Chapter 7
A Flexible Language for Exploring Clustered Search Results

Gloria Bordogna
CNR IDPA, Italy

Alessandro Campi
Politecnico di Milano, Italy

Stefania Ronchi
Politecnico di Milano, Italy

Giuseppe Psaila
Università di Bergamo, Italy

ABSTRACT

In this chapter the authors consider the problem of defining a flexible approach for exploring huge amounts of results retrieved by several Internet search services (like search engines). The goal is to offer users a way to discover relevant hidden relationships between documents. The proposal is motivated by the observation that visualization paradigms, based on either the ranked list or clustered results, do not allow users to fully appreciate and understand the retrieved contents. In the case of long ranked lists, the user generally analyzes only the first few pages. On the other side, in the case the documents are clustered, to understand their contents the user does not have other means that looking at the cluster labels. When the same query is submitted to distinct search services, they may produce partially overlapped clustered results, where clusters identified by distinct labels collect some common documents. Moreover, clusters with similar labels, but containing distinct documents, may be produced as well. In such a situation, it may be useful to compare, combine and rank the cluster contents, to filter out relevant documents. In this chapter the authors present a novel manipulation language, in which several operators (inspired by relational algebra) and distinct ranking methods can be exploited to analyze the clusters’ contents. New clusters can be generated and ranked based on distinct criteria, by combining (i.e., overlapping, refining and intersecting) clusters in a set oriented fashion. Specifically, the chapter is focused on the ranking methods defined for each operator of the language.

DOI: 10.4018/978-1-60566-858-1.ch007
INTRODUCTION

Retrieving useful and relevant information over the Internet is not an easy task by using current search engines. Too often, the relevant documents are merged and hidden in the long ranked list of retrieved documents. The list can span over hundreds of web pages, each one containing just few retrieved items.

To discover the relevant documents, users have to browse the titles of the documents, but generally only the first two or three web pages are analyzed, while the content of the successive ones is missed. Thus, if users do not find what they are looking for in the first pages, they reformulate a new query trying to capture what they are looking for in the top ranked items.

Some users turned to using meta-search engines, such as mamma, dogpile, Metacrawler etc., in an attempt to optimize their search effort. The assumption is that, if one regards a search engine as an expert in finding information, by using several experts together one should achieve better results. However this is not generally true, since meta-search engines fuse the individual ranked lists of documents retrieved by each underlying system by applying rigid and static fusion functions, applying criteria that are not transparent to the user. The side effect of list merging is to augment the number of retrieved documents, leaving the user skeptics as far as the actual correspondence of the ranking to her/his relevance judgments is concerned. Furthermore, the retrieved documents besides the first page will be hardly analyzed by users; thus, this makes much of the meta search engine’s effort useless.

To overcome this problem, some search services such as vivisimo, Snaket, Ask.com, MS AdCenter Labs Search Result Clustering etc., have shifted from the usual ranked list to the clustered results paradigm. This consists in organizing the documents retrieved by a query into containers (i.e., clusters), possibly semantically homogeneous with respect to their contents, and in presenting them labeled, so as to synthesize their main content focus (Osinski, 2003).

Clustering is often proposed as a viable way of narrowing a search into a more specific query, like in Ask.com (Chen & Dumais, 2000; Zamir & Etzioni, 1999; Coates et al., 2001).

On the other side, one problem users encounter with such clustered results, is the inability of fully understanding and appreciating the contents of the clusters. This is mainly due to the short and sometimes bad quality of the labels of the clusters, which generally consist of a few terms, or individual short phrases, which are automatically extracted from the documents of the cluster based on statistics and co-occurrence analysis. Often, several clusters have similar labels that differ just for a single term. To effectively explore the cluster contents, users have no other means than clicking on the cluster labels and browsing the clusters themselves.

This problem is much more apparent when submitting the same request to distinct search engines, each one producing a group of clustered results reflecting distinct criteria. For example, the Gigabits search engine clusters retrieved documents by their freshness dating (Last Day, Last Week, Last Month, Last Year, etc.), the vivisimo search service presents clustered documents. In such a situation, one may want to explore if a given cluster contains documents that are fresh or not; this necessity may occur quite frequently in analyzing news streams (RSS) to find out the frequency of a given news story reported by media as a function of time.

When the groups of clusters are generated by distinct search services, users may be faced with distinct clusters, possibly with the same labels. In this situation, it becomes necessary to explore the relationships between the contents of these clusters to identify common and distinct documents, and re-rank the contents of the clusters based on distinct criteria. This may require the application of several manipulation operations, such as the intersection and join of clusters, as well as their union and re-clustering,