1D4-07

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

Integration of Internet Directories using Hierarchical Knowledge and Similarity in Contents

濱崎雅弘[*1*2] 武田英明[*1] 市瀬龍太郎[*1]
HAMASAKI Masahiro TAKEDA Hideaki ICHISE Ryutaro

[*1]国立情報学研究所 National Institute of Informatics
[*2]総合研究大学院大学 Graduate University of Advanced Studies

In this paper, we propose a method to integrate concept hierarchies using the conceptual structures and content-based technology. Concept hierarchies such as ontologies and information categorizations are powerful and convenient methods to solve information flood. We had already proposed the system *Hical* that can input information from a source concept hierarchy and finds suitable location for them in a target hierarchy. However, it cannot integrate them without sharing information among source concept hierarchies. We modified this system using keyword matching, and conducted experiments using two internet directories. We map information instances from the source directory into the target directory, and show that our method can overcome the problem of *Hical*.

1. はじめに

近年の情報ネットワークの発達により，個人が入手できる情報は，飛躍的に大きくなり，それと共に多くの情報をいかにして管理するかが重要になってきた．人間が情報を入手，または作成する時は，概念階層を用いて情報を管理することが多い．顕著な例で言えば，図書館における本の分類がある．最近では，情報自体を概念階層を用いて管理する XML も多くの場所で利用され始めている．

情報を管理するときには，それぞれの情報利用者の目的，収集している情報などにより，情報管理方法への要求が異なる．そのため，同じような概念階層を利用して情報を管理しているにもかかわらず，それぞれの管理者，利用者などによって別々の概念階層を用いて情報が管理されている事が多い．それは，概念階層に一貫性が必要な点や，分散管理の許容性の点などを考えると，現実的な方法ではあるが，情報の再利用性という観点からは効率が悪い．一方で，情報を一カ所で集中的に管理するという手法もある．しかし，その場合には，全て情報利用者の目的，得られる全ての情報などについて考慮しながら概念階層の設計を行わなければならないため，一貫性の維持が非常に困難になるなどの問題が生じる．

このような状況に対して，武田らは知識共生 [Takeda 01] という概念を提唱している．これは知識が分散して存在し，互いに交換し合うことで自分の持つ知識を拡張していくような環境を指す．知識共生環境では知識の分散管理が前提となるため，一貫性の管理が容易になる．しかし，このような環境下では，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きたとしても，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し，利用することは難しい．

この問題の解決策として，市瀬らは概念階層が持つインスタンスに基づいて他の知識との相違を調整する規則を学習する手法を提案し，その実装である *Hical* を試作した [市瀬 02]. *Hical* では，2 つの概念階層がインスタンスをそれぞれどのように分類しているかという情報を元に，概念階層間の結合を行っている．この手法の問題として，2 つの異なる概念階層間で，十分な数の同じインスタンスを分類していなくてはならないという点が挙げられる．

連絡先: 濱崎雅弘，国立情報学研究所，tel:03-4212-2681, fax:03-4212-1916, hamasaki@grad.nii.ac.jp

そこで本研究では，概念階層間で共有されていないインスタンスを，機械的に求めたインスタンス間の類似性に基づいて自動的に共有させることで，同じインスタンスを持っていない概念階層間でも結合が行えるように *Hical* を改良した．本論文では，改良したシステム *WebHical* の説明とそれを用いた評価実験について解説をする．

2. 階層的知識源

本研究で仮定する階層的知識源についてモデル化を行う．ここで対象とするのは，概念階層に基づいて分類が行われ，管理されている情報源である．例えば，インターネットディレクトリ，図書の分類目録，オントロジーなどがそのようなものとして挙げられる．これらの情報源における概念階層は，もっとも一般的な概念を最上位として，順により詳細な分類を示す概念からなる階層構造を成している．個々の情報はこの概念階層の中のいずれかに割り当てられて管理される．なお，分野や目的によって最も下位の概念にしか個小野情が割り当てられていない場合と，任意の概念に割り当て可能な場合があるが，本研究ではより，任意の概念に割り当て可能であるとする．

本論文では，分類に使われる概念階層は木構造であると仮定する．先に述べたクラスライブラリーやインターネットディレクトリ，目録分類など様々な分野において，木構造の概念階層は多く用いられている．本研究では，この様な概念階層を利用することを目的としているため，その基本構造である木構造を対象とする．なお．複数の上位ノードを持つものは単一の上位ノードを持つものを複数作る形で展開を行うことで木構造として扱う．木構造で示された概念階層を単純化し，グラフを用いて表すと図 1 の様に表せる．この図では，概念階層が木構造で表され，インスタンスが木のノードに割り当てられる．黒点がある概念を表し，白点がインスタンスを表す．本論文では，以降，情報の実体をインスタンスと呼ぶ．ここで提供される知識は，概念階層の構造によって異なり，それぞれが分類を行う知識を表していると考えられる．本論文では，情報を階層的に分類するものを知識と呼び，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．

階層的知識源は，その中に含まれるインスタンスの種類，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．従って，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困難が伴う．例えば，図 2 では，2 つの知識源が存在

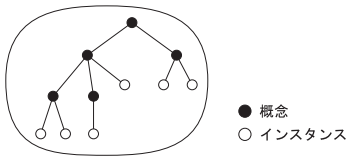

図 1: 階層的知識源のモデル

している．概念階層 H_1 にはインスタンスとして I_1, I_2, I_3 の 3 が存在しているが，概念階層 H_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．このとき，そのまま I_2 を H_2 に持ってきただけでは，H_2 上のどこに位置すべき情報なのかが H_2 にはわからない．従って，そのまま I_2 を利用することはできない．この問題を，*Hical* は知識源の間で共有しているインスタンス I_1, I_3 がどの様に概念階層に分類されているかということから，概念階層間の対応を学習することで解決している．この手法の問題点として，共有するインスタンスが存在しない場合には適用不可能だという点がある．そこで本論文では，インスタンスの内容を見ることでインスタンス間の類似関係を見つけ，インスタンスの共有を仮想的に作り出すことで，実際には共有しているインスタンスの無い知識源の間でも概念階層の対応を学習を可能にした *WebHical* を提案する．

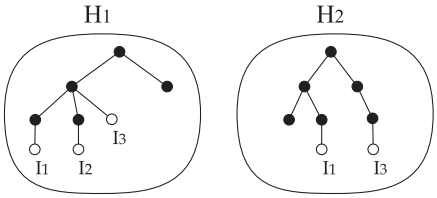

図 2: 複数の階層的知識源における問題点

3. 階層的知識源の統合手法

WebHical は階層的知識源の統合を行うシステムである．ここで，統合とは，ある知識源が持つ知識を異なる知識源に取り込むことを指す．そのためには両者の知識源の相違を吸収する必要がある．提案システムでは，その相違を吸収するために階層的知識と内容的類似性という二種類の情報を利用する．階層的知識とは階層的知識源においてあるインスタンスがどこに入っているかという情報を指し，内容的類似性とは階層的知識源が持つインスタンス間の似ている度合いを指すと，ここでは定義する．この章では，この二種類の情報それぞれを利用した場合の，階層的知識源の統合手法について説明する．

3.1 κ 統計量 (階層的知識の利用)

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ものである．ここで扱う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断できる．そのため，ある概念ノードを選択したときに，任意のインスタンスがその概念に適合するか否かを容易に判定することができる．すると，2 つの知識源における任意の 2 つの概念ノードに対して，共有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ことができるようになる．ここで，概念基準とは，ある概念に属するか否かの判断の基準のことを意味する．

Hical では，この概念基準の判定に κ 統計量 [Fleiss 73] を用いる．κ 統計量では，2 つの概念に対して表 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．N_1，N_2 は，それぞれ概念階層 H_1，H_2 中にある概念を表し，m_{**} はそれぞれの分類に含まれるインスタンスの数を示している．

表 1: インスタンス分割表

		概念 N_2	
		含まれる	含まれない
概念 N_1	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 と偶然に一致率 P' を次の式により計算する．

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

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

その時，κ 統計量は次式で表される．

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

この κ 統計量より，概念基準が一致しているかどうかを求める．概念基準が一致したということは，その概念間で知識の交換が可能であることを指す．このようにして階層的知識を用いた異なる階層的知識源の統合は行われる．

3.2 キーワード マッチング (内容的類似性の利用)

Hical の手法は，二つの概念階層が一致する URL を十分な数だけ持っていなくてはならない．そのため概念的に類似していてもカバーしているインスタンスの領域が全く別である場合には適用できない．そこで *WebHical* では，インスタンスの内容的類似性に基づいて概念階層間でインスタンスの追加を行い，仮想的に一致する URL を作ることで，この問題を解決する手法を用いる．

内容的類似性に基づくインスタンスの追加手法について説明する．まず概念階層 H_1 が持つインスタンス I_1 と類似したインスタンス I_{1s} を，概念階層 H_2 が持っているインスタンスの中から抽出する．なお，I_{1s} にはすでに H_1 が持っているものと同じインスタンスは含まれ得ない．これを H_1，H_2 が持つインスタンス全てに対して行う．類似したインスタンスの抽出は，インスタンスの内容を解析し，その中身を比較することで行う．提案システムでは階層的知識源としてインターネットディレクトリ，インスタンスとして Web ページを用いることを想定している．Web ページのテキストからキーワード抽出を行い，キーワードベクトル間の類似度を測ることで Web ページ間の類似判定を行う．実験で用いた類似度については次節で説明する．

次に，ある階層 C_1 が持つインスタンス $\{I_{C_1}\}$ と類似したインスタンス $\{I_{C_1S}\}$ を集める．この中からさらに選ばれたインスタンスが，階層 C_1 に追加される．これを概念階層 H_1，H_2 が持つ階層全てに対して行う．具体的な選択方法については次節で説明する．

以上の手続きにより，元々共有しているインスタンスが無かった概念階層間にも共有インスタンスができ，κ統計量による概念階層の統合が可能になる．

4. システム概要

本研究では，知識源の概念階層としてインターネットディレクトリの分類体系を対象とする，階層的知識と内容的類似性を用いた *WebHical* システムを試作した．本システムの概要を図 3 に示す．

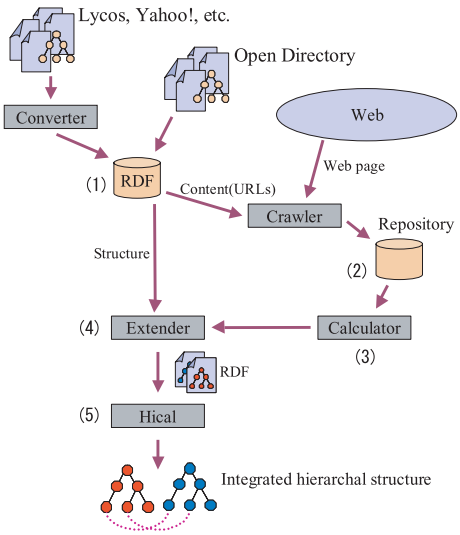

図 3: システム概要

システムの処理の流れを説明する．まずインターネットディレクトリの概念体系を取得する．Open Directory の場合は，データが 1 つにまとめられておりダウンロード可能なので，それをそのまま使う．Lycos や Yahoo!の場合は，クローラで Web サイトを巡回して Web ページを取得し，そこから概念体系を取り出す．これら手に入れた概念体系は RDF 形式で保存される．(1)

次に，RDF に記述された URL を元に，クローラにより Web ページを取得する．今回，クローラには Larbin[*1] を用いた．取得した Web ページから HTML タグを抜き取り，茶筅 [茶筅 99] により分かち書きをして名詞と未知語だけをキーワードとして抜き出し，リポジトリへ格納する．(2)

格納された Web ページ間の内容的類似性の計算には GETA[GETA 01] を用いる．GETA は汎用的な検索エンジンシステムであり，文書をキーとして類似する文書を検索する連想検索などが高速に行える．類似度には SMART[Singhal 96] の重み値を用いた．(3)

類似度により内容的類似性を計算できるようにした後で，3.2 節で説明したように概念階層 (インターネットディレクトリ) にインスタンス (URL) の追加を行う．まず，ある概念階層 H_1 の中の概念 C_1 が含む URLI_1 のキーワードベクトルをキーとして，概念階層 H_2 の中から URLI_{C_1S1} を取り出す．URLI_{C_1S1} を類似度で降順にソートし，その上位 10 個を追加候補 URLI_{C_1S2} とする．なお，URLI_{C_1S2} には H_1 が最初から持っている URL は入らないようにする．URLI_{C_1S2} の中からさらに閾値以上の類似度を持つ URLI_{C_1S3} を集め，これを追加 URL として概念 C_1 に追加する．(4)

最後に，追加処理を終えた概念階層 H_1 および H_2 を *Hical* により統合する．(5)

5. 実験

5.1 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の OpenDirectory 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) をインスタンスとして用いた．実験には，OpenDirectory から提供されている RDF データ (2002 年 9 月に生成) と，2000 年の 8 月から 9 月にかけて収集した LYCOS Japan の HTML ページを RDF に変換したものを用いた．ただし，インターネットディレクトリの全データは大き過ぎるので，その中の一部 (OpenDirecotry の文学ディレクトリと Lycos の Literature ディレクトリ) を用いた．

5.2 インターネットディレクトリの解析

Hical では，異なる階層構造の中のある階層間において共有している URL が多くある場合に，その概念は類似関係にあると判断した．今回新たに提案した手法は，異なる階層構造の中のある概念間において類似している URL が多くある場合に，その概念は類似関係にあると判断するものである．ここで問題になるのは，そもそも OpenDirectory や Lycos といった人手で作られた分類体系は，今回用いる内容的類似性に沿って作られた物であるのかどうかという点である．そこで予備実験として，同じ概念に入っている URL 間の内容的類似性を，今回用いた手法で見つけ出すことができるのかどうかを検証する．

予備実験は，次の手順で行った．まず階層構造を一つ用意し，それが持つ URL を GETA に検索対象として登録する．次に，3.2 節で解説した手法によって，階層構造に URL を追加する．この時，URL が元々入っていた階層の中に追加された場合に内容的類似性に基づく URL 追加が成功したとみなす．なお，キーと同じ URL は類似検索の結果から省くようにする．

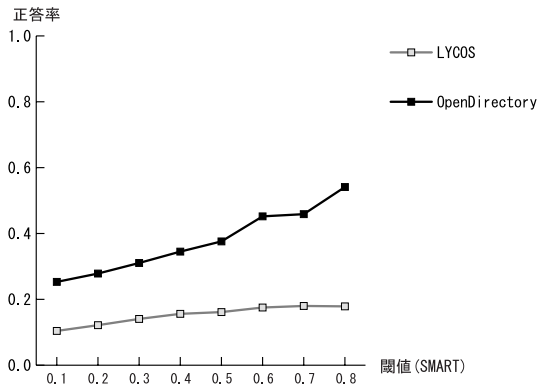

図 4: 内容的類似性による分類の正答率

図 4 は，OpenDirectory の文学カテゴリと Lycos の文学カテゴリとをデータとした，それぞれの場合の内容的類似性に基づく分類の正答率を示している．正答率とは，追加した URL が元々入っていた階層に追加された割合を指す．

*1 http://larbin.sourceforge.net/

*2 http://www.dmoz.org/

*3 http://www.lycos.co.jp/

閾値を 0.8 まで挙げた場合の正答率は OpenDirectory で 0.55，Lycos では 0.17 となっている (図 4)．これはつまり前者は半分，後者に至っては 8 割の追加 URL が誤ったデータであるということであり，補完データとして適切なものであるとは言えない．だが，OpenDirctory，Lycos 共に類似判定の閾値を高く設定することで追加 URL の精度が向上していることから，内容的類似性と階層的分類にはある程度の関係があることを示していると言える．

本研究での提案手法は，言語処理技術のみを用いたものではなく，人手によって作られた階層的知識と言語処理技術を併用したものである．次章では階層的知識と内容的類似性を併用した場合の実験結果について解説する．

5.3 インターネットディレクトリの統合

インターネットディレクトリの結合を Hical で行った．インターネットディレクトリには，事前に前章の手法でページを追加している．対象としたデータを表 2 に示す．内容的類似性を導入することで，246 個のリンク (URL) が新たに共有されている．

表 2: 実験に用いたデータ

	OD	Lycos
階層数	43	186
URL 数	454	1119
共有 URL 数	92	
追加共有 URL 数	246	

Hical を用いて発見した類似階層ペアの内訳を表 3 に，発見された類似階層ペアのうち，内容的類似性により類似ページを追加することで初めて発見されたものを表 4 に示す．

表 3: 発見された類似階層ペア数

	類似階層ペア数
WebHical	14
Hiacal	37
WebHical ∩ *Hical*	9

表 3 を見ると，*WebHical* によって見つけた類似階層ペア数は 14 であるのに対し，*Hical* で見つけた類似階層ペア数は 37 と 2 倍以上の差があることがわかる．内容的類似性の導入により本来見つけ出せなかった類似階層ペアを見つけ出せるようにすることが目的であったが，逆に発見できなくなった類似階層ペアを作り出してしまった．

この結果は，人手によって作られた階層的知識と言語処理により見つけ出せる内容的類似性の差が影響しているものと考えられる．予備実験により，今回用いた手法で導かれた内容的類似性による分類はあまり良い精度が出ないことがわかっている．これはつまり，内容的類似性に基づいて階層的知識源にインスタンスを追加するということは，多くの場合において，整然とインスタンスが配置された階層的知識源に対してでたらめな配置でインスタンスを追加するに等しい．これは結果として階層的知識源の作者が暗黙に持っていた分類ルールをあやふやにしてしまう．あやふやな分類の概念階層間で対応付けを行っても，見つけられる類似階層ペア数が少ないのは当然の事である．これが *WebHical* によって発見できた類似階層ペア数が *Hical* よりも少ない原因と考えられる．

そのような状況においてなお発見された類似階層ペア，その中でも特に *WebHical* で新たに見つけたペアはどのようなものかというと，これは内容的類似性に基づく URL 追加が，既にある階層的知識とバッティングしない形で行われた場合に，見つけ出されたものと考えられる．いわば，内容的類似性におけるノイズを階層的知識がフィルタリングした結果である．

表 4: 内容的類似性の導入で新たに発見された類似階層ペア

OpenDirectory/文学	Lycos/Literature
作家/小説家/横溝正史/	mistery/yokomizo/
俳句・和歌・川柳/	haiku/
小説/ファンタジー/	juvenile/

内容的類似性を導入した結果，新たに発見された類似階層ペアの中で「OpenDirectory/文学/作家/小説家/横溝正史」と「Lycos/Literature/mistery/yokomizo」に注目してみる．これらディレクトリを詳しく調べてみると，元々は同じ URL を一つも持っていない関係であることがわかった．このペアは内容的類似性に基づいて URL を追加したことで初めて発見できたものであり，これは典型的な階層的知識と内容的類似性を併用することの効果といえる．

6. まとめ

本研究では，階層的知識と内容的類似性を使った階層的知識源の結合手法を提案し，それを実装した *WebHical* システムを試作した．さらに階層的知識源としてインターネットディレクトリを用い，提案システムの実験を行った．結果，内容的類似性の導入が階層的知識のみを用いた場合には欠落してしまう部分を補完する効果が見られた一方で，内容的類似性による不正確な情報がかえって正しい発見を阻害してしまうという場面が見られた．

5.2 節にて内容的類似性を用いて人手により同一のカテゴリに分類されたページ間の関係を抽出できるか検証した結果，今回用いた手法では高い精度の抽出は行えないという結果となった．人手により作られた階層的知識を機械的処理により自動生成した内容的類似性によって模倣することは不可能であったが，後者により前者を補完するという点に関しては，5.3 節の実験結果より，その効果が確認できた．人手で作られた知識をより広範に活用するために，内容的類似性に基づく階層的知識の自動補完は有効な手段であったと言える．

今後は内容的類似性の精度向上を図ると共に，階層的知識と内容的類似性とのより有効な連携方法について考察を深め性能向上を目指す．

参考文献

[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.256-278, 2002.

[Takeda 01] 武田英明, 市瀬龍太郎, 村田剛志, 本位田真一: 知識共生プロジェクトーネットワーク情報の自律的生態系を目指してー, 情処研報, Vol. 2001, No. 41, pp. 25–33, 2001.

[茶筅 99] 松本裕次, 他:日本語形態素解析システム『茶筅』version2.2.0 使用説明書, NAIST Technical Report, NAIST-IS-TR99012, 1999.

[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, 1997.

[GETA 01] 汎用連想計算エンジン（第 2 版）導入・操作マニュアル, http://geta.ex.nii.ac.jp/getaN2001/gdoc/manual.html, 2001.

[Singhal 96] Singhal, A. and Buckley, C. and Mitra, M.: "Pivoted Document Length Normalization", Proc. of SIGIR, 1996, pp.21-29