International Journal of u- and e- Service, Science and Technology
Vol.9, No. 11 (2016), pp.95-104
http://dx.doi.org/10.14257/ijunesst.2016.9.11.09

# Automatic Summarization for Agricultural Product Review

Qingxi Peng[1, a], Qisheng Lu[2,b] and Ling Shen[1,c]

[1]*Computer School, Wuhan Donghu University*
[2]*Computer School, Huazhong University of Science and Technology*
[a]*qingxipeng@hotmail.com;*[b]*592973941@qq.com;*[c]*aleenapple@163.com*

***Abstract***

*Agricultural product review is playing increasing role in finance websites. Since manual text summarization needs large human efforts and also time consuming, we propose automatic summarization for agricultural product review in this paper. Firstly, we formulate the agricultural product as a quintuple, and present a framework to extract the features from agricultural product. Then we exploit learning to rank method to identify key sentence in product review. Finally, sentence analysis method has been proposed to extract comparative information. Experimental results show that the proposed approaches outperform the existing research.*

***Keywords****: Automatic Summarization, Agricultural Product Review, Feature Extraction, Sentence Analysis*

## 1. Introduction

In recent years, governments and organizations usually collect agricultural information and publish to the public on websites. The information includes production, process, circulation and sales of the agricultural product. Researchers rely on the available information to analyze the production, and forecast the price. Nevertheless, the reviews grow rapidly, which make it difficult for the potential customer to read the reviews carefully to make decision. Processing the review manually need large human effort. Even if people can utilize it manually, it is an exhausting work.

Previous studies mainly focus on the experts' experience to acquire knowledge from the agricultural product review. They usually collect the financial data from the trading market, and process it by mathematical methods such as regression and time series. If the public want to acquire the information, they usually get it from the experts. Otherwise, they will spend a lot of time to browse the agricultural product review. On the other hand, the useful information spread out in many financial websites. Without professional approaches, they can hardly been collected and processed. To our knowledge, Machine Learning methods have never been employed in this domain.

In this paper, we introduce several machine learning methods for agricultural product review analysis. Firstly, we formulate the agricultural product review as a quintuple tuple, which includes important features to depict the agricultural product review. Then a framework for agricultural product features identification has been proposed. In this framework, we present a bootstrapping method, which identify the agricultural product features with a set of feature seeds. Secondly, we exploit sentence analysis approaches to handle the agricultural product analysis. Key sentence and comparative sentence analysis are two effective methods in opinion mining and sentiment analysis. In this paper, we exploit these two methods in agricultural product review mining. Agricultural product review is different from common text mining and online product review. Common online text is longer than

ISSN: 2005-4246 IJUNESST

agricultural product review, while online product review is usually posted by customers. Agricultural product review usually posted by administrative department, and the review text is usually short. Therefore, different methods have been combined to settle the problem.

On the whole, there are three contributions of our works:

We formulate the agricultural product review, and propose a framework to summarize the agricultural product review. We also propose a hybrid algorithm which can extract features by means of a set of explicit feature seeds and lexicons.

We present a key sentence identification problem in agricultural product review summarization. Then a learning to rank method is proposed to handle the problem.

We also propose a comparative sentence analysis method which can identify the comparison information in the agricultural product.

To our knowledge, this is first study exploring agricultural product review by machine learning methods. The remainder of the paper is organized as follows. In Section 2, we first review the related works about automatic summarization. The techniques mentioned in this paper will also be introduced in this section. In section 3, several methods are proposed. Firstly, a framework will be proposed to extract the features of agricultural products. Secondly, two sentences analysis methods are shown to handle the key sentence and comparative sentence analysis. In section 4 we present the experiments to demonstrate the efficiency of proposed methods. Finally, conclusions and directions for future work are given in section 5.

## 2. Related Works

Automatic text summarization has been an important research in information retrieval. Many methods have been proposed to solve the problem. Early researches mainly focus on document level [12]. In recent year, feature extraction attracts researchers. In this paper, feature means the characteristics of product. Feature extraction is more difficult than document level summarization [10].

Agricultural product review is a new media wait us to mine. Even if enough mathematical tools have been devoted into agriculture analysis, text mining methods have never been used in this domain. Other domain such as stock, news, online reviews exploited text mining methods [1-2] and [13]. Inspired by this situation, we exploit text mining and machine learning methods to solve the problem in agricultural product review analysis.

The machine learning methods used in this paper mainly include bootstrap method and learning to rank. Bootstrap method is a resampling procedure [3], and it involves resampling from original data set. Learning to rank refers to machine learning techniques for training a model in a ranking task. It constructs a ranking model using training data, and the model can sort the document according to their degrees of relevance or preference [7].

## 3. The Proposed Method

Our goal in this paper is to summarize the agricultural product review by means of machine learning methods. In the first place, we give a formulation to the agricultural product review. Figure 1, is a piece of agricultural product review. It is translated from www.100ppi.com, which is an authority website mainly report international and domestic agricultural products.

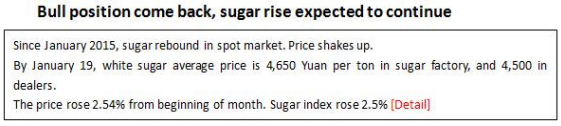


**Figure 1. Agricultural Product Review**

 

### 3.1. Agricultural Product Review Formulation

We use agricultural product review listed in Figure 1, to introduce the problem. From this review, we notice several important points:

1 The agricultural product review has several features: product name, time, place, price and trend. The basic step of automatic summarization is feature extraction.

2 Different sentences have different weights in review. Sentence (1) is the topic of the review. Sentence (2) shows the place and time of the product. Sentence (3) express the price and predict the trend.

3 Comparison information has been expressed in review. The comparison information mainly include features such as times and places.

4 The author usually gives some prediction of the price. In this review the price of sugar is expected to going up.

We are now ready to define agricultural product review as a quintuple.

Definition (Agricultural Product review): An agricultural product review is a quintuple, (*e, pl, pr, t, tr*), where *e* is the agricultural product, *pl* is the place the review mentioned, *pr* is the price of the agricultural product, *t* is the time the review mentioned, *tr* is the price trend that the author predicted.

This definition may be difficult to apply in agricultural product review, since the features may scattered in the review and difficult to extract correctly. Firstly, all the sentences in the review mentioned about the agricultural product. It is difficult to decide which sentence is important. Secondly, it seems that some sentences are redundant in the review. For example, price has been mentioned three times in figure 1. If we analyze it deeply, we will find that they present different meanings. So in this review, it is important to determine what the key meaning is. Thirdly, there are many comparative sentences in the review. Some of them are time comparison such as this year and last year, some of them are place comparison such as international and domestic, others are product features such as mature and immature.

We employ different machine learning methods to solve the problem. To extract the features from the review, we propose a hybrid feature extraction framework. The framework exploits lexicon and a bootstrapping method to extract the features iteratively. Then we will present a key sentence classification method to identify important meaning of the review. Finally, we also employ a comparative sentence analysis method to identify the useful information from the review.

### 3.2. Agricultural Product Summarization Framework

In this section, we propose a hybrid summarization framework. In this framework, we propose a bottom-up method which extract features individually and assemble it to a quintuple mentioned above. The framework consists of four levels: word level, feature level, sentence level and review level. In word level, a bootstrapping algorithm is proposed to extract the features. When the features have been collected, we combine it to quintuples in feature level. Then a key sentence identification method has been proposed to find out which sentence is important. Next step, since there are very much comparative sentences in the reviews, we employ a comparative sentence analysis method to determine the exact meaning of the sentence. Figure 2 shows the proposed framework.

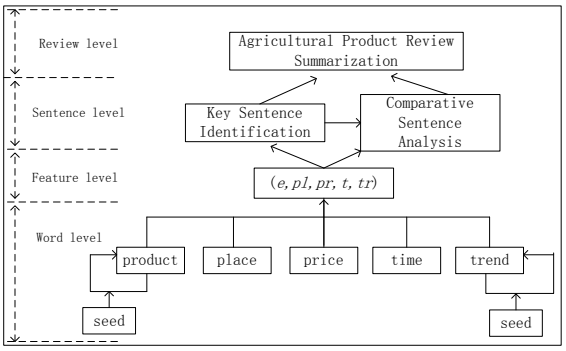


**Figure 2. Agricultural Product Summarization Framework**

We propose a hybrid feature extraction algorithm by exploiting lexicon and bootstrapping method to identify features of agricultural product review. We have observed that some features such as place and time can be predefined, while other features such as product name and price trend cannot be predefined. Although some features are listed explicitly in the agricultural product reviews, there are also many important features ignored by the webpage. For example, the review website may list chicken, egg as agricultural product. Some Chinese herbal medicines such as astragalus and gastrodia elata, however, are seldom listed. For those features that usually been talked about, we utilize the lexicon to extract it. In summary, we exploit a hybrid feature extraction algorithm to handle the agricultural product review extraction problem.

Some known agricultural product names was chosen to be seeds. Then we propose a boot-strapping method to extract the implicit features iteratively. An important observation is that different features in agricultural product modified each other. Therefore, some features can be inferred from other known features. For example, we analyze the second sentence on the figure1. -"By January 19, white sugar average price is 4,650 Yuan per ton in sugar factory, and 4,500 in dealers." Features such as "white sugar" and "price" occur on different sentences. If other features have been identified, such features can be identified. Our algorithm has been listed below.

Algorithm: Hybrid Feature Extraction Algorithm

Input: A collection of agricultural product reviews $\{d_1, d_2 \ldots d_{|D|}\}$, lexicon of place *PL* and time *T*, subset of known product name $F\{F_1, F_2 \ldots F_k\}$, feature distance *D*, selection threshold *p* and iteration step limit *I*

Output: Set of whole review content r(e, pl, pr, t, tr) $\in$R

Step 0: Split the reviews into sentences, $X = \{x_1, x_2 \ldots x_n\}$, build a word list *L*

Step 1: Check each sentence in X; remove unimportant words; add words W to L

Step 2: Match the word W in each sentence X with place lexicon PL and time lexicon T, record the position

Step 3: Find other word *W'* from *W* with feature distance *D*, feature i weight match number *Count (i)*

Step 4: Determine the feature with weight ai= argmaxi Count(i)

Step 5: Rank the feature words with respect to their IG value and join the top p feature words for each feature into their corresponding feature word list L(F)

Step 6: If the feature word list is unchanged or iteration exceeds I, go to Step 7, else go to Step1

Step 7: Combine the feature set to r(e, pl, pr, t, tr) $\in$R with respect to feature order

Step 8: Output the quintuple $r(e, pl, pr, t, tr) \in R$

In the beginning, we get a few agricultural product names such as pork, egg,

 

potato, onion, apple, *etc*., With our algorithm, we needn't care about unusual agricultural product names, since the algorithm will extract it by bootstrapping method. In our experiment, almost all the agricultural product names have been identified. Section 4 will show our experimental result.

### 3.3. Key Sentence Analysis

There are four sentences in figure 1 including review title. It is shown that there is different importance in each sentence. It is obviously that the first sentence (review title) and last sentence are more important than other sentences. This type of sentence is key sentence. In this section, we try to identify the key sentence in the agricultural product review. We employ a ranking based method to identify the key sentence. We also use annotation mentioned in hybrid feature extraction algorithm. Inspired by [13], we employ sentiment, position and keyword as features to identify the key sentence. But our method is different from [13] since our method is a machine learning method while [13] is score based method of three features.

Firstly, key sentence mainly represent the overall sentiment of author. Generally speaking, the sentiment and preference are expressed by sentiment words. Therefore, the main orientation of review can be determined by calculating the sentiment words. The positive and negative sentiment of one sentence can be calculated by Equation 1 and Equation 2 respectively.

$$f\_positive(x_i) = \frac{\left|\sum_{j=1}^{n} positive(w_{ij})\right|}{n} \tag{1}$$

$$f\_negative(x_i) = \frac{\left|\sum_{j=1}^{n} negative(w_{ij})\right|}{n} \tag{2}$$

We download sentiment lexicon from online sentiment (http://ir.dlut.edu.cn/). We also observe that the position is also very important for the sentence. It is shown that the beginning and end of review play an important role in a review. A parabola equation has been proposed to identify the key sentence.

$$f\_position(x_i) = \alpha \times pos(x_i)^2 + \beta \times pos(x_i) + c \tag{3}$$

We also collect the keywords such as overall, oiwa market, etc to build a keyword lexicon. The keywords are counted to show the importance of the sentence.

We employ ranking SVM in this section to handle learning to rank. The agricultural product reviews are split to sentences, and the sentences are annotated according to importance. Point wise method has been utilized to rank the sentences. Finally, the sentences are utilized ordering by importance in agricultural product review summarization.

### 3.4. Comparative Sentence Analysis

If we study carefully on agricultural product review, an interest question will attract us. There are many comparative sentences in the reviews. For example, this is a review content "Cotton Price is current level at 1.485, down from 1.506 last month and down from 2.005 one year ago." This review expresses the comparative relation: (cotton, down, {today, last month}), (cotton, down, {today, one year ago}). Comparative sentence was firstly proposed as a concept in [6]. In this paper, comparative sentence is different from [6]. Comparative relations are usually products. However, in this paper, a review mainly report just one agricultural product. The comparative relations are different times and places of this product.

To some extent, our solution depends on sequential rules. In 3.3 we have extracted quintuple r(e, pl, pr, t, tr), which means that the product name, place, price, time and trend have been identified. Now we extend the scope. The example mentioned above give us an example.

*(r(cotton, null,1.485 , now, null), down, r(cotton, null,1.506 , last month, null) )*

*(r(cotton, null,1.485 , now,* null*), down, r(cotton, null,2.005 , last year, null) )*

Actually, there are many missing data in example mentioned above. We use sequential rules to infer the missing data.

There are three steps in comparative sentence analysis.

(1) Classification based method to identify the comparative sentence. We collect agricultural product reviews to train classification model. We manually annotate the reviews as comparative sentences and non-comparative sentences. SVM has been employed to classify the data.

(2)Wrapper based method to extract features from comparative sentence. If one comparative sentence is identified, we use part-of-speech tagging method to build wrapper. Then the wrappers are utilized to extract the features.

(3)Linear interpolation can infer the missing data. Many features are not mentioned but implied in comparative sentence. In this step, missing features are filled based on wrapper generated in step 2.

With three steps mentioned above, the comparative sentences are identified, and the comparative relations are also extracted.

## 4. Experiments

In this section, experiments are presented to demonstrate the efficiency of the proposed methods. The agricultural product reviews are extracted from www.100ppi.com, www.cnhnb.com, and pfscnew.agri.gov.cn. We choose about one to two months reviews as datasets. The stopwords are erased in the dataset. Then we use Porter Stemming tool to stem all the word. Table 1, is the statistics of the data.

**Table 1. Statistic of Information for Datasets**

| Data set | Dataset 1 | Dataset 2 | Dataset 3 |
|---|---|---|---|
| Number of reviews | 1200 | 860 | 520 |
| Sentences per review | 7.5 | 14.3 | 18.7 |

### 4.1. Feature Identification Result

In section 3.2, hybrid feature extraction algorithm has been proposed to identify the feature. In this algorithm, we firstly give some feature seeds. Since there are differences in datasets, we give different feature seeds in Table 2.

**Table 2. Feature Seeds in Datasets**

| | |
|---|---|
| Seed for Dataset 1 | wheat, rape, egg, soybean, mung, cotton, corn, isatis radix, rice |
| Seed for Dataset 2 | spinach, garlic, milk, wheat, walnut, medlar, pineapple, lotus nut |
| Seed for Dataset 3 | mushroom, tea, fish, scallop, crucian, cabbage, pear, watermelon |

Two feature extraction methods are chosen to be baseline. Ibrahim (2012) proposes an automated approach for constructing feature model from available product descriptions found in online repositories has been proposed. We name this method as snowball. Another baseline is [4], an algorithm named opinion ranking (OR) has been proposed to analyze the overall sentiment of review, and rank the

 

sentiment strength. We use precision to compare the result our method and those of the baselines. Our method named HFE (hybrid feature extraction) algorithm.

From Figure 3, we can see that HFE algorithm outperform other algorithms in three dataset. With predefined feature seeds, HFE algorithm acquire above 80% precision. All feature extraction get low precision in dataset 3 since reviews in dataset3 have relative long text, and it seems complex than other datasets.

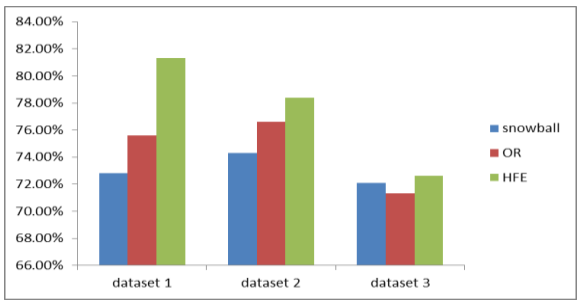


**Figure 3. Precision of Feature Extraction**

### 4.2. Key Sentence Analysis Result

Since there isn't another research focus on key sentence, we only use [13] as baseline to compare with our method. Zheng (2012) use a score based method, while our method employs learning to rank to identify key sentences. We use NDCG as measure to compare two methods. NDCG is a popular measure in information retrieval. It is defined as follows:

$$NDCG = \frac{DCG}{DCG\ of\ \ ideal\ ranked\ list}$$

Where DCG is defined as:

$$DCG = rel_i + \sum_{i=2}^{p} \frac{rel_i}{\log_2 i}$$

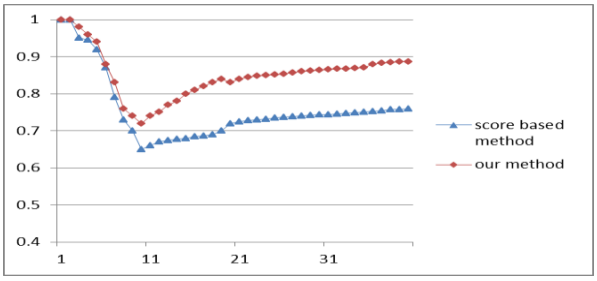


**Figure 4.1. NDCG Comparisons in Dataset 1**

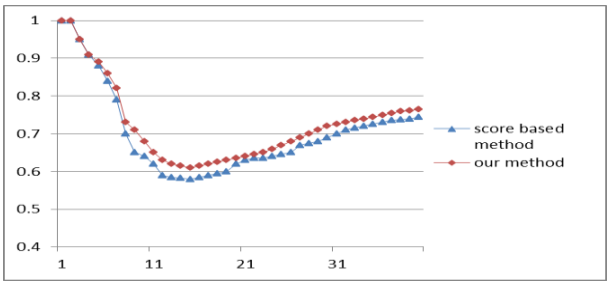


**Figure 4.1. NDCG Comparisons in Dataset 2**

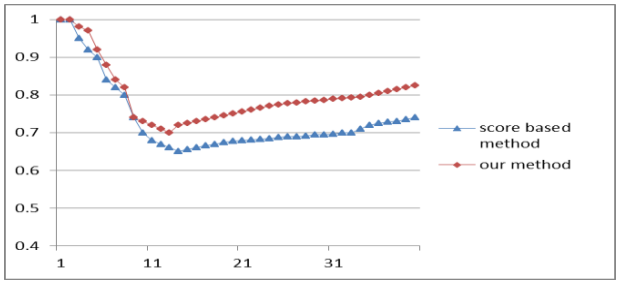


**Figure 4.1. NDCG Comparisons in Dataset 3**

NDCG comparison in different datasets is shown in Figure 4. From Figure 4, we can see that our method outperform score based method. Our method employs machine learning techniques which make its precision is higher than score based method.

### 4.3. Comparative Sentence Analysis Result

To validate the effectiveness of our method, we use method in [6] as baseline to compare with our method. Since method in [6] doesn't assemble the features to quintuple, we only compare the comparative sentence identification. Note that our method not only identifies the comparative sentences, but also assembles the features to quintuple. This is an important difference between method in [6] and ours.

We use Stanford POS Tagger in our experiment. The method in [6] has many steps to perform the comparative sentence identification. We collect reviews in our datasets, and then use the baseline method and our method to identify the comparative sentences. SVM has been used to classify the sentences. All the results are obtained through 10-fold cross discussions.

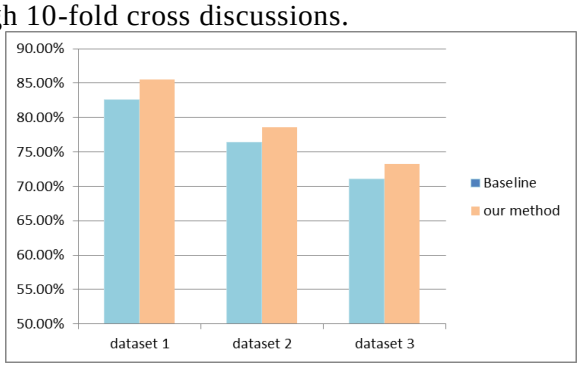


**Figure 4.1. Comparison of Classification Precision**

The experimental result has been shown in Table 3. Note that the precision is lower than that mentioned in [6], since agricultural product reviews is difficult than ordinary product reviews. In Figure 4, our method surpasses the baseline. The main reason is that Jindal (2006) ignore the inherent characteristic of agricultural product reviews.

## 5. Conclusion

In this paper, the summarization of agricultural product reviews has been discussed. The agricultural product review has been formulated as a quintuple, and a hybrid feature extraction algorithm has been proposed to identify the features. Key sentence and comparative sentence analysis approaches have been proposed to solve

 

the important problems. Experimental results demonstrate the efficiency of proposed method.

## Acknowledgments

This work was supported by the grants from Hubei Provincial Collaborative Innovation Centre of Agricultural E-Commerce (under Construction) (Wuhan Donghu university research [2014] No.4)

## References

[1] M. Acher, "On extracting feature models from product descriptions", Proceedings of the Sixth International Workshop on Variability Modeling of Software-Intensive Systems. ACM, **(2012)**
[2] J. Bollen, H. Mao and X. Zeng, "Twitter mood predicts the stock market", Journal of Computational Science, vol. 2, no. 1, **(2011)**, pp. 1-8.
[3] M. R. Chernick, "Bootstrap methods: A guide for practitioners and researchers", John Wiley & Sons, vol. 619, **(2011)**.
[4] M. Eirinaki, S. Pisal and J. Singh, "Feature-based opinion mining and ranking", Journal of Computer and System Sciences, vol. 78, no. 4, **(2012)**, pp. 1175-1184.
[5] R. Ibrahim, N.Salehi Dastjerdi and S. H. Ghorashi, "Product feature extraction using natural language processing techniques, Journal of Computing, vol. 4, no. 7, **(2012)**, pp. 39-43.
[6] N. Jindal and B. Liu, "Mining comparative sentences and relations", AAAI, vol. 22, **(2006)**.
[7] H. Li, "Learning to rank for information retrieval and natural language processing", Synthesis Lectures on Human Language Technologies, vol. 7, no. 3, **(2014)**, pp.1-121.
[8] L. Zheng, T. Songbo and C. Xueqi, "Sentiment classification analysis based on extraction of sentiment key sentence", Journal of Computer Research and Development, vol. 49, **(2012)**, pp. 2376-2382
[9] T. Y. Liu, "Learning to rank for information retrieval", Springer Science & Business Media, **(2011)**.
[10] A. M. Popescu and O. Etzioni, "Extracting product features and opinions from reviews", Natural language processing and text mining. Springer London **(2007)**, pp. 9-28.
[11] S. Gerard, "Automatic text structuring and summarization", Information Processing & Management, vol. 33, no. 2, **(1997)**, pp. 193-207.
[12] D. Shen, "Document Summarization Using Conditional Random Fields", IJCAI, vol. 7. **(2007)**.
[13] X. Zhang, H. Fuehres and P. A. Gloor, "Predicting stock market indicators through twitter, "I hope it is not as bad as I fear". Procedia-Social and Behavioral Sciences, vol. 26, **(2011)**, pp. 55-62.

 