2005－NL－170（5）
2005／11／21

blog ページ集合に対する話題語句抽出手法

関口 裕一郎† 佐藤 吉秀† 川島 晴美† 奥田 英範† 奥 雅博†

似た興味を持つ人々の間で注目されている事柄を話題と定義し，blog 文書中から話題語句を抽出することを目的とする．blog 記事は省略を多く含む口語的な記述がなされている為，文中の情報のみによる話題語句の判別は難しい．本論文では，発信者相互の興味の関連性を抽出し，ある語句を使用している発信者集合の持つ関連度の分布を見ることにより，高い関連度を持つ発信者間で使われる語句に高い話題度を算出する．blog 記事の集合を用いて実験を行った結果，記事中の話題を表す語句に対して，高い話題度を算出することができた．

Topic Words Detection for Blog Documents.

Yuichiro SEKIGUCHI†, Yoshihide SATO†, Harumi KAWASHIMA†, Hidenori OKUDA†, Masahiro OKU†

In this paper, we describe the method to detect the topic words from blog documents. The ‘topic words’ is defined as a word that gains the attention of people sharing same interest. While blog documents are written by ordinal people, their texts are written in abbreviated informal expression. We use the information of blogger to adjust this characteristic of blog documents. The proposed method extracts the relevancies of each blogger; compares the deviation of these relevancies; and calculates the topic scores for each word of a blog document. The experiment shown that the method can extract appropriate topic words from blog documents.

1. はじめに

インターネットへの常時接続環境の普及に従い，個人による Web 上での情報発信が一般化してきている．特に blog は技術的な知識を必要とせず，容易に情報発信が行える為，ここ数年で急速に普及を進めてきた．このような個人による情報発信が増大する傾向は今後数年に渡り続くと予測されている[1].

このような個人による情報発信は，個々人の意見や主観がその内容に含まれることが大きな特徴である．特に blog 記事はその傾向が強く，ニュースに対する世間の反応や，新製品の評判等を得るために blog サイトを巡回するという閲覧形式が定着してきている．しかし blog 記事は自由な立場で書かれているために，その評判内容は様々に異なる．従って評判の全体としての傾向を知る為には，同じ事柄について書かれた複数の blog 記事を探し出して目を通す必要があった．

本論文は，上記のような同一内容を扱う複数の blog 記事を巡回する閲覧スタイルの支援を目指し，blog 記事中から扱われている話題を表す語句を判別することを目的とする．

テキストマイニングの分野において，大量文書から同一内容を扱った文書群を抽出する手法は種々提案されている[2]. しかしこれらの手法を blog 記事に適応する際には， blog 記事の口語的な表現形態が問題となる．多くの blog 記事は短く主観的に書かれていることが多いため，一般的な TF-IDF による文書中の重要語句抽出手法で用いる，単語の出現頻度と重要性の相関が必ずしも成り立たないという問題点がある．また各

† 日本電信電話株式会社 サイバーソリューション研究所
NTT Cyber Solutions Laboratories, NTT Corporation

発信者の文章の書き方の癖や，顔文字等の記号使用も処理を難しくする要因となる．

本論文では，複数の blog 記事で扱われるような話題は，「似た興味を持つ人々の間で注目されている」と仮定する．その仮定に従い，各発信者間の興味の関係性を用いて，興味を同じくする発信者の間で特徴的に出現する語句を重要語句とみなして高い話題度を算出する手法を提案する．

以下，第 2 章で blog 文書中から話題語句を抽出する際の問題点，及び従来の研究について述べる．第 3 章では本研究で取り扱う話題語句の定義を行い，第 4 章で第 2 章の問題点に考慮した文書中話題語句抽出手法を説明する．第 5 章で blog 記事を対象に行った実験と結果を示し，第 6 章で考察を行う．最後に第 7 章でまとめを述べる．

2. 背景

2.1. blog 記事の特徴

本研究の処理対象となる blog 記事は，一般の人々が専用のシステムを使用して作成・公開した記事の集合である．その為，文書執筆の訓練を受けた書き手によって作成されているニュース記事とは異なる特徴を持つ．以下に，blog 記事に対して機械的処理を行う際に注意すべき特徴を述べる．

・了解事項や一度書いた事柄は省略される

例えばニュース記事であれば，首相の動静について書かれた記事では，「首相」という主語が繰り返し用いられる．しかし blog 記事においては，首相についての記事で首相のことが書かれるのは当たり前であるため，代名詞の使用や主語の省略がなされ，同じ表記を複数回使用する事は少ない．

その為，類似文書判定で一般的に用いられる文書内語句頻度(Term Frequency)を用いた重み付けが有効に作用しないことが多い．

・発信者ごとに文体の癖が存在する

発信者が好んで用いる略称や顔文字などの記号，または特徴的な一人称などが存在し，それらが文中で繰り返し用いられる．

その為，文書内容と直接的な関連を持たない語句においても文書内語句頻度が高くなる傾向がある．

・blog システムにより付加情報が与えられる

blog システムの大半は ATOM や RSS といったフォーマットやトラックバックリンクに対応しているため，それによる付加情報が存在する．これらの中には例えば作成時刻情報や発信者情報，トラックバックやコメント等の有用な情報も多く含まれる．

以上のような blog 記事の特徴から， TF-IDF よる重要語句抽出は blog 記事においてその精度が低下する．その為 blog を対象とした話題抽出においては，時刻情報やトラックバック情報を利用した話題抽出手法が多く検討されている．

2.2. 関連研究

blog 記事から話題を取得する方法としては，リンク構造を用いた手法と，語句の出現頻度を用いた手法の二つが主に用いられている．

リンクを用いた手法としては，トラックバックリンク等による blog 記事間の直接的なリンク構造を取得する方法[3]や，同一のリンク先を持つ blog 記事を纏める方法など[4][5]が存在する．これらは発信者が意図的に設置したリンク構造を利用する為，高い精度で同一話題記事の集合が得られる利点がある．一方，リンクを持つ blog 記事の割合は全体の 2 割程度であり（例えば第 5 章で用いる blog 記事群では 18.5%がリンクを持つ），リンクを張るようなある程度技術に明るい発信者の記事のみが処理対象となってしまう欠点がある．

一方，語句の出現頻度がある期間において特徴的に変化した場合に話題語句として抽出する手法がある[6][7]．これは blog 記事集合全体を処理対象と出来る為，その日のテレビ番組に対する感想記事の集合，といったようなリンクをほとんど持たないような記事集合を抽出できる利点がある．しかし，全体の中で目立って扱われている話題しか抽出できないため，幅広い話題を抽出する為には，処理対象文書をその内容の分野ごとに分割する必要がある．

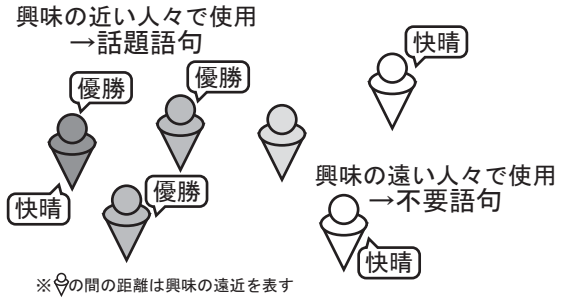

図１．分野特徴語句のイメージ

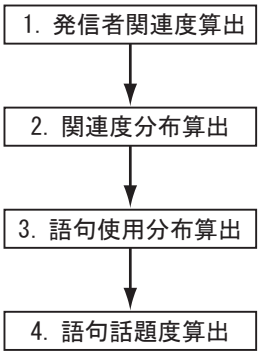

図２．話題度算出処理の概要

本論文における提案手法は，出現頻度による話題語句抽出の一応用となる．発信者間の興味の関連性から得られる語句の出現頻度の分野的な偏りを利用することにより，幅広い話題を抽出することを目指す．

3. blog 記事における「話題」

本来「話題」とは，ある文書で扱われている主題のことを指し示すが，本論文では同一内容を扱う blog 記事の発見支援という観点から，「同じ分野に興味を持つ人々の間で，注目・共有されている事柄」を「話題」と定義する．

定義された話題の概念図を図１に示す．例えば，「優勝」という言葉が野球ファンの間で用いられている場合，これは話題の定義に当てはまる．反対に「快晴」という言葉は，「優勝」と同程度使われているが，その使用者に共通性がない為話題とはみなさない．また「今日」という言葉のように，あらゆる人が一様に使用するような一般的な語句も，やはり話題としての特徴を持たない．

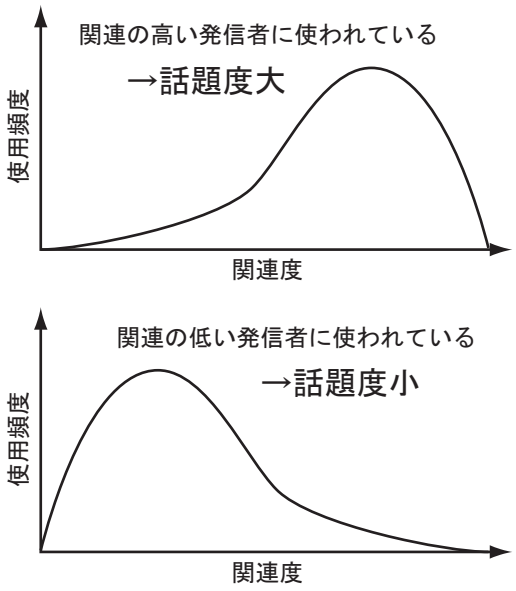

図３．話題度算出のイメージ

4. 提案手法

第 3 章における話題の定義から，各語句の使用者の特徴を見ることにより，blog 文書中の話題語句に高い重みを置く話題度算出手法を提案する．図 2 に提案手法の処理の流れを，図 3 に話題度算出のイメージを示す．

本手法は，ある発信者 i による blog 文書 d_i 中の語句に対し話題度を算出する．その際に，発信者 i とその語句を使用している他の発信者との関連度の分布を求め，その語句の使用者が図３の上の図のように関連度の高い範囲に偏っていた場合に高い話題度を算出する．また，話題度の算出対象となる語句は，文中に含まれる名詞と名詞の連続による複合名詞とに限定することにする．

具体的な処理フローとしては，まず各発信者の興味分野をベクトル値として抽出し，興味ベクトルの類似度を求めることにより，発信者 i とその他の発信者との興味の関連度を算出し（１），得られた関連度の分布を求める（２）．次に処理対象文書 d_i 中の各語句 w_k について，w_k を使用した事のある発信者のみに絞った場合の関連度の分布を求める（３）．最後に，語句 w_k に対応した関連度の分布と，関連度全体の分布とを比較

して，図３に示されるように関連度の高い範囲に分布している語句に高い話題度を算出する（４）.以下,（１）～（４）の個々の処理について詳細に述べる.

4.1. 発信者関連度算出

興味の似た発信者を特定する為に，各発信者の興味情報を語群ベクトルとして取り出し，得られた語群ベクトルの類似度を用いて発信者間の関連性を数値化する.

発信者は興味を持つ分野の特徴語句を複数の blog 記事に渡って使用していると仮定し，過去に各発信者が発信してきた文書から，語句の使用傾向を語群ベクトルとして抽出する．発信者 i について得られた語群ベクトルを，発信者 i の興味ベクトル Vi と呼ぶこととする.

Vi の値は，次の式によって求まる.

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

$$x_{ik} = ef_i(w_k) \cdot idf(w_k) \quad (k = 1,2,\cdots K)$$

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値で, 発信者 i が過去に発信した語句 w_k を含む blog 文書の数 $ef_i(w_k)$ と，語句 w_k の IDF 値の積で求まる.

一組の発信者間の関連度 R_{ij} を，興味ベクトルのコサイン類似度の値で定義する． 関連度 R_{ij} は以下の式で求まり，その値の範囲は 0 から 1 となる.

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

4.2. 関連度分布算出

話題度算出処理を行う処理対象文書の発信者 i と他の発信者との関連度集合の値の分布 RDi を，0〜1 を N 分割した各範囲の値を持つ $R_{ij}\,(i \neq j)$の数を集計することにより求める.

$$RD_i(n) = f_i(n) \quad n = 1,2,\cdots N$$

$$f_i(n) = \sum_{j \neq i} \begin{cases} 1 & if \quad \dfrac{n-1}{N} \le R_{ij} < \dfrac{n}{N} \\ 0 & else \end{cases}$$

本手法においては，語群ベクトルのコサイン類似度を元に関連度を算出している．その為，関連度の値は一組の発信者間での語句の共起数との相関を持つ為, その分布はポアソン分布に似た傾向を持つ.

4.3. 語句使用者分布の算出

処理対象文書 d_i に含まれるそれぞれの語句 w_k に対して, 発信者 i と語句 w_k を使用している他の発信者との関連度の値の分布 WDi を次の式で求める.

$$WD_i(n, w_k) = g_i(n, w_k) \quad n = 1,2,\cdots N$$

$$g_i(n, w_k) = \sum_{j \neq i} \begin{cases} ef_j(w_k) & if \quad \dfrac{n-1}{N} \le R_{ij} < \dfrac{n}{N} \\ 0 & else \end{cases}$$

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，WDi は関連度の高い方に分布が偏る．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，WDi の分布は 4.2 で求めた RDi の分布と近くなる.

4.4. 語句話題度算出

図３に示されているように, 本手法は語句 w_k の使用者が関連度の高い範囲に分布している際に語句 w_k に高い話題度を算出する手法である．しかし，関連度全体の分布がポワソン分布的な偏りを持つため，それを考慮した分布の評価手法が必要である.

その為，全関連度の分布である RDi を基準値として用い, それに対して語句 w_k についての分布 WDi がどれだけ関連度の高い範囲に偏っているかを数値化した値 $TS(w_k)$を語句 w_k の話題度とする.

話題度 $TS(w_k)$の値は，次の式により与えられる.

$$TS(w_k) = \sum_n \left\{ \left(\frac{WD_i(n)}{\sum_n WD_i(n)} - \frac{RD_i(n)}{\sum_n RD_i(n)} \right) \times \frac{n - n_0}{100} \right\}$$

ここで n_0 は分布 RDi の重心位置を示す値であり，次の式により求まる.

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

以上の式を用いた結果，RDi に比べ WDi が高い範囲に偏っていたときに話題度が正の値に，WDi が低

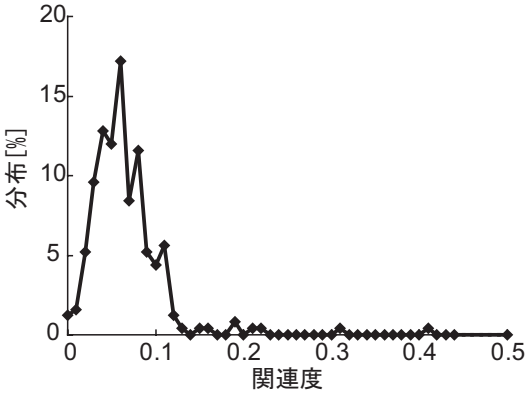

図 4. 関連度全体の分布

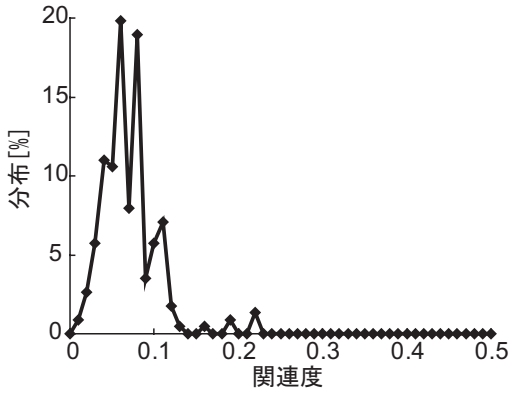

図５．「時間」を含む発信者のみでの関連度分布

い範囲に偏っていたときに話題度が負の値になる．

5. 実験

250 の blog サイトの一ヶ月分の blog 記事集合 7145 件に対して，提案手法を用いて処理を行なった．同一話題の記事集合が存在するように，処理対象はアニメ･マンガについて主に扱っているサイトに限定した．

「仮面ライダー響鬼」について書かれた blog 文書に対して処理を行なった際の，得られた関連度全体の分布を図４に示す．語句の共起をベースとして算出しているため，ポワソン分布的な傾向を持った分布となっていることが分かる．また，発信者の使用語句特徴の興味ベクトル作成においては，一人称や表記の癖といった話題と関連しない語句が入らないように，blog 上で話題となりうる語句の集合と考えられるはてなキーワード[8]に含まれる語句のみを用いて処理を行った．

「響鬼」「時間」という語句の使用者に絞った際の関連度の分布を図５，図６に示す．一般的な語句と考えられる「時間」は全体での分布とあまり変わらなく，「響鬼」は関連度の高い発信者に多く使われる傾向がある．処理の結果,「時間」には 0.0037,「響鬼」には 0.121 という話題度が算出された．

「仮面ライダー響鬼」第 3 話について書かれた２つの blog 記事についての処理結果から，話題度の上位 6 つの語句についての話題度・TF-IDF 値・TF 値を纏めた表を表 1 に示す．

話題の中心となる語句「響鬼」や，その放映回に特

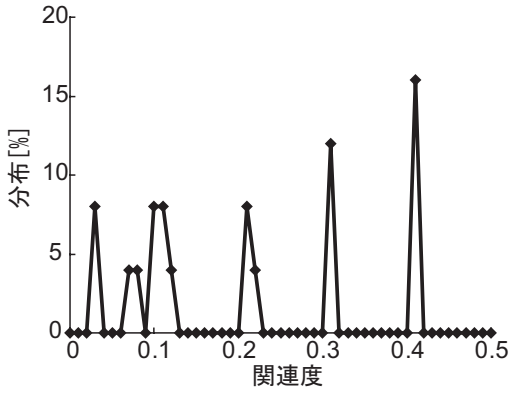

図６．「響鬼」を含む発信者のみでの関連度分布

徴的な語句(「ディスクアニマル」「ヤマビコ」「三之巻」)に高い話題度が算出されている．

日記的な明確な話題が存在しない blog 記事に対して話題度を求めた結果を表２に示す．全体的に表１の結果よりも低い話題度が算出されている．

6. 考察

発信者の情報を用いた話題度算出処理を行うことにより，blog 文書中の重要語句に高い話題度を与えることができた．また日々の活動記録のような明確な話題を持たない記事に対しては, 低い話題度が算出された．

このことから，提案手法は blog 記事中の話題語句をよく抽出しており，blog 記事からその特徴を現す語群ベクトルを作成する際の重み付けにも有効であると考えられる．また一定以上の話題度を持つ語句が存在しない記事を判別することによって，日々の活動記録のような内容の明確な話題を持たない記事を事前に取り

表１．話題度上位６語の話題度・TF-IDF・TF

仮面ライダー響鬼第三話感想記事１：1153 byte

語句	話題度	TF-IDF	TF
ディスクアニマル	0.229	4.730	1
三之巻	0.204	5.135	1
仮面ライダー	0.162	3.813	1
響鬼	0.146	3.749	1
マジレンジャー	0.115	4.123	1
バンダイ	0.062	4.219	1

仮面ライダー響鬼第三話感想記事２：1694 byte

語句	話題度	TF-IDF	TF
ディスクアニマル	0.164	4.729	1
魔化魍	0.163	9.458	2
手癖	0.132	10.856	2
ヤマビコ	0.126	34.968	6
変身	0.125	2.995	1
響鬼	0.121	18.745	5

除く作業への適応も可能になると考えられる．

本手法は同じ興味を持つ発信者の間で共有される語句に重み付けを行うため，特定の分野においては一般的に用いられる専門用語に高い話題度が算出されてしまう可能性がある．今回の実験は１ヵ月分の記事群について処理を行った為，そのような傾向は見られなかったが，長期間にわたって処理を行う場合には，出現頻度の変化を元にした時間的な重み付けモデル[9][10]を導入するような対策が必要になると考えられる．

7. まとめ

本論文は blog 記事の中の話題語句を抽出する手法として,各発信者の興味分野を抽出し,興味を同じくする人々の間で共通して用いられる語句を話題語句として抽出する手法を提案した．他の発信者との関連性を考慮した提案手法を用いることにより，一般的な TF-IDF 手法に基づいた話題語句抽出が精度良く働かない blog 記事に対しても，適切な話題語句の抽出を可能とした．

表２．日記的文書における話題度上位６語

語句	話題度	TF-IDF	TF
原稿	0.045	5.140	2
新刊	0.031	4.888	2
トーン	0.025	3.343	1
縁側	0.020	10.846	2
真正面	0.020	5.423	1
布団	0.018	2.832	1

また提案手法を用いて，blog 記事集合の各記事に対して語句の話題度を算出する実験を行い， TF-IDF 手法に比べ，適切な重み付けができていることを確認した．

今回提案した方法は，各発信者間の関連性を全て算出する為，その計算量コストが膨大になるという問題点を抱えている．今後は，関連度の分布がポアソン分布に近似できる特性を持つことを利用することにより，関連度の計算対象を絞り込んで計算量を抑制する手法の構築に取り組む．

参考文献

[1] 総務省, "ブログ・SNS の現状分析及び将来予測," 2005.

[2] J. Allan, "Topic Detection and Tracking," Kluwer Academic Publishers, USA, 2002.

[3] R. Kumar, J. Novak, P. Raghavan, A. Tomkins, "On the Bursty Evolution of Blogspace," In Proc. of WWW2003, pp. 568-576, 2003.

[4] はてなダイアリー：注目 URL, http://d.hatena.ne.jp/hoturl

[5] K. Ishida, "Extracting Latent Weblog Communities," Presented at the Workshop on the Weblogginng Ecosystem at the WWW2005, 2005.

[6] N. Glance, M. Hurst, T. Tomokiyo, "BlogPulse: Automated Trend Discovery for Weblogs," Presented at the Workshop on the Weblogginng Ecosystem at the WWW2004, 2004.

[7] T. Fukuhara, T. murayama, T. Nishida, "Analying concerns of people using Weblog articles and real world temporal data," Presented at the Workshop on the Weblogginng Ecosystem at the WWW2005, 2005.

[8] はてなキーワード, http://d.hatena.ne.jp/keywordlist

[9] J. Kleinberg, "Bursty and Hierarchical Structure in Streams," In Proc. of KDD 2002, pp. 91-101, 2002.

[10] 佐藤，川島，佐々木，奥，"時系列ニュース記事における最新話題語抽出方法," NL-168-1, 2005.