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

Community Creation Technique based on User-Interest Ontology extracted from Blog Knowledge Base

中辻 真[1]　三好 優[2]　大塚 祥広[2]

Makoto NAKATSUJI　Yu MIYOSHI
Yoshihiro OTSUKA

ユーザ興味を発信する手段としブログ利用が目覚しい．しかし，ユーザの興味情報を簡易に詳細化する手段がなく，大量のエントリが発信されるにも関わらず興味に即したエントリを容易に発見できない．そこでユーザ興味をクラス階層化し，各クラスに対する興味度を備える興味オントロジを導入する．まず，サービス毎のドメインオントロジへユーザの蓄積エントリを分類する事で興味オントロジを自動抽出し，次に各ユーザが自身のオントロジを調整する first top-down / second bottom-up な抽出手法を提案する．さらに，興味度を考慮した近似オントロジ計測手法を提案し，興味オントロジに存在しないが興味を持つ可能性の高い Innovative 情報の推薦とコミュニティ形成へ適用する．また，大規模ブログデータと音楽オントロジを用いた検証より，本手法が適切な興味オントロジとコミュニティを抽出できることを示す．

The use of blogs is remarkable as the means to publish the user interests. However, there is no means to extract the user interests in detail, it is difficult to find suitable information resources in spite of a large amount of blog entries are published every day. In order to resolve above problems, we firstly classify user entries into service domain ontology and create interest ontologies which express user's interests as a hierarchy of classes with interest weights. Next, users modify their interest ontologies to update their interests in more detail. Furthermore, we propose similarity measurement between ontologies based on the interest weights, then try to adopt it to innovative entry detection and creation of blog communities. We evaluate the performance of our techniques based on large-scale blog entries and music domain ontology.

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[1]．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[2]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，個人の興味情報を簡易かつ詳細に生成する機能を備えないため，自身の興味に即した検索目的語を適切に構成する必要があり，検索キーワード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，検索自体ができないことも多い．

こうした問題を解決するため本研究[3]では，ユーザ興味をクラス階層表現し，各クラス・インスタンスへの興味度合を示す値である興味度を備えた興味オントロジを各ユーザのブログ記述から自動抽出し，興味度やクラストポロジを考慮した興味オントロジの近似度計測に基づくユーザ興味に即したエントリ推薦やブログコミュニティ生成を試みる．

また，ブログポータルDoblog[4]における大規模データを用いた検証により，高精度な興味オントロジ生成やコミュニティ解析，およびユーザ興味に即したInnovativeエントリ推薦とコミュニティ形成への有効性を確認した．

2. ブログの概要と関連研究

ブログにおける情報検索の特徴は，エントリ公開時に生成されるRDF Site Summary（RSS）と呼ばれる各エントリのタイトル，要約などの簡易なメタデータを用い，公開エントリ群の更新情報を効率的に把握できる点である．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が，エントリに付加されるメタデータは上記簡易なものであり，Semantic Webで期待される詳細なクラス関係を持つオントロジとはいえない．そのため，エントリの検索には，ユーザが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．

これに対しSemblog[5]は，予めカテゴリを各ユーザが作成し，自身が記述・収集したエントリを手動分類することで，パーソナルオントロジというユーザ興味を体系化した情報を構築する．そして，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に基づく検索フレームワーク実現を試みる．一方，本研究はサービス毎の雛型オントロジを用い興味オントロジをトップダウンで生成するためユーザによるオントロジ設計・構築手間が少ない．

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として，文献[6]では，クラス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，クラスとインスタンスのトポロジを分離し，各トポロジへの興味度を考慮する．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Innovative情報として抽出・ユーザ推薦する手法を提案する．

3. 興味オントロジ生成手法の提案

本章では，サービスドメイン毎の雛型オントロジ設計について説明し，次に，興味オントロジ生成手法を提案する．

3.1 雛型オントロジの設計

以下，雛型オントロジの設計手順を説明する[3]．まず，(1)設計者はブログコミュニティ形成を行うサービスドメイン

1　正会員　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　nakatsuji.makoto@lab.ntt.co.jp
2　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　{miyoshi.yu,otsuka.yoshihiro}@lab.ntt.co.jp

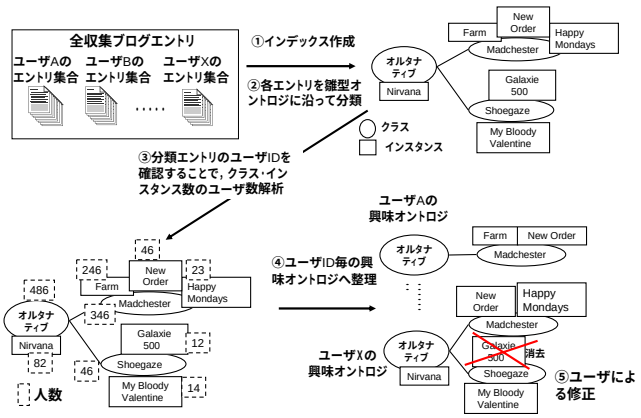

図 1　ユーザ分布解析と興味オントロジ生成手順

Fig.1　Procedure of interest ontology generation.

を選択する．次に，(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選択する．次に，(3) クラス階層を制約するメタデータをプロパティとして選択し，各クラスへその他メタデータを分類する．例えば，ジャンルをプロパティとしアーチストをインスタンスとして各クラスに分類する．なお，雛型オントロジは，サービス拡充に伴い徐々に増やしていくものである．

3.2　ユーザの興味分布解析と興味オントロジ生成

図 1 を用い，ユーザの興味分布解析と興味オントロジ自動抽出手順としてベーシックアルゴリズム(BA)を説明する．

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インデックスファイルを作成する．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従って分類する．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 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 1 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_e の持つ各インスタンスに興味を持つユーザ数を計測する．なお, ユーザ数算出の際, 同一ユーザが複数エントリに同一インスタンスを記述していても，ユーザ数は 1 とする．このようにしユーザ数を末端インスタンスからルートクラスまで再帰的に計測し，そのドメインに興味を持つユーザ分布を計測する．そして，(4)分類結果からユーザ ID の一致する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 1 にユーザ A のエントリ集合がインスタンス“ Farm ”や“ New Order ”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，興味に応じ修正を加える．さらに，こうした修正情報を収集し，雛型オントロジへフィードバックすることも可能である．

3.3　分類誤りのフィルタリング手法

ベーシックアルゴリズムでは，例えば図 2 において，“ Madchester ”配下のインスタンス“ Farm ”などの多義語に対しては，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 ”に分類してまう．そこで，本研究では，(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ち, (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オントロジの特徴，加えて，(3)ユーザの興味対象は一定期間継続されるものであり複数日時を跨がり興味対象概念の記述を行うというブログの特徴を利用し，分類誤りを除去するフィルタリングを 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 の名前が記述されている場合，そのユーザが複数日時を跨がって蓄積してきた全エントリに対し，C_i に所属する I_i 以外のインスタンスや C_i を分類決定要素とし，記述をチェックする．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 2 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，“ New Order ”の記述があれば， E_i は“ 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 以外のインスタンスや C_i を分類決定要素と捉え，記述が存在するかチェックする．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_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，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分類決定要素としフィルタリングの強さを調整する．

3.4　興味度の導入

さらに，興味オントロジを構成するクラス・インスタンスに対するユーザの興味度を導入し，同じクラス・インスタンスに興味を持つユーザの中でも興味度の近いユーザ間でのコミュニティ形成を試みる. まず興味度の定義を与える. (1)1 エントリ当たりのユーザの興味度は 1 であり, (2)あるエントリ E_i に登場するユーザの興味クラス･インスタンスの種類を $N(E_i)$ 個とすると，そのエントリにおける，ユーザの各クラス・インスタンスの興味度は $1/N(E_i)$ となる．(3)興味オントロジ内の各インスタンス I_i に対する興味度は，ユーザの全蓄積エントリ集合を E とし， $I_i = \sum_{I_i \in E_i}^{|E|} 1/N(E_i)$ となり，各ク

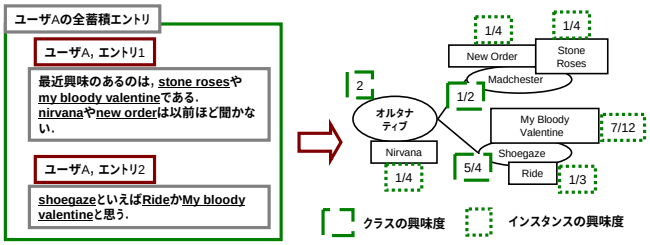

図 2　興味度の例

Fig.2　An example of user interest weight.

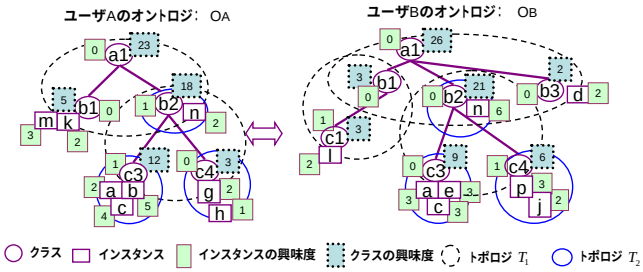

図 3　オントロジ間の近似度計測アルゴリズム

Fig.3　An algorithm of similarity calculation.

ラスに対する興味度は，$C_i = \sum_{C_i \in E_i}^{|E|} 1/N(E_i) + \sum_{I_i \in C_i} I_i$ となる．(4)インスタンスの興味度は所属クラスの興味度に反映され，子クラスの興味度は親クラスの興味度に反映される．例えば，図 2 では，``My bloody valentine"というインスタンスについては，興味度は 1/4+1/3=7/12 となり，``shoegaze" というクラスについては，興味度は 1/3+7/12+1/3=5/4 となる．さらに，``オルタナティブ"の興味度は配下クラス・インスタンスの興味度の和となり 1/2+5/4+1/4=2 となる．

4. 興味オントロジ間の近似度計測手法と Innovative エントリの推薦

まず，興味度を考慮したオントロジ間の近似度計測手法を提案し，近似度の高いユーザ間の興味オントロジ解析に基づく Innovative エントリ推薦について述べる．

図 3 を例にあげ近似度計測アルゴリズムを説明する．なお，ユーザ興味に従うコミュニティ生成には，オントロジを構成する各クラス・インスタンスへの興味度を考慮した近似度計測が重要である．更に，興味オントロジは雛形のクラススキーマに沿って作成されるため，近似度の計算量を低く抑えることが出来る．これは，オントロジマッピング技術を大規模ブログコミュニティに適用するには重要である．

ここで，親・子クラスからなるトポロジを T_1 とし，クラスとインスタンスからなるトポロジを T_2 と与える．さらに，両オントロジの共通クラスを C_i と定義し，共通インスタンスを I_i と定義する．特に，トポロジ T_1 に所属するクラス集合を $C(T_1)$，トポロジ T_2 に所属するクラス集合を $C(T_2)$ とする．また，共通クラスに対する興味一致度を $I(C_i)$ とし，共通インスタンスに対する興味一致度を $I(I_i)$ とし，共通クラスにより構成されるトポロジ間の興味一致度を $I_t(C_i)$ とする．(1)まず，共通クラスを分析しトポロジ T_1 を形成する共通クラスと，トポロジ T_2 を形成する共通クラスを抽出する．例えば，図 4 では，共通クラス a1，b1，b2 はトポロジ T_1 を形成し，b2，b3，c4 はトポロジ T_2 を形成する．(2)共通クラス C_i が両オントロジ間で共通インスタンス I_i を持つとすると，共通インスタンス I_i に対する興味一致度 $I(I_i)$ は，両オントロジのインスタンス I_i に対する興味度のうち小さい方の値とする．例えば，図 4 では，共通クラス c3 配下の共通インスタンス a に対する興味一致度は 2 となる．(3)同様に，共通クラスに関する興味一致度 $I(C_i)$ は，両オントロジのクラス C_i に対する興味度のうち小さい方の値とする．例えば，共通クラス c3 に対する興味一致度は 9 となる．(4)次に，クラス $C_i \in C(T_1)$ 配下の興味一致度 $I_t(C_i)$ は，O_A と O_B における C_i 配下の子クラスの積集合を $N(C_i)$ とし，和集合を $U(C_i)$ とすると，$I_t(C_i) = \sum_{C_j \in (N(C_i))} I(C_j)/U(C_i)$ で与える．例えば，共通クラス b2 配下に対する興味一致度は(9+3)/2=6 となる．そして，興味一致度 $I_t(C_i)$ をトポロジ T_1 を形成する全共通クラスで足し込んだ値 $\sum_{C_i \in C(T_1)} I_t(C_i)$ をトポロジ T_1 を形成する共通クラス集合に対する興味一致度 $S(T_1)$ とする．(5)一方，共通クラス $C_i \in C(T_2)$ に対し，O_A における C_i 配下のインスタンス集合を $I_A(C_i)$ とし，O_B における C_i 配下のインスタンス集合を $I_B(C_i)$ とし，C_i の興味一致度 $I_t(C_i)$ を，$\sum_{I_i \in C_i} I(I_i)/|I_A(C_i) \cup I_B(C_i)|$ で与える．例えば，共通クラス c3 配下に対する興味一致度は ((2+0+3+0)/4)=5/4 となる．そして，興味一致度 $I_t(C_i)$ をトポロジ T_2 を形成する全共通クラスで足し込んだ値 $\sum_{C_i \in C(T_2)} I_t(C_i)$ をトポロジ T_2 を形成する共通クラス集合に対する興味一致度 $S(T_2)$ とする．(6) $S(T_1)$ および $S(T_2)$，両トポロジ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 $S_O(AB) = S(T_1) + f(S(T_2))$ で与える．

次に，提案手法を Innovative 情報の推薦やコミュニティ形成へ適用する．まず，ユーザ A とその他のブログユーザ集合 $u \in U$ との間で近似度計測を総当りで行う．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 Innovative 情報として推薦する．

5. 提案手法の実装と評価

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の興味分布の検証を行なった[3]．実験では goo 音楽[7]などの Web ポータルの公開情報を参考とし，音楽ドメインの 114 ジャンルをクラスとし，各クラスに合計約 4300 のアーチストをインスタンスとして分類し，図 1 に一部示すような雛型オントロジを作成した．図ではクラス階層のみを表示しているが，実際には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，“ verve ”というインスタンスには，“ ヴァーヴ ”と“ verve ”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

表 1　提案手法による興味オントロジの精度（FA2，hop2）

Table.1　Accuracy of user interest Ontology.

	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

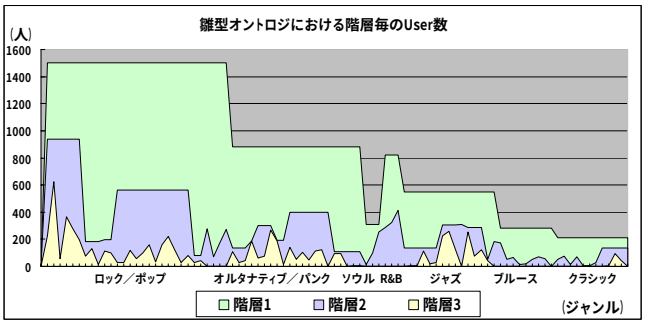

図 4　雛型オントロジに対するブログユーザの興味分布

Fig.4　Interest distributions of blog users.

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．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適合率向上がまず必須である．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[8]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 1）．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．また図 4 に，雛型オントロジ内の各クラスに対するユーザ分布をクラス階層毎に解析した結果を示す．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．

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（表 2）．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，FA1 で設定したユーザ興味の継続期間の仮定を，全蓄積エントリから時系列の近いエントリやトラックバック等関連エントリに短縮し，分類決定要素をチェックする事にし，適合率を維持しつつ正解数を増やす事を試みる．また FA2 に対しホップ数を変化させ，精度を比較したところ，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った．さらにホップ 0 と 4 を比較すると，

表 2　BA，FA1，FA2 に対する精度の比較

Table.2　Accuracy comparison among BA, FA1, and FA2.

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

ホップ 4 の方が若干正解数は増えるが，適合率は下がった．これは，今回用いた雛型オントロジが，末端クラスに多くインスタンスが存在する上，例えばクラス“メタル”配下の末端クラスに“北欧メタル”などがあり，親クラスに“ロック”があるなど，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と考えられる．

6. 結論と今後の課題

本稿では，ユーザ毎の興味オントロジ生成と興味度を考慮したオントロジ間の近似度計測手法を提案し，innovative 情報の推薦への適用を試みた．そして，ブログポータル Doblog の大規模データを基に提案手法の実現性を確認した．

今後，Doblog 上でコミュニティ形成支援実験サービスを実施し，アクセス履歴に基づく(1)Innovative エントリ推薦と(2)コミュニティ毎のユーザ分布の時間変化を検証する．

[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 チーム及び株式会社ホットリンク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] 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.
[2] goo homepage, http://www.goo.ne.jp.
[3] 中辻真，三好優，大塚祥広: 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, DEWS2006, 2006.3.
[4] Doblog homepage, http://www.doblog.com.
[5] Ohmukai, I. and Takeda, H.: Metadata-driven Personal Knowledge Publishing, ISWC2004, 2004.
[6] Maedche, A. and Staab, S.: Measuring Similarity between Ontologies, In Technical Report, E0448, University of Karlsruhe, 2001.
[7] goo音楽 homepage, http://music.goo.ne.jp/
[8] Namazu homepage, http://www.namazu.org/

中辻 真　Makoto NAKATSUJI
NTT ネットワークサービスシステム研究所．2003 京都大学大学院情報学研究科システム科学専攻修士課程了．電子情報通信学会，日本データベース学会など会員．

三好 優　Yu MIYOSHI
NTT ネットワークサービスシステム研究所．2000 早稲田大学大学院理工学研究科電子・情報通信学修士課程了．2003 度電子情報通信学会学術奨励賞，電子情報通信学会など会員．

大塚 祥広　Yoshihiro OTSUKA
NTT ネットワークサービスシステム研究所．1985 東京工業大学理工学研究科電子物理専攻修士課程了．1992 テレコムシステム技術奨励賞，電子情報通信学会会員．