自己組織化マップに基づく文書検索 *

福住 悟† 井越 昌紀‡
東京都立大学 工学研究科

1. はじめに

近年、ワープロをはじめとする入力環境の整備に加え，電子メールなどにより文書の電子化が急速に進んでいる。また，インターネットの普及により我々は膨大な量の文書データにアクセスすることが可能となってきた。こうしたことを背景に、大量のデータの中から有用な情報を発見し、抽出する技術の需要が高まっている。そのため、専門的な知識や技術を持たなくても必要な情報を探し出せる、使いやすい検索方法を実現することが重要となっている。

現在、文書データベースに対して用いられている検索手法は、WWW のサーチエンジンに代表されるように、カテゴリやキーワード等、あらかじめ文書に付与しておいた二次的情報を用いて AND／OR 演算を行い、結果のリストを得るものである。しかし、この方式では類義語や同義語等は考慮されない。既存の研究としてシソーラスを利用した検索式の拡張[1]などが行われてきたが、通常は固定であるシソーラスは、観点によって構造が変化できない。

また、結果が一次元的なリストとしてしか示されないため、文書群の関係を把握することが困難であり、必要な情報の探索に適しているとはいい難い。加えて従来の検索結果リストは基準が不明であったが、根拠となる単語間の関係を視覚化できれば、人間が考える指針になる。

そこで本研究では、文書群の傾向をつかみながら観点に基づいて柔軟な検索を行う手法を提案することを目的とする。そのためにニューラルネットモデルのひとつである、自己組織化マップ（Self-Organizing Maps：ＳＯＭ）[2]を利用する。上記目的の達成のために、次の四つの機能を実現することにした。

(1) 同義語、類義語、関連語によって検索が可能なこと
(2) 単語間の関係を観点に基づいて変化できること
(3) 単語間の関係を視覚化すること
(4) 検索結果を二次元のマップで表示すること

これらの機能を持つシステムを実装し検証を行った。

2. システム概要

Fig.1 にシステムの概要を示す。

2.1 単語辞書ＳＯＭ

単語の概念は広辞苑等の国語辞典の語義説明文から獲得できるものとし、また語義説明文中にある品詞の中で、名詞が最も主要な情報を表すという仮定の下に、ベクトル空間モデルを作成する手法を考案した。Fig.2 にその概要を示す。

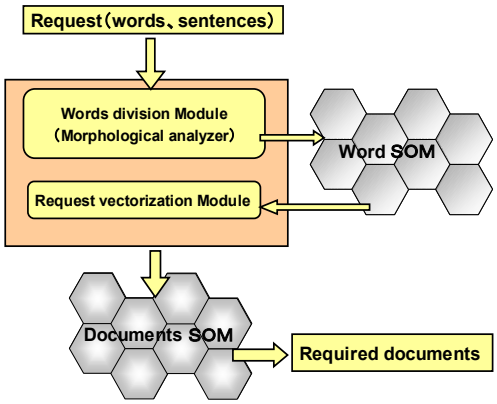

Fig.1 System Overview

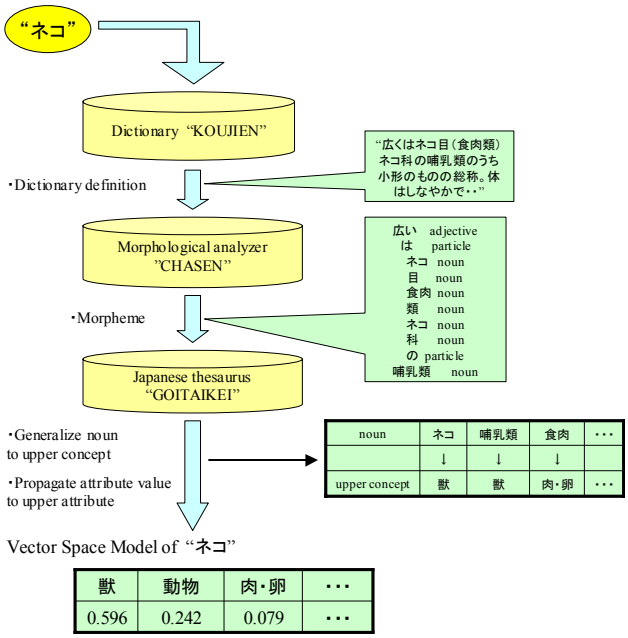

Fig.2 Vector Space Model of Word ”ネコ”

例えば単語「猫」の概念をベクトル空間モデルで表す際に、システムは広辞苑の CD-ROM から「猫」の語義説明文を取り出してきて、その説明文を形態素解析エンジン「茶筌」[3]によって形態素に分解する。分解された形態素から名詞を取り出す。次いで日本語語彙体系[4]の一般名詞意味属性体系を用いて、それぞれの名詞の上位概念を見つけ、汎化する。この操作によって単語「猫」は基底属性数 2715 次元のベクトル空間モデルであらわされる。ベクトルの各属性値は、基底属性の出現頻度によって重み付けされる。

属性の基底として日本語語彙体系シソーラスを用いているので、属性には階層性があることになる。しかし実際には語義説明文の表記ゆれ等の影響によって、そうならないことがあるため、一定の割合で下位属性から上位

---

* Document Retrieval based on Self-Organizing Maps
† Satoru FUKUZUMI, Tokyo Metropolitan University, Graduate school of Engineering
‡ Masanori IGOSHI, Tokyo Metropolitan University, Graduate school of Engineering

属性へと属性値を伝播させる。

ベクトル空間モデルで表された単語を、主成分分析初期化ＳＯＭアルゴリズム[5]にかけ、単語辞書ＳＯＭを作成する。主成分分析初期化ＳＯＭは、通常のＳＯＭと比べ、少ない学習回数で量子化誤差（ＭＱＥ）が収束することがわかっている。

単語辞書ＳＯＭは、2.2、2.3 で述べる文書群ＳＯＭや、検索質問のベクトル表現に影響するため、特定の属性に重みをおいた単語ＳＯＭを作成することで、観点を含めることができる。これは機能(2)に当たる。また単語間の関係はＳＯＭによって視覚化されているので、システムを利用した際に、その出力の根拠を追跡することができる。これは機能(3)に当たる。

2.2　文書群ＳＯＭ

文書特徴ベクトルの属性となるものは 2.1 で構築した単語辞書ＳＯＭの各ユニットであり、ユニットの個数と同数の属性数をもつ。属性値はそれぞれのユニットに適合している単語が文書中に現れる頻度、つまりヒストグラムによって重み付けされる。

まず、文書に含まれる単語で、単語辞書ＳＯＭに含まれるものを取り出す。ある単語が取り出されたとき、取り出された単語が適合しているユニットに該当する属性に重みをあたえる。また、その近傍にあるユニットに該当する属性にもある程度の重みを与える。

このことによって、同義語や類義語による表記ゆれに対して頑強な知識ベースＳＯＭを作成できると考えられる。これは機能(1)に当たる。出来上がった文書特徴ベクトルを主成分分析初期化ＳＯＭにかけ、文書群ＳＯＭを作成する。

2.3　検索質問と文書出力

検索したい文を入力することにより、構築した単語辞書ＳＯＭを通して検索ベクトルが作成される。検索ベクトルは文書群ＳＯＭと比較され、適合度によって文書群ＳＯＭの色が塗り分けられる。ユーザは文書群ＳＯＭで適合度が最も高い位置を選択することで最適合の文書を、またその周辺を選択することで関連があると思われる文書を検索することが可能となる。このように検索結果が二次元マップ上で表示されることで、文書群の分類の全体像を概観することができる。また二次元の広がりは何らかの方向を持っているため、単に適合度の違いだけはでなく、違いの方向性を表現できることになる。これは機能(4)に当たる。

3.シミュレーション

今回用いたサンプルは、BMIR-J2 テストパターン[6]に含まれる基本機能 F1 のテストに必要とされる、毎日新聞 CD-ROM’94 年版の新聞記事 551 件である。

551 記事に含まれる単語（自立名詞、動詞・形容詞の転生名詞）7729 語を 2715 次元の単語意味ベクトルで表し、40×34 ユニットの単語辞書ＳＯＭを構築した。その一部を Fig.3 に示す。

続いて構築した単語辞書ＳＯＭを基に 551 の記事を 40×34=1360 次元の特徴ベクトルで表し、20×19 ユニットの文書群ＳＯＭを構築した。

ここでテストパターンの検索質問の一つである「農薬」を入力した。文書群ＳＯＭの彩色の様子を Fig.4 に示す。また、検索質問「農薬」に対して「検索されるべきである」とされている文書が、文書群ＳＯＭ上でどのユニットにマップされているかを円柱で示した。円柱の高さはそのユニットにマップされている「検索されるべきである」文書の数を表す。

Fig.4 にて、白いユニットは検索質問「農薬」との適合度が高かったユニットを示す。この例では適合度が高いユニットの周辺に、テストパターンにて「検索されるべきである」文書が集まっている様子が伺える。

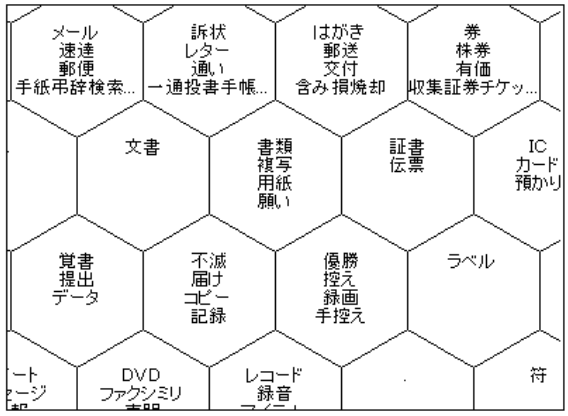

Fig.3 Word SOM

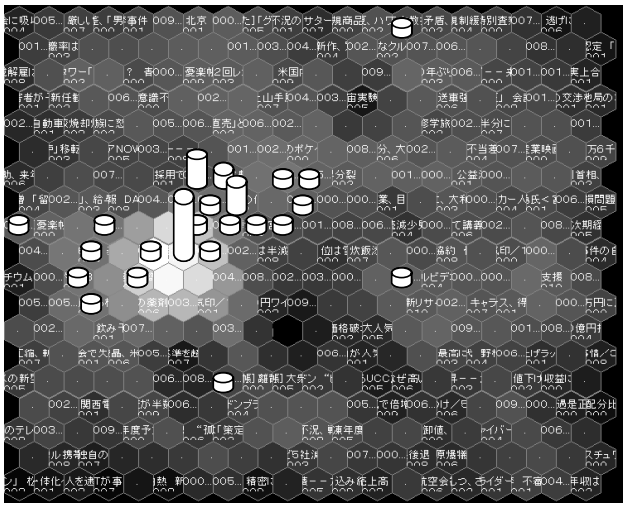

Fig.4 Documents SOM

4.結論

・ＳＯＭの可視性を利用した情報検索について考察した。
・テストパターンを用いて有効性を一部確認した。

参考文献

[1]　栗山　和子“シソーラスを用いた検索式拡張の評価”　情報処理学会研究報告 98-FI-52, pp.1-8 ,1998.

[2]　T.コホネン著　徳高平蔵他　訳　“自己組織化マップ”　シュプリンガー・フェアラーク東京　1996

[3]　日本語形態素解プログラム　茶筅
http://cactus.aist-nara.ac.jp/lab/nlt/chasen/distribution.html

[4]　日本語語彙体系
http://www.kecl.ntt.co.jp/icl/mtg/resources/GoiTaikei/index.html

[5]　福住　悟、井越　昌紀　“主成分分析初期化ＳＯＭの評価”　2002 年度精密工学会春季大会講演論文集 p183 ,2002

[6]　木谷　強、石川　徹也、木本　晴夫　ほか “日本語情報検索システム評価用テストコレクション BMIR-J2”　情報処理学会研究報告　98-DBS-114, pp.15-22, 1998