1C-02

オープンデータのRDF化のための類似項目間関係の推定

山根 昇平† 鵜飼 孝典†
†株式会社富士通研究所

1 はじめに

オープンデータの多くはPDFやスプレッドシート，CSVなどの形式で公開されている．このようなデータは，人間が読んで理解したり，特定のデータに特化したアプリケーションの開発を行う場合にはそのまま利用できる．しかし，他のデータと組み合わせて機械処理するためには，オープンデータの5star[1]が示すようにLinked Open Data(LOD)とすることが好ましい．

既存データをLODに変換するためには，データ間の関係を明らかにする必要がある．例えば，ある文書について作成者が列挙されていた場合，それらが単なる集合であるのか，順番に意味のあるリストであるのか明示する．ところが，既存のデータの多くは，表形式に代表されるような属性と値の組の単純な集合であるため，属性間の意味構造が明示されていない．

既存データから意味構造を獲得する手法として，複雑な表構造を一般化して意味構造を抽出する手法[2]，学習による手法[3]，データセット間のスキーママッピングを行う手法[4]などを用いることができる．これらの手法では，一般化した表構造に対するオントロジーや学習のためのデータを，専門家が与える必要がある．これらに対し，本研究では，単純な表形式のデータ単体から，属性間の意味構造を推定することを目指す．

本研究で示す手法は，与えられた属性群に対する値の一致関係に注目して，意味関係を推定する．実際のデータを用いて小規模な実験を行ったところ，正確な判定を機械的に行うことは困難であるが，人間の判断を効率化するための情報が得られることがわかった．

2 対象データ

本研究では，複数の事象と複数の属性名があり，各事象と属性名の組に対してひとつの値が記述されているデータを入力とする．例えば，表1のような文書に関する表形式データがある．この例では，各文書に対して書名やカテゴリという属性名があり，各セルには，ある文書におけるある属性の値が示されている．文書データでは，書名，カテゴリのほか，発表日，作成者，キーワードといった属性がある．

表1: 表形式データの一部

書名	カテゴリ1	カテゴリ2	カテゴリ3	カテゴリ4
A	計算機科学	知能情報学	人工知能	エージェント
B	計算機科学	知能情報学	人工知能	機械学習
C	計算機科学	知能情報学	自然言語処理	
D	計算機科学	プログラミング言語	パラダイム	関数型言語
E	計算機科学	プログラミング言語	コンパイラ	

このようなデータは，類似する属性名によってなんらかの関係を持つことが示唆される属性群を含むことがある．例えば，表1では，カテゴリ1からカテゴリ5が連番の属性名を持つ．本研究は，このような属性群が持ちうる意味構造を，集合，リスト，階層，同値，それ以外（無関係）の5種類に分類し，与えられた属性群が5種類のうちどれに該当するかを推定することを目的とする．

3 意味関係の推定

本稿で示す手法では，意味構造を集合，リスト，階層，同値，無関係の5つに分類し，与えられた属性群がそれらのうちどの意味構造を持つか推定する．以下に，それぞれの意味構造と判定基準を概説する．

集合． 文書におけるキーワードのように，複数の値の単なる列挙であり，順番がない．図1のように，異なる行，異なる列において同じ値が出現することで判定する．これは，各列に意味的な違いがないことを示唆している．ただし，同じ行に同じ値が出現する場合は除く．

リスト． 論文における著者のように，複数の値の列挙であるが，順番が定義されている．図1のように，異なるふたつの行について，値の入れ替わったデータが出現することで判定する．これは，属性名の意味は同じであるが，順番に意味があることを示唆する．ただし，同じ行に同じ値が出現する場合は除く．

階層． 文書における分類のように，上位・下位の概念によって構造を持つ．属性群を順に並べ，重複を除いた値の数が単調増加になっていることで判定する．空値を持ちうるため，属性数を2から属性群の大きさまで順に増やしていき，空値を含む行を除外する．そして，すべての属性数で要素数が単調増加となるとき，階層であると判定する．表1の例では，属性数2のときカ

Relation Inference Between Attributes for Making Linked Open Data
†Shohei Yamane †Takanori Ugai
†Fujitsu Laboratories Ltd.

属性1	属性2	属性3
あああ	いいい	うううう
あああ	うううう	
えええ	おおお	

属性1	属性2	属性3
あああ	いいい	
あああ	うううう	
いいい	あああ	

図 1: 集合の判定（左），リストの判定（右）

表 2: カテゴリ 1 から 5 の階層判定結果

属性数	1	2	3	4	5
2	11	65			
3	7	31	287		
4	4	12	25	96	
5	2	8	12	41	18

テゴリ 1 の要素数は 1，カテゴリ 2 の要素数は 2 となり，単調増加．属性数 3 のときカテゴリ 1,2,3 の要素数はそれぞれ 1,2,4 となり単調増加．属性数 4 のときは 3 行目と 4 行目を除外するため，カテゴリ 1,2,3,4 の要素数はそれぞれ 1,2,2,3 となり単調増加．以上により，この例は階層であると判定される．

同値． 文書における和文タイトル・英文タイトルのように，単一のものに複数の表記を与える．値が必ず同一の組となるか否かで判定する．

無関係． 上記いずれも該当しない場合，無関係，すなわち，それぞれの属性が別々の意味を持つと判定する．

4 実験と考察

予備実験として，社内文書メタデータ 6730 件を用いて判定を行った．判定を行った属性名は，カテゴリおよび作成者である．カテゴリはカテゴリ 1,2,3,4,5 という属性名で与えられ，カテゴリ 1 が上位，カテゴリ 5 が下位の階層構造になっている．作成者は作成者氏名 1,2,3,4,5 という属性名が与えられ，最大 5 人の作成者氏名の集合が表されているが，集合かリストかは明らかではない．まず，カテゴリについて本手法を適用したところ，階層構造については 2 に示した結果が，また，集合判定については 13833 組が発見された．その他は該当しなかった．表 2 の通り，属性数 5 のときのみ，カテゴリ 5 の要素数 18 が，カテゴリ 4 の要素数 41 を下回り，単調増加とならない．この結果により，カテゴリ 1 から 5 は階層でないと判定される．作成者について本手法を適用したところ，集合判定については 270 組，リスト判定については 7 組となり，共に該当した．それ以外は該当しなかった．

カテゴリについては，階層構造であることがわかっている一方，階層でないと判定された．しかしながら，表 2 の通り，属性数 4 までは階層構造の条件を満たしており，また，カテゴリ 5 の要素を見ると，上位階層が異なる行について “レベル 1”，“レベル 2” など抽象的な表現が繰り返し出現するため，このようなデータを見れば階層であると判断できる．また，作成者については，集合かリストか事前に明らかではなかったものの，リストである可能性を示す少数の要素を発見することができた．

5 おわりに

本稿では，オープンデータの LOD 化の支援を目的とし，複数の事象と複数の属性名があり，各事象と属性名の組に対してひとつの値が記述されているデータを対象として，属性間の意味構造を推定する手法を示した．この手法により，属性間の関係が示されていないデータ形式において，値の一致関係によって意味的構造を推定することが期待できる．また，同じ属性名で構造が異なる場合，例えば，“作成者” のように順番のない集合である場合と，論文のように順番のあるリストである場合があるものについても．値を参照することにより判定することも期待できる．

しかしながら，予備実験では，5 つの意味構造のうち複数に該当する場合があり，また，その精度も高くない．一方で，意味構造を判断する重要な示唆を与えることができた．したがって，今後は，精度向上のための手法の改善のほか，人間の判断を効率化するための情報の提示手法の検討を行う．

参考文献

[1] Tim Berners-Lee. Linked data. `http://www.w3.org/DesignIssues/LinkedData.html`.

[2] 田仲正弘, 石田亨. 表構造の一般化に基づくオントロジーの獲得. 情報処理学会論文誌, Vol. 47, No. 5, pp. 1530–1537, 2006.

[3] Julius Volz, Christian Bizer, Martin Gaedke, and Georgi Kobilarov. Discovering and maintaining links on the web of data. In *Proceedings of the 8th International Semantic Web Conference*, ISWC ’09, pp. 650–665, Berlin, Heidelberg, 2009. Springer-Verlag.

[4] Philip A. Bernstein, Jayant Madhavan, and Erhard Rahm. Generic schema matching, ten years later. *PVLDB*, Vol. 4, No. 11, pp. 695–701, 2011.