データベースシステム 125－83
（2001．7．19）

情報検索技術による構造化部分文書の抽出法

波多野 賢治† 絹谷 弘子†
吉川 正俊†,‡ 植村 俊亮†

構造化文書から，利用者の問合せに相応しい部分文書を抽出する方法として提案されている手法には，問合せ言語を利用するデータベース的なアプローチ法と，情報検索技術を用いる手法がある．しかし，これらの手法によって得られる構造化部分文書の葉ノードであるテキストノードの中には，利用者の問合せに相応しくないノイズとなるべきものも存在している．本稿では，検索システムによって検索された構造化部分文書から，こうしたノードを除去する手法を提案し，その有効性を確認した．また，構造化部分文書の検索精度評価のための手法についても提案し，いくつかの知見を得ることができた．これらの手法が確立すれば，現在で WWW で利用され始めている XHTML 文書から，利用者の問合せに相応しい部分文書を効果的に検索することが可能となる．

Extraction of Structured Partial Documents Based on IR Technique

Kenji Hatano† , Hiroko Kinutani† , Masatoshi Yoshikawa†,‡ and Shunsuke Uemura†

Until now, a lot of researches were appearard concerned with extraction of structured partial documents. These researches can be classified into two categories — database-based approach using query language and IR-based approach. However, some text nodes, leaf nodes of structured partial documents searched by information retrieval systems, are not suitable for user's query in many cases. In this paper, we proposeed an approach of removing such nodes from the structureed partial documents searched by the IR systems, and checked the validity of our proposed method. We also proposed an evaluation method for our system of structured partial documents, and got some useful knowlege for its establishment. If our proposed method is established, people can retrieve structured partial documents relevant to user's query effectively from XHTML documents which are emerging as the standard format representing documents on the Internet.

1. はじめに

XML の浸透による情報化社会の動きは目覚しく，XML 文書などの構造化文書に対する検索の要求はますます大きくなると予想される．XML に代表される構造化文書の検索の手法にはこれまで多くの研究が行われているが，それらは主に文書構造を利用しデータベースに格納したり，XML 問合せ言語を用いたりするデータベース的なアプローチと，出現単語の情報から索引を生成し検索を行う情報検索技術を用いるアプローチに大別される．前者の手法は，データベースの機能に盛り込まれたり，W3C (World Wide Web Consortium) の勧告となったりと非常に研究が盛んに行われているが，後者の手法は情報検索の研究分野の歴史を考えるとまだまだ日が浅く研究も始められたばかりである．しかし，後者によるアプローチは，前者に比べ問合せ言語を書くといった問合せ言語に関する専門的知識や，データベースに格納されている文書の文書構造についての知識を必要とせず，利用者にとっての親和性が高いという利点を持っている．つまり，インターネットの検索エンジンのように，利用者は検索キーワードを入力するだけで，求めている情報を検索することができ，それらはさらにランキングされることによって利用者にとってどれほど有用な情報なのかを知る手がかりを利用者は知ることができる．

さて，いずれの手法を用いても XML 文書から利用

† 奈良先端科学技術大学院大学情報科学研究科．
Graduate School of Information Science, Nara Institute of Science and Technology (NAIST).
‡ 国立情報学研究所ソフトウェア研究系．
Software Research Division, National Institute of Informatics (NII).

者の求める情報，すなわち利用者が問合せを発行することで問合せに相応しい部分文書を得ることができるわけであるが，最終的に得られる結果は，現在のところ XPath (XML Path Language) を用いた部分文書の根ノードの指定であるため，部分文書中に問合せの答えとして相応しくないテキストノードが含まれていても検索されてしまうという問題点がある．検索された部分文書にこのようなノイズとなるべきノードが多く含まれれば，利用者が求めている情報となる可能性は低いため，利用者の問合せに合致しないノードを除去する過程は検索システムの精度を左右する非常に重要な処理だと考えられる．また，この重要な処理は，単に XML 文書の検索システムとして利用されるだけではなく，今後，WWW で利用されるようになると考えられている XHTML 文書の検索システムとしても応用が可能であり，現在の WWW 検索エンジンでは実現できていない，WWW 上の情報からの部分文書検索も可能となるための要素技術としても重要であると考えられる．

そこで本研究では，利用者にとって有益な情報を効果的に検索するという利用者の立場に立ち，利用者にとって必要な情報を効果的に提示するような XML 検索システムの提案を行う．本論文では，検索システムの提案およびその具体的な実装法について述べる．また，その有効性を確認するための評価実験を行うために，構造化部分文書のための検索システムの性能評価尺度を定めたので，それについても報告する．

2. 関連研究

1 章で述べたように，これまでの XML 文書の検索手法では，検索結果の部分文書が検索された文書の XPath 式，すなわち部分文書の根ノードが指定されることが多いので，部分文書中に利用者の問合せに相応しくないノードが含まれていても検索されてしまうという問題点がある．この問題点を解決するための一つの方法として考えられるのは，XML 問合せ言語と情報検索技術との統合である．

このような研究は，近年，数多く行われるようになってきている．Fuhr らは，XQL において情報検索技術を利用できるように拡張した検索言語 XIRQL を提案している [4]．この論文では，XQL と情報検索技術を統合するために必要な要素技術として，出現単語に対する重み付けの手法や問合せに対する類似度を基にした検索の実現手法が必要であるとし，XQL に対して拡張を行っている．これによって，検索結果の部分文書の粒度を索引語に依らないよう工夫を行っている．また，Florescu らは，XML 問合せ言語 XML-QL を拡張し，XML-QL においてキーワード検索ができるよう工夫している [3]．これにより XML 文書の構造を知らない利用者でも，異なる DTD (Document Type Definition) を持つ XML 文書の検索を容易にできるよう工夫を行っているが，検索される部分文書の粒度を利用者側で指定しなければならない．

さらに，XML 検索エンジンもいくつか公開されるようになってきた．主記憶上のデータベースと検索エンジンを組み合わせた Xset[10]，XML 文書中の各要素の特徴を文書要素の葉にあるテキストノードの内容を索引付けすることによって行う BUS (Bottom Up Schema)[9] を利用した XRS[8]，利用者に対して検索結果のスキーマを示すことで利用者の問合せ発行を容易にさせる対話システムである XYZFind[2] などが代表例である．

我々の先行研究では，XML 部分文書の検索を二つのアプローチを用いて行ってきている．一つは XML 文書の文書構造を解析して文書の内容のひとまとまりを表すであろう部分文書の根ノードである文脈ノード (詳細は 3.1.1 項を参照) を探索し，その文脈ノードを根ノードとする部分文書を検索する手法である [6]．もう一つは，XML 文書の葉に存在するテキストノードの内容と問合せキーワードとの類似度を計算し，その類似度と元文書の持つ隣接構造から検索結果として相応しいテキストノードを持つ部分文書を検索する手法である [5]．しかし，これらの手法では，一方の長所が他方の短所である場合が多いため，現時点のシステムは，これら二つのアプローチを統合して，文書内容と文書構造を利用し元文書から問合せの内容に類似した部分文書を抽出し，さらにそれらをランキングして利用者に表示するシステムとなっている．現時点の我々の提案システムでは，他の手法と比較して大きく異なる点が一つあり，XML の文書構造を表す DTD (Document Type Definition) や XML Schema を利用しなくても，XML 文書の検索が可能である点である．

本稿で提案する手法は，これら手法を使って検索された部分文書から，利用者にとって有益な情報を含んでいるテキストノードを探索し，利用者に検索結果を効果的に提示するものである．つまり，問合せに対するテキストノードの類似度を利用して，問合せに相応しいテキストノードだけで形成された部分文書のスコアを計算し，それを基にランキングする検索システムを提案するものである．したがって，本稿で例を挙げているような XML 文書のための検索システムだけではなく，今後，WWW に広まると思われる XHTML

文書の検索システムにも，容易に本手法を組み込むことが可能となっている．

3. 検索結果の質の改善

本章では，我々が先行研究で開発したシステムに提案手法をどのように適用するかについて簡単に述べる．

3.1 検索システムの概要

我々が想定する XML 文書の検索システムでは，検索対象となる XML 文書の構造，すなわちスキーマは統一されていない．つまり，利用者ばかりか検索システムの管理者でさえも対象 XML 文書のスキーマを把握することはできないことを想定している．DTD や XML Schema によって文書の構造が定義されている妥当 (valid) な XML 文書を扱う場合は，システム管理者が利用者のためのインタフェースを開発すれば，既存の XML プロセッサや XML 問合せ言語を利用して部分文書検索が可能であるが，我々の想定している文書構造を定義していない整形式 (well-formed) な XML 文書の場合は，スキーマ情報を利用することができないため，検索の際にそれらの情報を利用することが難しく，検索システムの構築が非常に困難になる．

こうした問題を解決するために，先行研究では XML 文書中からその文書の文脈のまとまりを表現する文脈ノードを探索し，それを根ノードとする部分文書単位で検索することが可能な「文脈検索」と呼ばれる手法を提案した．

3.1.1 文脈ノード

文献 6) において，我々は文脈ノードを以下のように定義している．

定義: 文脈ノード XML 文書 D 中の葉に存在するテキストノードを n としたとき，その文脈ノード $Context(s)$ は以下のように定まる:

(1) D の葉に存在するテキストノード $s_k(k = 1,2,\ldots)$ に注目し，まず探索開始ノード $p(s_k)$ を設定する．
- n に兄弟要素ノードが存在する場合は，s_k の親ノードを $p(s_k)$ とする．
- そうでない場合は s_k の祖父母ノードを $p(s_k)$ とする．

(2) $p(s_k)$ から D の最上位の要素ノード間の経路をボトムアップに走査しながら，経路中の各ノードの要素名と同一の要素名を兄弟要素ノードに持つか否かを調べる．兄弟要素ノードを持つ場合には，そのノードを文脈ノード $Context(s_k)$ とし，持たない場合は D の最上位の要素ノードを $Context(s_k)$ とする．

ここで兄弟要素ノードを持たない場合，テキストノード s_k の親ノードを除外しているのは，s_k の文字列値はそのまま親の要素ノードの文字列値の一部となるが，この親要素ノードは構造上の単位としては粒度が小さすぎるためである．

図 1 は，文脈ノードを根ノードとした部分文書の例 $PD_i(i = 1,2,3,4)$ を示したものである．PD_1 は全てのテキストノード s_i，PD_2 および PD_3 はそれぞれテキストノード s_2 と s_3, s_4 の部分文書であり，その根ノードが文脈ノードを表している．

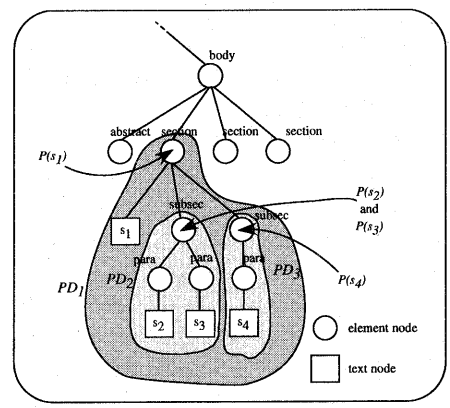

図 1 文脈ノード

3.1.2 検索システム

整形式の XML 文書から利用者の問合せに合致する部分文書を検索する手法として，最も利用者指向の検索システムとして考えられる方法は，現在公開されている WWW 検索エンジンのようにあらかじめ検索結果として返される文書を索引付けし，それを基に検索するというものである．つまり，整形式の XML 文書からあらかじめ検索結果となると予想される部分文書を抽出しておき，あらかじめそれらについて索引付けしたものに対して検索に関連する一連の処理を行うわけである．

整形式の XML 文書の部分文書を抽出する手法には様々な方法が考えられる．最も簡単な手法は，文書中の全ての要素ノードを根ノードとするように部分文書を抽出することであるが，各部分文書の重なりや要素数を考慮すると，一つの XML 文書からだけでも非常に多くの部分文書が抽出され，索引を生成する際の索引のサイズや検索時間など物理的な問題が多い．そこで，これらの問題を解決するために 3.1.1 項で述べた

文脈ノードを利用する手法を考えることにした．文脈ノードを根ノードとする部分文書は，その抽出アルゴリズムから文書作成者によってまとめられた意味的なまとまりを表現していると考えられる．つまり，全ての要素ノードを根ノードとする部分文書に比べ内容のまとまりを表現し，かつ物理的にも抽出される部分文書数は少なくなるため，索引に関連する問題点を克服することも可能となっている．

図 2 に構築したシステムの概略を示す．ここに示したように，本システムは XML 文書を XML パーザーを用いて解析し文脈ノード抽出することで検索対象となる部分文書を生成する部分と抽出された部分文書の特徴ベクトルを生成しその索引を生成する部分，そして利用者の問合せに対し各部分文書との類似度を計算しそれを基にランキング付きの検索結果を表示する三つの部分からなっている．

文脈ノード抽出・部分文書抽出

まず最初に，入力となる XML 文書の文字コードの統一などを行い (canonicalize)，XML プロセッサ (Apache Xerces version 1.2.2) を用いて XML 文書の展開を行い，主記憶上に DOM 木を生成する．生成された DOM 木を利用して XML 文書から 3.1.1 項で述べたアルゴリズムに従い文脈ノードを抽出し，その文脈ノードを根ノードとする部分文書を検索対象としてファイル出力する．

部分文書の特徴ベクトル生成

XML 文書から部分文書を抽出できれば，特徴ベクトルの生成や検索結果の表示は既存の手法を用いることが可能となる．本研究では，我々の先行研究 [5] で行った XML 文書の葉に存在するテキストノードごとにベクトルを生成する手法を利用した．なぜなら，Salton 等が提案したパッセージ検索 [7] でも示されているように，文書をある単位で分割しその要素単位でベクトルを生成・検索する手法を利用する場合のほうが，文書全体からベクトルを生成・検索する手法に比べ検索精度が良いからである．

具体的に説明すると，検索対象となる全 XML 部分文書 $\{PD_1, PD_2, \ldots\}$ 中に出現した単語の種類とその出現頻度を利用して文書要素ごとに特徴量を計算し，特徴ベクトルを生成する．全 XML 文書から単語 $w_1, w_2, \ldots, w_n$ (n は単語の種類を表現する整数) が合計 $N_{w_1}, \ldots, N_{w_n}$ 個抽出され，それぞれの単語がテキストノード s_{ij} (i は部分文書番号，j は部分文書 PD_i から抽出されたテキストノードの数を表す) 中に $N_{w_1}^{s_{ij}}, \ldots, N_{w_n}^{s_{ij}}$ 回出現したとすると，s_{ij} の特徴ベクトル $\boldsymbol{F}(s_{ij})$ は，

$$\boldsymbol{F}(s_{ij}) = \left(\frac{N_{w_1}^{s_{ij}}}{N_{w_1}}, \frac{N_{w_2}^{s_{ij}}}{N_{w_2}}, \ldots, \frac{N_{w_n}^{s_{ij}}}{N_{w_n}}\right) \quad (1)$$

となる．

検索結果の表示

問合せ Q が与えられたとき，検索システムはテキストノード特徴ベクトル $\boldsymbol{F}(s_{ij})$ と同じ基底をを持つ，問合せ Q の特徴ベクトル $\boldsymbol{q}$ を生成する．$\boldsymbol{q}$ は以下のように定義される．

$$\boldsymbol{q} = (q_{w_1}, q_{w_2}, \ldots, q_{w_n}) \quad (2)$$

ただし，q_{w_k} $(k = 1, 2, \ldots, n)$ は，検索語に w_k が含まれている場合は 1，含まれていない場合は 0 である．

次に，$\boldsymbol{F}(s_{ij})$ と問合せベクトル $\boldsymbol{q}$ との類似度を次式のような二つのベクトルの余弦によって計算する．

$$sim(\boldsymbol{q}, \boldsymbol{F}(s_{ij})) = \frac{\boldsymbol{q} \cdot \boldsymbol{F}(s_{ij})}{|\boldsymbol{q}||\boldsymbol{F}(s_{ij})|} \quad (3)$$

最後に，部分文書の評価値は，部分文書を構成するテキストノードが持つ類似度の平均値として計算する．

$$sim(Q, PD_i) = \frac{\sum_{k=1}^{j} sim(\boldsymbol{q}, \boldsymbol{F}(s_{ik}))}{j} \quad (4)$$

こうして計算された類似度を基にその値の高いものから順に並べ，利用者に検索結果として返す．

3.2 検索結果の精度改善

1 章でもで述べたように，通常，XML 部分文書を検索することが可能な検索システムは，XPath で指定されたある要素ノードを根ノードとする部分文書として検索結果を得ることができる．我々は，この検索される部分文書に含まれている問合せに相応しくないノードを除去するために，利用者にとって必要なテキストノードの情報だけを利用して，利用者に効果的に提示するような XML 検索システムの提案を行う．

提案する機能は，3.1.2 項で説明した検索システム上には容易に実装できる．なぜなら，XML 部分文書全体からその特徴ベクトルを生成しているのではなく，XML 部分文書を要素ごとに分割し問合せとの類似度を計算した上で XML 部分文書の類似度を計算しているからである．したがって，問合せに対する類似度が低いテキストノードの情報は利用せずに，部分文書の類似度を計算する手法が最も単純であると考えられる．

本稿では問合せに対するテキストノードの類似度が，部分文書中に存在する全テキストノードの平均値以上であるものだけを部分文書の類似度の計算に利用することにした☆．例えば，図 3 に示されるような XML

☆ もちろん，この閾値の計算手法は様々な方法が考えられるが，検索対象に関わらず自動的に設定されることが望ましいと考えたため，この手法を採用した．

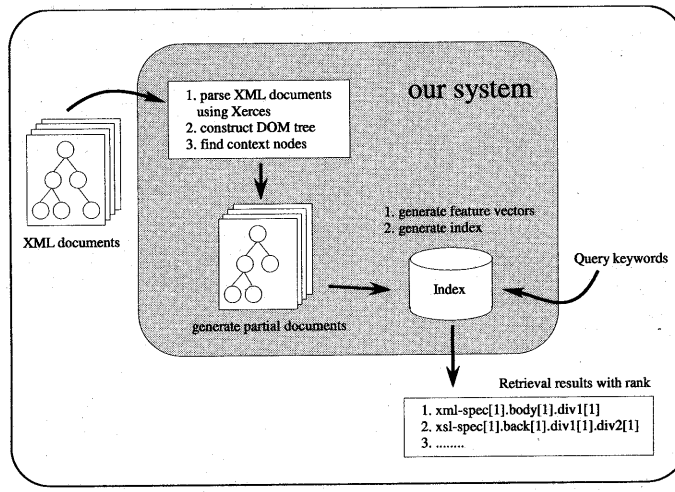

図 2 XML 文書検索システムの概略図

部分文書が与えられた場合，3.1 節で説明したこれまでの手法では，問合せ Q に対する類似度は 0.37 であるが，本稿で提案する手法を用いれば 0.75 と計算されることになる．つまり，後者の手法では，部分文書の類似度の計算の際に，部分文書中に存在する問合せに相応しくないノードから受ける影響を少なくすることが可能となり，より実際の類似度が計算できるようになると考えられる．

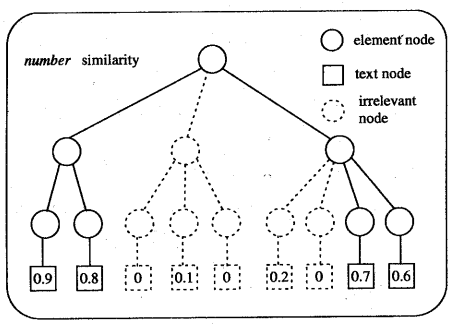

図 3 部分文書の類似度の計算

また，この考え方は利用者に対して検索結果を表示する際のユーザインタフェースにも応用可能である．WWW 検索エンジンの検索結果ように単に該当ファイルへのリンクが張られているだけでは，結局，利用者がそのリンク先へ検索結果を見に行き，そしてどの部分が問合せに該当しているのかを探さなければならず，利用者にとって非常に不便である．我々が提案する XML 部分文書検索システムは，利用者にとって使いやすく，かつ得られた結果には非常に多くの情報が含まれているようなシステムである．つまり，先に説明した問合せに対する要素ノードの類似度を利用して，検索結果として返された部分文書のどのノードが利用者にとって有益な情報でどのノードがそうでないかを表現できるようなシステムの構築を目指している．

図 4 は，我々が現在開発している利用者の問合せ結果を表示するユーザインタフェースである．利用者がキーワードを問合せ入力フィールドに入力すると，類似度の順に該当する部分文書がリスト表示されるが，さらに詳細に部分文書を閲覧しようとした場合，図のようなウィンドウが現れる．左側のウィンドウは，検索された部分文書が検索対象であるオリジナルの XML 文書のどの部分に当たるのかを，利用者にわかりやすく提示するための概観把握インタフェースであり，問合せに該当する部分文書の内容が文字列として展開されている点が特徴となっている．これに対して右側のウィンドウでは，XML 文書のどの部分が問合せに合致しているかを，該当部分を強調することで表しており，こちらは該当部分把握のためのインタフェースということができる．

これらいずれのインタフェースも，問合せに合致する部分を XML 文書から探索しなければならないが，検索対象が XML 文書であるため従来の HTML 文書の検索システムにはなかった問合せに該当する部分文書検索が可能であり，また，XSLT などを利用することで様々な種類の出力形式に容易に変更することも可能である．したがって，我々が採用した XML 文書の葉に存在するテキストノードと問合せとの類似度を利

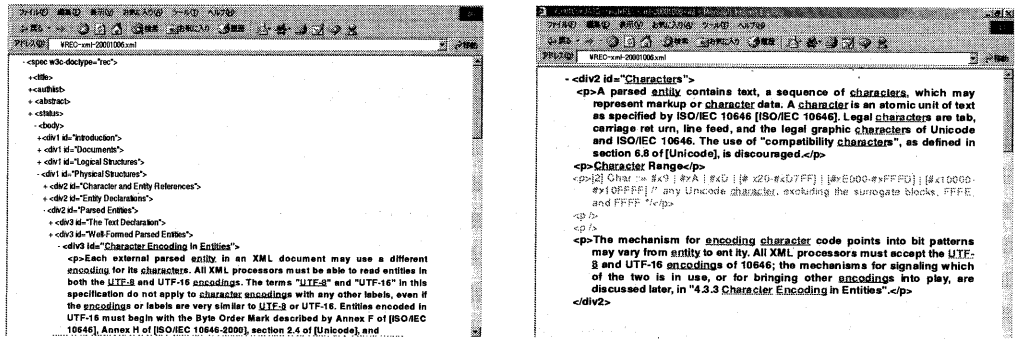

図 4　XML 部分文書の検索結果の例

用した検索システムは，利用者にとって利用しやすい検索システムのためのインタフェースの構築や次世代 WWW 記述言語である XHTML 文書の検索システムへの応用を容易に実現できると考えることができる．

4. 検索精度評価実験

4.1 実験概要

問合せに相応しくない部分文書中のテキストノードを，検索システムから返された検索結果から除去する手法の有効性を検証するために，本稿では簡単な評価実験を行った．

本実験の比較対象は，問合せに相応しくないテキストノードの情報を利用する従来型のシステムとそういった情報を利用しない我々が提案するシステムである．それぞれの検索システムに，我々がこれまでに行った研究で利用したテストコレクションを用いて適合率および再現率の算出を行い，その値を基に描かれるグラフから評価を行う．使用したテストコレクションは W3C の Technical Reports and Publications の WWW ページ☆から得られる HTML 文書を XML 文書に変換したものから構築されている．この DTD は W3C が記述する XML の DTD である XMLspec[1] の一部である．

テストコレクションを構成するオリジナルの XML 文書は 17 個であるが，前処理によって抽出される文脈ノードを根ノードとする部分文書，すなわち検索対象は 1,191 文書である．これらの部分文書に対して以下の 3 通りの問合せ/解答セットを用いて実験を行った．

- **問合せ/解答セット 1**
  質問文　XHTML の互換性の問題は将来どう解決されるのか?
  問合せキーワード　XHTML compatible issue future.
  解答　`REC-xhtml1-20000126.xml` の 5 章および 6 章.
- **問合せ/解答セット 2**
  質問文　XML のエンティティの文字コードは UTF-8 の他に何が利用できるのか?
  問合せキーワード　XML entity character encoding UTF-8.
  解答　`REC-xml-19980210.xml` の 2.2 節および 4.3.3 小節，付録 F 章と，`REC-xml-20001006.xml` の 2.2 節および 4.3.3 小節，付録 F.1 小節.
- **問合せ/解答セット 3**
  質問文　XML の要素型名や属性名に使える文字には何があるのか?
  問合せキーワード　attribute element type name character code.
  解答　`REC-xml-names-19990114.xml` の 2, 3, 4 章および付録 A.3 小節と `REC-xml-19980210.xml` および `REC-xml-20001006.xml` の 2.2, 2.3, 3.0 (3 章の枕詞の部分), 3.1 節および付録 B 節.

4.2 適合率・再現率の計算

本実験の評価するために，適合率・再現率を計算しそれを再現率–適合率グラフとして表現するが，本システムではこれまでの情報検索システムで扱わなかった部分文書を扱っているため，検索システムによって得られた部分文書が，テストコレクションの正解部分と粒度の異なる部分文書になる場合がある．したがって，これらの評価のために適合率・再現率を部分文書を検索するシステムのために再定義する必要がある．

図 5 を用いて本評価実験における適合率・再現

☆ http://www.w3.org/TR/

率の定義を説明する．テストコレクション文書集合 $\{D_1, D_2, \ldots, D_q, \ldots, D_p\}$ から，テストコレクションの解答部分文書として PD_l^r $(l = 1, 2, \ldots, a)$ が与えられ，また，検索システムから検索結果として部分文書 PD_m^s $(m = 1, 2, \ldots, b)$ が k 位にランキングされているとする (PD_m^s は m の順番で検索結果から返されるものとする)．

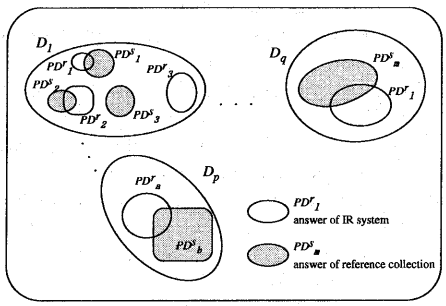

図 5　適合率・再現率の計算手法

このとき，PD_m^s の適合率・再現率を以下のように定義する．

$$\text{再現率} = \frac{\bigcup_{m,l=1}^{k,a}(PD_m^s \cap PD_l^r)}{\bigcup_{l=1}^{a} PD_l^r} \qquad (5)$$

$$\text{適合率} = \frac{\bigcup_{m,l=1}^{k,a}(PD_m^s \cap PD_l^r)}{\bigcup_{m=1}^{k} PD_m^s} \qquad (6)$$

この定義は，本研究で扱っている文書が部分文書であるために，これまで情報検索システムの評価として用いられてきた適合率・再現率と同様のアイデアを適用したに過ぎない．すなわち，適合率はシステムによって検索された部分文書 PD_m^s のうち，テストコレクションの解答とマッチングしている部分の割合を表し，再現率はテストコレクションの解答である PD_l^r 全体のうち，検索システムによって検索された部分文書の正解部分の割合を表している．

4.3 実験考察

実験結果として得られた適合率・再現率から，それぞれのシステムの評価を行うために再現率–適合率グラフを描いたところ，図 6 のようになった．このグラフは，4.1 節で述べた問合せ/解答セット 1 ～ 3 の適合率および再現率から得られる再現率–適合率グラフの各値の平均をとったものである．

このグラフを見れば分かるように，問合せに相応しくないテキストノードを検索結果の評価値の計算の際に考慮しない手法を適用した我々の提案手法のほう

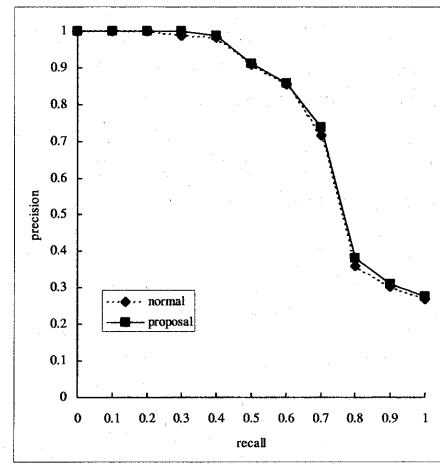

図 6　再現率–適合率グラフ (問合せ 1 ～ 3 の平均)

が，わずかながらではあるがすべての再現率において高い適合率を示すことがわかった．つまり，検索結果として得られる部分文書として，問合せに対する類似度が低いテキストノードは考慮しないほうが良い検索精度が得られるという知見が得られた．また，テストコレクションの解答となっている部分文書のサイズが小さい場合や，問合せに使われる検索語が検索対象の文書セット中に多く現れるような場合には，検索対象部分文書や問合せの特徴をベクトルで正確に表現することが難しいため，ランキングに影響を与えることがわかった．

最後に，図 6 で両手法の差が明確に現れなかった理由として挙げられる点を 2 つ示す．以後，部分文書の適合率・再現率を計算する場合にはこれらに対する考慮が必要になってくる．

- 検索システムの実装上，図 5 に示しているようなテストコレクションの解答部分文書 PD_l^r と検索システムが検索した部分文書 PD_m^s の一部だけが一致しているような場合はなく，共通部分を持つ場合は必ず PD_l^r と PD_m^s が合致もしくは PD_l^r が PD_m^s を包含する場合であった．本稿で定義した適合率・再現率は前者の場合も考慮して定義した式であったため，後者の場合しかなかった場合には，従来手法と提案手法の差がつきにくい．
- テストコレクションの解答部分文書の設定の仕方に問題がある．例えば，図 7 に示しているように，問合せに対する類似度が低いために，そのテキストノードを削ぎ落として生成された部分文書 PD_1 と，文脈ノードから生成された部分文書

PD_2 は，検索結果として相応しいとされているテキストノードは一致するが，部分文書が根ノード以下の内容を表現していることを考えると，表現している内容は微妙に異なるはずである．しかし，本テストコレクションではそうした違いを考慮した解答を有しておらず，従来手法にとって適合率の計算は優位に働いている．

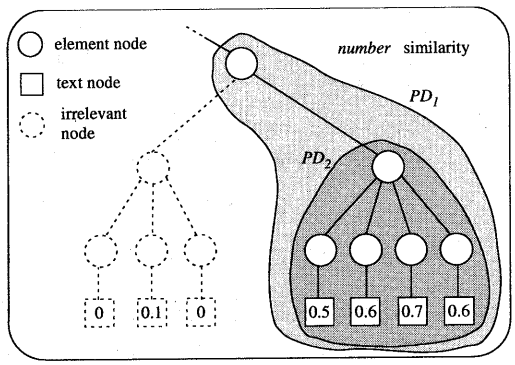

図 7　部分文書の意味の違い

5. おわりに

本稿では，XML 検索システムによって検索された構造化部分文書から，利用者の問合せに相応しくないテキストノードの除去を考慮して部分文書の評価値を再計算する手法を提案し，その具体的な実現方法について述べた．また，提案手法が検索システムの検索精度に影響することを確認し，提案手法の有効性を確認することができた．さらに，検索システムの評価を行う際の評価尺度を提案し，よりよい評価尺度の定義のための知見を得ることができた．本手法は，情報検索技術を利用した XML 文書検索手法の一手法であるが，今後 HTML の発展として利用されるであろう XHTML 文書の検索にも容易に応用でき，検索システムの要素技術として利用されると考えられる．

今後の課題として考えられる事項は，4.3 節でも述べたように，利用者が入力する問合せに使う検索語をより精錬されたものに拡張する問合せ拡張の手法を適用すること，およびテキストノードのサイズに左右されないような検索モデルを本システムに適用すること，そして，部分文書の意味の違いを考慮してテストコレクションの解答部分文書を再度見直し，それに基づいた評価を行うことが挙げられる．

謝　辞

本研究の一部は，文部科学省科学研究費基盤研究 (B)(2)，(C)(2) および奨励研究 (A) (課題番号はそれぞれ 11480088，12680417，12780309) によるものである．ここに記して誠意を表します．

参考文献

1) J. Bosak, et al. Guide to the W3C XML Specification ("XMLspec") DTD, version 2.1. http://www.w3.org/XML/1998/06/xmlspec-report-v21.htm, Feb. 2000.
2) D. Egnor and R. Lord. Structured Information Retrieval using XML. In *Proc. of the ACM SIGIR 2000 Workshop on XML and Information Retrieval*, July 2000. http://www.haifa.il.ibm.com/sigir00-xml/final-papers/Egnor/index.html.
3) D. Florescu, D. Kossmann, and I. Manolescu. Integrating Keyword Search into XML Query Processing. In *Proc. of the 9th International World Wide Web Conference*, May 2000. http://www9.org/w9cdrom/324/324.html.
4) N. Fuhr and K. Großjohann. XIRQL: An Extension of XQL for Information Retrieval. In *Proc. of the ACM SIGIR 2000 Workshop on XML and Information Retrieval*, July 2000. http://www.haifa.il.ibm.com/sigir00-xml/final-papers/KaiGross/sigir00.html.
5) 波多野賢治, 渡邉正裕, 吉川正俊, 植村俊亮. 情報検索技術を用いた部分文書構造の自動抽出. 情報処理学会論文誌: データベース, Vol. 42, No. SIG8 (TOD10), July 2001. (採録決定).
6) H. Kinutani, M. Yoshikawa, and S. Uemura. Identifying Result Subdocuments of XML Search Conditions. In *Proc. of the 2000 Kyoto International Conference on Digital Libraries: Research and Practice*, Nov. 2000.
7) G. Salton, J. Allan, and C. Buckley. Approaches to Passage Retrieval in Full Text Information Systems. In *Proc. of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 49–58, June/July 1993.
8) D. Shin. XRS: XML Retrieval System. http://dlb2.nlm.nih.gov/%7Edwshin/xrs.html.
9) D.Shin, H.Jang, and H.Jin. BUS: An Effective Indexing and Retrieval Scheme in Structured Document. In *Proc. of the 3rd ACM Conference on Digital Libraries*, pp. 235–243. ACM, June 1998.
10) B.Zhao. Xset 2.0. http://www.cs.berkeley.edu/%7Eravenben/xset/, June 2000.