XML検索技術を利用した検索結果の構成手法

欅　惇志†　　波多野　賢治‡　　宮崎　純§

†同志社大学大学院　　‡同志社大学　　§奈良先端科学技術大学院大学

1　はじめに

構造化文書の一つであるXMLはデータ交換の標準フォーマットとして策定されている．そのため，データ交換を目的として，さまざまなデータがXMLにて記述されるようになった．その結果，Web上には膨大な数のXML文書が存在するようになったため，ユーザが情報検索を行う際には多大な困難を伴う．このような経緯から，XML文書の情報検索技術への必要性は高まってきている．

構造化文書の持つ，木構造に見なすことができるという特性を利用して，XML文書はタグで囲まれたそれぞれの部分を一つの文書、すなわち部分文書と見立てて検索対象とすることができる．また，XML検索では一つの文書から複数の部分文書が抽出されるため，部分文書に対する情報検索では文書より細かい粒度での情報提示を行うことが可能である．つまり，従来の文書検索では検索結果として文書を提示してきたが，部分文書検索では文書中のユーザの求める部分を提示することが可能となる．

これまでの部分文書情報検索手法では，文書中のうち，クエリに対してもっとも適合する部分文書一つを検索結果として提示してきた．しかし，クエリに対する最適部分を一つの部分文書で表そうとした場合に，不要な部分が混じる，大きすぎる部分文書が抽出される，適合部分を抽出しきれない，といった問題があるため，クエリに対する適合部分を一つの部分文書で表すことは適切とは言い切れない．これらの問題を解決すべく，本稿ではクエリに対する適合部分として部分文書を一つ抽出するのではなく，複数の部分文書を抽出することで効果的な情報提示手法の提案を行う．

2　関連研究

情報検索における効果的な検索結果の提示方法に関する研究の一つに，スニペット(Result Snippet)が挙げられる．スニペットはWebページの要約文のことであり，ユーザがシステムによって返される検索結果から有用なページを選ぶ際の判断材料として用いる．重要文抽出技術によって生成され，情報検索を行う上で大きな成果を挙げている．また，クエリ依存と文書依存の二つのアプローチからスニペットの生成を行うことで，スニペットのパーソナライズ化を目指す研究も存在する[2].

一方，XML文書検索において，クエリに対する適合を効率的に発見する手法では，その検索精度が低下するという問題が発生している．eXtruct [1]はそのような問題を解決するために，効率的に抽出された適合部分に対して，タグの分類や意味的な照合，クエリの解析を行うことで有益な検索結果に再構成を行う．

3　提案手法

1節でも述べたように，クエリに対する最適部分を一つの部分文書で表すことが適切であるとは言い切れない．例えば，クエリに対する最適部分が文書中の離れた部分に複数個所存在する場合に，すべての箇所を含むと不要な部分が混じったりテキストサイズが大きすぎる部分文書が抽出される可能性がある，もしくは，いずれかの箇所だけを抽出すると他の適合部分を抽出できないという問題が起こる．そこで本稿では，一つの文書からクエリに適合する部分文書を複数抽出することでどのような効果があるのか確認を行った．

本提案手法の概略図を図1に示す．まず，既存の部分文書検索技術を利用して，各部分文書に対して得点付けを行う(1)．続いて，部分文書を文書番号ごとに纏め(2)，先ほど付与された得点を用いて部分文書の抽出を繰り返すことで文書ごとに検索結果，すなわち回答部分文書を作成する(3)．その際，大きな得点が付与された部分文書から順番に回答部分文書に対して挿入を行う．なお，部分文書抽出処理は，回答部分文書に含まれるテキストノードのテキストサイズが抽出閾値ELを超えない限り繰り返される．ただし，ELは以下の式で表される．

$$EL = \alpha \cdot textsize(root) \tag{1}$$

このとき，$\alpha(0 \leq \alpha \leq 1)$は文書中の適合文書の割合を表すパラメータであり，$textsize(root)$は回答部分文書の作成対象の文書のテキストサイズ，すなわち全索引語数である．文書中の適合部分は，文書全体のうちの

Atushi KEYAKI †, Kenji HATANO ‡, and Jun MIYAZAKI §
†Graduate School of Culture and Information Science, Doshisha University
‡faculty of Culture and Information Science, Doshisha University
§ Graduate School of Information Science, Nara Institute of Science and Technology

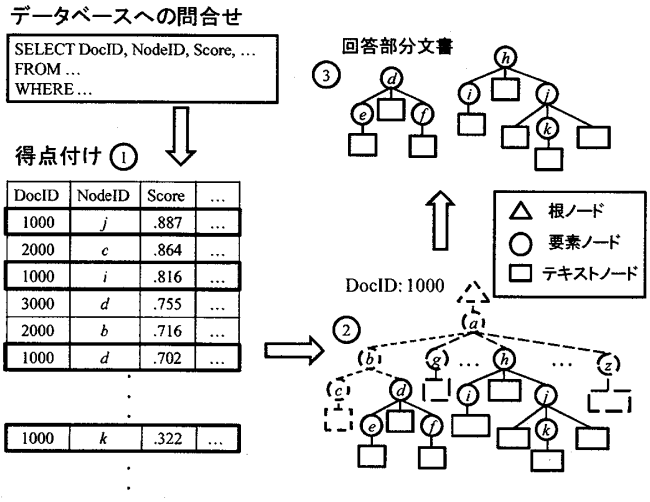

図 1: 回答部分文書作成手順

表 1: α の変動による精度への影響

α	0.1	0.2	0.3	0.4	0.5
MAiP	.119	.131	.138	.146	.151
α	0.6	0.7	0.8	0.9	1.0
MAiP	.158	.162	.164	.164	.167

一定の割合以下の部分であると考えられるため，特定の文書から大きすぎる部分文書が抽出されることを防ぐために α の設定を行う．

4 評価と考察

提案手法の効果を確認するため，部分文書検索技術 (従来手法) と提案手法の評価比較を行った．なお，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．従来手法と提案手法の比較を行う前に，まずはパラメータ α の最適値の調査を行った．0.1 から 1 までの 0.1 刻みでの精度評価実験を行った結果，$\alpha = 1$ において最高精度を示した (表 1 参照)．ここでは，評価尺度には MAiP (Mean Average interpolated Precision) を利用した．この結果から，回答部分文書の大きさを制限することよりも，解として相応しくない比較的スコアの低い文書を抽出しないようにすることで，効果的な検索に結びることが適切であると判明した．ただし，文書によってそのテキストサイズに大きなばらつきがあるため，今後 *EL* を文書中の割合ではなく定数で制限することでどのような効果が発生するのかについても考慮する必要がある．

また，評価実験の結果，最高得点の部分文書を一つ抽出する従来手法と比較し，複数の部分文書を抽出する提案手法は 3% 程度高い精度を示すという結果が得られた (図 2 参照)．このことから，部分文書検索において一文書から複数の部分文書を抽出することで検索精度の向上に結び付けられると考える．

*http://www.inex.otago.ac.nz/

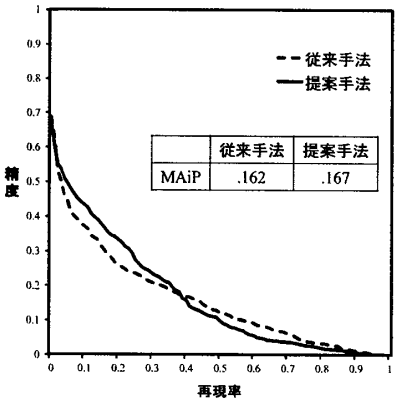

図 2: 精度比較

5 まとめ

本稿では，XML 検索技術を利用した検索結果の構成手法の提案を行った．評価実験の結果，一つ文書から複数の適合部分文書を抽出することで精度の向上が見られた．更に，文書中の適合部分の割合を制限せず，得点の高い部分文書を優先して回答部分文書として抽出することで，より高精度検索を実現できることが確認された．また，今回得られた知見より文書から複数の部分文書を抽出することの効果が認められたので，今後は具体的に部分文書を抽出する際の手法についても調査を行う必要がある．

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 (課題番号: 21013035)，日本学術振興会科学研究費補助金 若手研究 (B) (課題番号: 20700227) によるものである．ここに記して謝意を表す．

参考文献

[1] Y. Huang, Z. Liu, and Y. Chen. Query Biased Snippet Generation in XML Search. In *Proceedings of the 2008 ACM SIGMOD*, pp. 315–326, June/July 2008.

[2] 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．DEWS2007 論文集 (C7-4), March 2007.