文書クラスタの階層構造を利用した代表文の生成

倉田 早織・加納 敏行・齋藤 佳美
東芝ソリューション株式会社　ＩＴ技術研究所
{kurata.saori, kano.toshiyuki, saito.yoshimi}@toshiba-sol.co.jp

1　はじめに

近年、WWW からの収集や、パソコンの普及に伴う企業内での作成等により、大量の文書が蓄積されるようになってきた。蓄積された文書を活用するための第一歩として、大量の文書の全体像を把握するために、文書分類、特に、文書クラスタリングが活用されている[1]。

文書クラスタリングの結果に基づき、大量の文書の全体像を明確に把握するためには、上位下位関係や同階層の関係にある複数のクラスタが、それぞれ他のクラスタと比べてどのような位置づけにあるのかといった、クラスタに特有の情報を明確に理解する必要がある。

個々のクラスタに付与されたキーワードを見ることでクラスタの概要を把握するのが一般的だが、キーワードだけで概要を明確に理解するには困難な場合が多い。そこで、概要の明確な理解のためにはクラスタの概要として文を提示することが有効であると考えられる。

関連研究として、構文要素の共通部分を抽出することにより、クラスタの概要を表す文を提示する研究がある[2]。この研究は、複数のクラスタの存在を前提とした手法でないため、クラスタに特有の情報を提示できない場合がある。

本研究は、文書集合の全体像を明確に把握することを目的として、クラスタに特有の内容を表す文を生成する手法を開発した。このような性質を持った文を本研究では代表文と呼ぶ。2 章では代表文を生成する手法を述べ、この手法を用いて生成された代表文の例を 3 章で紹介する。4 章で生成された代表文の評価を行い、5 章で今後の課題を述べる。

2　代表文の生成

代表文の生成プロセスは、抽出パターンによる代表文候補の生成と、階層構造を利用した文の選択の 2 つから構成される。

以下、順に各プロセスの詳細を述べる。

2.1 抽出パターンによる代表文候補の生成

図 1 に抽出パターンによる代表文候補の生成の流れを示す。抽出対象クラスタ（代表文を生成したい文書クラスタ）とその上位クラスタに含まれる全ての文の構文木の各々に対し、抽出パターンを適用し、部分構文木を抽出する。そして、抽出した部分構文木から文を生成し、代表文候補とする。

抽出パターンは、主語、述語等の文の骨格として重要な単語と、単語間の係り受け関係を表したものであり、抽出パターンにマッチした部分が部分構文木として抽出される。

2.2 階層構造を利用した文の選択

生成された代表文候補の中から、抽出対象クラスタに特有の文を選択する。文の選択には、抽出対象クラスタとその親クラスタの間の重みつき自己相互情報量を用いる。

さらに、抽出対象クラスタに対する文書カバー率が高くなるように、代表文候補を選択し代表文とする。

3　代表文の例

本研究では、実験データとして、国土交通省が

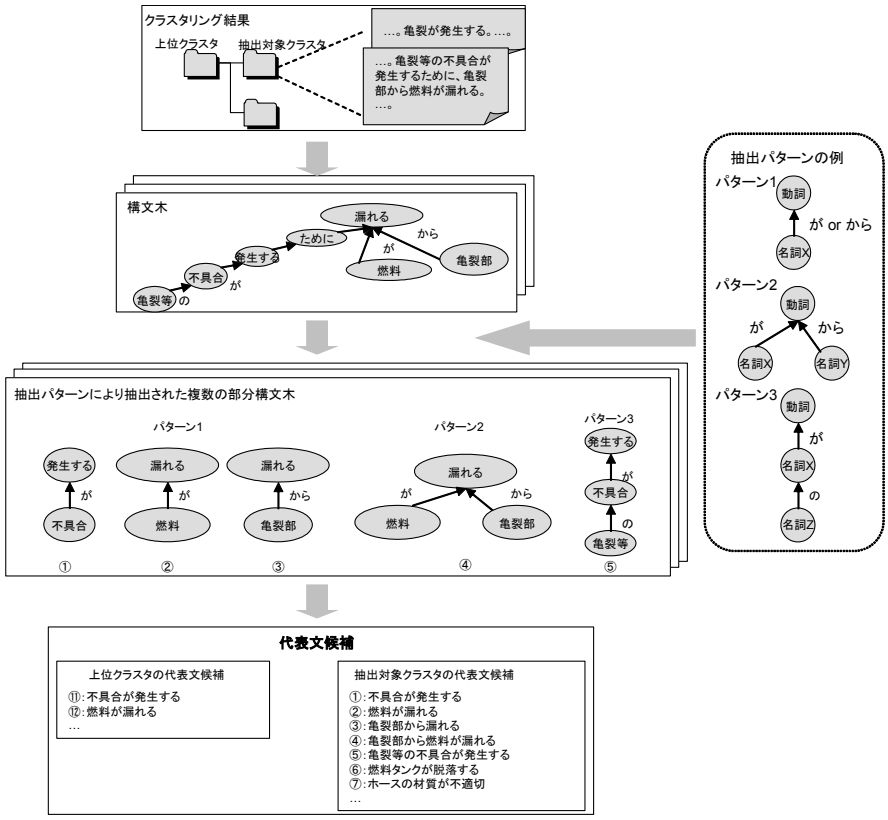

図 1 抽出パターンによる代表文候補の生成の流れ

公開している自動車リコール情報[1]を使用した。図2にリコール情報のクラスタリング結果の一部を示す。文書が幾つかのクラスタに自動分類され、クラスタの名前も自動生成される[3]。「燃料装置」、「ホース」、「タンク」、「ポンプ」等が生成されたクラスタであり、燃料装置クラスタの下位クラスタとして、燃料装置の構成要素のクラスタが生成されている。なお、下位クラスタに含まれる文書は、燃料装置クラスタにも含まれる。

表1に図2のホースクラスタから生成された代表文の例を示す。この例では、「燃料が漏れる」のように、ホースクラスタ、タンククラスタ、ポンプクラスタ等の下位クラスタに共通の内容は代表文として生成されず、「燃料が漏れる」の原因と考えられる内容、「燃料ホースが損傷する」が、ホースクラスタに特有な文として生成されている。

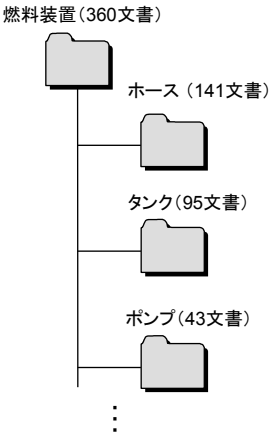

図 2 クラスタリング結果の一部

4 評価

本手法で生成されたホースクラスタの代表文が、ホースクラスタに特有な内容を表した文であるかを評価した。評価は2つの観点、即ち、人手で作成された代表文との比較、及び代表文とクラスタ間の類似度で行った。

[1]http://www.mlit.go.jp/jidosha/carinf/rcl/recall.html

表 1 ホースクラスタの代表文の例

取り回しが不適切
燃料ホースが損傷する
亀裂が進行する
ホースが損傷する
中速回転域において噴射パイプに共振が発生する
燃料パイプに亀裂が発生する
ホースが外れる
振動により燃料が漏れる
燃料ホースと燃料ポンプの組付けが不適切なものがある
ホースに亀裂が発生する

4.1 人手によって作成された代表文との比較

4.1.1 評価方法

人手で作成した代表文と比較し、両者の一致の度合いを算出した。自動車の不具合に精通していない日本語母語話者 2 名（以下、A、B とする）が作成した、ホースクラスタの代表文を正解代表文とした。作成者の行った作業を次に示す。

(1) ホースクラスタに含まれる全文書を読み、内容を把握する。
(2) 燃料装置クラスタに含まれる文書集合において、ホースクラスタに含まれない文書を全て読み、内容を把握する。
(3) 把握した内容に基づき、ホースクラスタに特有な内容を動詞文や形容詞文で作成する。

正解代表文として、A により 33 文、B により 36 文が得られた。ホースカテゴリの各代表文を、正解代表文と比較し、表記ゆれ、活用語尾、送り仮名、付属語以外の部分が一致しているとき、正解代表文と一致とした。ただし、付属語のみの違いでも、「振動が発生する」と「振動により発生する」、文の意味が異なる場合は、正解代表文と不一致とした。

A と B により作成された 2 つの正解代表文の集合には、共通する文もあったが、異なる文も含まれていた。A の正解代表文に対する B の正解代表文の適合率は 0.33 であった。

作成者 A と B は代表文を作成する過程で、対応する文書の頻度が小さい文を除外している。その除外する基準は、作成者の主観によるため、2 つの正解代表文の集合に差異が生じたと考えられる。

しかし、2 つの正解代表文の集合を目視で確認すると、それぞれの集合はクラスタに特有な概要を把握するには妥当な内容であった。

そこで、2 つの正解代表文の集合は、代表文の正解データとして妥当と考えられるので、それぞれを評価に用いた。

4.1.2 評価結果と考察

生成された代表文と正解代表文の一致の度合いとして、適合率、再現率と F 値を表 2 に示す。適合率は正解数を代表文の数（以降、M で表す。）で割ったもの、再現率は、正解数から一致する正解代表文の重複を除いたものを、正解代表文の数で割ったもの、F 値は適合率と再現率の調和平均である。ここで、正解数は正解代表文と一致した代表文の数である。

本手法では、M はユーザーによって設定可能なパラメータとしている。ここでは正解代表文と同程度の数である

表 2 ホースクラスタの代表文と人手で作成された代表文との比較

	作成者 A			作成者 B		
	$M=20$	$M=30$	$M=50$	$M=20$	$M=30$	$M=50$
適合率	0.45	0.43	0.40	0.50	0.50	0.42
再現率	0.27	0.36	0.45	0.25	0.39	0.50
F 値	0.34	0.40	0.43	0.33	0.44	0.46

$M=30$ の場合に注目して、代表文と正解代表文の一致の度合いを評価した。代表文の適合率は、A の正解代表文に対して 0.43、B の正解代表文に対しては 0.50 となった。このように、生成された代表文のほぼ半数がそれぞれの正解代表文と一致した。

本手法では、対応する文書の頻度が 3 以下である代表文候補が 1000 個以上あったが、そのうち代表文として選択したものは 16 個であった。1000 以上のものから 16 個を選択する方法は多数あり、選択結果は人間と一致しなかったと考えられる。

したがって、ほぼ半分である正解代表文との一致の度合いは、高いといえる。

4.2 代表文とクラスタ間の類似度

4.2.1 評価方法

生成された代表文とその抽出元であるクラスタの内容が類似しており、他のクラスタとは類似していないことを評価するために、代表文とクラスタの類似度を算出した。

類似度として、次の式で定義される値を用いた。

$$\text{類似度}=\frac{\vec{D}\cdot\vec{E}}{\left|\vec{D}\right|\left|\vec{E}\right|}$$

ここで、$\vec{D}$ は代表文を構成する全ての文が結合された文書の文書ベクトル、$\vec{E}$ はクラスタに含まれる文書が結合された文書の文書ベクトルである。

文書を構成する単語を文書ベクトルの素性とし、単語の $tf\cdot idf$ 値を単語の重みとした。ここでは、品詞が、名詞、動詞、副詞、形容詞である単語を素性として選択した。

4.2.2 評価結果と考察

表 3 に、代表文（$M=30$）とクラスタの類似度を示す。

ホースクラスタの代表文とその抽出元であるホースクラスタとの類似度は 0.69 であり、タンククラスタとの類似度は 0.55 である。このように、抽出元クラスタとの類似度が他のクラスタとの類似度より高くなっている。タンククラスタの代表文についても、上記と同様な傾向が得られた。

表 3 クラスタの代表文とクラスタの類似度

	ホースクラスタ	タンククラスタ
ホースクラスタの代表文	0.69	0.55
タンククラスタの代表文	0.51	0.73
ホースクラスタ	1.00	0.86
タンククラスタ	0.86	1.00

クラスタ間の類似度を表 3 の下 2 行に示す。含まれる文書が排他的であるホースクラスタとタンククラスタの類似度は 0.86 である。この値と比較して、ホースクラスタの代表文とタンククラスタの類似度 0.55 は小さい値であり、代表文が他のクラスタとは類似していないことを表しているといえる。

したがって、生成された代表文は、抽出元クラスタの内容を表しつつ、他のクラスタの内容を表さない文であり、クラスタに特有な内容を表した文であると考えられる。

5 今後の課題

文書集合の全体像を明確に把握するためには、読み手にとって意味を解釈しやすい文を提示することも重要である。今後、代表文の意味の解釈のしやすさを評価していく。

参考文献

[1] 奥村学、難波英嗣、『テキスト自動要約 （知の科学)』、オーム社、2005.

[2] 上田良寛、小山剛弘、「共通意味断片の抽出による複数文書要約」、言語処理学会第 6 回年次大会、pp360-363、2000.

[3] 宮部泰成、「ユーザーの意図を反映した対話型文書分類技術」、東芝レビュー、Vol.64 、No.52、pp.58-59、2009.