Machine Learning, 39, 99–101, 2000.

Special Issue of Machine Learning on Information Retrieval Introduction

JAIME CARBONELL, YIMING YANG
Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA 15213

WILLIAM COHEN
AT&T Labs—Research, Shannon Laboratory, 180 Park Ave, Florham Park, NJ 07922

As the field of machine learning (ML) has matured, it has reached out to several related fields, both for challenging applications and problems (e.g., robotics), and for new techniques and methods (e.g., statistics and databases). One especially interesting source of both problems and methods for ML is the field of Information Retrieval (IR).

IR has witnessed a boom in activity and attention in the 1990's, fueled in large part on the commercial side by web-based search engines, and in the scientific side by large-scale, repeatable evaluations on common tasks, as illustrated by the TREC, MUC, TDT, and SUMMAC conferences. One important trend in information retrieval has been toward exploration of new tasks. Most early IR research focused on document retrieval for same-language ad-hoc queries. More recent work has explored cross-language information retrieval, in which the documents retrieved are written in a different language than the users' query; automated text categorization; text summarization; fact extraction; topic tracking and detection; and multi-media IR. As part of this trend, IR has necessarily reached out to related communities, such as the natural language processing community, the speech recognition community, and the machine learning (ML) community. ML techniques have been applied to several IR problems, including text categorization, topic tracking and detection, and trainable cross-language IR.

Although IR and ML have both benefited from intellectual cross-pollination in recent years, the ML and IR scientific communities still largely move in different orbits. Different terminology is often used; existing results are sometimes inadvertently re-invented; different evaluation metrics are used; and the two communities often focus on different aspects many problems. For instance, consider the so-called "curse of dimensionality." In ML, learning in high dimensions is usually considered intractable, while IR systems typically deal with dimensions proportional to the size of vocabulary (e.g., 10,000 or more), and with instance spaces that populate these dimensions extremely sparsely. One might think, then, that existing ML techniques would simply be inapplicable to these IR problems. However, in practice, the same ML methods with intractable theoretical worst-case performance on high dimensional problems often work well (with some modification) on IR problems. Investigating why this is so, and how to best exploit and extend such methods, provides fertile

common ground for interdisciplinary collaboration. There are many other opportunities for joint work as well, including text/web mining, unsupervised learning for novelty detection, and multi-strategy learning.

In fact, interdisciplinary collaboration between the IR and ML communities is clearly on the rise. An increasing number of researchers specialize in the intersection of the two fields. There are joint workshops in IR and ML. Representatives of each field attend and publish in the others' conferences. And, more recently, joint publications are providing new venues for scientific exchange. For this special issue we sought high-quality representative papers illustrating potential and actual areas of collaboration. Although a special issue with only four papers is subject to inevitable sampling bias, the papers included here are certainly illustrative of the potential for work at the intersection of IR and ML.

The best-investigated area of interaction in IR and ML is text categorization, a classic supervised inductive-learning task. Two of the papers in this issue focus on different methods for text categorization, making different assumptions about the task domain. Automated extraction of facts from text represents another important area, accelerated by DARPA's TIPSTER program and the Message Understanding Conferences (MUC). Although many approaches to text extraction are not based on learning methods, ML approaches are starting to show good performance on MUC-like tasks. The third paper included in this issue focuses on a pure ML approach to text extraction. Finally, the fourth paper in this issue focuses on ML methods for building web-based search agents that exploit both web-page content and link-topology.

Areas not represented in this issue, but nonetheless fertile grounds for joint IR and ML approaches include:

— topic and event tracking—which is essentially a form of on-line text categorization with very few positive examples;
— new adaptive IR methods based on training language models from the collection and from queries. The language-models are statistically-trained to help predict correlations between queries and documents that may not share actual terms;
— novelty detection (e.g., detecting the onset of a new events of interest in a news-stream and routing related stories to the appropriate users);
— hierarchical text categorization (e.g., for assigning web pages to Yahoo-like ontologies);
— unsupervised clustering, including hierarchical clustering for discovering potentially useful ontologies;
— cross-language information retrieval, tracking and clustering, where ML methods are used to build up cross-language correspondences.

Returning to the four papers in this issue, let us see how they fit into the larger context of IR and ML.

Nigam, McCallum, Thrun and Mitchell investigate the very real problem of sparse training data for building text classifiers. Since the number of potential classes can be very large, it is unrealistic to expect very large numbers of human-labeled documents per class for training. Instead, a small number of human-labeled documents and a much larger set of unlabeled ones more closely matches many potential applications. The question, then, is how to exploit the larger number of unlabeled documents—some of which belong to the desired

class—in order to improve over training with just the small set of labeled documents. The authors demonstrate that using Expectation-Maximization (EM) with Naive-Bayes classifiers works well for this task. They show up to a 30% error reduction in test-set classification when compared to training with just the small number of labeled instances. Future work along this line would include testing the improvement hypothesis with classifiers more powerful than Naive Bayes, such as support vector machines and *k*-nearest neighbors, both of which perform better on sparse training data.

Schapire and Singer introduce a new boosting method which improves the accuracy of text categorization, especially for multi-class problems. Boosting essentially entails reweighting training examples and their category labels to favor those labels which are harder to predict, and then retraining on the reweighted data. The authors show that boosting indeed helps to improve text categorization accuracy on several standard test collections. The boosted method are in fact among the most accurate learning methods, as compared to several other text categorization methods, such as Naive Bayes and Sleeping Experts. New and interesting questions are also raised by the paper as to whether boosting in conjunction with other categorization methods or cross-method boosting could yield even better performance.

Freitag describes several ML approaches for extracting facts from text, especially for text that is not linguistically well formed, such as on-line postings. Text extraction is a somewhat more complex and more structured task than pure classification. Assigning a label to the extracted information is only part of the task; the start and end of each text unit to be extracted must also be correctly identified. Hence, extraction combines segmentation and classification, both of which must be learned concurrently. Freitag evaluates five methods: rote learning, Naive Bayes, a simple grammar induction method, a rule-based relational learner, and a multi-strategy approach which combines the classifiers produced by other four methods. He shows that the multi-strategy method outperforms each of the four basic methods on average. It is significant that a pure learning approach can perform well on this task; one possible step for further research would be to investigate whether a small amount of human assistance can significantly improve performance of an extraction system.

Menczer and Belew describe a set of methods for adaptive retrieval of information from the web, which they call "InfosSpiders" or "adaptive retrieval agents." Web-based retrieval should in principle exploit more than the textual context of web pages; the link topology is also a significant source of information. But, how should both sources of information be combined in an automated and adaptive manner? This is the question that the authors sought to answer, via reinforcement learning, evolutionary adaptation and user-driven relevance feedback. Standard search engines provide the starting point, and then learning local navigation patterns helps to pinpoint useful information. Experiments on a subset of the web show the utility of their novel localized adaptation via learning methods. Unlike the more heavily studied area of text categorization, there are probably many questions in this pioneering area not yet posed that will raise new opportunities for improving the performance of adaptive web-search agents.