Applications of Machine Learning to Information Access

Mehran Sahami
Gates Building 1A, Computer Science Department
Stanford University
Stanford, CA 94305-9010
sahami@cs.stanford.edu

The recent explosion of on-line information has given rise to a number of query-based search engines (e.g., *Alta Vista*) and manually constructed topic hierarchies (e.g., *Yahoo!*). But with the current rate of growth in the amount of available information, query results grow incomprehensibly large and manual classification in topic hierarchies creates an immense information bottleneck. Therefore, these tools are rapidly becoming inadequate for addressing users' information needs.

We address these problems with a system for topical information space navigation that combines the query-based and taxonomic systems. Our system is built within a unifying probabilistic framework, thereby harnessing the expressive power of this representation while also providing understandable semantics. We are developing a system to automatically create dynamic hierarchical document categorizations based on the full-text of networked articles. Employing clustering technologies, such as AutoClass (Cheeseman *et al.* 1988), we can organize document collections at a more "topical" level. This is similar in flavor to the Scatter/Gather system (Cutting *et al.* 1992), but allows for a variety of disparate, networked information sources to be dynamically accessed. Moreover, this approach is then extended to automatically organize collections of documents into a topic hierarchy, thereby providing a greater contextual organization to the document collection. A prototype of this system has already been implemented as part of the Stanford Digital Libraries testbed (Sahami, Yusufali, & Baldonado 1997).

Furthermore, we have also developed methods for the classification of new articles into such automatically generated, or existing manually generated, hierarchies (Koller & Sahami 1997). In contrast to standard classification approaches which do not make use of the taxonomic relations in a topic hierarchy, our method makes explicit use of the existing hierarchical relationships between topics. Using the hierarchy as a "contextual guide" we are able to achieve higher classification accuracy by solving a series of simple classification tasks rather than treating the problem as one monolithic and very difficult classification task. Much of this improvement is derived from the fact that these simpler decisions at each level of the hierarchy can be made by considering only the presence (or absence) of a small number of features (words) in the document. The selection of relevant words uses a novel information theoretic algorithm we have developed for feature selection (Koller & Sahami 1996). Moreover, the utilization of feature selection to reduce the input space enables us to employ more expressive classification models. To this end, we have developed an efficient method for inducing Bayesian classifiers which do not assume conditional independence of inputs (Sahami 1996), since such an independence assumption can be unrealistic for modelling words in text documents.

The integration of the hierarchical clustering and classification methods will allow large amounts of information to be organized and presented to a user in a comprehensible way, one which is tailored to his or her own particular needs. By alleviating the information bottleneck, we hope to provide users with a solution to the problems of the information access on the Internet.

Acknowledgments. I am indebted to Marti Hearst, Tom Mitchell, Nils Nilsson, and especially Daphne Koller, for their guidance with this research.

References

Cheeseman, P.; Kelly, J.; Self, M.; Stutz, J.; Taylor, W.; and Freeman, D. 1988. AutoClass: a bayesian classification system. In *Proceedings of ML*, 54–64.

Cutting, D. R.; Karger, D. R.; Pederson, J. O.; and Tukey, J. W. 1992. Scatter/gather: a cluster-based approach to browsing large document collections. In *Proceedings of ACM/SIGIR*, 318–329.

Koller, D., and Sahami, M. 1996. Toward optimal feature selection. In *Proceedings of ML*, 284–292.

Koller, D., and Sahami, M. 1997. Hierarchically classifying documents using very few words. In *Proceedings of ML*. To appear.

Sahami, M.; Yusufali, S.; and Baldonado, M. Q. W. 1997. Real-time full-text clustering of networked documents. In *Proceedings of AAAI*. This volume.

Sahami, M. 1996. Learning limited dependence bayesian classifiers. In *Proceedings of KDD-96*.