Research on Automatic Construction of Financial Ontology Using Chinese Encyclopedia Resource[1]

Mo Qian[1, a], Zhang Shu[1, b]

[1] Institute of Computer and Information Engineering,

Beijing Technology and Business University. Fucheng Road No. 33, Haidian District, Beijing, China

[a] moqian@th.btbu.edu.cn , [b]zhangshu0298@163.com

Keywords: Ontology, Financial, Encyclopedia.

Abstract. Ontology plays a dominant role in a growing number of different fields, such as information retrieval, artificial intelligence, semantic Web and knowledge management, etc. However, manual construction of large ontology is not feasible. This article discusses how to create Financial Ontology automatically from a resource of Chinese Encyclopedia. Financial Ontology includes Is-A relationship, Class-Instance relationship, Attribute-of relationship and Synonym relationship. Experimental Results show us that the constructed Financial Ontology has great advantages in the large scale, creation cost and the richness of semantic information.

Introduction

In the field of information science, ontology is a formal, shared conceptual model which has a clear description [1]. Domain ontology can be widely used in a variety of information applications, such as information retrieval, information extraction, summary and question answering system. However, most of the existing systems can only do reasoning and obtain new knowledge based on a large number of manual input knowledge provided by domain experts. And it is clear that creating ontology manually is very time-consuming, laborious, prone to bias errors, and difficult to be updated dynamically in time. As a new information resource, the open web-based encyclopedia has attracted more and more attentions from researchers. Encyclopedic resources (e.g. Wikipedia) have rich lexical information, semi-structured feature and good update capability. Therefore, using these resources to obtain ontology knowledge has become increasingly popular.

Interactive Encyclopedia (Hudong.com) is the largest Chinese Encyclopedia in the world. Compared with the Chinese Wikipedia, Interactive Encyclopedia has richer vocabulary resources. So we used Interactive Encyclopedia as a knowledge resource and proposed a method of creating large scale Financial Ontology automatically.

Related Work

Nakayama et al.[2] has proposed Wikipedia Mining which is a mining for Wikipedia. They have mentioned the following characteristics of Wikipedia: Dense link structure, URL as an identifier, Brief link texts, and live update.

Auer et al.'s DBpedia [3] constructed a large information database to extract RDF from semi-structured information resources of Wikipedia. However, their properties and classes are built manually and there are only 170 classes in the DBpedia.

Lian Li et al. [4] proposed a method to construct domain ontology from the Chinese Wiki automatically. The main idea is based on the entry segmenting and Feature Text (FT) extracting. Secondly, they extract the FTs from the descriptions of entries to eliminate the redundant information. Finally the similarity between pairs of FTs is calculated to revise the concept hierarchy and gain non-taxonomy relations between concepts.

Fabian et al. [5] proposed YAGO which enhanced WordNet by using the Conceptual Category.

[1] This paper is support by Beijing City Board of Education Science and technology innovation platform construction project (PSM2011_014213_113631).

Published by Atlantis Press, Paris, France.

However, this method cannot be applied to the Chinese language due to the fact that the techniques depend on English grammar.

Ponzetto et al. [6] tried to extract Is-A relationship from category tree of Wikipedia. The main method in this paper called Syntax-based methods is the simple matching method.

Building Financial Ontology

In this paper, Interactive Encyclopedia is used as a knowledge resource to construct ontology automatically in the domain of Financial, where the division of financial domain is mainly based on Interactive Encyclopedia category tree. There are 13 top nodes in the tree, which are nature, culture, people, history, life, society, art, economy, science, sports, technology, geography, and hot terms. We chose the financial class under economic category as the source data, and then did some analysis to derive ontology from the domain of financial, which is called Financial Ontology, including four relationships: Is-A relationship, Class-instance relationship, Attribute-of relationship, Synonyms relationship.

Building Is-A Relationship

We defined a conceptual category as a class, and it constitutes Is-A relationship with the Sub category. During the data collection, we found that Interactive Encyclopedia has a good structure of hierarchical relationship, in which there usually is a category page displaying the current category: the parent category and the sub category. Using this clear hierarchy, we can easily extract Is-A relationship.

It should be clear that not all the classes in all category levels of Interactive Encyclopedia can be considered as concept category, some of which are instances as confusion, for which we need to prune the classes in the class hierarchy. The basic idea of pruning is to cut off all non-conceptual categories in the class hierarchy using some experience rules:

Rule 1: a class whose name represents time such as year, etc;

Rule 2: a class whose name represents bibliography with a "《》";

Rule 3: a class whose name starts with numbers;

Rule 4: the class whose name is consisted of English letters without a subclass.

There is a basic understanding [7]: the frequency of domain concept occurred in the domain corpus is higher than that in the general corpus. Thus, if the frequency of a concept in the domain corpus is lower than or equal to that in the general corpus, it means that it probably has nothing to do with the domain. Based on the above rules and refinement methods, we obtained the following algorithm to strike classes of financial domain:

```
Function Fdomain(C)
Input: Hudong.com category set C in financial
Output: Financial conceptual class S
Var c: String; flag, nDomains: Boolean; FN, nFN: int;
Begin:
  While((c=C.next())!=null){
      flag=ConceptualClassFilter(c){
          if (c contains “《》”)   return false;
          else if (c matches pattern “ ^\d+.*”)   return false;
          else if (c matches pattern “.*(\d+)(year|period|era|century)[2].*”)   return false;
          else if (c matches pattern “\w+$” and c has no subclass)   return false;
          else return true; }
       if(flag){
          nDomains= ConceptualClassFilter2(c){
              FN= NumofFDomain(c);
              nFN=NumofNFDomain(c);
```

[2] This pattern is in accordance with the Chinese habit of writing, and may not apply to English.

Published by Atlantis Press, Paris, France.

```
                if(FN>nFN) return true;
                else    return false;}
          if(nDomains)    S.add(c);}
    }
  End
```

Building Class-Instance Relationship

We defined the terms which belong to the conceptual category as instances of the class. In the page structure of Interactive Encyclopedia, there is a kind of pages that gives all the terms affiliated with their classes, so taking the advantage of this structure, we can extract Class-Instance relationship easily. Usually, an instance of a subclass must also belong to its super-class. A class term page gives all the terms of one class, among which there may be duplicate ones that may be given in its subclass. We need to exclude them, and delete the terms in the super class which also belong to the subclass.

Acquisition of Attributes

By analyzing the Interactive Encyclopedia pages, we found that there are many differences in the content of different term articles but little in the generally structure. For instance, if a term is edited soundly, there will be a basic information box in the right of the content page in which its basic information will be listed. We can extract the public part of the terms belong to the same subcategory as the class attribute. The attribute extraction is gradually improved from the lower level to the higher one.

Extraction of Synonyms

From the view of information retrieval, Chinese synonym is mainly divided into scientific names and common ones, full names and short ones, new calls and former ones, model or code, Chinese-English translation word, abbreviations, transliterations, etc. In this paper, the method of feature pattern matching [8] is used to extract synonyms appearing in term articles. In this study, synonym extraction patterns are derived by artificial induction randomly from 10,000 term corpus.

Most of the patterns defined in feature pattern matching method, usually appear in the definition of the term. In order to improve the efficiency of the algorithm, in this paper, a sliding window with the size about 200 characters was selected to go further text anglicize according to Wikipedia's compilation rules, which means we only analysis the first 200 characters of the term text. We match the text in the fixed window with that predefined extraction pattern. If successful, we can directly extract the synonyms.

Experimental Results and Observation

An Overall Observation

Our Financial Ontology has 5,441 classes, 192,185 instances and 247 attributes. All relationships are shown in Table 1.

In view of the human factor in judging domains, this paper uses a user-based assessment method. Randomly, we extract 1,000 samples from gained classes, and let two domain experts mark the correct and incorrect results of samples, then contrast the two judge results. For the inconsistency, we make third party coordination, and finally a unified test results is achieved as assessment criteria.

Table1. All relationships of Financial Ontology

Relationship	Number
IS-A	25,578
Class-instance	790,998
Attribute-of	11,979
Synonyms	6,238
Total	834,793

We use the formula (1) as a 95% confidence level result evaluation function, in which, n indicates the number of samples, N indicates the number of the whole sets, and p indicates accuracy of experimental result. The classification accuracy of our algorithm is: $85\% \pm 1.09\%$.

$$[p-1.96\sqrt{(1-\frac{n}{N})\frac{p(1-p)}{n-1}}, p+1.96\sqrt{(1-\frac{n}{N})\frac{p(1-p)}{n-1}}]. \qquad (1)$$

Through the experiment, we know that the accuracy of synonyms is $81.2\% \pm 1.22\%$. Although the use of pattern matching can extract effectively synonyms from article text and the extraction accuracy could be further improved through continuously improving precision of the patterns, it is hard to make great improvement just from the pattern matching. The completeness of term information and the frequent appearance of attributes are the reasons why the number of attributes we extracted is small.

Comparative analysis

We designed a comparison experiment with “Chinese Category Thesaurus"(CCT). Some results are shown in Table 2, Table 3 and Figure 1, in which FO stands for Financial Ontology.

Table 2. Examples of comparing the Is-A relationship.

Sub Class	CCT Super Class	Level	FO Super Class	Level
Insurance	Financial	2	Financial terms	5
Investment management	Banking theory	5	Financial management	7
Currency	Financial	2	International Finance	3
Financial crisis	Financial market	3	Economic crisis	5
Stock	Securities market	4	Securities	9

Table 3. Examples of comparing the Class-instance relationship

Instance	CCT Class	FO Class
Industrial and Commercial Bank of China (ICBC)	Commercial Bank	Financial institution Financial company Bank of China
Venture Capital	Investment	Finance Financial theory
Gold reserve	Currency management	Finance

Contrasted the matched results, we found that Financial Ontology’s super class level is deeper, in addition, each of over 50% instances in Financial Ontology is related to more than two classes, it means that more detailed category and class-instance relationships are described in Financial Ontology.

To further evaluate our Financial Ontology, we also compared it with WordNet, results show that Financial Ontology defines more detailed class hierarchy specialized in a specific field, and it prefers to describe the specific concepts while as for abstract concept, it does not do well.

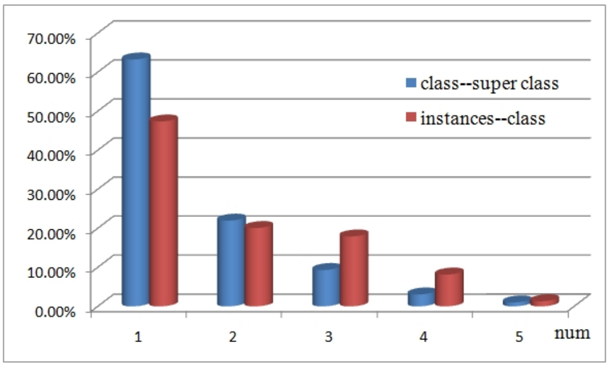

Figure 1: the overall analysis of FO
(Note: it presents the distribution situation that a class/instance relates to 1, 2, or more super class/class in FO.)

Conclusion

This paper proposed a method to automatically construct financial Ontology using Chinese Encyclopedia resources. It has been proved that the constructed ontology has great advantages in the large scale, creation cost and the richness of semantic information. However, it is inadequacies in other non-hierarchy relationship’s extraction, such as Part-of relationship. Its upper ontology classes mostly depend on the top category of the Encyclopedia resources, which makes it hard to be integrated into other exiting well-made ontology. Next step we will consider further optimize Financial Ontology with better existing language ontology and semantic information resources,

Published by Atlantis Press, Paris, France.

such as HowNet.

References

[1] R Studer, VR Benjamins, D Fensel, Knowledge engineering: Principles and methods, in: Data & Knowledge Engineering (DKE), 1998, pp. 161-197.

[2] Nakayama, K., Hara, T. and Nishio, S, Wikipedia mining for an association web thesaurus construction, in: Proceedings of the 8th international conference on Web information systems engineering, Springer Berlin Heidelberg, 2007, pp. 322-334.

[3] Sören Auer1, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, Zachary Ives, Dbpedia: A nucleus for a web of open data, Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2007, pp.722-735.

[4] Li Lian, Jun Ma, Jingsheng Lei, and Ling Song, Automated Construction Chinese Domain Ontology from Wikipedia. In: Proceedings of the 2008 Fourth International Conference on Natural Computation, IEEE Computer Society, Washington DC, USA, 2008, pp. 670-674.

[5] Fabian M.Suchanek, Gjergji Kasneci, Gerhard Weikum, YAGO: A Core of Semantic Knowledge Unifying WordNet and Wikipedia, in: Proceedings of the 16th international conference on WWW (Banff, Alberta, Canada), New York, 2007, pp.697-706.

[6] Simone Paolo Ponzetto, Michael Strube, Deriving a Large Scale Taxonomy from Wikipedia, in: Proceedings of the 22nd national conference on Artificial intelligence, AAAI Press, 2007, pp. 1440--1447.

[7] Hong bin, Wang, Da xin, Liu, Nian bin, Wang, Tong Wang, Research on sieving algorithm of domain-specific concept from ontology learning (in Chinese), Systems Engineering and Electronics, China , 2010.

[8] Lu Yong, Hou Hanqing, Research on Automatic Acquiring of Chinese Synonyms from Wiki Repository, in: Proceedings of Web Intelligence and Intelligent Agent Technology International Conference (WI-IAT '08, Sydney, NSW), 2008, pp: 287-290.

[9] Wang Lei, Zhou Kuanjiu, Qiu Peng, Automatic Domain Ontology Construction (in Chinese), Journal of the China Society for Scientific and Technical Information, China , 2010.