知能と複雑系　124−5
（２００１．　５．１７）

WWW における情報源に関する知識の共生

市瀬 龍太郎，武田 英明，本位田 真一
国立情報学研究所 知能システム研究系 知識処理研究部門

概要

本論文では，WWW で使われる情報源に関する知識を流通させるための手法を提案する．WWW の情報源に関する知識は，インターネットディレクトリーをはじめとして，分類階層を使って管理されることが多い．このような知識は異なる概念階層，語によって管理されているため，そのまま他者の知識を持って来たとしても，自分の持つ概念階層のどこに位置するべきものなのかを同定することは難しい．そこで，本研究では，情報の分類に基づいて，他者が持つ知識を自分が持つ知識に統合する手法を提案する．その提案手法を評価するためにシステムを作成して実験を行った．インターネットディレクトリーを用いた実験を行った結果，他者が持つ知識を適切なカテゴリーに取り込めることが示された．

Knowledge Symbiosis for the Information Resources of the WWW

ICHISE Ryutaro, TAKEDA Hideaki, and HONIDEN Shinichi
National Institute of Informatics

Abstract

There are much information managed via a system of hierarchical categorization on the internet. Such information is hard to align between different concept hierarchies, because of their variety of conceptual structure. Consequently, information in a concept hierarchy can not used in the other concept hierarchy. In this paper, we propose a method for aligning information from one concept hierarchy to another. In order to evaluate our method, we conducted experiments using internet directories. The results of these experiments show that the proposed method can induce appropriate alignment rules for concept hierarchies and classify information into appropriate categories within another concept hierarchy.

1　はじめに

近年の WWW の普及により，個人が入手できる情報は，飛躍的に大きくなってきている．これらの情報は，ブックマークやインターネットディレクトリーなど，概念階層を使って管理されることが多い．しかし，同じような概念階層を利用して管理しているにもかかわらず，それぞれの管理者は，別々の概念階層を用いて管理を行っているため，それらの情報を再利用するのは難しいという問題点がある．本論文では，それぞれの概念階層が持つ知識を分散した管理状態のままで相互に利用できるようにする手法を提案する．知識が分散している環境を知識共生環境としてとらえ，共生している他の知識源から知識を取り込むことで，自分の持つ知識を拡張していく手法である．そのようなアプローチを取ることで，他の知識源が持つ知識を自分の知識として取り込み，有効に利用できるようになると考えられる．また，共生環境においては，知識は分散管理されるため，管理が容

易になるという利点がある．しかし，他者の知識は，異なる概念階層，語によって管理されているため，そのまま他者から情報を持って来たとしても，自分の持つ概念階層のどこに位置するべき情報なのかを同定することは難しい．そこで，本研究では，情報のインスタンスに基づいて他者が持つ知識を自分の持つ知識に統合する規則を学習する手法を提案する．

次の第 2 章では，本研究で仮定する情報源についての定義をおこなう．第 3 章では，第 2 章で定義した情報源を統合して自分の持つ知識に取り込む手法について述べる．第 4 章では，提案する手法の有効性を確認するために，その手法に基づいて作られたシステム HICAL について述べ，インターネットディレクトリーを知識源として適用した実験について報告する．第 5 章では，本研究と関連研究の比較を行って，本研究の特徴を明らかにし，第 6 章で本研究をまとめる．

2　WWW における情報源

この章では，本研究で仮定する WWW における情報源についてモデル化を行う．ここで取り扱う情報源は，インターネットディレクトリーやブックマーク，リンク集などの概念階層に基づく分類が行われて管理されているものである．これらの管理手法には 2 つの要素がある．1 つは，情報のインスタンスとなる WWW ページであり，もう 1 つは概念階層である．概念階層は最も一般的な概念を最上位として，徐々に詳細化されている．インスタンスはその中である基準に従って分類が行われていき，分類がこれ以上できなくなる所で，ある概念を示すノードに割り当てられる．

これを単純化し模式的に表すと，図 1 のような形で表す事ができる．概念階層は木構造で表され，インスタンスが木構造のノードに割り当てられる事になる．図中では，黒点がある概念を表し，白点がインスタンスを表すこととなる．この分類手法は，概念階層の構造によって異なり，分類を行う知識を表していると考えられる．本論文では，そのように知識を提供できるものを階層的知識源と呼ぶことにする．

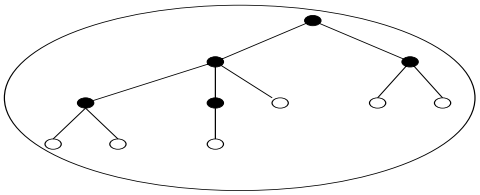

図 1: 階層的知識源のモデル

階層的知識源は，その中に含まれる情報の種類，概念階層の構築者などによって，異なった階層構造を有する．したがって，このような知識源が複数存在する時に，他者の知識を利用するのは，非常に困難が伴う．例えば，図 2 では，2 つの知識源が存在している．概念階層 C_1 には，インスタンスとして I_1, I_2, I_3 の 3 つが存在しているが，概念階層 C_2 には I_1, I_3 の 2 つのみが存在しており，I_2 は含まれていない．この時，C_2 が C_1 の持つ知識を利用する事で，I_2 の情報が利用できるようになれば，C_2 にとっての利益が非常に大きいと考えられるが，C_1 と C_2 の持っている概念階層が異なるため，そのまま I_2 を C_2 に持って来ただけでは，C_2 上のどこに位置するべき情報なのかが C_2 には分からない．したがって，そのまま I_2 を利用する事はできない．本研究では，これらの知識源における概念階層の対応を学習する事によって，他の知識源を利用する手法を提案する．このことにより分散する知識の共生環境が実現できることとなる．

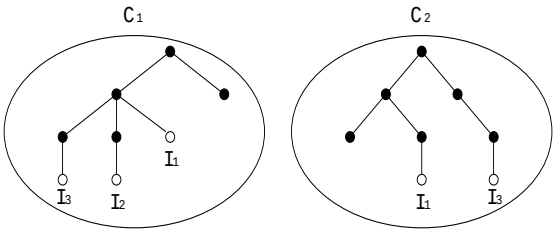

図 2: 2 つの階層的知識源の問題点

3　階層的知識源の統合手法

ここでは，階層的知識源を統合する手法について述べる．前章で述べたように，2 つの知識源が

	N_{1i} に含まれる	N_{1i} に含まれない
N_{2j} に含まれる	m_{11}	m_{12}
N_{2j} に含まれない	m_{21}	m_{22}

表 1: 分割表

ある時には，それぞれの対応関係が分からないため，他の知識源のもつ知識を利用する事ができない．本研究では，情報のインスタンスを利用する事で，知識源の持つ分類基準の類似性を発見し，類似する分類間の変換規則として学習する事で，他の知識源を利用できるようにする手法を提案する．

3.1 κ 統計量

知識源における概念階層は，インスタンスをある概念に従って分類した物となる．概念階層は，木構造をしているため，ある概念ノードより下に属するインスタンスはその概念ノードに属していると判定できる．そのため，あるインスタンスを選択した時に，その概念に適合するか否かを容易に判定する事ができる．そのように考えると，2 つの知識源における 2 つの任意の概念ノードに対して，インスタンスの分類を元に概念基準の類似性の判定を行う事ができるようになる．本研究では，この概念基準の類似性の判定に，κ 統計量 [1] を用いた．κ 統計量では，2 つの概念判定に対して，表 1 のような分割表を作成する．表 1 はある概念ノードに含まれるインスタンスの数と含まれないインスタンスの数を一覧にした物である．N_{1i}, N_{2j} は，それぞれ知識源 C_1, C_2 中のある概念ノードを表している．ここで，2 つの概念判定基準が等しければ，表 1 中の m_{11}, m_{22} の数が多くなり，m_{12}, m_{21} の数が少なくなる．逆に基準が正反対であれば，m_{12}, m_{21} の数が多くなり，m_{11}, m_{22} の数が少なくなる．κ 統計量では，このことを利用して 2 つの判断基準が等しいか否かをある有意水準で判定を行う．本手法では，類似性の判定にこれを利用する．

3.2 類似概念の発見

類似概念の発見は，2 つの知識源の最上位の概念階層から調べていくことで行われる．このアルゴリズムは図 3 で表される．まず 2 つの最上位の概念に対して，κ 統計量を用いて類似性の判定を行う．類似している場合には，その組合せを記録するとともに，その概念同士の下位の概念に対して組合せを作成し，その作成された組合せに対して再帰的に類似性の検査を行っていく．この時に，類似概念 N_{1i}, N_{2j} に対して生成される組み合わせは，以下の 3 つである．

- N_{1i} と N_{2j} の子の組み合わせ
- N_{1i} の子と N_{2j} の組み合わせ
- N_{1i} の子と N_{2j} の子の組み合わせ

類似した組合せが生成されなくなった時に，システムは停止し類似する概念の組合せを出力する．

3.3 規則の生成

図 3 のアルゴリズムにより生成された類似概念の組合せ集合に対して，インスタンスがある概念に含まれているならば，それと対応する類似概念にも含まれるという形式の規則を生成し，学習結果として利用する．たとえば，類似概念として，C_1 上の N_1 と C_2 上の N_2 が類似しているという規則を図 3 のアルゴリズムが出力した場合には,「C_1 上の N_1 に属するインスタンス I_i は，C_2 上の N_2 に属する」という規則が生成される．

4 インターネットディレクトリーを用いた実験

この章では，前章で述べた手法の妥当性を評価するために，SICStus Prolog を使って実装されたシステム HICAL を用いた実験についての報告を行う．

4.1 実験設定

実験を行う対象として，インターネットディレクトリーの Yahoo! Japan [2] と LYCOS Japan [3] の分類体系を知識源の概念階層として用い，そこに含まれる外部リンク (URL) を情報のインスタンスとして用いた．Yahoo!には，約 41,000 のカテゴリーがあり，約 224,000 個の URL が登録されてい

```
Input:   N10,   // C1 の最上位概念
         N20,   // C2 の最上位概念
         P,     // κ 統計量の閾値
Output:  R;     // 類似概念の組合せの集合
begin
  /* 子ノードなどを組合せ，候補と */
  /* なる組合せの集合を作る */
  X1 := combination(N10, N20);
  t := 1;
  R := φ;
  while Xt ≠ φ
    while Xt ≠ φ
      I := Xt の要素;
      N1, N2 := I 中の 2 つのノード;
      /* κ 統計量の計算 */
      if κ(N1, N2) ≥ P
        Xt+1 := combination(N1, N2);
        R := R + I;
      fi;
      Xt := Xt − I;
    end;
    t := t + 1;
  end;
  return R;
end;
```

図 3: 類似概念同定アルゴリズム

る．一方，LYCOS には，約 5,700 のカテゴリーがあり，約 48,000 個の URL が登録されている．登録 URL 数を見ると，Yahoo!の方が圧倒的に多く，知識源としての LYCOS は役に立たないように見えるが，LYCOS に収録されている外部リンクの半数である約 24,000 個しか，Yahoo!と共有されていない．このことは，巨大な知識源が 1 つあったとしても，中に含まれる知識には偏りがあるため，他者の知識が必要になるということを表していると考えられる．

Yahoo!と LYCOS からは以下の 3 つのカテゴリーとそのサブカテゴリーの概念階層を選択し，実験を行った．

- Yahoo! : Arts / Humanities / Literature
  LYCOS : 芸術と人文科学 / 文学
- Yahoo! : Business_and_Economy / Companies
  LYCOS : 経済・産業 / 企業
- Yahoo! : Recreation
  LYCOS : 趣味・スポーツ

4.2 実験手順

実験は図 4 で示される手順で行った．まず，インターネットディレクトリーから，Ruby で作成されたデータ変換ツールを用いて，分類のデータベースの作成を行った．その後，両者に共通して出現する Web 文書の URL のみを抽出し，データを 10 分割して訓練例とテスト例の作成を行った．これは 10-fold のクロスバリデーションを行うためである．そして，ここで作成した訓練データを Prolog で作成した HICAL システムに入力し，規則の学習を行った．規則の学習を行う際の κ 統計量の有意水準は 5%とした．その後，規則の妥当性を調べるため，Ruby で作成した評価器でテスト例を使った評価を行い，正答率を計算した．

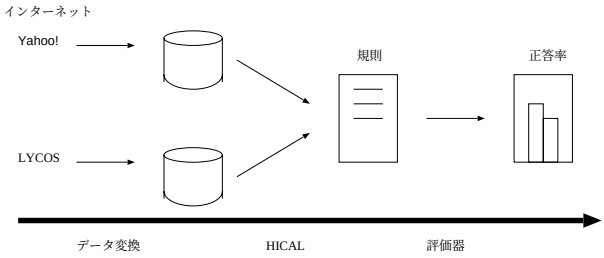

図 4: 実験手順

評価器で使われる正答の評価方法として，いくつかの手法が考えられる．この実験では，2 つの評価手法を提案して，それぞれの評価器を用いて実験を行った．その手法は以下の 2 つである．

1. インスタンスがテスト例と同じカテゴリーに分類された時に正答とする手法.
2. インスタンスがテスト例と同じカテゴリーに

分類されるか，その上位のカテゴリーに分類された時に正答とする手法.

評価法 1 では，元の概念階層に対して，目標の概念階層が十分な中間概念を含んでいなければならず，非常に厳密な評価方法であると言える．一方の評価法 2 は，元の概念階層と目標の概念階層のどちらが詳細な階層を持っているかということに依存しない一般的で現実的な評価方法であると言える．

4.3 実験 1

まず 1 番目の実験として，学習された規則の妥当性を調べるために，インスタンスが属しているカテゴリーの規則のみを使って評価を行った．この場合には類似概念が学習されたものだけがテストされるため，本研究で提案する「あるカテゴリーに属するものは，それと類似しているカテゴリーにも属する」という仮説の検証を行うのに適していると考えられる．この結果を表したものが図 5 である．上の図が Yahoo!から LYCOS への変換を

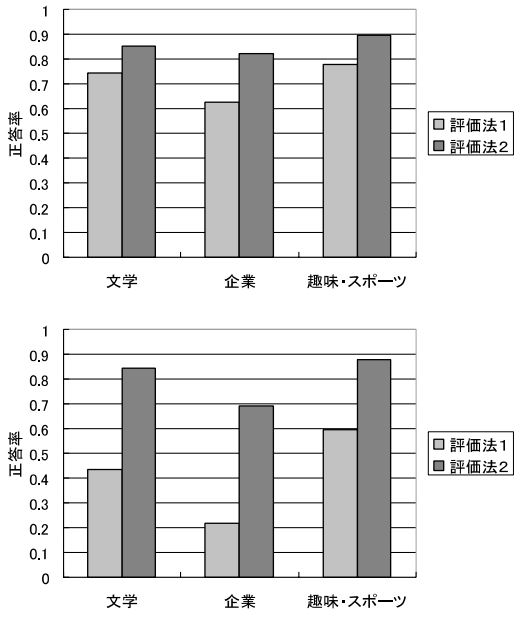

図 5: 実験 1 の結果

行った規則の評価であり，下の図が LYCOS から Yahoo!への変換を行ったものである．図 5 から分かるように，文学と趣味・スポーツの実験では 8 割を越える正答率を出している．また，企業の実験においても Yahoo!から LYCOS への実験は 8 割以上の正答率を出している．

4.4 実験 2

実験 1 で使った手法は，インスタンスが属する概念を利用した規則のみを使っているため，正確な規則を得られる事が期待されるが，類似概念が学習されなかった場合には，その概念に属するインスタンスの情報を利用できないという問題点がある．2 番目の実験として，全てのインスタンスに対して，この手法を適用した場合の性能について評価を行う．この実験では，規則が割り当てられなかったインスタンスが属する概念に対して，その概念より上位の概念に割り当てられた規則を適用することで対処して実験を行った．実験 2 は，実験 1 の時とは異なり，他者の持つ全てのインスタンスを利用できるようになるため，より実践的な評価であるとも言える．この結果を示したものが，図 6 である．実験 1 と同様に，上の図が Yahoo!から LYCOS への変換を行った規則の評価であり，下の図が LYCOS から Yahoo!への変換を行ったものである．実験 1 の結果と比較すると，全体的に多少正答率が低くなっているが，実験 1 の結果とほぼ同様の結果を出していると言える．

4.5 実験 3

次に規則の選択手法の違いによる生成規則の性能の違いを調べるために，HICAL の設定を変更して実験を行った．HICAL システムでは，κ 統計量により，信頼性の高い概念の組を選択して規則の生成を行う．しかし，数が少いため信頼性が低いと判定されても，より特殊な概念同士の組の方が，一般的な概念同士の組よりも有用であることが考えられる．そこで，一つの概念に対して，2 つ以上の類似する概念の組が κ 統計量により発見された場合には，より特殊な概念の組を選択して規則を生成する手法を使って実験を行った．それ以外の実験設定は実験 2 と同じである．結果は，図 7 のようになった．上の図が Yahoo!から LYCOS への

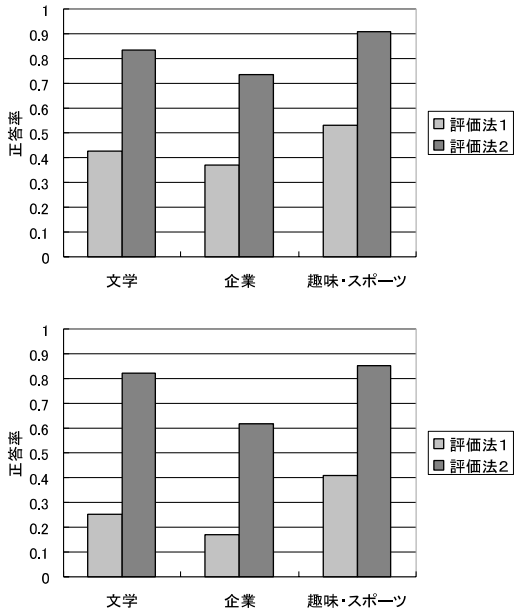

図 6: 実験 2 の結果

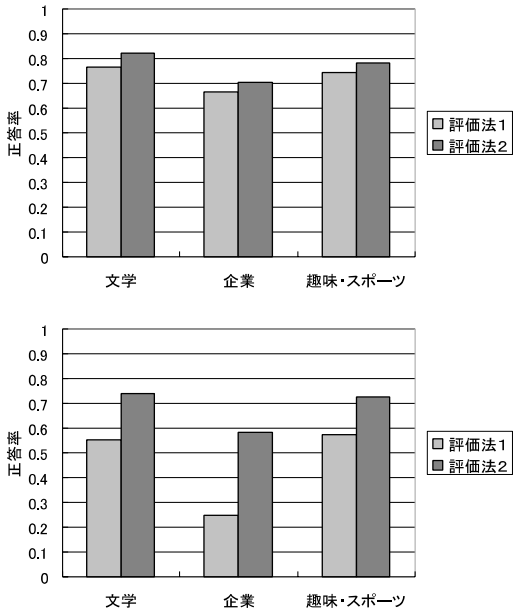

図 7: 実験 3 の結果

変換を行った規則の評価であり，下の図が LYCOS から Yahoo!への変換を行ったものである．他の実験と同様に，企業データでの正答率が他と比べて低い点は変わらないが，評価法 1 の正答率は大幅に向上する．しかし，評価法 2 では正答率が下がってしまう．

4.6 考察

実験 1 の良好な結果から,「あるカテゴリに属するインスタンスは類似するカテゴリにも属する」という本研究の仮定は，ほぼ正しいことが分かる．その仮定を基に，すべてのインスタンスを他の知識源に移行させる場合についての実験を実験 2 で行ったが，文学と趣味・スポーツのドメインにおいては 8 割以上の正答率を持ち，企業のドメインにおいても 6 割以上の正答率を示す結果となった．これらの結果から分かるように，本手法はどのドメインにおいてもうまく規則を学習していると言えるであろう．実験 3 では，類似概念を選択する手法について，実験 2 とは異なる新たな手法で実験を行った．この手法は，学習されたカテゴリーに対しての正答率を向上させるが，親のカテゴリーの規則を使った場合には，かえって正答率を低下させてしまう難点があるといえる．

「Yahoo!から LYCOS」と「LYCOS から Yahoo!」を比較すると,「Yahoo!から LYCOS」の方が，より正確であると言える．全体のカテゴリーの数を比較すれば分かるように，Yahoo!の方が LYCOS よりもより詳細な分類階層を持っていると考えられる．従って,「Yahoo!から LYCOS」では，複雑な分類体系から簡単な分類体系への規則を学習していると考えられる．このような場合には，比較的正確な規則を学習しやすいのではないかと考えられる．例えば，概念階層 A, B を考え，両方が分類 X を持っているとする．A はさらに X を分割する概念 S を持っていて，B にはそのような概念がないとする．そのような場合には，

$$A : /X/S- > B : /X$$

という規則は

$$B : /X- > A : /X/S$$

という規則よりも学習されやすい．なぜならば，$B : /X$ には $A : /X/S$ に入るインスタンスのみならず，S には入らない $A : /X$ 以下のインスタンスも含まれているからである．しかし，本研究で提

案する手法は，このような場合でも，うまく動作していると考えられる.「評価法 2」の結果から分かるように，C_1 から C_2 へと C_2 から C_1 へがほとんど同じ結果となっている．このことから

$$B : /X- > A : /X/S$$

という規則を学習するかわりに，

$$B : /X- > A : /X$$

という規則が正しく学習されていると考えられる．

定性的な評価を行うために，文学のドメインで学習された規則の一部について解析を行った．文学のドメインにおいては，10 回のクロスバリデーションの結果，平均で 137 個の規則を学習している．その中では,「Yahoo! : Arts / Humanities / Literature / Genres / Literary_Fiction / Authors / Murakami_Haruki」と「LYCOS : 芸術と人文科学 / 文学 / 小説/ 村上春樹」のように，概念のラベル名の対応を取るだけで，発見できるような規則もあったが,「Yahoo! : Arts / Humanities / Literature / Poetry / Waka / Kajin / Masters / Murasakisikibu」と「LYCOS : 芸術と人文科学 / 文学 / 日本文学 / 上代・中世文学 / 源氏物語」のように，概念のラベル名の対応だけでは関係を学習できないようなものに対しても，発見を行っていた．

一般的に使われる同義語辞書は同じものに対しての識別効果は高いが，分類ラベルとして使われる場合には，同じ分類基準を持っているにもかかわらず，この例のように必ずしも同義語が使われるとは限らない．本研究で提案している手法は，同義語辞書などを利用せずに形式のみを利用しているため，辞書に依存するような問題が起こることがなく，このような関係も抽出できていると考えられる．このような手法は，語によらない概念同士の関係を発見できるため，知識発見の手法などにも応用できるのではないかと考えられる．

次に，規則の質的な妥当性を調べるために，テスト例に対して分類の間違いを起こした規則の調査を行った．あるデータセットに対して，Yahoo!から Lycos への規則を対象として調査を行うと，間違えた分類の規則でも大きな間違えはさほど多くないということが分かった．たとえば,「大阪府立国際児童文学館」の Web ページは，テスト例では Lycos の「児童文学／児童文学館」というカテゴリーに分類されていたが，HICAL を使うと「文学／記念碑・記念館」に分類されてしまう．これは正答率を計算する際に分類間違えとして計算されることになるが，内容的には完全な分類間違えとは言いがたい．Yahoo!では，児童文学館に類似する分類を持っていないため,「大阪府立国際児童文学館」を「図書館・文学館」として分類している．そのため，LYCOS のみが持つ「児童文学館」というカテゴリーへの学習が行われず，類似している「図書館・文学館」への規則が学習されている．このように，テスト例と違う分類が行われているものでも，ある程度の妥当性のある答えを返している物が多かった．しかし,「ヤングアダルト/作家」に属するものは「SF・ホラー・ファンタジー」に属するという一般的な規則が学習されたため,「ミステリー」に属する作家のページを「SF・ホラー・ファンタジー」に誤分類してしまうという例も見受けられた．

5 関連研究

オントロジーの併合 (merging)/調整 (alignment) を行うシステムの Chimaera [4] や PROMPT [5] と比較すると，これらのシステムは，併合/調整の際にユーザの介在が欠かせない点で異なっている．また，これらのシステムは，類似した概念を発見するのに語の類似性を利用しているため，4.3.2 節で述べたような，語とは離れた類似概念を発見することができない上，使用する辞書などの影響を受けやすいと考えられる．一方，HICAL では，形式的な情報のみを利用するため，このような影響はないという利点がある．

ブックマーク共有システムの Siteseer [6] や Blink [7] と比較すると，HICAL が異なる知識源の知識を利用している点で，非常に似ている．しかし，これらのシステムとは，HICAL が概念階層

を利用しているという点で大きく異なる．SiteseerやBlinkは与えられたカテゴリーに含まれるインスタンスの数のみしか利用していないが，HICALは階層構造を利用することで，より有効に他者の知識を利用している．階層構造を用いた場合には，あるインスタンスにちょうど適するカテゴリーが存在しない場合に，親のカテゴリーを利用することができる．kMedia [8] は階層構造を用いたブックマーク共有システムであるが，PROMPTなどと同様に語の影響を受けてしまう．

6 むすび

本論文では，WWWの情報を管理する概念階層を一つの知識源としてみなし，その知識源のモデル化を行った．そして，その知識源に含まれる情報のインスタンスを用いることで，他の知識源が持つ知識を自分の持つ知識として取り込む手法について述べた．その手法は，インスタンスの分類の類似性に基づいて，各カテゴリー間の類似性を同定し，対応関係を規則として学習する機械学習の手法である．その手法に対する有効性を評価するために，システムHICALを構築し，インターネットディレクトリーのYahoo!とLycosを用いて実験を行った．実験によって，提案手法を用いると他の知識源から知識を適切な場所に取り込めることが示された．

参考文献

[1] Fleiss, J. L.: *Statistical Methods for Rates and Proportions*, John Wiley & Sons (1973), 佐久間 昭訳，邦題：「係数データの統計学」，東京大学出版会，1975.

[2] Yahoo! Japan, http://www.yahoo.co.jp/ (2000).

[3] LYCOS Japan, http://www.lycos.co.jp/ (2000).

[4] McGuinness, D. L., Fikes, R., Rice, J. and Wilder, S.: An Environment for Merging and Testing Large Ontologies., in Cohn, A. G., Giunchiglia, F. and Selman, B. eds., *Proceedings of the Conference on Principiles of Knowledge Representation and Reasoning (KR-00)*, pp. 483–493, S.F. (2000), Morgan Kaufman Publishers.

[5] Noy, N. F. and Musen, M. A.: PROMPT: Algorithm and Tool for Automated Ontology Merging and Alignment, in *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI-2000)*, pp. 450–455, Menlo Park (2000), AAAI Press.

[6] Rucker, J. and Polanco, M. J.: Siteseer: Personalized Navigation for the Web, *Communications of the ACM*, Vol. 40, No. 3, pp. 73–75 (1997).

[7] Blink.com, http://www.blink.com/ (2000).

[8] Takeda, H., Matsuzuka, T. and Taniguchi, Y.: Discovery of Shared Topics Networks among People – A Simple Approach to Find Community Knowledge from WWW Bookmarks, in Mizoguchi, R. and Slaney, J. eds., *Proceedings of the 7th Pacific Rim International Conference on Topics in Artificial Intelligence (PRICAI-2000)*, Vol. 1886 of *LNAI*, pp. 668–678, Berlin (2000), Springer.