inxight

Emerging Technologies in Knowledge Management

By Ramana Rao, CTO of Inxight Software, Inc.

This paper provides an overview of a presentation at the Internet Librarian International conference in London in March, 2002. Additional resources available at the time of the talk are available from the web sites above.

Summary

Knowledge Management has been a topic of interest to large corporations for the last 10 years. In essence, the desired goal has been to leverage the intellectual assets of a corporation in the process of improving its business performance. One aspect of achieving this goal is to provide effective tools to employees for accessing available content as they perform their knowledge tasks. Beyond the content available externally from a variety of publishers and electronic information sources, there are often large collections of scattered text documents, so-called "unstructured data." Typical estimates indicate 80 to 90% of corporate information is unstructured as opposed to the structured data available in enterprise databases. A number of new capabilities (beyond traditional search and directories) for leveraging external and internal content have emerged in the last few years.

Automated Processing in Support of Knowledge Work

Information Access capabilities can be understood as automatic support for many of the activities that knowledge workers perform as they process information in their knowledge work. All knowledge work can be viewed as an iterative series of stages starting with collecting information, analyzing it for relevant or useful facts or insights, organizing the materials, and eventually synthesizing them for delivery. In most serious knowledge work, of course, knowledge workers perform deeper analysis than what information professionals might as they prepare "research" for others to use in their actual planning or decision-making task. Nevertheless, the basic phases of preparing information for use provides a suitable framework for understanding automatic techniques.

Figure 1. portrays a processing chain that connects information sources to users. It is important to remember that ultimately a user must benefit from the entire pipeline, otherwise whatever can be said about any capability or its effectiveness will be meaningless. In corporate circles, of course, the organizational benefit derives from the cumulative benefit to the people that serve the organization over time.

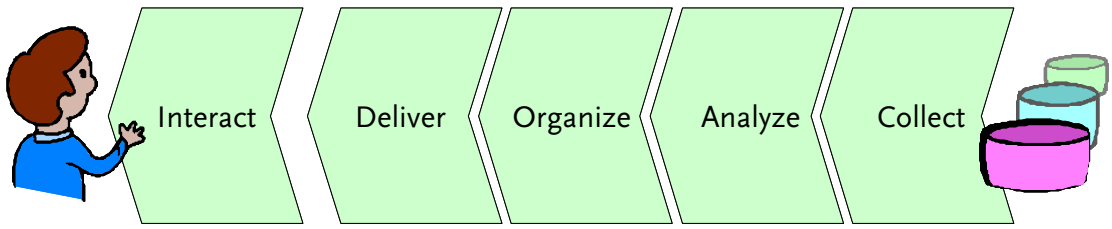

Figure 1: Processing Chain to Support Information Access

Search & Browsing and the Web

Traditional technologies for information access can be understood using the basic model of Figure 1. *Search engines* are really composed of two elements that together cover this entire chain in a specific way to support a very specific form of end user interaction based on queries in, matching documents out. Indexing engines collect the documents, analyze them for their words and phrases and organize them into an index structures. Then query engines take queries from end users that express their information need and process them against the index to deliver the documents that match the information need of the user. In the Internet environment, collection became a process of crawling or spidering Web sites and their pages as they sprouted up furiously, and processing queries became a series of battles with and refinements on doing well against extremely thin or at least thinly-expressed information needs.

The Web originally started as information-sharing technology that improved on the model of individual documents stored in directories on file servers or FTP server.
Links across documents provided the necessary mechanism for using documents to organize other documents. The "mostest for the leastest" style standards for URLs, HTML and HTTP adequately define mechanisms for universal addressing and requesting and serving documents. The simplicity of these protocols allowed for a incredibly rapid and widespread implementation and deployment of web servers and browsers. As an information access technology, the Web relies on the order being produced by humans operating against a set of conventions drawn from organizational, social or cultural milieus. Human authors did the work of producing documents that provided navigation and editorial structures over content published by them or others.

The early Web saw the rise of Internet search engines that indexed pages and sites, but it was Yahoo! that achieved the greatest success in helping users find what they were seeking using an *Information Directory* approach. Using humans to categorize sites into a well-organized taxonomy, they were able to provide a less exhaustive, but more reliable way, of finding resources on the Web. The directory provides an overlay structure on the Web quite independently of the linking or content structure of pages as produced by authors. This "library catalog" approach to web content worked quite well on dimensions of intelligent organization and usability, and provided a good alternative to the search engine approach in a broader range of situations.

Search, browsing and directories all provide an integrated approach across the processing chain, but are quite limited or specific in what they do in each phase. The Web provided a robust delivery mechanism. Search provides a precise but brittle method for access that works great in some situations and terribly in others. Meanwhile, human authoring, analysis and organization can provide quite usable and accurate results, but requires considerable expense and custom design. All of these techniques suffer as the amount and diversity of content increases. These traditional technologies are the initial basis for intranet content and information access, and only now are they being augmented with greater degrees of automatic support.

Enterprise Portals

Enterprise Portal technology emerged as an attempt to provide uniform access to available enterprise collections and resources. Enterprise Portal products typically provide an open framework that allows for integration of specialized components at both the delivery and collection ends of the chain:

- front-end "portlets" or "gadgets" present content to users, and are generally implemented as typical web components using HTML, web scripting and objects (e.g,. Java applets).
- a "catalog" or "metadata repository" captures information about available collections and documents
- back-end "adaptors" allow for collection or connection of back-end content repositories or enterprise applications

These frameworks are quite powerful in providing secure and uniform access, not only to content and documents but also to applications and collaboration technologies. However, they do not specifically address the problems associated with creating useful catalogs for content that is not already cataloged or otherwise structured in content or document management systems. In particular, human effort would be needed to produce high quality catalogs of internal document content.

Unstructured Data Management Capabilities

Search is the major automation technology widely used to access unstructured data. As described above, as powerful as search is, it doesn't fully address the requirements for information access support for knowledge workers. Nor does it begin to exhaust the possibilities for the uses of computational techniques in supporting more effective access. Beyond search, four capabilities are offered by a variety of vendors:

Categorization. Many vendors offer automatic categorization products that sit squarely. At the core, these products automate the process of assigning documents to "categories" which are typically organized in a hierarchical taxonomy. In essence, this can be thought of as trying to produce a "Yahoo! for the Enterprise" without the labor. In reality, as any library or information professional would know, any real method is going to require some

degree of human involvement. Thus to deploy this capability in a real setting, support must be provided for taxonomy management and evolution, and for the process of "training" or "tuning" the automatic behavior of the categorization product. Ultimately, the most effective techniques are going to be a human-efficient blend of computation and editorial involvement.

Information Extraction. These capabilities live within the analysis phase of the chain. Basic linguistic analysis pulls out words, phrases, and sentences, and tags these elements with type information that identifies key entities like people, organizations, products, dates and so on. This level of analysis is necessary to support higher-level functions including automatic summarization, and key concept or topic identification. In the next few years, there will be considerable progress on making even higher-level observations from and across documents using techniques for fact extraction or link analysis.

Personalization. At the delivery end of the chain, personalization is a capability that holds much promise. Its roots can certainly be seen in early day Internet fads around "push" and "recommendation" technologies, though clearly the drivers of those fads were related more to driving commerce or attention to Internet commerce or publishing efforts. Within enterprise knowledge management, the point of personalization is to provide information that is relevant to the individual on an ongoing basis. In essence, it is a flip of the kinds of techniques used in Search and Categorization from push to pull, and from specific projects at a given moment in time to general patterns of interest over time.

Information Visualization. Visualization capabilities provide an enhancement to the standard graphical user interface as provided by web browsers or desktop environments. Content Visualization can provide a map of large collections of content that help user both understand the entire collection as well as navigate to specific areas of interest. Common techniques include "Content Terrains Maps" which represent content collections as landscapes based on topical features, and "Wide Widgets" which provide visual/interactive "racks" for information structures like tables and hierarchies.

Emerging Information Access Infrastructure

In the next few years, enterprises will look to add "unstructured data management" capabilities to their deployed infrastructure whether it is based on portal frameworks, document or content management systems, or basic enterprise web or application servers. Unstructured Data Management vendors offer one or more of these capabilities as servers that can be rapidly deployed against the basic content/document infrastructure.

Meanwhile, vendors of document and content infrastructure, as well as the gorillas of enterprise systems including IBM, Microsoft, Oracle, and SAP, will start to offer these capabilities often by acquiring or reselling the capabilities of the companies that have proven capabilities in these areas.

About the Author

Ramana Rao is the CTO and Founder of Inxight Software and the editor of the monthly email newsletter, Information Flow. Past issues and other articles and papers can be found on www.ramanarao.com and www.inxight.com.