Introduction to the *JASIST* Special Topic Section on Web Retrieval and Mining: A Machine Learning Perspective

Hsinchun Chen, Guest Editor
McClelland Professor of MIS, Director, Artificial Intelligence Lab and Hoffman E-Commerce Lab, Management Information Systems Department, The University of Arizona, Tucson, AZ 85721. E-mail: hchen@bpa.arizona.edu

Web Retrieval and Mining: Introduction

Research in information retrieval (IR) has advanced significantly in the past few decades. Many tasks, such as indexing and text categorization, can be performed automatically with minimal human effort. Machine learning has played an important role in such automation by learning various patterns such as document topics, text structures, and user interests from examples.

In recent years, it has become increasingly difficult to search for useful information on the World Wide Web because of its large size and unstructured nature. Useful information and resources are often hidden in the Web. While machine learning has been successfully applied to traditional IR systems, it poses some new challenges to apply these algorithms to the Web due to its large size, link structure, diversity in content and languages, and dynamic nature. On the other hand, such characteristics of the Web also provide interesting patterns and knowledge that do not present in traditional information retrieval systems.

Machine Learning for Information Retrieval and Analysis: Pre-Web

Learning techniques had been applied in information retrieval (IR) applications long before the recent advances of the Web. In this section, we will briefly survey some of the research in this area, covering the use of machine learning in information extraction, relevance feedback, information filtering, text classification, and text clustering.

Information extraction is one area in which machine learning is applied in IR. Information extraction techniques aim to identify useful information from text documents automatically. Named-entity extraction is one of the most widely studied sub-fields. It refers to the automatic identification from text documents of the names of entities of interest, such as persons (e.g., “John Doe”), locations (e.g., “Washington, D.C.”), and organizations (e.g., “National Science Foundation”). It also includes the identification of other patterns, such as dates, times, number expressions, dollar amounts, email addresses, and Web addresses (URLs). The Message Understanding Conference (MUC) series has been the major forum for researchers in this area, where they meet and compare the performance of their entity extraction approaches (Chinchor, 1998). Machine learning is one of the major approaches. Machine learning-based entity extraction systems rely on algorithms rather than human-created rules to extract knowledge or identify patterns from texts. Examples of machine learning algorithms include neural networks, decision tree (Baluja, Mittal, & Sukthankar, 1999), Hidden Markov Model (Miller et al., 1998), and entropy maximization (Borthwick, Sterling, Agichtein, & Grishman, 1998). Instead of relying on a single approach, most existing information extraction systems combine machine learning with other approaches (such as a rule-based approach or a statistical approach).

Relevance feedback is a well-known method used in IR systems to help users conduct searches iteratively and reformulate search queries based on evaluation of previously retrieved documents (Ide, 1971; Rocchio, 1971). The main assumption is that documents relevant to a particular query are represented by a set of similar keywords (Salton, 1989). After a user rates the relevance of a set of retrieved documents, the query can be reformulated by adding a set of terms from the relevant documents and subtracting a set of terms from the irrelevant documents. It has been shown that a single iteration of relevance feedback can significantly improve search precision and recall (Salton, 1989). Probabilistic techniques have been applied to relevance feedback by estimating the probabilistic of relevance of a given document to a user. Using relevance feedback, the model can be applied to learn the common characteristics of the relevant documents in order to estimate the probability of relevance for the remaining documents in a collection (Fuhr & Buckley, 1991; Fuhr & Pfeifer, 1994; Chen, Shankaranarayanan, Iyer, & She, 1998).

Similar to relevance feedback, *information filtering and recommendation* techniques employ user evaluation to improve IR system performance. The main difference is that while relevance feedback helps users reformulate their search queries, information filtering techniques try to learn about users' interests based on their evaluations and actions, and then to use this information to analyze new documents. Information filtering systems are usually designed to alleviate the problem of information overload in IR systems. Decision tree has been used for news-article filtering (Green & Edward, 1996). Another approach is called collaborative filtering or recommender systems, in which collaboration is achieved as the system allows users to help one another perform filtering by recording their reactions to documents they read (Goldberg, Nichols, Oki, & Terry, 1992). One example is the GroupLens system which performs collaborative filtering on USENET news articles (Konstan, Miller, Maltz, Herlocker, Gordon, & Riedl, 1997). Many personalization and collaborative systems have been implemented as software agents to help different information systems users (Maes, 1994).

Text classification and *text clustering* have been studied extensively in traditional IR literature. Text classification is the classification of textual documents into predefined categories (supervised learning), while text clustering grouped documents into categories dynamically defined based on their similarities (unsupervised learning). Machine learning is the basis of most text classification and clustering applications. Text classification has been extensively studied at SIGIR conferences and evaluated on standard testbeds. For example, the Naïve Bayesian method has been widely used (e.g., Koller & Sahami, 1997; Lewis & Ringuette, 1994; McCallum, Nigam, Rennie, & Seymore, 1999). This method uses the joint probabilities of words and categories to estimate the probabilities of categories given a document. The *k*-nearest neighbor method is another widely used approach in text classification. For a given document, the k neighbors that are most similar to a given document are first identified (Iwayama & Tokunaga, 1995; Masand, Linoff, & Waltz, 1992). Neural network programs also have been applied to text classification, usually employing the feedforward/backpropagation neural network model (Wiener, Pedersen, & Weigend, 1995; Ng, Goh, & Low, 1997; Lam & Lee, 1999). Another new technique used in text classification is called support vector machine (SVM), a statistical method that tries to find a hyperplane that best separates two classes (Vapnik, 1995). Joachims first applied SVM to text classification (Joachims, 1998). It has been shown that SVM achieved the best performance on the Reuters-21578 data set for document classification (Yang & Liu, 1999).

Similarly to text classification, text clustering tries to assign documents into different categories based on their similarities. However, in text clustering, there are no predefined categories; all categories are dynamically defined. There are two types of clustering algorithms, namely hierarchical clustering and non-hierarchical clustering. The *k*-nearest neighbor method and Ward's algorithm (Ward, 1963) are the most widely used hierarchical clustering methods. For non-hierarchical clustering, one of the most common approaches is the K-means algorithm. Another clustering approach being used a lot in recent years is the neural network approach. Kohonen's self-organizing map (SOM), which produces a 2-dimensional grid representation for *N*-dimensional features, has been widely applied in IR (Lin, Soergel, & Marchionini, 1991; Kohonen, 1995; Orwig, Chen, & Nunamaker, 1997).

Web Retrieval and Mining

The term Web mining was coined by Etzioni (1996) to denote the use of data mining techniques to automatically discover and extract information from Web documents and services. Web mining research can be classified into three categories: Web content mining, Web structure mining, and Web usage mining (Kosala & Blockeel, 2000).

Web content mining refers to the discovery of useful information from Web contents, including text, image, audio, video, etc. Web content mining research includes resource discovery from the Web (e.g., Cho, Garcia-Molina, & Page, 1998; Chakrabarti et al., 1999), document categorization and clustering (e.g., Zamir & Etzioni, 1999; Kohonen, Kaski, Lagus, Salojärvi, Honkela, Paatero, & Saarela, 2000), and information extraction from Web pages (e.g., Hurst, 2001). Web structure mining studies the model underlying the link structures of the Web. It usually involves the analysis of in-links and out-links information of a Web page, and has been used for search engine result ranking and other Web applications (e.g., Brin & Page, 1998; Kleinberg, 1998). Web usage mining focuses on analyzing search logs or other activity logs (in a way similar to data mining) to find interesting patterns. One of the main applications of Web usage mining is to learn user profiles (e.g., Armstrong, Freitag, Joachims, & Mitchell, 1995; Wasfi, 1999).

There are a few major differences between Web retrieval and traditional IR. First, most Web documents are in HTML (HyperText Markup Language) format. HTML documents contain many markup tags, mainly used for formatting. Web mining applications must parse the HTML documents to remove these markup tags. But the tags also can provide additional information about the document. For example, a bold typeface markup (<b>) may indicate that a term is more important than other terms that appear in normal typeface. Such formatting cues have been widely used to determine the relevance of terms (Arasu, Cho, Garcia-Molina, Paepcke, & Raghavan, 2001).

Second, while traditional IR systems often contain structured and well-written documents (e.g., news articles, research papers, metadata), this is not the case on the Web. Web documents are much more diverse in terms of length, document structure, writing style, and many Web pages contain grammatical and spelling errors. Web pages are also very diverse in terms of languages and domains; one can find almost any language and any topic on the Web. In addition, the Web has many different types of content, including text, images, audios, videos, and executables. There are numerous formats, such as HTML, XML, PDF, MS Word, mp3, wav, ra, rm, avi, just to name a few. Web applications have to deal with these different formats and retrieve the desired information.

Third, while most documents in traditional IR systems tend to remain static over time, Web pages are much more dynamic; they can be updated every day, every hour or even every minute. Some Web pages do not even have a static form; they are dynamically generated on request, with content varying according to the user and the time of the request. Such dynamics make it much more difficult for retrieval systems such as search engines to keep an up-to-date search index of the Web.

Another characteristic of the Web, perhaps the most important one, is the hyperlink structure. Web pages are hyperlinked to each other, and it is through hyperlink that a Web page author "cites" other Web pages. Intuitively, the author of a Web page places a link to another Web page if he or she believes that it contains a relevant topic or is of good quality (Kleinberg, 1998). Anchor text, the underlined, clickable text of an outgoing link in a Web page, also provides a good description of the target page because it represents how other people linking to the page actually describe it. Several studies have tried to make use of anchor text or the text nearby to predict the content of the target page (Amitay, 1998; Rennie & McCallum, 1999).

Lastly, the size of the Web is larger than traditional IR collections by several orders of magnitude. The number of indexable Web pages has exceeded 2 billion, and is still growing at a rate of roughly 1 million pages per day (Lawrence & Giles, 1999; Lyman & Varian, 2000). Collecting, indexing, and analyzing these documents presents a great challenge. Similarly, the population of Web users is much larger than that of traditional IR systems. Collaboration among users can be more feasible because of the availability of a large user base, but it can also be more difficult because users are more diverse.

In This Issue

This special issue consists of six papers that report research in web retrieval and mining. Most papers apply or adapt various pre-web retrieval and analysis techniques to other interesting and challenging web-based applications.

The first paper, "Automatic Generation of English/Chinese Thesaurus Based on a Parallel Corpus in Laws," by Yang and Luk, describes a project that aims to address cross-lingual semantic interoperability by developing a cross-lingual thesaurus based on an English/Chinese parallel corpus. Their experiments showed that such a thesaurus is useful in suggesting relevant terms in a different language. The second paper, "DocCube: Multi-Dimensional Visualization and Exploration of Large Documents Sets," by Mothe, Chrisment, Dousset, and Alaux, presents a novel user interface that provides global visualization of large document sets to help users formulate query and access documents. Concept hierarchies are introduced to facilitate browsing. The third paper, "Relevant Term Suggestion in Interactive Web Search Based on Contextual Information in Query Session Logs," by Huang, Chien, and Oyang, proposes a query log-based term suggestion approach to interactive Web search. Using this approach, relevant terms suggested for a user query are those that co-occur in similar query sessions from search engine logs, rather than in the retrieved documents. Their experiments showed that the proposed approach can exploit the contextual information in a user query session to make useful suggestions. The fourth paper, "A Novel Method for Discovering Fuzzy Sequential Patterns Using the Simple Fuzzy Partition Method," by Chen and Hu, proposes a fuzzy data mining technique to discover fuzzy sequential patterns. The fifth paper, "Client-Side Monitoring for Web Mining," by Fenstermacher and Ginsburg, proposes a client-side monitoring system that is unobtrusive and supports flexible data collection. Moreover, the proposed framework encompasses client-side applications (such as standard office productivity tools) beyond the Web browser. The sixth and last paper, "HelpfulMed: Intelligent Searching for Medical Information over the Internet," by Chen, Lally, Zhu, and Chau, describes an "intelligent" web-based medical portal that supports meta searching, vertical search engine creation, term suggestion, and knowledge map browsing, all in an integrated web-based architecture. Initial user evaluations of the system were promising in comparison to other traditional medical search engines.

Conclusions and Future Directions

The Web has become the world's largest knowledge repository. Extracting knowledge from the Web efficiently and effectively is becoming increasingly important for various Web applications. The current Web still consists of more information than knowledge. Also, most of the Web mining activities are still in their early stages and will continue to develop as the Web evolves. We hope this collection of research papers will help advance our knowledge and understanding of this fascinating and evolving field of web retrieval and mining.

References

Amitay, E. (1998). Using common hypertext links to identify the best phrasal description of target Web documents. In Proceedings of the ACM SIGIR'98 Post-Conference Workshop on Hypertext Information Retrieval for the Web, Melbourne, Australia, 1998.

Arasu, A., Cho, J., Garcia-Molina, H., Paepcke, A., & Raghavan, S. (2001). Searching the web. ACM Transactions on Internet Technology, 1(1), 2–43.

Armstrong, R., Freitag, D., Joachims, T., & Mitchell, T. (1995). WebWatcher: a learning apprentice for the World Wide Web. In Proceedings of the AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments, Stanford, CA, March 1995.

Baluja, S., Mittal, V., & Sukthankar, R. (1999). Applying machine learning for high performance named-entity extraction. In Proceedings of the Conference of the Pacific Association for Computational Linguistics, Waterloo, Ontario, 1999.

Borthwick, A., Sterling, J., Agichtein, E., & Grishman, R. (1998). NYU: description of the MENE named entity system as used in MUC-7. In

Proceedings of the Seventh Message Understanding Conference (MUC-7), Washington, D.C., April 1998.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. In Proceedings of the 7th WWW Conference, Brisbane, Australia, April 1998.

Chakrabarti, S., van der Berg, M., & Dom, B. (1999). Focused crawling: a new approach to topic-specific Web resource discovery. In Proceedings of the 8th World Wide Web Conference, Toronto, May 1999.

Chen, H., Shankaranarayanan, G., Iyer, A., & She, L. (1998). A machine learning approach to inductive query by examples: an experiment using relevance feedback, ID3, genetic algorithms, and simulated annealing. Journal of the American Society for Information Science, 49(8), 693–705.

Chinchor, N.A. (1998). Overview of MUC-7/MET-2. In Proceedings of the Seventh Message Understanding Conference (MUC-7), Washington, D.C., April 1998.

Cho, J., Garcia-Molina, H., Page, L. (1998). Efficient crawling through URL ordering. In Proceedings of the 7th WWW Conference, Brisbane, Australia, April 1998.

Etzioni, O. (1996). The World Wide Web: quagmire or gold mine. Communications of the ACM, 39(11), 65–68.

Fuhr, N., & Buckley, C. (1991). A probabilistic learning approach for document indexing. ACM Transactions on Information Systems, 9, 223–248.

Fuhr, N., & Pfeifer, U. (1994). Probabilistic information retrieval as a combination of abstraction, inductive learning, and probabilistic assumption. ACM Transactions on Information Systems, 12(1), 92–115.

Goldberg, D., Nichols, D., Oki, B., & Terry, D. (1992). Using collaborative filtering to weave an information tapestry. Communications of the ACM, 35(12), 61–69.

Green, C.L., & Edwards, P. (1996). Using machine learning to enhance software tools for internet information management. In Proceedings of the AAAI-96 Workshop on Internet-Based Information Systems (pp. 48–55), Menlo Park, CA, AAAI, 1996.

Hurst, M. (2001). Layout and language: challenges for table understanding on the Web. In Proceedings of the 1st International Workshop on Web Document Analysis (pp. 27–30), Seattle, WA, September 2001.

Ide, E. (1971). New experiments in relevance feedback. In G. Salton (Ed.), The SMART retrieval system—experiments in automatic document processing. Englewood Cliffs, NJ: Prentice-Hall, pp. 337–354.

Iwayama, M., & Tokunaga, T. (1995). Cluster-based text categorization: a comparison of category search strategies. In Proceedings of the 18th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR'95) (pp. 273–281), Seattle, WA, July 1995.

Joachims, T. (1998). Text categorization with support vector machines: learning with many relevant features. In Proceedings of the European Conference on Machine Learning, Berlin, 1998, pp. 137–142.

Kahle, B. (1997). Preserving the Internet. Scientific American, March 1997, 82–83.

Kleinberg, J. (1998). Authoritative sources in a hyperlinked environment. In Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms, San Francisco, CA, January 1998, pp. 668–677.

Kohonen, T. (1995). Self-organizing maps. Berlin: Springer-Verlag.

Kohonen, T., Kaski, S., Lagus, K., Salojärvi, J., Honkela, J., Paatero, V., & Saarela, A. (2000). Self organization of a massive document collection. IEEE Transactions on Neural Networks, Special Issue on Neural Networks for Data Mining and Knowledge Discovery, 11(3), 574–585.

Koller, D., & Sahami, M. (1997). Hierarchically classifying documents using very few words. In Proceedings of the 14th International Conference on Machine Learning (ICML'97) (pp. 170–178), Nashville, TN, 1997.

Konstan, J.A., Miller, B., Maltz, D., Herlocker, J., Gordon, L., & Riedl, J. (1997). GroupLens: applying collaborative filtering to usenet news. Communications of the ACM, 40(3), 77–87.

Kosala, R., & Blockeel, H. (2000). Web mining research: a survey. ACM SIGKDD Explorations, 2(1), 1–15.

Lam, S.L.Y, & Lee, D.L. (1999). Feature reduction for neural network based text categorization. In Proceedings of the International Conference on Database Systems for Advanced Applications (DASFAA '99) (pp. 195–202), Hsinchu, Taiwan, April 1999.

Lawrence, S., & Giles, C.L. (1999). Accessibility of information on the Web. Nature, 400, 107–109.

Lewis, D.D., & Ringuette, M. (1994). A comparison of two learning algorithms for text categorization. In Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval (SDAIR'94) (pp. 81–93), 1994.

Lin, X., Soergel, D., & Marchionini, G. (1991). A self-organizing semantic map for information retrieval. In Proceedings of the 14th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'91) (pp. 262–269), Chicago, IL, 1991.

Lyman, P., & Varian, H.R. (2000). How much information. [Online]. Available at http://www.sims.berkeley.edu/how-much-info/. February 20, 2001.

Maes, P. (1994). Agents that reduce work and information overload. Communications of the ACM, 37(7), 31–40.

Masand, B., Linoff, G., & Waltz, D. (1992). Classifying news stories using memory based reasoning. In Proceeedings of the 15th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR'92) (pp. 59–64), Copenhagen, Denmark, 1992.

McCallum, A., Nigam, K., Rennie, J., & Seymore, K. (1999). A machine learning approach to building domain-specific search engines. In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI-99) (pp. 662–667), Stockholm, Sweden, 1999.

Miller, S., Crystal, M., Fox, H., Ramshaw, L., Schwartz, R., Stone, R., Weischedel, R., and the Annotation Group (1998). BBN: Description of the SIFT system as used in MUC-7. In Proceedings of the Seventh Message Understanding Conference (MUC-7), Washington, D.C., April 1998.

Ng, H.T., Goh, W.B., & Low, K.L. (1997). Feature selection, perceptron learning, and a usability case study for text categorization. In Proceedings of the 20th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR'97) (pp. 67–73), Philadelphia, PA, 1997.

Orwig, R., Chen, H., & Nunamaker, J.F. (1997). A graphical self-organizing approach to classifying electronic meeting output. Journal of the American Society for Information Science, 48(2), 157–170.

Rennie, J., & McCallum, A.K. (1999). Using reinforcement learning to spider the Web efficiently. In Proceedings of the 16th International Conference on Machine Learning (ICML-99) (pp. 335–343), Bled, Slovenia, 1999.

Rocchio, J.J. (1971). Relevance feedback in information retrieval. In G. Salton (ed.), The SMART Retrieval System—Experiments in automatic document processing. Englewood Cliffs, NJ: Prentice-Hall, pp. 337–354.

Salton, G. (1989). Automatic text processing. Reading, MA: Addison-Wesley.

Vapnik, V. (1998). Statistical learning theory. Chichester, GB: Wiley.

Ward, J. (1963). Hierarchical grouping to optimize an objection function. Journal of the American Statistical Association, 58, 236–244.

Wasfi, A.M.A. (1999). Collecting user access patterns for building user profiles and collaborative filtering. In Proceedings of the 1999 International Conference on Intelligent User Interfaces (IUI'99) (pp. 57–64), Los Angeles, CA, 1999.

Wiener, E., Pedersen, J. O., & Weigend, A.S. (1995). A neural network approach to topic spotting. In Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval (SDAIR'95) (pp. 317–332), Las Vegas, NV, 1995.

Yang, Y., & Liu, X. (1999). A re-examination of text categorization methods. In Proceedings of the 22nd Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR'99) (pp. 42–49), Berkeley, CA, 1999.

Zamir, O., & Etzioni, O. (1999). Grouper: a dynamic clustering interface to Web search results. In Proceedings of the 8th World Wide Web Conference, Toronto, May 1999.