The Semantic Annotated Documents - From HTML to the Semantic Web

Jason C. Hung
Department of Information Management
Technology and Science Institute of Northern Taiwan
No. 2, Xueyuan Rd., Peitou, 112 Taipei,
Taiwan, R.O.C
http://member.mine.tku.edu.tw/www/jhung

Abstract: -The current circumstance of the Semantic Web is that there is not much of a Semantic Web due to the lack of annotated web pages. There is such a lack because annotating web pages currently does not provide much practical benefit. In this work an automated approach to semantics extraction and annotation on textual data is proposed. Word sense disambiguation technique is used to identify the concepts, and RDF is used to annotate the semantics. A corresponding approach to retrieve data via ontology is also discussed. Finally a framework to integrate and automate these processes is demonstrated. In this fashion all the existing data on the Web can be processed and brought to the Semantic Web.

Key-Words: - Semantic Web, RDF, word sense disambiguation, semantic extraction, semantic annotation, ontology, Wordnet

1 Introduction

When looking for interested information through the traditional approaches such as category browsing or search engine, the users have to make lots of effort manually filtering out noise and prospecting for the interested pieces from the massive information feedback.

The major drawback of the nowadays most commonly used Web language, the HTML, is its inability to encode the semantic information essential to a large range of Web applications. This is the main reason why new languages such as XML and RDF [1] have been developed. Some corresponding schemas, as well as extensions or applications of these languages include DAML [2], OIL [3], OWL [4], SHOE [5] and many others.

Today's Web was initially designed for direct human processing. With its current structure, machine-based Web applications are not possible unless its content is transformed into a machine-readable format with the semantic information encoded. And this is the major intention of the Semantic Web [6]: to create a new type of Web content which is meaningful to machines. Semantic Web contains not just one single kind of relation (the hyperlink) between resources, but many various kinds of relations between the various types of resources.

The realization of the Semantic Web requires the widespread availability of semantic annotations for existing and new documents on the Web. Semantic annotations are to tag ontology class instance data and map it into ontology classes. The fully automatic creation of semantic annotations is still an unsolved problem. An approach that can effectively leverage semantically tagged data is definitely needed.

This work aims at facilitating the migration from today's Web to the future's Semantic Web. It tries to automate the process of generating and annotating semantic information for the existing general-domain textual data. The semantic retrieval to these data is demonstrated also. The proposed framework can be viewed at three levels: data format level, semantics level and ontology level.

The major concerns of this work are:
1. Extract semantic information from the data;
2. Annotate data with the semantics;
3. Search for data based on the semantics;
4. Make inference from the ontology (to help the searching).

2. Extract Semantic Information from Textual Data

Up to August 2005 there are over 8 billion HTML documents on the Internet; however there is relatively a very small quantity of semantic annotated documents. One reason is that Semantic Web related standards and techniques are still under construction. However defining and annotating the semantic information to the data is still tough to most of the ordinary information providers (or editors). Especially for the existing innumerable web documents, how to make them "semantic" is a critical concern. In this section, how to automatically extract (or identify) the semantic information from textual data according to its context is investigated. This makes it possible to represent these data by concepts (semantics) rather than keywords.

Because meaningful sentences are composed of meaningful words, any computer system that hopes to process natural languages as human do must have information about words and their meanings. This information is traditionally provided through dictionaries, and digitalized dictionaries are now widely available. But most dictionaries are designed and constructed for the convenience of human readers, not for machines. Fortunately, there are a few machine-readable dictionaries emerged and developed continuously. One of the most widely known is the Wordnet [7], developed by George A. Miller et al. at Princeton University.

Here, the idea is to construct the "mappings" between words in the content and their corresponding word senses defined in Wordnet. In Wordnet, synonymous words are grouped together into "synonym sets", or called "synsets". Each such synset represents a single distinct word sense or concept. For example, in Wordnet the synset consists of the word forms {car, auto, automobile, machine, motorcar} and the concept of "4-wheeled motor vehicle; usually propelled by an internal combustion engine". Synsets are in turn linked through semantic relations that determine word definitions and senses.

There are two situations when constructing the mapping:

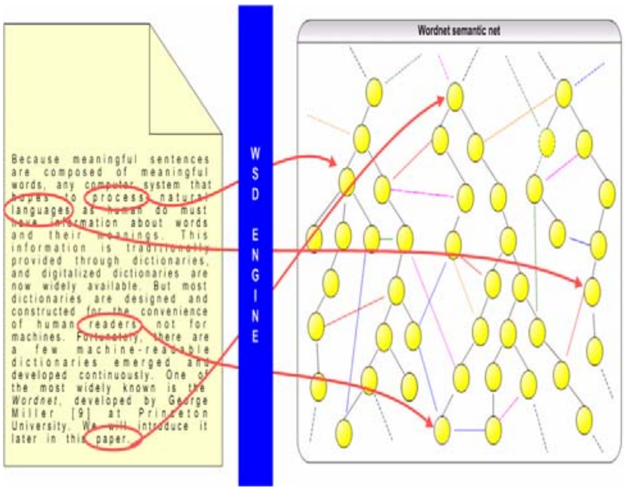
(1). A word has only one sense.

(2). A word has multiple senses (polysemy).

To the former, the mapping between a word and a sense can be made without additional effort. However to the latter, a proper sense has to be determined from multiple senses. All human languages have words that can mean different things in different contexts, such words with multiple meanings are potentially "ambiguous". For almost all applications of language technology, word sense ambiguity is a potential source of error.

Human beings are sophisticated at judging the semantics of content or meanings of words. For example, given the sentence "The bank holds the mortgage on my home", people immediately know that the bank here refers to a financial institution that accepts deposits and channels the money into lending activities. Whereas given the sentence "He sat on the bank of the river and watched the currents", the bank here means the sloping land beside a body of water. However it is very difficult for machines to do the same job effortlessly - all these words are just alphabetical strings to them. Especially tough situation is when there are issues of "polysems" and 'synonyms". Polysemy - a single word form having more than one meaning; synonymy - multiple words having the same meaning, are both important issues in natural language processing or artificial intelligence related fields.

The WSD task involves labeling a word with a tag from a pre-specified set of tag possibilities by using features of the context and other information. It is to form the "mappings" between words and word senses, as shown in figure 1.



Because meaningful sentences are composed of meaningful words, any computer system that hopes to process natural languages as human do must have information about words and their meanings. This information is traditionally provided through dictionaries, and digitalized dictionaries are now widely available. But most dictionaries are designed and constructed for the convenience of human readers, not for machines. Fortunately, there are a few machine-readable dictionaries emerged and developed continuously. One of the most widely known is the Wordnet, developed by George Miller [9] at Princeton University. We will introduce it later in this paper.

Figure 1. mappings between word forms and senses.

The research on WSD has been one of the popular issues in computational linguistics for a long while. Roughly speaking, recent advances benefit from machine learning techniques, sophisticated sense inventories (especially WordNet [7]), and large corpora to find relevant linguistic features.

3. Annotate Textual Data with Semantics

The semantic information can be extracted from text, the "interests" of the content is known. Now following issue is: how to annotate text with these interests in a form understandable or processable to the machine.

The Resource Description Framework (RDF) [1] is a language for representing the information about resources in the World Wide Web. RDF is recommended by W3C and is designed to represent information in a minimally constraining, flexible way. It can be used in isolated applications, where individually designed formats might be more direct and easily understood, but RDF's generality offers greater value from sharing. The value of information thus increases as it becomes accessible to more applications across the entire Internet. To facilitate operation at Internet scale, RDF is an open-world framework that allows anyone to make statements about any resource.

In RDF, the underlying structure of any expression is a collection of triples, each consisting of a "subject", a "predicate" and an "object". The assertion of an RDF triple says that some relationship, indicated by the predicate, holds between the things denoted by subject and object of the triple. A set of such triples is called an RDF graph. A statement can be illustrated by a node and directed-arc diagram, in which each triple is represented as a node-arc-node link, as shown in figure 2.

Figure 2. RDF graph data model to make statements about any resource.

A node may be a Uniform Resource Identifier (URI) with optional fragment identifier (URI reference, or URIref), a literal, or blank. A URI is a compact sequence of characters that identifies an abstract or physical resource. Properties are URI references. A URI reference or literal used as a node identifies what that node represents. A URI reference used as a predicate identifies a relationship between the things represented by the nodes it connects. A predicate URI reference may also be a node in the graph.

The RDF graph data model can be used to annotate data with semantics. Here the annotation can be on a "document basis". A textual document can correspond to a "subject", and a word sense in Wordnet can correspond to an "object". As to the "predicate", which indicates the relationship between "a textual document" and "a word sense", should be qualified to reflect the idea of "has the concept in its context" or "refers to the concept in its context". The RDF graph is as figure 3.

Figure 3. Using RDF graph to annotate a document with semantic information

Recently, Wordnet has been utilized in Semantic Web research community for use in annotation, reasoning, and as background knowledge in ontology mapping tools. Currently there exist several conversions of WordNet to RDF(S) or OWL representations. One standard conversion of Princeton WordNet to RDF/OWL was proposed by Mark van Assem et al. [21]. The aim is to provide a feasible fashion for Wordnet to be used in Semantic Web applications. Up to now, it is an editor's draft, considered for publication as First Public Working Draft by the Semantic Web Best Practices and Deployment Working Group, part of the W3C Semantic Web Activity.

In this RDF/OWL schema of WordNet ontology, there are three main classes: Synset, WordSense and Word. The former two have subclasses for the lexical groups present in WordNet, e.g. NounSynset and VerbWordSense. Each instance of Synset, WordSense and Word has its own URI. There is a pattern for the URIs so that (i) it is easier to determine from the URI the class to which the instance belongs; and (ii) the URI provides some information on the meaning of the entity it represents.

4. Search for Documents Based on the Semantics

The modern search engines operate based on the traditional information retrieval techniques. They directly and merely deal with the strings in the text, and do not take care of the semantic annotations in the documents. Therefore these techniques are far from sufficient for dealing with the Semantic Web documents which consist of rich semantic annotations such as RDF triples or RDF statements. As a result, to the nowadays search engines, the Semantic Web does not exist. This is a barrier to move toward Semantic Web from today’s Web, yet also an opportunity to make great advances.

One possible approach to search data utilizing the semantic annotations is through the ontology used to describe the data. First, the concepts of the information needs should be identified in the ontology. Then the searching is like to find the instances of these concepts (classes) through the semantic annotations. The ways to identify intended concepts in ontology are: searching or browsing through the ontology with proper interface, or mapping from other ontologies.

As the semantic information can be annotated to the documents using the RDF graph data model, some corresponding method to query for the information in the RDF graphs in needed. SPARQL [21] is a query language and a protocol for accessing RDF designed by the W3C RDF Data Access Working Group. As a query language, SPARQL queries the information held in the RDF graphs.

The SPARQL query language is based on matching graph patterns. The simplest graph pattern is the triple pattern, which is like an RDF triple, but with the possibility of a variable instead of an RDF term in the subject, predicate or object positions. Combining triple gives a basic graph pattern, where an exact match to a graph is needed to fulfill a pattern.

5. Make Inference from Ontology

The “inference”, here, does not refer to the assembly of pieces, nor the assessment of coherence. Instead, here the “inference” stands for “prospecting for relative items from and for a given item.” This process can be iterative: to find even more items from previous results. For instance, the inference may take a concept as input, use certain kind of rules or relations to acquire some relative concepts to the given one.

The Wordnet ontology is a kind of semantic net that consists of nodes (synsets) that represent unique concepts and are connected to each others through some proper semantic relations. These nodes and semantic relations can be treated as the inference rules for exploring concepts from one to others. Also, the previously inferred concepts can be used as the inputs to iterate the inference. This inference can be used to facilitate the retrieval or management.

Particular related information needs can be fulfilled through particular related concepts, and particular related concepts can be found through particular semantic relations. After obtained the data on “chair”, if a user (people or software) may want the data about further instances of “chair”, then the hyponym relation can be adopted for inference. The related concepts such as “folding chair”, “chaise lounge” and ”wheelchair” are then achieved. These particular concepts can then be use to query for the data contains them. Or if the user wants to find the data talking about things that have similar idea to “chair”, then the relations “coordinate” can be used.

In practical, for human user, this inference can be visualized as a nevigatable map. This can be done by visualizing the Wordnet ontology as a directed graph. In this way, the inference is just like “traveling” from one node to another. Since the data has been associated with proper nodes, the data thus can be obtained by user easily. For example, to the concept ”http://wordnet.princeton.edu/wn/chair-noun-1” (chair as a concept of “seat”), the local area of the Wordnet semantic net centered from the concept is shown in figure 4.

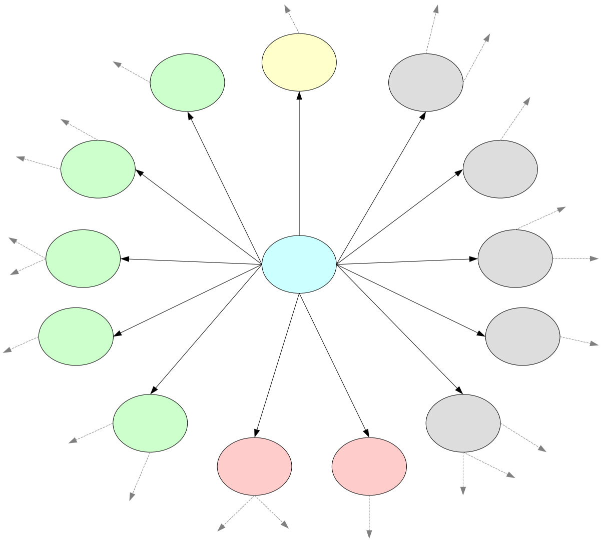


Figure 4. Local area of the Wordnet semantic net centered from the concept “chair-noun-1”

Wordnet defines rich set of semantic relations. Some more common ones are listed in the table 1.

Table 1. Some common semantic relations that can be used for inference

Relation	Description

coordinate	Coordinate terms are nouns or verbs that have the same hypernym
entailment	A verb X entails Y if X cannot be done unless Y is, or has been, done
holonym	The name of the whole of which the meronym names a part. Y is a holonym of X if X is a part of Y.
hypernym	The generic term used to designate a whole class of specific instances. Y is a hypernym of X if X is a (kind of) Y.
hyponym	The specific term used to designate a member of a class. X is a hyponym of Y if X is a (kind of) Y .
meronym	The name of a constituent part of, the substance of, or a member of something. X is a meronym of Y if X is a part of Y.
troponym	A verb expressing a specific manner elaboration of another verb. X is a troponym of Y if to X is to Y in some manner.
Antonym	f(X) is the opposite of X

6. A Semantic Annotation and Retrieval Framework

This section demonstrates a possible approach to application, and discusses some critical concerns in the implementation.

For the proposed framework of semantic information extracting, annotation, searching and inference, three resource and techniques are required:

- Word Sense Disambiguation
- Wordnet ontology
- Resource Definition Framework (RDF)

6.1 The Workflow

The flow of adding semantics to textual documents can be as follows:
1. Acquisition of textual document(s);
2. Text processing;
3. Semantic processing;
4. Semantic annotation and storage.

After identified concepts refer to in the content of the documents, the syntax and model of RDF can be used for annotation. Through this process not only the semantics of the content is realized, but also the annotations conform to the convention of the Semantic Web.

In the retrieval progress, documents can be retrieved according to the concepts contained in their content. Besides, through the inference from the ontology, related concepts can be achieved so that related documents can be further retrieved.

The flow of retrieving textual documents based on their annotations can be as follows:
1. Identify the concept(s) intended for searching;
2. Query for documents;
3. Infer the related concepts;
4. Output the results.

After identified the concept(s) intended for searching in the Wordnet ontology, the user (people or software) can then query the RDF graphs for certain documents. These RDF graphs consist of rich semantic annotations denoting the relationship between documents and concepts. Further inference of concepts can be made to retrieve related documents.

7. Conclusions

This research demonstrates an automated approach to extract and annotate the semantics of the data using techniques of word sense disambiguation, RDF and the resource of Wordnet ontology. After this the data are associated with the concepts in the ontology. In the retrieval, the query can also be first associated with the concepts in the ontology, then search for the instances data. Related concepts can be inferred from the ontology, the inference results can be use to facilitate the data management such as retrieval or classification. The approach is fully automated and can be apply to very large scale data. This work facilitates the migration from today's web to the future's Semantic Web in that the semantics can be annotated to all the existing textual data on the Web in this way.

The proposed framework can be viewed at three levels: data format, semantics and ontology. In this work, the data format level is given the example of textual data. However, other kinds of data, such as image, video, audio, can be dealt in a similar approach. In semantics level, the extraction of semantic information of these data formats has their own specific techniques. Meanwhile, the RDF can still be used to annotate the semantics.

How to share the semantic annotations against/between different ontologies will be a future research, e.g. through SUMO (Suggested Upper Merged Ontology) to search on Wordnet-annotated documents. This cross-ontology data sharing will concern the interoperation or mapping between various ontologies.

References:

[1] Resource Description Framework (RDF), http://www.w3.org/RDF/.

[2] The DARPA Agent Markup Language Homepage, http://www.daml.org/.

[3] Dieter Fensel, Frank van Harmelen, Ian Horrocks, Deborah L. McGuinness, and Peter F. Patel-Schneider. OIL: An ontology infrastructure for the semantic web. IEEE Intelligent Systems, 16(2):38–45, 2001.

[4] Michael K. Smith, Chris Welty, and Deborah L. McGuinness. OWL Web Ontology Language Guide. W3C Recommendation, 10 February, 2004. Available at http://www.w3.org/TR/owl-guide/.

[5] Jeff Heflin, Jim Hendler, Sean Luke. SHOE: A Knowledge Representation Language for Internet Applications. Technical Report CS-TR-4078 (UMIACS TR-99-71), Dept. of Computer Science, University of Maryland at College Park. 1999.

[6] Tim Berners-Lee, James Hendler and Ora Lassila. The Semantic Web. May 2001.

[7] Christiane Fellbaum, editor. 1998. WordNet: An Electronic Lexical Database. MIT Press, Cambridge, Massachusetts.

[8] J. Kahan and M.R. Koivunen. Annotea: an open RDF infrastructure for shared Web annotations. In World Wide Web, pages 623–632, 2001.

[9] Aditya Kalyanpur, James Hendler, Bijan Parsia, Jenni-fer Golbeck, SMORE - Semantic Markup, Ontology, and RDF Editor, available at http://www.mindswap.org/papers/SMORE.pdf

[10] L. Denoue and L. Vignollet. An annotation tool for Web browsers and its applications to information retrieval. In Proceedings of RIAO2000, Paris, April 2000.

[11] D. Maynard, V. Tablan, H. Cunningham, C. Ursu, H. Saggion, K. Bontcheva, and Y. Wilks. Architectural elements of language engineering robustness. Journal of Natural Language Engineering, Special Issue on Robust Methods in Analysis of Natural Language Data, 2002.

[12] M. Tallis, N. Goldman, and R. Balzer. The briefing associate: A role for cots applications in the semantic web. In Semantic Web Working Symposium (SWWS), Stanford, California, USA, August 2001.

[13] M. Tallis. Semantic Word Processing for Content Authors. In Proceedings of the Knowledge Markup & Semantic Annotation Workshop, Florida, USA, 2003. Part of the Second International Conference on Knowledge Capture, K-CAP 2003.

[14] N. F. Noy, M. Sintek, S. Decker, M. Crubezy, R. W. Fergerson, and M. A. Musen. Creating Semantic Web contents with protege-2000. IEEE Intelligent Systems, 2(16):60–71, 2001.

[15] S. Staab, A. Maedche, and S. Handschuh. An annotation framework for the Semantic Web. In S. Isjizaki, editor, Proceedings of the First Workshop on Multimedia Annotation, Tokyo, Japan, January 2001.

[16] T. Leonard and H. Glaser. Large scale acquisition and maintenance from the Web without source access. Available at http://semannot2001.aifb.unikarlsruhe.de/positionpapers/Leonard.pdf, 2001.

[17] 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.

[18] T. Leonard and H. Glaser. Large scale acquisition and maintenance from the Web without source access. Available at http://semannot2001.aifb.unikarlsruhe.de/positionpapers/Leonard.pdf, 2001.

[19] A. Broder and M. R. Henzinger. Algorithmic aspects of information retrieval on the Web. In M. G. C. R. J. Abello, P. M. Pardalos, editor, Handbook of Massive Data Sets. Kluwer Academic Publishers, Boston, to appear.

[20] K. Lerman, C. Knoblock, and S. Minton. Automatic data extraction from lists and tables in web sources. In IJCAI-2001 Workshop on Adaptive Text Extraction and Mining, August 2001.

[21] Mark van Assem, Aldo Gangemi, Guus Schreiber. RDF/OWL Representation of WordNet, Editor's Draft 2 February 2006. Available at http://www.w3.org/2001/sw/BestPractices/WNET/wn-conversion.html.