効率的な情報アクセスに向けた時系列情報の獲得手法

山本健一　谷岡広樹　殿井加代子

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

{kenichi_yamamot,hiroki_tanioka,kayoko_tonoi}@justsystem.co.jp

概 要

近年，電子化された情報の増加と検索技術の発展により，比較的単純な情報であれば検索技術を用いて容易にアクセスすることが可能となってきた．しかし，例えばある会社の株価の変動と同期している株価の変動をもつ会社を知りたい場合や，ある製品の売上の増加と共に使用されるようになった単語を知りたい場合，さらにはある単語の流行に合わせて使用される単語を知りたい場合など，時系列情報を伴った検索要求に応えることは未だ難しい．そこで我々は，株価やある製品の売り上げ傾向，単語の出現傾向など多様な時系列情報間の関連度を計算することにより，効率的に時系列情報を獲得する手法を提案する．

Keywords: 時系列情報，相関係数，ピアソンの積率相関係数，関連度，関連語

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 [4]．このような要求に答える技術のひとつとして，我々は，時系列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 [9]．

しかし，時系列情報を用いて様々な分析を行う際には，単一の時系列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時系列情報を同時に分析する必要がある．そして，内閣支持率の時系列グラフと日経平均株価の時系列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，多種多様な時系列情報を 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，関連し合う時系列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となる．

本稿では，我々が研究開発中のシステムの内，特に任意の時系列情報と関連する時系列情報の獲得方法に焦点を絞って説明する．

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，時系列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 節で関連する時系列情報の抽出実験に関して述べ，最後に第 5 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時系列情報を可視化する．
2. 可視化している時系列情報をキーにして，相関のある時系列情報を検索，可視化する．
3. 可視化している時系列情報を選択して，2 に戻る．

ここでは，時系列情報は，大きく以下の 2 種類に分類されるものとする．

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

頻度時系列情報： 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時系列情報．

本稿においては，目的とするシステムで必要となる技術のうち，特に任意の時系列情報と関連する時系列情報の獲得方法に焦点を絞って説明する．

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時系列情報を必要とする．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[11] 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，時系列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時系列情報だけでなく，内閣支持率や日経平均株価などの多様な統計時系列情報が表示できる点でも異なる．

Gruhl ら [12] 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．

我々はあらゆる時系列情報間の相関を算出し相関のある時系列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[10] 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，バースト，評判情報，男女推定，バーサス，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バ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．バースト度は，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 [13]．

一方，我々のシステムではバ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時系列情報や統計時系列情報を取得できる．

3 時系列情報間の関連度

一般に，データ列 $X = \{x_i\}$ と $Y = \{y_i\}$（$i = 1, 2, \cdots, n$）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 [5]．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，

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

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \quad (3)$$

とする．

今後，我々は時系列情報 X と Y との間の相関には式 (1) を用いることとする．

4 関連する時系列情報の検索実験

4.1 実験条件

ここでは，任意の時系列情報と関連する時系列情報の検索実験に関して報告する．

統計時系列情報は，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．

頻度時系列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

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
０．３３５	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
ｊａｓ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

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，統計時系列情報，及び頻度時系列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時系列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時系列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時系列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．

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．

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．図 1 に「オリンピック」と相関の高い頻度時系列情報のグラフを，図 2 に「PAD」と相関の高い頻度時系列情報のグラフをそれぞれ示す．これらの図より,「オリンピック」と相関の高い語は,「PAD」と相関の高い語と比較して頻度時系列情報の分散が大きそうなことが分かる．表 1，及び表 2 より，実際に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時系列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．

5 おわりに

本稿では，多様な時系列情報間の関連度を計算することにより，効率的に時系列情報を獲得する手法を提案した．その結果，任意の時系列情報と関連する時系列情報を効率的に取得できる可能性を示すことができた．

今後の課題として，以下の項目が考えられる．

- 統計時系列情報と相関のある時系列情報の検索実験
- 分散によるノイズ除去手法の開発
- 提案手法の評価方法の検討と評価
- ユーザインタフェースの開発

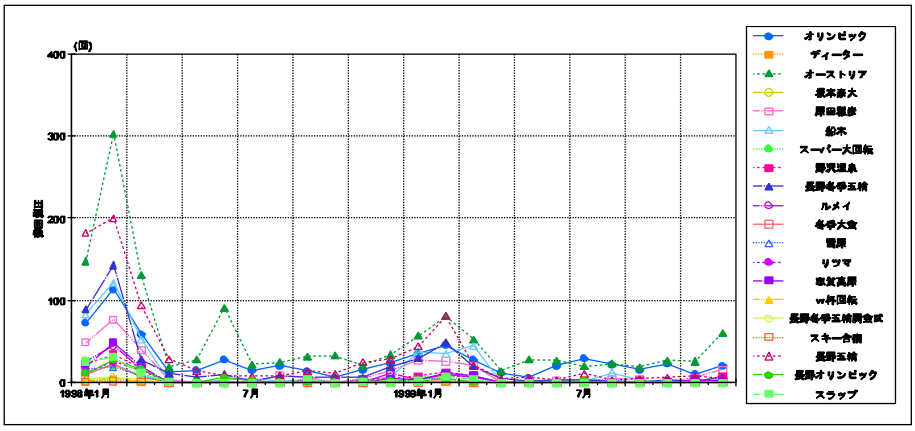

図 1: 「オリンピック」と相関の高いグラフ

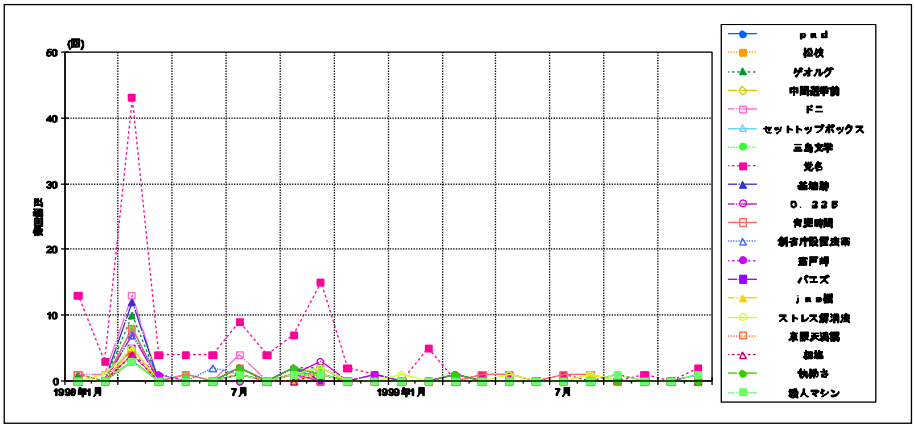

図 2: 「PAD」と相関の高いグラフ

謝辞

毎日新聞コーパスを提供して下さった MuST オーガナイザーに感謝します.

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] 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.

[4] 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.

[5] 竹内 哲 (編集委員代表), 統計学事典, 東洋経済新報社, pp.334-346.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] blogWatcher, http://blogwatcher.pi.titech.ac.jp

[11] kizashi.jp, http://kizasi.jp

[12] 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.

[13] 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.