2009 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Workshops

Specialized Review Selection for Feature Rating Estimation

Chong Long[†] Jie Zhang[‡] Minlie Huang[†] Xiaoyan Zhu[†] [§] Ming Li[‡] Bin Ma[‡]
[†] State Key Laboratory of Intelligent Technology and Systems, Tsinghua National Laboratory for Information Science and Technology, Department of Computer Science and Technology, Tsinghua University, China
[‡]School of Computer Science, University of Waterloo, Canada
[§] Corresponding Author: zxy-dcs@tsinghua.edu.cn

***Abstract*—On participatory Websites, users provide opinions about products, with both overall ratings and textual reviews. In this paper, we propose an approach to accurately estimate feature ratings of the products. This approach selects user reviews that extensively discuss specific features of the products (called specialized reviews), using information distance of reviews on the features. Experiments on real data show that overall ratings of the specialized reviews can be used to represent their feature ratings. The average of these overall ratings can be used by recommender systems to provide feature specific recommendations that better help users make purchasing decisions.**

***Keywords*-Data Mining; Text Mining; Kolmogorov Complexity; Information Distance**

I. INTRODUCTION

With the rapid development of Web2.0 and e-commerce that emphasizes the participation of users, Websites such as Amazon (www.amazon.com) encourage users to express opinions on products by posting overall ratings and textual reviews. These ratings are often used by recommender systems to recommend highly rated products, assisting users in making decisions [1]. However, overall ratings of a product tend to be generic and a user may be more interested in some particular feature of the product. Most Websites do not ask for feature ratings simply because it may cost users too much effort to provide detailed feature ratings. Even for Websites that do collect feature ratings such as TripAdvisor (www.tripadvisor.com), a large portion (approximately 43%) of users will still not provide feature ratings.

We propose a method to estimate feature ratings by making use of users' textual reviews. Our approach is distinguished from the existing work [2], [3] that predicts feature ratings through semantic orientation classification. We look for correlation between feature ratings and overall ratings. More specifically, we believe that overall ratings of products are heavily affected by users' feelings about a certain feature if they extensively discuss this feature in their reviews (called specialized reviews), and their ratings for the feature should be similar to their overall ratings. We select specialized reviews using the information distance measure based on Kolmogorov complexity. Reviews are ranked according to their information distance specialized on one feature. The overall ratings of the ranked top reviews are then used to represent the ratings for that feature.

We carry out experiments to support our approach using data collected from a popular travel Website TripAdvisor. For selected specialized reviews, the corresponding overall ratings are carefully verified to be more similar to their feature ratings. This result becomes the basis of our proposed feature rating estimation approach. Evidence from this dataset also indicates that feature ratings and overall ratings of selected specialized reviews are more similar to each other, respectively. We are then able to use the average of these overall ratings as the estimation of an average feature rating. This average feature rating actually represents an overall opinion of a set of users who care more about the feature and are more likely to be knowledgable (or "experts") on this feature. It can be used by recommender systems to recommend highly rated products based on this feature.

II. HYPOTHESES FOR FEATURE RATING ESTIMATION

A textual review written by a user for a product normally reveals how much the user cares about certain features of the product. Extensive discussion of a feature reveals that the user cares more about the feature. It is also intuitive that users' overall ratings of a product will be more heavily affected by the features that the users care about the most. We therefore believe that users who extensively discuss a certain feature in their textual reviews are likely to provide feature ratings that are close to overall ratings. These reviews are referred to as specialized reviews on the feature. For example, on the TripAdvisor Website, in her review a user strongly criticized rooms at a hotel:

Example 1: "PLEASE DO NOT STAY HERE!!" ... I reserved a non-smoking room, and was placed in a smoking room. The halls, even in non-smoking, stink of stale smoke. There was no hot water(!) in my bathroom. The beds have 1 sheet and no mattress pad. The pillows are stained and have 1 pillow case, no cover. The bathroom has mold all over the tub lining ...
This user gave "1" (in a range of 1-5) to both the feature "Rooms" rating and the overall rating. We formally state our argument in Hypothesis 1:

978-0-7695-3801-3/09 $26.00 © 2009 IEEE
DOI 10.1109/WI-IAT.2009.38

IEEE computer society

Hypothesis 1: The ratings for a feature provided by users who extensively discuss this feature in their reviews are more similar to the overall ratings given by the users.
This hypothesis indicates that, for a user who writes a specialized review on a feature, we can use the overall rating to estimate the rating for the feature if absent.

Statistical analysis by Talwar et al. [4] on real data shows that users who extensively discuss a certain feature are more likely to agree on a common rating for that feature. This evidence is formally stated in Hypothesis 2 as follows:

Hypothesis 2: The ratings for a feature that correspond to specialized reviews on this feature are more similar to each other than to the whole collection of ratings for the feature.

Together with Hypothesis 1, Hypothesis 2 can then be extended as follows:

Hypothesis 3: The overall ratings for reviews that are specialized on a feature are more similar to each other than to the entire collection of overall ratings.
Hypothesis 3 indicates that the users who amply discuss a certain feature of a product tend to converge to a common opinion about the product.

When users amply discuss a certain feature of a product, the feature is obviously important to the users. Since people tend to be more knowledgable in the aspects they consider important, these users' ratings will contain a subset of "expert" ratings for the feature. For the specialized reviews written by these users, their feature ratings are more similar to overall ratings (Hypothesis 1). Also, according to Hypotheses 2 and 3, feature ratings and overall ratings for selected specialized reviews are more similar to each other, respectively. The average of these overall ratings should also be closer to the average of these feature ratings. We are then able to use the average of the overall ratings to estimate an average feature rating, representing the overall opinion from more knowledgable users about the feature.

In summary, we have the feature rating estimation approach that works as follows. We select specialized reviews on a feature using a specialized review selection method. The overall ratings for these reviews will be averaged to represent an average feature rating. We verify the three hypotheses and evaluate our approach in Section IV.

III. Specialized Review Selection

The process of selecting specialized reviews can be viewed as measuring each feature's "weight" on a review. A review specialized on one feature probably means that this feature get a much higher "weight" than any other feature. Talwar et al. [4] proposed a method to compute the weight of a feature in a review. Firstly, a large number (approximately 40 to 50 on average) of words related to a feature are manually selected. Then, a feature's weight in a review is measured by the number of corresponding feature words appeared in this review. Finally, the speciality of a review on one feature depends on its weight.

We, however, select specialized reviews through a different way. Firstly, we minimize manual work by automatically generating a collection of associated feature words. Second, their method lacks a theoretical basis and ours is based on well established measurement of information distance. Our review selection method makes use of the information distance measure based on Kolmogorov complexity.

A. Kolmogorov Complexity and Information Distance

Fix a universal Turing machine U. The Kolmogorov complexity [5] of a binary string x condition to another binary string y, $K_U(x|y)$, is the length of the shortest (prefix-free) program for U that outputs x with input y. It can be shown that for different universal Turing machine U', for all x, y

$$K_U(x|y) = K_{U'}(x|y) + C,$$

where the constant C depends only on U'. Thus $K_U(x|y)$ can be simply written as $K(x|y)$. They write $K(x|\epsilon)$, where ϵ is the empty string, as $K(x)$. It has also been defined in [6] that the energy to convert between x and y to be the smallest number of bits needed to convert from x to y and vice versa. That is, with respect to a universal Turing machine U, the cost of conversion between x and y is:

$$E(x, y) = \min\{U(x, p) = y,\ U(y, p) = x\} \tag{1}$$

It is clear that $E(x, y) \leq K(x|y) + K(y|x)$. From this observation, the following theorem has been proven in [6]:

Theorem 1: $E(x, y) = \max\{K(x|y), K(y|x)\}$.

Thus, the max distance was defined in [6]:

$$D_{\max}(x, y) = \max\{K(x|y), K(y|x)\}. \tag{2}$$

This distance is shown to satisfy the basic distance requirements such as positivity, symmetricity, triangle inequality and is admissible.

Here for an object x, we can measure its information by Kolmogorov complexity $K(x)$; for two objects x and y, their shared information can be measured by information distance $D(x, y)$. In [7], the authors generalize the theory of information distance to more than two objects. Similar to Equation (1), given strings $x_1, \ldots, x_n$, they define the minimum amount of thermodynamic energy needed to convert from any x_i to any x_j as:

$$E_m(x_1, \ldots, x_n) = \min\{|p| : U(x_i, p, j) = x_j\}$$

Then, it is proven in [7] that:

Theorem 2: Modulo to an $O(\log n)$ additive factor,

$$E_m(x_1, \ldots, x_n) \leq \min_i \sum_{k \neq i} D_{\max}(x_i, x_k) \tag{3}$$

Given n objects, this equation may be interpreted as the most specialized object that is similar to all of the others. In the next section, we will use this to guide our practical work.

However, our problem is that neither the Kolmogorov complexity $K(\cdot,\cdot)$ nor $D_{max}(\cdot,\cdot)$ is computable. Therefore, we find a way to "approximate" these two measures. The most useful information in a review article is the English words that are related to the features. If we can extract all of these related words from the review articles, the size of the word set can be regarded as an estimation of information content (or Kolmogorov complexity) of the review articles.

Our method for selecting specialized reviews is outlined as follows. First, for each type of product or service (such as a hotel), a small set of core feature words (such as price and room) is generated statistically. Then this set of core feature words is used to generate the expanded words. Thirdly, a parser is used to find the dependent words associated to the occurrences of the core feature words and expanded words in a review. For each review-feature pair, the union of the core feature words, expanded words and dependent words in the review defines the related word set of the review on the feature. Lastly, information distance is used to select the the best specialized review.

B. Word Extraction

Feature words are the most direct and frequent words describing a feature, for example, price, room or service of a hotel. Given a feature, the core feature words are the very few most common English words that are used to refer to that feature. For example, both "value" and "price" are used to refer to the same feature of a hotel. In [3], the authors indicate that when customers comment on product features, the words they use converge. If we remove the feature words with frequency lower than 1% of the total frequency of all feature words, the remaining words, which are just core feature words, can still cover more than 90% occurrences. So firstly we extract those words through statistics; then some of those with the same meaning (such as "value" and "price") are grouped into one feature. They are just "core feature words".

Apart from core feature words, many other less-frequently used words that are connected to the feature also contribute to the information content of the feature. For example, "price" is an important feature of a hotel, but the word "price" is usually dropped from a sentence. Instead, words such as "$", "dollars", "USD", and "CAD" are used. We use information distance $d(.,.)$ based on Google to expand words [8]. Let α be a feature and $\mathcal{A}$ be the set of its core feature words. The distance between a word w and the feature α is then defined to be

$$d(w,\alpha) = \min_{v\in\mathcal{A}} d(w,v)$$

A distance threshold is then used to determine which words should be in the set of expanded words for a given feature.

If a core feature word or an expanded word is found in a sentence, the words which have grammatical dependent relationship with it are called the dependent words [9]. For example, in sentence "It has a small, but beautiful room", the words "small" and "beautiful" are both dependent words of the core feature word "room". All these words also contribute to the reviews and are important to determine the reviewer's attitude towards a feature.

The Stanford Parser [9] is used to parse each review. For review i and feature j, the core feature words and expanded words in the review are first computed. Then the parsing result is examined to find all the dependent words for the core feature words and expanded words, all of which are called "related words".

C. Computing Information Distance

Let S and T be two sets of words. Then the Kolmogorov complexity can be intuitively estimated by

$$K(S) = \sum_{w\in S} K(w), \;\; K(S|T) = \sum_{v\in S\setminus T} K(v)$$

Here w and v are the words in S and $S \setminus T$, respectively. Then we can use frequency count, and use Shannon-Fano code (Page 67, Example 1.11.2 in [**?**]) to encode a phrase which occurs in probability p in approximately $-\log p$ bits to obtain a short description. Therefore, a related word w's complexity can be estimated by

$$K(w) = -\log P(w|u) = -\log P(w) + \log P(u)$$

where w is in feature u's related word set. A word w or a feature u's probability can be approximated by its document frequency in a corpus.

Such intuition of estimating $K(S|T)$ can be extended to vectors of sets. For two vectors of sets $S_i = (S_{i1}, S_{i2}, \ldots, S_{in})$, $i \in \{1,2\}$, define

$$S_1S_2 = S_1 \cup S_2 = (S_{11} \cup S_{21}, \ldots, S_{1n} \cup S_{2n})$$

$$K(S_i) = \sum_{j=1}^{n} K(S_{ij}), \quad K(S_1|S_2) = \sum_{j=1}^{n} K(S_{1j}|S_{2j})$$

Then, $D_{max}(S_1,S_2) = \max\{K(S_1|S_2), K(S_2|S_1)\}$. Thus, we are able to use Equation (3) for review selection.

If there are m reviews $(x_1, x_2, \ldots, x_m)$ and n features $(u_1, u_2, \ldots, u_n)$, the best specialized review selection needs some minor changes to Equation (3). Without modification, the best specialized review i for a feature j would be such that

$$i = \arg\min_i \sum_{k\neq i} D_{max}(S_{ij}, S_{kj}).$$

However, for specialized review selection we want that (a) the review focus on the given feature only, and (b) a review article that does not discuss the feature should not be counted in the selection. Therefore, the above formula should be modified to be

$$i = \arg\min_i \sum_{S_{kj}\neq\emptyset} D_{max}(S_i, S_{kj}), \tag{4}$$

More specifically, S_{ij} is changed to S_i to penalize the content of review i not related to feature j; and the reviews with an empty word set on feature j are excluded from the selection. In the next section, Equation (4) is used to select specialized reviews.

IV. Experimental Results

In this section, we present a set of experimental results to justify our work. Our experiments are carried out using real data collected from the travel Website TripAdvisor(www.tripadvisor.com). This Website indexes hotels from cities across the world. It collects feedback from travelers. Feedback of each traveler consists of an overall rating (from 1, lowest, to 5, highest), a textual review written by the traveler, and numerical ratings for different features of hotels (e.g., value, service, rooms).

Table I
Summary of the Data Set

Location	# Hotels	# Feedback	# Feedback with feature ratings
Boston	57	3949	2096
Sydney	47	1370	879
Vegas	40	5588	3144

We crawled this Website to collect travelers' feedback for hotels in three cities: Boston, Sydney and Las Vegas. During this crawling process, we carefully removed information about travelers and hotels to protect their privacy. Their names were replaced by randomly generated unique numbers. For users' feedback, we recorded overall ratings, textual reviews, and numerical ratings for four features: Value(V), Rooms(R), Service(S) and Cleanliness(C). These features are rated by a significant number of users. Table I summarizes our data set. For each city, this table contains information about the number of hotels, the total amount of feedback and the amount of feedback with feature ratings. In general, each hotel has sufficient feedback with feature ratings for us to evaluate our work.

A. Specialized Review Selection

We first evaluate the performance of our specialized review selection approach using manually annotated data. More specifically, in our data set, 415 reviews for Boston hotels, 161 for Sydney hotels, and 420 for Las Vegas hotels (996 reviews in total) are selected for manual annotation. Two annotators look over each review and agree on whether the review is specialized or not. The reviews are annotated as "specialized" only when both of these two annotators believe that they are specialized. After that, each review has one of the following five labels: Specialized on feature Value (SV), Specialized on feature Room (SR), Specialized on feature Service (SS), Specialized on feature Cleanliness (SC), and Not Specialized (N).

Table II
The Best Specialized Reviews (Boston)

Top #	V	R	S	C
1	SV	SR	SS	SC
2	SV	SR	SS	SC
3	SV	N	SS	SC
4	SV	SR	SS	SC
5	N	SR	N	SC
⋮	⋮	⋮	⋮	⋮

The reviews for hotels in each city are ranked according to their information distances on each feature. For example, the best specialized review on feature "Value", which has the minimal information distance (see Equation 4) to this feature, is ranked No.1. Table II shows the annotated reviews for Boston hotels that are ranked on top five on each feature. It can be seen that most of these top reviews are labeled as specialized reviews on respective features. Our specialized review selection approach generally performs well.

To clearly present the performance of our specialized review selection approach, we use the measures of precision, recall and f-score. These three measures are formally defined as follows. Suppose there are N reviews in total. Let p_{jk} ($1 \leq k \leq N$) be the review ranked the kth specialized on feature j. Define

$$z_{jk} = \begin{cases} 1 & p_{jk} \text{ labeled specialized on feature } j; \\ 0 & \text{otherwise.} \end{cases}$$

The precision (P), recall (R), and f-score (F) of top k reviews specialized on feature j are formalized as follows:

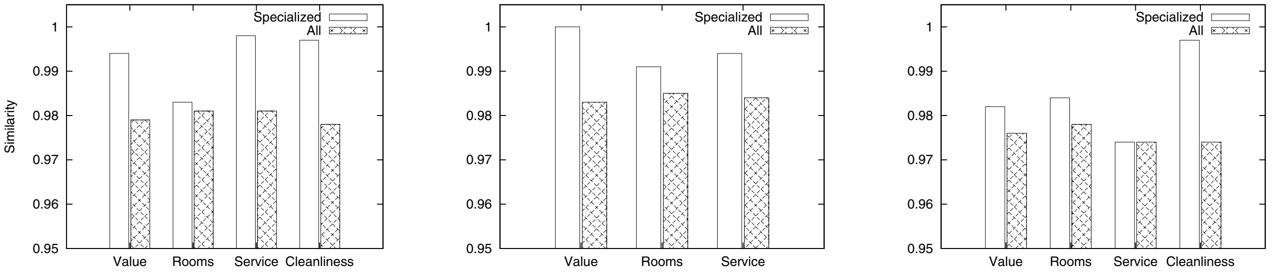
$$P_{jk} = \frac{\sum_{l=1}^{k} z_{jl}}{k}, \quad R_{jk} = \frac{\sum_{l=1}^{k} z_{jl}}{\sum_{l=1}^{N} z_{jl}}, \quad F_{jk} = \frac{2P_{jk}\, R_{jk}}{P_{jk} + R_{jk}}$$

For each ranked review set on a feature, the maximum f-score and its associated precision and recall are listed in the last three columns of Table III. It can be seen that for the best f-scores, the precision and recall values are mostly larger than 70%, that is, a great part of reviews labeled as specialized receive top rankings by using our specialized review selection. Note that there are no selected reviews specialized on feature "Cleanliness" for the selected hotel reviews in Sydney, so there are no results for this row. Also note that only very few reviews for hotels in Sydney are labeled specialized on the feature "Value" and they are all ranked on the top, therefore the precision, recall and f-score are high as 1.0. The column "# FW" in Table III lists the number of core feature words used by each method.

We also compare our approach with the method Talwar et al. [4] and the TF*IDF method. As mentioned in Section III, the method of Talwar et al. computes the weight of a feature in a review based the number of corresponding feature words appeared in this review. The speciality of a review on one feature depends on the weight. The TF*IDF method weights a feature based on the TF*IDF scores of the related

Table III
EVALUATION AND COMPARISON OF SPECIALIZED REVIEW SELECTION

City		Results of Talwar et al.				TF*IDF Method				Our Results			
		# FW	Precision	Recall	F-Score	# FW	Precision	Recall	F-Score	# FW	Precision	Recall	F-Score
Boston	V	34	0.833	0.667	**0.741**	5	0.571	0.533	**0.552**	5	0.833	0.667	**0.741**
	R	37	0.689	0.836	**0.756**	7	0.642	0.705	**0.672**	7	0.652	0.738	**0.692**
	S	59	0.857	0.737	**0.792**	4	0.738	0.789	**0.763**	4	0.762	0.842	**0.800**
	C	49	0.692	0.750	**0.720**	2	0.500	0.667	**0.571**	2	0.800	0.667	**0.727**
Sydney	V	34	0.667	1.000	**0.800**	5	0.667	1.000	**0.800**	5	1.000	1.000	**1.000**
	R	37	0.815	0.733	**0.772**	7	0.714	0.625	**0.667**	7	0.741	0.719	**0.730**
	S	59	0.882	0.750	**0.811**	4	0.484	0.750	**0.588**	4	0.684	0.650	**0.667**
Vegas	V	34	0.586	0.680	**0.630**	5	0.333	1.000	**0.500**	5	0.727	0.640	**0.681**
	R	37	0.750	0.672	**0.709**	7	0.677	0.657	**0.667**	7	0.691	0.701	**0.696**
	S	59	0.800	0.538	**0.644**	4	0.685	0.712	**0.698**	4	0.696	0.750	**0.722**
	C	49	0.414	0.600	**0.490**	2	0.650	0.650	**0.650**	2	0.813	0.650	**0.722**



Boston Sydney Vegas

Figure 1. Similarity between Overall and Feature Ratings (Using Manually Labeled Data)

words for the feature. The results of these two methods are also listed in Table III. Although the method of Talwar et al. produces generally similar results, a greater number of core feature words are manually selected and used for this method. Moreover, due to the bias of manual selection, it performs not so well for the feature “Cleanliness” on the Vegas review set. Our approach produces more stable and reliable results. Our approach also outperforms the TF*IDF method when using the same number of core feature words.

B. Verification of Hypotheses

In the previous section, we evaluated our review selection method and it performs better than the other methods. The main purpose of our work is to make use of this method in selecting specialized reviews for feature rating estimation. This idea is supported by the three hypotheses presented in Section II. Here, we carefully verify these hypotheses.

1) Hypothesis 1: We first verify Hypothesis 1 using the manually annotated data. As mentioned in Section IV-A, we manually label for randomly selected reviews as specialized or not on feature. For specialized reviews on each feature, the similarity between their feature ratings and overall ratings is calculated and compared with the similarity between feature and overall ratings for all reviews. More formally, if there are m reviews labeled as specialized on a feature, their feature ratings are formed as a vector X and the overall ratings are formed as Y. Cosine similarity [10] is used to measure the similarity between them as follows:

$$Sim(X,Y) = \frac{X \bullet Y}{\|X\| \, \|Y\|}$$

The similarity between overall ratings and feature ratings of all reviews for hotels in each city is also calculated. Figure 1 compares these similarities for each feature of hotels in each city (from left to right, Boston, Sydney and Vegas). It is clear that the similarities between overall ratings and feature ratings for specialized reviews are higher than those for all reviews. We do not have results for the feature “Cleanliness” of the hotels in Sydney because there are no reviews annotated as specialized on this feature among those randomly selected reviews for hotels in this city.

We also verify Hypothesis 1 using specialized reviews selected by our review selection approach. All reviews are ranked based on each feature for hotels in each city. Because of the space limitation we show here only the results for the city of Boston in Figure 2. Each diagram in this figure shows the similarities between overall and feature ratings for top 5, top 10, top 15, ..., top 995 and top 1000 reviews selected by our review selection method. From this figure, we can see that the overall and feature ratings of more specialized reviews tend to have higher similarities. Therefore, overall ratings of the best specialized reviews can be used to represent their feature ratings.

2) Hypothesis 2: In this experiment, we use our review selection approach to select specialized reviews to verify

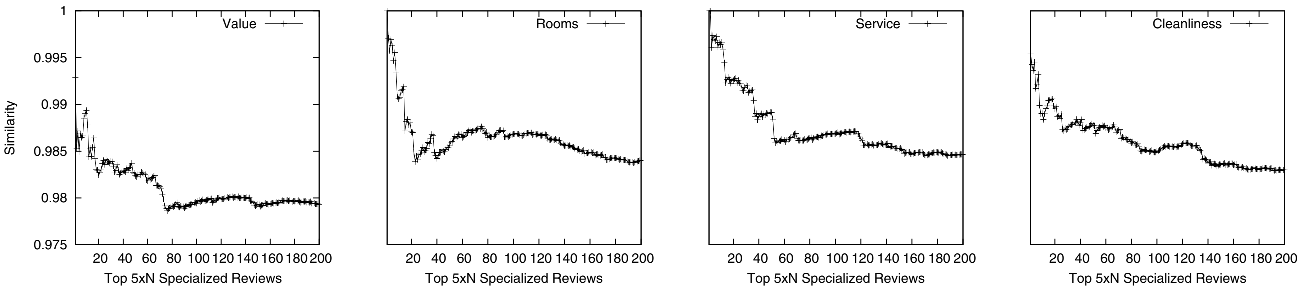


Figure 2. Similarity between Overall and Feature Ratings (Specialized Review Selection Method)

Hypothesis 2. More specifically, for each city, hotels that receive at least 10 reviews with feature ratings are selected. We use our specialized review selection approach to select top 20% and 50% specialized reviews on each feature for hotels in each city. We calculate the standard deviation of their feature ratings, as well as that of all feature ratings, for each hotel in a city. We then average these standard deviations over all the hotels in the same city. The average values are listed in Table IV. The feature ratings of specialized reviews have smaller average standard deviations. Standard T-test is used to measure the significance of the results between top 20% specialized reviews and all reviews, city by city and feature by feature. Their p-values are shown in the braces, and they are significant at the significant level of 0.05. Although for some items there does not seem to be a significant difference, the results are significant for the entire data set.

Table IV
DEVIATION OF FEATURE RATINGS

City	#Hotel	Feature	20%	50%	All
Boston	52	V	0.921 (0.0003)	1.040	1.136
		R	0.953 (0.1851)	1.028	1.013
		S	0.936 (0.0057)	1.070	1.144
		C	0.756 (0.0009)	0.866	0.949
Sydney	34	V	0.925 (0.0670)	0.991	1.054
		R	0.677 (0.0030)	0.879	0.945
		S	0.815 (0.0023)	1.006	1.115
		C	0.660 (0.0056)	0.737	0.907
Vegas	33	V	1.116 (0.0324)	1.210	1.291
		R	0.885 (0.0058)	1.161	1.175
		S	1.165 (0.1130)	1.246	1.269
		C	1.056 (0.1477)	1.058	1.158

Therefore, when these travelers write reviews that focus on one feature, their feature ratings tend to converge. These feature ratings can be averaged to represent a specific opinion of these travelers on that feature.

3) Hypothesis 3: Similar to our verification of Hypothesis 2, we verify Hypothesis 3 by first selecting the top 20% and 50% specialized reviews on each feature. We then calculate the average standard deviations of the overall ratings for the specialized reviews, as well as those for all reviews, as listed in Table V. The p-values are also calculated and shown in the braces. It is clear that, for any feature, overall ratings for specialized reviews have smaller standard deviation. Travelers who wrote these reviews tend to agree on their overall ratings for hotels.

Table V
DEVIATION OF OVERALL RATINGS

City		V	R	S	C
Boston	20 %	0.882 (0.0005)	0.982 (0.011)	0.930 (0.0026)	0.866 (0.0002)
	50%	1.016	1.123	1.061	1.005
	All	1.128			
Sydney	20 %	0.874 (0.050)	0.726 (0.0015)	0.751 (0.0024)	0.831 (0.017)
	50%	0.947	0.963	0.953	0.930
	All	1.058			
Vegas	20 %	1.075 (0.011)	0.928 (0.0005)	1.185 (0.111)	1.103 (0.026)
	50%	1.237	1.264	1.258	1.224
	All	1.299			

C. Estimating Average Feature Rating

Supported by the three hypotheses verified in the previous section, we can then use the average of overall ratings for specialized reviews on a feature to estimate an average rating for this feature. This single rating of the feature can represent a general opinion of users who are more knowledgable about this feature. We use an average of the feature ratings for top 20% specialized reviews to reflect a general opinion of knowledgable/expert users. Note that if we use a large number of specialized reviews, the error of estimating feature ratings from the overall ratings for these reviews will be large. On another hand, if we use a small number of specialized reviews, the average feature rating estimated based on the overall ratings for these reviews may not be representative. This average feature rating can be estimated using the overall ratings for the top 20% specialized reviews. In this section, we first provide statistical analysis to show the rationale of ranking hotels using an average of feature/overall ratings for specialized reviews. We then directly evaluate the performance of this process.

Given a set of reviews for one hotel, let F_{20} be the mean of the feature ratings for top 20% specialized reviews, and O_{20} be the mean of the overall ratings for these reviews. Similarly, the mean values of feature and overall ratings for

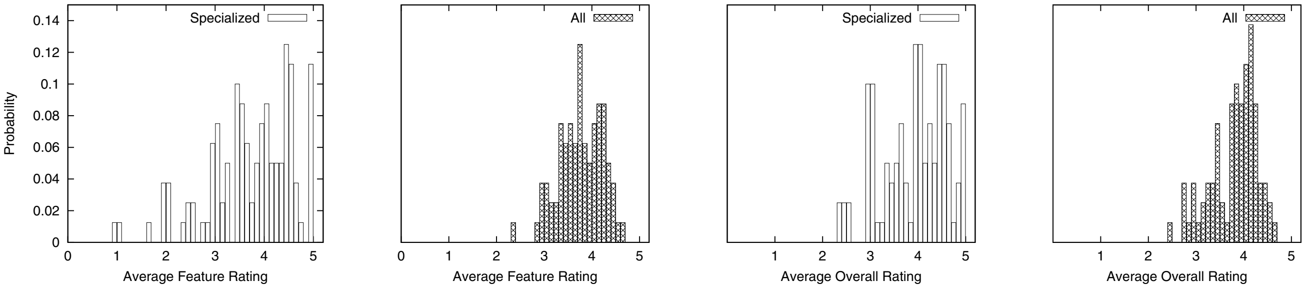


Figure 3. Comparison between Average Rating Distribution on Specialized Reviews and that on All Reviews

all reviews are referred to as F_{100} and O_{100} respectively. We plot in Figure 3 the distribution of these average feature ratings (F_{20} and F_{100} in the two left most figures) and average overall ratings (O_{20} and O_{100} in the two right most figures) of all hotels. We can see that the average values of feature or overall ratings for all reviews are mostly located around the rating 4. Ranking hotels based on these average ratings is not easily distinguishable. This may be caused by several reasons. The most obvious one is noise (for example, caused by users' subjectivity) in the data that reduces rating difference among hotels. Comparably, the average of ratings for specialized reviews is spread out because more knowledgable opinions tend to contain less noise. It is thus better to use the average of these ratings to rank hotels and to provide feature-specific recommendations.

We evaluate the performance of using overall ratings for specialized reviews to estimate an average feature rating. In this experiment, we randomly choose 20% reviews from the set of reviews for each hotel. We repeat this process for five times. In each time, we calculate the mean of feature and overall ratings. We then obtain the average of the mean values, referred to as Fr_{20} and Or_{20} respectively.

Table VI
RESULT OF ESTIMATING AVERAGE FEATURE RATING

Feature	V	R	S	C	AVG
$\|F_{20} - O_{20}\|$	0.316	0.191	0.231	0.300	0.260
$\|Fr_{20} - Or_{20}\|$	0.396	0.284	0.300	0.389	0.342
$\|F_{20} - F_{100}\|$	0.458	0.429	0.429	0.348	0.416
$\|F_{20} - Or_{20}\|$	0.654	0.635	0.624	0.625	0.634
$\|F_{20} - O_{100}\|$	0.455	0.519	0.430	0.522	0.481

For all hotels that receive no less than ten reviews with feature ratings, we calculate the average absolute differences between F_{20} and O_{20}, written as $|F_{20} - O_{20}|$ in the second row of Table VI. Compared with the average absolute difference between Fr_{20} and Or_{20} (written as $|Fr_{20} - Or_{20}|$ in the third row), $|F_{20} - O_{20}|$ is 23.98% lower. This result directly confirms our idea of using overall ratings for specialized reviews to estimate an average feature rating, resulting from the three verified hypotheses.

Second, we calculate the average difference between F_{20} and F_{100} ($|F_{20} - F_{100}|$) for all these hotels. The result shows that the average of feature ratings for specialized reviews is different from the average of all feature ratings, that is, the general opinion of all users is not the same as that of more knowledgable users. This inequality indicates that we cannot use the former to replace the later. Finally, the last two rows, $|F_{20} - Or_{20}|$ and $|F_{20} - O_{100}|$, show that neither Or_{20} nor O_{100} can be used to estimate F_{20}.

To evaluate the performance of our approach, we also show how much the estimation error $|F_{20} - O_{20}|$ will affect the result of ranking hotels based on their average feature ratings. This directly reflects how well our approach can assist a feature-specific recommender system in recommending hotels. We present both maximum and minimum errors in ranking hotels.

Table VII
ERROR RANGE OF HOTEL RANKING

Feature	V	R	S	C	AVG
Max	0.157	0.105	0.118	0.192	0.131
Min	0.084	0.044	0.050	0.098	0.063
AVG	0.120	0.074	0.084	0.145	0.097

Suppose that m hotels in one city are ranked according to their F_{20}, each of which is written as a_i $(1 \leq i \leq m)$. Let $g(a_i)$ be the ranking of an average feature rating a_i, and $e = |F_{20} - O_{20}|$. The maximum ranking error is formalized as follows:

$$\overline{rd} = \frac{1}{m^2} \sum_{i=1}^{m} \max(|g(a_i + e) - g(a_i)|, |g(a_i - e) - g(a_i)|)$$

The minimum ranking error is defined as follows:

$$\underline{rd} = \frac{1}{m^2} \sum_{i=1}^{m} \min(|g(a_i + e) - g(a_i)|, |g(a_i - e) - g(a_i)|)$$

The average ranking error is the mean of the maximum ranking error $\overline{rd}$ and the minimum ranking error $\underline{rd}$. The results are summarized in Table VII. From this table we can see that the average ranking error range is 9.7%. It is within 3-5 hotel ranks for a city, according to the number of hotels in each city listed in Table IV. Our feature rating estimation method provides sufficiently good performance.

V. Related Work

Our work is aimed at estimating feature ratings of a product based partially on textual reviews. It is related to review mining and sentiment classification. Existing work on review mining focuses mainly on product reviews. As the pioneering work, Hu and Liu [3] proposed to use word attributes, including occurrence frequency, part-of-speech, and synset in WordNet. Zhuang et al. [11] focused on movie reviews and introduced a multi-knowledge based approach to integrate WordNet, statistical analysis, and movie knowledge. The task of sentiment classification is to determine the semantic orientations of words, sentences or documents. [2] is the earliest work of automatic sentiment classification at the document level, using several machine learning approaches with common text features to classify movie reviews. Dave et al. [12] designed a classifier based on information retrieval techniques for feature extraction and scoring.

However, the above work makes use of only textual reviews for sentiment classification. While doing the evaluation, it requires a lot of annotation work from experts. Moreover, given the reviews or sentences which have already been classified as “positive” or “negative”, these researchers do not provide a method to obtain a final feature rating that accurately reflects the general opinion of users. Our feature rating estimation method instead makes use of the information of both textual reviews and overall ratings. It reduces the manual annotation work as much as possible, and is able to produce the general opinion of a special set of expert/knowledgeable users.

VI. Conclusion and Future Work

We developed a novel approach to accurately estimate feature ratings of products, making use of a review selection method for selecting specialized reviews based on information distance of reviews on features. Three hypotheses are verified through experiments. We evaluated our approach based on data collected from an e-commerce Website. Results show that this approach achieves sufficiently high performance. Our work is therefore a novel first attempt to accurately estimate feature ratings using users’ overall ratings and their textual reviews without requiring much effort from experts.

Our approach works well when a sufficient number of specialized reviews exist for each product. This might not be the case for some products, which may not have many reviews. Or, they may have features that are not easily distinguishable. For example, the hotel features of “Rooms” and “Cleanliness” are often both mentioned in many reviews because one important aspect of rooms is their cleanliness. For future work, we may make use of reviews that are not specialized. We plan to use syntax and semantic information and machine learning methods to solve this problem.

After having sufficient feature ratings, we plan to develop a recommender system that makes use of these ratings to provide feature specific recommendations.

Acknowledgment

The work was supported by NSFC under grant No.60621062 and 60803075, the National Basic Research Program (973 project in China) under grant No.2007CB311003, and Microsoft joint project “Opinion Summarization toward Opinion Search”. The work was also supported by IRCI project from the International Development Research Center, Canada.

References

[1] J. B. Schafer, J. Konstan, and J. Riedi, “Recommender systems in e-commerce,” in *1st ACM Conference on Electronic Commerce (EC)*, 1999, pp. 158–166.

[2] B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs up? sentiment classification using machine learning techniques,” in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, July 2002, pp. 79–86.

[3] M. Hu and B. Liu, “Mining and summarizing customer reviews,” in *10th ACM International Conference on Knowledge Discovery and Data Mining (KDD)*, 2004, pp. 168–177.

[4] A. Talwar, R. Jurca, and B. Faltings, “Understanding user behavior in online feedback reporting,” in *8th ACM Conference on Electronic Commerce (EC)*, 2007, pp. 134–142.

[5] M. Li and P. Vitányi, *An Introduction to Kolmogorov Complexity and its Applications (2nd Edition)*. Springer-Verlag, 1997.

[6] C. Bennett, P. Gacs, M. Li, P. Vitányi, and W. Zurek, “Information distance,” *IEEE Transactions on Information Theory*, vol. 44, no. 4, pp. 1407–1423, July 1998.

[7] C. Long, X. Zhu, M. Li, and B. Ma, “Information shared by many objects,” in *ACM 17th Conference on Information and Knowledge Management (CIKM)*, October 2008.

[8] R. L. Cilibrasi and P. M. Vitányi, “The google similarity distance,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 3, pp. 370–383, March 2007.

[9] M. C. de Marneffe, B. MacCartney, and C. D. Manning, “Generating typed dependency parses from phrase structure parses,” in *The fifth international conference on Language Resources and Evaluation (LREC)*, May 2006.

[10] J. Lewis, S. Ossowski, J. Hicks, M. Errami, and H. R. Garner, “Text similarity: an alternative way to search medline,” *Bioinformatics*, vol. 22, no. 18, pp. 2298–2304, 2006.

[11] L. Zhuang, F. Jing, and X. Zhu, “Movie review mining and summarization,” in *ACM 17th Conference on Information and Knowledge Management (CIKM)*, 2006, pp. 43–50.

[12] K. Dave, S. Lawrence, and D. M. Pennock, “Mining the peanut gallery: Opinion extraction and semantic classification of product reviews,” in *International World Wide Web Conference (WWW)*, May 2005, pp. 519–528.