ソーシャルタギングに基づくオントロジー構築支援システムの提案と評価 2H1-1

An Ontology Development System with Social Tagging

木村 牧人[*1] 手島 拓也[*1] 石川 雅之[*1] 森田 武史[*1] 和泉 憲明[*2] 山口 高平[*1]
Makito Kimura Takuya Tejima Masayuki Ishikawa Takeshi Morita Noriaki Izumi Takahira Yamaguchi

[*1] 慶應義塾大学 Keio University
[*2] 独立行政法人産業技術総合研究所 National Institute of Advanced Industrial Science and Technology

As it is necessary for us to develop ontologies with less cost, we propose an ontology development system based on social tagging. Although some researchers propose folksonomy tags as linguistic resourses for ontologies, there is still a big gap between folksonomy and ontologies, so it is hard to develop ontologies just by using folksonomy tags. To solve this problem, we propose the way of adding short definition sentences to tags so that we can define correct semantics from various semantics of the tags.

1. はじめに

ドメインオントロジー構築に利用できる既存リソースの一つとしてソーシャルタグの生成コストの低さ，情報の専門性などが近年注目されている．しかしソーシャルタグとオントロジーの間には大きな形式度の違いがあり，この差を埋めない限りタグを利用したオントロジー構築は難しい．

タグとオントロジーは相対的であり，形式度は低いが入力コストの低さゆえ一般ユーザからの利用の多いタグが存在する一方，構築コストは高いが企業など情報統制を必要とする組織の情報整理に高い能力を示すオントロジーが存在する．ソーシャルタグの情報量を生かすべくソーシャルタグの共起性に注目したクラスタリングによるオントロジー構築手法の研究がなされている[Mika 05]．しかしタグの利用が増加した反面，ユーザ毎に様々な記述方法が存在するためタグの形式度は非常に低く，知識共有のための共通語彙であるオントロジーに必要とされる秩序性は到底満たされていない．また爆発的とも言えるタグの増加量により，同一ラベルのタグがユーザ間で異なった意図として利用されるケースが目立ち始めた[Au Yeung 07]．タグのセマンティクスの認識がユーザ毎に異なる以上，タグ同士の関係を定義しても質の高いオントロジーの生成は困難である．

本研究の提案するシステムではソーシャルタギングに基づくセマンティック Web のためのドメインオントロジー構築支援を行う．ソーシャルタギングの持つボトムアップの性質を利用しつつ，複数のユーザの意見を採用することにより信頼性が高く情報の新規性に対応した構築手法を目指す．今回の手法ではタグとオントロジーの間に存在するものとして「意味情報」を定義し，本提案手法の実現可能性の高い対象ユーザ群としては，一般ユーザと組織の中間に存在する，お互いを認知しあった者同士で構成されるコミュニティを想定する．このようなコミュニティに対し，通常のタグ情報よりも形式度を高めた「意味情報」を用いることで，どのように情報共有がなされ，意味同定を経てオントロジー構築へと繋がるか検証する．

本論文ではシステムの主要部分となるセマンティクス同定手法について詳しく述べ，セマンティクス同定を行わない場合に生じる意味の非明確性の検証を，全体の部分実験として行う．まずタグの持つセマンティクスの同定を行い，また同一の意味に対して付与されるタグの入力支援によりタギング時のタグフィルタリングを行い形式度の向上を図る．セマンティクスが明確化されたタグに対して関係定義を行うことにより，利用価値の高いドメインオントロジーの構築手法を提案する．

連絡先：木村牧人,山口高平, 慶應義塾大学理工学部,
〒223-8522 神奈川県横浜市港北区日吉 3-14-1
Tel:045-566-1614,E-mail:{m_kimura,yamaguti}@ae.keio.ac.jp

2. 関連研究

関連研究として，V.Tanasescu らの Extreme Tagging がある[Tanasescu 07]．本研究の狙いと同様にソーシャルタギングの手法を利用しボトムアップ的なオントロジー構築方法を提案した．この研究ではタグ―タグ間，リソース―タグ間のプロパティをユーザに自由に記述させることでトリプルを取得するというものであるが，プロパティは特に定まった仕様に基づくわけではなく，プロパティの形式度は低い．またタグの多義性解消に関しては，トリプル内で共起されるタグを指定しトリプルを利用した検索を行うことで多義性解消が行うと主張しているが，その判断は機械的処理では行えず，あくまでもユーザの主観によるものであり，この手法ではセマンティクスの同定は不十分だと言える．タグが他のユーザにどのような意図で使われているかユーザの推測のもと利用されている．そのためタグのセマンティクスは暗示的に示されているに過ぎず，同じラベルを持つだけでは同一のセマンティクスを表しているのか正しく判別ができない．

3. オントロジー構築支援システムの設計

3.1 システム概要

ラベル情報に注目するだけでは背景に存在する暗黙的な情報の違いにより意味の差が生じる．それを明確にするのがオントロジーの本来の目的であり，ラベルよりもそのラベルの意味（概念）を重視すべきである[溝口 06]． フォークソノミータグが表す語彙情報はこのラベルに相当し，タグ情報からオントロジーを構築するためには，通常のソーシャルタグに加え，そのタグがどのような意味を持つのかという情報を加える必要がある．タグの持つセマンティクスを同定し，ユーザ間で一意の共通した認識が共有できた結果，プロパティを結び形式度が高く利用価値のあるオントロジーの既存リソースとすることが可能である．

今回提案するシステムでは，タグの持つセマンティクスの同定を第一に行い，タグに意味を付与した後にタグ間の階層関係をユーザに行わせオントロジー構築の基盤とする．従来のタギングシステムではタグを暗黙的なセマンティクスを表すラベルとして利用していたが，本システムでは

- 説明文 ID
- 説明文
- ラベル（タグ）の集合
- 定義ユーザ ID
- 採用ユーザ数

これらをまとめて「意味情報」として定義する．特にこの意味情報が示す内容がクラスを表すものである場合，この意味情報は説明文 ID に基づく URI，説明文による意味，付与されたタグ集合による同義語の集合という情報がそろっており，これは概念として扱うことが可能である．

モジュールとしては意味情報を定義するセマンティクス同定モジュール，意味情報が付与されたタグ同士の関係定義を行う関係定義モジュール，URL に意味情報を付与してブックマークを行うブックマーク登録モジュール，意味情報を利用し検索を行う検索モジュール，ユーザ毎のパーソナルオントロジーを構築するオントロジー構築モジュールが存在する．

以下は各モジュールについて述べる．

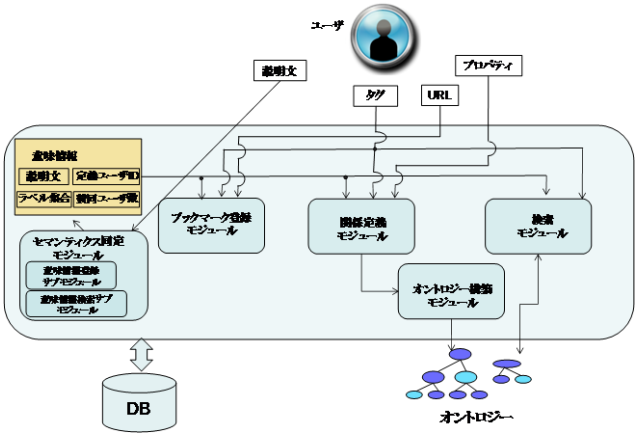

図 1：システム構成図

3.2 セマンティクス同定モジュール

セマンティクス同定モジュールとはタグのセマンティクスをユーザがタギング時に用いた正しい形で付与するモジュールである．タギング時にユーザが想定するセマンティクスを明示化することにより，タグの持つセマンティクスを一意に認定することが可能になる．

処理の流れとしては，セマンティクスを同定したいタグ（入力タグ）を入力する．データベースを照会し，その入力タグに対して既に登録されている説明文と，その説明文に対し付与されている他のタグ全てをユーザに返す．ユーザは入力タグの持つセマンティクスを最も的確に表す説明文を選択，もしくは新規登録を行う．選択した説明文と入力タグ，ユーザ ID を DB に格納することで最終的にタグの持つセマンティクスを入力ユーザの同意により明確に定義することが可能になる．タグをセマンティクスと合わせて識別することで多義性の解消も同時にこのモジュールで行うことが可能になる．

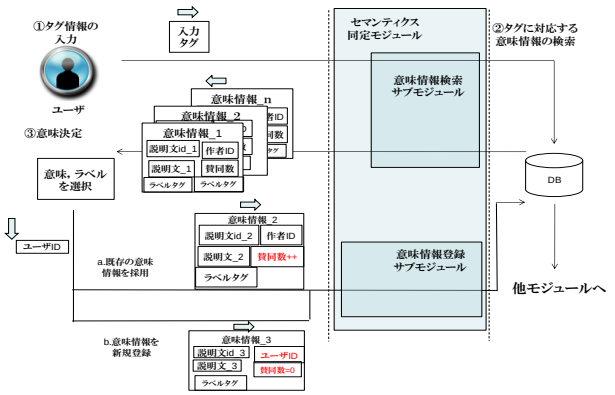

図 2：セマンティクス同定モジュール構成図

実際のセマンティクス同定の流れは以下の図3のようになる．ユーザがテレビ，雑誌に関するリソースに付けた”メディア”というタグに対してセマンティクス同定を行うとする．まず“メディア”をデータベースに照会し，“メディア”というタグについて既に登録されている説明文を取得する．その取得した説明文に付与されている他のタグ（ラベル）情報も併せてユーザに返す．この例ではユーザは意味情報_2 の意味を採用し，またその意味を表すラベルの採用数を参考にし，“メディア”よりも“マスメディア”というタグがこの概念を表すのにより適しているとユーザが判断した場合，入力したタグを“マスメディア”に変更して URL に付与することが可能である．このようにユーザ自身が入力したタグよりも適切に概念を表す可能性のあるタグへの変更を可能にすることで，ある概念の主なラベル表記として人気の高いタグを抽出することが可能になり，概念を表すタグとしてふさわしくないタグのフィルタ機能となる．

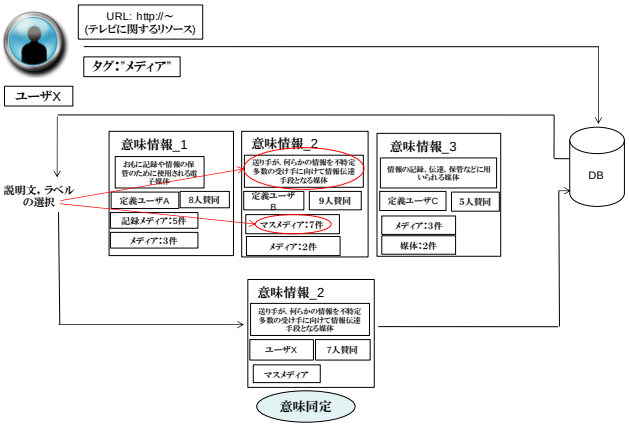

図 3：セマンティクス同定例

3.3 ブックマーク登録モジュール

ブックマーク登録モジュールでは，セマンティクス同定モジュールで生成された意味情報をリソースに付与する．先述したセマンティクス同定のアウトプットである意味情報と，付与する対象となるリソースの URL をインプットとし，対応付けを行う．

通常のソーシャルブックマークと異なる点は，ユーザ側の認識としてはタグと URL の入力後，そのタグの持つ意味を最も適切に表す説明文をユーザに選択もしくは新規入力させるという点である．この過程を追加することにより，システム側では URL に対してタグを付与するのではなく，ユーザが入力したタグに対応する意味を明示的にリソースに付与することが出来る．

3.4 関係定義モジュール

このモジュールでは説明文同士の関係定義を行う．ユーザは自分が登録済みのタグと，関係定義に利用するプロパティをインプットし，タグとタグの間をプロパティで結ぶ．その間のシステム側の流れとしては，まず入力されたタグとユーザ ID に対応する説明文をデータベースから参照し，説明文同士をプロパティで結ぶ．説明文を参照する際にその説明文の id に対応する全体のユーザから付与された別のタグ（ラベル）も取得する．説明文 id は URI で識別されており，この関係定義は URI,ラベル(タグ)，説明文の三つが揃った概念間の関係定義となり，形式度が非常に高くオントロジーへの転用のしやすいものである．

ユーザは付与したタグ間の関係定義をすることにより，自身のタグをツリー形式に整理を行うことが可能になり，現行のフォークソノミーで表示されるタグクラウドに比べ階層化された整理された状態での表示が可能になる．

また定義の際に既に他のユーザが提案した関係定義が自分の弁別の基準に適すると判断した場合，ユーザはその関係定義をそのまま採用することも可能であり，この概念間関係定義の共有機能により，ユーザの関係定義を行う際の負担を軽減することが可能である．

今回の提案手法のプロパティは以下の表に表す rdf,rdfs 語彙を利用する．タグがインスタンス情報を表す場合，クラス―インスタンス関係を定義することも可能である．

表 1:プロパティ説明

rdfs:subClassOf	クラス間の階層関係定義
rdf:type	クラス―インスタンス関係定義
rdfs:label	ラベル記述
rdfs:comment	コメント記述

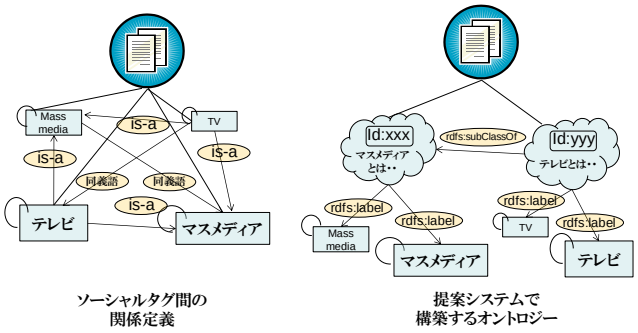

図 4：意味間の関係定義

3.5 オントロジー構築モジュール

このモジュールでは，関係定義モジュールで定義した概念間の関係を統合しセマンティック Web のためのオントロジーの構築を行う．関係定義モジュールのアウトプットである意味情報間の関係定義，ユーザ ID をインプットとし，アウトプットは一人のユーザの関係定義をまとめたパーソナルオントロジーとする．パーソナルオントロジーとはユーザ自身が登録した概念やインスタンスから構築されるオントロジーであり，ユーザ自身の情報整理の利便性に貢献する．

3.6 検索モジュール

検索モジュールでは，オントロジー構築モジュールにより生成されたオントロジーを参照し，入力したタグに対して対応するセマンティスクに基づいた意味検索を行う．検索を行いたいタグをインプットし，データベースから対応する意味情報と付属する他のタグ群を取得する．これにより入力したタグのセマンティクスを特定し，そのセマンティクスが付与されたドキュメント全体の検索が可能になる．

以下の図 5 で示すように，旧来のフォークソノミーのタグ検索は入力タグとのテキストマッチングによるタグの検索，さらに付与されているリソースの検索を行っている．この検索手法では”ツール”が表す概念を表す他のラベル表記として付与されている”tools”,”TOOL”などが検索対象から外れてしまっており，さらに”ツール”というタグもセマンティクス同定がなされていないためユーザが意図した”ツール”とは異なる意味でタグ付けされたリソースまで検索結果に挙げられてしまう．提案するシステムではまず入力タグに対してセマンティクス同定を行い，ユーザが検索したい意味情報を明確にする．さらにその意味情報に対して付与されたリソースをデータベースから検索するため，ユーザの意図を適切にした意味検索の実現が可能であり，また意味をそのまま検索するので異表記ラベルによる検索結果の減少の心配もない．

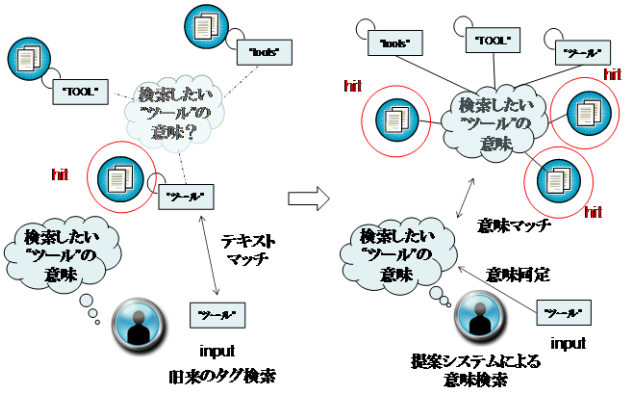

図 5：意味検索

4. ケーススタディ

今回の実験では既に実装済みの機能であるボトムアップの手法に基づく関係定義の精度について部分的に検証を行う．

4.1 実験方法

実験対象としては情報系の学生による二週間の試用を行い，実際にタグ間の関係取得を行った．実験概要は表 2 に示す．セマンティクスの同定を行わずに，タグの持つ暗黙的なセマンティクスに関する関係定義を行いソーシャルタギングの手法による複数のユーザによる関係定義を行い，性質継承の精度について検証した．

表 2：実験概要

テストユーザ	情報系学生 26 名
実験期間	2008.1.14～2008.1.28
手法	インターネット接続によるサービス利用
ブックマーク数	658
対象ブックマーク領域	IT 分野を推奨

4.2 実験結果

(1) 全体の性質継承率

性質継承率，取得数の結果を以下の表に示す．全体としての正解率は 80%を超えた．また is-a 関係の正誤の判断は，性質継承の有無を基準にし，性質継承が正しく行われている場合を正解とした．

表 3：上位下位関係取得結果

is-a 関係取得数	408
正解数	336
正解率（%）	82.4

(2) 階層関係定義の被共有数と性質継承率の関係

多くのユーザに共有された人気の高いタグ間関係と正解率の高さを検証するために，タグ間の関係定義の被共有数と性質継承率の関連を調べた．結果を図 6 に示す．

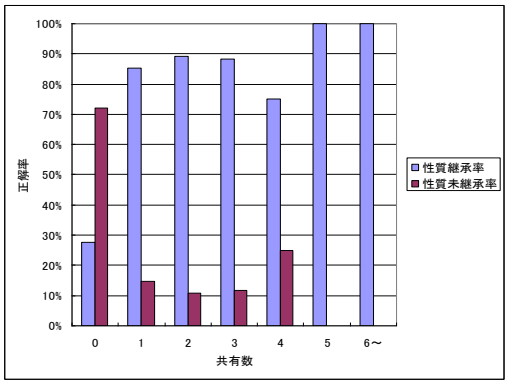

図 6：関係定義の被共有数と性質継承率の関係

結果としては，関係共有数が 0，すなわち定義したユーザ以外から誰にも共有されなかった場合と，関係共有数が 1,すなわち定義ユーザ以外のユーザ一人以上に共有された場合とでは大きな差があった．

(3) オントロジー取得結果

得られたオントロジーの一部分を図 7 で表す．この結果より，前述したようにユーザがタグ付けの際の意図を正しく反映できていない例が目立つ．また今回は上位下位関係を is-a 関係として定義し，クラスとインスタンス情報を厳密にはユーザに区別させなかったが，取得結果を見ると概念間の階層関係が定義されたものの末端にインスタンス情報が付与された例が多く存在した．

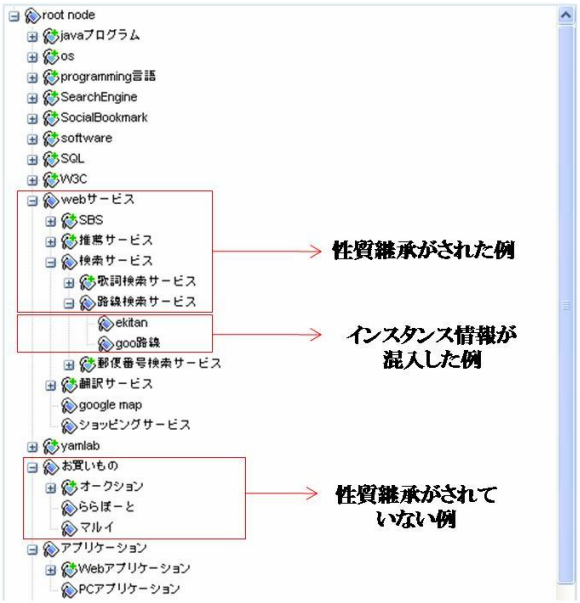

図 7：取得結果オントロジーの一部

4.3 考察

(1) 階層関係定義の非共有数と性質継承率の関係

結果をみると，関係共有数が 1 以上，すなわち定義ユーザ以外の最低一人以上のユーザに共有された階層関係定義は誰にも共有されなかった階層関係に比べ性質継承率が高くなった．ユーザ独自の価値観を強く反映した一般性の低く正しく性質継承を行われていない階層関係が他のユーザに共有されておらず，フィルタ作用が働いたことがわかる．

(2) オントロジー取得結果

取得した階層関係の一部を次の表に載せる．特に性質継承が正しく行われていない例に注目すると，”Wikipedia is-a 便利”などは，ユーザ自身の分類の為に付与されたタグであり，性質継承の観点からタグ付けが行なわれておらず，オントロジーへの利用価値の低い関係である．次に”youtube is-a 動画”，”photoshop is-a 画像”というこれらの関係は，ユーザがタグ付けの際の意図を正しく反映することができず，本来付与したい概念を曖昧に表すタグが付与されてしまっている．これらは”動画共有サービス”,”画像編集ソフトウェア”というタグであればユーザが意図した概念を正しく付与することができる．また”ajax is-a web2.0”という関係は，ajax,web2.0 という専門分野の中でも複数の捉え方，すなわち複数のセマンティクスを持つ単語に対し，ユーザ自身の主観を強く反映させて関係定義を行った例である．この場合は， ajax と web2.0 という概念がドメイン内でどのような概念として共通認識されているか明示化してから概念関係を定義すべきである．

表 4：性質継承の正誤例

性質継承が正しく行われている例		
photoshop	is-a	software
Editor	is-a	ソフトウェア
マスメディア	is-a	メディア
写真	is-a	画像
youtube	is-a	動画共有サービス

性質継承が正しく行われていない例		
ウィキペディア	is-a	便利
youtube	is-a	動画
photoshop	is-a	画像
ajax	is-a	web2.0

5. おわりに

今回の研究では，セマンティクス同定を行ったタグ間に関係定義を行い，ドメインオントロジー構築の基盤となる形式度の高い概念関係の取得方法を提案した．

また実験結果として，意味同定を行わず現状のタグを直接オントロジーへ利用することの問題を明らかにし，今回提案する方法論の有用性を示した．

今後の課題としては，現行のシステムに意味同定モジュールとオントロジー構築モジュールの実装を行い，パーソナルオントロジーの取得を試みる．また，複数のパーソナルオントロジーを統合するために必要となるアラインメント手法を考案し，システムの最終的なアウトプットであるセマンティック Web のためのドメインオントロジーの構築を行っていく予定である．

参考文献

[Mika 05] Petar Mika : “Ontologies Are Us:A Unified Model of Social Networks and Semantics” ,ISWC 2005,2005

[Au Yeung 07] Ching-man Au Yeung, Nicholas Gibbins, Nigel Shadbolt : “Understanding the Semantics of Ambiguous Tags in Folksonomies”, International Workshop on Emergent Semantics and Ontology Evolution 2007,2007

[Tanasescu 07] V.Tanasescu, Olga Streibel : “Extreme Tagging: Emergent Semantics through the Tagging of Tags”, International Workshop on Emergent Semantics and Ontology Evolution 2007,2007

[溝口 06] 溝口 理一郎著: オントロジー構築入門，オーム社 ,2006