DEWS2006 1B-i4

blogにおける人物に関する”旬な”話題の抽出

外間　智子†　　北川　博之†,††

† 筑波大学システム情報工学研究科　〒 305-8577 茨城県つくば市天王台１–１–１
†† 筑波大学 計算機科学研究センター　〒 305-8577 茨城県つくば市天王台１–1–１
E-mail: †{tomokoh,kitagawa}@kde.cs.tsukuba.ac.jp

あらまし　インターネットの発達とともに、既存のメディアに頼らず、個人が情報を発信できる環境が整ってきている。中でも blog は、手軽な情報発信の手段として最近特に注目を集めている。ツールの普及、ホスティングサービスの増加により、blog のユーザ数は急増を続けており、大手メディアを介さず、世の中の関心をリアルタイムに反映する新たなメディアとしても注目を集めている。blog を対象とした既存の研究には、コミュニティ抽出、バーストの検出などがあるが、一方で、特定の対象 (人物、企業、商品など) に特化した情報収集・集約も今後重要となってくると考えられる。本研究では、とくに人物情報の収集に焦点を当て、人物についてのリアルタイムな話題を抽出することを目指す。

キーワード　ブログ, トピック検出, クラスタリング

Detecting “Hot” Topics about a Person from Blogspace

Tomoko HOKAMA† and Hiroyuki KITAGAWA†,††

† Graduate School of Systems and Information Engineering, University of Tsukuba　Tennodai 1–1–1, Tsukuba-shi, 305–8577 Japan
†† Center for Computional Sciences, University of Tsukuba　Tennodai 1–1–1, Tsukuba-shi, 305–8577 Japan
E-mail: †{tomokoh,kitagawa}@kde.cs.tsukuba.ac.jp

Abstract　Explosive growth of the Internet technology allows indivisuals to publish information without depending on mass media. Especially Weblogs, or ”blogs” are gathering attention as a handy way to put out information. Proliferation of blog tools and hosting services has brought a sharp rise in the number of bloggers and blogs have gained much attention as new media reflecting real-time interest of people. While existing researches focus on community detection, burst detection and so on, it will become important to collect and summarize information about a particular object (e.g. person, business organization, product). In this paper, we focus especially on information collection and summarization about people. Our purpose is to extract topics about particular person.

Key words　Blogspace, Topic Detection, Clustering

1. はじめに

インターネットの発達とともに、既存のメディアに頼らず、個人が情報を発信できる環境が整ってきている。中でも blog は、手軽な情報発信の手段として最近特に注目を集めている。ツールの普及、ホスティングサービスの増加により、ブロガー (ブログを更新する人) 数は急増を続けており、BlogFan.org(注1)の集計によれば、2005 年 11 月現在、国内ホスティングサービスを利用しているユーザ数はのべ 130 万人を超えている(注2)。

日本における blog には、Web 上で綴る日記という側面と、個人が日常生活で触れた興味深いニュース・商品などを意見・感想とともに記録する媒体、という側面がある。後者としての blog は、大手メディアを介さず、世の中の関心をリアルタイムに反映する新たなメディアとして注目を集めている。blog から有用な情報を抽出したい、という需要は高く、現在、様々なアプローチから研究が始められている [2]〜[4], [8], [9]。

blog を対象とした既存の研究には、コミュニティ抽出、バーストの検出などがあるが、これらはいずれもすべての blog 記事を何らかの手法により集約して提示する、という大域的な視点に基づいている。しかし一方で、特定の対象 (人物、企業、商品など) に特化した情報収集・集約も今後重要となってくると

(注1)：http://www.blogfan.org
(注2)：一ヶ月以内に更新された blog を対象としている

考えられる。本研究では、特定の対象、とくに人物情報の収集に焦点を当てる。

人物に関する情報源としては、人物情報データベース、Wikipedia、公式 HP などがあるが、これらは基本的に静的・公式な情報源である。これらの情報源からは得られない動的で非公式な情報として、最近の話題や評判情報などがあるが、blog はその速報性や口コミ的な性質から、こうした情報を収集する際に非常に適した情報源といえる。

TDT をはじめとして多くのトピック抽出に関する研究があるが、特定の対象や人物に特化したトピックの抽出に関する研究はこれまで行われていない。従来のトピック抽出手法を用いて特定の人物に特化したトピック抽出を行う場合、まずはその人物に関する記事を収集する必要がある。しかし、blog 記事等 Web 上の一般的な文書においては、ある人物は様々な呼び名で参照され、また同姓や同名の人物が存在するため、目的人物に関する記事をもれなく、正確に収集することは困難である。このことにより、トピック抽出において「トピックが正しく抽出できるか」「それぞれのトピックの規模 (記事数) が測れるか」という問題が生じる。この問題に対し、本研究では、トピックの抽出とクラスタ規模の推測の 2 つのフェーズからなる手法を提案する。blog という情報源の性質を考えると、「トピックの規模 (関連記事数)」は、「その話題がどの程度関心を集めているか」を示す指標とみなすことができ、「どのようなトピックが抽出されるか」と同様、重要な情報といえる。

2. 関 連 研 究

新聞記事より話題の抽出を行うものに、TDT(Topic Detection and Tracking) の手法がある。タイムスタンプが付加されたニュース記事からトピックを検出する試みであり、文書クラスタリング手法を用いた手法など、いくつかのアプローチがある。蓄積した過去の記事が利用可能であり、リアルタイム処理を目的としないのであれば、階層的クラスタリングに時間的要素を組み合わせた手法が高い精度を示すことが報告されている [6]。

トピックを検出するのに、記事集合そのものではなく、記事集合からトピックを表現する特徴語を抽出し、それら用いるというアプローチもある。TimeMines システム [1] では χ^2 検定を用いて新聞記事から際だった出現頻度をもつ特徴語を抽出、それらのグルーピングを行い、特徴語の集合をトピックとして検出している。さらに、TimeMines システムは抽出したトピックを年表形式でユーザに提示するインタフェースも提供している。ただし、正しくトピックを検出するためには特徴語 (人名・地名などの固有表現、名詞句) が正確に抽出される必要があるが、そのためのツールがうまく機能しないために、テストコーパスの一部が使用できなかったと報告されている。口語表現が混在するブログ記事においては、固有表現抽出等が新聞記事以上に困難であると考えられる。

クラスタリングを用いずに記事集合からトピックを検出する手法として、他に burst の検出がある。これは、すべての単語の定常状態の出現頻度を保持しておき、それから大きく外れて高頻度で出現する単語をトピックワードとして提示する手法である [7], [9]。burst 情報を利用してトピックを検出する手法 [5] も提案されているが、burst を検出するためには大量のコーパスが必要となる。本研究では特定の人物に関する話題を抽出することが目的のため、必要な量のコーパスが各対象について用意できるとは考えにくく、従って単語の burst を利用する手法の適用は難しい。

これら既存の手法は通常、考慮する期間内に書かれた全記事を対象としてトピックの抽出を行う。これらの手法を適用するためには、まずその人物に関する記事を全記事の中から選択する必要があるが、ニュース記事等公式な文書を除き、一般に人物は様々な呼び名で参照されるため、必要な記事をトピック抽出以前にもれなく正確に収集することは困難である。

3. 研究の目的とアプローチ

3.1 研究の目的

本研究は、指定された期間内に書かれた blog 記事から特定の人物についての話題を抽出し、併せてその話題についての関連記事数を推測することを目的とする。以下、3.2 節で関連記事の収集、3.3 節でトピック抽出へのアプローチついて述べる。

3.2 関連記事の収集

新聞記事であれば、ある人物は通常フルネームで表記されるので、それらを用いて検索することで、特定の人物についての記事を収集することは比較的容易である。しかし blog など Web 一般の文書においては、一般にある人物に対する表記方法はフルネーム、姓のみ、愛称など様々なものがある。例えば、全ての記事もしくは人物の姓や名のみを含む記事すべてを対象とすると、関係のない記事 (ノイズ) が多数含まれてしまい、対象人物の情報が埋もれてしまう。逆に、対象人物のフルネームを含む記事のみを考慮すると、同姓同名の人物がいなければ確実に対象人物の記事が収集できるが、多くの関連記事を見落としてしまう。このように、対象人物に関する記事の収集が困難であるため、「トピックが正しく抽出できるか」「それぞれのトピックの規模 (記事数) が測れるか」という問題が生じる。この問題に対応するため、本研究では、トピック抽出と規模 (記事数) 推測の 2 段階に分けて処理を行う。具体的には、確実に対象人物について記述している少数の記事を用いてトピックの抽出を行い、次にそれらに対し類似している記事を付加することで関連記事の収集を行う。トピック抽出の対象とする記事は、対象人物のフルネームを含む記事、とする。現段階では、同姓同名の人物は考慮していない。

3.3 トピック抽出

文書クラスタリングの手法を用いることで、出現単語の分布が「似ている」記事の集合を抽出することが可能である。しかし現実のニュースや blog 記事を考えると、出現単語が似ていてもかなり期間が空いているような記事同士は、違うイベントについて言及している場合が多い。特定の人物に着目する場合、その傾向は顕著に現れる。たとえばあるスポーツ選手が関わる複数の試合についての記事は、記事に出現する単語という点では類似しているであろう。単純に似ている記事の集合 (カテゴ

リ) を抽出するか、時間軸をもつ現実のイベントに対応した記事の集合 (トピック) を抽出するかは、目的にもよるが、blog のようなその時々の世相を反映する情報源の場合、後者がより有用であると考えられる。これを踏まえ、提案手法ではタイムスタンプを考慮したクラスタリングを行う。

4. 提 案 手 法

本節では、提案手法の詳細について述べる。提案手法は、「目的人物のフルネーム」を入力としてとり、目的人物に関するトピック (記事のクラスタ) とその期間および関連記事数を出力する。

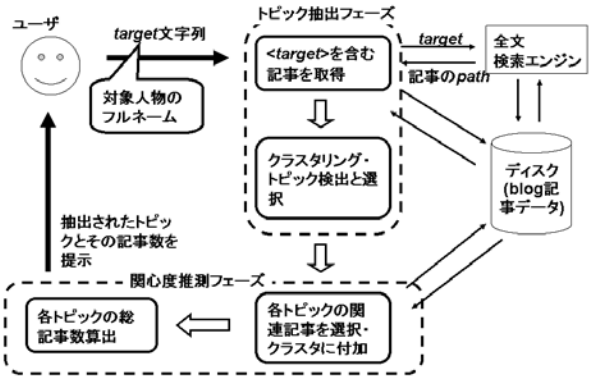

図 1　提案手法の流れ

4.1 提案手法の概要

提案手法は、トピック抽出と規模推測 (関連記事収集) の 2 つのフェーズからなる。トピックの関連記事数はそのトピックに対する世の中の関心の度合を示す指標であると考えられるため、関連記事の収集を行うフェーズを以下では「関心度推測フェーズ」と呼ぶ。まず目的人物のフルネーム $target$ が与えられると、$target$ を含む記事に対しトピック抽出フェーズを適用し、話題を抽出する。次に関心度推測フェーズを介することで各トピックについての関連記事数を収集する。最後に、抽出されたトピック (関連記事、期間) を記事数とともに出力する。提案手法の処理の流れを、図 1 に示す。以下、トピック抽出フェーズ、関心度推測フェーズについてそれぞれ詳しく述べる。

4.2 トピック抽出フェーズ

トピック抽出フェーズでは、まず目的人物のフルネーム $target$ を用いて対象期間内の blog 記事を検索し、記事集合を得る。次に、検索した全記事を $window$ 行単位でセグメントに分割する。1 セグメントを 1 文書とみなし、$target$ を含む文書を対象に形態素解析(注3)を行って文書ベクトルを作成する。1 記事の長さがまちまちであり、また 1 つの記事中で話題が変化することを考慮し、このように記事をセグメントに分割する。ベクトルの各要素は情報検索で一般的に用いられる tf/idf で重みづけられた値とし、文書 d における単語 t の重み w_t^d は以下の式により算出する。

$$w_t^d = \frac{\log_2(tf(t,d)+1)}{\log_2 |W_d|} \cdot (\log_2 \frac{N}{df(t)})$$

ここで、$tf(t,d)$ は文書 d 中に出現する単語 t の頻度、$|W_d|$ は d 中に出現する索引語の異なり数、$df(t)$ は単語 t が出現する文書数、N は文書集合中の全文書数である。

次に、作成した文書ベクトル群を対象に凝集型の階層的クラスタリングを適用する。凝集型の階層的クラスタリングは、一つの要素からなる初期クラスタ群から出発し、最も似ているクラスタペアの併合を繰り返すことでクラスタを徐々に大きくしていく手法である。最終的には全要素を含むような 1 つの巨大なクラスタが生成されるが、適当な終了条件を設定することで任意の階層でのクラスタを得ることができる。本手法では、最も似たクラスタペアの類似度がある閾値以上になった時点でループを終了する。3.3 節で述べたように、本手法では通常の階層的クラスタリングを拡張し、タイムスタンプを考慮したクラスタリングを行う。具体的には、クラスタごとに併合許容期間を定め、類似度計算の後に 2 つのクラスタの許容期間が重なっているかどうかを考慮し、併合するかどうかを決定する。併合しない、と判断された場合は、次に類似度の高いクラスタペアについて同様にチェックを行う。各クラスタの併合許容期間の算出方法は以下の通りである。

クラスタ C に含まれる記事のうち、最も古い記事のタイムスタンプ $ts_s(C)$ から最も新しい記事のタイムスタンプ $ts_e(C)$ までを、クラスタ C の期間とする。また、クラスタ C の期間の長さ $len(C)$ を $ts_e(C) - ts_s(C)$ と定める。このとき C の併合許容期間を、

$ts_s(C) - max(\alpha{\cdot}len(C), \beta)$ to $ts_e(C) + max(\alpha{\cdot}len(C), \beta)$

とする (式中の α,β はパラメータ)。基本的にクラスタの期間の長くなれば、併合許容期間も長くなる。クラスタが小さい初期の時点で $len(C)$ が 0 もしくはそれに近い値をとり、クラスタの期間と併合許容期間がほぼ等しくなってしまう場合を考慮するため、max をとっている。これにより各クラスタの併合許容期間は、$len(C)+2\beta$ 以上の長さをもつことが保証される。2 つのクラスタを併合する/しない場合の図を図 2 に示す。横軸は時間軸を表す。

クラスタリングが終了すると、次に、抽出されたクラスタのうち何らかのトピックを表現しているようなクラスタを選択する。blog 記事には書き手の個人的な内容を綴っているものも多いため、小規模なクラスタ (特に、1 つの記事しか含まないようなクラスタ) が多数出てくることが予想される。こうしたクラスタは一般的なトピックとは言い難いため、含む記事数が非常に少ないクラスタは無視し、一定数以上の記事を含むようなクラスタをトピックとして選択する。

4.3 関心度推測フェーズ

関心度推測フェーズでは、トピック抽出フェーズで抽出された各トピックについて、関連記事の収集を行う。関連記事の収

(注3)：形態素解析ツール「茶筌」
(http://chasen.naist.jp/hiki/ChaSen/) を用いる。

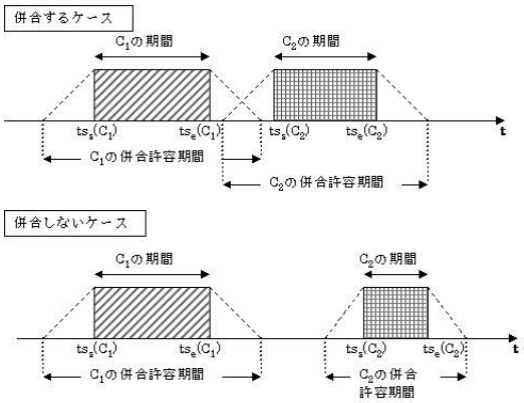

図 2　クラスタ C_1,C_2 を併合する/しない場合

集は、いずれのクラスタにも含まれなかった残り全記事より、各クラスタごとに付け加えるべき記事を選択することで行う。その際、各クラスタにはトピック抽出フェーズで定義したように「併合許容期間」があるため、実際には全記事を対象とするのではなく、その期間内に書かれた記事のみから選択すればよいと考えられる。ある記事 a がクラスタ C に追加されるかどうかの判定は、基本的にクラスタ C に含まれる全記事と記事 a から作成したベクトルとの平均類似度を計算することで行う。しかし単純に類似度の高い記事を全て付け加えてしまうと、関連のない記事が多数付加されてしまう。そのため、ベクトルの類似度とは別の観点から付加すべき記事を絞ることが必要である。絞り込む方法としては、人物の「姓」「名」のいずれかを含むかどうか、という条件を用いた。

具体的には、以下の手順でクラスタの拡張を行う。

（1） 目的人物の「姓」もしくは「名」を含む記事を検索

（2） 記事 a について、a のタイムスタンプ ts_a と各トピックの併合許容期間を比較し、a の追加先候補となるクラスタ集合 $S = C_1, C_2, ...C_n$ を得る

（3） a と S に含まれるクラスタとの類似度をそれぞれ計算

（4） a と最も類似度の高いクラスタ C_k を選択

（5） a と C_k の類似度が閾値以上であれば、a を C_k に追加

関心度推測フェーズの処理の例を、図 3 に示す。図で、横軸は時間軸を、a_i は記事を、C_j はトピック抽出フェーズで抽出されたトピックを、$sim(a_i, C_j)$ は a から作成したベクトルと C に含まれるベクトルとの平均類似度を表す。また、長方形は各クラスタの併合許容期間を表す。図は、記事 a_3 の追加先候補のクラスタが C_1,C_2,C_3 の 3 つある状況を示している。

5. 実　　験

提案手法の有効性を検証するため、実際の blog 記事を用いていくつかの実験を行った。本節では、まず使用したデータおよびパラメータについて述べ、次に 2 フェーズ処理の有効性と、タイムスタンプを考慮したクラスタリングの有効性を検証する。

5.1　データおよびパラメータ設定

実験に使用したデータとパラメータ設定について以下に示す。

実験データとして、Web をクローリングして取得した blog

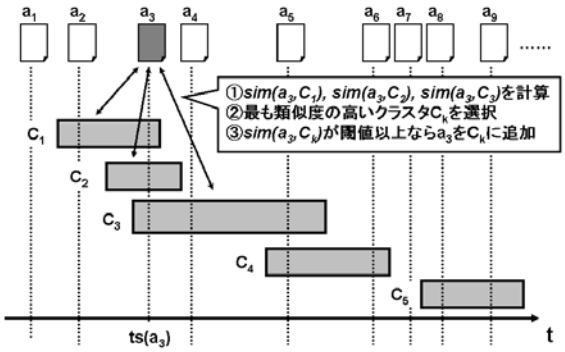

図 3　関心度推測フェーズ

記事 1639760 件 (2004 年 10 月 16 日-2004 年 12 月 31 日) を使用した。個々の記事にはタイムスタンプが付加され、また HTML タグ等は含まないプレーンテキストである。パラメータの設定については以下に示す。

- トピック抽出フェーズのパラメータ
  - 記事を分割する際の 1 セグメントの大きさ: 10 行
  - クラスタリング終了条件の類似度: 0.05
  - クラスタ併合許容範囲を算出する際のパラメータ α: 0.5, β: $60 \times 60 \times 24 \times 2$ (2 日)
  - トピックとして抽出するクラスタが含むべき最低記事数: 3 件
- フェーズ 2 のパラメータ
  - クラスタ内記事と記事ベクトルの平均類似度の閾値: 0.05

5.2　2 フェーズ処理の有効性の検討

2 フェーズ処理の有効性を検討するため、以下の 3 つの手法でトピック抽出を行い、関連記事数を求めた結果を比較した。

(**手法 1**)　目的人物の「姓」もしくは「名」を含む記事すべてを対象としてトピック抽出

(**手法 2**)　目的人物のフルネームを含む記事のみを対象としてトピック抽出 (提案手法のトピック抽出フェーズのみ)

(**手法 3**)　目的人物のフルネームを含む記事のみを対象としてトピック抽出、その後 4.3 節にしたがいクラスタを拡張 (提案手法)

まず、検出されたトピックについて比較を行う。手法 3 は手法 2 で抽出したトピックについて拡張を行うため、手法 2, 手法 3 は同じトピックが検出される。したがって、手法 1 および手法 3 について比較を行う。

例として、「松井秀喜」(MBL 選手) について手法 1 と手法 3 で抽出されたトピックを表 1,2 に示す。それぞれ、関連記事数の多かった上位 10 件を示している。同様に、「堀江貴文」(ライブドア前社長) についてトピック抽出を行った結果を表 3,4 に示す。

各トピックには、人手でラベルをつけている。また、表中の”???” はどのような話題を表しているのか判断できなかったもの、” 広告記事” は長い商品宣伝文句が末尾に付加されているような記事 (いわゆる blog スパム) からなるクラスタである。表 1,3 の () でくくられたトピックは、対象人物とは別人についてのトピックもしくは複数人物についての記事が混在し

表 1　*target* < 松井秀喜 >(手法 1)

クラスタリング対象記事数		1716			
検出トピック数		36			
ランク	トピック	期間	記事数	不適切な記事数	固有名詞上位 5 件
1.	MLB ア・リーグ優勝決定戦	10/16-12/28	179	62	[ヤンキース 日本 イチロー 米 日]
2.	松井&酒井美紀	10/23-12/1	53	10	[酒井 美紀 ヤンキース ゴジラ ニューヨーク]
3.	???(松井大輔、松井秀喜)mix	10/16-12/26	48		[フランス ルマン 田中 ル・マン ホアキン]
4.	(松井証券)	10/16-12/31	37		[松井証券 日本 ジャパン 中国 アメリカ]
5.	???(松井稼頭央、松井秀喜)mix	10/29-12/31	35		[メッツ ヤンキース 石井 ドジャース 央]
6.	???(松井稼頭央、松井秀喜)mix	10/18-11/23	19		[新潟 中越 日本 央 ヤンキース]
7.	(松井繁 (競艇))	11/18-12/1	11		[山崎 繁 豊 光太郎 智也]
7.	???	11/18-12/29	11		[阪神 新庄 日本ハム 辻本 新庄]
9.	広告記事	12/6-12/9	10		[日本 天田 コスタリカ 剛志 新庄]
9.	(松井繁 (競艇))	12/20-12/23	10		[田中 植木 今村 繁 信]

表 2　*target* < 松井秀喜 >(手法 3)

クラスタリング対象記事数		194			
検出トピック数		10			
ランク	トピック	期間	記事数	不適切な記事数	固有名詞上位 5 件
1.	MLB ア・リーグ優勝決定戦	10/17-10/22	146	5	[ヤンキース 日本 ボストン 大リーグ ニューヨーク]
2.	日米野球	11/3-11/7	32	3	[米 日本 日 ヤンキース 石井]
3.	松井&酒井美紀	11/2-11/3	30	0	[美紀 酒井 ニューヨーク ヤンキース 成田]
4.	田臥勇太 NBA デビュー	11/2-11/7	23	9	[ヤンキース 酒井 フェニックス 美紀 日本]
5.	松井&酒井美紀	11/8-11/11	10	1	[酒井 美紀 静岡 美樹 和夫]
6.	終身契約”100 億円”	12/16-12/16	9	0	[イチロー ヤンキース マリナーズ]
7.	松井&酒井美紀	10/24-10/24	8	1	[酒井 美紀 ニューヨーク ゴジラ 徳光]
7.	???	11/2-11/5	8		[長島 茂雄 笑 イチロー 真紀子]
9.	広告記事	12/6-12/6	7		[日本 新庄 剛志 天田 コスタリカ]
10.	???	11/10-11/13	6		[ヤンキース イチロー ビートたけし 胡麻 松坂]

ているトピックであり、どの人物 (組織) についてのトピックであるかを () 内に示している。対象人物と、別人についての記事が混在しているトピックには、ラベルの末尾に”mix” とつけている。「不適切な記事数」とは、人手で各トピックを評価した際に「このトピックの関連記事としては不適切」と判断された記事の数を表す。なお、「対象人物についての記述を含み」、かつ「そのトピックについてのニュース記事の引用、意見、コメント等を含む」場合に「トピック内関連記事として適切である」と判断している。「不適切な記事数」が空欄になっているのは、”???”、” 広告記事”、および複数人物についての記事が混在しているトピックなど、不適切な記事数の評価ができないものである。また、「固有名詞上位 5 件」は、記事中より抽出された人名、地名などの固有名詞(注4)のうち、出現頻度の高かった上位 (最大)5 個を示している。

表 2 の 2 位の「日米野球」は、松井選手が不在の日米野球についての話題であり、不在であることが話題になっているという点が興味深い。松井選手の国民的な人気ぶりをうかがわせるトピックであり、blog という情報源ならではの話題といえよう。表 2 の 4 位の「田臥勇太 NBA デビュー」は、松井と直接的な関わりはないが、田伏選手の NBA デビューにからめ、アメリカのスポーツ界で活躍する松井秀喜、イチローらに言及した話題である。これも「日米野球」と似た性質の話題である。

表 1 より、手法 1 を用いた場合では別人に関するトピック、対象人物と別人についての記事が混在しているトピックかなり含まれている (上位 10 件中 6 件) ことがわかる。また、「松井秀喜」「松井大輔」「松井稼頭央」という 3 名のスポーツ選手に関する記事が混ざっているため、複数の「松井」についての記事が含まれているクラスタも多くみられた。手法 3 で用いた記事集合は手法 1 で用いた記事集合のサブセットであるが、クラスタリング結果を比較すると、手法 3 で抽出されたが手法 1 では抽出されなかったトピックも多い。手法 1 の場合、職種の似ている別人の記事がクラスタリング時にノイズとなったためと考えられる。表 3 と 4 については、上位 10 件については、適切に検出されたトピックの数にそれほど違いがみられなかった。これは、同姓もしくは同名の、職種が似ている著名人が少ないためであろう。

次に、手法 2, 手法 3 について各トピックの関連記事数により比較を行った。その結果を図 4 について示す。手法 1 については、抽出されたトピックが手法 2, 手法 3 と異なるため、比較対象としていない。例として、「松井秀喜」「堀江貴文」「小泉純一郎」(首相)「宮里藍」(プロゴルファー) について比較した結果を図 4 に示す。それぞれのターゲットについて、上位 5 件 (“???” のトピックは除いている) のトピックについてのみ示

(注4)：「茶筌」の辞書に載っている固有名詞

表 3　*target* < 堀江貴文 >(手法 1)

クラスタリング対象記事数		1727			
検出トピック数		47			
ランク	トピック	期間	記事数	不適切な記事数	固有名詞上位 5 件
1.	IT 系企業の球界参入	10/18-12/27	179	12	[仙台 西武 ソフトバンク ダイエー フェニックス]
2.	著作の感想・広告	10/18-12/27	48	7	[近鉄 日本 オン 大阪 奈々]
3.	競馬界参入	10/24-12/28	45	8	[高崎 群馬 高知 ウラ 笠松]
4.	民間有人飛行	12/16-12/24	22	0	[日本 ロシア 米 ブッシュ 米国]
5.	???	10/23-12/24	20		[慶 ヨ 荻野目 笑 かをり]
6.	流行語大賞	12/1-12/4	14	0	[波田 陽 北島 浜口 康]
7.	(赤星貴文、サッカー選手)	11/22-12/6	12		[広島 赤星 磐田 浦和 前橋]
8.	???	12/13-12/29	11		[モン ソニー リエ 有馬 オン]
9.	???	10/23-11/16	10		[日経 木谷 オン ソフトバンク 新潮]
10.	TV 番組「平成教育委員会」出演	12/22-12/29	9	3	[平成 フジテレビ 健一 高島 鈴村]

表 4　*target* < 堀江貴文 >(手法 3)

クラスタリング対象記事数		226			
検出トピック数		17			
ランク	トピック	期間	記事数	不適切な記事数	固有名詞上位 5 件
1.	楽天 vs ライブドア	10/22-11/5	181	1	[仙台 フェニックス パ・リーグ 宮城 東北]
2.	西武球団の売却打診	10/27-11/9	105	26	[西武 コクド 仙台 香港 中国]
3.	著作の感想・広告	10/27-12/21	87	38	[近鉄 日本 オン 雄 ソフトバンク]
4.	流行語大賞	12/1-12/13	39	5	[北島 波田 浜口 陽 アテネ]
5.	???	11/3-11/6	31		[高崎 小寺 群馬]
6.	???	11/10-12/5	27		[ ]
7.	楽天 vs ライブドア	11/2-11/4	21	0	[中内 宮城 銀座 正]
8.	民間有人飛行	12/16-12/17	16	1	[日本 モン ロシア リエ ブッシュ]
8.	高崎競馬場買収問題	11/10-11/12	16	1	[高崎 群馬 弘之 小寺 高知]
8.	「カネで買えないものなんてあるわけない」発言	11/21-11/29	16	2	[日本 リエ モン 朝日新聞 アメリカ 楽天]

している。1 つの棒グラフが、1 つのトピックの記事数を表す。各棒グラフ中の、白い部分がトピック抽出フェーズ終了時での記事数、網掛け部分が関心度推測フェーズでトピックに付加されたの記事数を表す。すなわち、白い部分が手法 2 を用いた場合の記事数、白い部分と網掛け部分を合わせた全体が手法 3 を用いた場合の記事数を表している。ラベル末尾の () 内の数字は、人手で評価した際にそのトピック内の記事としては不適切と判断された記事数を示している。

図 4 より、関心度推測フェーズを介することで各トピックへかなりの数の記事の追加が行われていることがわかる。トピック抽出フェーズ終了段階では、各トピックの記事数は少ないもので 10 件足らず、多いものでも 30 件程度であり、世の中の関心の度合を反映しているとは言い難い。それに対し、関心度推測フェーズ終了時の記事数はトピック抽出フェーズ終了時に比べ格段に増えており、ある程度は関心度が推測できるようになっていると考えられる。例えば「堀江貴文」の場合に着目すると、1 位のトピックは記事数が 15 倍に増えている。また、トピックを抽出した段階では各対象人物についての総記事数は大差ないが、関心度推測フェーズ後では総記事数にかなりの差がみられた。4 人の中で記事数が最も多かったのは「小泉純一郎」であり、次いで「堀江貴文」となっている。これが妥当な結果であるのか、それとも関連記事収集方法に問題があるため総記事数に差が出ているのか、今後の検討課題としたい。また現段階では人物の「姓」もしくは「名」を含む記事のみを追加対象としているため、その人物が「愛称」等で参照されている記事については無視している。したがって実際には、これよりも多くの関連記事が存在していることが予想される。

なお、「小泉純一郎」における「イラク問題」とは、自衛隊派遣期間の延長に関する話題を中心に、米軍のファルージャ攻撃など、特定のイベントではなくイラク問題全般についての話題となっている。トピックの期間も実験データの期間全体にわたっていることが観察された。トピック「北朝鮮」も、拉致問題 – 特に、横田めぐみさんの遺骨が偽物だったという件 – を中心に、様々な北朝鮮問題についてのイベントの集合であった。トピック「日中関係」も同様 (ただし、中心的な話題は日中首脳会談) である。このように、「小泉純一郎」に対しては、トピック・イベントというよりは、「カテゴリ」に近いものが多く観察された。政治・外交のイベントは多くが複雑に絡み合っており、また blog 上で様々な立場から意見が交わされているため、厳密にイベントに分けることは困難であるためと考えられる。

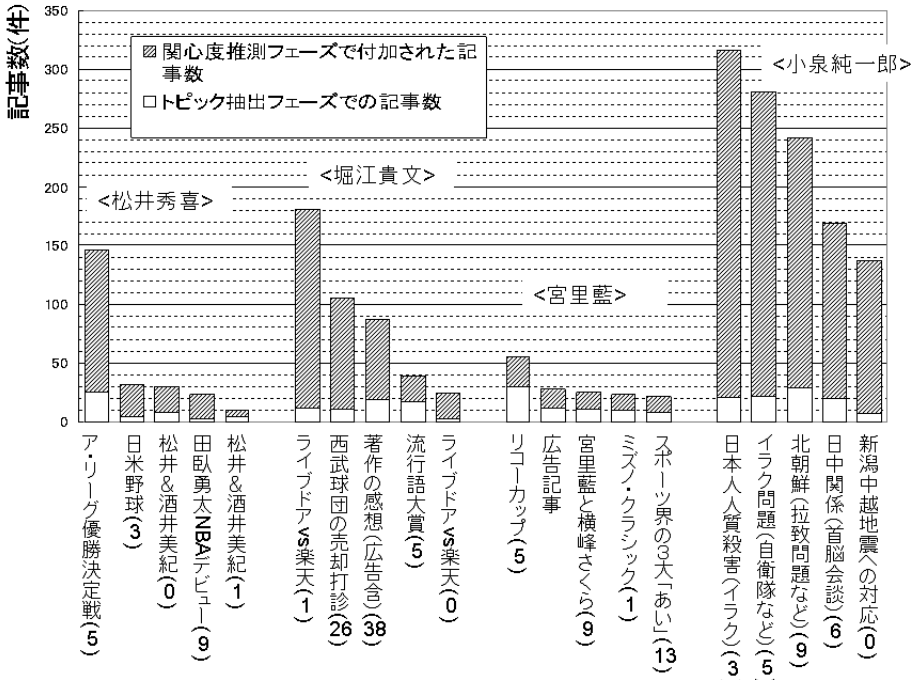

図 4　記事数の比較

5.3　タイムスタンプを考慮したクラスタリングの有効性の検討

提案手法では、トピック抽出フェーズにおいてタイムスタンプを考慮したクラスタリングを行っている。タイムスタンプを考慮しない場合との比較のため、通常のクラスタリングを行うとどのようなトピックが検出されるか実験を行った。タイムスタンプを考慮した場合との比較のため、「松井秀喜」「堀江貴文」「宮里藍」についての結果を表 5〜7 に示す。タイムスタンプを考慮する場合と似たようなトピックが検出されるケースもみられたが、まったく違うクラスタも多くみられた。

タイムスタンプを考慮しないクラスタリング結果の全体的な傾向として、時間軸とは関係の薄いクラスタ (「カテゴリ」に近いもの) が観察された。表 5 中の 3 位、5 位は、メジャーリーグに挑戦するプロ野球選手が増えていることに関するトピックであり、様々な選手のメジャー挑戦についての記事の集合となっている。これらは、個々のイベントとしては時期が違うので、タイムスタンプを考慮した場合は別々のクラスタとして検出される。表 6 中の 2 位のトピックは、表 4 中の「楽天 vs ライブドア」とは記事集合が少し違い、10 月後半から 11 月初旬の球界参入の話題に加え、12 月中旬に行われた「“ライブドアフェニックス” ファン感謝イベント」についての話題を包含していた。こちらも、タイムスタンプを考慮した場合は別のトピックとして検出された。タイムスタンプを考慮した場合の効果が最も顕著に現れたのは、「宮里藍」についての場合である。表 7 の 1 位のトピックは、ミズノクラシック、リコーカップ、マスターズ GC レディースなど様々な女子ゴルフツアーについての記事集合となっていた。これらは、提案手法ではそれぞれ別の話題として検出されている。また実験では、タイムスタンプを考慮しない場合は、どのようなトピックを表しているのかがわからなかったクラスタ (“???” とついているもの) の割合が高いことが観察された。

タイムスタンプを考慮する/しないはユーザの目的にもよるが、タイムスタンプを考慮することによってトピック検出にどのような影響があるのかは今後の検討課題としたい。

6. おわりに

blog より特定の人物に対する話題の抽出を行う手法を提案した。今後の課題としては、関連記事の収集方法の検討が挙げられる。本手法では人名と記事間の類似度を手がかりに関連記事を収集したが、トラックバック等のリンク情報を利用する、トピックごとに分類器を作成するなど、より精度・再現率の高い収集方法を検討していきたい。また関連記事の候補集合として、人物の「姓」および「名」のいずれかを含む記事を用いたが、所属や肩書など名前以外の人物の属性、Wikipedia 等の情報源を用いる手法も考えられる。こうした対象人物に関する情報の利用も今後の課題としたい。その他、人物以外 (企業など) を対象とした場合への拡張、本手法を用いてシステムを構築する際に必要となる blog 記事収集方法、タグ解析方法等が挙げられる。

謝辞　本研究の一部は、日本学術振興会科学研究費補助金基盤研究 (B)(#15300027) の助成による。

文　献

[1] Russell Swan and James Allan. Extracting significant time varying features from text, CIKM'99

[2] R.Kumar et al., On the bursty evolution of Blogspace,

表 5　*target* < 松井秀喜 >(通常クラスタリング)

クラスタリング対象記事数		194
検出トピック数		16
ランク	トピック	固有名詞上位 5 件
1.	ア・リーグ優勝決定戦	[ヤンキース 日本 ボストン 大リーグ イチロー]
2.	松井&酒井美紀	[酒井 美紀 ニューヨーク ヤンキース ゴジラ]
3.	メジャーに挑戦する日本プロ野球選手	[井口 ヤンキース イチロー 日本 ダイエー]
4.	広告記事	[新庄 剛志 日本 天田 中村 コスタリカ 大久保 俊輔 宮里 アル]
5.	メジャーに挑戦する日本プロ野球選手	[石井 ヤンキース ジョンソン ドジャース イチロー]
5.	???	[イチロー 央 日本 ヤンキース 田口]
5.	???	[能美 根上 北陸 寺井 星稜]
8.	日米野球	[米 日本 日 ヤンキース 石井]
8.	???	[石川]
8.	???	[ヤンキース 胡麻 寅 ビートたけし 松坂]

表 6　*target* < 堀江貴文 >(通常クラスタリング)

クラスタリング対象記事数		226
検出トピック数		23
ランク	トピック	固有名詞上位 5 件
1.	著作の感想など (広告記事含む)	[近鉄 日本 雄 オン 大阪]
2.	楽天 vs ライブドア	[フェニックス 仙台 宮城 東北 パ・リーグ]
3.	流行語大賞	[波田 浜口 北島 陽 カチュー]
4.	競馬界参入	[高崎 群馬 小寺 高知 弘之]
5.	「カネで買えないものなんてない」発言	[モン リエ 高崎 朝日新聞]
6.	著作の感想など	[ ]
6.	西武球団が売却打診	[西武 コクド 香港 仙台 中国]
8.	民間有人飛行	[日本 モン ロシア リエ ブッシュ]
8.	紅白裏番組出演	[細木 数子 和田 日テレ 日本テレビ]
8.	ソフトバンクホークス誕生	[ソフトバンク ホークス 福岡 鳥栖 正義]

表 7　*target* < 宮里藍 >(通常クラスタリング)

クラスタリング対象記事数		236
検出トピック数		24
ランク	トピック	固有名詞上位 5 件
1.	女子ゴルフ	[裕理 ミズノ 古閑 タイ リコー]
2.	広告記事	[日本 剛志 新庄 松井 清原]
3.	宮里藍と横峯さくら	[米 日本 ハワイ アメリカ 日]
4.	宮里聖志 (兄) が優勝	[聖 沖縄 優作 那覇 宜野湾]
4.	メガネベストドレッサー賞	[江梨子 佐藤 久米 日本 宏]
4.	???	[沖縄]
7.	???	[ゆかり 馬場]
7.	???	[裕理 リコー 丸山 アメリカ]
9.	賞金女王	[辻本 裕理 松坂 松坂 阪神]
9.	???	[谷 亮子 日本]

WWW2003

[3] D.Gruhl et al., Information Diffusion Through Blogspace, WWW2004

[4] E.Aar et al., Implicit Structure and the Dynamics of Blogspace, Workshop on the Weblogging Ecosystems, WWW2004

[5] Gabriel Pui Cheong Fung et al., Parameter Free Bursty Events Detection in Text Streams, VLDB2005

[6] Yiming Yang et al., Learning Approaches for Detecting and Tracking News Events, IEEE Inteligent Systems 1999(Vol.14, No.4) pp.32-43

[7] Jon Kleinberg. Bursty and hierarchical structure in streams, SIGKDD2002

[8] 中島伸介, 竹原幹人, 舘村純一, 日野洋一郎, 原良憲, 田中克己, blog 解析に基づく Web 情報検索の信頼性向上技術, 人工知能学会研究会資料 SIG-SWO-A401-05

[9] 藤木稔明, 南野朋之, 鈴木泰裕, 奥村学, document stream における burst の発見, 言語処理学会 第 11 回年次大会 2005