異種XMLデータに対するファセット検索システムの性能評価

駒水 孝裕† 天笠 俊之†‡ 北川 博之†‡

†筑波大学大学院システム情報工学研究科　‡筑波大学計算科学研究センター

1 はじめに

XML[1] は近年，データフォーマットの標準フォーマットとして多くのアプリケーションで利用されている．そのため，XMLデータに対する検索の要求が大きくなってきている．これらの問題点に対し，我々は[3] でXMLデータの検索にファセット検索 [2] を適用した．

本稿では先行研究 [3] にて提案したシステムの基本概念と概要を紹介し，システムの有効性を検証するためにその性能を評価する．提案システムではファセット検索に必要なクラス，クラス属性，オブジェクト，ファセットとキーの概念をデータベーススキーマとしたデータベースを用いる．提案システムはユーザインターフェース，検索モジュール，統合モジュールからなる．検索では利用者があらかじめユーザインターフェースに表示されたファセットとキーを選択し，選択されたファセットとキーに基づき検索モジュールが検索処理を行い，統合モジュールが表示する情報を統合・整形する．このような検索システムの有効性を検証するための実験を行い，その有効性を示す．本稿の構成は以下のようになる．まず，第2節で異種XMLデータに対するファセット検索の説明をし，第3節でこれを実現するシステムの概要を提示する．その後，第4節で実験とその評価について触れ，最後の第5節でまとめと今後の課題について述べる．

2 異種XMLデータに対するファセット検索

これまでにXMLデータに対するファセット検索がなされてこなかった理由の一つとして，検索対象オブジェクトやそれに対するファセットとキーが定義されていないことが挙げられる．そこで我々は [3] にて形式的な定義を与えた．本節ではこれらの定義を簡単に紹介する．これらの定義はDTDやXMLSChemaなどのスキーマ定義に基づいて作成されるスキーマ木と実際に検索されるXMLデータから得られる．

クラス
: スキーマ木における濃度が * や + のノードをクラスとして定義する．これらの濃度はそのノードが複数回出現することを示しており，レコードのような意味的なまとまりを表しているものと考えられるからである．

クラス属性
: クラスノード以下のノードで子ノードにテキストを持つものをそのクラスのクラス属性として定義する．これはテキストがクラスの付加的な情報を表していると考えることができるためである，

オブジェクト
: 実際のXMLデータ中でクラスとして定義されたノードと定義する．すなわち，クラスに対応する実データのことである，

ファセット
: オブジェクト中に出現する属性の和集合と定義する．これはクラス属性として定義された属性が必ずしも実データに出現するとは限らないことと，複数のオブジェクトで重複があった場合にその重複を取り除くためである．

キー
: 実データ中の属性に対するテキスト値として定義する．このテキスト値を基にオブジェクトの検索を行う．

これらの定義に基づいた検索方法について [3] にて議論した．以下でその説明をする．

検索する際の入力はファセットとキーのペアの集合である．利用者はこのペアを順次選択していくことで検索を繰り返し行い絞り込むことができる．検索処理ではペア毎に検索対象オブジェクト集合を絞込み，それらの積集合をることで現在選択されているファセットとキーのペアの集合に対する結果を得ることができる．その後のこの結果と結果におけるファセットとキーを表示する．また，このファセットとキーを表示する際にそれぞれが検索できるオブジェクトの数を算出す

Evaluating the faceted navigation system over heterogeneous XML data

Takahiro KOMAMIZU† Toshiyuki AMAGASA†‡ Hiroyuki KITAGAWA†‡

†Graduate School of Systems and Information Engineering, University of Tsukuba ‡Center for Conputational Sciences, University of Tsukuba

る．これらの処理の流れによりファセット検索における検索処理を実現することができる．

3 システムの概要

本節では2節で述べた手法を基にした異種XMLデータに対するファセット検索システムの概要を紹介する．本システムは以下のモジュールにより構成される．

- ユーザインターフェース
- 検索モジュール
- 統合モジュール

以下で各モジュールについて説明する．

まず，ユーザインターフェースは利用者とのインタラクションを行う．利用者に検索可能なオブジェクトとそれに関連するファセットとキーを表示し，利用者の検索行動を促す．利用者が選択したファセットとキーをペアとして現在選択されているすべてのペアを検索モジュールに渡す．

次に検索モジュールはユーザインターフェースから渡されたファセットとキーのペアの集合を基にデータベースへの検索を行う．検索の結果として次の2つが得られる．1つは検索されるオブジェクト集合で，もう一つはそれらに関するファセットとキーである．これらはこの時点では別々のデータであり，整形されていないため統合モジュールに送る．

最後に統合モジュールでは検索モジュールにより送られてきた検索結果に関するデータをまとめる．ファセットとキーは利用者の見やすいように整形し，検索結果オブジェクト集合はXMLデータとして見える形に成形する．なお，この検索結果オブジェクト集合はXMLであるためXSLTなどで形式を定義することで利用者の好みの形に変換することが可能である．

4 実験と評価

本節では提案システムの有効性を実験により検証する．実験環境は次のようになる．CPUはIntel Core2 Duo 2.33GHzでメモリは2GB，オペレーティングシステムはFedora 11 (Linux)であり，プログラム言語はJava言語である．実験方法は実際のシステムを使用し，各検索行動における検索モジュールと統合モジュールの実行時間を合計したものを算出する．ユーザインターフェースにて応答時間を測るとネットワークの遅延等の他の要因が検証に関与してしまうことが想定されるため．全モジュールの実行時間を計測した．結果は図1の様になった．

横軸検索されたオブジェクト数で縦軸が実行時間である．青い線は全体の実行時間の平均を示している．

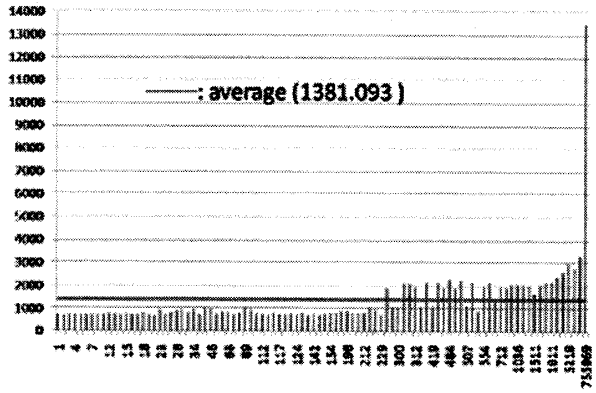

図1: 実行結果

全オブジェクトを検索する処理（オブジェクト数755869）においておよそ14秒ほどかかっているが，これは何も選択していない状態であるためキャッシュすることでこの時間を0.1秒以下に抑えることができる．また，平均時間を見ると約1.4秒ほどなので検索システムとしては十分な速度であると考える．検索可能オブジェクト数が増えるにつれ実行時間が増えるが，その中で実行時間が前後する理由は選択したファセットとキーのペアの数により処理時間が前後するためである．

5 まとめと今後の課題

本稿では[3]で議論した手法がプロトタイプの実験により十分なパフォーマンスが得られることを示した．今後の課題としては検索手法の拡張とファセットの効率的な配置順序の考慮が挙げられる．

謝辞 本研究の一部は科学研究費補助金特定領域研究（# 21013004）と科学研究費補助金若手研究（B）（# 21700093）による．ここに記して謝意を示す．

参考文献

[1] XML1.0. http://www.w3.org/TR/REC-xml/.

[2] Eyal Oren, Renaud Delbru, and Stefan Decker. Extending faceted navigation for RDF data. In Isabel F. Cruz, Stefan Decker, Dean Allemang, Chris Preist, Daniel Schwabe, Peter Mika, Michael Uschold, and Lora Aroyo, editors, *International Semantic Web Conference*, Vol. 4273 of *Lecture Notes in Computer Science*, pp. 559–572. Springer, 2006.

[3] 駒水孝裕, 天笠俊之, 北川博之．異種XMLデータに対するファセット検索手法の提案．情報処理学会研究報告「デジタルドキュメント (DD)」, Vol. 2009-DD-73, No. 7, pp. 1–8, 2009.