Information Discovery:

Needles and Haystacks

Carl Lagoze
Cornell University

Amit Singhal
Google

For thousands of years, people have realized the importance of archiving and finding information. With the advent of computers, it became possible to store large amounts of information in electronic form — and finding useful needles in the resulting haystacks has since become one of the most important problems in information management.

Many systems exist to help users navigate the considerable information available to them over the Internet, which is arguably the biggest information haystack around. From personal email search systems to large corporate information-management systems, from small library collections to the whole Web, search is everywhere. Yet, much work remains. In this issue of *IC*, we showcase some emerging techniques that are helping to improve this vibrant research area.

The Information Retrieval Tradition

We can trace the practice of archiving written information back to around 3000 BC, when the Sumerians designated special areas to store clay tablets with cuneiform inscriptions. Realizing that proper organization and access to the archives was critical for efficient use of information, the Sumerians even developed special classifications to identify every tablet and its content. (See www.libraries.gr for a great historical perspective on modern libraries.)

The need to store and retrieve data has become increasingly important over the ensuing centuries, particularly as inventions such as paper, the printing press, and computers have made it easier to generate larger and larger amounts of written records. In 1945, Vannevar Bush published a ground-breaking article titled "As We May Think," which introduced the idea of automatic access to large amounts of stored knowledge.[1] By the mid-1950s, researchers had built on this idea and created more concrete descriptions of how text archives could be searched automatically with a computer. One of the most influential was H.P. Luhn's proposal for (put simply) using words as indexing units for documents and measuring word overlap as a criterion for retrieval.[2]

Over the years, such efforts have

 Published by the IEEE Computer Society 1089-7801/05/$20.00 © 2005 IEEE

matured into the vibrant field we know as information retrieval. IR researchers explore all aspects of information management and access, applying expertise in a wide variety of topics, including digital libraries, natural-language understanding, statistics, computer science, hypertext, and the Web.

Modern search systems largely use keyword-based algorithms, which years of scrutiny have shown to be the most effective and efficient method for practical, general-purpose search. Although simple on the surface, this approach has led to the development of very sophisticated search algorithms that are tailored to individual domains (such as the Web or companies' intranets).

The Metadata Tradition

An IR system extracts keywords directly from a document corpus. In the Web context, this keyword indexing has been enhanced by deriving indexing information from link structure. However, a rich tradition also exists for using external metadata supplied by authors or third parties. By and large, information discovery in the traditional "bricks and mortar" library context depends on professionally created metadata, which is collected in catalogs. Tools for searching these catalogs have matured through the past several decades and are now dominated by a few commercial library management system (LMS) vendors. In this domain, search depends on well-structured catalog records that include controlled subject vocabularies and name authorities.

Using the terminology from this issue's theme, we can characterize individual library catalogs as separate haystacks. Along with the spread of the Internet, there has been considerable work on federating catalog searches across these haystacks. The most notable product of this work is the Z39.50 protocol (www.niso.org/z39.50/z3950.html), a US National Information Standards Organization (NISO) standard that nearly all LMS vendors support. Using Z39.50 gateways such as the one run by the US Library of Congress (www.loc.gov/z3950/gateway.html), a user can even submit a single query to search across library catalogs worldwide.

The Web's rapid growth over the past decade has challenged this catalog-based search paradigm and the systems and standards that support it. The federated searching community has responded to the Web's maturation by releasing the so-called "next generation Z39.50," which includes a search/retrieve Web (SRW) service (www.loc.gov/z3950/agency/zing/srw/).

Another response, promoted primarily by the library community, has been the development of metadata standards that are less strict than those used in traditional cataloging records. The dominant effort in this area is the Dublin Core Metadata Initiative (www.dublincore.org). The guiding belief underlying the DCMI and related efforts is that metadata remains important, but the complexity and cost of traditional library cataloging must be reduced in the digital context (for both Web and digital library efforts). This is due to several factors, including the Web's massive scale and the ephemeral and informal nature of much of the content on it. Although the DCMI was intended to be a foundation for improving Web search, however, major search engines scarcely use it in practice.

As search has matured into one of the most-used Web applications, commercial interests have made it hard for Web search engines to use metadata in search algorithms. Web metadata often comes from page authors, some of whom provide misleading metadata to search engines for the sole purpose of "Web spamming." In a world in which Web traffic is money, some vendors have considerable incentive for spamming search engine indices to get higher rankings and thus generate more traffic. Nonetheless, a trusted source of metadata, such as a library, publisher, or institutional repository, can still be very valuable for Web search.

Independent of whether metadata is simple or complex, machine- or human-generated, the library, publishing, and institutional-repository communities have shown strong interest in the notion of harvesting metadata to allow "cross-haystack" information discovery. In contrast to federated searching, as in the context of Z39.50 and SRW, metadata harvesting is a relatively simple model through which information providers use a common protocol to expose structured information about their information objects. The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH; www.openarchives.org) is the most widely used protocol for this purpose. It is designed to let service providers access any type of metadata (in fact, any type of XML-structured data) related to any form of information object. Thus, developers of Web-based services could use OAI-PMH to harvest a Dublin Core metadata record about a digital document in an institutional repository, or an XML representation of a portion of a scientific database. As such, OAI-PMH provides an access point for search engine providers and their crawlers to extract indexable information from the

"deep" or "dark" Web, such as scientific databases or publishers' repositories.

Theme Features

The three theme articles in this issue take different perspectives on information discovery.

In "Search Adaptations and the Challenges of the Web," Michael P. Evans and colleagues present a historical survey of IR and examine Web search within that context. They describe the unique challenges and opportunities of Web search, including some areas for potential advances over the next few years. The Evans article presents a good overview of the breadth of the challenges in this field.

Fillipo Menczer explores the issue of semantic similarity among Web pages in "Mapping the Semantics of Web Text and Links." He looks at how well content- and link-based measures of similarity — the two cues most widely used by search engines and other tools — approximate true (that is, human-determined) semantic similarity between Web pages. The ability to capture document and cross-document semantics has been an issue ever since Luhn raised the notion of token-based retrieval. This article presents a nice study of the relationship between document content and link structure and user-perceived document semantics.

With "Ranking Complex Relationships on the Semantic Web," Boanerges Aleman-Meza and colleagues describe information discovery in the context of the emerging Semantic Web, a subject of considerable interest in the W3C and general Web communities. As envisioned, the Semantic Web would provide additional search-relevant cues, such as taxonomies and concept relationships. This article describes how semantically rich metadata could help improve results ranking in response to queries. Improved search is one of the primary justifications for the Semantic Web effort, and this article makes some progress in documenting the possibilities in this area.

These articles certainly don't cover the entire topic space, but they effectively complement the results of a broad spectrum of researchers whose work continues to advance our ability to find information in the increasingly rich and diverse Web.

The distinctions between the Web and traditional libraries are increasingly blurring. This is exemplified by efforts such as Google Print (http://print.google.com), which is currently scanning large portions of major library collections for indexing and access, and Google Scholar (http://scholar.google.com), which indexes a large portion of the scholarly literature. As a result, commonality is increasing in the methods for finding needles in their respective haystacks. In addition, there is a growing need to fully bridge the gap between traditional structured metadata search (exemplified by library catalogs) and full-text indexing (exemplified by modern Web search engines). Efforts such as XQuery 1.0 and XPath 2.0 Full-Text (www.cs.cornell.edu/database/) demonstrate the interest in query languages and indexing technology that seamlessly bridges the gaps between fully-structured, semistructured, and unstructured data.

As the amount of information available online continues to grow at a dramatic rate, information discovery becomes ever more important. One key challenge that lies ahead is personalization of search and discovery. Although commercial sites like Amazon.com already have mechanisms that use our previous actions and choices to influence the results to our new queries, the question remains: how can we accomplish this at a more general Web scale and without compromising privacy? This and many other questions will be the focus of much research and development in the coming years.

References

1. V. Bush, "As We May Think," *Atlantic Monthly*, vol. 176, no. 1, 1945, pp. 101–108.
2. H.P. Luhn, "A Statistical Approach to Mechanized Encoding and Searching of Literary Information," *IBM J. Research and Development*, 1957, pp. 309–317.

Carl Lagoze is senior research associate at Cornell University. His research interests include information and document architectures, scholarly publishing, and digital libraries. Lagoze received an MSE from the Wang Institute of Graduate Studies. He was recently awarded the Frederick G. Kilgour Award for Research in Library and Information Technology. Contact him at lagoze@cs.cornell.edu.

Amit Singhal is a distinguished engineer at Google. His research interests include information retrieval, Web search, and Web data mining. He received a PhD in computer science from Cornell University, where he studied with the late Gerard Salton, one of the founders of the field of modern information retrieval. Prior to joining Google, Singhal was a researcher at AT&T Labs. Contact him at singhal@google.com.