AberdeenGroup

Finding Relevant Information Requires a Lot More Than Search

An Executive White Paper

August 2003

Aberdeen Group, Inc.
260 Franklin Street
Boston, Massachusetts 02110-3112 USA
Telephone: 617 723 7890
Fax: 617 723 7897
www.aberdeen.com

Finding Relevant Information Requires a Lot More Than Search

Preface

Over the past 25 years, enterprises have become much better at extracting information from databases. Data warehouses, online analytical processing (OLAP), analytical applications, and executive dashboards are among the many mechanisms that companies now use to monitor and manage the health of the organization.

Unfortunately, enterprises trying to mine their unstructured data with equal effectiveness have been stymied until recently — despite the fact that the vast majority of a corporation's knowledge capital is stored in memos, articles, and e-mails. Admittedly, keyword search has been of some help, but fails when the technology cannot decide whether "jaguar" is an animal, a car, or a sports team, or does not recognize that "International Business Machines" and "IBM" are one and the same.

What has historically been a frustration with search is now turning into a crisis — as enterprises continually work to increase their productivity, they are recognizing that they can no longer afford to "forget" what they already know. In addition, various constituencies are now holding enterprises to a higher "knowledge retrieval" standard than was tolerated in the past:

- *Stakeholders* — Business owners and managers, having ruthlessly improved the effective use of physical assets (e.g., production lines and factories), are now turning their attention to improving returns on knowledge assets such as patents. IBM, for example, earns more than $1 billion a year in patent sales and royalties, while pharmaceutical companies race to discover the next drug worth billions of dollars in annual sales. An enterprise's consistent ability to leverage the appropriate patent or article can now have a significant impact on the bottom line.
- *Customers* — Customers addicted to the quick response time of the Web get frustrated when they (1) cannot find an answer to their problem on the Web site or (2) get varying answers from the different customer support representatives (CSRs) that they come in contact with. Improving the ability of customers and CSRs to mine a knowledge base can significantly improve customer satisfaction and decrease support costs. For example, when Gateway, the PC manufacturer, replaced keyword search with a natural language search solution on its Web site, it saw online resolution rates increase by 37%. In addition, Aberdeen conservatively estimates that deflecting support phone calls by answering questions on the Web is saving Gateway $480,000 a month.
- *Regulators* — In the Sarbanes-Oxley Act of 2002, the U.S. Congress amended the Securities Exchange Act of 1934 to require that regulated companies disclose material changes in their financial condition or operations rapidly and in plain English. Thus, major business discontinui-

> ties documented in e-mails or memos must swiftly bubble up to corporate management so that they can decide if they must notify authorities and the public. A corporation's failure to act on what it "knows" somewhere within the organization can now literally be a crime.

This Aberdeen *Executive White Paper* discusses how enterprises can strengthen their corporate memory. It investigates the difficulties inherent in finding relevant information, outlines the infrastructure necessary to remedy the situation, and describes one vendor's response to this information retrieval problem.

Yesterday's Information Matchmaker — The Corporate Librarian

Just a short decade ago, well-heeled corporations optimized the information retrieval process by funding corporate libraries. The corporate librarian — armed with a Master's in Library Science, an understanding of the company's business, and personal knowledge of employees' interests and tasks — served as an information matchmaker. This subtly — but significantly — accelerated information retrieval. A librarian's casual aside of, "Oh, I heard you're working on Project Excalibur — you should read these three articles that just came in," would save literally hours of research time. The library also fostered information matchmaking by being a physical place — employees could sit in comfortable chairs and browse through the latest journals, as well as mingle and exchange tidbits about articles and experts.

This is not to say the corporate library was a perfect mechanism. If the requested information was not a high-enough corporate priority to involve the librarian, workers made do by rummaging through the library themselves, making a few phone calls, or sometimes making a decision without the necessary information.

Today's Information Matchmaker — Technology

Because of the growth of the Web and tight corporate budgets, many enterprises have, over time, cut back or abolished their corporate libraries. Non-librarians — or more specifically, anyone with a PC and a Web browser — are today's researchers. The human form of the information matchmaker — the corporate librarian — has either been removed from the equation or evolved into an information facilitator with a much broader constituency. The personal touch has been diminished, with the result that although enterprises can ask many more questions, they are not always answered. When a user queries Google for information, the search engine recommends thousands of articles, rather than just the three most relevant articles. This deluge of information means that workers must still make do. Rather than not asking for the information, now they just ignore it. Ten years have passed, businesses have replaced the human touch with the technological touch, but they are *still* not using available information to its fullest potential.

260 Franklin Street
Boston, Massachusetts 02110-3112
Telephone: 617 723 7890
Fax: 617 723 7897
www.aberdeen.com

Must corporations resign themselves to being ignorant, just in a different way? The answer is no. But to recreate the librarian's knowledgeable touch, companies need to depend on subtle technologies, mechanisms that summarize, find, and suggest relevant information in much the same way that the librarian did, as well as take into account the various ways that people explore and search for information.

Requirement One: Relevance

The number one requirement of any information retrieval solution — whether based on human or technological means — is to find *relevant* information. Relevance is in the eye of the beholder — an article that one worker finds vital can be completely uninteresting to the next. Therefore, understanding a user's interests, vocabulary, and tasks is crucial to matching content with users. If understanding the user is not challenging enough, the solution must also not be misled by language ambiguity. For example, "apple" can mean a company or a fruit, and users who query for "car" may be interested in the affiliated concepts of truck and vehicle. Librarians understood that these personalization and translation tasks were part of their job; a technological solution must be built with the same attitude.

Requirement Two: Supporting Three Ways to Search for Information

A technological solution must also be process-friendly. Finding relevant information is very much a process — and a complicated one at that. People use a mix of different strategies, depending on whether they are a domain expert or neophyte, their personal preferences, and the amount of time they have. Search strategies fall into three main categories: shortcutting, wandering, and navigating.

Shortcutting

In this case, the users know exactly what they are looking for, and they want to go right to it. Domain experts with a deep understanding of their areas' vocabulary and sources will typically use this method to retrieve information quickly. Scientists who know a journal article's author and publication date or CSRs who have memorized a bug report number are examples of these types of users.

Wandering

However, not everyone is an expert, and even experts had to train themselves initially. A much more roundabout approach to finding information is to wander about in it — to peruse the taxonomy (or hierarchical organization) of the content, to look at some abstracts here, and to sample an article there. At times, the users do not know what they are looking for — but they will know it when they see it. The emphasis is not so much on finding something specific, but rather on becoming familiar with the area of interest: How is it organized? Who are the experts? What are the best sources?

260 Franklin Street
Boston, Massachusetts 02110-3112
Telephone: 617 723 7890
Fax: 617 723 7897
www.aberdeen.com

Navigating
Navigating is a cross between shortcutting and wandering — the user may not know the unique identifier of a piece of content, and so cannot jump to it, but at the same time has a good idea of the appropriate "information neighborhood." Therefore, navigating down a content hierarchy makes a lot of sense. This drilling down into the content offers relative speed while also allowing the user to see what other content is available.

These three different strategies are frequently mixed and matched, and they can blend into each other. An expert who typically uses shortcutting may put aside some time during the day to wander among the content as a way to see what is new in the subject area. Another user may take a shortcut to a specific point in the subject hierarchy and then use navigation to drill deeper into the content. It is important to note that any information retrieval system that does not take this infinite variability into account is ultimately hampering the ability of its users to work effectively.

The Required Mechanisms for Information Matchmaking
The resultant challenge then is to create a technological infrastructure that enables these human search strategies as a way to find relevant information. This task is a tall order that a system can fulfill only if certain building blocks are in place: content profiles, user profiles, and mechanisms that link them together, both among themselves and between each other (Figure 1).

Content Profiles
A content profile lists an article's concepts. A system may demand that an editor tag each article individually, offer a categorization engine that does the task automatically, or offer a hybrid solution that attempts to tag content and highlights any troublesome articles for human intervention. In some cases, systems can perform entity extraction, which is the recognition of company and personal names within the content. When that occurs, the tagging becomes more sophisticated, as the system recognizes that "Apple" refers to a computer company. Whatever the mechanism used, an article on Tiger Woods, for example, could be tagged as an article on Tiger Woods, sports, golf, a multimillionaire, and a Stanford graduate.

User Profiles
A user profile is a distillation of a user's tasks and interests. These attributes can either be explicitly declared by the user — "I'm interested in all articles on HP; please send them to me when they come in" — or can be inferred from the user's query patterns or place in the corporate hierarchy. This user-profiling capability is crucial for improving relevance. For example, if a user is looking for the term "ATM," knowing whether the user is a banker or a network engineer would help

260 Franklin Street
Boston, Massachusetts 02110-3112
Telephone: 617 723 7890
Fax: 617 723 7897
www.aberdeen.com

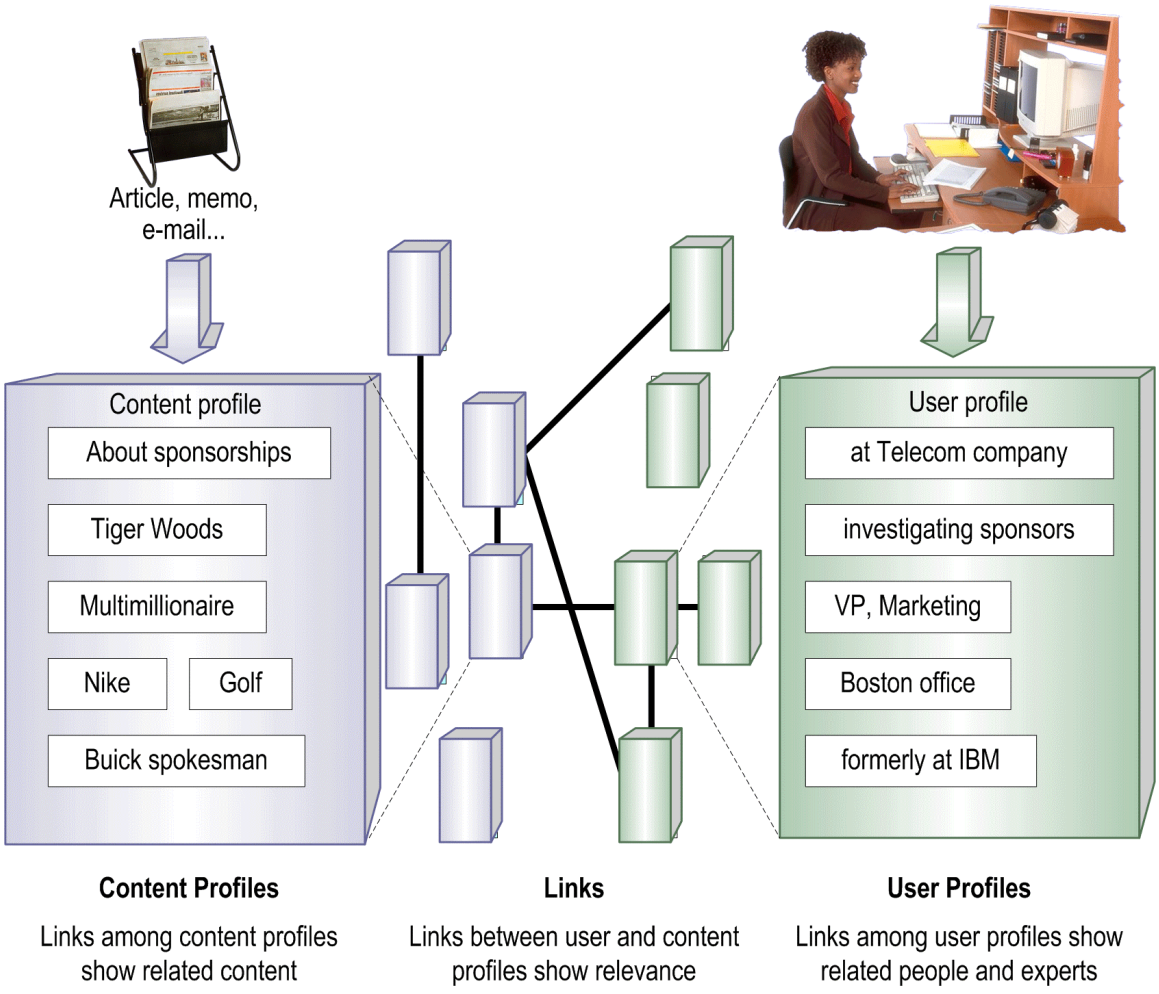
the engine decide whether to lead with articles on automatic teller machines or asynchronous transfer mode.

Linking Content to Users, Users to Users, and Content to Content

Content and user profiles, however, unleash their value only when interrelationships between them are declared. These logical links can take various forms. For example, links among user profiles can lead to a grouping of experts or employees working on a project together. Links among content profiles can map to a taxonomy (enabling a navigation process) or create a list of related content.

This linking can also help the system understand context — for example, it may discover that one set of articles is read by experts, while another set is read by

Figure 1: Required: Content Profiles, User Profiles, and Affiliated Links



Source: Aberdeen Group, August 2003

neophytes. By exploiting this understanding, a system can go beyond suggesting articles based solely on concepts. For example, if an expert reads a set of articles, other experts with similar user profiles may find those articles relevant as well.

Inxight's Solution to the Information Retrieval Problem: SmartDiscovery™
One company that has used these logical building blocks to create an information retrieval solution is Inxight Software, Inc., of Sunnyvale, CA. Inxight's SmartDiscovery software leverages the company's 20-plus years of research in natural language processing (originally as part of Xerox Palo Alto Research Center) to enable corporations to easily search, summarize, categorize, mine, and visualize content. Its capabilities are described below.

Document Decomposition via Linguistic Analysis
SmartDiscovery utilizes a natural language processing platform to analyze text in more than 20 major business languages and perform functions such as splitting compound words into their component parts (de-compounding), identifying noun phrases, and locating sentences and paragraphs. This ability to decompose a document into its component parts is also displayed in the software's ability to perform entity extraction — that is, identify and index entities such as people, companies, places, and dates. By identifying the multiple objects within unstructured content, this software makes it possible to later link them together in meaningful ways — to identify the skeleton of a document, as it were, as well as identify all the documents that mention HP, for example.

Document Profiling via Categorization and Taxonomy Development
SmartDiscovery can profile documents by summarizing and categorizing them. It can also help enterprises create their own taxonomies. Users can define which category or categories a document should fall into by using any combination of representative words and phrases, sample documents that reflect the category's meaning, and rules about appearance (or lack thereof) of words or phrases in a document. Users can also explicitly include or exclude a given document from a specific category. These mechanisms, leveraging natural language processing in a number of cases, enable users to concentrate on navigating the content, rather than thinking of how to adjust so that the system will better understand their query. For example, when presented with a single word query on "bugs," the SmartDiscovery software will generate a vastly different list of relevant documents depending on whether the query is coming from the context of "farming" or "computing" — a way of productively resolving language ambiguity that keyword search is incapable of.

Information Visualization and Navigation
SmartDiscovery also includes visualization technology to help knowledge workers view the resultant information. The software's Star Tree mechanism generates eas-

260 Franklin Street
Boston, Massachusetts 02110-3112
Telephone: 617 723 7890
Fax: 617 723 7897
www.aberdeen.com

ily navigated graphs representing the structure of the information. Boxes represent documents or categories of documents, and lines represent relationships between documents — a metaphor somewhat similar to an organizational chart. However, unlike an org chart, the hierarchy dynamically rearranges itself. When the user clicks on a document box, it moves to the center of the graph, and the affiliated lines and boxes rearrange themselves accordingly. Star Tree helps users understand large collections of information, and it is especially useful when workers are wandering or navigating to the relevant information.

Mapping to the Information Matchmaking Requirements

Inxight's SmartDiscovery offers key technologies required for information matchmaking: content profiling and linking mechanisms. The software's linguistic analysis, entity extraction, summarization, taxonomy, and categorization capabilities are all focused on content profiling — that is, understanding a document's structure, meaning, and place in the information universe. This deep understanding of a document's essence then enables Inxight to link them together, making it easier for users to filter and retrieve information from a list of highly relevant documents.

Inxight also supports the three main ways of searching for information. Sophisticated document tagging and summarization enable shortcutting; that is, letting users get to the relevant document quickly. Inxight's visualization, taxonomy, and categorization capabilities support users who would rather wander or navigate to the information. In summary, the company's solutions make it easy for users to find relevant information *their* way.

Aberdeen Conclusions

At a visceral level, ubiquitous Web search has made businesses think that all that is required when searching for information is a search engine text box. Put another way, the beguiling simplicity of the user interface sometimes makes companies forget that a sophisticated behind-the-scenes infrastructure is required to make searching look easy.

Unfortunately, it often takes several information retrieval project failures before enterprises realize this fact. Only after companies have failed at manually categorizing reams of data or have become frustrated at trying to understand how content is interrelated by scrolling down long lists, do they realize that relatively arcane technologies such as automated categorization, taxonomy building, and "visual content maps" can offer significant business value.

The amount of delivered value varies widely among firms. For single-location companies that have a small set of content, content that does change quickly, or few knowledge workers, such search infrastructure is overkill. However, for dispersed enterprises that maintain large, dynamic, and valuable content repositories, a sophisticated search infrastructure is part of the cost of doing business.

Telephone: 617 723 7890
260 Franklin Street
Fax: 617 723 7897
Boston, Massachusetts 02110-3112
www.aberdeen.com

Examples include multinational manufacturers, pharmaceutical companies, law firms, and large consulting practices — companies that, not surprisingly, were among the first to hire corporate librarians.

It was almost a century ago that a small group of corporate and other specialized librarians founded the Special Libraries Association — proof that businesses have depended on information matchmakers for a long time. As today's volume and speed of information, as well as the number of eager information consumers, overwhelm the librarian, savvy information-rich enterprises are responding by shifting toward information retrieval technologies — a notable example being Inxight's — that mimic the human touch of the librarian. By using categorization and taxonomies, increasing their accuracy through entity extraction, and making content relationships visible through visualization tools, these corporations are delivering relevant information to their employees whether the workers are shortcutting, wandering, or navigating to the information.

These companies know that remembering and reusing what they already know is not an accident, but instead a core competency that must be continually nurtured, especially in this high stakes environment full of impatient stockholders, demanding customers, and vigilant regulators. The result is a corporation that responds to customers quickly, trains its employees rapidly, reacts to business threats with dispatch, and notifies regulators promptly — all strategic advantages in today's highly competitive and regulated marketplace.

To provide us with your feedback on this research, please go to www.aberdeen.com/feedback.

Aberdeen Group, Inc.
260 Franklin Street
Boston, Massachusetts
02110-3112
USA

Telephone: 617 723 7890
Fax: 617 723 7897
www.aberdeen.com

Aberdeen Group is a computer and communications research and consulting organization closely monitoring enterprise-user needs, technological changes, and market developments.

Based on a comprehensive analytical framework, Aberdeen provides fresh insights into the future of computing and networking and the implications for users and the industry.

Aberdeen Group performs projects for a select group of domestic and international clients requiring strategic and tactical advice and hard answers on how to manage computer and communications technology. This document is the result of research performed by Aberdeen Group, which was underwritten by Inxight Software, Inc. Aberdeen Group believes its findings are objective and represent the best analysis available at the time of publication.