AN ONTOLOGY-BASED AND COOPERATIVE ANNOTATION SYSTEM

Wenjuan Wu, Xiaoyong Du, He Hu, Ning Ma
Information School, Renmin University of China, Beijing, P.R. China 100872

wendywood@ruc.edu.cn, duyong@ruc.edu.cn, hehu@ruc.edu.cn

Abstract: The Semantic Web is charming but not easy to realize, since one of the prerequisite steps is to create semantic and precise data adhere to traditional web pages. We provide an annotation system ConAnnotator, which is an ontology-based annotation system and allows cooperative working. It aims at supporting the annotation process as well as the evolvement of the ontology. By semi-automatically creating ontology-based annotations and managing the statistic information about the annotation history, it facilitates the annotation process, makes the annotated documents connected to the ontology and further constructs the Semantic Web, and ultimately helps the users to evolve the ontology itself.

Key words: ontology-based annotation, domain ontology, collaborative annotation, ontology evolvement.

1. INTRODUCTION

The semantic web [1] is more prominent for it contains information that is not only readable for human but also can be understood by the computer. Therefore, to make the semantic web come true, the first and the most important step is to add semantic and precise data to traditional web pages. The data is added is called “annotation”. However, this is an arduous, time consuming and error-prone task [9]. In this paper, we mainly introduce our

ConAnnotator - an Ontology-based annotation system bundling role-based cooperation function.

In this paper, we first introduce related work in brief. We describe the structure of Cooperative Ontology Developing Environment and Repository System (CODERS) [2] in which our ConAnnotator plays an important role respectively in section 3. In section 4, we describe the framework of ConAnnotator[3] in detail. Finally, we discuss the future work and draw a conclusion.

2. RELATED WORK

There are several tools used to create semantic annotation, such as SHOE, Annotea[13], Ontomat, SMORE and so on. [4][5] SHOE [10] was one the earliest systems for adding semantic annotations to web pages. SHOE Knowledge Annotator allows users to mark up pages in SHOE guided by ontologies available locally or via a URL. These marked up pages can be reasoned about by SHOE-aware tools such as SHOE Search. Such tools are described in [11, 12]. Annotea is a W3C tool (and protocol) that enhances collaboration via shared metadata but it does not support information extraction nor is it linked to an ontology server; Ontomat is quite meaningful for the future HTML editors; SMORE is a tool that allows users to markup their documents in RDF using web ontologies in association with user-specific terms and elements.

ConAnnotator improves the annotation efficiency by introducing a fully automotive method. The other mentioned systems provide useful tools in annotation processes; but they all lack automatic features in their implementations and hinder the large scale deployment. Further more, it applies itself to Chinese resource.

3. COOPERATIVE ONTOLOGY DEVELOPING ENVIRONMENT AND REPOSITORY SYSTEM (CODERS)

We have been applying ourselves to building a demo Economics Semantic Web, which is a subject –oriented semantic web described in [2].

Cooperative ontology developing environment and Repository System (CODERS) is based on role-based collaborative development method (RCDM) [2]. Our ConAnnotator plays an important role in it.

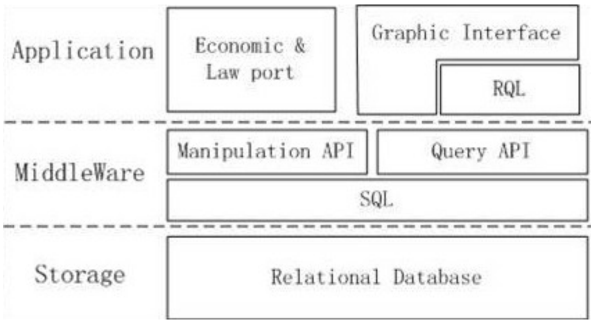


Figure 1. Hierarchy of CODERS

4. FRAMEWORK OF CONANNOTATOR

Figure 2 depicts the architecture of ConAnnotator. The Google Web API is used to crawl resource from WWW, the resource, maybe web page or other formatted files, are moved to the Crawled Repository, then the domain filters act on them, resource focus on the domain is saved in the domain repository. Next step comes the ConAnnotator, it will automatically annotate the resources using the domain ontology after tokenizing and doing Part-Of-Speech (POS) tagging on resources.

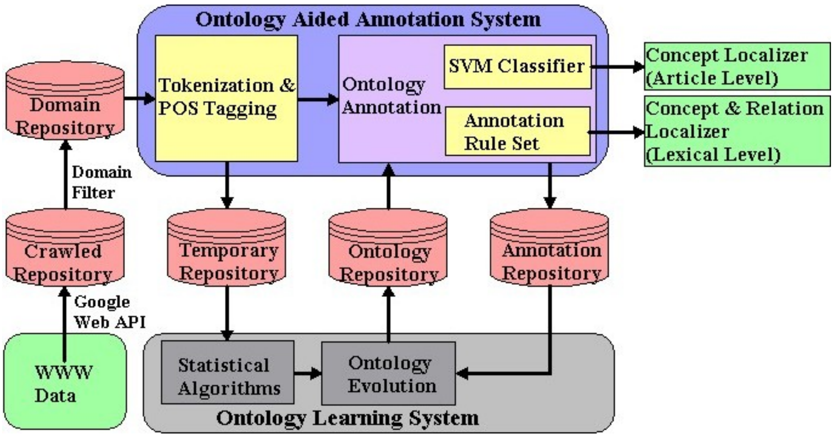


Figure 2. Framework of ConAnnotator

4.1 Semi-automatic Annotation

The semi-automatic process contains 2 levels:

1. Classification at Article Level
   Article level annotations describe the semantic linking between a particular resource and the domain ontology concepts. We use SVM (Support Vector Machine) as the classifier. SVM has been proven to be both precise and effective in solving text classify problem [7]. In the user interface, we use different color to show the confidence for the classification result, by this means we facilitate the annotate process.
2. Information Extraction at Lexical Level
   The IE function offers two ways to support semi-automatic annotate:
   a) Extract “°basi”¡± information about the document like title author, abstract, keyword, and class number. We take advantage of Regular Expressions to describe the characters of the “basic” information.
   b) Extract keyword candidates.
      We use a free Chinese lexical analysis system ICTCLAS developed by Institute of Computing Technology, Chinese Academy of Sciences. Details are in [3]. We provide an algorithm to find keyword candidates based on word frequency.

4.2 Evolution of Ontology

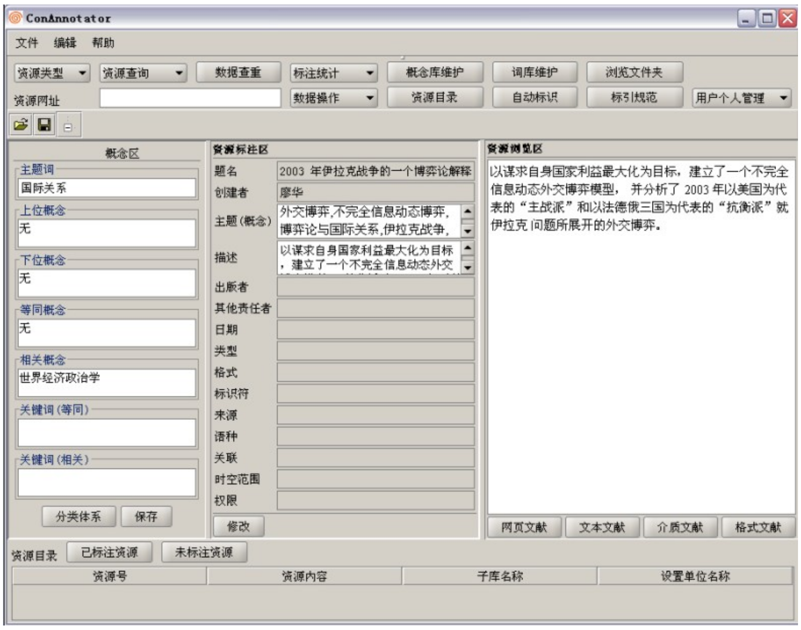
We maintain a Post-controlled word repository which stores the Post-controlled vocabulary and the statistic information about the “concept” term of annotations, and an Ontology Repository to store the ontologies what underpin the system. The structure of the Post-controlled vocabulary is as same as that of ontology.

Repositories maintenance module helps in building the mapping between the Post-controlled repository and the ontology repository, and showing the statistic data of one given keyword. Users can decide which concept in the ontology the keyword should relate to, according to their understanding, experience, or type of the document, and so on. The statistic information is used to help annotators to decide if they should add a keyword to the ontology. In this way, we support the evolution of ontology.

4.3 User Interface of ConAnnotator

The user interface of ConAnnotator is divided into 5 parts: Function Bar, Concept Browser (including the Keyword Browser), Resource Annotation

Editor, Resource Browser, and Resource List. The interface is showed in Figure 3.



主题词：国际关系

上位概念：无

下位概念：无

等同概念：无

相关概念：世界经济政治学

题名：2003 年伊拉克战争的一个博弈论解释

创建者：廖华

主题（概念）：外交博弈,不完全信息动态博弈,博弈论与国际关系,伊拉克战争,

描述：以谋求自身国家利益最大化为目标，建立了一个不完全信息动态外交

以谋求自身国家利益最大化为目标，建立了一个不完全信息动态外交博弈模型， 并分析了 2003 年以美国为代表的“主战派”和以法德俄三国为代表的“抗衡派”就伊拉克 问题所展开的外交博弈。

Figure 3. User Interface of ConAnnotator

5. FUTURE WORK

In the future, firstly we will develop a content annotation module. This module will help us create annotation at the content level and save them, thereby provide more information about not only the whole document but also its content, and then these documents can support applications more freely and adequately. Secondly, we will improve our IE algorithm and try to make it more intelligent.

6. CONCLUSIONS

ConAnnotator is an annotation system for facilitating the annotation process. It helps users, Digital Library Team members at present, to create annotation easily and efficiently. The annotated documents are connected to the domain ontology, thus they can support applications based on Semantic Web. It plays an important role in the CODERS, and has been proven to be practical. We will improve ConAnnotator both in function and efficiency in future, to make it more universal and perfect.

REFERENCES

1. T. B. Lee, J. Hendler and L. Ora, *The Semantic Web*, The Scientific American, May 200.
2. M. Li, D. Wang, X. Du, S. Wang, “° Ontdogy Constructio for Semantic Web: A Role-based Collaborative Development Method”±*Proceedings of the 7th Asia Pacific Web Conference (APWeb 2005), Lecture Notes in Computer Science series 3399,* Shanghai, China, 2005, pp.609-619.
3. He Hu and Xiaoyong Du, ConAnnotator: Ontology-aided Collaborative Annotation System, Proc. Of the 10th International Conference on CSCW in Design (CSCWD 06), IEEE Press， Nanjing, China, 2006
4. M. Vargas-Vera, E. Motta, J. Domingue, etc., MnM: Ontology Driven Semi-Automatic and Automatic Support for Semantic Mark-up. The 13th International Conference on Knowledge Engineering and Management (EKAW 2002), 2002, pp379-391
5. S. Mukherjee, G.Z. Yang, and I. V. Ramakrishnan. Automatic Annotation of Content-Rich HTML Documents: Structural and Semantic Analysis. In *Proceedings of the Second International Semantic Web Con-ference (ISWC 2003),* Sanibel Island, Florida, October, 2003. pp. 533-549.
6. K. Crammer, Y. Singer, “° Onthe algorithmic implementatio of multi-class kernel-based vector machines”±*Machine Learning Research*, 2:265-292, 2001.
7. J. Frank, M. Radermacher, P. Penczek, etc,. SPIDER and WEB: Processing and Visualization of Images in 3D Electron Microscopy and Related Fields. J. *Structural Biol.*, 116, 190-199 (1996)
8. M. Erdmann, A. Maedche, H. Schnurr, and S. Staab. From manual to semi-automatic semantic annotation: About ontology-based text annotation tools. In P. Buitelaar and K. Hasida, editors, Proceedings of the COLING 2000 Workshop on Semantic Annotation and Intelligent Content, August 2000.
9. J. Heflin and J. Hendler. Searching the web with shoe. In AAAI-2000 Workshop on AI for Web Search, 2000.
10. T. Leonard and H. Glaser. Large scale acquisition and maintenance from the web without source access. http://semannot2001.aifb.unikarlsruhe.de/positionpapers/Leonard.pdf, 2001.
11. M. Vargas-Vera, E. Motta, J. Domingue, M. Lanzoni, A. Stutt, and F. Ciravegna. MnM: Ontology driven semi-automatic and automatic support for semantic markup. In The 13th International Conference on Knowledge Engineering and Management (EKAW 2002), 2002.
12. J. Kahan, M. R. Koivunen, E. P. Hommeaux, and R. Swickd. Annotea: An Open RDF Infrastructure for Shared Web Annotations. In *Proceedings of the Tenth International World Wide Web Conference*, Hong Kong, China, May, 2001. pp. 623-632.