Chapter V

Interactive Query Expansion With Automatically Generated Category-Specific Thesauri*

Fabrizio Sebastiani
Istituto di Elaborazione dell’Informazione, Italy

The categorization of documents into subject-specific categories is a useful enhancement for large document collections addressed by information retrieval systems, as a user can first browse a category tree in search of the category that best matches her interests and then issue a query for more specific documents “from within the category.” This approach combines two modalities in information seeking that are most popular in Web-based search engines, i.e., category-based site browsing (as exemplified by, e.g., Yahoo™) and keyword-based document querying (as exemplified by, e.g., AltaVista™). Appropriate query expansion tools need to be provided, though, in order to allow the user to incrementally refine her query through further retrieval passes, thus allowing the system to produce a series of subsequent document rankings that hopefully converge to the user’s expected ranking. In this work we propose that automatically generated, category-specific “associative” the-
sauri be used for such purpose. We discuss a method for their generation and discuss how the thesaurus specific to a given category may usefully be endowed with “gateways” to the thesauri specific to its parent and children categories.

INTRODUCTION

We here report work in progress within the Eurosearch project, whose purpose is the design and implementation of a European federation of search engines $\varepsilon_1, \ldots, \varepsilon_n$, each addressing a national Web space of documents expressed in the respective languages $L(\varepsilon_1), \ldots, L(\varepsilon_n)$. Each search engine $\varepsilon_i$ in the federation will be capable of answering queries $q_j$ worded in $L(\varepsilon_i)$ that ask for documents written in either of a set of languages $L(\varepsilon_{i,j}), \ldots, L(\varepsilon_{m,j})$ contained in $\{L(\varepsilon_1), \ldots, L(\varepsilon_n)\}$. The search engine will achieve this by translating, in collaboration with search engines $\varepsilon_{i,j}, \ldots, \varepsilon_{m,j}$, the query from $L(\varepsilon_i)$ to each of the languages $L(\varepsilon_{i,j}), \ldots, L(\varepsilon_{m,j})$, dispatching the translated queries to the appropriate engines, and presenting to the user the results returned by them. Of course, $\varepsilon_i$ will also search within its own national Web space in case, as most of the times will indeed happen, $L(\varepsilon_i) \in \{L(\varepsilon_{i,j}), \ldots, L(\varepsilon_{m,j})\}$.

Modalities of Information Seeking

One key aspect of the Eurosearch architecture is the combination of two information access modalities that have traditionally been addressed by separate tools.

The first modality is keyword-based document querying, exemplified by the AltaVista™ search engine (http://www.altavista.digital.com/), in which a user types in a list of keywords or phrases and receives back as a result a list of documents, ranked in decreasing probability (or estimated degree) of relevance to the query, in which these keywords play a (decreasingly) significant role (e.g., appear in the title, in headers, or in the body of the document). The second modality is category-based site browsing, exemplified by the Yahoo™ search engine (http://www.yahoo.com/), in which a user is allowed to browse a hierarchy of subject categories, each of which contains pointers to a select number of Web sites highly relevant to it; after navigating down to the category she is interested in, the user can then browse the sites referred therein.

The Eurosearch architecture integrates the two modalities by allowing the user to first browse a hierarchy of categories, thus docking to the category
Normalization of Relations with Nulls in Candidate Keys
www.igi-global.com/article/normalization-relations-nulls-candidate-keys/3282?camid=4v1a