Vol.2010-MBL-56 No.3
Vol.2010-ITS-43 No.3
2010/11/11

クラスタリングによる 実世界行動モデルの階層構造学習

深澤 佑介[†,††]　　太田 順[††]

本稿では，ユーザの実世界行動モデルの構築を軽減するため，クラスタリング手法の適用可能性を検証する．検証実験の結果，マニュアルで構築したモデルと50%程度重複するモデルを構築することができた．現状，実世界行動モデルをマニュアルで構築するには１ヶ月程度かかることから，クラスタリングとマニュアルを組み合わせることで構築に要する時間を劇的に削減できると考えられる．

Learning Hierarchical Model of User's Real World Activity by Clustering

Yusuke Fukazawa[†,††]　and Jun Ota[††]

To construct user's real world activity model in an automatic way, in this paper, the possibility to use the clustering method is verified. As the experimental result, the taxonomic overlap between model constructed with clustering and manually constructed reference model becomes almost 50%. Considering that we need one month to construct the real world activity model buy manual, this result indicates that the time required constructing the model will be reduced dramatically by using the clustering with manual way of constructing the model.

1. はじめに

近年，携帯端末で閲覧可能なコンテンツやメディアは劇的に増加している．乗り換え案内，チケット購入，天気予報，及び災害情報等，実世界でユーザが直面する様々な問題や要望をリアルタイムで解決可能な問題解決型のコンテンツも充実してきている．このようなコンテンツを探しやすくするため，ユーザの問題解決行動をモデル化したタスクモデルを利用する方法が提案されている[1][4][5][6]．Naganuma らはユーザが実世界で抱える問題をインタビュー形式でヒアリングしそれをブレークダウンすることによってタスク知識ベースを構築している．タスク知識ベースとモバイル向けのサービスを関連付けることにより，サービスの検索を効率化している [1][4]．Masuoka らは，ユーザの周りで利用可能なサービスを，それらのサービスで達成可能なタスクと一緒に表示することにより，ユーザがタスク達成に必要なコストを削減する環境 Task Computing Environment(TCE)を提案している[5]．Bernstein らが提案する MIT Business Process Handbook では，サービスをタスクの観点で索引化する方法を提案している．Naganuma らの研究は，トップダウンでユーザの行動をブレークダウンするのに対し，Masuoka，Bernstein らは，実行可能なサービスからボトムアップで必要なタスクを洗い出している点が異なる[6]．本研究では，Naganuma らと同様に実世界行動モデルの構築に着目していることから，Naganuma らの研究について詳細を述べる．

Naganuma らの提案する実世界行動モデルとは，ユーザが実世界で抱える問題の構造を分析し，ユーザの問題の解決行動（タスク）を構造化したものである[1][4]．実世界行動モデルでは，実世界の場所（ドメイン）ごとに，その場所に関連して発生するタスクがツリー状に構造化されている．映画館ドメインにおける実世界行動モデルの一部を図１に示す．タスク間の関係は達成関係 *is-achieved-by* で表現されている．上位のタスクではユーザの抽象的な要求や問題を表現しており，下位のタスクではより具体的な解決策を表現している．末端のタスクには該当するタスクを解決可能なモバイル上のサービス（コンテンツの Uniform Resource Identifier:URI）が関連付けられている．

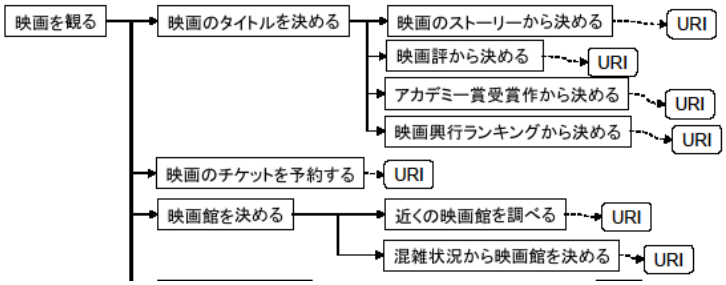

Figure 1 View of a part of task-model

文献[1]では，ユーザが実世界で直面する問題をインタビューなどから集め，それをマニュアルで階層化することで実世界行動モデルを構築している．しかしながら，未知の新しいドメインにおいて，実世界行動モデルを構築するためには専門家による長時間の構築・精査が必要である．そこで，本稿では，実世界行動の集合から実世界行動

† NTT ドコモ
NTT DOCOMO, Inc.
†† 東京大学
The University of Tokyo

ⓒ2010 Information Processing Society of Japan

Vol.2010-MBL-56 No.3
Vol.2010-ITS-43 No.3
2010/11/11

間の階層構造を自動で構築する方法を提案する．従来，名詞概念については，クラスタリング手法を用いて自動的に階層化する方法が提案されている．一方，実世界行動間の階層構造の自動獲得に関する研究は行われていない．そこで本稿では従来のクラスタリング手法を用いて実世界行動間の階層化が実現可能であるか検証する．

2 章では，クラスタリングによる実世界行動の階層構造を獲得するための手法について述べる．3 章では，dvd ドメインにおいて実際にクラスタリングを行い，マニュアルで構築した正解モデルと比較しその精度を評価する．4 章で結論を述べる．

2. 実世界行動間の階層構造の構築

2.1 手法概略

クラスタリングを利用して実世界行動を階層化する手法について述べる[2]．一般的に，名詞概念，実世界行動を問わず，概念の集合のクラスタリングを行う場合には，以下の手順が必要となる．

Step1）クラスタリングの対象となる概念の集合を用意する．
Step2）用意したクラスタリングの対象となる概念を特徴表現する．
Step3）特徴表現を基にクラスタリングアルゴリズムを用いて概念間の階層構造を構築する．

ここで，名詞概念を対象にしたクラスタリングの場合と異なり，実世界行動を対象にしたクラスタリングを行う場合には，Step1 および Step2 に工夫が必要となる．具体的には，Step1 において，名詞概念の場合には既に名詞概念の集合である辞書や Web 上に蓄積されたデータセットを利用することができるが，実世界行動としてまとまったものは公開されていない．また，Step2 において，名詞概念の場合には，名詞が登録されている辞書の説明文を対象に特徴表現したり，文書中の名詞の周囲の語を特徴量として利用したりすることが可能である．一方，実世界行動の場合には，実世界行動向けの辞書が存在しないこと，名詞概念と比較し文書中に実世界行動が現れる確率が非常に低いことから，別の特徴表現の方法を検討する必要がある．なお，Step3 に関しては，本稿では名詞概念で効果があるとされる 3 つのアルゴリズムを利用する[2]．以下 Step ごとに詳細な方法を説明する．

2.2 実世界行動の収集方法

Step1 について述べる．M.Strohmaier らは，AOL の検索エンジンの検索クエリのデータから行動目的に関連するクエリのみを抽出している[3]．行動目的に関連するクエリとは，名詞単体ではなく，自然文で入力されたクエリ（e.g. 映画を見に行きたい）を意味する．本研究では，彼らからデータを借り当該データを利用したクラスタリングを行う．ただし，クエリ数が非常に多いため，ここでは，クエリ内に DVD および Movie という単語が入っているクエリのみを対象とする．実際に利用したクエリを Table1 に示す．

Table 1 List of human-goal queries

how to get original tv movies roll of thunder hear my cry on dvd
how to find dvd movies
how can I find out what movies are released on dvd on tuesdays
movies on dvd to buy
how to copy dvd movies
how to burn movie dvd from computer file
how to share dvd movies
buy dvd movies
buy dvd vhr movies
how to burn dvd movies
dvd movies to buy at discounts
how to make highlight movies from dvd
blank dvds movies buy complain
make dvd movie
how to download dvd movies
how to get original tv movies on dvd
blank dvds movies buy
buy used dvds movies in houston tx
how to record home movies to dvd

2.3 特徴抽出方法

Step2 の特徴抽出の方法について述べる．実世界行動の特徴表現方法には，Table2 に示すとおり 2 つの方法がありうる．第一に，Step1 で抽出された実世界行動をクエリとして検索エンジンに入力，検索結果を得，その検索結果に含まれる重要語を抽出することにより当該行動の特徴を表現する．第二に，行動の表現自体から，名詞または動詞を抽出しそれを用いて特徴表現する．ここでは，簡単のため，第二の方法を用いる．

Table 2 Features extraction methods

	Features	
Using search results	1)Search human-goal query using Google 2)Pick up important keyword from search results and make them features.	Noun only
		Verb only
		Noun plus Verb
Using query itself	1) Use the word that consists of human-goal query as features	Noun only
		Verb only
		Noun plus Verb(The feature we used in this test)

ⓒ2010 Information Processing Society of Japan

Vol.2010-MBL-56 No.3
Vol.2010-ITS-43 No.3
2010/11/11

2.4 クラスタリングアルゴリズム

Step3 で利用するクラスタリングアルゴリズムについて述べる．クラスタリングアルゴリズムの候補を Table3 に示す．これらのアルゴリズムは文献[2]において，名詞概念間の階層構造を構築するために利用されており，本稿でもこの 3 つの手法を実装し比較する．

Table 3 Clustering method

	Worst case Complexity	Traceability	Size
FCA(Formal Concept Analysis)	$O(2^n)$	Good	Large
Hierarchical Clustering	$O(n^2)$	Fair	Small
Bi-Section K-means	$O(n^2)$	Weak	Small

2.4.1 FCA: Formal Concept Analysis（形式概念分析）による階層構造の獲得

FCA とは，概念間の関係性を束論（Lattice）で分析する方法である．また，形式概念とは，共通の属性（Attribute）を持つ概念（Object）の集合で表現される抽象度の高い概念を表す．形式概念分析では，多くの概念を有する形式概念を上位概念とし，少ない概念を有する形式概念を下位概念とすることで，形式概念を階層構造化することができる．さらに，形式概念の階層構造を概念の木構造に変換することにより，概念間の階層構造を獲得することができる．以下，具体的な手順について述べる．

まず，概念と概念の属性を列挙する．具体例を Figure2 に示す．ここでは概念として旅行のドメインに現れる 6 つの概念を対象としている．また，概念の属性としては，表に記載の 6 つの属性を対象とする．Figure2 では，概念ごとに有する属性をまとめている．

		Attributes				
		bookable	rentable	drivable	rideable	Joinable
objects	Hotel	*				
	Apartment	*	*			
	Car	*	*	*		
	Bike	*	*	*	*	
	Excursion	*				*
	trip	*				*

Figure 2 Example of tourism-related object-attribute pair

次に，形式概念を導出する．前述のとおり，形式概念は，共通の属性（Attribute）を持つ概念（Object）の集合で表現される．形式概念は形式的に（O,A）で表され，O,A はそれぞれ以下の条件を満足する必要がある．

・O:概念の部分集合，A:属性の部分集合．
・O はすべての A を持つ．
・O は A 以外に共通して持つ属性がない．

例えば，Figure3 に示すように，概念 Car, Bike は，共通の属性（bookable, rentable, driveable）を持つため，一つの形式概念をなす．Figure2 の表からは以下の 5 つの形式概念を導くことができる．

$$(\{hotel, apartment, car, bike, excursion, trip\}, \{bookable\})$$
$$(\{aprtment, car, bike, excursion, trip\}, \{bookable, rentable\})$$
$$(\{car, bike\}, \{bookable, rentable, dribable\})$$
$$(\{bike\}, \{bookable, rentable, dribable, rideable\})$$
$$(\{excursion, trip\}, \{bookable, joinable\})$$

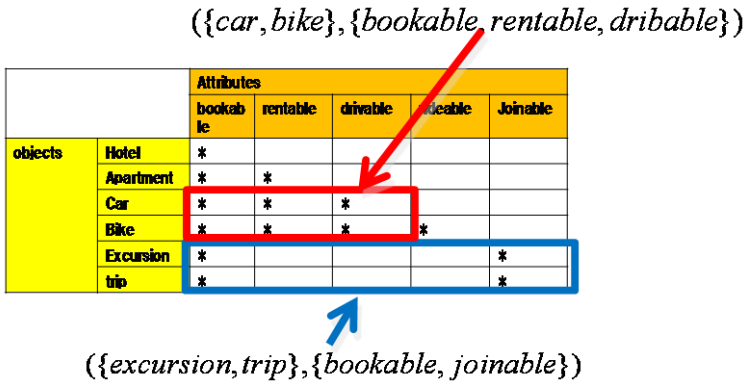

Figure 3 View of a part of task-model

次に，獲得した形式概念間の関係を Lattice 構造で表現する．ここでは，多くの Object を有する形式概念を上位概念とし，少ない Object を有する形式概念を下位概念とする．これは，概念の階層化においては一般に，上位ノードほど包含する概念が多く，下位ノードほど包含する概念が特殊化されていくことに由来する．Figure4 に具体例を示す．

次に，獲得した Lattice 構造を簡略表現する．具体的には，形式概念（O,A）について，O の中から，下位のノードに出現する概念については上位のノードから削除し，A については，上位のノードに出現する属性については下位のノードから削除する．これは，概念は各最上限を，属性は各最下限をとることと同意である．Figure5 に具体

ⓒ2010 Information Processing Society of Japan

Vol.2010-MBL-56 No.3
Vol.2010-ITS-43 No.3
2010/11/11

例を掲載する．

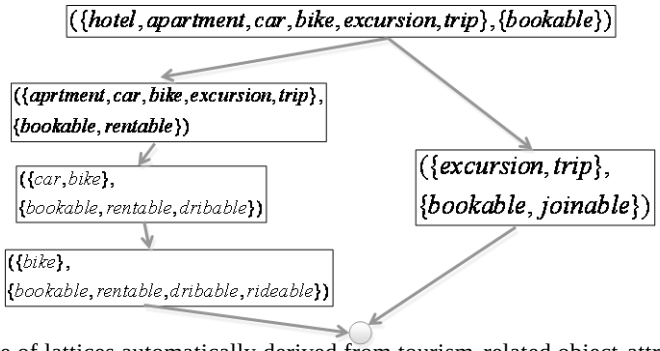

Figure 4 Example of lattices automatically derived from tourism-related object-attribute pair

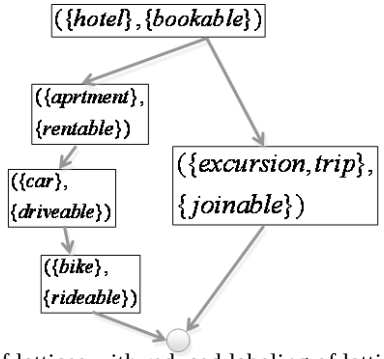

Figure 5 Example of lattices with reduced labeling of lattice shown on 図 4

最後に，Lattice 構造を分解し，概念の階層構造を獲得する．具体的には，形式概念（O,A）について A を上位概念，O を下位概念として分解することにより，概念の階層構造を獲得することができる．具体例を Figure6 に掲載する．

2.4.2 階層型クラスタリングによる階層構造の獲得

階層型クラスタリングでは，概念単位の小さなクラスタを仮定しておき，近いクラスタをまとめる処理を繰り返すことで概念間の階層構造（デンドログラム）を獲得する．具体的なアルゴリズムは以下のとおりである．1)まず初めに，概念の数分のクラスタを用意する．クラスタの代表ベクトルは概念に対する属性をベクトルの要素とする．具体的には属性を持つ場合は値 1，持たない場合は 0 として表現される．属性の要素によっては連続値で表現される場合もありうる．次に，クラスタ間の距離 d(Cx, Cy) を計算し，最もこの距離の近い二つのクラスタを逐次的に併合する．この処理を全ての対象が一つのクラスタに併合されるまで繰り返す. Figure2 に掲載した概念と属性の関係に基づき階層型クラスタリングを実行した場合，最終的に得られる概念の木構造は Figure7 に示すとおりである．

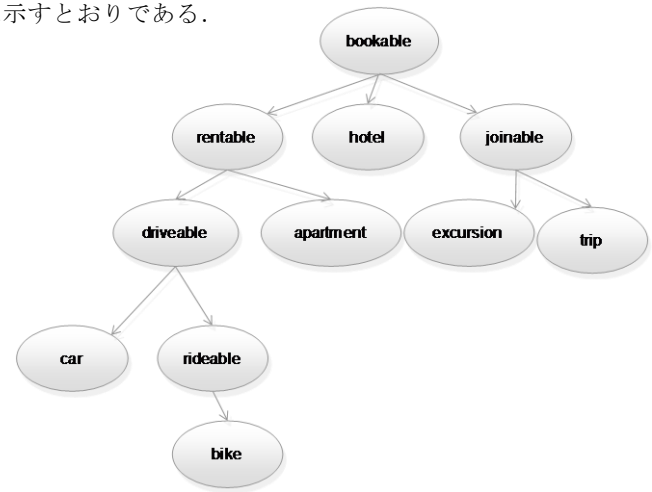

Figure 6 Example of acquired concept hierarchy derived from reduced labeling (Figure5)

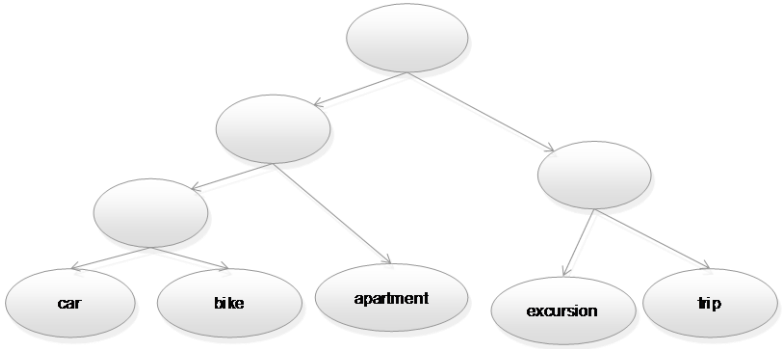

Figure 7 Example for an automatically acquired concept hierarchy derived from Hierarchical agglomerative clustering

2.4.3 KMeans による階層構造の獲得

KMeans は，階層型クラスタリングとは異なり，複数の概念が含まれるクラスタを仮定し，互いの距離が最適になるようクラスタ分割を繰り返すことで階層構造を得る方法である．クラスタ分割の評価関数を定め，その評価関数を最適にする分割を探索

ⓒ2010 Information Processing Society of Japan

Vol.2010-MBL-56 No.3
Vol.2010-ITS-43 No.3
2010/11/11

する方法である．具体的なアルゴリズムは以下のとおりである．1)まず初めに k 個の概念をランダムに選択し，その概念を初期クラスタとして選択する．初期クラスタの代表ベクトルは概念の重心である．2)次に，概念（Xi）と初期クラスタ（Ci）間の距離を計算し，概念を C=arg min d(Ci,Xi) なるクラスタ C に割り当てる．この処理を全ての概念について実行する．3)全ての概念に割り当てられたクラスタが変化しないならば終了し，そうでなければ各クラスタの重心をクラスタの代表ベクトルにしてステップ 2)へ戻る. Figure2 に掲載した概念と属性の関係に基づき階層型クラスタリングを実行した場合，最終的に得られる概念の木構造は Figure8 に示すとおりである．

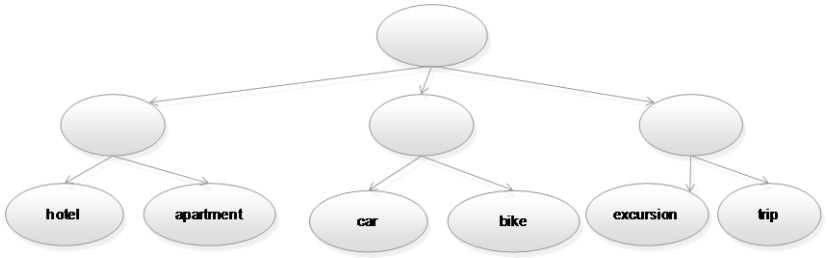

Figure 8 Example for an automatically acquired concept hierarchy derived from KMeans clustering

3. 実験と評価

本章では，前章で述べたクラスタリング手法を用いて，実世界ユーザ行動の階層構造を構築し，その精度を評価する．3.1 節で手動で正解モデルを構築する方法について述べる．3.2 節でクラスタリングで得られたモデルと正解モデルとの比較により精度を評価する方法について述べる．

3.1 マニュアルでの正解モデルの構築方法

マニュアルでの正解モデルの構築方法について述べる．本稿では Naganuma らが提案する方法 [4]を利用する．Naganuma らのモデルの構築指針は，「上位のタスクではユーザの抽象的な要求や問題を表現しており，下位のタスクではより具体的な解決策を表現している」であり，モデルの構築方法は本方針を具現化したものである．1)第一に，全ての実世界ユーザ行動から複数のメインのトピックを洗い出し，トピックごとに実世界のユーザ行動を分類する．2)次に，トピックごとに，トピックに含まれるユーザ行動の階層構造化を行う. まず, クラスタ内で最上位のユーザ行動を決定する．ユーザ行動モデルの最上位のノードには，その他のユーザ行動を意味的に包含するため，最も一般的なユーザ行動を選択する．ユーザ行動が動詞と目的語から構成されるが，一般的なユーザ行動は，目的語が 1 つでかつ一般名詞であることが多い．3)最後に，最上位で選択されたユーザ行動以外の他のノードを階層化する．まず，最上位のユーザ行動を具体化するユーザ行動を最上位のユーザ行動の子ノードとして配置する．具体化は，WHAT の観点と HOW の観点の 2 通り存在する．この処理をクラスタ内のすべてのユーザ行動が配置されるまで繰り返す．Table1 のユーザ行動に対してマニュアルで構築した場合，Figure9 に示すユーザ行動のモデルが得られた．

Figure 9 Manually constructed Reference concept hierarchy of user’s activity in dvd-related domain

3.2 評価方法

クラスタリングによって獲得された階層構造のモデルを評価するする手法について述べる． [2]では，獲得したモデルと正解のモデルがどの程度重複しているかによって，獲得したモデルの精度の評価を行う方法が提案されており[2]，本研究においても評価手法として用いる．この方法によれば，重複度が高いほど高精度で階層構造を構築できるといえる．以下，正解のモデル O1 と，学習によって得られたモデル O2 間の重複度 TO(O1,O2)を測る方法について述べる．モデルの重複度を測るためには，定義されている概念単体だけを比較するのではなく，モデル内で概念が他の概念とどのようなつながりを持っているかを比較する必要がある．そこで，モデル内で定義されているすべての概念 C をモデル内の他の概念を用いて表現する．具体的には，Semantic Cotopy=（C，C の親ノード，C の子ノード）の組み合わせで表現する．次に，正解モデル内の概念 C と，学習モデル内の概念 C’がどの程度近いかを計算する．

$$TO(c,c',O_1,O_2)=\frac{|SC(c,O_1)\cap SC(c',O_2)|}{|SC(c,O_1)\cup SC(c',O_2)|}$$

ⓒ2010 Information Processing Society of Japan

Vol.2010-MBL-56 No.3
Vol.2010-ITS-43 No.3
2010/11/11

ここで，TO（c,c’,O1,O2）は，正解モデル O1 内の概念 c と，比較対象のモデル O2 内の概念 c’の概念の重複度を表す．また，SC(c,O1)とは，概念 c の O1 上で得られる SemanticCotopy を意味する．例えば，Figure10 に示す正解モデルの中の Vehicle と，Figure6 に示す FCA で得られた学習モデルの中の Driveable がどの程度近いかを計算する．正解モデルの中の Vehicle の SemanticCotopy は以下のとおりとなる．

$$SC(c,O_1)=\{car,bike,two-wheeled-vehicle,vehicle,object-to-rent\}$$

一方，FCA で得られた学習モデルの中の Driveable の定義は以下のとおりとなる．

$$SC(c',O_2)=\{bike,car,rideable,driveable,bookable\}$$

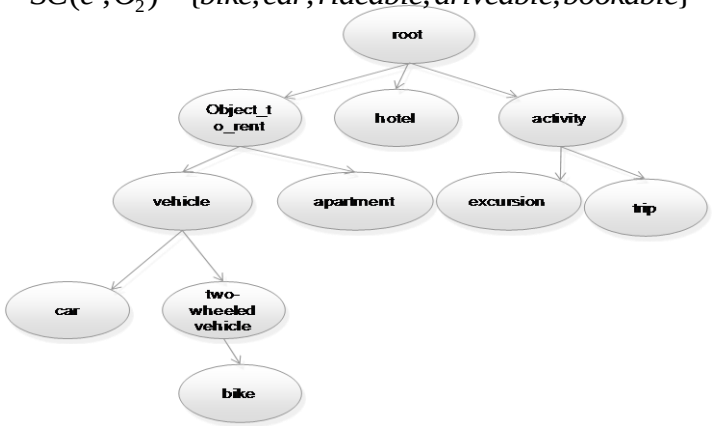

Figure 10 Example for an reference concept hierarchy

ここで，重複度 TO の分子である，概念 c と c’の積集合の概念数は，Car, Bike の 2 個であり，分母である概念 c と c’の和集合の概念数は 6 個 であることから，正解モデルの Vehicle と，学習モデルの Driveable の類似度は 2/6=33%となる．この計算を学習で獲得されたモデルに含まれる全ての概念に対して行い，正解のモデル内の概念 c について，学習で獲得されたモデル内で最も類似する概念 c’を見つける．

$$TO(c,O_1,O_2)=\max_{c'\in C_2}\frac{|SC(c,O_1)\cap SC(c',O_2)|}{|SC(c,O_1)\cup SC(c',O_2)|}$$

最後に，正解モデル内のすべての概念 C について上記の計算を行い平均値を計算する．

$$\overline{TO}(O_1,O_2)=\frac{1}{|C_1|}\sum_{c\in C_1}TO(c,O_1,O_2)$$

ここで C1 は正解モデル O1 に含まれる全ての概念集合を意味する．

4. 結果と考察

2 章で述べたクラスタリングアルゴリズムごとに得られた階層構造を図 11，12，13 に示す．図 11，図 12 から分かるように，いずれのクラスタリングアルゴリズムを用いた場合でも，主な 3 つのトピック（Make/burn/copy dvd, buy dvd, find dvd）に分類することができている．これは，Figure9 に示す，正解データと同様である．また，Table4 に各アルゴリズムで生成された階層構造と正解の階層構造間の重複度を示す．

Table 4 Evaluation result of Taxonomic overlap

	BiSection kmeans	FCA	Hierarchical Clustering
Taxonomic Overlap(%)	39.9	46.5	39.9

Table4 からわかるように，FCA による階層構造が最も正解例と重複度が高いことがわかる．文献[2]の報告でも FCA による名詞概念の階層構造の精度が高いことが報告されており，実世界行動間の階層構造の構築においても FCA が有力であると考えられる．ただし，FCA には，アルゴリズムの性質上，Attribute の数分だけ空のノードが増加するため，モデルとしての視認性に欠けるという欠点がある．したがって，重複度に違いがあまり見られない場合は，K Means もしくは階層型クラスタリングを利用するほうがよい場合もある．例えば，クラスタリングで得られた階層構造を基に，マニュアルとで正確なモデルを構築する場合には，クラスタリング結果が作業者が見て分かりやすいモデルとなっている必要がある．参考までに Figure14 に結婚のドメインに関する実世界行動を収集し，それを階層型クラスタリングによって階層化した例を示す．

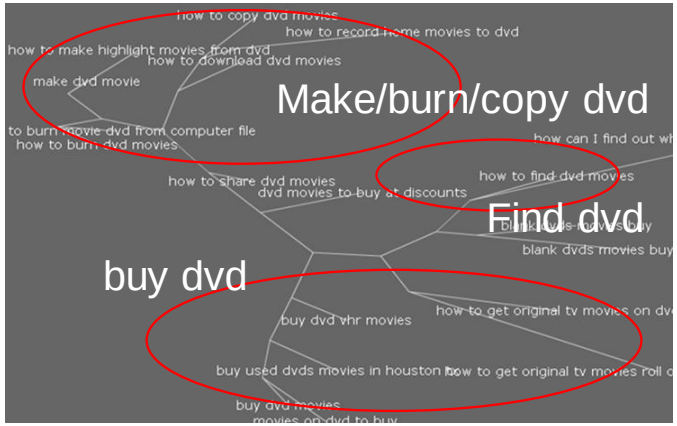

Figure 11 Concept hierarchy of user’s activity in dvd-related domain constructed by Hierarchical agglomerative clustering

ⓒ2010 Information Processing Society of Japan

Vol.2010-MBL-56 No.3
Vol.2010-ITS-43 No.3
2010/11/11

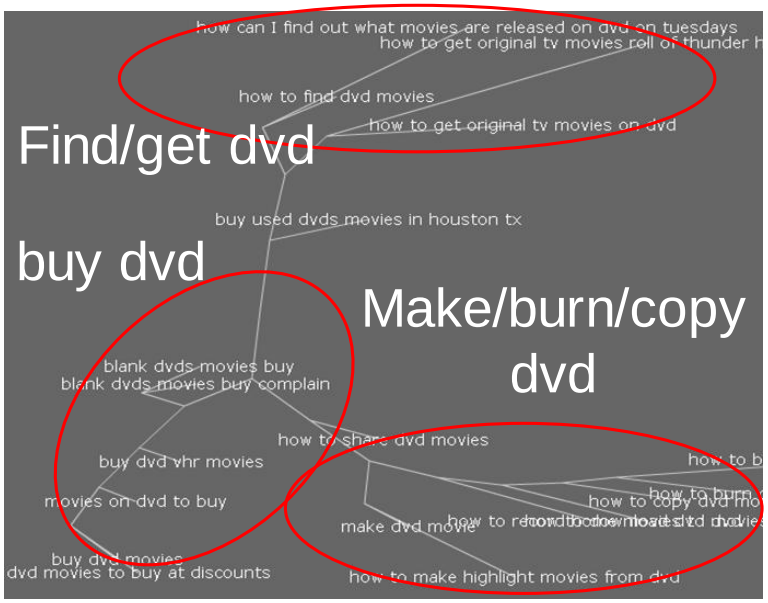

Figure 12 Concept hierarchy of user's activity in dvd-related domain constructed by KMeans clustering

5. おわりに

本稿では，ユーザの実世界行動モデルを自動で構築するため，クラスタリング手法の適用可能性を検証した．検証実験の結果，マニュアルで構築したモデルと 50%程度重複するモデルを構築することができた．現状，実世界行動モデルをマニュアルで構築するには 1 ヶ月程度かかることから，クラスタリングとマニュアルを組み合わせることで構築に要する時間を劇的に削減できると考えられる．今後は，クラスタリングによる実世界行動モデルの構築精度を高め，よりマニュアルで構築したモデルに近づけるため，ユーザ行動の特徴抽出方法，クラスタリングアルゴリズムの改良を行っていく予定である．

謝辞　本研究を遂行するにあたり，AOL 社の検索クエリデータから実世界行動に関するクエリのみを抽出したクエリセットを提供頂いた Graz University of Technology の Markus Strohmaier 教授に謝意を表する．

Figure 13 Concept hierarchy of user's activity in dvd-related domain constructed by FCA

参考文献

1) Y. Fukazawa, T. Naganuma, K.Fujii and S.Kurakake: Construction and Use of Role-ontology for Task-based Service Navigation System, ISWC2006 5th International Semantic Web Conference, pp. 806-819, 2006.
2) Philipp Cimiano: Ontology Learning And Population from Text: Algorithms, Evaluation And Applications, Springer-Verlag, 2006.
3) M. Strohmaier, P. Prettenhofer, M. Kroell, Acquiring Explicit User Goals From Search Query Logs, International Workshop on Agents and Data Mining Interaction ADMI'08, in conjunction with WI'08,2008.
4) T. Naganuma and S. Kurakake. Task Knowledge Based Retrieval for Service Relevant to Mobile User's Activity. ISWC2006 5th International Semantic Web Conference, pp. 959–973, 2005.
5) R. Masuoka, B. Parsia, and Y. Labrou. Task computing - the semantic web meets pervasive computing. In Proc. ISWC2003, pages 866–881, 2003.
6) A. Bernstein and M. Klein. Towards high-precision service retrieval. In Proc. ISWC2002, pages 84–101, 2002.

ⓒ2010 Information Processing Society of Japan

Vol.2010-MBL-56 No.3
Vol.2010-ITS-43 No.3
2010/11/11

Figure 14 Concept hierarchy of user's activity in marriage-related domain constructed by Hierarchical agglomerative clustering

ⓒ2010 Information Processing Society of Japan