Association for Information Systems

AIS Electronic Library (AISeL)

ICEB 2002 Proceedings

International Conference on Electronic Business (ICEB)

Winter 12-10-2002

Knowledge Discovery Model in Chinese Industrial News

Ju-Yu Huang

Huey-Ming Lee

Knowledge Discovery Model in Chinese Industrial News

Ju-Yu Huang
Department of Information Management
Jin-Wen Institute of Technology
Taipei, Taiwan
stella@jwit.edu.tw

Huey-Ming Lee
Department of Information Management
Chinese Culture University
Taipei, Taiwan
hmlee@faculty.pccu.edu

Abstract

With prevalence of Internet, users can easily retrieve the information what they want from Internet. Information explosion shows that efficient information summarization is aspired to all users. Therefore, an efficient knowledge management methodology becomes very important. Some technologies, such as text mining, for acquiring knowledge from huge amount of electronic documents are recognized as important technology in this field.

This work focuses on text-mining applications on Chinese industrial news and knowledge discovery. We use information extract method to extract news into companies, event keyword, time, location, and person categories based on the characteristics of news. The set of five extracted categories is called information template. The templates are summarized by rule induction. We can discover unexpected knowledge from these summarized rules. We built an integrated industrial news text-mining model by using induction rule learner. This model is suitable to manipulate rules in bag-of-word form. Furthermore, we proposed interestingness to measure interesting strength of rules. The users can analyze the discovered rules based this measure. These are helpful to discover unexpected knowledge. It is meaningful to commercial activities if we can discover valuable rules. Besides industrial news application, we believe this model is suitable for knowledge discovery application in other fields.

1. Introduction

With prevalence of Internet, users can easily retrieve the information what they want from Internet. Unfortunately, people could not promptly acquire knowledge from this information source without computer processing technology. Therefore, an efficient knowledge extraction technology becomes very important. Text mining is recognized as one of the most important technologies for discovering knowledge from huge amount of electronic documents. In recent years, the technologies of knowledge discovery from database (KDD) have been well developed. Most researchers proposed text-mining models based on well-developed data mining technologies [7]. The purpose of knowledge discovery is to extract implicit, unknown, and unexpected information [2].

Figure 1 depicts a text-mining model extended from data mining. In order to transform ordinary text document into predefined database, an information extraction (IE) process is developed to extract necessary information from text documents. IE process involves natural language process technologies [3]. After transforming unstructured text documents into structured information database, the system applies suitable data-mining technologies to post-process the extracted information for completing knowledge discovery.

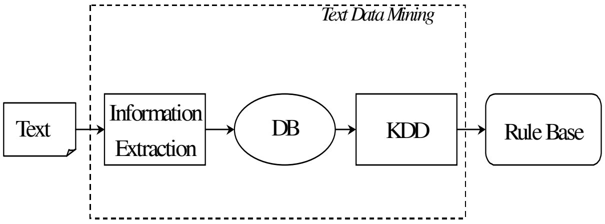


Figure 1 Text-mining model extended from data mining [7]

Text mining is to discover useful information pattern from huge amount of electronic documents. The discovered information pattern can be transformed into specific knowledge presentation form. Once the extract information can be transformed into knowledge presentation database, we achieve the knowledge discovery purpose. In this study, we focus on industry related news report. We applied mutual information (MI) technology to segment Chinese text document into phrases and statistical method to select keywords of the text documents. The extracted information is transformed into predefined rule templates to represent the extracted knowledge fact. The extracted rule templates are stored in database and we applied knowledge discovery technologies, namely rule learners, to discover unexpected knowledge rule form the database. The produced knowledge rules are stored in database. As a result, our knowledge

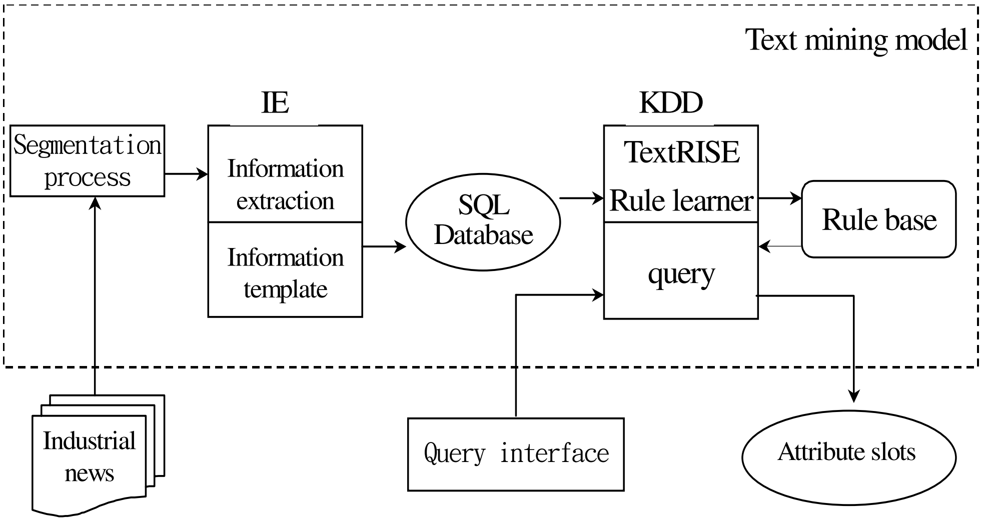


Figure 2 The knowledge discovery model in Chinese Industrial News

database is created and offers convenient query interface for knowledge rule querying. We have to remind readers that not every minded knowledge rule is interesting to all users. Furthermore, we introduced a measure, interestingness, to help users find out the knowledge rules what they are interested.

In Section 2, we introduce our Knowledge Discovery model in Chinese industrial news with text-mining. The globalview of the proposed model is discuessed.The detail discussion of the text preprocess is shown in Section 3. In section 4, we discuss the rule learner algorithm TextRise what we use in our model in detail. We introduced the interestingness measure in Section 5.

2. The Knowledge Discovery Model in Chinese Industrial News

In this section, we give the global view of the knowledge discovery model in Chinese industrial news. Our knowledge discovery model is derived from the general text data mining model which is discussed in previous section. As shown in Figure 2, the model is divided in two parts: pre-process and post-process. The pre-process part takes the Chinese news reports are input data. The pre-processes include Chinese segmentation and information extraction. The extracted information is stored in predefined format database to represent the knowledge template. The post-process, KDD, part is applied a rule learner, named TextRise, to induce the knowledge templates into a set of rule base. The users find out the interesting or helpful knowledge rules aided by a proposed measure, interestingness, from the rule set. In such a way, the users can easily find out useful knowledge or information without reading huge amount of text documents from Internet or other sources.

Unlike pre-process for English text documents, Chinese text documents are composited from Chinese characters without spaces. The process to divide Chinese text into segments, or phrases, are called segmentation process. There are three major approaches for Chinese segmentation. The first approach is dictionary-base with maximum matching. That is, the process segments Chinese text by looking up a pre-defined Chinese dictionary. In general, the process takes the phrase with maximum length from all candidate phrases. The second approach is based on statistical methodology. The system uses a pre-produced characters mutual information database to divide Chinese text into proper phrases. The character mutual information is a statistical information from an exist corpus. The third approach integrates the first two approached to have both advantages of these two approaches. We use the third approach as part of our Chinese text segmentation process. We will discuss the segmentation in detail in next section.

The information extraction categorizes the segmented phrases into pre-defined bags of words, or BOWs, and stores the extracted information into database. We called the set of categories as 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 rule representation we use. The rule learner is designed for BOW-base rule learning. Actually, the users are not interested to all mined rules. We introduce a measure interestingness to help users find out knowledge rules what they are interested. We shall discuss this part in Section 4.

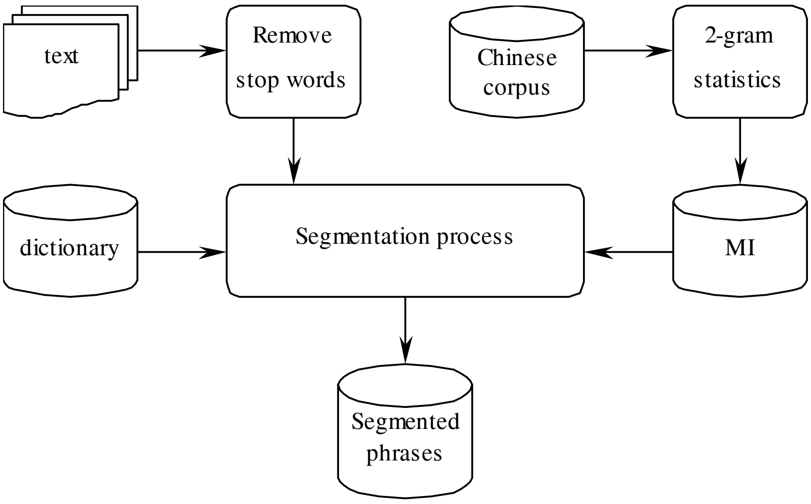


Figure 3 The integrated Chinese segmentation process

3. The Chinese segmentation and information extraction

Figure 3 depicts our integrated Chinese segmentation process in the proposed model. We prepared Chinese character mutual information from huge amount of Chinese corpus. We also prepared a Chinese stop word list for removing meaningless Chinese stop words first. In the segmentation process, we use dictionary-base and MI-base to segment the same text. The segmentation process takes the longest phrase as a result.

We use Sporat and Shih's approach [9] 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 equation (1):

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

where Chinese character *b* appears after character *a*. $f(ab)$ represents the times what character *b* appears after character *a*. $f(a)$ and $f(b)$ represent the appearing times of character *a* and *b* 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 such a way, we can possibly find a new n-gram phrase. This approach solves the deficiency of phrase-base segmentation approach for new phrase. For example, we find that the both MI(范巽) and MI(巽綠) are high, 15.9 and 14.4 respectively, and 3-gram “范巽綠” is not in our dictionary. Due to the high MI values of both bi-gram “范巽” and “巽綠”, our segmentation process can successfully segment the present Taiwan vice-minister of education “范巽綠”.

Table 1 Example of finding new Chinese phrase

bi-gram	MI value
長范	4.8009
范巽	15.9413
巽綠	14.4544

After segmentation process, information process categories all segmented phrases into pre-defined BOWs and form rule templates. The definition of BOWs is based on the characteristics of processed text documents. In this work, we are interested in text-mining application of industrial news reports. We categorize the phrases into five BOWs, namely people, company, location, time, and event. Table 2 shows such an possible template example which is extracted from an industrial new report.

Table 2 The template example

Bag of words	Contents
Company	文化公司
Event	股票上市、開高走高
Time	民國 89 年第 3 季
Location	文化資訊公司
People	王財得、林大義

4. The TextRise algorithm

The TextRise is derived from Rise algorithm [8]. In order to process text information, the rule distance is calculated as the text similarity. The TextRise algorithm is shown in Figure 4. The algorithm uses TextAccuarcy value to determine whether to terminate the induction process or not. The TextAccuracy is defined as the average accuracy of rule set to the given example set. The algorithm stops when none of rule generalization can improve the TextAccuracy of rule set on given example set.

```
Input: ES is the training set.
Output: RS is the rule set.
Function TextRISE (ES)
RS := ES.
Compute TextAccuracy (RS,ES).
Repeat
For each rule R∈ RS,
     Ê:=arg max Similaity(E,R)
        E∈ES'
      where ES' = { E':E'∈ES and E' is not covered by R}
       R':=Most_Specific_Generalization(R, Ê)
       RS':=RS with R replaced by R'
       If TextAccuracy(RS',ES)≥TextAccuracy(RS,ES)
          Then RS:=RS'
          If R' is identical to another rule in RS,
                Then delete R' from RS.
Until no increase in TextAccuracy(RS,ES) is obtained.
Return RS.
```

Figure 4 The TextRise algorithm [8]

We have to remind readers that the size of the produced rule set is smaller than size of the given example set. The new generalized rule may be pruned if there exists identical rule in present rule set.

5. Interestingness measure

The induction rule base produced from the TextRise will be huge when the system processes mass amount of text news reports. We introduce an interestingness measure to help users finding valuable knowledge from our system. Actually, a single rule in the rule base may not be interesting to everybody. We define the proposed interestingness measure as:

$$I(k,\tau) = \frac{\left|\{b \in D \mid \forall A_i, a_i, sim(A_i, a_i) \geq \tau, sim(C,c) \geq \tau\}\right|}{\left|\{b \in D \mid \forall A_i, a_i, sim(A_i, a_i) \geq \tau\}\right|} \quad (2)$$

where $\boldsymbol{k} \equiv \boldsymbol{A} \rightarrow \boldsymbol{C}$ is a predicate rule. *A* is antecedence of the predicate rule *k*, and *C* is the consequence. *D* is the knowledge rule base produced by our system. The τ is the user given threshold. The predicate rule *k* is interesting to user if its interestingness measure is high. Otherwise, the rule is useless to a user.

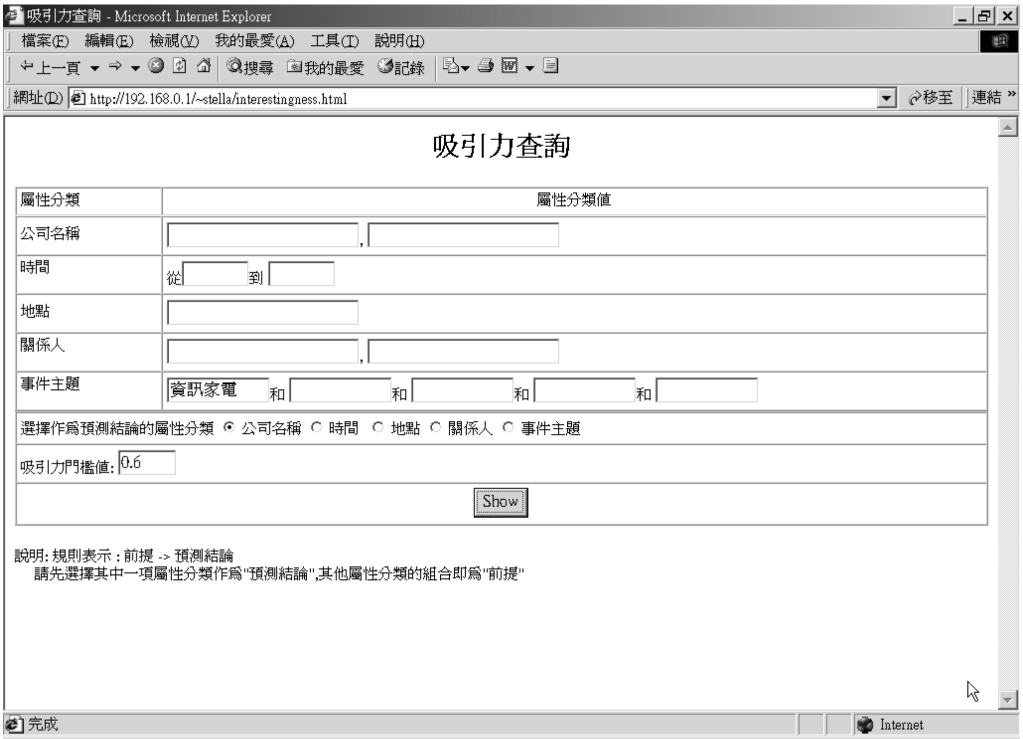
Figure 5 depicts an interestingness query example. The query result is shown in Figure 6. The example shows query about "information consumer electrical products". The system returns four related rules from rule base with their interestingness measure.

6. Conclusion

In this study, we propose an automatic knowledge discovery model in Chinese Industrial news with text mining. The proposed model can process mass amount of Chinese text documents and induce them into knowledge rule base. The major contributions are:

1. Define a useful measure, interestingness, to help users find out the useful or unexpected knowledge from our rule base.
2. The system can automatically process mass amount of electronic Chinese text documents. The system induces these documents into knowledge rule base. Such that, the users can easily discover knowledge by query the rule base.
3. Using bag of words has better and complete knowledge representation.

Our approach is not only for industrial news reports. It is also suitable for other field of text documents if we can properly define the bags of words for the specific field. The post-process of the proposed model can be applied other approach to produce better knowledge database. The neural fuzzy might be a possible approach to replace the TextRise algorithm.



吸引力查詢 - Microsoft Internet Explorer
檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)
上一頁 搜尋 我的最愛 記錄
網址(D) http://192.168.0.1/~stella/interestingness.html
移至 連結
完成
Internet

吸引力查詢

屬性分類	屬性分類值
公司名稱	,
時間	從 到
地點	
關係人	,
事件主題	資訊家電 和 和 和 和

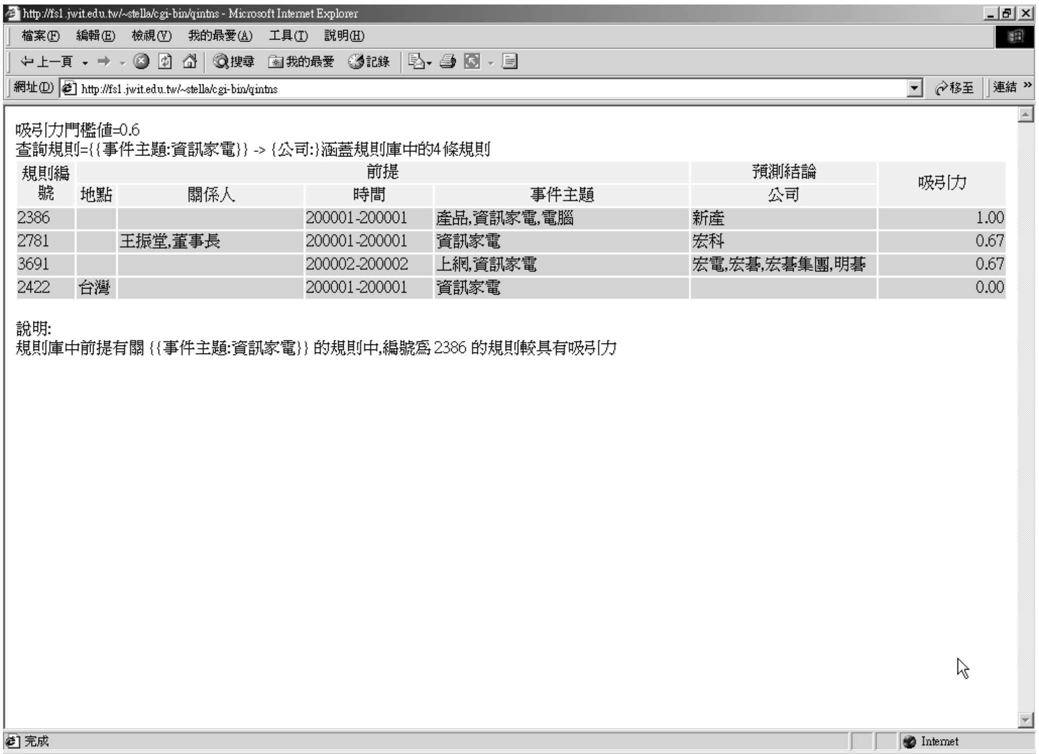
選擇作爲預測結論的屬性分類 ● 公司名稱 ○ 時間 ○ 地點 ○ 關係人 ○ 事件主題

吸引力門檻值: 0.6

Show

說明: 規則表示 : 前提 -> 預測結論
請先選擇其中一項屬性分類作爲"預測結論",其他屬性分類的組合即爲"前提"

Figure 5 The interestingness query example



http://fs1.jwit.edu.tw/~stella/cgi-bin/qintns - Microsoft Internet Explorer
檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)
上一頁 搜尋 我的最愛 記錄
網址(D) http://fs1.jwit.edu.tw/~stella/cgi-bin/qintns
移至 連結
完成
Internet

吸引力門檻值=0.6
查詢規則={{事件主題:資訊家電}} -> {公司:}涵蓋規則庫中的4條規則

規則編號	前提				預測結論	吸引力
	地點	關係人	時間	事件主題	公司	
2386			200001-200001	產品,資訊家電,電腦	新產	1.00
2781		王振堂,董事長	200001-200001	資訊家電	宏科	0.67
3691			200002-200002	上網,資訊家電	宏電,宏碁,宏碁集團,明碁	0.67
2422	台灣		200001-200001	資訊家電		0.00

說明:
規則庫中前提有關 {{事件主題:資訊家電}} 的規則中,編號爲 2386 的規則較具有吸引力

Figure 6 The interestingness query result

References

[1] Berson, A., Smith, S., & Thearling, K. (2000), Building Data Mining Applications for CRM, The McGraw-Hill Companies, New York.

[2] Brachman, R. J., & Anand, T. (1996). In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, & R. Uthurusamy (Eds.), Advances in knowledge discovery and data mining: The process of knowledge discovery in databases (pp.38-56). AAAI Press / The MIT Press, Menlo Park, California.

[3] DeJong, G. (1982). An overview of the frump system. In W. B. Lehnert & M. H. Ringle (Eds.), Proceedings of Strategies for Natural Langue Processing (pp. 149-176), Erlbaum.

[4] Hsu, C. C., & Chen, J. K. (2001). Data Mining in Chinese News Articles, Journal of Information Management,7(2)，103-122。

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

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

[7] Nahm, U. Y., & Mooney, R. J. (2000). Using information extraction to aid the discovery of prediction rules from text. In KDD Committee (Ed.), Proceedings of the KDD-2000 Workshop on Text Mining (pp. 51-58), Boston, Massachusetts.

[8] Nahm, U. Y., & Mooney, R. J. (2001). Mining soft-matching rules from textual data. In IJCAI Committee (Ed.), Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI-01) (pp. 979-984), Seattle, Washington.

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