Guest Editorial
Machine Learning for the Internet

GARY WILLIAM FLAKE
Yahoo! Research Labs
PAOLO FRASCONI
Università di Firenze
C. LEE GILES
Pennsylvania State University
and
MARCO MAGGINI
Università di Siena

INTRODUCTION

Our ability to obtain efficient algorithmic solutions is often limited by factors such as a poor understanding of the underlying problem domain, the intrinsic presence of uncertainty, or, in some cases, the unaffordable computational cost of finding an exact solution. For all of these factors, problem instances of an enormous size can be both a curse and a blessing. As problem instances become larger, many of these confounding factors are often magnified, hence size can be a curse. However, enormous problem instances may also yield an unexpected source of power in finding solutions when size can be leveraged in nontrivial ways.

The World Wide Web has been at the center of a revolution in how algorithms are designed with massive amounts of data in mind. The essence of this revolution is conceptually very simple: real-world massive data sets are, more often than not, highly structured and regular. Regularities can be used in two complementary ways. First, systematic regularities within massive data sets can be used to craft algorithms that are potentially suboptimal in the worst-case, but highly effective for expected cases. Second, nonsystematic regularities—those that are too subtle to be encoded within an algorithm—can be discovered by

Authors' addresses: Gary William Flake (http://labs.yahoo.com/~flakeg/); Paolo Frasconi (http://www.dsi.unifi.it/~paolo/); C. Lee Giles (http://clgiles.ist.psu.edu/); Marco Maggini (http://www-dii.ing.unisi.it/~maggini/).

automated methods so that the solutions are actually determined by the underlying data. In both cases, the existence of enormous problem instances that arise from a highly regular source is key to building more effective methods.

Machine learning is the study of how to build adaptive solutions that make explicit use of regularities in data. Machine learning is rapidly becoming a very influential discipline, virtually pervading all other fields where intelligent processing of information is necessary or useful. The study of the Internet and its related technologies is no exception. After less than a decade, since the first World Wide Web Conference, the literature on the theory and application of machine learning to the Internet is already large and variegated.

During the 1990's, the Web became the object of several new directions of scientific investigation. It is no coincidence that a significant portion of these efforts focused on the Web's hyperlink network properties that has transversely been developed across multiple fields including computer science, sociology, physics, and biology. In many of these studies, the network structure of the Web's hyperlinks played a prominent role; however, many other efforts grounded the more theoretical studies by explicitly exploring the relationship between hyperlinks and Web content. In addition, several hyperlink analysis algorithms were devised for locating important Web pages or identifying communities of related information. Although none of these studies can be technically seen as belonging to the subject of machine learning, they all explore the Web's systematic structural regularities. Together these form a solid bulk of results underpinning many of the present approaches to intelligent information processing on the Web.

Information retrieval is, perhaps, the research area that has been most significantly revitalized during the last few years because of the need for new methods for the Internet. Machine learning is presently playing an important role in solving several retrieval-related problems. For example, text categorization is most successfully solved using methods such as support vector machines or Bayesian classifiers. In the case of Web documents, the relational information contained in the hyperlinks can be useful to improve classification accuracy, as demonstrated in various frameworks such as cotraining, inductive logic programming, and, more recently, probabilistic relational models. Probabilistic versions of latent semantic analysis are easily described as Bayesian networks with parameters learned from data. Document clustering is a form of unsupervised learning with great potential usefulness in Web applications (for example, to group documents returned by a search engine). Machine learning has also been successfully applied to information extraction from Web documents, as well as to focused crawling of the Web with the purpose of retrieving and indexing documents related only to a given topic of interest.

Data mining is a second broad research area that has received significant attention in conjunction with the analysis of large amounts of Web-related data. In this context, human behavior can be studied from various angles in order to build data-driven predictive models. For example, various forms of adaptive Markov models have been proposed for exploring navigation patterns on the Web, and Bayesian networks have been used to represent the query actions of a search engine's user. Another interesting area of application is related

to e-commerce, where machine learning can provide personalized automatic recommendation systems, for example, through collaborative filtering.

IN THIS ISSUE

We received 28 submissions for this special issue from 11 countries in Asia, Europe, and North America. Because of the high number of high-quality manuscripts, a second issue is scheduled in this journal later on this year. Four articles appear in this issue.

"Learning to Find Answers to Questions on the Web" by Agichtein, Lawrence, and Gravano is about question-answering. Ideally, users of a Web search engine should be able to formulate their queries in natural language. However, the available retrieval technology used by general purpose engines is largely based on keyword matching rather than semantic matching. In this article, the authors present the prototype of a system called TRITUS that uses machine learning to transform natural language queries into keyword-based queries that are likely to contain an answer to the original question. The system is focused on four question types: *who*, *how*, *where*, and *what*. Examples of question-answer pairs for training the system are derived in this case from a large collection of Frequently Asked Questions.

"Selective Markov Models for Predicting Web-Page Accesses" by Deshpande and Karypis presents an improved algorithm for predicting what page will be accessed next by a surfer browsing a certain Web site, given the history of previously accessed pages. A natural approach for modeling the conditional probabilities of pages given user sessions is based on higher-order Markov chains. However, model complexity (as measured by the number of states) grows exponentially with the length of the session history that is taken into account. Deshpande and Karypis propose and compare three alternative pruning policies for reducing the size of the state space. In the first model, states that rarely occur are removed. In the second model, only those states that sufficiently discriminate between the most probable next page and the second most probable page are retained. Finally, in the third model, states are selected based on the generalization performance measured on a validation set.

"PageCluster: Mining Conceptual Link Hierarchies from Web Log Files for Adaptive Web Site Navigation" by Zhu, Hong, and Hughes complements the previous work by showing how user traversals from log files can be used to infer the semantic relationship between Web pages on a site. In this framework, traversals are treated as weights which allow pages on a site to be clustered into different conceptual hierarchies. Zhu, Hong, and Hughes describe a prototype for adaptive Web site navigation that can help users find information more efficiently and effectively for many information retrieval tasks. Their results also show how user data nicely complements bibliographic data for identifying pages with semantically related content.

"A Web Mining Approach to Hyperlink Selection for Web Portals" by Fang and Liu Sheng focuses on the automatic identification of a small set of documents in a large Web site that maximally facilitates access to the information contained in the site assuming the user will follow hyperlinks rather than perform

keyword searches. The algorithm proposed by Fang and Liu Sheng identifies two types of relationships among documents: structural-based, defined by the network structure of the site, and access-based, defined by monitoring user behavior on the Web site. In particular, the second form of relationship is identified by mining association rules from log files. Empirical results on the University of Arizona Web site indicate that the automatic selection procedure outperforms both human experts and the heuristic approach of selecting the most accessed documents in the site.