Webからの関係情報マイニング

森 純一郎

1 はじめに

近年のWebにおける情報の爆発的な増加を受けて，Webから有用な情報や構造を抽出するWebマイニングに関する研究が盛んに行われてきている．特に，検索エンジンを利用したWebマイニング手法が注目されている．

検索エンジンを用いたWebマイニングの一例として，エンティティの自動抽出 [5] があげられる．エンティティ抽出とは，あるWebページに出現する人名，地名や組織名などのエンティティをWeb上における出現パターンや頻度を元に自動で抽出するものである．また，人と人，組織と組織といったエンティティ間の関係，ネットワークをWebから抽出する研究 [6] も行われている．

エンティティとエンティティのつながりが得られたときに，興味深いことは，その関係に関するさらなる情報である．関係を自動抽出する際に単に関係の強さだけでなく，その関係の背後にある情報も含めて抽出することで，関係構造だけでは浮かび上がってこない多様な意味づけと解釈を社会ネットワークに与えることできる．

本研究では，人と組織，人と地名といった，あるエンティティとエンティティの間の関係をあらわすような情報を関係情報として，それらの情報をWeb上からキーワードとして自動的に抽出する手法を提案する．エンティティ間の関係を表す情報とは，例えば政治家と地名というエンティテペアであれば，その政治家が元首，出身，選出など地名とどのような関係にあるかをあらわすものである．提案手法では同じ関係を持ったエンティティペアは同様の文脈でWeb上に表れるとの仮定に基づき，エンティティペアをクラスタリングすることで関係情報を抽出することを行う．抽出された関係情報は社会ネットワーク，セマンティックWebにおけるメタデータの自動生成などへの応用が考えられる．

2 関係情報の抽出

本研究で提案するWebからのエンティティ間の関係情報の抽出は,「Web上に出現する文脈が類似しているエンティティのペアは類似した関係を持っている」という仮説に基づいている．文脈の類似性が意味的な類似性に寄与するという同様の仮説は従来研究においても指摘されいている [3]. この仮説に基づいて，類似した文脈で表れるエンティティのペアをまとめ，同じ関係を持つエンティティペアが共通して持つ重要語を関係を表す情報として抽出するというのが提案手法の基本的なアイデアである．この時，個別のエンティティのペアを対象に処理を行うのではなく，ペアの集合を扱うことにより得られる大局的な情報を用いる点が提案手法の重要な点である．

本研究で提案するWebからの関係情報抽出の手順は以下の通りである．

1. エンティティペアの集合を取得
2. 各エンティティペアの文脈モデルを取得
3. エンティティペア間の文脈モデルの類似度を計算
4. 類似度に基づきエンティティペアをクラスタリング
5. 各クラスタから関係情報となるラベルを抽出

図1は提案手法の手順を図示したものである．まず，関係抽出の対象とするエンティティのペア集合を取得する．例えば，人物 (PERSON) と組織 (ORGANIZATION) や人物 (PERSON) と地名 (GPE) などのエンティティのペアである．次に各エンティティペアを検索エンジンのクエリーとして検索をおこない，エンティティペアを含むWebページを取得する．取得したWebページの中でエンティティペアの出現する周囲の語を用いて，ペアの文脈ベクトルを作成する．各エンティティペアについて文脈ベクトルを

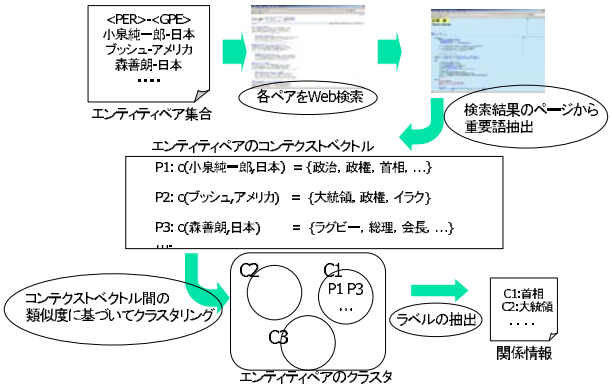

図 1: Web からの関係情報抽出手法の概要図

作成し，文脈ベクトル間の類似度に基づいてクラスタリングを行う．クラスタリングの結果生成された各クラスターからラベルを抽出し，最終的にそのラベルをクラスターに属するエンティティペアの関係情報とする．先の仮説に基づけば，類似した文脈で表れるエンティティのペアは同一のクラスターに属し，そのクラスターの各ペアは同様の関係を持っているはずである．

3 実験

提案手法を用いて，関係情報抽出の実験を行った．実験にはエンティティペアとして人物 (PERSON) と地名 (GPE) を対象とし，特に政治家と地名のペアを使用した．エンティティペアの総数 143 である．各ペアに対して正解データとして関係のラベル付けを行った．ここで関係は首相や議員などのように地名に対する人物の政治的な立場，役割を表すものであり，その内訳は首相が 22，大統領が 17，知事が 47，市長が 13，議員が 44 ペアであった．

各エンティティペアを Web 検索[1]し，上位 100 件の Web ページを用いて文脈ベクトルを作成した．各 Web ページの中でエンティティペアが 10 語以内表れる箇所を対象にしてエンティティペアに挟まれるすべての語とエンティティの前後 10 語を用いて文脈ベクトルを作成した．

生成するクラスターの数を 6 つとした時に，各クラスターから抽出した関係情報を表 1 に示す．

表 2 に生成されたクラスターの Precision と Recall の値を示す．Precision については非常に高い値が出

[1]検索エンジンには google を使用した

表 1: エンティティペアのクラスターから抽出した関係情報

ラベル (手動判別)	クラスターから抽出した関係情報
市長	市長　県知事　市　知事　区
大統領	大統領　政権　首相　知事　選挙
首相	首相　総理　議員　大統領　選挙
知事	県知事　知事　県　市長　市
議員	区　議員　自民　衆院　比例
その他	区 議員　県　県知事　知事

表 2: クラスターの Precision と Recall

Precision	Recall
92.30%	76.44%

ているが，知事クラスタにおける市長のエンティティペアや議員と知事の混同などが観察された．現状では，文脈ベクトルを作成する際にペア周辺の限られた語を用いている．また，エンティティペアと語の関連が考慮されていない．今後は，Web ページの構造情報も考慮して文脈モデルを構成する語を選択すると伴に，共起情報を用いたスコアリングを行うことによって文脈モデルの表現が向上させる必要がある．またエンティティペア間の類似尺度の計算においては文脈ベクトル間の意味的な類似性を考慮する必要もある．

4 まとめ

本研究では，Web からの関係情報の抽出手法を提案した．出現する文脈が類似しているエンティティペアは類似した関係を持っているという提案手法の基本的な考え方である．この仮定の基づき文脈間の類似性によりエンティティペアのクラスタリングを行い，クラスターからエンティティ間の関係をラベルとして抽出するのが提案手法の特徴である．

参考文献

[1] G. A. Miller and W. G. Charles. Contextual correlates of semantic similarity. 6(1):1–28, 1991.

[2] 森 純一郎, 松尾 豊, and 石塚 満. Web からの人物に関するキーワード抽出. 人工知能学会誌, 20(5):337–345, 2005.

[3] 松尾 豊, 友部 博教, 橋田 浩一, 中島 秀之, and 石塚 満. Web 上の情報からの人間関係ネットワークの抽出. 人工知能学会誌, 20(1):46–56, 2005.