Vol.2009-DBS-148 No.1
Vol.2009-FI-95 No.1
2009/7/28

索引語の統計量を用いた XML 部分文書検索法の組合せ利用とその効果

欅　悼志[†1]　波多野　賢治[†2]　宮崎　純[†3]

本稿では，構造化文書に対する高精度検索のための重要部分抽出技術を取り入れた情報検索法について述べる．我々がこれまでに行った研究では，部分文書中に含まれるクエリキーワード数を考慮することで検索精度に改善が見られるという知見が得られたが，その一方で，部分文書に対してスコアリングを行う際に考慮すべき要件のうち，クエリに対する部分文書の適正性や部分文書のや大きさに関する要件は依然として満たせていなかった．そこで，提案手法ではこれらの問題に対して，テキストノードの持つ索引語数や頻度など，これまで情報検索の分野で用いられてきた文書から抽出される統計量を利用することで解決を目指す．評価実験を行った結果，提案手法は既存の構造化文書に対する検索技術よりも高い精度を示すということが判明した．さらに，提案手法は従来手法に比べて，テキストサイズの大きな部分文書ほど高スコアと判定される傾向があるという知見が得られた．

Advantages of XML Fragment Retrieval Method Considering Query-Oriented Statistics

Atsushi Keyaki,[†1] Kenji Hatano[†2] and Jun Miyazaki[†3]

In this paper, we report the advantages of a scoring method for searching XML fragment considering the query-oriented statistics. We believe that retrieved XML fragments should be scored considering not only traditional retrieved-document-oriented statistics like the tf-ipf scoring but also query-oriented ones such as constituent rate of query keywords and statistics of the query results. From our experimental evaluation, we could find that considering the query-oriented statistics helped to improve the retrieval accuracies of XML search engine. We also fould that the XML fragments containing large-size text nodes have a greater tendency to be given a large score compared with ones containing small-size text nodes.

1. はじめに

構造化文書の中でも特に XML (Extensible Markup Language)[1] はデータ交換のための標準フォーマットとしてさまざまな分野で利用されており，今後莫大な量のデータが XML で作成されることが見込まれる．従って，XML 文書に対する情報検索技術は非常に重要な技術として考えられ，現在盛んに研究が行われている．

XML 文書を木構造と見立てた場合に，それぞれの要素ノード以下，すなわち，それぞれの部分木を一つの単位と見なすことが可能となる．つまり，XML 文書に対する情報検索では，文書単位よりも細かな単位を検索粒度とした情報検索を行うことができる [⋆1]．

これまでの検索粒度を文書単位とする情報検索では，文書全体に対してユーザの情報要求を満たす部分，すなわち正解部分のサイズが小さい場合に，その文書は正解として見なされずに見落とされる可能性があった．その一方で，検索粒度を部分文書とする検索では如何なる部分文書単位においても検索を行うことが可能であるため，正解がごくごく小さい場合にも適切に正解部分文書を抽出することができる．また，仮に正解文書が正しく提示されたとしても，ファイルサイズの大きな文書が提示された場合に文書中から正解部分を発見することはユーザにとって大きな負担となるという問題が存在する．このような問題に対しても，検索単位を部分文書とする検索では，ユーザの求める粒度での文書中の重要部分を提示することで，文書から正解部分を抽出する際に大きな手助けになると考えられる．

これまでの構造化文書に対する情報検索に関する研究では，検索精度の向上を目的とした効果的な検索を目指す研究よりも，検索速度の向上を目的とした効率的な検索を目指す研究に主眼が置かれて研究されてきた．なぜなら，部分文書単位での情報検索はその性質上，一つの文書から複数の部分文書が取り出されるため，検索対象となる部分文書数とデータ量が元文書と比較して膨大になってしまい，その結果検索速度が低下するという問題が生じるからである．しかし，昨今のコンピュータやデータベースマネージメントシステムの高性能化の影響で，大規模データに対してもクエリが発行されてから検索結果を返すまでの

†1 同志社大学大学院文化情報学研究科
Graduate School of Culture and Information Science, Doshihsa University
†2 同志社大学文化情報学部
Faculty of Culture and Information Science, Doshisha University
†3 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

⋆1 本稿では部分木からなる文書を部分文書と呼ぶ．

Vol.2009-DBS-148 No.1
Vol.2009-FI-95 No.1
2009/7/28

処理時間は短縮されてきているため，近年は高精度検索にも注目が集まっている．高精度検索では，検索結果のうちユーザの情報要求を満たす部分文書をより上位に配置することを目指す．その際，我々は重要部分抽出の分野の研究で得られた知見を利用する．なぜなら，文書中からもっともクエリに相応しい部分文書を抽出する際に，その部分文書は文書中の重要部分に相当すると考えられるため，重要部分抽出の際の要件を満たしている部分文書はクエリに対する最適部分となると考えられるためである．これまでの高精度検索を目指す研究においては，各部分文書に対してスコアリングを行い，高スコアと判定された部分文書をクエリに対する最適部分としているが，このスコアリング法は文書中からの重要部分抽出を目的としていない．そのため，これらの部分文書が実際にクエリに対して最適部分であるとは言い切れない．そのような経緯を踏まえ，我々は過去の研究で，重要部分抽出に関する研究で得られた知見を利用した三種類のスコアリング法と，それらのスコアを統合するためのスコア統合法を提案した[11]．しかし，評価実験の結果，クエリに対する部分文書の適正性や部分文書のや大きさを適切に考慮できていないという問題が明らかになったため，したがって，重要部分抽出を考慮した情報検索を目指す上では，部分文書の持つ情報を上手く利用してこれらの問題を解決しなければならない．

そこで本稿では，過去の研究で有用性が確認できた手法を活かしつつ，依然未解決の課題の解決に取り組む．それらの試みが成功すれば，これまで考慮されていなかった重要部分の要件を考慮した構造化文書に対する情報検索が可能となる．その結果，ユーザのクエリに対して検索結果を返す際に，正解文書を提示するだけではなく，ユーザの情報要求に合致すると思われる部分を文書中から抽出してユーザに提示できる．

2. 関 連 研 究

2.1 効果的な検索を目指す研究

XML 文書には，データ指向 (data centric) XML と呼ばれる一つのテキストノードに含まれる索引語数が少ない XML 文書と，文書指向 (document centric) XML と呼ばれる一つのテキストノードに含まれる索引語数が多い XML 文書の二種類存在すると言われている．データ指向 XML ではクエリキーワードを含むノード周辺が正解となると考えられるため，クエリに対する正解部分文書を特定することが比較的容易である．そのため，データ指向 XML に対する検索では情報抽出の分野において効率的な検索，つまり主に検索速度の高速化を目指している．これに対して，文書指向 XML はクエリキーワードが含まれる箇所を発見するだけでは正解を特定することが難しく，クエリに対する正解部分文書を決定することが比較的困難である．そのため，文書指向 XML に対する検索では情報検索分野において効果的な検索，つまり主に検索精度を向上させる研究が進められている．

ところで，これまで情報検索分野で行われてきた構造化文書に対する情報検索に関する研究について述べる．文書検索では，多くの検索システムにおいてキーワードを入力してクエリとすることが一般的であったが，文書指向 XML に対する部分文書検索では文書の論理的構造とキーワードの両方を指定するクエリが一般的に用いられる．そのため，これまでの研究で文書に対する情報検索技術として用いられてきた TF-IDF 法[3] を部分文書検索用に拡張し，索引語が出現するノードへの XPath を考慮した索引語の頻度情報から求める TF-IPF 法[4] が用いられることが多い．TF-IPF 法とは，部分文書内での索引語の重要度を表す TF (Term Frequency) 値とある XPath において索引語自体の重要度を表す IPF (Inverse Path Frequency) 値を掛け合わせることで部分文書の重要度を算出する方法であるため，この値を各部分文書に対して計算することでクエリに対して適切な部分文書を選択することができる．クエリキーワード群を T とし，ある部分文書 s の索引語 $t_i(\in T)$ の TF 値を $tf_f(s,t_i)$，IPF 値を $ipf_f(s,t_i)$ とすると，その部分文書の TF-IPF 値 $S_f(s)$ は以下の式で表すことができる．

$$tf_f(s,t_i) = \frac{n(s,t_i)}{l(s)},\quad ipf_f(s,t_i) = 1 + \log\frac{M(s)}{m(s,t_i)} \tag{1}$$

$$S_f(s) = \sum_{t_i} tf_f(s,t_i)\cdot ipf_f(s,t_i) \tag{2}$$

ただし，$n(s,t_i)$ は部分文書 s 中の索引語 t_i の出現数であり，$l(s)$ は部分文書 s の索引語数，$M(s)$ をクエリによって指定される XPath を満たす部分文書数，$m(s,t_i)$ をクエリによって指定される XPath を満たしつつクエリキーワード t_i を含む部分文書数とする．TF-IPF 法を用いた構造化文書に対する情報検索は，高精度検索において大きな成果が上がっている[12] が，複雑な計算式を用いる必要があることや，検索を行う前に予め多くの処理を行っておく必要があるためデータの更新が困難であるという問題がある．

なお，TF-IPF 法を更に高精度情報検索手法に発展させた研究も存在する．波多野らの研究では，部分文書のスコアである TF-IPF 値とクエリに関するスコアである TF-IAF 値を組み合わせる手法を提案した[5]．TF-IAF 値は，XQuery Full-Text クエリの持つ論理的な構造情報を利用して，指定された構造を満たす部分文書の数を考慮することで求められ，以下の式で表すことができる．

Vol.2009-DBS-148 No.1
Vol.2009-FI-95 No.1
2009/7/28

$$tf_q(t_i) = w(t_i), \quad iaf_q(t_i) = 1 + \log \frac{V(p)}{v(p,t_i)} \tag{3}$$

$$S_q(s) = \sum_{t_i} tf_q(t_i) \cdot iaf_q(t_i) \tag{4}$$

ただし，$w(t_i)$ は クエリキーワード t_i の出現回数，$V(p)$ を指定された XPath を満たす部分文書の数，$v(p,t_i)$ を指定された XPath を満たしつつクエリキーワード t_i を持つ部分文書の数とする．以上より，最終的に部分文書 s のスコア $S_{f\cdot q}(s)$ は以下の式で表される．

$$S_{f\cdot q}(s) = S_f(s) \cdot S_q(s) \tag{5}$$

2.2 重要部分抽出に関する関連研究

我々の提案手法では，重要部分抽出分野の研究で得られた知見を利用する．なぜなら，文書中からもっともクエリに相応しい部分文書を抽出することは，文書中から重要部分を抽出することと同義であると見なせるからである．そのため，重要部分抽出の際に考慮しなければならない要件を考慮し，文書中から部分文書を抽出する．重要部分抽出技術 (重要文抽出技術) はさまざまな研究に応用されており，その一つとして情報検索分野でも利用されている．Web 検索システムのスニペット (Results Snippet) は Web 文書の自動要約文であるが，これは重要部分抽出技術の一種である．スニペットは情報検索を行う上で大きな成果を上げており，高見らの研究らでは自然言語処理によるアプローチによって文書中から重要文を抽出し，スニペットを生成している[9]．また，クエリ依存のよる重要文抽出手法と文書依存による重要文抽出手法を組み合わせている．本提案手法が文書全体のうち文書中クエリに最適な一部分を抽出するのに対して，高見らの研究では文書全体から重要と判断される文を抜き出すという違いはあるが，情報検索精度向上のために重要部分抽出に関する研究を踏襲しているという面では，同じ方向を目指しているといえる．

2.3 重要部分抽出を利用した情報検索技術に関する関連研究

2.1，2.2 節を踏まえて，我々は過去に構造化文書に対する重要部分抽出を利用した情報検索法を提案した[11]．情報検索に関するさまざまな研究を包括的に纏めた書籍である Introduction to Information Retrieval[8] によると，重要部分を抽出する際には三つの要件があり，具体的には，

- クエリキーワードに関する最大限の情報を盛り込む
- 短くまとめて要約する
- 文書中の用語を用いる

の三つである．これらを考慮して重要部分を抽出するには，以下の三つの課題が存在する．

(1) 仮に部分文書中においてクエリキーワードが頻出していたとしても，特定のクエリキーワードのみが出現している場合には，その部分文書が重要部分とは言い切れない．

(2) 部分文書中に不要な情報が大量に含まれていると，ユーザのクエリに対する最適部分とは言い難い．

(3) 部分文書の大きさを考慮し，必要以上に大きすぎる部分がクエリに対する最適部分とならないようにしなければならない．

過去の研究[11] では，部分文書の構造情報を最大限利用しつつ極力計算量を抑えた手法にて課題解決を試みた．(1) に対しては，部分文書中に含まれるクエリキーワードの種類数，(2) に対しては部分文書に含まれるすべてのテキストノードのうち，クエリキーワードを含むテキストノードの割合，(3) に対しては部分文書の最小接続木[6] のエッジの本数を利用することで課題の解決を試みたが，(2), (3) の課題は適切に解決できていないということが判明した．その原因として，(2), (3) の課題を解決するためのスコアリング法において，テキストノード中にクエリキーワードが出現するかどうかのみに着目し，そのテキストサイズや索引語数は考慮しなかったことが不適切であったためであると考えられる．つまり，索引語数が多いテキストノードも索引語数が少ないテキストノードもすべて同じ重みとして扱ったために，部分文書を木と見立てた際のエッジの本数が同じであっても部分文書の索引語数に大きな差が生まれてしまい，中には索引語数が少ない部分文書が大きな部分文書だと判定されたり逆に索引語数が大きな部分文書にも関わらず小さな部分文書であると判定されてしまった．このように，これらの手法では課題 (2), (3) を解決するため必要な統計量を上手く抽出できていないことが原因となって，適切に課題を解決できないということが判明した．そのため，適切に各部分文書に対してスコア付けするためには，各部分文書のテキスト全体に対して評価を行う必要がある．また，未解決の二つの課題の解決にはそれぞれ部分文書のテキストノードの索引語の生起状況の統計量や部分文書のテキストサイズを考慮することで解決することができるということが判明した．また，課題 (1) を解決できたスコアリング法キーワード含有率スコア $S_c(s)$ は以下の式で表すことができる．

$$S_c(s) = \frac{j(s)}{k} \tag{6}$$

ただし，$j(s)$ は部分文書 s に含まれるクエリキーワード数，k を T に含まれるクエリキーワード数とする．

Vol.2009-DBS-148 No.1
Vol.2009-FI-95 No.1
2009/7/28

3. 提案手法

2.3 節の未解決の二つの課題の解決法を提案する．課題 (2) の解決には，部分文書のテキスト全体のうち，クエリに関する部分の割合を抽出する必要がある．また，課題 (3) の解決には，部分文書のテキストサイズを抽出する必要がある．これらを言い換えれば，部分文書中にクエリに関する索引語が多ければ多いほど，クエリに関係のない索引語が少なければ少ないほど部分文書に高スコアを与えればよいということである．ここで，先ほども述べたように，テキストノードが持つ索引語数には大きな幅がある．そのため，部分文書に含まれるテキストノード数からは各部分文書の索引語数を推測することができず，部分文書のテキスト全体が持つ索引語の生起状況の統計量やテキストサイズを利用して上記の二つの課題を解決しなければならない．その際，2.1 節で挙げた既存の情報検索技術を利用することで二つの課題を解決できると考えることができる．なぜなら，既存の情報検索技術では，部分文書のクエリに関する索引語数をテキスト全体に含まれる索引語数で正規化しており，これはクエリに関する索引語の割合を求めることと同義であるため課題 (2) を解決していると考えられる．また，既存の情報検索技術では，正解とは考えにくいテキストサイズの小さな部分文書にはスコアが大きくなりすぎないように規制をかけており，これは部分文書のテキストサイズを考慮していると考えられるために課題 (3) も解決することができる．なお，クエリに関してキーワードと論理構造両方をスコアリングに利用しているという点で，より厳密にクエリに関する情報を考慮しているモデルである TF-IAF を利用する．よって，式 (5) と，式 (6) を組み合わせることで，重要部分抽出の三つの要件を満たすための三つの課題を解決する．従って，最終的な部分文書のスコアは以下で求める．

$$S(s) = S_c(s) \cdot S_{f \cdot q}(s) \tag{7}$$

4. 評価実験

4.1 実験の目的と評価尺度

提案手法の妥当性を検証するため，後述するテストコレクションを用いて評価実験を行った．また，重要部分抽出を考慮していない従来手法[5] と検索精度を比較することで，重要部分抽出の要件を満たすことの効果を検証した．その際，評価尺度には MAiP (Mean Average interpolated Precision) と MAgP (Mean Average generalised Precision) を用いる[7]．本評価実験における MAiP とは，情報検索において従来から用いられてきた再現率と精度から求められる値であり，x 軸に再現率，y 軸に精度を用いて描かれるグラフ，つまり再現率–精度グラフにおける再現率が 0% から 100% までの一定の区間ごとの各点における精度の値を求め，すべての点での精度の値を平均した値である．今回の実験では，適合率を 1% 刻みに取る，計 101 点に対して MAiP を求めた．また，MAgP とは MAiP における iP (補完適合率) の代わりに gP (ページ中の正解のうち，抽出出来ている割合) を用いて 11 点で評価したものである．

4.2 テストコレクション

評価実験には，2008 年版の INEX (INitiative for the Evaluation of XML Retrieval) テストコレクションを用いた．INEX テストコレクション は，XML 部分文書検索のための国際プロジェクトである INEX Project[⋆1] によって 2002 年より構築作業が行われている XML 部分文書検索システム用の性能評価データセットである．検索対象となる約 66 万個の XML 文書で構成される 2008 Wikipedia document collection[2]，285 個のクエリの集合 INEX topics，それらクエリに対する解答部分文書集合及びその評価が記述されている INEX relevance assessments の三つで構成されている．なお，INEX topics には XPath とクエリキーワードを指定する CAS (Content-And-Structure) クエリと，クエリキーワードのみを指定する CO (Content-Only) クエリの二種類が存在する．

4.3 評価用タスク

INEX によって配布されている評価ツールには三種類の評価方法が存在する[7]．一つ目の評価方法は Focused タスクであり，一つの文書から正解として最適な部分文書を一つだけ抽出する．我々がこれまで取り組んできた研究はこのタスクにおいて成果を出すことである．二つ目の評価方法は Relevant in Context タスクであり，文書中から正解部分に該当する箇所をより多く抽出する．三つ目の評価方法は Best in Context タスクであり，こちらも文書のうち正解部分をより多く抽出する．ただし，Best in Context タスクでは正解部分の開始位置との距離から精度評価を行う．よって，Focused タスクで満たさなければならない条件が最も多く，Relevant in Context タスク, Best in Context タスクとなるにつれて満たさなければならない条件が少なくなる．なお，Relevant in Context タスクと Best in Context タスクでは，クエリと関連がある部分文書を列挙するのが目的であるため，部分文書同士が重複しない限りは一つの文書から複数の部分文書を抽出することが可能である．

4.4 評価実験

INEX テストコレクションに含まれているのクエリのうち，解答が用意されているクエ

⋆1 http://www.inex.otago.ac.nz/

Vol.2009-DBS-148 No.1
Vol.2009-FI-95 No.1
2009/7/28

表 1 精度上昇率 (Focused)

	All	CAS	CO
上昇率	1.048%	1.062%	1.043%

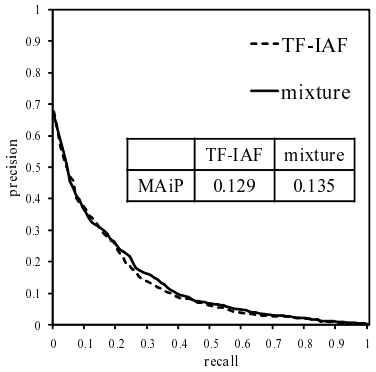

図 1 Focused タスク精度比較 (すべてのクエリ)

リ 70 個に対して 3 節で述べた検索法により高スコアと判断された部分文書上位 1500 件と，比較対象となる従来手法[5] によるスコアリング法により高スコアと判断された部分文書上位 1500 件を抽出して精度を求める．

Focused タスクにて提案手法と従来手法を比較した結果を図 1 に示す．再現率が 0.0% , 0.1% の地点では従来手法の精度が上回り，その後何度か逆転を繰り返した後に 0.15% 以降は常に提案手法の精度が上回り続ける．MAiP を比較した結果，従来手法よりも提案手法の優位性を示せた．また，CAS 検索のみを対象とした精度評価 (図 2) と CO 検索のみを対象とした精度評価 (図 3) を行った結果，提案手法は CAS 検索において，より高い精度を示すということが判明した (**表 1**)．

4.5 考察と今後の課題

4.5.1 包含関係に関する調査

図 1–3 を見れば分かるように，MAiP では従来手法に対して精度で勝っていたが，その一方で検索結果上位における精度が従来手法に劣るという問題が起こった．検索結果のうちユーザが確認するのは上位数件と言われているため[10]，検索結果上位での精度が劣っていることは重大な問題である．そこで，従来手法と提案手法では回答部分文書群，すなわち検索結果上位と判定された部分文書群にどのような関係があるのかを調査した．つまり，

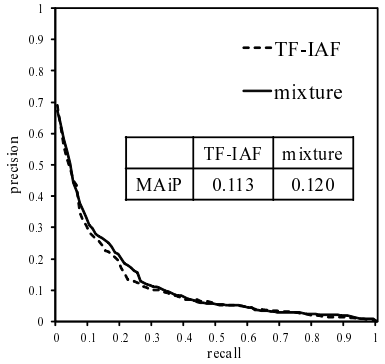

図 2 Focused タスク精度比較 (CAS クエリ)

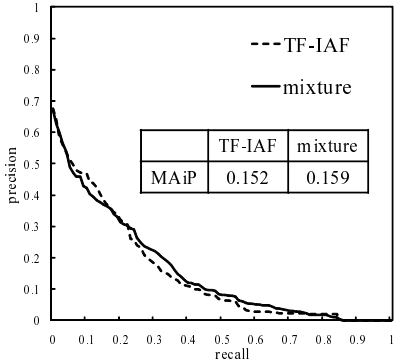

図 3 Focused タスク精度比較 (CO クエリ)

抽出された部分文書のうち同じ文書から抽出された部分文書数や，同じ文書から抽出された部分文書同士にどのような包含関係があるのかを調査した．なお，部分文書間の関係は，回答部分文書群のうち図 1 の精度において従来手法が勝っている上位 1% (15 件) と，提案手法が勝っている上位 2% (30 件)，従来手法と提案手法の精度差がもっとも大きい点 (従来手法がもっとも大きく勝っている 6% ，提案手法がもっとも大きく勝っている 23%) の計 4% でそれぞれの部分文書の関係を比較し (**表 2**–**表 5**)，共通に出現した文書の割合を求めた (**表 6**)．なお，従来手法のスコアリングで抽出された回答部分文書群を A，A に含まれる要素を a，提案手法のスコアリングで抽出された回答部分文書群を B，B に含まれる要素を b とし，回答部分文書同士に同じ文書から抽出された文書があればそれぞれの部分文書の包含関係を調査した．また，いずれか一方の回答部分文書群のみに出現した部分文書はその文書数のみカウントした．上記の 4 点で計測した理由は，任意の 2 点で比較した結果何らかの傾向があったとして，それは手法による影響から起こっているのか，比較件数による影響から起こっているのか判断できないためである．

これらの結果からは，手法や件数によって抽出される回答文書に傾向は見られなかった．また，いずれの点においても $a \supset b$ という関係は全く見られなかった．これは，提案手法におけるスコア結果を算出するために，従来手法である $S_{f \cdot q}(s)$ とキーワード含有率スコアである $S_c(s)$ を掛け合わせる際に，ある文書中の $S_{f \cdot q}(s)$ においてある先祖ノードを根とする

Vol.2009-DBS-148 No.1
Vol.2009-FI-95 No.1
2009/7/28

表 2 部分文書の包含関係 (上位 15 件)

関係	確率	同一文書内確率
a⊂b	0.797%	1.73%
a⊃b	0.00%	0.00%
a=b	43.2%	93.8%
a ∩ b = ϕ	2.07%	4.50%
A only	27.6%	
B only	26.4%	

表 3 部分文書の包含関係 (上位 30 件)

関係	確率	同一文書内確率
a⊂b	1.03%	2.28%
a⊃b	0.00%	0.00%
a=b	42.0%	93.3%
a ∩ b = ϕ	1.97%	4.39%
A only	28.1%	
B only	26.9%	

表 4 部分文書の包含関係 (上位 90 件)

関係	確率	同一文書内確率
a⊂b	1.63%	3.24%
a⊃b	0.00%	0.00%
a=b	47.0%	93.2%
a ∩ b = ϕ	1.77%	3.51%
A only	25.3%	
B only	24.2%	

表 5 部分文書の包含関係 (上位 345 件)

関係	確率	同一文書内確率
a⊂b	1.84%	3.56%
a⊃b	0.00%	0.00%
a=b	47.9%	92.8%
a ∩ b = ϕ	1.87%	3.63%
A only	24.8%	
B only	23.6%	

表 6 共通回答文書割合

	上位 30 件	上位 60 件	上位 90 件	上位 345 件
共通	46.1%	45.0%	50.4%	51.6%
片方	53.9%	55.0%	49.6%	48.4%

部分文書 ⋆1が対応する子孫ノードを根とする部分文書 ⋆2 よりもスコアが高かった場合に，子孫部分文書に含まれるクエリキーワードは必ず先祖部分文書にも含まれるため，$S_c(s)$ を掛け合わせたとしてスコアに差が開くことはあっても，決して差が縮まることはない．よって $S_c(s)$ を掛け合わせても決してそのスコアの優劣は変わらないためである．それに対して $S_{f\cdot q}(s)$ において子孫部分文書が先祖部分文書よりもスコアが高かった場合は，$S_c(s)$ を掛け合わされた際に先祖部分文書が子孫部分文書よりも多くのクエリキーワードを含んでいた場合にスコアの優劣が逆転する可能性があるため，$a \subset b$ が確認された．さらに，表 6 のように，共通回答文書の差からも何かしらの傾向は確認できなかった．そのため，ランキング上位において提案手法が従来手法に対して劣っているのは部分文書の包含関係や抽出される文書以外の原因があると思われる．そのため，他の視点から原因を探ることを試みた．

⋆1 以降，これを先祖部分文書と呼ぶ．
⋆2 以降，これを子孫部分文書と呼ぶ．

表 7 回答部分文書の統計量

	TF-IAF	mixture
平均テキストサイズ	180	268
分散	45577	137328
最大	4428	7724
最小	26	26

4.5.2 Relevant in Context タスクと Best in Context タスクの評価実験

表 1 から読み取れる結果として，提案手法を用いた場合に CO クエリよりも CAS クエリに対して優位性が確認された．そのため，そのような結果となった要因を調査した．表 2–表 5 の a⊂b と a⊃b の割合から，提案手法と従来手法では提案手法でより大きな部分文書が含まれている可能性が高いということが示唆された．従って，実際に部分文書のテキストサイズに差があるのか，あるとすればどのような差があるのかを確認するためにそれぞれの手法の回答文書にの統計量を求めた (**表 7** 参照)． その結果，想定した通り提案手法における平均テキストサイズは大きいということが判明した．それは，キーワード含有率スコアが高い部分文書というのは部分文書中に含まれるクエリキーワード数が多いため，比較的テキストサイズの大きな部分文書が該当するためである．また，提案手法の分散は従来手法の分散よりも大きくなったことより，提案手法では抽出される部分文書のテキストサイズに大きく散らばりが生じるということが判明した．つまり，提案手法の回答部分文書にはテキストサイズの大きな部分文書からテキストサイズの小さな部分文書まで幅広く存在しているといえる．そのうち，テキストサイズの小さな部分文書は共通して出現しているため，テキストサイズの大きな部分文書の影響で検索精度が向上したと考えられる．従って，テキストサイズの大きな部分文書が利用者の情報要求に合致する部分文書となる可能性があると推測される．しかし，現時点でそれを断定することはできず，利用者の情報要求に合致する部分文書がどのようなサイズの部分文書として生起しているかを調査するため，INEX テストコレクションを詳細に分析する必要がある．

上記の結果からは CAS クエリにおける優位性を説明できない．そこで，Relevant in Context タスクと Best in Context タスクにおいても精度評価を行った．ただし，その目的は Relevant in Cntext タスクと Best in Context タスクにおける提案手法の有効性を計るためではなく，Focused タスク，すなわち最適部分のみを抽出するタスクに対して，実際どの条件を満たすことができているのか，また，どの条件を満たせていないのかを確認するための評価実験である．そのため，回答部分文書はすべてのタスクにおいて同じ文書を用いた．

Vol.2009-DBS-148 No.1
Vol.2009-FI-95 No.1
2009/7/28

表 8 精度上昇率 (Relevant in Context)

	All	CAS	CO
上昇率	1.021%	1.081%	1.136%

表 9 精度上昇率 (Best in Context)

	All	CAS	CO
上昇率	1.025%	1.048%	1.278%

表 10 テキストサイズ上昇率

	All	CAS	CO
上昇率	1.393%	1.361%	1.448%

つまり，Focused タスクと同様に一つの文書から一つの部分文書のみを抽出した．Relevant in Cntext タスクの評価実験の結果が図 4–6，Best in Context タスクの評価実験の結果が図 7–9 である．また，**表 8**, **表 9** の結果から，いずれのタスクにおいても CAS クエリよりも CO クエリに対して検索精度の上昇率が高いということが判明した．以上のことをまとめると，

- 最適部分のみを抽出する Focused において，CO クエリよりも CAS クエリに対して有効
- 正解部分をより多く含むことを目的とする Relevant in Context において，CAS クエリよりも CO クエリに対して有効
- 回答部分文書の開始位置と正解部分の開始位置を近づけることを目的とする Best in Context において，Relevant in Context よりも更に顕著に CAS クエリよりも CO クエリに対して有効

となる．つまり，満たさなければならない条件が少ないタスクほど CO クエリに対して有効性を発揮するということがわかる．これと先ほど得られた，提案手法は従来手法よりもテキストサイズが大きくなるという傾向があることから，タスクにおいて満たさなければならない条件が少ないほど大きな部分文書が高スコアとなりやすいという知見が得られた．また，**表 10** より，CAS クエリよりも CO クエリにおいて部分文書のサイズは大きくなるということも判明した．これは，CAS クエリでは，ユーザの指定した構造を持たないノードはいくら多くのクエリキーワードを含んでいてもスコアリング対象とならないためであると考えられる．これらすべてを踏まえると，Focused タスクの CO クエリにおける検索精度を向上させるためには，部分文書の大きさが大きくなりすぎないように制限しなければならないということが示唆される．そのため，クエリに関係しない部分が多くを占める部分文書には，何らかのペナルティを設けて高スコアと判定されないように規制する必要がある．

上記から，全体的な検索精度を向上させるための取り組み方法は示唆されたが，依然として検索結果上位の精度を向上させるための取り組み方法に関しては不明である．そのため，検索結果上位において精度を向上させるために必要な条件を調査する必要がある．

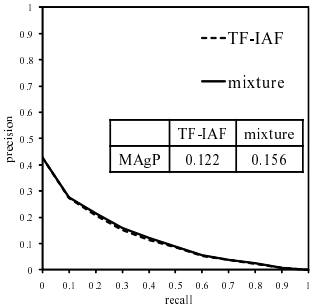

図 4 Relevant in Cntext タスク精度比較 (すべてのクエリ)

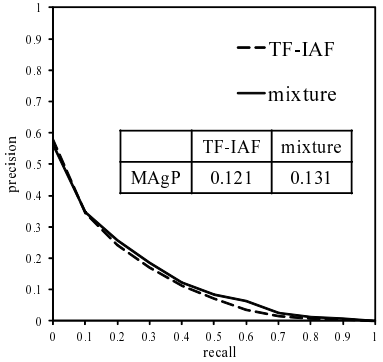

図 5 Relevant in Cntext タスク精度比較 (CAS クエリ)

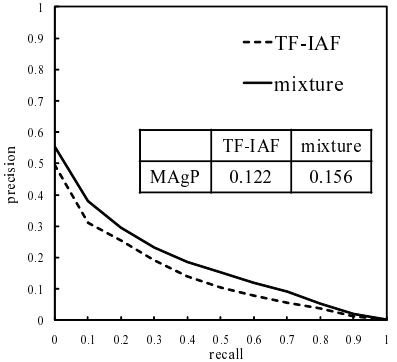

図 6 Relevant in Cntext タスク精度比較 (CO クエリ)

5. おわりに

本稿では，構造化文書に対する高精度検索を目指すための過去に考案した検索法を改良し，その評価を行った．すなわち，従来の手法では考慮されてこなかった重要部分抽出の際に考慮すべき要件を考慮し，部分文書中に含まれるクエリキーワードの数から得られる統計量と既存の情報検索技術の統合を行った．その結果，既存の検索法よりも精度における優位

Vol.2009-DBS-148 No.1
Vol.2009-FI-95 No.1
2009/7/28

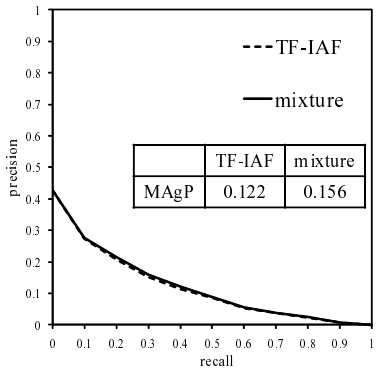

図 7 Best in Context タスク精度比較 (すべてのクエリ)

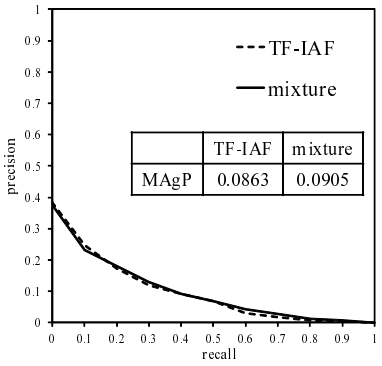

図 8 Best in Context タスク精度比較 (CAS クエリ)

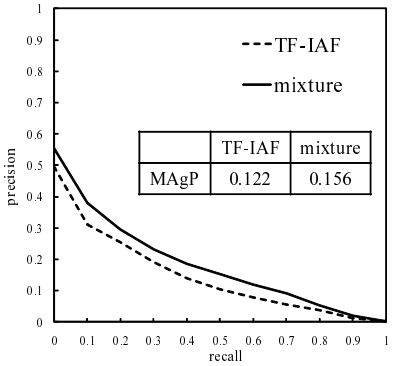

図 9 Best in Context タスク精度比較 (CO クエリ)

性を示した．なお，提案手法では従来手法に比べて部分文書のテキストサイズが大きくなるということが判明した．

今後の課題として，Focused タスクにおいて CO クエリにおける回答部分文書のテキストサイズが大きくなりすぎないように考慮しなければならず，また，検索結果上位において高精度を実現するために必要な条件を調査する必要がある．

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」(課題番号: 19024058) によるものである．ここに記して謝意を表す．

参 考 文 献

1) Tim Bray, Jean Paoli, C.M. Sperberg-McQueen, Eve Maler, and Francois Yergeau. Extensible Markup Language (XML) 1.0 (Fifth Edition). `http://www.w3.org/TR/REC-xml/`, November 2008.
2) Ludovic Denoyer and Patrick Gallinari. The Wikipedia XML Corpus. *SIGIR Forum*, 2006.
3) Salton Gerard and Buckley Christopher. Term-weighting approaches in automatic text retrieval. *Information Processing and Management: an International Journal*, Vol.24, No.5, pp. 513–523, 1988.
4) Torsten Grabs and Hans-Jörg Schek. PowerDB-XML: A Platform for Data-Centric and Document-Centric XML Processing. In *Proceedings of the First International XML Database Symposium*, Vol. 2824 of *Lecture Notes on Computer Science*, pp. 100–117. Springer-Verlag, September 2003.
5) 波多野賢治, シーハムアメルヤヒア, ディベッシュスリバスタバ. XML 情報検索における構造問合せを利用した部分文書スコアリング. 電子情報通信学会技術研究報告, 第 107 巻, pp. 13–18, October 2007. DE2007-117.
6) Vagelis Hristidis and Nick Koudas. Keyword proximity search in xml trees. *IEEE Transaction on knowledge and data engineering*, Vol.18, No.4, pp. 525–539, April 2006.
7) Jaap Kamps, Jovan Pehcevski, Gabriella Kazai, Mounia Lalmas, and Stephen Robertson. Inex 2007 evaluation measures. In *INEX*, pp. 24–33, 2007.
8) ChristopherD. Manning, Prabhakar Raghavan, and Hinrich Schutze. *Introduction to Information Retrieval*. Cambridge University Press, July 2008.
9) 高見真也, 田中克己. 類似性を考慮したスニペットの再生成による検索結果のパーソナライズ. 第 18 回データ工学ワークショップ (DEWS2007) 論文集, March 2007.
10) 中村聡史. 情報信頼に対する信頼性調査および結果. 人工知能会誌, No.6, pp. 767–774, 2008.
11) 欅惇志, 波多野賢治, 宮崎純. 構造化文書の重要部分抽出のためのスコアリングアルゴリズム. 第 1 回データ工学と情報マネジメントに関するフォーラム (DEIM2009) 論文集, March 2009.
12) 清水敏之, 寺田憲正, 吉川正俊. 関係データベースを用いた XML 情報検索システムの開発. 情報処理学会論文誌: データベース, Vol.48, No. SIG11, pp. 224–234, June 2007.