Chapter 12
An Ontology-Based Search Tool in the Semantic Web

Constanta-Nicoleta Bodea
Academy of Economic Studies, Romania

Adina Lipai
Academy of Economic Studies, Romania

Maria-Iuliana Dascalu
Academy of Economic Studies, Romania

ABSTRACT

The chapter presents a meta-search tool developed in order to deliver search results structured according to the specific interests of users. Meta-search means that for a specific query, several search mechanisms could be simultaneously applied. Using the clustering process, thematically homogenous groups are built up from the initial list provided by the standard search mechanisms. The results are more user-oriented, thanks to the ontological approach of the clustering process. After the initial search made on multiple search engines, the results are pre-processed and transformed into vectors of words. These vectors are mapped into vectors of concepts, by calling an educational ontology and using the WordNet lexical database. The vectors of concepts are refined through concept space graphs and projection mechanisms, before applying the clustering procedure. The chapter describes the proposed solution in the framework of other existent clustering search solutions. Implementation details and early experimentation results are also provided.

INTRODUCTION

The Web users are asking for intelligent services in order to discover and access the content they need. The mechanisms for discovering Web documents are powerful search engines, with specialized discovery services, indexes, and databases.

A simple query could have produce hundreds even thousands of results making it practically impossible for the user to check the relevance of all of them. Even when the list of results is ordered by a rank, most of the time it is not sufficient support for the user to identify the most relevant resources. A first solution for this issue

DOI: 10.4018/978-1-4666-2494-8.ch012

was to sort the results based on a relevance criteria (the more relevant the result is, the higher in the list it is displayed). Even so, the required result is sometimes hard to find because it is not in the first 20 – 50 displayed results. The algorithm for clustering search results presented in this chapter addresses this issue.

Trying to keep up with the continuous growth of World Wide Web (WWW) the searching tools are engaged in a permanent race for ever faster development in order to reach better performances. In the initial stages the general trend of development was concentrated on bigger databases, bigger document bases, in order to store the web pages accordingly.

When the document storage reached considerable sizes, the problem of better indexation was addressed. The bigger the storage capacity becomes, the more efficient the indexing algorithm had to be in order to keep the Web pages properly ordered. However, the WWW was still growing with increasingly speed, so the crawler module had to be developed to reach higher speed in finding and downloading new pages.

For many years, it was believed that the bigger the database of the search engine is, the more performing it will be. The more and more efficient crawler was downloading pages at a higher speed and proper indexer algorithm was constructing the permanently increasing document base. However, when document bases reached billions and tens of billions of documents, and the crawlers were downloading new documents at a speed of hundreds, or even thousand a day, a new problem appeared. With such large quantity of pages, the indexer was retrieving and presenting to users longer and longer lists as result to queries. The simpler and more common the query is, the more results will be returned, rendering the user unable to check all of them in order to identify the Web pages that best fit his needs. Thus, another efficiency criterion was introduced: easy retrieval of the relevant information within the results provided by the search tool. The "easy retrieval" is evaluated both from the speed perspective and from the relevance of the results.

RELATED WORK

Existing Systems for Search Results Clustering

A result clustering of Web searches represents a technique for retrieving information that had relatively little success in comparison with other techniques used in the Web. The first system which implemented clustering was developed by Scatter and Gather. They proposed a system, which searches documents ordered by clustering. The main drawback of the clustering algorithms is that they are slow, not attractive for Web search applications. To combat this disadvantage, Cutting (1992) proposed algorithms with linear processing time, demonstrating their usefulness in retrieving information. After Scatter and Gather, the Grouper system was designed as an interface for clustering the results of a meta-search engine, named HuskySearch (Zamir, 1999). The algorithm used by Grouper used suffix trees. Two years later, Carrot system has emerged as an infrastructure for the development of new clustering algorithms (Weiss, 2001). As an open source platform, Carrot is currently developed by adding new clustering algorithms. Clustering technology applied to the Web search results still remains partially unknown to the general public, despite numerous research and publications have been made in this direction. Next, we present brief characteristics of some of the existing clustering systems.

Carrot

It is a search engine that provides results clustering of a meta search. It is located at: www.carrotsearch.com. Search engines used by Carrot system are: Yahoo, Google, MSN, Open Search, Lucene, and Dmoz. It implements several search algorithms