氏名	Bui (ブイ) Quang (クアン) Hung (フン)
博士の専攻分野の名称	博士（工学）
学位記番号	第　２２４９９　号
学位授与年月日	平成20年９月25日
学位授与の要件	学位規則第４条第１項該当 基礎工学研究科システム創成専攻
学位論文名	A Study of Extraction of Anchor-Related Text and Its Application for Ranking Web Pages （アンカー関連テキストの抽出とWebページのランキングへの応用に関する研究）
論文審査委員	（主査） 教授　西田　正吾 （副査） 教授　新井　健生　　教授　飯國　洋二　　講師　土方　嘉徳

論文内容の要旨

This doctoral dissertation describes a fundamental method for extracting anchor-related text and an application of anchor-related text for ranking web pages which is one of the major techniques of Web Information Retrieval.

With the exponential growth of information on the World Wide Web (Web), we are facing the information overload problem. Web Information Retrieval is a research area which tries to solve the information overload problem. The major techniques of Web Information Retrieval are web page categorization, web page summarization, and Web page ranking. There are three approaches for Web Information Retrieval techniques: content analysis, structure analysis, and usage analysis. These approaches are different in what kind of Web data they exploit for realizing Web Information Retrieval techniques. The content analysis approach exploits the contents of Web pages. The structure analysis approach exploits the link structure of the Web. The usage analysis approach exploits the information generated by interactivities between Web users and the Web. This dissertation focuses on the structure analysis approach.

One of the most important characteristics of the link structure of the Web is that authors of original pages create links because they think the links are useful for the Web users. A link from an original page to a target page can be seemed as a recommendation about the target page by the author of the original page. The author also writes some texts around the anchor to explain the target page to the Web users from his own viewpoint. These texts are semantically related to the target page. Recent researches have shown that these anchor-related texts can be used for Web Information Retrieval techniques.

This research studies the issue of extraction of anchor-related text and the applications of anchor-related text for Web Information Retrieval techniques. We first have a definition of Semantic Text Portion (STP). A STP in an original page is a text

portion which is semantically related to the anchor pointing to the target page. We propose a method for extracting STPs. We then investigate the effectiveness of using STPs for ranking web pages. The main contents of this doctoral dissertation consist of 2 parts. The first part describes the investigation of STPs and our proposed method for extracting STPs which is based on the result of the investigation. We also evaluate our method from the viewpoint of extracting anchor-related text. Experimental results show that our proposed method extracts STPs in high accuracy. The second part investigates the effectiveness of using STPs for improving the HITS algorithm which is a popular algorithm for ranking web pages. We first study how much we can improve the HITS algorithm using STPs. Then we compare STPs with other kinds of anchor-related text from the viewpoint of improving the HITS algorithm. Experiment results show that STPs are best for improving the HITS algorithm.

論文審査の結果の要旨

Webの出現以来，Web上で公開される情報の量は増え続けている．大量の情報から，必要な情報を獲得することが困難な問題は，深刻なものとなっている．Webではページ同士がリンクで結ばれており，リンク元ページにはリンクが埋め込まれているアンカーが存在する．近年，アンカーの周辺テキストを用いた検索・探索行動の支援に関する研究が盛んに行われている．これまでの研究は，ある特定のルールで抽出したテキストをリンク先ページに関連のあるテキストとして用いているが，リンク元ページ中のどの部分がリンク先ページと意味的に関連しているかについて，詳しく調査を行った研究はなかった．

本論文は，上記の調査を初めて行った研究である．第１章では，これまでアンカー関連テキストがどのような技術に用いられてきたかを説明している．第２章では，アンカー関連テキストがリンク元ページ中のどこに存在するかを調査した実験，提案するアンカー関連テキストの抽出方法，前記抽出方法によるアンカー関連テキストの抽出実験について述べている．この調査は，無作為に獲得した1000ページ以上のWebページに対して，どこがアンカー関連テキストかを調査している．この調査結果に基づき，アンカー関連テキストを抽出する方法を提案している．評価実験では13もの既存の手法と比較を行い，高い精度と再現率で抽出可能なことを示している．

第３章では，代表的なWebページの検索結果のランキングを行うアルゴリズムであるHITSアルゴリズムに対して，提案手法で抽出したアンカー関連テキストを適用している．HITSアルゴリズムでは，Webのリンクに重みを付けて，反復的にWebページの重要度を計算するが，このリンクの重み付けにアンカー関連テキストを用いている．これも従来の既存の手法と比較を行い，従来手法よりも高い精度でランキングを行えることを示している．第４章では，アンカー関連テキストについて深く分析することが，より性能の良い検索支援システムを構築できる可能性について言及している．

以上のように，本論文は，Webの情報獲得技術に共通して使える基本技術を開発しており，さらに検索エンジンのランキングにおいて高い有用性を持つことを示している．本論文を，博士（工学）の学位論文として価値のあるものと認める．