User Interface Design for Search-Term Recommendation and Interactive Query Expansion Services

A general problem with searching via keyword matching is the so called „vocabulary problem“ (Furnas et al., 1987): The same idea or search query can be expressed in a variety of ways. Using current web search engines by typing in the right term typically leads to a list of documents where the “right one” is also included. However, even if you are using a “wrong” term there is a high probability of getting a non-empty result set – simply saying that you will find documents that contain the wrong term. Now let’s have a look at todays Digital Library (DL) systems or specific databases where controlled vocabularies, usually a Thesaurus, is used for classifying publications. They often contain the metadata of publications but lack the full texts or even an abstract. In this situation the vocabulary problem can become quite serious. If the search doesn’t use one of the controlled terms the document was indexed with the chance of getting relevant documents is very low, there is a higher chance than in the Web of getting an empty result set. For many users this leads to a search strategy to reflect the unnatural nature of keyword querying. In Aula et al. (2005) one expert articulated: “I choose search terms based not specifically on the information I want, but rather on how I could imagine someone wording […] that information.”

Petras (2006) presented a method to overcome the vocabulary problem called search term recommendation. She used specialized search term recommender (STR), which were trained to map free terms from the users mind to one or more specialized controlled vocabularies (e.g. thesauri or subject headings) using a statistical method (Gey 2001). We implemented such a STR by using a probabilistic latent semantic analysis (pLSA) and support vector machine (SVM) based classification tool.

Modern information-seeking support systems (ISSS) try to make use of automated transformations and expansion of textual queries by using e.g. stopword lists or stemming (Hearst 2009, Petras et al. 2007) but there is often a need for query specification via entry form interfaces. A dynamic query term suggestion is implemented in many modern web search engines but these systems only try to make suggestions based on prefix matches (user types “canc” and the system suggests “cancer”, “cancel” and so on).

White and Marchionini (2007) performed a study on a similar interactive method which they called “real time query expansion”. After the user types a word and presses the space bar, the system presents terms based on the surrogates of the ten top-ranked documents. Using a very different data foundation for their suggestions this supporting method might be adaptable for the STR. So far, it lacks an evaluated user interface.

A pretest using 125 CLEF topic sets (2003 and 2007) showed, that the training corpus (taken from the social science literature database SOLIS (GESIS 2009) of 370.000 documents (title, abstract and intellectually indexed via a thesaurus) is too large and too scattered to present contextually mapped terms for one-to-one mappings: the pretest showed that looking up only one single free term presents to many controlled terms which leads to an ineffective support for the searcher. This was due to a missing context (e.g. “politics” can be in the context of “foreign affairs“, „disenchantment with politics“ etc.) the free terms can belong to. By adding only 1-2 additional terms, which describe the context in which the original term was meant, the STR returned effective term mappings. Even if Jansen et al. (2007) showed that a normal web searcher uses 2.8 terms per query there still is a need for adding more contextual information to a STR by using an intelligent user interface for this ISSS.

For the NKOS workshop, we plan to present an overview of the implemented STR and the general techniques to build and design query expansion tools for web search engines and DLs. We will focus on state-of-the-art solutions in interactive query expansion and will show a prototype for such an interactive search term recommendation tool.

This work was done in the Value-Added Services for Information Retrieval (IRM) project funded by Deutsche Forschungsgemeinschaft (DFG - INST 658/6-1).

References:

A. Aula, N. Jhaveri, and M. Käki. Information search and re-access strategies of experienced web users. Proceedings of the 14th International Conference on World Wide Web (WWW'05), pages 583–592, 2005.

GW Furnas, TK Landauer, LM Gomez, and ST Dumais (1987). The vocabulary problem in human-system communication. Communications of the ACM, 30(11):964–971, 1987.

GESIS – Leibniz Institute for the Social Sciences (2009). SOLIS - Social Science Literature Information System, http://www.gesis.org/en/services/specialized-information/databases-and-information-systems/solis-social-science-literature-information-system/

Fredric Gey, Michael Buckland, Aitao Chen, and Ray Larson (2001). Entry vocabulary – a technology to enhance digital search. In Proceedings of HLT2001, First International Conference on Human Language Technology, San Diego, pages 91–95, March 2001.

B.J. Jansen, A. Spink, and S. Koshman (2007). Web searcher interaction with the Dogpile.com metasearch engine. Journal of the American Society for Information Science and Technology, 58(5):744–755, 2007.

Gary Marchionini, Ryen White, Nick Belkin, Gene Golovchinsky, Diane Kelly, Peter Pirolli, mc schraefel (2008). Information Seeking Support Systems: An invitational workshop sponsored by the National Science Foundation, http://ils.unc.edu/ISSS/

Vivien Petras (2006). Translating Dialects in Search: Mapping between Specialized Languages of Discourse and Documentary Languages. PhD thesis, University of California, Berkeley, 2006.

Petras, Vivien and Baerisch, Stefan and Stempfhuber, Maximillian (2008). The Domain-Specific Track at CLEF 2007. In Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers, pages 160-173, 2008

Ryen W. White and Gary Marchionini. Examining the effectiveness of real -time query expansion. Inf. Process. Manage., 43(3):685–704, 2007.