Why is it that searching an intranet is so much harder than searching the Web?

RAJAT MUKHERJEE AND
JIANCHANG MAO, VERITY

The last decade has witnessed the growth of information retrieval from a boutique discipline in information and library science to an everyday experience for billions of people around the world. This revolution has been driven in large measure by the Internet, with vendors focused on search and navigation of Web resources and Web content management. Simultaneously, enterprises have invested in networking all of their information together—to the point where it is increasingly possible for employees to have a single window into the enterprise. Although these employees seek Web-like experiences in the enterprise, the Internet and enterprise domains differ fundamentally in the nature of the content, user behavior, and economic motivations.

Enterprise Search: TOUGH STUFF

Enterprise Search: TOUGH STUFF

Our principal focus here is on outlining the demands on information retrieval in enterprises and various technologies that are employed in an enterprise content infrastructure. We define an enterprise to mean any collaborative effort involving proprietary information, whether commercial, academic, governmental, or nonprofit. The term *search* is usually used to mean *keyword search*. In this article, we use a broader definition that encompasses advanced search capabilities, navigation, and information discovery.

The overwhelming majority of information in an enterprise is unstructured—that is, it is not resident in relational databases that tabulate the data and transactions occurring throughout the enterprise. This unstructured information exists in the form of HTML pages, documents in proprietary formats, and forms (e.g., paper and media objects). Together with information in relational and proprietary databases, these documents constitute the enterprise information ecosystem.

Web Search Experience (Google)

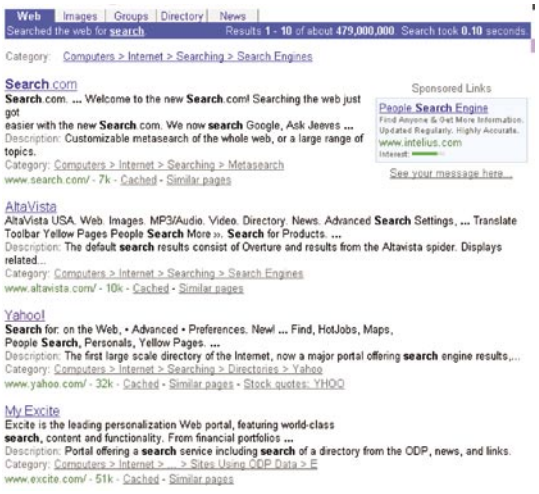


FIG 1

Arguably, it is the structured information in an enterprise that is the most valuable; enterprises thus seek to enhance the value of their unstructured information by adding structure to it. Creating, aggregating, capturing, managing, retrieving, and delivering this information are core elements in an enterprise content infrastructure. Enterprise information delivery must clearly meet the performance that users have come to expect on the Internet. While some techniques for scaling and performance developed on the Web can be adapted to the enterprise, many techniques for searching, organizing, and mining information on the Web are less applicable to the enterprise.

ENTERPRISE VERSUS INTERNET SEARCH

Enterprise search differs from Internet search in many ways.[1,2,3] First, the notion of a "good" answer to a query is quite different. On the Internet, it is vaguely defined. Because a large number of documents are typically relevant to a query, a user is often looking for the "best" or most relevant document. On an intranet, the notion of a "good" answer is often defined as the "right" answer. Users might know or have previously seen the specific document(s) that they are looking for. A large fraction of queries tend to have a small set of correct answers (often unique, as in "I forgot my Unix password"), and the answers may not have special characteristics. The correct answer is not necessarily the most "popular" document, which largely determines the "best" answer on the Internet. Finding the right answer is often more difficult than finding the best answer.

Second, the social forces behind the creation of Internet and intranet content are quite different.[1] The Internet reflects the collective voice of many authors who are free to publish content, whereas an intranet generally reflects the view of the entity that it serves. Intranet content is created for disseminating information, rather than attracting and holding the attention of any specific group of users. There is no incentive for content creation, and all users may not have permission to publish content.

Content from heterogeneous repositories—for example, e-mail systems and content management systems—typically do not cross-reference each other via hyperlinks. Therefore, the link structure on an intranet is very different from the one on the Internet. For example, on the Internet the *strongly connected component* (pages that can reach each other by following the links) accounts for roughly 30 percent of crawled pages. On corporate intranets, this number is much smaller (10 percent

on IBM's intranet, for example[1]). The popular PageRank[4] and HITS[5] algorithms are thus not as effective on an intranet as on the Internet.[6] Other techniques have to be employed to improve search relevance on an intranet.

Enterprise content and processes have different characteristics that make information retrieval within the enterprise substantially different from Web search. This in itself has caused enterprise search to evolve differently from Internet search. (See the sidebar, "Enterprise Characteristics.") The different needs result in dramatically different experiences on the Internet (figure 1) and, say, a corporate intranet (figure 2).

Deployment environments for these domains also differ: an Internet search engine, including hardware and software, is fully controlled and managed by one organization as a *service*. Enterprise search software is licensed to and deployed by a variety of organizations in diverse environments. This imposes varying requirements: hardware constraints, software platforms, bandwidth, firewalls, heterogeneous content repositories, security models, document formats, user communities, interfaces, and geographic distribution. Enterprise search software needs—high flexibility/configurability and ease of use with ease of deployment—are often at odds with each other.

Though a search service can incorporate new technologies in quick cycles, this is often not the case for enterprise deployments. Economic and time constraints in enterprises sometimes prevent quick upgrade cycles. Often, enterprises use old software versions, although they are fully cognizant that they are not employing certain technologies that they could. This is more pronounced in cases where search software is embedded in third-party enterprise applications with extended release cycles. This sometimes leads to the end users being dissatisfied with the quality of search provided within the enterprise. A search service, however, cannot effectively be bundled into an enterprise offering, since enterprises demand flexibility, security, and custom application access.

Sample User Experience on an Intranet

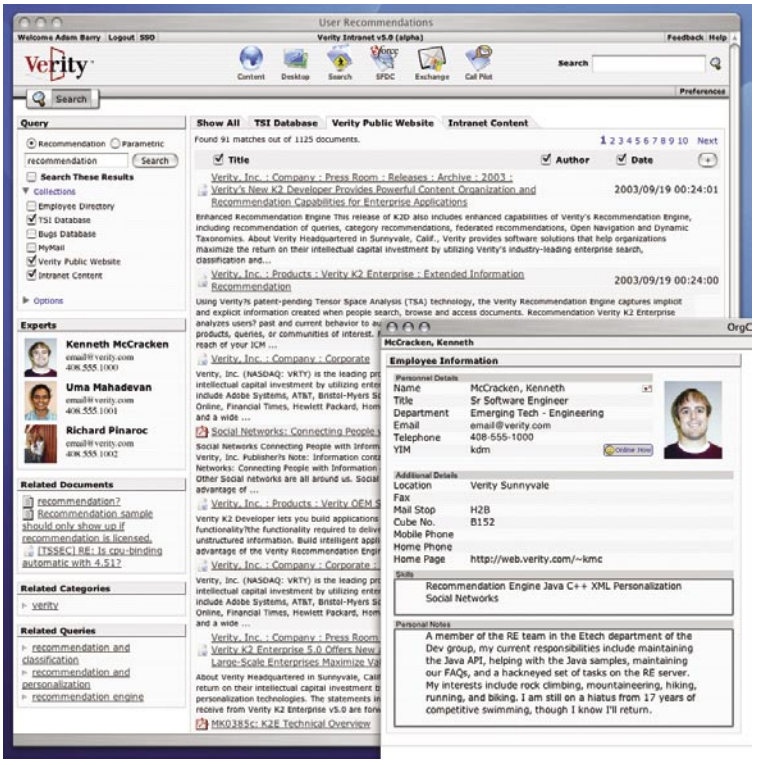


FIG 2

TECHNOLOGIES

Specific techniques can improve enterprise search. Figure 3 depicts key ingredients of an enterprise search system.

Spidering and Indexing. Data must be accumulated (spidered) and indexed before it can be searched. This requires knowledge of where the critical information exists, and access to these repositories, which can be secure. Many current spiders run on predefined schedules that do not match the rates at which information is changing. Adaptive refresh of indexes is required, which involves more sophisticated change-detection mechanisms. Most spiders use a pull model, which is harsh on network resources and on the target repositories. Future spiders will take greater advantage of triggers and targeted

Enterprise Search: TOUGH STUFF

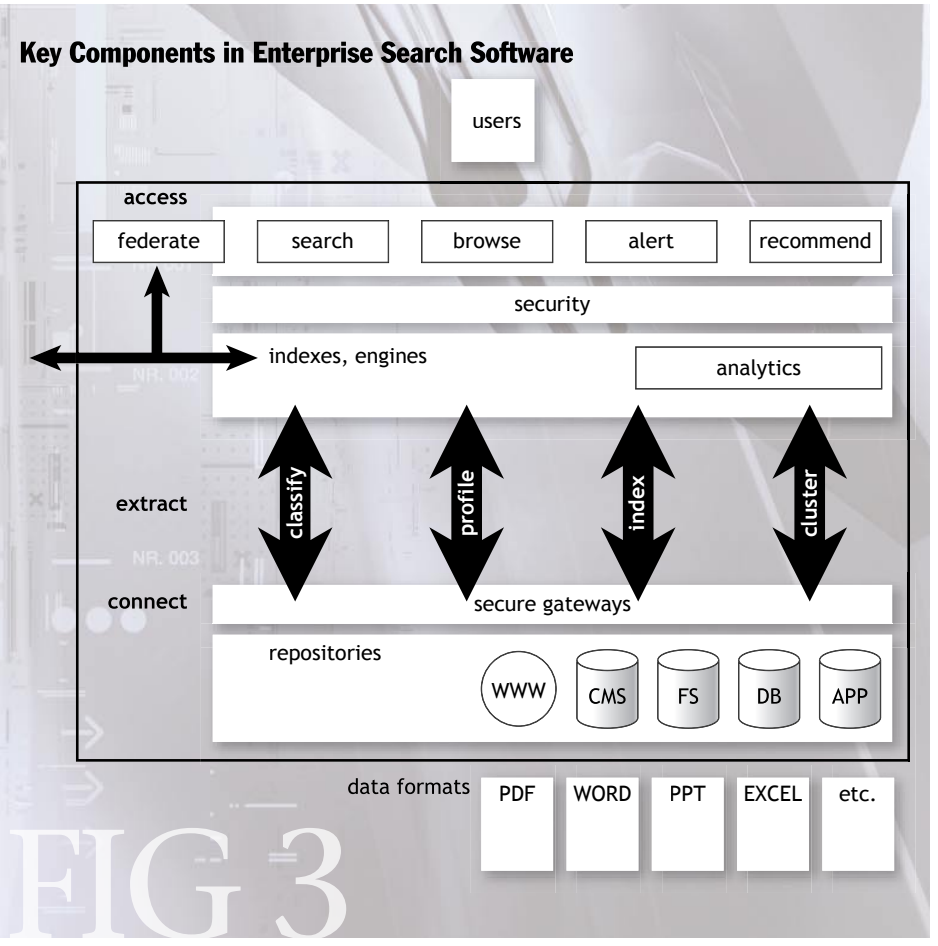
crawling. One issue faced here is that most applications do not expose information about what has changed; for example, they are not designed for exploiting external search technology, since they often build search into the application. Adoption of search standards by application vendors can help solve this problem.

Documents in multiple languages can reside in the same index, and techniques for automatic language detection can be used for language-based content routing and partitioning. Indexes are already taking into account information on hyperlinks, anchor-text, etc. Metadata will automatically be extracted during indexing to improve retrieval quality. Application and content management vendors will do well to flag content that has been modified to eliminate unnecessary reindexing.

Data Filtering. One of the keys to quality is to weed out information that is dated, irrelevant, or duplicated. Clean data implies better search relevance. Moreover, automatic classification, feature extraction, and clustering technologies will be more accurate when the data presented to them is cleaned up by a pre-processor. Techniques such as link-density analysis can be used to detect the differences between content and link-rich menus on Web pages. Entity-extraction techniques can be used to add relevant information before indexing occurs. Stripping out advertisements, menus, and so forth improves the quality of the subsequent ranking algorithms that operate on the content. This is important for enterprises that are indexing external content.

Search Relevance. Certain Internet search strategies, such as hyperlink analysis, cannot be carried over directly to the enterprise. Some strategies are actually being abused on the Internet; for example, Internet search engines are constantly tuned to offset the effects of spam and the doctoring of Web pages to take advantage of search algorithms. Other characteristics of enterprise content must be exploited to achieve higher search relevance. For example, since intranets are essentially spam-free (because of the lack of incentives for spamming), anchor-text and title words are reliable sources of information for ranking documents. The rank-aggregation approach proposed by Fagin et al provides an effective way of combining ranks derived from separate sources of information.[1]

Many intranet queries (60 to 80 percent) are targeted to retrieve "stuff I've seen." A user may remember some attributes of the target results, such as date or author. Search engines must provide a way of specifying attributes in conjunction with the



query. Sorting on an attribute also helps to locate information quickly.

User role and context can improve the relevance of search. Session-based personalization techniques are already available in enterprise search software from leading vendors. Whereas a public Web site gathering user information can flag privacy concerns, users in enterprises have far fewer concerns if their access—for example, clickstreams on the intranet—is tracked, because it is likely business related. In some enterprises—for example, in the finance industry—even IM (instant messaging) is fair game for regulatory reasons.

The perceived relevance of a result can be dramatically changed by providing better titles (using techniques to create titles automatically if none exists), dynamic summaries, category information, and so forth. Usability is tightly coupled to relevance.

Structured versus Unstructured Information. In product catalogs, each item has unstructured text, as well as structured attributes. For example, an automobile typically has a description and attributes, such as year, model, and price. A typical query is a conjunction of an arbitrary

Enterprise Characteristics

Diversity of Content Sources and Formats
Enterprises must ingest and extract structured and unstructured information from heterogeneous content sources, for example, Microsoft Exchange, Lotus Notes, Documentum, as well as file systems and intranets. Furthermore, documents exist in a myriad of file formats and several languages; a single document could contain multiple languages, or attachments with multiple MIME-types. At this time, less than 10 percent of enterprise content by volume is HTML.

Secure Access
An individual's role in an enterprise dictates what documents can be accessed. Sophisticated enterprises demand a more stringent notion of security in which search result lists are filtered to display only the documents accessible to the user. Doing this in conjunction with the native security of the repositories is a particularly difficult challenge.

Combined Structured and Unstructured Search
Information that is considered to be unstructured is in fact semi-structured, with metadata such as author, title, date, size, and so forth. Conversely, much structured information in an RDBMS (relational database management system) is unstructured—for example, blobs of text and VARCHAR fields. XML is ubiquitous in content and applications. It is essential to provide high-performance parametric search that allows the user to navigate information through a flexible combination of structured and unstructured data.

Flexible Scoring and Ranking Mechanisms
No single scoring and ranking function will work for all enterprise search contexts. Many of the powerful link-based scoring and ranking algorithms that have been honed for the Web are unlikely to be germane to the enterprise. Enterprise content is fundamentally different from Web content, enterprise users are different in their goals and expectations from Web users, and enterprise search imposes layers of complexity that the Web lacks.

Federated and Peered Results
Federated search enables a single point of access to multiple sources (internal indices, Web search, and subscription sources—for example, realtime newsfeeds). The key challenge here is to merge sets of results from all sources for unified presentation. This must be done even though the sets typically have no documents in common and employ different scoring and ranking schemes.

Content Generation Processes
While the Internet tends to grow democratically, intranets are often governed by bureaucracies. Content creation on an intranet is normally centralized to a small number of people. While content published on an intranet may need to comply with specific policies (reviews, approvals), consistency is not guaranteed, since there may be multiple organizational units whose policies differ.

People/Roles/Behaviors
It is well known that some of the most valuable knowledge in an enterprise resides in the minds of its employees. Enterprises must combine digital information with the knowledge and experience of employees. An important distinction between the enterprise and the Internet is that while Internet users are anonymous for the most part, enterprise users are answerable and guided by specific controllable processes. Privacy issues are also very different in an enterprise, since people are usually engaged in enterprise-specific behavior and are being compensated for their engagement.

Enterprise Search: TOUGH STUFF

text query ("Leather Trim" AND "All-wheel Drive") and a parametric query on structured fields (Manufacturer = Toyota AND price < $30,000 AND year > 2000). Whereas the *text query* is within the purview of classic information retrieval systems, the *parametric query* is traditionally handled using relational database systems. Modern search tools perform both functions for applications such as e-commerce and marketplaces where scalability and cost-performance are critical.

Using an RDBMS (relational database management system) to solve the problem would result in unacceptably poor query responses. Text search extensions in RDBMSs do not support powerful free text query capabilities—for example, fuzzy search—and are not cost effective for search.[7] In addition to being able to sort along attribute values, it is crucial to be able to rank the results based on the query. This enables efficient *guided* navigation of results, allowing the user to progressively refine (or relax) the query.

The information explosion has reached the point where information architects often lack an adequate grasp of all the themes and topics represented in the corpus.

Parametric refinement[8] provides a solution to this problem by augmenting a full-text index with an auxiliary parametric index that allows for fast search, navigation, and ranking of query results. The main issue with this powerful technology is that data preparation is very important, and organizations need to invest time in augmenting and normalizing the data. Classification and entity-extraction techniques will be used to augment information with attributes for improved search and navigation.

Another key characteristic of data is structure within the data itself. With the increased adoption of XML, the ability to search and retrieve specific portions of documents—for example, specific elements in XML—is mandatory. Query semantics like XQuery will be supported, but with the added ability to handle unstructured text and fuzzy constructs that databases do not handle elegantly (e.g., spelling errors). The ability to dynamically construct virtual documents that can consist of relevant portions of many documents will be critical. What the end user will want in the future is not just a matching document, but something that represents an answer.

Classification, Clustering, and Taxonomy Navigation. Search provides an efficient way of finding relevant information from huge amounts of data only if the users know what to search for. With large corpora, queries can have large result sets. It is important to facilitate users in forming effective queries through browsing and navigating information in a relatively small, manageable space. Examples of such spaces are *taxonomies*. They organize documents into navigable structures to assist users in finding relevant information. Searches within a category typically produce higher relevance results than unscoped searches. Research has shown that presenting results in categories provides better usability.[9]

Most taxonomies are built and maintained by humans because domain expertise is required. Well-known examples include the directory structures of Yahoo! and the Open Directory Project. Manual taxonomy construction is time consuming and expensive, however. Further, in many enterprises, the information explosion has reached the point where information architects often lack an adequate grasp of all the themes and topics represented in the corpus. They need automated systems that first mine the corpus, extract key concepts, organize the concepts into a concept hierarchy,[10] and assign documents to it. Visualization tools are helpful here, rendering a *thematic map* of the concepts found and the relationships (parent–child, sibling, etc.) between them. A further desirable feature is to label these concepts with succinct human-readable labels. Finally, this idea can be extended to suggest a taxonomy that allows users to browse the corpus.

The state of the art in this area is still young and cannot be relied on to extract a taxonomy such as that produced by professional library scientists. It is, however, reasonable to expect systems to generate a strawman taxonomy for refinement by domain experts, making the domain experts dramatically more productive by automatically mining the patterns and discovering associations in the data.

Classification rules for assigning documents into categories can be either manually constructed by knowledge workers, or automatically learned from pre-classified

training data. Once these classification rules are defined, a document can be automatically assigned to categories from multiple taxonomies.[11] Navigation of a single taxonomy is limited at best, however, assuming disjointed nodes organized according to the taxonomy creator. Consider the topic "Sushi Restaurants in San Jose." Would we navigate the taxonomy tree to this node starting with the *Regional* heading, to *United States,* then *San Jose*, under which we would look for *Restaurants*? Or would we first select *Travel and Leisure*, then *Dining*, *Japanese Cuisine*, and, finally, expect to find restaurants classified by region? Ideally, the user should not be forced into making a choice at the top level of the tree. *Relational taxonomies* solve this problem, when classification metadata exists on multiple axes (e.g., regional *San Jose*, as well as functional *Sushi*). The system must be able to render this combined classification information into a navigational experience where users are not forced to make choices aligned with those of the taxonomy creator, but can instead navigate the way they want.

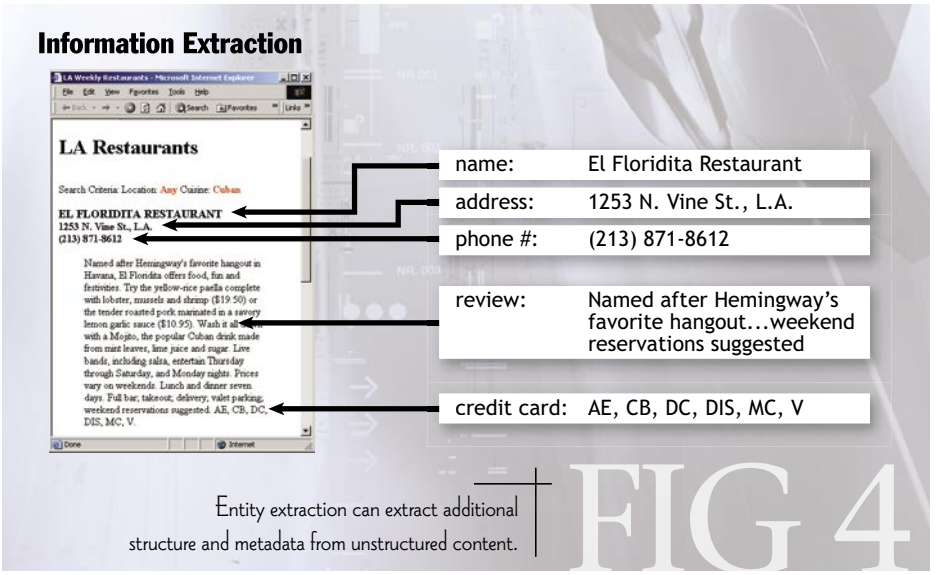
Although intranet search does not normally return millions of documents as in Internet search, a result set may contain a large number of documents. Sifting through a long list is very tedious. In this case, on-the-fly result list clustering is desirable to help users navigate the results. Realtime result list clustering organizes search results by query-dependent topics that are dynamically generated from search results.

Information Extraction and Text Mining. Metadata in semi-structured documents brings tremendous value to content search and organization. *Metatags* can relate to the document (subject, author) or can apply to in-line content (company, zip code, gene). Once documents are tagged, parametric search or OLAP (online analytical processing)-style analysis of multidimensional data sets is possible in order to reveal interesting details within a larger corpus. Subject matter experts can be hired to tag or annotate documents manually. Manual tagging, however, does not scale to large volumes of information; therefore, automation is mandatory.

Information extraction and text mining are useful tools for reducing tagging costs. Text mining uses linguistic, semantic, syntactic, statistical, and structural analysis to classify documents or extract persistent entities, facts, events, and their relationships. Linguistic analysis includes parsing, tokenization, and parts-of-speech tagging. Semantic analysis can disambiguate the meaning of polysemous words based on context. Syntactic approaches define entities as patterns that can be expressed as regular expressions or context-free grammars. Statistical analysis can be used to discover hidden patterns that are not easily expressed by human experts and to discover correlation between entities. Structural analysis can exploit proximity and layout information in order to link entities.

The effectiveness of information extraction and text mining depends on document quality and the homogeneity of the target information entities. In almost every application of automatic tagging, domain-specific heuristics will be employed to improve effectiveness.

Figure 4 shows an example of applying information extraction techniques to obtain a restaurant's name, address, phone number, review, and accepted credit cards from an online advertisement. Such automatic extraction can dramatically improve a yellow-pages application by enabling search and navigation of restaurants using specific directives—for example, cuisine, payment mode, and telephone area code—as well as by providing additional information, such as a map or driving directions.



Entity extraction can extract additional structure and metadata from unstructured content.

Federation. Information spans organizational boundaries. Not all the information required for a task is available as indexed content. Even if an organization has access

Enterprise Search: TOUGH STUFF

to relevant content (e.g., on the Web), there are cases where it cannot be indexed (e.g., security) or is forbidden from being indexed because of legal constraints. Further, in large organizations, different departments commonly index silos of information via different software systems or applications.[3]

In such cases, federated search is the only way to provide a single point of access to data from enterprise repositories and applications, as well as external subscription sources and realtime feeds. Sophisticated systems add further value via ranking, filtering, duplicate detection, dynamic classification, and realtime clustering of results from disparate sources that may not be under the jurisdiction of the enterprise.[12] Database vendors provide federation across disparate relational databases, but federated search for unstructured data provides different challenges (e.g., ranking across independent systems) and opportunities (e.g., classification and clustering).

Sources could include personal workstations, as in a peer network, requiring asynchronous and incremental behavior and scheduling support. For example, users can schedule an overnight search that provides a blended, filtered, and classified result set before their arrival in the morning.

Social Networks and Use-Based Relevance. An enormous proportion of an organization's intellectual capital resides as tacit knowledge. Social networks include the human element in the information ecosystem. Information usage patterns can be analyzed to discover the patent and latent relationships between the people in an organization and the documents they create, modify, access, search, and organize. Web search engines such as Google use the structure of hyperlinks between Web pages, which reflects relevance of Web assets to communities. This is a static approach. A system that analyzes communications patterns between people and the dynamic usage of information in the enterprise will deliver a richer personalized experience, based on a combination of content and context, to individuals and groups.

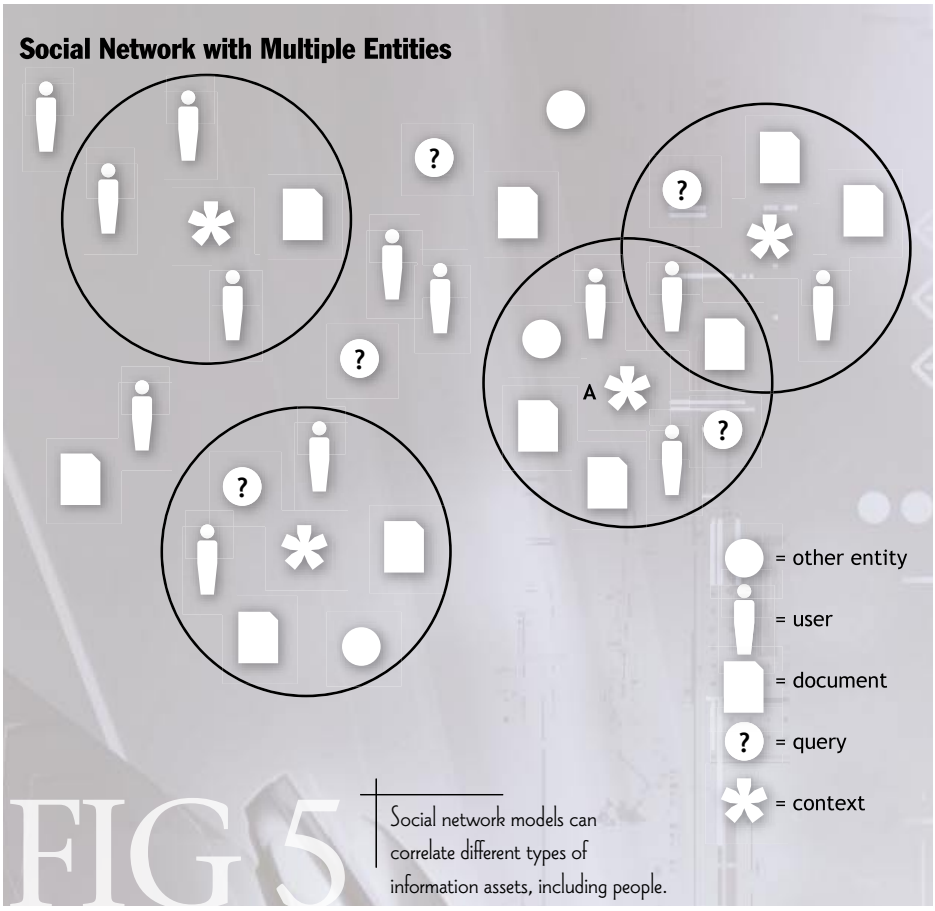


FIG 5 Social network models can correlate different types of information assets, including people.

Figure 5 depicts a social network, including multiple entities of different types (e.g., documents, queries, categories, and users). It is possible to represent these disparate entities using a consistent model (e.g., a set of keywords or feature vector). This enables detection of useful correlations among entities of different types. The input context can be a combination of multiple entities—for example, a user's profile, the input query, and the current document the user is browsing. Further, these representations can change based on user interactions, improving relevance over

time. Although many activities on the Internet are unreliable (e.g., spam), you can be more confident of interactions that occur within enterprises, allowing systems to take advantage of this vital input.

Traditionally, the scoring and ranking of documents is based on the content in each document matching the query. The social network can be exploited to augment content analysis with the historical behavior of users, changing result ranking. This *adaptive ranking* could be simplistic (boost a document's rank if a previous user elected to view/rate it after issuing the same search) or more sophisticated (boost rank if selected from the second results page for a similar query). Incorporating dynamic feedback also allows for the infusion of new terms to document representations, allowing relevant information to be returned for query terms that do not even exist in the content (*concept-based retrieval).*

User profiles can enhance the input context to provide personalization and targeted search. Persistent profiles can exploit a user's role in addition to historical patterns of access. A session-based profile provides realtime personalization, improving relevance to the current task. Such systems can exploit a user's queries, clickstreams, entry in the company directory, and so forth.

Both physical assets (e.g., documents, users) and virtual assets (e.g., categories, groups) can be included in the social network, facilitating information discovery relevant to the user's context. Such a system enables users to participate in the taxonomy building process, enabling personal and community taxonomies. The vicinity of the input can include experts within the organization. Social-network research suggests that portal users form into overlapping cliques of tightly knit communities. This drives the need for discovering communities of other users most germane to a user's current information context.

Analytics. Reporting and analytic modules supplied by the software can provide concrete metrics of search relevance and efficacy, and are powerful tools for improving user experience. These metrics can help to validate improvements in the search implementation—for example, adding dynamic query-based summaries to result lists, enabling user feedback, and creating synonym lists and predefined queries to avoid the dreaded "*No results found.*" Without reporting tools and analytics, measurements of relevance and user satisfaction are difficult to make. Analyzing what categories are being navigated and what documents are popular can be useful for organizations in evaluating their information needs and evolution.

THE OUTLOOK

Several opportunities exist for developing better enterprise search platforms, but certain challenges must be faced in exploiting these opportunities.

Algorithms for content-based search relevance will improve. However, exploiting user interactions to further tune system performance—for example, users rating documents or providing and updating their profiles—implies changes in user behavior. Providing incentives to employees to participate can result in different community behavior. Already, companies are being established to extract social networking benefits, such as contact list management. Cultural changes in organizations will facilitate the efficacy of such technologies.

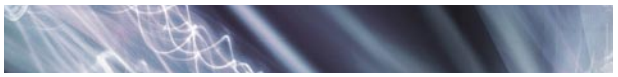
High-quality automated systems will be the norm of the future. Any system that depends on explicit human intervention (e.g., human tagging, annotation) without recourse to automation is destined to fail over the long term. As enterprise content increases, manual intervention will not scale. Automation implies that only a small percentage of content is siphoned off for human oversight, based on stringent thresholds. As algorithms get more accurate, an automated system can be more reliable and consistent than the collective output of multiple humans.

As the necessity for clean, organized content is recognized as crucial by more organizations, content publishing processes will become more stringent. Additional tools will be employed to capture information that is currently lost. Further, these tools will use newer technologies, such as XML, thereby enhancing consistency and automated processing. While data quality is likely to improve over time, however, we will still need to deal with noisy legacy data.

Studies of enterprise data show that important metadata in documents (e.g., author) is often incorrect, as it is set to some default value (e.g., organization name or template creator). Enforcement of correct metadata, either via technology assists or via policy, will go a long way toward improving data quality. Removal of redundant/obsolete data (as high as 20 to 30 percent) will benefit relevance. Techniques such as duplicate detection and near-duplicate detection can ensure that irrelevant data is eliminated from active corpora.

Internet search engines have become popular and clearly demonstrate the power of having information at one's fingertips. Enterprise search, while having similar desiderata, is faced with a different set of challenges. Besides having to deal with multiple heterogeneous repositories and myriad data formats, enterprises also

Enterprise Search: TOUGH STUFF



have to deal with security, compliance, and deployment issues. Many of these challenges can be addressed very effectively by technology.

While search relevance is an important yardstick, there are other key characteristics that make for effective search, such as navigation, classification, entity extraction, recommendation, summarization, query language, and semantics. Systems that incorporate user behavior will become the norm, yielding higher relevance, better personalization, and higher utilization of human assets and tacit information. Enterprise systems typically cannot compile the large-scale statistics that Internet search engines use to weed out noisy data, and other techniques will be employed to address the special needs of the enterprise. Q

REFERENCES

1. Fagin, R., Kumar, R., McCurley, K., Novak, J., Sivakumar, D., Tomlin, J. A., and Williamson, D. P. Searching the workspace Web. *Proceedings of the 12th International World Wide Web Conference*, Budapest, Hungary (May 2003) 366-375.
2. Raghavan, P. Structured and unstructured search in enterprises. *Data Engineering* 24, 4 (Dec. 2001) 15-18.
3. Hawking, D. Challenges in enterprise search. *Proceedings of the Australasian Database Conference*, Dunedin, New Zealand (Jan. 2004) 15-24.
4. Brin, S., and Page, L. The anatomy of a large-scale hypertextual Web search engine. *Proceedings of the 7th International World Wide Web Conference*, Brisbane, Australia (1998) 107-117.
5. Kleinberg, J. M. Authoritative sources in a hyperlinked environment. *Journal of the ACM* 46, 5 (1999) 604-632.
6. These algorithms are often exploited on the Internet via spam and page manipulation, making them less effective.
7. Many applications derive immediate ROI (return on investment) when offloading the search component from the database engine to enterprise search software.
8. Abrol, M., Latarche, N., Mahadevan, U., Mao, J., Mukherjee, R., Raghavan, P., Tourn, M., Wang, J., and Zhang, G. Navigating large-scale semi-structured data in business portals. *Proceedings of the 27th VLDB Conference*, Rome, Italy (2001) 663-666.
9. Dumais, S. T., Cutrell, E., and Chen, H. Optimizing search by showing results in context. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Seattle, Washington (March 2001) 277-284.
10. Chung, C., Lieu, R., Liu, J., Luk, A., Mao, J., Raghavan, P. Thematic mapping—from unstructured documents to taxonomies. *Proceedings of the Conference on Information and Knowledge Management*, McLean, Virginia (2002) 608-610.
11. Dumais, S. T., Platt, J., Heckerman, D., and Sahami, M. Inductive learning algorithms and representations for text categorization. *Proceedings of the Conference on Information and Knowledge Management*, Bethesda, Maryland (1998) 148-155.
12. Choo, K., Mukherjee, R., Smair, R., and Zhang, W. The Verity federated infrastructure. *Proceedings of the Conference on Information and Knowledge Management*, McLean, Virginia (2002) 621.

LOVE IT, HATE IT? LET US KNOW

feedback@acmqueue.com or www.acmqueue.com/forums

RAJAT MUKHERJEE is the principal software architect at Verity in Sunnyvale, California, working in the area of social networks. After completing his B.Tech. in electrical engineering at the Indian Institute of Technology, Madras, India, and his M.S. and Ph.D. in parallel computing at Rice University, Houston, Texas, he joined IBM's Thomas J. Watson Research Center. There he worked on clustered computing, scalability, and high-availability and scalable Web servers, and helped design infrastructures used during the 1996 Atlanta Olympics and the Deep Blue–Kasparov chess match. He moved to IBM's Almaden Research Center in 1997, where he worked on distributed digital libraries and content management. He then worked with Purpleyogi, a Silicon Valley content discovery startup, for a year before joining Verity.

JIANCHANG MAO is the principal software architect and manager in the Emerging Technologies Group at Verity. Prior to joining Verity, he worked at IBM's Almaden Research Center for more than six years. He earned his Ph.D. in computer science from Michigan State University in 1994. A senior member of IEEE, Mao has served as associate editor for *IEEE Transactions on Neural Networks* and as an editorial board member of *Pattern Analysis and Applications*. He received the Outstanding Technical Achievement Award and three Research Division Awards from IBM between 1996 and 2000.