字幕テキストの利用によるブログで引用されたテレビ番組の推定

及川 孝徳[†1]　中島 泰[†2]　松崎 勝彦[†3]　黒木 さやか[†4]　山名 早人[†5, †6]

†1, †2, †3, †4 早稲田大学大学院基幹理工学研究科 〒169-8555 東京都新宿区大久保 3-4-1

†5 早稲田大学理工学術院 〒169-8555 東京都新宿区大久保 3-4-1

†6 国立情報学研究所 〒101-8430 東京都千代田区一ツ橋 2-1-2

E-mail: {t_oikawa,tai,matuzaki,kuroki,yamana}@yama.info.waseda.ac.jp

あらまし インターネットの普及に伴い，ウェブ上に存在する情報は増え続けている．2003 年頃から急速に普及してきたブログは，現在注目されている情報源の 1 つである．ブログからの情報抽出を目的とした研究が数多く行われており，その中には，ブログ記事内で取り扱っている話題を抽出する研究もある．しかし，その多くがウェブ上の情報のみで完結しており，ウェブ以外の情報と結びつけている研究は少ない．そこで本稿ではブログ記事の解析にあたり，ウェブ以外の情報である，テレビ放送の字幕テキストを利用する手法を提案する．テレビ番組について書かれたブログ記事に含まれる特徴語と，字幕テキストとを照らし合わせることで，番組名を含まないブログ記事からも正確な番組名を推定する。実験の結果，検索エンジンでは推定できない番組の推定が可能であり，字幕テキストが検索の精度・網羅性向上に有効な情報源であることがわかった．

キーワード ブログ，話題語抽出

Detecting TV Program quoted in Blog by using Closed Caption Streams

Takanori OIKAWA[†1]　Tai NAKAJIMA[†2]　Katuhiko MATUZAK[†3]

Sayaka KUROKI[†4]　Hayato YAMANA[†5, †6]

†1, †2, †3, †4 Graduate School of Fundamental Science and Engineering, Waseda University 3–4–1 Okubo, Shinjuku-ku, Tokyo, 169–8555, Japan

†5 Science and Engineering, Waseda University 3–4–1 Okubo, Shinjuku-ku, Tokyo, 169–8555, Japan

†6 National Institute of Informatics 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, 101-8430, Japan

E-mail: { t_oikawa,tai,matuzaki,kuroki,yamana }@yama.info.waseda.ac.jp

Abstract As the Internet spreads, information in the web keeps increasing. The blog that has spread since 2003 is one of the important sources. Studies that extract information from blog are active. One of them is extraction of topic of blog. However most studies use only sources in the web, and studies used sources not in the web is few. In this paper, we analyze blog by using closed caption streams. We can get accurate TV program name by using words used in blog and closed caption. By using our way, we can detect the TV program that cannot be detected by using the search engine. Result of the experiment shows that closed caption is an effective source for improving precision and recall of the search engine.

Keyword Weblog, Topic Extraction

1. はじめに

インターネットの普及に伴い，ウェブ上に存在する情報は増え続けている．2003 年頃から急速に普及してきたブログは，現在注目されている情報源の 1 つである．総務省[1]によると，ブログは 2008 年 1 月の時点で国内において 1690 万に達し，記事総数は 13 億 5000 万件に上る．ブログは，個人やグループによって運営されるウェブページであり，著者の日記・ニュース・商品等の評価が書かれている．ブログの特徴として，著者の意見，興味が反映されやすく，速報性・リアルタイム性があることが挙げられる．また，トラックバックやコメントといった機能をもつため，ユーザー同士のコミュニティも形成されている．このように様々な情報を持つため，ブログは有用な情報源としてとら

えられている．

ブログに関する研究は様々な種類がある．代表的な研究として,特定の対象について書かれた文書の抽出,ブログ記事内で話題となっている要素の抽出，リンクや共起によるコミュニティの抽出，ブログの信頼度算出と推薦，ブログの監視と収集，スパム除去を目的とした記事のフィルタリングなどが存在する．本稿では,ブログ記事内で話題となっている要素の抽出に関する研究を提案する．

ブログ記事は,他のブログ記事,書籍,テレビ番組,商品といった特定の対象についての話題を扱うものが多い．よって，企業においてはマーケティングへ利用が期待されている．しかし，ブログは企業が提示する情報とは異なり，正確でわかりやすくする必要がないため，話題の対象となっている対象について、正式な名称が書かれていないこともある．そのため，特定の対象について書かれたブログ記事を収集しようとしても，正式名称で検索するだけでは対象について書かれたブログ記事全体の一部しか得ることはできない．よって，ブログから情報を抽出する際には，特定の対象と関連性の高い関連語群を生成し，関連語群でも検索を行い，ブログ記事の収集を補助する必要がある．

関連語の生成に関しては様々な研究が存在する．「対象」の別称を関連語とみなす手法として，複数の省略語を推測することで「対象」の別称に対応した研究[2][3]がある．また,「対象」を表す語と共起する単語を関連語とみなし，ブログ集合の共通の要素から関連語を推定する研究[4]も行われている．しかし，多くの研究はウェブ上の情報のみで行われており，ウェブ以外の情報を利用した研究はない.ウェブ上の情報に,ウェブ以外の情報源を加えて利用することにより，扱う情報の幅が広がるため，精度・網羅性の向上が可能となる．

本稿では，ブログ記事内で取り扱われているテレビ番組に焦点をあて，ウェブ以外の情報として字幕テキストを関連語群として利用した話題対象・引用元推定手法を提案する．字幕テキストとは，テレビの字幕放送において表示されるテキストを示す．字幕放送は聴覚障害者のテレビ視聴支援のために開発された，テレビ画面文字表示技術である.字幕テキストは 1985 年から始まり,現在は NHK においてはほぼすべての番組,他の民間放送局でもゴールデンタイムはほとんどの番組で行われている．デジタル放送においてはすべての受信機で字幕放送に対応しているため，今後テレビの地上波放送がデジタル化するにあたり，字幕放送は広く普及していくと考えられる．字幕放送は，番組内のすべての音声を文字情報にしている．つまり，番組における関連語を多く含んでいる．本稿では，字幕情報とブログ記事の対応から，話題対象・引用元となっているテレビ番組の推定を行う．

提案手法は 2 つの段階からなる．1 段階目では，対象のブログ記事から制約条件の設定と特徴語群の抽出を行う．制約条件は，2 段階目で字幕テキストを探索する際，探索範囲を絞り込むための条件であり，対象ブログ記事が書かれた時間，対象ブログ記事に含まれるテレビ局名・番組ジャンル名で設定される．例えば,「昨日見た」のように時間情報が含まれていた場合,ブログ記事の書かれた時間とあわせ，制約条件を設定する．特徴語群の抽出は，ブログ記事に形態素解析を行い，得られた名詞を特徴語群とみなして行う．

2 段階目では対象のブログ記事がどのテレビ番組を話題対象・引用元にしているかを，字幕テキストから推定する．推定は，1 段階目で得られた制約条件によって絞り込まれた字幕テキストから，特徴語群と一致率の高い番組を抽出することで行う．推定の際，一致率が一定以上の番組が存在しない場合，対象のブログ記事はテレビ番組を話題に含まないと判定する．

提案手法によって，特定の番組のウェブ上における影響調査や，番組で出現した話題がウェブに反映される時間経過の把握が可能となる．

本稿の構成は以下のとおりである．第 2 節では，関連研究について述べる．第 3 節では提案手法について説明する．第 4 節では提案手法の評価実験について述べ，第 5 節ではまとめを述べる．

2. 関連研究

2.1. 省略語の推定に関する研究

関口ら[2]，村山ら[3]は特定対象の関連語として，対象の正式名称の略語を生成する研究を行った．

関口ら[2]は正式名称の文字列から任意の文字を抜くことで生成できる全ての略語候補を生成し，元となる正式名称を含むブログ記事集合と，略語候補を含むブログ記事集合の内容の類似度によって，略語であるかどうかを判定した．

村山ら[3]は，略語候補が大量になるのを防ぐため,文字の省略や削除といった，正式名称から略語が生成されるプロセスをモデル化し，確率的に高いモデルを使用して略語候補の生成を行った．結果，略語候補の正誤判定の計算コストを小さくすることに成功した．

関口ら，村山らともに精度よく省略語を獲得している．ただし，略語を関連語として用いる手法は，代名詞が用いられている等，対象の名前を記述していない文書には対応できない．

2.2. ブログ集合を利用した関連語抽出に関する研究

関口ら[4]は，特定対象の関連語として，特定対象と頻出する単語を抽出する研究を行った．

関口らは，ある著者と同じ興味をもつ人々のブログ集合を利用し，ブログ集合の共通要素から対象の関連語の抽出を行った．まず，対象著者のブログ記事と同じ語句を扱っているブログ記事を集め，それぞれ対象著者のブログ記事と一致する語句の量で関連度を算出する．そして，関連度の高いブログ記事ほど出現度が高くなっている語句を，特徴的な関連語とみなし抽出する．

ブログ集合を利用して関連語抽出を行う手法は，対象語句の関連語なのか，対象語句を含むジャンル全体の関連語なのかの判別が難しいという欠点がある．

2.3. 異なったデータベース間の対応付けに関する研究

池田ら[5]はブログ記事とニュース記事の対応付けを行った．まず，ブログ記事，ニュース記事それぞれで特徴的な語を抽出する．次に，各特徴語に出現頻度で重みづけを行う．重みづけの際，ニュース記事という特性を生かし，ニュース記事が配信された直後に出現頻度が増した単語には別に重みづけを行っている．最後に特徴語群の類似度でブログ記事とニュース記事の対応付けを行っている．

結果，ニュース記事の特性を生かした重みづけによって精度の向上に成功した．

3. 提案手法

本章では提案手法について述べる．提案手法の流れを図 1 に示す．

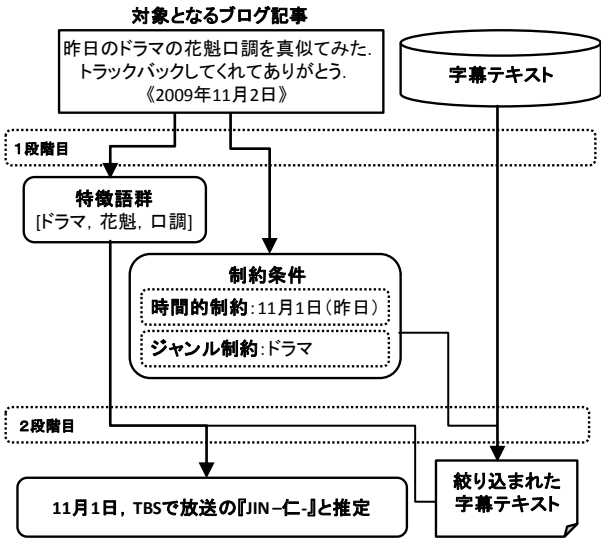

図 1 提案手法処理の流れ

処理は 2 段階に分かれ，1 段階目で対象のブログ記事から，制約条件の設定と特徴語群の抽出を行う．2 段階目では 1 段階目で得られた特徴語群を，制約条件で絞り込みを行った字幕テキストと照らし合わせ，一致率の高い番組を抽出する．各節で処理について述べる．

3.1. 字幕テキスト探索範囲絞り込みのための制約条件設定

対象のブログ記事に含まれる表現によって時間・テレビ局・番組ジャンル, 3 つの制約条件の設定を行う．設定された制約条件は，3.3 節のテレビ番組推定で探索される字幕テキストの絞り込みに利用される．

時間制約の設定は「昨日，今日」という単語についてのみ行った．2 つのうちどちらかが含まれていた場合，探索範囲をブログ記事が書かれた日を基準に 2 日前までとした．どちらも含まれていなかった場合は探索範囲をブログ記事が書かれた日を基準に 7 日前までとしている．

テレビ局・番組ジャンル制約の設定は対象のブログ記事にテレビ局名・番組ジャンル名が含まれていた場合に行う．テレビ局名は字幕テキストを持っている 7 局分，番組ジャンル名にはジャンルのうち一部を利用した．テレビ局名・番組ジャンル名が含まれていた場合は，字幕テキストの探索範囲を含まれていたテレビ局，含まれていたジャンルに絞り込む．

制約条件に使用した単語を表 1 に示す．ただし，テレビ局名は表記ゆれを考慮している．

表 1 制約条件設定に利用する単語

時間制約	今日，昨日
テレビ局制約	NHK 総合，NHK 教育，TBS，日本テレビ，テレビ朝日，テレビ東京，フジテレビ
番組ジャンル制約	ドラマ，ニュース，バラエティ，アニメ

3.2. ブログ記事からの特徴語抽出

まず，ブログ記事に形態素解析を行い，すべての名詞を特徴語候補として抽出する．

次にブログ記事において頻出する名詞を除去する．ブログ記事で頻出する名詞を特徴語群に含めてしまうと, 常に同じ番組が推定される危険があるためである．実験では, ランダムに選んだ 5000 件のブログ記事内での出現頻度によって頻出名詞を決定した．

最後に，各候補に字幕テキスト内での IDF 値によって重みづけを行い, 重みが一定値以下の候補を除去し，残ったものを特徴語群とする．

3.3. 字幕テキストからのテレビ番組推定

まず，3.1 節で設定した制約条件で字幕テキストの絞り込みを行う．絞り込みのイメージを図 2 に示す．

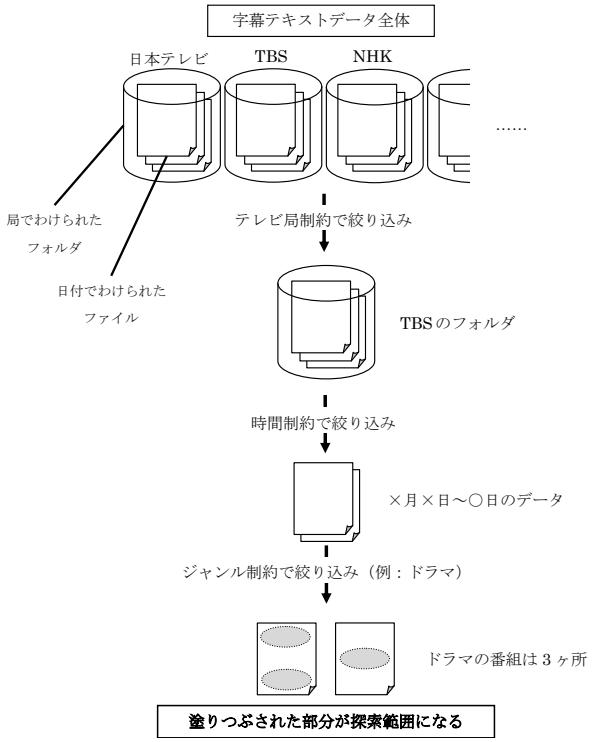

図 2　字幕テキスト絞り込みのイメージ

次に，絞り込まれた字幕テキストから 3.2 節で得られた特徴語群と一致率の高い番組を抽出する．絞り込まれた字幕テキスト内で各特徴語の探索を行い，特徴語を含んでいる番組にその特徴語の重み分スコアを加算する．特徴語群すべてにおいて同様の処理を行い，スコアの最も高いものを抽出する．実験ではスコアの上位 10 件を求めた．

4.　評価実験

4.1.　使用データ・実験環境

字幕テキストには，2005 年 1 月から 2009 年 11 月までの，関東で放送された 7 局のテレビ番組おける字幕テキストを用いた．また，ブログ記事集合として，2007 年 7 月～2009 年 7 月までの 2 年間のブログデータを使用した．各データ量を表 2 に示す．

表 2　実験に使用したデータ

	収集期間	データ量
字幕テキスト	2005.1～2009.11	472,031 番組
ブログ記事	2007.7～2009.7	約 10TB

実験環境として，形態素解析には Yahoo!JAPAN が提供している日本語形態素解析[6]を利用した．

4.2.　評価方法

実験対象として，テレビ番組について書かれているブログ記事 100 件を人手で収集した．収集の際，字幕テキストを持っていないテレビ局の番組等，推定不可能な番組についてかかれたブログ記事は除外している．また，収集したブログ記事には，記事内に番組正式名称が含まれていた場合，正式名称の除去を行った．収集したブログ記事の内訳を表 3 に示す．

表 3　実験に使用したブログ記事内訳

ドラマ	バラエティ	アニメ	情報	その他	合計
24 件	25 件	23 件	22 件	6 件	100 件

作成した実験対象ブログ記事を提案手法のプログラムに入力し，推定番組上位 10 件を表示し，評価を行った．

4.3.　評価結果

評価実験の結果，提案手法の精度を表 4 に示す．

表 4　提案手法精度

精度―候補 1 位	精度―候補 5 位以内
55%	72%

以上の結果より，字幕テキストを利用することでブログ記事の話題番組推定が可能であることがわかる．

また，2 つの正解番組推定例を挙げる

> 庵ｗｗｗｗゴルゴｗｗｗｗｗｗまぁｗわからないからなｗｗ皆ｗｗｗ　ザゴールデンゴールデンｗｗｗｗｗ　サバンナｗｗｗｗあるよｗｗ目が刺さってﾃﾝｼｮﾝ下がるｗｗｗｗ

上記のブログ記事からは，特徴語として

[庵，ゴルゴ，ザゴールデン，サバンナ]

が抽出され，提案手法は正解番組「爆笑レッドカーペット」を推定する．同じ特徴語を検索エンジンに投げた場合，検索エンジンは上位に「爆笑レッドカーペット」を返す．

> 昨日の続きですが、住民税の話。タイムリーな事に今朝のＮＨＫの番組で取り上げられておりました。…

上記のブログからは，特徴語として

[住民税，タイムリー，NHK，…]

が抽出され，提案手法は正解番組「家計診断　おすすめ悠々ライフ」を推定する．しかし，同じ特徴語を検索エンジンに投げた場合，ブログ記事が書かれた時間等をクエリに付加しても正解を上位には返さない．

理由として，字幕テキストは検索エンジンに比べ，限定的な情報源から探索を行うことができるため，少

しの特徴しか持たない単語からも番組を推定できることが挙げられる．よって，字幕テキストの利用は検索の精度・網羅性の向上に有効な情報源であることがわかる．

5. まとめ

本稿では，テレビ放送の字幕テキストを利用した，ブログ内で引用された番組名推定の提案手法について説明し，実験を行った．結果として，候補 1 位の精度は 55%，候補 5 位以内の精度では 72%を出すことができた．また，検索エンジンでは対応しきれいないブログ記事に対しても，適切な推定が可能であった．よって，字幕テキストは検索の精度・網羅性向上に有効な情報源といえることがわかった．

今後の課題としては，提案手法の精度向上の他，提案手法とは逆に，特定番組の字幕テキストの特徴語群を利用することで，対象の番組について書かれたブログ記事を抽出する研究を行いたい．また，特徴語に，番組で使われる独特な台詞･言い回しを用いることで，潜在的に番組の影響をうけている著者の抽出も可能ではないかと考えている．

その他，より時間との結びつきが強い，Twitter 等のマイクロブログへの適用など，他の情報源への応用も行っていきたい．

謝　辞

本研究は，文部科学省・次世代ＩＴ基盤構築のための研究開発「多メディア Web 解析基盤の構築及び社会分析ソフトウェアの開発」及び科学研究費補助金（基盤研究（Ｂ）21300038）の補助によるものである．なお，解析用データ提供において国立情報学研究所の佐藤真一様，アクセラテクノロジ株式会社様にお世話になりました．ここに感謝いたします．

参　考　文　献

[1] 総務省 情報通信政策研究所，ブログの実態に関する調査研究の結果，http://www.soumu.go.jp/iicp/chousakenkyu/data/research/survey/telecom/2009/2009-02.pdf

[2] 関口裕一郎，佐藤吉秀，川島晴美，奥田秀範，“ブログ文書集合を用いた省略語抽出手法の検討”，電子情報通信学会技術研究報告，データ工学 107，pp.207-210，2007.

[3] 村山紀文，奥村学，“Noisy-channel model を用いた略語推定”，言語処理学会第 12 回年次大会，pp.837-840，2006.

[4] 関口祐一郎，佐藤吉秀，川島晴美，奥田英範，奥雅博，“blog ページ集合に対する話題語句抽出手法”，情報処理学会研究報告，Vol.2005，No.117，pp.27-32，2005.

[5] 池田大介，藤木稔明，奥村学，“blog とニュース記事の自動対応付け”，言語処理学会第 11 回年次大会発表論文集，pp.1030-1033，2005.

[6] Yahoo!デベロッパーネットワーク，http://developer.yahoo.co.jp/