XML 統合を支援する軽量 XQuery プロセッサの開発
―XBird: 省メモリでギガバイト XML 問合せ処理を実現―

1. 背景

現在，XML コンテンツを利用する場合，開発者は DOM や SAX といった低レベルの XML 操作 API を利用するのが主流である。低レベルの API であるがため，複雑な XML 操作を行う上での開発者負担は大きい。

W3C で標準化が推進された XQuery は，XML の為の照会言語であり，集約処理や結合処理，および柔軟な選択処理をサポートする。そのため，XQuery は XML を扱うドメイン特化言語(DSL)として有望である。2007 年 1 月 23 日付けで W3C 勧告となったことで，XPath 1.0 や XSLT 1.0 の先例と同様にユーザレベルでの認知度も高まっていくと考えられる。

本プロジェクトでは，企業におけるユーザの多い Java プラットフォームを対象とし，アプリケーションに組み込み可能な XQuery プロセッサを開発する。XQuery が利用される場面としては，XML データ交換や Web サービスのメッセージ処理といったアプリケーションサーバ上での利用が想定される。そのため，アプリケーションに組み込んで利用できるかどうかは重要な事項である。

これまでに公開されている Java プラットフォーム向けの主要な XQuery プロセッサとしては，Saxon と Qizx がある。特に前者のフリー版である Saxon-B のシェアがもっとも高く，Saxon はいくつかの商用 XQuery エンジンでも内部的に利用されている。しかし，Saxon はメモリ消費量が大きく，大きな XML 文書を扱う場合の性能に劣るという問題がある。また，扱えるデータ量の上限も 500MB 程度までと限界があり，また，充分な Join 最適化が実装されていないため，計算量の大きい問合せにおいて性能は望めない。本プロジェクトではこれらの不満点を解消する XQuery プロセッサを開発する。

2. 目的

本プロジェクトの目的は，アプリケーションに組み込み可能で，かつ少ないメモリ消費量で動作する XQuery プロセッサを開発することである。競合実装に対して優位に立ち，デファクトの XQuery プロセッサとして世界中で広く利用されることを目指す。大学発のデータベースソフトウェアという面で，XML データベースにおける PostgreSQL のような地位を獲得することが最終的な目標である。

ライセンスにはオープンソースライセンスを適用し，研究者及び開発者が柔軟に利用できる XQuery プロセッサを提供する。開発したソフトウェアは CPL(Common Public License) 1.0 で公開する。一部の機能については，論文として発表されるまでソースコードを難読化するが，機能面・性能面での制限事項は設けない。

3. 開発の内容

プロジェクトでの開発項目は次の通りである。

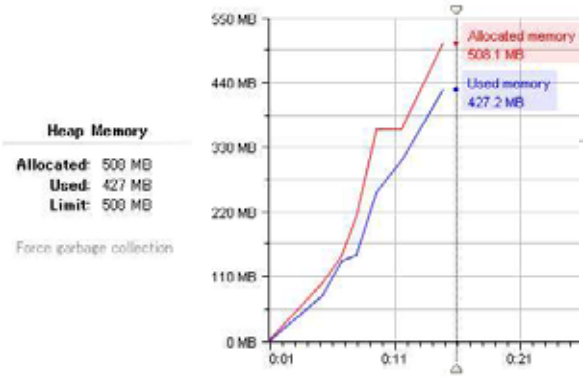
- XQuery プロセッサ機能
  W3C の XQuery 1.0 仕様 (http://www.w3.org/TR/xquery/) を満たす XML 問合せ処理機。スキーマインポート機能などの Optimal とされる機能については開発対象外である。
- HTML 文書の解析機能
  厳格でなく HTML 文書を極力 XML として解析し，XQuery プロセッサで処理する機能である。
- 動的 Web ページ生成機能
  XQuery を用いた動的 Web ページ生成技術である。
- ビューとクエリの統合機能
  ユーザ・クエリとビュー・クエリを統合し，下位のデータソースに効率的にアクセスすることを可能とする機能，XML コンテンツ統合を実現する機能である。
- XQTS テストスイート対応
  W3C が提供する XQuery のテストスイートのテスト通過率を高める目的の開発項目である。テストスイートを自動実行するテストコードの開発と，テストスイートの実施により見つかった XQuery プロセッサのバグ対処を行う。
- プロジェクトサイトとドキュメントの作成
  プロジェクトの情報発信を行なうための Web サイト、及びプロジェクトサイトのコンテンツ

4. 従来の技術 (または機能) との相違

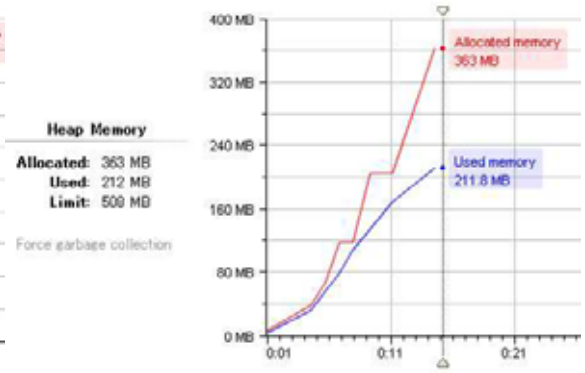
ここでは，開発したソフトウェアの特徴について，メモリ消費量と Join 最適化機能について説明する。

- メモリ消費量

XBird では，DTM(Document Table Model)というコンパクトなデータ構造を用いて XML の木構造を表現することで，少ないメモリ消費量で XML 文書を扱うことに成功した。



(A) SAXON-SA のヒープ消費量



(B) XBird のヒープ消費量

上記の図は，競合実装である SAXON-SA と XBird について，XMark テストスイートを用いて生成した 113MB の XML 文書を，XQuery プロセッサがそれぞれ読み込んだ際に利用されるヒープ消費量の推移を示すグラフである。(A)の SAXON-SA が元の文書の約 4 倍となる 427.2MB のヒープメモリを消費するのに対して，(B)の XBird では元の文書の約 2 倍となる 211.8MB にヒープの消費量が抑えられている。

- Join 最適化機能

XBird では，Join 処理を最適化することによって競合実装に対して，Join を含む問合せにおいて 24 倍から 91 倍程度の性能が得られた。

	XBird	Saxon-SA	Saxon-B
Q8	10.266	11.344	592.594
Q9	10.265	914.672	739.313
Q10	16.672	13.422	88.906
Q11	20.422	766.344	1480.188
Q12	15.594	351.812	523.235

XML ベンチマークツール XMark による性能比較 (対象データ: 113MB, 計測単位: 秒)

5. 期待される効果

Java 環境で動作する XQuery プロセッサとして，2GB 以上のデータを円滑に扱えるものはこれまで存在しなかった。XBird は，10GB 以上の XML データを扱うことができる。また，省メモリ環境においても動作するという特性は，他の XQuery プロセッサにはなかった特徴である。競合実装と比較して，少ないメモリでより大きな XML 文書を扱うことが可能なため，ハードウェアへの投資を抑えることができる。また，XBird はオープンソースで開発

されるため，XBird を利用することで開発者は特定の企業の製品にロックオンされる（依存する）事態を回避することができる。

今回開発した XBird は，最新の XML データベース研究の成果を遅延なしに実システムに適用していく試みの第一歩でもある。UC バークレーで開発された Postgres の前進とそのプロトタイプ実装は，関係データベース技術の発展に貢献し，Michael Stonebraker 教授らの研究グループの知名度の向上にも大きく役立った。将来的に XBird が広く利用されるようになれば，同様の効果が得られることが期待できる。

6．普及（または活用）の見通し

XQuery は，2007 年 1 月 23 日付けで W3C 勧告となったが，そのプレスリリースで仕様準拠度をまとめた 14 の実装のうちの一つという形で XBird が参照され，XQuery 開発関係者に一定の認知を得たと考えている。

参考). XQuery 1.0、XSLT 2.0、XPath 2.0 の各仕様群の公開について（W3C 勧告）
http://www.w3.org/2007/01/qt-pressrelease

今後，XQTS テストスイートの通過率を上げ，充分な通過率が得られ次第，オープンソースソフトウェアとして XBird を公開する予定である。
競合実装の Saxon は通算 33 万ダウンロードを達成しており，eXist は通算 12 万ダウンロードとなっている。認知度が上がるまでには時間がかかると予想されるが，最大で 10 万ダウンロード以上を期待することができる。まずは 1 万ダウンロードを目指し，普及に向けた課題としてテスト通過率の向上とドキュメントの整備を進めていく。

XBird で開発した技術の産業展開していく試みの一環で，未踏の発表の場を通じて知り合った国内の XML データベースベンダと技術交流を行っていく。切磋琢磨することは，国内の XML データベースベンダの国際的な競争力を高めることに繋がる。

7．開発者名（所属）

油井　誠（奈良先端科学技術大学院大学 情報科学研究科）
（参考） XBird プロジェクトサイト
http://db-www.naist.jp/~makoto-y/proj/xbird/