科学研究費助成事業　　研究成果報告書

平成　２６　年　　６　月　２２　日現在

機関番号：　３２６８５
研究種目：　若手研究(B)
研究期間：　2011～2013
課題番号：　２３７００１０７

研究課題名（和文）視覚的コンテキストによる検索結果の提示とナビゲーション

研究課題名（英文）Search Engine Result Page with Visual Context and WWW Navigation

研究代表者
丸山　一貴（Maruyama, Kazutaka）

明星大学・情報学部・准教授

研究者番号：３０３７７０１４

交付決定額（研究期間全体）：（直接経費）　　3,300,000円、（間接経費）　　990,000円

研究成果の概要（和文）：web検索結果ページの改善として、web検索結果として示された各ページを表示し、検索キーワードをハイライトした画像（VPと呼ぶ）だけを用いて、検索結果ページを構成する方法について研究を行った。
従来のVP生成手法の問題点を解決して、検索APIサービスに頼らず通常の検索ページを利用してVPを生成するとともに、VPから所望の検索結果を発見しやすくするため、検索キーワードのHTMLタグ階層に基づいた分類を試みた。複数のニュース記事から生成されたVPが同一グループに分類されるなど、有効な場合があることを確認した。

研究成果の概要（英文）：To improvement search engine result pages, we built a method of generating such pages including only Visual Patches (VP), which are clipped images from search result pages around search queries.
We resolved some problems of our prior system to generate VPs from normal user interface of Google search instead of web search API services. We also group VPs together based on HTML DOM hierarchy of search queries in search result pages. In some cases, the grouping method produced a useful result. For example, VPs from more than one news sites were put together.

研究分野：　情報推薦、ユーザインタフェース

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　ユーザインタフェース

１．研究開始当初の背景

一般的な検索エンジンの検索結果ページ（以下、SERP: Search Engine Result Page という）では、検索キーワード（以下、クエリという）が含まれる web ページのタイトルと URL に加えて、文字ベースのスニペットが出力される。SERP の出力要素の中では、多くのユーザがスニペットに注目することが知られている。しかし文字ベースのスニペットでは、同じ字面で他の意味を持つ言葉は除外しやすいが、ページの中でそのクエリがどのように扱われているか（見出し語か、関連語句一覧の一部か、広告の一部か、等）は分からない。結果として、ユーザは SERP で表示されるページタイトルや URL も含めてページの妥当性を判断しなければならない。そうして開いたページであっても所望の情報を含まないことは多く、ページ遷移により SERP へ戻って別のページを確認する作業はユーザの負荷が高い。SERP の段階で、どのページを選択すべきかは明白に示されるべきであり、より情報量の多いスニペットを提示することが必要である。

２．研究の目的

(1) web 検索エンジンの SERP において、ページを実際にレンダリングした上で検索キーワード周辺を切り取って提示する、視覚的コンテキストに基づくスニペット（以下、VP: Visual Patch という）を用いて結果を提示することを目的とする。

(2) VP をクラスタリングによって分類し、ページという構成単位にとらわれない検索結果提示とナビゲーション基盤の実現を目指す。

３．研究の方法

(1) サーバ・クライアント双方で通常のブラウザを動作させる従来の実装方式に代わる、実装及び運用負荷の少ない実装方法として、レンダリングエンジン等による実装を検討した。検討すべき課題が主に 2 つ挙げられる。第 1 は検索サービスの利用方法である。従来方式では、根本となる web 検索を Yahoo!社が提供する検索 API サービスに依存していたが、当該サービスは 2013 年に終了となり継続利用できなくなった。他社も同様に検索 API サービスを終了させてきた経緯があることから、API サービスには頼らず、ユーザが手動で入力することを想定した通常の検索サービス（例えば、Google 社の検索のトップページ）を利用する方法に切り換えることとした。第 2 はサーバ上で検索結果である web ページ群から VP を生成する方法である。従来は GUI を持つ通常のブラウザに追加するアドオンを開発して実現してきたが、ブラウザのバージョンアップが頻繁であり継続的に追随することが困難であった。実行時のサーバ負荷を下げる目的でブラウザのレンダリングエンジンのみを直接利用する方法も試みたが、バージョンアップへの追随も含め、開発コストが高いことから不適当であった。そこで、web アプリケーションの自動テストに用いられているヘッドレスブラウザである CasperJS を利用して、web 検索から VP 生成までを自動化した。

(2) 従来方式では、VP はその元となったページに紐付けて表示する方式をとっており、その SERP から発生するナビゲーションは web の情報がページ単位で構成されていることに依存してしまう。本研究では VP の発生元ページに限らず、複数のページをまたいで VP を分類するための方法を検討した。(1)により VP を生成する際に、web ページに含まれるクエリが HTML 要素として持つ DOM 階層を用いて分類する。ページ内のコンテンツは HTML タグをルートとする木構造を構成しており、箇条書きの項目であれば<li>タグ、表の要素であれば<tr>等のタグが付く。この構造はページの表示を整えるためのものであるが、クエリを含む箇所がページ内でどのような位置づけにあるかを示しているとも考えられる。この仮定に基づいて、クエリを含む箇所から階層を遡り、直近の 3 つのタグ（以下、分類タグという）を用いて VP を分類することとした。その際、や は意味の構造に関連しないことが多いと考え、対象から除外した。

４．研究成果

(1) web 検索を利用・改善しようとする研究で持続的に利用可能な、システムの実装方法を確立した。本研究の成果に基づいて実装可能なシステムの構成図を**エラー！参照元が見つかりません。**に示す。従来方式では検索を行う部分に PHP プログラムを、検索には検索エンジン運営者が開発者向けに提供する検索 API サービスを、検索結果に含まれる web ページのレンダリングと VP の生成には GUI 付きブラウザと専用に開発したアドオンを

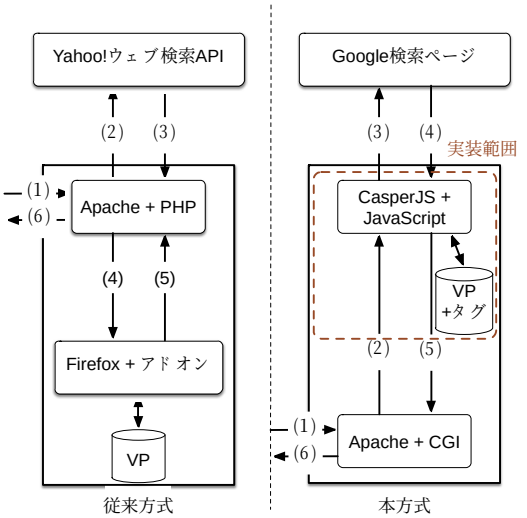

図 1　システム構成図

利用していた。研究の方法(1)で述べた通り、検索 API サービスに依存した実装方法では持続的に利用可能なシステムを実装できないので、ユーザが手動で操作することを前提としたインタフェースである通常の検索ページを API として利用することとした。本研究で実装した範囲は図 1 で破線により示した。CasperJS を制御する部分は JavaScript を用いて実装し、(i)外部から与えられるクエリを用いて Google 検索ページで検索を行い、(ii)SERP から検索結果の web ページを抽出し、(iii)各ページをレンダリングし、(iv)ページ内に含まれるクエリを VP として保存すると同時に、研究の方法(2)で述べた分類用タグを出力する。
本実装は、Google 検索ページや SERP の構成と、CasperJS に依存した実装となっている。Google のページ構成が変更される可能性はあり、その場合には(i)及び(ii)を実装し直す必要がある。しかし、検索 API のように検索ページそのものが利用不可能となることは考えられず、また、SERP も含めてそのページソースは公開される。また、インタラクティブな web サービスの重要性はますます高まっていくことから、CasperJS や、そのベースとなっている PhantomJS の開発が終了する懸念は低く、代替となるシステムの登場も十分予見されること。以上のことから、持続的に利用可能なシステムを構築できたと言える。

(2) 本実装のソースコードは再利用可能な形で公開しており、利用希望者が外部から検索要求を受け付ける CGI 部と合わせて実装することで、VP を含む独自の SERP を開発可能となる。我々は VP のみを用いた SERP を想定して研究を行っているが、従来通りの文字ベースのスニペットに対して VP を付加した SERP を実装することも可能となり、web 検索を対象とする研究者に広く貢献できるものとなった。
ただし、研究成果の(3)で述べる通り、VP 抽出・分類でクエリを含まない不適切な VP を除外する処理に問題があることも確認されており、多様な web ページを対象としながら改善していくことが必要である。

(3) VP を分類する方法として、研究の方法(2)で述べたタグによる分類を行い、有効な局面

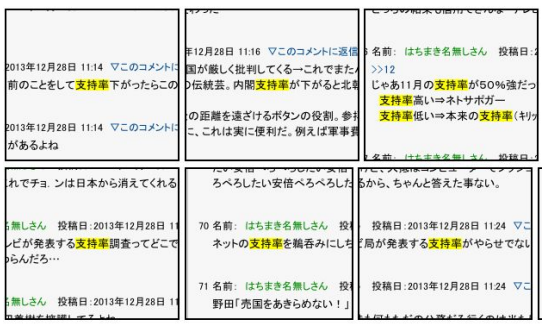

図 2 VP 分類例 1（最上位サイト関連）

図 3 VP 分類例 1（第 2 位サイト関連）

があることを示した。VP の分類に当たっては、まず VP をその元となった検索結果の web ページごとにまとめ、検索結果で最上位となったページの VP を出現順に並べる。その際、同じ分類タグを持つ VP は繰り上げて、まとめて表示する。さらに、第 2 位以降の web ページでも同じ分類タグの VP があれば、最上位ページの VP と同じ位置に繰り上げて、まとめて表示する。以降も同様にして VP を分類した。
「安倍総理 支持率」で検索した結果の一部を図 2 及び図 3 に示す。この場合の最上位サイトは匿名掲示板のまとめサイトであり、同じサイトの VP だけがまとめて表示される結果となった。他の検索結果には図 3 のようなニュースサイトも含まれていたが、分類タグが同一となる VP は存在しなかった。これに対して検索結果で第 2 位となったページはニュースサイトであり、そこで生成された VP と同じ分類タグを持つ VP は、第 3 位以降に登場する他のニュースサイトでも現れており、複数の web サイトから VP が集まる結果となった。
次に、「”surface 2” 寸法」で検索した結果の一部を図 4 及び図 5 に示す。このときの

図 4 分類例 2（最上位サイト関連、クエリ 1）

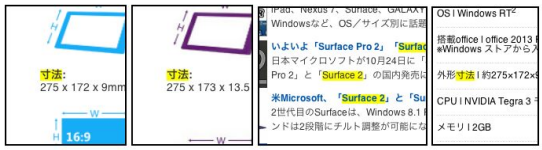

図 5 分類例 2（最上位サイト関連、クエリ 2）

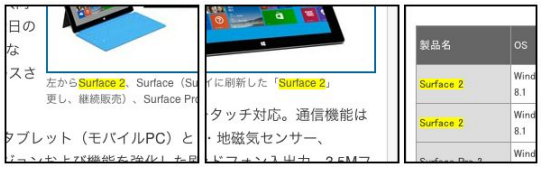

図 6　分類例 2（第 5 位サイト関連）

最上位サイトは Surface 2 のメーカである Microsoft 社の公式サイトであり、図 4 は 1 つ目のクエリである surface 2 をハイライトしており、公式サイト、Surface 2 発表時にニュースサイトが取り上げた記事、ニュースサイトの記事の端に並ぶ他の記事へのリンクが分類されている。左上の 3 つは公式サイトであるが VP 内にクエリが表示されていない。タイトルにクエリを含む場合など、不適切な VP を除外する処理が不十分なことが原因と考えられる。一方、最上位サイトに関連づけられた他の VP はニュースサイト記事であり、公式サイトに紐付けられた VP としては妥当なものであると言える。図 5 は 2 つ目のクエリである寸法をハイライトしており、公式サイトの仕様表にあたる部分、ニュースサイトのピックアップ記事の一覧、通信販売サイトの仕様表が近い場所に配置された。今回のクエリでは第 2 位から第 4 位のサイトに関しては対応する VP がほとんどなく、大部分が最上位サイトである公式ページに紐付けられる結果となり、サイト横断的に分類することに成功したと言える。図 6 に第 5 位サイトの VP と分類されたものの一部を示す。第 5 位のサイトはニュースサイトの紹介記事であり、インライン画像のキャプションの一部と仕様表が近くに配置されている。いずれも HTML の表として構成されているためであるが、画像に丁寧なキャプションが付けられていることから、機器の紹介記事であることが伺える。特定の IT 機器について調べている場合は紹介記事も有用である場合が多いと考えられ、一般的な文字ベースのスニペットよりも関連性を直感的に把握することができると言える。

ニュースサイトのピックアップ記事一覧が VP として多数現れる結果となったが、こうした VP が示す箇所には本来ユーザが欲しいと考えた情報は提示されていないと考えられるため、今後は順位を下げることや除外も検討する必要がある。その場合は、リンクが密集した箇所はインデクスの可能性が高いと言ったヒューリスティクスを導入することも有効であると考えている。

５．主な発表論文等

〔雑誌論文〕（計 2 件）

[1] 丸山 一貴，井桁 正人，寺田 実，“視覚的コンテキストによる検索結果提示とナビゲーションの可能性”，情報処理学会研究報告，査読無，Vol.2014-HCI-157, No.26, pp.1-6, 2014.

[2] Kazutaka Maruyama, Masato Igeta, Minoru Terada, “Search Engine Result Page with Visual Context and Already Rendered Snippets”, Proceedings of the 7th International Conference on Web Informational Systems and Technologies, 査読有, pp.340-345, 2011.

〔その他〕

ホームページ等

http://www.sanpo-lab.jp/~kazutaka/research/hci157/abstract.html

６．研究組織

(1)研究代表者

丸山 一貴（MARUYAMA, Kazutaka）

明星大学・情報学部・准教授

研究者番号：30377014