2009－NL－190 (13)
2009／3／26

多重トピックを用いたブログ空間の情報伝搬解析

横山 正太朗† 江口 浩二† 大川 剛直†

† 神戸大学工学部情報知能工学科
〒 657–8501 神戸市灘区六甲台町 1–1
E-mail: s-yokoyama@cs25.scitec.kobe-u.ac.jp

要旨 近年ブログの利用が爆発的に増加しており，重要な情報源のひとつになりつつある．ブログは，ハイパーリンクを利用することで，参考にした情報を明示的に参照することが可能であり，このネットワークを対象にした研究が最近注目されつつある．しかし，こういった研究のほとんどが，リンク情報のみを対象にしており，本文の情報を参照していない．そこで本研究では，リンク構造だけでなく，本文のトピックを推定し，適切に情報伝播を捉える手段を確立することを目的とする．文書集合の潜在的なトピックを統計的に推定するのに用いられるトピックモデルの代表的なものに，潜在的ディリクレ配分法 (Latent Dirichlet Allocation ; LDA) が挙げられ，広く用いられている．本研究では，この LDA を用いてポストのトピックを推定し，リンク間のトピック分布を比較することで，正確な情報伝搬の単位（カスケード）を抽出する枠組みを提案した．実際，日本語ブログデータを用いた実験において，提案手法の有効性を示唆する結果を得た．

Analyzing Information Diffusion in Blogosphere using Probabilistic Topics

Shotaro Yokoyama† Koji Eguchi† Takenao Ohkawa†

†Kobe University
Department of Computer Science and Systems Engineering
1-1 Rokkoudai, Nada-ku, Kobe, 657-8501, Japan
E-mail: s-yokoyama@cs25.scitec.kobe-u.ac.jp

Abstract Recently, the number of blogs has been explosively increased, and has become one of the crucial resources for us to get useful information. A blog post can be linked to other blog posts using hyperlink to give reference information. Analyzing such networks of blogs has recently attracted considerable attentions of researchers. However, most of them focused only on link structure of the blog networks, not using the content of the blog posts. In this paper, we propose a method for appropriately analyzing information diffusion in the blog networks using both the link structure and text content. Latent Dirichlet Allocation (LDA) is a standard topic model to estimate latent topics in a document collection. We propose a framework to accurately extract units of the information diffusion, cascades, by estimating latent topics of blog posts using LDA and then comparing distributions of the topics in the blog posts between each hyperlink. We demonstrate, through an experiment using Japanese blog data, that our proposed method works effectively.

1 はじめに

近年ブログサービスの利用数が爆発的に増加しており，2008 年 1 月の時点で日本国内のブログ総数は約 1,690 万，ブログの記事 (ポスト) 総数は約 13 億 5,000 万，データ量は 42 テラバイトであると言われている [1]．ブログとは，日々更新される日記的な Web サイトの総称であり，タイムリーな公開や更新が容易に行える．内容としては，ブログ著者 (ブロガー) が関心を持ったニュースや出来事について書かれることが多い．

また，ブログは，ハイパーリンクを利用することで，参考にした情報を明示的に参照することが可能であり，インターネット上に形成される社会ネットワークから，情報や影響がどのように伝播していくのかを表わしたリソースであると言える．こういった背景から，ブログは情報を手に入れるための重要なメディアのひとつであると言え，ブログを対象にした研究が現在注目を浴びている．

ところで，ブログポストのリンクの遷移によって生成される情報伝播グラフはカスケードと呼ばれており，ブログ空間を対象にした研究のひとつに，このカスケードに関する研究が行われている [3][4][5]．こういった研究のほとんどがカスケードを抽出する際，リンク情報のみを対象にしており，ブログポストの本文を参照していない．

これらは，ブロガーがポストを投稿する際，他のポストやウェブ上のリソースにリンクを貼るという行為は，リンク先の情報に影響を受けたからであるという考えに基づいている．しかし，ブロガーが意図的にリンクを貼ることや単なる間違いなどにより，情報伝搬が起こっているとは言い難いリンクが少なからず存在する．更に，現在ブログの利用者数の増加に伴い，スパムブログ (splog) の数が増えているという問題がある．splog は国内ブログの約 12 %を占めていると言われ [1]，こういった問題を考慮しなければ，情報の伝播を示していないリンクも捉える可能性が高い．

以上の問題を解決するために，ブログポストのリンク情報だけでなく，本文の情報を利用して，情報伝播を適切に捉えたカスケードを抽出する手段を確立する必要がある．

ところで，与えられた文書集合から潜在的なトピックを推定するのに用いられるモデルでトピックモデルが存在する [8]．トピックとは，ある特徴を持った単語の分布であり，例えば，アメリカの大統領選挙に関する単語の分布などがトピックモデルを用いることで抽出できる．このトピックモデルは，情報検索やデータマイニングなどの様々な応用が存在し，代表的なトピックモデルとして，潜在的ディリクレ配分法 (Latent Dirichlet Allocation, LDA)[7] が挙げられ，広く用いられている．

本研究では，この LDA を用いてブログポストのトピックを推定し，リンク間のトピック分布を比較することで，正確なカスケードを抽出する枠組みを提案する．

2 関連研究

ここでは，本研究に必要なトピックモデル，ブログや情報カスケードに関連した研究について以下で述べる．

2.1 トピックモデル

トピックモデルとは「文書がどのようなトピックについて書かれているのか」という考えに基づいた確率的モデルである [8][7][9]．Blei ら [7] は，潜在的ディリクレ配分法（Latent Dirichlet allocation: LDA）を提案した．彼らは，LDA モデルを推定するのに変分ベイズを用いた．変分ベイズに対し，Griffiths ら [10] は，ギブス・サンプリングを推定に用いた．本研究では，この Griffiths らが利用したギブス・サンプリング法を用いる．Wang ら [2] は，トピックを文書が書かれたタイムスタンプと関連付けることで，時間の変化に対応した Topics-over-Time モデルを開発した．

2.2 ブログと情報カスケード

Leskovec ら [3] は，ブログ空間におけるリンク構造に着目し，ブログ空間で情報がどのように伝播するのかを分析をした．また，ブログ空間での突発的な動きを効率的に検出する方法について研究を行った．[4]．谷口ら [5] は，ブログポストの本文中のハイパーリンクから，ブログコミュニティを抽出・分析した．これらの研究はリンク構造に着目しており，ブログポストの本文情報を扱っていない点で本研究と異なる．戸田ら [6] はブログ空間での話題の変遷を抽出し，カテゴリごとの特徴的なパターンを抽出する手法を提案した．この研究では，ブログポストの本文情報を用いてクラスタリングを行っており，リンク構造には着目していない点で本研究と異なる．

本研究では，ブログ空間のネットワーク構造およびポストの本文情報に着目した研究を行う．

3 LDAに基づくカスケードの抽出

本研究は，まずブログ空間からポスト間のリンクによって構成される情報伝播ネットワークであるカスケードの抽出を行う．そして，ポストの本文情報からポストのトピックを推定し，抽出したカスケードのリンクで結ばれたポスト同士のトピック分布を比較する．比較した結果，トピック分布が大きく異なるリンクは，正しい情報伝播を表さないと考えられるので，そのリンクを除去し，正しい情報伝播を捉えたカスケードを抽出するというのが今回の目的である．そこで，まずブログやカスケードに関する基本的な概念について紹介する．そして，今回ブログポストのトピックを推定するために用いたLDA，およびポスト間のトピックの比較に用いたJSダイバージェンスについて説明を行う．

3.1 ブログ空間におけるカスケード

ブログ空間で，ポストは他のポストやウェブ上のリソースにリンクしている．このとき，ポスト間のリンクによって生じるグラフ構造をポストネットワークと呼ぶ．図1でノードはポスト，辺はリンクに対応して

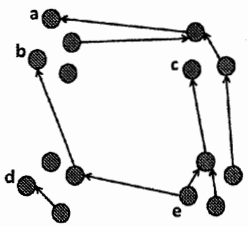


図1: ポストネットワーク

いる．ポスト間のリンクに注目すると，ポストはそれぞれ投稿された時間を表すタイムスタンプを持っており，未来のポストにはリンクできない．このポストネットワークからカスケードを得られる．

カスケードは開始点となる一つの開始ポストが存在し，開始ポストは他のポストを参照(アウトリンク)していない．図1ではノードa,b,c,dが開始ポストである．ポストが開始ポストにリンクすることでカスケードが生成され，新しいポストがカスケード内のポストに時間順でリンクしていくことでカスケードは大きくなっていく．図2は図1から得られたカスケードのリストであり，このとき情報伝播の流れは辺の向きと逆である．

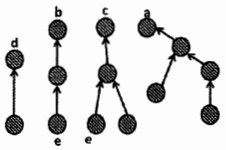


図2: 図1のカスケード

ここで，単一のポストのみで形成されるカスケードは情報の伝播を表さないため，以降は2つ以上のポストで形成されるカスケードのみを扱う．

また，あるポストが色々な話題について要約している場合，複数のカスケードが1つのポストでつながることが起こりうる．こういった複数のカスケードをひとつにまとめるポストを連結ポストと呼び，図1ではノードeが連結ポストである．図2を見てわかるように，この連結ポストは2つのカスケードに現れる．この例のように連結ポストにおいてカスケードを分割する．

3.2 LDAにおけるギブスサンプリング

本節では，ギブスサンプリングを用いたLDAの説明を行う．ここで，Dは文書の数，Tはトピックの数，Wは文書集合の異なり語数，N_iは文書iの延べ語数，α,βはディリクレ事前分布の超パラメータ，θは文書におけるトピックの多項分布，ϕはトピックにおける単語の多項分布，z_jは単語w_jのトピック割り当て，つまり語w_jにトピックt_kが割り当てられていることを$z_j = t_k$と表せる．

図3のLDAについて説明する．LDAの生成プロセスは，

(1)超パラメータαを与えたディリクレ分布から各文書iについてのトピックの多項分布θ_iをサンプリングする．

(2)超パラメータβを与えたディリクレ分布から各トピックt_kについての語の多項分布ϕ_kをサンプリングする．

(3)文書i内のN_i個の語w_jそれぞれに対して

(a)超パラメータθ_iを与えた多項分布からトピックz_jをサンプリングする．

(b)超パラメータϕ_{z_j}を与えた多項分布から語w_jをサンプリングする．

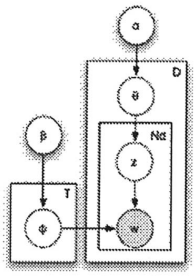


図 3: LDA のグラフィカルモデル

となる．LDA モデルにおける完全同時分布は，

$$\begin{aligned} & P(\mathbf{W}, \mathbf{Z}, \theta, \phi | \alpha, \beta) \\ & = P(\phi|\beta) \prod_{i=1}^{D} P(\theta_i|\alpha) P(z_i|\theta_i) P(w_i|z_i, \phi) \\ & = \left\{ \prod_{i=1}^{D} \frac{\Gamma(T\alpha)}{\Gamma(\alpha)^T} \prod_{k=1}^{T} \theta_{ik}^{c_D(i,k)+\alpha-1} \right\} \\ & \quad \times \left\{ \prod_{k=1}^{T} \frac{\Gamma(W\beta)}{\Gamma(\beta)^W} \prod_{j=1}^{W} \phi_{kj}^{c_V(j,k)+\beta-1} \right\} \end{aligned} \tag{1}$$

となる．ここで，$c_D(i,k)$ は，i 番目の文書の中でトピック z_k が割り当てられている回数，$c_V(j,k)$ は，トピック z_k に語 w_j が割り当てられている回数である．

では，実際にギブスサンプリングを用いた LDA のトピック推定の方法について説明する．ギブスサンプリングを用いて，i 番目の文書のある語のトピックを t_k だと推定する確率は，

$$\begin{aligned} & P(z_{il} = t_k | \mathbf{W}, \mathbf{Z}_{-il}, \alpha, \beta) \\ & \propto \frac{c_D(i,k) - 1 + \delta(k \neq k') + \alpha}{\sum_{k=1}^{T} c_D(i,k) - 1 + T\alpha} \\ & \quad \times \frac{c_V(j,k) - 1 + \delta(k \neq k') + \beta}{\sum_{j=1}^{W} c_V(j,k) - 1 + \delta(k \neq k') + W\beta} \end{aligned} \tag{2}$$

となる．なお，$\delta()$ は，() 内の条件が成り立つときは 1 となり，成り立たないときは 0 となるものである．

ここで，θ_{ik} は i 番目の文書を構成する空欄に，トピック t_k が割り当てられる確率であり，ϕ_{kj} はトピック t_k が割り当てられた空欄が，語彙 w_j で満たされる確率であるとすると，これらはそれぞれ式 (2) の右辺第 1 項と第 2 項に対応し，

$$\begin{aligned} \theta_{ik} &= \frac{c_D(i,k) - 1 + \delta(k \neq k') + \alpha}{\sum_{k=1}^{T} c_D(i,k) - 1 + T\alpha} \\ \phi_{kj} &= \frac{c_V(j,k) - 1 + \delta(k \neq k') + \beta}{\sum_{j=1}^{W} c_V(j,k) - 1 + \delta(k \neq k') + W\beta} \end{aligned} \tag{3}$$

式 (2) を用いて，各文書の各語について，トピックを割り当てる確率分布（最初はランダムな分布）を求め，求められた分布からトピックをひとつ選び，式 (2) に従ってその語へトピックを割り当てなおす．その結果をさらに別の語についてトピック割り当ての確率分布を求めるために利用する ... という計算を繰り返すのが，LDA におけるギブスサンプリングの実際の計算である．

3.3 JS ダイバージェンス

この節では，ポストのトピック分布を比較するために用いる JS ダイバージェンスの説明を行う．ブログポスト間の類似度は，各ポストに対応するトピック分布 θ の類似度から導出することが可能である．2 つの確率分布 p,q 間の類似度を求める方法のひとつとして KL ダイバージェンスがあり，

$$D(p,q) = \sum_{k=1}^{T} p_j \log \frac{p_j}{q_j} \tag{4}$$

と表せる．KL ダイバージェンスは非対称な関数であり，これを対称化した JS ダイバージェンス [11] がある．

$$JS(p,q) = \frac{1}{2}[D(p,(p+q)/2) + D(q,(p+q)/2)] \tag{5}$$

これは，p と q の類似度を p と q の平均を用いて導出している．本研究では，JS ダイバージェンスの計算の際には，各ブログポストのトピック分布を用いる．

JS ダイバージェンスによりポスト同士の類似度が導出され，この値の順にリンクを並べることで，(a) ある閾値のもとで境界線を引き，必要なリンク，不要なリンクに分ける．(b) 扱うブログ空間で切れるべきリンク，切る必要のないリンクの比率が既知であると仮定した場合，それを用いることで提案手法の有効性を確かめる，などの評価方法が考えられる．なお，本研究における提案手法の精度を評価する方法の詳細については 5 章で述べる．

4 ブログ空間のカスケードの抽出

実験に扱うデータセットとしては，2007年6月から8月の間の日本語で書かれた215万ブログポストを利用した．このデータセットから，トピック推定に必要な本文，および必要なハイパーリンクの抽出方法について以下に述べる．

4.1 本文テキストの抽出

ブログには多くのハイパーリンクが含まれているが，ブログを個人が情報を発信する場であると考えると，ブログの本文に含まれるハイパーリンクが情報の伝播を表していると考えることができる．そこで本研究では，まずブログの本文部分を抽出して，本文に含まれるハイパーリンクだけを抽出する．

ブログは，各ブログサービスで基本的な構造は似ていて，例えば，$< div\ class = post\_body >$ 本文 $< div\ class = post\_tail >$ のように，本文部分を特定できるようなテンプレートが存在することが多い．この特徴に注目し，各ブログサービスに対応したテンプレートを40個ほど用意し，本文を抜き出した．

さらに，本文に形態素解析[1]を行い，名詞を抽出した．ここで用いる名詞は，非自立語，接尾語，数，代名詞を除いている．また，文書中に稀にしか現れない語はトピックモデルの推定精度を低下させるので，10文書より少ない出現数の語も除去した．

4.2 ハイパーリンクの整理

次に，抽出した本文に含まれるハイパーリンクの選別を以下の方法で行う．

URLの標準化 URLの最後が"/"や"/index.html"で終わる場合，同一のリソースであると推測できるので，一意になるように統一する．さらに，クリックしたハイパーリンクの履歴をサーバ側に残すためにURLを符号化したパラメータとして指定している場合は(例，Yahoo!Japan)，URLを復号化して使用する．

データセット外へのリンクの除去 データセット以外のブログポストやウェブ上のリソースへのリンクは取り除く．これは，データセット外のURLが作成された時間が分からないので，未来リンクの発見ができないからである．

[1]形態素解析のツールとしてはmecab(http://mecab.sourceforge.net/)を用いた．

未来リンクの除去 未来の情報を参照することはできないので，未来のブログポストへのリンクはすべて除去する．こういった未来リンクは，ポストが更新されたり，意図的なバックポスト，時差などが原因であると考えられる．

自身へのリンクの除去 自身へのリンクは情報の伝播を表さないので，除去する．しかし，同一ブロガーの他のポストへのリンクは残しておいてよい．

以上の過程から，実際に利用するデータとして表3のデータセットを作成した．

表1: データセット

D	ブログポスト数	25604
L	リンク数	22150
C	カスケード数	7951
W	一般語の語彙数	19599
W_{num}	一般語の総数	3112607

5 評価実験

ここでは，ブログポストのトピックの推定精度を評価し，また，JSダイバージェンスにより，ポスト間の類似度がどれくらい正確に求められたのかについて評価する．

5.1 対数尤度

本実験には4章で作成したデータセットを使用し，各ブログポストの90%の単語をトレーニングデータとし，残り10%をテストデータとした．また，LDAのトピック数を $T = 50, 100, 200, 300, 400$ と設定し評価をおこなった．更に，ディリクレ事前分布の超パラメータは $\alpha = 50/T$，$\beta = 0.01$[9] に設定した．

推定されたモデルのテストデータに対する尤度は，

$$p(\mathbf{w}^{test}) = \prod_{ij} \sum_{k} \theta_{i,k} \phi_{k,w_{i,j}^{test}} \qquad (6)$$

から導出することができる．対数尤度は，上式の対数をとることにより求まる．この値が高いほど，推定モデルは高精度であることを示している．

各トピック数ごとに導出した対数尤度のグラフを図4に示す．横軸は繰り返し回数で，縦軸は1単語あたりの対数尤度である．

各グラフを比較すると，どのトピック数でも繰り返し回数100回ほどで収束が確認され，トピック数が増加すると結果が良くなることが分かる．しかし，LDAの計算コストはトピック数に依存するので，本実験では，トピック数400，繰り返し回数200回で推定されたモデルを用いる．

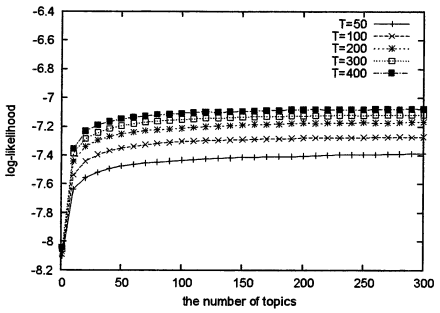


図 4: 対数尤度

5.2 ポスト間類似度の評価

JS ダイバージェンスにより求めたポスト間の類似度の正確さについて評価する．評価方法としては，再現率 (recall) と精度 (precision)，そして平均精度 (Average Precision) や R 精度（R-precision, Break even point）を用いる．

再現率とは検索における「もれ」の少なさを，精度は検索における「ごみ」の少なさを意味する．本研究において「もれ」とは，本来不必要なリンクで，提案手法で除去できなかったリンク,「ごみ」とは，本来必要なリンクで，提案手法により取り除いてしまったリンクのことである．再現率と精度は，

$$recall = \frac{\text{正しく取り除けたリンク数}}{\text{実際に不必要なリンク数}} \quad (7)$$

$$precision = \frac{\text{正しく取り除けたリンク数}}{\text{取り除いた全リンク数}} \quad (8)$$

となる．

また，再現率と精度が同じになるときの値を R 精度と呼び，平均精度は，検索された全ての正解について精度を計算して足し合わせ，最後に全正解数で割るというもので，

$$AvPrec = \frac{\text{検索された各正解の順位における精度の和}}{\text{全正解数}} \quad (9)$$

5.2.1 評価結果

上で述べた評価方法から，提案手法の評価を行う．評価に用いたデータとして，表 3 からランダムに 100 個のカスケードを選び，選んだカスケードのリンク間のポストの内容を実際に確認した．本実験では，不必要なリンクを発見したいため，不必要なリンクを正解リンクとし，必要なリンクを不正解リンクとした．提

表 2: データセット

C	カスケード数	100
D	ブログポスト数	520
L	リンク数	400
T	正解リンク数	142
F	不正解リンク数	258

案手法として，このデータのリンクを JS ダイバージェンスの値が大きい順，すなわちトピック分布が異なる順に並べた結果 (*proposed result*)，および比較実験のため，リンクをランダムに並べた結果 (*random result*) を用意した．なお，以下に示す結果の *random result* は 30 個の *random result* の平均である．

まず，横軸を再現率，縦軸を精度として表したグラフを図 5 に示す．このとき，R 精度は，

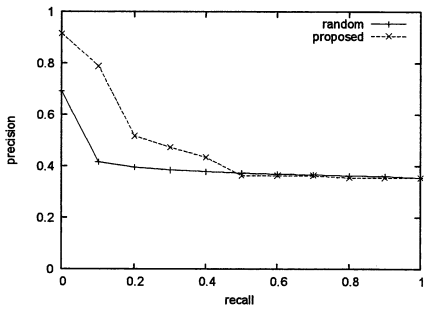


図 5: 再現率と精度

$$Rprec_{(proposed)} = 0.4225$$

$$Rprec_{(random)} = 0.3535$$

また，平均精度は，

$$AvPrec_{(proposed)} = 0.4557$$

$$AvPrec_{(random)} = 0.3649$$

となり，どちらの評価においても提案手法は良好な結果が得られたことが分かる．しかし，図 5 を見てわか

るように，再現率が 0.5 以上，正解リンクの半数が発見されて以降は *proposed result* と *random result* で同じ精度になっている．この理由としては，正解リンク (不必要だと判断したリンク) の JS ダイバージェンスの値が小さい，すなわちトピック分布が似ているという矛盾が起こっている考えられる．実際に，この状況に当てはまるリンクを確認したところ，考えられる原因は以下の 3 通りであった．

(1)本文が短いポスト同士のリンクである．

(2)本文抽出の際，本文以外の部分まで抽出している．

(3)ポストが複数のトピックのリンク集である．

まず，(1) の問題は，本文が短いポスト同士の場合，仮に内容が違うトピックについて書かれていたとしても，トピック分布が近いものになってしまうために起こりうる．

次に，(2) の問題に関しては，本文部分を表すテンプレートの多様性が考えられ，確実に本文部分だけを抽出することは，非常に困難である．したがって，本文以外の部分も抽出したり，本文とは全く違う部分だけを抽出することが起こりうる．これにより，本文以外の部分が同じ内容で書かれていた場合，トピック分布が近い値になってしまう．

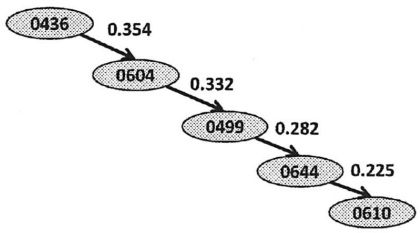
最後に (3) の問題は，本文がリンクのみを集めたものである場合がある．こういったリンク集には，(a) 同じトピックのリンクのみを集めたもの，(b) 様々なトピックのリンクのみを集めたもの，がある．本実験では，(a) の場合，リンクは残すべきだと判断し不正解リンクとした．一方，(b) では生成されるカスケードが，様々なトピックが含まれてしまうため，リンクを切るべきだと判断し正解リンクとした．(b) の場合，リンク先のポストの本文が短いと，(1) と同様に JS ダイバージェンスの値が小さくなってしまう．

一方で，切れるべきだと判断したリンクで実際に JS ダイバージェンスの値が大きかった事例で共通することは，本文の情報量が長いことである．当然のことながら，のべ語数の多いポストは LDA で正確なトピック推定が行えるので，ポストの類似度も正確になる．

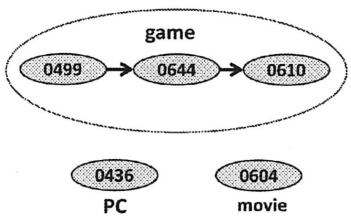
以上の点から，今後の課題としては本文の抽出精度の向上，および，利用するポストののべ語数が少ないものは除去することで提案手法は改善されるのではないかと思われる．

5.3 例

実際，提案手法により不要と判断されたリンクを取り除いた結果，カスケードがどのように変化したのかを例を以下に示す．今回は，精度を重視し，図 5 の再現率が 0.1，精度が 0.8 になる地点で境界線を引き，不要なリンクを取り除いた．その結果を以下に示す．実際はひとつのカスケード (図 6(a)) であったが，提案手法により 3 つに分離された (図 6(b))．各ポストの本文を確かめたところ，ポスト ID が 0499，0644，0610 はゲーム，0436，0604 はそれ以外のトピックであり，情報伝搬の単位となるカスケードが適切に抽出できたと言える．なお、図 6 に示された各ブログポストのトピック分布を表 3 に，主なトピックの単語分布のそれぞれにおいて確率値の大きい 3 語を表 4 に示す．



(a) 手法適用前



(b) 適用後

図 6: カスケードの例

表 3: トピック分布

0499

Topic ID	Prob.
Topic 267	0.1143
Topic 83	0.0505
Topic 366	0.0434

0644

Topic ID	Prob.
Topic 267	0.1273
Topic 39	0.0830
Topic 355	0.0704

0610

Topic ID	Prob.
Topic 267	0.1296
Topic 354	0.0559
Topic 235	0.0498

0436

Topic ID	Prob.
Topic 199	0.0879
Topic 362	0.0831
Topic 355	0.0637

0604

Topic ID	Prob.
Topic 312	0.1403
Topic 24	0.0733
Topic 67	0.0677

表 4: 単語分布

Topic ID	Topic 199	Topic 267	Topic 312	Topic 362
Rank 1	成海	ゲーム	ハリー	メモリ
Rank 2	璃子	プレイ	ポッター	カード
Rank 3	モデル	キャラ	シリーズ	DVD

6 結論

本論文では，トピックモデルである LDA を利用し，ブログポストのトピックを推定することで，不要なリンクを取り除き，正確な情報伝播を捉えたカスケードを抽出する枠組みを提案した．結果，約半数の不適切なリンクに対して，提案手法に一定の有効性が確認された．とりわけ提案手法において，JS ダイバージェンスによりトピック分布が異なると判断された上位１０％では良好な結果が得られた．

今後の課題としては，本研究の主目的ではないが，ブログポストの本文を抽出する精度の向上，および，JS ダイバージェンスによるポスト間の類似度比較がより効果的に行える条件の設定が挙げられる．また，統計的に提案手法が優れたものであるのかを調査する必要がある．更に，LDA 以外のトピックモデル，例えば文書が書かれた時間を考慮に入れた Topics-over-time[2] などと比較検討を行うことで，結果にどれくらいの差異や精度の違いが見られるかを確認することが考えられる．

そして，今後の展望としては，本研究によりトピックごとのカスケードの分布を導出することが可能となったので，カスケード単位の検索サービスの開発やリンク予測が挙げられる．

謝辞

本研究の一部は，科学研究費補助金特定領域研究「情報爆発 IT 基盤」(19024055)，基盤研究 (B) (20300038) の援助による．

参考文献

[1]総務省情報通信政策研究所調査研究部: "ブログの実態に関する調査研究の結果", http://www.soumu.go.jp/iicp/chousakenkyu/seika/houkoku.html (2008)

[2]Xuerui Wang, Andrew McCallum: Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends (2006).

[3]Jure Leskovec, Mary McGlohon, Christos Faloutsos. Cascading Behavior in Large Blog Graphs. Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA. Natalie Glance, Matthew Hurst Neilsen Buzzmetrics, Pittsburgh, PA.(2007).

[4]Jure Leskovec,Andreas Krause,Carlos Guestrin,Christos Faloutsos,Jeanne VanBriesen and Natalie Glance Cost-effective Outbreak Detection in Networks (2007).

[5]谷口智哉, 松尾豊, 石塚満: Blog コミュニティの抽出と分析, 第 6 回 セマンティックウェブとオントロジー研究会, 人工知能学会研究会資料 (2004).

[6]戸田 智子, 福田 直樹, 石川 博: Blog 記事のクラスタリングに基づいたカテゴリ別話題変遷パタンの抽出 (2007).

[7]D. M. Blei, A. Y. Ng and M. I. Jordan, "Latent Dirichlet allocation", Journal of Machine Learning Research, Vol. 3, pp. 993.1022 (2003).

[8]T. Hofmann, "Probabilistic latent semantic indexing", In Proceedings of the 22nd Anuual International ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley, California, USA, pp. 50.57 (1999).

[9]M. Steyvers and T. Griffiths, "Handbook of Latent Semantic Analysis", chapter 21: Probabilistic Topic Models, Lawrence Erbaum Associates, Mahwah, New Jersey and London (2007).

[10]T. L. Griffiths and M. Steyvers, "Finding scientific topics", Proceedings of the National Academy of Sciences of the United States of America, Vol. 101, pp. 5228.5235 (2004).

[11]Lillian Lee, "Measures of Distributional Similarity", Proceedings of the 37th ACL,(1999).