ブログ発信者の特徴を利用した話題抽出手法

Topic Detection from Blog Documents Using Bloggers' Interest

関口 裕一郎♠　川島 晴美♦
奥田 英範♦　奥 雅博♦

Yuichiro SEKIGUCHI　Harumi KAWASHIMA
Hidenori OKUDA　Masahiro OKU

本論文では，ブログ文書群から特定の興味分野における話題語句を抽出する手法について論ずる．興味を同じくする人々の間で頻繁に用いられる語句をその興味分野での「話題語」と定義し，各ブログサイトからブログ発信者の興味と発信者間の興味の関連度合いを算出し，関連する興味を持った発信者の間で特徴的に出現する語句を話題語として抽出する．ブログ文書を用いて話題語句抽出の精度を評価した結果，一般的な TF-IDF による話題語句抽出に比べ 4.4%の精度の向上が確認された．

In this paper, we describe a method to detect topic words from blog documents. We define "topic words" as words frequently used by people who share the same interests. In this method, each blogger's interests are extracted from each blog site, and interest similarities between bloggers are calculated. Unusual words that are used by bloggers who have a high level of similarity are then extracted as topic words. We evaluated the precision of this method using blog documents, and the results show that the proposed method is superior (by 4.4 %) to the traditional TF-IDF method in terms of precision.

1. はじめに

近年ブログサイトをはじめとする個人が情報を発信するウェブサイトが多く開設され，様々な意見や体験の情報をウェブ上で閲覧できるようになってきた．そのような状況の中で，ニュースに対する人々の意見や，新製品の評判などを知るためにブログ記事を閲覧するという利用形態が現れてきた．このような閲覧を支援する為に，多数のブログサイトに書かれる内容を解析して，最近多くの人々に注目されている話題を表示するサービスの試みが行われている[1][2]．

このようなブログ文書中から注目されている事柄を抽出する手法として，ある期間において文書群中での使用回数が優位に増加している語句を，その期間に注目されている話題を表す語句として抽出するものがある[3][4]．このような語句の使用文書数の時間変動に注目した話題抽出手法は，多くの人々が注目している事柄を発見するに有効である．

その一方でブログサイトは個人の視点で書かれる為，マスメディアに扱われないような様々な趣味分野における話題も書き込まれ，そのような情報を得るためにブログを閲覧するという利用法が存在する．このような閲覧を支援するには，発信している人数が少ない分野の話題語句を抽出せねばならないため，従来の文書数変動による話題抽出手法では対応が困難であった．本論文では，各ブログ発信者興味分野を抽出し，似た興味を持つ人々の間でのみ使用される語句を抽出することにより，様々な分野に対して話題となっている語句を抽出する手法を論じる．

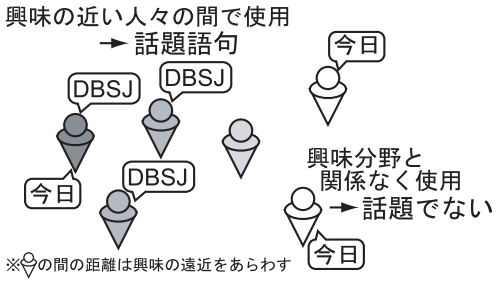

図 1　話題語のイメージ図
Fig.1 Concept image of topic word

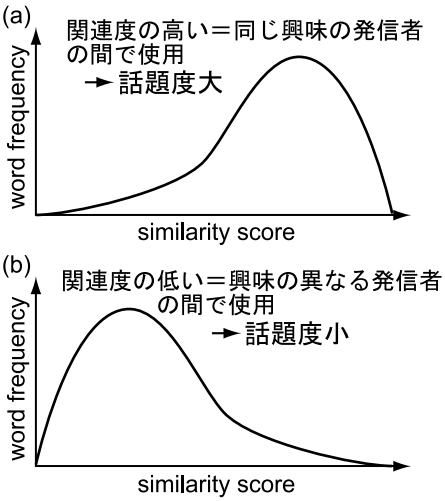

図 2　話題度算出のイメージ
Fig.2 Definition of high & low topic scores

2. 抽出対象とする「話題」

本論文においては，興味を同じくする人々の間で頻繁に用いられる語句を，その興味分野での「話題語」と定義する．定義された話題語の概念図を図 1 に示す．図 1 においては，興味分野が似ている発信者はより近くに配置されているとする．例えば，「DBSJ」という言葉はデータベース技術に興味のある発信者の間でのみ特徴的に用いられる為，これは先に定義した「話題語」に当てはまる．一方で「今日」という言葉は，「DBSJ」と同頻度で使われているが，その使用者に共通性がなく，誰しもが用いるような一般的な語句であり，少なくともデータベース技術に興味のある発信者にとっての話題語ではないとみなす．

以上の定義に基づいた話題語句の抽出を行うため，各語句

♠ 正会員　日本電信電話株式会社　ＮＴＴサイバーソリューション研究所　sekiguchi.yuichiro@lab.ntt.co.jp
♦ 日本電信電話株式会社　ＮＴＴサイバーソリューション研究所　{kawashima.harumi, okuda.hidenori, oku.masahiro}@lab.ntt.co.jp

について，その語句をブログサイトで使用した発信者達の興味の関連度合いを集計し，話題語句か否かの判別を行う．ある語句がある発信者の興味分野においての話題語句かを判別するために，その語句をブログで使用している人々が，対象となる発信者とどの程度興味が関連しているかを，関連度として数値化して集計する．ある語句が処理対象発信者の興味分野における話題語句であれば，その語句を使用している人は対象となる発信者と似た興味を持ち高い関連度持つと考えられる．従って，各関連度の値におけるその語句の使用人数の分布を集計すると，図 2(a)のようなグラフの分布になる，一方ある語句が処理対象発信者の興味分野と関連が無い場合においては，図 2(b)のように関連度の低い人々に多く使われているような分布になる．以上のように，関連度と語句使用者数の分布を見て，高関連度の人々で使われる語句を取得することにより，処理対象となる発信者の興味分野における話題を取得可能となる．

3. 提案手法

提案手法は 5 段階の処理で構成される．(1)各発信者の興味分野を語句ベクトルの形式で抽出する．(2)抽出された語句ベクトル間の類似度を求めることにより，発信者間の興味の関連度合いを求める．(3) 各語句 w_k について，w_k の各関連度の範囲における語句使用数の分布を語句関連度分布 $WD_{i,wk}$ として求める． (4) 発信者 i とその他の全て発信者との間の関連度の値を集計し，関連度に対する人数の分布を基準関連度分布 BD_i として求める． (5)最後に，図 2(a)に示されるように関連度に対する語句の使用者の分布が高い範囲に分布している語句に高い話題度を算出する．

3.1 ブログ発信者の興味ベクトル抽出

各発信者の興味を表す語句に対して高い重みを設定した語群ベクトルを，発信者の興味ベクトルとして抽出する．

ブログ発信者は興味を持つ分野における特徴的な語句を複数のブログ記事に渡って使用すると仮定し，興味ベクトルを抽出する際に，ブログサイト中の複数の文書に渡って用いられる語句に発信者の興味を表す語句として高い重みをつけることとする．また，多くの発信者が使用する語句は一般的な語句として，その重みを下げることとする．

発信者 i についての興味ベクトル Vi の値は，次の式によって求まる．

$$V_i = \begin{pmatrix} x_{i1} & x_{i2} & x_{i3} & \cdots \end{pmatrix}$$

$$x_{ik} = ef_i(w_k) \times \log\left(\frac{N_u}{uf(w_k)}\right)$$

ここで， x_{ik} は発信者 i の語句 w_k への興味の度合いを表す重みで，発信者 i が過去に発信した語句 w_k を含むブログ記事の数 $ef_i(w_k)$ と，一般的な語句の重みを下げる要素である語句 w_k の使用したユーザ数 $uf(w_k)$ の逆数に全ユーザ数 N_u をかけてログをとった値と，をかけ合わせることによって求まる．

3.2 ブログ発信者間の関連度算出

一組の発信者間の興味がどれだけ類似しているかを表す関連度 R_{ij} を，全ての発信者の組み合わせについて求める．

もし発信者 i と j の興味が似ていれば，その興味ベクトル Vi と V_j も類似すると考えられる．その考えにもとづき，次式に表されるような興味ベクトルのコサイン類似度の値で関連度 R_{ij} を定義する．

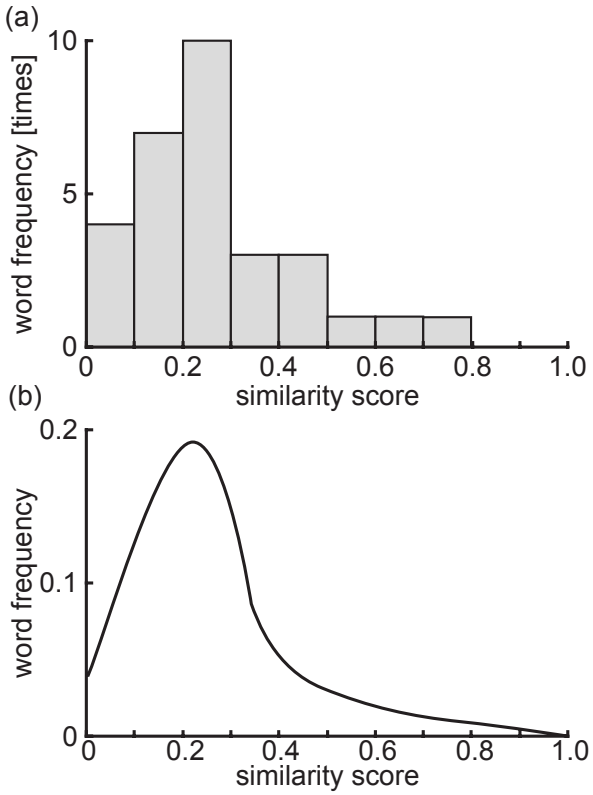

図 3 「ACL」「サッカー」「部分」の 3 語における語句関連度分布の例

Fig.3 Example of word similarity distribution for three words

$$R_{ij} = \frac{V_i \times V_j}{\|V_i\| V_j|}$$

3.1 で述べたように，興味ベクトルの各要素は全て正の値をとる為，R_{ij} の取る範囲は 0 から 1 となる．

3.3 語句関連度分布

発信者 i に対して，語句 wk を使用している人々がどの程度の関連度を持つ人々であるかを表す，関連度に対する語句 w_k 使用頻度の分布を集計した語句関連度分布 $WD_{i,wk}$ を求める．具体的には，図 3(a)のように関連度の取り得る範囲 0～1 を N 分割した集計区間毎に，その区間に対応する関連度を持つ人々が何回語句 w_k を使用しているかを集計することにより求める．この際，語句ごとに全体での使用頻度は異なるため，各集計区間における使用頻度を全体の使用数で割ることにより，図 3(b)のように正規化を行う．

語句関連度分布 $WD_{i,wk}$ を求める式は下のようになる．

$$WD'_{i,wk}(n) = \begin{cases} ef_j(w_k) & if \quad \dfrac{n-1}{N} \le R_{ij} < \dfrac{n}{N} \\ 0 & else \end{cases}$$

$$WD_{i,wk}(n) = \frac{WD'_{i,wk}(n)}{\sum_n WD'_{i,wk}(n)}$$

語句 w_k が発信者 i と似た興味の持つ人々の間で共有される語句である場合には，$WD_{i,wk}$ は関連度の高い部分にピークが

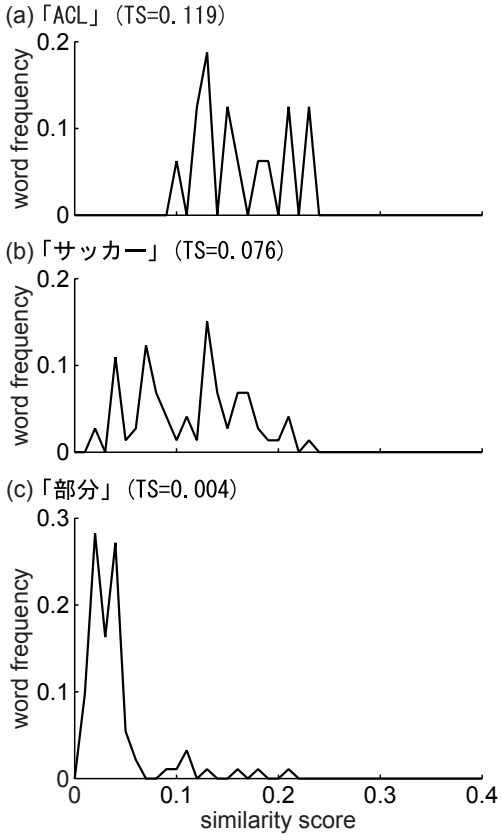

図 4 「ACL」「サッカー」「部分」の 3 語における語句関連度分布の例

Fig.4 Example of word similarity distribution for three words

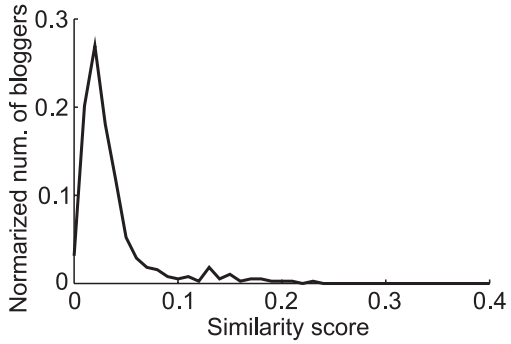

図 5 基準関連度分布の例

Fig.5 Example of base similarity distribution

くるような分布となる．一方，w_kの使用者に特徴がなく，あらゆる人に使われる語句であれば，$WD_{i,wk}$の分布は関連度の低い部分にピークがくるような分布となる．

図 4 に，サッカーファンに対する「ACL（アジアチャンピオンシップリーグ)」,「サッカー」,「部分」の 3 語の語句関連度分布を算出した例を示す．サッカーファンにとって興味分野の話題語である「ACL」は，関連度が比較的高い部分に分布のピークがきていることが分かる．興味分野の言葉ではあるが，より一般的な内容である「サッカー」は「ACL」と比べると関連度の低い部分に分布のピークがあり，興味分野の話題語ではない「部分」は関連度の低い部分にピークが現れるようになっている．

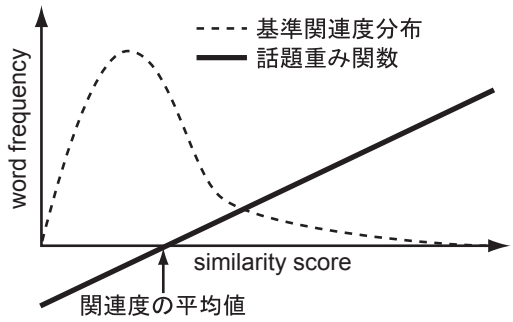

図 6 話題重み関数

Fig.6 Topic weight function

3.4 基準関連度分布の取得

語句関連度分布の形状は，処理対象発信者と似た興味を持つ発信者がどの程度いるのかに強く依存する．その為，語句の話題度合いを算出する基準として，全ての発信者が用いる語句における関連度分布を，基準関連度分布 BDi(n)として求める．語句関連度分布 $WD_{i,wk}$と同様に，基準関連度分布 BD_iも発信者数で正規化を行うこととする．

基準関連度分布を求める式は次のようになる．

$$BD'_i(n) = \begin{cases} e) & if \quad \dfrac{n-1}{N} \le R_{ij} < \dfrac{n}{N} \\ 0 & else \end{cases}$$

$$BD_i(n) = \frac{BD'_i(n)}{\sum_n BD'_i(n)}$$

図 5 に発信者が 500 人いた場合の，あるサッカーファンに対する基準関連度分布を求めた例を挙げる． 関連度の値は興味ベクトルのコサイン類似度であるため，その値は 2 人の発信者間で興味語句がどれだけ一致したかを表す．その為関連度の分布を集計したグラフは，ポワソン分布に似た傾向を持つ．

3.5 話題度算出

図 2 に示されているように，本手法はある発信者 i と関連の高い発信者の間で語句w_kが使用されている際に発信者 i にとっての語句 w_k の話題度を高く算出する．具体的には語句関連度分布 $WD_{i,wk}$に，図 6 に示すような関連度の平均値 n_0 での重みをゼロとなる線形の話題重み関数を掛け合わせた結果を足し合わせた値を，発信者 i にとっての語句w_kの話題度 $TS_i(w_k)$をとする．

その際に，語句関連度分布 $WD_{i,wk}$がポワソン分布的傾向を持つため，その影響をする必要がある．その為に，語句関連度分布 $WD_{i,wk}$から基準関連度分布 BD_iを引いた値に,話題重み関数を掛け合わせることとする．その算出式は，以下のようになる．

$$TS_i(w_k) = \sum_n \left\{ \left(WD_{i,w_k}(n) - BD_i(n) \right) \cdot \frac{n - n_0}{100} \right\}$$

$$n_0 = \frac{\sum_n \{ BD_i(n) \cdot n \}}{\sum_n BD_i(n)}$$

表 1　あるブログ記事における話題度上位 5 語句
Table 1　Example of to 5 topic score words in an entry

語句	話題度	TF-IDF	TF
マリノス	0.125	6.36	1
ACL	0.119	14.00	2
過密日程	0.095	15.94	2
アジア	0.091	21.56	3
Ｊリーグ	0.091	6.00	1
平均	0.104	－	－

4. 評価実験

提案手法によって，発信者の興味分野における話題語句が抽出されるかの評価実験を行った．ブログ記事は発信者の興味分野について書かれると仮定し，提案手法を用いて各記事からその発信者に取っての話題度語句を抽出し，その精度を測定した．具体的には各ブログ記事の話題を表す語句を５つユーザに提示することにより，閲覧記事の選択を容易にするような支援サービスを想定し，提案手法により抽出される高話題語句５つと，同じ記事に対して人手で選択した話題語句との一致率により精度を測定した．また比較対照として，TF-IDF の上位 5 語句についても同様の評価を行った．

実験対象データとして，500 ブログサイトにおいて 2005 年 4 月 19 日～5 月 18 日に発信された 11513 記事を用意した．このうち 5 月 12 日～18 日の 2530 記事について，3 名の作業者によって正解となる話題語句を抽出した．この際，特定の相手に対して書かれた私信や単純な行動記録といったような，明確な話題を含まない記事については話題語句の抽出を行わず，「話題なし」とタグ付けした．その結果，3 名の作業者全てが話題を含むと判断した記事について，2 名以上が話題語句と判断した語句を正解データとして，計 745 記事に対する正解データを作成した．

また，提案手法は各語句の周囲での使用状況を用いて話題度を判別するため，データセット中に同じ話題について扱った記事が 1 つしかない場合には正確な処理が行われない．その為，上位 5 語句の話題度の平均が一定以下の記事については，データセットの範囲が限られているため他に類似する記事がないか，明確な話題を含まない場合と判断し，精度評価から除外した．その判別の閾値は，人手によって話題有りと判定された記事が 9 割含まれる点に設定し，上位 5 語句に与えられた話題度の平均が 0.06 以上の記事とした．

上位 5 語句の話題度平均が 0.06 以上かつ人手で話題度有りと判定された 492 記事を用いて，話題抽出精度の評価を行った．提案手法における話題語の抽出精度は 46.2%，TF-IDF おける精度は 41.8%となり，他に同一の話題を扱っている記事が存在する場合には TF-IDF よりも適切な重み付けができていることを確認できた．

また抽出結果から，提案手法と TF-IDF とにおいて取得される話題語の傾向の違いが見られた．ブログによく見られる短い記事においては提案手法が効果的であった．このような記事においては，各語句の TF の値がほとんど 1 に近くなるため，周囲での語句使用状況を考慮することが有効に働いたためと考えられる．一方で記事が長い場合には TF-IDF が有効になる傾向があった．

短文において提案手法が効果的に働いた例として，表 1 にサッカーファンが ACL(アジアチャンピオンシップリーグ)について書いたブログ記事から話題度上位 5 語句を示す．出場チーム名やリーグ名といった語句に高い話題度が算出されていることが分かる．同じ記事に対して TF-IDF が高くなる語句は，日本（話題度 0.026，TF-IDF22.7），アジア（話題度 0.091，TF-IDF21.6），代表戦（話題度 0，TF-IDF18.5）となり，TF が高くなった語句が上位に来る．

5. まとめと今後の課題

発信者同士の興味の関連度を勘案した話題算出手法を行うことにより，話題を持つ記事に対してのみ話題度を算出できるようになった．また記事単位で TF-IDF よりも適切な話題語句抽出が可能になることを確認した．

今回の実験においてはすでに定まったデータに対して一括での処理を行ったが，今後は日々投稿されるブログデータに対して連続的に処理していくことが可能になるよう，関連度の算出方式を中心に処理量の軽減化の検討を行っていく．

また従来検討されてきた，語句の使用頻度の時間変動による要素を加えることにより，「サッカー」のような興味分野自体を表す語句と，「ACL」のようなその分野で起きているイベントを表す語句を，より明確に区別することが可能になると考えられる．そのような精度向上手法についても今後さらに検討を行っていく．

[文献]

[1] BlogPulse, http://www.blogpulse.com/
[2] kizashi.jp, http://kizasi.jp/
[3] Glance, N., Hurst, M., Tomokiyo, T.: “BlogPulse: Automated Trend Discovery for Weblogs,” Presented at the Workshop on the Weblogging Ecosystem at the 13th International World Wide Web Conference, (2004).
[4] 佐藤吉秀，川島晴美，佐々木努，大久保雅且，“文書の類似度と新鮮度に基づく話題語抽出，”情報処理学会自然言語処理研究会発表資料，2005-NL-165, pp. 29-35 2005.

関口 裕一郎　Yuichiro SEKIGUCHI
NTT サイバーソリューション研究所所属．2004 年東京大学大学院情報理工学系研究科修士課程修了．同年日本電信電話（株）入社．現在インターネットにおける情報抽出の研究開発に従事．日本ヒューマンインタフェース学会，日本データベース学会各会員．

川島 晴美　Harumi KAWASHIMA
NTT サイバーソリューション研究所主任研究員．1990 年山梨大学大学院工学研究科修士課程修了．同年日本電信電話（株）入社．現在インターネットからの話題情報抽出技術の研究開発に従事．電子情報通信学会会員．

奥田 英範　Hidenori OKUDA
NTT サイバーソリューション研究所主幹研究員．1988 年東京大学大学院工学系研究科修士課程修了．同年日本電信電話(株)入社．1994 年スタンフォード大学コンピュータ科学科修士課程修了．現在 CGM における情報抽出の研究開発に従事．電子情報通信学会，映像情報メディア学会各会員．

奥 雅博　Masahiro OKU
NTT サイバーソリューション研究所主幹研究員．博士(工学)．1984 年大阪府立大学大学院工学研究科博士前期課程修了．現在は検索をはじめとするインターネットサービスの研究開発に従事．現在電子情報通信学会，情報処理学会，言語処理学会各会員．