

氏 名	おおしま ひろあき 大 島 裕 明
学位(専攻分野)	博 士 (情 報 学)
学 位 記 番 号	情 博 第 246 号
学位授与の日付	平 成 19 年 3 月 23 日
学位授与の要件	学 位 規 則 第 4 条 第 1 項 該 当
研 究 科 ・ 専 攻	情 報 学 研 究 科 社 会 情 報 学 専 攻
学 位 論 文 題 目	Web からの概念構造発見と検索への応用

論文調査委員 (主 査)
教授 田 中 克 己 教授 石 田 亨 教授 吉 川 正 俊

論 文 内 容 の 要 旨

本論文では、Web における情報取得手段として最も重要であると考えられる検索システムを、より知的で高度な情報取得システムとすることを目的とし、Web に存在する様々な情報から概念構造を発見し、発見された概念構造を活かした情報取得を行うための方式について述べている。

近年のインターネット環境、特に World Wide Web (WWW) は、新たな情報を取得するのに非常に有用であり、今や我々の生活に欠かせないものになってきている。Web の中でも、特に Web 検索エンジンの占める位置は非常に重要であり、あらゆる情報につながる Web の入口になっている。現在の Web 検索エンジンの機能は、クエリとして与えられたキーワードを含む Web ページ群を何らかのランキング手法によって順位付けして列挙するというものである。しかし、検索エンジンを利用するユーザの目的は、辞書的な利用、社会的評価の取得、集約的情報の取得、比較などの目的に応じた情報検索、個人化された検索など、単なる文書取得ではない多岐にわたるものである。Web 情報取得に関わる技術がこれらの目的に応じるべきであるという社会的要請は極めて高い。

本論文における研究は、そのような社会的状況に対する回答として、Web の情報から様々な形式での概念構造の発見を行い、それを利用して Web 上における情報取得をより良くするというアプローチをとっている。概念構造とは、概念どうしの関係性であり、本論文では、人間の認知する概念構造について、コンピュータで扱える情報として推測されたもののことをいう。例えば、ある語に対して上位関係、下位関係にはどのような語があるのかといったものや、ある概念を特徴ベクトルで表したときに、上位関係にあるような概念がどのような特徴ベクトルで表せるか、というような情報である。Web には様々な情報が存在しており、工夫すれば様々な目的に応用できる概念構造が発見できる。そして、発見された概念構造を活かした情報取得を行えるようにするのである。

本論文では、具体的には、Web 検索を用いた同位語とそのコンテキストの発見、文書事例を問合せとする兄弟カテゴリ文書の検索、および、個人保有文書からの概念構造の発見とその利用という3つの課題について研究を行い、その有効性の検証を行っている。

本論文は6章から構成されている。その概要は以下の通りである。

第1章は序論であり、本論文の研究の背景、本論文の研究を行うに至った動機、および、本論文の研究の全体の概要について述べている。

第2章では、1章で行った広い意味での本論文の位置づけから、第3～5章において説明するそれぞれの課題について特に深く関連する従来の研究について説明し、さらに位置づけを明確にしている。

第3章では、Web 検索エンジンのインデックス情報を利用した同位語の発見手法を提案している。これは、ある語が与えられたときに、その語の同位語を発見するものである。さらに、発見された同位語が持つコンテキストを発見する手法についても提案している。

第4章では、文書事例を問い合わせとする兄弟カテゴリ文書の検索手法について提案している。まず、ユーザがいくつかの文書をクエリとして与えたときに、それらの兄弟カテゴリ文書はどのような特徴を有するべきかや、どのような特徴を有すべきではないかということを、特徴ベクトルを用いて表す。そして、与えられたクエリに対してある文書が兄弟カテゴリ文書としてどの程度適合しているかを判定する方式を提案している。

第5章では、個人保有文書中の語の意味関係の発見とそのWeb文書検索への応用について提案している。個人のPC上に存在する文書がディレクトリツリーのような構造上に分類されている時に、その構造上における語の分布の仕方を考慮することによって、その個人が語と語の関係をどのようにとらえているかを、概念構造として発見する方式を提案している。そして得られた概念構造をクエリ拡張や再ランキングに用いる方法を提案している。

最後に、第6章において、本研究で得られた研究成果をまとめ、さらに今後の展開について述べている。

論文審査の結果の要旨

本論文は、Webに存在する様々な情報から概念構造を発見し、発見された概念構造を活かした情報取得を行う手法について検討したものである。

従来のWeb検索エンジンは、利用者がキーワードを入力し、そのキーワードを含んだ文書を羅列するというものであった。一方、ユーザがWebから情報を取得する際には様々な目的を持っており、現在のWeb検索エンジンだけでは十分に対応できていないという状況にある。学位申請者はユーザが現在潜在的に持っている様々な目的に対応するため、Webから概念構造を発見し、それを検索に応用することによってこれまでに困難であったような新たな情報取得を可能にする手法を提案している。さらに、提案手法それぞれについて、それらの有効性検証のために評価実験を行い興味ある結果を提示している。

具体的には、本論文の研究によって得られた成果は以下のように要約される。

1. 概念構造を取得する情報源としてWeb検索エンジンのインデックス情報に着目し、ある語が与えられたときにその語の同位語を発見する手法について提案した。様々な粒度の語を用いて実験を行い、一般的な語からあまり有名ではないような語においても提案手法が有効性を持つことを示した。さらに、いくつかの同位語が共通して持つ話題となるような語を、同位語が持つコンテキストとして抽出・提示する手法も提案し、実験によってその有効性を示した。
2. ユーザが提示するいくつかの文書事例に対して、それらが属するカテゴリを想定したときに、兄弟関係にあるカテゴリに分類されるような文書を検索する手法について提案した。従来の文書検索は、基本的にはクエリとして与えられた文書の類似文書を検索するというものであるのに対し、ユーザが提示した文書に対して、求める文書は、その上位カテゴリが共通しているが、兄弟関係にあるようなカテゴリに属するというような、新たな文書検索手法についての提案を行った。提案手法と従来手法として疑似適合性フィードバックを用いた手法との比較実験を行い、この提案手法が兄弟関係にあるカテゴリに分類されるような文書をより良く得られることを明らかにした。
3. 個人が自身のローカルコンピュータ上に保有する文書群とそれらが分類されているディレクトリ構造から、個人が持っている語の関係性の発見する手法について提案した。具体的には、分類されている文書に現れる語が、分類構造上でどのような出現分布をしているかに着目し、出現分布の仕方に応じて4種類の語の関係性を抽出する手法を提案した。また、得られた語の関係性を従来のWeb検索の個人化への利用として、クエリ拡張と検索結果の再ランキング手法について提案し、実験によってその有効性を示した。

以上、本論文は、発見した概念構造を利用して様々な目的に応じた情報取得手法を提供するという新たなフレームワークを提案し、その中で新たな概念構造発見手法や、従来は無かった新たな目的の情報取得が可能となる手法を提案しており、学術上、實際上寄与するところが少なくない。よって、本論文は、博士（情報学）の学位論文として価値あるものと認める。また、平成19年2月7日に論文内容とそれに関連した事項について試問を行った結果、合格と認めた。