2I3-NFC4-1

Wikipedia シソーラス Ver. 3

Wikipedia Thesaurus Ver. 3

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 University of Tokyo

Wikipedia, a collaboratively developed online-encyclopedia, has become an invaluable resource for knowledge extraction. In our previous research works, we developed "Wikipedia Thesaurus," a huge scale association thesaurus with its own APIs. The Version 3 now supports new information named "concept class" in addition to the associational relations. We introduce the overview of the feature and construction algorithms (improvements based on Web search) with some demonstrations.

1. はじめに

大規模オンライン百科事典である「Wikipedia」は，人工知能研究，Web 研究，自然言語研究など，幅広い分野で利用される有用な知識リソースとして成長してきた．筆者らは，Wikipedia を解析することで大規模な連想シソーラス「Wikipedia シソーラス」とその API を構築・公開してきた．本稿では，最新バージョンである「Wikipedia シソーラス Ver. 3」について，概要とその構築方法を解説する．

2. Wikipedia シソーラス

Wikipedia は，幅広い分野に対する網羅性や，密なリンク構造など知識リソースとして魅力的な特徴を持つ．このことから，連想関係抽出や関係抽出などの研究に幅広く利用されてきた．筆者らは，Wikipedia を解析することで他の研究やアプリケーションに利用可能なリソースの構築・公開を進めている．その一つが「Wikipedia シソーラス」である．Wikipedia シソーラスは，Wikipedia のリンク構造を解析する pfibf [Nakayama 07] やリンク共起性解析 [Ito 08] に基づくアルゴリズムによって概念間の関係度を数値化し，連想関係を抽出した辞書である．

pfibf では，概念をノード，ハイパーリンクをエッジとしたネットワーク構造において，概念間のパスの数が多いほど，またそれらのパスが短いほど，それらの概念が強く関連していると見做す．それに加えて，ある概念が他の概念からリンクを多く張られている（バックワードリンク数が多い）ほど，その概念が一般的な概念であるとして，関連度を弱くする．

リンクの共起解析では，記事内に出現するリンクの共起性から概念の関連度を計算している．pfibf はリンク構造を繰り返し探索するという解析方法であったが，この手法では各記事を一通り解析すれば十分であるため，pfibf と同程度の精度を実現しつつも計算量が大幅に改善されている．

さらに，上記の pfibf やリンク共起情報に加え，共通カテゴリへの経路長や Web ヒット件数などの素性を加えて SVM で学習することにより，重要な素性の組み合わせを発見し，精度を向上させることが可能であることを示した [Nakayama 09]．

このようにして構築された Wikipedia シソーラスは，他研究やシステムなどで利用できるように，SOAP や JSON に基づく API も公開しており，国内外で利用されている．

連絡先: 中山浩太郎，東京大学知の構造化センター，〒 113-8656 東京都文京区本郷 7-3-1，03-5841-0462，03-5841-0454，nakayama@cks.u-tokyo.ac.jp

3. 構造化データを利用した概念分類

上述のとおり，Wikipedia シソーラスに関する研究では概念間の連想関係を抽出し，各種アプリケーションへ適用可能な大規模連想辞書を構築してきた．しかし，筆者の研究グループにおいて，連想関係抽出は研究の全体像においては第一ステップに過ぎず，より幅広い情報の取得を目指している．その一つが，「構造化データを利用した概念分類」である．概念分類とは，ある概念（単語）が与えられた時に，その概念が人物なのか，企業なのか，製品なのか，など，どのようなクラスに属するかを推定する問題である．Wikipedia では，この問題に対して利用可能な二つの強力な構造化データが存在する．カテゴリツリーとインフォボックスである．以降，二つの情報の特徴と問題を説明する．

3.1 カテゴリツリー

Wikipedia は，概念を分類するために「カテゴリツリー」と呼ばれる大規模な分類体系を持つ．ページ-カテゴリ，カテゴリ-カテゴリの間には,「カテゴリリンク」と呼ばれる所属関係を示すリンクで結ばれ，日本語で 200 万以上，英語では 1,000 万以上のカテゴリリンクが存在する（2009 年 6 月）．Wikipedia の場合，各概念ページは一つまたは複数のカテゴリに所属可能な上に，一部では祖先カテゴリが子孫カテゴリに再帰して所属するようなループも存在し，複雑なネットワーク構造を持つ．そのため，あるカテゴリに所属する概念を取得する手法として，そのカテゴリ以下の概念を再帰的に取得するという階層的な手法は適用できない [Shirakawa 09]．

3.2 インフォボックス

インフォボックスとは，各記事において属性情報を記述するためのテンプレートであり，人物に関する記事であれば「生年月日」「血液型」「身長」といった属性情報が記載される．また，都市や国に関する記事であれば,「首都」や「隣接する国」といった属性情報が記述される．インフォボックスは，構造化されたデータを高精度に入手できることから，関係抽出の研究で利用されることが多い．このインフォボックスを単純にクラスへマッピングすることで，高精度にクラスの推定が可能である．しかし，インフォボックスをクラス分類のタスクで利用するは，二つの問題がある．一つ目の問題は網羅性である．インフォボックスが用意されているページは少なく，Wikipedia

の中で約 20%程度のページしかインフォボックスが定義されていない（英語版 2009 年 6 月調べ.）．二つ目の問題は多様性である．インフォボックスは 7,000 以上の種類が存在し，日々増加しているが，これを網羅的に追従してマッピングするのには多大な労力を必要とする．

3.3 構造化データの複合利用によるクラス分類

本研究では，インフォボックスとカテゴリツリーを相互補完して利用し，クラス分類のタスクに利用する手法を提案する．本手法の主な流れは図 1 に示すとおりである．以下に各ステップについて説明する．

本手法は，まず少量のインフォボックスを正解集合として指定するところから処理をスタートする．たとえば，「Infobox_Artist」や「NFL_player」は人物クラス，「Infobox_Company」は企業，「Geobox」は地理情報といったように正例集合として指定する．

次に，上記インフォボックスを含むページをすべて抽出し，各クラスを特徴づけるカテゴリとそのスコアを解析する．このとき，お互いのクラスに含まれるデータが互いに負例となるようにした．これは，Wikipedia 全体で統計を取ると，データが疎（スパース）で，特徴量がうまく抽出できないためである．各クラスに特徴的なカテゴリを抽出するには，Robinson 法によるベイジアンフィルタを利用した．

そして，すべてのページに対して，カテゴリ情報を利用して各クラスへの所属の強さを算出すると同時に，各クラスに強く所属するページ集合 P_c から，頻出のインフォボックス I_c を抽出する．

最後に I_c を最初に人間が与えた正解集合に追加し，再度上記の処理を実行する．これを新しいインフォボックスが発見されなくなるまで実行し，最後の実行で抽出された P_c を各クラスに所属する概念として採用する．この結果，Wikipedia の全ページに対して，各クラスへの所属の強さが確信度（0.0-1.0）としてスコア付けされる．

このようにして得られたクラス情報を解析したところ，高精度にクラス分けができていることがわかった．特に，スコアが 0.8 ポイント以上のデータ（2009 年 6 月英語版において，約 100 万件）に関しては，正解率 95%以上の精度で分類ができていた．

各クラスを特徴付けるカテゴリを分析してみたところ，「XXX 年生まれ」などは人物を特定するのに強力なヒントとして利用されていることがわかった．また，企業だと「NASDAQ」などが重要な手がかりとして利用できることがわかった．

4. まとめ

本稿では，Wikipedia から抽出した大規模連想辞書「Wikipedia シソーラス」のバージョン 3 についてその構築方法を解説した．本辞書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．

- SIG-WP
  (Special Interest Group on Wikipedia)
  http://sigwp.org/
- Wikipedia Thesaurus V3
  http://dev.sigwp.org/WikipediaThesaurusV2
- Wikipedia Ontology API SOAP End Point
  http://dev.sigwp.org/WikipediaOntologyAPIv3/Service.asmx?WSDL

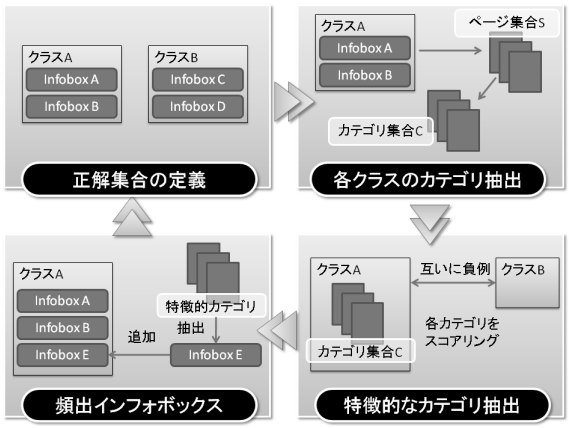

図 1: 処理フロー

本稿では紙面の関係上説明できなかったが，本研究では Web 情報との融合による精度向上も実現している．Wikipedia 単独で網羅される情報は Web 空間に比べると限界があり，間違いや虚偽の情報などが掲載・放置される場合もあるが，Wikipedia マイニングで得られた結果を Web 検索エンジンによってダブルチェックすることで信頼性の低い情報をフィルタリングできることがわかっている．今後の展開では，さらに Web 情報との融合を進め，精度だけでなく網羅性の向上も目指す．

謝辞：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および，科学研究費補助金基盤研究 C(20500093)，科学研究費補助金基盤研究 B(21300032) によるものである．ここに記して謝意を表す．

参考文献

[Ito 08] Ito, M., Nakayama, K., Hara, T., and Nishio, S.: Association Thesaurus Construction Methods based on Link Co-occurrence Analysis For Wikipedia, in *Proceedings of Conference on Information and Knowledge Management (CIKM)*, pp. 817–826 (2008)

[Nakayama 07] Nakayama, K., Hara, T., and Nishio, S.: Wikipedia Mining for an Association Web Thesaurus Construction, in *Proceedings of International Conference on Web Information Systems Engineering (WISE)*, pp. 322–334 (2007)

[Nakayama 09] Nakayama, K., Ito, M., Hara, T., and Nishio, S.: Wikipedia Relatedness Measurement Methods and Influential Features, in *Proceedings of IEEE International Symposium on Mining and Web (MAW)* (2009)

[Shirakawa 09] Shirakawa, M., Nakayama, K., Hara, T., and Nishio, S.: Concept Vector Extraction from Wikipedia Category Network, in *Proceedings of Ubiquitous Information Management and Communication (ICUIMC)*, pp. 71–79 (2009)