F-035

二項関係に基づく階層関係抽出と構造的知識の自動構築手法
Extraction Hierarchical Relationships based on a Binary Relation and Method of Constructing Structured Knowledge

金盛 克俊† 大和田 勇人†
Katsutoshi Kanamori Hayato Ohwada

1. はじめに

人工知能研究において，与えられたデータから自動的に知識を構築する方法論の確立は困難な課題の一つである．

本研究の目的は，オブジェクトの集合とオブジェクト間の関係が定義されたデータが与えられたとき，それらを効率よく整理した構造的知識を構築することである．より具体的には，大量のオブジェクトや関係をそのまま並列的に蓄えるのではなく，似た性質をもつものをまとめてグループとし，オブジェクト間の関係をグループ間の関係に一般化して整理することである．

与えられた多量のデータをそのまま蓄えるのではなく，対象や関係を一般化して新たな知識として蓄積することは，知識獲得の本質ともいえる．一般化された知識は，データを効率よく保存するのに役立つだけでなく，未知の対象にたいしての予測にも役立つ．

本研究では，二項関係が定義されたオブジェクトの集合に対して，オブジェクト間の階層関係を定義し，さらにその階層関係の類似性からグループを生成し，グループ間の階層関係を求めることにより，新たな知識を構築する手法を提案する．

データや対象の類似性により，教師なしでグループ（クラスタ）にわける処理はクラスタリングとよく似ているが，本研究で目指す知識は通常のクラスタリングと異なり，グループ間でオブジェクトの重複を許す．また，グループは階層関係という定性的な性質をうまくまとめるために生成されるのであり，定量的な対象同士の近さをもとにまとめることが目的ではない．本稿でグループをあえてクラスタと呼ばないのはそのためである．

概念や概念間の関係を整理した知識表現として，オントロジーがよく知られている．オントロジーは主に自然言語処理技術によって構築される，概念同士の関係が記述された辞書で，さまざまな分野で具体的なオントロジーが提案・構築されている[1,2]．しかしながら，多くの実装は概念同士の関係が並列的に羅列さている状態で，その階層関係をグループとしてまとめているものは少ない．また，多くの実用的なオントロジーの構築は人手により行われており，そのための有用なツール[3,4]は提案されているものの，テキストデータからオントロジーを自動構築する試みについては十分な成果が得られているとはいいがたい．

また，自然言語文から単語概念の意味内容をベクトルで表現する手法[5,6]も提案されているが，概念同士の意味的関係を定量的尺度で比較することはできても，大量のデータをその定性的な性質から整理するためのものではない．

†東京理科大学理工学部経営工学科 Department of Industrial Administration, Faculty of Science and Technology, Tokyo University of Science.

本研究では対象を自然言語に限らず，二項関係が定義できる対象領域であればどのようなものに対しても適用可能な，二項関係により定まる対象の階層関係抽出手法と，それによる構造的知識を構築したい．

一般に二項関係は組の集合で表現されるが，これはグラフと同一視できる．グラフを対象としたデータマイニングはグラフマイニング[7]と呼ばれる．グラフマイニングは主にグラフに頻出する部分パターンの抽出を目的として行われており，本研究が目指すような，グラフをさらに体系的に整理するという発想のものは少ない．

本稿ではまず，二項関係をもとにした階層関係を定義し，それらが階層構造を成すことを示す．また，それをもとにしたグループの定義，さらにグループ同士の階層関係を定義する．

次に，これらの階層関係とグループ，グループ間の階層関係を求める手法をそれぞれ提案する．また，その有用性を示すために，ユーザとアイテムの評価値行列を用いた応用例を示す．

2. 問題の定式化

本研究では，二項関係が定義された対象領域のオブジェクトに対して，オブジェクト間の階層関係を明らかにし，その相互関係をもとにしてオブジェクトをグループ化し，さらにグループ間の階層関係を求めることが目的である．

本章では，まず求めるべきオブジェクト間の関係を定義し，それが階層関係になっていることを示す．さらに，その階層関係を用いて，他のオブジェクトに対して同じ階層関係を持つオブジェクトをまとめたグループと，そのグループ間の階層関係を定義する．このグループ間の階層関係こそが，本研究で求めるべき解である．

2.1 二項関係に基づく階層関係

本研究で求めるオブジェクト同士の階層関係は，以下のように定義されるものである．

2.1.1 階層関係の定義

オブジェクトの集合OとTに対して二項関係$R \subset O \times T$が定義されていて，オブジェクト$a, b \in O$に対して

$$\forall x \in T : aRx \rightarrow bRx$$

かつ

$$\exists x \in T : bRx \text{ かつ } (aRx \text{ でない})$$

が成り立つとき，$a \succ b$と書き，a は b の上位であるという．

2.1.2 階層関係の例

自然数を用いた簡単な例を示す．

$O = T = \{2,3,4,5,6,7,8,9,10\}$とし，二項関係 aRb を”a は b で割り切れる”とすると，4R2 や 6R2 などが成り立つ．このとき，4 を割り切るものは全て 8 も割り切るので，

$$\forall x \in T : 4Rx \rightarrow 8Rx$$

かつ

$$\exists x \in T : 8Rx \text{ かつ } (4Rx \text{ でない})$$

が成り立つので，$4 \succ 8$である．

この関係を図にすると，図１のようになる．

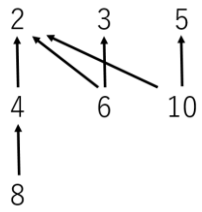

図１　”割り切る”関係による階層関係

2.1.3 階層関係の性質

このように定義された階層関係が，実際に階層構造を構成することを示す．この関係が可換でなく，推移的であることは定義より明らかである．すなわち次の定理が成り立つ．

定理１（非可換）：任意の a.b に対して，$a \succ b$であれば$b \succ a$でない．

定理２（推移性）：任意の a.b.c に対して，$a \succ b$かつ$b \succ c$であれば$a \succ c$である．

また，次の定理も成り立つ．

定理３（閉路）：任意の集合O，Tと任意の二項関係$R \subset O \times T$に対して，階層関係$\succ$に閉路は存在しない．

証明：閉路が存在すると仮定すると，推移性より，あるオブジェクト$o \in O$が存在して，o よりも上位かつ下位であるようなo'が存在することになる．しかし，これは$\succ$が非可換であることに矛盾する．よって閉路は存在しない．QED.

2.2 グループ

前節で定義した階層関係はオブジェクト同士の相互関係を表している．本研究ではこの相互関係を用いて，同じ性質を持つものはグループとしてまとめ，さらにグループ間の階層関係を求め，データ全体をグループで構成される構造的な知識体系として整理したい．ここではまずグループの定義と，グループ間の階層関係について定義し，簡単な例を示す．

2.2.1 グループの定義

次の２つの性質を満たすオブジェクトの集合 G をグループという．

$$\forall a, b \in G, x \in O : (a \succ x \rightarrow b \succ x) \text{かつ} (b \succ x \rightarrow a \succ x)$$
$$\forall a, b \in G, x \in O : (x \succ a \rightarrow x \succ b) \text{かつ} (x \succ b \rightarrow x \succ a)$$

2.2.2 グループ間の階層関係の定義

次を満たすとき，G_1はG_2の上位であるといい，$G_1 \succ G_2$と書く．

$$\forall a \in G_1, \forall x \in T : (\text{aRx ならば } \forall b \in G_2 : bRx)$$

$G_1 \succ G_2$であるとき，上位グループG_1に含まれるオブジェクトと関係のあるものは，必ず下位グループG_2に含まれる全てのオブジェクトと関係がある．

2.2.3 グループの例

自然数を用いた簡単な例を示す．

$O = T = \{2,3,4,5,6,7,8,9,10\}$とし，二項関係 aRb を”a と b が 1 以外の共通の約数を持つ”とすると，図２に示すような階層関係が定義される．

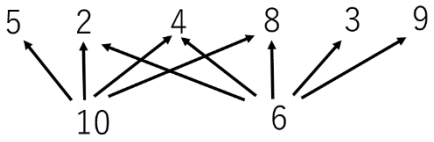

図２　”共通の約数を持つ”関係による階層関係

この関係より，グループとその階層関係を求めると，図３のようになる．

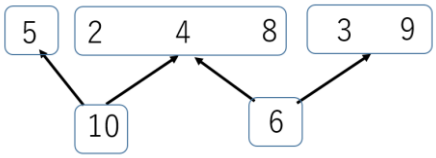

図３　”共通の約数を持つ”関係によるグループ

図３では共通のグループに含まれるオブジェクトが囲まれており，図２とは異なり，オブジェクト同士ではなくグループ間の階層関係が記述されている．同じ階層関係を持つオブジェクト同士がグループにまとまり，全体として，オブジェクト同士が成す構造的な関係が体系的に整理されているのがわかる．

3. 提案手法

実際のデータを用いて前章で定義した体系的知識を定義通りに求めるのは，現実的には困難である．なぜなら，オブジェクト間の階層関係やグループ内のオブジェクト間の階層関係の定義には一つの例外も許さず，実際のデータに含まれるノイズや不完全性に対して対応できないためである．そこで，本研究では，ある程度の誤差を許した階層関係とグループを求める手法を示す．

3.1.1 階層関係を求める手法

2.1.1 節の定義では，全てのターゲットに対してaRx → bRxが成り立つことが必要であったが，ここでは閾値th_hを定め，次の２つの式を満たすオブジェクト a,b についてa ≻ bであるとする．

$$\left|\{x|aRx \text{かつ} bRx\}\right| / |\{x|aRx\}| \geq th_h$$
$$\left|\{x|aRx \text{かつ} bRx\}\right| / |\{x|bRx\}| < th_h$$

この閾値を 1 とすると，一つの例外も許さない階層関係が求められることになる．一般にこのしきい値が高いほど階層関係は生成されにくく，低いほど生成されやすい．

3.1.2 グループとグループの階層関係を求める手法

グループを求める手法を説明する前に，オブジェクト同士の類似度と，グループの類似度を定義する．

オブジェクト a,b の類似度sim(a, b)は次の式で定義される．

$$\begin{aligned}&\text{sim}(a, b)\\ &= (|\{x|x \succ a \text{かつ} x \succ b\}|/|\{x|x \succ a \text{または} x \succ b\}|\\ &+ |\{x|a \succ x \text{かつ} b \succ x\}|/|\{x|a \succ x \text{または} b \succ x\}|)/2\end{aligned}$$

グループ G,H の類似度 sim_g(G,H)は次の式で定義される．

$$\text{sim\_g}(G, H) = \left|\{x|x \in G \text{かつ} x \in H\}\right| / \left|\{x|x \in G \text{または} x \in H\}\right|$$

オブジェクト a,b が他のオブジェクトに対して全く同じ階層関係を持っていれば，sim(a,b)の値は 1 となり，グループ G,H に含まれるオブジェクトが全く同じであれば sim_g(G,H)の値は 1 となる．

グループとその階層関係を求める手法を以下に示す．
手順１：グループの集合 LG を空集合とする．
手順２：オブジェクトを順番に一つずつ選んで o とし，すべてのオブジェクトに対して手順３～手順５を繰り返す．
手順３：{o}を新たなグループ G とする．
手順４：o 以外の任意のオブジェクト x に対して，

$$\mathrm{sim}(\mathrm{o},\mathrm{x}) > th_g$$

であれば，G ∪ {x}を新たに G とする．これをすべてのオブジェクトに対して行う．th_gは閾値である．
手順５：LG に含まれるグループのうち，G との類似度がしきい値th_f以上となるものがあれば，そのうちで最も類似度の高くなるグループG_0に対して，$\mathrm{G}_0 \cup \mathrm{G}$を新たに$\mathrm{G}_0$とする．類似度がしきい値を超えるグループが LG に含まれていなければ，LG ∪ {G}を新たに LG とする．
手順６：LG に含まれるすべてのグループの組み合わせ(G, H)に対して，次式を満たすとき G ≻ H とする．

$$\sum_{a\in G} Av_{rate}(\mathrm{a},\mathrm{H})/|G| > th_{gh}$$

ただし，$Av_{rate}(\mathrm{a},\mathrm{H})$はa ∈ G と H に対して、aRx であるような x ∈T における，$R_{rate}(\mathrm{H},\mathrm{x})$の平均であり，次式で与えられる．

$$Av_{rate}(\mathrm{a},\mathrm{H}) = \sum_{x\in\{x|aRx\}} R_{rate}(\mathrm{H},\mathrm{x}) \,/\, |\{x|aRx\}|$$

また，$R_{rate}(\mathrm{H},\mathrm{x})$は，H の任意の要素 b に対して bRx であるものの割合を表し，次式で定義される．

$$R_{rate}(\mathrm{H},\mathrm{x}) = |\{\mathrm{b}|\mathrm{bRx},\mathrm{b}\in\mathrm{H}\}|/|\mathrm{H}|$$

以上の手順を用いると，定めた閾値によってある程度誤差や例外を許したグループとグループ間の階層構造が求められる．閾値th_gが低いとオブジェクトがグループにまとまりやすく，閾値th_fが低いとグループ同士が統合されやすく，閾値th_{gh}がグループ間の階層関係が生成されやすい．
次章では，具体的な応用例を示す．

4. ユーザ・アイテム行列への応用実験

複数のユーザによるアイテムへの評価値を二次元表にまとめた評価値の行列は，協調フィルタリングをはじめとした推薦システムの対象データとしてよく用いられる．これらのシステムの多くは，一人一人のユーザに対して推薦すべきアイテムを抽出するもので，評価行列を用いてユーザ同士の関係を体系的に整理するようなものではない．そこで，ここではユーザ・アイテム行列の異なる応用例として，提案手法を適用してユーザやアイテムを体系的にグループ化する方法について述べる．

4.1 映画の評価値を用いたユーザの体系的整理

ユーザによる映画に対する 5 段階評価値を持つ Movielens[4]のデータを用いる．このデータは 138493 人のユーザによる 27278 の映画に対する評価値を持つ．ただし，各ユーザは全ての映画に評価値をつけているわけではなく，多くの未評価を含むスパースな行列である．
ここでは，ユーザとそのユーザが最高評価をつけている映画に対して二項関係を定め，この関係をもとにユーザ同士の階層関係を抽出することを考える．すなわち，ユーザ a,b の階層関係a ≻ bは，ユーザ a が高評価をつけている映画はユーザ b も必ず高評価をつけているときに成り立つことになる．このような二項関係と階層関係に基づくグループとその階層関係を実験により求める．
Movielens に含まれる全てのデータのうち，100 人のユーザを抽出し，これらのユーザとこれらのユーザが高評価をつけている全ての映画を対象に実験を行う．
それぞれのしきい値は $th_h = 0.9, th_g = 0.9, th_f = 0.9, th_{gh} = 0.6$とした．
このとき，25 のグループが生成され，いくつかのグループ同士の階層関係が形成された．
表 1 に 25 のグループの構成をいくつか示す．

表１　映画高評価関係によるユーザグループ

ID	含まれるユーザ（ID）
1	1, 4, 6, 9, 14, 17, 18, 25, 27, 28, 30, 31, 33, 35, 37, 38, 39, 40, 41, 42, 43, 44, 47, 49, 55, 60, 61, 63, 66, 67, 68, 69, 70, 76, 77, 78, 79, 80, 81, 83, 84, 85, 90, 91, 95, 96, 97, 98, 99, 100
2	2, 34, 7, 57, 93
4	65, 5, 71, 75, 82, 21, 22, 24, 89, 32, 48, 51, 62
7	64, 11, 29
12	72, 74, 46, 19, 53
17	50, 92

グループ 1 に最も多くのユーザがまとめられた．これらのユーザは，高評価をつけた映画が極端に少なく，他のグループとの階層関係は形成されなかった．ユーザの高評価の相互関係から，データの少ないユーザを自動的に一つのグループにまとめて抽出できたことがわかる．
表 1 に示さなかったグループの多くは，ユーザ一人だけが含まれるグループであった．このようなグループは，複数のユーザが含まれる他のグループの上位となっていたり，下位となっていたりするものが多かった．複数のユーザの上位にあるユーザがいるということは，全体の構造として影響力の強いユーザが抽出できたと考えられる．
図 5,6 に生成されたグループとグループ間の階層関係を示す．四角い枠がグループを表し，中の数字はユーザ ID を表す．

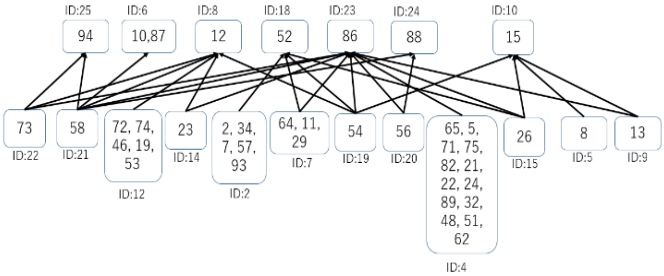

図５　”高評価”関係によるユーザグループ階層関係１

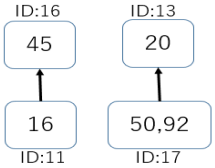

図６　”高評価”関係によるユーザグループ階層関係２

それぞれのグループ間において，視覚的にもわかりやすい階層関係が構成されていることがわかる．これまで，一人一人のユーザに対しての推薦やアプローチしかできなかったアイテムユーザ行列に対して，グループ単位でその性質を抽出し，対応できる可能性を示した．

4.2 映画の評価値を用いた映画の体系的整理

4.1 節ではユーザによる映画の高評価関係を用いてユーザをオブジェクトとしてユーザグループとその階層関係を生成したが，ここではユーザと映画の関係を逆転させて，映画をオブジェクトとして扱い，映画のグループとその階層関係を生成することを考える．すなわち，オブジェクト（映画）同士の階層関係は，映画 A に高評価を与えるユーザは必ず映画 B にも高評価を与えるという条件が満たされるとき，$A \succ B$である．

前節と同様に，100 人のユーザに対する映画評価データを用いたところ，27278 の映画が 571 のグループにまとめられ，例えば図 7 に示すような映画グループ同士の関係が得られた．それぞれのしきい値は$th_h = 0.9, th_g = 0.8, th_f = 0.8, th_{gh} = 0.6$とした．

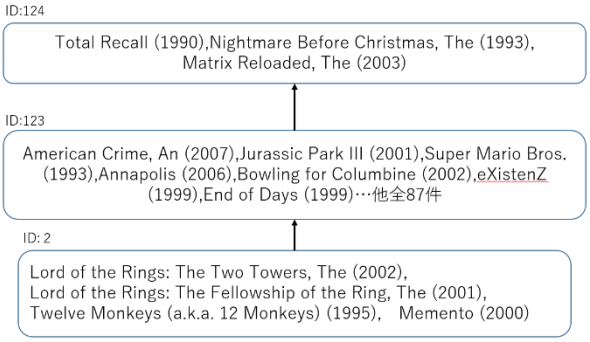

図７　”高評価”関係による映画グループ階層関係

前節とは異なり，映画グループ同士の階層関係が獲得できているのがわかる．図 7 からは，ID124 番のグループに属す映画を好むユーザは，ID123 番のグループに属す映画も好む傾向にあるという関係が読み取れる．さらに，ID123 の映画を好むユーザは ID2 に属す映画も好む傾向にある．

例えばこのようなグラフを用いることにより，あるグループに属す映画に高評価をつけたユーザに対して，その下位グループに属す映画を推薦する手法が考えられる．また，そのような個人に対するアプローチだけでなく，多くの商品の全体の傾向を求めることにより，商品をどのように配置すべきかという問題に対するヒントと成りうる．

5. おわりに

二項関係が定義された対象空間において，オブジェクト間の階層構造を定義し，その階層関係からオブジェクトのグループを定義し，グループ間の階層関係を定義した．さらにそれを求めるアルゴリズムを提案した．

このアイデアと手法の有効性を確かめるため，映画の評価データを用いた応用例を示すことにより，ユーザ-アイテム行列に提案手法が有効であることを示した．

本稿で紹介した応用例はひとつにとどまったが，提案した手法は一般的な手法であり，対象領域が何であれ，何らかの二項関係が定義されていれば適用が可能である．当然のことながら，同じ対象領域でも用いる二項関係がかわると結果が大きくかわる．

一方で，閾値パラメータを４つも設定しなければならないという問題点もある．それぞれのしきい値による影響は大きく，適切な値を設定しないとグループが多く生成されすぎたり，グループ間の階層関係がほとんど生成されないことがある．それぞれのしきい値同士が相互に関係しあうこともあり，適切なしきい値を定めるのは容易ではない．例えばオブジェクトの類似度のしきい値を下げるとオブジェクトがグループにまとまりやすくなるが，その分グループ間の階層関係はできにくくなり，これを解消するにはグループの階層関係のしきい値を下げなければならない．

今後はオントロジーやシソーラスの自動構築や，論文のリファー関係，WEB サイトのリンク関係など，さまざまな応用研究を進め，本手法の有効性を確かめる必要がある．

本研究の目指すところは，実用的なデータマイニング技術としての手法よりも，与えられたデータをもとに知識を自動的に構築するという，人工知能研究が本来解決すべき根本的な方法論の解決である．本研究で提案した，オブジェクトのグループ化と関係の一般化は，知識構築に関わる新たな問題提起でもある．本手法を用いて，実用的なシステムが構築できることを示すだけでなく，人間の知能活動を再現するような成果が望まれる．近年滞りがちな，人工知能研究が抱える基礎的な問題を解決する糸口になることを願う．

参考文献

[1] 桝屋 啓志, 溝口 理一郎, “遺伝学オントロジー”, 人工知能学会論文誌, Vol. 29 (2014) No. 3 p. 311-327 (2014).
[2] Michael Ashburner et al., “Gene Ontology: tool for the unification of biology”, Nature Genetics 25, 25 - 29 (2000).
[3] 砂川 英一, 古崎 晃司, 來村 徳信, 溝口 理一郎: “コンテキスト依存性に基づくロール概念組織化の枠組み”, 人工知能学会論文誌, Vol. 20, No. 6, pp.461-472 (2005)
[4] Naoki Sugiura, Yoshihiro Shigeta, Naoki Fukuta, Noriaki Izumi, and Takahira Yamaguchi: Towards On-the-fly Ontology Construction - Focusing on Ontology Quality Improvement, 1st European Semantic Web Symposium(2004)
[5] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean: Efficient Estimation of Word Representations in Vector Space, arXiv:1301.3781,ICLR(2013)
[6] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, Jeff Dean: Distributed Representations of Words and Phrases and their Compositionality, arXiv:1310.4546(2013)
[7] Deepayan Chakrabarti, Christos Faloutsos,Graph mining: Laws, generators, and algorithms, ACM Computing Surveys (CSUR), Volume 38 Issue 1, 2006, Article No. 2(2006)
[8] F. Maxwell Harper and Joseph A. Konstan, “The MovieLens Datasets: History and Context”, ACM Transactions on Interactive Intelligent Systems (TiiS), Vol.5, No.4, Article 19 (2015).