A CONCEPT-BASED RETRIEVAL TOOL: THE COOPERATIVE WEB

Daniel Gayo Avello, Darío Álvarez Gutiérrez, Juan Manuel Cueva Lovelle
Department of Informatics, University of Oviedo
Calvo Sotelo s/n 33007 Oviedo (SPAIN)
{dani, darioa, cueva}@lsi.uniovi.es

ABSTRACT

The Web is a colossal document repository that is nowadays processed by humans only. Machines' role is limited to transmission and layout processing, barely being able to do something else with contents of documents. Therefore, information retrieval in the current Web is a difficult task where many results provide little relevance. The Semantic Web tries to solve this and other problems by means of new technologies to build ontologies and to annotate semantically the documents. Many of the ontological initiatives try to achieve automatic ontologies construction and semantic annotation. However, such tasks require close human supervision. In addition to that, although the Semantic Web can be very useful to retrieve information from semistructured repositories from e-business, digital libraries and corporate intranets, it is difficult to turn it into a shell over the Web as a whole because the Web size and heterogeneity hinder the development of ontologies to satisfy any conceivable query. We propose a different, although complementary, approach to the Semantic Web called the Cooperative Web. With this approach it would be possible to extract semantics from the Web providing better information retrieval mechanisms without the need of ontological artifacts.

KEYWORDS

Semantic Web, Cooperative Web, Web semantics, collaborative filtering, content-based recommendation, user profiling.

1. INTRODUCTION

The Web is a colossal document repository that is nowadays processed by humans only. Machines' role is just to transmit and display contents. It is indeed very little what a computer can do autonomously with the Web contents. The problem arises because only the text of the documents can be processed, and not the semantics, as the language in which the documents are authored does not allow to attach meaning to the contents.

The Semantic Web (Berners-Lee, 1998) is a proposal from Tim Berners-Lee that tries to partially solve these problems. A number of techniques were proposed in the beginnings of the Semantic Web to solve the lack of semantic markup. Some suggested to use HTML/XML tags (van Harmelen and van der Meer, 1999), whereas others used extensions of HTML (Luke and Heflin, 2000). All these projects had two things in common. The first common point was the need for ontologies to provide a conceptual framework for the semantic markup to have meaning. The second was the possible use of inference systems to obtain new knowledge.

The Semantic Web has maintained this evolution by defining an architecture that offers solutions to many problems of the Web. However, other semantic problems are out of the scope of its approach, but can be solved by using the approached proposed in this paper.

2. SEMANTIC WEB AND WEB SEMANTICS

The Semantic Web tries to facilitate semantic information processing in the Web to machines. To achieve this, technologies to define ontologies and to express concepts with these ontologies are being developed,

thus providing software agents with the ability to "understand" those concepts and to infer new information from them.

Technologies such as RDF/RDFS, OIL or DAML+OIL do allow to explicitly express a semantic for Web documents that was lacked previously. Nevertheless, that kind of Semantic Web, although useful and necessary in fields like e-business, digital libraries or knowledge management in corporate intranets does not cover all the Web semantics issues.

Briefly, a Semantic Web application requires an ontology that describes the fundamental concepts of a particular field in order to semantically markup the documents. Obviously, ontologies can be generated semi-automatically (Maedche and Staab, 2000), as well as the documents semantic markup (Erdmann *et al*, 2001). However, there are situations in which this is very difficult to apply. It may be the case that building the ontology is not easy or possible (Maedche and Staab, 2000), or there is no economic interest, or documents can not be tagged because they do not belong to the ontology author, etc. Current Web, because of its size and heterogeneity, makes the global implementation of a Semantic Web shell not possible. However, we think that it is possible to make a different approach.

3. COOPERATIVE WEB

As a complement to the Semantic Web we propose what we call the Cooperative Web, supported by three basic points: concepts, document taxonomies and cooperation between users (actually between agents acting on behalf of the users).

3.1 Concepts vs. Keywords and Ontologies

Keyword-based information retrieval used by current search engines has the problems of relatively low precision and excessive recall. Ontologies could improve precision in some cases but developing ontologies to support any conceivable query on the Web would be insurmountably hard. However, there is a middle point: use of concepts.

A concept would have more semantics than a keyword but it would not require complex artifacts such as ontological layers. A concept can be seen as a cluster of words with related meaning in a given scope, ignoring tense, gender, and number. For instance, in a knowledge field related to cinema or films could exist the concept `(actor,actress,artist,celebrity,star)`.

Concepts would be useful if they add semantics in an analogous way as ontologies, whereas they should be able to be automatically generated and processed as keywords. There are enough techniques able to be adapted to carry out this task, such as Latent Semantic Indexing (Foltz, 1990) or Concept Indexing (Karypis and Han, 2000).

3.2 Document Taxonomies

Semantic Web relies heavily on ontologies. The Cooperative Web, instead, would use the whole text of the document, with no markup, as the source for semantic meaning. How could this be done without the need to "understand" the text?

A document can be seen as an individual from a population. Among living beings an individual is defined by its genome, composed of chromosomes, divided into genes constructed upon genetic bases. Alike, documents are composed of passages, divided into sentences built upon *concepts*. It is evident that two documents are semantically related if their "genome" are alike. Big differences mean that the semantic relationship between documents is low.

We think that this analogy can be put into practice by adapting some computational biology algorithms. In a gross way, these kind of algorithms work with long character strings representing DNA fragments. Similar individuals or species show similitudes in their genetic codes so it is possible to classify them into taxonomies without the need to know what every gene "does". Documents could be classified into taxonomic trees depending on the similitudes found in their "conceptual genome". The important thing about such a

classification is that it would provide semantics (conceptual similitudes) without requiring the classification process to use any ones.

3.3 Collaboration between Users

The current Web has also another problem. Each time a user browses the Web, she establishes a path that could be useful for others. Besides, many others could have followed that path before. However, that experimental knowledge is lost. The Cooperative Web intends to extract useful semantics from user experiences.

Each user in the Cooperative Web would have an agent (see figure 1) with two main goals: to learn from its master, and to retrieve information for her. Reaching the first goal involves the task of developing a user profile that describes her interests in terms of concepts, it would be constructed upon the documents the user stores in her computer, visits frequently, are in her browser's bookmarks, etc.

Another aspect not considered by the current Web and the Semantic Web is the "utility" of a document. Documents are searched and processed by humans depending on the usefulness they expect to get from them. Utility does not reside in the contents but it is a subjective judgement that a particular user assigns to a specific document.

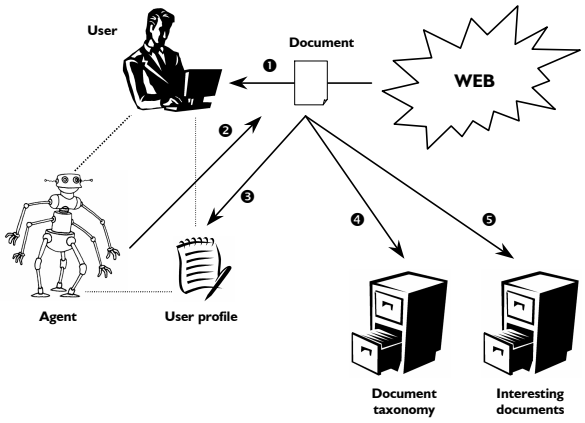

Figure 1. Basic operation of the Cooperative Web

The Cooperative Web, having each user attached to a profile, could assign to each par `(profile,documentPassage)` a utility level. Each user agent would be responsible for deciding that utility level. In order for this to be really practical, the utility level should be determined in an implicit way just by observing users' behavior.

Most projects related to users' resource rating require voluntary participation, e.g., AntWorld (Meñkov *et al*, 2000) or Fab (Balabanovic and Shoham, 1997). However, there are interesting experiences in the field of implicit rating. (Morita and Shinoda, 1994) describe an experimental study that treated the problem of providing interesting USENET posts to a group of users depending on their preferences. The technique used to implicitly determine users' rating was based on reading times, actions made upon the environment, and actions made upon the posts text. GroupLens (Konstan *et al*, 1997) is a similar approach.

We think that implicit rating is more adequate for a practical implementation. Moreover, a thorough research of the psychological attention and learning mechanisms along the browsing process would contribute interesting results to the field of implicit rating.

Regarding the retrieving of information for the master, the agent would have two different ways to do it: to find information satisfying a query, or to explore on behalf of the user to recommend unknown documents. A hybrid of two reputed techniques could be applied for both cases: Collaborative Filtering (CF) (Goldberg *et al*, 1992) and Case/Content-Based Recommendation (CBR).

In a nutshell, CF provides a user with what other individuals alike have found useful (one example is the Amazon service "Customers who bought this book also bought:"). CBR, on the other hand, provides elements similar to a start element as a recommendation. In our case, if the agent used CF, documents with a high utility level for the user profile would be recommended, without regard to the conceptual relationship between the document and the profile. Using CBR, documents similar to the description of the user profile, to a query or to a start document would be returned, without regard to the utility level of these documents.

Hybrid techniques facilitate the finding of new elements and the operation of a user community (profile members) when they have not rated many documents yet (Burke, 1999).

4. CONCLUSION AND FUTURE WORK

We have briefly described the concept of the Semantic Web, pointing some aspects that hinder its application to the Web as a whole.

We propose a different approach, Cooperative Web, based on the automatic extraction of concepts from free text to establish a document taxonomy in an automatic way. Besides, our approach integrates users as another system element classifying them into different profiles in order to extract valuable information about documents' utility. Cooperative Web would allow better retrieval and recommendation mechanisms than current search engines and would enlarge application scope of Semantic Web.

We are making a deeper study about the Cooperative Web that is the subject for a Ph.D. thesis. In order to get a full operative prototype the following subsystems would be developed: text filtering, conceptual distilling, taxonomic classification, user profiling, implicit rating, retrieval and recommendation.

REFERENCES

Balabanovic, M., and Shoham, Y. (1997), "Fab: Content-Based, Collaborative Recommendation", Communications of the ACM, Vol. 40, No. 3, pp. 66-72.

Berners-Lee, T. (1998), "Semantic Web Road map", http://www.w3.org/DesignIssues/Semantic.html

Burke, R. (1999), "Integrating Knowledge-based and Collaborative-filtering Recommender Systems", Proceedings of the AAAI Workshop on AI and Electronic Commerce, Orlando, Florida, EE.UU., pp. 69-72.

Erdmann, M., Maedche, A., Scnurr, H.P., and Staab, S. (2001), "From Manual to Semi-automatic Semantic Annotation: About Ontology-based Text Annotation Tools", ETAI Journal – Section on Semantic Web, 6.

Foltz, P.W. (1990), "Using Latent Semantic Indexing for Information Filtering", Proceedings of the ACM Conference on Office Information Systems, Boston, USA, pp. 40-47.

Goldberg, D., Nichols, D., Oki, B.M., and Terry, D. (1992), "Using Collaborative Filtering to Weave an Information Tapestry", Communications of the ACM, Vol. 35, No. 12, pp. 61-70.

Karypis, G., and Han, E. (2000), "Concept indexing: A fast dimensionality reduction algorithm with applications to document retrieval and categorization", Technical Report TR-00-0016, University of Minnesota.

Konstan, J.A., Miller, B.N., Maltz, D., Herlocker, J.L., Gordon, L.R., and Riedl, J. (1997), "GroupLens: Applying Collaborative Filtering to Usenet News", Communications of the ACM, Vol. 40, No. 3, pp. 77-87.

Luke, S., and Heflin, J. (2000), "SHOE 1.01. Proposed Specification", http://www.cs.umd.edu/projects/plus/SHOE/spec1.01.html

Maedche, A., and Staab, S. (2000), "Discovering Conceptual Relations from Text. Technical Report 399", Institute AIFB, Karlsruhe University.

Meñkov, V., Neu, D.J., and Shi, Q. (2000), "AntWorld: A Collaborative Web Search Tool", Distributed Communities on the Web, Third International Workshop, pp. 13-22

Morita, M., and Shinoda, Y. (1994), "Information filtering based on user behavior analysis and best match text retrieval", Proceedings of the 17th Annual International Retrieval, Dublin, Ireland.

van Harmelen, F., and van der Meer, J. (1999), "WebMaster: Knowledge-based Verification of Web-pages", Practical Applications of Knowledge Management, PAKeM'99