Using Lexical Chain in Ontology-Based Information Extraction

Chunyu Cong[1, a], * Rui Gao[2, b], Zhongying Wang[3], Xiao Meng[4]

[1]Changchun University of Chinese Medicine, Changchun, Jilin, 130117, China

[2]Aviation University of Air Force, Changchun, Jilin, 130022, China

[3]Continental Automotive Changchun Co.,Ltd , Changchun, Jilin, 130033, China

[4]Chinatelcom Jilin Corporation, Changchun, Jilin, 130000, China

[a]644497215@qq.com, [b]gaor_088@163.com

Keywords: Ontology, Lexical Chain, Information Extraction

Abstract. Due to the establishment of domain ontology is prior to information extraction activity, therefore it is impossible to extract the relevant information when new issue from new field is coming. These documents which include new issues become isolated text in text set. The information in these documents cannot be extracted by the original ontology. To solve this defect the relevant information in isolated text will be extracted by using of the method of lexical chain in extraction module in this paper. The new extracted information can extend the original ontology which can lead the possibility of ontology self-perfection. The experiment results show that the refined ontology can extract more accurate information.

Introduction

The IEs (Information Extraction) based on ontology is dedicated to certain field, not only it can extract the entity but also the relationship to entity. By using of the ontology could link between statement text information and semantic effectively and extract the semantic information of implicit text to make the text understanding naturally.

The methods of ontology-based IE fall in two broad categories: document-driven IE and ontology-driven IE.

The model of ontology-driven IE is proposed by D.W.Embley for the first time [1]. Later David W.Embley optimized it, a record extractor was added which can extract the content of the HTML website automatically. At the meanwhile applied it to some different fields which had to be proved performance excellent for extraction [2] [3]. But what the user interested to ontology is changing always which can be seen by the text set they use. Although David W.Embley's model is ontology-driven, ontology will not change once as input to system, the adaption capability need to be improved. This is the reason why Burcu Yildiz and Silvia Miksch invented a new ontology-driven [4] based on David's model. The principle is a new ontology management module adding which can make the ontology in system learning from text set and in the meanwhile continuously optimize ontology to adapt text set by adding or getting rid of some assembly. The ontology management module had increased the automation and adaption rate for IES. Later Burcu Yildiz and Silvia Miksch applied it to specific application which named ontoX system [5].

The information extraction module in this paper extract out new content from isolated text by using of the lexical chain method and extension then optimizing ontology. The system is always performing the repeat work, the new information extracted by lexical chain can extend the ontology, the seeded vocabulary and relevant words can be generated by using of the optimized ontology in learning module which also make it possible to extract the needed text in extraction module [6] [7].

The establishing of the extraction module

In the paper three modules are applied, they are learning module extraction module and extension module, the relationship among them in Fig.1. The hazard can be extracted accurately

from food complaint documents by doing the cooperation among the three modules.

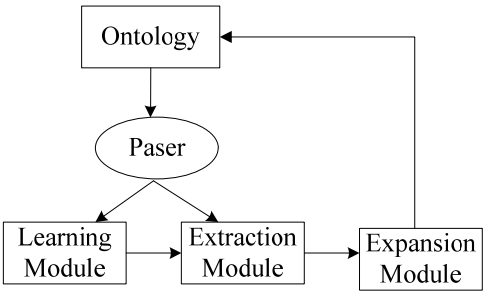

Fig.1. System architecture

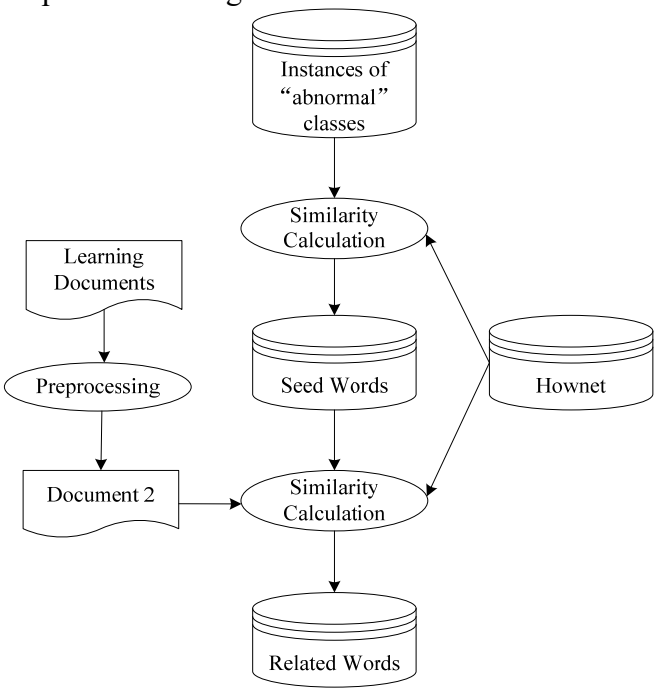

Fig.2. The learning module

Learning module

First, preparation of a certain number of dairy complaint documents which are used as learning documents downloaded from the internet must be completed. The principle of learning module is shown in Figure 2. As it is shown in the picture, the learning documents need to be preprocessed, and the ontology needs to be parsed. After that, similarity calculation needs to be done among the parsing results of the ontology in order to generate seed words. In the following step, we calculate similarity between seed words and preprocessed words to generate related words of each seed word. The selection algorithm is as follows:

```
Input:
        Seed word and candidate seed words
Output:
        Seed word and the window of that seed word
Procedure:
        If (there are other seed words in candidate seed words)
           {
           Count the number of other seed words and denote the number by n;
           Sort the other seed words according to the similarity;
           If (n>10)
              {
               Take the top 10 words as the related words of that seed word;
               }
            Else
               {
                Put the other seed words in the window;
                Sort non-seed words according to the similarity;
                Take the top 10-n words as the related words of that seed word;
               }
           }
        Else
          {
           Sort non-seed words according to the similarity;
           Take the top 10 words as the related words of that seed word;
           }
        End
```

Fig.3. Selection of related words

Extraction module

Now we can extract hazard information from food complaint documents with the knowledge we have learned in the learning module. Three "extraction" processes are essential to this module, the first one is the extraction of background knowledge, the second one is the extraction of negative information, and the third one is the extraction of hazard information in each single document. At last we combine the three types of information together to give a better explanation of the food complaint document.

Extension module

The extraction module is used for the text in complaint documents which exist and can not be recognized in extraction module. The research indicated that the hazard information from complaint document which can not be extracted can not state as not existing but the hazard information could not be reflected in ontology database. Due to the ontology database was created in former time the hazard information was also outdated. With the time passed new food complaint will coming but the hazard information is not in the ontology, so it only can extract the background information but not hazard information. The extraction module extracts the new hazard information by lexical chain extend ontology to realize ontology self-perfection. The framework of the extension module as showed in Fig.4.

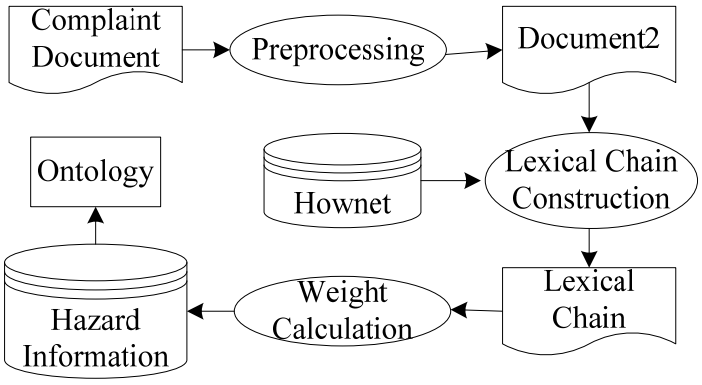

Fig.4. The extension module

Experimental Results

The traditional IE evaluation was adopted in this paper to see the performance regarding hazard information extraction, they are accurate rate P, recall rate R and measuring value F, the formula(1)(2) as following:

$$\mathrm{P} = \frac{manually\_selected \cap machine\_selected}{machine\_selected} \tag{1}$$

$$\mathrm{R} = \frac{manually\_selected \cap machine\_selected}{mannually\_selected} \tag{2}$$

The "manually_selected" means the hazard information in complaint documents was artificially extracted, "machine_selected" means the hazard information in complaint documents was extracted based on document-driven method. Accurate rate and recall rate only can describe the extraction performance partly, therefore a method(F measuring value) that considered both accurate rate and recall rate is necessary, the formula(3) as following

$$F\_measure = \frac{2 \times P \times R}{P + R} \tag{3}$$

The paper [8] download 500 complaint documents over internet as learning documents and the

1500 complaint documents as experimental text ,the experiment showed good performance. In this paper additional 200 complaint documents were involved, totally 700 complaint documents as learning documents and additional 200 complaint documents were involved, totally 1700 complaint documents as experimental documents.

In learning module, keep the vocabulary which visible frequency over 2 in ducument2. Classify the extracted hazard information into 66 clusters, each cluster contain one or more hazard vocabularies. Artificially choosing the representative seeded vocabulary for the clusters, if there is a hazard vocabulary in the cluster it is the seeded vocabulary, if there are several hazard vocabularies in one cluster select the seeded vocabulary artificially. Later generating of the relevant words to seeded vocabulary by using of algorithm1 and at the meanwhile setting of the upper threshold index as 0.788361 and lower value as 0.2857139.

In extraction module, the 1700 complaint documents as experimental text which among them 1200 documents belong to dairy product and 500 belong to other fields. The results show in table1.

Table 1 The results of extraction module.

P	R	F_measure
95.47%	93.5%	94.38%

After finishing of the extraction module 70 complaint documents were sent to extension module for extraction of the new hazard information. In extension module the threshold was set as 0.2857139, α、β、γ were set as 1, 26 out of 70 complaint documents has indication description "sexual prematurity" incident, therefore "sexual prematurity" was extracted and extended the ontology. After the extension of the ontology finishing, the hazard information extraction had been re-performed towards 1700 complaints documents, results showed in table2.

Table 2 The results of extraction by refined ontology

P	R	F_measure
97.5%	96.2%	96.76%

References

[1] Embley D W, Campbell D M, Smith R D, et al. Ontology-based extraction and structuring of information from data-rich unstructured documents[C]. Proceedings of the seventh international conference on Information and knowledge management. ACM, 1998: 52-59.

[2] Embley D W, Campbell D M, Jiang Y S, et al. A conceptual-modeling approach to extracting data from the web[C]. International Conference on Conceptual Modeling. Springer Berlin Heidelberg, 1998: 78-91.

[3] Embley D W, Campbell D M, Jiang Y S, et al. Conceptual-model-based data extraction from multiple-record Web pages[J]. Data & Knowledge Engineering, 1999, 31(3): 227-251.

[4] Yildiz B, Miksch S.. Motivating ontology-driven information extraction [C]. International Conference on Semantic Web and Digital Libraries. Indian Statistical Institute Platinum Jubilee Conference Series, 2007: 45–53.

[5] Yildiz B, Miksch S. ontoX-a method for ontology-driven information extraction[C]. International conference on computational science and its applications. Springer Berlin Heidelberg, 2007: 660-673.

[6] Victoria Uren , Philipp Cimiano , Jos´e Iria , Siegfried Handschuh , Maria Vargas-Vera , Enrico Motta , Fabio Ciravegna. Semantic annotation for knowledge management: Requirements and a survey of the state of the art [J]. In Journal of Web Semantics: Science, Services and Agents on the World Wide Web 2005.

[7] Buitelaar P, Eigner T. Topic extraction from scientific literature for competency management[C]. The 7th International Semantic Web Conference. 2008.

[8] Gao R, Zhang Y, Deng H, et al. Ontology Self-Perfect Based Lexical Chain for IE[C]. Applied Mechanics and Materials. Trans Tech Publications, 2014, 644: 1972-1975.