ブログデータに基づくユーザの興味オントロジ自動生成とコミュニティ形成支援手法の提案

中辻　真[†]　三好　優[†]　大塚　祥広[†]

† 日本電信電話株式会社 NTT ネットワークサービスシステム研究所
〒 180-8585 東京都武蔵野市緑町 3-9-11
E-mail: †{nakatsuji.makoto,miyoshi.yu,otsuka.yoshihiro}@lab.ntt.co.jp

あらまし　近年，ユーザ興味を発信する手段としてブログの利用が目覚しい．しかし現状，ユーザ興味を詳細に情報化する手段がないため，大量のブログエントリが日々発信されているにも関わらずユーザ興味に即したエントリやユーザの発見が実現できておらず再利用性が低い．そこで本研究では，ユーザの興味情報をクラス階層として詳細に表現する興味オントロジを導入する．そして，ドメインごとに形成される雛型オントロジに対し，蓄積ブログデータを分類することでユーザごとの興味オントロジを自動生成する手法を提案する．その上で，興味オントロジ間の近似度計測に基づくユーザ興味に即した新しい形態のコミュニティ形成支援手法の提案を行う．さらに，ブログポータル Doblog における大規模な実ブログデータと，音楽ドメインにおける雛型オントロジを利用した検証により，本手法が適切な興味オントロジを自動生成できることと，コミュニティ形成を支援できる可能性を持つことを示した．

キーワード　オントロジ，オントロジマッピング，情報推薦，コミュニティ形成，セマンティック Web，ブログ

A Proposal of Automatic Generation of User Interest Ontology and Formalization Support of Weblog Community

Makoto NAKATSUJI[†], Yu MIYOSHI[†], and Yoshihiro OOTSUKA[†]

† NTT Network Service Systems Laboratories, NTT Corporation
9-11 Midori-Cho 3-Chome, Musashino-Shi, Tokyo, 180-8585 Japan
E-mail: †{nakatsuji.makoto,miyoshi.yu,otsuka.yoshihiro}@lab.ntt.co.jp

Abstract　Recently, the use of Weblog is remarkable as the means to publish the user interest. However, there is no means to informationize the user interest in detail, it is difficult to find suitable information resources in spite of a large amount of weblog entries are published every day. In order to create suitable user interest information, we classify user entries to domain ontology and create interest ontology which expresses user interest in detail as a class hierarchy. Furthermore, we try to formalize weblog community based on the degree of similarity between interest ontologies. We evaluate the performance of our proposed automatic interest ontology generation and community formalization support technique based on large-scale weblog entries on Weblog portal Doblog and music domain ontologies.

Key words　Ontology, Ontology Mapping, Information Retrieval, Community Formalization, Semantic Web, Weblog

1. はじめに

近年，インターネット上でユーザの興味対象を発信しユーザ間での議論を促進する Weblog（ブログ）サービスや互いに友人として承認し合ったユーザ間で興味対象を議論する Social Networking（ソーシャル・ネットワーキング）サービス等が注目されており，今後ますますユーザ数やこれらを利用したサービスは拡大していくと考えられる [12]．そして，この種の情報流通サービスは，ユーザが自身の興味に近いユーザの発信記事やコミュニティでの議論内容を閲覧する事を通じ，各自の興味対象を拡大する基盤となる可能性を持つため，興味深い．

しかし，現状のブログサービスにおける情報検索は，goo(注1)などの Web ページ検索エンジンや，RDF Site Summary（RSS）という簡単なメタデータ記述を利用したキーワード検索でしかない．更に，個人の興味情報を自動的に生成する機能を備えていないため，自身の興味に即した検索目的語を適切に構成する必

(注1)：http://www.goo.ne.jp

要があり，検索キーワードの選択に手間がかかる．また，事前に検索対象をある程度把握していないとキーワード自体を構成できないため，興味を持つ可能性があるがキーワードを特定できない場合は，情報検索自体ができないことも多い．

こうしたユーザの興味情報の生成や興味に基づく情報推薦に関する研究は，従来の Web 検索においても様々な研究が行われている [4,8,10]．例えば，個人の興味を検索ログや個人が分類した Web ページに対するブックマーク情報などから推定し検索に活用する研究がある [4,10]．しかし，個人のブックマークなどは音楽や映画などの興味分野ごとに詳細化された階層情報を持たない事が多いため，興味に即した情報推薦へは利用できない場合もある．一方，Web 上の情報に対しその背景となる意味情報を機械処理可能なオントロジとして記述する事で，様々なソフトウェアが自動でオントロジによる処理を実行する事を目指すセマンティック Web [2] に関する研究では，オントロジに基づく自動的な情報推薦やオントロジマッピングに基づく情報統合に対する研究が盛んである [6–8]．セマンティック Web 技術をブログ検索に適用する研究として，Semblog [8] ではユーザの興味情報をオントロジとして構築する事で，オントロジマッピングに基づく興味に即した検索を試みている．このように，セマンティック WEB 技術をブログ検索に適用する事の有効性が唱え始められているが，その核となるオントロジ構築が難しい．

こうした問題を解決するため本研究では，ブログからユーザ興味を興味オントロジとして自動生成する手法や，興味オントロジ間の近似度計測によるコミュニティ形成支援を行う手法を提案する．そして，個人が興味を持つ可能性が高い未知な情報を意外な情報と定義し，一般的な Web 利用者が意外な情報を自然と発見し，新たな興味に組み込むことを支援する基盤を提供することを目的とする．

具体的には，個人の興味を詳細に情報化するため，近年急速に利用が進んできたブログを利用し，個人の持つ興味概念（クラス）を階層的に記述する興味オントロジを自動生成する．つまりブログは一般的な Web ページや BBS と異なり，個人としての興味を記述していることが多いため，個人の興味特定に利用する．なお，ブログでは例えば音楽と映画など複数ドメインに跨り興味が混在した形態で自由記述されている事が多いため，各ドメイン毎の情報をクラス階層として記述した雛型オントロジを予め用意し，雛型オントロジに対し個人のブログエントリを分類する事で，興味オントロジをドメイン毎に分離し自動生成する．そして，複数ユーザの興味オントロジ間でクラスやクラスの接続形態であるトポロジの近似度を基に，オントロジ間の近似度を計測する．さらに，近似度が高いオントロジ間で一部トポロジが異なるクラスに属するエントリを，意外な興味エントリとしてユーザ推薦することで，ユーザの興味幅の拡大と，他ユーザ間とのコミュニケーション促進を狙う．

また，ブログポータル Doblog(注2) における大規模データ (約 5 万 5 千ユーザ，160 万エントリ) を用い提案手法の検証を行い，本提案が高精度な興味オントロジ生成やブログコミュニティ解析，およびユーザ毎の興味に即したエントリ推薦によるコミュニティ形成に対し有効性を持つことを確認した．

以下，2. 章では，本論文の背景となるブログの概要説明とその問題点について述べ，関連研究の紹介も行う．3. 章では，雛型オントロジを用い，ブログユーザ毎の興味オントロジを自動生成する手法を提案し，4. 章において，興味オントロジを用いたユーザ興味に即したエントリ推薦と，コミュニティ作成支援について述べる．5. 章では，実ブログデータを用いた興味オントロジ生成とユーザ分布解析，および推薦エントリの検証を行い，6. 章の結論と将来の課題で結ぶ．

2. ブログの概要と関連研究

ブログの定義は明確ではないが，主に MovableType(注3) などのブログ作成ツールや，Doblog などの Weblog ホスティングサービスを用い個人が発信するニュースサイト，または日記サイトという位置づけで理解されることが多い．その特徴を説明すると [9,14]，(1) 個人が Web 上で発信する記事（エントリ）の集合体であり，(2) エントリが時系列に表示されている．そして，(3) ブログ作成ツールを用いることで作成したエントリを簡単に公開できる．また，(4) サイト内外のエントリを元情報にしたエントリを発信する trackback といわれる機構を持つ．これにより，ブログサイトを跨りエントリ一覧（スレッド）を構成する事ができ，ユーザ間のコミュニケーションやエントリ毎にテーマを絞った議論が行いやすい．それ以外に，(5)RDF Site Summary（RSS）と呼ばれるメタデータをエントリ記述の際に生成し，ブログの更新情報を集め提供しているサーバである ping サーバや Weblog ホスティングサービスの運営するサーバに登録することで，エントリの存在や簡単な内容，更新情報などを他のユーザへ公開できる．RSS は，Web サイトの各ページのタイトル，アドレス，要約などをメタデータ記述できるものであり，公開 RSS に対し RSS フィードというサービスを用いることで，多数 Web サイトの更新情報を効率的に把握できる．

このように RSS はメタデータを構成し，流通させる仕組みを持つため，Semantic Web におけるオントロジ普及に期待されている．しかし，RSS はユーザがブログを公開するときに最低限必要なメタデータのみを提供するものであり，Semantic Web で期待される詳細なクラス関係を持つオントロジを簡単には生成できない．RSS フィードを用いた検索でも，メタデータが上記単純なものであるため，ブログエントリの発見には，ユーザが検索キーワードを予め構成する必要があることに変わりない．そのため，キーワードが明確でない限り興味に即したエントリであっても発見できない事が多い．

これに対し Semblog [8] では，パーソナルオントロジというツリー構造を持つカテゴリ体系をユーザが生成し，自身が記述もしくは収集したエントリを分類する．つまり，ユーザの興味体系であるパーソナルオントロジをボトムアップアプローチで生成することで，他ユーザの持つパーソナルオントロジや各種トピックディレクトリとのマッピングができる．一方，本研究

（注2）：http://www.doblog.com/weblog/PortalServlet

（注3）：http://www.movabletype.org/

はドメイン毎に用意する雛型オントロジを用い興味オントロジをトップダウンアプローチで自動生成するため，ユーザによるオントロジ設計・構築の手間がない．

その他，個人の登録ブックマークや保持フォルダなどの階層構造と，ブックマークやフォルダの格納ファイルに基づき個人の興味情報を階層的に構築し，協調フィルタリングに基づきユーザにとって興味の近い別ユーザの興味階層に属する情報を推薦する研究がある [4,10]．しかし，ブックマークなどの階層構造には様々な分野の興味が混在する事も多いため，ブログコミュニティ形成に適用するには，分野毎の詳細な情報を適切に切り出す技術が必要となる．本研究では，様々な対象を記述しているブログエントリに対し，ドメイン毎に適切な粒度の雛型オントロジを用意する事で，ドメイン毎に分離しユーザ興味を適切に反映した興味オントロジを生成できる．また，複数オントロジ間で近似度の近いクラスに基づく情報を推薦するだけでなく，近似度が高いオントロジ間で一部トポロジが異なるクラスに属する情報を，意外な情報として推薦を試みる点も異なる．

また，トピックディレクトリに対し Web ページを分類する研究 [3] が存在する．本研究の興味オントロジ生成の際の雛型オントロジに対するエントリ分類法との違いは，本研究では (1) オントロジのみを利用し，従来手法で必要な大量な Web ページより構成されるディレクトリやページ間のリンクを必要としない点，(2) オントロジの持つクラス特性を利用し分類誤りを除去する事で高精度な分類を実現できる点，および (3) ブログに適用しユーザ毎の興味オントロジを自動生成できる点である．

一方，エントリ間のリンク構造を分析し関係のあるブログサイト集合をコミュニティと捉え，抽出を試みる研究がある [1,5,14]．これらは，従来の Web コミュニティ抽出手法 [3] をブログへ適用するものが多い．上記研究のコミュニティ形成支援への適用課題は，抽出されるコミュニティの興味対象が明示できない点と，既にリンク関係が形成されているページ集合を抽出するためユーザに対し意外な情報を推薦するとは言えない点である．

また，オントロジ間の近似度計測やマッピングに関する研究がある [6,7]．例えば，文献 [6] では，クラスの接続形態であるトポロジを考慮した近似度計測手法を提案している．本稿では，上記手法を雛型オントロジを介した計測手法に拡張する事で計算量を抑えている．また，近似度の高い興味オントロジ間で共起するクラスやトポロジを分析し，あるユーザの興味オントロジには出現しないが，そのユーザと近似度の高いオントロジに頻出するクラスを抽出する．そして，それらに基づくエントリを意外な情報としユーザ推薦する事でブログユーザ間のコミュニケーション促進を試みる．

3. 興味オントロジ自動生成手法

本章では，音楽や映画といったサービスドメイン毎の雛型オントロジ設計法について説明し，それを用いた興味オントロジ生成法について説明する．

3.1 雛型オントロジの設計

本節では，OWL(Web Ontology Language) [11] を基にオントロジの説明をした上で，雛型オントロジの設計法を説明する．

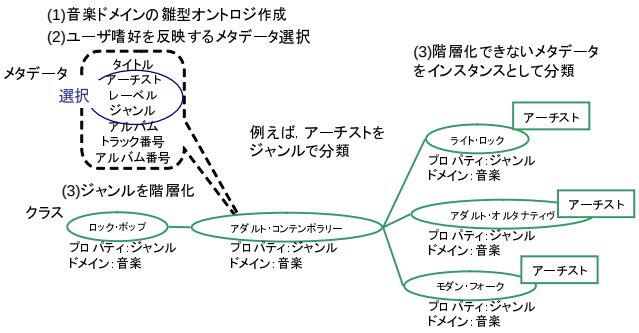

図 1　雛型オントロジ構築手順

OWL におけるクラスは，同様の性質を持つ個体をグループ化しその性質を論理的に表現するための機能を提供する．クラスは，クラスの持つ個体であるインスタンスの列挙などのクラス表現を用い定義される．また個体同士の関係や個体とデータ値の関係を定義するプロパティは，RDF スキーマの rdfs:range や rdfs:domain で値域，定義域を記述できる．さらに，クラスのインスタンスに関する公理を用い，例えば owl:sameAs により 2 つのインスタンスが同値である事を記述できる [11]．

このように OWL を利用すれば，ドメイン毎のオントロジを詳細に設計できる．さらに，OWL を人間が手書きで記述するのは困難なため，*protégé*(注4) などのオントロジ記述サポートツールの研究も進んでいる．とはいっても，やはり詳細なオントロジ設計や記述を一般ユーザが行うのは負担が大きく，オントロジ生成・流通を阻害すると著者らは考えている．そのため，本研究ではまずは雛型オントロジを，OWL 記述法則の中でもクラスの階層関係（subClassOf 記述）とクラスに所属するメンバーであるインスタンスの列挙（oneOf 記述），クラスのドメインやレンジの指定（domain 記述，range 記述），階層構造の基準となるメタデータを指定するプロパティ記述のみを用いるライトウェイトなオントロジ [13] として設計する．そして，ユーザの興味オントロジは，雛型オントロジへのユーザエントリ分類を通じ自動生成し，ユーザはオントロジ記述を行わない．

なお，雛型オントロジの設計には，クラス間の階層関係やユーザ興味を細やかに反映するための末端クラスの粒度調整が必要である．幸い，goo 等のポータルサイトにおけるトピックディレクトリは詳細化が進んでおり，例えば，音楽サービスドメインのジャンルを例に挙げると Web で公開されるジャンルの階層情報はユーザ興味に従う検索を考慮し，粒度を細かく設定している．そのためまずは，これらのトピックディレクトリを基に雛型オントロジを構築し，本研究における分析を通じ，適切な粒度考察を進める．

以下，雛型オントロジの設計手順を図 1 に示す例を基に説明する．まず，(1) 設計者は興味オントロジとしてどのドメインのオントロジを生成するかを選択する．その上で，(2) そのドメインにおいてユーザ興味を反映するメタデータを選択する．選択材料としては，掲示板などの既存コミュニティの傾向を分析すればよい．例えば，音楽ドメインは，ジャンル・アーチス

(注4)：http://protege.stanford.edu/

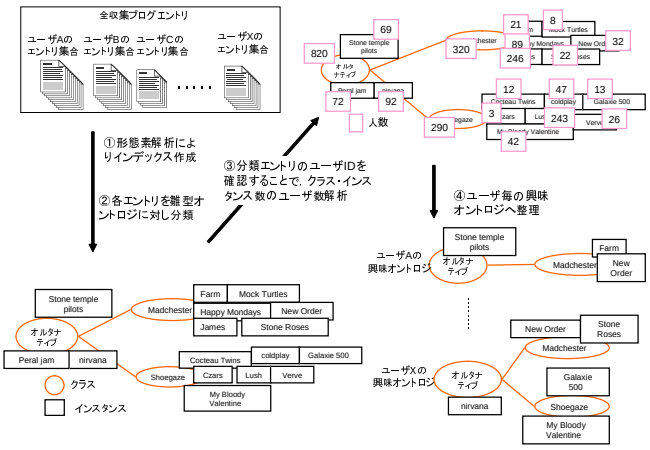

図 2　ユーザ分布解析と興味オントロジ自動生成手順

トなどでコミュニティが生成されていることを考慮し，上記メタデータがユーザ嗜好を反映すると想定し，選択する．次に，(3) クラスツリーを作成可能なメタデータを選択し，クラス階層を形成する．この際，選択されたメタデータをクラスの性質を制約するプロパティとしてクラス階層間で継承する．例えば，ジャンルをプロパティとして継承するクラス階層を構築しアーチストなどをインスタンスとして各クラスに分類する．

なお本稿では，興味オントロジは雛型オントロジを基に作成する．これは著者らは，興味オントロジをユーザ自身が生成するのは困難な作業であるため，まずは雛型に沿った興味オントロジを自動生成し，それに基づいた情報推薦などのサービスを普及した上で，興味オントロジの改変を実現することを考えてるためである．更に，雛型オントロジは，サービス設計者の意図するドメインから構築していけば良く，サービス拡充に伴い徐々に増やして行けばよい．また，雛型オントロジを設計者が恣意的に作成する事で，設計者の意図するコミュニティ生成やコミュニティ分布の時間推移を取得できる．

3.2　ユーザの興味分布解析と興味オントロジ自動生成

本節では，図 2 に示す雛型オントロジ例を用い，ユーザの興味分布解析と興味オントロジ自動生成手順を述べる．

まず，(1)ping サーバなどを通じ収集した全ブログエントリに対し形態素解析を行いインデックスを作成する．ここで，収集されたブログエントリは，一意なユーザ ID を持つとする．

その上で，(2) 全ブログエントリを雛型オントロジに対し分類する．分類方法としては，あるエントリ内の記述に雛型オントロジのあるクラス C_i の名前属性があれば，そのエントリを C_i に分類し，また，C_i に所属するインスタンス $I_i(\in C_i)$ の名前属性があれば，エントリをクラス C_i のインスタンス I_i に分類する．なお，エントリが複数クラスに分類されても良い．例えば，図 2 において，エントリ内の記述に“ Charlatans ”という文字列がある場合，そのエントリはクラス“ Madchester ”のインスタンス“ Charlatans ”に分類される．

次に，(3) 雛型オントロジを形成する最下層クラス C_l の持つ各インスタンスに対し興味を持つユーザ数を計測する．なお，クラス C_l のインスタンス I_l に興味を持つユーザ数を算出する際，同一ユーザが複数エントリにおいてインスタンス I_l を記述していたとしても，ユーザ数は 1 と計測する．次に上記計測を最下層クラスに対しても実施し，最下層クラスに興味を持つユーザ数を，最下層クラス配下の全インスタンスに興味を持つユーザ数と最下層クラス C_l 自身に興味を持つユーザ数の総和で計測する．この場合も，同一ユーザが複数インスタンスに興味を持っていたり，最下層クラスとそのクラスに所属するインスタンスに同時に興味を持つとしても，ユーザ数は 1 と計測する．このようにしてユーザ数をルートクラスまで再帰的に計測する事で，そのドメインに興味を持つユーザ分布を計測できる．

そして，(4) 分類結果からユーザ ID の一致するエントリの分類体系のみを抽出すれば，そのユーザに対する興味オントロジを生成できる．例として，図 2 にユーザ A のエントリ集合がインスタンス“ stone temple pilots ”や“ New Order ”,“ Farm ”を記述している場合に生成される興味オントロジを示す．

以上をベーシックアルゴリズム (BA) と名づける．

しかしベーシックアルゴリズムでは，例えば図 3 において，クラス“ Madchester ”配下のインスタンス“ Farm ”などの多義語に対しては，Madchester というジャンルのアーチストである“ Farm ”でなく，農場という意味の“ Farm ”を記述するエントリをも，クラス“ Madchester ”のインスタンス“ Farm ”に分類してしまい誤りが多い．そこで，本研究では，オントロジの持つ (1) 同一クラスに所属するインスタンスは同一の性質を持つという特性と，(2) クラス階層の近いクラス間の性質は近く，両者のインスタンス間の性質も近いという特性を利用し，分類誤りを除去するフィルタリングアルゴリズムを 2 種類提案する．

以下，フィルタリングアルゴリズムを説明する．

ベーシックアルゴリズムの手順 (2) を細分化し，(2-1) あるユーザのあるエントリ E_i 内に雛型オントロジのあるクラス C_i に所属するインスタンス $I_i(\in C_i)$ の名前が記述されている場合，そのユーザの蓄積する全エントリに対し，C_i に所属する I_i 以外のインスタンス $I_k\{(I_k) \in C_i\}$ や C_i の記述があるかどうかをチェックする．そして，(2-2) 記述がある場合にエントリ E_i はクラス C_i に所属するインスタンス I_i を話題にするエントリとして分類し，ない場合は誤りとする．図 3 を用い説明すると，“ Farm ”に対する記述が，あるユーザのエントリ E_i に存在し，そのユーザの全蓄積エントリ内に例えば,“ Milltown Brothers ”の記述がある場合，E_i はクラス“ Madchester ”のインスタンス“ Farm ”に関するエントリとし分類する．以上をフィルタリングアルゴリズム 1(FA1) と名づける．

更に FA1 よりも分類制約の強いアルゴリズムとし，以下のフィルタリングアルゴリズム 2(FA2) を提案する．FA2 では，FA1 の手順 (2-1) において，同一エントリ E_i 内にクラス C_i に所属するインスタンス $I_i(\in C_i)$ 以外のインスタンス $I_k\{\in C_i\}$ や C_i の記述が存在するかどうかをチェックする．そして，記述がある場合に E_i はクラス C_i に所属するインスタンス $I_j(\in C_i)$ に関するエントリとし分類し，記述がない場合は誤りとする．

更に FA1 と FA2 に対し，オントロジのクラス階層を利用しフィルタリングの調整を行う機構を与える．つまり上記で説明した FA1 や FA2 では，エントリ E_i 内の記述に雛型オントロジのあるクラス C_i に所属するインスタンス $I_i(\in C_i)$ の名前が記

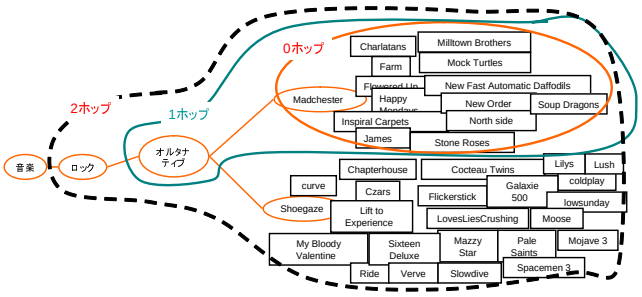

図 3　フィルタリングアルゴリズムにおけるホップ数

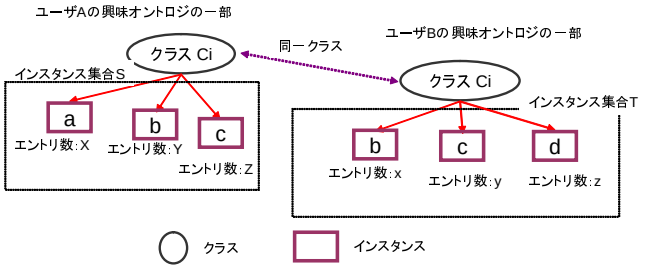

図 4　ユーザの興味オントロジ例

述されている場合，エントリ E_i を分類する際に同一クラス C_i に存在するインスタンス $I_k\{(I_k) \in C_i\}$ やクラス C_i を分類決定要素として利用している．しかし，あるエントリ E_i 内での興味対象は，同一クラスのインスタンスと一緒に現れるだけでなく，近隣のクラスのインスタンスとも一緒に出現する可能性が高いことを考慮し，図 3 のように，ホップ数 0 の時は同一クラスと同一クラスに所属するインスタンスのみを分類決定要素とし，加えてホップ数 1 の場合は親クラスと親クラスに所属するインスタンスを，ホップ数 2 の場合は祖父クラスと兄弟クラス，およびそれぞれに所属するインスタンスまでを分類決定要素とすることでフィルタリングの強さを調整する．

こうして生成した興味オントロジをブログに適用する事で，従来の単純なキーワード検索でなく，オントロジの近似度に基づく意外なエントリ推薦によるコミュニティ形成を支援でき，ユーザ興味を自然と広げる可能性を持つ．更に，クラスや複数クラスを跨がるユーザ分布を基にコミュニティの活性状況を解析できる．

4.　興味オントロジに基づくコミュニティ形成支援

本章では，興味オントロジを利用し，ユーザ興味に即したエントリを推薦する事を通じたコミュニティ形成支援手法の提案を行う．まず，あるクラス C_i に興味を持つユーザに対し，同一クラス C_i に興味を持つユーザやそのエントリを推薦する手法について説明する．次に，興味オントロジ間の近似度を計測し，近似度の高いユーザの興味クラスやインスタンスに対するエントリを推薦する手法を説明する．

4.1　同一興味クラスに対するエントリ推薦

同一クラスに興味を持つユーザを解析することは，3.2 節で提案したベーシックアルゴリズムの手順 (3) において実現している．本節では，ユーザ毎に，同一クラスに興味を持つ他ユーザのエントリをスコアリングして推薦することを試みる．

図 4 に示すユーザ A と B のクラス C_i に関する興味オントロジを用い説明する．ユーザ A はインスタンス集合 S とし，インスタンス a，b，c に興味を持っており，a に対しては X 個，b に対しては Y 個，c に対しては Z 個のエントリを蓄積しているとする．同様に，ユーザ B はインスタンス集合 T とし，インスタンス b，c，d に興味を持っており，b に対しては x 個，c に対しては y 個，d に対しては z 個のエントリを蓄積している．このとき，ユーザ A からみたユーザ B の適切さ $S_u(AB)$ を，興味の一致するインスタンス数 $|S \cap T|$，興味の一致するインスタンスに対する記述エントリの比率 $\frac{y+z}{X+Y}$ および，2 パラメータの重みを決定する関数 $f(X)$ を用い，次式（1）により与える．

$$S_u(AB) = N(|S \cap T|) \times f(N(\frac{x+y}{Y+Z})) \quad (1)$$

ここで，C_i に興味を持つ全ユーザを母集合とし，重みを考慮した関数 N として，平均 $\overline{x}$，分散 ${\sigma_x}^2$ をもつ式 (2) を与える．

$$N(x_i) = \frac{x_i - \overline{x}}{\sigma_x} \quad (2)$$

更に，ユーザ A からみたユーザ B の持つエントリ E_i の適切さ $S_{E_i}(AB)$ を，雛型オントロジ全体に対する E_i の所属クラス数とインスタンス数の和 $Num(total)$，クラス C_i における E_i の所属インスタンス数 $Num(C_i)$，2 パラメータの重みを決定する関数 $f(X)$，および C_i に対し書き込みを行う全エントリを母集合とし式（2）で与えられる関数 N を用い，以下の式（3）により与える．これは，エントリの中には，他の複数ブログサイトや Web サイトの情報を単純にコピーして表示するエントリなどがあり，そうしたエントリは特定のユーザ興味を反映したエントリとはいえずインスタンスを列挙しているだけのエントリとなる可能性が高い点を考慮し，スコアを下げ，C_i が話題の中心であるエントリを優先的にユーザ推薦するためである．

$$S_{E_i}(AB) = S_u(AB) \times f(N(\frac{Num(C_i)}{Num(total)})) \quad (3)$$

その上で，ヒューリスティックな閾値 θ を用い，$S_{E_i}AB > \theta$ となるエントリ集合をユーザ A の興味に即したエントリ集合 $G(E)$ として蓄積する．そして，$G(E)$ の所属するインスタンスを解析することで，ユーザ A の興味オントロジを形成するインスタンス集合 S に対し共起性を持つインスタンス集合 I を解析する．そして，I に関するエントリ集合 $G(I)$ を，ユーザ A の興味に即しかつ意外なエントリとして，ユーザ A に推薦する．

本手法により実現できるコミュニティ形成支援サービスイメージを図 5 を例にあげ説明する．例えば，従来のキーワードベースのエントリ検索では，“ Madchester ”というジャンルに対して，キーワード検索を行っても少数の結果しか得ることができず，“ Madchester ”というジャンルに対するコミュニティはユーザにとって把握しづらい．それに対し，本研究では，“ Madchester ”というジャンルに所属するインスタンスをグループ化することで，“ Madchester ”というクラスを形成する．これにより，キーワードベースでは発見できなかった“ Madchester ”

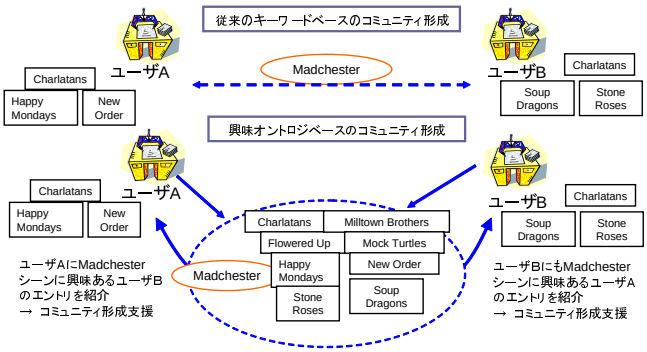

図 5　従来サービスと興味オントロジによる実現サービスの比較

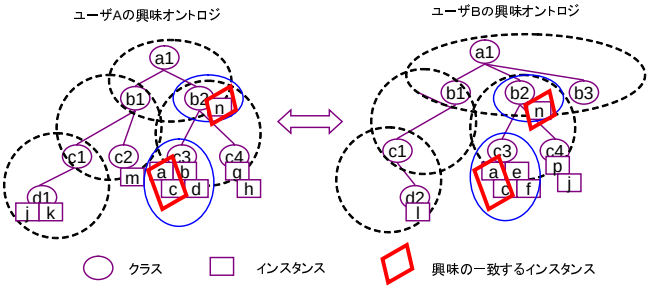

図 6　オントロジ間の近似度計測アルゴリズム

に興味を持つユーザやエントリ集合を発見することができ，更に本節のエントリ推薦アルゴリズムを通じ，“ Madchester ”グループの中でもユーザの興味に即したエントリ紹介を行う．

4.2　オントロジ間の近似度計測に基づくエントリ推薦

本節では，クラスのみでなくトポロジも考慮しオントロジ間の近似度を求め意外なエントリをユーザ推薦する事を提案する．図 6 を例にあげ近似度計測アルゴリズムについて説明する．

(1) オントロジ間で末端クラスを除く共通クラスを分析し，共通クラスを親クラスとし，親子クラスからなるトポロジを抽出する．図 6 では，点線で囲まれた 4 種類が抽出できる．

(2) 次に，オントロジ間で各トポロジを形成する子クラス集合間の近似度を Jaccard 係数に基づき計測する．Jaccard 係数では，集合 X と Y 間の近似度は，$\frac{X \cap Y}{X \cup Y}$ となる．そして，各トポロジの子クラス集合間の近似度を足し合わせ最終的なトポロジの近似度 S_T を計測する．図 6 では 1/3+1/2+0/2+2/2=13/6 となる．

(3) 一方，オントロジ A と B 間で共通クラス間の近似度を計測する．図 6 では 6 種のクラスが共通する．ここで，クラスの近似度をクラスの所属インスタンス集合を用い Jaccard 係数より計測する．そして共通クラス間の近似度を足し合わせ，最終的な近似度 S_C を与える．図では 1/1+2/6=1/2 となる．

(4) そして，トポロジの近似度 S_T とクラスの近似度 S_C および，トポロジとクラスに対する重要度に応じた評価関数 $f(X)$ を用いオントロジ間の近似度 S_O を以下の式（4）で与える．

$$S_O(AB) = S_T + f(S_C) \tag{4}$$

次に，オントロジ間の近似度計測アルゴリズムを基にユーザ A が興味を持つ可能性のある意外な情報推薦法を説明する．

まず，近似度計測アルゴリズムをユーザ A とその他のブログユーザ集合 $u\{\in U\}$ との間で総当りで行う．そして，ヒューリスティックな閾値 δ を用い，$S_O(Au) > \delta$ を満たすユーザグループ G_U を導出し，グループ G_U に属する各ユーザの興味オントロジとユーザ A の興味オントロジの差分クラスとインスタンスを分析する．そして，オントロジ間の近似度が近いにも関わらずユーザ A に存在しないクラス，インスタンスに関するエントリをユーザ A の興味に即している意外な情報として推薦する．

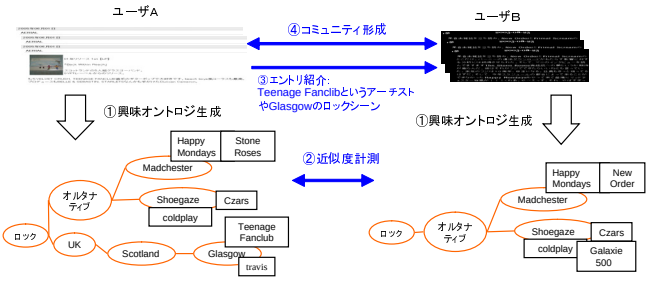

図 7　興味オントロジの近似度計測による実現サービスイメージ

オントロジ間の近似度計測手法に基づくサービスイメージを図 7 を用いて説明する．ユーザ間の興味オントロジの近似度を計測し，近似度の高いオントロジ間で共起するクラスやインスタンスを分析することで，例えば，“ Madchester ”等のクラスや“ Happy Mondays ”等のインスタンスに興味を持つユーザは“ Glasgow ”クラスやその“ Teenage Fanclub ”というインスタンスにも興味を持つ可能性が高い事が分かる．こうしたトポロジが異なるにも関わらず興味を持つ可能性が高い情報を他ユーザのエントリを介して意外な情報としてユーザに推薦できる．

5.　提案手法の実装と評価

本章では，ブログポータルサイト Doblog における大規模データ (約 5 万 5 千ユーザ，160 万エントリ) に対し，興味オントロジの自動生成とユーザ全体の興味分布の検証を行う．また，コミュニティ形成支援プロトタイプの実装を通じ，ユーザ毎の興味オントロジを用いたエントリ推薦の有効性を検証する．

実験にあたり音楽ドメインの雛型オントロジを作成した．作成した雛型オントロジは，goo 音楽(注5)などの Web ポータルの公開情報を参考とし，114 ジャンルをクラスとし，末端クラスに約 4300 のアーチストをインスタンスとして分類し，図 2 に一部示すような雛作成した．図 2 ではクラス階層のみを表示しているが，実際には末端クラスにインスタンスを配置している．なお，各クラス，インスタンスには名前属性を複数与えている．例えば，“ verb ”というインスタンスには，“ ヴァーヴ ”と“ verv ”という 2 つの名前属性を与える．このようにし，4300 のアーチストに対し約 7600 の名前属性を与えた．

そして，ベーシックアルゴリズム (BA)，フィルタリングアルゴリズム (FA1，FA2) により生成される興味オントロジの精度を測定した．なお，検証方法としては，興味オントロジを構成するクラス・インスタンスに分類されるユーザのエントリを人手で確認した．正解の導出根拠としては，実際にそのクラス・インスタンスの名前属性の記述があるエントリを正解とした．

(注5)：http://music.goo.ne.jp/

表 2　提案手法による興味オントロジの精度 (FA2, ホップ数 2)

	Rock	Jazz･Classic･その他	Total
正解数	911	1440	2351
適合率	911/1001=91.0%	1440/1520=94.7%	2351/2521=93.2%
クラス数	36	78	114
インスタンス数	2133	2158	4291

表 3　1 単語の名前属性を持つインスタンスの分類精度

	Rock	Jazz･Classic･その他	Total
正解	138	97	235
適合率	138/204=67.6%	97/125=77.6%	235/329=71.4%
1単語よりなるインスタンスの割合	455/2133=21.3%	458/2158=21.2%	913/4291=21.2%

精度の尺度としては，正解数と分類結果中の正解数の割合 (適合率) を用いた．正解数が多いほど，ユーザが記述した興味がカバーされるが，適合率が低いと興味オントロジに誤りが含まれ，ユーザへの推薦情報の信頼性が落ちるため本稿では適合率向上がまず必須と考える．更に，1 単語から形成される名前属性を持つインスタンスやクラスが特に多義語となる可能性が高い事を考慮し，そうしたインスタンスやクラスにフィルタリングアルゴリズムを適用した．なお，ブログエントリのインデックス作成には全文検索エンジン Namazu(注6)を用いた．

FA2 の精度を，分類結果の 1/10 のデータをランダムに抽出し検証した（表 2）．これによると，適合率は 90%以上にまで達しており，フィルタリングが，エントリ分類や興味オントロジ生成に効果を持つことが確認できた．また，表 3 に 1 単語の名前属性を持つクラス・インスタンスへの分類精度を示す．こうした語は多義語である可能性が高く適合率が落ちている．

また図 2 に，雛型オントロジ内の各クラスに対するユーザ分布の一部を示す．これによると，末端クラスに所属するユーザ数であっても 200 程度存在する．末端クラスに分類されたエントリ集合を調査すると，そのクラスを特徴付ける語が多く頻出する事が確認できた．例えば，デス・メタル配下にはデスヴォイスという語などが頻出する事が分かった．雛型オントロジ作成においては，サービス毎に適切な粒度設計がポイントになると考えられるため，今後はこうしたクラスの特徴語や実験サービスを通じたユーザ数分析を通じ適切な粒度の定量化を試みる．

更に，フィルタリングアルゴリズムの性能を検証するため，ロックジャンルの 1 単語よりなる名前属性を持つ 1/4 のエントリをランダムに抽出し BA，FA1，FA2(ホップ数 2) の正解数・適合率を比較した．なお検証を公正にするため，FA2 におけるインスタンスへの分類数が 4 以下のものは検証対象外とした．その結果 83 のインスタンスについて，正解数，適合率を得る事ができた．なお，BA に対してはエントリ数が膨大であったためそのうち 17 インスタンスのみに検証を実施している．

図 9 は，83 のインスタンスを横軸に各適合率を示す．また，表 4 に BA における 17 インスタンスに対する正解数，適合率の比較を示す．本結果より，BA，FA1，FA2 の順で適合率が向上することが分かる．また正解数は，FA1 は BA よりそれほど落

(注6)：http://www.namazu.org/

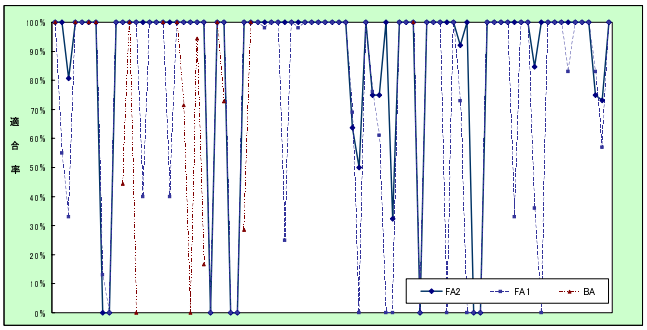

図 9　1 単語の名前属性を持つインスタンスの精度比較

表 4　BA，FA1，FA2 に対する正解数，適合率の比較

	FA2	FA1	BA
正解	14	40	43
適合率	0.7	0.597	0.189

ちないが，FA2 と比較すると FA2 は大きく減少する．そのため今後は，ブログの特徴である時系列の近いエントリやトラックバック等の関連エントリを利用し，FA2 における分類決定要素を同一エントリのみでなく分類決定要素が出現する確率の高い上記エントリまで参照する機構を付加し，適合率を維持しつつ正解数を増やす事を試みる．なお，グラフより FA2 においても適合率をあげることができないインスタンスが 8 個存在し，全体の適合率を下げる事が分かる．そこで FA1 から FA2 にした際，急激に分類数が増えるインスタンスは，ユーザが興味を持ち多数エントリ記述しているにも関わらず，そのインスタンスの分類決定要素と全く共起しないため誤りである可能性が高いと考え，FA1 と FA2 で 10 倍以上結果が増えるインスタンスを自動抽出し分析した．結果，28 個抽出でき，その内 5 個が FA2 でも適合率が 0%となる事が分かった．これより，FA1 と FA2 の比較を通じ急激に分類数の増えるインスタンスを分析し，雛型オントロジから削除する事が，適合率向上に有効と考える．

また FA2 に対しホップ数を変化させ，精度を比較した（表 5）．表 5 によると，ホップ 0 と 2 を比較するとホップ 2 が正解数，適合率ともに良くなる．さらに，ホップ 0 と 4 を比較すると，ホップ 4 の方が若干正解数は増えるが，適合率は下がる．これは，今回用いた雛型オントロジが，例えばクラス“ メタル ”を例に挙げると“ メタル ”配下の末端クラスに“ 北欧メタル ”や“ ポップメタル ”などがあり，“ メタル ”の親クラスが“ ロック ”であるなど，末端クラスとその親クラス間の概念間の距離と親クラスと祖父クラスの距離が遠くなるように設計されているため，ホップ 0 よりも分類決定要素が多いホップ 2 が最も適合率がよくなり，距離が遠くなるクラスのインスタンスまで分類決定要素に入れるホップ 4 が最も適合率が低くなると考えられる．また，ホップ数 2 と 4 の間で 10 倍以上分類エントリ数に変化がある 1 語からなる名前属性を持つインスタンスを自動抽出し分析した．結果，それらのインスタンスに分類された増加エントリはすべて誤りである事がわかった．例として“ ヨーロッパ ”というクラス“ 北欧メタル ”配下のインスタン

表 5　ホップ数変化による興味オントロジの精度比較

Rock	正解数	適合率	検証数/総分類エントリ数	総分類エントリ数
ホップ0	475	475/533=89.1%	0.05	10662
ホップ2	495	495/544=91.0%	0.05	10886
ホップ4	477	477/557=85.6%	0.05	11133

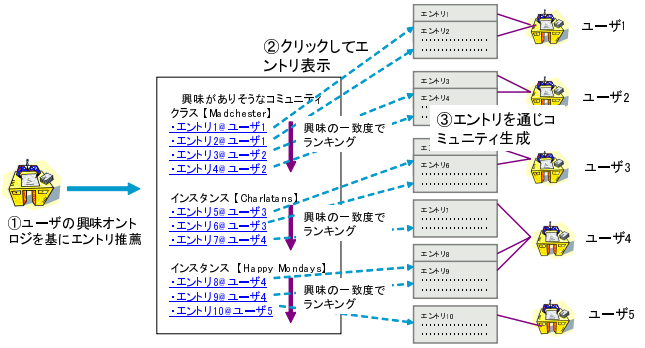

図 10　コミュニティ形成支援プロトタイプ

表 6　推薦エントリの精度

正解数	適合率
7164	7164/7552＝94.9%

スを調べると，ホップ 4 では，クラス“ メタル ”の親クラスであるクラス“ ロック ”配下の“ アダルトコンテンポラリー ”配下のインスタンスも“ ヨーロッパ ”の分類決定要素に使われるようになるため，ヨーロッパ等の“ ヨーロッパツアー ”の話題などと共起してしまうような語を名前属性とするインスタンスの場合，誤りが急激に増える事がわかった．これより，ホップ数を変化させた際，急激に分類数が増えるインスタンスは，そのホップ数における分類は誤りである可能性が高いため，興味オントロジ生成に利用しない機構を付加する事で，正解数を増やしつつ適合率を維持できると考える．

更に，コミュニティ形成支援の検証として，図 10 に示すコミュニティ形成支援プロトタイプを作成し検証を行うことで，ユーザの興味オントロジに即し推薦されるエントリの精度を検証した．推薦精度の尺度としては，推薦エントリ中にユーザ毎の興味オントロジを形成するクラス・インスタンスに関する記述があるエントリを正解とし，正解数と推薦結果中の正解の割合 (適合率) を用いた．なお今回，クラス毎のエントリをユーザ毎にランキングするため式 (1) を用いている．結果を表 6 に示す．表によると，適合率は 95%近くにまで達することが分かり，今回生成した興味オントロジに基づきユーザ毎に興味に即したエントリ推薦ができることが確認できた．

6.　結論と今後の課題

本稿では，ブログユーザのエントリからユーザの興味オントロジを自動生成し，オントロジ間の近似度計測に基づく興味に即した意外な情報推薦やコミュニティ形成を提案した．そして，ブログポータル Doblog の大規模データを基に高精度な興味オントロジ生成やコミュニティ解析およびコミュニティ形成の実現性を確認した．

今後，4.2 章で述べたようなユーザ毎の興味エントリの統計処理に基づく意外なエントリ推薦方法を検証する．そして，提案するコミュニティ形成支援実験サービスを実施し，(1) 意外なエントリ情報推薦によるコミュニティ活性化の有効性と (2) コミュニティのユーザ分布の時間変化の検証を進める．更に，ユーザのブログ閲覧履歴を用いた興味オントロジの拡張手法についても検討を深める．

謝辞

本研究の検証は，株式会社 NTT データのブログポータル Doblog のデータを利用させて頂いている．データ提供とコミュニティ形成サービスのブレインストーミングに快くご協力頂きました Doblog チームには大変お世話になりましたことを感謝致します．

文　　献

[1] Adar, E., Zhang, L., Adamic, L. and Lukose, R. M.: Implicit Structure and the Dynamics of Blogspace, *WWW 2004 Workshop on the Weblogging Ecosystem* (2004).
[2] Berners-Lee, T.: An attempt to give a high-level plan of the architecture of the Semantic Web (1998).
[3] Chakrabarti, S., van den Berg, M. and Dom, B.: Distributed Hypertext Resource Discovery through Examples, *Proceedings of the 25th VLDB*, pp. 375–386 (1999).
[4] Jung, J. J., Yoon, J. S. and Jo, G.: Collaborative Information Filtering by Using Categorized Bookmarks on the Web, *INAP* (2001).
[5] Kumar, R., Novak, J., Raghavan, P. and Tomkins, A.: On the bursty evolution of Blogspace, *Proceedings of the twelfth international conference on World Wide Web (WWW2003)*, pp. 568–576 (2003).
[6] Maedche, A. and Staab, S.: Measuring Similarity between Ontologies, *In Technical Report, E0448, University of Karlsruhe* (2001).
[7] Noy, N. F. and Musen, M. A.: Anchor-PROMPT: Using Non-Local Context for Semantic Matching, *In Proceedings of the Workshop on Ontologies and Information Sharing at the International Joint Conference on Artificial Intelligence (IJCAI)* (2001).
[8] 大向一輝, 武田英明: Semblog: RDF メタデータによる Web 情報の共有支援プラットフォーム, The 18th Annual Conference of the Japanese Society for Artificial Intelligence (2004).
[9] 奥村学, 南野朋之, 藤木稔明, 鈴木泰裕: blog ページの自動収集と監視に基づくテキストマイニング, 第 6 回人工知能学会セマンティック Web とオントロジー研究会 (2004.7).
[10] 佐保田圭介, 波多野賢治, 宮崎純, 植村俊亮: ブックマークの階層構造を考慮した協調フィルタリングによる Web ページの推薦手法, Vol. 3, No. 1, DBSJ Letters (2004).
[11] 神崎正英: セマンティック・ウェブのための RDF/OWL 入門, 森北出版株式会社 (2005).
[12] 総務省:　ブログ・SNS の現状分析及び将来予測, http://www.soumu.go.jp/s-news/2005/050517_3.html (2005).
[13] 橋本大也: ライトウエイト・メタデータの応用事例とその可能性, 第 10 回人工知能学会 SIGSWO 研究会　招待講演資料 http://www.ringolab.com/note/daiya/archives/003614.html (2005).
[14] 谷口智哉, 松尾豊, 石塚満: Blog コミュニティの抽出と分析, 第 6 回人工知能学会 SIGSWO 研究会 (2004).