Journal of Machine Learning Research 3 (2003) 1023-1024 Published 2/03

Introduction to the Special Issue on Machine Learning Methods for Text and Images

Jaz Kandola JAZ@CS.RHUL.AC.UK
Royal Holloway College
University of London
Egham, Surrey TW20 0EX, UK

Thomas Hofmann TH@CS.BROWN.EDU
Brown University
Providence, RI 02912 USA

Tomaso Poggio TP@AI.MIT.EDU
Massachusetts Institute of Technology
Cambridge, MA 02142 USA

John Shawe-Taylor JOHN@CS.RHUL.AC.UK
Royal Holloway College
University of London
Egham, Surrey TW20 0EX, UK

The amount of unstructured and semi-structured data available in the form of text documents, images, audio, and video files by far exceeds the volume of data stored in relational databases. Yet, in order to optimally utilize this data, it is necessary to devise methods and tools that extract relevant information and support efficient, content-oriented access to it. There is considerable interest, commercial as well as academic, in this domain which has sparked research in machine learning methods for information access. Some of the devised techniques have already led to improved tools that are incorporated in today's information retrieval and knowledge management systems.

This special issue arose from a NIPS 2001 workshop on Machine Learning Methods for Text and Images held at the Whistler Resort, Vancouver, Canada. The aim of the workshop was to present new perspectives and new directions in information extraction from structured and semi-structured data for machine learning. Contributions to the special issue were also open to researchers who had not presented their work at the workshop, and 18 papers were submitted. After rigorous reviewing, only 5 were accepted — an acceptance rate of less than 30% for this special issue. The original workshop was jointly sponsored by *KerMIT*,[1] an EU funded project investigating the use of kernel based methods in the analysis of text and images and *NSF-ITR/IM*[2] an NSF funded project on statistical learning technologies for digital information management search.

The papers in this issue cover a range of topics in machine learning algorithms applied to both text and images. "A Family of Additive Online Algorithms for Category Ranking" by Crammer and Singer describes a new family of topic-ranking algorithms, inspired by online learning algorithms, for multi-labeled text documents. They also provide a unified analysis of the algorithms in the mistake bound model. Extensive experiments results, on the Reuters-21578 and *new* Reuters datasets,

[1] See http://www.euro-kermit.org for more details.
[2] See http://www.ai.mit.edu/projects/cbcl/projects/index-projects.html

highlight the advantages of the proposed topic-ranking algorithms.

“Word Sequence Kernels” by Cancedda et al. addresses the problem of categorizing documents using an extension of the *string kernel* for Support Vector Machines. The string kernel attempts to compute document similarity based on matching non-consecutive subsequences of characters. The approach considered in this paper uses a sequence of *words* rather than characters. Cancedda et al. argue that this approach is not only computationally more efficient it, but that it also has a natural interpretation with standard linguistic pre-processing techniques. They provide extensive experimental evaluation of the approach using the Reuters-21578 dataset.

“Kernel Methods for Relation Extraction” by Zelenko et al. presents an application of kernel methods for extraction relations from unstructured natural language sources. The authors define kernels over shallow parse representations of text, and present a number of efficient algorithms for computing these kernels. The support vector machine and voted perceptron algorithm are trained using these kernels for extracting relations from text. The authors present experimental results for the developed algorithms along with comparisons to other feature-based learning algorithms.

“Matching Words and Pictures” by Barnard et al. deals with probabilistic modeling techniques for multi-modal data sets. More specifically, it presents and discusses a number of related statistical approaches that utilize latent variable models to learn a joint distribution over image regions and corresponding text annotations. As the paper impressively demonstrates, these methods have widespread applications in content-based image indexing, for example, automatically annotating images or image regions with relevant keywords. In addition, the proposed models may be instrumental in designing intelligent tools to structure and browse through multimedia content.

“A Neural Probabilistic Language Model” by Bengio et al. presents a highly innovative approach to language modeling based on neural networks, which overcomes many of the limitations of standard techniques like n-grams. The key idea is to learn a distributed representation for individual words that is able to capture semantic similarities between words. A probabilistic model for word sequences is then expressed and learned in terms of this representation. Both steps are performed simultaneously by optimizing a joint objective (penalized likelihood) over a layered neural architecture.

We feel these papers provide a useful snapshot of the current state-of-the-art techniques being employed for text and image-based problems. The editors thank the reviewers for their valuable help in producing this special issue, as well as Leslie Pack-Kaebling and David Cohn of JMLR for their unflagging support. We would also like to thank the EPSRC, KerMIT and NSF for providing financial support.