2F2-2

ニュース記事における時間変化する話題の抽出

Extraction of Drifted Topics with Time from News Articles

森 幹彦[*1]
Mikihiko Mori

[*1]京都大学 学術情報メディアセンター
Academic Center for Computing and Media Studies, Kyoto University

Although readers of news articles can obtain precise information, a series of pages to which is related mutually can be hardly found out. The readers want to see that a topic separate and that some topics are confluent. They also need to see how this topic is now and which topics are releted this topic.

In this paper, I propose novel clustering method for news articles. The method clusters articles serially to topic clusters, divides a cluster, and merges some clusters appropriately.

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．また，2005 年から 2006 年にかけて起きた構造計算書偽造問題を調べることも考えられる．偽造発覚をきっかけにして，次第に設計会社への話題と住人への対応についての話題などに分岐していたが，行政による会社への処分と住人への対策によって行政の対応という話題の収束につながった．しかし，今後も裁判などの話題で分岐や収束を繰り返すことが予想される．

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クラスタリング法を提案する．

2. 提案手法

2.1 ニュース記事の類似度

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 $d_i = (w_{i1}, ..., w_{ij}, ..., w_{in})$ 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．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，文書ベクトルの内積を正規化した

$$s(d_i, d_k) = \frac{d_i \cdot d_k}{\sqrt{|d_i||d_k|}} \tag{1}$$

とする．

2 つの記事を取り出したとき，式 (1) に示した単純な類似度で同じ値になったとしても，記事の直感的な関連性からすると異なるはずである．すなわち，時間的に離れた記事間の類似度は，近い時期に書かれた記事同士の類似度よりも小さくなると考えるのが自然で，そのために期間をまたがる類似度は忘却を考慮する．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 にある記事と期間 $t + a$ にある記事では，Ebbinghaus の忘却曲線 [2] を用いて類似度を以下の計算式で算出する．

$$s(d_i, d_k) = \lambda^a \frac{d_i \cdot d_k}{\sqrt{|d_i||d_k|}} \tag{2}$$

ここで，λ は忘却定数を表し，$0 < \lambda < 1$ である．実際には，$a < a_0$ のときには $\lambda^a = 0$ として計算対象を少なくする．

2.2 話題の抽出

ある話題に関する記事群は，互いに類似度が高いと考えられる．また，多くの話題は継続的に語られる．そこで，記事群から類似度の高い記事群を抽出し，それをある話題に関係する記事群とする．このような記事群を話題クラスタと呼ぶことにする．

さらに，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．これを話題クラスタに置き換えるなら，話題の分岐は話題クラスタの分割であり，話題の収束は話題クラスタの併合であると考えることにする．適宜，話題クラスタの分割や併合を行うことで話題の変化に適応でき，それぞれの時点で適切な話題クラスタを維持することが可能になる．

したがって，基本的な話題クラスタの生成法としては，既存クラスタとの類似性をもとに逐次的に処理していくことを考える．すなわち，どれかの話題クラスタの記事に類似していれば，新たにその話題クラスタの一員にし，もしどの話題クラスタと

連絡先: 森 幹彦，京都大学学術情報メディアセンター，〒606-8501 京都市左京区吉田二本松町，075-753-9052

も類似していなければ新たにその記事のみが所属する話題クラスタを生成する．

i 番目の話題クラスタ D_i，D_i の j 番目の記事を d_{ij} とする．このとき，記事 d_{ij} が期間 t の記事であるなら，$j \leq k$ である d_{ik} は期間 t もしくはそれ以降の期間の記事である．

記事を時系列に直列に並べた記事行列を Ds とし，Ds の記事を古い順に取り出すことを考える．Ds から新たに記事 d^{new} を取り出したとき，d^{new} の所属を決めるために，各 D_i の d_{ij} に対して式 (2) 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．

$$s(d^{new}, d_{ij}) > \theta_s \tag{3}$$

ここで，θ_s は閾値を表し，$0 \leq \theta_s \leq 1$ の任意の値とする．

話題クラスタ D_i の重心 Dc_i は，

$$Dc_i = \frac{1}{n}\sum_{j=1}^{n} d_{ij} (n \text{ は } D \text{ に含まれる記事の数}) \tag{4}$$

と表せる．また，話題クラスタの分散 V_D は，

$${\sigma_D}^2 = \frac{1}{n}\sum_{i=1}^{n} (d_i - Dc_i)^2 \tag{5}$$

である．

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 (3) 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．

$$a\sigma_{D_x} + b\sigma_{D_y} > |Dc_x - Dc_y| \tag{6}$$

ここで，a, b は任意の係数である．

一方，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場合を想定すると，ある話題クラスタ D_i が潜在的に分割の可能性を考慮しなければならない．そこで，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 (6) で評価する．

最後に，長い期間，すなわち期間 t から期間 $t+a$（a は定数）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実質的には変わらないが，新しい記事の追加の見込みがないにもかかわらず残っていると，継続中の話題との区別ができなくなるため定義する．

以上をまとめると，ある期間 t について次のような手順を実行する．

1. Ds から順次記事を取り出す．取り出した記事 d^{new} とする．
2. d^{new} と各 D_i の各 d_{ij} との類似度を計算し，式 (3) を満たす d_{ij} があったときに，d^{new} を D_i に所属させる．
3. もし複数の話題クラスタに所属するとされたなら，併合対象としてその話題クラスタに印を付ける．
4. 期間 t が終了するなら，
   （a）話題クラスタの重心を求める．
   （b）話題クラスタの分割を試みる．
   （c）併合対象の話題クラスタがあるなら，併合を試みる．
   （d）次の期間 $t+1$ に対し，1. から始める．

3. 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[3] 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 [5] 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[6] があるが，時間的な変遷を明示的に示すに至っていない．

4. 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，話題の分岐や収束を再現している．また，終了した話題と継続している話題の判別も可能である．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] Ebbinghaus, H.: *Memory: A Contribution to Experimental Psychology*, Dover Publications (1987).

[3] 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).

[4] 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).

[5] 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).

[6] Takano, A., Niwa, Y., Nishioka S., Iwayama, M., Hisamitsu, T., Imaichi, O., and Sakurai, H.: Associative Information Access Using DualNAVI. Kyoto International Conference on Digital Libraries 2000 (ICDL'00), pp.285-289, IEEE Computer Society (2000).