Design of a user interface for searching documents indexed with controlled terms

R. E. de Vries and N.J.I. Mars

Netherlands Institute for Scientific Information Services (NIWI)
PO Box 95110, 1090 HC Amsterdam, Netherlands
Email: repke.de.vries@niwi.knaw.nl nicolaas.mars@niwi.knaw.nl
Http://www.niwi.knaw.nl

Introduction

Though information systems using controlled language for indexing and searching (classification systems, subject headings and thesauri) have a longer tradition, many retrieval systems today are based on free-text searching: the natural language of words in titles, abstracts or the full text of documents. Both research and common experience have by now identified the relative weaknesses and strengths of these two approaches. For example from a summary given by Aitchison et al.: for controlled language " an artificial language has to be learned by a searcher, [but] the burden of searching is eased [because it] controls synonyms [..] and leads [from] specific natural language concepts to the nearest preferred terms [..] [and] avoids precision loss through over-exhaustivity" whereas for natural language "words and phrases used by searcher are [his own], [but] the intellectual effort is placed on searcher [and] exhaustivity may lead to loss of precision".[1]
Some of the NIWI databases[2] hold thesaurus indexed documents but the user interface design for searching these documents so far only offered users free-text searching. Not only for reasons as given above by Aitchison et al., but certainly also by looking at users'of the NIWI databases own preferences, we decided for a different design. This interface uses the same collection of controlled terms used for indexing. By accommodating expert and non-expert user alike, the design seeks to reconcile the often quite complicated nature of collections of controlled terms with the fact of a web environment open to all users. The interface offers assistance in two different ways. One approach helps users choose a particular controlled term from all possible terms in the collection and searches in a document only the terms applied to it in indexing. This assistance comes to the user in two forms: walking the structure in relationships between terms in a top-down fashion and picking a term, or alternatively users' own choice of terms, which are standardized to controlled terms before searching. The other approach is an extra and searches all of a document without distinction but first expands the users' query words with equivalent terms. Evidence for equivalence is taken from the collection of terms.
These two approaches are based upon the two types of information present in a thesaurus: the first type being the structure of the domain, as evidenced by the relations between the various concepts. The second type the (possibly many) ways in which each concept can be expressed in natural language. Often, one of the possibly many terms associated with one concept is selected to play the role of preferred term; this preferred term is then used to refer to the concept whenever needed.

Implementation for two web databases with bibliographic references at NIWI

The design is currently[3] undergoing implementation in the following way:
users start with a search form with two or more entry fields. Each entry field has a pick list where the user chooses which part of the document is searched and a button indicating that assistance can be given in choosing a controlled term from all possible terms, when searching "applied indexing terms" in documents. This assistance helps a user understanding the domain (subject field) to which the documents to be searched belong, by showing the conceptual structure of the domain as it is represented by the hierarchical structure in the thesaurus. Alternatively when searching "applied indexing terms" in documents each entry field can be searched free-text but now users own search words will first be standardized to preferred terms in the thesaurus. Lastly each entry field can be searched free-text in combination with the choice "anywhere in document": in this case users own search words will first be expanded with synonyms taken from the thesaurus.

The web interface offering assistance in choosing controlled terms is considered to be the main design challenge. Programmed reduction or expansion of query terms by drawing upon information in the thesaurus, needs very careful explanations to the user when presenting search results but is otherwise hidden. Offering assistance needs dialogue and a clear communication to the user about when the thesaurus is searched and when the thesaurus indexed documents. Presently the thesaurus can both be searched and browsed. Browsing starts with only the (limited number of) top terms to avoid long alphabetical lists familiar in published thesauri.[4] Searching and browsing are in the same interface and have the same text oriented presentation with a central location for the current term and position in the thesaurus hierarchy. Broader terms are to the left, narrower to the right and "use for", related terms and scope notes also in the middle. As further information to the user, a count is given of the number of documents that would be selected by each of the presented terms. Hypertext linking allows for walking the thesaurus and at the same time hides its complexity. With an also centrally placed button the user starts the document search, with button labeling both reminding the user of going from thesaurus to document searching and repeating the currently chosen thesaurus term.

Thesaurus presentation with a web interface in practice takes quite different approaches. For example HASSET has a design with several web pages and a complicated presentation of its thesaurus.[5] Another example SOSIG has a reasonably clear presentation of thesaurus terms and of the difference between searching its thesaurus and the SOSIG indexed information, but again needs several web pages and does not allow for walking or browsing the thesaurus.[6]
It is hoped that the NIWI design where the user has one web page with minimal or no scrolling, an integration of searching, browsing and presenting a thesaurus together with launching the document search, can further contribute to a more standardized approach for an as general as possible web audience.
Several usability tests are planned to guide and where necessary correct the implementation under way.

[1] [Aitchison 97] Aitchison, Gilchrist and Bawden, *Thesaurus construction and use: a practical manual*, Third edition, ASLIB, 1997, Table with comparison at page 6.
[2] http://www.niwi.knaw.nl/   NIWI's website
[3] October 1998
[4] [Aitchison 97] pp 91-126
[5] HASSET web thesaurus: http://dasun1.essex.ac.uk/services/zhasset.html
[6] SOSIG subject oriented gateway with web thesaurus:
http://sosig.esrc.bris.ac.uk/roads/cgi/thesaurus.pl