F-025

Web ページのテーブルに着目したオントロジ構築及び，比較情報抽出方式の提案

Comparison information extraction using ontology generated from web tables

境　美樹†　佐藤　宏之†　村山　隆彦†
Miki Sakai　Hiroyuki Sato　Takahiko Murayama

1. はじめに

人が情報検索を行う際の行動には，1) 課題の選定，2) 情報探索戦略の決定，3) 情報源へのアクセス，4) 情報の獲得，5) 情報の統合，6) 評価，の 6 段階のプロセスがあるとされている[1]．例えば，「新しい住居を探す」という課題を選定した場合，情報探索の戦略として「場所」や「予算」などの条件を決める．その後，Web での検索や直接不動産屋に行くなどして情報源にアクセスし，必要な情報を獲得し，それらの情報を何らかの観点で統合し，評価を行うというプロセスを踏む．最終的に気に入った物件が見つかるまで，獲得した情報を比較しながら 3) ～ 6) の行動を繰り返す．この様な場合，人は「バルコニーがある or ない」「床がフローリング or 和室」などの比較観点を利用し，情報を選択する．一方で，情報を収集し整理している段階で気がつかない比較観点もあるだろう．

本稿では，この様な Web を使った検索行動を支援するため，Web ページから情報の比較観点と値のセットを自動的に抽出し，比較表を自動生成する方式の提案を行う．

2. テーブルからの情報抽出

本研究では，課題の定まったユーザが Web ページを用いて情報探索を行う際，その目的に対する比較表を提供することが目的である．例えば，ユーザのキーワード入力に応じて比較表が提示したり，常にブラウザの横に表示されていて閲覧ページに応じて比較表が変動していくと，ユーザの利便性が向上する．

Web ページを閲覧していると，表として比較情報が提供されていることがある．この有益な情報を利用したいと考える．その一方で，ページ作成者の意図に応じた情報であるため，全情報が網羅されておらず，比較対象や比較項目が不足している場合がある．

そこで，不足している情報は，別の表や表以外の部分から抽出し補う．そのため，表から生成した辞書を利用し，情報抽出に活用することを考える．

3. Web ページからの情報抽出

本稿では，利用者がある課題に対して閲覧した情報から表形式の情報を探してオントロジ（辞書）を構築し，表形式以外の情報からも関連情報を抽出することで，不足した情報を補う方式を提案する．

ここで述べるオントロジは，RDFSchema[2]を利用しプロパティ(property)に対して，サブジェクト(subject)とオブジェクト(object)の取り得る概念が定義されているものを指す(図 1 の上部)．プロパティは比較項目，サブジェクトは比較する対象，オブジェクトは比較項目に対応する値のことである．

具体的には，以下の手順を取る．

†日本電信電話株式会社　NTT 情報流通プラットフォーム研究所
NTT Information Sharing Platform Laboratories, NTT Corporation

まず，情報抽出の種となるオントロジを構築するため，Web ページのテーブル構造に着目し，オントロジを構築する．Web ページにおいては，様々な著者が，情報を表として提示している場合があるため，より信頼性の高いオントロジが構築できると考える．

オントロジ構築では，着目する表の対象が何であるか分かっていると仮定して，以下の手順で行う．

1. セルのスコアリング
2. 比較項目の特定
3. オントロジの構築

以下，それぞれの詳細手法を記述する．

3.1 セルのスコアリング

表において，一般的に比較項目は，同一の並びに存在すると仮定する．各表のセルに対して，次に示すルールを適用しスコアを与え，行・列ごとに集計したスコアが高い行，又は列を比較項目の並びとする．

【表の目的に対するオントロジが無い場合】

(1) セル内の語彙の品詞に応じてスコアリングする．
(2) セルと同一行又は列に同じ語彙が複数ある場合，スコアを下げる．
(3) 文章構造を持つセルはスコアを下げる．

【オントロジがある場合】

(4) オントロジ内のプロパティとセル内の語彙が同一概念ならば，スコアを上げる．
(5) セルと同一行又は列にオブジェクトと同一概念がある場合，スコアを上げる．

例えば，図 1 のオントロジが存在し，表 1 をスコアリングする場合，1 行 4 列のセル「色」は，オントロジのプロパティと一致し，更に，「色」のセルと同一列にはオントロジのオブジェクトのインスタンス「シルバー」があるので高いスコアが与えられる．

3.2 比較項目の特定

比較項目の並びが確定すれば，それに対応する行又は列は自動的に決まる．しかし，候補となる行又は列が複数あり，かつ行と列どちらにもある場合は，行と列のどちらが一つのインスタンスについて述べているのかを調べる必要

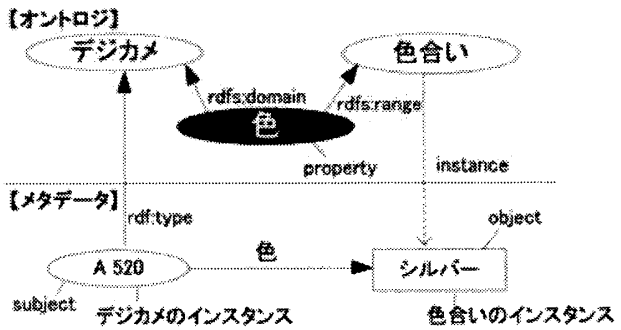

図 1　オントロジとメタデータ

表 1　デジカメに関する表

メーカー	型番	画素数	色	電池
A社	A-01	310万	ホワイト	充電池
A社	A 520	520万	シルバー	充電池
B社	B200	200万	ブラック	単3
C社	C300	310万	シルバー	単4

がある．

そこで，以下のルールを適用する．

ある行又は列に類似した内容のセルが続くならば，その行又は列と直交する列又は行が一つのインスタンスについての記述であるとして，最終的な比較項目の並びを特定する．

比較項目が特定できたら，その比較項目をプロパティとする．更に，比較項目の行又は列と直交し，かつスコアの高い行又は列はサブジェクトとみなして，各行・列ごとに情報抽出を行う．

例えば，表 1 の場合，1 行目と 2 列目のスコアが高くなるが，他の行・列の並びを見ると行単位で一つのインスタンスを表している事と判断され，1 行目が比較項目だと特定され，1 列目と直交する 2 列目をサブジェクトとみなすと，図 2 のようなメタデータが抽出される．

3.3 オントロジの構築

上記方式で抽出されたメタデータから，図 1 と同形式のオントロジ候補を生成する．

本稿では，抽出対象の表が何について記述されているか予め分かっていると仮定されているので，表 1 の場合，プロパティは比較項目，サブジェクトの取り得る概念は「デジカメ」となる．オブジェクトは，プロパティに対応する値を見て，数値などは定式化し，それ以外はそのまま定義し，オントロジ候補とする．

候補としたオントロジは，更に別の表からオントロジ構築を行い，同様のオントロジ候補が生成された時点で正式にオントロジ化する．

この様な手順でオントロジ化するのは，表にある比較項目がその対象に対して一般的ではない場合や，誤認識して間違ったオントロジが生成された場合を考慮しているからである．

4. 予備実験

3.1 節の(1)〜(3)を用いてサブジェクト及びプロパティの並びを特定する予備実験を行った．対象とする表は「デジカメ」に関するもので，8 つの異なる Web ページから合計 17 個の表を抽出し，利用した．

その結果，プロパティの並びを特定できたものは 64.7%，サブジェクトの並びを特定できたものは 47.1%だった．プロパティの並びを特定する場合，比較項目として現れる語彙が特徴的なので，スコアリングのパラメータを調整することで精度を向上させることが可能と考える．

一方，サブジェクトの並びを特定する場合，それだけでは上手くいかない．今回，同一ページ内に同一構造を持った表が複数出現しているものがあり，本来は同一の行と列がサブジェクト・プロパティとして判定される筈である．しかし，行又は列数が極端に少ない表では,プロパティの並びは正確に判定されるがサブジェクトの並びは誤った判定がなされ，行又は列数が適度な場合は，サブジェクトの並びも正確な判定となっており，この様な場合，精度は行又は列数に依存することが分かった．

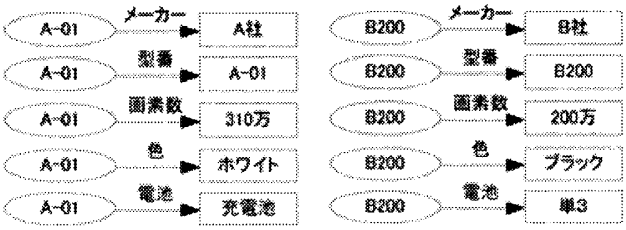

図 2　表 1 から抽出したメタデータの一部

そこで，3.1 節(4)(5)で提案した，オントロジを利用した概念のマッチングによるスコアリング方式を用いれば，サブジェクト及びプロパティの並びのスコアを高めることが可能となり，より正確にサブジェクト及びプロパティ位置の特定が出来ると考える．

5. 関連研究

本稿の関連研究として，表の取り得る構造を予め定義しておき，セル間の配置などからオントロジを獲得する研究がある[3]．定義とマッチする表では，高い精度が得られると報告されている．また，セル内の語彙の配置から，表の構造を認識し，クラスタリングを用いて複数の表をマージするという研究がある[4]．本稿では，抽出したオントロジを利用して，構造認識をより容易にするというアプローチを取っている．

6. おわりに

比較観点とその値を抽出するために，テーブル形式に着目した比較情報抽出及びオントロジを構築の手法を提案した．この手法により，オントロジが構築され，様々な情報源からメタデータ抽出が出来るようになり，ユーザの目的に対応した比較表の提供を行うことができる．また，Semantic i タウンページ[5]における店舗情報抽出方式への適用も可能となる．

今後は，提案方式の有効性の検証，ユーザが閲覧している情報の内容を認識，対応した情報を抽出するための方式などを検討する．また，抽出したメタデータの利用方法として，比較表だけではなく，情報間の関連性を考慮した検索やある目的に対するページの自動生成などへの応用を検討したい．

参考文献

[1] 三輪: 情報検索のスキル, 中央公論社, 2003.
[2] Dan Brickley, R.V.Guha :
RDF Vocabulary Description Language 1.0: RDF Schema,
http://www.w3.org/TR/2004/REC-rdf-schema-20040210/
[3] 田中, 石田: 表形式からのオントロジーの構築, 人工知能学会研究会資料 SIF-SWO-A404-05, 2005.
[4] 吉田,鳥澤,辻井: 表形式からの情報抽出手法, 言語処理学会第 6 回年次大会発表論文集, pp. 252-255, 2000.
[5] 境, 佐藤, 向垣内, 村山: オントロジを活用したポータルサービス〜Semantic i タウンページ 〜, INTAP セマンティック Web コンファレンス 2005 , http://www.net.intap.or.jp/INTAP/s-web/data/conference2005/08-ntt_kagami.pdf , 2005.