3C3-03

Web 検索結果の概観提示による情報収集支援インタフェース

Information Gathering Support Interface by the Overview Presentation of Web Search Results

小林拓海[*1] 佐藤大介[*1] 三末和男[*1] 田中二郎[*1]
Takumi Kobayashi Daisuke Sato Kazuo Misue Jiro Tanaka

[*1]筑波大学大学院システム情報工学研究科
Graduate School of Systems and Information Engineering, Univeristy of Tsukuba

It is hard work to retrieve necessary information from huge numbers of Web pages because the Internet consists of several billion Web pages. Traditional search engines which divide results into dozens of pages regardless of the genre and present them in the form of a text-based list is not necessarily useful.

In this paper, we propose the new visual interface based on Hyperbolic Tree, which promotes user's intuitive understanding of the entire Web search results. Our system classifies the Web search results and visualizes them on one screen.

1. はじめに

我々が日常の Web 検索を行なう際、特定のトピックに関する 1 件もしくは複数件の Web ページを獲得することを目的とした検索を行なう場合がある。例えば「福岡県」に関する様々なジャンルの情報を得ることを目的とした検索などである。このような検索を IBM の Andrei Broder は情報指向型検索と呼んでいる [Broder 02]。

情報指向型の検索を行なう場合、ユーザは膨大な数の Web 検索結果を様々な観点から眺め、必要な情報を取捨選択する必要がある。「福岡県に関する情報」という検索要求は抽象的であり、県政に関する情報、交通に関する情報、観光名所に関する情報など様々である。そのためユーザは特定の Web ページのみから情報を得るのではなく、様々な複数の Web ページを巡回しながら情報を収集することとなる。

以上の事柄を考慮すると、情報指向型検索を行う際に利用する一般的なキーワード検索インタフェースには様々な問題点が存在し、ユーザの情報収集を困難にしていると考えられる。それらで用いられているような検索結果をテキストによる一次元のリストでユーザに提示するインタフェースは様々なジャンルの Web ページが混在しているために、ユーザが効率よく情報収集できない場合がある。また、検索結果が数十ページにわたって分割される提示法は、検索結果全体の特徴を直観的に理解できず、有益な情報を見落としてしまうなどといった問題がある。さらに検索結果の中にはユーザの全く意図しない情報を含む Web ページが存在する場合がある。このような Web ページの存在も、ユーザの情報収集を困難にする要素となっていると考えられる。

そこで本研究では Web 検索結果をページの内容によって適切に分類し、分類結果にラベルを付加した情報を 1 画面に納めてユーザに提示するインタフェースを提案、試作した。本インタフェースを用いることでユーザは Web 検索結果の全体像を直観的に理解することが容易になり、必要な情報を効率よく取捨選択することが可能となる。

連絡先: 小林拓海，〒 305-8573 茨城県つくば市天王台 1-1-1
筑波大学大学院システム情報工学研究科
コンピュータサイエンス専攻
TEL/FAX 029-853-5165，takumi@iplab.cs.tsukuba.ac.jp

2. 概観提示インタフェースの提案

本研究では上記の問題点を解決する以下のような要件を満たす Web 検索結果提示インタフェースを提案する。

要件 1 類似ページを二次元空間において近傍に配置

要件 2 検索結果を一画面に納めてユーザに提示

要件 1 を満たし、検索結果を二次元空間を用いて提示することで、リストを順に見ていく必要がある一次元による提示インタフェースよりも提示表現の幅を広げることが可能となり、ユーザはより柔軟に Web 検索を行なうことができるようになる。また、類似ページを近傍に配置するために検索結果の Web ページをページの内容によってクラスタリングする。これにより様々なジャンルのページをあちこちに散在させることなく近傍に配置することで、ユーザはより効率よく情報を収集することができる。さらに、クラスタリングを行なうことにより、検索結果に含まれるユーザの意図しない Web ページもジャンルごとに近傍に配置されることになる。このため、意図しない Web ページの散在による情報収集の際の弊害を取り除くことができる。

要件 2 を満たすことで分類された検索結果は一画面に納まってユーザに提示される。このため検索結果が数十ページに分割されるインタフェースよりもユーザの検索結果の全体像理解が容易になる。また、情報を提示する際に、本システムでは分類した Web ページをそのまま提示するのではなく、分類結果にラベルを付加して提示する。ラベルはクラスタの特徴を表す単語で構成されている。ユーザはラベルを参考に必要な情報が存在すると考えられる複数の Web ページを容易に発見することができる。

3. 概観提示インタフェースの試作

本研究における概観提示インタフェースは、検索結果のクラスタリング部とクラスタリング結果提示部の 2 つの部分から構成されている。

3.1 検索結果のクラスタリング部

検索結果のクラスタリング部では、Web 検索エンジンによって与えられる検索結果である複数の Web ページを分析し、分析結果に基づいてクラスタリングを行う。

本システムはユーザに検索クエリを与えられると GoogleAPI を用いて複数の検索結果の URL を得る。次にシステムは URL に対応した HTML ファイルを取得し分析を行う。

HTML ファイルの分析は HTML ファイル中に出現する単語を形態素解析を用いて抽出し、tf/idf 法を用いて各 HTML ファイルをベクトル空間モデルで表現した。HTML ファイルは単純な文章ではなく、タグと呼ばれるコマンドを用いて木構造的に構成されている。我々はこの HTML ファイルの構造情報を積極的に利用することで分析対象である Web ページの特徴をより顕著に抽出することができると考えた。TITLE タグや H タグ、STRONG タグなどで修飾された部分は、Web ページの要点や作者が強調して表現したかった部分であるので、よりページの特徴を表す単語を含んでいると考えられる。本システムではこのような部分に出現する単語により大きな重みを与えている。また、META タグには Web ページ上には直接記述されないページの説明や特徴を表す単語が記述されている場合がある。さらに、FRAME タグや IFRAME タグを使用している Web ページの分析は、フレームやインフレームを参照している URL を得て同様に分析を行うことで適切な分析が可能となる。

次にシステムは分析された各文書ベクトルを元にクラスタリングを行う。クラスタリングには階層的クラスタリングを用いた [S.Everitt 93]。まずクラスタリング前の各 Web ページをそれぞれ 1 つのクラスタと見なす。分析によって求められた各 Web ページの特徴を表すベクトル同士の内積を比較し、最も内積値が小さい (最も類似している)2 つの Web ページを合成したものを新たにクラスタとして見なす。本システムではこのとき 2 つのクラスタ間において結びつきが強い上位 3 単語を求める。この単語はクラスタリング結果提示部において 2 つのクラスタのラベルを表現する単語となる。このような処理をすべての Web ページが 1 つのクラスタにまとまるまで繰り返すことで類似ページが近傍に配置された樹形図が完成することになる。

3.2 クラスタリング結果提示部

一般に Web 検索結果は膨大な数となるために、クラスタリングの結果として得られる樹形図もまた巨大なものとなる。そのような巨大な樹形図を 1 画面に納めてユーザに提示するために本システムでは Hyperbolic Tree[Lamping 95] を用いた。Hyperbolic Tree とは John Lamping らによって提唱された双曲空間上に樹形図を配置する手法である。

Hyperbolic Tree は中央に近いノードほど大きく、中央から遠いノードほど小さく表示される。また、本システムではマウスドラッグによって周辺部の部分木を中央に移動させることでフォーカスの移動が可能となっている。このような特徴のため、通常の樹形図よりも 1 画面に多くの情報を収めることが可能となり、ユーザは必要な情報にフォーカスを移動させることで概観を保ったまま必要な情報に注目することができる。

Hyperbolic Tree を用いてクラスタリング結果の樹形図を表示しただけではユーザが情報収集する際に必要な情報を十分に与えているとはいえない。そこで本システムでは、2 つのクラスタ同士において関連の強い上位 3 単語を 2 つのクラスタの親ノードとすることでユーザの情報収集を支援している。図 1 はクラスタ A,B,C と単語 1〜6 の関係を示す例である。

図 1 においてクラスタ A とクラスタ B は親ノードである単語 1, 単語 2, 単語 3 に対して強い関連があることを示している。さらにクラスタ A とクラスタ B からなるクラスタ AB(図 1 中の赤丸で囲まれた部分) とクラスタ C は、単語 4, 単語 5, 単語 6 に対して強い関連があることを示している。このようにクラスタの特徴を表すラベルを付加することで、ユーザはより効率的に情報を収集することが可能となるのである。

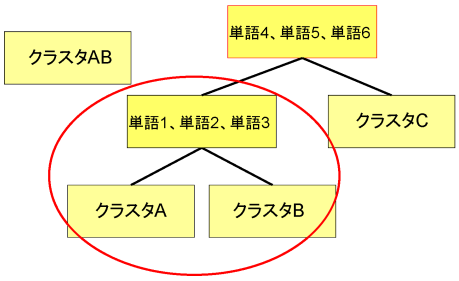

図 1: 地理に関する部分木

本システムのクラスタリング手法には、あまり類似していない Web ページ同士がクラスタを形成してしまい、樹形図が階段状になってしまう場合があるという問題点がある。この問題を表示インタフェース的に改善するため、それらの Web ページを「その他」というラベルをつけた 1 つのクラスタにまとめることで樹形図を変形してユーザにとってより見やすい提示画面を提供した。

4. 試作インタフェースの利用例

ユーザが日本の歴史について様々な観点から調べたいと考えて検索要求「日本　歴史」を用いて Web 検索を行なう場合を考える。この場合、もし一般の検索エンジンを用いて検索した場合、「日本」と「歴史」という単語が含まれるページで、検索エンジンが重要と判断したページから順に表示される。そのためにユーザは様々なジャンルが混在するテキストのリストを 1 つずつ順に見ていく必要がある。また、ユーザの意図とは違うページが多数含まれてしまう場合もある。この例では「日本の歴史」について調べたいというユーザの意図に反して意図していない「占い」のページが多数検索結果のリスト中に紛れ込んでしまう。一次元のリストを順に見ているユーザにとってこれは目障りであり、ユーザの情報収集を困難にしていると考えられる。

本システムに検索クエリ「日本　歴史」を与えた場合の提示画面を図 2 に示す。システムはアルゴリズムにしたがってクラスタリングし、結果をユーザに提示する。図 2 を見ると右下の部分木は「日本」と「歴史」という単語を含む占いについてのページが集合していることが分かる。ユーザは意図していない占いに関するページが右下の部分木に集合していることが一目で分かるため、余計なページに情報収集を阻害されることはない。

図 2 の左下の部分木を見ると「その他」を親ノードとする Web ページが複数ある。ここにはアルゴリズムではクラスタリングしきれなかった Web ージが集合している。検索結果の Web ページの中では特殊な Web ページであるといえる。

図 2 の上の部分木はかなり大きなものとなっている。ユーザはラベルを見てマウスドラッグによってフォーカスを移動し部分木を辿って情報を収集することができる。図 3 は「日本」と「歴史」を含み「地理」に関する Web ページが集合している部分木である。図 4 は「日本」と「歴史」を含み「教科書」に関する Web ページが集合している部分木である。さらに詳しく見ると従軍慰安婦と教科書の問題について述べているページが多いことが分かる。その他にも日本の歴史に関する書籍に関する Web ページから構成される部分木や、日本の歴史を扱

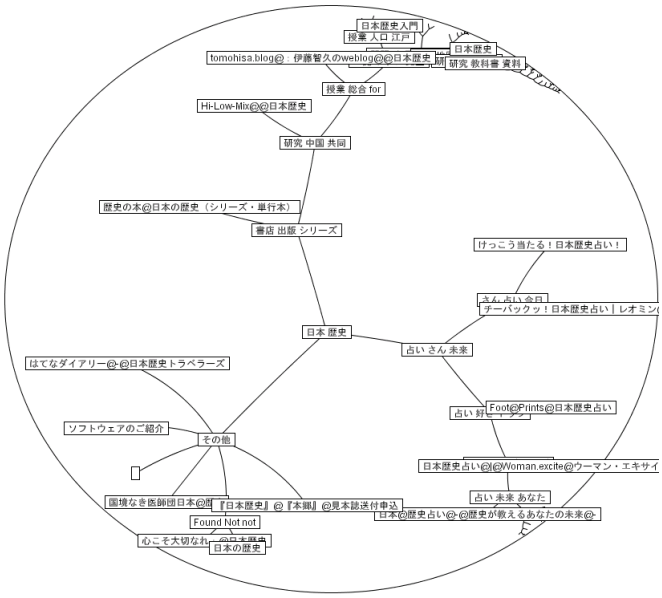

図 2: 検索要求「日本　歴史」に対する提示画面

う教育機関に関する Web ページから構成される部分木、日本の歴史に関する学会や論文を取り扱う Web ページから構成される部分木なども存在する。

一次元のテキストリストを用いて検索結果を提示するインタフェースではユーザは様々なジャンルの Web ページを含むリストに対して、1 つの Web ページを単位に順に確認する必要があるのに対して、本システムでは類似した Web ページによって構成される部分木を単位に見ていけばよい。そのため様々なジャンルから情報を収集しなければならない場合にユーザは思考を切り替えることなく部分木ごとに必要な情報を効率よく取捨選択することができるのである。また、検索結果が 1 画面に納められていることで、ユーザは部分木を参考にして容易に検索結果の全体像を理解することが可能となる。

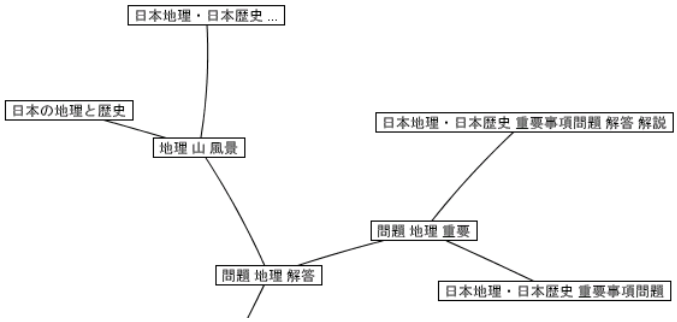

図 3: 地理に関する部分木

5. 関連研究

Web 検索結果を一次元のテキストリスト以外の提示方法を用いて提示する検索インタフェースとして三次元空間に Web 検索結果を配置する Poznan University of Economics の開発する Periscope[Wiza 04] や Web ページをホスト名によりクラスタリングし二次元空間に配置する University of Kent at Canterbury

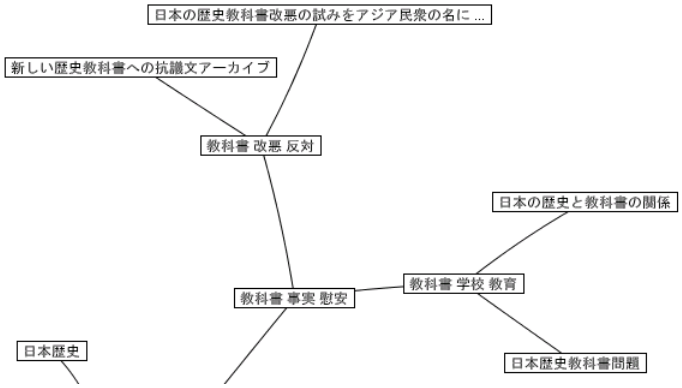

図 4: 教科書に関する部分木

の開発したシステム [Roberts 02] などがあるが、これらのインタフェースは Web ページの内容を考慮したクラスタリングを行っておらず、Web ページを単なる二次元または三次元オブジェクトとして画面に配置しているためユーザに十分な情報を提供しているとは言えない。

本研究で提案したシステムでは Web ページのタグ情報を活用し、内容を考慮したクラスタリングを行なった。また、分類結果を特徴を表す付加情報と共に一面に納めて提示する。

6. まとめ

本研究では Web ページの増加に起因する情報指向型検索に対する既存インタフェースの問題点を考察し、問題を解決するためのインタフェースの提案と試作を行なった。提案手法を用いることで Web 検索結果を適切にクラスタリングすることができ、ユーザは Hyperbolic Tree のラベルを参照しながら効率よく情報指向型検索を行なうことが可能となった。今後の課題としては、ユーザインタフェースの充実やユーザ評価を行なうことなどが挙げられる。

参考文献

[Broder 02] Broder, A.: A taxonomy of web search, *SIGIR Forum*, Vol. 36, No. 2, pp. 3–10 (2002)

[Lamping 95] Lamping, J., Rao, R., and Pirolli, P.: A focus+context technique based on hyperbolic geometry for visualizing large hierarchies, in *CHI '95: Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 401–408, ACM Press/Addison-Wesley Publishing Co. (1995)

[Roberts 02] Roberts, J., Boukhelifa, N., and Rodgers, P.: Multiform Glyph Based Web Search Result Visualization, the Sixth International Conference on Information Visualisation (IV '02), pp. 549–554IEEE (2002)

[S.Everitt 93] S.Everitt, B.: *Cluster analysis*, London:E.Arnold, 3rd edition (1993)

[Wiza 04] Wiza, W., Walczak, K., and Cellary, W.: Periscope: a system for adaptive 3D visualization of search results, in *Web3D '04: Proceedings of the ninth international conference on 3D Web technology*, pp. 29–40, ACM Press (2004)