Application of Aggregated Knowledge Concept in Automatic Knowledge Acquisition from Chinese Web Pages∗

Haiyan Che, Jigui Sun, Xi Bai, Lian Shi
College of Computer Science and Technology, Jilin University, Changchun, 130012, China
Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun. Jilin, 130012 China
chehaiyan001@gmail.com

Abstract

Proper ontology definition is the prerequisite for efficient knowledge acquisition. For the complex knowledge that could not be described by simple binary relation, we advocated a methodology for aggregated knowledge acquisition, describing how to define the ontology for such aggregated knowledge concept and how to acquire knowledge basing on such definition. Experiment shows that this methodology is effective in automatic knowledge acquisition from Chinese free text.

1. Introduction

V. R. Benjamins has pointed that the availability of content is the first challenge for Semantic Web [1]. Since currently, there is little Semantic Web content available, without which, computers can not understand the meanings of information on the Web and thus can not provide users knowledge-based services. Semantic annotation is the solution to this problem, which is adding semantic information to the data to make it machine understandable. For semantic annotation, the most important task is to map the data on the Web to particular ontology.

In this paper, we explore the problem of not only how to map the Web content to domain ontology but also how to compose these content into knowledge triples, which we call it knowledge acquisition (KA for short) in this paper. Automatic KA is very difficult, especially for Chinese, since there’s no such powerful tool as WordNet [2] for Chinese. Besides, not all the Web content can be depicted simply by a binary relation. There’re many cases concerning the relation among more than two individuals, and acquiring knowledge from them is much more complex. This paper advocated an aggregated knowledge concept based KA methodology focusing on extracting knowledge automatically from Chinese Web pages, which are in free text and may contain not only the simple but also the complex knowledge. As far as we know, this is the first time to try to handle such complex knowledge in Chinese free text.

The rest of the paper is organized as follows: in section 2, we introduce some related works. In section 3, we discuss some problems in KA from Chinese Web pages. In section 4, we describe our aggregated knowledge concept based knowledge acquisition methodology. In section 5, we conduct experiment to evaluate the effectiveness of our methodology, and conclude our work in section 6.

2. Related works

Semantic annotation can be seen as the basis for KA and it has received much attention in the research community. Many methods have been proposed, for example, manual annotation [3, 4], rule learning based annotation [5, 6], classification based annotation [7] and sequential labeling based annotation [8]. But all these automatic annotation methods are only applicable for Web content that have certain structure, and not for the free text in natural language.

ALPHA [9] and MUMIS [10] take advantage of the natural language processing technology to try to annotate the free text. But ALPHA only deals with the domain specific sentences that should meet several rigid conditions. At the same time, both of them use a semantic lexicon like WordNet for English or EuroWordNet [11], which links words between many European languages through a common inter-lingua of concepts. For Chinese, HowNet [12] can be used as a general lexical database, but there’re great differences between the specific domain vocabularies and the general language ontology words. So we can not annotate the content depending only on it.

The framework of iOkra [13] is about acquiring knowledge automatically from Chinese documents. It extracts knowledge basing on domain ontology and a linguistic ontology [14], and the resolution of the syntactic ambiguities and the recognition of unknown roles depend on both of them. However, the Chinese words and expressions are so rich and colorful that we could not define everything in a Chinese lexical ontology.

∗ Supported by the Natural Science Foundation of China under Grant No. 60496321; the National Research Foundation for the Doctoral Program of Higher Education of China under Grant No. 20050183065.
Meanwhile, iOka doesn’t deal with the complex relation among more than two individuals.

3. Problems in automatic knowledge acquisition from Chinese Web pages

3.1. Knowledge organization problem

Besides linking two individuals or an individual and a value by a binary relation, we also need to represent the relations among more than two individuals. Take the following paragraph about the annual financial report for example:

“本报讯岳阳兴长(000819)今日公布2006年年度报告,报告称该公司实现销售收入18.49亿,同比增长12.01%;净利润4377万元,同比增长45.49%;每股收益0.265元,同比增长45.6%。(Yueyang Xingchang(000819) released the annual financial report of 2006 today, stating that the company’s sales income was 1849 million, increased 12.01%; net profit was 43.77 million, increased 45.49%; earns per share was 0.265, increased 45.6%.)”

Paragraph 1

This kind of structured information is very common in financial news and we must find a way to deal with it efficiently. After thorough analyzing a quantitative amount of such Web pages, we can make the following conclusions:

- The aggregated knowledge is about inter-related information about more than two individuals;
- The knowledge in the aggregated structure will always appear together, and they should also be returned together when user queries;
- Every piece of knowledge in the aggregated structure is only meaningful when it is considered as part of the whole aggregated structure;
- One deputy should be selected for the whole aggregated structure, for example the annual financial report of Yueyang Xingchang of 2006 in Paragraph 1.

As illustrated in RDF Primer [15] and a W3C Working Draft [16], this involves dealing with an n-ary relation between all the related things, and the n-ary relation must be broken up into a group of separate binary relations. But the method described is not systematic and there’s not any note about how to use this kind of definition in KA. In this paper, we advocated an aggregated knowledge concept based methodology, including how to define the aggregated knowledge concept systematically and how to use them in the automatic KA.

3.2. Triple recognition and composition problem

How to recognize and compose triple if the triple elements don’t appear literally at all? It is a major problem in KA. Take Paragraph 1 for example again, the phrase “2006” should be the value of property: Date of report of the annual financial report entity and “43.77 million” should be the value of property: Data of the net profit entity. But these two properties don’t exist literally in the paragraph at all. How can we recognize these elements and compose the knowledge triples correctly? For KA from English Web pages, for example the project of Artequakt [17], WordNet can be used to give semantic meaning to words, and its three lexical chains (synonyms, hypernyms and hyponyms) are very useful in synonym expansion. But WordNet is also powerless if the word corresponding to the triple elements doesn’t appear. iOka deals with this problem by using a Chinese lexical ontology [14], in which the word “净利润(net profit)” may has a “quantity” relation to the phrase “43.77 million”. However, it is unpractical to define all such things in a linguistic ontology.

To solve this kind of problem, we have advocated a theme-based KA method [18, 19], the main idea is not trying to match every word in the sentence to domain ontology any more. Instead, according to the information already extracted and the theme information defined in the domain ontology, we can infer the description theme of the sentence, which can then provide clues to extract more knowledge. Elementary experiment shows that this method can increase the recall rate and precision rate to a satisfactory degree. The key of this theme-based KA method is the theme definition in the domain ontology. How to choose the accurate themes, how to define the mapping from concepts to themes and from themes to properties deserve further exploration.

In this paper, we advocated an aggregated knowledge concept based knowledge acquisition algorithm, which can infer the “invisible” properties according to the domain ontology and the already recognized properties and instances.

4. Aggregated knowledge concept based knowledge acquisition methodology

Proper domain ontology can provide powerful support to KA. For example, a someValuesFrom restriction can suggest the program to look for the property this restriction concerns. So, our method focuses on define domain ontology with proper restrictions, which characterizes the knowledge precisely and can assist in KA.

4.1. AKC ontology definition method
Definition 4.1. (Aggregated Knowledge Concept) (AKC for short) AKC is such a concept that the only function of which is to organize all the related information together into an aggregated structure. Since the instances of AKC are only generated as intermediate objects to denote the group of knowledge, they are blank nodes in nature. But for the convenience of narration and implementation, we will assign them URIs, which are generated according to some AKC naming rule and are intermediate URIs. So AKC instances can not be identified by their URIs as the named entities do (two AKC instances with different URIs may denote the same group of aggregated knowledge), and their property values are the true meaningful content of the aggregated structure knowledge. In short, the AKC instances are blank nodes in nature but with special URIs.

Considering the possibility that there will also be some aggregated structure inside an AKC, we categorize the AKC further into Outer-AKC and Inner-AKC.

Definition 4.2. (Outer-AKC) The Outer-AKC is the outmost AKC in an aggregated structure and is the deputy of the whole structure, so the instance of which may be referred to from outside of a particular graph and should be identified uniquely by some way. The Outer-AKC instances may be the property values of some named entity but will never be the property values of other AKC instances.

Definition 4.3. (Inner-AKC) The Inner-AKC is the internal AKC in an aggregated structure, the instance of which can only be the property value of some Outer-AKC instance, so it is subordinate to the corresponding Outer-AKC instance and need not be identified.

4.1.1. Ontology definition method. Outer-AKC definition method:
For an Outer-AKC Oakc, \{{P}_{Oakc}\} (i=1..n) is its property set, marked as P(Oakc),
1. Create a new concept Oakc in the proper position in the class hierarchy;
2. Identify the key property set, marked as Key (P(Oakc));
3. Create necessary restrictions for Oakc:
   a) For each property \(p \in \text{Key}(P(Oakc))\), create an exactly cardinality restriction with the value of 1; if p is an object property, add an alternative someValuesFrom restriction is also right;
   b) For each non key property \(p \in P(Oakc)\-\text{Key}(P(Oakc))\)
      a). Create an exactly cardinality restriction with the value of n if every instance of Oakc must have at least n values for p; if n is 1 and p is an object property, add an alternative someValuesFrom restriction is also right;
  b). Create a minCardinality restriction with the value n if every instance of Oakc must have at least n values for p; a someValuesFrom restriction is also available if n is 1 and p is an object property;
  c) Create a maxCardinality restriction with the value of n if every instance of Oakc can have at most n values for the property p;
  c) Specify some other property restriction, such as Functional, Inverse Functional, Symmetric or Transitive.
4. Improve the ontology definition further, for example, we can adjust the class hierarchy and property hierarchy to abstract some parent class or parent property.

Inner-AKC definition method:
For an Inner-AKC Iakc, \{\{Iakc\}_i\} (i=1..n) is its property set, marked as P(Iakc),
1. Create a new concept Iakc in the proper position in the class hierarchy;
2. Create necessary restrictions for Iakc:
   a) Create an exactly cardinality restriction with the value of n if every instance of Iakc must have exact n values for p; if n is 1 and p is an object property, add an alternative someValuesFrom restriction is also right;
   b) Create a minCardinality restriction with the value of n if every instance of Iakc must have at least n values for p; a someValuesFrom restriction is also available if n is 1 and p is an object property;
   c) Create a maxCardinality restriction with the value of n if every instance of Iakc can have at most n values for the property p;
   c) Specify some other property restriction, such as Functional, Inverse Functional, Symmetric or Transitive.
3. Specify some other property restriction, such as Functional, Inverse Functional, Symmetric or Transitive.
4. Improve the ontology definition further, for example, we can adjust the class hierarchy and property hierarchy to abstract some parent class or parent property.

4.1.2. A use case. Take paragraph 1 for example, to represent such a group of aggregated knowledge, an Outer-AKC: Annual financial report entity is defined to delegate the outmost aggregated structure, with the key properties of Report releaser and Date of report. The ontology definition of it is depicted as Figure 1.
Further, the Net profit entity, Total income entity and Earnings per share entity should be treated as Inner-AKCs, since instances of them are property values of some Outer-AKC instance, here is the Annual financial report entity’s instance, and they have no key properties to identify themselves. The ontology definition of Net profit entity is depicted as Figure 2.
4.2. AKC based automatic knowledge acquisition method

According to the ontology definition method described in 4.1, we can recognize “invisible” properties or instances from the already recognized instances, properties and the restrictions of the concepts. For example, if an instance of concept $C$ is recognized and there is a restriction on $C$ saying that every instance of $C$ must have at least one property value for $P$, then $P$ should be added into the current recognized property set if no $P$ or sub property of $P$ in it. Further, if the range type of $P$ is concept $C_p$, but no instance of $C_p$ or sub class of $C_p$ has been recognized, then we should create a new instance of $C_p$ and put it in the current recognized instance set, so that it can be used later to compose knowledge triples. The aggregated knowledge concept based knowledge acquisition algorithm is depicted as the algorithm of AKC_KAA.

**Algorithm 1 AKC_KAA**

**Input:** A sentence $S$ to be processed

1. Perform named entity recognition and word segmentation to get a word list: wordlist;
2. Recognize all the properties: propertyList and instances: instanceList in wordlist by direct word-matching method according to the synonym table and domain ontology;
3. Recognize “invisible” properties and instances according to the restrictions of the recognized instances, the propertyList and the instanceList;
4. Compose candidate triples according to the instanceList, propertyList and the literals already recognized from the former steps;
5. Check the validity of these candidate triples according to the domain ontology definition and get rid of the invalid ones.

(Due to the space limitation, we omitted the concrete details here.)

5. Experiment and discussion

To evaluate the effectiveness of the proposed approach, we applied it to a practical project: CRAB-Second Generation Browser [20]. In CRAB, we aimed at extracting knowledge from the Web pages which are financial news and free text in Chinese.

We collected 208 Chinese news Web pages about the company’s annual financial reports from http://finance.sina.com.cn/. Five domain experts extracted the knowledge manually by using Bugle, a manual semantic annotator developed by our team, and the result set is used as the standard valid knowledge triples. The total workload is 80 person*hour. Then, we performed the knowledge acquisition automatically from the same set of Web pages by the program basing on the algorithm of AKC_KAA and compared the results. As usual, we conducted evaluations in terms of precision rate, recall rate and F1-measure.

Table 1 shows the experimental results on the data set. **TNM** denotes the total number of triples extracted manually by experts, and **TNA** and **VNA** denote the total number and hit number of triples extracted automatically by our program. $P$, $R$, and $F1$ respectively represent the precision rate, recall rate, and F1-measure. Since this is the first time to try to extract knowledge from Chinese web pages which are in free text, there’s no other method we can use as baseline, so we only gave the result of our method.

<table>
<thead>
<tr>
<th></th>
<th>TNM</th>
<th>TNA</th>
<th>VNA</th>
<th>P (%)</th>
<th>R (%)</th>
<th>F1 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>4158</td>
<td>3241</td>
<td>2907</td>
<td>89.69</td>
<td>69.91</td>
<td>78.58</td>
</tr>
<tr>
<td>T2</td>
<td>4158</td>
<td>3656</td>
<td>3285</td>
<td>89.85</td>
<td>79</td>
<td>84.08</td>
</tr>
</tbody>
</table>

After a thorough analysis of the result, we found that:

- If the program could not recognize the instance of the company described in the Web page, there will be no corresponding triples extracted (because there’s no subject for the triple). There’re many such cases in the first experiment (denoted by $T1$). After we have pre-furnished the knowledge base (KB for short) such instances, the performance is better (denoted by $T2$).
- Most of the errors come from the wrong selection of
the values for datatype properties, and the main reason is the inaccuracy and immaturity of the NLP tools. It can only be improved as the NLP tools improve.

- The experimental result is satisfactory and demonstrates the feasibility of our approach.

6. Conclusion and future work

In this paper, we investigated the problem of acquiring knowledge automatically from Chinese Web pages which are free text and may contain some complex knowledge with complex structure, and proposed an aggregated knowledge concept based methodology, including how to define domain ontology which can characterize the factual knowledge correctly, and how to extract knowledge efficiently according to such definition. As far as we know, this is the first time to handle the factual knowledge with complex structure in Chinese free text. The proposed approach has been applied to acquire knowledge from Chinese news about company’s annual financial reports and the experimental results show that our approach is satisfactory.

However, there’re some problems related with this methodology. First, how to merge the factual knowledge newly extracted into KB? We should recognize and discard the redundant and contradictory ones to ensure the KB’s consistency. But the AKC instances’ names can not be used to identify them, so we need to compare the content of them, which are the property values of them, to decide whether the same content exist or whether they are contradictory to the ones already in the KB. Second, how to query the KB? Since this AKC definition method makes it complex to query the knowledge deep in the structure, so we need to explore the query method applicable for such definition.

References


