Web ニュース上のホットトピックスの効率的な検索手法

坪川 貴和† 五島 洋行‡

†長岡技術科学大学 工学研究科 経営情報システム工学専攻 〒940-2188 新潟県長岡市上富岡町 1603-1
‡長岡技術科学大学 経営情報系 〒940-2188 新潟県長岡市上富岡町 1603-1
E-mail: †s073368@stn.nagaokaut.ac.jp, ‡hgoto@kjs.nagaokaut.ac.jp

あらまし 本研究では，Web 上に膨大な数存在するニュース記事の中から，ユーザーの興味がある話題に関連する記事を，効率的に検索する手法を提案する．具体的には，Web ニュースストリームのバースト性に注目して，1．注目されている話題とその注目されている期間に関する理解を補助するグラフ，2．検索の際に用いることで，それらの話題それぞれを容易に取得可能にする単語，の二つをユーザーに提示する．もし，ユーザーが提示されたグラフを参照したとき，特定の話題に興味を持ったならば，提案手法が提示した単語を検索エンジンに入力して検索を行なうことで，その話題に関連する記事を効率的に検索することができる．

キーワード クエリ推薦，Web ニュース，文書ストリーム，バースト検出，データマイニング

A Method of Extracting Assistive Words for Retrieving Hot Topics from Web News

Takakazu TSUBOKAWA† and Hiroyuki GOTO‡

†Faculty of Management and Information Systems Engineering, Nagaoka University of Technology Nagaoka 1603-1, Kamitomiokamachi Nagaoka, Niigata 940-2188, Japan
‡Department of Management and Information Systems Science, Nagaoka University of Technology Nagaoka 1603-1, Kamitomiokamachi Nagaoka, Niigata 940-2188, Japan

1. はじめに

本研究では，Web 上のニュース記事から，目的の記事を効率的に検索する方法論について考察する．検索エンジンを用いた Web ニュース検索において，情報検索に不慣れな，または話題の動向に詳しくないユーザーがキーワードを直感的に決定した場合，漠然とした内容となる傾向があるため，検索結果の記事の数が多くなりすぎてしまう．この際，検索結果には意図した内容の記事だけではなく，それ以外の記事が多く含まれていることが多い．よって，漠然とした内容のキーワードでは，意図した内容の情報が記載された記事のみに検索結果を絞ることは難しい．つまり，情報検索に慣れている，または興味ある話題の動向に詳しいユーザーでない限り，目的の記事のみに結果をすばやく絞れるキーワードを的確に設定することは困難である．したがって，情報検索に不慣れな，もしくは興味のある話題に詳しくないユーザーであっても，目的の記事を効率的に得ることができる手法が必要である．

そこで我々は，Web ニュースの効率的な検索方法を実現するために，1. 検索を行う前に，ユーザーの取得したいニュースの内容を具体化させる，2.その情報に関する記事の取得を可能にする，の二つを満たす方法論を検討すべきだと考えた．1.を満たすためには，Web ニュースにおける話題の動向を理解させる必要がある．この際，それらの理解に必要な時間は極力短い方が好ましい．人間がある物事を短時間で理解する場合，一般的にその概要や特徴を必要とする．そこで，Web ニュースにおける話題の動向の概要や特徴を，検索を行う前にユーザーに提示することにより，取得したいニュースの内容を具体化させることを目指す．この際，ユーザーに提示する話題の動向の概要や特徴を抽出するために，Web ニュースを文書ストリームの一種とみなし，Web ニュースストリームのバースト性が高い部分に属する記事に注目した．これらをふまえ，最近我々は，ユーザーが入力した検索キーワードに対して，以下に示す二つを出力する方法論を考案した[1].

① 入力されたキーワードに関するホットトピックスと，それらの注目度合いと注目された期間を提示する．

② ①のホットトピックスに関する記事の中から，それらの記事の特徴をよく表す単語を抽出する．

ここで，ホットトピックスとはバースト性が高かった話題のことを指す．①より，ユーザーは，キーワードに関する話題を理解できる．もし，①の中に興味がある話題があれば，②で抽出された単語のうち，興味を持った話題を検索するための単語を，検索キーワードとして検索エンジンに追加すると，キーワードに関するホットトピックスの記事が効率的に検索できる．

従来法[1]では，①と②の出力は，ユーザーがキーワードを入力した後に行っていた．しかし，ユーザーがキーワードを入力していない状態でも，ホットトピックスとそれらを検索可能にする単語は提案されるべきである．なぜなら，キーワードの入力をする必要がない分，従来法[1]に比べてより効率的な検索を行えるからである．つまり，キーワードが未入力の状態であっても，いくつかのホットトピックスに関する情報をユーザーに提示することにより，特定のホットトピックスに興味を促すことができれば良い．そうすれば，興味を持ったホットトピックスに関する記事を検索可能にする単語を推薦し，それを検索に利用してもらうことで,ユーザーはキーワードを自ら決定すること無く，興味のある内容の記事を検索できる．そこで本論文では，ユーザーがキーワードを直接入力せずとも，興味があるホットトピックに関連する記事を，効率的に検索できる方法論を新たに提案し，その有用性の評価を行う．

2. 提案手法

2.1. 概要

Web ニュースストリームのバースト性に注目して，１．ホットトピックスとそれらの注目されている期間に関する理解を補助するグラフ，２．検索の際に用いることで，それらの話題それぞれを容易に取得可能にする単語，の二つをユーザーに提示する．１．のグラフと２．の単語を，それぞれトピックグラフと検索支援語と呼ぶ．トピックグラフは，ホットトピックスに対するユーザーの興味を促し，検索支援語は，ホットトピックスに関する記事を取得可能にすることが期待される．なお本研究では，検索エンジンを用いて Web ニュース検索を行うことを前提としている．

クエリ推薦に関する多くの先行研究では，今井ら[2]の研究のように，クエリログを解析するものが多い．しかし本研究では，Web ニュース記事を解析して，ユーザーに提案する単語を抽出する．

2.2. 全体フロー

提案手法は，大きく分けて，以下に示す５つの処理からなっている．

① インデキシング．
② グループ分け．
③ 注目度と注目期間算出．
④ 検索支援語抽出．
⑤ 検索支援語の評価．

①では,解析対象の全ニュース記事それぞれに対して，インデキシングを行い，それらをベクトル空間モデル[3]で表現する．インデキシングは，対象となる記事に対して MeCab[4]を用いて形態素解析を行い，一般名詞または固有名詞のみを索引語として扱う．この際，索引語の重みを TF(Term Frequency)[5]法より与える．つまり，n 件の記事に対しインデキシングを行い，m 個の索引語を得た場合，n 件の記事を，n 個の文書ベクトル $\boldsymbol{d}_1,\boldsymbol{d}_2,\cdots,\boldsymbol{d}_n$ それぞれが m 個の索引語 $w_1,w_2,\cdots,w_m$ を持つ，サイズ $m\times n$ の索引語-文書行列 $\boldsymbol{D}$ として以下のよう表現できる．

$$\boldsymbol{D}=\begin{pmatrix}\boldsymbol{d}_1 & \boldsymbol{d}_2 & \cdots & \boldsymbol{d}_n\end{pmatrix}=\begin{pmatrix}d_{11} & d_{12} & \cdots & d_{1n}\\ d_{21} & d_{22} & \cdots & d_{2n}\\ \vdots & \vdots & \ddots & \vdots\\ d_{m1} & d_{m2} & \cdots & d_{mn}\end{pmatrix}.$$

ここで，d_{ij} は索引語 w_i の文書 $\boldsymbol{d}_j$ における重みである．また，$d_{ij}=\mathrm{tf}(w_i,\boldsymbol{d}_j)=f_{w_i,d_j}$ であり，f_{w_i,d_j} は w_i の $\boldsymbol{d}_j$ における出現回数である．②では，全ニュース記事をその内容が似ている記事同士のいくつかのグループに分割する．それらのグループを，それぞれ話題として扱う．グループ分けには，PDDP (Principal Direction Divisive Partitioning)[6]アルゴリズムを用いる．ただし，②の処理の場合のみ，索引語の重みを，TF[5]法ではなく，TF-IDF(Term Frequency Inverse Document Frequency) [5][7]法を用いて，$d_{ij}=\mathrm{tf}(w_i,\boldsymbol{d}_j)\times\mathrm{idf}(w_i)=f_{w_i,d_j}\times\log(n/n_{w_i})$ で与える．ただし，n は文書の総数，n_{w_i} は w_i を含む文書の数である．③では，すべての話題の注目度とその注目期間を算出する．この際，②で生成された各クラスタに属する記事の，全記事に対する出現頻度に基づいて，各クラスタの注目度と注目期間を算出する[1]．④では，それぞれの話題からその特徴を表す単語を抽出する．最後に⑤では，評価関数を用いて検索支援語を評価し，その評価が高い検索支援語を優先的にユーザーに出力する．また，トピックグラフは，処理③の後に，出力できる．検索支援語を用いて，AND 検索することで，Web ニュースの効率的な検索を実現する．ここで AND 検索とは，指定したす

べてのキーワードに関する記事の検索を行う機能を指し,検索エンジンはこれを基本機能として備えている.

2.3. 注目度と注目期間の算出

この処理では，2.2 の②の処理で生成された話題それぞれの，注目度とその注目期間を求める．話題の注目度とその注目期間は，クラスタに属している記事の出現数のバースト性に注目して，特定クラスタに属する記事が集中的に出現している期間とその集中具合を定量的な値で算出する.

有名な文書ストリームのバースト検出法として，Kleinberg[8]の手法がある．しかし，この手法は，特定単語を含む文書の出現頻度に基づいてバーストの検出を行なっているため，提案手法にそのまま適用することができない．そこで我々は，Kleinberg の手法を拡張して，特定クラスタに属する記事の出現頻度に基づいたバースト検出法を考案した[1]．つまり，特定クラスタに属する記事が集中的に出現している期間を話題の注目期間，その集中具合の定量的な値を注目度として扱う．以下で定義を示す.

各時刻 $T_1,T_2,\cdots,T_n$ に対応して，記事集合 $\boldsymbol{D}=\{\boldsymbol{D}_{T_1},\boldsymbol{D}_{T_2},\cdots,\boldsymbol{D}_{T_n}\}$ が与えられたとする．同様に，あるクラスタ $\boldsymbol{C_k}$ を，各時刻 $T_1,T_2,\cdots,T_n$ それぞれにおける記事集合の集まりであると考えると $\boldsymbol{C}_k=\{\boldsymbol{C}_{k,T_1},\boldsymbol{C}_{k,T_2},\cdots,\boldsymbol{C}_{k,T_n}\}$ と表せる．ただし，$\boldsymbol{C_k}\in\boldsymbol{D}$，$\boldsymbol{C}_{T_i}\in\boldsymbol{D}_{T_i}(i=1,2,\cdots,n)$ である．全期間 $[T_1,T_n]$ において，記事が一定の確率で定常的に生成されるとすれば，$\boldsymbol{C_k}$ の属する記事の生成確率は，以下のように推定できる.

$$p_{k,0}=\sum_{t=T_1}^{T_n} m_{k,t}\Big/\sum_{t=T_1}^{T_n} M_t .$$

ただし，M_t は $\boldsymbol{D}_t$ に属する記事の数，つまり，ある時刻 t における生成された記事の総和であり，$m_{k,t}$ は t における $\boldsymbol{C}_{k,t}$ に属する記事の数である.時刻 t において，M_t および $p_{k,0}$ が与えられている場合，$m_{k,t}=p_{k,0}\times M_t$ を満たす確率 $q_{M_t}(m_{k,t}:q_{k,0})$ は，次の二項分布に従う.

$$q_{M_t}(m_{k,t}:q_{k,0})={}_{M_t}C_{m_{k,t}}\times p_{k,0}^{m_{k,t}}\times(1-p_{k,1})^{M_t-m_{k,t}} .$$

一方，ホットトピックス期間においては，$\boldsymbol{C_k}$ に属する記事が生成されやすい注目状態にあるとして，その出現確率を以下のように定義する.

$$P_{k,1}=s\times P_{k,0}\geq P_{k,0}\quad(s\geq 1).$$

$p_{k,0}$ および $p_{k,1}$ を用いて，任意の期間 $[t_1,t_2]$ における注目度 $G(t_1,t_2:k)$ を，二項分布の尤度比として以下に定義する.

$$\begin{aligned}G(t_1,t_2:k)&=\prod_{t=t_1}^{t_2} q_{M_t}(m_{k,t}:q_{k,1})\Big/\prod_{t=t_1}^{t_2} q_{M_t}(m_{k,t}:q_{k,0})\\&=\prod_{t=t_1}^{t_2}(q_{k,1}/q_{k,0})^{m_{k,t}}(1-q_{k,1}/1-q_{k,0})^{M_t-m_{k,t}}\end{aligned}\quad\cdots\quad(1)$$

実際の計算では，$G(t_1,t_2:k)$ の対数尤度比 $g(t_1,t_2:k)=\ln G(t_1,t_2:k)$ を算出し，$g(t_1,t_2:k)$ が最も大きいときの $[t_1,t_2]$ をクラスタ C_k の注目期間 $[t_{g,1},t_{g,2}]$ として抽出する．ただし，$t_{g,1}<t_{g,2}$ である.

このようにして求めた各クラスタの注目度と注目期間を，話題の注目度と注目期間とみなす.

2.4. 検索支援語の抽出

検索支援語は，ホットトピックスに関する記事を容易に取得するための単語である.その抽出方法を示す.

PDDP[6]アルゴリズムによって生成された n 個のクラスタ $\boldsymbol{C_1},\boldsymbol{C_2},\cdots,\boldsymbol{C_n}$ を定義する.各クラスタ $\boldsymbol{C_1},\boldsymbol{C_2},\cdots,\boldsymbol{C_n}$ に属する記事から，そのクラスタが持つ注目期間の範囲内のタイムスタンプを持つ記事のみを抽出し，その記事のグループを，それぞれ $\boldsymbol{C}_1',\boldsymbol{C}_2',\cdots,\boldsymbol{C}_n'$ とする．そして，$\boldsymbol{C}_1',\boldsymbol{C}_2',\cdots,\boldsymbol{C}_n'$ それぞれの重心ベクトルの要素のうち，その値が大きいいくつかの索引語を，$\boldsymbol{C_1},\boldsymbol{C_2},\cdots,\boldsymbol{C_n}$ それぞれのための検索支援語とする.

言い換えると，$\boldsymbol{C}_i=(\boldsymbol{d}_{i,1}\ \ \boldsymbol{d}_{i,2}\ \cdots\ \boldsymbol{d}_{i,p})((i\leq n),(p=\boldsymbol{C}_i$ に属する記事の数))に，$\boldsymbol{C}_i$ の注目期間 $[t_{i,1},t_{i,2}](t_{i,1}<t_{i,2})$ の範囲のタイムスタンプを持つ記事が $r(\leq p)$個あった場合，$\boldsymbol{C}_i$ から抽出された記事の集合であるクラスタ $\boldsymbol{C}_i'$ は，$\boldsymbol{C}_i'=(\boldsymbol{d}_{i',1}\ \ \boldsymbol{d}_{i',2}\ \cdots\ \boldsymbol{d}_{i',r})$ となる．この時，$\boldsymbol{C}_i'\in[t_{i,1},t_{i,2}]$ である．$\boldsymbol{C}_i$ のための検索支援語は，$\boldsymbol{C}_i'$ の重心ベクトルより抽出されるので，$\boldsymbol{C}_i'$ の重心ベクトルを $\boldsymbol{G}_i'$ とすると，$\boldsymbol{G}_i'=\boldsymbol{C}_i'\boldsymbol{e}/r=(g_1'\ \ g_2'\ \cdots\ g_m')^T$ より，$g_1'\sim g_m'$ のうちその値が大きいものを検索支援語として抽出する.

2.5. 検索支援語の評価

抽出したいくつかの検索支援語に優先順位を与えるために，評価関数を定義し，評価の高い検索支援語を優先してユーザーに提案する．この評価関数を優先度と定義する．優先度は，下記に示す二つの概念を考慮する.

- 話題のバースト性
- 情報の価値の時間的な減衰

話題のバースト性の概念は，式**(1)**を用いる．情報の価値の時間的な減衰については，以下で説明する.

情報の質は時間の経過と共に低下するはずである.よって，検索支援語を用いて得られることができるニュースは，検索時刻に近いものであることが好ましいと仮定する．従って優先度は，ホットトピックスの注目期間の終了時刻から検索時刻までの間に，時間の経過とともに減衰するものとして扱う．優先度の時間的減衰は，石川ら[9]の研究で用いられている，人間の記

憶の忘却曲線を用いて表現する．これは，指数関数減衰モデルと一致する．モデルには，以下に示す仮定を反映させる．

- ホットトピックスの注目期間の終了時刻 t_{g2} からユーザーが検索を行う時刻 t までの間に，優先度は減衰する．
- 優先度は注目期間 $[t_{g,1},t_{g,2}]$ が長い検索支援語ほど減衰しにくく，$[t_{g,1},t_{g,2}]$ が短いほど減衰しやすい．

以上の仮定より，減衰度 $\sigma(t)$ を以下のように定義する．

$$\sigma(t)=\exp[-a(t-t_{g,2})/(t_{g,2}-t_{g,1})]. \quad (2)$$

ただし，t は検索時刻，a は減衰定数である．また，$0<\sigma(t)\leq 1$ である．$\sigma(t)$ が小さいほど，優先度は小さくなる．

ここで優先度について説明する．注目期間が $[t_{g,1},t_{g,2}]$ のある話題 C_k を，容易に検索するための検索支援語を考える．この検索支援語を，ユーザーに優先的に提案する優先度合いを，下記の評価関数 Z を用いて算出する．

$$Z=\sigma(t,t_{g,1},t_{g,2})\cdot G_{c_k}(t_{g,1},t_{g,2}). \quad (3)$$

ただし，$\sigma(t,t_{g,1},t_{g,2})$ は式**(1)**と，$G_{c_k}(t_{g,1},t_{g,2})$ は式**(2)**とそれぞれ一致する．つまり，評価関数 Z は，C_k に属する記事の，$[t_{g,1},t_{g,2}]$ における集中的な出現度合いと，時間的な減衰度合いを乗じることにより求まる．言い換えると，C_k の注目度と注目期間，およびその話題の新鮮度により，C_k に対応している検索支援語の評価は高くなる．よって，注目度が高い，注目期間が長い，または最近の話題の検索支援語ほど，Z の値が大きくなる．つまり，話題性が高い話題に関する記事を検索可能にする検索支援語を，優先する評価関数である．

3. 実験

提案手法の有用性を確かめる目的で，実験を二つ行った．一つは，テストコレクションを用いた検索実験で，NTCIR-1（情報検索用テストコレクション）[10]を用いた．この実験は，検索支援語を用いることで，正解文書を検索結果の上位に取得することができるかどうかを確認する目的で行った．もう一つは，実際に検索エンジンを用いた検索実験で．この実験は，提案手法を用いて，ホットトピックスに関連する記事を取得することができるかどうかを確認する目的で行った．テストコレクションを用いた実験では，ユーザーが意図した内容の文書を取得可能かどうかを評価する．一方，検索エンジンを用いた実験では，ホットトピックスに関する内容の記事を取得可能かどうかを評価する．上記二つの実験で，良い評価を得ることが出来れば，提案手法は，ユーザーが興味のあるホットトピックスに関連する内容の記事を取得可能な方法論であると考えられる．

3.1. テストコレクションを用いた実験

テストデータと検索質問のみが異なる二つの実験を行った．一つ目の実験では，テストデータとして，文書番号'gakkai-0000000001'から'gakkai-0000002000'までの 2,000 件の文書を用いた．また，検索質問として，'Topic-0053:「電波の人体への影響について」'を用いた．なお，上記のテストデータの中に含まれる'Topic-0053'の正解文書は 1 件であった．

まず前処理として，上記のテストデータから，「電波」,「人体」，もしくは「影響」のいずれかを含んだ文書のみを抽出した．抽出した文書に対して提案手法を適用した結果，最も優先度が高かった検索支援語として,「周波数」,「電磁」,「周波数」,「オシロスコープ」,および「クーロン」が得られた．この際，検索日時を，1988 年 3 月 30 日と仮定し，式**(2)**中の変数 t には，この日付を代入し，減衰定数は a=0.15 とした．また，式**(1)**中のパラメータ s は，s=2 とした．

上記五つの検索支援語と，「電波」,「人体」，および「影響」の，合計 8 つの索引語を，検索質問ベクトルとして扱い，検索質問ベクトルと 2,000 件の全文書に対応する文書ベクトルそれぞれとのコサイン類似度を算出した．その結果，正解文書とのコサイン類似度の値が,その他の文書とのコサイン類似度の値と比べて，最も大きい値を得ることができた．つまり，検索支援語を用いることで，ユーザーは意図した内容の情報を得ることができる．

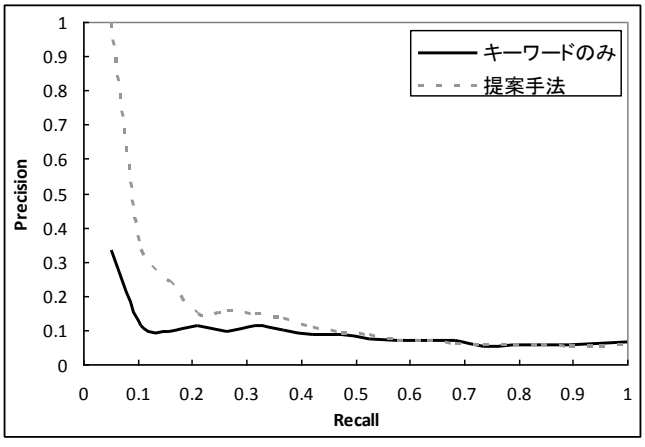

図 1　再現率-適合率曲線

次に，二つ目の実験について述べる．テストデータとして，’gakkai-000020001’から’gakkai-000022001’の文書 ID を持つ 2,000 件の文書を扱った．また，ID として’Topic:0054’をもつ検索質問を採用した．この検索質問の内容は，「光ファイバの通信速度」である．上記のテストデータのうち，19 件の文書が，検索質問として’Topic:0054’を用いた場合の，正解文書である．

前処理として，上記のテストデータから，「光」，「ファイバ」，もしくは「速度」のいずれかを含んだ文書のみを抽出した． 抽出された文書に対して，提案手法を適用した結果，45 個のクラスタを得ることができ，そのうち優先度の値が大きい三つを抽出した．この際，検索日時を，1992 年 5 月 30 日と仮定した．式**(2)**中の変数 *t* には，この日付を代入し，減衰定数は *a*=0.15 とした．また，式**(1)**のパラメータ *s* は，*s*=2 とした．

それぞれのクラスタから，五つの検索支援語と，「光」，「ファイバ」，および「速度」の合計８つの索引語を，検索質問ベクトルとして扱い，検索質問ベクトルと 2,000 件の全文書に対応する文書ベクトルそれぞれとのコサイン類似度を算出した．その結果を基に，再現率[11]と適合率[11]を算出した．

算出された再現率と適合率を用いて，再現率-適合率曲線[11]を描いたものを，図 1 に示す．図の’キーワードのみ’は，提案手法を用いず，「光」，「ファイバ」，および「速度」の合計三つの索引語を，検索質問ベクトルとして扱い，検索質問ベクトルと 2,000 件の全文書に対応する文書ベクトルそれぞれとのコサイン類似度を算出した結果を用いて描いた曲線である．一方’提案手法’は，関連度の値が大きい三つのクラスタより抽出された検索支援語を用いた場合の，結果の平均を用いて描いた曲線である．この曲線は，低い再現率の時に非常に高い適合率を示しているが，再現率が大きくなるに従って，’キーワードのみ’の適合率を下回っている．よって，提案手法が優先して提案した検索支援語を用いることで，一部の正解文書を容易に取得することができると考えられる．二つの実験の結果より，提案手法はユーザーの意図した内容の文書を効率的に取得可能にする手法であるといえる．ただし，ユーザーがトピックグラフを見て興味を持った話題を検索するための検索支援語と，提案手法が優先して提案した検索支援語とが一致したと仮定する．

3.2. 検索エンジンを用いた実験

データソースとして，Yahoo!ニュースで配信されている RSS[12]のうち，Yahoo!トピックストップにおける，2010 年 12 月 1 日から 2011 年 1 月 7 日の 38 日間に配信された 1,461 件の記事を利用した．また，索引語の数は 7,047 個となった．つまり，1,461×7,047 の索引語-文書行列が解析対象のニュース群である．また，検索エンジンは Yahoo!ニュース[13]のものを用いた．実験は 2011 年 1 月 8 日に行い，式**(2)**中の変数 *t* には，この日付を代入し，減衰定数は *a*=0.15 とした．また，式**(1)**のパラメータ *s* は，*s*=2 とした．

以上の条件の下，検索支援語の抽出とそれらの優先度の算出を行った．分割されたクラスタの数は全部で 41 個である．41 個のクラスタうち，高い優先度の値を得た五つをピックアップして，それらのステータスを図 2 および表 1 に示す．

表 1 の，Day1，Day2 ，および word1～5 は，それぞれ，注目開始日，終了日，および検索支援語を示す．ただし，ID=4 の word4,5 は，word1～3 の他に，word4,5 に相当する重みを持った索引語も加えて，AND 検索を行ったところ，検索不能となったため表記していない．また，注目度，減衰度，および優先度は，それぞれ，式**(1)**，式**(2)**，および式**(3)**より算出された値である．図 2 は，トピックグラフである．これは，縦軸が対数目盛の対数グラフであり，log(bursty point)は注目度の値の対数をとった値を指す．横軸は注目期間である．また，図 2 と表 1 における ID は同一のクラスタのことを示している．本研究ではクラスタを話題として扱っているので，図 2 より，ホットトピックスの注目性の高さと，その期間を理解することができ，表 1 の検索支援語より，ホットトピックスの内容を推定することができる．

表 1 の検索支援語を，Yahoo!ニュースの検索エンジンに入力して検索を行ったところ，表 2 に示すような結果が得られた．本実験では，検索目的の記事を，ホットトピックスに関連する記事とした．つまり，表 2 における意図した記事は，注目期間のタイムスタンプを持つ記事を指す．表 2 の適合率は，(意図した記事の数)/(検索数)より算出し，上位 20 は，検索結果の上位 20 件のうちの，意図した記事の件数である．また，再現率の算出は行っていない．なぜなら，再現率の算出に必要な，Yahoo!ニュースにて配信されているすべての記事から，目的の記事の数を正確に調べることは，現実的に不可能なためである．

表 1 検索支援語とパラメータ

ID	Day1	Day2	注目度	減衰度	優先度	word1	word2	word3	word4	word5
1	1/2	1/4	58322	0.78	45421	箱根	駅伝	往路	早大	東洋大
2	12/29	12/31	9258	0.64	5903	NHK	紅白	リハーサル		
3	12/22	12/27	852	0.72	615	選手権	男子	SP	全日本	中京
4	12/22	12/28	431	0.77	333	麻木	山路	タレント	大桃	久仁子
5	12/25	12/27	69	0.52	36	ワゴン	事故	池	太宰府	福岡

表 2 検索結果

ID	検索数	意図した記事	適合率	上位20
1	79	65	0.82	15
2	133	125	0.94	12
3	17	15	0.88	15
4	75	33	0.44	0
5	9	8	0.89	8

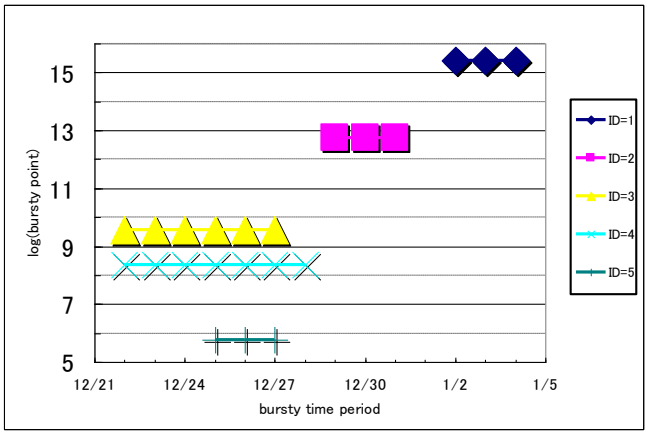

図 2 トピックグラフ

提案手法の有用性を示すために，提案手法の使用例を示す．最も優先度の値が大きかった ID=1 のホットトピックスが，ユーザーの興味を惹きつけたと仮定する．そこで，ID=1 のホットトピックスに関する内容の記事を目的に，ID=1 の検索支援語を用いて AND 検索を行ったところ，検索結果として得た約 80 件の記事の内，約 60 件が目的の記事であった．また，検索結果の上位 20 件の記事のうち，15 件が目的の記事であった．これらの結果は，検索結果として得られた記事の中から，目的の記事を容易に探し出せる量と考えられる．よって，提案手法を用いることで，ユーザーが興味を持ったホットトピックスに関する内容の記事を容易に取得することができる．

表 2 より，ID=4 を除く，ID=1~3,5 の四つの検索支援語を用いた場合，適合率が 0.8 以上であることがわかる．各検索支援語を用いた場合の適合率の平均を，(ID=1~5 の意図した記事の数)/(ID=1~5 の検索数)，より算出したところ，0.79 を得た．これらの値より，高い正確性を実現できたと評価できる．また，検索結果の上位 20 件中に含まれる正解文書の数の平均を，(ID=1~5 の上位 20)/5 より算出したところ，10 を得た．この値より，検索結果の上位にユーザーの意図した内容の記事を得ることができたと考えられる．以上より，提案手法を用いることで，ホットトピックスに関する内容の記事を取得できることがわかる．

4. まとめ

本研究では，Web ニュース検索における，ホットトピックスに関連する記事を効率的に取得する手法を提案した．提案手法は，１．注目されている話題とその注目されている期間に関する理解を補助するグラフ，２．検索の際に用いることで，それらの話題を容易に取得可能にする単語，の二つを出力する．１.より，特定の話題に興味を惹かれたなら，２.より，その話題を検索可能にする単語を用いて検索を行うことで，その話題に関連する記事を容易に取得できる．提案手法の有用性を評価するために，テストコレクションを用いた検索実験と検索エンジンを用いた検索実験の二つの実験を行ったところ，興味のあるホットトピックスに関する記事を容易に取得できることを確認できた．

謝辞

本研究の実験に用いたテストコレクションを提供していただいた国立情報学研究所様に感謝致します．

参 考 文 献

[1] Takakazu TSUBOKAWA and Hiroyuki GOTO, "A Method of Extracting Assistive Words for Web News Search with the Concept of Time Axis", Proc. The 11th Asia Pacific Industrial Engineering and Management Systems Conference (APIEMS '10), In

Melaka, Paper ID 18 (6 pages), (2010).

[2] 今井良太，戸田浩之，関口裕一郎，望月崇由，鈴木智也，今井桂子，“Web 検索サービスにおける多義的なクエリ推薦手法”，DBSJ Journal,Vol.9,No.1,pp.7−11,(2010).

[3] Salton, G., Wong, A., and Yang, C.S. ”A vector space model for automatic indexing”, Communications of the ACM, Vol. 18, No. 11, pp.613–620,(1975).

[4] T. Kudou, “ MeCab : Yet another part-of-speech and morphological analyzer. ” , http://mecab.sourceforge.net , (2009).

[5] Luhn, H. P. ”Statistical approach to mechanized encoding and searching of literary information.”, IBM Journal of Research and Development, Vol.1, No.4, pp.309–317, (1957).

[6] D.Boley ,”Principal direction divisive partitioning”, Data Mining and Knowledge Discovery,Vol.2, No.4, pp.325−344,(1998).

[7] Jones, K. S. ”A statistical interpretation of term specificity and its application in retrieval.“, Journal of Documentation, Vol.28, No.1, pp.11–21, (1972).

[8] J. Kleinberg, ” Bursty and hierarchical structure in streams”, Proc. The 8th annual international ACM Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD’02) International Conference, pp.91–101, (2002).

[9] Yoshiharu Ishikawa, Yibing Chen, and Hiroyuki Kitagawa, “An on-line document clustering method based on forgetting factors”, Proc. the 5th European Conf. on Research and Advanced Tech. for Digital Libraries (ECDL 2001), pp.325−339, (2001).

[10] 神門 典子,栗山 和子,野末 俊比古,大山 敬三,“NTCIR-1 : 情報検索システム評価用テストコレクション構築の方針と実際”,情報処理学会研究報告. 情報学基礎研究会報告 99(20), pp.33−40, (1999).

[11] R. Baeza-Yates and B. Ribeiro-Neto, ”Modern Information Retrieval”, Addison Wesley, (1999).

[12] Yahoo! ニュース – RSS , http://public.news.yahoo.co.jp/rss/, [accessed 2011/01/24].

[13] Yahoo!Japan ニュース, http://headlines.yahoo.co.jp/, [accessed 2011/01/24].