Searching the Web: A Semantics-Based Approach

<u>T. H. Cao</u>[1], T. H. D. Nguyen[1], and T. C. T. Qui[1]

Abstract: Search engines are important for finding useful information over the extremely large virtual pool of data over the World Wide Web. However, current engines such as Google are keyword-based only, whereby a query is represented by a set of keywords. Such a query language is not expressive enough to allow a user to represent the subject of the websites that he/she really wants to find. Consequently, as we are experiencing, one often receives many useless results when searching the web.

To overcome that shortcoming, i.e., to increase the percentage of the correct websites out of the hit ones, this paper proposes a semantics-based approach to web search engines. At this first step, we limit the study in the domain of scientific publications only, but the framework could be applied to other domains of web documents as well. The key assumption of our approach is that the subject of the documents that one wants to search for can be represented by a set of concepts and relations between them.

For example, the subject "Application of Internet Computing to High Performance Scientific Computing" has two concepts, namely, "Internet Computing" and "High Performance Scientific Computing", and one relation "Application-To" between them. As such, it is different from the subject "Application of High Performance Scientific Computing to Internet Computing" and, thus, the two queries with these two subjects are expected to have different publications returned as their answers. We note that, for current keyword-based search engines, one cannot express the difference between the two subjects, whose queries comprise the same set of keywords, namely, "Internet Computing", "High Performance Scientific Computing" and "Application".

We study and build up an ontology of primary topics of scientific publications and an ontology of conceptual relations between them, and propose to use conceptual graphs ([1], [2]) to represent both document queries and topics. In order to reduce the computational cost, a keyword-based search engine is first exploited as a filter to return the documents that contain the concepts in a query graph, and then those documents among them whose topic graphs match to the query graph are selected. A demonstration of our proposed semantics-based search engine is also presented.

Main References:

1. Cao, T.H. (2001), *Fuzzy Conceptual Graphs for the Semantic Web*. Invited to the Berkeley Initiative in Soft Computing International Workshop on Fuzzy Logic and the Internet.

2. Martin, P. and Eklund, P. (1999). *Embedding Knowledge in Web Documents*. In Proceedings of the 8th International World Wide Web Conference.

---

[1] Faculty of Information Technology
Ho Chi Minh City University of Technology
268 Ly Thuong Kiet St., Dist. 10, HCM City, Vietnam
tru@dit.hcmut.edu.vn