Knowledge Base Discovery Tool*

Erik Eilerts, Kathleen Lossau, Christopher York

Austin Info Systems, Inc.
303 Camp Craft Road
Austin, TX 78746
{eilerts, lossau, yorkc}@ausinfo.com

Abstract

The Knowledge Base Discovery Tool (KBDT) is a suite of tools and components to improve the indexing of and search for documents. KBDT extracts and displays content from documents and builds knowledge indexes based on meaning, rather than keywords. KBDT uses the indexes to perform more intelligent searches. It also includes visualization technology to display relevant results using multi-media, rather than plain text. This paper describes prototypes of two tools in this suite that use components for searching, extraction, and display of requested information. The tools are the Knowledge Base Editor and the Intelligent Information Retrieval Engine.

Overview

KBDT is a suite of tools and components under development by Austin Info Systems, Inc. (AIS) to improve the indexing of and search for on-line documents. This suite consists of the following components:

- **Extraction tool** - extracts content (not keywords) from documents in multiple formats (e.g. HTML, PDF, Microsoft Word)
- **Knowledge base** - contains an ontology that is used for content extraction and stores content-based indexes to documents processed by the extraction tool
- **Search portal** - provides a single interface for searching distributed information sources (world-wide web, ODBC databases, etc.) using multiple search strategies (including existing search engines)
- **Results engine** – organizes search results based on the document's content
- **Rendering engine** - converts search results into various multi-media formats, including tables, maps, charts, video, HTML, as well as plain text.

Knowledge Base Editor

The Knowledge Base Editor is used by humans to browse the ontology in the knowledge base and by other KBDT components to find content. The ontology is the core of the knowledge base. It is used for parsing and storing content-based indexes. The initial ontology was derived from the WordNet database (Fellbaum 1998). The primary component of the ontology is a "concept. "

In the knowledge base, a concept denotes a collection of synonyms plus a description that indicates the concept's usage. One definition for "concept" is:

something that exists or that can be thought about.

The following example best illustrates this:

- *tank, army tank - a military tank*
- *tank, storage tank - container holding gases or liquids*

The word tank has at least two different meanings or *senses*, making it impossible to specify its usage using the word tank alone. By combining the word tank with its synonyms and a description, a common meaning can be determined. Information in the knowledge base is stored based on the synonym collections or *senses*, rather than as single words. The synonym collections are called Concepts and are organized into the following part of speech categories: Noun, Verb, Adverb, and Adjective. Concepts greatly improve the representative power of the knowledge base by allowing information to be attached to a words' usage, rather than just to the individual word.

The job of the KB Editor is to enable browsing of the Concepts, their parts of speech, and the synonyms that make up each one. The specific functionality that is demonstrated by the Knowledge Base Editor includes:

- Finding a Concept
- Selecting a Concept with multiple senses
- Moving between Concepts by following links.

Finding a Concept

The tool starts by asking for a word that is related to the Concept being searched for (Concept To View). Selecting one of the part of speech buttons or "All" to view Concepts in all parts of speech categories begins the search for the Concepts related to this word. Three possible results are displayed:

- A list of Concepts that make use of the word entered, including each Concept's parts of speech
- Information related to the Concept, if the word entered is only used in one Concept
- An error message indicating that no Concept in the knowledge base corresponds to the word that was entered.

* This work is supported in part by a CECOM C2SID Phase II SBIR *Contract number:* DAAB07-98-C-D027.

Selecting a Concept with Multiple Senses

If multiple Concepts are related to an individual word, a list of the Concepts is displayed, including the part of speech, synonyms, and a description of each concept. The purpose of this step is to help the user select the appropriate Concept based on usage. Based on the previous example, the user has to decide whether they are interested in the Noun-category Concept for an army tank or a tank of water.

Moving between Concepts by following Links

Once a concept has been selected, the user is presented with the following information:

- **Name** - a concatenation of the concept's synonyms and its part of speech
- **Description** - a textual description of the concept, potentially including some sample sentences showing its usage
- **Connections** - a list of connections between the concept and other related concepts

To view one of the connections, the user selects its hyperlink. All connections are links to other concepts, except for the Synonym connections. When a synonym is selected, the knowledge base is searched for all Concepts that use the selected synonym word. The result of the search is displayed according to the options in "*Finding A Concept.* "

Intelligent Information Retrieval Engine

This Intelligent Information Retrieval Engine provides the capability for the user to discover information relevant to a given Concept or set of Concepts. It makes extensive use of the Concepts that are stored in the knowledge base for both query preparation and post-processing of query results.

The specific functionality that is demonstrated by the Intelligent Information Retrieval Engine includes:

- Entering A Query
- Providing Semantic Contexts for a Query's Terms
- Conducting a Search using Semantic Contexts

These tasks provide further details about how the concepts are used.

Entering A Query

The tool starts by asking for a short description of the item to search for. At this point, a typical search engine extracts keywords from this description and then runs the query. Instead, the IIR engine retrieves Concepts related to the words in the description. This allows documents to be located based on the meanings of the words, rather than just the words themselves.

Providing Semantic Contexts for a Query's Terms

The tool next asks for a clarification of the search terms, if multiple Concepts are found for individual words. This is the step where the quality of the search is significantly improved, since the search is now focused on Concepts, not just words. For example, if *tank* was entered, the user must now decide whether they are interested in an army tank or a tank of water. The Concepts used at this step are extracted from the knowledge base.

Conducting a Search using Semantic Contexts

After the user has selected meanings for each of the query's terms, this tool conducts a search of the world-wide web to find documents related to the search topic. The search is focused on documents that contain the selected concepts and do not contain the discarded concepts. For example, if the user had selected *army tank*, then the documents must contain the phrase *tank* or *army tank*. Documents that contain the phrase *storage tank* are discarded.

The quality of the results provided by semantic-context information retrieval depend on the amount and kind of information in the knowledge base related to the user's topic of interest.

Conclusion

The demonstration covers early prototypes of two tools that are part of a larger project devoted to knowledge discovery and intelligent information retrieval. The early prototypes indicate that Concepts provide a useful foundation for discovering relevant information from a wide variety of sources, and encourage future project development in this area.

Future

The next set of tools to be developed include a parsing engine for extracting concepts form documents and a concept builder to help average users expand the contents of the knowledge base. The concept builder presents the user with a list of words extracted from a document that are not yet in the knowledge base and helps the user to create a definition for the new terms. Prototypes of these tools should be available by mid-year 1999.

Demo URL

The demonstration can be viewed via the Internet by accessing the following URL:

http://www.ausinfo.com/kbdt/aaai99.html

References

Fellbaum, C. ed. *1998 WordNet: An Electronic Lexical Database.* Cambridge, Mass.: MIT Press.

Mayk, et.al. 1998 *A Knowledge Based Doctrine Tool for Command and Control*, Proceedings of the Command and Control Research and Technology Symposium.