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

The Internet and the Web are continuously evolving giving rise to a rich and extremely dynamic environment where an increasing number of users require and expect new and more sophisticated services. Because of this, a field called “ Web Intelligence” is starting to receive interest from the Artificial Intelligence community. The Web and Internet pose new challenges to AI algorithms, which have been successfully applied in many other fields, at the same time stimulating the development of new techniques. In particular, as pointed out in the introduction to the first part of this special issue (Vol. 4, no. 2, May 2004), machine learning methods have been extensively studied and have been applied to create intelligent systems that are actively involved with the Internet and Web.

The characteristics of the Internet require improvements in the robustness and efficiency of classical learning schemes. The large availability of data which can be thought to be a blessing for effectively building training sets for learning machines, can result in other difficulties such as increased training times. Large amounts of data require efficient sampling schemes so that training sets of reasonable size can be selected, thus reducing training times. The robustness of the algorithms is also an important issue: large amounts of data can contain many irrelevant features and can have more noise than expected. An example

Authors’ addresses: G. W. Flake, http://research.yahoo.com; P. Frasconi, http://www.dsi.unifi.it/~paolo/; C. L. Giles, http://clgiles.ist.psu.edu/; M. Maggini, http://www.dii.unisi.it/~maggini/.

is data from malicious attacks from users. Moreover, though data is simple to obtain, obtaining enough labeled data for supervised learning algorithms is not that straightforward. This issue stimulated the development of learning algorithms that take advantage of both labeled and unlabeled examples.

This special issue covered applications of machine learning to the Internet. Among the 28 submissions received, seven were selected as high quality manuscripts, which were split into two different issues. These articles discuss some of the problems that arise when developing intelligent systems for Internet and Web services. The papers deal with very different issues: information retrieval tasks like question answering or focused crawling on the Web; data mining and personalization for information access such as the prediction of Web navigation patterns and the automatic organization of hyperlinks in a Web portal; collaborative recommendation for automatic marketing in e-commerce applications.

Machine learning has also been applied to many other tasks on the Internet. For instance, network security and monitoring is an important application; intrusion detection systems can exploit learning capabilities.

In This Issue

This issue completes the special issue on Machine Learning for the Internet. The first four papers were published in Vol. 4, No. 2, May 2004. This issue presents three more papers.

"Collaborative Recommendation: A Robustness Analysis" by O'Mahony, Hurley, Kushmerick and Silvestre analyzes collaborative recommendation from a theoretical perspective. A recommendation system can be used in e-commerce sites as a marketing solution; it recommends products to users by mining the data collected from the other users' actions. A product is suggested to a particular user if it was liked by other users having similar characteristics. Unfortunately, such a system might be prone to malicious attacks of users willing to bias the choices towards or away from particular products. The article analyzes when mining algorithms can show robustness in terms of accuracy and stability when data is corrupted by noise, in this case modeling a set of attacks delivered to bias the system decisions. The authors provide a model to evaluate the robustness of the system with respect to the number of fake ratings.

"Topic-Driven Crawlers: Machine Learning Issues" by Menczer, Pant and Srinivasan presents a detailed evaluation and comparison of different algorithms for designing focused Web crawlers. Topical Web crawlers are particularly useful for dealing with scalability problems in Web search engines, especially when coverage and freshness of information are important issues. These agents can be trained to retrieve relevant pages by visiting only the most promising regions of the web graph. The authors address the issue related to the tradeoff between *exploitation* and *exploration* for Web crawlers that use machine learning to guide their search. They propose an extension of the InfoSpider crawler, which uses an evolutionary approach to obtain a good bias between these two behaviors. Exploitation of the quality of the information allows the crawler to concentrate on more promising regions, but at the same time exploration leads the crawler to other regions that might initially seem sub-optimal.

"Market-based Recommendation: Agents that Compete for Consumer Attention" by Bohte, Gerding and La Poutré describes an agent based system that is used to allocate *consumer attention space* for e-commerce applications. The system optimizes the allocation of the available advertising space in order to maximize the utility both for customers and suppliers. Customers need to receive recommendations for finding relevant shops and products in a large marketplace; suppliers would like to advertise only to those customers who are likely to be interested in their products. A central system acts as a trusted third party that assigns the available space by performing an auction where each shop is represented by an agent that bids for each consumer's space. An evolutionary scheme is proposed to adapt the bidding strategy of the agents.