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

2005－ＮＬ－168（１）
2005／7／22

時系列ニュース記事における最新話題語抽出方法

佐藤　吉秀†　　川島　晴美†　　佐々木　努†　　奥　　雅博†

† 日本電信電話株式会社 NTT サイバーソリューション研究所
〒 239-0847 神奈川県横須賀市光の丘 1-1
E-mail: †{sato.yoshihide,kawashima.harumi,sasaki.tsutomu,oku.masahiro}@lab.ntt.co.jp

あらまし　逐次増加するニュース記事中に含まれる話題情報を効率的に把握するため，新鮮で可能な限り多くの幅広い話題情報を最新話題語と呼ぶキーワードの形態で抽出する手法について報告する．ニュース記事中の話題を扱うにあたり，世間の注目度が高い出来事を伝える記事数が増加する「話題の広がり」と，広がり状態が続報記事発行によって時間的に持続する「話題の伸び」の２つの側面に注目する．提案手法では，話題の整理のために記事をジャンル分類・クラスタリングした後，記事のタイムスタンプから算出する記事新鮮度，および記事間類似度を用いて各クラスタを代表する最新話題語を抽出する．ニュース記事（2164 記事）を対象にした評価実験の結果，提案手法はクラスタ中の新鮮かつ代表的な話題を表し，さらに受容性も高い語句を抽出可能であることを確認した．
キーワード　ニュース，最新話題語，クラスタリング，類似度，新鮮度

Latest Topic Words Detection from Chronological News Stream

Yoshihide SATO†, Harumi KAWASHIMA†, Tsutomu SASAKI†, and Masahiro OKU†

† NTT Cyber-Solutions Laboratories, NTT Corporation
1-1, Hikarino-oka, Yokosuka-shi, Kanagawa, 239-0847 Japan
E-mail: †{sato.yoshihide,kawashima.harumi,sasaki.tsutomu,oku.masahiro}@lab.ntt.co.jp

Abstract　We propose a method to detect 'Latest Topic Words' from incoming news articles to understand topical overview in them. Each word is representation of latest topical incident. Two aspects of topic in news stream have to be considered. One is accession of articles which play up same event, and the other is a continuation of follow up articles. At first we apply clastering method to news collection, evaluate significance of each article, and then extract a few latest topic words from each cluster using the significance. We estimated our method by experiment using 2164 news articles with time stamp and made sure the effect of it.
Key words　topic, keywords, clustering, news, similarity, freshness

1. はじめに

インターネットへの常時接続環境の普及や検索エンジンの高精度化，人々の情報収集スキル向上などにより，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．これに伴い，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たいと願う最新情報取得欲求も大きくなっている．特にニュース記事はリアルタイム性が高く，最新の情報でなければ価値が下がってしまうため，最新情報取得の意義が大きい文書の１つである．

このように次々と新規の文書が増加する最近の状況において，最新の状況を把握するため，RSS（RDF Site Summary）が注目されている．ニュース記事のヘッドラインや新製品情報など，最新の情報が RSS 形式で提供されており，ユーザは RSS リーダなどの RSS 閲覧ソフトを用いて各種最新情報を閲覧することができる．閲覧ソフトにも様々なインタフェースが存在し，その多くは文書のタイトル，概要，タイムスタンプなどの組を羅列した中からユーザが選択して本文へと到達する形式である．しかし，こういった方法ではユーザは多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閲覧コストを要する．登録したキーワードにマッチする文書のみをフィルタリングして閲覧することも可能であるが，母数の増加に伴いマッチ文書数が増加すれば，同様の問題が発生する．

そこで，このような問題を解決するための話題情報抽出手法が求められる．言語処理分野において，大量文書から話題情報を獲得，提示する手法は種々提案されているが，獲得対象の話題の規模をいかに設定するかが難しい．通常，文書集合中には

万人が認める圧倒的に大規模なものから，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ものまで，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．

これらの課題を踏まえ，本研究では複数の話題を含む大量の時系列ニュース記事中から，最新の話題情報を可能な限り多く，偏りなく抽出し，抽出した話題情報を端的に提示する手法の実現を目的とする．

以下，第2章で本研究の関連技術について，第3章では本研究における「話題」の定義を述べる．第4章で提案手法の詳細を，第5章で本提案手法の有効性を調べるために行った評価実験とその結果を述べる．第6章で実験結果を考察し，最後に第7章でまとめを述べる．

2. 関連技術

大量文書から重要なキーワードを抽出する手法は複数提案されている．

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って“話題”を単位として扱わないため，例えば大事件を伝えるニュースが含まれていた場合に，その事件に関わるキーワードばかりが偏って抽出される可能性がある．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現在時刻から近い特に新鮮な期間における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り，話題を“文書クラスタ”の単位で扱っている．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が，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．また，時間軸を考慮した抽出方式ではないため，新鮮な話題情報を提示することができない．

3. ニュース記事における「話題」

本研究では，ある情報に対する認知度が増し，ある時期において多数の人々に注目されている状態を“話題”と捉える．概念的には，図1のように，社会的認知性が増加した状態が時間的に持続し，社会的認知性および時刻の両軸に対しての大きさを持った領域が話題である．社会的認知性の大きさを話題の「横方向への広がり」と呼ぶならば，その状態の時間的継続は「縦方向への伸び」と呼ぶことができる．

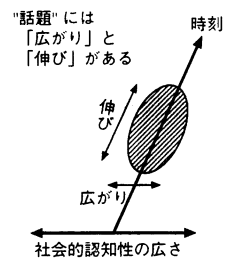

図1 話題の「広がり」と「伸び」

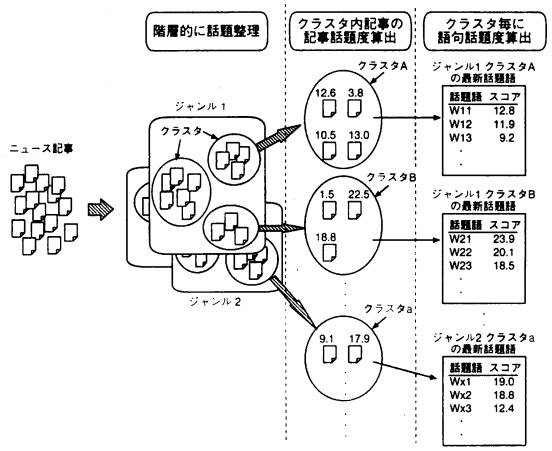

図2 最新話題語抽出処理の概要

タイムスタンプ付きの多数のニュース記事で構成される時系列データ中では，各記事自体が持つ話題性の大きさは，同一の出来事を伝える時間的に近い関連記事の量をもとに判断することができる．世間の注目度が集まることが推測される出来事ほど，各新聞社が続報記事を含めた多くの記事を発行するためである．この点に注目することで，膨大なニュース記事の中での話題を検出することが期待できる．

本研究では，上述のように「広がり」と「伸び」を持った話題を扱い，各話題を最新話題語と呼ぶキーワードによって代表させることで，膨大なニュース記事データ中に含まれる最新の話題情報を容易に取得できるようにする．

4. 提案手法

膨大なニュース記事中に含まれる話題情報を効率的かつ的確に伝達するため，本研究では話題を「最新話題語」と呼ぶ語句の形態で代表させて提示するアプローチをとる．これは，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である．語句での提示はタイトル等の文による提示に比べ，文書量が増加した場合の閲覧コストが少ない．

提案する最新話題語抽出処理の概要を図2に示す．まず，膨大なニュース記事中に含まれる話題の中から，可能な限り多く

の情報を提示するため，最初に記事単位での話題の整理を行う．整理は，記事をジャンル分類し，続いて各ジャンル中の記事をクラスタリングすることで階層的に行う．

記事単位で話題整理を行う理由は次の通りである．ニュース記事は，事実を正確かつ効率的に伝達する目的により，余分な記述を省いた必要十分な表現が用いらるため，関連する題材を扱った記事は，出現する単語の分布も類似しやすい．また，1記事に複数の題材を扱わず，扱う題材が異なれば複数の記事に分割して発行するのが通常である．このことから，1記事を1つの題材と捉え，記事の単位で関連記事のグルーピングを行うことで話題を整理することができる．

その後，話題整理のプロセスで生成された各クラスタ内の記事について，話題性の大きさ（記事話題度と呼ぶ）を算出し，最後に記事中に含まれる語句の重要性を表す語句話題度を算出して，各クラスタを代表すべき最新話題語を抽出する．

以下，各プロセスについて詳細に述べる．

4.1 話題の整理

ニュース記事の分類に一般に用いられる「政治」「スポーツ」「経済」「国際」などのジャンル体系に基づいて分類した後，さらに各ジャンル内で記事クラスタリングを行うことで，階層的に話題を整理することとした．

ジャンル分類

ニュース記事はあらかじめ上に挙げた「政治」「スポーツ」等の分類情報を持っている場合もあるが，記事の多義性を考慮すれば，重複を許した分類を行うのが望ましい．また，今回は分類情報未付与のニュース記事にも対応するため，自動的なジャンル分類手法を用いた．本研究で使用したのは，パラメトリック混合モデル（PMM）に基づく非排他的な多重トピック分類手法[4]である．あらかじめ人手でジャンル分類した記事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すべき1以上のジャンルを推定する．なお，ニュース記事本文中の単語の取得には文献[5]の形態素解析機能を利用し，名詞，カタカナの未知語，漢字の未知語として取得された形態素のみをジャンル推定に用いた．

記事間類似度

各ジャンル内で記事クラスタリングを行うために，記事間の類似度を算出する．同一の題材を扱ったニュース記事は出現する単語の分布が類似しやすいため，ベクトル空間モデルにおけるコサイン類似度を記事間類似度とする単純なクラスタリングでも，高い精度で同一題材の記事を集約することができる．

ベクトル空間モデルで文書をベクトル表現する際に用いられる代表的な単語の重み付け法にはTF-IDF法がある．しかし，ニュース記事では主題となりうる重要な単語（例えば首相，大統領等の人名や，会議等のイベント名称など）が記事の先頭付近に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）と単語の重要性との相関が必ずしも高くない．そこで，本手法では単語の出現記事数（Document Frequency）のみに基づいて，記事数 N の記事集合中の記事 d_i の記事ベクトルを以下のように定義する．

$$\boldsymbol{d}_i = (x_{i1},\ x_{i2},\ \cdots,\ x_{iV}) \quad (i = 1, 2, \cdots, N) \tag{1}$$

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

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

V は全記事中に出現するユニーク単語数であり，$df(w_v)$ は全 N 記事中で単語 w_v が1度でも出現する記事の数，$tf(d_i, w_v)$ は記事 d_i における単語 w_v の出現頻度である．本定義を用い，ジャンル内の任意の記事 i，j 間の類似度 S_{ij} を次式で求める．

$$S_{ij} = \frac{\boldsymbol{d}_i \cdot \boldsymbol{d}_j}{|\boldsymbol{d}_i||\boldsymbol{d}_j|} \tag{2}$$

記事クラスタリング

あらかじめクラスタ数が未知の場合にも適用可能なクラスタリング手法である最長距離法[6]を用いてジャンル内の記事クラスタリングを行い，話題を細分化する．記事間距離は，記事の非類似度 $(1 - S_{ij})$ を用いる．

クラスタリング手順は次の通りである．任意の1記事を中心とするクラスタを生成し（1），既存クラスタの中心から最も遠い記事を探索する（2）．一定の条件を満たせばその記事を中心とするクラスタを生成し（3），満たさなければ処理を終える（4）．

（1） 記事 d_1 を中心とするクラスタを生成．

（2） $l = \max_i \{\min_j d(d_i, \overline{Z_j})\}$ 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．

4.2 記事話題度算出

クラスタリング処理を終えた時点で，関連する題材を扱った記事からなるクラスタが生成されている．以後の処理はクラスタ毎に行う．

本ステップでは，語句に対する評価値を算出して最新話題語を抽出する前処理として，クラスタ内の各記事が持つ話題性の大きさを表すスコアである記事話題度を算出する．記事話題度は，クラスタ内でも特に新鮮で重要な記事中の語句を，そのクラスタを代表する最新話題語として抽出するために必要な指標である．

以下，記事話題度の算出に用いる記事新鮮度について述べた後，記事新鮮度と記事間類似度に基づく記事話題度算出方法について述べる．

記事新鮮度

新鮮な話題の抽出のため，記事の作成時刻に基づく記事新鮮度を定義する．作成時刻が t_i である記事 d_i の新鮮度 $Fresh(i)$ を式(3)のように与える．続々と新着記事が到着する状況において常に最新の話題情報を抽出するため，時刻に比例する重み

表 1　記事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}	$\cdots$	$\sum S_{ij}/N_C$
$\vdots$		$\vdots$			$\vdots$

ではなく，指数関数として与え，記事の新鮮さを特に強調する．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．

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

記事話題度

記事の新鮮度，記事間類似度を用い，各記事のクラスタ内における重要性を表す記事話題度を算出する．

クラスタリングの結果，関連記事がグルーピングされたクラスタ中に，他との類似性が比較的低い記事が混在することがある．何らかの単語が大きな影響を与えて集約してしまうことが原因だが，クラスタリングが一意な解を持たない問題である以上，クラスタリング手法によらず起こりうる．

クラスタ内の他の記事との類似性が低い記事は，クラスタを代表する重要な記事とは呼べない．そこで，クラスタ中の代表的な記事に重みを与えるため，類似度を用いた以下の手法で記事話題度を算出する．まず，表 1 に示すように，注目する記事 d_i とクラスタ内の他の記事との類似度の平均値 $\sum_j S_{ij}/N_C$ を求める．N_C は注目記事が属するクラスタサイズ（クラスタ構成記事数）である．クラスタを構成する記事群の重心に近い中心的な記事であるほどこの値は大きい．これに上述の記事新鮮度 $Fresh(i)$ を乗じ，重要性が高く，かつ新鮮な記事を評価する記事話題度 $DT(i)$ を得る．(式 (4))

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

(C は d_i の所属クラスタ)

4.3　語句話題度算出

高い記事話題度を持つ記事中に高頻度で含まれ，かつクラスタを特徴付ける語を最新話題語として抽出するため，式 (5) に従って語句話題度 WT を算出する．語句話題度算出の対象語句は，[5] により取得した名詞ならびにカタカナ，漢字の未知語であり，文中で連続している場合にはそれらを連結し，複合語化した．

$$WT(w) = WA(w) \cdot ICF(g,w) \cdot IGF(w) \tag{5}$$

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 度でも出現する記事の記事話題度の総和で，話題性の高い記事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 が出現する記事を 1 記事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 が出現する記事を 1 記事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．

以上の処理により，時系列ニュース記事がクラスタで整理され，語句話題度がクラスタ内の記事に含まれる各語句について計算された状態になる．クラスタ中で高い語句話題度を持つ語句が，そのクラスタが指し示す話題の概要把握に役立つ語句である．

5.　評価実験

本手法の最新話題語抽出精度について調べるため，評価実験を行った．実験方法は，90 名の被験者による主観評価であり，以下の 2 つの観点で実施した．

- 新鮮で代表的な話題を表す語句を抽出できているか
- 抽出された最新話題語は受容されるか

前者は，意味に基づく話題整理の結果である各クラスタから，新鮮かつ代表的な話題を表す語句が最新話題語として抽出される割合（抽出成功率）を調べる実験であり，後者は，最新話題語として抽出した語句が，元のクラスタを代表する話題語として相応しいか否かを調べる実験である．それぞれを評価 A，評価 B として行った．

実験条件

2005 年 1 月 10 日～17 日の 8 日間に goo ニュース [7] から収集した記事を，[4] により「スポーツ」「社会」「国際」「経済」「企業」「政治」の 6 ジャンルに分類した合計 2164 記事を使用した．さらに記事クラスタリングを行った後，式 (5) に基づいて語句話題度を算出し，各クラスタの語句話題度上位 10 語を最新話題語として抽出した．なお，各パラメータの値は，経験的に $r = 1.05$，$T = 2$（日）とした．

提案手法が $Fresh$（新鮮度），S（類似度），$IGF \times ICF$（ジャンル・クラスタ特徴度）の 3 要素に基づいて語句話題度を算出するのに対し，$Fresh$ のみに基づいてスコアリングした場合（以下，FRE），$Fresh$ および S に基づいてスコアリングした場合（以下，SIM），これとは別に 1 クラスタに含まれる記事を 1 文書とみなして TF-IDF 法を適用した場合（以下，

表 2 「国際」ジャンル内のクラスタから抽出された語句上位 10 語

順位	提案手法	FRE	SIM	TF-IDF
1	投票所	16 日	16 日	バグダッド
2	ハマス	武装勢力	イスラエル	投票所
3	武装闘争	イスラエル	死亡	国民議会選挙
4	イスラエル	死亡	イラク	イラク
5	アッバス氏	イラク	アッバス氏	意思表示
6	本誌	アッバス氏	表明	ウサマ
7	過激派	表明	14 日	イラク暫定政府
8	議長	明らか	議長	選挙
9	首謀者	攻撃	明らか	フセイン政権
10	テロ停止	可能性	可能性	アラウィ首相

TF-IDF）も含めた 3 手法を比較手法とし，それぞれ 10 語ずつ抽出したものを用いた．各ジャンルに多くのクラスタが生成されるが，クラスタサイズが 5 以上のクラスタを各ジャンルからランダムに 5 クラスタ選択し，6 ジャンル計 30 クラスタをテストセットとした．表 2 に，「国際」ジャンル内のあるクラスタから各手法によって抽出された語句上位 10 語を示す．

評価 A： 新鮮かつ代表的話題からの抽出率

時間および他記事との類似性に基づく提案手法の効果を確認するため，以下の手順で評価を行った．

（A1） クラスタ内記事の見出しと本文を被験者に提示する
（A2） クラスタ内の代表的な話題を構成する記事を選択させる
（A3） クラスタ内記事のうち，タイムスタンプが新しい記事を 1/3 取得する
（A4） A2，A3 の両条件を満たす記事を正解記事，その他の記事を不正解記事とする
（A5） 不正解記事での出現頻度（注目語句を含む不正解記事の数）が 2 未満の語句を正解とする

（A1）および（A2）が被験者に課したタスクである．各被験者の（A2）の回答結果の多数決を取り，クラスタ内での代表的な話題を構成する記事を決定する．また，タイムスタンプによりソートしたクラスタ内記事のうち，特に新しい記事を（A3）で取得する．続いて（A2），(A3）の両方の条件を満たす記事，すなわちクラスタ内の新鮮かつ代表的な話題の構成記事を正解記事とし，それ以外を不正解記事として分類した．最後に，各手法で抽出した 10 語それぞれについて，(A5）の方法で正解/不正解を判断し，各語句の正解率（抽出成功率）の平均値を算出した．なお，本研究の目的は，クラスタ内に含まれる新しくて代表的な話題を表す語句（不正解記事よりも正解記事に偏って出現する語句）の抽出である．この観点から語句の正解/不正解を判断するために，(A5）のように不正解記事数に基づいた判定を行うこととした．(注1)

結果を表 3 に示す．TF-IDF に比べて他の 3 手法が新鮮かつ代表的話題からの抽出成功率が高いことがわかったが，それら 3 手法の間に大きな差は見られなかった．

(注1)：正解記事での出現頻度で判定すると，クラスタ内の高頻度語（例えば「昨日」などの一般語）を高く評価することになってしまう．また，不正解記事に出現する語句が全て不正解というわけではなく，たとえ不正解記事に含まれていても，それが新鮮かつ代表的な話題に関わる語であれば正解と判定すべきである．

表 3 新鮮かつ代表的話題からの抽出成功率 （単位: %）

提案手法	FRE	SIM	TF-IDF
67.3	66.0	66.8	44.3

表 4 最新話題語の受容性 （単位: %）

	提案手法	FRE	SIM	TF-IDF
◎	42.5	27.2	26.7	51.9
○	28.5	29.7	29.6	26.7
△	15.3	20.6	21.0	11.4
×	13.6	22.5	22.7	10.1

◎: 適当　　△: あまり適さない
○: まあまあ適当　×: 不適
（※）四捨五入により合計値が 100%にならない場合がある

評価 B： 最新話題語の受容性

続いて，本手法による最新話題語がクラスタ内の話題情報を表すキーワードとして適するか否かを評価するため，クラスタ内の記事から 4 手法により抽出した語句各 10 語を混合したリストを被験者に提示し，話題語としての適性を回答させるタスクを課して実験を行った．評価 A の実験手順（A1）で記事の見出しと本文を読ませた後，対応するクラスタから抽出された語句の混合リストを提示した．回答は，各語句が話題語として「適当」「まあまあ適当」「あまり適さない」「不適」の 4 択とした．

なお，評価 A の（A2）で記事を選択した後に本評価を行うと，評価 A での選択記事を意識した回答になってしまう可能性があることから，先入観を排除するために，実際の手順では評価 B，評価 A の順に実験した．つまり，評価 A の（A1）で記事を提示した後に語句リストを提示して本タスクの受容性評価を行い，その後（A2）で記事を選択させた．

本評価実験の結果を表 4 に示す．ポジティブな評価（「適当」および「まあまあ適当」）の合計値は提案手法，FRE，SIM，TF-IDF の順に 71.0，56.9，56.3，78.6 となり，提案手法は FRE，SIM に比べて高いが，さらに TF-IDF が高い値を示した．

6. 考　　察

以上の評価結果をまとめると，評価 A では TF-IDF を除く 3 手法が新鮮かつ代表的な話題の抽出精度で優れ，評価 B では TF-IDF，提案手法の順に受容性が高いことがわかった．以下，各評価実験の結果を詳細に考察し，それらの結果を踏まえた総合判定を行う．

評価 A の考察

新鮮度に基づいて語句をスコアリングするのが FRE，これに記事間類似度の効果を加え，クラスタ重心に近い共通項的・中心的な記事に重みを与えたことで代表的な話題からの抽出率向上が期待されるのが SIM である．実験を行った結果，ニュース記事の場合は，ベクトル間距離による単純なクラスタリング手法でも同一題材の記事がある程度精度よくグルーピングされ，クラスタ中に異質な記事が混入しにくかったため，FRE から

表 5　評価実験結果のまとめ

	提案手法	FRE	SIM	TF-IDF
評価 A	○	○	○	×
評価 B	○	×	×	○
総合判定	○	×	×	×

SIM に対して期待したほどの改善効果が見られなかった．しかしながら，クラスタ内にいくつか見られた多少異質な記事の記事話題度は全般的に低く，クラスタを代表するのに相応しい記事を得る方法として，類似度を用いた記事話題度算出方法は有効だと考えられる．

また，提案手法も上記 2 手法に対して大きな違いが見られなかったが，原因として記事新鮮度の影響が強かったことも挙げられる．今回，時系列ニュース記事データに含まれる特に新しい話題を取り扱うために，式 (3) のように指数関数として記事新鮮度を定義した．その結果，時間軸に対して敏感な抽出を行える反面，その他の効果が多少埋もれてしまったようである．

評価 B の考察

語句の受容性については TF-IDF，提案手法の順に高く，FRE と SIM は低かった．TF-IDF の受容性が提案手法より高かったのは，「国民議会選挙」などの比較的長い文字列や，人名，地名等の固有名詞によるところが大きかった．これらの語句は，インパクトが強いために印象に残りやすい．また，クラスタ内の記事に共通する用語であることが多く，評価 A の (A1) でクラスタ内の記事を読んだ被験者にとっては，目に触れる頻度が高かった語句だと言える．そのため，提示された時に元の記事が容易に想起しやすく，ポジティブな評価を行うことが多かったものと推測される．なお，提案手法でも固有名詞は抽出されていたが，全体に占める割合は TF-IDF に比べると低かった．

提案手法は，時間軸を考慮した処理の効果により，同一クラスタに属する記事の中でも特に新しい記事に目立って用いられる語句が多く抽出されていた．例えば，犯罪容疑者の逮捕後の供述内容を伝えるニュースから，その後実況検分が行われたことを伝えるニュースまで，一連の続報記事からなるクラスタがあったが，その中で時間的に新鮮な情報である「実況検分」が提案手法では抽出されていた．ところが，この語を含む記事はクラスタ内の最新の数記事であったため，クラスタ内での頻出語が抽出されやすい TF-IDF では抽出されなかった．

一方，FRE および SIM の抽出結果には，「表明」「明らか」「可能性」などの一般名詞や「16 日」などの日付表現などが比較的多く含まれていた．単独では話題の内容を類推不可能なため，受容性が低かったものと思われる．

総合判定

評価実験結果のまとめを表 5 に示す．評価 B の結果では，提案手法より TF-IDF の受容性が高かったが，クラスタ内の新鮮かつ代表的な話題を表す語句の抽出成功率は，評価 A の結果から提案手法が優れている．たとえ受容性が高い語句であっても，伝えるべき最新の話題情報を表現していない場合は，最新話題語とは呼べない．時々刻々と進展する状況において，常に最新の話題情報を提示し続けるのが本研究の目指すところである．例として挙げた「実況検分」は，全体から見た出現頻度は決して高くないが，一連の報道の最新状況を表す重要な最新話題語である．

このように，提案手法は，最新の話題情報を表し，かつ受容性の高い語句を抽出する方法として優れている．逆に，TF-IDF は，時系列ニュース記事中から最新の話題情報を抽出・端的に提示するという本研究の目的において最適な手法ではない．

しかしながら，評価 B の結果が示す TF-IDF の受容性の高さを無視してよいわけではない．この結果は，固有名詞のように，ある程度話題の内容を類推できる語句が望まれたことを示している．よって，大量記事中に含まれる話題情報を効率的に伝達するためには，評価 A の結果の向上のみに注力するのではなく，語句の受容性も考慮して抽出手法を改善する必要がある．

7. ま と め

同一の題材を扱ったニュース記事における単語の出現分布の類似性に注目し，各記事クラスタから最新話題語と呼ぶキーワードを抽出する手法を提案した．初期ステップとして，ジャンル分類およびクラスタリングによる話題整理を行うため，様々な話題から抽出された最新話題語を同時に提示することが可能になる．これは，幅広い話題情報の効率的な把握につながる．

今回，語句話題度という単一の尺度で語句をスコアリングし，上位語を抽出するアプローチを取ったが，評価実験の結果，新鮮で話題性の高い語句が必ずしも受容されるわけではないことがわかった．今後は，さらに受容性の高い話題情報提示方法の検討を行う．これには，例えば，TF-IDF が抽出を得意とするクラスタ内での共通的な用語と，本手法が得意とする新鮮な事象を表す語を組み合わせて提示するなど，単一の語句による表現以外の方法も視野に入れた幅広い検討が必要である．

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] 齋藤邦子, 永田昌明, "HMM に基づく多言語固有表現抽出システムの開発", 言語処理学会第 9 回年次大会, pp.5-8, 2003

[6] 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

[7] goo ニュース http://news.goo.ne.jp/