1G2-1

Wikipediaマイニングによる大規模Webオントロジの実現

Towards a Huge Scale Web Ontology by Mining Wikipedia

中山浩太郎
Kotaro Nakayama

東京大学 知の構造化センター
Center for Knowledge Structuring, the Univeristy of Tokyo

Wikipedia has become an invaluable corpus for AI researchers. In particular, Web ontology construction from Wikipedia is one of the hottest topics on Wikipedia mining research area. Although the importance of this research direction is widely recognized and a number of researches have been conducted, there still remain many technical issues such as reliability of information, scalability and accuracy. In this challenge, we try to accelerate the new research area; Wikipedia Mining for Web ontology construction.

1. はじめに

Wikipedia は人工知能研究者が最近手に入れた新しいフロンティアである．Wikipedia は，幅広い分野において膨大な概念が網羅されているだけでなく，知識抽出のためのコーパスとして興味深い特徴を数多く持っている．密なリンク構造や，URL により語彙の意味が一意に特定できる点，200 以上の言語をサポートしている点，質の高いアンカーテキストなどは Wikipedia の持つ Web コーパスとしての特徴の一例である．これら Wikipedia の持つ特徴は，知識抽出のコーパスとして極めて有利に働くことが各種の研究によりここ数年で急速に解明されてきた．特に，Wikipedia は概念同士の関係度を数値化した連想シソーラスの構築に極めて有効であることが種々の研究で証明されている [Strube 06, Nakayama 07].

本チャレンジでは，連想シソーラスの研究からさらにもう一歩踏み出し，Wikipedia マイニングによる大規模 Web オントロジの自動的構築に関する研究分野の確立を目指す．本チャレンジによって大規模 Web オントロジが実現されたときの社会的インパクトは大きい．これは，Wikipedia は一般的な概念から最新の概念まで幅広く網羅しているため，WordNet などの既存オントロジの弱点を補う新しい情報基盤が実現できる可能性を持っているためである．

しかし，その一方で技術的な課題も存在する．例えば，Wikipedia に掲載されている情報の信頼性が挙げられる．Wikipedia に虚偽の情報が掲載され，社会問題になるケースが増加している．また，日に日に増加する一方の Wikipedia に対して解析処理を行う場合，スケーラビリティの高い解析手法を検討することは必要不可欠である．本チャレンジでは，これらの技術的課題を解決し，Web オントロジ構築の新しい方向性を切り開く研究分野の確立を目指す．その第一歩として，本論文では，Wikipedia からの Web オントロジ構築に関する研究を俯瞰するとともに，筆者らが提案する Web オントロジ構築の精度向上のための手法を解説し，今後の本チャレンジの方向性を示す．

連絡先: 中山浩太郎，東京大学知の構造化センター，東京都文京区本郷 7-3-1，TEL：03-5841-8800，FAX：03-5841-8917，kotaro.nakayama@acm.org

2. Semantic Web と Wikipedia

概念間の意味関係に基づいた情報検索の仕組みにより，高度な WWW を実現する取り組みである Semantic Web は，次世代の Web プラットフォームとして注目を集めている．近年では，知識モデルの表現方法，推論手法，統合手法などをはじめとする種々の研究が行われ，その結果，多種多様な Web オントロジが分散的に WWW 上に構築・公開されてきた．しかし，これら分散的に構築された Web オントロジの問題として，「マッピング情報の不足」が挙げられる．個々のオントロジ同士が相互運用可能な状態で，推論可能な Web を実現するためには，オントロジ同士がマッピングされる必要がある．そのため，ここ数年でオントロジマッピングの研究領域が活発化してきている．オントロジマッピングの手法は，1）グローバルオントロジとローカルオントロジのマッピング，2）ローカルオントロジ間のマッピング，3）オントロジ統合の三つに分類される [Choi 06].

この 3 つのアプローチの中でも，本研究ではグローバルオントロジを利用したマッピングに着目する．グローバルオントロジとは，多くの概念を保持する一種の巨大な共有オントロジであり，仲介オントロジとして機能する．小規模なローカルオントロジは，グローバルオントロジとのマッピングを記述することにより，グローバルオントロジを仲介して他のローカルオントロジと間接的にマッピングされる．グローバルオントロジを利用するアプローチは，WWW 上に共有の語彙を構築して利用するため，他の手法と比較してマッピングの作成や検索などが容易であるという利点がある．しかし，このような巨大なオントロジを構築するためには多大なコストを必要とし，メンテナンス性が悪いという欠点があるため現実的なアプローチではなかった．

しかし，Wikipedia の登場により，この状況は一転しつつある．Wikipedia では，一般的な概念から専門的な概念に至るまで新旧の幅広い概念が網羅されており，すべての言語を合わせると 1,000 万記事を超える（2008 年 4 月）ほどの非常に膨大な量のコンテンツが存在している．また，Wikipedia の一つのページ（URL）が一つの概念に対応していることや，質の良いカテゴリ構造が構築されているという特徴を利用することにより，網羅性の高い Web オントロジを構築する研究が最近行われている．

3. Wikipedia からの Web オントロジ構築

Wikipedia から Web オントロジなどの構造化された概念関係を構築する研究としては，DBPedia[Auer 07] と YAGO[Suchanek 07] などが挙げられる．

DBPedia では，主に定型化されたテーブル情報（Infobox）の部分を解析することにより，人や音楽の属性を抽出し，RDF 化することで，意味検索を可能にしている．DBPedia には，いくつかのインタフェースが存在するが，SPARQL での検索も可能である．例えば，「ベルリンで 1900 年以降に生まれた人を検索する」といったクエリを SPARQL で記述し，DBPedia に渡すことにより，条件に適合する人物だけの情報を抽出可能である．DBPedia では，80,000 以上の人物，70,000 以上の地理情報，35,000 以上の音楽検索などが可能である．

YAGO では，WordNet の概念（Synset）と Wikipedia のページをマッチングすることにより，WordNet と Wikipedia の網羅性を持った巨大な Web オントロジを構築することを目指している．Wikipedia のカテゴリは一種のタクソノミであり，ページとカテゴリ間の関係を上位・下位関係としてリンク（カテゴリリンク）で定義している．そのため，カテゴリ構造をそのまま利用した場合，単なる包含関係に基づく関係抽出が主体となる．

4. 重要文解析による精度向上

本章では，筆者らが提案している Web オントロジの構築の精度向上に関する一手法を解説する [中山 08]．本手法では，Wikipedia の記事内に記述されているテキスト部分を解析し，概念同士の関係性を発見する．本研究と同様，テキストを構文解析するアプローチを採用する先行研究として，Dat ら [Nguyen 07] の研究が挙げられる．しかし，現在の統計的な自然言語解析に基づく意味関係抽出の手法では，文章に書かれている内容を理解しながら解析しているわけではないため，曖昧性の高い単語の処理や代名詞の意味の同定などが困難であるという問題があった．そのため，本研究では，記事内から重要な文だけを選択的に解析することで，精度向上を目指す．これは，記事にとって重要な文は，記事の主題と深く関係するため，すべての文を解析する場合に比べ，主題に意味的に関係の深い概念が抽出できる可能性があるという仮定に基づく手法である．

重要文の抽出プロセスでは，Wikipedia の記事から重要文を抽出するために LSP（Lead Sentence Parsing）法と ISP（Important Sentence Parsing）法の二つの手法を提案する．以下，二つの手法について詳述する．

4.1 LSP（Lead Sentence Parsing）法

LSP 法は，記事のリード部分（冒頭文）を重要文と見做して解析する手法である．これは，Wikipedia の各記事において，リード部分が多くの場合に他の概念との明確な意味関係を定義した文であることを利用した手法である．特に，Wikipedia におけるリード部分は，他の概念に対する is-a 関係が豊富に定義されていることがこれまでの調査によって判明している．リード部分に関する統計情報（2006 年 9 月のデータ）を表 1 に示す．

Wikipedia 全体では，約 158 万のページ（リダイレクトページとカテゴリページを除く）が存在するが，情報の信頼性が低い「ノイズページ」を除くためにバックワードリンク数が 100 以下のページを削除したところ，約 6 万 5 千ページを抽出できた．バックワードリンク数に応じてノイズページを除外する方法は，Gabrilovich らの研究 [Gabrilovich 07] でその有効性が証明されている．次に，リード部分が「is-a」関係（is/are/was/were）を定義した文であるかを解析したところ，実に 86.3% (P_b/P_a) ものページが「is-a」関係を定義したページであることが判明した．さらに，95.7 % (P_c/P_a) のページは，他のページに対するリンクがリード部分に存在していた．そして，85.5 % ($(P_b \cap P_c)/P_a$) のページは，リード部分に「is-a」関係と他のページに対するリンクを保持していることが判明した．この統計情報は，Wikipedia の各ページにおいて，リード部分は，他の概念に対しての「is-a」関係を抽出するために有用な情報を含んでいる可能性が高いことを示している．

表 1: Wikipedia のリード部分に関する統計

# of concept pages (exc. redirect and category pages)	1,580,397
# of pages having more than 100 backward links: P_a	65,391
# of pages (in P_a) begin with is-a definition sentence: P_b	56,438
# of pages (in P_a) that the 1st sentence has links: P_c	62,642
# of $P_b \cap P_c$	56,411

4.2 ISP（Important Sentence Parsing）法

ISP 法は，記事の中から重要な文章を抽出して解析する手法である．ここで，重要な文章とは，その記事の中で重要なリンクや単語を含む文章のことである．重要なリンクや単語を抽出する方法には，リンクの共起性解析や TF-IDF などの手法が利用可能であるが，本研究では，筆者らが提案する連想シソーラスの構築手法 $pfibf$[Nakayama 07] を利用する．$pfibf$ は，高い精度を実現しつつ，スケーラビリティを考慮した連想シソーラスの構築手法であり，特定の記事の中に含まれる重要なリンクを抽出することが可能である．以下に本手法の詳細を説明する．

$pfibf$ は，リンク構造解析手法であり，グラフ $G = \{V, E\}$（V はページの集合，E はリンクの集合）内において n ホップ以内のノード同士の関係性を数値化することを目的としている．ここで，Wikipedia では一つの記事（ページ）が一つの概念に対応するため，二つの記事間の関係性を抽出することは，二つの概念間の関係性を抽出することと同義である．二つの記事間（v_i，v_j）の関係の強さを計測する問題を考えた場合，関係の強さは以下の二つの要素に依存すると考えられる．

- 記事 v_i から記事 v_j へのパスの多さ
- 記事 v_i から記事 v_j への最短距離

ここで，パスとはハイパーリンクを伝って到達できるページ間の経路のことである．つまり，記事 v_i から記事 v_j へのパスが多ければ多いほど（共通のリンク先や共通の参照元が多いほど），記事間の関係性は強く，またそのパスの長さが短ければ短いほど強く関係すると考えられる．v_i から v_j への n ホップ先の全経路 $T = \{t_1, t_2, ..., t_n\}$ が与えられたとき，記事 v_i から記事 v_j の関係性 $pfibf$（Path Frequency Inversed Backward link Frequency）を以下の式により表現する．

$$pfibf(v_i, v_j) = \sum_{k=1}^{n} \frac{1}{d(|t_k|)} \cdot \log \frac{N}{bf(v_j)}. \qquad (1)$$

表 2: $pfibf$ によって抽出された連想関係の例

Query	Extracted association terms		
Sports	Basketball	Baseball	Volleyball
Microsoft	MS Windows	OS	MS Office
Apple Inc.	Macintosh	Mac OS X	iPod
iPod	Apple Inc.	iPod mini	iTunes
Book	Library	Diamond Sutra	Printing
Google	Search engine	PageRank	Google search
Horse	Rodeo	Cowboy	Horse-racing
Film	Actor	Television	United States
DNA	RNA	Protein	Genetics
Canada	Ontario	Quebec	Toronto

d は経路 t_k の経路長に応じて増加する関数であり，単調増加関数を利用する．N は全記事数，$bf(v_j)$ は記事 v_j が持つ他の記事からのリンク数とする．つまり，$pfibf$ は多くのリンク先を共有するが，他の記事とはリンク先を共有しない記事により高い値を示す．また，同じ距離（例えば距離 1，直接リンク関係にある）の記事であっても，より多くリンク先を共有する記事に対して高い値を示す．$pfibf$ で得られた連想シソーラスの例を表 2 に示す．

ISP 法では，$pfibf$ で得られた概念間の連想関係を利用して，ページ（概念）の中で重要なリンクと単語を含むセンテンスを抽出し，解析対象とした．

4.3 構文解析

構文解析のステップでは，上述のステップで選択された重要文に対して構文解析を行った．構文解析のツールとしては，確率的構文解析法を採用している，Stanford Parser[Klein 03] を利用した．Stanford Parser は，与えられたセンテンスの構文を解析し，POS タグが付与された構文木を生成する．例えば，以下のようなセンテンスからは，

```
Lutz_D._Schmadel is [[Germany|German]] [[astronomer]]
```

以下のような構文木が生成される．

```
(S (NP (NN Lutz_D._Schmadel)
    (VP (VBZ is)
     (NP (NN [[Germany|German]]) (NN [[astronomer]])
))))
```

提案手法では，生成された構文木から「(NP ...) (VP (VBZ/VBD/VBP ...) (NP ...))」パターンを抽出し，最初の NP を主語，二つ目の NP を目的語，VP を述語とする意味的三つ組みを抽出する．例えば，上述の「Lutz_D._Schmadel」に関するセンテンスでは，二つ目の NP は二つの NN から構成され，そのどちらもリンクを保有する．一つ目の NN はページ「Germany」へのリンクであり，もう一つはページ「astronomer」へのリンクである．この場合，最後の NN である「astronomer」が NP の語幹であるため，NP 全体を「astronomer」で置き換える．そして，最後に「Lutz_D._Schmadel」「is」「astronomer」を意味的三つ組みとして抽出する．

4.4 解析結果

表 3 に ISP 法と LSP 法によって抽出された意味関係の例を示す．また，解析精度を計測したところ，単にすべての文章を解析するよりも，ISP や LSP を利用することで，構文解析の精度が向上することを確認した．特に，LSP は他の手法より大幅に高い適合率を示したことから，リード文が精度の高い意味関係を抽出するための有用なリソースであることが実証されている [中山 08]．

表 3: ISP 法と LSP 法によって抽出された意味関係の例

Subject	Predicate	Object
Apple	is	Fruit
Bird	is	Homeothermic
Bird	is	Biped
Cat	is	Mammal
Computer	is	Machine
Isola_d'Asti	is	Comune
Karwasra	is	Gotra
Nava_de_Francia	is	municipality
Sharon_Stone	is	Model
Sharon_Stone	is	Film_producer
Al Capone	was	gangster

5. 近未来チャレンジ

セマンティック Web の研究領域では，前章で紹介した研究をはじめとし，共通の概念大系としての Wikipedia の可能性が示されている．しかし，その一方で情報の信頼性やスケーラビリティなど研究者が解決すべき課題（チャレンジ）も数多く残っている．

5.1 情報の信頼性

Wikipedia は，Web ブラウザを利用して誰でも更新が可能であるため，間違いが迅速に修正され，その結果信頼性が高いコンテンツが実現できていると主張するユーザは多い．これを裏付けるように，2005 年 12 月に公開された英 Nature 誌の調査によれば，Wikipedia は世界最大の百科事典の Britannica と同等の規模と精度を持つと報告されている [Giles 05]．しかし，最近では各種の虚偽の書き込み事件をきっかけにその信頼性に対する問題点が明らかにされてきた．

しかし，Wikipedia の情報の信頼性を数値化する手法としては，編集者の編集履歴や，編集コンテンツの生存時間などを元に編集者の質を推測し，利用することが可能である [Hu 07]．また，他にも記事のバックワードリンク数が記事の信頼性を示す指標として利用できる [Nakayama 07] ほか，編集者同士のソーシャルネットワークや，ディスカッション状況，閲覧回数など記事の信頼性に影響を与えると考えられるパラメーターは数多く存在する．そのため，これらの指標を統合的・学習するような方法は，重要なチャレンジとなると考えられる．

5.2 解析手法のスケーラビリティ

これまでに公開されているほとんど全ての Wikipedia に関する研究において，スケーラビリティの問題に関する議論は行われてこなかったが，Wikipedia のように日に日に増加するデータを解析する上では，極めて重要な技術的課題である．増加の一途を辿る Wikipedia のリンク構造を解析するためには，総記事数 n に対して計算量が線形に増加するような手法（$O(n)$）か，最低でも $O(n \cdot logn)$ 程度の計算量で解析が可能なアルゴリズムが必要だといえる．

また，記事の更新情報をインクリメンタルに解析する手法も有効だと考えられる．インクリメンタルな解析手法とは，過去の解析結果に対してデータの更新箇所のみ解析し，必要な箇所のみを変更するアプローチである．

並列処理に適しているアルゴリズムの検討も重要だと考えられる．複数台の計算機を利用して効率よくリンク構造を解析するためには，アルゴリズムやデータモデルも最適化する必要がある．例えば，Web サイトにおけるリンク構造はしばしば

隣接行列によって表現されるが，Wikipedia では記事数が 150 万記事存在することを考えると，150 万行× 150 万列の膨大な行列が必要となり，通常の計算機リソースではデータを保持することすら不可能である．このように問題を解決するためには，データの圧縮方式や解析手法などの検討が重要となる．筆者らの研究では，Wikipedia のリンク構造を隣接行列によって表現し，二重に二分木を利用して高速に行列積の計算を行うアルゴリズムを提案している [Nakayama 07]．計算上の工夫をすることで，計算量を抑え，並列処理を可能にすることが Wikipedia の解析では重要な技術的課題である．

5.3 アプリケーション

構築した Web オントロジの精度や網羅性は，アプリケーションやタスクによって評価指標が異なるため，今後の Wikipedia マイニングの研究では，アプリケーションへの適用が重要なチャレンジとなる．例えば，実際に意味関係を考慮した情報検索や質疑応答，文書要約などのアプリケーションに適用することで，用途に応じた最適化や精度の評価が可能となる．

興味深いタスクとして，文章中の単語の意味の曖昧性解消に Wikipedia のデータを利用する研究が行われている [Bunescu 06]．この研究では，Wikipedia では一つのページが一つの概念に対応することを利用し，曖昧性のある単語リストを抽出したのち，周辺のリンクやコンテンツ内のテキストを利用することで，コンテキストに応じた曖昧性解消を行う．

このように，タスクやアプリケーションを定めて解析結果を適用していくことで，さらに実用性の高い Web オントロジの構築が可能になると考えられる．また，WordNet や OpenCYC などの既存オントロジとの融合やローカルオントロジの統合なども大規模かつ完成度の高いオントロジを構築する上で欠かせないチャレンジである．

6. まとめ

本論文では，次世代 WWW であるセマンティックを実現するための一つのアプローチとして，Wikipedia からの大規模な Web オントロジの構築手法について俯瞰した．本チャレンジに関する情報やシステムは，以下の URL からアクセス可能である．

- Wikipedia Lab.
  http://wikipedia-lab.org
- Wikipedia Thesaurus
  http://wikipedia-lab.org:8080/WikipediaThesaurusV2
- Wikipedia Ontology
  http://wikipedia-lab.org:8080/WikipediaOntology

Wikipedia からの Web オントロジ構築は，大きなポテンシャルを持っている一方で，情報の信頼性確保など研究者が解決すべき研究課題は未だ多く残されている．この近未来チャレンジでは，これらの問題をひとつずつ解決し，幅広いアプリケーションに適用される実用的な大規模 Web オントロジを実現することを目指す．

謝辞：本研究を推進するにあたって，東京大学松尾豊准教授，大阪大学原隆浩准教授，大阪大学西尾章治郎副学長理事に大きなご指導・ご助言をいただいている．また，本研究の一部は，マイクロソフト産学連携研究機構 CORE 連携研究プロジェクトの助成によるものである．ここに記して謝意を表す．

参考文献

[Auer 07] Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. G.: DBpedia: A Nucleus for a Web of Open Data, in *International Semantic Web Conference (ISWC2007)*, pp. 722–735 (2007)

[Bunescu 06] Bunescu, R. C. and Pasca, M.: Using Encyclopedic Knowledge for Named entity Disambiguation, in *Proc. of Conference of the European Chapter of the Association for Computational Linguistics (EACL2006)* (2006)

[Choi 06] Choi, N., Song, I.-Y., and Han, H.: A survey on ontology mapping, *SIGMOD Rec.*, Vol. 35, No. 3, pp. 34–41 (2006)

[Gabrilovich 07] Gabrilovich, E. and Markovitch, S.: Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis., in *Proc. of International Joint Conference on Artificial Intelligence (IJCAI 2007)*, pp. 1606–1611 (2007)

[Giles 05] Giles, J.: Internet Encyclopaedias Go Head to Head, *Nature*, Vol. 438, pp. 900–901 (2005)

[Hu 07] Hu, M., Lim, E.-P., Sun, A., Lauw, H. W., and Vuong, B.-Q.: Measuring article quality in wikipedia: models and evaluation, in *Proc. of ACM conference on Conference on information and knowledge management (CIKM 2007)*, pp. 243–252, New York, NY, USA (2007), ACM

[Klein 03] Klein, D. and Manning, C. D.: Accurate Unlexicalized Parsing, in *Proc. of Meeting of the Association for Computational Linguistics (ACL 2003)*, pp. 423–430 (2003)

[Nakayama 07] Nakayama, K., Hara, T., and Nishio, S.: Wikipedia Mining for An Association Web Thesaurus Construction., in *Proc. of International Conference on Web Information Systems Engineering (WISE 2007)* (2007)

[中山 08] 中山浩太郎, 原隆浩, 西尾章治郎：自然言語処理とリンク構造解析を利用した Wikipedia からの Web オントロジ自動構築に関する一手法, データ工学ワークショップ (DEWS) (2008)

[Nguyen 07] Nguyen, D. P. T., Matsuo, Y., and Ishizuka, M.: Relation Extraction from Wikipedia Using Subtree Mining, in *Proc. of National Conference on Artificial Intelligence (AAAI-07)*, pp. 1414–1420 (2007)

[Strube 06] Strube, M. and Ponzetto, S.: WikiRelate! Computing Semantic Relatedness Using Wikipedia, in *Proc. of National Conference on Artificial Intelligence (AAAI-06)*, pp. 1419–1424 (2006)

[Suchanek 07] Suchanek, F. M., Kasneci, G., and Weikum, G.: Yago: a Core of Semantic Knowledge, in *Proc. of International Conference on World Wide Web (WWW2007)*, pp. 697–706 (2007)