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

2005－ＦＩ－ 80 (6)
2005－ＮＬ－169 (6)
2005／9／29

ネット上の会話からの話題即時抽出技術の評価について

石井 恵† 井沢味奈子† 片岡 良治†

† 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: †{ishii.megumi, izawa.minako, kataoka.ryoji}@lab.ntt.co.jp

あらまし 本稿では 1 つの会話の場から盛り上がりがあり利用者の興味をひきそうな話題を提示する提案手法の話題取得容易性の性能を示すため、提案手法に対して行った性能比較の結果を報告する。提案手法は、ネット上の複数の会話のそれぞれの場の盛り上がりがあり利用者の興味をひきそうな話題を同時に把握するシステムの基本モジュールとして利用することを目的とする。比較の結果、最新のメッセージの近傍における話題の発生間隔と話題の部分マッチを用いた提案手法は話題の出現数で提示する話題を決定する典型的な手法に比べ、利用者の話題取得容易性が高く、各会話の場に対して少数の話題しか提示しない状況においてもその性能の低下は小さいという優位性があるという結果を得た。

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

Evaluation of the Topic Prompt Extraction Method for a Message Stream

Megumi ISHII†, Minako IZAWA†, and Ryoji KATAOKA†

† NTT Cyber Solutions Labolatories, NTT Corporation
1-1 Hikarinooka Yokosuka-shi, Kanagawa, 239-0847, Japan
E-mail: †{ishii.megumi, izawa.minako, kataoka.ryoji}@lab.ntt.co.jp

Abstract This paper describes our proposed topic extraction method and reports results of comparison between our method and a typical conventional method based on topic frequency. Our method promptly selects attractive topics from a communication channel by using interval of each topic around the latest message in that channel and matching a word to the topic. It works as a module of a system which enables a user to grasp attractive topics on multiple communication channels concurrently. The results show that our method enables a user to collect attractive topics more easily than the conventional method.

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テキストで会話を行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。

インターネット上には会話の場は多数存在する。その中から自分の興味をひきそうな話題を探すためには、利用者は既知である自分のお気に入りの会話の場を 1 つ 1 つ訪問し、メッセージをよんでいかなければならない。同時進行する複数の会話の場がある場合は、各場で次々に発生するメッセージを同時に読むのは集中力を要し、利用者に大きな負担を要する。

我々は利用者が自分の興味をひきそうな話題の容易な取得を可能とするために、メッセージの登録に応じて 1 つの会話の場の規定の数の最新のメッセージから盛り上がりがあり利用者の興味をひきそうな話題を抽出し提示する話題即時抽出手法を提案した [1]。提案手法は図 1 に示すシステムでの利用を想定したものである。図 1 の適用例(注1)ではメッセージの登録が新しい場から順に上から並ぶ。画面上の各会話の場の話題はメッセージの登録に応じて決定される。利用者は画面に表示される話題の移り変わりを適宜、眺めることで自分の興味をひきそうな話題を取得する。このように会話の各場に対して提案手法を適用

(注1)：2 ちゃんねる「ニュース速報+、http://news19.2ch.net/newsplus/のデータを利用

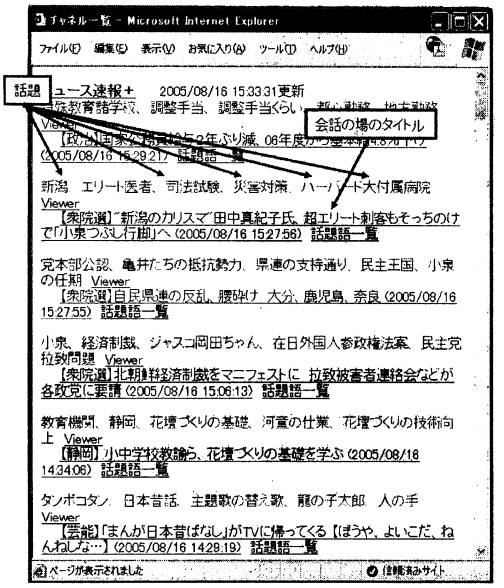

図 1　適用システム画面例

し、各会話の場毎に会話の場と提案手法で提示される話題を組にして利用者に提示することで、利用者は盛り上がりがあり利用者の興味をひきそうな話題を多くの会話の場に対して同時に把握でき、自分の興味をひきそうな話題の探索における負担が軽減される。

本稿では前記提案手法の話題取得容易性の性能を示すため、提案手法と話題の出現数で提示する話題を決定する典型的な手法に対して行った性能比較の結果を報告する。以後、本稿では話題即時抽出手法は前記我々の提案の話題即時抽出手法を示すこととする。以下、2 章では話題即時抽出法を説明し、3 章では比較について述べる。4 章はまとめである。

2. 話題即時抽出手法

本稿では盛り上がりのある話題、および利用者の興味をひきそうな話題を以下のように定義する。

■盛り上がりのある話題: メッセージリストにおいて、そのリスト中のある 2 つのメッセージ間に以下のパタン (1),(2) いずれかが存在する時、話題 α を盛り上がりのある話題と呼ぶ。

(1) 盛り上がりが感じられるパタン 2 つのメッセージ間で後ろのメッセージ (システムへの登録時刻が新しいメッセージ) で前のメッセージ (システムへの登録時刻が古いメッセージ) の話題 α をフォローしている。

(2) 会話の場で興味が高まっているパタン 2 人以上の利用者のメッセージで話題 α に着目している。

■利用者の興味をひきそうな話題: 話題をみて思わず場に入ってしまう話題

話題即時抽出手法では、1 つの会話の場からメッセージの登録に応じて逐次、盛り上がりがあり利用者の興味をひきそうな話題を抽出し提示する。

2.1 話題の表現

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、自分の既知の情報に関する意外性のあ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。よって名詞句および固有名詞を利用者の興味をひきそうな話題として扱う。

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ので、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扱う。

具体的には、形態素解析プログラム jtag [6] 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扱う。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。

2.2 話題のスコアリング手法

提示される話題を逐次眺めることにより興味をひきそうな話題を取得する話題取得方式では、提示される話題の変化は利用者の感覚にあったものであることが好ましい。最も基本的なものは、登録メッセージに盛り上がりのある話題が含まれる場合はその話題の提示順位が上昇することである。盛り上がりのある話題が多数ある場合は、全てを利用者に提示することは現実的ではない。よって提示する話題の高順位における話題の順位変化の制御が課題となる。その際、自分の興味をひきそうな話題を取得しようとしている利用者は、盛り上がりのある話題の中でも既知の話題よりは未知の話題を取得できた方が嬉しいということを考慮することが望まれる。また、会話を楽しむことを目的とする利用者は、提示された話題を見て会話に加わるため、できるだけ最新のメッセージの近傍で盛り上りが密にある話題が提示されることが望ましい。

メッセージが追加されるメッセージリストから話題を求める場合、処理範囲のメッセージを定める window を設け、メッセージの投稿に応じてこの window を 1 メッセージづつずらしてメッセージリストをオーバーラップさせながら連続的に話題を抽出する方法と、一定期間毎にその期間に登録されたメッセージを処理の範囲とし、メッセージのオーバーラップなしで話題を抽出する方法がある。本手法は利用者にメッセージの登録に応じて話題を提示するシステムでの利用を目的とするため、window を 1 メッセージづつずらしてメッセージをオーバラップさせながら連続的に話題を抽出する手法を選択する。

メッセージリスト中の話題に対する既存の話題のスコアリング手法の研究として余 [2]、山田 [3]、斉藤 [4]、松村 [5] らの手法がある。これらの手法では話題の発生数の加算を基本とし、話題の発生の間隔を扱っていない。

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り、利用者の感覚にあった提示とはいえない。さらにこの新しい話題が取得しにくい状況は話題の取得効率面からみても好ましくない。

一方、新しい話題の取得を狙って、頻度の少ないものを提示する手法も考えられる。通常頻度の少ない話題は多数存在する。そのため、提示の優先順位を頻度情報で決定しようとすると同頻度の話題の中では提示の優先順序を設定できず、同じ話題が継続して提示されつづけるという状況を回避する根本的な解決にはならない。

これに対し我々は最新のメッセージを端点とした話題の発生の間隔に着目する。最新のメッセージを端点とした話題の発生の間隔はメッセージの登録毎に変化するため、最新のメッセージの近傍での盛り上がりのある話題の提示を優先させる働きを期待できるからである。

2.2.1 基本提案スコアリング

本スコアリングでは最新のメッセージN件を用いて話題のスコアを求める。盛り上がりのある話題を優先して提示するために、各メッセージの各話題に対して、そのメッセージ以降に登録されたメッセージでどれだけフォローのメッセージが発生したか、また、同じ話題を話すメッセージが発生したかを支持数として利用する。その支持数を最新のメッセージとそのメッセージの発生間隔の長さで割り、支持率とする。各話題に対して支持率の最大値をその話題のスコアとし、スコアの高い上位n件を利用者に提示する話題とする。

発生間隔を用いた支持率の最大値をとることにより、頻度によらず最新のメッセージの近傍で盛り上がりが密にある話題を優先できる。話題のスコアは、最新のメッセージを基準とした話題の支持率の最大値である。そのため、最新のメッセージの近傍における盛り上がりの強さをイメージでき、話題の勢いとみなせる。本稿ではこのスコアを「**勢いスコア**」と呼ぶ。

話題のスコアの式を以下に示す。チャットや掲示板サービスでは、フォロー先のメッセージや投稿者を識別できるものがあるが、本手法ではどのシステムにも共通するメッセージの投稿順序のみを利用して話題のスコアを計算する。本スコアでは利用者を識別しないため、前記盛り上がりのある話題の定義以外の話題に対してもスコアが付与される。

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

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

$Score_t$: メッセージリスト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。
$Support_{t_{m_i}}$: メッセージm_iにおける話題tの支持率。$R_{m_i}=0$の時は$Support_{t_{m_i}}=0$とする。
$C_{t_{m_i}}$: 支持数を示す。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}: メッセージリストMにおいてメッセージm_iに後続するメッセージ数。

スコアリングの例

図2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。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における話題t_kの支持率$Support_{t_{m_i}}$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3は、メッセージリストに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、

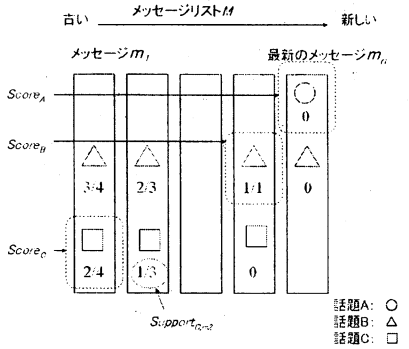

図2 メッセージのスコアリングの例

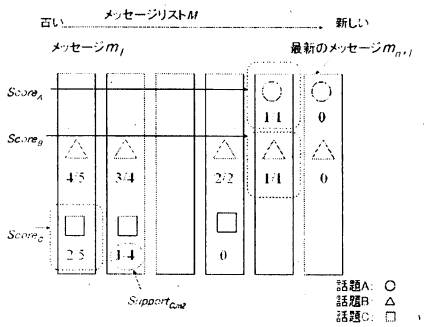

図3 メッセージ追加後のスコアリングの例

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。

2.2.2 スコアへの意外性の導入

意外性がある話題はユーザの興味を引く。前記、勢いスコアでは最新メッセージの近傍での盛り上がりの強さ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きそうな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。新しい話題は出現頻度が少ない。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。本稿ではこのスコアを「**意外性スコア**」と呼ぶ。以下にスコアの式を示す。MF_tは、メッセージリストにおける話題tを含むメッセージの数である。

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

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

図3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を上位に提示できるため、本スコアリング手法はメッセージリストから盛り上がりのある利用者の興味をひきそうな話題の取得を容易にする。

2.2.3 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、具体的には名詞の支持率を用いた以下の式によるスコアリングを行う。本稿ではこのスコアリングを「**部分マッチスコア**」と呼ぶ。

$$Score''_t = \max_{m_i \in M} Support''_{t_{m_i}} \tag{5}$$

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

$Score''_t$: メッセージリスト 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。
$Support''_{t_{m_i}}$: メッセージ m_i における話題 t の支持率。
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: メッセージリストにおける名詞 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 における支持率 $Support''_{t_{m_i}}$ は、話題 t を構成する名詞の意外性により重み付けをした支持率の平均値である。

3. 性能比較

話題即時抽出手法の利用者の話題の取得の容易性の性能を示すため、以下の観点から前記スコアリング手法と既存手法の性能を比較する。

(A) 一定期間における盛り上がりのある話題の提示性能　一定期間に盛り上がりのある話題を多く提示できる手法は利用者の話題の取得容易性が高いといえる。

(B) 盛り上がりのある話題の出現通知性能　提示される話題の変化から盛り上がりのある話題の出現を利用者が把握できる手法は利用者の話題の取得容易性が高いといえる。

(C) 提示する話題数による性能変化　メッセージ登録毎に提示する話題の数が性能に与える影響を調べる。性能差が少ない手法の方が優れた手法と言える。

比較する手法は、(a) メッセージリスト中に現れる話題の出現回数をスコアとする頻度手法 (TF 法)、(b) 勢いスコア手法 ($Proposed_P$ 法)、(c) 意外性スコア手法 ($Proposed_N$ 法)、(d) 部分マッチスコア手法 ($Proposed_S$ 法) である。話題どの手法も 2 章で定義した抽出話題パタンで抽出する。更に話題の表現とスコアリングによる利用者の興味のひきやすさも調査する。

3.1 準　　備

本比較では Yahoo の掲示板サービス(注2)と掲示板サービス「2 ちゃんねる」(注3)から著者が選んだ 4 つの会話の場の先頭 100 メッセージを用いた。各会話の場の基本データを図 4 に示す。Yahoo のデータに関しては、各メッセージに付けられているタイトルもメッセージに含めて扱う。

データセット名	会話の場(タイトル)	URL	平均メッセージ長(byte)	情報源
Vuitton	ルイヴィトンが大好きな方、お話しましょう	http://messages.yahoo.co.jp/bbs?mm=GN&action=1&board=1000002&tid=a5ka5a4a5ta5a3a5ha5sa4actbg9a5a4ada4jjfda1a2a4aaoca47a4dea47a4ga4a6&sid=1000002&mid=1	411.46	Yahoo
Smart	スマートかっちゃいました。	http://messages.yahoo.co.jp/bbs?mm=AU&action=1&board=1834966&tid=a59a5dea1bca5ha4aba4ca4aa4ca4a4a4dea47a4bfa1a3&sid=1834966&mid=1	385.73	Yahoo
Loewe	Loewe★ロエベ	http://yasai.2ch.net/shop/kako/996/996903959.html	196.26	2ちゃんねる
Eunos	今更、ロードスターを買う奴ってどうよ？	http://natto.2ch.net/car/kako/979/979558932.html	116.39	2ちゃんねる

図 4　会話の場の基本データ

[盛り上がりのある話題の正解データ作成]

メッセージリストに対する盛り上がりのある話題を手動で抽出する方法として、メッセージリストのメッセージを全て読んだ後、その中で盛り上がっていると思われる話題を書き出すという方法が考えられる。しかし、数十メッセージを対象とする場合、代表的な話題は記憶に残り書き出せるとしても、直後に発生したフォローメッセージ 1 つで終わってしまうような細かい話題の盛り上がりは記憶にのこりにくく書き出すことは難しい。また、各メッセージ毎にキーワードに相当するものを話題として書き出し、書き出した話題を眺めて盛り上がっていると思われる話題を盛り上がりのある話題として書き出すとう手法も考えられる。しかし、チャットや掲示板ではあるメッセージではキーワードとはならないような話題に対してその話題をフォローするメッセージが発生する場合がある。メッセージ毎にそのメッセージでキーワードとなっている話題を書き出す方法では、フォロー先のメッセージ中における話題が抽出されていないことがあるため、適さない。そこで本稿では以下の手順で与えられたメッセージリスト M に対する盛り上がりのある話題の正解データを作成する。

■盛り上がりのある話題の作成手順

(1) n 個のメッセージからなるメッセージリスト M の各メッセージ $m_i (1 \leqq i \leqq n)$ に対してメッセージ m_i 中の名詞と話題と各話題毎にその話題を構成する名詞を求める。ここで名詞は固有名詞も含む。

(2) 各メッセージ m_i の各話題 w_{is} に対して、そのメッセージより後ろの全てのメッセージ $m_j (i < j \leqq n)$ を対応づけたメッセージ対 $p[m_i, w_{is}, m_j] \in P$ を作成する。

(3) P 中のメッセージ対 p のうち、条件 1 を満たさないものを P から削除する。

(条件 1)m_j に w_{is} を構成する名詞や固有名詞が存在する。

(4) 各評価者は P の各メッセージ対 p に対し後述する**盛り上がり判定手順**に従い、盛り上がり「有」、「無」を付与する。

(5) P の各メッセージ対のうち、過半数の評価者が「有」を付与したメッセージ対 $p[m_i, w_{is}, m_j]$ を話題 w_{is} で盛り上がりをもつメッセージ対とする。盛り上がりをもつメッセージ対に現れる話題 w_{is} を、メッセージリスト M における盛り上がりのある話題の正解データとする。

■盛り上がり判定手順

メッセージ m_i、メッセージ $m_j (i < j)$、話題 w_{is} に対して以下をチェックし、条件 1〜2 のいずれかを満たした場合に「有」、満たさない場合「無」を付与する。盛り上がりの基準がぶれないよう評価者に条件 1、2 を陽に意識させるために、「有」の場合は条件 1〜2 のどちらを満たしたかを参考情報として残す。着目のガイドラインは、発言者が話題 w_{is} をさらりとながすのではなく、(a) 発言者が話題 w_{is} の説明をしていると感じられる部分がある、または、(b) 発言者が話題 w_{is} に対する感想、印象、考え等を述べていると感じられる部分がある、または、(c) 発言者が話題 w_{is} に興味をもていることが感じられる部分があ

(注2)：http://messages.yahoo.co.jp/index.html
(注3)：http://www2.2ch.net/2ch.html

m_iメッセージ番号	m_iタイトル＋本文	話題	m_jメッセージ番号	親番号	m_jタイトル	m_j本文	盛り上がり判定	該当条件
46	"春にレザージャケットをオーダーしました。ナッパのオリーブグリーンで23万くらいか。出来上がるのは11月くらいから、それが待ち遠しいよ。"	レザージャケット	47			ロエベのレザーもの私も欲しいです。(衣類) でも受注会終わっちゃいましたよね。	有	条件1
34	">>33 よく ご存知ですね。その通りです。やれグッチだプラダなど言われる中でロエベのマークをみてちゃんと答えられる人はほんとの意味で通だとお聞きしたことがあります。私は個人的にはシャネルとLVが好きなんですが お母ちゃんがロエベが好きなので、お誕生日か敬老の日に何かプレゼントしてあげようと思っています。本人はキーホルダーが欲しいらしくて 皮ものよりかは金属製のものにしてあげようと思うのですが デザインのお洒落ないいものあったらどなたか教えて下さい。"	ロエベのマーク	47			ロエベのレザーもの私も欲しいです。(衣類) でも受注会終わっちゃいましたよね。	無	

図 5　評価者が評価したメッセージ対の例

る、または、(d) 発言者が話題 w_{is} に思い (話題 w_{is} に対する気持ち) を持っていると感じられる部分がある、または、(e) 上記以外で発言者が話題 w_{is} に着目していると感じられる部分がある、である。話題 w_{is} の文字列そのものがでてこなくてもメッセージ m_j で話題 w_{is} が着目されていると感じられるものは着目されているとみなす。例えば、話題が「黒の鞄」でメッセージ m_j が「私もヴィトンの鞄をもっています。色は黒で大型のものです。」だった場合、メッセージ m_j では「黒の鞄」が着目されているとみなす。

(条件 1) 盛り上がりが感じられるパタン

以下の条件 1-a〜1-c 全てを満たす場合、条件 1 を満たすとする。ただし、メッセージ m_j で話題 w_{is} を受けている部分が、メッセージ m_i とメッセージ m_j で共に話題 w_{is} より長い同じ名詞句となっている場合は盛り上がり有としない。例えば、話題 w_{is} が「ヴィトン」であり、メッセージ m_j で参照しているメッセージ m_i の箇所が「ヴィトンの鞄」、メッセージ m_j でメッセージ m_i を受けている箇所が「ヴィトンの鞄」の場合、「ヴィトン」ではなく、名詞句である「ヴィトンの鞄」で受けているとし、盛り上がりは無とする。

(条件 1-a) 話題 w_{is} は話題として不自然な文字列だったり、意味的に不自然ではない。

(条件 1-b) メッセージ m_j で話題 w_{is} を受けている。

(条件 1-c) メッセージ m_j で話題 w_{is} に着目している。

(条件 2) 会話の場で興味が高まっているパタン

以下の条件 2-a〜2-c 全てを満たす場合、条件 2 を満たすとする。ただし、話題 w_{is} ではなく、話題 w_{is} を含む同じ名詞句がメッセージ m_i とメッセージ m_j で着目されていると解釈した方が自然に感じる場合は盛り上がりが有とはしない。例えば、話題 w_{is} が「ヴィトン」であり、メッセージ m_i での着目が「ヴィトンの鞄」、メッセージ m_j での着目が「ヴィトンの鞄」の場合、着目されているのは「ヴィトンの鞄」とし、「ヴィトン」は着目されていないとし、盛り上がりは無とする。

(条件 2-a) 話題 w_{is} は話題として不自然な文字列だったり、意味的に不自然ではない。

(条件 2-b) メッセージ m_i 話題 w_{is} に着目をしている。

(条件 2-c) メッセージ m_j で話題 w_{is} に着目している。

(4) で評価者が評価したメッセージ対の例を図 5 に示す。本比較では 3 名の評価作業者を雇い、正解データを作成した。

3.2　一定期間における盛り上がりのある話題の提示性能

一定期間に利用者に提示できた盛り上がりのある話題の数を比較するため、メッセージ番号 1 からメッセージ番号 100 までのメッセージを登録した期間において、利用者に提示できた盛り上がりのある話題の数を比較した。利用者に提示する話題数は、提示される話題の特徴をとらえるためには提示する話題数はある程度の数で扱うのがよいと考え、20 とした。window サイズは掲示板サービスが提示する最新のメッセージ件数と同じに設定するのが妥当と考え、代表的な掲示板サービス「2 ちゃんねる」のその件数 50 と同じにした。

メッセージ m_k が登録された時に提示されるべき盛り上がりのある話題の集合 W_k を以下のように定める。データセット毎に各メッセージ $m_k(1 \leqq k \leqq 100)$ に対してメッセージ m_k より前の連続するメッセージ 49 個からなるメッセージ数 50 個のメッセージリスト M_k を作成する。$k < 50$ ではリスト中のメッセージ数は 50 ではなく k 個である。各メッセージリスト M_k に対して、前記「盛り上がりのある話題の作成手順」で作成したメッセージ対 p$[m_i, w_{is}, m_j]$ の i,j が $k - 50 < i, j \leqq k$ である対の話題 w_{is} を M_k に対する盛り上がりのある話題とし、それら話題の重複を除いた和集合を W_k とする。$k - 50 < 1$ の場合は $1 \leqq i, j \leqq k$ として盛り上がりのある話題を定める。

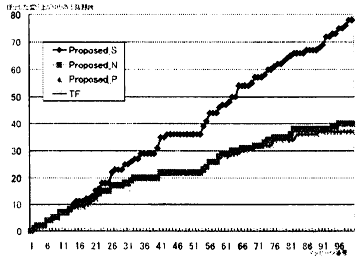

図 6　盛り上がりのある話題の提示累積数 (Vuitton, 提示話題数 20)

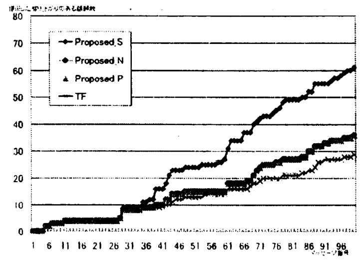

図 7　盛り上がりのある話題の提示累積数 (Smart, 提示話題数 20)

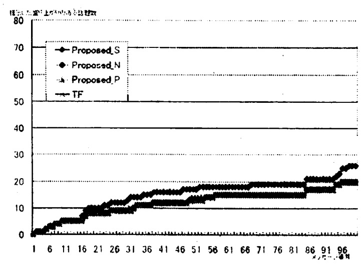

図 8　盛り上がりのある話題の提示累積数 (Loewe, 提示話題数 20)

3.2.1　出現回数と勢いのスコア性能比較

スコアに勢いを用いることによる性能向上を調べるため、TF 法と Proposed_P法を比較した。

図 6〜図 9 はメッセージ m_k が登録された時点で、それまでに各手法が提示した盛り上がりのある話題の累積数を示す。累積数は以下の手順で求めるものである。

メッセージリスト M_k に対して手法が提示した話題のうち、盛り上がりのある話題集合 W_k の話題と一致したものを話題集合 W'_k とする。メッセージ m_k の累積数は、$W'_k(1 \leqq i \leqq k)$ の

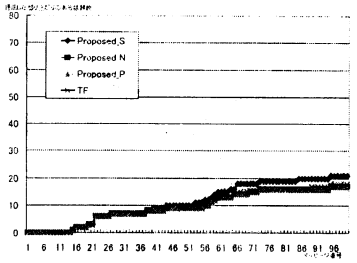

図 9　盛り上がりのある話題の提示累積数 (Eunos, 提示話題数 20)

	Proposed_S	Proposed_P	Proposed_Sのみ提示
Vuitton	78	40	38
Smart	61	35	26
Loewe	26	19	7
Eunos	21	17	4

図 10　部分マッチによる盛り上がりのある話題の提示累積数の比較

要素の重複を除いた和集合の要素数である。

データセット Eunos, Vuitton, Loewe では、提示できた盛り上がりのある話題の数にはほとんど差はないかあっても若干である。以下、提示できた盛り上がりのある話題数に差がでた Smart のデータセットにおいてスコアリングの「勢い」の働きについて述べる。働きとして (a) 盛り上がりのある話題の早期検出、(b) 話題の提示網羅性の向上、の 2 つが得られた。

盛り上がりのある話題の早期検出

メッセージ 47 に最初に出現した「右ハンドル」という話題は、その後連続して 5 メッセージに出現した。TF 法では「右ハンドル」はメッセージ 48 の時点で 8 位に提示され、その後、徐々に順位が上昇して行きメッセージ 57 で首位になった。少しづつ順位が上がっていくため、利用者は話題の動きに気づきにくい。一方、$Proposed_P$ 法ではメッセージ 48 で首位に提示され、メッセージ 59 まで首位を維持した。$Proposed_P$ 法は連続して話される話題の立上りを TF 法よりも早いタイミングで利用者に通知できるといえる。

話題の提示網羅性の向上

一方、会話の中には一瞬の盛り上がりも存在する。例えば、メッセージ 43 で初めて出現した「ドアミラーのミラー」に対してメッセージ 44 で反応があった。この際、$Proposed_P$ 法は「ドアミラーのミラー」に対して高いスコアを割り付けることが可能であるため、「ドアミラーのミラー」は上位の話題とでき、利用者に提示できた。それに対して、TF 法ではスコアが小さいため提示される話題の上位には入らず、利用者に提示されなかった。$Proposed_P$ 法は利用者の一部で一瞬盛り上がる話題の提示も可能であり、利用者に TF 法に比べ盛り上がりのある話題の網羅的な提示が可能と言える。

3.2.2　意外性導入の効果

意外性をスコアリングに導入したことによる性能の向上を調べるため、$Proposed_P$ 法と $Proposed_P$ 法に意外性を導入した $Proposed_N$ 法の性能を比較した。図 6〜図 9 に示すように $Proposed_P$ 法と $Proposed_N$ 法では提示できた盛り上がりのある話題数には差がない。これは 1 度に提示する話題数が 20 と多く、最新のメッセージの近傍で密に現れる話題を全て提示できたためと考えられる。

3.2.3　部分マッチへの拡張の効果

$Proposed_P$ 法を部分マッチのスコアリングへ拡張したことによる性能の向上を調べるため、$Proposed_P$ 法と $Proposed_S$ 法の性能を比較した。図 10 に各手法で提示できた盛り上がりのある話題の累積数を示す。いずれのセットにおいても $Proposed_S$ 法の方が $Proposed_P$ 法より提示した盛り上がりのある話題の累積数が多い。$Proposed_P$ 法のみが提示した盛り上がりのある話題は全てが複数の単語から構成されていた。よって、部分マッチは利用者が取得する盛り上がりのある話題の数を増加させる効果があるといえる。また図 4 の平均メッセージ長と図 10 よりメッセージ長が長い会話の場ほど部分マッチの効果がある傾向にあることがわかる。

	TF	Proposed_P	Proposed_N	Proposed_S
Vuitton	13.61	14.08	13.90	13.24
Smart	11.10	12.73	12.71	12
Loewe	6.47	5.65	5.65	7.35
Eunos	8.41	8.41	8.41	11.86

図 11　スナップショットにおける盛り上がりのある話題の提示数 (提示話題数 20)

3.2.4　ま と め

一定期間に取得できる盛り上がりのある話題が多くても利用者がある一瞬、システムを見た時に得られる盛り上がりのある話題の数が TF 法に比べて著しく少ないようでは問題である。そこで、一定期間だけではなくスナップショットすなわちメセージ投稿毎の盛り上がりのある話題の提示数も比較した。メッセージリスト $M_k(50 \leq k \leq 100)$ に対して各手法が提示した話題と W_k の中で一致した話題の数の平均を図 11 に示す。利用者へ提示する話題数が 20 では TF 法, $Proposed_P$ 法、$Proposed_N$ 法 、$Proposed_S$ 法、どれも大差はなかった。

以上より、話題 20 個を提示する比較的多めの話題を提示する利用においては、スナップショットにおいて提示する盛り上がりのある話題の数は TF 法同程度で一定期間に利用者が取得できる盛り上がりのある話題の数は部分マッチを用いる $Proposed_S$ 法では多く、また部分マッチをもちいない $Proposed_P$ 法、$Proposed_N$ 法でも同程度か若干多く話題を利用者に提示でき、提案手法は TF 法に比べ優位である。

3.3　盛り上がりのある話題の出現通知性能

提示される話題の変化が利用者の感覚にあったものであるかという観点から比較を行った。利用者の感覚にあう話題の変化の最も基本的なものは登録メッセージに盛り上がりのある話題が含まれる場合は、その話題の提示順位が上昇することである。この変化が話題の提示で適切に起これば、その提示法は盛り上がりのある話題の出現の通知性能がよく、利用者の話題取得容易性の性能がよいとみなすことができる。本比較ではメッセージ A に対する盛り上がりのある話題の正解データを以下のようにして作成した。Vuitton, Smart, Loewe, Eunos のデータセット毎に各メッセージ $m_k(1 \leq k \leq 100)$ に対してそのメッセージより前の連続するメッセージ 49 個からなるメッセージ数 50 個のメッセージリスト M_k を作成する。$k < 50$ ではリスト中のメッセージ数は 50 ではなく k 個である。メッセージリスト M_k に対して前記の「盛り上がりのある話題の作成手順」で作成したメッセージ対 $\text{p}[m_i, w_{is}, m_j]$ の i,j が $j = k$ かつ $k - 50 < i \leq k$ となる対の話題 w_{is} からなる集合をメッセージ m_k が登録された時の利用者の感覚にあう変化を示す正解の話題集合 $W_{k_{correct}}$ とする。本比較では手法 A がメッセージリスト M_k で提示した話題 a がメッセージリスト M_{k-1} での提示順位より順位が上昇した場合に、メッセージ m_k の登録により、話題 a に盛り上がりが発生したことを手法 A は利用者に通知したとみなす。

各データセットに対してメッセージ 1〜メッセージ 100 までの登録における各手法の話題の盛り上がりのある話題の出現の通知性能を示す精度と再現率を図 12 に示す。通知性能の精度 Accuracy、再現率 Recall は以下の式により規定される。

		TF	Proposed $_P$	Proposed $_N$	Propsoed $_S$
Vuitton	Accuracy	0.14	0.25	0.24	0.18
	Recall	0.26	0.348	0.29	0.43
Smart	Accuracy	0.09	0.18	0.16	0.18
	Recall	0.19	0.28	0.25	0.43
Loewe	Accuracy	0.03	0.04	0.06	0.09
	Recall	0.14	0.16	0.25	0.50
Eunos	Accuracy	0.07	0.10	0.11	0.15
	Recall	0.30	0.40	0.47	0.65

図 12　盛り上がりのある話題の出現の通知性能比較 (提示話題数 20)

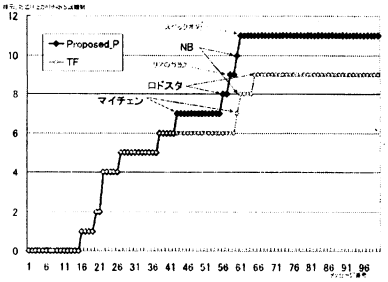

図 13　話題の先行提示 (提示話題数 5)

$$Accuracy = \frac{\sum_{1 \leq k \leq 100} Hit_k}{\sum_{1 \leq k \leq 100} Show_k}$$

$$Recall = \frac{\sum_{1 \leq k \leq 100} Hit_k}{\sum_{1 \leq k \leq 100} Correct_k}$$

Hit_k は、メッセージ m_k 登録時に手法が提示する話題の中でメッセージ m_{k-1} を登録した時より順位の上昇があった話題のうち $W_{k_{correct}}$ と一致した話題の数を示す。$Show_k$ はメッセージ m_k を登録した時に手法が提示した話題の中でメッセージ m_{k-1} 登録時より順位が上昇した話題の数である。$Correct_k$ は $W_{k_{correct}}$ の要素数を表す。

図 12 よりいずれのセットにおいても提案の手法は全て精度、再現率とも TF 法より高い。よって提案手法は盛り上がりのある話題の出現通知性能面で、TF 法より性能がよいといえる。よって提案手法は、話題の変化が利用者の間隔によりあっており、利用者の話題取得容易性がより高いといえる。

3.4　提示する話題数による性能変化

提示する話題数が手法の性能に与える影響を比較した。提示する話題数を 5 として、TF 法と提案手法 $Proposed_P$, $Proposed_N$, $Proposed_S$ について一定期間における盛り上がりのある話題の提示性能を比較した。

図 13 より、勢いによる提案手法 $Proposed_P$ は TF 法より先行して話題を提示できていることがわかる。また図 6〜図 9、図 14〜17 より提案法は TF 方法に対し、提示数を 20 から 5 へ減らした際の提示正解話題数の低下率が低いといえる。これは TF 法が提示数が多い時には提示できていた低頻度の正解の話題が提示数が少ない時には提示できなかったのに対し、提案手法は頻度の影響を受けにくいため、提示数を少なくしても盛り上がりのある話題を提示できたことを示している。

TF 法は話題数の多い Vuitton, Smart ではメッセージ 50 以降、新規の正解話題をほとんど提示できていない。これはメッセージの投稿により、頻度の高い話題が数個できてしまうと他の話題が上位に入りにくい状況が発生していることを示してい

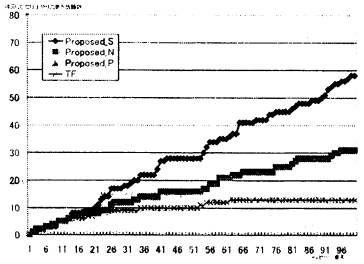

図 14　盛り上がりのある話題の提示累積数 (Vuitton, 提示話題数 5)

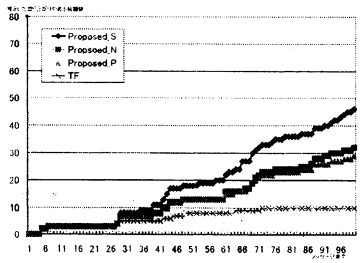

図 15　盛り上がりのある話題の提示累積数 (Smart, 提示話題数 5)

る。このことは利用者に対して多数の話題の変化を知るには下位の話題を見ることを要求する。それに対し提案手法は 話題数の多い Vuitton, Smart でメッセージ 50 以降も新規の正解の話題を提示できている。すなわち利用者は提示される上位の話題を眺めて多数の話題の変化を知ることができ、利用者にとって容易な話題の取得となっていると言える。

しかし、スナップショットにおける正解の話題数は図 18 に示すように上位 5 件個の話題の提示の場合、提案する手法はいずれも TF 法に比べ少なかった。部分マッチを用いる Propsoed$_S$ は盛り上がりのある話題の通知や話題提示累積数において TF 手法や $Propsoed_P$ より優位性をもっている。そのため今後取り組むべき課題の 1 つは、部分マッチを用いる $Proposed_S$ のスナップショットにおける精度向上である。

現在の部分マッチでは、メッセージに出てきた名詞は文脈に関係なく、そのメッセージより古いメッセージ中の名詞句の話題に支持数を加算している。例えば、名詞句内の名詞を他の名詞句の重みづけに利用している。そのため関係ない名詞の重みを加算する場合がある。例えば加算誤りとして「雪道の走り」の「走り」を「走りの問題」へ加算、「日本の気候」の「日本」を「カプリオ日本版」への加算である。日本語の語の構造に着目した支持率の細かい加算制御によりこの問題点を解決し、精度向上を計る予定である。

3.5　提示方式が利用者に与える興味の強さの違い

話題に利用する表現パタンとどの話題を提示するかのスコアリングが利用者に与えるの興味の強さの比較を行った。平成 17 年第 67 回情報処理学会全国大会デモセッションの発表(注4)にて 2 つの手法で生成した話題を表示したシステムの画面キャプチャをみせ、どちらのキャプチャの方が会話の場への興味をひくかを選択してもらうアンケートを実施した。発表では 2 手法で動作するシステムが提示する話題の違いや話題の変化の違いを示

(注4)：デ-10 ネットコミュニケーションの話題の勢いを用いたコンテンツナビゲーションの提案

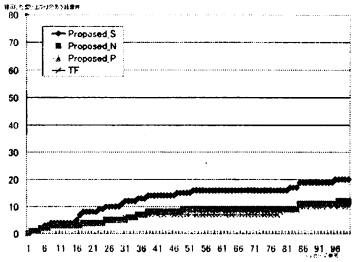

図 16　盛り上がりのある話題の提示累積数 (Loewe, 提示話題数 5)

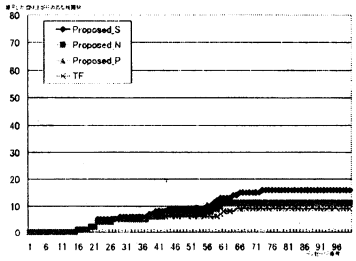

図 17　盛り上がりのある話題の提示累積数 (Eunos, 提示話題数 5)

	TF	Proposed_P	Proposed_N	Proposed_S
Vuitton	5	4.04	4.23	3.55
Smart	4.06	3.69	3.61	3.53
Loewe	2.12	1.96	1.86	1.88
Eunos	4.51	4.35	4.35	3.35

図 18　スナップショットにおける盛り上がりのある話題の提示数 (提示話題数 5)

すデモを行った。先入観が入らないよう、デモを見てもらう前にアンケートを実施した。比較した手法は話題を名詞句と固有名詞で表した Proposed_S と、最も単純な手法と考えられる TF 法のスコアリングで話題を 1 単語の名詞および固有名詞で表した手法である (名詞 TF 法)。

アンケート用紙の例を図 19 に示す。アンケート用紙に提示されている話題はメッセージ番号 1〜メッセージ番号 100 までのメッセージリストから各手法が作成したものである。アンケートは Yahoo と 2 ちゃんねるから著者が選んだ 14 個のデータセットに対して行った。

回収 42 名分のうち全てもれなく回答されている 34 名のデータにおいて、Proposed_P 法は (1) アンケート者の大半が名詞 TF 法より興味をひくと回答、(2) 新規利用者の獲得可能性は名詞 TF 法の 3 倍、という結果を得た。スコアリングが同一ではないので話題に用いた表現形式による違いであるとは厳密には言えないが、名詞句や固有名詞を用いた話題の表現の方が概ね利用者に与える興味は強い傾向がある。

4. おわりに

本稿では多くの会話の場に対して盛り上がりがあり利用者の興味をひきそうな話題を同時に把握するシステムへの適用を目的とした、1 つの会話の場から盛り上がりがあり利用者の興味をひきそうな話題を提示する我々の既提案手法と頻度の多い順

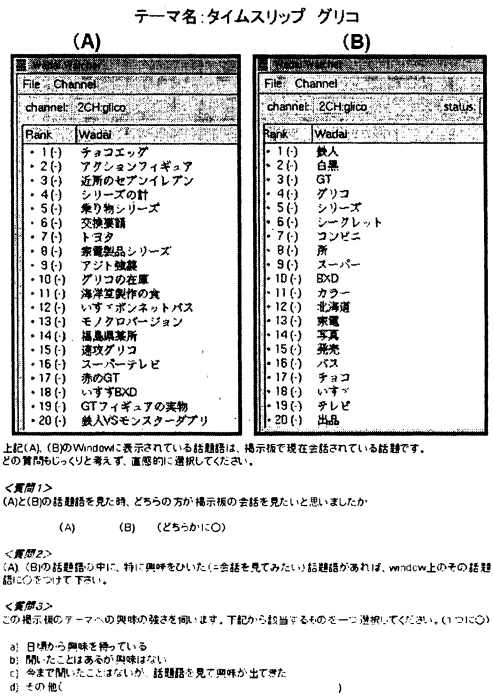

図 19　提示方式が利用者に与える興味の強さ調査の用紙例

に話題を提示する TF 法の性能の比較を行った。比較より、提案手法の方が TF 法に比べ利用者の話題の取得容易性が高く、各会話の場に対して少数の話題しか提示しない状況においても一定期間に提示できる盛り上がりのある話題の数の低下は小さいという優位性をもつことを得た。しかしながら提案手法は提示話題数が少ない場合においてスナップショットで提示される話題に関しては、TF 法よりも性能が低いという問題点が明らかになった。今後の課題はスナップショットで提示される話題の精度向上である。

文　献

[1] 石井、中渡瀬、富田, 名詞句と単語の勢いを用いた話題抽出手法の提案, 情報処理学会, 自然言語処理研究会, 情処研報 2004-NL-160, pp.79-84 (2004).

[2] 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).

[3] 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).

[4] 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).

[5] 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).

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