TELKOMNIKA Indonesian Journal of Electrical Engineering
Vol.12, No.3, March 2014, pp. 2212 ~ 2218
DOI: http://dx.doi.org/10.11591/telkomnika.v12i3.4489

Key Technology of Agricultural Production and Market Information Matching in Big Data Era

Shuo Wang*[1], Shihong Liu[2]
Key Laboratory of Digital Agricultural Early-Warning Technology, Ministry of Agriculture
Beijing, P.R. China 100081
*Corresponding author, e-mail: wangsurecn@163.com[1], lius@mail.caas.net.cn[2]

Abstract

This article describes challenges faced by agricultural production and marketing in the era of big data, and then builds the agricultural market information matching platform based on HADOOP & NUTCH combining cloud computing technology, finally details its layers and key technologies, including the use of open source search engine to capture the whole network market information to build agricultural production and market data sources, users' interest model and the combination of matching algorithm and HADOOP environment. The aim of this paper is to make the agricultural market information matching platform more suitable for Chinese agricultural production and marketing system in order to solve the bottleneck problems such as information overload, lack of storage space, scalability and efficiency of analysis and calculation. As a result, this thesis provides a useful reference and new strategy for analysis and mining of big data in agricultural production and marketing areas.

Keywords*: information matching, hadoop, user interest model, matching algorithm, agricultural products*

1. Introduction

With the accelerating global agriculture marketization and internationalization, the circulation of agricultural market is increasing and huge amounts of information is scattered and messy, which leads to increasing unstructured data in agricultural market information. Agriculture big data presents features as follows: Volume, Velocity, Variety, Value, Veracity. These features make it more difficult to obtain the required information rapidly, timely and accurately, and also make more and more prominent contradiction that agricultural production does not match market circulation. The personalization system can solve the problem of overloading information well, but as the information explosion phenomenon deepening in the big data era, the personalized recommendation system is confronted with some bottleneck problems such as the extensibility of storage space and analysis calculation efficiency. As a new business model, cloud computing is especially suitable for processing large volumes of data and has achieved widely attention and recognition. Among them, Hadoop is a cloud computing platform which can processing mass data parallel with strong expansion ability, low cost, high efficiency and good reliability. Therefore, it is imperative to build the accurate cloud computing and analysis model based on the users' interest model and provide high precision and high speed agricultural market information matching system through combing the existing personalization matching recommendation system with the Hadoop platform, which can collect entire network market data and build a big data center of market data. To build such a matching platform of gathering and analyzing massive data will have far-reaching significance to solve the contradiction between small-scale production and big market.

2. The Design of the Matching System

The strict direction of data flow and information flow is important in the system, as well as the division and logical association between the various business levels. Therefore, the matching system is divided into four levels in this article: getting and parsing source information, building user interest model, information matching algorithm and recommendation, and user interface. The source data fetching and parsing. Authoritative market information release site is chosen as the source URL, using open source search engine Nutch to build foundation

Received August 21, 2013; Revised September 25, 2013; Accepted October 16, 2013

framework of this set of recommendation service, which is integrated into the Hadoop distributed environment to finish the basic distributed acquisition work of source information. Then the system grasps the original market data and pretreats the source data. The main body of web information will be transformed from HTML format into a structured text document easily to deal with, which is obtained by the information parsing module. At last, feature extraction will be done, etc. The construction of user interest model. It is mainly used to lay the foundation for providing personalized services, which builds user interest model by the Vector Space Model through analyzing the user registration and WEB logs. The basic algorithm of information matching recommendation. This paper designs a matching recommendation algorithm through combining model technology after analyzing the existing agricultural information matching systems. The user interfacehas two effects: displaying the results to users and collecting log information. We sent test results to farmers, government agencies, research institutions and consumers through mobile phone, computer, TV and other media mediation tools, which totally realizes the goal of “combine production and marketing, service the whole society”. The technology roadmap is shown in Figure 1.

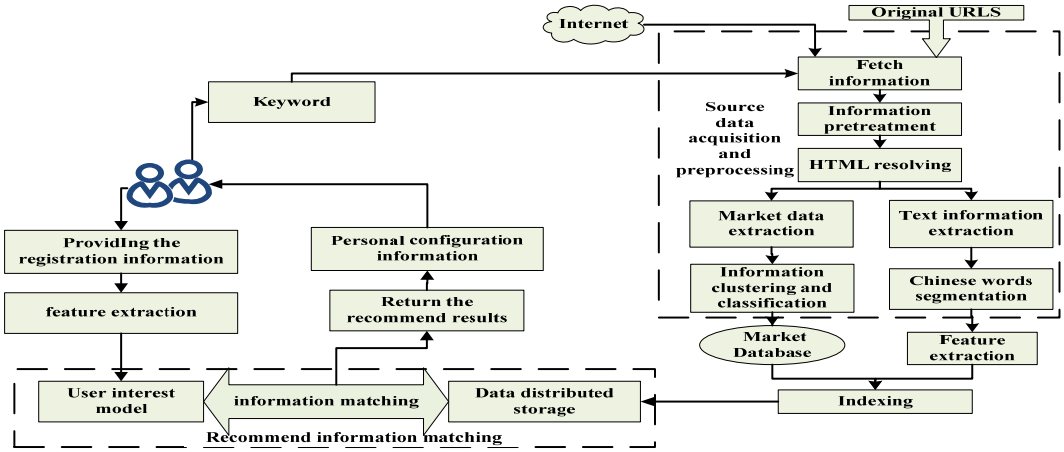

Figure 1. The Technology Roadmap

2.1. The Marketdata Acquisition

Specific steps are as follows: First, we need to provide a source of seed-websites for the source information access. In order to meet the statistical regularity seed, we selected a large number of agriculture-related websites, and then a unified summarized, finally get a list of seed-websites. In this list, including three sites: the purchase information, agriculture official government website, and agricultural technology and seed informational sites. Then, start the virtual machine software and the Nutch, in order to achieve crawl tasks. This module provides the whole system with source information, which provides the basis for recommended services. The system uses open source search engine program as the way of source information access, which allows us to see the complete source code. Nutch is used as the special open source search engine in this article, which is a top-level Apache Software Foundation project, programmed in JAVA.

2.2. The Data Pre-treatment

Source information pre-processing module is mainly responsible for pre-processing source information, captured by source information fetching module, and its main workflow includes eliminating duplicates, analyzing HTML, Chinese word segmentation, stop word processing [2].

Eliminating duplicates is mainly used to filter out reproduces pages or mirror site in the source information, which is characterized as the same URL of these pages. This part is realized by writing code to compare the URL string. The purpose of HTML parsing is to obtain information about the page, such as the URL, title, content and so on. The main purpose of

obtaining this information is to prepare for the next feature extraction, which can be completed by the HTML parser of Nutch. Resolution can be done by creating your own analytic categories, and then calling the class in org.apache.nutch.parse.HTML. In the Chinese text processing, the Chinese word segmentation is an integral part. This article uses ICTCLAS Chinese word segmentation system based on Hidden Markov Model. We build our own Segmentation class in the field of agricultural production and marketing based on the API provided by this Chinese word segmentation system, with specific functions implemented in Segment function. In the Chinese text processing, two types of words should be special treated. The first is the words which are common in every document, and the second is the words which are rarely used alone to express feature information of a document, which are regarded as "stop words". Generally, these words will be dealt with abandon to reduce the dimension of feature vectors, increasing search efficiency, improving the retrieval results and avoiding a lot of interference.

2.3. The Data Parsing and Feature Extraction

The information analysis is mainly to obtain the content of the website, and then change it from HTML into text documents, completed by the HTML parsing modules.

The parsed text will be carried out feature extraction to reduce the dimensionality of the data processed by subsequent recommendation algorithm, improving computational efficiency. In summary, we select TF-IDF algorithm because its design ideas and principles can satisfy several features of characteristic data, which is calculated as follows:

$$W(t,d) = \frac{tf(t,d) \times \log\left(\frac{N}{n_t} + 0.01\right)}{\sqrt{\sum_{t \in d}\left[tf(t,d) \times \log\left(\frac{N}{n_t} + 0.01\right)\right]^2}} \quad (1)$$

W(t, d) means the weight of the feature word t in the document d, TF is the word frequency of the feature word t in the document d, N is the total time of training, ni is the number of texts containing t in the training library, the denominator is a normalized factor [3]. In this article, functions are implemented by the feature extraction class, in which extraction function is responsible for completing the task of feature extraction.

2.4. User Interest Model Construction

The user interest model is particularly important as a recommendation standard in the agricultural production and marketing recommendation layer. First user interest information have been obtained through multiple channels and extracted feature after Chinese word segmentation by using TF-IDF algorithm to calculate the weight.

We will extract text feature after parsing and Chinese word segmentation of the screened web pages. Here we adopt the TF-IDF algorithm to calculate the weight of each character word and sort them according to the weight. Because the system is used to represent multiple agricultural production and marketing information, we have to extract the screened text feature uniformly in a certain field to get the related characters and we use the top-N ranked by weights as the character information in the field. Thus the amount of the character item needed by n topics in the model is: n*N. We take the vector space model to represent the model for the convenience of the subsequent computation. The VSM is presented as follow:

$$U_{vsm}=\{(t_1, w_1), (t_2, w_2)\ldots\ldots(t_n, w_n)\} \quad (2)$$

Among them, t represents character words, w represnts nomalized weights of character words in the whole training text sets based on TF-IDF algorithm.

We will also adjust user interest model according to changes of users interests. The system will get and analyze users browsed logs in the users' using process and the information is mainly about web pages information, residence time, clicks and download times and so on. If users pay long time or click many times for some pages, the system will deal with the pages and get character text sets.

2.5. The Information Matching Algorithm

In the design of recommendation algorithm part, we use recommendation algorithm, based on content and collaborative, to design a neural network-based information recommendation method, which uses BP neural network algorithm and SOM algorithm to achieve mixed information recommended. Its implementation block diagram is shown in Figure 2.

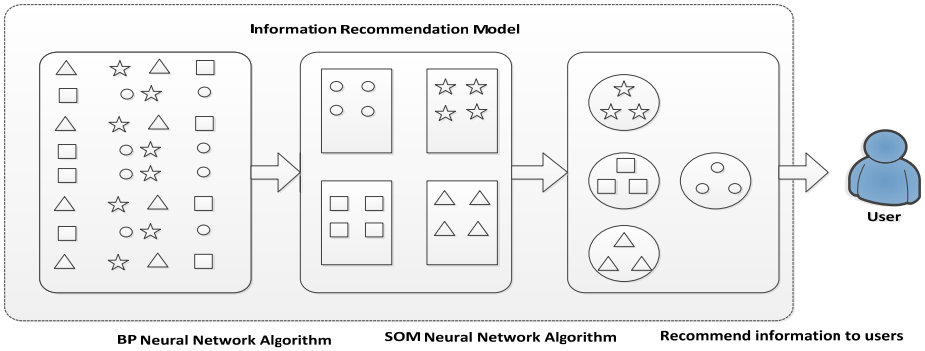

Figure 2. Implementation Block Diagram of Mixed Recommended Algorithm

Entire hybrid information recommendation algorithm has two main parts. The first part is to classify the retrieved information in term on subject, clustering the same information pages together achieved by BP neural network algorithm, which uses the thought of the information recommendation based on content. The second part is to match the clustered Web pages with the user interest model, and then recommend the pages that its calculated value is greater than the set value to the users, which uses the thought of the information recommendation based on collaborative filtering, achieved by SOM neural network. We use the SOM algorithm in the implementation of recommendation algorithm based on collaborative filtering [4]. The basic structure of the algorithm includes the input layer and output (Competition) layer, and the connection type between input layer neurons and output layer neurons is the whole interconnect. The neurons in the output layer are arranged two-dimensionally, and each of nerves element represents an input sample. The matching system in the article can not only recommend the most relevant topic information retrieved by users, but also recommend other information related to the topic information. In the system, the input value is the character vector information of web pages and the weight of the character attach the input with the output value. If the results of calculation is greater than the preset threshold, we will recommend the information to users otherwise not.

3. The Implementation of the System

3.1. The Construction of Experimental Environment

We choose Hadoop version 0.2 in constructing the experimental environment and use jdk1.6.0_24 which can support Hadoop. The specific development environment is Eclipse + Hadoop eclipse plugin, and the configuration of the hardware experiment platform is OS: CentOS5.5 x64; CPU: Intel(R) Xeon(R) E5420@2.50GHz; Memory: 4GB RAM. In the test we mainly adopts four PCs(PC1~PC4) to build the cloud computing environment, in which the PC1 is used as the Namenode and the Jobtracker and PC2∼PC4 are used as the Datanode and the Tasktracker. The /etc/hosts directory configurations for each PC in the experiment are as follows: masters: 192.168.10.1,slaves: 192.168.10.1,192.168.10.2,192.168.10.3,192.168.10.4. In addition, we use ssh-keygen to generate the key pair on PC1 and then copy the public key to the /home/.ssh directory of each machine so that we can log in each machine from PC1 without password. In the critical configuration of Hadoop, we modify the localhost under the conf/masters and slaves of each machine to the corresponding IP address, and configure IP addresses of the namenode and the jobtracker under conf/mapred-site.xml. We can realize the critical configuration parameters by modifying conf/core-site.xml, conf/mapred-site.xml and conf/hdfs-site.xml, etc.

3.2. The Acquisition of Market Data Source

Firstly, we need provide seed sites for the system to grab source information. In order to meet the statistical regularity of site selection, we choose a large number of relevant agricultural sites and then summarize them uniformly to get a list of seed sites. In the list there are three kinds of sites: bidders, agricultural government's official websites and agricultural technology and seed information websites. Secondly, we start the virtual software and Nutch to achieve grasping task. After inputting the source information fetching instruction "bin/nutch crawl url.txt -dir crawtest -depth 3 -threads 4 >&crawl.log", the system will start the garbing process. In the instruction, url.txt storages seed sites also stored in our database for the subsequent updates; depth 3 means the crawling depth is 3 layer; threads 4 means starting 4 thread process to grab at the same time; crawl.log records the crawl log information about the system running state.

3.3. Market Data Processing

In order to facilitate the calculation, we have to carry on feature extraction and vectorization of texts and the calculated value of the vectorization is come from the users' interest model. During this process, we firstly need to extract features of text information in all areas. Also for the statistical regularity, we collect many texts related to bidders, government, seed and planting techniques and divide them into two parts: one is for the construction of the interest model and the other is for the test.

Then we can get top-N features in all areas to build our users' interest model through the feature selection algorithm mentioned above. In our system, we let N be 30 by reference to the experience of predecessors' research. Thus, we can not only meet the requirement of the efficiency needed by the recommendation algorithm but also keep the basic information of texts as far as possible. We get the users' interest model as shown in Table 1.

Table 1. Users' Interest Model

	Feature item	Weight	Forgetting factor	Present time	Modification time
Market information	Purchasing	0.1683	1	2013.5.3	2013.5.3
	Supply	0.0492	1	2013.5.3	2013.5.3
	Market	0.0324	1	2013.5.3	2013.5.3
	…	…	…	…	…
Policy	Increase	0.16338	1	2013.5.3	2013.5.3
	Decrease	0.04596	1	2013.5.3	2013.5.3
	Impact	0.02939	1	2013.5.3	2013.5.3
	…	…	…	….	…
Seed	cultivation	0.0754	1	2013.5.3	2013.5.3
	Judgement	0.0687	1	2013.5.3	2013.5.3
	varieties	0.0632	1	2013.5.3	2013.5.3
	…	…	...	…	…
Planting	Prevention and cure	0.1783	1	2013.5.3	2013.5.3
	planting	0.0492	1	2013.5.3	2013.5.3
	management	0.0324	1	2013.5.3	2013.5.3
	…	…	…	…	…

From Table 1, we know feature items, weight, forgetting factors and related time parameters of the users' model. Initial forgetting factor factors are 1, which is decided by its calculation formula that t=Y at the initial time. In order to facilitate the calculation, we present the users' interest model as the VSM form Uvsm={(t1,w1),(t2,w2)…(tn,wn)}, and make the arrangement according to topics of bidders, policy guidance and market forecasting, good seed information and planting technology. The effect is shown as follows: Uvsm={(purchasing,0.1683)…(increase,0.1633)…(cultivation,0.0754)…(Prevention and cure,0.1733)…}.

3.4. Information Recommendation

We use the hybrid recommendation algorithm based on the neural network and the input as the algorithm must be calculable vector value. So we need to return to pages with processed retrievals and carry on vectorizations. Firstly we extract features from page information, and then compare obtained feature items with users' interest model one by one, if the same, set the position of the corresponding feature item in users' interest model be 1,

otherwise 0. In this way, we will eventually get a vector only containing 0s and 1s. The vectorized text information can be used as input information of recommendation algorithm. The first BP neural network layer as the mixed recommendation algorithm.

Settings the number of the input layer nodes as 50, the hidden layer 30, the output layer 5, the learning coefficient 0.7 and the expected error 0.05. The error adjustment curve finally becomes stable and the maximum output error is 0.96, the minimum output error is 0.037, all within the scope of the expected error, which means the parameter adjustment has finished. We have done 6000 trainings in the test.

Then, we use the output of the first layer as the input of SOM algorithm in the second layer, in which carry on the visualization operation of vectorization texts to make a more intuitive display. The domain name will show the related topic domain information including agricultural production and marketing information, market and policy guidance, good seed information, planting techniques and so on when we click “import” button to choose training library. The test name will present vectorization texts and the presentation of vectorization texts is the vector form converted from texts. Set the number of input layer nodes as 90, input layer 120, learning coefficient 0.1, training times 10000 and click the button to begin the training.In the hybrid algorithm layer, we will recommend the webpage to users when the weight calculated by texts and users’ model is greater than the threshold value. The threshold value is set as 0.012, the maximum recommend output 0.5, the minimum value 0.039.

4. The Experimental Results and Analysis

First of all, we will build a test-text library. The selection rules of test-text are similar with that of building the users’ interest model. Secondly, we deal with these texts according to the process of building users’ interest model, such as selecting feature, calculating weights and so on, then we get results used as the output of the recommendation algorithm. Eventually we get final recommendation results by setting recommendation algorithm close value. And we can calculate recall ratio and precision ratio shown in Table 2.

Table 2. Test Result

Threshold value	8	9	10	11	12	13	14	15	16	17	18	19	20
Recall ratio(%)	61.3	70.6	71.4	75.1	77.5	80.3	82.1	85.6	85.9	86.2	86.3	86.6	86.8
Precision ratio(%)	81.3	80.8	80.3	79.8	78.3	78.0	77.2	75.8	74.6	74.1	73.5	73.2	72.6

From test results above, we know that in the test environment of small amount of data in the laboratory, when the threshold value is around 0.012, the recall radio can reach 77.5% and the precision radio can reach 78.3%, which both meet the requirement of the practical scope of system effectiveness.

5. Conclusion

The existing agricultural market information matching systems are discussed and analyzed in this thesis firstly, and then the spatial agricultural production and market data sources are built using the open search engine. Next, this thesis puts forward the user interest model and matching algorithm more suitable for users in agricultural fields. Under the big data circumstance, this article combines the existing matching system with Hadoop to solve the bottleneck problems of the matching system, such as limit of the storage space and the efficiency of analysis and calculation. And then, the paper provides a new method to optimize the proper matching and processing of the mass information about agricultural production and marketing and furnish a beneficial outlook for the application of cloud computing technology in agricultural fields. Finally, we test the system and analysis the recall ratio and the precision ratio as well as the impact of cloud computing on recommendation system. In future, we plan to design and improve our system under cloud environment, and study data acquisition and preprocessing technologies related to application, and explore other more suitable recommendation algorithm combing with cloud computing technology.

References

[1] Meng X, Ci X. The Concepts Techniques and Challenges ofBig Data Management. *J. Comput Sci Technol. Sciencep.* 2013; 50(1): 146-169.
[2] Huang L, Dai L, Wei Y. *A Recommendation System Based on Multi-agent. J. Expert Systems with Applications*. Proceedings of the 2rd Intl. Second International Conference on Genetic and Evolutionary Computing (WGEC '08). Hubei. 2008: 223-226.
[3] Li X, Zhao L, Wu L. *A Feature Extraction Method Using Base Phrase and Keyword in Chinese Text.* Proceedings of the 3rd Intl. 3rd International Conference on Intelligent System and Knowledge Engineering (ISKE 2008). Xiamen. 2008; 1: 680-684.
[4] Foster I, Kesselman C, Nick J, Tuecke S. *A SOM combined with KNN for classification task*. International Joint Conference on Neural Networks (IJCNN). San Jose, CA. 2011: 2368- 2373.