# Of Crawlers, Portals, Mice, and Men: Is there more to Mining the Web?

**Minos N. Garofalakis**
Bell Laboratories
*minos@bell-labs.com*

**Sridhar Ramaswamy**
Epiphany Inc.
*sridhar@epiphany.com*

**Rajeev Rastogi**
Bell Laboratories
*rastogi@bell-labs.com*

**Kyuseok Shim**
Bell Laboratories
*shim@bell-labs.com*

The World Wide Web is rapidly emerging as an important medium for transacting commerce as well as for the dissemination of information related to a wide range of topics (e.g., business, government, recreation). According to most predictions, the majority of human information will be available on the Web in ten years. These huge amounts of data raise a grand challenge for the database community, namely, how to turn the Web into a more useful information utility. This is exactly the subject that will be addressed by this panel.

**Panel objective.** To debate methods for improving information and knowledge discovery on the Web. The following questions are germane to such a debate.

**What are the challenges to efficient and effective knowledge discovery on the Web?** Crawlers, search engines and Web directories like Yahoo! constitute the state-of-the-art tools for information retrieval on the Web today. Crawlers for the major search engines retrieve Web pages on which full-text indexes are constructed. A user query is simply a list of keywords (with some additional operators), and the query response is a list of pages ranked based on their similarity to the query.

Today's search tools are plagued by the following four problems: (1) the *abundance problem*, that is, the phenomenon of hundreds of irrelevant documents being returned in response to a search query, (2) *limited coverage* of the Web, (3) a limited query interface that is based on syntactic keyword-oriented search, and (4) limited customization to individual users. These problems, in turn, can be attributed to the following characteristics of the Web. First and foremost, the Web is a huge, diverse and dynamic collection of interlinked hypertext documents. There are about 300 million pages on the Web today with about 1 million being added daily. Furthermore, it is widely believed that 99% of the information on the Web is of no interest to 99% of the people. Second, except for *hyperlinks*, the Web is largely unstructured. Finally, most information on the Web is in the form of HTML documents for which analysis and extraction of content is very difficult. Furthermore, the contents of many internet sources are hidden behind search interfaces and, thus, cannot be indexed – HTML documents are dynamically generated by these sources, in response to queries, using data stored in commercial DBMSs.

The question therefore is: how can we overcome these and other challenges that impede the Web resource discovery process?

**How can more structure be imposed on the Web?** Topic hierarchies like the one provided by Yahoo! give a hierarchical classification of documents. Searches in the context of a specific topic help to eliminate clutter by focusing the search to only documents pertaining to the topic. Unfortunately, since these hierarchies are manually generated, they cover only a small portion of the Web. The challenge here is to automate the classification and clustering of millions of dynamically changing Web documents with diverse authorship.

**How can customization be improved?** Customization involves learning about an individual user's preferences/interests based on access patterns or alternately, based on explicit directives from the user. Thus, customization aids in providing users with pages, sites and advertizements that are of interest to them. It may also be possible for sites to automatically optimize their design and organization based on observed user patterns.

**Will XML solve all our information discovery problems?** Information extraction from web resources is complicated since HTML annotations provide very little semantic information. Furthermore, a number of sources hide information behind search interfaces. Thus, the majority of today's information-extraction systems rely on "hand coded" wrappers to access a fixed set of Web resources. Obviously, such a manual approach cannot scale, and new techniques for automating the extraction process from unfamiliar documents need to be devised. It is widely believed that HTML will be replaced by XML, forcing documents to become self-describing through the specification of tag sets (referred to as the Document Type Definitions, or DTDs). Thus, the contents of each XML document can be extracted by consulting the DTDs to which the document conforms. Web sites will also be able to describe their query capabilities through XML – thus enabling structured queries like "find the cheapest airline ticket from New York to Chicago" or "list all jobs with salary $>$ 50K in the Boston area". Will XML be able to transform the entire Web into one unified database ?

**What is the vision for the future?** How will users interact with the Web in the future? Will structured, declarative querying ever become the de-facto standard for the Web? Will traditional mining techniques like clustering, correlation, causality, and classification be able to cope with the scale, heterogeneity, and dynamic nature of the Web? What are the key innovations necessary to facilitate knowledge discovery on the Web? The Web contains a wealth of information on diverse topics, and the discovery of "interesting" patterns could be of immense benefit to mankind (e.g., excessive oil drilling can cause contamination of ground water, periods of high economic growth are typically followed by brutal recessions).

**Panelists.**

- Rakesh Agrawal (IBM Almaden)
- Surajit Chaudhuri (Microsoft Research)
- Umeshwar Dayal (HP Labs)
- Jiawei Han (Simon Fraser University)
- Raghu Ramakrishnan (University of Wisconsin)
- Jeff Ullman (Stanford University)