文書群からの時間的変化する話題の抽出

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 (以降，Web と呼ぶ) 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．すなわち，次のような要求ということができる：

- 特定の話題の時間的変遷：特定の話題に注目して，時間遷移による話題の内容の変遷を知りたい．
- 話題の分岐・収束：内容の変遷として話題の分岐や収束の様子を知りたい．

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．例えば，2002 年から 2003 年にかけて起きた高病原性鳥インフルエンザについて，事件の発生から終了までを系統的に調べることが考えられる．

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文書クラスタリング法を提案する．

2 ニュース記事の話題

本研究では，ニュース記事を bag-of-words として扱う．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 $d_i = (w_{i1}, w_{i2})\ (i = 1, ..., n)$ 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．このとき，i 番目の記事と k 番目の記事の類似度 $s(d_i, d_k)$ は，文書ベクトルの余弦 $s(d_i, d_k) = d_i \cdot d_k$ とする．また，記事群 D の重心 D_c は，$D_c = (\sum_{i=1}^{n} d_i)/n$ (n は D に含まれる記事の数) と表せる．記事群の類似度は，重心 D_c が代表すると考えて重心の余弦で求める．

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．ある期間 t で内容が類似する記事群に分けることを考える．n 個の記事群 $D(i,t)\ (i = 1...n)$ に分けられる場合，$D(i,t)$ の内容で表しているものを話題と見なし，話題 $T(i,t)$ とする．

ある時点において，ある 1 つの話題として扱えるニュース記事群も，時間が進むと複数の異なる話題として扱った方が適切であるようになることがある．また，時間が進むと今まで複数の話題として扱っていた内容の記事群を 1 つの話題として扱った方が適切になることもある．期間 t に m 個の話題 $T(i,t)\ (i = 1, ..., m)$ があり，期間 $t+1$ に n 個の話題 $T(j,t+1)\ (j = 1, ..., n)$ があったとする．ある話題が t から $t+1$ で変化しないとすることは，話題 $T(i,t)$ を引き継いだとする話題 $T(j,t+1)$ の間では，$D_c(i,t)$ と $D_c(j,t+1)$ の位置がほとんど変わらないことである．一方，期間 $t+1$ において $D_c(i,t)$ と十分に類似する記事群が複数ある場合，話題 $T(i,t)$ は分岐したと考える．また，期間 t において $D_c(j,t+1)$ と十分に類似する記事群が複数ある場合，話題 $T(j,t+1)$ に収束したと考える．さらに，$D_c(i,t)$ と十分に類似する記事群の重心がなかった場合，話題 $T(i,t)$ は終了したと考えるのが適当であろう．

このように，期間 t と $t+1$ において話題が対応付けられる模式図を図 1 に示す．実際には，記事群に含まれる語数分の次元があるが，簡便のために 3 次元で表現している．

Extraction of Topics Drifted with Time from Documents, Mikihiko Mori, Academic Center for Computing and Media Studies, Kyoto University

3 ニュース記事のクラスタリング法

一般的に，文書群に含まれる話題が k 個であることが分かっている場合，k-means 法を用いることができる．しかし，本研究が扱いたい，ある期間 t における全ニュース記事に含まれる話題の数は未知である．このような不定数のクラスタリング法として x-means 法が提案されている [3]．x-means 法は，十分に小さい数 k で始める k-means 法で分割した後に，各クラスタを $k=2$ の k-means 法で分割を試み，分割が適当でないと判断されるまで繰り返す方法である．このときの停止基準として BIC (Baysian Information Criterion) を用いている．

各期間における全記事に対し，x-means 法を適用して複数のクラスタに分割する．このとき，期間の区切り方を十分に小さくすると，期間 t と $t+1$ の間でクラスタ数に大きな変化はなく，それぞれの期間の重心同士で対応する重心は近いと考えられる．ただし，分岐や収束の起きている重心を考慮しなければならない．そこで，期間 t と $t+1$ における重心を頂点とする二部グラフを考える．t 側のそれぞれの頂点からは，$t+1$ の頂点で一番近い，すなわちもっとも類似する重心に辺を結ぶ．同様にして，$t+1$ 側から t へ辺を結ぶ．この手続きの結果，いかなる辺も隣接していない辺で結ばれた頂点同士，すなわち重心同士は対応付けられた話題と考える．また，t 側の頂点を共有する辺で結ばれた $t+1$ 側の頂点との対応関係を分岐と呼び，逆に $t+1$ 側の頂点を共有する対応関係を収束と呼ぶことにする．ただし，この方法では，もっとも類似する重心を必ず見つけようとするため，類似するとはいえない重心を対応付けるかもしれない．そこで，あらかじめ決めておく閾値 θ 以上の類似度を対象とする．

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]．本研究は広義の TDT と考えることもできるが，TDT がデータからの話題の判別を主眼におくのに対し，本研究では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く．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 [4] がある

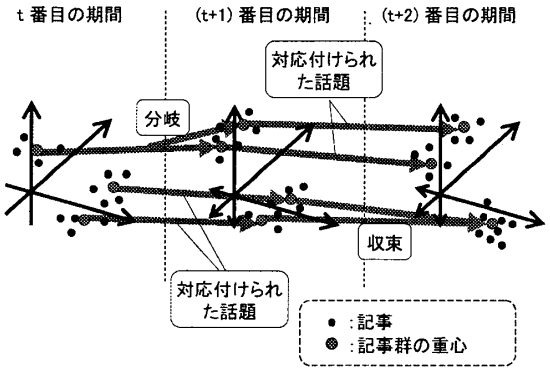

図 1　クラスタの重心としての話題とその対応付け

が，本研究では話題の時間的な変遷に焦点をあてる．

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．この方法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．今後は，実際のニュース記事を対象として有用性の検証をしていきたい．

参考文献

[1] Allan, J., Papka, R. and Lavrenko, V.: *On-line New Event Detection and Tracking*, Proc. of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, 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] Pelleg, D., Moore, A.: X-means: Extending K-means with Efficient Estimation of the Number of Clusters, Proceedings of the Seventeenth International Conference on Machine Learning, pp. 727–734 (2000).

[4] 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).