基于XML的WEB数据挖掘系统研究

周晓兰

（湖南科技大学 计算机科学与工程学院，湖南 湘潭 411201）

摘 要：在论述WEB数据挖掘技术理论后，详细探讨了WEB数据挖掘的内容、流程、任务。根据用户的行为建立用户的兴趣模型，在此基础上对中文网页进行推荐，使用户最感兴趣的信息显示在最前面。

关键词：WEB数据挖掘；XML；用户兴趣模型

中图分类号：TP311 **文献标识码**：A **文章编号**：1673-9833(2008)04-0037-04

Research on WEB Data Mining System Based on XML

Zhou Xiaolan

(School of Computer Science and Engineering, Hunan University of Science and Technology, Xiangtan Hunan 411201, China)

Abstract: After elaborating the data mining technology theory, its content, flow and the duty is analyzed. An interest model is established according to the user's behavior. On the base of recommendation of the Chinese WEB page, the information which the people found well and interesting will be put on the most front.

Key words: WEB of data mining; XML; user interest model

WEB数据挖掘具有半结构化的数据结构、异构数据库环境以及解决半结构化的数据源问题等特点[1]，而XML(eXtensible Markup Language)的出现为解决WEB数据挖掘难点提供了很好的解决方法[2]。XML解决了Internet发展速度快而接入速度慢的问题，以及可利用的信息多，但难以找到自己需要的那部分信息的问题。XML能增加结构和语义信息，使计算机和服务器即时处理多种形式的信息。XML没有固定的标记，但能描述数据的形式和结构，XML还将数据和显示分开，从而能方便地实现网络应用和信息共享[3]。XML给基于WEB的应用软件赋予了强大的功能和灵活性，给开发者和用户带来许多好处[4]。

本系统重点研究用户兴趣模型的建立和XML代码的内容挖掘。根据用户浏览网页所反馈的信息和用户浏览网页的动作来对用户的兴趣进行分析，建立用户兴趣模型；根据XML的特点，抽取网页的特征向量；计算XML文档的特征向量与用户兴趣向量的相似度，将网页文档按相似度的大小排序，把大于规定阀值的网页推荐给用户。

1 系统简介

为提高挖掘系统的整体性能和提供挖掘的集成环境，基于XML的WEB挖掘系统原型由3个逻辑层次构成，WEB挖掘系统的逻辑架构如图1所示。数据获取层是对半结构化的WEB数据进行模型抽取和转换，用结构化数据表示，建立多层次的WEB数据库，并对WEB服务器日志数据进行预处理形成WEB日志数据库。数据存储层把非XML的网页转换成XML的网页，并对网页信息库进行存储。数据挖掘层是系统功能实现的核心，提供各种数据挖掘算法和有效的解决方案以及结合用户的兴趣模型，最终为用户挖掘到所需要的信息，有效完成各种数据挖掘任务。

收稿日期：2008-06-19

作者简介：周晓兰（1974-），女，湖南洞口人，湖南科技大学教师，硕士，主要研究方向为计算机应用技术及信息管理.

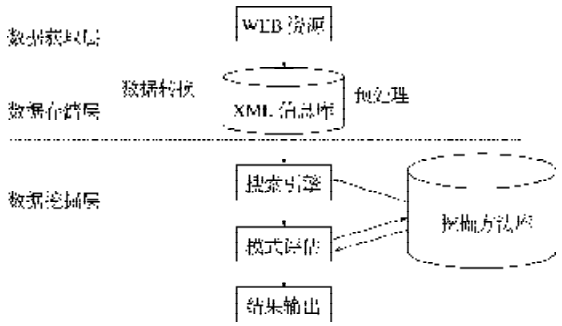

图 1 基于 XML 的 WEB 挖掘逻辑层次

Fig. 1 WEB data mining logic level based on XML

2 建立和更新用户兴趣模型

不同知识结构的用户对文档相关性的判断以及检索结果的要求是不同的，即使同一个用户，在不同时期也有所侧重[5]。用户兴趣建模是为用户提供个性化信息服务的信息检索或信息过滤系统的核心组成部分，能获取每个用户的不同信息需求。为了跟踪用户的兴趣与行为，要求检索系统为每个用户建立用户描述文件，用来保存用户兴趣。在检索过程中，结合用户兴趣对检索结果进行过滤，以实现检索结果的个性化。经实验、统计和调查可知：用户访问 WEB 的动机能反映出用户的兴趣；用户访问网页时的相关反馈信息和用户浏览网页的行为，也能反映用户的兴趣。

2.1 用户浏览网页的静态分析

静态分析主要是对 Cookies 和收藏夹进行分析。Cookies 是 WEB Server 在连接时保存在 WEB 客户端的信息，由浏览器建立并维护，用户不得自行更改其内容。通过读取用户电脑上 Cookies 的内容，把记录的登陆次数和记录的点击次数存入数据库中，以备建立模型时计算页面权值。

2.2 用户浏览网页的动作分析

用户的兴趣度与用户浏览的动作有密切关系，如查询、浏览页面、反馈信息、标记书签、点击鼠标、拖动滚动条、前进、后退等都能暗示用户的喜好；用户访问时的停留时间、访问次数、编辑、保存、修改、浏览时间和翻页 / 拉动滚动条次数等能揭示用户兴趣。用户的这些动作中只有 2 个关键因素可以反应用户对网页 P 的兴趣度 $d(P)$ ：网页 P 上的浏览时间 $T(P)$（简称为 T 行为）和翻页 / 拉动滚动条次数 $V(P)$（简称 V 行为）[6]。T 行为和 V 行为都可通过钩子函数求得。

用多元线性回归模型描述用户浏览网页的兴趣度 $d(P)$ 与两种行为之间的关系，其回归方程表示为 ：

$$d(P) = a*T(P) + b*V(P) + c + \alpha, \quad (1)$$

式（1）中，a、b、c 都是与 $T(P)$ 和 $V(P)$ 无关的未知参数，α 是随机误差，服从正态分布 $N(0, \sigma)$，公式（1）为多元正态线性回归模型。

$$d(P) = a*T(P) + b*V(P) + c, \quad (2)$$

式（2）中，a 和 b 称为回归系数，称 a、b、c 为行为影响因子（根据站点类型不同而取不同的值，通常为一组经验值），参数 a、b、c 的估计可采用最小二乘法。公式(2)为线性回归方程。

动态权值是指对 WEB 挖掘中用户浏览网页动作的挖掘，用前面计算出来的用户浏览网页的兴趣度来表示。计算动态权值公式为 ：

$$d(P) = a*T(P) + b*V(P) + c。 \quad (3)$$

2.3 建立和更新用户兴趣模型

本系统根据用户的浏览内容和浏览行为自动构建用户兴趣模型，构建方式有：1）通过用户浏览的网页，对搜索结果的反馈信息建立和更新用户兴趣；2）在用户没有明确参与的情况下，通过观察用户的浏览行为建立和更新用户兴趣。

2.3.1 用户兴趣模型的建立

用户初次进入系统时，要求用户在分类信息中选择其兴趣的大致所在，系统根据用户注册信息获得用户的初始兴趣，建立一个稍有针对性的用户兴趣初始模型。系统利用用户留在服务器上的信息（即日志文件），以及通过观察用户的行为来建立和更新用户兴趣模型[7]，并采用基于 XML 的资源描述框架（RDF，Resource Description Framework）来表达用户描述文件，利用数据库系统来存储用户模板文件。

1）用户兴趣向量表示

用户兴趣集 C 由用户所有的兴趣类别构成，表示为：$\{c_1, c_2, \cdots, c_m\}$，其中 $c_i\ (1 \leqslant i \leqslant m)$ 为用户感兴趣的兴趣类别名称，m 表示用户兴趣类别总数。兴趣类特征词集 $T(c_i)$ 由类 c_i 中的特征词构成，表示为：$\{t_1, t_2, \cdots, t_n\}$，其中 $t_i(1 \leqslant i \leqslant n)$ 表示特征词名称，k 为特征词总个数。在构造用户模型时，兴趣数 m 应该适当。

选取 1 组适合表示用户兴趣的特征词集（T_1，T_2，T_3，$\cdots$，T_n），根据关键词 T_i 在用户浏览过的网页文件中重要程度求出权值 W_i（i = 1，2，$\cdots$，n），然后把用户兴趣类向量用 1 个加权特征词向量表示。

特征词权值的计算过程如下：

Ⅰ）统计兴趣类中所有内容页面的数目 N；

Ⅱ）求出所有页面的特征词的并集 $K=\{K_1, K_2, \cdots, K_m\}$ 作为用户兴趣类向量的候选特征词；

Ⅲ）统计特征词 K_i 在多少篇文档中出现，将其数目记为 N_i；

Ⅳ）用 TF-IDF 方法计算用户兴趣各特征词权值。

研究发现，兴趣模型与 4 个变量（词语、词语的权值、文档、文档的建立时间）相关，可以得出时间衰减后词 t_i 的权值。权值 $W(t_i)$ 的计算公式为：

$$W(t_i) = \sum_{j=1}^{n} \frac{tf(t_i, d_j)}{age(d_j)}, \quad (4)$$

式中，$tf(t_i, d_j)$为词 t_i 在文档 d_j 中的词频；$age(d_j)$为第 j 个文档的年龄；N 为文档的个数。可得用户兴趣的个性化向量 $U(T_1, W_1; T_2, W_2; \cdots; T_n, W_n)$。

2）用户兴趣存储方式

为区分用户的不同兴趣类别，把用户的兴趣表示成用户兴趣树，用以保存用户的兴趣类型信息，也可保存用户兴趣特征词的信息。树中除虚拟根结点外，中间两层表示用户兴趣类别的结点称为兴趣结点，最底层的结点称为特征词结点。为适应用户兴趣的变化，用户的 2 棵兴趣树（稳定兴趣和偶然兴趣）可单独用来进行个性化分析，也可综合起来使用。

3）用户兴趣模型表示

用户兴趣模型由表示用户兴趣的以关键字为主体的一些对象组成，每个对象都有 1 个权值信息，权值越高，表明用户对这个关键字方面的信息越感兴趣。对象还可包括：该关键字对象的父对象信息、本对象信息以及扩展联系信息等。

用户兴趣树中的所有兴趣结点构成用户兴趣全集，可表示为{$Node(c_1)$，$Node(c_2)$，…，$Node(c_m)$}，记作 $U(C)$。其中，$c_i \in C$，m 为用户感兴趣的类别总数。

用户的偶然兴趣集记为 $U(C\text{short})$，稳定兴趣集记为 $U(C\text{long})$。用户的兴趣可以用户短期兴趣和长期兴趣来共同表示为：$U(C)=U(C\text{short})+U(C\text{long})$。其中：兴趣集 $C=C\text{short} \cup C\text{long}$。

兴趣类 c_i 的兴趣度：

$Node(c_i).v = x*Node(c_i\text{short}).v + y*Node(c_i\text{long}).v$。其中，$c_i \in C$，$c_i\text{short} \in c\text{short}$，$c_i\text{long} \in c\text{long}$，$\{x, y | x>0, y>0, x+y=1\}$。

建立模型以后，就要进行抽词（取出个性词典中的前 M 个高权值词条），一般选取 M=100 就足够了。

2.3.2　用户兴趣模型的更新

用户使用过程中，系统不断记录用户的使用情况，并分析记录的使用情况，不断修正用户兴趣模型。

用户兴趣模型的更新过程分三步：第一步是更新偶然兴趣，加入用户的最新兴趣和剔除最老的兴趣；第二步是偶然兴趣向稳定兴趣转化。将偶然兴趣中相对比较重要的特征词及兴趣度超过一定阈值的兴趣类转成稳定兴趣；第三步是更新稳定兴趣。随着时间的推移，逐渐淘汰用户不感兴趣的主题。这一过程一般安排在有效任务执行完后、系统空闲时进行。

通过优化，用户兴趣模型会更好地为系统的智能支持提供帮助。以用户兴趣词典中前 N 个词条在兴趣词典中的归一化权值为特征项，同样可以构造一个反映用户兴趣的 N 维特征向量 $V(V_1, d_1; V_2, d_2; \cdots; V_n, d_n)$，$d_i$ $(1 \leqslant i \leqslant n)$是词条在词典中的归一化权值。通过不断地学习，系统会根据一定的原则或用户的要求对分类信息进行修改和优化。

3　根据用户兴趣模型对中文网页进行自动推荐

对网页进行挖掘的目的是为用户检索到用户所需要的信息，能把用户感兴趣的信息推荐给用户。协同过滤(Col-borative Filtering)技术可以服务于这一目的，它通过参考与当前用户类似的兴趣和偏好，为当前用户提供访问的建议。为了更好地服务用户，根据建立好的用户兴趣模型，对网页进行推荐，使检索到的信息基本符合用户的需求。把用户的兴趣向量看作用户提交的标准文档，该文档能够最大程度地反映用户的兴趣爱好。其它需要过滤的文档与该文档进行比较，与标准文档越相似的文档其获得推荐的机会就越大。

3.1　抽取 XML 文档的特征向量

XML 网页与普通的 WEB 文档相比，有明显的标识符，结构信息特别明显，对象的属性更为丰富。首先来看如何从 XML 文档中提取特征项。

XML 的第一部分包含了文件中的声明文字部分、注释文字部分和属性文字部分。文字部分没有分层结构信息，一般特征向量由以上 3 个单元产生。在剔除完停用词和所有训练文件被读完之后，得到初始一般特征表，利用 TF−IDF[8]方法运算，得到其权值。

XML 的第二部分包括单元及属性 2 个重要的标记部分。随着 XML 特有的分层结构性，单元及属性标记部分将随着层级的不同而呈现不同的重要性。1 份 XML 文件每个单元及属性标签的位置都有它的层次性。计算不同层次所各自拥有的重要权重（weight）公式（5）：

$$W_{\text{level}}=N^{(5-i)}\ , \qquad (5)$$

式中 i=1，2，3，4（只计算 XML 文件前 4 层的内容），由根结点算起由上到下分别为第 1、2、3、4 层，W_{level} 为各个层次所拥有的权重值，N 是整数变量，可以根据单元及属性标记的内容的重要程度自行指定，这里设定 N=2 作为权重值的评分标准。

产生出一般特征词表后，就建立起层次特征词表。它不仅会记录所有该分类的训练文件中单元标签及属性标签的文字内容，而且也会填入各自所属的权重值，如果在处理过程中遇到相同的字符，将保留最大的权重值。该词表的特点是随着词出现的位置和层次不同，而具有不同的权重。

本系统参考新浪网页分类体系（一级分类，共 28 类），将这些分类提取出来作为知识特征词表。这个特征词表配合前面提到的 2 个词表，成为作为分类依据的第 3 个特征词表。当 3 个特征词集准备完成后，就进行分类计算。当读入 1 个 XML 文件时，会先将此文件拆成 2 部分。第 1 部分包含注解文字、单元文字、属性文字等纯文字部分，称为一般测试集。第 2 部分包

括单元标签与属性标签2个具有层次性的内容，称为分层测试集，这2个部分被分别存放为链表形式。经过计算，1个XML文档的特征向量可以用V（T_1，W_1；T_2，W_2；…；T_m，W_m）表示。

3.2 进行网页推荐

根据用户兴趣向量与网页文档向量的相关度可以知道网页与用户兴趣的相关程度。把目标文件U看作用户兴趣的特征向量文件，未知文件V看作网页的特征向量文件，未知文件V与目标文件U的相似度越高，未知文件就越符合用户兴趣的要求，也就与用户兴趣越接近，网页上的信息就越符合用户的需求。从而把用户兴趣特征向量和未知网页特征向量的匹配问题转化为向量空间中的向量匹配问题。这里采用相似度分类算法（计算待分类网页与各类别的相似度，选取相似度最大的类别作为待分类文档的类别）。

2个特征矢量的相似度$sim(d_k, c_i)$用2个特征矢量之间的夹角余弦来度量。夹角越小说明相似度越高，未知文件V与目标文件U就越相似，也就是此网页就越满足用户的兴趣。

设用户特征向量为U，未知网页的特征向量为V，使用向量距离分类法计算相似度(计算网页文档与用户兴趣词典相关度)的公式如式（6）。

$$sim(V,U)=\frac{\left(\sum_{k}^{m}(W_{rk}\times W_{sk})\right)^{2}\times\sum_{k=1}^{m}W_{rk}}{L\times\sqrt{\sum_{k=1}^{m}W_{rk}^{2}}}。\qquad(6)$$

对每一个网页文档构造完特征向量后，利用公式（6）计算其与用户兴趣向量的相关度，然后按照相关度从大到小排序，将相关度大的推荐给用户。

4 实验检验

对本系统的使用效能进行检验，采取比较分析法。首先统计不用本系统的情况下，湖南科技大学图书馆的借阅量；然后统计应用本系统后，湖南科技大学图书馆借阅量；进而进行比较研究。

湖南科技大学图书馆总藏书量约100万册。表1是不应用本系统的2个月（2006-09-01~2006-10-31）和应用本系统的2个月（2006-11-01~2006-12-31）的数据对比表。

从表格中可看出，应用本系统后图书馆的总借阅效率提高了。具体体现在以下2个方面：

1）图书馆的采编部根据服务器的数据对采购任务进行调整。对于借阅得比较多的图书采购的多些，对于借阅的极少的图书，就少采购。通过应用此系统，图书馆不再出现以前的怪现象：有些图书堆积如山也没有人看，有些图书总是不能满足读者的要求。

2）文科图书借阅率变化不大；计算机图书借阅率提高较快，特别是查询电子图书的读者，一般都应用了本系统，从而图书的借阅率大大提高了。

表1 应用系统前后2个月的纸质图书借阅数据对比表

Tab. 1 Contrast table of 2 month-long paper books borrowing data reference between pre-and post application system

借阅量	使用系统情况（册）		增加率
	使用前2个月	使用后2个月	
图书总借阅量	182 200	200 386	9.98 %
计算机图书借阅量	15 268	17 526	14.79 %
文科图书借阅量	54 462	56 507	3.75 %

5 结语

基于XML的WEB数据挖掘是一个具有前瞻性的研究课题。本文提出的基于XML的WEB挖掘系统的原型对指导实际的WEB挖掘系统的开发具有一定的参考价值，对WEB挖掘的理论研究也将起到一定的推动作用。研究了基于XML的WEB数据挖掘，实现了对WEB数据库的访问和异构数据库之间的结构化转换。重点讨论了如何建立用户兴趣模型以及如何根据用户兴趣查找用户所需要的信息。随着网络的发展，有关XML研究正处于不断发展中，新的研究领域和各种应用技术也不断出现，实现对XML整合的实际数据源或某领域的专用数据源进行挖掘，以获取有用的知识将成为未来Internet环境中主流的网络计算技术。

参考文献：

[1] 曼丽春，朱 宏. WEB数据挖掘研究与探讨[J]. 西南民族大学学报：自然科学版，2005，31(2)：305.

[2] 王玉珍. WEB数据挖掘技术与XML[J]. 信息技术，2005(10)：142-143.

[3] 雷筱珍. XML技术在数据库中的应用初探[J]. 电脑知识与技术，2006(5)：17-18.

[4] 刘晓鹏，邢长征. 基于WEB文本数据挖掘的研究[J]. 计算机与数字工程，2005，33(9)：76-78.

[5] 曾 春，邢春晓，周立柱. 个性化服务技术综述[J]. 电脑知识与技术，2006(2)：17-18.

[6] Han Jing，Zhang Hongjiang，Cai Qingsheng. Prediction for Visiting Path on WEB[J]. Journal of Software，2002(6)：1041-1043.

[7] 余 侠，朱 林. 根据用户反馈建立和更新数字图书馆用户兴趣模型[J]. 情报杂志，2004(11)：21-22.

[8] 罗 欣，夏德麟，晏蒲柳. 基于词频差异的特征选取及改进的TF-IDF公式[J]. 计算机应用，2005，25(9)：2031-2033.

（责任编辑：罗立宇）