情報潮流抽出のための分類精度の改善手法について

3Y－5

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

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、インターネットのネットニュースなど日々刻々と新しいテキストが発信されている。これらの情報メディアから、自分が必要な情報を取り逃さないようにするためにメディアから発信されるすべての情報に目を光らせようとす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、時間的にどのように変化しているかを視覚化する手法を研究している [1]。情報潮流を抽出するための分類手法について検討を行っており、本報告ではその検討内容について報告する。

2　情報潮流とその抽出手法

最初に、情報潮流とその抽出手法について説明する。対象としている情報は、インターネット上のネットニュースや新聞社などの World Wide Web 上のニュースサービスのような、常に新しく発信されるテキストである。これらのテキストには、その内容を端的に表現するキーとなる複数の単語 (キーワード) が存在しており、新たに発信されるテキストには過去のテキストと同一のものや類似したキーワードが混在し、時間が経過するにつれ登場しなくなるキーワードもある [2]。このキーワードの時間的な変化の様子を「情報潮流」と読ぶ。

情報潮流を抽出するためには、ある時間において類似する内容のテキストのまとまり (カテゴリ) が、次の時間においてどのように変化したかを知る必要がある。類似した情報のカテゴリを自動的に抽出するために、テキストの自動分類技術を利用する。

また、情報潮流の抽出手法は、各時間の区切りをオーバーラップさせてテキスト情報集合を作成し分類する手法を考案した (図 1)。この手法では、時間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するため、その記事を基にしてそれぞれの時間の区切り毎から類似するカテゴリを一つの情報潮流として収集することができる。

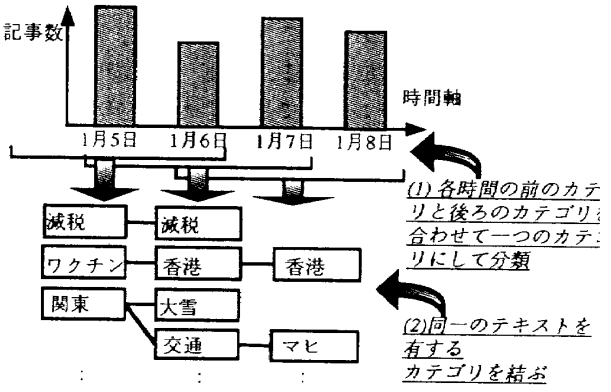

図 1: 情報潮流の概念図

図 2は、自動的に抽出した情報潮流の一つの例である。データは、ロイタージャパンの記事で、1997 年 6 月 27 日から 7 月 2 日の約一週間分の記事 971 件を用いた。図 2において、時間は左から右へ進んでおり、四角一つが各時間毎に抽出されたカテゴリで、表示している単語はそのカテゴリを代表する単語である。この情報潮流は「小学生殺人事件」関する記事が多く割り当てられていた。

3　潮流抽出時の問題点

各時間でのテキスト分類手法には最近傍決定則 (図 3) を利用している。これはテキスト情報間に類似度を設定し、類似しているテキスト同士を同じカテゴリに分類する手法である。このとき閾値を導入し、テキスト間の類似度が閾値を超えるまで分類を行う。

適切な情報潮流を抽出する際に問題となるのは、(1) 分類結果の精度 (2) 情報潮流抽出時の精度の 2 点がある。図 1にある情報潮流抽出技術は、各時間毎における分類結果を利用して情報潮流を抽出しており、分類結果に誤りがある場合、情報潮流の抽出がうまくいかない場合が生じる。その例が図 2である。

図 2は、「小学生殺人事件」に関する情報潮流であったが、「殺人事件」とは直接関係ない「デンバー」や「サミット」のカテゴリが混在していた。この理由を調べる

Improving clustering accuracy for topic stream extraction
Masayuki SUGIZAKI, Takafumi INOUE,
Masaaki OHKUBO, and Kazuo TANAKA
NTT Human Interface Laboratories

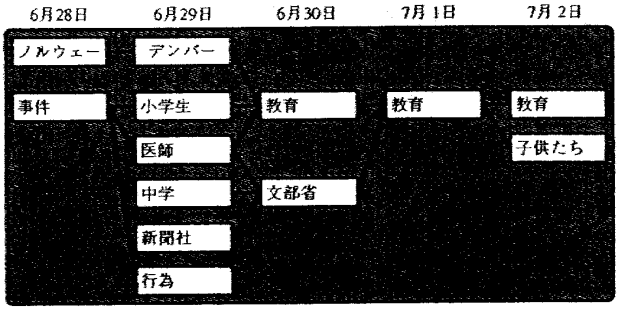

図 2: 情報潮流の抽出結果 (1)

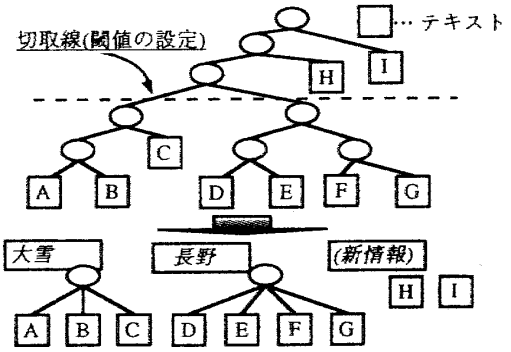

図 3: テキスト自動分類技術

と、6 月 28 日の「サミット」には“首相のサミット参加”の記事が分類されていた。6 月 29 日の「デンバー」には“首相のサミット参加”の記事と“首相の殺人事件へのコメント”の記事が分類されていた。6 月 30 日の「教育」には“首相の殺人事件へのコメント”の記事があった。すなわち、6 月 29 日の「デンバー」のカテゴリに“首相の殺人事件へのコメント”の記事が存在したために、これらが一つの情報潮流として抽出されていた。

システムが自動的にカテゴリ抽出を行うために、システム利用者が考えている分類結果を抽出する事は困難であり、実際に上記の例は閾値の調整で適切に分類出来ない。

4 潮流抽出手法の改善

前記の問題点を解決するために、図 1の情報潮流抽出手法において各カテゴリを繋ぐ際に距離を導入する。各カテゴリに存在する複数のテキスト情報を用いて、カテゴリを代表する単語とその重みが対となった特徴ベクトルを作成し、図 1の (1) でカテゴリ同士を繋ぐ際にそのカテゴリの代表ベクトル間で距離計算をする。ここで閾値を導入し、距離と閾値との大小を評価して、条

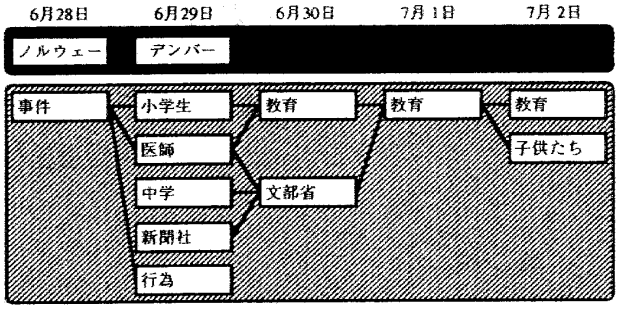

図 4: 情報潮流の抽出結果 (2)

件を満たす場合にそのカテゴリは同じ情報潮流のカテゴリだと判断する。

この手法を評価するため実験を行った。今回の実験では、カテゴリの代表ベクトルは、カテゴリ内に存在する各テキストの特徴ベクトルの平均とした。各テキストの特徴ベクトルは Salton[3] の $tf * idf$ を用いて作成した。また、カテゴリ間の距離はベクトルの内積を用いた。図 2を抽出した時のデータを用いて潮流の抽出を行ったところ、その結果は図 4になった。この例では「デンバー」と「教育」とのカテゴリの距離が閾値の条件を満たさず、別の情報潮流として区別することができた。

5 考察

今回、時間毎の各カテゴリを結ぶ際に距離と閾値を導入することにより、一つの情報潮流内で複数の情報潮流を区別できることを確認した。これにより、「一つの大きな情報潮流の中の細かな複数の情報潮流」であるか「個々が別々の複数の情報潮流が集まった一つの情報潮流」であるかの判断はシステム利用者に依存するが、それをシステム側から提示することが可能であると思われる。今後は、今回導入した閾値の調整手法とその分類結果の評価、および、閾値を用いた情報潮流の分割結果の視覚化手法を検討していきたい。

参考文献

[1] 杉崎ら: トレンド・アウェアネスのための情報潮流の抽出, 第 55 回情処全大会, 分冊 3, 3-226, 1997

[2] 巌寺, 菊井: トレンド・トラッキング型テキスト自動分類の試み, 情処研報 NL119-4, pp.19-24, 1997

[3] Gerard Salton: Automatic Text Processing, pp.280, Addison Wesley, 1989