自然言語処理　144－9
（２００１．７．１６）

Simple PCA を用いたベクトル空間情報検索モデルの次元削減

黒岩眞吾　柘植 覚　田仁 宏典　Tai Xiaoying　獅々堀 正幹　北 研二

徳島大学　工学部　知能情報工学科
〒 770-8506 徳島市南常三島町２－１

ベクトル空間モデル (VSM) は情報検索における代表的な検索モデルである．同モデルでは文書が単語の出現頻度に基づくベクトルで表現されるため，そのベクトル空間は一般にスパースかつ高次元となりメモリや検索時間の増大を招くとともに，文書中に含まれる無意味な単語がノイズ的な影響を及ぼし検索精度を低下させるという問題を生じる．これに対し特異値分解 (SVD) を用い次元数を削減した空間で類似度を計算する潜在的意味インデキシング (Latent Semantic Indexing; LSI) が提案され，その効果が報告されている．本稿では SVD に比べより少ない演算量で近似的に主成分分析を行うことが可能な Simple Principal Component Analysis(SPCA) を次元削減に適用する．MEDLINE コレクションを用いた検索実験を行った結果，SVD と同等以上の検索性能を SPCA により達成した．

Simple PCA，情報検索，LSI，ベクトル空間モデル，次元削減

Dimensionality Reduction of Vector Space Model for Information Retrieval using Simple Principal Compornent Analysis

Shingo Kuroiwa　Satoru Tsuge　Hironori Tani　Tai Xiaoying
Masami Shishibori　Kenji Kita

The University of Tokushima
2-1, Minami-Josanjima, Tokushima, 770-8506

Abstract　The Vector Space Model (VSM) is a popular information retrieval model, which represents a document collection by a term-by-document matrix. Since term-by-document matrices are usually high-dimensional and sparse, they are susceptible to noise and are also difficult to capture the underlying semantic structure. Additionally, computing resources necessary for the storage and processing of such data is enormous. Dimensionality reduction is a way to overcome these problems. Principal Component Analysis (PCA) and Singular Value Decomposition (SVD) are popular techniques for dimensionality reduction based on matrix decomposition. However, such methods consume a large amount of computation resources. In the work described here, we use Simple Principal Component Analysis (SPCA), which is a data-oriented fast method, for dimensionality reduction of the vector space model. Experiments based on the MEDLINE collection showed that SPCA achieved significant improvement compared to the conventional vector space model.

key words　Simple PCA, Information retrieval, LSI, VSM, Dimensionality reduction

1. まえがき

インターネット技術の急速な発展と普及により，World Wide Web (WWW) を代表とする，オンラインテキストデータ量は増加の一途をたどっている．これに伴い，莫大なテキストデータ中から必要かつ十分な情報のみを効果的に検索できる検索エンジンへの期待が高まっている．その期待に答えるべく多くの研究機関や企業によって，様々な手法を用いた情報検索に関する研究・開発が積極的に進められている．これらの研究については，米国における TREC (Text Retrieval Conference)[1] や，日本における IREX (Information Retrieval and Extraction Exercise)[2]，NTCIR (NII-NACSIS Test Collection for IR Systems)[3] のワークショップにおいてその成果が報じられているが，今なお発展途上の段階である．

代表的な情報検索モデルの一つとして，検索対象文書と検索質問を多次元ベクトルで表現するベクトル空間モデル (VSM; Vector Space Model)[4] がある．VSM を用いた情報検索システムは，質問ベクトルと各文書ベクトル間の距離 (類似度) を計算し，距離の近い文書を検索結果として出力するものであり，検索精度を上げるため様々な距離尺度が提案されている [5].

しかし，同手法では全文書中に含まれる単語が，各文書中にどの程度出現したかに基づき多次元ベクトルを構成するため，そのベクトルの次元数は巨大なものとなる．また，各文書に出現する単語は限られているため，ベクトルは要素に 0 の多いスパースなベクトルとなる．文書全体をこのようなベクトルで表現すると，莫大な記憶容量が必要になると同時に，類似度計算を行う際の計算コストも増加してしまう．また，文書中に含まれる無意味な単語の有無が距離に影響を与え検索精度を低下させてしまうという現象も生じやすい．このため，これらのスパースなベクトルで表現された文書全体 (単語・文書行列) の次元数を削減する手法が数多く提案されている [5, 6, 7].

単語・文書行列の次元数を削減する最も代表的な手法として，特異値分解 (SVD; Singular Value Decomposition) を用いた潜在的意味インデキシング (LSI; Latent Semantic Indexing) がある [5, 6, 7]. この手法は，単語・文書行列に対し特異値分解を行い，元の単語・文書行列より低いランクの基底行列を求め，その基底に各ベクトルを射影することにより，次元数を削減する方法である．次元削減された空間では潜在的な意味の近さに文書ベクトル間の距離が反映されると考えられ，実際この空間で検索を行うことで多くの場合，検索性能は高くなる [6, 7]. しかし，特異値分解は計算コストが高いため，大規模な行列に対して適用することが困難な場合もある．

そこで，本稿では Simple Principal Component Analysis (SPCA)[8] を用いたベクトル空間モデルの次元削減手法を提案する．SPCA は，文書ベクトルの重みつき平均演算を繰り返すことで主成分分析を近似的に行う手法であり，少ない計算量で寄与率の大きい順に主成分を計算可能である．また，同手法における繰り返し演算は k-means クラスタリング等の教師なしクラスタリングによる手法とも類似性があり，通常の主成分分析を用いて次元削減を行った場合に比べ，クラスタリングの効果により高い性能を達成できる可能性もある．

以下，2 節において，SPCA のアルゴリズムを説明する．3 節では，SPCA を用いた次元削減法の有効性を検証するため，英文情報検索テストコレクション MEDLINE を用いた情報検索実験を行うとともに，それらの結果に対する考察を行う．最後に，4 節において，本稿のまとめと今後の課題について述べる．

2. Simple PCA

Simple Principal Component Analysis (SPCA) は，主成分分析の高速化を目的に Patridge らによって提案され手法であり，手書き文字の情報圧縮においてその効果が確認されている [8]．同手法は，単純な繰り返し演算により主成分を近似的に抽出する手法で，SVD 等の行列演算に基づく手法 (Matrix Method) とは異なり，ニューラルネットワーク法 [9] 等のデータに基づく手法 (Data Method) に分類される．

本節では SPCA のアルゴリズムを文献 [8] に従い説明する．ベクトル集合 $\boldsymbol{V} = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_m\}$ の主成分を抽出する問題を考える．まず，通常の主成

分分析と同様に，集合の重心を原点とするために各ベクトルから全ベクトルの平均ベクトルを減じ新たなベクトル集合 $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_m\}$ (以下，入力データと呼ぶ) を求める．

$$\boldsymbol{x}_i = \boldsymbol{v}_i - \frac{1}{m}\sum_{j=1}^{m}\boldsymbol{v}_j$$

これらの入力データの第一主成分を表す固有ベクトルを $\boldsymbol{\alpha}_1$ としたとき，以下の出力関数 (output funciotn) y を考える．

$$y_1 = \boldsymbol{\alpha}_1^T \boldsymbol{x}_j \tag{1}$$

この関数は，$\boldsymbol{\alpha}_1$ と直交する原点を通る超平面に対し，入力ベクトルが $\boldsymbol{\alpha}_1$ と同じ側に存在すれば正の値を，反対側に存在すれば負の値を返す関数である．この関数を用い次の閾値関数 (threshold function) を導入する．

$$\Phi_1(y_1, \boldsymbol{x}_j) = \begin{cases} \boldsymbol{x}_j & \text{if } y_1 \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

以上の関数を用い，初期値として適当に与えた任意のベクトル $\boldsymbol{a}_1$ を繰り返し演算により，$\boldsymbol{\alpha}_1$ と同じ方向に近づけることを考える．この繰り返し演算は以下の式で表現される．

$$\boldsymbol{a}_1^{k+1} = \frac{\sum_j \Phi_1(y_1, \boldsymbol{x}_j)}{\left\|\sum_j \Phi_1(y_1, \boldsymbol{x}_j)\right\|}; \qquad y_1 = (\boldsymbol{a}_1^k)^T \boldsymbol{x}_j \tag{3}$$

ここで，k は繰り返し回数であり，出力関数 y_1 は 1 つ前の繰り返しで推定した $\boldsymbol{a}_1^k$ により計算される．この演算は $\boldsymbol{a}_1$ を分散のより大きな方向に修正する性質があるので，同演算を繰り返すことにより，$\boldsymbol{a}_1^k$ は $\boldsymbol{\alpha}_1$ と同じ方向に収束すると期待できる．

以上の演算により第一主成分が計算できるが，引き続き第二主成分を求めるために入力データより第一主成分を取り除く必要がある．そこで，下記の式により各データ $\boldsymbol{x}_i$ から第一主成分を除去する．

$$\boldsymbol{x}_i' = \boldsymbol{x}_i - (\boldsymbol{a}_1^T \boldsymbol{x}_i)\boldsymbol{a}_1 \tag{4}$$

この $\boldsymbol{x}_i'$ に対し式 (3) により $\boldsymbol{a}_2$ を求めることで第二主成分が得られる (ただし，$\boldsymbol{a}_1$ は $\boldsymbol{a}_2$ に，$\boldsymbol{x}_j$ は $\boldsymbol{x}_j'$ に置き換える)．以下，同様の操作を繰り返すことで寄与率の大きい順に主成分を得ることができる．

なお文献 [8] においては，閾値関数として式 (2) に加え以下の 3 つの関数を閾値関数として提案している．

$$\Phi_2(y, \boldsymbol{x}_j) = \begin{cases} +\boldsymbol{x}_j & \text{if } y \geq 0 \\ -\boldsymbol{x}_j & \text{otherwise} \end{cases} \tag{5}$$

$$\Phi_3(y, \boldsymbol{x}_j) = y\boldsymbol{x}_j \tag{6}$$

$$\Phi_4(y, \boldsymbol{x}_j) = \frac{1}{\|\boldsymbol{a}^k\|} y\boldsymbol{x}_j \tag{7}$$

3 節の検索実験では，式 (2) に加え，この 3 つの閾値関数についても性能比較を行う．

なお，SPCA の計算量は $O(dmn)$(ただし，d は求める主成分数，m はデータ数，n はベクトルの次元数，また収束回数が定数となる) であり，SVD 等の行列演算に基づく方法に比べ格段に少ない演算量で主成分を近似的に求めることが可能である．

3. 検索実験

情報検索における SPCA を用いた次元削減手法の有効性を検証するため，情報検索評価用テストコレクション MEDLINE[6] を用い以下の 3 つの手法による情報検索実験を行った．

(1) ベクトル空間モデル

(2) SVD を用いた LSI

(3) SPCA を用いた LSI（提案手法）

3.1 実験条件

MEDLINE は，医学・生物学分野における英文の文献情報データベースである．このテストコレクションは，検索対象文書 1,033 文書で構成される約 1Mbyte の容量を持つテキストデータである．情報検索評価用データとして，30 個の検索質問文と各検索質問に対する正解 (適合) 文書が用意されている．各検索質問に対する平均関連文書数は 23.2 文書である．

前処理として，このテストコレクションに含まれる 1,033 文書全体から，“a” や “about” などの一般的な 439 語を，文書の内容と関連の無い単語 (不要語) として削除した．また，全文書中に 1 回しか出現しなかった単語も削除した．この処理により削除されなかった単語に対し，Porter 法による接辞処理 [10] を施し，語幹への変換を行った．以上の処理の

結果，文書全体に存在した異なり単語数は 5,526 から 4,329 単語に削減された．実験ではこれら 4,329 単語を検索に用いる索引語とした．

VSM　前処理によって得られた索引語を用い，4,329 次元の文書ベクトルを構成しベクトル空間モデル (VSM) に基づく情報検索システムを構築した．文書 j に対応する文書ベクトル $\boldsymbol{v}_j$ の要素 v_{ij} は，文書 j の索引語 i のローカル重み L_{ij} と索引語 i のグローバル重み G_i の積で表現される．すなわち，

$$v_{ij} = L_{ij} \cdot G_i \tag{8}$$

となる．

各々の重みとしては，対数エントロピーに基づく重み [11] を用いた．この重みは，

$$L_{ij} = \begin{cases} 1 + \log f_{ij} & (f_{ij} > 0) \\ 0 & (f_{ij} = 0) \end{cases} \tag{9}$$

$$G_i = 1 + \sum_{j=1}^{m} \frac{\frac{f_{ij}}{F_i} \log \frac{f_{ij}}{F_i}}{\log m} \tag{10}$$

で与えられる．m はテストコレクション中の文書数を，f_{ij} は文書 j における索引語 i の出現頻度を，F_i は全文章中における索引語 i の出現頻度を各々表す．

SPCA　まず VSM の文書ベクトル集合 $\{\boldsymbol{v}_j\}$ を用い 2 節に示したアルゴリズムにより寄与率の高い順に N 個の主成分を求めた．次にこれらの主成分を基底ベクトルとして次元数の削減を行い，次元削減されたベクトル間の cosine 距離により検索を行った．閾値関数としては式 (2)，式 (5)，式 (6)，式 (7) の 4 種類を用い収束性および次元数に関し平均適合率を尺度に比較を行った．なお，初期ベクトル $\boldsymbol{a}^0$ は次元すべての値を 1 とした．

SVD　比較のために，SVD により SPCA と同次元に次元数を削減した場合の実験も行った．SVD にはランチョス法 (SVDPACK)[12] を用いた．

評価方法　検索システムの精度の評価には，“trec_eval” プログラム [1] を用いて平均適合率 (non-interpolated average precision) および再現率・適合率曲線 (recall-precision curve) を求めた．ただし，平均適合率を求めるにあたっては，実用性を考慮し上位 50 位までの検索結果を用いた．なお，適合率および再現率の定義は以下のとおりである．

$$\text{適合率} = \frac{\text{検索できた関連文書数}}{\text{検索文書数}}$$

$$\text{再現率} = \frac{\text{検索できた関連文書数}}{\text{関連のある文書数}}$$

3.2　収束性

SPCA の繰り返し演算による収束性を調べるために，各閾値関数を使った場合の繰り返し回数と平均適合率の関係を調べた．次元削減後の次元数を 50 とした場合 (この次元数において SVD では，最高の平均適合率が得られた) の結果を図 1 に示す．図中の Eq.(2)，Eq.(5)，Eq.(6)，Eq.(7) は，式 (2)，式 (5)，式 (6)，式 (7) の閾値関数を用いた場合の結果を各々示している．比較のために SVD の結果も示した．どの閾値関数を用いた場合も同様に，SPCA は繰り返し回数 10 で SVD の平均適合率を上まわっており，SPCA の有効性が確認できる．その後，繰り返し回数 30 を超えた段階で平均適合率は収束している．安定性や頑健性という点を考慮に入れれば，式 (7) の閾値関数が実用的である．なお，次元数を 20〜200 程度に振った場合でも，収束性に関しては同様な結果が得られている．

3.3　最適次元数

繰り返し回数を 10 に固定して次元数と平均適合率の関係を調べた．結果を図 2 に示す．比較のために SVD を用いた場合の結果もプロットした．SPCA を用いることで，すべての次元数において SVD より高い平均適合率が得られた．SPCA を用いた場合の最高の平均適合率は式 (5) の閾値関数を用いた次元数 20 のときで 0.683 である．これに対し SVD では次元数 50 のときの 0.663 が最高の平均適合率であり，SPCA を用いることで誤り率を 5%削減できたことになる．なお，VSM を用いた場合の平均適合率は 0.494 で図 2 の範囲外となっている．

3.4　再現率・適合率曲線

SPCA，SVD，VSM の性能の違いをより詳細に調査するために，各手法において平均適合率が最高となった条件下で再現率・適合率曲線を作成した．結果を図 3 に示す．図中の () 内の数字は各手法での次元

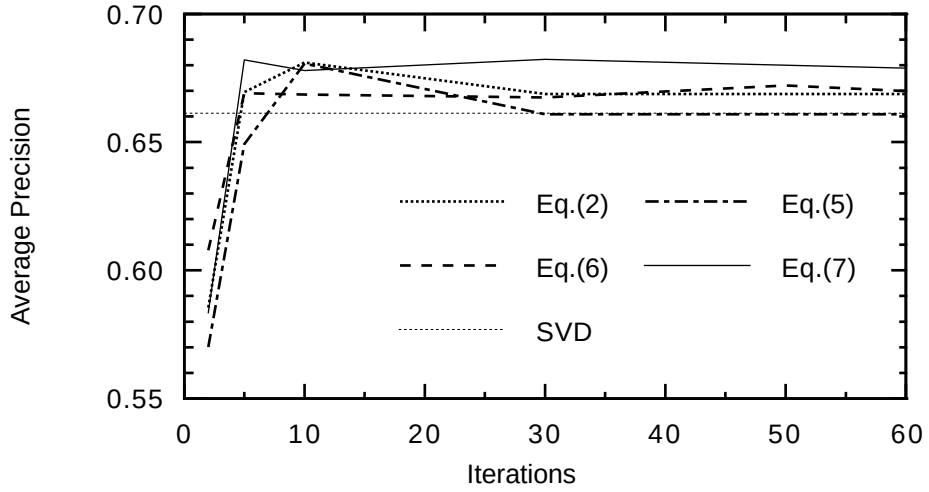

図 1　繰り返し回数と平均適合率

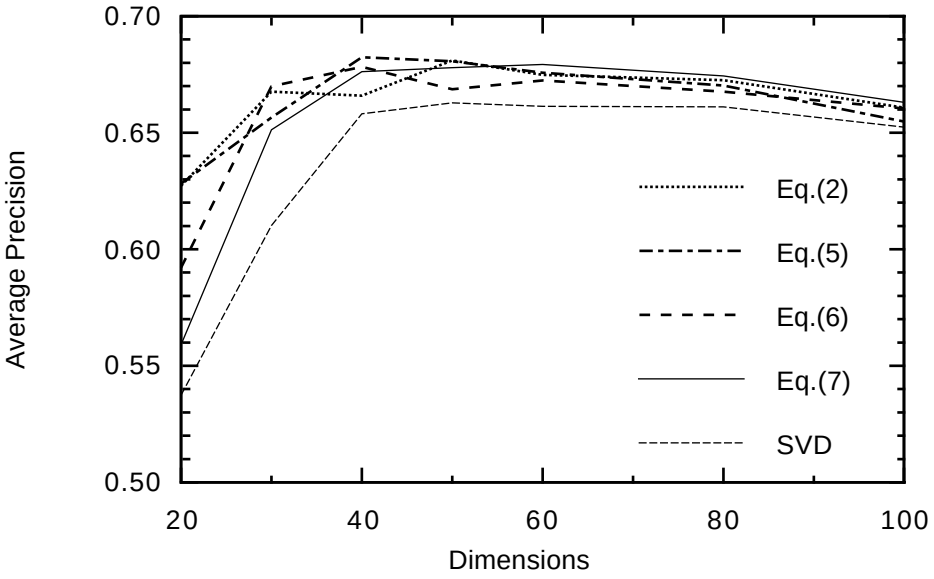

図 2　次元数と平均適合率

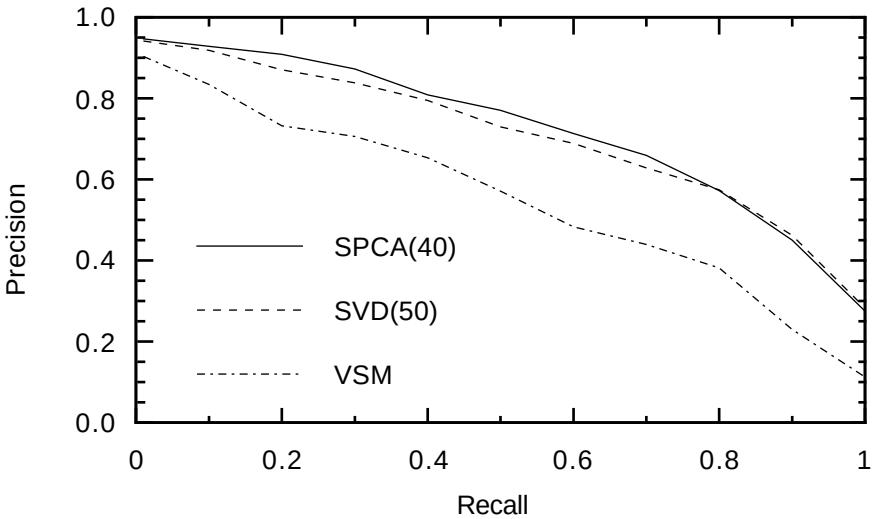

図 3　再現率・適合率曲線

数である．VSM に比べ，次元数を削減した SPCA，SVD では顕著な性能の向上が見られる．SPCA と SVD を比較した場合，SPCA が平均的に高い適合率を示しているが，再現率 0.8 以上では SPCA よりも SVD の適合率がやや高くなる．しかし，一般的な情報検索の問題を考えた場合，検索結果として利用者に提示できる文書数は限られる場合が多く，実用的範囲では SPCA が SVD よりも高い性能を示すと考えられる．

3.5 考察

SPCA の計算量 $O(dmn)$ の定数項となる繰り返し演算回数は 10 程度で十分であり，計算量という観点からの SPCA の有効性は確認できたと思われる．一方，検索精度の点からも SPCA は SVD と同等以上の性能を示しており有効な手法であると結論できる．なお，SPCA は本来主成分分析を近似的に求める手法として提案されたものであり，その性能の上限は主成分分析と等価な手法である SVD に等しいと考えられる．これに対し本実験で SPCA が SVD 以上の性能を示したことは，SPCA の繰り返し演算にクラスタリング的要素が含まれているため，通常の主成分分析以上に潜在的意味を反映した次元削減が行われている可能性を示唆するものと考えられる．

4. まとめ

本稿では，情報検索の代表モデルであるベクトル空間モデル (VSM; Vector Space Model) の次元削減手法として，Simple Principal Component Analysis (SPCA)[8] を用いることを提案した．SPCA は簡単な繰り返し演算により，主成分を寄与率の大きい順に求めていく近似的な主成分分析手法であり，SVD 等の行列演算を基本とする手法に比べ計算量が格段に少ないという特長がある．また，同手法はクラスタリングに基づく次元削減手法と解釈することも可能であり通常の主成分分析法以上に潜在的意味が各基底ベクトルに反映される可能性がある．

MEDLINE を用いた検索実験を行った結果，繰り返し回数 10 回で削減した次元数によらず，SVD に比べて同等以上の平均適合率を得た．SPCA の計算量は $O(dmn)$(ただし，d は求める主成分数，m はデータ数，n はベクトルの次元数) であり，繰り返し回数 10 回程度で十分な性能が得られたことから，計算量に関して SPCA が有効であることが確認できた．

また，この実験における最高の平均適合率は SPCA で 0.682(繰り返し回数 10，次元数 40)，SVD で 0.663(次元数 50) であり，SPCA を用いることで誤り率は 5% 削減されている．これは，次元削減においてクラスタリング的な手法を導入することで検索精度が改善される可能性を示唆した結果と考えられる．そこで，現在，教師無しクラスタリング手法の 1 つであるベクトル量子化を用いた次元削減法の実験を進めている．

参考文献

[1] TREC Homepage. http://trec.nist.gov/.

[2] IREX Homepage. http://cs.nyu.edu/cs/projects/proteus/irex.

[3] NTCIR Homepage. http://www.rd.nacsis.ac.jp/~ntcadm/.

[4] G. Salton and J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill Book Company, 1983.

[5] M. W. Berry, Z. Drmac, and E. R. Jessup. Matrices, vector spaces, and information retrieval. *SIAM Review*, 41:335–362, 1999.

[6] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[7] T. A. Letsche and M. W. Berry. Large-scale information retrieval with latent semantic indexing. *Information Sciences – Applications*, 100:105–137, 1997.

[8] M. Partridge and R. Calvo. Fast dimensionality reduction and simple PCA. *IDA*, 2(3):292–298, 1997.

[9] K. I. Diamantaras and S. Y. Kung. *Principal component neural networks : theory and applications*. Wiley, New York, 1996.

[10] W. B. Frakes and R. Baeza-Yates. *Information Retrieval: Data Structures and Algorithms*. Prentice Hall, 1992.

[11] E. Chisholm and T. Kolda. New term weighting formulas for the vector space method in information retrieval. *Technical Memorandum ORNL-13756*, 1999.

[12] M. W. Berry, T. Do, G. O'Brien, and V. D. Pietra. *SVDPACKC (Version 1.0) User's Guide*. Department of Computer Science, University of Tennessee, 1993.