JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 26, 1973-1990 (2010)

Feature-based Product Review Summarization Utilizing User Score*

JUNG-YEON YANG, HAN-JOON KIM+ AND SANG-GOO LEE
Department of Computer Science and Engineering
Seoul National University
Gwanak-gu, Seoul, 151-742 Korea
+*School of Electrical and Computer Engineering*
University of Seoul
Dongdaemun-gu, Seoul, 130-743 Korea

With the steadily increasing volume of e-commerce transactions, the amount of user-provided product reviews is increasing on the Web. Because many customers feel that they can purchase product based on the experiences of others that are obtainable through product reviews, the review summarization process has become important. In particular, feature-based product review summarization is needed in order to satisfy the detailed needs of some customers. To achieve such summarization, numerous techniques using natural language processing (NLP), machine learning, and statistical approaches that can evaluate product features within a collection of review documents have been studied. Many of these techniques require sentiment analysis or feature scoring methods. However, existing sentiment analysis methods are limited when determining the sentiment polarity of context-sensitive words, and existing feature scoring methods are limited when only the overall user score is used to evaluate individual product features. In our summarization approach, context-sensitive information is used to determine sentiment polarity while opinioned-feature frequency is used to determine feature scores. Based on experiments with actual review data, our method improved the accuracy of the calculated feature scores and outperformed existing methods.

Keywords: product review summarization, opinion mining, sentiment analysis, feature scoring, user score

1. INTRODUCTION

With the continuously increasing volume of e-commerce transactions, the amount of product information and the number of product reviews are increasing on the World Wide Web (Web). Because many customers feel that they can make better decisions based on the experiences of others that are obtainable through these product reviews, such reviews have become an important source of enabling information in e-commerce. However, as the number of reviews increases, it becomes more difficult for users to read all of the relevant review documents. In order to alleviate this problem, there have been several studies into ways to summarize product evaluations from reviews [2, 3, 7, 13, 17, 26, 27].

Since automated evaluation summaries can influence buying decisions, accuracy is an important performance metric of review summary methods. In addition, the summary

Received June 9, 2009; revised August 6 & October 5, 2009; accepted November 18, 2009.
Communicated by Chin-Teng Lin.
* This research was supported by the MKE (The Ministry of Knowledge Economy), Korea, under the ITRC (Information Technology Research Center) support program supervised by the NIPA (National IT Industry Promotion Agency) grant No. NIPA-2009-C1090-0902-0031.

information that is shown to users is important. For example, a detailed summary that includes evaluations for several product features, such as cost, size, and design, may be more useful than a summary that only shows an average score for all of the product's features. A product feature characterizes a product and a feature can represent a detailed specification of the product or be a more general term. To prepare feature-level review summaries, numerous techniques to evaluate opinions on product features from a collection of review documents have been studied using natural language processing (NLP) [1-3, 8], machine learning [5], and statistical approaches [7]. Here, we propose a new way of summarizing features from a large number of product reviews.

To summarize evaluations for each product feature, sentiment analysis [1-3, 5, 6, 8, 19] and feature scoring methods [7] have been used. However, sentiment analysis methods are limited when analyzing the sentiment polarity of context-sensitive words [1, 2, 19]. A context-sensitive word is one that has an alternate sentiment polarity according to the context of the word. Context is important because some opinion words that modify some features are evaluated as a positive opinion while it is evaluated as negative in other cases. Furthermore, current feature scoring methods are limited as they use the overall user rating for a product to evaluate individual product features. However, an overall user rating may not be an evaluation for a specific product feature, but rather, is an evaluation of the entire product. As a result, the application of the overall score to a specific feature can be incorrect.

In our work, to obtain more accurate feature scores, both sentiment analysis and feature scoring methods are used for feature evaluation. In addition, we have enhanced the techniques currently used in both methods. These enhancements include the use of sentiment polarity information and opinioned-feature frequency data. As well, overall user ratings for a product are used during the processing of feature scores. To achieve higher accuracy during sentiment analysis, sentiment polarity of a context-sensitive opinion word is assessed using positive and negative sentiment dictionaries that have been constructed from the available user ratings.

In section 2 of this paper, we present a brief review of related works, while the motivation for this work and the definition of the problem are described in section 3. In section 4, we describe a novel review summarization process for use in producing feature-level summaries of product reviews. The detailed methods used for sentiment analysis and feature scoring in that novel approach are shown in section 5 while section 6 reports on experimental results using actual review data. A conclusion is provided in section 7.

2. RELATED WORK

To summarize opinions from documents in an accurate way, many methods have been studied. Lee *et al.* used fuzzy ontology to summarize news documents [28]. In such documents, term ambiguity is an important issue because the meaning of a term may vary with the situation. To match the meaning with the terms, they used a fuzzy ontology approach to summarize opinions from the news documents. Similarly, a fuzzy ontology approach was used by Lee *et al.* as a decision support agent to monitor and control project processes [29].

When performing product review summarization, both sentiment analysis and fea-

ture scoring methods are important to the production of an accurate summary. To determine sentiment polarities, word corpora resources, such as WordNet [21] and SentiwordNet [16] have been used by various authors [1, 2, 8, 16, 23]. For example, Hu and Liu [1] used sets of adjective synonyms to predict the sentiment orientation of opinions while Sista and Srinivasan [5] created a lexicon table using a general inquiry function of WordNet and expanded that lexicon to classify the sentiments of opinion words. Commonly, in such sentiment-based approaches, domain experts define the seed words before analyses can be conducted. Such manual efforts can affect the accuracy of a sentiment prediction as the result can change depending on which words are included in the seed word corpus. Another problem is the difficulty of handling context-sensitive words in opinions. For instance, when the opinion word "big" modifies an LCD screen feature in an MP3 player, its sentiment polarity can be positive. However, when body size is modified by the opinion word "big", the opinion can be negative. Thus, determining the polarities of context-sensitive opinion words is important.

Another common approach to determine sentiment polarity is to use a point-wise mutual information (PMI) value. Such values indicate an association between two words by using the co-occurrences of those words as described in Eq. (1).

$$PMI(word1, word2) = \log_2 \frac{p(word1, word2)}{p(word1) \cdot p(word2)} \tag{1}$$

To calculate PMI values, Turney *et al.* used the probability that two words appear in the same Web document [19]. As in sentiment-based approaches, the PMI approach requires domain experts to define the polarities of seed words before performing analyses.

Recently, a feature scoring approach that evaluates the scoring of product features has been studied [7]. In this approach, user scores in product reviews are used to calculate the feature scores of the various features of that product. A user score is a rating given to a product by a user. Specifically, the frequency of opinioned-features is used to weight the strength of the users' opinions and the distribution of user scores is used to calculate a feature score. This approach utilizes users' product rating scores during review summarization to calculate an overall feature score.

Fig. 1 shows an example of the feature scoring method used by Scaffidi *et al*. [7]. In that method, all features mentioned in a review are given the same score as the user assigned to the entire product. As shown in the figure's example (user review number 5), the size, design, utility, and battery time features all have the same score (*i.e.*, 4, which was given by the user to the whole product). The assigned scores for each feature among the reviews were then averaged. Note that feature size, design, utility, and battery time all have positive overall polarity while color had a negative polarity. Also note an error that although user review number 5 had a negative comment on battery time, that feature was assigned a high (4) rating.

3. MOTIVATION AND PROBLEM DEFINITION

Fig. 2 illustrates a conventional summarization process, which typically consists of two main steps: feature extraction and feature evaluation. That figure also shows that the

<Example> Review No.5 **User score : ★★★★**

The size of camera is good to hold in one hand and comfortable. a design is so cool, nice body!!. But battery time is short. So, in outdoor, additional batteries are needed. This camera is almost perfect!!

Review No.	User score	Product Feature							
		Size	Cost	Design	Utility	Shutter speed	Battery time	A/S	Color
1	★★★★★(5)	5	-	5	-	-	-	-	-
2	★★★★★(5)	-	-	5	-	5	-	-	-
3	★★★★★(5)	5	-	-	5	5	5	-	-
4	★★★★(4)	4	-	-	4	4	4	-	-
5	★★★★(4)	4	-	4	-	-	4	-	-
6	★★★★(4)	-	-	4	4	4	4	-	-
7	★★★(3)	-	-	3	3	-	-	-	-
8	★★★(3)	-	-	3	-	-	-	3	-
9	★★★(3)	-	3	-	-	-	3	-	-
10	★(1)	-	-	-	-	-	-	1	1
11	★(1)	-	1	-	-	-	-	-	1
Average score		4.5	2	4	4	4.5	4	2	1
Overall polarity		pos	neg	pos	pos	pos	pos	neg	neg

pos : positive, neg : negative

Fig. 1. Examples of the previous feature scoring method.

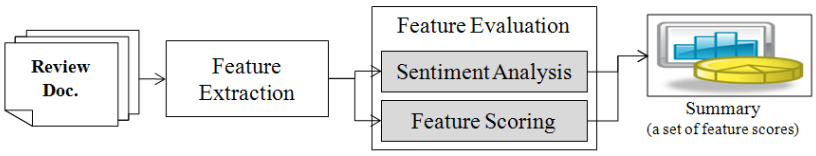

Fig. 2. Summarization procedure of conventional approaches.

review summary for a product is derived from the feature scores that are contained in multiple review documents. Conventionally, the feature score for a product is the overall evaluation for a feature among all of the reviews for the product. The feature extraction step determines the features that are representative of the product and their corresponding opinion words. In that step, an opinion word is the sentiment word that the user applied to the feature. For instance, in the Fig. 1 example review by user 5, 'good', 'cool', and 'short' are opinion words. The subsequent feature evaluation step evaluates the features among all available review documents and considers the semantics of the opinion words and the associated user score for a product. Conventionally, this step requires the use of sentiment analysis [5, 8, 10, 16, 18, 19, 22] or feature scoring [7] methods.

Sentiment analysis is used to determine the sentiment polarity of an opinion about a specific product feature. The feature score is then derived from the sentiment polarities of the various opinions on a feature. For this step, NLP and statistical approaches can be used [1-3, 5-8, 10]; however, NLP techniques require a considerable amount of manual effort and comprehensive domain knowledge. On the other hand, the feature scoring method is based on users' overall product evaluation [7]. Such scores are comprehensive for the entire product and are not individually determined for each product feature. Thus,

a user score represents a generalized evaluation derived from many opinion words on a variety of features. In addition, although there is only one score value, there may be both positive and negative evaluations on a specific feature within the same review.

To overcome those problems, our approach calculates a feature score by analyzing both the user score and the sentiment polarities of the opinion words. Fig. 3 indicates how sentiment polarities of opinion words can affect feature scores. In the example sentences from reviewer number 5, the user's opinion on the size and design features have positive sentiment polarities. However, the opinion on the battery time feature has a negative sentiment polarity. In the conventional approach, the battery time feature has the same score as the overall score from the review (Fig. 1). Assigning such a high score to this feature results in a sentiment error. Using a sentiment polarity approach to the reviews that produced Fig. 1 would result in a distribution of sentiment polarities such as that shown in Fig. 3. That figure shows that the overall polarities of the 'utility' and 'battery time' features are different from those in Fig. 1. This result shows that there can be erroneous summary opinions calculated if only overall user scores are used to determine feature scores. Thus, feature-level sentiment polarity needs to be considered when attempting to obtain accurate feature scores.

<Example> Review No.5 **User score : ★★★★**

The size of camera is good to hold in one hand and comfortable. a design is so cool, nice body!!. But battery time is short. So, in outdoor, additional batteries are needed. This camera is almost perfect!!

Review No.	User score	Product Feature							
		Size	Cost	Design	Utility	Shutter speed	Battery time	A/S	Color
1	★★★★★(5)	pos	-	pos	-	-	-	-	-
2	★★★★★(5)	-	-	pos	-	pos	-	-	-
3	★★★★★(5)	pos	-	-	pos	pos	neg	-	-
4	★★★★(4)	pos	-	-	neg	pos	neg	-	-
5	★★★★(4)	pos	-	pos	-	-	neg	-	-
6	★★★★(4)	-	-	pos	neg	pos	neg	-	-
7	★★★(3)	-	-	neg	pos	-	-	-	-
8	★★★(3)	-	-	pos	-	-	-	neg	-
9	★★★(3)	-	neg	-	-	-	neg	-	-
10	★(1)	-	-	-	-	-	-	neg	neg
11	★(1)	-	neg	-	-	-	-	-	neg
Overall polarity		pos	neg	pos	neu	pos	neg	neg	neg

pos : positive, neg : negative, neu : neutral

Fig. 3. Example concerning sentiment polarities of features.

Another problem is the presence of bias in user reviews. Based on an examination of user-provided review data from Epinions (http://www.epinions.com/), the number of product reviews with high user scores was 5-10 times higher than the number of reviews with low user scores [30]. This suggests the presence of bias within the available user-provided reviews. If bias is present in user-provided reviews and the biased scores are used as the source of feature scores, then most features will have high scores. Therefore,

not only do user scores as a source of feature scores ignore sentiment polarity, but they may also be biased.

In previous studies that calculated feature scores through sentiment analyses, they were unable to handle context-sensitive words that had different sentiment polarities depending on the feature they were modifying [2, 8, 19]. Both NLP- and PMI-based approaches are able to determine the sentiment polarity of an opinion using common opinion words; however, for context-sensitive words, sentiment polarity may change depending on the product or feature. To correctly perform feature scoring, opinions on each feature need to be assessed. Therefore, analysis of the sentiment polarity of each opinion is an important basic process when calculating feature scores.

Fig. 4 shows the system architecture of our novel method of review summarization. Initially, features and opinion words are extracted from product reviews. Feature-opinion pairs are the obtained using the extracted features and their corresponding opinion words. Next, sentiment polarities for each pair are classified via a sentiment analysis process. To determine the polarity of an opinion word in the sentiment analysis step, positive and negative sentiment dictionaries are used. Subsequent to deriving the sentiment polarity of each pairs, scores for each product feature are calculated through a new feature scoring process.

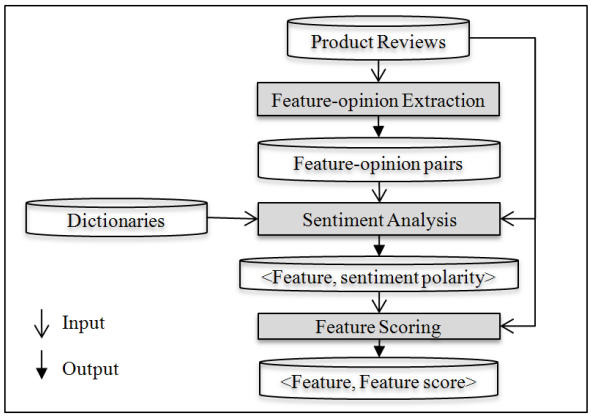

Fig. 4. Product review summarization system.

However, there were two problems encountered when attempting to correctly summarize product reviews through the sentiment analysis and feature scoring steps. First, context-sensitive words needed to be appropriately interpreted during sentiment analysis in order to obtain accurate sentiment polarities of the opinion words. As the sentiment polarity of an opinion word can change depending on the context of a word, conventional approaches, such as PMI or NLP, could not handle the context-sensitive opinion words. Second, during feature scoring, the focus was not on document-level evaluations, but on feature-level evaluations. In most previous work, the distribution of document-level user scores was used to calculate feature scores. Thus, in order to calculate feature scores more accurately, feature-level evaluations were necessary.

To solve these problems, we developed a product review summarization method that can handle context-sensitive words during sentiment analysis and that considers fea-

ture-level evaluations during feature scoring. To handle context-sensitive words, we consider the contexts of the opinion words. In addition, we use a method that utilizes user scores, frequency of opinioned-features, and sentiment polarities of opinion words to increase the accuracy of the feature scores derived during the feature scoring step. Details of this review summarization method are provided in the next section.

4. REVIEW SUMMARIZATION MODEL

In this section, we introduce our novel review summarization model. The entity-relationship diagram for the product review documents is provided (Fig. 5). The figure shows that a particular product may have many (*n*) review documents that have *reviewer*, *review title*, *user score*, and *review date* attributes. Furthermore, each review document may comprise several (*n*) opinioned-features. Within each opinioned-feature is a corresponding opinion word and feature word, and each opinion has a strength indicator and a sentiment polarity. Generally, the feature and opinion are represented by a short phrase or sentence (see example sentences in Fig. 1). Within the product review database that is being processed, we conduct both sentiment analysis and feature scoring for review summarization.

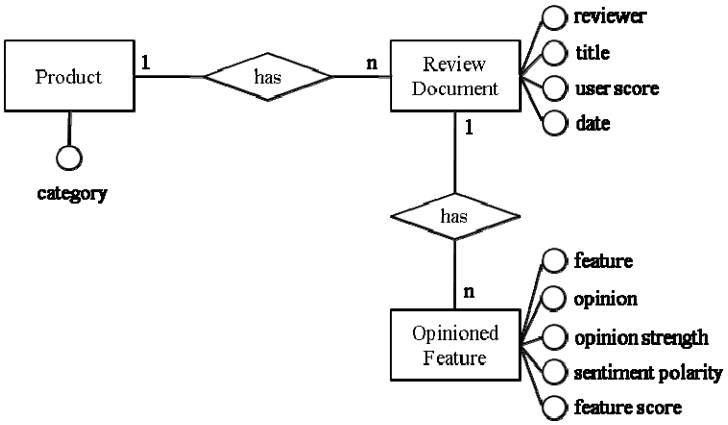

Fig. 5. ER-diagram of product review data.

Fig. 6 illustrates our review summarization model. With a set (*j*) of reviews, a review document R_j contains a user score s_j and a set (*i*) of opinioned-features. Each opinioned-feature comprises a feature word f_{ij} and an opinion word o_{ij} which appear as a pair in a review document. The pair (f_{ij}, o_{ij}) is used as a unit when summarizing review R_j, and strength w_{ij} and sentiment polarity p_{ij} of the pair are derived by sentiment analysis. Opinion strength *w* is based on the frequency of occurrence of an opinioned-feature in a review while sentiment polarity *p* is derived from sentiment classification using the feature-opinion pair and the sentiment dictionary that was constructed from the user scores. The feature score e_{ij} for feature f_{ij} is derived by a feature scoring process from the review's overall user score s_j, the opinion strength w_{ij}, and the sentiment polarity p_{ij}. The

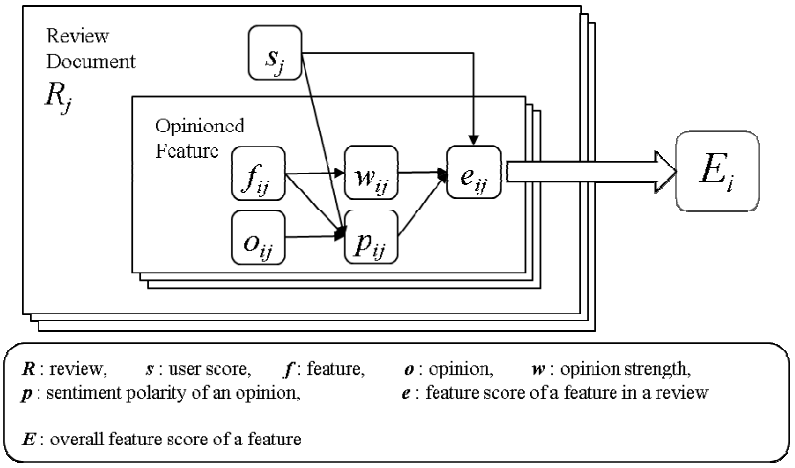

Fig. 6. Review summarization model.

overall feature score E_i of the ith feature within the whole review collection is derived from a feature scoring process using the set of feature scores e for each feature.

Thus, in our model, a review summary is represented as a set of f_i and E_i values where f_i is a feature, and E_i is a feature score. Details on the sentiment analysis and feature scoring methods are discussed in the next section.

5. ANALYTICAL TECHNIQUES

In this section, we explain the methods used for feature scoring and sentiment analysis in the review summarization model. To derive a feature score for a feature various factors extracted from the review documents are used. Both user scores, which are explicit evaluation indices from users and the frequency of occurrence of the opinioned-features within the reviews are used. In addition, opinions on each feature are analyzed and sentiment polarities of those opinions are derived. Thus, in order to obtain the overall feature score of each feature, we use feature frequencies, user scores, and sentiment polarities.

5.1 Preprocessing

The process for product review summarization is shown in Fig. 7. During review summarization at the feature level, feature-opinion pairs are extracted from review documents using part-of-speech (POS) tagging [25]. Through the POS tagging, the most frequent noun words that describe the feature within the set of reviews are selected. Our intuition to extract opinion words is similar to that of [1] and [3]. It is that an opinion associated with a product feature will be mentioned in its vicinity. We use a window of size k, opinion words that are within two words of the selected feature word are extracted. Because a word that is be-verb or preposition may not indicate any special opinion, we exclude these words in word counting. For instance, in the sample sentences of the example 3, ‘is’ and ‘of’ are ignored. The POS tags used for feature and opinion word iden-

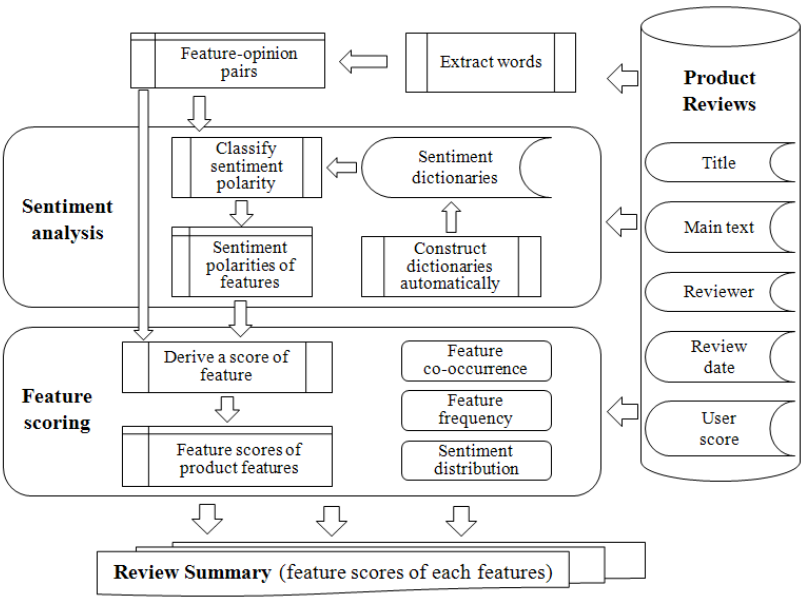

Fig. 7. Review summarization process.

Table 1. Description of POS tags.

Tag	Usage	Description
NN	Feature word	noun, common, singular
NNS	Feature word	noun, common, plural
VBN	Opinion word	verb, past participle
RB	Opinion word	adverb
RBR	Opinion word	adverb, comparative
RBS	Opinion word	adverb, superlative
JJ	Opinion word	adjective or numeral, ordinal
JJR	Opinion word	adjective, comparative
JJS	Opinion word	adjective, superlative

tification are presented in Table 1. Subsequently, we derive sentiment polarities for the opinions on each feature through sentiment analysis. During the feature scoring step, scores for each feature are calculated using sentiment polarities, user scores, opinion strengths, and feature frequencies.

5.2 Summarization

The product review summarization process can be represented as the mean value of the feature scores (E_i) within a set of feature scores (e_{ij}) as in Eq. (2).

$$E_i = \frac{\sum_{j=1}^{n} e_{ij}}{n} \tag{2}$$

where n is the number of reviews that contain the ith feature.

5.2.1 Feature scoring

To obtain the feature score for each feature, the user scores from the set of reviews and the distribution for sentiment polarities of opinions on each feature are used. An example of the distribution of sentiment polarities for a set of features in a set of reviews is shown in Fig. 8. In review R_1 of that figure, the user score is 4 and the sentiment polarities for features f_1, f_2, and f_6 are positive. On the other hand, features f_4 and f_n have negative sentiment polarities. Because the user score represents an overall evaluation composed of various opinions on many features, f_1, f_2, and f_6 may have higher feature scores than the user score while f_4 and f_n scores may be lower. Thus, the original user score of the review needs adjustment when calculating individual feature scores. By considering the interferences among opinions in the same review, we weight the feature score calculations to reflect opinion strength using data on the occurrence of features in the reviews.

	S	f_1	f_2	f_3	f_4	f_5	f_6	f_7	...	f_n
R_1	4	P	P		N		P			N
R_2	5		P	P						P
R_3	4	N			P	P	P			
...										
R_n	2			P		N		N		

$f_1 \sim f_n$: features, $R_1 \sim R_n$: reviews, P: pos. opinion, N: neg. opinion, S: user score of a review

Fig. 8. Example of sentiment distribution.

Eq. (3), which is based on the above approach, is used to calculate the feature score of the ith feature in the jth review.

$$e_{ij} = \begin{cases} s_j + \frac{D}{2} \cdot \frac{N_{negative}(R_j)}{N_{total}(R_j)} \cdot w_{ij}, & \text{if } p_{ij} \text{ is a positive number} \\ s_j - \frac{D}{2} \cdot \frac{N_{positive}(R_j)}{N_{total}(R_j)} \cdot w_{ij}, & \text{if } p_{ij} \text{ is a negative number} \end{cases}, \tag{3}$$

where $N_{total}(R_j)$ is a number of all opinions in the jth review, $N_{positive}(R_j)$ is a number of positive opinions in the jth review, $N_{negative}(R_j)$ is a number of positive opinions in the jth review, and D is the difference between the maximum and minimum user scores. In Eq. (3) instead of directly using a review's user score, the user score is adjusted using the distribution of sentiment polarities and by assigning a weight w to the user score. For example, if the sentiment polarity of an opinion on a specific feature is positive, then we raise the user score assigned to the feature based on the ratio of positive to negative

opinions in the same review. Conversely, when the sentiment polarity of an opinion is negative, the feature score weighting is decreased according to the ratio of positive to negative opinions in the same review. The maximum adjustment is limited in order to avoid bias that may occur when all adjusted user scores have an extreme value. In this work, D is 4 because the maximum user score is 5 and the minimum user score is 1.

During feature scoring, we have assumed that if a feature is mentioned many times in a review document, then the feature had greater opinion strength than other features that are mentioned less often. Therefore, the higher the opinioned-feature frequency, the larger the weight w of that feature. The weighting for opinion strength, w, is calculated using Eq. (4).

$$w_{ij} = \frac{F(f_{ij}, R_j) \cdot N_f(p_{ij}, R_j)}{F_{sum}(p_{ij}, R_j)} \tag{4}$$

where $F(f_{ij}, R_j)$ is the frequency at which the ith feature is mentioned in the jth review, $F_{sum}(p_{ij}, R_j)$ is the sum of the frequency of features that have the same sentiment polarity as p_{ij} in the jth review, and $N_f(p_{ij}, R_j)$ is the number of different features that have the same sentiment polarity as the feature is p_{ij} in the jth review. Using Eq. (4), weight w is more than 0 and less than $N_f(p_{ij}, R_j)$ in order to total sum of adjustments in both positive and negative sentiment has to be zero. For example, if there is only one feature with a given sentiment polarity then w is 1. On the other hand, if there are more than two features which have the same sentiment polarity then the weighting of the stronger opinion approaches $N_f(p_{ij}, R_j)$. Conversely, the weight of the weaker opinion approaches 0.

5.2.2 Sentiment analysis

The sentiment polarity p_{ij} determines whether a user score is increased or decreased is calculated during the sentiment analysis step. To handle cases in which the opinion word is context-sensitive, a PMI value takes into account the context of the words. In previous research that considered PMI values [19], only the opinion word, not its context, was considered. The method can be expressed as $SP = so(o)$, where SP is the sentiment polarity of the opinion word (o), and so is the function that derives the sentiment polarity. This expression is only valid for the common usage of the opinion word and cannot account for sentiment polarity changes when other features are being modified by that opinion word. In the expression $SP = so(f, o)$, even if feature f is being modified by the opinion word o, the calculation still cannot handle context-sensitive words because the sentiment polarity of the opinion word can change according to the type of product (product category) it is modifying. Thus, we have used Eq. (5) to perform sentiment analysis that considers context-sensitive words.

$$SP = s(c, f, o, e) \tag{5}$$

where c is product category, f is the product feature, o is the opinion word, and e is the feature score evaluation of a user for the feature being considered.

By considering the contexts related to a feature (*i.e.*, product category, opinion word,

and user evaluation), more correct sentiment polarities are obtainable. To apply context consideration to sentiment analysis, sentiment dictionaries, constructed automatically from the review data, are used. Fig. 9 illustrates the process of creating the positive and negative dictionaries from the review documents.

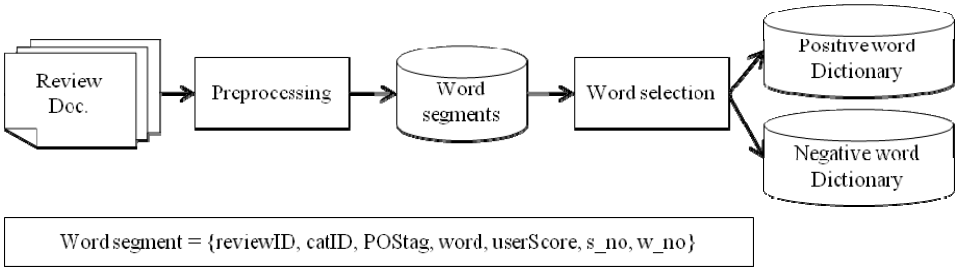

Fig. 9. Building process of sentiment dictionary.

During such dictionary development, word segments are extracted from all review documents through a preprocessing stage. Each word segment includes the context of the word and includes the review identification (ID) number, the product category ID, the word, a POS tag for the word, sentence and word numbers, and the review's user score. Because feature words and opinion words are used for the sentiment classification, we select word segments that have specific POS tags (see Table 1). Following selection, word segments are classified into two dictionaries based on the user score that is associated with the word segment. For example, if a user score is 5 or 4 then the word segment is stored in the positive word dictionary, while word segments associated with user scores of 1 or 2 are stored in the negative word dictionary. The sentiment polarity of opinion o on feature f can be calculated from the difference between the PMI values by using the positive and negative word dictionaries as shown in Eq. (6).

$$\begin{aligned} p_{ij} &= Sentiment\ Polarity(f_{ij},\ o_{ij}) \\ &= PMI_{positive}(f_{ij},\ o_{ij}) - PMI_{negative}(f_{ij},\ o_{ij}) \\ &= \log_2 \frac{P_{positive}(f_{ij}, o_{ij})}{P_{positive}(f_{ij}) \cdot P_{positive}(o_{ij})} - \log_2 \frac{P_{negative}(f_{ij}, o_{ij})}{P_{negative}(f_{ij}) \cdot P_{negative}(o_{ij})} \end{aligned} \tag{6}$$

where p_{ij} is the sentiment polarity of the opinion on the ith feature in the jth review document, $PMI_{positive}$ is the PMI value from the positive word dictionary, and $PMI_{negative}$ is the PMI value from the negative word dictionary, $P_{positive}(x)$ is the probability that x appeared in the positive dictionary and $P_{negative}(x)$ is the probability that x is appeared in the negative dictionary. These two probabilities are the proportions of the number of reviews in which x appeared within the total number of reviews. Using such an approach, context-sensitive words can be considered during the sentiment analysis process.

6. EXPERIMENTS

To determine whether the method can produce accurate summaries of a large num-

ber of reviews, experiments involving both the sentiment analysis and feature scoring methods in our model were performed using existing reviews. Online review data extracted from http://www.epinions.com/ were used for the experiments. For assessment of the sentiment analysis method, we used 16,000 review documents on 1,385 products in the Epinions "mobile phone" and "digital camera" categories (Table 2). When assessing the feature scoring method, we used 17,000 review documents on 422 products in the Epinions "digital camera", "mobile phone", and "hotel" categories (Table 4). To compare the performance of our model with that of existing methods, we used Turney's approach [19] for sentiment analysis and the Red Opal system [7] for feature scoring comparisons. In particular, we focused on the performances of the two methods on context-sensitive opinion words. For the comparison of feature scoring processes, the performance measure was the precision of the overall feature score in comparison with the score in the review data.

Details of the data used in the assessment of the capacity of our model to handle context-sensitive words are shown in Table 2. We extracted 48 and 37 features from the "mobile phone" and "digital camera" categories, respectively, along with their corresponding opinions. In these data, the percentage of context-sensitive words within the opinion word total was approximately 15%. The list of context-sensitive words selected for use in this experiment was:

big, *small*, *high*, *low*, *long*, *short*, *many*, *much*,
few, *several*, *huge*, *tiny*, *heavy*, *light*, *loud*, and *silent*.

Table 2. Review data.

Product category	reviews	positive reviews	negative reviews	Product feature	<*f*, *o*> pair	Context-sensitive word
Hand phone	2947	2196 (74.5%)	418 (25.5%)	48	734	124 (16.9%)
Digital camera	12917	9940 (76.9%)	1740 (23.1%)	37	974	137 (14.1%)

Table 3. Precision of the sentiment classification.

Precision		Mobile phone	Digital camera
Previous method (PMI using Web doc. Search)	All	0.786	0.817
	Context-independent	0.834	0.866
	Context-sensitive	0.508	0.515
Our method	All	0.784	0.764
	Context-independent	0.775	0.758
	Context-sensitive	0.847	0.797
Combined method	All	0.845	0.857
	Context-nonsensitive	0.848	0.88
	Context-sensitive	0.831	0.717

These words were selected manually by 20 experts during the observation on the review documents. On the other hand, for Turney's approach, sentiment polarity was derived from PMI values calculated during a search of the Web documents [19]. As shown in Table 3, the precision for all words for both products was similar in the two

methods. The PMI value based method had a slightly higher precision for context independent opinion words. However, our method produced a markedly higher precision for context-sensitive words in both product categories. We then combined the two methods and re-assessed the data using the mean value of both sentiment polarity values as a new sentiment polarity value in order to reflect the good point of the each method. As a result, the precision increased for all words, context independent and context sensitive words in both products. The results indicate that our combined method performs well in all cases, and that the handling of context-sensitive words is important when classifying the sentiment polarity of an opinion.

Table 4. Experiment data for feature scoring.

	Product (selected randomly)	Feature (selected manually)	Distribution of sentiment polarities in evaluation set	
			Positive (%)	Negative (%)
Mobile phone	222	15	83	17
Digital camera	100	17	73	27
Hotel	100	22	81	19

In the experiment on feature scoring, 54 features and 422 products from 3 product categories were used (Table 4). The 'mobile phone' and 'digital camera' categories were devices, while the 'hotel' is a service. In case of devices, opinions on a product specification are typically mentioned, which can result in objectively-based opinions on many features. However, in service categories, opinions on several kinds of services may be issued in a review and subjectively-based opinions may be more prevalent.

In contrast to general review forms, the Epinions reviews include itemized evaluations of a product's features: with 'Pros' representing good features and 'Cons' indicating poor features. To evaluate the feature scoring performance of our method and that used in Red Opal [7], the 'Pros' and 'Cons' fields within the Epinions reviews were used. If a feature is mentioned in a 'Pros' field then a score of 5 was assigned to that feature. In contrast, if a feature appears in a 'Cons' field then the feature score was 1. The overall score of a feature is its mean value within all reviews.

In the Red Opal system [7], the user score and the feature frequency are used while in our method user scores, opinioned-feature frequency and a distribution of sentiment polarities are used. In order to evaluate the performances of the two approaches, we measured how closely the overall feature scores that are derived by each method match the overall feature scores in the original evaluation set. A match was deemed to occur when the method-derived feature score and the original feature score had the same sentiment polarity. We regarded a feature score over 3 as positive, a feature score under 3 as negative.

The precisions of both scoring methods are presented in Table 5. Our adjustment-based scoring method (see section 5.2.1) resulted in a higher precision than the Red Opal method in the positive, negative, and combined cases of the three product categories tested. In the three categories, our method shows precisions that are 2-12% and 5-13% higher in positive and negative opinioned-features, respectively while within all opinioned-features, our method resulted in 3-12% higher precision than the Red Opal method.

Table 5. Precision of both scoring methods.

Precision	Previous Scoring Method			Our Scoring Method		
	Positive	Negative	Total	Positive	Negative	Total
Mobile phone	0.71	0.27	0.63	0.83	0.32	0.75
Digital camera	0.84	0.17	0.66	0.88	0.30	0.73
Hotel	0.72	0.42	0.67	0.74	0.53	0.70

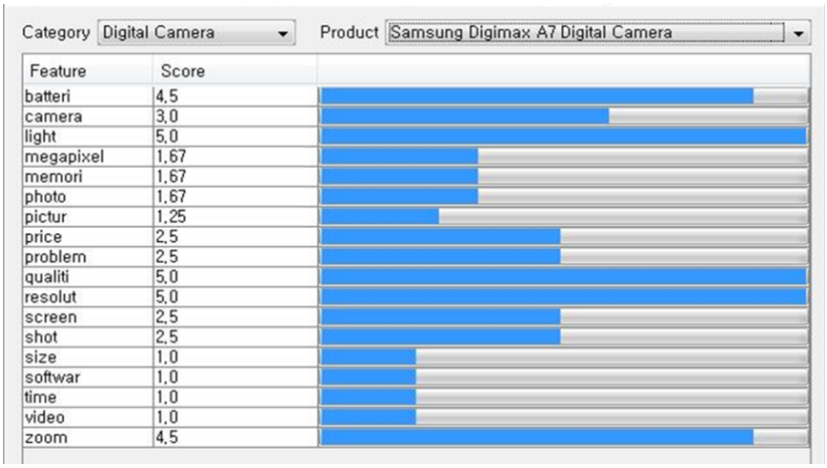

Fig. 10. Screenshot of a review summary.

The results indicate that the sentiment polarity and opinion strength weighting adjustments can effectively calculate a user's feature evaluation. In our approach, because features with positive opinions in a low-scored review and features with negative opinions in a high-scored review are scored according to the opinion for that feature, the errors that may occur when all feature scores are inherited from the overall user score are decreased and precision increases. However, in all three product categories the precision among the negative opinioned-features was lower than that among the positive opinioned-features. This phenomenon may be the result of scoring bias in the review documents because, in many review datasets, the number of reviews with a high user score is typically larger than the numbers of reviews with a low user score [30]. This bias has influence on calculation precision because the feature scoring method is based on the overall user score. Fig. 10 shows a typical result for a multiple featured product using our novel review summarization approach. The summary provides a set of calculated scores for each product feature as well as showing the scores in a bar graph format.

7. CONCLUSION

Here, we present a novel method to summarize individual feature scores from product reviews. In order to provide greater precision than that in existing methods, we analyzed the sentiment polarity of an opinion by considering the review's context-sensitive opinion words. Subsequently, feature scores are obtained by adjusting the original user

scores by the sentiment polarities of the opinion words and by the frequency of the opinioned-feature within the available reviews. We were able to show, through experiments with actual reviews, that our method can accurately perform product review summarization at the product's feature-level. In addition, our method resulted in higher precision than existing methods. To perform additional testing of our summarization approach, we will apply our methods to a set of Korean language product reviews. Further development of the review summarization system will focus on issues that are relevant to the Korean language. In addition, we will assess the impact of reviewer leniency by including this factor in the feature scoring process.

REFERENCES

1. M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2004, pp. 168-177.
2. B. Liu, M. Hu, and J. Cheng, "Opinion observer: Analyzing and comparing opinions on the web," in *Proceedings of the 14th International Conference on WWW*, 2005, pp. 342-351.
3. A. Popescu and O. Etzioni, "OPINE: Extracting product features and opinions from reviews," in *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 2005, pp. 339-346.
4. H. Nakagawa and T. Mori, "A simple but powerful automatic term extraction method," in *Proceedings of International Workshop on Computational Terminology*, 2002, pp. 1-7.
5. S. Sista and S. Srinivasan. "Polarized lexicon for review classification," in *Proceedings of International Conference on Machine Learning*, *Models*, *Technologies and Applications*, 2004, pp. 867-872.
6. K. Dave, S. Lawrence, and D. Pennock, "Mining the peanut gallery: opinion extraction and semantic classification of product reviews," in *Proceedings of the 12th International Conference on WWW*, 2003, pp. 519-528.
7. C. Scaffidi, K. Bierhoff, E. Chang, M. Felker, H. Ng, and C. Jin, "Red opal: Product-feature scoring from reviews," in *Proceedings of the 8th ACM Conference on Electronic Commerce*, 2007, pp. 182-191.
8. X. Ding and B. Liu, "The utility of linguistic rules in opinion mining," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2007, pp. 811-812.
9. J. Liu, G. Wu, and J. Yao, "Opinion searching in multi-product reviews," in *Proceedings of the 6th IEEE International Conference on Computer and Information Technology*, 2006, pp. 25.
10. S. Patwardhan, S. Banerjee, and T. Pedersen, "UMND1: Unsupervised word sense disambiguation using contextual semantic relatedness," in *Proceedings of the 4th International Workshop on Semantic Evaluations*, 2007, pp. 390-393.
11. Y. Matsuo and M. Ishizuka, "Keyword extraction from a single document using word co-occurrence statistical information," in *Proceedings of the 16th International Florida AI Research Society*, 2003, pp. 392-396.

12. M. Gamon, A. Aue, S. Corston-Oliver, and E. Ringger, "Pulse: Mining customer opinions from free text," *Lecture Notes in Computer Science*, Vol. 3646, 2005, pp. 121-132.
13. N. Archak, A. Ghose, and P. G. Ipeirotis, "Show me the money!: Deriving the pricing power of product features by mining consumer reviews," in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2007, pp. 56-65.
14. G. Mishne, "Experiments with mood classification in blog posts," *Stylistic Analysis of Text for Information Access*, 2005.
15. R. Ghani, K. Probst, Y. Liu, M. Krema, and A. Fano, "Text mining for product attribute extraction," *ACM SIGKDD Explorations Newsletter*, Vol. 8, 2006, pp. 41-48.
16. E. Courses and T. Surveys, "Using SentiWordNet for multilingual sentiment analysis," in *Proceedings of the International Conference on Data Engineering Workshop*, 2008, pp. 507-512.
17. D. Lee, O. Jeong, and S. Lee, "Opinion mining of customer feedback data on the web," in *Proceedings of the 2nd International Conference on Ubiquitous Information Management and Communication*, 2008, pp. 230-235.
18. B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 2002, pp. 79-86.
19. P. Turney and M. Littman, "Measuring praise and criticism: Inference of semantic orientation from association," *ACM Transactions on Information Systems*, Vol. 21, 2003, pp. 315-346.
20. T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 2005, pp. 347-354.
21. G. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miler, "Introduction to WordNet: An on-line lexical database," *International Journal of Lexicography*, 1990, pp. 235-244.
22. N. Godbole, M. Srinivasaiah, and S. Skiena, "Large-scale sentiment analysis for news and blogs," in *Proceedings of International AAAI Conference on Weblogs and Social Media*, 2007, pp. 219-222.
23. X. Ding, B. Liu, and P. S. Yu, "A holistic lexicon-based approach to opinion mining," in *Proceedings of International Conference on Web Search and Web Data Mining*, 2008, pp. 231-240.
24. J. Yang, J. Myung, and S. Lee, "The method for a summarization of product reviews using the user's opinion," in *Proceedings of International Conference on Information*, *Process*, *and Knowledge Management*, 2009, pp. 84-89.
25. J. Myung, J. Yang, and S. Lee, "PicAChoo: A tool for customizable feature extraction utilizing characteristics of textual data," in *Proceedings of International Conference on Ubiquitous Information Management and Communication*, 2009, pp. 650-655.
26. J. Yang, J. Myung, and S. Lee, "A holistic approach to product review summarization," in *Proceedings of International Workshop on Software Technologies for Future Dependable Distributed Systems*, 2009, pp. 150-154.
27. J. Myung, D. Lee, and S. Lee, "A Korean product review analysis system using a semi-automatically constructed semantic dictionary," *Journal of KIISE: Software*

and Applications, Vol. 35, 2008, pp. 392-403.
28. C. S. Lee, Z. W. Jian, and L. K. Huang, "A fuzzy ontology and its application to news summarization," *IEEE Transactions on Systems*, *Man and Cybernetics – Part B*, Vol. 35, 2005, pp. 859-880.
29. C. S. Lee, M. H. Wang, and J. J. Chen, "Ontology-based intelligent decision support agent for CMMI project monitoring and control," *International Journal of Approximate Reasoning*, Vol. 48, 2008, pp. 62-76.
30. J. Yang, J. Myung, and S. Lee, "A sentiment classification method using context information in product review summarization," *Journal of Korean Institute of Information Science and Engineers: Databases*, Vol. 36, 2009, pp. 254-262.

Jung-Yeon Yang (梁晶淵) received the B.S. degree in Computer Science and Engineering from Chung-nam National University at Daejeon, Korea, in 2002. Currently, he is a candidate for the degree of Ph.D. of Computer Science and Engineering in Seoul National University and a member of the Intelligent Database Systems Lab. His research interests include opinion mining, databases, intelligent information retrieval, semantic technology, and e-business technology.

Han-Joon Kim (金漢峻) received the B.S. and M.S. degrees in Computer Science and Statistics from Seoul National University, Seoul, Korea in 1994 and 1996 and the Ph.D. degree in Computer Science and Engineering from Seoul National University, Seoul, Korea in 2002, respectively. He is currently an associate professor at the School of Electrical and Computer Engineering, University of Seoul, Korea. His current research interests include data/text mining, databases, machine learning, and intelligent information retrieval.

Sang-Goo Lee (李相求) received his Ph.D. and M.S. degrees in Department of Computer Science from Northwestern University, Illinois, U.S.A., in 1990 and 1987, respectively, and his B.S. degree in Computer Science and Statistics from Seoul National University, Seoul, Korea, in 1985. He is a professor of Computer Science at Seoul National University, Seoul, Korea. He is the Director of Center for E-Business Technology which specializes in R&D in the fields of e-catalogs and Semantic Web services. He also heads the GLOBE R&D Group which specializes on RFID-based methods for logistics applications. His research search interests are in semantic technology, context-aware personalization, e-catalogs, and database design.