デジタル・ドキュメント 6－2
（1997. 3. 28）

情報分類を用いたトレンド・アウェアネスの支援

杉崎 正之　井上 孝史　大久保 雅且　田中 一男

NTT ヒューマンインタフェース研究所

本稿では、情報分類技術を用いた情報潮流の知覚(トレンド・アウェアネス)を支援するための手法を提案する。近年、情報発信の手軽さと高速性から、コンピュータネットワークを用いたテキスト情報の流通が盛んになってきている。しかし、自分が必要な情報を取り損ねないように、新たに発信される情報を常に見続けることは困難である。この問題を解決するために、どのようなテキスト情報が発信されているかを視覚化する手法を考案した。テキスト情報の自動分類技術を用いてテキスト集合の中から類似するテキストのまとまりを抽出し、そのまとまりを時間的に並べることでこれを実現した。また有効性を確認するために支援システムを試作し、その評価を行った。

An Assistance Method for Trend Awareness using Text Classification and Clustering

Masayuki Sugizaki, Takafumi Inoue, Masaaki Ohkubo and Kazuo Tanaka

NTT Human Interface Laboratories

This paper proposes a method for extracting trend of topics using automatic text classification and clustering techniques. In the recent information society, people need to keep watching a variety of text information media in order to acquire useful information. But it costs time and moreover there is a possibility of losing important information. To solve this problem, we propose an assistance method for making trend of topics visible using automatic text classification and clustering. It extracts categories including similar topics from a set of text documents which are newly sent, and displays those categories in accordance with time. To confirm the effectiveness, we constructed an experimental system and evaluated the results.

1 はじめに

近年、情報発信の手軽さと高速性からコンピュータネットワークを用いた電子化されたテキスト情報の流通が盛んになり、誰もが新しい情報を任意の時間で発信することができるようになった。また、これらの情報を検索するサービスが行われており、自分が必要な情報を探し出し入手することが可能となった。しかし、大量の情報を得ることができるために、すべてを処理できないという情報過多の問題が生じてしまっている。

さらに、電子ニュースや電子メール、文字放送などの情報発信メディアの多様性が問題を複雑化している。例えば、必要な情報を漏れなく収集するために様々な情報メディアに注目しようとしても、すべてのメディアを対象にできず、結果的に必要な情報を取り逃がしてしまう。また、情報源をある一つのメディアに絞ったとしても、情報は常に変化し続け必要な情報がいつ発信されるか分からないために、不必要な情報も見続ける必要が生じてしまう。

本研究は、時間情報を持つテキスト情報に注目し、どのような情報が含まれているか、その情報が時間変化と共にどのように変化しているかを視覚化することで、ユーザが情報潮流を見極める(トレンド・アウェアネス)ための支援システムの構築を目指している。

2 情報潮流

最初に情報潮流とは何かを説明する。研究対象とする情報は、インターネット上のネットニュースや、新聞社などがWorld Wide Web上で情報発信しているニュースサービスのような、情報メディアから常に新しく発信されるテキスト情報である。これらの情報メディアが発信する各記事にはその内容を端的に表現するキーとなる複数の単語が存在している(以後、これらをキーワードと呼ぶ)。新たに発信される記事には過去の記事と同一のものや類似したキーワードが混在するという特徴があり、記事集合内におけるキーワードの時間的な変化の様子は以下のようにまとめることができる。

(1) 新情報の発信 — 過去に発信されていない新たなキーワードを持つ記事が発信される。

(2) 情報の継続 — 同一のキーワードを持つ複数の記事が次々に発信される。

(3) 情報の転換 — 新たな記事が発信され続けているうちに、別のキーワードが徐々に出現してくる。

(4) 情報の分岐 — 一つの記事に対し、視点の違いによってキーワードが異なる複数の記事に分かれていく。

(5) 情報の統合 — 複数の記事に存在していたキーワードが一つの記事の中に集められて発信される。

このキーワードの時間的な変化の様子を「情報潮流」と呼ぶことにする。5つの潮流を抽出し視覚化することによって、情報潮流(トレンド)の大きな流れや変化の様子をユーザに気付かせること(アウェアネス)の支援となる。

情報潮流の大きな流れの変化を知るには、ある時間において類似した記事のまとまり(これをカテゴリと呼ぶ)が、次の時間においてどのように変化したかを監視しなければならない。ある時点における情報集合の中から類似した情報のカテゴリを自動的に

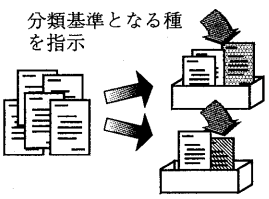

図 1: 教師あり分類

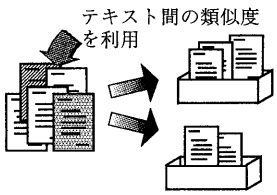

図 2: 教師なし分類

抽出するために分類技術を用いる。まず分類技術について述べ、次にその技術を用いた情報潮流抽出手法を説明する。

3 情報分類技術

テキストの自動分類技術として従来からいくつかの研究が行われている。大きく分けると、別に用意された情報を利用する分類技術と、分類対象とする情報集合のみを利用した分類技術がある。

前者には、第三者が用意した類義語辞書や単語間の階層構造情報を利用して類似するテキストを分類する方法がある [1]。この方法の短所として、辞書のメンテナンスを行う必要があり、また、分類対象となるテキスト情報に応じた辞書を用意する必要がある。

後者のほうは「教師あり分類 (text classification)」(図 1)「教師なし分類 (text clustering)」(図 2) の 2 種類の手法に分けることができる。ここでいう教師あり分類とは、あらかじめ分類する規則やカテゴリを分類対象となる情報から選択して与え、それらに応じて自動的に分類する技術である。また、教師なし分類とは情報集合から規則やカテゴリを自動的に抽出し分類する技術である。

3.1 教師あり分類

分類の条件や分類対象となるテキストの特徴により分類方法がいくつかある。

一つには、キーワードの組合せによる分類規則を明示的に与える方法があるが、あらかじめ適切な規則を構成することは難しい。

これに対し、ユーザが各カテゴリに分類されるテキストの例をシステムに与え、類似するテキストを自動的に分類する方法がある。システムは、割り当てられたテキストからそのカテゴリを代表する特徴を自動的に抽出し、その特徴を用いて分類する (一般に特徴は単語とその重要度の組のベクトルで表現される)。この方法は、ユーザがいくつかのテキスト情報を与えることにより分類を行うため、明確なキーワード等を指定する必要がなくユーザには負担が少ない。代表的な手法として最近傍決定則 (NN 法) やニューラルネットワークを用いる手法がある。

3.2 教師なし分類

教師なし分類は、分類カテゴリを自動的に抽出し分類する方法で、ユーザが分類対象の情報集合に関する知識をあらかじめ持つ必要がない。一般には、すべてのテキストに対し特徴ベクトルを作成しておき、それを用いて分類を行う。

その利用法として、一つは統計的処理を利用する方法がある。特徴ベクトルの要素の値の分布を調べ、その分布状況によりテキストを分類する。これには、多変量解析で用いられている主成分分析やクラスター分析などがある。特徴が数的に表現された情報を整理する場合に従来から多く行われている手法である。

また、従来からのアプローチと異なり、Kohonen [3] の提案した自己組織化マップの手法を利用する方法もある。これは、中間層のない 2 層型のニューラルネットワークで教師なし競合学習モデルである。問題点としては、分類結果が初期値に依存する部分が多く、また計算コストも高い。

3.3 分類技術の選択

従来の方法から自動分類アルゴリズムを検討する際に、考慮した点が 2 点ある。

一つは分類アルゴリズムの高速性である。各個人が持つ情報の意味の捉え方がさまざまであり、分類結果がシステム利用者の意図と異なる場合がある。そこで、システム利用者が分類結果を修正し、今後その修正を反映させた分類を高速に提示できる方が望ましいと考えた。ニューラルネットワークを用いた手法 [4] は学習に時間がかかり、自己組織化マップや主成分分析は分類結果を出すまでの処理時間が長い。

もう一点は、どのカテゴリにも割り当てられないテキストの抽出を考慮した。新規に発信される記事は従来から存在しないキーワードに関するものも存在し、それらは新規の情報としてどのカテゴリにも割り当てないようにしなければならない。

また、抽出する情報潮流として、テキスト集合に対しユーザがあらかじめ興味を抱いている情報に対する情報潮流と、ユーザが陽に指示していない情報潮流がある。2 種類の情報潮流を抽出するために、あらかじめ分類したい情報を与えて分類する教師あり分類手法と、自動的に分類する教師なし分類手法の 2 つが必要となる。

処理の高速性とどのカテゴリにも割り当てられな

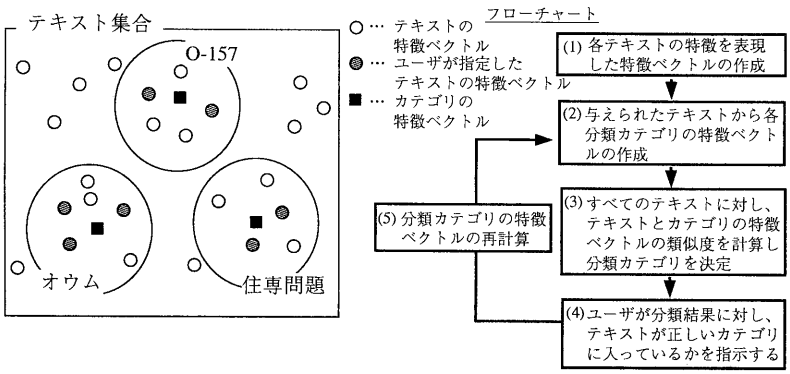

図 3: NN 法を用いた分類の概略図

いテキストの抽出が可能な手法として、教師あり分類手法としてNN法、教師なし分類手法としてクラスター分析を用いることにした。

4 トレンド・アウェアネス支援手法

4.1 特徴ベクトルの作成

分類を行うための前処理として、各テキストに対し、その特徴を表現した特徴ベクトルを作成する特徴ベクトルを自動的に作成するため、テキスト内に存在する単語とその出現頻度を抽出し、以下の値を獲得した。

$$TF_{ij} \quad \ldots \quad \text{テキスト } i \text{ での単語 } j \text{の出現数} \quad (1)$$

$$IDF_j \quad \ldots \quad \text{単語 } j \text{が出現したテキスト数} \quad (2)$$

$$L_j \quad \ldots \quad \text{単語 } j \text{の長さ (文字数)} \quad (3)$$

単語の切り出しには形態素解析エンジン InfoBee/TC [5] を使用した。以上を用いてテキスト i に対する特徴ベクトル$\vec{FV_i}$を、

$$\vec{FV_i} = (w_{i1}, \cdots, w_{ij}, \cdots, w_{iN}) \quad (4)$$

$$w_{ij} = TF_{ij} \cdot log\left(\frac{M}{IDF_j}\right) \cdot log(L_j) \quad (5)$$

とした (Nはテキスト集合における全単語数、Mはテキストの総数)。テキスト内において単語の出現回数が多い場合、w_{ij}の値は大きくなる。しかし、その単語がどのテキストにも存在するような単語で在った場合、IDF_jを用いてw_{ij}の値を小さくする。L_jの項は、単語の文字列が長いものほど分類するための要因として利用できると考え使用した。また、各テキストの長さによる影響を減らすために、特徴ベクトルの要素の値を正規化しておく。

次に、各テキスト間の類似度を定義する。テキスト i とテキスト jの類似度 $Sim(i,j)$ はベクトルの内積を利用して、

$$Sim(i,j) = \sum_{k=1}^{N} w_{ik} \cdot w_{jk} \quad (6)$$

とした。この関数を用いて、$Sim(i,j)$ の値が大きいテキスト同士は類似しているとする。

4.2 NN 法

NN 法とは、各テキストと分類するカテゴリに対して特徴ベクトルを作成し、カテゴリとテキストの特徴ベクトル間の類似度に応じて分類する手法である。ユーザがシステムに対し関心があるキーワードを含むテキストとそれを割り当てるカテゴリを指定すると、システムが指定されたテキストから各カテゴリの特徴ベクトルを作成し分類を行う。今回、カテゴリの特徴ベクトルはそれぞれ一つとし、その値はユーザによって与えられた各テキストの持つ特徴ベクトルの平均とした。カテゴリとテキストの類似度は、テキスト間の類似度と同様、ベクトルの内積を用いた。アルゴリズムの概略を図3に示す。

また、どのカテゴリとも類似しないテキストを抽出するために閾値を導入した。値は 0 から 1 の間の実数値で、テキストとカテゴリの類似度が閾値より

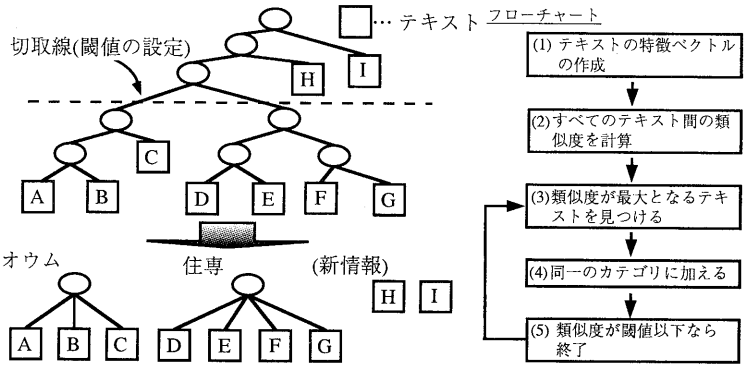

図 4: クラスター分析を用いた分類の概略図

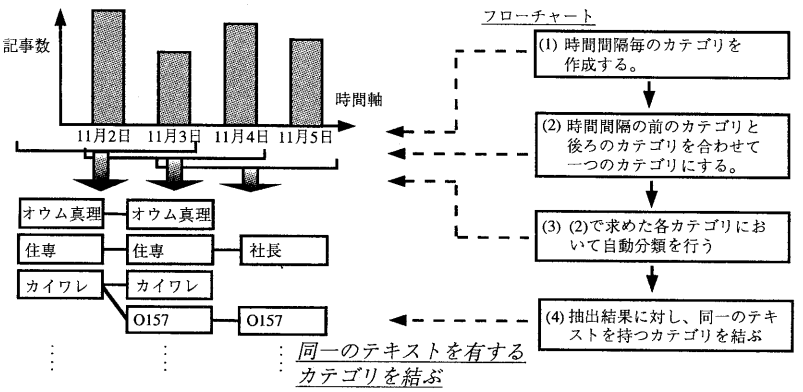

図 5: 情報潮流の抽出の概念図

小さい値であれば、そのカテゴリには割り当てない。この値は、すべてのカテゴリを通して共通とした。

4.3 クラスター分析

本手法で用いているクラスター分析は、テキストの特徴ベクトルを用いてテキスト間の類似度を計算し、その値が大きいものから同一のカテゴリに割り当てていくというものである。この手法は、どのカテゴリにも属さないテキストを抽出するのが容易で、また、分類結果の出力が木構造になるため、木の分岐点で切り取ることにより、分類されるカテゴリの数を自由に変化させることができる。このアルゴリズムの概略を図4に示す。

一般的なクラスター分析の手法では、図4のフローチャートの (3) から (5) のループにおいて、カテゴリに新たにテキストが加わった場合、そのカテゴリを代表する特徴ベクトルを更新し、次の分類に利用する。しかし、特徴ベクトルの値の更新を行うと、そのカテゴリと分類されていない他のテキストとの類似度を計算する必要が出てくる。そのため、本手法ではカテゴリの特徴ベクトルを更新せずに、最初に求めた各テキスト間の類似度のみを用いて分類することにより高速化を図った。

また、自動抽出されたカテゴリには、どのような

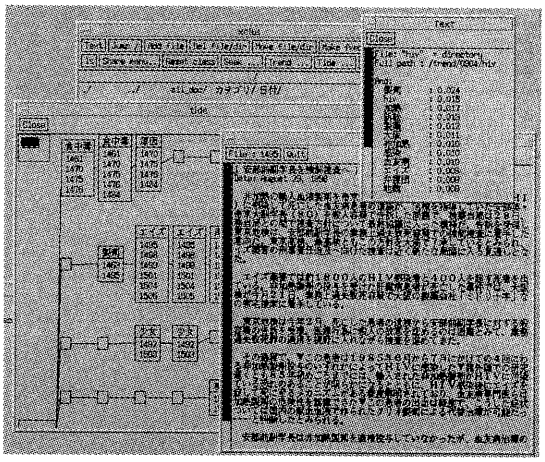

図 6: Xwindow 上の試作システム

キーワードが入っているか把握できることが望ましい。今回はカテゴリの特徴ベクトルの要素で値が最大となった単語を、カテゴリを代表するキーワードとした。

4.4 情報潮流抽出手法の検討

情報潮流を視覚化するには、各時間毎のカテゴリが次の時間にどのように変化したかを示さなければならない。最初に次のような手法を考案した。

各テキストが作成された時間を用いて、ある時間間隔毎にテキストを区別しておく。各時間間隔毎に分類を行いカテゴリを抽出する。類似するカテゴリ同士を時間間隔の順に結合させれば、時間毎にどのように情報が変化し分岐していったかが抽出できる。

しかし、この手法では時間間隔毎のテキスト集合から抽出したカテゴリを時間毎に結びつけるのは正確さに欠ける。なぜなら、情報の転換を例に挙げると、時間間隔を境にして抽出されたキーワードが異なった場合、実は過去の情報から派生したカテゴリにも関わらずそのカテゴリ同士を結ぶことができなくなる恐れがある。そこで、さらに時間間隔毎の区切りをオーバーラップさせて分類する手法を考案した。潮流抽出手法の概念図を図 5に示す。

この手法では、図 5の (2) において時間間隔毎に区別したカテゴリとその前後のカテゴリ内に同一の記事が存在し、その記事を基にしてそれぞれの時間間隔毎のカテゴリで類似するテキスト情報の収集ができる。そのため、時間間隔毎でキーワードが変化したカテゴリ (情報の転換) や、過去のキーワードから複数の情報に分岐したカテゴリ (情報の分岐) などを正確に結びつけることができる。

新聞記事数	465 件
正解カテゴリ (RC_i)	70 種類
システムの出力 (SC_j)	61 種類
$RC_i \subseteq SC_j$ となった組 (i, j) の割合	70% (49/70)
システムのみ抽出したカテゴリ	31%(19/61)

表 1: 試作システムの分類結果

	抽出結果	
	SC_j $(j = 1, \cdots, 61)$	未分類カテゴリ
正しい分類	185 件 (1) 39.8%	136 件 (2) 29.2%
誤った分類	98 件 (3) 21.1%	46 件 (4) 9.9%

表 2: 分類の正誤表

5 試作システムの評価

本手法の有効性を確認するため、図 6に示す情報潮流抽出システムを試作し評価を行った。

5.1 分類精度

ここでは、教師なし分類 (クラスター分析) による分類評価について説明する。テキスト集合はロイタージャパンの記事を利用し、1997 年 2 月 8 日から 2 月 14 日の 465 件 (約 1 週間分) を対象とした。各記事は、本文と見出し、簡単なカテゴリ (記事を代表するキーワード) が一つ付加されていたが、分類技術の汎用性を調べるためにテキスト情報は本文のみを利用した。あらかじめ付加されているカテゴリ情報は 267 種類あり、一つのテキストから成るカテゴリが 197 種類、複数のテキストから成るカテゴリは 70 種

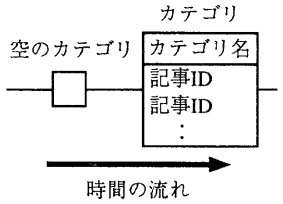

図 7: 出力図の見方

類であった。この 70 種類のカテゴリを今回正解カテゴリとする (これを $RC_i(i = 1, \cdots, 70)$ とする)。

閾値を 0.2 として試作システムに自動分類させた結果を表 1に示す。抽出されたカテゴリ (これを SC_j とする) は 61 種類で、分類されなかった記事は 182 件であった。システムが自動的に抽出したカテゴリに含まれる正解カテゴリの数、すなわち

$$RC_i \subseteq SC_j \text{となった組 } (i,j) \text{ の数}$$

は 70 種類中 49 種類であり、正解カテゴリの 7 割をシステムが抽出できたといえる。さらに、49 種類のカテゴリのうち 27 種類のカテゴリについては $RC_i = SC_j$ となった。抽出できなかったカテゴリとしては、「交通事故」(12 件)「焼死」(4 件)「受賞」(2 件) などがあり、これらのカテゴリ内のテキストを見ると、それぞれ単発に発信された記事がほとんどで、ある特定の内容を継続して発信しているものではなかった。

次にシステムが抽出したカテゴリ SC_j について評価する。どのカテゴリにも割り当てられなかったテキストはすべて未分類カテゴリに割り当てられているとした。システムの抽出した結果に対し、正解カテゴリを参考にして正誤を付けた結果が表 2である。正誤の基準は、SC_j 内において各テキストの持つ最も多い RC_i のカテゴリ名を SC_j の正しいカテゴリとし、それ以外のカテゴリ名を持つテキストは分類誤りとした。

表 2より、誤って分類された記事数は (3)+(4)=144 件で 31.0%となった。記事ではなくカテゴリを用いて誤りを考えると、誤りを含んでいる SC_j は 34 種類 (55.7%) であった。すなわち、各カテゴリに広く渡って記事が少数づつ誤って分類されていると考えることができる。この誤りは自動分類の際に用いる閾値の値により変化させることができ、閾値の調整を行ってより向上させる必要があると思われる。

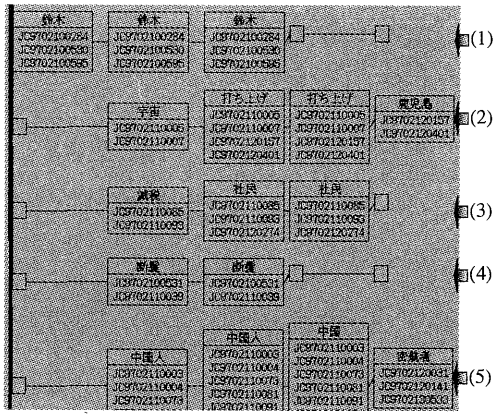

図 8: 新聞記事に対する処理結果

5.2 情報潮流

次に抽出された潮流について説明する。図 8は、ロイターの新聞記事の出力結果である。情報潮流抽出の際の時間間隔は一日毎とし、分類技術は高速化したクラスター分析手法のみを用いている。

出力図の見方は図 7となっている。記事の ID が複数書かれた四角一つがカテゴリで、その一番上にカテゴリを代表するキーワードを表示している。横方向に左から右へ時間が進んでおり、カテゴリのつながりが特定の情報に対する潮流を表している。空のカテゴリは、表示の際に時間間隔毎にそろえるために利用し、空のカテゴリから記事 ID のあるカテゴリへのつながりは、新情報の発信を表している。

図 8の (2) は、分類された記事の内容を見ると「M5 ロケット打ち上げ」に関する情報潮流であった。各カテゴリ内のテキストを見ると、「宇宙」のカテゴリには衛星 MUSES-B の説明の記事があり、「打ち上げ」のカテゴリで衛星を積んだ M5 ロケットの打ち上げの延期の記事が、「鹿児島」のカテゴリでは鹿児島宇宙観測所の記事があった。また、各カテゴリ内で上位を占めるキーワードには「打ち上げ」「鹿児島」「ロケット」「観測所」などがあり、キーワードを見ることによって (2) がロケット打ち上げに関する情報潮流であることが把握できると思われる。

インターネット上のネットニュースの記事に対しても実験を行った。データは fj.rec.autos の 10 日分 (記事数約 400 件) の記事を利用した。図 9がその抽

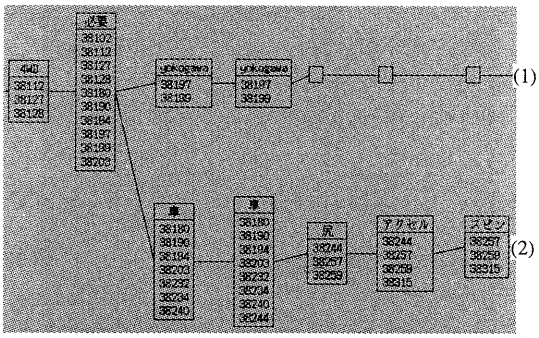

図 9: NetNews に対する処理結果

出例である。これは、「北国で生活する場合 4WD が必要か」という内容の記事が発端となっている。図9において、(1) では「スノータイヤの必要性」、(2) では「4WD の車の方が安全」「急ブレーキではスピンする」「ある車はスピンしやすい」という記事が続いている。この例では情報の分岐が抽出できていることを示している。このように、最初に提案した情報潮流が抽出できた。

6 今後の課題

情報潮流として提案した 5 つの潮流において、どのカテゴリにも属さない新情報の発信の抽出が最も困難である。本手法では唯一つのテキスト情報に対して新たなカテゴリを作成していない。どのカテゴリにも属さないテキスト情報を唯一つの情報としてしまうと、分類技術の誤りによるテキスト情報も唯一つの情報として判断しかねないからである。多くの類似したテキストが発信される情報に対してだけではなく、唯一つのテキスト情報も抽出する手法を検討する必要があると考えている。

また、各個人の持っている単語に対する意味付けの違いから、一般的なテキストの分類は困難である。今後は、各個人がシステムの分類結果に対して正誤の指示を与えることにより、分類手法に導入した閾値の調整などを含めて、分類結果を各個人にカスタマイズしていけるシステムを構築したい。

また、図 8や図 9で示した情報潮流は、各カテゴリに付加されたキーワードによってある程度判断が可能である。しかし、各個人の持つ単語の意味の違いがあるため、カテゴリを代表するキーワードの提示方法を考える必要がある。このカテゴリに付加されたキーワードを利用することによって、時間軸方向の情報圧縮として提示できるように考えていきたい。

7 まとめ

本稿では、常に新たなテキスト情報を発信する情報メディアに対して見い出した 5 つの情報潮流を提示し、それを抽出するために必要となる情報分類技術について述べ、情報潮流抽出手法を提案した。提案した手法を確認するためにシステムを試作し、実際の新聞記事や NetNews の記事を用いて実験を行い、提案した情報潮流が抽出できている事を示した。

謝辞

本実験のためにデータを提供していただいたロイタージャパン株式会社に感謝致します。

参考文献

[1] 山本, 増山, 内藤: 分類体系相互の関係を利用したテキストの自動分類, 情報処理学会 NL106-2, pp.7-12(1995)

[2] C. Apte, F. Damerau: Automated Learning of Decision Rules for Text Categorization, ACM TIS, Vol 12, No.3, pp.233-251(1994)

[3] Teuvo Kohonen: The Self-Organizing Map, Proceedings of the IEEE, Vol.78, No.9, pp.1464-1480(1990)

[4] 豊浦, 小船, 有田: 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生成, 情報処理学会 NL88-6, pp.41-48(1992)

[5] 田中: InfoBee 検索エンジンを用いたディレクトリ検索サービス, NTT 技術ジャーナル 8 月号, pp.24-27(1996)