F-051

SR 法を利用した文書ストリームのホットトピック抽出

Extracting Hot Topics In Doucument Streams Using the SR-method

東野正行 † 木村昌弘 †,†† 斉藤和巳 †††
Masayuki Higashino Masahiro Kimura Kazumi Saito

1 はじめに

近年の情報通信技術の発展により、インターネット上には様々な文書ストリームが大量に存在するようになった。そこで文書ストリームからホットトピックを自動抽出する有効な技術が求められている。

Kleinberg[1] は、バースト度に基づいて、文書ストリームデータからホットトピックとそのホット期間を抽出する手法を提案している。また Zhao ら [4] は、電子メールデータのような、送信者と受信者などの社会的関係性情報をも含む特殊な文書ストリームデータに対して、文書ネットワークを含む複数のネットワークを分析しそれらの結果を統合することにより、主要なイベントを精度良く抽出する手法を提案している。ところで、Saito ら [2] は、ネットワーク内の密結合するコア部を効率良く抽出する SR 法を提案し、ブログのトラックバックネットワークから、主要トピックや SEO スパムを検出することに対して成果を上げている。本論文では、通常の文書ストリームデータのホットトピックとそのホット期間を、より高精度に同定する手法の開発を目指して、文書ネットワーク構造、SR 法、およびバースト度に基づいた手法を提案する。実文書ストリームデータを用いた実験により、提案法の Kleinberg 法に対する有効性を示す。

2 Kleinberg 法

2.1 バースト度

全期間で文書が一様に出現すると仮定した通常状態時の文書出現確率を $p_{k,0}$、あるトピック k に関する文書が生成されやすいバースト状態においての文書出現確率を $p_{k,1}$ とし、それぞれの二項分布尤度比でバースト度を定義する。各時刻 t に出現した総文書数 N_t に対し、トピック k に関連した文書数が $n_{k,t}$ となる確率を通常状態では $P_N(n_{k,t};p_{k,0})$、バースト状態では $P_N(n_{k,t};p_{k,1})$ とする。トピック k の期間 $[t_1,t_2]$ におけるバースト度は以下の式で求められる。

$$B(t_1,t_2;k)=\frac{\prod_{t=t_1}^{t_2}P_{N_t}(n_{k,t};p_{k,1})}{\prod_{t=t_1}^{t_2}P_{N_t}(n_{k,t};p_{k,0})} \quad (1)$$

2.2 抽出アルゴリズム

Kleinberg 法では、ホットトピック文書群とそのホット期間を、バースト度に基づいて次のように抽出する。

単語 w_k を含む文書群をトピック k に関する文書群候補と考え、各文書群候補のバースト度が最大となるホット期間 $[t_1,t_2]$ を求める。バースト度の大きい順に単語 w_k をランキングして、ホットトピック文書群とそのホット期間を抽出する。

† 龍谷大学大学院 理工学研究科 電子情報学専攻
†† 龍谷大学 理工学部 電子情報学科
††† 静岡県立大学 経営情報学部

3 提案法

本研究では、まず文書ストリームデータから文書ネットワーク構造を構築する。次に，拡張 SR 法によりホットトピック文書群を抽出する。そして，式 (1) で定義されるバースト度に基づいてそのホット期間を求める。

3.1 文書ネットワーク構築

文書データを TF-IDF 表現で表現し、コサイン類似度で文書間類似度を定義する。そして各文書をノード、文書間の類似度を重み付きリンクとした、重み付きネットワークを構築する。ただし、設定された閾値 r 未満の重みをもつリンクは、リンク重みを零とする。

3.2 ホットトピック抽出

SR 法を拡張した手法を用いて、文書ネットワークからホットトピック抽出を行う。

我々は、重み付きネットワークにおいて重み付きリンク密度の高い部分（コア部）にホットトピックが存在すると考え、そのようなコア部を抽出する。文書ストリームの文書全体の集合を $S=\{1,\cdots,N\}$ とする。$A=(a_{i,j})$ を構築した文書ネットワークの隣接行列とする。ここに、$a_{i,j}$ はノード i とノード j 間のリンク重みである。ノード集合 $C\subset S$ に対し、そのリンク密度を以下の式で定義する。

$$G(C)=\frac{1}{2}\sum_{i\in C}\sum_{j\in C}\frac{a_{i,j}}{|C|}$$

$|C|$ は集合 C の要素数である。$G(C)$ を最大にするノード集合 C を探索する。しかし、単純な網羅的探索では大規模ネットワークにおいて組み合わせ爆発が起こる。そこで、緩和問題、量子化問題を解くことによりコア部 C を推定する [2]。

まず、リンク密度が最大となるコア部 C_1 を抽出する。次に、C_1 内のすべてのリンク重みを零とした重み付きネットワークを構築し、リンク密度が最大となるコア部 C_2 を抽出する。以下同様の手順を繰り返すことによって、K 個のホットトピック文書群

$$\{C_k;k=1,\cdots,K\}$$

を抽出する。

3.3 アルゴリズム

すなわち、抽出するホットトピック数 K が与えられたとき、抽出アルゴリズムは以下となる。

1. 文書を TF-IDF 表現で表現し、コサイン類似度を求め、閾値 r に基づいて文書ネットワークを構築する。
2. 拡張 SR 法によって K 個のホットトピック文書群 $\{\ C_k;\ k=1,\cdots,K\ \}$ を抽出する。
3. 式 (1) で定義されるバースト度に基づいて、抽出した各ホットトピック文書群 C_k のホット期間 $[t_{k,1},t_{k,2}]$ を求める。

4　実験評価

大規模実データで、提案法の性能を評価した。

4.1　評価データ

実験では、トピックが付与された 1994 年の 1 月から 6 月に記載された毎日新聞の国際面記事データ [3] を用いた。総文書数は 2695、語彙総数は 18070、トピック数は 45 個であった。

記事データには各記事が実際に掲載された日時、各記事における出現単語 ID とその出現頻度が記載されている。

4.2　評価尺度

真のホットトピック群を $\{H_\ell\ ; 1 \leq \ell \leq 45\}$ とする。抽出したホットトピック群 $\{C_k\ ; 1 \leq k \leq K\}$ の性能 $F(K)$ を、情報検索などでよく使用される F 値に基づいて、

$$F(K) = \frac{1}{45} \sum_{\ell=1}^{45} F_{\ell,k^*_\ell}$$

で評価した。ここに、

$$k^*_\ell = \arg\max_k F_{k,\ell}, \quad F_{\ell,k} = \frac{2\,|H_\ell \cap C_k|}{|H_\ell| + |C_k|}$$

である。

4.3　実験結果

図 1 に、提案法と Kleinberg 法との性能比較結果を示す。提案法は Kleinberg 法より高い性能を示していた。

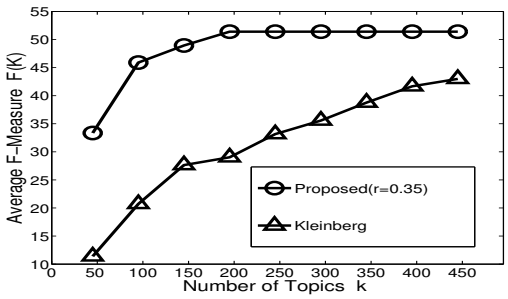

図 1　SR 法と Kleinbeg 法との性能比較

図 2 に、閾値 r を 0.30 から 0.40 の範囲で変化させた時の提案法の性能変化を示す。$r = 0.35$ の時が最も性能が高かった。ところで、$r = 0.35$ のネットワークのリンク数 (5008) は、$r = 0$ のネットワークのリンク数 (2910653) の 0.17 %であった。$r < 0.35$ ならば、より大規模なネットワークが構築され、抽出するコアのサイズが大きくなりすぎて性能が下がり、$r > 0.35$ ならば、より小規模なネットワークが構築され、抽出すべきコアが抽出できず性能が下がっているものと考えられる。

表 1 に提案法で抽出した上位 5 位までのホットトピックを示す。抽出したホットトピックのアノテーションとして、そのホットトピックに典型的な単語群をフィッシャー検定法により抽出している。

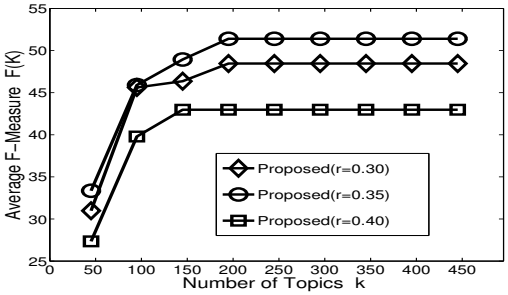

図 2　r の変化による提案法の性能変化

表 1　ホットトピック抽出結果

順位	ホット期間	抽出単語
1	6 月 1 日から 6 月 24 日	制裁, 北朝鮮, 朝鮮民主主義人民共和国, 核, カーター
2	5 月 3 日から 5 月 21 日	先行, ガザ, パレスチナ, 自治, 警察
3	2 月 6 日から 4 月 28 日	セルビア, サラエボ, 空爆, ボスニア, 勢力
4	6 月 17 日から 6 月 29 日	カーター, 日成, 南北, 洪, 板門店
5	4 月 5 日から 6 月 17 日	ゴラジュデ, 幸治, 上村, 幸彦, 町田

5　まとめ

文書ネットワーク構造、SR 法、およびバースト度に基づいて、文書ストリームデータからホットトピックとそのホット期間を抽出する手法を提案した。トピックが付与されている新聞記事データを用いた実験により、提案法の Kleinberg 法に対する有効性を実証した。

謝辞

本研究は科学研究費補助金基盤研究 (C)(No.20500147) の補助を受けた。

参考文献

[1] Kleinberg. J.: Bursty and hierarchical structure in streams, *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-03)*, pp. 91–101 (2002).

[2] Saito, K., Ueda, N., Kimura, M., Kazama, K., and Sato, S.: Filtering search engine spam based on anomaly detection approach, *Proceedings of the KDD2005 Workshop on Data Mining Methods for Anomaly Detection*, pp. 62–66 (2005).

[3] 斉藤和巳, 木村昌弘, 上田修功: 文書トピックに関する認知科学的実験, 人工知能学会研究会資料 *(SIG-KBS-A405-10)* pp. 57-62 (2005).

[4] Zhao, Q., Mitra, P., and Chen, B.: Temporal and information flow based event detection from social text streams, *Proceedings of the 22nd AAAI Conference on Artificial Intelligence (AAAI-07)*, pp. 1501–1506 (2007).