Opinion Integration Through Semi-supervised Topic Modeling

Yue Lu
Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, IL 61801
yuelu2@uiuc.edu

Chengxiang Zhai
Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, IL 61801
czhai@uiuc.edu

ABSTRACT

Web 2.0 technology has enabled more and more people to freely express their opinions on the Web, making the Web an extremely valuable source for mining user opinions about all kinds of topics. In this paper we study how to automatically integrate opinions expressed in a well-written expert review with lots of opinions scattering in various sources such as blogspaces and forums. We formally define this new integration problem and propose to use semi-supervised topic models to solve the problem in a principled way. Experiments on integrating opinions about two quite different topics (a product and a political figure) show that the proposed method is effective for both topics and can generate useful aligned integrated opinion summaries. The proposed method is quite general. It can be used to integrate a well written review with opinions in an arbitrary text collection about any topic to potentially support many interesting applications in multiple domains.

Categories and Subject Descriptors: B.3.3 [Information Search and Retrieval]: Text Mining

General Terms: Algorithms

Keywords: opinion integration, semi-supervised, probabilistic topic modeling, expert review

1. INTRODUCTION

As Web 2.0 applications become increasingly popular, more and more people express their opinions on the Web in various ways such as customer reviews, forums, discussion groups, and Weblogs. The wide coverage of topics and abundance of opinions make the Web an extremely valuable source for mining user opinions about all kinds of topics (e.g., products, political figures, etc.). However, with such a large scale of information source, it is quite challenging for a user to integrate and digest all the opinions from different sources.

In general, for any given topic (e.g., a product), there are often two kinds of opinions: the first is opinions expressed in some well-structured relatively complete review typically written by some expert about the topic and the second is fragmental opinions scattering around in all kinds of sources such as blog articles and forums. For convenience of discussion, we will refer to the first kind as *expert opinions* and the second *ordinary opinions*. The expert opinions are relatively easy for a user to access through some opinion search website such as CNET. Because a comprehensive product review is often written carefully, it is also easy for a user to digest expert opinions. However, finding, integrating, and digesting ordinary opinions pose significant challenges as they are scattering in many different sources, and are generally fragmental and not well structured. While expert opinions are clearly very useful, they may be biased and often out of date after a while. In contrast, ordinary opinions tend to represent the general opinions of a large number of people and get refreshed quickly as people dynamically generate new content. For example, a query "iPhone" returns 330,431 matches in Google's blogsearch (as of Nov. 1, 2007), suggesting that there are many opinions expressed about iPhone in blog articles within a short period of time since it hit the market. To enable a user to benefit from both kinds of opinions, it is thus necessary to automatically integrate these two kinds of opinions and present an integrated opinion summary to a user.

To the best of our knowledge, such an integration problem has not been studied in the existing work. In this paper, we study how to integrate a well-written expert review about an arbitrary topic with many ordinary opinions expressed in a text collection such as blog articles. We propose a general method to solve this integration problem in three steps: (1) extract ordinary opinions from text using information retrieval; (2) summarize and align the extracted opinions to the expert review to integrate the opinions; (3) further separate ordinary opinions that are similar to expert opinions from those that are not. Our main idea is to take advantage of the high readability of the expert review to structure the unorganized ordinary opinions while at the same time summarizing the ordinary opinions to extract *representative* opinions using the expert review as guidance. From the viewpoint of text data mining, we are essentially to use the expert review as a "template" to mine text data for ordinary opinions. The first step in our approach can be implemented with a direct application of information retrieval techniques. Implementing the second and third steps involves special challenges. In particular, without any training data, it is unclear how we should align ordinary opinions to an expert review and separate similar and supplementary opinions. We propose a semi-supervised topic modeling approach to solve these challenges. Specifically, we cast the expert review as a prior in a probabilistic topic model (i.e., PLSA[6]) and fit the model to the text collection with the ordinary opinions with Maximum A Posterior (MAP) estimation. With the estimated probabilistic model, we can then naturally obtain alignments of opinions as well as ad-

WWW 2008, April 21–25, 2008, Beijing, China.
ACM 978-1-60558-085-2/08/04.

ditional ordinary opinions that cannot be well-aligned with the expert review. The separation of similar and supplementary opinions can also be achieved with a similar model. We evaluate our method on integrating opinions about two quite different topics. One is a popular product "iPhone", and the other is a popular political figure Barack Obama. Experiment results show that our method can effectively integrate the expert review (a produce review from CNET for iPhone and a short biography from Wikipedia for Barack Obama) with ordinary opinions from blog articles.

This paper makes the following contributions:

1. We define a new problem of opinion integration. To the best of our knowledge, there is no existing work that solves this problem.
2. We propose a new semi-supervised topic modeling approach for integrating opinions scattered around in text articles with those in a well-written expert review for an arbitrary topic.
3. We evaluate the proposed method both qualitatively and quantitatively. The results show that our method is effective for integrating opinions about quite different topics.

Collecting and digesting opinions about a topic is critical for many tasks such as shopping, medical decision making, and social interactions. Our proposed method is quite general and can be applied to integrate opinions about any topic in any domain, thus potentially has many interesting applications.

The rest of the paper is organized as follows. In Section 2, we formally define the novel problem of opinion integration. After that, we present our Semi-supervised Topic Model in Section 4. We discuss our experiments and results in Section 5. Finally, we conclude in Section 7.

2. PROBLEM DEFINITION

In this section, we define the novel problem of opinion integration.

Given an expert review about a topic T (e.g., "iPhone" or "Barack Obama") and a collection of text articles (e.g., blog articles), our goal is to extract opinions from text articles and integrate them with those in the expert review to form an integrated opinion summary.

The expert review is generally well-written and coherent, thus we can view it as a sequence of semantically coherent segments, where a segment could be a sentence, a paragraph, or other meaningful segments (e.g., paragraphs corresponding to product features) available in some semi-structured review. Formally, we denote the expert review by $R = \{r_1, ..., r_k\}$ where r_i is a segment. Since we can always treat a sentence as a segment, this definition is quite general.

The text collection is a set of text documents where ordinary opinions are expressed and can be represented as $\mathcal{C} = \{d_1, ..., d_{|C|}\}$ where $d_i = (s_{i1}, ..., s_{i|d_i|})$ is a document and s_{ij} is a sentence. To support opinion integration in a general and robust manner, we do not rely on extra knowledge to segment documents to obtain opinion regions; instead, we treat each sentence as an opinion unit. Since a sentence has a well-defined meaning, this assumption is reasonable. To help a user interpret any opinion sentence, in real applications, we would link each extracted opinion sentence back to the original document to facilitate navigating into the original document and obtaining context of an opinion.

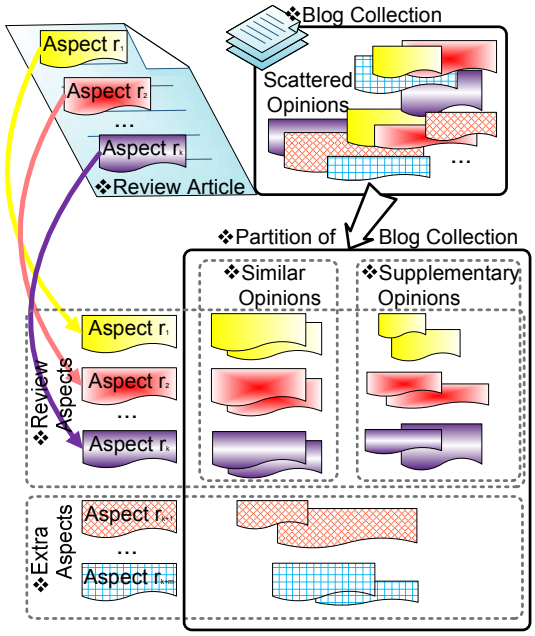

Figure 1: Problem Setup

We would like our integrated opinion summary to include both opinions in the expert review and those most representative opinions in the text collection. Since the expert review is well written, we keep their original form and leverage its structure to organize the ordinary opinions extracted from text. To quantify the representativeness of an ordinary opinion sentence, we will compute a "support value" for each extracted ordinary opinion sentence. Specifically, we would like to partition the extracted ordinary opinion sentences into groups that can be potentially aligned with all the review segments $r_1, ..., r_k$. Naturally, there may also be some groups with extra ordinary opinions that are not alignable with any expert opinion segment, and these opinions can be very useful to augment the expert review with additional opinions.

Furthermore, for opinions aligned to a review segment r_i, we would like to further separate those that are similar to r_i from those that are supplementary for r_i; such separation can allow a user to digest the integrated opinions more easily.

Finally, if r_i has multiple sentences, we can further align each ordinary opinion sentence (both "similar" and "supplementary") with a sentence in r_i to increase the readability.

This problem setup is illustrated in Figure 1. We now define the problem more formally.

Definition (Representative Opinion (RO)) A *representative opinion(RO)* is an ordinary opinion sentence extracted from the text collection with a support value. Formally, we denote it by $o_{ij} = (\beta, s_{ij})$ where $\beta \in [1, +\infty)$ is a support value indicating how many sentences this opinion sentence can represent, and s_{ij} is a sentence in document d_i.

Since ordinary opinions tend to be redundant and we are primarily interested in extracting representative opinions,

the support can be very useful to assess the representativeness of an extracted opinion.

Let $RO(\mathcal{C})$ be all the possible representative opinion sentences in $\mathcal{C}$. We can now define the integrated opinion summary that we would like to generate as follows.

Definition (Integrated Opinion Summary) An *integrated opinion summary* of R and $\mathcal{C}$ is a tuple $(R, S^{sim}, S^{supp}, S^{extra})$ where (1) R is the given expert review; (2) $S^{sim} = \{S_1^{sim}, ..., S_k^{sim}\}$ and $S^{supp} = \{S_1^{supp}, ..., S_k^{supp}\}$ are similar and supplementary representative opinion sentences, respectively, that can be aligned to R, and $S_i^{sim}, S_j^{supp} \subset RO(\mathcal{C})$ are sets of representative opinion sentences; (3) $S^{extra} \subset RO(\mathcal{C})$ is a set of extra representative opinion sentences that cannot be aligned with R.

Note that we define "opinion" broadly as covering all the discussion about a topic in opinionate sources such as blog spaces and forums. The notion of "opinion" is quite vague; we adopt this broad definition to ensure generality of the problem set up and its solutions. In addition, any existing sentiment analysis technique could be applied as a post-processing step. But since we only focus on the integration problem in this paper, we will not cover sentiment analysis.

3. OVERVIEW OF PROPOSED APPROACH

The opinion integration problem as defined in the previous section is quite different from any existing problem setup for opinion extraction and summarization, and it presents some special challenges: (1) How can we extract representative opinion sentences with support information? (2) How can we distinguish alignable opinions from non-alignable opinions? (3) For any given expert review segment, how can we distinguish similar opinions from those that are supplementary? (4) In the case when a review segment r_i has multiple sentences, how can we align a representative opinion to a sentence in r_i? In this section, we present our overall approach to solving all these challenges, leaving a detailed presentation to the next section.

At a high level, our approach primarily consists of two stages and an optional third stage: In the first stage, we retrieve only the relevant opinion sentences from $\mathcal{C}$ using the topic description T as a query. Let $\mathcal{C}_O$ be the set of all the retrieved relevant opinion sentences. In the second stage, we use probabilistic topic models to cluster sentences in $\mathcal{C}_O$ and obtain S^{sim}, S^{supp} and S^{extra}. When r_i has multiple sentences, we have a third stage, in which we again use information retrieval techniques to align any extracted representative opinion to a sentence of r_i. We now describe each of the three stages in detail.

The purpose of the first stage is to filter out irrelevant sentences and opinions in our collection. This can be done by using the topic description as a keyword query to retrieve relevant opinion sentences. In general, we may use any retrieval method. In this paper, we used a standard language modeling approach (i.e., the KL-divergence retrieval model [20]). To ensure coverage of opinions, we perform pseudo feedback using some top-ranked sentences; the idea is to expand the original topic description query with additional words related to the topic so that we can further retrieve opinion sentences that do not necessarily match the original topic description T. After this retrieval stage, we obtain a set of relevant opinion sentences $\mathcal{C}_O$.

In the second stage, our main idea is to exploit a probabilistic topic model, i.e., Probabilistic Latent Semantic Analysis (PLSA) with conjugate prior [6, 11] to cluster opinion sentences in a special way so that there will be precisely one cluster corresponding to each segment r_i in the expert review. These clusters are to collect opinion sentences that can be aligned with a review segment. There will also be some clusters that are not aligned with any review segments, and they are designed to collect extra opinions. Thus the model provides an elegant way to simultaneously partition opinions and align them to the expert review. Interestingly, the same model can also be adapted to further partition opinions aligned to a review segment into similar and supplementary opinions. Finally, a simplified version of the model (i.e., no prior, basic PLSA) can be used to cluster any group of sentences to extract representative opinion sentences. The support of a representative opinion is defined as the size of the cluster represented by the opinion sentences.

Note that what we need in this second stage is semi-supervised clustering in the sense that we would like to constrain many of the clusters so that they would correspond to the segments r_is in the expert review. Thus a direct application of any regular clustering algorithm would not be able to solve our problem. Instead of doing clustering, we can also imagine using each expert review segment r_i as a query to retrieve similar sentences. However, it would be unclear how to choose a good cutoff point on the ranked list of retrieved results. Compared with these alternative approaches, PLSA with conjugate prior provides a more principled and unified way to tackle all the challenges.

In the optional third stage, we have a review segment r_i with multiple sentences and we would like to align all extracted representative opinions to the sentences in r_i. This can be achieved by using each representative opinion as a query and retrieve sentences in r_i. Once again, in general, any retrieval method can be used. In this paper, we again used the KL-divergence retrieval method.

From the discussion above, it is clear that we leverage both information retrieval techniques and text mining techniques (i.e., PLSA), and our main technical contributions lie in the second stage where we repeatedly exploit semi-supervised topic modeling to extract and integrate opinions. We describe this step in more detail in the next section.

4. SEMI-SUPERVISED PLSA FOR OPINION INTEGRATION

Probabilistic latent semantic analysis (PLSA) [6] and its extensions [21, 13, 11] have recently been applied to many text mining problems with promising results. Our work adds to this line yet another novel use of such models for opinion integration.

As in most topic models, our general idea is to use a unigram language model (i.e., a multinomial word distribution) to model a topic. For example, a distribution that assigns high probabilities to words such as "iPhone", "battery", "life", "hour", would suggest a topic such as "battery life of iPhone." In order to identify multiple topics in text, we would fit a mixture model involving multiple multinomial distributions to our text data and try to figure out how to set the parameters of the multiple word distributions so that we can maximize the likelihood of the text data. Intuitively, if two words tend to co-occur with each other and one word is

assigned a high probability, then the other word generally should also be assigned a high probability to maximize the data likelihood. Thus this kind of model generally captures the co-occurrences of words and can help cluster the words based on co-occurrences.

In order to apply this kind of model to our integration problem, we assume that each review segment corresponds to a unigram language model which would capture all opinions that can be aligned with a review segment. Furthermore, we introduce a certain number of unigram language models to capture the extra opinions. We then fit the mixture model to $\mathcal{C}_O$, i.e., the set of all the relevant opinion sentences generated using information retrieval as described in the previous section. Once the parameters are estimated, they can be used to group sentences into different aspects corresponding to the different review segments and extra aspects corresponding to extra opinions. We now present our mixture model in detail.

4.1 Basic PLSA

We first present the basic PLSA model as described in [21]. Intuitively, the words in our text collection $\mathcal{C}_O$ can be classified into two categories (1) background words that are of relatively high frequency in the whole collection. For example, in the collection of topic "iPhone", words like "iPhone", "Apple" are considered as background words. (2) words related to different aspects which we are interested in. So we define $k+1$ unigram language models: θ_B as the background model to capture the background words, $\Theta = \{\theta_1, \theta_2, ..., \theta_k\}$ as k theme models, each capturing one aspect of the topic and corresponding to the k review segments $r_1, ..., r_k$. A document d in $\mathcal{C}_O$ (in our problem it is actually a sentence) can then be regarded as a sample of the following mixture model.

$$p_d(w) = \lambda_B p(w|\theta_B) + (1-\lambda_B)\sum_{j=1}^{k}[\pi_{d,j} p(w|\theta_j)] \quad (1)$$

where w is a word, $\pi_{d,j}$ is a document-specific mixing weight for the j-th aspect ($\sum_{j=1}^{k}\pi_{d,j} = 1$), and λ_B is the mixing weight of the background model θ_B. The log-likelihood of the collection $\mathcal{C}_O$ is

$$\log p(\mathcal{C}_O|\Lambda) = \sum_{d\in\mathcal{C}_O}\sum_{w\in V}\{c(w,d)\times \log(\lambda_B p(w|\theta_B) + (1-\lambda_B)\sum_{j=1}^{k}[\pi_{d,j}p(w|\theta_j)]\} \quad (2)$$

where V is the set of all the words (i.e., vocabulary), $c(w,d)$ is the count of word w in document d, and Λ is the set of all model parameters. The purpose of using a background model is to "force" clustering to be done based on more discriminative words, leading to more informative and more discriminative theme models.

The model can be estimated using any estimator. For example, the Expectation-Maximization (EM) algorithm [3] can be used to compute a maximum likelihood estimate with the following updating formulas:

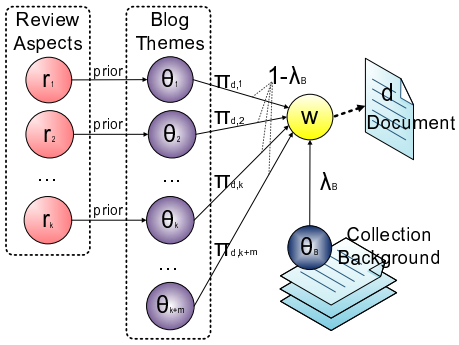

Figure 2: Generation Process of a Word

$$p(z_{d,w,j}) = \frac{(1-\lambda_B)\pi^{(n)}_{d,j}p^{(n)}(w|\theta_j)}{\lambda_B p(w|\theta_B) + (1-\lambda_B)\sum_{j'=1}^{k}\pi^{(n)}_{d,j'}p^{(n)}(w|\theta'_j)}$$

$$p(z_{d,w,B}) = \frac{\lambda_B p(w|\theta_B)}{\lambda_B p(w|\theta_B) + (1-\lambda_B)\sum_{j'=1}^{k}\pi^{(n)}_{d,j'}p^{(n)}(w|\theta'_j)}$$

$$\pi^{(n+1)}_{d,j} = \frac{\sum_{w\in V} c(w,d)p(z_{d,w,j})}{\sum_{j'}\sum_{w\in V} c(w,d)p(z_{d,w,j'})}$$

$$p^{(n+1)}(w|\theta_j) = \frac{\sum_{d\in\mathcal{C}_O} c(w,d)p(z_{d,w,j})}{\sum_{w'\in V}\sum_{d\in\mathcal{C}_O} c(w,d)p(z_{d,w',j})}$$

4.2 Semi-supervised PLSA

We could have directly applied the basic PLSA to extract topics from $\mathcal{C}_O$. However, the extracted topics in this way would generally not be well-aligned to the expert review. In order to ensure alignment, we would like to "force" some of the multinomial distribution component models (i.e., language models) to be "aligned" with all the segments in the expert review. In probabilistic models, this can be achieved by extending the basic PLSA to incorporate a conjugate prior defined based on the expert review segments and using the Maximum A Posterior (MAP) esatimator instead of the Maximum Likelihood estimator as we did in the basic PLSA. Intuitively, a prior defined based on an expert review segment would tend to make the corresponding language model similar to the empirical word distribution in the review segment, thus the language model would tend to attract opinion sentences in $\mathcal{C}_O$ that are similar to the expert review segment. This ensures the alignment of the extracted opinions with the original review segment.

Specifically, we build a unigram language model $\{p(w|r_j)\}_{w\in V}$ for each review segment r_j ($j \in \{1, ..., k\}$) and define a conjugate prior (i.e., a Dirichlet prior) on each multinomial distribution topic model, parameterized as $Dir(\{\sigma_j p(w|r_j)\}w \in V)$, where σ_j is a confidence parameter for the prior. Since we use a conjugate prior, σ_j can be interpreted as the "equivalent sample size" which means that the effect of adding the prior would be equivalent to adding $\sigma_j p(w|r_j)$pseudo counts for word w when we estimate the topic model $p(w|\theta_j)$. Figure 2 illustrates the generation process of a word W in such a semi-supervised PLSA where the prior serves as some "training data" to bias the clustering results.

The prior for all the parameters is given by

$$p(\Lambda) \propto \prod_{j=1}^{k+m} \prod_{w \in V} p(w|\theta_j)^{\sigma_j p(w|r_j)} \quad (3)$$

Generally we have $m > 0$, because we may want to find extra opinion topics other than the corresponding segments in the expert review. So we set $\sigma_j = 0$ for $k < j \leq k + m$.

With the prior defined above, we can then use the Maximum A Posterior (MAP) estimator to estimate all the parameters as follows

$$\hat{\Lambda} = \arg\max_{\Lambda} p(\mathcal{C}_O|\Lambda)p(\Lambda) \quad (4)$$

The MAP estimate can be computed using essentially the same EM algorithm as presented above with only slightly different updating formula for the component language models. The new updating formula is:

$$p(w|\theta_j)^{(n+1)} = \frac{\sum_{d \in \mathcal{C}_O} c(w,d)p(z_{d,w,j}) + \sigma_j p(w|r_j)}{\sum_{w' \in V} \sum_{d' \in \mathcal{C}_O} c(w',d')p(z_{d',w',j}) + \sigma_j} \quad (5)$$

We can see that the main difference between this equation and the previous one for basic PLSA is that we now pool the counts of terms in the expert review segment with those from the opinion sentences in $\mathcal{C}_O$, which is essentially to allow the expert review to serve as some training data for the corresponding opinion topic. This is why we call this model semi-supervised PLSA.

If we are highly confident of the aspects captured in the prior, we could empirically set a large σ_j. Otherwise, if we need to ensure the impact of the prior without being over restricted by the prior, some regularized estimation techniques are necessary. Following the similar idea of regularized estimation [19], we define a decay parameter η and a prior weight μ_j as

$$\mu_j = \frac{\sigma_j}{\sum_{w' \in V} \sum_{d' \in \mathcal{C}_O} c(w',d')p(z_{d',w',j}) + \sigma_j} \quad (6)$$

So we could start from a large σ_j (say 5000) (i.e., starting with perfectly alignable opinion models) and gradually decay it in each EM iteration by equation 7, and we stop the decaying of σ_j until the weight of the prior μ_j is below some threshold δ (say 0.5). Decaying allows the model to gradually pick up words from $\mathcal{C}_O$. The new updating formulas are

$$\sigma_j^{(n+1)} = \begin{cases} \eta\sigma_j^{(n)} & \texttt{if } \mu_j > \delta \\ \sigma_j^{(n)} & \texttt{if } \mu_j \leq \delta \end{cases} \quad (7)$$

$$p(w|\theta_j)^{(n+1)} = \frac{\sum_{d \in \mathcal{C}_O} c(w,d)p(z_{d,w,j}) + \sigma_j^{(n+1)} p(w|r_j)}{\sum_{w' \in V} \sum_{d' \in \mathcal{C}_O} c(w',d')p(z_{d',w',j}) + \sigma_j^{(n+1)}} \quad (8)$$

4.3 Overall Process

In this section, we describe how we use the semi-supervised topic model to achieve three tasks in the second stage as defined in Section 3. We also summarize the computational complexity of the whole process.

4.3.1 Theme Extraction from Text Collection

We start from a topic T, a review $R = \{r_1, ..., r_k\}$ of k segments, a collection $\mathcal{C}_O = \{d_1, d_2, ..., d_N\}$ of opinion sentences closely relevant to T. We assume that $\mathcal{C}_O$ covers a number of themes each about one aspect of the topic T. We further assume that there are $k + m$ major themes in the collection, $\{\theta_1, \theta_2, ..., \theta_{k+m}\}$, each being characterized by a multinomial distribution over all the words in our vocabulary V (also known as a unigram language model or a topic model).

We propose to use review aspects as priors in the partition of $\mathcal{C}_O$ into aspects. We could have used the whole expert review segment to construct the priors. But if so, we could only get the opinions that are most similar to the review opinions. However, we would like extract not only opinions supporting the review opinions but also supplementary opinions on the review aspect. So we use only the "aspect words" to estimate the prior. We use a simple heuristic: opinions are usually expressed in the form of adjectives, adverbs and verbs while aspect words are usually nouns. And we apply a Part-of-Speech tagger[1] on each review segment r_i and further filter out the opinion words to get a r_i'. The prior $\{p(w|r_i')\}_{w \in V}$ is estimated by Maximum Likelihood:

$$p(w|r_i') = \frac{c(w,r_i')}{\sum_{w' \in V} c(w',r_i')} \quad (9)$$

Given these priors constructed from the expert review $\{p(w|r_i')\}_{w \in V}$, $i \in \{1, ..., k\}$, we could estimate the parameters for the semi-supervised topic model according to Section 4.2. After that, we have a set of theme models extracted from the text collection $\{\theta_i | i = 1, ..., k + m\}$, and we could group each sentence d_i in $\mathcal{C}_O$ into one of the $k + m$ themes by choosing the theme model with the largest probability of generating d_i:

$$\arg\max_j p(d_i|\theta_j) = \arg\max_j \sum_{w \in V} c(w,d_i)p(w|\theta_j) \quad (10)$$

If we define $g(d_i) = j$ if d_i is grouped into $\{p(w|\theta_j)\}_{w \in V}$, then we have a partition of $\mathcal{C}_O$:

$$\mathcal{C}_O = \{S_i | i = 1, ..., k + m\} \quad (11)$$

where each S_i is a set of sentences $S_i = \{d_j | g(d_j) = i, d_j \in \mathcal{C}_O\}$ with the following two properties:

$$\mathcal{C}_O = \bigcup_{i=1}^{k+m} S_i \quad (12)$$

$$S_i \cap S_j = \emptyset \qquad \forall i, j \in \{1, ..., k + m\}, i \neq j \quad (13)$$

Thus each S_i, $i = 1, ..., k$, corresponds to the review aspect r_i and each S_j, $j = k + 1, ..., k + m$, is the set of sentences that supplements the expert review with additional aspects. Parameter m, the number of additional aspects, is set empirically.

4.3.2 Further Separation of Opinions

In this subsection, we show that how we further partition each S_i, $i = 1, ..., k$ into two parts:

$$S_i = \{S_i^{sim}, S_i^{supp}\} \quad (14)$$

[1] http://l2r.cs.uiuc.edu/~cogcomp/asoftware.php?skey=LBPPOS

such that S_i^{sim} contains sentences that is similar to the opinions in the review while S_i^{supp} is a set of sentences that supplement the review opinions on the review aspect r_i.

We assume that each subset of sentences S_i, $i = 1, ..., k$, covers two themes captured by two subtopic models $\{p(w|\theta_i^{sim})\}_{w \in V}$ and $\{p(w|\theta_i^{supp})\}_{w \in V}$. We first construct a unigram language model $\{p(w|r_i)\}_{w \in V}$ from review segment r_i using both the feature words and opinion words. This model is used as a prior for extracting $\{p(w|\theta_i^{sim})\}_{w \in V}$. After that, we estimate the model parameters as described in Section 4.2. And then, we could classify each sentence $d_j \in S_i$ into either S_i^{sim} or S_i^{supp} in the way similar to equation 10.

4.3.3 Generation of Summaries

So far, we have a meaningful partition over $\mathcal{C}_O$:

$$\mathcal{C}_O = \{S_1^{sim}, ..., S_k^{sim}\} \cup \{S_1^{supp}, ..., S_k^{supp}\} \cup \{S_{k+1}, ..., S_{k+m}\} \quad (15)$$

Now we need to further summarize each block P in the partition $P \in \{S_1^{sim}, ..., S_k^{sim}\} \cup \{S_1^{supp}, ..., S_k^{supp}\} \cup \{S_{k+1}, ..., S_{k+m}\}$ by extracting representative opinions $RO(P)$. We take a two-step approach.

In the first step, we try to remove the redundancy of sentences in P and group the similar opinions together by unsupervised topic modeling. In detail, we use PLSA (without any prior) to do the clustering and set the number of clusters proportional to the size of P. After the clustering, we get a further partition of $P = \{P_1, ..., P_l\}$ where $l = |P|/c$ and c is a constant parameter that defines the average number of sentences in each cluster. One representative sentence in P_i is selected by the similarity between the sentence and the cluster centroid (i.e. a word distribution) of P_i. If we define rs_i as the representative sentenced of P_i, and $\beta_i = |P_i|$ as the support, we have a representative opinion of P_i which is $o_i = (\beta_i, rs_i)$. Thus $RO(P) = \{o_1, o_2, ..., o_l\}$.

In the second step, we aim at providing some context information for each representative opinion o_i of P to help the user to better understand the opinion expressed. What we propose is to compare the similarity between opinion sentence rs_i and each review sentence in segment corresponding to P and assign rs_i to the review sentence with the highest similarity. For both steps, we use KL-Divergence as the similarity measure.

4.3.4 Computational Complexity

PLSA and semi-supervised PLSA have the same complexity: $O(I \cdot K(|V| + |W| + |\mathcal{C}|))$, where I is the number of EM iterations, K is the number of themes, $|V|$ is the vocabulary size, $|W|$ is the total number of words in the collection, $|\mathcal{C}|$ is the number of documents. Our whole process makes multiple invocations of PLSA/semi-supervised PLSA, and we suppose we use the same I across different invocations.

"Theme Extraction from Text Collection" makes one invocation of semi-supervised PLSA on the whole collection $\mathcal{C}_O$, where the number of cluster is $k + m$. So the complexity is $O(I \cdot (k + m) \cdot (|V| + |W| + |\mathcal{C}_O|) = O(I \cdot (k + m) \cdot |W|)$.

There are k invocations of semi-supervised PLSA in "Further Separation of Opinions", each on a subset of the collection $S_i (i = 1, ..., k)$ with only two clusters. And we know from equation 11 that $\bigcup_{i=1}^{k} S_i \subseteq \bigcup_{i=1}^{k+m} S_i = \mathcal{C}_O$. Suppose W_{S_i} is the total number of words in S_i. So the total complexity is $O(\sum_{S_i} I \cdot 2 \cdot (|V| + |W_{S_i}| + |S_i|))$ which in the worst case is $O(I \cdot 2 \cdot (k|V| + |W| + |\mathcal{C}_O|)) = O(I \cdot (k|V| + |W|))$.

Finally, "Generation of Summaries" makes $2k + m$ invocations of PLSA, each on a subset of the collection $P \in \{S_1^{sim}, ..., S_k^{sim}\} \cup \{S_1^{supp}, ..., S_k^{supp}\} \cup \{S_{k+1}, ..., S_{k+m}\} = \mathcal{C}_O$. In each invocation, the number of clusters is $\frac{|P|}{c}$, and W_P is the total number of words in P. So the total complexity in this stage is $O(\sum_P I \cdot \frac{|P|}{c}(|V| + |W_P| + |P|))$, which in the worst case is $O(\frac{I}{c} \cdot (|\mathcal{C}_O| \cdot |V| + |\mathcal{C}_O| \cdot |W| + |\mathcal{C}_O|^2)) = O(\frac{I}{c} \cdot |\mathcal{C}_O| \cdot |W|)$.

Thus, our whole process is bounded by the computational complexity $O(I \cdot ((k + m + 1)|W| + k|V| + \frac{|\mathcal{C}_O| \cdot |W|}{c}))$. Since k, m, and c are usually much smaller than $|\mathcal{C}_O|$, the running time is basically bounded by $O(I \cdot |\mathcal{C}_O| \cdot |W|)$.

5. EXPERIMENTAL RESULTS

In this section, we first introduce the data sets used in the experiment. Then we demonstrate the effectiveness of our semi-supervised topic modeling approach by showing two examples in two different scenarios. Finally, we also provide some quantitative evaluation.

5.1 Data Sets

Topic Desc.	Source	# of words	# of aspects
iPhone	CNET	4434	19
Barack Obama	wikipedia	312	14

Table 1: Basic Statistics of the REVIEW data set

Topic Desc.	Query Terms	# of articles	N
iPhone	iPhone	552	3000
Barack Obama	Barack+Obama	639	1000

Table 2: Basic Statistics of the BLOG data set

We need two types of data sets for evaluation. One type is expert reviews. We construct this data set by leveraging the existing services provided by CNET and wikipedia, i.e., we submit queries to their web sites and download the expert reviews on "iPhone" written by CNET editors[2] and the introduction part of articles about "Barack Obama" in wikipedia[3]. The composition and basic statistics of this data set (denoted as "REVIEW") is shown in Table 1.

The other type of data is a set of opinion sentences related to certain topic. In this paper, we only use Weblog data, but our method can be applied on any kind of data that contain opinions in free text. Specifically, we firstly submit topic description queries to Google Blog Search[4] and collect the blog entries returned. The search domain are restricted to spaces.live.com, since schema matching is not our focus. We further build a collection of N opinion sentences $\mathcal{C}_O = \{d_1, d_2, ..., d_N\}$ which are highly relevant to the given topic T using information retrieval techniques as described as the first stage in Section 3. The basic information of these collections (denoted as "BLOG" is shown in Table 2. For all the data collections, Porter stemmer [18] is used to stem the text and stop words in general English are removed.

[2]http://reviews.cnet.com/smart-phones/apple-iPhone-8gb-at/4505-6452_7-32309245.html?tag=pdtl-list
[3]http://en.wikipedia.org/wiki/Barack_Obama
[4]http://blogsearch.google.com

5.2 Scenario I: Product

Gathering opinions on products is the main focus of the research on opinion mining, so our first example of opinion integration is a hot product, iPhone. There are 19 defined segments in the "iPhone" review of the REVIEW data set. We use these 19 segments as aspects from the review and define 11 extra aspects in the semi-supervised topic model.

Due to the limitation of the spaces, only part of the integration with review aspects are show in Table 3. We can see that there is indeed some interesting information discovered.

- In the "background" aspect (which corresponds to the background introduction part of the expert review), we see that lots of people care about the price of iPhone, and the sentences extracted from blog articles show different pricing information which confirms the fact that the price of iPhone has been adjusted. In fact, the first two sentences only mention the original price while the third sentence talks about the cut down of the price but the actual numbers are incorrect.

- The displayed sentence in the "activation" aspect describes the results if you do not activate the iPhone. A piece of very interesting information related to this aspect, "unlocking the iPhone" is never mentioned in the expert review but is extracted from blog articles by using our semi-supervised topic modeling approach. Indeed, we know that "unlock" or "hack" is a hot topic since the iPhone hit the market. This is a good demonstration that our approach is able to discover information which is highly related and supplementary to the review.

- The last aspect shown is about battery life. There is a high support ($support = 19$ in the column of similar opinions) of the life of battery described in the review, and there is another supplementary set of sentences ($support = 7$) which gives a concrete number of battery in hours under real usage of iPhone.

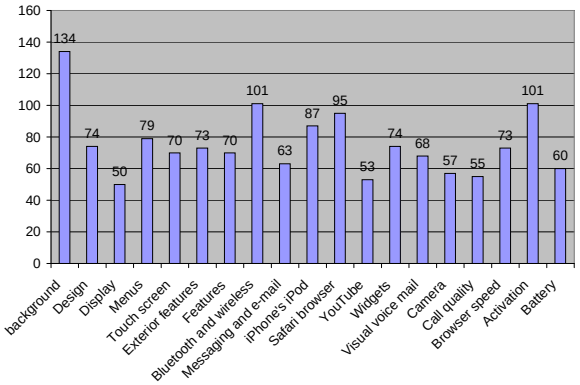

Figure 3: Support Statistics for iPhone Aspects

Furthermore, we may also want to know which aspects of iPhone people are most interested in. If we define the support of an aspect as the sum of the support of representative opinions in this aspect, we could easily get the support statistics for each review aspects in our topic modeling approach. As can be seen in Figure 3, the "background" aspect attracts the most discussion. This is mainly caused by the mention of the price of iPhone in the background aspect. The next two aspects with highest support are "Bluetooth and Wireless" and "Activation" both with support 101. As stated in the iPhone review "The Wi-Fi compatibility is especially welcome, and a feature that's absent on far too many smart phones.", and our support statistics suggest that people do comment a lot about this unique feature of iPhone. "Activation" is another hot aspect as discovered by our method. As many people know, the activation of iPhone requires a two-year contract with AT&T, which brings much controversy among customers.

In addition, we show three of the most supported representative opinions in the extra aspects in Table 4. The first sentence points out another way of activating iPhone, while the second sentence brings up the information that Cisco was the original owner of the trademark "iPhone". The third sentence expresses a opinion in favor of another smartphone, Nokia N95, which could be useful information for a potential smartphone buyer who did not know about Nokia N95 before.

5.3 Scenario II: Political Figure

If we want to know more about a political figure, we could treat a short biography of the person as an expert review and apply our semi-supervised topic model. In this subsection, we demonstrate what we can achieve by an example of "Barack Obama". There is no definition of segments in the short introduction part in wikipedia, so we just treat each sentence as a segment.

In Table 5, we display part of the opinion integration with the 14 aspects in the review. Since there is no short description of each aspect in this example, we use ID in the first column of the table to distinguish one aspect from another.

- Aspect 0 is a brief introduction of the person and his position, which attracts many sentences in the blog articles some directly confirming the information provided in the review, some also suggest his position while stating other facts.

- Aspect 1 and 3 talk about his heritage and early life, and we further discover from the blog articles supplementary information such as his birthplace is Honolulu, his parents' names are Barack Hussein Obama Sr. and Ann Dunham, and even why his father came to the US.

- For aspect 10 about his presidential candidacy, our summaries not only confirm the fact but also point out another democratic presidential candidate Hillary Clinton.

- A brief description of his family is in review aspect 12, and the mention of his daughters has attracted a piece of news related to young daughters of White House aspirants.

After further summing up the support for each aspect, we display two of the most supported aspects and one least supported aspect in Table 6. The most supported aspect is aspect 0 with $Support = 68$, which as mentioned above is a brief introduction of the person and his position. Aspect 2 talking about his heritage ranks as the second with $Support = 36$, which agrees with the fact that he is special among the presidential candidates because of his Kenyan

Aspect	Review	Similar Opinions	Supplementary Opinions
Background	Even with the new $399 price for the 8GB model (down from an original price of $599), it's still a lot to ask for a phone that lacks so many features and locks you into an iPhone-specific two-year contract with AT&T.		[**support=19**]The iPhone will come in two versions, a 4GB 499 model, and an 8GB 599 model with a two year contract.
			[**support=16**]The Price: 499 (4GB) or 599(8GB) with a two year contract , by the time the contract is over your iPhone will probably be scratched all over like the Nano or be made obsolete by better phone on the market.
			[**support=12**]Recently, Apple decided to cut down price of iPhone from 399 to 200 , giving rise to much rage from consumers bought the phone before.
Activation	You can make emergency calls, but you can't use any other functions, including the iPod music player.		[**support=10**]Several other methods for unlocking the iPhone have emerged on the Internet in the past few weeks, although they involve tinkering with the iPhone hardware or more complicated ways of bypassing the protections for AT T's exclusivity.
Battery	Battery life The Apple iPhone has a rated battery life of 8 hours talk time, 24 hours of music playback, 7 hours of video playback, and 6 hours on Internet use.	[**support=19**] iPhone will Feature Up to 8 Hours of Talk Time, 6 Hours of Internet Use, 7 Hours of Video Playback or 24 Hours of Audio Playback	[**support=7**]Playing relatively high bitrate VGA H.264 videos, our iPhone lasted almost exactly 9 freaking hours of continuous playback with cell and WiFi on (but Bluetooth off).

Table 3: iPhone Example: Opinion Integration with Review Aspects

Supplementry Opinions on Extra Aspects
[**support=15**]You may have heard of iASign (http: iphone.fiveforty.net wiki index.php IASign), an iPhone Dev Wiki tool that allows you to activate your phone without going through the iTunes rigamarole.
[**support=13**]Cisco has owned the trademark on the name "iPhone" since 2000, when it acquired InfoGear Technology Corp., which originally registered the name.
[**support=13**]With the imminent availability of Apple's uber cool iPhone, a look at 10 things current smartphones like the Nokia N95 have been able to do for a while and that the iPhone can't currently match...

Table 4: iPhone Example: Opinion Integration on Extra Aspects

ID	Review	Support
0	Barack Hussein Obama (born August 4, 1961) is the junior United States Senator from Illinois and a member of the Democratic Party.	68
1	Born to a Kenyan father and an American mother, Obama grew up in culturally diverse surroundings.	36
12	He married in 1992 and has two daughters.	3

Table 6: Obama Example: Support of Aspects

origin and indicates that people are interested in it. The least covered aspect is aspect 12 about his family, since the total support is only 3.

5.4 Quantitative Evaluation

In order to quantitatively evaluate the effectiveness of our semi-supervised topic modeling approach, we designed a test which consists of three tasks, each asks a user to perform a part of our processing. The main goal is to see to what extent can our approach reproduce the human choice. The test is designed based on the above-mentioned "Barack Obama" example. In order to reduce the bias, we collect the evaluation results from three users, who are all PhD students in our department, two males and one female.

In the first designed task, we aims at evaluating the effectiveness of our approach in identifying the extra aspects in addition to review aspects. Towards this goal, we generate a big set of sentences S_{all} by mixing all the sentences in $\{S_1^{sim}, ..., S_k^{sim}\} \cup \{S_1^{supp}, ..., S_k^{supp}\}$ with seven most supported sentences in $\{S_{k+1}, ..., S_{k+m}\}$. There are $|S_{all}| = 34$ sentences in S_{all} in total. The users are asked to select seven sentences from randomly permutated S_{all} that do not fit into the k review aspects. In this way, we could see how is the human consensus on this task and how our approach could recover the choice of human.

User	Sentence ID of the 7 sentences
Our Approach	2, 6, 9, 21, 22, 25, 30
User 1	1, 6, 9, 13, 16, 25, 30
User 2	9, 11, 16, 20, 21, 30, 31
User 3	2, 6, 8, 9, 24, 25, 31

Table 7: Selection of 7 Sentences on Extra Aspects

Table 7 displays the selection of the seven sentences on extra aspects by our method and the three users. The only sentence out of seven that all three users agree on is sentence number 9, which suggests that grouping sentences into extra aspects is quite a subjective task so it is difficult to produce results satisfactory to each individual user. However our method is able to recover 52.4% of the user's choices on average.

In the second task, we try to evaluate the performance of our approach in grouping sentences into k review aspects. we randomly permutate all the sentences in $\{S_1^{sim}, ..., S_k^{sim}\} \cup \{S_1^{supp}, ..., S_k^{supp}\}$ to construct a S_{review} and remove the aspect assigned to each sentence. For each of the 27 sentences, the users are asked to assign one of the 14 review aspects to it. In essence, this is a multi-class classification problem where the number of classes is 14.

The results turn out to be

- Three users agree on 13 sentences about the class label, which means that more than half of the sentences are controversial even among human users.
- On average, our method could recover the user's choices by 10.67 sentences out of 27. Note that if we randomly

ID	Review	Similar Opinions	Supplementary Opinions
0	Barack Hussein Obama (born August 4, 1961) is the junior United States Senator from Illinois and a member of the Democratic Party.	[**support=9**]Senator Barack Hussein Obama is the junior United States Senator from Illinois and a member of the Democratic Party .	[**support=21**]Barack Obama, another leading Democratic presidential hopeful, campaigns for more dollars with "Dinner With Barack." [**support=11**]A Chicago, Illinois, radio station recently conducted a live survey on a man called Barack Obama. [**support=10**]In fact, there is not a single metropolitan area in the country where a family earning minimum wage can afford decent housing, said Senator Barack Obama.
1	The U.S. Senate Historical Office lists him as the fifth African American Senator in U.S. history and the only African American currently serving in the U.S. Senate.		[**support=16**]Barack Obama is an African American whose father was born in Kenya and got a sholarship to study in American.
3	He lived for most of his childhood in the majority-minority U.S. state of Hawaii and spent four of his pre-teen years in the multi-ethnic Indonesian capital city of Jakarta.		[**support=12**]Obama was born in Honolulu, Hawaii, to Barack Hussein Obama Sr., a Kenyan, and Kansas born Ann Dunham.
10	He is among the Democratic Party's leading candidates for nomination in the 2008 U.S. presidential election.	[**support=2**]Mr Obama will contest the Democrat presidential nomination	[**support=14**](AP) Democratic presidential candidate Barack Obama said Sunday that the front runner for his party's nomination, Hillary Rodham Clinton, does not offer the break from politics as usual that voters need.
12	He married in 1992 and has two daughters.		[**support=3**]MARCH 4 Senator Barack Obama is threatening legal action against a self describ--ed pedophile who has posted photos of the Democratic politician's young daughters on a web site that purports to handicap the 2008 presidential campaign by evaluating the "cuteness" of underage daughters and granddaughters of White House aspirants

Table 5: Obama Example: Opinion Integration with Review Aspects

assign one aspect out of 14, (1) the probability of recovering k sentences out of 27 is

$$\binom{27}{k} \times pr^k \times (1-pr)^{27-k}$$

where $pr = \frac{1}{14}$. When $k = 10$, the probability is only around 0.00037; (2) the expected number of sentences recovered would be

$$\sum_{k=0}^{27} \binom{27}{k} \times pr^k \times (1-pr)^{27-k} = 1$$

- Our method and all three users assigned the same label to 8 sentences.
- Among the many mistakes our method made, three users only agree on 5 sentences. In other words, they assigned the same label to the 5 sentences which is different the label assigned by our method.

Again, this task is subjective, and there is still much controversy among human users. But our approach performs reasonably : in the 13 sentences with human consensus, our method achieves the accuracy of 61.5%.

In the third task, our goal is to see how well we can separate similar opinions from supplementary opinions in the semi-supervised topic modeling approach. We first select 5 review aspects out of 14 which our method has identified both similar and supplementary opinions; then for each of the 5 aspects, we mix one similar opinion with several supplementary opinions; the users are supposed to select one sentence which share the most similar opinion with the review aspect. On average, our method could recover 60% of the choices of human users. Among the different choices between our method and the users, only one aspect has achieved consensus of three users. That is to say, this is a "true" mistake of our method, while other mistakes do not have agreement in the users.

6. RELATED WORK

To the best of our knowledge, no previous study has addressed the problem of integrating a well-written expert review with opinions scattering in text documents. But there are some related studies which we will briefly review in this section.

Recently there has been a lot of work in opinion mining and summarization especially on customer reviews. In [2], sentiment classifiers are built from some training corpus. Some papers [8, 7, 10, 17] further mine product features from reviews on which the reviewers have expressed their opinions. Zhuang and others focused on movie review mining and summarization [22]. [4] presented a prototype system, named Pulse, for mining topics and sentiment orientation jointly from customer feedback. However, these techniques are limited to the domain of products/movies, and many are highly dependent on the training data set, so are not generally applicable to summarize opinions about an arbitrary topic. Our problem setup aims at shallower but more robust integration.

Weblogs mining has attracted many new research work. Some focus on sentiment analysis. Mishne and others used the temporal pattern of sentiments to predict the book sales [14, 15]. Opinmind[16] summarizes the weblog search results with positive and negative categories. On the other hand, researchers also extract the subtopics in weblog collections,

and track their distribution over time and locations [12]. Last year, Mei and others proposed a mixture model to model both facets and opinions at the same time [11]. These previous work aims at generating sentiment summary for a topic purely based on the blog articles. We aim at aligning blog opinions to an expert review. We also take a broader definition of opinions to accommodate the integration of opinions for an arbitrary topic.

Topic model has been widely and successfully applied to blog articles and other text collections to mine topic patterns [5, 1, 21, 9]. Our work adds to this line yet another novel use of such models for opinion integration. Furthermore, we explore a novel way of defining prior.

7. CONCLUSIONS

In this paper, we formally defined a novel problem of opinion integration which aims at integrating opinions expressed in a well-written expert review with those in various Web 2.0 sources such as Weblogs to generated an aligned integrated opinion summary. We proposed a new opinion integration method based on semi-supervised probabilistic topic modeling. With this model, we could automatically generate an integrated opinion summary that consists of (1) supporting opinions with respect to different aspects in the expert review; (2) opinions supplementary to those in the expert review but on the same aspect; and (3) opinions on extra aspects which are not even mentioned in the expert review. We evaluate our model on integrating opinions about two quite different topics (a product and a political figure) and the results show that our method works well for both topics. We are also planning to evaluate our method more rigorously. Since integrating and digesting opinions from multiple sources are critical in many tasks, our method can be applied to develop many interesting applications in multiple domains. A natural future research direction would be to address the more general setup of the problem – integrating opinions in arbitrary text collections with a set of expert reviews instead of a single expert review.

8. ACKNOWLEDGMENTS

This work was in part supported by the National Science Foundation under award numbers 0425852, 0428472, and 0713571. We thank the anonymous reviewers for their useful comments.

9. REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[2] D. Dave and S. Lawrence. Mining the peanut gallery: opinion extraction and semantic classification of product reviews.

[3] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of Royal Statist. Soc. B*, 39:1–38, 1977.

[4] M. Gamon, A. Aue, S. Corston-Oliver, and E. K. Ringger. Pulse: Mining customer opinions from free text. In *IDA*, volume 3646 of *Lecture Notes in Computer Science*, pages 121–132, 2005.

[5] T. Hofmann. Probabilistic latent semantic analysis. In *Proc. of Uncertainty in Artificial Intelligence, UAI'99*, Stockholm.

[6] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of SIGIR '99*, pages 50–57.

[7] M. Hu and B. Liu. Mining and summarizing customer reviews. In *KDD*, pages 168–177.

[8] M. Hu and B. Liu. Mining opinion features in customer reviews. In *AAAI*, pages 755–760.

[9] W. Li and A. McCallum. Pachinko allocation: Dag-structured mixture models of topic correlations. In *ICML '06: Proceedings of the 23rd international conference on Machine learning*, pages 577–584.

[10] B. Liu, M. Hu, and J. Cheng. Opinion observer: analyzing and comparing opinions on the web. In *WWW '05: Proceedings of the 14th international conference on World Wide Web*, pages 342–351.

[11] Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai. Topic sentiment mixture: Modeling facets and opinions in weblogs. In *Proceedings of the World Wide Conference 2007*, pages 171–180.

[12] Q. Mei, C. Liu, H. Su, and C. Zhai. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 533–542.

[13] Q. Mei and C. Zhai. A mixture model for contextual text mining. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 649–655.

[14] G. Mishne and M. de Rijke. MoodViews: Tools for blog mood analysis. In *AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW 2006)*, pages 153–154.

[15] G. Mishne and N. Glance. Predicting movie sales from blogger sentiment. In *AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW 2006)*.

[16] Opinmind. http://www.opinmind.com.

[17] A.-M. Popescu and O. Etzioni. Extracting product features and opinions from reviews. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 339–346.

[18] M. F. Porter. An algorithm for suffix stripping. pages 313–316, 1997.

[19] T. Tao and C. Zhai. Regularized estimation of mixture models for robust pseudo-relevance feedback. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 162–169.

[20] C. Zhai and J. Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of CIKM 2001*, pages 403–410.

[21] C. Zhai, A. Velivelli, and B. Yu. A cross-collection mixture model for comparative text mining. In *Proceedings of KDD '04*, pages 743–748.

[22] L. Zhuang, F. Jing, and X.-Y. Zhu. Movie review mining and summarization. In *CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 43–50.