Chapter 19
Ontology-Based Clustering of the Web Meta-Search Results

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, as a result of 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. Implementation details and early experimentation results are also provided.

INTRODUCTION

Information retrieval refers to the "representation, storage, organization and access to information items" and its success is strongly related to users' needs (Baeza-Yates & Ribeiro-Neto, 1999), (Heisig, Caldwell, Grebici, & Clarkson, 2010), (Domingo-Ferrer, Bras-Amorós, Wu, & Manjón, 2009). Nevertheless, defining users' needs is not a straightforward issue. Building a query with a set of keywords, as an expression of users' needs and applying that query to a large set of data is not enough. The users have to receive the most relevant results, according to the query. The task became more challenging once the World Wide Web came into scene: "The Web is becoming a

DOI: 10.4018/978-1-4666-1833-6.ch019

universal repository of human knowledge and culture which has allowed unprecedented sharing of ideas and information in a scale never seen before" (Baeza-Yates & Ribeiro-Neto, 1999). Trying to keep up with the continuous growth of the World Wide Web (WWW), the retrieval tools are engaged in a permanent race for faster development in order to reach better performances (Ajayi, Aderounmu, & Soriyan, 2010), (Wang, Tsai, & Hsu, 2009), (Tu & Seng, 2009). Information retrieval doesn't just mean information access (summarization, filtering, search, categorization), but also knowledge acquisition (visualization, mining, extraction, clustering). Thus, besides simple retrieval application, mining and learning applications are needed. Many operations in information retrieval can be automated, such as document indexing or query refinement, but classifications are more often performed manually. For saving time, algorithms were developed for mining documents (Qiu, 2010), (Jeng, Chuang, & Tao, 2010) (Chen, Tseng, & Liang, 2010). These algorithms are based on machine learning," a dynamic, burgeoning area of computer science which is finding application in domains ranging from 'expert systems', where learning algorithms supplement—or even supplant—domain experts for generating rules and explanations (Langley, & Simon, 1995), to 'intelligent agents', which learn to play particular, highly-specialized, support roles for individual people and are seen by some to herald a new renaissance of artificial intelligence in information technology (Hendler, 1997)" (Cunningham, Littin, & Witten, 2001). A good example of machine learning algorithm used in information retrieval is the case in which knowledge bases are built as mirrors of WWW in local computer, thus optimizing the search process (Craven, et al., 2000).

Langley & Simon (1995) identify five major paradigms in machine learning research: rule induction, instance-based learning, neural networks, genetic algorithms and analytic learning. First four of them can be applied in information retrieval: their mechanism is based on learning from information with very simple structure, such as lists of symbolic or numeric attributes. Genetic algorithms are applied to generate structures that represent relationships implicit in the data. According to Lewis (1991), the information retrieval process can be divided into four distinct stages: indexing, query Equationtion, comparison and feedback. Usually, when a researcher tries to improve the retrieval process, one focuses on one of these stages. In clustering techniques links are built between related documents so that indexing becomes more effective (Martin, 1995).

THE RESEARCH CONTEXT

A common way of dealing with efficient information retrieval in web environments is using an ontology approach (Yang, 2010), (Segura, Sánchez, García-Barriocanal, & Prieto, 2011), (Park, Cho, & Rho, 2010). In clustering, the methods which use the ontology approach identify the concepts instead of the index words occurring in web documents (Hotho, Staab, & Stumme, 2003), (Bloehdorn, & Hotho, 2004). The corresponding concepts are identified with the aid of ontology class labels and the WordNet lexical database. In WordNet the terms are organized in synsets, which are sets of synonyms. Using concepts instead of index words reduces the clustering dimensionality, considering that we can have multiple index words replaced by the same concept. Some methods recommend adding the identified concepts to the index words list, and increasing the dimensionality of the process (Jing, Zhou, Ng, & Huang, 2006). As a consequence, it is not possible to consider the concept-based clustering as an implicit solution for the dimensionality issue. The clustering solution proposed in this paper is based on the concepts' identification, via WordNet and the replacement of the index words by the corresponding concepts.

For clustering purpose, a document is represented as a vector of index words in a vector