Web 查询技术研究

Research on Web Query Technologies

孟小峰 曹 巍 王 珊

(中国人民大学信息学院数据与知识工程研究所 北京 100872)

Abstract Now the prevalent technique of Web query is information retrieval based on keyword matching, but this way can not adapt itself to the advanced, complex, structural, and semantic querying demands raised by various Web users. At present, the Web query goes mainly in two directions: one is to perfect information retrieval technique to support structure search as well as semantically content search, the other one is to introduce database query technique into the Web context, such as declarative query language (e. g. SQL, OQL etc), a formal semantics in the form of either calculus or algebra, and exploitation of database schema etc. A new tide is basing the query over XML data, so the XML data query is another active research object.

We introduce some prototypes of Web query languages, and made a comparison between these languages. We also mentioned somistructured data query languages (e. g. Lorel, UnQL). As a paradigm of XML data query language, we presented XML-QL, and a unified approach querying both kinds of data smoothly. In the end, we analyzed both the theoretical flaw and the practical deficiency of semistructured data model and its query language.

Keywords Semistructure, Information retrieval, Database queries, Web query, XML-QL, Unified approaches

一、引言

WWW 的迅速发展，使其成为全球信息传递与共享的日益重要和最具潜力的资源，如何管理 WWW 上的大量信息，以满足用户不断增长的高质量的信息需求？WWW 作为一种新的环境资源，为新技术的产生开辟了新的领域，同时也为传统技术(如数据库、人工智能等)的研究提出了新方向。

在 WWW 被广泛运用之前，较普遍的查询技术主要有：对文档的基于关键词匹配的检索技术、对数据库有结构数据的说明性查询语言(如关系数据库的 SQL、对象数据库的 OQL 等)。但是由于互联网的发展，网上数据不断激增，对网上信息的应用需求也不断提高。原有的对文本文件的链接浏览和关键词检索已无法满足一些复杂的应用需求。近年来大量的研究致力于将数据库技术应用于网上数据的管理和查询，使得查询可以在更细的粒度上进行，并集成多个数据源的数据。而将传统数据库技术直接应用于网上数据的最大困难在于，网上数据缺乏统一固定的模式，数据往往是不规则的并且经常变动。研究这种新类型数据的半结构数据模型应运而生，由于其无模式及自描述的特点适宜于描述网上数据。事实上，日益普及的 XML 数据就是一种自描述的半结构数据，它的出现推动了万维网在电子商务、电子数据交换和电子图书馆等多方面的应用，然而，如何有效地存储管理和查询这类数据目前却莫衷一是。仅靠已有的数据库技术，如关系数据库，面向对象数据库，都不能完全适应于新的应用需求，而专用的半结构数据管理系统仍处于初步的实验阶段。

我们可以预言 XML 将成为数据组织和交换的事实标准，并且大量的 XML 数据将很快出现在 Web 上。实质上，XML 为 Web 的数据管理提供了新的数据模型，可以预见，很多成熟的数据库技术将进入 Web 信息处理领域，把 Web 变为一个巨大的数据库。XML 是朝这个方向迈出的第一步。这种变化给数据库研究界带来了巨大的机会，使得将数据库技术和研究扩展到 Web 数据的管理成为可能。目前对 XML 数据的存储和查询方面的研究正方兴未艾，XML 数据模型与半结构数据模型有着很多的相似性，可以说，XML 是 WWW 上的半结构数据。它既为半结构数据的研究提供了广泛的应用前景，同时也推动了半结构数据研究的发展。本文从 Web 信息的特点出发，分析了数据库技术用于 WWW 相比于信息检索技术的优势，分析比较了正处于研究阶段的结合了数据库技术的各种

Web 相关的查询语言的特点，简要介绍了一个崭新的融合了数据库查询和 XML 的 XML 数据的查询语言——XML-QL 的功能及其特点，最后提出了从一个统一的角度透明或半透明地访问有结构和半结构（可以是 XML）混合数据的构想和范例。

二、数据库查询与信息检索技术的比较

WWW 目前还只是一个巨大的分布的信息检索系统，大多数 WWW 上的搜索引擎是基于信息检索技术，数据库技术与信息检索技术有很多不同，详见下表：

表 1　数据库查询与信息检索

	数据库查询	信息检索(IR)
数据	有结构	无结构
模型	有确定性的模型	基于概率
查询语言	人工的(如 SQL 等)	自然的
查询规范	完全的	不完全的
匹配	精确匹配	部分匹配、最佳匹配
所需条目	基于匹配	基于相关
出错报告	敏感的	不敏感
推理	演绎	归纳
类属	单向度(Monothetic)	多向度(Polythetic)
数据更新	完全支持	不支持
事务	支持	不支持
使用	面向应用	面向人

二者最重要的一个区别是数据库的数据结构性更强，比信息检索的数据包含更多的语义。在一定意义上，信息检索技术更适合于处理无结构数据，数据库则是管理结构数据的最好途径，本质上，信息检索使用近似方法为用户的浏览需求查找相关信息。其中"近似"的含义包括：

(1)近似的查询条件说明。数据库查询中包括了用户所需信息的完全的条件说明，但在信息检索中条件说明总是不完全的，用户有时不能完全描述条件，有时根本无法确定自己要找什么。

(2)近似匹配。数据库的查找基于对条件的完全匹配；在信息检索中，这种匹配也有意义，但是通常用户需要找出部分的匹配查询要求的项，并从中选出最佳匹配的项。

(3)近似结果。信息检索的最终结果传递给用户用于浏览，结果是近似匹配得到的，表征着项之间相关的可能性，所以查询结果无须也不可能非常精确，用户可以进一步分析筛选系统返回的结果，并且信息检索系统中，匹配的失误通常并不显著影响系统性能；而数据库对失误更敏感，匹配的失误意味着系统的整体失败。

数据库中的简单演绎推理的形式为：如果 aRb 并且 bRc，那么 aRc，在信息检索技术中更经常使用归纳推理，关系只由确定或不确定的程度表达，因此推理的可信度是个变量，这个区别导致数据库被描述为确定性的，而信息检索是概率性的。在信息检索中，经常用贝叶斯定理进行推导。

另外一个区别以类属为依据。数据库的类属关系中的类由组成一个类的所有必要和充分的处理属性定义；在信息检索中，类的一个个体将只拥有该类所有个体的所有属性的一部分，类属没有充分或必要的属性。

数据库的查询语言通常是人工语言，有严格的语法和词汇表；在信息检索中，经常使用的是自然语言。

随着电子数据的数量激增和 Web 规模的快速增长，传统的信息检索方法在这样一个无限的信息海洋中准确快速定位所需信息时，越来越显得力不从心，在未来的 Web 进展中如何提高信息检索的准确性和效率成为关键；另一方面，目前出现了超越浏览方式而使信息面向应用访问的迫切需要，从而可以为各种服务提供自主性、互操作性和面向 Web 的应用模式，因为可以确信在不远的将来很多新的应用将在网络上，无结构的 HTML 文档及其相应的信息检索技术将不再适应下一代更复杂的 Web 应用。

基于以上的讨论，可以得出结论：未来的 Web 信息将由更近似于数据库的方式进行管理，而不是目前采用的单一的信息检索的方式。因此，Web 资源需要以有结构的方式进行组织和访问。

三、Web 查询语言分析与比较

目前 Web 上的查询主要基于搜索引擎的关键词索引技术，这种技术的搜索范围可以很大(甚至是整个 Web)，但是也存在一些不足：无法进行对页面内结构和页面间连接的查询、查询的结果重复页面多、查询结果格式的重构能力弱、无法利用用户已知的知识缩小查询范围、无法利用已有的字符串处理或文档处理的库函数、无法反映 Web 的动态变化等等。因此，很多研究侧重在提供功能更强大的 Web 查询和重构语言，如 WebSQL[2]、WebOQL[5]、WebLog[4]、StruQl[6]、ULIX-ES 与 PENELOPE[7]、W3QL[3]等。此外，与 Web 查询有关的理论(如 Web 查询的可计算性[8,9]、路径遍历的优化算法[15]、Web 站点的完整性约束机制等)也都是很受关注的课题；Web 查询的用户界面的研究也很有意义，在这方面，Lore 系统的 DataGuide 为半结构化数据查询提供了一个交互式用户界面，帮助用户利用抽取出来的模式构造查询[17,18]。

3.1　Web 查询语言分析

Web 的最主要的特性是超连接，关于 Web 的模型比较一致的看法是"边标记图"模型，该图模型还可

以描述其他 Web 特有的结构(如支持顺序、嵌套数据结构、集合类型等)。Web 查询是两个任务的合成:基于内容的查询(根据页面内容查询符合条件页面)和基于页面之间连接结构的查询。根据图模型的描述信息的不同粒度和查询语言的功能,可大致将 Web 查询语言分为第一代 Web 查询语言和第二代 Web 查询语言[5]。

1. 第一代查询语言,其主要特点包括:①图的结点是 Web 页面、边是页面之间的连接;②利用现成的搜索引擎实现基于内容的查询,并从借鉴数据库的技术实现基于结构的查询。使用文本模式和描述连接的图模式分别用来描述两种类型的条件,这种查询语言未考虑 Web 页面的内部结构和查询结果的重构。它的典型代表有 WebSQL、WebLog 和早期的 W3QL。(WebSQL 还对连接进行了区分,分为站点内部连接和外部连接,作为分析查询代价的基础[2])

2. 第二代查询语言,其主要特点包括:①图的结点是粒度小于页面内部的数据,边既可以是页面内部的连接也可以是页面之间的连接。一些模型还支持表达更自然的有序集合和记录;②支持查询结果重构成复杂的结构,更多地依赖半结构化数据的查询。典型代表是 WebOQL、StruQL。

3.2 Web 查询语言的评价因素[26]

1)查询语言的表达能力

· 当传统的数据用新的数据模型表示时,查询语言应具有与关系演算或 SQL 相同的表达能力,可以表达传统查询语言中的操作或限定(如投影、选择、连接、分组、排序、聚集、集合操作、数量限定等);

· 对特殊的结构(如超连接),Web 查询应能够通过"导航"遍历连接路径,需要查询语言支持路径表达式;

· 对半结构的特征,语言应能够通过正则路径表达式(或递归的规则)匹配事先不清楚的结构,并通过文本的模式表达式匹配内容,语言对数据类型的限制是宽松的;

· 查询语言的结果重构能力也应该作为评价语言表达能力的一个方面。

2)查询语言的语义 查询语言应该有精确的语义,实现查询语言的转换和优化,

3)语言的合成能力(compositionality) 合成即查询的输出可以作为另一个查询的输入,查询语言支持合成对建立视图很有用处。同时也可以用来评价语言是否有重构能力和语法上的引用透明性。

4)模式与查询的构造 在事先不知道模式的情况下,查询语言应支持对部分模式信息的查询(如 Lorel 中的 path-of 函数),构造更复杂的查询则需依赖模式抽取技术。可计算的 Web 查询(主要是第一代查询语言)的构造更多依赖引入点(初始页面),通过导般的方式构造、

5)可生成性 可以看到,各查询语言的语法非常复杂,可行的做法是由界面更友好的应用程序(如自然语言理解系统或图形处理系统)生成需要的查询语言。

6)查询语言与其他技术的集成 web 查旬语言与其他技术合作,开发更完善的功能(如文档检索、机器学习等),或者支持用户自定义的函数或谓词等。

表 2 给出了我们对目前主要语言的特点的对比分析[1]。

表 2 Web 查询语言比较

语言名称			数据模型	语言风格	路径表达式	结果的重构
Web数据查询语言	侧重查询	WebSQL	关系	SQL	支持	无
		W3QL	边标记多图	SQL	支持	无
		WebLog	关系	DataLog	不支持	有
	侧重重构	WebOQL	超树(hypertree)	OQL	支持	有
		StruQL	边标记图	DataLog	支持	有
		Ulixes & Penelope	关系和页面模式	ADM DDL 和 SQL	支持	有
半结构化数据语言		Lorel	边标记图	OQL	支持	有
		UnQL	"树"	结构递归	支持	有
XML 语言		XML-QL	标记图	类似 SQL 或 OQL	支持	有

四、XML 及其查询语言

近年来,WWW 委员会为 Web 上结构文档的交换提出了一种新的数据格式标准——XML[19]。XML 与半结构数据非常近似。和 HTML 相比,HTML 语言是面向显示的,信息的内容靠用户来理解;而 XML 语言是面向内容的,其语义隐含在语言的标记中,因此 XML 更适合用来表示数据。XML 形式的 Web 数据不仅是一种新的 Web 数据组织形式,而且它的面向数据交换的特性推动了 Web 应用模式的发展,它反映的是有结构并具有自描述能力的数据。已经有很多研究是关于 XML 数据存储、XML 数据的查询和 XML 的系统实现与应用模式等方面,以 XML 家族为基础的新一代的 WWW 环境是直接面对 Web 数据的,不仅可以很好地兼容原有的 Web 应用,而且可以更优地实现 WWW 这一分布计算环境下的信息共享与交换。因此,它成为 Web 信息发展的可喜趋势。

数据的检索、转换和集成都是已解决了的数据库问题,它们都依赖一种查询语言,关系的 SQL 或面向对象的 OQL。但 XML 数据与关系数据和面向对象数

据均不同，因此上述的查询语言不能直接用于 XML。然而 XML 数据模型与近来数据库界研究的半结构化数据模型很相似，一些处于研究阶段的查询语言已被设计并运用于半结构化数据，在此基础上 AT&T 实验室的专家们提出了一种基于 XML 的查询语言称为 XML-QL[24]，它用查询的方式可实现 XML 数据的检索、转换和集成。

4.1 XML-QL 简介

XML-QL 是在查询语言(UnQL 和 StruQL)基础上设计的，它能对 XML 文档进行查询、构造、转换和集成。XML-QL 集中了查询语言技术和 XML 语法格式，它通过说明路径表达式和模式的方式，给出 XML 数据的提取条件(WHERE 子句)。同时 XML-QL 中可以给出构造查询输出的 XML 数据的模板，其输出结果仍为 XML 文档(CONSTRUCT 子句)。

XML-QL 有类似 SELECT-FROM-WHERE 的结构(WHERE-IN-CONSTRUCT)，与 SQL 很相似。但 XML-QL 有一些很重要的区别于基于结构化数据查询语言的特点。其 WHERE 子句由两部分组成：模式和条件表达式，这意味着被选出的数据项要满足两个条件：(1)数据项的类型(或 schema)和值必须与指定的模式匹配；(2)数据项的值要满足条件表达式。

在查询条件中加入模式匹配是 XML-QL 与半结构化查询语言和结构化查询语言最大的不同之处。下面是 XML-QL 的标准结构。

```
WHERE<tag1>
        <tag2>abcdefg</tag2>
        ......
        <tag3>$a</tag3>
  </tag1>IN URL(Data Source)
  $a>100
  and.....
CONSTRUCT<result>
          ......
        </result>
```

模式

条件表达式

结果(XML 形式)

XML-QL 可以利用绑定变量、嵌套查询等特点实现关系代数中的选择、连接、投影、分组、排序等操作，在模式中用正则路径表达式支持特殊的图结构的数据。XML-QL 用嵌套查询处理含有可选数据的查询。另外，XML-QL 还支持 tag 变量，实现对结构的查询。在查询结果的构造中，变量既可以绑定到元素的内容上，也可以绑定到元素本身。前者可以实现灵活的重构功能；后者则避免了在结果中重复创建相同的元素。XML-QL 既支持对输入元素的顺序的限定，也支持对结果数据的排序。XML-QL 可用来实现 XML 数据的重构，和不同数据源集成的数据视图。但是也有一些不足的地方，如目前不支持聚集函数和数值表达式的计算等。为查询结果自动推导 DTD 也是一个值得研究的问题。

五、Web 数据查询的统一方法[26]

半结构数据和 XML 数据的查询取得了可喜的研究成果，但是单纯的半结构数据查询却排除了数据的有结构部分的严格的类型系统和有效的实现机制。在实际应用中，随着越来越多 Web 数据进入传统的应用，包含了有结构和无结构部分的混合数据的应用将逐渐普遍，因此有必要研究一个可行的统一访问方法，可以同时利用这两种类型数据的优点。

XML 可以看成是 Web 上数据交换的语法，XML 的查询语言应该方便并具有高度表达能力以访问 XML 数据。考虑到数据交换的应用，XML 的数据模型及其相关的查询语言应具有明确定义的简洁的语义和可接受的标准性。其语义既反映 XML 的目前状态，又可灵活地适应未来的扩展，并且不应忽视近年来已开展的关于数据模型和查询语言的研究和开发。

首先我们来看数据模型。Web 上的数据具有很大的差异性，从具有松散结构的不规则的文档(如 HTML 主页)到具有良好结构的信息(如从关系数据库管理系统中抽取出来的数据)。应用只有在理解了数据资源的语义时才会充分利用丰富的数据资源。因此，XML 基于的数据模型应允许表达有松散结构的数据，同时，该模型应可以同样描述有良好结构的数据。

再来考虑查询语言。语言应提供标准的基于关键词的模式搜索(基于全文索引)，这些方法可见于信息检索和 Web 上的标准搜索引擎。该语言还应通过 XML 的标记结构提供导航的形式。更进一步，该语言还可以采用标准数据库的风格(例如使用 SQL、OQL 或某一变种)对有结构数据的进行说明性的和具有表达能力的访问，同时还应具有一种简洁正规的语义。OQL 被认为是 XML 查询语言的合适的基础。

因此，与其提出一种新的数据模型和查询语言，不如考虑从对象数据库的事实标准 ODMG 模型和查询语言 OQL 展开研究。目标是该模型的扩展和可以处理结构松散数据(即缺乏类型)、模式搜索和导航(即沿循标记或连接)的语言。这种对象风格的模型和语言方案的好处是可以遵循 DOM 和 XML 的规范实现标准的功能。

由斯坦福大学设计开发的 Ozone 系统，扩展了 ODMG 数据模型，使其不仅能表示标准的有结构的对象数据，还可以描述基于 OEM 模型的半结构数据(或者 XML 数据)。它保持了 ODMG/OQL 和 OEM/Lorel 各自的优点，并通过二者的结合，可以为 ODMG 数据提供 OEM 模型的语义，同样也支持标准的 ODMG 应用通过 ODMG 视图访问半结构数据(或

XML 数据；对于 XML 数据，该系统还考虑了 DTD 对数据模型和查询语言的影响，可以利用 ODMG 丰富的类型系统实现 DTD 中定义的数据。Ozone 系统的查询语言 OQL` 扩展了 OQL 语言，可以查询混合数据。整个系统是在 ODMG 兼容的对象数据库系统 O2 上开发的[27]。从 Ozone 的经验中可以得出两方面的经验：一是在 XML 数据模型和查询语言的设计中，数据库界多年来的研究和开发成果不容忽略。ODMG 和 OQL 是 XML 的优良的方案。二是适宜的数据模型和查询语言应该允许有结构数据和 XML 的简洁的集成，同样 ODMG 和 OQL（如 Ozone 系统中的扩展）也是良好的备选方案。

总结 尽管目前已经有很多半结构化数据模型以及查询语言被提出，但存在许多不足之处。如目前所提出的查询语言，虽然具有一定的描述能力，但是还存在一些不足。主要表现在：在理论上没有给出半结构数据查询的表达能力的标准及其形式描述；半结构查询语言的语法繁杂：表现在路径的正则表达式的表达复杂，特殊的数据结构（如数据引用）的表达也很复杂，使查询只能为一些非常熟悉系统的专业人士构造；这些查询语言基于的模型各有差别，语言的语法各异，语义功能也不相同，没有一个实现半结构数据查询的统一的方法，以适应于半结构数据集成的需要；半结构数据查询系统对于用户来说是被动的，系统只能回答由用户提出的也许可能不规范（相对于本系统的数据模型而言）的查询，系统的容错性虽然可以掩盖这些查询的不正确，但是不能够从不正确查询推导出符合用户意图的正确查询，主动地引导或者迎合用户所需的能力更是薄弱；针对不同的网络终端，如台式机、便携机和掌上机等不同类型，查询构造和结果的重构与显示方面没有因地制宜地专门研究。

参 考 文 献

1 Florescu D, Levy A, Mendelzon A. Database Techniques for the World-Wide Web: A Survey. ACM SIGMOD Record, 1998, 27(3)

2 Mendelzon A O, Mihaila G A, Milo T. Querying the World Wide Web

3 Konopnicki D, Shmueli O. W3QS: A Query System for the World-Wide Web. Available at: http://www.cs.technion.ac.il/~konop

4 Lakshmanan L V S, et al. A Declarative Language for Querying and Restructring the Web

5 Arocena G O, Mendelzon A O. WebOQL: Restructuring Documents, Databases and Webs, IEEE 1998. 24～33

6 Fernandez M, Suciu D, et al. A Query Language for a Web-Site Management System

7 Atzeni P, et al. To Weave the Web. In: Proc. of 23rd VLDB Conf. Athens Greece, 1997

8 Abiteboul S, Vianu V. Queries and Computation on the Web, 1997

9 Mendelzon A O, Milo T. Formal Models of Web Queries. ACM-PODS, 1997

10 Christophides V, et al. From Structured Document to Novel Query Facilities. In: Proc. of ACM SIGMOD Conf. on Management of Data. Minneapolis, Minnesota, 1994. 313～324

11 Abiteboul S, et al. The Lorel Query Language for Semistructured Data. Available at: http://www-db.stanford.edu/~lore

12 Buneman P, Davidson S, Suciu D, et al. A Query Language and Optimization Techniques for Unstructured Data. SIGMOD '96, Montreal Canada. 505～516

13 Suciu D. An Overview of Semistructure Data. SIGACT News, 1998, 29(4): 28～38

14 Nestorov S, Abiteboul S, Motwani R. Extracting Schema from Semistructured Data. In: Proc. of ACM SIGMOD Conf. on Management of Data. Seattle, WA., 1998

15 Milo T, Suciu D. Index Structures for Path Expressions. ICDT, 1999

16 Abiteboul S. Querying Semistuctured Data. In: Proc. of Intl. Conf. on Database Theory, 1997

17 Goldman R, Widom J. DataGuide: Enabling Query Formulation and Optimization in Semistructured Databases. In: Proc. of the Int. Conf. on Very Large Data Bases (VLDB). Athens, Greece, 1997

18 Goldman R, Widom J. Interactive Query and Search in Semistructured Databases. In: Proc. of the International Workshop on the Web and Databases. Valencia, Spain, 1998. 42～48

19 XML 1.0, W3C Recommendation 10-February-1998. Available at: http://www.w3.org/XML

20 Suciu D. Semistructured Data and XML, 1999. Avaiable at: http://www.research.att.com/~suciu

21 Deutsch A, et al. Xml-ql: A query language for xml, 1998. Available at: http://www.w3.org/TR/NOTE-xml-ql/

22 Widom J. Data Management for XML: Research Directions, 1999. Available at: http://www-db.stanford.edu/~widom

23 Shanmugasundara J, et al. Relational Databases for Query XML Documents: Limitation and opportunities. In: Proc of the 25th VLDB. Edinburgh, Scotland, 1999. 302～314

24 McHugh J, Widom J. Query Optimization for XML. In: Proc. of the 25th VLDB. Edinburgh, Scotland, 1999. 315～326

25 Abiteboul S, Buneman P, Suciu D. Data on the Web-From Relations to Semistructured Data and XML. Morgan Kauffmann Publishers, 2000

26 Abiteboul S, Widom J, Lahiri T. A Unified Approach for Querying Structured Data and XML. 1998. Available at: http://www.w3.org/TandS/QL/QL98/

27 Lahiri T, Abiteboul S, Widom J. Ozone-Integrating Structured and Semistructured Data

28 Yang W. XML and Mobile Agent Technology Enabling the Virtual Enterprise: A Model for Distributed Document Management. In: Proc. of AI'2000. Innsbruck, Austria, Feb. 2000

29 孟小峰. The Framwork of Querying Web by Natural Language. In. Prco. of APPLIED INFORMATICS-AI'2000. Innsbruck, Austria, February 2000. 14～17