Enriched thesauri as networked knowledge bases for people and machines

Paper presented at the CENDI Conference Controlled Vocabulary and the Internet
Bethesda, MD
1999 September 29

Dagobert Soergel
College of Library and Information Services
University of Maryland
College Park, MD 20742
301-405-2037 ds52@umail.umd.edu www.umd.clis/faculty/soergel

Abstract

The presentation will address the opportunities offered through automation generally and the Web environment in particular for structuring thesaurus databases and presenting thesaurus data It will argue for much richer thesaurus structures with much more information - differentiated relationships that allow an extension of thesauri to include precise representation of large amounts of factual information (some of which is included now, but only vaguely, such as organism RT disease rather than organism causes disease); full definitions and not just usage notes; priority levels for thesaurus information to guide display, such as having a short definition with the user being able to access a longer definition and definitions from many different places, including links into texts (parts of documents) that explicate a concept, and links into graphical representation of concept relationships, such as causal influence graphs; and maintenance of information on meaningful sequencing of concepts It will argue for more powerful displays that let the user explore hierarchic and network structures at various levels of detail and amount of information, such as coupled overview and detail windows, choice between linear/text and graphical displays, use of colors. As mentioned, adaptation of the level of detail and amount of information to the user's needs requires support from the thesaurus structure. The presentation will argue for connectedness - clickable relationships within one thesaurus and, more importantly, to specific entries in other thesauri (this requires a standard on how such links should be established and maintained in the face of constant change, including a standard how to create anchors inside a thesaurus Web page and a standard on how to link to specific entries in a thesaurus that exists in form of a Web accessible database). Ultimately, this would lead to a utility that would provide simultaneous access to many thesauri and integrate the information for the user. The presentation will argue for using the Web to support users in maintaining their own personal thesauri (possibly embedded in some large public thesaurus) and to create mechanisms for collaborative maintenance of thesauri. It will also argue for a thesaurus registry that would always direct the user or other systems to the proper URL - URIs for thesauri; such a registry could be used in conjunction with the Dublin Core facility for the identifying the vocabulary of origin for subject metatags to let the user interact with any of these vocabularies directly The presentation will also address the marriage of thesauri and other knowledge organization systems with dictionaries for natural language processing to create more powerful tools for sophisticated text understanding, translation, and retrieval.

Enriched thesauri
Networked knowledge bases
for people and machines

Dagobert Soergel
College of Library and Information Services
University of Maryland

Exploit the possibilities of the new medium

- Data structures of adequate complexity for rich content
- Searchability and selectivity
- Flexibility of display
- Processing power and inference
- Linkage

Expanded functions of thesauri

- Convey meaning, orientation, and structure

- Provide rich relationships and definitions
  Give facts

- Knowledge-based assistance for indexing and searching, behind the scenes or collaboratively with the user

- Linkage to thesaurus entries from text.
  Linkages among thesauri
  Integrated access system

- Assistance to users in maintaining their own thesauri
  Collaborative development and maintenance

Convey meaning, orientation, and structure

- Assists any user thinking about a problem

- Helps with better query formulation

- Requires good methods for displaying structure. Most thesaurus interfaces provide local views but not views of the structure at large

- Examples

    Hierarchical display

    Concept graph

    Facets to elicit query

Convey meaning, orientation, and structure. Continued

- Meaningful arrangement. There is no need for alphabetical arrangement in online environments

- Requires intensive effort in developing meaningful structure

Definitions

- A thesaurus should give full definitions, not just usage notes
- Multiple definitions
- Links to document segments that elaborate on the concept

Rich relationships. Give facts

- Examples

  Cancer *combine with* Body part (RC)

    When cancer is indexed or searched, the system posts a reminder about body part

  Bromocriptin *treats* Alcohol withdrawal

    Now shown, if at all, as

    Alcohol withdrawal agents NT Bromocriptin

  Early behavior disorder *is risk factor for*
  Alcohol or other drug disorder

  Alcohol *causes* Liver disease

Rich relationships. Give facts. Cont.

- Problem: The very richness of information will be overwhelming; too many types of relationships, too many relationships for any one term (there can be 50 or more risk factors)

- Solution: Flexible display. User can select information to be displayed by

  type of relationship and

  priority of relationship

Knowledge-based assistance for indexing and searching, behind the scenes or with the user

Searching

- Expand use fo common techniques:

  Synonym expansion (query term mapping)

  Hierarchic expansion

- Knowledge-based elicitation of user requirements

- Knowledge-based clustering of search results

Knowledge-based assistance, continued

Indexing

- Example: MedIndex

  Can be used for assisting human indexers and for improved automated indexing

- Natural language processing using tools that combine linguistic dictionary information with hierarchy and other thesaurus information.
  Example: UMLS and its Specialist Lexicon

Example: MedIndex (Susanne Humphrey, NLM)

Indexer enters **Bone Neoplasms**

System displays the Neoplasms frame which shows the facets to be considered when indexing a document on neoplasms. The frame is already specialized for bone neoplasms:

Bone Neoplasms - Current Frame

ANATOMICAL STRUCTURE
    Bone and Bones

SECONDARY-FROM

ETIOLOGY

COMPLICATION

PROCEDURE

PROCESS

HISTOLOGIC TYPE

Indexer decides to work further on ANATOMICAL STRUCTURE, clicks on it, and is presented with a hierarchy.

Body Areas
. Back
. Extremities
. . Arm
. . Leg
. . . Foot
. . . Knee
. Head
. . Face
. Neck
. Pelvis
. Thorax
Bone and Bones
. Facial Bones
. . .Palate
. Leg Bones
. . **Femur**
. . Fibula
. . Tibia
ETC

Indexer selects **Femur**

System checks its knowledge base and responds

Femur not permitted.
The correct MeSH heading is

Femoral Neoplasm

Linkage to thesaurus entries from text.

- Assist readers in understanding text be seeing a definition or seeing a concept in its hierarchical context.

- See a subject descriptor recorded in a metatag in the context of the scheme it comes from.

  This would require a thesaurus registry with URIs for thesauri.

Linkages among thesauri
Integrated access system

- Useful for cross-database searching

- Integrated access useful for getting more information.

- Ideally: A “Virtual Thesaurus” that would provide transparent access to multiple thesauri, dictionaries, and other lexical resources and provide an integrated display of the information about a concept or term.

  The challenge: Do this integration automatically

Assistance to users in maintaining their own thesauri

Collaborative development and maintenance of thesauri