D-042

可変クラスタリングウィンドウによるトピック追跡システムの試作
An Implementation of the Topic Tracking System Using Variable Clustering Windows

平田紀史 † 大囿忠親 †† 新谷 虎松 ††
Norifumi Hirata, Tadachika Ozono, Toramatsu Shintani

1 はじめに

現在，ポータルサイトや新聞社，通信社などのサイトで，大量のニュース記事が配信されている．このような状況下では興味のある分野における流行や変化を把握することは困難である．この問題は記事からトピックを抽出し，トピック追跡を行うことでの解決が図られる．このようなトピック分析は TDT(Topic Detection and Tracking)[1][2] という分野で広く扱われている．

記事を時系列で区切り，それらをクラスタリングすることで，トピックの抽出と追跡が可能になる [3]．従来手法では，クラスタリングにおけるウィンドウを固定しており，トピックの抽出が失敗することがある．ウィンドウとは対象となる記事の時間的な区間のことである．この問題は，トピックによってそれを報じる記事の頻度が異なるために起こる．

本稿ではタイムスタンプを持つ記事を対象にクラスタリングを行い，トピック追跡を行うための理想的なウィンドウの決定手法を示す．

2 可変クラスタリングウィンドウによるトピック抽出

2.1 理想的なウィンドウ

トピック追跡における理想的なウィンドウは，記事の内容に対応して決定されたウィンドウである．図 1 は，例えばオリンピックのような，特定のトピックにおける時間と記事の関係を表したものである．t_1 から t_2 は開催地の決定やそれまでの準備期間，t_2 から t_5 は開幕中の期間，t_5 から t_6 は閉幕後の期間である．t_1 から t_2 のような記事の少ない期間では長期的なウィンドウが理想的なウィンドウとなる．t_2 から t_5 のように，記事数が増えた場合には変化に対応するために短期的なウィンドウが理想的なウィンドウと考えられる．このトピックについて図 2 の t_1 から t_5 の様にウィンドウを一定間隔にしてしまうと，トピックの変化に対応できないため，適切なクラスタを抽出することができなくなると考えられる．このように，トピック追跡を適切に行うためには，トピックの変化に対応したウィンドウの決定が重要である．

2.2 クラスタリング手法と類似度

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．記事を表すベクトルは，各次元に索引語を割り当て，各成分に索引語の評価値を割り当てたものとする．この索引語の評価値は tf-idf の値とする．また，クラスタに含まれる記事のベクトルの平均ベクトルをクラスタのベクトルとする．

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

I_1, I_2 は各ベクトルを表す．(1) 式に表すコサイン尺度は 2 つのベクトルの角度が小さい場合に大きな値を取る．

クラスタリング手法としては，非階層的手法で代表的な k-means 法を使用する．

† 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 工学先攻科 情報工学専攻

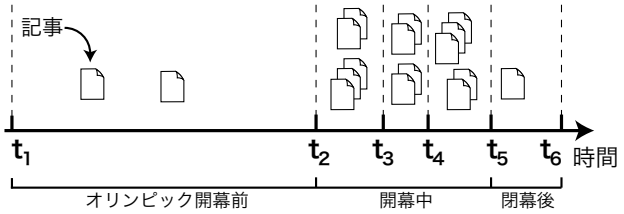

図 1: 理想的なウィンドウ

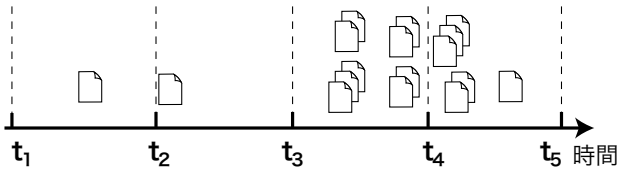

図 2: 一定間隔のウィンドウ

3 実験

クラスタリングウィンドウを固定した場合，ウィンドウによって追跡結果がどのように変化するのかを観察するための実験を行う．

3.1 実験方法

クラスタリングウィンドウを 1 日，3 日，5 日間隔とした場合でトピック追跡を行う．共同通信社[1]の 2006 年 9 月 1 日から 9 月 15 日までに配信された社会のジャンルの記事を対象とする．クラスタリング手法は k-menas 法を用いる．

トピック追跡は (1) 式で得られるクラスタ間の類似度によって行う．クラスタ間の類似度が閾値以上ならば同一トピックとし，閾値以下ならば新たなトピックとする．

3.2 評価方法

対象記事の期間内に起こった主なトピックとして，悠仁親王の誕生に関するトピックについて比較する．図 3 に期間内に起きたこのトピックに関する主なイベントを示す．

内容の変化に対応したウィンドウを理想とするので，理想的なウィンドウは図 3 中の $w1$ から $w4$ で示す期間とする．

3.3 実験結果

ウィンドウを 1 日間隔とした場合のトピック追跡の結果の一部を図 4 に，3 日とした場合を図 5 に，5 日とした場合を図 6 に示す．図中の単語はクラスタ内で tf-idf の値が大きかったものである．

図 5 の結果では 9/1 から 9/7 の間にもこのトピックに関する記事は存在するが，トピック追跡ができていない．これは，このトピックに関するクラスタが，ウィンドウ内に複数存在したためであると考えられる．

固定間隔のウィンドウが理想的なものではないので，トピック追跡ができていないが，図 5 のクラスタと図 6 のクラスタの一部を組み合わせると，図 7 に示すようにトピックの内容

[1] http://www.kyodo.co.jp/

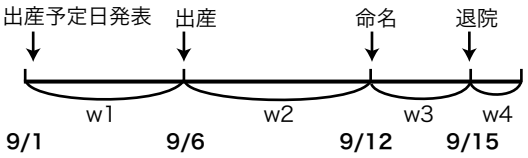

図 3: 期間内の悠仁親王誕生に関するイベント

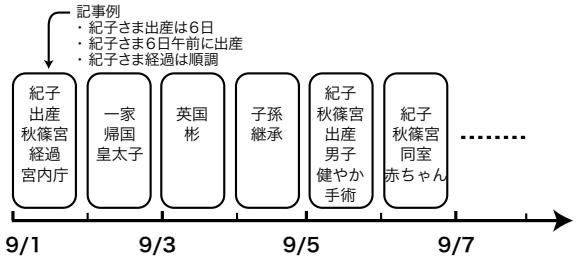

図 4: 1 日間隔のウィンドウでの実験結果

図 5: 3 日間隔のウィンドウでの実験結果

図 6: 5 日間隔のウィンドウでの実験結果

の変化に対応できる．このように，異なるウィンドウから得られたクラスタの組み合わせによって，トピック追跡を行うことが可能になる．

4　トピック追跡におけるクラスタの選択手法

クラスタの自己類似度に基づく可変クラスタリングウィンドウの決定手法を示す．複数の間隔のウィンドウを適切に組み合わせることによって，より正確なトピック追跡が可能になる．

4.1　自己類似度

クラスタの選択時に使用する自己類似度を定義する．これはクラスタ内の記事の分散を示す指標である．クラスタ i のベクトルとクラスタ i 内の記事 j のベクトルとのユークリッド距離を r_{ij} と表す．このとき，クラスタ i の自己類似度を以下のように定義する．N はクラスタ内の記事数である．

$$s_i = \frac{\sum_{j=1}^{N} r_{ij}}{N} \tag{2}$$

4.2　クラスタの選択手法

異なるウィンドウによるクラスタ間の自己類似度の差が小さければ，より大きなウィンドウによるクラスタを選択する．図 8 にクラスタ選択の様子を示す．

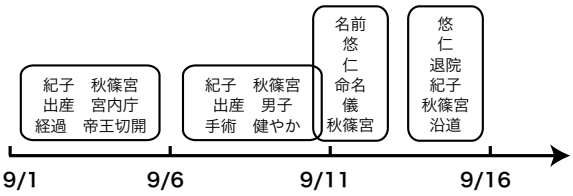

図 7: 3 日間隔と 5 日間隔のウィンドウの組み合わせ

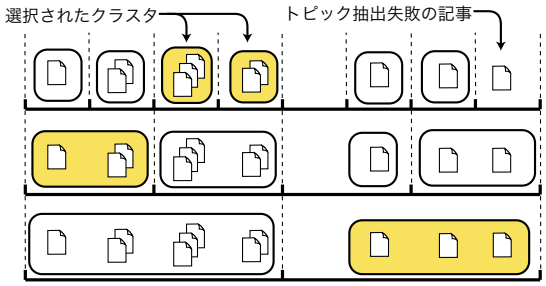

図 8: クラスタ選択の様子

1. ウィンドウ間隔を複数設定し，各ウィンドウにおいてクラスタリングを行い，トピックごとにクラスタを分類
2. 各クラスタにおいて自己類似度を計算
3. 最も細かいウィンドウのクラスタ c_{small} と，同じ時間帯の 1 つ大きなウィンドウのクラスタ c_{big} の自己類似度の差を計算
4. その差が閾値以上なら c_{small} を選択
5. c_{big} のウィンドウよりも大きなウィンドウが存在しなければ c_{big} を選択
6. c_{big} を c_{small} として，すべての時間帯でクラスタが選択されるまで 2 以降の繰り返し

図 8 の一番右上の記事に示すように，ウィンドウによってトピックに関する記事をクラスタとして抽出できない可能性がある．同様に，記事数の少ない長期的なトピックについても抽出できない可能性がある．したがって，大きさの異なるウィンドウでクラスタリングを行うことが必要となる．

5　おわりに

本稿では，ウィンドウを一定間隔とした場合のトピック追跡を行った．一定間隔の場合，トピックにより適切なウィンドウが異なり，トピック追跡を行えない場合があることを確認できた．そして，トピック追跡におけるクラスタの選択手法を提案した．今後は，より長期間のデータを用いたトピック追跡を行う必要がある．

参考文献

[1] J. Allan, J. Carbonell, G. Doddington, J. Yamron, Y. Yang, “ Topic detection and tracking pilot study final report ”, Proc. of the DARPA broadcast news transcription and understanding workshop, pp.194-218, 1998.

[2] NIST(National Institute of Standards and Technology) http://www.nist.gov/speech/tests/tdt/

[3] 大川原雄也, 大囿忠親, 新谷虎松“ 言語モデルに基づく階層型クラスタリングを用いたトピック分析 ”, 情報処理学会第 69 回全国大会, 2007.