Simultaneous Word and Document Clustering

JOHN LAFFERTY PETER VENABLE
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
<lafferty,pvenable>@cs.cmu.edu

Abstract — **In this paper we outline some of our current research on unsupervised learning algorithms for clustering words and documents. The focus of our work is on the use of statistical methods and builds upon previous work on mutual information word clustering, document clustering, and feature selection for text classification.**

I. Introduction

Word and document clustering algorithms play important roles in a wide variety of natural language processing and information retrieval tasks. In current and ongoing research, we are investigating methods for unsupervised, simultaneous clustering of words (or terms) and documents. The motivation for this comes from several different directions. In information retrieval applications, "vector space" methods for document clustering and text classification suffer from sparse statistics, so it is desirable to reduce the dimension of the term space when clustering. We believe that a probabilistic approach may offer certain advantages over more algebraic techniques such as latent semantic indexing [3]. In natural language and speech processing applications, word clustering algorithms tend to yield semantically impoverished classes, so it is desirable to introduce a dependence on topic. We are investigating a unified approach to unsupervised clustering that we consider to be promising for improving the state of the art in both document and word clustering.

II. A Language Modeling Approach

The starting point for our approach is a natural probabilistic model for text generation and a simple observation. Suppose that we group words into disjoint classes $\Phi = \{\phi\}$, and that we similarly group documents into disjoint classes $\Pi = \{\pi\}$. The probability that we assign to a document $d = w_1 w_2 \cdots w_{n_d}$ is expressed in terms of a topic-dependent bigram model:

$$p(d \mid \Pi, \Phi) = \prod_{w_1, w_2} \left(p(w_2 \mid \phi_2)\, p(\phi_2 \mid \phi_1, \pi)\right)^{c(w_1, w_2; d)} \qquad (1)$$

where $\phi_i = \phi(w_i)$, $\pi = \pi(d)$, and $c(w_1, w_2; d)$ is the number of times that the bigram (w_1, w_2) occurs in document d. A little fiddling shows that the log-likelihood $L(\{d\} \mid \Pi, \Phi) = N \sum_d \log p(d \mid \Pi, \Phi)$ of a corpus of documents $\{d\}$ containing a total of N words decomposes as

$$L(\{d\} \mid \Pi, \Phi) = I(\Phi_1; \Phi_2 \mid \Pi) + I(\Phi; \Pi) - H(W) \qquad (2)$$

The first term on the righthand side is the conditional mutual information between adjacent word classes given the class of the document; the second term is the average mutual information between word and document classes; $H(W)$ is the unigram word entropy.

Ignoring the document classes leads to the class-based language model and mutual information criterion for word clustering developed in [2]. When we assume that document classes are given, and use unigrams rather than bigrams, finding good word classes amounts to maximizing the average mutual information $I(\Phi; \Pi)$ between word classes and document labels. This quantity has been explored in recent text classification research [1, 5]. In our work we are carrying out unsupervised clustering of both words and documents in the framework of the general class-based language model.

III. A Greedy Algorithm

To simultaneously cluster words and documents, we are experimenting with natural agglomerative techniques which seek to maximize the mutual information

$$I(\Phi_2, \Pi; \Phi_1) = I(\Phi_1; \Pi) + I(\Phi_2; \Phi_1 \mid \Pi) \qquad (3)$$

One such algorithm begins by first clustering documents using a vector space model where each word is in its own class. Since the statistics are sparse, this clustering is noisy, but on a large corpus it can be expected to give a useful starting point. We then hold these document classes fixed, and perform agglomerative clustering on words, where each merge is chosen to minimize the reduction in mutual information given by (3). Given these word classes, we go back and reclassify the documents, again using an agglomerative algorithm with (3) as the objective function. The algorithm proceeds in this manner, going back and forth between word and document classification, but always using the same objective function.

IV. Work in Progress

We have begun experiments on the Switchboard corpus of telephone conversation transcripts. Since the conversations were restricted to 70 topics, this corpus affords a means of quantitatively evaluating our clustering algorithms. Preliminary results using the mutual information $I(\Phi; \Pi)$ as the objective function have yielded encouraging results. We plan to carry out experiments on other corpora, as well as to investigate the use of distributional clustering techniques [4].

References

[1] L. D. Baker and A. K. McCallum. Distributional clustering of words for text classification. Preprint.

[2] P. Brown, V. Della Pietra, P. de Souza, J. Lai, and R. Mercer. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479, 1992.

[3] S. Deerwester, S. Dumais, T. Landauer, G. Furnas, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[4] F. Pereira, N. Tishby, and L. Lee. Distributional clustering of English words. Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics, pp. 183–190, 1993.

[5] Y. Yang and J. Pedersen. Feature selection in statistical learning of text categorization. Proceedings of the Fourteenth International Conference on Machine Learning (ICML'97), 1997.