大規模文書検索結果のクラスタリングと可視化

小田　良治　　土井　章男
岩手県立大学ソフトウェア情報学部ソフトウェア情報学科

1. はじめに

大規模な文書集合を対象とした検索と，その結果の可視化について述べる．ユーザからの情報要求は自然言語の文や索引語（キーワード）などで与えられるものと仮定している．

検索対象となる複数の文書を文書集合（document collection），自然言語文などで記述されたユーザの情報要求を検索質問（query）と呼ぶ．

従来の一般的な検索結果の表示手法は，検索質問を入力することで，文書のランク付けを行い，ランク順に文書（タイトルや内容の一部等）をリスト表示していた．この手法の場合，検索結果が膨大になった場合に，上位の検索結果しか見ることができずに終わってしまうことが多い．

そこで，本研究では，検索結果全体を見渡し，全体の特徴を視覚的に把握しながら検索を行うために，検索結果の可視化を行った．

関連研究としては，文書集合を多次元ベクトル空間として２次元への可視化を行った研究[1]，文書情報を可視化することにより，検索質問（検索式）の作成支援を行った研究[2]，goo で実験的に行われている 3D 検索[3]などがある．

本研究は，文書間の類似度を用いて階層型のクラスタ分析を行うことで文書集合を組織化し，検索結果を３次元空間上で可視化を行う．

2. システム設計

2.1. ベクトル空間モデル

文書集合の検索にはベクトル空間モデル（vector space model）を用いた[4]．

ベクトル空間モデルでは，索引語の重みを要素とするベクトルで文書を表現する．いま，検索対象となる文書を d_1，d_2，…，d_n とし，これら文書集合全体を通して全部で m 個の索引語があるとする．このとき，文書 d_j は，次のようなベクトルで表現されることになる．これを文書ベクトル（document vector）と呼ぶ．

$$d_j = \begin{bmatrix} d_{1j} \\ d_{2j} \\ \vdots \\ d_{mj} \end{bmatrix}$$

2.2. 処理の流れ

（１） 文書集合を準備する．文書に特有のタグがある場合は，この時点で取り除いておく．

（２） 準備した文書集合に対して形態素解析を行う．形態素解析には Chasen[5]を用いた．そして，Chasen の出力結果から名詞だけを抜き出し，索引語として利用する．

（３） 索引語を用いて，文書ベクトルを作成する．索引語の重み付けには TF・IDF 値を用いた．

（４） 文書ベクトルは非常に巨大なベクトルであり，値が０の要素が多いという特徴があるため，そのまま利用するのは効率的ではない．そのため，文書ベクトルの次元削減を行う．次元削減には coveigen[6]を用いた．

（５） 文書の検索では，与えられた検索質問をベクトルへと変換し，各文書ベクトルとの内積を計算して類似度を求める．

（６） 検索結果のクラスタリングには階層型クラスタ分析手法の一つである最長距離法を用いた．文書間の類似度をユークリッド距離で定義し，距離が近い文書から順番にクラスタリングを行った．

（７） 検索結果の可視化にはデータ宝石箱[7]を用いた．データ宝石箱は階層型のデータ全体を一望することができるツールである．

3. 適用例

本研究では NTCIR-1[8]の文書集合を用いてシステムのテストを行った．NTCIR は情報検索の検索性能を評価するために用いる検索実験用のセットであり，（１）文書データベース，（２）

Clustering and Visualization for Search Result of Large Document Collections
† Ryoji Oda and Akio Doi, Iwate Prefectural University Faculty of Software and Information Science

利用者の検索要求を記述した検索課題群，（３）各検索課題に適合する正解文書の網羅的なリストからなる．

本研究では，NTCIR から正解文書を含んだ文書集合を意図的に作成して検索を行った．

3.1. 可視化結果

文書数３００件の文書集合を準備した．この文書集合には，「テキストからのキーワードの自動抽出について知りたい」という検索質問に対する５０件の正解文書が含まれている．この文書集合から得られた索引語の数は２１７３語であった．これを用いて文書ベクトルを作成し，その文書ベクトルの次元数を１０００次元に削減した．これを用いて検索と可視化を行ったのが図１である．

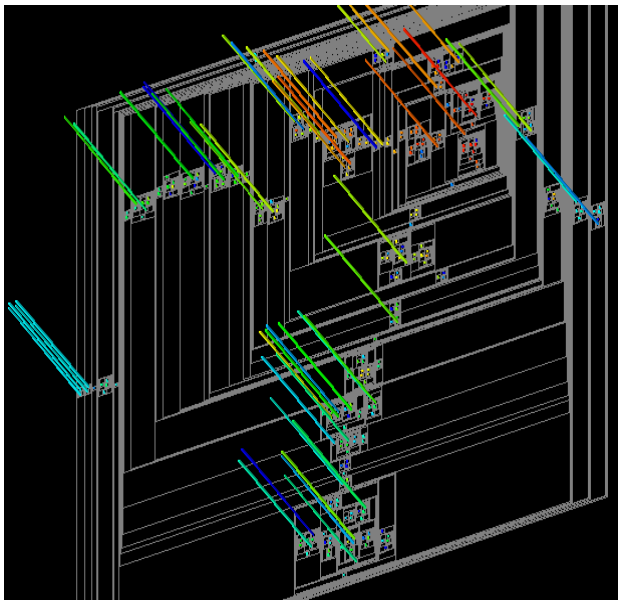

図１

この図では，検索結果の順位と各文書のクラスタリング結果が可視化されている．

各点（または棒）は文書を表している．それらを取り囲んでいる長方形の枠はクラスタを表している．

色は順位を表している．青いものほど順位が高く，赤いものほど順位が低い．

高さは文書の正解／不正解を表している．正解文書の場合は高さがあり，不正解文書の場合は高さがない．

3.2. 考察

色データの分布を見ると，順位が高い文書と順位が低い文書が大雑把ではあるが分かれていることがわかる．しかし，高さデータも考慮してみると，この色データの分布は検索結果として得られたデータが反映されてはいるが，本質的な意味での正解／不正解によって文書同士がクラスタリングされているわけではないということがわかった．

4. おわりに

本研究では，文書間の類似度や検索によって得られる順位などを基にして検索結果を可視化した．これによって，全体の特徴を直感的に感じ取ることが可能だとわかった．

今回のクラスタリングでは，すべての文書が強制的にクラスタリングされてしまうという問題がある．これでは，それほど似ていない文書同士がクラスタリングされてしまう可能性もある．そのため，より柔軟性の高いクラスタリング手法の導入を検討したい．

さらに，順位（類似度），クラスタリング結果以外の情報をユーザに提示するために．今後は文書ベクトルの可視化も行ってみたい．

参考文献

[1] 渡辺正裕，石川佳治，吉川正俊，植村俊亮：多次元ベクトル空間の視覚的探索機能を有する情報検索，http://www.kde.is.tsukuba.ac.jp/~ishikawa/publications/papers/9607watanabe.pdf

[2] 林一成，岩佐英彦，竹村治雄，横矢直和：文書情報の可視化による検索絞込み支援，2000 年電子情報通信学会研究報告　パターン認識・メディア理解, Vol.99, No.263, pp.15-20, March 2000.

[3] 3D ウェブ検索，http://goo.ntt-infolead.net/

[4] 北研二，津田和彦，獅々堀正幹：情報検索アルゴリズム

[5] 形態素解析システム茶筅，http://chasen.aist-nara.ac.jp/

[6] 青野雅樹，小林メイ：ベクトル空間モデルに基づく次元削減による大規模文書データの検索と可視化

[7] 伊藤貴之，山口裕美：「データ宝石箱」～ビジュアルデータマイニングの実現に向けて～

[8] NACSIS Test Collection for Information Retrieval systems：日本語情報検索システム評価用テストコレクション