

テキストからの単語重要度変動パターンの発掘

鈴木裕一郎[†] 森本 康彦[†]

[†] 広島大学大学院工学研究科 〒739-8511 広島県東広島市鏡山一丁目7番1号

E-mail: [†]{m082352,morimo}@hiroshima-u.ac.jp

あらまし 我々はスーパーなどで購入される商品の時系列売り上げ変動のパターンに対して相関ルールを調べる研究を行ってきた。この技術をテキストマイニングに応用し、ある単語重要度の増減がまた別の単語重要度の増減に影響しているとした場合、この連動を一つの変動パターンとみなすことができ、本研究ではインターネット上の blog から得られる単語重要度の変動パターンを発掘した。単語重要度は文書の特徴づけを行う際に有用であり、それを用いることで文書中のトピック等を求めることができる。しかし、単語重要度は季語のように時系列順において増減変動が起こる場合があり、トピックに選ばれなかった重要度の低い単語にも価値があると考えた場合、時系列順において単語重要度がどのように増減変動したかについて求めることで、より多くの情報を用いてトピック等の評価を行うことができる。本論文では、任意に定めた閾値以上となる単語重要度の変動パターンを求めることを目的とし、手順として blog から得た時系列情報を有するテキストデータから単語重要度を求め、その増減変動パターンの発掘を行い、さらに推移的変動パターンを求めることで、変動パターンの最適化を行った。また、一般的な動的計画法を用いて推移的変動パターンを求めた場合、多量の記憶領域が必要であるため、効率的な記憶を行う CE-ハッシュ木を用いて推移的変動パターンを求めた。

キーワード blog 分析, 相関ルール, 単語重要度

Fluctuation pattern mining of keywords in web text

Yuichiro SUZUKI[†] and Yasuhiko MORIMOTO[†]

[†] Graduate School of Engineering, Hiroshima University Kagamiyama 1-7-1, Higashi-Hiroshima, Hiroshima, 739-8521, JAPAN

E-mail: [†]{m082352,morimo}@hiroshima-u.ac.jp

1. はじめに

近年、web2.0 の普及により誰でも手軽にインターネット上で様々なコンテンツを利用したり、また、自身が情報源として世界に情報を与えるといったことが非常に容易に行うことが可能となった。特に、後者の代表例である blog の発展・普及により以前とは比較にならない量のテキスト情報が現在 Web 上に存在している。総務省情報通信政策研究所が行った「ブログの実態に関する調査研究」の調査結果によると、2008 年 1 月現在における国内の blog は約 1690 万存在しており、記事総数は約 17 億 9000 万にのぼる。

blog とは、個人の日記や批評を手軽に書きこみ、更新することができる web サイトのことであり、書き込まれる内容は主として、ニュースや TV 番組、ある商品のレビューといった個人または特定集団による批評・分析を行った文書であるものが多い。また、個人で blog を運営している場合には、そのような

批評を行う文章以外にも日々の出来事の感想といった日記的な記録を同時に残している blog が多数存在している。

これら大量に存在するテキスト情報は消費者の生の声として非常に価値があり、現在売られている商品や放映されている TV 番組といったものの評判分析を行うマーケティング調査やトレンド分析にしばしば利用されている。

本論文では、blog 等のテキスト情報に出現する単語の重要度が時系列において増減していることに注目し、単語重要度の増減変動パターンを発掘することを目的とする。これにより、ドラマに関する単語とスポーツに関する単語といった一見全く関係のない単語間において、ドラマに関する単語の重要度が増加したとき、スポーツに関する単語の重要度が減少したといった変動パターンを見つけることが可能となる。また、2.3.2 章で詳細を説明する推移的変動パターンについても扱うため、単純に見ると確信度が低い変動パターンだが、ある変動パターンを経由することで、確信度が高くなる変動パターンの算出を行う。

以下に本論文の構成を示す．第2章では，本論文で提案する単語重要度変動パターンの抽出方法について述べる．第3章では，blog 等におけるテキストマイニングに関する研究概要について述べる．第4章では，実際に存在する blog においてどのような単語重要度変動パターンが発掘されたかについて述べる．最後に第5章でまとめを行う．

2. 単語重要度変動パターンの発掘アルゴリズム

2.1 単語重要度の変動

blog 等における web 上テキストにおいて，ある単語の文書中における重要度を測る指標として tf-idf がある．ある文書における単語出現頻度を tf (time frequency) と呼び，これはページ内で単語がどれだけ多く使用されているかを示す．単語が文書に出現する割合を idf (inverse document frequency) と呼び，これはどれだけの数のページでその単語が使用されているかを示す．この2つの積によって単語重要度が算出される．単語重要度では，ある文書内での出現頻度が高い単語のうち，他の文書での出現数が少ない単語を，その文書における特徴的な単語として評価を行う．

tf-idf は次のように定義される．

$$w(T, D) = tf(T, D) \cdot idf(T)$$

$$tf(T, D) = tfreq(T, D)$$

$$idf(T) = \log\left(\frac{M}{dfreq(T)}\right)$$

$w(T, D)$: 文書 D における重要度

$tfreq(T, D)$: 文書 D における単語 T の出現頻度

$dfreq(T)$: 単語 T を含む文書数 (M を全文書数)

時間の変化により単語の重要度も変化していることを単語重要度の増減変動と呼び，テキストデータが持つ時系列情報が $t_0, \dots, t_i, \dots, t_n$ と存在したとき，時間 t_{i-1} から時間 t_i となったとき，両時間にある単語 A が存在し，時間 t_i 中の単語 A の重要度が時間 t_{i-1} に対して増加したならば，単語 A の重要度が増加したとする．逆も同様．

2.2 変動パターンの相関ルール

相関ルールはスーパーマーケットにおける商品の売り上げに関するルールを求める際によく使われており，例えば，コーヒーの売り上げが上がった翌週にマヨネーズの売り上げが上がるといった相関ルールの発掘を行う．本論文では単語重要度の変動に関する相関ルールの発掘を行うため，ここで相関ルールについて定義をする．

全文書集合を D としたとき D 中に存在する形態素解析によって得られた単語の集合 (*itemset*) を $T = \{w_1, w_2, \dots, w_n\}$ とする．文書集合 D_t は時間 t における文書の集合であり，各時間 t における文書集合は単語集合 T 中の部分集合 X から成り立つ．相関ルールとはある時間 t_i における文書集合 D_{t_i} が $X \subset T$ となる単語集合 X_a と， $Y \subset T$ となる単語集合 Y_a を持つ

ているとき，全時間 t_0 から t_n において単語集合 X_a からの任意の単語 x 存在したとき，単語集合 Y_a からの任意の単語集合 y が高い確度で存在したならば，これを $x \rightarrow y$ という相関ルールと呼ぶ．

相関ルールは，確信度 (confidence) と 支持度 (support) という2つの指標を用いて評価される．変動パターンの相関ルールにおける確信度は，ある時間 t から時間 $t+1$ において，左辺事象として単語 X の重要度の増減が起こり，右辺事象として単語 Y の重要度の増減変動が起きたとき，単語 Y の重要度の増減が発生する条件付き確率を表し， $\text{Conf}(X \rightarrow Y) = C\%$ と表記する．支持度は，上の左辺事象が起こった後に右辺事象が起こる確率はどれだけかを表し， $\text{Sup}(X \rightarrow Y) = s\%$ と表記する．確信度，支持度はそれぞれ，相関ルールの強さ，相関ルールの発生する確率を表していると言える．

2.3 時系列変動パターン

2.3.1 時系列変動パターンの算出

ある時間 t 中の文書集合 D_t において D_{t-1} から単語重要度が増加したか，減少したかを判定するために閾値を用いる．例えば，閾値を 0.3 と設定したとき，単語重要度が 0.3 以上増加したとき単語重要度が増加した $\{A^+\}$ とする．逆に， D_t において重要度が 0.3 以下減少したとき単語重要度が減少した $\{A^-\}$ とする．単語重要度の増減が閾値未満の場合，増減なしとする．

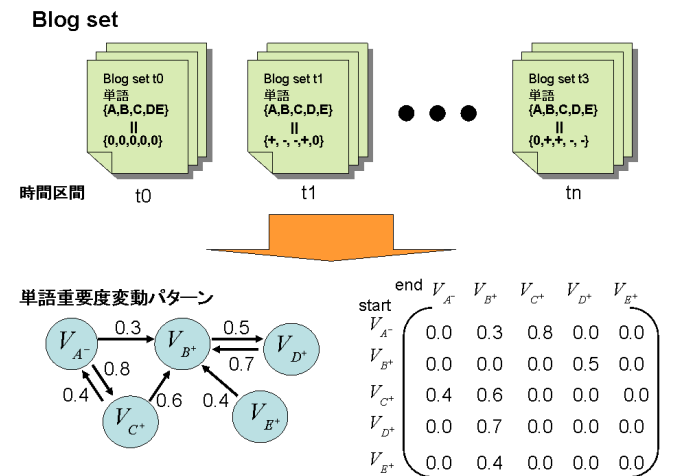


図1 単語重要度変動パターン

2.3.2 推移的時系列変動パターンの算出

変動パターンは { 左辺事象 右辺事象 } と表現されるため，各事象をノード，変動パターンの確信度を重みとする重み付き有効グラフを用いて表現することができる．

図1において各ノードは事象を表しており，この場合単語 A の重要度が下がったことを $\{V_{A^-}\}$ と表す．有向辺は変動パターンを表しており，その重みは変動パターンの確信度である．図1において，ノード $\{V_{A^-}\}$ からノード $\{V_{B^+}\}$ へと向かう有向辺は，変動パターン $\{V_{A^-} \rightarrow V_{B^+}\}$ を表し，その有向辺の重みである値 0.3 は変動パターンの確信度を示している．有向グラフを用いて表現することにより，変動パターンの積にお

いて、推移的時系列における変動パターンの確信度を算出することができる。

変動パターンは連続した時系列順の事象のみしか扱っていないため、単純な結果しか得ることしかできない。そこで、最適変動パターンを求めるため本論文では、推移的な変動パターンについても算出を行う。これにより、単純に連続事象を見ただけでは見つけることのできない変動パターンについても求めることができ、その評価は上で述べた確信度・支持度を用いて行う。ここで述べている推移的変動パターンとは、変動パターンである{左辺事象 右辺事象}が存在しているとき、左辺と右辺以外の事象を経由してよい変動パターンである。

図1の{ V_{A-} V_{B+} }の確信度は0.3だが{ V_{A-} }が{ V_{C+} }を経由した場合、{ V_{B+} }への確信度は次のようになる。 $\text{conf}(\{V_{A-} \quad V_{C+}\}) = 0.8$ 、 $\text{conf}(\{V_{C+} \quad V_{B+}\}) = 0.6$ から、 $\text{conf}(\{V_{A-} \quad V_{C+} \quad V_{B+}\}) = 0.48$ となり、単純に{ V_{A-} V_{B+} }となる変動パターンよりも{ V_{C+} }を経由した方が確信度が高いことがわかる。この手法は直接繋がった経路が存在しない変動パターンにおいても有用である。図1では{ V_{A-} V_{D+} }という変動パターンは存在しないが{ V_{A-} V_{B+} }、{ V_{B+} V_{D+} }を考慮すると{ V_{A-} V_{B+} V_{D+} }という{ V_{B+} }を経由した変動パターンが存在することがわかる。これにより、多様な変動パターンの発掘が可能となる。

上のように、ある事象を経由した変動パターンを求める際、求める変動パターンの事象間に複数の経路が存在していることがある(図1では{ V_{A-} }から{ V_{D+} }). 其中で、最も確信度の高い経路の変動パターンを最適時系列変動パターンと呼び、これを求めるには確信度を重みとする重み付き有向グラフにおける最短経路問題を解かなくてはならない。

確信度を重みとした重み付き有向グラフにおいて、この問題は全ての頂点のペアに対して最短経路を求める問題とみなすことができ、動的計画法を応用した全頂点間の最短経路探索アルゴリズムを用いて効率的に解くことだできる。しかし、この手法では計算過程において全ての頂点間の最短経路の確信度を記憶する必要があり、大量の記憶領域を使用する。

全頂点数 n からなる重み付き有向グラフの全頂点間の最短経路は以下のようにして求める。全頂点に1から n まで番号を与え、全頂点間のどの頂点も経由しない始点から終点への直接の有向辺を0時点での最短経路とする。1番から k 番までの頂点のみを経由する最短経路を k 時点での最短経路とする。このとき図2のように、 $k+1$ 時点で既に存在する k 時点での経路{ V_i V_j }間に新しい頂点{ V_p }が追加されたとき{ V_i }が{ V_p }を経過しない場合と{ V_i }が{ V_p }を経過する場合の、2つ場合がある。前者の場合は、確信度の変更を行う必要がないが、後者の場合は{ V_i V_p V_j }経路の確信度と{ V_i V_j }経路の確信度を比較し、値が高い方を最適経路の確信度として記憶する必要があるが、これを図3で示す構造を持つCE-ハッシュ木を用いて記憶する。

図3では、図1にて有向グラフを用いて表した単語重要度変動パターンをCE-ハッシュ木を用いて最短経路の記憶を行った様子を示している。CE-ハッシュ木では必要な最短経路のみ記

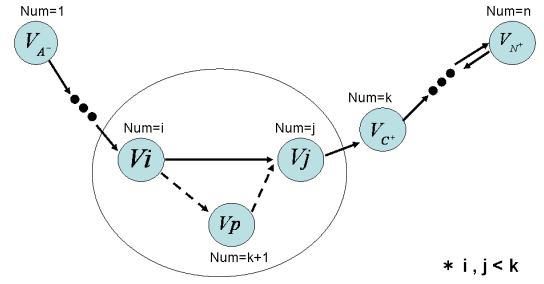


図2 最短経路の判定

憶を行っており、変動パターンの最小確信度を設定した場合は、それを満たす経路のみ記憶を行うので、全ての経路を記憶する単純な記憶方式よりも効率的な記憶を行うことができる。

例えば、図3において{ V_{A-} }から{ V_{B+} }へ向かう有向辺を持つ変動パターン{ V_{A-} V_{B+} }では{ V_{B+} }から破線として最短経路を通過するノードの記憶を行っているため{ V_{C+} }を経由した{ V_{A-} V_{C+} V_{B+} }が最短経路であることを示している。

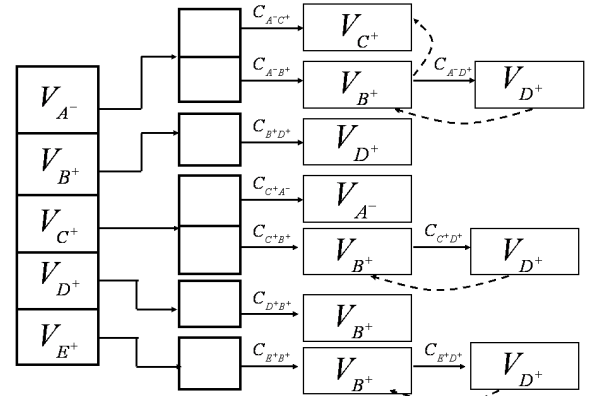


図3 CE ハッシュ木

3. 関連研究

blogを対象とした研究内容は、主に特徴抽出、評判判定、生起時間による評価といったものが多数を占める。

戸田ら[4]は、blog記事群から多視点からのトピック抽出として品詞情報に注目し、動詞・形容詞により特徴語に重み付けを行い、行動中心、感想中心といった同一トピックに対して異なる視点による評価を与えている。倉島ら[5]は、批評評価文から対象間の優劣関係を有効グラフで表し、より良い商品を探る移動する顧客の行動をモデリングしており、ノードを対象、有効辺で対象間の優劣関係を表現し、重み付けとして評価・優劣の遷移確率を与えている。福原ら[6]は、blogや新聞記事、メールマガジン等の時系列テキスト集合から社会的関心解析システムとして、共起語や感情表現、実世界データとの関連といった視点で分析し、時間軸上で社会的関心がどのように変化していくかを調査している。Tianmingら[7]は、2部グラフを用いてクラスタリングを行い、よりの確かなトピックを単語・文

書における相関ルールに割り当てる．Tianming らの手法ではノードに単語と文書におけるある閾値以上の確信度を持つ変動パターンを持っている．単語と文書の変動パターン集合をそれぞれ 2 部グラフの両端に置き、そのつながりを求め、適切なトピックを割り当てる．

本論文では、blog といったテキスト中の単語重要度が時間軸上において増減変動が起こることに注目し、これの変動パターンを発掘することを目的としているため、文書中における単語の重要度を評価対象とする．また、倉島らのように有効グラフを用いてモデル化を行うが、各ノードには単語重要度の増減状態を設置し、推移的変動パターンの発掘を行うため、より確信度の高い経路を求めるアルゴリズムについて提示をした．

4. 実験

4.1 実験準備

今回、単語重要度増の減変動からルールが得られるかを調べる実験として忍者ブログ [2] から平成 21 年 2 月 14 日におけるブログ集合から推移的変動パターンを求めた．また、過去 blog データから変動パターンを求める実験として、google ブログ検索 [3] より得られた平成 20 年 8 月から平成 21 年 1 月までの "サッカー" に関するブログ集合から推移的変動パターンを求めた．なお、単語重要度を求める際は形態素解析ツール chasen [1] を用いてテキストから単語集合を求めた．実験の手順として、以下の順に行う．

1) ブログ検索によりブログの記事 URL を取得、2) 取得したブログ URL からブログテキストの抽出、3) テキストを形態素解析ツール chasen を用いて、形態素解析を行う、4) 形態素解析により得られた形態素集合をデータベースに挿入、5) データベースから tf-idf 値を算出し、それを用いて増減変動、変動パターンを求める、6) 5) より得られた変動パターンから推移的変動パターンの算出

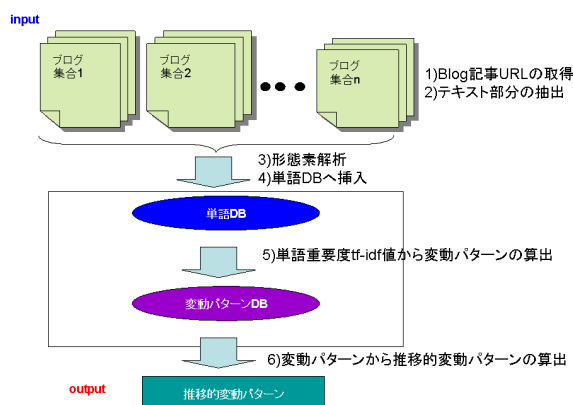


図 4 推移的変動パターン発掘フロー

4.2 実験結果

以下が 2 月 14 日における blog 集合の実験結果である．解析 blog 数は 700、解析対象の時間区間を 1 時間ごととし、単語重要度の増減判定の閾値を 0.2、そして推移的確信度を求める際

の閾値を 0.3 とした．ここでは変動パターンが得られるかを調べているため、増減変動を判定する閾値を低く設定している．時間別において得られた単語重要度 (表 1) と、その変動パターン (表 2) について示す．

word (20 時)	tf-idf	word (19 時)	tf-idf	word (18 時)	tf-idf
兄弟	1.900	チョコ	3.691	今日	2.808
今日	1.780	金	3.355	チョコ	1.962
や	1.723	合唱	1.678	仕事	1.549
日	1.711	今日	1.501	バレンタイン	1.527
無料	1.710	スキ	1.468	恋	1.257
バレンタイン	1.705	バレンタイン	1.445	設定	1.257
ホーン	1.585	母	1.372	生活	1.257
ドラゴン	1.585	思い	1.300	薬	1.234
何	1.577	好き	1.280	ホテル	1.234
キャラ	1.559	女の子	1.067	友達	1.120
万	1.535	妨害	1.048	日記	1.120
方法	1.520	悩み	1.048	写真	1.120
ミス	1.520	音楽	0.915	気分	1.120
的中	1.359	母親	0.762	引っ越し	0.987

表 1 時間別単語重要度

$word_s$	$word_e$	sup	conf
写真-	苦手-	0.4	0.200
金+	もの+	0.3	0.333
歌詞-	朝-	0.3	0.333
報告+	女の子-	0.2	0.500
誕生+	限定+	0.2	0.500
チョコ+	是非+	0.2	0.333
嘘-	可能+	0.1	1.000
福岡+	ラブラブ+	0.1	1.000
ラブラブ+	小説-	0.1	1.000
異常+	送信+	0.1	1.000
編+	ドキドキ+	0.1	1.000
幸せ+	買い物-	0.1	1.000
映画+	お待ち-	0.1	1.000
幸せ-	医者+	0.1	0.500

表 2 単語重要度変動パターン

表 1 では 18 時から 20 時までの各時間における単語重要度を示している．この表からチョコ、バレンタイン、恋といった単語の重要度が高くなっていることがわかる．このように各時間における単語全てにおいて単語重要度を算出し、その増減変動を求め、変動パターンを求める．表 2・表 3 では、閾値以上の変動パターンとそれから得られた推移的変動パターンを示しており、 $word_s$ は変動パターンの始まり、 $word_e$ は変動パターンの終わり、 $conf$ はその変動パターンの確信度、 $root$ ではその変動パターンが通る経路を示す．

表 2 では表 1 から得られた各時間における単語重要度から得られた単語重要度の増減に対する変動パターンを示しており、" { 誕生 (+) 限定 (+) } " といった変動パターンを得ることができる．

以下に、CE ハッシュ木を用いて得られた確信度の高い変動

$word_s$	$word_e$	$conf$	$root$
ああ +	可能-	0.5	
ああ +	好き +	0.333	> 意外 +
ああ +	意外 +	0.666	
ええ +	多い-	0.5	
ええ +	好き +	0.5	
ええ +	意外 +	0.375	> 好き +
ええ +	楽しい +	0.75	
おそろしい-	楽しい +	0.5	
おそろしい-	馬鹿-	0.5	
便利-	ああ +	0.375	> 可能-
便利-	可能-	0.5	0
便利-	意外 +	0.375	> 可能-
便利-	馬鹿-	0.5	
可能-	ああ +	0.75	
可能-	便利-	0.5	

表 3 変動パターンと推移的変動パターン

パターンと推移的変動パターンを示す。

表 3 より推移的増減変動パターンとして次の結果が得られる。今回の場合 { 便利 (-) 意外 (+) } という変動パターンは実際に存在していないが { 可能 (-) } を経由することで { 便利 (-) 可能 (-) 意外 (+) } 変動パターンが発掘されており、直接の経路が存在しない変動パターンを求め、また、ある変動パターンを経由することにおいて確信度が高くなる推移的変動パターンを発掘している。

次に "サッカー" に関する blog 集合から得られた特徴的な変動パターン、及び推移的変動パターンについて示す。解析 blog 数は 2400、解析対象の時間区間を 1ヶ月ごととし、増減判定の閾値を 0.4 とした。先ほどの実験と同様に単語重要度とその増減変動を求め、変動パターンと推移的変動パターンを算出し、以下にその結果を示す。

word	tf-idf (1 月)	tf-idf (12 月)	tf-idf (11 月)
シュート	1.038	0.283	0.422
パーレーン	4.646	0.436	0.340
ジャパン	1.412	0.283	0.220
五輪	0.314	0.850	0.440
Jリーグ	1.255	0.850	0.422
世界	0.627	1.367	0.661
オーストラリア	0.784	0.283	0.440

表 4 時間別単語重要度 (keyword:サッカー)

表 4 では、単語の重要度が時系列においてどのように遷移しているかを示しており、単語 "パーレーン" は 1 月に急激に増加している。表 5 では、単語重要度の増減変動より得られた変動パターンを示す。表 6 では、閾値以上の変動パターンとそれから得られた推移的変動パターンを示す。スポーツと関連性の低い単語も変動パターンとして発掘されており、サッカーに関する話題と全く関連性の見られない単語に関する推移的変動パターンの発掘が確認できる。

今後の課題として、今回の実験では解析対象 blog が限定的期間であるため、より長い期間で実験を行うべきである。今後、

$word_s$	$word_e$	sup	$conf$
セット-	心配+	0.100	1.000
ジャパン+	世界-	0.100	1.000
カタール-	柏+	0.100	1.000
柏+	目-	0.100	1.000
タイトル+	大宮-	0.100	0.500
海外-	千葉+	0.100	0.333
相手+	甲府-	0.100	0.250
世界+	ガンバ+	0.100	0.250
リーグ-	代表-	0.400	0.200
クラブ+	サッカー+	0.400	0.200
大会-	現状+	0.100	0.167

表 5 単語重要度変動パターン (keyword:サッカー)

$word_s$	$word_e$	$conf$	$root$
五輪+	日本人+	0.5	
五輪+	状況+	0.5	> 日本人+
ストライカー+	バック-	1	> 登場+
ストライカー+	登場+	1	
インタビュー+	ユース-	0.5	> 7+
アルゼンチン+	埼玉+	1	> 会-> キックオフ+
大宮+	パス-	1	
チャンピオンズ+	復帰-	0.5	> 最終-> 話-
テレビ+	パス-	1	
監督+	事業+	0.25	

表 6 変動パターンと推移的変動パターン (keyword:サッカー)

長期間にわたるリアルタイムの blog の確保や、既に存在している大量過去 blog データを用いて実験を行い、さらに、時間区間を様々な期間を設定し、どの期間で行うとどのような特徴が出るかについて調べる必要がある。

5. ま と め

本論文では、文書中に存在する単語の単語重要度を求め、その増減に関する変動パターンの発掘を行い、さらに推移的変動パターンを効率的に記憶する手法を用いた。単語重要度の増減変動に注目することにより、世間の注目度に即した形で変動パターンを発見でき、また、スポーツとドラマのような全く関係のない単語間における相関を発見することができるため、隠れたニーズや関係性の把握を行うことができる。さらに、推移的な変動パターンを用いることで、連続的時系列上全く無関係に見えるがある事象を経由することで確信度が高くなる事象の変動パターンを見つけだすことが行え、より詳しく事象の関連性を知ることができる。また、本論文の提案手法は blog だけでなく様々なテキスト情報において応用可能であるため、アンケートから得られた満足商品の推移や、メールの単語重要度変動パターン等においても利用でき、blog と同様に隠れたニーズや重要情報を得ることができると考えられる。

今回は blog 集合全体から解析を行ったが、tf-idf 等を用いた blog の特徴付けにより文書集合のクラスタリングを行った上で本提案手法を実行することで、似通った blog を書く人の単語重要度変動パターンを求めることができるため、より精度の高

い結果が得られることが期待できる。

謝辞

この研究は科研費(課題番号 19500123)の助成を受けて行った。

文 献

- [1] ChaSen, <http://chasen-legacy.sourceforge.jp/>
- [2] 忍者ブログ, <http://blog.ninja.co.jp/>
- [3] Google ブログ, <http://blog.fc2.com/>
- [4] 戸田 智子, 黒田 晋矢, 福田 直樹, 石川 博:”ブログにおける多視点からのトピック抽出手法の提案”, DEWS2008
- [5] 倉島 健, 別所 克人, 内山 俊郎, 片岡 良治:”比較評価情報の抽出とそれに基づくランキング手法の提案”, DEWS2007
- [6] 福原 知宏, 中川 裕志, 西田 豊明:”時系列テキスト集合からの社会的関心の分析”, 第 16 回インテリジェント・システム・シンポジウム・2006 年
- [7] Tianming Hu, Hui Xiong, Sam Yuan Sung:”Co-Preserving Patterns in Bipartite Partitioning for Topic Identification”, SIAM2007
- [8] Y. Morimoto:”Optimized Transitive Association Rule: Mining Significant Stopover between Events”, Proc. of the ACM SAC, oo.547-548, 2005.
- [9] 森本 康彦, 八木 一光:”POS データからの売り上げ変動パターン発掘”
- [10] 福田 剛志, 森本 康彦, 徳山 豪:データマイニング. 共立出版, 2001