Comparing Feature Selection Criteria For Term Clustering Applications

Miles Efron, Junliang Zhang, and Gary Marchionini
{efrom*, march}@ils.unc.edu, junliang@email.unc.edu
School of Information and Library Science
University of North Carolina, Chapel Hill, 27599

ABSTRACT

This poster compares the quality of term clustering results obtained by using two feature selection criteria. Two sets of representative terms were extracted from a corpus of 23,530 documents about labor statistics. The first set consisted of terms with high term-discrimination value as described in [9]. The second set contained the first set's terms plus the 100 most commonly occurring terms (after removal of stop-words). Using each set of terms, *kmeans* clustering was performed, and the validity of the clusters was ascertained by a panel of nine evaluators. We found that high-frequency terms were especially indicative of the corpus' subject matter.

1. INTRODUCTION

Data clustering involves classifying data points (observations) into self-similar groups. A classic unsupervised learning problem, clustering has been employed in service to a variety of data and makes use of a similar variety of statistical approaches (cf. [5, 6]). Term clustering involves the classification of natural language terms into semantically consistent groups that may enhance information retrieval system performance (cf. [8, 1]). This study concerns the problem of feature selection for term clustering. "Feature selection techniques," write Jain *et al.*, "identify a subset of the existing features for subsequent use [in clustering]" [5]. Although feature selection is well-studied in supervised learning applications (cf. [2, 3]), it remains an open question in unsupervised learning. Thus the problem involves discovering in a corpus a set of terms that lead to a useful partitioning of the term space and that characterize the corpus' semantic domain.

This poster compares the quality of term clustering derived from the output of two feature selection rules. In clustering C_1 we include those terms whose document frequency fall within Salton's model of high term-discrimination value [9]. In clustering C_2 we supplement Salton's rule by adding the 100 terms with highest document frequency (after removal of stop words). Thus C_2 contains 100 terms in addition to those defined by C_1. Salton's model is intended for automatic indexing applications, and we hypothesized that supplementing it with high-frequency terms would improve clustering by adding important (but common) terms to the process.

*To whom correspondence should be addressed

2. DATA COLLECTION

The clustering reported here was undertaken as part of the GovStat project (`www.ils.unc.edu/govstat`), an NSF-funded research program intended to improve access to governmental data. Our corpus is the collection of HTML documents that comprised the Bureau of Labor Statistics (BLS) website (`www.bls.gov`) as of December 2002. The corpus contains 23,530 documents. To reduce the term count, we removed all terms that occur on the SMART stop-word list (cf. [1]). To remove numerals and other non-textual material we validated each term, retaining only those words that appear within the WordNet database [4]. This left us with 26,772 terms, to which we applied the Porter stemming algorithm ([7]) to obtain each term's root form.

Having obtained this corpus, our goal is to extract $k < 26,772$ terms that are representative of the corpus as a whole, and that occur with sufficient frequency to enable effective clustering.

3. FEATURE SELECTION CRITERIA

Our first clustering, C_1, uses T_1, the term-document matrix selected by Salton's term discrimination model. Salton's approach retains terms whose document frequency lies on the interval $[\frac{n}{100}, \frac{n}{10}]$, where n is the number of documents. Thus Salton suggests that terms that occur with middling frequency are the best indicators of document content. Applying Salton's model led to a sample of 1882 terms and 15,231 documents, represented as an 1882×15231 matrix of *tf.idf* weights. However, Salton's model was developed in the context of automatic indexing. It thus omits commonly occurring terms, which are poor discriminators between documents, but which may entail globally significant information for the entire corpus.

Thus our second clustering, C_2, uses terms T_2, derived by supplementing Salton's model with the 100 terms that occur in the most documents. Clustering C_2 operates on a matrix with 100 more rows than C_1 (we restrict the documents for C_2 to those that appear in C_1). A sample of terms

annual	state
employment	transportation
metropolitan	wage
national	workforce
production	year

Table 1: Additional Terms under the T_2 Criterion

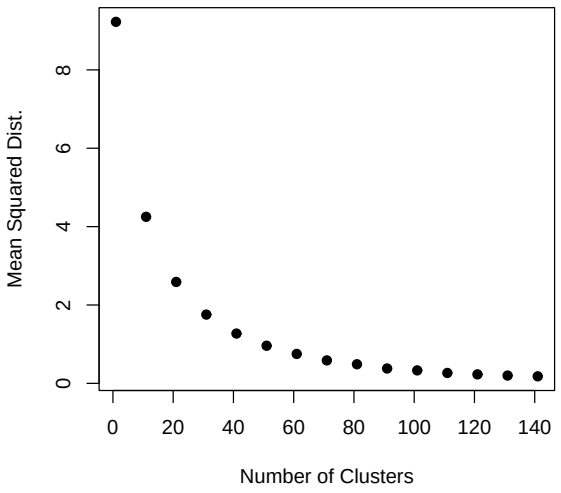

Figure 1: Cluster Density for $k = 10 \cdots 140$

that appear in C_2 but not C_1 are shown in Table 1. These terms are highly indicative of the subject matter of the BLS website, and thus their inclusion in a clustering process is intuitively appealing.

4. CLUSTERING PROCEDURE

Before submitting T_1 and T_2 to the clustering process, we projected each term-document matrix onto the first 100 principal components of its terms. Thus each term was represented as a vector in 100-space. One hundred components were retained by application of the Kaiser-Guttman rule. This projection mitigated the sparse data problem and provided a more robust term-term similarity model for the unsupervised learning task.

To cluster the terms selected by T_1 and T_2 we used the kmeans clustering algorithm (cf. [6]). Kmeans is a partitional clustering procedure wherein the researcher requests k clusters. Given an $n \times p$ input matrix (n observations on p features), kmeans defines k random cluster centroids in p-space. At each iteration the algorithm assigns each of the observation o_i to the cluster k_j that is closest to it (measured by Euclidean distance in our case). The centroid of each cluster is then recomputed by averaging the coordinates of its members, and the data are re-classified. The process stops when the classifications converge on a steady state.

During kmeans clustering the researcher must select k, the desired number of mutually exclusive clusters to derive. Figure 1 shows the within-cluster mean squared distance (MSD) derived for T_1 as we take $k = 10 \cdots 140$ by increments of 10. We selected $k = 80$ due to the decrease in MSD slope near this point. This led to a clustering with many good clusters. However, it also produced spurious clusters. Thus our final clusterings, C_1 and C_2 were derived by retaining those clusters whose within-cluster sum of squared distance was less than the 65^{th} quantile of the cluster sum of squared distances. We also removed all clusters with only a single term. Finally, 2-term clusters were judged for quality manually by the researchers. This left us with 35 clusters containing 293 terms for C_1 and 34 clusters containing 321 terms for C_2.

annual	detail	office	service
area	median	**publish**	**transportation**
code	metropolitan	question	**wages**
construction	number	**research**	**workers**

Table 2: High-Frequency Terms in C_2

5. DATA ANALYSIS

To evaluate the quality of each term selection method we asked a panel of nine domain experts to rate the output of each clustering. For each term selction criterion, raters analyzed each cluster and assigned it an integrity (i.e. self-similarity) grade and a distinctness (i.e. difference from other clusters in the output) grade, where 1 indicates high quality, 0 indicates neutrality, and -1 indicates poor quality. In addition to this, each rater marked the term in each cluster that he deemed the best exemplar of that cluster's semantic content.

A χ^2 test on the number of 1, 0, and -1 ratings incurred under each term selection method suggests that T_2 did not degrade performance over Salton's T_1 approach. For the null hypothesis that including high-frequency terms in the clustering does not affect integrity scores we observed $p = 0.246$. For distinctness, we found $p = 0.43$. Thus including high-frequency terms did not lead to inferior performance by these measurements.

As mentioned in Section **??** we removed many spurious clusters to achieve our final output, C_1 and C_2. An important question is thus how many of the high-frequency terms included during C_2 actually remained in the final output. Of the 100 high-frequency terms, 16 appeared in the final output. These 16 terms appear in Table 2. The eight boldface terms in Table 2 were chosen by one or more of our judges as exemplars of their clusters. Several of these (*area, construction, research, transportation,* and *wages* were exemplars according to all raters, or all raters but one). Clustering C_2 operated on 321 terms, of which 96 were selected as exemplars by one or more judges. Thus the chances of sixteen randomly chosen terms yielding eight exemplars are low ($p = 0.07$), suggesting that these high-frequency terms were judged to be of especially high quality by our raters.

6. CONCLUSION

Our findings suggest that adding high-frequency words to a term discrimination value-based feature selection criterion for clustering merits future work. Using these high-frequency terms did not impede the ability of the kmeans algorithm to partition the term space intuitively, according to our judges. Moreover, the extra terms included in T_2 were often identified as cluster exemplars. This implies that

high-frequency terms convey important information about the semantic domain of a corpus.

The proposed poster will supplement the analysis presented here with a richer description of our experiences in deriving high-quality terms for the clustering problem. In particular, we will focus on the problem of labeling individual clusters. We found a high degree of overlap among our human judges' choices for exemplary terms. However, this consensus was difficult for automated methods to anticipate. Moreover, we found that in many cases, no member term exemplified a cluster's content. For instance, a cluster of month names has a clear semantics despite lacking a member term that summarizes its domain. The full version of this poster will explore this problem and report several means of addressing it.

7. REFERENCES

[1] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press, 1999.

[2] C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford, 1995.

[3] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. Wiley, 2nd edition, 2001.

[4] C. Fellbaum. *Wordnet: An electronic lexical database*. MIT Press, 1998.

[5] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323, September 1999.

[6] Leonard Kaufman and Peter J. Rosseeuw. *Finding Groups in Data: an Introduction to Cluster Analysis*. Wiley, 1990.

[7] M. F. Porter. An algorithm for suffix stripping. In *Program*, pages 130–137, 1980.

[8] C. J. Van Rijsbergen. *Information Retrieval*. Butterworths, 2nd edition, 1979.

[9] G. Salton and C. T. Yu. A theory of term importance in automatic text analysis. *Journal of the American Society for Information Science*, 26(1):33–44, September 1975.