A grand research challenge:
Plug and play architecture for knowledge management

ANDREI BRODER AND ARTHUR CICCOLO
IBM RESEARCH

The web revolution has exposed hundreds of millions of people to the experiences of search and taxonomy browsing, and has reshaped their expectations of the knowledge retrieval experience *outside the web*, in their workplaces. But study after study shows that at the enterprise level, these expectations are not met. Knowledge management in the enterprise setting and even simply document search are universally perceived as disappointing.

Why is that so? The search technology per se has made enormous strides: web search engines return excellent results on one word queries on a 15 TB corpus even though not so long ago this would have been labeled impossible *in principle*, not as a matter of computational cost.

On the other hand, a number of techniques from Natural Language Processing (NLP) such as statistical text analysis, computational linguistics, speech recognition, machine learning, and taxonomy generation and classification, have been combined with classic search methods and have shown significant benefit. As a result, there is growing confidence that many may move from the status of cutting edge research to commercial application in the near term. For example, the augmentation of standard search query methods with lexical information, such as thesauri, has been shown to improve precision by approximately 50% for short queries. (Mandala et al, Proc SIGIR (1999), W.A Woods et al , Proc 6th ANLP Conf (2000)). Automatic document categorization and classification became more accurate than human processing in the late 1990's and is now considered as an essential means of organizing large corpora for knowledge management systems. Automated summarization of documents based upon information extraction techniques has been demonstrated to improve search efficiency by supporting more focused examination of retrieved documents. 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. 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, at this point in time, some of these technologies, their computational demand might be too high for application to "the entire web", this should be less of a problem in the enterprise where the corpora are usually much smaller.

So it seems that search and knowledge management in the enterprise should 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. For example, 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. But the most important factor is independent of the public Web versus the enterprise, and rests with a fundamental character of the technologies. The advanced technologies described above, for the most part, simply do not work together. 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 their developers being "algorithmic centric". The computational requirements of these technologies are so great, that their developers tended to engage in "programming–in–the–small". That is to say, that they built highly integrated, optimized, and hence closed and rigid, narrow applications of their core competency. To build an end systems to be used by consumers of information, not by programmers, such narrow applications are usually piled on top of each other. If there is any cooperation at all, then 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 be 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. And so on, cascading inefficiency on inefficiency.

There is an alternative to the traditional process described above, which capitalizes on the computational power of distributed systems. We propose as a "grand challenge" building the means to enable these technologies to synergize and thus solve the enterprise KM and search challenge. We submit that the missing part is an architecture that enables the integration of the technologies described above with search and retrieval

Such an architecture would not only enable dramatic new capability in the enterprise, it will also provide an infrastructure which will accelerate progress in the individual technologies. As the individual technologies mature, today's computational barriers may yield to the point where widespread application to the external Web may be practical. This will accelerate the ability to create a truly Semantic Web. We envision the following key elements of this architecture:

- A set of Frameworks which allow Annotators (the key NLP components which perform the individual functions) to interoperate in a computationally efficient manner.
- A set of Web Services which allow these Frameworks or individual Annotators to communicate amongst each other and with the Search and Store systems.
- A plurality of highly scalable Search engines which are customized to the data types being operated upon (text, image, audio, video, parametric data) and interoperate dynamically with the Annotators to index new semantic classes.
- A highly efficient Store which supports persistent storage of both retrieved documents and the results of Annotator's analysis.
- An abstract data structure which serves as a representation of common annotation labels and is the basis for a Common annotation System.
- An Ontology Management System, which would support the use of domain knowledge in to the Search and NLP functions, reducing ambiguity, and hence improving precision and relevance.

To *architect* such a system will require establishing a new set of metrics for evaluating performance as well as agreement on new standards for ontological, linguistic, and system components. To *build* such a system will require a new class of middleware both open and flexible to change that can come from any many sources. The *installation and maintenance* of such a system will present new challenges given the potential size and federated nature of its data sources and the size and distributed nature of its processing structure. The *operating performance* both in terms of throughput and reliability and availability present new challenges when one considers that these systems will become the corporate KM resource for major enterprises. Finally from a research point of view, this system might enable for the first time the direct comparison of various methods (for instance, statistical versus grammarian approaches) in a precisely identical environment.

Returning to the web context, it is worth noting that efforts related to the "semantic web" are predicated on the existence of a large number of annotated documents. How will these documents be produced? We postulate that the semantic web will really take off only if the human effort to create annotated documents will be at most slightly higher than the effort to create plain old html and the benefit to have such documents will far outweigh the costs. This is a chicken-and-egg problem: there is little annotated material on the web (in fact there is not much XML altogether), hence no large scale search engine takes advantage of semantic annotations, hence there is little incentive to put them in. Here is how we can break the cycle: The plug and play architecture that we are proposing will stimulate the production of annotators and the adoption of standards. Thus people will have the tools to produce annotated documents mostly automatically. On the other hand, the progress made by search engines in using these annotations will transfer to large scale web search engines. So we will have a fairly large collection of automatically annotated documents, and the technology to take advantage of them. This will start the snowball.

Thus we are going full circle: the unstructured search paradigm on the web represents an unusual example of a technology that 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 consumer world. The challenge is to make it happen.

Bios

Andrei Broder is an IBM Distinguished Engineer in the Research Division of IBM working on search technology and knowledge management applications. From 1999 until very recently he was Vice President for Research and Chief Scientist at the AltaVista Company. Previously he has been a senior member of the research staff at Compaq's Systems Research Center. He was graduated Summa cum Laude from Technion, the Israeli Institute of Technology, and obtained his M.Sc. and Ph.D. in Computer Science at Stanford University under Don Knuth. His main research interests are the design, analysis, and implementation of probabilistic algorithms and supporting data structures, in particular in the context of web-scale information retrieval and applications. Broder is co-winner of the Best Paper award at WWW6 (for his work on duplicate elimination of web pages) and at WWW9 (for his work on mapping the web). He has published more than seventy papers and has been awarded fourteen patents.

Arthur C. Ciccolo is a Department Group Manager in the Research Division of IBM. He currently has responsibility for the **Information and Knowledge Management Department** as well as being responsible for the Research Division’s world-wide activities in the areas of Unstructured Information Management. This includes the work of several hundred researchers in the area of natural language processing (NLP), including advanced search and information retrieval, text analysis, machine translation, and document generation, as well as the application of Human Computer Interaction (HCI) and User Centric Design to a range of applications from Web navigation and search, to collaboration systems.

Prior to assuming his current responsibilities, Arthur served in a number of senior technical management positions, including Research Division Assignee, **IBM Corporate Strategy Group**, Armonk, NY, where he contributed technical expertise in the development of business strategy and plans.

Prior to these assignments, he 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 or imaging systems.

Before joining IBM Research, Arthur held senior technical management positions at **MIT's Instrumentation Laboratory**, where he was responsible for the development of advanced real-time control systems for ICBM’s, aircraft, spacecraft, and satellites, and the **Charles Stark Draper Laboratory**, where he headed the Air Force Programs Computer Science Division. In addition, he led the Lab’s efforts in diversification, establishing significant businesses in manufacturing automation and manufacturing systems.