社団法人 情報処理学会　研究報告
IPSJ SIG Technical Report

2004－NL－160　(12)
2004／3／5

名詞句と単語の勢いを用いた話題抽出手法の提案

石井　恵[†]　中渡瀬秀一[†]　富田　準二[†]

† 日本電信電話株式会社 NTT サイバースペース研究所

〒 239–0847 神奈川県横須賀市光の丘 1–1

あらまし　本稿では掲示板のような場につぎつぎ書き込まれるメッセージの列における利用者の興味をそそる話題や、メッセージの書き込みに即応したそれら話題の掛け合いの発生や盛り上がり、流行の兆しなどの意味のある変化を利用者に飽きがこないように提示するための話題抽出手法を提案する。本手法ではユーザの興味をそそる話題として名詞句や固有名詞を話題として抽出する。そして、話題の勢いを扱える話題のスコアリング手法の提案により、それら話題の意味のある変化を利用者に飽きがこないように提示することを実現する。

キーワード　話題抽出、掲示板システム、時系列テキストマイニング

Topic Extraction from a Message Stream using Noun Phrase and Word Pressure around the latest Message

Megumi ISHII[†], Hidekazu NAKAWATASE[†], and Junji TOMITA[†]

† NTT Cyber Space Labolatories, NTT Corporation

1–1 Hikarinooka Yokosuka-shi, Kanagawa, 239–0847, Japan

Abstract　This paper proposes a topic extraction method for a message stream such as BBS. This method extracts noun phrases and proper nouns as topics attracting users and ranks those topics using their pressure around the latest message. As a result, the method can tell users interesting topics and a moment when a meaningful change happens on those topics.

Key words　Topic extraction, BBS, Text stream mining

1. はじめに

インターネットの普及により、掲示板システムやチャットシステム等同じ興味をもつ人々がコミュニケーションを行なう場の利用者の裾野は爆発的に広がった。これら場では、企業の評判、商品へのニーズ、問題軽決、お勧め情報等、世の中の様々な話題がのぼるため、企業にとっては無視出来ないものとなりつつある。また、一般の利用者にとっては情報収集や娯楽の場として非常に魅力的である。

このような状況で、利用者の興味をそそる話題やメッセージの書き込みに即応したそれら話題の意味のある変化は特に魅力的な情報であり、利用者はそれら情報を探し求めている。利用者の興味をそそる話題として、利用者が以前から探し求めていた話題や自分の興味に関連する思いもしなかった話題があげられる。意味のある変化として、メッセージの書き込みに即したそれら話題の勢いの強弱変化が上げられる。書き込みに応じた話題の勢いの増加はその時点で、コミュニティにおいてその話題に対する興味が強まったことを表し、話題に対する掛け合いの発生や盛り上がり、流行の兆しを示すことがあるからである。

これら魅力的な情報を利用者に提示することにより、多くの利用者を場へ引き込み、これにより新しい視点が場に加わり、新たな魅力的な情報が次々に生成される魅力的な場になることが期待できる。積極的に書き込みを行なうその分野に興味をもつ利用者は、場の盛り上げに貢献する。よって提示においては、彼らが提示される話題を逐次眺めていても飽きがこないよう、彼らを惹きつける提示法が必要となる。

そこで本稿では掲示板のような場につぎつぎ書き込まれるメッセージの列における利用者の興味をそそる話題や、それら話題のメッセージに即応した意味のある変化を利用者に飽きがこないように提示する話題抽出手法を提案する。

2. 既存の研究

ニュースアーカイブや掲示板やメーリングリストなどの場から話題を抽出する研究は多く行なわれている。主なアプローチ方針として文書指向とラベル指向の２つがある。

文書指向のアプローチでは、ある期間の文書集合が与えられ、その中の文書（記事やメッセージ等）を内容で分類（クラスタリング）して、各グループに対してそのグループの内容を表す文字列をそのグループの中の文字列から選び、分類グループの内容を表す話題としてそのグループにラベルづけする。この手法の目的は分類であるので、グループにつけられた話題の順位づけは通常扱われない。順位づけする場合は、例えば、内容が類似した文書が集まっているグループほどスコアを高くしたり、その他の期間に比べてどれだけその期間に特徴的現れた内容かに着目してスコアリングをする。

余 [1]、山田 [2] らの手法はこのアプローチである。共に話題を利用者が理解しやすい表現となるよう名詞句で表している。余が名詞句である話題を１つのシンボルとして重みを割り当てるのに対し、山田の手法は名詞句である話題をその名詞句を構成する名詞単位で重みづけすることにより、話題に対して部分マッチによる重みづけを可能にしている。

ラベル指向のアプローチでは、まず、文書集合から話題となるラベル候補を抽出し、出現文書数をもとにそれらラベル候補に対して重み付けを行なう。斉藤 [3]、松村 [4] の手法はこのアプローチである。

斉藤らは、メーリングリストのメールからそのコミュニティが触発をうけた話題の抽出を目的としている。メッセージの引用部分から、パターンにマッチする文字列を話題の候補として抽出する。話題の寿命をメッセージにその話題を含む最も古いメッセージから最も新しいメッセージまでの日数とし、話題の候補を寿命が３０日以上のグループと３０日未満のグループの２つに分ける。各グループで出現メッセージ数が多い上位１割をコミュニティが触発を受けた話題とする。

松村らは、掲示板において他のコメントの内容を強く支配するような影響をもつコメント、語（話題）、オピニオンリーダを見つけることを目的とする。話題の重みは、コメントチェーン上のコメント相互が共有する語（話題）の割合にもとづく重みを加算して求めたコメントの重みを共有する語で分配し、再度、コメントチェーン上で加算することにより求める。

3. 提案手法

3.1 アプローチ方針

文書指向のアプローチは、ユーザに出力する話題はグループのできに依存する。生成されるグループは、生成するグループの個数やグループの境界を制御する閾値に依存し、人間の直観にあうようなグループが生成されるように閾値の調整をするのは難しい。グループの作成はテキスト集合の細分化であるので、排他的や階層的にグループを作成するのに適する。一方、グループ同士が非階層でオーバラップするようにグループを作成するのには不向きである。よって、１つのテキストに包含関係にない複数の話題が存在するテキスト集合への適用は向かない。また、異なるグループに対して同じ話題が付与される可能性がある。これはグループ毎に話題候補を作成するためである。

一方、ラベル指向のアプローチでは、ラベル相互を独立に扱うため、１つのテキストに包含関係にない複数の話題が存在するテキスト集合に適する。このアプローチでは重複なく作成された話題候補から話題を選択するため、話題が重複することはない。

掲示板などのコミュニケーションの場では、短いメッセージへの対応が必要である。文書指向のアプローチでは、メッセージ相互の単語の共通性を基準にするので、共通する単語が少ない短いメッセージへの適用は適さない。また、メッセージ内で話題を転換したり、複数のメッセージに対するコメントを１つのメッセージで行なったりするので、包含関係にない複数の話題が存在するメッセージへの対応が必要である。よって本手法では、ラベル指向のアプローチを採用する。

3.2 話題の表現

ユーザに提示する話題は、思わず場に入ってしまうようなユーザの興味を引くものがよい。我々は興味を引く話題の表現条件として、文字列から自分が知っているかどうかをある程度判断できるものがよいと考える。知っ

ているかどうかの判断は、情報が具体的であるほど判断しやすい。そこで、本論では名詞句および固有名詞を話題として抽出する。

名詞句は複数の名詞の組合せであるため、名詞１つで表すより話題が具体化する。また各名詞は既知のものであってもその組合せが未知であれば、意外性のある話題となりうるため、ユーザの興味を引く話題の表現として適する。固有名詞は具体的な対象物を表すので、単体でも具体性をもつと考えることができる。また、新商品など興味の対象となりやすいものを抽出できるようにするためにも、固有名詞を話題として抽出する。具体的には、形態素解析プログラム jtag [5] で得られる主品詞および副品詞情報にもとづき、以下に示す品詞パタンに最長マッチする単語列を話題として抽出する。以下の抽出パタンにおいて、? は直前の表現の 0 または 1 回の繰り返しを、+ は直前の表現の 1 回以上の繰り返しを表す。| は選択を表す。

抽出話題パタン (正規表現)

$$p?(n|N)s?(a?p?(n|N)s?)+|N$$

p**（接頭辞）：** 主品詞が「冠名詞」。

N**（固有名詞）：** 品詞に「固有」を含むもの。ただし、年号を除く。カタカナの連続とアルファベットの連続は固有名詞として扱う。

n**（名詞）：** 固有名詞を除き、主品詞が「名詞」で副品詞が「連用」、「Kana」、「代名詞」、「形容」、「非複合」ではないもの。

s**（接尾辞）：** 主品詞が「名詞接尾辞」か「名詞接尾辞名詞」で、副品詞が「名詞」のもの。

a**（各助詞「の」）：** 主品詞が「各助詞」で文字列が「の」のもの。

3.3 話題のスコアリング手法

既存の研究の話題のスコアリング手法では基本的には話題の発生頻度のみを扱い、話題の発生の間隔を扱っていない。その結果、話題の勢いの強弱変化が扱えないため、意味のある変化に応じた話題のスコアリングができない。

話題の発生の間隔を考慮しないスコアリングでは、同じ頻度で発生した話題はいつ発生しても基本的には同じ扱いである。そのため、頻度情報にもとづき一度大きなスコアを得た話題は長期に渡り提示され続ける傾向となる。その結果、提示される話題は利用者にとってあたり前の話題ばかりになり、利用者の飽きに繋がる。また、間隔のあいた弱い掛け合いや盛り上がり小さいものなのか、間隔の細かい強い掛け合いや盛り上がりの大きいものかを利用者は把握することはできず、場の娯楽性の１つである掛け合いや盛り上がりに即する場へのかけつけを支援することを目的とした利用には不適である。そこで我々は最新のメッセージの近傍での話題の発生の間隔に着目し、話題の勢いを扱える話題のスコアリング手法を提案する。

3.3.1 基本提案スコアリング

ある時点のある事象の勢いは、その時点に向かってより密にその事象が発生しているものほど勢いが強いとみなせる。そこで、我々は各話題に対して、最新のメッセージに向かってその話題が最も密に発生しているメッセージ区間を求め、その区間の話題の密度すなわち勢いをその話題の勢いを表す代表スコアとする。ただし、掛け合い、盛り上がり、流行等、話題の勢いの強弱の変化は複数の利用者が同じ話題について話すことにより発生する。よって話題のあるメッセージと最新メッセージの間の勢いは、そのメッセージ以降に発生したメッセージの列における話題の発生密度とする。本スコアリングの式を以下に示す。

$$Score_t = \max_{m_i \in M} Pressure_{t_{m_i}} \tag{1}$$

$$Pressure_{t_{m_i}} = \frac{C_{t_{m_i}}}{R_{m_i}} \tag{2}$$

$Score_t$: メッセージ列 M の最新メッセージが投稿された時点の話題 t のスコア。

$Pressure_{t_{m_i}}$: メッセージ m_i における話題 t の後続区間における発生密度。話題 t のメッセージ m_i から最新メッセージの間の勢いを示す。$R_{m_i} = 0$ の時は $Pressure_{t_{m_i}} = 0$ とする。

$C_{t_{m_i}}$: メッセージ m_i に後続するメッセージ区間における話題 t を含むメッセージ数。

R_{m_i}: メッセージ m_i に後続するメッセージ数。

$Pressure_{t_{m_i}}$ は、後続メッセージにおける話題 t を含むメッセージの割合である。よって、メッセージ m_i の話題 t の後続メッセージでの支持率ともみなせ、盛り上がりの強さを表すと考えることができる。

スコアリングの例

図１を用いてスコアリングの例を示す。図中の○、△、□は各々、話題A、B、Cを表す。長方形の列はメッセージの列を表し、個々の長方形は１つのメッセージである。メッセージはメッセージの投稿順に図の左手より順に並ぶ。図の左手のメッセージが最も古く、図の右手のメッセージへ行くほど新しいメッセージとなり、最も右手にあるのが最新のメッセージである。長方形内に複数の図形が存在するものは、そのメッセージが複数の話題を含むことを示す。例えば、最新メッセージ m_n は話題Aと話題Bを含む。各図形の下の値は、メッセージ m_i におけ

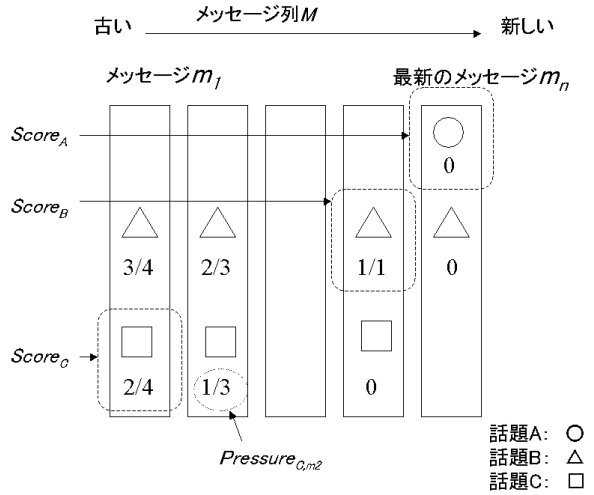

図 1　メッセージのスコアリングの例

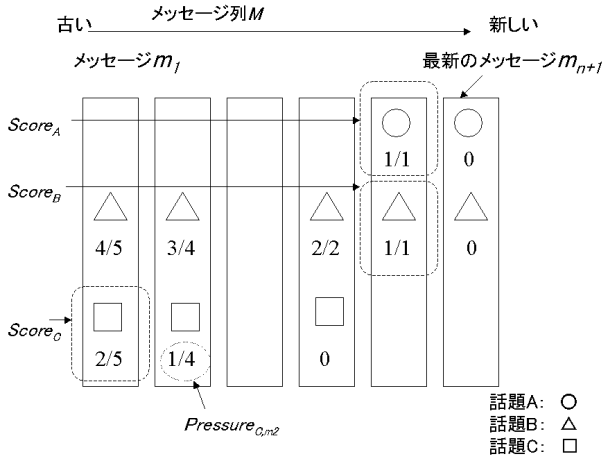

図 2　メッセージ追加後のスコアリングの例

る話題 t_k の後続区間における勢い $Pressure_{t_{m_i}}$ である。話題Aのスコア $Score_A$ は、$max\{0\}$ で $Score_A = 0$ となる。話題Bのスコア $Score_B$ は、$max\{3/4, 2/3, 1/1, 0\}$ で $Score_B = 1/1$ となる。話題Cのスコア $Score_C$ は、$max(\{2/4, 1/3, 0\}$ で $Score_C = 2/4$ となる。よって話題Bが最もスコアが高く、話題A、B、Cは $B > C > A$ の順に順位づけられ、話題Bがもっとも強い勢いをもつとみなされる。

次に最新メッセージ m_n の後に1つメッセージが書き込まれた場合のスコアの変化を説明する。図2は、メッセージ列に1つメッセージが書き込まれ、最新メッセージが m_{n+1} となった図である。各話題のスコアは以下のようになる。話題Aのスコア $Score_A$ は、$max\{1/1, 0\}$ で $Score_A = 1/1$ となる。話題Bのスコア $Score_B$ は、$max\{4/5, 3/4, 2/2, 1/1, 0\}$ で $Score_B = 1/1(= 2/2)$ となる。話題Cのスコア $Score_C$ は、$max\{2/5, 1/4, 0\}$ で $Score_C = 2/5$ となる。よって話題A、B、Cは、$A = B > C$ の順に順位づけられる。ここで注目すべきことは、頻度ではなく出現パタンによりスコアは決まるため、最新のメッセージ近くで密に発生している頻度が最も小さい話題Aにも高いスコアを割り付けることができている点である。

3.3.2　スコアへの意外性の導入

場であたりまえでない話題、すなわち意外性がある話題はユーザの興味を引く。前記、基本スコアリングでは最新メッセージの近傍での勢いを扱うことはできているが、話題の意外性を扱うことができない。そこでユーザの興味を引くものがより高いスコアとなるようスコアリングに意外性を導入する。出現の少ない話題は意外性をもつと考えることができるので、各話題のスコアを出現頻度と反比例の関係になるように式 (2) に出現頻度の逆数を導入する。

$$Score'_t = \max_{m_i \in M} Pressure'_{t_{m_i}} \tag{3}$$

$$Pressure'_{t_{m_i}} = \frac{(C_{t_{m_i}} \times \frac{1}{MF_t})}{R_{m_i}} \tag{4}$$

MF_t は、メッセージ列における話題 t を含むメッセージの数である。新しい話題は当然出現頻度が少ない。よって出現頻度の逆数の導入は新しい話題の優先にも繋がる。

図2において、式5によりスコアを計算した場合、話題A、B、Cのスコアは $Score'_A = 1/2$、$Score'_B = 1/5$、$Score'_C = 2/15$ となり、順位は、$A > B > C$ となる。よって、新しく、意外性があり、最新メッセージの近傍で勢いのある話題Aに最も高いスコアがつく。$\frac{1}{MF_t}$ の導入より、新しい話題の盛り上がりの発生に対してよりスコアが敏感に反応するため、本スコアリング手法は、メッセージ列から、利用者の興味をそそる話題やそれら話題のメッセージに即応した意味のある変化を利用者に飽きがこないように提示できる。

3.3.3　部分マッチへのスコアリングの拡張

掲示板やチャットなどの場では、メッセージの流れを前提としてメッセージの書き込みが行なわれる。そのため話題の一部分が省略されて話されることも多い。例えば、「洗濯機の購入を考えているのですが、洗濯機の乾燥機能は便利でしょうか。」というメッセージに対して「うちのは乾燥機能がついているけど、よく壊れます。便利だけど壊れてばかり、実質使いものにならない。」といったメッセージが書き込まれる場合である。同一文字列の場合に同じ話題とみなす前記提案スコアリング手法では、話題の一部分を省略して話された話題を話題のスコアリングに反映することができない。スコアリングに話題の部分マッチを導入するため、名詞句を構成する単語、具体的には名詞の勢いを用いた以下の式によるスコアリン

グ手法を提案する。

$$Score''_t = \max_{m_i \in M} Pressure''_{t_{m_i}} \quad (5)$$

$$Pressure''_{t_{m_i}} = \frac{\sum_{w \in W_t}(C_{w,m_i} \times \frac{1}{MF_w})}{L_t \times R_{m_i}} \quad (6)$$

$Score''_t$: メッセージ列 M の最新メッセージが投稿された時点の話題 t のスコア。

$Pressure''_{t_{m_i}}$: メッセージ m_i における話題 t の後続区間における発生密度。話題 t のメッセージ m_i から最新メッセージの間の勢いを示す。

W_t: 話題 t に含まれる名詞の集合。

C_{w,m_i}: メッセージ m_i に後続するメッセージ区間において名詞 w が発生したメッセージ数。

MF_w: メッセージ列における名詞 w を含むメッセージの数。

L_t: 話題 t に含まれる名詞の数。

R_{m_i}: メッセージ m_i に後続するメッセージ数。

話題 t のメッセージ m_i から最新メッセージの間の勢いを示す $Pressure''_{t_{m_i}}$ は、話題 t を構成する名詞の勢いの平均値である。平均値を用いているため、少ない名詞数で構成される話題の方が有利である。本稿では固有名詞の場合は１つでも話題として抽出している。固有名詞１つで表される話題を出にくくするには、話題を構成する名詞数が１つの場合は、L_t に定数を加算すればよい。構成する名詞の数が多い話題ほど有利にしたい場合は、名詞の勢いの平均値ではなく、名詞の勢いの合計とすればよい。

3.4 手法の特徴のまとめ

本提案の手法の特徴を以下にまとめる。

- ユーザの興味を引きやすい表現の話題を抽出する。具体的には固有名詞と名詞の連続や名詞間が「の」で繋がる名詞句である。
- 最新メッセージ投稿時の話題の勢いにより話題を順位づけする。その結果、話題に対する掛け合いの発生や盛り上がり、流行の兆しといった意味のある変化のある話題に対して高いスコアをつけることができる。勢いの変化にスコアが敏感に反応するようになっているため、新しい話題の出現や話題の順位の入れ替わりなど、提示される話題が変化に富むので、利用者を飽きさせない。
- メッセージ内の話題の出現頻度を用いず後続メッセージにおいて話題の出現メッセージの割合を利用して話題の勢いを求めるので、メッセージ内で同じ言葉を連呼して場を荒らす記事や、他の利用者が興味を示さない独りよがりなメッセージの影響を受けにくい。
- 時系列のあるテキストから話題を抽出する場合、固定サイズのバッファを設け、バッファ内で話題を抽出して話題の変化を把握する手法が一般的である。それに対し、本手法では最新メッセージからの時間的な距離に応じたスコア計算を行なう。すなわち、ユーザに提示される話題はバッファサイズの影響を受けにくいスコア計算となっている。そのため固定サイズバッファの設定を必要としない。固定サイズのバッファを設けた場合は、処理対象となるメッセージ数を制限できるため処理の高速化が可能である。

図 3 適用システム画面例

4. 適用システム例

図 3 に本手法を利用したプロトタイプのシステムの画面例を示す。このシステムは掲示板の各板（Yahoo(注1)の掲示板のトピク、2 ちゃんねる(注2)のスレッドに相当）の話題の最新の状態を表示する。本例は 2 ちゃんねるの「イタリアについて語るスレッド http://yasai.2ch.net/oversea/kako/972/972181882.html」のメッセージに適用した例である。システムはメッセー

(注1)：http://www.yahoo.co.jp/

(注2)：http://www.2ch.net/2ch.html

ジの書き込みに即して掲示板の各板の話題を勢い順に表示している。利用者は提示される話題を見ることにより、利用者の興味を引くような表現をもつ話題で意外性のあるものや盛り上がりのあるもの、新しいものを次々に知ることができる。プルダウンメニューで対象とする板の切り替えができるため、テレビで面白そうな番組を探すように、面白そうな話題で盛り上がっている掲示板の板を探すことができる。

5. おわりに

本稿ではメッセージの列における利用者の興味をそそる話題やそれら話題のメッセージに即応した意味の変化を、利用者に飽きがこないように提示する話題抽出手法を提案した。本提案手法は様々な特性をもつ。それら特性を定量的に示すことが必要である。現在、以下の観点での評価を検討している。

- 本提案の話題抽出パタンの話題切り出し精度。
- 本提案の話題表現と１つの単語による話題の表現との利用者の興味を引くという点での比較。
- 提示される話題における利用者に有益な話題の割合や、提示される話題を一定期間眺めた場合に取得できる有益な話題の数の既存手法（話題の出現頻度や文書アプローチのクラスタリング手法等）との比較。

これら評価の結果はまとまり次第、追って報告していきたい。

謝　　辞

本研究をすすめるにあたり、話題のスコアの定義における深い議論やプロトタイプの実装をいただいた NTT サイバースペース研究所の牛島浩一氏に深く感謝いたします。

文　　献

[1] 余、石川, コミュニティウェブにおける掲示板からのトピック抽出, FIT（情報科学技術フォーラム）2002, E-17, pp.115-116 (2002).

[2] 山田、金淵、柴田、浦谷, ニュース記事を利用したトピック抽出の検討, 言語処理学会第５回年次大会発表論文集, pp.116-119(1999).

[3] 斉藤、水澤、山本、山口, 話題の自動抽出による電子メールの情報組織化手法, 情報処理学会論文誌, Vol.39, No.10 pp.2907-2913(1998).

[4] 松村、大澤、石塚, テキストによるコミュニケーションにおける影響の普及モデル, 人工知能学会論文誌, Vol.17, No.3 pp.259-267(2002).

[5] T. Fuchi, S. Takagi. Japanese Morphological Analyzer using Word Co-occurence -JTAG, COLING-ACL pp.409-413, 1998.