

The Cooperative Web: a step towards Web Intelligence

Daniel Gayo-Avello, Darío Álvarez-Gutiérrez, Agustín Cernuda-del-Río, José Gayo-Avello, Luis Vinuesa-Martínez, Néstor García-Fernández

Department of Informatics, University of Oviedo, Calvo Sotelo s/n 33007 Oviedo (SPAIN)
{dani, darioa, guti, vinuesa, nestor}@lsi.uniovi.es

Abstract. The Web is mainly processed by humans. The role of the machines is just to transmit and display the contents of the documents, barely being able to do something else. Nowadays there are lots of initiatives trying to change this situation; many of them are related to fields like the Semantic Web or Web Intelligence. This paper describes a new proposal towards Web Intelligence: the Cooperative Web, which would allow us to extract semantics from the Web in an automatic way, without the need of ontological artifacts, with language independence and, besides of this, allowing the usage of browsing experience from individual users to serve the whole community of users.

1 Introduction

Although the Web provides access to a huge amount of information it is not a perfect information retrieval mechanism. Search engines perform a really useful task but we can say that they are toying out since they provide a view of the Web quite poor to get a more powerful use. Besides of this, the current Web shows a problem as serious as its lack of semantics: each time a user browses the Web, he opens a path which could be useful for others and, in the same way, other users can have yet followed such path and have found its worth or its uselessness. However, all that experimental knowledge is lost. Through this paper we will show how we think this situation can be changed in order to provide intelligence and semantics to the Web in an automatic way.

2 The Web as an Information Retrieval System

The main goal of the Web was to avoid the loss of information as well as making the access to it easier. The initial proposal [1] suggested developing the Web starting from a semantic ground showing the drawbacks of keyword-based information retrieval. However, the Web was finally developed in a simpler way, similar to traditional hypertext and with no document retrieval mechanisms.

In 1994 first search engines appeared. They showed that links directories were not enough, albeit they raised two problems: (1) The users had to try their queries with several search engines. (2) Most of the returned documents had little relevance.

The main problem of the pair Web + search engines remains in the use of keyword-based queries. It is known that the probability of two users employing the same keyword to refer a unique concept is below 20% [2]. So, if the query keywords are only looked for in the HTML META tags or in the document title the results are quite poor while if the search is performed using free text from the documents the recall is larger but at the expense of a serious lack of precision [3] that lies, mainly, in the ambiguity of the words, even in well defined domains [4].

3 The Semantic Web

In spite of better search engines, the increasing number of documents hinders the precision of the results and keeps the users under a flood of information. In 1998 Tim Berners-Lee started to outline the Semantic Web. The main idea is to mark up the documents on the Web with “semantic tags” that would provide metainformation about the tagged text. In a sense, this idea is quite similar to the use of “concept nodes” described in the original Web proposal [1] and now again the crux of the matter is the way to provide such semantic tags and state the relationships between them. To perform this task ontologies and ontological languages are used.

Other approaches were proposed before the Semantic Web itself and have contributed greatly to it (e.g. SHOE [5], and Ontobroker [6]). Later, a more elaborated version of the Semantic Web [7] was introduced; in this one, ontologies take a leading role similar to the one played in the proposals mentioned above.

The Semantic Web is not widespread enough as to provide search engines comparable to those from traditional Web. However, some solutions have been proposed (e.g. SquishQL [8] or RQL/Sesame [9]). In spite of differences on syntax or architecture, these “search engines” can be seen as a kind of inference engines that accept queries expressed in terms of one or more ontologies and return as results objects belonging to such ontologies. Thus, the Semantic Web depends heavily on ontologies; because of that, many efforts are being made to provide semi-automatic generation of ontologies [10] and automatic semantic markup of documents [11].

We think that the Semantic Web will make the access to information much easier in well-defined environments such as corporate intranets but it would be really difficult to apply the same techniques to the Web as a whole.

4 The Cooperative Web

Traditional keyword-based approaches to fight against information overload are not suitable to help the user in his information searches on the Web. As for the Semantic Web, it will play a vital role in well-defined domains but it is difficult to apply it to the whole Web in an automatic way. Thus, we propose a complementary solution to contribute towards the Web Intelligence, the so-called Cooperative Web:

“The Cooperative Web is a layer on top of the current Web to give it semantics in an automatic, global, transparent and language independent way. It does not require explicit user participation but implicit feedback that would be acquired by software

agents. The Cooperative Web relies on the use of concepts and document taxonomies, both of them can be obtained with no human supervision from free text.”

Keywords provide poor information retrieval while ontologies can improve precision. However, developing ontologies to support any query on the Web would be really hard. There is a middle point: the use of concepts. A concept would be a more abstract entity 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 related meaning in a given scope, ignoring tense, gender, and number. We think that techniques such as Latent Semantic Indexing [12] or concept indexing [13] could serve to automatically generate and process concepts.

So, 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, divided into sentences built upon concepts.

Using this analogy, it seems clear that two documents are semantically related if their “genome” is alike. Big differences between genomes mean that the semantic relationship between documents is weak. We think that it is possible to adapt some algorithms used in computational biology to the field of document classification. The important thing about such a classification is that it would provide semantics without requiring the classification process to use any. In fact, it should be able to cluster documents in categories similar to the ones that a person would build.

Besides this, the Cooperative Web intends to employ user browsing experience, extracting useful semantics from it. Each user in the Cooperative Web would have an agent that would learn from its master (building a user profile) and retrieve information for him. Having each user attached to a profile, it is possible to assign to each pair (`profile`, `document`) a utility level. In order for this utility valuation to be really practical, the utility level should be determined in an implicit way.

The agent would have two ways to perform information retrieval: to retrieve information for a query formulated by the user, or to explore in the background on his behalf to recommend him unknown documents. To perform both tasks we want to employ two well-known techniques: Collaborative Filtering (CF) and Content-Based Recommendation (CBR). If the agent uses CF it would recommend the user documents that have obtained a high utility level from users with similar profiles. If the agent uses CBR it would retrieve documents that would be conceptually related with the user profile, a query or an initial document, without prioritizing the utility level.

5 Conclusion

We have described a proposal to provide Web Intelligence: the Cooperative Web. We have compared it with the Semantic Web to avoid information overload in the Internet. In more detail we have describe the information retrieval techniques that the Cooperative Web could provide. If they are compared with modern search engines we

think it is clear that our proposal would obtain less but more relevant results since it would employ conceptual taxonomies.

Many researchers involved in the Web Intelligence field share similar views and proposals. Nishida introduces the concept of “virtualized ego” [14], software agents quite similar to the ones proposed for the Cooperative Web. Han and Chang explain the need for automatic building of documents taxonomies [15]. Cercone et al state the relevance of recommender systems and software agents in the Intelligent Web [16].

References

1. Berners-Lee, T.: Information Management: A Proposal
<http://www.w3.org/History/1989/proposal.html> (1989)
2. Furnas, G.W., Landauer, T.K., Gómez, L.M., Dumais, S.T.: The vocabulary problem in human-system communication. *CACM*, Vol. 30, No. 11 (1987) 964-971
3. Pinkerton, B.: Finding what people want: Experiences with the WebCrawler. Proc. of the Second International World Wide Web Conference. Chicago, IL, USA (1994)
4. Krovetz, R., Croft, W.B.: Lexical Ambiguity and Information Retrieval. *ACM Transactions on Information Systems*, Vol. 10, No. 2 (1992) 115-141
5. Luke, S., Spector, L., Rager, D.: Ontology-Based Knowledge Discovery on the World-Wide Web. Working Notes of the Workshop on Internet-Based Information Systems at the 13th National Conference on Artificial Intelligence (AAAI96) (1996)
6. Fensel, D., Decker, S., Erdmann, M., Studer, R.: Ontobroker: Or How to Enable Intelligent Access to the WWW. Proc. of the 11th Workshop on Knowledge Acquisition, Modeling, and Management. Banff, Canada (1998)
7. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. *Scientific American*, 284 (5) (2001) 34-43
8. Brickley, D., Miller, L.: RDF: Extending and Querying RSS channels. ILRT discussion document. <http://ilrt.org/discovery/2000/11/rss-query/> (2000)
9. Karvounarakis, G., Christophides, V., Plexousakis, D., Alexaki, S.: Querying RDF Descriptions for Community Web Portals. The French National Conference on Databases. Agadir, Maroc (2001)
10. Maedche, A., Staab, S.: Discovering Conceptual Relations from Text. Technical Report 399". Institute AIFB, Karlsruhe University (2000)
11. Erdmann, M., Maedche, A., Scnurr, H.P., Staab, S.: 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)
12. Foltz, P.W.: Using Latent Semantic Indexing for Information Filtering. Proc. of the ACM Conference on Office Information Systems. Boston, USA (1990) 40-47
13. Karypis, G., Han, E.: Concept indexing: A fast dimensionality reduction algorithm with applications to document retrieval and categorization. Technical Report TR-00-0016. University of Minnesota (2000)
14. Nishida, T.: Social Intelligence Design for the Web. *IEEE Computer*, IEEE Computer Society, Washington, D.C., 35(11) (2002) 37-41
15. Han, J., Chang, K.C.-C.: Data Mining for Web Intelligence. *IEEE Computer*, IEEE Computer Society, Washington, D.C., 35(11) (2002) 64-70
16. Cercone, N., Hou, L., Keselj, V., An, A., Naruedomkul, K., Hu, X.: From Computational Intelligence to Web Intelligence. *IEEE Computer*, IEEE Computer Society, Washington, D.C., 35(11) (2002) 72-76