Exploration of Dimensionality Reduction for Text Visualization

Shiping Huang, Matthew O. Ward and Elke A. Rundensteiner
Computer Science Department
Worcester Polytechnic Institute
Worcester, MA 01609
{shiping,matt,rundenst}@cs.wpi.edu *

ABSTRACT

In the text document visualization community, statistical analysis tools (e.g., principal component analysis and multidimensional scaling) and neurocomputation models (e.g., self-organizing feature maps) have been widely used for dimensionality reduction. Often the resulting dimensionality is set to two, as this facilitates plotting the results. The validity and effectiveness of these approaches largely depend on the specific data sets used and semantics of the targeted applications. To date, there has been little evaluation to assess and compare dimensionality reduction methods and dimensionality reduction processes, either numerically or empirically. The focus of this paper is to propose a mechanism for comparing and evaluating the effectiveness of dimensionality reduction techniques in the visual exploration of text document archives. We use multivariate visualization techniques and interactive visual exploration to study three problems: (a) Which dimensionality reduction technique best preserves the interrelationships within a set of text documents; (b) What is the sensitivity of the results to the number of output dimensions; (c) Can we automatically remove redundant or unimportant words from the vector extracted from the documents while still preserving the majority of information, and thus make dimensionality reduction more efficient. To study each problem, we generate supplemental dimensions based on several dimensionality reduction algorithms and parameters controlling these algorithms. We then visually analyze and explore the characteristics of the reduced dimensional spaces as implemented within a linked, multi-view multi-dimensional visual exploration tool, XmdvTool. We compare the derived dimensions to features known to be present in the original data. Quantitative measures are also used in identifying the quality of results using different numbers of output dimensions.

Keywords: Dimension reduction, multidimensional scaling (MDS), self-organizing maps (SOM), text visualization.

1 INTRODUCTION

With the rapid growth of the Internet, wireless communication, multimedia home and office servers, virtually everyone is faced with huge amount of information coming from digital libraries, web sites and other sources [14, 27]. Much of this information comes in the form of unstructured text. We simply cannot read or skim this information in a traditional way. To an ever increasing extent we depend on analysis and visualization tools to get insight into those documents.

The curse of dimensionality and the empty space phenomenon are unavoidable challenges in the text visualization and information retrieval communities. Text documents are often represented by a vector of word counts in a vector-space model of documents, where the dimensionality could be over 10,000. On the one hand, the sample size needed to estimate a function of several variables to a given degree of accuracy (i.e., to get a reasonably low-variance estimate) grows exponentially with the number of variables. On the other hand, the high-dimensional spaces are inherently sparse. For example, a word that appears in one document over 100 times may not appear in any of the other documents. An example is Figure 1. Here only the top 228 words are used to visualize a document collection with 98 records, although over 10,000 unique words are very common for even small document collections.

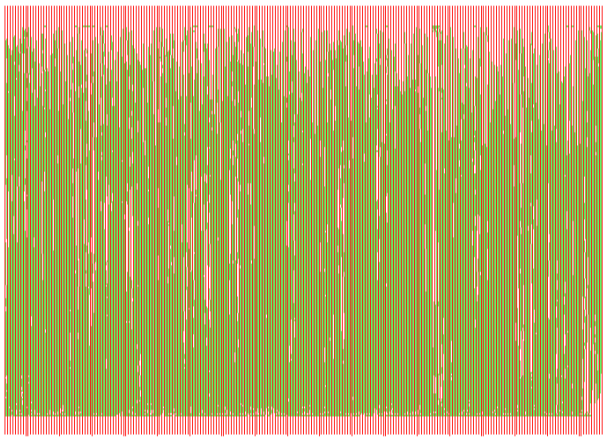

Figure 1: Parallel coordinates display showing counts for the top 228 words in a collection of documents. Clearly little structure can be seen.

To overcome these problems intrinsic in text visualization and classification, a widely used method is dimension reduction. The main idea behind these techniques is to map each text document into a lower dimensional space that explicitly takes the dependencies between the terms into account. The associations present in the lower dimensional representation can then be used to perform visualization, classification and categorization more efficiently.

While the reasons for performing dimension reduction are clear, it is not without problems. Open issues include [5, 6]:

- Unknown intrinsic dimensionality. We have no effective way to find the minimum number of dimensions sufficient to represent the data.
- Non-linear relationships among data. Underlying relationships among the variables may be very complicated.
- Unknown relevance of information. The case where dimension reduction is performed without losing information is ideal. Very often however dimension reduction will not be possible without a certain amount of loss.

Due to the complex nature of the dimension reduction process, there is no single method to deal with all situations. Thus, a large number of dimension reduction approaches have been developed

*This work is supported under NSF grant IIS-0119276.

and tested in different application domains and research communities. These dimension reduction techniques can be classified into three categories. One refers to the set of techniques that take advantage of class-membership information while computing the lower dimensional space. Examples of such techniques include a variety of feature selection schemes that reduce the dimensionality by selecting a subset of the original features [3], and techniques that derive new features by clustering the terms [1, 34, 33]. These dimension reduction techniques aim to minimize the information loss compared to the original data or to maintain the inter-record distances found in the data set. The second class of dimension reduction techniques are computational algorithms based on statistical analysis. principal component analysis (PCA), MDS and latent semantic indexing (LSI) belong to this category of dimension reduction techniques. They are appropriate to use in situations when the relationships among the dimensions are linear [9, 7, 19]. The third type of dimension reduction technique is self-organizing maps (SOMs) that use a neurocomputational approach.

It is widely accepted that there is no precise evaluation method for dimension reduction techniques even though a large number of algorithms have been developed. This paper attempts to address this problem. We try to evaluate several dimension reduction techniques both visually and statistically when applied to unstructured text documents. In addition, we explore the effectiveness and computational load of these dimension reduction techniques in terms of the number of distinct input dimensions used for the dimensionality reduction.

The remainder of this paper is organized as follows: Section 2 presents dimension reduction techniques for text visualization and information retrieval. Section 3 describes how we use XmdvTool [28, 29] to visually explore the effectiveness of some of these dimension reduction methods when applied to unstructured text documents. These dimension reduction techniques are also evaluated in terms of class clustering and statistical analysis. Section 4 describes related work on dimension reduction in different areas. Section 5 summarizes our work and presents possible future research.

2 Description of existing dimension reduction methods

We define dimension reduction as any operation that maps high dimensional data into a lower dimensional space, while attempting to preserve characteristics and relationships in the raw data. We now review the dimension reduction techniques analyzed in this paper.

2.1 Principal component analysis

Principal component analysis is a widely used technique for dimension reduction [9, 19, 7, 18]. Given an n x m document-term matrix (the number of documents and terms are n and m respectively), PCA uses the k-leading eigenvectors of the n x n covariance matrix as the axes of the lower k-dimensional space. These leading eigenvectors correspond to linear combinations of the original variables that account for the largest amount of term variability. One disadvantage of PCA is that it has high memory and computational requirements. It requires $O(n^2)$ memory for the dense covariance matrix, and $O(kn^2)$ for finding the k leading eigenvectors. These requirements could be unacceptably high when the number of documents (n) is very large, for example, tens of thousands.

The effectiveness of PCA in empirical studies is often attributed to reduction of noise, redundancy, and ambiguity [10]. The terms of a text document are typically not independent. The noise and redundancy could show in the term-matrix text data. This could lead to the conclusion that PCA is suitable for text document data, but the resulting dimensions lack semantic meaning.

2.2 Multidimensional scaling

Multidimensional Scaling (MDS) is a set of mathematical techniques that enable a researcher to uncover hidden structure in data. Common applications include fields such as psychology, sociology, economics, educational research and document visualization [7, 5, 2].

Suppose we have a set of objects (e.g., a number of text documents) and that a measure of the similarity between objects is known. This measure, called proximity, indicates how similar or how dissimilar two objects are or are perceived to be. It can be obtained in different ways, e.g., by computing the correlation coefficient or Euclidean distance from the vector representation of the text documents. What MDS does is to map to a lower dimensional space in which each object is represented by a point and the distances between points resemble the original similarity information; i.e., the larger the dissimilarity between two objects, the farther apart they should be in the lower dimensional space. This geometrical configuration of points reflects the hidden structure of the data and may help to make it easier to understand.

2.3 Self-organizing maps

Self-organizing maps (SOM) is a neurocomputational algorithm to map high-dimensional data to a lower (typical two) dimensional space through a competitive and unsupervised learning process [21, 22]. This algorithm is frequently used to visualize and interpret large high-dimensional data sets. It has also been employed to visualize very large unstructured text document archives [24, 16].

Self-organizing maps take a set of objects (e.g., text documents), each object represented by a vector of terms (keywords from the original text), and then maps them onto the nodes of a two-dimensional grid. The map is represented initially by a matrix of nodes, where each node is represented by a codebook vector with the same length as the input vectors. Fitting of the model vectors is usually carried out by a sequence of best match and neighborhood modification processes. For a specific input vector, a distance measure is used to find the best match codebook that has the closest distance to the input vector. Then the neighborhood codebooks are modified based on the input vector and neighborhood function. These processes are iterated over the available input vectors.

2.4 Similarity-based dimension clustering

Another approach to dealing with high-dimensional data is to group the dimensions based on a similarity measure. In XmdvTool [34, 33] an agglomerative clustering algorithm is used to create a dimension hierarchy. Given the hierarchy a radial space-filling (RSF) technique called InterRing [34, 33] is then used to provide interactive operations such as dimension hierarchy navigation and modification.

Figure 2 shows the clustering of the top seventy one words extracted from a document collection. The center ring corresponds to the cluster containing all the terms, and each successive ring is broken into branches of the cluster hierarchy. Individual terms are found on the outer (terminal) nodes. After calculating the correlation coefficients among the word vectors (in a term-document matrix, each document was represented by a vector of word counts that indicate the number of occurrences in the corresponding document), we see that the clustering algorithm groups ”study” and ”problems”, ”discussed” and ”methods”, ”paper” and ”presented”, and so on together. InterRing provides flexibility and a rich assortment of user interactions so that the user can gain more understanding about the dimension reduction process and use her domain knowledge to reorganize the clusters if desired. Dimensionality reduction is achieved by either selecting clusters of similar dimensions or a subset of representative dimensions for visual analysis.

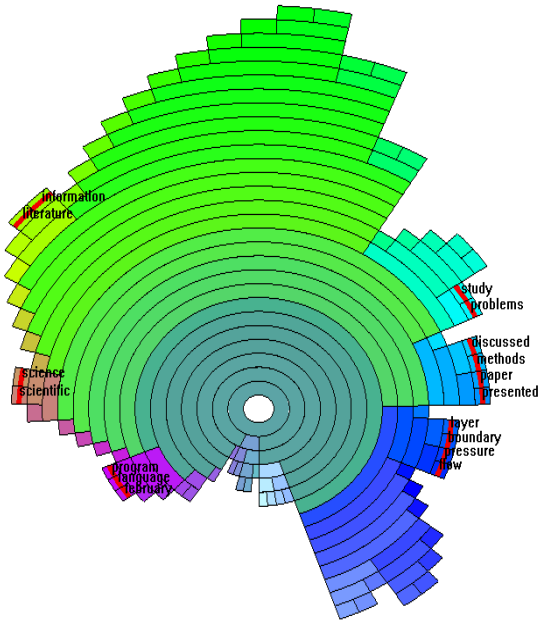

Figure 2: An agglomerative clustering of the top 71 words in a set of documents, displayed with InterRing. Labeled nodes convey the quality of the clustering.

3 VISUAL EXPLORATION

In this section, we describe a process by which the effectiveness of the dimension reduction techniques including MDS, SOM and agglomerative clustering can be visually evaluated. Then the effect of using different number of input dimensions are assessed to ascertain if reducing the input vector size can yield comparable results.

First, the goal was to visually assess the quality of proposed word and document clusters by exploring their multi-dimensional nature and examining the capabilities of a dimension reduction technique to construct the necessary decision boundaries that separate the groups in the text data. Direct interactions among MDS, SOM and agglomerative clustering help to enable this task. The interesting regions in one display can be highlighted and the corresponding data items in the derived dimension space can be examined. Alternatively, samples or regions in the derived dimension space that are suspected of being problematic or exhibit clustering can be selected and the data samples giving rise to them can be explored via the other dimension spaces. This is an example of the use of an established exploratory technique called linked brushing; what is new here is that visualization of raw data points is linked with the display of output from different dimensionality reduction techniques.

The second goal of this paper was to test the effects of using using different numbers of input terms for these dimension reduction techniques. All these dimensionality reduction techniques are computationally intensive and are sensitive to the number of original dimensions. The final goal was to explore the sensitivity of the number of output dimensions used in MDS. We feel that many researchers in document visualization routinely default to using two output dimensions, and perhaps do not realize the amount of information loss this can generate. We hope to encourage more use of higher order MDS output to generate better cluster boundaries.

The test data on which we ran experiments are from standard test document collections in the information retrieval community [30]: CRAN (1398 document abstracts on Aeronautics from Cranfield Institute of Technology), CACM (3204 abstracts of articles in the Communications of ACM), MED (1033 abstracts from the National Library of Medicine), TIME (546 documents), LISA (6004 text collections), and CISI (1460 abstracts from the Institute of Scientific Information). Each of these text collections is broken into a number of separate files with about one hundred abstracts or text collections for each file.

An available public domain tool, Rainbow [25], was employed for text extraction. The text was tokenized using common tokenization options: the words from the SMART stop-list (524 common words) [4], such as ”the” and ”of”, are neglected before tokenization; the Porter stemming algorithm [12] was applied for all words before they are counted. After tokenization, a document-term matrix was acquired and processed by the dimension reduction algorithms mentioned above for analysis, visualization, and comparative study.

3.1 Effectiveness study of MDS, SOM and InterRing

The variant of MDS we employed was the Shepard-Kruskal algorithm [7]. We used the principal components as the initial configuration. An optimization process was carried out until the stress difference between two iterations was less than 0.001. We also computed SOMs consisting of 10 x 10 codebook vectors. Figure 3 presents examples from the medical abstracts archive based on selecting records with high level of occurrences for particular terms (’coronary’ for the first and ’tumor’ for the second). It is clear in both cases that MDS resulted in better clustering than SOM. On the other hand, in Figure 4 we see in a similar search (based on the term ’myocardial’) the clustering in SOM space seems to be better than in MDS space. Finally, Figure 5 shows that a small, dense region in MDS space can map to several nodes in SOM space, while a set of outliers in MDS space may map to a single node in SOM space.

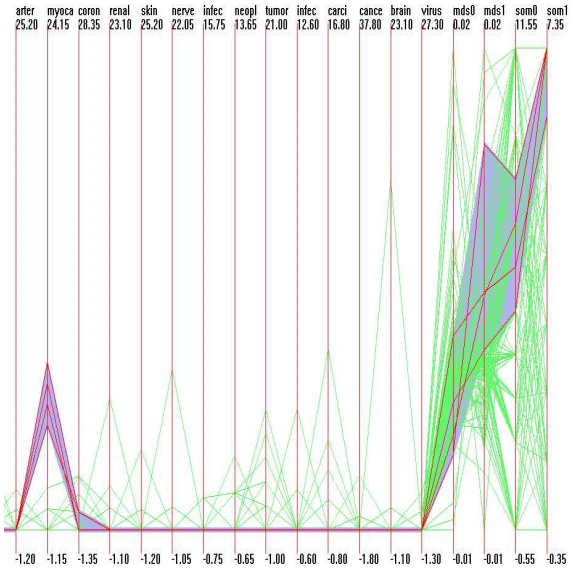

Figure 4: Comparing MDS and SOM: clustering data with high numbers of occurrences of the term ’myocardial’. The derived dimensions (mds0, mds1, som0, som1) are the last four dimensions. For this example we see tighter clustering in SOM space.

To facilitate studying sub-clustering activities in documents from different sources (based on the assumption that good dimensionality reduction would enable users to differentiate clusters), we assigned a numerical label arbitrarily for each document so that the differences between the documents from the same source are small while

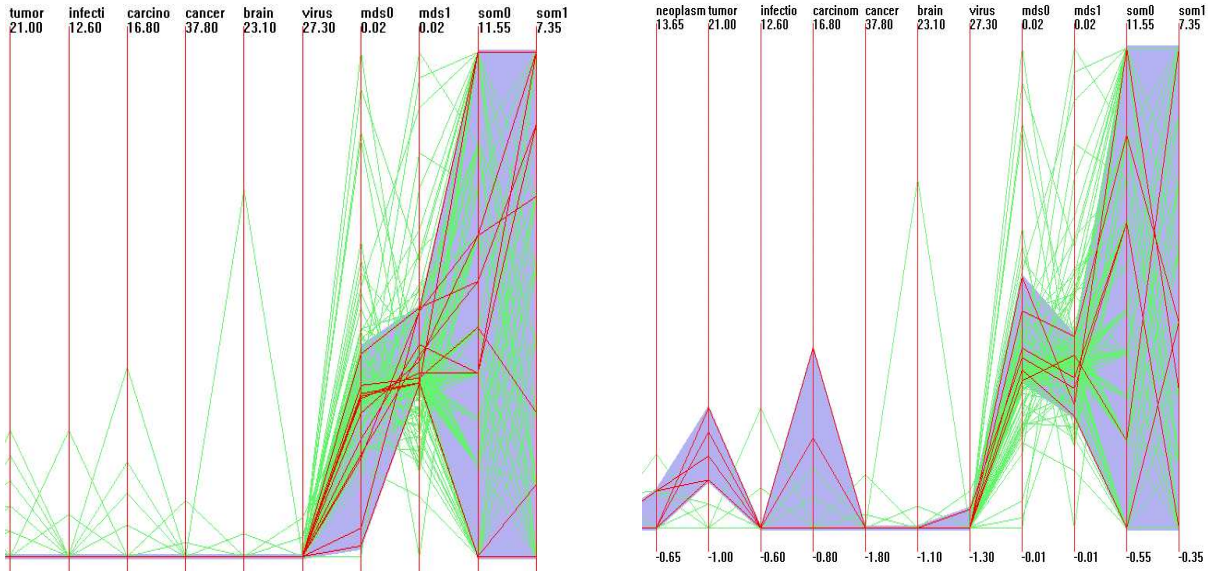

Figure 3: Comparing MDS and SOM: clustering data with high numbers of occurrences of the terms 'coronary' (first plot) and 'tumor' (second plot). The derived dimensions (mds0, mds1, som0, som1) are the last four dimensions. Selected points are highlighted in red, and their envelope (the hyperbox containing all selected points) is shown in grey. For these examples we see tighter clustering in MDS space.

those between documents from different sources are large. This label corresponds to the first dimension in some of our test data sets.

Upon investigation we can find differences in the cluster results in MDS and SOM spaces. Figure 6 shows mutual clustering activities among word clusters (nodes), single words and derived dimensions of MDS and SOM. The discretization and the rigidity of MDS's output space are clearly visible if one compares them with output maps given by SOM's output space. In addition, the clustering appears better in word cluster (nodes) space than single word space.

Most text document visualization systems only use the first two principal components as the lower dimensional space. It is good for visualization implementation because a terrain surface is a convenient metaphor to convey information hidden in the text collection. However, the loss of information is often significant. Figure 7 plots the logarithm of the principal values for the data set mentioned above. All but one principal value are positive, albeit the first 5 principal values are much bigger than the others.

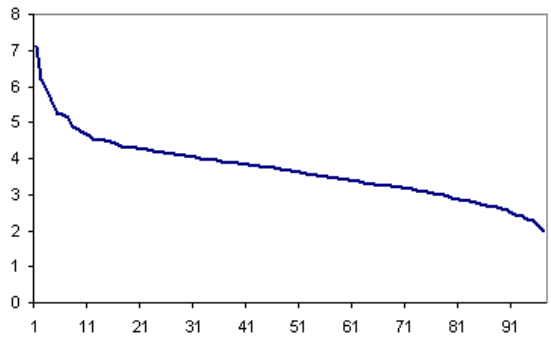

Figure 7: Principal values. The first 98 eigenvalues are positive, though only the first 5 are much bigger than the rest.

On the other hand, SOM can also lead to information loss. Figure 8 shows the hierarchical document clustering in the derived dimension space, which includes word clusters (nodes), single words, MDS and SOM derived dimensions. It is not difficult to find that only two document clusters exist in SOM space but six distinct clusters appear in other dimension spaces (which is correct, as for this case, we used samples from six sources). This could happen when there is only a limited number of nodes in the SOM algorithm. A higher number of nodes might help against the negative effect of the discretization of SOM's output space.

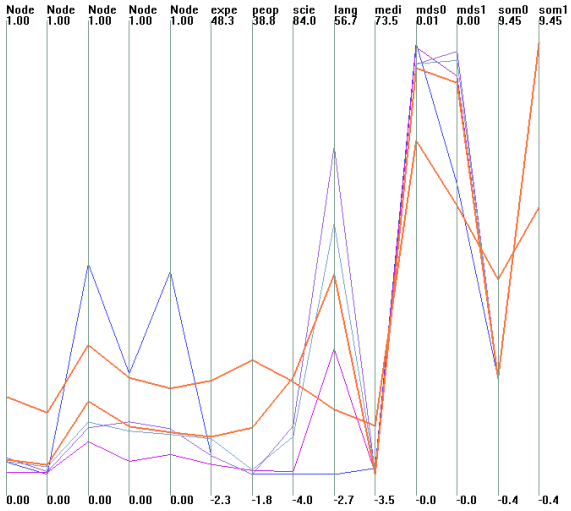

Figure 8: Degenerate problems of SOM. Only two clusters exist in SOM derived space while roughly six clusters are discernible in the other dimension spaces.

This inspired us to investigate these dimension reduction algorithms in more depth. For MDS, rather than reduce to two dimensions, we reduced to 3, 4, 5 and more dimensions. Second, we combined these MDS dimensions with derived dimensions from SOM. We expected to detect more details that exist in the space defined by the original document collections. Figure 9 contains two views of the derived dimensions for the data set generated by merging docu-

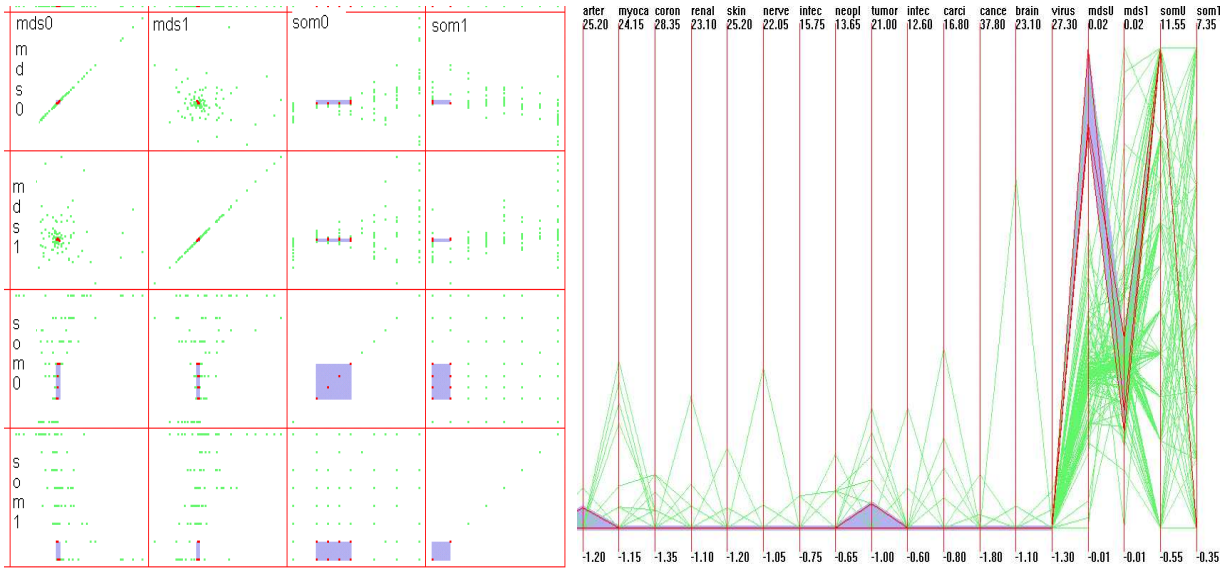

Figure 5: Comparing MDS and SOM: a) A scatterplot matrix of the derived dimensions. Several adjacent nodes in SOM space map to the center of the main cluster in MDS space. b) Outliers in MDS space map to a single node in SOM space.

ments from six sources; the first figure uses two output dimensions for MDS and the second uses four. A tight cluster in the 2D version maps to several adjacent nodes in the SOM space, while using the 4D version we can isolate the cluster at a single node in SOM space. In Figure 10 we see for a section of the keywords the original envelope (in grey) and the data points selected by the 4D query,thus verifying that a more refined cluster has been isolated. On the other hand, Figure 11 shows that the cluster formed by selecting high occurrences of the word 'saigon' is clearly isolated both in the SOM space and MDS with two output dimensions. The third MDS output dimension resulted in no additional discernment ability.

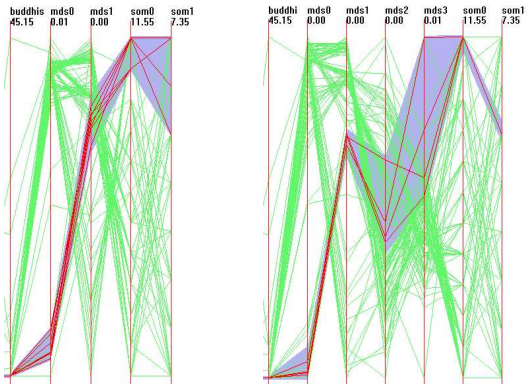

Figure 9: Derived dimensions (MDS and SOM) for the CACM data set, using two and four output dimensions for MDS, respectively. A distinct cluster in 2D MDS space is highlighted. A more focused selection in 4D isolates a tighter cluster, as seen in SOM space.

3.2 Computation exploration for MDS and SOM

Text mining, visualization and analysis are processes that often require a short response time. That means that when a user specifies the document collection needed for analysis, the system should be able to process, analyze and present a visual interpretation for the document collection in a short time span. The challenge is that all

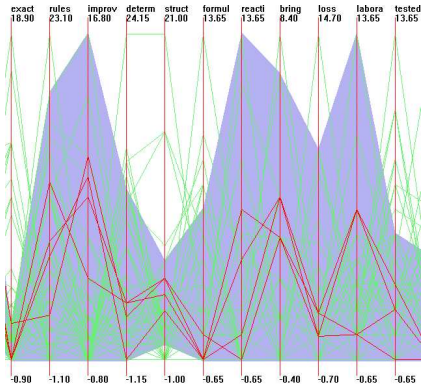

Figure 10: Section of raw data initially selected with two dimensions in MDS space and then refined using MDS dimension four. A much tighter group of similar documents has been isolated.

the dimension reduction techniques discussed in this paper are time consuming. The computational complexity depends on the number of data records and the number of dimensions that are used for dimension reduction. There is generally very little flexibility in terms of the number of data records (document collections) used for computation, although we might get reasonable approximations using sampling. The alternative option is to explore the computational complexity and effectiveness of these algorithms when using different numbers of input dimensions for dimension reduction.

We explored the computational resource requirements and effectiveness for MDS and SOM with different numbers of input dimensions. For simplicity only a subset of the top words (terms) was used as input in our experiments in a document vector space model. We generated results using 71/228/1634 words for the dimension reduction algorithms. The clustering activities are shown in Figure 12. Their computation time and stress from MDS are shown in Table 1. We found that the clustering activities of text documents were not significantly improved with an increased number of input dimensions. However, significant difference exists in terms of computational time when computing with different numbers of

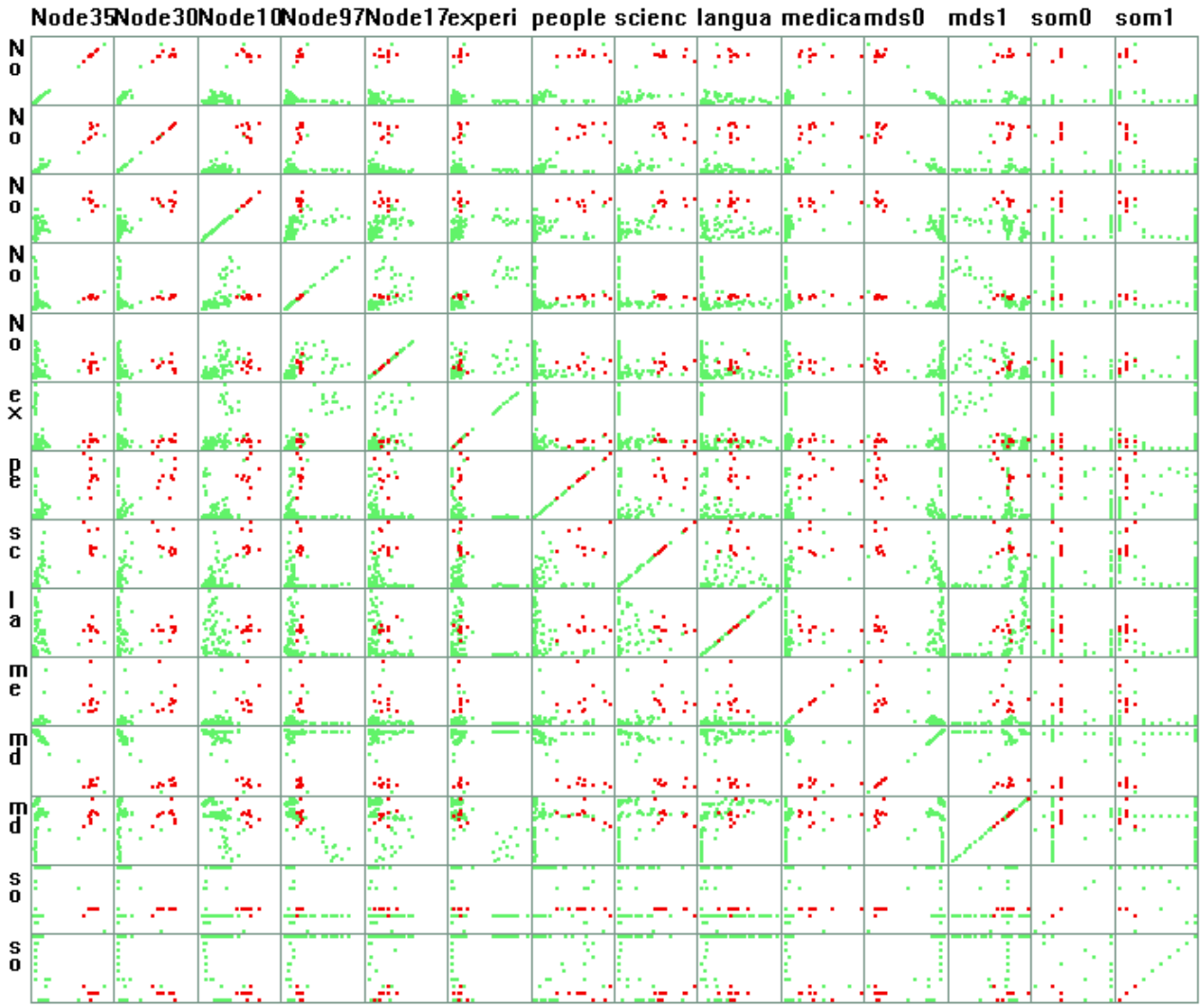

Figure 6: Scatterplot matrix: Several word clusters and individual words, followed by four derived dimensions (2D MDS and 2D SOM). A cluster isolated (circled) in the plot of the two MDS dimensions is highlighted (red) and clearly identifiable in many of the other views.

Extracted document Data	Time (sec)	Stress (MDS)
Top 71 keywords	300	0.281
Top 228 keywords	5000	0.247
Top 1634 keywords	60000	0.217

Table 1: Computation time, stress from MDS for document data.

input dimensions.

4 Related work

The major approaches for dimensionality reduction in the text visualization community belong to topology preserving algorithms, which include PCA, MDS and SOM. Topology preserving algorithms aim to represent high dimensional data spaces in a low dimensional space while preserving as much as possible the structure of the data in the high dimensional data space. This is achieved by mapping "points in one space to points in another space such that nearby points map to nearby points (and sometimes in addition far-away points map to far-away points)" [11].

Galaxies [32, 31] visualization displays clusters and document interrelatedness by reducing a high dimensional representation of documents to a two dimensional scatterplot. The documents are clustered in the high dimensional space through a metric of similarity such as Euclidean distance or cosine measures. Then the documents are projected to a 2D space that reflects document clusters with cluster centroids. In ThemeScape [31] two different dimension reduction techniques were applied. For small document sets (up to 1.5k), the Shepard multidimensional scaling algorithm was used, while for large document sets, an Anchored Least Stress algorithm was developed. The ground plane was employed to project the document sets, where the peaks represent the large number of document clusters and the valleys represent the distances between these document clusters as found in the raw document sets.

A number of papers have been published on the utilization of self-organizing maps for interactive exploration of document collections [23, 24, 26], i.e., the WEBSOM project [16, 15, 17]. Self-organizing maps are used to represent documents on a map that provides an insightful view of the document collections. This view visualizes similarity relations between the documents. The com-

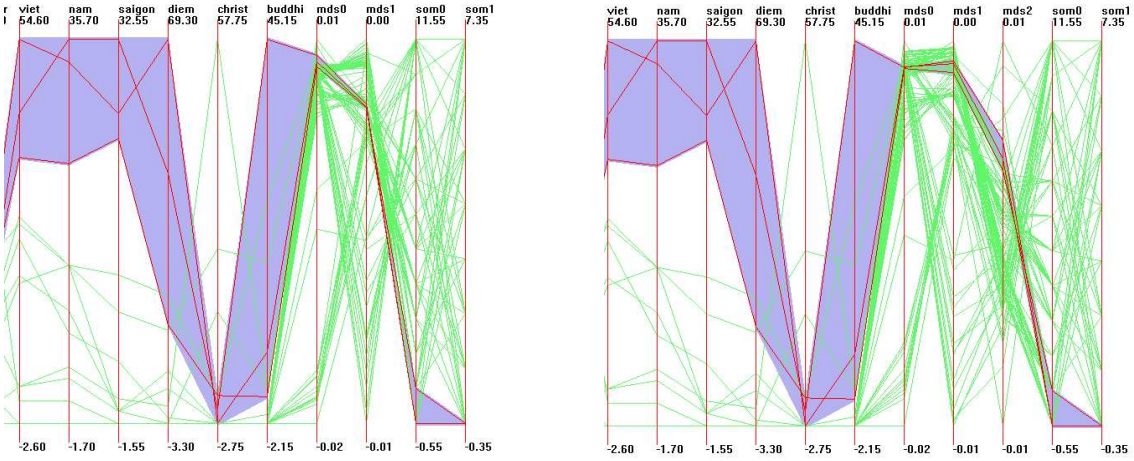

Figure 11: A section of data dimensions followed by the derived dimensions (MDS and SOM) for the CACM data set, using two and three output dimensions for MDS, respectively. Clustering based on high numbers of occurrence of the term 'saigon' is easily distinguished in both views.

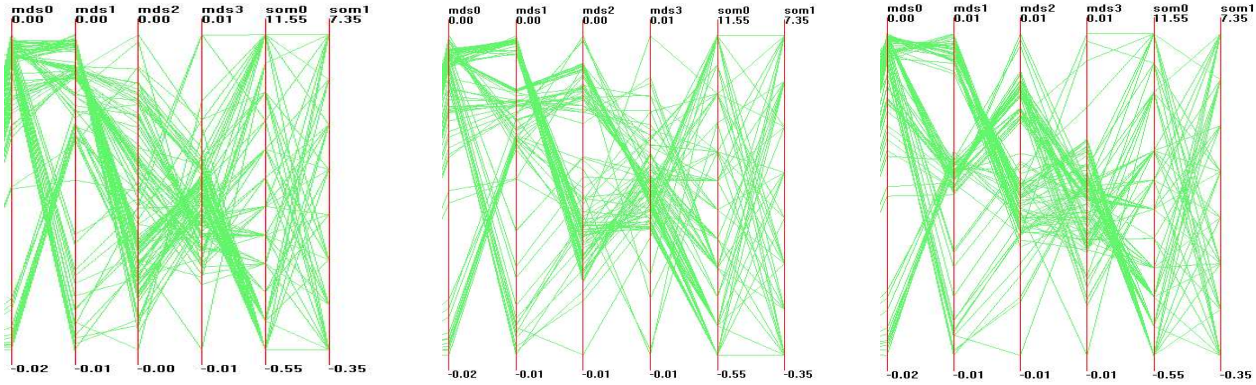

Figure 12: Document clusters in MDS and SOM spaces with different numbers of input dimensions. (a), (b) and (c) represent the derived dimensions computed using the top 1638/232/75 words. In MDS derived space, clusters in (a) and (b) are more With the increased number of input dimensions, the clustering activities were not improved significantly. discernible than in (c), while in SOM derived space, there are three clusters in (a) and (c) but only one cluster in (b).

plete WEBSOM method involves a two-level SOM architecture comprised of a word category map and a document map. SOMs were used to construct a word category map. Usually interrelated words that have similar context appear close to each other on the map. Then the documents are encoded by mapping their text onto the word category map. The document map is then formed with a SOM algorithm using the document vectors in word category map space.

In [11], the use of self-organizing maps for clustering and visualization was discussed in depth. A comparative study on the quality and effectiveness of SOMs and Sammon's mapping when applying to classification and visualization was reported.

A number of other dimension reduction algorithms have been reported in other communities. The computation of dimensions using principal component analysis through singular value decomposition (SVD) is a popular approach for numerical attributes. In information retrieval, latent semantic indexing uses SVD to project textual documents represented as document vectors. SVD is shown to be the optimal solution for a probabilistic model for document/word occurrence [10]. An adaptive dimension reduction algorithm that attempts to avoid local minima was used for clustering high dimensional data in [9]. They claimed that if the data distribution is far from Gaussian, the dimensions selected using PCA will deviate substantially from the optimal. A nonlinear dimension reduction method with minimal loss of (mutual) information contained in the original data was proposed for text classification [13]. In addition, dimension reduction by random mapping was also reported [26, 20].

5 Conclusion and Future Work

In this paper, several existing dimension reduction techniques were explored and evaluated for text document visualization. The effectiveness and computational complexity of these techniques were also compared. We conclude:

- Visualization can be useful for comparing and evaluating different dimensionality reduction methods by linked brushing between the derived dimensions and the original data.
- The first two principal components that are commonly used for text document visualization in many systems often lead to significant information loss. The 3rd and 4th and sometimes even 5th or more components could contribute to the accurate classification and visualization of the text documents.

- The discretization problem of SOM is not avoidable. Increasing the grid number may improve this problem to some degree, however the computation load can become unacceptable.

- In assessing the tradeoff between computational load and precision for MDS, we found that for many data sets a significant number of input dimensions could be eliminated without seriously degrading the quality of the results of dimensionality reduction.

Future research work could include:

- In addition to derived dimensions from dimension reduction techniques such as MDS and SOMs, the metrics that were used to evaluate the quality of dimension reduction algorithms, such as stress from MDS, could be used as derived dimensions. This makes it possible to evaluate how much individual documents contribute in terms of the total stress.

- To overcome the discretization of SOMs, a relatively new algorithm for performing topology preserving non-linear dimension reduction, Curvilinear Components Analysis (CCA) [8, 6], could be explored in such situations where there are large number of text documents.

- Additional dimensionality reduction techniques found in information retrieval and text classification, such as LSI [10, 9], could be incorporated into future studies.

6 ACKNOWLEDGMENTS

We gratefully acknowledge our colleagues in the XmdvTool group at WPI for their contributions to this research. The research funds from NSF and NSA are graciously appreciated.

REFERENCES

[1] M. Ankerst, S. Berchtold, and D. A. Keim. Similarity clustering of dimensions for an enhanced visualization of multidimensional data. *Proc. of IEEE Symposium on Information Visualization, InfoVis'98, p. 52-60*, 1998.

[2] W. Basalaj. Proximity visualisation of abstract data. Technical Report UCAM-CL-TR-509, University of Cambridge, Computer Laboratory, 15 JJ Thomson Avenue, Cambridge CB3 0FD, United Kingdom, phone +44 1223 763500, January 2001.

[3] T. Bayer, H. Mogg-Schneider, I. Renz, and H. Schafer. Daimler benz research: System and experiments routing and filtering. In *Text REtrieval Conference*, pages 329–346, 1997.

[4] C. Buckley. Implementation of the smart information retrieval system. Technical Report TR85-686, Cornell University, Computer Science Department, may 1985.

[5] M. A. Carreira-Perpinan. A review of dimension reduction techniques. Technical Report CS–96–09, Dept. of Computer Science, University of Sheffield, January 1997.

[6] A. Choppin. Unsupervised classification of high dimensional data by means of self-organizing neural networks. Master's thesis, Universit catholique de Louvain (Belgium), June 1998.

[7] M. Davison. *Mutidimensional scaling*. John Wiley & Sons, 1983.

[8] P. Demartines and J. Hérault. Curvilinear component analysis: A self-organizing neural network for nonlinear mapping of data sets. *IEEE Transactions on Neural Networks*, 8(1):148–154, January 1997.

[9] C. Ding, X. He, H. Zha, and H. Simon. Adaptive dimension reduction for clustering high dimensional data. In *Proc. 2nd IEEE Int'l Conf. Data Mining*, pages 147–154, December 2002.

[10] C. H. Ding. A probabilistic model for dimensionality reduction in information retrieval and filtering. In *Proc. of 1st SIAM Computational Information Retrieval Workshop*, October 2000.

[11] A. Flexer. On the use of self-organizing maps for clustering and visualization. *Intelligent-Data-Analysis*, 5:373–84, 2001.

[12] W. B. Frakes. Stemming algorithms. In W. B. Frakes and R. Baeza-Yates, editors, *Information Retrieval: data structures and algorithms*, pages 131–160. Prentice Hall, Englewood Cliffs, US, 1992.

[13] A. Globerson and N. Tishby. Sufficient dimensionality reduction - a novel analysis method. In *ICML*, pages 203–210. IEEE Service Center, July 2002.

[14] S. Havre, E. Hetzler, P. Whitney, and L. Nowell. ThemeRiver: Visualizing thematic changes in large document collections. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):9–20, January 2002.

[15] T. Honkela. Comparisons of self-organized word category maps. In *Proceedings of WSOM'97, Workshop on Self-Organizing Maps, Espoo, Finland, June 4–6*, pages 298–303. Helsinki University of Technology, Neural Networks Research Centre, Espoo, Finland, 1997.

[16] T. Honkela, S. Kaski, K. Lagus, and T. Kohonen. Exploration of full-text databases with self-organizing maps. In *Proceedings of the ICNN96, International Conference on Neural Networks*, volume I, pages 56–61. IEEE Service Center, Piscataway, NJ, 1996.

[17] T. Honkela, S. Kaski, K. Lagus, and T. Kohonen. WEBSOM–self-organizing maps of document collections. In *Proceedings of WSOM'97, Workshop on Self-Organizing Maps, Espoo, Finland, June 4–6*, pages 310–315. Helsinki University of Technology, Neural Networks Research Centre, Espoo, Finland, 1997.

[18] J. Jolliffe. *Principal Component Analysis*. Springer Verlag, 1986.

[19] G. Karypis and E.-H. Han. Concept indexing: A fast dimensionality reduction algorithm with applications to document retrieval and categorization. Technical report tr-00-0016, University of Minnesota, 2000.

[20] S. Kaski. Dimensionality reduction by random mapping: Fast similarity computation for clustering. In *Proceedings of IJCNN'98, International Joint Conference on Neural Networks*, volume 1, pages 413–418. IEEE Service Center, Piscataway, NJ, 1998.

[21] T. Kohonen. The self-organizing map. *Proc. of IEEE, p. 1464-80*, 1978.

[22] T. Kohonen. *Self Organizing Maps*. Springer Verlag, 1995.

[23] T. Kohonen, S. Kaski, K. Lagus, and T. Honkela. Very large two-level SOM for the browsing of newsgroups. In C. von der Malsburg, W. von Seelen, J. C. Vorbrüggen, and B. Sendhoff, editors, *Proceedings of ICANN96, International Conference on Artificial Neural Networks, Bochum, Germany, July 16–19, 1996*, Lecture Notes in Computer Science, vol. 1112, pages 269–274. Springer, Berlin, 1996.

[24] X. Lin, D. Soergel, and G. Marchionini. A self-organizing semantic map for information retrieval. In *Proceedings of the Fourteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Semantic Models, pages 262–269, 1991.

[25] A. K. McCallum. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. http://www.cs.cmu.edu/ mccallum/bow, 1996.

[26] H. Ritter and T. Kohonen. Self-organizing semantic maps. *Biological Cybernetics*, 61(4):241–54, 1989.

[27] J. Thomas, K. Cook, V. Crow, B. Hetzler, R. May, D. McQuerry, R. McVeety, N. Miller, G. Nakamura, L. Nowell, P. Whitney, and P. Wong. Human computer interaction with global information spaces - beyond data mining, 1999.

[28] M. Ward. Xmdvtool: Integrating multiple methods for visualizing multivariate data. *Proc. of Visualization '94, p. 326-33*, 1994.

[29] M. Ward, J. LeBlanc, and R. Tipnis. N-land: a graphical tool for exploring n-dimensional data. *CG194 Proc: Insight Through Computer Graphics, p. 130-41*, 1996.

[30] R. W. White, I. Ruthven, and J. M. Jose. Finding relevant documents using top ranking sentences: an evaluation of two alternative schemes. In M. Beaulieu, R. Baeza-Yates, S. H. Myaeng, and K. Järvelin, editors, *Proceedings of the 25th annual international ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 57–64, New York, Aug. 11–15 2002. ACM Press.

[31] J. A. Wise. The ecological approach to text visualization. *JASIS, Vol. 50, No. 13, p. 1224-1233*, 1999.

[32] J. A. Wise, J. J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, and V. Crow. Visualizing the non-visual: Spatial analysis and interaction with information from text documents. *Proc. of Information Visualization '1995, p. 51-58*, 1995.

[33] J. Yang, M. O. Ward, and E. A. Rundensteiner. Interring: An interactive tool for visually navigating and manipulating hierarchical structures. *IEEE Symposium on Information Visualization (InfoVis'02), p. 77-84*, 2002.

[34] J. Yang, M. O. Ward, E. A. Rundensteiner, and S. Huang. Visual hierarchical dimension reduction for exploration of high dimensional datasets. *VisSym 2003, accepted*, 2003.