An Adaptive Model for Probabilistic Sentiment Analysis

Xiaohui Yu†∗, Yang Liu†, Aijun An‡

†Department of Computer Science and Engineering
York University, Toronto, Canada
ahyan@cse.yorku.ca

‡School of Information Technology
York University, Toronto, Canada
xyu@yorku.ca

∗School of Computer Science and Technology
Shandong University, Jinan, Shandong, China
{xyu,yliu}@sdu.edu.cn

I. INTRODUCTION

Abstract—Online reviews, which are getting increasingly prevalent with the rapid growth of Web 2.0, have been shown to be second only to “word-of-mouth” in terms of influencing purchase decisions. It is therefore imperative to analyze them and distill useful knowledge that could be of economic values to vendors and other interested parties. Previous studies have confirmed that the sentiments expressed in the online reviews are strongly correlated with the sales performance of products. In particular, a model called ARSA has been proposed for predicting sales performance using a model called S-PLSA. In this paper, we build upon that work, and present an adaptive sentiment analysis model called S-PLSA+, which not only can capture the hidden sentiment factors in the reviews, but has the capability to be incrementally updated as more data become available. We show how the proposed S-PLSA+ model can be applied to sales performance prediction using the ARSA model. A case study is conducted in the movie domain, and results from preliminary experiments confirm the effectiveness of the proposed model.

Keywords—sentiment analysis; review mining; prediction

Whereas marketing plays an important role for the newly released products, public opinion about the products might be crucial to determine their success in the long run. Such effect is largely magnified thanks to the rapid growth of Web 2.0 which encourages user participation. In fact, online reviews have been shown to be second only to word-of-mouth in a study that compares the factors influence purchase decisions [17]. Therefore, online reviews can be very valuable, as collectively such reviews reflect the “wisdom of crowds” and can be a good indicator of the product’s future sales performance.

Researchers have also recognized the impact of online reviews, 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 that are available online have a significant effect on actual customer purchasing [1], [7], [13], [12]. Various economic functions have been utilized to examine the relationship between opinions discovered from product reviews and revenue growth, stock trading volume change, as well as the bidding price variation in commercial Websites, such as eBay [4], [8], [24]. In particular, Gruhl et al. [12] show that the volume of relevant postings can help predict the sales rank of books on Amazon, especially the spikes in sales ranks. In contrast to above work which captures review sentiments with explicit rating indication such as the number of stars, there are also a few studies that attempt to exploit text mining strategies for sentiment understanding. For example, Ghose et al. [9] demonstrate that the subjectivity of reviews can have an impact on sales performance, and review texts contain rich information that cannot be easily captured using numerical ratings.

In the same vein, Liu et al. [20] study the important problem of mining opinions and sentiments from reviews, and utilize the extracted patterns for predicting future product sales. Based on the understanding that simply classifying reviews as positive or negative, as most current sentiment-mining approaches are designed for, does not provide a comprehensive explanation of the sentiments reflected in reviews, they propose a probability model called Sentiment PLSA (S-PLSA for short) based on the assumption that sentiment consists of multiple hidden aspects. They develop a model called ARSA (which stands for Auto-Regressive Sentiment-Aware) to quantitatively measure the relationship between sentiment aspects and reviews.

Our experience with running ARSA on several online review datasets reveals that the model is highly sensitive to the sentiment factors, which are constantly changing over time as new reviews become available. It is therefore essential to allow the S-PLSA model to adapt to newly available review data.

To this end, we take a Bayesian approach, and propose an adaptive version of the S-PLSA model that is equipped with the incremental learning capability for continuously updating the model using newly observed reviews. The proposed model is motivated by the principle of quasi-Bayesian
(QB) estimation, which has found successful applications in various domains such as adaptive speech recognition and text retrieval [5]. We call the proposed model the S-PLSA model, in which the parameters are estimated by maximizing an approximate posterior distribution. One salient feature of our modeling is the judicious use of hyperparameters, which can be recursively updated in order to obtain up-to-date posterior distribution and to estimate new model parameters. This modeling approach makes it possible to efficiently update the model parameters in an incremental manner without the need to re-train the model from scratch each time as new reviews become available.

The rest of the paper is organized as follows. In Section 2, we provide a review of related work. In Section 3, we start with presenting the general framework of S-PLSA for sentiment understanding based on reviews. In Section 4, we propose S-PLSA+, a Bayesian adaptive model which effectively update the system parameters without incurring much cost. We present the experimental results on a movie review dataset in Section 5, and conclude this paper in Section 5.

II. RELATE WORK

A. Sentiment Mining

Instead of being informed only by manufacture’s agents, potential consumers are now able to objectively evaluate the product by viewing others’ opinions online before making a decision. Nonetheless, the task of understanding knowledge of interest from online reviews can be difficult. With more and more users becoming comfortable with the Web, a large quantity of people are writing reviews online. Consequently, the number of reviews grows rapidly. When trying to locate information on a product, a general Web search would retrieve a large collection of documents; however getting an overall sense of the reviews can be daunting and time-consuming. To solve these problems, recent years have seen a growing interest in sentiment mining, whose objective is to find opinions, feelings, and attitude expressed in text, rather than facts. In the literature, sentiment mining also goes under various names, such as opinion mining [6], [11], [21], sentiment analysis [26], [22], [23], etc. Its related work may come from both computer science and linguistics, and its immediate applications may involve data mining, market intelligence, and customer relationship management.

In literature, existing sentiment analysis techniques can be generally divided into three sub-directions: determining subjectivity [27], [28], determining orientation, and determining the strength of orientation [25], [26], and most of the studies focus on investigating the sentiment orientation of words, phrases, and documents. For example, Hatzivassiloglou et al. [14] first proposed to determine the semantic orientation of adjectives by analyzing pairs of adjectives (conjoined by and, or, but, either-or, or neither-nor) extracted from large unlabeled corpora. Kamps et al. [18] tried to evaluate the semantic distance from a word to good/bad with WordNet. They first defined a graph on the adjectives appeared in both the WordNet and the target term list. If two adjectives in WordNet display a synonym relation, a link will be added between them. In turn, the semantic orientation of a word \( w \) is decided by its relative distance to \emph{good} and \emph{bad}. Liu et al. [19] built 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 were visualized with an \emph{Opinion Observer}. Observed that simply classifying reviews as being positive or negative, as most of the previous work is designed for, does not provide a comprehensive understanding of sentiments reflected, our previous work [20] assumes that sentiment consists of multiple hidden aspects, and use a probability model to quantitatively measure the relationship between sentiment aspects and reviews.

B. Latent Semantic Model

Latent semantic modeling has become very popular as a completely unsupervised technique for topic discovery in large documents. These models, such as PLSA [15] and LDA [3], exploit co-occurrence patterns of words in documents to understand semantically meaningful probabilistic clusters of words. These models assign a probabilistic membership to documents in the latent topic space, assisting us for viewing and processing the data in a lower-dimensional space. PLSA was shown to be a special variant of LDA with a uniform Dirichlet prior in a maximum a posterior model [10], and has been successfully applied to content-based recommendation and collaborative filtering [2], [16]. However, one limitation of the model is its incapacity of adapting itself as new data become available, and the problem will get worse when the data arrive in a stream. This is due to the fact that the PLSA model is estimated only for documents that appear in the training set, and re-training model using both existing training data and new data from scratch is highly inefficient. Motived by the idea of quasi-Bayes estimate, Chien et al. [5] propose an incremental learning method to estimate the model parameters by maximizing an approximate posterior distribution, and expect that such an approach can effectively absorb the domain knowledge from the newly arrived data. Here, we take a similar methodology, but explore the possibility of developing a model for accurately predicting the product sales using the sentiments that dynamically change as new online reviews come in.

III. S-PLSA

Many existing models and algorithms for sentiment mining are developed for the binary classification problem, i.e., to classify the sentiment of a review as positive or negative. However, sentiments are often multi-faceted, and can differ from one another in a variety of ways, including
polarity, orientation, graduation, etc. Moreover, sentiments are often multifaceted, and can differ from one another in a variety of ways, including polarity, orientation, graduation, etc. Therefore, for applications requiring a more accurate understanding of the sentiments, it would be too simplistic to just classify the sentiments expressed in a review as either positive or negative. Moreover, mining opinions and sentiments present unique challenges that cannot be addressed easily with traditional text mining algorithms, due to the fact that opinions and sentiments, which are usually written in natural languages, are often expressed in subtle and complex ways. All these concerns call for a model that can extract the sentiments in a more accurate manner.

To this end, Liu et al. [20] propose the S-PLSA model, in which a review can be considered as being generated under the influence of a number of hidden sentiment factors. Inspired the PLSA model [15], [16], the use of hidden factors in S-PLSA provides the model the ability to accommodate the intricate nature of sentiments, with each hidden factor focusing on one specific aspect. The use of a probabilistic generative model, on the other hand, allows for dealing with sentiment analysis in a principled way.

What differentiates S-PLSA from conventional PLSA is its use of a set of appraisal words [26] as the basis for feature representation. In order to represent a given review as an input to the mining algorithm, the traditional way would compute the (relative) frequencies of various words in a review and use the resulting multidimensional feature vector as the representation of the document. In the S-PLSA model, we follow the same methodology. However, instead of using the frequencies of all words appearing the blogs (barring any stop words), we choose to focus on the set of 2,030 appraisal words extracted from the lexicon constructed by Whitelaw et al. [26]. As a concrete example of appraisal words, the lexical entry for the appraisal word beautiful can be described as follows:

- **Attitude:** appreciation/reaction-quality
- **Orientation:** positive
- **Force:** neutral
- **Focus:** neutral
- **Polarity:** unmarked

where the adjective is fully described with four types of attributes. In this context, the attitude of an appraisal word provides the appraisal expressed as either **affect**, **appreciation**, or **judgment**. Orientation describes whether the appraisal is **positive** or **negative**. Graduation presents the intensity of appraisal wrt. two independent folds of **force** and **focus**. Finally, polarity is **marked** if it is confined with a polarity marker (such as ‘not’), or **unmarked** otherwise [26]. In this study, we adopt the above lexicon, and utilize the frequencies of such appraisal words in a review as the basis for the feature vector. The rationale is that those appraisal words, such as “**good**” or “**terrible**”, are more indicative of the review’s sentiments than other words.

For a given set of \( N \) reviews \( D = \{ d_1, \ldots, d_N \} \), and the set of \( M \) appraisal words \( \mathcal{W} = \{ w_1, \ldots, w_M \} \), the S-PLSA model dictates that the joint probability of observed pair \(( d_i, w_j) \) is generated by

\[
P(d_i, w_j) = P(d_i) \sum_{k=1}^{K} P(w_j | z_k) P(z_k | d_i),
\]

where \( z_k \in Z = \{ z_1, \ldots, z_K \} \) corresponds to the latent sentiment factor, and where we assume that \( d_i \) and \( w_j \) are independent conditioned on the mixture of associated sentiment factor \( z_k \). The set of parameters \( \theta \) of this model consist of \( \{ P(w_j | z_k), P(z_k | d_i) \} \), where \( \sum_{j=1}^{M} P(w_j | z_k) = 1 \) and \( \sum_{k=1}^{K} P(z_k | d_i) = 1 \), and there totally exist \( KM + KN \) probabilities in \( \theta \). If we consider the number \( c(d_i, w_j) \) of word \( w_j \) occurring in document \( d_i \) and accumulate the log likelihood of training data \( X = \{ d_i, w_j \} \) using \( \theta \), then

\[
\log P(X|\theta) = \sum_{i=1}^{n} \sum_{j=1}^{M} c(d_i, w_j) \log P(d_i, w_j).
\]

S-PLSA parameter set \( \theta \) thus can be found by maximizing the accumulated log likelihood

\[
\theta_{ML} = \arg \max_{\theta} \log P(X|\theta).
\]

As the hidden parameter \( z_k \) is embedded in the above function, the expectation-maximization (EM) algorithm ([15], [16]) can be adopted to estimate the probabilities.

### IV. Adaptive S-PLSA

When the characteristics of the underlying data are relatively stable and do not evolve significantly over time, it is possible to train the S-PLSA model in a batch manner on a collection of reviews, and apply the trained model on unseen reviews encountered in the future. In many cases, however, the reviews are continuously becoming available, with the sentiment factors constantly changing. Moreover, a model trained on a corpus consisting of mixed reviews on different types of products may be too general and not “accurate” enough to capture the specific characteristics of the newly available reviews. We thus hope to adapt the model to the newly obtained reviews, in order to make it more suitable to the changing contexts. A naïve way to perform the adaptation is to re-train the model from scratch using all data available including the newly obtained data, which has two drawbacks: (i) it is clearly highly inefficient, especially when the data volume is high; and (ii) the out-of-date reviews from a long time ago and not relevant anymore may actually harm the performance of the model if they are not included in training.

An alternative solution is to develop a S-PLSA model that only takes the newly available data into consideration and discards all old data. This approach, however, may suffer
from the problem of not having sufficient amount of training samples, as it is very likely that only a few reviews are written within a short period of time. Also, discarding the old data in its entirety may be unwise, because knowledge obtained from those data (which is reflected in the model parameters) is lost.

Here, we propose a model called S-PLSA+, which performs incremental learning based on the principle of quasi-Bayesian (QB) estimation. The basic idea is to perform updating using the new data and fading away the out-dated data at the same time by:

1) incrementally accumulating statistics on the training data, and

2) fading out the out-of-date data.

Let \( D_n \) be the set of reviews made available at epoch \( n \) (e.g., the reviews published on a certain day, but the time unit used can be set to finer or coarser based on the need), and denote by \( \chi^n = \{ D_1, \ldots, D_n \} \) the set of reviews obtained up to epoch \( n \). In order to support parameter update based on new knowledge, we take a Bayesian approach and perform maximum a posteriori (MAP) estimation instead of maximum-likelihood estimation as stated in the preceding section. The MAP estimates for S-PLSA+ at epoch \( n \) are determined by maximizing the posterior probability using \( \chi^n \):

\[
\theta^{(n)} = \arg \max_{\theta} P(\theta | \chi^n) = \arg \max_{\theta} P(D_n | \theta) P(\theta | \chi^{n-1})
\]

(4)

The learning (i.e., update of parameters) is expected to be done repeatedly at different epochs.

In order to allow closed-form recursive update of \( \theta \), we use the closest tractable prior density \( g(\theta | \phi^{(n-1)}) \) with sufficient statistics to approximate the posterior density \( P(\theta | \chi^{n-1}) \), where \( \phi^{(n-1)} \) is evolved from review set \( \chi^{n-1} \). This leads to

\[
\hat{\theta}^{(n)} \approx \arg \max_{\theta} P(D_n | \theta) g(\theta | \phi^{(n-1)}).
\]

(5)

Note that at epoch \( n \), only the new reviews \( D_n \) and the current statistics \( \phi^{(n-1)} \) are used to update the S-PLSA+ parameters, and the set of reviews \( D_n \) are discarded after new parameter values \( \phi^{(n)} \) are obtained, which results in significant savings in computational resources.

The particular choice of the prior \( g(\theta | \phi) \) in our model is the Dirichlet density, which can be expressed by

\[
g(\theta | \phi) = \prod_{k=1}^{K} \left[ \prod_{j=1}^{M} P(w_j | z_k)^{\alpha_{j,k} - 1} \prod_{i=1}^{N} P(z_k | d_i)^{\beta_{k,i} - 1} \right]
\]

(6)

where \( \phi = \{ \alpha_{j,k}, \beta_{k,i} \} \) are the hyperparameters of the Dirichlet distribution. This choice of conjugate prior allows for a closed-form solution for fast model adaptation.

Assuming for the moment that \( \phi^{(n-1)} \) is known, we can show that \( \hat{\theta}^{(n)} \) can be obtained through an EM algorithm [5], and \( \hat{\theta}^{(n)} \) can be obtained by

\[
\hat{\theta}^{(n)}(w_i | z_k) = \frac{\sum_{j=1}^{N} c(d_i^{(n)}, w_j^{(n)}) P(z_k | d_i^{(n)}, w_j^{(n)}) + (\alpha_{j,k}^{(n-1)} - 1)}{\sum_{k=1}^{K} \sum_{i=1}^{M} c(d_i^{(n)}, w_j^{(n)}) P(z_k | d_i^{(n)}, w_j^{(n)}) + (\alpha_{j,k}^{(n-1)} - 1)}
\]

(7)

\[
\hat{\beta}_{k,i}^{(n)} = \sum_{j=1}^{M} c(d_i^{(n)}, w_j^{(n)}) P(z_k | d_i^{(n)}, w_j^{(n)}) + \beta_{k,i}^{(n-1)}
\]

(8)

A major benefit of S-PLSA+ lies in its ability to continuously update the hyperparameters. We can show that the new hyperparameters are given by

\[
\alpha_{j,k}^{(n)} = \sum_{i=1}^{|D_n|} c(d_i^{(n)}, w_j^{(n)}) P(n | z_k | d_i^{(n)}, w_j^{(n)}) + \alpha_{j,k}^{(n-1)}
\]

(9)

\[
\beta_{k,i}^{(n)} = \sum_{j=1}^{M} c(d_i^{(n)}, w_j^{(n)}) P(n | z_k | d_i^{(n)}, w_j^{(n)}) + \beta_{k,i}^{(n-1)}
\]

(10)

where the posterior \( P(n | z_k | d_i^{(n)}, w_j^{(n)}) \) is computed using \( D_n \) and the current parameters \( \hat{\theta}^{(n)} \), and \( c(d_i^{(n)}, w_j^{(n)}) \) denotes the number of \( (d_i^{(n)}, w_j^{(n)}) \) pairs.

To summarize, S-PLSA+ works as follows. In the startup phase, initial estimates of the hyperparameters \( \phi^{(0)} \) are obtained. Then, at each learning epoch \( n \), (i) new estimates of the parameters \( \hat{\theta}^{(n)} \) are computed based on the newly available data \( D_n \) and hyperparameters obtained from epoch \( n-1 \); and (ii) new estimates of the hyperparameters \( \phi^{(n)} \) are obtained using (9) and (10). This way, the model is continuously updated when new reviews \( (D_n) \) become available, and at the same time fades out historical data \( \chi^{n-1} \), with the information contained in \( \chi^{n-1} \) already captured by \( \phi^{(n-1)} \). The implementation procedure for parameter update at each epoch \( n \) is shown in Algorithm 1.
Algorithm 1: S-PLSA parameter update at epoch $n$

Input: $\theta^{(n-1)}$, newly available data $D_n$, hyperparameters $\phi^{(n-1)}$, and EM algorithm threshold $\epsilon$

Output: $\theta^{(n)}$, $\phi^{(n)}$

1 while $\frac{|\theta^{(n)} - \theta^{(n-1)}|}{\theta^{(n)}} > \epsilon$ do

2 compute $\hat{P}(n) (z_k | d_i^{(n)}) = \frac{\sum_{i=1}^{M} c(d_i^{(n)}, w_j^{(n)}) P(z_k | d_i^{(n)}, w_j^{(n)}) + (\alpha_k^{(n-1)} - 1)}{\sum_{k=1}^{K} \sum_{i=1}^{M} c(d_i^{(n)}, w_j^{(n)}) + (\alpha_j^{(n-1)} - 1)}$;

3 compute $\hat{P}(n) (z_k | d_i^{(n)}) = \frac{\sum_{i=1}^{M} c(d_i^{(n)}, w_j^{(n)}) P(z_k | d_i^{(n)}, w_j^{(n)}) + (\beta_k^{(n-1)} - 1)}{\sum_{i=1}^{M} \sum_{j=1}^{W} (\beta_{i,j}^{(n-1)} - 1)}$;

4 end

5 update $\alpha_{j,k}^{(n)} = \sum_{i=1}^{M} c(d_i^{(n)}, w_j^{(n)}) P(n) (z_k | d_i^{(n)}, w_j^{(n)}) + \alpha_{j,k}^{(n-1)}$;

6 update $\beta_{k,i}^{(n)} = \sum_{j=1}^{W} c(d_i^{(n)}, w_j^{(n)}) P(n) (z_k | d_i^{(n)}, w_j^{(n)}) + \beta_{k,i}^{(n-1)}$;

7 return $\theta^{(n)}$, $\phi^{(n)}$;

V. APPLICATION TO SALES PREDICTION

The proposed S-PLSA model can be employed in a variety of tasks, e.g., sentiment clustering, sentiment classification, etc. As a sample application, we plug it into the ARSA model proposed in [20], which is used to predict sales performance based on reviews and past sales data. The original ARSA model uses S-PLSA as the component for capturing sentiment information. With the proposed S-PLSA model, the ARSA model can be formulated as follows:

$$y_t = \sum_{i=1}^{P} \varphi_i y_{t-i} + \sum_{i=1}^{q} \sum_{j=1}^{R} \rho_{i,j} \omega_{t-i,j} + \epsilon_t,$$ (11)

where

1) $y_t$ denotes the sales figure at time $t$ after proper preprocessing such as de-seasoning,
2) $p$, $q$, and $R$ are user-chosen parameters,
3) $\varphi_i$ and $\rho_{i,j}$ are coefficients to be estimated using training data, and
4) $\omega_{t,i,j} = \frac{1}{K} \sum_{k=1}^{K} \sum_{d \in R_t} p(z_j | d)$, where $R_t$ is the set of reviews available at time $t$ and $p(z_j | d)$ is computed based on S-PLSA. It reflects the sentiment “mass” that can be attributed to factor $z_j$.

The ARSA model can be trained using linear least squares regression. Note that the notion of time ($t$) in the ARSA model is different from the epoch ($n$) in S-PLSA. For example, sales prediction can be made for each day using ARSA, whereas the model adaptation of S-PLSA can happen every other day.

VI. EXPERIMENTS

Experiments were conducted on an IMDB dataset to evaluate the effectiveness of the proposed S-PLSA model, and the prediction power of ARSA using S-PLSA. The dataset was obtained from the IMDB Website and consists of two parts. Part 1, denoted by IMDB-REVIEW, is obtained by collecting 28,353 reviews for 20 drama films released in the US from May 1, 2006 to September 1, 2006, and Part 2, denoted by IMDB-BO contains the daily box office revenues of those films. For each review, we extracted the title, free text contents, time stamp, etc., and then indexed them using Apache Lucene.

A. Perplexity Evaluation of S-PLSA

We evaluate the effectiveness of the proposed S-PLSA model by computing its perplexity on the IMDB-REVIEW dataset. Perplexity is a commonly used measure of goodness for statistical language models. It is defined as the inverse of the probability of the test set as assigned by the language model, normalized by the number of words. Roughly speaking, it corresponds to the weighted average word branching factor of a language model. Lower perplexity indicates better modeling capability of the model on the given corpus (dataset). We varied the number of latent factors in the original S-PLSA and the proposed S-PLSA model, and compared their perplexities. As discussed in preceding sections, only appraisal words are employed to construct the feature vectors used in those models.

We first use the reviews from IMDB-REVIEW that correspond to 10 randomly chosen films to train an S-PLSA model. This model is then adapted using the proposed method in four epochs, with one-fourth of the remaining reviews used as adaptation reviews at each epoch. We perform 10-fold validation over training and adaptation sets.

Figure 1 shows the perplexities of the original S-PLSA model and the S-PLSA model at different adaptation epochs with the number of latent factors $K = 4, 8, 16$ respectively. It is clear from the graph that the perplexities are consistently reduced with incremental model adaptation as more adaptation data is introduced at each epoch. This testifies to the fact that model adaptation does help the sentiment modeling of the reviews. Increasing the number of hidden factors $K$ also has a positive effect on the perplexities of all models, with a consistent decrease of the perplexities when $K$ is increased from 4 to 8 and then to 16.

B. Effectiveness of S-PLSA when used for prediction

The effectiveness of S-PLSA for sales performance prediction is evaluated by replacing the S-PLSA component in the original ARSA model. Like in the previous perplexity experiments, reviews for half of the movies are used for

1http://www.imdb.com
2http://lucene.apache.org
batch training. For the original ARSA, the trained model is then used to make predictions in the testing data consisting of the other half the movies. For the proposed model, adaptation of the S-PLSA$^+$ component is performed for each movie in the testing set, in four epochs on four different days $v$ ($v = 2, 4, 6, 8$) using the review data available up to day $v$. The up-to-date model at day $v$ is then used for subsequent prediction tasks. We also compare the proposed incremental adaptation method with an alternative where the S-PLSA component in ARSA is completely re-trained from scratch using all data available including the original training data. This represents a batch adaptation approach.

The mean absolute percentage error (MAPE) is used to measure the prediction accuracy:

$$MAPE = \frac{1}{T} \sum_{i=1}^{T} \left( \frac{|Pred_i - True_i|}{True_i} \right),$$

where $T$ is the number of instances in the testing set, and $Pred_i$ and $True_i$ are the predicted value and the true value respectively.

Figure 2 shows the MAPE of the original ARSA with S-PLSA, the ARSA using S-PLSA$^+$ updated at Epochs 1-4 ($v = 2, 4, 6, 8$), and the ARSA with S-PLSA component completed re-trained at the four epochs. It is apparent from the figure that accuracy of the proposed model is much superior to that of the other two approaches. The accuracy of the model improves significantly as the it is getting updated in the first two epochs, which demonstrates the benefits of having an incremental model to absorb new information; especially in our case, S-PLSA$^+$ allows the models to be adapted to the individual movies. The rate of increase in accuracy get slower from Epoch 2 through Epoch 4, indicating that no significant new information is available from Epoch 2 to Epoch 4. The proposed model even outperforms the re-training approach where the S-PLSA is completed re-trained at each epoch. This is due to the fact that some information in the original training set may be out-of-date and not as relevant as the newly available reviews that focus more on the individual movies that we are making the prediction for. The proposed model can discount such out-of-date and irrelevant information, whereas the re-training approach cannot. Aside from the accuracy advantage, the proposed model also enjoys a much lower cost in terms of memory and computation.

VII. CONCLUSIONS AND FUTURE WORK

In this paper, we have presented an adaptive S-PLSA$^+$ model that is capable of incrementally updating its parameters and automatically downdating old information when new review data become available. This model has been used in conjunction with the ARSA model for predicting sales performance. Experimental results on a movie dataset show that by allowing the model to be adaptive, we can capture new sentiment factors arising from newly available reviews, which can greatly improve the its modeling capability as well as the accuracy when used for prediction. For future work, we plan to study the performance of S-PLSA$^+$ in other information retrieval and data mining tasks.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (60903108), the Program for New Century Excellent Talents in University (NCET-10-0532), and Discovery Grants from the Natural Sciences and Engineering Research Council of Canada (NSERC).

REFERENCES


