The Web Semantic Analysis of Port Customer based on Hadoop

Song Zhang[a], Lei Huang[b]

School of Beijing Jiaotong University, Beijing 100044, China

[a]14120581@bjtu.edu.cn, [b] lhuang@bjtu.edu.cn

Abstract. With the changing of ports services market' relationship between supply and demand, companies are trying to assist their decision-making through traditional business intelligence technology. Ports through the internal data collection can play a limited role. Then it needs to be studied through external information. In this paper, we study the port customers Web semantic analysis method based on Hadoop, using distributed computing methods to study extraction port customer-related information.

Keywords: Ports; Business Intelligence; External information; Semantic analysis; Distributed computing.

1. Introduction

Since the 1990s, the biggest change in the world port industry is its development focus to Asia, especially move to China. According port charts and follow-up observations in Hong Kong in recent years, the network continuously released: the world's port cargo throughput achieved 10 in 2010, Chinese mainland port throughput accounting for 78.2%, in 2011 the proportion rose to 79.33% and in 2012 the proportion rose to 80.19%, accounting for eighty percent in 2013 to 81.11 percent continue to improve, reflecting the Chinese port "Legion" in the global top ten ports absolute advantage and components.

More and more companies began to focus on customer value analysis and mining, but in addition to the internal relatively fixed production business information, for a lot of content on the page, still uses a manual for web browsing and analysis of state. With the increasing amount of information in the Internet age, the use of computer intelligence for Web content analysis and processing is critical. The port enterprise customers' external pages related information extraction and analysis, improve relevant theories and methods to correctly measure the development of port enterprise customers, optimize port customer strategy has important significance.

Through the research on distributed computing and semantic analysis theory, pointing out that semantic reasoning mechanisms by means of external data port customer value analysis of superiority. On the basis of the relevant semantic analysis, information analysis to further improve the mechanism of external ports for port enterprises Web information drawing customer value analysis provides a theoretical basis.

2. Method and Experimental

2.1 Method description and experimental procedure

(1)Semantic web

In order to solve the conventional Web publishing information, which it can't be understood by the machine, and thus can be automated reasoning and knowledge discovery computers and other issues. Tim Berners-Lee et al., The definition of the Semantic Web is: "not another Web Semantic Web, which is an extension of the existing Web, where the corresponding page information is given well-defined, so that the machine can be a good synergy with others jobs" [1].

(2)Distributed computing

In this paper, the open source Hadoop distributed system infrastructure is widely used nowadays. Hadoop was developed by the Apache Foundation to create the data store level implements a distributed file system (Hadoop Distributed File System Acronym: HDFS), with high fault-tolerance

features. Hadoop distributed system was designed with consideration to the application of cost, which can be a good hardware configuration to the low basis; Hadoop distributed system can provide very high bandwidth to ensure the application of real-time data access, suitable for large amounts of unstructured data storage and access requirements. [2] External entire distributed data extraction flow chart shown in *Figure 1*:

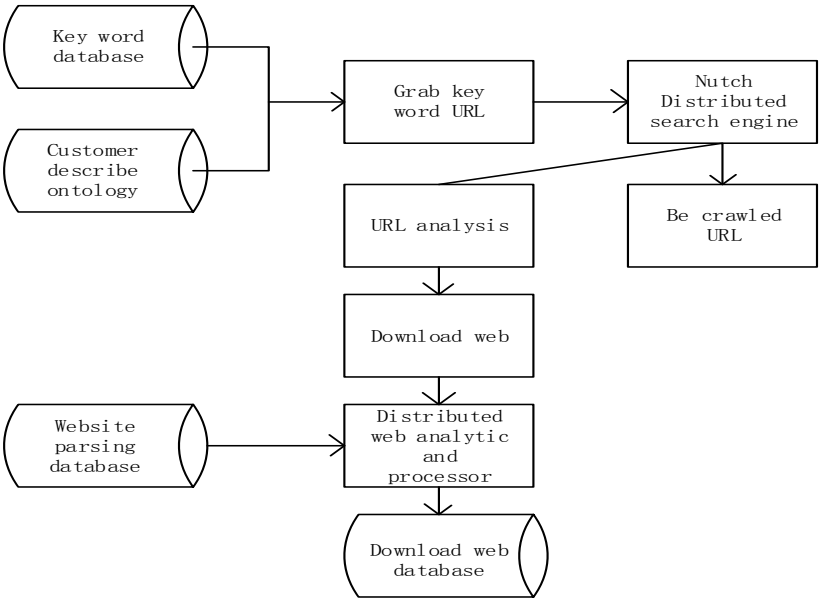

Figure 1 Distributed crawling and preprocessing mechanism based on Nutch

(3)Text data pre-processing

The first is text de-noising. The text is usually accompanied by a lot of noise data, such as advertising and promotion, regardless of content with the main content by reptile crawled. The presence of these data will seriously affect the accuracy and precision of subsequent Web semantic analysis and other related work, and therefore noise must be removed. [3] The present study carried out a lot of method and we use more mature and widely method called Defrost algorithm.

The second is Chinese word. It aims to dig out the text of potentially valuable information with words as the smallest segmentation granularity. In order to achieve fast and accurate word, we need to rely on efficient string matching algorithms based on the efficient and accurate word dictionary.

The third is Removal of stop words. It refers to the text processing. If you encounter such words, it will immediately stop processing.

The last is Data storage. For information extraction port external pages and word processing, this article will be relevant pages according to certain information content stored in the database for further semantic analysis.

2.2 Mathematical formulas and equations

Based on these theory and method, we need to further reconstruction of semantic Web content to obtain the desired web content including paper weights, containment relationship and reference relationship.

This paper weights calculated mainly using popular $tf-idf$ algorithm. Equation 1, the right w_{ik} to represent the concept of the word weight, tf_{ik} conceptual word frequency in the document.

$$w_{ik}=\frac{\sum_{k=1}^{j} tf_{ik}\times\log(N/df_k)}{\sqrt{\sum_{t_k\in i}[\sum_{k=1}^{j} tf_{ik}\times\log(N/df_k)]^2}} \tag{1}$$

Equation 1 is synonymous in the context of the relationship between the document and the corresponding similarity to quantify the weight calculation. In addition to synonyms, the concept description has two major reference relationship.

A containment relationship among the body means, a concept of contents includes the concepts expressed by b, i.e. the concept of a description of the contents of a larger than b, b concept more

precise, there is contained between the two concepts relationship. [4] In this paper, the concept of a concept called b comprises an upper, b is called a lower-level contains. For the right to existence of the relationship between the weights calculated as shown in Equation 2.

$$\Delta w_q^c = \lambda w_{q^f} + \sum_{s=1}^{n} \theta w_{q^s} \tag{2}$$

Reference relationship, that the concept of a field of knowledge and concepts described in b belong to the same category of knowledge. [5] For the right to existence of the relationship between the weights calculated as shown in Equation 3.

$$\Delta w_q^r = \sum_{d=1}^{n} \alpha w_{q^d} \tag{3}$$

Through semantic reconstruction of the text and the calculation of similarity, external customer-related information resources to conduct a preliminary calculation of a certain degree of classification and corresponding indicators of each category. [6] After the description of the body through the field of each module of text classification, this paper, the accuracy (Precision) commonly used in text classification, recall (Recall) and Fb-score to classify the results of testing the validity of the text index.

In addition, there is another formula Customer Evaluation Index Calculation. In this paper, through the acquisition of external information, building and semantic analysis of the text through the body. We can create a customer evaluation based on external information, the level of external information from the customer to evaluate the auxiliary port further decisions on customers [7].

$$\max\left[-\sum_{k=1}^{p} w_k \ln w_k\right] \tag{4}$$

$$s.t.,\ \lambda = -\sum_{k=1}^{p} \frac{p-k}{p-1} w_k\ ,\ \sum_{k=1}^{p} w_k = 1, w_k \in [0,1]$$

3. Result and Discussion

The text is try to constructing a domain ontology for external information semantic analysis for port customer, calculate customer evaluation, and provide reference by indicators for port enterprise customer management policies. But these indicators in the actual production and operation of the port whether it can play a role, therefore, the text try to make the results of semantic analysis to empirical analysis.

Firstly, we choose a port in the south of a large customer as a typical case, the client in 2014 is one of the major customers of the port. In this paper, we choose the typical case of the customer to tap external information customers and study the impact of external information on the port enterprise customer management decisions, as well as demonstrate the proposed ontology semantic analysis system for evaluation of customers are reference customers in port management.

In this paper, the customer data crawling. Specific data for each indicator as shown in *Table 1*:

Table 1 Port a customer external information index data table

Index	Text number	positive text	Negate text	Result
Finance	19	1	18	-12.2
Develop	2	0	2	-0.92
Industry	674	196	316	-14.2
Competition	3	1	2	-0.5

As can be seen from the results of each index, the customer's overall rating is poor, it was in the current downturn in the presence of a weak state in coal industry and it has a downward trend. Poor financial situation, there is a financial risk. This article from the port to get the customer's production business systems 2015 arrears each month (the amount of accounts receivable of customers, blue) and

transactions with the port (transfer amount of revenue that month trading volume, red) case data such as observation and analysis, shown in *Figure 2*:

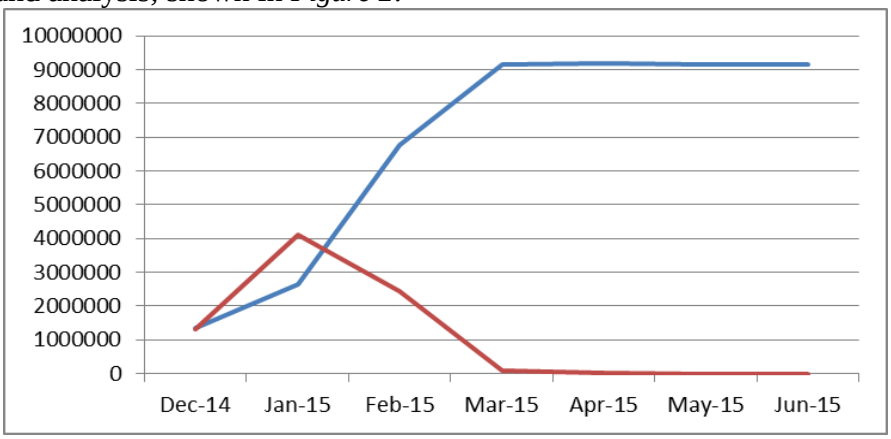

Figure 2 Financial and trading of a customer in 2015

As can be seen from the figure, the client continued to increase in 2015, not only the amount of accounts receivable is still not eased, while its port and trading volume in 2015 after the sharp drop in January 2015 and increased to May trading volume is zero. To customers, semantic analysis system able to alert port enterprises to take appropriate measures as soon as possible. They can make the financial dunning, reducing promotions and other strategies to reduce the loss of free heap. Port enterprises should quickly respond to external information, and to develop appropriate strategies.

4. Summary

For ports, the use of distributed platforms and ontology semantic analysis techniques to customers to organize and describe external information, which is still a relatively new attempt. [8] In this study, the use of the current relatively objective news text easily accessible as the main source of external data customers, achieved a certain effect. However, there are still some issues that need to be further improved.

Firstly, the external data is only conducted research related to two aspects of the industry from the customer and related customer. The customer still has a lot of external data in the relevant information about the customer's concern port [9].

Secondly, in the description of the customer's body, because of personal study and understand of the limitations of the field, there are inaccuracies and incomplete to some extent in establishing ontology.

Finally, as the body primarily responsible and logical layers semantic Web pages of information and intelligence, which also involves a lot of very complex technology, closely related to the development of knowledge engineering, artificial intelligence and other related areas. Therefore, semantic Web technology will eventually move toward standardization, which is a relatively lengthy process. [10] How to determine the logical relationship between the semantics, semantic division and more accurately analysis and research needs to be further improved.

References

[1] Sure Y Angele J, Erdmann M, Wenke D. OntoEdit: Collaborative ontology engineering for the semantic Web[C]. In: Horrock I. Proceedings of ISWC, 2002: 221-235

[2] F. Bander,D. McGnirmess. D, Nardi. The Description Logic Handbook: Theory Implementation and Applications [M]. Cambddge University Press, 2003

[3] Steffen Staab, Rudi Studer. Handbook on Ontologies [M]. Springer, 2004

[4] Skin E, Parsia B. Pellet: A Practical OWL-DL reasoned [J]. Journal of Web Semantics, 2007, 5(2): 51-53

[5] Haarslev V Moiler R. Racer System Description[C]. International Joint Conference on Automated Reasoning, 2001: 701-705.

[6] Forgy. Pete: A Fast Algorithm for the Many Pattern/Many Object Pattern Match Problem [J]. Artificial Intelligence, 1982: 17-37.

[7] Tao Li, Yi Zhang, Vikas Sindhwani. A non-negative matrix tri-factorization approach to sentiment classification with lexical prior knowledge[C]. Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP. Suntec, Singapore.2009.244-252.

[8] Shi Feng, Jun Pang, Daling Wang, Ge Yu, Feng Yang, DongpingXu. A novel approach for clustering sentiments in Chinese blogs based on Hadoop similarity [J] - Computers & Mathematics with Applications, 2011 7:2775-2777

[9] Xiangdong Li, Cheng Zhang. Research on enhancing the effectiveness of the Chinese text automatic categorization based on ICTCLAS segmentation method [C] 2013:267-269

[10] Yang Y, Pedersen J O. A comparative study on feature selection in text categorization[C]//ICML. 1997, 97: 412-420.