7-15-2012

A Rule-Based Approach For Effective Sentiment Analysis

Chin-Sheng Yang
Department of Information Management, Yuan Ze University, Chung-Li, Taiwan, R.O.C., csyang@saturn.yzu.edu.tw

Hsiao-Ping Shih
Department of Information Management, Yuan Ze University, Chung-Li, Taiwan, R.O.C., s986334@mail.yzu.edu.tw

A RULE-BASED APPROACH FOR EFFECTIVE SENTIMENT ANALYSIS

Chin-Sheng Yang, Department of Information Management, Yuan Ze University, Chung-Li, Taiwan, R.O.C., csyang@saturn.yzu.edu.tw

Hsiao-Ping Shih, Department of Information Management, Yuan Ze University, Chung-Li, Taiwan, R.O.C., s986334@mail.yzu.edu.tw

Abstract

The success of Web 2.0 applications has made online social media websites tremendous assets for supporting critical business intelligence applications. The knowledge gained from social media can potentially lead to the development of novel services that are better tailored to users' needs and at the same time meet the objectives of businesses offering them. Online consumer reviews are one of the critical social media contents. Proper analysis of consumer reviews not only provides valuable information to facilitate the purchase decisions of customers but also helps merchants or product manufacturers better understand general responses of customers on their products for marketing campaign improvement. This study aims at designing an approach for supporting the effective analysis of the huge volume of online consumer reviews and, at the same time, settling the major limitations of existing approaches. Specifically, the proposed rule-based sentiment analysis (R-SA) technique employs the class association rule mining algorithm to automatically discover interesting and effective rules capable of extracting product features or opinion sentences for a specific product feature interested. According to our preliminary evaluation results, the R-SA technique performs well in comparison with its benchmark technique.

Keywords: Sentiment Analysis, Opinion Mining, Class Association Rules, Consumer Reviews.

1 INTRODUCTION

The success and explosion of Web 2.0 applications have made social media (e.g., blogs, forums, social networking sites) and user-generated contents tremendous assets for supporting critical business intelligence applications. The knowledge gained from social media can potentially lead to the development of novel services that are better tailored to users' needs and at the same time meet the objectives of businesses offering them. Among the various user-generated contents, online consumer reviews are the critical one. According to a survey of 2,400 U.S. adults (Horrigan 2008), 81% of Internet users employ Web to do research about a product they are considering to purchase. Another survey, of more than 2,000 U.S. Internet users conducted by the comScore and the Kelsey Group in October 2007 (comScore & Kelsey Group 2007), reveals that nearly one-quarter of users (24%) consult consumer reviews prior paying a service and more than three-quarters (79%) of review users report that consumer reviews have a significant influence on their purchase decision. Moreover, for different type of services considered, customers are willing to pay from 20% to 99% more for services receiving an "Excellent" rating than for those receiving a "Good" rating.

Numerous websites have been established for collecting consumer reviews. Such consumer reviews are essential and beneficial for customers, merchants, and product manufacturers. For example, consumer reviews help merchants or product manufacturers better understand general responses of customers on their products for product or marketing campaign improvement. Furthermore, consumer reviews can enable merchants better understand specific preferences of individual customers and facilitates effective marketing decision making. From the perspective of customers, consumer reviews provide valuable information to facilitate their purchase decisions.

As the number of consumer reviews expands rapidly, it becomes difficult for users (e.g., retailers, product manufacturers, customers) to obtain a comprehensive view of consumer opinions pertaining to the products of interest through a manual analysis. Consequently, it is essential and desirable to develop an efficient and effective sentiment analysis (aka opinion mining) technique capable of summarizing the sentiments of consumer reviews automatically. Sentiment analysis involves two main tasks: 1) product feature extraction that extract all product features mentioned in the consumer reviews and 2) opinion orientation identification that determines the sentiments on product features expressed in the consumer reviews (Hu & Liu 2004; Popescu & Etzioni 2005; Pang & Lee 2008; Liu 2010).

Although many studies concentrate on the investigation of sentiment analysis, there are some critical issues that have not been addressed adequately. From the product feature extraction perspective, the identification of implicit product features and categorization (or grouping) of product feature synonyms are essential and not well-studied yet (Liu, 2010). For the opinion orientation identification subtask, using opinion words to determine sentiment categories of extracted product features should properly handle the problems of absence of opinion words (Yang et al. 2011) and context/domain variation (Pang & Lee 2008).

Few studies (e.g., Su et al. 2008; Guo et al. 2009) have attempted to address the abovementioned problems in sentiment analysis but only focus on a single point. For example, Su et al. (2008) and Guo et al. (2009) address the extraction of implicit product features and categorization of product feature synonyms respectively, but do not consider them as a whole and propose an integrated framework for effective sentiment analysis. Consequently, the objective of this study is to design a sentiment analysis technique capable of dealing with most of the abovementioned limitations. Specifically, we propose a rule-based sentiment analysis (R-SA) technique which learns a set of product feature extraction rules for a focal product feature. The R-SA technique takes the description of the product feature f_j (e.g., price, design, battery, etc.) as the only human-provided input and incorporates with the WWW, a set of seed opinion words, and a lexical dictionary to automatically extract opinion sentences which express users' opinions on f_j and then determine the sentiment categories of these opinion sentences.

The remainder of this paper is organized as follows. Section 2 reviews existing sentiment analysis techniques. In Section 3, we depict the detailed design of the proposed R-SA technique. Subsequently, we discuss our experimental design and preliminary evaluation results in Section 4. Finally, we conclude with a summary and some future research directions in Section 5.

2 LITERATURE REVIEW

Sentiment analysis aims to generate review summaries on the basis of product features and sentiments expressed in review sentences. Existing product feature extraction techniques can broadly be classified into two major approaches: supervised and unsupervised (Yang et al. 2009). Supervised product feature extraction techniques require a set of preannotated review sentences as training examples. A supervised learning method is then applied to construct an extraction model, which is capable of identifying product features from new consumer reviews. For example, Wong and Lam (2008) employ hidden Markov models and conditional random fields for extracting product features from auction websites. Yang et al. (2010) adopt two supervised learning algorithms (i.e., class association rules and naïve Bayes classifier) to classify opinion sentences into appropriate product feature classes. Although the supervised techniques can achieve reasonable effectiveness, preparing training examples is time consuming. Additionally, the effectiveness of the supervised approach greatly depends on the quality of the training examples. On the other hand, Yang et al. (2009) adopt the information retrieval approach for product feature extraction. A user has to give a query to express the product feature of interest. Takes as its input the prespecified product feature, the task can be considered as sentence retrieval which identifies review sentences that discuss the target product feature.

In contrast, the unsupervised approaches automatically extract product features from consumer reviews without training examples. For example, the technique developed by Hu and Liu (2004) assumes that product features must be nouns or noun phrases and employs the association rule mining algorithm to extract all product features within a target set of consumer reviews. Wei et al. (2010) extend Hu and Liu’s study by incorporating a set of positive and negative semantic words and three heuristic rules to improve the effectiveness of product feature extraction. In addition to association rule mining, information extraction provides another method commonly employed by the unsupervised product feature extraction approach. For instance, Popescu and Etzioni employ KnowItAll (a Web information extraction system) and propose OPINE to extract product features from consumer reviews automatically (Popescu & Etzioni 2005). Other information-extraction–based product feature extraction techniques have also been proposed (Kobayashi et al. 2005). Their process is similar to that of OPINE, differing only in the representation and construction of extraction patterns and the plausibility test adopted.

Following product features extraction, the opinion orientation identification determines the sentiment categories (i.e., positive or negative) of the extracted product features. The intuitive approach for opinion orientation identification is considered it as sentence-level sentiment classification. However, it is common that a customer may comment on several product features with different semantic orientations in a single review sentence. Alternatively, a lexicon-based approach, which employ a set of opinion words and some linguistic rules to identify the semantic orientations of product features in opinion sentences, is proposed (Hu & Liu 2004; Ding et al. 2008). Specifically, the sentiment category of a product feature in a particular opinion sentence is determined by aggregating the semantic orientations weighted with inverse word distances of surrounding opinion words.

3 THE RULE-BASED SENTIMENT ANALYSIS TECHNIQUE

The overall design of the rule-based sentiment analysis (R-SA) technique is illustrated in Figure 1. The R-SA technique consists of three main phases: 1) product feature extraction rule learning, 2) opinion sentence extraction, and 3) opinion orientation identification. In this study, we accomplish the detailed design of the first two phases and will discuss the preliminary design of the last phase.

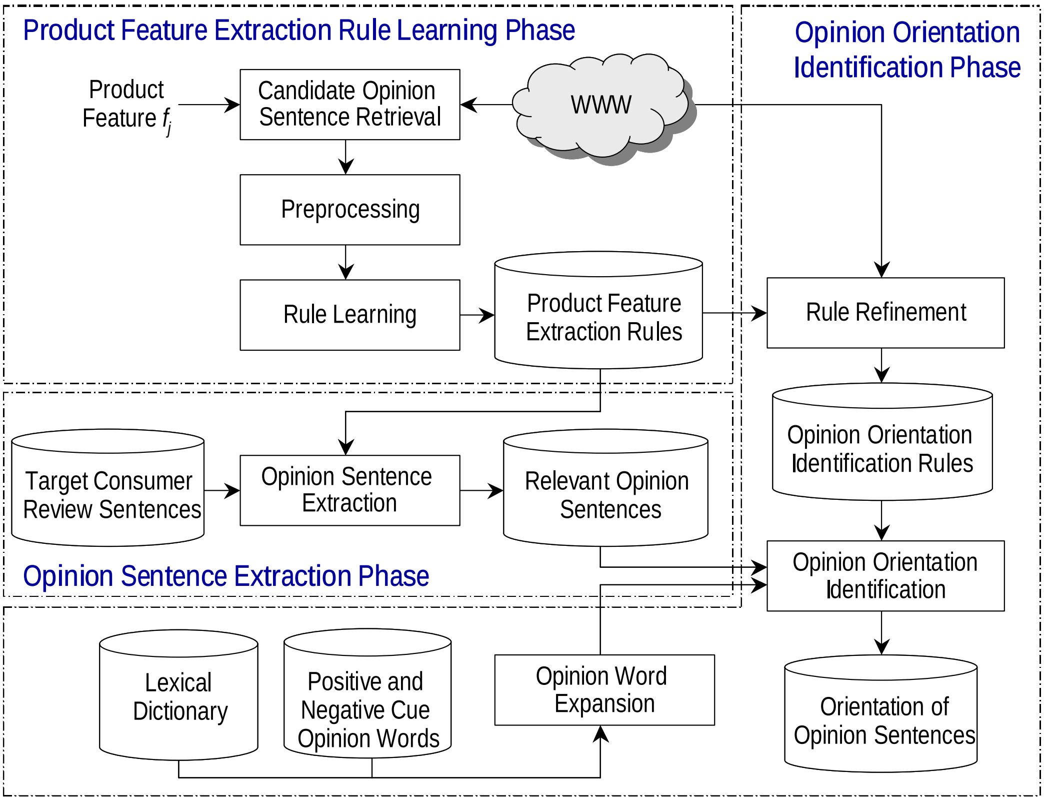

Product Feature Extraction Rule Learning Phase
Product Feature f_j
Candidate Opinion Sentence Retrieval
WWW
Preprocessing
Rule Learning
Product Feature Extraction Rules
Opinion Orientation Identification Phase
Rule Refinement
Opinion Orientation Identification Rules
Target Consumer Review Sentences
Opinion Sentence Extraction
Relevant Opinion Sentences
Opinion Sentence Extraction Phase
Opinion Orientation Identification
Lexical Dictionary
Positive and Negative Cue Opinion Words
Opinion Word Expansion
Orientation of Opinion Sentences

Figure 1. Design of the Rule-based Sentiment Analysis (R-SA) Technique

3.1 Product Feature Extraction Rule Learning Phase

The purpose of this phase is to learn a set of product feature extraction rules which serves as the input of the subsequent opinion sentence extraction and opinion orientation identification phases.

Candidate Opinion Sentence Retrieval:

The purpose of this step is to retrieve some review sentences that are candidate opinion sentences of a particular product feature f_j. First, we employ the WWW as the information source to search a set of consumer reviews *CR* which discuss the focal product type (e.g., digital camera). In this study, we use RateItAll (http://www.rateitall.com/) as our underlying search tool. After the dataset *CR* of the focal product type is collected, we retrieve those sentences, in which keyword f_j (e.g., “battery” of digital camera) occurs, as the candidate opinion sentences (COS_j) of product feature f_j from this dataset. The set of candidate opinion sentences are assumed as review sentences which express opinions on product feature f_j and applied for the subsequent rule learning task.

Preprocessing:

The preprocessing step aims at converting the free-text review sentences into a set of words and, at the same time, enriching their semantic meaning. Three subtasks involved are part-of-speech (POS) tagging, stemming, and meaningful word selection. We employ the TreeTagger (Schmid 1995) for POS tagging and stemming. Subsequently, the meaningful word selection eliminates semantically meaningless words and retains only those helpful for extraction rule learning. Specifically, only nouns, adjectives, adverbs, and verbs are selected for subsequent analysis. Shortly, for each review sentence s_i in *CR*, it is converted into a set of words $W_i = \{w_1, w_2, \ldots, w_{ni}\}$, where n_i is the number of meaningful words in s_i. Moreover, the POS tag $T(w_k)$ of each word w_k in W_i is also attached.

Rule Learning:

The purpose of this step is to induce a set of product feature extraction rules for f_j. Because W_i of each candidate opinion sentence s_i in COS_j always contains the keyword f_i, we can learn a set of extraction rules in the form of $A \rightarrow f_i$, where *A* is a word in the union of words in COS_j. Such a rule implicates that if word *A* appears, the review sentence has a high probability to be an opinion sentence of f_j. Considering the following extraction rules: lithium-ion → battery, mAh → battery, rechargeable → battery; they indicate that if “lithium-ion,” “mAh,” or “rechargeable,” appears in a review sentence, we have high confidence to believe that this sentence contains the opinion of a specific reviewer on the product feature “battery” and should be regarded as an opinion sentence. To learn the desired set of product feature extraction rules from the candidate opinion sentences COS_j automatically, we adopt of the class association rule mining algorithm (Yang et al. 2010) as the underlying technique. Given the prespecified minimum support (*min-supp*) and minimum confidence (*min-conf*), the class association rule algorithm generates a set of product feature extraction rules $PFE\text{-}R_j$ for the product feature f_j.

3.2 Opinion Sentence Extraction Phase

The set of product feature extraction rules $PFE\text{-}R_j$ learned in the previous phase is the only knowledge source to bias the opinion sentence extraction task. Currently, we adopt the simplest and naïve approach for opinion sentence extraction. Specifically, we check whether the left side (i.e., *A*) of any extraction rule appears in a review sentence or not. If true, this sentence is considered as an opinion sentence of f_j. However, the confidences of the extraction rules vary greatly. They should not be treated equally in the opinion sentence extraction process. Furthermore, multiple extraction rules may be applied to a single review sentence. An alternative approach, which is proceeded now, should be designed to take these two critical concerns into consideration.

3.3 Opinion Orientation Identification Phase

This phase aims at determining the opinion orientation of each opinion sentence extracted previously. We do not complete the design of this phase yet but we would like to discuss the general idea of this phase. Our proposed approach integrates the lexicon-based approach (Hu & Liu 2004) and the set of product feature extraction rules $PFE\text{-}R_j$ to accomplish the opinion orientation identification task. Specifically, some extraction rules may be good indicators for opinion orientation identification. Given the observation that adjectives, adverbs, or verbs are usually used to express subjective opinions, they should be useful in determining the opinion orientations. For example, given the extraction rule: “rechargeable → battery,” we have a high confidence to believe that a review sentence satisfies this rule expresses a positive opinion. Because these opinion orientation identification rules ($OOI\text{-}R_j$) are domain-dependent, they may have the ability to mitigate the negative effect of context/domain variation problem. Moreover, unlike traditional approaches in which only adjectives are adopted as opinion words, the proposed technique extends opinion words by including adverbs and verbs and consequently reduces the chance of absence of opinion words.

The words in the left side (i.e., *A*) of the $OOI\text{-}R_j$ are not semantically annotated yet. Accordingly, the rule refinement step concentrates on determining the semantic orientations of the left words of our $OOI\text{-}R_j$. Turney & Littman’s idea (2003) will be adopted. However, the $OOI\text{-}R_j$ may not be sufficient to identify the sentiment categories of all opinion sentences. We design the opinion word expansion step to generate a comprehensive set of positive and negative opinion words and combine it with the opinion orientation identification rules ($OOI\text{-}R_j$) to conduct the opinion orientation identification task.

4 PRELIMINARY EVALUATION RESULTS

This section reports our preliminary evaluation results of the proposed R-SA technique. Specifically, the performance of the opinion sentence identification (aka product feature extraction) is evaluated.

4.1 Data Collection and Evaluation Criteria

We use Yang et al.’s (2009; 2010) dataset to conduct the experiments. This dataset consists of consumer reviews of digital camera collected from Amazon.com. Particularly, there are 3,000 review sentences, concerning eight product features. 149, 83, 273, 241, 95, 156, 102, and 67 review sentences are identified to discuss Battery, Flash, Image, Lens, Memory, Price, Screen, and Video of digital cameras, respectively. We collect another set of consumer reviews, which consists of 442,509 unannoated review sentences of digital camera, from RateItAll for product feature extraction rule learning.

Moreover, the well-known F_1 measure is adopted as evaluation criteria. The F_1 measure with respect to a specific product feature f_j is the harmonic average of its precision and recall, defined as $(2 \times P_j \times R_j)/(P_j+R_j)$, where $P_j=TP_j/(TP_j+FP_j)$, $R_j=TP_j/(TP_j+FN_j)$, TP_j is the number of opinion sentences of f_j correctly identified as opinion sentences, FP_j is the number of non-opinion sentences incorrectly identified as opinion sentences, and FN_j is the number of opinion sentences of f_j incorrectly identified as non-opinion sentences. Since our dataset covers eight product features, the macro- and micro-measurements (Yang et al. 2010) are employed to estimate the overall performance among the eight product features. The macro- precision, recall, and F_1 measures among a set of product features F are defined as:

$$\textit{Macro-P} = \frac{\sum_{f_i \in F} P_j}{|F|},\ \textit{Macro-R} = \frac{\sum_{f_i \in F} R_j}{|F|},\ \text{and } \textit{Macro-F}_1 = \frac{2 \times \textit{Macro-P} \times \textit{Macro-R}}{\textit{Macro-P}+\textit{Macro-R}}.$$

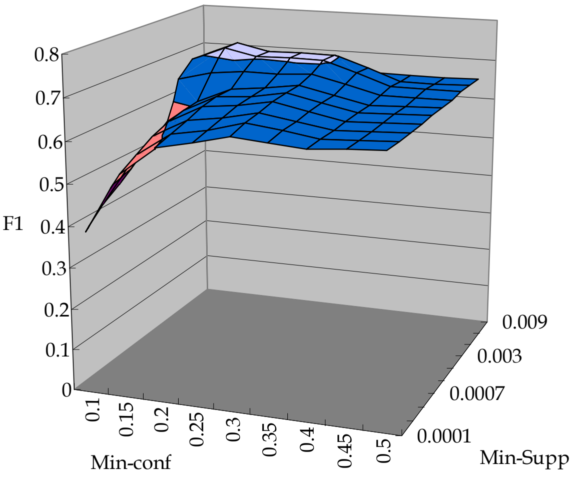
On the other hand the micro- precision, recall, and F_1 measures are defined as:

$$\textit{Micro-P} = \frac{\sum_{f_i \in F} TP_j}{\sum_{f_j \in F} TP_j + \sum_{f_j \in F} FP_j},\ \textit{Micro-R} = \frac{\sum_{f_i \in F} TP_j}{\sum_{f_j \in F} TP_j + \sum_{f_j \in F} FN_j},\ \text{and}$$

$$\textit{Micro-F}_1 = \frac{2 \times \textit{Micro-P} \times \textit{Micro-R}}{\textit{Micro-P}+\textit{Micro-R}}.$$

4.2 Evaluation Results

The values of minimum support (*min-supp*) and minimum confidence (*min-conf*) should have great effects on the effectiveness of the proposed R-SA technique. Thus, we examine the effects of these two parameters. We investigate the *min-supp* at 0.0001, 0.0003, 0.0005, 0.0007, 0.0009, 0.001, 0.003, 0.005, 0.007 and 0.009 and *min-conf* at 0.1 to 0.5 in increments of 0.05. The macro- and micro- F1 measures with different min-supp and min-conf values of the proposed R-SA technique are shown in Figure 2 and Figure 3 respectively. The best macro- and micro- F1 measures were 72.04% and 68.96% respectively. The performance of our R-SA technique is slightly worse than Yang et al.'s (2010) study (i.e., 72.26% and 70.77% for macro- and micro- F1 measures). However, the proposed R-SA technique does not need the manual preparation of training data for each product feature in each product type. Moreover, the R-SA technique is relatively stable to the values of min-supp and min-conf than Yang et al.'s (2010) technique. Overall, we believe that the cost, i.e., slight decrease of F1, of the R-SA technique is reasonable and acceptable. In addition, the R-SA technique can indeed discover some meaningful and interesting product feature extraction rules automatically. Table 1 shows example rules for product feature "battery."



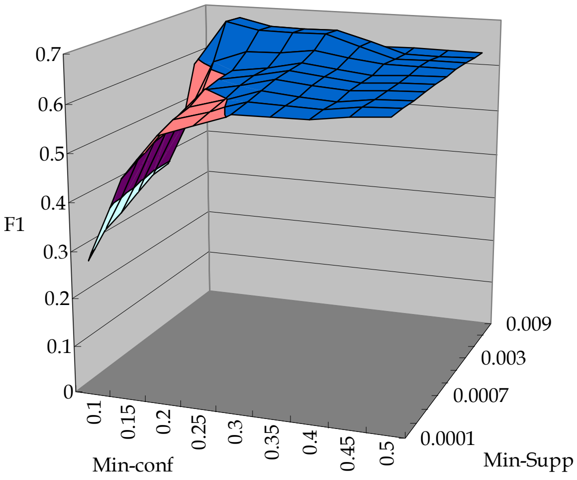


Figure 2. Macro- F1 Measure with Different Min-supp and Min-conf Thresholds

Figure 3. Micro- F1 Measure with Different Min-supp and Min-conf Thresholds

Table 1. Example Product Feature Extraction Rules for "Battery" in Digital Camera

Left Side of Rules	Support	Confidence	Left Side of Rules	Support	Confidence
Rechargeable	0.006	0.906	AA	0.006	0.839
Drain	0.002	0.877	Life	0.011	0.826
Lithium	0.002	0.869	NiMH	0.003	0.819
Alkaline	0.002	0.867	Charger	0.005	0.765

5 CONCLUSION AND FUTURE RESEARCH DIRECTIONS

We propose the rule-based sentiment analysis (R-SA) technique which learns a set of effective rules for product feature extraction without the need to prepare training data manually. The evaluation results reveal that the R-SA technique achieves comparable performance with a fully supervised approach in which the manual preparation of training data is critical. As a result, the R-SA technique is more practical in the world where there are many product types and each of which has diverse product features. Some ongoing and future research directions are discussed briefly. First, only some preliminary experiments are conducted in this study. It is essential to perform some more in-depth analysis of the R-SA technique. Second, as mentioned previously, the design of an alternative approach which considers the number and importance of rules for opinion sentence extraction phase is critical. Third, detailed design and empirical experiments of the proposed opinion orientation identification are desirable. Last but not least, additional empirical evaluation involving more product types from diverse information sources is one of our future research directions.

ACKNOWLEDGEMENT

This work was supported by the National Science Council of the Republic of China under the grant NSC 100-2410-H-155-013-MY3.

References

comScore and The Kelsey Group (2007). Online consumer-generated reviews have significant impact on offline purchase behaviour. http://www.comscore.com/press/release.asp?press=1928.

Ding, X., Liu, B., and Yu, P.S. (2008). A holistic lexicon-based approach to opinion mining. In *Proceedings of the International Conference on Web Search and Web Data Mining (WSDM'08)*, 231-239.

Guo, H., Zhu, H., Guo, Z., Zhang, X., and Su, Z. (2009). Product feature categorization with multilevel latent semantic association. In *Proceeding of the 18th ACM Conference on Information and Knowledge Management (CIKM'09),* 1087-1096.

Horrigan, J. B. (2008). Online shopping. Pew Internet & American Life Project. http://www.pewinternet.org/Reports/2008/Online-Shopping.aspx

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD'04),* 168-177.

Kobayashi, N., Iida, R., Inui, K., and Matsumotto, Y. (2005). Opinion extraction using a learning-based anaphora resolution technique. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCNLP-05)*, 173-178.

Liu, B. (2010). Sentiment analysis and subjectivity. In *Handbook of Natural Language Processing*, (Indurkhya, N. and Damerau, F.J. Eds.), NY, Chapman and Hall/CRC.

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2 (1-2), 1-135.

Popescu, A. and Etzioni, O. (2005). Extracting product features and opinions from reviews. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP 2005),* 339-346.

Schmid, H. (1995). Improvements in part-of-speech tagging with an application to German. In *Proceedings of the ACL SIGDAT-Workshop*.

Su, Q., Xu, X., Guo, H., Guo, Z., Wu, X., Zhang, X., and Swen, B. (2008). Hidden sentiment association in Chinese Web opinion mining. In *Proceeding of the 17th International Conference on World Wide Web (WWW'08),* 959-968.

Turney, P.D. and Littman, M.L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21 (4), 315-346.

Wei, C., Chen, Y.M., Yang, C.S., and Yang, C.C. (2010). Understanding what consumers concern: A semantic approach for product feature extraction from consumer reviews. *Information Systems and E-Business Management*, 8 (2), 149-167.

Wong, T.L. and Lam, W. (2008). Learning to extract and summarize hot item features from multiple auction Web sites. *Knowledge and Information Systems*, 14 (2), 143-160.

Yang, C.C., Tang, X., Wong, Y.C., and Wei, C. (2010). Understanding online consumer review opinions with sentiment analysis using machine learning. *Pacific Asia Journal of the Association for Information Systems*, 2 (3), 73-89.

Yang, C.S., Wei, C., and Yang, C.C. (2009). Extracting customer knowledge from online consumer reviews: A collaborative-filtering-based opinion sentence identification approach. In *Proceedings of the 11th International Conference on Electronic Commerce (ICEC'09)*, 64-71.

Yang, C.S., Chen, C.H., and Chang, P.C. (2011). Improving the effectiveness of opinion orientation identification in sentiment analysis of consumer reviews: A gloss-based approach. In *Proceedings of the 5th China Summer Workshop on Information Management (CSWIM 2011).*