第4週B4 ゼミ

情報の意味的な統合とオントロジー写像

出典

市瀬 龍太郎, 情報の意味的な統合とオントロジー写像. 人工知能学会誌, Vol. 22, No. 6, pp. 818-825, (2007)

背景

- 多様なデータをどのように連携させるか
  - メタデータを利用
    - 異なるメタデータを持つ場合連携できない

- 意味的な情報統合
  - メタデータ(データスキーマ、オントロジー)間で対応関係を発見し、統合、変換を自動的に行う

目的

・この研究分野では複数のスキーマやオントロジーに対象として「写像」、「対応」、「調節」を決める

・研究の状況を概観し、特にオントロジーを対象とした意味的な情報統合にどのようなアプローチがあるかをシステム例とともに解説する。

意味的な統合の事例

・会社合併に伴い、顧客のデータベースを統合しなければならない。データスキーマが異なるデータベースをどのように統合するか。

・スキーマ対応

　-スキーマ間にどのような関係があるのか発見し、　統合 しなければならない

・A:「住所」　　B:「都道府県」、「市町村名」

「住所」を2つに分解しなければならない

意味的な統合の事例

・工業部品を出す会社が多数ある。各会社は独自の体系で部品を分類、カタログを作成している。販売店がそれらをまとめて新たなカタログを作るにはどうすべきか

・カタログ統合

　-概念体系の間にどのような関係があるか考慮

・A社:「赤色ダイオード」　　販売店:「ダイオード」

「赤色ダイオード」を「ダイオード」に統合すればよいが、逆はできない

意味的な統合の事例

・Webサービスがある。セマンティックWebではオントロジーでどのようなサービスであるか記述を作る。しかし、すべてを共通のオントロジーに基づいて記述することは難しい。多数のWebサービスを連携させるにはどのようにすればよいか。

・Webサービス合成

-オントロジー間にどのような関係があるか考慮しなければならない

・これらの問題をオントロジー写像問題と呼ぶ

基本的な要素技術

- オントロジー写像問題
  - -片方のオントロジーに存在する概念、プロパティ、インスタンスをもう1つのオントロジーのどこに対応するかを発見すること

- 文字列に基づく手法
- グラフに基づく手法
- 知識資源に基づく手法
- インスタンスに基づく手法

文字列に基づく手法

- 概念、プロパティに付いているラベルを利用する
- 得られた文字列がハイフンなどで複数の語が接続されたものの場合分解する(PyramidTheory)
- 類似度を測定し、対応するか否か判定する
- インスタンス、概念の説明文を利用できる場合がある

グラフに基づく手法

・一方の木構造で表されたオントロジーが、グラフ的にもう一方のどこに対応するかを見つける

・グラフ的に隣接するものを利用することができる

- 親子関係、兄弟関係など上下関係が逆転することがないもの

知識資源に基づく手法

- 外部の知識源(辞書、シソーラス)を対応の発見に利用する
  - 類義語辞典を用いて異なる語の対応関係を発見する

- 別のオントロジーや共通語彙を経由して対応関係を発見

インスタンスに基づく手法

・インスタンスの分類に着目し、概念の対応を決定する

・インスタンスの分類方法が似ていれば、分類している概念同士も近い

代表的なシステム

- GLUE
  - オントロジーの概念とインスタンスに、機械学習手法による学習と分類を行う(文字列に基づく手法)
- 分類した確率分布から概念間の類似度を計算
- 与えられたヒューリスティックから対応が計算される

代表的なシステム

- S-Match
  - 外部の知識資源を利用して、より詳しい概念記述を作成する
- 記述は論理記述として扱える
  - 関係発見の精度向上(制約充足問題)

代表的なシステム

- HICAL-NB
  - HICALにNBを組み合わせ拡張(インスタンスに基づく手法)
- HICAL
  - 統計によりインスタンスの類似性を測る

- NBにより対応関係の詳細化を行う
  - 精度の向上

オントロジー写像発見システム

・Malfom-SVM

・接頭辞一致、接尾辞一致、編集距離、nグラムをもちいて類似度を計算

・語の組み合せを利用したラベル

-リストを作成し、リストの類似度を計算する

オントロジー写像発見システム

・知識資源として同義語(synset)、Wu&Palmer、説明(description)

・同義語
  – WordNetの同義語のパスの長さ
・Wu＆Palmer
  – 深さと最小共通上位概念
・説明
  – WordNetにおける説明

オントロジー写像発見システム

・パスを利用した類似度

・最上位の概念からの関係

　-パスに対する類似度を計算する

・対象概念の付近の概念をもちいて局所的な構造に対する類似度を計算

性能評価

- 3グラムを用い、学習器としてSVMを適用
- 2160個の正例、2327個の不例、計4487個の概念対

	hmatch	falcon	RiMOM	Malfom-SVM
適合率	32.4	40.5	39.3	52.5
再現率	13.4	45.5	40.4	92.5
F値	18.9	42.9	39.8	67.0

まとめ

・情報の意味的な統合をどのようにするかという問題について要素技術、代表的なシステムの手法について概説した

・