文書ベクトルの次元削減に基づく類似文書判定の一考察

梅澤香矢乃† 小林 一郎†

† お茶の水女子大学大学院人間文化創成科学研究科理学専攻 〒 112–8610 東京都文京区大塚 2-1-1
E-mail: †{umezawa.kayano,koba}@is.ocha.ac.jp

あらまし 文書検索技術において，テキストデータを文書ベクトルで表現した際，一般的に高次元ベクトルのデータとなる．高次元データをそのまま扱うと処理が困難になるため，文書ベクトルの次元を縮小して扱う必要がある．本研究では，性能の良い次元削減が報告されているランダムプロジェクションを用いて次元削減を行い，類似性判別の検証を行った結果についての考察を述べる．

キーワード 類似文書判定，ベクトルの次元削減，ランダムプロジェクション

A Study on Identifying Similar Documents based on the Dimension Reduction of a Document Vector

Kayano UMEZAWA† and Ichiro KOBAYASHI†

† Advanced Sciences, Graduate School of Humanities and Sciences, Ochanomizu University
2-1-1 Ohtsuka, Bunkyo-ku, Tokyo, 112–8610 Japan
E-mail: †{umezawa.kayano,koba}@is.ocha.ac.jp

Abstract As for the similar document retrieval technology, each document is usually expressed with a document vector. However, the number of the dimension of document vectors would become considerably large, as the number of target documents increases. Since it is quite hard to deal with high dimensional document vectors because of expensive calculation cost. Therefore, in this study, we reduce the dimension of a document vector using the random projection method. We show the result of an experiment and discuss it.

Key words identifying similar documents, dimension reduction of a document vector, random projection

1. はじめに

近年，大量のテキストデータがインターネット等を通じて蓄積され，アクセスして利用することが可能となった．テキストデータが増大するにつれ，必要な文書を検索する文書検索技術の要求が高まっている．たとえば，研究者は自己の論文の内容に類似する過去の論文を検索することが必須であり，企業においては自社の特許明細書に類似する他社の特許明細書の調査は必須と言える．このように，ある文書群（論文，特許公報，資料等）から，自己の文書と類似する文書を検索する方法が文書検索においての重要な課題となっている．

文書検索技術においては，テキストデータは文書ベクトル空間モデルとして表現され，検索対象テキストとのベクトルの類似性を測ることにより所望のテキストを探す [1]．しかし，文書ベクトルは検索対象となる文書内に含まれる語彙の数だけ次元を持つため，一般的に高次元ベクトルのデータとなる．高次元データをそのまま扱うと実時間応答が困難になるため，文書ベクトルの次元を縮小して扱う必要がある．これまでに LSI(Latent Semantic Indexing) などの次元圧縮手法 [2] [3] [4] や，主成分分析 (Principal Component Analysis) や LSI を改良し文書と単語の潜在的な共起関係を捉えることによる次元圧縮を行う pLSI(Probablistic LSI) [5] などが提案されている．本研究では，それらの手法と比較して性能の良い次元削減が報告されている [7] ランダムプロジェクション [6] を用いて，低次元数に文書ベクトルの次元削減をした場合の類似性判別の有効性について考察を述べる．

本稿では，2 章で関連研究を，3 章では類似文書判定処理を，4 章ではデータの次元縮小手法であるランダムプロジェクションについて説明する．そして 5 章では，実際のデータを用いてランダムプロジェクションで次元削減した場合の類似文書判定の有効性を検討する．最後に 6 章で本研究のまとめと今後の課題について述べる．

2. 関連研究

ランダムプロジェクションによる次元縮小の性能を調べた研

究としては，佐々木らによる研究 [7] や渡邉らによる研究 [8] がある．前者は， 1033 件の文書から取り出した 4329 個の索引語を要素とする文書ベクトルの次元削減を行った場合の検索の評価がなされており，LSI や pLSI など他の次元削減手法と比較してランダムプロジェクションの性能の良さを報告している．後者は，ランダムプロジェクションと線形計画法を組み合わせることで効率の良いアルゴリズムを得られることを報告している．

ランダムプロジェクションによる次元削減手法を導入し，文書検索を行った研究としては，大内らによるニューストリームなど時系列テキストに適用した研究がある [9] [10]．この研究では，ニューストリームの差分情報に対しランダムプロジェクションによる次元縮小を行い，増大する情報を効率的に格納し，検索性能や処理時間の向上を図っている．また，LSI との比較も行われており，テキストストリームの検索においてもランダムプロジェクションの方が優れていることも述べている．

本研究では，今後もテキストデータの蓄積が増大していくことに鑑み，より多数の文書を対象とした場合においても，効率的な類似文書検索がランダムプロジェクションによる文書ベクトルの次元縮小により実現可能であることを前提に，低次元数に文書ベクトルの次元削減をした場合でも，検索精度を保つことが可能かどうかを考察する．

3. 類似文書判定処理

図 1 に提案する類似文書判定の処理概要を示す．

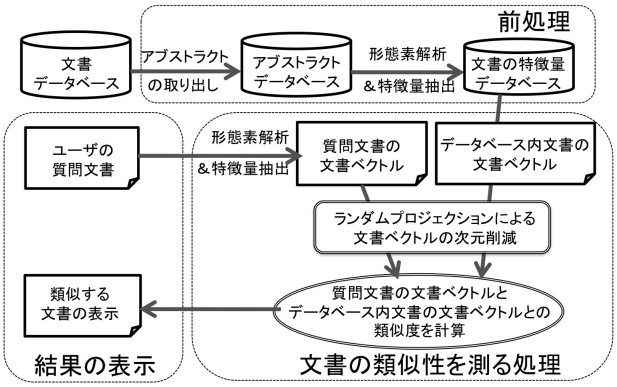

図 1　類似文書判定処理の流れ

大量の文書から各文書の特徴量を取り出して，文書ベクトルとして格納しておく．クエリとなる文書と対象となる文書における文書ベクトルとのコサイン類似度を求め，類似度の高い順に並べる．

ここで文書ベクトルとは，ある文書において，その文書の特徴を表す文書中の語彙を特徴語と定義し，特徴語の重要度を表す値である特徴量をもったベクトルで表される．そのため，文書ベクトルは特徴語の数に相当する高次元のベクトルとして表される．

3.1 形態素解析

特徴量抽出のため，クエリとなる文書と検索対象となる文書に対し，初めに形態素解析を行う．形態素解析とは，テキストを文法の最小単位 (形態素) に分割し，各形態素の品詞を特定する作業である．本研究では形態素解析を MeCab [11] を用いて行う．形態素解析の結果，得られた語彙のうち，文書群中の複数の文書に出現する，品詞が名詞，動詞である語彙を特徴語として選ぶ．名詞や動詞，形容詞は文書の内容を直接表現するものであり，その他の助詞などは文書を文法的に成立させるための機能語とされる [12]．なお，研究論文や特許公報の類似文書を検索する際には，品詞が名詞，動詞である語彙のみを特徴語として用いる．さらに，名詞の中でも非自立（「うち」,「ため」など），数（「１」,「2010」など），接尾辞（「個」,「年」など）は，名詞であっても文書の内容を直接表現するものではないため，対象から除く．

3.2 特徴語の重要度算出

特徴語の文書内での重要度を考慮して，特徴量を求めることが必要である．そのために，本研究では，$tfidf$ を用いる．$tfidf$ は, tf (Term Frequency) と idf (Inverse Document Frequency) の 2 つの指標を利用し，その積によって文書中の特徴語の重要度を計算する．tf は文書中における特徴語の出現頻度であり，文書中に多く現れる語ほど値が大きくなる．idf は全文書中に，特徴語が出現する文書数の逆数であり，出現する文書数が多い語ほど値が小さくなる．つまり，多くの文書に出現する特徴語より，一部の文書にのみ含まれている特徴語ほど idf 値は大きくなる．文書 d における特徴語 t の重要度である $tfidf(d,t)$ は，以下の式によって与えられる．

$$tf(d,t) = \frac{n_t}{W_d} \tag{1}$$

$$idf(t) = \log \frac{N}{w_i} + 1 \tag{2}$$

$$tfidf(d,t) = tf_{d,t} \times idf_t \tag{3}$$

n_t は単語 t の出現回数であり，W_d は文書 d における全特徴語数である．N は総文書数，w_i は単語 t を含む文書の数であり，対数の底を 2 とする．$tf(d,t)$ と $idf(t)$ の積により，$tfidf(d,t)$ が求まる．

さらに，文書の長さによる影響を調整するため，得られた文書 d の文書ベクトル $\mathbf{x}_d$ の値を正規化することも必要である．正規化する方法として，コサイン正規化を用いる．コサイン正規化では，各文書ベクトルのノルムを計算し，その文書ベクトルの各要素をノルムで割る．文書 d の文書ベクトル $\mathbf{x}_d$ のノルム $\|\mathbf{x}_d\|$ は，以下の式で表される．

$$\|\mathbf{x}_d\| = \sqrt{\sum {x_i}^2} \tag{4}$$

ノルム $\|\mathbf{x}_d\|$ の値で，文書 d の文書ベクトル $\mathbf{x}_d$ の各要素を割る．以上の計算により，各文書の特徴量を文書ベクトルで表せる．

3.3 類似度判定

各文書の類似の度合いを測るために，各文書の特徴量ベクトル同士のコサイン類似度を求める．コサイン類似度はテキスト処理で多用される類似度の指標である．文書 d_1 の特徴量ベクトル $\mathbf{x}$ と文書 d_2 の特徴量ベクトル $\mathbf{y}$ のコサイン類似度は以下の式で与えられる．

$$s_{\cos(\mathbf{x},\mathbf{y})} = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}} \tag{5}$$

コサイン類似度の値が大きいほど，ベクトルで表された文書同士の類似の度合いが大きいと判定される．

4. ランダムプロジェクション

ここでは，文書ベクトルの次元を縮小する方法として本研究で用いる，ランダムプロジェクション手法について述べる [6] [13] [14]．高次元のベクトル空間においては，直交ベクトルに近いベクトルが多く存在する．そのことから，ランダムな方向を持ったベクトルは，十分に直交ベクトルに近いと推測される．このことから，ランダムプロジェクション手法は，通常は正規直交系に座標を変換する射影行列を用いて次元を縮小するものを，要素をランダムに決定した行列 R に変更することにより計算コストを抑え，文書ベクトル行列 X を低次元の部分空間に射影するという次元縮小の手法である．

特徴語数 d，文書数 n の文書行列 $X_{d\times n}$ は，行列の大きさが d 行 n 列となり，それぞれの列ベクトルが 1 件の文書を表す．i 行 j 列の要素 x_{ij} は，文書 j における単語 i の正規化した $tfidf$ 値である．

縮小後の次元数を k とした場合，要素をランダムに決定したランダムプロジェクション行列 $R_{k\times d}$ を大きさ $k\times d$ となるように作成する．$d\times n$ の行列 X を，$k\times n (k \ll d)$ の行列 R に射影するためである．ランダムプロジェクション行列の i 行 j 列の要素 r_{ij} は，通常ガウス分布に従うように設定されるが，Achlioptas [6] によって以下の式で表される単純な独立した分布に置き換えることにより，計算効率の向上が図れることが示されている．

$$r_{ij} = \begin{cases} +1 & \text{確率 } 1/6 \\ 0 & \text{確率 } 2/3 \\ -1 & \text{確率 } 1/6 \end{cases} \tag{6}$$

行列 $X_{d\times n}$ の ランダムプロジェクション手法による次元縮小は，次の計算で行われる．

$$X^{RP}_{k\times n} = R_{k\times d} \times X_{d\times n} \tag{7}$$

この処理の計算量は $O(dkn)$ である [4]．すなわち，次元数を縮小するほど計算時間は短縮される．

検索の際には，次の計算により質問となるベクトルも低次元空間に射影して類似計算を行う．

$$q^{RP}_{k\times 1} = R_{k\times d} \times q_{d\times 1} \tag{8}$$

文書ベクトルとの類似度を計算し，検索結果として類似順位を決定する．

次元縮小による誤差は，ベクトル間のユークリッド距離に対して定義される．今，ϵ を $0 < \epsilon < 1$，n を整数として，k' を次のようにおく．ここで，k' は誤差との範囲で保証される縮小後の次元数である．

$$k' \geqq \frac{4+2\beta}{\epsilon^2/2 - \epsilon^3/3} \log n \tag{9}$$

また，行列 $X_{d\times n}$ から 行列 $X_{k'\times n}$ へのランダムプロジェクション行列 $R_{k'\times d}$ による写像を $f : X_{d\times n} \rightarrow X_{k'\times n}$ と表わすとする．そして，行列 $X_{d\times n}$ の任意の 2 つの列ベクトル $\mathbf{u}$ 及び $\mathbf{v}$ をとると，少なくとも $1 - n^{-\beta}$ の確率で，式 (10) を満たす [6].

$$(1-\epsilon) \parallel \mathbf{u} - \mathbf{v} \parallel^2 \leqq \parallel f(\mathbf{u}) - f(\mathbf{v}) \parallel^2 \leqq (1+\epsilon) \parallel \mathbf{u} - \mathbf{v} \parallel^2 \tag{10}$$

これは，式 (9) を満たす k' 次元に縮小した場合において，ある 2 つのベクトルのユークリッド距離が 1 $\pm\epsilon$ の誤差の範囲で保存されることを示して いる．

5. 実　　験

ここでは，類似文書の判定において有効な次元削減数を調べる実験を行う．まず，実験環境，使用する文書データの詳細，及び，実験結果の評価指標について説明し，実験結果を示すと共に類似検索の結果の評価，さらに有効な縮小次元数について考察する．

5.1 実 験 環 境

実験で使用する計算機の構成を表 1 に示す．

表 1　実験で使用する計算機の構成

OS	Mac OS X 10.6.4
CPU	2.53 GHz Intel Core 2 Duo
メモリ	4 GB 1067 MHz DDR3
開発言語	Ruby

5.2 用いたデータセット

実験には，評価型ワークショップ NTCIR [15] によって提供されているデータセット NTCIR-1 に含まれる論文を用いた．NTCIR-1 には，過去に発表された 339,501 件の論文のデータが収録されており，各論文の ID 番号，題目（日本語），題目（英語），著者名（日本語），論文アブストラクト（日本語），論文アブストラクト（日本語），キーワード（日本語），キーワード（英語），発表された学会名（日本語），発表された学会名（英語）等の情報を利用できる．1 論文が発表された学会は 1 種類である．実際に類似判定をする対象文書には，論文の内容を最も表している論文アブストラクト（日本語）を用いた．この中から論文アブストラクト 3,774 件を選んで実験に用いる．これらの論文は，26 種類の学会のいずれか 1 つの学会で発表されている．学会ごとの発表論文数は，無作為に選んでおり，論文数が最も少ない学会で 4 論文，最も多い学会で 774 論文である．論文アブストラクトを形態素解析して得た名詞，動詞は 9,939 個

であり, 一度しか出現しない語を除いて特徴語を選ぶと 6,109 個となった. これは, 文書ベクトル空間が 6,109 次元で表されることを示す. さらに, 6,109 個の特徴語からなる次元数に対して, ランダムプロジェクションによる次元削減後の精度の保証が示されている次元数を式 (9) より求める. $\epsilon = 0.1$ を式 (9) に代入すると式 (11) となり, この式において $\beta > 0$ となる k の最小値を 100 のオーダで求めると 2,100 となった.

$$k' \geqq \frac{4+2\beta}{0.1^2/2 - 0.1^3/3} \log 6109 \qquad (11)$$

検索の適合文書には, 質問文書と同じ学会で発表された論文アブストラクトを選ぶ.

5.3 評価方法

3,774 件 の論文アブストラクトから学会を 1 種類選び, さらに, その学会で発表された論文アブストラクトを無作為に 10 個選んで質問文書とする. 1 個の論文アブストラクトだけを質問文書とすると, 内容次第で検索結果が偏ると考えられるため, 質問文書は 10 個選んだ. そして, 10 個それぞれの論文アブストラクトを質問文書として, 様々な次元に削減した文書行列を用いた検索を 10 回行い, 11 点平均適合率の平均値を求めた. ランダムプロジェクションは次元削減のための射影行列の要素がランダムに決まり, 次元削減の度に結果も異なるため, 検索は 10 回行って平均値を求めることとした. 10 回の 11 点平均適合率の平均値を算出し, そこからさらに同じ学会で発表された 10 個の質問文書の平均値も求め, 精度評価を考察する.

5.4 結果と考察

質問文書は, 実験 (1) では論文数が最も多い学会「西日本支部大会」で発表された論文アブストラクト, 実験 (2) では論文数が 181 個(注1) である「電子通信用電源技術」で発表された論文アブストラクト, 実験 (3) では論文数が最も少ない学会「撮影分科会」で発表された論文アブストラクトを用いた.

各検索結果についての 11 点平均適合率の平均値を, 図 2, 図 3, 図 4 に示した. グラフの横軸は削減後の次元数を, 縦軸は 11 点平均適合率の値を表す. 式 (10) による精度の保証がされていない次元数 100 次元, 200 次元, 300 次元, 400 次元, 500 次元, 600 次元, 700 次元, 800 次元, 900 次元, 1,000 次元に次元削減して検索を行った結果と, 精度の保証がされている次元数 2,100 次元に削減して検索を行った結果, さらに次元削減しないで検索を行った結果についての 11 点平均適合率の平均値を表した.

その結果, 理論的に精度保証がされていない低次元に削減した場合であっても, 精度保証がなされている高次元の場合に近い類似判定の精度を保っていることが分かった. また, 低次元に削減するほど, 検索速度が高速になっていることを確認した.

ただし, 一部の結果においては, 理論的に精度保証がされている次元数に削減した場合であっても, 次元削減前の検索結果との誤差が理論値よりも下がるものも見られる. その理由としては, 式 (9) で示されるランダムプロジェクションによって保証される誤差の理論値は, あくまでも次元削減におけるユークリッド距離の保証をしているためであると考えられる. 類似文書精度も同じ程度の誤差で保証されるかどうかは, 作成されたプロジェクション行列の内容に依存する可能性があり, より詳細な調査が必要になる. このことに関連して, プロジェクション行列の内容を検索対象文書に合わせて作成する研究 [16] もなされている.

また, 図 2, 図 3, 図 4 に示す次元数削減前の検索結果と誤差の大きさは, 同じ学会で発表された論文アブストラクトの数による, トピック数の違いにも起因すると考えられる. 次元削減前の検索結果との誤差が理論値よりも大きく下がる実験 (1) や実験 (2) は, 同じ学会で発表された論文アブストラクトの数が多く, 同じ学会で発表された論文の中に多様なトピックの論文が存在すると考えられるため, 精度が低く見えると考えられる. 逆に, 実験 (3) は同じ学会で発表された論文アブストラクトの数が少なく, 同じ学会の中で発表された論文に少数のトピックの論文しか存在しないと考えられるため, 精度が高くなったと考えられる.

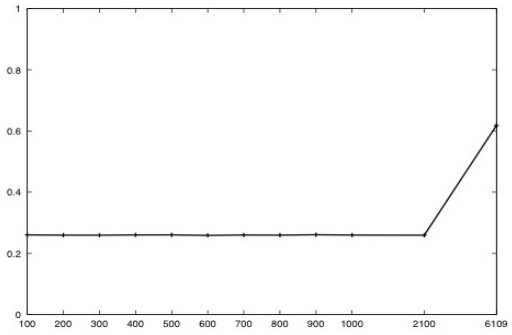

図 2 実験 (1) の結果

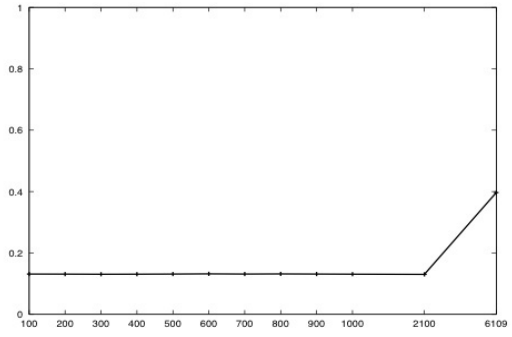

図 3 実験 (2) の結果

6. おわりに

類似文書の検索において, 検索精度をある程度保ちつつ様々な次元数の下で文書ベクトルの次元削減を試みる実験を行い, 類似性判別の検証を行う実験を行った. その結果, 理論的に精度保証がされていない低次元に削減した場合であっても, 精度

(注1): 選んだ論文アブストラクトの中で, 学会ごとの発表論文数の平均値に最も近い値が 181 であったため, 実験の質問文書に用いた.

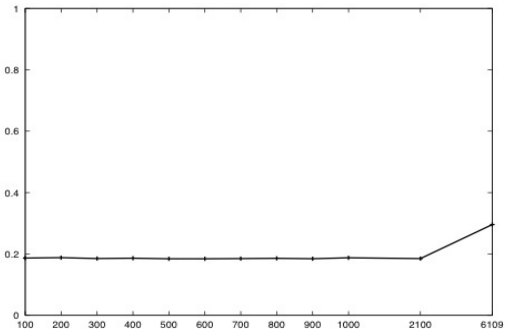

図 4　実験 (3) の結果

保証がなされている高次元の場合に近い検索精度を保っていると思われる結果を得た.

今後は, ランダムプロジェクションによる誤差の保証が, 類似文書判定の精度にどの程度の影響を与えるのかを詳しく調べる予定である.

文　　献

[1] 北 研二, 津田 和彦, 獅子堀 正幹: “情報検索アルゴリズム”, 共立出版, 2002.

[2] Deerwester, S. C., Dumais, S. T., Furnas, G. W., Landauer, T. K. and Harshman, R. A.: “Indexing by latent semantic analysis ”, journal of the American Society for Information Science, Vol 41, No. 6, pp. 391-407, 1990.

[3] 大内 浩仁, 三浦 孝夫, 塩谷 勇: “多義性を考慮した文書検索”, データ工学ワークショップ (DEWS), 電子情報通信学会データ工学研究会, 2003.

[4] Papadimitriou, C. H., Raghavan, P., Tamaki, H. and Vempala, S.: “Latent semantic indexing: A probabilistic analysis ”, In Proc. 17th ACM Symp. on the Principles of Database Systems, pp 159-168, 1998.

[5] T. Hofmann: “Probabilistic Latent Semantic Analysis”, Proc. Uncertainty in Artificial Intelligence, 1999.

[6] Achlioptas, D.: “Database-friendly random projections ”, In Proc. ACM Symp. on the Principles of Database Systems, pp 274-281, 2001.

[7] 佐々木稔, 北研 二: “ランダム・プロジェクションによるベクトル空間モデルの次元削減”, 自然言語処理, Vol.8, No.1, 2000.

[8] 渡邉辰也, 瀧本英二, 丸岡章: “ランダムプロジェクションによる次元圧縮”, 電子情報通信学会技術研究報告. COMP, コンピュテーション 101(707), 73-79, 2002-03-04.

[9] 大内浩仁, 三浦孝夫, 塩谷勇: “ランダムプロジェクションを用いたニュースストリームの検索”, 日本データベース学会論文誌 (DBSJ Letters) Vol.3, No.3, 2004, pp.1-4.

[10] 大内浩仁, 三浦孝夫, 塩谷勇: “ランダムプロジェクションによるテキストストリームの検索”, データ工学と情報マネジメントに関するフォーラム (DEIM), 2004.

[11] Mecab,http://mecab.sourceforge.net/

[12] 車: “R 嘉呼估ζ篤蕉”, 俎, 2008

[13] Bingham, E. and Mannila, H.: “Random projection in dimentionality reduction: Applications to image and text data ”, Proc. 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2001), pp 245-250, 2001.

[14] Kaski, S.: “Dimensionality reduction by random mapping: Fast Similarity Computation for Clustering ”, In Proc. Int. Joint Conf. on Neural Networks (IJCNN), Vol 1, pp. 413-418, 1998.

[15] http://research.nii.ac.jp/ntcir/index-en.html

[16] 大内浩仁, 三浦孝夫, 塩谷勇: “頻度分布に基づくプロジェクションを用いた文書検索”, データ工学ワークショップ (DEWS), 2005.