第 25 卷 第 2 期
2010 年 4 月

天津科技大学学报
Journal of Tianjin University of Science & Technology

Vol.25 No. 2
Apr. 2010

利用 XML 的一种因果模式 Web 挖掘模型

李孝忠，赵国桦
（天津科技大学计算机科学与信息工程学院，天津 300222）

摘 要：定义了一种因果关系模式，包括因集、果集和各种因果关系，以及影响度、分类权值等参数. 再结合统计分析等各种数据挖掘算法和 XML 的优势组成了一种 Web 挖掘模型，该模型可用来发现 Web 上事物之间的内在联系以及发生规律，以便为任务的执行提供有力的预测和决策依据. 应用实例表明，该模型是有效的，能够在预测和决策中发挥作用.

关键词：Web 挖掘；因果模式；因集；果集；影响度
中图分类号：TP18 **文献标志码**：A **文章编号**：1672-6510(2010)02-0065-03

A Causal Pattern Web Mining Model Based on XML

LI Xiao-zhong，ZHAO Guo-hua
（College of Computer Science and Information Engineering，Tianjin University of Science & Technology，Tianjin 300222，China）

Abstract：A causal pattern was defined，including resulting set，factor set，a variety of causal relations and kinds of parameters such as impact degree，category weight. Then combined with statistical analysis of various data mining algorithms and the advantages of XML，a Web mining model was formed to find the intrinsic link and the occurrence between things on the Web. The model could be used to provide a strong basis of the forecasting and decision-making for the future implementation of the mandate. Application example shows that the model is effective and benefits the forecasting and decision-making.

Keywords：Web mining；causal pattern；factor set；result set；impact degree

Web上包含了巨大的信息，如何从浩瀚的Web 信息资源中快速、准确且高效地搜索和发现用户感兴趣的信息和知识已经成为一个急待解决的问题[1–3]. 这些信息中，有一些有价值的信息是可以被用来进行预测和分析的，可是到目前为止，Web 数据挖掘的研究在这方面还没有形成一个明确的领域. 只是在传统的方法中，如关联规则(仅只是从事物发生的相互关联度来进行推测)等内容中有所涉及，但都没有把它作为一个确定的领域对待.

本文把对于有用信息的 Web 挖掘作为一个确定的目标，从新的角度考虑问题，建立了一种因果模式来进行 Web 挖掘，从而能很好的发现 Web 上事物之间的内在联系以及发生规律，以便于预测. 最后给出了一个利用 XML 的 Web 数据挖掘模型及具体的应用实例，实验结果表明该模型是能够在预测和决策中发挥重要作用的.

1 XML与 Web数据挖掘

1.1 XML

XML 是 HTML 的扩展，是一种扩展性标记语言，它定义了数据的真正物理含义，是新一代网络数据表示、传输和交换的标准，是 Internet 环境中跨平台的、依赖于内容的技术[4–5]. 它具有如下的优点：(1)更有意义的搜索；(2)开发灵活的 Web 应用软件；(3)

收稿日期：2009–10–31；**修回日期**：2009–12–13
基金项目：国家自然科学基金资助项目(70571056)
作者简介：李孝忠（1962—），男，山东人，教授，lixz@tust.edu.cn.

不同来源的数据集成；(4)多种应用得到的数据；(5)本地计算和处理；(6)数据的多样显示；(7)粒状的更新；(8)在 Web 发布数据；(9)升级性；(10)压缩性.

1.2　Web数据挖掘

Web 数据挖掘是数据挖掘技术在 Web 环境下的应用，它将数据挖掘技术应用在 Web 上，从大量的 Web 文档集合和在站点内进行浏览的相关数据中发现蕴涵的、未知、有潜在应用价值的过程[6–7]. 它所处理的对象包括：静态网页、Web 数据库、Web 结构、用户使用记录等信息[8]. 通过对这些信息的挖掘，可以得到仅通过文字检索所不能得到的信息[9].

2　利用 XML 的 Web 数据挖掘流程

XML 出现后极大地方便了 Web 数据挖掘. 原始的通过爬行器搜索到的 Web 资源是 Html 格式的，只需简单的转换成 XML 后再通过 Java 编程解析，将 XML 数据变成结构化的数据存入数据库当中，就便于挖掘了. 在 Web 挖掘中加入因果模式之前需要了解传统的挖掘过程，其挖掘流程见图 1[10].

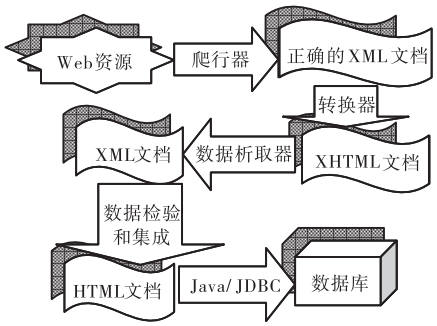

图 1　利用 XML 的 Web 数据挖掘流程

Fig.1　Web data mining process based on XML

3　利用 XML 的因果模式 Web 挖掘模型

网络中信息众多，如何从其中挖出有用的知识成为关键，本文将建立一种利用因果模式的挖掘模型，使得建立的挖掘模型能够帮助使用者进行一些分析与预测. 其基本思想是，先建立因果集(因集和果集)，然后分析因集中的各个子项(也就是各种原因)在果集的子项(也就是各种结果)中发挥的作用(用影响度这个参数来度量)，以此为依据帮助使用者进行一些判断和预测. 其中，因集和果集中需要分类，甚至子类还可以再往下分，并为每个子类分配索引和权值(分类效果权值)，通过实验，来验证分类效果，进而修改分类效果权值，最后将分类效果权值最大的推荐给使用者以作预测用.

由于因果关系是一个复杂的关系，其中可能遇见的几种情况如下：

(1)因素与结果一一对应；

(2)一种因素出现在多种结果中，或一个结果由多种因素所形成；

(3)有时候某个结果还可能成为其他结果的原因(或某个原因是其他原因的结果)，甚至以此类推，最终形成一种链状结构，也就是因果链；

(4)各种因果关系相互交织，形成了一种多维的网状结构.

由于以上原因，这就需要分门别类的为每一种情况(因集和果集每一个子类之间)建立索引、关系对照表或关系矩阵，包括分类效果权值、影响度等各类参数和备注. 这样，就形成了一个庞大的数据库——因果智能库. 而且，这个智能库可以通过反复训练和实验进行修改和完善，并最终能达到帮助使用者进行分析和预测的理想效果.

以上提到的一些概念解释和图例如下：

因集(Factor Set) = {(子类 1)，(子类 2)，(子类 3)，…}

果集(Result Set) = {(子类 1)，(子类 2)，(子类 3)，…}

$$\text{影响度}=\frac{\text{某结果发生时该因素出现的次数}}{\text{某结果所有发生的次数}}\times 100\%$$

分类效果权值：用来判断各种分类的效果，根据每次事件结果是否应验了该分类而增加或减少其权值，最后将权值较大的分类推荐给使用者，以帮助未来的分析和预测.

加入因果关系后，建立起来的利用 XML 的 Web 数据挖掘模型就分成了三大模块：

(1)使用者界面模块，负责接收用户的命令，修改意见和对结果的呈现.

(2)预处理模块，按照请求将所需的 Web 上的各种数据转换成便于挖掘的结构化数据.

(3)挖掘模块，根据因果模式规则和挖掘方法(可能用到的 Web 挖掘方法有统计分析、关联规则等方法)挖掘出想要的结果并返回给用户，同时也接受修改意见而进行自我完善.

挖掘过程是在图 1 的基础上加入了因果模式规则后而形成了一个整体框架，详细过程见图 2.

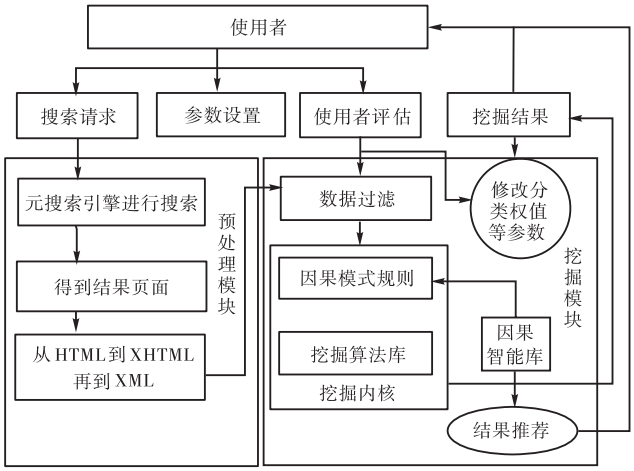

图 2　利用 XML 的一种因果模式 Web 挖掘模型

Fig.2　A causal pattern Web mining model based on XML

4　实　例

实例采用天津科技大学易佳影视网，Web 挖掘方法采用统计分析法.

将影视节目分为科幻、文艺、动作、战争、动画等因素（因集的组成就是{（科幻），（文艺），（动作），（战争），（动画）}），将结果分为收看人次在（0 ~ 10 000 次），（10 000 ~ 20 000 次），（20 000 次以上）三个类别（果集的组成就是{（0 ~ 10 000 次），（10 000 ~ 20 000 次），（20 000 次以上）}）. 经统计发现（每一类都抽取前 10 名作为代表），科幻类影视节目的收看人次多（21 435 次），然后在网站上多上传一些科幻类影视节目进行观测，用未来 10 天的收看人次的增加数量作为结果（增加了 1 298 次）进行权衡.

又将影视节目分为国产、欧美、日韩、港台等因素（这样因集就是{（国产），（欧美），（日韩），（港台）}），结果分类不变，统计发现欧美类影视节目的收看人次最多（19 163 次），于是就上传欧美影视节目进行未来 10 天的观测，发现增加的人次数（为 912 次）不如第一种分类的多.

根据两种分类增加人次数的比例分配权值，显然第一种分类的权值大，因此推荐给网站建设者：要按第一种分类上传影视节目，而且多上传科幻类的影视节目. 其中，为了达到更佳的效果，可以以每种分类中不同种类影视节目的比例分配图为依据，不断调整各类节目上传比例. 将不同分类中的因素进行组合，效果更佳. 比如：科幻类的欧美影视节目最受欢迎.

正如上面所提到的其中一种情况：一个结果可能由几个因素所促成，因此该模型最终结果还是令人满意的. 由此可见，分类是一个至关重要的步骤，分类方法对未来预测效果有直接影响. 本模型通过不同形式分类方法建立因果集，再结合统计分析等挖掘方法发现因果关系进而进行预测，比仅用关联规则等传统预测具有更加全面、高效的功能.

5　结　语

由于 Web 上的信息越来越庞大，能够从其中挖掘出有用的知识来满足用户需要将会越来越重要，也是必然的发展趋势. 本文通过建立一种因果模式，发现事物之间的一些联系，利用这些联系能够帮助使用者对未来情况进行分析和预测. 当然，本模型预测的准确性还有待进一步的完善.

参考文献：

[1] 范明，孟小峰. 数据挖掘概念与技术[M]. 北京：机械工业出版社，2007：1–2.

[2] 蒋望东，黄发良. 基于 Web 的数据挖掘研究综述[J]. 湖南工程学院学报：自然科学版，2007（1）：61–64.

[3] 朱德利. Web 结构挖掘的 XML 实现策略[J]. 计算机工程与设计，2006，27（23）：4447–4449.

[4] Chen Y，Tsai F S，Chan K L. Machine learning techniques for business blog search and mining[J]. Expert Systems with Applications，2008（35）：581–590.

[5] Facca F M，Lanzi P L. Mining interesting knowledge from weblogs：A survey[J]. Data & Knowledge Engineering，2005，53（3）：225–241.

[6] 蔡飞，贝佳，潘金贵. 一种简单高效的 XML 与关系数据库信息交换的方法[J]. 计算机科学，2004，31（12）：72–75.

[7] Han J W，Chang K C-C. Data mining for Web intelligence [J]. Computer，2002，35（11）：64–70.

[8] 赵葵. 基于 XML 与 Web 的数据挖掘[J]. 计算机与网络，2008，22（6）：170–171.

[9] 徐立宇. 利用 Web 的数据挖掘技术研究及其应用[J]. 电脑知识与技术，2009，5（8）：1804–1805. .

[10] 崔建群，何炎祥，郑世珏，等. 利用 XML 的 Web 数据挖掘关键技术的研究[J]. 计算机工程，2006，32（20）：43–77.