Real-time Full-text Clustering of Networked Documents

Mehran Sahami Salim Yusufali Michelle Q. Wang Baldonado
Gates Building 1A, Stanford University, Stanford, CA 94305-9010
Phone: (415) 725-8784 {sahami,yusufali,michelle}@cs.stanford.edu

With the recent explosion of available on-line information, query-based search engines (e.g., *AltaVista*) and manually constructed topical hierarchies (e.g., *Yahoo!*) have proven to be valuable. However, these tools alone are becoming inadequate as query results grow unwieldy and manual classification in topic hierarchies creates an immense information bottleneck.

We address these problems with a system for *topical* information space navigation that combines query-based and taxonomic systems by employing Machine Learning to create dynamic document categorizations based on the full-text of articles that are germane to a user's query. Our system, named SONIA (Service for Organizing Networked Information Autonomously), has been implemented as part of the Stanford Digital Libraries Testbed [Gro95].

SONIA takes as input a list of document handles (generally URLs for Web documents, although other distributed data sources, such as DIALOG are supported) and employs a document retriever (i.e., Web crawler) capable of robust, real-time retrieval of the full text of up to 250 documents in parallel. Upon retrieving documents, SONIA parses the text into alphanumeric terms (i.e., words), and uses this term set to transform textual documents to a vector-based representation. The dimensions of each vector represent terms encountered in the text of the set of all the retrieved documents and feature values are simply the (normalized) count of that term in the given document.

Since the number of distinct terms in text is very large (10^5 for even small collections), feature selection becomes necessary. SONIA uses multi-stage feature selection, using both Natural Language phenomena as well as statistical techniques to successfully reduce the feature set by as much as an order of magnitude or more. Initial feature selection involves *stopword* (non-meaningful term) removal. Next, SONIA employs a feature selection method based on a Zipf's Law analysis of word occurence over the collection of document vectors. Finally, a Term Frequency-Inverse Document Frequency (TFIDF) metric [SB87] is used to eliminate features (terms) that appear too (in)frequently to have much distinguishing power.

Next, a clustering algorithm is applied to the resulting vector set. Currently, we use K-Means clustering [KK82], but in future work we hope to integrate Multiple Cause Mixture Model clustering [SHS96]. The motivation for clustering stems from the observation that documents that are 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. These results allow users to quickly identify the document clusters which satisfy their information needs, as well as helping to determine what additional keywords might be of use in future queries.

Currently, SONIA is accessed through the *SenseMaker* [BW97] 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. As an illustrative sample run, the query "Mars" was sent to several Web search engines and 33 URLs were then sent on to SONIA, which performed its complete document full-text retrieval and clustering in less than 1 minute. This process involved successfully retrieving the full-text from all 33 URLs, parsing the documents into an initial feature set of 1335 terms, reducing the features set size to 669, and then applying K-Means clustering. The clusters very clearly delineated between documents about the planet Mars (with two clusters having keywords such as "global, planet, surface, viking" and "martian, crust, image, landing", respectively), whereas another cluster captured documents about the book *Men Are From Mars, Women Are From Venus* (with the keywords "successful, marriage, principles, understands"). Thus the user could get a very quick handle on the numerous query results. Other experiments with different queries that have yielded collection sizes of over 100 URLs and over 10,000 features have led to similar results, albeit requiring a little more, but still quite reasonable, running time.

SenseMaker also provides a mechanism whereby users can limit a collection of documents to only those clusters that are of interest and then request a re-clustering of only those documents, thus capturing the Scatter/Gather information access model [CKPT92], but in a much more dynamic and broadly distributed fashion. In future work, we seek to allow users to provide relevance feedback [SB90] to the system and thereby extend SONIA to employ *supervised* Machine Learning techniques for feature selection [KS96] and document classification as well.

References

[BW97] Michelle Q. Wang Baldonado and Terry Winograd. Sensemaker: An information-exploration interface supporting the contextual evolution of a user's interests. In *Proceedings of CHI*, 1997. To appear.

[CKPT92] D. R. Cutting, D. R. Karger, J. O. Pederson, and J. W. Tukey. Scatter/gather: a cluster-based approach to browsing large document collections. In *Proceedings of the 15th International ACM/SIGIR Conference on Research and Development in Information Retrieval*, pages 318–329, 1992.

[Gro95] Stanford Digital Libraries Group. The stanford digital libraries project. *Communications of the ACM*, April 1995.

[KK82] P. R. Krishnaiah and L. N. Kanal. *Classification, Pattern Recognition, and Reduction in Dimensionality.* Amsterdam: North Holland, 1982.

[KS96] Daphne Koller and Mehran Sahami. Toward optimal feature selection. In *Machine Learning: Proceedings of the Thirteenth International Conference*, pages 284–292. Morgan Kaufmann, 1996.

[SB87] Gerard Salton and Chris Buckley. Term weighting approaches in automatic text retrieval. Technical Report 87-881, Cornell University Computer Science Department, November 1987.

[SB90] Gerard Salton and Chris Buckley. Improving retrieval performance by relevance feedback. *Journal of the American Society of Information Science*, 41(4):288–297, 1990.

[SHS96] Mehran Sahami, Marti Hearst, and Eric Saund. Applying the multiple cause mixture model to text categorization. In *Machine Learning: Proceedings of the Thirteenth International Conference*, pages 435–443. Morgan Kaufmann, 1996.