XMLディスクとその応用事例

高度メディア蓄積・管理手法研究グループ（奈良先端科学技術大学院大学　筑波大学　同志社大学）

研究の背景

近年のXMLデータの急速な普及

バイオデータベース、Webサービス、Wikiペディア、…

ホスト計算機でのXMLデータ処理および管理コスト

ストレージ側にXMLデータ処理をオフロード

研究目的

XMLディスクの開発

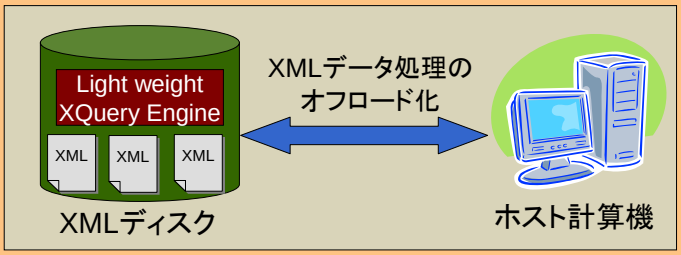
XQueryエンジンの軽量化と自律ディスクへの組込み

XMLディスクの応用例

バイオ情報データの検索

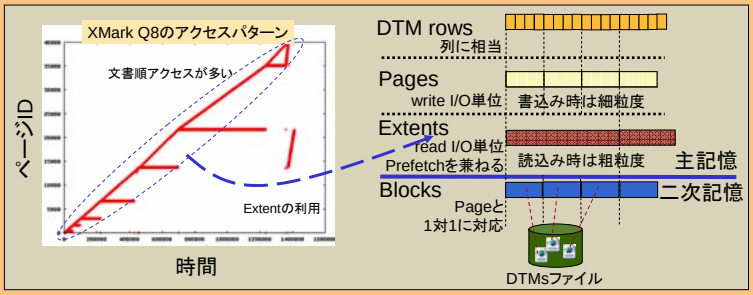
XMLディスク

コンセプト



I/O粒度の最適化

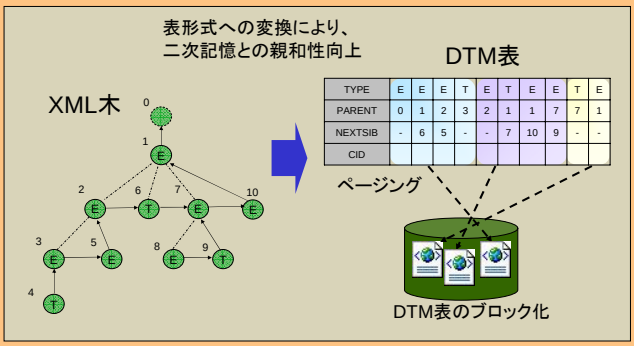
XQueryのページアクセスパターンからI/O粒度の決定



XQueryエンジンの省メモリ化

従来のXQueryエンジンは大量のメモリを消費

Document Table Model (DTM)を二次記憶への格納に拡張



性能評価

XMarkベンチマークによる評価

	XMark113MB (ScaleFactor: 1)				XMark340MB (ScaleFactor:3)			
	DTM	可変長DTMs	固定長DTMs	Saxon-SA	DTM	可変長DTMs	固定長DTMs	Saxon-SA
Q1	9.938	3.531	3.609	9.875	28.64	10.704	9.93	28.575
Q2	9.922	4.296	3.891	10.203	28.593	11.75	9.265	27.968
Q3	9.89	5.219	4.687	10.328	30.281	14.359	12.47	27.641
Q4	9.812	5.453	5.266	10.375	29.641	15.282	12.797	27.219
Q5	9.203	3.719	3.953	10.25	29.172	10.14	8.953	25
Q6	10.968	10.156	8.578	9.906	32.625	29.375	22.703	27.328
Q7	12.969	15.562	14.62	11.047	39.203	61.906	45.453	27.734
Q8	10.672	5.657	5.515	11.344	30.5	14.875	13.203	27.875
Q9	10.953	8.937	9.25	914.672	31.31	25	22.422	DNF
Q10	16.813	23.157	27.422	13.422	49.704	78.453	73.796	36.828
Q11	19.328	16.907	16.938	766.344	89.687	67.297	84.172	DNF
Q12	15.703	12.984	12.656	351.812	67.563	60.156	57.656	DNF
Q13	9.781	3.437	3.563	10.719	28.781	9.84	8.391	30.484
Q14	11.719	12.563	11.484	10.781	34.47	37.922	30.218	32.234
Q15	9.765	3.25	3.156	9.828	28.969	8.578	7.719	28.719
Q16	9.812	3.39	3.391	10.5	29.15	9.406	8.11	25.797
Q17	10.109	4.156	3.984	10.469	29.891	11.63	9.578	27.016
Q18	10.25	4.969	4.75	10.672	28.922	13.422	11.93	27.36
Q19	11.532	9.594	9.797	11.5	33.657	26.922	24.844	34.625
Q20	10.312	5.344	5.156	10.313	30.219	14.765	13.156	27.938

heapサイズの影響
(ScaleFactor:10)

	1024m	256m
Q1	37.62	35.656
Q2	40.984	41.625
Q3	38.656	39.31
Q4	41.641	43.688
Q5	30.469	31.718
Q6	93.32	88.187
Q7	151.704	153.735
Q8	50.25	44.578
Q9	101.562	110.984
Q10	328.531	400.875
Q11	731.344	747.688
Q12	468.781	501.63
Q13	33.687	33.547
Q14	152.437	143.25
Q15	32.234	27.32
Q16	27	27.36
Q17	31.14	32.47
Q18	38.235	38.14
Q19	139.47	ERR
Q20	56.219	51.453

大多数のクエリにおいてSaxon-SAより高速
Q9,Q11,Q12ではより顕著

メモリが少なくても安定して動作
（Saxon-SAは双方ともメモリ不足エラー）

デモンストレーション

複数のバイオデータベースの情報統合

ウェット実験者のための検索サポート
アノテータのデータベース参照支援

バイオデータベースの問題点

同一塩基、タンパク質でもデータベースで異なるIDが使用
複数のデータベースとの情報統合が困難

デモ内容

BLASTで得られたアミノ酸配列の相同性から
複数のデータベースの柔軟な情報統合を行う