XML 检索方法研究

郑 宏,李 年,丁 凯
(北京理工大学 计算机学院,北京 100081)

摘 要:随着 XML 在互联网信息传输、数字图书馆领域的广泛应用,XML 检索成为了 XML 应用过程中需要解决的一个关键问题。研究一种高效、准确的 XML 检索方法对提高系统效率、可用性有非常重要的意义。文中针对 XML 检索技术进行研究,对目前已有的检索算法进行了对比总结。重点从 XML 检索亟待解决的重要问题:XML 查询语言,XML 索引和 XML 检索模型等方面对 XML 检索方法进行了论述。XML 检索模型目前是通过传统模型改进而来,可以通过研究传统索引方法来进一步改进 XML 存储与索引。

关键词:XML 检索;查询语言;索引;检索模型
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2013)10-0015-04
doi:10.3969/j.issn.1673-629X.2013.10.004

Research on XML Retrieval Approach

ZHENG Hong, LI Nian, DING Kai
(College of Computer, Beijing Institute of Technology, Beijing 100081, China)

Abstract: With the extensive application of XML in Internet information exchange and digital library, XML retrieval draws increasing attention from researchers. Research an efficient, precise retrieval approach is of great significance to enhance system efficiency and availability. In this paper, gave an overview of research state in this field, and made a comparison between kinds of approaches. Discussed the XML retrieval approaches majorly from XML query language, XML indexing and XML retrieval model. Recently, get the XML retrieval model from the improvement of traditional model, can further improve XML storage and index from the study of traditional index methods.

Key words: XML retrieval; query language; indexing; retrieval model

0 引 言

传统信息检索往往关注非结构化数据的检索,XML 作为一种典型的半结构化数据,早期不被信息检索领域的研究者们关注。但是,自从 XML 在 1998 年成为 W3C 推荐标准以来,XML 在互联网信息传输、数字图书馆等领域取得了广泛的应用。同时,人们也越来越重视 XML 检索的研究。

1 XML 检索概述

1.1 研究背景

对 XML 检索的研究始于 2002 年成立的 INEX[1](Initiative for the Evaluation of XML Retrieval)评测会议。INEX 提供一种统一的平台,是专门为 XML 检索的不同方法提供统一评测的会议(类似于 TREC 会议)。在目前看来,INEX 是研究 XML 检索最为全面、深入的活动。

1.2 XML 检索问题

XML 检索与传统信息检索的区别在于,XML 检索不仅需要实现文档级的检索,而且还需要实现元素级的检索[2]。同时,XML 检索还需要处理 XML 属性和各元素的关系。对于特定的用户查询,XML 检索系统不但需要返回相关的文档,而且需要进一步返回相关的元素。这样,就可以让用户更容易地获得他们更关心的内容。

具体地说,XML 检索需要研究的问题有:

- XML 查询语言;
- XML 元素的存储与索引;

收稿日期:2012-12-27 **修回日期**:2013-04-02 **网络出版时间**:2013-07-24
基金项目:总装预研基金(9140A04020411BQ0111)
作者简介:郑 宏(1963-),男,副教授,研究方向为计算机集成制造与云计算;李 年(1988-),男,硕士,研究方向为计算机集成制造与云计算。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20130724.1005.040.html

●XML 检索模型方法。

2 XML 查询语言

根据是否需要用户提供查询结果的结构限制,目前 XML 查询语言主要分为两类。一类是 CO (Content Only),CO 是目前传统信息检索最常使用的查询语言,使用这类查询语言的用户只能对查询结果的内容进行限制,使用方法简单,但是不能充分利用文档的结构来获取更精确的查询结果。目前 CO 语言的代表主要有 XRANK[3]、XKSearch[4] 等。另一类是 CAS (Content And Structure),CAS 查询语言是 XML 检索特殊的语言,使用这类语言的用户需要对查询结果元素的内容和结构同时进行限定。这种方法可以更充分利用文档的结构信息,从而查得更精确的结果,缺点是需要用户对 XML 本身比较熟悉,使用起来有一定难度。这类查询语言的典型代表有 XSEarch[5]、XIRQL[6] 等。

表 1 为两类 XML 查询语言对比。

表 1 两类 XML 查询语言对比

XML 查询语言	易用性	准确性
CO	高,只面向内容,用户不需要知道关于 XML 文档内部结构的信息	低,XML 文档结构不能被用于准确表达查询请求
CAS	低,面向内容和 XML 结构。用户需要了解 XML 相关知识和被检索的 XML 文档结构	高,可以通过限制 XML 元素结构来更精确地查询 XML 元素

3 XML 索引与存储

同传统信息检索一样,XML 检索也需要对信息进行索引和存储。与传统信息检索对整个文档进行索引不同,XML 检索需要对元素进行索引和存储。这样才能满足对 XML 文档进行元素级检索的需求。

不同的索引和存储方式会对检索算法的实现造成一定的影响。目前,已有的 XML 索引与存储方式主要有如下几种。

3.1 文档级索引

文档级索引是指对整个 XML 文档进行索引而不考虑 XML 文档的内部结构,只是把 XML 标签也作为索引项。这样,可以利用现有的非结构化索引工具来对 XML 建立索引。目前,Web 搜索引擎大都采用这种索引方式。这种索引方式的优点是能够充分利用传统信息检索索引方法的研究成果,实现简单。缺点是基于这种索引的检索模型不能充分利用文档的内部结构来进行检索。

3.2 直接元素索引

直接元素索引是指把 XML 元素当作传统信息检索中的文档来进行索引。这种索引方式考虑了文档的内部结构。采用这种索引的方法可以对每一种元素建立索引,然后在检索之后合并检索结果[7]。但是这样导致的问题是对于大规模的 XML 文档集索引文件将会很大,导致检索效率随之降低。因为其中包含了很多不必要的索引。

对于上述问题,人们提出了改进的直接元素索引,如根据元素长度过滤掉一些不可能成为检索结果的短元素[8];或者根据用户反馈的频率,选择被用户检索频率高的元素来进行索引[9]。更进一步,因为所有的内容都存在于叶子节点中,人们又提出了只对叶子节点进行索引[10],这样索引文件可以大大减小。这是原始直接元素索引与文档级索引的一个折中。

3.3 数据库索引

文档级索引和直接元素索引都是把传统索引方法应用到 XML 索引的解决方案,而数据库索引则是一种 XML 检索特有的索引方法。

数据库索引是利用目前已有的关系数据库来为 XML 元素建立索引的方法[10-11]。具体地说,是从 XML 文档中抽取出一部分信息,存入数据库,在执行查询处理的时候,通过把 XML 查询语言映射成 SQL 语言进行查询。这种索引方法需要根据 XML 文档的结构来设计数据库表项。一般是将 XML 元素作为 XML 树中的一个节点存入数据库中。对于这种方法设计的数据库表项一般包含:<元素名称,元素路径信息,元素内容>。其中,元素名称和元素内容很容易理解;元素路径信息可以直接是元素的整个路径,也可以是元素路径的一种变换。如 Dewey 编码就广泛应用于一些支持 XML 的商业数据库中。

3.4 基于路径的索引

这种索引方式也是 XML 检索特有的索引方式。

基于路径的索引跟数据库索引类似,需要存储元素的<元素名称,元素路径信息,元素内容>,不同的是元素路径信息一般是元素的完整路径或者元素的 XPath 表示的路径[12]。基于路径的索引方法一般是将索引信息以文件的形式存储。

3.5 多维索引

多维索引是指类似于多维数据库中的并行分析处理的方法来建立索引[13]。多维索引往往是将 XML 文档划分为块来建立索引,在处理查询的时候,可以在各个数据块上并行处理,然后再将结果合并。这有利于提高整个系统的检索性能。

3.6 各种 XML 索引与存储方法对比

表 2 对各种 XML 索引与存储方法进行了对比。

表 2　各种 XML 索引与存储方法对比

方法	复杂性	索引表示	存储	性能
文档级索引	简单,类似传统信息检索索引方法	直接从文档中抽取索引项,跟传统信息检索一样	文件	较差,无法利用 XML 元素位置信息进行索引,只能实现 CO 型查询
直接元素索引	简单,是基于传统信息检索索引方法的改进	利用传统索引模型对每一元素建立索引	文件	较差,索引文件较大
数据库索引	较复杂,基于现有的关系数据库系统,但是需要设计索引表的结构	<元素名称,元素路径信息,元素内容>	关系数据库	较高,利用关系数据库已有的查询优化策略可以取得较好的时间性能
基于路径的索引	较复杂,需要设计索引的表示,并通过文件来管理索引信息	<元素名称,元素路径信息,元素内容>	文件	较高,基于路径的元素索引结构可以有效地减小索引文件的大小
多维索引	复杂,需要对 XML 文档进行块划分,设计索引多维结构,并且要并行查询处理的方法设计	多维索引结构	多维数据库	高,可在每个数据块上并行查询

4　XML 检索模型

XML 检索模型与传统的非结构化检索模型的最大区别是,XML 检索模型不但需要返回相似度最高的元素,而且需要确定返回元素的粒度。目前 XML 检索模型的研究还没有取得重大进展,已有的解决方案大都是在传统信息检索模型的基础之上来建立的。下面介绍几种简单的 XML 检索模型。

4.1　向量空间模型

在文献[14]中作者对传统的向量空间模型进行了扩展,使之能够适应 XML 检索。该模型计算相似度定义为元素向量 C 和查询向量 Q 之间的余弦值,公式如下:

$$\mathrm{Sim}(Q,C)=\frac{\sum_{t_i\in Q\cap C} W_Q(t_i)\times W_c(t_i)}{\|Q\|\times\|C\|} \tag{1}$$

其中:

$$W_{x\in\{Q,C\}}(t)=\log(\mathrm{TF}_x(t)\times\log(\frac{N}{\mathrm{CF}(t)})) \tag{2}$$

公式(2)中,$\mathrm{TF}_x(t)$ 是关键词 t 在元素 C 中的出现频率;N 为数据集中的元素总数;$\mathrm{CF}(t)$ 是包含关键词 t 的元素总数。

由上面的公式可以看出,该向量空间模型在传统向量空间模型的基础之上将查询 Q 与文档 D 的相似度计算公式改成了查询 Q 与元素 C 的相似度计算公式。

4.2　加权布尔模型

使用直接元素索引,布尔模型可以检索包含特定元素的文档。对于那些更高级的查询,当它们描述了 XML 文档中索引项之间的关系时,布尔模型就需要扩展一个新的二目运算符,叫做 contain[15]。这种新的运算符不满足交换率,第一个操作数是 XPath 的类型,第二个操作数是布尔表达式,下面的表达式可以展示 contain 运算符的语法:

Bexprc ::= XPath contain Bexpre　　(3)

Bexprc ::= XPath　　(4)

其中,Bexpre 是一个代表布尔表达式的变量,这种扩展了的布尔模型就允许用户使用 XPath 来完全表达查询元素的结构信息,使用 contain 元素符也可以表述结构与内容之间的包含关系。

例如(UNIX and //title contains (network or TCP/IP) and //topic_area contains (IT or "computer systems"))。

通过这种方式,布尔模型也就能够适用于 XML 检索了。

4.3　统计语言模型

在文献[16]中,作者提出了使用统计语言模型应用到 XML 检索中的方法。该方法配以 Jelinek-Mercer 平滑方法,使用多项语言模型(即每一元素对应一种模型)。元素的语言模型如公式(5)所示[16]:

$$p(e\mid q)\propto p(e)\times p(q\mid e) \tag{5}$$

假设查询输入 q 中的关键词 $t_1,t_2,\cdots,t_k$ 之间相互独立,由公式(5)可进一步转换得到:

$$p(e\mid q)\propto p(e)\times\sum_{i=1}^{k}p(t_i\mid e) \tag{6}$$

这里,可以对元素的语言模型分别用元素所在文档的语言模型和数据集的语言模型进行线性插值,从而防止出现数据稀疏的问题。则 $p(t_i\mid e)$ 可以表示为:

$$p(t_i\mid e)=\lambda_e\times p_{\mathrm{mle}}(t_i\mid e)+\lambda_d\times p_{\mathrm{mle}}(t_i\mid d)+(1-\lambda_e-\lambda_d)\times p_{\mathrm{mle}}(t_i) \tag{7}$$

公式(7)中,$p_{\mathrm{mle}}(*\mid e)$,$p_{\mathrm{mle}}(*\mid d)$,$p_{\mathrm{mle}}(*)$ 均采用的是最大似然估计来分别对元素 e,元素 d 和数据

集建模;λ_e 和 λ_d 是平滑用的插值参数。

5 结束语

文中主要介绍了 XML 检索与传统信息检索的区别,指出了 XML 检索主要解决的问题。然后,从 XML 查询语言、XML 索引与存储、XML 检索模型三个方面对目前研究现状进行了介绍,并进行了简单对比。

XML 查询语言主要的问题是是否需要用户提供 XML 元素结构信息,XML 索引与存储目前主要有传统索引改进的和 XML 特有的方法,而 XML 检索模型方面目前几乎都是通过传统信息检索模型扩展而来,有待进一步研究。

参考文献:

[1] Lalmas M, Trotman A. XML retrieval[M]. Berlin: Springer, 2009.

[2] 张 博,耿志华,周傲英. 一种支持高效 XML 路径查询的自适应结构索引[J]. 软件学报,2009,20(7):1812-1824.

[3] Guo L, Shao F, Botev C, et al. XRANK: ranked keyword search over XML document[C]//Proceedings of the 22nd ACM International Conference on Management of Data. New York, NY, USA: ACM, 2003: 16-27.

[4] Xu Y, Papakonstantinou Y. Efficient keyword search for smallest LCAs in XML database[C]//Proceedings of the 24th ACM International Conference on Management of Data. Baltimore, Maryland, New York, NY, USA: ACM, 2005: 527-538.

[5] Cohen S, Mamou J, Kanza Y, et al. XSearch: a semantic search engine for XML[C]//Proceedings of the 29th ACM International Conference on Very Large Data Base. New York, NY, USA: ACM, 2003: 45-56.

[6] Fuhr N, Gro-johann K. XIRQL: a query language for information retrieval in XML documents[C]//Proceedings of the 24th Annual International ACM SIGIR Conference. New York, NY, USA: ACM, 2001: 172-180.

[7] Mass Y, Mandelbrod M. Retrieving the most relevant XML components[C]//Proceedings of the 2nd Initiative on the Evaluation of XML Retrieval Workshop. Berlin: Springer, 2004: 53-58.

[8] Sigurbjornsson B, Kamps J, Rijke M. The effect of structured query and selective indexing on XML retrieval[C]//Proceedings of the 4th Initiative on the Evaluation of XML Retrieval Workshop. Berlin: Springer, 2006: 104-118.

[9] Liu J, Lin H, Han B. Study on reranking XML retrieval elements based on combining strategy and topics categorization [C]//Proceedings of the 6th Initiative on the Evaluation of XML Retrieval Workshop. Dagstuhl Castle, Germany: INEX, 2007: 170-176.

[10] Sauvagnat K, Hlaoua L, Boughanem M. XFIRM at INEX 2005: Ad-Hoc and relevance feedback tracks[C]//Proceedings of the 4th Initiative on the Evaluation of XML Retrieval Workshop. Berlin: Springer, 2006: 88-103.

[11] 吉训遵,钟 声. 关系数据库中 XML 索引技术研究[J]. 科技传播,2010(14):233-234.

[12] Geva S. GPX-gardens point XML information retrieval at INEX 2004[C]//Proceedings of the 3rd Initiative on the Evaluation of XML Retrieval Workshop. Berlin: Springer, 2005: 211-223.

[13] Luk R, Leong H, Dillon T S, et al. A survey in indexing and searching XML documents[J]. Journal of the American Society for Information Science and Technology, 2002, 53(6): 426-437.

[14] Mass Y, Mandelbrod M. Using the INEX environment as a test bed for various user models for XML retrieval [C]//Proceedings of the 4th Initiative on the Evaluation of XML Retrieval Workshop. Berlin: Springer, 2006: 187-195.

[15] Alin D. XML-QL: a query language for XML[EB/OL]. 1998-08-19[2012-12-23]. http://www.w3.org/TR/NOTE-xml-ql/.

[16] Sigurbjornsson B, Kamps J, Tijke M. Anelement-based approach to XML retrieval[C]//Proceedings of the 2nd Initiative on the Evaluation of XML Retrieval Workshop. Berlin: Springer, 2004: 19-26.

(上接第 14 页)

[8] 乔海泉,田新华,黄柯棣. 将 Simulink 模型用于 HLA 仿真[J]. 系统仿真学报,2006,18(2):335-337.

[9] 郭 斌,熊光楞,陈晓波,等. MATLAB 与 HLA/RTI 通用适配器研究与实现[J]. 系统仿真学报,2004,16(6):1275-1279.

[10] Pawletta S, Drewelow W, Pawletta T. HLA-based simulation within an interactive engineering environment[C]//Proc of Fourth IEEE International Workshop on Distributed Simulation and Real-time Application. Piscataway, NJ: IEEE, 2000: 97-102.

[11] 郭志强,黄 燕,吴 平. Java-MATLAB 集成方法的分析与探讨[J]. 农业网络信息,2006(6):15-17.

[12] Naderlinger A, Templ J, Resmerita S, et al. An Asynchronous java interface to matlab[C]//Proc of 4th International ICST Conference on Simulation Tools and Techniques. Brussels, Belgium: ICST, 2011: 57-62.

XML检索方法研究

作者： 郑宏， 李年， 丁凯， ZHENG Hong， LI Nian， DING Kai
作者单位： 北京理工大学 计算机学院,北京,100081
刊名： 计算机技术与发展
ISTIC
英文刊名： Computer Technology and Development
年，卷(期)： 2013(10)