クラスタクソノミを活用したブログ記事のポジティブ・ネガティブ・感性判定技術

中辻　真[†]　吉田　誠[†]　平野　美貴[†]

† 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,yoshida.makoto,hirano.miki}@lab.ntt.co.jp

あらまし　近年，ユーザ興味などのユーザコンテクストを用い，ユーザに対しその場その時の状況に沿った適切なコンテンツを NW を介し提供するサービスが注目されている．また，ユーザ興味を発信する手段としブログ利用が目覚しい．そこで著者らは，コンテクストの 1 要素としてユーザ興味を自動抽出するため，ブログに着目し，プロバイダが与えるサービスオントロジへユーザの蓄積記事を分類する事で，記事へその内容を端的に示すタグを自動的に与えるオートタギング手法と，ユーザ興味をクラス階層化した興味オントロジを自動抽出する手法を提案してきた．本稿では興味オントロジの抽出精度を向上するため，ブログ記事からユーザ興味を抽出するにあたり，ユーザがブログ内で記述しているインスタンスがユーザにとってポジティブかネガティブか，どのような感性で記述されているかをインスタンスの所属するクラス知識を用い判定する PNM 判定技術を導入する．PNM 判定を実現するためには，ポジティブ・ネガティブ・感性辞書の自動生成が核となるが，これまでの PNM 判定技術では，クラスに関係なく同一辞書が生成されてきた．それに対し本稿では，サービスオントロジに属する各クラスの特徴に沿ったポジティブ・ネガティブ・感性辞書を自動生成する手法を提案する．そして実ブログデータを利用した実験により，ユーザの興味対象となるクラスに沿った PNM 判定を高精度で実現可能であることを示す．

キーワード　感性判定，PN 判定，オントロジ，情報推薦，セマンティック Web，ブログ

Positive, Negative and Mood Analysis of Blog Entries based on Class Taxonomy.

Makoto NAKATSUJI[†], Makoto YOSHIDA[†], and Miki HIRANO[†]

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

Abstract　Recently, services that use the user contexts such as the user interests and offer appropriate contents along the situation at the that time of the place for the user are paid to attention. Furthermore, the use of blog is remarkable as the means to publish the user interests. In order to automatically extract user interests as one element of user contexts, we proposed interest ontology generation technique that classifies user blog-entries to domain ontology and extracts interest ontology which expresses user interests in detail as a class hierarchy. In this paper, we introduce PNM analyzer that analyzes whether the instances that the user described in blog-entries are positive or negative or by what mood they are described by using the knowledge of class taxonomy. The key of PNM analyzer is creating the positive, negative and mood dictionaries. However, the current PNM analyzer creates the same dictionary regardless of the class knowledge of instances described in blog-entries. We propose the method that automatically creates those dictionaries along the class taxonomy of service ontology. We evaluated the performance of our proposed method based on actual blog entries on blog portal Doblog and music service ontologies.

Key words　Mood analyzer, PN analyzer, ontology, Information Retrieval, Semantic Web, Blog

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 [15]．また，Amazon [1] や Last.fm [3] などユーザによる商品の購買・視聴履歴を基にユーザの興味対象をユーザプロファイルとして自動構築し，プロファイルに基づく商品の推薦を行うサービスも登場するなど，ユーザの多様な発信情報に基づく極め細やかな情報推薦を行うサービスへの関心も高まっている．そして，この種の情報流通サービスは，興味の近いユーザコミュニティの発信情報を通じ，ユーザの興味対象に対する評判を参考にしつつ，各自の興味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，Google [2] などの Web ページ検索エンジンや，RDF Site Summary（RSS）(注1) 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ため，大量に発信されるブログ記事から自身の興味に沿った情報を掲載する記事を適切に選択するのは困難なことが多い．つまり，ユーザは自身の興味に従う情報が掲載されたブログ記事を検索するためには，検索のたびに，自身の興味に即した検索目的語を適切に構成する必要があり，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．

こうした問題を解決するため，著者らは，ユーザの興味対象としてユーザが興味を持つキーワードのみでなく，そのキーワードの背景となる概念（クラス）情報をも保持する興味オントロジをユーザのブログ記事から自動構築し，興味オントロジに基づく情報検索を提供する試みを行ってきた [9, 14]．具体的には，音楽や映画などのサービスプロバイダが与えるサービスオントロジへユーザの蓄積記事を分類する事で，記事に対しその話題対象であるインスタンスとインスタンスの背景クラスを自動的にタグ付けするオートタギング手法と，ユーザ興味をクラス階層化した興味オントロジを自動抽出する興味オントロジ生成手法を提案してきた．そして，興味オントロジ間の近似度を計測し，近似度が高いオントロジ間で一部クラス階層の異なるクラスを検出し，そのクラスに属する記事を，意外な興味記事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検証する実験を実ブログサイト上で実施してきた [9](注2)．

しかし，興味オントロジをより詳細に抽出するためには，ブログ記事からユーザ興味を抽出するにあたり，記事内での興味対象をユーザがどのような感性で記述しているかを判定し，例えば，ネガティブな興味対象を興味オントロジより除去する必要がある．また，ブロガー全体の興味インスタンスに対する感性情報を抽出し，インスタンスへ自動的にタグ付けすることができれば，感性に基づくブログやインスタンスの検索にも応用でき，ユーザによるブログ・コンテンツ検索の際の参考情報として利用できると考えられる．

そこで，本研究ではサービスオントロジにおけるクラスタクソノミに沿った意外記事推薦に加え，ユーザがブログ記事で記述している興味対象に対しポジティブなのか，ネガティブなのか，どのような感性で記述しているのかを判定する PNM 判定技術を導入する．なお，本研究における感性は，ポジティブ，ネガティブ，どちらでもないの 3 種類があると考えている．そのため，どのような感性で記述しているのかの判定結果にはポジティブ，ネガティブな記述も含む．

PN 判定を実現するためには，ポジティブ・ネガティブ辞書の自動生成が核となるが，これまでの PN 判定技術では，話題インスタンスの所属クラスに関係なく同一辞書が生成されてきた．それに対し本稿では，インスタンス決めうちではなく，インスタンスに付随するクラス情報まで考慮することで，クラスとして特徴的なポジティブ・ネガティブ・感性辞書を自動生成し PNM 判定の精度を向上することを試みる．更に，実ブログデータを用い提案手法の検証を行い，本提案が，ユーザの興味対象となるクラスに応じた PNM 判定を高精度で実現可能であることを確認した．なお，本研究におけるサービスオントロジは，同様の性質を持つインスタンスをグループ化しその性質を表現したクラス階層を持つのに対し，辞書は単純に単語が登録されクラス・インスタンスの関係を持たない．

以下，2. 章では，本論文の背景となるブログ検索の概要説明とその問題点を述べ，関連研究の紹介と本研究との比較を行う．3. 章では，著者らの先行研究として，文献 [9] におけるサービスオントロジに基づくブログ記事へのオートタギングや興味オントロジ生成手法を説明し，4. 章において，サービスオントロジにおけるクラスタクソノミを利用し，クラス毎の PNM 辞書の自動生成手法について述べ，5. 章では，実ブログデータを用い提案手法の検証を行い，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

ブログの特徴として，ブログ記事のタイトル，アドレス，要約などに対するメタデータ記述である RDF Site Summary（RSS）を用いた情報検索について説明する．ユーザは，ブログ記述の際に自動的に記事に付与される RSS を，ブログの更新情報を集め提供しているサーバである ping サーバに登録することで，記事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．そして，ブログを閲覧するユーザは公開 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．

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，ブログ検索に際しては，ユーザが検索キーワードを構成しないといけない事に変わりはない．

(注1)：http://blogs.law.harvard.edu/tech/rss

(注2)：http://music.doblog.com/で 2006 年 8 月から 12 月まで実施

一方，Amazon [1] や last.fm [3] では，購買履歴や楽曲の再生履歴からユーザ興味を自動的にプロファイル化し，協調フィルタリング技術 [10] を適用し，ユーザ毎へアイテムの推薦を実施する．しかし，協調フィルタリングでは，ユーザ興味間の近似度の計算を，ユーザの興味アイテム等のインスタンスから構成される興味ベクトル間の近似度計算のみを基に行っており，インスタンスに付随するクラス知識を利用していない．そのため結果として，ユーザに同種のクラスに属するインスタンスを推薦をすることが多い．また，抽出されたユーザ興味をブログ検索へ適用する試みは行われていない．

こうした問題に対し著者らは，ユーザが日々自身の興味対象を記述しているブログからユーザプロファイルを興味オントロジとして自動抽出する手法を提案してきた [9]．具体的には，ブログでは例えば音楽と映画など複数サービス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，各サービスドメイン毎の情報をクラス階層として記述したサービスオントロジを予め用意し，それに対し個人のブログ記事を分類しタグ付けしていく事で，ユーザ毎の興味オントロジをサービスドメイン毎に分離しトップダウン的に生成する．その上で，ユーザは興味オントロジを，自身の興味により適合したものへとボトムアップ的に更新する．このようにして興味オントロジを生成した上で，複数ユーザの興味オントロジ間の近似度を計測し，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記事を，意外な興味記事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実験サービスとして実施してきた．

しかし，ユーザ興味を詳細に抽出し，ブロガーによる情報検索に再利用するためには，ブログ上のユーザの興味対象がどのような感性で記述しているのかを判定し，少なくとも明示的にネガティブに記述されている対象は興味オントロジから除去しなければならない．また，有名な対象を記述している場合は，ブログ記述をしていてもニュース的な情報に過ぎないことも多く，興味としては低い場合もある．PN 判定は，そうした興味の強さを測る指標の一つとして期待できる．図 1 に本研究におけるブログからのユーザ興味抽出・意外情報推薦，インスタンスへの感性情報付与に対するシステム構成の概要を示す．本稿は，著者らが過去に文献 [9] などを通じ蓄積したブログからのユーザ興味抽出（3. 章参照）に加え，図における点線部の PNM 判定やブログ解析に基づくインスタンスへの感性情報のタギングを新規に提案する．

こうした PN 判定に関する研究 [4, 11, 12] として，文献 [12] では，記事集合に対し，ポジティブ語やネガティブ語との共起性を基に，フレーズの PN 判定を実施する．そして，記事内のフレーズをチェックし平均的にポジティブであればポジティブな記事と分類し，それ以外はネガティブと判定する PN 判定を実施し，比較的高精度を得ている．しかし映画記事に対する検証では，記事内における”more evil”などのフレーズは，ホラーなどのジャンルによってはポジティブな記述であることも多いにも関わらず，映画全体ではネガティブフレーズとなるなど，記事内での話題対象となるインスタンスの背景クラスまで考慮

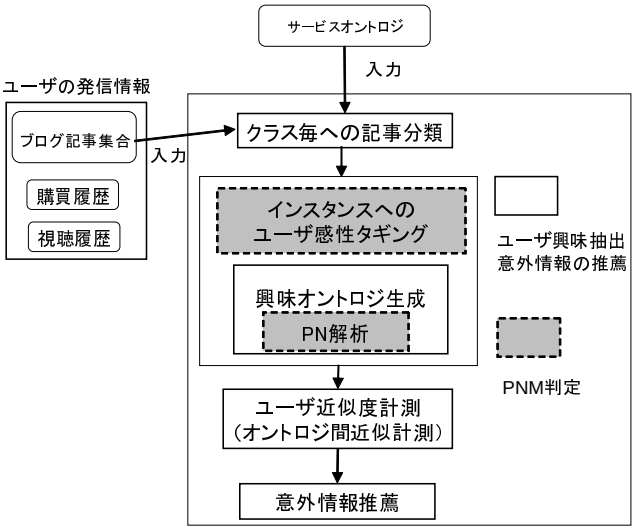

図 1　本研究のシステム構成の概要

した PN 判定を実施できてるとはいえない．

一方，ブログ記事から感性情報を抽出する研究としては，Mishne らは文献 [8] で，ある一定期間におけるユーザ全体の感性の動向を高精度に予測している．方法としては，ブログ記事を投稿する際にユーザが記事の持つ感性語をタギングすることができるブログサービスである LiveJournal [5] で公開されているユーザのブログ記事をコーパスとし，感性語に対し影響力を持つ特徴語を感性語との共起性を基に学習し，感性語や特徴語で感性情報をモデリングした上で，感性語に合致するブログ記事を抽出している．また，文献 [7] では，同種のコーパスを用いブログ記事毎の感性予測を実施しているが，これについては，精度は低い．このように，単一ブログ記事では高精度な PNM 判定はできず，複数ブログ記事集合から PNM 判定を行うのが適切と考えられる．このことを考慮し，我々はインスタンス毎に対するブログ記事では記事数が少なくなり，高精度な感性辞書の生成を阻害する上，インスタンスの背景となるクラス知識を利用できないと考え，インスタンスの背景となるクラス毎に分類された記事集合を用い各クラスに特徴的な PNM 辞書を生成した上で，辞書に登場する単語が，記事内での話題対象に対し，ポジティブ，ネガティブまたは感性表現としての効果を与えるものであるとし，分類を行うというアプローチを取っている．

3.　興味オントロジの生成

本章では，文献 [9] で提案してきたユーザの蓄積ブログ記事を，音楽や映画といったサービスドメイン毎のサービスオントロジへ分類することによるブログ記事へのオートタギング手法と興味オントロジ生成手法を紹介する [9, 14]．本章で提案する方法により得られるサービスオントロジ内のクラスごとへの記事分類結果を利用することで，クラス毎に特徴的な PNM 辞書を生成できる．

3.1　サービスオントロジの設計

本節では，OWL(Web Ontology Language) [6] を基にオントロジの説明をした上で，サービスオントロジの設計法を説明する．

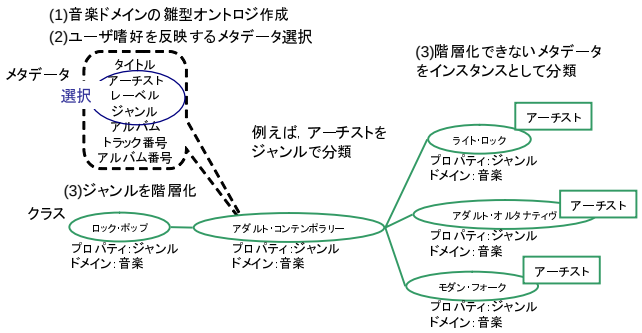

図 2　サービスオントロジ構築手順

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を，特定のクラスとともに使用することでクラスの特徴を詳細に記述することができる．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 [6].

OWL を利用すれば，サービスオントロジを詳細に設計できるが，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．そのため，本研究ではまずはサービスオントロジを，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），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シンプルなオントロジとして設計し，ユーザの興味オントロジは，サービスオントロジへのユーザエントリ分類を通じ自動生成する．幸い，goo 音楽(注3)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，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サービスオントロジを構築すればよい．

以下，サービスオントロジの設計手順を図 2 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サービス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) 選択したメタデータの中で，ユーザ興味を直接反映するものをインスタンスと捉え，その他のメタデータでインスタンスを分類することで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特徴付けるプロパティとしてクラス階層間で継承する．例えば，音楽に関するユーザ興味を抽出する際，アーチストなどをインスタンスとし，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する．

（注3）：http://music.goo.ne.jp/

3.2　ブログへのオートタギングと興味オントロジ抽出手法

図 3 に示すサービスオントロジ例を用い，ブログ記事へのオートタギングと興味オントロジ抽出手順を説明する．

(1) まず，ping サーバなどを通じ収集した全ブログ記事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，収集されたブログ記事は，一意なユーザ ID を持つとする．

(2) その上で，全ブログ記事をサービスオントロジに対し分類することで各記事へその内容を端的に示すタグを自動的に与える．タギング方法としては，ある記事内の記述にサービスオントロジのあるクラス C_i に所属するインスタンス $I_i(\in C_i)$ の名前属性があれば，その記事はクラス C_i のインスタンス I_i に関するものとしてタギングする．なお，記事が複数クラスに関するとしてタギングされても良い．例えば，図 3 において，記事内の記述に“ Charlatans ”という文字列がある場合，その記事はクラス“ Madchester ”のインスタンス“ Charlatans ”に関するものとしてタギングされる．

(3) 次に，サービスオントロジを形成する最下層クラス C_r 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_r のインスタンスに興味を持つユーザ数を算出する際，同一ユーザが複数記事において同一インスタンス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_r 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 の一致する記事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 3 にユーザ A の記事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．

最後に，(5) 自動抽出された興味オントロジをユーザが閲覧し，興味インスタンスの追加や，興味の無いインスタンスを削除することで，各個人に適切な興味オントロジを生成する．

以上をベーシックアルゴリズムと名づける．

3.3　分類誤りのフィルタリング手法

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 ”を記述する記事をも，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提案する．以下，フィルタリングアルゴリズムを説明する．

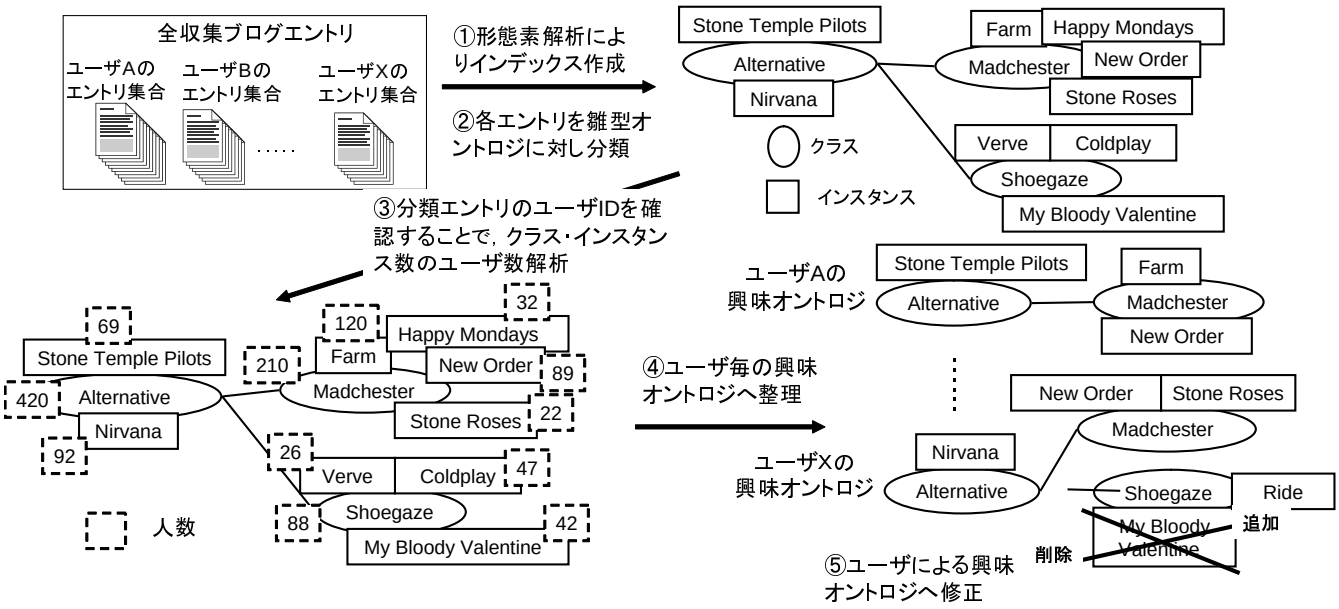

図 3　興味オントロジ自動生成手順

ベーシックアルゴリズムの手順 (2) を細分化し，(2-1) あるユーザのある記事 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，E_i と関係の深い記事として，時系列の近い記事などに対し，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 を話題にする記事として分類し，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，あるユーザの記事 E_i に存在し，E_i と時系列の近い記事内に例えば，“ Happy Mondays ”の記述がある場合，E_i はクラス“ Madchester ”のインスタンス“ Farm ”に関する記事とし分類する．

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記事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．本章で述べた興味オントロジの生成や意外な情報推薦の精度については，文献 [9] で検証し，有効性を確認してきている．

4.　クラスタクソノミを活用した PNM 判定技術

本章ではクラス毎に分類されたブログ記事の話題対象となるインスタンスに対するユーザの興味がポジティブなのかネガティブなのか，どのような感性で対象インスタンスを記述しているのかを判定する PNM 判定技術について手順を追って説明する．まず，クラス毎の特徴を反映した PNM 辞書を生成する辞書生成アルゴリズムについて述べ，生成辞書に沿って記述対象の PNM 判定をする．

4.1　PNM 辞書の生成

ブログ記事の PNM 判定には，PNM 辞書の生成が核となる．しかし，これまでの研究では話題となるインスタンスに対する辞書の生成が中心であり，インスタンスの所属するクラス知識を利用した辞書の生成は試みられていない．例えば,「とんがった」という語は，ロッククラスにおいては，ポジティブな意味を持つとしても，クラッシクやオペラクラスにおいては，ネガティブな意味を持つ場合もある．こうした各クラスの特徴を反映した語の分析は，そのクラスの特徴を把握するためには重要である．そこで，本節では 3.2 章の手法に従いサービスオントロジに沿って分類された記事集合を用い，各クラスに特徴的な PNM 辞書を生成する．以下，PNM 辞書の生成手順を述べる．

4.1.1　ベース辞書の生成

各クラス毎に特徴的な PNM 辞書を自動的に生成するため，まずポジティブ，ネガティブ，感性それぞれに対し，ベースとなる辞書（ベース辞書）を人手で生成する．ベース辞書に登録する語としては，品詞として形容詞や形容動詞であり，かつ多様なクラスに汎用的に適用可能な形態素を選択する．例えば,「素晴らしい」という形態素は多数のクラスにおいて一般的にポジティブな意味を持つが,「甘い」という形態素は，ポジティブな意味を持つ場合もネガティブな意味を持つ場合もある．こうした形態素をベース辞書に登録すると，クラス毎に生成される PNM 辞書の精度に影響を与えるためである．

4.1.2　PNM 辞書の自動生成

次に，3.2 章で各クラスに分類された記事集合を用い，各クラスの特徴に沿った PNM 辞書を自動生成する．以下，辞書自動生成の手順を図 4 を用い説明する．

(1) クラス C_i とその配下に属するクラスからなるクラス集合 $S(C(C_i))$ に分類された記事集合 $S(E(C_i))$ に所属する記事 $E_i(\in S(E(C_i)))$ に対し，$S(C(C_i))$ に所属するインスタンス $I_i(\in S(C(C_i)))$ に対する記述箇所をチェックする．そして，その記述箇所の前後 X 個の形態素に対し，その基本形と品詞情報をチェックし，ポジティブベース辞書内の形態素と一致するものがあるかどうかをチェックする．そして，ある場合に限りクラス集合 $S(C(C_i))$ に特徴的なポジティブ辞書 $P(C_i)$(P 辞書)

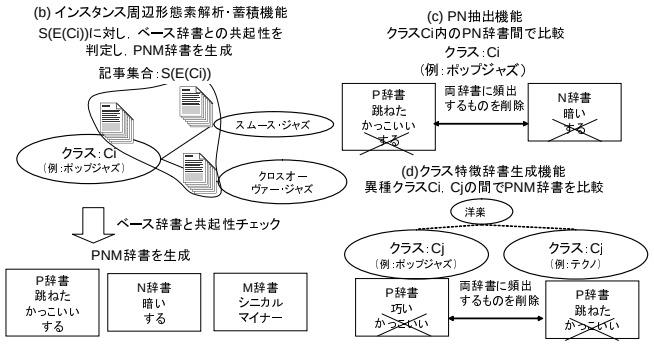

図 4　PNM 辞書の生成手順

として蓄積する．これを，$S(E(C_i))$ に所属する全記事に対し実行する．また，蓄積にあたり，各形態素の基本形の出現回数も保持する．これを，ネガティブベース辞書，感性ベース辞書に対しても実行し，それぞれに対し蓄積された特徴的な辞書を $N(C_i)$(N 辞書)，$M(C_i)$(M 辞書) とする．

(2) 生成されたポジティブ特徴辞書 $P(C_i)$ は，単純にポジティブベース辞書内の登録形態素と共起して出現した形態素を登録しただけであり，必ずしもポジティブ・ネガティブを判別するために利用できる辞書であるとは限らない．そこで，$P(C_i)$ に対し，ポジティブベース辞書とのみ頻出する形態素を抽出するため，$P(C_i)$ と $N(C_i)$ で登録された形態素 $m(\in P(C_i), \in N(C_i))$ に対し，式 (1) を実行し，ヒューリスティックな閾値 α を下回る場合は，m は両辞書に頻出する語でありポジティブでもネガティブでもないと捉え，削除する．ここで，形態素 m の $P(C_i)$ における出現回数を $|m\in P(C_i)|$ とし，$N(C_i)$ における出現回数を $|m\in N(C_i)|$ とする．同様の処理をネガティブ特徴辞書 $N(C_i)$ に対しても実行する．この処理を経て $P(C_i)$ と $N(C_i)$ に残った形態素は，ポジティブベース辞書とのみ頻出する形態素，ネガティブベース辞書とのみ頻出する形態素を集めたものであるため，PN 判定に有効と考える．

$$\frac{\frac{|m\in P(C_i)|}{\sum_{n\in P(C_i)}|n\in P(C_i)|}}{\frac{|m\in P(C_i)|}{\sum_{n\in P(C_i)}|n\in P(C_i)|}+\frac{|m\in N(C_i)|}{\sum_{o\in N(C_i)}|o\in N(C_i)|}} < \alpha \qquad (1)$$

(3) 次に，各クラスの特性のみを反映したポジティブ特徴辞書を生成するため，異なるクラス集合 $S(C(C_j))$ に対し生成された $P(C_j)$ と $P(C_i)$ を比較する．ここで，クラス C_i に対し特徴的な辞書を構築することを狙うため，クラス集合 $S(C(C_i))$ とクラス集合 $S(C(C_j))$ は互いに疎な関係であるとする．比較方法としては，手順 (2) と同様に，$P(C_i)$ と $P(C_j)$ で登録された形態素 $m(\in P(C_i), \in P(C_j))$ に対し，式 (2) を実行し，ヒューリスティックな閾値 β を下回る場合，形態素 m は複数クラスの辞書に跨って出現し一般的にポジティブな形態素であると捉え，$P(C_i)$ から削除する．これを，ネガティブベース辞書，感性ベース辞書に対しても実行する．

$$\frac{\frac{|m\in P(C_i)|}{\sum_{n\in P(C_i)}|n\in P(C_i)|}}{\frac{|m\in P(C_i)|}{\sum_{n\in P(C_i)}|n\in P(C_i)|}+\frac{|m\in P(C_j)|}{\sum_{o\in P(C_j)}|o\in P(C_j)|}} < \beta \qquad (2)$$

ここで，クラス C_i と C_j がクラス階層として近い場合，つまり，意味的に近い場合，各クラスにより特徴的な語が情報集合に残るが，残る単語数が少なくなる．一方，クラス C_i と C_j がクラス階層として遠い場合，つまり，意味的に遠い場合，より一般的な語のみが情報集合から削除され，残る単語数は多い．

(4) こうして生成された特徴辞書をベース辞書に追加し，新たなベース辞書とし，再度 (1) から (3) を繰り返し，特徴辞書内に登録される形態素の数を増やしていく．

このようにして，クラス毎に特徴的な辞書を生成する．

5. 提案手法の実装と評価

本章では，著者らが過去に実施した，ブログ記事へのオートタギングと興味オントロジ自動生成に関する実験 [9] で得られたデータを基に，辞書の自動生成手法および，生成辞書による PNM 判定の検証を行う．

5.1 実験で用いたデータセット

本研究では，ブログポータルサイト Doblog(注4)における実際の記事データ (約 5 万 5 千ユーザ，160 万記事) を 3.2 章で提案する手法に基づき，洋楽サービスのサービスオントロジに分類した際の結果のうち，ポップジャズクラスに分類された 2067 件の記事とテクノクラスに分類された 1028 件の記事をを検証用データとして用いた．そして，4.1 章の PNM 辞書生成手法に従い各記事からポジティブ，ネガティブ，感性に沿った形態素を抽出した．そして，そのうち 1/4 形態素に対し，各形態素が文書内で記述されているポップジャズとテクノクラスのインスタンスに対し適切な効果を持っているかを検証した．今回の検証で用いたテクノとポップジャズの記事は，図 4 にイメージを示すように互いに素な階層関係を持ち，下位クラスの記事を集約して保持している．

5.2 検 証 方 法

次に生成された PNM 辞書を用い，文書の PNM 分類を実施する際の検証の基準を説明する．

ブログからユーザ興味を抽出する際に重要なこととして，ユーザが記述対象に対して明示的にネガティブな記述をしている場合は，興味オントロジからその対象を削除することである．そのためにも，インスタンスをネガティブに表現する形態素を発見することは重要である．また，例えば「テクノのアーチスト A は，シニカルで近未来的な視線と方向性は今聴いても全く古さを感じさせない.」という文章例のように，対象「テクノ」の持つ感性的な情報を「シニカル」などで表現しているのであるが，ポジティブに記述しているのかネガティブに記述しているのか単一の形態素のみでは明確に判断できないことも多い．さらに，例えば「素晴らしい味付けのハンバーグを食べながら，アーチスト A の音楽を聴いた.」や「大麻所持という恐ろしい罪を犯し逮捕されたアーチスト B は … 」という文章例のように，ユーザの行う多様な記述の中にアーチストの情報が少しだけ掲載されているということも多く,「素晴らしい」や「恐ろしい」という PN 語が出現するが，必ずしもアーチストに効果を

(注4)：http://www.doblog.com/weblog/PortalServlet

表 1　PN 判定の精度の尺度

正解	記述インスタンス(クラス)に対しLP(N)辞書内の形態素がポジティブな(ネガティブな)効果を持つ場合
不正解	記述インスタンス(クラス)に対しLP(N)辞書内の形態素がネガティブな(ポジティブな)効果を持つ場合
感性表現	記述インスタンス(クラス)に対しLP(N)辞書内の形態素がポジティブ(ネガティブ)な効果を持たないが感性的な効果は与える場合
無関係	記述インスタンス(クラス)に対しLP(N)辞書内の形態素が無関係である場合

表 2　感性判定の精度の尺度

正解	記述インスタンス(クラス)に対しM辞書内の形態素が効果を持つ場合
不正解	記述インスタンス(クラス)に対しM辞書内の形態素が無関係である場合

持つ形態素ではないことも多い.

そこで，本研究では，ブログユーザが対象インスタンスについて記述をするときは，(1) ポジティブな記述，(2) ネガティブな記述，(3) ポジティブでもネガティブでもないが感性的な表現を与える記述，(4) 多様な記述の中でインスタンスの名前が登場するがポジティブでもネガティブでも感性的な表現も与えない記述の 4 種類とし，自動生成された PNM 辞書に基づき提示される記述対象を表 1 に示す精度の尺度にしたがって検証する.

また，感性辞書の生成に関しては，記事内での話題クラス（インスタンス）を表現する感性語を抽出することが目的であるため，表 2 に示す精度の尺度に沿って検証を進める.

次に，本研究における PNM 判定の判定対象を示す．本研究では以下の 2 パターンに対し判定を実施する.

（1） インスタンスレベル：インスタンスに対し PNM 辞書に登録された形態素がかかっている場合

（2） クラスレベル：インスタンスの所属するクラスに対し PNM 辞書に登録された形態素がかかっている場合

対象（1）は話題としているインスタンスに直接辞書内の形態素がかかっている場合の結果であり，インスタンスに対する PNM 判定を直接的に実施できる．一方対象（2）は，該当インスタンスに対する興味の PNM 判定をインスタンスの所属クラスに対する PNM 判定を基に実施する．対象（2）の例としては，ユーザが複数インスタンスについて記述している場合，特徴辞書内の形態素が該当インスタンスとは別のインスタンスにかかっており，かつそのインスタンスの所属クラスが該当インスタンスの所属クラスである場合などである．こうした対象（2）についての検証は，本研究の特徴であるクラスレベルの PNM 判定の有効性を評価するのに重要と考える.

以上の方法論に従い，検証では，ベース辞書の生成において，4.1 章における手順 (1) の取得形態素数 X を 50 で検証を実施した．これは，X を 100 とした場合と比較した場合，X が 50 の方が PN 判定結果がよかったためである．また，式 (1) における閾値 α は P 辞書の抽出には 0.6，N 辞書の抽出には 0.8 とした．これは，ブログではポジティブな記述の方が多いため，確実にネガティブな形態素を集めるために N 辞書については，閾を大きく設定している．また，式 (2) における閾値 β は 0.6 とした．なお，4.1.2 章の手順 (4) は実施していない.

表 3　クラス知識を利用しない際の結果

		正解	不正解	感性表現	無関係
テクノ	クラスレベル	16/41	9/41	0/41	18/41

表 4　PN 判定の結果

			正解	不正解	感性表現	無関係
ポップジャズ	インスタンスレベル	ポジティブ	45/98	3/98	3/98	47/98
		ネガティブ	7/16	2/16	7/16	0/16
	クラスレベル	ポジティブ	86/98 (88%)	3/98 (3%)	6/98	3/98
		ネガティブ	12/16 (75%)	2/16 (13%)	0/16	2/16
テクノ	インスタンスレベル	ポジティブ	44/102	4/102	10/102	44/102
		ネガティブ	3/26	4/26	6/26	13/26
	クラスレベル	ポジティブ	64/102 (63%)	4/102 (3.9%)	25/102	9/102
		ネガティブ	12/26 (46%)	4/26 (1.5%)	9/26	1/26

5.3　PNM 判定の検証

前節の検証方法に従い，特徴辞書に登録された形態素毎に対し PNM 判定を実施した.

5.3.1　PN 判定の検証

辞書生成法として，4.1 章における PN 辞書生成時に，手順 (1) のみを実施し，手順 (2) と (3) を実施しない場合の結果を表 3 に示す．手順 (2) と (3) を実施していないため，生成される PN 特徴辞書内の形態素はほぼ同じになるため，P 辞書の結果のみ示す．この結果によると，クラスレベルで見てみても生成された P 辞書の精度は低く，このままでは PN 判定に利用できないことがわかる.

次に，今回自動生成した PN 辞書内での形態素が，PN 辞書に沿った効果を持つかどうかを，表 1 の精度の尺度に従い，記事の中での話題対象となるインスタンスと，インスタンスの所属クラスの 2 パターンに対し実施をした．なお，今回の結果はベース辞書と PNM 辞書で生成された特徴辞書がブログ記事内の対象インスタンスの前後 50 個の形態素に現われたものを抽出しており，そうした記事は，各クラスの記事数の約 1/5 であった．表 4 にテクノとポップジャズにおける結果を示す．本結果によるとポップジャズに関しては，インスタンスレベルで見ると正解の割合が 50%程度であるが，クラスレベルで見るとポジティブで 88%，ネガティブで 75%にまで向上しており，PN 辞書に登録された語が，ポップジャズに関して良い効果を発揮していることが分かる．これは，提案手法がクラスレベルの PNM 特徴辞書を解析できているため，インスタンスレベルでは無関係と分類されている結果であってもクラスレベルでは関係のある結果が多く，有効な結果を多く得ることができているからである．また，ポップジャズ，テクノの両クラスに対し，クラスレベルでの不正解率を小さく押さえることができている．ブログ記事からユーザ興味を抽出し，ユーザへ情報推薦などを実施するためには，特に PN 判定における不正解を抑制するすることが重要であるため，本結果は有効であると考えられる．さらに，ブログ記事の中で，クラスをネガティブに判定できる語を比較的高精度で抽出できていることも分かり，ネガティブに記述されているインスタンスを興味オントロジから削除するのに

表 5　感性判定の結果

		正解	不正解
ポップジャズ	インスタンスレベル	20/41	21/41
	クラスレベル	33/41 (80%)	8/41 (20%)
テクノ	インスタンスレベル	71/142	71/142
	クラスレベル	120/142 (85%)	22/142 (15%)

有効であると考えられる．

しかし，テクノに関しては，そのジャンルを形容する感性表現ではあるが，ジャンルに対しポジティブともネガティブとも判定できない語が多く PN 辞書に登録されていることが分かった．例として,「テクノのアーチスト A は，シニカルで近未来的な視線と方向性は今聴いても全く古さを感じさせない.」という記事の一部を基に説明をする．この記事における「シニカル」という語は，テクノクラスに対し感性的な表現を与えてはいるが，テクノに対しポジティブ・ネガティブを表現する語と考えることができない．しかし，こうした感性表現もテクノというジャンルを表現する上で必要になると思われるため，PN 判定には利用できないが，ユーザが感性情報に基づいて記事を検索する際の手助けとしての利用も可能と考える．更に，このような感性表現を行う形態素を辞書から取り除けばクラスレベルの精度はテクノにおいてもポジティブな形態素に対しては，約 78%にまで，ネガティブな形態素に対しても約 71%以上にまで向上することも分かった．

5.3.2　感性判定の検証

次に，今回生成された感性辞書の精度を検証した．検証結果を表 5 に示す．本結果においても PN 判定の結果と同様に，インスタンスレベルで解析を行うと精度は 50%程度であるが，クラスレベルで解析することで精度は 80%から 85%にまで達し，抽出された感性語が記事内での話題対象にクラスレベルで影響を持つことが分かる．本結果も，提案手法がクラスレベルでの感性語の抽出に効果を示すものである．

以上の結果より，記事内での話題対象となるインスタンスの背景となるクラス情報まで考慮することで，従来のインスタンスのみに対する PNM 判定よりも高精度な PNM 判定を実現できることを示した．しかし，感性的な表現を明確に PN 分類することや除去することは難しいとの結論も得られた．但し，こうした感性情報は話題となるクラスを表現するためには，重要な情報と考えらる．例として，記事に対しその話題対象のクラスレベルの感性情報をタギングすることで，ユーザがこうした直感的なタグを参照し自身の興味を持つ可能性のある記事を発見することへの利用なども有効と考えている．

6.　結論と今後の課題

本稿では，ユーザ興味をクラス階層化した興味オントロジをブログ記事から自動抽出する際に，従来より極め細やかな興味オントロジ抽出を実現するため，記事におけるユーザの記述対象がユーザにとってポジティブであるかネガティブであるか，どのような感性で記述されているかを判定する PNM 判定技術について提案した．PNM 判定においては PNM 辞書の生成が核となるが，本研究では特に，記述インスタンスの背景知識としインスタンスの所属するクラスタクソノミを活用することで，各クラスに特徴的な PNM 辞書を生成する．そして，実ブログデータを用いた検証を通じ，本提案がブログ記事内での話題対象をクラスレベルで PNM 分類するに有効であることを示した．

現在，洋楽のみでなく邦楽やファッションをも対象とした汎用性確認のための検証を継続している [13]．また，PNM 辞書内の単一の形態素でなく，複数の形態素の出現性を確認することで記述対象の PNM 判定の精度を向上することを試みる．

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] Amazon web site: *http://www.amazon.com/*.
[2] Google web site: *http://www.google.com/*.
[3] last.fm web site: *http://www.last.fm*.
[4] Liu, B., Hu, M. and Cheng, J.: Opinion observer: analyzing and comparing opinions on the Web., *WWW*, pp. 342–351 (2005).
[5] LiveJournal web site: *http://www.livejournal.com/*.
[6] McGuinness, D. L. and v. Harmelen, F.: Web Ontology Language (OWL): Overview, W3C Recommendation, http://www.w3.org/TR/-owl-features/ (2004).
[7] Mishne, G.: Experiments with Classification in Blog Posts, *Style2005 - the 1st Workshop on Stylistic Analysis Of Text For Information Access, at SIGIR 2005*, SIGIR, ACM (2005).
[8] Mishne, G. and de Rijke, M.: Capturing Global Mood Levels using Blog Posts.
[9] Nakatsuji, M., Miyoshi, Y. and Otsuka, Y.: Innovation Detection Based on User-Interest Ontology of Blog Community., *International Semantic Web Conference*, pp. 515–528 (2006).
[10] O'Donovan, J. and Dunnion, J.: Evaluating Information Filtering Techniques in an Adaptive Recommender System, *Proceedings of Adaptive Hypermedia and Adaptive Web-Based Systems*, pp. 312–315 (2004).
[11] Pang, B., Lee, L. and Vaithyanathan, S.: Thumbs up? Sentiment Classification using Machine Learning Techniques, *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 79–86 (2002).
[12] Turney, P. D.: Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews, *Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics*, Philadelphia, Pennsylvania, pp. 417 – 424 (2002).
[13] 報道発表：情報検索とコミュニティを融合させたモバイルサイト『BRAND COLLECTION（ブランド・コレクション）』における感性検索トライアルの実施について: *http://www.ntt.com/release/2007NEWS/0002/0215.html*.
[14] 中辻真, 三好優, 大塚祥広: ユーザ興味オントロジ抽出によるブログコミュニティ形成手法, 日本データベース学会 Letters, Vol. 5, No. 1 (2006).
[15] 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).