情報爆発時代における
文書構造を考慮した検索システムのユーザインタフェース

伊藤　智博†　　宮崎　純†　　中島　伸介†　　植村　俊亮‡　　加藤　博一†

奈良先端科学技術大学院大学　情報科学研究科†　　奈良産業大学　情報学部‡

1．はじめに

現在、検索エンジンによって作成される検索結果一覧は見易いとは言えず、この結果一覧に含まれる要約を見ても内容を理解できない物が多い。そのため得られた検索結果の中で検索対象の語が web ページ内のどこで使われているかをユーザ自身が確認し、ページの適合性を判断しなければならない。さらに、現在の検索エンジンを利用した情報検索において、対象とする web ページ内において複数の話題を取り扱っていることが多く、情報量が膨大である。これらを解決するために、我々は、web ページを検索としての単位とするのではなく web ページ内の部分文書を検索としての単位とすることにより単位あたりの情報を絞ることが可能となり効率的な情報提示が可能ではないかと考えた。さらに、web ページの文書構造情報を利用し、検索対象の語がページ内でどの部分に出現しているかをユーザに知らせるためのインタフェースを考えた。これらにより検索エンジンのユーザビリティの向上を図る手法を提案する。

2．文書構造の表示方法の検討

従来の検索エンジンでは www から web ページの情報をクローリングし、その情報を web ページ単位で保存し利用している。本研究ではこれらの情報を web ページ単位ではなく、各ページ中の部分文書単位で保存し、利用する。なお、この方法は、処理時間が多少必要ではあるが、既存の検索エンジンから得られる情報から動的に部分文書を抽出することによってシステムを実現することも可能である。

例えば、『情報』と『検索』をキーワードに検索を行い、図１の様な文書が検索されたとする。既存の検索エンジンの表示方法、例えば Google にスニペットでは、リンクを開くまでキーワードとして使用された語同士がどのような関係を持つかを知りえることは難しい。また、従来の手法として web ページ内に現れるキーワード間の距離（単語数）を計り、キーワード間の関係を求める研究があるが、文書中でキーワード間の距離が近くても、構造上でも近いとは限らない。構造上の関係を示すために、図２のような表示形式が考えられる。なお、図２は図１の部分文書が得られたときの表示例である。

```
<div>
    <div>
        <p>情報</p>
        …検索…
    </div>
    <p>検索…情報…</p>
</div>
```

図 1 得られた部分文書例

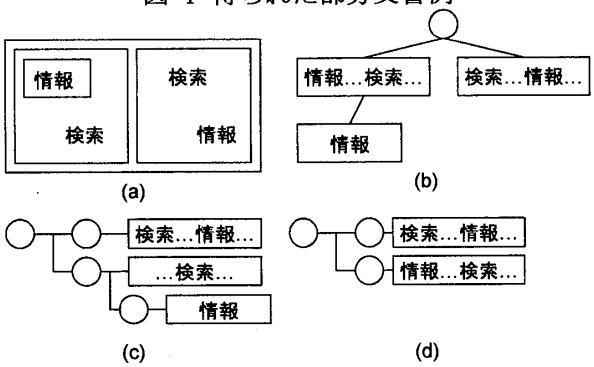

図 2 部分文書の表示形式の例

図 2 の(a)は、文書構造の一階層をそれぞれ矩形で表したものである。この方法ではグラフィカルに表示することが可能であるが深い階層を表す時に大変見難くなる。さらに、矩形の位置と実際の表示位置が異なるためユーザが誤解を招きやすいと考えられる。

(b)は構造を木構造によって表したものである。木構造は情報科学に関する知識を持つユーザにとっては理解し易いが、知識を持たない一般ユーザからすると直感的な理解が難しい。

(c)は Windows のエクスプローラ等でも用いられている形式を利用したものである。これは(b)

User Interface of a Search Engine considering Document Structures in the Info-plosion Era.

† Graduate School of Information Science, Nara Institute of Science and Technology

‡ Faculty of Informatics, Nara Sangyo University

の表示を、情報科学に関する知識の無い一般ユーザにも馴染みの深い構造で表した形であるが、(b)と同等の表現力を持つ。

この(c)の方法においても、深い階層を表す場合には(a)と同様の問題が生じる。この問題は(d)の様に特定のノード以下は一つのテキストノードとして表示する事によって解消できる。

一般に、検索エンジンでは、検索キーワードが全て出現する文書が必ずしも高いランキングであるとは限らない。そこで、エクスプローラ型の構造表示方法に加えて、ユーザに各部分文書に含まれる検索キーワードの数を、色や濃淡を使用して視覚的に表す。例えば部分文書に検索キーワードとして使用された語が全て含まれている場合は赤、含まれていないキーワードが増えるほど青に近づけていく等で表現することにより、ユーザが検索意図に近い部分文書を容易に発見できるよう表示を行う。

3．検索結果のランキング方法に関する考察

次に、全体の検索結果のランキングの表示方法について検討する。本研究では検索結果から得られた部分文書を tf-idf を用いてスコア値を計算するものとする。

従来の検索エンジンでは、各ページをその検索エンジン特有のスコア付けによりスコアが高い順に表示している。同様に、web ページに関係なく、その部分文書ごとにスコア付けを行い、web ページとは独立にスコアの高い部分文書を表示する方法が考えられる（図３参照）。図３のように、URL は各部分集合が含まれる web ページの URL を示し、その下は web ページ内での部分文書を表すパスである。空白の矩形は前節で議論した文書構造を表示する領域である。

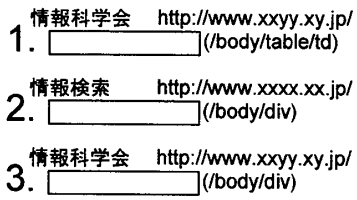

図 3　スコア順による部分文書検索の表示例

この図３の方法では一つの web ページごとに情報がまとまっておらず、部分文書同士はオーバラップを許すため、同一の情報に複数回アクセスしてしまう可能性があり、ユーザビリティが低下する原因となるかも知れない。この問題を解決するには、同一 web ページに存在する部分文書をグルーピングして、各 web ページ内での部分文書の絶対スコア順に並べる方式が考えられる。図４はその表示例である。左の数字が web ページ単位のランキング、カッコ内がその web ページの部分文書と部分文書の絶対ランキングである。ユーザは web ページ単位のランキングと、部分文書のランキングを一度に閲覧することができ、ユーザは検索意図に合った部分文書を探すことが可能となる。

この方式は、既存の検索エンジンを web ページ単位のスコアリングに利用することで、容易に実現することが可能である。

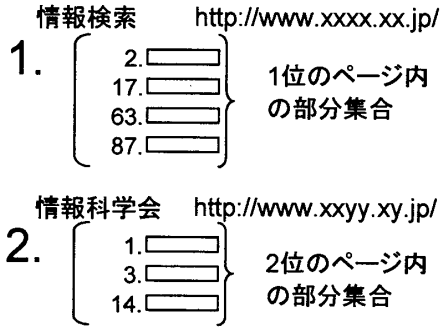

図 4 提案手法による検索結果の表示例

4．おわりに

本研究は、ユーザが web 検索を効率的に理解することが可能な検索システムのユーザインタフェースを提案した。検索システムがユーザの求める情報検索意図を完全に判断することは不可能であり、検索システムはいかに文書の持つ多様な情報を、ユーザが判断しやすい形で提供するかが重要であり、本インタフェースはそれをサポートする。

今後提案したユーザフェースを持つシステムの実装を行い、システムの評価を行う予定である。

参考文献

1) 高見真也，田中克己，ウェブページに対する定量的評価の視覚化による情報検索支援，データベースと Web 情報システムに関するシンポジウム（DBWeb2006）論文集，pp.67-74，2006 年 12 月

2) Martin Theobald, Ralf Schenkel, Gerhard Weikum, An Efficient and Versatile Query Engine for TopX Search, Proceedings of the 31st VLDB Conference, pp.625-636, 2005

3) 波多野賢治，絹谷弘子，吉川正俊，植村俊亮，XML 文書検索システムにおける文書内容の統計量を利用した検索対象部分文書の決定，電子情報通信学会論文誌，Vol.J89-D，No.3，pp.422-431，2006