様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２７　年　　５　月　２２　日現在

機関番号：　１２６０８
研究種目：　基盤研究(C)
研究期間：　2011～2014
課題番号：　２３５００１２１

研究課題名（和文）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

研究課題名（英文）High Precision and Fast Structured Document Retrieval in the Many-core Processor Era

研究代表者

宮崎　純（Miyazaki, Jun）

東京工業大学・情報理工学（系）研究科・教授

研究者番号：４０２９３３９４

交付決定額（研究期間全体）：（直接経費）　　4,000,000円

研究成果の概要（和文）：本研究では、構造化文書、特にWikipediaに代表されるXML文書に関して、それらが多くの一般ユーザにより自由に書き換えられるという特徴から、メニーコアプロセッサの利用を考慮した高速なXML文書の更新、それに伴う高い検索精度の維持、ならびに高速検索について研究を行なった。
文書の更新を効率化するために、新たな索引スキーマやノイズとなる語を排除するための二つのフィルタを提案した。さらに、文書統計量をGPGPUにより高速に計算する手法を提案した。これらにより、検索精度を保ちつつ文書更新コストを25%削減し、GPGPUにより文書統計量の計算をCPU実装よりも10倍以上高速化可能であることを示した。

研究成果の概要（英文）：In this project, we have studied on an efficient calculation of document statistics taking account of use of manycore processors and a method for fast updates of structured documents, in particular, XML documents such as Wikipedia, in response to their frequent modifications by many users with keeping its effective retrieval as well as fast query processing, so that these dynamically updated documents can always be retrieved precisely and efficiently.
In order to improve the efficiency of the updates of documents, we have proposed new term indexing schema and two filters to avoid inserting noisy terms into the indices. In addition, we have also proposed a method to efficiently calculate document statistics by using a manycore GPGPU. The experimental results showed that the cost of document updates can reduce up to 25% due to the new indices and filters without deteriorating its precision, and the GPGPU can lead to more than 10x faster calculation of document statistics than a CPU.

研究分野：　データ工学

キーワード：　情報検索　構造化文書　XML　文書統計量　メニーコアプロセッサ

１．研究開始当初の背景

近年、ハードウェアの微細化が進み、動作周波数を上げるよりも、一つのチップに多数の CPU コアを搭載するプロセッサ、すなわちマルチコア・メニーコアプロセッサにより計算処理の高速化を目指す方向へ進んでいる。コア数は増加の一途をたどり、GPU のように機能を簡略化したコアを 2000 個以上搭載するプロセッサも現れており、汎用プロセッサよりも格段に高速な計算が可能となってきた。

一方、Web、blog、センサ情報等、タグ付けされた様々な構造化文書、その中でも特に XML 文書が世界中で大量に作成されており、これらはインターネット上に多数遍在している。XML 文書が最も利用されているのはバイオデータベース等の科学データベースや Wikipedia をはじめ、最近ではオフィス文書や PDF 形式も格納形式として XML 形式を利用しており、今後ますます XML 文書に対する高精度かつ高性能な検索が不可欠となる。XML 文書検索の利点は、内包するタグ情報を利用して検索キーワードに適合する文書中の一部分の検索、すなわち部分文書検索が容易である点にある。その反面、一つの文書は複数の部分文書からなるため、検索対象の数が文書総数の数十から数百倍に増加することから様々な計算コストが大きくなるという問題がある。

XML 文書検索の研究は、これまで検索の高精度化、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、にのみ重点がおかれ、従来のマクロな文書検索と比べて、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、処理時間についてはあまり考慮されてこなかった。さらに、現実の世界に目を向ければ、Wikipedia 等のように文書が頻繁に更新される場合も多い。各文書統計量を効率良く再計算を行わなければ、検索精度が低下するが、これまで、計算量の問題から更新の問題についてはほとんど検討されてこなかった。

２．研究の目的

先述した背景を鑑み、本研究ではメニーコアプロセッサを利用し、高精度検索を維持するための文書統計量の高速計算を行ない、高精度かつ高速な XML 文書検索を実現する。また、本研究を含めて、これまで研究してきた高精度な XML 文書検索手法を、様々な応用するための試みも行なう。具体的には、

(1) 文書の動的な更新に対応するための、高精度検索を維持しつつ、高速な文書更新と問合せを可能とする文書統計量の計算方法、維持管理手法の開発
(2) マルチコア・メニーコアプロセッサ向けの効率の良い汎用データ構造の開発
(3) メニーコア GPU を利用した、高精度検索を維持するための高速文書統計量の計算方法の開発
(4) 高精度かつ高性能な部分文書検索技術の新しい応用分野への展開

の各研究項目を実現することを目的とする。

３．研究の方法

文書の動的な更新に対応した高精度かつ高速な XML 部分文書検索技術に関して、各索引語を効率良く検索、維持するための新しい索引スキーマの設計方法を明らかにする。また、文書の更新操作を効率よく行なうために、検索の対象となり得ないような索引語の除去するための方法や、文書統計量の推定により、計算量を減らすための手法について研究を行なう。これと並行して、メニーコアプロセッサで計算処理を実現するための効率の良い汎用データ構造とその性質を明らかにする。

次に、XML で記述された約 67 万文書の英語版 Wikipedia からなる INEX テストコレクションを利用し、提案手法の検索精度や処理性能の評価を行なうとともに、問題点を明らかにし、世界トップレベルの XML 部分文書検索の精度を維持しつつ、ユーザの満足しうる応答時間を目指して適宜提案方法の改善を行なっていく。

以上の基礎的な研究成果を利用し、メニーコアプロセッサを用いて高精度検索を維持するために、文書統計量を高速計算する手法について研究を行ない、その評価を行なう。これと並行して、高精度な部分文書検索技術を新しい事例に適用し、本研究の成果の展開を試みる。

４．研究成果

オンライン文書が動的に多数のユーザから更新されることを考慮し、高速な更新処理、高速な問合せを可能としつつ検索精度を保つために、新しい索引スキーマと更新コストを低減するための二つのフィルタを提案した。67 万個の XML 文書からなる INEX Wikipedia テストコレクションを利用して評価実験を行った結果、提案した索引スキーマを利用することで新たに追加された文書へも適切に重み付け可能であり、また、提案したフィルタを利用することで検索精度を維持しつつ索引構築時間を約 4 割、索引サイズを約 1 割削減できることを示した。

さらに、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、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。この手法を組み入れて同一のテストコレクションを利用して評価したところ、検索精度を維持しつつ、XML 文書の更新を考慮したナイーブな実装と比較して、更新処理が 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秒程度に短縮化した。これに検索結果を文書構造を利用して再構成するアルゴリズムを適用することで、検索精度を従来よりも4%向上することを明らかにした。

次に、高い検索精度維持のために、リアルタイムな文書統計量計算を目指して、メニーコアGPUを利用した文書統計量計算の高速化を試みた。クラウドコンピューティングでしばしば使用されるMapReduceフレームワークのGPU上での実装の一つであるMarsを利用して、Okapi BM25に基づく文書統計量計算アルゴリズムを実装した。その際に、MapReduceフレームワークのシャッフルフェーズをソートと集約演算とを分離する工夫も施し、全計算を1フェーズで実行可能とした。実験結果から、250MBの文書サイズ時に、2880コアのGPU実装の方が、4コアのCPUよりも11.6倍高速であることを明らかにした。GPUが一般に非数値演算にあまり適合しないにも関わらず、一桁以上の高速化を実現できることを示した。

以上の技術を発展させるために、部分文書検索技術のWeb文書への応用を試みた。Webの文書内容の論理構造と文書の物理構造の一致のための再構造化、ならびに検索結果として不要な箇所を特定するためのフィルタを提案した。評価実験から、再構造化によって、より適切な粒度の部分文書を抽出可能となったため、定義やQA検索の精度の向上が達成できた。またフィルタによって文書中のメインコンテンツ以外の箇所の除外を行なったことにより、一般的なWeb文書検索と比較して再現率が低下するものの、より高精度な検索を実現できることが判明した。

さらに、XML部分文書検索を高精度化するための技術として、複数の適合度指標、例えばテキスト情報の適合度や文書構造の適合度などの統合が必要であるが、適合度指標間には通常、複雑な依存関係があり、その依存関係に応じた統合方法が要求される。本研究では統計学での接合関数を利用するアプローチを試み、従来手法と比較して若干高精度であることを明らかにした。しかし、この関数の推定には時間を要するため、GPUに代表されるメニーコアプロセッサの利用が要求される。すなわち、今後のさらなる高精度な構造化文書の検索には、必然的にメニーコアプロセッサを利用した手法が強く望まれることを示唆した。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計５件）

(1) 欅惇志，宮崎純，波多野賢治，山本豪志朗，武富貴史，加藤博一．文書の更新を考慮した高精度XML部分文書検索手法の提案，情報処理学会論文誌データベース，査読有，Vol. 6, No. 4, pp.1-16, Sep. 2013.

(2) Atsushi Keyaki, Jun Miyazaki, Kenji Hatano, Goshiro Yamamoto, Takafumi Taketomi, Hirokazu Kato. Fast Incremental Indexing with Effective and Efficient Searching in XML Element Retrieval, International Journal of Web Information Systems, Emerald, 査読有, Vol. 9, No. 2, pp.142-164, Jun. 2013.

(3) Atsushi Keyaki, Kenji Hatano, Jun Miyazaki. Result Reconstruction Approach for More Effective XML Element Search, International Journal of Web Information Systems, 査読有, Vol. 7, No. 4, pp.360-380, 2011.

(4) 松原裕貴，宮崎純，藤澤誠，天野敏之，加藤博一．CC-PAID: CPUキャッシュを有効利用した並列時系列パターンマイニングアルゴリズム，情報処理学会論文誌データベース，査読有，Vol. 4, No. 2, pp.88-100, 2011.

(5) Atsushi Keyaki, Kenji Hatano, Jun Miyazaki. Relaxed Global Term Weights for XML Element Search, Comparative Evaluation of Focused Retrieval, LNCS, Springer, 査読有, Vol. 6932, pp. 71-81, 2011.

〔学会発表〕（計16件）

(1) 森谷祐介，欅惇志，宮崎純．GPUを用いたMapReduceによる高精度検索のための高速な重み計算，第7回データ工学と情報マネジメントに関するフォーラム（第13回日本データベース学会年次大会），福島県郡山市，Mar. 2015.

(2) 小松田卓也，清水伸幸，田島玲，欅惇志，宮崎純．Copulaを用いたスコア統合手法とその有効性の検証，第7回データ工学と情報マネジメントに関するフォーラム（第13回日本データベース学会年次大会），福島県郡山市，Mar. 2015.

(3) Atsushi Keyaki, Jun Miyazaki, Kenji Hatano. An Expansion Method of XML Element Retrieval Techniques into Web Documents, IIAI 3rd International Conference on Advanced Applied Informatics 2014, Proc. of IIAI 3rd International Conference on Advanced Applied Informatics 2014, 査読有, pp. 853-858, Kokura, Fukuoka, Sep. 2014.

(4) 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部分文書検索技術のWeb文書への適用，第6回データ工学と情報マネジメントに関するフォーラム(DEIM2014)，兵庫県淡路島，Mar. 2014.

(5) Atsushi Keyaki, Jun Miyazaki, Kenji Hatano, Goshiro Yamamoto, Takafumi

Taketomi, Hirokazu Kato. XML Element Retrieval@1CLICK-2, NTCIR-10, Proc. of NTCIR-10, pp.237-242, Tokyo, Jun. 2013.
(6) 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, 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(DEIM2013), 福島県郡山市, Mar. 2013.
(7) Atsushi Keyaki, Jun Miyazaki, Kenji Hatano, Goshiro Yamamoto, Takafumi Taketomi, Hirokazu Kato. Fast and Incremental Indexing in Effective and Efficient XML Element Retrieval Systems, 14th International Conference on Information Integration and Web-based Applications & Services, Proc. of the 14th International Conference on Information Integration and Web-based Applications & Services, 査読有, pp. 157-166, Bali, Indonesia, Dec. 2012.
(8) 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, 第5回 Web とデータベースに関するフォーラム (WebDB Forum 2012), 第5回 Web とデータベースに関するフォーラム (WebDB Forum 2012) 論文集, 秋葉原ダイビル, Nov. 2012.
(9) Yuki Matsubara, Jun Miyazaki, Goshiro Yamamoto, Yuki Uranishi, Sei Ikeda, Hirokazu Kato. CCDR-PAID: More Efficient Cache-Conscious PAID Algorithm by Data Reconstruction, 27th ACM Symposium On Applied Computing, Proc. of the 27th ACM Symposium On Applied Computing, 査読有, pp. 193-198, Riva del Garda, Italy, Mar. 2012.
(10) 宮崎純, 鬼塚真. データクラウドを支える技術と研究動向, 第 10 回 先端的データベースと Web 技術動向講演会 (ACM SIGMOD 日本支部 第 47 回大会), 東京工業大学, Jun. 2011.
(11) 欅惇志, 宮崎純, 波多野賢治, 山本豪志朗, 加藤博一. XML 索引の更新コスト削減のための部分文書の統計量に基づくフィルタの評価とその最適化, 第4回データ工学と情報マネジメントに関するフォーラム (DEIM2012), 兵庫県神戸市, Mar. 2012.
(12) 福澤優, 宮崎純, 山本豪志朗, 加藤博一. カラムストアとローストアを利用した OLAP 問合せ処理における消費電力と処理速度の関係について, 第4回データ工学と情報マネジメントに関するフォーラム (DEIM2012), 兵庫県神戸市, Mar. 2012.
(13) 吉武亮, 宮崎純, 山本豪志朗, 加藤博一. スカイラインの近傍探索を可能とする拡張スカイライン演算の実装と評価, 第4回データ工学と情報マネジメントに関するフォーラム (DEIM2012), 兵庫県神戸市, Mar. 2012.
(14) 松原裕貴, 宮崎純, 山本豪志朗, 浦西友樹, 池田聖, 加藤博一. データアクセスの改良による時系列パターンマイニングアルゴリズムの高速化, 情報処理学会データベースシステム研究会, 情報処理学会研究報告データベースシステム(DBS), pp. 1-8, 工学院大学, Nov. 2011.
(15) 欅惇志, 宮崎純, 波多野賢治, 山本豪志朗, 加藤博一. XML 情報検索のための動的な索引管理手法の一提案, 情報処理学会データベースシステム研究会, 情報処理学会研究報告データベースシステム (DBS), pp. 1-8, 工学院大学, Nov. 2011.
(16) 吉武亮, 宮崎純, 藤澤誠, 天野敏之, 加藤博一. 情報推薦のための Skyline 演算の拡張, 平成 23 年度情報処理学会関西支部大会, 平成 23 年度情報処理学会関西支部大会講演論文集, 大阪大学中之島センター, Sep. 2011.

〔図書〕(計0件)

〔産業財産権〕
○出願状況 (計0件)

○取得状況 (計0件)

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
宮崎　純 (MIYAZAKI, Jun)
東京工業大学・大学院情報理工学研究科・教授
研究者番号：40293394