Vol.2011-IFAT-102 No.2
Vol.2011-DD-80 No.2
2011/3/28

ソーシャルブックマークを基にした Twitter ユーザの興味語抽出・推薦手法の提案と評価

齋藤準樹[†1]　湯川高志[†1]

本稿では，ソーシャルブックマークに含まれるタグの共起関係から階層的な類語辞書を作成し，それを用いてユーザの興味語を抽出する手法および興味語の類似性によりユーザ推薦を行う手法を提案する．また提案した手法について，Twitter を対象として興味抽出および推薦の精度と意外性に関する評価実験を実施し，有用性の確認を行った結果についても述べる．

Interest Extraction and User Recommendation based on Social Bookmark

Junki Saito[†1] and Takashi Yukawa[†1]

In the present paper, as a means of extracting user interest for the purpose of user recommendation, the authors propose a method for constructing the hierarchy of words based on SBM tags and to emphasize characteristic word by using this relation. Additionally, the user recommendation system based on this interest extraction is proposed. As a result of a survey on Twitter, the authors discovered that the tags in SBM and their hierarchy have a rich vocabulary for extracting the interests of Twitter users. Moreover, experimental results have indicated that the user recommendation system attains approximately 0.41-0.48 precision if the friend relations in SNS are also utilized as a user preference data.

†1 長岡技術科学大学
Nagaoka University of Technology

1. 序　　論

推薦システムにおいて，ユーザの興味・嗜好・属性をいかに抽出するかということは重要な研究課題である．抽出のための根拠となるデータとして何を用いるか，またそのデータからどのように興味などを推定するのかによって，推薦のパフォーマンスは変化する．

近年では，推薦システムが推薦対象として扱うアイテムはウェブページや商品に限らず，ソーシャルネットワーキングサービス (Social Networking Service, SNS) 内におけるユーザ推薦のように，ユーザ自身へとその範囲を拡大しつつある．しかし，アイテムやウェブページと比較した場合，ユーザはそれらよりも非常に多くの特徴（多様な興味・嗜好）を有しているため，SNS 上の日記・コメントなど（以降「メッセージ」と呼ぶ）から TF·IDF のような従来のキーワードベースの指標で興味語の抽出を試みたとしても，ノイズとなるキーワードが数多く存在し，ユーザがどのような対象や領域に興味を持っているのか検出することが困難である．そのためキーワードベースでのユーザ推薦を行うには，ユーザの興味語について，単語自体の頻度や TF·IDF だけではなく，その単語と関連性がある語まで考慮する範囲を拡大して重み付けを行うような手法が必要となる．

関連性がある語も考慮した単語の重み付け，ということを考えたとき，まず挙げられるのは日本語語彙大系のような類語辞書を用いる方法である．しかし SNS 上などでユーザが投稿するメッセージには頻繁に新語・略語・俗語が含まれるため，従来の類語辞書ではこれらの単語を網羅することが難しい．

そこで本研究では，ソーシャルブックマークサービス (Social BookMarking service, SBM) と呼ばれるウェブサービスで使用されている「タグ」に注目する．SBM において，タグとはユーザが自身のブックマークに付与する短い単語や注釈のことである．この情報には一般的な単語以外に，そのブックマークの内容に関連するものであれば，ブックマークが作成された時点での新語や略語も豊富に含まれている．したがって，その共起情報を基に類語辞書を構築することにより，従来の類語辞書では対応できなかった単語も検出が可能となる．またその語彙によって，ユーザが用いた単語に対し関連語まで考慮に入れた重み付けを行うことで，より興味や嗜好の度合いを強調して特徴的な語を抽出することができると予想される．

以上のことから，本稿では SBM のタグを基に構築した類語辞書を用いてユーザの興味・嗜好・属性をキーワードベースで抽出し，自身と類似するユーザの推薦を行う手法を提案する．また Twitter を対象として行った実験的評価についてもあわせて述べる．

Vol.2011-IFAT-102 No.2
Vol.2011-DD-80 No.2
2011/3/28

2. 背景要素と既存研究

ここでは本研究の背景にある要素として，フォークソノミーと呼ばれる新しい情報分類とSBM の概要，および本研究で実験的評価の対象とする SNS として選定した Twitter について述べる．またそれらに関わる既存研究についてもそれぞれ説明する．

2.1 フォークソノミー

フォークソノミー (folksonomy)[1] とは，情報の受信者（ユーザ）自身が情報の分類を行うボトムアップ型の分類法である．具体的には，情報の分類やグループ化は「タグ」と呼ばれるいくつかの短い単語やフレーズが情報に対して付与されることにより行われる．

従来の分類学 (taxonomy) とは対称的に，フォークソノミーでは単語に上位概念や下位概念を定めないため，情報につけられるタグはただの単語の集合であり，その分類体系はフラットである．またその単語も，予め分類のために規定されたものではない．その結果，ある情報にどのような単語をタグとして付与するのかという判断は，原則として個々のユーザの語彙や価値観に基づいて自由に行われる．

2.2 ソーシャルブックマークサービス

ソーシャルブックマークサービス (Social BookMarking service, SBM) とは，ブックマークへのタグ付け機能を有するオンラインブックマークサービスの一種である．SBM では基本的に各ブックマークの情報[⋆1]が公開され，互いに他のユーザのブックマーク情報を閲覧・共有できるように設定されていることから，フォークソノミーとしての環境が成立する．

各ブックマークにおいてタグは自由に付与されるため，同じページが違ったタグで表現されるということは頻繁に発生する．このようなタグの表記のばらつきがブックマーク数の増加にともなって増え続けると，タグによるそのページの分類は混沌としたものになってしまうという可能性が考えられるが，実際にはそうした事態は起こりにくい．Golder らは delicious [⋆2]のブックマーク情報の分析を通して，特定のページが多くのユーザにブックマークされ続けたとしても，その中で多数派となるタグやその割合が一定の比率に固定化されていく傾向にあることを示した[2]．これは，後にブックマークを行うユーザが，以前にそのページをブックマークした他のユーザのタグの組合わせを模倣することなどが原因であると推定されている．

⋆1 ブックマークされている情報のアドレス，タイトル，タグなどのメタデータ
⋆2 http://www.delicious.com/

2.3 フォークソノミーと情報推薦・情報分類との関わり

フォークソノミーによって作られたメタデータを，情報検索システムや情報推薦システムの構築・既存のアルゴリズムの改善，情報の分類のために用いる研究は，この概念が登場した 2004 年以降，活発に行われるようになった[3),4)]．それらの研究は，フォークソノミーの条件を満たすリソースとして，前述した SBM のブックマーク情報を利用するものがほとんどである．ここでは，日本語を主言語とするユーザが多数派となっている SBM をリソースとして用いたいくつかの研究について，その概要を述べる．

丹羽らは，SBM のブックマーク情報を基に，インターネット全体を対象としたウェブページ推薦システムを提案・構築した[5]．推薦システムの対象範囲が特定のサイト・分野内などに限定されていないという点で，それまでの推薦システムとは異なっている．また丹羽らは，システムを構築する過程において，SBM のタグをページとタグ間で定義した TF·IDF ベースの指標などを用いて抽象化し，SBM のユーザの嗜好をそれらの抽象化したタグで表現した．なお，丹羽らは評価実験を通して，SBM を使用したことがないユーザであっても，彼らが利用しているブラウザのブックマーク情報と SBM 内のブックマーク情報を照らし合わせ，いずれにも共通して存在するブックマークを嗜好データとすることにより，精度 0.4〜0.6 程度のページ推薦が行えることを示した．しかし，ウェブページではなく，被験者の嗜好と近いユーザを推薦対象として選び出すようなタスクについては試されていない．

タグを基にして分類のためのカテゴリを構築する研究には，江田らが行ったものがある[6]．江田らは，PLSI(Probabilistic Latent Semantic Indexing) に基づく folksonomy の索引付け手法をベースに，タグのクラスタリングと is-a 関係の抽出を試みている．またその結果から，類似したタグ同士を集めたグループタグクラウドと自動カテゴリの構築法を提案し，その実行例を示した．しかし，そのカテゴリを用いて実際にウェブページなどの情報を適切に分類できるかどうかということについては実験が行われていないため，カテゴリの有用性が十分に確認されているとは言えない．

2.4 Twitter

Twitter[⋆3]は Twitter, Inc. が運営するマイクロブログサービスの一種である．Twitter ユーザは，140 文字以下の短いメッセージを投稿することによって，自身の現在の状況，ニュースなどに対する意見や感想等を伝えることができる．この短いメッセージは Twitter において「ツイート (tweet)」と呼ばれている．

⋆3 http://twitter.com/

Vol.2011-IFAT-102 No.2
Vol.2011-DD-80 No.2
2011/3/28

標準の設定では各ユーザのツイートは公開されており，そのプロフィールにアクセスすることで閲覧することができる．また，特定のユーザを友人として登録すると，そのユーザの新しいツイートをリアルタイムで読めるようになる．この登録行為は「フォロー」と呼ばれる．なお Twitter では,「あるユーザ “が” フォローしているユーザ」をそのユーザにとっての「フレンド」と呼び，また「あるユーザ “を” フォローしているユーザ」をそのユーザにとっての「フォロワー」と呼んで区別している．

Twitter について，Java らは主に SNS のユーザネットワークの面から分析を行っている[7]．Java らは Twitter のネットワークが他の SNS と同様に，高次の相関性や相互関係を有していることを発見した．またユーザの関心やユーザ間リンクによって推定されるコミュニティ[⋆1]の構造についても考察を行ない，その結果友人関係にあるユーザをグループ化できる可能性を示した．

また，ツイートに含まれる語を基に類似するユーザの発見やツイートの分類を試みる研究には桑原ら[8] や田中ら[9] の研究がある．特に桑原らは，ブログやニュースなどのテキスト内で表現される生活体験に基づいて半自動的に作成したシソーラスを用いることでトピックの抽出と類似ユーザの発見・推薦を行う手法を提案しており，本研究と類似している．しかし，この手法を用いて実際にユーザ推薦を行うまでには至っておらず，有効性の検証が十分になされていない．また，シソーラスの作成過程で一部人手による分類と整理が必要となる点において，SBM のタグから自動的に類語辞書を構築する本研究とは異なる．

3. 提案するシステムの概要

3.1 SBM のタグを利用した類語辞書の構築

本研究では，次のような手順にしたがって，SBM のタグの共起関係から関連語の推定と類語辞書の構築を行う．

(1) **ブックマーク情報からタグに関する情報を抽出**：SBM のブックマーク情報のデータセットから，タグとその頻度，また共起しているタグの組み合わせとその頻度を集計する．どのタグとも共起関係を持っていない孤立したタグについてはここで除外する．

(2) **タグに用いられている単語の関連度を計算**：集計したタグの共起頻度から，各共起ペアの関連度の強さを算出する．この際，関連度の強さを判断するためのスコアとして，MI-score, t-score, G-score(log likelihood[10]) の 3 種類のスコアそれぞれについて計算を行う．これらのスコアは，いずれも語の共起関係の強さを示す指標である．共起関係にあるタグ T_A, T_B について，MI-score は式 (1), t-score は式 (2) によって求められる．また，G-score は**表 1** に示す 2×2 の共起頻度対照表に基づき，式 (3) によって求められる．なお各式において，N は少なくとも一つのタグと共起関係を有するタグの数（SBM に含まれる全種類のタグから，どのタグとも共起関係を持っていなかった孤立したタグの種類数を引いたもの）である．

$$\text{MI-score}(T_A, T_B) = \log_2 \frac{freq(T_A \cap T_B) \times N}{freq(T_A) \times freq(T_B)} \tag{1}$$

$$\text{t-score}(T_A, T_B) = \frac{freq(T_A \cap T_B) - (freq(T_A) \times freq(T_B))/N}{\sqrt{freq(T_A \cap T_B)}} \tag{2}$$

$$\begin{aligned}\text{G-score}(T_A, T_B) &= 2 \sum_{i,j} O_{ij} (\log O_{ij} - \log M_{ij}) \\ &= 2 \Bigg\{ a \log \frac{aN}{(a+b)(a+c)} + b \log \frac{bN}{(a+b)(b+d)} \\ &\quad + c \log \frac{cN}{(a+c)(c+d)} + d \log \frac{dN}{(b+d)(c+d)} \Bigg\}\end{aligned} \tag{3}$$

表 1 G-score の計算に用いる共起頻度対照表
Table 1 Contingency table of co-occurrence frequency

	Tag B(T_B)	¬Tag B($\neg T_B$)
Tag A(T_A)	a	b
¬Tag A($\neg T_A$)	c	d

(3) **上位レベルタグの設定**：それぞれのタグごとに上位レベルタグを設定する.「上位レベルタグ」とは，そのタグと共起している全てのタグの中で最も強い共起関係，すなわち最も高いスコアを有しており，かつ自身よりも多くの種類のタグと共起しているタグのことである．ここでは，より多くの種類のタグと共起しているタグはそうでないタグよりも広い意味で用いられる概念や事物であると仮定している．なお，あるタグ間で互いにもう一方が最も共起関係の強いタグであり，また共起している他のタグの種類数も同じであった場合には，SBM 全体で出現頻度が高かったタグの方を上位レベルタグとする．このような処理を共起関係を持つ全てのタグに対して適用すること

⋆1 一般的な SNS における「コミュニティ」機能とは異なる．Twitter は他の SNS で見られるような明示的なコミュニティを作成・登録する機能を有していない

Vol.2011-IFAT-102 No.2
Vol.2011-DD-80 No.2
2011/3/28

で，タグに用いられている単語の階層的なカテゴリが自動で構築される．

3.2 構築した類語辞書による興味抽出とユーザ推薦

SBM のタグから構築した類語辞書を用いて SNS のユーザの興味をキーワード単位で抽出し，ユーザ推薦を行う手順は以下の通りである．

(1) SNS 上でユーザが書いているメッセージなどを取得する．本研究では，Twitter ユーザの各種ステータスを対象とし，Twitter API を用いてユーザの自己紹介文 (description) やツイートを収集する．

(2) 収集した自己紹介文とツイートに対して形態素解析を行い，名詞を抽出する．形態素解析器には MeCab [*1]を用いる．なお抽出した名詞のうち，SBM のタグとして存在しないものについては除外する．

(3) 自己紹介文中の名詞とツイート中の名詞を，SBM から構築した階層的な類語辞書と単語の出現頻度に基づいて強調する．ある名詞 n の重み w_n は次の式 (4) により求められる．

$$w_n = \sum_{n_{rel} \in RelatedNouns} \frac{freq(n_{rel})}{1 + distance(n, n_{rel})} \tag{4}$$

なお，式 (4) において，各項は次の意味を表したものである．

- $freq(n_{rel})$：自己紹介文中またはツイート中における n_{rel} の出現頻度
- $distance(n, n_{rel})$：類語辞書内の階層構造における n と n_{rel} の距離
- $RelatedNouns$：$distance(n, n_{rel})$ が 3 以下の範囲にある名詞．本手法では，この範囲にある名詞を n の関連語 n_{rel} とみなす

(4) ユーザごとに，メッセージ中の単語の重み w_n を成分とする嗜好ベクトルを次のように定義する．なお，この嗜好ベクトルは，自己紹介文中の単語とツイート中の単語のそれぞれについて別々に生成する．

$\boldsymbol{v}_A = (w_{1A}, w_{2A}, \ldots, w_{NA})$　：ユーザ A の嗜好ベクトル

$\boldsymbol{v}_B = (w_{1B}, w_{2B}, \ldots, w_{NB})$　：ユーザ B の嗜好ベクトル

そして，これらのベクトルの類似度を cosine 類似度を用いて計算し，その上位 20 位までのユーザを推薦候補として被験者に提示する．またここでは比較のために，TF·IDF により単語の重み付けを行った場合の嗜好ベクトルも生成し，これについても類似度上位 20 位のユーザを推薦候補として提示する．

4. 類語辞書の構築・興味抽出手法に対する実験的評価

4.1 評価実験に用いたデータセット

類語辞書の構築と興味抽出手法に対する評価実験に用いたデータセットには，SBM のデータセットと Twitter のデータセットがある．これらについて以下にその概要を述べる．

4.1.1 SBM のデータセット

類語辞書の構築と興味抽出の評価実験では，SBM のデータセットとして，株式会社ライブドアが提供する EDGE Datasets [*2]を使用した．このデータセットには，タグ付けされているブックマークが 1,856,348 件，ユニークタグが 189,258 件含まれている．

4.1.2 Twitter のデータセット

興味抽出の評価実験において，Twitter のデータセットに含まれているユーザの条件は，次の通りである．

- 2010 年 7 月 19 日（データセットの収集を行った日）に少なくとも一回はツイートを投稿している（登録されているがユーザが放置しているようなアカウントではない）こと
- 累計 200 件以上のツイートを投稿していること

この評価実験では，これらの条件を満たす日本語ユーザ 4,161 人の当時における直近 200 件までのツイートを収集し，Twitter のデータセットとして使用した．また，4,161 人のうち自己紹介文 (description) を書いていたユーザ 3,749 人については，これもツイートと同様に収集した．

4.2 実験結果と考察

4.2.1 階層ごとのタグ数分布についての実験結果

表 2 に，それぞれのスコアに基づいて構築した類語辞書におけるタグの階層別頻度分布を示す．いずれのスコアにおいても，最上位の階層（深さ 0）のタグが最も高い頻度となり，非常に多数のカテゴリに分かれている．特に MI-score ではこの傾向が顕著であり，ほとんど全てのタグが最上位の階層とその一つ下の階層に集中している．また階層の深さについても，t-score では最大深さ 8，G-score では最大深さ 7 まで単語の階層関係が構築されたのに対して，MI-score では最大深さ 5 にとどまっている．

次に，属するタグの種類数が上位 50 位までのタグカテゴリについて各スコア間で比較

⋆1 http://mecab.sourceforge.net/

⋆2 http://labs.edge.jp/datasets/, 今回は 2009 年 12 月までのデータを使用

Vol.2011-IFAT-102 No.2
Vol.2011-DD-80 No.2
2011/3/28

表 2 構築した類語辞書におけるタグの階層別頻度分布
Table 2 Frequency of hierarchically structured tags

階層の深さ	MI-score	t-score	G-score
0	93,670	61,441	69,059
1	74,330	36,394	49,691
2	434	37,296	32,192
3	12	22,173	13,400
4	1	8,366	3,414
5	-	2,220	603
6	-	470	85
7	-	82	3
8	-	5	-

した結果を**表 3** に示す．t-score と G-score での共通タグカテゴリ数に対して，それらと MI-score を比較した際の共通タグカテゴリ数の少なさは際立っている．

表 3 スコアごとでの共通タグカテゴリ数（上位 50 位までの比較）
Table 3 The number of common tag categories (top 50)

比較したスコア	共通タグカテゴリ数
MI-score と t-score	13
MI-score と G-score	15
t-score と G-score	40

これらの結果から，MI-score では他の 2 つのスコアに比べて，階層関係の構築とカテゴリの形成がほとんど行われていないと言える．

4.2.2 SBM のタグによる名詞の網羅性についての実験結果

ユーザの興味を抽出するという目的において，ユーザがメッセージ中などに書いた名詞を SBM のタグがどの程度検出できるのかということは，その有効性に直接影響するため，調査される必要がある．ここでは，そのような網羅性を調べるために行った実験結果について述べる．

図 1 は，各ユーザの自己紹介文中またはツイート中に出現した名詞がどの程度 SBM 内のタグと一致するのかについて計算した結果を基に，割合別のユーザ数を集計したものである．なお “Coverage” はユーザごとの自己紹介文中やツイート中にある名詞が，SBM にもタグとして存在している割合である．

Coverage が 0.5 を超えるユーザは，自己紹介文の名詞を対象とした場合は 3,218 人，ツイート中の名詞を対象とした場合は 3,808 人存在した．これらはそれぞれ，自己紹介文を書いているユーザの 85.8 %と，ツイートを書いているユーザ（収集した全ユーザ）の 91.5 %に相当する．またツイート中の名詞よりも，自己紹介文中の名詞の方がタグによる網羅性は高く，8 割以上の名詞がタグと一致したユーザは自己紹介文を書いているユーザの 42.2 %に達した．したがってこれらの結果から，SBM のタグは Twitter 上で実際にユーザが使用する名詞のかなり広範囲を網羅し，その検出が可能であると言える．

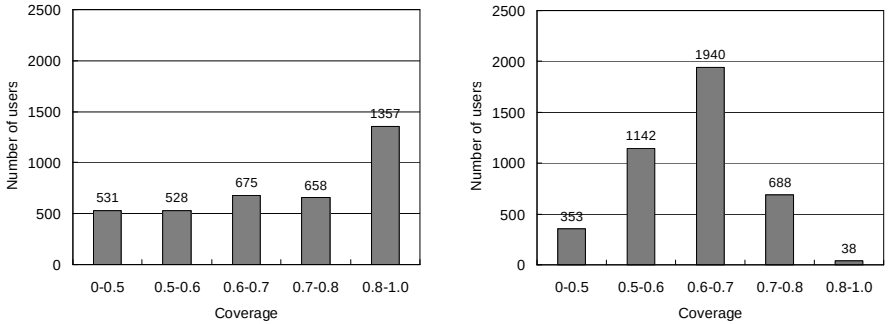

図 1 SBM のタグによる名詞の Coverage 別ユーザ数分布（左：自己紹介文中, 右：ツイート中）
Fig. 1 User distribution with noun coverage by SBM tags (left: description, right: tweet)

4.2.3 興味語の抽出についての実験結果

提案した興味抽出手法において抽出された語が，実際にユーザの興味や属性と合致するものであるかどうかは次のように評価した．

(1) 収集したユーザのうち，bot ではないユーザを 300 人無作為に選出する
(2) 各ユーザの自己紹介文中の名詞について抽出と強調を行い，重み上位 3 位までの名詞がユーザの興味や属性を表しているか人手で確認する

評価のため，本実験では抽出された興味語の精度 (interest precision) p_i を次式で定義した．

$$p_i = \frac{|W_c|}{|W_e|} \tag{5}$$

式 (5) において，W_e は自己紹介文中に含まれる強調された名詞のうち，単語の重みが上位 3 位までの単語の集合，W_c は W_e のうち実際にユーザの興味や属性に関連していた単語の集合である．**表 4**, **表 5** にこの評価結果を示す．

Vol.2011-IFAT-102 No.2
Vol.2011-DD-80 No.2
2011/3/28

表 4 スコアごとの p_i
Table 4 p_i with each score

	MI-score	t-score	G-score
p_i	0.618	0.657	0.628

表 5 p_i が 0.5 以上もしくは 0.8 以上のユーザの割合
Table 5 User ratio whose p_i exceeds 0.5 or 0.8

	MI-score	t-score	G-score
$p_i \geq 0.5$	0.747	0.747	0.757
$p_i \geq 0.8$	0.383	0.420	0.397

提案した手法は非常にシンプルなものであるが，いずれのスコアにおいても 0.6 を超える精度が得られた．またこれらのうち t-score を用いた場合が最も高く，$p_i = 0.657$ となった．MI-score はこの実験で求めた精度では他のスコアとそれほど大きな差がつかなかったが，個々の単語の重みに注目すると，t-score や G-score では幅広い強調が行われたのに対して，MI-score では全く同じ表記の単語が出現しない場合はほとんど重みの強調が行われなかった．したがって，ユーザのツイートを継続的に収集して興味・属性を推定するような場合は，関連語の影響が MI-score よりも反映される t-score や G-score を用いる方が適切であると考えられる．

5. ユーザ推薦に対する実験的評価

5.1 評価実験に用いたデータセット

ユーザ推薦の評価実験に用いたデータセットのうち，類語辞書を構築するための SBM のデータセットは前節と共通のものを使用したが，Twitter のデータセットについては異なるものを用いている．これは次の通りである．

- 2010 年 10 月 12 日（データセットの収集を行った日）に少なくとも一回はツイートを投稿していること
- 累計 1,000 件以上のツイートを投稿していること

この評価実験では，これらの条件を満たす日本語ユーザ 11,104 人の当時における直近 1,000 件までのツイートを収集し，Twitter のデータセットとして使用した．

5.2 評 価 基 準

推薦された候補が適切であるかどうかについては，次のような基準で 10 人の被験者[⋆1]に評価ランクをつけてもらい，それに基づいて Precision を算出した．ここで Precision は，以下の基準での評価結果において評価ランク 1 のユーザ数と評価ランク 2 のユーザ数の合計を推薦された候補数で除したものである．

⋆1 被験者はいずれも日常的に Twitter を利用しているユーザである

(1) ユーザの名前，自己紹介文だけを見て，興味を持った／フォローしてみようと思った（評価ランク 1）
(2) 名前や自己紹介文だけでは十分に興味を持たなかったが，ツイートも読んだところで興味を持った／フォローしてみようと思った（評価ランク 2）
(3) 自己紹介文・ツイート両方を読んでも興味を持たなかった／フォローしてみようとは思わなかった（評価ランク 3）

またこれとは別に，推薦された候補が未知のユーザ・既知のユーザのどちらであるかという Discovery についても評価を行った．この評価基準は次の通りである．

(1) 推薦された候補は未知のユーザであり，自身の自己紹介文やツイートからも予想できないユーザである (Unknown, Unpredictable)
(2) 推薦された候補は未知のユーザであるが，自身の自己紹介文やツイートから予想しうる範囲のユーザである (Unknown, Predictable)
(3) 推薦された候補は既知のユーザ[⋆2]である (Known)

5.3 実験結果と考察

提案手法による語の重み付けに基づくユーザ推薦および TF·IDF による語の重み付けに基づくユーザ推薦での Precision を**表 6** に示す．

自己紹介文から興味を抽出して比較した場合は，提案手法が TF·IDF より 2〜3 倍程度の Precision を達成しており，推薦性能の向上が見られるのに対して，ツイートから興味を抽出して比較した場合は，逆に提案手法が TF·IDF よりも 0.08 低い結果となった．またスコアごとの違いについては，自己紹介文中の名詞を基に推薦した場合，前節の表 4 と同じ順番に並んでおり，t-score で最大 0.31 の Precision が得られた．一方，ツイートの場合の Precision では全てのスコアが 0.23 で横並びとなり，差が現れなかった．

表 6 ユーザ推薦の Precision
Table 6 Precision of user recommendation

	自己紹介文			ツイート		
	MI-score	t-score	G-score	MI-score	t-score	G-score
提案手法	0.25	0.31	0.27	0.23	0.23	0.23
TF·IDF	0.10			0.31		

⋆2 ここで言う「既知のユーザ」とは，現在自分と双方向のフォロー関係を結んでいないがその存在自体は知っているユーザのことである．例えば，有名人のアカウントなどがこれにあたる

Vol.2011-IFAT-102 No.2
Vol.2011-DD-80 No.2
2011/3/28

次に，提案手法による語の重み付けに基づくユーザ推薦および TF·IDF による語の重み付けに基づくユーザ推薦での Discovery の比率を**表 7** に示す．提案手法による結果では，自己紹介文から興味を抽出した場合とツイートから興味を抽出した場合のいずれにおいても，被験者にとって未知でありかつ推薦を予想できないユーザの率が 0.66〜0.69 と推薦候補の 7 割近くを占めた．またツイートから興味を抽出した場合は，未知であるが推薦を予想できるユーザの率が t-score で 0.32, G-score で 0.29, MI-score で 0.27 となり，これも表 4 と同じ順に並んだ．一方 TF·IDF による結果では，自己紹介文を基にした場合で，被験者にとって未知でありかつ推薦を予想できないユーザの率が 0.76 となり，提案手法による結果よりも増加している．しかしツイートを基にした場合はこれが 0.57 と大きく減少し，未知であるが推薦を予想できるユーザの率が 0.39 と 4 割近くに達した．

表 7 ユーザ推薦の Discovery の比率
Table 7 Discovery of user recommendation

			Unknown		Known
			Unpredictable	Predictable	
自己紹介文	提案手法	MI-score	0.66	0.25	0.09
		t-score	0.68	0.25	0.08
		G-score	0.67	0.25	0.08
	TF·IDF		0.76	0.22	0.03
ツイート	提案手法	MI-score	0.69	0.27	0.04
		t-score	0.66	0.32	0.02
		G-score	0.68	0.29	0.04
	TF·IDF		0.57	0.39	0.04

以上の実験結果より，自己紹介文から興味を抽出した場合とツイートから興味を抽出した場合とで，提案手法による推薦と TF·IDF による推薦の Precision がそれぞれ互い違いにもう一方よりも良い結果となった原因について考察する．

まず，自己紹介文から興味を抽出した場合について考える．Twitter での自己紹介文は，公式の指針では 65 文字以内で記入することが推奨されている．また長すぎる自己紹介文は読みにくいため，ほとんどのユーザたちは長くとも 1 ツイートの最大文字数より少ない範囲で自己紹介文を記入している．その影響で，自己紹介文では本人が現在興味や関心を持っている事柄や属性について，名詞を列挙するような形で書かれているものが比較的よく見受けられる（例：[洋楽][Mac][写真][旅行]）．一見すると，ここから名詞を取り出せばそれがそのまま興味語になるため，TF·IDF による重み付けでも問題はないように思える．

しかし，文が短いことから基本的に各名詞の出現頻度は 1 回のみであり，TF での重みには差が現れない．また，タグと同様にこれらの名詞の表記はユーザごとでぶれがあるため，IDF での重みにも差が生じにくい．その結果，TF·IDF では自己紹介文中のどの名詞にも同じような値で重み付けがなされてしまい，被験者との類似度を計算した際，同じ類似度で多数のユーザが横並びになるという現象が見られた．

一方，提案手法では自己紹介文中とツイート中の関連語も考慮するため，よりユーザの関心の強さを反映した重み付けが自己紹介文中の名詞に対して行われる．この場合，自己紹介文中に同じ名詞を有するユーザが複数いたとしても，被験者との類似度計算では類似度に差がつきやすくなることから，より強くその分野や事物に関心を持つユーザを発見することができ，ひいては Precision の増加に寄与したと考えられる．

次に，ツイートから興味を抽出した場合について考える．こちらではユーザごとに直近の 1,000 ツイートを収集したため，出現する名詞の種類・頻度がともに自己紹介文より遥かに多い．したがって，被験者とその他のユーザ間で一致する名詞も多くなり，TF·IDF で重み付けした名詞から生成した嗜好ベクトルの類似度計算においても，ユーザごとで差がつきやすくなったと考えられる．これは，自己紹介文に基づく推薦とツイートに基づく推薦を比較した際，Discovery について「未知であるが推薦を予想できるユーザ」の比率が後者で大きく増加していたことからも推測できる．

一方，提案手法ではツイート中に含まれる名詞のうち，SBM にタグとして登録されていない名詞は類語辞書による重み付けが行えないため，除去している．ツイートから興味を抽出した場合はこれがマイナスに働き，被験者のものと一致する名詞まで除いてしまうことから，結果として TF·IDF よりも Precision がやや劣る推薦結果になったと考えられる．

6. フォロー関係と類語辞書を組み合わせたユーザ推薦への拡張

Precision の向上を図るため，Twitter でのフォロー関係の類似性を基に候補となるユーザを絞り込んだ後，興味抽出を行う形に推薦システムを拡張した．ここでは拡張したユーザ推薦システムの概要と評価結果について述べる．

6.1 評価実験に用いたデータセット

フォロー関係のデータセットは，被験者の「フレンドのフレンド」（2 次のフレンド）1,424,070 ユーザがフォローしているユーザの ID リストを取得・使用した．一方ユーザのメッセージのデータセットには，被験者とその 2 次のフレンドとでそれぞれフォローしているユーザの ID リストの類似性を比較し，上位 1,000 位までのユーザについて自己紹介文

Vol.2011-IFAT-102 No.2
Vol.2011-DD-80 No.2
2011/3/28

と直近 1,000 件のツイートを収集したもの[⋆1]を使用した．なお，ID リストの類似性の比較には cosine 係数, jaccard 係数, dice 係数, 閾値つき simpson 係数[⋆2]をそれぞれ用いた．

6.2 実験結果と考察

拡張したユーザ推薦システムでの Precision を**表 8** に示す．興味語の比較に用いたソース（自己紹介文・ツイート）の違い，スコアの違いにより値に若干の差は生じているが，全てのケースにおいてメッセージのみを基にしたユーザ推薦よりも Precision が大きく向上した．この結果から，拡張した推薦システムでは，友人関係の類似性を一種の嗜好フィルタとして機能させることにより，比較できる興味語がユーザ間で不足するという問題に起因した Precision の低下が改善されている．

表 8 拡張したユーザ推薦の Precision
Table 8 Precision of extended user recommendation

	自己紹介文			ツイート		
	MI-score	t-score	G-score	MI-score	t-score	G-score
拡張前	0.25	0.31	0.27	0.23	0.23	0.23
拡張後	0.44	0.48	0.47	0.41	0.42	0.41

7. 結　　論

本稿では，SNS 上でユーザたちが用いる新語・略語・俗語の抽出も可能にするため，SBM のタグに基づいた類語辞書の構築手法およびそれを用いた興味抽出手法とユーザ推薦手法を提案し，実際の SBM のデータセットおよび Twitter 上でユーザが投稿したメッセージを用いてそれらの手法の有効性を実験的に評価した．

実験結果より，SBM のタグはユーザが実際に使用する名詞の広い範囲を網羅し，興味抽出のためのリソースとして有用であることが確認された．また提案した手法が，自己紹介文のようなごく短い文章で特に興味抽出とユーザ推薦を有効に行えることを示した．さらに，Twitter 上のフォロー関係を用いてユーザの絞り込みを行った後に興味抽出を適用する形にシステムを拡張することで，ユーザ推薦の Precision の向上が可能となることを示した．

今回の提案手法では，ユーザ推薦の Precision が類語辞書による名詞の網羅性に大きく影響を受けるため，今後はより大規模な SBM データセットや Wikipedia など他の言語資源を利用した名詞網羅性および語彙の強化などが課題である．

謝辞 ソーシャルブックマークのデータセットとして，EDGE Datasets を提供していただいた株式会社ライブドア及びライブドア ラボ EDGE の関係者の皆様に感謝いたします．また，形態素解析器には MeCab を利用させていただきました．京都大学情報学研究科－日本電信電話株式会社コミュニケーション科学基礎研究所共同研究ユニットプロジェクトの関係者の皆様に感謝いたします．

参 考 文 献

1) Mathes, A.: Folksonomies - Cooperative Classification and Communication ThroughShared Metadata (online), available from ⟨http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html⟩ (accessed 2011-02-16)
2) Golder, S.A. and Huberman, B.A.: The Structure of Collaborative Tagging Systems, *Journal of Information Science*, Vol.32, No.2, pp.198–208 (2006).
3) Voss, J.: Tagging, Folksonomy & Co - Renaissance of Manual Indexing? (online), available from ⟨http://arxiv.org/abs/cs.IR/0701072⟩ (accessed 2011-02-16)
4) Milicevic, A.K. et al.: Social tagging in recommender systems: a survey of the state-of-the-art and possible extensions, *Artificial Intelligence Review*, Vol.33, No.3, pp. 187–209 (2010).
5) 丹羽智史，土肥拓生，本位田真一：Folksonomy マイニングに基づく Web ページ推薦システム，情報処理学会論文誌，Vol.47, No.5, pp.1382–1392 (2006).
6) 江田毅晴，吉川正俊，山室雅司：Folksonomy のタグを用いた自動分類体系構築へ向けて，情報処理学会研究報告，2007-DBS-143, pp.405–410 (2007).
7) Java, A. et al.: Why we twitter : understanding microblogging usage and communities, *WebKDD/SNA-KDD '07 Proceedings*, San Jose, California, ACM, pp.56–65 (2007).
8) 桑原雄，稲垣陽一，草野奉章，中島伸介，張建偉：マイクロブログを対象としたユーザ特性分析に基づく類似ユーザの発見および推薦方式，情報処理学会研究報告，Vol.2009-DBS-149, No.18, pp.1–3 (2009).
9) 田中淳史，田島敬史：twitter のツイートに関する分類手法の提案，第 2 回 データ工学と情報マネジメントに関するフォーラム (DEIM Forum 2010), A5-4 (2010).
10) Dunning, T.: Accurate Methods for the Statistics of Surprise and Coincidence, *Computational Linguistics*, Vol.19, No.1, pp.61–74 (1993).

⋆1 収集日は 2011 年 1 月 22 日
⋆2 少なくとも 10 ユーザ以上をフォローしているユーザを対象とした