論 文

Technical Papers

影響の普及モデルに基づくオンラインコミュニティ参加者のプロファイリング

Profiling Participants in Online-Community Based on Influence Diffusion Model

松村 真宏
Naohiro Matsumura
東京大学大学院情報理工学系研究科電子情報学専攻
Graduate School of Information Science and Technology, The University of Tokyo
matumura@kc.t.u-tokyo.ac.jp, http://www.kc.t.u-tokyo.ac.jp/~matumura/

大澤 幸生
Yukio Ohsawa
筑波大学大学院ビジネス科学研究科 / 科学技術振興事業団
GSBS, University of Tsukuba / Japan Science and Technology Corporation
osawa@gssm.otsuka.tsukuba.ac.jp, http://www.gssm.otsuka.tsukuba.ac.jp/staff/osawa/

石塚 満
Mitsuru Ishizuka
東京大学大学院情報理工学系研究科電子情報学専攻
Graduate School of Information Science and Technology, The University of Tokyo
ishizuka@miv.t.u-tokyo.ac.jp, http://www.miv.t.u-tokyo.ac.jp/~ishizuka/

keywords: participant profiling, online-community, influence diffusion model

Summary

Text-based communication in an online-community obscures the characteristics of the participants that aid social interaction. In this paper, we propose a new method for profiling participants in an online-community to help the participants gain a better grasp of their social milieu, i.e., who are the other participant, what are their characteristics, and what are their roles. The proposed algorithm is based on Influence Diffusion Model (IDM), a method for discovering influential comments, opinion leaders, and interesting terms from threaded online discussions. We applied the proposed algorithm to eight electronic message boards, and confirmed higher precision and coverage values than other traditional keyword-based profiling methods.

1. はじめに

近年，電子掲示板・チャット・メーリングリストなどインターネットを介したコミュニケーションツールの浸透により，見知らぬ人たちと情報交換したりディスカッションすることが当り前のように行われている．例えばYahoo!掲示板*1 には数千ものトピックごとに電子掲示板があり，そのトピックに興味のある人たちが自然と集まってコミュニティを形成している．このようなインターネットを介して興味・価値観を共有する人たちの集まりを本論文では**オンラインコミュニティ**と呼ぶ．

オンラインコミュニティにおいて行われるコミュニケーション (Computer-Mediated Communication, CMC) には，実社会における対面でのコミュニケーションにはない大きな特徴として，参加者の匿名性が指摘されている [Kiesler 84]．これは，対面状況だと相手の表情や声のトーン，場の雰囲気など顔が見えることによって伝わる様々な情報によって自然と相手の人物像を捉えることができるが，CMC ではそのような情報は伝わらないので相手の人物像が捉えにくいということである．オンラインコミュニティにおいてしばしば観察される議論が盛り上がらない，もしくは議論の方向が定まらないといった現象はそのようなメディアの特性に依存するところも大きい．このような理由から，オンラインコミュニティにおいてはそのコミュニティが主に扱っている話題や誰がどの話題に詳しいのかといった情報がコミュニケーションを促進するためには重要であることが指摘されている [Gaines 94, 前田 98]，また，コミュニティで交される情報は玉石混交であるので，誰の情報が面白く信用できるのかを予め知ることができれば他の参加者にとって非常に有益となる．会社などの組織においても，社員ごとの特性を知ることができれば，人間関係の維持や発展だけでなく，分散化された専門知識をうまく活用するためのナレッジマネージメント [野中 01] などに活用できる．

では，そのような有益な情報はどうすれば得ることができるのであろうか．ひとつのアプローチとして，面白くない，もしくは間違った情報は淘汰され，本当に面白く信用できる情報はオンラインコミュニティ中に広まることを利用することが考えられる．そこで本論文では，オンラインコミュニティ内でやり取りされたコミュニケーションデータから，参加者に大きな影響を与えた語を参

*1 http://messages.yahoo.co.jp/index.html

加者の特徴を表す**プロファイル**として自動的に抽出する新しい手法について述べる．

本論文では，まず 2 章で関連研究を紹介し，3 章で参加者のプロファイリングを求める本研究のアプローチについて述べる．4 章では本論文で提案するアルゴリズムの元になる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 (Influence Diffusion Model, IDM)[松村 02] について述べ，5 章でコミュニケーションデータから参加者のプロファイルを求める提案手法について述べたあと，6 章で実験による評価について述べる．

2. 関 連 研 究

オンラインコミュニティにおけるコミュニケーションの様子を視覚的に表示すると，参加者のプロファイルを直観的に理解することができる．例えば，電子掲示板やメールソフトに広く採用されているメッセージの返信関係をスレッド表示する機能を使うと，誰が投稿したコメントに返信がたくさんついているのかが一目見てわかる．また，メッセージと参加者との関係を分かりやすく視覚化することにより，オンラインコミュニティにおける参加者の特徴や役割が分かるようにしたシステムもあり，実際のニュースグループやチャットシステム [Donath 94, Viegas 99, Netscan] に採用されている．

また，数理社会学の分野では，人間関係を表すネットワークの構造的な特徴から中心となっている人物を求めようとするネットワーク分析 [Freeman 88, 安田 99] が盛んに研究されており，メーリングリストに適用されている事例もある [金子 96, 北山 97, 高橋 99]．しかし，これらの手法では参加者のプロファイルをキーワードとして抽出していない点が本研究の目的とは異なる．

参加者のプロファイルをキーワードとして獲得する方法には，大きく分けて，質問項目を用意して参加者自身に記入してもらう方法と，自動的に構築する方法の 2 通りある．参加者自身にプロファイルを入力してもらう方法は手軽なために広く用いられている．しかし，参加者にとっては手間がかかるうえ，プライバシーに関わる情報を公開したくはないという気持ちも働くので，あたりさわりのないプロファイルしか得られないことが多い．したがって，このようにして得られたプロファイルでは参加者の特徴を把握することは難しい．

また，プロファイルを自動的に構築する方法として，ユーザが読んだり書いたりしたメールなどのテキストからキーワードを抽出することによってユーザの興味をプロファイルすることや [Foner 97]，参加者が Web 上に公開している情報から参加者のプロファイルを自動抽出することが試みられている [吉田 97]．また，参加者の発言から抽出したキーワードの関係を視覚化することにより，グループディスカッションにおける話題の整理に利用している研究もある [角 97]．我々のアプローチも自動的にプロファイルをキーワードとして獲得することを目指すが，メッセージの返信関係に着目してコミュニティに広く伝搬した語をプロファイルとして抽出する点でこれらの従来研究とは異なる．

3. 参加者のプロファイリング

誰がどの話題に詳しいのかといった参加者ごとの特徴は，オンラインコミュニティに投稿されるコメントを読んでいくうちに自然と明らかになってくるものである．例えば，鉄腕アトムやエンターテイメントロボット AIBO についてよく発言する人ならばロボットに関心があることが推測できるし，ゴミの分別や地球温暖化に詳しい人ならば環境問題に関心のあることが推測できる．

このようにして得られる参加者ごとの特徴は，他の参加者とのコミュニケーションによって引き出されることが多く，参加者自身でさえも意識していなかった潜在的な興味まで表出していることが多い [博報堂 00]．したがって，年齢・性別などのデモグラフィック属性やあらかじめ記述してもらった情報とは質的に異なる．我々は，このようなコミュニケーションの中から得られた特徴こそ参加者のプロファイルにふさわしいと考えている．

オンラインコミュニティではテキストを媒介としたコミュニケーションが行われているので，そこでやり取りされる情報は必然的に文字で表現される．したがって，コミュニケーションにおいて文字すなわち語への興味が伝播していく過程を観察すれば，他のコメントの内容を強く支配するような影響力のある語を参加者ごとに見つけることができると考えられる．そこで本論文では，コミュニティ内に浸透した語に注目して，参加者のプロファイルを自動的に抽出することを狙う．

4. 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の概要

オンラインコミュニティにおいて，ある人の発した語への興味が他の人に伝わっていくプロセスに着目したモデルに，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 (IDM) [松村 02] がある．このモデルは，ある話題が他のコメントの内容を強く支配しているときに，この話題を盛り上がる話題，そのような話題を提供している人をオピニオンリーダ [Katz 55, Rogers 62] とみなす．IDM の詳細は [松村 02] に譲るが，そのアイデアの概要を述べると次のようになる．

掲示板上のコメントは誰でも読むことができるので，厳密に誰がどのコメントから影響を受けたのかを知ることはできない．しかし，一連の返信関係は返信先のコメントについて述べているので，返信関係は参加者間の興味の連鎖を表していると見なすことができる．つまり，ある人 x が投稿したコメント C_x に返信している人 y は，

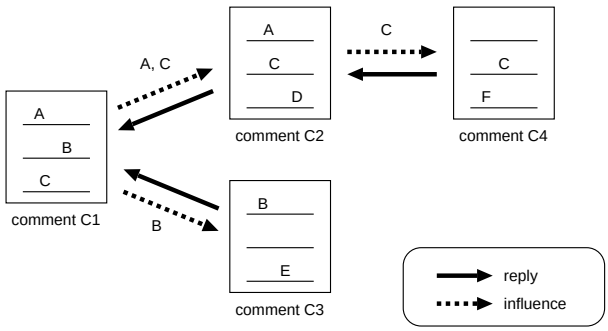

図 1 コメントチェーン.

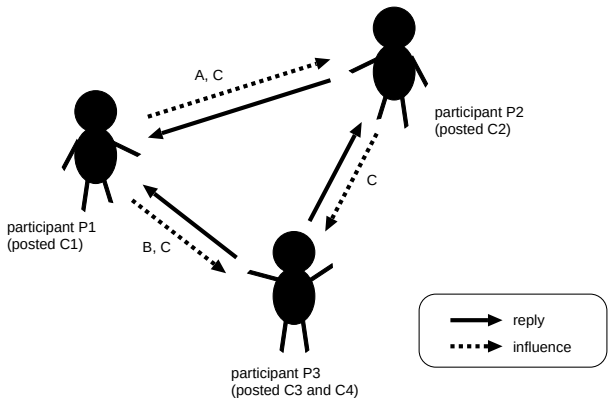

図 2 ヒューマンネットワーク.

C_x を読んで何らかの影響を受けた人である可能性が高い．また，C_x の影響は，伝播先のコメントを介して更にその先にまで，C_x を起点とするコメントチェーン（コメントの返信関係の連なり）上を連鎖的に伝播することになり，このコメントチェーンが長く続けば C_x の影響は広範囲に及ぶと考えられる．

ここで，コメントチェーン上を伝搬していくことにより伝わる影響量を**媒介影響量**と呼ぶことにすると，コメントチェーン上を伝播していく語の割合に着目することにより，C_x が他のコメントに及ぼす媒介影響量を定式化することができる．まず，コメント C_x 中の語の集合を w_x，C_x に返信しているコメント C_y 中の語の集合を w_y とすると，C_x から C_y に伝播した媒介影響量 $i_{x,y}$ は次のようになる．

$$i_{x,y} = \frac{|w_x \cap w_y|}{|w_y|} \tag{1}$$

また，C_y に C_z がさらに返信している場合において，C_x から C_z に伝播した媒介影響量 $i_{x,z}$ は次のようになる．

$$i_{x,z} = \frac{|w_x \cap w_y \cap w_z|}{|w_z|} \cdot i_{x,y} \tag{2}$$

式 (1) は文脈支配の関係を継承する単語の比率で表しており，式 (2) はこの比率をコメントチェーンに沿って掛け合わせていくことを表している．これは，ある単語がある人にとって支配的な文脈になり，その支配的な文脈がさらにその次にメッセージの伝搬する人にとって支配的になっていく比率を求めることを意味している．以上の定式化により，C_x が他のコメントに及ぼす媒介影響量を測ることができるようになる．

4·1 コメントの媒介影響量

例えば，図 1 で表されるコメント C_1,C_2,C_3,C_4 からなるコメントチェーンを考える．C_1 から C_2 には語 A，C が，C_1 から C_3 には語 B が伝播しており，さらに C_2 から C_4 へ語 C が伝播している．このコメントチェーンにおいて，C_1 が他の各コメントに与える媒介影響量は次のようになる．

C_1 から C_2 へ伝わった媒介影響量： C_1 から C_2 へ伝わった語は 2 語であり，C_2 が発している語はこれを含む 3 語であるから，C_1 から C_2 へ伝わった媒介影響量は 2/3 となる．

C_1 から C_3 へ伝わった媒介影響量： C_1 から C_3 へ伝わった語は 1 語であり，C_3 が発している語はこれを含む 2 語であるから，C_1 から C_3 へ伝わった媒介影響量は 1/2 となる．

C_1 から C_4 へ伝わった媒介影響量： C_2 から C_4 へ伝わった語は C_2 が C_1 から受け取った 2 語のうち 1 語であり，また C_1 から C_2 へ伝わった媒介影響量は 2/3 であるから，C_1 から C_4 へ伝わった媒介影響量は $2/3 \cdot 1/2 = 1/3$ となる．

ここで，コメントの媒介影響量を次のように定義する．

【定義 4·1】（コメントの媒介影響量） コメントの媒介影響量は，そのコメントが他のコメントに及ぼした媒介影響量の総和とする．

定義 4·1 により，図 1 のコメントチェーンにおいて C_1 が発した媒介影響量は，(C_1 から C_2 へ伝わった媒介影響量）＋（C_1 から C_3 へ伝わった媒介影響量）＋（C_1 から C_4 へ伝わった媒介影響量）＝ $2/3+1/2+1/3 = 3/2$ となる．同様の手続きにより，C_2 が発した媒介影響量は $1/2$，C_3 が発した媒介影響量は 0，C_4 が発した媒介影響量は 0 と計算される．したがって，図 1 のコメントチェーンにおいて影響力をもつコメントは順に C_1,C_2,C_3,C_4 (C_3,C_4 は同順) となる．

4·2 参加者の媒介影響量

次に，図 1 において C_1，C_2，C_3，C_4 がそれぞれ参加者 P_1，P_2，P_3，P_3 によって投稿されたと仮定すると，参加者間の関係を表すヒューマンネットワークは図 2 のようになる．このヒューマンネットワークにおいて，P_1 が他の参加者に及ぼす媒介影響量は次のようになる．

P_1 から P_2 へ伝わった媒介影響量： P_1 から P_2 に伝わった媒介影響量は C_1 から C_2 に伝わった媒介影

響量と同じになるので 2/3 となる．

P_1 **から** P_3 **へ伝わった媒介影響量：** P_1 から P_3 に伝わった媒介影響量は C_1 から C_3 に伝わった媒介影響量と，C_1 から C_2 を経由して C_4 に伝わった媒介影響量との和になるので，$1/2+2/3\times1/2=5/6$ となる．

ここで，コメントチェーンの場合と同様に，参加者の媒介影響量を以下のように定義する．

【定義 4·2】（参加者の媒介影響量） 参加者の媒介影響量は，その参加者が他の参加者に及ぼした媒介影響量の総和とする．

定義 4·2 により，図 2 のヒューマンネットワークにおける P_1 の媒介影響量は，(P_1 から P_2 へ伝わった媒介影響量）＋（P_1 から P_3 へ伝わった媒介影響量）$=2/3+5/6=3/2$ となる．同様の手続きにより，P_2 の媒介影響量は $1/2$，P_3 の媒介影響量は 0 となる．したがって，図 2 において影響力をもつ参加者は順に P_1,P_2,P_3 となる．

4·3 語の媒介影響量

式 (1), (2) はいずれも，コメント間を伝搬する媒介影響量は伝搬している語の数に比例することを意味している．ここで，全ての語が均等に媒介影響量を伝搬すると仮定すると，語の媒介影響量を以下のように定義できる．

【定義 4·3】（語の媒介影響量） 語の媒介影響量は，その語が伝搬した媒介影響量の総和とする．

定義 4·3 により，図 1 における語 A，B，C，D，E，F の媒介影響量は次のようになる．

A **の媒介影響量：** A は C_1 から C_2 へ伝わった媒介影響量 2/3 を C と共に伝搬しているので，A が伝搬した媒介影響量は $2/3\times1/2=1/3$ となる．

B **の媒介影響量：** B は C_1 から C_3 へ伝わった媒介影響量 1/2 を B だけで伝搬しているので，B が伝搬した媒介影響量は 1/2 となる．

C **の媒介影響量：** C は C_1 から C_2 へ伝わった媒介影響量 2/3 を A と共に伝搬し，C_1 から C_4 へ伝わった媒介影響量 1/3，C_2 から C_4 へ伝わった媒介影響量 1/2 を C だけで伝搬しているので，C が伝搬した媒介影響量は $(2/3\times1/2)+1/3+1/2=7/6$ となる．

D，E，F **の媒介影響量：** D，E，F はどのコメントにも伝搬していないから，D，E，F の媒介影響量はそれぞれ 0 となる．

したがって，図 1 において影響力のある語は順に C，B，A，D，E，F (D，E，F は同順) となる．

5. 提 案 手 法

5·1 プロファイリングのアイデア

4 章で紹介した IDM は，参加者 P_1 が発言した語 A も参加者 P_2 が発言した語 A も同じ語 A と見なして媒介影響量を求めていたが，語がもつ影響力はその語を発言した参加者によって異なると考える方が自然であろう．例えば，人工知能の研究者と言語学の研究者が参加しているオンラインコミュニティにおいて「機能文法」という語が影響力をもつのは，多くの場合，言語学の研究者が発言したときであろう．そのような，影響力があり，かつ参加者を特徴づけるような語は参加者の特徴を表すプロファイルにふさしい．そこで本論文では，参加者のプロファイルを以下のように定義する．

【定義 5·1】（参加者のプロファイル） 参加者ごとに求めた媒介影響量の高い語の集合を，その参加者のプロファイルとする．

図 2 のヒューマンネットワークにおいては，参加者 P_1 が発した語 A は参加者 P_2 に伝搬しているが，参加者 P_2 が発した語 A は誰にも伝搬していないことがわかる．したがって，この場合だと語 A は P_1 のプロファイルになる可能性があるが，P_2 のプロファイルにはならないということになる．ここで仮に媒介影響量が 0 より大きい語をプロファイルと見なすと，P_1,P_2,P_3 のプロファイルは次のようになる．

P_1 **のプロファイル：** P_1 が用いた語は A，B，C の 3 語である．A の媒介影響量は $2/3\times1/2=1/3$，B の媒介影響量は 1/2，C の媒介影響量は $2/3\times1/2+2/3\times1/2=2/3$．したがって，$P_1$ のプロファイルは A, B, C となる．

P_2 **のプロファイル：** P_2 が用いた語は A，C，D の 3 語である．A，D の媒介影響量は 0，C の媒介影響量は 1/2．したがって，P_2 のプロファイルは C となる．

P_3 **のプロファイル：** P_3 が用いた語は C_3 で B，F，C_4 で C，F であるが，いずれの語も媒介影響量は 0．したがって，P_3 のプロファイルとなる語は得られない．

5·2 アルゴリズム

5·1 節で述べたアイデアを定式化したプロファイリングのアルゴリズムは次のようになる．

コメント C_i からコメント C_z に至るコメントチェーンを $\xi_{i,z}=\{C_i,C_j,C_k\cdots C_q,C_r\cdots C_y,C_z\}$ $\{i<j<k\cdots q<r\cdots y<z\}$ とすると，$\xi_{i,z}$ において C_i が C_r に及ぼした影響量 $i_{i,r}$ は式 (1)(2) を拡張した式 (3) で表される．

$$i_{i,r}=\frac{|w_i\cap w_j\cap\cdots\cap w_r|}{|w_r|}\cdot i_{i,q} \tag{3}$$

ここで，w_r は C_r 中の語の集合であり，$\{w_i\cap w_j\cap\cdots\cap w_r\}$ は C_i から C_r まで伝播した語の集合である．

式 (3) は，C_r が発した語のうち C_r が C_i から受け取った語の割合だけ $i_{i,q}$ の影響量が $i_{i,r}$ に伝播することを意味している．ここで，各語が均等に影響を伝播しているとすると，1 語あたりの影響量すなわち $t\in\{w_i\cap w_j\cap\cdots\cap w_r\}$ なる語 t が C_i から C_r へ伝える影響量 $j_{i,r,t}$ は式 (4) で表される．

$$j_{i,r,t}=\frac{1}{|w_i\cap w_j\cap\cdots\cap w_r|}\cdot i_{i,r} \tag{4}$$

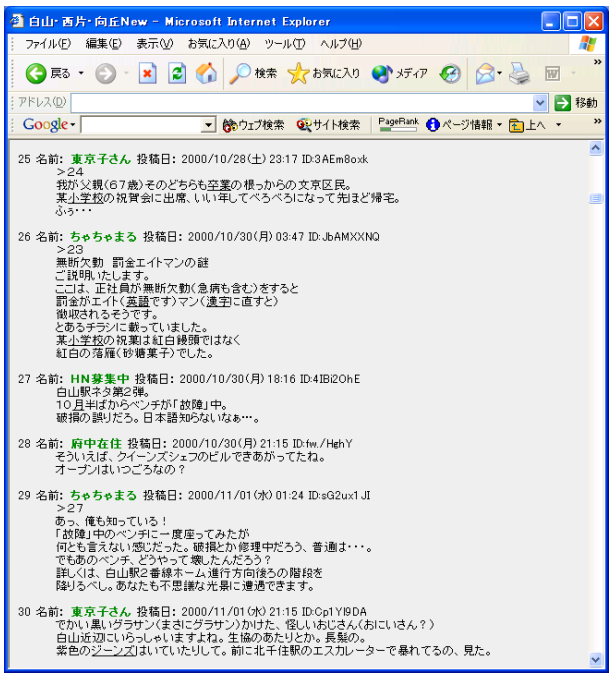

図 3 「白山・西片・向丘スレ」掲示板.

すると，コメントチェーン $\xi_{i,z}$ 上を C_i から C_z まで伝わる t の影響量 $J_{\xi,t}$ は

$$J_{\xi,t} = j_{i,j,t} + j_{i,k,t} + \cdots + j_{i,y,t} + j_{i,z,t} \qquad (5)$$

で表される．ここで，C_i の投稿者を P_x とすると，$J_{\xi,t}$ は C_i において P_x が発した語 t の影響量を表していることになる．

参加者 P_x が発した語 t の媒介影響量 $D_{P_x,t}$ は，P_x が投稿したコメントを起点とした全てのコメントチェーン ξ_{P_x} についての $J_{\xi,t}$ の総和であるから，

$$D_{P_x,t} = \sum_{\xi \in \xi_{P_x}} J_{\xi,t} \qquad (6)$$

となる．つまり，$D_{P_x,t}$ の上位の語 t の集合が P_x を特徴づけるプロファイルとなる．

本論文ではこのアルゴリズムを **PDT 法** (Profiling by Diffusing Terms) と呼ぶことにする．

6. プロファイリング実験とその評価

6·1 実験に用いた電子掲示板

5·2 節で提案した PDT 法の評価を行うために，ある特定の地域について語り合う「まち BBS」掲示板[*2]の中から，白山・西片・向丘に住んでいる多くの参加者によって情報提供や意見交換が活発に繰り広げられている「白山・西片・向丘スレ」掲示板を分析した．図 3 はそのスナップショットである．この掲示板は投稿数が 300 に達すると新しい掲示板を立ててそちらに移行するようになっており，2002 年 9 月 4 日現在では 9 つめの掲示板が立っている．本論文では，8 つめまでの各掲示板を分析した．コメントを投稿するときに名乗るハンドルネームでカウントした投稿者数は平均 41.6 人であった．

なお，この掲示板は自動的に返信がつかない仕様になっており，ユーザはあるコメントに返信する際には「> 75」のように引用符を用いて手動で返信先のコメント番号を指定している．引用符を記さないで返信しているコメントもあるが，これらのコメントの返信関係を調べるためにはコメントの内容から察するしかないので，今回は引用符による返信関係だけを利用した．また，コメントは日本語で書かれているので，茶筌 [茶筌] により形態素にわかち書きし，名詞，未知語，動詞，形容詞，副詞を抜きだして解析を行った．

*2 http://www.machibbs.com/

6·2 比 較 手 法

実験に用いた掲示板は白山，西片，向丘に関する話題を扱っているので，これらの語はよく使われている．また，「行く」「買う」などの一般的な語もよく使われている．しかし，これらの語では参加者間のプロファイルの違いをうまく特徴づけることは難しい．

この問題を解決するために，語が特定のコメントに出現する程度を表す IDF (Inverse Document Frequency) [Sparck-Jones 72] を利用することを考える．IDF は式 (7) で定義される．

$$idf(t) = log\frac{N}{df(t)} + 1 \qquad (7)$$

ここで，N は掲示板中の全コメント数，$df(t)$ は語 t を含むコメントの数であり，$idf(t)$ は語 t が少数のコメントにしか出現しないほど大きな値をとる．

そこで，PDT 法に IDF を考慮した **PDTIDF 法**を以下の式 (8) で定義し，$D'_{P_x,t}$ の上位の語 t の集合を P_x を特徴づけるプロファイルとして PDT 法との比較を行う．

$$D'_{P_x,t} = D_{P_x,t} \times idf(t) \qquad (8)$$

また，従来の代表的な手法として，参加者の発言したコメント集合において出現頻度の高い語をプロファイルとする TF 法 [Luhn 57] と，TF 法に IDF を考慮した TFIDF 法 [Salton 83] との比較も行う．

6·3 実 験 事 例

プロファイリングの例として，「白山・西片・向丘スレ NEW7」掲示板において媒介影響量の最も大きかった参加者の発言の中から，6·2 節で述べた PDT 法，PDTIDF 法，TF 法，TFIDF 法により抽出したキーワードを表 1 に示す．なお，この掲示板の参加者は白山・西片・向丘地域に関する話題に興味をもっているので，参加者の興味はこの地域の名所などに集まりやすい．しかし，表 1 で取り出した語は必ずしもこの地域の名所を取り出すこ

表 1 PDT 法, PDTIDF 法，TF 法，TFIDF 法により抽出されたプロファイルの例．* 印はこの掲示板全体において出現頻度が 100 位以下の語を表している．

順位	PDT	PDTIDF	TF	TFIDF
1	行く	歯科*	白山	前
2	歯科*	近道*	思う	白山
3	ラーメン	千石自慢*	前	言う
4	千石自慢*	ラーメン	近く	ラーメン
5	巣鴨	東洋大	言う	近く
6	東洋大	巣鴨	行く	食う
7	近道*	戸隠*	情報	歯科*
8	白山	そば*	ラーメン	移転
9	思う	京北*	向丘	思う
10	戸隠*	ゴミ*	いい	ケーキ

とを狙っているわけではなく，参加者の興味を惹いた話題に着目していることを付け加えておく．

6･4 節で後述する実験結果によると，この参加者が注目を集めたキーワードは,「歯科」「ラーメン」「千石自慢」「近道」「戸隠」「そば」「京北」「ゴミ」「ケーキ」であった．白山には歯科が多く，筆者を含めて歯科の評判を知りたがっている人は多い．この掲示板でも歯科に関する話題は繰り返し登場しており,「歯科」はそのようなユーザの興味を反映している語であった．また，地元では有名な知る人ぞ知る「千石自慢ラーメン」，白山神社の渡り廊下の下を抜ける意外な「近道」，隠れファンが多い「戸隠そば」，最近名称が白山高校に変わった「京北」学園，コンビニ周辺にちらかった「ゴミ」，おいしくて安い「ケーキ」屋に関する話題など，いずれもこの掲示板で盛り上がった話題であった．

これらの語の出現頻度は「ラーメン」「ケーキ」以外はいずれも 100 位以下 (出現頻度は 8 回以下) であるので，表 1 を見れば明らかなように TF 法や TFIDF 法では取り出すことが難しい語であった．このような出現頻度は低いけれども参加者の興味を捉えた語の多くを PDT 法，PDTIDF 法が取り出せていることは大変興味深い．「ケーキ」は TFIDF 法でしか取り出せていないが，これは「ケーキ」という語の代わりにケーキ屋さんの名称やケーキの種類が使われて返信されたためであった．

なお，PDT 法で得られていた「行く」「白山」は一般的な語でありこの参加者のプロファイルには不適切であるが，PDTIDF 法ではこれらの語の重要度は下がり，代わりに重要な「そば」「京北」「ゴミ」が得られている．また，TF 法では得られなかった重要な「歯科」「ケーキ」が TFIDF 法で得られている．これらの結果は，IDF がプロファイルに有効であることを示唆している．

6･4 実験による評価

評価実験は次のようにして行った．まず，各掲示板から媒介影響量の大きい上位 5 名の参加者について 6･2 節で示した PDT 法，PDTIDF 法，TF 法，TFIDF 法でキーワードをそれぞれ 5 語ずつ抽出した．実験に用いた掲示板は 8 つなので，合計 40 名分のプロファイルを 4 手法で獲得したことになる．次に，それらのキーワードを各参加者ごとに集めてシャッフルし，各キーワードがどの手法により抽出されたのかを分からないようにした．そのようにして参加者ごとに得られたキーワード集合について，各掲示板を読み込んでもらった大学院生 5 名で手分けして，その参加者のプロファイルとして相応しいかどうかを判別してもらった．また，これとは別に各参加者について PDT 法，PDTIDF 法，TF 法，TFIDF 法で 20 語ずつ抽出したキーワードを合わせたキーワード集合の中から，大学院生の回答と照らし合わせてプロファイルの正解集合を人手で作成した．なお，大学院生が回答していないキーワードについては，別途大学院生に手分けして評価してもらった．最後に評価されたキーワードを各手法ごとに再集計し，precision と coverage で評価した．なお，precision は抽出されたキーワードに対する適切だと判断されたキーワードの割合，coverage は正解キーワード集合に対する適切だと判断されたキーワードの割合であり，precision と coverage の値は 1 に近い方が良い結果を意味している．precision と coverage による結果を表 2 に示す．

表 2 precision と coverage による PDT 法，PDTIDF 法，TF 法，TFIDF 法の評価（5 語をプロファイルとしたとき）．

手法	precision	coverage
PDT	0.665	0.507
PDTIDF	0.723	0.551
TF	0.212	0.159
TFIDF	0.394	0.295

表 2 を見ると，明らかに PDT 法，PDTIDF 法は TF 法，TFIDF 法よりもよい結果となっており，語がコミュニティに普及した程度をその人のプロファイルとする提案手法の有効性を示している．また，PDT 法よりも PDTIDF 法，TF 法よりも TFIDF 法の方がよい結果になっており，一般的な語の重要度を下げる IDF が参加者のプロファイリングに有効であることが分かる．

しかし，表 1 を見ても分かるように，PDT 法，PDTIDF 法が抽出するキーワードは，TF 法，TFIDF 法が抽出するキーワードとは大きく異なっている．これは，TF 法，TFIDF 法が PDT 法，PDTIDF 法で取り出せていないキーワードを取り出している可能性を示唆している．

そこで，各参加者について PDTIDF 法，TFIDF 法でそれぞれ抽出した 5 語を合わせた 10 語をその参加者のプロファイルとして再集計した結果を表 3 に示す．また比較のために，PDT 法，PDTIDF 法，TF 法，TFIDF 法で 10 語ずつ取り出してプロファイルとしたときの結果も表 3 に示す．取り出す語数を増やせば coverage が上がるの

はある意味当然だが，10 語取り出してプロファイルとしたときは，単独の手法よりも PDTIDF 法+TFIDF 法の方が precision も coverage も高くなることが興味深い．特に PDTIDF 法+TFIDF 法では，非常に高い coverage が得られている．

表 3 precision と coverage による PDT 法，PDTIDF 法，TF 法，TFIDF 法，PDTIDF 法+TFIDF 法の評価（10 語をプロファイルとしたとき）．

手法	precision	coverage
PDT	0.450	0.648
PDTIDF	0.492	0.678
TF	0.211	0.291
TFIDF	0.348	0.480
PDTIDF + TFIDF	0.565	0.784

6･5 考　　察

人工知能のコミュニティにおいて運営の中心的な役割を果たしている研究者が，町内会の集まりでは宴会係として活躍しているかもしれない．このように人にはさまざまな側面があるので，人の特徴を表すプロファイルは同じ人でもコミュニティによって異なると考えられる．そのような理由から本研究では，参加者の特徴は他者との相互作用によって定まると考え，複雑に交錯するコミュニケーションの中から，多くの参加者の興味を惹いた話題に着目して参加者のプロファイルを行っている．本論文で提案している PDT 法，PDTIDF 法が参加者のプロファイルを的確に獲得できていることは，コミュニケーションにおいて他者との相互作用が人物像の形成に重要な役割を果たしていることを示している．

なお，提案手法はコメント間の明示的な返信関係に基づいているが，返信先のコメントを明示せずに返信しているコメントも少なくないので，提案手法では取り出せない語もある．明示的な返信関係のないコミュニケーションデータにも適用できるように，文脈からコメント間の関連を特定する拡張を施すことは今後の課題である．

また，提案手法では表記に揺れがあると，同じ意味で用いられていても異なる語として認識されてしまう問題が残っている．これについては，類義語辞典や分類語彙辞典などのシソーラスを使うことである程度は対応できると考えているが，今後の課題である．

また，提案手法では他の参加者に全く影響を及ぼしていない参加者のプロファイルは求めることができない．そのような場合でも TF 法や TFIDF 法など他の手法を用いれば参加者のプロファイルは求めることができるが，取り出されるキーワードの性質は異なっている．また，プロファイルとして提示する語の数を増やすとプロファイルの precision や coverage も変化する．したがって，何語くらい取り出すのがプロファイルとして適当であるかは，手法の種類や精度に加えてユーザの理解度に依存する．本稿では，プロファイルの語数を予備実験での結果から経験的に 5 語と 10 語に定めたが，手法や語数によるプロファイルの違いがコミュニケーションにどのような影響を与えるのかを検討することは，今後の課題である．

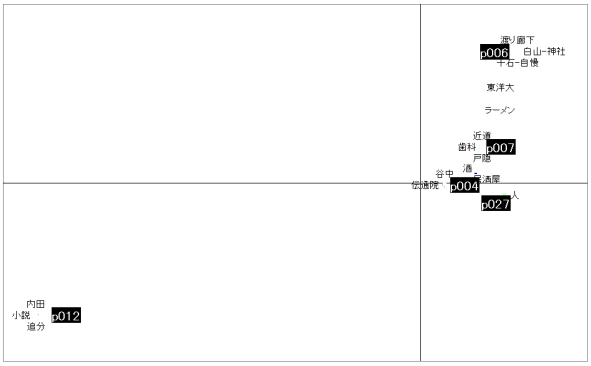

図 4 数量化 III 類による参加者とプロファイルの視覚化の一例．p004, p006, p007, p012, p027 は参加者を表しており，参加者同士の関係や参加者とプロファイルの関係が分かる．

今回はプロファイルの評価を行ったが，ここで得られたプロファイルを利用してオンラインコミュニティにおける参加者間の関係を視覚化し，参加者の役割や興味の分布を知ることも今後の重要な課題である．例えば，6･3 節で紹介した「白山･西片･向丘スレ NEW7」掲示板における媒介影響量の上位 5 名について PDTIDF 法+TFIDF 法で抽出したプロファイルを数量化 III 類 [東京大学 99] により視覚化すると図 4 のようになる．この図を見れば誰がどのような話題に詳しいのか，誰と誰の興味が近いのかを一目見て把握できるようになる．

7. む　　す　　び

本論文では，オンラインコミュニティ内でやり取りされたコミュニケーションデータから参加者のプロファイルを自動的に抽出するための新しい方法について述べ，実験により提案手法が有効であることを確認した．

提案手法はコミュニティの他の参加者に大きな影響を与えた語を参加者のプロファイルとするので，オンラインコミュニティにおいて誰がどの話題に詳しいのかを把握するのに役立つ．そのような情報はオンラインコミュニティを活性化する上で非常に有用であると考えられているので，今後は実際にオンラインコミュニティの参加者に参加者ごとのプロファイルを提示したり，6･5 節で示した参加者間の関係を視覚的に提供するシステムを構築し，その効果を確認していきたい．

◇ 参　考　文　献 ◇

[茶筅]　茶筅，http://chasen.aist-nara.ac.jp/

[Donath 94] J. Donath and N. Robertson: The Sociable Web, *Proc. of World Wide Web Conference*, 1994.
[Foner 97] L. Foner: Yenta: A Multi-Agent Referral-Based Matchmaking System, *Proc. of the First International Conference on Autonomous Agents (Agents 97)*, pp. 301–307, 1997.
[Freeman 88] L. Freeman: Computer Programs in Social Network Analysis, *Commect.*, Vol, 11, pp. 26–31, 1988.
[Gaines 94] B.R. Gaines, M.L.G. Shaw: Using Knowledge Acquisition and Representation Tools to Support Scientific Communities, *AAAI-94,* pp. 707–714, 1994.
[博報堂 00] 博報堂インタラクティブカンパニー：インターネットマーケティング，日本能率マネジメントセンター，2000.
[金子 96] 金子郁容・ＶＣＯＭ編集チーム：「つながり」の大研究，NHK 出版，1996.
[Katz 55] E. Katz, P.F. Lazarsfeld: *Personal Influence*, The Free Press, 1955.
[Kiesler 84] S. Kiesler, J. Siegel, and T. McGuire: Social Psychological Aspects of Computer-Mediated Communications, *American Psychologist*, **39**, pp. 1123–1134, 1984.
[北山 97] 北山聡：フォーラムの生態，編集工学研究所（編）：電緑交響主義，NTT 出版，pp. 34–67, 1997.
[Luhn 57] H.P. Luhn: A Statistical Approach to the Mechanized Encoding and Searching of Literary Information, *IBM Journal of Research and Development*, Vol. 1, No. 4, pp. 309–317, 1957.
[前田 98] 前田晴美，梶原史雄，足立秀和，沢田篤史，武田英明，西田豊明：弱い情報共有を用いたコミュニティの情報共有システム，システム制御情報学会論文誌, Vol. 11, No. 10, pp. 568–575, 1998.
[松村 02] 松村真宏，大澤幸生，石塚満：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.
[Netscan] Netscan Web site. `http://netscan.research.microsoft.com`
[野中 01] 野中郁次郎，梅本勝博：知識管理から知識経営へ－ナレッジマネージメントの最新動向－，人工知能学会誌 Vol. 16, No. 1, pp. 4–14, 2001.
[Rogers 62] E.M. Rogers: *Diffusion of Innovations*, The Free Press, 1962.
[Salton 83] G. Salton and M.J. McGill: Introduction to Modern Information Retrieval, *McGraw-Hill*, 1983.
[Sparck-Jones 72] K. Spark-Jones: A Statistical Interpretation of Term Specificity and Its Application in Retrieval, *Journal of Documentation*, Vol. 28, No. 5, pp. 111–121, 1972.
[角 97] 角康之，西本一志，間瀬健二：共同発想と情報共有を促進する対話支援環境における情報の個人化，電子情報通信学会論文誌, Vol. J80-D-I, No. 7, pp. 542–550, 1997.
[高橋 99] 高橋正道，北山聡，金子郁容：ネットワーク・コミュニティにおける組織アウェアネスの計量と可視化，情報処理学会論文誌 Vol. 40, No. 11, pp. 3988–3999, 1999.
[東京大学 99] 東京大学教養学部統計学教室 (編)：自然科学の統計学，東京大学出版会，1999.
[Viegas 99] F.B. Viegas and J.S. Donath: Chat Circles, *Proc. of CHI'99*, pp. 9–16, 1999.
[安田 99] 安田雪：ネットワーク分析，新曜社，1999.
[吉田 97] 吉田仙，亀井剛次，横尾真，大黒毅，船越要，服部文夫：潜在的なコミュニティの可視化，第 6 回マルチエージェントと協調計算ワークショップ (MACC'97) オンライン予稿集, `http://www.kecl.ntt.co.jp/msrg/macc97/sen.html`, 1997.

〔担当委員：武田英明〕

2002 年 9 月 13 日　受理

著　者　紹　介

松村　真宏（正会員）

1998 年大阪大学基礎工学部システム工学科卒業．2000 年同大学院修士課程修了．2003 年東京大学大学院工学系研究科博士課程修了．博士（工学）．同年より東京大学大学院情報理工学系研究科ポスドク．最近は人間の意思決定のプロセスやオンラインコミュニティにおける集団のダイナミズムに興味がある．情報処理学会，日本グループ・ダイナミックス学会の会員．2001 年，2002 年人工知能学会 MYCOM 優秀プレゼンテーション賞受賞．

大澤　幸生（正会員）

1990 年東京大学工学部電子卒業．1995 年同大学院博士課程修了．博士（工学）．大阪大学基礎工学部助手を経て 1999 年より筑波大学社会工学系助教授，現在に至る．ATR 知能ロボティクス研究所，イリノイ大学客員研究員，科学技術振興事業団さきがけ２１研究員．チャンス発見研究に従事．情報処理学会，AAAI，IEEE などの会員．人工知能学会では 1994 年，1999 年全国大会優秀論文賞，1998 年論文賞受賞．

石塚　　満（正会員）

1971 年東京大学工学部電子卒業．1976 年同大学院博士課程修了．工学博士．同年 NTT 入社，横須賀研究所．1978 年東京大学生産技術研究所助教授．1992 年工学部電子情報工学科教授．2001 年より情報理工学系研究科電子情報学専攻．研究分野は人工知能，知識処理，マルチモーダル擬人化インタフェース/コンテンツ，WWW インテリジェンス．IEEE，AAAI，情報処理学会，電子情報通信学会，映像情報メディア学会，画像電子学会等の会員．