

動向情報の検索による情報編纂

The Information Compilation by Searching Trend Information

山本 健一*¹ 谷岡 広樹*¹ 殿井 加代子*¹
 Kenichi YAMAMOTO Hiroki TANIOKA Kayoko TONOI

*¹株式会社ジャストシステム
 JustSystems Corporation

In recent years, We can gather enough information of our basic interest based on search technology. But it is still difficult for us to gather information of our complex interest: for example “What company has the synchronizing stock price with the given company’s one?” or “What term has the synchronizing frequency in a newspaper with the sales of given product?”. So we propose a visualization system and interaction. The proposed system is based on similarity score between various trend information.

1. はじめに

計算機の処理能力の向上や高速ネットワーク環境の普及に伴い、電子化された情報は増加の一途を辿っており、この傾向は今後も継続するものと思われる。そのため、ユーザの関心や興味に合致する情報に直接的かつ簡便にアクセスするための技術が求められている [6]。このような要求に答える技術のひとつとして、我々は、動向情報の変化とその変化要因とを視覚的に表示するシステムを研究してきた [11][12]。

我々の以前のシステムは、毎日新聞コーパスと内閣支持率をシステムに入力することにより、図 1 のような SVG (Scalable Vector Graphics) ファイルを出力する。図 1 は、横軸に時間を、縦軸に支持率をとった内閣支持率の折れ線グラフである。さらに、ユーザ実験の結果、内閣支持率の変化の要因を知りたい箇所として明らかになった値の変化が大きい部分とその前後や、値が最大の位置と最小の位置、グラフの最初と最後に内閣支持率の変動に影響を与える新聞記事から抽出した重要語を表示している。なお、これらの箇所以外でも、マウスオーバーすることで重要語を確認することができる。また、表示されている重要語をクリックすることにより、表示内容が重要語から新聞記事のタイトルに変化し、表示された新聞記事のタイトルをクリックすることにより、実際に新聞記事を閲覧できるシステムとなっている。

しかし、動向情報を用いて様々な分析を行う際には、単一の動向情報のみを用いて分析を行うことは少なく、例えば内閣支持率と日経平均株価の変動など複数の動向情報を同時に分析する必要がある。そして、内閣支持率の動向グラフと日経平均株価の動向グラフの形状が正の相関があることがわかれば、内閣支持率を維持するためには、日経平均株価を維持する必要があるという知見が得られる。だが、多種多様な動向情報を 1 つのグラフ描画領域上に表示したのでは、無数のグラフが重なり合い見にくいことが問題となる。そこで我々は、関連し合う動向情報のみを効率的に提示し、分析の支援を行うシステムの研究を行っている。我々のシステムを用いることにより、例えば「ある会社の株価の変動と同期している株価をもつ会社を探したい」や、「ある製品の売り上げの変動と共に使用されるようになった単語を知りたい」といったニーズに答えることが可能

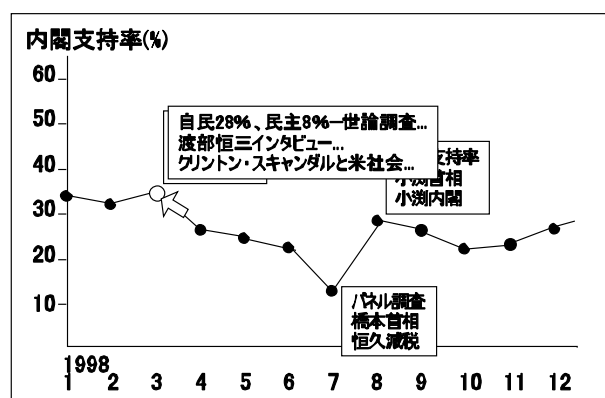


図 1: システムの出力イメージ

となる。

次節では研究の目的と関連研究に関して説明し、第 3 節では、動向情報間の関連度の算出手法に関して説明する。そして、第 4 節で関連する動向情報の抽出実験に関して述べ、第 5 節で提案システムのインターフェースとインタラクションに関して提案する。最後に、第 6 節でまとめと今後の課題に関して述べる。

2. 研究の目的と関連研究

2.1 研究の目的

我々は、先に述べたように「ある会社の株価の変動と同期している株価をもつ会社を探したい」や、「ある製品の売り上げの変動と共に使用されるようになった単語を知りたい」といったニーズに答えるシステムの開発を研究の目的としている。より具体的には、以下のようなシナリオを想定している。

1. あらかじめ準備された動向情報を可視化する。
2. 可視化している動向情報をキーにして、相関のある動向情報を検索、可視化する。
3. 可視化している動向情報を選択して、2 に戻る。

ここでは、動向情報は、大きく以下の 2 種類に分類されるものとする。

連絡先: 山本健一, 株式会社ジャストシステム,
 〒 771-0189 徳島市川内町ブレインズパーク,
 TEL: 088-666-1000, FAX: 088-666-1010,
 e-mail: kenichi_yamamoto@justsystem.co.jp

表 1: 「オリンピック」に相関の高い語

相関の高い単語	相関度 R_{XY}	分散
ディーター	0.9612	0.91
オーストリア	0.9515	3872.19
坂本豪大	0.9512	1.33
原田雅彦	0.9512	344.46
船木	0.9500	907.71
スーパー大回転	0.9495	26.98
野沢温泉	0.9470	72.90
長野冬季五輪	0.9454	1038.89
ルメイ	0.9451	108.96
冬季大会	0.9420	24.87
雪原	0.9410	25.77
リツマ	0.9403	38.21
志賀高原	0.9400	114.72
w杯回転	0.9389	1.75
長野冬季五輪開会式	0.9386	3.29
スキー合宿	0.9381	0.25
長野五輪	0.9367	2755.85
長野オリンピック	0.9365	42.08
スラップ	0.9360	64.71

表 2: 「PAD」に相関の高い語

相関の高い単語	相関度 R_{XY}	分散
松枝	0.9930	0.71
ゲオルグ	0.9697	4.19
中間選挙前	0.9457	0.64
ドニ	0.9337	7.19
セットトップボックス	0.9311	2.58
三島文学	0.9258	2.67
党名	0.9243	78.50
基壇跡	0.9243	5.85
0 . 3 3 5	0.9239	2.25
育児時間	0.9237	2.62
新省庁設置法案	0.9216	2.16
室戸岬	0.9211	0.92
バエズ	0.9197	1.06
j a s 機	0.9197	1.06
ストレス解消法	0.9192	1.16
京阪天満橋	0.9180	0.71
粗塩	0.9180	0.71
執拗さ	0.9173	0.65
殺人マシン	0.9167	0.47

統計動向情報: 時間情報と何らかの統計情報とのペアから構成される動向情報 (例: 内閣支持率, 日経平均株価など) .

頻度動向情報: 時間情報が付与されたコーパス (例: 新聞コーパス, blog など) において, ある単語の出現頻度を単位時間ごとに集計した動向情報 .

本稿では, 任意の動向情報と関連する動向情報の獲得方法と, 提案システムのインタフェースとそのインタラクションに関して説明する .

2.2 関連研究

我々のシステムにおいては, blog や新聞記事など予め時間情報が付与されたコーパスと, 内閣支持率や日経平均株価などの統計動向情報を必要とする . 時間情報が付与されたコーパスを研究の対象とし, 我々のシステムに関連すると思われるものには, 次のようなものがある .

kizashi.jp[3] では, blog をコーパスとし, ある任意の単語をシステムに入力すると, 横軸を時間, 縦軸をその単語の blog 中での出現回数としたグラフを得ることができる . また, その単語と等しい文脈情報を伴って出現した単語を関連語として得ることができる .

このシステムは, 我々のシステムと良く似ているが, 我々のシステムでは, 動向情報の相関に基づき関連語を取得するので, 関連語の取得方法が異なる . さらに, グラフには単語の出現回数に基づく頻度動向情報だけでなく, 内閣支持率や日経平均株価などの多様な統計動向情報が表示できる点でも異なる .

Gruhl ら [13] は, blog と Amazon sales rank data (<http://www.amazon.com/gp/aws/landing.html>) を用いて blog でのある製品の発言回数とその製品の実際の売り上げとの相関を調べ, blog での発言回数の推移から今後の売り上げを予測することを目的とした研究を行っている . 例えば, ある本に関連する blog での発言回数を調べるためには, 人手で本に関連するキーワードをシステムに入力するか, 本のタイトルや著者名をキーワードとする単純なルールを用いる .

我々は, あらゆる動向情報間の相関を算出し相関のある動向情報のみをユーザに提示するのに対して, Gruhl らは, 人手で, またはルールで作成されたキーワードと売り上げデータ間のみの相関を提示する点で異なる .

blogWatcher[4] では, blog をコーパスとし, ある任意の単語をシステムに入力することにより, パースト, 評判情報, 男女推定, パーサス, もしかして?, 行動分析, 関連ニュースといった多様な機能を利用することができる . これらの機能のうち我々の研究に関連するのはパーストに関連する機能である . これは, システムに入力された単語の blog 中での盛り上がり度をグラフとして確認できる機能である . パースト度は, blog データを document stream とみなし, document stream に入力された単語に関連するドキュメントが出現する傾向を元に確率的に算出される値である [14] .

一方, 我々のシステムではパースト度の代わりに, 単に入力された単語の出現回数を用いている . また, 我々のシステムでは, 関連する単語の頻度動向情報や統計動向情報を取得できる .

3. 動向情報間の関連度

一般に, データ列 $X = \{x_i\}$ と $Y = \{y_i\}$ ($i = 1, 2, \dots, n$) が与えられたときに, X と Y との相関を示す値としてピアソンの積率相関係数がある [7] . ピアソンの積率相関係数では, データ列 X と Y との間の相関 R_{XY} を以下のように定義する .

$$R_{XY} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (1)$$

ただし,

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \quad (2)$$

とする .

今後、我々は動向情報 X と Y との間の相関には式 (1) を用いることとする。

4. 関連する動向情報の検索実験

4.1 実験条件

ここでは、任意の動向情報と関連する動向情報の検索実験に関して報告する。

統計動向情報は、1998年1月から1999年12月までの24ヶ月分の統計情報を人手で与えることとする。

頻度動向情報は次のように作成する。まず、1998年、1999年の毎日新聞コーパス(全220,087記事)に対して形態素解析を用いて名詞句抽出を行う。形態素解析器は、隠れマルコフモデルに基づき独自に開発したものである。その結果、異なり語数は1,280,313語であった。その後、それぞれの単語の月ごとの出現回数を算出し、24次元の特徴ベクトルを作成した。なお、24次元のうち22次元以上が0の場合は対象外とした。

以上のように、統計動向情報、及び頻度動向情報の特徴ベクトルを作成し、式(1)を用いて任意の特徴ベクトル間の相関を算出する。

4.2 実験結果

実験の例として、単語「オリンピック」に相関が高かった頻度動向情報を持つ単語を表1に示す。次に、同様に単語「PAD」に相関が高かった頻度動向情報を持つ単語を表2に示す。相関度は、式(1)に基づく値で、分散はそれぞれの単語の頻度動向情報の分散値である。

表1では、「オリンピック」に関連する単語が上手く取れているが、表2では、ノイズが多く提案手法が上手く機能していないことが分かる。

4.3 考察

提案手法はまだまだノイズが多く改善を行う必要がある。ここでは、改善の方向性に関して考察する。表1、及び表2より、それぞれの単語に関して分散を見ると、「オリンピック」と相関の高い語の分散(0.25~3872.19)が、「PAD」と相関の高い語の分散(0.47~78.50)よりも大きいことが分かる。従って、ノイズの除去の1つの指標として分散が利用できそうなことが分かる。また、本システムでは頻度動向情報の変動に着目していることを考えれば、分散をノイズ除去の指標とすることは直感にも一致する。

なお、提案手法の評価に関しては評価手法も含めて今後の課題とする。

5. インタフェースとインタラクション

図2に提案システムのインタフェースを示す。ここでは、ユーザとシステムとのインタラクションを、「 \uparrow 」という統計動向情報の先行指標として「単語B」の頻度動向情報が見られるかもしれない」という知見を獲得するまでを例として説明する。

1. #ユーザ システム
システムを起動する。
2. #システム ユーザ
初期画面を表示する。
3. #ユーザ システム
図2のプルダウンメニュー「B」において、統計動向を選択する(これで検索クエリが統計動向情報を示す)
4. #ユーザ システム
図2のテキストフィールド「C」に、クエリ を入力し「検索」ボタンを押下する。

5. #システム ユーザ
統計動向情報である と関連度の高い動向情報が図2の「A」にはグラフとして、リスト「D」には関連度と動向情報のタイトルを表示する(動向情報のタイトルは、統計動向情報の場合にはその統計量名が、頻度動向情報の場合には単語の表記がそれぞれ該当する。)
6. #ユーザ システム
図2のリスト「D」から「単語A」の頻度動向情報を選択する。
7. #システム ユーザ
「単語A」の頻度動向情報を図2の「A」において強調表示する。
8. #ユーザ システム
図2のリスト「D」から重要と判断した「単語A」の頻度動向情報を選択し、「 \leftarrow 」ボタンを押下する。
9. #システム ユーザ
「単語A」の頻度動向情報を、図2のリスト「E」に追加し、以後「クリア」ボタンが押下されるまで「D」の動向情報とは違った色で表示する(「E」から削除したい場合は、削除したい動向情報を選択し、「 \leftarrow 」ボタンを押下する。)
10. #ユーザ システム
「E」の中から「単語A」の頻度動向情報を選択する。
11. #システム ユーザ
「単語A」の頻度動向情報を図2の「A」において強調表示する。
12. #ユーザ システム
「E」の中から「単語A」の頻度動向情報を選択し、「 $<$ 」ボタンを2回押下する。
13. #システム ユーザ
図2の「A」において、「単語A」の頻度動向情報が2ヶ月分左に移動する(「 $<$ 」ボタンが1回押下されるごとにグラフは右に移動する。)
14. #ユーザ システム
「E」の中から「単語A」の頻度動向情報をダブルクリックする。
15. #システム ユーザ
図2のプルダウンメニュー「B」を毎日新聞に変更し、「単語A-2」と「C」に表示する(プルダウンメニュー「B」を毎日新聞とすることで、クエリを頻度動向情報と解釈する。)
16. #ユーザ システム
「検索」ボタンを押下する。
17. #システム ユーザ
「単語A」の頻度動向情報を2ヶ月過去にずらした動向情報と関連度の高い動向情報が図2の「A」にはグラフとして、リスト「D」には関連度と動向情報のタイトルを表示する。
18. #ユーザ システム
図2のリスト「D」から「単語B」の頻度動向情報を発見する。

以上のインタラクションにより、ユーザは という統計動向情報と「単語A」の頻度動向情報が同期していること、及び「単語A」の頻度動向情報の周期を2ヶ月遅らすと「単語B」の頻度動向情報があることを知り、「 \uparrow 」という統計動向情報の先行指標として「単語B」の頻度動向情報が見られるかもしれない」という知見を獲得できる。

また、今回のインタラクションでは使用しなかったが、図2のリスト「D」、及びリスト「E」から任意の動向情報を2つ選択し、「計算」ボタンを押下することで、選択した動向情報間の相関度を算出し、算出結果をテキストフィールド「F」で確認することも可能である。

6. おわりに

本稿では、多様な動向情報間の関連度を計算することにより、効率的に動向情報を獲得する手法を提案した。その結果、任意の動向情報と関連する動向情報を効率的に取得できる可能性を示すことができた。また、システムのユーザインタフェース

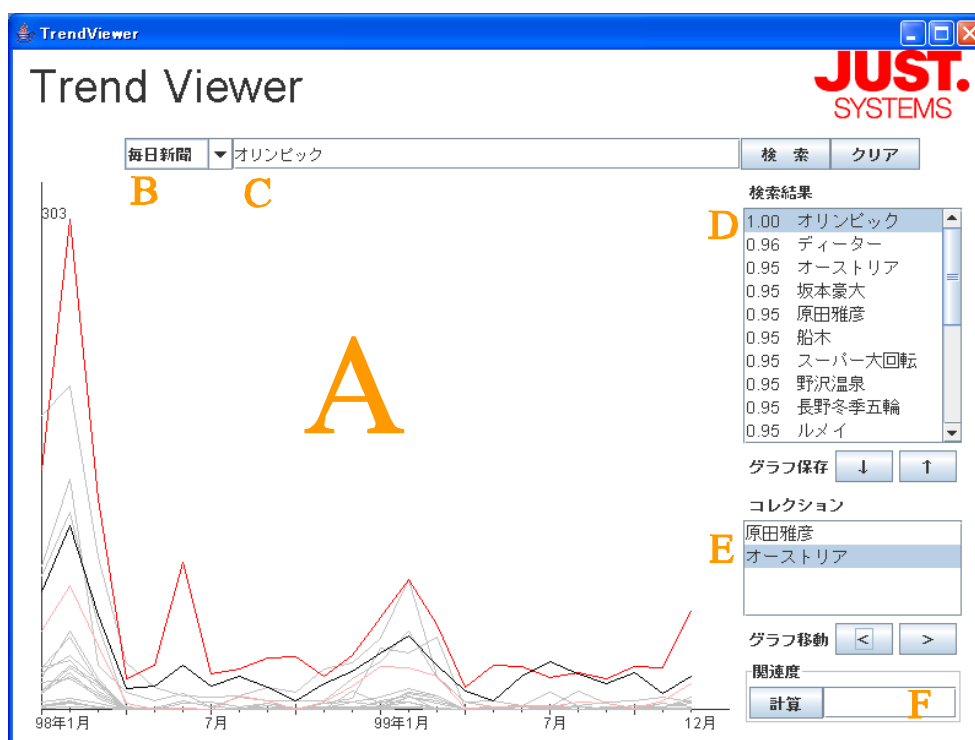


図 2: 提案システムのインターフェース

スとインタラクションの方法を提案した。その結果、関連する動向情報のみを効率的に可視化することができた。今後の課題として、以下の項目が考えられる。

- 統計動向情報と相関のある動向情報の検索実験
- 分散によるノイズ除去手法の開発
- 関連する動向情報の獲得手法の評価方法の検討と評価
- ユーザインタフェースとインタラクションの評価

謝辞

本研究は、動向情報の要約と可視化に関するワークショップ (MuST: A Workshop on Multimodal Summarization for Trend Information) [1] から毎日新聞コーパスやデータの提供を受けました。毎日新聞コーパスやデータを提供して下さった MuST オーガナイザーに感謝します。

第 2 回 MuST 成果進捗報告会では、本研究に対して多数のご意見、コメントを頂きました。ご意見、コメントを下さったすべての方に感謝します。

参考文献

- [1] 加藤恒昭, 松下光範, 神門典子, 動向情報の要約と可視化に関するワークショップ ホームページ, <http://must.c.u-tokyo.ac.jp>
- [2] 加藤恒昭, 松下光範, 平尾努, 動向情報の要約と可視化に関するワークショップの提案, 情報処理学会自然言語処理研究会, 2004-NL-164 (15), pp.89-94, 2004.
- [3] kizashi.jp, <http://kizashi.jp>
- [4] blogWatcher, <http://blogwatcher.pi.titech.ac.jp>
- [5] 加藤恒昭, 松下光範, 平尾努, 神門典子, 評価なきワークショップの試み — 「MuST: 動向情報の要約と可視化に関するワークショップ」を例に —, 言語処理学会全国大会併設ワークショップ「評価型ワークショップを考える」, 2005.
- [6] 松下光範, 加藤恒昭, 動向情報に基づく情報可視化の基礎検討, 人工知能学会第 19 回全国大会, 2005.
- [7] 竹内 哲 (編集委員代表), 統計学事典, 東洋経済新報社, pp.334-346.
- [8] 難波英嗣, 国政美伸, 福島志穂, 相沢輝昭, 奥村学, 文書横断文間関係を考慮した動向情報の抽出と可視化 情報処理学会自然言語処理研究会, NL-168, pp.67-74, 2005.
- [9] 難波英嗣, WWW 上のテキスト情報の知的統合, 『人工知能学会誌』, 19 巻 3 号, 2004.
- [10] 難波英嗣, 複数テキスト情報の可視化: 研究事例の紹介, 電子情報通信学会 Web インテリジェンスとインタラクション研究会, WI2-2005-28 ~ 49, pp.109-115, 2005.
- [11] 山本健一, 殿井加代子, 谷岡広樹, タグ付きコーパスを用いた動向情報とその要因の可視化, 言語処理学会第 12 回年次大会ワークショップ, 「言語処理と情報可視化の接点」, 2006.
- [12] 山本健一, 谷岡広樹, 殿井加代子, 動向情報に対する変化要因の抽出手法, 第 6 回 Web インテリジェンスとインタラクション研究会, 2006.
- [13] Daniel Gruhl, R. Guha, Ravi Kumar, Jasmine Novak and Andrew Tomkins, The predictive power of online chatter, Proceeding of the eleventh ACM SIGKDD, pp.78-87, ACM Press, New York, NY, USA, 2005.
- [14] 藤木稔明, 南野朋之, 鈴木泰裕, 奥村学, document stream における burst の発見, 情報処理学会研究報告, 2004-NL-160, pp.85-92.