社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－ＦＩ－90 (6)
2008－ＤＤ－65 (6)
2008／3／28

文書構造情報を利用したWeb情報検索

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
†〒630-0192 奈良県生駒市高山町 8916-5
‡〒630-8503 奈良県生駒郡三郷町立野北 3-12-1

本論文では膨大な情報にあふれる現代社会において、Web検索エンジンの検索単位をWebページからその部分文書にすることにより、ユーザの負担を減らすことを目的とする。現在の検索エンジンによって作成される検索結果一覧に含まれる要約は、すぐに内容を理解できない場合が多い。そのため検索対象の語がWebページ内のどこで出現し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、Webページ内には複数の話題を取り扱っていることも多く、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、ページ内の部分文書を検索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な検索手法を提案する。

A Web Search Engine Considering Document Structure Information

Chihiro Ito†, Jun Miyazaki†, Shinsuke Nakajima†, Shunsuke Uemura‡, Hirokazu Kato†
†Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara 630-0192, Japan
‡Faculty of Informatics, Nara Sangyo University
3-12-1 Tatsuno-kita, Sango-cho, Ikoma-gun, Nara 630-8503, Japan

We propose a high precision Web page retrieval method by decomposing Web pages into partial documents which are regarded as a unit of information. It is not easy to understand the contents of the retrieved Web pages obtained by a existing Web search engines because the contents of most of obtained Web pages cannot intuitively be captured even by reading their abstracts and/or snippets. Therefore, we must manually ensure where query terms are appeared in a page and whether the page is really relevant or not by reading each page. In addition, a Web page may contain several topics which might be irrelevant to user's information needs. To deal with these problems, we treat a partial document which has one topic, instead of a page, as a unit of information, and then, we propose an improved Web page search method based on the partial document search.

1. はじめに

WWW (World Wide Web) の爆発的な普及により、多様で膨大な情報がインターネット上に氾濫している。その代表的なものの一つに電子化された文書がある。当初、電子化された文書は、WWW の登場とともに急速に普及したHTML (Hyper Text Markup Language)形式[4]で記述されたものが主流であった。しかし、

HTML は、電子的な文書管理を目的として開発された言語である SGML (Standard Generalized Markup Language) をベースに開発された言語であったため、WWW 上でのデータ交換という利用領域においてその拡張性や柔軟性等の面で限界が生じてきた。

こうした状況を踏まえて W3C (WWW Consortium) が推奨したメタ言語が XML (Extensible Markup Language) [5]である。XML は、SGML のサブセットとして簡明な表現が行えるよう言語仕様が規定されるとともに、電子的なデータ交換の役割も担えるよう設計が行われた。このため、XML は文書の表現形式のみにとどまらず、より広い情報交換の形式として急速に普及しつつある。XML は構造化文書であり、内部に記述するタグとその入れ子構造によって木構造で表現することができる。また、タグ内の要素や属性が名前を持つことから、自己記述的な表現が可能であり、様々な種類のデータが一つの XML データに記述されることも多い。データ交換への対応に加え、これらの柔軟な表現力により、様々な領域のデータが XML で記述されている。

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場合が多い。そのため、得られた検索結果中から検索対象の語が 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、対象とする一つの Web ページ内で複数の話題を取り扱っていることも多く、情報量が多いため、目的とする情報へたどり着くのは非常に困難である。

そこで我々は、Web ページを検索の単位とするのではなく Web ページ内の部分文書を検索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な Web 検索手法を提案する。本手法により、ユーザに対してより効率的な情報提示が可能となる。

本論文では、プロトタイプシステムを構築し、Web 文書の部分文書の重み付けならびに正規化方法を検証し、その検索精度を評価した。

本論文の構成は以下の通りである。まず 2 節で関連研究を挙げ、提案手法との比較を行う。3 節では提案する Web 文書の部分文書検索手法ならびにその実装に関して述べる。4 節で Web 文書の部分文書の重み付け、正規化について検証し、最終節で本論文をまとめる。

2. 関連研究

高見らは、検索語間の文字数を単語間の距離として計算し、距離と出現頻度によってページの特徴を表現しているが[1]、この手法では、複数のトピックを含む文書や、文書の構造上では遠い単語同士が、単語間の距離では近い関係となるような場合もある。そのため、検索語の距離を測るのではなく文書構造を考慮する必要があるのではないかと考えられる。本研究では Web ページを、タグをもとに部分文書に分解し、タグによる構造情報を利用することにより検索語同士の関係を明確にすることが可能となる。

Martin らは、厳密な XML に対して文書構造を用い、検索速度を重視した部分文書の検索手法を提案している[2]。これは厳密な XML 文書の部分文書を対象としているが、Web ページには XML のような厳密さがないばかりか、構造情報ではなく装飾のためのタグの利用も多く、XML 部分文書検索手法をそのまま Web に適用することはできない。本研究では HTML で書かれた Web 文書のように、厳密ではない構造を持つ文書に対して検索精度を重

視した部分文書検索を目指している。

波多野らは厳密なXML文書中から情報量の少ない部分文書を前もって除去することで、XML部分文書検索の精度を上げる研究を行っている[3]。Webページは XML ほど厳密ではないが波多野らの手法は本研究でも適用可能であり、本研究においても更なる精度を上げるために適用可能な技術であると考えられる。

3. 部分文書を利用した情報検索手法

本研究では、部分文書を検索対象とする、新しい Web 検索システムを提案する。現行のWeb検索システムではWebページを検索の単位としてユーザへ情報を提供するが、提案手法ではWebページの部分文書を検索の単位としてユーザへ情報を提供する。これにより、ユーザはWeb 文書全体を見ることなく、ユーザが入力した検索語に関連した部分だけを閲覧することが可能となり、ユーザの負担が少なくなる。

しかしながら、Web 文書の主要な記述言語のHTMLはXMLとは異なり、HTMLのタグの利用は、文書の構造を表す以外にもデザイン上の装飾としてのタグの利用も多い。HTMLの部分文書検索の検索精度を上げるためには、構造とは関係のないタグや冗長な構造を除去するクレンジング処理が必要となる。

クレンジング処理の後、各 Web 文書を部分文書に分解し、各部分文書中の索引語の局所重み付けを行い、データベースに格納する。検索エンジンはこのデータベースから部分文書と統計情報を取り出し、大域重み付けと正規化を行い、得られた検索結果をランキングしてユーザに提示する(図1参照)。

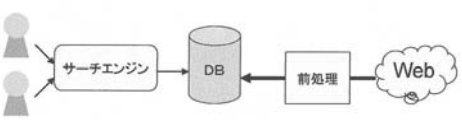

図 1：システムの概要

3.1 前処理

3.1.1 HTML から XHTML への変換

HTML のタグ付けは厳密ではないものが多く処理しにくいが、これらを厳密な構造をもつ言語へ変換することで部分文書による情報検索が容易となる。そこで、我々はHTML文書をすべて XHTML に変換して処理することにした。XHTMLはHTMLをXMLで定義し直した言語であり、HTMLからXHTMLへの変換はいくつものツールが利用でき、容易に行うことができる。また、XHTMLはXMLのサブセットであるため厳密な構造を持つだけでなく、XPath や XQuery など文書構造を利用した問合せのツール群を利用することができる利点も大きい。

3.1.2 文書のクレンジング

前述したように、Web ページから部分文書へ分解する前にデータのクレンジングを行う必要がある。これは、i)文書構造上冗長となる子孫にテキストノードを含まないノードの削除、ii)文書の構造とは関係の無い装飾に関するタグの削除ならびにノードのマージを行う。

i)に関して、子孫にテキストノードを含まないノードから重みを求めることはできない。そこで、葉ノードにテキストノードを含まないノードの削除を行う。ii)に関しては、HTML には<table> <p>のように構造を表すタグと<b><i><font>などのように装飾を表すタグに分けられる。しかし、ユーザの主観やHTMLによる制約によってHTMLは記述されているため文書構造を利用するにあたり装飾タグに該当するノードは不要である。

図 2 を例に考える。カレントノード b は装飾タグである（図 2 左）。そのため b の子ノードである p の親ノードを div に変更する。その後 div と b のリンクを削除する（図 2 右）。

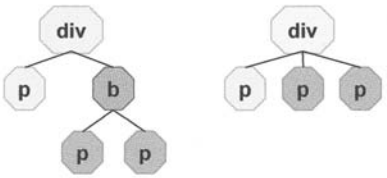

図 2： 装飾タグ削除の例

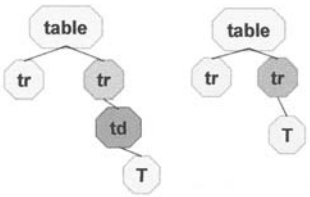

図 3： ノードのマージの例

一方、もしあるノードに子ノードが一つしかない場合は、そのノードと子ノードが表す部分文書は同一のものとなる。そのため、このようなノードの削除も行う。

図 3 を例に挙げると、カレントノードは tr であり、その子ノードは td のみである（図 3 左）。この場合、td の子ノードであるテキストノードは tr の子ノードになるように親ノードを変更すればよい（図 3 右）。

3.1.3 統計情報の計算とデータベースへの格納

データのクレンジングの後、文書を部分文書に分解し、分解された部分文書の統計情報の計算を行う。前処理として行う統計情報の計算は局所重み付けのみである。大域重み付けと正規化の計算は対象となる文書群が決定しなければ計算できないため実際にユーザから問い合わせがあった際に計算を行う。

得られた部分文書とその統計情報である局所重み付けの結果はデータベースに格納され、ユーザからの問い合わせの際に検索エンジンによって取り出される。

XML 部分文書検索の局所重み付けの研究は多くの研究があるが、Web 文書の部分文書に関しては、どのような局所重み付けが有効であるか明らかではない。よって、本研究では対数化索引語頻度（1）、対数化索引語頻度の 2 乗（2）、対数化索引語頻度をさらに対数化した重み付け（本論文では重対数化索引語頻度と呼ぶことにする）（3）を局所的重み付けとして準備した。

$l_{ij}=\log(f_{ij}+1)$ ・・・（1）

$l_{ij}=(\log(f_{ij}+1))^2$ ・・・（2）

$l_{ij}=\log(\log(f_{ij}+1)+1)$ ・・・（3）

3.2 問い合わせ処理

検索エンジンはユーザから問合せを受けると問合せ条件を満たす部分文書と前もって計算された統計情報をデータベースから取り出す。それらをもとに大域重み付けと正規化を行い各部分集合のスコアを計算する。そのスコアによってランキングを行いユーザへ検索結果の提示を行う。本研究では、大域重み付けに IDF、正規化法には L1、L2 ならびにコサイン正規化を準備し、Web 部分文書に適した重み付けの組合せを明らかにする。

例として、『奈良』と『大学院』を問合せ条件として検索エンジンに与えた場合、検索エンジンは『奈良』と『大学院』の両方、もしくはどちらかを含む部分文書とその統計情報をデータベースから取り出す。この時はじめて対象となる部分文書数が決定し、大域重み付けと正規化が計算可能となる。大域重み付けと正規化により各部分文書のスコアを求め、検索結果のランキングを行いユーザに提示を行う。

3.3 実装

本研究は、提案する検索エンジンのシステム全体を実装する事が目的ではないため、Google を図1のデータベースの代わりとして利用することで実装を行った。これは Web 上にある膨大な情報を格納するのが困難なためである。

まず、Google の検索用ライブラリを利用して、問合せ条件を満たす上位 n 件の Web ページの URL を取り出し、各 Web ページのソースコードを取得する。Google の検索結果の上位を取得するのは、検索結果一覧の下位の情報はそれほど重要ではないと考えられるためである。その後、取得した Web ページのソースコードを厳密な構造を持つ XHTML に変換する。

文書のクレンジングに関して、子孫にテキストノードを含まないノードから重みを求めることはできない。そこで、葉ノードにテキストノードを含まないノードの削除を行う。これは、XHTML に図 4 のような XPath 問合せを適用することで簡単に実現できる。この XPath 問合せは、XHTML の木構造から、ルートからのテキストを含んでいる全てのパスの集合を取得する。

```
/descendant-or-self::text()
```

図 4　全テキストノードを取得する XPath 問合せ

装飾を表すノードの削除は以下のように実現した。

1. XHTML の木構造を幅優先探索によりノードを一つ取り出しカレントノードとする
2. カレントノードが装飾を表すタグかどうかを判断
3. 構造を表すタグならば1へ
4. カレントノードの子ノードを取り出しそれらの親ノードをカレントノードの親ノードに差し替える
5. 幅優先探索により全てのノードを探索するまで繰り返す

一方、ノードのマージについては以下の方法で実現した。

1. XHTML ツリーを幅優先探索によりノードを一つ取り出しカレントノードとする
2. カレントノードに子ノードが一つ以上あるかを判断
3. 複数の子ノードを持つならば1へ
4. カレントノードの子ノードを取り出しそれらの親ノードをカレントノードの親ノードに差し替える
5. 幅優先探索により全てのノードを探索するまで繰り返す

クレンジングを行った XHTML に、図5で示す XPath 問合せを適用し、検索語 w_i を含むテキストノードを取得する。得られたテキストノードセットからノードを取り出し各検索語の出現頻度を計算する。その値を自分自身からルートノードまでにある全ノードに加算していく。これにより、それぞれのノード以下に含まれる各検索語数が求まり、各検索語の局所重みが計算可能となる。同時に各検索語を含む部分文書数を検索語ごとに求め、大域的重み付けと正規化を行う。

```
/descendant-or-self::text()[contains(.,'wi')]
```

図 5　検索語 W_i を含む XPath 問合せ

以上の統計情報から各部分文書のスコアを求めランキングを行いユーザに検索結果の提示を行う。

4. 評価

本節では、Web 文書の部分検索システムを実用化する上で重要となる、単語の重み付けと正規化法の最適な組合せを、実験を通して明らかにする。

局所重み付けには対数化索引語頻度、重対数化索引語頻度、対数化索引語頻度を 2 乗した頻度の 3 通りを用いる。また、大域重み付けには IDF を用いた。文書正規化にはコサイン正規化、L1 正規化、L2 正規化を用いた。

Web 検索エンジンをはじめ、情報検索システムを評価するためには、通常テストコレクションが利用される。しかし、Web 文書の部分文書検索にはテストコレクションが存在しない。そのため問合せ集合は独自に検索語が二語のものを三種類、三語と五語のものをそれぞれ一種類ずつ作成した。

実験の方法として、GoogleAPI により Google の検索一覧上位 50 件の情報を取得し、それから部分文書集合を求めランキングを行い、その上位 30 件を抽出する。求めた部分文書上位 30 件に関して、R 適合率を利用して評価を行った。

実験の結果、対数化索引語頻度とそれを 2 乗した頻度による局所的重み付けではいかなる正規化法を適用しても適合率が著しく低かった。これは正規化により出現回数が少なく文書の長さが短い部分文書が上位になるためと考えられる。例えば、検索語を一つしか含まないような小さなテキストノードが上位に現れてしまう。特に L1、L2 正規化の場合、索引語の重み付けや検索語の多少に関わらずこの現象が見受けられた。

R 適合率により検証した結果、重対数化索引語頻度とコサイン正規化の組み合わせが最も良い結果を示すことが分かった（図 7）。これらから、局所重み付けは出現頻度の高い索引語の影響を少なくすることで良い結果が得られることが分かる。

全般的に、検索結果の上位 30 件中に似通った文書がいくつか選ばれる可能性があった。そのため、適合率のランクに対する変動が顕著に現れてしまっている。これから、3 節で挙げたデータのクレンジング方法だけではまだノイズの除去が不十分と考えられる。

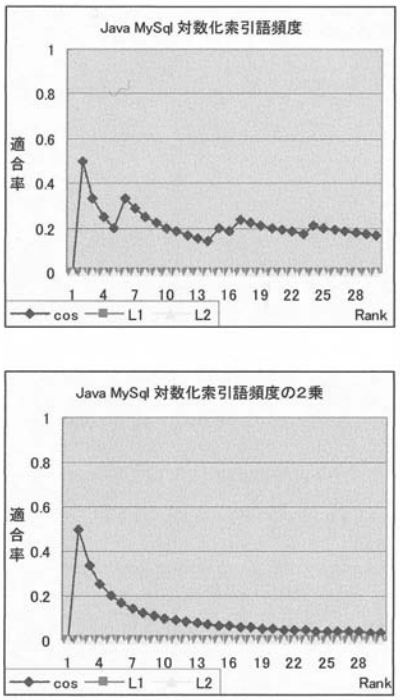

図 6　対数化索引語頻度、対数化索引語頻度の 2 乗の場合の R 適合率

図 8 で示すように、『ジャガー Mac』で検索した場合、重対数化索引語頻度とコサイン正規化の結果が悪かった。これは、Mac という文字の並びが全く関係の無い英単語にも含まれていることが多いため、この単語では検索意図を十分に絞れなかったためであると考えられる。しかしながら、検索エンジンを使う多くのユーザは検索結果の上位数十件程度しか利用

しない傾向があるが、この結果から見る限り、上位ランクの適合率が高いため、この重み付け方法でも十分に有効であると言える。

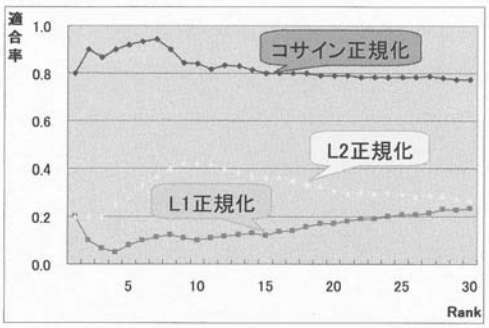

図 7　重対数化索引語頻度による R 適合率の平均

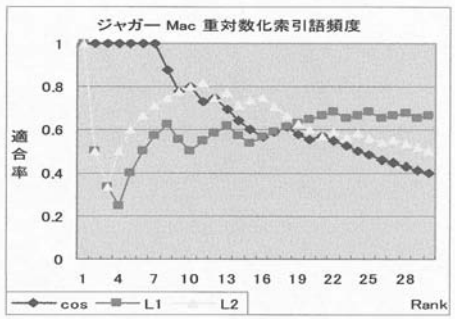

図 8　「ジャガー」「Mac」の場合の R 適合率
（重対数化索引語頻度）

5. おわりに

本論文は、Web 検索の単位を Web ページからページ中の部分文書にすることにより、Web 文書群からより容易にユーザが目的とする部分文書を検索することが可能なシステムを提案し、評価を行った。現在の検索エンジンを利用した場合、対象とする Web ページ内に複数の話題を取り扱っていることも多く、情報量が膨大である。このため目的とする情報へたどり着くには非常に時間を要するが、部分文書を評価の単位とし局所的重み付けに重対数化索引語頻度、正規化係数にコサイン正規化を用いることで、上位 10 件の R 適合率が 80%を超える結果となることを示した。これは Web ページの情報を部分文書としてユーザへ提示することの可能性ならびに有効性を示している。

今後の課題として、検索結果の部分文書をユーザへ提示するためのインタフェースが挙げられる。

謝辞

本研究の一部は、文部科学省科学研究費補助金特定領域研究「情報爆発時代に向けた新しい IT 基盤技術の研究」の公募研究(課題番号 19024058)の支援による。ここに記して謝意を表します。

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", 情報処理学会データベースと 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, and Gerhard Weikum: "An Efficient and Versatile Query Engine for TopX Search", Proc. 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 D, Vol.J89-D, No.3, pp.422-431, 2006

[4] World Wide Web Consortium. Hyper Text Markup Language (HTML) 4.01. http://www.w3.org/TR/REC-html40/, W3C Recommendation, December 1999 (2008 年 2 月 URL 確認)

[5] World Wide Web Consortium. Extensible Markup Language (XML) 1.0 (Third Edition). http://www.w3.org/TR/2004/REC-xml-20040204/, W3C Recommendation, February 2004 (2008 年 2 月 URL 確認)