情報学基礎 28－3
（1992. 11. 9）

テキストの内容を表すワードマップ作成の試み

豊浦 潤　有田 英一
三菱電機（株）中央研究所

本稿では電子化テキストの斜め読み的要約化手法を検討した。まず要約に対する適切性を表す尺度として「連想性」「専門性」導入し、これらを数値的に付与する方式が示される。そしてテキストに出現する単語をこれらを尺度に2次元平面上にマッピングすることによりワードマップを作成した。この方式によるテキストの要約度は10%程度であった。ワードマップは電子化テキストの内容を自動的に要約化し情報検索や情報獲得に有効である。

Word Mapping Method for Text Representation

Jun Toyoura, Hidekazu Arita
Central Research Laboratory
Mitsubishi Electric Corporation
8-1-1 Tsukaguchi-Honmachi, Amagasaki, Hyogo, 661, Japan

In this paper, method for text summarization like skiming through a book is introduced. At first, in order to measure adaptiveness of summarization, "associativity" and "speciality" are introduced and quantification methods for them are shown. Words in the text are ploted on the 2-D *Word-Map* by using these criteria. Sumarization rate of Word-Map is about 10 percent on an average. Word-Map summarises contents of text automatically and helps information retrieval and information acquisition.

1 はじめに

テキストの電子化に伴うペーパーレス化が進行している。ワードプロセッサの普及により個人が作成した文書はテキストファイルの形でフロッピーディスクなどに保存されることが多くなり、また広辞苑などの辞書類は勿論のこと、新聞1年分の記事といった従来商品化が困難だった大容量のテキストがCD−ROM化され本屋の店頭に並べられるようになっている。

今後もペーパーレス化は様々な局面で展開されていくだろう。しかし、かつてレコードがCDに駆逐されたように、本屋から本が完全に姿を消しフロッピーディスクやCD−ROMがそれに取って変わるためには技術的に克服せねばならない課題が多く残されている。その課題の1つに、これら電子化テキストのメディア自体としての表現力の低さが挙げられる。テキストはプログラムのように、実行可能で常に同じ結果を与えるものではなく解釈されるものである。それゆえテキストメディアにおいては解釈の自由度の高さが保証されねばならない。しかし、新聞にざっと目を走らせるように、本にさっと目を通すように、電子化テキストを眺める方法は現在のところ存在しない。

快適な情報環境は、

1. 欲しい本が決まっていれば、すぐに探し出せる
2. 検索対象となる分野が分かっていれば、求める内容の本を見つけ出せる
3. 面白そうな本を発見できる

ような、良く整理された大型書店のフロアーに準えられる。そこは、1、2のように検索条件がハッキリしている場合でも、3のように漠然としている場合でも客の要求に対応できる万能な情報検索の場である。特に3のように従来の情報検索パラダイムに対応する技術がない発見的情報検索を、筆者は情報散策と呼んでいる。情報源の複数化や情報サイクルの短縮により不必要に大量のテキストに囲まれた今日の私達の情報環境では、情報獲得のボトルネックを避けるためにもこうした情報散策環境の構築が必要なのである。

筆者は発見的情報検索の基本技術は、本の狩人たちがテキストの概略を迅速に把握する技術、言うなればテキストの斜め読み手法であると考える。本稿ではこのような問題意識から、電子化テキストを計算機上で斜め読みし、縮約的な表現を与える方法について考察する。そして、その具体例としてキーワードを2次元平面上にマップ形式に配置したテキスト表現を提案する。このマップは筆者らが既に提案しているテキストの自動分類機構 [1] より抽出される単語の連想関係を表すテンプレートパタンを利用して自動作成される。

2 テキストの要約化

2.1 要約文作成の困難

一般にテキストの要約化を人が行なう場合、要約文の作成を意味する場合が多い。しかし要約文作成を自然言語処理により計算機上で実行するためには

1. テキストを構成する個々の文の解釈
   ⟶ テキスト全文の解釈
2. 幾つかの話題の抽出
   ⟶ 文章の主題の抽出
3. 個々の話題の重要度を評価
   ⟶ 主題に基づく作文

などの高度な処理が必要となる。しかし1 → 3の過程をボトムアップに実行するためには莫大な背景知識と推測が必要とされ、例えば「私は鰻だ。」という短文の解釈さえ不可能な現状では実現が困難である。

この問題を回避するためにはテキスト内容を解釈抜きに直観的に捉えることができれば都合がよい。

2.2 斜め読みテキスト処理

本を読むという人間の情報処理活動に関しては心理学実験などで多くの研究があり、そのメカニズムの深層的解明はそれ自体困難な問題であるが、ここでは斜め読みシステムを下のように簡単にモデル化した。

1. テキストから単語を抽出する
2. 特徴的な単語を選別する
3. 選ばれた単語からテキストの種別を決定する
4. 決定した種類に対応した方法でテキストを読み直し情報を補う

このシステムは例えば入力テキストが交通事故の新聞記事である場合は

1. 記事を眺め
2. 「事故」、「衝突」など特徴的な単語を発見し
3. 交通事故に関するテキストと判断し
4. 5W1Hなどの情報を穴埋め的に補う

のように動作する。

斜め読みテキスト処理は浅い情報からテキストの主題を決定し、その決定に基づき深い情報獲得を行なうトップダウン的機構である。以下では2の特徴的な単語の選別法、3のテキストの種別の決定法、4の情報の補足をどのように行なうのかについて具体的に検討を進める。

2.3 単語の専門性

単語の特徴を測る尺度として、ここで専門性を導入する。一般に用語には**専門用語**と**一般用語**の範疇があり、テキストの内容表現には前者が適していると言われている。専門用語の定義に関しては幾つかの見解があるが[2]、ここでは「専門用語とは特定の分野に属するテキストに集中的に出現し、他の分野のテキストには殆んど登場しない単語である」と定義する。逆に一般用語は「複数の分野のテキストに共通して出現する単語」ということになる。

ここで定義した専門性を求めるためにはテキストの種別化が必要である。これは前節で述べた、2 → 3の順序関係と一見矛盾しているが、筆者は実際の斜め読みでは2 → 3の過程は、2 → 3 → 2 → 3 → ··· → 3 → 2のように循環的に収束していく過程だと考えている換言すれば「最初は一般用語を捨て」次に「選別された単語群に適合する可能性のある分野以外の種別を捨て」その次に「選別された分野以外の専門用語を捨て」・・・のように消去法的にテキストの種別が決定されると考えている。即ちここで言う専門性は単語に *apriori* に備わっている属性ではなく、多くのテキストを分類する過程で獲得されていく属性なのである。

2.4 テキストの種別化

今回は筆者が既に提案している教師なしニューラルネットワークによるテキストの自動分類機構をテキストの種別化に用いることにする。図

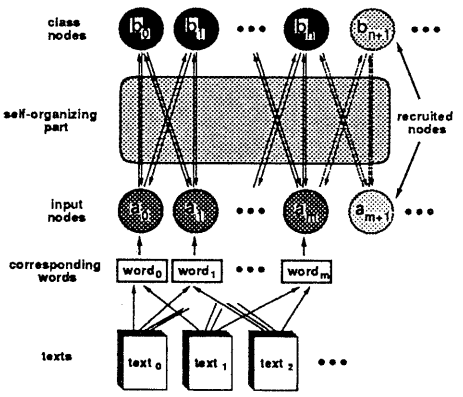

図 1: テキストの自動分類モデル

1に、この機構の概略を示す。今、テキストは
分類されるテキストの総数：n
全てのテキストから抽出した単語の種類：m
抽出単語のリスト：$\vec{kw} = (kw_1, kw_2, \ldots, kw_m)$
i番目のテキスト：$\vec{t_i} = (t_{i1}, t_{i2}, \ldots, t_{im})$
$t_{ij} = i$番目のテキストにkw_jが出現する頻度
のようにコーディングされる。そして、図中のニューラルネットは$\vec{t_i}$を入力層：$\vec{a}$に対する入力とし、出力層：$\vec{b}$がWTA^*の原理で分類を実行すると同時に、同一出力ノードに分類されたテキスト中のワードの共起関係の強さを重み付きパタンの形式でリンク上に保存する。このパタンの内で重みの大きいノードに対応するキーワードはその種別を代表するワードであると解釈される。以下これをテンプレートと呼び、j番目のテンプレートを$\vec{c_j}$などと書く。

一般にテンプレート中には専門用語と一般用語が混在していると考えられるがこのテンプレートから前節で導入した専門性を導出するモデルを図2に示す。このモデルでは各テンプレートを構成する単語のうち同一の単語間には抑制性結合を張ることにより、一般用語の活性度が抑制される。即ち、c_j上の
単語の専門性：$\vec{s} = (s_1, s_2, \ldots, s_m)$は

$$s_k = c_{jk} - \alpha \sum_{i \neq j} c_{ik} \qquad (1)$$

と定式化される（$\alpha > 0$は抑制の強さの係数）。図2では注目するテンプレートのうち「可能」「場合」「利用」は一般用語として抑制され、「知識ベース」「AI」「ルール」が専門性が高いと

*Winner Takes All

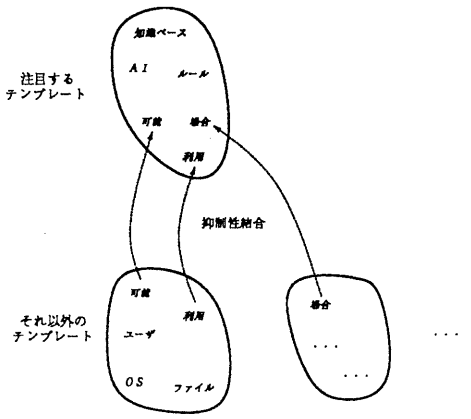

図 2: テンプレートからの専門性評価モデル

判定される。

2.5 情報の補足

今回はは５Ｗ１Ｈなどの抽出は行なわず、テンプレートに対し補足されるる情報として $\vec{t_i}$ 自体を用いることにした。即ち $\vec{t_i}$ が j 番目の種別に分類されたときの、補足された情報：p を、

$$p_k = \beta t_{ik} + c_{jk} \qquad (2)$$

の形で表される。以下ではこの補足された情報を連想性と呼ぶ。

ところで図 1のニューラルネットでは、入力層・出力層の各ノードは入力パタンに対し競合的に動作し、$\vec{t_i}$ がクラス j に分類されたとすると入力ノードの活性度は

$$a_k = \frac{t_{ik} + B c_{jk}}{t_{ik} + A} \qquad (3)$$

に収束する†。式 2は式 3の分子に対応しており、今回は連想性として式 3を用いた。

2.6 マッピング

要約化の結果は、テキストに出現した単語を前述の尺度を用いて２次元平面上にマッピングして表現することにした。以下、この手法で単語をマッピングした平面をワードマップと呼ぶ。

単語を何らかの尺度の下にマッピングした例は国内では [5]、[6] などがあるが、これらでは単語の属性や単語間の類似度・親近性が主観的に付与されていた。それに対しワードマップは客観的に自動作成することが可能である。

テキスト中への単語出現の統計情報にのみ注目したマッピングの例としては [4] が知られているが、マッピングのパラメータが２つの単語の共起性のだけなので、単語の２次元配置に恣意性があった。

†$1 \simeq A \gg B > 0$ はネットワークのパラメータで B は連想の強度を表す

3 実験

3.1 実験の設定

３６９件のＡＩ関係の技術記事をサンプルとしてワードマップを作成した。これらのテキストは [1] の自動分類で予め約３０種類に分類されている。

以下では対照実験として行なった、同じテキストに対する数量化３類による解析とその結果を述べた後にワードマップの作成結果を示す。

3.2 数量化３類による解析

数量化３類は、各データが幾つかのカテゴリに属すか否かの形式で得られている場合に適用される。今データを $\vec{t_i}$、カテゴリを $\vec{kw}$ とすれば、数量化３類を直接テキストの数量化に適用でき、数量化３類に非常に似た方法でテキスト分類を行なった例も報告されている [7]。

一般に数量化３類の処理結果は相関行列から導かれる固有ベクトルを固有値の大きさの順に、$\vec{e_1}, \vec{e_2}, \ldots$ とするとき、$\vec{e_1}, \vec{e_2}$ の要素の値で XY 平面上にカテゴリをプロットした図で表現されるが、今回のサンプルではカテゴリが約２０００語と大きいので、各ベクトルの中で成分の大きい単語、上位１０語だけを表 1、表 2に示す。ここで表の頻度は「その単語のテキスト全体での出現回数」である。

表 1、表 2に登場する単語の特性として、

1. 共起関係が強い
2. 出現頻度が低い

ことが指摘できる。これは $\vec{e_0}, \vec{e_1}$ のなかで成分の大きい要素に対応する単語が専門性の高い単語であることを示唆する。しかしテキストに関する共起関係の観点からは、$\vec{e_0}$ はテキスト ($A1$, $A2$) の特徴を代表する単語であるのに対し、$\vec{e_1}$ にはテキスト ($B2$, $B3$, $B4$) に特徴的な単語とテキスト ($B1$, $B5$) に特徴的な単語が混在している

表 1: $\vec{e_1}$ の成分表

重み	キーワード	頻度	出現テキスト
0.118	計算機科学分科会	3	A1
0.111	分科会	3	A1,A2
0.111	第五世代	3	A1,A2,A3
0.111	新情報処理技術調査研究委員会	3	A1,A2
0.111	新機能構想分科会	3	A1,A2
0.111	後継プロジェクト	3	A1,A2
0.111	基礎技術分科会	3	A1,A2
0.111	WG	4	A1,A2
0.081	情報処理	4	A1,A2,A3,A4
0.075	調査・研究	3	A1,A2,A5

表 2: $\vec{e_1}$ の成分表

重み	キーワード	頻度	出現テキスト
0.061	ブラジル	4	B1
0.045	東京都大田区	3	B2,B3,B4
0.045	エー・エス・ビー	5	B2,B3,B4
0.045	RT－Prolog	9	B2,B3,B4
0.043	イタリア	4	B1,B5
0.043	受託開発業務	7	B4,B6,B7
0.040	AZ－Prolog	6	B4,B8,B9,B10
0.040	WING	5	B8
0.040	MS－WINdows3	3	B3,B4,B11
0.037	Expertech社	3	B5,B12,B13

ため、異なった分野の専門用語が混在することになるため、その特徴を説明することは難しい。また、$A1 \sim A5, B1 \sim B13$ 以外のテキストの特徴化をどのように行なうかという問題もあるため、テキストの要約化の指標には適切でないと考えられる。

3.3 自動分類によるテンプレート

自動分類より得られたテンプレートの内、$\vec{e_0}, \vec{e_1}$ に対応するものを、テンプレート1、テンプレート2とする。ただし $\vec{e_1}$ に対してはテキスト ($B2$, $B3$, $B4$) の特徴に対するテンプレートを選んでいる。これらの中で重みが大きい単語を表3、表4に示す。

表 3: テンプレート1

重み	キーワード	頻度
0.2108	通産省	16
0.1064	検討	80
0.1042	プロジェクト	43
0.1044	組織	18
0.0992	米国	78
0.0964	システム	527
0.0890	WG	4
0.0844	組織変更	5
0.0820	開催	68
0.0791	意見	16
0.0776	研究者	35
0.0725	研究開発	20
0.0724	技術	40
0.0701	計算機科学分科会	3
0.0684	設置	33
0.0682	研究成果	13
0.0656	分科会	3
0.0656	新機能構想分科会	3
0.0656	基礎技術分科会	3
0.0610	新情報処理技術調査研究委員会	3
0.0610	後継プロジェクト	3
0.0610	第五世代	3
0.0609	状況	22
0.0599	ニューラル・ネットワーク	46
0.0568	計画	124
0.0565	ICOT	9
0.0558	中核	12
0.0539	調査・研究	3
0.0527	支援	70
0.0515	活動内容	7

表3、表4より分かるように、テンプレートには出現頻度の高い単語と低い単語が混在している。

3.4 ワードマップ

前述のテンプレートを用い、式3と式1を軸に作製したワードマップの一例を図3に示す。

サンプルテキストとしてAI関係の技術記事を用いたため、図3のマップでは一般的尺度では専門性が高いと考えられる、「システム」、「ユ

表 4: テンプレート 2

重み	キーワード	頻度
0.2796	販売	207
0.1549	ＲＴ－Ｐｒｏｌｏｇ	9
0.1401	Ｗｉｎｄｏｗｓ３	14
0.1165	移植	58
0.1060	Ｅｇｅｒｉａ	11
0.1095	ＴＡＯ	15
0.0955	記述	111
0.0939	ＨＯＯＰＳ	7
0.0920	パソコン	37
0.0889	Ｘｉ Ｐｌｕｓ	7
0.0888	エー・エス・ピー	5
0.0812	ＶＡＲ契約	4
0.0803	提供	142
0.0716	アプリケーション	111
0.0789	日本語版	27
0.0673	処理系	20
0.0654	ソフト・ハウス	25
0.0634	ユーザー	265
0.0632	ＡＩツール	93
0.0622	普及	25
0.0574	利用	499
0.0544	計画	124
0.0537	ファモティク	4
0.0530	ＯＳＦ／Ｍｏｔｉｆ	8
0.0529	ＧＵＩ	6
0.0529	ＮＥＷＳ	19
0.0525	稼働	133
0.0518	オブジェクト指向	23

ーザ」などの専門度は低いと判定されている。マップの右上に分布するワード、例えば図の破線の右側の数ワードは連想度・専門度が高く、テキストの内容を代表するのに特に適切なキーワードと言える。そして、これらの中でも「知識ベース」のように左上部に位置するワードは、このテキストに限って言えば連想性は余り高くないが、分類されたクラスに固有のキーワードであること、「ＣＬ」のように右下に位置するワードは、分類されたクラスに対する専門性は低く、このテキスト固有のキーワードであることが推察される。実際にこのテキストは、「ルールベースを利用したＡＩツール：ＣＬ」に関する記事であり、マップは上に述べた内容を良く

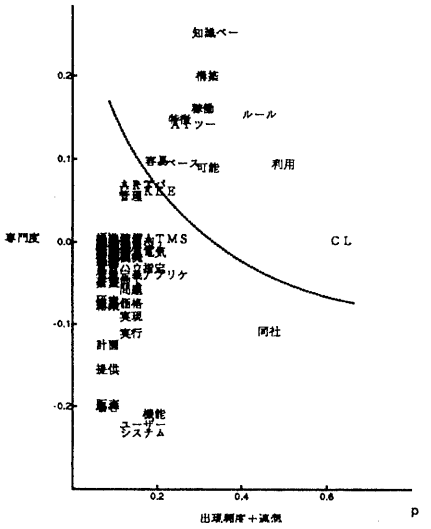

図 3: ワードマップの例

反映している。

また、Ａ１に対するワードマップを図 4に、Ｂ３に対するワードマップを図 5に示す。

ワードマップは右上、左上、左下、右下に 4等分したとき、右上に位置する単語数を N_{UR}、左上、左下、右下に位置する単語数を N_{UL}、N_{DL}、N_{DR}、として要約率を下のように定義する。

$$\text{狭要約率} = \frac{N_{UR}}{N_{UR} + N_{UL} + N_{DL} + N_{DR}}$$

$$\text{広要約率} = \frac{N_{UR} + N_{UL} + N_{DR}}{N_{UR} + N_{UL} + N_{DL} + N_{DR}}$$

作成した３６９枚のワードマップについてこれら要約率の平均値は

$$\overline{\text{狭要約率}} \Longrightarrow 7.2$$

$$\overline{\text{広要約率}} \Longrightarrow 12.6$$

であった。

4 おわりに

本稿で述べた内容を以下にまとめる。

1. テキストの内容を斜め読み的に要約化するモデルを示した
2. 単語の要約に対する適切度の尺度として連想性、専門性を導入した
3. テキストと単語テンプレートから連想性、専門性を定量化する方式を示した

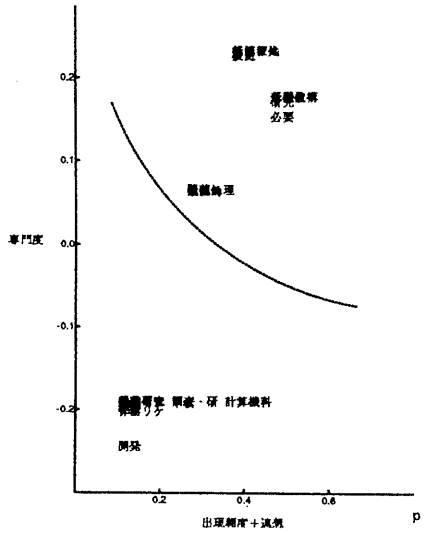

図 4: A 1のワードマップ

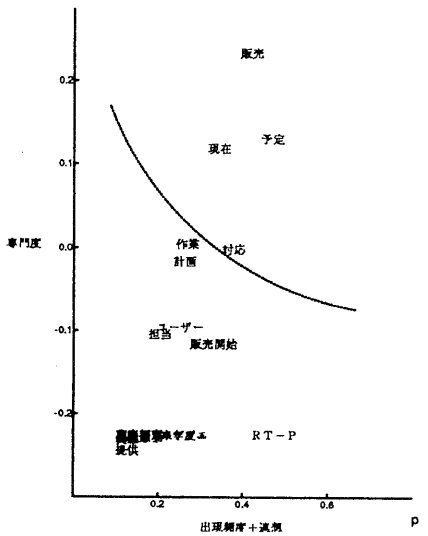

図 5: B 3のマップ

4. ワードマップを作成し、テキスト上の単語が1割程度に要約化されることを確認した

[1] のテキストの自動分類の結果は必ずしも人間が見て分かり易い分類とは言えなかったが、ワードマップはこの自動分類の結果を説明する。マップ上の単語配置は機械の解釈を表しており、これは場合によっては人間に新しい視点を与えるが、マップの最終的解釈はそれを見る個々人に開かれている。

今後はパラメータチューニングなどにより、ワードマップの改良を進めていく一方

1. 検索過程でのテキスト選別
2. テキストへのキーワード付与

に対する有効性を具体的に確証していく予定である。

参考文献

[1] 豊浦潤, 小船隆一, 有田英一, 自己組織型ニューラルネットワークによるドキュメントの自動分類, 情処NL研資, *92*, 21, pp.41-48, 1992.

[2] 石井, 専門用語を抜き出す試み, 専門用語研究, *3*, pp.32-37, 1991.

[3] Rumelhart, D., McClelland, J.,and the PDP Research Group, Parallel Distributed Processing: Explorations in the Microstructure of Cognition, *MIT Press*, 1986.

[4] Doyle, L. B., Indexing and Abstracting by Association, *American Documentation, Oct.*, pp.378-390, 1962.

[5] 堀浩一, 単語の意味の学習について, コンピュータソフトウェア, *3*, 4, pp.65-72, 1986.

[6] 田村淳, 記号間の力学に基づく概念マップ生成システムSPRINGS, 情処学論, *33*, 4, pp.465-470, 1992.

[7] Can, F., Ozkarahan, E. A., Concepts and Effectiveness of the Cover-Coefficient-Based Clustering Methodology for Text Databases, *ACM Trans. Database Syst, 15*, 4, pp.483-511, 1990.