Logical Structure Based Semantic Relationship Extraction from Semi-Structured Documents

Zhang Kuo
Department of Computer Science,
Tsinghua University,
Beijing, 100084, China
zkuo99@mails.tsinghua.edu.cn

Wu Gang
Department of Computer Science,
Tsinghua University,
Beijing, 100084, China
wug03@mails.tsinghua.edu.cn

Li JuanZi
Department of Computer Science,
Tsinghua University,
Beijing, 100084, China
ljz@keg.cs.tsinghua.edu.cn

ABSTRACT
Addressed in this paper is the issue of semantic relationship extraction from semi-structured documents. Many research efforts have been made so far on the semantic information extraction. However, much of the previous work focuses on detecting 'isolated' semantic information by making use of linguistic analysis or linkage information in web pages and limited research has been done on extracting semantic relationship from the semi-structured documents. In this paper, we propose a method for semantic relationship extraction by using the logical information in the semi-structured document (semi-structured document usually has various types of structure information, e.g. a semi-structured document may be hierarchical laid out). To the best of our knowledge, extracting semantic relationships by using logical information has not been investigated previously. A probabilistic approach has been proposed in the paper. Features used in the probabilistic model have been defined.

Categories and Subject Descriptors
I.2.4 [**Artificial Intelligence**]: Knowledge Representation Formalisms and Methods -- Relation systems, Semantic networks

General Terms
Algorithms, Performance, Languages.

Keywords
Semi-structured document, Logical structure, Relationship extraction, Ontology

1. INTRODUCTION
Currently, there are a few semantic annotation platforms which extract information from web pages and annotate them with ontology. For example, S-CREAM [1] supports the semi-automatic annotation for web pages. KIM [2] provides a novel Knowledge and Information Management infrastructure and services for automatic semantic annotation, indexing, and retrieval of documents. WebKB [3] extracts instances of classes and relations based on web page contents and their linkage path.

However, few of the previous works focus on detecting semantic relationships. Furthermore, the existent semantic annotation systems mainly discover the relationships making use of web linkage or sentence structures. As a result, only relationships between pages or within a sentence can be extracted.

WWW 2006, May 23–26, 2006, Edinburgh, Scotland.
ACM 1-59593-323-9/06/0005.

There also exist a large amount of semi-structured documents other than web pages, such as academic papers, enterprise reports. This kind of documents is different from web pages, because they usually do not contain hyperlinks, and authorized in strict logical structure. Therefore, we need a new algorithm that fits into the features in this kind of documents.

In this paper, we propose an approach that exploits document logical structure to extract relationships. We first extract text pieces as data type property values with iASA [4]. Then we compute the probability that two property values are related by the same instance using logistic regression. And then we find the relationships between the property values that maximize the loss function.

2. Problem Statement
Now we formally define the relationship extraction problem that we are solving.

We first give the definition of knowledge base in our scenario. A knowledge base can be viewed as a three tupel:

$$KB = (I, C, P)$$

where C denotes a set of concepts; P denotes a set of property; I denotes a instance set of all concepts. Specifically, let $c \in C$ denote a concept, $p \in P$ denote a property and $i_c \in I$ denote an instance of concept c, i.e. $i_c \in c$.

We now illustrate the problem of relationship extraction by an example. Say we have a document snippet about hotel information:

1. Hotel description:
   - 1.1 Name: Holiday inn
     - 1.1.1 Address: Beilishi road.
     - 1.1.2 Phone number: 12345678
   - 1.2 Name: Beijing hotel
     - 1.2.1 Address: Chang'an road.
     - 1.2.2 Phone number:87654321

The task is to annotate the snippet by the following ontology:

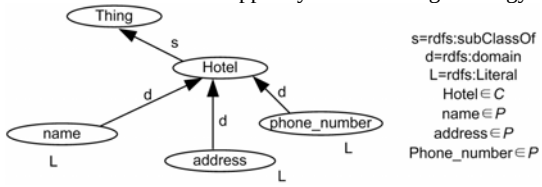

Figure 1: Ontology snippet

By semantic information extraction tool (e.g. iASA), we can obtain:

Name: Holiday inn, Beijing Hotel Address: Beilishi road, Chang'an road Phone number: 12345678, 87654321

Then the task is to associate the information correspondingly (i.e. in this example, we need associate the hotel name, address and phone number). This is exactly the problem semantic relationship extraction addresses. Finally, the output might be:

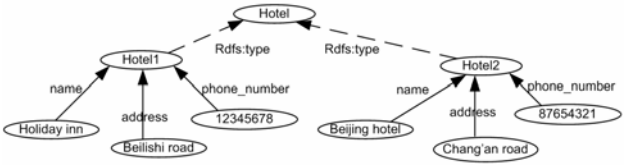

Figure 2: Constructed instances

3. Our Approach

Before explaining our approach in detail, we give two assumptions:

1. Property values for the same instance are usually in the same relative position in logical structure. For instance, hotel name is usually in the parent logical level of hotel address.
2. Two property values usually appear in a document in a constant order. For instance, hotel address usually appears before phone number.

Definition1. For any property values l_1 and l_2, they are called relevant if and only if there exist $i \in c$, $c \in C$, $p_m \in P$, $p_n \in P$ having $p_m(i, l_1)$, $p_n(i, l_2)$. In other words, l_1 and l_2 are property values for the same instance. We use $r_{mn}(l_1, l_2)$ to denote the relevant relation.

Our approach has two main steps. At the first step, we use property values extracted by iASA and their logical structure information as input, and exploit logistic regression to predict the probability of $r_{mn}(l_i, l_j)$ for any property value pair (l_i, l_j). At the second step, we use the relevant probabilities to construct the instances by maximizing a loss function defined in 3.2 section. The output is constructed instances which is similar to figure 2.

3.1 Relevant Probability Estimation

We consider one implementation of our approach. We employ logistical regression [5] in the relation probability estimation. It has not been investigated previously to the best of our knowledge.

The learning based probability estimation consists of two stages: training and prediction.

In training, we train a regression model λ_{mn} for each property pair p_m, p_n that have the same domain concept. Table 1 shows the major features used in the regression model.

Table 1: Features used in regression model

Features	comments
Higher_logic_level	Whether l_1 is in a higher logical level than l_2
Same_logic_level	Whether l_1 is in the same logical level than l_2
Lower_logic_level	Whether l_1 is in a lower logical level than l_2
Appear_before	Whether l_1 appears before l_2
Logical_distance	The distance in the logical structure tree
Same_sentence	Whether l_1 and l_2 are in the same sentence
Same_paragraph	Whether l_1 and l_2 are in the same paragraph

Where, the logical distance is defined:

$$logical\_distance = \frac{2 * level(l_1 \cap l_2)}{level(l_1)+level(l_2)}$$

where $l_1 \cap l_2$ denotes their closest common ancestor, and *level*(*l*) means the length from l to the root node.

3.2 Instance construction

For each concept *c*, we associate property values with instances so as to maximize the loss function:

$$Loss = \sum_{\substack{\{p_m, p_n | domain(p_m)=c \\ ,domain(p_n)=c\}}} \sum_{\substack{\{l_a, l_b | \exists i, \\ p_m(i,l_a) \wedge p_n(i,l_b)\}}} \log(\mathrm{P}(r_{mn}(l_a, l_b))) + \sum_{\substack{\{p_m, p_n | domain(p_m)=c, \\ domain(p_n)=c\}}} \sum_{\substack{\{l_a, l_b | \forall i, \\ \neg p_m(i,l_a) \vee \neg p_n(i,l_b)\}}} \log(1 - \mathrm{P}(r_{mn}(l_a, l_b)))$$

where $p(i, l)$ means the value of property p of instance i is l, and $P(r_{mn}(l_a, l_b))$ represents the probability that l_a and l_b are relevant.

Obviously, it is impossible to enumerate all the instance list candidates $\{i_{c1}, i_{c2}, \ldots, i_{ck}\}$, and select the one which maximize the loss function. So we propose an algorithm to construct the instances:

Step1. for each text value l of property p, construct instance i, s.t. $p(i,l)$.

Step2. for each text value l of property p_n, find instance i* that maximize:

$$loss(l) = \sum_{p_m} \sum_{\{l_a | p_m(i,l_a)\}} (\log(\mathrm{P}(r_{nm}(l, l_a))) - \log(1 - \mathrm{P}(r_{nm}(l, l_a))))$$

then attach l to i*, i.e., set l as the value of property p_n of i*, and detach l from the original instance.

Step3. compute $Loss^{(k)}$. If $Loss^{(k)} - Loss^{(k-1)} < \varepsilon$, then stop, otherwise repeat step2.

The complex of step2 is $O(n_l^2)$, where n_l is the number of property values. A property value l may be reattached to instances more than one time, because the attachment changing of other property values may affect l.

4. Conclusion

In this paper, we investigated the problem of semantic relationship extraction from semi-structured documents. We give a definition of relationship extraction problem. We proposed an approach for the problem by using logistic regression.

REFERENCES

[1] S. Handschuh, S. Staab, and F. Ciravegna. S-CREAM -- Semi-automatic CREAtion of Metadata. In Proceedings of EKAW 2002.

[2] P. Borislav, K. Atanas, K. Angel, M. Dimitar, O. Damyan, G. Miroslav: KIM - Semantic Annotation Platform. International Semantic Web Conference 2003: 834-849.

[3] C. Mark, D. Dan, F. Dayne, M. Andrew, M. Tom, N. Kamal and S. Seán. Learning to Construct Knowledge Bases from the World Wide Web, Artificial Intelligence, 118(1-2): 69-113.2000.

[4] J. Tang, JZ. Li, HJ. Lu, BY. Liang, XT. Huang, KH. Wang. iASA: Learning to Annotate the Semantic Web. Journal on Data Semantics (4): 110-145.2005

[5] D.H. Freeman. Applied Categorial Data Analysis. Dekker, New York, 1987