基于 WEB挖掘的 SWMS模型的研究与实现

崔国庆, 乔佩利
(哈尔滨理工大学 计算机科学与技术学院,黑龙江 哈尔滨 150080)

摘 要: 万维网的出现使得计算机拥有了海量的资源,但也困扰着人们获取其中有用信息. Web挖掘的应用为解决此问题指明了方向. 对 Web挖掘的最新技术及发展方向进行了论述,包含 Web挖掘分类,Web挖掘的特征和两个具体的 Web挖掘算法,最后提出一个具体的 Web挖掘系统 SWMS模型.

关键词: Web挖掘; 万维网; 数据挖掘
中图分类号: TP393 **文献标识码**: A **文章编号**: 1007- 2683(2006)05- 0014- 05

The Model SWMS Research and Implementation on Web Mining

CUI Guo-qing, QIAO Pei-li
(Computer Science & Technology College, Harbin Univ. Sci. Tech., Harbin 150080, China)

Abstract: With the arising of the World Wide Web, a computer has the huge amount of information available online, but people are puzzled about how to get knowledge that they need in the huge amount of information from Internet nowadays. The application of Web mining is a way to solve the problem. This paper focuses on the latest technology and perspective of Web mining, including Web mining category, Web mining speciality and the concrete Web mining algorithm, at last, a concrete SWMS model is shown.

Key words: Web mining; world wide Web; data mining

1 引 言

万维网已经成为人们获取信息资源的最直接便捷的途径. 网络是博大的、多样的、动态的,它容纳的是大规模形形色色的相对短暂的各种数据. 人们需要从繁杂冗余的海量数据中快速及时地检索到自己感兴趣的信息. 由于数据挖掘技术的日益完善,数据挖掘技术应用于 Web便成为可能.

1 Web挖掘的分类

1.1 Web挖掘的定义

Web挖掘是数据挖掘在 Web上的应用,它是一项综合技术,不同的领域有不同的定义. 文[1]将 Web挖掘定义为:针对包括 Web页面内容、页面之间的结构、用户访问信息、电子商务信息等在内的各种 Web数据,应用数据挖掘方法以发现有用的知识来帮助人们从 WWW中提取知识,改进站点设计,更好地开展电子商务的过程. 本文将 Web挖掘定义为:Web挖掘是指在 WWW上挖掘潜在的、有用的模式以及隐藏信息的过程[1].

1.2 Web挖掘的分类

1.2.1 Web内容挖掘

Web内容挖掘是对 Web页面内容进行挖掘,从 Web文档的内容信息中抽取知识的过程. Web内容挖掘的重点是页面分类和聚类[2].

1.2.2 Web结构挖掘

Web结构挖掘是从 WWW的组织结构和链接关系中推导信息知识. 这方面工作的典型代表有

收稿日期: 2005 - 09 - 16
作者简介: 崔国庆 (1979 -),男,哈尔滨理工大学硕士研究生.

PageRank和 CLEVER.此外,在多层 Web数据仓库(MLDB)中也利用了页面的链接结构[3].

1.2.3 Web使用记录的挖掘

Web使用记录挖掘的主要目标是从 Web的访问记录中抽取感兴趣的模式.图 1给出了 Web挖掘的一般分类图.

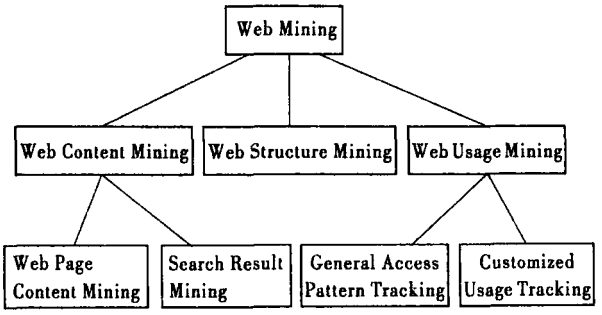

图 1　Web挖掘的分类

2　Web挖掘的特征

万维网目前是一个巨大的、分布广泛的和全球性的信息服务中心,因此,基于 Web的数据挖掘具有以下特征.

2.1　对有效的数据仓库和数据挖掘而言,Web过于庞大

Web的数据量目前以几百兆兆字节计算,而且仍然在迅速的增长.许多机构和社团都在把各自大量的可访问信息置于网上.

2.2　Web页面的复杂性远比任何传统的文本文档复杂得多

Web页面缺乏同一结构,它包含了远比任何一组书籍或其他文本文档多得多的风格和内容.Web可以看作一个巨大的数字图书馆,然而,这一图书馆中的大量文档并不是根据任何有关排列次序加以组织.它没有分类索引,更没有按标题、作者、封面页、目次等的索引.对在这样一个图书馆中搜索希望得到的信息是极具挑战性的.

2.3　Web是一个动态性极强的信息源

Web不仅以极快的速度增长,而且其信息还在不断地发生着更新.新闻、股票市场、公司广告和Web服务中心都在不断地更新着各自的页面.链接信息和访问记录也在频繁的更新之中.

2.4　Web面对的是一个广泛的形形色色的用户群体

目前因特网上连接有约 5 000万台工作站,其用户群仍在不断地扩展当中.各个用户可以有不同的背景、兴趣和使用目的.大部分用户并不了解信息网络结构,不清楚搜索的高昂代价,极容易在"黑暗"的网络中迷失方向,也极容易在"跳跃式"访问中烦乱不已或在等待一段信息时失去耐心[3].

2.5　Web上的信息只有很小部分是相关的或有用的

个人只是关心 Web上的很小的一部分信息,Web所包含的其余信息对用户来说是不感兴趣的,而且可能会淹没所希望得到的搜索结果.

3　数据挖掘算法

由于 Web挖掘主要可以分为 3类,对应不同的分类就有不同的算法,本文主要讨论 Web结构挖掘下的相应算法.

Web结构挖掘是对 Web的链接结构进行分析,以对超链接分析来评估基础 Web资源,从而发现有有用模式,提高搜索质量.目前 WSM (Web结构挖掘)用到的算法主要有 2种:一个是独立于查询的算法 PageRank,另一个是与查询相关的 HITS迭代算法,当然还有 SALSA等算法.它们的共同点是都利用特征向量作为理论基础和收敛性依据[4].

3.1　PageRank算法

PageRank算法的主要思想是来源于引用分析技术,就是如果一个页面被多次引用,则它可能很重要;一个页面可能没有被多次引用,但却被一个重要页面引用,则它可能很重要.一个页面的重要性被均分并被传递到它所引用的页面.PageRank算法的定义如下:如 $T_1, \cdots, T_n$为指向页面 A的页面;d为 0～1之间的阻尼系数;$C(A)$为 A所引用的页面数,则页面 A的 PageRank为

$$PR(A) = (1-d) + d(PR(T1)/C(T_1) + \cdots + PR(T_n)/C(T_n))$$

PageRank算法是独立于内容的静态链接算法,离线计算,开销较小,其值是通过遍历它搜索到的所有网页得到的.查询时,通过把基于内容的 IR合并起来给出排名顺序.

3.2　HITS算法

Web结构挖掘的另一个重要算法便是著名的 HITS算法.HITS(Hyper-link Induced Topic Search)的具体内容为:将查询 q提交给普通的基于相似度的搜索引擎,搜索引擎返回很多页面,从中取前 n个页面作为根集(root set),用 S表示.通过向 S中加入被 S引用的页面和引用 S的页面将 S扩展成一个更大的集合 T.以 T中的 hub页为顶点集 V_1,以

Authority页为顶点集 V_2, V_1中的页面到 V_2中的页面的超链接为边集 E, 形成一个二分有向图 SG = (V_1, V_2, E). 对 V_1中的任一个顶点 v, 用 $h(v)$表示页面 v的 hub值, 对 V_2中的顶点 u用 $a(u)$表示页面 u的 Authority值. 开始时, $a(u) = h(v) = 1$, 对 u执行 I操作,修改它的 $a(u)$. 对 v执行 O操作,修改它的 $h(v)$.

I操作: $$a(u) = \sum_{V:(V,U)\in E} h(v) \quad (1)$$

O操作: $$h(v) = \sum_{V:(V,U)\in E} a(u) \quad (2)$$

每次迭代后,对 $a(u)$和 $h(v)$进行规范化处理:

$$a(u) = \frac{a(u)}{\sqrt{\sum_{q\in V_2}[a(q)]^2}}$$

$$h(v) = \frac{h(v)}{\sqrt{\sum_{q\in V_1}[h(q)]^2}}$$

式 (1)反映了若一个页面由很多好的 hub所指,则其 Authority权重会相应增加 (即权重增加为所有指向它的页面的现有 hub权重之和). 式 (2) 反映了若一个页面指向许多好的权威页,则 hub权重也会相应增加 (即权重增加为该页面链接的所有页面的 Authority权重之和).

HITS算法输出一组具有较大 hub权重的页面和具有较大 Authority权重的页面. 许多实验表明,该算法对许多查询具有非常良好的搜索结果.

在实际应用中, 由 S生成 T的代价可能是很昂贵的. 为了找出 S所引用的页面,需要将 S中的所有页面下载,为了找出引用 S的页面,对 S中的页面需要搜索引擎能根据给出的 URL能找出引用该 URL的页面,另外需要排除重复的页面. 一般情况下 $\beta S\beta = 200$, 而 $\beta T\beta$可能达到 5 000. 因此根据 T生成有向图 SG可能是昂贵的[5].

虽然基于链接的算法可以带来很好的结果,但这种方法由于忽略文本内容,也遇到一些困难. 例如,当 hub页包含多个话题的内容时, HITS有时会发生偏差. 这一问题可以按如下的方法加以克服,即将式 (1) 和式 (2) 置换为相应权重的和, 降低同一站点内多链接的权重, 使用 Anchor文本 (Web页面中与超链接相连的文字)调整参与 Authority计算的链接的权重,将大的 hub页面分裂为小的单元.

基于 HITS的系统包括 CLEVER. Google也基于同样原理.

4 SWMS系统模型

目前存在的挖掘工具一般都针对一种 Web对象, 人们往往希望一种数据挖掘工具能挖掘多种 Web数据. 并且,随着结构化标记语言 XML越来越流行,并被人们接受和采纳. 可以预计,未来将会有大量的 Web页面用 XML书写. 而目前的挖掘工具都是面向 HTML的. 基于这种考虑和以上对 Web挖掘的研究,在这里提出了一个综合内容挖掘和结构挖掘功能的 Web挖掘系统 SWMS,如图 2所示.

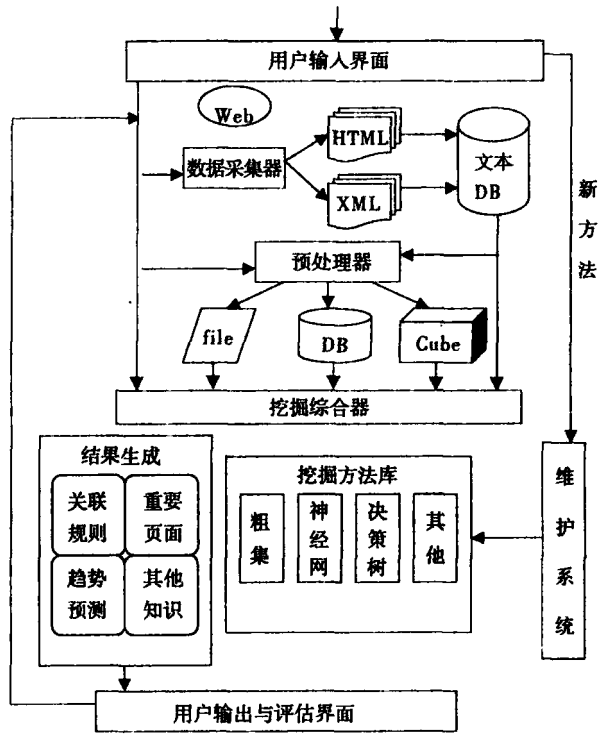

图 2 SWMS 系统模型

4.1 数据采集器

用来按用户要求从网上采集数据并将其存入文本数据库中. 数据采集器类似一个搜索引擎,但它比现有的搜索引擎功能要强大的多. 将数据采集器建立在搜索引擎之上. 首先从用户要求中提取出关键词,然后向 YAHOO等搜索引擎发出查询请求, 对查询结果进行过滤. 可以多种方式提出要求,如用户可以给出目标网页的一个例子, 据采集器通过提取它的特征作为关键词. 当然用户也可以直接给出查询主题词.

4.2 预处理器

SWMS仍然使用传统的挖掘工具,所以将文本

数据库中的 HTML文档和 XML文档组织成传统的挖掘工具可用的数据格式至关重要.预处理器正是要对文本数据库中的数据进行分类、提取和组织等操作,按挖掘要求分别生成数据立方、数据库或数据文件等数据形式.

4.3　挖掘综合器

挖掘综合器是一个挖掘驱动引擎.它根据挖掘要求和挖掘方法的选择策略到挖掘方法库中去选择合适的挖掘算法,并使用该方法去执行挖掘任务.不同的挖掘算法有不同的适用情况,如遗传算法较适于优化问题,而粗集理论适于处理模糊数据等.

4.4　挖掘方法库和维护系统

挖掘方法库存放各种挖掘方法,如粗集、决策树、神经网等.各种挖掘算法高度模块化,以便很容易地加入新方法,升级整个系统.维护系统正是提供给用户的一个增加新方法的接口.

4.5　用户输出与评估界面

将挖掘结果提交给用户,若用户满意,则挖掘过程结束,否则可以重提挖掘要求,然后重新进行挖掘.作为一个系统,各个元素是相互联系协同工作的.用户首先通过用户输入界面输入自己的挖掘要求,数据采集器根据用户要求搜集网页并存入文本数据库中,文本数据库存储网页的内容和地址等信息.然后预处理器将文本数据库中的内容取出,按照挖掘需要将数据组织成文件、数据库或是数据仓库供挖掘综合器使用.由于结构挖掘可以在网页上进行,故挖掘综合器可以直接对文本数据库操作.取得数据后,数据挖掘综合器仍然根据挖掘要求从挖掘方法库中选择适当的方法进行挖掘,将挖掘的结果通过用户输出与评估界面呈现给用户.用户根据自己的满意程度,结束任务或调整挖掘要求并进入新一轮挖掘.考虑到不断地有新的挖掘方法出现,为了用户对系统升级方便,提供给用户一个维护接口,用户可以把新的方法加入到挖掘方法库中.

5　实验结果及分析

以 http://www.speednet.net.cn从 2004年 6月到 2005年 6月的数据作为实验数据,该站点包括若干个主题(如:产品信息、营销活动信息、企业概貌、公司新闻、公司资质、招聘信息、解决方案及案例),以一个滑动窗口大小为 2的用户会话中的推荐机的使用为例,得到的推荐结果如表 1所示(CREC和 SREC分别为由本文算法得到的内容推荐值和结构推荐值,"/"表示值小于 0.5,未列出).

表 1　由本文模型得出的推荐结果

当前会话	推荐页面	CREC	SREC
公司概述	企业文化	0.58	/
	企业资质	0.56	/
	荣誉证书	0.53	/
	招贤纳士	0.50	/
服务器/集群产品信息	小型机产品信息	0.54	/
	交换机产品信息	0.53	/
	解决方案及案例	/	0.67
	企业资质	/	0.52
解决方案及案例	企业资质	/	0.54
	企业文化	0.53	/
	荣誉证书	0.51	/

可以看出,结合内容挖掘和结构挖掘的个性化推荐可以为用户提供附加的结果.如在内容挖掘推荐中,会话"服务器/集群产品信息"的推荐结果为一些关于其他类别网络产品的信息,同时运用结构挖掘推荐,则可以推荐出一些客户感兴趣其他内容,如解决方案及案例、企业资质等.同样,在基于结构挖掘的推荐结果中,会话"公司概述"没有能够产生任何大于阈值 0.5的推荐结果,而在基于内容挖掘的推荐中,则产生了一些相应的推荐页面.

6　结　语

本文从 Web挖掘的内容、体系结构及相关的各种技术等方面做了详细论述,并给出了 Web结构挖掘下的 2个经典算法和一个具体模型.实践证明,此模型对于提高挖掘速度和效率具有一定的应用价值.尽管 Web挖掘发展迅速并取得了不错的进展,但毕竟是一个新兴领域,还有许多问题有待于进一步的研究.

参 考 文 献:

[1] HAN J, MICHELINE K. 数据挖掘概念与技术[M]. 北京:机械工业出版社, 2004.
[2] 李亚飞,刘业政. Web挖掘的体系研究[J]. 合肥工业大学学报, 2004, 27(3): 306 - 308.
[3] 韩家炜. Web挖掘研究[J]. 计算机研究与发展, 2001, 38(4): 407 - 408.
[4] RAYMOND K, HENDRIK B. Web Mining Research: A Survey[J]. ACM SIGKDD, 2005, 2(1): 6 - 8.
[5] AJITH A. Bussiness Intelligence from Web Usage Mining[J]. IEEE Press, 2003, 11(1): 94 - 107.

(编辑:付长缨)