Implications of the Recursive Representation Problem for Automatic Concept Identification in On-line Governmental Information *

Miles Efron, Gary Marchionini, and Julinang Zhiang
School of Information and Library Science
CB#3360, 100 Manning Hall
University of North Carolina
Chapel Hill, NC 27599-3360
{efrom, march}@ils.unc.edu, junliang@email.unc.edu

October 28, 2003

Abstract

This paper describes ongoing research into the application of unsupervised learning techniques for improving access to governmental information on the Web. Under the auspices of the GovStat Project (http://www.ils.unc.edu/govstat), our goal is to identify a small number of semantically valid and mutually exclusive "concepts" that adequately span the intellectual domain of a web site. While this is a classic instance of the clustering problem [14] the task is complicated by the dual-representational nature of term-document relationships. Since documents are defined in term-space and vice versa, we may approach this as a document- or term-clustering problem. The current study explores the implications of pursuing both term- and document-centered representations. Based on initial work, we argue for a document clustering-based approach. Describing completed research, we suggest that term clustering yields semantically valid categories, but that these categories are not suitably broad. To improve the coverage of the clustering, we describe a process based first on document clustering.

1 Introduction

The GovStat Project is a joint effort of the University of North Carolina Interaction De-

*This research was supported by NSF EIA grant 0131824.

sign Lab and the University of Maryland Human-Computer Interaction Lab. Citing end-user difficulty in finding governmental information online (especially statistical data), the project seeks to create an integrated model of user access to US government statistical information that is rooted in realistic data models and innovative user interfaces. To enable such models and interfaces, we propose a data-driven approach, based on unsupervised machine learning techniques. In particular, our work analyzes a given website (e.g. http://www.bls.gov, the Bureau of Labor Statistics' site) in efforts to discover a small number of linguistically meaningful, mutually exclusive "concepts," or "bins," that collectively summarize the semantic domain of the site.

The project goal is to classify the site's web content according to these inferred concepts as an initial step towards data filtering via active user interfaces (cf. [15]). Governmental websites already make use of content classification, both explicitly and implicitly; they divide their resources manually by topical relation; they organize web content into hierarchically oriented file systems. The goal of the present research is to develop yet another means of browsing the content of these sites. By analyzing the distribution of terms and the relationships among documents, our goal is to supplement (not replace) the agencies' pre-existing information structures. Statistical learning technologies are desirable in this context insofar as they stand to define a data-driven—as opposed to an agency-driven—navigational structure for a site.

This goal is distinct from supervised learning approaches to document organization that attempt to classify documents into appropriate categories. Our task is less structured; before classifying documents, we must define the classes that exist. The spirit of the work is in close step with the automatic discovery of facets described by Anick in [1]. Insofar as we aim to identify terms (words or phrases) that capture the high-level semantics of an entire site, our work appears to be a classic instance of the data clustering problem (cf. [14]): given n documents in p terms we seek the k terms (where $k \ll p$) that are maximally expressive, according to some criterion.

However, the task is complicated by the nature of web data, where documents and terms are recursively defined (cf. [18, 6]). Unlike more traditional clustering applications, in text analysis it is unclear what constitutes a variable and what is an observation. Speaking geometrically, terms are represented in document space, and vice versa. To discover semantically valid "bins" in a website, we might select the k terms closest to the centroids derived by *k-means* clustering carried out on terms in m-dimensional document space. Alternatively, we might derive k document clusters from the n-dimensional term space, subsequently selecting representative labels for each document cluster.

The remainder of this paper pursues the implications of the fluidity of perspective inherent in concept extraction via text clustering. In Section 2 we formalize our terms, detailing the types of evidence and operations that inform term- or document-based concept discovery. Next we turn to a description of completed research into bin identification based on term clustering. Section 4 describes our current efforts, which employ a document-centric approach. Finally, we conclude in Section 5 by outlining upcoming work on the project.

2 Dual Representations for Text Clustering

Let $\mathbf{A}$ be the $n \times p$ data matrix with n observations in p variables. Thus a_{ij} shows the measurement for the i^{th} observation on the j^{th} variable. As described in [14], the goal of cluster analysis is to assign each of the n observations to one of a small number k mutually exclusive groups, each of which is characterized by high intra-cluster correlation and low inter-cluster correlation. Though the algorithms for accomplishing such a partitioning are legion, our analysis focuses on so-called *k-means* clustering[1], during which, each observation o_i is assigned to the cluster c_k whose centroid is closest to it, in terms of Euclidean distance. The model is fitted by minimizing the sum of squared error (sse), given in Equation 1:

$$sse = \sum_{k=1}^{K} \sum_{i=1}^{|k|} = \|x_{ik} - \overline{x}_k\|^2 \qquad (1)$$

where x_{ik} is the i^{th} observation in the k^{th} cluster, and $\overline{x}_k$ is the centroid of the k^{th} cluster.

Medioid-based clustering by *k-means* is well-studied in the statistical literature, and has shown good results for text analysis (cf. [12, 18]). However, applying *k-means* to concept extraction from Web documents demands that we confront the recursively defined notions of documents and terms in common information retrieval models. Let $\mathbf{A}$ be the $n \times p$ data matrix for a corpus containing n documents in p terms. Here we consider each document as a p-dimensional vector of term observations. We may submit this matrix $\mathbf{A}$ for clustering, in which case we assign each document to a given class based on the words that it contains. Alternately, we may begin with the transpose of the problem, wherein we cluster the $p \times n$ matrix $\mathbf{A}'$, yielding a clustering of terms.

Clustering terms in document space bears an obvious appeal for the task at hand; we desire a small number of terms that describe the most general topics in our corpus. By assigning each of the p terms to a cluster and choosing for each cluster c_k the term t_k that is closest to its centroid $\overline{x}_k$, we achieve a principled set of artificial "concepts." In Section 3 we describe initial results obtained in this manner.

On the other hand, clustering documents in term space provides an indirect, but perhaps more viable route to artificial concept extraction from corpora of Web data. To acquire the desired key terms, a document-centric approach demands a two-stage process. First, we must cluster the documents. Having grouped the documents into

[1] We have focused on *k-means* as opposed to hierarchical clustering for several reasons. Chief among these is the computational efficiency enjoyed by the *k-means* approach. Because we need only a "flat" clustering there is little to be gained by the more expensive hierarchical algorithms.

k clusters, we then select the most representative term (or terms) from each cluster. The advantages of a document-centric approach lie in its potential introduction of extra-linguistic information into the clustering problem. That is, by clustering Web documents, we are able to use not only term co-occurrence information, but also information supplied by hyperlink analysis, and conceivably by information culled from external sources such as Web server logs. In Section 4 we report on the current state of our research into such a document-centered clustering approach.

3 Term Clustering: a Pilot Study

Although the aegis of the GovStat project spans several governmental agencies, our work on text clustering initially focused on the website of the Bureau of Labor Statistics. As of our web crawl of December 2002, this site contained 23,530 documents. Based on this initial corpus, our goal was to extract $k \approx 15$ key terms for use in high-level information filtering and browsing.

To reduce the term count prior to analysis, we removed all terms that occur on the SMART stop-word list (cf. [2]). Removing numerals and other non-textual material, we validated each term by retaining only those words that appear within the WordNet database [8]. This left us with 26,772 terms, to which we applied the Porter stemming algorithm ([17]) to obtain each term's root form.

3.1 Variable Selection via Salton's Term Discrimination Model

Faced with a $26,772 \times 23,530$ matrix, we initially sought to identify a subset of "important" terms as a means of limiting the problem and reducing noise in the system prior to clustering. We pursued two methods for identifying putatively important terms for clustering purposes. Our first clustering used 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. 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, but which may entail globally significant information for a corpus. Thus our second term-based clustering used term-document matrix T_2, derived by supplementing Salton's model with the 100 terms that occur in the most documents. The decision to augment the classic term discrimination model by precisely 100 terms was settled by inspection of the ranked list of all terms. Those terms with rank much greater than 100

Table 1: Additional Terms under the T_2 Criterion

annual	state
employment	transportation
metropolitan	wage
national	workforce
production	year

were deemed to have little semantic weight by the researchers. In future work, such a heuristic will of course be studied further.

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 that appear in C_2 but not C_1 is 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.

The terms in Table 1 are omitted by the classic term discrimination model because their frequency makes it unlikely that they will appear in any focused subset of the collection. In other words, they are too general to be good discriminators. However, it is precisely their generality that makes them attractive for the concept extraction task. Because they are globally significant to the corpus, their role in describing the corpus' most high-level information is likely to be high.

3.2 Procedure for Term Clustering

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 (cf. [13]). One hundred components were retained by application of the Kaiser-Guttman rule [11]. This projection mitigated the sparse data problem and provided a robust term-term similarity model for the clustering.

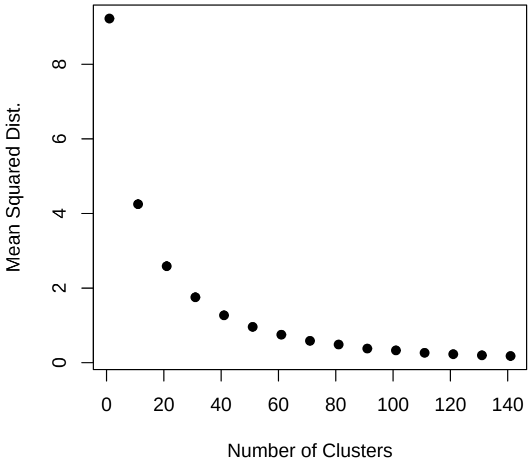

Figure 1: Mean squared distance for $k = 1 \ldots 140$

Having projected the terms onto the first 100 principal components, we submitted the data to *k-means* clustering, retaining $k = 80$ clusters. We selected $k = 80$ after inspecting the change in mean-squared within cluster distance for many values of k.

Figure 1 plots MSD, the average within-cluster distance (i.e. the average error) for a variety of k-dimensional models. For $k > 80$ the change in MSD by adding another cluster is subtle, suggesting that setting $k = 80$ adequately captures the orientation of our term space.

Setting $k = 80$ 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 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 by the researchers. This left us with 35 clusters containing 293 terms for C_1 and 34 clusters containing 321 terms for C_2.

3.3 Comparing Term Selection Criteria

To evaluate each term selection method we asked a panel of nine domain experts to rate the output of each clustering. For each term selection criterion, raters analyzed each cluster and assigned it an integrity (i.e. self-similarity) grade and a distinctness (i.e. difference from other clusters) grade, where 1 indicates high quality, 0 indicates neutrality, and -1 indicates poor quality. In addition to this, each rater marked the term that he or she deemed the best exemplar of each 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 adding high-frequency terms to the clustering does not affect integrity we observed $p = 0.246$. For distinctness, we found $p = 0.43$. Thus including high-frequency terms did not lead to inferior performance measurements.

Table 2: High-Frequency Terms in C_2

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

As mentioned earlier we removed spurious clusters to achieve our final output, C_1 and C_2. An important question, then, is how many of the high-frequency terms included during C_2 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 boldface terms in Table 2 were chosen by one or more of our judges as exemplars of their clusters. Five of these (*area, construction, research, transportation,* and *wages*) were exemplars according to at least eight raters. Clustering C_2 operated on 321 terms, of which 96 were selected as exemplars by one or more judges. The chances of sixteen randomly chosen terms yielding eight exemplars are low ($p = 0.07$), suggesting that these high-frequency terms were of especially good quality for our raters. Although this work is preliminary, we count this as an encouraging sign that high-frequency terms bear further scrutiny, and probable inclusion during term clustering, despite their notoriously weak utility as discriminators during IR.

4 Current Work: Document Clustering

Performing *k-means* clustering on terms did appear to yield self-consistent, mutually orthogonal clusters of terms. However, the method described above suffered a serious defect, which we call the "fallout" problem. That is, by discarding so many terms—both during the initial variable selection by Salton's model, and by discarding spurious clusters—the remaining clusters covered only a small part of the BLS website's semantic domain. Thus, while concepts such as **education**, **time**, and **benefits** were well represented by the clusters that we derived, issues such as **inflation** and **productivity** happened to drop out. Thus if we tried to abstract from each cluster a label, such that these labels could inform high-level browsing and information filtering, our set of labels would fail to provide avenues into large portions of the collection.

To remedy this problem—and in efforts to improve the clustering in general—our current efforts are focused on the transpose of the task described in Section 3. Currently, we are involved in a two-stage approach to concept extraction that is predicated on a document-centric clustering.

Clustering documents for automatic concept identification is desirable insofar as it allows the process to draw on more information than a strictly term-based approach admits. In particular, our current efforts supplement an analysis of document-term relationships with information about the inter-document hyperlink structure. Thus the document-based approach synthesizes a variety of sources of evidence in service to automatic concept extraction.

In addition to $\mathbf{A}$ the $n \times p$ document-term matrix (n documents in p terms) described above, we also define $\mathbf{AA}'$, the $n \times n$ document co-occurrence matrix. If a_{ij} is binary—recording 1 if term j appears in the i^{th} document, 0 otherwise—then $\mathbf{AA}'$ gives the number of common terms shared by each pair of documents[2].

We supplement the document co-occurrence matrix with hyperlink information prior to clustering. Let $\mathbf{M}$ be the $n \times n$ adjacency matrix for the corpus. Thus $m_{ij} = 1$ if the i^{th} document contains a hyperlink pointing to document j. Otherwise $m_{ij} = 0$. Given matrix $\mathbf{M}$, we also define the cocitation matrix $\mathbf{M'M}$. The i, j^{th} element of this matrix gives the number of documents that contain hyperlinks to documents i and j.

We thus obtain three sources of information for the document-based clustering. We have $\mathbf{AA}'$, the document co-occurrence matrix. Additionally we have, $\mathbf{M}$ and $\mathbf{M'M}$. Each of these matrices is $n \times n$. Thus our starting point for the clustering process is matrix $\mathbf{D}$, given in Equation 2:

$$\mathbf{D} = \alpha\mathbf{AA}' + \beta\mathbf{M} + \gamma\mathbf{M'M} \qquad (2)$$

[2] In actual application $\mathbf{A}$ captures *tf.idf* scores.

where α, β, and γ are tunable parameters. The final input to the *k-means* would thus be the projection of each document onto the first principal components of $\mathbf{D}$.

Web page clustering based on a fusion of textual and hyperlink-based analysis has been explored to good effect by He *et al.* in [9]. However, He *et al.* pursue the matter in service to categorization of search results on the Web at large. Although our task—automatic concept extraction from a focused domain—is different than theirs, we hypothesize that a fusion of textual and hyperlink information will yield a useful clustering of documents. This hypothesis is also borne out by the successes of fusion-based approaches to query-specific information retrieval (cf. [3, 4, 7]).

Unlike the more direct term clustering approach described in Section 3, the document-centric technique yields k sets of putatively similar documents. Thus the extraction of exemplary terms remains as a final stage in the process. That is, for each document cluster C_k, we wish to identify a label, a small set of terms that are characteristic of documents in the cluster. To generate these labels, we turn to information theory. Having partitioned the documents into k disjoint sets, we identify the most indicative terms for a given cluster C_k by calculating the the information gain for each term t_i with respect to cluster C_k. Let D_k be the set of documents assigned to cluster C_k. Also let D'_k be all documents not in the cluster (i.e. in all other clusters). Following [16], we define the information gain of term i on cluster C_k:

$$G(t_i, C_k) = H(C) - H(C|t_i) \tag{3}$$

where where $H(C)$ is the Shannon entropy of the class variable, and $H(C|t_i)$ is the specific conditional entropy between the two variables [5]. To choose labels for cluster C_k, we choose the p terms with the highest information gain for the cluster.

5 Future Work

We hypothesize that a document-based approach to text analysis will yield superior results to the method based on an explicit clustering of terms. The already formidable problem of concept extraction is made all the more difficult in large applications by the sparseness of textual data, and its attendant violations of common statistical assumptions (e.g. the assumptions of normality inherent in linear models). By supplementing term-document co-occurrence information with data derived from hyperlink analysis, we hope to gain additional insight into the document-document correlations at work in a given corpus. And since terms and documents are inherently interrelated, we anticipate that this improved model of inter-document similarity will lead to superior clustering and, ultimately, to superior extraction of key terms.

Most immediately, then, our efforts are focused on testing this hypothesis. We have currently put in place software to generate the matrices called for in Equation 2. After the requisite pre-processing, we aim to cluster the BLS data described in Section 3. Our goal then will lie in comparing the quality of the clusters and their labels obtained via each method. To accomplish this, we anticipate utilizing a combination of user-based and statistical approaches.

What is especially important to note, however, is the close relationship between the term and document clustering problems. In an abstract sense, each task involves modelling the correlational structure of the data. That is, in both cases, we infer inter-item similarity by analyzing patterns of variable occurrence across observations. Because the starting point of each approach is the term-document matrix, the attendant correlation matrices (i.e. of terms and of documents) are intimately related.

Projecting the data onto their principal components reflects our underlying concern with modelling correlational tendency. That is, by retaining, say, the first 100 principal components, we obtain the best 100-dimensional model of inter-variable correlation, in the least squares sense (cf. [13]). The crucial point is that the eigensystems of term and document co-occurrence matrices are deeply related. A well known result from linear algebra (cf. [19]) holds that the eigenvalues of these matrices are equal. Equation 4 gives the singular value decomposition of the $n \times p$ matrix $\mathbf{A}$:

$$\mathbf{A} = \mathbf{T\Sigma D} \tag{4}$$

where matrices $\mathbf{T}$ and $\mathbf{D}$ contain the eigenvectors of the term and document co-occurrence matrices, respectively, while $\mathbf{\Sigma}$ is diagonal, giving the positive square roots of the co-occurrence matrix eigenvalues. By definition, then, the first principal component of the terms captures the same amount of variance as the first document principal component. Thus we hypothesize that adopting a document-centric approach to clustering will still give us implicit access to term-term co-occurrence information, while simultaneously allowing the analysis to include extra-linguistic information, such as hyperlink data.

In future work we plan to experiment with alternative models of correlational structure. In particular, we are currently implementing a clustering algorithm that operates not on the principal components of the data, but rather on their so-called independent components ([10]), an approach to correlational modelling that avoids the problematic assumption of normality common to least-squares methods.

Regardless of the algorithm used to model these relationships, however, the problem of concept extraction via clustering reminds us keenly of the unique problems inherent in applications of data mining technologies to natural language text. Not only does text defy common parametric assumptions, but linguistic data also point out the mutually reinforcing relationships between observations and variables. Ultimately this fluidity allows the researcher an unusual degree of liberty in analysis, by enabling the sorts of synthesis and data fusion that inform our current analysis.

References

[1] P. G. Anick. *Automatic Construction of Faceted Terminological Feedback for Context-Based Information Retrieval.* PhD thesis, Brandeis University, 1999.

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

[3] K. Bharat and M. Henzinger. Improved algorithms for topic distillation in a hyperlinked environment. In *Proceedings of the 21st Annual Conference on Research and Development in Information Retrieval (SIGIR 1998)*, New York, 1998. ACM.

[4] S. Chakrabarti et al. Mining the link structure of the world wide web. *IEEE Computer*, August 1999.

[5] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* Wiley, 1991.

[6] C. H. Q. Ding. A similarity-based probability model for latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1999.

[7] P. Domingos and M. Richardson. The intelligent surfer: Probabilistic combination of link and content information in PageRank. In *Advances in Neural Information Processing Systems 14*, volume 14, 2002.

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

[9] X. He, H. Zha, C. H. Q. Ding, and H. Simon. Automatic topic identification using webpage clustering. In *Proceedings of the IEEE International Conference on Data Mining*, pages 195–202. IEEE, 2001.

[10] A. Hyvarinen, J. Karhunen, and E. Oja. *Independent Component Analysis.* Wiley Interscience, 2001.

[11] J. E. Jackson. Stopping rules in principal components analysis: A comparison of heuristical and statistical approaches. *Ecology*, 74:2204–2214, 1993.

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

[13] I. T. Jolliffe. *Principal Component Analysis.* Springer, 2 edition, 2002.

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

[15] G. Marchionini, S. Haas, C. Plaisant, B. Shneiderman, and C. Hert. Toward a statistical knowledge network. In *Proceedings of the National Conference on*

Digital Government Research, pages 27–32, Boston, 2003. National Science Foundation.

[16] T. Mitchell. *Machine Learning.* McGraw Hill, 1997.

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

[18] E. Rasmussen. Clustering algorithms. In W. B. Frakes and R. Baeza-Yates, editors, *Information Retrieval: Data Structures and Algorithms*, pages 419–442. Prentice Hall, 1992.

[19] G. Strang. *Linear Algebra and its Applications.* International Thompson Publishing, 1988.