様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２４年６月 12 日現在

機関番号：12601
研究種目：基盤研究（B）
研究期間：2009～2011
課題番号：21300032
研究課題名（和文）　Wikipedia マイニングによる大規模 Web オントロジの構築

研究課題名（英文）　Wikipedia Mining for Constructing a Huge Scale Web Ontology

研究代表者
中山　浩太郎（NAKAYAMA KOTARO）
東京大学・知の構造化センター・特任講師
研究者番号：00512097

研究成果の概要（和文）：
Wikipedia は，「群衆の叡智」と呼ばれるほどのソーシャルメディアとして成長し，WWW・人工知能・自然言語処理・情報検索など幅広い分野の研究者の間で，重要な情報リソースとして認識されるに至った．しかし，オントロジなどの構造化された知識集合を抽出するためのリソースとしては，情報の信頼性向上やスケーラビリティ，精度などが課題であった．そこで，本研究では，これらの問題を解決するために，信頼性の数値化手法，スケーラビリティの高い解析手法，Web などの既存リソース解析との融合手法などについて研究し，手法を確立した．また，研究計画で予定していたとおり，成果報告のため積極的に論文投稿・学会発表を行うことで，成果の公表に努めた．その結果，Web Intelligence や CSCW，EMNLP といった難関会議に論文が採択され，ACM Transaction にも論文が採録された．これらは，成果報告の公表という側面では高い効果があったと確信している．

研究成果の概要（英文）：
Wikipedia, as known as wisdom of crowds, has become one of the most remarkable resources in various areas such as WWW, AI, NLP and IR. However, difficulties on handling data such as reliability of information and scalability of process still remain due to the nature of heterogeneous environments. In this research work, we have proposed a number of solutions on reliability calculation, scalability improvements and integration with other resources like the Web. Furthermore, we actively submitted research papers and they are published in a number of competitive conferences and journals.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2009 年度	3,000,000	900,000	3,900,000
2010 年度	2,500,000	750,000	3,250,000
2011 年度	2,200,000	660,000	2,860,000
年度			
年度			
総　計	7,700,000	2,310,000	10,010,000

研究分野：総合領域
科研費の分科・細目：情報学・メディア情報学・データベース
キーワード：①人工知能　②情報システム　③情報検索

１．研究開始当初の背景

Wikipedia は，Wiki をベースにした Web 百科事典であり，誰もが自由に編集可能であるという特徴から大量の情報が書き込まれ，普遍的な概念から新しい概念に至るまで，膨大な情報が網羅されている．この結果，Wikipedia は「群衆の叡智」と呼ばれるほどのソーシャルメディアとして成長し，WWW・人工知能・自然言語処理・情報検索など幅広い分野の研究者の間で，新しい知識抽出のための情報リソースとして認識されるに至った．事実，この２年間で WWW や ISWC，AAAI，ACL，SIGIR といったトップカンファレンスを中心に Wikipedia の解析に関する論文が急増している．

その中でも，現在最も研究が盛んに行われているのが，概念間の関係度（Relatedness）解析である．国外の研究としては，Strube らの研究である WikiRelate や Gabrilovich らの研究などが有名である．これらの研究では，カテゴリ情報や特徴ベクトルの比較などによって語彙同士の関係性を数値化するが，これらの研究では意味関係を抽出した機械処理可能な概念辞書を構築するものではない．Semantic Wikipedia は，Wikipedia の拡張アーキテクチャとして，リンクに意味情報を手動で付与する仕組みを提案しているが，このようにユーザに新たな労力を強いるアプローチが，コミュニティに受け入れられるかという問題には，未だ議論の余地があり，実際は小規模な概念体系しか構築できていないのが現状である．DBPedia は，Wikipedia のデータを Semantic Web のフォーマットに変換するプロジェクトであり，単純なマッチングによりインフォボックスと呼ばれる構造化されたテーブルから意味関係を抽出している．しかし，その適用範囲が限定的で高度な意味情報は抽出していない．つまり，Wikipedia の有用性には注目が集まっているものの，意味抽出の State-of-the-art と呼べる手法はいまだ確立されていない上に，他のアプリケーションに適用可能な形で成果を提供しているプロジェクトは極めて少ない．

２．研究の目的

提案者らはこれまでの研究で，Wikipedia を解析することで実用性の高い知識抽出が可能であることを証明し，実際に大規模な Web コンテンツ（特に Web 百科事典「Wikipedia」）を解析することで，大規模かつ精度の良い連想シソーラス辞書を構築してきた．この連想シソーラスは，Wikipedia のリンク構造を解析して作成したものであるが，既に英語で 300 万概念，日本語で 77 万概念をサポートしており，世界最大規模の連想シソーラスを実現している．本連想シソーラスは，与えられた単語に対して，連想関係にある単語を高精度かつ高速に抽出することが可能であり，情報検索の効率化などを実現する基盤技術として利用可能である．しかし，提案者らにとって，連想シソーラスは，今回の提案内容である大規模知識ベースを構築するための第一ステップにしか過ぎない．本提案では，従来研究までで実現した連想シソーラスからさらに飛躍し，より高度な意味関係を持つ概念辞書を Wikipedia から抽出することを目指す．しかし，Wikipedia のような大規模かつ不特定多数のユーザが編集するようなソーシャルメディアを解析して有用な知識を抽出するには，①情報の信頼性と②スケーラビリティ二つの大きな技術的課題をクリアする必要がある．また，WordNet や OpenCYC，EDR 電子化辞書などの③既存の概念辞書との概念マッピングも本研究の重要なポイントである．これは，Wikipedia には幅広い概念に対する網羅性は高いものの，間違った情報も存在するため，（網羅性は低いが）制度の高い既存の概念辞書と融合することで，互いの弱点を補完し，さらに高い網羅性と精度向上が望めるためである．

３．研究の方法

本研究では，三つの解決するべき技術的課題があるが，研究初期には，情報の信頼性向上およびスケーラビリティに関する研究に注力して研究を進める．特に，信頼性向上に関する研究は比較的困難な問題であるため，プロジェクトの開始とともに着手し，プロジェクト全体でも最も多くの時間をかけて研究する．具体的には，今までの研究で判明している Wikipedia の情報の信頼性を計測するためのいくつかの指標（被参照数，PageRank，編集履歴，コンテンツ位置，著者ソーシャルネット，連想関係，Web ヒット数など）を統合的に計測する機械学習モデルを構築し，信頼性の数値化技術を実現する．次に，スケーラビリティについては，主に自然言語処理がボトルネックになるため，PC クラスタを利用した自然言語処理とリンク構造解析の分散処理を実現するルーチンの開発を進める．また，これらの研究がある程度進めば，概念辞書のプロトタイプとして利用可能な成果物ができるため，また，既存辞書とのマッピングに関する研究を一部開始する．

研究後期においては，基礎研究分野としては特に既存辞書とのマッチングに注力して研究を進める．ここでは，YAGO のように，述語論理の整合性をチェックする手法だけでなく，コンテキスト解析や別名（同義語）判

断により精度と網羅性の向上を図ることを目指している．これらの情報は，リンク構造（リダイレクトリンクとアンカーテキスト）を解析することである程度抽出できることは既に判明しており，さらに，自然言語処理と統合することで高度化が期待できる．また，本プロジェクトにとって重要なステップの一つである実アプリケーション開発を進める．具体的には，構築した大規模概念関係を利用した情報検索システムを構築し，その有用性を示す．これは，実社会に対する真の貢献を達成するためには，概念辞書を利用する実アプリケーションを提示してその有効性を示すことが必要不可欠なためである．そのため，本提案では，単に大規模な概念辞書を構築するだけでなく，実際のアプリケーションを開発することで，研究成果の有効性を示すと同時に，実応用の知見をさらにフィードバックして完成度の高い概念辞書を構築する．

４．研究成果

本研究の成果は多岐にわたる．まず，Wikipedia 上の構造化データを教師データとして活用し，少量の情報を元に多量のクラス分類を行う手法を確立した．この結果，本研究に先立ち構築していた大規模連想辞書「Wikipedia シソーラス」に対し，上位概念を付与することに成功した．また，RDF サポートおよびAPI の充実など，アプリケーションを考慮したインタフェースを提供することにより，さらに， Web を主とする他リソースとの融合による精度・網羅性向上を果たした点も大きな成果であった．

研究発表という面では，研究後期においては，研究計画に基づき（国際）会議での発表や論文発表を通じて積極的に成果の公開に尽力した．その結果，論文誌 9 本，口頭発表 30 件の成果を挙げることができたことは特筆すべき成果であると言える．これらの中には，マルチメディア分野で世界最高峰の論文誌である ACM. Trans. on Multimedia Computing, Communications and Applications の掲載論文や，知識処理分野および情報検索分野でトップレベルの国際会議である WI の発表論文が含まれている．これらの成果は，研究開始当初に想定していた以上のものである．

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 9 件）

① M. Erdmann, K. Nakayama, T. Hara, and S. Nishio, Improving the Extraction of Bilingual Terminology from Wikipedia, ACM Trans. on Multimedia Computing, Communications and Applications, 査読有, Vol. 5, No. 4, 2009.

② M. Ito, K. Nakayama, T. Hara, and S. Nishio, Semantic Relatedness Measurement based on Wikipedia Link Co-occurrence Analysis, International Journal of Web Information Systems (IJWIS), 査読有, Vol. 7, No. 1, pp. 44 – 61, 2011.

③ 中山浩太郎, 神経細胞移動に着想を得た自己組織化マップによるwikipediaリンクデータの可視化, 日本データベース学会論文誌, 査読有, Vol. 9, No. 3, pp. 19-24, 2011.

④ 白川真澄, 中山浩太郎, 荒牧英治, 原隆浩, 西尾章治郎, WikipediaとWebの情報を組み合わせたオントロジ構築の試み, 電子情報通信学会和文論文誌, 査読有, Vol. 94, No. 3, pp. 525 – 539, 2011.

⑤ 白川 真澄, 中山 浩太郎, 原 隆浩, 西尾 章治, Wikipediaのカテゴリグラフ解析による語句の確率的分類とその応用, 情報処理学会論文誌データベース, 査読有, 2012 （採録決定）

⑥ 白川 真澄 中山 浩太郎 原 隆浩 西尾 章治, Wikipedia とへ▯イス▯理論を用いた関連エンティティ推測と短文クラスタリン▯への応用, 日本データベース学会論文誌, 査読有, Vol.11, No.1, 2012 （採録決定）

〔学会発表〕（計 30 件）

① K. Nakayama, M. Ito, T. Hara, and S. Nishio, Wikipedia relatedness measurement methods and influential features, IEEE International Symposium on Mining and Web (MAW), pp. 738-743, 2009.

② M. Shirakawa, K. Nakayama, T. Hara, and S. Nishio, Relation Extraction between Related Concepts by Combining Wikipedia and Web Information for Japanese Language, Asia Information Retrieval Societies Conference (AIRS), 2010.

③ M. Shirakawa, K. Nakayama, T. Hara, and S.

Nishio. Relation extraction between related concepts by combining wikipedia and web information for japanese language. In Proceedings of Asia Information Retrieval Societies Conference (AIRS), December 2010.

④ K. Nakayama and Y. Matsuo. A self organizing document map algorithm for large scale hyperlinked data inspired by neuronal migration. In Proceedings of International World Wide Web Conference (WWW) Poster, pp. 95-96, 2011.

⑤ Masumi Shirakawa, Kotaro Nakayama, Takahiro Hara, and Shojiro Nishiol, Wikipedia Sets: Context-Oriented Related Entity Acquisition from Multiple Words, Proc. of IEEE/WIC/ACM Int'l Conf. on Web Intelligence (WI), 2011

〔その他〕

受賞等

① 白川真澄, 中山浩太郎, 荒牧英治, 原隆浩, 西尾章治郎, Web とデータベースに関するフォーラム(WebDB Forum 2011), 最優秀論文賞, 2010.

② 中山浩太郎, Web とデータベースに関するフォーラム (WebDB Forum) 最優秀論文, 2010

③ 中山浩太郎, 情報処理学会山下記念賞, 2012

④ Masahiro Ito, Kotaro Nakayama, Takahiro Hara, Shojiro Nishio, Emerald Highly Commended Paper Award (Awards for Excellence), 2012

ホームページ等

- 成果公開用 Web サイト (SigWP)
  http://sigwp.org

６．研究組織

(1)研究代表者

中山　浩太郎 (NAKAYAMA KOTARO)
東京大学・知の構造化センター・特任講師
研究者番号：00512097

(2)研究分担者

荒牧　英治 (ARAMAKI EIJI)
東京大学・知の構造化センター・特任講師
研究者番号：70401073

岡　瑞起 (OKA MIZUKI)
東京大学・知の構造化センター・特任研究員
研究者番号：10512105

増田　勝也 (MASUDA KATUYA)
東京大学・知の構造化センター・特任研究員
研究者番号：2012105

松尾　豊 (MATSUO YUTAKA)
東京大学・大学院工学系研究科・准教授
研究者番号：30358014
（研究計画調書当時：研究連携者）

原　隆浩 (HARA TAKAHIRO)
大阪大学・大学院情報科学研究科・准教授
研究者番号：20294043

(3)連携研究者

（　　）

研究者番号：