

平成 21 年 5 月 31 日現在

研究種目：若手研究 (B)
 研究期間：2007～2008
 課題番号：19700083
 研究課題名 (和文)
 e サイエンスのためのオーバレイネットワークに基づく XML データ流通機構の研究
 研究課題名 (英文)
 Study on XML data distribution system based on overlay networks for e-Science
 研究代表者
 天笠 俊之 (AMAGASA TOSHIYUKI)
 筑波大学・大学院システム情報工学研究科・講師
 研究者番号：70314531

研究成果の概要：新しい科学の形として、ICT 技術を通じて新たな科学的発展を目指す e サイエンスが注目されている。そのインフラとしては、よくインターネットなどの公共ネットワークの上に独自のネットワークを構築する技術である、オーバレイネットワークが用いられる。他方、XML (Extensible Markup Language) は、データを記述するための標準フォーマットとして広く利用されている。このことから、本研究では e サイエンスの基盤として、オーバレイネットワークにおける XML データ流通機構の研究を行う。具体的には、分散ハッシュ表における XML データの格納・検索機構、XML データのトレーサビリティ、e サイエンスのための XML 流通システムプロトタイプ作成に焦点を絞って研究を行なった。

交付額

(金額単位：円)

	直接経費	間接経費	合計
2007年度	1,600,000	0	1,600,000
2008年度	1,600,000	480,000	2,080,000
年度			
年度			
年度			
総計	3,200,000	480,000	3,680,000

研究分野：データベース

科研費の分科・細目：情報学・メディア情報学・データベース

キーワード：XML, トレーサビリティ, オーバレイネットワーク, e サイエンス

1. 研究開始当初の背景

本研究課題を準備していた 2006 年当時、新しい科学の形として、ICT 技術を通じて情報資源、計算資源、人的資源の交流を促進し、新たな科学的発展を目指す「e サイエンス」が科学の諸分野、例えば物理、天文、科学、バイオなどで注目され始めていた。関連して、新たなネットワークインフラとして、あるネットワークの上に独自のネットワークを構築できる「オーバレイネットワーク」技術が注

目されていた。これによりインターネットなどセキュアではないネットワークを利用して、独自のプライベートなネットワークを構築することができ、あるグループに閉じたコンテンツの流通などに利用することが可能となる。他方、XML (Extensible Markup Language) は、データを記述するための標準フォーマットとして広く利用されるようになっていた。e サイエンスの各分野においても、実験データや観測データの記述、あるいはそれらのメタデータを記述するために、XML が多く用いら

れていた。このため、XML 形式のデータが多く蓄積され、その利活用が望まれていた。

オーバーレイネットワーク上での XML データの利用については、格納と検索に関する研究はいくつか見ることができた。しかしながら、XML データの出自や系統の管理に関する研究は見ることができなかった。

2. 研究の目的

e サイエンスでは一般のファイル交換アプリケーションと異なり、流通するデータから導かれる生成物の系統管理が極めて重要である。例えば、万が一不正なデータが紛れ込んでいたり、プログラム等の不具合によって、意図せず不正なデータが混入してしまった場合でも、そこから影響を受けるデータの範囲を明確に切り分けることが可能でなければならない。物流に関しては、IC タグなどを利用したトレーサビリティに関する研究が盛んであるが、XML データのトレーサビリティに関してはあまり例がなく、e サイエンスを支える基盤技術として研究を進める必要がある。本研究では e サイエンスの基盤として、オーバーレイネットワークにおける XML データ流通機構の研究を行う。具体的には、分散ハッシュ表における XML データの格納・検索機構、XML データのトレーサビリティ、e サイエンスのための XML 流通システムプロトタイプ作成に焦点を絞って研究を行なった。

3. 研究の方法

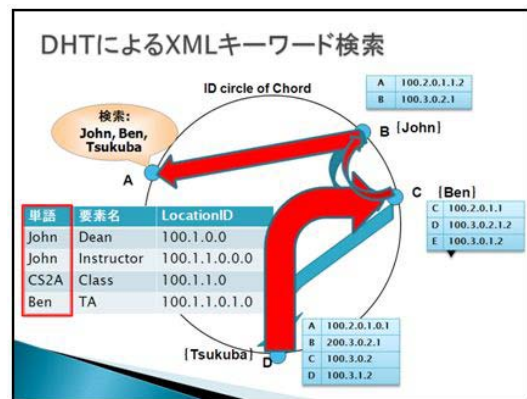
(1) 分散ハッシュ表に基づくオーバーレイネットワークを用いて、XML で記述された各種データを格納・検索する手法を開発する。基本的には、これまでに提案してきた分散ハッシュ表を利用した XML データの格納、検索手法をベースに、これを拡張することによって進める。

(2) XML データのトレーサビリティについては、関係データベースにおけるトレーサビリティ機構に関する研究を参考に、XML データの各項目（要素または属性）に系統情報を管理するためのアノテーションを埋め込み、問合せ処理やデータ変換の際にもそれらを適切な方法で処理することによって、データの追跡を可能にする。また、この検索を高速化するための索引手法についても検討する。

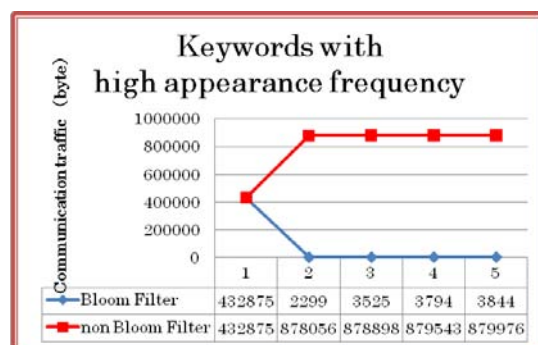
(3) 上記 2 項目で作成した手法を組み合わせ、XML データ流通機構のプロトタイプシステムを実装する。

4. 研究成果

(1) 分散ハッシュ表に基づくオーバーレイネットワークを用いて、XML で記述された各種データを格納・検索する手法を検討した。特に、これまでは P2P ネットワークではあまり扱われてこなかった XML データの内容に基づく検索（キーワード検索）を可能にすることを目標に、1) 分散ハッシュ表の上に XML データを対象とした転置索引の構築、2) キーワード検索の際にネットワークトラフィックを低減するための、bloom フィルタを用いたデータ転送方式と検索アルゴリズムの開発を行った。基本的には、各ピアに転置インデックスを用意しておき、問合せキーワードと中間結果をピア間でやり取りすることにより、採取的な結果を絞り込む。また、その際、bloom フィルタにより通信量の削減を行う。

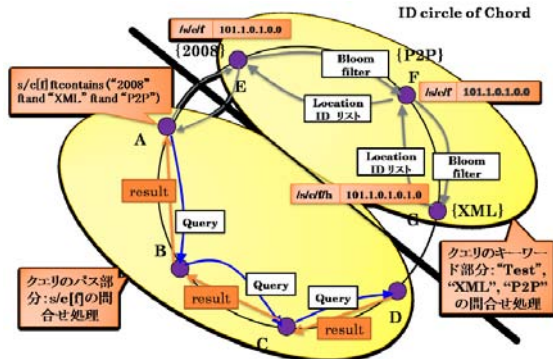


実験により提案手法の有効性を評価した。その結果、分散ハッシュ表を利用して、XML データに対するキーワード検索を効率的に処理できることが示された。検索キーワードが増えても、bloom フィルタの効果により通信量が削減できている。



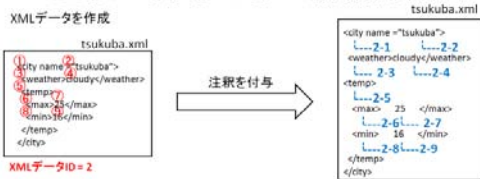
さらに、先行研究課題で提案した、分散ハッシュ表に基づく XPath 問合せの処理方式と、本研究課題で提案したキーワード検索方式を組み合わせ、キーワードを条件に含むような XPath 問合せを処理するアルゴリズムを提案

した。基本的なアイデアは次のとおりである、まず問合せを XPath 部分とキーワード部分に分ける。次にそれぞれを同時に分散ハッシュ表を用いて処理する。最後に得られた部分的な解を構造結合することによって、最終的な解を得る。



(2) XML データのためのあらたなトレーサビリティ手法を提案した。このため、まず XML データモデルを拡張し、各 XML ノードに注釈（アノテーション）を付与したモデルを提案した。問合せ処理の際には、XML ノードの注釈を伝搬させることによって、XML データの系統管理が可能となる。

注釈の付与とその格納方法



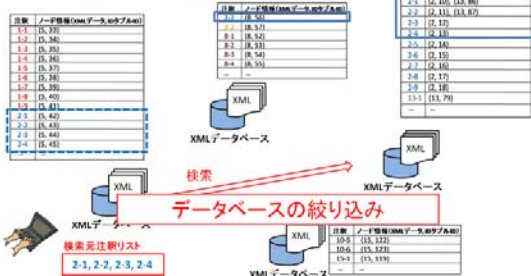
・格納データに注釈を格納する属性を追加

tid	did	type	name	value	nodenum	path	anno
...
10	2	element	city	null	1	/city	2-1
11	2	atts	name	tsukuba	1	/city/@name	2-2
12	2	element	weather	sunny	1.1	/city/weather	2-3
...

また、複数のデータベースに分散して格納された注釈付きの XML データに対して、注釈をキーとした検索を可能にするための索引手法を提案したこれにより広域に流通した XML データを、その系統を元に高速に検索することができる。

検索処理

- 検索元の注釈 (2-1, 2-2, 2-3, 2-4) を全て蓄えているデータを検索したい
- 問題点
 - 注釈インデックスを実際に検索してみないと、検索結果があるかどうか分からない
 - 対象データベースが多数存在する場合、非効率



(3) 上記手法をプロトタイプシステムとして実装した。

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 2 件)

- 李曉晨, 天笠俊之, 北川博之, 「構造型 P2P ネットワークにおけるキーワードを含む XPath による XML 文書検索」, 日本データベース学会論文誌, Vol. 7, No. 3, pp. 67-72, 2008 年 12 月. (査読あり)
- 吳俊輝, 天笠俊之, 北川博之, “構造型 P2P ネットワークにおける負荷分散を考慮した XML データ処理”, 日本データベース学会 Letters, Vol. 6, No. 1, pp. 93-96 2007 年 6 月. (査読あり)

[学会発表] (計 27 件)

- 永元芳幸, 天笠俊之, 北川博之, 「注釈を利用した XML データのためのトレーサビリティ機構の提案」, 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 (DEIM 2009), セッション D1, 静岡県掛川市, 2009 年 3 月 8 日~11 日.
- 李曉晨, 天笠俊之, 北川博之, 「DHT における XPath に基づく XML キーワード検索の性能評価」, 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 (DEIM 2009), セッション B5, 静岡県掛川市, 2009 年 3 月 8 日~11 日.
- Imam Machdi, Toshiyuki Amagasa, and Hiroyuki Kitagawa, “XML Data Partitioning Strategies to Improve Parallelism in Parallel Holistic Twig Joins”, The 3rd International Conference on Ubiquitous Information Management and Communication (ICUIMC 2009), SKKU, Suwon, Korea, January 15-16, 2009.
- Imam Machdi, Toshiyuki Amagasa, and Hiroyuki Kitagawa, “GMX: An XML Data Partitioning Scheme for Holistic Twig Joins”, The 10th International Conference on Information Integration and Web-based Applications & Services (iiWAS2008), pp. 137-146, Linz, Austria, 24-26 November, 2008.
- Chantola Kit, Toshiyuki Amagasa, and Hiroyuki Kitagawa, “Algorithms for Efficient Structure-Based Grouping in XML-OLAP”, The 10th International

Conference on Information Integration and Web-based Applications & Services (iiWAS2008), pp. 170-177, Linz, Austria, 24-26 November, 2008.

- ⑥ Imam Machdi, Toshiyuki Amagasa, and Hiroyuki Kitagawa, "An Algorithm for Parallel Holistic Twig Joins on a PC Cluster", iDB フォーラム 2008, 福島県飯坂市, 2008年9月21日~23日.
- ⑦ 李曉晨, 天笠俊之, 北川博之, 「構造型 P2P ネットワークにおけるキーワードを含む XPath による XML 文書検索」, iDB フォーラム 2008, 福島県飯坂市, 2008年9月21日~23日.
- ⑧ Xiaochen Li, Toshiyuki Amagasa, and Hiroyuki Kitagawa, "Searching XML Documents by Keywords in Structured P2P Networks", 3rd International Workshop on XML Data Management Tools and Techniques (XANTEC'08) in conjunction with DEXA 2008, pp. 274-278, Turin, Italy, Sept. 1-5 2008.
- ⑨ Toshiyuki Amagasa, Kentarou Kido, and Hiroyuki Kitagawa, "Querying XML Data using PC Cluster System", 2nd International Workshop on XML Data Management Tools and Techniques (XANTEC'07) in conjunction with DEXA 2007, pp. 5-9, Regensburg, Germany, 3-7 September, 2007.
- ⑩ Toshiyuki Amagasa, Lianzi Wen, and Hiroyuki Kitagawa, "Proximity Search of XML Data using Ontology and XPath Edit Similarity", 18th Int'l Conference on Database and Expert Systems Applications (DEXA2007), pp. 298-307, Regensburg, Germany, 3-7 September, 2007.
- ⑪ Toshiyuki Amagasa, Chunhui Wu, and Hiroyuki Kitagawa, "Retrieving Arbitrary XML Fragments from Structured Peer-to-Peer Networks," Joint Conference of the 9th Asia-Pacific Web Conference and the 8th International Conference on Web-Age Information Management (APWeb/WAIM 2007), pp. 317-328, HuangShan (Yellow Mountains), China, 16-18 June, 2007.

[その他]

ホームページ等

<http://www.kde.cs.tsukuba.ac.jp/~amagasa/>

6. 研究組織

(1) 研究代表者

天笠 俊之 (AMAGASA TOSHIYUKI)

筑波大学・大学院システム情報工学研究科・講師

研究者番号 : 70314531