ニュース記事閲覧のための複数ウィンドウ方式を用いた特定トピック追跡システムの試作

平田 紀史† 児玉 政幸‡ 伊藤 正都‡ 大囿 忠親‡ 新谷 虎松‡
名古屋工業大学工学部情報工学科† 名古屋工業大学大学院工学研究科情報工学専攻‡

1 はじめに

ポータルサイトや新聞社などのサイトでは，ニュース記事が大量に配信されており，流行やトピックの変化を把握することが困難である．本稿では，様々なトピックが混在する大量の記事集合から，特定のトピックを追跡するシステムについて述べる．

本稿では，トピックを互いに直接関連するイベント，または活動と定義する．イベントは特定の時間，場所で起こった出来事を，活動は共通の関心や目的を持った行動の連鎖を指す．また，トピックは各々のイベント，活動を表すサブトピックによって構成されているとする．そして，トピック追跡とはトピックに含まれるサブトピックを時系列に提示することで，トピックの変化を把握することである．

2 複数ウィンドウからのトピック追跡手法

本稿で提案するトピック追跡手法は，クラスタリングによりサブトピックを得て，サブトピックをさらにクラスタリングし，トピック抽出を行い，トピック追跡を行う手法である．

2.1 記事間の類似度の定義

クラスタリングを行うために，記事間の類似度，クラスタ間の類似度を定義する．類似度計算にはベクトル空間モデルに基づいて，2 つのベクトルを比較することで実現する．文書をベクトルで表現し，ベクトルの各次元に単語を割り当て，各次元の大きさには単語の評価値を割り当てる．評価値は TF-IDF による値とする．類似度を式 (1) に示す．

$$\sigma(I_1, I_2) = \frac{I_1 \cdot I_2}{\|I_1\|\|I_2\|} \quad (1)$$

I_1, I_2 は記事を表す各ベクトルである．式 (1) はコサイン尺度と呼ばれ，2 つのベクトルの各成分が類似するほど大きな値を取る．

また，文を単語に分割するために形態素解析エンジンである MeCab[1]を用いる．MeCab により得られた名詞のみを類似度計算に用いる．

2.2 ウィンドウによるトピック追跡への影響

サブトピックの抽出を行うためのクラスタリングにおいて，記事全てを対象にクラスタリングを行うと計算時間が膨大になる．そこで，計算対象を制限するため，時間的な区間であるウィンドウを設定する．このウィンドウに含まれる記事を対象にそれぞれクラスタリングを行う．理想的なウィンドウとは，図 1 に示すようにサブトピックごとに分割されている時間的な区間である．ここでは，殺人事件が起こり，容疑者逮捕までの記事が記事集合中に存在する場合を考える．この

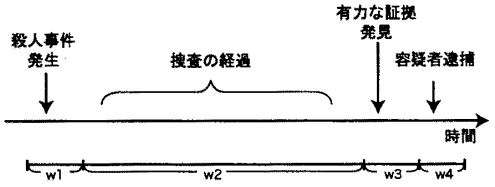

図 1: 理想的なウィンドウ

場合，理想的なウィンドウは w1 から w4 で示すそれぞれの期間となる．しかし，サブトピックを表すイベント，活動の発生は不定期であることが多く，ウィンドウサイズやウィンドウの位置を設定しておくことは困難である．ここで，ウィンドウサイズとは，ウィンドウの示す時間的期間の長さである．また，あるトピックにおいて理想的なウィンドウが得られたとしても，それが別トピックにおいて理想的であるとは限らないという問題もある．

2.3 複数ウィンドウからのサブトピック抽出手法

階層的クラスタリングによりサブトピックを抽出する．また，ウィンドウの影響を軽減するために，スライディングウィンドウ方式 [1] を用いる．スライディングウィンドウ方式とは一定サイズのウィンドウを少しずつ移動する手法である．ウィンドウを移動させることにより，ウィンドウの位置による問題は解決できる．本研究ではさらに，スライディングウィンドウ方式におけるウィンドウサイズを複数設定する．そして，ウィンドウごとにそれぞれクラスタリングを行う．

2.4 トピック抽出手法

複数ウィンドウからのクラスタリングにより，サブトピックの候補となるクラスタが得られる．これらのクラスタを対象にクラスタリングを行うことで，トピックを抽出する．トピック抽出には単一パスクラスタリングの一種である leader-follower 法 [2] を用いる．これにより，複数のウィンドウから得られたクラスタを含むトピックを得ることが可能となる．

2.5 クラスタの選択

トピック内のクラスタからサブトピックに対応したクラスタを選択する．各クラスタに優先度を付け，優先度の高いクラスタから順に選択を行う．そして，クラスタの記事に重複がある場合，後から選択されたクラスタからその記事を除く．

クラスタに含まれる記事が類似しており，多くの記事を含むクラスタを優先的に選択する．まず，各クラスタのまとまりを式 (2) によって表す．

$$s = \frac{1}{N_C}\sum_d \sigma(C, I_d) \quad (2)$$
$$C = \frac{1}{N_C}\sum_d I_d$$

ここで，N_C をクラスタに含まれる記事数，C をクラスタのベクトル，I_d をクラスタに含まれる記事 d を表すベクトルとする．この値 s が大きいほど内容の類似した記事を多く含むクラスタであると判断できる．しかし，s はクラスタに含まれる記事数が少ない方が大きな値を取ることが多い．そ

An Implementation of a Topic Tracking System using Multiple Windows for Browsing News Articles

Norifumi HIRATA, Masayuki KODAMA, Masato ITO, Tadachika OZONO and Toramatsu SHINTANI

† Dept. of Computer Science, Nagoya Institute of Technology

‡ Dept. of Computer Science and Engineering, Graduate School of Engineering Nagoya Institute of Technology

[1] http://mecab.sourceforge.net/

のため，式 (3) で示すように，クラスタに含まれる記事数を影響させるため，$N_C{}^k$ を掛けた値を優先度とする．

$$s' = N_C{}^k \cdot s \tag{3}$$

k は記事数の影響を調節する定数である．

そして，選択されたクラスタを時系列に提示することでトピック追跡が可能となる．

3 提案手法によるトピック追跡実験

3.1 実験手法

毎日 jp[2]が 2007 年 10 月 1 日から 2007 年 11 月 30 日までに配信した 10692 個の記事を対象に，トピック追跡実験を行う．提案手法において組み合わせるウィンドウサイズは 4 日と 16 日とし，ウィンドウの移動量はウィンドウサイズの 2 分の 1 とする．また，式 (3) における k は 1 とした．これらの値は，対象とする記事数を 5000 個とした予備実験により決定した．

実験環境は OS が Windows XP Home Edition，CPU が Pentium M 1.20GHz，メモリが 1GB DDR2 SDRAM，Java の実行環境が JRE1.5.0 である．

毎日 jp の配信する記事の特徴として，記事のタイトルにコロンで区切り，何に対する記事かを示す文字列がある．この文字列をトピック名とみなし，再現率と精度の調和平均である F 尺度 $F_m(K_h, C_k)$ を求める．あるトピックを表す正解のクラスタが K_h，抽出結果のクラスタが C_k である．クラスタリング全体の評価は，正解のトピックごとに最大となる F 尺度を求め，記事数で重み付けした式 (4) で表す値 [3] とする．

$$F_s = \sum_h \frac{\tilde{n}_h}{N} \max_k F_m(K_h, C_k) \tag{4}$$

N は全記事数，$\tilde{n}_h$ は正解集合 K_h に含まれる記事数である．この値が 1 に近いほど良い結果であると判断する．

また，抽出精度を比較するためにウィンドウを設定せず，全範囲から階層的クラスタリングを行う．

3.2 実験結果

トピック追跡実験によって得られたトピックの，式 (4) で表す F 尺度による評価値 F_s は約 0.625 で，実行時間は約 4590 秒であった．また，ウィンドウを設定せず，全範囲からトピック抽出を行うと F 尺度による評価値 F_s は約 0.661 で，実行時間は約 5551 秒であった．実行時間の差は約 1000 秒であり減少した．

パキスタンに関するトピック追跡を行った例を表 1 に示す．表 1 のサブトピックを表すタイトルとは，クラスタのベクトルと最も類似した記事のタイトルであり，配信時間とは，その記事が配信された時間を表す．異なるウィンドウサイズから得られたクラスタが選択されることが確認できる．このトピックを表すクラスタは 71 個の記事から構成されていた．また，同時期のパキスタンに関する主なイベントと活動を表 2 に示す．

3.3 考察

提案手法によるトピック抽出の精度は，全範囲からクラスタリングした場合より減少した．これは，異なるクラスタリング手法を用いた影響であると考えられる．したがって，トピック抽出という観点からすると，精度を低下させることになる．トピック追跡を行うためには，トピック抽出ができていることが前提となるため，トピック抽出の精度は重要となる．表 1 の例では，別のトピックを示すクラスタにパキスタンに関する記事が含まれることがあった．

ウィンドウサイズを複数設定することにより，サブトピックに記事の時間の範囲を可変とすることが可能となった．本来のサブトピックに近いクラスタを得ることができたと言える．表 1 と表 2 を比較すると，実験から得られたサブトピックは細かなイベントについても追跡を行っている．これは，式 (3) の k を変化させることで，選択するクラスタを調節できると考えられる．

提案手法では，より多くのウィンドウサイズを設定した方が，サブトピックに対応したクラスタを得る可能性が増加する．しかし，多くのウィンドウサイズを設定するほど，計算時間が増加する．これは，対象をウィンドウに分割するという本来の利点を失う．

[2]http://mainichi.jp/

表 1: 複数ウィンドウからのトピック追跡結果

配信時間	ウィンドウサイズ	サブトピックを表すタイトル
10 月 19 日	4 日	史上最悪級のテロ　ブット氏、間一髪
11 月 4 日	4 日	ムシャラフ大統領、全土に事実上の戒厳令
11 月 5 日	4 日	「ムシャラフ大統領は辞職すべきだ」元...
11 月 7 日	16 日	ブット元首相、ＮＳＣ最高幹部と会談
11 月 10 日	4 日	軟禁解かれたブット元首相が外出　デモ...
11 月 13 日	4 日	ブット元首相のデモ行進は不許可
11 月 19 日	16 日	大統領選の違憲審理再開　最高裁が５件...
11 月 20 日	4 日	来年１月に総選挙実施　候補者届け出締...
11 月 23 日	4 日	シャリフ元首相、２６日にもサウジから...
11 月 30 日	16 日	ムシャラフ大統領、非常事態解除へ

表 2: パキスタンに関する主なイベントと活動

発生時間	主なイベントと活動
10 月 18 日	ブット元首相が帰国
10 月 19 日	ブット元首相を狙ったテロ発生
11 月 3 日	非常事態宣言発令
11 月 9 日	ブット元首相の自宅軟禁
11 月 10 日	ブット元首相の軟禁解除
11 月 13 日	滞在先でブット元首相を再び軟禁
11 月 26 日	シャリフ元首相が帰国
11 月 30 日	非常事態宣言解除

4 おわりに

本稿ではウィンドウの影響を軽減してトピック追跡を行う手法を提案した．ウィンドウサイズを複数設定し，スライディングウィンドウ方式を用いる手法である．そして，実際の記事を対象にトピック追跡を行い，サブトピックに対応したクラスタを得られることが確認できた．

実際にトピック追跡システムを使用する場合は，事前にすべての記事を対象にトピック追跡を行っておき，興味のあるトピックを検索して，トピック追跡を行うことになる．

問題点として，ウィンドウサイズの設定する数により計算時間が増加すること，トピック抽出の精度が若干減少することがある．今回の実験では階層的クラスタリングと leader-follower 法を用いたが，別のクラスタリング手法を用いた場合についても検討する必要がある．

参考文献

[1] Thorsten Brants and Francine Chen : A System for new event detection, Proc. ACM SIGIR, pp.330-337, 2003.

[2] 岸田和明：大規模文献集合に対して階層的クラスタ分析法を適用するための単連結法アルゴリズム［短報］, Library and Information Science, No.47, pp.27-38, 2002.

[3] 白砂健一, 小山聡, 田島敬史, 田中克己 : Web の構造情報とプロファイル抽出を用いたオブジェクト識別, 電子情報通信学会第 17 回データ工学ワークショップ, 2C-i7, 2006.