Web Mining in Thematic Search Engines

Massimiliano Caramia
University of Rome "Tor Vergata," Italy

Giovanni Felici
Istituto di Analisi dei Sistemi ed Informatica IASI-CNR, Italy

INTRODUCTION

In the present chapter we report on some extensions on the work presented in the first edition of the Encyclopedia of Data Mining. In Caramia and Felici (2005) we have described a method based on clustering and a heuristic search method- based on a genetic algorithm - to extract pages with relevant information for a specific user query in a thematic search engine. Starting from these results we have extended the research work trying to match some issues related to the semantic aspects of the search, focusing on the keywords that are used to establish the similarity among the pages that result from the query. Complete details on this method, here omitted for brevity, can be found in Caramia and Felici (2006).

Search engines technologies remain a strong research topic, as new problems and new demands from the market and the users arise. The process of switching from *quantity* (maintaining and indexing large databases of web pages and quickly select pages matching some criterion) to *quality* (identifying pages with a high quality for the user), already highlighted in Caramia and Felici (2005), has not been interrupted, but has gained further energy, being motivated by the natural evolution of the internet users, more selective in their choice of the search tool and willing to pay the price of providing extra feedback to the system and wait more time to have their queries better matched. In this framework, several have considered the use of data mining and optimization techniques, that are often referred to as *web mining* (for a recent bibliography on this topic see, e.g., Getoor, Senator, Domingos, and Faloutsos, 2003 and Zaïane, Srivastava, Spiliopoulou, and Masand, 2002).

The work described in this chapter is bases on clustering techniques to identify, in the set of pages resulting from a simple query, subsets that are homogeneous with respect to a vectorization based on *context* or *profile*; then, a number of small and potentially good subsets of pages is constructed, extracting from each cluster the pages with higher scores. Operating on these subsets with a genetic algorithm, a subset with a good overall score and a high internal dissimilarity is identified. A related problem is then considered: the selection of a subset of pages that are compliant with the search keywords, but that also are characterized by the fact that they share a large subset of words different from the search keywords. This characteristic represents a sort of semantic connection of these pages that may be of use to spot some particular aspects of the information present in the pages. Such a task is accomplished by the construction of a special graph, whose maximum-weight clique and k-densest subgraph should represent the page subsets with the desired properties.

In the following we summarize the main background topics and provide a synthetic description of the methods. Interested readers may find additional information in Caramia and Felici (2004), Caramia and Felici (2005), and Caramia and Felici (2006).

BACKGROUND

Let P be a set of web pages, and indicate with $p \in P$ a page in that set. Now assume that P is the result of a standard query to a database of pages, and thus represents a set of pages that satisfy some conditions expressed by the user. Each page $p \in P$ is associated with a score, based on the query that generated P, that would determine the order by which the pages are presented to the user who submits the query. The role of this ordering is crucial for the quality of the search: in fact, if the dimension of P is relevant, the probability that the user considers a page p strongly decreases as the position of p in the order increases. This may lead to two major drawbacks: the pages in the first positions may be very similar (or even equal) to each other; pages that do not have a very high score but are representative of some aspect of set P may appear in

a very low position in the ordering, with a negligible chance of being looked at by the user.

Our method tries to overcome both drawbacks, focusing on the selection, from the initial set *P*, of a small set of pages with a high score and sufficiently different one from each other. A condition needed to apply our approach is the availability of additional information from the user, who indicates a search context (a general topic to which the search is referred to, that is not necessarily linked with the search keywords that generated the set *P*), and a user profile (a subjective identification of the user, that may be either provided directly by choosing amongst a set of predefined profiles, or extracted from the pages that have been visited more recently by that user).

MAIN THRUST OF THE CHAPTER

The basic idea of the method is to use the information conveyed by the search context or the user profile to analyze the structure of *P* and determine in it an optimal small subset that better represents all the information available. This is done in three steps. First, the search context and the user profile are used to extract a finite set of significant words or page characteristics that is used to create, from all pages in *P*, a vector of characteristics (page vectorization). Such vectorization represents a particular way of "looking" at the page, specific of each context/profile, and will constitute the ground on which the following steps are based.

Second, the vectorized pages are analyzed by a *clustering algorithm* that partitions them into subsets of similar pages. This induces a two-dimensional ordering on the pages, as each page *p* can now be ordered according to the original score within its cluster. At this point the objective is to provide the user with a reduced list that takes into account the structure identified by the clusters and the original score function.

This is done in the third step, where a *genetic algorithm* works on the pages that have higher score in each cluster to produce a subset of them that are sufficiently heterogeneous and of good values for the original score. In the following we describe the three steps in detail.

We then report on the technique proposed to select pages that, beside being relevant for the search, are induced by a set of strongly connected words, with a proper definition of connection; in addition, we are inclined to select in this set those words that have a high degree of similarity with the search keywords, to enhance the significance of the induced pages. This problem is formulated as a maximum-weight clique problem on a graph whose nodes are associated with the initial set of words and whose arcs convey a weight based on the cardinality of page subsets associated with the word nodes.

Page Vectorization

The first step of the method is the representation of each page that has been acquired by a vector of finite dimension *m*, where each component represents a measure of some characteristic of the page (page vectorization). Clearly, such representation is crucial for the success of the method; all the information of a page that is not maintained in this step will be lost for further treatment. For this reason it is very important to stress the thematic nature of the vectorization process, where only the information that appears to be relevant for a context or a profile is effectively kept for future use. In the most plain setting, each component of the vector is the number of occurrences of a particular word; one may also consider other measurable characteristics that are not specifically linked with the words that are contained in the page, such as the presence of pictures, tables, banners and so on. As mentioned above, the vectorization is based on one context, or one profile, chosen by the user. One may then assume that, for each of the contextes/profiles that have been implemented in the search engine, a list of words that are relevant to that context/profile is available and a related vectorization of the page is stored. In Caramia and Felici (2005) we propose two methods to determine the initial list of words.

Page Clustering

There has been extensive research on how to improve retrieval results by employing clustering techniques. In several studies the strategy was to build a clustering of the entire document collection and then match the query to the cluster centroids (see, e.g., Willet, 1988). More recently, clustering has been used for helping the user in browsing a collection of documents and in organizing the results returned by a search engine (Leuski, 2001, and Zamir, Etzioni, Madani, and Karp, 1997), or by a metasearch engine (Zamir and Etzioni, 1999) in