文書間の相関の可視化による文書検索支援

5Q－6

林 一成　岩佐 英彦　竹村 治雄　横矢 直和
奈良先端科学技術大学院大学　情報科学研究科

1　はじめに

近年，大量の文書データベース中からの文書検索の研究が数多く行われている．最も古典的な方法であるキーワード型検索法は，欲しい文書に関連すると思われるキーワードの入力に対して，キーワードを含む文書をデータベースから取り出し，その一覧を表示する検索方法である．ユーザが欲しい文書集合に必ず含まれるようなキーワードを見つけることが比較的容易であるため，再現率 (正解総数に対する検索結果として取り出した正解数の割合) の高い方法であるといえる．しかし，簡単に思いつくような単語をキーワードとすると，検索結果の文書集合が非常に大きくなり，不要な文書が多く含まれてしまうことがしばしばあり，適合率 (検索結果の総数に対する正解数の割合) が低くなってしまう．そのとき，ユーザは再現率を維持したまま適合率をあげるための検索式を新たなキーワードを使って作成する必要がある．しかし，そのような検索式を書くには，検索対象に対する専門的な知識や書くための技術が必要となり，一般のユーザにとっては難しい作業である．結果として，ユーザが大きな文書集合を個々に内容を確認していくことが必要となり，非常に手間がかかる．

そこで本稿では，数量化 IV 類という統計手法を用いて文書間の類似度をユークリッド距離に変換することによって文書集合を 2 次元平面上に可視化する手法を提案し，可視化結果を用いた文書の絞り込みを試みる．

2　可視化による文書検索支援法

2.1　文書ベクトルの作成と文書間の類似度計算

本節では，文書を文書ベクトルとして表現し，文書間の類似度を計算する方法について説明する．Salton[1] は，対象になる文書集合に含まれる全単語数を N とするとき，文書 D_i を文書ベクトル d_i として表現した．

$$d_i = (w(i,1), w(i,2), \ldots, w(i,N))$$

ここで，$w(i,k)$ は文書 D_i に対する単語 W_k の重みで，次式で定義される．

$$w(i,k) = tf(i,k) \cdot idf(k)$$

Document Retrieval Support Based on Visualization of Correlation among Documents
Kazushige Hayashi, Hidehiko Iwasa, Haruo Takemura and Naokazu Yokoya
Nara Institute of Science and Technology (NAIST)
8916-5 Takayama, Ikoma, Nara 630-0101, Japan.

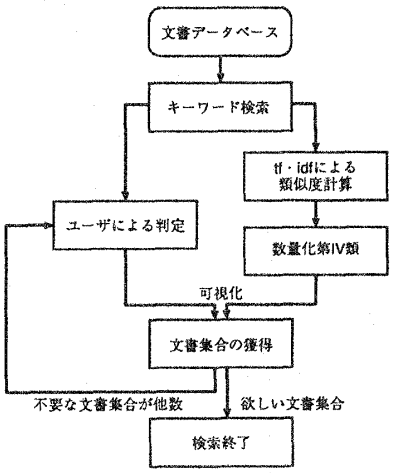

図 1: 可視化による文書検索支援手順

ここで用いた $tf \cdot idf$ 値は文書に対する単語の特徴量を表した値で，次式で定義される．

$$tf(i,k) = (\text{文書 } D_i \text{における単語 } W_k \text{の出現頻度})$$

$$idf(k) = \log\left(\frac{\text{全文書数}}{\text{単語 } W_k \text{が現われる文書数}}\right) + 1$$

すなわち，特定の文書中に集中して出現する単語は大きい値を，逆に多くの文書にまんべんなく出現する単語は低い値を取る．文書 D_i と D_j の類似度 $sim(D_i, D_j)$ は，文書ベクトル間のなす角度として次式で定義される．

$$sim(D_i, D_j) = \frac{\sum_{k=1}^{N} w(i,k) \cdot w(j,k)}{\sqrt{\sum_{k=1}^{N} w(i,k)^2 \cdot \sum_{k=1}^{N} w(j,k)^2}}$$

2.2　数量化 IV 類を用いた可視化手法

文書の絞り込みを行うためには，前節で計算した文書間の類似度をユーザが直観的に把握できる必要がある．そこで，本稿では文書間の類似度の値をユークリッド距離として表現して，多次元空間内に文書を配置するために数量化 IV 類の手法を用いる．

数量化 IV 類 [2] とは，n 個の個体に対して i 番目の個体と j 番目の個体との相互の類似度を表す何らかの数量 e_{ij} $(i, j = 1, 2, \ldots, n)$ が与えられているとき，その個体を多次元空間に，類似度が高い個体間の距離が小さく，類似度が低い個体間の距離が大きくなるように，各個体を多次元空間に配置する方法である．

本稿では，数量化IV類を用いて，得られるn次元空間の中から2次の部分空間を選択し，文書ベクトルを2次元平面上に配置して可視化する．

2.3 可視化を用いた文書検索支援手法

本節では数量化IV類に基づいて，2次元平面上に可視化された結果を用いて，文書検索を行う手法について述べる．検索手法の概略を図1に示す．

はじめに，ユーザは欲しい文書に関連するキーワードを入力して文書検索を行う．その検索結果内のいくつかの文書をユーザが読んで，欲しい文書を正解文書サンプル，不要な文書を不正解文書サンプルとそれぞれ指定する．そして，文書集合を正解サンプル，不正解サンプル，それ以外の文書の3種類に分けて，2次元平面上の点として可視化する．可視化は数量化IV類で求められた，多次元空間内から選択された複数の2次元平面に対して行う．

次に，正解文書サンプルが集中して分布し，かつ不正解文書サンプルが正解サンプルの分布領域以外に分布している2次元平面を選択する．これは，正解文書サンプルが集中している近傍には，それらと類似しているユーザが欲しい文書が集まっていると考えられるためである．

続いて，ユーザは正解文書サンプルの近傍の点として表示された文書集合内の文書を読んで，欲しい文書集合かどうか確認を行う．欲しい文書集合であると判断できたときは，検索を終了する．不要な文書が多く含まれていると判断したときは，再度ユーザが正解・不正解文書サンプルの指定を行い，サンプルが理想的な分布になっている2次元平面を，提示された中から探す．欲しい文書集合が得られるまで，この過程を繰り返す．

以上の手法により，専門的な知識がなくとも，ユーザは文書の2次元平面の選択と提示された文書が自分の欲しい文書かどうかの判断から，欲しい文書集合を容易に得ることができると考えられる．

3 類似文書検索実験

提案手法による文書検索実験を行った．文書のベクトル化に必要な形態素解析は茶筌[3]を使って行い，全文書中の全名詞2278語を抽出した．そして，「画像」という語が含まれる論文30件に対して本手法を用いた.「血管」に関する文書を抽出するために，「血管」という単語が含まれる論文3件を正解文書サンプルとし，内容的に正解文書と離れていると思われる論文3件を不正解文書サンプルとして,「血管」に関する論文が抽出できるか実験した．正解文書サンプル間の類似度は，正解文書サンプルと不正解文書サンプルとの類似度に比べて高かった．

図2の可視化結果は，数量化IV類での第1軸と第2軸から計算された2次元空間である．正解文書サンプルと不正解文書サンプルが同じ領域に分布しているために，文書集合の絞込みが行えなかった．

図3の可視化結果は，数量化IV類での第5軸と第10軸から計算された2次元平面である．図2に比べて，正解文書サンプルが集中して分布し，不正解文書サンプルが正解文書サンプルの分布領域以外の領域に分布しており，ユーザが欲しい文書に関する類似度が距離に反映された2次元平面である可能性が高いと考えることができる．内容を確認した結果「血管」という単語が含まれる文書が，正解文書サンプルと近い距離にあることが確認された．しかし，近い距離に「血管」という単語を含まない論文も発見された．なお，様々な2軸について検討した結果，図3のように正解・不正解が分離された空間は非常に少なく，そのような組合せを発見することは困難であった．

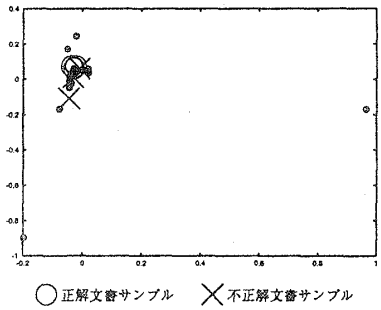

図 2: 1,2軸から成る2次元平面における文書の分布

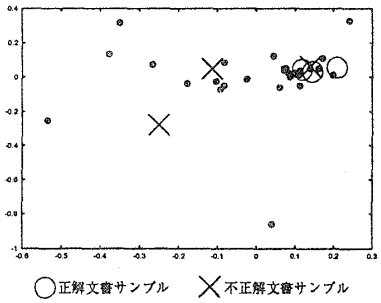

図 3: 5,10軸から成る2次元平面における文書の分布

以上の結果は，文書集合を2次元平面上の点として可視化する手法では，多様な文書の類似性を判断することは困難であることを示していると考えられるため，可視化手法を改良する必要がある．

4 まとめと今後の課題

本論文では，文書ベクトルを元にした文書間の相関の可視化結果を表示することにより，キーワード型検索法での検索結果に対して，絞り込みを直感的に行える文書検索支援を試みた．実験結果からは，文書の大半が2次元平面内の同一領域に集中して存在する場合が非常に多くなり，2次元平面を用いた可視化では不十分であることが確認された．今後は3次元空間や，色，アニメーションの利用など可視化手法の改良を行いたい．また，ユーザによる判定結果から予測される，ユーザの求める文書集合に関連する重要語を探し出し，その出現頻度を強く類似度に反映させる方法を考えたい．

参考文献

[1] Salton,G. and Allen,J.: Text Retrieval Using the Vector Processing Model, Proc 3rd Annual Symposium on Document Analysis and Information Retrieval, 1994.

[2] 多変量解析入門 II: 河口至商，森北出版株式会社

[3] 松本裕治，北内啓，山下達雄，平野善隆，今一修，今村友明:" 日本語形態素解析システム『茶筌』version1.5 使用説明書 " , 1997.