An Automatic Reply to Customers' E-mail Queries Model with Chinese Text Mining Approach

JU-YU HUANG [1], HUEY-MING LEE [1], SHU-YEN LEE [2]

[1] Department of Information Management, Chinese Culture University,
55, Hwa-Kung Road, Yang-Ming-San, Taipei 11114, Taiwan

[2] Dep. of Private Participation in Infrastructure, China Engineering Consultants, Inc.
5Fl., 300, Sec. 4, ChungHsiao E. Road, Taipei 10694, Taiwan

Abstract: - With the prevalence of the Internet, most companies provide technical support and customer service via the net. In general, companies' websites provide keyword search interfaces to access pre-written documents. Unfortunately, customers may not always find the correct information that they need. To fulfill a customer's needs, a quick response model is needed. We proposed a Chinese e-mail questioning quick response model to enhance customer service on the Internet. This model not only can process all customer queries, transform e-mails into pre-defined template rule set, discover a suitable reply from pre-defined knowledge-based documents, but also can provide an efficient and complete customer service for commercial companies, as well as government departments and organizations.

Key-Words: - Text-mining; FAQ

1 Introduction

With the prevalence of the Internet, most companies provide technical support and/or customer service via the net. In general, companies' websites provide keyword search interfaces to access pre-written documents. Unfortunately, customers may not always find the correct information that they need. To fulfill a customer's needs, a quick response model is needed.

The 80-20 rule is based on the idea that only a few factors account for a large percentage of the total number of questions [9]. According to this rule, 80 percent of the questions come from 20 percent of the problems. If we prepare reply for these questions in advance, then the questioning quick response model is possible.

In this study, we proposed a Chinese e-mail questioning quick response model to enhance customer service on the Internet. The model is an extended work of our previous research on Chinese text mining- the knowledge discovery model in Chinese news documents [3]. The model tries to discover a suitable reply from pre-defined knowledge-based documents. If a proper pre-defined reply document can not be found, the model will forward the un-reply e-mail to the customer services department. Once the proper response is provided, the model transforms the new information into a pre-defined knowledge-based document. This model can provide an efficient and complete customer service for commercial companies, as well as government departments and organizations.

We applied text-mining technology to build an automatic reply model. The result of text-mining can be presented in different formats. We adopt TextRise algorithm [6] to induce fewer rules from many texts. These rules mean that several texts have the same trend in the sample base. In this model, we applied this characteristic to find out the terms of a specific question. People usually use the same question description in most texts which mean that they use the same terms to ask the same questions. After text mining induction, we can get a rule that is composed of keywords about a specific question. So, the rule set is the frequently-asked-question set.

The design of the automatic reply mechanism is based on the rule set. We collected correct reply for each rule. When a new e-mail is received, keywords will be extracted and compared with the rule set to decide which reply is the most suitable. We can

therefore process a large amount of e-mails using the automatic reply mechanism, thereby allowing more time for specialized queries to be processed.

The remainders of this paper are as follows: Section 2 introduces the Knowledge Discovery model in Chinese documents with text-mining. We introduce the automatic reply mechanism in Section 3, and the global view of the proposed model is discussed. Section 4 concludes the paper and discusses future research.

2 The knowledge discovery model in Chinese documents

In this section, we give a global view of the knowledge discovery model in Chinese documents. Our knowledge discovery model is derived from the general text data mining model which was discussed in the previous section. As shown in Figure 1, the model is divided into two parts: pre-process and post-process. The pre-process takes the Chinese documents as input data. The pre-process includes Chinese segmentation and information extraction. The extracted information is stored in pre-defined databases to represent the knowledge template. The post-process, Chinese Knowledge Discovery, CKD, is applied to a rule learner, named TextRise, to induce the knowledge templates into a set of rule base. Users discover interesting or helpful knowledge rules aided by a proposed interestingness measure from the rule set [3]. Therefore, users can easily obtain useful knowledge or information without having to read large text documents from the Internet or other sources.

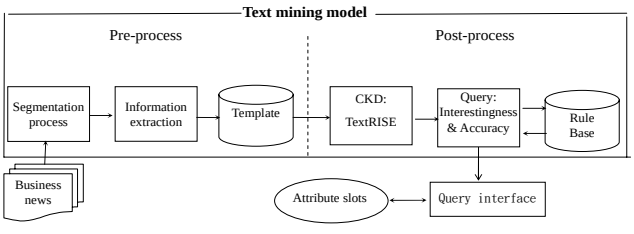

Figure 1 The knowledge discovery model in Chinese Documents

Unlike pre-process for English text documents, Chinese text documents are composed from Chinese characters without spaces. The process to divide Chinese text into segments, or phrases, is called segmentation process. There are three major approaches [1, 2] to Chinese segmentation. The first approach is dictionary-based with maximum matching. That is, the process segments Chinese text by using a pre-defined Chinese dictionary. In general, the process takes the phrase with maximum length from all possible phrases. The second approach is based on statistical methodology. The model information is to divide Chinese text into proper phrases. The characters mutual information is statistical information derived from an existing corpus. The third approach integrates the first two approaches. We modified the third approach to the Chinese text segmentation process in this uses pre-produced characters database by mutual model.

The information extraction categorizes the segmented phrases into pre-defined bags of words, or BOWs, and stores the extracted information in a database. We called the set of categories the knowledge template. The post-process of our model uses a rule learner TextRise to induce the knowledge template into knowledge rule base. The TextRise is suitable to process the rule representation that we use. The rule learner is designed for BOW-base rule learning. Therefore, we can get many key terms about one subject. This will then help us to evaluate the main crux of the query.

3 Automatic reply to e-mail queries model

This proposed model can be divided into two parts, saying the reply knowledge base and the automatic reply mechanism.

The first part is the reply knowledge base. We collected large e-mail documents to produce a rule base. First, we processed Chinese segmentation and keywords for all collected documents and then focused on the keywords in each to present a topical subject of an event. We then used dictionary-based method to extract product, and item number, and finally we got an information template.

The templates are composed of a database of relevant topics. Text mining algorithm helps us to

find out the trend for the templates. We adapt TextRise algorithm to induce lots of templates into fewer generalized rules. These rules present common queries found in many e-mail documents. In other words, many e-mail queries follow one rule. Therefore, these queries will be referred to company experts to make an reply knowledge base.

The second part is the automatic reply mechanism. When an e-mail is sent by a customer, Chinese segmentation processing is done first. We extract information from the e-mail to form a template, then compare it with the antecedence and consequence in the rule base to find a proper rule. If matching is possible, the model will provide a proper reply that is matched to the reply knowledge base.

3.1 The Chinese segmentation and information extractionFigure 2Figure 2 depicts the integrated Chinese segmentation process [4] in the proposed model. We prepared mutual information from large corpus of Chinese characters. We also prepared a stop word list for removing meaningless characters. In the segmentation process, we use dictionary-bases and MI-bases to segment the same text. The segmentation process takes the longest phrase as a result.

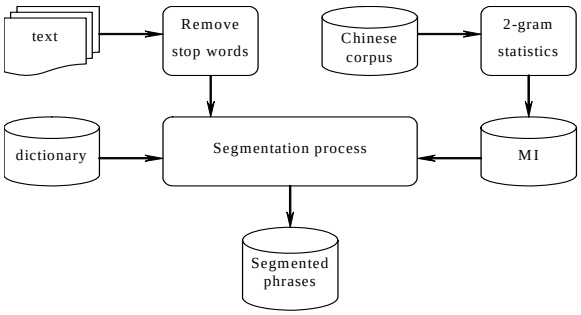

Figure 2. The integrated Chinese segmentation process

We use Sporat and Shih's approach [8] to calculate the Chinese mutual information (MI). The MI measure represents the concatenation strength of two Chinese character a and b. The MI value is calculated by the following equation [8]:

$$MI(ab) = \log_2(N) + \log_2\left(\frac{f(ab)}{f(a)\times f(b)}\right) \quad (1)$$

where Chinese character *b* appears after character *a*. *f*(*ab*) represents the times that character *b* appears after character *a*. *f*(*a*) and *f*(*b*) represent the number of times that characters a and b appear, respectively. *N* is the total number of Chinese characters in the corpus. Chinese phrase *ab* could be a Chinese phrase if their MI value is high. Chinese character sequence abc could be highly possible a true Chinese phrase if both MI(*ab*) and MI(*bc*) are high. In this way, we can possibly find a new n-gram phrase. This approach solves the deficiency of phrase-based segmentation approaches for new phrases.

Table 1. The template example.

Bag of word	Contents
Product	洗衣機 (washing machine)
Item number	W20033
Event	噪音(noise)、漏水(leak)、轉動不順(turn over problem)
Time	2007/02/14

After the segmentation process, information process categories segment phrases into pre-defined BOWs and form rule templates. The definition of BOWs is based on the characteristics of processed text documents. In this study, we are interested in the text-mining application of documents. We categorized the phrases into four BOWs, namely product, item number, time, and event. Table 1 shows a possible template example extracted from an e-mail query.

3.2 Extracting Key Terms

We use keywords to represent a subject. Therefore, there are several important terms to represent e-mail content. Our work employed three methods to compute the score of each term [1]. The several highest score terms are keywords in query.

(1) Remove single Chinese words

In general, based on experience, a single

Chinese word, for example '是' (is), can not explain a complete meaning, so it should be removed.

(2) Select important terms

We can select key terms by using TFIDF (term frequency; inverse document frequency) [7]. Term frequency is the number of times a particular term occurs in a given document or query. Inverse document frequency is a measure of how often a particular term appears across all of the documents in a collection. So, common words will have a low IDF and words unique to a document will have a high IDF.

(3) TFIDF

The TFIDF weighting scheme is used to assign higher weights to distinguished terms in a document. TFIDF makes two assumptions about the importance of a term. First, the more a term appears in the document, the more important it is. Second, the more it appears through out the entire collection of documents, the less important it is since it does not characterize that particular document very well [7]. The weight for term *ti* in a document *di*, W_i is defined as follows:

$$W_i = tf_i \times \log_2 \frac{N}{n} \qquad (2)$$

where tf_i is the frequency of term *ti* in document *di*, *N* the total number of documents in the collection, and *n* the number of documents where term *ti* occurs at least once.

3.3 The Rule Base

Induction methods induced original examples into rules to produce specific patterns, in other words, they induce each specific rule in an example set into fewer generalized rules in order to predict e-mail content. The knowledge rules of the induction method model can be expressed by antecedence and consequence. The rule is composed of antecedence and consequence. The antecedence is one or more of the conditions. The consequence is true only if the antecedence is true. Our work applied TextRise algorithm [6] to generate a rule base.

The size of the produced rule set is smaller than the size of the given example set. The new generalized rule may be pruned if an identical rule exists in the present rule set.

In our study, the rule antecedence is product and item number, the consequence is event subject. A rule example as Table 1 shows that it can be { Product:洗衣機(wash machine)},{Item number: W20033} → { Event: 噪音(noise)、漏水(leak)、轉動不順(turn over problem)}.

4 Conclusion and future works

In this study, we proposed an automatic reply model for Chinese e-mail documents with text mining. The proposed model can process a mass amount of Chinese text documents and induce them into a knowledge rule base. Its major contribution includes:

(1). The model can automatically process mass amounts of electronic Chinese text documents. The model induces these documents into a knowledge rule base, so that, users' queries can easily be processed by the automatic reply model.

(2). Using Bag of Words enables better and more complete knowledge representation.

(3). The automatic reply model can respond to customers immediately which would result in a need for less resources and ultimately lead to large cost saving.

The model is not only useful for e-mail documents, it is also suitable for other text documents if we can properly define the Bags of Words for that specific field. The post-process of the proposed model can be applied to other approach to produce better knowledge database. The neural fuzzy might also be a possible alternative to the TextRise algorithm.

Acknowledgment

The authors gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

References:

[1] Hsu, C.-C., Chen, J.-K.,:.“*Data Mining in Chinese News Articles*”, Journal of Information Management,7(2), 2001, 103-122

[2] Hu, S.-J. & Hsu, C.-C.,:” *Word Segmentation in Chinese News Articles*”, Proceedings of the 10th International Conference on Information Management, 1999, pp. 968-974, Taiwan.

[3] Huang, J.-Y., Lee H.-M., and Chen, W.-Y.,: “*Industrial News Knowledge Discovery with Text Mining Approach*”, Proceedings of the 7th Conference on Information Management and Practice (CSIM2001)[CD-ROM], 2001, Taipei, Taiwan.

[4] Huang, J.-Y., Lee, H.-M. (2002), *Knowledge discovery model in Chinese industrial news*. Proceedings of the Second International Conference on Electronic Business (ICEB-2002) (pp.412-414), Taipei, Taiwan.

[5] Huang, J.-Y., Lee, H.-M.,: “*Automatic information extraction in Chinese industrial news*”, Proceeding of the Third Conference on Information Management, pp.861-869, 2002

[6] Nahm, U. Y., and Mooney, R. J.,:”*Mining soft-matching rules from textual data*”, Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI-01) , 2001, pp. 979-984, Seattle, Washington.

[7] Salton, G.,: ”*Automatic text processing: The transformation, analysis and retrieval of information by computer*”, Massachusetts: Addsion-Wesley. (1989)

[8] Sporat, R., Shih, C. A.,: “*A statistical method for finding word boundaries in Chinese text*”, Computer Processing of Chinese and Oriental Languages, 4(4), 1990, 336-351.

[9] Stevenson, W. J., : *Operations Management*, seventh edition. McGraw-Hill Companies. 2002