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 24 年 3 月 31 日現在

機関番号：24506
研究種目：若手研究(B)
研究期間：2010～2011
課題番号：22700108
研究課題名（和文）　把握容易性に基づく手法掲載ページの検索とランキング

研究課題名（英文）　Search and Ranking of How-to Information Pages on the Basis of Understandability

研究代表者
湯本　高行（YUMOTO TAKAYUKI）
兵庫県立大学・大学院工学研究科・助教
研究者番号：20453152

研究成果の概要（和文）：本研究では，料理のレシピやソフトウエアのインストール方法などの手法情報を対象とした Web ページの検索とランキング方法について研究を行った．まず，手法情報に共通する表現および HTML の構造に着目して手法情報の掲載の有無と該当部分を発見すると共に，文単位の分析により動作と対象のペアとして手法の構成要素を抽出する方法を開発した．さらに，重要語や画像の出現状況によって把握容易度を定義し，これによるランキング方法を開発した．

研究成果の概要（英文）：In this research, we developed searching and ranking method for how-to information pages such as recipes for dishes and installation procedure of software. Firstly, we focused on expressions and HTML structures which are common in how-to information, and developed algorithm to detect how-to information and extract how-to information parts. We express elements of how-to information as pairs of operations and targets, and extract them by analyzing sentences in how-to information. We also defined understandability based on appearance of important words and images, and developed the ranking algorithm using it.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2010 年度	1,000,000	300,000	1,300,000
2011 年度	800,000	240,000	1,040,000
総　計	1,800,000	540,000	2,340,000

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，WWW

１．研究開始当初の背景

ユーザが望む情報をより見つけやすくするためにさまざまな検索技術が研究・開発されており，特定の用途に特化した vertical search engine(専門検索エンジン)および検索方式が開発されている．検索の用途の１つに「手法」の検索がある．本研究における「手法」とは，何かをつくったり，行ったりするための方法の説明全般を意味する．この手法情報の検索に対する需要の大きさは，中村らによって行われた情報検索に対する信頼性に関する調査の結果からもわかる．この調査は，2006 年に 20～50 代の男女 1000 人に対して実施されたアンケートである．このアンケートの中の「Q4.検索した時に知ろうとしているもので多いものは何ですか」という問いに対して，多いと思う順に 5 位までを回答者が答えた結果を集計したところ，「検索キーワードの意味・定義」，「人・組織」に次いで，「方法・手段」が 3 位となっており，回答者の 50%以上が「方法・手段」を検索することが多いと回答している．この調査におけ

る「方法・手段」は本研究で「手法」と呼んでいる概念とほぼ等価であり，このことから手法検索に対する需要は非常に大きいと言える．

手法の検索に関連する研究としては，レシピ検索があるが，手法情報にはレシピだけではなく，ソフトウエアのインストール，折り紙の折り方などさまざまなドメインが存在する．それにも関わらず，さまざまなドメインの手法に対して汎用的に適用できる検索方式が，研究開始当初は確立されていなかった．そこで，ドメインに依存せずに，手法に特化した検索方式を確立することが重要であると考えた．

２．研究の目的

手法情報の検索は，手法情報の収集，手法掲載部分の抽出，手法のランキングの 3 段階で実行することを想定した．手法情報の収集については，クエリ拡張による既存研究があり，本研究では対象としない．次に，手法掲載部分の抽出については Web ページからの本文抽出や箇条書き部分の抽出などの既存研究があるが，手法掲載部分や手法の構成要素を抽出するには不足している．さらに手法のランキングについてであるが，手法情報の検索においてはわかりやすさを考慮したランキングが重要であるが，Web ページのわかりやすさを自動的に判断する方式は確立されていない．

そのため，本研究では，手法部分の抽出と手法のランキングについて，「(1)手法の構成要素の抽出方法」と「(2)手法情報の把握容易性に基づくランキング」に取り組む．これによって，手法の把握しやすさに基づく新たな検索方式の実現を目指す．

３．研究の方法

(1)手法の構成要素の抽出方法の開発

手法の構成要素の抽出について以下の 2 つの課題に取り組む．

①　手法掲載部分の抽出

Web ページから手法情報が掲載されているかどうかを判定し，掲載されている場合は，その部分を抽出するアルゴリズムを開発する．

②　手法の構成要素の抽出

手法の構成要素をモデル化し，上記の①で抽出した手法掲載部分に対して文単位で分析を行うことで構成要素を抽出するアルゴリズムを開発する．

(2)手法情報の把握容易性に基づくランキング方法の開発

把握容易性の尺度として，概要を迅速に把握するのに役立つかどうかを表す概要把握と詳細を深く把握するのに役立つかどうかを表す詳細把握に分け，ランキングの指標を開発する．さらにこれを利用して，手法情報の検索システムのプロトタイプシステムを開発する．

４．研究成果

(1)手法の構成要素の抽出方法の実現

①　手法掲載部分の抽出アルゴリズム

手法を説明している文章に共通している，内容的な特徴と HTML の構造的な特徴を用いて手法について説明している部分の特定と抽出を行うアルゴリズムを開発した．このアルゴリズムでは，まず，HTML のブロック要素に注目して，ページを分割する．次に，分割した領域内に手法に共通する特徴が存在するかを，(a)手順ごとに番号がふられている，(b)「まず」，「次に」などの順序を表す語が使われている，(c)過去形の文の割合が少ない，の 3 条件を用いて検証する．なお，(a)と(b)はどちらか一方を満たすものとし，(c)は必ず満たすものとする．最後に，手法情報では，手順が HTML 構造内で並列関係になることが多いことに注目し，手法情報を含む部分木を複数，子に持つノードが存在する場合，各部分木が手順に対応しているとし，そのノードを根とする部分木を手法掲載部分とみなす．これによって，図 1 に示すような手法について説明していない箇所の誤検出を訂正することが可能になった．

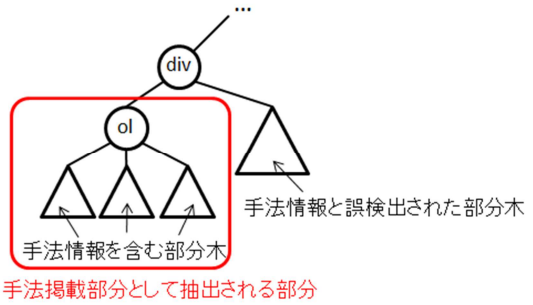

図 1　手法掲載部分の抽出

②　手法の構成要素の抽出アルゴリズム

手法の構成要素は「動作」と「対象」のペアとして表現する．抽出は，上記の①で抽出した手法掲載部分に対して文ごとに述語とその対象となる目的語を発見し，それぞれを「動作」と「対象」とすることによって行う．具体的には，まず，文末表現に注目し，過去形表現や状態を述べている文，感想を述べている文などを除外して消去法によって手法について説明している文(説明文)を発見する．次に説明文を形態素解析し，文末からさかのぼって分析を行い，各文の述語を発見すると共に格助詞に注目して目的語を抽出する．それぞれを「動作」と「対象」とする．たとえば，「じゃがいもを切る」という文では，「切る」が「動作」，「じゃがいも」が「対

象」として抽出され，(切る，じゃがいも)というペアが抽出される．最後に，目的語が省略されているなどの理由で「対象」が抽出できない文については，直前の文から「対象」を補う．これによって，　手法の手順を最小単位に分割することが可能になり，手法情報間の比較が容易になった．

(2)手法情報の把握容易性に基づくランキング方法の実現

把握しやすさは文章による説明と説明に用いられる画像がそれぞれ関係していると考えた．そのため，概要把握と詳細把握のそれぞれで文字情報と画像情報にを用いた指標を定義した．

まず，概要把握については，手法掲載部分に含まれる画像の数を手法掲載部分の集合内の最大数で正規化したものと手法掲載部分に含まれる重要語の網羅度の平均を指標u_oとして定義した．なお，重要語については，手法掲載部分の集合中でその語を含んでいる手法掲載部分が多いものをその手法における重要語とした．

また，詳細把握については，手法掲載部分内で画像を含んでいる手順の割合と重要語の出現数を手法掲載部分の集合内の最大数で正規化したものの平均を指標u_dとして定義した．

概要把握と詳細把握の指標を以下のように統合して把握容易度とする．

$$u = \alpha u_o + (1-\alpha) u_d \quad \cdots \text{(A)}$$

なお，αは$0 \leq \alpha \leq 1$の範囲で，ユーザが検索時に自由に決定および変更できるパラメータである．

この把握容易度を用いたランキングと既存の検索エンジンである Google の検索結果と比較実験を行った．実験には，料理のレシピやソフトウエアのインストール方法など 10 種類のクエリを用いた．実験では，7 名の大学院生と学部生がクエリごとに検索結果の Web ページを「大まかな流れが把握しやすい順」と「詳細が把握しやすい順」に並び替えた結果を作成した．前者は概要把握の評価に用い，後者は詳細把握の評価に用いた．これらのランキングと提案法によるランキング，Google によるランキングに対してケンドールの順位相関係数を求め，正の相関がある被験者の割合によって評価を行った．10 種類のクエリの平均は表 1 のようになり，提案法によるランキングは Google のランキングよりも良好な結果となった．

また，被験者間でクエリごとの並べ替えの結果の相関をとったところ，相関が低くなったり，弱い負の相関が見られる場合もあった．そのため，わかりやすいと考える基準は，人およびクエリによって異なっていることがわかった．

表 1 ランキング結果の評価

	提案法	Google
概要把握	0.73	0.46
詳細把握	0.70	0.51

さらに，図 2 に示すようなプロトタイプシステムを開発した．このプロトタイプシステムでは，ユーザが式(A)のパラメータ α を自由に変更してランキングを変更できる．

図 2 手法情報検索のプロトタイプシステム

また，各指標の重みを動的に決定するアルゴリズムを開発した．具体的には，手法掲載ページ集合内の各ページの指標を用いて主成分分析を行うことにより，重みを決定した．指標としては，手法の構成要素の抽出によって利用可能になった指標を合わせ，画像数，手順数，構成要素数，重要語の網羅度，画像網羅度，手順あたりの平均構成要素数を検討した．

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）

〔学会発表〕（計 4 件）

① 湯本高行，“ハウツー情報間の共通性に基づく検索結果の提示手法”，第 22 回電子情報通信学会 Web インテリジェンスとインタラクション研究会，査読無，東京都文京区，2012 年 3 月 18 日，pp. 83-84.

② 湯本高行，“Web 上のハウツー情報からの構成要素の抽出”，第 153 回情報処理学会 データベースシステム研究発表会，査読無，東京都新宿区，2011 年 11 月 3 日，A-1-3.

③ Ryoji Nonaka, Takayuki Yumoto, Manabu Nii, Yutaka Takahashi , “Finding

How-to Information Web Pages and Their Ranking by Readability”, IADIS International Conference on Internet Technologies & Society 2010, 査読有, Perth, Australia, 30 November 2010, pp.155-163.

④ 野中諒志, 湯本高行, 新居学, 高橋豐, “概要・詳細の見やすさに基づく手法情報のランキングと閲覧支援”, WebDB Forum 2010, 査読有, 東京都新宿区, 2010 年 11 月 11 日, 2A-1.

６．研究組織

(1)研究代表者

湯本　高行（YUMOTO TAKAYUKI）
兵庫県立大学・大学院工学研究科・助教
研究者番号：20453152