

強連結成分分解を利用した電子番組表からの話題抽出

山崎 智弘[†]

[†] (株) 東芝 研究開発センター 知識メディアラボラトリー
〒 212-8582 川崎市幸区小向東芝町 1
E-mail: †tomohiro2.yamasaki@toshiba.co.jp

あらまし 一般に世の中の全ての事象をチェックすることは不可能だが、世の中の関心の移り変わりを簡単に知りたいという欲求がある。そこで我々は電子番組表から「時事性」に着目して話題を抽出する手法を開発した。本手法は電子番組表におけるキーワードの共起グラフを強連結成分分解することによって実現されている。本稿では本手法を説明するとともに、実験で得られた知見と今後改良すべき点について考察を加える。

キーワード キーワード抽出, 話題抽出, 共起グラフ, 強連結成分分解

Topic extraction from Electronic Program Guides by using decomposition of the co-occurrence graph into strongly connected components

Tomohiro YAMASAKI[†]

[†] Corporate Research & Development Center, Toshiba Corporation
1. Komukai Toshiba-cho, Saiwai-ku, Kawasaki, 212-8582, Japan
E-mail: †tomohiro2.yamasaki@toshiba.co.jp

Abstract Generally speaking, it is impossible to find every sort of news, but we want to know changes of the concerns of the world easily. Then we focused the viewpoint on “the relation to current events”, and developed a method to extract topics from Electronic Program Guides (EPGs). In this method, we calculate the contexts where keywords occur and decompose the co-occurrence graph of keywords into strongly connected components. In this paper, we describe our method and the results of experiment. Moreover, we consider the findings obtained from the experiment and the points that should be improved in future.

Key words topic extraction, co-occurrence graph, decomposition of the strongly connected components

1. はじめに

インターネットの普及に伴い、我々が手に入れることができる情報は刻々と増え続けている。しかし世の中の話題を知るためにあらゆるニュースをチェックすることは困難であり、

- 世の中の関心の移り変わりを簡単に知りたい
- ニュースや関連情報などを話題ごとにまとめたインデックスがほしい

といったニーズが高まっている。そのようなニーズに応えるべく我々が取り組んでいる技術がテキストからのキーワード抽出・話題抽出技術である。

世の中の話題を幅広く抽出するという観点からは、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し、話題として提示するといった手法が行なわれている [3], [5]。ただし単純な検索回数によるランキングでは “Yahoo”, “mixi”, “Google” のように常に上位

であるためユーザにとってはほとんど情報量がないキーワードまで抽出されてしまうという問題がある。

その一方で web 上にある膨大な情報を集合知として捉え、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も行なわれている。中でもきざし [6] では、ブログ上のキーワードの出現頻度の時系列を約 10 分ごとに解析することで話題を構成するキーワード群の抽出を実現している。ただし最近では、話題のキーワードを含んだブログ記事を手動ないし自動で大量に生成することでトップページのリストを乗っ取るスパマーが出現しており、情報源の信頼性を確保するためにさまざまな対策が講じられていることも知られている。

話題という単語を「話の主題となる事柄」として捉えた場合、一つの事柄でもさまざまな切り口によってさまざまな主題となりえる。そこで今回我々は「時事性」(流行や現実の出来事との関連性) という切り口に注目することにより、電子番組表

(Electronic Program Guide, EPG) のテキストから話題の時間的な継続性や変化の特徴を捉えるキーワード (時事キーワード) を抽出する技術を開発した. 中でも情報源として EPG のテキストに着目した理由は,

- 記述されたテキストは放送局が多く話題からあらかじめ取捨選択した結果を表している
- 記述できる文字数が制限されているため話題ごとの非常に良い要約となっている

と考えられるためである. [6] に見られるように一般的な web 上の文書を情報源とする限り信頼性の確保は大きな課題であるが, 我々のアプローチではクリアされると考えられる. また話題が多岐にわたる web 上の情報を情報源とするよりも精度よくキーワードの抽出, 話題の分類が行なえるものと期待される.

本稿では EPG のテキストから話題の時間的な継続性や変化の特徴を捉えるキーワードを抽出する技術について, ならびに抽出したキーワードを話題ごとに分類する技術について実験・開発した結果を報告する.

2. 電子番組表 (EPG)

電子番組表 (EPG) とは放送番組表をテレビなどの画面に表示するシステムである. EPG のフォーマットは方式によって多少異なるものの, 放送局名, 放送日時, 番組名, 番組内容といった情報が記載されている. 図 1 は SONY が開発したテレビ番組録画予約方式である iEPG の例である. 2007/08/13 に放送された, 女子サッカー北京五輪アジア地区の最終予選である日本対タイの試合を表している.

```
Content-type: application/x-tv-program-info;
charset=shift-jis
version: 1
station: 日本テレビ
year: 2007
month: 08
date: 13
start: 01:50
end: 03:20
program-title: サッカー女子・北京五輪アジア地区最終予選
「日本×タイ」
program-subtitle:
description: ~ 国立競技場 (録画)
performer: [解] 大竹奈美 [解] 川上直子 [実] 右松健太

サッカー女子・北京五輪アジア地区最終予選「日本×タイ」
```

図 1 iEPG の例

図 1 では番組のジャンル情報が記載されていないが, テレビ朝日が開発した EPG 送信サービスである ADAMS-EPG では 60 種類ほどにわけられたジャンル情報が記載されている. 近年の多くのデジタルテレビや HDD レコーダは EPG を検索する機能を持っているため, 番組名や出演者名, ジャンル情報などによって EPG を検索することで視聴したい番組を手軽に検索することができる.

表 1 は我々の研究所がある横浜・川崎エリアで受信可能な放

表 1 横浜・川崎エリアで放送された番組数

ジャンル	番組数
ニュース・ワイドショー	8464
音楽・バラエティー	5535
スポーツ	1864
映画・ドラマ・アニメ	1168
ドキュメンタリー, 教養, その他	2815

送局で半年間に放送された番組数を, 大まかなジャンルごとに集計したものである. 時事キーワードを抽出するという観点からは, ニュース・ワイドショーのような時事的な話題を扱うジャンルを重視すべきだと考えられる. 一方音楽・バラエティージャンルは番組数も比較的多く, 出演している芸能人の名前が EPG のテキストに多く含まれるという意味では情報量も多い. しかし流行や現実の出来事とは直接的な関連が薄いと考えられるため, 今回の時事キーワード抽出の入力としては利用しないものとした. またスポーツ, 映画, ドキュメンタリーなどのジャンルも流行や現実の出来事とは直接的な関連が薄いと考えられるうえ, 番組数も比較的少ないため, 今回の時事キーワード抽出の入力としては利用しないものとした.

3. 時事キーワード抽出

文書集合からキーワード (特に複合語) を抽出するためには C-value という手法が古くから知られている [2]. この手法は文書集合における単語間の結合度を計算するものであり, 語 w の C-value は w の出現頻度 $n(w)$, w を含むより長い複合語の出現頻度 $t(w)$, w を含むより長い複合語の異なり数 $c(w)$ を用いて

$$C\text{-value}(w) = (|w|_{\text{morph}} - 1)(n(w) - t(w)/c(w))$$

で定義される. ここで $|w|_{\text{morph}}$ は w の形態素数である. w がより長い複合語の一部としてしか出現していない場合は $n(w) \approx t(w)/c(w)$ となるため, C-value は小さな値となる.

3.1 文書集合としての EPG の縮退

しかしながら一般の文書集合と異なり EPG のテキストは

- 同じ内容の番組を何度も再放送している
- ニュースなど時事的な番組は当日以外は詳細な情報を提供できない

などが原因で, チャンネルや時間帯によってはほとんど同じものが頻出するという特徴がある. そのためこれらの番組をそのまま処理すると, これらの番組の重要度を相対的に高く扱うことになってしまう. 同様に, EPG に記述できる文字数は制限されているため, 放送の長さが短い番組を長い番組と同列に処理すると短い番組の重要度を相対的に高く扱うことになってしまう. そこで情報源の公平性を保つため, 再放送などほとんど同じテキストである番組や長さが 15 分以下である番組は削除し, 文書集合を縮退するものとした.

表 2 は 2007/07/23 の EPG に含まれる番組数を放送波ごとに集計したものである. 縮退前は BS デジタルの番組数は地上波デジタルとほとんど同じであるが, 縮退後は地上波アナログ

と同じくらいまで減少することがわかる。これは BS デジタルでは他の放送波と比較して同一内容の番組の再放送が多いためだと考えられる。

表 2 2007/07/23 の EPG に含まれる番組数 (縮退前と縮退後)

放送波	地アナ	地デジ	BS	CS
縮退前	2589	6364	6557	13866
縮退後	788	1280	988	7509

また EPG のテキストは、番組のメタデータを表す角括弧でくくられた表現 ([再] = 再放送, [解] = 解説者など) や放送関連用語 (速報, スタジオ, 中継など), 誇大表現 (緊急, 最新, 衝撃など) の出現頻度が一般の文書よりもかなり高いという特徴がある。これらの表現はキーワードとしては意味がなく, C-value の計算においてノイズとなる可能性が高いため, 前処理としてあらかじめ除去するものとした。

3.2 複合語抽出のための新たな指標 C'-value

前節で述べたように情報源の公平性を保つため番組の縮退を施しても, 短いキーワードの場合は $n(w)$, $c(w)$ が $t(w)$ と比較して大きな値となりやすく, その結果 C-value が不当に大きな値となることがある。これは C-value では短いキーワードが不当に抽出されやすいということを意味している。そこで我々は短すぎるキーワードの抽出を抑制するため, C'-value という新たな手法を提案する。C'-value は

$$C'\text{-value}(w) = (|w|_{char} - 1 - t(w)/c(w)) \times (n(w) - t(w)/c(w))$$

で定義される。ここで $|w|_{char}$ は w の文字数である。C-value との大きな差は形態素数ではなく文字数を用いている点である。C-value では形態素が 1 文字からなる場合も複数文字からなる場合も同一に扱っていたが, C'-value では文字数を用いているため長いキーワードが抽出されやすくなっている。ただし文字数をそのまま用いると長いキーワードが不当に抽出されやすくなってしまうため, ヒューリスティックとして文字数に対して $t(w)/c(w)$ を補正するものとした。

3.3 時事性の判定

前節で述べたように, 短すぎるキーワードの抽出を抑制するため C'-value という新たな手法を提案した。一方, 文書集合に含まれるキーワードの中でも時事的なキーワードの場合, 例えば最近 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 C-value を用いてキーワードを抽出するだけでなく, 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。

あるキーワードの出現確率が期間によらず一様であると仮定すると, 出現頻度は期間の長さに比例する。そのため最近 N 日間の出現頻度を u とすると, 最近 n 日間の出現頻度 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。生起確率

p , 試行回数 u の二項分布の確率関数 $P(v) = {}_u C_v p^v (1-p)^{u-v}$ 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おり, Z 検定によって平均がある値に等しいかどうか (すなわちキーワードの出現確率が一様であるか) を検定することができる。

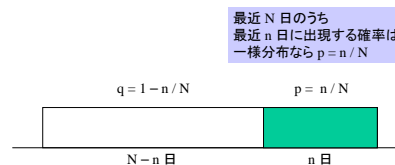


図 2 出現確率が一様であるときの出現頻度

表 3 は 2007/07/23 時点の EPG に対し $(N, n) = (28, 7)$ を与え, Z 検定によって時事的であると判定されたキーワードの例である。「原発」「避難」「倒壊」など, 短すぎて単体では話題がわからないキーワードに対する C-value は大きな値となっているが, C'-value は小さな値となっており, 短すぎるキーワードの抽出が抑制されていることが確かめられる。また Z 検定によって, 「震度 6 強」(2007/07/16) のようにすでに時期が過ぎたキーワードや「参院選」(2007/07/29) のようにまだ時期が来ていないキーワードをうまくふるい落とすことができていることがわかる。

表 3 2007/07/23 の EPG から抽出されたキーワードの例

keywords	C-value	C'-value	Z-value
中越沖地震	89.50	53.14	8.49
被災地	62.00	31.00	9.31
参院選	52.82	10.88	<u>0.94</u>
震度 6 強	37.80	20.16	<u>0.38</u>
登場	30.03	0.79	<u>-2.78</u>
原発	20.90	<u>-2.04</u>	8.12
地震から 1 週間	18.00	18.00	3.00
フジ子杯	15.00	10.00	<u>0.98</u>
避難	15.00	<u>0.00</u>	6.58
平塚 5 遺体	11.20	7.84	3.46
倒壊	10.79	<u>-2.25</u>	6.00
...

4. キーワードの話題ごとの分類

前章では時事キーワード抽出について説明したが

- 関連のあるキーワードがバラバラに抽出される
- 単体では抽出されたキーワードの話題がわからない場合がある

などの問題がある。そのため抽出されたキーワード間の関連を表示することができれば話題の把握がより容易になると考えられる。そこで本節では, キーワードの共起関係を利用し意味的なまとまりにわけること, 自動的にキーワードを話題ごとに分類する実験を行なった結果について説明する。

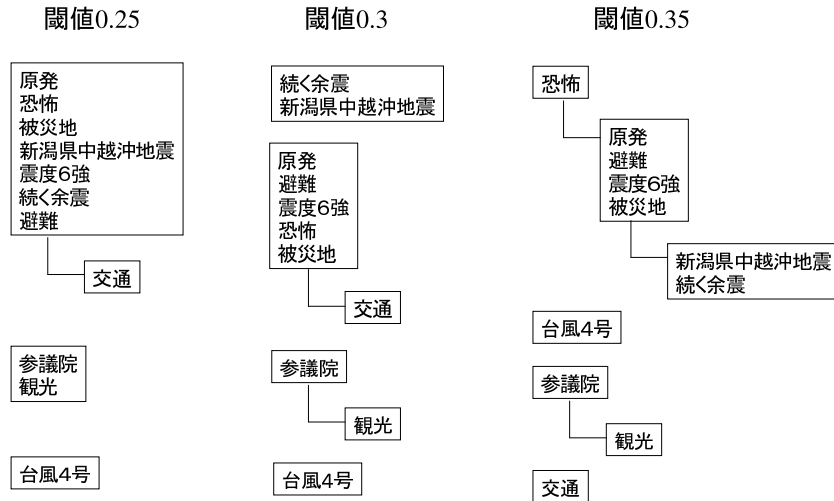


図 3 閾値を変化させたときの強連結成分分解による話題分類

4.1 共起グラフの強連結成分分解による話題分類

EPG における時事キーワード w_1, w_2 の出現確率を $P(w_1), P(w_2)$ とする。仮に w_1 と w_2 に意味的なつながりがあれば同じ番組に出現する確率が増加するので、 w_1 が出現する番組に w_2 が出現する条件付確率 $P(w_2|w_1) = P(w_1, w_2)/P(w_1)$ は $P(w_2)$ より大きくなると考えられる。時事キーワードを点で表し、時事キーワード間の条件付確率が与えられた閾値よりも大きいときに有向枝を接続することによって、時事キーワード間の向きを持った共起関係を表す有向グラフを構築することができる。

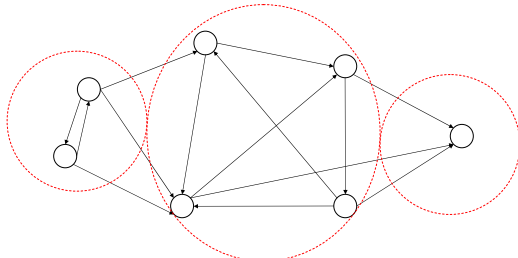


図 4 有向グラフの強連結成分分解

時事キーワード間の共起関係を表す有向グラフを強連結成分分解した結果得られる強連結成分は、それぞれの成分に含まれる2点間には必ず行き来できることを考えると、共起関係における同値類であるとみなすことができる。同じ話題に含まれる時事キーワードは同じように共起していると考えられるため、この手法によって強連結成分を抽出することで時事キーワードを話題ごとに分類できる可能性がある。

図 3 は 2007/07/23 時点のデータに対し $(N, n) = (7, 3)$ を与え、それぞれ時事キーワード間の条件付確率が閾値 0.25, 0.3, 0.35 以上の場合に有向枝を接続した有向グラフを強連結成分分解した結果である。各列が閾値による差を、四角で囲まれたキーワードの集合が一つの強連結成分を、ツリーの分岐が強連結成分の順序関係を表している。図 3 から閾値が 0.3 や 0.35 のときは比較的うまく話題ごとにツリーがわかれていることがわかる。

実際さまざまな日付のデータで閾値を変えて実験してみたところ、経験的には 0.3 くらいのときが一番うまくいくことがわかった。ただし抽出された時事キーワードの数や微妙な条件付確率の差でグラフの構造が大きく変化してしまうため、この手法はあまりロバストではないことも確かめられた。

一方閾値が 0.25 のときは多くのキーワードが一つの話題として固まりすぎてしまっていることがわかる。これは「不安」「会見」のような多くの番組に一般的に出現する単語が共起グラフにおけるハブの役割を果たし、一つの強連結成分が非常に大きくなってしまったためだと考えられる。本来関連のないキーワードである「参議院」と「観光」がつながってしまっているのも原因は同じである。

そこでこれらの点を踏まえ、単純な強連結成分分解ではなく枝の本数や枝の重み（条件付確率）を積極的に利用した実験を行なった。枝が 2 本以上でつながった強連結成分、および枝の重みが 1.0 以上となる強連結成分による分解の結果を図 5 に示す。共起グラフを構築するために用いた閾値は 0.25 である。

この図を見ると、単純な強連結成分分解では一つの話題として固まりすぎてしまっていたキーワードが、閾値を 0.3 や 0.35 に変化させたときと同じくうまく話題ごとにわかれていることがわかる。また図 3 では閾値を変化させても正しく分類することができていなかった「交通」が「台風4号」のツリーに正しく分類されるようになるなど、話題の分類の精度自体も改善することができていることが確かめられる。

4.2 考 察

共起グラフの強連結成分分解による分類は、微妙な条件付確率の差でグラフの構造が大きく変化してしまうためあまりロバストではないものの、閾値によってはうまく話題ごとにグラフをわけることができることが確かめられた。また枝の本数や枝の重み（条件付確率）を利用することで話題の分類の精度を改善することができた。しかし枝の本数や枝の重みを利用しても、一般的に出現する単語がグラフにおけるハブの役割を果たし本来関連していない単語までくっついてしまう問題は完全には解

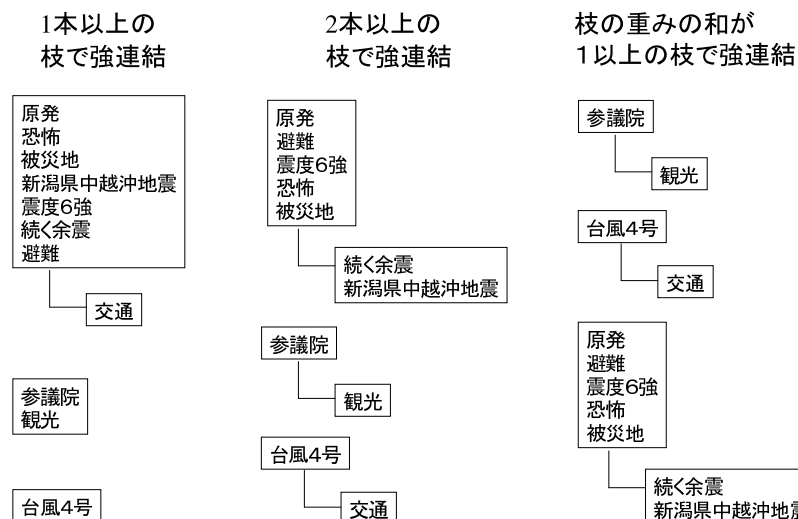


図 5 枝の本数や枝の重みを利用したときの強連結成分分解による話題分類

消できていない。たとえば今回の例ではあがっていないが、秋田男児殺害事件における「豪憲君」とW杯における「日豪戦」がつながってしまうという問題も確認された。これは「豪憲」が人名辞書に載っていないため「豪」「憲」と形態素解析され、「日豪戦」の「豪」と同じ単語として扱われてつながってしまったためである。そのため枝の本数や重みだけでなく、時事キーワードの出現頻度や文字列長によるフィルタリング、あるいはidfのようなキーワード自体の重要度の推定値を利用することも必要になると考えられる。

5. 関連研究

文書検索においては、単語を各次元の単位ベクトルとみなした高次元ベクトル空間を想定し、そこに配置された検索対象の文書とベクトル表現された検索クエリの距離によって検索対象の文書と検索クエリの関連性の強さを計算するベクトル空間モデルという手法がよく用いられている。

しかし一般には単語の生起は独立ではなく、単語を単位ベクトルとみなしたときにそれぞれが独立である保証はない。そこで考案された手法が潜在意味解析 (Latent Semantic Analysis, LSA) である [1]。もともとのベクトル空間に対して統計処理 (特異値分解) を施し、次元を縮約することでさまざまな文脈において単語の意味がどのように使用されているかを導出する。

しかしどちらの手法も EPG をそのまま文書集合として扱うことは難しいと考えられる。なぜなら EPG に記述できる文字数はかなり制限されていて一つの番組の文書長は通常のテキストに比べて非常に短いにもかかわらず、複数のまったく関連のない話題がまざっていることが多い。そのため一つの文書が一つの文脈を表しているという前提が成り立っていないことが多いためである。

他方ネットワーク上の協調作業によって発生するメッセージ列からユーザの興味を引く話題を抽出する手法も提案されている。例えば石井 [4] らのアプローチは掲示板につきつぎと書き込まれるメッセージにおける語の発生密度に着目した手法であり、斎藤 [7] らのアプローチはメーリングリストにおいて大きく

議論となった話題を示す語を抽出する手法である。EPG に含まれる個々の番組情報には放送局名やジャンル情報といった間接的な関連性はあるが、チャットや掲示板におけるメッセージ列のような直接的な関連性はない。そのため、[4] や [7] はメッセージ間の関連性を積極的に利用している点が我々の手法とは異なる。

6. おわりに

今回我々は、話題抽出技術の中でも「時事性」という切り口に注目することによって話題の時間的な継続性や変化の特徴を表すキーワードを抽出する技術を開発した。中でも EPG における出現頻度を統計的に検定して時事性を判定し、かつ出現した文脈を計算することでうまく話題ごとに分類することができた。今後は抽出した時事キーワードが最新の話題へのインデクスとしてうまく働くかどうか検索キーワードとしての有用性を検証していく予定である。また「時事性」以外の「地域」や「個人の嗜好」といった切り口に注目することによって、時事以外の観点からの話題を表すキーワードを抽出する技術を開発していく。

文 献

- [1] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. In *Journal of the Society for Information Science* 41(6), pp. 391–407, 1990.
- [2] K. T. Frantzi and S. Ananiadou. Extracting nested collocations. In *Proceedings of 16th International Conference on Computational Linguistics*, pp. 41–46, 1996.
- [3] goo ランキング. <http://ranking.goo.ne.jp/keyword/>.
- [4] 石井, 中渡瀬, 富田. 名詞句と単語の勢いをを用いた話題抽出手法の提案. 情報処理学会研究報告. 自然言語処理研究会報告, Vol. 2004, No. 23, pp. 79–84, 2004.
- [5] livedoor キーワード. <http://keyword.livedoor.com/>.
- [6] きざし. <http://kizasi.jp/>.
- [7] 斎藤, 水澤, 山本, 山口. 話題の自動抽出による電子メールの情報組織化手法. 情報処理学会論文誌, Vol. 39, No. 10, pp. 2907–2913, 19981015.