RD-004

Web ページへのタグ付けによる類似ユーザ群を利用した意味情報の抽出

Extracting Ontology from Tags on Web Pages Using a Similar User Group

伊藤 真也 † 小河 真之 † 原田 史子 ‡ 島川 博光 ‡
Masaya Ito Masayuki Ogawa Fumiko Harada Hiromitsu Shimakawa

1. はじめに

膨大な Web ページの中から有益なものを探す手段として検索エンジンがある．一般の検索エンジンは，ユーザが入力するキーワードと Web ページに含まれる文字列の一致に基づいて検索する．同一キーワードであっても，ユーザによってその言葉の使い方は異なる．キーワードの個人にとっての使い方を考慮できれば，よりユーザに有益な情報を検索結果として提示できる．

個人の言葉の使い方を抽出する方法としてはてなブックマーク [5] や livedoor クリップ [6] ソーシャルブックマーク (SBM) の利用が考えられる．SBM では，ユーザはブックマークの整理にタグと呼ばれる自由記述のキーワードを用いる．あるユーザおよび，そのユーザに似た嗜好をもつユーザ群のブックマークに対するタグ付与には，個人の言葉の使い方が表れる．本論文では，嗜好の類似するユーザ群内の SBM へのタグ付与から個人の言葉の階層関係と類義関係を抽出する手法を提案する．

2. 研究背景

2.1 個人の言葉の使い方を考慮した検索エンジン

一般的な検索エンジンは，ユーザが入力するキーワードと Web ページに含まれる文字列の一致に基づいて検索する．しかしながら，入力されたキーワードのユーザにとっての使い方を一般的な検索エンジンは考慮していない．ユーザは，同一のキーワードであっても辞書に含まれない暗黙的な意味を付与している場合がある．また，同一のキーワードであっても，ユーザによってその指し示す範囲は異なることがある．例えば，“情報推薦”というキーワードを考える．“情報推薦”という言葉は一般的には，検索エンジンなどの pull 型システムやポップアップ広告などの push 型システムを含む．ここで，ユーザ A が “情報推薦”というキーワードに “所属研究チームの研究内容”という意味を暗黙的に付与していたとする．検索エンジンは，ユーザ A の所属チームについての情報を優先的に検索結果として提示できない．またユーザ B が “情報推薦”というキーワードを pull 型システムのみを指し示す言葉として使用していたとする．検索エンジンは，ユーザ B の望まない push 型システムについてのページも提示してしまう．検索エンジンが個人の言葉の使い方を考慮できれば，よりユーザに有益な情報を提示できる．

個人の言葉の使い方を考慮した検索エンジンを作るためには，まず個人の言葉の使い方を把握し，計算機で処理できる形式で表現する必要がある．表現方法として知識や用語を体系化するのに用いられるオントロジが適用できる．文献 [1] では，オントロジの自動構築にあたり広義語，狭義語，同義語に着目している．本論文では，言葉の広義語，狭義語の関係を言葉の上位，下位の関係と定義する．また同義語の関係を類義関係と定義する．個人の言葉の使い方を表現するためには，言葉の階層関係と類義関係を表現するオントロジが必要である．

2.2 SBM を利用した言葉の意味情報の抽出

個人の言葉の使い方を抽出する方法として SBM の利用が考えられる．SBM におけるタグ間の関係を解析することで言葉の使い方を抽出できる．SBM におけるタグ間の関係を抽出する研究がいくつか存在する [2][3]．文献 [2] では，ユーザ，リソース，タグの三組のデータを Probabilistic Latent Semantic Indexing を用いて計算し，SBM で使用されるタグ同士の関係を木構造で表現している．ある個人の付与したタグのみに着目すれば，個人の言葉の階層関係を抽出できると考えられる．

タグは，自身のブックマークの整理のために付与するので，個人が付与したタグ群には，類義関係が存在し難い [3]．暗黙的な意味の付与や言葉の指し示す範囲の変化は，人と言葉や情報をやりとりし合うことで起こる．やりとりは嗜好が似通っている人同士ほど頻繁である．そのため，嗜好が似た人同士は，似た言葉の使い方をすると考えられる．嗜好が似たユーザ群から，その個人が使用する言葉と類義語である言葉を抽出することで，言葉の類義関係を抽出することが期待できる．嗜好が似たユーザ群は，例えば，文献 [4] で提案されている方法を用いて SBM から抽出できる．そこで，嗜好が似たユーザ群の SBM におけるタグ付与から個人の言葉の意味情報を抽出する手法を提案する．言葉の意味情報が抽出できれば，個人の言葉の使い方を反映した検索エンジンが実現できる．ここで類似ユーザ群とは嗜好が似たユーザ群，言葉の意味情報とは，あるユーザの言葉同士の階層関係および類義関係と定義する．以後，言葉の意味情報を，意味情報と略記する．

3. タグに基づく特定ユーザの意味情報の抽出

3.1 Web ページ集合からのタグラベル間の関係づけ

本論文では，個人の言葉の使い方を考慮した検索エンジンを実現するために，ある類似ユーザ群内での SBM へのタグ付与に基づいて，類似ユーザ群に所属する特定ユーザの意味情報を抽出する手法を提案する．

本手法では，ユーザがブックマークに付与したタグを用いて，個人が使用する言葉を収集する．ユーザが Web ページにタグを付与することで，図 1 のように，各タグごとに，そのタグが付与された Web ページの集合ができる．各タグに対し，そのタグの文字列をラベル，そのタグが付与された Web ページの集合をクラスタと呼ぶことにする．SBM では，同じラベルでも，異なる意味で付与されたタグが存在する．そのため，本論文では，タグのラベルではなく，各タグによってできるクラスタが，そのユーザにとっての各ラベルの言葉の意味を示す

†立命館大学大学院理工学研究科
‡立命館大学情報理工学部

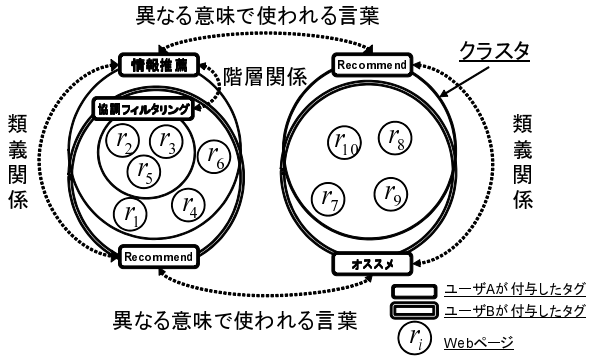

図 1: クラスタとラベル間の関係の例

と考える．

2 つの言葉の類義関係は，それらの言葉がラベルである 2 つのタグのクラスタが似ていることで判断できる．本論文では，たとえ異なるユーザがタグに同じラベルをつけたとしても，各クラスタが似通っていなければ，異なる言葉として扱う．一方，2 つのタグが持つクラスタが似通っていれば，異なるユーザによって異なるラベルを付けられたタグでも近い意味を持つ言葉として捉える．

言葉同士の階層関係についても，1 タグのもつクラスタに基づいて判断できる．ある 2 つの関連ある言葉について階層関係を考えた場合，上位の言葉は広義な言葉であり，下位の言葉は狭義な言葉である．さらに，関連ある 2 つの言葉であれば，同一の Web ページにタグラベルとして付与される可能性が高い [3]．広義な言葉であれば，そのタグが付与されるクラスタの大きさは比較的大きくなると考えられる．反対に狭義な言葉であれば，タグとして付与できる Web ページは比較的少なくなり，クラスタは比較的小さくなると考えられる．

3.2 類似ユーザ群に属する特定ユーザの意味情報抽出

3.1 節の議論に基づき，類似ユーザ群に所属する特定ユーザの意味情報を抽出する手法を提案する．本手法では，ユーザが使用するタグ一覧とそれぞれのタグのクラスタをまとめた情報をそのユーザのタグ情報と呼ぶ．図 2 に提案手法の概念図を示す．図 2 において，例えば，タグ D2 とは，ユーザ D が付与した 2 番目のタグであることを表す．手法の手順は (1) 各ユーザの Web ページへのタグ付け，(2) 各ユーザのタグ情報の抽出，(3) 特定ユーザのタグのラベル間の階層関係の抽出，(4) 特定ユーザのタグのラベルと類義関係にあるラベルの類似ユーザ群内で使用されているタグからの抽出，となる．これにより類似ユーザ群に所属する特定ユーザの意味情報を抽出する．なお，手順 (3) は 3.3 章，手順 (4) は 3.4 章で詳述する．

3.3 タグ間の階層関係の抽出

特定ユーザの付与した任意の 2 タグのラベルについて，階層関係を同定する．本節では，階層関係の同定手法を提案する．ある類似ユーザ群 Sim に n 人のユーザが所属するとき，ユーザ群を $\{u_1, u_2, ..., u_n\}$ とおく．SBM で l 件の Web ページが共有されているとき，これを $\{r_1, r_2, ..., r_l\}$ とする．特定ユーザ u_i が付与したタグ群を $\{T_{i1}, T_{i2}, ..., T_{im_i}\}$ とする．ユーザ u_i のタグ T_{ix} に対

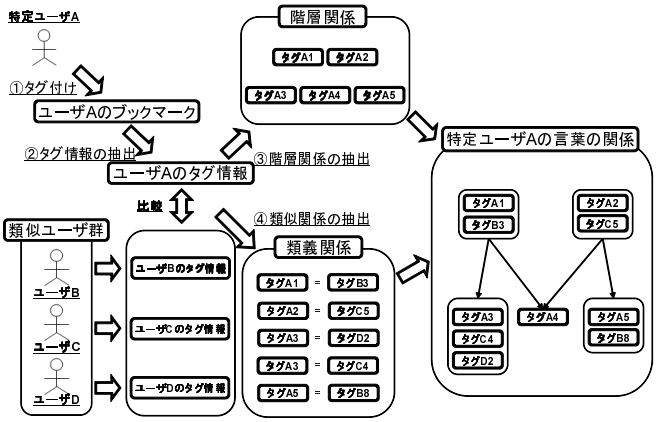

図 2: 手法の概念図

応するクラスタを $Cls_i(T_{ix})$ と書く．特定ユーザ u_i の付与した m_i 個のタグから，${}_{m_i}C_2$ 種類の相異なる 2 タグの組み合わせ (T_{ix}, T_{iy}) を生成し，各組み合わせ (T_{ix}, T_{iy}) の階層関係を判定する．

本節では完全内包式と共起式の 2 種類の手法による階層関係の判定を提案する．完全内包式では，タグ T_{ix} のクラスタがタグ T_{iy} のクラスタを完全に内包するとき，T_{ix} のラベルは，T_iy のラベルの意味を含み，より広義な意味をもつと考える．したがって，$Cls_i(T_{iy}) \subset Cls_i(T_{ix})$ のとき，タグ T_{ix} をタグ T_{iy} の上位の言葉として関連づける．

一方，共起式では，完全に内包でなくとも 2 つの言葉に関連があり，かつ，その 2 つの言葉の指し示す範囲が異なれば，階層関係が成り立つ可能性が高いと考える．任意の 2 タグの組み合わせ (T_{ix}, T_{iy}) について，$|\ Cls_i(T_{ix}) \cap Cls_i(T_{iy})\ | \geq HT_1 > 0$ かつ $\Big|\big|Cls_i(T_{ix})\big| - \big|Cls_i(T_{iy})\big|\Big| \geq HT_2$ のとき親子関係があると判定する．HT_1, HT_2 は，それぞれ共起式でタグのラベル間の階層関係を抽出する際に用いる閾値を表す．$|\ Cls_i(T_{ix})\ | > |\ Cls_i(T_{iy})\ |$ のとき，タグ T_{ix} が上位，タグ T_{iy} を下位のラベルをもつタグとする．$|\ Cls_i(T_{ix})\ | < |\ Cls_i(T_{iy})\ |$ のとき，タグ T_{ix} を下位，タグ T_{iy} を上位のラベルをもつタグとする．

3.4 タグ間の類義関係の抽出

特定ユーザ u_i のもつタグ情報と，他ユーザ群 $Sim - \{u_i\}$ の中の各ユーザが持つタグ情報を比較し，類義関係にある 2 タグの組を抽出する．特定ユーザ u_i とユーザ $u_j \in Sim - \{u_i\}$ のタグ情報を比較する場合を考える．ユーザ u_i が m_i 個，u_j が m_j 個のタグをそれぞれ使用したとき，u_i が使用したタグを $\{T_{i1}, T_{i2}, ..., T_{im_i}\}$，$u_j$ が使用したタグを $\{T_{j1}, T_{j2}, ..., T_{jm_j}\}$ とおく．特定ユーザ u_i とユーザ u_j がそれぞれもつタグの組み合わせ $(T_{i1}, T_{j1}), (T_{i2}, T_{j1}), ..., (T_{i1}, T_{j2}), (T_{i1}, T_{j3}), ..., (T_{im_i}, T_{jm_j})$ を生成する．各 (T_{ix}, T_{jy}) に対して $Cls_i(T_{ix})$ と $Cls_j(T_{jy})$ が似通っている場合，すなわち，タグ T_{ix} とタグ T_{jy} が同じブックマークに対して付与されている率が高い場合，それらのタグのラベルを同義語とみなす．タグ T_{ix} とタグ T_{jy} が類義関係か否かの判断は式 (1) および (2) の指標を用いる．ここで，ユー

ザ u_i のブックマークしている Web ページ数を N_{u_i}，類義語かどうかの判定のさいに使用する閾値をそれぞれ ST_1,ST_2 と定義する．

$$\frac{|\ Cls_i(T_{ix}) \cap Cls_j(T_{jy})\ |}{|\ Cls_i(T_{ix}) \cup Cls_j(T_{jy})\ |} \geq ST_1 \quad (1)$$

$$\frac{|\ Cls_i(T_{ix}) \cap Cls_j(T_{jy})\ |}{N_{u_i}} \geq ST_2 \quad (2)$$

3.5 階層関係と類義関係からの意味情報の同定

3.3 節で述べた方法により，特定ユーザの使用する言葉の階層関係を抽出できる．さらに，3.4 節で述べた方法により，特定ユーザの使用する言葉と類義関係にある言葉を類似ユーザ群のタグから抽出できる．抽出した階層関係と類義関係を結合し，特定ユーザの意味情報を抽出できる．抽出された両関係から意味情報を表す意味情報グラフを作成する．類義関係にある言葉は，同一のノードとして結合される．この結果，図 3 のような意味情報を表すグラフが出力される．本論文では，図 3 のようなグラフを意味情報グラフと定義する．

4. 評価実験

本手法を用い，特定ユーザおよび特定ユーザの所属する類似ユーザ群のブックマークに付与されているタグから，そのユーザの言葉の意味情報を正確に抽出できているかを検証する．被験者 7 名に同一のテーマを与え，そのテーマについて調べてもらい，気に入った Web ページを実験用に構築した SBM システムを使って共有してもらった．各被験者に対して，各被験者を特定ユーザ，その他のユーザを類似ユーザ群とみなして，提案手法を適用した．

提案手法により意味情報を正確に抽出できているか評価するために，各被験者の意味情報に含まれるタグのラベルの階層関係と類似関係の正確性を検証する．

4.1 階層関係の評価

タグのラベルの階層関係の正確性を検証する．被験者にタグのラベルとして使用した言葉の階層関係を手動で構築してもらった．構築の形式には，非巡回有向グラフを採用した．グラフ内の各エッジの向きは上位から下位の言葉である．また，被験者がグラフを構築するさいには，ある言葉とその孫にあたる言葉を直接エッジで結ばないという制約を設けた．ここで，この制約を持つ非巡回有向グラフを順次接続意味情報グラフと呼ぶことにする．一方，提案手法を用いて，図 3 のような意味情報グラフを抽出する．ここで，意味情報グラフにおいて類義関係にある 2 ノードをひとつのノードに縮約したグラフをノード縮約意味情報グラフと呼ぶ．

提案手法を用いて抽出したノード縮約意味情報グラフを順次接続意味情報グラフに変換する．変換によって得られた順次接続意味情報グラフと，被験者が構築した順次接続意味情報グラフの類似度を調べる．類似度

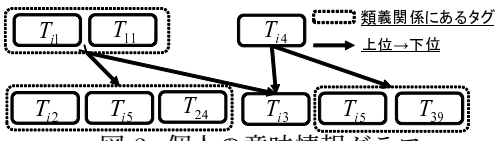

図 3: 個人の意味情報グラフ

が高いならば，提案手法により各被験者の言葉の階層関係を正しく同定できているといえる．ある被験者のタグ付与の結果に基づいて提案手法で抽出した順次接続意味情報グラフを G_s，その被験者が手動作成した順次接続意味情報グラフを G_m とおく．グラフ G_s，G_m が n 個のノード V を持つとき，グラフ G_s, G_m のノードを $\{V_1, V_2, ..., V_n\}$ とおく．ここで，あるノード V_i からノード V_j への有向パスがある場合，2 ノード (V_i, V_j) の関係を子孫関係，逆を，祖先関係と呼ぶ．グラフ内の 2 ノード $(V_i, V_j)(i \neq j,\ \ i, j = 1, ...n)$ の祖先関係と子孫関係の一致数を算出し，その適合率と再現率を 2 つのグラフ G_s, G_m の類似度の指標とする． 各グラフ内の 2 ノード (V_i, V_j) の関係を $n \times n$ 行列で表現する．(V_i, V_j) において祖先関係が成立していれば，行列の (i, j) 成分を 1，子孫関係が成立していれば-1，そうでなければ 0 とする．提案手法で抽出したグラフ G_s の 2 ノード (V_i, V_j) の関係を表現した行列を Mat_{G_s}，被験者の手動で構築したグラフ G_m の 2 ノード (V_i, V_j) の関係を表現した行列を Mat_{G_m} とおく．適合率と再現率を次の (I),(II) の手順にしたがって算出する．(I) 行列 Mat_{G_s} と Mat_{G_m} の差を求める．このとき，(V_i, V_j) の関係がグラフ G_s, G_m で関係が一致する 2 ノードに対応する $Mat_{G_s} - Mat_{G_m}$ の成分は 0 になる．
(II) ある行列 Mat のうち，値 x になる成分の数を，$Cnt(Mat, x)$ と表記する．$M := Mat_{G_s} - Mat_{G_m}$, $Elems(Mat, x) := \{(i, j) \mid Mat := (Mat_{ij}),\ Mat_{ij} = x\}$,$E :=\mid Elems(Mat_{G_m}, 0) \cap Elems(Mat_{G_s}, 0)\ \mid$ とおくとき，式 (3) で適合率を，式 (4) で再現率を算出する．

$$\frac{Cnt(M, 0) - E}{Cnt(Matrix_{G_s}, 1) + Cnt(Matrix_{G_s}, -1)} \quad (3)$$

$$\frac{Cnt(M, 0) - E}{Cnt(Matrix_{G_m}, 1) + Cnt(Matrix_{G_m}, -1)} \quad (4)$$

4.2 類義関係の評価

次に，言葉の類似関係の正確性を検証する．3.4 節でも述べたようにユーザ u_i, u_j がタグをそれぞれ使用したとき，タグの順序つき組み合わせは $(T_{i1}, T_{j1}), (T_{i2}, T_{j1}), ..., (T_{i1}, T_{j2}), (T_{i1}, T_{j3}), ...$,(T_{im_i}, T_{jm_j}) となる．タグの順序つき組み合わせ (T_{ix}, T_{jy}) について，特定ユーザ u_i である被験者が類義語と判断したものと，提案手法が抽出した類義語との一致を調べる．全ての順序つき組み合わせ (T_{ix}, T_{jy}) について，被験者に類義語かどうかを示してもらえば，提案手法が抽出した類義語の適合率，再現率を算出できる．しかしながら，順序つき組み合わせ (T_{ix}, T_{jy}) の数があまりにも膨大なため，提案手法が同義語と判断しうるすべての順序つき組み合わせ，すなわち $\mid Cls_i(T_{ix}) \cap Cls_j(T_{jy})\ \mid > 0$ となる順序つき組み合わせのみを抽出し，適合率のみを算出した．

5. 結果と考察

5.1 提案手法での階層関係抽出の評価結果と考察

3.3 節で述べた完全内包式と共起式を用いてタグ間の階層関係を抽出した．共起式においては，閾値 $HT_1 = 1, HT_2 = 1$ に設定した．2 つの手法について式 (3) で

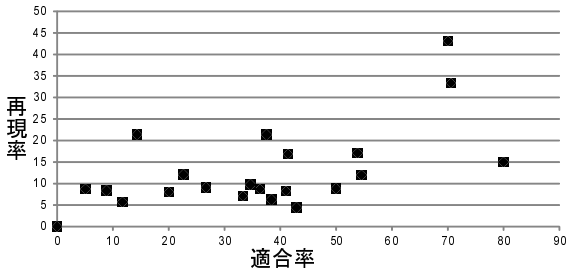

図 4: 完全内包式の適合率と再現率

適合率を，式 (4) で再現率をそれぞれ算出した．その結果，完全内包式は図 4, 共起方式は図 5 のような結果となった．

提案手法の完全内包式，共起式ともに原理上，1 つのブックマークに対して 2 つ以上タグが付与されなければ，階層関係は抽出されない．そのため，1 ブックマークへの平均タグ付与数が少なく，階層関係が抽出されなかった被験者に対しては，タグのラベル間の階層関係を，正確に抽出できなかった．

さらに被験者が 1 つのブックマークに対してどれだけタグを付与しているかが提案手法の精度に影響する要素ではないかと考え，各被験者の 1 ブックマークへの平均タグ付与数に着目して考察した．今回の実験環境と現実の SBM の環境を比較検討する．比較対象となる SBM サービスとして，livedoor クリップを採用する．提案手法が，livedoor クリップの平均タグ付与数と近い平均付与タグ数の被験者に対して高い精度で階層関係を抽出できていれば，提案手法は現実の SBM の環境において，正確に階層関係を抽出できるといえる．livedoor クリップの 2009 年 6 月時点での 25370 人のユーザの 1 ブックマークに対するタグ付与数の平均をとった結果，平均 2.618 個のタグが付与されていた．各被験者ごとに | 平均タグ付与数 − 2.618 | を求める．ここで，| 平均タグ付与数 − 2.618 | をタグスコアと呼ぶ．タグスコアと各手法の適合率との間の相関係数を求めた．その結果，相関係数は完全内包式では，−0.479，共起式では，−0.4548 という値が算出された．タグスコアと提案手法の適合率には，中程度の相関があるといえる．そのため，現実的な SBM の環境に近ければ，比較的高い適合率での階層関係の抽出が期待できる．

5.2 提案手法での類義関係抽出の評価結果と考察

3.4 節で述べた方法で，閾値を $ST_1 = 0.0426, ST_2 = 0.0508$ に設定した．その結果，提案手法は被験者全体で合計 239 件の類義関係を抽出した．各類義関係 (T_{ix}, T_{jy}) について，それぞれのタグ T_{ix} を付した被験者に，その類義関係が正しいと判断できるか否かを回答してもらった．239 件のうち，類義関係があると判断されたものは，合計 60 件あった．タグのクラスタからの情報のみを用いて，25.10% の適合度で類義語を抽出した．

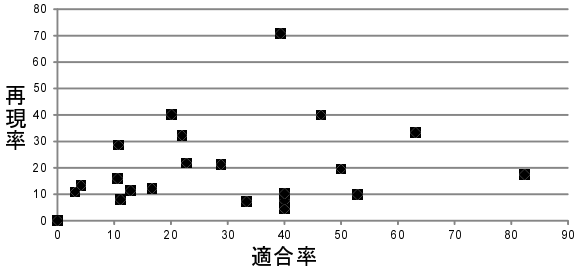

図 5: 共起方式の適合率と再現率

今回の実験環境では，提案手法は十分な精度で類義関係を抽出することができなかった．3.4 節で述べたように本手法では 2 タグ (T_{ix}, T_{jy}) が類義語かどうかの判定に，2 人の被験者がもつタグのクラスタの重なりの大きさ，すなわち $Mth :=| Cls_i(T_{ix}) \cap Cls_j(T_{jy}) |$ の大きさに着目している．今回，Mth は高々3 であった．Mth と，タグ T_{ix} とタグ T_{jy} のラベルが類義語である確率の相関を調べた結果，相関係数は 0.987 と強い正の相関があることがわかった．そのため，Mth を類義関係の判定に用いる式 (1), 式 (2) には正当性があるといえる．

今回の実験で，高精度の類義語抽出ができなかった原因として，実験時間が，1 時間と短かった点が挙げられる．今回の実験では，Mth が高々3 件であった．Mth が広い範囲をとらなかったため，閾値を詳細に設定できず，類義語である 2 タグ (T_{ix}, T_{jy}) とそうでない 2 タグ (T_{ix}, T_{jy}) を正確に判別できなかったと考えられる．

ユーザのブックマーク件数が増えれば，それに伴いタグクラスタの重なり Mth は大きくなる．大きくなれば，Mth と類義語の関係が顕著に現れ，閾値も詳細な設定が可能になり，精度の高い類義語の抽出が期待できる．

6. おわりに

本論文では，個人の言葉の使い方を考慮した検索エンジンを実現するために，類似ユーザ群の SBM におけるタグ付与に基づいて，類似ユーザに所属する特定ユーザの言葉の意味情報を抽出する手法を提案した．言葉の意味情報がもつ，言葉の階層関係と類義関係をタグのラベルではなく，タグが付与された Web ページのクラスタに基づいて抽出した．本手法により一般的な言葉の使い方ではなく，ある特定ユーザ独自の言葉の使い方を抽出できる．

実験結果を検証したことで，今回の実験の期間は十分でなかったことがわかった．今後，提案手法の有用性を評価するために，十分な期間をとり再度実験を試みたい．

参考文献

[1] 内田英里，石野武志：オントロジーの自動構築に関する基礎的研究，人工知能学会，第 3 回セマンティック Web とオントロジー研究会，pp.05-1–05-10，2003 年 6 月．

[2] Xian Wu,Lei Zhang,Yong Yu: Exploring Social Annotations for the Semantic Web, In WWW,pp.417–426,2006.

[3] 丹羽智史，土肥拓生，本位田真一: Folksonomy の 3 部グラフ構造を利用したタグクラスタリング，合同エージェントワークショップ & シンポジウム 2006，2006 年 10 月．

[4] 矢島 健太郎，井上 潮: ソーシャルブックマークにおける文書解析を利用した類似文章および類似ユーザの推薦方法の提案，第 18 回データ工学ワークショップ，第 5 回日本データベース学会 年次大会，C9-3，2007 年 3 月．

[5] はてなブックマーク: http://b.hatena.ne.jp/

[6] livedoor クリップ: http://clip.livedoor.com/