様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成 25 年 6 月 1 日現在

機関番号：３４３１０
研究種目：若手研究（B）
研究期間：2010～2012
課題番号：22700248
研究課題名（和文）　Web データに対する情報検索における情報単位に関する研究

研究課題名（英文）　A Study on Information Unit in Information Retrieval for Web Data

研究代表者
波多野　賢治（HATANO　KENJI）
同志社大学・文化情報学部・准教授
研究者番号：80314532

研究成果の概要（和文）：本研究では，多くの Web 検索エンジンが検索結果の表示形式にとらわれず，Web データに対する情報検索における新しい検索結果の形式を Web 文書間のリンク構造，Web 文書内の文書構造，そして利用者が入力する情報要求から算出される索引語の重みにより決定する方法の提案を行った．

本研究の成果は二つあり，一つは索引語の重みを正確かつ高速に再計算可能なアルゴリズムの提案，もう一つは誰もが容易に使用可能な情報アクセス機構の開発を行ったことである．その結果，索引語の重み再計算コストを約 64%に減らすことができ，またどの文書のどの部分に利用者の情報要求を満たす部分が存在するのかを判断できる有用なツールとなり得ることが判明した．さらに，ドメイン内検索においてはこのような提示形式が非常に有効に機能することが判明し，本研究を行う意義は十分にあったということができる．

研究成果の概要（英文）：Current Web search engines are usually return a search result as a list of Web pages. In this study, however, we try to propose a novel method for indicating “Information Unit” as a new search result of the Web search engines.

The research outcome of this study is composed of two parts. One is an efficient algorithm for calculating weights of index terms in Web pages. This has only a small effect in the time cost. The other is an information access tool which can help users to access and understand information sources easily. As a result, it can be said that we can perform this highly meaningful research in the research field of developing an efficient and effective search engine.

交付決定額　（金額単位：円）

	直接経費	間接経費	合　計
2010 年度	900,000	270,000	1,170,000
2011 年度	1,000,000	300,000	1,300,000
2012 年度	1,000,000	300,000	1,300,000
総　計	2,900,000	870,000	3,770,000

研究分野：総合領域
科研費の分科・細目：情報学・図書館情報学・人文社会情報学
キーワード：情報検索，表示方式，情報単位

１．研究開始当初の背景

Web 検索エンジンを支える要素技術は，検索エンジン開発当初のような URL のリスト表示だけではなく，その隣にスニペットをも表示されるようになるなど，利用者の要求に耐えうるものになりつつある．しかしながら，検索結果として検索エンジンから表示されるものは，未だ該当 Web 文書に関するデータが単にリスト形式で表示されたものであるため，利用者はそれぞれの Web 文書に一つ一つアクセスすることによって利用者自身が欲している情報かどうかを判断しなければならない．

このような形式では，それら Web 文書のどの部分が利用者の欲している情報なのかを利用者自身が瞬時に把握することが難しく，特に末端利用者にとって検索エンジンの使いにくさを感じる大きな原因の一つとなっている点は否めない．このような末端利用者が感じる検索エンジンの使いにくさは，多かれ少なかれ，検索エンジン開発者は誠実に対処していかねばならない問題である．

２．研究の目的

本研究では，多くの Web 検索エンジンが検索結果の表示形式として採用している「検索結果＝Web 文書のリスト」という従来の枠組みにとらわれず，Web データに対する情報検索における新しい検索結果の表示形式である情報単位を，Web 文書間のリンク構造，Web 文書内の文書構造，そして利用者が入力する情報要求から算出される索引語の重みにより決定し，それらを元に表示する方法の提案を行う．

この提案により，Web 検索エンジンの検索結果の表示形式は「検索結果＝情報単位のリスト」となり，従来の Web 検索エンジンのように利用者の情報要求に合致する情報を得るために複数の Web 文書を何度もアクセスする必要がなくなる．したがって，これまで以上に利用者に使い易い Web 検索エンジンを提供することができるようになることが期待される．

３．研究の方法

研究期間の一年目は手始めとして「検索結果＝情報単位のリスト」という形式を追求するために，情報単位の重要度計算アルゴリズムの開発を行った．二年目はリスト形式による表示よって利用者が抱える問題の一つである検索結果の見通しの悪さを解決するための提示手法として，クエリとして入力された内容に合致する部分をハイライト表示する方法の提案を，そして最終年度は前年度の提案を実際の Web 検索システムに実装し，実運用に耐えうるものかどうかの評価実験を行うことで，遂行するに値する研究を行ってきたかの評価を行った．

４．研究成果

「検索結果＝情報単位のリスト」という新しくかつ複雑な表示形式を使用する以上，検索対象となる Web 文書数が膨大になることは必然的である．そのため情報単位ごとの索引語の重み計算や新しい Web 文書の追加による索引語の重み再計算には，非常に大きなコストがかかるという問題が生じる．本研究の遂行により，従来の計算手法と比較して索引語の重み計算に要する時間コストは約 64%にまで軽減されたが，それでも検索精度が若干低下するなど未だ解決出来ていない問題も残っている．この根本的な問題の解決には一台の計算機ですべてを処理するのではなく，分散問合せ処理化することによる

高速化が今後必要となってくる．

その一方で，提案した情報単位表示方式により，入力したクエリキーワードを基準として Web テキスト上に着目すべき部分がハイライト表示されるようになったため，末端利用者であってもどの文書のどの部分に利用者の情報要求を満たす部分が存在するのかを容易に判断できるようになり，容易な情報アクセスのための有用なツールとなり得ることが明らかになった．

さらに，このような提示形式を使用できるような場面として，例えばある組織内だけで公開されている Web 文書，すなわちその組織内 Web 文書の検索をするというドメイン内検索のような場面においては，非常に有効に機能することも判明し，本研究を行う意義は十分にあったということができる．

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 5 件）

① K. Tamura, K. Hatano, and H. Yadohisa, A Retrieval Method Based on Language Model Considering Neighboring Contents, Journal of Digital Information Management (JDIM), reviewed, Vol.10, No.1, 2012, pp.1-9.

② A. Keyaki, K. Hatano, and J. Miyazaki, Result-reconstruction Method to Return Useful XML Elements, International Journal of Web Information Systems (IJWIS), reviewed, Vol.7, No.4, 2011, pp.360-380, DOI: 10.1108/17440081111187556.

③ S. Kitahara, K. Tamura, and K. Hatano, Extraction of the Contents in the Web Texts by Content-Density Distribution, International Journal of Knowledge Engineering and Soft Data Paradigms, reviewed, Vol.3, No.2, 2011, pp.108-120, DOI: 10.1504/IJKESDP.2011.045723

④ 欅惇志，波多野賢治，宮崎純，有益な検索結果提示のための部分文書再構成手法の提案，情報処理学会論文誌：データベース，査読有，Vol.4, No.1, 2011, pp.1-13.

⑤ A. Keyaki, K. Hatano, and J. Miyazaki, A Query-oriented XML Fragment Search Approach on A Relational Database System, Journal of Digital Information Management (JDIM), reviewed, Vol.8, No.3, 2010, pp.175-180.

〔学会発表〕（計 20 件）

① S. Kitahara and K. Hatano, A Report on the Size of Information Unit to Extract Contents on the Web text, International MultiConference of Engineers and Computer Scientists 2013 (IMECS2013), 2013-03-15, Hong Kong, China.

② 欅惇志，宮崎純，波多野賢治，山本豪志郎，武富貴史，加藤博一，更新を考慮した XML 部分文書検索システムの精度の改善，第 5 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)，2013-03-03, 福島県.

③ 北原沙緒理，波多野賢治，Web ページにおけるハイパリンクを考慮した内容抽出とその評価，第 5 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), 2013-03-03, 福島県.

④ 嫁兼弘修，北原沙緒理，波多野賢治，レ

ファレンスデータを用いた情報探索過程段階化手法の検討, 第 5 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), 2013-03-03, 福島県

⑤ A. Keyaki, J. Miyazaki, K. Hatano, G. Yamamoto, T. Taketomi, and H. Kato, Fast and Incremental Indexing in Effective and Efficient XML Retrieval Systems, 14th International Conference on Information Integration and Web-based Applications & Services (iiWAS 2012), 2012-12-05, Indonesia.

⑥ 欅惇志，宮崎純，波多野賢治，山本豪志郎，武富貴史，加藤博一，XML 部分文書検索における索引の高速な差分更新と高精度検索，WebDB Forum，2012-11-20, 東京都．.

⑦ 北原沙緒理，波多野賢治，単語位置を考慮した単語単位で行う Web テキストの内容抽出に対する一考察, 平成 24 年度情報処理学会関西支部支部大会，2012-09-21，大阪府.

⑧ 欅惇志，宮崎純，波多野賢治，山本豪志郎，加藤博一，XML 索引の更新コスト削減のための部分文書の統計量に基づくフィルタの評価とその最適化, 第 4 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), 2012-03-03, 兵庫県.

⑨ S. Kitahara, K. Tamura, and K. Hatano, Extraction of Web Texts using Content-Density Distribution, 7th Asia Information Retrieval Societies Conference (AIRS2011), 2011-12-19, UAE.

⑩ 欅惇志，宮崎純，波多野賢治，山本豪志郎, 加藤博一, XML 情報検索のための動的な索引管理手法の一提案，情報処理学会 DBS 研究会，2011-11-03, 東京都.

⑪ A. Keyaki, K. Hatano, and J. Miyazaki, Relaxed Global Term Weights for XML Fragment Search, iDB Workshop 2011, 2011-08-01, Kyoto Pref., Japan.

⑫ 白井涼子，欅惇志，波多野賢治，RDF を利用した和歌データの管理に関する提案, 情報処理学会 DD 研究会，2012-03-26, 東京都.

⑬ K. Tamura, K. Hatano, and H. Yadohisa, Calculating Query Likelihood based on Web Data Analysis, 3rd International Conference on Intelligent Decision Technologies (KES IDT 2011), 2011-07-20, Greece.

⑭ 白井涼子，欅惇志，波多野賢治，和歌データの構造化とその格納手法の一考察, 電子情報通信学会 DE 研究会，2011-06-06, 神奈川県.

⑮ 北原沙緒理, 田村航弥, 波多野賢治, Web テキストにおける内容密度分布の抽出とその評価, 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，2011-02-27，静岡県.

⑯ 田村航弥，波多野賢治，宿久洋，リンク情報に基づく周辺文書の索引語尤度を考慮した文書検索手法の提案と評価，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，2011-02-27，静岡県.

⑰ A. Keyaki, K. Hatano, and J. Miyazaki, Result Reconstruction Method for Effective XML Fragment Search at INEX 2010, INEX 2010 Workshop, 2010-12-13, Netherlands.

⑱ A. Keyaki, K. Hatano, and J. Miyazaki, Result Reconstruction Approach for More Effective XML Fragment Search, 12th International Conference on Information Integration and

Web-based Applications & Services (iiWAS 2010), 2010-11-09, France.

⑲ 欅惇志，波多野賢治，宮崎純，再構成された XML 部分文書に対するランキング手法の提案，電子情報通信学会 WI2 研究会，2010-09-16，新潟県

⑳ K. Tamura, K. Hatano, and H. Yadohisa, Characterizing Web Pages based on the Query Likelihoods of Neighboring Pages, The 5th International Conference on Digital Information Management (ICDIM2010), 2010-07-07, Canada.

６．研究組織

(1) 研究代表者

波多野　賢治（HATANO KENJI）

同志社大学・文化情報学部・准教授

研究者番号： 80314532

(2) 研究分担者

（　　）

研究者番号：

(3) 連携研究者

（　　）

研究者番号：