Comparing Dimension Reduction Techniques for Document Clustering

Bin Tang, Michael Shepherd, Malcolm I. Heywood, Xiao Luo

Faculty of Computer Science,
Dalhousie University, 6050 University Avenue,
Halifax, Nova Scotia, Canada, B3H 1W5
{btang, shepherd, mheywood, luo}@cs.dal.ca

Abstract. In this research, a systematic study is conducted of four dimension reduction techniques for the text clustering problem, using five benchmark data sets. Of the four methods -- Independent Component Analysis (ICA), Latent Semantic Indexing (LSI), Document Frequency (DF) and Random Projection (RP) -- ICA and LSI are clearly superior when the k-means clustering algorithm is applied, irrespective of the data sets. Random projection consistently returns the worst results, where this appears to be due to the noise distribution characterizing the document clustering task.

1 Introduction

Document clustering is a fundamental and enabling tool for efficient document organization, summarization, navigation and retrieval. The most critical problem for document clustering is the high dimensionality of the natural language text, often referred to as the "curse of dimensionality". While various dimension reduction techniques (DRTs) have been proposed [1, 2], there are two major types, feature transformation and feature selection [2]. Feature transformation methods project the original high dimensional space onto a lower dimensional space, while feature selection methods select a subset of "meaningful" dimensions from the original ones.

In this research, we compare DRTs in a systematic manner for the text clustering task. We investigate the relative effectiveness and robustness of four dimension reduction techniques; one feature selection method, Document Frequency (DF) [3], and three feature transformation methods, including Latent Semantic Indexing (LSI) [4], Random Projection (RP) [5] and Independent Component Analysis (ICA) [6].

This paper is organized as follows. Section 2 describes our data sets, Section 3 our experimental procedure and evaluation methods, Section 4 our results and Section 5 presents our conclusions and directions for future research.

2 Characteristics of the data sets

We used five data sets used widely in information retrieval and text mining research. The number of classes ranges from 4 to 50 and the number of documents ranges between 4 and 3807 per class. The WebKB4 data set consists of WWW-pages. Reuters-2 and Reuters-10 are derivatives of the Reuters-215781 newswire stories data set. Reuter-2 is a collection of documents, each with a single topic label. The version of Reuter-2 that we used eliminates categories with less than 4 documents, leaving only 50 categories. We derive Reuters-10 from Reuters-2, consisting only of the ten most frequent categories. 20NG-4 is a subset of the 20-Newsgroup data set, and only includes 4 categories. The fifth data set consists of technical reports (CSTR). Details of the datasets are found elsewhere [7]

The data sets were pre-processed to remove tags, non-textual data and stop words[1]. The remaining words were stemmed[2] and those stems with low document frequency were removed. For example, the cutoff for the Reuter-2 data set is 4. The stem-weighting scheme used is the most commonly used "*ltc*" variant of the *tfidf* function [8].The document vectors were then normalized to unit length.

3 Experimental Procedures and Evaluation

3.1 Experimental Procedures

All experiments were conducted in the Matlab 6.5.1 environment. Each data set was split into training and test sets in a ratio of 3:1. For each number of experimental dimensions, the reduced dimension version of the data was generated as per each of the DRTs. This data was renormalized to unit length for each document, and k-means clustering was applied, to generate the clusters using only the training set. The choice of k is *ad hoc*, larger than the number of classes in general. Each cluster was given a class label using majority voting (using training set) and the classification accuracy (only for test set) was determined as described below.

At each number of reduced dimensions, the seeds for the k-means clustering were generated for each data set by sampling the data set and finding the set of points around which the rest of the sample are tightly grouped (details in [7]).

3.2 Evaluation methods

To judge the relative effectiveness of the DRTs, we apply them to text clustering tasks on different data sets. Based on the quality of their clustering results, we rank them

[1] http://www.dcs.gla.ac.uk/idom/ir_resources/linguistic_utils/stop_words
[2] http://www.tartarus.org/~martin/PorterStemmer/

accordingly. There are two perspectives to the ranking, the absolute clustering results and the robustness of the method. Here, good robustness implies that when using a certain DRT, reasonably good clustering results should be found across a relatively wide range of dimensions (reduced), i.e., the clustering results should degrade gracefully if non-optimal reduced dimensions are used.

To measure the quality of text clustering, we choose to use *Purity* as introduced in [9]. We modify the calculation of *Purity* as follows. Each cluster i is assigned a class label, T_i, based on a majority vote by its members using only the training set. Then, the purity of cluster i is defined as the proportion of points assigned as members of cluster i in the test set whose class labels agree with T_i. It is easy to establish that *Purity* is the clustering-version of the micro-average of classification accuracy. Hereafter, we refer to the cluster quality measure as *classification accuracy (CA)*.

To judge the relative robustness of DRTs, we combine a heuristic observation and student t test. We first plot the CA curves of the DRTs against the dimensionalities. Based on the CA values, it is visually possible to clearly establish the relative effectiveness of the DRTs based on these curves. For situations when more than one curve shares very similar CA values over "an interesting range of dimensions" (defined later), such that we cannot visually resolve performance levels, we perform a paired student t test. For each data set, the relative ranks of the DRTs are determined by the combination of visual observation and paired student t tests on the CA curves of the DRTs.

To ensure that the results are representative and systematic, many precautions have to be taken in the process of comparison. First, the choice of data sets has to be made in such way that a broad genre of text collections is covered in our test. The second issue concerns the usage of the clustering algorithm. We choose to use k-means, since k-means or its variants are the most commonly used clustering algorithms used in text clustering. A well-known problem for k-means is that poor choices of initialization often lead to poor convergence to sub optimal solutions. To ameliorate the negative impact of poor initialization, we devised a simple procedure, described in Section 3.1.

4 Experimental Results and Discussions

4.1 Comparisons of the four DRTs

The DRT comparisons over each data set are conducted by the combination of visual inspection and paired student t tests. We are only interested in comparing their performance on the most "interesting" dimension range. By "interesting" dimension range, we refer to the dimension range within which the methods produced the best clustering results. Hereafter, we will use [a, b] to denote the "interesting" dimension range under investigation. To detect the "good range of reduced dimensions", we also plot the LSI performance against its singular values. Since ICA uses PCA as a pre-

processing stage to "whiten" the raw data and determine the number of components (dimensions) to reduce to, we are also interested in the correlation between ICA performance and the number of eigenvectors (number of reduced dimensions) used in the PCA whitening step. This correlation may suggest how to determine the "good range of dimensions to reduce to" by ICA.

Due to space limitations, we cannot present all our results, which can be found elsewhere [21]. Since the DRTs show very similar performances over the 5 data sets, we only present the results on Reuter-2 in detail (Figure 1).

From our results we can make a number of observations. RP is inferior to DF for the whole range of dimensions being investigated. DF peaks around a dimension of 657 with CA of 0.85 and then settles around 0.8 with increasing dimensionality. ICA and LSI achieve their best results with lower dimensionalities ([30, 93]). The result of the t-test indicates superior performance of ICA over that of LSI. ICA also maintains very good performance over a much larger range of dimensions than LSI and, therefore, appears to be more robust.

The correlation between singular values and LSI performance (and eigenvalues and ICA performance) is not clear. We observe that, both the singular and eigen values decrease very rapidly within the first few to few tens of dimensions, after which there is general reduction. Hereafter, we refer to the part of the singular/eigen value curve that transits from very rapid reduction to slow reduction as the transition zone. This transition zone seems to correspond to the best performance of LSI/ICA. In all cases, it appears that over the transition zone, the CA curve of ICA reaches its peak and keeps at a constant level over a wider range of dimensions than that of LSI, indicating less feature sensitivity of ICA. For Reuter-2, considering all the factors, we rank the DRTs in the order of ICA > LSI > DF> RP, where ">" denotes better.

5 Conclusions and Future Work

In this research, we compared four well-known dimension reduction techniques, DF, RP, LSI and ICA, for the document clustering task using five benchmark data sets. In general, we can rank the four DRTs in the order of ICA >LSI >DF >RP. ICA demonstrates good performance and superior stability compared to LSI. Both ICA and LSI can effectively reduce the dimensionality from a few thousands to the range of 100 to 200 or even less. The best performances of ICA/LSI seem to correspond well with the transition zone of the eigen/singular value curve. The experiments with DF clearly indicate to us that most of the raw dimensions in the text data are very noisy and meaningless with respect to the document clustering task, which further explains the relatively poor performance of RP.

Our future research includes comparing the semantic meanings of the latent variables derived from ICA and LSI, and using DF to pre-screen the raw dimensions for LSI/ICA to further reduce the computational cost of LSI/ICA.

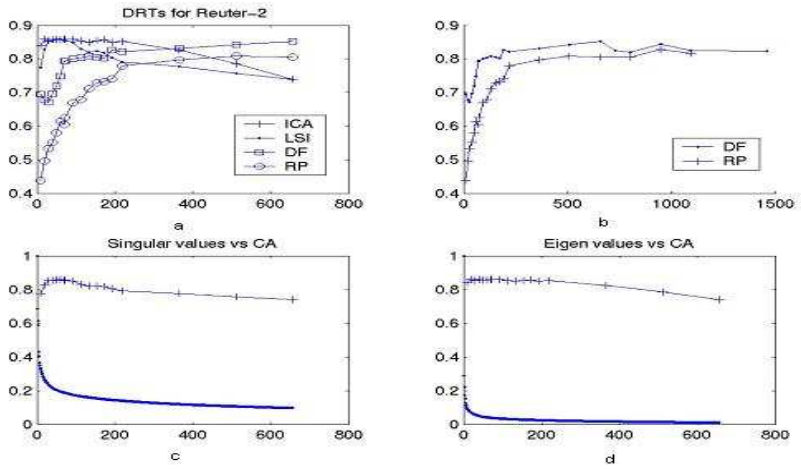

Fig. 1. DRT performance summary for Reuter-2.a. parallel comparison of four DRTs, x-axis: dimensionality, y-axis: CAs for DRTs. b. comparisons between DF and RP with extended dimensionality. c. correlation of classification accuracy and normalized singular value for LSI, '+' denotes the CA curve and '.' denotes the normalized singular value. d. correlation of classification accuracy of ICA and the normalized eigenvalues of its PCA step, '+' denotes the CA curve and '.' denotes the normalized eigenvalues

References

1. Fodor I.K.: A survey of Dimension Reduction Techniques. LLNL technical report, UCRL ID-148494, (2002) URL: http://www.llnl.gov/CASC/sapphire/pubs.html
2. Parsons L. et al.: Subspace Clustering for High Dimensional Data: a Review. ACM SIGKDD Explorations Newsletter**.** 6(1) (2004) 90 - 105
3. Yang Y. Pedersen J. O.: A Comparative Study on Feature Selection in Text Categorization. Proc. ICML (1997) 412-420
4. Berry M.W., Dumais S.T., and O'Brien G.W.: Using linear algebra for intelligent information retrieval. SIAM Review. 37(4) (1995) 573-595
5. Bingham E., Mannila H.: Random Projection in Dimensionality Reduction: Applications to Image and Text Data. Proc. SIGKDD (2001) 245-250
6. Hyvärinen A. Oja E.: Independent Component Analysis**:** Algorithms and Applications. Neural Networks. (4-5) (2000) 411-430. FastICA package: http://www.cis.hut.fi/~aapo/
7. Tang B. Luo X. Heywood M.I. Shepherd M.: A Comparative Study of Dimension Reduction Techniques for Document Clustering. TR # CS-2004-14, Faculty of Computer Science, Dalhousie University. (2004), http://www.cs.dal.ca/research/techreports/2004/CS-2004-14.shtml
8. Buckley C., Salton G., Allan J., Singhal A.: Automatic Query Expansion using SMART: TREC-3. Proc. TREC-3. (1995) 500-225.
9. Zhao Y. Karypis G.: Empirical and Theoretical Comparisons of Selected Criterion Functions for Document Clustering. Machine Learning.55 (3) (2004) 311 – 331.