COMPARISON OF THE SEMANTIC ANNOTATION SYSTEMS FOR THE TEXT-BASED WEB DOCUMENTS

Shabnam Azari*

Tooraj Samimi Behbahan*

Abstract: *In recent years, internet has become one of the most important sources of information. Due to the amount of deviation of available information, searching for web content via keywords is inefficient. To some extent this is because unconstructed HTML web pages has been created for human understanding and cannot be processed directly by machine. The aim of semantic web in line with automatizing of duties and processes is improving the structural condition of web from the readable level for machine to understandable level for it. For achieving this prospect, some metadata should be added to existing data in web. These metadata include an explanation about content or function of sources. One of the main guidelines for linking such metadata is annotation.*

*Department of Computer Engineering, Behbahan Branch, Islamic Azad University, Behbahan, Iran

1- INTRODUCTION

Semantic web promises of some functions such as concept searching, custom web page generation and question answering systems. Partial semantic annotation is a key for actualizing semantic web. Available content and existing documents on web cause difficulty for manual annotation. Semi-automatic semantic annotation systems have been called platform due to extensibility and comparability of services. These systems have been designed for reduction of workload of text-based web documents. Semantic annotation platforms offer services for supporting annotation such as ontology, access and storage of knowledge base, information extraction, programming interfaces and final user interfaces [1].

2- COMPARISON OF SEMANTIC ANNOTATION PLATFORMS

This section presents an overview of semantic annotation platforms according to platform properties. Then according what was mentioned, a framework for presentation of differences between platforms has been defined. An outline of some representative platforms will be offered and each platform will be briefly analyzed by using framework of platform description.

2-1- Development of semantic annotation platforms

The semi-automatic annotation systems which were compared in this section present semantic annotation of text-based web documents. Such systems are largely called platform for their extensibility and computability. In addition, in some research they will be referred as platform.

2-2- Platform architecture

Figure 1 shows a general architecture of semantic annotation platforms (SAP) as a constructible system [1]. Often, SAPs are extensible. This means that their different components can be replaced by other implementations. The advantage of extendible annotation platform is that it can be adjusted with many requirements like changing of domain and language or creating scaling.

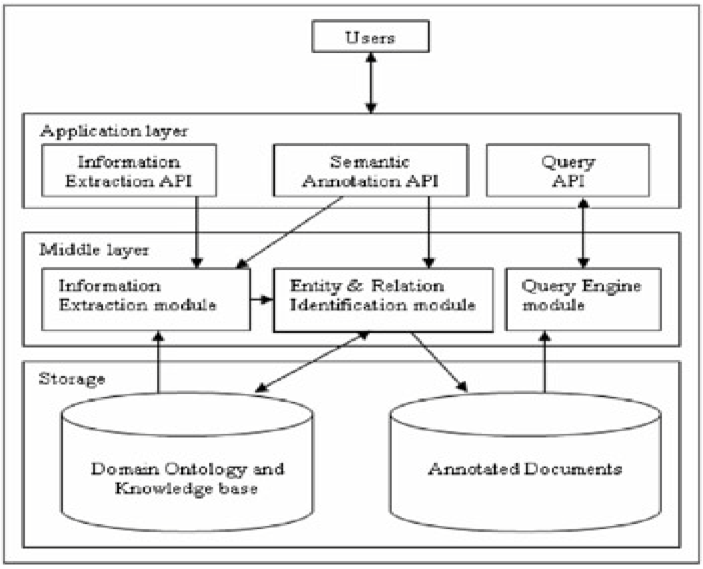
Users
Application layer
Information Extraction API
Semantic Annotation API
Query API
Middle layer
Information Extraction module
Entity & Relation Identification module
Query Engine module
Storage
Domain Ontology and Knowledge base
Annotated Documents

Figure 1: general architecture of a semantic annotation platform

Application layer is responsible for creating a final interface user for prepared services by SAP. The samples include facilities for annotating a document or collection of document and finally acknowledging annotations before their stabilizations. This architecture creates an investigating interface for finding annotations and an interface user for forming information extraction section. Application layer is a layer for interfaces of primary application program. A collection of general programming interface has been designed, and covers the application of middle layer defined in this layer. Due to actions of platform of an application, they are called defined API which can be multiple. Middle layer includes an original part which can perform an action for an application like information extraction for concept identification (names and relations). But middle layer has been created and/or adopted for an existing tool. Storage layer has been used for making storage and management of storage facilities for long-term saving of data such as ontologies, annotations of documents and knowledge base.

2-3- Platform categorization

Recent annotation platforms use various methods of information extraction from web documents. Figure 2 shows a hierarchy categorization of annotation platforms on the basis of their IE component [1]. This categorization schema can be used for organizing complier platforms of semantic annotation.

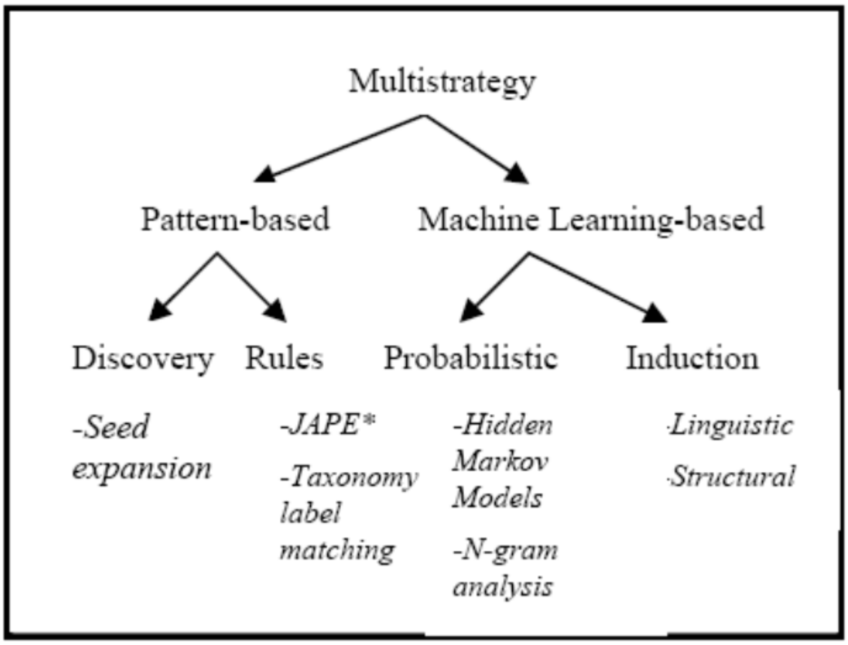

Figure 2: A categorization of semantic annotation platforms on the basis of used information extraction method

During past years, many tools and systems have been designed for semantic annotation. These tools and systems, called as annotation platforms, are categorized on the basis of used annotation method in them. For this, platforms are divided into two major groups: pattern-based and machine learning-based. These two groups illustrated in Figure 2 [1].

Platforms can use existing methods in both groups. This is for strengthening and compensation of reduction of existing methods in each group which is called multi-strategy. Pattern-based methods can perform discovery of pattern and also use patterns which have been defined manually. Machine learning-based techniques use probabilistic and induction approaches. Platforms with probabilistic approach use statistical models for predicting the existing place within the text.

2-4- Pattern-based methods

Patterns have been extensively used in semantic annotation platforms. Activities of pattern discovery find some pattern-based existence by reception of several examples. Receiving examples are extended by some patterns from new found existence. This process will be repeated till no more sample found or user stops repeated process. In order to find existences within the text, recognized language pattern can be used, like Hearst patterns [1].

3- AN OUTLINE OF SEMANTIC ANNOTATION PLATFORMS

In this section, some of semantic annotation platforms have been surveyed [1]:

- ✓ AeroDAML: [2] is designed for mapping proper nouns and common relations, DARPA (DAML) agent marking for categories and corresponding properties in language ontology.
- ✓ Armadillo [3] has been used for searching home pages of computer teachers in order to find information of private calls, like name, place, home page and email.
- ✓ Kim [4] is a place for managing knowledge and information. This tool contains ontology, knowledge base, semantic annotation, indexing server and recovery besides final software for interface server.
- ✓ MnM [5] offers a platform for manual annotation of didactic writing. They are given to an induction cover system on the basis of Amyl care [6]. Once platform is taught and rules are inducted from didactic writings.
- ✓ MUSE [7] applies an applicable rule-based approach for annotating. Text properties are used for conditional performing of various processing sources like different cultures on a document.
- ✓ SemTag [8] includes Seeker semantic annotation as a general platform for web pages' annotation at large scale. SemTag is used as specific tool of semantic annotation independent of domain. This tool annotates 264 million web pages and produces 434 million semantic annotations which have been automatically cleared.

4- CONCLUSION

In order to achieve semantic web, semantic annotations should be used extensively. The advantage of adding meaning to web includes query process by concept searching, custom web page generation in impaired vision, and using information with different concepts, development of needs and user viewpoint and answering query. Manual annotation is difficult for some reasons. Manual annotation is not scalable for document volume on web and suffered from some matters like motivation and knowledge of annotator domain.

REFERENCES

[1] Taniar D., WennyRahayu J.,”Web Semantics and Ontology”,IGI Global 2006.

[2] Kogut P., Holmes W.,"AeroDAML: "Applying Information Extraction to Generate DAML Annotations from Web Pages.", First International Conference on Knowledge Capture (K-CAP 2001). Workshop on Knowledge Markup and Semantic Annotation, Victoria, B.C., Canada, October 2001.

[3] Ciravegna F., Chapman S., Dingli A., Wilks Y., "Learning to harvest information for the semantic web. In Proceedings of the First European Semantic Web Symposium", May 2004. Crete.

[4] Popov B., Kirayakov A., Ognyanoff D., Manov D., Kirilov A., " KIM—a semantic platform of information extraction and retrieval", Nat. Lang. Eng. 10 (3/4) (2004) 375–392.

[5] Vargas-Vera M., Motta E., Domingue J., Lanzoni M., Stutt A., Ciravegna F., "MnM: Ontol`ogy driven semi-automatic and automatic support for semantic markup". In The 13th International Conference on Knowledge Engineering and Management (EKAW 2002), 2002.

[6] Ciravegna F. A., Iria J., Wilks Y.," Multi-strategy definition of Annotation Services in Militia", Department of Computer Science, University of Sheffield, Regent Court, 211, Portobello Street, Sheffield S1 4DP,2002.

[7] Maynard D., Tablan V., Bontcheva K., Cunningham H., Wilks Y.," MUSE: a MUlti-Source Entity recognition system", Department of Computer Science, University of Sheffield, 25 July 2003, 2003 Kluwer Academic Publishers, Printed in the Netherlands.

[8] Dill S., Eiron N., "A Case for Automated Large-Scale Semantic Annotation", Journal of Web Semantics, 2003.