

XML 数据管理技术新进展

周军锋

XML 技术在当前的互连网络和 IT 环境中扮演越来越重要的角色，它事实上已经成为数据交换的标准、SOA 架构的基石。Gartner 预测，XML 文件的使用率在 2007 年达到 40%，在 2008 年将占据支配地位。IDC（国际数据公司）最近发布的一份报告显示，在 500 家受访企业的 IT 部门中，有 29% 的企业宣称正在大量使用 XML 存储库和数据库。XML 的广泛应用使得高效的 XML 数据管理成为一种迫切的需求。

据作者统计，每年数据库相关的国内外会议期刊上发表 300 篇左右与 XML 相关的文章。而已有的综述类文章针对的都是 XML 数据管理技术的某个方面，例如查询或者索引，不够系统性。另外，发表时间大都在 07 年之前，涉及的文章则在 06 年之前。08 年是 XML 标注推出的第十个年头，在这样一个历史点上对 XML 数据管理技术进行全面、系统的总结具有实际意义的事情。

本综述所要达到的目的可表述为四个方面：（1）全面性。指该综述涉及 XML 数据管理的大部分技术点，并非已有文献的简单罗列。（2）系统性。指该综述所述技术点之间构成一个系统，并非技术点的简单罗列。（3）前瞻性。指该综述力求全面、深入的分析目前技术所面临的问题并准确的预测未来的研究点。

本综述作为初步总结，主要涉及总结材料的收集工作。其中的文章源均出自中国计算机学会推荐的会议期刊列表。详细内容见后续的 ppt。

XML数据管理技术新进展

周军锋

大纲

- **综述简介**
 - 必要性
 - 目标
 - 意义
 - 信息源
 - 内容提炼
- **综述流程**
- **内容介绍**
- **总结**

综述简介——必要性

- XML应用无处不在
- XML数据大量涌现
 - Gartner[1]预测，XML文件的使用率在
 - 2007年达到40%，
 - 2008年将占据支配地位
 - IDC（国际数据公司）报告显示，在500家受访企业的IT部门中，有29%正在大量使用XML数据库
- XML研究如火如荼
 - 每年各种学术会议期刊发表XML相关论文多达300篇
- 没有系统的总结和比较
 - 发表时间早：大部分出现在06年左右
 - 内容局限性：主要涉及查询，索引
- XML 10 Years !

[1]<http://egovstandards.gov.in/summit/eform/technical-papers/gartneruseofxml.pdf/view>

2008-12-21

3/18

综述简介——目标

- 全面性
 - 技术点，不是文章的简单罗列
- 系统性
 - 技术点之间的内在联系
- 前瞻性
 - 比较分析，提炼令人信服的研究点
- 实用性
 - 从系统层理解XML数据管理技术点之间的逻辑关系
 - 了解XML数据管理技术面临的挑战
 - 了解XML相关的新应用
- 技术路线
 - 全面搜集相关文献
 - 分类整理
 - 总结归纳
 - 比较分析
 - 展望

2008-12-21

4/18

综述简介——意义

- 勾勒清晰的架构
- 全面总结问题
- 指导未来**一段时间**的研究

2008-12-21

5/18

综述简介——信息源

- 要求
 - 全面性
- 06-08年各种会议期刊
 - 国际会议
 - 国际期刊
 - 国内会议
 - 国内期刊

2008-12-21

6/18

综述简介——信息源

- **国际会议**
 - (ACM) **SIGMOD** : (Association for Computing Machinery) Special Interest Group on Management of Data
 - **VLDB** : International Conference on Very Large Data Bases
 - **ICDE** : International Conference on Data Engineering
 - **EDBT** : International Conference on Extending Database Technology
 - **WWW** : International Conference on World Wide Web
 - **CIKM** : International Conference on Information and Knowledge Management
 - **DASFAA** : Database Systems for Advanced Applications
 - **ER** : International Conference on the Entity Relationship Approach
 - **PODS** : Symposium on Principles of Database Systems
 - **SIGIR** : International Conference on Research and Development in Information Retrieval
 - **ICDT** : International Conference on Database Theory
 - **DEXA** : Database and Expert Systems Applications
 - **CIDR** : Conference on Innovative Data Systems Research
 - **WISE** : Web Information Systems Engineering
 - **WAIM** : International Conference on Web-Age Information Management
 - **APWeb** : Asia-Pacific Web Conference
 - **WebDB** : International Workshop on the Web and Databases
 - **INEX** : INitiative for the Evaluation of XML Retrieval
 - **XIME-P** : Workshop on XQuery IMplementation, Experience and Perspectives
 - **XSym** : International XML Database Symposium (08年不存在了)
 - [XML Conference](#) : **应用相关的会议**

2008-12-21

7/18

综述简介——信息源

- **国际期刊**
 - [VLDBJ](#) : **The VLDB Journal**
 - [TODS](#) : **ACM Transactions on Database Systems**
 - [TKDE](#) : **IEEE Transactions on Knowledge and Data Engineering**
 - [TOIS](#) : **ACM Transactions on Information Systems**
 - [JACM](#) : **Journal of the ACM**
 - [CACM](#) : **Communications of the ACM**
 - **IS** : **Information System**
 - **IR** : **Information Retrieval**
 - **KIS** : **Knowledge and Information System**
 - [SIGMOD-Record](#)
 - **DKE** : **Data & Knowledge Engineering**
 - **JDM** : **Journal of Database Management**
 - [WWWJ](#) : **World Wide Web**
 - [JCST](#) : **Journal of Computer Science and Technology**

2008-12-21

8/18

综述简介——信息源

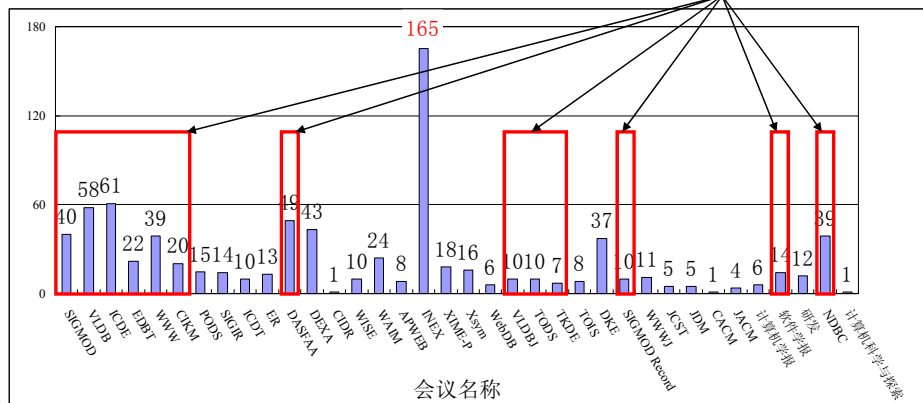
- 国内会议
 - NDBC
- 国内期刊
 - 计算机学报
 - 软件学报
 - 计算机研究与发展
 - 计算机科学与探索

2008-12-21

9/18

综述简介——内容提炼

- 06-08年文献数目: **812** (精读)
- 自己读过的文章: **100/200(150:50)/350** (国内)



2008-12-21

10/18

综述简介——内容提炼

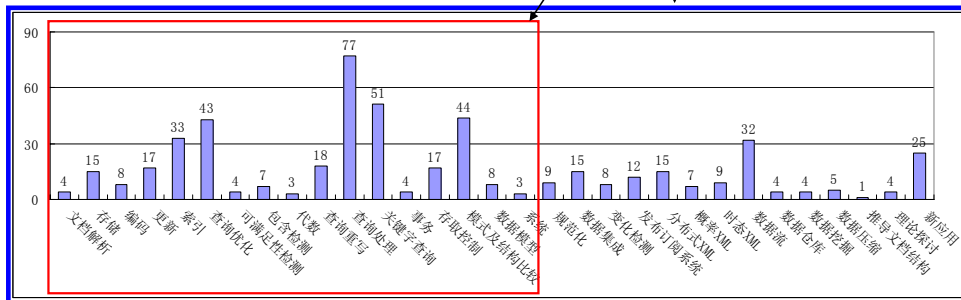
- 如何压缩内容？
 - 06-08: 200/812, 2005年以前的？
 - 如何全面？
 - 一篇综述能容纳多少文章？
 - 如何尽快完成？
 - 还有多少内容需要了解？
 - 要写多少内容？

2008-12-21

11/18

综述简介——内容提炼

- 分类整理, 去除重复: 150/360/500/800



2008-12-21

12/18

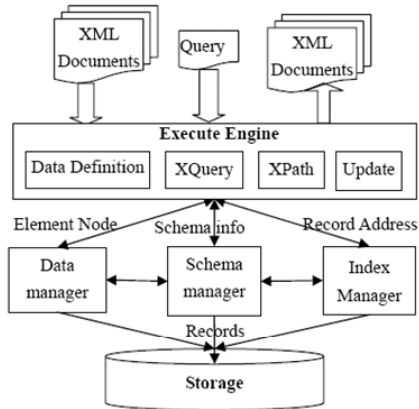
综述流程

● 介绍系统框架

- 4个例子
 - 文档解析
 - 结构化查询
 - 关键字查询
 - 结构化近似查询

● 目的

- 不同模块的作用
- 模块之间的关系



2008-12-21

13/18

内容介绍

● 存储

- 文档解析
- 编码方案
 - 运算速度快，区间编码
 - 支持更新，前缀编码
 - 如何权衡？根据什么权衡？
- 存储策略
 - 关系存储
 - 文档存储
 - 子树存储：划分策略，聚类策略
 - 元素存储
 - 如何存储
 - 单一存储策略：选那个？
 - 混合存储策略：根据什么？
- 更新策略
 - 编码更新
 - 存储更新
 - 视图更新
 - 如何最小化代价？

2008-12-21

14/18

内容介绍

- **模式信息管理**
 - 模式推导
 - 根据文档，倒推模式信息
 - 模式信息查询
 - XPath? 还是别的途径
 - 模式信息比较
 - 规范化

2008-12-21

15/18

内容介绍

- **索引**
 - 结构索引
 - 1-index, F-B index, structural summary
 - 用于匹配查询的结构部分
 - 值索引
 - 倒排表
 - 用于查询的值谓词处理
- **问题**
 - 如何选择合适的索引
 - 如何将索引的处理代价纳入查询计划

2008-12-21

16/18

内容介绍

- **查询处理：结构化，关键字，近似查询**
 - 查询改写：帮助用户提交合适的查询
 - 查询优化
 - 可满足性检测：检测给定查询是否有解
 - 代数优化：查询模式最小化，代价最小化
 - 包含检测：基于视图的优化
 - 实现算法
 - 二元结构连接，整体匹配连接，基于序列的匹配
- **问题**
 - 如何准确提交查询表达式
 - 用户反馈， schema summary， 近似查询
 - 如何返回符合用户查询意图的结果
 - 结果排序， 利用schema附带的语义信息， 定义更好的语义
 - 如何快速求解
 - 涉及高效算法

2008-12-21

17/18

总结

- **XML数据管理技术发展快**
- **全面性和系统性兼顾的综述**
- **已完成论文的收集和分类**
- **接下来进一步归纳分析**

2008-12-21

18/18