

4 U-3 色彩を用いた複数のテキスト情報の差分分析

片山佳則

(株) 富士通研究所 ドキュメント処理研究部

e-mail: kata@flab.fujitsu.co.jp

1. はじめに

大量で、雑多なテキスト情報から役に立つ情報を取り出すための技術としてテキストマイニングが注目されている。テキストマイニングでは、情報を検索する技術に加えて、検索された情報を役立たせるための、分析技術も必要である。

本稿では、これらの分析技術を備えたテキストマイニングツールの出力結果を、読み取り易くするための差分表示方法を提案する。

2. 情報分析技術

検索された情報を役立たせるための、分析作業に活用できる技術として、既に文書群から抽出された単語の間の「連想関係」を単語の共起性に基づいて計算し、マップとして可視化するテキストマイニングツール「ACCENT」¹⁾²⁾³⁾が開発されている。

ACCENTは、出力結果である単語の連想マップを作成する際の、(1)マップ構成単語の選択、(2)単語間の関連度計算、(3)マップレイアウトのそれぞれに対して、分析目的を反映する機能が用意されている。これらの機能を用いて、連想関係が抽出されるスコープを制御し、分析の目的や視点を反映した結果が得られる事例が示されている³⁾。

また、このACCENTの連想関係の可視化方法として、ランキングリストや連想関連マップが提案され、操作として、頂点固定や関連の間引きが提案されている。これらの可視化技術により、個々のテキスト情報からは把握が困難な情報が得られる事例が示されている¹⁾。

これらのテキスト情報の分析では、複数のテキストが対象となり、それらには系列的な関係が含まれることが多い。従って、情報分析出力

の系列的な変化(差分情報)を利用者に分かりやすく提示することが重要になる。

3. 差分分析対象の観点

分析出力を利用者に分かりやすく提示するためには、出力の観点を規定させる必要がある。ここでは、“要素の選択”と“見せ方の選択”の2つの切り口で活用すべき観点を明らかにさせる。前者の“要素の選択”は、文献3)の単語集合の設定機能を活用する。

(1) 表示要素規定による観点の明確化

次の2種類の絞り込みを行い観点を規定する。

- 単語の属性指定により出力結果である連想マップに用いる単語を規定することで観点を絞り込む。例えば、単語の属性として語尾を「サ-ビス」、「システム」、「戦略」、「事業」などに規定することで観点を絞る。
- 指定した特徴キーワードや属性単語に関連する単語に規定することで、観点を絞り込む。例えば、社名や商品名などの特徴キーワードや年代や、性別などの属性単語を指定し、それらに関連する単語に規定することで観点を絞る。

(2) 見せ方規定による観点の明確化

次の2種類の分析方法により観点を規定する。

- 入力テキスト情報に与えられている時系列に合わせて出力する連想マップを規定する。
- 入力テキスト情報を時系列以外の属性で区別させておき、出力する連想マップを規定する。

これらの規定方法により、差分分析結果の活用すべき観点を明確にし、分析出力である連想マップの分かりやすさを実現する。

4. 差分分析への色彩の活用方法と事例

複数のテキスト情報から構成される連想マップを比較するこれまでの方法は、(1)個々の連想マップを並べる(例えば、年代毎の連想マップを並べる)、(2)特徴部分(共通部分/差分)を切

Knowledge Discovery from Text Information Using
Color Discrimination
Yoshinori KATAYAMA
FUJITSU Laboratories LTD.

り出して提示する(例えば、連想マップの要素となる重要キーワードのランキング)などが採られてきた。個々の連想マップからは、情報を容易に得られるが、系列のある連想マップ間の情報を読み取ることは、困難であった。

ここでは、以下の2種類の色彩利用を行うことで、系列のある複数の連想マップを利用者に分かりやすく提示する。

(A) 系列連想マップの差分に使用
連想マップが、時系列で得られる場合、新たに生まれた要素や、すべてに共通する要素などを強調するために色彩を用いることで、全体の中での差分情報を容易に識別できる。

(B) 連想マップの要素の位置変化に使用
連想マップとして頂点固定マップを用いた場合に、系列内での要素の位置変化を色彩として強調させることで、動向や関連の変化のキーワードとなる要素を容易に識別できる。

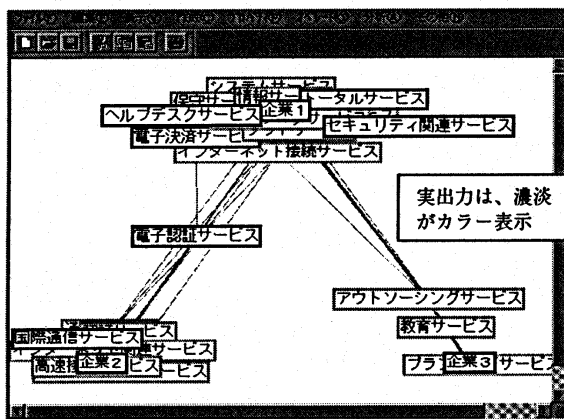


図1 出現キーワードの時系列差分表示結果

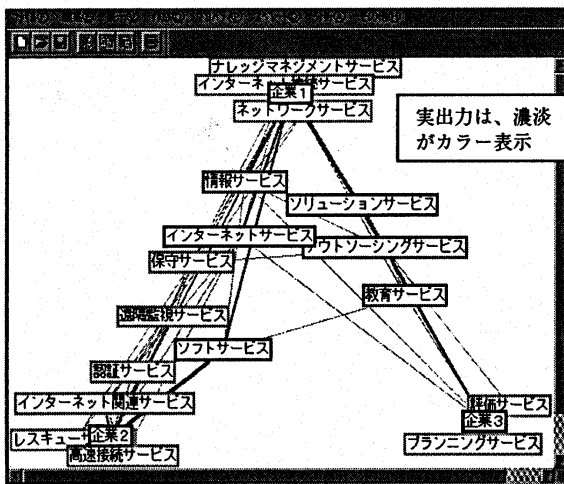


図2 系列的関連位置関係の差分結果

以下の事例は、あるメーカーが関連するニュース記事(1998年~1999年)をWEB上で集めたもので、1194記事、12733単語を半期ごとの時系列として連想マップを作成したものである。ここでは、要素の観点を「~サービス」に絞り込んで分析を進めた。

図1は、(A)の事例であり、時系列の連想マップの差分結果を、1998年上半期の連想マップ上で色彩表示させたものである。各企業のこの時期の新たな取り組みのキーワードを把握でき、各社の必要なアクションの判断に役立てられる。

図2は、(B)の事例であり、1999年上半期から下半期にかけてある一定以上の位置変化のあった要素に色づけされている。位置変化状況の確認は必要だが、これにより変化のあった要素を容易に把握でき、業界動向の判断や戦略的な企画立案につなげることができる。

5. 効果の整理

連想マップに観点と色彩を取り入れたことで、以下のビジネス的観点での効果を期待できるとの意見を頂いている。

- ・ ビジネス動向の明確化: 他社動向や業界動向などの変化を際立たせられる。
- ・ ビジネス情報の差別化: 共有情報と特殊情報を区別し、ビジネス戦略に役立てられる
- ・ 注目ポイントの明確化: ビジネス関連情報などでは、積極的に見せたい方向(トップの意向)を際立たせるために役立てられる。

6. おわりに

差分情報の色彩活用は、分かりやすさと漏れを無くし、動向や状況把握には有効な機能である。さらに観点の絞り込みは、情報の利用価値を高めるために、ノウハウとして蓄積される切り口である。

今後は、テキスト情報の処理に不可欠な分類技術のソリューション提供にこれらの技術を活用させる。

参考文献

- 1) 三末, 渡部: テキストマイニングのための連想関係の可視化技術, 情報処理学会 第55回情報学基礎研究会, 1999年7月
- 2) Symfoware Text Mining Server(富士通(株))
- 3) 渡部, 三末: 単語の連想によるテキストマイニング, 情報処理学会 第55回情報学基礎研究会, 1999年7月