様式 C-19

科学研究費補助金研究成果報告書

平成　23 年　5 月 10 日現在

機関番号：14401
研究種目：基盤研究（C）
研究期間：2008　～　2010
課題番号：20500093
研究課題名（和文）　ソーシャルメディア解析による高精度連想記憶エンジンの研究

研究課題名（英文）
Social Media Analysis for Accurate Association Relational Engine

研究代表者
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

研究成果の概要（和文）：
ソーシャルメディアは，Web などを通じて多くのユーザが協調的にコンテンツを作り上げる仕組みであり，新しい知識抽出の基盤として注目されている．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高いスケーラビリティの実現と②情報の信頼性の向上という二つの技術的課題をクリアする必要があった．本研究課題では，これらの問題を解決するために，１）スケーラビリティの高い連想関係抽出手法，２）機械学習による重要素性の発見，３）テストコレクションの整備，４）多様なリソースの融合的解析手法などの研究を行い，成果を挙げることに成功した．

研究成果の概要（英文）：
Social media systems allow users to construct contents collaboratively on the Web. Because of the unique characteristics such as coverage and accuracy, it has become one of the promising corpora for knowledge extraction. However, there exist two major technical issues on social media analysis; scalability and credibility. In this research project, to solve these problems, we have conducted a series of research; 1) scalable association extraction, 2) machine learning for finding important features, 3) a large scale test collection construction and 4) integrated analysis across different types of resources.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2008　年度	1,200,000	360,000	1,560,000
2009　年度	1,100,000	330,000	1,430,000
2010　年度	1,100,000	330,000	1,430,000
総　計	3,400,000	1,020,000	4,420,000

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

ソーシャルメディアは，Web を通じて多くのユーザが協調的にコンテンツを作成する手法であり，網羅性と精度の高いコンテンツが形成されている．Wiki をベースにした Web 百科事典「Wikipedia」は，ソーシャルメディアの代表例であり，文化，歴史，数学，科学，社会，テクノロジなどの幅広い分野のト

ピックをカバーしている．これらのソーシャルメディアは，「群集の叡智」と呼ばれる形式の新しいメディアであり，知識抽出のためのコーパスとして，その有用性が研究者の間でも急速に注目を集めはじめている．

研究代表者らはこれまでの研究で，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（http://sigwp.org 参照）．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ことが判明している．研究代表者らは，連想シソーラス辞書構築に関する研究を通じて，Wikipediaのリンク構造を解析することで極めて実用性の高い知識抽出が可能であることを証明してきており，Wikipediaが人工知能研究の新しいパラダイムシフトを引き起こす可能性を秘めたコーパスであることを証明してきた．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) 高いスケーラビリティの実現と(2) 情報の信頼性の向上という二つの技術的課題をクリアする必要があった．

２．研究の目的

上記の研究背景に基づいて，本研究ではこれら二つの技術的課題を解決することを目的とし，幅広い分野で利用可能な汎用連想エンジンに関する研究を行う．具体的には，以下の二つの観点から研究開発を推進する．

(1) スケーラビリティの高い連想記憶エンジンの実現
(2) ソーシャルメディアからの信頼性の高い情報の抽出

３．研究の方法

ソーシャルメディアからの知識抽出に関する研究は黎明期を迎えたばかりである．事実，国内ではWikipediaが知識抽出のコーパスとして利用されている研究は非常に少ない．しかし，国外ではいくつかの先行研究が存在する．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．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の指標を用いて語彙同士の関係性を0から1までの数値で表す．しかし，これらの研究ではスケーラビリティに関する考察も情報の信頼性に対する考察も無い．そのため，実用的なアプリケーションで利用するには技術的な課題が残されているのが現状である．また，情報の信頼性についても考慮がなく，ソーシャルメディアからの情報抽出手法としては問題がある．

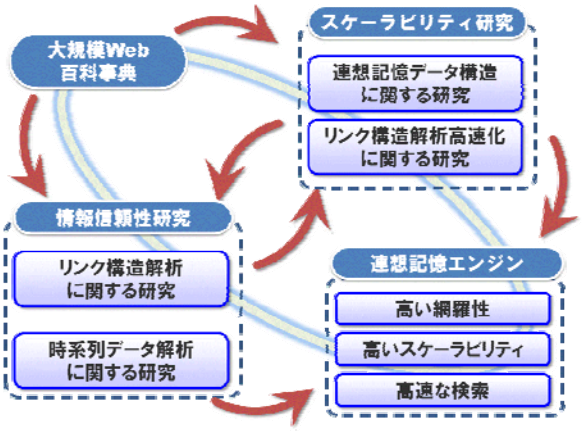

図１：本研究課題の全体概要

そこで，本研究では図１に示すとおり，(1) スケーラビリティの高い連想記憶エンジンに関する研究と(2) ソーシャルメディアからの信頼性の高い情報の抽出に関する研究を行う．研究の方法として，まずはスケーラビリティの高い連想記憶エンジンについて注力して研究を進める．これは，抽出した連想記憶の効率的な解析などについても，高速に処理する必要があるためである．つまり，信頼性の高い情報抽出の研究を行う際には，頻繁に精度を計測する必要があるが，このときにスケーラビリティの高い連想記憶の仕組みが完成していなければ，研究を円滑に進めることができない．その後，ソーシャルメディアからの信頼性の高い情報抽出に関する研究を進める．特に，どのような要素が情報の信頼性に影響を与えるのかについて，実際にWikipediaの記事データを調査し，アルゴリズムの設計を進める．

最後に，信頼性の高い情報抽出手法についての研究を重点的に推進する．最終的にはスケーラビリティ技術との融合を果たし，ソーシャルメディアから信頼性の高い情報を抽出し，精度の高い連想記憶エンジンを開発する．

また，本研究開発では，得られた知見などを積極的に論文化し，公開することで他の研究者と交流し，迅速な研究推進を図る．これは，Wikipediaマイニングという極めて新しい研究分野では，非常に重要なポイントであ

る．しかし注意しなければならない点としては，最初から必要以上に研究発表数や論文投稿数を重視した場合，重要な研究開発部分に注力できない可能性があることである．そのため，研究初期にはあえて研究発表・論文投稿数を絞り，質の高い研究を実現することに注力する．逆に，研究後期では，積極的に研究発表と Web を通じた情報公開を進め，研究成果の公開に努める．

４．研究成果

(1) スケーラビリティの高い連想抽出手法

研究初期においては，特にスケーラビリティの高い連想記憶エンジンについて注力して研究を進めてきた．これは，抽出した連想記憶を実際のアプリケーションに適用するためには効率的な解析手法および連想データの抽出手法が必要であるためである．

その成果として，ソーシャルメディアにおけるハイパーリンクの共起性に基づく連想関係抽出手法を確立した（図 2）．本手法は，同じページや同じセンテンスなど，定められた領域の中に共に出現するハイパーリンク同士は関係性が強い，という戦略に基づく手法である．この結果，従来手法で課題であったスケーラビリティの面で高いパフォーマンスを実現することができた．これは，共起性解析がデータの数に対して線形にスケールする特性を利用したものである．

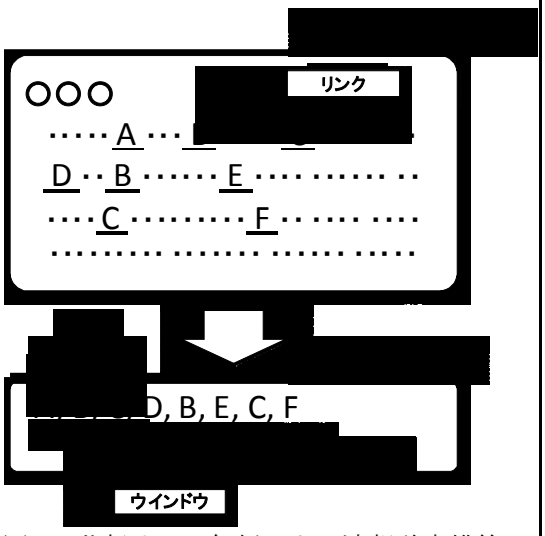

図 2：共起リンク解析による連想辞書構築

(2) 機械学習による重要素性の発見

また，詳細な実験により，各種の共起性解析モデルを評価し，高いスケーラビリティを実現しつつ，従来手法と同等の精度を実現することにも成功した．具体的には，SVM（Support Vector Machine）により，多属性（リンク共起性，TF-IDF，相互リンク，pfibf，リンクの出現位置，所属カテゴリ）を考慮した連想関係抽出アルゴリズムを構築し，さらなる精度向上を実現した．この結果，共起性解析やページ間のリンク解析手法，相互リンクなどの情報が連想関係を抽出するには重要な要素であることが判明した．

(3) テストコレクションの整備

本研究の重要な成果の一つが，大規模テストコレクションを利用した連想関係抽出の精度向上である．テストコレクションを利用することにより，手法の精度を客観的・機械的に評価できるようになった．この結果，パラメータを変更しながらアルゴリズムを最適化することが可能となり，精度向上に貢献した．また，正解集合としてテストコレクションを用意することで，Web 情報や，リンク共起性解析などの素性を利用した機械学習手法の適用が可能となり，より高い精度を実現できた．重要な素性の組み合わせを発見することができることから，計算量を押さえつつ，精度を向上するための素性の組み合わせを調べることができた．

表１：テストコレクションのデータ例

Concept 1	Concept 2	Relatedness
Mobile_phone	Cell_phone	9.8
Thailand	Thai_people	9.6
Photography	Photographic_paper	8.4
Thailand	Buddhist	8
Photography	Photojournalism	7.6
Mobile_phone	AT&T	7.2
Mobile_phone	Bluetooth	6.6
Photography	Black-and-white	6
Mobile_phone	Car_phone	5.8
Thailand	Myanmar	4.6
Mobile_phone	Las_Vegas,_Nevada	1.2
Thailand	Europe	1.2
Photography	Greek_language	1.0
Photography	France	0.8
Mobile_phone	Postage_stamp	0.2

(4) 多様なリソースの融合的解析手法

最後の重要な研究成果は，ソーシャルメディア外の情報との融合および信頼性の数値化手法である．昨年度の成果である大規模テストコレクションを利用したことで，多様な情報の中から信頼性や連想関係に影響を与える要素を発見した．特に，ソーシャルメディア外の情報として，Web 情報（Web 共起）情報の有効性が確認され，手法の精度と網羅性の向上に貢献した．さらに，大規模なデータを効率的に取り扱うための基盤アルゴリ

ズムの研究にも進展があり，Web ドキュメントのように大規模なリンク情報を扱う技術の方向性を開けたことは，今後の研究を進める上で重要なポイントであった．

最終年度においては，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，最終年度だけで論文誌 4 本，口頭発表 12 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．研究期間全体では，雑誌論文 9 件，学会発表 24 件の成果を挙げた．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 CIKM 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

〔雑誌論文〕（計 9 件）

① 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.

② 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.

③ 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.

④ 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.

〔学会発表〕（計 24 件）

① 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 年 12 月 6 日.

② K. Nakayama, M. Ito, T. Hara, and S. Nishio, Wikipedia relatedness measurement methods and influential features, IEEE International Symposium on Mining and Web (MAW), 2009 年 5 月 26 日.

③ M. Ito, K. Nakayama, T. Hara and S. Nishio, Association Thesaurus Construction Methods based on Link Co-occurrence Analysis for Wikipedia, Conference on Information and Knowledge Management (CIKM), 2008 年 10 月 28 日.

④ K. Nakayama, T. Hara, and S. Nishio, Wikipedia Link Structure and Text Mining for Semantic Relation Extraction, International Semantic Search Workshop (SemSearch), 2008 年 6 月 2 日.

〔その他〕

受賞等

① 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.

② 白川真澄，中山浩太郎，荒牧英治，原隆浩，西尾章治郎，データ工学と情報マネジメントにするフォーラム(DEIM 2010)，最優秀論文賞，2010.

③ 中山浩太郎，iDBフォーラム 2008，優秀若手研究者賞，2008.

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
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

(2)研究分担者
中山 浩太郎（NAKAYAMA KOTARO）
東京大学・知の構造化センター・特任助教
研究者番号：00512097