From: AAAI-97 Proceedings. Copyright © 1997, AAAI (www.aaai.org). All rights reserved.

Dynamic Organization of Search Results Using a Taxonomic Domain Model

Wanda Pratt
Section on Medical Informatics
Stanford University, Stanford, CA 94305
pratt@smi.stanford.edu

Current information-retrieval tools return the results of a search as a flat list of documents. Some tools order the documents according to relevance criteria, such as date created or number of matching search terms, but few group them in a meaningful way. The returned list of search results is often so long that it is too time consuming for the user to browse and understand all of it. I propose that organizing search results into categories will help users to explore and understand the information space related to their query.

I am developing an approach that automatically generates a hierarchical organization of search results and assigns the appropriate documents to each category based on the type of query, on the documents retrieved, and on a taxonomic model of the domain. The goal of my approach is to combine the main advantage of clustering techniques (organization dependent on the document set) with the main advantage of classification techniques (meaningful groupings and labels for those document groups). I call this approach **dynamic categorization** because it dynamically generates the categorization structure, as well as the category labels.

The categorization generated by this approach should help users to find specific information efficiently, and to learn more about the information that is available. For example, consider a woman with breast cancer who wants to know what possible complications of a mastectomy have been investigated in the medical literature. A search in the medical literature using the keywords *mastectomy adverse effects* yielded over 350 relevant documents. Without a more specific question, it is not possible to narrow the search criteria without artificially eliminating relevant documents. If the user were merely to browse through the list, she might never form an accurate model of all possible complications. A tool that categorized the documents according to the adverse effects discussed would help her to see the possibilities, and would enable her to browse the documents easily for the effects that are of most concern to her.

Dynamic categorization is based on three key premises: (1) an appropriate categorization depends both on the user's query and on the documents returned from the query, (2) the type of query can provide valuable information about the expected types of categories and about the criteria for assigning documents to those categories, and (3) taxonomic knowledge about terms in the document can enable useful and accurate categorization. This approach requires two domain models: a terminology model and a query model. The **terminology model** must be a hierarchical model of domain terms, where **terms** may be single words, abbreviations, acronyms, or multiword phrases. The **query model** provides a mapping between the types of queries that a user may enter and the categorization criteria for generating categories that correspond to the user's query. The **categorization criteria** are the taxonomic constraints that must be satisfied by the document's keywords for that document to be placed in the category. The category labels are generated from the keywords that matched the categorization criteria.

The resulting list of categories and their corresponding documents is organized into a hierarchy based on the terminology model and on the distribution of documents from the search results. When there are many categories at one level in the hierarchy, the categories are grouped under a more general label. The system generates the more general label by traversing up the terminology model to find a term that is a parent to the category label of several document categories.

I have developed a prototype of this approach for the medical domain. The system uses the terminology model created by the National Library of Medicine, the Unified Medical Language System (UMLS), which provides information on over 500,000 biomedical terms, and a query model that I have developed. As an example, consider the query: *What are the complications of a mastectomy?* The query type is *treatment: complications*. One of the categorization criteria for that query type specifies that the keywords must be a *disease or syndrome*. If it finds a document with the keywords *lymphedema, arthritis, diagnostic imaging,* and *middle age*, the system will categorize that document under both *lymphedema* and *arthritis* because they are diseases. It will not categorize it under *diagnostic imaging,* or *middle age* because those terms are not diseases or syndromes. Note that *lymphedema* and *arthritis* were not predefined category labels in the query model. Rather, they were generated dynamically because they satisfied the categorization criteria in the query model.

In summary, I have presented a new approach to organizing search results. Dynamic categorization should help users to understand and explore the results of their search by hierarchically organizing the documents into categories with meaningful labels, and by making the categories and categorization structure a function of the user's query.

Acknowledgments

This work was conducted with the support of the NLM grant LM-07033 and the NCI contract N44-CO-61025. I thank my advisors, Larry Fagan and Marti Hearst, for helping me to formulate this research. Thanks to Lyn Dupre and Ramon Felciano for comments on this manuscript.

Copyright ©1997, American Association for Artificial Intelligence (www.aaai.org). All rights reserved.