Mining Online Reviews for Predicting Sales Performance: A Case Study in the Movie Domain

Xiaohui Yu, *Member*, *IEEE*, Yang Liu, *Member*, *IEEE*,
Jimmy Xiangji Huang, *Member*, *IEEE*, and Aijun An, *Member*, *IEEE*

Abstract—Posting reviews online has become an increasingly popular way for people to express opinions and sentiments toward the products bought or services received. Analyzing the large volume of online reviews available would produce useful actionable knowledge that could be of economic values to vendors and other interested parties. In this paper, we conduct a case study in the movie domain, and tackle the problem of mining reviews for predicting product sales performance. Our analysis shows that both the sentiments expressed in the reviews and the quality of the reviews have a significant impact on the future sales performance of products in question. For the sentiment factor, we propose Sentiment PLSA (S-PLSA), in which a review is considered as a document generated by a number of hidden sentiment factors, in order to capture the complex nature of sentiments. Training an S-PLSA model enables us to obtain a succinct summary of the sentiment information embedded in the reviews. Based on S-PLSFA, we propose ARSA, an Autoregressive Sentiment-Aware model for sales prediction. We then seek to further improve the accuracy of prediction by considering the quality factor, with a focus on predicting the quality of a review in the absence of user-supplied indicators, and present ARSQA, an Autoregressive Sentiment and Quality Aware model, to utilize sentiments and quality for predicting product sales performance. Extensive experiments conducted on a large movie data set confirm the effectiveness of the proposed approach.

Index Terms—Review mining, sentiment analysis, prediction.

1 Introduction

With the advent of Web 2.0 that centers around user participation, posting online reviews has become an increasingly popular way for people to share with other users their opinions and sentiments toward products and services. It has become a common practice for e-commerce websites to provide the venues and facilities for people to publish their reviews, with a prominent example being Amazon (www.amazon.com). Reviews are also prevalent in blog posts, social networking websites as well as dedicated review websites such as Epinions (www.epinions.com). Those online reviews present a wealth of information on the products and services, and if properly utilized, can provide vendors highly valuable network intelligence and social intelligence to facilitate the improvement of their business. As a result, review mining has recently received a great deal of attention.

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], [3], [4]. Since what the general public thinks of a product can no doubt influence how well it sells, understanding the opinions and sentiments expressed in the relevant reviews is of high importance, because collectively these reviews reflect the "wisdom of crowds" (what the general public think) and can be a very good indicator of the product's future sales performance. In this paper, we are concerned with generating actionable knowledge by developing models and algorithms that can utilize information mined from reviews. Such models and algorithms can be used to effectively predict the future sales of products, which can in turn guide the actions of the stakeholders involved.

Prior studies on the predictive power of reviews have used the volume of reviews or link structures to predict the trend of product sales [1], [5], failing to consider the effect of the sentiments present in the blogs. It has been reported [1], [5] that although there seems to exist strong correlation between the volume of reviews and sales spikes, using the volume or the link structures alone do not provide satisfactory prediction performance. Indeed, as we will illustrate with an example, the sentiments expressed in the reviews are more predictive than volumes. In addition, another important aspect that has been largely overlooked by 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

- *X. Yu is with the School of Computer Science and Technology, Shandong University, 1500 Shun Hua Lu, High-Tech Development Zone, Jinan 250101, China, and the School of Information Technology, York University, 4700 Keele Street, Toronto, ON M3J 1P3, Canada. E-mail: xhyu@yorku.ca.*
- *Y. Liu is with the School of Computer Science and Technology, Shandong University, 1500 Shun Hua Lu, High-Tech Development Zone, Jinan 250101, China. E-mail: yliu@sdu.edu.cn.*
- *J.X. Huang is with the School of Information Technology, York University, 4700 Keele Street, Toronto, ON M3J 1P3, Canada. E-mail: jhuang@yorku.ca.*
- *A. An is with the Department of Computer Science and Engineering, York University, 4700 Keele Street, Toronto, ON M3J 1P3, Canada. E-mail: aan@cse.yorku.ca.*

Manuscript received 21 Feb. 2010; revised 18 Aug. 2010; accepted 6 Sept. 2010; published online 23 Dec. 2010.
Recommended for acceptance by H. Wang.
For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number TKDE-2010-02-0104.
Digital Object Identifier no. 10.1109/TKDE.2010.269.

can post virtually anything, the quality of reviews can vary to a great extent. Examples of "bad" reviews include very short insulting comments with no substance like "*This book sucks*," or long and tedious reviews that are simply duplicates of the product descriptions. 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.

We believe that prediction of product sales is a highly domain-driven task, for which a deep understanding of various factors involved is essential. In this paper, using the movie domain as a case study, we investigate the various issues encountered in modeling reviews, producing sales predictions, and deriving actionable knowledge. To this end, we identify three factors that play important roles in predicting the box office revenues in the movie domain, namely, public sentiments, past sales performance, and review quality, and propose a framework for sales prediction with all those factors incorporated.

We start with modeling sentiments in reviews, which presents unique challenges that cannot be easily addressed by conventional text mining methods. Simply classifying reviews as positive or negative, as most current sentiment-mining approaches are designed for, does not provide a comprehensive understanding of the sentiments reflected in reviews. In order to model the multifaceted nature of sentiments, we view the sentiments embedded in reviews as an outcome of the joint contribution of a number of hidden factors, and propose a novel approach to sentiment mining based on Probabilistic Latent Semantic Analysis (PLSA), which we call Sentiment PLSA (S-PLSA). Different from the traditional PLSA [6], S-PLSA focuses on sentiments rather than topics. Therefore, instead of taking a vanilla "bag-of-words" approach and considering all the words (modulo stop words) present in the blogs, we focus primarily on the words that are sentiment related. To this end, we adopt in our study the appraisal words extracted from the lexicon constructed by Whitelaw et al. [7]. Despite the seemingly lower word coverage (compared to using "bag of words"), decent performance has been reported when using appraisal words in sentiment classification of movie reviews [7]. In S-PLSA, appraisal words are exploited to compose the feature vectors for reviews, which are then used to infer the hidden sentiment factors.

The second factor we consider is the past sale performance of the same product, or in the movie domain, past box office performance of the same movie. We capture this effect through the use of an Autoregressive (AR) model, which has been widely used in many time series analysis problems, especially in econometric contexts [8]. Combining this AR model with sentiment information mined from the reviews, we propose a new model for product sales prediction called the Autoregressive Sentiment Aware (ARSA) model. Extensive experiments show that the ARSA model provides superior predication performance compared to using the AR model alone, confirming our expectation that sentiments play an important role in predicting future sales performance.

Since online reviews are of varying quality and, thus, carry different predictive power, we should not treat them equally in producing the prediction. This motivates our study of the quality factor in sales prediction. We consider both cases in which quality indicators are readily available (e.g., in the form of user ratings), and cases in which they are not. Our focus is on the latter case, for which we develop a model that is able to automatically predict the quality of a review based on its syntactical characteristics. The quality factor is then incorporated into the ARSA model, resulting in an Autoregressive Sentiment and Quality Aware (ARSQA) model for sales prediction.

In summary, we make the following contributions:

- Using the movie domain as a case study, we approach the problem of predicting sales performance using online reviews as a domain-driven task, and identify the important the factors involved in generating prediction.
- We model sentiments in reviews as the joint outcome of some hidden factors, answering the call for a model that can handle the complex nature of sentiments. We propose the S-PLSA model, which through the use of appraisal groups, provides a probabilistic framework to analyze sentiments in reviews.
- We develop a model for predicting the quality of reviews in the absence of readily available quality indicators.
- We propose the ARSA and ARSQA models for product sales prediction, which reflects the effect of sentiments, and past sales performance (and in the case of ARSQA, the quality of reviews) on future sales performance. There effectiveness is confirmed by experiments.
- We discuss how actionable knowledge can be derived through utilizing the proposed models, explaining the practical impact of the proposed approach.

The rest of the paper is organized as follows. Section 2 provides a brief review of related work. In Section 3, we discuss the characteristics of online reviews which motivate the proposal of S-PLSA in Section 4. In Section 5, we propose ARSA, the sentiment-aware model for predicting future product sales. Section 6 considers the quality factor and presents the ARSQA model. Section 7 reports on the experimental results, and Section 8 concludes this paper.

2 Related Work

2.1 Domain-Driven Data Mining (D^3M)

In the past few years, domain-driven data mining has emerged as an important new paradigm for knowledge discovery [9], [10]. Motivated by the significant gap between the academic goals of many current KDD methods and the real-life business goals, D^3 advocates the shift from data-centered hidden pattern mining to domain-driven Actionable Knowledge Discovery (AKD). The work presented in this paper can be considered as an effort along this direction in that 1) we aim to deliver actionable knowledge by making predictions of sales performance, and 2) in developing the prediction model, we try to integrate multiple types of intelligence, including human intelligence, domain intelligence, and network intelligence (Web intelligence).

2.2 Review Mining

With the rapid growth of online reviews, review mining has attracted a great deal of attention. Early work in this area was primarily focused on determining the semantic orientation

of reviews. Among them, some of the studies attempt to learn a positive/negative classifier at the document level. Pang et al. [11] employ three machine learning approaches (Naive Bayes, Maximum Entropy, and Support Vector Machine) to label the polarity of IMDB movie reviews. In follow-up work, they propose to first extract the subjective portion of text with a graph min-cut algorithm, and then feed them into the sentiment classifier [12]. Instead of applying the straightforward frequency-based bag-of-words feature selection methods, Whitelaw et al. [7] defined the concept of "adjectival appraisal groups" headed by an appraising adjective and optionally modified by words like "not" or "very." Each appraisal group was further assigned four type of features: attitude, orientation, graduation, and polarity. They report good classification accuracy using the appraisal groups. They also show that the classification accuracy can be further boosted when they are combined with standard "bag-of-words" features. We use the same words and phrases from the appraisal groups to compute the reviews' feature vectors, as we also believe that such adjective appraisal words play a vital role in sentiment mining and need to be distinguished from other words. However, as will become evident in Section 4, our way of using these appraisal groups is different from that in [7].

There are also studies that work at a finer level and use words as the classification subject. They classify words into two groups, "good" and "bad," and then use certain functions to estimate the overall "goodness" or "badness" score for the documents. Kamps and Marx [13] propose to evaluate the semantic distance from a word to good/bad with WordNet. Turney [14] measures the strength of sentiment by the difference of the Mutual Information (PMI) between the given phrase and "excellent" and the PMI between the given phrase and "poor."

Extending previous work on explicit two-class classification, Pang and Lee [15], and Zhang and Varadarajan [16] attempt to determine the author's opinion with different rating scales (i.e., the number of stars). Liu et al. [17] build a framework to compare consumer opinions of competing products using multiple feature dimensions. After deducting supervised rules from product reviews, the strength and weakness of the product are visualized with an "Opinion Observer."

Our method departs from conventional sentiment classification in that we assume that sentiment consists of multiple hidden aspects, and use a probability model to quantitatively measure the relationship between sentiment aspects and reviews as well as sentiment aspects and words.

2.3 Economic Impact of Online Reviews

Whereas marketing plays an important role in the newly released products, customer word of mouth can be a crucial factor that determines the success in the long run, and such effect is largely magnified thanks to the rapid growth of Internet. Therefore, online product reviews can be very valuable to the vendors in that they can be used to monitor consumer opinions toward their products in real time, and adjust their manufacturing, servicing, and marketing strategies accordingly.

Academics have also recognized the impact of online reviews on business intelligence, and have produced some important results in this area. Among them, some studies attempt to answer the question of whether the polarity and the volume of reviews available online have a measurable and significant effect on actual customer purchasing [18], [19], [20], [5], [1]. To this end, most studies use some form of hedonic regression [21] to analyze the significance of different features to certain function, e.g., measuring the utility to the the consumer. Various economic functions have been utilized in examining revenue growth, stock trading volume change as well as the bidding price variation on commercial websites, such as Amazon and eBay.

In most of the studies cited above, the sentiments are captured by explicit rating indication such as the number of stars; few studies have attempted to exploit text mining strategies for sentiment classification. To fill in this gap, Ghose and Ipeirotis [2] argue that review texts contain richer information that cannot be easily captured using simple numerical ratings. In their study, they assign a "dollar value" to a collection of adjective-noun pairs, as well as adverb-verb pairs, and investigate how they affect the bidding prices of various products at Amazon.

Our work is similar to [2] in the sense that we also exploit the textual information to capture the underlying sentiments in the reviews. However, their approach mainly focuses on quantifying the extent of which the textual content, especially the subjectivity of each review, affects product sales on a market such as Amazon, while our method aims to build a more fundamental framework for predicting sales performance using multiple factors.

Foutz and Jank [22], [23] also exploit the wisdom of crowds to predict the box office performance of movies. The work presented in this paper differs from theirs in three ways. First, we use online reviews as a source of network intelligence to understand the sentiments of the public, whereas their approach uses virtual stock markets (prediction markets) as an aggregated measure of public sentiments and expectations. Second, we use a parametric regression model to capture the temporal relationships, whereas their approach uses nonparametric functional shape analysis to extract the important features in the shapes across various trading histories and then uses these key features to produce forecasts. Third, the prediction of our model is ongoing as time progresses and more reviews are posted, whereas their approach is limited to forecasting the box office performance in the release week.

2.4 Assessing the Review Helpfulness

Compared to sentiment mining, identifying the quality of online reviews has received relatively less attention. A few recent studies along this direction attempt to detect the spam or low-quality posts that exist in online reviews. Jindal and Liu [24] present a categorization of review spams, and propose some novel strategies to detect different types of spams. Liu et al. [25] propose a classification-based approach to discriminate the low-quality reviews from others, in the hope that such a filtering strategy can be incorporated to enhance the task of opinion summarization. Elkan [26] develops a complete framework that consists of six different components, for retrieving and filtering online documents with uneven quality. Similar to [25], a binary classifier is constructed in order to discriminate the documents with the *Classifying Component*. Our work can be considered complimentary to those studies in that the spam filtering model can be used as a preprocessing step in our approach.

2.5 Ranking and Recommender Systems

Recommender systems have emerged as an important solution to the information overload problem where people find it more and more difficult to identify the useful information effectively. Studies in this area can generally be divided into three directions: content-based filtering, Collaborative Filtering (CF), and hybrid systems. Content-based recommenders rely on rich content descriptions of behavioral user data to infer their interests, which raises significant engineering challenges as the required domain knowledge may not be readily available or easy to maintain. As an alternative, collaborative filtering takes the rating data as input, and applies data mining or machine learning algorithms to discover usage patterns that constitute the user models. When a new user comes to the site, his/her activity will be matched against those patterns to find like-minded users and items that could be of interest to the users are recommended.

Various CF algorithms ranging from typical nearest neighbor methods [27] to more complex probabilistic-based methods [28], [29] have been designed to identify users of similar interests. A few variations and hybrid methods that combine both content information and collaborative filtering have also been proposed to solve the cold-start problem [30], [31].

Similar to recommender systems, our work also accounts for textual contents and peer votes in those reviews to effectively construct and evaluate the prediction model; however, one of the objectives of this study is to investigate the quality of movie reviews, which is different from the above work.

2.6 Time Series Analysis

Time series analysis is a well-established field with a large body of the literature [8]. Its purpose is to reveal the underlying forces and structure that produced the observed data, based on which a model can be fit, and tasks such as forecasting and monitoring can be carried out. It has been widely applied in a large variety of areas, such as economic forecasting, sales forecasting, stock market analysis, etc. Three broad classes of models that are most widely used are autoregressive models, the Moving Average (MA) models, and the integrated (I) models. In real applications, they are often combined to produce models like Autoregressive Moving Average (ARMA) and Autoregressive Integrated Moving Average (ARIMA) models. When the time series data are vector valued, these models can be extended to their vectorized versions, which constitutes multivariate time series analysis.

Our method utilizes times series analysis in that we use the AR model to capture the temporal relationship in the box office data. The main difference between our work and conventional time series analysis is that while in conventional time series analysis the data series are generally directly observed, in our work some data (such as the sentiments embedded in the reviews, and the quality of reviews) are not directly observable and are inferred from the underlying data (reviews). Our focus is to identify appropriate methods to "recover" those latent data so that accurate prediction can be made.

This paper is built upon our previous work on the predictive power of sentiments [3]. In particular, we have extended the ARSA model proposed in [3], and incorporated the important factor of review quality into the model. We have also considered how to predict the quality of reviews using content features, and conducted more experiments to evaluate the effectiveness of the proposed models.

3 Characteristics of Online Reviews

As a domain-driven task, it is essential to understand the characteristics of online reviews and their predictive power. To this end, we investigate the pattern of reviews and its relationship to sales data by examining a real example from the movie sector. In particular, we are interested in reviews posted in the blogsphere, as they usually serve as a good sample of reviews published in various forms (e.g., bulletin boards, review websites, etc.).

3.1 Number of Blog Mentions

Let us look at the following two movies: *The Da Vinci Code* and *Over the Hedge*, which were both released on May 19, 2006. We use the name of each movie as a query to a publicly available blog search engine[1] and limit the scope of search to a popular blog website (`blogspot.com`). In addition, as each blog is always associated with a fixed time stamp, we augment the query input with a date for which we would like to collect the data. For each movie, by issuing a separate query for each single day in the period starting from one week before the movie release till three weeks after the release, we chronologically collect a set of blogs appearing in a span of one month. We use the number of returned results for a particular date as a rough estimate of the number of blog mentions published on that day.

In Fig. 1a, we compare the changes in the number of blog mentions of the two movies. Apparently, there exists a spike in the number of blog mentions for the movie *The Da Vinci Code*, which indicates that a large volume of discussions on that movie appeared around its release date. In addition, the number of blog mentions are significantly larger than those for *Over the Hedge* throughout the whole month.

3.2 Box Office Data and User Rating

Besides the blogs, we also collect for each movie one month's box office data (daily gross revenue) from the IMDB website.[2] The changes in daily gross revenues are depicted in Fig. 1b. Apparently, the daily gross of *The Da Vinci Code* is much greater than *Over the Hedge* on the release date. However, the difference in the gross revenues between the two movies becomes less and less as time goes by, with *Over the Hedge* sometimes even scoring higher toward the end of the one-month period. To shed some light on this phenomenon, we collect the average user ratings of the two movies from the IMDB website. *The Da Vinci Code* and *Over the Hedge* got the rating of 6.5 and 7.1, respectively.

3.3 Discussion

It is interesting to observe from Fig. 1 that although *The Da Vinci Code* has a much higher number of blog mentions than *Over the Hedge*, its box office revenue are on par with that of *Over the Hedge* save the opening week. This implies that the number of blog mentions (and correspondingly, the number of reviews) may not be an accurate indicator of a product's

1. http://blogsearch.google.ca/blogsearch.
2. http://www.imdb.com/.

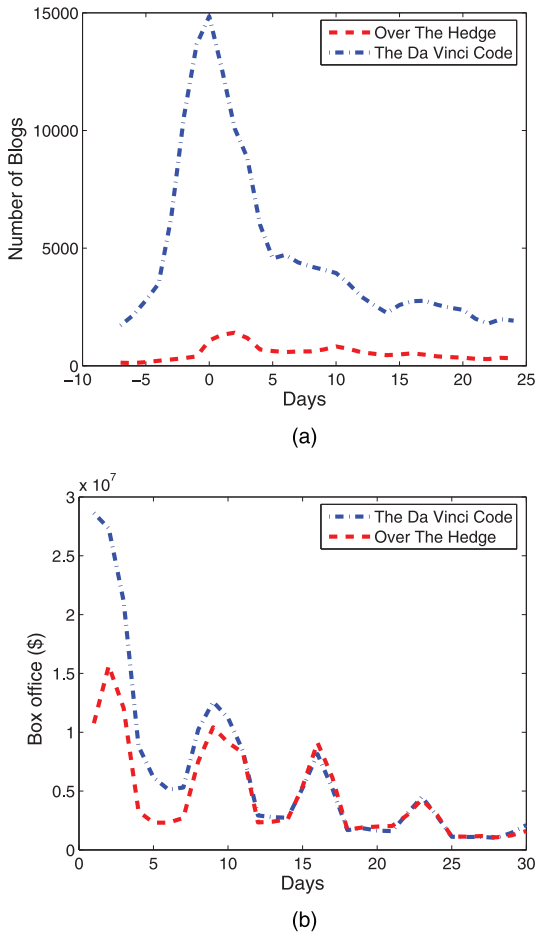


Fig. 1. An example of the relationship between the number of blog mentions and the box office revenues. (a) The number of blog mentions over time. (b) The box office revenues over time.

sales performance. A product can attract a lot of attention (thus a large number of blog mentions) due to various reasons, such as aggressive marketing, unique features, or being controversial. This may boost the product's performance for a short period of time. But as time goes by, it is the quality of the product and how people feel about it that dominates. This can partly explain why in the opening week, *The Da Vinci Code* had a large number of blog mentions and staged an outstanding box office performance, but in the remaining weeks, its box office performance fell to the same level as that for *Over the Hedge*. On the other hand, people's opinions (as reflected by the user ratings) seem to be a good indicator of how the box office performance evolves. Observe that, in our example, the average user rating for *Over the Hedge* is higher than that for *The Da Vinci Code*; at the same time, it enjoys a slower rate of decline in box office revenues than the latter. This suggests that sentiments in the blogs could be a very good indicator of a product's future sales performance.

4 S-PLSA: A Probabilistic Approach to Sentiment Mining

In this section, we propose a probabilistic approach to analyzing sentiments in reviews, which will serve as the basis for predicting sales performance.

4.1 Feature Selection

We first consider the problem of feature selection, i.e., how to represent a given review as an input to the mining algorithms. The traditional way to do this is to compute the (relative) frequencies of various words in a given document (review) and use the resulting multidimensional feature vector as the representation of the document. Here, we follow the same methodology, but instead of using the frequencies of all the words appearing the reviews, we choose to focus on the set containing 2,030 appraisal words extracted from the lexicon constructed by Whitelaw et al. [7], and construct feature vectors based on their frequencies in reviews. The rationale behind this is that for sentiment analysis, sentiment-oriented words, such as "good" or "bad," are more indicative than other words [7]. It is noted in [7] that "*... the appraisal taxonomies used in this work are general purpose, and were not developed specifically for sentiment analysis or movie review classification.*" Therefore, we consider the appraisal groups developed by Whitelaw et al. a good fit to our problem, and the same lexicon can be applied to other domains as well.

4.2 Sentiment PLSA

Mining opinions and sentiments present unique challenges that cannot be handled easily by traditional text mining algorithms. This is mainly because the opinions and sentiments, which are usually written in natural languages, are often expressed in subtle and complex ways. Moreover, sentiments are often multifaceted, and can differ from one another in a variety of ways, including polarity, orientation, graduation, and so on. Therefore, it would be too simplistic to just classify the sentiments expressed in a review as either positive or negative. For the purpose of sales prediction, a model that can extract the sentiments in a more accurate way is needed.

To this end, we propose a probabilistic model called Sentiment Probabilistic Latent Semantic Analysis (S-PLSA), in which a review can be considered as being generated under the influence of a number of hidden sentiment factors. The use of hidden factors allows us to accommodate the intricate nature of sentiments, with each hidden factor focusing on one specific aspect of the sentiments. The use of a probabilistic generative model, on the other hand, enables us to deal with sentiment analysis in a principled way.

In its traditional form, PLSA [6] assumes that there are a set of hidden semantic factors or *aspects* in the documents, and models the relationship among these factors, documents, and words under a probabilistic framework. With its high flexibility and solid statistical foundations, PLSA has been widely used in many areas, including information retrieval, Web usage mining, and collaborative filtering. Nonetheless, to the best of our knowledge, we are the first to model sentiments and opinions as a mixture of hidden factors and use PLSA for sentiment mining.

We now formally present S-PLSA. Suppose we are given a set of reviews $\mathcal{B} = \{b_1, \ldots, b_N\}$, and a set of words (appraisal words) from a vocabulary $\mathcal{W} = \{w_1, \ldots, w_M\}$. The review data can be described as a $N \times M$ matrix $D = (c(b_i, w_j))_{ij}$, where $c(b_i, w_j)$ is the number of times w_j appears in review b_i. Each row in D is then a frequency vector that corresponds to a review.

S-PLSA is a latent variable model for co-occurrence data ((b, w) pairs) that associates with each (w, b) observation an unobserved hidden variable from the set of hidden sentiment factors, $\mathcal{Z} = \{z_1, \ldots, z_K\}$. Just like in PLSA where hidden factors correspond to the "topics" of the documents, in S-PLSA those factors may correspond to the sentiments embodied in the reviews (e.g., joy, surprise, disgust, etc.). Such sentiments are not directly observable in the reviews; rather, they are expressed through the use of combinations of appraisal words. Hence, we use hidden factors to model sentiments.

For a word-review pair (w, b), S-PLSA models the co-occurrence probability as a mixture of conditionally independent multinomial distributions

$$\Pr(b, w) = \sum_{z \in \mathcal{Z}} \Pr(z)\Pr(w|z)\Pr(b|z),$$

where we consider both review b and word d to be generated from the latent factor z in similar ways, using the conditional probabilities $\Pr(b|z)$ and $\Pr(w|z)$, respectively. The assumption made here is that b and w are independent given the choice of the latent factor.

To explain the observed (b, w) pairs, we need to estimate the model parameters $\Pr(z)$, $\Pr(b|z)$, and $\Pr(w|z)$. To this end, we seek to maximize the following likelihood function:

$$L(\mathcal{B}, \mathcal{W}) = \sum_{b \in \mathcal{B}} \sum_{w \in \mathcal{W}} c(b, w) \log \Pr(b, w),$$

where $c(b, w)$ represents the number of occurrences of a pair (b, w) in the data.

A widely used method to perform maximum likelihood parameter estimation for models involving latent variables (such as our S-PLSA model) is the Expectation-Maximization (EM) algorithm [32], which involves an iterative process with two alternating steps.

1. An Expectation step (E-step), where posterior probabilities for the latent variables (in our case, the variable z) are computed, based on the current estimates of the parameters.
2. A Maximization step (M-step), where estimates for the parameters are updated to maximize the complete data likelihood.

In our model, with the parameters $\Pr(z)$, $\Pr(w|z)$, and $\Pr(b|z)$ suitably initialized, we can show that the algorithm requires alternating between the following two steps:

- In E-step, we compute

$$\Pr(z|b, w) = \frac{\Pr(z)\Pr(b|z)\Pr(w|z)}{\sum_{z' \in \mathcal{Z}} \Pr(z')\Pr(b|z')\Pr(w|z')}.$$

- In M-step, we update the model parameters with

$$\Pr(w|z) = \frac{\sum_{b \in \mathcal{B}} c(b, w)\Pr(z|b, w)}{\sum_{b \in \mathcal{B}} \sum_{w' \in \mathcal{W}} c(b, w')\Pr(z|b, w')},$$

$$\Pr(b|z) = \frac{\sum_{w \in \mathcal{W}} c(b, w)\Pr(z|b, w)}{\sum_{b' \in \mathcal{B}} \sum_{w \in \mathcal{W}} c(b', w)\Pr(z|b', w)},$$

and

$$\Pr(z) = \frac{\sum_{b \in \mathcal{B}} \sum_{w \in \mathcal{W}} c(b, w)\Pr(z|b, w)}{\sum_{b \in \mathcal{B}} \sum_{w \in \mathcal{W}} c(b, w)}.$$

It can be shown that each iteration above monotonically increases the complete data likelihood, and the algorithm converges when a local optimal solution is achieved.

Once the parameter estimation for the model is completed, we can compute the posterior probability $\Pr(z|b)$ using the Bayes rule

$$\Pr(z|b) = \frac{\Pr(b|z)\Pr(z)}{\sum_{z \in \mathcal{Z}} \Pr(b|z)\Pr(z)}.$$

Intuitively, $\Pr(z|b)$ represents how much a hidden sentiment factor $z(\in \mathcal{Z})$ "contributes" to the review b. Therefore, the set of probabilities $\{\Pr(z|b)|z \in \mathcal{Z}\}$ can be considered as a succinct summarization of b in terms of sentiments. As will be shown in the Section 5, this summarization can then be used in the predication of future product sales.

5 ARSA: A Sentiment-Aware Model

We now present a model to provide product sales predications based on the sentiment information captured from reviews. Due to the complex and dynamic nature of sentiment patterns expressed through online chatters, integrating such information is quite challenging.

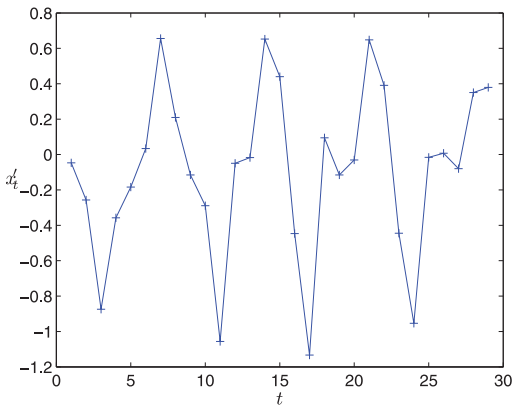
We focus on the case of predicting box office revenues to illustrate our methodologies. Our model aims to capture two different factors that can affect the box office revenue of the current day. One factor is the box office revenue of the preceding days. Naturally, the box office revenue of the current day is strongly correlated to those of the preceding days, and how a movie performs in previous days is a very good indicator of how it will perform in the days to come. The second factor we consider is the people's sentiments about the movie. The example in Section 3 shows that a movie's box office is closely related to what people think about the movie. Therefore, we would like to incorporate the sentiments mined from the reviews into the prediction model.

5.1 The Autoregressive Model

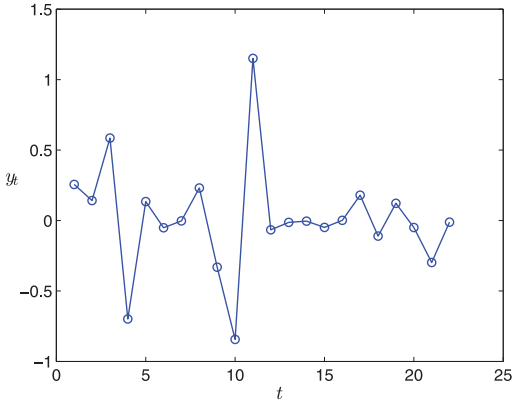
We start with a model that captures only the first factor described above and discuss how to incorporate the second factor into the model in the Section 5.2.

The temporal relationship between the box office revenues of the preceding days and the current day can be well modeled by an autoregressive process. Let us denote the box office revenue of the movie of interest at day t by x_t ($t = 1, 2, \ldots, N$, where $t = 1$ corresponds to the release date and $t = N$ corresponds to the last date we are interested in), and we use $\{x_t\}(t = 1, \ldots, N)$ to denote the time series $x_1, x_2, \ldots, x_N$. Our goal is to obtain an AR process that can model the time series $\{x_t\}$. A basic (but not quite appropriate, as discussed below) AR process of order p is as follows:

$$X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + \epsilon_t,$$



(a)



(b)

Fig. 2. An example of preprocessing the box office data (for the movie *The Da Vinci Code*). (a) Result after log transformation and differencing (x'_t). (b) Result after removing seasonality (y_t).

where $\phi_1, \phi_2, \ldots, \phi_p$ are the parameters of the model, and ϵ_t is an error term (white noise with zero mean).

Once this model is learned from training data, at day t, the box office revenue x_t can be predicted by $x_{t-1}, x_{t-2}, \ldots, x_{t-p}$. It is important to note, however, that AR models are only appropriate for time series that are stationary. Apparently, the time series $\{x_t\}$ are not, because there normally exist clear trends and "seasonality" in the series. For instance, in the example in Fig. 1, there is a seemingly negative exponential downward trend for the box office revenues as the time moves further from the release date. "Seasonality" is also present, as within each week, the box office revenues always peak at the weekend and are generally lower during weekdays. Therefore, in order to properly model the time series $\{x_t\}$, some preprocessing steps are required.

The first step is to remove the trend. This is achieved by first transforming the time series $\{x_t\}$ into the logarithmic domain, and then differencing the resulting time series $\{x_t\}$. The new time series obtained is, thus,

$$x'_t = \Delta \log x_t = \log x_t - \log x_{t-1}.$$

We then proceed to remove the seasonality [8]. To this end, we apply the lag operator on $\{x'_t\}$ and obtain a new time series $\{y_t\}$ as follows:

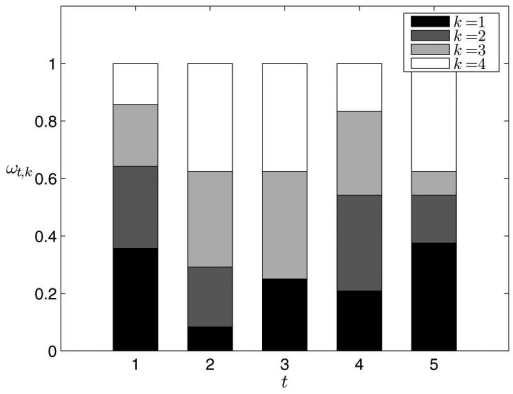


Fig. 3. The value of $\omega_{t,k}$ for different days (for the movie *The Da Vinci Code*).

$$y_t = x'_t - L^7 x' t = x'_t - x'_{t-7}.$$

By computing the difference between the box office revenue of a particular date and that of seven days ago, we effectively removed the seasonality factor due to different days of a week. We use the box office data for the movie *The Da Vinci Code* as an example and illustrate the results of the preprocessing steps in Fig. 2. As evident from Figs. 2a and 2b, the preprocessing steps make the box office data closer to stationary so that it is more amenable to the subsequent autoregressive modeling.

After the preprocessing step, a new AR model can be formed on the resulting time series $\{y_t\}$

$$y_t = \sum_{i=1}^{p} \phi_i y_{t-i} + \epsilon_t. \quad (1)$$

It is worth noting that although the AR model developed here is specific for movies, the same methodologies can be applied in other contexts. For example, trends and seasonality are present in the sales performance of many different products (such as electronics and music CDs). Therefore, the preprocessing steps described above to remove them can be adapted and used in the predicting the sales performance.

5.2 Incorporating Sentiments

As discussed earlier, the box office revenues might be greatly influenced by people's opinions in the same time period. We modify the model in (1) to take this factor into account. Let v_t be the number of reviews for a given movie posted on day t, and $\psi_{t,j,k}$ be the probability of the kth sentiment factor conditional on the jth review posted on day t, i.e., $\psi_{t,j,k} = p(z = k|t, j)$. Then, the average probability of factor $z = k$ at time t is defined as

$$\omega_{t,k} = \frac{1}{v_t} \sum_{j=1}^{v_t} \psi_{t,j,k} = \frac{1}{v_t} \sum_{j=1}^{v_t} p(z = k|t, j).$$

Intuitively, $\omega_{t,k}$ represents the average fraction of the sentiment "mass" that can be attributed to the hidden sentiment factor k. As an example, Fig. 3 shows the values of $\omega_{t,k}$ for different days $t(t = 1, \ldots, 5; k = 1, \ldots, 4)$ for the movie *The Da Vinci Code* using parameters obtained from the ARSA model to be presented in the sequel. The number of factors is set to 4. As can be observed from Fig. 3, the distribution of factors varies from one day to another,

reflecting the different charateristics of the reviews in terms of sentiments on different days.

Our new model, which we call the ARSA model, can be formulated as follows:

$$y_t = \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{i=1}^{q} \sum_{k=1}^{K} \rho_{i,k} \omega_{t-i,k} + \epsilon_t, \quad (2)$$

where p, q, and K are user-chosen parameters, while ϕ_i and $\rho_{i,k}$ are parameters whose values are to be estimated using the training data. Parameter q specifies the sentiment information from how many preceding days is taken into account, and K indicates the number of hidden sentiment factors used by S-PLSA to represent the sentiment information.

In summary, the ARSA model mainly comprises two components. The first component, which corresponds to the first term in the right hand side of (2), reflects the influence of past box office revenues. The second component, which corresponds to the second term, represents the effect of the sentiments as reflected from the reviews.

5.3 Training the ARSA Model

Training the ARSA model involves learning the set of parameters $\phi_i (i = 1, \ldots, p)$, and $\rho_{i,k} (i = 1, \ldots, q; k = 1, \ldots, K)$, from the training data that consist of the true box office revenues, and $\omega_{t,k}$ obtained from the review data. As we will show below, the model can, after choosing p and q, be fitted by least squares regression to estimate the parameter values.

For a particular movie $m (m = 1, 2, \ldots, M)$, where M is the total number of movies in the training data, and a given date t, let us add the subscript m to y_t and $\omega_{t-i,k}$ in (2) to be more precise. Let $\alpha_{m,t} = (y_{m,t-1}, \ldots, y_{m,t-p}, \omega_{m,t-1,1}, \ldots, \omega_{m,t-q,k})^T$. Then, (2) can be rewritten as

$$\alpha_{m,t}^T \theta = y_{m,t}.$$

Let A be a matrix composed of all $\alpha_{m,t}$ vectors corresponding to each movie and, for each movie and each t, i.e., $A = (\alpha_{1,1}, \alpha_{1,2}, \ldots)^T$. Similarly, let $\mathbf{c}$ denote the vector consisting of all possible $y_{m,t}$, i.e., $\mathbf{c} = (y_{1,1}, y_{1,2}, \ldots)$. Then, based on the training data, we seek to find a solution $\hat{\theta}$ for the "equation"

$$A\theta \approx \mathbf{c}.$$

More precisely, we seek to minimize the euclidean norm squared of residual $A\theta - \mathbf{c}$. This is exactly a least squares regression problem, and be solved using standard techniques in mathematics.

Once the model is trained, (2) can be used to predict the box office revenue of day t based on the box office revenues of the preceding days (which have already observed before day t), and the sentiments mined from the reviews.

6 The Quality Factor in Prediction

As discussed in Section 1, reviews vary a great deal in quality; thus, it is desirable to differentiate the reviews in terms of quality in order to achieve greater prediction accuracy. In this section, we analyze how to predict the quality of review in the absence of explicit quality indicators, and propose an extension of ARSA, which we call the ARSQA model.

6.1 Modeling Quality

Quality indicators for reviews are often 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 useful/helpful*." In such cases, we simply use $h = \frac{x}{y}$ as the approximation to the "true" quality factor $\mu_{t-i,j}$. Apparently, $h \in [0, 1]$. This measure has also been widely used in previous studies on helpfulness prediction [16], [33]. For robustness, we only consider using this h value for reviews that have received at least 10 votes.

In cases where user-supplied quality information is absent or too little to be considered useful (e.g., reviews that have received less than 10 votes), we propose to obtain this information through the use of a trained prediction model. In particular, we use the *writing style* of a review (which has been shown to be among the most influential factors on the helpfulness of a review [16], [33]) to help predict its quality. Due to the large variation of the reviewers' background and language skills, the online reviews are of dramatically different qualities. 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. A proper representation of writing style must be identified and incorporated into the prediction model.

Other factors are also considered in previous studies on the related problem of helpfulness prediction, such as reviewer expertise and timeliness of reviews [33]. 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 for quality prediction.

Our modeling of the writing style is based on a previous study which shows that shallow syntactical features like part of speech can be a good indicator of the utility of the review [16]. 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:

1. Qualifiers: one that modifies, reduce, tempers, or restrains (e.g., quite, rather, enough).
2. Modal auxiliaries: a type of verbs used to indicate modality. They give additional information about the function of the main verb that follows it (e.g., can, should, will).
3. Nominal pronouns (e.g., everybody, nothing).
4. Comparative and superlative adverbs: indicators of comparison (e.g., harder, faster, most prominent).
5. Comparative and superlative adjectives: indicators of comparisons, again (e.g., bigger, chief, top, largest).
6. Proper nouns: reference to a specific item, which will begin with a capital letter no matter where it occurs in a sentence (e.g., Caribbean, Snoopy).
7. Interjections/exclamations: strong signs of opinion (e.g., ouch, well).

8. wh-determiners (e.g., what, which), *wh-pronouns* (who, whose, which), and *wh-adverbs*: wh-words that signify either questions or other interesting linguistic constructs such as relative clauses. (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[3] part-of-speech tagger. Prior work has revealed that the relationship between the syntactical features and the quality of review is highly nonlinear [16]. Therefore, we use ϵ-Support Vector Regression with Radial Basis Function (RBF) kernels to learn the relationship between $\mathbf{y}$ and the quality of the review. This model can be trained using reviews with known labels (quality ratings h), and then used to predict the quality of any given review.

6.2 The ARSQA Model

Recall that $\{y_t\}$ denotes the time series representing the sales figures after proper treatment as described in the preceding section. Let v_t be the number of reviews posted at day t. Also, recall that $\psi_{t,j,k}$ is the inferred probability of the kth sentiment factor in the jth review at time t, which we assume can be obtained based on S-PLSA. Denote by $\mu_{t,j}$ the quality of the jth review (either readily available or predicted by some model) on day t. Then, the prediction model can be formulated as follows:

$$y_t = \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{i=1}^{q} \frac{1}{v_{t-i}} \sum_{j=1}^{v_{t-i}} \mu_{t-i,j} \sum_{k=1}^{K} \rho_{i,k} \psi_{t-i,j,k} + \epsilon_t, \quad (3)$$

where p, q, and K 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 K specifies how many sentiment factors we would like to consider. What differentiates ARSQA from ARSA is that in (3), 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 on the prediction.

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 (3) can be done using least squares regression in a fashion similar to that for ARSA.

7 Empirical Study

In this section, we report the results obtained from a set of experiments conducted on a movie data set in order to validate the effectiveness of the proposed model, and compare it against alternative methods.

7.1 Experiment Settings

The movie data we used in the experiments consists of three components. The first component is a set of blog documents on movies of interest collected from the Web, the second component contains the corresponding daily box office revenue data for these movies, and the third component consists of movie reviews and their helpfulness scores that are obtained from the IMDB websites.

Blog entries were collected for movies released in the United States during the period from May 1, 2006 to August 8, 2006. For each movie, using the movie name and a date as keywords, we composed and submitted queries to Google's blog search engine, and retrieved the blogs entries that were listed in the query results. For a particular movie, we only collected blog entries that had a timestamp ranging from one week before the release to four weeks after, as we assume that most of the reviews might be published close the release date. Through limiting the time span for which we collect the data, we are able to focus on the most interesting period of time around a movie's release, during which the blog discussions are generally the most intense. As a result, the amount of blog entries collected for each movie ranges from 663 (for *Waist Deep*) to 2,069 (for *Little Man*). In total, 45,046 blog entries that comment on 30 different movies were collected. We then extracted the title, permlink, free text contents, and time stamp from each blog entry, and indexed them using Apache Lucene.[4]

We manually collected the gross box office revenue data for the 30 movies from the IMDB website.[5] For each movie, we collected its daily gross revenues in the United States starting from the release date till four weeks after the release.

In each run of the experiment, the following procedure was followed:

1. We randomly choose half of the movies for training, and the other half for testing; the blog entries and box office revenue data are correspondingly partitioned into training and testing data sets.
2. Using the training blog entries, we train an S-PLSA model. For each blog entry b, the sentiments toward a movie are summarized using a vector of the posterior probabilities of the hidden sentiment factors, $\Pr(z|b)$.
3. We feed the probability vectors obtained in Step 2, along with the box revenues of the preceding days, into the ARSA model, and obtain estimates of the parameters.
4. We evaluate the prediction performance of the ARSA model by experimenting it with the testing data set.

In addition, to evaluate the effectiveness of our quality-aware model, we collected movie reviews that were published on the IMDB website. Specifically, we also selected the reviews for movies released during May 1, 2006 to August 8, 2006. We intentionally selected the time that is not very close to the present time in the hope that the voting of helpfulness has stabilized, as less and less reviews are expected to appear as time increases across the whole time span. To ensure the robustness of the predictive model, we only consider the reviews that have received at least 10 votes. Also, for the purpose of training and testing, only the reviews with a usefulness score available are used. The number of such movie reviews is 18,652.

3. www.lingpipe.com.

4. http://lucene.apache.org.
5. http://www.imdb.com.

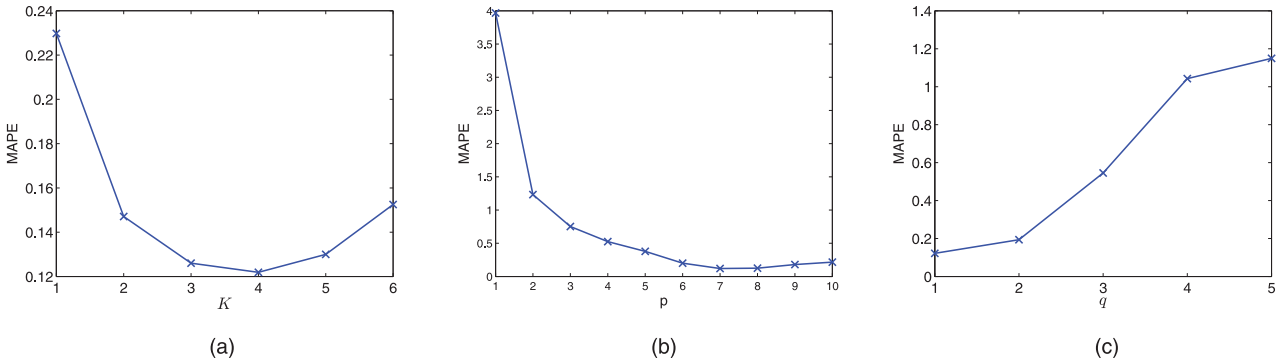


Fig. 4. The effects of parameters on the prediction accuracy. (a) Effect of K. (b) Effect of p. (c) Effect of q.

In this paper, we use the *Mean Absolute Percentage Error* (*MAPE*) [34] to measure the prediction accuracy

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|Pred_i - True_i|}{True_i},$$

where n is the total amount of predictions made on the testing data, $Pred_i$ is the predicted value, and $True_i$ represents the true value of the box office revenue. In statistics, *MAPE* is a measure of accuracy in a fitted time series value, specifically trending. The difference between actual value and the forecast value is divided by the actual value. The absolute value of this calculation is summed up for each fitted or forecast point and divided again by the number of fitted points. This makes it a percentage error so we can compare the error of fitted time series. All the accuracy results reported herein are averages of 30 independent runs.

Note that in order to evaluate the performance of ARSA, we conduct empirical studies on both blog documents and IMDB movie reviews. As similar trends are observed on those two data sets, only the experimental results for the blog documents are demonstrated from Sections 7.2 to 7.4. To verify the effectiveness of our quality-aware model, we only adopt the IMDB movie reviews, as each post in this data set is associated with a clear helpfulness score which can be used for training and testing. However, such information is not available for the blog data set. The performance of ARSQA model is shown in Section 7.5.

7.2 Parameter Selection for ARSA

In the ARSA model, there are several user-chosen parameters that provide the flexibility to fine tune the model for optimal performance. They include the number of hidden sentiment factors in S-PLSA, K, and the orders of the ARSA model, p and q. (Recall that p denotes the order of autoregression, and q specifies the sentiment information from how many preceding days is taken into account.) We now study how the choice of these parameter values affects the prediction accuracy.

We first vary K, with fixed p and q values ($p = 7$, and $q = 1$). As shown in Fig. 4a, as K increases from 1 to 4, the prediction accuracy improves, and at $K = 4$, $ARSA$ achieves a MAPE of 12.1 percent. That implies that representing the sentiments with higher dimensional probability vectors allows S-PLSA to more fully capture the sentiment information, which leads to more accurate prediction. On the other hand, as shown in the graph, the prediction accuracy deteriorates once K gets past 4. The explanation here is that a large K may cause the problem of overfitting [35], i.e., the S-PLSA might fit the training data better with a large K, but its generalization capability on the testing data might become poor. Some tempering algorithms have been proposed to solve the overfitting problem [6], but it is out of the scope of our study. Also, if the number of appraisal words used to train the model is M, and the number of blog entries is N, the total number of parameters which must be estimated in the S-PLSA model is $K(M + N + 1)$. This number grows linearly with respect to the number of hidden factors K. If K gets too large, it may incur a high training cost in terms of time and space.

We then vary the value of p, with fixed K and q values ($K = 4, q = 1$) to study how the order of the autoregressive model affects the prediction accuracy. We observe from Fig. 4b that the model achieves its best prediction accuracy when $p = 7$. This suggests that p should be large enough to factor in all the significant influence of the preceding days' box office performance, but not too large to let irrelevant information in the more distant past to affect the prediction accuracy.

Using the optimal values of K and p, we vary q from 1 to 5 to study its effect on the prediction accuracy. As shown in Fig. 4c, the best prediction accuracy is achieved at $q = 1$, which implies that the prediction is most strongly related to the sentiment information captured from blog entries posted on the immediately preceding day.

To better illustrate the effects of the parameter values on the prediction accuracy, we present in Fig. 5 the experimental results on a particular movie, *Little Man*. For each parameter, we plot the predicted box office revenues and the true values for each day using different values of the parameter. It is evident from the plots that the responses to each parameter are similar to what is observed from Fig. 4. Also note that the predicted values using the optimal parameter settings are close to the true values across the whole time span. Similar results are also observed on other movies, demonstrating the consistency of the proposed approach for different days.

7.3 Comparison with Alternative Methods

To verify that the sentiment information captured with the S-PLSA model plays an important role in box office

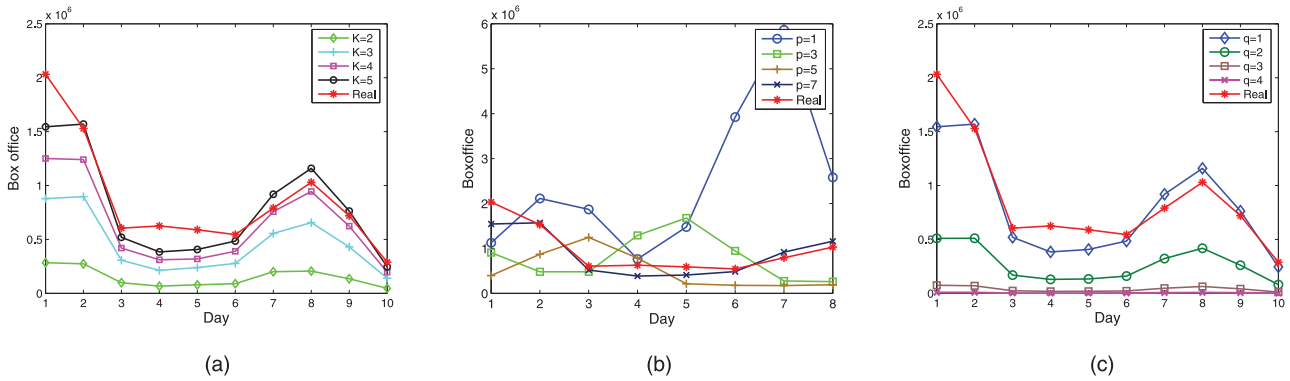


Fig. 5. The effects of parameters for the movie *Little Man*. (a) Effect of K. (b) Effect of p. (c) Effect of q.

revenue prediction, we compare ARSA with three alternative methods which do not take sentiment information into consideration.

We first conduct experiments to compare ARSA against the pure autoregressive model without any terms on sentiments, i.e., $y_t = \sum_{i=1}^{p} \phi_i y_{t-i} + \epsilon_t$. We call this model *AR-only*. The results are shown in Fig. 6. We observe the behaviors of the two models as p ranges from 3 to 7. Apparently, although the accuracy of both methods improves with increasing p, ARSA constantly outperforms the AR-only model by a factor of 2 to 3.

We then proceed to compare ARSA with an autoregressive model that uses the volume of blog mentions as the basis for prediction (which we call *volume-based*). In Section 3, we have illustrated the characteristics of the volume of blog mentions and its connection to the sales performance with an example, showing that although there exists a correlation between the volume of blog mentions and the sales performance, this correlation may not be strong enough to enable prediction. To further demonstrate this, we experiment with the following autoregressive model that utilizes the volume of blogs mentions. In contrast to ARSA, where we use a multidimensional probability vector produced by S-PLSA to represent bloggers' sentiments, this volume-based model uses a scalar (number of blog mentions) to indicate the degree of popularity. The model can be formulated as

$$y_t = \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{i=1}^{q} \rho_i u_{t-i} + \epsilon_t,$$

where y_t's are obtained in the same way as in ARSA, u_{t-i} denotes the number of reviews on day $t-i$, and ϕ_i and ρ_i are parameters to be learned. This model can be trained using a procedure similar to what is used for ARSA. Using the same training and testing data sets as what are used for ARSA, we test the performance of this model and compare it with ARSA. The results are shown in Fig. 6. Observe that although this method yields a moderate performance gain over the pure AR model (which proves that the volume data do have some predictive power), its performance is still dominated by the ARSA model.

We finally compare ARSA with a baseline method which uses true average values of previous sales for prediction. In particular, we predict the current box office revenue as the moving average in the preceding Z days. We adopt the moving average as a benchmark as it has been widely used in time series analysis to smooth out short-term fluctuations. We expect that it could capture the general trend of the sales performance. In this experiment, we vary the value of Z, and study how the prediction accuracy is affected by the choice of Z, and also compare the result of this solution against ARSA under two different parameter settings ($p = 5$, $q = 1$, $K = 4$, and $p = 7$, $q = 1$, $K = 4$).

As shown in Fig. 7, the prediction accuracy varies with different Z values, and the optimal performance is obtained at $Z = 7$. In addition, we observe that the accuracy ARSA is generally superior to the moving average method, confirming the effectiveness of our modeling approach.

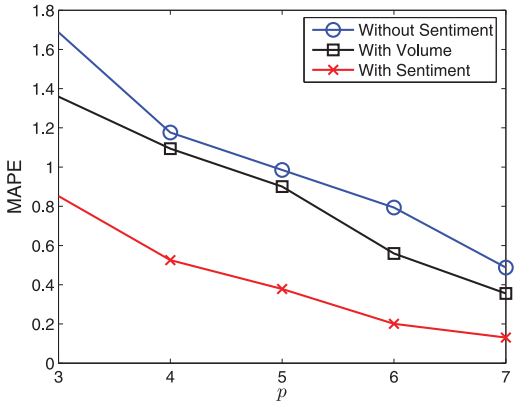


Fig. 6. ARSA versus alternative methods.

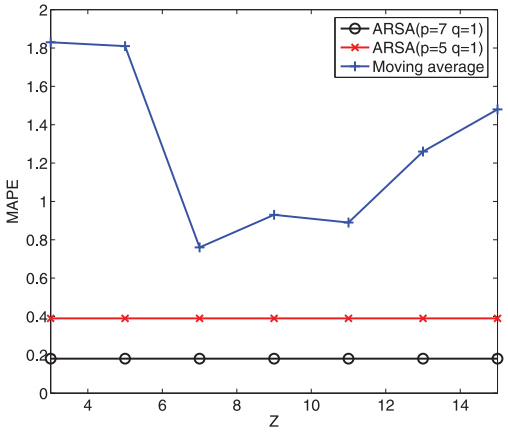


Fig. 7. Comparison with moving average prediction.

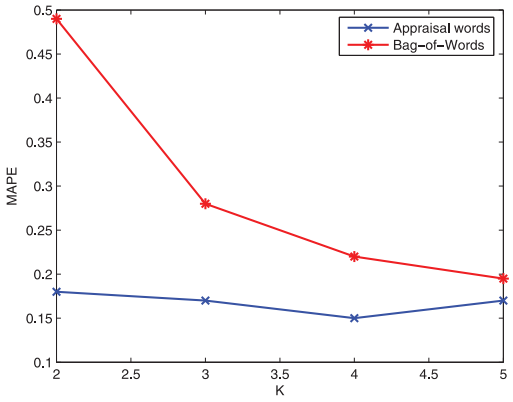


Fig. 8. Comparison with bag of words.

7.4 Comparison with Other Feature Selection Methods

To test the effectiveness of using appraisal words as the feature set, we experimentally compare ARSA with a model that uses the classic bag-of-words method for feature selection, where the feature vectors are computed using the (relative) frequencies of all the words appearing in the blog entries. That is, instead of using the appraisal words, we train an S-PLSA model with the bag-of-words feature set, and feed the probabilities over the hidden factors, thus, obtained into the ARSA model for training and prediction. Using $p = 7$ and $q = 1$, we vary K from 2 to 5 and compare the performances of both feature selection methods. As shown in Fig. 8, using appraisal words significantly outperforms the bag-of-words approach. Similar trends can be observed when other values of the parameters p, q, and K are used.

7.5 Incorporating Review Quality

To make a better estimation of future revenue, in Section 6.1, we proposed a method that can quantitatively evaluate the review quality; in Section 6.2, we developed the ARSQA model which explicitly incorporates the review quality factor into the ARSA model. To verify the effectiveness of these methods, we first evaluate the performance of our quality prediction model, and then compare the ARSQA model with the ARSA model.

To verify the effectiveness of using ϵ-Support Vector Regression to predict review quality, we compare it with the conventional Linear Regression (LR) model. To this end, we first formulate feature vectors in the same way as we described in Section 6.1, and feed them into two approximators, respectively. We then compare their performance in terms of the squared correlation coefficient r^2 and mean squared error σ^2. Let us denote the approximateors' output for review i is $\hat{\mu}_i$, and the true helpfulness value is μ_i. We have

- Squared correlation coefficient

$$r^2 = \frac{\left(\sum_{i=1}^{n}(\mu_i - \bar{\mu})(\hat{\mu}_i - \bar{\hat{\mu}})\right)^2}{\sum_{i=1}^{n}(\mu_i - \bar{\mu})^2 \sum_{i=1}^{n}(\hat{\mu}_i - \bar{\hat{\mu}})^2}.$$

- Mean squared error

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(\mu_i - \hat{\mu}_i)^2.$$

TABLE 1
Performance Comparison of Two Approximators

Approximator	r^2	σ^2
Linear regression	0.0865	0.0722
ϵ-SVR	0.2698	0.0612

As shown in Table 1, ϵ-Support Vector Regression demonstrates clear advantage over LR irrespective of the two evaluation metrics. This might be because linear approximator is not reliable if the true relation between the inputs and the output is nonlinear, although it may enjoy the benefit of being straightforward with a lower computational cost.

To further evaluate the performance of ARSQA model, we compare its performance with that of the original ARSA model. In this set of experiments, we call the model that utilizes the scores calculated with our proposed method $ARSQA_1$, and the one that directly adopts the true helpfulness values from the IMDB website $ARSQA_2$. Again, we use $p = 7$ and $q = 1$, and vary K from 2 to 5 to compare the performances of both ARSQA models and the ARSA model.

As shown in Fig. 9, incorporating true helpfulness values in $ARSQA_2$ generally outperforms the original $ARSA$ model which does not consider the effect of review quality. In addition, $ARSQA_2$ is only slightly superior to $ARSQA_1$ which again indirectly proves the effectiveness of our quality prediction model. Similar trends are observed when other values of the parameters, i.e., p and q are used.

7.6 The Impact of Other Factors on Review Quality

In this study, we have exploited the *writing style* to help assess the quality of a review. In fact, we have also considered other possible factors that may affect the helpfulness values, including the length of the review, the polarity of the review, the number of responses the review received, the subjectivity of the review, and the average rating of all reviews on the movie. Some of these factors have been studied in the previous literature to measure the quality of product reviews, e.g., reviews on digital cameras and MP3 players posted on commercial websites such as Amazon and Ebay [2], [36]. Using a set of experiments, we

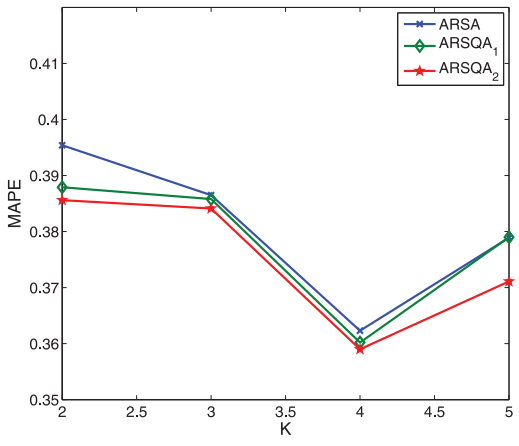


Fig. 9. Incorporating quality.

TABLE 2
Correlation Coefficient to Helpfulness Rating

Factor	Correlation Coefficient
Comment length	0.1396
Polarity	0.0947
Movie average rating	0.0070
Number of responses	0.1829
Review subjectivity	0.0307

investigate if they are effective indicators of review quality in the movie domain.

In our experiments, we use the number of sentences in a review to estimate the length of a comment in this study; the polarity of a review is considered to be the ratio of the number of positive words to the number of all appraisal words in a post;[6] the number of responses to a review is considered to be proportional the number of people who rated the review, i.e., the value of y in "x out of y people found the following review helpful"; the review subjectivity is calculated as the ratio of the number of appraisal words to the number of sentences in a post of interest.

We use the correlation coefficient to measure the strength of the linear relationship between the true helpfulness rating and each individual factor we are considering here. Specifically, for each factor (e.g., the review length), we compute the correlation coefficient between the helpfulness rating $\{\mu_i\}_{i=1}^N$ and the corresponding values of the factor $\{f_i\}_{i=1}^N$ on the set of N reviews in the movie data set. The result is recorded in Table 2.

As shown in Table 2, none of the above factors demonstrates strong correlation to the helpfulness ratings in the movie domain, which suggests that those factors may not be good indicators of review quality. This might be because the task of review quality mining is inherently nontrivial [37]; only utilizing those simple 1D features is not powerful enough to capture its intricate nature. We, therefore, opt to use the *writing style*, which is represented as a multidimensional vector, to model the quality of reviews.

8 Practical Impact of the Proposed Models

As pointed out in [38], a KDD process should not be a purely data-driven process; it should be a domain-driven process where the ultimate goal is to produce actionable knowledge. We have kept this mind when developing the framework of prediction, and we expect that the outcome of the proposed models can be readily used to support decision making in the real world. In this section, we discuss the practical impact of the proposed models, and explain how various players in the movie business can benefit from the deployment of those models in a number of different ways.

For theaters, accurate prediction of future box office revenues can help better allocating resources. For example, if a movie is poorly received (as reflected in online reviews) and is predicted by our model to have declining box office revenues, then the manager of a theater could decide to use smaller showing rooms for this movie, while allocating larger rooms to more popular movies. It is worth noting that, in general, the contracts between the distributors and the theaters favor the former in terms of revenue splitting in early weeks, and shift to the theaters later on. Therefore, for the theaters, if they find out that the box office performance in the early weeks does not meet their expectation, they could initiate some last-minute remedies on their own to help boost the box office revenue, e.g., through increased advertising efforts and other marketing campaigns.

In our proposal, we have used "day" as the unit for box office revenues to illustrate our methodology. In practice, one might wish to use larger time units. For example, it might be too late to make any managerial decisions today for actions to be taken tomorrow based on its predicted box office revenue. Fortunately, our model is general enough to accommodate larger time units. In general, larger time units (3-day, week, month, etc.) can be used, with only minor modification needed in most cases. For example, if "week" is used as the time unit, then all reviews collected are grouped and analyzed by weeks, and all box office revenues (both past and predicted) become weekly totals. The main difference from using "day" as the unit is that the seasonality removal procedure is no longer needed, as the seasonality within a week is no longer a problem when box office revenues within a week are now considered as a whole.

Although in this paper we have used the movie domain as a case study, the same methodology can be adapted to the task of predicting sales performance in other domains. The exact models and preprocessing steps used might have to be modified to accommodate the specific characteristics of the target domains. For example, the time series for sales performance in other domains (such as electronic products) may not exhibit the same periodic fluctuation as in the movie domain, and therefore, the preprocessing step of deseasonality becomes unnecessary. As another example, in some domains, the autoregressive model may not be a good fit to the time series; other models such as the autoregressive moving average model [8], may be better candidates. In such cases, we simply have to replace the AR component in the ARSA model with the proper time series models.

Equipped with the proposed models and the actionable knowledge they produce, decision makers will be better informed in making critical business decisions, resulting in significant competitive advantages. For example, if an online retailer (such as Amazon) finds out that a particular product (e.g., a book) is expected to generate more sales, it could increase its stock of that product to accommodate the increasing demand. On the other hand, if the sale performance of a particular product is lower than expected, the retailer could either decrease its stock of that product, or try to revive its sales through focused marketing campaigns.

In summary, the proposed models can be used to obtain actionable knowledge, and are flexible enough to be deployed in a variety of settings.

9 Conclusions and Future Work

The wide spread use of online reviews as a way of conveying views and comments has provided a unique opportunity to understand the general public's sentiments and derive business intelligence. In this paper, we have explored the

6. We adopted the lexicon of the appraisal words from [7].

predictive power of reviews using the movie domain as a case study, and studied the problem of predicting sales performance using sentiment information mined from reviews. We have approached this problem as a domain-driven task, and managed to synthesize human intelligence (e.g., identifying important characteristics of movie reviews), domain intelligence (e.g., the knowledge of the "seasonality" of box office revenues), and network intelligence (e.g., online reviews posted by moviegoers). The outcome of the proposed models leads to actionable knowledge that be can readily employed by decision makers.

A center piece of our work is the proposal of S-PLSA, a generative model for sentiment analysis that helps us move from simple "negative or positive" classification toward a deeper comprehension of the sentiments in blogs. Using S-PLSA as a means of "summarizing" sentiment information from reviews, we have developed ARSA, a model for predicting sales performance based on the sentiment information and the product's past sales performance. We have further considered the role of review quality in sales performance prediction, and proposed a model to predict the quality rating of a review when it is not readily available. The quality factor is then incorporated into a new model called ARSQA. The accuracy and effectiveness of the proposed models have been confirmed by the experiments on two movie data sets. Equipped with the proposed models, companies will be able to better harness the predictive power of reviews and conduct businesses in a more effective way.

It is worth noting that although we have only used S-PLSA for the purpose of prediction in this work, it is indeed a model general enough to be applied to other scenarios. For future work, we would like to explore its role in clustering and classification of reviews based on their sentiments. It would also be interesting to explore the use of S-PLSA as a tool to help track and monitor the changes and trends in sentiments expressed online. Also note that the ARSA and ARSQA models are general frameworks for sales performance prediction, and would certainly benefit from the development of more sophisticated models for sentiment analysis and quality prediction.

ACKNOWLEDGMENTS

This work was supported in part by National Natural Science Foundation of China Grants (No. 61070018, No. 60903108, No. 61003051), the Program for New Century Excellent Talents in University (NCET-10-0532), grants from Natural Sciences and Engineering Research Council of Canada (NSERC), the Early Researcher Award of Ontario, the Independent Innovation Foundation of Shandong University (2009TB016), and the SAICT Experts Program. Y. Liu is the corresponding author for this paper.

REFERENCES

[1] D. Gruhl, R. Guha, R. Kumar, J. Novak, and A. Tomkins, "The Predictive Power of Online Chatter," *Proc. 11th ACM SIGKDD Int'l Conf. Knowledge Discovery in Data Mining (KDD),* pp. 78-87, 2005.

[2] A. Ghose and P.G. Ipeirotis, "Designing Novel Review Ranking Systems: Predicting the Usefulness and Impact of Reviews," *Proc. Ninth Int'l Conf. Electronic Commerce (ICEC),* pp. 303-310, 2007.

[3] Y. Liu, X. Huang, A. An, and X. Yu, "ARSA: A Sentiment-Aware Model for Predicting Sales Performance Using Blogs," *Proc. 30th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR),* pp. 607-614, 2007.

[4] Y. Liu, X. Yu, X. Huang, and A. An, "Blog Data Mining: The Predictive Power of Sentiments," *Data Mining for Business Applications,* pp. 183-195, Springer, 2009.

[5] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins, "Information Diffusion through Blogspace," *Proc. 13th Int'l Conf. World Wide Web (WWW),* pp. 491-501, 2004.

[6] T. Hofmann, "Probabilistic Latent Semantic Analysis," *Proc. Uncertainty in Artificial Intelligence (UAI),* 1999.

[7] C. Whitelaw, N. Garg, and S. Argamon, "Using Appraisal Groups for Sentiment Analysis," *Proc. 14th ACM Int'l Conf. Information and Knowledge Management (CIKM),* pp. 625-631, 2005.

[8] W. Enders, *Applied Econometric Time Series,* second ed. Wiley, 2004.

[9] L. Cao, C. Zhang, Q. Yang, D. Bell, M. Vlachos, B. Taneri, E. Keogh, P.S. Yu, N. Zhong, M.Z. Ashrafi, D. Taniar, E. Dubossarsky, and W. Graco, "Domain-Driven, Actionable Knowledge Discovery," *IEEE Intelligent Systems,* vol. 22, no. 4, pp. 78-88, July/Aug. 2007.

[10] L. Cao, Y. Zhao, H. Zhang, D. Luo, C. Zhang, and E.K. Park, "Flexible Frameworks for Actionable Knowledge Discovery," *IEEE Trans. Knowledge and Data Eng.,* vol. 22, no. 9, pp. 1299-1312, Sept. 2009.

[11] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs Up? Sentiment Classification Using Machine Learning Techniques," *Proc. ACL-02 Conf. Empirical Methods in Natural Language Processing (EMNLP),* 2002.

[12] B. Pang and L. Lee, "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts," *Proc. 42nd Ann. Meeting on Assoc. for Computational Linguistics (ACL),* pp. 271-278, 2004.

[13] J. Kamps and M. Marx, "Words with Attitude," *Proc. First Int'l Conf. Global WordNet,* pp. 332-341, 2002.

[14] P.D. Turney, "Thumbs Up or Thumbs Down?: Semantic Orientation Applied to Unsupervised Classification of Reviews," *Proc. 40th Ann. Meeting on Assoc. for Computational Linguistics (ACL),* pp. 417-424, 2001.

[15] B. Pang and L. Lee, "Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales," *Proc. 43rd Ann. Meeting on Assoc. for Computational Linguistics (ACL),* pp. 115-124, 2005.

[16] Z. Zhang and B. Varadarajan, "Utility Scoring of Product Reviews," *Proc. 15th ACM Int'l Conf. Information and Knowledge Management (CIKM),* pp. 51-57, 2006.

[17] B. Liu, M. Hu, and J. Cheng, "Opinion Observer: Analyzing and Comparing Opinions on the Web," *Proc. 14th Int'l Conf. World Wide Web (WWW),* pp. 342-351, 2005.

[18] N. Archak, A. Ghose, and P.G. Ipeirotis, "Show Me the Money!: Deriving the Pricing Power of Product Features by Mining Consumer Reviews," *Proc. 13th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD),* pp. 56-65, 2007.

[19] J.A. Chevalier and D. Mayzlin, "The Effect of Word of Mouth on Sales: Online Book Reviews," *J. Marketing Research,* vol. 43, no. 3, pp. 345-354, Aug. 2006.

[20] C. Dellarocas, X.M. Zhang, and N.F. Awad, "Exploring the Value of Online Product Ratings in Revenue Forecasting: The Case of Motion Pictures," *J. Interactive Marketing,* vol. 21, no. 4, pp. 23-45, 2007.

[21] S. Rosen, "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition," *J. Political Economy,* vol. 82, no. 1, pp. 34-55, 1974.

[22] N.Z. Foutz and W. Jank, "The Wisdom of Crowds: Pre-Release Forecasting via Functional Shape Analysis of the Online Virtual Stock Market," Technical Report 07-114 Marketing Science Inst. of Reports, 2007.

[23] N.Z. Foutz and W. Jank, "Pre-Release Demand Forecasting for Motion Pictures Using Functional Shape Analysis of Virtual Stock Markets," *Marketing Science,* to be published, 2010.

[24] N. Jindal and B. Liu, "Opinion Spam and Analysis," *Proc. Int'l Conf. Web Search and Web Data Mining (WSDM),* pp. 219-230, 2008.

[25] J. Liu, Y. Cao, C.-Y. Lin, Y. Huang, and M. Zhou, "Low-Quality Product Review Detection in Opinion Summarization," *Proc. Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP),* pp. 334-342, 2007.

[26] C. Elkan, *Method and System for Selecting Documents by Measuring Document Quality.* US patent 7,200,606, Washington, D.C.: Patent and Trademark Office, Apr. 2007.
[27] B. Sarwar, G. Karypis, J. Konstan, and J. Reidl, "Item-Based Collaborative Filtering Recommendation Algorithms," *Proc. 10th Int'l Conf. World Wide Web (WWW),* pp. 285-295, 2001.
[28] T. Hofmann, "Probabilistic Latent Semantic Indexing," *Proc. 22nd Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR),* pp. 50-57, 1999.
[29] A. Popescul, L.H. Ungar, D.M. Pennock, and S. Lawrence, "Probabilistic Models for Unified Collaborative and Content-Based Recommendation in Sparse-Data Environments," *Proc. 17th Conf. in Uncertainty in Artificial Intelligence (UAI),* pp. 437-444, 2001.
[30] J. Basilico and T. Hofmann, "Unifying Collaborative and Content-Based Filtering," *Proc. 21st Int'l Conf. Machine Learning (ICML),* p. 9, 2004.
[31] X. Jin, Y. Zhou, and B. Mobasher, "A Maximum Entropy Web Recommendation System: Combining Collaborative and Content Features," *Proc. 11th ACM SIGKDD Int'l Conf. Knowledge Discovery in Data Mining (KDD),* pp. 612-617, 2005.
[32] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum Likelihood from Incomplete Data via the *em* Algorithm," *J. Royal Statistical Soc.,* vol. 39, no. 1, pp. 1-38, 1977.
[33] Y. Liu, X. Huang, A. An, and X. Yu, "Modeling and Predicting the Helpfulness of Online Reviews," *Proc. Eighth IEEE Int'l Conf. Data Mining (ICDM),* pp. 443-452, 2008.
[34] W. Jank, G. Shmueli, and S. Wang, "Dynamic, Real-Time Forecasting of Online Auctions via Functional Models," *Proc. 12th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD),* pp. 580-585, 2006.
[35] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet Allocation," *J. Machine Learning Research,* vol. 3, pp. 993-1022, 2003.
[36] S.-M. Kim, P. Pantel, T. Chklovski, and M. Pennacchiotti, "Automatically Assessing Review Helpfulness," *Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP),* pp. 423-430, 2006.
[37] B. Liu, M. Hu, and J. Cheng, "Opinion Observer: Analyzing and Comparing Opinions on the Web," *Proc. 14th Int'l Conf. World Wide Web (WWW),* pp. 342-351, 2005.
[38] L. Cao, "Domain Driven Data Mining (d3m)," *Proc. IEEE Int'l Conf. Data Mining Workshops (ICDM),* pp. 74-76, 2008.

Xiaohui Yu received the BSc degree from Nanjing University, China, the MPhil degree from The Chinese University of Hong Kong, and the PhD degree from the University of Toronto, Canada. His research interests include the areas of database systems and data mining. He has published more than 30 papers in premier venues, including SIGMOD, VLDB, ICDE, EDBT, SIGIR, and ICDM. He has served on the program committees/review boards of various conferences and workshops, such as VLDB '11 and CIKM '09, and was the program cochair of the OMBI '10 Workshop, and the WI-IAT '10 Doctoral Workshop. He is an associate professor in the School of Information Technology, York University, Canada, and a member of the IEEE.

Yang Liu received the BSc degree from the Harbin Institute of Technology, China, and the MSc and PhD degrees from York University, Canada. Currently, she is working as an associate professor in the School of Computer Science and Technology, Shandong University, China. Her main areas of research include data mining and information retrieval. She has published in top conferences and journals, including ICDM, SIGIR, WWW, and the *Journal of the American Society for Information Science and Technology*. She has served on the program committees of various conferences and workshops, such as CIKM '10 and WI-IAT '10, and was the program cochair of the OMBI '10 Workshop. She is a member of the IEEE.

Jimmy Xiangji Huang received the PhD degree in information science at City University in London. Currently, he is working as a professor in the School of Information Technology and the founding director of the Information Retrieval and Knowledge Management Research Lab at York University. Previously, he was a postdoctoral fellow in the School of Computer Science, University of Waterloo. He also worked in the financial industry in Canada, where he was awarded a CIO Achievement Award. He has published more than 110 refereed papers in top ranking journals, book chapters, and international conference proceedings. In April 2006, he was awarded tenure at York University. He received the Dean's Award for Outstanding Research in 2006, an Early Researcher Award, formerly the Premier's Research Excellence Awards in 2007, the Petro Canada Young Innovators Award in 2008, the SHARCNET Research Fellowship Award in 2009, and the Best Paper Award at the 32nd European Conference on Information Retrieval in 2010. He was the general conference chair for the 19th International ACM CIKM Conference and the general program chair for IEEE/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology in 2010. He is a member of the IEEE.

Aijun An received the PhD degree in computer science from the University of Regina in 1997. Currently, she is working as an associate professor in the Department of Computer Science and Engineering at York University, Toronto, Canada. She held research positions at the University of Waterloo from 1997 to 2001. She joined York University in 2001. She has published more than 80 papers in refereed journals and conference proceedings. Her major research area includes data mining. She has worked on various data mining problems including classification, clustering, indirect association mining, transitional pattern mining, diverging pattern mining, review mining, data stream mining, and bioinformatics. She is a member of the IEEE.