ニュース記事の話題分岐を時系列で追跡可能な可視化法

森 幹彦

京都大学学術情報メディアセンター

1 はじめに

World Wide Web の文書が爆発的に増加している要因の一つに，オンラインニュースのサイト数の増加やブログや日記の普及による公開数の増加がある．このようなニュース記事から様々な事件やイベントに関して調べる場合，これまでの経緯，現在の状況，今後の展開を見つけ出したいという要求がある．

従来，事件などの全体像を知りたいという利用者の要求に対して，キーワード検索によって提示される記事群から前後関係を読み取り，手作業または頭の中で関連づけの作業を利用者が行う方法が多かった．したがって，特定の事件に途中から興味を持った者にとって，大きな事件の全体像を掴むことは難しく，事件の初期から注目している者にとっても，後から系統的に思い起こすのが困難であった．

そこで本稿では，文書群として時間とともに変化する話題を扱うニュース記事を対象にして，記事の話題の分岐や収束に注目できるクラスタリング法を示し，それをもとにした可視化法を提案する．

2 クラスタリング法

2.1 ニュース記事の類似度

ニュース記事を bag-of-words として扱い，あるニュース記事 d_i は，そこに含まれる語の重みを用いて文書ベクトル $d_i = (w_{i1}, ..., w_{ij}, ..., w_{in})$ として表す．ここで，w_{ij} は i 番目の記事における j 番目の単語の重みである．また，記事群を一定期間ごとに分割し，それぞれの期間を t とする．i 番目の記事が期間 t に現れ，i 番目の記事が期間 $t+a$ に現れたときの類似度 $s(d_i, d_k)$ は，忘却を考慮して次の式で算出する．

$$s(d_i, d_k) = \lambda^a \frac{d_i \cdot d_k}{\sqrt{|d_i||d_k|}} \quad (1)$$

ここで，λ は忘却定数を表し，$0 < \lambda \leq 1$ である．

2.2 話題クラスタ

ある話題に関する記事群は互いに類似度が高いと考え，記事群から類似度の高い記事群を抽出することで各話題を切り取ることを検討する．ここで，ある話題に関係する話題クラスタと呼ぶことにする．また，話題は継続的に語られると考える．ただし，ある時点において1つの話題であっても，時間が進むと複数の異なる話題とした方が適切になることがある．このような場合を話題の分岐と呼び，話題クラスタの分割を行う．一方，今まで複数の話題として扱っていた内容の記事群を1つの話題とした方が適切になることもある．これを話題の収束と呼び，話題クラスタの併合を行う．

本クラスタリング法では，記事が追加されるごとに既存クラスタの類似性をもとに逐次的に処理をする．i 番目の話題クラスタ D_i における j 番目の記事を d_{ij} とする．記事 d_{ij} が期間 t の記事であるなら，$j < k$ である d_{ik} は期間 t もしくはそれ以降の期間の記事とする．

記事を時系列に直列に並べた記事列から古い順に記事を取り出すとき，新たに取り出した記事を d^{new} とすると，d^{new} の所属を決めるために各 D_i の d_{ij} に対して式(1)を計算する．次の式が成り立つとき，その d_{ij} が所属する D_i に d^{new} も所属するとする．

$$s(d^{new}, d_{ij}) > \theta_s \quad (2)$$

ここで，θ_s は閾値で $0 < \theta_s \leq 1$ の範囲とする．

話題クラスタ D_i の重心 Dc_i は，D_i の記事数を $|D_i|$ として次のように表せる．

$$Dc_i = \frac{1}{|D_i|} \sum_{j=1}^{|D_i|} d_{ij} \quad (3)$$

Timeline Visualization for Tracking Topic Drifts from News Articles,
Mikihiko Mori, Academic Center for Computing and Media Studies, Kyoto University

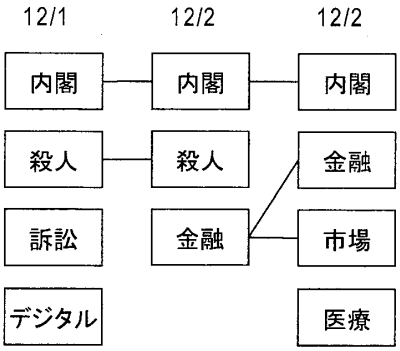

図1　話題クラスタの表示例

また，話題クラスタの分散 ${\sigma_{D_i}}^2$ は，次の式のようになる．

$$ {\sigma_{D_i}}^2 = \frac{1}{|D_i|}\sum_{i=1}^{|D_i|}(d_i - Dc_i)^2 \qquad (4) $$

ある記事が2つの話題クラスタ D_x と D_y のどちらに対しても式 (2) が成り立つなら，D_x と D_y を併合の対象とする．この場合，次式を満たすときに2つの話題クラスタを併合する．

$$ a\sigma_{D_x} + b\sigma_{D_y} > |Dc_x - Dc_y| \qquad (5) $$

ここで，a, b は任意の係数である．

一方，期間が遷移する時点で分割を試み，分割がもっともらしい場合はそれ以降は別の話題クラスタとする．文書群の分割には様々な手法が提案されているが，本稿は k-means 法を適用することにする．分割数は2とし，分割後のクラスタに対して式 (5) で評価する．

最後に，長い期間，すなわち期間 t から期間 $t+\tau$ (τ は任意の定数) までに新しい記事が追加されない話題クラスタを終了した話題とみなす．忘却を考慮した類似度の計算によって，話題の終了を定義しなくても新規の記事は追加されないため，実質的には問題にならない．

3　可視化法

本可視化法では，2次元平面上に横軸を時間軸をとり，期間 t を1ずつ進めたときのそれぞれの期間での話題クラスタを縦方向に表示する．継続的に記事の加わっているクラスタと記事の追加がなくなったクラスタを区別するため，期間 t の記事が加わったクラスタのみを期間 t の位置に表示する．このとき，$t-1$ から継続している話題クラスタであれば，クラスタ間を線で結ぶ．1期間を1日とした場合の3期間分の表示例を図1に示す．表示例のように話題の分岐も表現できる．

4　関連研究

Allan らは，ニュースデータを話題ごとに分割して同一の話題の再出現を追跡する TDT (Topic Detection and Tracking) を提唱した [1]．本研究は TDT のひとつであるが，話題内の文書間の関係や話題の分岐と収束に関する計算に主眼をおいている．TDT 関連の研究として，短期間に特定語が大量に発生する現象（バーストと呼ばれている）をもとに話題の抽出を行う BlogWatcher[2] や，トピックのバーストと支持率などの別の時系列データとの相関を求める研究 [3] があるが，本研究では話題の時間的な変遷に焦点をあてる．一方，文書間の関係を可視化するインタフェースとして DualNavi[4] があるが，時間的な変遷を明示的に示すに至っていない．

5　おわりに

本稿では，ニュース記事における話題を時間の経過とともに分岐や収束が起こるものと考え，このような話題の変化に追従できるようなニュース記事のクラスタリング法を提案した．逐次的にクラスタに記事を追加することと，一定期間ごとに分割や併合を検討することにより，話題の分岐や収束が表される．さらに，このクラスタを用いた可視化表現を提案した．これによりニュースの背景や前後関係の把握や話題の追跡が容易になることを期待している．

参考文献

[1] Allan, J., Papka, R. and Lavrenko, V.: *On-line New Event Detection and Tracking*, ACM SIGIR, pp. 37–45 (1998).

[2] Nanno, T., Suzuki, Y., Fujiki, T. and Okumura, M.: *Automatic Collection and Monitoring of Japanese Weblogs*, WWW2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics (2004).

[3] 張 一萌, 何 書勉, 小山 聡, 田島敬史, 田中克己: 時系列データに意味的に関連するニューストピックの発見, 日本データベース学会 Letters, Vol.5, No.1, pp. 133–136 (2006).

[4] Takano, A., Niwa, Y., Nishioka S., Iwayama, M., Hisamitsu, T., Imaichi, O., and Sakurai, H.: *Associative Information Access Using DualNAVI*, ICDL'00, pp.285-289 (2000).