Journal of Intelligent Information Systems, 18:2/3, 103–105, 2002


# Guest Editors' Introduction to the Special Issue on Automated Text Categorization

Information retrieval (IR), broadly intended as the discipline concerned with computerized access to data with poorly understood semantics, has grown in the last decade from a relatively small academic specialty to a richly articulated field at the forefront of computer science. The most conspicuous trend of the '90s within IR, aside from the emergence of Web search engines, has been the proliferation of a number of subtasks that depart from the mainstream "text search" paradigm, and tackle information access and use from a larger perspective. Tasks such as automated text categorization, information filtering (or routing), information extraction, text mining, question answering, text summarization, and topic (or event) detection and tracking, are no longer the "next frontier" of IR, but have become its pulsating heart.

Central to the development of these novel (or relatively novel) IR tasks is the application of techniques from machine learning (ML). The key idea is that there is a lot of knowledge out there that can be brought to bear on information access tasks, but this knowledge hardly ever comes in the form of a systematized knowledge base, since such knowledge bases have to be manually built, and for most applications are thus unavailable or too costly to develop. Rather, knowledge manifests itself in the data, and has to be extracted from it; the shift of focus is thus from an "intensional" to an "extensional" notion of semantics, whereby the semantics of a concept is no more a declarative description of it or a set of rules for recognizing its instances, but is the set of contexts (e.g. documents) in which it is instantiated. In this setting, the goal of ML techniques is to help structure data in various respects.

Probably the best known among the disciplines that lie at the crossroads of IR and ML is automated text categorization (ATC), the task of building software tools capable of classifying text documents under one or more of a set of predefined categories or subject codes. ATC dates back to the early '60s, when it was mainly viewed as a means to alleviate the task of indexing scientific literature by controlled vocabulary terms. However, it was only in the early '90s that ATC fully flourished, under the pressure caused by the increased availability of ever larger numbers of text documents in digital form and by the ensuing need to organize them for easier use. ATC is now being applied in many different contexts, ranging from the usual automatic or semi-automatic (i.e. interactive) indexing of texts by controlled vocabulary terms, to personalized information delivery, filtering unsuitable content, Web page categorization under hierarchical catalogues, spoken document categorization, automated meta-data generation, ontology learning, detection of text genre, or even authorship detection for documents of disputed paternity.

From the early '90s the effectiveness of text classifiers has dramatically improved, mainly due to the arrival in the ATC arena of ML methods that are backed by strong theoretical

motivations. Learning methods based on multiplicative weight updating, adaptive resampling (e.g. boosting), or support vector machines, provide a sharp contrast to the relatively unsophisticated and weak methods of the old days. Furthermore, ML researchers have found ATC a challenging application, since datasets consisting of hundreds of thousands of documents and characterized by tens of thousands of terms are widely available. This means that ATC often provides an important benchmark for checking whether a given learning technique can scale up to substantial sizes, and this means in turn that the technological gap between the state-of-the-art in ML and the ML techniques being applied to ATC is becoming increasingly smaller.

This rapid and exciting progress has lead to the growing adoption of automatic or semi-automatic classification systems in applicative contexts in which manual work has been the rule. It is to be expected that this progress will bring about systems of increased cost-effectiveness, will constitute an important testbed for the technologies that are applied to ATC, and will see the increased application of ATC to other advanced IR tasks.

This Special Issue of JIIS brings together 6 contributions selected from 22 papers submitted in response to an open call, and representative of a wide spectrum of ATC techniques and applications.

The paper "A Dynamic Probabilistic Model to Visualise Topic Evolution in Text Streams", by **Ata Kabán and Mark Girolami** (University of Paisley, UK), proposes a new technique for topographically visualizing document streams that develop over time. This technique can be applied, for example, in the analysis of discussion and decision processes. Unlike visualization techniques for static collections, by incorporating knowledge about the coherence of the stream over time, the new method can effectively take into account the sequential dependency structure of the data.

The paper "Latent Semantic Kernels", by **Nello Cristianini, John Shawe-Taylor, and Huma Lodhi** (University of London, UK), explores the question of how to construct kernels (a key component in methods like support vector machines, kernel PCA, etc.) for text. Besides systematically analysing existing representations from IR and relating them to a unified view as kernels, they propose a new method for generating kernels for text based on the latent semantic structure of the term-document matrix. They show how such kernels can be derived from data efficiently and evaluate their performance.

The paper "A Probabilistic Framework for the Hierarchic Organisation and Classification of Document Collections", by **Alexei Vinokourov and Mark Girolami** (University of Paisley, UK), introduces a hierarchical mixture model for documents, extending previous flat models. Vinokourov and Girolami show how their method can derive hierarchic models of a document collection. They also show and evaluate how such models can be exploited as a Fisher kernel in support vector machines.

The paper "PVA: A Self-Adaptive Personal View Agent", by **Chien Chin Chen, Meng Chang Chen, and Yeali Sun** (Academia Sinica, TW), describes an implemented system for personalized management of online reading preferences. The authors especially focus on the question of how to maintain the user model. Since a user's interests typically change over time, it is necessary to continually update the user model and to "forget" outdated preferences. The paper proposes a method for handling such changes in the user's interests based on a structured model.

The paper "Text Categorization for Multi-page Documents: A Hybrid Naive-Bayes HMM Approach", by **Paolo Frasconi, Giovanni Soda, and Alessandro Vullo** (University of Firenze, IT), recognizes that in some situations documents are not isolated entities, but occur in a sequential context and often consist of multiple parts. This is, for example, true in digital library applications, where a document might consist of a sequence of OCR-scanned pages. The authors propose a Hidden Markov Model approach to segmenting and classifying such multi-page documents. A key asset of the model is that it takes advantage of sequence information, exploiting structural regularities present in the page sequence.

The paper "A Study of Approaches to Hypertext Categorization", by **Yiming Yang, Seán Slattery, and Rayid Ghani** (Carnegie Mellon University, US), reports on a number of experiments applying and comparing several representations of hypertext documents previously proposed in the literature, in the context of a hypertext categorization problem. The representations tested here encode different intuitions on the nature and structure of hypertext corpora, and had performed somehow inconsistently in individual experiments previously reported in the literature. The comparative experiments reported here indicate that different hypertext corpora may exhibit different "regularities" (i.e. linkage patterns), and show that picking the best representation technique for a given application requires previous understanding of which among these regularities is present in the corpus.

Many people have contributed to bringing this Special Issue to life. A special word of thank goes to the external referees, whose job was instrumental in providing timely and high quality feedback to the authors. It is thus a great pleasure to acknowledge the help of Ah-Hwee Tan, Alessandro Sperduti, Andreas Rauber, Carlo Strapparava, Daniel Billsus, David Eichmann, David Hull, Dunja Mladenić, Elizabeth Liddy, Georges Siolas, Gerhard Widmer, Harris Drucker, Isabelle Moulinier, Jacques Savoy, James Allan, Jason Rennie, Johannes Fürnkrantz, Joydeep Ghosh, Justin Picard, Kamal Nigam, Ke Wang, Leah Larkey, Marcello Federico, Mark Craven, Marko Grobelnik, Markus Junker, Martin Szummer, Mehran Sahami, Michael Littman, Miguel Ruiz, Nicola Bertoldi, Norbert Fuhr, Ralf Klinkenberg, Ravi Kumar, Soumen Chakrabarti, Stan Matwin, Susan Dumais, Susan Gauch, Sven Schmeier, Thomas Hofmann, Toni Jebara, Umberto Straccia, Vineet Agarwal, Wai Lam, and William Cohen. Above all, we should like to thank Larry Kerschberg, Zbigniew Ras and Maria Zemankova, who originally conceived the project of a Special Issue of JIIS devoted to ATC and invited us to carry it out.

**Thorsten Joachims**
Department of Computer Science
Cornell University
Ithaca, NY, USA

**Fabrizio Sebastiani**
Istituto di Elaborazione dell'Informazione
Consiglio Nazionale delle Ricerche
Pisa, IT