2L－1

大量テキストのネットワーク表現と可視化手法の検討

豊浦 潤　津高 新一郎　小中 裕喜
RWCP 情報ベース機能三菱研究室

1　はじめに

大量文書を扱うための新たな情報処理技術の必要性が年々高まっている．我々は，この問題に対する1つの答えとして，ユーザが大量文書を鳥瞰しながらアクセスできる情報散策システム：3D-Stroller をこれまでに開発した[1]．情報散策システム中では，最初にいくつかの検索キーワードを与えて，これにマッチする文書を文書はキーワードの出現に基づくベクトル空間モデルで表現され，キーワードの共起など統計情報に基づき自己組織化マップ上に自動分類される．この自己組織化マップは3次元に可視化され，キーワードと文書がユーザに提示される．ユーザはこれらの操作を通じ，はっきりとした検索目的を持たなくても発見的に興味ある情報を得ることができる．

しかし，ここ数年のインターネットの普及などによりアクセス可能な文書の量は爆発的に増大している．これらは，同一の情報ソースからの編集や引用を持つ文書を多く含んでいるため，同一キーワードで検索される大量の文書は，キーワードの出現頻度の統計情報だけでは，妥当な空間配置を行なうために十分な分解能が得られず，統計的情報のみに基づく自己組織化にも限界が見えてきた．

そこで我々は，文書の文法情報や，文書集合の時間関係や参照関係などを文書の文脈情報と位置付け，統計情報で絞り込まれた文書に対して，文脈情報を用いた自己組織化を行なって，更に絞り込む方法を提案する．本報告では，その例として文書上の単語の出現順序を反映したネットワークを文脈情報とすて用いた自己組織化と可視化を行なった結果について述べる．

2　記号系列の自己組織化

文書を単語の連続系列とみなすとき，似た内容を記述した箇所では，同じ様な単語の系列が出現するし，ある単語に続く単語は意味が近くなることが多い．そのため，単語の連接関係を表現するネットワークが作成できれば，文書の内容を解析する上で有益である．この目的に合致するのが，RWC つくば研で提案されている，IPM(*I*ncremental *P*ath *M*ethod)[2],[3],[4] である．このアルゴリズムの基本規則を図1に示した．

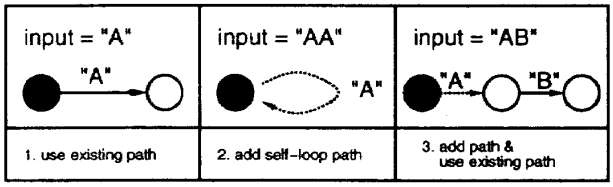

図 1: IPM 基本規則

単語の連続系列に対して IPM を適用することにより，単語の連接関係を表現するネットワークが自動的に構築できる．このとき入力が複数の文書であれば生成されたネットワークは複数の文書の広い意味での文脈を表現していると考えられる．

複数の文書を形態素解析などで単語系列に変換し，これを入力として IPM でネットワークを生成する実験を行った．実験の目的より，入力文書は互いに関連を持ち，単語も共有している必要があるので，RWC-DB-TEXT-95-3[5] の文書分類コードを用いて，新聞記事の中から記載の内容の分野が同じである記事を選んで入力した．この新聞記事は，1994 年の毎日新聞を研究用に利用しているものである．

実験の結果，約 100 種類の分類カテゴリーについて．入力記事数 100 ～ 500 に対して，生成ネットワークのノード数 1000 ～ 10000 のネットワークが作成された．

3　区間配置の準備

IPM ネットワークの可視化については，すでに [3],[4] などでも提案されているが，文書からネットワークを構築する場合には，語彙数が多くなるため，ノード数が多くなる，単語のみでなく，文書も空間上に配置する，などの条件を考慮する必要がある．

さて IPM ネットワークという文脈上では，間に挟むノードが少ない単語ほど親近度が高いと考えられる．そこで単語 $word_i, word_j$ 間の距離 $d(word_i, word_j)$ を式1で定義する．

$$d(word_i, word_j) \quad = \quad word_i \text{と} \ word_j \text{の最短}$$

A study of network expression and visualization method for large number of texts
Jun Toyoura, Shin-ichirou Tsudaka, Hiroki Konaka
Real World Computing Partnership, Information-Base functions Mitsubishi Laboratory

path 上の node 数 (1)

一方，文書 $text_i$ と単語 $word_j$ 間の距離は，文書 $text_i$ に出現する単語と $word_j$ 間の距離の平均と考えると式 2 で定義できる．

$$d(text_i, word_j) = \sum_{word_k \in text_i}^{N_i} d(word_k, word_j)/N_i \quad (2)$$

同じ考え方で，文書 $text_i, text_j$ 間の距離 $d(text_i, text_j)$ が式 3で定義できる．

$$d(text_i, text_j) = \sum_{word_k \in text_j}^{N_i} d(text_i, word_k)/N_i \quad (3)$$

対象間の距離が分かっている場合，この距離に基づき，距離の小さい対象同士は近傍に，低い対象同士は遠くにと空間内へ配置する方法としては一般に数量化法 [6] が用いられる．今回も数量化 IV 類を用いることにした．

4 ネットワーク可視化実験

あらかじめ IPM で生成したネットワークに対し，前節に示した方法を適用して文書情報空間を構築する実験を行なった．数量化 IV 類の行列に対して主成分分析を行ない，累乗法を用いて固有値の大きい順に第 1 ～ 50 成分まで求めた．計算は SUN Sparc Station 20 上で実行した．行列計算にメモリを大量に必要としたため，ノード数が 10000 を越える計算が実行できなかったので，実験はノード数が 1000 ～ 3000 程度のネットワークについて行なった．このときの入力文書数は約 100 ～ 200 程度である．第 1 ～ 3 成分より作成された文書情報空間の表示例を，図 2に示す．

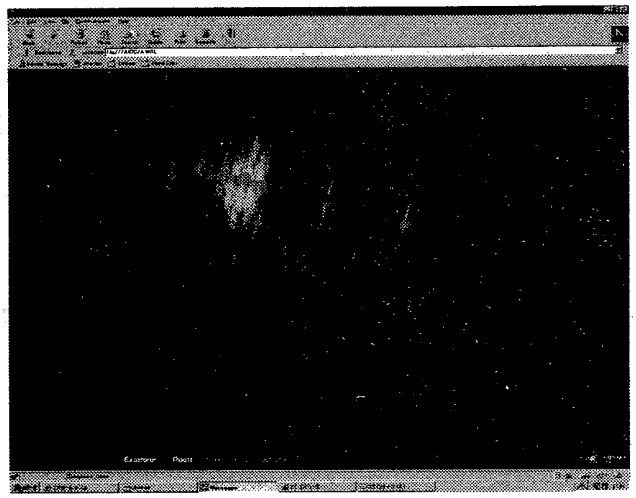

図 2: IPM 可視化の例

図に示したように，少数の単語，文書オブジェクトが座標軸に沿って離散的に，残り大部分のオブジェクトは原点付近に集中する形で配置されている．これは実験を行なった他の例にも当てはまる傾向で，結論から述べると，今回の実験ではユーザが見てわかり易いような配置は得られなかった．

5 評価

前章に述べたように，今回の実験からは意図するような文書可視化空間は構成できなかった．原因として式 1~3の定義が適切でなかったことが考えられる．そこで式 1に代えて式 4の d' を距離として用いる実験も行なったが，特に良い結果は得られなかった．

$$d'(word_i, word_j) = -1/d(word_i, word_j) \quad (4)$$

一方，式 2,3も改善の余地は有り，例えば文書との距離は数量化の計算では用いずに，後から出現単語の重心などに配置する方法が考えられるが，分布の大勢は単語間の距離で決定するので，根本的な改善は得られない．対象アイテムが少数の場合に比べ，1000 を越すと数量化による可視化は困難なのかも知れない．

6 おわりに

以上，大量テキストのネットワーク表現と可視化手法の一例を示した．その結果ネットワーク上の距離に基づき単純に可視化を行なっただけでは，文書・単語は十分分散されて配置されず，分解能が不十分であることがわかった．可視化方法と空間の構成に関しては今後も多くの改良が必要である．

今後は，上に述べたように時間などの文脈情報を利用した自己組織化方式の検討を進めるとともに，文法情報を使った自己組織化方式の研究開発も進めていく予定である．

参考文献

[1] H.Arita, T.Yasui and S.Tsudaka: "3D Stroller: Strolling in the self-organized information space", *Proc.of RWC Symposium*, pp53-58, 1997.

[2] 豊浦潤, 岡隆一, "テキストの知識ベース化のための自己組織化ネットワークの提案", 信学技報,*NLC96-59*, pp.23-30, 1997.

[3] 遠藤隆, 高橋裕信, 豊浦潤, 向井理朗, 岡隆一, "動画像の自己組織化ネットワークによるモデル化とその動的特徴の可視化 - Video Intra-structure Visualization -", 信学技報,*PRMU97-78*, pp.49-54, 1997.

[4] 向井理朗, 西村拓一, 遠藤隆, 岡隆一, "ジェスチャー動画像の自己組織化ネットワークによるモデル化と要素動作の自動抽出", 信学技報,*PRMU97-128*,pp.55-61, 1997.

[5] 豊浦潤, 徳永健伸, 井佐原 均, 岡隆一, "RWCにおける分類コード付きテキストデータベースの開発", 信学技報,*NLP96-13*, pp.23-32, 1996.

[6] 林知己夫, "数量化 ー理論と方法ー", 朝倉書店,1993.