Chapter IX
Contextualized Clustering in Exploratory Web Search

Jon Atle Gulla
Norwegian University of Science and Technology

Hans Olaf Borch
Bekk Consulting AS, Norway

Jon Espen Ingvaldsen
Norwegian University of Science and Technology

ABSTRACT

Due to the large amount of information on the Web and the difficulties of relating users’ expressed information needs to document content, large-scale Web search engines tend to return thousands of ranked documents. This chapter discusses the use of clustering to help users navigate through the result sets and explore the domain. A newly developed system, HOBSearch, makes use of suffix tree clustering to overcome many of the weaknesses of traditional clustering approaches. Using result snippets rather than full documents, HOBSearch both speeds up clustering substantially and manages to tailor the clustering to the topics indicated in a user’s query. An inherent problem with clustering, though, is the choice of cluster labels. Our experiments with HOBSearch show that cluster labels of an acceptable quality can be generated with no supervision or predefined structures and within the constraints given by large-scale Web search.

INTRODUCTION

With the explosive growth of the Internet, Web search engines need to be able to handle billions of documents and deal with queries that may be very short, vague, or even ambiguous. As a consequence, a typical user query may easily result in millions of hits that are only partly related to each other. Overwhelmed by the number of results, many users investigate only the two or three top ranked documents. If nothing interesting shows up, most users reformulate their query rather than
sift through numerous pages of search results. Studies of query logs from large commercial search engines reveal that users tend to type short queries and only look at the first results (Silverstein, Marais, Henzinger, & Moricz, 1999; Spink, Wolfram, Jansen, & Saracevic, 2001).

The idea behind exploratory search is to shift the focus from one-shot document retrieval to a multi-step learning and investigation process. This process is characteristic to situations in which users’ information needs are imprecise or evolving or when their domain knowledge is limited (Kules, 2006). According to Rose and Levinson (2004), the exploratory queries constitute about 20-30% of all Web queries today. Traditional lists of ranked documents do not seem to be sufficient for these exploratory search tasks (White, Kules, Drucker & Schraefel, 2006), and we need additional techniques to help the users analyze the search results efficiently and drill down to the information they were looking for. We would like the users to explore the information relevant to them in a constructive manner, discovering new patterns, new entities, and knowledge they did not even realize they needed. Their initial queries set the stage for this exploration process, but other techniques are needed to extract the knowledge from the search results that will provide further insight into the domain.

A whole range of text mining techniques have been applied to extract additional patterns or knowledge from search result sets. These can be roughly classified according to function, resources, and content:

- **Function**: Whereas some techniques address the documents as a whole, others extract individual concepts or phrases from the documents. The techniques may either describe the content of each document or categorize and compare the documents according to some attributes or terms.
- **Resources**: The analysis may reflect the content of the result documents only, though many techniques adapt the analysis to the user’s information needs or context. Also, the techniques may involve some degree of supervision.

  - **Content**: Whereas some techniques view words as incomprehensible tokens, other include linguistic analyses that take into account morphosyntactic or semantic aspects. For structuring the mining results, everything from incomprehensible tokens to well-defined concepts may be used.

One way of assisting users in finding the desired information quickly is to group the search results by topic. The user does not have to reformulate the query, but can merely click on the topic that most accurately describes his specific information need. The function here is the clustering of documents according to terms found in the documents, but the resources and approach used may differ depending on the way this grouping is implemented. The process of grouping documents is called clustering and was originally used for analyzing numerical data in the field of data mining. It has since been adapted to suit the needs in textual document collections.

The two main challenges with adapting clustering to the needs of Web search engines and textual data have been to find good descriptive labels to the clusters and to cluster documents on-the-fly in response to a particular user query. Traditional data mining approaches are not concerned with labeling clusters, but in return they are often very good at grouping data. Unfortunately, regardless of how good the document grouping is, users are not likely to use a clustering engine if the labels are poor. Clustering performance is also a major issue because Web users expect fast response times. To deal with this linear time, clustering algorithms that can cluster hundreds of documents per second have been developed.

Several commercial clustering engines have been introduced in recent years. Vivisimo, which was rewarded by SearchEngineWatch.com for best