To the Graduate Council:

I am submitting herewith a thesis written by Farial Shahnaz entitled "A Clustering Method Based on Nonnegative Matrix Factorization for Text Mining." I have examined the final paper copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Computer Science.

Michael W. Berry

Michael W. Berry, Major Professor

We have read this thesis and
recommend its acceptance:

R C Ward

George S. Jordan

Accepted for the Council:

Vice Chancellor and
Dean of Graduate Studies

A Clustering Method Based on Nonnegative Matrix Factorization for Text Mining

A Thesis

Presented for the

Master of Science Degree

The University of Tennessee, Knoxville

Farial Shahnaz

August 2004

Thesis
2004
.S48

Dedication

To my family. Of course.

Acknowledgement

First and foremost, I would like to express my deep gratitude to Dr. Michael Berry: my thesis would have remained a distant possibility without his constant support and guidance. I would like to thank Drs. Paul Pauca and Robert Plemmons at Wake Forest University for their enormous contribution to this study, and my committee members, Drs. Samuel Jordan and Robert Ward, for their time and involvement. I would also like to extend my gratitude to Murray Browne for his invaluable help.

Last, but not least, I would like to thank the entire Computer Science Department for providing me with a home and making my journey through college a successful and enjoyable one.

Abstract

This study presents a methodology for automatically identifying and clustering semantic features or topics in a heterogeneous text collection. The methodology involves encoding the text data using a low rank nonnegative matrix factorization algorithm to retain natural data nonnegativity, thereby eliminating the need to use subtractive basis vector and encoding calculations present in other techniques such as principal component analysis for semantic feature abstraction. Existing techniques for nonnegative matrix factorization are reviewed and a new hybrid technique for nonnegative matrix factorization is proposed. Performance evaluations of the proposed method is conducted on a few benchmark text collections used in standard topic detection studies.

Contents

1 Introduction 1

2 Motivation 4

3 Algorithm 6

3.1 Multiplicative method 8

3.2 Sparse Encoding 9

3.3 A Hybrid Method 10

3.4 Algorithm for GD-CLS 11

4 Software Implementation 12

4.1 GTP 12

4.2 LAPACK 13

5 Performance 14

5.1 Reuters 14

5.2 TDT2 20

6 Observations **23**

7 Conclusion and Future Work **29**

Bibliography **31**

Appendices **36**

A GD-CLS algorithm in Matlab **37**

B Topic File for Reuters **38**

C Topic File for TDT2 **39**

D SGML Tags for Reuters Files **41**

E SGML Tags for TDT2 Files **42**

Vita **43**

List of Tables

5.1 Document-topic relationship in Example 5.1.1 16

5.2 Topic list produced for Example 5.1.1 16

5.3 The 2×5 H matrix for Example 5.1.2 19

5.4 The A matrix for Example 5.1.2 19

5.5 Comparison between original cluster labels and GD-CLS generated labels for Example 5.1.2 19

6.1 Results for Reuters 25

6.2 Results for TDT2 26

6.3 CPU time for $k = 15$ for different λ values 27

6.4 A comparison of results with different cluster sizes 28

Chapter 1

Introduction

Text mining refers to the detection of trends, patterns, or similarities in natural language text. Given a collection of text documents, often the need arises to classify the documents into groups or clusters based on similarity of content. For a relatively small collection, it may be possible to manually perform the partitioning of documents into specific categories. But to partition large volumes of text, the process would be extremely time consuming. Moreover, automation also greatly reduces the time needed to perform the classification.

When the categories or topics for classification are predefined, the process of classification is considered *supervised*; there are several methods in use that satisfactorily automate the task of supervised classification [7]. However, in absence of any information regarding the nature of the data, the problem of classification

becomes much more difficult. For *unsupervised* classification of text data, only one valid assumption can be made, which is that the text collection is completely unstructured. The task then becomes organizing the documents into a structure based solely on patterns learned from the collection itself. This structure can be *partitional* or *hierarchical* [7]. The hierarchical organization of documents has a tree-like structure with the entire collection situated at the root level. In subsequent levels of the tree, the collection is partitioned into smaller groups and eventually each document is represented as a separate group at the bottom level.

If the text collection is given a partitional structure, then the documents in the collection are flatly partitioned or clustered into groups that are non-overlapping. The proposed Nonnegative Matrix Factorization (NMF) method for text mining introduces a technique for partitional clustering that identifies semantic features in a document collection and groups the documents into clusters on the basis of shared semantic features. The algorithm used in the NMF method for this study was developed by Drs. Paul Pauca and Robert Plemmons at Wake Forest University. The factorization can be used to compute a low rank approximation of a large sparse matrix along with preservation of natural data nonnegativity.

In the *vector space model* of text data, documents are encoded as n-dimensional vectors where n is the number of terms in the dictionary, and each vector com-

ponent reflects the importance of the corresponding term with respect to the semantics of a document [3]. A collection of documents can, thus, be represented as a term-by-document matrix. Since each vector component is given a positive value (or weight) if the corresponding term is present in the document and a null or zero value otherwise, the resulting term-by-document matrix is always nonnegative. This inherent data nonnegativity is preserved by the NMF method as a result of constraints (placed on the factorization) that produce nonnegative lower rank factors that can be interpreted as semantic features or patterns in the text collection. The vectors or documents in the original matrix can be reconstructed by combining these semantic features, and documents that have common features can be viewed as a cluster. As shown by Xu et al. [22], NMF outperforms traditional vector space approaches to information retrieval (such as latent semantic indexing) for document clustering on a few topic detection benchmark collections.

Chapter 2

Motivation

Nonnegative matrix factorization differs from other rank reduction methods for vector space models in text mining, e.g., principal component analysis (PCA) or vector quantization (VQ), due to use of constraints that produce nonnegative basis vectors, which make possible the concept of a *parts-based representation* [14]. Lee and Seung first introduced the notion of parts-based representations for problems in image analysis or text mining that occupy nonnegative subspaces in a vector-space model. Techniques like PCA and VQ also generate basis vectors – various additive and subtractive combinations of which can be used to reconstruct the original space. But the basis vectors for PCA and VQ contain negative entries and cannot be directly related to the the original vector space to derive meaningful interpretations. In the case of NMF, the basis vectors contain no negative entries – this allows only additive combinations of the vectors to reproduce the original.

So the perception of the whole, be it an image or a document in a collection, becomes a combination of its parts represented by these basis vectors. In text mining, the vectors represent or identify semantic features, i.e., a set of words denoting a particular concept or topic. If a document is viewed as a combination of basis vectors, then it can be categorized as belonging to the topic represented by its principal vector. Thus, NMF can be used to organize text collections into partitional structures or clusters directly derived from the nonnegative factors.

Recently Xu et al. [22] have demonstrated that NMF outperforms methods such as singular value decomposition and is comparable to graph partitioning methods that are widely used in clustering text documents. The tests were conducted on two different datasets: the Reuters data corpus[1] and TDT2 corpus[2], both considered benchmark collections for topic detection. These two data corpora are also used in this study to observe the results of using nonnegative factorization for text mining or document clustering. The algorithm used to derive the factorization introduces a new parameter to control the number of basis vectors used to reconstruct the document vectors, thereby providing a mechanism to balance the tradeoff between accuracy and computational cost (including storage).

[1] Reuters-21578 at http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html.
[2] http://www.lcd.upenn.edu.

Chapter 3

Algorithm

Given a set of documents S, in the standard vector space model S can be expressed as a $m \times n$ matrix V, where m is the number of terms in the dictionary and n is the number of documents in S. Each column V_j of V is an encoding of a document in S and each entry v_{ij} of vector V_j is the significance of term i with respect to the semantics of V_j, where i ranges across the terms in the dictionary. The NMF problem is defined as finding a low rank approximation of V in terms of some metric (e.g., the norm) by factoring V into the product (WH) of two reduced-dimensional matrices W and H. Each column of W is a basis vector, i.e., it contains an encoding of a semantic space or concept from V and each column of H contains an encoding of the linear combination of the basis vectors that approximates the corresponding column of V. Dimensions of W and H are $m \times k$ and $k \times n$ respectively, where k is the reduced rank or selected number

of topics. Usually k is chosen to be much smaller than n, but more accurately, $k << min(m, n)$. Finding the appropriate value of k depends on the application and is also influenced by the nature of the collection itself [9].

Common approaches to NMF obtain an approximation of V by computing a (W,H) pair to minimize the Frobenius norm of the difference $V - WH$. The problem can be cast in the following way [18] — let $V \in R^{m \times n}$ be a nonnegative matrix and $W \in R^{m \times k}$ and $H \in R^{k \times n}$ for $0 < k << min(m, n)$. Then, the objective function or minimization problem can be stated as

$$\min_{W,H} \|V - WH\|_F^2, \tag{3.1}$$

with $W_{ij} > 0$ and $H_{ij} > 0$ for each i and j.

The matrices W and H are not unique. Usually H is initialized to zero and W to a randomly generated matrix where each $W_{ij} > 0$ and these initial estimates are improved or updated with alternating iterations of the algorithm. In the following sections some existing NMF techniques are discussed and a new algorithm is proposed.

3.1 Multiplicative method

The NMF method proposed by Lee and Seung is based on multiplicative update rules of W and H. This scheme is referred to as the multiplicative method (MM). Algorithm 3.1.1 contains a formal statement of the method [18].

Algorithm 3.1.1: Algorithm for MM

1. Initialize W and H with nonnegative values, and scale the columns of W to unit norm.

2. Iterate for each c, j, and i until convergence or after l iterations:

   (a) $H_{cj} \leftarrow H_{cj} \frac{(W^T V)_{cj}}{(W^T W H)_{cj} + \epsilon}$

   (b) $W_{ic} \leftarrow W_{ic} \frac{(V H^T)_{ic}}{(W H H^T)_{ic} + \epsilon}$

   (c) Scale the columns of W to unit norm.

In steps 2(a) and (b), ϵ, a small positive parameter equal to 10^{-9}, is added to avoid division by zero. As observed from Algorithm 3.1.1, W and H remain nonnegative during the updates. Simultaneous updating of W and H generally yield better results than updating each matrix factor fully before the other. In the algorithm, the columns of W or the basis vectors are normalized at each iteration; in case of W, the optimization is performed on a unit hypersphere with the columns of W effectively being mapped to the surface of the hypersphere by

repeated normalization [18].

The computational complexity of MM can be shown to be $O(kmn)$ operations (for a rank-k approximation) per iteration [18]. Once the term-by-document matrix V has been factored into W and H, if new data needs to be added, then the data can be a direct addition to W with a minor modification to H if k is not fixed. In case of a fixed k, the new data can be integrated by further iterations with W and H as the initial approximations. In [14] it is shown by Lee and Seung that under the MM-update rules, the objective function (3.1) is monotonically non-increasing and becomes constant if and only if W and H are at a stationary point. This multiplicative method is related to expectation-maximization approaches used in image restoration, e.g. [19], and can be classified as a diagonally-scaled gradient descent method [9].

3.2 Sparse Encoding

A new nonnegative sparse encoding scheme, based on the study of neural networks has been suggested by Hoyer [10]. This scheme is applicable to the decomposition of datasets into independent feature subspaces by Hyvärinen and Hoyer [11]. The method proposed by Hoyer [10] has an important feature that enforces a statistical sparsity of the H matrix. As the sparsity of H increases, the basis vectors become

more localized, i.e., the parts-based representation of the data in W become more and more enhanced. Mu, Plemmons and Santago [17] have put forth a regularization approach that achieves the same objective of enforcing statistical sparsity of H by using a point-count regularization scheme that penalizes the number of non-zero entries rather than the sum of entries $\sum_{ij} H_{ij}$ in H.

3.3 A Hybrid Method

The NMF algorithm used in this study [18] is a hybrid method that combines some of the better features of the methods discussed in the previous sections. In this approach, the multiplicative method, which is basically a version of the gradient descent optimization scheme, is used at each iterative step to approximate the basis vector matrix W. H is calculated using a constrained least squares (CLS) model as the metric. It serves to penalize the non-smoothness and non-sparsity of H; as a result of this penalization, the basis vectors or topics in W become more localized, thereby reducing the number of vectors needed to represent each document. The method for approximating H is similar to the methods described in [10] and [17] and related to the least squares Tikhonov regularization technique commonly used in image restoration [19]. This hybrid algorithm is denoted by GD-CLS (gradient descent with constrained least squares) in [18].

3.4 Algorithm for GD-CLS

1. Initialize W and H with nonnegative values, and scale the columns of W to unit norm.

2. Iterate until convergence or after l iterations:

   (a) $W_{ic} \leftarrow W_{ic} \frac{(VH^T)_{ic}}{(WHH^T)_{ic} + \epsilon}$, for c and i [$\epsilon = 10^{-9}$]

   (b) Rescale the columns of W to unit norm

   (c) Solve the constrained least squares problem:

$$\min_{H_j} \{\|V_j - WH_j\|_2^2 + \lambda \|H_j\|_2^2\},$$

   where the subscript j denotes the j^{th} column, for $j = 1, \ldots, m$. Any negative values in H_j are set to zero. The parameter λ is a regularization value that is used to balance the reduction of the metric

$$\|V_j - WH_j\|_2^2$$

   with enforcement of smoothness and sparsity in H.

Chapter 4

Software Implementation

Two software packages, namely GTP and LAPACK, are used in the C-based implementation of GD-CLS — the NMF algorithm used in this study.

4.1 GTP

The General Text Parser (GTP) is a software environment developed at the University of Tennessee by Giles et al. [8]. One of the functions of GTP is to parse text documents and construct a sparse matrix data structure, i.e., a term-by-document matrix that defines the relationship between the documents and the parsed terms [16]. The GTP software can be used to parse single files or entire directories and is fitted with the capability to process both raw text and HTML files. The user can also integrate external filters into the software to process other forms of tagged data. Currently there are two versions of the software available

— one in C++ and another in Java — both of which are designed to facilitate users with all ranges of expertise. For this study, the C++ version of GTP was used.

4.2 LAPACK

The linear algebra package LAPACK was developed and is maintained by the Innovative Computing Lab (ICL) at the University of Tennessee, Knoxville, and is used to solve linear algebra problems. LAPACK has different routines, which can be individually downloaded from the LAPACK website[1], for solving different types of linear equations. For the C version of NMF, the *dposv* software routine of LAPACK is used to derive solutions (in double precision) to linear systems of the form AX = B, where A is a symmetric positive definite matrix.

[1] http://www.netlib.org/lapack/

Chapter 5

Performance

Originally written in Matlab by Pauca and Plemmons (see Appendix A), the proposed NMF algorithm or GD-CLS has been converted to C in this study for scalability. Performance evaluations are conducted using two different datasets — the Reuters Document Corpus and TDT2. This chapter comprises a description of the methodology used for evaluation, while the actual results[1] are discussed in Chapter 6.

5.1 Reuters

The Reuters data corpus[2], contains 21578 documents and 135 topics or document clusters created manually and each document in the corpus is been assigned one or

[1] All results are collected on a Sun Microsystems SunBlade 1000 workstation with 500 MHz UltraSPARC-IIe processor, 256KB L2 cache, 512MB DRAM and 20GB internal disk.

[2] Reuters-21578 at http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html.

more topics or category labels based on its content. The manually created cluster sizes, i.e., the number of documents assigned to the topics, range anywhere from less than ten to nearly four thousand topics. The documents are in SGML format (see Appendix D) with meta tags denoting title, topic(s), and beginning and end of content.

For this experiment, documents associated with only one topic are used and topics with cluster sizes smaller than five are discarded. To achieve this, a Perl script is used to traverse through the corpus and create an index of topics with associated cluster sizes, where a document is considered part of a cluster only if it has a single topic. Example 5.1.1 illustrates the steps to creating the index.

Example 5.1.1.

Data corpus, $D = \{doc_1, doc_2, doc_3, doc_4, doc_5\}$
Topic set, $T = \{A, B, C\}$

```
Script (D, T) {
    TopicList = { } // TopicList[TopicName] = Cluster Size
    for each document doc_i in D {
        if doc_i has only 1 topic X {
            if( topic X not in TopicList) TopicList[X] = 1
            else increment TopicList[X]
        }
    }
}
```

The relationship between D and T is shown in Table 5.1 and the generated topic

Table 5.1: Document-topic relationship in Example 5.1.1

Document	Topic(s)
doc_1	A
doc_2	A, B
doc_3	C
doc_4	A
doc_5	A, C

Table 5.2: Topic list produced for Example 5.1.1

Topic	Cluster Size
A	2
C	1

list is displayed in Table 5.2.

This preprocessing script is used only once to generate the index or topic list. The topic list, thus generated, contains 50 topics with cluster sizes ranging from 3735 to 5 (see Appendix B). Topics with fewer than five documents are discarded from the list. Once the list from the document collection is generated, the GTP software creates a term-by-document matrix in Harwell-Boeing (HB) format [2] from the document collection. The HB matrix is then used in the NMF algorithm to automatically generate clusters. However, instead of using the entire document collection spanned by the 50 chosen topics, several subsets of the document collection with varying number of topics are created in order to observe the performance of GD-CLS as the number of topics increases. The subsets are created by adding

a filter to the GTP software. This filter processes the Reuters SGML files as text stream input and uses a *topicfile* containing a list of chosen topics to suppress documents that do not belong to the chosen topics. Thus, by modifying the topic file, i.e., adding or deleting topics to this file, various document subsets can be created. For instance, in case of Example 5.1.1, if the topic file contains topic A, only documents 1 and 4 are used to create the term-by-document HB matrix, while if the topic file contains topics A and C, documents 1, 3, and 4 are used to create the HB matrix.

In order to observe the performance of the GD-CLS implementation of NMF as the complexity of the problem increases, i.e., as the number of clusters or the parameter k is incremented, seven different k values 2, 4, 6, 8, 10, 15, 20 are chosen. For each k, three different document collections or subsets are generated by the filter using different topic files, which result in creation of three term-by-document HB matrices for each k. After the HB matrices are generated, the NMF clustering algorithm is performed on all 21 matrices ($7k$ values $\times$ 3 document subsets each) to produce the W and H factors for each HB matrix. For any given HB matrix V, with k topics and n documents, matrix W has k columns or basis vectors that represent the k clusters, while matrix H has n columns that represent the n documents. A column vector in H has k components, each of which denotes the contribution of the corresponding basis vector to that column or document.

The classification or clustering of documents is then performed based on the index of the highest value of k for each document. So, for document i ($i = 1, ..., n$), if the maximum value is the jth entry ($j = 1, ..., k$), document i is assigned to cluster j. After the documents are clustered into k topics, the NMF generated k clusters are compared to the original k clusters using a mapping function. The mapping is performed using a Perl script that assigns the original cluster labels to the NMF clusters based on a similarity measure. Example 5.1.2 provides an explanation of the mapping process for $k = 2$.

Example 5.1.2.

Original Topic Set $T = \{A, B\}$
Document subset $D = \{d_1, d_2, d_3, d_4, d_5\}$
$Cluster_A = \{d_2, d_3\}$, $Cluster_B = \{d_1, d_4, d_5\}$

Using GD-CLS on the HB matrix generated from D with topic set T yields WH, where $W \in R^{m \times 2}$ and $H \in R^{k \times 7}$. Assuming H has the value shown in Table 5.3, the clustering based on the maximum column entry is

$Cluster_1 = \{d_2, d_3, d_5\}$,
$Cluster_2 = \{d_1, d_4\}$.

The values of the mapping function is used to form a matrix A (Table 5.4) where $A_{iX} = similarity(Cluster_i, Cluster_X)$ = number of documents in $Cluster_i$ that appear in $Cluster_X$, $i = (1, 2)$ and $X = \{A, B\}$.

Each $Cluster_i$ is assigned the original cluster label to which it is the most similar. $Cluster_1$ and $Cluster_2$ are assigned labels A and B respectively and the documents are reassigned to topics based on the new clustering. A comparison of the original clustering to the GD-CLS generated cluster labels is shown in Table 5.5.

Table 5.3: The 2×5 H matrix for Example 5.1.2. (maximum entries are represented in boldface)

d_1	d_2	d_3	d_4	d_5
0.3	**1.2**	**0.2**	0.01	**2.1**
1.4	0.9	0.01	**1.4**	1.9

Table 5.4: The A matrix for Example 5.1.2

	$Cluster_A$	$Cluster_B$
$Cluster_1$	2	1
$Cluster_2$	2	1

Table 5.5: Comparison between original cluster labels and GD-CLS generated labels for Example 5.1.2

Document	Original label	GD-CLS label
d_1	B	B
d_2	A	A
d_3	A	A
d_4	B	B
d_5	B	A

Once the relabeling is accomplished, the accuracy of the classification or clustering is assessed using the metric AC [22] defined by

$$\mathrm{AC} = \sum_{i=1}^{n} \delta(d_i)/n,$$

where $\delta(d_i)$ is set to 1 if d_i has the same topic label for both NMF and the original classification, and set to 0 otherwise, and n is the number of documents in the collection. So, for Example 5.1.2

$$\begin{aligned}\mathrm{AC} &= \{\delta(d_1) + \delta(d_2) + \delta(d_3) + \delta(d_4) + \delta(d_5)\}/5 \\ &= \{1+1+1+1+0\}/5 = 4\ /\ 5 = 0.8.\end{aligned}$$

In the GD-CLS implementation of NMF, the contribution of the λ parameter with which the sparsity of H is controlled is also of interest. Hence, for each k, results for three different λ values (0.1, 0.01, 0.001) are calculated.

5.2 TDT2

The second data corpus TDT2, obtained from the Language Data Consortium at The University of Pennsylvania[3], contains transcripts from a total of six news sources[4] in 3440 files, with each file containing several transcripts or documents. Although the corpus consists of about sixty-four thousand documents in SGML format (see Appendix E), some fourteen thousand of these are actually assigned

[3] http://www.lcd.upenn.edu.
[4] ABC, CNN, VOA, NYT, PRI, and APW.

a topic label and the rest are not classified. Among the preclassified documents, 7919 documents are single topic documents, i.e., these documents only have a single topic or category label. The SGML markup tags for each document denote a unique document ID or identification number and the beginning and end of text content. The document-topic relationships are described in a separate file that contains a line in it for each document with a category label. A line corresponding to a particular document consists of the document ID, topic label, and the name of the file containing that document.

In order to make the document collection from this corpus comparable to the Reuters dataset, some preprocessing with the use of Perl scripts is applied to the SGML files. First, the file containing the document-topic relationships is parsed and a *topic file* or a file containing a list of 73 topics that have cluster sizes of at least five documents is created. Here also, as with the Reuters collection, documents containing multiple topic labels are not deemed relevant. Since the entire document corpus consists of 64,000 documents and only 7919 are relevant to the experiments, another preprocessing step is taken to reduce the runtime of GTP by traversing the entire collection once and writing the relevant documents to a single file. For all subsequent testing, only this file is then used in order to avoid traversing thousands of irrelevant documents for each test run. Once the topic file and the reduced set of 7919 documents are at hand, several subsets are

created to monitor the decline of accuracy for the NMF algorithm as complexity or the k values increase. As before, 7 different k values (2, 4, 6, 8, 10, 15, 20) are chosen with 10 different topic sets or document subsets each. After application of GD-CLS and the accuracy metric, this selection of datasets produces results presented in the following chapter.

Chapter 6

Observations

The results from TDT2 and Reuters data corpora bring to attention trends such as the decline in accuracy in relation to the increase in complexity or the value of k. Results from both document collections indicate that as more and more topics or document clusters are added to the dataset being clustered by GD-CLS, the accuracy of the clustering decreases. For the Reuters collection, in case of $k = 2$, i.e., when dealing with only two topics, the algorithm performs with above 99% accuracy, but in case of $k = 20$, the accuracy drops down to just above 54% (Table 6.1). However, in case of TDT2, the drop in accuracy is much less precipitous than for Reuters (Table 6.2). For TDT2, for $k = 20$, accuracy is just above 80%, which seems like a significant improvement from 54% for Reuters. This disparity can be attributed to the differences in content of the two collections. Documents in the Reuters collection are categorized under broad topics

(such as "earn," "interest," "cocoa," "potato," etc., listed in Appendix B), while for TDT2, the topic labels are much more specific ("The Asian economic crisis," "Tornado in Florida," "Oprah lawsuit," etc., listed in Appendix C). The very specificity of the topics in the TDT2 guarantees a heterogeneity in the document collection that is not present in the Reuters collection. In the case of Reuters, while "potato" and "zinc" may constitute very distinct clusters, "interest" and "money-fixes" do not. In fact, as noted by Xu et al. [22], there is a degree of overlapping of content across topics in the Reuters collection that contributes to the much more rapid decline of accuracy in case of Reuters than it does for TDT2.

Another notable trend that also points to the sensitivity of the GD-CLS algorithm for NMF to the contents of the document collections is the differences in accuracy for the different λ values. In case of TDT2, the different λ values for each k do not affect the performance by any noticeable amount. But for Reuters, the drop in accuracy for increasing values of the λ parameter suggests that text collections that are somewhat homogeneous in content, are more sensitive to the changes of the λ parameter (or the sparsity of the H matrix). The primary reason for using a larger λ value (or an increase in the sparsity of H) is to achieve faster computation times. Inspection of the results from Table 6.1 and 6.2 suggests that that is indeed the case, especially in higher complexity problems (Table 6.3).

Table 6.1: Results for Reuters (AC = Accuracy measure defined in Section 5.1)

k	λ	AC	CPU time (sec)
2	0.100	0.962256	2.63
2	0.010	0.963440	2.76
2	0.001	0.962262	3.19
4	0.100	0.758630	3.86
4	0.010	0.774503	4.43
4	0.001	0.777460	5.51
6	0.100	0.716229	6.51
6	0.010	0.722549	8.01
6	0.001	0.726186	10.54
8	0.100	0.572499	9.73
8	0.010	0.555926	12.79
8	0.001	0.560444	18.39
10	0.100	0.657349	30.65
10	0.010	0.673601	36.79
10	0.001	0.666243	47.75
15	0.100	0.609148	56.53
15	0.010	0.613033	74.89
15	0.001	0.618249	104.18
20	0.100	0.545806	57.26
20	0.010	0.567711	87.77
20	0.001	0.571387	122.13

Table 6.2: Results for TDT2 (AC = Accuracy measure defined in Section 5.1)

k	λ	AC	CPU time (sec)
2	0.100	0.993629	2.93
2	0.010	0.993629	2.94
2	0.001	0.978329	3.00
4	0.100	0.906264	9.42
4	0.010	0.908873	9.48
4	0.001	0.925784	10.04
6	0.100	0.878919	23.38
6	0.010	0.858782	23.60
6	0.001	0.860544	25.81
8	0.100	0.858497	46.86
8	0.010	0.859123	47.42
8	0.001	0.853479	52.48
10	0.100	0.840443	97.39
10	0.010	0.836955	98.34
10	0.001	0.847155	110.26
15	0.100	0.869069	135.66
15	0.010	0.872499	140.08
15	0.001	0.870932	172.06
20	0.100	0.832097	303.54
20	0.010	0.835903	315.64
20	0.001	0.840977	405.16

Table 6.3: CPU time for $k = 15$ for different λ values

λ	Reuters	TDT2
0.1	56.5433	135.6620
0.01	66.0033	140.0820
0.001	93.3900	172.0670

It can be inferred from Table 6.3 that an increase in the sparsity of H results in a significant increase in computational speed and this holds for both TDT2 and Reuters. As for accuracy, the λ values do affect performance for Reuters but not for TDT2. However, when compared to the gain in computational time, the 2 to 3% decrease in accuracy can be considered a very reasonable tradeoff.

An aspect of GD-CLS that cannot be directly observed from the result tables is the change in performance of the factorization with regards to disparate cluster sizes. When creating document subsets for each value of k from the preclassified clusters of the Reuters or TDT2 corpus, attention is given to keep the cluster sizes within a reasonable bound of one another. This constraint, which is not imposed by Xu et al. in [22], is enforced due to results obtained from experiments similar to those described in Table 6.4.

The imbalance in the cluster sizes in $dataset_1$ has a definite effect on the performance of GD-CLS regardless of the document corpus being used. In case of

Table 6.4: A comparison of results with different cluster sizes

Corpus	Dataset	Cluster	Original cluster sizes	GD-CLS generated cluster sizes
Reuters	$dataset_1$	$cluster_1$	2125	1690
		$cluster_2$	45	480
	$dataset_2$	$cluster_1$	114	112
		$cluster_2$	99	101
TDT2	$dataset_1$	$cluster_1$	1476	1231
		$cluster_2$	31	276
	$dataset_2$	$cluster_1$	110	109
		$cluster_2$	120	121

the original clusters from $dataset_1$, the ratio of $cluster_1$ to $cluster_2$ is approximately 48:1, while the clusters produced by GD-CLS have a ratio of 3:1. This implies GD-CLS performs much better on datasets that have balanced cluster sizes, such as $dataset_2$, where clustering is performed with almost 100% accuracy.

Chapter 7

Conclusion and Future Work

This study demonstrates how GD-CLS, a hybrid NMF algorithm, can be effectively used to classify text collections in an unsupervised or automated manner. The proposed algorithm can be used to construct a parts-based representation of the text data, in which the localization of the parts or features can be regularized to create a balance between computational cost and accuracy.

In its current stage, the GD-CLS algorithm for NMF is not equipped to handle updating in an efficient manner. Once the document collection has been clustered via NMF, adding a small number of documents to the collection can be achieved by comparing each of the new documents (represented by a vector) to the basis vectors and associating the new document to the basis vector or topic to which it is the most similar. But this updating technique is not scalable and would produce poor results if used to add a large number of documents that cannot be

associated with any of the basis vectors.

NMF, in general, has mostly been applied to image analysis and text mining. Another field that could benefit from this technique is bioinformatics. Problems such as identifying motifs or significant features in protein sequences (partial strings of DNA) are a natural candidate for application of NMF. In such problems, protein sequences can be viewed as analogous to text documents and the basis vectors or topics to motifs that control gene expression [20].

Strictly from a usability standpoint, the NMF software can be fitted with a better user interface to enable users easier access to clusters and perhaps also create tools for query matching. Although the primary function of NMF is not information retrieval but actual classification, the clusters can be used to provide retrieval capabilities. Much in the style of limited updating discussed earlier, a user query can be represented by a term vector, which is then used to compute a similarity measure (using cosine measurement) between the query and the basis vectors. The basis vector or topic that yields the highest value is deemed the most relevant and documents belonging to that topic is provided to the user.

Bibliography

Bibliography

[1] A. Berman and R. Plemmons. *Non-Negative Matrices in the Mathematical Sciences*, SIAM Press Classics Series, Philadelphia, 1994.

[2] M. Berry and M. Browne. *Understanding Search Engines: Mathematical Modeling and Text Retrieval*, SIAM, Philadelphia, 1999.

[3] M. Berry, Z. Drmač, and E. Jessup. "Matrices, Vector Spaces, and Information Retrieval," *SIAM Review*, Vol. 41, pp. 335-362, 1999.

[4] *Concise Columbia Encyclopedia.* Columbia University Press, New York, Second Edition, 1989.

[5] M. Cooper and J. Foote, "Summarizing Video using Non-Negative Similarity Matrix Factorization," *Proc. IEEE Workshop on Multimedia Signal Processing* St. Thomas, US Virgin Islands, 2002.

[6] D. Donoho and V. Stodden. "When does Non-Negative Matrix Factorization Give a Correct Decomposition into Parts?," preprint, Department of Statistics, Stanford University, 2003.

[7] M.H. Dunham. *Data Mining: Introductory and Advanced Topics.* Prentice Hall, Upper Saddle River, NJ, 2003.

[8] J.T. Giles, L. Wo, and M.W. Berry. "GTP (General Text Parser) Software for Text Mining," in *Statistical Data Mining and Knowledge Discovery*, H. Bozdogan (Ed.), CRC Press, Boca Raton, FL, pp. 455-471, 2003.

[9] D. Guillamet and J. Vitria. "Determining a Suitable Metric when Using Non-Negative Matrix Factorization," *16th International Conference on Pattern Recognition (ICPR'02)*, Vol. 2, Quebec City, QC, Canada, 2002.

[10] P. Hoyer. "Non-Negative Sparse Coding," *Neural Networks for Signal Processing XII (Proc. IEEE Workshop on Neural Networks for Signal Processing)*, Martigny, Switzerland, 2002.

[11] A. Hyvärinen and P. Hoyer. "Emergence of Phase and Shift Invariant Features by Decomposition of Natural Images into Independent Feature Subspaces," *Neural Computation*, Vol. 12, pp. 1705-1720, 2000.

[12] I. Jolliffe. *Principle Component Analysis*, 2nd Ed., Springer Series in Statistics, Springer-Verlag, New York, 2002.

[13] D. Lee and H. Seung. "Learning the Parts of Objects by Non-Negative Matrix Factorization," *Nature*, Vol. 401, pp. 788-791, 1999.

[14] D. Lee and H. Seung. "Algorithms for Non-Negative Matrix Factorization," *Advances in Neural Information Processing Systems*, Vol. 13 , pp. 556-562, 2001.

[15] W. Liu and J. Yi. "Existing and New Algorithms for Non-negative Matrix Factorization," preprint, Computer Sciences Department, The University of Texas at Austin, 2003.

[16] S. Mironova. "Integrating Network Storage into Information Retrieval Applications," Master's Thesis, Department of Computer Science, The University of Tennessee, Knoxville, 2003.

[17] Z. Mu, R. Plemmons and P. Santago. "Iterative Ultrasonic Signal and Image Deconvolution for Estimating the Complex Medium Response," preprint, submitted to *IEEE Transactions on Ultrasonics and Frequency Control*, 2003.

[18] V. Pauca, F. Shahnaz, M. Berry, and R. Plemmons. "Text Mining Using Non-Negative Matrix Factorizations," *Proceedings of the Fourth SIAM International Conference on Data Mining*, Lake Buena Vista, FL, April 22-24, 2004.

[19] S. Prasad, T. Torgersen, V. Pauca, R. Plemmons, and J. van der Gracht. "Restoring Images with Space Variant Blur via Pupil Phase Engineering," Optics in Info. Systems, Special Issue on Comp. Imaging, SPIE Int. Tech. Group Newsletter, Vol. 14, No. 2, pp. 4-5, 2003.

[20] G.W. Stuart and M.W. Berry. "Comprehensive Whole Genome Bacterial Phylogeny Using Correlated Peptide Motifs Defined in a High Dimensional Vector Space," *Journal of Bioinformatics and Computational Biology*, Vol. 1, No.3, pp. 475-493, 2003.

[21] S. Wild, J. Curry and A. Dougherty. "Motivating Non-Negative Matrix Factorizations," *Proceedings of the Eighth SIAM Conference on Applied Linear Algebra*, Williamsburg, VA, July 15-19, `http://www.siam.org/meetings/la03/proceedings`, 2003.

[22] W. Xu, X. Liu, and Y. Gong. "Document-Clustering based on Non-Negative Matrix Factorization," *Proceedings of SIGIR'03*, July 28-August 1, Toronto, CA, pp. 267-273, 2003.

Appendices

Appendix A

GD-CLS algorithm in Matlab

```
[W, H] = gdcls(V, k, maxiter, lambda, options)

myeps = 10^-9;
if strcmp(options, 'nonneg')
   neg = 1;
else
   neg = 0;
end
[m,n] = size(V);
W = rand(m, k);
H = zeros(k, n);
for j = 1 :  maxiter,
    A = W' * W + lambda * eye(k);
    for i = 1 :  n
      b = W' * V(:,i);
     H(:,i) = A / b;
    end
   if neg == 1
     H = H .* (H > 0);
    end
   W = W .* (V * H') ./ (W * (H * H') + myeps);
   W = W ./ (ones(m, 1) * sum(W));
end
```

Appendix B

Topic File for Reuters

acq 2125	earn 3735	interest 211	money-supply 97	silver 11
alum 45	fuel 11	ipi 41	nat-gas 42	strategic-metal 19
bop 24	gas 22	iron-steel 46	oilseed 9	sugar 135
carcass 11	gnp 73	jobs 48	orange 18	tea 6
cocoa 55	gold 99	lead 8	pet-chem 21	tin 30
coffee 114	grain 45	lei 10	potato 5	trade 333
copper 54	heat 14	livestock 20	reserves 42	veg-oil 37
cotton 26	housing 15	lumber 12	retail 18	wpi 23
cpi 68	income 6	meal-feed 7	rubber 39	yen 6
crude 355	instal-debt 5	money-fx 259	ship 156	zinc 15

Appendix C

Topic File for TDT2

Rev Lyons Arrested 5
Great Lake Champlain?? 5
Capps Replacement Elections 5
Nazi-plundered Art 5
$1 million Stolen at WTC 6
Tello Maryland Murder 6
Strike in Germany 6
Marcus Allen Retires 6
Mountain Hikers Lost 7
Spanish Dam Broken 7
Buffett buys Silver 8
POW Memorial Museum 8
DiBella Treatment CURES Cancer? 8
Job incentives 8
Saudi Soccer coach sacked 8
Food Stamps 9
Akin Birdal Shot and Wounded 9
Cubans returned home 9
Goldman Sachs - going public? 9
JJ the Whale 11
Mary Kay LeTourneau 12
Race Relations Meetings 12
Puerto Rico phone strike 13
Anti-Chinese Violence in Indonesia 14
Fossett's Balloon Ride 15
Dr Spock Dies 15
Denmark Strike 15
David Satcher confirmed 16
Bird Watchers Hostage 16
Tony Awards 16
World Figure Skating Champs 17
McVeigh's Navy Dismissal and Fight 19
Babbitt Casino Case 20
World AIDS Conference 21
Afghan Earthquake 23
Grossberg baby murder 26
Diane Zamora 30
Asteroid Coming?? 31
Quality of Life-NYC 33
State of the Union Address 34
China Airlines Crash 36
John Glenn 37
Shevardnadze Assassination Attempt 38
Upcoming Philippine Elections 41

Karla Faye Tucker 48
James Earl Ray's Retrial? 49
Tornado in Florida 53
German Train derails 54
Casey Martin Sues PGA 56
Rats in Space! 60
Nigerian Protest Violence 61
Oprah Lawsuit 70
NBA finals 83
Superbowl '98 84
Clinton-Jiang Debate 84
Viagra Approval 93
Bombing AL Clinic 99
Cable Car Crash 110
India Parliamentary Elections 120
Unabomber 120
Violence in Algeria 125
Jonesboro shooting 125
Segt Gene McKinney 126
GM Strike 142
Pope visits Cuba 151
Israeli-Palestinian Talks (London) 210
National Tobacco Settlement 281
Anti-Suharto Violence 297
India - A Nuclear Power? 473
Winter Olympics 535
Monica Lewinsky Case 954
Asian Economic Crisis 1083
Current Conflict with Iraq 1476

Appendix D

SGML Tags for Reuters Files

```
<REUTERS TOPICS="" LEWISSPLIT="" CGISPLIT=""
OLDID="" NEWID="">
   <DATE> </DATE>
   <TOPICS> </TOPICS>
   <PLACES> </PLACES>
   <PEOPLE> </PEOPLE>
   <ORGS> </ORGS>
   <EXCHANGES> </EXCHANGES>
   <COMPANIES> </COMPANIES>
   <UNKNOWN> </UNKNOWN>
   <TEXT>
     <TITLE> </TITLE>
     <DATELINE> </DATELINE>
     <BODY> </BODY>
   </TEXT>
</REUTERS>
```

Appendix E

SGML Tags for TDT2 Files

```
< DOC >
   < DOCNO >  < /DOCNO >
   < DOCTYPE >  < /DOCTYPE >
   < TXTTYPE >  < /TXTTYPE >
   < TEXT >  < /TEXT >
< /DOC >
```

Vita

Farial Shahnaz was born in Dhaka, Bangladesh on March 19, 1979. After receiving her high school diploma from Holy Cross Girls' High School, Dhaka, Bangladesh, she attended the University of Tennessee at Knoxville, where she graduated with a Bachelor of Science degree in Mathematics and Computer Science in May 2002.

1331 4702 7
10/20/04 MRB