1A3-4

Webからのエンティティ間の関係情報の抽出

Extracting Relationships among Named Entities from the Web

森 純一郎[*1] 辻下 卓見[*1] 松尾 豊[*2] 石塚 満[*1]
Junichiro Mori Takumi Tsujishita Yutaka Matsuo Mitsuru Ishizuka

[*1]東京大学 The University of Tokyo
[*2]産業技術総合研究所 National Institute of Advanced Industrial Science and Technology

With the currently huge amount of information on the Web, Web mining methods that obtain useful information and structures from the Web have been gained interest. We propose a novel Web mining method that automatically extracts relational information among named entities from the Web. The basic idea is to cluster similar pairs of named entities based on their contextual similarity on a Web document. Relational information among named entityes is obtained from the result of clustering process. Our experiments conducting on entity pairs of politicians and places achives clustering of the entity pairs with high recall and precision, and find appropriate relational information among the entities.

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．基本的な考え方は，検索エンジンにおけるヒット件数や検索結果のページを用いてある語やフレーズがどの程度Web上で用いられているかの統計情報を取得し有用な情報を抽出するというものである．検索エンジンを用いたWebマイニング手法は，Web全体を巨大なコーパスと見なした言語処理であり，Webマイニングのみならず自然言語処理やセマンティックWebなどさまざまな分野から多様な応用が研究されてきている．

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 [Cimiano 05, Etzioni 04]．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から抽出する研究も行われている．松尾らは，氏名のWeb上における共起情報から研究者間の関係や企業間の関係をWebから自動的に抽出する手法を提案している [松尾 05]．近年の社会ネットワークへの関心の増大から，Web上における共起情報を用いてエンティティ同士の関係を抽出する手法は大きく着目されており，他にもさまざまな研究がなされてきている [Mika 05, Culotta 04, 森 05]．

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．松尾らは研究者間の関係を抽出する際に，その関係が共著，同所属など研究上でどのような関係にあるのかを判別している．企業間の関係抽出において金らは提携や訴訟などの関係を同定している．このように，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，人と人といった，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，さらに情報検索や質問応答などへの応用が考えられる．

以下，2章ではWebからの関係情報抽出の手法を述べる．3章では実験について述べ，4章では評価を行う．最後に5章においてまとめを行う．

連絡先: 森 純一郎、東京大学大学院情報理工学研究科，東京都文京区本郷 7-3-1，03-5841-6755，jmori@mi.ci.i.u-tokyo.ac.jp

2. Webからの関係情報の抽出

関係情報の抽出は情報抽出タスクの一つとして，MUC(Message Understanding Conference) における Template Relation Task や ACE(Automatic Content Extraction) meetings における Relation Detection and Characterization などで扱われてきた．これらのタスクで対象とする関係とは人物や組織などのエンティティ[*1]間における所属，役割，位置，Part-Whole， 社会的関係を指すものであり，例えば，ACE の関係抽出タスクにおいては場所の関係を表す located, near, part-whole や社会的関係を表す business, family や雇用関係を表す executive, staff などの関係が定義されている．例えば「日本の小泉純一郎首相は...」という記述に対しては，PERSON エンティティである「小泉純一郎」と GPE エンティティ[*2]である「日本」との関係である「首相」がエンティティ間の関係を表すことになる．

2.1 提案手法のアイデア

あるエンティティとエンティティの関係情報をWebからどのように抽出できるだろうか．ここで，例として，政治家(PERSON) と地名 (GRE) という二つのエンティティ間にある関係を抽出することを考えてみよう．政治家と地名での間に

*1 ACE における固有表現抽出タスクでは Person, Organization, Facility, Location, GPE, Vehicle, Weapon のエンティティが定義されている．

*2 GPE は Geo political entity であり地名を政治的な意味として用いられるものである．

表 1: ” 小泉純一郎 AND 日本”，” 森善朗 AND 日本”，” 小泉純一郎 AND 神奈川”，” 森善朗 AND 石川” の各検索結果の上位ページから $tfidf$ によって得られた重要語

クエリー	$tfidf$ によって抽出された重要語
小泉純一郎 AND 日本	病理 藤原 **首相** 小泉 光文社 政治 宰相 参拝 ページ, **総理** バックス 野郎, 商品 **内閣** 国民 改革 **大臣** ワルシャワ アメリカ 靖国 靖国神社 再生 社会
森善朗 AND 日本	ラグビー **首相** 会長 招致 **大臣** 協会 科学政権 **総理** 館長 サッカー アフリカ 世界 宇宙 競技 ページ スポーツ 失言 関連 メディア 毛利 弘之 敦子 **内閣** 理事
小泉純一郎 AND 神奈川	**選挙** **首相** 横須賀 **候補** つよし **議員** 斉藤 **自民党** 三浦 小泉 民主党 衆議院 ページ 関連強敵 政治 公認 自由民主党 一家 総裁 出馬 補選 地元 同志
森善朗 AND 石川	一川 保夫 **首相** **選挙** 自民 奥田 **候補** 小松 **議員** **自民党** 祐士 能美 加賀 金沢 西村 ページ 新進 回答 松任 公明党 当選 委員 民主 政治 県議 比例 支配 衆院 開票

は，その地名の出身，選挙区から選出された，その地名の首長，元首などさまざまな関係が存在する．これらの関係はエンティティとともに Web 上の文書に表れているはずである．エンティティ間の関係を情報抽出する際の単純な方法として，エンティティのペアが Web ページ上で表れる箇所を調べて，関係を表すような情報を見つけるといったものが考えられる．

表 1 は，” 小泉純一郎 AND 日本”，” 森善朗 AND 日本”，” 小泉純一郎 AND 神奈川”，” 森善朗 AND 石川” という 4 つの検索クエリーそれぞれに対して得られた検索結果の上位ページの文書から検索クエリーの近傍に出現する重要語を抽出した結果である．ここで，得られた重要語群は，検索クエリーである政治家と地名のエンティティペアが出現する文脈を bag of words で単純に表したものとみなせる．なお，重要語の抽出には $tfidf$ を用いてスコアリングを行った．

” 小泉純一郎 AND 日本” と” 森善朗 AND 日本” という検索クエリーは，どちらも” 首相” または” 総理大臣” といった関係を含む政治家と地名のエンティティペアであるが，それぞれのエンティティペアに対して得られた重要語の中で共通している語を見てみると” 首相”，” 総理”，” 内閣” といった関係を表す語が共通して含まれていることがわかる．また，” 小泉純一郎 AND 神奈川” と” 森善朗 AND 石川” は，どちらも政治家と選挙区という関連を持つエンティティペアであるが，こちらも重要語として” 選挙”，” 首相”，” 候補”，” 議員” といった関係を表すような語が共通して含まれている．一方，” 小泉純一郎 AND 日本” と” 小泉純一郎 AND 神奈川” に対するそれぞれの重要語を見比べると，同一人物であるにもかかわらず” 日本” と” 神奈川” という，クエリーにおける地名の違いにより，異なる重要語が表れていることがわかる．

以上のことから，Web からのエンティティ間のの関係情報の抽出において，「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」という仮説を考えることができる．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 [Miller 91]．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．

以下では，このアイデアに基づく Web からの関係情報抽出の手法について具体的に述べていく．

2.2 提案手法の詳細

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．

以下では，各手順の詳細について説明を行う．

2.3 エンティティペア集合の取得

関係抽出の対象とするエンティティのペアの集合は人物と組織，人物と地名のような同一種類のペアの集合とする．これは提案手法のアイデアに基づき，同一種類のエンティティペア集合から，同様の関係を持つエンティティペアをまとめていき関係を取り出すという処理を行うためである．そのためには，まず対象とするエンティティの判別を行う必要がある．

エンティティの判別・抽出は，我々は行ってきた Web からの人物に関するキーワード抽出 [森 05] と固有表現抽出により行う．文書集合から人物，地名，組織などのエンティティを抽出し，関係抽出の対象とする人物と組織，人物と地名，人物と人物，組織と組織などといったエンティティのペアの集合を事前に作成しておく．

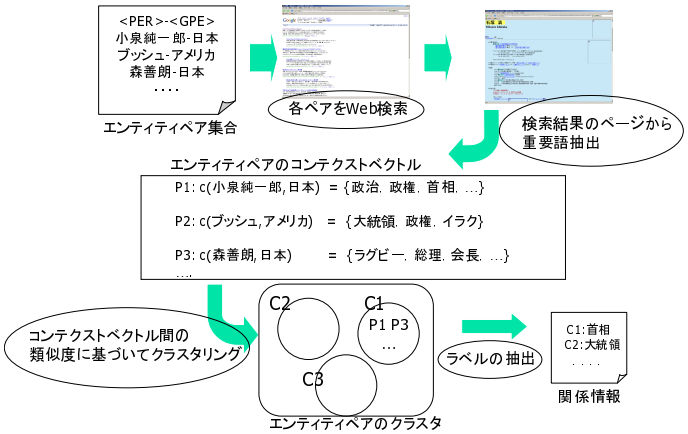

図 1: Web からの関係情報抽出手法の概要図

2.4 エンティティペアモデルの取得

エンティティペアを同士を類似度に基づきクラスタリングするために，エンティティペア集合の各ペアが Web 上に出現する文脈を何らかのモデルで表現する必要がある．提案手法では，エンティティペアがある距離内で共起している時に，その間の語およびエンティティの前後の語を用いてエンティティペアの文脈を表現する．

エンティティペアモデルを作成するために，エンティティ集合の各ペアを検索エンジンのクエリーとして検索を行う．例えば，人物と組織のエンティティペアを対象とする場合は，人物名と組織名を AND 検索する．この時，検索ヒット件数が，あらかじめ定めた閾値以下のエンティティペアは Web 上において出現が少ないと見なして処理から除くようにする．

検索結果から上位の Web ページを取得し，各 Web ページからエンティティペアがある語数以内で共起する箇所を抽出し，エンティティの間の語およびエンティティの前後の語をエンティティペアの文脈として取得する．語の品詞としては名詞，未知語を使用する．また，ストップワードとして低頻度および高頻度の語は除くようにする．

エンティティペア $e1-e2$ を含む文脈から取得した各語 t に対して，$tfidf$ (Term Frequency-Inverse Document Frequency) を用いて次の式で重み付けを行う．

$$tfidf(t) = tf(t) \cdot idf(t)$$

ここで，$tf(t)$ は $e1-e2$ を含むすべての文脈における語 t の出現頻度，$idf(t)$ は全エンティティペアの文脈ベクトルの作成に用いた Web ページの内，どれぐらいの Web ページに語 t が出現するかの尺度である．以上により，エンティティペアモデルは，ペアをなすエンティティ$e1$,$e2$ の文脈ベクトル $C(e1,e2)$ として以下のように表される．

$$C(e1,e2) = \{t_1, t_2, ..., t_k, ...\}$$

ここで，文脈ベクトルの要素となる語 t_k は $tfidf(t_k)$ により重み付けがなされている．

2.5 エンティティペア間のクラスタリングとラベル抽出

各エンティティペアの文脈ベクトルを用いて，ペアのクラスタリングを行う．クラスタリングを行う際のエンティティペア間の類似度は文脈ベクトル C_i 同士の内積

$$cos(C_i, C_j) = \frac{C_i C_j}{|C_i||C_j|}$$

によって求める．クラスタリングの手法は，生成すべきクラスター数が事前にわからないため階層化クラスタリングを用いる．

エンティティペアをクラスタリング後，クラスター内のエンティティペアの文脈ベクトルに多く含まれているような語を，そのクラスターのラベルとして抽出する．その際に，文脈ベクトル作成の時と同様に $tfidf$ を用いてラベルの重要度を決定する．ただし，この時の tf はクラスターにおける語の出現頻度を用いる．これは各クラスターに特徴的な語を，そのクラスターに属するエンティティペアの関係情報となるラベルとして抽出していることに相当する．

3. 実験

提案手法を用いて，実際に Web から関係情報の抽出実験を行った．実験に用いたエンティティペアは，人物 (PERSON) と地名 (GPE) を対象とした．特に政治家と関連する地名のペアを使用し，” 首相” や” 議員” のように地名に対して人物が政治的な立場，役割，関わりをもったエンティティペアのデータを作成した．対象としたエンティティペアの総数は 143 であり，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．その関係の内訳は首相が 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*3 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 ページからエンティティペアが含まれる文脈を取得する際に，2 つのパラメータを用意した．一つはエンティティ間の語数 n，もう一つはエンティティの前後の語数 m である．文脈ベクトルの作成において，エンティティペアが語数 n 以内で共起する箇所を対象に，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の m 語を用いた．2 章において示した表 1 は，実験で用いたエンティティペアとその文脈ベクトル要素である語の例を示している．

クラスタリングには最長距離法を用い，生成するクラスターの数は事前に付与した正解ラベルの数である 5 つとした．表 2 は，n を 30，m を 10 とした時に各クラスターから抽出されたラベルを示しており，関係情報として重要度の高い順に左から並んでいる．

4. 評価

まず，クラスタリングの結果について Precision と Recall を用いた評価を行う．生成された各クラスタごとに，手動判別したクラスターのラベルと一致するクラスター内のエンティティペアを正解とし，クラスタ cl における正解ペアの数を $N_{correct,cl}$，不正解であったペアの数を $N_{incorrect,cl}$ とする．また，関係 r ついて正解であったペアの数を $N_{correct,r}$，関係 r の正解ラベルがついてるペアの数を $N_{true,r}$ とする．この時，クラスタリングの結果の Precision(P) と Recall(R) を以下のように求める．

$$P = \Sigma_{cl} \frac{N_{correct,cl}}{N_{correctmcl} + N_{incorrect,cl}}, R = \Sigma_r \frac{N_{correct,r}}{N_{true,r}}$$

また，P と R から F 尺度も同時に求める．

図 2 はエンティティペアの文脈ベクトル作成時に用いる語数のパラメータである n と m を変化させた時のクラスタリング結果の F 尺度を表している．エンティティペア間の語数の

*3 検索エンジンには google を使用した

表 2: エンティティペアのクラスターから抽出した関係情報

クラスター	ラベル (手動判別)	クラスターから抽出した関係情報
1	市長	市長　市民 開催　会長　事務　局員　つぶやき　回答
2	大統領	大統領　政権　世界　日本　経済 政策　戦争　主義
3	首相	首相 政権 政治 記事 選挙 総理 政府 和平
4	知事	県知事　知事　会長 県庁 委員 平成　県政 市長
5	議員	議員 選挙 自民 自民党 候補 衆議院 衆院 民主党

表 3: 文脈ベクトルに使用する語とクラスタリングの性能

文脈ベクトルに使用する語	Precision	Recall	F 尺度
エンティティペアの前後 10 語および間 30 語まで	0.992	0.995	0.994
エンティティペアの前後 5 語および間 10 語まで	0.88	0.85	0.86
エンティティペアを含むページ全体	0.76	0.677	0.716

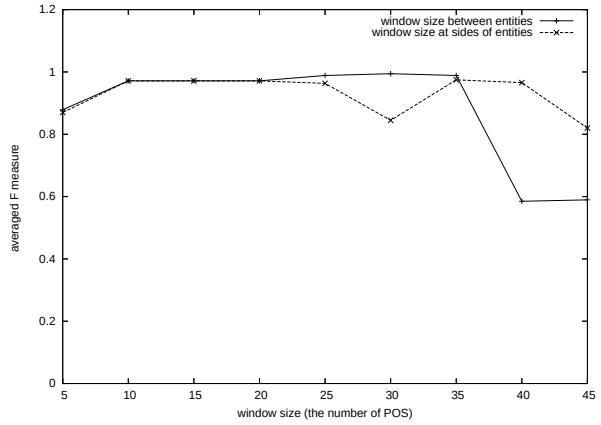

図 2: 文脈ベクトルに用いる語数とクラスターの F 尺度の関係

上限である n は 30 語，エンティティペアの両端の語数 m は 10 語とした時にもっともクラスタリング結果の F 値が高いことがわかる．この時，表 3 に示すようにクラスタリングの結果は Precision および Recall ともに 99%を示している．文脈ベクトル作成に使用する語数の幅を増減させると F 値は減少する．ページ全体を対象とした場合は，エンティティペアの文脈とあまり関係のない語が含まれるようになるためクラスタリング結果の精度は大きく落ち込む．エンティティペアの文脈を適切にモデル化するには，使用する近傍の語数が大きく影響していることがわかる．

つぎに表 2 を見ると，関係情報となる抽出されたラベルについて，各クラスターから抽出された重要度の高いラベルと正解ラベルはよく一致していることがわかる．しかし，抽出したラベルには関係情報とは関連のない一般的な語や他のクラスターのラベルが含まれているなどしているため，ラベルの重要度計算については今後改良を行う必要がある．また，ラベルの評価に関して概念距離や意味的な類似度を用いることで適切な関係情報が抽出できているかを定量的に評価を行えるようにする必要がある．

5. 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．Web 上において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．この仮定の基づいて，Web から取得したエンティティペアモデル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抽出するのが提案手法の特徴である．今後は，議論で述べたように手法の改善を行うとともに，多様なエンティティペアの種類に対して適用を行っていく．

参考文献

[Cimiano 05] Cimiano, P., Ladwig, G., and Staab, S.: Gimme' the context: Context-driven automatic semantic annotation with cpankow, in *Proc. of the 14th World Wide Web Conference* (2005)

[Culotta 04] Culotta, A., Bekkerman, R., and McCallum, A.: Extracting social networks and contact information from email and the web, in *Proc. of CEAS* (2004)

[Etzioni 04] Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A., Shaked, T., Soderland, S., Weld, D., and Yates, A.: Web-scale information extraction in KnowItAll(preliminary results, in *Proc. of the 13th World Wide Web Conference*, pp. 100–109 (2004)

[松尾 05] 松尾 豊, 友部 博教, 橋田 浩一, 中島 秀之, 石塚 満：Web 上の情報からの人間関係ネットワークの抽出, 人工知能誌, Vol. 20, No. 1, pp. 46–56 (2005)

[Mika 05] Mika, P.: Flink:Semantic web technology for the extraction and analysis of social networks, *Journal of Web Semantics*, Vol. 3, No. 2 (2005)

[Miller 91] Miller, G. A. and Charles, W. G.: Contextual correlates of semantic similarity, *Language and Cognitive Processes*, Vol. 6, No. 1, pp. 1–28 (1991)

[森 05] 森 純一郎, 松尾 豊, 石塚 満：Web からの人物に関するキーワード抽出, 人工知能誌, Vol. 20, No. 5, pp. 337–345 (2005)