OmniPaper Smart Information Retrieval Prototype

Bert Paepen, Jan Engelen
Katholieke Universiteit Leuven, Department of Electrical Engineering
Research Group on Document Architectures
Kasteelpark Arenberg 10, 3001 Heverlee, Belgium
bert.paepen@esat.kuleuven.ac.be

The OmniPaper project has implemented several information retrieval prototypes in the area of electronic news publishing. One prototype uses SOAP as communication protocol between the central system and a number of distributed news archives. The second prototype uses an RDF metadata database, enabling direct metadata queries to the central system. Finally the Topic Map prototype uses query expansion and semantic linking for smart metadata search. The Topic Map prototype enhances the search experience by implementing a knowledge layer that combines the semantic content of a lexical database, consisting of concepts and keywords, with a metadata-set of newspaper articles.

After developing and testing three smaller prototypes, the OmniPaper consortium has combined these prototypes in one. In this final prototype a kind of "enhanced full-text search" engine is implemented. This means that the prototype is an interface on top of existing search engines.

When a user submits a query, this query is forwarded to several distributed news archives to retrieve relevant news articles. Next to this, the system: 1) translates queries to enable multilingual search, 2) provides a query refinement mechanism, both in graphic and text-based form, allowing users to adapt their query and 3) provides uniform result ranking algorithm across the different news archives.

In this prototype querying and navigation are considered as alternative methods to find relevant information. Both interact with each other and together they produce a combined user experience that can be expressed as *find what you were looking for and then browse away from it*. In fact, the prototype considers both querying and navigation as a kind of search action and tries to integrate both.

In concrete, keywords in a query are looked up in a dictionary and shown to the user. In the background, the keywords are translated and expanded to related terms. These expanded queries are sent to the underlying full-text search engine(s) in all requested languages.

In the graphical tool ("web of concepts") users can redefine the meaning of their query words, resulting in an updated query and result set. Both disambiguation (choosing one meaning of a word out of many) and refinement (browsing to related words) are possible. Figure 1 shows the web of concepts for the query "poll Indonesia". The word "Indonesia" is recognized in only one concept, "Dutch East Indies", whereas the word "poll" has many different meanings. If you select the meaning "canvass" for example, this word is replacing the word "poll" in the original query. After selection the concept "canvass" can again be expanded to related concepts, be it more general or more specific in meaning.

In the textual tool only refinement is possible. The user gets a list of words that are related to the words appearing in the query, grouped into more similar, more specific and more general terms. Then the user can change his/her query using these proposed words.

Figure 1: graphical query refinement for the query "poll Indonesia"

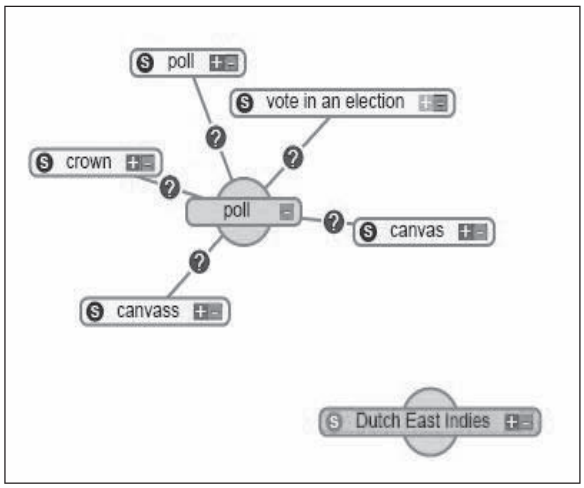

Figure 2: textual query refinement for the query "poll Indonesia"