Real-time Full-text Clustering of Networked Documents

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

Salim Yusufali
Gates Building 1A
Stanford University
Stanford, CA 94305-9010
yusufali@cs.stanford.edu

Michelle Q. W. Baldonado
Gates Building 3B
Stanford University
Stanford, CA 94305-9035
michelle@cs.stanford.edu

With the recent explosion of available on-line information, there is an enormous need for methods that allow users to easily access this information. We address this problem with a system, named SONIA (Service for Organizing Networked Information Autonomously), which enables *topical* information space navigation by employing machine learning to create dynamic document categorizations based on the full-text of articles that are germane to users' queries.

SONIA takes as input a list of document handles (URLs for Web documents) and robustly retrieves and parses the corresponding documents. Documents are then represented as vectors, where each dimension represents a term from the retrieved corpus. Since the number of distinct terms in text is very large, SONIA uses a multi-stage feature selection approach. First meaningless *stop words* (i.e., "the") are eliminated, followed by a Zipf's Law analysis of word occurences that eliminates very infrequent terms. Finally, terms with minimal entropy are eliminated as having insufficient resolving power between documents.

Next, clustering is applied to the resulting vector set. We have used both K-Means clustering (Krishnaiah & Kanal 1982) and AutoClass (Cheeseman *et al.* 1988), which are sufficiently fast to maintain reasonable system interaction time. (Data subsampling can also be employed to speed the clustering process.) The motivation for clustering stems from the observation that documents about similar topics tend to cluster in the document space. A dynamic categorization of the documents is thus created through clustering. For increased interpretability, SONIA also returns for each cluster a small list of characteristic terms. Other researchers (Pirolli *et al.* 1996) have found that such clustering is successful at conveying topical information to users. Moreover, it allows users to quickly identify the document clusters which satisfy their information needs, as well as suggesting additional keywords that might be useful in future queries.

Currently, SONIA is accessed through the *SenseMaker* (Baldonado & Winograd 1997) interface, which can query multiple information sources (Web search engines, DIALOG databases, etc.), and then organize the results itself or ask SONIA to cluster the results. SenseMaker allows users to select clusters to re-cluster thus enabling finer grained distinctions. This captures the Scatter/Gather interaction model (Cutting *et al.* 1992), but allows for a variety of disparate, networked information sources to be dynamically accessed.

As an illustrative example, the query "Mars" was sent to several Web search engines and 45 URLs were then sent on to SONIA, which performed document retrieval, parsing, feature selection and clustering (with AutoClass) in less than 1 minute. The clusters very clearly delineated topical structure within the query results. For example, one exemplary cluster having keywords including "martian, global, surveyor, nasa, pathfinder", captured documents about the NASA Pathfinder mission to Mars, whereas another cluster having keywords including "life, meteorite, planet, nasa, surface", clearly referred to articles regarding evidence of life on Mars in a meteorite found by NASA. Furthermore, on a more disparate topic, one cluster, with keywords such as "cd, rom, images, planet, solar", captured documents selling CD-ROMs with images of planets in the solar system. Other experiments with different queries have yielded larger collections and still led to similar results in reasonable running time.

References

Baldonado, M. Q. W., and Winograd, T. 1997. Sensemaker: An information-exploration interface supporting the contextual evolution of a user's interests. In *Proceedings of CHI.* To appear.

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.

Krishnaiah, P. R., and Kanal, L. N. 1982. *Classification, Pattern Recognition, and Reduction in Dimensionality.* Amsterdam: North Holland.

Pirolli, P.; Schank, P.; Hearst, M.; and Diehl, C. 1996. Scatter/gather browsing communicates the topic structure of a very large text collection. In *Proceedings of CHI.*