Chapter 3.12
Automatic Semantic Annotation Using Machine Learning

Jie Tang
Tsinghua University, China

Duo Zhang
University of Illinois, USA

Limin Yao
Tsinghua University, China

Yi Li
Tsinghua University, China

ABSTRACT

This chapter aims to give a thorough investigation of the techniques for automatic semantic annotation. The Semantic Web provides a common framework that allows data to be shared and reused across applications, enterprises, and community boundaries. However, lack of annotated semantic data is a bottleneck to make the Semantic Web vision a reality. Therefore, it is indeed necessary to automate the process of semantic annotation. In the past few years, there was a rapid expansion of activities in the semantic annotation area. Many methods have been proposed for automating the annotation process. However, due to the heterogeneity and the lack of structure of the Web data, automated discovery of the targeted or unexpected knowledge information still present many challenging research problems. In this chapter, we study the problems of semantic annotation and introduce the state-of-the-art methods for dealing with the problems. We will also give a brief survey of the developed systems based on the methods. Several real-world applications of semantic annotation will be introduced as well. Finally, some emerging challenges in semantic annotation will be discussed.

DOI: 10.4018/978-1-60960-818-7.ch3.12

INTRODUCTION

Semantic annotation of the web documents is the only way to make the Semantic Web vision a reality. The current Semantic Web meets a bottleneck

that there is not much of a Semantic Web due to the lack of annotated web pages. There is such a lack that the Semantic Web is still submerged in the sea of the un-meaningful (un-annotated) web pages.

Semantic annotations are to tag ontology class instance data and map it onto ontology classes. Manual annotation is more easily accomplished today, using authoring tools such as OntoMat (Handschuh, Staab, and Ciravegna, 2002) and SHOE (Heflin, Hendler, and Luke, 2003). However, the use of human annotators is often fraught with errors due to factors such as annotator familiarity with the domain, amount of training, and complex schemas. Manual annotation is also expensive and cannot be used to deal with the large volume of the existing documents on the Web. Automatic semantic annotation is an ideal solution to the problem. However, the fully automatic creation of semantic annotations is also an unsolved problem. Hence, semi-automatic creation of annotations is the method mostly used in current systems.

There are many automatic annotation methods have been proposed, including: A. supervised machine learning based method, B. unsupervised machine learning based method, and C. ontology based method.

A. The supervised machine learning based method consists of two stages: annotation and training. In annotation, we are given a document in either plain text or semi-structured (e.g. emails, web pages, forums, etc.), and the objective is to identify the entities and the semantic relations between the entities. In training, the task is to learn the model(s) that are used in the annotation stage. For learning the models, the input data is often viewed as a sequence of units, for example, a document can be viewed as a sequence of either words or text lines (depending on the specific applications). In the supervised machine learning based method, labeled data for training the model is required.
B. The unsupervised machine learning based method tries to create the annotation without labeled data. For example, Crescenzi, Mecca, and Merialdo (2001) propose a method for automatically generalizing the extraction patterns from the web pages. The generalized patterns can then be used to extract the data from the Web.
C. The ontology based method employs the other knowledge sources like thesaurus, ontology, etc. The basic idea is to first construct a pattern-based ontology, and then use the ontology to extract the needed information from the web page. Some systems also utilize the human general knowledge from common sense ontologies such as Cyc (Lenat and Guha, 1990) and WordNet (Fellbaum, 1998).

In this chapter, we will focus on the first topic: how to create semantic annotation by using supervised machine learning. Figure 1 shows our perspective on semantic annotation. It consists of three layers: Theoretical layer, Annotation layer, and Advanced application layer. The bottom layer is the basic theories including machine learning, statistical learning, and natural language processing; based on these theories, the annotation layer (the middle layer) is mainly comprised of four subtasks: entity extraction, relation extraction, relation discovery, and annotation; based on the annotated results (i.e. semantic data), different advanced applications can be developed (the top layer), for example: semantic integration, semantic search, semantic mining, and reasoning. In semantic annotation, by entity extraction, we aim at identifying and pulling out a sub-sequence that we are interested in from a web page. The identified sub-sequence is viewed as an instance (Appelt, 1999; MUC, 1999). By relation extraction, given a pair of entities, the objective is to decide whether a particular relation holds between the entities (ACE, 2003; Culotta and Sorensen, 2004). By relation discovery, we aim at discovering unknown