Exploiting Markup Structure for Intelligent Search (Dissertation Abstract)

Udo Kruschwitz
Department of Computer Science, University of Essex
Wivenhoe Park, Colchester, CO4 3SQ, United Kingdom
udo@essex.ac.uk

Collections of digital documents can nowadays be found everywhere in institutions, universities or companies. Examples are Web sites or intranets. But searching them for information can still be painful. Searches often return either large numbers of matches or no suitable matches at all.

Such document collections can vary a lot in size and how much structure they carry. What they have in common is that they typically do have *some* structure and that they cover a limited range of topics. The second point is significantly different from the *Web* in general.

Our aim is to assist a user in the search process so that information can be located more easily. The type of search system that we propose will go beyond a standard search engine. Apart from displaying the best matches in some ranked order it can also suggest ways of refining or relaxing the query. We guide a user through the information available.

In order to suggest sensible query modifications we would need to *know* what the documents are about. Explicit knowledge about the document collection that has been encoded in some electronic form is what we need. However, typically such knowledge is not available. So we construct it automatically.

This thesis demonstrates three main aspects:

- It demonstrates how document markup structure can be used to construct domain models for collections of partially structured documents.
- We show how such knowledge can be utilized when searching the document collections. We introduce a simple dialogue manager that can apply the model as part of an online search system.
- Two implemented search systems - *UKSearch* and the YPA - demonstrate the usefulness of this approach. The systems have been implemented to work on different document collections. There is a particular focus on the evaluation of the applications.