コミュニティ構造を利用した話題ナビゲーション手法の検討

関口裕一郎† 川島 晴美† 奥田 英範† 奥 雅博†

† 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 〒 239–0847 神奈川県横須賀市光の丘 1–1
E-mail: †{sekiguchi.yuichiro,kawashima.harumi,okuda.hidenori,oku.masahiro}@lab.ntt.co.jp

あらまし ブログの爆発的な普及により人々の興味や評判を Web 上から取得することが可能となった．これらの情報を解析することにより，人々が注目している話題を抽出する試みが現在行われているが，その多くは大きく取り扱われている話題を抽出するものであり，ブログ記事に多く見られるようなある限られた趣味分野での話題といったような，小規模な分野における話題を抽出するには向いていなかった．本論文は発信者間の興味分野の関連性を抽出し，ある興味を共有する人々の間で注目されている語句を話題語句として抽出することにより，従来取得が難しかった小規模な話題も取得可能な話題抽出を行うことを目的とする．また興味分野ごとに話題を抽出することにより，話題に対して興味を持っている人物によって書かれている記事のみを抽出することも目指す．本論文は各発信者が過去に発信した情報の蓄積を利用し各発信者間の興味の関連を抽出し，興味の類似する発信者で注目されている話題語句に高い話題度を算出する手法を提案する．また提案手法を用いて実際のブログ記事から話題を抽出した結果に対し，話題を含む記事に対して話題度が算出されているか，記事中の適切な語句を話題語句として取得しているかを評価し，提案手法の有効性を確認した．

キーワード 話題抽出，ブログ文書，テキストマイニング

Topic Detection from Blog Documents Using User Interests

Yuichiro SEKIGUCHI†, Harumi KAWASHIMA†, Hidenori OKUDA†, and Masahiro OKU†

† NTT Cyber Solutions Laboratories, NTT Corporation Hikarino-oka 1–1, Yokosuka-City, Kanagawa, 239–0847 Japan
E-mail: †{sekiguchi.yuichiro,kawashima.harumi,okuda.hidenori,oku.masahiro}@lab.ntt.co.jp

Abstract 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.

Key words Topic Detection, Blog Documents, Text Mining

1. はじめに

インターネットの普及とともにネットワーク上で入手可能なテキスト情報の量は増え続け，今や最も有用な情報取得源の一つとなっている．特に近年のブログの爆発的な普及は，様々な人々の思いや体験といった生の声までがテキスト情報としてネットワーク上に蓄積される状況を生み出した．この傾向は今後も続くと考えられ，総務省の調査によると 2005 年末に約 335 万人とされる国内ブログ利用者は，2007 年末には約 782 万人に達すると予測されている [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．これに対応して，ブログサイトから得られるテキストデータを処理することによって，現在世の中で話題となっている事柄を表示するサービスの試みがいくつかなされてきている．

だがこれらの試みにおいては，ブログ記事集合中で多く扱われたリンクやキーワードを現在話題の事柄として抽出するために，多くの人々によって注目されている話題が主に抽出される．その為，特定の趣味の分野で話題になっている事柄のような，

個人によって書かれるメディアであるブログ記事特有の話題を取得することができていなかった．

このような事を可能とするためには，ブログ記事集合を分野ごとに分割した上で，細分化された各分野ごとに話題語句を抽出することによって対処が可能であると考えられる．しかし分野の分割基準を人手で設定することは，各分野に対する知識が必要となる難しい作業であり，時間の経過により新たな分野が現れた際に再度分野設定をする必要があるという点で問題がある．また分野の設定がなされたとしてても，例えばスポーツの中のプロ野球の中のあるチームのみに関する話題といったような，設定分野の子分野・孫分野といった話題の取得は難しいという問題もある．

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．これによって分野設定を与えることなしに，細かな分野で話題となっている事柄を抽出することが可能となる．

以下，2 節では関連研究について述べる．3 節では本論文で抽出対象とする「話題」の定義を行う．第 4 節では提案手法についての説明を行う．そして，第 5 節ではブログ記事を用いて行った評価実験について述べ，第 6 節でその考察・今後の検討について述べる．第 7 節では本論文のまとめを記す．

2. 従来技術とその課題

ブログ記事集合から現在の話題を抽出する技術は，大きく分けて語句使用頻度分析による手法，キーワードマッチングによる手法，リンク構造を利用した手法が存在する．

語句使用頻度分析による手法は，話題が存在する場合にはその話題を表す語句の使用頻度が記事全体中で有意に増加する，という仮定に基づいた手法である [3][4][5][6]．記事群の中に出てくる語句の単位時間ごとの出現数を分析し，その値が急上昇した場合に話題として抽出する．これによりニュースに出るようなタイムリーな出来事を精度良く抽出することが可能になる．

この手法は全体の中の頻度の増減を見るために，突発的に起きた大きな話題の取得は得意であるが，小規模な話題が大きな話題に隠れて発見できないという問題点がある．また，話題語句を含めばその話題を扱っている記事と識別するために，単に話題語句が含まれているだけの私信や個人的な日記のような不適切な記事が入り込んでしまい，話題に対してどのような反応があるかを知ろうとした際にユーザが不便を感じることが多い．

キーワードマッチングによる手法は，話題となりえる語句は一定の範囲の決まった語句に限られているという仮定に基づいた手法である [7]．あらかじめ話題候補となるキーワードを人手で登録した辞書を用意しておき，その語句を含む記事は話題を扱っているという単純な仮定に基づき，各キーワードを含む記事の集合を自動で収集する．また各キーワードの使用回数を取得し，その大小をみることによって特に注目されている話題を抽出することも可能である．

これは話題候補の辞書データの作成やメンテナンスにかかる人的なコストが問題となる．また比較的簡単に実装ができる一方で，得られた記事集合に，たまたまその語句を使っているだけで，話題に対する感想や意見等が書かれていない記事が混ざってしまうという問題も存在する．

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 [8] や，同一のリンク先を持つ blog 記事を纏める方法など [9][10] 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．一方外部ページにリンクを持つ blog 記事の割合はそれほど多くはなく，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．

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，似た興味を持つ発信者に特徴的に注目されている語句の判定を行い，幅広い話題を抽出することを目指す．

3. 抽出対象とする「話題」

本来「話題」とは，ある記事中で扱われている主題のことを指し示す．しかし従来の話題抽出技術においては，閲覧者にとって興味を引きうる話題である必要があるため,「ある程度の数の記事・発信者の間で扱われている事柄」を話題と定義していた．例えば「今日の朝食」というようなある一人の人しか扱わないような事柄は話題として含まれなく,「野球の勝敗」といったような複数の人の注目を集めるような事柄を話題とする．

本論文では特定の分野における注目事件のような分野的な特異性をもつ話題の抽出，及び話題を扱った記事を精度良く取り出すといった目的から，抽出対象とする「話題」を「興味を同じくする人々の間で注目されている事柄」と定義する．このように定義することによって，従来の頻度変化の分析では埋もれてしまったような小さな分野の話題を取得可能になると共に，話題と記事の関連付けの精度の向上も期待される．例えば専門的な趣味の話のような扱っている記事数が少ない事柄においても，それについて書いている発信者間での興味の一致度合いをみることにより，単に頻度の低い語句と区別して話題として扱うことが可能になる．また，興味の定まった集団においては，その中で扱われる話題についてきちんとした意見や内容が書かれているだろうという仮定ができる．

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．

4. 提 案 手 法

第 3 節における話題の定義から，各発信者の興味分野を見る

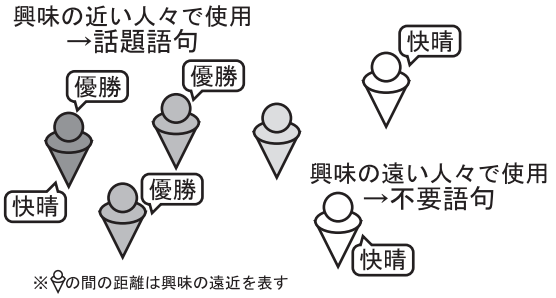

図 1　分野特徴語句のイメージ

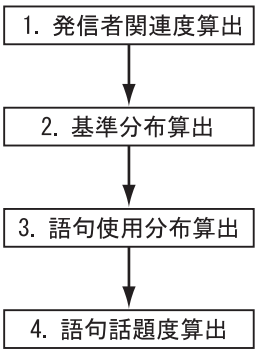

図 2　話題度算出処理の概要

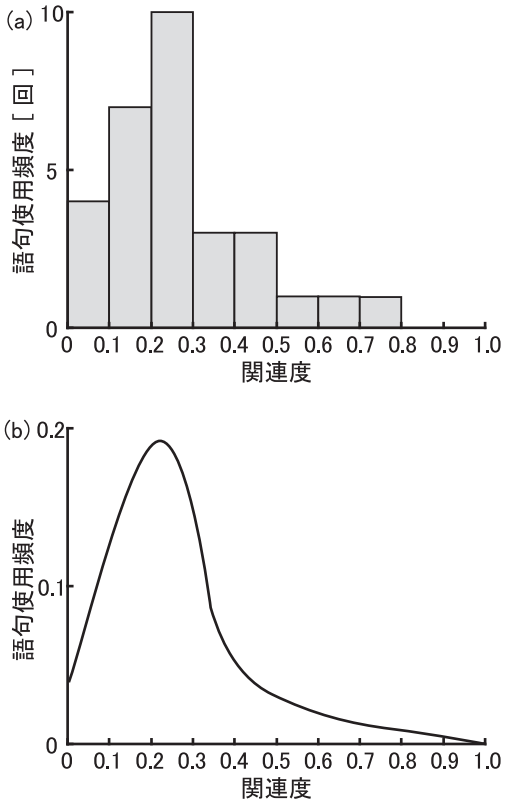

図 3　関連度に対する語句使用頻度の例

ことにより，ブログ記事で扱われている話題を表している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示す．

本手法は，ある発信者 i にとって語句 w_k がどれだけ話題としての意味を持つかの度合いを話題度として算出する．話題度の算出に当たっては，あらかじめ発信者 i と他の発信者との興味の一致度合いを関連度として算出しておく．その後に，その関連度が高い人の間で語句 w_k が使用されているかを，関連度に対する語句使用頻度の分布として算出する．例えば，発信者 i にとってある関連度 0.1〜0.2 の範囲の他の発信者が 5 名いて，それぞれが語句 w_k を 3 回，2 回，0 回，1 回，1 回使用した場合は，関連度 0.1〜0.2 の範囲の語句使用頻度は 7 回となる．このようにして図 3 の (a) のような関連度に対する語句使用頻度の分布図が得られる．実際には，もっと細かい区分で集計を行った後に全体での合計使用頻度を元に正規化を行うため，図 3 の (b) のような関連度と使用頻度の分布の図ができる．この分布が図 4 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．

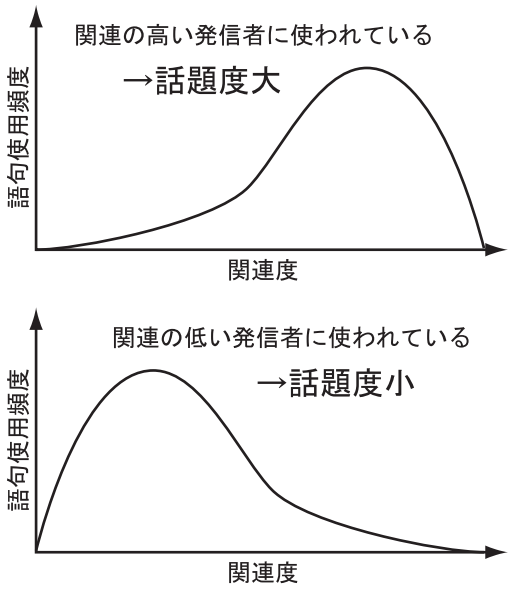

図 4　話題度算出のイメージ

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，（1）発信者 i とその他の発信者との興味の関連度を算出し，得られた関連度の分布から，（2）全ての発信者がある語句を 1 回使用した場合の，関連度に対する語句の使用頻度の分布を基準分布として求める．（3）次にある発信者 i による記事中の各語句 w_k について，w_k の関連度に対する使用頻度の分布を求める．（4）最後に，図 4 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．この際，全体での関連度の分布は図 3 の (b) のようにポアソン分的な傾向を持つため，その影響を受けて大半の語句において関連度の低いほうに分布が偏る．その影響を除去するため，語句 w_k に対応した関連度の分布と基準分布とを比較することにより話題度の高低を算出する．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

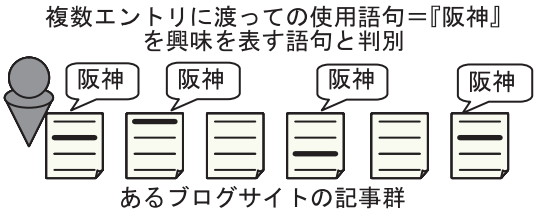

図 5　興味ベクトル算出のイメージ

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ブログ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記事から，語句の使用傾向を語群ベクトルとして抽出する．例えば図 5 に示したように，複数のエントリで「阪神」という言葉を用いていた場合，この語句を発信者の興味を表す語句として高い重みをつけることにする．

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 V_i と呼ぶこととし，興味ベクトル V_i の値は，次の式によって求まる．

$$V_i = (x_{i1}, x_{i2}, x_{i3}, \ldots) \tag{1}$$

$$x_{ik} = ef_i(w_k) \times log\left(\frac{N_u}{uf(w_k)}\right) \tag{2}$$

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を掛け合わせて求まる値である．一般的な語句の重みを下げる要素には，IDF 値をユーザ単位での処理に合うように変更した，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を使用した．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．

$$R_{ij} = \frac{V_i \times V_j}{|V_i||V_j|} \tag{3}$$

例として，ブログサイト 250 サイトにおける，ある発信者と他の発信者 249 人との関連度を，0.01 刻みで集計した分布のグラフを，図 6 に示す．

4.2　基準分布算出

語句の使用分布の偏りを元に話題度を算出するため，全発信者が一回だけ使用した語句の関連度ごとの語句使用頻度の分布を，基準分布 RD_i として抽出する．この基準分布は，各関連度範囲の発信者数に語句の使用頻度 1 を掛けることにより求まるので，先にあげた関連度の分布のグラフ図 6 と同じ形の分布となる．

話題度算出処理を行う処理対象記事の発信者 i と他の発信者との関連度集合の値の分布 RD_i を，0〜1 を N 分割した各範囲の値を持つ $R_{ij}(i \neq j)$ の数を集計し，それを全体の語句使用頻度で正規化することにより求める．

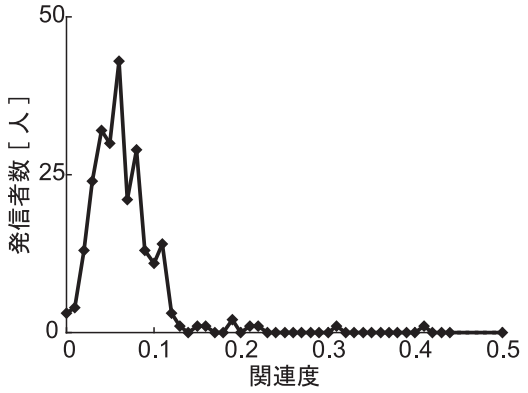

図 6　関連度の分布例

$$RD'_i(n) = \begin{cases} 1 & if \quad \frac{n-1}{N} \leqq R_{ij} < \frac{n}{N} \\ 0 & else \end{cases} \tag{4}$$

$$RD_i(n) = \frac{RD'_i(n)}{\sum_n RD'_i(n)} \tag{5}$$

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　語句使用頻度分布の算出

発信者 i による記事に含まれるそれぞれの語句 w_k に対して，語句 w_k が発信者 i とどの程度の関連度を持つ他の発信者の間でよく使用されていたかを表す，関連度に対する語句使用頻度の分布 WD_i を次の式で求める．基準分布 RD_i と同様に，WD_i も全体での語句 w_k の使用総数で正規化を行うこととする．

$$WD'_i(n) = \begin{cases} ef_j(w_k) & if \quad \frac{n-1}{N} \leqq R_{ij} < \frac{n}{N} \\ 0 & else \end{cases} \tag{6}$$

$$WD_i(n) = \frac{WD'_i(n)}{\sum_n WD'_i(n)} \tag{7}$$

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 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，WD_i の分布は 4.2 で求めた RD_i の分布と近くなる．

4.4　語句話題度算出

図 4 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，それを基準分布 RD_i でキャンセルした形での話題度の算出手法を構築する．話題度の値は，図 7 に示したように，RD_i を基準として，それに対して語句 w_k についての分布 WD_i が左側に偏っていた場合には負の話題度を，右側に偏っていた場合には高い話題度を算出するようにする．

その為に，図 8 に示すような基準分布 RD_i における関連度

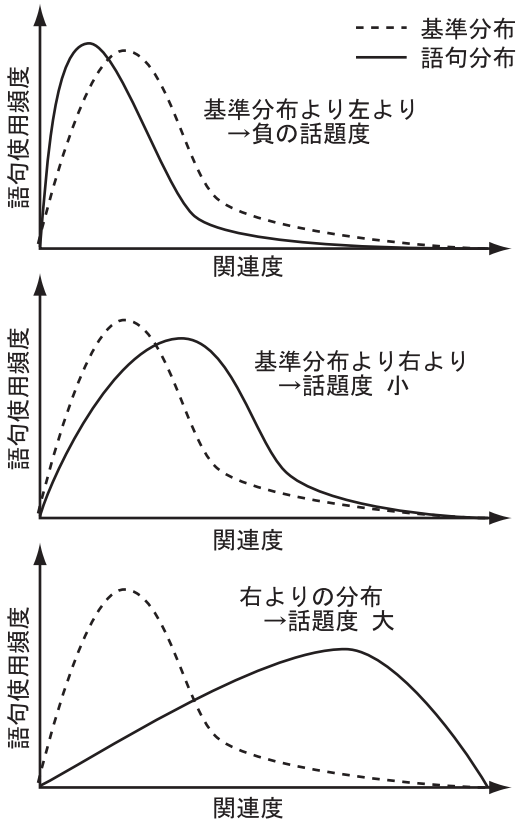

図 7　関連度に対する語句使用頻度の例

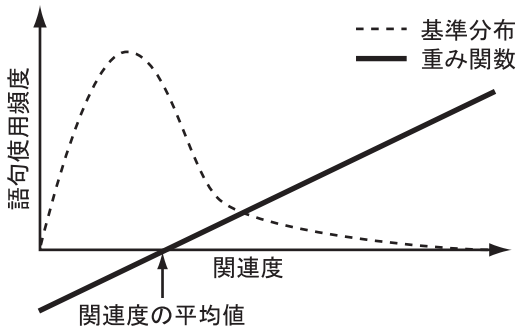

図 8　関連度に対する語句使用頻度の例

の平均値 n_0 での重みをゼロとする線形の重み付け関数を導入し，分布の各関連度の範囲において語句の関連度分布と基準分布との差に重み付け関数の値を掛けた値を累積したものが話題度になる．発信者 i にとっての語句 w_k の話題度 $TS_i(w_k)$ を算出する式は，次のようになる．

$$TS_i(w_k) = \sum_n \left\{ (WD_i(n) - RD_i(n)) \cdot \frac{n - n_0}{100} \right\} \qquad (8)$$

ここで n_0 は次の式により求まる．

$$n_0 = \frac{\sum_n \{RD_i(n) \cdot n\}}{\sum_n RD_i(n)} \qquad (9)$$

5. 実験と考察

提案手法によってブログ記事中の話題が適切に取得されているかを確認するために，人手でブログ記事中の話題の有無を判別した正解データを用いた評価実験を行った．実験は 2 段階で行われ，まず話題を含む記事についてのみ話題度が算出されているかの評価を行い，次に話題を含む記事について適切な語句に対して高い話題度が重み付けられているかを確認した．

5.1　ブログコーパス

実験対象の処理データとして，ライブドアブログのスポーツジャンルとアニメ・漫画関連ジャンルに所属するアクティブなブログサイト 250 サイトの 2005 年 4 月 20 日〜5 月 19 日の記事の集合である 11513 記事を用意した．この際，アクティブなブログサイトの基準として「月１０件以上の投稿」という条件を設定した．このうちの最後の一週間である 5 月 13 日〜19 日の 2530 記事について，３名の作業者による正解データの作成を行った．

各記事について，記事が話題を含むか否かの判断である話題有無フラグと，記事中の話題を表す語句を複数選択した話題語句群との 2 種類の正解データを作成した．話題語句群は，話題有無フラグが正である「話題を含む」と正解付けされた記事についてのみ行った．また話題語句とする対象には，形態素解析で得られる名詞と，名詞の連続からなる複合名詞のみを設定した．

話題を含むか否かの判定においては，最終的に話題の閲覧を目標とすることから，不特定の他者と共有され得る話題を持つものを選ぶこととした．その為，個人的な行動を記録したような日記や，特定の人物に対して書かれた私信といった内容の記事に対しては，話題有無フラグが偽に設定される．例えば,『今携帯で学校から書いてます．隣で某ギタリストが騒いでます．これを見たらメールくれるとうれしいです.』といった内容は特定の相手に書かれた私信と考え話題の存在しない記事と判定し，テレビ番組の感想記事やスポーツの観戦記などは話題を持つ記事と判断し，話題有無フラグを真にする．３名の作業者によって作成された正解データのうち，３名ともが話題を含むと判断した記事 756 件と３名ともが話題を含まないと判断した 551 件の計 1307 件を話題有無判定の正解データとして採用した．

記事中の話題語句の抽出においては，記事中で扱われている事柄におけるキーワードを選択するようにした．例えば野球の試合であれば，チーム名や活躍した選手名等が話題語句として抽出されることとなる．例えば,『横浜のクルーンが 159 キロをマーク』という記事であれば,『横浜』『クルーン』といった言葉を話題語句として抽出する．複数の話題を含む記事については，両方の話題に対応する語句を抽出することとした．上で述べた３名が話題を含むとした 755 件の記事について，３名の作業者のうち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採用した．うち 10 件の記事については 2 名以上が話題語句と判断した語句が存在しなかったため除外し，計 745 件の正解データを採用した．

5.2　話題語句抽出処理

正解データの付与対象となった 2530 件に対して，提案手法を用いて各語句に対して話題度の算出を行った．話題度の算出対象とする語句は，形態素解析を行って得られた名詞と，名詞

表 1　話題の有無による平均話題度の差異

記事種別	話題度平均
話題含む	0.105
話題なし	0.030

の連続からなる複合名詞を処理対象と設定した．また代名詞については処理対象から削除を行った．

算出時の条件として，各発信者の興味ベクトルと関連度の算出においては 1ヶ月間の記事を用いて処理を行うことにより，長期間にわたる興味を抽出するようにした．また興味ベクトル作成時の構成語句には，Web 上で得られる注目語句辞書であるはてなキーワード [7] に載っている語句を対象とし，関連度抽出の処理軽減と精度の向上を図った．話題度を算出する際の，語句使用分布の算出の範囲は一週間に設定した．また，比較対象として TF-IDF による語句の重み付けも行った．IDF 値の算出は，一ヶ月間の全ドキュメント群から算出を行った．

5.3　話題あり記事選択

話題を含む記事に対しては高い話題度を算出し，話題を含まないような記事に対しては話題度を低く算出しているかを確認する実験を行った．話題判別の正解データの各記事に対して，提案手法による話題度を算出し，その上位５つの語句の話題度の平均を取る．この際話題を含む記事における話題度平均が高くなっていれば，話題を含む記事についてのみ話題度が算出されているため，日記的な記事の除去が成功していると考えられる．

得られた結果は表 1 のようになった．

また，ある話題度平均がある閾値以上の記事のみを話題有り記事と選んだ場合の，話題有無フラグが真となっている正解データに対する精度と再現率の値も求めた．話題度平均が 0.06 以上の記事 556 件を話題有り記事と選んだ場合，その中に話題有りと正解付けされている記事数が 500 件合ったため，精度は 90 %，再現率は 66 %になった．

5.4　話題語句の評価

各記事に対する話題語句の抽出が適切にできていたかを判断するために，正解データとの話題度上位語句の比較検討を行った．話題語句の正解データと，各記事の話題度上位語句との一致率を精度として算出した．話題度上位語句の選出は，１件のみを選んだ場合から，５件選んだ場合までの５通りについて算出を行った．比較対象として，TF-IDF を用いて各記事の重要語句を重み付けした際の上位１件から５件を選んだ場合の精度も算出した．

評価対象の正解データとしては，前の実験において話題度 0.06 以上の値を得た記事中の話題を持つという判定を得た 500 記事のうち話題語句が設定されている 492 記事使用した．

得られた結果を表 2 に表す．

6.　考察と課題

6.1　考　　察

提案手法において，話題の存在する記事に対して高い話題度を算出することができていることが確認された．この結果は，ある閾値以上の話題度平均を持つ記事のみを対照とすることにより，話題表示時に不必要な記事を表示してしまうという問題を改善するのに有効であることを示唆している．

表 2　話題度上位語句の話題語句抽出精度

選出数	TF-IDF	提案手法
1	51.4	52.0
2	48.1	48.4
3	45.8	49.1
4	43.3	48.2
5	41.8	46.2

表 3　全話題有り記事に対する話題語句抽出精度

選出数	TF-IDF	提案手法
1	55.3	51.8
2	51.3	48.8
3	48.7	48.2
4	46.2	46.4
5	44.9	44.5

表 4　話題度の算出例

語句	話題度	TF-IDF	TF
マリノス	0.99	6.36	1
ＡＣＬ	0.77	14.00	2
過密日程	0.59	15.94	2
アジア	0.31	21.56	3
Ｊリーグ	0.87	6.00	1

また，話題度が一定以上の記事に対しては，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．精度の算出対象を話題有無フラグが真になっている全記事に拡大すると，表 3 のような結果となり，TF-IDF と同等もしくは若干劣る値にその精度が低下する．このことはその分野に興味を持つ発信者が複数存在しないような記事に対して，正しい重み付けを行えないことによるものと考えられる．

また TF-IDF と比較すると，得意とする記事に差異があることが確認された．短い記事で記事中の各語句の TF の値がほとんど１に近くなるような記事に対しては提案手法が有効となる傾向があった．一方である程度記事が長くなり TF の値にばらつきが出る場合では TF-IDF が有効になる傾向があった．実際に抽出した例を表 4 に示す．サッカーのアジアチャンピオンシップリーグについて書かれたブログ記事における話題度上位５つの語句と，その話題度・TF-IDF 値・TF 値について纏めた表である．

6.2　今後の課題

今回の処理は全ての発信者間の関連度を算出し，また発信者ごとに語句の話題度を算出する手法を用いたため，計算量が非常に大きくなっている．ブログサイト全体に対して処理を行うには，関連度算出部分の処理量を軽減する必要がある．提案手法では話題度の算出において，平均以上の関連度を持つ場合に正の重みを与えるような処理を行っていた．この際，最終的には話題度の高い語句のみを抽出するため，負の重みがかかる平均以下の関連度を持つ部分については高関連度の部分に比べ関

連度の算出精度が求められない．その為，経済について書いている発信者とゲームについて書いている発信者といったような，まったく違う分野について書いている発信者間については，経済・ゲーム・ＩＴといった大雑把な分野分類を用いて一括で低い関連度を与えるといった方法を用いることにより，計算量の削減が可能になると考えられる．

今回は分析対象の期間が 1 週間と短かったため，話題度の算出に時間的な要素を加える必要がなかった．しかし長期間のデータを処理するに当たっては，従来行われてきた語句使用頻度の時間的な変動による話題抽出手法を提案手法に組み込んでいくことが必要である．

7. おわりに

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，取得した話題情報をどのように活用していくかの情報も重要であると考えている．本論文では話題抽出の手法についてのみ検討を行ってきたが，今回取り扱ったような細かな分野における話題は従来のように一覧で表示にするには数が多すぎる．今後は話題を検索するための方法や，ブログを閲覧している際に各記事に話題を用いたナビゲーション機能を埋め込むといったような，話題情報を用いたより便利なインタフェースの構築方法も検討していきたい．

文　献

[1] 総務省,“ ブログ・SNS の現状分析及び将来予測，”(2005).

[2] Chris, A.,“ The Long Tail, ”Wired Magazine，Issue 12-10, (2004).

[3] Kleinberg, J., “ Bursty and hierarchical structure in streams, ”In Proc. the 8th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining, (2002).

[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).

[5] 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).

[6] BlogPulse, http://www.blogpulse.com/

[7] “ はてなキーワード, ” http://d.hatena.n.jp/keywordlist.

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

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

[10] K. Ishida, ”Extracting Latent Weblog Communities,” Presented at the Workshop on the Weblogginng Ecosystem at the WWW2005, (2005).