著者話題モデルを用いたウェブからのエンティティ-属性推定

Modeling Entities and Attributes from the Web using the Author-Topic Model

森純一郎　　松尾豊
Junichiro Mori　Yutaka Matsuo

東京大学大学院工学系研究科総合研究機構
Institute of Engineering Innovation, School of Engineering, The University of Tokyo

Given the large amount of information available on the Web, it is important to structure the information so that a user can find desired information. We propose a method that enables to model entities and attributes from the Web. The proposed model is based on the Author-Topic model which is a generative model that reduces the generation of documents to a simple series of probabilistic steps. We focus on a person entity and collected the information about the person entities which include classes and attributes of a entity. We applied our proposed model to a data set of the person entities on a task of estimating attributes of person entities. Our results show that the proposed model is able to estimate some attributes of the entities correctly. The results also show that some classes of the entities are closely related to a particular attribute while others depend on a set of attributes.

1. はじめに

Web に我々の実世界を投影した膨大な情報が蓄積されている現在、どのように所望の情報を見つけ出し、効率よくアクセスするかは大きな課題である。このような現状のもとで、情報に意味付け、構造化を行い機械可読にするセマンティックウェブは、今後の Web の一つの方向性を示している。近年では実用的なセマンティックウェブアプリケーションとそれに付随する”Linked Data” と呼ばれる Web 情報を構造化したデータが普及してきている。セマンティックウェブではオントロジーを用いて、Web に現れた実世界の情報を記述する。例えば、セマンティックウェブでもっともよくつかわれる FOAF は人物を記述するためのオントロジーであり、人物を”name”,”interest” や関係を表す”knows” といった属性により記述する。このように実世界の対象（エンティティ）は、属性と関係という構成要素を用いて概念化されることが多い。エンティティ、属性、関係は ER モデルのようなデータベースにおけるモデリングにおいても基本的な構成要素である。セマンティックウェブはオントロジーを用いることによりトップダウンに実世界の情報をエンティティ、属性、関係という構成要素を用いて構造化している。

セマンティックウェブのアプローチとは対照に、Web 上の情報を用いてエンティティ、属性、関係という構造を抽出する研究も行われている。著者らは検索エンジンを用いて Web 上の情報から人物の属性や関係を抽出する手法を提案している [森 05, 松尾 06]。膨大な Web 上の記号情報を用いて高次の情報を抽出するこのような Web マイニングのアプローチは情報をエンティティ、属性、関係といった構成要素で構造化するボトムアップなアプローチと考えられる。

Web に我々の実世界を投影した膨大な記号情報が蓄積されている現在、既存のオントロジーには明に包含されない潜在的に構造化されうる多くの情報が Web 上には存在している。Web マイニングを用いてそれらの情報を抽出することで、Web 上の情報の構造化をさらに押し進めることができるだろう。そして、その構造化のためには、エンティティが持つ属性、エンティティが属するクラス（カテゴリ）がどのように生成されるかをモデル化する必要がある。このモデルは、Web の記号情報（一般には膨大なテキスト情報）からエンティティの属性、クラスを抽出し情報を構造化するための基本となるものである。本論文では、Web におけるエンティティの属性とクラスの生成のモデルを提案する。提案手法は、従来文書中のトピックをモデル化するため用いられてきた著者話題モデルをエンティティの属性生成モデルに適用したものであり、エンティティとして人物の情報を対象に実験によりモデルの検証を行う。本論文の成果は、セマンティックウェブにおける情報の構造に貢献するとともに、情報検索の面からはエンティティの情報検索、例えば人検索、などに応用可能である。

連絡先: 森純一郎，東京大学大学院工学系研究科総合研究機構，〒 113-8656 東京都文京区弥生 2-11-16，03-5841-1161，jmori@ipr-ctr.t.u-tokyo.ac.jp

2. エンティティと属性

2.1 ウェブにおけるエンティティと属性

エンティティとは実世界に存在するものである。DOLCE などの上位オントロジーにおいてはエンティティはすべての階層のトップに位置するものである。ウェブにおけるエンティティの扱いついて、松尾らは検索エンジンを用いた情報アクセスの文脈で、実世界の対象物を正解として捉えたときのクエリーがシステムにとってのエンティティの表現となると定義している [松尾 06]。特に、検索エンジンのクエリーの 30%は人名である [Artiles 05] ように、人物は Web 上における重要なエンティティの一つである。そのため、近年では Web における人物の同姓同名の解消、人物の属性抽出などエンティティとして人物の情報に関する研究が多く行われてきている。人物の属性抽出タスクの国際ワークショップ [関根 08] においては、“生年月日”、“職業”、“所属” などの属性が扱われている。ここで、人物を、例えば” 研究者” として具体化すると” 研究分野” や” 所属学会” などの属性が、“音楽家” であれば” 楽器” や” ジャンル” などの属性がその人物を特徴付けることになる。ある研究者をエンティティ、“研究者” を人物のクラス、“研究分野” や” 所属学会” を人物の属性、そして” ウェブマイニング” や” 人工知能学会” を属性の値とすると、Web 上において観測できるのはクエリーとしてのエンティティ（具体的な人名）と検索結果に含まれる属性値（キーワード）である。本論文の基

本的な問いは、観測されるエンティティと属性値からエンティティの属性やクラスが Web 上の情報からどのようにモデル化されうるかということである。

2.2 エンティティと属性のモデル

エンティティの属性とクラスの生成に関する本論文の基本的な考え方は、あるクラスのエンティティに集合において、エンティティを他のエンティティと区別して特徴づける性質が属性である、というものである。そして、そのような属性はエンティティの属するクラスを特徴付け、他のクラスとの区別を可能にする。

エンティティの属性とクラスのモデル化について考えるために、ここで著者話題モデル [Steyvers 04] を考える。著者話題モデルは、著者による文書の生成を確率モデルによって捉えたものであり、著書は複数の話題の組み合わせ上に、話題は複数の語の上に確率的分布として表される。著者話題モデルの確率モデルにおいて話題を属性とすると、クラスは複数の属性の上に、属性は属性値の上の確率的に分布するものである、というように著者話題モデルの上でエンティティの属性とクラスを捉えることができる。実際、新聞記事におけるエンティティのトピック抽出 [Newman 06]、人物の同姓同名解消 [Bhattacharya 06]、人物のトピック抽出 [Tang 08] など、著書話題モデルを応用したエンティティのモデルに関する研究が近年さまざま行われている。

3. 実験

本実験では、実データに対して提案モデルを適用し人物エンティティの属性推定問題を解くことによりモデルの検証を行う。まず、データとして人物エンティティの属性およびクラスの情報を Freebase [Metaweb 09] から取得する。Freebase は Wikipedia, NNDB, MusicBrainz などの Web 上の情報源からエンティティに関する情報を集約し構造化したものである。エンティティは、それらが属するクラスとそれらが持つ属性により表される。なお、Freebase のデータではエンティティをオブジェクト、クラスをカテゴリと表現しているが、本論文ではこれまでの議論を踏まえてエンティティ、クラス、属性と表記する。例えば、人物は”people/person” クラスに属し、“誕生日”、“職業”、“性別”、“国籍” などの属性を持つ。本実験では特に、Freebase において頻出する上位 30 の人物クラスに属する人物エンティティを対象し、各人物クラスからランダムに 100 個ずつ抽出した計 3000 の人物エンティティをエンティティ集合とする。エンティティ集合は計 3443 個、51 種類の属性およびそれらの値である 2385 のキーワードを含む。このエンティティと属性のデータ集合に対して、提案モデルを適用し属性の推定を行う。

対象データに提案モデルを適用し、エンティティデータ集合に含まれる属性と同数の 51 の話題を抽出した。図 1 は、エンティティが属するクラスごとに話題の分布を示したものである。クラスごとに異なる話題と関連があることがわかる。特に、“pro_athelete” や”baseball_player” は特定の話題と強く関連している。表 1 に示すように、これらの話題に関連しているキーワードは年代やポジションなどそのクラスを特徴づけるものである。一方、“artist” や”writer” は複数の話題と関連している。これらの話題に関連しているキーワードはさまざまなものが混在している。本論文では、エンティティを区別するための特徴的な性質を属性と仮定した。その点で、抽出された話題は属性に相当するものと考えれ、提案モデルによりエンティティ属性の推定がなされたと考えられる。しかしながら、

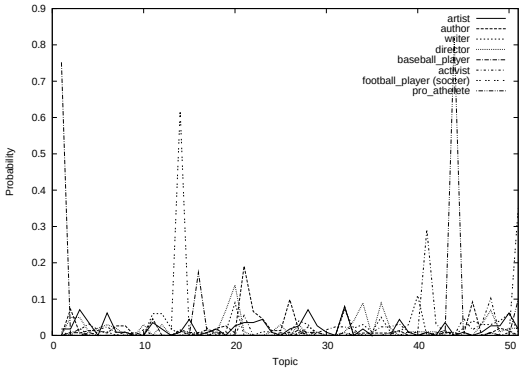

図 1: エンティティが属するクラスごとの話題分布.

表 1: 話題の上位関連キーワード.

クラス	Topic	キーワード
pro_athelete	44	2007,2003,1998
baseball_player	1,16	Pitcher,Catcher,Outfielder
artist	3,6,28,38	CVScooperation,Houston,
writer	11,12,20,40	GlobalGiving,Linguists,Kolkata

本実験結果は Freebase というすでに構造化された、いわば正解データに対して得られたものであり、Web 上の多様な情報に対するモデルの適用可能性を今後検討する必要がある。

4. おわりに

本論文では、Web 上の情報の構造化を目的に、著者話題モデルを用いたエンティティの属性とクラスの生成のモデルを提案した。エンティティとして特に人物を対象にしてモデルを実際のエンティティデータに適用した結果、人物エンティティの属性を推定可能であることを確認した。今後は、提案モデルを洗練させ、エンティティとエンティティの関係も考慮することで、エンティティ、属性、関係により情報が構造がされるモデルの構築を行う。また、エンティティ情報抽出や情報検索などへの応用に取り組む予定である。

参考文献

[Artiles 05] Artiles, J., Gonzalo, J., and Verdejo, F.: A testbed for people searching strategies in the WWW, in *SIGIR ’05*, pp. 569–570 (2005)

[Bhattacharya 06] Bhattacharya, I. and Getoor, L.: A Latent Dirichlet Model for Unsupervised Entity Resolution., in *SDM* (2006)

[Metaweb 09] Metaweb, : Freebase Data Dumps, http://download.freebase.com/datadumps/ (2009)

[Newman 06] Newman, D., Chemudugunta, C., and Smyth, P.: Statistical entity-topic models, in *KDD ’06*, pp. 680–686 (2006)

[Steyvers 04] Steyvers, M., Smyth, P., Rosen-Zvi, M., and Griffiths, T.: Probabilistic author-topic models for information discovery, in *KDD ’04*, pp. 306–315 (2004)

[Tang 08] Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., and Su, Z.: ArnetMiner: extraction and mining of academic social networks, in *KDD ’08* (2008)

[関根 08] 関根 聡, Artiles, J., Gonzalo, J.：人名の曖昧性解消評価型プロジェクト WePS, 言語処理学会第 14 回年次大会 (2008)

[松尾 06] 松尾 豊, 山川 宏：ネットワーク-予測性-属性生成, 人工知能学会全国大会 (2006)

[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, 人工知能学会論文誌, Vol. 20, No. 5, pp. 337–345 (2005)