A Bayesian Based Search and Classification System for Product Information of Agricultural Logistics Information Technology

Dandan Li [1],Daoliang Li [1,3], Yingyi Chen [1,3], Li Li [1], Xiangyang Qin [3],
Yongjun Zheng [1,*]

[1] China Agricultural University, P.O. Box 121, Beijing, 100083, P.R. CHINA

[2] Key Laboratory of Modern Precision Agriculture System Integration, Ministry of Education, P.O. Box 121, Beijing, 100083, P.R. CHINA

[3] Beijing agricultural information technology research center, Beijing, P.R. CHINA

Abstract. In order to meet the needs of users who search agricultural products logistics information technology, this paper introduces a search and classification system of agricultural products logistics information technology search and classification. Firstly, the dictionary of field concept word was built based on analyzing the characteristics of agricultural products logistics information technology. Secondly, the system used meta-search engine to search related pages on the Internet based on keywords collections, and then used Web mining to analyze and filter the relevant pages. Finally, classify the agricultural products logistics information technology by web text classification according to different users' needs. The results showed that the system could efficiently and accurately search the required information, and classification with good results.

Keywords. Agricultural products logistics; Web mining; information technology; classification of web text

1 Introduction

Agricultural products logistics refers to moving material objects and related information from producer to consumer physically for meeting customer's needs and achieve the value of agricultural products [1]. It mainly includes agricultural production, purchase, transport, storage, loading and unloading, handling, packaging, circulation, processing, distribution, information activities and many other aspects. Each aspect will be involved in many information technologies and products, also new technology will come out continually, and relevant information on the network has become increasingly rich. People want to know the existing information technology and products and hope to fully use them, but in the sea of information in the network, it is a great difficulty to find the information needed quickly and accurately.

According to the above requirements, Web mining and Web text classification were used to retrieve and classify agricultural product logistics information technology. Web mining can generally be divided into three types [2], which are Web content mining, Web structure mining and Web usage mining. Web content mining is a process of getting useful knowledge from the summary and the document content of pages, generally, including text files and multimedia documents mining [3-4]. Web text classification is an important technology of text mining, which refers to that each document of documents collection, will be included in a pre-defined category [5-6]. At present, the main classification algorithm includes the decision tree based on inductive learning, the K-nearest neighbor based on vector space model, Bayes classification based on probabilistic models, neural

*Corresponding author, Tel: +86-10-62736385, Fax: +86-10-62736591, Email: zyj@cau.edu.cn

networks, the support vector machines based on statistical learning theory, etc. [7].

The aim of this paper is to design a system of agricultural products logistics information technology search and classification based on the above analysis. Firstly, the dictionary of field concept word was built based on analyzing the characteristics of agricultural products logistics information technology. Secondly, the system used meta-search engine to search related pages on the Internet based on keywords collections, and then used Web mining to analyze and filter the relevant pages. Finally, classify the agricultural products logistics information technology by web text classification according to different users' needs. The classified information can be a good decision support for the future.

2 Materials and Methods

2.1 Systems framework analysis and design

The agricultural product logistics information technology search and classification system was designed to help users in the field of agricultural logistics to search required information from the Internet and make full use of this information more easily. The system's main functions included the following aspects:

1) According to the user's search request, match the agricultural product logistics concept word dictionary which the system has been built, and get effective keywords collection.
2) Considering the custom search scheme of the system, use meta-search technology to search the page information that meet the search request from the Internet.
3) Use web services method based on semantic vector model to judge match degrees between the information searched from the Internet and demanded by users. Then filter out irrelevant information, and store useful information.
4) The system achieved automatic classification function for searched useful information, and provided decision support for users to select technologies and products.

System work flow is shown in fig.1.

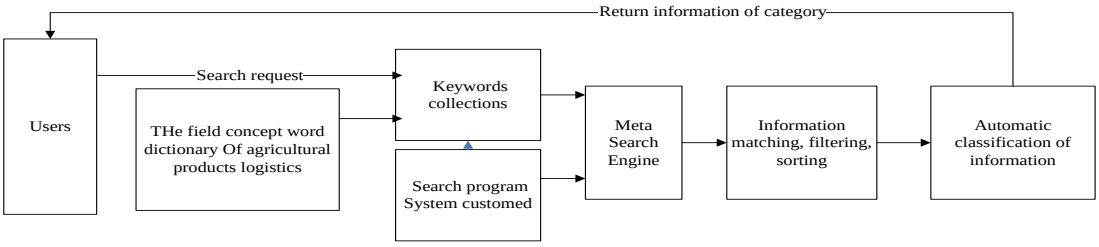

Fig.1. System workflow chart.

2.2 Analysis of agricultural products logistics information technology

Agricultural product logistics technology refers to the machine, equipment, facilities and other hardware and software and a variety of opportunities method which are used in the process of agricultural product from producers to consumers. Informatization of agricultural products logistics namely agricultural product logistics technology is applied to the logistics field. Agricultural product

logistics can be divided into agricultural product production logistics, sales logistics and waste logistics according to supply chain function; and can be divided into food logistics, economic crops logistics, fresh food logistics, livestock product logistics, aquatic product logistics, forest product logistics and other agricultural product logistics according to concrete object; while, can be divided into room temperature chain logistics, cold chain logistics and fresh chain logistics and so on accordance with the logistics storage and transportation conditions. Agricultural product logistics information technology is very different, for that the system provided a diverse and differentiated information services according to the different demand and the needs of different levels to achieve the system's utility and meet the information needs of individual use.

According to the above analysis of agricultural product logistics information technology fields, the field concept word dictionary has been built. The system can obtain a valid set of search keywords by extracting valid search terms from users' input, semantic analyzing and matching with the the concept diction. This will reduce irrelevant information from the returned search results.

2.3 Search method framework

As the general search engine is limited coverage of the entire web, and search results will return many useless information, so the system selected meta search engine technology. Because the system is aim to search agricultural products logistics information technology, users' needs are logistics technology and product-related information, such as technical characteristics of the products, scope and price information, and so on. Based on the above considerations, the system specifically customized search scheme, defining a term including three words such as warehouse management system, system features and the price. When a keyword is inputted, for example, warehouse management system, the system will search uses the term which the keyword relevant in the Internet and return relevant information to the user. Users send search requests to meta search engine [8], and the meta search engine send the actual search requests to multiple search engine according to the users' requests, and multiple search engines perform search requests from the meta search engine, and sent search results to the meta-search engines by the response form. The meta-search engines send the obtained and search results to the actual users.

The system filters valuable information by judging relevance between the search results and user's queries to. There are many kinds of methods or models to judge relevance between the search results and user's queries, such as vector-based, based on probability, fuzzy set, latent semantic models, and so on. Here draw lessons from the vector model based on semantic Web service matching method for decision making [9].

In the model, data items to be matched compose of the concept a public body, and set the concept of space vector ($c_1, c_2, \cdots, c_n$) finally getting the data items to be matched vector model is $d_j = (w_{1,j}, w_{2,j}, \cdots, w_{n,j})$, User query is $q_j = (w_{1,q}, w_{2,q}, \cdots, w_{nj})$, The formula of weight is shown as follows.

$$w_{i,j} = \begin{cases} 1 & freq_{i,j} > 0 \\ \max_t S(c_i, c_t) & freq_{i,j} = 0 \cap c_i \operatorname{Re} lc_t \\ 0 & others \end{cases} . \quad (1)$$

$freq_{ij}$ is the frequency of concept in the data item d_j, $\mathrm{Re}\ 1$ represent the relationship between the concepts, $S(c_i, c_t)$ is the semantic similarity between c_i and c_t,

The formula is as follows.

$$S(c_i, c_t) = e^{-al} \times \frac{e^{\beta h} - e^{-\beta h}}{e^{\beta h} + e^{-\beta h}} . \quad (2)$$

In the formula, l is the shortest path length between c_i and c_t ,h is the depth of the deepest common ancestor concept of c_i and c_t ,α ∈ (0,1), β ∈ (0,1) are the Impact coefficient of the two factors which are the length of the shortest path and the concept depth to the concept of semantic similarity.

Finally, the formula of the matching degree of the data item d_j and q_j, being the same with the cosine of two data items vectors, is as follows.

$$Sim(d_j, q) = \frac{d_j \times q}{|d_j| \times |q|} = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{t} {}^{2}_{i,q}}} . \quad (3)$$

$|d_j|$ and $|q|$ are the model of data item vector and inquires the vector, $Sim(d_j, q)$ is between 0 and 1, finally, matching result is sorted according to $Sim(d_j, q)$.

Through this method, the system can effectively extract relevant information of agricultural logistics information technology from the results of general search engines.

2.4Automatic text classification

After searching agricultural products logistics information technology and related information, the system would automatically classify the obtained information. In the process of Web text classification, including the four key steps as follows, namely the text pretreatment, the text says, characteristic dimension reduction, training methods and classification algorithms. The text pretreatment process take out some HTML or XML tags, a key link of Chinese text classification of is the Chinese automatic segmentation.Web document content is described in natural language, computer difficult to handle its semantic, to facilitate the computer process, so must transform the content of the text features into the computer can process format. After word segmentation and removing stop words and high frequency from the training text and the text to be classified, the dimension of vector space and category vector for said text is very big, so the need for feature dimension reduction. The job of training algorithm is to statistics each text corresponding word table in training set of documents, calculate category vector matrix simultaneously normalization, finally save the table get from the training, namely classification knowledge base. The classification algorithm was designed based on the classification knowledge base.

Now many text classification algorithm has come up and improvement, such as based on group classification method, multiple classifier fusion method, based on RBF network text categorization model, latent semantic classification model , K-neighbor algorithm and support vector machine, etc.

This paper adopted the bayes classification algorithm [10].

The bayes classification algorithm as follows,

Step 1, Take out entry from the entry set T_x one by one, and match with the word in the feature vocabulary, if t_{xk} and $t_i(c_j)$ match, then give category c_j to t_{xk} , recorded as $t_{xk}(c_j)$; if t_{xk} and $t_i(c_j)$ not match, then take out next entry until T_x is null, at last, get the entry set classified is

$$\{t_{xk}(c_j)\} \quad k = (1,2,\cdots,n; j = 1,2,\cdots mx).$$

Step 2, Calculate conditional probability of each entry in the category, the formula as follows.

$$P(t_{xk} \mid c_j) = \frac{n_{kj} + mp}{n_{xj} + m} \tag{4}$$

m is a constant, called the equivalent sample size; P is the priori estimates of the probability to be defined.

Step 3, Calculate conditional probability of text x, the formula as follows.

$$P(t_x \mid c_j) = P(t_{x^1} \mid c_j) \times P(t_{x^2} \mid c_j) \times \cdots \times P(t_{x^n} \mid c_j) = \prod_{k-1}^{n} P(t_{xk} \mid c_j) \tag{5}$$

Step 4, Calculate the probability of text x belong to category c_j , the formula as follows.

$$P(c_j \mid T_x) = \frac{P(c_j)P(T_x \mid c_j)}{P(T_x)}. \tag{6}$$

and

$$P(T_x) = \sum_{j-1}^{m} P(c_j)P(T_x \mid c_j) \tag{7}$$

Step 5, Take text category when

$$\max\{P(c_1 \mid T_x), P(c_2 \mid T_x), \cdots, P(c_{mx} \mid T_x)\} \tag{8}$$

as the category of text x.

This system takes two indexes for classification of evaluation methods, namely the accuracy and recall ratio. Let the correct number of text classification be num1, and let the actual number of text classification be num2, and let the number of should have text be num3.

The definition of accuracy, accuracy $= \frac{num1}{num2}$

The definition of recall ratio, recall ratio $= \frac{num1}{num3}$

2.5 Results and analysis

The test aimed to prove the search and classification effect of the system. The search object of this

system was agricultural product logistics information technology and related product, including the transportation, loading and unloading handling, storage, packaging, circulation processing, collection and processing of information and containers unitization in agricultural logistics activities. Here selected three items as search request, and compared the precision ratio with general search engines. Precision ratio is the ratio of effective search page to the total number of pages. The results are shown in table 1.

Table 1. The comparison of search results.

Search request	Precision ratio of general search engines	Precision ratio of this system
Automatically lead machine	90.3%	92.5%
Automation warehouse	91.2%	93.7%
Warehousing management system	90.5%	93.2%

Table 1 shows that the precision ratio of the system is improved than the precision ratio of general search engines.

Through detailed analysis of agricultural logistics information technology and the characteristics of the products, the system divided agricultural product logistics information technology into seven parts according to the function, as follows.

(1) Transportation, including rail, road, water transport, air and pipeline transportation.

(2) Material handling, including loading and unloading machinery, transportation machinery and material handling machinery, such as forklifts, automated guided machines, lifts, stackers, etc.

(3) Storage, including storage of materials, storage equipment, such as automated warehouses, shelves, trays, temperature and humidity control equipment.

(4) Packaging, including filling machines, sealing machines, labeling machines, sterilization machine, and multi-functional packaging machinery.

(5) Distribution processing, means the professional machinery and equipment used in the activity such as packaging, split, measurement, sorting, assembling, pay the price stickers, labels pay and so on.

(6) Information collection and processing, including computer and network related hardware and software, information identification devices, communication equipment.

(7) The equipment of container unitization includes containers, trays, slide, FIBC, container network goods bundle, container handling equipment, transport equipment, container, and container identification system.

The system pre-set six categories such as transport, handling, storage, packaging, distribution processing, information collection and processing. and then selected the 600 piece of pages in the above searched page, the 480 of them as a training text, the other 120 as a test text. The test results of the system automatically classified shown as Table 2.

Table 2. The test results of the system automatically classified

Category	Accuracy	Recall ratio
Transportation	91.5%	90.3%
Material handling	95.2%	94.6%
Storage	90.1%	89.7%
Packaging	92.6%	93.2%
Distribution processing	87.9%	88.6%
Information collection and processing	94.7%	93.5%

3 Conclusions

A system of agricultural products logistics information technology search and classification was designed in order to meet the needs of users in the field of agricultural products logistics information technology. Firstly, the dictionary of field concept word was built based on analyzing the characteristics of agricultural products logistics information technology. Secondly, the system used meta-search engine to search related pages on the Internet based on keywords collections, and then used Web mining to analyze and filter the relevant pages. Finally, classify the agricultural products logistics information technology by web text classification according to different users' needs. The results showed that the system could search the required information efficiently and accurately, and classification with good results.

Acknowledgements

This work was supported by Special Fun for Agro-scientific Research in the Public Interest (200903009). The research was also financially supported by the national science and technology support plan (2009BADC4B01).

Reference

1. Dejun Liu,Guangsheng Zhang.Modern technology and management of agricultural product logistics[M]. China Logistics Publishing House,2009.
2. Juan D. Velasquez, Luis E. Dujovne, Gaston L'Huillier, Extracting significant Website Key Objects: A Semantic Web mining approach, Engineering Applications of Artificial Intelligence, 2011(2)1:1-10
3. Oguz Mustapasa, Dilek Karahoca, Adem Karahoca, Ahmet Yucel, Huseyin Uzunboylu, Implementation of Semantic Web Mining on E-Learning, Procedia - Social and Behavioral Sciences.Innovation and Creativity in Education, 2010:5820-5823.
4. Oguz Mustapasa, Dilek Karahoca, Adem Karahoca, Ahmet Yucel, Huseyin Uzunboylu, Implementation of Semantic Web Mining on E-Learning, Procedia - Social and Behavioral Sciences, Innovation and Creativity in Education, 2010(2)2:5820-5823.
5. Jingnian Chen, Houkuan Huang, Shengfeng Tian, Youli Qu, Feature selection for text classification with Naive Bayes, Expert Systems with Applications, 2009(36)3:5432-5435.
6. Shuchuan Lo, Web service quality control based on text mining using support vector machine. Expert Systems with Applications.2008,34(1):603-610.
7. Selma Ayse Ozel, A Web page classification system based on a genetic algorithm using tagged-terms as features, Expert Systems with Applications, 2011,38(4) : 3407-3415.
8. Mohamed Salah Hamdi, SOMSE: A semantic map based meta-search engine for the purpose of web information customization, Applied Soft Computing, 2011,11(1): 1310-1321.
9. Xue Mao,Jiehong Guan,Fubao Zhu.Web services matchmaking approach based on semantic vector space model[J]. Application Research of Computers, 2010, 27 (10):3754-3758.
10. Hyoungdong Han, Youngjoong Ko, Jungyun Seo, Using the revised EM algorithm to remove noisy data for improving the one-against-the-rest method in binary text classification, Information Processing & Management, 2007(43)5:1281-1293.