社団法人　情報処理学会　研究報告
IPSJ SIG Technical Report

2005－NL－165 (5)
2005/1/11

文書の類似度と新鮮度に基づく話題語抽出

佐藤 吉秀　　川島 晴美　　佐々木 努　　大久保 雅且
NTT サイバーソリューション研究所

大量文書中において，人々の持つ関心が増大することによって関連文書数が増加する「情報の広がり」と，そのような注目状態が時間的に継続する「情報の伸び」の 2 つの側面に注目した最新話題語抽出手法を提案する．文書クラスタリングを用いて同一話題を集約した後，類似文書の量によって各文書の持つ話題性の大きさを判定する．続いてその結果を用い，文書中に含まれる語句に対する話題性の大きさを評価する．Web 上に掲載されたニュース記事を用いて実験を行った結果，記事集合中に含まれる異なる複数の話題性の高い情報について，各話題を連想させる語句を抽出することができた．

Topic Words Detection by both Documents' Freshness and Similarity

Yoshihide SATO, Harumi KAWASHIMA, Tsutomu SASAKI, Masaaki OHKUBO
NTT Cyber Solutions Laboratories

"Topic" has two aspects. One is a status which many people are concerned about the information and the other is a continuation of it. The algorithm, we propose in this paper, detects topic keywords in a large collection of documents by taking them into account. The procedure consists of two steps. At first, it combines several documents on the same subject by clustering and evaluates significance of each document. Next, it extracts a few topic keywords from each document cluster by using the significance. After the experiment using news articles, we could confirm the effectiveness of the method.

1　はじめに

人々の情報収集スキル向上や情報取得環境の改善により，我々が大量の文書に接する機会は以前に比べて格段に増加している．時々刻々と変化する情報世界の最先端に身を置き，常に最新の話題に追随していたいという最新情報取得欲求もこれに伴って大きくなっている．ニュース記事のようなリアルタイム性の高い文書の閲覧に適した方法として注目される RSS(RDF Site Summary) も，最新情報把握のための 1 手段である．RSS は，最近では多くの新聞社サイトやポータルサイト等が配信しており，RSS リーダやティッカーなどの RSS 閲覧ソフトを用いて情報を監視することができる．このような閲覧方法は文書単位での情報提示になるため，ユーザは提示された無数のタイトル等を概観して情報を取捨する必要があり，文書数が増加した場合に大きな労力を要する．この状況の中，大量文書の概要把握や話題の発見支援に対するニーズは大きい．

テキストマイニングの分野において，大量文書から話題情報を獲得し提示する手法は種々提案されているが，話題情報獲得の難しさは含まれる話題の規模の差異にある．通常，文書集合中には万人が認める圧倒的に大規模な話題から，少数派にしか認識されないが着実に盛り上がりを見せる話題まで，大小様々な規模の話題が存在する．話題獲得の焦点を大規模な話題に当てると小規模な話題が埋没して陰に隠れてしまい，小規模な話題に焦点を当てて獲得しようとすると誤差が増大して獲得精度が下がってしまう．

本研究では複数の話題を含む大量文書中から，最新の話題情報を可能な限り多く，端的に提示する手法の実現を目指す．それにあたり，情報量が膨大な場合のユーザの概要把握コストを削減するため，話題を “語句” で代表させて提示するというアプローチをとる．語句での提示はタイトル等の短文による提示に比べ，量が膨大な場合の閲覧コストが少ない．

以下，第 2 章で本研究の関連研究について，第 3 章では本研究で抽出する “話題” の定義を述べる．第 4 章で提案手法を説明し，第 5 章でニュース記事を対象に行った実験と結果を示し，第 6 章で考察する．最後に第 7 章でまとめを述べる．

2 関連研究

大量文書からの話題キーワード抽出のためのマイニング手法は多数提案されている．

時系列文書データにおいて，話題を含む文書や話題性の高い固有名詞が短期間に集中的に出現することを利用し，確率モデルを用いて異常出現を検出する手法 [1][2] がある．これらは “語” を単位として扱う手法であるが “話題” を 1 つの単位として扱わないため，抽出したキーワードがいずれも規模の大きな特定の話題に関する語句で占められる可能性を排除しきれない．このことは文書集合から可能な限り多くの異なる話題を抽出・提示するという本研究の目的に合致しない．また，これらはいずれもある期間における話題語を抽出する手法であり，常に最新の話題語を提示するために時間的な出現傾向にも注目する本研究とは，目的およびアプローチが異なる．

“Scatter/Gather”[3] は明確な検索意図を持たない場合の文書検索支援方法である．これは，文書集合をクラスタに分割し各クラスタの代表語を提示する Scatter ステップと，ユーザに選択させた 1 以上のクラスタをマージし部分文書集合を生成する Gather ステップとを繰り返しながら対話的に文書の絞り込みを行うものである．この手法は，話題を “文書クラスタ” の単位で扱っており，複数の語句によって各クラスタの概要を提示する点で本手法の目指す方向性と非常に類似する．しかし研究の主眼は対話的ナビゲーションの効率向上にあり，各クラスタを代表させる語の抽出方法はクラスタ内の高頻度語を選択するという単純な手法に留まっている．

3 大量文書における “話題”

“話題” とは，情報が人から人へと拡散 (伝播) し，ある期間において多数の人々に注目されている状態を言う．すなわち，話題は情報がいかに人々に浸透しているかという社会的認知性と，認識された情報の注目状態がいかに継続するかという時間的継続性の 2 つの側面で捉えることができる．図 1 のように，社会的認知性を話題の「横方向への広がり」と呼ぶならば，時間的継続性は話題の「縦方向への伸び」と言える．ある一時期において，ある特定の題材を主題として書かれた文書が多ければその題材は注目度が高く，社会的認知性が大きい．また，社会的認知性の高い話題が，時刻の経過とともに持続あるいは発展する様子は時間的継続性の側面として説明できる．例えば，全世界の新聞社がイラク戦争関連のニュースを報道し，全世界のインターネットユーザがイラク戦争に関して個々の見解を発言を重ねる状況が長期に渡って継続する時,「イラク戦争」は社会的認知性・時間的継続性が共に極めて高い，すなわち「広がり」と「伸び」とを兼ね備えた話題だと言える．

本研究では，上述のように一定の「広がり」と「伸び」をあわせ持つ一連の文書群が指し示す題材を “話題” と捉え，これを話題語と呼ぶキーワードによって代表させる．

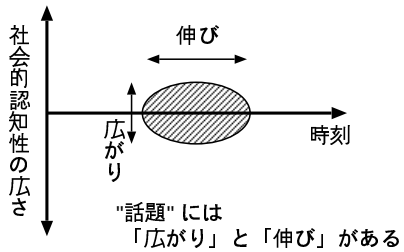

図 1: 話題の「広がり」と「伸び」

4 提案手法

本稿では，情報の「広がり」「伸び」という空間的・時間的両方向への拡散に着目した最新話題抽出手法を提案する．

前章ではある話題に注目した場合の「広がり」と「伸び」について述べたが，話題を構成する個々の

文書に注目しても同様の議論が展開できる．すなわち，ある文書が話題性の大きな情報を含む場合，同時期に記述された他の文書中に類似する文書が多く存在し，またその前後の時刻に記述された文書中にも類似，または関連する文書が多く存在する．よって，文書を話題の最小単位として扱うことで話題の存在およびその規模を的確に捉えることができる．

ところで，文書集合から可能な限り多くの異なる話題を表す語句を抽出するためには，単語の頻度などの統計情報のみを利用し，直接的に単語の話題性を評価する方法は好ましくない．なぜならば，この方法では，際立って大きな話題を成す文書に含まれる語の評価が全般的に高くなり，一部の大きな話題の影響を強く受けた抽出結果になってしまうためである．これを防ぐため，前処理として，文書集合を内容に応じて複数の部分集合に分類することで文書レベルでの同一話題の集約を行う．続いて各部分集合内の文書の話題性の大きさを評価し，その結果を用いて文書中から話題性の高い語句を抽出するという段階的な話題語抽出を行う．これにより，意味に基づいて分類された部分集合の各々を代表する話題語を抽出でき，文書集合中に含まれる複数の話題の提示が可能になる．

提案する話題語抽出処理の概要を図 2 に示す．処理は，文書の持つ話題性の大きさ (以後，文書話題度と呼ぶ) を算出する処理 (1〜5) と，その結果を用いて文中の語句の持つ話題性の大きさ (以後，語句話題度と呼ぶ) を算出する処理 (6) に大別される．前者の処理を 4.1 節で，後者を 4.2 節で説明する．

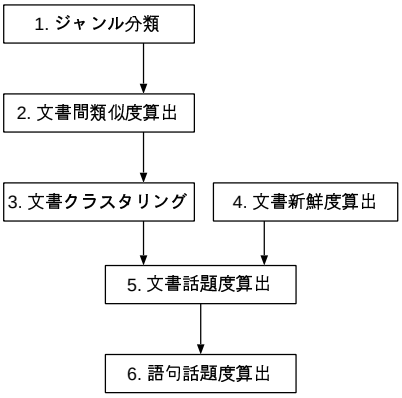

図 2: 話題語抽出処理の概要

4.1 文書の話題性評価

語句の話題性評価の前処理として，文書の話題性 (文書話題度) を評価する．図 2 において，最初に文書をジャンル分類 (1) し，ベクトル空間モデルに基づく文書間の類似度 (2) を利用した文書クラスタリング (3) を行う．この時点で各文書は内容に応じてジャンル・クラスタの 2 段階に分類され，話題の大きさに応じた文書クラスタが形成された状態となる．一方で文書の作成時刻から文書新鮮度を決定 (4) し，(2)〜(4) の結果を統合して文書話題度を算出 (5) する．これにより，注目度の高さおよび新鮮さという 2 つの観点で文書を評価することができる．以下，(1)〜(5) の個々の処理について述べる．

ジャンル分類

文書をその内容に応じて複数のジャンルに分類する．ニュース記事のようにあらかじめ「社会」「スポーツ」等に分類されている場合はその分類情報を用いてもよいが，文書の多義性を考慮すれば，重複を許したジャンル分類を行うのが望ましい．

ここでは，パラメトリック混合モデルに基づく非排他的な多重トピック分類手法 [4] を用いることとした．あらかじめ人手でジャンル分類した文書データで学習を行い，文中の単語出現頻度分布を用いて分類ジャンルを推定する．

文書間類似度

文書クラスタリング (3) および文書話題度算出 (5) の処理には，文書間の関連性の強度 (文書間類似度) が必要となる．そこで，同一ジャンル内に分類された文書の全ての組み合わせに対して文書間類似度を算出する．ここでは，ベクトル空間モデルにおけるコサイン類似度を文書間類似度とすることとした．

ベクトル空間モデルで文書をベクトル表現する際に広く用いられる単語の重み付け法には TF-IDF 法がある．しかし，インターネット上で収集される多くのニュース記事や blog の投稿記事のように比較的短い文書では，主題となりうる重要な単語が文章の先頭付近に 1 度だけ出現するような場合も多く，単語の出現頻度 (Term Frequency) と重要性との相関が必ずしも高くない．したがって，本手法では出現文書数 (Document Frequency) のみに基づい

て重みを決定する．文書数 N の文書集合中の文書 d_i の文書ベクトルは以下のように定義する．

$$\boldsymbol{d}_i = (x_{i,1},\ x_{i,2},\ \cdots,\ x_{i,V})\ \ (i = 1, 2, \cdots, N) \quad (1)$$

$$x_{i,v} = \begin{cases} \log\{M/df(w_v)\} & \text{if} \quad tf(d_i, w_v) \neq 0 \\ 0 & \text{else} \end{cases}$$

$$\text{ただし} \quad M = \max_{v=1,\cdots,V}\{df(w_v)\}$$

V は文書集合全体におけるユニーク単語数であり，$df(w_v)$ は全 N 文書中で単語 w_v が 1 度でも出現する文書の数，$tf(d_i, w_v)$ は文書 d_i における単語 w_v の出現頻度である．

これを用いて，ジャンル内の任意の文書 i，j 間の類似度 S_{ij} を式 (2) によって求める．

$$S_{ij} = \frac{\boldsymbol{d}_i \cdot \boldsymbol{d}_j}{|\boldsymbol{d}_i||\boldsymbol{d}_j|} \quad (2)$$

文書クラスタリング

同一の話題について触れた文書を集約するために文書クラスタリングを行う．文書間の非類似度 $(1 - S_{ij})$ を文書間距離とし，あらかじめクラスタ数が未知の場合にも適用可能な階層的クラスタリングの手法である最長距離法 [5] を用いてクラスタリングを行う．手順は次の通りである．任意の 1 文書を中心とするクラスタを生成 (1) し，既存クラスタの中心から最も遠い文書を探索 (2) する．一定の条件を満たせばその文書を中心とするクラスタを生成 (3) し，満たさなければ処理を終える (4)．

1. 文書 d_1 を中心とするクラスタを生成．
2. $l = \max_i \left\{\min_j d(d_i, \overline{Z_j})\right\}$ を求める．最大値を与える i を k と表す．
3. $\max_{ij} d(\overline{Z_i},\ \overline{Z_j}) < lr$ なら d_k を中心とするクラスタを作成．2 以降を繰り返す．
4. 3 の条件を満たさなくなった時点で終了．

ここで，$\overline{Z_i}$ はクラスタ Z_i の中心，$d(\overline{Z_i}, \overline{Z_j})$ はクラスタ Z_i，Z_j の中心間の距離，r は継続条件を決定するパラメータである．

文書新鮮度

ここでは文書の作成時刻に基づいて文書新鮮度を決定する．ある話題に関する新情報が出現すれば，それまでの情報は既知の情報となり重要性は低下する．今回は新しい文書を特に強調した最新話題語抽出を行うために，作成時刻が t_i である文書 d_i の新鮮度 $Fresh(i)$ を式 (3) のように指数関数で与える．t_0 は現在時刻，T は新鮮度の減衰速度を決定するパラメータである．

$$Fresh(i) = \exp\{(t_0 - t_i)/T\} \quad (3)$$

文書話題度

クラスタリングにより，同一の題材を扱った文書は同一クラスタに集約される．ここではクラスタ内の各文書がどの程度の話題性を持つか，すなわちクラスタを代表する話題文書としてどの程度相応しいかを評価するため，文書間類似度，文書の新鮮度，ならびにクラスタリングの結果を利用して文書話題度 (Document Topic) の算出を行う．

ある文書に注目したとき，同一クラスタ内に属する他の文書との類似性が低ければ，その文書は本来別のクラスタに属するべき文書であった可能性が高い．このような文書はクラスタを代表する文書とは呼べない．そこで表 1 に示すように，注目文書 d_i とクラスタ内の他の文書との類似度の平均値 $\sum_j S_{ij}/N_C$ を求める．N_C は注目文書が属するクラスタのサイズ (クラスタの構成文書数) である．クラスタを構成する文書群の重心に近い中心的な文書であるほどこの値は大きい．これに文書新鮮度 $Fresh(i)$ を乗じ，重要性が高く，かつ新鮮な文書を評価する指標である文書話題度 $DT(i)$ を得る．(式 (4))

$$DT(i) = Fresh(i) \cdot \sum_{j \in C,\ j \neq i} S_{ij} \cdot \frac{1}{N_C} \quad (4)$$

(C は d_i の所属クラスタ)

表 1: 文書間類似度からの重要度算出

	d_1	d_2	d_3	$\cdots$	$\sum/N_C$
d_1		S_{12}	S_{13}	$\cdots$	$\sum S_{1j}/N_C$
d_2	S_{21}		S_{23}	$\cdots$	$\sum S_{2j}/N_C$
$\vdots$		$\vdots$			$\vdots$
d_i	S_{i1}	S_{i2}	S_{i3}	$\cdots$	$\sum S_{ij}/N_C$
$\vdots$		$\vdots$			$\vdots$

4.2 語句の話題性評価

続いて文書中の語句に対する話題性評価に移るが，各クラスタを代表しうる語句はクラスタ内で文書話題度が高い文書中に含まれると考えるのが自然である．また，ジャンル分類，文書クラスタリングによって各話題はそれぞれ分離されているため，話題語は各ジャンル・各クラスタに特有の語だと言える．この点を考慮して式 (5) で語句話題度 WT(Word Topic) を算出する．

$$WT(w) = WA(w) \cdot ICF(g,w) \cdot IGF(w) \quad (5)$$

ただし

$$WA(w) = \sum_i DT(i) \cdot f(d_i, w_m)$$

$$f(d_i, w_m) = \begin{cases} 1 & \text{if} \quad tf(d_i, w_m) \neq 0 \\ 0 & \text{else} \end{cases}$$

$$IGF(w) = \log\{G/GF(w)\}$$

$$ICF(g,w) = \log\{C_g/CF(g,w)\}$$

$WA(w)$ は語句 w が 1 度でも出現する文書の文書話題度の総和で，話題性の高い文書での出現が多い語句ほど値が大きい．G はジャンル総数，$GF(w)$ は語句 w が出現する文書を 1 文書以上含むジャンル数 (出現ジャンル数: Genre Frequency)，C_g はジャンル g に含まれるクラスタ総数，$CF(g,w)$ はジャンル g 中で語句 w が出現する文書を 1 文書以上含むクラスタ数 (出現クラスタ数: Cluster Frequency) である．$IGF(w)$, $ICF(g,w)$ はそれぞれジャンル内，クラスタ内での特徴的な語句を抽出するための要素である．

各ジャンル，各クラスタにおいて，高い語句話題度を持つ語句が，文書集合中の話題概要把握に役立つ語群である．

5 実験

2004 年 11 月 1 日〜15 日の期間中に Web 上に掲載されたニュース記事を，[4] によって「社会」「国際」「経済」「企業」「政治」の 5 ジャンルに分類した合計 4758 記事を用い，現在時刻 t_0 を変えながら実験を行った．各パラメータの値は，今回は経験的に $r = 1.05$，$T = 2$(日) と固定した．

t_0 を 2004 年 11 月 9 日 00:00 としたときのジャンル分類による分類文書数とクラスタリングによる分類クラスタ数を表 2 に示す．現在時刻 t_0(= 11/9 00:00) 以前に収集された文書のみが処理対象になるため，表 2 に示した文書数の合計は実験に用いた文書総数 4758 より少なくなっている．

表 2: 文書分類結果 (合計 1479 記事)

ジャンル	文書数	クラスタ数
社会	363	71
国際	349	58
経済	126	29
企業	239	61
政治	402	67

ジャンル内文書数に占めるクラスタ構成文書数が最大のクラスタは「経済」ジャンルの原油価格急落を伝える記事で構成されるクラスタで，構成文書数は 15 であった．一方，クラスタ構成文書数の最小値は 1 で，各ジャンルに散見された．

次にクラスタを構成する各文書の文書話題度を算出した結果，文書話題度の大小は文書新鮮度の大小の順序とほぼ一致しており，互いに関連する文書群においては，新鮮度の高い文書ほど文書話題度が高くなる傾向が得られた．一方で，人間の印象とは多少反するクラスタが形成されるケースもあった．表 3 は,「国際」ジャンル内のクラスタにおける文書話題度算出結果の 1 例である．No.1〜4 の 4 文書がイラクからのハンガリー軍の撤退を伝える記事であり，No.5 はイラクで発生した警察官襲撃事件を伝える記事であった．No.1〜4 の記事と No.5 の記事とは,「イラク」というキーワードで同一クラスタに属すると判断されたものと思われるが，内容の関連度は極めて低い．各文書の作成日は，No.1〜4 がいずれも 11/4，No.5 が 1/7 であり，No.5 の文書新鮮度が圧倒的に高いにも関わらず，文書話題度は 5 記事中最下位である．これは，同一クラスタ内において他との類似度が低く孤立状態にある文書は，たとえ新鮮であってもクラスタを代表する文書とは言い難く，重要度が低いとみなす本手法の特徴による結果である．

続いて各記事中の名詞を取得し，文中で連続している場合には連結，複合名詞化して得た語群を評価の対象とし，式 (5) で与えられる語句話題度を算出した.「国際」ジャンル内のある 2 クラスタについて，クラスタ内の語句話題度上位 10 語を表 4 に示

表 3: クラスタ構成文書の文書話題度

No.	文書話題度 ($\times 10^{-2}$)	時刻	記事概要
1	2.5	11/4 01:15	イラクからのハンガリー軍の撤退
2	2.4	11/4 05:15	同上
3	1.6	11/4 01:45	同上
4	1.2	11/4 01:15	同上
5	0.8	11/7 14:45	イラクでの警察署襲撃事件

す．クラスタ C_1，C_2 とも，一部を除いてクラスタの内容を連想させる語句が上位を占めていた．

構成文書数の多い主なクラスタについて，それぞれのクラスタ内で評価された語句話題度の上位 3 語をピックアップした結果を表 5 に示す．例えば「社会」ジャンルでは新潟県での大地震の余震を伝える記事からなるクラスタがあり，そこでの語句話題度の上位 3 語は「魚沼市」「守門村」「震源地」であった．

最後に，特定の話題語に注目して語句話題度の時間推移を追跡した結果を図 3 に示す.「国際」ジャンルにおいて，語句「アラファト」の語句話題度 WT(“アラファト”) を，現在時刻 t_0 を 11/4〜11/15 まで 6 時間間隔で変えながら追跡している．11/5 には，アラファト議長の病状を伝える記事が多く発行された．一部では死亡説が出るなど情報が錯綜したこともあり，類似文書が増加し高いスコアを記録した．その後一旦沈静化するが，11 日に死亡が発表された後，葬儀関連のニュース (12 日)，次期議長選挙の日程決定 (15 日) などの関連記事が続き，再度スコアは上昇している．アラファト議長死去を伝える記事によるピークは，速報の受信から話題としての検出までのタイムラグにより葬儀のニュースのピークに隠れている．

表 4: 語句話題度 (上位各 10 語)

クラスタ	順位	語句	語句話題度 ($\times 10^{-4}$)
C_1	1	空爆	226.5
	2	ユーフラテス川	162.4
	3	中部ファルージャ	119.2
	4	武装勢力掃討	117.4
	5	イラク駐留米軍	109.2
	6	米武装勢力	108.9
	7	主要病院	103.0
	8	同市北部一帯	94.8
	9	救護施設	94.5
	9	同夜	94.5
C_2	1	ジュルチャーニ首相	11.5
	2	ブダペスト	10.1
	3	補給部隊	9.3
	3	撤退表明	9.3
	5	ハンガリー	8.3
	5	式典	8.3
	7	派兵	7.9
	8	ハンガリー通信	7.6
	9	それ以上	7.1
	10	爆弾攻撃	6.3

6 考察

文書クラスタリングにより，内容が徐々に変化する続報記事も含め，概ね良好に同一話題を集約することができた．仮にクラスタリング精度が低くても，クラスタを代表するに相応しくない文書の文書話題度は極端に小さくなるため，以後の処理への影響は小さい．したがって，本手法はクラスタリング精度に関してロバストな手法と言える．しかし，本手法をニュース記事以外の情報ソースに適用する場合には注意が必要である．ニュース記事は，迅速・的確に情報を伝達する目的から，使用する語句や文章の形態が必要かつ十分に精錬されており，既存のクラスタリング手法で高精度の分類を達成しやすい情報ソースである．他の情報ソースに対しては今回のようなクラスタリング精度は必ずしも期待できないため，適用する情報ソースに応じたクラスタリング手法が必要だと考えられる．

表 3 に示した結果は 1 例であるが，他のクラスタにおいてもクラスタを代表するに相応しい文書が高く評価されており，文書話題度による評価法は直感的に有効な手法と言える．また，表 5 に示した各クラスタは，それぞれ異なる題材を扱った記事で構成されており，クラスタリングと文書話題度算出による文書の話題性評価方法は，複数話題を含む文書集合の概要把握を行う上で有効な手法と言える．表 4 や表 5 に示した語句話題度のランク上位語の大部分も各クラスタの内容を連想しうる語句であり，話

表 5: 各ジャンルの話題語 [2004/11/9 00:00]

ジャンル	クラスタ	語句話題度の上位 3 語	クラスタ構成文書の概要
社会	1	魚沼市, 守門村, 震源地	新潟中越地震の余震
	2	請託, 議員証言法違反, 北海道開発庁長官	元議員の収賄事件の判決
	3	北海道苫前町沖, 砂利運搬船, エビかご漁船	漁船衝突事件の不明者捜索情報
	4	阪神大震災以来, 陛下, 皇后	天皇陛下の被災地訪問
国際	1	訪仏, パレスチナ解放機構, 容体	アラファト議長の見舞い延期
	2	ハマス, イスラム聖戦, 主流派ファタハ	アラファト議長死後の構想
	3	イラク聖戦アルカイダ組織, カイーダ組織, 聖戦アル	国際テロリストの犯行声明
	4	大統領主席補佐官, 結果判明, 有権者登録名簿	アメリカ大統領選挙の開票速報
経済	1	就業者数, 円買いドル売り, 景気動向	円高ドル安の加速
	2	円ユーロ, ドルポンド, 前週末	東京外為市場の値動き
	3	米国産標準油種, 暖房油, インターミディエート	NY 原油価格が大幅反落
	4	東京株式市場, 値幅制限, 東京外国為替市場	管理ポストに置かれた企業の株価が急落

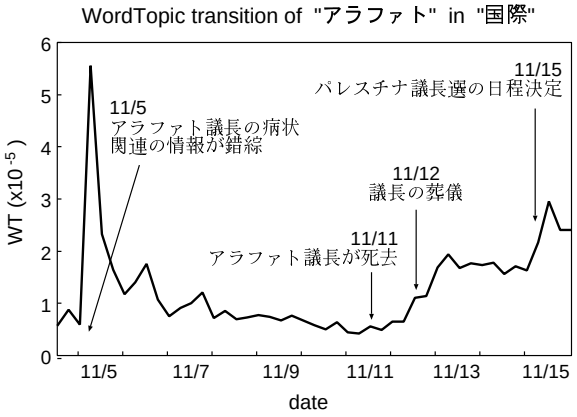

図 3: 語句話題度の推移

題語として受容されるものと考えられる．詳細な評価は今後行う予定である．

図 3 のグラフでは，各事柄の発生に追随するように語句話題度が上昇している．議長死去のニュースに比べて他の事柄の発生後のピークが際だって高いが，実際に議長の病状や次期議長選挙の注目度も高く記事数が増加しており，情報の「広がり」と「伸び」によって話題性を判定する本手法に照らし合わせても妥当な結果である．しかしながら，この結果が人間の直感的な印象と多少反するものであるのも事実であり，語句の話題性評価方法については再考の余地がある．

7 まとめ

話題性の高い事柄に対しては，社会的に注目される状態が時間的に継続する．今回はこのような情報の「広がり」と「伸び」に注目した最新話題語抽出手法を提案した．話題語の抽出にあたり，前処理として文書単位での同一話題集約を行うことで，複数話題を含む文書集合からの効果的な話題抽出を行うことができた．今後は，評価実験を行い各クラスタからの話題語抽出精度向上をめざすとともに，異なる情報ソースにも適用範囲を拡大する予定である．

参考文献

[1] Jon Kleinberg, “Bursty and Hierarchical Structure in Streams”, *Proc. the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002

[2] 仲村大也, 梅村恭司, “Katz’s K mixture による固有表現の異常出現の検出”, 情処研報 2001-NL-141, 2001

[3] Douglass R. Cutting, David R. Karger, Jan O. Pedersen and John W. Tukey, “Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections”, *Proc. 15th Annual ACM SIGIR Conf. on Research and Development in Information Retrieval*, pp.318-329, 1992

[4] 上田 修功, 斉藤 和巳, “多重トピックテキストの確率モデル – パラメトリック混合モデル – ”, 信学論 D-II Vol.J87-D-II No.3 pp.872-883, 2004

[5] 長尾真，他， “岩波講座 マルチメディア情報学 2 情報の組織化”, 岩波書店, pp.192-193