The Cooperative Web: A Complement to the Semantic Web

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

Abstract

The Web is a colossal document repository that is nowadays processed by humans only. The machines' role is just to transmit and display the contents, barely being able to do something else. The Semantic Web tries to change this status so that software agents can manipulate the semantic contents of the Web. There are some technologies proposed for this task that facilitate the definition of ontologies and the semantic markup of documents based on that ontologies. However, although the Semantic Web can be very useful in fields such as e-business, digital libraries or knowledge management inside corporate intranets, it is difficult to apply to the global Web. We propose a different, although complementary, approach: The Cooperative Web. With this approach, it would be possible to extract semantics from the Web without the need of ontological artifacts. Besides, the experience of the users would also be leveraged.

1. Introduction

The Web is a colossal document repository that is nowadays processed by humans only. The machines' role is just to transmit and display the contents. It is indeed very little what a computer can do autonomously with the Web contents.

This situation is painfully obvious whenever any user needs to get some information by means of a search engine. Initially, thousands of documents can be returned[1]. Only after successive refinement of the query the result set is manageable, although it is not usually what was looked for.

The problem lies in the way the search engine processes the documents. Only the text of the documents is 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 [1][2] is a proposal from Tim Berners-Lee that tries to partially solve these problems. It is described as "a web of data that can be processed directly or indirectly by machines". It would not be a new Web, but an evolution of the current Web by the use of "tags" that provide semantics instead of layout structure (like HTML tags).

A number of techniques were proposed in the beginnings of the Semantic Web to solve this lack of semantic markup. Some suggested to use HTML/XML tags [3], while others used extensions of HTML [4][5].

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 an inference system (more or less powerful) to obtain new knowledge. The Semantic Web has maintained this evolution by defining an architecture that offers a solution to many of the problems of the Web. However, other semantic problems are out of the scope of this 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.

These technologies do allow to explicitly express a semantic for Web documents that was lacked previously. Nevertheless, that kind of Semantic Web, although useful and necessary, does not cover all the Web semantics issues.

2.1. Technologies for the Semantic Web

There are already some technologies that make possible important parts of the Semantic Web. This section overviews the main ones and how they are related.

RDF [6] is a W3C recommendation that provides support for the description of resources available in the Web, the relationships between them, and an XML syntax

[1] A Google search of the phrase "semantic web" returns 44,600 documents (20th January, 2002).

for its codification and serialization. Metadata described using RDF can be easily processed and exchanged by agents, and therefore a number of semantic services can be created. However, although RDF can use attributes and relationships, no mechanisms are provided to declare them. This task is done by RDF Schema [7] using RDF.

OIL [8] is a product of the On-To-Knowledge[2] project. It is a standard for the definition and exchange of ontologies. It extends RDF Schema and allows the definition of classes, relationships, and the possibility of doing inference as well.

DAML+OIL[3] [9] is a semantic markup language based on OIL and on the previous version of the ontology language DAML-ONT. It is similar to OIL. Both of them can be deemed as RDF Schema extensions.

2.2. There are more Semantics in Web than are Managed by the Semantic Web

The Semantic Web as described before is very useful in fields such as e-business, digital libraries or knowledge management in corporate intranets. Nevertheless, there is more useful semantic information out of the reach of the Semantic Web. Summarizing, a Semantic Web application requires an ontology that describes the fundamental concepts of a particular field in order to semantically markup the documents. Obviously, the ontologies can be generated semi-automatically [10][11], as well as the documents semantic markup [12].

However, there are situations in which this is very difficult to apply. For example, it may be the case that building the ontology is not easy or possible [13] (especially in the case of free text), or that there is no economic interest, or that the documents can not be tagged because they do not belong to the entity that developed the ontology, etc. These cases are very common, as the current Web, because of its size and heterogeneity, makes the global implementation of a Semantic Web shell not possible.

It is possible, and urgent, to apply the Semantic Web in many Web Engineering fields. Anyway, the Web as a whole is not among these fields. However, we think that it is possible to make a different and complementary approach to the Semantic Web that can be applied in fields where it can not do so.

3. The Cooperative Web

As a complement to the Semantic Web we propose what we call the Cooperative Web, supported by three basic points: using concepts instead of keywords and ontologies, the classification of documents based on these concepts into a taxonomy, and the cooperation between users (actually between agents acting on behalf of the users).

3.1. Concepts vs. Keywords

The retrieval of information using keywords and keyphrases used by current search engines has the problems of a relatively low precision and a high recall value[4]. The use of ontologies can improve precision in some cases. However, developing ontologies to support any conceivable query on the Web would be insurmountably hard.

There is a middle point: the use of concepts. A concept would be a more abstract entity (and with more semantics) than a keyword. It would not require complex artifacts such as ontology languages or inference systems. A concept can be seen as a cluster of words with similar meaning in a given scope, ignoring tense, gender, and number. So, in a given knowledge field the concept `(computer, machine, server)` would exist, while in another field `(actor, actress, artist, celebrity, star)` would be a valid concept.

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. Currently there are enough techniques able to be used or adapted to carry out this automatic extraction task, such as Latent Semantic Indexing[5] [14] or others that were already mentioned for the semi-automatic generation of ontologies [10][11]. In the next section we will examine how semantics can be obtained using concepts without resorting to any ontology support.

3.2. Document Taxonomies

To give meaning to a document the Semantic Web needs an ontology defining a number or terms and the relationships between them, in order to then tag parts of

[2] On-To-Knowledge is an European project that has the goal of developing methods and tools that allow to exploit the potential of ontologies in the field of knowledge management. http://www.ontoknowledge.org/

[3] DAML (DARPA Agent Markup Language) is a DARPA program similar in some ways to the On-To-Knowledge project. The main goal of DAML is the developing of languages and tools to facilitate the implementation of the Semantic Web. http://www.daml.org/

[4] Precision and recall concepts defined in [17].

[5] "Latent Semantic Indexing (LSI) is an information retrieval method that organizes information into a semantic structure. It takes advantage of some of the implicit higher-order associations of words with text objects. The resulting structure reflects the major associative patterns in the data while ignoring some of the smaller variations that may be due to idiosyncrasies in the word usage of individual documents. This permits retrieval based on the the "latent" semantic content of the documents rather than just on keyword matches." [14]

the document based on these terms. Instead, the Cooperative Web would use the whole text of the document without using any 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, which is composed of chromosomes, divided into genes constructed upon genetic bases. Alike, documents are composed of passages (groups of sentences related to just one subject), which are divided into sentences built upon concepts. Using this analogy, it is evident that two documents are semantically related if their ”genome” are alike. Big differences between genomes mean that the semantic relationship between documents is low.

We think that this analogy can be put into practice, and that it is possible to adapt some algorithms used in computational biology [15][16] to the field of document classification. In a gross way, these kind of algorithms work with long character strings representing fragments of individuals’ genome from same or different species. Similar individuals or species have similitudes in their genetic codes so it is possible to classify individuals and species into taxonomies without the need to know what every gene “does”.

In the same way, 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 (similitudes at the conceptual level between documents or between documents and user queries) without requiring the classification process to use any semantics.

3.3. Collaboration between Users

The current Web has also another problem at least as serious as its lack of semantics. 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 utilize user experiences, extracting useful semantics from them. Each user in the Cooperative Web would have an agent with two main goals: to learn from its master, and to retrieve information for her.

3.3.1. Learning from the Master

Reaching the first goal, to learn from its master, involves the task of developing a user profile that describes her interests. This description would be done in terms of concepts, and would be constructed upon the documents the user stores in her computer, visits frequently, are in her browser’s bookmarks, etc.

Once the user is attached to a given profile, it is possible to use this information to give a semantic to Web documents that does not depend only on the document, but on the user browsing the document herself. One 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. That utility does not reside in the contents but it is a subjective judgement that a particular user assigns to a specific document.

The Cooperative Web, having each user attached to a profile, could assign to each par `(profile, document)` a utility level. Having an agent for each user it would be responsible for deciding that utility level. In order for this utility valuation to be really practical, the utility level should be determined in an implicit way (just by observing users’ behavior, without querying them). The utility level should also be assigned to individual passages within a document, and not to the document as a whole.

Most of the projects related to users’ resource rating require a voluntary participation of the user, as for example in AntWorld [18] and Fab [19][20]. The main goal of AntWorld was to utilize the users’ experience to facilitate other users the searching task. It used document explicit ratings, making suggestions depending on the query the user was formulating at the moment. Fab, on the other hand, was a web page recommendation system. It did lexical analysis of texts, requesting from users a rating of the suggested recommendations.

However, there are some interesting experiences in the field of implicit rating. Reference [21] describes 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 the user rating was based on reading times, actions made upon the environment, and actions made upon the text of the posts. GroupLens [22] describes a similar system, asserting that using the reading time as the implicit rating system obtains similar recommendations to the ones obtained using explicit rating, thus confirming findings in [21].

We think that the implicit rating approach is more adequate for a practical implementation. A thorough research of the psychological attention and learning mechanisms along the browsing process will probably contribute very interesting results to the field of implicit rating.

3.3.2 Retrieving Information for the Master

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 then unknown documents. A hybrid of two reputed techniques would be very interesting to apply for both cases: Collaborative Filtering [23] and Case/Content-Based Recommendation.

In a nutshell, Collaborative Filtering (CF) provides a user with what other individuals alike have found useful (one example is the Amazon[6] service "Customers who bought this book also bought:").

Case/Content-Based Recommendation (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 (or similar to a query or a start document) would be recommended, without regard to the utility level of these documents.

Using hybrid techniques facilitates the finding of new elements and the operation of a user community (profile members) when they have not rated many documents yet [24]. This hybrid approach has been used in some projects. For example, [25][26] describe how a combination of both techniques is used for a musical recommendation system. The CASPER project (Case-based Agency: Skill Profiling and Electronic Recruitment)[7] researches these techniques in the field of content customization. In the first case, the goal was to recommend songs that users would probably like. The system was able to indicate songs that other users with similar taste found interesting (CF), or to find songs that "sounded" similar to other songs the user had already liked (CBR). CASPER tries to develop an environment that offers searches by content similitude, as well as user profiling to provide customized contents, related in this case to employment offers.

4. Conclusion

We have briefly described the concept of the Semantic Web, pointing some aspects that hinder its application to the Web as a whole. As a complement to the Semantic Web we propose the Cooperative Web, which is based on the automatic extraction of concepts from document text to establish a document taxonomy in an automatic way.

Besides, the Cooperative Web integrates users as another system element. Users are classified into different profiles, and extracting valuable information that links users and documents with a utility relationship.

These metadata would allow the implementation of information retrieval and recommendation mechanisms in the global Web more accurate and effective than current search engines and that can not be provided by the Semantic Web.

5. Future Work

We are making a deeper study about the Cooperating Web that is the subject for a PhD. thesis. The following subsystems would be developed for a full operative prototype:

- *Text filtering:* Natural Language Processing (NLP) systems that eliminate stop words, and text features such as gender, tense, and number. These systems would have to be adaptable to different languages.
- *Conceptual Distilling:* Systems to extract the concepts present in the filtered text. They do not obtain a "concepts bag", but a "conceptual genome" for each document.
- *Taxonomic Classification:* Systems that, based on that "genome", are able to classify it into a document tree with conceptual similitude criteria.
- *User Profiling:* Agents that establish a user profile based on the documents the user "processes", and that classify that profile in a taxonomy of user profiles.
- *Implicit Rating:* Agents that determine the utility level for a document, or for part of a document, and a user profile, based on the actions of the user.
- *Retrieval:* Systems that provide documents that conceptually satisfy the information requests made by the user. They apply the conceptual filtering and distilling systems upon the query and taxonomically classify that query in the document tree.
- *Recommendation:* Agents that explore the document tree and cooperate with other agents from their profile to find items of interest for its master.

6. References

[1] T. Berners-Lee, "Semantic web road map," Internal note, World Wide Web Consortium, http://www.w3.org/DesignIssues/Semantic.html, 1998.

[2] T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web," *Scientific American*, 2001.

[3] F. van Harmelen, and J. van der Meer, "WebMaster: Knowledge-based Verification of Web-pages," *Proceedings of "Practical Applications of Knowledge Management" PAKeM'99*, The Practical Applications Company, London, 1999.

[6] http://www.amazon.com

[7] http://kermit.ucd.ie/casper

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

[5] S. Decker, M. Erdmann, D. Fensel, and R. Studer, "Ontobroker: Ontology based access to distributed and semi-structured information," in R. Meersman et al., editor, *DS-8: Semantic Issues in Multimedia Systems*, Kluwer Academic Publisher, 1999 pp. 351-369.

[6] O. Lassila, and R. Swick, "Resource Description Framework (RDF) Model and Syntax Specification," W3C Recommendation, World Wide Web Consortium http://www.w3.org/TR/REC-rdf-syntax, 1999.

[7] D. Brickley, and R.V. Guha, "Resource Description Framework (RDF) Schema Specification 1.0," W3C Candidate Recommendation, World Wide Web Consortium, http://www.w3.org/TR/rdf-schema, 2000.

[8] I. Horrocks, et al., "The Ontology Inference Layer OIL," Technical report, On-To-Knowledge, http://www.ontoknowledge.org/oil/TR/oil.long.html, 2000.

[9] F. van Harmelen, P.F. Patel-Schneider, and I. Horrocks, "Reference Description of the DAML+OIL (March 2001) Ontology Markuk Language," DAML+OIL Document, http://www.daml.org/2001/03/reference.html, 2001.

[10] P. Clerkin, P. Cunningham, and C. Hayes, "Ontology Discovery for the Semantic Web Using Hierarchical Clustering," Semantic Web Mining Workshop, 2001.

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

[12] M. Erdmann, A. Maedche, H.P. Scnurr, and S. Staab, "From Manual to Semi-automatic Semantic Annotation: About Ontology-based Text Annotation Tools," *ETAI Journal – Section on Semantic Web* (Linköping Electronic Articles in Computer and Information Science), 6, 2001.

[13] C. Kwok, O. Etzioni, and D.S. Weld, "Scaling Question Answering to the Web," In *Proceedings of the Tenth International World Wide Web Conference*, Hong Kong, China, 2001, pp. 150-161.

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

[15] L. Arvestad, "Algorithms for Biological Sequence Alignment," PhD thesis, 1999.

[16] A. Ben-Dor, R. Shamir, and Z. Yakhini, "Clustering Gene Expression Patterns," *Journal of Computational Biology* 6, 1999, pp. 281-297.

[17] G. Salton, *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*, Addison Wesley, 1989.

[18] V. Meñkov, D.J. Neu, and Q. Shi, "AntWorld: A Collaborative Web Search Tool," In *Proceedings of Distributed Communities on the Web*, Third International Workshop, 2000, pp. 13-22.

[19] M. Balabanovic, and Y. Shoham, "Fab: Content-Based, Collaborative Recommendation," *CACM* 40(3), 1997, pp. 66-72.

[20] M. Balabanovic, "An Adaptive Web Page Recommendation Service," In *Proceedings of the First International Conference on Autonomous Agents*, 1997.

[21] M. Morita, and Y. Shinoda, "Information filtering based on user behaviour analysis and best match text retrieval," In *Proceedings of the 17th ACM Annual International Conference on Research and Development in Information Retrieval*, Dublin, Ireland, 1994, pp. 272-281.

[22] J.A. Konstan, B.N. Miller, D. Maltz, J.L. Herlocker, L.R. Gordon, and J. Riedl, "GroupLens: Applying Collaborative Filtering to Usenet News," *CACM* 40(3), 1997, pp. 77-87.

[23] D. Goldberg, D. Nichols, B.M. Oki, and D. Terry, "Using Collaborative Filtering to Weave an Information Tapestry," *CACM* 35(12), 1992, pp. 61-70.

[24] R. Burke, "Integrating Knowledge-based and Collaborative-filtering Recommender Systems," In *Proceedings of the AAAI Workshop on AI and Electronic Commerce*. Orlando, Florida, 1999, pp. 69-72.

[25] I. Goldberg, S.D. Gribble, D. Wagner, and E.A. Brewer, "The Ninja Jukebox," In *Proceedings of USITS' 99: The 2nd USENIX Symposium on Internet Technologies & Systems*. Boulder, Colorado, USA, 1999.

[26] M. Welsh, N. Borisov, J. Hill, R. von Behren, and A. Woo, "Querying Large Collections of Music for Similarity," Technical Report UCB/CSD00-1096, U.C. Berkeley Computer Science Division, 1999.