Towards the next generation of enterprise search technology

by A. Z. Broder
A. C. Ciccolo

Unstructured information represents the vast majority of data collected and accessible to enterprises. Exploiting this information requires systems for managing and extracting knowledge from large collections of unstructured data and applications for discovering patterns and relationships. This paper elucidates the differences between search systems for the Web and those for enterprises, with an emphasis on the future of enterprise search systems. It also introduces the Unstructured Information Management Architecture (UIMA) and provides the context for the unstructured information management (UIM) papers that follow.

The Web revolution has exposed hundreds of millions of people to the experiences of searching and taxonomy browsing and has reshaped their expectations of the knowledge retrieval process, not only while browsing the Web, but more importantly, while at work, performing their jobs. Unfortunately, study after study shows that at the enterprise level, these expectations are not being met.[1] Knowledge management in the enterprise setting and even simple document search functions are often perceived as disappointing.

Why is this so? Search technology per se has made enormous strides. Web search engines can return excellent results on single-word queries of a 15-terabyte corpus, though this would have been considered impossible *in principle* not so long ago, regardless of computing power or computational cost. Furthermore, a number of techniques from natural language processing (NLP), such as information extraction, automatic identification of named entities (such as mentions of people, places, and organizations), the identification of relationships between entities, machine translation, and taxonomy generation and classification have been combined with classic search methods and have shown significant benefits. Automatic document categorization and classification became more accurate than human processing in the late 1990s and is now considered an essential means of organizing large corpora for knowledge management systems.[2] Automated summarization of documents based upon information extraction techniques has been demonstrated to improve search efficiency by supporting more focused examination of retrieved documents.[3] Finally, statistical machine translation, while still far below the capabilities of skilled human translators, may be good enough to support cross-lingual information retrieval on the Web or across enterprise document collections.[4] Given these results, there is growing confidence that many of these technologies may move from the status of cutting-edge research to commercial application in the near term. Although the computational demands of some of these technologies might be too high for application to the entire Web currently, this should be less of a problem in the enterprise, where the corpora are usually much smaller.

Thus, it seems that search and knowledge management in the enterprise should be improving and may indeed be easier than on the Web. The demand is there. The technologies are there. What is the missing part? Where is the problem? The answer lies in part in the essential differences between the public Web and the internal environment of the enterprise. One factor is that although enterprise corpora are smaller, they lack the highly hyperlinked nature of the Web, and thus some of the most successful techniques for the Web, based on link analysis, do not apply in the enterprise. This results in lower relevancy of retrieved documents. Another factor is that in the enterprise there are additional security, reliability, and performance issues that complicate the problem. A well-publicized example is the need to protect the privacy of individuals' personal data. The implications of this issue on search and text-analytic applications is a current popular research area, with legislatively mandated compliance monitoring eliciting heated debate, both pro and con.

Nonetheless, the most important factor is independent of the differences between the public Web and the enterprise, and rests on the fundamental character of the technologies. The advanced technologies described above, for the most part, simply do not work together easily or well. Typically, each one of these technologies has a completely different view of the world, represents the underlying documents in different ways, and is concerned with performance in different areas. This situation arises in part from the developers of technologies being "algorithm-centric." The computational requirements of these technologies are so great that their developers tend to engage in "programming-in-the-small," that is to say, building highly integrated, optimized, and hence closed and narrow applications based on their core technologies. To build systems to be used by consumers of information, rather than programmers, such narrow applications are usually awkwardly integrated, using ad hoc approaches. If there is any cooperation at all, it takes the form of one narrow application that consumes documents, performs its magic on their contents, and produces new documents as output. That output is then consumed by another narrow application, which starts by repeating much of the text parsing, tokenization, and so on, to convert the data to *its* representation. This process continues in subsequent stages, cascading inefficiency on inefficiency.

There is an alternative to the traditional process described above, one which capitalizes on the computational power of distributed systems. We submit that the "missing part" is the architecture that enables the integration of the technologies described above with search and retrieval. Such an architecture has been developed within IBM Research—namely, the Unstructured Information Management Architecture (UIMA). Various aspects of the UIMA, a software architecture for supporting the development, integration, and deployment of UIM technologies, are described in the first group of papers in this issue. This engineering foundation has been adopted by both IBM Research and the IBM Software Group as a delivery platform for advanced UIM technology.

The first paper, by Ferrucci and Lally, presents UIMA "by example." Starting from a high-level overview of the architecture, they take the reader through all the steps required to build a simple UIM application, and in the process, they highlight some of the major UIMA concepts and methodologies.

Götz and Suhre describe the design and implementation of the Common Analysis System (CAS), the subsystem of UIMA that provides data modeling, creation, and access. The CAS supports data modeling via a type system that is programming-language-independent and provides a powerful and portable indexing mechanism. In a sidebar, Marshall Schor delineates an effective approach to working with the CAS from within Java; Schor's approach has many desirable properties, including type safety, maintainability, readability, performance, and composability.

Turning to applications, Mack et al. present BioTeKS, a system for text analytics for life science using the UIMA platform. BioTeKS integrates research technologies from multiple IBM Research labs and is the first major application of the UIMA. The paper describes the system and some of its applications and highlights the role played by the UIMA framework in developing BioTeKS.

The second group of UIM papers in this issue presents research and applications that predate the wide adoption of the UIMA across IBM Research. Despite this, they exemplify the need for combining multiple tools and technologies to build high performance UIM applications. There is no doubt that such combinations will be greatly facilitated by the "plug-and-play" capabilities of the UIMA.

Uramoto et al. describe MedTAKMI, a system for knowledge discovery from biomedical documents. MedTAKMI is the first production text-mining sys-

tem in the world that can deal with the entire MEDLINE** database of abstracts. It consists of two main components: information extraction and relationship mining. In the preprocessing stage, keywords are extracted and categorized, and binary and ternary relationships are identified. At program execution, MedTAKMI uses this information to provide mining functions to users in an interactive manner.

Su et al. present an information portal for market intelligence management called MIP. One component of this system gathers daily market information from multiple sources: Web sites, file systems, mail servers, etc. A second component extracts and organizes information according to user-given requirements. Customized interaction with the user is enabled via a presentation server and a search and indexing component.

Kozakov et al. take on the difficult problem of creating a glossary from documents in a specialized technical field, when the terms presented may not be commonly used or found in general dictionaries. They focus on glossary extraction and utilization for the IBM Technical Support information search and delivery system, but their ideas have general applicability.

The final UIM paper in this issue, by Wolf et al., exemplifies some central issues in this area: how do we evaluate a UIM technology, and how do we choose the best approach to satisfy the users of our systems? Wolf et al. performed an evaluation of four methods for summarizing technical support documents, as used in an actual search system: programmatic sentence extraction summaries, summaries of terms highlighted in context (THIC), existing summaries of varying quality, and search of document titles only (that is, without additional summary text). It is notable that THIC summaries, although currently widely popular on the Web, do not represent the best approach for technical support documents in terms of task completion time. This raises the question as to whether better summarization techniques using deeper analytical methods and possibly user intent inferences might not yield more effective Web searches as well.

The field of UIM may come full circle: while the unstructured search paradigm on the Web exploded in the consumer sphere before being adopted in the enterprise, we believe that the combination of semantic and linguistic annotations with unstructured search will follow the more conventional path of first being developed in the enterprise sphere before becoming pervasive in the Web world. Regardless of the sequence of events, the advantages of these hybrid approaches are already evident.

**Trademark or registered trademark of United States National Library of Medicine.

Cited references

1. R. Mukherjee and J. Mao, "Enterprise Search: Tough Stuff," *Queue* **2,** No. 2, 36–46 (2004), http://doi.acm.org/10.1145/988392.988406.
2. D. Radev and W. Fan, "Automatic summarization of search engine hit lists," *Proceedings, ACL Workshop on Recent Advances in NLP and IR*, Hong Kong, (October 2000).
3. M. Franz, J. S. McCarley, and S. Roukos, "Ad hoc and Multilingual Information Retrieval at IBM," in *NIST Special Publication 500-242: The Seventh Text REtrieval Conference (TREC-7)*, Gaithersburg, MD. (1998).
4. F. Sebastiani, "Machine learning in automated text categorization," *ACM Computing. Surveys, ACM Press* **34,** No. 1 (March 2002), 1–47.

Accepted for publication May 24, 2004.

Andrei Z. Broder *IBM Thomas J. Watson Research Center, 19 Skyline Drive, Hawthorne, NY 10532 (abroder@us.ibm.com).* Dr. Broder is an IBM Distinguished Engineer and the CTO of the IBM Institute for Search and Text Analysis. From 1999 until early 2002, he was Vice President for Research and Chief Scientist at the AltaVista Company. Previously, he was a senior member of the research staff at Compaq's Systems Research Center in Palo Alto, California. He graduated summa cum laude from the Technion, the Israeli Institute of Technology, and obtained his M.Sc. and Ph.D. degrees in computer science at Stanford University under Don Knuth. His main research interests are the design, analysis, and implementation of randomized algorithms and supporting data structures, in particular in the context of Web-scale information retrieval and applications. Dr. Broder was the SIGIR (ACM Special Interest Group on Information Retrieval) keynote speaker in 2003, and is co-winner of the best paper award at WWW6 (for his work on elimination of duplicate Web pages) and at WWW9 (for his work on mapping the Web). He has published more than 70 papers and was awarded seventeen patents. He serves as the chair of the IEEE Technical Committee on Foundations of Computer Science.

Arthur C. Ciccolo *IBM Thomas J. Watson Research Center, 19 Skyline Drive, Hawthorne, NY 10532 (ciccolo@us.ibm.com).* Mr. Ciccolo is a Department Group Manager in the Research Division of IBM and co-leader of the IBM Institute for Search and Text Analysis. He currently has responsibility for the information and knowledge management department as well as the Research Division's world-wide strategy in the area of unstructured information management. This includes the work of several hundred researchers in the area of natural language processing (NLP), advanced search and information retrieval, text analysis, machine translation, automated document generation, and the application of human-computer-interaction User-Centered Design to a wide range of applications. Prior to assuming his current responsibil-

ities, he served in a number of senior technical management positions, including Research Division assignee at the IBM Corporate Strategy Group in Armonk, New York, where he contributed technical expertise in the development of business strategy and plans. He has held numerous senior technical management positions within IBM Research, including the management of technical groups working in the areas of operations research, manufacturing systems, electro-optics, and robotics. Combining skills from many of these areas, he formed and served as the CEO of an internal company which developed an advanced rapid prototyping system for producing 3D physical models directly from CAD (computer-assisted design) or imaging systems. Before joining IBM Research, he held senior technical management positions at the Massachusetts Institute of Technology's instrumentation laboratory, where he was responsible for the development of advanced real-time control systems for ICBMs, aircraft, spacecraft, and satellites, and the Charles Stark Draper laboratory, where he headed the Air Force Program's computer science division. In addition, he led the lab's efforts in diversification, establishing significant businesses in manufacturing automation and manufacturing systems.