

ウェブ検索エンジンのためのテキスト種別特定に関する研究

著者	吉田 光男
学位授与大学	筑波大学 (University of Tsukuba)
学位授与年度	2013
報告番号	12102甲第6868号
URL	http://hdl.handle.net/2241/00122401

氏名(本籍)	吉田 光男 (和歌山県)			
学位の種類	博士 (工学)			
学位記番号	博 甲 第 6868 号			
学位授与年月日	平成26年 3月25日			
学位授与の要件	学位規則第4条第1項該当			
審査研究科	システム情報工学研究科			
学位論文題目	ウェブ検索エンジンのためのテキスト種別特定に関する研究			
主 査	筑波大学 教授	博士 (工学)	山本幹雄	
副 査	筑波大学 教授	工学博士	徳永隆治	
副 査	筑波大学 教授	博士 (工学)	狩野 均	
副 査	筑波大学 准教授	博士 (工学)	滝沢穂高	
副 査	筑波大学 准教授	博士 (工学)	天笠俊之	
副 査	筑波大学 助教	博士 (工学)	乾 孝司	

論 文 の 要 旨

本論文では、ウェブ検索エンジンを高度化するために必要なテキスト種別特定に関する2つのアルゴリズムを提案し、有効性を実証した。

一つ目の提案は、ウェブページのコンテンツ並びにポスト及びコメントを自動抽出するアルゴリズムである。「コンテンツ部分はウェブサイト内で唯一の存在であり、繰り返し出現する部分は不要部分である」というアイデアを用い、コストがかかる人手による抽出ルールの作成や、機械学習のための正解データ作成を行わず、HTMLをテキスト処理するだけで分離抽出を行う手法を提案した。評価実験では、従来手法及び従来手法を組み合わせた手法と比較を行い、提案手法の方が高精度に抽出できることを示した。

二つ目の提案は、突発的かつ集中的に検索されるクエリを対象としたカテゴリ特定アルゴリズムである。突発的に増加する検索クエリは、何らかの社会的イベントが原因となることが多いが、初期段階でそのイベントとの関連を推定することはこれまで困難であった。提案手法では、比較的すばやく社会的イベントに反応するソーシャルメディアのデータを利用することにより、初期段階でのカテゴリ特定を可能としている。評価実験では、検索のピーク日に識別を試みるという実用的な環境において、ニュースデータやクリックログを利用する手法と比較して、提案手法が高精度にカテゴリ特定を行えることを示した。

審査の要旨

【批評】

本論文では、ウェブ検索エンジンの高度化に必要な技術の中でも特に重要かつ困難な課題に挑み、オリジナリティの高いアイデアで2つの課題を解決している。一つ目の提案では、コストの高い人手による関与をほとんどなくし、ウェブページの集合を与えるだけで自動的にコンテンツの分離抽出を行える手法を考案している。劇的なコストダウンに成功しており、今後標準的な手法として普及する可能性が期待できる。二つ目の提案では、ソーシャルメディアの即時性を利用し、これまで困難とされていた検索クエリのピーク日のカテゴリ特定に成功している。提案手法は、新しいアプローチとして本分野の技術発展に寄与することが期待できる。

本論文における提案手法はウェブ検索エンジンの要素技術であり、今後実際に検索エンジンに組み込みその有効性を実証する必要がある。しかし、要素技術としてのオリジナリティと完成度は高く、検索エンジンだけでなくその他の分野にも影響を与えうる潜在的価値を持っており、高く評価できる。

【最終試験の結果】

平成26年2月3日、システム情報工学研究科において、学位論文審査委員の全員出席のもと、著者に論文について説明を求め、関連事項につき質疑応答を行った。その結果、学位論文審査委員全員によって、合格と判定された。

【結論】

上記の学位論文審査ならびに最終試験の結果に基づき、著者は博士（工学）の学位を受けるに十分な資格を有するものと認める。