In the earliest days of search engines, people were amazed and delighted to stumble across the information they sought; eureka moments were unexpected surprises, said Yahoo!’s cofounder Jerry Yang. ‘Today their attitudes are much more presumptive. They presume that the information they’re looking for is certainly available and that it’s just a matter of technologists making it easier to get to, and in fewer keystrokes,’ he said. (Friedman, 2005 p. 154)

Web mining is the search for relevant information from the World Wide Web. The found Web pages in a search are relevant if they provide an accurate answer to the searcher’s information need. As all users of Web search engines are aware, accurate answers are not always at the top of the results list.

Accuracy is a function of three elements—the searcher, the page author, and the search engine. The searcher has a sense of what they are looking for on the Web and develops a keyword query to represent that information need. Meanwhile, Web search engines are capturing and interpreting the semantic meaning from the content of Web pages; expressing the page author’s intent again as keywords. The search engine matches the query keywords to the page keywords. Thus, the quality of search results are dependent upon the accuracy of the keyword representation of both the searcher’s query and author’s page, and the search engine’s ability to understand both.

The extraction of meaning from a Web page requires an understanding of the page elements and an understanding of the relationships between those elements. Also, an understanding of the relationship between Web pages can expedite finding the appropriate page. Understanding the searcher sometimes requires assisting the searcher with hints about the information need domain or what other searchers with similar needs have found.

This issue is dedicated to research on intelligent search of the World Wide Web aimed at improving search results and meeting the expectations of people using the Web. We begin with an article that presents a technique to improve the understanding of the page author’s intent. The next two articles address page similarity to help satisfy the searcher’s need.
The fourth article looks at shortening the number of hops needed to reach a desired page, assisting the searcher further. The final article takes a practical look at ranking not the pages found, but rather at the search engines. The technologist and their contributions to this issue are:

Werner and Böttcher who consider in the first article that a text document, such as a Web page, contains information about itself that goes beyond the text itself. Content is emphasized in various ways to bring out the author’s tone and intentions. When evaluating the meaning of a document, whether it is a Web page or other text document, the search engine must consider what text was emphasized and how it was emphasized. The article, “Supporting Text Retrieval by Typographical Term Weighting,” provides approaches for improving the classification of a document by considering the typographical information in the document’s format.

Web searches are improved when Web pages that are similar in content can be identified. This allows the presentation of search results, which more accurately reflect the needs of the searcher. This similarity can be accomplished by classifying the Web pages into known classes, or clustering them based solely on the content of the page. Zhongmei Yao and Ben Choi have developed a bi-directional hierarchical clustering algorithm that clusters similar Web pages in an unsupervised manner. This algorithm, presented in “Clustering Web Pages into Hierarchical Categories,” proves to be better than established methods and provides visualization at different granularities of similarity. Additionally, this article addresses the question of the number of clusters optimally needed for a given collection of Web pages.

The user’s ability to express their information need is a critical element in any search. It is also the most difficult to control. Search engines are needed that can compensate for the user not knowing exactly for what they are searching. This can be accomplished by relevance feedback. Picariello and Rinaldi present an ontology-based relevance feedback technique for ranking search results in “User Relevance Feedback in Semantic Information Retrieval.” They consider the domain of the search, develop a dynamic semantic network, and calculate a syntactic-semantic grade for each term and a semantic grade for each document. These are then used to rank the documents, improving recall and precision.

As the Web is accessed more and more by mobile devices it becomes more important to efficiently navigate from Web page to Web page. “Improving Mobile Web Navigation Using N-grams Prediction Models” presents this ability. Using Web usage mining, Fu, Paul and Shetty are able to discover shortcuts to Web pages based on previous pages viewed. These shortcuts can be dynamically provided to users when navigating from a Web page and looking for a desired destination Web page. Their MINCOST algorithm is especially effective in reducing the time spent by mobile users as they search the Web.

While search engines rate Web pages to return an ordered set to a searcher, Meng, Xing and Clark look at rating the search engine performance. The traditional ranking metrics for information retrieval are dependent on knowledge of
the content of the information repository. This knowledge is never available with the World Wide Web. In “An Empirical Performance Measurement of Microsoft’s Search Engine and its Comparison with other Major Search Engines,” the Rank-Power metric is used, along with response time, to compare the performance of various search engines.

These articles are primarily a direct result of the 2005 International Conference on Data Mining (DMIN ’05) and its session on Web mining. The conference articles have been expanded and again double blind reviewed by experts in the field. I am grateful for the encouragement of the DMIN ’05 chairman, Hamdi Arabnia, for suggesting this special issue.

REFERENCE

Anthony Scime is a 1997 graduate of George Mason University with an interdisciplinary doctorate in information systems and education. Currently he is an associate professor of computer science at the State University of New York College at Brockport. Prior to joining academia he spent over 20 years in industry and government applying information systems to solve large-scale problems. Idea Group Publishing has recently published his book Web Mining: Applications and Techniques. He is also a member of editorial review boards of the Information Resources Management Journal (IRMJ), the Journal of IT Education and the Journal of Issues in Informing Science and Information Technology. His research interests include the World Wide Web as an information system and database, information retrieval, knowledge creation and management, decision making from information, data mining and computing education.