文書ベクトルの次元削減に基づく有効な類似文書判定

梅澤香矢乃　(指導教員：小林一郎)

1　はじめに

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，テキストのベクトル同士の類似性を測ることにより所望のテキストを探す．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，ベクトルの次元を削減して扱う必要がある．本研究では，これまでに研究されてきた LSI(Latent Semantic Indexing) や，pLSI(Probablistic LSI) など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 [1] ランダムプロジェクションを用いて，低次元数に文書ベクトルの次元削減をした場合の類似文書判定の有効性について考察を述べる．

2　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．

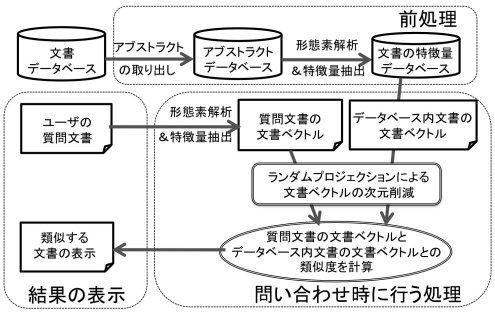


図 1: 類似文書判定処理の流れ

大量の文書から各文書の特徴を取り出して，文書ベクトルとして格納しておく．質問となる文書と検索対象となる文書の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．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ある．

2.1　形態素解析

特徴量抽出のため，質問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．本研究では，形態素解析器 MeCab[2] 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，品詞が名詞または動詞である語彙を特徴語として選ぶ．

2.2　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 $tfidf$ 法を用いる．$tfidf$ 法は，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．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．

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

$$idf(t) = \log \frac{W}{w_t} + 1 \tag{2}$$

$$tfidf(d,t) = tf(d,t) \times idf(t) \tag{3}$$

n_t は特徴語 t の出現回数であり，N_d は文書 d における全特徴語の出現回数である．W は総文書数，w_t 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 $\mathbf{x}_d$ の値をコサイン正規化する．コサイン正規化では，ベクトルのノルムを計算し，ベクトルの各要素をノルムで割る．ベクトル $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ のノルム $\|\mathbf{x}\|$ は，式 (4) で表される．

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

以上の計算により，各文書を文書ベクトルで表せる．

2.3　類似度判定

各文書の類似の度合いを測るために，各文書の文書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 の文書ベクトル $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ と文書 d_2 の文書ベクトル $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ 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．

3　ランダムプロジェクション

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．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削減するものを，要素をランダムに決定した行列 R を射影行列と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削減の手法である．

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，ランダムプロジェクション行列 $R_{k\times d}$ は要素をランダムに決定し，大きさ $k \times d$ の行列となるように作成する．$d \times n$ の行列 $X_{d\times n}$ を，$k \times d (k \ll d)$ の行列 $R_{k\times d}$ に射影するためである．射影行列の 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，式 (6) で表される単純な独立した分布に置き換え，計算効率の向上が図れることが示されている [3].

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

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削減は，式 (7) によって行われる．

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

次元数を削減するほど計算時間は短縮される．

検索の際には，式 (8) 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検索を行う．

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

次元を削減した検索対象文書の文書ベクトルとの類似度を計算し，その値から類似順位を決定する．

次元削減による誤差は，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削減後の次元数である．

$$k' \geq \frac{4+2\beta}{\epsilon^2/2 - \epsilon^3/3} \log n \quad (9)$$

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 $f : X_{d\times n} \to X_{k'\times n}$ と表すとする．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 [3].

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

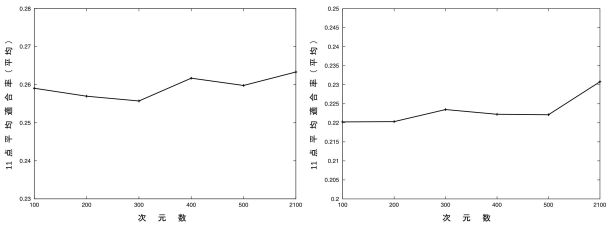
これは，式 (9) を満たす k' 次元に縮小した場合において，任意の 2 つのベクトルのユークリッド距離が $1\pm\epsilon$ の誤差の範囲で保存されることを示している．

4 実験

類似文書の判定において本手法の有効性を調べる実験を行う．検索対象の文書には，評価型ワークショップ NTCIR [4] によって提供されているデータセット NTCIR-1 の論文データ 339,501 件からランダムに選んだ 3,774 件の論文アブストラクト（日本語）を用いた．各論文は 26 の学会のうち，いずれか 1 つの学会で発表されている．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 個あり，その中から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．

4.1 性能評価

性能評価の指標として検索結果の 11 点平均適合率を算出し，類似文書判定が適切かどうか評価した．質問文書に対する適合文書は，次元削減せずに同じ質問文書で検索をして，類似度が高いと判定された順に並べた文書群とする．



(a) の場合　　(b) の場合

図 2: 実験の結果

4.2 結果

質問文書は，実験 (a) では論文数が最も多い学会で発表された論文アブストラクト，実験 (b) ではその次に論文数が多い学会で発表された論文アブストラクトを用い，検索結果についてそれぞれ 11 点平均適合率を求める．検索は 5 回繰り返し，その結果についての 11 点平均適合率の平均値を図 2 に示した．グラフの横軸は削減後の次元数を，縦軸は各次元削後の類似文書判定結果についての 11 点平均適合率の平均値を表す．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として，式 (9) より求めた 2,100 次元に削減した場合[1] の値を示した．さらに，精度の保証がされない次元数 100 次元，200 次元，300 次元，400 次元，500 次元に削減した場合の値も示した．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判定の精度を保っていることが分かった．そして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．

5 おわりに

今回は，類似文書の判定において，ランダムプロジェクションを用いて，検索精度をある程度保ちながら検索速度を向上させる手法を試した．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が出た．

今後は，より一般的な次元削減と類似文書判定精度の関係を考察するために，検索対象となる文書の種類や数を増やして実験するつもりである．

参考文献

[1] 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

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

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

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

[1] このとき $\epsilon = 0.1$ とし，$\beta > 0$ となる 100 のオーダでの最小の k の値を求めた．