特集論文

階層的分類データを統合するための規則学習機構

A Rule Learning Mechanism for Integration of Concept Hierarchies

市瀬 龍太郎
ICHISE Ryutaro
国立情報学研究所 知能システム研究系／ 総合研究大学院大学 情報学専攻
Intelligent Systems Research Division, National Institute of Informatics / Department of Informatics, Graduate University for Advanced Studies
ichise@nii.ac.jp, http://research.nii.ac.jp/~ichise/

濱崎 雅弘
HAMASAKI Masahiro
総合研究大学院大学 情報学専攻
Department of Informatics, Graduate University for Advanced Studies
hamasaki@grad.nii.ac.jp, http://www-kasm.nii.ac.jp/~hamasaki/

武田 英明
TAKEDA Hideaki
国立情報学研究所 実証研究センター／ 総合研究大学院大学 情報学専攻
Research Center for Testbeds and Prototyping, National Institute of Informatics / Department of Informatics, Graduate University for Advanced Studies
takeda@nii.ac.jp, http://www-kasm.nii.ac.jp/~takeda/

keywords: machine learning, data integration, catalog, concept hierarchy, web

Summary

With the rapid advance of information technology, we are able to easily and quickly obtain a great deal of information on almost any topic. One method by which to managing such large amounts of information is to utilize catalogs which organize information within concept hierarchies. However, the concept hierarchy for each catalog is different because one concept hierarchy is not sufficient for all purposes. In the present paper, we address the problem of integrating multiple catalogs for ease of use. The primary problem lies in finding a suitable category in a catalog for each information instance in another catalog. Three approaches can be used to solve this problem: ontology integration approach, instance classification approach and category alignment approach based on categorization similarity. The main idea of this paper is a multiple strategy approach to combine the instance classification approach and the category alignment approach. In order to evaluate the proposed method, we conducted experiments using two actual Internet directories, Yahoo! and Google. The obtained results show that the proposed method improves upon or is competitive with the integration method based only on category alignment or instance classification. Therefore, the proposed catalog integration method is shown to be an effective combination of the instance classification approach and the category alignment approach.

1. はじめに

近年のインターネットの普及により，子供から老人まで多くの人々がインターネットを使うようになってきた．そのような状況を受け，さまざまな人がさまざまな情報を受信するのが容易になり，逆に，情報の発信も非常に容易になってきた．これを背景として，膨大な量の情報交換がインターネットでなされるようになると同時に，すべての情報がインターネットを経由してやりとりされるようになりつつある．例えば，従来まで専用回線でやりとりされていた音声電話もインターネットを通したVoIPに取って代わられるようになっているし，RFIDの普及により，さまざまな物に関する情報などもインターネットでやりとりされることになるであろう．つまり，求めるすべての情報がインターネット上で手に入るようになってきているのである．しかし，ユーザがある情報を基に状況の判断，意志決定を下すには，単一の情報を捜し出せば十分なのではなく，多数の情報を分析しなければならないことが多い．現状において，これらの作業はユーザに任されており，得られる情報が多くなった分，情報の分析の手間が増えてしまうという状況を生んでいる．そこで，ユーザの負担を減らすために，複数の情報を統合し，ユーザに統一的に提供するデータ統合技術が必要となってくる．

インターネット上のデータは，情報の形式，粒度が異なっているため，単純に統合をしようとしても難しい．そのような中で，一番重要な問題は，統合データの整合性をいかにして保つかということである．たとえば，レストランの場所を表すデータに対して「住所」と書かれたデータと「Address」と書かれたデータは，同じものとして統合しても問題はないであろう．また，同様に，分散するWebデータを統合して宿泊施設の住所録を作る場合には，旅館の住所は，宿泊施設の住所録に入れてもよいであろうが，レストランの住所は入れてはいけない．このように，住所を示している同じデータであっても，データの意味によって統合できるか否かが決まる．本論文で

は，整合性を持ったデータ統合を実現するために，カタログ統合問題として，複数のデータを統合する問題を定義し，計算機による自動統合の実現手法について提案する．カタログ統合問題とは，インターネットディレクトリのように階層的に分類された複数のデータを統合する問題である．

次の第 2 章では，本研究で取り扱う問題であるカタログ統合問題について定義する．第 3 章では，カタログ統合問題に対して，どのようなアプローチが採られて来たかについての関連研究を述べる．第 4 章では，本論文で提案するカタログ統合問題の解決法について述べる．第 5 章では，提案手法の有効性を確認するために行った実験について述べる．この実験では，実世界で使われているインターネットディレクトリを用いて，従来手法よりも高い精度で統合が行えることを示す．最後の第 6 章では，本研究をまとめると同時に，今後の課題について述べる．

2. カタログ統合問題

本論文で取り扱う問題を明確化するために，この章では，カタログ統合問題と名付けて問題の定式化を行う．この問題では，2 つのカタログが与えられることを想定する．1 つを元カタログとし，もう 1 つを目標カタログとする．それぞれのカタログは，下記の要素により構築されているものとする．

- 元カタログ S_c は，カテゴリ集合 $C_{s1}, C_{s2}, \ldots, C_{sn}$ を含み，それらのカテゴリは，”is-a” 関係によって階層化されている．各々のカテゴリは，何かの属性を持つ情報インスタンスを含むことができる．
- 目標カタログ T_c は，カテゴリ集合 $C_{t1}, C_{t2}, \ldots, C_{tm}$ を含み，それらのカテゴリは，”is-a” 関係によって階層化されている．各々のカテゴリは，何かの属性を持つ情報インスタンスを含むことができる．

ここで，元カタログのインスタンスを目標カタログに統合する問題を考える．全てのインスタンスを目標カタログの適切なカテゴリに割り当てることができれば，目標カタログの概念階層を使用して，全てのインスタンスを利用することができるようになり，2 つのカタログを統合したものを目標カタログの概念体系を利用して作ることが可能となる．例えば，ある E コマースサイトが独自の商品分類体系を持っており，その商品群の中に，ある会社の使用している商品群を追加するような場合が，これに相当する．この場合には，E コマースのサイトが使っている商品分類体系が目標カタログ，ある会社の使用している商品分類体系が元カタログ，それぞれの商品がインスタンスとなる．

ここでの定式化では，元カタログ S_c に含まれるインスタンス I_{si} に対して，目標カタログ T_c に含まれる適切なカテゴリ C_t を割り当ててやる問題になる．ただし，目標カタログ T_c にすでに含まれているインスタンスは，適切なカテゴリを探す必要がないため，S_c に含まれていて，T_c に含まれていない全てのインスタンスに対してだけ，適切なカテゴリを割り当ててやればよいこととなる．

3. 関 連 研 究

第 2 章で述べた問題に対する最も簡単なアプローチは，一般的な機械学習手法を適用してやることである．カテゴリの階層性，元カタログの分類情報を使わない場合には，Naive Bayes [Mitchell 97] のような機械学習手法が適用可能となる．目標カタログ中の各カテゴリに含まれるインスタンスを訓練データとして，各カテゴリに対する分類器を学習し，その分類器で元カタログに含まれるインスタンスを分類してやれば統合を実現できる．このアプローチは，すべてのカテゴリを等価に扱うため，目標カタログに含まれるカテゴリ数が多くなると，適用が困難になる．このような問題を解決するために，目標カタログで階層的に分類されている特徴を利用した文書分類のアプローチ [Koller 97, McCallum 98, Wang 99, Dumais 00, Sun 01] がある．これらのシステムでは，階層性を利用して各カテゴリにインスタンスの分類を試みるが，元カタログにある分類の情報を使うことはできない．元カタログの分類情報を用いるアプローチとして，Enhanved Naive Bayes[Agrawal 01] がある．Enhanced Naive Bayes は，元カタログの分類情報を利用することで，従来の Naive Bayes を用いた場合よりも高い精度で分類をすることが可能となり，信頼性の高いカタログ統合を実現している．同様に LSD [Doan 03] では，元カタログにあるカテゴリ間の関係を制約として使うことで，Naive Bayes 手法を拡張している．しかし，この手法では，基本的な分類性能は，Naive Bayes の性能に依存することになる．HICAL [市瀬 02] は，インスタンスに含まれる情報を利用せずに，カタログ同士の分類の類似性に注目をして統合規則を学習し，カタログ統合を実現する．HICAL の手法を用いると，Enhanced Naive Bayes などの文書分類アプローチよりも高い性能が得られることが分かっている [Ichise 03]．しかし，インスタンスの属性情報を利用することで，さらに高い精度のカタログ統合を期待できる．本研究では，このようなアプローチで HICAL を拡張する手法を提案する．HICAL については，次章で詳しい解説を行う．

この問題に対する別のアプローチとして，オントロジー統合の手法がある．オントロジー統合手法は，階層的概念で作られた 2 つのオントロジーを統合するためのものであり，Chimaera [McGuinness 00] や PROMPT [Noy 00] がその例として挙げられる．これらは，2 つのカタログ間で概念ラベルのマッチを取る処理を行い，統合できそうな概念を人間に提示するシステムである．そこでは，統合には人間の介在が必要となる．別のタイプのオントロジー統

合システムである FCA-Merge [Stumme 01] は，概念の属性を使って異なったオントロジーを統合する．この手法では，はじめにあった 2 つのオントロジーと関係なく統合を行うため，はじめからある概念階層と全く異なる概念階層を作ってしまうことがある．また，Calvanese らは，局所的オントロジーと大局的オントロジーの観点から，オントロジー統合について論じている [Calvanese 01]．カタログ統合の観点からは，Omelayenko らが，抽象的な概念階層を構築して，2 つの階層を統合する手法 [Omelayenko 01] について述べている．この手法では，元カタログから，目標カタログに対する統合を抽象的な概念階層を経由することで実現する．この手法は，多くのカタログの統合を実現する時に有用な方法となるが，共通して使える抽象的な概念階層を構築することが困難な問題点がある．

応用的な側面からの類似研究としては，Siteseer [Rucker 97] や kMedia [濱崎 02] などのブックマーク共有システムが挙げられる．これらのシステムでは，元ブックマークにあり，目標ブックマークにない URL の情報を共有することを目的としている．これらのシステムでは，ページ内の同じ意味の語の数や同一の URL の数を用いて，カテゴリ間の対応関係を抽出している．

4. 提 案 手 法

第 2 章で述べた問題を解決するシステムとして，HICAL [市瀬 02] がある．本論文では，HICAL の手法を拡張することで，第 2 章で述べた問題に対して，さらに性能を改善することを試みる．この章では，HICAL の手法についてまず述べ，次にその拡張方法 HICAL-NB について述べる．

4·1 HICAL

HICAL [市瀬 02] は，第 2 章で述べたカタログ統合を実現するためのシステムとして提案された．その基本的なアイデアは，インスタンスの分類の類似性に着目することである．例えば，元カタログ中のカテゴリ C_{si} に分類されているインスタンスの多くが，目標カタログ中のカテゴリ C_{tj} に含まれるのならば，これらの分類基準は似ていると判断できるので，C_{si} と C_{tj} は類似カテゴリだと判断できる．その時，C_{si} にあって C_{tj} にないインスタンス I は，分類が類似しているので，C_{tj} にも分類される可能性が高いであろう．HICAL では，この性質を使って，カタログ統合を実現する． 図 1 は，その例を図示したものである．この図は，左に元カタログ S_c，右に目標カタログ T_c があり，下層にある 2 つのカテゴリ C_s と C_t が類似カテゴリの組と判定された場合を示している．この時，元カタログ中の C_s にあるインスタンス I は，目標カタログ中の C_t にそのまま割り当てることが可能となる．

次に，この分類の類似性をはかる手法について述べる．HICAL では，分類の類似性をはかるために κ 統計量 [Fleiss 73] を用いる．κ 統計量では，2 つのカテゴリに対して，表 1 のような分割表を作成する．表 1 はあるカテゴリ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ものである．C_s, C_t は，それぞれカタログ S_c, T_c の中にあるカテゴリを示し，m_{**} はそれぞれに含まれるインスタンスの数を示している．階層の中位にあるカテゴリに対して，この数を計算するには，HICAL では概念の階層性を利用する．下位のカテゴリに含まれるインスタンスは，上位のカテゴリにも含まれるとして計算する． 表 1 では，2 つの分類基準が近ければ，m_{11}, m_{22} の数が多くなり，m_{12}, m_{21} の数が少なくなる．逆に分類基準が遠ければ，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，この性質を利用して 2 つの分類基準が等しいか否かをある有意水準で判定する．

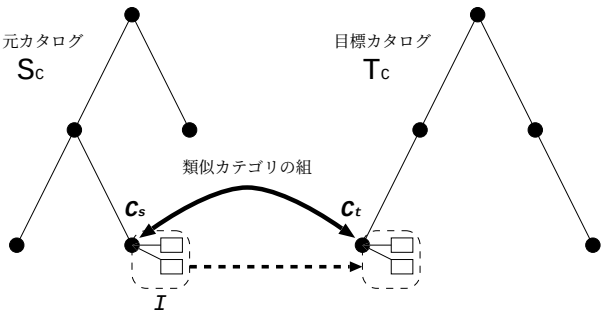


図 1 目標カタログへのインスタンスの割り当て方法

表 1 2 つのカテゴリによるインスタンスの分割

		カテゴリ C_t	
		含まれる	含まれない
カテゴリ C_s	含まれる	m_{11}	m_{12}
	含まれない	m_{21}	m_{22}

κ 統計量では，まず，概念基準の一致率 P と偶然の一致率 P' を次の式により計算する．

$$P = \frac{m_{11} + m_{22}}{m_{11} + m_{12} + m_{21} + m_{22}}$$

$$P' = \frac{(m_{11} + m_{12})(m_{11} + m_{21})}{(m_{11} + m_{12} + m_{21} + m_{22})^2} + \frac{(m_{21} + m_{22})(m_{12} + m_{22})}{(m_{11} + m_{12} + m_{21} + m_{22})^2}$$

その時，κ 統計量は，次式で表される．

$$\kappa = \frac{P - P'}{1 - P'}$$

次に，二つの概念基準の一致率が 0 である事を意味する $\kappa = 0$ であるかの検定を行う．そのために，次の値 Z を計算する．

$$Z = \kappa \sqrt{\frac{(m_{11} + m_{12} + m_{21} + m_{22})(1 - P')}{P'}} \quad (1)$$

Z は正規分布に従うため，有意水準を 5%とした時に，次の式を満たせば，概念基準の一致率が 0 であるとの帰無仮説が棄却される．

$$Z \geq 1.64486$$

仮説が棄却される時には，概念基準が一致している，すなわち，類似カテゴリの組と判定できる．

κ 統計量は，ある危険率で，2 つの概念分類に対する判断基準が一致しているかを判定する数学的な検定手法の一つである．そのため，統計的に有意と言える数のデータが無い場合には，帰無仮説が棄却できなくなり，結果として類似カテゴリとの判定は行われない．もし，あるカテゴリに対して，類似カテゴリが見付からない場合には，HICAL は，そのカテゴリの親カテゴリに対する規則を利用して，目標カタログのカテゴリを推定する．

HICAL では，類似カテゴリの組を見付けるためにすべてのカテゴリの組合せについて，κ 統計量を計算するのではなく，階層の上位から必要そうな場所だけを探索するようになっている．この部分の詳細を知りたい場合には，[市瀬 02] を参考にして頂きたい．

4·2 HICAL-NB

HICAL では，インスタンスを分類しているカテゴリの類似性だけを利用しており，インスタンスが持っている属性は，全く利用せずにカタログ統合をする．一方，Enhanced Naive Bayes [Agrawal 01] では，インスタンスが持っている属性と元カタログの分類の両方の情報を使って統合をする．HICAL では，Enhanced Naive Bayes で利用している分類情報のみを利用して，Enhanced Naive Bayes を上回る性能を出していると言えるが，Enhanced Naive Bayes と同様にインスタンスの属性も利用することができれば，さらに性能を向上させることが可能であると考えられる．

本論文では，HICAL でインスタンスの属性も利用できるようにするために，階層的文書分類手法と HICAL を組み合わせる多戦略アプローチを提案する．HICAL では，元カタログにあるカテゴリに一致するカテゴリが目標カタログにない時には，それを包含する上位のカテゴリに分類することが多い．したがって，最も適切な分類カテゴリは，HICAL が目標カタログ中で割り当てたカテゴリよりも下の階層に出現する可能性が高いと言える．このような例を図 2 で説明する．図 2 のように，2 つのインスタンスがあり，元カタログでは，1 つのカテゴリ C_s に，目標カタログでは共通の親カテゴリ C_{t1} を持つ別々のカテゴリ C_{t2}, C_{t3} に分類されているとする．その時，HICAL では，類似カテゴリの組として，C_s と C_{t1} の組合せを発見し，統合規則として提示する．なぜならば，C_s と C_{t2}，C_s と C_{t3} の組合せでは，どちらも片方のインスタンスを含まなくなってしまうからである．このように HICAL では，最適なカテゴリよりも一般的なカテゴリに統合されやすい．この問題を解決するためにインスタンスの属性を用いることを考える．

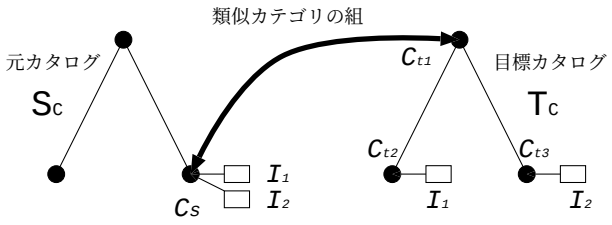


図 2 学習された規則の例

本論文で提案する手法では，HICAL によって一般的なカテゴリに統合された後に，インスタンスの属性を使ってより特殊なカテゴリに再分類することを考える．つまり，HICAL で統合先が決まった後に，そこを基準カテゴリとして，基準カテゴリから下のカテゴリに対して，文書分類手法を使って，再分類をしてやることで最終的な統合先のカテゴリを決めてやるというアプローチである．この提案手法を実現するために，HICAL システムに，文書分類システムである Naive Bayes を統合した HICAL-NB システムを構築した．このアプローチで用いることが可能な文書分類手法は，Naive Bayes 手法に限らないが，本論文では，このアプローチが有効であるかどうかを調べるため，広く使われており，容易に実現可能なパチンコ型 [McCallum 98]Naive Bayes を採用した．

パチンコ型 Naive Bayes とは，階層のトップから順に分類先を決めて行く機構を用いた Naive Bayes 手法である．Naive Bayes は，インスタンスの属性から分類器を作る学習手法の一つであり，インスタンスの属性から最も確率が高いカテゴリを推定して分類を行う．分類したいインスタンスが属性 $a_1, a_2, \ldots, a_n$ を持ち，カテゴリ $C_1, C_2, \ldots, C_m$ のどれかに分類をする時，Naive Bayes は，次式 V が最も高くなるカテゴリに分類を行う．

$$V = P(C_j | a_1, a_2, \ldots, a_n)$$

この式は，ベイズ規則を用いると

$$V = \frac{P(a_1, a_2, \ldots, a_n | C_j) P(C_j)}{P(a_1, a_2, \ldots, a_n)} \qquad (2)$$

となる．式 (2) では，分母は各カテゴリに関係なく一定になるので，分子が最大になるカテゴリを求めれば，分類先のカテゴリを決定できる．Naive Bayes では，式 (2) の値が最大になるカテゴリを計算するために，各属性は各カテゴリに対して独立であるとの仮定をおく．その時，式 (2) の分子は，次式で書ける

$$V' = P(C_j) \prod_i P(a_i | C_j) \qquad (3)$$

これを計算するために，訓練例のインスタンスの属性を用いて $P(a_i|C_j)$ の値を計算してやり，分類先のカテゴリを決定する．第 2 章で定義したカタログ統合問題においては，元カタログ S_c に含まれていて，目標カタログ

T_c に含まれないインスタンスの属性に対して，目標カタログ T_c に含まれるインスタンス $I_{t1}, I_{t2}, \ldots, I_{tv}$ の属性を用いて式 (3) の $P(a_i|C_j)$ を計算し，分類先のカテゴリを決定する．

一方，パチンコ型とは，パチンコのように上から順番に行き先を分けて行く方法であり，概念階層の内部ノードにおいて，そのノードの直下にあるカテゴリに対して分類器を作成し，その分類器を使ってインスタンスを下位のカテゴリに分類することを葉ノードにたどり着くまで繰り返し続けて行く手法である [McCallum 98]．内部ノードにおいてカテゴリを選択する際に，Naive Bayes 手法を用いたのが，パチンコ型 Naive Bayes である．ただし，実装システムでは，階層の中間に出現するカテゴリに対してもインスタンスの分類ができるようにするため，階層の中間カテゴリに分類されているインスタンスで一つのカテゴリを作り，下位のカテゴリと同列で分類規則を作ることとした．これらのシステムの実装には，HICAL および Naive Bayes システムの Rainbow[McCallum 96] を利用している．

5. 実　　験

5·1 実 験 方 法

本論文で提案した手法の有効性を確認するために実験を行った．実験では，カタログとして，実際に使われているインターネットディレクトリ Yahoo! [Yah 03] と Google [Goo 03]*1を用い，それぞれのインターネットディレクトリ中で分類されている URL をインスタンスとして用いた．データは，2003 年 12 月から 2004 年 1 月までの間に収集したものを用いた．Yahoo!と Google で使った概念階層は次に示すカテゴリとそのサブカテゴリである．

- Yahoo! : Recreation / Automotive
  Google : Recreation / Autos
- Yahoo! : Entertainment / Movies_and_Film
  Google : Arts / Movies
- Yahoo! : Recreation / Outdoors
  Google : Recreation / Outdoors
- Yahoo! : Arts /Visual_Arts / Photography
  Google : Arts / Photography
- Yahoo! : Computers_and_Internet / Software
  Google : Computers / Software

表 2 は，それぞれのインターネットディレクトリに含まれるカテゴリの数とインスタンスの数，両方のディレクトリに含まれるインスタンスの数を表している．この実験では，両方のディレクトリに出現しているインスタンスを使って 10-fold のクロスバリデーションを実施して，HICAL と Naive Bayes，HICAL-NB のそれぞれで正答率を計測した．ここで使用したインスタンスは両方のディレクトリに出現しているため，両方の分類体系において，どのカテゴリに分類するべきかが既知である．このインスタンスの集合をランダムに 10 分割し，そのうちの 9 個を訓練例として規則を学習し，学習された規則を残りの 1 個のテスト例に適用して，1 回の正答率を計測した．テスト例を変えることで合計 10 回の実験を行い，得られた正答率の平均をこの実験での正答率とした．そして，HICAL-NB でどの程度の性能向上が見られるかを評価した．ここで用いた HICAL は，第 4·1 節で述べた手法であり，Naive Bayes は，第 4·2 節の式 (3) で説明した手法である．Naive Bayes を使う際には，HICAL-NB と同様に，中間階層も一つのカテゴリとして扱い，パチンコ型の Naive Bayes を利用した．Naive Bayes を用いるには，インスタンスとなる文書から属性となるキーワードを抽出し，インデクシングする必要があるが，今回はまず URL が指す Web ページを取得し，そこから Rainbow を用いて，アルファベット以外の文字を含む語，2 文字以下の語，ストップワード*2を取り除いた上で，残る語全てをインデクスとした．これは Naive Bayes と HICAL-NB で同じ手法を用いた．なお，アクセス不能などで，Web から取得できなかったページは実験に用いていない．

実験時には，カテゴリの数と分類精度の関係を見るために，階層の深さ毎にデータを取った．各階層の深さと，実験に使ったカテゴリの数，およびその階層に含まれるインスタンスの数は，表 3 のようになっている．表に記載されている数はその階層にあるカテゴリの数を示し，括弧内の数はその階層に割り当てられているインスタンスの数である．ここで深さ 1 とは，上記で挙げた対象カテゴリの子カテゴリを表している．例えば，Yahoo!の Autos のドメインにおいて，深さ 2 で統合する場合には，深さ 2 までのカテゴリの総計 128(=1+23+104) が統合対象のカテゴリ数となる．この表で共通のインスタンスの数と各階層にあるインスタンスの数の合計は，同じインスタンスが複数のカテゴリに割り当てられている場合があるので，一致しないこともある．

5·2 実 験 結 果

実験結果は，図 3 のようになった．縦軸が正答率を表し，横軸が使った階層の深さを表している．ここで正答とは，目標カタログで分類されているカテゴリに一致した場合を指し，もし，正答のカテゴリが実験で使用したカテゴリの深さよりも下位に位置する場合は，正答のカテゴリの上位にあるカテゴリの中で，実験対象とするカテゴリの一番下位のカテゴリを正答とした．また，目標カタログで複数のカテゴリに分類されている場合には，どれかに分類されれば正答とした．ここでは，目標カタログで分類されているカテゴリに完全に一致するもの以外にも，意味的には親カテゴリなども正答であると考え

*1 Google のデータは，dmoz [Dmo 03] から作られているため，データは Dmoz を使って収集した．

*2 SMART[Buckley 85] のストップワードリストを利用

表 2 インターネットディレクトリの統計データ

	Yahoo!		Google		共通の
	カテゴリ数	インスタンス数	カテゴリ数	インスタンス数	インスタンス数
Autos	885	5134	874	9702	544
Movies	5297	19192	7947	36288	1480
Outdoors	2590	7960	1221	17065	362
Photography	578	5548	278	5443	305
Software	513	3268	2339	41883	353

表 3 各階層で共通インスタンスが存在するカテゴリの数とその階層に存在する共通インスタンスの数

深さ	Autos		Movies		Outdoors		Photography		Software	
	Yahoo!	Google	Yahoo!	Google	Yahoo!	Google	Yahoo!	Google	Yahoo!	Google
0	1(4)	1(0)	1(0)	1(0)	1(3)	1(0)	1(0)	1(0)	1(2)	1(0)
1	23(50)	11(24)	30(158)	27(74)	21(69)	18(25)	23(71)	9(19)	19(39)	21(10)
2	104(180)	99(155)	95(203)	118(210)	63(232)	60(115)	31(67)	26(57)	42(102)	59(67)
3	81(109)	154(272)	219(161)	729(840)	85(46)	62(98)	46(158)	65(143)	44(81)	102(109)
4	115(186)	50(94)	682(888)	178(234)	90(244)	42(62)	8(42)	37(74)	53(115)	101(115)
5	35(67)	8(10)	270(341)	105(136)	40(41)	57(69)	-	3(12)	23(42)	39(36)
6	3(5)	-	62(85)	3(3)	12(23)	4(5)	-	1(4)	2(6)	24(31)
7	-	-	3(4)	-	3(4)	-	-	-	-	2(8)
8	-	-	-	-	-	-	-	-	-	-

られる．[市瀬 02] では，カテゴリが完全に一致するものを正答とした場合と，親カテゴリも正答とした場合の両方の結果について報告している．また，[Ichise 03] では，[Agrawal 01] に掲載されている結果と比較をするために，親カテゴリも正答とした結果を掲載している．しかし，本論文では，もっともふさわしいカテゴリを自動的に発見して統合する精度の高い手法の開発を目的にしているため，カテゴリが完全に一致するもののみを正答と取り扱うこととした．その結果，[市瀬 02, Ichise 03] が用いた正答の条件よりも，正答になるカテゴリ数が減っているため，これらの論文よりも正答率が低くなっている．なお，親カテゴリも正答と見なすような大まかな統合を行いたい場合には，厳密な解を発見しようとする HICAL-NB よりも，階層の上位に分類されやすい HICAL の方が向いている．

図 3 のグラフは，上から順に Autos, Movies, Outdoors, Photography, Software の結果であり，左側のグラフが Google を元カタログ，Yahoo!を目標カタログにした場合，右側のグラフが Yahoo!を元カタログ，Google を目標カタログにした場合を示している．図 3 では，HICAL の結果と Naive Bayes の結果，HICAL-NB の結果が併記してあり，HICAL-NB がこの論文で提案している手法での結果である．また，詳しくは後述するが，HICAL を適用した後に，HICAL-NB が到達可能な正答率の理論的上限値を Upper Bound として記載している．

5･3 考　　察

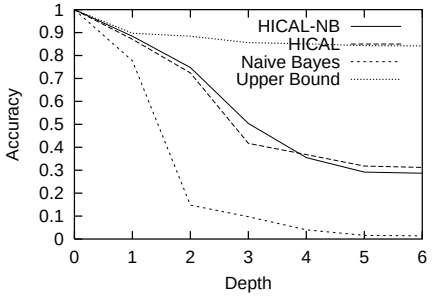
まず，図 3 より，階層が深くなるにつれて，正答率が低下していくのが分かる．表 3 が示しているように，階層が深くなると急激に統合先のカテゴリ数が多くなる．例えば，Movies を対象とした問題では，深さが多くなると，千以上のカテゴリがある分類問題となる．このような多数のカテゴリのどこかにインスタンスを分類して正解とするのは，非常に困難な問題設定であると言える．そのような困難な問題でも，この手法ではある程度の正答率を保っていると言えるであろう．

次に，Naive Bayes と比較すると，全ての場合において，提案手法の方が正答率が高くなっていることが分かる．この結果より，全体として提案手法の方が Naive Bayes よりも優れていると言えるであろう．

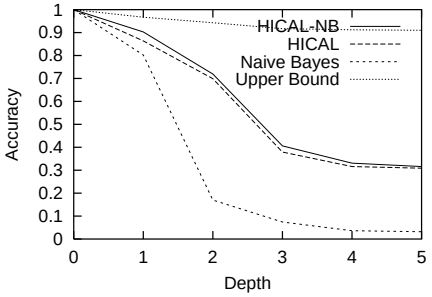
HICAL と比較すると，Yahoo!の Automotive，Outdoors に統合する場合以外は，ほぼ提案手法の方が正答率で上回っている．また，その 2 つの分野においても，差は非常にわずかであり，HICAL と比較しても提案手法はよい手法であると言えるであろう．提案手法は，Naive Bayes の拡張と受け取ることもできるが，HICAL が Naive Bayes の拡張手法 [Agrawal 01] と比べて，良い性能を出している [Ichise 03] ことを勘案すると，ここで提案した手法は，Naive Bayes 手法の拡張としても，十分な能力を持っていると言えるであろう．

次に，図 3 から分かるように，提案手法は，深さが小さい場所や大きい場所よりも，中間階層で HICAL よりも性能が改善していることが多いと言える．HICAL-NB は，大まかな分類を行う概念階層の上位では，HICAL の機構を利用している．このため，階層の上位で HICAL が分類先を間違えてしまった場合には，HICAL-NB では正答を発見することが不可能となる．この割合を明示的に示すために，HICAL を適用した後に，HICAL-NB において到達可能な正答率の上限を計算したものが，図 3 の Upper Bound である．[市瀬 02] では，HICAL に

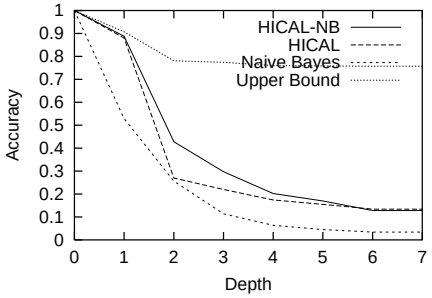
Yahoo!: Automotive に統合



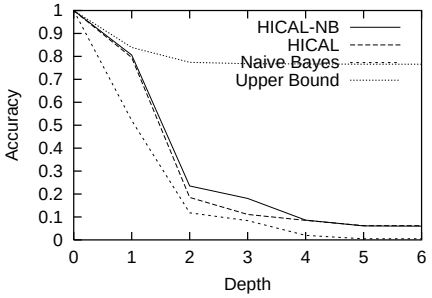
Google: Autos に統合



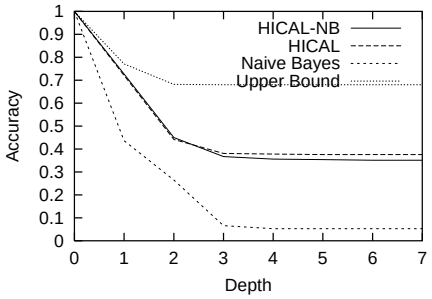
Yahoo!: Movies_and_Film に統合



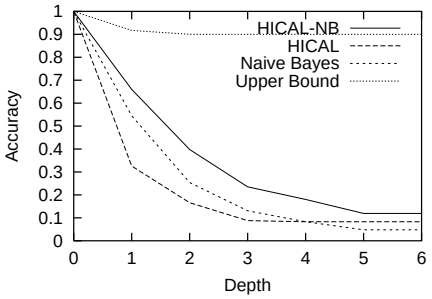
Google: Movies に統合



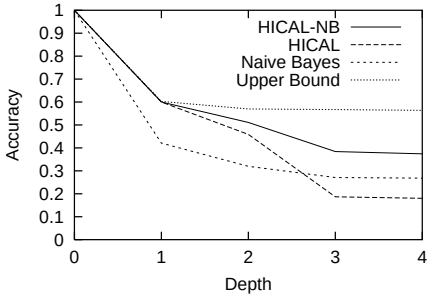
Yahoo!: Outdoors に統合



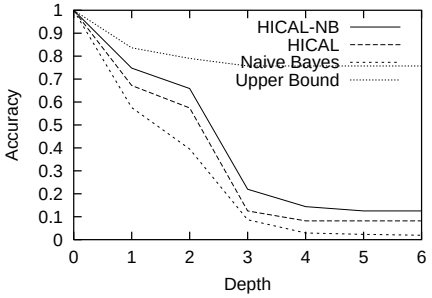
Google: Outdoors に統合



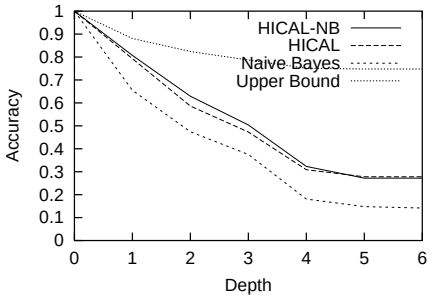
Yahoo!: Photography に統合



Google: Photography に統合



Yahoo!: Software に統合



Google: Software に統合

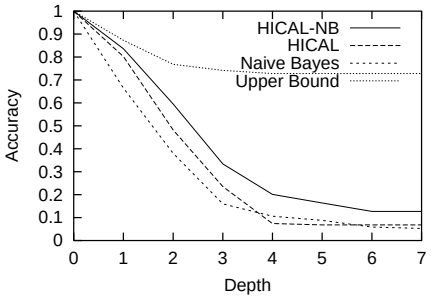


図 3 実験結果

対して，親のカテゴリに分類した時も正答とみなす正答率も報告している．この論文によると，親カテゴリも正答と見なした場合には，HICAL の正答率はおよそ 6 割から 9 割となっている．一方，親カテゴリを正答と扱わない場合には，それに比べておよそ 3 割から 5 割低い正答率となっており，HICAL-NB では，およそ 4 割程度の精度向上の余地が残されていることが分かる．これは，第 4･2 節の図 2 で示されたような状況が約 4 割のインスタンスで起こっているとも言える．このような状況は中間階層で起こりやすい．全く同じ条件を用いた実験でないため，一概には言えないが，本実験では，数%から 2 割程度の性能向上が見られるため，HICAL-NB では，HICAL で起こっているこのような問題状況の半分程度までを解決していると考えられる．特に，中間階層では，深い階層に比べて性能の向上が見られており，提案アプローチが有効に働いていると考えられる．

最後に，Naive Bayes, HICAL の結果と HICAL-NB の結果を比較すると，Naive Bayes の結果がよい時には，HICAL-NB の結果も向上すると言える．Naive Bayes の影響は，図 3 の Upper Bound を上限とした HICAL-NB と HICAL の差で見ることができる．例えば，Photography の階層を使った実験では，提案手法が HICAL よりも大きく改善している．これは，インスタンスの属性を利用する分類手法，つまり，Naive Bayes 手法が大きな役割を果たしているためであると考えられる．このことは，深い階層において Naive Bayes 手法が HICAL を上回る性能を出している点からも分かる．一方，Naive Bayes の結果が悪い時には，提案手法は，少なくとも HICAL と同等の性能を出しており，Naive Bayes のために性能低下を引き起こすような事態を招いていない．例えば，Autos の領域では，Naive Bayes の正答率が非常に低いが，HICAL と HICAL-NB の性能がほぼ同じになっている．本論文で提案した HICAL-NB は，Naive Bayes や HICAL を上回る性能を多くの領域で出しており，インスタンスの属性に基づく文書分類のアプローチと分類の類似性を利用するアプローチをうまく統合していると言えるであろう．

6. ま と め

本論文では，階層的に分類されたデータを統合する問題をカタログ統合問題として定式化し，その問題に対処するための新たな手法の提案を行った．提案手法は，カテゴリの類似性を利用したアプローチと文書分類を利用したアプローチを組み合わせた多戦略のアプローチを用いており，この手法によって，一般的な階層に統合されやすいという以前の手法の問題に対処することが可能となった．このことを示すために実際に使われているインターネットディレクトリを用いて実験を行った．実験の結果，本論文で提案した手法は，ほとんどの実験で，従来手法よりも性能で上回ることが分かった．

本論文で提案した手法は，以前の手法よりも有効であることが示されたが，まだ研究が必要な部分がいくつか残っている．まず 1 点目として挙げられることは，他の文書分類手法と組み合わせることの検討である．本論文では，文書分類手法と組み合わせた場合に性能向上が見られるか否かを検討するために，Naive Bayes を用いた．しかし，Naive Bayes 以上の性能が得られる文書分類手法として SVM[Cristianini 00] などの手法もある．このような手法も HIACL-NB と同様なアプローチによって組み合わせることが可能である．そのような組み合わせについても検討を行っていきたいと考えている．次に挙げられることは，この手法を 3 つ以上のカタログ統合に展開していくことである．現在は，2 つのカタログを想定しており，3 つ以上の場合には，2 つずつ順次統合していくことで対応できると考えている．しかし，元カタログが複数あった場合には，それらの情報をうまく利用してやることによって，より正確なカタログ統合が可能になると考えられる．最後に，共通のインスタンスが必要という制約の緩和である．[濱崎 03] では，インスタンスに出現する属性によって，共通のインスタンスが少ない場合でも，カタログ統合がうまくいくようにすることを提案している．インスタンスの属性だけでなく，オントロジー統合で用いられるようなカテゴリの属性も利用することで，この制約を緩和する手法について，検討していきたいと考えている．

謝　辞

本研究では，Naive Bayes のプログラムとして Rainbow[McCallum 96] を用いた．有用なプログラムを提供して下さった McCallum 博士に感謝する．

◇ 参 考 文 献 ◇

[Agrawal 01] Agrawal, R. and Srikant, R.: On Integrating Catalogs, in *Proceedings of the Tenth International World Wide Web Conference*, pp. 603–612 (2001)

[Buckley 85] Buckley, C.: Implementation of the SMART information retrieval system, Technical Report TR85-686, Department of Computer Science, Cornell University, Ithaca, NY 04853 (1985)

[Calvanese 01] Calvanese, D., Giacomo, G. D., and Lenzerini, M.: A framework for ontology integration, in *Proceedings of the First Semantic Web Working Symposium*, pp. 303–316 (2001)

[Cristianini 00] Cristianini, N. and Shawe-Taylor, J.: *An Introduction to Support Vector Machines*, Cambridge University Press (2000)

[Dmo 03] Dmoz, http://dmoz.org/ (2003)

[Doan 03] Doan, A., Domingos, P., and Halevy, A.: Learning to Match the Schemas of Data Sources: A Multistrategy Approach, *Machine Learning*, Vol. 50, No. 3, pp. 279–301 (2003)

[Dumais 00] Dumais, S. T. and Chen, H.: Hierarchical classification of Web content, in Belkin, N. J., Ingwersen, P., and Leong, M.-K. eds., *Proceedings of the 23rd ACM International Conference on Research and Development in In-*

formation Retrieval, pp. 256–263, Athens, GR (2000), ACM Press, New York, US
[Fleiss 73] 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
[Goo 03] Google, http://www.google.com/ (2003)
[濱崎 02] 濱崎 雅弘, 武田 英明, 松塚 健, 谷口 雄一郎, 河野 恭之, 木戸出 正継：Bookmark からの共通話題ネットワークの発見手法の提案とその評価, 人工知能学会論文誌, Vol. 17, No. 3, pp. 276–284 (2002)
[濱崎 03] 濱崎 雅弘, 武田 英明, 市瀬 龍太郎：階層的知識と内容的類似性を用いたインターネットディレクトリの統合, 第 17 回人工知能学会全国大会論文集 (2003), 1D4-07
[市瀬 02] 市瀬 龍太郎, 武田 英明, 本位田 真一：階層的知識間の調整規則の学習, 人工知能学会論文誌, Vol. 17, No. 3, pp. 230–238 (2002)
[Ichise 03] Ichise, R., Takeda, H., and Honiden, S.: Integrating Multiple Internet Directories by Instance-based Learning, in *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, pp. 22–28 (2003)
[Koller 97] Koller, D. and Sahami, M.: Hierarchically classifying documents using very few words, in Fisher, D. H. ed., *Proceedings of the 14th International Conference on Machine Learning*, pp. 170–178, Nashville, US (1997), Morgan Kaufmann Publishers, San Francisco, US
[McCallum 96] McCallum, A. K.: Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering, http://www.cs.cmu.edu/ mccallum/bow/ (1996)
[McCallum 98] McCallum, A. K., Rosenfeld, R., Mitchell, T. M., and Ng, A. Y.: Improving text classification by shrinkage in a hierarchy of classes, in Shavlik, J. W. ed., *Proceedings of the 15th International Conference on Machine Learning*, pp. 359–367, Madison, US (1998), Morgan Kaufmann Publishers, San Francisco, US
[McGuinness 00] 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 7th International Conference on Principles of Knowledge Representation and Reasoning*, pp. 483–493, Morgan Kaufman Publishers (2000)
[Mitchell 97] Mitchell, T. M.: *Machine Learning*, McGraw Hill (1997)
[Noy 00] 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*, pp. 450–455, Menlo Park (2000), AAAI Press
[Omelayenko 01] Omelayenko, B. and Fensel, D.: An Analysis of B2B Catalogue Integration Problems, in *Proceedings of the International Conference on Enterprise Information Systems*, pp. 945–952 (2001)
[Rucker 97] Rucker, J. and Polanco, M. J.: Siteseer: Personalized Navigation for the Web, *Communications of the ACM*, Vol. 40, No. 3, pp. 73–75 (1997)
[Stumme 01] Stumme, G. and Maedche, A.: FCA-Merge: Bottom-Up Merging of Ontologies, in *Proceedings of the 17th International Joint Conference on Artificial Intelligence*, pp. 225–230 (2001)
[Sun 01] Sun, A. and Lim, E.-P.: Hierarchical Text Classification and Evaluation, in Cercone, N., Lin, T. Y., and Wu, X. eds., *Proceedings of IEEE International Conference on Data Mining*, pp. 521–528, San Jose, CA (2001), IEEE Computer Society Press, Los Alamitos, US
[Wang 99] Wang, K., Zhou, S., and Liew, S. C.: Building Hierarchical Classifiers Using Class Proximity, in Atkinson, M., Orlowska, M. E., Valduriez, P., Zdonik, S., and Brodie, M. eds., *Proceedings of the 25th international Conference on Very Large Data Bases*, pp. 363–374, Los Altos, CA 94022, USA (1999), Morgan Kaufmann Publishers
[Yah 03] Yahoo!, http://www.yahoo.com/ (2003)

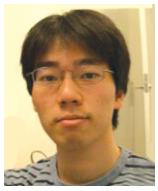
〔担当委員：中川裕志〕

2004 年 4 月 8 日　受理

著　者　紹　介

市瀬　龍太郎(正会員)

2000 年東京工業大学大学院情報理工学研究科計算工学専攻博士課程修了．博士 (工学)．同年より国立情報学研究所知能システム研究系助手．2001 年から 2002 年までスタンフォード大学言語情報研究所客員研究員．機械学習，知識発見，知識共有などの研究に従事．AAAI，電子情報通信学会，情報処理学会，日本認知科学会，各会員．



濱崎　雅弘(学生会員)

総合研究大学院大学博士後期課程在学中．2002 年奈良先端科学技術大学院大学情報科学研究科博士前期課程修了．オンラインコミュニティシステムの研究・開発に従事．情報処理学会，日本データベース学会，各学生会員．

武田　英明(正会員)

1991 年 3 月東京大学大学院工学系研究科博士課程修了．1993 年 4 月奈良先端科学技術大学院大学助手．1995 年 4 月同助教授．2000 年 4 月国立情報学研究所助教授．2003 年 5 月同教授．現在に至る．総合研究大学院大学情報学専攻教授兼務．人工知能特に知識共有，ネットワークコミュニティ，実世界エージェントなどの研究に従事．AAAI，電子情報通信学会，情報処理学会など各会員．