F-050

特定ジャンルのブログに対する共起とユーザ別特徴語抽出を用いた話題抽出

Topic Extraction using Co-occurrence and Specialized Words of Blog Articles in Specific Catetories

山本　尚央† Nao Yamamoto
佐藤　進也‡ Shin-ya Sato
菅原　俊治† Toshiharu Sugawara

1　はじめに

現在，情報発信の手段としてブログが広く普及している．ブログの普及により，多くのユーザが個人の意見を発信できるようになり，世の中の様々な人の意見が記述されている．

ブログの特徴として二つの側面があげられる．第一に，ウェブ上における個人の日記という側面，第二に，イベントに対する個人の意見を表現するメディアの一つという側面である．これらはブログ中に記述される様々な興味分野に対する話題として現れる．従って，ブログ全体の傾向だけではなく，興味分野に注目して傾向を解析することにより，話題発生の特徴や，その興味分野において最も注目を浴びている話題の抽出が可能である．

ブログから抽出したキーワードを話題語として表示する取り組みとして，kizasi.jp[2] では，共起頻度の高い話題語や単語を関連のあるものを提示し，その頻度をグラフ化している．また，ブログの話題の推移を抽出する研究も多くなされており，たとえば [10] ではブログを対象としてクラスタリングにより話題推移抽出を行っている．しかし，ブログの個別のジャンルでのみ興味が上がっている話題を抽出する手法は行われていない．

そこで本研究ではブログサービスで特定ジャンル（例：音楽ジャンル）に分類されているブログだけに注目し，ブログを書くユーザごとに語を重み付けして解析する．実データを使った実験を行い，特定ジャンルに対する他の解析手法と比較して，本手法が特定ジャンルに関する話題をより多く提示できることを示す．

2　関連研究

Kleinberg[8] は新聞などの時系列文書集合からの話題抽出の方法を提案した．時系列文書集合とは，新聞，電子掲示板，ブログに付属する時間の情報を時間順序に並べて構成できる集合である．文章の出現間隔を利用し出現間隔の短い箇所 (*burst*) を検出する手法を提案した．この手法では，時系列文書はそのときの *burst* 度に応じた頻度で，ランダムに出現すると仮定する．これは，文書の出現間隔が指数分布に従うことと等しい．つまり，文書の出現間隔が短いほど単位時間あたりの文書数は増加し，その時の burst 度が高くなる．この burst 度を用いて，それぞれの文書について最適な burst 度を知ることができる．しかし，この値をそのまま各文書の burst 度として使用した場合，ゆらぎとしてこの値が一時的に高くなることがある．そのため，Kleinberg の手法では，各 burst 度を隠れマルコフモデルの状態に割り当て，確率密度関数の値をそれぞれの状態に対するスコアとみなす．さらに状態が遷移する場合にはコストを与えて，状態遷移を起こりにくくしている．これによって，burst 度に関する境界値付近でも burst 度が連続して変化しないようにしている．この手法を用いてある単語の注目度を計算する場合，その単語を含む時系列文書のみからなる集合をつくり，その集合に対してこの手法を適用することで，その単語が盛り上がっていた時期と盛り上がりの度合いを調べている．

この Kleinberg の burst 度発見手法は，計算量が少なく，単語の出現頻度が大きくなる期間を特定できるという特徴を持つが，ブログのように急速にその数を増加させる対象に対してこの手法を適用すると，文書の出現頻度が一定であるという仮定が満たされず，うまく検出できないという課題がある [7]．この問題は，特に長期間にわたって収集されたブログに対して適用する場合に顕著となる．

そこで [7] では，Kleinberg の手法で単語と burst 度にのみ依存していた部分を時間の関数に拡張し，時期によってブログ総数が異なるという課題に対応できるように改良している．それにより，ブログ総数が一ヶ月の中では変動しないという仮定のもと，一ヶ月単位で出現する文書数の期待値を変化させることで注目単語が抽出可能であることを示した．しかし，一ヶ月ブログの総数が変動しないという仮定には無理がある．

また，[9] では話題抽出の手法として，トラックバックの仕組みによるネットワーク性に着目し，ある時点で急激に成長する記事クラスタとそのクラスタの扱う話題を特徴語の形で特定する手法を提案している．その結果，ブログ記事ネットワークにおけるトポロジ的なクラスタは，現実の出来事に呼応して急激に成長する時期を持つことを示した．そのため，ある時期で急激に成長しているクラスタと，そこに現れる特徴語に着目することで，その時点で盛り上がっていた話題（emerging topic）とそれを言及する記事との関係性を含めて抽出することに成功している．しかし，盛り上がった話題はクラスタの特徴語によって表されており，特定ジャンルに属する人しか興味を持たない話題を抽出する観点では記事数の多い話題でないと適切な特徴語が抽出できないという問題がある．

†早稲田大学基幹理工学研究科情報理工学専攻
‡NTT 未来ねっと研究所

本論文で提案する手法は，ウェブ上に存在する全ブログを対象にするのではなく，あらかじめユーザが決めた特定のジャンルのブログのみを用いて解析を行う点，ユーザ別に特徴語を重み付けし，共起解析を行うことで，全ブログで話題になっていない話題も抽出できる点でこれらの研究と異なっている．

3 提案手法

本手法では，特定ジャンルのブログを解析し，一定期間ごとの話題を抽出する．対象とする解析期間を I と表し，それを順に一定期間で分割したものを $I_1, I_2, \cdots, I_K$ と表す．期間の最小単位は 1 日で，一定期間を N 日，つまり $|I_n| = N$ とした場合，$|I| = N \times K$ である．本手法では，以下の大きく分けて 5 つの処理を行う．

3.1 特定ジャンルに属しているあるユーザの 1 日の記事に対して形態素解析を行い，共起対象名詞のリストを作成する．

3.2 3.1 で作成された名詞のリストに対して共起解析を行う．

3.3 共起解析の結果を元に，記事に出現する共起対象名詞の TF・IDF 値を算出し，共起語ペアのスコアを算出する．

3.4 3.1〜3.3 の処理をその日に記事を書いている特定ジャンルに属する全てのユーザに対して行い，各 I_n ごとに集計し，日ごとの共起語ペアのスコアの平均を計算する．

3.5 3.1〜3.4 を繰り返し，結果の差分から I_n ごとの特徴語ペアを抽出し，特徴語ペアに対する補足語を抽出して対象期間の話題を示す．

以下，各処理の詳細について述べる．なお上記の処理の項番は以下の節の番号に対応する．

3.1 形態素解析

特定ジャンルに属すあるユーザの 1 日の記事（1 日に複数の記事を書いている場合は一つの記事として扱う）に対して，そこに記述されている内容を抽出するために形態素解析を行う．本論文でユーザーとは，特定ジャンルに記事を書いている者を意味する．形態素解析ツールとしては，Sen[5] を用いる．ブログには，辞書に含まれていない多くの固有名詞が書かれているので，本手法で用いた Sen の辞書は，IPA 辞書 [1] に Wikipedia[3] のすべての項目名を固有名詞として追加した辞書を使用する．Sen によって形態素解析を行い，結果から名詞のみを抽出し，記事に出てきた順でリストを作成する．その際，代名詞，数詞，非自立名詞はリスト作成には除くものとする．また，連続して出現している名詞は複合語である名詞が分割されたものと考え，結合し 1 つの名詞としてリストに登録する．リストに登録された名詞を本論文では共起対象名詞と呼ぶ．

3.2 共起解析

単語同士の共起の回数をカウントする．例として「日本の男性ピアニストはオスカー・ピーターソンに憧れてピアノを始める」という文があるとき，形態素解析から “日本”，“男性”，“ピアニスト”，“オスカー・ピーターソン”，“ピアノ” という共起対象名詞が抽出される．このとき，本手法では共起の発生を以下のように定義する．

- 共起対象名詞によるリストがあるとき，ある名詞とその直後に出現する名詞の共起が発生したとする．

つまり，“日本” と “男性ピアニスト”，“男性ピアニスト” と “オスカー・ピーターソン”，“オスカー・ピーターソン” と “ピアノ” が，一回ずつ共起したとする．このようにして共起解析を行い，ユーザの 1 日の記事で発生した共起をカウントする．以下，単語 w_1 と w_2 が共起したことを $p = (w_1, w_2)$ と表わし，共起ペアと呼ぶ．

3.3 共起語ペアのスコア算出

3.3.1 ユーザ別 TF・IDF 値の算出

ユーザの 1 日の記事で抽出されたすべての共起対象名詞に対して TF・IDF 値を算出する．本研究では，あるブログ記事 A における単語 w の TF・IDF 値 $tfidf_w^A$ を式 (1) によって算出する．

$$tfidf_w^A = \log(tf(w, A) + 1) * \log\left(\frac{N}{df(w)}\right) \quad (1)$$

ここで，$tf(w, A)$ はブログ記事 A 中に単語 w が出現する頻度，$df(w)$ は Web において単語 w が出現しているページ数，N は全 Web ページ数を示す．提案手法では，

A をユーザの 1 日の記事，w を A における一つの共起対象名詞，$tf(w, A)$ を A 中に単語 w が出現する回数，$df(w)$ を Yahoo が提供する API[4] によってレスポンスされる w の web 検索結果数，N を 1 兆とした．これは，Google が 2008 年 7 月時点でインデックスしたページ数 [6] である．式 (1) の後半の $\log\left(\frac{N}{df(w)}\right)$ は $19.152364382498185 \geq \log\left(\frac{N}{df(w)}\right) \geq 0$ の値をとる．

なお，TF 値を求める際，$\log(tf(w, A) + 1)$ を用いた．この理由として，特定ジャンルに関する特徴的な語であっても，それが Web 上で多く使われていると IDF 値が大きくならず，一般的な語の IDF 値と差が生まれないことがある．その際，log を取らず $tf(w, A)$ を使うと，TF 値だけに TF・IDF 値が大きく左右され，特定ジャンルに関する語の TF・IDF 値が上位になりにくくなるからである．

3.3.2 共起語ペアのスコア算出

本手法ではある日 d に共起した $p = (w_1, w_2)$ に対して式 (2) によってスコアを算出する．

$$m_p^d = (tfidf_1^A \times tfidf_2^A) \times |p| \quad (2)$$

表 1: 実験手法一覧

No.	TF・IDF 対象	集計方法	スコア計算手法
1	TF・IDF 未使用	(4)	共起回数
2	TF・IDF 未使用	(3)	共起回数
3	全体	(4)	(5)
4	全体	(4)	(2)
5	全体	(3)	(5)
6	全体	(3)	(2)
7	ユーザごと	(4)	(5)
8	ユーザごと	(4)	(2)
9	ユーザごと	(3)	(5)
提案手法	ユーザごと	(3)	(2)

ここで式 (2) に含まれる $tfidf_1^A$ と $tfidf_2^A$ はそれぞれブログ記事 A における共起ペア $p = (w_1, w_2)$ の w_1 と w_2 の TF・IDF 値である．

式 (2) で，$tfidf_1^A$ と $tfidf_2^A$ を乗算しているため，二つの語の TF・IDF 値が共に大きくないと m_p^d の値は大きくならない．よって，二つの語が共に，特定ジャンル内で重要とみなされる場合のみペアの m_p^d の値が大きくなる．

3.4 スコア集計

対象の日に記事を書いたすべてのユーザの m_p^d を加算する．次に，m_p^d の I_n での平均 $m_p^{I_n}$ を次式を用いて算出する．

$$m_p^{I_n} = \frac{\sum_{d \in I_n} m_p^d}{\text{期間 } I_n \text{ で } p \text{ が発生した日数}} \quad (3)$$

3.5 話題の抽出

I_n ごとの特徴語ペアと補足語を抽出し，話題を示す．この処理では，すべての p において式 (3) を用いて，$(m_p^{I_{n-1}}) - (m_p^{I_n})$ が最も高かった p を I_n の特徴語ペアと呼び，$t_{I_n} = (w_1, w_2)$ と表わす．$t_{I_n} = (w_1, w_2)$ のそれぞれの対象期間に共起した語の内，IDF 値が高いもの上位 5 つ (高いものが 5 つ未満であればそれら全て) を抽出し，補足語とする．特徴語ペアと補足語を組み合わせて話題とする．

4 実験

4.1 実験手法

本研究では，本手法の特定ジャンルのブログに対する効果を明確にするため，

他の単純な手法を含めて全通りの組み合わせを比較対象とし実験を行った．

本実験では特定ジャンルのブログデータセットとして，「音楽」ジャンルのブログ 2008 年の約 1 年分（ユーザ数 2694 記事数 366736）を使用した．また，$|I| = 7$ として話題を抽出する．特定ジャンルのブログで話題抽出を行った関連研究はないので，本提案で使用したスコア計算手法，集計法ならびにスコア集計対象を変えて，それぞれの効果を比較した．具体的には，(a) 本手法が TF・IDF を「ユーザごと」に行っているのに対し，TF・IDF を行わない「TF・IDF 未使用」，ユーザごとに区別せず 1 日に書かれた記事すべてをまとめて TF・IDF を計算する「全体」を用いた手法，(b) 本手法が第 3.4 節で式 (3) を使用しているのに対し，式 (4) を用いた手法，

$$m_p^{I_n} = \sum (m_p) \quad (4)$$

さらに，(c) 本手法が第 3.3.2 節で式 (2) を使用しているのに対し，$tfidf_1^A$ と $tfidf_2^A$ を加算する式 (5) を用いた手法，

$$a_p = (tfidf_1^A + tfidf_2^A) * |p| \quad (5)$$

をそれぞれ組み合わせたものを比較手法とした．評価方法として，359 日（1/8〜12/31）分の抽出された話題を提示されることによって，対象期間に何があったか推測可能かを判定し，推測できた総日数を有効話題提示数と呼び，結果とした．提示された語によって対象期間の音楽ジャンルの話題が推測できたとする条件は，表 2 の条件の「」内に，提示された特徴語ペアと補足語の中の一つを選択して入れ（「」内には少なくとも一つは $t_{I_n} = (w_1, w_2)$ の w_1 もしくは w_2 を選択しなければならない），音楽ジャンルに関係する文が完成できたものとした．

表 2: 話題が推測できたとする条件と推測できた例

条件	抽出できた例
「いつ」「誰（何）」が「何」で「何」をした．	「今日」「バーミンガム市交響楽団」が「ヴァイオリンリサイタル」で「後期弦楽四重奏曲」をした．
「誰（何）」が「何」を「何」した．	「ヘイセイジャンプ」が「真夜中のシャドーボーイ」を「発売」した．「ハンス・フォンク」が「チャイコフスキー」を「レコーディング」した．
「いつ」「何」が○○した．	「今日」「雪」が降った．「明日」「バレンタインデーライブ」がある．「昨日」「バレンタインデー」だった．
「誰（何）」が「何」で「何」をした．	「アカデミー室内管弦楽団」が「ネヴィル・マリナー指揮」で「クリスマスコンサート」をした．「YOSHIKI」が「東京ドーム公演」で「発売記念ライブ」をした．
「誰（何）」が「何」をした．	「Buono!」が「リリース記念イベント」をした．「THE ポッシボー」が「ファーストソロコンサート」をした．

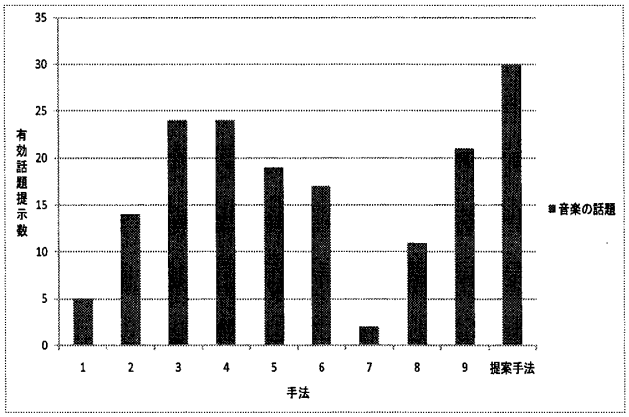

図 1: 実験結果

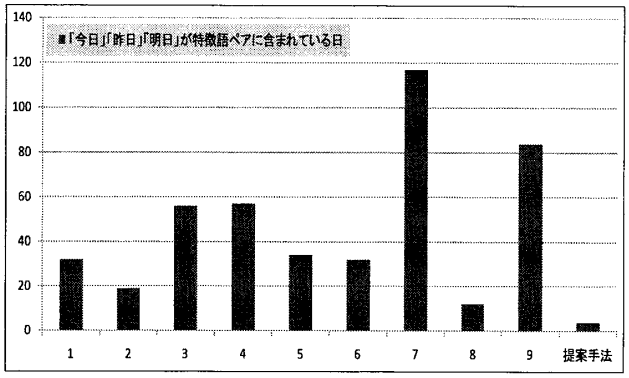

図 2: 特徴語ペアに含まれる「昨日」,「今日」,「明日」の個数と割合

4.2 実験結果

結果を図 1 に示す．図 1 より，本研究の提案手法は，特定ジャンルの有効話題提示数が大きいことがわかった．実際に抽出できた例を表 2 に示す．

結果の中で特徴的なものを二つあげる．第 1 は，12 月 17 日の各手法の $t_{I_n} = (w_1, w_2)$ は「集計方法」に式 (4) を用いた手法 1，3，4，7，8 は，すべて w_1 =「送料無料」，w_1 =「買い物マラソン」という記事内に書かれた広告と考えられる語が t_{I_n} となっていた．逆に,「集計方法」に式 (3) を用いた手法 2，5，6，9 と提案手法の中では，手法 9 以外で広告と考えられる語が t_{I_n} になることを回避している．さらに，手法 2 と提案手法では広告が入った日でも話題を推測することができている．

第 2 に，傾向として，式 (5) を使用した手法の t_{I_n} には，式 (2) を使用した場合に比べ「昨日」,「今日」,「明日」という語が含まれる割合が増え，特に「TF・IDF 対象」を「ユーザごと」にしたものは，顕著にその傾向が現れている（図 2 参照）.「昨日」,「今日」,「明日」が含まれる特徴語ペアの相手は,「ひな祭り」「バレンタインデー」などの行事関係の語が多く，他には「雪」「オリンピック開催」などが存在した．

4.3 考察

前節で述べた，提案手法が比較手法よりも多く話題を提示できた理由を考察する．第 1 に，の特徴的な結果が出た理由として，1 日に大量の同じ広告が記事内に書かれた場合，広告に用いられた語の共起語ペアのスコアは高くなるが，平均をとることによって急激な共起語ペアのスコアの上昇を抑えることができたと考えられる．よって「集計方法」に式 (3) を用いると，1 日に大量の同じ広告が記事内に入り間違った特徴語が上位になることを軽減でき，特定ジャンルの有効話題提示数が増加したと考えられる．

第 2 の特徴的な結果の理由として,「昨日」,「今日」,「明日」は TF・IDF 値が低いが，行事や，ほぼすべてのユーザに起こったイベントでは，そのイベントに関係する語の急激な TF・IDF 値の上昇につられて「昨日」,「今日」,「明日」が含まれる p のスコアが大きくなったと考えられる．式 (2) を使用した場合では，第 3.3.2 項で述べたように，二つの語の TF・IDF 値が共に大きくならないと m_p^d も大きくならないので,「昨日」,「今日」,「明日」という語が含まれる t_{I_n} の割合が少ないと考えられる．代わりに，特定ジャンルに関する t_{I_n} が多くなり，式 (5) を使用した手法と比べ，式 (2) を使用した手法は特定ジャンルの有効話題提示数が増加したと考えられる．

5 結論

本研究では，特徴語抽出の手法として一般的に用いられる TF・IDF を特定ジャンルの各ユーザに対して用い，ブログの共起関係を調べることによって対象期間の話題を抽出する手法を提案した．

音楽ジャンルのブログ 1 年分に対し，評価を行い，他の手法と比較した結果，提案手法がより多くの特定ジャンルに関する話題を提示する点で他の手法を上回る結果を示した．このことから，目的であった特定ジャンルに関する話題をより多く抽出する手法の提案が達成できた．

現在は音楽のジャンルを対象としたが，今後は多様なジャンルのブログをデータセットとし，評価・検証する予定である．また，本研究では提示する話題の精度や推移に関しては考慮していないので，二部グラフやコメントによるブログネットワークの作成を行い，精度の向上や話題推移の発見の手法も検討していきたい．

参考文献

[1] http://sourceforge.jp/projects/ipadic/.

[2] http://kizasi.jp/.

[3] http://ja.wikipedia.org/.

[4] http://developer.yahoo.co.jp/.

[5] https://sen.dev.java.net/.

[6] Jesse Alpert. We knew the web was big... *The Official Google Blog*, 20080725. http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html.

[7] Toshiaki Fujiki, Identification of bursts in a document stream, *Proceedings of First International Workshop on Knowledge Discovery in Data Streams, 2004*, 2004.

[8] J. Kleinberg, Bursty and hierarchical structure in streams, *Proc. of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002*, 2002.

[9] Makoto Uchida, Extracting and visuzalization of a emerging topic from the blogspace, *Proceedings of the Annual Conference on JSAI (CD-ROM)*, 2006.

[10] 戸田智子, Blog 記事のクラスタリングによるカテゴリ別話題変遷パタンの抽出, 電子情報通信学会データ工学ワークショップ DEWS2007, 2007.