様　式　Ｆ－７－１

科学研究費助成事業（学術研究助成基金助成金）実施状況報告書（研究実施状況報告書）（平成２４年度）

1．機関番号　1 4 6 0 3　　2．研究機関名　奈良先端科学技術大学院大学

3．研究種目名　基盤研究(C)　　4．補助事業期間　平成２３年度～平成２５年度

5．課題番号　2 3 5 0 0 1 2 1

6．研究課題　メニーコアプロセッサ時代における構造化文書の高精度かつ高速検索の実現

7．研究代表者

研究者番号	研究代表者名	所属部局名	職名
4 0 2 9 3 3 9 4	ミヤザキ　ジュン 宮崎　純	情報科学研究科	准教授

8．研究分担者

研究者番号	研究分担者名	所属研究機関名・部局名	職名

9．研究実績の概要

XML部分文書検索の研究は、これまで静的な文書集合に対して検索の高精度化と高性能化、すなわち検索結果中に含まれる適合文書の割合をいかに高めるか、あるいはいかに高速に検索を行うかにのみ重点がおかれ、従来の文書検索と比べてXML部分文書検索は根本的に計算コストが大きいにも関わらず、文書集合の更新と高精度・高性能検索とを両立させる研究がなされてこなかった。

本研究では、高精度XML部分文書検索に関して、実用性の観点から、文書の更新に対しても高い検索精度を保ちつつ、高速な更新処理と検索処理を行う手法について研究を行った。XML部分文書検索では、扱うデータ量が従来の文書検索と比較して数十倍程度多く、また、新たに出現したトピック中の索引語を正確に重み付けするには、更新処理に工夫が必要である。特に、単語の重み付けを正確に行うためには、統計的に十分な量の文書が必要となる。そのため、更新により新たに出現したトピック中の索引語についても正確な重み付けが可能なよう、本研究では類似クラスの部分文書を集約して、十分な統計情報を得る手法を提案した。また、更新オーバヘッドを軽減するために不要な部分文書と索引語を排除するフィルタを提案した。67万文書からなるテストコレクションを用いた評価実験の結果、XML文書を更新可能な単純なシステムと比較して更新処理を25%高速化した。検索処理に関してもtop-kアルゴリズムを適用して処理時間を0.5秒程度に短縮化し、さらに検索結果を再構成するアルゴリズムを適用して検索精度を従来よりも4%向上させることが可能であることを示した。

以上により、動的にXML文書が更新される現実的な環境においても、実用的な性能で動作可能なXML部分文書検索システムを実現可能であることを示した。

10.キーワード

(1) XML文書検索	(2) 更新処理	(3) 問合せ処理	(4) 高性能計算
(5)	(6)	(7)	(8)

11.現在までの達成度

（区分）（2）おおむね順調に進展している。

（理由）

本研究の目標である、文書の更新を考慮しつつ、高速かつ高精度のXML部分文書検索を可能とする課題に関して、本年度は検索処理の速度向上ならびに検索精度の向上を目指した。達成した研究目標の概要は以下の通りである

まず検索処理を高速処理するために、top-kアルゴリズムを適用し、それを効率よく処理するためのランダムアクセス索引を取り入れた。これにより単純な検索処理は0.5秒程度となり、実時間処理を可能とした。その一方で、従来の静的な文書集合のみ扱うシステムと比較して、正確な統計量を計算できないため検索精度が下がるという問題に対して、索引語の重み計算方法で代表的なものを精査して最適な計算方法を選定し、さらに検索結果を再構成して結果をランキングし直す手法を適用し、インタラクティブ性を損なうことなく静的な文書集合を扱うシステムよりも高い精度を実現した。

以上のように現実的なXML部分文書検索システムの構成手法を明らかにできたことから、研究の達成度は十分であると考えている。

12.今後の研究の推進方策　等

（今後の推進方策）

今年度の研究成果より、文書更新可能な高精度かつ高速なXML部分文書検索システムの構成方法を明らかにしたが、文書の統計量計算はそれでも非常に負荷のかかる処理であることが判明している。文書の統計量計算には、大きく分けて、更新された文書から索引語を抽出する過程と、索引語の出現頻度情報を算出する過程の二つに分けられる。後者は単純な数値計算であり、GPUなどの多数の単純コアからなるメニーコアプロセッサが適しており、前者は比較的少数の高機能コアからなるマルチコアプロセッサが向いていると考えられる。

次年度は、メニーコアプロセッサとマルチコアプロセッサを併用して協調処理させることにより、大規模XML文書群からの高速な統計量計算や、更新文書に対する高速な統計量再計算のためのデータ構造ならびにアルゴリズムの設計を行っていく計画である。また、XML文書にとどまらず、さらに現実的な問題であるWeb文書に研究を展開していく予定である。

（次年度の研究費の使用計画）

未使用額が生じた大きな要因は、当初計画していた研究論文の出版が延期され、予算執行計画を変更したことに伴うものである。また、次年度の請求額と合わせての執行計画は以下の通りである。

次年度は、XML文書の統計量計算について、それぞれの処理モジュールをマルチコアプロセッサとメニーコアプロセッサに適材適所で割り当て、協調処理による高速化を目指す方針である。そのために、8コアのマルチコアプロセッサと1500コア程度のメニーコアGPUを搭載する実験用計算機ならびにソフトウェア開発ツール等を購入し、文書統計量計算のための並列処理アルゴリズムの設計と評価を行う計画である。また、研究成果の論文の出版ならびに対外発表にも研究費を使用する予定である。

13.研究発表（平成24年度の研究成果）

〔雑誌論文〕 計(2)件　うち査読付論文 計(2)件

著者名	論文標題 【掲載確定】
Atsushi Keyaki, Jun Miyazaki, Kenji Hatano, Goshiro Yamamoto, Takafumi Taketomi, Hirokazu Kato	Fast Incremental Indexing with Effective and Efficient Searching in XML Element Retrieval

雑誌名	査読の有無	巻	発行年	最初と最後の頁
International Journal of Web Information Systems	有	9	2013	-

掲載論文のDOI（デジタルオブジェクト識別子）
なし

著者名	論文標題
Atsushi Keyaki, Jun Miyazaki, Kenji Hatano, Goshiro Yamamoto, Takafumi Taketomi, Hirokazu Kato	Fast and Incremental Indexing in Effective and Efficient XML Element Retrieval Systems

雑誌名	査読の有無	巻	発行年	最初と最後の頁
Proc. of the 14th International Conference on Information Integration and Web-based Applications & Services	有	-	2012	157 - 166

掲載論文のDOI（デジタルオブジェクト識別子）
なし

〔学会発表〕 計(2)件　うち招待講演 計(0)件

発表者名	発表標題
欅惇志, 宮崎純, 波多野賢治, 山本豪志朗, 武富貴史, 加藤博一	更新を考慮した XML 部分文書検索システムの精度の改善

学会等名	発表年月日	発表場所
第5回データ工学と情報マネジメントに関するフォーラム	2013年03月03日	福島県郡山市

発表者名	発表標題
欅惇志, 宮崎純, 波多野賢治, 山本豪志朗, 武富貴史, 加藤博一	XML部分文書検索における索引の高速な差分更新と高精度検索

学会等名	発表年月日	発表場所
第5回Webとデータベースに関するフォーラム	2012年11月19日	東京都千代田区

〔図書〕 計(0)件

著者名	出版社

書名	発行年	総ページ数

14.研究成果による産業財産権の出願･取得状況

〔出願〕 計(0)件

産業財産権の名称	発明者	権利者	産業財産権の種類、番号	出願年月日	国内･外国の別

〔取得〕 計(0)件

産業財産権の名称	発明者	権利者	産業財産権の種類、番号	取得年月日 出願年月日	国内･外国の別

15.備考