トレンド・アウェアネスのための情報潮流の抽出

5Q−1

杉崎 正之　大久保 雅且　井上 孝史　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技術を研究しており、本報告ではその抽出手法について述べる。

2 情報潮流

最初に、情報潮流とは何かについて説明する。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が World Wide Web 上で発信している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情報である。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が存在している (以後、これらをキーワードと呼ぶ)。また、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存在するという特徴がある [1]。テキスト集合内におけるキーワードの時間的な変化の様子を以下のようにまとめた。

(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り、情報潮流（トレンド）の大きな流れや変化の様子を視覚化し、その結果を見ることによりどのような情報が有益であるか気付くこと（アウェアネス）が可能となる。

3 テキスト自動分類技術

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する内容のテキストのまとまり (カテゴリと呼ぶことにする) が、次の時間においてどのように変化したかを監視する必要がある。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抽出するために、テキストの自動分類技術を導入する。分類結果の加工の容易さとどのカテゴリにも割り当てられないテキストの抽出が可能な手法として、クラスター分析を用いることにした。

この分類手法は、2つのテキスト間に類似度を導入し、その値を用いて類似するテキストを共通のカテゴリに分類するという手法である。この処理により、図1のように各テキストを葉に持つ木構造を作り出すことができ、閾値を導入して枝を切り取ることにより、類似したテキストのカテゴリを作成することができる。

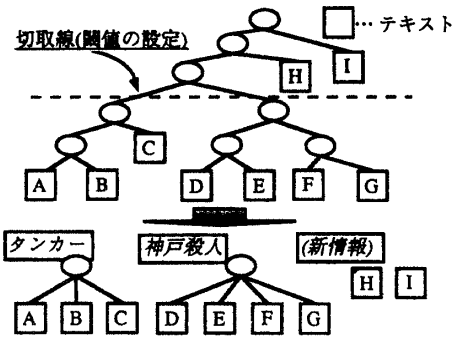

図 1: クラスター分析を用いた分類結果のイメージ

A Method for Trend Awareness based on Extracting Topic Streams
Masayuki SUGIZAKI, Masaaki OHKUBO, Takafumi INOUE, and Kazuo TANAKA
NTT Human Interface Laboratories

4　情報潮流抽出手法

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き、それぞれにおいて自動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。すなわち、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、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、図2に示すよう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。

この手法では、図2の(1)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。

5　抽出結果

本手法の有効性を確認するために、情報潮流抽出システムを試作した。図3が出力例であり、インターフェースはWWWのブラウザを用いた。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3日までの一週間分の記事971件を用いた。図3において、時間は左から右へ進んでおり、表示されている単語は抽出されたカテゴリを代表する単語である。(1)～(7)がそれぞれ抽出された情報潮流であり、例えば(4)は「女子プロゴルフのトーナメント」に関する情報潮流であり、(6)は「神戸殺人事件」に関する情報潮流であった。

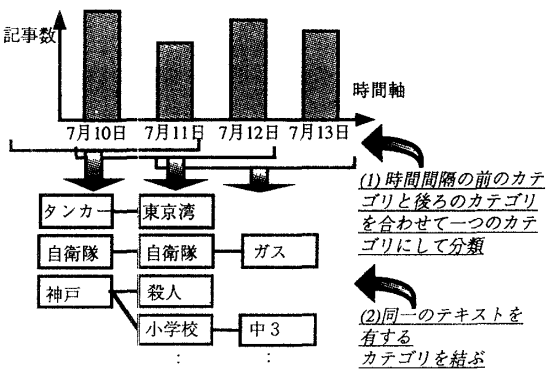

図 2: 情報潮流の概念図

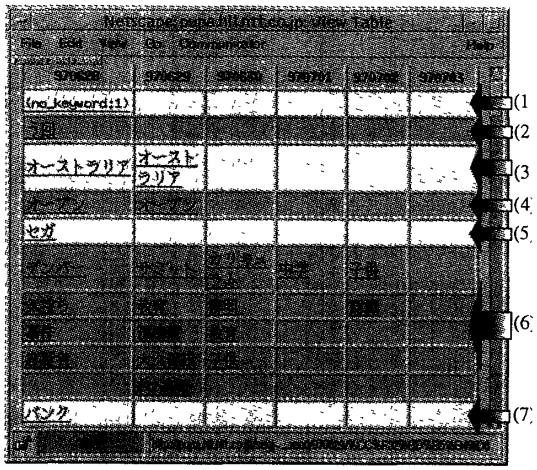

図 3: 情報潮流の抽出結果

6　考察

出力結果を見ると、図3の(6)において「神戸殺人事件」とは直接関係ない“デンバー”や“サミット”のカテゴリが存在する。これは「神戸殺人事件」に関する「首相のコメント」の記事が存在し、その記事がきっかけでこれらが一つの情報潮流にまとまっている。この問題は単純に分類精度の向上により解決する部分ではなく、“サミット”に関する記事と「神戸殺人事件」に関する記事を別の情報潮流とするには、これらを切り放すための情報をなんらかの形で実現する必要がある。ユーザが求めている情報潮流を抽出する手法を検討する上で、切り放すための情報を獲得する手段を考えるのが今後の課題である。

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] 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