基于 XML 的 Web 数据挖掘的研究*)

刘振岩 王万森
(首都师范大学信息工程学院 北京100037)

Research of Web Data Mining Based on XML

LIU Zhen-Yan WANG Wan-Sen
(College of Information Engineering, Capital Normal University, Beijing 100037)

Abstract The paper advances a system framework of Web data mining based on XML. This system framework integrates Information Retrieval with Information Extraction, and utilizes traditional data mining methods to complete Web data mining through XML.

Keywords Web data mining, XML, Information retrieval, Information extraction

1. 引言

传统的数据挖掘方法一般是针对数据库或数据仓库中的结构化数据进行的,但在现实世界中,人们面对的数据绝大部分是属于非结构化或半结构化的,例如 Web 页面。我们知道,Web 的数据量目前至少可以用数百兆兆字节计算,且仍在迅速增长。这些数据一方面为数据挖掘提供了丰富的资源,另一方面也对数据挖掘技术提出了严峻的挑战。与传统的数据挖掘相比,实现 Web 数据挖掘的主要困难表现在以下三个方面:第一,Web 页面缺乏统一的结构,Web 上的每一个站点就是一个数据源,这些数据源都是异构的;第二,Web 页面的数据是属于半结构化的,它既不是完全无结构的,也不是完全结构化的;第三,Web 上的信息是在不断更新变化的。因此,Web 数据挖掘正在成为数据挖掘领域中的一个极富挑战性的新的研究领域。

目前,对 Web 数据挖掘的分类有两种不同观点,一种观点认为 Web 数据挖掘可分为 Web 内容挖掘、Web 结构挖掘及 Web 使用记录挖掘三类;另一种观点认为,Web 结构挖掘可以作为 Web 内容挖掘的一部分,即 Web 数据挖掘可以分为 Web 内容挖掘和 Web 使用记录挖掘两类。本文依据后一种观点,重点讨论 Web 内容挖掘,并在此基础上提出了一种基于 XML(eXtensible Markup Language)的 Web 数据挖掘的系统构架(XWDM)。该构架主要结合信息检索(Information Retrieval,IR)、信息抽取(Information Extraction,IE)和传统的数据挖掘(Data Mining,DM)方法,以 XML 为媒介进行 Web 数据挖掘。

2. 系统构架

2.1 HTML 页面及其缺陷

HTML 作为 Web 页面信息的主要载体,它可以在用户界面这个层次上提供丰富的显示效果,是目前被广为接受的一种网络上的流行语言,具有简单、易用的特点。但是,HTML 无法提供管理数据的标准方式,在数据管理方面的功能明显不足。并且,由于 HTML 标记几乎不含任何数据信息,因此很难支持对数据的搜索,即 HTML 只是描述了页面的外观形式,而不能显示其数据。

目前,人们正在试图采用各种方法去尝试对 HTML 页面的数据抽取。其中的多数方法都是先采用一些专用查询语言把 HTML 页面的各个部分映射成为代码,然后再用这些代码将 Web 页面上的信息填入到数据库中。尽管这些方法都能实现一定的数据抽取功能,但往往是不切实际的。其主要原因有以下两个方面:第一,它们需要开发人员花费一定的时间去学习一种无法在其它情况下使用的查询语言;第二,在健壮性方面存在严重缺陷,当目标 Web 页面有所改动时,哪怕只是很简单的更改,它们都难以处理。

2.2 XML 及其特性

XML 是由万维网协会(W3C)设计的一种中介标示语言(Meta-markup Language),它提供了描述结构化数据的格式,可以通过独立运行程序的方法来共享数据。同时,XML 又是一种是用来自动描述信息的新的标准语言,它能使计算机通过 Internet 的强大功能把信息传递到人类的各种活动中去。

与 HTML 相比,XML 具有内容与形式分离的特性,以及良好的可扩展性、跨平台移植性和自描述性。

(1)内容与形式的分离性 在 HTML 中,数据内容和表现形式是混在一起的。这样,当数据的表现形式需要改变时,文档更新的工作量就比较大。而对于 XML 文档而言,标记是包含信息的,比如关键字、继承关系等,这些信息对于数据的检索、描述将起到极大的简化作用。利用 XML 的这一特性,当数据的表现形式有所改变时,仅需修改从 XML 文档中分离出的用于描述数据表现形式的样式单就可以了。

(2)良好的可扩展性 XML 允许程序员制定自己的标记集,允许一个行业或某一个特定领域制定在本范围内的通用标记集。这样,XML 就可以轻松地适应每一个领域而不需要对语言本身作大的修改。另外,由于 XML 的数据定义和数据本身也是分离的,这就使得 XML 的标记集不会无限扩大。

(3)良好的跨平台移植性 XML 语言可以定义各种数据,如文本、图像、声音等。虽然这些数据的格式不同,但 XML

*)本文的研究工作得到北京市自然科学基金(4012006)的资助。刘振岩 大学在读博士生,主要研究领域为人工智能、专家系统、数据挖掘。王万森 硕士,主要研究领域为数据挖掘、专家系统。 教授,西北工业

能通过一种用于交换数据格式的文件—XML 文档，来处理由 XML 标注的各种数据，从而实现不同格式数据的跨平台交换。

(4)良好的自描述性 HTML 良好的自描述性使得其数据能够被不同的应用程序分析处理。并且，人们可以通过标记及元素之间的关系，很清楚地看出数据要表达的内容。

2.3 系统构架

从以上分析可以看出，虽然 XML 本身也是一种标记语言，但由于它是直接面向 Web 数据的，因此在数据描述方面大大优于 HTML。为克服 HTML 的缺陷，利用 XML 的特性，我们可以以标准的 Web 技术—XML 为媒介进行 Web 数据挖掘。采用这种方法，需要首先将与 Web 挖掘任务相关的 HTML 页面转化为 XML 数据源，然后再利用传统的数据挖掘算法在 XML 数据源上进行知识的获取。

这种基于 XML 的 Web 数据挖掘方法的系统构架如图1所示。在该图中，信息检索(即 IR)的主要任务是搜索到与 Web 挖掘任务相关的 HTML 页面。信息抽取(即 IE)的主要任务是，首先定制生成 XML 元数据模板，然后将 HTML 页面的相关数据填充到 XML 元数据模板中，形成 XML 数据源。数据挖掘(即 DM)的主要任务是在 XML 数据源上应用传统的数据挖掘算法去获取有用的知识，形成知识库。

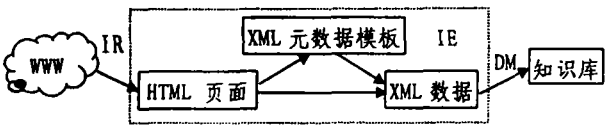


图1 基于 XML 的 Web 数据挖掘(XWDM)的系统构架

3. 实现方法

3.1 将 HTML 页面转化为 XML 数据源

由图1可以看出，将 HTML 页面转化为 XML 数据源的工作是由 IE 来完成的。IE 的具体任务就是将 HTML 页面的非结构化或半结构化的数据转为结构化的数据—XML 数据。而要实现这种转化，IE 必须事先定制一个用于容纳转化后的数据的 XML 元数据模板。另外，并非所有的 HTML 页面都与数据挖掘任务有关，因此在执行 IE 之前需要先执行 IR，获取与数据挖掘任务相关的 Web 页面。

IR 的首要任务是考虑所获得的数据源是否可靠。所谓可靠是指数据源是否有可靠的网络连接?它的生存期有多久?网页的布局结构是否稳定?其中，布局结构的稳定性尤为重要。由于一些网站的布局结构可能会经常更新，这就给 XML 元数据模板的定制带来了一定的困难。

IR 最简单的实现方法是采用静态定制方式，即先由设计人员凭经验确定若干个相关的 HTML 页面的网址，观察其页面布局结构，从中找到相应的数据引用点。这种方法的最大优点是简单，但其不能跟踪动态变化的信息，局限性太大。要解决这个问题，需要引入数据挖掘技术。其大致过程是：先由专家收集一定数量的相关的 Web 页面形成训练数据集，并在此训练数据集上应用数据挖掘的某种分类算法生成一个具有一定智能的分类器；然后再利用这个智能分类器对更多的 Web 页面进行筛选分类。这样，就可以得到最新的数据源，并能够跟踪动态变化的信息。

当 IR 获得了可靠的数据源，并搜集到相关的 Web 页面后，就可以把控制权交给 IE。IE 获得控制权后，首先定制生成所需要的 XML 元数据模板，然后利用这个元数据模板将 HTML 页面的相应数据转化为 XML 数据，形成 XML 数据源。

通常，可将 IE 分为两种类型，一种是从非结构化的文本中进行数据抽取，称为传统的 IE；另一种是从半结构化或结构化的数据中进行数据抽取，称为结构化的 IE。这两类 IE 有着很大的差别，传统的 IE 着重于诸如语义、语法分析等方面的研究，而结构化的 IE 主要是利用一些元数据信息，像 HTML 标记等，来辅助数据抽取。例如，对 HTML 页面，所需的数据通常会处在 HTML 树中深度嵌套的单个〈table〉或〈div〉标记里，这就使得数据抽取的过程可以只在文档的一小部分执行。本文所讨论的 IE 主要是指结构化的 IE。由于它充分利用了 HTML 的标记，因此使数据抽取工作变得相对简单。

在 IE 所定制的 XML 元数据模板中，主要包括相关的 Web 页面的网址以及 Web 页面上的数据引用点的位置，即 HTML 嵌套树中相关数据所在的某个〈Table〉或〈div〉等。XML 元数据模板实质上是放置了特殊标记的 XML 文档，这些特殊的标记以后就可以用 Web 页面的具体的数据值代替，而其定义可通过一个 XML schema 元语言文件来实现，即用一个 XML schema 文件来设定这个 XML 元数据模板的文件结构和数据类型等。而用 Web 页面的具体的数据值代替这些特殊标记的过程，即填充 XML 元数据模板的过程，就完成了将 HTML 页面转化为 XML 数据源的功能。

3.2 在 XML 数据源上进行数据挖掘

XML 提供了 DOM(Document Object Model)和 SAX(Simple Application for XML)两种数据访问接口。外部应用程序可以通过这两种接口很方便地访问 XML 文档。

DOM 接口是由 W3C 制定的。该接口定义了一系列用来实现对 XML 文档数据进行访问和修改的对象，并将 XML 文档转换为树型的文档结构。这些对象树是 XML 文档内元素之间关系的反映，通过它们，就可以访问和修改 XML 文档的全部数据。应用程序也可以通过这种树型结构对 XML 文档数据进行层次化访问。对于文档中的信息，如数据、数据的意义和数据的关系等，都可以由 DOM 接口将其转换为树型结构中的节点或节点之间的关系。由于所有的 XML 文档信息都能被树型结构所包含，这就使得对 XML 文档数据的随机访问变得十分方便。目前，DOM 的标准是分两级制定的，包括 DOM 级别1和 DOM 级别2。

DOM 接口是对 XML 的全面解析，它需要将完整的 XML DOM 树全部放入内存，其随机访问速度非常快。但是，当 XML 文档比较复杂、庞大时，需要占用的内存空间较多，并且对 DOM 树的访问速度也会大大降低。

SAX 接口可以避免 DOM 接口的这一缺陷，它不需要将所有的文档数据全部放入内存，并对数据文档采用了时间驱动的顺序访问方式。在 SAX 接口中，当 XML 解析器遇到特定的事件时，会调用相应的函数来处理该事件。当然，SAX 接口也只仅仅是调用相应函数，至于具体的数据的处理过程是通过函数的执行来完成的。SAX 不是 W3C 的标准，它是由 Internet 上一群热衷于 XML 技术的人共同研究出来的。作为一种基于事件的 XML 编程接口，SAX 已被各种 XML 团体广泛认可。

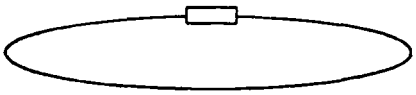
Microsoft .NET Framework 提供了对 XML DOM 对象

(下转第70页)

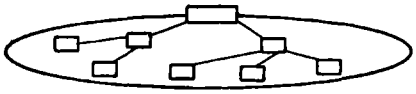
模、异构性、欺骗行为普遍、不易协调等等都会被淡化消减，一些适合更小规模、需要密集利用人工劳动的技术也能够应用，建造更加智能灵活的搜索引擎。

综合采用上述技术，并配合人工/专家劳动和用户参与，在限定领域范围，领域型搜索引擎将提供比大规模综合型搜索引擎更智能、高质量的 IR 服务。一个我们身边的例子是科学院的 FTP 搜索引擎，检索质量可以说是令人满意的。

结论和未来的工作 传统搜索引擎将 Web 看作一个无结构文档库，对每个文档、文档中的每个词一视同仁地进行处理，力图索引尽量多的文档，搜索引擎是用户与 Web 这个巨大文档库交互的唯一接口，该视图可以用图4(a)表示。



(a)传统的搜索引擎与 Web 的关系



(b)在理想的搜索引擎服务框架下的搜索引擎视图

图4 不同视图下搜索引擎与 Web 的关系

而在相互合作的搜索引擎服务框架下，Web 被看作一个分门别类的知识库，如图4(b)所示。综合型搜索引擎的目的是把握全貌，而各种领域型搜索引擎则能够向用户提供深入、智能、更新及时的检索服务。每个搜索引擎都会就用户查询建议其他适合的搜索引擎。

由于不同的搜索引擎处理的信息，面向的用户群、发挥的作用都不用，它们分别适合采用不同的技术。结构分析技术适合应用于综合型搜索引擎；Metadata 演算、数据库、人工智能等技术一方面由于需要的人工劳动多、耗费资源大、要求原始数据规范性好，另一方面能够构造深入细致服务，适合于在领域型搜索引擎上应用。各种搜索引擎不再是各自为政，它们建立一个松散耦合的合作网络，以相对于整个合作网络最低的代价为用户的各种 IR 需求提供适当的服务。

事实上搜索引擎服务框架的核心思想不是一个全新的技术发明，当面临一个大型问题时，人们往往采用建立一个各司其职、相互合作的层次结构(Hierarchy)这种思路，如企业或国家的管理。搜索引擎服务框架也是我们在 Web 从完全的"无政府状态"向适当的有序化发展的一个设想和建议。

在我们的原型检索系统 SAInSE[1]中，我们尝试着对理想服务框架中的部分思想进行了检验。实验的效果是明显的。如果能够实现这样一个服务框架的话，那么对于 Web 上的查询，无论是查全，还是查准，无疑都是十分理想的解决方案。

参 考 文 献

1 马国臻. 基于结构分析的大规模 WWW 文本信息检索技术的研究:[2001中科院计算所博士论文]

2 Ricardo Baeza-Yates Berthier Ribeiro-Neto. Modern Information Retrieval. ACM Press, 1999

3 CLEVER. CLEVER Project, 1999

4 Collins, Collins J A, Schweitzer R. Applying metadata to the search interface: Constructing effective local and distributed searches of web-based scientific data. In: proc of the 5th WWW conf. 1994. On line at: http//www. cdc. noaa. gov/~jac/www. 95/paper. html

5 COOL links 1995 L. Jay Wantz, Micheal Miller. Towards user-centric navigation of the Web: COOL links using SPI. In: proc of the 6th WWW conf. 1995

6 COOL linds 1996 Michael Miller, L. Jay Wantz. COOL links: ride the wave. In: proc of the 7th WWW conf. 1996

7 Dean 1999 J. Dean, M. R. Henzinger. Finding related pages in the world wide web. In: 8th World Wide Web Conference, Toronto, May 1999

8 Silverstein 1998 Craig Silverstein, Monika Henzinger, Hannes Marias, Michael Moricz, Analysis of a very large AltaVista query log, DEC System Research Center (SRC) technical note, Oct. 1998

(上接第43页)

模型的支持，这种支持是通过一系列相关的类来实现的。对于 SAX 接口，在.NET 中也有相应的模拟实现。在 XML 数据源上，借助于 XML 提供的 DOM 接口或 SAX 接口，利用传统的数据挖掘算法即可进行 Web 数据挖掘，获取有用的知识，形成知识库。

我们已经利用 Microsoft VS. NET 的开发工具，实现了对若干农业网站上的农产品供求信息的 Web 数据挖掘。实践证明，本文提出的利用 XML 这种半结构化的数据模型辅助进行 Web 数据挖掘的方法是行之有效的。

结束语 Web 数据挖掘是一个新的很有前途的研究领域。由于 Web 中的数据是半结构化的，因此 Web 数据挖掘不同于传统的数据库中的结构化数据挖掘。如何针对 Web 上的数据为半结构化的这一特点，寻找一种半结构化的数据模型是 Web 数据挖掘必须解决的一个首要问题。当然，仅有这种半结构化的数据模型还不够，还需要有相应的半结构化模型抽取技术，即自动地从现有数据中抽取半结构化模型的技术。面向 Web 的数据挖掘必须以半结构化的数据模型和半结构化数据模型抽取技术为前提。XML 可看作一种半结构化的数据模型，它可以很方便地将 XML 的文档描述与关系数据库中的属性一一对应起来，实现精确的查询与模型抽取。

本文结合 IR 和 IE，提出的基于 XML 的 Web 数据挖掘方法已在实际系统中得到了成功应用。我们相信，此项研究一定会对 Web 数据挖掘起到积极的促进作用。同时，随着 XML 这一标准 Web 技术的不断发展，Web 数据挖掘的研究也必将取得越来越多的成就。

参 考 文 献

1 (加)Han J, Kamber M 著，范明，孟小峰 等译. 数据挖掘概念与技术(Data Mining: Concepts and Techniques). 机械工业出版社，2001

2 kosala R, Blockeel H. Web Mining Research: A Survey. ACM SIGKDD, July, 2000

3 Jackson J, Myllymaki J. Automatically extract information with HTML, XML, and Java. www-900. ibm. com , 2001

4 王超，张鹏. ASP. NET/XML 深入编程技术. 北京希望电子出版社，2002

5 徐振航，刘莉芹. XML 与面向 Web 的数据挖掘技术. www. ASP-Cool. com , 2001

6 刘振岩，王万森. 急切分类与懒散分类的研究. 小型微型计算机系统，2002待发