Model-driven indexing: Indexing by ignoring content

Kishore Swaminathan
Center for Strategic Technology Research
Andersen Consulting
Chicago IL 60606

Indexing can be viewed is the process of segmenting a large information with the indices acting as descriptors for the subspaces. In the case of information retrieval systems, the indices are used to retrieve documents based on a user query, whereas in the case of knowledge navigation systems, the indices circumscribe browsable spaces of documents. In this paper, I take the position that traditional approaches to indexing are "data-driven"; an alternate "model-driven" approach to indexing has considerable promise, particularly within an enterprise setting.

Traditional approaches to information retrieval assume that the most valuable indices for a document can be derived from the content of the document. While statistical techniques attempt to derive the content from statistically significant keywords, AI techniques go one step further to derive content through parsing and inferencing. In either case, the indices are completely determined by the documents in a document collection. I refer to this approach as "data-driven."

When a user attempts to locate information from a document collection, the collection itself is only one part of the equation, with the user forming the other part of the equation. For the user, the effectiveness of an indexing scheme depends on how well it supports his/her goals, task requirements and level of expertise, not on abstract measures of effectiveness such as precision and recall. An alternate approach to indexing, therefore, is to start from the user's end, deriving indices based on user's mental models and task models. I refer to this approach as "model-driven[1]."

In the most general case, model-driven approaches are difficult to instantiate for large document collections because of different mental and task models for different users. However, the problem takes on a different hue when we consider information retrieval and navigation within an enterprise setting. Unlike the general populace, mental models and task models exist within enterprises; further, unlike the general populace, an enterprise also wields power and influence over its employees to train them for common mental models and mandate consistent forms of usage through procedures, guidelines and methodologies.

For example, the organizational structure of an enterprise is well-known among its employees, and the employees are generally aware of the practice areas of different organizational within the enterprise. This common mental model can be exploited for indexing documents based on which organization created (or owns) a document, leaving the employee to make the inferential leap from a topical area in which he or she is interested to the likely organizational unit that may own documents in that area.

Further, task models derived from conventions, procedures, guidelines and methodologies constitute an important source of information for indexing. For example, within the same content area, task models can distinguish the information requirements of a salesman from that of a product developer or a project manager. In one of the prototypes we have built in Andersen Consulting, an indexing scheme based on the format and source of documents (e.g., proposals, presentations, reports, white papers, brochures, memos, workplans, newspaper articles etc., the kind of information that is typically ignored by traditional approaches) turned out to be more important than the actual content of the documents: for instance, project managers often look for past workplans for estimating guidelines, sales people look for newspaper and magazine articles about the company for establishing their credentials, and brochures for product information.

In conclusion, the position taken in this paper is that the prevalent approach to indexing a large information space (for both information retrieval and knowledge navigation) is data-driven, essentially ignoring the user. An alternative is a model-driven approach based on common mental models and task models among the users. The latter is less sensitive to changing content and vocabulary and is particularly promising within an enterprise setting.

[1] While processing mechanisms such as scripts and goals do constitute models, they are models of content, rather than models of user or usage of the information.