

**Text Clustering and Active Learning Using a  
LSI Subspace Signature Model and Query Expansion**

A Thesis

Submitted to the Faculty of

Drexel University

by

Weizhong Zhu

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

June 2009

**©Copyright 2009**  
**Weizhong Zhu. All Rights Reserved.**

## ACKNOWLEDGEMENTS

I appreciate the support of my Dissertation Committee, my family, and the College.

Firstly, I would like to express my greatest thanks to my supervisor, Dr. Robert B. Allen, who has guided my research directions and shaped my goals with his profound knowledge in diverse fields of information studies. My future career and personal life will benefit from his advice. I am also very grateful to my committee members, Dr. Xia Lin, Dr. Xiaohua Hu, Dr. Christopher C. Yang and Dr. Michael Christel from CMU. As mentors, Dr. Lin and Dr. Hu raised my interests in the major topics of the dissertation. Dr. Yang and Dr. Christel provided very valuable suggestions to accomplish the dissertation.

My collaboration with Dr. Chaomei Chen promoted my understanding on the research issues for information studies and motivated several key ideas of the dissertation. Dr. Il-Yeol Song is the first one to teach me how to be a scholar, to write more polished research papers and to make good conference presentation. Without the encouragement of these professors, my dream would not have come true.

Secondly, I would like to thank my mother, Guangrong Sun, and my parents-in-law, Hongming Cui and Fengying Li, for their selfless support and encouragement when I face problems. I would like to express many thanks to my wife, Weihong Cui, for her love and sacrifice. I dedicate this dissertation to my little girl, Ivy. Her lovely smiles

always refresh my mind and have given me energy to continue the five-year journey in the PhD program. Without the support of my family, the journey is endless.

Finally, the researches in this dissertation have been supported in part by the National Visualization and Analytics Center (NVAC) through the Northeast Visualization and Analytics Center (NEVAC), the National Science Foundation (NSF) under Grant No. SEIII-0612129 and an Institute of Museum and Library Services (IMLS) grant. My travels to attend many academic conferences are generously sponsored by the College of Information Science and Technology and the Office of Graduate Studies at Drexel University. These financial supports helped me to complete the research successfully.

## TABLE of CONTENTS

<b>LIST of TABLES .....</b>	<b>viii</b>
<b>LIST of FIGURES .....</b>	<b>xii</b>
<b>ABSTRACT .....</b>	<b>xiv</b>
<b>CHAPTER1: Introduction.....</b>	<b>1</b>
<b>1.1 Latent Semantic Indexing .....</b>	<b>4</b>
<b>1.2 Visual Exploration of the LSI Subspaces.....</b>	<b>8</b>
<b>1.3 Research Questions.....</b>	<b>13</b>
<b>1.4 Organization of the Thesis .....</b>	<b>15</b>
<b>CHAPTER2: Literature Survey .....</b>	<b>16</b>
<b>2.1 Data Models for Semantic Content Representation .....</b>	<b>16</b>
<b>2.2 Text Clustering.....</b>	<b>21</b>
<b>2.3 Active Learning.....</b>	<b>23</b>
<b>2.4 Query Expansion.....</b>	<b>25</b>
<b>2.5 Social Network Analysis.....</b>	<b>27</b>
<b>CHAPTER3: The LSI Subspace Signature Model.....</b>	<b>29</b>
<b>3.1 LSI Subspace Term Signatures and Document Signatures .....</b>	<b>31</b>
<b>3.2 LSI Subspace Signature Ranking.....</b>	<b>35</b>
<b>3.2.1 Introduction.....</b>	<b>35</b>

3.2.2	Experiments, Text Preprocessing and Data Sets .....	36
3.2.3	Experiment Results.....	38
3.2.4	Conclusion .....	41
3.3	Similarity Measures between Signatures.....	42
<b>CHAPTER4: Text Clustering Using LSISSM .....</b>		<b>45</b>
4.1	Introduction.....	45
4.1.1	Model-based K-means and SOM.....	46
4.1.2	Model-based Two-stage K-means.....	47
4.2	Evaluation Methods and Datasets .....	50
4.3	Experiment Results.....	52
4.3.1	K-means Using Top Ranking Terms.....	52
4.3.2	K-means Using Top Ranking Term Clusters .....	56
4.3.3	Two-stage K-means.....	58
4.3.4	SOM Using Top Ranking Terms.....	60
4.3.5	Comparisons with VSM and Traditional LSI.....	62
4.3.6	Tuning the Model Parameters .....	65
4.4	Conclusions for Text Clustering.....	68
<b>CHAPTER5: Active Learning Using LSISSM.....</b>		<b>69</b>
5.1	Introduction.....	69

<b>5.2</b>	<b>Evaluation Methods, Text Preprocessing and Datasets .....</b>	<b>69</b>
<b>5.3</b>	<b>Experiment Results.....</b>	<b>71</b>
<b>5.3.1</b>	<b>Sampling Selection with Outlier Categories.....</b>	<b>71</b>
<b>5.3.2</b>	<b>The Effect of the Model-based Dimension Reduction on the Text Classifiers.....</b>	<b>75</b>
<b>5.3.3</b>	<b>The Effect of the Model-based Sample Reduction on the Learning Curves of the Text Classifiers .....</b>	<b>77</b>
<b>5.3.4</b>	<b>Combined Effect of the Model-based Dimension Reduction and Sample Reduction on the Text Classifiers .....</b>	<b>82</b>
<b>5.4</b>	<b>Conclusions for Active Learning .....</b>	<b>84</b>
<b>CHAPTER6: Query Expansion Using Domain Ontologies .....</b>		<b>86</b>
<b>6.1</b>	<b>Term Re-weighting Strategy with UMLS Co-concepts and Synonyms</b>	<b>86</b>
<b>6.1.1</b>	<b>Query Expansion Strategies.....</b>	<b>87</b>
<b>6.1.2</b>	<b>Evaluation Methods and Datasets .....</b>	<b>90</b>
<b>6.1.3</b>	<b>Experiment Results.....</b>	<b>90</b>
<b>6.1.4</b>	<b>Conclusions for Query Expansion Strategies.....</b>	<b>92</b>
<b>6.2</b>	<b>User Relevance Feedback Expanded by the IPTC Hierarchical Structure.....</b>	<b>94</b>
<b>6.2.1</b>	<b>Introduction.....</b>	<b>94</b>
<b>6.2.2</b>	<b>User Relevance Feedback Expanded by Hyponyms in IPTC Codes</b>	<b>94</b>

6.2.3	Evaluation Methods, OCR Preprocessing and Datasets .....	96
6.2.4	Experiment Results.....	98
6.2.5	Conclusion for the IPTC Study .....	99
<b>CHAPTER7: Weighted PageRank Enhanced by Betweenness Centrality .....</b>		<b>100</b>
7.1	Introduction.....	100
7.2	Evaluation Methods and Datasets .....	104
7.3	Experiment Results.....	105
7.4	Conclusion .....	109
<b>CHAPTER8: Overall Conclusions and Future Work.....</b>		<b>110</b>
8.1	Contributions of the Thesis .....	110
8.2	Potential Research Directions.....	112
<b>LIST of REFERENCES .....</b>		<b>115</b>
<b>VITA .....</b>		<b>122</b>

## LIST of TABLES

<b>Table 3-1: Top 20 Themes which Make the Highest Contribution to the LSI Term Subspace of the Reuters Corpus Ranked by GLCR. ....</b>	<b>39</b>
<b>Table 3-2: Discussion Topic Evolution in the URI Working Group from 1994 Dec to 1998.....</b>	<b>40</b>
<b>Table 3-3: Discussion Topic Evolution in the URI Working Group from 1999 Dec to 2004 May.....</b>	<b>41</b>
<b>Table 4-1: The Clustering Evaluation Matrix of the Basic K-means Using the Similarity CDSS and the Reuters Corpus. ....</b>	<b>53</b>
<b>Table 4-2: The Clustering Evaluation Matrix of the Basic K-means Using the Similarity CDSS and the TDT1 Corpus.....</b>	<b>54</b>
<b>Table 4-3: The Clustering Evaluation Matrix of the Basic K-means Using the Similarity CDSS and the TDT2 Corpus.....</b>	<b>55</b>
<b>Table 4-4: The Clustering Evaluation Matrix of the Basic K-means Using the Term Clusters and the Reuters Corpus.....</b>	<b>56</b>
<b>Table 4-5: The Clustering Evaluation Matrix of the Basic K-means Using the Term Clusters and the TDT1 Corpus. ....</b>	<b>57</b>
<b>Table 4-6: The Clustering Evaluation Matrix of the Basic K-means Using the Term Clusters and the TDT2 Corpus. ....</b>	<b>57</b>

<b>Table 4-7: Comparison of the Running Time (seconds) between the Basic K-means and the Two-stage K-means Using the Top Ranking Terms and the Top Ranking Term Clusters with the Reuters corpus, TDT1 and TDT2. ....</b>	<b>58</b>
<b>Table 4-8: Comparison of the Clustering Performance between the Basic K-means and the Two-stage K-means Using the Similarity CDSS and the Top Ranking Terms with the Reuters Corpus, TDT1 and TDT2.....</b>	<b>59</b>
<b>Table 4-9: Comparison of the Clustering Performance between the Basic K-means and the Two-stage K-means Using the Similarity CDSS and the Top Ranking Term Clusters with the Reuters Corpus, TDT1 and TDT2. ....</b>	<b>60</b>
<b>Table 4-10: The Clustering Evaluation Matrix of SOM Using the Similarity GCDSS and the Reuters Corpus. ....</b>	<b>60</b>
<b>Table 4-11: The Clustering Evaluation Matrix of SOM Using the Similarity GCDSS and the TDT1 Corpus.....</b>	<b>61</b>
<b>Table 4-12: The Clustering Evaluation Matrix of SOM Using the Similarity GCDSS and the TDT2 Corpus.....</b>	<b>61</b>
<b>Table 4-13: Comparison of the Clustering Performance between VSM and the Traditional LSI Using the Basic K-means with the Reuters Corpus, TDT1 and TDT2. ....</b>	<b>62</b>
<b>Table 4-14: Comparison of the Clustering Performance between VSM and the Traditional LSI Using the SOM with the Reuters Corpus, TDT1 and TDT2.....</b>	<b>62</b>

<b>Table 4-15: Comparison of the Clustering Performance between the VSM baseline and LSISSM Using the Basic K-means, the Similarity CDSS and the Top Ranking Terms with the Reuters Corpus, TDT1 and TDT2.....</b>	<b>63</b>
<b>Table 4-16: Comparison of the Clustering Performance between the VSM baseline and LSISSM Using the Basic K-means, the Similarity CDSS and the Top Ranking Term Clusters with the Reuters Corpus, TDT1 and TDT2. ....</b>	<b>63</b>
<b>Table 4-17: Comparison of the Clustering Performance between the VSM Baseline and LSISSM Using the SOM, the Similarity GCDSS and the Top Ranking Terms with the Reuters Corpus, TDT1 and TDT2.....</b>	<b>64</b>
<b>Table 5-1: The Sampling Distribution across 10 Categories of the Reuters Corpus Selected by GLCR. ....</b>	<b>71</b>
<b>Table 5-2: The Sampling Distribution across 25 Categories of the TDT1 Corpus Selected by GLCR.....</b>	<b>72</b>
<b>Table 5-3: The Sampling Distribution across 30 Categories of the TDT2 Corpus Selected by GLCR.....</b>	<b>73</b>
<b>Table 5-4: The Average Learning Accuracy of the Three Classifiers, Naïve Bayes, KNN and Rocchio, on the Full Training Sets of the Reuters Corpus, TDT1 and TDT2 Using Feature Reduction.....</b>	<b>76</b>
<b>Table 5-5: The Average Learning Accuracy of the Three Classifiers, Naïve Bayes, KNN and Rocchio, on the Independent Testing Sets of the Reuters Corpus, TDT1 and TDT2 Using Feature Reduction. ....</b>	<b>76</b>

<b>Table 5-6: Comparison of the Average Learning Accuracy between the Full Training Set and the Training Subsets of the Reuter Corpus Using the Three Classifiers, Naïve Bayes, KNN and Rocchio.....</b>	<b>77</b>
<b>Table 5-7: Comparison of the Average Learning Accuracy between the Full Training Sets and the Training Subsets of TDT1 and TDT2 Using the Three Classifiers, Naïve Bayes, KNN and Rocchio.....</b>	<b>80</b>
<b>Table 5-8: Comparison of the Average Learning Accuracy of the Three Classifiers, Naïve Bayes, KNN and Rocchio, Trained by the Training Subsets of the Reuter Corpus, TDT1 and TDT2 with Feature Reduction and Tested by the Independent Testing Sets. ....</b>	<b>82</b>
<b>Table 6-1: Comparison of P@200 between the VSM baseline and the User Relevance Feedback with the Hyponyms in IPTC Codes. ....</b>	<b>98</b>
<b>Table 7-1: Top 10 Actors in the URI Working Group Ranked by the Ranking Algorithms, BW, CL, DE, T_CL, and T_DE. ....</b>	<b>105</b>
<b>Table 7-2: Top 10 Actors in the URI Working Group Ranked by the Ranking Algorithms, W_PR, PR_BW, T_BW, PR_TBW and TE_BW. ....</b>	<b>106</b>
<b>Table 7-3: The Spearman Correlations among the Nine Ranking Algorithms except TE_BW.....</b>	<b>107</b>

## LIST of FIGURES

<b>Figure 1-1: Architecture of Storylines. ....</b>	<b>8</b>
<b>Figure 1-2: The Themes and Named Entities Projected to the Fourth LSI Latent Concept Dimension. ....</b>	<b>9</b>
<b>Figure 1-3: The Term Signatures for “Mad Cow Disease” and “Boynton Laboratory”.....</b>	<b>10</b>
<b>Figure 1-4: The Concept Map of the Story “Mad Cow Disease” and the Story “Boynton Lab”. ....</b>	<b>11</b>
<b>Figure 3-1: Architecture of the LSI Subspace Signature Model. ....</b>	<b>29</b>
<b>Figure 3-2: The Term Signature of the Key Topical Theme: “Taliban” (the Upper Image) and the Document Signature of a News Article (the Lower Image). ....</b>	<b>33</b>
<b>Figure 4-1: Document Clustering Based on Top Ranking Terms versus Top Ranking Term Clusters. ....</b>	<b>46</b>
<b>Figure 4-2: The Architecture of Two-stage K-means.....</b>	<b>48</b>
<b>Figure 4-3: The Variation of the Evaluation Matrix of the Basic K-means with <math>T_g</math> Using the Reuters Corpus and CDSS.....</b>	<b>66</b>
<b>Figure 5-1: Comparison of the Learning Curves between the Subset RL5080 and the Full Training Set R100 Using the KNN Classifier.....</b>	<b>78</b>

<b>Figure 5-2: Comparison of the Learning Curves between the Subset RL5080 and the Full Training Set R100 Using the Naïve Bayes Classifier.....</b>	<b>79</b>
<b>Figure 5-3: Comparison of the Learning Curves between the Subset RL5080 and the Full Training Set R100 Using the Rocchio Classifier.....</b>	<b>79</b>
<b>Figure 5-4: Comparison of the Learning Curves between the Subset T150 and the Full Training Set T1100 Using the KNN Classifier.....</b>	<b>81</b>
<b>Figure 5-5: Comparison of the Learning Curves between the Subset T230 and the Full Training Set T2100 Using the Rocchio Classifier.....</b>	<b>81</b>
<b>Figure 6-1: Query Expansion Procedures of Local Analysis.....</b>	<b>87</b>
<b>Figure 6-2: Query Expansion Procedures of Global Analysis and Term Re-weighting Strategies.....</b>	<b>88</b>
<b>Figure 6-3: The Interface Allows Students to Make Relevance Ratings about the IPTC Categories for the OCR Text of the Historical Newspaper.....</b>	<b>97</b>
<b>Figure 8-1: A Schematic Framework which Integrates NLP, IE, LSISSM, TM, IR, SNA and User Interfaces.....</b>	<b>112</b>

## ABSTRACT

### **Text Clustering and Active Learning Using a LSI Subspace Signature Model and Query Expansion**

Weizhong Zhu  
Robert B. Allen, Ph.D. Supervisor

In this dissertation research, we developed a novel Latent Semantic Indexing Subspace Signature Model (LSISSM) for semantic content representation of unstructured text based on the Singular Value Decomposition (SVD). The model represents the meanings of the terms according to the distribution of their statistical contribution across the top ranking LSI latent concept dimensions. Each LSI latent concept dimension is related to one or more themes with their contexts which are composed of semantically coherent topics, entities and social indicators and are supported by a set of related documents. The model provides feature reduction and finds a low-rank approximation for the scalable and sparse term-document matrix. Firstly, the top ranking conceptual terms or term clusters are selected to represent the corpora according to their global statistical contribution to the LSI term subspace. Secondly, terms and documents are defined as spectral signatures which are represented by the distribution of their local statistical contribution on the identical LSI latent concept dimensions. Then, two novel similarity measures are defined to evaluate the associations between the concept signatures and the document signatures by reducing noise. Finally the model bridges the LSI subspaces naturally and produces a low-dimension term-document matrix.

Experiments suggest that this model significantly improves the performance of the clustering algorithms such as basic K-means and Self-organized Mapping (SOM) efficiently and effectively, compared with the Vector Space Model and the traditional LSI model. Our model is also suitable for active learning which significantly decreases the number of the training examples through bootstrapped sampling and iterative learning without degrading the performance of the classifiers. The LSI Subspace Signature Model selects the document samples iteratively according to their statistical contribution to the LSI document subspace. The sampling method is evaluated in the context of text categorization with three classic classifiers on three standard news corpora. The results indicate that our approach improves the selection of the sampling subsets from the perspectives of sampling distribution and learning performance of the classifiers. This method picks the most important samples and keeps the sampling distribution on the text categories, even outlier categories. The tests demonstrated that the sample subsets with the optimized feature sets substantially improve the performance of the three classifiers, Naïve Bayes, K-Nearest Neighbor and Rocchio effectively and efficiently.

Four types of ontology-based query expansion strategies are applied in MEDLINE abstracts and OCR news text. UMLS-based term re-weighting and user relevance feedback with IPTC hyponyms significantly improve the performance of IR models, VSM and BM25. In addition, a novel random-walk centrality measure is developed to overcome the rank sink problem of the PageRank algorithm.



## **CHAPTER1: Introduction**

Knowledge structure extraction, exploration and discovery are essential for information science. The tasks are difficult because the domains of the knowledge subjects are diverse and the relationships among the subjects are very complex. Knowledge structures can be thought of as the semantic networks in different domains which represent the relatedness among knowledge subjects like key concepts, indicators, documents, image and so on. For instance, the Unified Medical Language System (UMLS) collects millions of bio-medical concepts and the International Press and Telecommunications Council (IPTC) codes cover thousands of metadata tags for daily news. The indicators include social indicators like person name, attributes like geo-spatial locations and temporal factors, and verbs for actions and the types of the semantic relationships. A complete ontology which organizes the relationships between knowledge objects for a domain is rare and often lag compared to the development of a field. Systematically and dynamically extracting structured information from evolving unstructured text like news and email is very difficult. In this research, we design and develop components for automatic knowledge structure extraction, exploration and discovery which require the integrated processes of natural language processing (NLP), information extraction (IE), data models for content semantic representation, information retrieval (IR), text mining (TM) and social network analysis (SNA). These processes fit into the three basic schemes under which knowledge may be organized: (1) Declarative knowledge is about what the knowledge objects are and when, where, how and why they work the way together.

Extracting stories from newspaper articles is a good example. Most news articles deal with events. In general, events in news are contemporary happenings of significance. The events are defined as structured objects with many properties and the structures are evolved with spatiotemporal and causal conditions. These objects, properties and conditions are interpreted as questions like What, Who, Where, When and Why. A story is a sequence of related events. In processing and understanding large scale of news text, it is helpful to automatic identify components of events such as What, Who, Where, When and to construct semantic relationships among them for answering "Why". Zhu et al. (2007) designed a framework to target these types of tasks. (2) Procedural knowledge details steps or activities required to perform a task automatically. For instance, document clustering and text categorization techniques define the procedures to automatically group and label the news articles. Tasks like Topic Detection and Tracking (TDT, NIST, 1998) target such automatic online text categorization problems. A topic in TDT is defined to be a seminal event or activity, along with all directly related events and activities. The notion of topic in TDT is similar to the definition of Story, but it usually refers to human focus and interests on certain domains. (3) To match the requirements of the tasks, it is necessary to create strategies and setting conditions for different procedures.

Knowledge structure extraction is the automatic extraction of knowledge units from unstructured text, such as news articles, MEDLINE abstracts, email, OCR text and citation abstracts. POS tagging, stemming and stop-word filtering are standard NLP

techniques. LingPipe<sup>1</sup> and Gate<sup>2</sup> are the IE tools applied for the named entity extraction. The fusion of the two tools might produce more accurate results. Presenting the semantic meaning of the knowledge subjects are the essential tasks. In general there are two ways for the semantic representation, statistical data models and interpretation from domain ontology. The semantic association between knowledge subjects could be hierarchical or associative within certain contexts. Domain ontology and the statistical data models are useful to catch these semantic relationships. This thesis designs and develops a new statistical data model and tests its effect on the automatic text categorization. It also utilizes the domain ontology in medicine (UMLS) and news (IPTC code) to expand user queries and improve the performance of the IR search engines.

---

<sup>1</sup> <http://alias-i.com/lingpipe/>

<sup>2</sup> <http://gate.ac.uk/>

## 1.1 Latent Semantic Indexing

The key issues of semantic content representation models for unstructured text are term representation and document representation, dimensionality reduction and similarity measures. The Vector Space Model (VSM, Salton et al., 1975) uses TF (Term Frequency) or TFIDF (Term Frequency times the Inverse of Document Frequency) to present terms and documents, which cannot capture the semantic relationships between terms and documents effectively and cannot reduce the dimensions. Recent research shows that data models for semantic content presentation (Griffiths et al., 2007) include two major types, probabilistic topic models (Beil et al., 2003; Hofmann, 1999) derived from the statistical language model (Ponte and Croft, 1998) and Latent Semantic Indexing (LSI) (Deerwester et al., 1990) based on the linear algebra technique Singular Value Decomposition (SVD). However, traditional LSI has several problems which are related to information loss, noise reduction and the senses of the latent concept dimensions.

Information loss occurs when the number of the LSI concept dimensions selected for the low-rank approximate subspaces is small. Empirically, the number of the latent concept dimensions selected is usually less than 400 even when the corpus is large enough. Can a small number of latent concept dimensions sufficiently represent the whole latent space? Ding (2005) developed a probabilistic model which provides insight for LSI based on the dual relationship between words and documents. The model has established the amount of contributions of dimensions to the latent semantic space. Specifically, the singular value squared is the amount of

contributions of the corresponding dimension. This quadratic dependence indicates that LSI dimensions with small singular values are overrepresented by the linear relationship as previously thought. Furthermore, they demonstrated that the importance of LSI dimensions follows the Zipf's distribution, which explains why a small number of the most important dimensions can adequately approximate the overall semantic space. Derived from the conclusion that the singular value squared is proportional to the statistical contribution of the latent concept dimensions, the normalized contribution of the approximation subspace to the overall semantic space is the ratio between the sum of the singular value squared to the subspace to the sum of the singular value squared to the whole space. The experiments in Section 3.2 show that even if selecting a subspace which makes a high percent contribution, the number of the latent dimensions is petty large. For instance, we test a corpus with 2527 Reuters news articles and the results indicate that 917 top latent concept dimensions make an 80 percent contribution. This indicates that a few hundred latent concept dimensions which are usually used in LSI may not cover all the important topics which dominate the performance of automatic text categorization techniques.

LSI produces noise because it takes account of co-occurrence and second-order co-occurrence (Kontostathis and Pottenger, 2006). Second-order co-occurrence means that although two concepts do not appear in one same document and their association will be concerned if there are hidden links between them. For instance, if Concept A and Concept B co-occur in a few of documents, and Concept C co-occurs with the Concept B in a few of documents, the association between Concept A and the

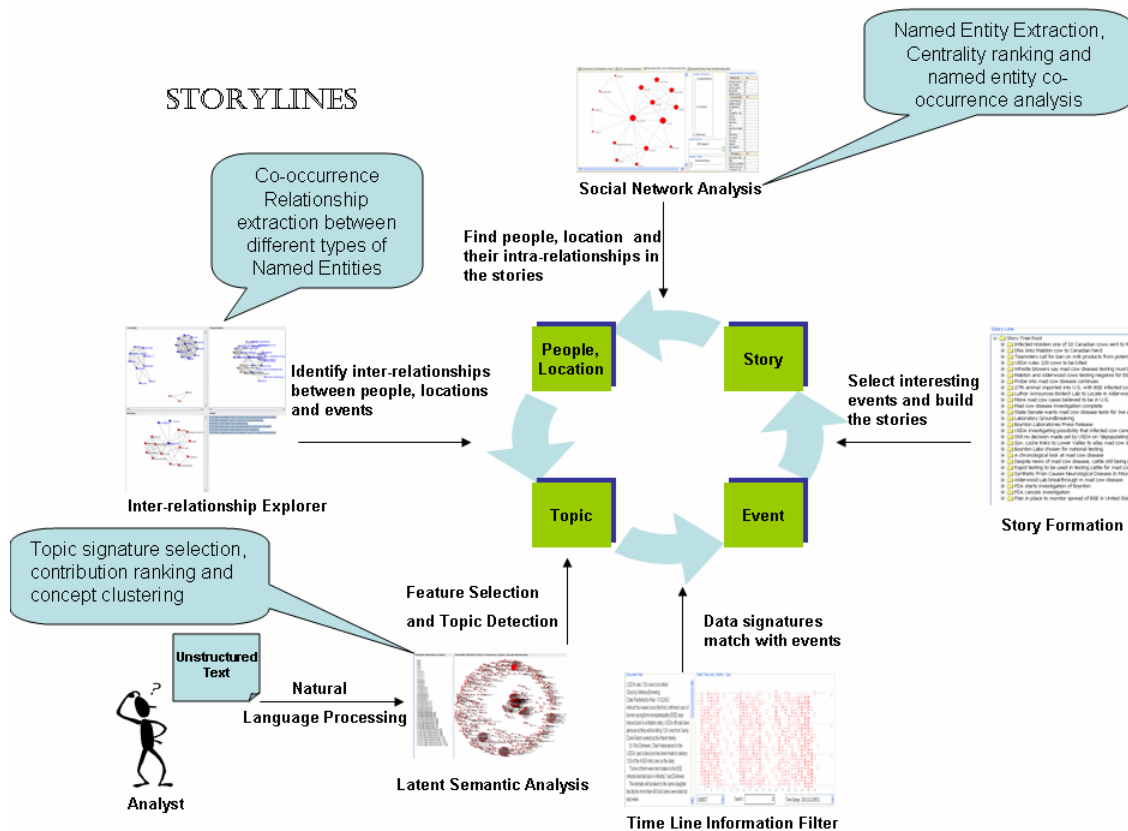
Concept C will be accounted by the LSI processing even though the two concepts does not co-occur in any document. This type of co-occurrence might be useful for knowledge discovery but in most of the cases it is noise and reducing that noise is an open question.

Another issue is whether LSI latent concept dimensions reflect the topics directly? By observation of the projection of terms in the LSI term subspace, the answer is that a latent concept dimensions does not often correlate to one topic explicitly. Almost every term has non-zero projection on each of the latent concept dimensions. A group of terms that has higher projection scores in the same dimensions with the same sign might be related to one topic and its context. But, two different topics often map to the same latent concept dimension with different signs of projection scores. In another word, each latent concept dimension can be thought of a mixture of different topics. How to use these latent concept dimensions to present terms and documents is another question.

Targeting these two open difficulties for LSI, we propose a novel semantic representation model which is motivated by the LSI probabilistic model (Ding, 2005) and the visual patterns of the concept mapping and the distribution of term contribution and document contribution on the latent concept dimensions explored by Storylines (Zhu and Chen, 2007). In this model, each term or each document is defined by an LSI subspace signature which represents the distribution of its local statistical contribution on the top ranking LSI concept dimensions. Then two novel similarity measures which bridge the LSI term subspace and the LSI document

subspace are constructed between the term signatures and the document signatures. Finally, this model transforms the initial term-document matrix into a low-dimension approximate term-document matrix using the latent concept dimensions as intermediate layers and boost-strap term ranking to control information loss. The model is applied to two types of text mining applications to demonstrate its efficiency and effectiveness, text clustering and active learning. The model should also be able to be applied to text retrieval and social network construction, but those are not addressed in this thesis.

## 1.2 Visual Exploration of the LSI Subspaces

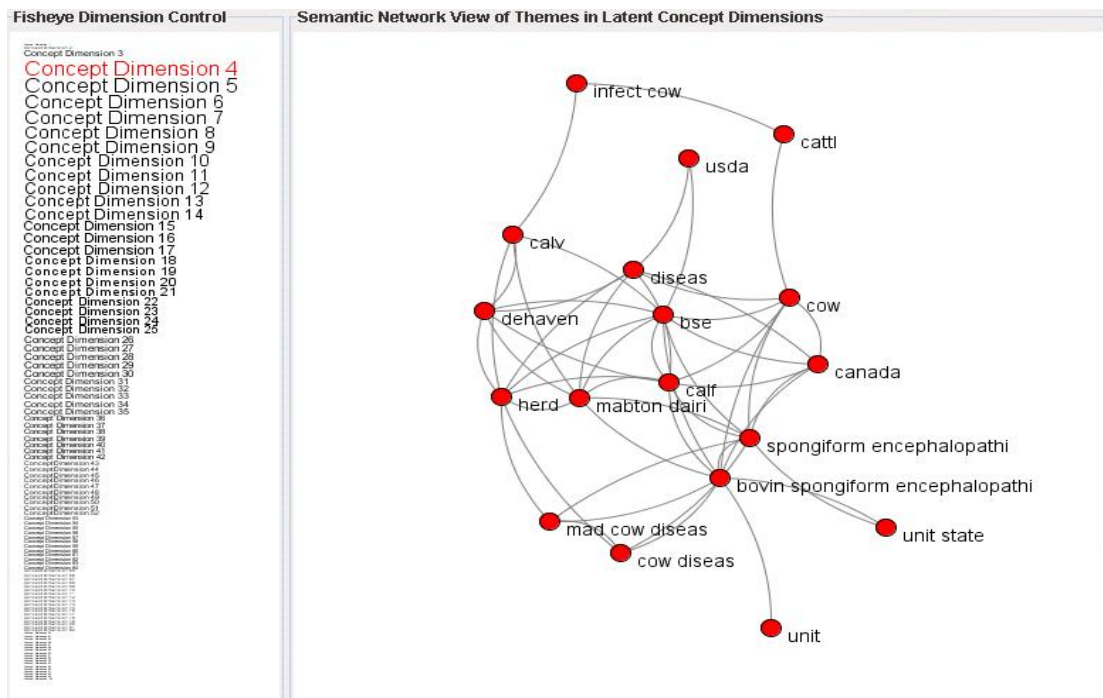


**Figure 1-1: Architecture of Storylines.**

We developed a prototype system for visual analytics (Chen, 2008; Thomas and Cook, 2005) based on the LSI Subspace Signature Model which we called Storylines (Zhu and Chen, 2007). The goal of the system is to show how well the system can capture news event threads to form stories. As Fig. 1-1 shows, Storylines enables analysts visually and systematically explore the concept maps generated by the model and study unstructured text without prior knowledge of its thematic structure. The system integrates the LSI Subspace Signature Model, natural language processing,

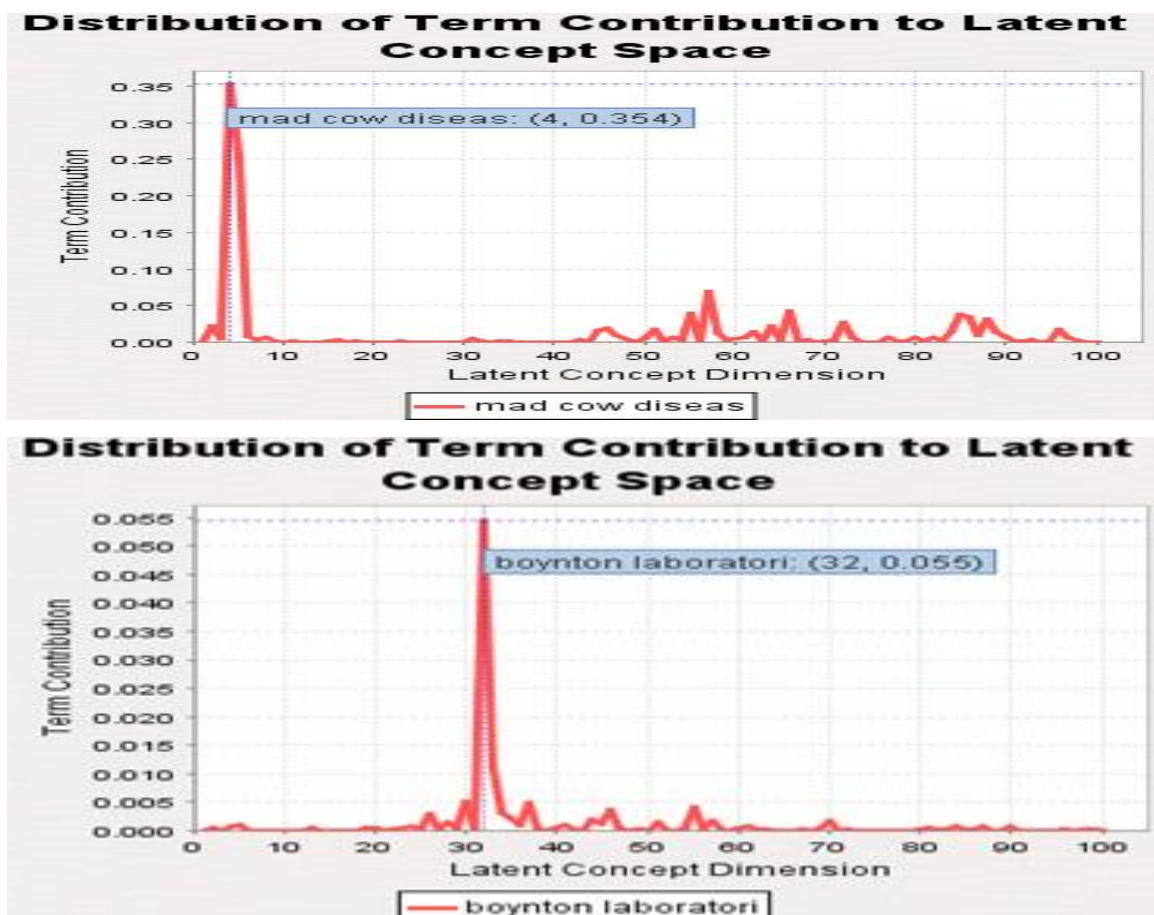
information retrieval and social network analysis. The contributions of the work include providing an intuitive and directly accessible representation of a latent semantic space using concept mapping, an integrated process for identifying salient lines of stories, and coordinated visualizations across a spectrum of perspectives in terms of people, locations, and events involved in each story line. The system was tested with the portion of news articles in the 2006 VAST contest data, and successfully identifies the topics and key players in the plots of the tasks.

By observation on the system, the LSI latent concept dimensions reflect the semantic relationships in different contexts which are represented by semantically coherent topics and indicators, see Fig.1-2.



**Figure 1-2: The Themes and Named Entities Projected to the Fourth LSI Latent Concept Dimension.**

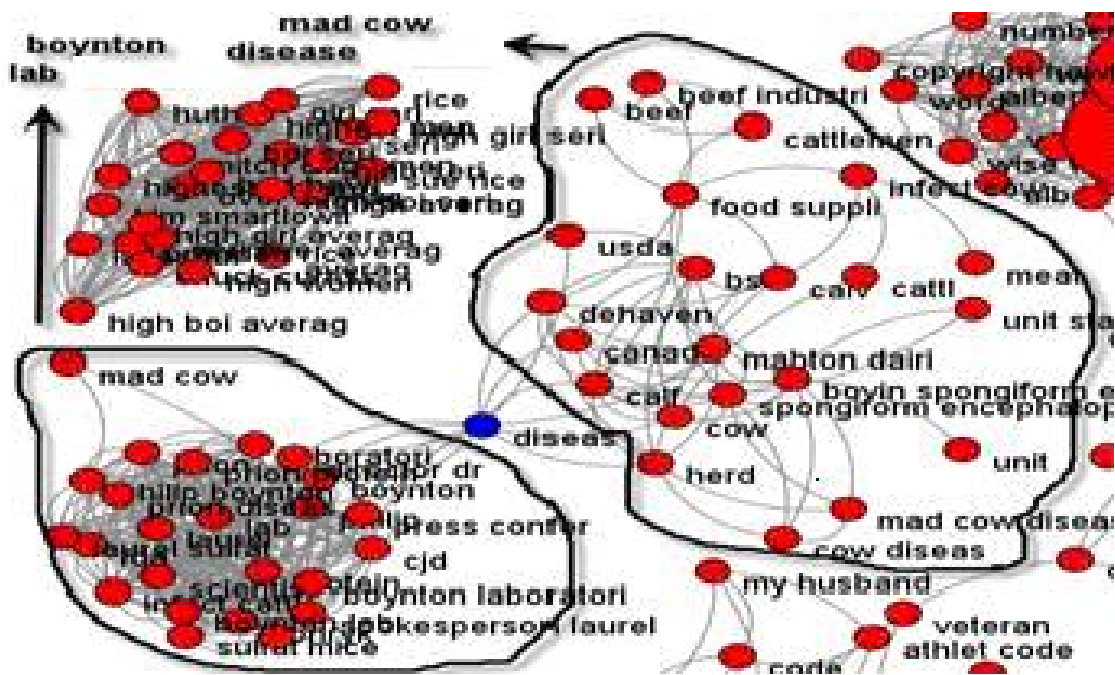
The term cluster in the Fig 1-2 includes the themes (e.g., “Mad Cow Disease”), location name (e.g., “Canada”) and organization names (e.g., “USDA”). The signature representation is helpful to differentiate and address the senses of the terms and identify the distinct topics, see Fig. 1-3.



**Figure 1-3: The Term Signatures for “Mad Cow Disease” and “Boynton Laboratory”.**

The popular topic terms link different themes and contexts together in the concept map, see Fig.1-4. The formula to calculate the signatures is introduced in Section 3.1.

And, these popular terms have specific meanings in different contexts. For example, in Fig.1-3, the term “disease” links the two term clusters which represented the story “Mad Cow Disease” and the story “Boynton Lab” respectively. In the two stories the specific senses of the term “disease” indicate different diseases.



**Figure 1-4: The Concept Map of the Story “Mad Cow Disease” and the Story “Boynton Lab”.**

However, Storylines focuses on the interface design and has limited functions for knowledge searching and categorization. To gain better performance, we apply the LSI Subspace Signature Model with text clustering algorithms and design a novel scenario for active learning. Several query expansion strategies are applied to improve the accuracy of the IR engines. Storylines also suggests that Social Network

Analysis (SNA) is a key component for knowledge discovery. The construction of the social network is based on the either content or context. In content-based social network, the nodes in a social network are named entities and the links between them are weighted by co-occurrence frequency in the same document/sentence. The approach is used in Storylines. In context-based social network, the level of communication between actors, for instance send-reply frequency of email conversations, could be used as the association weight between nodes. Identification of the influential players in the social networks is a big challenge for SNA. Thus, we developed a new random-walk centrality measure to detect the key players in the social networks in email conversations.

### 1.3 Research Questions

The goals of the thesis are to develop key components for the semantic representations of the unstructured text from the perspectives of the knowledge structures of how documents are related to topics, people, organization, location and time. These components are required to identify key textual features in documents, construct networks for these features and provide general model of the semantic relationships between them. Based on these goals the key research question is:

**How to represent semantic information of concepts and documents in unstructured text?**

Our LSI Subspace Signature Model addresses the question and divides it into three sub-questions from the perspectives of semantic content representation, dimension reduction and similarity.

**(Q1) How to represent the concepts and the documents with unified signatures in an unsupervised manner?**

**(Q2) How to reduce noise and transform the high scalable term-document matrix into a low-dimension approximation term-document matrix?**

**(Q3) How to compute the similarity between the signatures?**

For text categorization, we verify the effectiveness and efficiency of the model. Text categorization techniques need to emphasize the statistical identification, generation and connection of feature subspaces, active learning mechanisms, information self-organization methods and so on. Automatic clustering – unsupervised learning and

active learning – machine learning integrated with the LSI Subspace Signature Model are the major techniques to identify and represent the categories of the documents with the key concepts. For text categorization, the questions are:

**(Q4) How to integrate the LSI Subspace Signature Model with the traditional text categorization techniques?**

Specifically, which sample article should be picked first for training if there is no class label for the corpus? How does the model solve the initialization problem of the basic K-means clustering algorithm?

**(Q5) Does the model improve the performance of these techniques?**

**(Q6) Does the model outperform other models?**

Specifically: Compared to VSM, does our LSI Subspace Signature Model significantly enhance the performance of the basic clustering algorithms? Accordingly, query expansion using the domain ontology and SNA has the similar questions.

**(Q7) Do the re-weighted terms extended by UMLS co-concepts and synonyms improve the performance of VSM and Okapi BM25?**

**(Q8) Do user relevance feedback and hyponyms extracted from International Press Telecommunication Council (IPTC) codes enhance the VSM?**

**(Q9) How does the novel random walk centrality measure solve the rank sink problem of PageRank (Page et al., 1998)?**

## 1.4 Organization of the Thesis

Chapter 2 reviews the related work including semantic content representation models, text clustering, active learning, query expansion approaches to text retrieval, and centrality measures for SNA.

In Chapter 3, we propose the LSI Subspace Signature Model that incorporates explicit statistical meanings to the semantic representation of concepts, indicators and documents. This chapter first introduces the notion of subspace signatures and the general mechanism to calculate the association between signatures. Two types of similarity schemas are described. After that, we present a step-wise ranking algorithm to reduce the noise of the LSI subspaces and pick the concepts which make the major contributions to the LSI term subspace. The algorithm is also used for sampling selection. Then the LSI Subspace Signature Model transforms a high scalable term-document matrix weighted by TFIDF into a low rank term-document matrix weighted by the two signature-based similarity measures. Finally, the model is applied to concept mapping, concept ranking and evolution, text clustering and active learning on news articles, emails and ISI records.

Chapter 3 answers Research Questions 1-3 for the LSI Subspace Signature Model. In Chapters 4 and 5 we describe the TM applications of the LSI Subspace Signature Model. They are text clustering and active learning, respectively. In these applications, we evaluate the answers for the Research Questions 4, 5, and 6. Chapter 6 present two applications of query expansion strategies and answers Research Questions 7 and 8. Chapter 7 addresses Question 9 for SNA.

## CHAPTER2: Literature Survey

### 2.1 Data Models for Semantic Content Representation

To differentiate the ambiguity of word senses and understand unstructured text, the data models for semantic content representation need to simulate the semantic associations between knowledge subjects. Each subject is represented as a unified signature which is constructed as a weighted vector and has a statistical meaning. The signatures could be used to predict and compare related knowledge objects, differentiate them from others and disambiguate themselves. The similar signatures group the knowledge objects with closer associations and reflect the semantic structures which could be applied for IR, TM and SNA applications and could be extended by domain ontologies. The models include three major types, probabilistic topic models, latent semantic indexing (LSI) and connectionist models.

Hoffman (1999) proposed a probabilistic topic model in his probabilistic latent semantic indexing (PLSI) model which is based on the notion that a latent topic can be represented as a distribution over terms and a document is a mixture of the latent topics. The model specifies the relationships between a term  $t$  and the latent topic set  $Z$  with  $T$  topics within a document:

$$P(t_i) = \sum_{j=1}^T P(t_i | z_i = j) P(z_i = j) \dots\dots\dots(2.1)$$

However, the model is not generative, which makes it difficult to predict a new document. Blei et al. (2003) extended the model and smoothed the topic distribution

by placing a Dirichlet prior to it, so called Latent Dirichlet Allocation (LDA) generative model. The probability density of a  $T$  dimensional Dirichlet distribution over multinomial distribution  $P = (p_1, \dots, p_T)$  is defined by:

$$Dir(\alpha_1, \dots, \alpha_T) = \frac{\Gamma(\sum_j \alpha_j)}{\prod_j \Gamma(\alpha_j)} \prod_{j=1}^T p_j^{\alpha_j - 1} \dots\dots\dots(2.2)$$

In the Eq. 2.2,  $\alpha_j$  is the prior observation count that topic  $j$  is sampled in a document before observing any terms from the document. Using Expectation-Maximization (EM) or Gibbs Sampling, the probabilistic topic models can directly extract topics those are represented by the most associated terms (Griffiths et al., 2007). The effectiveness of topic models relies on two factors, the estimate of the term distribution – how many topics in the corpus and the smoothing which is dependent on the quality of the training sets.

LSI (Deerwester et al., 1990) is the truncated SVD which decomposes the term-document matrix into a term-concept matrix, a diagonal concept-concept matrix and document-concept matrix:

$$C = UDV^T \dots\dots\dots(2.3)$$

Usually each term is represented by a vector which includes the top  $k$  dimensions in the term-concept matrix and each document is represented by a vector which contains the top  $k$  dimensions in the document-concept matrix. Using matrix factorization, PLSI and LDA can also be split into two matrixes, term-topic matrix  $\Phi$  and

document-topic matrix  $\Theta$ . The feature values in PLSI and LDA are non-negative and sum up to 1. In LSI, the projection scores in the subspaces are least square distance and can be positive or negative, which indicates these values do not have direct statistical meanings.

A method called non-negative matrix factorization (NMF) (Xu et al., 2003; Xu et al., 2004) catches the topics with the positive projection to the latent concept dimensions and linearly combines these topics to represent the documents. Cai et al. (2005) propose an unsupervised Linear Discriminant Analysis method which is called Local Preserving Indexing (LPI), which projects the documents to a lower dimension space and grabs both geometric and discriminating structures of the documents. However, with this approach we still need to determine how many dimensions to use in the subspace.

The signatures in our model are non-negative and have clear statistical meanings. Assuming that the representative features and the discriminating features make the major statistical contribution to the LSI term subspace, our LSI Subspace Signature Model has control on the upper boundary of the number of latent concept dimensions selected and transfers the dimension reduction to a feature reduction problem by the step-wise term-picking algorithm (Zhu and Chen, 2007). A small number of the top terms ranked by their global contribution to the overall term subspace represent the corpus. The size of the optimized feature subset is controlled by the accumulated global contribution of the terms. Our model does not require any training set and

smoothing because the model can calculate the similarity between any pair of concept signature and document signature.

Besides these probabilistic models, many approaches are based on the connectionist model (Rumelhart, McClelland and the PDP Research Group, 1986) represent the semantic relationships between concepts and documents using artificial neural network (ANN). Based on the connectionist model, information is more or less uniformly and dynamically distributed between all of the components in the system. The Rumelhart feed-forward network (Rumelhart, 1990) proposes a three-layer architecture to represent the semantic relationships between concepts and their attributes. The input layer includes the concepts and relations and the output layer contains the attributes of the concepts. The hidden layer between the input layer and the output layer is composed of the neurons which adapt the Back-Propagation (BP) supervised learning algorithm (Werbos, 1994) and learn the patterns of the semantic associations of concepts and attributes within a series of contexts. The BP algorithm use non-linear sigmoid activation functions to estimate the weights of the inputs and presents the outputs as the weighted-sum of the inputs with a minimized error. The approach has been applied to multiple-label text categorization (Zhang and Zhou, 2006) which achieves the comparable performance to that of Support Vector Machine (SVM, Elisseff and Weston, 2002).

For unsupervised learning, the Self-Organized Mapping (SOM, Kohonen, 1990) does not include the hidden layer and adjusts the weights of the inputs so that the similar inputs cause similar outputs. The experiments in Section 4.3.3 indicate that integrated

with the LSI Subspace Signature Model the performance of the SOM is significantly improved.

## 2.2 Text Clustering

Document clustering faces several key challenges: feature selection, feature and document representation, similarity schemas, the number of the clusters, clustering accuracy and efficiency, description of the clusters, scalability and so on. A text corpus has very rich features which make noise and affect the performance of the clustering algorithms.

Clustering methods fit into two categories, hierarchical clustering and non-hierarchical clustering. Non-hierarchical clustering, in general, partitions a set of objects into a set of non-overlapping groups so as to maximize the within-cluster inter-object similarities and minimized the between-cluster similarities due to some heuristic criterion of ‘goodness of clustering’. Currently, the best K-means partition algorithm is called bisecting-k-means. It divides the corpus into two clusters with K-means first. Then, it keeps partitioning the currently largest cluster into two clusters, again using K-means, until k clusters have been discovered (Steinbach et al. 2000). However, these K-means document clustering algorithms require a predefined number of the clusters. Obviously, the size of the document clusters is difficult to predict. Self-Organizing Maps (SOM) (Kohonen, 1990; Lin et al., 1991) automatically decide the number of the regions or the clusters. Willett (1990) introduces the trend in the hierarchical agglomerative clustering, such as complete linkage and single linkage.

The approaches mentioned above emphasize on the discovery of the boundaries of the clusters but do not pay attention to the interpretations of the clusters. Wang et al.

(1999) introduce an association rule based non-hierarchical clustering algorithm, which compare the similarity among the frequent feature sets occur in the documents instead of the pair-wise similarity between documents. Beil et al. (2002) extend the idea to create a hierarchical clustering method called HFTC which greedily picks frequent feature sets to minimize the overlapping among the documents. Fung et al. (2003) propose the FIHC algorithm which applies terms in the global frequent feature sets to reduce dimensionality and significantly improve the clustering accuracy. These methods provide interpretation to the document clusters with frequently used terms but have weakness in explaining the explicit relationships between concepts and documents. Schütze and Silverstein (1997) developed an approach to use the LSI projections to the LSI document space and improved the efficiency of the document clustering. Each document is represented as a vector of the least-square distance from the latent concept dimensions to the original documents with a fixed number of the top ranking dimensions. Ampazis and Perantonis (2004) introduce a novel method called LSISOM which applies the LSI term subspace as input to improve the SOM document clustering.

### 2.3 Active Learning

The performance of the machine learning methods on text categorization is generally determined by three aspects: (1) classifiers embedded with classification algorithms, (2) optimized feature sets and (3) optimized training samples. Active learning differs from "learning from examples" in that the learning algorithm assumes at least some control over what part of the input domain it receives information about (Cohn et al., 1994). The traditional machine learning approaches require large amount of training examples. However the creation of the training examples is very expensive, which motivates the active learning methods to significantly decrease the size of the training examples through bootstrapped sampling and iterative learning without degrading the performance of the classifiers.

Lewis et al. (1994) introduce a method to decrease the size of the training set and the error of the classifier C4.5. McCallum et al. (1998) integrate a Query-By-Committee learning approach with an EM classifier which reduces more than one third of the training samples without degrading the performance of the classifier. Tong and Koller (2001) developed a SVM-based active learning approach significantly decrease the size of the training examples through iterative sampling. However, most of these studies did not concern to automatically keep the distribution of text categories in the sample subsets and these methods are originally designed for each single category, which supposes that the distribution is known or predefined.

Compared to pool-based methods like Tong's which study each category at one time and reported the learning accuracy for each of the top categories rather than overall,

our method studies all the categories at once, which does not necessarily require prior knowledge of the class labels. Second, the ranking algorithm in our model identifies all the categories even outliers with a fairly large scope of the parameters. And Tong's method only studies the most popular categories which include many samples in the training corpus. We intend to use the three corpora which contain many outlier categories which have sample size that is less than 5. Compared to the random sampling, the random sampling can't consistently include all these categories in the sample subsets when the size of the subsets is smaller than 50 percent of the full sample set.

## 2.4 Query Expansion

Query Expansion (QE) is the process of reformulating a seed query to improve retrieval performance in information retrieval operations (Vectomova and Wang, 2006). Many query expansion methods have been proposed to improve the precision and/or recall of Information Retrieval. There are two general strategies for query expansion. One is ontology-based; the other is statistical. Previous studies showed that expanding a query with synonyms or hyponyms has a limited effect on biomedical information retrieval performance (Guo et al., 2004; Hersh et al., 1995 and 2000; Leroy et al., 2001). There are several reasons for that: ontological methods use no notion of weight; they do not consider actual documents but use only prior knowledge. Some authors (Cesarano et al., 2003; Richardson et al., 1995; Zhu et al., 1999) explored different weighting methods but did not report how the weights were used in conjunction with the ontology.

Statistical methods focus on documents and they can be further divided in two main sub-categories: global analysis and local analysis (Xu and Croft, 1996). Global analysis considers the whole collection of documents to extract the co-occurrence of related terms. Global analysis methods include term clustering, latent semantic indexing, and similarity thesauri. One of the major drawbacks of global analysis methods is that the methods require semantic similarity and disambiguation of terms.

Local analysis extracts highly-related terms from the relevant documents retrieved by an initial query or from data mining results. Xu and Croft (2000) introduced Local Context Analysis which uses the top documents returned by an initial query but

selects the terms based on co-occurrence with query terms. This approach, because it usually requires less human intervention, assumes that the certain numbers of top documents returned by the initial query are actually relevant (“pseudo-relevance feedback”). However, these methods are not robust because it is almost impossible for all search engines or mining methods to return only relevant documents. So studies on a hybrid of global analysis and local analysis may be more promising.

## 2.5 Social Network Analysis

Social Network Analysis (SNA) investigates the interactions among people, organizations or communities. Two factors are essential for understanding the social status of an actor --- popularity and prestige. Popularity can be measured by the quantity of endorsements the actor receives from other actors, whereas the prestige is shown by the quality of the received endorsements, for example, the prestige of endorsing actors (Bollen et al., 2006). The quality of scholarly communication is often assessed in terms of the number of citations it has received. We extend this notion to the study of the influence of an individual in a network of email communication. A common criticism of social network research is that the study of prestige has not directly addressed the dynamic information flow in such networks (Friedkin, 1991; Page et al.1998). We develop a similarity measure which incorporates time and simulates the speed and frequency of email conversations between nodes. This measure is particularly useful for discovering long-term active experts and contemporary experts.

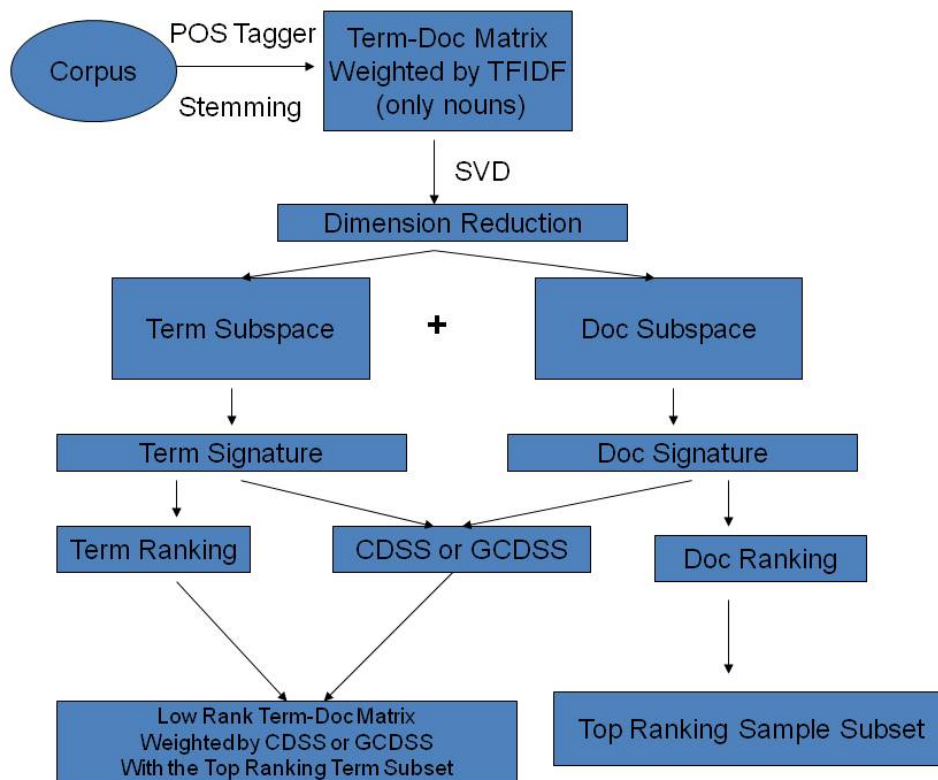
The Degree and Closeness (Sabidussi, 1966) centrality are generally accepted as indicators of influential status, and are based on the number of neighbors for a node in a network and the distances between nodes. However, they primarily indicate the popularity rather than prestige. A potential measure of prestige is Betweenness centrality (Freeman, 1979 and 1997), which is based on the critical members in the shortest paths between any pair of nodes in a network. Another possible measure of prestige is the PageRank algorithm (Brin and Page, 1998; Page et al., 1998), which

computes the influence of a web page based on a combination of the number of hyperlinks that point to the page and the influence of the pages that the hyperlinks originate from. PageRank, restricted to random walks, is essentially a special case of eigenvector centrality. All the four measures, Degree Centrality, Closeness, Betweenness Centrality, and PageRank, assume that influence propagates via restricted paths.

Bonacich (1972) and Borgatti (2005) address the problem of discovering key players by explicitly measuring the contribution of a set of actors to the cohesion of a network with two analytical functions. White and Smyth (2003) define the most important nodes in the network by considering the referral links like PageRank (Page et al., 1998) and HITS (Kleinberg, 1999). A linear model (Faloutsos et al., 2004) is produces sub-graphs on the basis of electrical circuit formula. Huberman and Wu (2004) use the same approach and exploit Kirchhoff's Laws to model a social network. Other approaches such as (Newman, 2004) use Betweenness to find crucial central nodes. Pujol et al. (2002) proposed a PageRank style ranking algorithm that uses the out-degree that could be thought as a slight variant of absolute out-degree centrality to weigh the random jumping probability. In this study, we extend weighted Page Rank algorithm to social network analysis which follows the same traversing mechanism. Our novel method propagates weighted Page Rank with Betweenness to solve the "tank sink" problem of random walks. Our study aims to address the extent one can identify the influential status of group members based on the structure of their email communications.

### CHAPTER3: The LSI Subspace Signature Model

We propose a novel LSI Subspace Signature Model (LSISSM) for semantic content representation. The architecture of LSISSM shown in Fig. 3-1 depicts the three key components: dimension reduction, term/document ranking and similarity between signatures.



**Figure 3-1: Architecture of the LSI Subspace Signature Model.**

The model provides spectral signature representation for terms and documents and ranks them according to their contribution and builds relatedness between the signatures by their matching patterns on the identical LSI latent concept dimensions.

Compared to the traditional LSI (Deerwester et al., 1990), our method has several advantages: the number of the top ranking dimensions are picked based on their global statistical contribution to the LSI subspaces instead of a predefined fixed number; we use a optimized feature set to present the documents instead of a number of the LSI latent concept dimensions because the meaning of the latent concept dimensions is implicit; the similarity schemas count on the association between term signatures and document signatures and reduce the noise in the associations by bridging the LSI term subspace and the LSI document subspace. The output of the model is a low-dimension term-document matrix weighted by the similarity between the term signatures and the document signatures.

### 3.1 LSI Subspace Term Signatures and Document Signatures

Our model is motivated by the LSI probabilistic subspace model (Ding, 2005), which points out that the singular value squared is proportional to the statistical contribution of the corresponding LSI latent concept dimensions. The importance of these LSI dimensions follows the Zipf's distribution.

Based on this conclusion, we built feature and document representations according to their local statistical contribution to the LSI latent concept dimensions. This is calculated by the squared product of the singular value of the latent concept dimension and the least-square distance from the dimension to the original term or the original document. The values of the signatures in our data model are non-negative and have explicit statistical meaning, statistical contribution to the latent concept dimensions. The terms and documents are represented separately by the projections to the top  $K$  dimensions in the identical LSI latent concept dimensions.

This following section describes how the signatures are generated:

**Step 1:** Use Singular Value Decomposition (SVD) to decompose the term-by-document matrix  $A$  into a term-concept matrix  $U$ , a diagonal matrix  $D$  and a document-concept matrix  $V$ .

**Step 2:** In the matrix  $U$  or  $V$ , the top  $K$  dimensions are selected as the proximity of the overall document space. In this study, the value of  $K$  is determined by the ratio:

$$\sum_{j=1}^K D_j / \sum_{n=1}^M D_n \dots\dots\dots(3.1)$$

In the Eq. 3.1,  $D$  is the square of the singular value  $S$  and  $M$  is the total number of the latent concept dimensions with a non-zero contribution. We select a threshold  $T_d$  for the ratio in Eq. 3.1 and get the value of  $K$  when the accumulated  $D_j$  value in the top  $K$  dimensions reaches the value of  $T_d$ . Here  $T_d$  defines the contribution portion of the top  $K$  dimensions that contribute to the overall latent subspaces. Users can define it according to their statistical confidence on errors.

**Step 3:** For a dimension  $n$ , the contribution value  $W_{in}$  of a term or a document  $X_i$  is calculated by Eq. 3.2, where  $X_{in}$  is the  $n^{th}$  dimension projection score, the least-square distance. The overall contribution of a term or a document  $X_i$  to the overall term/document subspace is:

$$W_{in} = \sum_{n=1}^K D_n X_{in}^2 \dots\dots\dots(3.2)$$

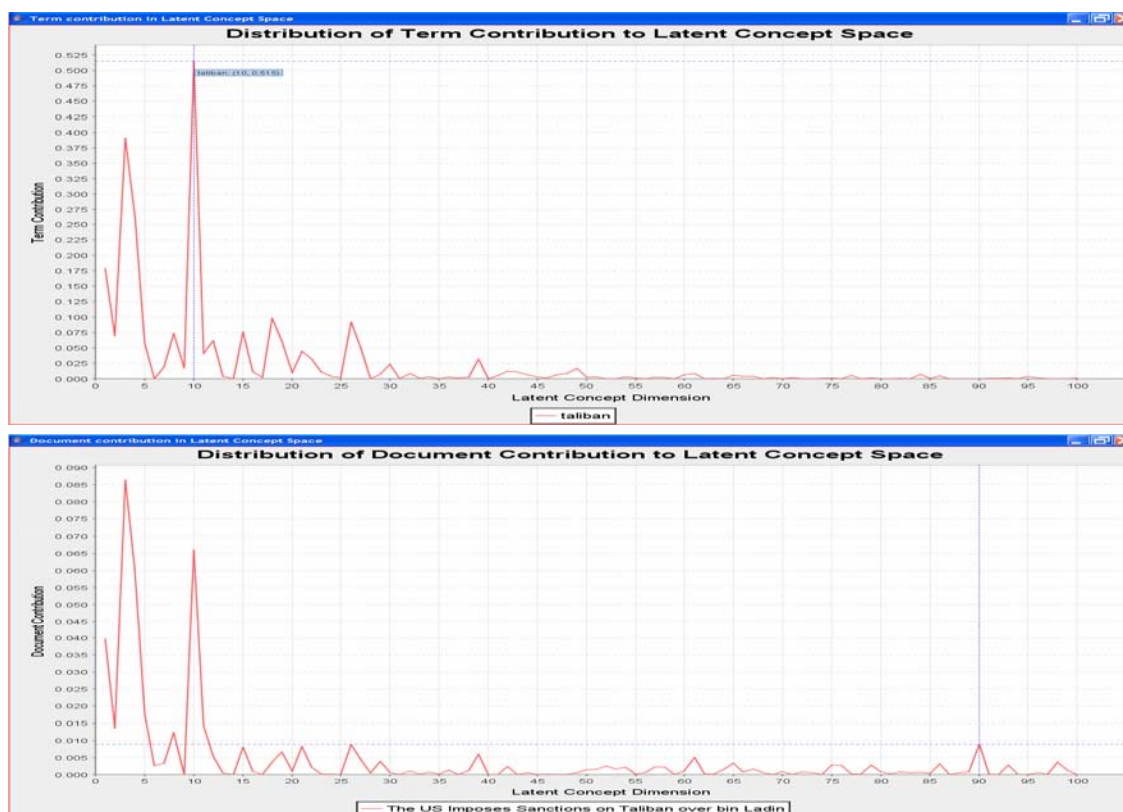
In the Eq. 3.2,  $K$  is selected from Step 2. Using percentages the contribution of a term/document to the selected latent subspaces is:

$$\sum_{n=1}^K W_{in} / \sum_{i=1}^t \sum_{n=1}^K W_{in} \dots\dots\dots(3.3)$$

In the Eq. 3.3,  $t$  is total number of the terms or the documents in the corpus.

**Step 4:** In the top-K dimensions the contribution distribution of a term or a document  $X_i$  is represented a vector of  $W_{ij}$  as a signature, where  $j$  belongs to  $\{1, K\}$ , see Fig. 3-2.

In the figure, the X axis in each image indicates the LSI latent concept dimensions and the coordinates in Y axis are weighted by  $W_{ij}$ . The data were taken from the analysis of a collection of 948 news articles extracted from a terrorism news resource<sup>3</sup>.



**Figure 3-2: The Term Signature of the Key Topical Theme: “Taliban” (the Upper Image) and the Document Signature of a News Article (the Lower Image).**

<sup>3</sup> <http://www.ict.org.il/>

The coherence between the position and height of the peaks in the two signatures make us believe that these signatures capture the semantics of the document and a combination of term signatures represents a document signature. This phenomenon motivates us to create a similarity metric which links the term signatures and document signatures through the projections to the same LSI latent concept dimensions. Another important phenomenon is that even though the projection weight is non-negative, the signs of the projection scores affect the match between the term signatures and document signatures. For instance, if the projection score of a term is negative in one dimension, the most related documents will have negative projection scores in the same dimension. This suggested that the association will be counted only when the term signatures and the document signatures have higher project scores with the same signs. The signature images in Fig. 3-2 show that each signature has many peaks and some of them are much lower than others. The lower peaks are probably noise.

## 3.2 LSI Subspace Signature Ranking

### 3.2.1 Introduction

The signature ranking algorithm, which concerns the global and the local contribution ranking and so called GLCR, iteratively picks terms/documents based on their rankings until these terms and documents collectively make a predefined threshold,  $T_g$ , to the ratio in Eq. 3.3. The term-frequency distribution in a corpus follows Zipf's Law (Gelbukh and Sidorov, 2001) which suggests that a small portion of terms makes major statistical contribution to the whole corpus. According to this notion, when using GLCR to select terms, the value of  $T_g$  are set high enough to make sure the term subset that has major contribution to the overall semantic space is kept.

GLCR includes two steps and concerns both the global contribution and the local contribution of documents and terms to the LSI subspaces. The global contribution of one term or one document is calculated by Eq. 3.2 and the local contribution of one term or one document is estimated by the absolute value of its projection score to one latent concept dimension. GLCR first selects a threshold for the local contribution,  $T_l$ .  $T_l$  is experiment-driven and is generally initialized as a ratio to the mathematic mean of the absolute value of the projection score in the  $k^{th}$  latent concept dimension,  $x_{ik}$ , across the top  $K$  dimensions. Some terms have higher scores of the global contribution, but the projection value to each of the top dimensions is lower than  $T_l$ . Such terms are lack of the discriminating power and are ignored. GLCR scans the top  $K$  dimensions. The term and the document selection start from the dimension with a higher singular value score and follow a step-wise strategy. That means that if a term

or a document has high projection scores in many dimensions, when it is selected in a dimension with a higher singular value, it will not also be counted in later dimensions. Then, each term or each document in the subset is ranked by the value calculated by Eq. 3.2. GLCR has advantages of selecting the documents which belong to an outlier category. For instance, even in a large corpus, one category might include only one or two samples. Each of those has a small global contribution, but might have a higher local contribution in certain dimensions because it represents distinct topics and has high discriminating power.

### ***3.2.2 Experiments, Text Preprocessing and Data Sets***

GLCR ranking can be applied to both term signatures and document signatures. In this section, the experiments are designed to evaluate the term signature ranking. The document signature ranking is applied for active learning and the experiment results are described in Chapter 5.

The term signature ranking in the LSI Subspace Signature Model picks the distinct terms which makes the major contribution to the LSI term subspace. The ranking algorithm is applied to news articles, Email Conversation and the ISI citation abstracts related to Sloan Digital Sky Survey<sup>4</sup> (SDSS).

Three news corpora, the Reuters 21587, TDT1 and TDT2 collections, are used. These were selected because they category labels were provided along with the articles. From the Reuters21578–Apte-90Cat corpus, 2527 training news articles are selected

---

<sup>4</sup> <http://www.sdss.org/>

and belong to 10 categories<sup>5</sup>, acq, coffee, interest, iron-steel, oat, palmkernel, sugar, sun-meal, veg-oil and wheat. The sample size for each category varied from 1 to 1646. For TDT1, we generate a subset of that consisting of all the 25 categories containing 1131 documents which are evaluated as “YES” rather than “BRIEF” in the evaluation sheet of TDT1. The sample size for each category varied from 2 to 273. The full TDT2 corpus has 100 categories (news story topics). We applied a random selection to generate a subset of that consisting of 30 categories (topics) containing 3349 documents which are evaluated as “YES” rather than “BRIEF” in the evaluation sheet of TDT2 while the selection make sure that the subset includes the outlier categories. The sample size for the categories varied from 1 to 1132.

This study on email conversations explores the evolution of the key discussion topics over an extended period of time in the W3C URI working group. The dataset is a subset of the testing corpus of the TREC Enterprise 2005. Our idea is to represent each actor in the social network of the working group with a document which contains all the emails the actor had sent in a given time frame and then to extract and rank concepts from these documents. At first, 4257 emails of the W3C URI working group among 388 members are divided into 388 separate documents. Each document is sub-divided into 11 pieces according to the years from 1994 to 2004. Next, we applied the Stanford part-of-speech (POS) tagging (Toutanova and Manning, 2000; Toutanova et al., 2003), stop-word filtering using the Google stop word list and Porter

---

<sup>5</sup><http://www.daviddlewis.com/resources/testcollections/reuters21578/>

stemming to the corpus. A total of 8,647 noun concepts are selected for the subsequent text analysis. All the nouns with a less than 2 occurrences are excluded. The initial associative relationship between a noun term and a document is weighted by traditional TFIDF.

### 3.2.3 *Experiment Results*

The sizes of the term subsets are determined by the values of  $T_g$ ,  $T_l$  and  $T_d$  which are between 0.0 and 1.0.  $T_d$  determines how many top dimensions are selected, which reduces noise by eliminating more dimensions. In general a larger  $T_l$  value means less noise.  $T_g$  decides the upper boundary of the overall contribution of the term/document subset if  $T_l$  and  $T_d$  are predefined. We changed  $T_d$ ,  $T_l$  and  $T_g$  to maximize the information included while minimizing the noise.

For example, from TDT1, GLCR selects a subset of 1614 concept terms out of the whole set (8338 terms), which makes an 82.30% (Maximum value of  $T_g$ ) statistical contribution to the overall LSI term subspace. From the TDT2, GLCR selects a subset of 3007 concept terms out of the whole set (17083 terms), which makes an 85.14% ( $T_g$ ) statistical contribution to the overall LSI term subspace. From the Reuters dataset, if  $T_d$  is set to 0.95 and  $T_l$  is set to 0.5, GLCR selects a subset of 2673 concept terms out of the whole set (9070 terms). This makes an 83.3% (Maximum value of  $T_g$ ) statistical contribution to the selected LSI term subspace. For the Reuters dataset, if  $T_d$  is set as 0.8 and keep  $T_l$  the same, GLCR selects a subset of 2048 concept terms out of the whole set (9070 terms), which makes an 82.5% (Maximum value of  $T_g$ ) statistical contribution to the selected LSI term subspace. The result indicates that the

size of the term subset increases by 625 after  $T_d$  grows from 0.8 to 0.95, which suggests many distinct topic terms have high local contribution to the low-rank latent concept dimensions. These terms might be useful distinguishing the documents and grouping them.

The top 20 terms listed in Table 3-1 ranked by the GLCR method from the subset of the Reuters 21578 corpus cover most of the class labels, for instance, “acq, coffee, interest, iron-steel, sugar, wheat, palmkernel, oat, sun-meal, and veg-oil”.

**Table 3-1: Top 20 Themes which Make the Highest Contribution to the LSI Term Subspace of the Reuters Corpus Ranked by GLCR.**

<b>Rank</b>	<b>Term</b>	<b>Rank</b>	<b>Term</b>
1	stg	11	reserve
2	tonn	12	Taft
3	rate	13	Oil
4	bank	14	Bill
5	sugar	15	Twa
6	wheat	16	federal
7	coffee	17	export
8	gencorp	18	purol
9	cyclop	19	chemlawn
10	usair	20	money

The topical terms in the email conversations of the URI working group are ranked by GLCR and the most important concepts are highlighted with bold in Table 3-2 and Table 3-3. For each year, the top ten themes are listed. These terms cover the key themes for URI, for instance, “urn, iri, character, base uri, fragment, ipv, resource” (see <http://gbiv.com/protocols/uri/rfc/rfc3986.html>). Because TFIDF weights are used in text processing, these noun terms with an IDF=0 are excluded in the ranking list.

For instance, the ranking list for 1994-2004 excludes a list of terms, “uri, url, name, scheme, and example”. Obviously, these terms are very important and are considered for understanding the concept evolution. The highlighted terms reflect the concept building history in the URI working group in the ten-year period. The concepts shift from “url, uri” to “urn” and then “iuri, iri”, which matches the concept development history of URI<sup>6</sup>. Overall, the study demonstrates the evolution of the conceptual temporal structures in the social network of the URI working group.

**Table 3-2: Discussion Topic Evolution in the URI Working Group from 1994 Dec to 1998.**

<b>Rank</b>	<b>1994-2004</b>	<b>1994</b>	<b>1995</b>	<b>1996</b>	<b>1997</b>	<b>1998</b>
1	Fragment	Some	<b>Urn</b>	<b>Urn</b>	Rtsp	Academy
2	<b>Urn</b>	Body	Rate	Vemmi	Utf	Nntp
3	Character	Davenport	Cid	Irc	Character	Lb
4	Lid	Ics	Initiative	Wnetc	Div	Encode
5	Rate	Herald	Range	Fragment	Chri	Script
6	Utf	Norwegian	Digest	Mud	Imap	Gaymen
7	Base	Usenet	Ipv	Draft	Base	Utf
8	<b>Iuri</b>	Alvestrand	Finger	Deployment	Susan	Axiom
9	<b>Iri</b>	Cmu	Docid	Local	Numer	Cesuscd
10	Vemmi	Usage	Lyco	Acct	Fragment	Networkd
Actor Size	388	11	137	54	70	37

---

<sup>6</sup> see, <http://www.w3.org/Addressing/>

**Table 3-3: Discussion Topic Evolution in the URI Working Group from 1999 Dec to 2004 May.**

<b>Rank</b>	<b>1999</b>	<b>2000</b>	<b>2001</b>	<b>2002</b>	<b>2003</b>	<b>2004</b>
1	<b>Urn</b>	Lid	Null	Lm	<b>Urn</b>	Snmp
2	Admin	Utf	Webdav	Smb	Mm	File
3	Error	Xml	Ark	Base	Fragment	Dollar
4	Palceum	Base	Dav	Query	Openurl	Namespace
5	Busy	Reagl	Protozilla	Rdf	Catalog	<b>Iri</b>
6	Termin	Sysrcus	Tftp	Offer	Tld	Sm
7	Nature	Idn	Index	<b>Iri</b>	Ni	Associative
8	Product	<b>Iuri</b>	Christian	Oai	Thing	Fragment
9	Paper	Entity	<b>Iri</b>	Identity	Pgp	Resource
10	Leslie	Gerald	<b>Urn</b>	Yahoo	Dan	Info
Actor Size	19	39	78	67	97	52

CLCR extracts the top five concepts from the 61 ISI records of Dr. Michael Vogeley, like “void galaxy”, “power spectrum”, “genu curve”, “largescale” and “release”. He verifies that these concepts are good summaries for his research.

### **3.2.4 Conclusion**

The LSI term signature ranking method, GLCR, selects the themes which make the major contribution to the corpus and controls the level of the dimension reduction. The top-ranking terms reflect the most popular topics. Using CLCR in the concept mapping, the system like Storylines helps users to understand the corpus even without prior knowledge. Combined with the temporal factor, the evolution of the themes suggests the conceptual structure change in the corpus.

### 3.3 Similarity Measures between Signatures

A vector of  $DX^2$  in a LSI term subspace and a LSI document subspace represents a term signature and a document signature respectively. The signature similarity between a concept  $r_a$  and a document  $r_u$  (CDSS) is calculated by a variation between cosine similarity and Pearson Correlation and the formula is shown in Eq. 3.4:

$$CDSS(r_a, r_u) = \frac{\alpha \sum_{i=1}^K (r_{a,i} - \beta \bar{r}_a)(r_{u,i} - \beta \bar{r}_u)}{\sqrt{\sum_{i=1}^K (r_{a,i} - \beta \bar{r}_a)^2 \sum_{i=1}^K (r_{u,i} - \beta \bar{r}_u)^2}} \quad \dots\dots\dots (3.4)$$

In Eq. 3.4,

$$\bar{r}_x = \frac{\sum_{i=1}^K r_{x,i}}{K} \quad \alpha = \begin{cases} 1 & \text{if } d_{a,u} \geq d_{avg} \\ \frac{d_{a,u}}{d_{avg}} & \text{if } d_{a,u} \leq d_{avg} \end{cases}$$

$K$  is decided by  $T_d$ ,  $\alpha$  is the significance weight which is related to the number of co-projected latent concept dimensions of two signatures with a range from 0.0 to 1.0 and  $\beta$  is a pruning parameter which decides how many dimensions in the top  $K$  latent concept dimensions are included in the similarity accounting. If  $\beta$  equals 0.0, all the top  $K$  dimensions are concerned in the similarity measure. And if  $\beta$  equals 1.0, only the dimensions with a higher value than that of  $\bar{r}$  are concerned. The value of  $\beta$  could be larger than 1.0.  $\bar{r}$  is the average score of the signature across the top  $K$  dimensions and  $d_{avg}$  in  $\alpha$  is the average number of the selected dimensions over every pair of the

signatures. The association is counted only when the two signatures have projection scores with the same sign on one dimension (e.g., both negative / both positive).

In a traditional term-document matrix, the association between terms and documents is weighted by term frequency or TFIDF, which does not reflect the relationship strength between the documents and the concepts directly. The LSI signatures concern both global statistical contribution and local statistical contribution to the whole corpus. The similarity between term signatures and document signatures represents the normalized association between a theme and a document by reducing noise. And in general, the most important themes are more likely to be used as the labels of the documents. For instance, the top 20 terms listed in Table 3-1 ranked by GLCR include most of the class labels in the according news corpus. So CDSS weighted by the global statistical contribution of the concept signatures, is more easily to identify the representative terms for the documents. Thus, we propose a variation of the CDSS measure -- global contribution enhanced CDSS (GCDSS), see Eq. 3.5, which simply multiplies the CDSS similarity score by Eq. 3.4 with the global contribution score of each concept signature calculated by Eq. 3.2. GCDSS is only used between a concept signature and a document signature.

$$GCDSS(r_a, r_u) = \sum_{n=1}^K D_n X_{an}^2 CDSS(r_a, r_u) \dots\dots\dots (3.5)$$

The document-term matrix is re-constructed and the association between a term and a document is weighted by CDSS and GCDSS. In this stage, the similarity measures only count the relationships between the terms and the documents where the terms

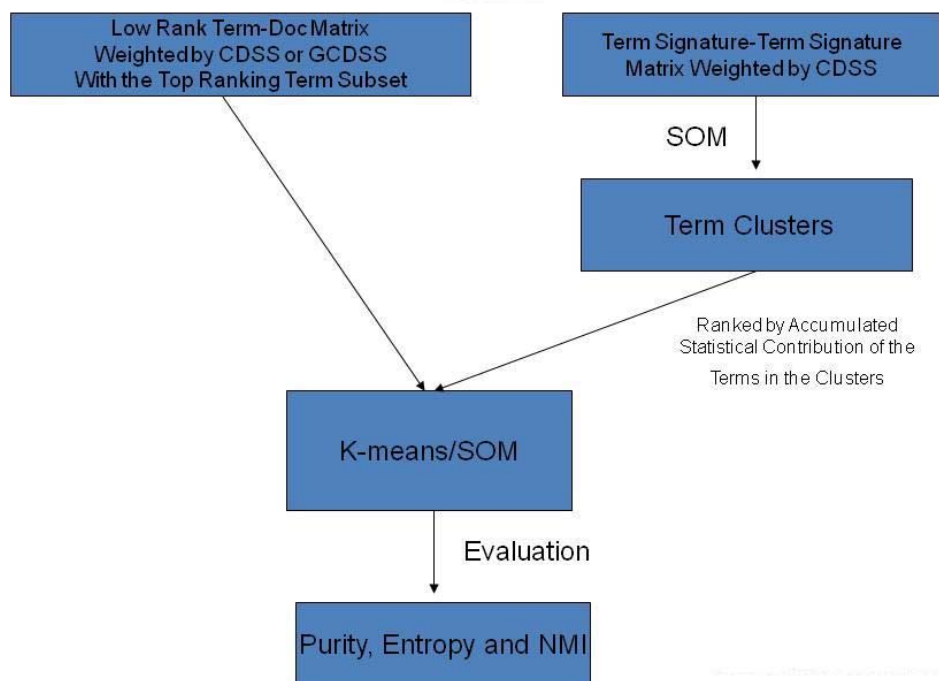
appear. Using GCDSS the top ranking terms have higher scores compared to those generated by CDSS, which suggests the representatives for the documents. By observation, many terms used as the document labels are ranked highest within the documents. The groups of the top-ranking terms represent the document clusters. Thus, the representative terms are potentially useful to identify the multiple labels of documents.

## **CHAPTER4: Text Clustering Using LSISSM**

### **4.1 Introduction**

The LSI Subspace Signature Model has controls on the level of the dimension reduction. The term-signature ranking narrows the scope of the features to a limited number of conceptual terms and the parameters of the model make sure the term subset make the main global contribution to the LSI term subspace. This mechanism is determined by the characteristics of the term distribution in a text corpus, the Zipf's distribution or power-law distribution. Using the term subset to present the overall semantic space, a highly scalable term-document matrix is transformed into a low-dimension term-document matrix weighted by CDSS or GCDSS. Two novel similarity schemas, CDSS and GCDSS match any concept signature with any document signature and prune the association between the signatures. So, the model reduces a lot of noise with parameter controls. Thus, experiments are designed to evaluate how the model affects the performance of the basic clustering algorithms because how to select the discriminated terms and how to reduce the noise from the relationships between term and documents are the major challenges for text clustering. Two standard clustering algorithms are tested: basic K-means and SOM. Moreover, to solve the initialization problem of the basic K-means, a new K-means algorithm, called two-stage K-means, is proposed.

#### 4.1.1 Model-based K-means and SOM



**Figure 4-1: Document Clustering Based on Top Ranking Terms versus Top Ranking Term Clusters.**

There are two ways to apply the LSI Subspace Signature Model to the basic K-means and SOM, see Fig. 4-1. One way is to use the low-dimension term-document matrix as input for the two algorithms. Another way is that the top ranked conceptual terms are grouped first by the clustering algorithms and then the term clusters are ranked according to the sum-up global contribution of the terms in the clusters. The term clustering algorithm is SOM (Lin et al., 1991) because the number of term clusters cannot be pre-defined. The term-term similarity matrix is constructed by the CDSS

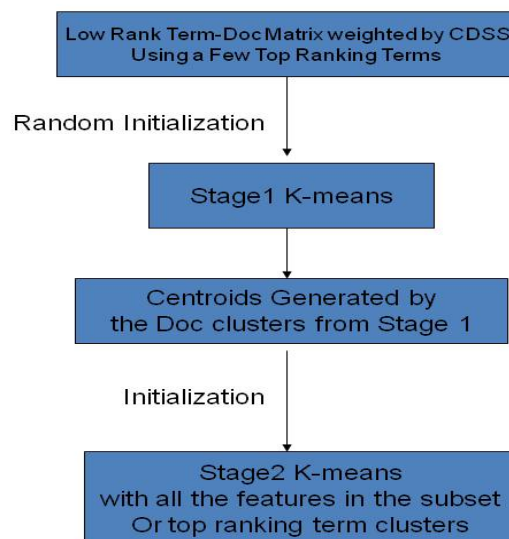
measure between each pair of the term signatures. The value of  $\beta$  is set from 1.0 to 1.5 in the experiments for term clustering. After term clustering, each concept cluster is ranked by the accumulated value of the global statistical contribution of all the terms in one cluster to the LSI latent term subspace. Only those terms which appear in the top ranking term clusters are used in the input matrix.

#### ***4.1.2 Model-based Two-stage K-means***

Basic K-means (Dubes and Jain, 1988) minimizes the sum of the squared distances between each point in the dataset and the closest center. The K centers are predefined to start from the randomization seeding. The random initialization procedure dramatically affects the performance of the K-means algorithm and sometimes causes very poor performance. Our approach indicates a different initialization procedure based on the LSI Subspace Signature Model, so called two-stage initialization. In the first stage, with a fixed number of the clusters and a small feature subset with the top ranking terms generated by the model, the basic K-means is processed. Second, the centroids of the K clusters produced by the first step are used as the initial centroids of the second stage using a large feature set or even full feature set. Then, the standard K-means algorithm is preceded again. The results show that this strategy improves the performance of the basic K-means on both efficiency and accuracy. K-mean Plus (Arthur and Vassilvitskii, 2007) mathematically shows that the initial clusters positioned in the dense data regions which make the major statistical contribution improve the performance of the basic K-means efficiently and effectively. And, the method in that work is applied to no-document data. Silic et al.

(2008) used the seeding method in K-means++ (Arthur and Vassilvitskii, 2007) but did not report the effect.

From the perspective of concept representation our approach targets to find and use the key conceptual term sets to represent the dense regions of the document clusters. Our assumption for the two-stage K-means algorithm is that the key term subset which makes the main contribution to the semantic concept space is most likely associated with the dense regions of document clusters. The stochastic processing of the basic K-means converges very fast if the key term set is very small. The resulting centroids are more likely better positions than those generated randomly. The conceptual feature subset is selected by the term signature ranking algorithm of LSISSM. The architecture is demonstrated in Fig. 4-2.



**Figure 4-2: The Architecture of Two-stage K-means.**

In the first stage, the feature subset is composed by the top ranking terms, for instance, the top 100 terms. We propose two strategies to define the second stage of the two-stage K-means algorithm. One strategy is to use the full feature subset. Another strategy is to use the term clusters. The full feature subset is clustered by Self-organizing Mapping (Lin et al. 1991) and the term clusters are ranked by the accumulated statistical contribution of the terms inside the clusters. And, the terms in the top ranking clusters are used to in the second stage.

The procedures of the two-stage K-means work as follows:

**Step 1:** Arbitrarily choose  $k$  initial centers  $C = \{c_1, \dots, c_k\}$ .

**Step 2:** For each document  $\chi_i$ , set the cluster  $C_i$  to be the set of documents in  $\chi$  that are closer to  $c_i$  than any other centers.

**Step 3:** For each document  $\chi_i$ , set  $c_i$  to be the center of mass of all points in  $C_i$ :

$$c_i = 1 / |C_i| \sum_{x \in C_i} x \dots\dots\dots(4.1)$$

**Step 4:** Repeat Steps 2 and 3 until  $C$  no longer changes.

The two-stage K-means focuses on the first step of the algorithm. In the first stage, the  $k$  initial centers are generated randomly and each document is represented by a small set of the top ranking terms picked with LSISSM. After the algorithm converges, the centers of the clusters produced by the first stage are calculated by Eq. 4.1. Then these centers are used as the initial seeds for the second stage to repeat Steps 2 to 4. In the second stage, the corpus is represented by a much larger set of the top terms, say 1500 of them.

## 4.2 Evaluation Methods and Datasets

We used the same three news collections, Reuters 21587, TDT1 and TDT2. The class labels were provided along with the articles. The details of the three datasets are introduced in Section 3.2.2.

Data pre-processing directly influences the quality of the sampling. First, the Stanford part-of-speech (POS) tagging (Toutanova and Manning, 2000; Toutanova et al., 2003), stop-word filtering is applied to the corpus using the Google stop word list and Porter stemming. Each article for one category has to include at least one noun term that could be identified by the POS tagging. All the noun terms appear at least once in the corpus. Only noun terms are included in the analysis because our study emphasizes the concept representations of the documents. The association between a noun term and a document in the term-document matrix is weighted by traditional TFIDF. Each column of the matrix is normalized to 1.0.

For the non-hierarchical clustering, Purity (Manning et al., 2008) and the Entropy measure (Beil et al., 2002) are applied to evaluate the pureness of the clusters. Xu (2003) and Cai (2005) use normalized mutual Information (NMI) metrics which is applied to every clustering method. The advantage of NMI is that its value is not affected by the number of the clusters. For Purity and NMI, higher values are better. And a lower entropy value means a better performance. In our study the entropy value is normalized by the size of the categories to make sure that it has a range between 0.0 and 1.0, and so called relative entropy.

The VSM baselines of the clustering algorithms are generated with the full-feature set and the similarity schema TFIDF. The experiments are designed to demonstrate the enhancement on the effectiveness and efficiency of the clustering algorithms from the LSI Subspace Signature Model.

### 4.3 Experiment Results

In the following tables the  $T_g$  column represents the accumulated statistical global contribution of the feature sets to the LSI term subspace calculated by the GLCR ranking method. For example, in Table 4-1, the top 200 terms make 28.0% contribution to the LSI term subspace of the Reuters data set. The Top Terms column lists the number of the top conceptual terms included in the subset. Across the three datasets, the basic K-means algorithm uses the same parameters. The predefined number of clusters for the K-means is 10, 25 and 30 respectively for the three data sets with 10 iterations. SOM also applies the same parameters across the three datasets. Epsilon is 0.25, and each run goes for 2500 iterations with 12 nearest neighbors. The value of  $\beta$  is set as 0.0 in CDSS for every run in this chapter comparing the term signatures and the document signatures.

#### 4.3.1 *K-means Using Top Ranking Terms*

In Table 4-1, the basic K-means reaches the best points with the top 2600 terms on both Purity and Entropy measures, and receive the maximum value of NMI with a feature subset with the top 600 terms. The peak points are shown in bold in the following tables.

**Table 4-1: The Clustering Evaluation Matrix of the Basic K-means Using the Similarity CDSS and the Reuters Corpus.**

<b>Top Terms</b>	<b>T<sub>g</sub></b>	<b>Purity</b>	<b>Entropy</b>	<b>NMI</b>
5	0.030	0.658	0.414	0.173
10	0.050	0.695	0.351	0.261
20	0.077	0.665	0.355	0.257
30	0.097	0.696	0.353	0.292
50	0.129	0.740	0.303	0.358
100	0.191	0.756	0.303	0.342
200	0.280	0.783	0.277	0.394
300	0.349	0.774	0.283	0.377
400	0.404	0.810	0.248	0.431
500	0.450	0.816	0.242	0.448
600	0.489	0.835	0.224	0.475
700	0.524	0.830	0.227	0.472
800	0.554	0.796	0.249	0.431
900	0.582	0.813	0.236	0.445
1000	0.608	0.800	0.243	0.442
1100	0.633	0.838	0.221	0.462
1200	0.655	0.795	0.246	0.438
1300	0.676	0.841	0.218	0.464
1400	0.695	0.814	0.239	0.444
1500	0.712	0.833	0.228	0.457
1600	0.728	0.818	0.234	0.453
1700	0.743	0.829	0.227	0.465
1800	0.757	0.833	0.225	0.455
1900	0.770	0.831	0.226	0.464
2000	0.781	0.827	0.219	0.444
2200	0.802	0.827	0.229	0.456
2400	0.819	0.848	0.207	0.471
2600	0.830	<b>0.853</b>	<b>0.200</b>	<b>0.462</b>
2673	0.833	0.835	0.224	0.463

In Table 4-2, the basic K-means reaches the best scores by the subset with the top 1500 terms evaluated by the three evaluation methods.

**Table 4-2: The Clustering Evaluation Matrix of the Basic K-means Using the Similarity CDSS and the TDT1 Corpus.**

<b>Top Terms</b>	<b>T<sub>g</sub></b>	<b>Purity</b>	<b>Entropy</b>	<b>NMI</b>
10	0.070	0.462	0.510	0.372
20	0.110	0.556	0.387	0.546
30	0.137	0.734	0.257	0.688
40	0.161	0.744	0.235	0.707
50	0.184	0.729	0.245	0.696
70	0.223	0.767	0.216	0.722
100	0.271	0.854	0.151	0.808
200	0.390	0.867	0.126	0.832
300	0.467	0.805	0.150	0.813
400	0.527	0.835	0.132	0.818
500	0.575	0.859	0.119	0.843
600	0.617	0.817	0.142	0.816
700	0.653	0.862	0.112	0.848
800	0.684	0.861	0.111	0.843
900	0.711	0.837	0.138	0.822
1000	0.735	0.867	0.120	0.837
1100	0.756	0.809	0.141	0.823
200	0.774	0.833	0.125	0.833
1300	0.790	0.878	0.108	0.840
1400	0.804	0.863	0.112	0.840
1500	0.815	<b>0.884</b>	<b>0.097</b>	<b>0.849</b>
1600	0.823	0.782	0.170	0.788

In Table 4-3, the basic K-means reaches the best scores with the top 2400 terms in either of Purity and Entropy. If using the NMI measure, the basic K-means reaches the maximum score with the subset that includes the top 2000 terms.

**Table 4-3: The Clustering Evaluation Matrix of the Basic K-means Using the Similarity CDSS and the TDT2 Corpus.**

<b>Top Terms</b>	<b>T<sub>g</sub></b>	<b>Purity</b>	<b>Entropy</b>	<b>NMI</b>
10	0.052	0.460	0.475	0.265
20	0.081	0.657	0.315	0.492
30	0.102	0.673	0.286	0.564
50	0.139	0.872	0.152	0.719
70	0.171	0.871	0.122	0.773
100	0.212	0.890	0.112	0.782
200	0.311	0.898	0.098	0.753
300	0.381	0.929	0.088	0.819
400	0.437	0.923	0.084	0.822
500	0.483	0.906	0.093	0.806
600	0.522	0.911	0.087	0.815
700	0.556	0.918	0.083	0.813
800	0.586	0.924	0.078	0.815
900	0.613	0.925	0.073	0.823
1000	0.636	0.924	0.074	0.823
1100	0.658	0.918	0.090	0.773
1200	0.677	0.904	0.087	0.795
1300	0.695	0.927	0.071	0.806
1400	0.711	0.936	0.064	0.817
1500	0.726	0.940	0.066	0.810
1600	0.740	0.899	0.077	0.791
1700	0.753	0.924	0.073	0.804
1800	0.765	0.913	0.078	0.807
1900	0.776	0.922	0.080	0.805
2000	0.786	0.942	0.064	0.829
2100	0.796	0.925	0.076	0.794
2200	0.805	0.939	0.066	0.810
2300	0.813	0.924	0.082	0.806
2400	0.820	<b>0.943</b>	<b>0.063</b>	<b>0.823</b>
2500	0.827	0.929	0.067	0.817
2600	0.833	0.939	0.066	0.810
3000	0.851	0.928	0.075	0.818

#### 4.3.2 *K-means Using Top Ranking Term Clusters*

The value of  $\beta$  is set to 1.0 in CDSS to produce term-term similarity matrix in this section. In Table 4-4, the basic K-means reaches the best points with the top 200 clusters on both Purity and Entropy measures, and receive the maximum value of NMI with the top 175 clusters.

**Table 4-4: The Clustering Evaluation Matrix of the Basic K-means Using the Term Clusters and the Reuters Corpus.**

Top Clusters	$T_g$	Purity	Entropy	NMI
C5	0.148	0.788	0.306	0.324
C10	0.225	0.825	0.258	0.395
C15	0.280	0.805	0.267	0.385
C20	0.320	0.820	0.254	0.397
C25	0.351	0.800	0.260	0.401
C30	0.377	0.803	0.254	0.402
C50	0.469	0.809	0.262	0.385
C100	0.618	0.761	0.264	0.393
C150	0.710	0.808	0.242	0.444
C175	0.743	0.812	0.238	0.475
C200	0.770	<b>0.833</b>	<b>0.224</b>	<b>0.471</b>
C250	0.809	0.772	0.257	0.427

In Table 4-5, the basic K-means reaches the best scores by the subset with the top 100 clusters evaluated by either of the three evaluation methods.

**Table 4-5: The Clustering Evaluation Matrix of the Basic K-means Using the Term Clusters and the TDT1 Corpus.**

Top Clusters	$T_g$	Purity	Entropy	NMI
C5	0.161	0.790	0.216	0.718
C10	0.257	0.811	0.165	0.774
C15	0.324	0.813	0.150	0.801
C20	0.383	0.810	0.173	0.788
C25	0.431	0.852	0.120	0.833
C50	0.567	0.845	0.128	0.840
C100	0.702	<b>0.882</b>	<b>0.098</b>	<b>0.850</b>
C150	0.778	0.843	0.128	0.817
C200	0.817	0.823	0.140	0.815

In Table 4-6, the basic K-means achieves the maximum scores of the three evaluation methods with the top 50 clusters.

**Table 4-6: The Clustering Evaluation Matrix of the Basic K-means Using the Term Clusters and the TDT2 Corpus.**

Top Clusters	$T_g$	Purity	Entropy	NMI
C5	0.215	0.873	0.135	0.728
C10	0.319	0.912	0.093	0.807
C15	0.380	0.887	0.105	0.778
C25	0.461	0.909	0.086	0.807
C30	0.490	0.865	0.107	0.761
C50	0.581	<b>0.936</b>	<b>0.068</b>	<b>0.825</b>
C100	0.707	0.915	0.083	0.802
C150	0.773	0.935	0.073	0.815
C200	0.810	0.936	0.069	0.811

### 4.3.3 Two-stage K-means

In the experiments, the two-stage K-means generally picks a term subset which size is no more than 100 in the first stage because the document clustering algorithm can converge very fast. In Tables 4-7, 4-8 and 4-9, a plus (+) sign in the Dataset column denotes a run using the two-stage K-means.

**Table 4-7: Comparison of the Running Time (seconds) between the Basic K-means and the Two-stage K-means Using the Top Ranking Terms and the Top Ranking Term Clusters with the Reuters corpus, TDT1 and TDT2.**

Corpus	Dataset	Running Time: 1st stage	Running Time: 2nd stage	Running Time: Overall
Reuters	2600	-----	-----	9368.8
Reuters	20+2600	<b>43.2</b>	<b>2842.5</b>	<b>2885.7</b>
Reuters	C200	-----	-----	6991.8
Reuters	20+C200	<b>42.6</b>	<b>2437.7</b>	<b>2480.3</b>
TDT1	1500	-----	-----	1360.4
TDT1	100+1500	<b>114.7</b>	<b>111.6</b>	<b>226.3</b>
TDT1	C100	-----	-----	1124.4
TDT1	100+C100	<b>114.7</b>	<b>75.5</b>	<b>190.2</b>
TDT2	2400	-----	-----	15804.6
TDT2	40+2400	<b>497.3</b>	<b>3998.5</b>	<b>4495.8</b>
TDT2	C200	-----	-----	13298.9
TDT2	40+C200	<b>492.0</b>	<b>2766.8</b>	<b>3258.9</b>

For instance, 20+2600 denotes in the first stage a subset with the top 20 terms is used to represent the documents and in the second stage the feature subset is expanded to 2600 features. And, 20+C200 denotes in the first stage the documents are represented by a top ranking feature set with a size 20 and in the second stage 200 concept clusters are used as the document representation. The concept clusters used in this section are the same as ones in Section 3.3.2. In Table 4-7, the running time for both

stages is recorded. Compared to the running time of the basic K-means, like Reuters (C200) and TDT1 (1500), the runs of the two-stage K-means Reuters (20+C200) and TDT1 (20+1500) are much faster. TDT1 (20+1500) saves 83% time compared to TDT1 (1500).

The runs in Table 4-7 not only save a huge amount of time but improve the clustering accuracy according to the three evaluation methods. For instance, in Table 4-8, compared TDT1 (1500) and TDT2 (2400), TDT1 (100+1500) and TDT2 (40+2400) outperform on each of the three evaluation measures. And, the three runs with the Reuters corpus have better scores only for NMI.

**Table 4-8: Comparison of the Clustering Performance between the Basic K-means and the Two-stage K-means Using the Similarity CDSS and the Top Ranking Terms with the Reuters Corpus, TDT1 and TDT2.**

Corpus	Dataset	Purity	Entropy	NMI
Reuters	2600	<b>0.853</b>	<b>0.200</b>	0.462
Reuters	20+2600	0.845	0.211	<b>0.470</b>
TDT1	1500	0.884	0.097	0.849
TDT1	100+1500	<b>0.913</b>	<b>0.090</b>	<b>0.864</b>
TDT2	2400	0.943	0.063	0.823
TDT2	40+2400	<b>0.947</b>	<b>0.059</b>	<b>0.827</b>

And, in Table 4-9, compared to TDT1 (C100) and TDT2 (C200), TDT1 (100+C100) and TDT2 (40+C200) are better on the scores with any of the three evaluation measures. For the Reuters corpus, with the top 200 concept clusters, the two-stage K-means has better scores in Purity and Entropy.

**Table 4-9: Comparison of the Clustering Performance between the Basic K-means and the Two-stage K-means Using the Similarity CDSS and the Top Ranking Term Clusters with the Reuters Corpus, TDT1 and TDT2.**

Corpus	Dataset	Purity	Entropy	NMI
Reuters	C200	0.833	0.224	<b>0.471</b>
Reuters	20+C200	<b>0.841</b>	<b>0.218</b>	0.455
TDT1	C100	0.882	0.098	0.850
TDT1	100+C100	<b>0.906</b>	<b>0.095</b>	<b>0.861</b>
TDT2	C200	0.936	0.069	0.811
TDT2	40+C200	<b>0.947</b>	<b>0.060</b>	<b>0.836</b>

#### 4.3.4 SOM Using Top Ranking Terms

In Table 4-10, SOM finds the turning point from the feature subset with the top 2600 terms according GCDSS if evaluated by Entropy. The best score for Purity is received by the subset with the top 30 terms and the best score for NMI is generated by the subset with only 10 top terms.

**Table 4-10: The Clustering Evaluation Matrix of SOM Using the Similarity GCDSS and the Reuters Corpus.**

Top Terms	$T_g$	Purity	Entropy	NMI
5	0.030	0.730	0.413	0.243
10	0.050	0.818	0.258	0.430
20	0.077	0.799	0.239	0.405
30	0.097	0.864	0.209	0.388
50	0.129	0.810	0.222	0.349
100	0.191	0.838	0.211	0.323
500	0.450	0.849	0.183	0.347
1000	0.608	0.835	0.193	0.331
1800	0.757	0.852	0.168	0.345
2600	0.830	<b>0.848</b>	<b>0.166</b>	<b>0.355</b>

In Table 4-11, SOM gets the maximum scores of NMI and Entropy measures with the top 1000 terms respectively and reaches the peak score of Purity with the top 1600 terms.

**Table 4-11: The Clustering Evaluation Matrix of SOM Using the Similarity GCDSS and the TDT1 Corpus.**

Top Terms	Contribution	Purity	Entropy	NMI
10	0.070	0.453	0.488	0.477
30	0.137	0.688	0.238	0.659
50	0.184	0.743	0.195	0.695
100	0.271	0.768	0.172	0.710
500	0.575	0.755	0.157	0.731
1000	0.735	<b>0.808</b>	<b>0.137</b>	<b>0.755</b>
1600	0.823	0.820	0.147	0.731

In Table 4-12, SOM achieves the maximum scores of the three evaluation methods with the top 3000 terms.

**Table 4-12: The Clustering Evaluation Matrix of SOM Using the Similarity GCDSS and the TDT2 Corpus.**

Top Terms	$T_g$	Purity	Entropy	NMI
10	0.052	0.450	0.491	0.310
30	0.102	0.670	0.297	0.486
50	0.139	0.727	0.242	0.525
100	0.212	0.757	0.191	0.576
500	0.483	0.739	0.203	0.551
1000	0.636	0.761	0.188	0.579
2000	0.786	0.773	0.191	0.563
2500	0.827	0.762	0.201	0.564
3000	0.851	<b>0.785</b>	<b>0.165</b>	<b>0.602</b>

#### 4.3.5 Comparisons with VSM and Traditional LSI

Traditional LSI has been shown to improve the efficiency of document clustering (Schütze and Silverstein, 1997). However it does not guarantee the effectiveness, see Tables 4-13 and 4-14:

**Table 4-13: Comparison of the Clustering Performance between VSM and the Traditional LSI Using the Basic K-means with the Reuters Corpus, TDT1 and TDT2.**

Corpus	Category	Similarity	Top K	Purity	Entropy	NMI
Reuters	10	TFIDF	-----	0.795	0.250	0.389
Reuters	10	LSI	10	<b>0.888</b>	<b>0.223</b>	<b>0.492</b>
TDT1	25	TFIDF	-----	0.815	<b>0.133</b>	<b>0.789</b>
TDT1	25	LSI	10	<b>0.828</b>	0.170	0.745
TDT2	30	TFIDF	-----	<b>0.913</b>	<b>0.072</b>	<b>0.765</b>
TDT2	30	LSI	10	0.905	0.091	0.719

**Table 4-14: Comparison of the Clustering Performance between VSM and the Traditional LSI Using the SOM with the Reuters Corpus, TDT1 and TDT2.**

Corpus	Category	Similarity	Top K	NMI
Reuters	10	TFIDF	-----	0.257
Reuters	10	LSI	20	<b>0.315</b>
TDT1	25	TFIDF	-----	<b>0.690</b>
TDT1	25	LSI	30	0.631
TDT2	30	TFIDF	-----	<b>0.517</b>
TDT2	30	LSI	30	0.475

The results indicate that VSM outperforms LSI in the two of the three corpora. With the increase of the category size in the corpora, traditional LSI becomes less effective. Thus, we picked VSM as the baseline. Compared to VSM, when using the top

ranking terms, the Purity test in Table 4-15 is significant across the three corpora ( $t(2)=4.508, p<0.05$ ) and NMI is also significant ( $t(2)=13.54, p<0.05$ ).

**Table 4-15: Comparison of the Clustering Performance between the VSM baseline and LSISSM Using the Basic K-means, the Similarity CDSS and the Top Ranking Terms with the Reuters Corpus, TDT1 and TDT2.**

Corpus	Similarity	Top Terms	Purity	Entropy	NMI
Reuters	TFIDF	ALL	0.795	0.250	0.389
Reuters	CDSS	2600	<b>0.853</b>	<b>0.200</b>	<b>0.462</b>
TDT1	TFIDF	ALL	0.815	0.133	0.789
TDT1	CDSS	1500	<b>0.884</b>	<b>0.097</b>	<b>0.849</b>
TDT2	TFIDF	ALL	0.913	0.072	0.765
TDT2	CDSS	2400	<b>0.943</b>	<b>0.063</b>	<b>0.823</b>

If using top ranking term clusters, the Purity test in Table 4-16 is significant across the three corpora ( $t(2)=4.508, p<0.05$ ) and NMI is also significant ( $t(2)=43.6, p<0.001$ ). The value of  $\beta$  is set as 1.5 in CDSS for term clustering for every run in table 4-16.

**Table 4-16: Comparison of the Clustering Performance between the VSM baseline and LSISSM Using the Basic K-means, the Similarity CDSS and the Top Ranking Term Clusters with the Reuters Corpus, TDT1 and TDT2.**

Corpus	Similarity	Top Term Clusters	Purity	Entropy	NMI
Reuters	TFIDF	ALL	0.795	0.250	0.389
Reuters	CDSS	C200	<b>0.846</b>	<b>0.210</b>	<b>0.460</b>
TDT1	TFIDF	ALL	0.815	0.133	0.789
TDT1	CDSS	C100	<b>0.923</b>	<b>0.069</b>	<b>0.865</b>
TDT2	TFIDF	ALL	0.913	0.072	0.765
TDT2	CDSS	C200	<b>0.938</b>	<b>0.064</b>	<b>0.836</b>

**Table 4-17: Comparison of the Clustering Performance between the VSM Baseline and LSISSM Using the SOM, the Similarity GCDSS and the Top Ranking Terms with the Reuters Corpus, TDT1 and TDT2.**

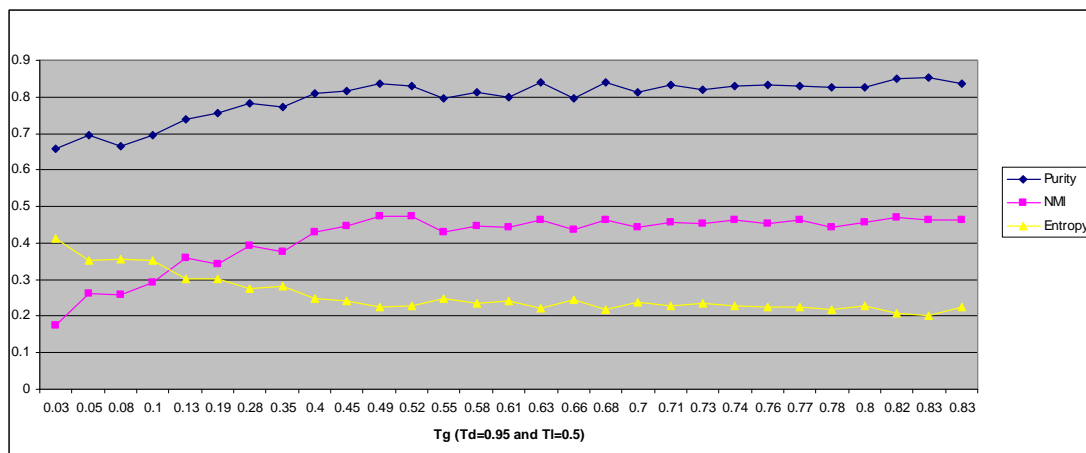
Corpus	Similarity	Top Terms	NMI
Reuters	TFIDF	ALL	0.257
Reuters	GCDSS	2600	<b>0.355</b>
TDT1	TFIDF	ALL	0.690
TDT1	GCDSS	1000	<b>0.755</b>
TDT2	TFIDF	ALL	0.517
TDT2	GCDSS	3000	<b>0.602</b>

In Table 4-17, NMI is significant ( $t(2)=8.613$ ,  $p<0.05$ ) across the three corpora. Because SOM produces different number of the clusters for the same corpus with the different feature subsets, only NMI is suitable to be used to compare the performance. The overall result indicates that the GCDSS measure significantly improves the performance of SOM compared to the baseline and the CDSS measure. And, the CDSS measure significantly enhances the performance of the basic K-means compared to the baseline and the GCDSS measure. However, there is no obvious difference between the baseline and GCDSS measure for the basic K-means algorithm, and there is no obvious difference between the baseline and CDSS measure for the SOM algorithm. That might be because the mechanisms of the two algorithms, K-means and SOM, are different. The GCDSS measure highlights the key terms and it is helpful for SOM to pick the documents as the winning nodes through competitive learning if the documents have stronger relationships with these key terms. The K-means averages the emphasis of these key terms on the centroid while the GCDSS does not have that effect.

#### **4.3.6 Tuning the Model Parameters**

Besides parameters in the clustering algorithms themselves, the parameters such as  $T_g$ ,  $T_l$ ,  $T_d$  and  $\beta$  in CDSS and GCDSS change the performance of the algorithms substantially.

The values of  $T_g$ ,  $T_l$  and  $T_d$  in LSISSM are between 0.0 and 1.0.  $T_d$  determines how many top dimensions are selected, which balances noise by eliminating more dimensions. The value range of  $T_d$  in our experiments varies from 0.5 to 0.98.  $T_d$  is often set as 0.95, which means the top  $K$  latent concept dimensions selected make 95% contribution to the LSI subspaces. The value range of  $T_l$  varies from 0.3 to 1.0. The initial value of  $T_l$  is often set as 0.5. If  $T_l$  and  $T_d$  are predefined, the upper boundary of  $T_g$  is determined and decides how many terms are included in the feature subset. Thus the performance of the clustering algorithms is changed with  $T_g$ . Actually, the results in Tables 4-1 to 4-9 demonstrate that the scope of  $T_g$  in which the clustering performance of the basic K-means and SOM are improved is very large. For instance, The Fig. 4-3 demonstrates the variation of the three measures, Purity, Entropy and NMI, with the values of  $T_g$  using the data in Tables 4-1. The trends shows that if  $T_g$  is larger than 0.5, the performance of the basic K-means is consistently higher than that of the VSM baseline.



**Figure 4-3: The Variation of the Evaluation Matrix of the Basic K-means with  $T_g$  Using the Reuters Corpus and CDSS.**

The pruning parameter  $\beta$  in CDSS or GCDSS affects the clustering performance by determining how many latent concept dimensions are used in the comparison of the two signatures. For instance, shifting the value of  $\beta$  from 0.0 to 0.5 and then to 1.0, the NMI scores for the basic K-means decreases from 0.407 to 0.389 and then increase to 0.455 with a term subset of 1800 top ranking terms and the Reuters corpus. Using the same term subset and the same corpus above with a  $\beta$  value of 0.5, if the  $T_d$  value that decreases from 0.95 to 0.75, the NMI score for the basic K-means drops from 0.389 to 0.370. Empirically the value of  $\beta$  is set as 0.0 comparing a term signature and a document signature. If matching two term signatures, the scope of  $\beta$  is 1.0 to 1.5. The effect of the term clusters on the performance of the basic K-means varies with the value of  $\beta$  while CDSS is used to calculate the similarity between the term signatures. For instance, if the value of  $\beta$  is shifted from 1.0 to 1.5, for the corpus TDT1 with the top 100 term clusters, the purity increases from 0.882 to 0.923.

LSISSM does not require strict parameter tuning and provides each parameter a large scope of values which ensure the significance of the clustering results.

#### 4.4 Conclusions for Text Clustering

CDSS significantly improves the effectiveness of the basic K-means compared to VSM and traditional LSI using top ranking terms or top ranking term clusters. With the top ranking terms, GCDSS significantly improves the effectiveness of SOM compared to those of VSM and traditional LSI. The experiments show that LSISSM consistently enhances the performance of the clustering algorithms until it reaches the maximum by increasing  $T_g$ . Our model decreases the dimensions at least 71.3% (2600 out of 9070). The running time of the algorithms drops from overnight to just a few hours.

The two-stage K-means improves the efficiency compared to the model-based basic K-means and speeds the clustering up to 5 times. Two-stage K-means solves the initialization problem of the basic K-means. Compared to the VSM baseline, the two-stage K-means algorithm runs faster in one or two orders of magnitude.

## CHAPTER5: Active Learning Using LSISSM

### 5.1 Introduction

The goal of this study is to develop an active learning method to improve the sampling selection process for a large unlabelled text corpus. We apply the document signature ranking method in the LSI Subspace Signature Model which selects the samples iteratively according to the proportion of the statistical contribution of the sample set to the overall LSI document subspace. The samples ranked in the top will be selected first for training.

### 5.2 Evaluation Methods, Text Preprocessing and Datasets

The classification procedure follows the three steps:

**Step1:** Select the training sets through either the LSI Subspace Signature ranking algorithm in Chapter 3.2 or by random sampling.

**Step2:** The learning curves of the classifiers are estimated through 10-fold cross validation (Breiman et al., 1992) to determine the optimal size of the training set for improving the categorization. 13 Learning Points are chosen to perform the 10-fold cross validation: {0.0, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0}. This sampling and testing procedure was repeated a total of 10 times for each fold.

**Step3:** In each learning point, the independent testing sets are tested in each 10-fold.

In addition, we considered three standard classification algorithms: Naïve Bayes (McCallum and Nigam, 1998), Rocchio (Ittner et al., 1995) and K-nearest neighbor

(KNN) (Cover and Hart, 1967) to determine if there is an interaction between out manipulation and the type of classifier. The learning accuracy for each learning point is measured by the average value of the classification accuracy in each fold. The classification accuracy is defined as the number of the document classified as “yes” divided by the size of the testing corpora. The overall classification accuracy for a classifier on a corpus is the average accuracy score across the 13 learning points.

Text pre-processing is the same as that in Section 4.2. We used the Reuters 21587, TDT1 and TDT2 collections for training. The training sample sets are the same as those in Section 4.2. Additionally, for each corpus, there is a sample set for independent testing. The test sets include 1031, 274 and 1790 news articles for Reuters, TDT1 and TDT2 respectively.

## 5.3 Experiment Results

### 5.3.1 Sampling Selection with Outlier Categories

We believe that the LSI subspace document signature ranking is an effective method for generating training samples. We test that across the three corpora and in those tests, we vary explore a number of parameters. After ranking by GLCR, the sampling distributions of the documents in the subset of Reuters news articles with 10 text categories are listed in Tables 5-1. R100 denotes the sample distribution in the full training set. RL25 denotes the sample subset selected by GLCR with  $T_g$  as 0.25,  $T_d$  as 0.95 and  $T_l$  as 0.5. RL5080 denotes the sample subset selected by GLCR with  $T_g$  as 0.50,  $T_d$  as 0.8 and  $T_l$  as 0.6. The results suggest that the variation of  $T_d$ ,  $T_g$  and  $T_l$  caused the change on the sample distribution of the selected subsets.

**Table 5-1: The Sampling Distribution across 10 Categories of the Reuters Corpus Selected by GLCR.**

Category\Subset	RL25	RL5080	R100
Acq	518	991	1646
Coffee	9	53	110
Interest	41	58	339
Iron-steel	7	14	40
Oat	1	1	8
Palmkernel	1	2	2
Sugar	10	29	125
Sun-meal	1	1	1
Veg-oil	7	29	87
Wheat	13	28	205

As a replication of the effect of inclusion of small categories found for the training samples in the Reuters collection, we considered the TDT1 and TDT2 corpus which have many more categories (topics).

**Table 5-2: The Sampling Distribution across 25 Categories of the TDT1 Corpus Selected by GLCR.**

Category\Subset	T1395	T150	T190	T1100
Aldrich Ames	5	6	8	8
Carlos the Jackal	4	4	8	10
Carter in Bosnia	8	15	32	34
Cessna on White House	6	9	13	14
Clinic Murders (Salvi)	18	23	37	41
Comet into Jupiter	15	21	43	45
Cuban riot in Panama	1	1	2	2
Death Kim Jong-il	11	15	46	58
DNA in OJ trial	56	72	105	114
Haiti ousts observers	1	3	9	12
Hall's copter (N. Korea)	36	44	80	97
Humble, TX, flooding	8	8	18	22
Justice-to-be Breyer	4	4	7	8
Karrigan/Harding	1	1	1	2
Kobe Japan quake	31	39	65	84
Lost in Iraq	15	22	43	44
NYC Subway bombing	9	10	17	24
OK-City bombing	87	108	222	273
Pentium chip flaw	3	3	4	4
Quayle lung clot	6	7	11	12
Serbians down F-16	15	23	59	65
Serbs violate Bihac	18	25	77	91
Shannon Faulker	5	5	7	7
USAir 427 crash	7	10	33	39
WTC Bombing trial	8	10	21	22

T1100 in Table 5-2 denotes the full training set of TDT1. T1395 in Table 5-2 denotes the sample subset selected by GLCR with  $T_g$  as 0.395 (378 out of 1131),  $T_d$  as 0.65 and  $T_l$  as 0.3. T150 in Table 5-2 denotes the sample subset selected by GLCR with  $T_g$

as 0.50 (488 out of 1131),  $T_d$  as 0.65 and  $T_l$  as 0.3. T190 in Table 5-2 denotes the sample subset selected by GLCR with  $T_g$  as 0.90 (968 out of 1131),  $T_d$  as 0.65 and  $T_l$  as 0.3.

**Table 5-3: The Sampling Distribution across 30 Categories of the TDT2 Corpus Selected by GLCR.**

Category\Subset	T230	T238	T2100
20001	194	322	1132
20005	25	29	41
20010	2	3	7
20012	9	26	151
20013	206	281	540
20014	1	1	2
20017	16	17	20
20019	52	55	110
20022	30	30	30
20023	34	46	125
20025	1	1	1
20026	66	67	70
20027	1	1	1
20030	2	2	2
20036	5	5	5
20040	6	6	6
20044	29	47	280
20047	63	71	93
20050	11	11	11
20060	8	8	8
20065	55	55	60
20068	8	8	8
20071	36	40	203
20074	22	28	50
20076	19	44	324
20082	3	3	4
20084	1	1	5
20087	58	61	79
20092	3	3	3
20098	9	9	9

In Table 5-3, T2100 denotes the full training set of TDT2. T230 denotes the sample subset selected by GLCR with  $T_g$  as 0.307,  $T_d$  as 0.73 and  $T_l$  as 0.6. T238 denotes the sample subset selected by GLCR with  $T_g$  as 0.384,  $T_d$  as 0.95 and  $T_l$  as 0.6. The  $T_g$  values of T230 to T238 shift from 0.307 to 0.384. Those values reflect the lower boundaries for GLCR to pick a subset which includes every category in the corpora if  $T_d$  varies from 0.73 to 0.95 and  $T_l$  is set as at 0.6. In another word, if the value of  $T_g$  is set higher than these values, GLCR generates a subset with all the categories.

If a subset is selected randomly, that subset probably will not have good coverage on every category. For instance, we randomly select 608 documents for 5 times which have the same size as the set RL25. But the data set excludes at least the sample for the category “sun-meal” that has only one sample in the training set. Conversely, GLCR includes it consistently if  $T_g$  is no less than 0.25 with  $T_d$  as 0.95 and  $T_l$  as 0.5.

If randomly selected 378 documents which have the same size as T1395, 2 categories with the smallest sample size, “Karrigan/Harding” and “Cuban riot in Panama”, are missed out of 25. If randomly selected 488 documents which have the same size as T150, the 2 categories with the smallest sample size are missed out of 25 based on the random seed occasionally.

If randomly selected 975 and 1226 documents which have the same size as T230 and T238, 9 and 4 categories are missed out of 30 respectively. Because so many categories are missed, we did not study learning curves of the random selection of TDT2.

The thresholds,  $T_d$ ,  $T_l$  and  $T_g$  in the sampling selection algorithm affect the scope of the sampling selection. If  $T_d$  or  $T_g$  increases or  $T_l$  decreases, the scope of sampling is expanded. GLCR provides a larger scope to select sampling sets without losing any category. For instance, GLCR finds all the categories with  $T_g$  between 0.3 and 0.384 (T238),  $T_d$  as 0.95 and  $T_l$  as 0.6.

The sample distribution shows that the ranking strategy GLCR favors the majority categories. For instance, in T230 to T238, the samples for the category “20013” are picked with about 38% to 52%. The most striking result replicated in the three corpora from the manipulation of the model is that for these outlier categories which have a few samples and these samples have high projection scores on one or more LSI latent concept dimensions, those categories are included in the training samples generated by our method but are excluded by randomly generated sub-sets of training examples.

### ***5.3.2 The Effect of the Model-based Dimension Reduction on the Text Classifiers***

Feature reduction can improve the performance of the classifiers effectively and efficiently. In this study GLCR ranking is applied on the term signatures.  $T_d$  is set as 0.95 and  $T_l$  is set at 0.5. The three datasets listed in experiment section are utilized the GLCR ranking. For the Reuters dataset, GLCR selects a subset of 2673 concept terms out of the whole set (9070 terms), which makes the 83.3% ( $T_g$ ) statistical contribution to the trucked LSI term subspace. For TDT1, GLCR selects a subset of 1614 concept terms out of the whole set (8338 terms), which makes the 82.30% ( $T_g$ ) statistical contribution to the trucked LSI term subspace. For the TDT2, GLCR selects a subset

of 3007 concept terms out of the whole set (17083 terms), which makes the 85.14% ( $T_g$ ) statistical contribution to the trucked LSI term subspace.

**Table 5-4: The Average Learning Accuracy of the Three Classifiers, Naïve Bayes, KNN and Rocchio, on the Full Training Sets of the Reuters Corpus, TDT1 and TDT2 Using Feature Reduction.**

Top Terms	Dataset	Naïve Bayes	KNN	Rocchio
ALL	R100	.873	.832	.914
2673	R100	<b>.874</b>	<b>.846</b>	.913
1847	R100	<b>.875</b>	<b>.848</b>	.913
1000	R100	<b>.874</b>	<b>.851</b>	.912
ALL	T1100	.866	.810	.753
1614	T1100	<b>.874</b>	.808	.751
1579	T1100	<b>.875</b>	.808	.751
900	T1100	<b>.877</b>	.808	.750
ALL	T2100	.870	.858	.896
3007	T2100	<b>.877</b>	.857	.894
2528	T2100	<b>.877</b>	.856	.894
1600	T2100	<b>.879</b>	.856	.892

**Table 5-5: The Average Learning Accuracy of the Three Classifiers, Naïve Bayes, KNN and Rocchio, on the Independent Testing Sets of the Reuters Corpus, TDT1 and TDT2 Using Feature Reduction.**

Top Terms	Dataset	Naïve Bayes	KNN	Rocchio
ALL	R100	.875	.822	.910
2673	R100	<b>.877</b>	<b>.842</b>	<b>.911</b>
1847	R100	<b>.877</b>	<b>.843</b>	.909
1000	R100	.874	<b>.851</b>	.905
ALL	T1100	.768	.728	.710
1614	T1100	<b>.794</b>	.727	.708
1579	T1100	<b>.793</b>	.726	.708
900	T1100	<b>.786</b>	.723	.700
ALL	T2100	.787	.762	.820
3007	T2100	<b>.789</b>	<b>.764</b>	.816
2528	T2100	.787	<b>.764</b>	.816
1600	T2100	.787	<b>.764</b>	.815

In Tables 5-4 and 5-5, ALL denotes all features and the baseline runs. And bold numbers denote the runs with higher scores than those of the baseline runs. The first columns of these tables denote the size of the feature subsets used in the training documents. The values in the columns 3, 4 and 5 denote the average accuracy of the classifier across the 13 learning points. Using these feature subsets, we study their effects on the learning curve performance of the three classifiers on the three full training sets (see table 5-4) and the independent test performance with the full training sets (see table 5-5). The results indicate in either test the feature subsets consistently improve the performance of the classifier, Naïve Bayes. With the feature reduction, Rocchio and KNN achieve equal or only slightly better results.

### 5.3.3 *The Effect of the Model-based Sample Reduction on the Learning Curves of the Text Classifiers*

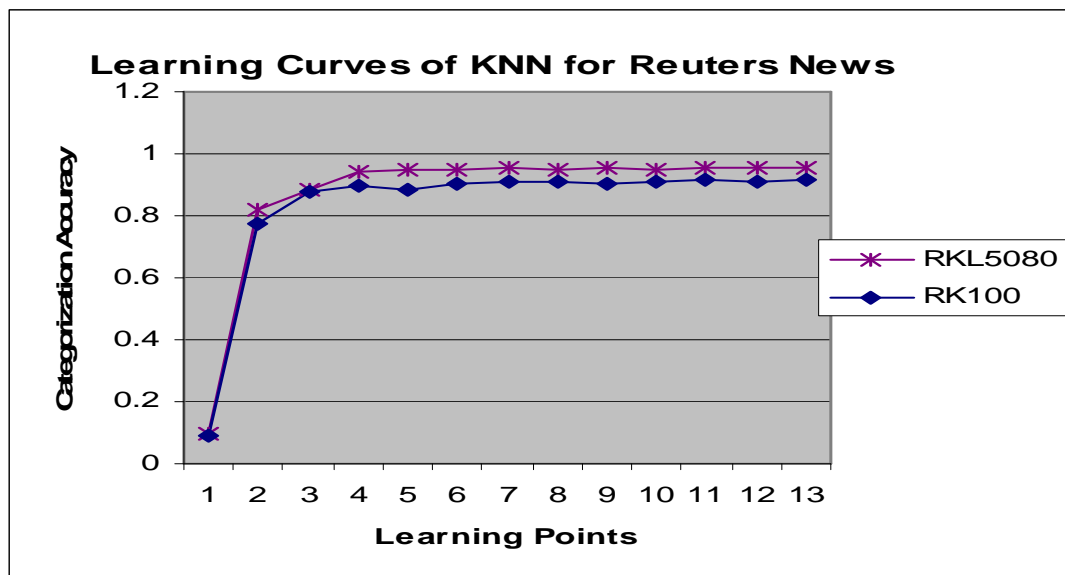
Indeed, we repeatedly found that the training samples produced better performance on the learning curves than those of the full training sets across corpora and classifiers.

**Table 5-6: Comparison of the Average Learning Accuracy between the Full Training Set and the Training Subsets of the Reuter Corpus Using the Three Classifiers, Naïve Bayes, KNN and Rocchio.**

Dataset	Naïve Bayes	KNN	Rocchio
R100	.873	.832	.914
RL25	.869	<b>.847</b>	<b>.924</b>
RL8050	<b>.893</b>	<b>.872</b>	<b>.945</b>

For example, in Table 5-6, the average accuracy of the three classifiers with the two datasets, RL25, RL5080 and R100, across the 13 learning points is listed. R100 denotes the full training set and its learning accuracy is the baseline. The results

indicate that the average learning accuracy of the three classifiers with the subset RL8050 is better than that of the full dataset, see Fig. 5-1, 5-2 and 5-3.



**Figure 5-1: Comparison of the Learning Curves between the Subset RL5080 and the Full Training Set R100 Using the KNN Classifier.**

In Fig. 5-1, RKL5080 denotes the learning curve generated for the subset RL5080; RK100 denotes the learning curve generated for the full training set R100. RKL5080 outperforms RK100 in every learning point.

In Fig. 5-2, RNL5080 denotes the learning curve generated for the subset RL5080; RN100 denotes the learning curve generated for the full training set R100. RNL5080 outperforms RN100 at every learning point.

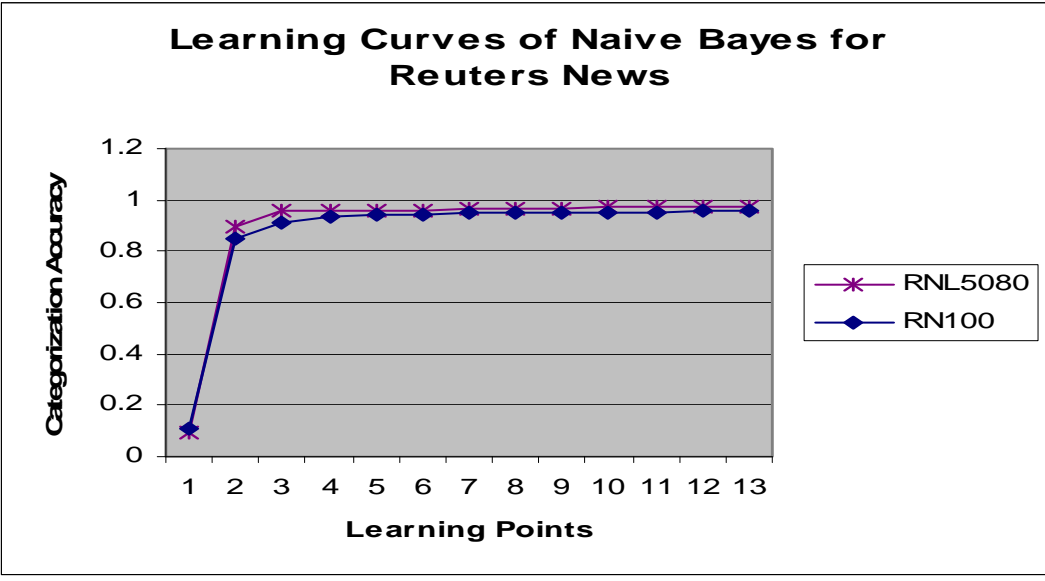


Figure 5-2: Comparison of the Learning Curves between the Subset RL5080 and the Full Training Set R100 Using the Naïve Bayes Classifier.

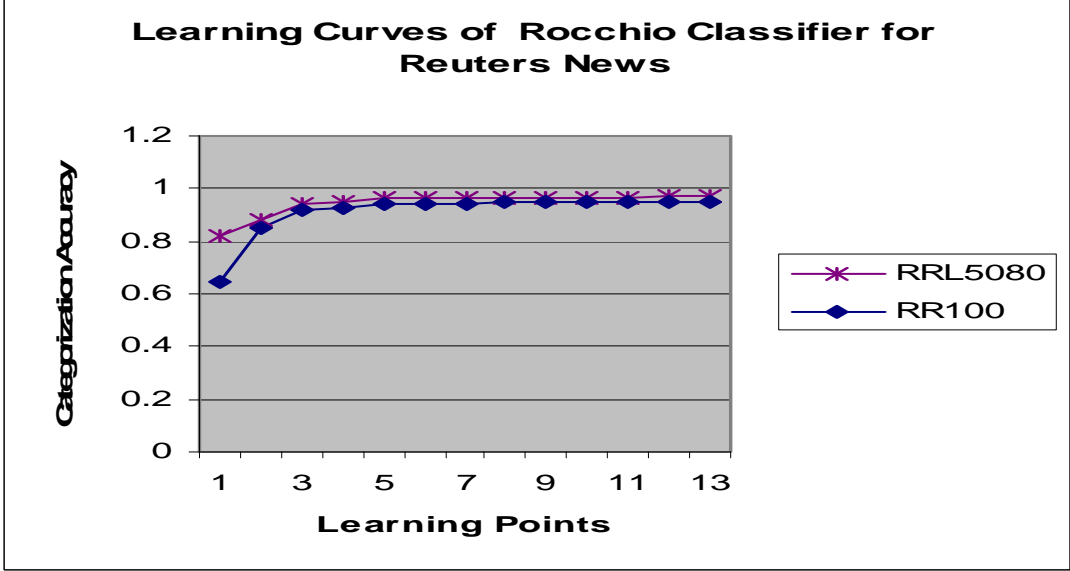


Figure 5-3: Comparison of the Learning Curves between the Subset RL5080 and the Full Training Set R100 Using the Rocchio Classifier.

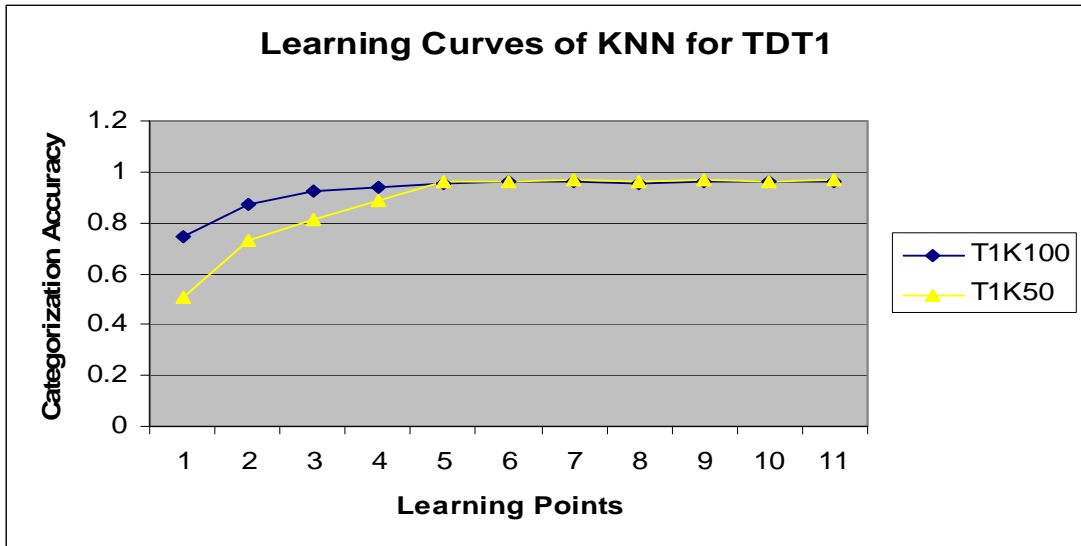
In Fig. 5-3, RRL5080 denotes the learning curve generated for the subset RL5080; RR100 denotes the learning curve generated for the full training set R100. RRL5080 outperforms RR100 at every learning point.

For TDT1 and TDT2, the average learning accuracy of the selected subsets is a little lower than those of the full training sets, see Table 5-7.

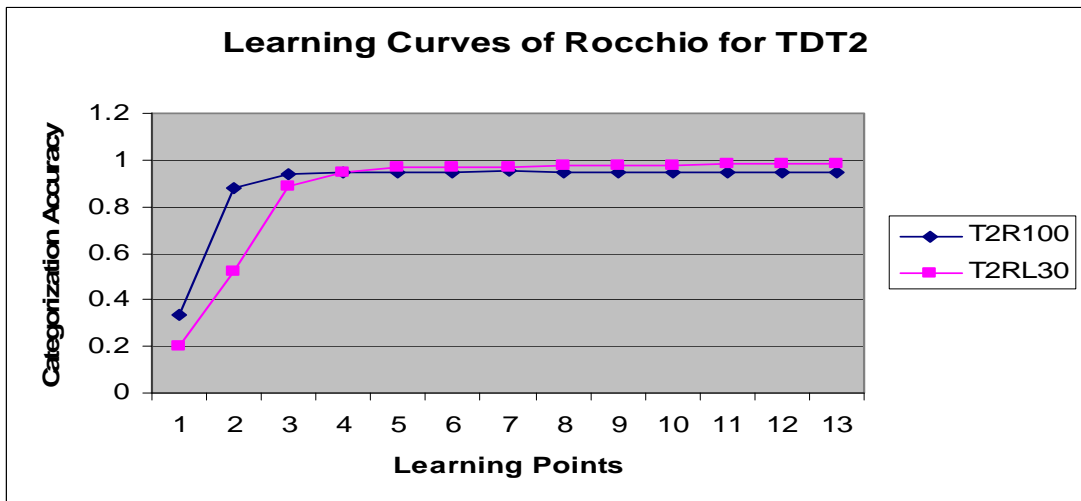
**Table 5-7: Comparison of the Average Learning Accuracy between the Full Training Sets and the Training Subsets of TDT1 and TDT2 Using the Three Classifiers, Naïve Bayes, KNN and Rocchio.**

<b>Dataset</b>	<b>Naïve Bayes</b>	<b>KNN</b>	<b>Rocchio</b>
T1100	.866	.810	.753
T150	.832	.758	-----
T190	.862	.806	.752
T2100	.870	.858	.896
T230	.848	.810	.872
T238	.858	.828	.887

However in many cases the subsets have better scores on most of the learning points, see Fig. 5-4 and 5-5. T1K50 in Fig. 5-4 denotes the learning curve generated for the subset T150; T1K100 denotes the learning curve generated for the full training set T1100. T1K50 outperforms T1K100 beyond learning point 7. T2RL30 in Fig. 5-5 denotes the learning curve for subset T230; T2RL100 denotes the learning curve for the full training set T2100. T2RL30 has higher learning accuracy than that of T2100 beyond learning point 5.



**Figure 5-4: Comparison of the Learning Curves between the Subset T150 and the Full Training Set T1100 Using the KNN Classifier.**



**Figure 5-5: Comparison of the Learning Curves between the Subset T230 and the Full Training Set T2100 Using the Rocchio Classifier.**

### 5.3.4 Combined Effect of the Model-based Dimension Reduction and Sample Reduction on the Text Classifiers

**Table 5-8: Comparison of the Average Learning Accuracy of the Three Classifiers, Naïve Bayes, KNN and Rocchio, Trained by the Training Subsets of the Reuter Corpus, TDT1 and TDT2 with Feature Reduction and Tested by the Independent Testing Sets.**

Top Terms	Dataset	Naïve Bayes	KNN	Rocchio
ALL	RL25	.806	.772	.839
2673	RL25	<b>.826</b>	<b>.789</b>	<b>.861</b>
ALL	RL8050	.836	.802	.879
2673	RL8050	<b>.846</b>	<b>.813</b>	<b>.885</b>
ALL	T150	.719	.695	-----
1614	T150	<b>.735</b>	.692	-----
ALL	T190	.769	.727	.712
1614	T190	<b>.800</b>	.724	.709
ALL	T230	.726	.702	.775
3007	T230	<b>.738</b>	<b>.704</b>	<b>.778</b>
ALL	T238	.770	.740	.801
3007	T238	<b>.774</b>	<b>.743</b>	.798

The results in Table 5-8 indicate that the performance of the subsets RL25, RL8050, T150, T230 and T238 on the independent tests are not good as those of the full training sets, R100, T1100 and T2100 (see Table 5-5), but they are comparable. Moreover, feature deduction shirks the difference because in most of the cases, the feature subsets improve the performance of the three classifiers on the sample subsets. Naïve Bayes and Rocchio classifier have a better performance on T190 than that of T1100. This instance reflects the ideal scenario of this research that a sample subset (968 out of 1131) with a small feature set (1614 out of 8338) trained by a classifier achieves a better performance compared to that of the full training set with a full feature set.

Apparently, the feature deduction approach is not only effective in the sense of classification accuracy, but it can actually improve classification processes efficiently. For instance, with the KNN classifier and the dataset R100, the total training time decreases from 5.71s to 2.72s and the training time per sample decreases from 0.71ms to 0.66ms. Obviously the sample selection suggests an optimized sample candidate subset for human labeling and could dramatically decrease the time to make a training corpus with class labels.

## 5.4 Conclusions for Active Learning

The LSI document signature ranking algorithm in the model picks the most important samples and features and keeps the sampling distribution on the text categories, even outlier categories. There is no need for our method to know the labels of these categories. Compared to randomly selected subsets within the procedure of the n-fold cross validation, for many cases our approach has better performance in the learning curves, especially when the size of the subset is much smaller than that of the full training set. That is because the ranking algorithm selects the samples following the order of their importance concern both the global and local statistical contribution of the samples/features to the LSI subspaces. The samples/features with the most important statistical contribution to the corpus will be included in the subset first. Each sample/ feature picked is the statistical representative of the corpus. So, the average contribution for each sample in the subset generated by our approach is higher than the average. Even compared to the independent testing of the full training set, the performance from the subsets does not decrease always and even out-performs in some cases.

Our approach is particularly useful for finding efficient feature/sample sets for a large unlabeled corpus. The feature subsets have been applied and successfully enhance the performance of the clustering algorithms from the perspective of unsupervised classification. And the results from this study indicate that the subsets with a small feature subset improve and stabilize the performance of the classifiers without affecting on the sample distribution to the text categories. The effect of the LSI

subspace term signature ranking on the unsupervised classification and supervised classification are consistent and identical. Our studies follow the ideal scenario to build a process for automatic text categorization. First, a training corpus is generated by using the LSI subspace document signature ranking algorithm and then the training corpus is labeled by user. Finally, machine learning methods are applied to classify all the unlabeled documents by the training examples. Hopefully, the process maximizes the effective of text categorization and will minimize the cost of generating of training sets.

## **CHAPTER6: Query Expansion Using Domain Ontologies**

### **6.1 Term Re-weighting Strategy with UMLS Co-concepts and Synonyms<sup>7</sup>**

Many search results contain a large number of irrelevant results or may contain only some of the aspects of topics requested by the users. In many cases, novice users simply do not know how to construct efficient and effective queries. Even experienced users do not always create efficient and effective queries when searching an unknown domain. Query expansion helps users solve this problem. Query expansion can add critical terms beyond original query terms to improve the precision and/or recall. A search tool with embedded query expansion add terms to the original query automatically or provide high-level information about the collections to the users and suggest the user to refine the original query. In this research, three query expansion strategies are compared. The three methods are local analysis, global analysis, and ontology-based term re-weighting - integrated with the UMLS<sup>8</sup> (Unified Medical Language System) are compared. These methods are applied to the Ad Hoc Retrieval task of the TREC 2004 Genomics task.

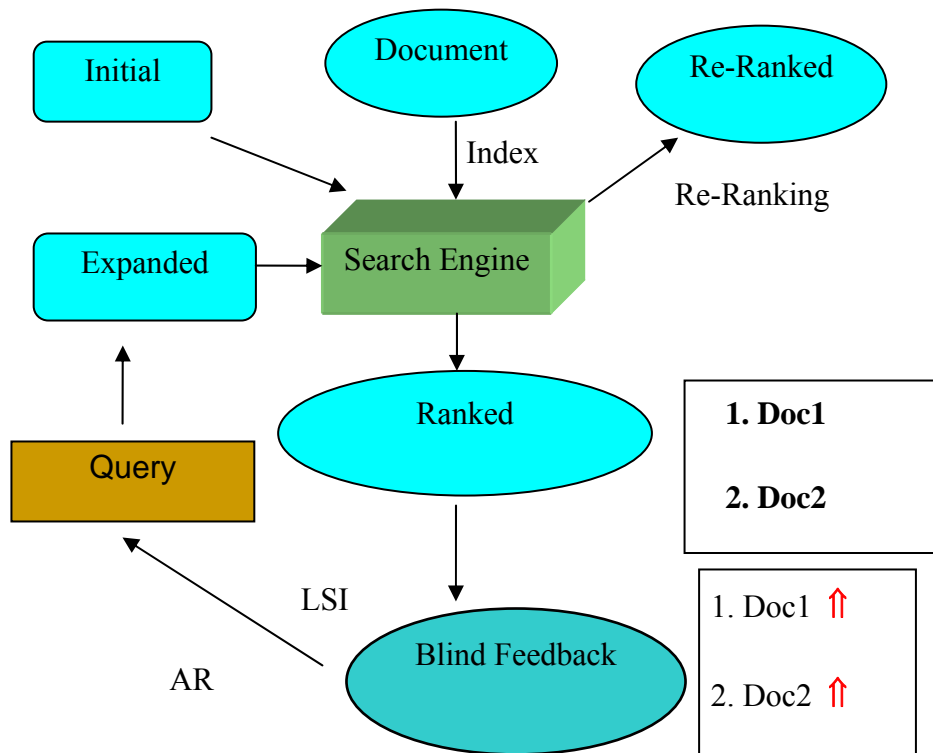
---

<sup>7</sup> The work in this section was published in Zhu et al., 2006

<sup>8</sup> <http://www.nlm.nih.gov/research/umls/>

### 6.1.1 Query Expansion Strategies

Local Analysis determines the expanded terms based on pseudo-relevance feedback. By examining the top N documents retrieved from the initial query, the traditional LSI (Deerwester et al., 1990) and Association Rule (AR) (Agrawal et al., 1995) are used to seek the top co-terms of the original terms from the top N retrieved documents. Co-terms are weighted according to the total term frequency in the top N retrieved documents. The processes of local analysis are depicted in Fig. 6-1:



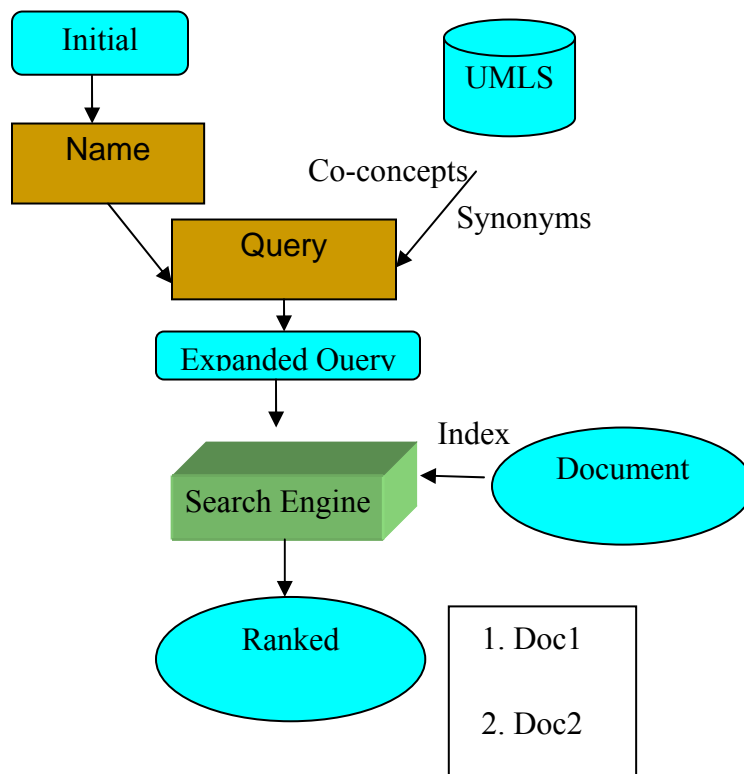
**Figure 6-1: Query Expansion Procedures of Local Analysis.**

The steps to expand the query terms based on LSI and AR algorithms are the following:

**Step 1:** Derive a term matrix based on the top N (N=300) retrieved documents for one initial query returned by the search engine.

**Step2:** Run LSI or AR algorithms to acquire the terms related to the original query terms. We choose the top M (M=5) co-terms of the original terms from the top N (N=300) retrieved documents.

**Step3:** Expand the query terms including the related terms acquired above.



**Figure 6-2: Query Expansion Procedures of Global Analysis and Term Re-weighting Strategies.**

The processes of global analysis and term re-weighting are depicted in Fig. 6-2.

In global analysis, terms to be added are extracted from all the documents of the whole collection. The initial query will be expanded by UMLS co-concepts of the original key terms with the same semantic types. UMLS provides co-concepts and related co-occurred frequencies for many of the medical terms appeared in MEDLINE during the past years. For instance, term “transgenic mice” in a query has a most frequent co-concept “mice, knockout” in UMLS with the same semantic type. Term “mice, knockout” will be added to the initial query. The key term “transgenic mice” in the initial query is extracted by LingPipe.

Term re-weighting enhances the weights of key original terms or co-terms according to their relative importance in queries. Specifically, we employed two principles for determining the weights. The first is that if an original term has a higher term frequency in the initial query with a specific major UMLS semantic type, the term should be given a higher weight in the expanded query. The second one is that if a key original term or an expanded term has a pre-selected major UMLS semantic type, its preferred MeSH synonym defined in UMLS is expanded and given a higher weight. MeSH<sup>9</sup> (Medical Subject Headings) is the U.S. National Library of Medicine's controlled vocabulary used for indexing articles for MEDLINE/PubMed. MeSH terminology provides a consistent way to retrieve information that may use different terminology for the same concepts. The selection of key original terms is decided by the tagging of LingPipe. In the TREC 2004 Genomics Ad Hoc Retrieval

---

<sup>9</sup> <http://www.nlm.nih.gov/mesh/meshhome.html>

Task, most topics discuss the functionality of certain genes or proteins. So, the major UMLS semantic type for these topics is selected as “Amino Acid, Peptide”, or “Protein”. For instance, protein “NEIL1” is located in the initial query and we used a boost score of 4. Moreover its preferred MeSH term “NEIL1 protein human” extracted from UMLS is expanded and a boost score of 8 is assigned. Generally, the preferred MeSH terms in UMLS are used to index the MEDLINE abstracts. Thus, it is reasonable to give them a higher weight.

### **6.1.2 Evaluation Methods and Datasets**

The goal of TREC 2004 Genomics Ad Hoc Retrieval Task was to find all the relevant documents to the 50 topics from the whole corpus. The structure of this task was a conventional searching task based on a 10-year subset of MEDLINE (about 4.5 million documents and 20 gigabytes in size with NLM XML format) and 50 topics derived from information needs obtained via interviews of biomedical researchers. Recall and precision for the ad hoc retrieval task were calculated in the classic IR way, using the preferred TREC statistics of mean average precision (average precision at each point a relevant document is retrieved, also called MAP).

### **6.1.3 Experiment Results**

We studied the three strategies (Zhu et al., 2006) for query expansion using two search engines, Lucene<sup>10</sup> and Lemur<sup>11</sup> to ensure the generality of our findings. With

---

<sup>10</sup> <http://lucene.apache.org/>

<sup>11</sup> <http://www.lemurproject.org/>

Lemur, TF-IDF based Vector Space Model and Okapi BM25 Model were selected to verify the term re-weighting approach. Ranking algorithms for both of the models are calculated by Okapi term frequency (TF) formula (Jones et al., 2000). Based on Okapi query TF formula, key terms or phrases in queries are re-weighted simply by increasing their query term frequency according to assigned boost scores. Lucene search engine supports customized term boost. The boost factor is a part of the Lucene rank algorithm.

Co-term expansion with local analysis increases average precision by only 0.5% (LSI). AR gives an even worse result and decreases the average precision by 6.0%. Here, AR mining may not be efficient. Instead, the association rules at sentence level may produce more precise term associations.

The term re-weighting strategy is applied to two runs on Lucene, Baseline and Baseline+Context (title + key terms or phrases in context of query). The results show this approach increases average precision in both conditions by 4.5% and 20.3%. Apparently the term re-weighting approach improves the average precision of the Baseline+Context than that of Baseline run. This implies context information of queries is critical to enhance the performance of retrieval. If the Baseline+Context run is treated as a baseline, the term re-weighting approach will eventually improve the retrieval performance by 16.2%. The term re-weighting strategy is also applied to four runs on Lemur which queries were both formed from Baseline and Baseline+Context. But, these four runs are based on two different information retrieval models, Vector Space Model and Okapi BM25 Model. The results indicate

that Okapi Model that increases average precision by 12.1% could more empower the term re-weighting approach than the Vector Model that increase average precision by 7.5% with context taken into account. Without context, the term-reweighting approach on the Okapi Model only increases average precision by 4.0% and the performance enhancement on Vector Space Model is about 2.2%. If the Baseline+Context runs are treated as baselines, through term re-weighting, The Okapi Model could elevate average precision by 8.6%, but the Vector Space Model could only enhance average precision only by 0.7%.

Co-concepts expanded from global analysis make the results worse. The average precision of the UMLS+Global run on Lemur decreases by 0.6% (Vector Space Model) and 14.9% (Okapi BM25 Model) compared to Baseline run and the one on Lucene decreases by 22.7% (Vector Space Model). Perhaps, the top-ranking co-concepts from UMLS co-occur frequently with original terms in certain context that is totally different from that of TREC Genomics topics.

The average precision of our baseline runs are better than the results of the 2004 TREC Genomics Ad Hoc task where the average precision of 47 runs submitted by 32 teams was 20.74%. Our best runs are 33.74% and 28.91% for Lemur and Lucene, respectively. The best run from Lemur outperforms the top 6<sup>th</sup> run (33.24%) in the contest of 2004 TREC Genomics.

#### ***6.1.4 Conclusions for Query Expansion Strategies***

We compared and explored three query expansion strategies for bio-medical domain. The term re-weighting strategy showed great potential to improve precision and recall

across different search engines and information retrieval models. We explicitly showed how to find most important terms in the queries with the IE techniques and how to apply domain ontology to extend these terms. This strategy could be utilized in other bio-medical information retrieval systems. The architecture of the system could be applied any other domain besides the biology and medicine.

## **6.2 User Relevance Feedback Expanded by the IPTC Hierarchical Structure**

### **6.2.1 Introduction**

Hundreds of thousands, perhaps millions, of pages of historical newspapers are now being digitized each year (Murray, 2005). In most cases, this digitization from preservation quality microfilm. While this digitization is probably every state of the US and in many countries around the world, the largest project is the NDNP (National Digital Newspaper Program), a project of the National Endowment of the Humanities and the Library of Congress (LC). In first year of Phase I of NDNP, six state libraries have been digitizing newspapers in the timeframe 1900 to 1910. For NDNP, the TIF image files from the digitization are also saved as JPEG2000 and as PDF. In addition, they are processed with OCR. The OCR files were coded in XML with the METS ALTO format.

### **6.2.2 User Relevance Feedback Expanded by Hyponyms in IPTC Codes**

The International Press Telecommunication Council (IPTC)<sup>12</sup> has developed subject codes for classifying news stories (Dhillon et al., 2002). Altogether, there are about 1000 categories. The IPTC categories include some modern terms which are not relevant to the historical context, for instance, “nanotechnology”. Nonetheless, the set also provides reasonable coverage for the historical newspapers. While it might eventually be desirable to refine, the IPTC with some additional categories specific to

---

<sup>12</sup> <http://www.iptc.org/>

historical eras, there is also an advantage to using a system which spans both historical and current newspapers. The IPTC subject codes are organized into a hierarchy. There are ten top-level labels and one or, sometimes, two levels below that. For instance, the top-level of “Crime, Law, and Justice” (CL&J) has “Crime” as a second-level category and types of Crime (e.g., “homicide”) as third-level categories.

In this study firstly we classify the articles with the top level categories in the baseline. Then, all hyponyms, the second-level children of the top IPTC labels, are selected to add to the query with an equal weight that is the same as the weight of the original query. For instance, CL&J is extended by “crime, judiciary (system of justice), police, punishment, prison, laws, justice and rights, trials, prosecution, organized crime, international law, corporate crime, war crime, inquest, inquiry, tribunal”. Finally, based on the ratings of the students, we do User Relevance Feedback (URF) by applying the well-known, Rocchio formula (Rocchio, 1971).

$$q_{i+1} = \alpha q_i + \frac{\beta}{|D_r|} \sum_{d_j \in D_r} d_j - \frac{\gamma}{|D_n|} \sum_{d_j \in D_n} d_j \dots\dots\dots(7.1)$$

for expanded query  $q_{i+1}$

$q_i$ : the initial query

$D_r$ : a set of relevant documents among retrieved documents

$D_n$ : a set of non-relevant documents among retrieved documents

$\alpha, \beta, \gamma$ : tuning constants

In this model the information in relevant documents is treated more as more important than the information in non-relevant documents ( $\gamma < \beta$ ). For pseudo-

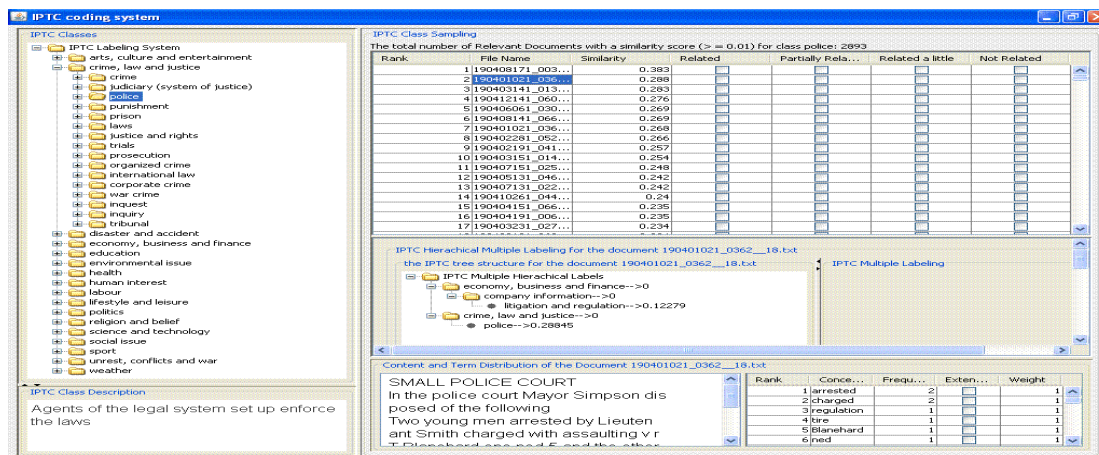
relevance feedback the top N retrieved documents are supposed to be all relevant:  $\gamma=0$ .

### **6.2.3 Evaluation Methods, OCR Preprocessing and Datasets**

The label for each IPTC code is treated as a query and the relevant documents ranked in the top are selected for training. For CL&J, the description was:

“Establishment and/or statement of the rules of behavior in society, the enforcement of these rules, breaches of the rules and the punishment of offenders. Organizations and bodies involved in these activities.”

We used these simple descriptions to help the trainer label the documents. We picked two top-level IPTC categories, “Crime, Justice, and Law” (CL&J) and “Disasters and Accidents” (D&A) on which to focus. We had two students complete make ratings of articles on these categories across the various conditions of the research as described below. The Spearman Correlation coefficient between the two raters is 0.574. The level of the coherence is average because the quality of the OCR texts might affect the judgments of the raters. For each of the two test categories, we then took the top 200 ranked articles which make sure that there are enough negative examples for user relevance feedback. We had the student raters use the interface in Figure 6-3 to indicate their judgments of the relatedness of articles.



**Figure 6-3: The Interface Allows Students to Make Relevance Ratings about the IPTC Categories for the OCR Text of the Historical Newspaper.**

Sample NDNF image files and OCR were obtained for pages from the *Washington Times* for the years 1900 to 1910. The OCR of this text varies greatly in quality from year to year and across newspapers. Thus, we improved the readability of the text simply by passing it through a standard word-processing spell correction program first. The OCR files do not identify article regions. An LSI-based automatic linear segmentation technique (Choi, 2000) was applied to divide the OCRd text extracted from the images of the newspaper into blocks. Then the edges of the blocks are further verified by the closest lines of capital letters which might indicate the titles. In our analysis each text block is treated as a separate article. For 1904 there were 50979 text blocks in the corpus. Among them, 50788 text blocks have relationships to at least one of the 1354 IPTC subject codes and a threshold of 0.01.

#### 6.2.4 Experiment Results

Baseline denotes the basic Vector Space Model. Baseline+URF denotes the user relevance feedback on the baseline. Baseline+Hier denotes the query expansion described in Section 6.3.2. Baseline+Hier+URF denotes the user relevance feedback on the expanded baseline with the children of the original query. P@200 denotes the top 200 retrieved documents. The results are shown in Table 6-1.

**Table 6-1: Comparison of P@200 between the VSM baseline and the User Relevance Feedback with the Hyponyms in IPTC Codes.**

<b>Runs</b>	<b>CL&amp;J</b>	<b>D&amp;A</b>
Baseline	0.73	0.77
Baseline+Hier	0.68	0.72
Baseline+URF	0.88	0.85
Baseline+Hier+URF	0.90	0.85

It can be seen that Baseline+URF and Baseline+Hier+URF are substantially better than the other two approaches. For CL&J, Baseline+Hier+URF achieves the best performance, 23% improvement compared to Baseline. Baseline+URF achieves 21% enhancement. For D&A, Baseline+Hier+URF and Baseline+URF both obtain the 10% improvement compared to Baseline. So in average, the best overall improvement on the P@200 across the two categories is 16.5%.

### **6.2.5 Conclusion for the IPTC Study**

Even with difficulties in the OCR, relatively simple methods allow pretty good categorization of the articles. The baseline has an average 75% precision on the top 200 retrieved documents. Furthermore, user relevance feedback extended with hyponyms in the IPTC hierarchal structure substantially improves the performance by 16.5% on average across two categories.

## CHAPTER7: Weighted PageRank Enhanced by Betweenness

### Centrality<sup>13</sup>

#### 7.1 Introduction

Identification of key players is one of the major challenges to SNA. There are four centrality measures to rank players' influence, i.e., Degree Centrality, Closeness (Sabidussi, 1966), Betweenness Centrality (Freeman, 1979), and PageRank (Page et al., 1998), assume that influence propagates via restricted paths. We compare and evaluate all these measures of influence of members in an enterprise according to their historical email conversations. First, we compare the results of these measures and identify the inter-relationships between them. Then we integrate two of them to improve the performance because the correlation relationships among these measures are statistically significant. We show that PageRank enhanced by time-sensitive Betweenness improves influence ranking by solving the rank sink problem of PageRank.

The original PageRank assumes that prestige is equally distributed across all the links of a Web page. In a social network, however, not all edges are equal; some actors interact more often and/or more profoundly with others. In this context, the PageRank equation should take into account weighted communication links and to what extent they should transfer PageRank values. In our weighted PageRank equation, a

---

<sup>13</sup> This study was published in Zhu et al. 2008

propagation proportion is defined as  $w(a_j, a_i)$  between actors  $a_i$  and  $a_j$  by normalizing the link weights emanating from a particular actor  $a_j$  as follows:

$$w(a_j, a_i) = W(a_j, a_i) / \sum_k W(a_j, a_k) \dots\dots\dots(7.1)$$

For any particular actor  $a_j$ ,  $w(a_j, a_i)$  is defined as the ratio between the number of email conversations between  $a_i$  and  $a_j$  to the number of all the email conversations of  $a_j$ . Therefore it can be used to determine the fraction of an actor's PageRank that transfers to other linked actors.

The PageRank algorithm computes the importance scores of Web pages through a stochastic irreducible Markov transition matrix that is constructed from all hyperlinks between Web pages. Directly defining the matrix as a normalized adjacency matrix of the Web graph always produces a sparse and reducible matrix, yielding the "rank sink" problem. To solve this problem, Brin and Page (1998) introduce a uniform matrix and linearly interpolates it with the normalized adjacency matrix with a fixed random jumping probability  $\beta$ . A surfer would be better off following the out links of a high-quality hub page rather than a low-quality one. This motivates us to think that a dynamic  $\beta$  value based on the page properties can be a better choice. If PageRank is used in social network analysis, the parameter  $\beta$  should reflect the communication properties of an actor. Interestingly, we observe that Betweenness centrality  $B(i)$  can be defined as the average probabilities across all possible pairs of nodes that the shortest path between any two nodes will pass through the given node  $i$  (Freeman, 1997). The Betweenness score could be seen as the average probability that any other

node goes through the selected node. Driven by this definition, we hypothesize that the score of Betweenness could dynamically be used as the value of the parameter  $(1-\beta)$ , and model a PageRank Markov matrix more accurately. This assignment assumes any pair of nodes in the network communicates through shortest paths. We use this approach to extend our weighted PageRank to rank the actors in a social network. Then the Weighted PageRank for an actor  $a_i$  is defined as follows:

$$PR_w(a_i) = (1 - \lambda) / N + \lambda \sum_j PR_w(a_j) * w(a_j, a_i) \dots\dots\dots(7.2)$$

According to Eq. 7.2, the transfer of prestige from one actor to the other is modulated by the propagation proportion  $w(a_j, a_i)$ . The parameter  $\lambda$ , which equals  $B(a_i)$ , represents the attenuation of prestige values as they are transferred from one actor to the other.

We also evaluate the temporal effect on the centrality measures. The simulation study (Friedkin, 1991) reported that the centrality of nodes is affected by the characteristics of dynamics of information flows. Motivated by this study, we developed a time series analysis. We divide a long period of time into a number of shorter, consecutive time slices. Participating actors in a social network are presented as the vectors of the email conversation frequencies in a time series. The network of the groups of actors is clustered based on the similarity between these vectors. A linkage is defined by a send-reply chain between two actors in a time slice such as a month. The linkage is weighted by the cosine similarity between the vectors. With a chosen threshold, the graph for the network is generated and analyzed with the small-world network model

in Pajek<sup>14</sup>. Centrality scores are obtained through Degree/Closeness/Betweenness analysis in Pajek. The Betweenness analysis is an implementation of Brandes's algorithm (Brandes, 2001).

The original centrality measures assume the impacts of the conversations are equally important over time, which may not consider an expert who contributed to the enterprise community during an early decade. If we are supposed to find an “emerging expert”, this measure may not also be accurate. So we develop another measure by assigning each conversation a delaying weight depending on its age. The modified conversation frequency is divided by  $(T(\text{current}) - T(i) + 1)^\delta$ . If  $\delta$  is set to 1, the conversation frequency is divided by the age. The measure favors a recently active member in a community.

---

<sup>14</sup> <http://vlado.fmf.uni-lj.si/pub/networks/pajek/>

## 7.2 Evaluation Methods and Datasets

To evaluate the consequences of such a change on the assessment of influence, we use the dataset of W3C mail collection from the TREC Enterprise 2005. The W3C email collection used for the experiment was crawled from w3c.org in June 2004. We automatically identified threads (in-reply-to chains) in emails to organize the groups of people in a social network. In the in-reply-to fields of the email messages, there are three types of information, unique message ID, nontrivial subject lines, and null (not a reply). By linking messages with unique message ID and nontrivial subject lines, the pairs of senders and receivers of the messages are treated as discussion threads. This yielded 3032 discussion threads/links among 330 members from 4257 emails from the URI working group at W3C. According to the evaluation of expert search task in TREC Enterprise 2005 competition, we select Dan Connolly, Michael Mealling, and Leslie Daigle as URI experts in addition to well-known members, Tim Berners-Lee, Larry Minster (URI chair), Roy T. Fielding and Martin Duerst. The criteria of goodness on those centrality measures is how many experts in the defined expert set are included in their top 10 ranking list because ideally the seven candidates of the experts should be included in the list.

### 7.3 Experiment Results

**Table 7-1: Top 10 Actors in the URI Working Group Ranked by the Ranking Algorithms, BW, CL, DE, T\_CL, and T\_DE.**

Rank	BW	CL	DE	T_CL	T_DE
1	Larry Masinter*	Larry Masinter*	Larry Masinter*	Larry Masinter*	Larry Masinter*
2	Roy T. Fielding*	Roy T. Fielding*	Roy T. Fielding*	Roy T. Fielding*	Roy T. Fielding*
3	Michael Mealling*	Martin Duerst*	Martin Duerst*	Dan Connolly*	Dan Connolly*
4	Martin Duerst*	Dan Connolly*	Michael Mealling*	Michael Mealling*	Michael Mealling*
5	Dan Connolly*	Michael Mealling*	Dan Connolly*	Martin Duerst *	Martin Duerst *
6	Paul Hoffman	Al Gilman	Al Gilman	Al Gilman	Al Gilman
7	Al Gilman	Graham Klyne	Patrick Stickler	Paul Hoffman	Paul Hoffman
8	Patrick Stickler	Paul Hoffman	Graham Klyne	Daniel LaLiberte	Harald Tveit Alvestrand
9	Daniel LaLiberte	Daniel LaLiberte	Paul Hoffman	Graham Klyne	Daniel LaLiberte
10	Aaron Swartz	Patrick Stickler	Daniel LaLiberte	Ronald E. Daniel	Leslie Daigle*

\*indicates the most influential members

In Tables 7-1, BW means Betweenness. CL denotes Closeness. DE denotes output Degree Centrality. T\_CL denotes time-sensitive Closeness. T\_DE denotes time-sensitive Degree Centrality.

**Table 7-2: Top 10 Actors in the URI Working Group Ranked by the Ranking Algorithms, W\_PR, PR\_BW, T\_BW, PR\_TBW and TE\_BW.**

Rank	W_PR	PR_BW	T_BW	PR_TBW	TE_BW
1	Larry Masinter*	Larry Masinter *	Larry Masinter*	Larry Masinter*	Roy T. Fielding*
2	Roy T. Fielding*	Roy T. Fielding*	Roy T. Fielding*	Roy T. Fielding*	Michael Mealling*
3	Martin Duerst *	Michael Mealling*	Dan Connolly*	Michael Mealling*	Patrick Stickler
4	Michael Mealling*	Martin Duerst *	Michael Mealling*	Martin Duerst*	Al Gilman
5	Al Gilman	Paul Hoffman	Martin Duerst*	Dan Connolly*	Martin Duerst *
6	Patrick Stickler	Al Gilman	Al Gilman	Al Gilman	Chris Lilley
7	Dan Connolly*	Dan Connolly*	Harald Tveit Alvestrand	Paul Hoffman	Graham Klyne
8	Graham Klyne	Patrick Stickler	Paul Hoffman	Daniel LaLiberte	Daniel LaLiberte
9	Daniel LaLiberte	Daniel LaLiberte	Daniel LaLiberte	Harald Tveit Alvestrand	Larry Masinter*
10	Paul Hoffman	Aaron Swartz	Leslie Daigle*	Leslie Daigle*	John Cowan

\*indicates the most influential members

In Tables 7-2, W\_PR denotes weighted PageRank with a fixed parameter  $\lambda=0.85$ . PR\_BW denotes weighted PageRank with a dynamic parameter  $\lambda$  generated from BW. T\_BW denotes time-sensitive Betweenness. PR\_TBW denotes weighted PageRank with a dynamic parameter  $\lambda$  generated from T\_BW. TE\_BW denotes Betweenness centrality with delaying weights on time. The top 10 ranking lists of BW, CL, DE, W\_PR, PR\_BW and T\_CL include 5 influential members. T\_BW, T\_DE and

PR\_TBW identify 6 influential members, including one more influential member, Leslie Daigle. TE\_BW identify four active experts recently, which excludes Leslie Daigle and Dan Connolly. It suggested these two experts might be more active in the early development stage of this working group. Tim Berners-Lee does not appear in the top ranking list because his conversation frequency is ranked as 25<sup>th</sup> in our dataset. If W\_PR is extended by T\_BW, his ranking is 18<sup>th</sup>. But if measured by T\_BW, his best ranking, 15<sup>th</sup>, is achieved.

**Table 7-3: The Spearman Correlations among the Nine Ranking Algorithms except TE\_BW.**

	BW	CL	DE	T_BW	T_CL	T_DE	W_PR	PR_BW
CL	.65*	----	----	----	----	----	----	----
DE	.85*	.78*	----	----	----	----	----	----
T_BW	.73*	.73*	.81*	----	----	----	----	----
T_CL	.65*	.57*	.71*	.81*	----	----	----	----
T_DE	.64*	.55*	.70*	.80*	.99*	----	----	----
W_PR	.83*	.65*	.88*	.76*	.65*	.64*	----	----
PR_BW	.22*	.17*	.16*	.22*	.57*	.55*	.18*	----
PR_TBW	.17*	.11*	.12*	.02	.71*	.70*	.16*	.48*

\*indicates statistically significant with a 95% confidence interval

The results in Table 7-3 indicate that most of the Spearman correlations between the 9 algorithms tend to be statistically significant. Even the correlation scores are changed from 0.11 to 0.99. Correlation coefficients among BW, CL, DE, T\_BW, T\_CL, T\_DE and W\_PR are larger than 0.50 from 0.64 to 0.99. Most of the coefficients among PR\_BW, PR\_TBW and other algorithms are less than 0.50 from 0.11 to 0.48 except T\_CL and T\_DE. The results indicate the integration of PageRank and Betweenness

dramatically changes the topology of the graph. Through a node with a higher Betweenness score, information flows more likely follow the shortest pathways that are linked to the most influential actors.

## 7.4 Conclusion

In summary, there is no substantial difference among the six centrality measures, BW, CL, DE, T\_CL and W\_PR. Weighted PageRank integrated with time-sensitive Betweenness (PR\_TBW), time-sensitive Betweenness (T\_BW), and time-sensitive (T\_DE) perform 60% (6 out of the top 10 ranks) accuracy. They appear to be the best measures to identify influential members from email conversations compared to any other algorithm. Although TE\_BW identify 4 out of the top 10, it emphasizes the discovery of the contemporarily active experts. So including the time attribute improves the centrality measures. Betweenness Centrality is validated to be a good estimator of random jumping probabilities in a social network and it partly solved the rank-sink problem of PageRank.

## **CHAPTER8: Overall Conclusions and Future Work**

### **8.1 Contributions of the Thesis**

We designed and developed key components for knowledge structure extraction and discovery from unstructured text. The central component is a novel LSI Subspace Signature Model which follows Zipf's Law, the term frequency distribution rule in the documents. The model gives unified and comparable spectral signature representations for terms and documents in an unsupervised manner. A unique ranking mechanism for signatures sorts the terms and documents and controls information loss during dimension reduction and sample selection. The similarity measures between the signatures reflect the coherence of term maps and document clusters in the LSI latent concept dimensions.

The model is fundamentally different from traditional LSI. The traditional LSI represents the terms or the documents as the vectors of the projection scores which don't have explicit statistical meanings. The signatures in our model mean the global statistical contribution to either the specific LSI latent concept dimension or the overall LSI subspaces.

Overall, the value of our model was demonstrated across several text mining applications. In this thesis we demonstrate that its applications to visual analytics, concept mapping and evolution, text clustering and active learning. Concept mapping and evolution highlight the representative themes and contexts demonstrate their

semantic network and the trends, which improves the users' understanding of a corpus.

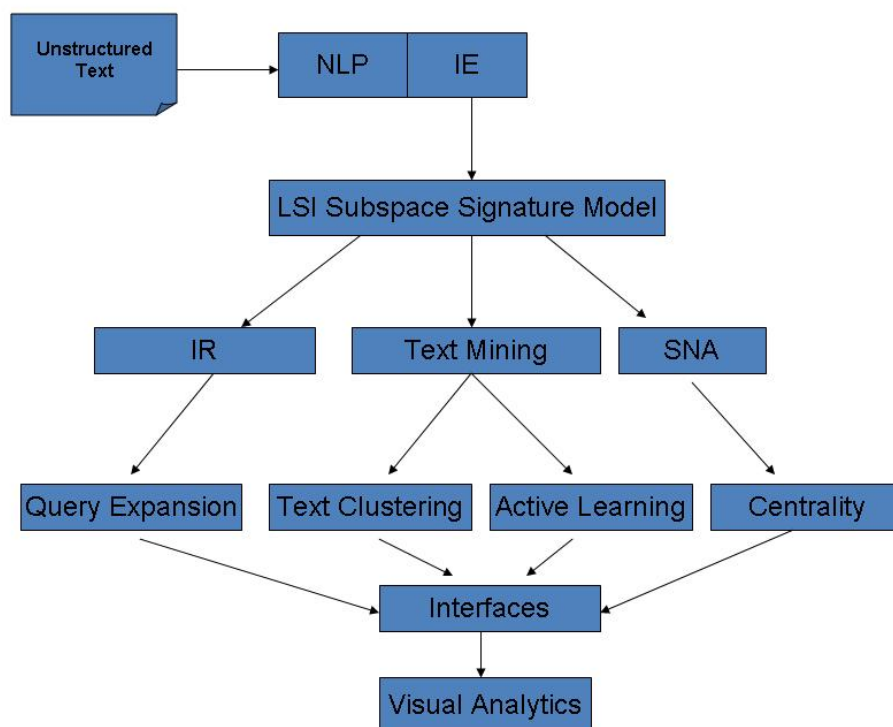
The model proposes two ways to measure similarity, CDSS and GCDSS. The results show that CDSS is beneficial to basic K-means and GCDSS is effective on the competitive learning approach, SOM, compared to VSM and the traditional LSI. Using SOM, the model-based concept clustering substantially improves the document clustering with the basic K-means. Moreover, a new K-means algorithm based on LSISSM, which we term the two-stage K-means, solves the initialization problem of the standard K-means.

In addition to the LSISSM model and its applications, we also study query expansion strategies and social network analysis. As discussed below, The LSI Subspace Signature Model could be suitable for IR. Due to limited time, we only study query expansion strategies on other IR models. The two ontology-based query expansion strategies, UMLS-based term re-weighting and user relevance feedback with IPTC hyponyms, significantly enhance the performance of the VSM and BM25.

Social network analysis captures the characteristics and roles of social indicators and is the indispensable component for knowledge discovery. We developed a novel centrality measure to identify influential member in the social networks. This algorithm integrates Betweenness centrality with the traditional PageRank and solves the rank sink problem of PageRank. Interestingly, the new measure demonstrates a propagation mechanism in which the positive effect of the time factor on the Betweenness score is transformed into PageRank.

## 8.2 Potential Research Directions

Our work on the LSI Subspace Signature Model includes novel ways to calculate the similarity between terms and documents. This suggests that the model could be used as the basis of a search engine. For the large of document sets, for instance the web, sentence-based LSISSM could be used. For the local analysis of query expansion, our model could predict the most related terms to the query.



**Figure 8-1: A Schematic Framework which Integrates NLP, IE, LSISSM, TM, IR, SNA and User Interfaces.**

Given these future directions, LSISSM can be viewed as a complete framework for knowledge extraction, exploration and discovery, see Fig. 8-1. The model can also be applied to the construction of social networks and estimate the textual content-based

influence of the social indicators if using the named entities instead of the terms in the model.

In a sense text clustering and active learning are micro-level classification. At the macro-level, we would like to define the relationships between the class labels. The relationships include hierarchical and associative relationships. The hierarchical relationships between the concept labels can often be obtained from the relevant ontology. One issue is the accuracy of the ontology and another is how to clearly define the associative relationships between concepts in a given context. Unified Medical Language System (UMLS), for instance, provides co-concept pairs and does not explain when they co-occurred together. There can be many kinds of associative relationships; the causal relationship is one of the most important. Structure Equation Modeling (SEM) (Duncan, 1975) models the causal relationships between multiple related variables that contain the latent constructs. Thus, testing the feasibility of SEM on text content analysis is the initial task. If we are able to identify causal relationships with this technique, we may be able to describe causal links among a sequence of news events or the semantic field of the verbs. Further, we can investigate the evolution of the news events and how that semantic field evolves across time. From the perspective of information integration, the content of one news event could be enriched by the same event distributed by different media or online domain thesauri. One good example would be Wikipedia. Combining with these macro level dimensions, our framework could be served as the core of the

information systems on different domains for knowledge discovery and sense making.

## LIST of REFERENCES

- [Agrawal et al., 1996] Agrawal, R., Mannila, H., Srikant, R., Toivonen, H. and Verkamo, A. I. (1996). Fast discovery of association rules. *In Advances in Knowledge Discovery and Data Mining*, U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Eds. American Association for Artificial Intelligence, Menlo Park, CA, 307-328.
- [Ampazis and Perantonis, 2004] Ampazis, N. and Perantonis, S.J. (2004). LSISOM — A Latent Semantic Indexing Approach to Self-Organizing Maps of Document Collections, *Neural Processing Letters*, 19(2),157-173.
- [Arthur and Vassilvitskii, 2007] Arthur, D. and Vassilvitskii, S. (2007). K-means++: The Advantages of Careful Seeding. *Proceedings of the 18<sup>th</sup> Annual ACM-SIAM Symposium on Discrete Algorithms*, 1027-1035.
- [Beil et al., 2002] Beil, F., Ester, M. and Xu, X. (2002). Frequent term-based text clustering. *Proceedings of the 8<sup>th</sup> International Conference on Knowledge Discovery and Data Mining*, 436-442.
- [Blei et al., 2003] Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993–1022.
- [Bollen et al., 2006] Bollen, J., Rodriguez, M. A., and Van de Sompel, H. (2006). Journal status. *Scientometrics*, 69(3), 669-687.
- [Bonacich, 1972] Bonacich, P. (1972). Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology*, 2, 113-120.
- [Borgatti, 2005] Borgatti, S.P. (2005). Centrality and network flow. *Social Networks*, 27, 55-77.
- [Brandes, 2001] Brandes, U. (2001). A Faster Algorithm for Betweenness Centrality. *Journal of Mathematical Sociology*, 25(2), 163-177.
- [Breiman and Spector, 1992] Breiman, L. and Spector, P. (1992) Submodel selection and evaluation in regression: The X-random case, *International Statistical Review*, 60, 291-293
- [Brin and Page, 1998] Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks*, 30(1-7), 107-117.

- [Cai et al, 2005] Cai, D., He, X. and Han, J., (2005). Document Clustering Using Locality Preserving Indexing. *IEEE Transactions on Knowledge and Data Engineering*, 17(12), 1624-1637.
- [Cesarano et al., 2003] Cesarano, C., d’Acierno, A. and Picariello, A. (2003). An Intelligent Search Agent System for Semantic Information Retrieval on the Internet. *Proceedings of the 5th ACM international workshop on Web information and data management*, New Orleans, Louisiana, USA, 111-117.
- [Chen, 2008] Chen, C. (2008) An information-theoretic view of visual analytics. *IEEE Computer Graphics & Applications*, 28(1), 18-23.
- [Choi, 2000] Choi, Y.Y. (2000) Advances in domain independent linear text segmentation. *In Proceedings of NAACL*, Seattle, USA.
- [Cohn et al., 1994] Cohn, D., Atlas, L. and Ladner, R. (1994). Improved Generalization with active learning. *Machine Learning*, 15, 201-221.
- [Cover and Hart, 1967] Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, IT-13, 21–27
- [Deerwester and Dumais et al., 1990] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. and Harshman R. (1990). Indexing by Latent Semantic Analysis. *Journal of the Society for Information Science*, 41(6), 391-407.
- [Dhillon et al., 2002] Dhillon, I.S., Mallela, S. and Kumar, R. (2002). Enhanced word clustering for hierarchical text classification, *ACM CIKM*, 191-200.
- [Ding, 2005] Ding, C. H. (2005). A probabilistic model for Latent Semantic Indexing. *Journal of the Society for Information Science*, 56(6), 597-608.
- [Dubes and Jain, 1988] Dubes, R.C. and Jain, A.K. (1988). *Algorithms for Clustering Data*, Prentice Hall, New York.
- [Duncan, 1975] Duncan O. D. (1975). *Introduction to Structural Equation Models*, ISBN 0122241509, Academic Press, Inc. (London) Ltd.
- [Elisseeff and Weston, 2002] Elisseeff A. and Weston J. (2002) A kernel method for multi-labelled classification, in *Advances in Neural Information Processing Systems 14*, T. G. Dietterich, S. Becker, and Z. Ghahramani, Eds. Cambridge, MA: MIT Press, 681–687.
- [Faloutsos et al., 2004] Faloutsos, C., McCurley, K. and Tomkins, A. (2004). Fast discovery of connection subgraphs. *ACM SIGKDD*, 118-127.

- [Freeman, 1997] Freeman, L. C. (1997). A set of measures of centrality based on Betweenness. *Sociometry*, 40, 35-41.
- [Freeman, 1979] Freeman, L.C. (1979). Centrality in social networks: Conceptual clarification. *Social Networks*, 1, 215–239.
- [Friedkin, 1991] Friedkin, N.E. (1991). Theoretical foundations for centrality measures. *American Journal of Sociology*, 96(66), 1478-1504.
- [Fung et al., 2003] Fung, B. C. M., Wang, K. and Ester, M. (2003). Hierarchical Document Clustering Using Frequent Itemsets, *SIAM International Conference on Data Mining (SDM'03)*, San Francisco, CA, 59-70.
- [Gelbukh and Sidorov, 2001] Gelbukh, A. and Sidorov, G. (2001). Zipf and Heaps Laws' Coefficients Depend on Language. *Proc. CICLing-2001, Conference on Intelligent Text Processing and Computational Linguistics*, February 18–24, 2001, Mexico City. Lecture Notes in Computer Science N 2004, ISSN 0302-9743, ISBN 3-540-41687-0, Springer-Verlag, 332–335.
- [Griffiths et al., 2007] Griffiths, T. L., Steyvers, M. and Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, 114, 211-244.
- [Guo et al., 2004] Guo, Y., Harkema, H. and Gaizauskas, R. (2004). Sheffield University and the TREC 2004 Genomics Track: Query Expansion Using Synonymous Terms, *13th Text Retrieval Conference (TREC 2004)*.
- [Hersh and Hickam, 1995] Hersh, W. and Hickam, D. (1995). Information Retrieval in Medicine-The Sapphire Experience, *Journal of the American Society for Information Science*, 46(10), 743-747.
- [Hersh et al., 2000] Hersh, W., Price, S. and Donohoe, L. (2000). Assessing thesaurus-based query expansion using the UMLS Metathesaurus, *Proc AMIA Symposium 2000*.
- [Hofmann, 1999] Hofmann, T. (1999). Probabilistic latent semantic indexing. *ACM SIGIR*, 50–57, New York: ACM Press.
- [Huberman and Wu, 2004] Huberman, B. and Wu, F. (2004). Finding communities in linear time: a physics approach. *The European Physical Journal B - Condensed Matter*, 38(2), 331-338.
- [Ittner et al., 1995] Ittner, D.J., Lewis, D. D. and Ahn, D. D. (1995). Text categorization of low quality images. *In Symposium on Document Analysis and Information Retrieval (Las Vegas, NV)*. 301–315.

- [Jones et al., 2000] Jones, K. S., Walker, S., and Robertson, S. E. (2000). A Probabilistic Model of Information Retrieval: Development and Comparative Experiments (parts 1 and 2). *Information Processing and Management*, 36(6), 779-840.
- [Kleinberg, 1999] Kleinberg, J. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46, 604-632.
- [Kohonen, 1990] Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE* 78 9, 1464–1480.
- [Kontostathis and Pottenger, 2006] Kontostathis, A. and Pottenger, W. M., (2006). A framework for understanding latent semantic indexing (LSI) performance. *Information Processing and Management*, 42, 1, 56-73.
- [Leroy and Chen, 2001] Leroy, G. and Chen, H. (2001). Meeting Medical Terminology Needs: The Ontology-Enhanced Medical Concept Mapper, *IEEE Transactions on Information Technology in Biomedicine*, 5(4), 261-270.
- [Lewis and Catlett, 1994] Lewis, D. and Catlett, J. (1994). Heterogeneous uncertainty sampling for supervised learning. In *Proceedings of the Eleventh International Conference on Machine Learning*, 148-156, New Brunswick, NJ. Morgan Kaufmann.
- [Lin et al., 1991] Lin, X., Soergel, D. and Marchionini, G. (1991). A self-organizing semantic map for information retrieval, *ACM SIGIR*, 262-269.
- [Manning et al., 2008] Manning, D. C., Raghavan, P. and Schütze, H., (2008). *Introduction to Information Retrieval*, Cambridge University Press, ISBN 0521865719, 357-359.
- [McCallum and Nigam, 1998] McCallum, A. and Nigam, K. (1998). Employing EM and Pool-Based Active Learning for Text Classification. *Proceedings of the Fifteenth International Conference on Machine Learning*, 350-358.
- [Murray, 2005] Murray, R.L. (2005). Toward a metadata standard for digitized historical newspapers, *ACM/IEEE JCDL*, Denver, 330-331.
- [NIST, 1998] National Institute of Standards and Technology (NIST) (1998). The Topic Detection and Tracking Phase 2 (TDT2) Evaluation Plan Version 3.7.
- [Newman, 2004] Newman, M. (2004). Who is the best connected scientist? A study of scientific co-authorship networks. In E. Ben-Naim, H. Frauenfelder, and Z. Toroczkai, (eds.) *Complex Networks*, Springer, 337–370.

- [Page et al., 1998] Page, L., Brin, S., Motwani, R. and Winograd, T. (1998). The PageRank citation ranking: Bringing order to the Web. Technical report, Stanford Digital Library Technologies Project.
- [Ponte and Croft, 1998] Ponte, J. and Croft, W.B. (1998). A Language Modeling Approach to Information Retrieval, *ACM SIGIR*, 275-281.
- [Pujol et al., 2002] Pujol, J. M., Sangüesa, R. and Delgado, J. (2002). Extracting reputation in multi agent systems by means of social network topology. *Proceedings of the first international joint conference on Autonomous agents and multi-agent systems*, 467-474.
- [Richardson and Smeaton, 1995] Richardson R. and Smeaton A.F (1995). Using Wordnet in a knowledge-based approach to information retrieval, *In Proceedings of the BCS-IRSG Colloquium*, Crewe.
- [Robertson et al., 1994] Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M. and Gatford, M. (1994). Okapi at TREC-3. *In Proceedings of the Third Text Retrieval Conference (TREC)*. Gaithersburg, USA.
- [Rocchio, 1971] Rocchio, J. J., Relevance feedback in information retrieval, (1971) *SMART Retrieval System - Experiments in Automatic Document Processing*, New York, Prentice Hall.
- [Rumelhart, 1990] Rumelhart, D. E. (1990). Brain style computation: Learning and Generalization. In Zornetzer, S. F., Davis, J. L., and Lau, C. (eds.), *An Introduction to Neural and Electronic Networks*, 405-420. San Diego, CA: Academic Press.
- [Rumelhart, McClelland and the PDP Research Group, 1986] Rumelhart, D. E., McClelland, J.L. and the PDP Research Group (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Volume 1: Foundations. Cambridge, MA: MIT Press.
- [Sabidussi, 1966] Sabidussi, G. (1966). The centrality index of a graph. *Psychometrika*, 31, 581-603.
- [Salton et al., 1975] Salton, G., Wong, A. and Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18, 11, 613-620.
- [Schütze and Silverstein, 1997] Schütze, H. and Silverstein, C. (1997). Projections for efficient document clustering. *SIGIR Forum* 31, SI, 74-81.
- [Šilić et al., 2008] Šilić, A., Moens, M.F., Žmak, L. and Bašić, B.D., (2008). Comparing Document Classification Schemes Using K-Means Clustering, *Knowledge-Based Intelligent Information and Engineering Systems*, 615 – 624.

- [Steinbach et al., 2000] Steinbach, M., Karypis, G. and Kumar, V. (2000). A Comparison of Document Clustering Techniques, *Proc. TextMining Workshop in KDD 2000*.
- [Zhang and Zhou, 2006] Zhang, M. and Zhou, Z. (2006) Multilabel Neural Networks with Applications to Functional Genomics and Text Categorization, *IEEE Transactions on Knowledge and Data Engineering(TKDE)*, 18(10), 1338-1351.
- [Thomas and Cook, 2005] Thomas, J.J. and Cook, K.A. (2005). *Illuminating the Path: The Research and Development Agenda for Visual Analytics*, National Visualization and Analytics Center.
- [Tong and Koller, 2001] Tong, S. and Koller, D. (2001). Support Vector Machine Active Learning with Applications to Text Classification. *Journal of Machine Learning Research*. 2, 45-66.
- [Toutanova and Manning, 2000] Toutanova, K. and Manning, D. C. (2000). Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. *In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000)*, 63-70.
- [Toutanova et al., 2003] Toutanova, K., Klein, D., Manning, D. C. and Singer, Y. (2003). Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. *In Proceedings of HLT-NAACL 2003*, 252-259.
- [Vectomova and Wang, 2006] Vectomova, O. and Wang Y. (2006). A study of the effect of term proximity on query expansion. *Journal of Information Science*, 32 (4), 324–333.
- [Wang et al., 1999] Wang, K., Xu, C. and Liu, B. (1999). Clustering transactions using large items. *ACM CIKM*, 483–490.
- [White and Smyth, 2003] White, S. and Smyth, P. (2003). Algorithms for estimating relative importance in networks. *ACM SIGKDD*, 266–275.
- [Willett, 1998] Willett, P. (1998). Recent trends in hierarchic document clustering: a critical review. *Information Processing and Management*, 24(5), 577-597.
- [Werbos, 1994] Werbos, P. J. (1994). *The Roots of Backpropagation: from Ordered Derivatives to Neural Networks and Political Forecasting*. Wiley-Interscience.
- [Xu and Croft, 1996] Xu, J. and Croft, W. (1996). Query expansion using local and global document analysis, *ACM SIGIR*, 4-11.

- [Xu et al., 2000] Xu, J. and Croft, W.B. (2000). Improving the Effectiveness of Information Retrieval with Local Context Analysis, *ACM Transactions on Information Systems*, 18(1), 79-112.
- [Xu and Gong, 2004] Xu, W. and Gong, Y. (2004). Document clustering by concept factorization, *ACM SIGIR*, 202–209.
- [Xu et al., 2003] Xu, W., Liu, X. and Gong, Y. (2003). Document clustering based on non-negative matrix factorization. *ACM SIGIR*, 267–273.
- [Zhang and Zhou, 2006] Zhang, M. and Zhou, Z. (2006) Multilabel Neural Networks with Applications to Functional Genomics and Text Categorization, *IEEE Transactions on Knowledge and Data Engineering(TKDE)*, 18(10), 1338-1351.
- [Zhu et al., 1999] Zhu, A., Gauch, S., Lutz, G., Kral, N. and Pretschner, A. (1999). Ontology-Based Web Site Mapping for Information Exploration, *ACM CIKM*, 188-194.
- [Zhu et al., 2007] Zhu, W. and Chen, C. (2007). Storylines: Visual exploration and analysis in latent semantic spaces, *International Journal of Computers and Graphics, Special Issue on Visual Analytics*. 31(3), 338-349.
- [Zhu et al., 2008] Zhu, W., Chen, C. and Allen, R. B. (2008). Analyzing the Propagation of Influence and Concept Evolution in Enterprise Social Networks through Centrality and Latent Semantic Analysis. Washio, T. et al. (Eds.): *Advances in Knowledge Discovery and Data Mining, PAKDD 2008*, Osaka, Japan, LNCS, 5012, 1090-1098.
- [Zhu et al., 2006] Zhu, W., Xu, X., Hu, X., Song, Il Y. and Allen, R.B. (2006). Using UMLS-based Re-Weighting Terms as a Query Expansion Strategy, *Proceedings of the 2<sup>nd</sup> IEEE Conference on Granular Computing*, 217-222.

## VITA

### EDUCATION

- 09/04 to 06/09 Ph.D. in College of Information Science and Technology  
Drexel University, Philadelphia, PA 19104
- 08/00 to 05/02 M.S. in Computer Science  
Saint Joseph's University, Philadelphia, PA19131
- 08/96 to 07/99 M.S. in Physical Chemistry  
Shandong University, Jinan, P. R. China
- 08/91 to 07/95 B.A. in Analytical Chemistry  
Shandong University, Jinan, P.R. China

### RESEARCH INTERESTS

Information Visualization, Information Retrieval, Text Mining and Data Mining, Knowledge Management, Digital Library, Bioinformatics and Social Network Analysis

### AWARDS and HONORS

- Weizhong Zhu and Xia Lin, "Mapping bio-medical concept space", Honorable Mention at Seventh Annual Drexel Research Day, 2005.
- Weizhong Zhu and Chaomei Chen, "Visual analysis of terrorism events extracted from the public knowledge bases", Dean's Honorable Mention in *i*-school Research Gallery at Ninth Annual Drexel Research Day, 2007.
- Honorable Mention in the Official Newsletter at DHS Student & Alumni Network (<http://www.dhsnetwork.org/documents/Nov07-visualanalysisresearchfeature.pdf>), November 2007.

### PUBLICATIONS

2 Journal Papers, 11 Conference Papers, 2 Demos and 2 Posters in Information and Computer Science, 2 Journal Papers in Cognitive Psychology, 3 Journal Papers in Chemistry

### PROFESSIONAL SERVICE ACTIVITIES

#### Conference Reviewer

- JCDL Ad Hoc Reviewer

#### Journal Reviewer

- Information Visualization
- Journal of Computing Science and Engineering

### TEACHING EXPERIENCE

03/08-06/09 TA for IR Systems, Database Management Systems at Drexel University

08/01-05/02 TA for JAVA Programming, Algorithms at Saint Joseph's University

