计算机科学 2006Vol.33№.12(增刊)

Web挖掘技术研究

王彤彤　强龙江　王　航
(解放军信息工程大学　郑州 450002)

摘　要　信息挖掘技术是当代计算机领域的热门话题，Internet/Web技术的快速普及和迅猛发展，产生出了海量的信息，如何在这个全球最大的数据集合中发现有用信息成为数据挖掘研究的热点。作为从浩瀚的Web信息资源中发现潜在的、有价值知识的一种有效技术，Web挖掘正悄然兴起，备受关注。目前，Web挖掘的研究正处于发展阶段，尚无统一的结论，本文结合当前Web挖掘的状况，介绍了一个基于Agent的Web挖掘模型，重点分析了Web挖掘的方法，概要介绍了Web数据挖掘在三个研究领域的研究现状及发展。

关键词　数据挖掘，Web挖掘

1　引言

数据挖掘是从数据库中挖掘出有用的、潜在的信息和知识，有一组操作组成，它被认为是解决知识爆炸，数据丰富但信息缺乏的一种有效方法。通常运用选定的知识发现算法，为被挖掘的数据提供摘要，形成预报和分类模型。近年来，随着Internet/Web技术的快速普及和迅猛发展，产生了大量的信息，在这些大量、异质的Web信息资源中，蕴含着具有巨大潜在价值的知识。人们迫切需要能够从Web上快速、有效地发现资源和知识的工具。Web挖掘作为数据挖掘的一个新主题，引起了人们的极大兴趣。Web挖掘就是利用数据挖掘技术，自动地从网络文档以及服务中发现和抽取信息的过程。信息挖掘有别于传统的信息检索，能够在异构数据组成的信息库中，从概念及相关因素的延伸比较上找出用户需要的深层次的信息。但是，数据挖掘的绝大部分工作所涉及的是结构化数据库，很少有处理Web上的异质、非结构化信息的工作。

对Web挖掘的定义[9]：Web挖掘是指从大量Web文档的集合C中发现隐含的模式p。如果将C看作输入，将p看作输出，那么Web挖掘的过程就是从输入到输出的一个映射N：C→p。

2　Web挖掘流程

数据挖掘可粗略地理解为三部曲：数据准备、数据挖掘，以及结果的解释评估。同样，Web挖掘也要经历这三个步骤，但是，由于与传统数据和数据仓库相比，Web上的信息是非结构化或半结构化的、动态的，并且是容易造成混淆的，因此很难直接以Web网页上的数据进行数据挖掘，而必须经过必要的数据预处理。典型Web挖掘的处理流程如下：

(1)查找资源：任务是从目标Web文档中得到数据，值得注意的是有时信息资源不仅限于在线Web文档，还包括电子邮件、电子文档、新闻组，或者网站的日志数据甚至是通过Web形成的交易数据库中的数据。

(2)信息选择和预处理：任务是从取得的Web资源中剔除无用信息和将信息进行必要的整理。例如从Web文档中自动去除广告连接、去除多余格式标记、自动识别段落或者字段并将数据组织成规整的逻辑形式甚至是关系表。

(3)模式发现：自动进行模式发现。可以在同一个站点内部或在多个站点之间进行。

(4)模式分析：验证、解释上一步骤产生的模式。可以是机器自动完成，也可以是与分析人员进行交互来完成。

Web挖掘作为一个完整的技术体系，在进行挖掘之前的信息获得(Information Retrieval)和信息抽取(Information Extraction)相当重要。信息获得的目的在于找到相关Web文档，它只是把文档中的数据看成未经排序的词组的集合，而信息抽取的目的在于从文档中找到需要的数据项目，它对文档的结构合表达的含义感兴趣，它得一个重要任务就是对数据进行组织整理并适当建立索引。

由于Web数据量非常大，而且可能动态变化，用原来手工方式进行信息收集早已经力不从心，目前的研究方向是用自动化、半自动化的方法在Web上进行IR和IE。在Web环境下既要处理非结构化文档，又要处理半结构化的数据，最近几年在这两方面都有相应的研究成果和具体应用，特别是在大型搜索引擎中得到了很好的应用。

3　Web挖掘分类及各自的研究现状及发展

根据对Web数据的感兴趣程度不同，Web挖掘一般可以分为三类：Web内容挖掘(Web Content

mining)、Web 结构挖掘(Web structure mining)、Web 使用挖掘(Web usage Mining)。

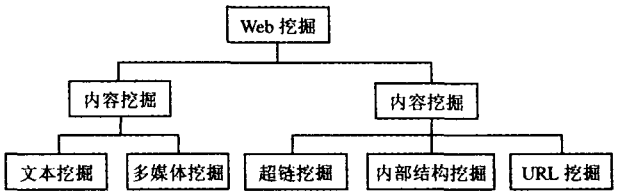

图 1 Web 挖掘的分类

3.1.1 Web 内容挖掘

即从网络的内容/数据/文档中发现有用信息的过程。网络信息资源类型众多,从网络信息源的角度看,大量的网络信息资源可以直接从网上抓取、建立索引、实现检索服务,但是还有一些网络信息是"隐藏"的,如由用户的提问而动态生成的结果,或是存在 DBMS 中的数据,或是那些私人数据,它们无法被索引,从而无法提供对它们有效的检索方式;从资源形式看,网络信息内容是由文本、图像、音频、视频、元数据等形式的数据组成的,因此网络内容挖掘是一种多媒体数据挖掘形式。

Web 内容挖掘的对象包括文本、图像、音频、视频、多媒体和其他各种类型的数据。其中针对无结构化文本进行的 Web 挖掘被归类到基于文本的知识发现(KDT)领域,也称文本数据挖掘或文本挖掘,是 Web 挖掘中比较重要的技术领域,也引起了许多研究者的关注。最近在 Web 多媒体数据挖掘方面的研究成为另一个热点。

Web 内容挖掘一般从两个不同的观点来进行研究。从资源查找(IR)的观点来看,Web 内容挖掘的任务是从用户的角度出发,怎样提高信息质量和帮助用户过滤信息。而从 DB 的角度讲 Web 内容挖掘的任务主要是试图对 Web 上的数据进行集成、建模,以支持对 Web 数据的复杂查询。

3.1.2 Web 结构挖掘

即挖掘 Web 潜在的链接结构模式。这种思想源于引文分析,即通过分析一个网页链接和被链接数量以及对象来建立 Web 自身的链接结构模式。可以用于网页归类,并且可以由此获得有关不同网页间相似度及关联度的信息,有助于用户找到相关主题的权威站点。

Web 结构挖掘的对象是 Web 本身的超连接,即对 Web 文档的结构进行挖掘。对于给定的 Web 文档集合,应该能够通过算法发现它们之间连接情况的有用信息,文档之间的超连接反映了文档之间的包含、引用或者从属关系,引用文档对被引用文档的说明往往更客观、更概括、更准确。

Web 结构挖掘在一定程度上得益于社会网络和引用分析的研究。把网页之间的关系分为 incoming 连接和 outgoing 连接,运用引用分析方法找到同一网站内部以及不同网站之间的连接关系。在 Web 结构挖掘领域最著名的算法是 HITS 算法和 PageRank 算法。它们的共同点是使用一定方法计算 Web 页面之间超连接的质量,从而得到页面的权重。著名的 Clever 和 Google 搜索引擎就采用了该类算法。

3.1.3 Web 使用挖掘

即 Web 使用记录挖掘,在新兴的电子商务领域有重要意义,它通过挖掘相关的 Web 日志记录,来发现用户访问 Web 页面的模式,通过分析日志记录中的规律,可以识别用户的忠实度、喜好、满意度,可以发现潜在用户,增强站点的服务竞争力。Web 使用记录数据除了服务器的日志记录外还包括代理服务器日志、浏览器端日志、注册信息、用户会话信息、交易信息、Cookie 中的信息、用户查询、鼠标点击流等一切用户与站点之间可能的交互记录。可见 Web 使用记录的数据量是非常巨大的,而且数据类型也相当丰富。根据对数据源的不同处理方法,Web 用法挖掘可以分为两类,一类是将 Web 使用记录的数据转换并传递进传统的关系表里,再使用数据挖掘算法对关系表中的数据进行常规挖掘;另一类是将 Web 使用记录的数据直接预处理再进行挖掘。Web 用法挖掘中的一个有趣的问题是在多个用户使用同一个代理服务器的环境下如何标识某个用户,如何识别属于该用户的会话和使用记录,这个问题看起来不大,但却在很大程度上影响着挖掘质量,所以有人专门在这方面进行了研究。通常来讲,经典的数据挖掘算法都可以直接用到 Web 用法挖掘上来,但为了提高挖掘质量,研究人员在扩展算法上进行了努力,包括复合关联规则算法、改进的序列发现算法等。

上述三个网络信息挖掘类型的比较见表 1。

根据数据来源、数据类型、数据集合中的用户数量、数据集合中的服务器数量等将 Web 用法挖掘分为五类。

(1)个性挖掘:针对单个用户的使用记录对该用户进行建模,结合该用户基本信息分析他的使用习惯、个人喜好,目的是在电子商务环境下为该用户提供与众不同的个性化服务。

(2)系统改进:Web 服务(数据库、网络等)的性能和其他服务质量是衡量用户满意度的关键指标,Web 用法挖掘可以通过用户的拥塞记录发现站点的性能瓶颈,以提示站点管理者改进 Web 缓存策略、网络传输策略、流量负载平衡机制和数据的分布策略。此外,可以通过分析网络的非法入侵数据找到系统弱点,提高站点安全性,这在电子商务环境下尤为重要。

(3)站点修改:站点的结构和内容是吸引用户的关键。Web 用法挖掘通过挖掘用户的行为记录和反馈情况为站点设计者提供改进的依,比如页面连接情况应如何组织、那些页面应能够直接访问等。

(4)智能商务:用户怎样使用 Web 站点的信息无疑是电子商务销售商关心的重点,用户一次访问的周期可分为被吸引、驻留、购买和离开四个步骤,Web 用法挖掘可以通过分析用户点击流等 Web 日志信息挖掘用户行为的动机,以帮助销售商合理安排销售策略。

(5)Web 特征描述:这类研究跟关注这样通过用户对站点的访问情况统计各个用户在页面上的交互情况,对用户访问情况进行特征描述。

4 一个 Web 文本挖掘原型 Web Miner

下面介绍一个 Web 文本挖掘系统原型 Web Miner(如图 2 所示)。Web Miner 采用了多 agent 的体系结构,将多维文本分析与文本挖掘这两种技术有机地结合起来,以帮助用户快速、有效地挖掘 Web 上的 HTML 文档。

表 1 三种 Web 挖掘方法的比较

	网络信息挖掘			
	网络内容挖掘		网络结构挖掘	网络用法挖掘
	信息检索观点	数据库观点		
数据形式	非结构化、半结构化	半结构化、数据库形式的网站	链接结构	交互形式
主要数据	文本文档、超文本文档	超文本文档	链接结构	服务器日志记录 浏览器日志记录
表示	Bag of words、n-grams、词、短语、概念或实体、关系型数据	边界标志图(OEM)、关系型数据	图形	关系型表、图形
方法	TFIDF 和变体、机器学习、统计学(包括自然语言处理)	Proprietary 算法、ILP、(修改后)的关联规则	Proprietary 算法	机器学习、统计学、(修改后)的关联规则
应用	归类、聚类、发掘抽取规则、发掘文本模式、建立模式	发掘高频的子结构、发掘网站体系结构	归类、聚类	站点建设、改进与管理、营销、建立用户模式

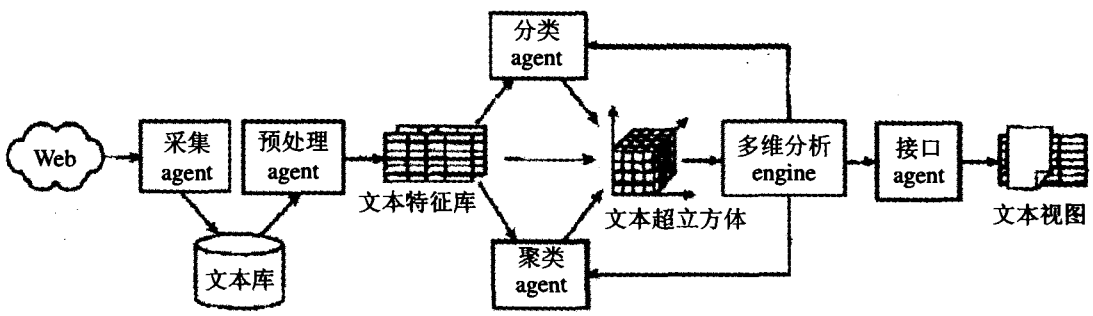

图 2 Web 文本挖掘系统原型 Web Miner

Web Miner 的系统组件

(1)文本搜集 agent:利用信息访问技术将分布在多个 Web 服务器上的待挖掘文档集成在 Web Miner 的本地文本库中。

(2)文本预处理 agent:利用启发式规则和自然语言处理技术从文本中抽取出代表其特征的元数据,并存放在文本特征库中,作为文本挖掘的基础。

(3)文本分类 agent:利用其内部知识库,按照预定义的类别层次,对文档集合(或者其中的部分子集)的内容进行分类。

(4)文本聚类 agent:利用其内部知识库,对文档集合(或者其中的部分子集)的内容进行聚类。

(5)多维文本分析引擎:Web Miner 引入了文本超立方体模型和多维文本分析技术,为用户提供关于文档的多维视图。多维文本分析引擎还具有统计分析功能,从而能够揭示文档集合的特征分布和趋势。此外,多维文本分析引擎还可以对大量文档的集合进行特征修剪,包括横向文档选择和纵向特征投影两种方式。

(6)用户接口 agent:在用户与多维文本分析引擎之间起着桥梁作用。它为用户提供可视化接口,将用户的请求转化为专用语言传递给多维文本分析引擎,并将多维文本分析引擎返回的多维文本视图和文档展示给用户。

每个 agent 作为系统的一个组件,能够完成相对独立的工作。这些部件可以位于同一台计算机上,也可以分布在网络中的多台计算机上。此外,由于系统高度模块化,因此易于加入新的部件。同

(下转第 145 页)

语义网络给 IE，特别是基于本体的信息抽取(Ontology-Based Information Extraction，OBIE)带来了全新的挑战。传统 IE 和 OBIE 最重要的不同就是后者使用了形式的本体而不是平面词典或专有名词字典结构，并且 OBIE 不仅是找到特定类型的待抽取实体，同时能通过在本体中把它链接到其语义描述上来识别它，使得实体可以在文档内被跟踪，并在 IE 过程中丰富其描述。尽管存在本体统一标准的问题，但在类似于互联网的"求同存异"的思想下，OBIE 技术必将得到更大的拓展空间，其发展也将对 IE 的语义资源建设及篇章分析与推理等深层次的问题的解决产生促动。另外，研究将多文档文摘和信息抽取相结合，从相似的文本中自动获取模板，将模板合并生成多文档文摘，再将 IE 与多文档文摘的优点结合起来应用于开放域信息检索，也具有很大的应用前景。早些时文[17]也做过类似的研究即把多文档文摘技术看作 IR 的后处理，IE 的技术应用[17]。

结束语 本文旨在分析比较并介绍信息抽取与信息检索技术各自发展历史及相关重要问题。在如此短小的篇幅下去穷尽这两个领域当前的所有问题，几乎是不可能的；仅希望通过这一基础性调研工作对今后的研究有所启发帮助。文中论述存在偏颇之处，恳请批评指正。

致谢 衷心感谢中科院计算机语言信息工程研究中心冯冲博士提供的宝贵意见与资料！

参 考 文 献

1 王晓龙，关毅. 自然语言处理[M]. 北京：清华大学出版社，2005. 136～137
2 TIPSTER Text Program. [EB]http://www. itl. nist. gov. iaui. 894. 02/related_projects/tipster
3 Cunningham H，Bontcheva K. Named Entity Recognition (presentation). [Z] RANLP (European NLP Conf. on recent Advances in Natural Language Processing). Borovets，Bulgaria，Sep. 2003
4 孙斌. 信息提取技术概述. 语言信息处理，2002
5 李保利，陈玉忠，俞士汶. 信息抽取研究综述. 计算机工程与应用，2003. 10
6 The ACE02 Evaluation Plan. [EB] ftp://jaguar. ncsl. nist. gov/ace/doc/ACE-EvalPlan- 2002. pdf
7 NIST 网站[EB] www. nist. gov/speech/tests/ace
8 TREC 网站[EB] http://trec. nist. gov/
9 863 评测网站[EB] http://www. 863data. org. cn
10 CLEF 网站[EB] http://www. clef-campaign. org/
11 NII 网站[EB] http://research. nii. ac. jp/ntcir/index-en. html
12 Gaizauskas R，Wilks Y. Information Extraction：Beyond Document Retrieval. [J] Computa- tional Linguistics and Chinese Language Processing，1998，3 (2)：17～60
13 Lee L. Similarity-Based Approaches to Natural Language Processing：[Ph. D. thesis]. Harvard University Technical Report TR-11-97
14 Yates R B，Neto B R. Modern Information Retrieval. [M]. Addison-Wesley，New York，1999
15 Maynard D. Multi-Source and Multilingual Information Extraction. ppt. [J]. September 2003
16 Bontcheva K，Maynard D，Tablan V，et al. [J/OL] GATE：A Unicode-based Infrastructure Supporting Multilingual Information Extraction，2003
17 White M，Korelsky T，Cardie C，et al. Multidocument Summarization via Information Extraction [A]. In：Proceedings of the First International Conference on Human Language Technology Research [C]. 1998. 36～44

(上接第 132 页)

时，各个 agent 之间通过相互协作来完成挖掘的全过程。其中，多维文本分析引擎以文本预处理为基础，以文本挖掘为支撑。文本超立方体中的维来自于文本预处理所得到的文本特征属性，例如时间、作者等。而文档主题类别的生成以及文档之间关系的聚类分析又依赖于文档挖掘技术。反过来，多维文本分析引擎又为文本挖掘提供了有效的可视化手段和特征修剪工具。文档集合的特征修剪结果可以展现给用户，也可以作为挖掘对象输入到文本分类 agent 和文本聚类 agent，如图 2 所示。

结束语 在 Web 信息充斥的情况下，Web 挖掘是一个具有极大潜力的研究方向。一些国际会议，例如 KDD97、IJCA I'99 等，已经或即将举行有关 Web 挖掘的专题讨论，对其理论、体系结构、算法等展开研究，Web 挖掘的一个重要应用方向将是电子商务系统。而与电子商务关系最为密切的是用法挖掘(Usage Mining)，也就是说在这个领域将会持续得到更多的重视。另外，在搜索引擎的研究方面，结构挖掘的研究已经相对成熟，基于文本的内容挖掘也已经有许多研究，下一步将会有更多的研究者把多媒体挖掘最为研究方向。基于 Agent 的研究也大量的出现，总之，Web 挖掘已经成为数挖掘和 Internet 结合的一个重要方向，吸引着越来越多的研究者投入到这种新的技术中去。

参 考 文 献

1 Kosla R，Blockeel H. Web mining research a survey. SIGKDD Explorations，2000，2：1～15
2 Pal S K，Talwar V，Mitra P. Web Mining in SoftComputing Framework：Relevance，State of the Art and Future Directions. IEEE Transactions on Neural Networks，2002，13(5)：1163～1177
3 Madria S K，Bhowmick S S. Research issues in web data mining. In：The 1st International Conference of Data Warehousing and Knowledge Discovery，1999. 303～312
4 Srivastava J，Cooley R. Web usage mining：discovery and applications of usage patterns from Web data. SIGKDD Explorations，2000，1(2)：12～23
5 陈莉，焦李成. Internet/Web 数据挖掘研究现状及最新进展. 西安电子科技大学学报，2001，28(1)
6 王继成，潘金贵，张福炎. Web 文本挖掘技术研究. 计算机研究与发展，2000，37(15)