様式 C-19

科学研究費補助金研究成果報告書

平成 23 年 3 月 31 日現在

機関番号：14301
研究種目：若手研究（B）
研究期間：2009 年度～2010 年度
課題番号：21700105
研究課題名（和文）　オンデマンド利用を目的とするＷｅｂからの知識発見に関する研究

研究課題名（英文）　Research on Knowledge Acquisition from the Web for On-demand Usage

研究代表者
大島　裕明（OHSHIMA HIROAKI）
京都大学・大学院情報学研究科・助教
研究者番号：90452317

研究成果の概要（和文）：本課題では、リアルタイムに知識発見を行う技術の開発を行った。主に、2 つの異なる方向性を持つ構文パターンを用いて Web 検索を行い、その検索結果のテキストから効率的に目的の知識を獲得する手法を用いて、様々な知識発見手法を開発した。様々な種類の知識を発見するために、適切な構文パターンを発見する手法についても開発した。さらに、様々な知識発見手法を用いたアプリケーションの作成を行った。

研究成果の概要（英文）：This research has developed a bunch of technologies for real-time knowledge acquisition. Two different directional syntactic patterns are used to obtain Web search results and knowledge can be efficiently extracted from the search results. A method for finding appropriate syntactic patterns for various kinds of knowledge has been also developed. The acquired knowledge can be used in many applications.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2009 年度	1,900,000	570,000	2,470,000
2010 年度	1,400,000	420,000	1,820,000
総　計	3,300,000	990,000	4,290,000

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：情報検索，情報抽出

１．研究開始当初の背景

本研究を開始した当初、様々なリソースから知識発見を行う研究が数多く行われている状況であった。例えば、1 語ないしは数語が与えられたときに、その上位語、下位語、類義語、同位語（共通の上位概念を持つ兄弟概念にあたる語）を取得するような技術や、そのような関係にある語のペアを大量に収集するような研究がなされていた。また、これらのような辞書的に定義される関係ばかりでなく、話題語や詳細語といったような定義があいまいな語や、略語、人物の呼称や愛称など、多様な関係にある知識の発見が研究の対象であった。知識発見に用いるリソースには、大規模テキストコーパス、Web のクローリングによって収集された大量の HTML 文書、ログデータのような大規模データ、Web 検索を利用して収集した文書などがあげられる。このような知識発見の研究の多くはリアルタイムでの知識発見を目的とはせず、ある程度時間をかけて大規模な辞書を作成するような利用を前提としたものであった。

２．研究の目的

本研究では、オンデマンドに利用されることを目的として、Web 上の種々のリソースから知識発見を行う手法とその応用についての研究を行う。

おおむね 10 秒を目標として、要求が行われてからリアルタイムに知識発見を行う手

法を開発する。リアルタイムに知識発見を行うことができれば、あらかじめ想定しておくことができないほど多様な要求がユーザから行われるサービスやアプリケーションにおいて、要求に応じて得られた知識を利用することができるようになる。リアルタイムに知識発見を行うため、Web 検索エンジンを利用し、知識発見に必要なデータを効率よく収集する手法や、少量のデータからでも精度良く知識発見を行う手法について研究を行う。テキストからの知識発見には、主に構文パターンを用いるものとする。例えば、上位語や下位語を発見するためには「〈上位語〉 such as 〈下位語〉」という構文パターンが、同位語を発見するためには「〈同位語〉 や 〈同位語〉」という構文パターンが有用であることが知られている。しかし、様々な種類の知識に対して、その発見において有用な構文パターンを思いつくことは容易ではない。そこで、正解の語のペアを与えることによって、知識発見に有用な構文パターンを自動的に発見する手法についても研究が必要となる。さらに、Web から発見される様々な知識を用いた様々なアプリケーションについても検討する。

３．研究の方法

(1) 多様な知識を対象とするリアルタイムな知識発見手法の確立

まず、様々な種類の知識について、リアルタイムな知識発見が実現可能であるかを明らかについて明らかにする。次に、それぞれの知識について、どの程度の精度で、どの程度の網羅的に発見できるかを明らかにする。Web 上のリソースとして、一般的な Web 検索はもちろんのこと、ニュース検索、ブログ検索や、Wikipedia のデータなど、多くの異なるリソースを用いて、多様な知識発見の可能性を検討する。

(2) 多様な知識発見の実現を容易にする機構の検討

Web から発見できる知識には、非常に多くの種類が存在する。各々の知識を発見するための手法はある程度異なるが、完全に異なる手法で実現されるわけではなく、ある程度の共通点を持っている場合があると考えられる。そこで、そのような共通点について明らかにする。そして、その共通点を考慮して、様々な種類の知識発見の実現を容易にするためのプラットフォームの構築などを検討する。

(3) 発見された知識が応用可能なサービスやアプリケーションの検討

本研究で対象とする知識発見は、様々なサービスやアプリケーションにおいてオンデマンドに利用するためのものである。Web 検索エンジンのように、ユーザがキーワードを入力するようなサービスの場合、ユーザの要求は多岐にわたり、あらゆる要求をあらかじめ想定しておくことはできない。そのような場合には、本研究で確立された知識発見手法を用いることによって、その場で必要な知識を求め、それを用いたサービスを展開することが可能となる。そのような、本研究の特性を活かすことが可能な応用分野について検討を行う。

４．研究成果

(1) 多様な知識のリアルタイムな発見手法の実現

本研究で開発した知識発見手法は、典型的には、1 語を与えると、ある特定の関係にある語を数語から十数語程度発見するものである。

主に、2 つの異なる方向性を持つ構文パターンを用いて検索を行い、その Web 検索結果から目的の語を抽出する技術を用いて、様々な知識発見手法を作成した。2 つの異なる方向性を持つ構文パターンとは、例えば、上位語発見においては、

・「〈上位語〉 である 〈下位語〉」
・「〈下位語〉 は 〈上位語〉」

というような構文パターンのことである。発見したい語は、この場合上位語であるが、1 つ目の構文パターンではそれが先頭に現れており、2 つ目の構文パターンでは、それが最後に現れている。この場合、上位語を発見する際には、下位語が与えられる。上記のパターンにおいて、〈下位語〉の部分に、与えられた語を代入し、〈上位語〉の部分を除いて、検索エンジンに対するクエリとする。そのクエリを用いて得られた検索結果のテキストの中から、構文パターンに適合する語を抽出することで上位語の発見が行われる。その際に、語を正しく切り出すことが必要となる。上記のような 2 つの異なる構文パターンは、それぞれ、切り出す語の最初と最後を決定することができる。それぞれ単独では、切り出す語の最初または最後を決めることができないが、両者を用いることで、形態素解析などを用いなくても、正しく語を切り出すことが可能となる。日本語の場合、語と語の間には、スペースなどが存在しないが、ある程度のストップワードリストを用意することによって、多くの場合、問題なく切り出すことが可能であることが確かめられた。

構文パターンを変更することによって、上位語、下位語、同位語など、非常に多様な関係にある語を発見することが可能であることが確かめられた。

構文パターンを用いる手法以外にも、Wikipedia のダンプデータを用いて、専門的

な語を発見する手法や、Wikipedia 項目の時間・空間・時空間的な関連度を計算する手法を開発した。専門的な語の発見についての基本的なアイデアは、ある同一のトピックに関するWikipedia項目集合が与えられたときに、その集合内からの被リンクが、それ以外の部分からの被リンクよりも優位に多い場合には、その語を専門的であるとみなすものである。ある同一のトピックに関する Wikipedia 項目集合の取得には、上記で述べた構文パターンを用いた話題語発見手法などを用いる。時空間的な関連度の計算については、例えば、ある人物のある時代、ある場所における重要度を算出するというような技術である。ある人物が生きた時代に対して、遠い時代から参照されている場合には、そのような人物は重要とみなすというアイデアに基づいて、重要度を計算する。ここで、参照としては、Wikipedia 項目間のリンクを用いる。リンク分析手法の１つであるBiased PageRankを用いたアルゴリズムによって、各時間や空間におけるWikipedia項目の関連度を計算する手法を提案した。

(2) 知識発見手法のための構文パターンの発見手法の確立とそれを応用した関係の類似性に基づく知識発見手法の確立

構文パターンを用いた知識発見手法では、良い構文パターンを用いなくてはならない。そのために、良い構文パターンを用意する必要があるが、場合によっては、良い構文パターンを発見することが困難である場合がある。そこで、良い構文パターンを自動的に発見するための手法の開発を行った。ユーザは、ある語と、その語が与えられたときに、発見されるべき正解の語を与える。それらが Web 上でどのような現れ方をしているかを、構文パターンを用いた知識発見と同様に、Web 検索を行うことで取得する。このようにして発見された構文パターンには、与えられた語のペアにのみ特化して適合しているものがあるため、様々な入力語においてうまく機能する構文パターンであるかどうかを評価し、最終的により良い構文パターンを発見する。

本手法を用いることによって、1 語を入力とする知識発見とは別に、与えられた語のペアの関係と類似するような語を発見するという、新しい知識発見が可能となった。例えば、(京都, 八つ橋) という語のペアを与え、さらに、(大阪, X) と与えられたときに、X に当てはまるような語を取得することができるようになった。提案手法では、まず、「京都」から「八つ橋」を発見できるような構文パターンを発見し、それを用いて、今度は「京都」を「大阪」に置き換えた際に発見できる語が上記のクエリに対する回答となる。このような新しい知識発見の問題に取り組むため、構文パターンを用いない手法についても検討、開発を行った。

構文パターンを用いる手法は、比較的明確に定義されるような関係にある場合にはうまく機能するが、説明が困難な関係の場合にはうまく機能しない。そこで、与えられた語のペアが共起する際に、有意によく共起する別の語が存在するというアイデアに基づく手法を開発した。共起の頻度については、Web 検索のヒットカウントを利用する手法や、Web 検索結果の上位における共起から求めた。さらに、構文パターンを用いる手法と、共起を用いる手法を組み合わせた手法を開発し、本問題に対してより良い精度を実現することに成功した。

(3) 知識発見プラットフォームの実現

Webからの知識発見は、

・Webからの情報収集
・テキスト処理
・データ集約

という３つの段階からなることが多い。通常は、Web からの知識発見のために、これらの処理を行うプログラムを作成する必要があるが、様々なデータソースを用いたり、様々な構文パターンを用いたりするために、いちいちプログラムを修正することはコストがかかると考えられる。そこで、このような処理が比較的容易に行えるプラットフォームの開発を行った。具体的には、既存機能としてデータの集約機能を標準的に保有する関係データベース環境に、Web からの情報収集機能や、テキスト処理機能を付加することで、上記の 3 つの処理が行える環境を作成した。SQL という宣言型言語によって知識発見のための処理を記述することが可能となり、通常のプログラミング言語を用いる場合に比べて非常に少量の記述で同様の処理が記述できる。

知識発見のために用いられるリソースは Web 上のものに限定する必要はなく、ローカルデータベースに保有する特許データベースや、Wikipedia のダンプデータのデータベースなどを用いることもある。本環境では、それらもシームレスにアクセス可能な環境であり、Web リソースとローカルリソースを橋渡しする環境であるということもできる。

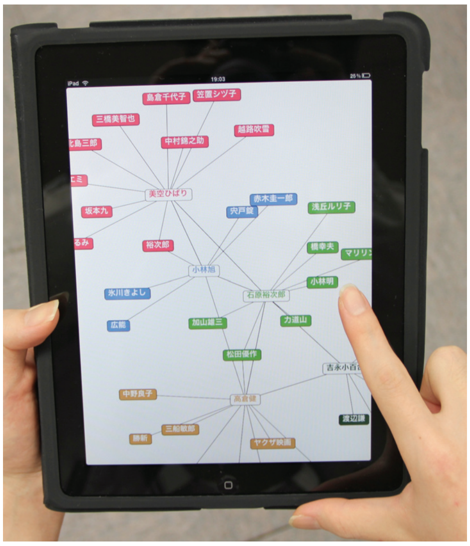
(4) 知識の時代的変化を閲覧するアプリケーションの開発

Web が誕生してから時間が経過したことによって、アーカイブされたWeb情報を利用することが可能になってきている。また、Google News Archive のように、既存のメディアのアーカイブの中にも、Web からアクセスできるものがある。それらのアーカイブされた情報を用いて、時代ごとに分けて知識発

見を行うことで、知識がどのように変化してきているかを閲覧することができるアプリケーションの開発を行った。
例えば、Google や Yahoo!と関連する語が最近 10 年でどのように変化してきているかを閲覧することができるなど、興味深い知識を閲覧可能である。

(5) クエリが思い付かない場合の Web ページ検索支援

小中学生が検索を行う際や、あまり知らない分野について検索を行う際には、どのような語をクエリとして用いれば良いか思い付きにくい場合がある。そこで、そのようなユーザが Web 検索を行うことを支援することを目的として、ユーザがはじめに適当な 1 語を与えるだけで、次々と関連する語を発見していくことができるアプリケーションの開発を行った。



本研究では、特に、より親しみやすく操作できる環境として iPhone や iPad を用い、それらの機器上で動作するアプリケーションを作成した。上図は、iPad 上で動作している例である。最初に 1 語を入力すると、関連する語がグラフとして表示される。新たに求められた語をダブルタップすると、さらにその語の関連語を求められるようになっている。一度関連語を求めた語をさらにダブルタップすると、その語が使われている文章が表示され、その中から気になるものを選択すると、その文章が実際に使われている Web ページを閲覧することが可能になっており、従来の Web 検索とは全く異なる新しい Web 検索インタフェースとして機能している。実際に、中学生や、一般の方々に利用してもらいながら改良を行った。

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 10 件）

① Mitsuo Yamamoto, Yuku Takahashi, Hirotoshi Iwasaki, Satoshi Oyama, Hiroaki Ohshima, Katsumi Tanaka, “Extraction and Geographical Navigation of Important Historical Events in the Web,” Proceedings of the 10th International Symposium on Web and Wireless Geographical Information, 査読有, pp. 21-35, 2011.
② Makoto P. Kato, Ohshima Hiroaki, Satoshi Oyama, Katsumi Tanaka, “Search as if You were in Your Home Town: Geographic Search by Regional Context and Dynamic Feature-space Selection,” Proceedings of the 19th ACM Conference on Information and Knowledge Management (CIKM 2010), 査読有, pp. 1541-1544, 2010.
③ Natsuki Takata, Hiroaki Ohshima, Satoshi Oyama, Katsumi Tanaka, “Searching the Web for Alternative Answers to Questions on WebQA Sites,” Lecture Notes in Computer Science, Web-Age Information Management - WAIM 2010, 査読有, Vol. 6184, pp. 441-452, 2010.
④ Ryohei Takahashi, Satoshi Oyama, Hiroaki Ohshima, Katsumi Tanaka, “Evaluating Truthfulness of Modifiers Attached to Web Entity Names,” Lecture Notes in Computer Science, Web-Age Information Management - WAIM 2010, 査読有, Vol. 6184, pp. 429-440, 2010.
⑤ 高橋 良平, 小山 聡, 大島 裕明, 田中克己, Web テキストと修飾表現との適合度判定手法, 日本データベース学会論文誌, 査読有, Vol. 9, No. 1, pp. 41-46, 2010.
⑥ Hiroaki Ohshima, Satoshi Oyama, Katsumi Tanaka, “Cloud as Virtual Databases: Bridging Private Databases and Web Services,” Lecture Notes in Computer Science, Database Systems for Advanced Applications - DASFAA 2010, 査読有り, Vol. 5981, pp. 491-497, 2010.
⑦ Hiroaki Ohshima, Katsumi Tanaka, “High-speed Detection of Ontological Knowledge and Bi-directional Lexico-Syntactic Patterns from the Web,” Journal of Software, 査読有,

Vol.5, No.2, pp.195-205, 2010.
⑧ Hiroaki Ohshima, Satoshi Oyama, Hiroyuki Kondo, Katsumi Tanaka, "Web Information Credibility Analysis by Geographical Social Support," Proceedings of the 3rd International Universal Communication Symposium (IUCS 2009), 査読有, pp.193-196, 2009.
⑨ Makoto P. Kato, Hiroaki Ohshima, Satoshi Oyama, Katsumi Tanaka, "Query by Analogical Example: Relational Search Using Web Search Engine Indices," Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM 2009), 査読有, pp.27-36, 2009.
⑩ Hiroaki Ohshima, Adam Jatowt, Satoshi Oyama, Katsumi Tanaka, "Seeing Past Rivals: Visualizing Evolution of Coordinate Terms over Time", Lecture Notes in Computer Science, Web Information Systems Engineering - WISE 2009, 査読有, Vol.5802, pp.195-205, 2009.

〔学会発表〕(計 15 件)

① 内藤 稔, 大島 裕明, 高橋 亜希子, 田中 克己, 複数文書閲覧時の文書間の意味的関係の抽出と提示による文書ナビゲーション, 第 3 回データ工学と情報マネジメントに関するフォーラム, F8-4, 静岡県伊豆市, 2011 年 2 月 28 日.
② 高橋 侑久, 大島 裕明, 山本 光穂, 岩崎 弘利, 小山 聡, 田中 克己, Wikipedia リンク構造を用いた歴史エンティティの重要度計算, 第 3 回データ工学と情報マネジメントに関するフォーラム, B3-2, 静岡県伊豆市, 2011 年 2 月 27 日.
③ 高田 夏希, 大島 裕明, 田中 克己, Web と QA コンテンツの相互補完, 第 3 回データ工学と情報マネジメントに関するフォーラム, D2-1, 静岡県伊豆市, 2011 年 2 月 27 日.
④ 川野 悠, 大島 裕明, 田中 克己, Web 閲覧行動に応じたマルチファセットの動的生成と比較ページの検索, 第 3 回データ工学と情報マネジメントに関するフォーラム, F1-6, 静岡県伊豆市, 2011 年 2 月 27 日.
⑤ 高田 夏希, 大島 裕明, 田中 克己, Web と QA コンテンツの相互補完に基づくソーシャルサーチ, Web とデータベースに関するフォーラム 2010 (WebDB Forum 2010), 2A-3, 東京都新宿区, 2010 年 11 月 11 日.
⑥ 川野 悠, 大島 裕明, 田中 克己, Web からの知識抽出による閲覧ページの動的なマルチファセット生成, 平成 22 年度 情報処理学会関西支部 支部大会, F-12, 大阪府大阪市, 2010 年 9 月 22 日.
⑦ 高橋 侑久, 大島 裕明, 小山 聡, 田中 克己, リンク構造分析と時空間詳細度制御に基づくイベント情報の一般性・専門性発見と提示, 平成 22 年度 情報処理学会関西支部 支部大会, E-06, 大阪府大阪市, 2010 年 9 月 22 日.
⑧ 高橋 良平, 小山 聡, 大島 裕明, 田中 克己, 発信者分析による修飾語の信憑性判定, 平成 22 年度 情報処理学会関西支部 支部大会, E-05, 大阪府大阪市, 2010 年 9 月 22 日.
⑨ 加藤 誠, 大島 裕明, 小山 聡, 田中 克己, アナロジーに基づく地理情報検索, 情報処理学会 第 72 回全国大会, 6ZC-9, 東京都文京区, 2010 年 3 月 11 日.
⑩ 高田 夏希, 小山 聡, 大島 裕明, 田中 克己, 質問に対する回答を含む Web ページの発見手法, 情報処理学会 第 72 回全国大会, 6ZC-3, 東京都文京区, 2010 年 3 月 11 日.
⑪ 高橋 良平, 小山 聡, 大島 裕明, 田中 克己, Web オブジェクトの修飾語表現の信憑性検証, 情報処理学会 第 72 回全国大会, 5ZN-1, 東京都文京区, 2010 年 3 月 11 日.
⑫ 高田 夏希, 大島 裕明, 小山 聡, 田中 克己, QA コンテンツに対する別解の発見とランキング, 第 2 回データ工学と情報マネジメントに関するフォーラム (DEIM Forum 2010), A8-3, 兵庫県淡路市, 2010 年 3 月 1 日.
⑬ 加藤 誠, 大島 裕明, 小山 聡, 田中 克己, 地物間の距離を考慮した動的な類似性尺度に基づく地理情報例示検索, 第 2 回データ工学と情報マネジメントに関するフォーラム (DEIM Forum 2010), D7-3, 兵庫県淡路市, 2010 年 3 月 1 日.
⑭ 高橋 良平, 小山 聡, 大島 裕明, 田中 克己, Web テキストと修飾表現との適合度判定手法, 第 2 回データ工学と情報マネジメントに関するフォーラム (DEIM Forum 2010), C3-3, 兵庫県淡路市, 2010 年 2 月 28 日.
⑮ 川野 悠, 大島 裕明, 田中 克己, クエリに応じたファセットの動的抽出による Web 画像検索結果の提示, 第 2 回データ工学と情報マネジメントに関するフォーラム (DEIM Forum 2010), A1-5, 兵庫県淡路市, 2010 年 2 月 28 日.

〔図書〕(計 0 件)

〔産業財産権〕
○出願状況（計０件）

○取得状況（計０件）

６．研究組織
(1)研究代表者
　大島　裕明（OHSHIMA HIROAKI）
　京都大学・大学院情報学研究科・助教
　研究者番号：90452317