

他者の持つ知識を利用した階層的分類知識の獲得

Acquiring Classification Rules for a Concept Hierarchy using Category Alignment

市瀬 龍太郎

ICHISE Ryutaro

武田 英明

TAKEDA Hideaki

本位田 真一

HONIDEN Shinichi

国立情報学研究所 知能システム研究系

National Institute of Informatics

In the present paper, we propose a classification system using category alignment approach. We conducted an experiment to show the performance of our approach using real-world internet directories. The results of this experiment show that the proposed method can be used to classify information into appropriate categories within a concept hierarchy in internet directories.

1. はじめに

近年の情報技術の進展に伴い、さまざまな情報がネットワークを通じてやりとりされるようになってきている。それに伴い、利用者にとって情報を入手するためのコストは以前と比べて、非常に低くなってきた。しかし、情報のコスト的な入手が容易になるにつれ、膨大な情報のやりとりが安易に行われるようになり、かえって必要な情報が手に入り辛くなるという問題点が明らかになっている。このような問題は、特に Web ページを利用した情報流通の点で顕著であると言える。そこで、本研究ではこの問題に対する解決法の一つとして、情報の情報、すなわち知識を流通させることを提案する。現在の情報流通システムでは、このようなことが想定されていないため、情報だけが過多になってしまうと考えられるからである。

情報の情報を表す知識には、多数の種類があるが、本研究では階層的な分類知識を対象とする。階層的な分類知識は情報の管理情報としてインターネットディレクトリー、図書館の本の分類、オントロジーなどさまざまな分野で利用されている。これらの階層的な分類知識では、整合性を保たねばならないため、通常一つの分類体系だけが用いられており、違う分類体系を持っている他者の知識を利用することは難しい。本論文では、他者の持つ分類知識を利用するために、自分の持つ概念体系にあった分類規則を学習する手法を提案する。そうすることで、他者の持つ知識を自分に合せた形式で取り込むことができ、知識の流通が行えるようになって考えられるからである。

2. 階層的な分類知識の構造

まず、本論文で対象とする階層的な分類知識の定式化を行う。情報を階層的に分類する際には、最も一般的な概念を最上位概念として、ある基準にしたがって分類を行っていき、徐々に詳細化していく概念階層を利用する。詳細な情報になるほど、下位にくることになる。インスタンスとなる情報は、この分類階層の最上位から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。ここで、インスタンスとなる情報とは、Web ページなどが相当することになる。

このモデルを示した例が図 1 である。2 つの概念階層 C_1 , C_2 に対して、情報インスタンス I_1 , I_2 , I_3 が存在し、それぞれが

概念階層に割り当てられている。この時に、情報インスタンス I_2 は C_1 では分類が行われているが、 C_2 では分類が行われていない。このようなインスタンスの分類方法が C_2 に伝達されれば、知識の流通が行われたと考えられる。そこで、 C_1 とは概念階層の構造が異なる C_2 で、階層的な分類規則を学習させることが、この論文で行う学習タスクとなる。

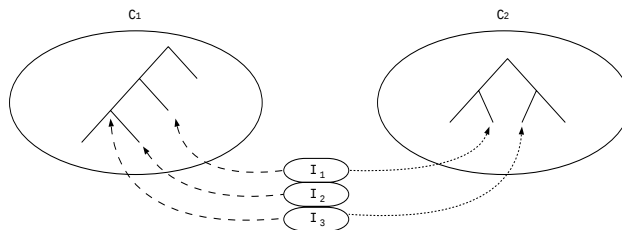


図 1: 階層的知識のモデル

3. 学習手法

前章で述べたように、2 つの異なる知識源がある時には、それぞれの対応関係が分からないため、他の知識源の持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、2 つの知識源が持つ分類基準の類似性を発見し、それらの分類間の変換規則として取り扱う事で、他が持つ知識を利用できるようにする手法を提案する。

2 つの概念階層から任意のカテゴリーを選んだ時に、ある情報インスタンスに対する判定はそのカテゴリーに属するか属さないかの 2 通りである。情報インスタンスがいくつかある時には、この判定の一致性を利用する事で、2 つのカテゴリーの分類基準の類似性を判定することができる。分類の類似性の判定には、統計を利用した検定手法の一つである κ 統計量 [Fleiss 73] を用いた。この手法を 2 つの概念階層の最も一般的なカテゴリーに適用し、一致性が閾値以上ならばその下位同士のカテゴリーの組合せについても判定をくりかえし行っていくことで、類似したカテゴリーの組を抽出することができる。次に、この抽出された 2 つの分類体系における類似カテゴリーについて、学習規則の生成を行う。規則は、ある概念体系 C_1 のあるカテゴリーに属する情報インスタンスは、別の概念体系 C_2 の類似カテゴリーにも属するという形式で作成が行われる。より詳しい学習機構については、[市瀬 01] を参照されたい。

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, Tel:03-4212-
2573, Fax:03-3556-1916, E-mail:ichise@nii.ac.jp

4. 実験

前章で提案する学習手法の有効性を確認するため、提案手法を計算機上に実装したシステム HICAL を作成して評価を行った。評価を行うためのデータとして、インターネットディレクトリーを利用した。インターネットディレクトリーは Web 文書を階層的に分類したものであり、Web 文書に対するメタ情報の一つとして広く用いられている。この実験では、Yahoo! Japan [Yahoo 00] と LYCOS Japan [Lycos 00] の「文学」より下の概念階層を利用した。

実験は図 2 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。

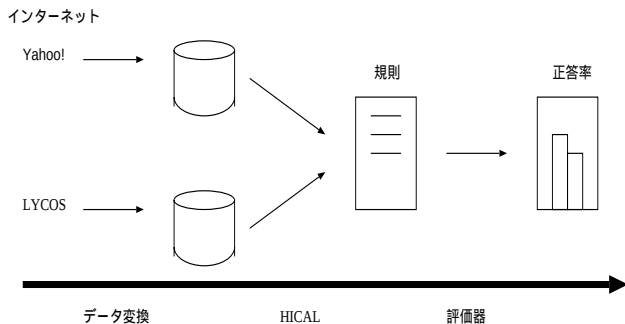


図 2: 実験手順

10 個のデータセットから得られた結果を平均した正答率は図 3 のようになった。縦軸は、テスト例で正しく分類された割合を表している。正しい分類は、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、その上のカテゴリーに分類される場合を利用した。テスト例と全く同じカテゴリーを利用することも考えられるが、その場合は、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ないため、一般的であるとは言えない。そこで、今回の評価尺度では、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現実的な評価方法をとった。「Exact Rules」は、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た場合の結果である。「Parent Rules」は、インスタンスが属しているカテゴリーに規則が割り当てられていない時に、その親の階層の規則を利用した場合の結果である。「Yahoo ⇒ Lycos」は Yahoo! にある情報インスタンスを Lycos に移動させる規則を評価した時の結果を示しており、「Lycos ⇒ Yahoo」はその逆の場合を示している。

図 3 の結果から分かるように、どの場合も 8 割以上の正答率があるため、うまく規則を学習していると言える。Yahoo! から Lycos への規則と、Lycos から Yahoo! への規則を比較すると、Yahoo! から Lycos の方がより正確な規則を学習している。Yahoo! と Lycos のカテゴリー数を比較すると、圧倒的に Yahoo! の方が多い。従って、Yahoo! から Lycos への規則は相

対的に詳しい分類体系から簡単な分類体系への規則となるため、学習が行いやすかったと考えられる。

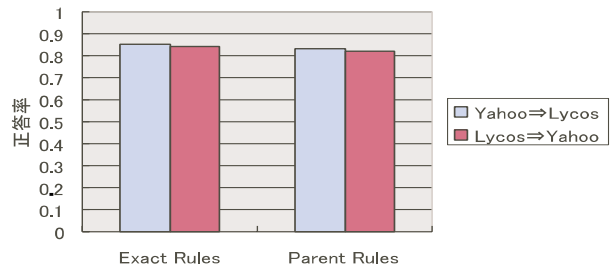


図 3: 実験結果

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い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い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いとは言いがたい。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. おわりに

本論文では、知識を流通させるための一つの手法として、他者が持つ知識を利用する分類規則の学習を提案した。そして、それを行うシステムを構築し、有効性を示した。本研究では、2 者の関係を取り扱ったが、今後は複数の概念体系から知識を学習していくような手法について検討を行っていく。

参考文献

- [Fleiss 73] Fleiss, J. L.: Statistical Methods for Rates and Proportions, John Wiley & Sons (1973), (邦訳: 係数データの統計学, 佐久間 昭 訳, 東京大学出版会 (1975))
- [市瀬 01] 市瀬 龍太郎, 武田 英明, 本位田 真一: インスタンスに基づく階層的知識源の統合, 人工知能学会 AI チャレンジ研究会 (2001) .
- [Yahoo 00] Yahoo! JAPAN, [http://www.yahoo.co.jp/\(2000\)](http://www.yahoo.co.jp/(2000)) .
- [Lycos 00] LYCOS JAPAN, [http://www.lycos.co.jp/\(2000\)](http://www.lycos.co.jp/(2000)) .