2I1-1

ブロガの発信話題履歴を利用した関連話題抽出手法

Discovery of Related Topics Using Serieses of Blogsites' Entries

関口 裕一郎[*1] 川島 晴美[*1] 内山 匡[*1]
Yuichiro SEKIGUCHI Harumi KAWASHIMA Tadasu UCHIYAMA

[*1]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
NTT Cyber Solutions Laboratories, NTT Corporation.

In this paper, we propose a novel approach to extract topics which relate to the base topic from blogosphere. In the past research, it extract related words or topics by analyzing word cooccurance in each documents. In our approach, we treat blogger's recent posts as a set of documents, and we analyze word cooccurence and words' repeat rate in each document set. Using these characteristics, we extract related topics which belong to similar class to base topic. We verified that our approach is effective to extract related topics using 9 million blog documents.

1. はじめに

近年のブログの浸透により，多数の一般の人々が自分の感想や体験を綴った記事をネットワーク上に発信するようになってきた．このようなブログに含まれる体験情報は，商品やサービスに対するクチコミ情報として幅広いユーザに利用されており，その評判の内容が人々の購買判断に大きく影響を与えるようになってきている．商品を販売している企業においても，日々更新されるブログ記事の中で自社の商品や関連する事柄がどのように語られているかは売れ行きを左右する要素となってきており，ブログにおける商品について書かれた記事の調査がマーケティング活動の一要素となってきている．

評判情報を求めてブログ記事を検索している途中に，ある商品の比較対象として語られている他の商品名を見つけることが往々にしてある．このような元々調べていた内容と関連する話題の発見は，一般ユーザにとっては商品の検討の幅を広げる上で有用であると考えられるし，マーケティング分析を行っている企業ユーザにとっても競合商品を知るという意味で有用であると考えられる．本論文では，このような元々の検索していた事柄と関連する事柄を関連話題としてユーザに提示することによってユーザの調査活動の支援を行うことを目指し，そのための関連話題抽出を行うことを目的とする．

従来にも同義語辞書の拡張や検索クエリの作成支援を目的として，ウェブの文書集合から関連語句を抽出する試みが多く行われてきている．ウェブ閲覧ログ情報を利用する手法や [1]，ウェブ検索結果の類似性を元にする手法 [2]，URL 情報の類似性を手がかりとする手法 [3] などが提案されている．これらの手法では，各語句がウェブ文書集合中のどのような文書中で使われているかを元に，ウェブ文書集合中で語句間の関連性を算出する．これらの手法は汎用的に関連性の高い語句を抽出することができる．

一方本論文の目的とする関連話題抽出では，ブログでの評判情報の調査を支援することを目的とするため，野球のチーム名に対して「ホームラン」といった語句が出ることは望ましくなく，調べている語句に関連してなおかつ評判の対象となるような語句を取得することが課題となる．これに対応するため，ブログサイト中での使用パターンから関連語句の意味的な階層を推定することにより，商品を表す語句を入力して関連する商品名を得られる関連話題抽出手法を提案する．また 1 か月のブログ文書を用いて，提案手法の有効性の評価を行う．

A: 関口 裕一郎，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，神奈川県横須賀市光の丘 1-1，046-859-2333，046-859-5552，sekiguchi.yuichiro@lab.ntt.co.jp

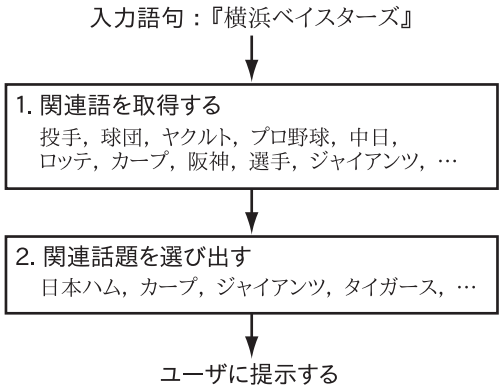

図 1: 提案手法の処理の流れ

2. 提案手法

提案手法は，ブログ文書集合からある語句に対する関連話題を抽出する．以下，関連話題を取得する元となる語句を入力語句と表記する．関連話題とは，入力語句と似通った分野で語られる言葉であって，なおかつ入力語句と同等の意味的な階層に所属する語句と定義する．似た分野で語られる言葉とは，野球のチーム名に対して「ホームラン」「イニング」「三振」といった分野特有の単語や他のチーム名や選手名などの固有名詞等を意味する．また意味的な階層が同等の単語とは，チーム名に対して他のチーム名といったように，固有名詞のクラスとして同じか近い所に所属する単語を意味する．以下，似た分野で語られる言葉を「関連語」，関連語であってなおかつ意味の階層が同等の言葉を「関連話題」と区別して用いる．

提案手法の処理の流れを図 1 に示す．提案手法は，入力語句に対して関連語を抽出する処理と，得られた関連語の中から「関連話題」を抽出する処理の 2 段階で構成される．また関連話題の抽出において，固有名詞のクラス抽出といった技術は用いないこととする．これは，提案手法が対象とするブログ記事集合は新たな文書が高頻度に提供され記述も口語に近い形式であるため，これらの技術で用いる辞書の構築や学習といった作

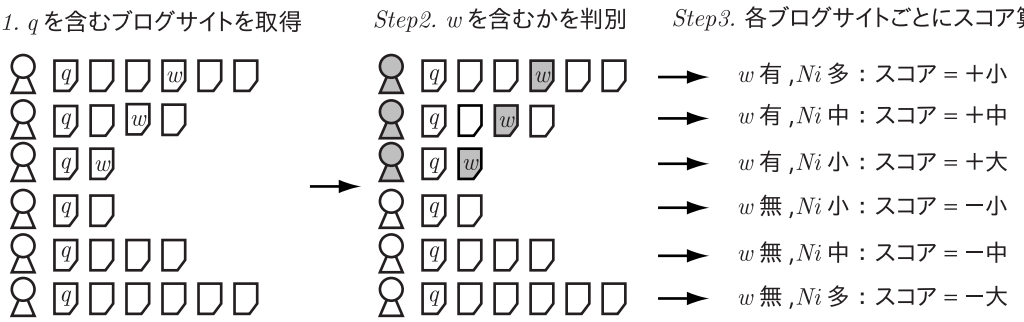

図 2: 関連度算出処理の流れ

業を行わねばならなく，これらの作業が今現在の話題を調べるという作業において望ましくない為である．

提案手法は，関連語を抽出する為に入力語句に対する関連度を様々な語句に対して計算する処理と，算出された関連度の高い語句の中から意味的階層が近い語句を関連話題として選び出す処理から構成される．以下，それぞれの処理について詳細に説明する．

2.1 関連度算出アルゴリズム

入力語句と関連の高さを合わす関連度を算出する手法を説明する．この処理では，入力語句を含むブログサイトの集合を抽出し，それらのブログサイトに含まれる文書中で特徴的に多く使用されている語句を，入力語句に関連する話題語句とみなして高い関連度を算出する．

関連度の算出において，入力語句を含むブログ記事を持つブログサイトに含まれる記事全体の集合を解析対象とする．入力語句を含むブログ記事のみを処理対象にしないのは，ブログ記事においては 1 つの事柄に対する感想や評判は 1 つの記事として書かれる傾向が存在し，入力語句と似た別の事柄についての感想や評判は別の記事として書かれていることが多い為である．このように処理範囲を広げることにより，入力語句は含んでいない他の記事で書かれている類似商品を表す語句も関連度算出の対象とすることを目的としている．

入力語句として q が与えられた時，q とある語句 w の関連度を算出する処理の流れを，図 2 に示す．最初に q を含む文書を持つブログサイトの集合 $U \ni \{u_1, u_2, ..., u_i, ..., u_n\}$ を抽出する．次にブログサイト集合 U の中で，語句 w を含む文書を持つブログサイトと持たないサイトを判別する．この処理が入るため，関連度を算出す語句 w は，U に含まれる文書集合中の語句に限られる．w を持つか否かの判別の後に，ブログサイトごとの語句 w のスコアを算出する．記事数が少ないサイトが w を含んでいた時に関連度が高くなり，記事数が多いサイトが w を含んでいなかった時に関連度が低くなるように，w を含む文書を持つブログサイトに対してはプラスのスコアを，持たないブロガに関してはマイナスのスコアを算出し，それを足し合わせた値を w の関連度スコアとする．具体的には，ある記事に語句 w が含まれる確率 $p(w)$ となる場合に，記事数 N_i のブログサイトが w を含む確率は $1-(1-p(w))^{N_i}$，w を含まない確率は $(1-p(w))^{N_i}$ となるので，この確率から求まる情報量に w を含むか否かに応じて正負の符号をつけたものをスコアとする．最後に，ブログサイトごとのスコアを式 1 のように足し合わせて，w の関連度スコアとする．$p(w)$ は w の df 値を全文書数で割った値で近似する．

表 1: 『横浜ベイスターズ』に対する関連度上位 10 語

順位	語句	関連度スコア
1	投手	890.45
2	球団	870.14
3	ヤクルト	859.77
4	プロ野球	849.88
5	中日	812.94
6	ロッテ	791.31
7	カープ	784.53
8	阪神	767.72
9	選手	748.90
10	ジャイアンツ	748.43

$$score(w) = \sum_{i=1}^{n} \begin{cases} -log(1-(1-p(w))^{N_i})) & u_i \text{ が } w \text{ を含む} \\ log((1-p(w))^{N_i}) & \text{含まない} \end{cases} \quad (1)$$

後述するデータセットから,「横浜ベイスターズ」で分析した結果得られた関連度の上位 10 の単語とその関連スコアを表 1 に示す．野球関係の語句やチーム名が関連度の高い語句として抽出されていることが分かる．

2.2 関連話題語の絞り込み

上記の式 1 で求められる関連度においては，分析対象語句との関連の高さを算出しているため，例えば野球チーム名を解析対象語句とした際に，同一の意味階層に所属他の野球チーム名の他にも,「投手」や「球場」といったその分野特有の用語等も関連語句に含まれてしまう．そのため，関連度の高い語句の集合から，関連話題となる語句を絞り込む処理を行う．

多くのブログはそのブロガの興味に従って書かれるため，分野的に近い内容について繰り返し記事が書かれる傾向がある．その為，1 つのブログサイトの中で複数の記事に渡って繰り返し使用される語句はその分野特有の語句である可能性が高い．一方で個々の記事の中心的なトピックとなる語句は，その記事のみにし書かれない傾向がある．例えば図 3 に示すように，食べ歩いた感想を載せているブログサイトにおいて「パスタ」といった料理特有の単語は複数記事にわたって書かれるが，個別の料理店の名称等は個々の記事にしか出現しない．

以上の想定から，上記の式 1 で高い関連度を持つ語句に対して，その語句を含む各ブログサイトでの語句の使用記事数の

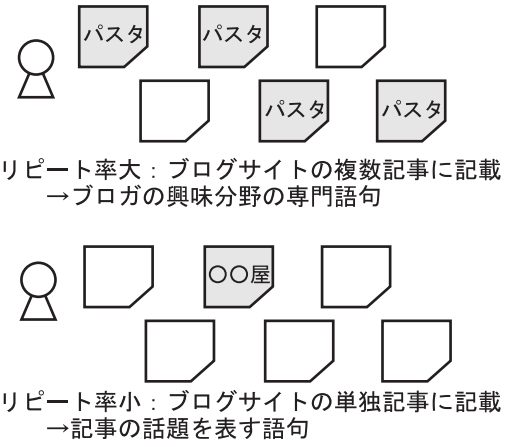

図 3: リピート率と語句の種類

表 2: 『パンパース』と関連度上位 10 語をリピート率で並び変えた表

順位	語句	リピート率
1	赤ちゃん	0.297
2	おむつ	0.268
3	体重	0.224
4	ベビー	0.224
5	しり	0.220
6	オムツ	0.208
-	パンパース	0.191
7	メリーズ	0.188
8	新生児	0.174
9	パース	0.165
10	ムーニー	0.158

割合を求め，その平均をリピート率とする．このリピート率が入力語句の値と近い語句を選ぶ事により，意味的階層の近い関連話題を抜き出すことが出来ると考えられる．

後述するデータセットから，入力語句を「パンパース」とした際の関連度の上位 10 単語をそのリピート率で並び変えた結果を表 2 に示す．育児に関連する記事でよく使われる用語はリピート率が高く,「パンパース」と同等の商品である「メリーズ」や「ムーニー」はリピート率が低くなっている．

3. 検証実験

提案手法の有効性を確認するための検証実験を行った．12 語の入力語句を選び出し，それぞれに対してブログ文書集合を用いて関連語及び関連話題の抽出を行い，その精度を評価した．

検証実験の条件として，2007 年 11 月 15 日段階での関連話題抽出処理を行う状況を想定し，処理対象として 2007 年 10 月 15 日から 11 月 15 日までの一ヶ月間に収集したブログ記事 9,706,974 記事を用いた．その中に含まれるブログサイト数は 1,239,721 サイトであり，1 サイト辺りの平均記事数は 7.8 記事となる．解析を行うに辺り，大量の投稿しているブログサイトの影響を除去するため，関連度スコア抽出時の処理に利用する 1 ブログサイトあたりの記事数は，最大で最近 20 記事に限定することとした．

表 3: 精度評価に用いた語句

ジャンル	語句
スポーツ	横浜ベイスターズ
スポーツ	浦和レッズ
スポーツ	鈴木啓太
スポーツ	中村俊輔
ゲーム	任天堂
ゲーム	Wii
ゲーム	PS3
ゲーム	スーパーマリオギャラクシー
ゲーム	Wii スポーツ
育児	パンパース
育児	ムーニー
育児	メリーズ

また入力語句には,「育児」「ゲーム」「スポーツ」の 3 分野の商品名，スポーツチーム名，選手名を合わせて 12 語を用いた．使用した入力語句は表 3 に示す 12 語となる．

3.1 関連度スコアの評価

それぞれの語句に対する関連度上位 10 語，30 語を抽出し，抽出された各語句に対して入力語句と関連があるかどうかを人手で判別し，その適合率を求めた．評価対象を上位 10 語，30 語と設定したのは，本手法は関連語句をユーザに提示することを目的とするため，人が一目で見られる程度の数の語句数に多くの正解が含まれていること望ましい為である．

また正解の判別においては，各語句が入力語句と同等の意味階層に所属する関連話題，入力語句と同分野で使われる語句である関連語，非関連語の 3 つのタイプに分類した．関連度スコアの算出の段階では，抽出された語句が関連語か関連話題かの区別は行っていないため，10 語もしくは 30 語中に含まれる関連語と関連話題の数の和が抽出結果の適合率となる．

全体でと各ジャンルごとでの評価結果を，図 4 に示す．全体として関連語と関連話題を合わせて，上位 10 語中で適合率 0.85，上位 30 語中で適合率 0.77 という結果になった．全体的に上位 30 語において上位 10 語に比べ若干値が低下する傾向があった．この原因としては，形態素解析で不適切に抽出された語句や日付表現，時刻表現が入っていたためであった．

分野別で見ると，上位 30 語においてスポーツと育児についてはそれぞれ 0.87，0.81 という結果になった一方，ゲームにおいては 0.67 と他に比べて低い値となった．これはゲームジャンルにおいては英語の商品名が多い為，商品名が途中で分割されてできた不適切な語句がいくつか存在し，それが適合率を低下させていた．

3.2 関連話題抽出の評価

関連語句の抽出において得られた上位 30 語に対しリピート率を求め，入力語句のリピート率との差異を元に並べ替えた上位 10 語を得ることにより，関連話題の抽出を行った．得られた関連話題のリストと，関連度の上位 10 語のリストに対して，平均逆順位を求めて評価を行った．平均逆順位 (Mean Reciprocal Rank, MRR) は，最も上位である適切な関連話題語句の順位の逆数を 12 語の抽出結果に対して平均した値であ

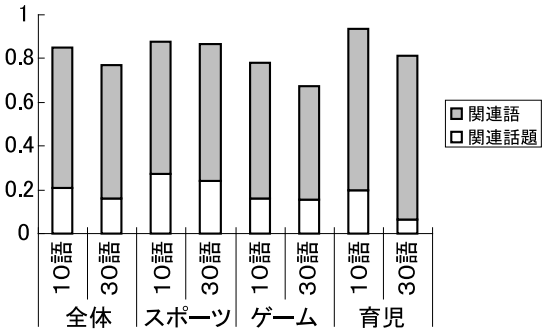

図 4: 関連度の評価結果

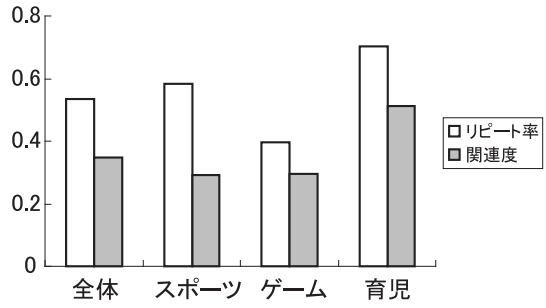

図 5: リピート率順での平均逆順位

り，適切な語句が上位に抽出できていると値が大きくなる評価値である．

全体での平均逆順位と，各ジャンルごとの平均逆順位による評価結果を図 5 に示す．全体として，全体では，リピート率による並び替えを行った場合が 0.54 で関連度順では 0.35 となった．各ジャンルにおいてもリピート率による並び替えを行うことにより，平均逆順位が向上している．これから，リピート率による並び替えで，関連話題が順位上位に来ていることが分かる．

またリピート順に並び変えた上位 10 後における関連話題，関連語の適合率の評価を，3.1 と同様に行った．その結果を図 6 に示す．リピート率順にした際に，全体で関連話題の適合率が 0.21 から 0.28 に向上した．これから，リピート率を用いて関連語を並び変えることにより，関連話題となる語句が上位にくると同時に多数の関連話題が上位 10 語に含まれるようになっていることが分かる．

一方で，関連語も含めての適合率は 0.85 から 0.78 へと低下した．これは形態素解析ミス等で混入したノイズ的な語句が，上位に来ていることによる．この原因は，これらの不適切な語句は偶発的に抽出されるため，ブログサイト中で複数の記事にわたって出現することが少なく，話題語句と似たようなリピート率を持つためと考えられる．

4. まとめ

本論文では，ブログサイトの発信記事集合を利用することにより，入力語句に関する関連話題を取得する手法について論じた．入力語句を含むブログサイト中の使用人数に応じた語句の関連度の算出手法と，ブログサイト中での語句のリピート率

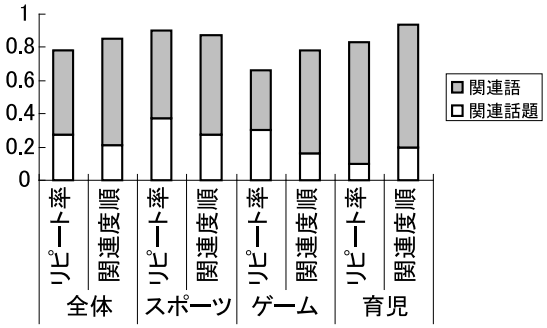

図 6: リピート率順での適合率

に注目した関連話題の抽出手法を提案した．

3 分野 12 語の語句を用いて，実際のブログ文書集合から関連語，関連話題を抽出する評価実験を行い，関連語句抽出において上位 10 語中で適合率 0.85，上位 30 語中で適合率 0.77 を実現した．また関連話題の絞り込みにおいては，平均逆順位で 0.54，関連話題語句の適合率において 0.28 を実現し，関連度順に並べた場合に比べ値が向上することを確認した．しかし関連話題の絞り込みを行うことにより，関連語句でない語句が上位に入るという問題が生じた．これについては関連度スコアとリピート率の双方を考慮した抽出アルゴリズムを構築するなどして，対処を行う必要がある．

今回の手法では，ブログ文書に与えられている時刻情報については考慮を行わなかった．ブログ文書集合からの話題語句抽出においては，時刻情報を利用してその時々に話題となっている語句を抽出する手法がいくつか提案されている [4][5]．これらの手法と組み合わせることによって，さらに精度の高い関連話題抽出が可能になると考えている．

参考文献

[1] “大域ウェブアクセスログを用いた関連語の発見法に関する一考察,” 大塚真吾，豊田正史，喜連川優: 情報処理学会論文誌 TOD, Vol. 46, No.SIG8, pp.82-92 (2005).

[2] “ウェブを利用した関連語収集,” 小原恭介，山田剛一，絹川博之，中川裕志: FIT2004, pp.183-184 (2004).

[3] “URL の類似性に着目した WWW 空間からの関連語自動収集手法,” 獅々堀正幹，山本一晴，小泉大地，北 研二: DBSJ Letters, Vol.6, No.1, pp.57-60 (2007).

[4] “単語出現の意外性に基づく話題性評価手法,” 佐藤吉秀，坂井俊之，川島晴美，奥田英範: 情報処理学会自然言語処理研究会 2007-NL-181, pp.87-92 (2007).

[5] “気づきに注目した情報提供システム「HotWindow」の開発,” 川島晴美，佐藤吉秀，関口裕一郎，佐々木努，大久保雅且，奥雅博: 電子情報通信学会論文誌D, Vol. J89-D, No.11, pp.2445-2457 (2006).