到着頻度と関連性を考慮した時系列文書のトピック分析

Topic Analysis of Document Streams Based on Document Arrival Rate and Document Relevance

崔 春花♥ 北川 博之♦

Chunhua CUI Hiroyuki KITAGAWA

近年ネットワークを介して大量の文書の配信や交換が行われており，それらコンテンツの分析技術の重要性が増加している．重要なコンテンツ分析の一つとして，電子メールやニュース記事などの大規模時系列文書のトピック分析がある．本研究では，与えられたキーワード集合で特徴付けられるトピックの時間的な活性度の変化の分析を対象とする．対象とするトピックとの関連性が高い文書が高い頻度で到着するのは，そのトピックの活性度が高い状態であり，そうでない場合には活性度が低い状態と見なす．時系列文書のトピック分析においては，ニュース記事などが到着するたびに連続的に過去の一定期間のトピックの活性度をモニターしたいという状況が考えられる．本論文では，Kleinberg の手法をもとに，このような時系列文書に対する連続的なトピック分析の手法を提案する．また，実データを用いた実験によりその有効性を検証する．

Dissemination and exchange of a large amount of documents have become popular according to the advance of network technology in recent years. Thus, importance of content analysis techniques is increasing. Topic analysis in a series of large-scale document streams such as E-mails and news articles is one of such. important research issues. Our research especially aims at the analysis of time varying activation levels of topics. When documents of high relevance with a specific topic arrive very frequently, then the activation level of the topic is regarded high, otherwise the activation level is considered to be low. In addition, it is required to continuously analyze activation levels in the document streams. In this paper, we propose a new method to attain this extending Kleinberg's analysis method. Moreover, we evaluate the effectiveness of the proposed method by experiments using real data.

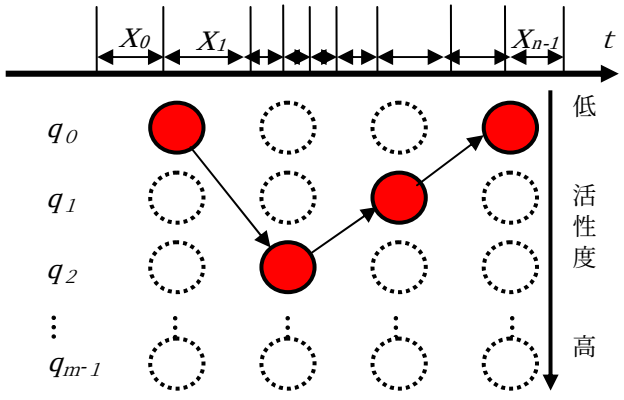
1. はじめに

近年ネットワークを介した大量の文書の配信や交換が急増しつつあり，電子メールやニュース記事などのような時系列文書のコンテンツ分析技術の重要性が増加している．そのようなコンテンツ分析の一つとしてトピック分析があり，またトピック分析の一種として，あるトピックの活性度の分析がある．例えば，ニュース記事文書列においてあるトピックに関する記事が頻繁に到着する場合には，一般に，そのトピックの活性度が高い状態であり，そうでない場合には活性度が低い状態と見なすことができる．このような活性度の分析は，時間軸と関連付けた大規模時系列文書の構造解析，要約，傾向分析等において極めて重要である．活性度分析においては，文書の内容分析と文書の到着頻度の分析の両者が重要である．トピックの活性度を分析する手法としては Kleinberg の提案によるもの[4]があるが，文書の到着頻度のみを考慮しており，類似度は考慮されていないという問題がある．さらに，Kleinberg の手法は既に配信済みの文書群を対象にバッチ的に分析処理を行うことを想定している．しかし，時系列文書のトピック分析においては，文書記事などが到着するたびにその記事から過去一定期間を対象として連続的にオンラインでトピックの活性度を分析しモニターしたいというような状況が考えられる．

そこで，本研究では，Kleinberg の手法をベースとして，あるトピックに対する文書の関連性と文書の到着頻度の両者を考慮した時系列文書に対するインクリメンタルな活性度分析手法を提案する．また，CNN ニュース記事文書を対象とした実験により，本提案手法の有効性を検証する．

2. Kleinberg の分析手法

Kleinbergは，時系列文書中の特定のトピックに関する文書の到着頻度に着目し，そのトピックの活性度を分析する手法を提案した．詳細については論文[4]に述べられている．Kleinbergによる手法では，内部状態に応じて文書の到着時間間隔が確率的に決定される隠れマルコフモデルを用いた分析を行う．簡単な例を図１に示す．図１では時間軸 t にそった時系列文書の到着を示す．縦線は文書の到着を表し，x_0, x_1, …, x_{n-1}は文書間の到着時間間隔を表す．Kleinbergの手法では，個々の到着時間間隔x_iは付随する隠れマルコフモデルの内部状態に応じて確率的に出力される記号であるとみなす．いま，q_0からq_{m-1}までのm個の状態を持つ隠れマルコフモデルを仮定し，q_0からq_{m-1}まで状態番号が増加する程，確率的により短い到着時間間隔を与えるものとする．したがって，この順に活性度は高くなることになる．例えば，図１では文書の到着時間間隔が長くなった場合に，活性度は低い状態にあると見なせるが，到着時間間隔が短くなった場合には，活性度の高い内部状態への遷移が生じたと見なすことが



状態遷移列：$\{q_0, q_2, q_1, q_0\}$

図１ Kleinberg の分析手法

Fig. 1 Kleinberg's analysis method

♥ 学生会員 筑波大学理工学研究科修士課程
hana@kde.is.tsukuba.ac.jp
♦ 正会員 筑波大学システム情報工学研究科
kitagawa@cs.tsukuba.ac.jp

できる．このように，文書の到着時間間隔を活性度に応じた内部状態遷移に反映することで，時間的な活性度の変化をモデル化するというのが基本的な考え方である．

より具体的には，各状態q_iにおいて，文書の到着時間間隔は所定の指数分布に基づいて確率的に決定されるものとする．つまり文書iとそれに続いて到着する文書$i+1$の間の時間間隔x_iは指数確率密度関数 $f_k(x_i) = \alpha_k e^{-\alpha_k x_i}$ によって決定されるものとする．ここで$\alpha_k = (n/T)\beta^k$ （nは全文書数，Tは全時間幅，$\beta > 1$はパラメタ，kは適当な整数）は単位時間当たりの文書の平均到着率であり，その逆数が平均到着時間間隔となる．したがって，α_kが大きい状態ほど頻繁に文書が到着する活性度が高い状態を表す．

一連の文書到着時間間隔列$x = (x_0, x_1, \cdots, x_{n-1})$が与えられた時，最適の状態遷移列$s = (s_0, s_1, \cdots, s_{n-1})$（各$s_i$は状態番号を表す）は下記のコスト関数を最小にするものとして求める．

$$c(s \mid x) = (\sum_{i=0}^{n-2} \tau(s_i, s_{i+1})) + (\sum_{i=0}^{n-1} -\ln f_{s_i}(x_i))$$

ここで，この式の第１項は，内部状態が状態s_iからs_{i+1}に遷移する際のコストの総和であり，第２項は，各状態s_iが到着時間間隔x_iを発生しやすい程小さいコストとなる．

文書到着時間間隔列$x = (x_0, x_1, \cdots, x_{n-1})$に対して上記コスト関数を最小化する状態遷移列$s$は，隠れマルコフモデルに対するビタビアルゴリズムを用いて求めることができる．

3. 提案手法

3.1 バッチ手法

Kleinberg の手法では，あるトピックと関連がある文書をキーワード検索等の何らかの手法で特定した後に，上記の方法で文書の到着頻度だけ考慮して，そのトピックの活性度を分析する．しかし，同じトピックに関する文書でも，当然そのトピックとの関連性が強い文書と弱い文書が存在する．したがって，個々の文書は関連性に応じてそれぞれの重要度を区分することができる．

そこで，我々は，文書の到着頻度と関連性の度合いの両者を考慮するよう，Kleinberg の手法を拡張する．具体的な方法として，各文書の関連性の度合いを考慮して文書の到着時間間隔を補正することを試みる．一般に，ある時刻に到着した文書の影響力は時間と共に次第に逓減すると考えることができる［5］［6］．この考え方を本研究に当てはめると，トピックとの関連性の度合いが弱い文書が到着した時点での状態は，より強い関連性の度合いをもつ文書が到着してから一定の時間が経過した後の状態と同等と見なすことができる．

具体的な文書や情報の影響力の時間的逓減の割合を表すためのモデルとしては，指数関数逓減モデルがこれまでの研究の中でしばしば用いられている［5］［6］．すなわち，文書の影響力は時間tの経過と共に，$R=e^{-\delta t}$のように逓減するとみなす．ここでδ は影響力の時間的減少の割合を決定するパラメタである．すなわち，トピックとの関連性の度合いR_iの文書が到着してから時間x_i経過後に次の文書が到着した場合，関連性の度合いがρ の文書が到着してから時間$x_i+\tau_i$経過後に次の文書が到着したものとみなす．ただし，$\tau_i = (\ln\rho - \ln R_i)/\delta$ であり，ρ はこのような補正のための正規化された関連性の度合いとする．また，$R_i > \rho$ の場合は$\tau_i = 0$とする．このようにしてもともとの到着時間間隔列$x = (x_0, x_1, \cdots, x_{n-1})$は$(x_0+\tau_0, x_1+\tau_1, \cdots, x_{n-1}+\tau_{n-1})$に変換される．このような変換後の最適状態遷移列は，もともとのKleinbergモデルと同様に隠れマルコフモデルに対するビタビアルゴリズムを用いて求めることができる．

ビタビアルゴリズムの概要は次の通りである．文書ストリームの到着時間間隔列$(x_0, x_1, \cdots, x_i)$に対応する状態遷移で状態q_jで終了するものの最小コストを$C_j(i)$で表す．$C_j(i)$は，初期状態がq_0にあるものとして，iを順次増加させながら次の式を計算することで求めることができる．

$C_j(i) = -\ln f_j(x_i) + \min_l(C_l(i-1) + \tau(l,j))$

この式からわかるように，ビタビアルゴリズムのよる計算は文書到着時間間隔列$x = (x_0, x_1, \cdots, x_{n-1})$に対して，$x_0$から$x_{n-1}$の順番で行われる．

このような時間的補正を行う手法は既に配信済みの文書群を対象にバッチ的に分析処理を行うため，以下ではバッチ手法と呼ぶ．

3.2 移動ウインドウ方式

時系列文書のトピック分析においては，新たな文書記事が到着するたびにその記事から過去一定期間を対象として連続的にオンラインでトピックの活性度を分析しモニターしたいというような状況が考えられる．この場合，記事の到着時刻より一定の時間幅をもつウインドウ内の文書群を対象とした分析を行うこととなる．時間経過に伴う新たな記事の到着によりこのウインドウは移動するため，移動ウインドウ（moving window）方式と呼ぶ．

最も単純な方法は，各ウインドウ内の文書群を対象に上記のバッチ処理手法を毎回適用することである．すなわち，全体の文書到着時間間隔列$x = (x_0, x_1, \cdots, x_{n-1})$が与えられたとき，現在分析対象のウインドウ$W_{n-1}$内に含まれる部分文書到着時間間隔列を$x' = (x_{n-m}, x_{n-m+1}, \cdots, x_{n-1})$とする．この$x'$にバッチ手法に適用し，状態遷移列$s' = (s'_{n-m}, s'_{n-m+1}, \cdots, s'_{n-1})$を求める．しかし，この方法では，各ウインドウ毎に毎回その中に含まれる全到着時間間隔を対象とした分析を行わなければならず，効率の面で望ましくない．すなわち，新たな記事の到着に伴う時間間隔x_{n-1}が与えられた時，このx_{n-1}に対する処理のみを行うことにより，x'に対する状態遷移列を導出するようなインクリメンタルな手法が望ましい．しかし，バッチ手法の状態遷移列計算そのものをインクリメンタルに行うことは難しい．そこで，バッチ手法における状態遷移列計算の方法を若干変更し，インクリメンタルに実行可能な方法を考える．

インクリメンタルに各ウインドウに対する状態遷移列を求めるために，次のような方法を試みる．すなわち，全到着時間間隔列上のx_iに対応する確率密度関数 $f_{k,i}(x_i) = \alpha_{k,i} e^{-\alpha_k x_i}$ を考える際，$\alpha_{k,i} = (n/T)_i \beta^k$とする．ここで，$(n/T)_i$はウインド$W_i$における平均到着頻度$n/T$である．上記に述べたビタビアルゴリズムの性質により，この計算は文書到着時間間隔列$x = (x_0, x_1, \cdots, x_{n-1})$に対応する状態遷移列$s = (s_0, s_1, \cdots, s_{n-1})$を毎回このような $f_{k,i}(x_i) = \alpha_{k,i} e^{-\alpha_k x_i}$ を用いながらインクリメンタルに行うことができる．そして，状態遷移列$s = (s_0, s_1, \cdots, s_{n-1})$の中で，$W_{n-1}$に対応した状態遷移列$s' = (s_{n-m}, s_{n-m+1}, \cdots, s_{n-1})$を結果とする．この計算は，インクリメンタル，かつウインドウ内の状態遷移列情報とコスト情報を保持していれば実行可能である．

4. 実験

本節では提案手法に対する実験評価を示す．実験１では実データを用いたバッチ手法と Kleinberg の手法を用いた分析結果を示し，3.1 節に述べた本研究における時間補正の有効性を示す．次の実験２では提案手法ともう一つのインクリメンタル計算手法である静的インクリメンタル手法の比較を行う．静的インクリメンタル手法については後述する．実験データとしては，TDT（Topic Detection and Tracing Evaluation）用の評価データの一部である 1998.1.1～1998.6.30 の CNN ニュース記事 21587 件を使用した．

4.1 実験１

実験１では，パッチ手法と Kleinberg の手法による活性度の分析結果の比較を行う．具体的な実験方法としては，最初にあるトピックを記述するキーワード集合を１つ与え，全ての記事について tf/idf による重み付けを考慮した余弦尺度を用いて関連性の度合いを計算する．そして，関連性の度合いが０より大きいすべての記事を対象とする．

本実験では，CNN ニュース記事中に現れるトピック「Oprah Lawsuit」を特徴付ける 10 単語を tf/idf による重み付けを用いた手法を用いて求め[8]，キーワード集合とした．具体的なキーワード集合は {Winfrei，oprah，rancher，cattl，Amarillo，defame，beef，texa，cow，cattlemen} である．また，各パラメタ値は，$\beta=1.1$，$\delta=1$，$\rho=0.2$ とした．

図２は，バッチ手法と Kleinberg 手法による分析結果を示す．横軸は 1998.1.1 から起算した日数を示し，縦軸は状態を表しており上にいくほど活性度が高い状態を示す．また，図３はトピック「Oprah Lawsuit」のラベルがついた記事がそれぞれの日に実際に到着した件数を示している．図３に示されているように，Kleinberg の手法で得た状態遷移では，全期間に渡って活性度の高い状態が出現している．一方，バッチ手法では，図４に示した記事の到着パターンとの整合性が高い結果が得られている．このように，Kleinberg の手法では，現実の活性度からの乖離が大きい．それに対して，バッチ手法から得た活性度は現実の活性度の変化により近いものとなっている．

4.2 実験２

実験２では，提案手法と静的インクリメンタル手法を比較する．静的インクリメンタル手法とは次の手法である．すなわち，全到着時間間隔列 $x = (x_0, x_1, \cdots, x_{n-1})$ 上の x_0 を含む最初のウインドウを対象として n/T を求める．この n/T の値を以降のすべてのウインドウに対して用いる（すなわち，$f_k(x_i) = \alpha_k e^{-\alpha_k x_i}$ を用いる）以外は，提案手法と全く同じ計算を行う．静的インクリメンタル手法はインクリメンタルな計算が可能である点は提案手法と同様であるが，ウインドウ毎の到着頻度の変化を考慮しないより単純な方法である．いずれの手法の場合も，各ウインドウ毎にバッチ手法を適用した結果により近い結果が得られる方が望ましい．

本実験では，実験１と同一のキーワード集合を与える．ここでは，ウインドウの幅を 60 日とした場合の実験結果において，最後の記事を含むウインドウに対する分析結果を示す．図４，図５，図６は，それぞれ，バッチ手法，静的インクリメンタル手法，提案手法による結果である．なお，バッチ手法の結果は，対象とするウインドウ内の文書群に対してバッチ手法を適用して得たものである．これらのグラフにおいて，横軸は 1998.4.27 から起算した日数を示し，縦軸は状態を示す．これらのグラフからわかるように，提案手法はバッチ手

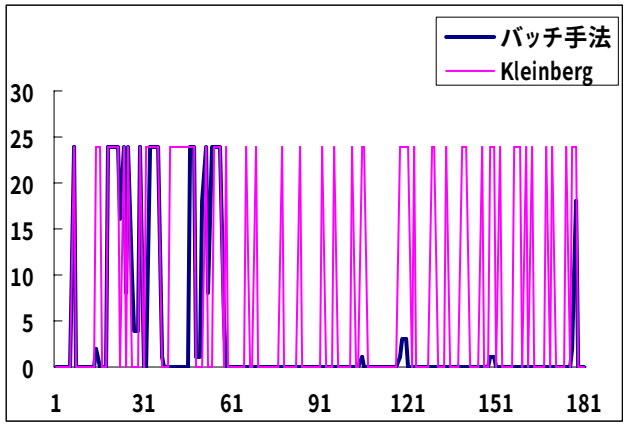


図２　トピック「Oprah Lawsuit」の活性度の変化
Fig. 2　Topic activation levels for “Oprah Lawsuit”

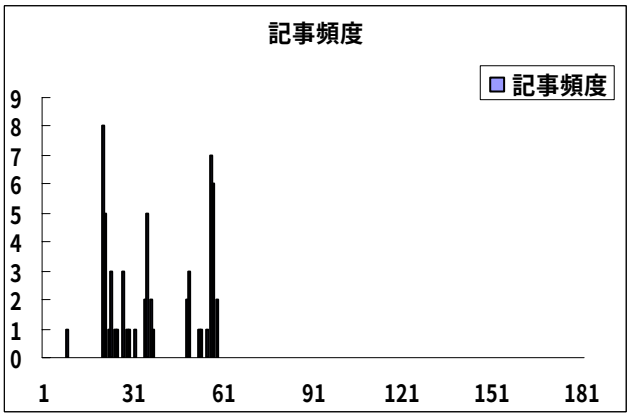


図３　トピック「Oprah Lawsuit」の記事の到着数
Fig. 3　Arrivals of documents for “Oprah Lawsuit”

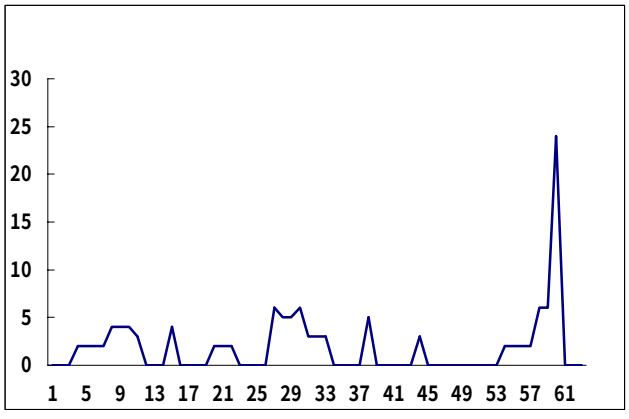


図４　バッチ手法による分析結果
Fig. 4　Result by batch scheme for “Oprah Lawsuit”

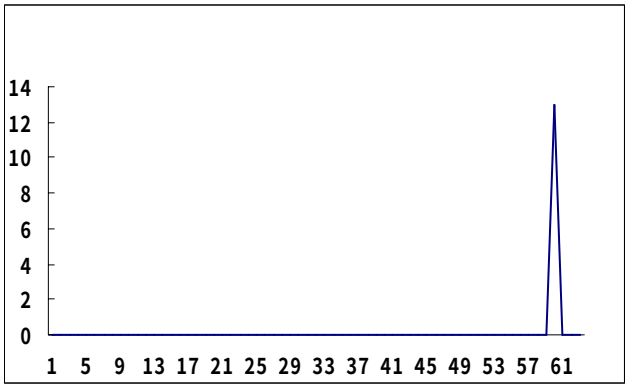


図５　静的インクリメンタル手法による分析結果
Fig. 5　Result by static incremental scheme

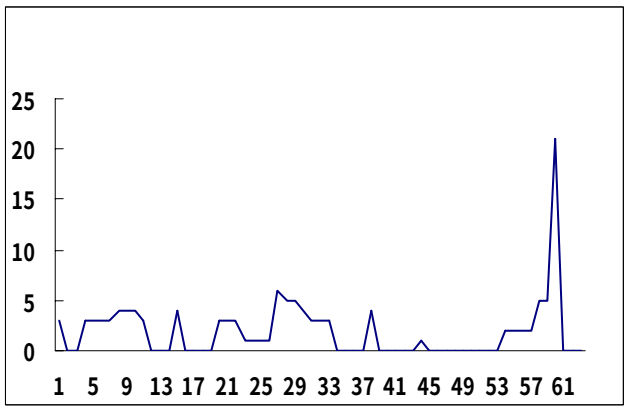


図６　提案手法による分析結果
Fig. 6　Result by proposed scheme

法を各ウインドウに適用した場合とほぼ同じ結果を与えている．これに対して，静的インクリメンタル手法の分析結果は，バッチ手法を適用した場合との乖離が大きい．

5. まとめと今後の課題

本研究では，文書の関連性と到着頻度の両者を考慮した時系列文書ストリームに対するトピック活性度の分析をインクリメンタルに行う手法を提案した．また，ＣＮＮニュース記事を用いた実験により，提案手法を用いることで，バッチ手法とほぼ同様の分析結果を連続的に得ることが可能であると見通しを得た．

今後の課題としては，より多様なデータやトピックを用いた詳細な実験検討や，適切なパラメタの選択方法の検討などがある．

[謝辞]

本研究の一部は，科学研究費基盤研究(B)（#15300027）ならびに特定領域研究（#16016205）による．

[文献]

[1] F. Walls, H. Jin, S. Sista, and R. Schwartz, "Topic Detection in Broadcast News", Proc. DARPA Broadcast News Workshop, 1999.

[2] J. M. Schultz and M. Liberman, "Topic Detection and Tracking using Idf-Weighted Cosine Coefficient", Proc. DARPA Broadcast News Workshop, 1999.

[3] H. Li and K. Yamanishi, "Topic Analysis using Finite Mixture Model", Information Processing and Management, Vol. 39, 2003.

[4] J. Kleinberg, "Bursty and Hierarchical Structure in Streams", Proc. ACM SIGKDD, 2002.

[5] Y. Ishikawa, Y. Chen, and H. Kitagawa, "An On-Line Document Clustering Method Based on Forgetting Factors", Proc. 5th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2001), September 2001.

[6] B. K. Yi, et al., "Online Data Mining for Co-Evolving Time Sequences", Proc. 16th International Conference on Data Engineering, 2000.

[7] 崔春花，北川博之，「到着頻度と関連性を考慮した文書ストリームのトピック分析」電子情報通信学会第 15 回データ工学ワークショップ(DEWS2004)，2004 年 3 月．

[8] 濱本雅史，北川博之，Jia-Yu Pan, and Christos Faloutsos，「独立成分分析を用いたテキストデータからのトピック検出」電子情報通信学会第 15 回データ工学ワークショップ(DEWS2004)，2004 年 3 月．

崔 春花　Chunhua CUI
筑波大学大学院理工学研究科在学中．日本データベース学会学生会員．

北川 博之　Hiroyuki KITAGAWA
筑波大学大学院システム情報工学研究科，計算科学研究センター教授．1980 年東京大学大学院理学系研究科修了，理学博士（東京大学）．異種情報源統合，文書データベース，WWW の高度利用などの研究に従事．著書「データベースシステム」（昭晃堂），「Unnormalized Relational Data Model」(共著，Springer-Verlag)等．ACM，IEEE-CS，情報処理学会，電子情報通信学会，日本データベース学会，日本ソフトウェア科学会，各会員．