Research on User Requirements Elicitation Using Text Association Rule

Dong Lili, Zhang Xiang, Ye Na, Wan Xiaoge
College of Information and Control Engineering
Xi'an University of Architecture and Technology
Xi'an, 710055, China
donglilixjd@163.com, zhangxiang1001@126.com, yenanaye@126.com, wanxiaoge@sina.com

***Abstract*—User requirements obtained through text data mining are very important to improve the competitiveness of enterprises. In this paper an algorithm of acquiring user requirements in machinery products by using text association rule is proposed. In the algorithm, the user requirement documents are represented by vector space model. The feature words matrix is obtained by transposing the documents matrix. An improved text association rule theory based on gray association rule is used to calculate the correlation degree between feature words and proper nouns of machinery industry. Then the matrix of candidates for proper noun is constructed by selecting a higher correlation degree word as a threshold. Finally, user requirements are obtained by using the weighted matrix. The experimental results suggest that the proposed method is feasible for user requirement elicitation.**

Keywords-text association rule; user requirements;vector space model;gray association rule;mechanical products

I. INTRODUCTION

With the rapid development of the network technology, Internet has become the main means of publishing and obtaining information. The user requirement is the motivation of technology development. Making full use of web resources to get user requirements has great importance to grasp the market trends correctly and improve the competitiveness of enterprises. The technology of acquiring user requirements based on text association rule is to put together all the requirement information of mechanical products, which is scattered on the Internet, and then extract the genuine and professional user requirements. Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner. Text data mining is concerned with data mining methodologies applied to textual sources [1].

In text mining text association is most widely used, in which semi-structured text data is transformed into computer representation model and then the association rule algorithm is used to extract knowledge. This paper presents a new text association rule algorithm based on gray theory to obtain the user requirements. First, documents of user requirements are represented by the vector space model, and then an improved gray association rule theory is used to calculate the correlation degree between feature words and proper nouns of machinery industry. Finally, user requirement is obtained by using the weighted matrix.

The rest of this paper is organized as follows: In section 2, we will discuss the process of obtaining user requirements. Data preprocessing of user requirement is discussed in secion3. In section 4, representation model of user requirements is introduced. In section 5， we will illustrate how to obtain proper nouns of machinery industry based on Gray association rule theory. In section 6, experimental steps and results are presented. Some conclusions and ideas for further research are described finally.

II. THE PROCESS OF OBTAINING USER REQUIREMENTS

The process of obtaining user requirements for machinery industry could be divided into three steps. The first step is data preprocessing. In this step the messy user requirements are collected from Internet and rough dimension reduction of user requirements is achieved by Chinese words segmentation and the stop words elimination. The second step is constructing requirements presentation model. In this step the user requirements is represented by vector space model (VSM). The third one is user requirements elicitation. According to the improved gray association rule, the correlation degree matrix of feature words can be acquired by considering many user requirement documents and we can extract the real needs of users by weighted calculation .The process of obtaining user requirements is shown in figure 1.

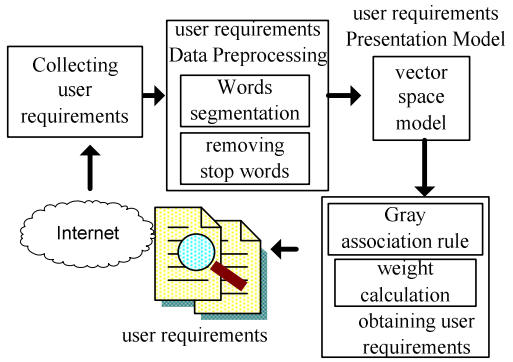

Figure 1. Flow diagram of the process of obtaining user requirements

978-0-7695-4196-9/10 $26.00 © 2010 IEEE
DOI 10.1109/IPTC.2010.72

III. USER REQUIREMENTS PREPROCESSING

Because there is no obvious separator between Chinese words, so it is necessary to divide the user requirement documents which are collected from the Internet into proper words by Chinese words segmentation. ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System) is used to segment Chinese words, which is developed by Institute of Computing Technology, Chinese Academy of Sciences. After segmentation, a table which is called text representation dictionary and composed by a series of words is formed. In order to reduce data noise of text feature vectors and the scale of text representation dictionary, the stop words and low frequency words should be removed from the dictionary and synonymous words are merged, this procedure is called rough dimensional reduction. After this process, the representation dictionary whose dimensions have been reduced can be obtained [2].

IV. REPRESENTATION MODEL OF REQUIREMENTS

The representation of user requirements is a task of converting the dictionary presentation into vector form which the computer can process. At present, the text representation model mainly uses the VSM (Vector Space Model) [3].Its basic idea is representing each text as a vector:

$$d = \{w_1, w_2, ..., w_n\} \quad (1)$$

Where w_i is the weight of feature term i in document d.

According to the vector space model, the representation model of user requirement documents can be expressed by:

$$D = \{d_1, d_2, ..., d_m\} \quad (2)$$

Each document d_i could be denoted as $d_i = \{w_{i1}, w_{i2}, ..., w_{in}\}\ i = 1,2,...m$; the weight w_{ij} is adopted Boolean Weight, if the feature term appears in the document, then the weight is 1,otherwise the weight is 0.

The D^T is the transpose of the matrix D:

$$D^T = \{t_1, t_2, ..., t_n\} \quad (3)$$

The t_i denotes a feature term which appears in the user's requirement, $t_i = \{w_{i1}, w_{i2}, ..., w_{im}\}\ i = 1...m$ .The D^T represents the situation, that is, how often the feature terms appear in every document.

V. THE ALGORITHM OF REQUIREMENT OBTAINING BASED ON TEXT ASSOCIATION RULE

A. Calculation of Correlation Degree

In 1982, the gray theory is put forward by Deng Julong, and today gray theory has developed into a new subject. Gray association rule is the uncertain association among the things or the uncertain factor associated with the main action[4] According to the matrix D^T ,this paper proposes a text association formula (4) based on the Gray association rule.

$$\varepsilon_{0i} = \frac{1 + 2s_{0i}}{1 + 2s_{0i} + |s_i - s_0|} \quad i = 1,..., m \quad (4)$$

Where $|s_i - s_0| = \sum_{k=1}^{m} |w_{ik} - w_{ok}|$; $s_{oi} = \sum_{k=1}^{m} w_{0k} \bullet w_{ik}$; s_0 is an industrial proper noun which is proposed by mechanical experts, s_i denotes the feature terms that come from the segmentation of user requirement, the correlation degree between feature words and proper nouns can be calculated by formula(4). Given the threshold T, if $\varepsilon_{0i} > T$, then the feature term s_i is selected as candidates for proper nouns.

B. Calculation of Feature Weight

Matrix of candidates for proper nouns is constructed by selecting a higher degree word from the correlation matrix. Matrix of candidates can be simplified to a single column vectors by multiplying the weight column vector of proper nouns. Higher weight proper nouns can be selected from this single column vector.

1) The candidate proper nouns matrix construction

ε_{0i} is the correlation degree between feature word s_i and s_0 , where s_i is the feature word and s_0 is the proper noun of machinery industry. If the correlation degree $\varepsilon_{0i} > T$, the feature word s_i is selected as the row of the candidate proper nouns matrix M, the column of matrix M is the proper noun w_i which is used to calculate the correlation degree of feature words. The candidate proper nouns matrix is deified as follows:

$$M = \begin{bmatrix} M_{00} & ... & M_{0n} \\ ... & ... & ... \\ M_{0n} & ... & M_{nn} \end{bmatrix} \quad (5)$$

Where M_{ij} denotes the correlation degree between feature word s_i and proper noun w_i .

2) The calculation of the weight of candidate proper nouns

The n-dimensional vector wei is constituted by the proper nouns $w_1 ... w_n$, which are selected from a database of proper nouns.

$$wei = \{weight_1, ...weight_n\} \quad (6)$$

Where $weight_i$ is the weight of w_i , which denotes the importance of the proper feature noun for some machineries, and it is given by the experience of Machinery industry experts. An n-dimension column vector v is the result of the multiplication of the M and the wei

$$M \times wei^T = v \quad (7)$$

Where v^T denotes the comprehensive weighted value $weight'$ of feature word t_i. The value v^T considers both the correlation degree between the feature word and the proper feature noun and the weight of proper nouns .The comprehensive weighted value is normalized as

$$weig_i = \frac{weight_i'}{\sum_{j=1}^{n} weitht_j} \quad (8)$$

In order to obtain a set of feature words which can describe a document, the feature words are often sorted by the comprehensive weighted values, from which the pre-k words are selected as feature words. The value k is decided by experts. Generally, about 30 words are needed to describe certain machine equipment.

C. Algorithm description

Algorithm 1 Text association rule for user requirements

Input: the feature words matrix of user requirements D^T, $Thre$ the threshold of correlation degree, The specialized feature terms $W = \{w_1, ...w_n\}$ and their weights $wei = \{wight_1, wight_2..., wight_n\}$, the count of the final feature nouns k;

Output: a set of specialized feature terms $T = \{t_1, t_2...t_k\}$;

Step1: For $i = 1$ to n do

Step 1.1: Separately use w_i as s_0, calculate the correlation degree of feature words which occur in the matrix D^T.

$$\varepsilon_{0i} = \frac{1 + 2s_{0i}}{1 + 2s_{0i} + |s_i - s_0|}$$

Step 1.2: if $\varepsilon_{0i} > \text{Thre}$, then select w_i to constitute a set of candidate feature words s ;

Step 2: Constitute the M of candidate feature words which come from S;

Step 3: $M \times wei^T = v$, according to the value of comprehensive weighted vector, the feature words t_i should be ordered;

Step 4: Select pre-k feature words as $T = \{t_1, t_2, ...t_k\}$;

Step 5: Output T.

VI. EXPERIMENT AND RESULT

In our experiment, Java is adopted as the programming language. The developing platform is Eclipse 3.3. The version of JDK is 1.5. The experimental dataset is 657 user requirements for excavator of machinery industry. After user requirements preprocessing and model representation, correlation degree of the feature words and proper noun can be calculated. When threshold of correlation degree is 0.7, there are 82 candidate feature words. Mechanical experts think that 30 proper noun are enough for describing the user requirements of mechanical equipment. So we select 30 proper nouns after weight calculation. According to 30 proper nouns, the mechanical expert constructs excavator by machine-aided design system, which is shown in figure 2.The result is consistent with the requirement document which is analyzed by experts.

Figure 2. Result of experiment

VII. CONCLUSION AND FUTURE WORKS

The paper proposes a method for obtaining user requirement in machinery industry based on text association rule. The first step is data preprocessing of user requirement. Vector space model is used to describe the user requirement. Secondly, an improved gray association rule theory is used to calculate the correlation degree between feature words and proper nouns of machinery industry. Then the matrix of candidates for proper nouns is constructed by selecting a higher correlation degree word. Finally, user requirement is obtained by using the weighted matrix. The experimental results show that the method for obtaining the user requirement is effective and efficient. Currently the dataset of user requirement on machinery industry is still relatively small. In the future, the study should be done on a large number of dataset to obtain more accurate experimental result.

REFERENCES

[1] Jiawei Han. Data Mining: Concept and Techniques [M], Morgan Kaufmann Publishers Inc.2001

[2] Zhang Xiang , Zhou Mingquan , Geng Guohua and Ye Na. A Combined Feature Selection Method for Chinese Text Categorization, 2009 International conference on Information Engineering and Computer Science.wuhan18-20 December2009:405-408.

[3] Salton G, Wong A, and Yang C S. A vector space model for automated indexing. Communications of the ACM.1975, 18(1):613-620.

[4] Deng Julong. Foundation of Gray Theory [M]. wuhan ,Hong Zhong university of Science and thechology press,2002（in Chinese）