3F1-04

Web 文書を階層的に分類するための複数の分類器の利用

Hierarchical Classification of Web Documents Using Combination of Multiple Classifiers

市瀬 龍太郎[*1] Ryutaro Ichise　濱崎 雅弘[*2] Masahiro Hamasaki　武田 英明[*3] Hideaki Takeda

[*1]国立情報学研究所 知能システム研究系
Intelligent Systems Research Division, National Institute of Informatics

[*2]総合研究大学院大学 情報学専攻
Deparment of Informatics, Graduate University for Advanced Studies

[*3]国立情報学研究所 実証研究センター
Research Center for Testbeds and Prototyping, National Institute of Informatics

One method by which to managing large amounts of information is to utilize catalogs which organize information within concept hierarchies. In the present paper, we address the problem of integrating multiple catalogs for ease of use. 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 the previous method.

1. はじめに

Web のように大量の情報がある場合に，情報の整理法として，しばしば階層的分類手法が用いられる．本論文では，複数の階層的情報分類があった時に，これらを整合性を持った状態で統合する方法を提案する．本論文では，まずこの問題をカタログ統合問題として定式化する．次に，この問題に対応するために，文書分類のアプローチと分類の類似性に着目したアプローチを統合する多戦略手法を提案する．そして，提案手法を実装したシステムで従来の手法との比較を行い，提案手法の有効性を示す．

2. カタログ統合問題

本論文で対象とする問題を定式化するために，カタログ統合問題について述べる．カタログ統合問題では，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．

本論文では，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．つまり，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ついて取り組む．

連絡先: 市瀬 龍太郎, 国立情報学研究所 知能システム研究系, 〒101-8430 東京都千代田区一ツ橋 2-1-2, ichise@nii.ac.jp

表 1: 二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}

3. 提案手法

カタログ統合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，カタログ統合問題に対して，さらに性能を改善することを試みる．

3.1 HICAL

HICAL 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．

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，二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．表 1 では，二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_{12} 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．

3.2 HICAL-NB

この節では，HICAL を Naive Bayes(NB)[Mitchell 97] と組み合わせた HICAL-NB について述べる．HICAL では，インスタンスが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 NB[Agrawal 01] のような手法では，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 NB で利用している分類情報のみを利用して，Enhanced NB を上回る性能を出している [Ichise 03] が，Enhanced NB 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．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 NB を統合した HICAL-NB システムを構築する．このアプローチで用いることが可能な文書分類手法は，NB 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，容易に実装可能なパチンコ型 NB を採用した．

パチンコ型 NB 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 NB 手法である．NB 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．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．内部ノードにおいてカテゴリを選択する際に，NB 手法を用いたのが，パチンコ型 NB 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．

4. 実験

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!と Google[*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! : Arts /Visual_Arts / Photography
  Google : Arts / Photography

実験結果は，図 1 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[*2]．NB, HICAL の結果

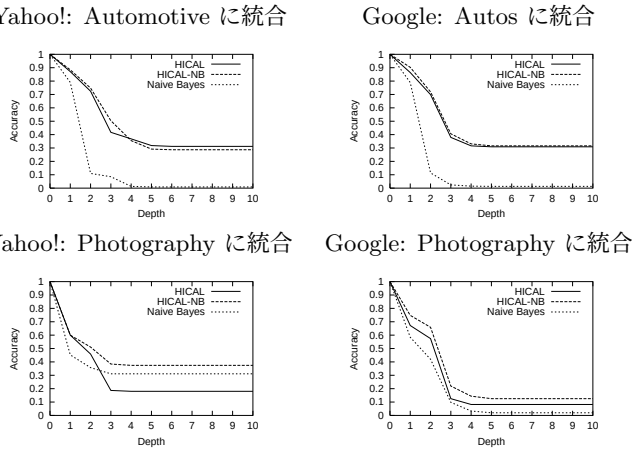

図 1: 実験結果

と HICAL-NB の結果を比較すると，NB の結果がよい時には，HICAL-NB の結果も向上すると言える．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．一方，NB 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，NB 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，NB 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．よって，本論文で提案した HICAL-NB は，両方のアプローチをうまく統合していると言えるであろう．

5. 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，実験の結果，本論文で提案した手法は，従来の手法の良い側面を統合できることが分かった．今後は，さらなる性能の向上を目指して，他のアプローチとの融合方法について考えて行きたい．

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)

[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

[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 AI*, pp. 22–28 (2003)

[Mitchell 97] Mitchell, T. M.: *Machine Learning*, McGraw Hill (1997)

∗1 Google のデータは，dmoz(http://www.dmoz.org/) から作られているため，データは dmoz を使って収集した．

∗2 本論文では，意味的には他のカテゴリも正答と考えられる場合でも，それを考慮せずに正答率を計算した．そのため，[市瀬 02],[Ichise 03] で用いた正答の条件よりも厳しい条件となり，これらの論文よりも正答率が低くなっている．