階層型クラスタリングを用いた時系列テキスト集合からの話題推移抽出

Extraction of Topic Transition through Time Series Document based on Hierarchical Clustering

菊池 匡晃 ♥ 岡本 昌之 ♦
山崎 智弘 ♠

Masaaki KIKUCHI Masayuki OKAMOTO
Tomohiro YAMASAKI

ニュース記事などの時系列テキストを対象に，話題の推移を表現したキーワード群を抽出する手法を提案する．本手法は，階層型クラスタリングと複合語判定のための C-value 法に基づいて話題の推移を抽出することが特徴である．提案手法を電子番組表 (EPG) に適用した際の話題把握に関する被験者評価を行った結果，32 日間で抽出された 640 件の話題のうち 94.3 %の話題を把握でき，68.6 %の話題については推移まで把握できることがわかり，本手法による話題把握支援の有用性が確認された．

We propose a new method for extracting keywords that express topic transition through time series document analysis. Our method is based on hierarchical clustering and C-value method that are used for compound word detection. In our experiments, we used 640 topics that had been extracted for 32 days, and the users correctly understood 94.3% of the topics and 68.6% of the topic transitions.

1. はじめに

インターネットの発展に代表される情報発信メディアの多様化に伴い，我々が入手可能な情報は日々増大している．そのような中であらゆる情報をチェックして最新の話題を知ることは困難であり，世間の関心を集めている話題を簡単に知りたい，最新の話題をまとめて知ることができる情報源が欲しいと言ったニーズが高まっている．そのような中，世間で話題となっているトピックをキーワードで表現しユーザに提示するサービスが増えつつある．例えばブログ上にある膨大な情報を集合知としてとらえ，頻出するキーワードを解析することで世の中の関心やニュースなどへの反響を探ろうとする試みがある [1]．一方世の中の関心事を表す時事キーワードを抽出するという意味では，検索サイトやニュースサイトなどで特定の期間に検索に利用されたキーワードをジャンルごとに集計するといった手法も広く行なわれている [2][3]．このような手法によって世の中で話題になっているキーワードを知ることはできるが，そのキーワードがなぜ話題になっているのかといった経緯まではわからないことが多く，ユーザ自身がキーワードについて調べる必要がある．

そこで本稿では話題把握支援のアプローチとして，話題を表すキーワード群を話題の推移グラフと合わせて提示する手法を提案する．本手法の特徴は，一連の話題の中から短期間のイベントを抽出し，話題の時系列推移グラフ上にイベントに関するキーワードを配置することにより話題推移の把握を支援すること，および話題の時事性が高い順に提示することで効率良く世の中の話題を知ることができることである．時系列テキストとしてはニュース記事やブログ記事などがあるが，我々はテレビ番組表に着目し，話題を抽出する対象として用いることとした．電子番組ガイド (EPG: Electronic Program Guide) はニュースなどの話題や番組内容が簡潔にまとめられており，広くテレビ視聴者全体を対象とした情報であるため，ブログなどよりも一般的な話題を抽出できることが期待される．本稿では EPG を対象に話題推移を抽出する手法を提案する．また提案手法を用いて実験を行い，被験者による話題把握に関する評価を行った結果について報告する．

以下 2. 節で提案手法の詳細について説明し，3. 節で実験を行った結果について，4. 節で提案手法により抽出された話題の被験者評価を行った結果について，5. 節では関連研究について述べる．

2. 話題推移抽出手法

本手法は時系列テキスト集合を入力とし，話題ごとへのクラスタリング，および話題クラスタからのキーワード，イベントを抽出する．各話題をキーワード群と話題の推移を表すグラフによって表し，時事性の高い話題順に提示することで話題把握を支援する．本手法の処理の流れを図 1 に示す．

時系列テキスト集合として EPG を用いる場合の話題推移抽出は以下の手順で行われる．

1 N 日分の EPG の文書を形態素解析し単語ベクトルを作成する．単語ベクトルは過去の EPG 文書集合からあらかじめ求められた idf 値によって重み付けされる．
2 各文書を 1 つのクラスタとみなして凝集型の階層型クラスタリングを適用する．クラスタ間類似度の最大値が閾値以下になるまでクラスタリングを行う．
3 抽出された話題クラスタから C-value 法を用いてキーワード群を抽出する．
4 話題クラスタからイベントを抽出するために生起時刻の近い文書が同一のクラスタを形成しやすいよう類似度計算式を重み付けした上で，各話題クラスタを手順 3 と同様に再度クラスタリングする．この際にクラスタに含まれる文書集合から求めた idf 値を用いて単語ベクトルをさらに重み付けする．
5 抽出された話題内のイベントクラスタから手順 4 と同様に

♥ (株) 東芝 masaaki11.kikuchi@toshiba.co.jp
♦ (株) 東芝 masayuki4.okamoto@toshiba.co.jp
♠ (株) 東芝 tomohiro2.yamasaki@toshiba.co.jp

イベントキーワード群を抽出する．

6 文書数の時系列推移を基に Z 検定によって話題の時事性を検定する．時事性の高い話題順にソートし，話題を表すキーワード群および話題の推移グラフを出力する．

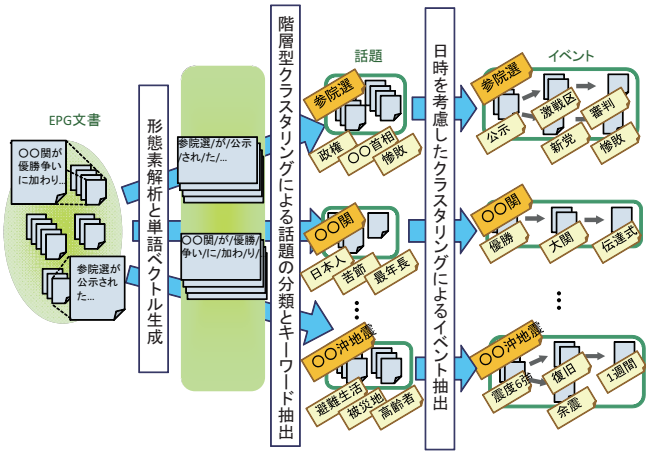


図 1　提案手法の処理の流れ

Fig. 1　An overview of extracting topic transition

2.1　クラスタリングによる話題抽出

文書間の類似度を計算する手法として，文書の単語ベクトルの余弦尺度を用いる．文書 a と文書 b の余弦尺度 s は以下の式で表される．

$$s(\mathbf{x}_a, \mathbf{x}_b) = \frac{\mathbf{x}_a \cdot \mathbf{x}_b}{\|\mathbf{x}_a\| \cdot \|\mathbf{x}_b\|}$$

ただし $\mathbf{x}$ は各文書の単語ベクトルである．単語ベクトルには，以下の式で表される idf で各単語を重み付けしたものを用いる．

$$\mathrm{idf(w)} = \log_{10} \frac{N}{\mathrm{df(w)}} + 1$$

ただし

N : 文書の総数

$\mathrm{df(w)}$: 単語 w が出現する文書数

よって文書 a の単語ベクトルは以下のように表される．

$$\mathbf{x}_a = \begin{pmatrix} \mathrm{idf(w_1)} \\ \mathrm{idf(w_2)} \\ \vdots \\ \mathrm{idf(w_n)} \end{pmatrix}$$

ただし，$\mathrm{w_1} \ldots \mathrm{w_n}$ は文書 a に含まれる単語群である．

上記の単語ベクトルとして表された文書群を凝集型の階層型クラスタリングによって分類し話題を抽出する．単語ベクトルの類似度 s が最大となるクラスタを再帰的に併合し，s の最大値が閾値 θ_T 以下になるまで繰り返す．併合された際の単語ベクトルは以下の式に従って更新される．

$$\mathbf{x}_{ab} = \mathbf{x}_a + \mathbf{x}_b$$

2.2　キーワード抽出

話題ごとに分類された文書の集合から話題を代表するキーワードを C-value 法 [4] を用いて抽出する．複合語 cw の C-value は以下の式で定義される．

$$\text{C-value(cw)} = (\text{length(cw)} - 1)\left(n(\text{cw}) - \frac{t(\text{cw})}{c(\text{cw})}\right)$$

ここでパラメータは以下の通りである．

$n(\text{cw})$: 複合語 cwの出現頻度

$t(\text{cw})$: 複合語 cwを含むより長い複合語の出現頻度

$c(\text{cw})$: 複合語 cwを含むより長い複合語の異なり数

length(cw) : 複合語 cwの文字列長

C-value の本来の定義では length(cw) は複合語 cw の形態素数であるが，この定義によると $\text{length(cw)} = 1$，すなわち cw が単形態素の場合に C-value が 0 になるため，length(cw) を文字列長として定義し直している．これによって $n(\text{cw})$ が大きくなりやすい短いワードの C-value が不当に高くなることを抑制できるが，逆に長いワードほど C-value が高くなるという特徴がある．このような特徴を補正した複合語抽出手法として C'-value が提案されている [5]．文書集合内から形態素 N-gram による複合語候補を作成し，それぞれの C-value を計算する．包含関係にある複合語候補から C-value の低い方を除去し，残った複合語の中で C-value が最大のキーワードを話題の代表キーワード，それ以外をサブキーワードとする．

2.3　イベント抽出

話題ごとに分類したときと同様に，話題クラスタ内部で再度クラスタリングを行うことで話題内の短期的なイベントを抽出する．クラスタ間類似度の計算時にクラスタ内文書の平均生起時刻の近さによって重み付けを行う．クラスタ a とクラスタ b の類似度の重み付けには以下の減衰関数を用いる．

$$W(a,b) = \exp\left(-\alpha(t_a - t_b)^2\right)$$

ただしパラメータは以下の通りである．

t_a, t_b : クラスタ a, b の平均生起時刻

α : 定数

$W(a,b)$ によって生起時刻の近い記事ほど同じクラスタに分類されやすくなる．最終的にクラスタ a と b の類似度 $\hat{s}$ は

$$\hat{s}(\mathbf{x}_a, \mathbf{x}_b) = s(\mathbf{x}_a, \mathbf{x}_b) \cdot W(a,b)$$

と定義する．

また，クラスタ内部に含まれる文書集合における各単語の idf を計算し，その値で単語ベクトルをさらに重み付けしてクラスタリングに用いる．これはクラスタ内部の idf を用いることで，話題の主題となる単語の重み付けが小さくなり，主題以外の単語の重み付けが相対的に大きくなるため，再度クラスタリングするこ

とにより話題内のイベントが抽出されることが期待できるためである．重み付けされた単語 w の単語ベクトル成分 x_{w} は，単語 w のクラスタ内部の idf 値を $\mathrm{idf}_L(\mathrm{w})$ として

$$\hat{x}_{\mathrm{w}} = x_{\mathrm{w}} \cdot \mathrm{idf}_L(\mathrm{w})$$

とした．ここでパラメータは以下の通りである．

$$\mathrm{idf}_L(\mathrm{w}) = \log_{10} \frac{N_L}{\mathrm{df}_L(\mathrm{w})} + 1$$

N_L : クラスタ内の文書の総数

$\mathrm{df}_L(\mathrm{w})$: クラスタ内で単語 w が出現する文書数

以上の類似度，単語ベクトルに基づき，類似度 $\hat{s}$ の最大値が閾値 θ_E に達するまで話題クラスタを 2.1 節と同様の手順で再帰的にクラスタリングすることでイベントを表すクラスタを抽出する．このイベントクラスタから 2.2 節で述べた C-value を用いてキーワードを抽出する．イベントクラスタ内部の文書のうち生起時刻の最も早い時刻とキーワードをマッピングし，日時情報をキーワードに付与する．これらのイベントキーワード群を，話題クラスタ内の文書数の推移と重畳して表示することで推移グラフを作成する．

2.4 時事性判定

時事性のある話題の場合，たとえば最近 7 日間の方が最近 28 日間よりも 1 日あたりの出現頻度は上昇するはずである．そのため時事性を評価するためには時系列にそった出現頻度の分布から短期的な出現頻度が長期的な出現頻度よりも有意に上昇していることを判定する必要がある．そこで長期的な出現確率は一様分布に従っていると仮定し,「短期的な出現確率も平均が同じ一様分布に従っている」という帰無仮説の検定を行なうことで話題の時事性を判定する．仮説検定を行う方式として Z 検定を用いる．

文書の出現確率が一様分布に従っている場合，最近 N 日の出現文書数を u とすると最近 n 日の出現文書数 v の確率分布は，生起確率 $p = n/N$，試行回数 u の二項分布に従うと考えられる．生起確率 p，試行回数 u の二項分布の確率関数 $Pr(v) = {}_uC_v pv(1-p)^{u-v}$ は平均 up，分散 $up(1-p)$ の正規分布で近似できることが知られているので，最近 n 日の出現文書数の観測値が $v0$ であったとき Z の値は

$$Z = \frac{(v0 - up)}{\sqrt{(up(1-p))}}$$

と計算される．検定表より Z の値が 1.96 より大きいとき 5 %の有意水準で，2.58 より大きいとき 1 %の有意水準で帰無仮説を棄却することができ，Z 値が大きいほど時事性が高いと考えることができる．話題クラスタを Z 値の降順にソートすることで時事性の高い話題順に並び替える．

3. 実験

2007 年 7 月 23 日から 8 月 23 日までの 32 日間を対象として 1 日ごとに EPG から話題を抽出する実験を行った．短期的な話題と長期的な話題の 2 種類を扱うために，それぞれ当日までの 7 日間，28 日間分の EPG を入力として用いた．また時事の話題を抽出することが目的であるため，番組ジャンルがニュース，ワイドショーに該当する EPG 文書をデータセットとし，Z 検定のパラメータは短期については $N = 7, n = 3$，長期については $N = 28, n = 14$ を，その他の各パラメータとして $\theta_T = 0.22, \theta_E = 0.1, \alpha = 1$ を用いた．

実験結果の例として 2007 年 7 月 23 日の出力結果を図 2 に示す．図 2(a) は短期，図 2(b) は長期の抽出結果を示し，それぞれ Z 値の上位 3 件の話題が示されている．それぞれの話題は代表キーワード 1 個と 5 個以内のサブキーワードの計 6 個までのキーワード群と，話題クラスタの記事数推移グラフにイベントキーワードを重畳表示した推移グラフによって表される．次節で抽出された話題の被験者評価を行った結果について述べる．

4. 被験者評価

抽出されたキーワード群および推移グラフから話題を把握できるのか，およびキーワード群が時事の話題を適切に表すキーワードによって構成されているかについて被験者による評価を行った．評価は 2007 年 7 月 23 日から 2007 年 8 月 23 日の 32 日間における短期 (7 日間) および長期 (28 日間) の両期間で上位の話題を 1 日につきそれぞれ 10 件，合計 640 個の話題を対象とし，6 人の被験者が 1 日あたり 3〜4 人ずつ出力結果を確認した．抽出されたそれぞれの話題は 1〜6 個，平均して 5.8 個のキーワード群により構成され，32 日間でのべ 3761 個のキーワードが抽出された．

4.1 話題把握の評価

話題の把握の評価として以下の 3 種類の評価を行った．

(1) キーワード群から話題が想起できるか
話題を想起できるかどうかに応じて以下の 3 段階で回答.
・キーワード群から知っている話題を想起できる
・ニュースだと理解できるがニュース自体は知らない
・明らかにニュースとして意味をなさない

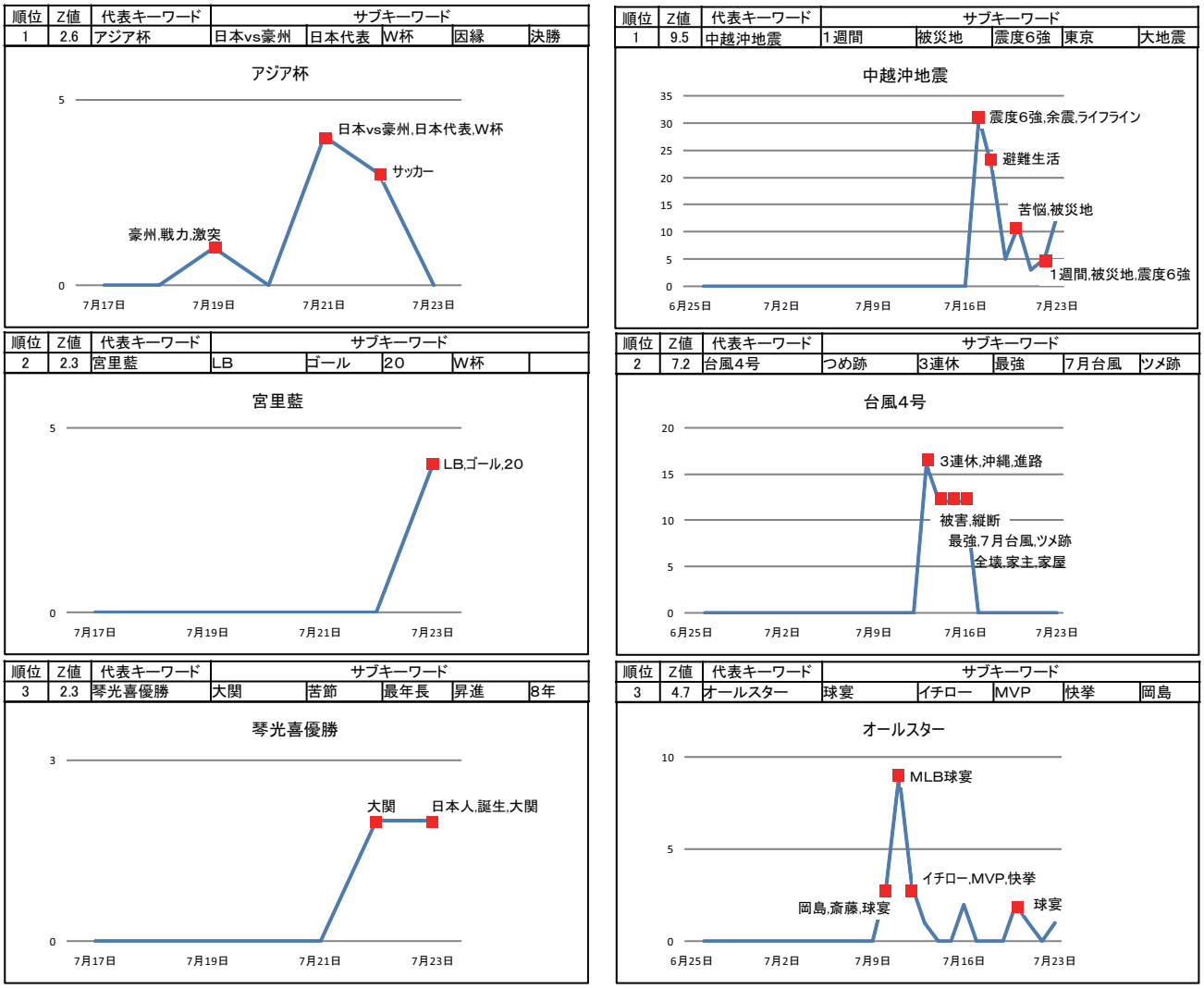
(2) 話題の推移が把握できるか
推移グラフを見た時に話題の推移・流れまで分かるかどうかで以下の 3 段階で回答.
・話題が推移も含めて分かる
・推移は分からないが話題は分かる
・話題も分からない

(3) キーワードを検索に用いて関連 Web ページに到達できるか
見出し，サブ見出しによる Yahoo！ JAPAN の Web ページ検索を行い，関連する話題が何位に出現したか回答.

評価にあたり我々は以下のような仮説を立てた．被験者による評価結果を受けてこれらの仮説の検証を行う．

(a) キーワード群の提示だけでもユーザは十分話題を把握することができる.

(b) 長期の話題でイベントが多いほど，より話題の推移を把握することができる.



(a) 短期 (7 日間) (b) 長期 (28 日間)

図 2 抽出された話題例 (2007/07/23)

Fig. 2 An example of extracted topics(2007/07/23)

(c) キーワード検索については，代表キーワードだけでは他の話題やニュース記事が多数検索されるが，サブキーワードを加えることにより，より適切な話題が上位に来る．

4.2 話題把握の評価結果と考察

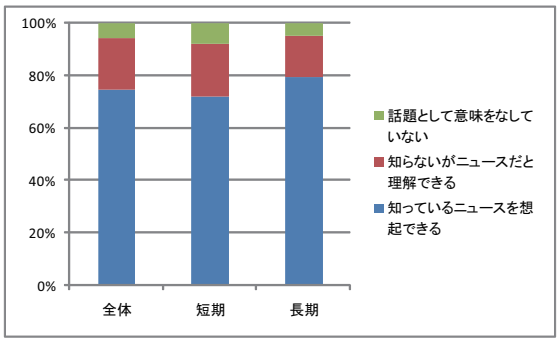
評価結果を図 3 に示す．まず，キーワード群から時事の話題を想起できるか調べた結果，全体として「時事の話題を想起できる」が 74.7 %,「知らないが話題であると分かる」が 19.6 %と，合計 94.3 %の話題はユーザが理解可能な程度にまとまっていることが分かった（図 3(a)）．

次に，それぞれの話題の推移グラフを見た時に，話題の推移まで分かるか調査した結果，全体として 67.4 %の話題については推移まで把握できることが分かった．特に，イベントが 4 個以上の話題 244 個に関しては 68.6 %の話題に対して推移まで把握でき，推移が分からない場合も含め 95.8 %の話題は把握できることが分かった（図 3(b)）．

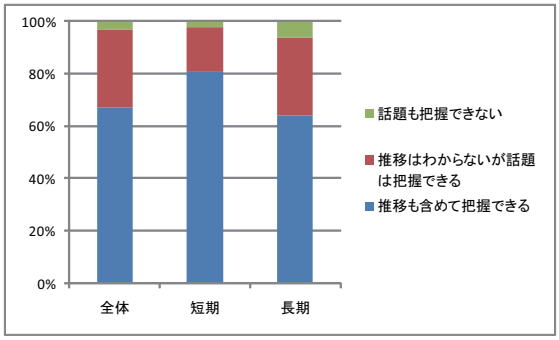
また，それぞれの代表キーワード 640 個とサブキーワード 3,120 個について，当日の午前 11 時に取得した 10 位以内の検索結果に適切な話題が含まれるか調査した結果，代表キーワードだけで検索する場合は 79.0 %，サブキーワードを含めた検索では 87.4 %と，サブキーワードを用いることでその話題の特定に約 8.4 %の効果があることが分かった（図 3(c)）．以上の結果より，提案手法による話題提示は，キーワードや推移グラフの提示により，時事の話題やその推移の理解に寄与していると言える．

これらの結果より，4.1 節で挙げた仮説に対する検証結果は以下の通りであると考えられる．

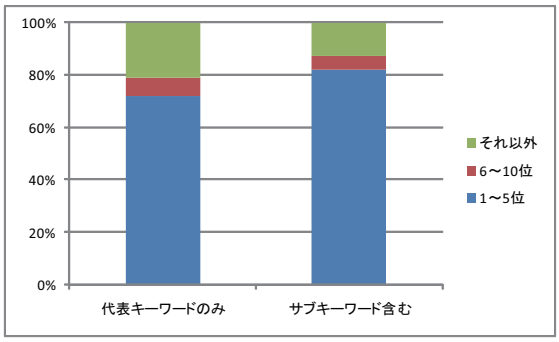
(a) キーワード群の提示による話題把握に関して，94.3 %の話



(a) 話題の想起



(b) 推移の把握 (イベントが 4 個以上の話題)



(c) 関連 Web ページ検索 (短期・長期の平均)

図 3　話題把握の評価結果

Fig. 3　A result of evaluation about topic understanding

題を一つのまとまりとして把握できたので，ユーザは十分話題を把握することができているものと考えられる．

(b) 話題推移の把握に関しては，全体平均と比べ，サブキーワードが多い方が若干把握できる割合が増加したが，その差はあまり大きくない．また，当初は長期の話題の方が出来事のフェーズに応じたイベント抽出が行われると考えていたが，被験者評価の結果では，短期の方がより把握しやすいことが分かった．これは，短期間でもイベントとして分割できる話題が，長期間でひとまとまりになった結果逆に把握しづらくなっているものと考えられ，今後の検討課題である．

(c) キーワード検索では，サブキーワードを加えることにより 5 位以内正解率が約 9.6 %，10 位以内正解率が約 8.4 %改善しており，サブキーワードにより適切な絞り込み検索の効果があることが分かる．

4.3　話題構成の評価

次に，各話題のキーワード群，話題クラスタの粒度が適切に構成されているかの評価として以下の 2 種類の評価を行った．

(1) 各キーワードは話題を表す言葉として適切か
　話題を表す各キーワード見たときに同じ話題のキーワードとして適切かどうかで以下の 3 段階で回答．
　・同じ話題として適切
　・別の話題のキーワードである
　・話題として意味が通らない

(2) 話題の粒度は適切か
　各話題について，同じ内容だと思う他の話題を回答．

4.4　話題構成の評価結果と考察

評価結果を表 1 に示す．表 1 より全てのキーワードが適切に話題を表している割合が 69.4 %，8 割以上のキーワードが適切に話題を表している割合が 96.4 %であることがわかる．1 つの話題には平均して 5.8 個のキーワードが含まれるため，8 割以上が適切であるということは，適切でないキーワードが 1 つの話題に平均 1 つ以下しか含まれないことに相当する．意味が通らないと評価されたキーワードは，大きく

- EPG に頻出する表現由来のキーワード
- 番組の出演者と思われるキーワード
- 話題とは言えないクラスタ由来のキーワード
- カタカナ語の一部など区切りが不適切なキーワード

に大きく分けることができる．これらは区切りが不適切な場合を除いて，入力として用いた EPG の特徴によるものと言える．このような EPG 特有の問題に対しては，EPG 頻出表現をストップワードとして指定する，EPG の出演者情報に含まれる文字列をフィルタする，話題抽出に用いる情報源のジャンルを絞る，などの改善策が考えられる．区切りが不適切であるパターンについては形態素解析を失敗していると考えられ，こういった場合への対処は今後の課題である．

また，1 日あたりに提示される 20 個の話題の中で重複を除いていくつの話題を出力できているのかを調査した結果，平均して 1 日あたり 15.2 個の異なる内容の話題が出力されており，重複を含まない話題の割合は全体の 76.4 %であることがわかった．今回の評価期間において重複していると評価された話題の多くは「選挙」「地震」「大臣更迭」などに関する話題であった．選挙などの長期間で大きな話題は内部に様々な話題があるため，例えば選挙準備段階の話題，選挙中の出口調査の話題，選挙後の当落の話題などに分割されて出力されているが，被験者から見ると全て「選挙」の話題であると判断されたと考えられる．話題把握の観点から，どの程度の話題の粒度が適切であるのかを調査，考察することは今後の課題である．

表 1　適切なキーワードから構成される話題の割合
Table 1　The rate of topics constructed by suitable keywords

期間	全話題数	適切なキーワードのみの話題数	キーワードの 8 割以上が適切な話題数
短期	320	221 (69.1 %)	309 (96.6 %)
長期	320	223 (69.7 %)	308 (96.3 %)
全体	640	444 (69.4 %)	617 (96.4 %)

5. 関連研究

時系列テキスト集合からの話題抽出手法としては，対象の出現間隔を利用して burst を検出する手法が挙げられ [6]，ブログ記事などに適用する研究も行われている [7]．ブログなどの記事は，あるイベントによって記事の総数自体も増加する性質があり，burst 検出による手法はこの性質を利用したものであると言えるが，EPG のように記事数の上限があらかじめ決められている情報源からの話題抽出には適用が難しいと考えられる．またブログのトラックバックによる記事間のネットワーク構造を利用した話題抽出の研究も行われている [8]．これは記事ネットワークの時系列成長により盛り上がっているトピックを可視化するものであるが，話題把握の観点では課題があると考えられる．

話題の推移を抽出する研究としては，トピック追跡に関する研究 [9] など様々な取り組みが行われている．文献 [10] による手法はブログを対象として，クラスタリングにより話題推移抽出を行う点で本手法と類似しているが，関連性が高い複数のトピックの話題変遷パターンを比較することを目的としており，話題把握支援を目的とする本手法とは用途が異なる．

また，本手法ではキーワード抽出に C-value 法 [4] を用いたが，専門用語抽出において，より適合率に優れる手法も提案されており [11]，精度向上に向けて比較検討していく必要がある．

6. おわりに

本稿では時系列テキストから話題推移を抽出する手法を提案し，本手法を EPG 文書に適用した際の被験者評価を通じた話題抽出の有効性について報告した．評価結果から，94.3 %の話題がユーザが理解可能な程度にまとまっており，推移グラフに関しても一定の割合で推移まで把握でき，推移が分からない場合も含め 95.8 %の話題はグラフを用いて把握できることが分かった．検索精度に関しても 87.4 %のサブ見出しは 1 ページ目に収まる 10 位以内に関連話題を検索できることが分かった．また抽出された話題のうち，全てのキーワードが適切に話題を表している割合が 69.4 %，8 割以上のキーワードが適切に話題を表している割合が 96.4 %であることが確認された．

以上の評価結果から，本手法により世の中の話題把握支援が有効に行われていると考えられる．しかしながら，他手法との比較検討は未実施であり今後の課題である．

今後は話題推移抽出のさらなる精度向上を目指すとともに，ニュース記事やブログ記事など EPG 以外の対象への適用，他手法との比較評価を進める予定である．

[文献]

[1] http://kizasi.jp
[2] http://searchranking.yahoo.co.jp/
[3] http://ranking.goo.ne.jp/keyword/
[4] Frantsi, K. and Ananiadou, S. “Extracting Nested Collocations.”, COLING 96, pp.41-46, 1996.
[5] 山崎智弘, “強連結成分分解を利用した電子番組表からの話題抽出”, DEWS2008, A1-Web, 2008.
[6] J. Kleinberg, “Bursty and Hierarchical Structure in Streams”, In Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002.
[7] T. Fujiki, T. Nanno, Y. Suzuki, M. Okumura, “Identification of Bursts in a Document Stream”, First International Workshop on Knowledge Discovery in Data Streams (in conjunction with ECML/PKDD 2004), 2004.
[8] M. Uchida, N. Shibata, “Extracting and Visuzalization of a Emerging Topic from the Blogspace”, Proceedings of the Annual Conference on JSAI (CD-ROM), Vol.20, pp. 3D2-3, 2006.
[9] NIST. Topic Detection and Tracking (TDT), http://www.nist.gov/speech/tests/tdt/
[10] 戸田 智子, 福田 直樹, 石川 博, “Blog 記事のクラスタリングに基づいたカテゴリ別話題変遷パタンの抽出”, DEWS2007, A8-Blog, 2007.
[11] H. Nakagawa, T. Mori, “A Simple but Powerful Automatic Term Extraction Method”, COLING-02 on COMPUTERM 2002, pp. 1-7, Morristown, NJ, USA, 2002.

菊池 匡晃　Masaaki KIKUCHI
(株) 東芝研究開発センター知識メディアラボラトリー所属．2006 年大阪大学大学院工学研究科修士課程修了．主にコンテクストアウェア技術および情報抽出の研究開発に従事．

岡本 昌之　Masayuki OKAMOTO
（株）東芝 研究開発センター知識メディアラボラトリー研究主務．2003 年京都大学大学院情報学研究科博士後期課程修了．博士（情報学）．主にコンテクストアウェア技術および情報抽出の研究開発に従事．情報処理学会，人工知能学会，ACM 各会員．

山崎 智弘　Tomohiro YAMASAKI
（株）東芝 研究開発センター知識メディアラボラトリー所属．2002 年東京大学大学院理学系研究科修士課程修了．テキストからのキーワード抽出および話題抽出の研究開発に従事．