

## 文書の自動分類のための、 4E-5 辞書を使わないテキストのパターン化方式<sup>1</sup>

安井 照昌<sup>2</sup>  
RWCP<sup>3</sup>新機能三菱研究室<sup>4</sup>

### 1 はじめに

全文データベースなどの発展・普及とともに、膨大なテキストの中から必要なものだけを選択する技術が重要になってきている。文書データベースなどの検索を目的として収集・構造化されたものにおいてさえ、実際にある目的をもって、関連文献を手に入れることは多大な労力を要する。

このためにさまざまな検索方式が研究され、一部は実用化されているが、それらの共通の基礎技術としてテキストのインデキシングが重要である。これは、自然言語でかけられたテキストを、計算機で処理するためにパターン化する手法である。例えば、テキストの内容を表す重要語句をテキスト中から取り出し、これに重みをつけるなどしたベクトルによってテキストを表現しようというものである。それ以降の検索処理はこのベクトル表現に対して行われる。このため、インデキシングの方法の良否が検索・分類などのテキスト処理全体に占める割合は大きい。

われわれは、自己組織型情報ベースの要素技術として、ニューラルネットワーク等を用いた文書分類方式<sup>[1]</sup>を研究してきたが、これまで、テキストのパターン化には既存のツールを利用してきた。このシステムは、巨大な単語辞書を利用してテキスト中から単語を切り出すものであるが、辞書の作成・メンテナンスなどは費用や労力の点で大変であり、また品質の点でも新出語の不足、質的なバラツキなどがありうる。このため、ある基準に沿って自動的にテキストから語句を抽出する方式が望まれる。

本研究では、プレーンなテキストから、定型的な処理によって意味を持つ文字列を取り

<sup>1</sup> A method to index texts into patterns without dictionary, aiming at automatic document clustering.

<sup>2</sup> Terumasa YASUI

<sup>3</sup> Real World Computing Partnership

<sup>4</sup> Novel Functions Mitsubishi Laboratory

出す方式について述べる。

### 2 文字列の切り出し

テキストのインデキシングに用いる文字列は、テキストの内容を表す概念的な内容を持ったものが好ましい。辞書によって意味を与えずに意味のまとまりを表す文字列を取り出すため、次の仮定を設ける。

「意味のまとまりを表す文字列は、複数のコンテキストで使われる」

この仮定に基づき、対象のテキストから以下の条件を満たす文字列を取り出し、辞書として使用するというものである。

(a) それを含み、それと同じ回数出現するより長い文字列が存在しない

全文を最長で、出現回数1回の要素として含めば、以下のように言い替えられる。

(a) 2回以上現れる

(b) それを含み、それと同じ回数出現するより長い文字列が存在しない

これによって得られた文字列によってそれぞれのテキストをパターン化する。

### 3 実験

テキストの全長を $L$ とすると、長さ $L-1$ から開始して、長さ1のものまで、すべてについてそれを含む文字列を数え、比較すれば良い。実際にはこれは計算量があまりに多いので、予備実験により2回以上現れる文字列の最大の長さを見積り、さらに、全テキストを分割して、それぞれの中で取り出した文字列を後でマージするという方法をとった。

以下にテキストからの文字列の抽出と、それによる分類実験の結果を示す。既存の人手で作った辞書を用いる方式（以下、方式A）と1月1日の114記事、95273文字から上記の手順により取り出した9511の文字列を用いる本方式（方式B）とを比較する。

まず、1415文字からなる一つの記事を、それぞれの方式でベクトル化し、比較してみ

た。以下にその傾向を列挙する。

- Aでは単語数は79種類131個であったのに対し、Bでは938種類2532個抽出された。
- Aで抽出されたもののうち半分(39/79)はBにも含まれる。
- Aで2回以上抽出されるもののうち90%(18/20)はBにも含まれる。
- Aに1回のみ含まれるものうち35%(21/59)はBにも含まれる。
- Aで抽出された要素で、Bでは抽出されなかったもののほとんどは(39/40)Bで抽出された文字列と包含関係にある。
- Bで抽出され、Aで抽出されないもののうち、Aの文字列と包含関係にあるものは、30%(269/899)

方式Aの辞書は、名詞を中心とした自立語を用いるが、本方式によるものは、附属語や活用語尾、言い回しなどのような文字列も繰り返し現れれば同様に抽出する。このため、Aの辞書より要素が多い。Aで取り出されたもの(特に出現頻度の高いもの)に関しては、Bによってもほとんどが抽出できる。また、全く同一でなくとも、包含関係に文字列が取り出されることもある。

逆に、Aでは取り出されないような文字列も多く取り出す。このうち30%はAの文字列と包含関係にあるが、単独の文字や活用語尾や助詞・助動詞の連続のような平仮名の文字列(「だった」、「さんは」、「ていたことが」など)、数字など(「8月22日」、「6300円」)のほか、「万円で」、「の利」などといった良く使われそうな断片を多く含む。現在の辞書は元になっているテキストが小さいため、特にカタカナやアルファベットの文字列に弱く、既存のツールと比較すると、抜けが多い。ただし、めずらしい固有名詞や流行語などは拾ってくる。

これをもとに簡単な分類実験を行なった。文字列によってベクトル化された114個のテキストを最大距離法によって30クラスタに分類した。各記事中の文字列の出現頻度を要素とするベクトルで各記事を表現する。ベクトルは長さ1に正規化してある。ベクトルの要素数は114記事から拾い出された文字列の種類数であり、方式Aでは5439要素、方式Bでは9507要素である。一記事の場合BはAの10倍以上要素数があったことからも、方式Bの方が記事間で共通の文字列を取り出していることがわかる。

方式Aで同じクラスタに分類されたものが、方式Bでも同じクラスタに分類されている数を点数とし、乱数によって作ったデータと比較すると、幾分高いことがわかる(表1)。

乱数	146点
方式B	220点

表1: 分類の比較

#### 4 考察

従来の方法に比べて最大の特徴は、あらかじめ単語辞書を用意する必要がないことである。また、処理対象のテキストから辞書を生成するため、そのテキスト中に含まれる新出語彙に対応できる。

一方、辞書を用いる場合には分類やインデキシングにおいて重要かどうかという基準で単語が選ばれるが、この方式では、テキスト全体の中で、他のテキストと比べてそのテキストが偏って多く含む文字列を選ぶ。つまり、他のテキストとの区別するのに有効かどうかという基準で、文字列を選ぶ。これは、対象となる全テキストの傾向にも依存する。例えば、対象が新聞記事なら、「コンピュータ」という語はインデクスとして重要なが、コンピュータ関係の文献ではあまり重要でないといった具合である。

また、この方式は、テキストを単語の並びに分割するのではなく、意味のまとまりごとの文字列を用いるので、例えば「株式会社」は「株式」「会社」「株式会社」のいずれでも、また重複した複数の文字列でインデクスされることが期待できる。特に日本語のような膠着語において、単語の境界が曖昧な場合に特によい。

今後は、テキスト情報ベースの自己組織化という全体目標の要素技術である、シソーラスや概念辞書の自動生成に対してこの方式が利用できないか検討していく方針である。

なお、本実験には、朝日新聞社提供の紙面データを用いた。

#### 参考文献

- [1] 豊浦潤、有田英一「テキストの内容を表すワードマップ作成の試み」、FI 28-3、情報処理学会研究報告(Nov. 1992).