (e.g., bigger, top), *proper nouns* (e.g., Caribbean, Snoopy), *interjections/exclamations* (e.g., ouch, well), *wh-determiners* (e.g., what, which), *wh-pronouns* (e.g., who, whose, which), and *wh-adverbs* (e.g., how, where, when).

We represent each review as a vector $\mathbf{y}$, each element of which represents the normalized frequency of a particular part-of-speech tag in the review, which can be obtained by parsing the review using LingPipe(www.lingpipe.com) part-of-speech tagger. The relationship between $\mathbf{y}$ and the quality of the review can be learned by ϵ-Support Vector Regression with radial basis function (RBF) kernels, which is capable of handling the non-linear relationship between the target values and the feature vectors. This model can be trained using reviews with known labels (quality ratings), and then used to predict the quality of any given review.

With the sentiment and quality factors already known (in the case of available quality ratings) or predicted, parameter estimation (for ϕ_i, $\mu_{t-i,j}$, and $\rho_{i,k}$) in Equation (1) can be done using least squares regression. The detailed algorithm is omitted due to space limit.

3. EXPERIMENTS

Extensive experiments were conducted on the IMDB dataset to evaluate the effectiveness of the proposed approach. The dataset was obtained from the IMDB Website. Specifically, we collected the reviews for 504 movies released in the United States during the period from January 6, 2006 to November 21, 2007, along with the daily gross box office revenues of those movies. We use 10-fold validation to evaluate the model. The quality model is first trained on the reviews with true quality ratings (by the users) in the training data and then used to predict the quality for reviews without ratings. We use the mean absolute percentage error (MAPE) to measure the prediction accuracy: $MAPE = \frac{1}{n}\sum_{i=1}^{N} \frac{|Pred_i - True_i|}{True_i}$, where N is the number of instances, and $Pred_i$ and $True_i$ are the predicted value and the true value of the box office revenues respectively.

We experimentally compare ARSQA with a baseline autoregressive (AR) model ($x_t = \sum_{i=1}^{p} \phi_i x_{t-i} + \epsilon_t$) and the original ARSA model [2]. The model that uses the quality ratings predicted with our proposed method is called $ARSQA_1$, and the one that uses the true quality ratings (as voted by users) $ARSQA_2$. We use $p = 7$ and $q = 1$, and vary r from 2 to 5. As shown in the table below, $ARSQA_1$ and $ARSQA_2$ significantly outperform the AR model and the original $ARSA$ model that does not consider review quality. In addition, $ARSQA_2$ is only slightly superior to $ARSQA_1$ which indirectly proves the effectiveness of our quality prediction model. Similar trends are observed when other values of the various parameters (i.e., p and q) are used.

	$r = 2$	$r = 3$	$r = 4$	$r = 5$
AR	0.623	0.623	0.623	0.623
$ARSA$	0.394	0.307	0.256	0.295
$ARSQA_1$	0.271	0.233	0.182	0.218
$ARSQA_2$	0.244	0.201	0.165	0.179

We also studied the effect of the parameter values (p, q, and r) on the prediction accuracy. Due to space limitation, detailed results are not presented here. Our main finding is that there are optimal choices for p, q, and r ($p = 7$, $q = 1$, $r = 4$ in our case).

4. CONCLUSIONS AND FUTURE WORK

In this paper, we have presented a quality-aware model for predicting sales performance. Experiments show that by considering the quality of reviews, we can greatly improve the prediction accuracy. Note that the proposed model is general enough so that different sentiment and quality models can be plugged in as needed. For future work, we plan to study how to further improve the prediction accuracy through the use of more sophisticated quality models.

Acknowledgement

This work is supported by NSERC Discovery Grants, an Early Researcher Award of Ontario, and an NSFC Grant (No. 60903108).

5. REFERENCES

[1] Anindya Ghose and Panagiotis G. Ipeirotis. Designing novel review ranking systems: predicting the usefulness and impact of reviews. In *ICEC*, pages 303–310, 2007.

[2] Yang Liu, Xiangji Huang, Aijun An, and Xiaohui Yu. ARSA: a sentiment-aware model for predicting sales performance using blogs. In *SIGIR*, pages 607–614, 2007.

[3] Yang Liu, Xiangji Huang, Aijun An, and Xiaohui Yu. Predicting the helpfulness of online reviews. In *ICDM*, pages 443–452, 2008.

[4] Casey Whitelaw, Navendu Garg, and Shlomo Argamon. Using appraisal groups for sentiment analysis. In *CIKM '05*, pages 625–631, 2005.

[5] Zhu Zhang and Balaji Varadarajan. Utility scoring of product reviews. In *CIKM*, pages 51–57, 2006.