新聞記事集合に対する時系列のトピック抽出

水落 大史† 井上 悦子‡ 吉廣 卓哉‡ 村川 猛彦‡ 中川 優‡

†和歌山大学大学院 システム工学研究科 〒640-8510 和歌山県和歌山市栄谷 930
‡和歌山大学 システム工学部 〒640-8510 和歌山県和歌山市栄谷 930

E-mail: †s091056@sys.wakayama-u.ac.jp, ‡{etsuko, tac, takehiko, nakagawa}@sys.wakayama-u.ac.jp

あらまし 本研究では,複数社の新聞記事に対する特徴語に基づくクラスタリングを用いた時系列のトピック抽出手法を提案する．複数社の記事を用いることで，各トピックの盛り上がりの変化や，どのような関連記事に派生していくかが把握できる．提案手法では,一日単位で記事をトピックに分類し,時系列でトピックの関連付けを行う．本手法で抽出した時系列のトピックが妥当であるか,評価をしたところ,被験者からは概ね良好な回答が得られた．

キーワード 時系列のトピック，クラスタリング，新聞記事

A Method for Chronological Topic Extraction from Newspaper Articles

Hirofumi MIZUOCHI†, Etsuko INOUE‡, Takuya YOSHIHIRO‡,
Takehiko MURAKAWA‡, and Masaru NAKAGAWA‡

†Graduate School of Systems Engineering, Wakayama University
930 Sakaedani, Wakayama-shi, Wakayama, 640-8510 Japan
‡Faculty of Systems Engineering, Wakayama University
930 Sakaedani, Wakayama-shi, Wakayama, 640-8510 Japan

E-mail: †s091056@sys.wakayama-u.ac.jp, ‡{etsuko, tac, takehiko, nakagawa}@sys.wakayama-u.ac.jp

Abstract We propose a method for extraction of chronological topics from the newspaper articles of multiple newspaper sites. By using the method together with sufficient quantities of articles, we can understand what topic rises and falls, and find the related, derivative topics. The proposed method analyzes every article to extract the feature words, and clusters the articles into the topics by the day. After that, it relates the topics over the successive two days. We received favorable responses when conducting assessment of our method in terms of the adequacy of related topics.

Keyword Chronological topic，Clustering，Newspaper

1. はじめに

新聞は，世の中の情報を得るための媒体として広く普及しており,現在様々な新聞社から発行されている．しかし，新聞社ごとに扱っている内容には差異が存在する場合があり，過不足のない情報を得るためには，複数新聞社の新聞を読むことが望ましいとされている．公共や大学の図書館に行けば，手間はかかるが，読み比べが可能である．

複数新聞社の新聞を読み比べる必要性は，インターネット上の新聞記事にも当てはめることができ，近年では，複数新聞社のインターネット上の記事を読み比べるサイトが，新聞社によって運営されている[1]．しかし，上記サイトでは，新聞社ごとに主要な同一事件や話題のまとまり（トピック）について，新聞記事を並列に提示しているだけであり，それ以外のトピックについての記事同士の関連を把握するためには，個々の記事の中身を確認する必要がある．

また，新聞から世の中の流れをつかむためには，複数日の新聞記事から，どんな事件があって，どのトピックに注目が集まっているかと同時に，各トピックが時間の経過とともにどのように変化していくかを確認する必要があると考える．

そこで本研究では，トピックの盛り上がりや変遷を,俯瞰的に確認できるようにする前段階として，インターネット上にある複数新聞社の記事を自動的に分類し,時系列のトピック抽出を行う手法を提案する．

2. 関連研究

新聞記事から，社会現象や話題語を抽出する研究は盛んに行われている[2][3]．これらの研究では，複数日の新聞記事集合を解析し，最新話題語や課題発見を行っている．新聞記事の分類という点では本研究と類似

しているが，複数日の新聞記事集合を解析し，その中での最新話題語や課題の発見を目的としており，時系列での変遷は抽出対象としていない．

複数新聞記事から類似記事を抽出する研究も行われている[4]．これは，ユーザからの検索質問に該当する記事を提示すると同時に，提示した記事と類似した記事を，提示した記事との差異とともに提示するシステムである．類似記事との差異を表示することを目的としており，トピックの時系列での変遷を追うことを目的としている本研究とは目的が異なる．

また，新聞記事以外の時系列テキスト集合からの話題抽出の研究も行われている[5]．これは，電子番組表（EPG）から抽出したキーワードに対して，時系列での盛衰や推移を確認しているものであるが，本研究ではキーワードごとではなくトピックごとに時系列での変遷の抽出を行う．

3. 本研究で抽出する時系列のトピック

3.1. トピック

本研究では，日ごとの新聞記事集合の中で同じ事件や話題を扱っており，かつ，僅かな差異しか持たないような，類似性が非常に高い記事集合を「トピック」として扱う．類似性が非常に高い記事をまとめることによって，同じトピックの記事を重複して読むことがなくなる．

さらに，同じ事件や話題について述べているトピック同士をまとめたトピック集合を「トピックグループ」として扱う．同じトピックグループに分類される各トピックは，同じ事件や話題について述べているが，述べる事柄が違っているなどの理由から，同一トピックとみなすほどは類似性が高くないトピック集合である．

このようにトピックとトピックグループを定義すると，各日の新聞記事集合には，少なくとも一つのトピックグループが存在し，各トピックグループには少なくとも一つのトピックが含まれていることになる．

なお，本研究では，複数新聞社の新聞記事を対象とする．複数新聞社から記事を収集することにより，一つの事件でも複数の新聞社の記事が収集されるため，あるトピックに含まれる記事数を確認することで，そのトピックの盛り上がりや，個々のトピックグループの中でどのようなトピックに注目が集まっているかを把握できる．また，新聞社によって記事を掲載する事件，掲載しない事件などの差異があるが，複数新聞社の記事を使用することにより，報じているトピックの網羅性とトピック内の記事内容の網羅性を確保することができ，世の中全体としてどのような事件や話題が報じられているかを把握できる．

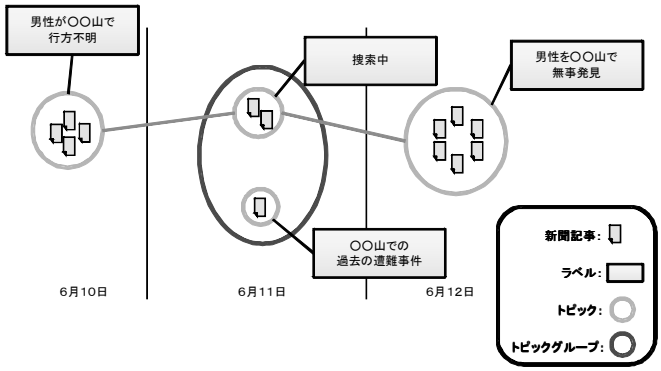

図 1　時系列のトピックイメージ

3.2. 時系列のトピック

本研究では一日単位で抽出した，トピックとトピックグループを，連続する日で同一事件や話題に関したトピック同士を時系列で関連付けたものを「時系列のトピック」として扱う．連続する日で関連付けることにより，同一事件や話題の内容の推移や，記事数の変化，つまり注目の度合いなどが追跡可能となる．

本手法で抽出する時系列のトピックのイメージを図 1 に示す．図 1 は，「男性が○○山で行方不明」という時系列のトピックの一連の流れとなっている．6 月 10 日に時系列のトピックが発生し，翌日の 6 月 11 日には，前日のトピックと関連があるトピックは「捜索中」のトピックとなっている．「○○山での過去の遭難事件」のトピックは，前日のトピックとは内容が異なり「○○山」に注目して述べている．「○○山での過去の遭難事件」のような，前日のトピックとは内容が違うトピックは，前日のトピックと関連付けされないが，関連付けされているトピックと同じグループトピックに含まれているので，同一の時系列のトピックとして追跡可能である．そして，その翌日の 6 月 12 日には「男性を○○山で無事発見」のトピックが関連付けられている．なお，図 1 でトピックに付属しているラベルの作成や抽出は，本研究では行わない．

ところで，日ごとに分類せずに，複数日の新聞記事をまとめてトピックまたはトピックグループに分類し，その結果を時系列に並べる手法が考えられる．しかし，新聞記事の性質上，時系列で内容が変遷していくので，発行日の差が開くほど内容に差異が出てしまうため，適切な時系列のトピックの抽出が困難となる．そこで，本手法では分類に支障が出ない，一日単位に分類し，時系列での関連付けを行う．

なお，本稿では，連続する日でのトピックのみの関連付けのみを紹介しているが，n 日後のトピックについて同様の処理を行えば，n-1 日間続報が途切れる場合でも時系列のトピックとして抽出可能である．

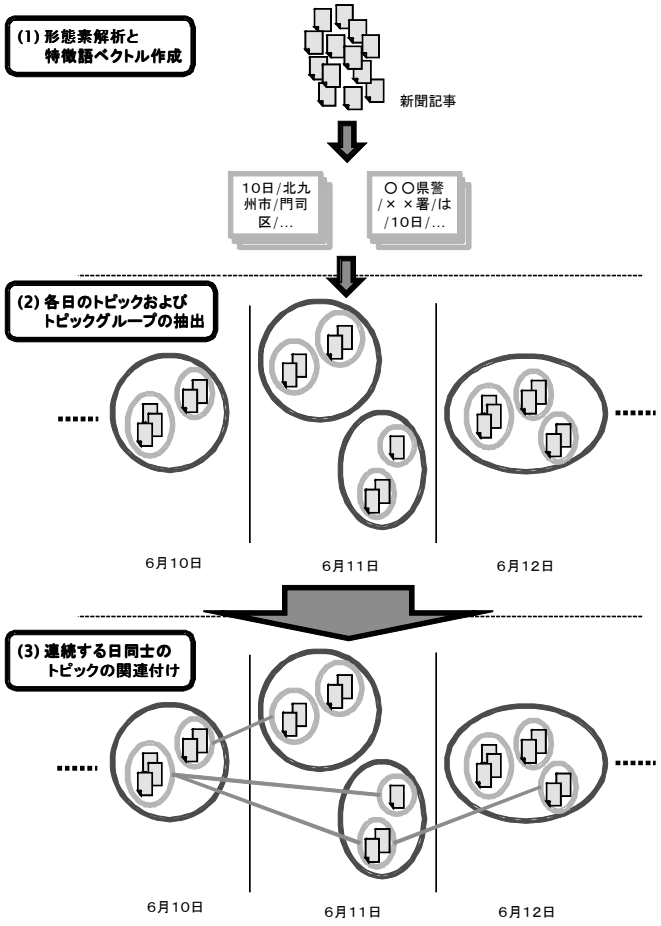

図 2　時系列のトピック抽出の流れ

4. 時系列のトピック抽出手法

4.1. 抽出の流れ

本手法では，新聞記事集合のクラスタリング結果から，トピックやトピックグループを抽出する．図 2 は本手法による時系列のトピックの抽出の流れを示したものである．その流れを簡単に説明する．それぞれの項目について詳しい説明は次節以降で行う．

（1）形態素解析と特徴語ベクトル作成

各新聞記事に対して形態素解析を行い，特徴語ベクトルを作成する．

（2）各日のトピックおよびトピックグループの抽出

各日の新聞記事集合から，各記事を 1 つのクラスタとみなし，各記事の特徴語ベクトルから記事間の関連度を求めて，階層型クラスタリングを適用する．その結果から，トピックとトピックグループを抽出する．

（3）連続する日同士のトピックの関連付け

手順(2)を複数日において行い，作成されたトピック同士を，連続する日ごとに関連付けたものを時系列のトピックとして抽出する．

4.2. 形態素解析と特徴語ベクトル作成

記事間の関連度を求めるために，それぞれの記事に対して形態素解析を行い，特徴語ベクトルを作成する．特徴語ベクトルを作成する手順は以下のとおりである．

まず，記事見出しと本文を，形態素解析器 MeCab [6] により解析し，単語と対応する品詞を判別する．さらに，判別した単語の中から，品詞が名詞である単語と，それらを複合語処理した語を特徴語として抽出する．

本手法では，複合語処理を行っているが，複合語を特徴語に使用すると，記事内の表記の揺れに対応出来ないという問題がある．例えば，田中太郎という氏名の裁判官についての記事があったとき，A 新聞社では「田中裁判官」と表記され，B 新聞社では「田中太郎裁判官」と表記された場合，複合語処理を行わない場合は「田中」,「裁判官」という特徴語が共通していると判定される．しかし，複合語処理した「田中裁判官」や「田中太郎裁判官」は，表記の揺れによって共通していないと判定される．本研究は，複数新聞社の記事を対象としているため，新聞社ごとの語句の表記の揺れは考慮すべき問題である．そこで，本手法では，形態素と複合語の両方を特徴語として使用する．

次に，特徴語ベクトルの重みとなる *TF-IDF* 値を求める．特徴語ベクトルを作成する記事を A，*TF-IDF* 値を求める特徴語を W_i としたときに，以下の式により記事 A に出現する特徴語 W_i の *TF-IDF* 値を求める．

$$TFIDF(A,W_i) = TF(A,W_i) \cdot IDF(W_i) \tag{1}$$

記事 A における特徴語 W_i の出現頻度を $tfreq(A,W_i)$ とすると，$TF(A,W_i)$ は以下のように定義される．

$$TF(A,W_i) = tfreq(A,W_i) \tag{2}$$

$IDF(W_i)$ は，過去の新聞記事集合の記事総数 N と，特徴語 W_i が 1 回以上出現する記事数 $dfreq(W_i)$ から，以下のように定義される．

$$IDF(W_i) = \log\left(\frac{N}{dfreq(W_i)}\right) \tag{3}$$

記事 A の特徴語ベクトル X_A は以下のように表す．ただし，$W_1 \cdots W_n$ は記事 A に含まれる特徴語群である．

$$X_A = \begin{pmatrix} TFIDF(A,W_1) \\ TFIDF(A,W_2) \\ \vdots \\ TFIDF(A,W_n) \end{pmatrix} \tag{4}$$

4.3. 各日のトピックおよびトピックグループの抽出

特徴語ベクトルで表された記事集合から，階層型ク

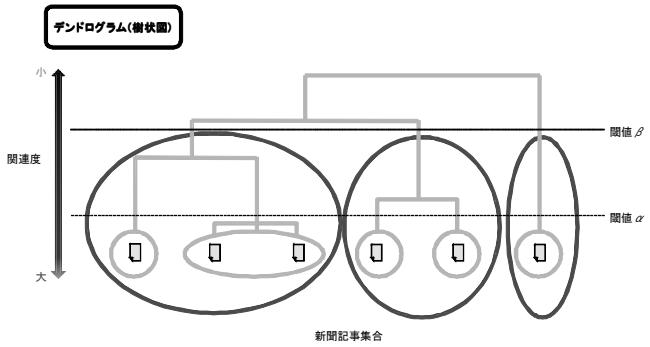

図 3　二段階でのトピック抽出

ラスタリングによってトピッククラスタを抽出する．記事間の関連度は特徴語ベクトルの余弦尺度を用いる．記事 A と記事 B の余弦尺度 S は,以下の式で表される．

$$S(X_A, X_B) = \frac{X_A \cdot X_B}{\|X_A\| \cdot \|X_B\|} \tag{5}$$

特徴語ベクトルで表された記事集合から，階層型クラスタリングを用いて，図 3 に示すようなデンドログラム（樹状図）のように関連度の高い記事から順に結合していく．ここで結合された特徴語ベクトルの値は以下の式で求める．

$$X_{AB} = \frac{X_A + X_B}{2} \tag{6}$$

階層型クラスタリングの結果から，関連度が閾値 α 以上のクラスタをトピック,閾値 β 以上のクラスタをトピックグループとして抽出する．クラスタリング結果から閾値 α で抽出されたトピックを T，閾値 β で抽出されたトピックグループを G で表す.図 3 の例では，6 つの新聞記事から二段階の閾値でトピックおよびトピックグループを抽出している．この例では，5 つのトピックと 3 つのトピックグループが抽出される．

4.4. 連続する日同士のトピックの関連付け

日ごとに抽出されたトピックを，連続する日のトピックに関連付ける．連続する二日間のトピックの関連付けには，各日のトピック同士の関連度を用いる．トピック同士の関連度は，トピック内の全記事の組み合わせに対する関連度の平均値とする．トピック同士の関連度は以下のように定義される．なお，n_i（i=1,2）は T_i 中の記事数を表す．

$$S(T_1, T_2) = \frac{1}{n_1 n_2} \sum_{X_{1i} \in T_1} \sum_{X_{2j} \in T_2} S(X_{1i}, X_{2j}) \tag{7}$$

$S(T_1,T_2)$が閾値 γ 以上であれば，トピック同士を関連トピックとみなす．図 4 は連続する特定の日同士を関連付ける流れを示したものである．

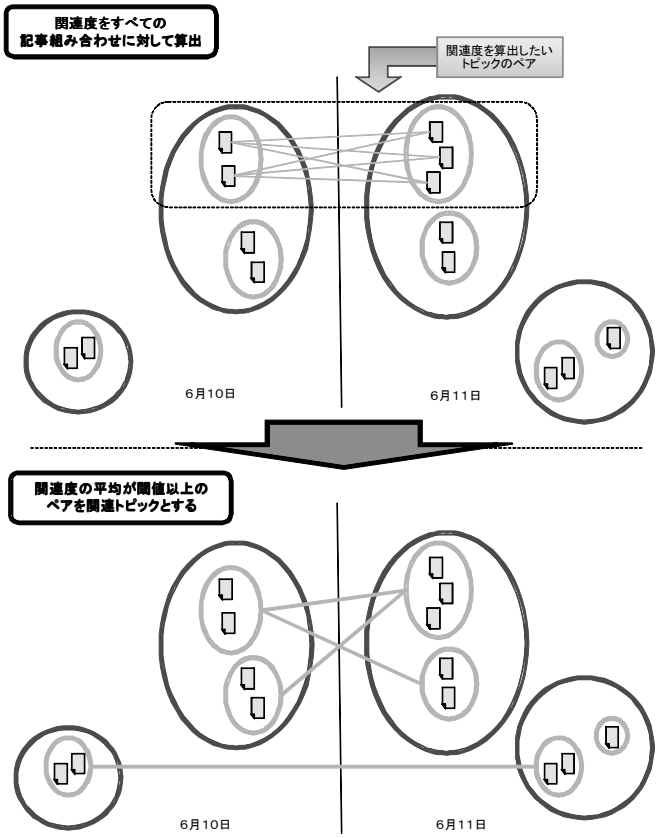

図 4　連続する日同士の関連付け

5. 評価

5.1. 評価実験

本手法を用いて抽出した時系列のトピックが妥当であるか確認するために，評価実験を行った．

実験に使用したのは，後述する 7 つの新聞記事サイトにおいて 2009 年 11 月 1 日から 2009 年 11 月 7 日までの一週間に掲載された新聞記事（1921 件）から，本手法で抽出した時系列のトピックである．それらの時系列のトピックを，筆者が所属する研究室の学生 7 名に確認してもらい，以下の項目を指摘してもらった．また，時系列のトピックの確認後に，アンケートに回答してもらった．

1. 分類が不適切なトピック
2. 分類が不適切なトピックグループ
3. 関連付けるべきではないトピック

なお，被験者には，各記事の記事見出しを提示し，記事見出しだけでは内容が判別できないような記事の場合は，随時，記事本文を確認できる状態で実験を行った．今回の実験で確認してもらった時系列のトピック数は 73，その中に含まれるトピック数は 281，トピックグループ数は 167，関連付け数は 116 であった．

今回の実験では，*IDF* 値を算出する際に必要となる記事集合の収集に使用した新聞記事サイト（新聞社）は，以下の 7 つである．新聞記事の収集期間は 2008

年 6 月 1 日から 2009 年 12 月 31 日まで，記事総数は 485,645 件となっている．

- 時事ドットコム（時事通信）
- 毎日 jp（毎日新聞）
- asahi.com（朝日新聞）
- MSN 産経ニュース（産経新聞）
- NIKKEI NET（日経新聞）
- TOKYO WEB（東京新聞）
- YOMIURI ONLINE（読売新聞）

また，今回の実験の対象記事として，記事 URL から判別できるカテゴリの中から，「社会」カテゴリの記事のみ使用している．これは，社会カテゴリの記事は，社会的な記事が多く，続報も多いため，経時変化を確認する本手法に適しているカテゴリだと判断したためである．

5.2. 結果および考察

評価結果を表 1，表 2，表 3 に示す．表 1 は，被験者ごとの指摘数をまとめたものである．この結果から，トピックと関連付けに関しては被験者によって不適切と判断した数にばらつきがあることがわかった．トピックグループに関しては，ほとんど指摘が見られなかった．

表 2 は，不適切であると指摘された箇所が，被験者間でどの程度共通しているかをまとめたものである．トピック，トピックグループ，関連付けのいずれに対しても，3 人以上の被験者が共通して指摘した箇所は無かった．表 2 からわかるように，複数の被験者が共通して不適切であると指摘した箇所は非常に少数であり，抽出結果は概ね良好であった．

表 3 は，実験終了後に実施したアンケート結果をまとめたものである．アンケートには，トピック，トピックグループ，関連付けが適切であったかを，5 段階評価で回答してもらった．適切であれば 5，不適切であれば 1 としている．全体的に，良好な回答が得られ，時系列のトピック抽出に関して，概ね問題ないといえる．

これらの結果では，全体的にトピックグループと関連付けに対して，トピックの抽出精度が低くなっている．しかし，不適切なトピックであると指摘されたトピックの中に，他のトピックとすべきと指摘された記事はなく，全て他のトピックと結合すべきという指摘であった．したがって，不適切であると指摘されたトピックは，過剰に分割されているトピックであるといえる．一方で，トピックグループと関連付けは適切に抽出されているので，トピックグループを追うことで，事件の変遷や関連を確認すること可能である．

ここで，不適切と判断されるトピックに共通する傾向を 2 つ述べる．1 つめは，同義語の表記の違いがあげられる．「ミスド，スティックパイのマロンとアップル間違え販売」という記事と「『リンゴとクリ』混同し販売，ミスド自主回収へ」という 2 つの記事は，被験者の評価では同じトピックにすべきと指摘されたが，抽出結果では別のトピックとなっていた．原因として，「リンゴ」と「アップル」，「クリ」と「マロン」という表記の違いにより違うトピックとして抽出された可能性がある．このような問題を解決するためには，同義語や類義語を考慮する必要がある．

もう 1 つは，記事数が多いトピックグループ中のトピックはうまく抽出できない傾向にある．これは，記事数が多くなると記事間の類似度がばらつくため，記事数の少ないトピックと同一の閾値では適切に抽出することが困難であるためと考えられる．したがって，今後，トピック，トピックグループを抽出する際の関連度の閾値の設定方法を見直す必要がある．

今回の実験では 7 名の被験者に対し，時系列のトピックの抽出精度を評価してもらったが，今後，より多くの被験者に時系列のトピックを確認してもらい，さらなる評価が必要である．また，今回の実験では抽出した時系列のトピックについてのみ評価を行っているので，時系列のトピックとして抽出されていない記事について今後は，確認し，本手法で抽出できていない時系列トピックを分析する必要がある．

表 1　被験者ごとの指摘数の集計結果

確認項目	被験者							平均
	A	B	C	D	E	F	G	
不適なトピック数	6	18	4	6	5	0	2	5.9
不適なトピックグループ数	0	1	0	0	0	0	0	0.1
不適な関連付け数	0	0	0	0	6	2	0	1.1

表 2　指摘箇所ごとの被験者数の集計結果

確認項目	指摘人数			総数
	1人	2人	指摘なし	
トピック数	29(10.3%)	6(2.1%)	246(87.5%)	281
トピックグループ数	1(0.6%)	0(0.0%)	166(99.4%)	167
関連付け数	6(5.2%)	1(0.9%)	109(94.0%)	116

表 3　アンケート結果

質問項目	被験者							平均点数
	A	B	C	D	E	F	G	
トピックが適切に分類されていたか	5	5	5	5	4	5	4	4.7
トピックグループが適切に分類されていたか	4	3	4	5	5	5	3	4.1
関連付けは適切であったかどうか	3	4	5	4	5	5	4	4.3

6. おわりに

本研究では，複数新聞記事集合から時系列のトピックを抽出する手法を提案した．また，提案手法を用いて抽出した時系列のトピックを被験者に確認してもらうことで，本手法の精度評価を行った．その結果，抽

出したトピックと連続する日同士のトピックの関連付けに関しては，被験者によって評価にばらつきがあることものの，抽出した時系列のトピックは概ね適切であるとの評価を得た．

今後は，時系列のトピックとして抽出されていない記事についての評価や，同義語への対応，トピックを抽出する際の閾値を選定方法の改良，また，抽出結果を視覚的に確認するためのシステムの構築などを行う予定である．

参 考 文 献

[1] あらたにす，http://allatanys.jp/

[2] 佐藤 吉秀，川島 晴美，佐々木 努，奥 雅博，“単独記事フィルタリングを用いた時系列ニュース記事分類法の提案”，情処研報 2005-NL-168，2005.

[3] 橋本 泰一，村上 浩司，乾孝 司，内海 和夫，石川 正道，“文書クラスタリングによるトピック抽出および課題発見”，社会技術研究論文集 Vol.5 pp.216-226，2008.

[4] 山田 剛一，大熊 耕平，増田 英孝，中川 裕志，“複数新聞記事サイトの横断検索とトピックのドリフト支援システム”，社会技術研究論文集 Vol.1 pp.100-105，2003.

[5] 菊池 匡晃，岡本 昌之，山崎 智弘，“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”，DEWS2008，2008.

[6] Taku Kudo，Kaoru Yamamoto，Yuji Matsumoto，“Applying Conditional Random Fields to Japanese Morphological Analysis，” Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004)，pp.230-237，2004.