2C2-2

属性の相互関係を利用したWebからの属性抽出

Attribute Extraction from the Web based on Correlation between Attributes

谷 直紀[*1] Naoki Tani
ボッレーガラ ダヌシカ[*1] Bollegala Danushka
石塚 満[*1] Mitsuru Ishizuka

[*1] 東京大学大学院情報理工学系研究科
Graduate School of Information Science and Technology, the University of Tokyo

Attribute extraction is important to organize information on the Web. However, choosing correct attributes from numerous candidates is difficult. In this paper, we propose an attribute extraction method using the correlation between attributes. For instance, an entity cannot have both "occupation": "sports player" and ”award”: "Nobel prize". A reason for this is that typically "sports player do not receive the Nobel prize". Experiments using Freebase, which is a large database, show that attributes are mutually related.

1. はじめに

1.1 研究の背景

現在，インターネット上には大量の情報が存在する．その内容は多岐にわたり，ほぼ全ての分野の情報を見ることができる．しかし，個々のユーザが独自に情報発信をしているため情報は整理されておらず，関連する情報が様々な Web ページに分散して保存されている．さらに，一定のフォーマットも定まっておらず，必要な情報がどこにあるか分かりにくくなっている．Web 上の大量の情報を整理してユーザに提示するためには，関係する文章を見つける「情報検索」ではなく，関係する情報を文章から抜き出す「情報抽出」の技術が求められている．

属性抽出とは情報抽出の一種であり，あるクラスを与えられた時にそのクラスの所有している属性を見つけることだ．属性は質疑応答タスクなど他のタスクにも利用できるため，重要な技術とされている．

属性抽出について様々な研究が行われているが，ほとんどの手法は属性を個別に抽出している．しかし，同じクラスの所有している属性は独立ではなく，共起しやすい属性や共起しにくい属性が存在すると考えられる．そこで本論文では，属性同士の関係性を利用した新しい属性抽出手法を提案する．

1.2 本論文の構成

第 1 章では，本研究の背景について述べた．第 2 章では，属性抽出に関する研究動向について述べる．第 3 章では提案手法を説明する．第 4 章では属性同士の関係を確認する実験を行う．第 5 章では本研究の成果，および今後の展望を論ずる．

2. 関連研究

人物に関する情報抽出のワークショップである Web People Search Evaluation Workshop (WePS)[1] において，属性抽出タスクが行われている．そのタスクは回毎に多少異なっているが，大きな目的はウェブ上に出現する人名の曖昧性を解消することである．すなわち，入力として人物の名前と，検索エンジンによって集められたドキュメントが与えられ，そのドキュメント集合を，同姓同名ではあるが異なる人物ごとにクラスタリングを行い，同姓同名の人物が何人居て，それぞれのドキュメントはどの人物に属するかを当てることが目的である．渡部らは，正規表現を利用して属性値候補のタグ付けを行った後に，人名からの距離などによって属性値を選び出す手法を提案している [Watanabe09].

ある特定のコーパスから属性を抽出する研究も存在する．Wu らは，Wikipedia の記事から Infobox を自動作成するシステムを作成した [Wu07]. Infobox とは，属性と属性値の組み合わせから成る Wikipedia 記事のまとめである．Wu らの手法は，属性抽出と属性値抽出の 2 段階で構成されており，1 番目の属性抽出ステップでは対象記事のクラス分類を行う．Wikipedia ではクラスごとに属性が定められているため，クラス分類によって記事の属性を決定できる．クラス分類は対象記事に付けられたタグ・リストを利用する．2 番目の属性値抽出ステップでは，属性値の周囲の文章を特徴量とした機械学習を利用する．

3. 提案手法

既存の手法でも属性を取得することはできるが，精度は十分とは言えない．精度向上の方法として，属性どうしの相関が利用できる．例えば，「1 歳の子供は政治家になれない」ことは自明である．この文章から，「職業」「年齢」という 2 つの属性の間に相関があるということが分かる．つまり，1 つのエンティティに「職業」:「政治家」と「年齢」:「1 歳」の 2 つの属性値が共起することはできないと言える．このように，属性同士の関係を利用することで，属性抽出の精度を高めることができる．

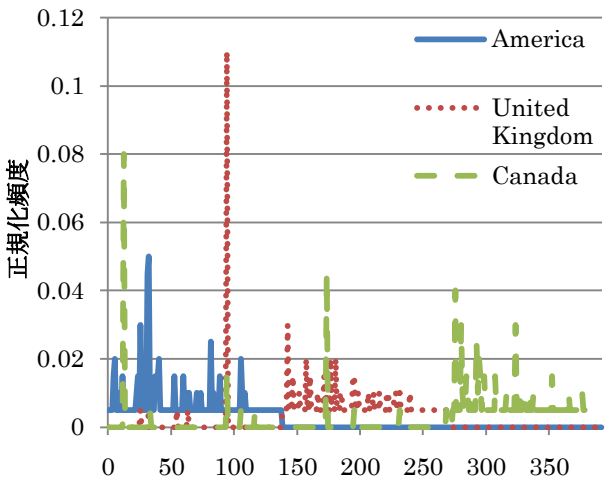
本論文では，下記の 3 段階の手法を提案する．入力は文章・エンティティ名・属性であり，出力は属性値である．Step1 では渡部らの手法 [Watanabe09] と同様，入力された文章に正規表現・抽出パターンを適用して属性値候補を集める．Step2 では，Freebase 上のエンティティが所有する属性値を調べ，属性値の共起しやすさを取得する．Step3 では，Step2 で調べた属性値同士の共起しやすさを利用し，属性値候補のランク付けを行う．

例えば「職業」「年齢」の 2 つの属性の値を求める場合，Step1 では「政治家」「1 歳」「50 歳」など複数の属性値候補が得られる．Step2 では「政治家」「1 歳」という属性値は共起しないが，「政治家」「1 歳」は共起する事が分かる．Step3 では，Step2 の関係を利用して「政治家」「50 歳」が正しい属性値であると推定する．

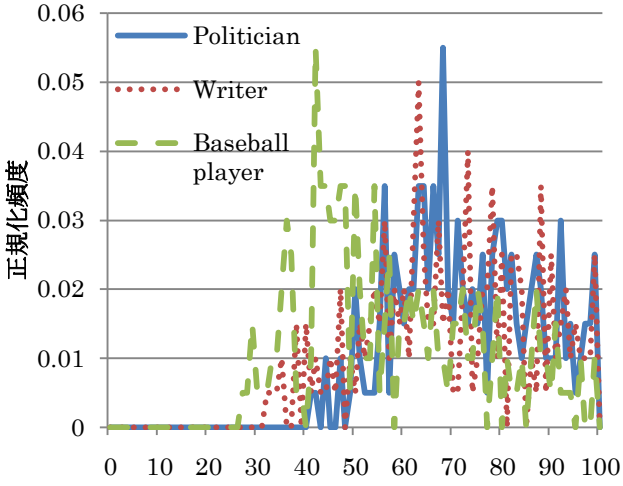
Step1: 入力された文章から属性値の候補となる語を見つける
Step2: 属性値同士の相関を測定する
Step3: 属性値候補から正しい属性値を選択する

連絡先：谷直紀，東京大学大学院情報理工学系研究科，
東京都文京区本郷 7−3−1，tani@mi.ci.i.u-tokyo.ac.jp

[1] http://nlp.uned.es/weps/



(a)国籍と出身地の関係



(b)職業と年齢の関係

図 1. 属性どうしの相関

4. 実験

提案手法では属性間の共起を利用して属性値候補のランク付けを行う. そこで, 属性どうしに関係があることを確認するための実験を行った.

4.1 データセット

データセットとして Freebase[1]を利用した. Freebase とは世界の情報を集めたオープンデータベースであり, 3000 以上の RDB と 1200 万件以上のエントリーによって構成されている. 不特定多数の人々が自由に編集するという点では Wikipedia と同じだが, Wikipedia では情報を文章で記述するのに対して Freebase では情報は属性:属性値のペアで記述されている. そのため, Freebase は属性抽出実験のためのデータセットに適していると. 今回の実験では, Metaweb Query Language (MQL) と呼ばれるデータベース言語を使って Freebase データを取得した.

4.2 実験方法

nationality(国籍)と place_of_birth(出身地), profession(職業)と age(年齢)という 2 種類の属性ペアについて, 属性間に関係があるかを調べた.「国籍」と「出生地」の関係についてはアメリカ・イギリス・カナダの 3 種類の国籍を持つ人物データを 200 件ずつ取得し, どのような出身地が表れるか調べた. 同様に,「年齢」と「職業」に関しても政治家・作家・プロ野球選手という 3 種類の職業について 200 件ずつデータを取得した. ただし, Freebase に年齢という項目はないため, 2010 年から誕生年を引くことで年齢を計算した. その際に年齢を 0〜100 歳に限定するため, 誕生年が 1910〜2010 年の人物のみを選んだ.

4.3 実験結果

国籍ごとの出身地分布を図 1 (a) に示す. 国籍によって出身地には偏りがあることが分かる. 例えば, アメリカ国籍を持つ人物はボストン (ID:31) やシカゴ (ID:25) などの出身地を持っていることが多い. 同様に, イギリス国籍を持っている人物にはロンドン (ID:94) 出身者が多い. また, 図 1 (b) の職業別年齢分布を見ると, 政治家と作家の年齢分布はどちらも似ていることが分かる. 一方, プロ野球選手の年齢分布は 42 歳にピークがあることから, 残りの 2 つの職業と区別することができる. このように, 図 1 から属性どうしの相関が見て取れる.

図 1 (b) を全体的に見ると, 実際よりも年齢が高く表されていることが見て取れる. 政治家・作家の半分以上は 60~90 歳代となっており, プロ野球選手の年齢ピークも 42 歳と高齢になっている. この理由として, すでに死亡した人物や引退したスポーツ選手を計算に入れていることが挙げられる. Freebase では死亡した年の情報や引退した情報は記載されていないため, どの人物が生きているかを判定することは難しい. Freebase だけでなくインターネット上の情報も利用し, 信頼度を高める必要がある.

5. おわりに

インターネットから属性を抽出する手法については研究されていたが, 抽出した属性候補を選択する手法に特化した研究は少ない. また, 属性同士の関係もあまり利用されていない. 本研究では, 属性同士の相関を利用したインターネットからの属性抽出手法を提案した. オープンデータベースである Freebase を利用した実験により, 属性同士が関係していることを示した.

しかし, 今回は属性抽出システム全体の評価を行っていない. 今後は属性抽出システム全体を実装し, インターネットから属性候補を見つける Step1 も含めた評価を行う必要がある.

また, 本実験では, 他の属性が決まった状態で残り 1 つの属性を決めるという問題設定をしている. しかし, 複数の属性が分からない実際の状態では, どの属性から決定していくかという問題が残っている. それぞれの属性の信頼度を決定することで問題を解決し, システムの性能向上を目指したい.

参考文献

[Watanabe09] K. Watanabe, D. Bollegala, Y. Matsuo, and M. Ishizuka, "A Two-Step Approach to Extracting Attributes for People on the Web," In Proceedings of the 2nd Web People Search Evaluation Workshop (WePS-09) at the 18th International World Wide Web Conference, 2009

[Wu07] F. Wu and S. Daniel, "Autonomously semantifying wikipedia," Proceedings of the sixteenth ACM conference on Conference on information and knowledge management - CIKM '07, 2007

[1] http://www.freebase.com/