A Novel Ranking Model for Product Name Extraction

Ling Shen[a], Qingxi Peng[b,*], Nian Li[c]

Wuhan Donghu University,Jiangxia, Wuhan,Hubei,PRC

[a]aleenapple@163.com, [b]pqx@hubu.edu.cn, [c]linian0728@163.com

* Corresponding Author

Keywords: Learning to Rank, Product Name, Extraction, SVM

Abstract. The Internet contains large amounts of information, which are valuable to the researchers. Since manual product name collection needs large human efforts and also time consuming, extracting keywords automatically become a hot topic. This paper addresses the issue of automatic extraction for agricultural product name. Previously, this problem was formalized as classification and learning methods for classification were utilized. In this paper, we transform the product name extraction problem into a ranking problem and employ a learning to rank method to solve the task. Specifically, we utilize the ranking SVM to extract the agricultural product name. Experimental results on real life datasets demonstrate that our novel ranking method outperforms the baseline methods.

1 Introduction

With the advent of Web 2.0 technology, there is an explosion of electric commerce on the Internet, which propels the economics. The data, however, scatter on different websites, which hamper the development of the E-commerce. Therefore, a large amount of studies have been devoted into the information integration. Agricultural product information is different from common text mining and product bulletin board. Common online text is longer than agricultural product information, while online product review is usually posted by customers. Agricultural product information is usually posted by administrative department, and the review text is usually short. Therefore, new method should be adopted to identify the agricultural product information. The core task in the information integration is entity name identification. In this paper, we focus on the agricultural product name extraction. Previous studies mainly focus on the classification methods to identify the produce name from the agricultural product. Otherwise, the public will spend a lot of time to browse the agricultural product news. On the other hand, the useful information spread out in many financial websites. Without professional approaches, they can hardly been collected and processed. In this paper, we employ a novel ranking method to extract the agricultural product name. To our knowledge, ranking methods have never been employed in product name extraction. This is the first study exploring agricultural product name extraction by a ranking method. The ranking SVM is the state-of-the-art approach of ranking method. We employ it to perform the extraction problem.

In this paper, we have made contributions listed below:

1. We transform the product name extraction problem into a ranking problem for the first time.
2. We take advantage of information gain method in feature selection to get ranking features.
3. Learning to rank has been employed for product name extraction. Experiments show that our method improves the accuracy of the task.

The rest of this paper is listed as follows. Section 2 give related work about our research. Section 3 give proposed method, which include feature selection, learning to rank method. Comparative experiments demonstrate the effectiveness of proposed method in section 4. Section 5 is the conclusion of the paper.

2 Related Works

Automatic keywords extraction has been an important research in information retrieval [1]. Many methods have been proposed to solve the problem. Early researches mainly focus on document level [2]. In recent years, product extraction attracts researchers. Agricultural product information 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, industry information exploited text mining methods [3, 4]. The existing approaches [5-7] rely heavily on the full-text retrieval. These methods, however, don't utilize machine learning methods, which result in low accuracy.

Learning to rank has been proposed in [8]. Then many methods have been proposed, which are point-wise, pair-wise and list-wise [9]. In [10], ranking SVM has been employed in document retrieval and gain good result. In feature selection research, many methods have been proposed. These methods include Information Gain, document frequency (DF), mutual information (MI), χ^2 statistic (CHI), and term strength (TS). The authority research is [11], which compare all the feature selection methods, and give commendation.

3 Proposed Method

3.1 A Novel Ranking Model

Previously the keyword extraction is often been regarded as a classification problem. The training data are collected manually. We classify the word as keyword and non-keyword. Then a classifier has been trained through the training data. Common used classification models are decision tree, Naive Bayes and SVM. The agriculture product name is professional and different from common text. In this regard, the product name extraction is by nature a ranking problem rather than a classification problem. We prefer to choose a learning to rank method than a classification method in this task. In this paper, we employ ranking SVM to the candidate agriculture product name. Ranking SVM is a state-of-the-art method in learning to rank model. Learning to rank is a model or a function for ranking objects in machine learning. Learning to rank is mainly applied in document retrieval, collaborative filtering, expert system and recommendation system. Recently, learning to rank technology has been intensively studied in document retrieval. However, learning to rank method has never been put into the keyword extraction. Learning to rank is different with classification and regression. The goal of learning to rank is to learn a function that can rank objects according to their degree of preference, importance, or relevance. In this paper, we devote learning to rank method in agriculture product name extraction problem for the first time. Firstly, the process of keyword extraction has been transformed into learning to rank process. In this model, the agriculture information has been used as document through 'bag-of-words' model. Secondly, a ranking model has been trained via learning to rank method. Finally, when we get new product information, the product text is treated as query. We get the ranking label through learning to rank model.

3.2 Feature Selection

In this section, we concern on feature selection. There are many feature selection methods such as document frequency (DF), information gain (IG), mutual information (MI), a χ^2 statistic (CHI), and term strength (TS). In this paper, we firstly employ Term Frequency Inverse Document Frequency (TF-IDF) as feature. TF-IDF is usually used to evaluate how relevant a word in a corpus is to a document. It may be thought of as a statistical measure. The important of a word to a document depends on its appearing number in the document, but is offset by the frequencies of the words in the other document. The second feature is the first occurrence position, which can discriminate the keyword and non-keyword. The third feature is Information gain. According to [4], information gain is an effective method in feature selection method, which is frequently employed as a term goodness criterion in the field of machine learning. It measures the amount of information obtained for category prediction by knowing the presence of a term in a document. This definition is more general than the one employed in binary classification models. Given a corpus with agriculture information, we can

compute information gain of each unique keyword. The terms with less information gain than predefined threshold are removed. In this section, we try the feature selection threshold manually.

3.3 Extraction Framework

There are three steps in our proposed agriculture product name extraction. The first step is use NLP tools to extract all the noun phrases in the agriculture information. Since many agriculture products are professional such as garlic, onion, pepper etc. New product name often occur on the agricultural information. In this step, we assure that all the product names are collected as candidate. Then in the second, we manually classify the training data as keyword and non-keyword. We use our proposed method and the training data to train a novel ranking model. In this model, many features have been employed. In the third step, we employ the ranking model to candidate keywords.

There are many other classifiers such as Naïve Bayes, support vector machine (SVM) and decision tree. A Naïve Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong independence assumption. A more description term for the underlying probabilistic model would be "independent feature model". SVM analyze data and recognize patterns, used for classification and regression analysis. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier. Previous studies [4] show that MaxEnt and SVM outperform other classifiers. In this section, we implement the maximum entropy tool for the learning to rank method.

4 Experiments

4.1 Experiment Setup

In order to evaluate the proposed method, we collect several agriculture websites. The agricultural product information is extracted from www.100ppi.com, www.cnhnb.com, and pfscnew.agri.gov.cn. We firstly utilize Stanford Parser to get what part of speech. All the nouns are collected as candidate product name. Then we get three datasets. Table 1 is the statistics of the data.

TABLE 1. Statistic of information for datasets

Data set	**Dataset 1**	**Dataset 2**	**Dataset 3**
Candidate Produce Name	2564	2138	2860

4.2 Experiment Result

In this section, we conduct three steps to finish the experiment. First step is feature selection. As shown in section 3, three features have been employed in feature selection. Then we manually construct the training data and testing data. Once we put it to the ranking system, the corresponding label will be given by the learning to rank model.

TABLE 2 Comparison of different methods

Method	Dataset 1			Dataset 2			Dataset 3		
	MAP	P@1	P@5	MAP	P@1	P@5	MAP	P@1	P@5
Navie Bayes	0.232	0.245	0.167	0.201	0.383	0.432	0.433	0.445	0.423
SVM	0.276	0.276	0.187	0.239	0.404	**0.489**	0.464	0.476	0.503
Ranking SVM	**0.302**	**0.334**	**0.201**	**0.248**	**0.554**	0.476	**0.512**	**0.464**	**0.576**

We choose two methods as baselines. The first baseline method is SVM. LibSVM has been chosen to classify the product name. The second baseline method is Naive Bayes. During the learning process, each dataset was separated into training data and test data. In evaluation, the keywords of each agricultural text were generated by ranking SVM and the baselines, and the performances of keyword extraction were evaluated in two measures: Precision at position n (denoted as P@n), Mean Average Precision (MAP). Table 2 shows the comparison of different methods.

As shown above, our ranking model acquires the highest effectiveness. For the classification methods, SVM is better than Naive Bayes. Our ranking model outperforms classification method in the three measures P@1, P@5 and MAP.

5 Conclusions

In this paper, we transform the product extraction problem into a novel ranking problem. Learning to rank has been employed to solve the problem. Three features have been exploited in feature selection. The result indicate that our novel ranking method outperform the baselines methods.

Acknowledgements

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 Document)

References

[1] Salton Gerard, et al., 1997. Automatic text structuring and summarization. Information Processing & Management 33.2 pp: 193-207.

[2] Shen, Dou, et al., 2007. Document Summarization Using Conditional Random Fields. IJCAI. Vol. 7.

[3] Acher, Mathieu et al., 2012. On extracting feature models from product descriptions. Proceedings of the Sixth International Workshop on Variability Modeling of Software-Intensive Systems. ACM.

[4] Bollen, Johan, Huina Mao, and Xiaojun Zeng., 2011. Twitter mood predicts the stock market. Journal of Computational Science 2.1 : 1-8.

[5] Sakaguchi, Motohiko, et al. Design and Implementation of an Automatic Text-referencing System by Keyword. Information Processing Society of Japan (IPSJ), 1995:169-170.

[6] Huang, Xiangji, and A. An. Design and implementation of a Chinese full-text retrieval system based on a probabilistic model. TENCON '93. Proceedings. Computer, Communication, Control and Power Engineering.1993 IEEE Region 10 Conference on IEEE, 1993:1090-1093 vol.2.

[7] Zhang, Xiao Wei, and Q. M. Zhu. Design and Implementation of a Web Full-text Information Retrieval System Based on Lucene. Computer & Modernization (2006).

[8] Trotman, Andrew. "Learning to rank." Information Retrieval 8.3 (2005): 359-381.

[9] Cao, Zhe, et al. "Learning to rank: from pairwise approach to listwise approach." Proceedings of the 24th international conference on Machine learning. ACM, 2007.

[10] Cao, Yunbo, et al. "Adapting ranking SVM to document retrieval." Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2006.

[11] Yang, Yiming, and Jan O. Pedersen. "A comparative study on feature selection in text categorization." *ICML*. Vol. 97. 1997.