基于知识库的 Web 文本挖掘模型 K-WebMiner

K-WebMiner: A Web Text Mining Model Based on Knowledge Base

胡安安/HU An-an,陈　晋/CHEN Jin

复旦大学管理学院,上海　200433

School of Management, Fudan University, Shanghai 200433, China

[摘要]　Web 文本挖掘在人们的日常生活和决策分析过程中起到了越来越重要的作用。介绍了 Web 挖掘的概念和基本特征,在此基础上重点研究了 Web 文本挖掘方法,引出了 Web 文本挖掘的模型 WebMiner。结合知识库概念,尝试对 WebMiner 模型进行改进,提出了基于知识库的 Web 文本挖掘模型 K-WebMiner,希望可以提高 Web 内容挖掘的效果。

[关键词]　决策分析;K-WebMiner 模型;Web 文本挖掘;WebMiner 模型　　[中图分类号]　N945.12

[文献标识码]　A　　[文章编号]　1000-7857(2006)04-0068-04

Abstract: The role of Web Text Mining technology is becoming more and more important in people's daily life and decision analysis process. We briefly introduce the concept and the features of Web Mining and pay more attention on Web Content Mining technology and Web Text Mining model-Webminer. We put forward the K-WebMiner, an improved Web Text Mining model based on knowledge base, and hope it will ameliorate the effect of Web Content Mining technology.

Key Words: decision analysis; K-WebMiner model; Web Text Mining technology; WebMiner model

CLC Number: N945.12　　**Document Code:** A　　**Article ID:** 1000-7857(2006)04-0068-04

1 概述

人类社会进入 20 世纪 90 年代后,日常积累的数据量以每月大于 15%的速度在持续增长,但数据的海洋并不能帮助人类产生决策意志。为了能更好地进行决策分析,人们不断地增强数据库的处理能力,并提出了数据挖掘技术,希望能从庞大的数据库中发现知识,用于决策。

伴随着 Internet 的飞速发展,越来越多的机构、团体、个人在互联网上发布信息、查找信息。虽然网上有大量的数据、信息,但其缺点也是明显的,由于 Web 是无结构的、动态的,并且网页的复杂程度远远超过了传统的文本文件,人们想找到自己需要的数据犹如大海捞针一般困难。为了能更便捷地在互联网上搜寻到有用的信息,人们开始尝试将数据挖掘技术用于 Web 检索中,提出了 Web 挖掘技术。本文从分析 Web 挖掘的特征入手,介绍 Web 文本挖掘技术和 WebMiner 模型,并引出基于知识库的 Web 文本挖掘模型 K-WebMiner。

2 Web 挖掘

区别于一般的信息源,Web 页面具有其自身的特点,这也给 Web 挖掘带来了相当的难度[1-3]。① 庞大性。互联网络上的网页数量以成百亿计算,增长速度不断地提高,现在仍以每月近千万的数量在飞速递增。② 动态性。Web 数量以极快的速度增长,其所提供的信息也在不断地发生改变,需要更新。不光 Web 的内容,相关的链接信息和访问记录也在频繁更新,这是一般信息源所不具备的。③ 异构性。如果从数据库的角度出发,那么所有网站上的信息都可以看作是一个数据库,每个站点就是一个数据源,每个数据源都是异构的,构成了一个巨大的异构数据库环境——Web 环境。④ 半结构化的数据结构。Web 上的数据非常复杂,不能用特定的模型来描述。每个站点按通用标准设计,数据具有一定的结构性;同时,各个站点又有其独立的层次性结构,使相关数据带有不同的网站结构特点,形成了 Web 数据半结构化的特征。人们越来越关注如何开发和利用 Web 上的数据资源,也针对 Web 的特点开发出了各种搜索引擎,这是 Web 挖掘技术的最简单应用。传统的基于关键字的搜索引擎目前仍存在以下缺陷[1-3]。

1) 覆盖面有限。根据权威机构的调查研究显示,即使是目前功能最完善的搜索引擎也只能搜索所有 Web 网页的 1/3 而已,Web 资源的挖掘利用率还很低。

2) 查全率不高。① 经常使用网络的用户可能会发现,一般情况下,无论怎么选择关键词,都会返回大量并不需要的结果,很多与话题的相关性并不大,或所包含的内容质量不高。比如,选择查询一项技术名词"数据挖掘",很多时候得到的搜索结果是含有数据挖掘这一章内容的书籍的广告。② 利用搜索引擎的另一类误差是,有时候很多与主题相关的文档可能并不包含相应的关键字,利用搜索引擎检索关键字就无法发现文档,这类问题称为多义问题,在 Web 挖掘中也是常见的。

3) 搜索速度不理想。搜索引擎的速度也影响着 Web 挖掘技术的应用。在通常情况下,使用一般文件检索或关键字检索的速度还让人可以接受,但对于复杂检索、多次检索、海量检索等情况来说,现有引擎的检索速度还远不能令人满意。

收稿日期:2005-12-13

作者简介:胡安安,男,上海市国顺路 670 号复旦大学管理学院,博士生,主要研究方向为企业信息化实施战略、知识管理、决策支持系统;E-mail:huanan@fudan.edu.cn

由此可见,传统的搜索引擎对于 Web 资源的利用还很不够,而如果连 Web 搜索引擎对 Web 资源的查找利用还不够充分,又何以谈得上 Web 挖掘? 因此,Web 挖掘技术还有很大的发展、提高空间。一般认为,Web 挖掘(Web Mining)就是指利用数据挖掘技术从 Web 文档和 Web 活动中抽取人们感兴趣的、潜在的有用模式和隐藏信息,它实现了对 Web 存取模式、Web 结构规则以及动态 Web 内容的查找。Web 挖掘技术分为 3 个方面:Web 内容挖掘(Web Content Mining)、Web 结构挖掘 (Web Structure Mining)、Web 使用记录的挖掘(Web Usage Mining)[1-4]。

Web 文本挖掘是 Web 内容挖掘中的一个重要组成部分。Web 挖掘技术的结构如图 1 所示[1-4]。

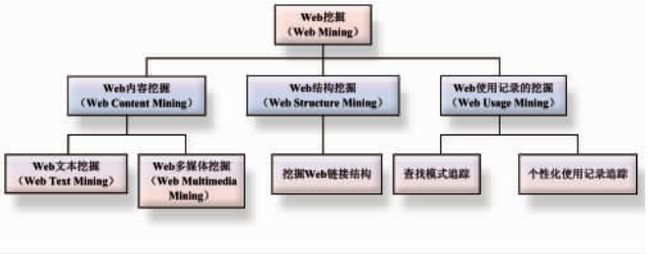

图 1 Web 挖掘技术结构图

Fig. 1 Structure of Web Mining technology

3 Web 文本挖掘和 WebMiner

Web 文本挖掘是指对 Web 上大量文件集合的内容进行文本总结、文本分类、文本群集和文本关联分析,利用 Web 文件的关键信息进行分配分析、趋势预测等操作的数据挖掘技术。

文本总结是指从文件中筛选关键性信息,用简洁的形式对文件内容进行摘要或解释。用户并不需要浏览全文,就可以了解文件或文件集合的总体内容。目前绝大部分查找引擎采用的文本总结方法是截取文件的前几行。

文本分类是指按照预先定义的主题式类别,为文件集合中的每个文件确定一个类别,用户不但能够方便地浏览文件,而且可通过限制搜索范围来使文件的查找更为容易。利用文本分类技术可以对大量文件进行快速、有效的自动分类,很多搜索引擎都采用文本分类方法提高搜索范围和搜索速度。

文本群集与文本分类的不同之处在于群集没有预先定义好的主题类别,它的目标是将文件集合分成若干个簇,要求同一簇内的文件内容的相似度尽可能地大,而不同簇间的相似度尽可能地小。在实际应用中,可以利用文本群集技术将查找引擎的检索结果划分为若干个簇,用户只需要考虑那些相关的簇,大大缩小了所需要浏览的结果数量。目前 Web 挖掘中的文本群集算法大体可以分为两种类型: 以 G-HAC 等算法为代表的层次凝聚法; 以 K-means 等算法为代表的平面划分法[2]。

关联分析是指从文件集合中找出不同词语之间的关系。比如:通过作者名和书名同时出现的模式来发现一些平时找不到的新书籍;通过电影名、导演名、演员名出现的模式来快速定位电影等。

分配分析与趋势预测是指通过对 Web 文件的分析, 得到特定数据在某个历史时刻的情况或将来的取值趋势。现在应用比较广泛的是通过分析相关的评论来预测每日的股票市场。

尽管在很多方面与传统的平面文本挖掘方法比较类似,Web 文本挖掘技术仍有很多独特之处。比如,Web 文件中的标记通常采用 Html 语言,包含大量<Title>、<Heading>等额外信息,这些信息对于 Web 挖掘非常重要, 我们可以利用这些信息来提高 Web 挖掘的性能和速度,达到事半功倍的效果。

为了更好地进行 Web 挖掘, 实现 Web 文本挖掘的总结、分类、群集、关联分析、趋势预测等功能,学术界提出了 Web 文本挖掘的系统模型 WebMiner[4-5],很多学者对其进行了补充和完善[2]。WebMiner 采用了多重"信息代理程序"(agent)的结构,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 文件。WebMiner 的系统组件和结构如图 2 所示[1-5]。

3.1 系统组件

1) 文本挖掘 agent:利用信息查找技术将分布在多个 Web 服务器上的尚待挖掘的文件综合在 WebMiner 的本地文本库中。

2) 文本预先处理 agent:利用启发式规则和自然语言处理从文本中筛选出的、代表其自身特征的元数据,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:使用其内部知识库,按照预定义的类别层次,对文件集合(或者其中的部分子集)的内容进行分类处理。

4) 文本群集 agent:利用其内部知识库,对文件集合(或者其中的部分子集)的内容进行群集处理。

5) 多维文本分析引擎:WebMiner 引入了文本超立方体模型和多维文本分析技术,为用户提供关于文件的多维视觉图。多维文本分析引擎还具有统计分析功能,从而能够揭示文件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件的集合进行特征修剪,包括横向文件选择和纵向特征投影两种方式。

6) 用户接口 agent: 用户接口在用户与多维文本分析引擎之间发挥了联系的功能。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达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觉图和文件展示给用户。

在系统组件中,每个 agent 可作为系统的一个组件,能够完成相对独立的工作。这些组件可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,因此易于加入新的组件,各个 agent 之间通过相互协作来完成挖掘的全部过程。其中,多维文本分析引擎以文本预先处理为基础,以文本挖掘(分类、群集)作为支持,文本超立方体中的维来自于预先处理得到的文本特征属性(例如时间、作者等),而文件主题式类别的生成以及文件之间关系的群集分析依赖于文件挖掘技术。相应地,多维文本分析引擎又为文本采集提供了有效的可视化手段和特征修剪工具, 可以将文件集合的特征修剪结果展现给用户,也可将其作为挖掘对象重新输入到文本分类 agent 和文本群集 agent 中。

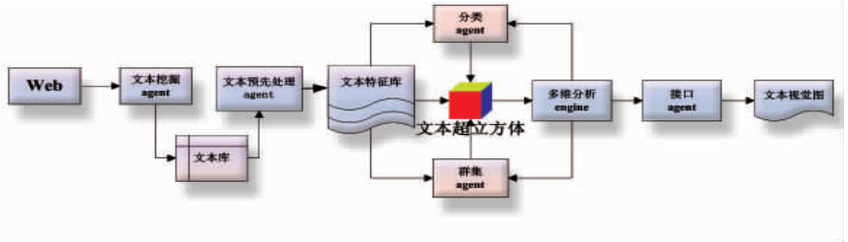

图 2 WebMiner 的系统级件和结构图

Fig. 2 System components and structure of WebMiner

3.2 系统模型运作流程

用户通过与系统中各个组件进行互动来执行 Web 文本挖掘的全部过程。首先,用户给出搜集策略(例如起始的 URL 图标、地址、指定的主题或网络等), 指导文本挖掘 agent 进行 Web 文件搜集。然后,文本预先处理 agent,从搜集到

的 Web 文件中筛选出描述性特征和语义性特征。此时用户有以下几种方案供选择。① 使用多维分析引擎对文件进行多维分析,得到多维文件视觉图(每个视觉图对应于文件集合的一个子集)。② 按照预先定义的类别层次,对文件集合(或者其中的部分子集)的内容进行分类。当预先定义的类别层次与文件集合的内在层次不符合时,用户可以修改或重新创造文本分类 agent 的类别层次预定义,或者利用文本群集 agent 对文件集合进行群集得到文件簇。③ 利用文本群集 agent 对簇进行处理。由于簇也是文件的集合,因此当用户对某个簇感兴趣,而这个簇中又包含很多文件时,可再次使用文本群集 agent 将簇进一步划分为子簇,直到每个簇中包含的文件数目适中为止。这 3 个过程可以多次交互往复进行,直到获得满意的结果,以文本视觉图的形式输出给用户。

4 基于知识库的 Web 文本挖掘系统模型 K-WebMiner

WebMiner 是一个 Web 文本挖掘系统模型,在互联网海量的文本信息中,它可以帮助用户有效地找到需要的信息,并实现用户的各种要求。然而在实际应用过程中,我们发现还有很多问题值得进行进一步的探讨。比如,对于有特定使用习惯的用户群,如何进一步提高数据挖掘的效率;对于大规模的文件集合,能否利用启发式的规则加强信息预处理的能力,以进一步提高检索的速度等。WebMiner 是一个 Web 挖掘的系统原型,它是由不同的挖掘组件所组成的,可以对应于不同需求,设计更多的 Web 挖掘组件,以丰富 WebMiner 的功能。针对于此,可以加强 Web 挖掘在知识库方面的功能,对 WebMiner 进行一些改进,使之成为基于知识库(Knowledge Base)的 Web 文本挖掘系统模型——K-WebMiner。具体的模型改进如图 3 所示[6-7]。

K-WebMiner 对于 WebMiner 的改进之处在于增加了 2 个独立的子模块:实例库 agent 和知识库 agent[7-11],有 2 个最大的特点。① 实例库利用以往的知识积累和数据库,对于那些检索要求相同或类似的文本进行处理,直接给出以往使用挖掘工具的基本信息和处理方法,充分利用已有的资源,方便下面模块的处理,使挖掘的速度大大提高。针对某一特定用户群使用系统时,K-WebMiner 的效果最为明显,比如对于常常调用某一类学术名词的用户群来说,K-WebMiner 可以在实例库中直接调出其处理的基本信息和方法,对 Web 文本进行有效的预处理(如提取元数据等),并告知后面的模块以后的处理过程和方法,使得整个挖掘的过程更加快速、有针对性。同时,针对用户每一次的不同要求,实例库将相关的处理过程和基本信息储存入数据库中,并为之建立快速有效的匹配机制,保证以后类似的要求可以在文本挖掘方法的选用中得到体现。② K-WebMiner 增加了一个专门的知识库,它可以实现原来 WebMiner 中分散在各模块中的知识库功能,换句话说,就是把各模块中的知识库功能整合起来,进行集中处理。知识库对于要进行挖掘的文本进行知识处理,把相关的信息传递给后面的模块,这样就保证了知识处理的集中进行,不再像原来分模块进行处理时那样占用系统的处理资源,保证了整个模型的处理速度和处理效率,而原来的分模块只需将重点花在分类、群集、分析等核心业务上,不必再分散资源进行知识处理[12]。

4.1 K-WebMiner 系统的扩展组件

1) 实例库 agent:根据以往进行的挖掘实例,分析用户可能的需求,对可以匹配的查询或要求可以直接调用实例库中的实例,对文本进行相同或类似的处理,方便后面的 agent 进行操作,提高系统运行的效率。同时,把新的处理不断更新到实例库之中,保证以后类似的文本挖掘可以更加有效地进行。

2) 知识库 agent:利用知识库对于知识的处理能力,对文本进行集中的知识预处理。一方面分离出文本中的元数据,输入文本预先处理子模块;另一方面,利用启发式规则和以往知识的推理过程来补全文本的特征信息,方便后期的处理。举个例子来说明,文本中出现"雅虎"字样的文本,为了方便后面模块的处理,可以利用知识库已有的知识推理过程来将之补全或替换为"Yahoo",这样后面的模块再进行处理时就方便得多了。

3)文本预先处理 agent:在知识库的帮助下任务减轻了很多,主要工作是将知识库从文本中筛选出来的元数据进行进一步预处理(考虑后期的处理和用户的要求),并存放在文本特征库中,作为文本挖掘的基础。

4) 文本分类 agent:利用前期知识库预处理的结果,按照预定义的类别层次,对文件集合(或者其中的部分子集)的内容进行分类。

5) 文本群集 agent:利用前期知识库预处理的结果,对文件集合(或者其中的部分子集)的内容进行群集。

6) 其它模块 agent:功能与没有扩展前的 WebMiner 功能基本一致,在此不再赘述。

4.2 K-WebMiner 系统模型运作流程

用户还是通过与 K-WebMiner 系统中各个组件的互动来执行 Web 文本挖掘的全部过程。首先,用户给出搜集策略对相关的 Web 文件进行搜集,系统对实例库与文本库中的文本进行实例匹配分析,与以往的用户要求和文本结构进行匹配,如果符合匹配的要求,则从实例库中调用相应的处理方法和过程,并将之传递到下一个模块中作为重要的参数。如果不符合匹配要求,则实例库跟踪用户要求和文本挖掘的全过程,将之记录于库中,为以后的匹配调用做好准备。经过实例库处理之后,知识库对 Web 文件进行启发式知识分析,从其中找出符合知识库内在推理模型的相关特征,或是补足、替换掉其中的一些文本,将这些文本挖掘的原材料传递给文本预先处理子模块,文本预先处理子模块根据用户的具体要求对元数据进行处理,将处理好的文本传递给文本特征库,由其来决定文本需要进入接下来的哪一个子模块。后面子模块的处理过程与 WebMiner 相类似,其核心特征就是通过用户与系统的交互,保证用户获得满意的结果,并以此作为整个 Web 挖掘结束的信号。

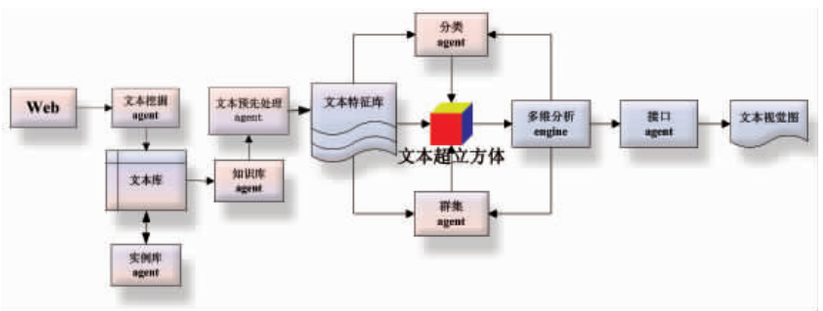

图 3 K-WebMiner 的系统组件和结构图

Fig. 3 System components and structure of K-WebMiner

5 结论与展望

在 Web 信息爆炸的今天,Web 挖掘技术本身就是一个具有极大挖掘潜力的宝藏,从 WebMiner 的子模块的交互式挖掘到基于知识库的 K-WebMiner 挖掘模型,我们追求的目标就是以最简单、有效的方法去检索尽可能多的 Web 文本,从中找到我们所需要的信息。WebMiner 模型本身是一个由不同 Web 挖掘组件构成的系统模型,特征就在于可以不断扩充、完善。基

科技导报 SCIENCE & TECHNOLOGY REVIEW

中国粮食补贴政策的演进与绩效分析

Study on Performance and Advance of Food Subsidy Policy in China

何忠伟/HE Zhong-wei

北京农学院经济贸易系，北京 102206

Department of Economy & Trade, Beijing Agricultural College, Beijing 102206, China

[摘要] 分析了中国粮食补贴政策的演进，对粮食补贴政策的绩效进行评价，提出了改善粮食补贴政策的建议，如合理分摊粮食补贴政策的成本、实施粮食价格市场化、改善粮食补贴对象与结构、增加农业支持总量等。

[关键词] 粮食补贴政策；演进；绩效分析 [中图分类号] F323.8, F326.11

[文献标识码] A [文章编号] 1000-7857(2006)04-0071-05

Abstract: It is very effective to put the food subsidy policy in practice for people. The paper analyzed advances of the food subsidy policy, appraised performances of the food subsidy policy and put forward some advices: cost apportion, reforming the price of food, improving the structure and object, increasing support-gross.

Key Words: food subsidy policy; advance; performance **CLC Numbers:** F323.8, F326.11

Document Code: A **Article ID:** 1000-7857(2006)04-0071-05

粮食是一种同时具有一般商品和公共品属性的特殊商品，它的供给弹性大、需求弹性小，受自然风险和市场风险双重制约。按照产权经济学家诺斯的理论，具有公益性的产品，其社会收益远远大于生产者的个人收益，如果没有外力支持保护，其供给量将少于正常需求量[1]。这就是粮食产业自身产业缺陷造成的粮食市场供给失败。因此，在市场经济条件下，政府必须修补粮食产业缺陷，应对粮食产业进行补贴，缩小粮食社会收益与企业、个人收益之间的差距，激励农民种植粮食与企业销售粮食，以确

收稿日期：2005-12-06

基金项目：2004 年湖南省软科学研究计划项目(04ZH3083)

作者简介：何忠伟，男，北京市昌平区朱辛庄北农路 7 号北京农学院经济贸易系副教授，财政部财政科学研究所博士后，主要研究方向为农业政策、农业技术经济；E-mail: h2w28@126.com

于知识库的 K-WebMiner 模型结构也需要不断地增加功能模块，以实现更有效的 Web 挖掘功能。

对于 Web 挖掘技术的未来展望，其重点在于两点。① 随着 Web 利用 XML 规范文档的不断普及，对于文档元数据的描述和筛选方法变得更加容易和统一，Web 数据挖掘可以充分利用这一特征，使大规模 Web 文本挖掘的瓶颈得到有效的解决，这是 Web 挖掘技术发展的主流方向。② 安全性问题。随着挖掘技术的不断更新和发展，人们可以从 Web 上获得越来越多的信息，甚至包括那些别人并不愿公开的信息，因此就涉及到安全隐私方面的问题，这也是伴随着挖掘技术的发展人们所要特别关注的重点。随着 Web 挖掘技术的不断进步，一定会有更多的 Web 文本挖掘组件(XML 方面、安全性方面等)来丰富 WebMiner 的功能[1-3]。

参考文献(References)

[1] HAN JIAWEI, KAMBER MICHELINE. Data Mining Concepts and Techniques[M]. 范明，孟小峰，等译. 北京：机械工业出版社，2002.

[2] 林杰斌，刘明德，陈湘编著. 数据挖掘与 OLAP 理论与实务[M]. 北京：清华大学出版社，2003.

[3] 陈京民，等编著. 数据仓库与数据挖掘技术[M]. 北京：电子工业出版社，2002.

[4] COOLEY R, MOBASHER B, SRIVASTAVA J. Web mining: Information and pattern discovery on the world wide web[R]. Technical Report T R97-027. 1997.

[5] MOBASHER B, JAIN N, HAN E, SRIVASTAVA J. Web mining: Pattern discovery from world wide web transactions[R]. Technical Report.T R96-050, 1996.

[6] 李东. 一个基于知识的 DSS 模型管理系统概念框架 [J]. 决策与决策支持系统，1995，5(2)：26-32.

[7] 胡军军，肖人彬，周济. 面向模块的知识库管理技术[J]. 华中理工大学学报，1997，25(1)：12-15.

[8] 董军，肖少拥. 知识库系统的现状与发展趋势 [J]. 计算技术与自动化，1995，14(3)：1-4.

[9] 陆汝钤，等. 面向 agent 的常识知识库 [J]. 中国科学，2000，30(5)：453-463.

[10] 匿名. 企业管理决策分析系统中知识库的设计 [J]. 微机发展，1994(1)：19-22.

[11] 汪虹，李龙树. 基于约束 agent 的知识库结构研究[J]. 计算机工程与设计，1999，20(4)：12-15.

[12] 张绍华，等. 基于样本实例的 Web 信息抽取 [J]. 河北大学学报，2001，21(4)：431-437.

(责任编辑 胡春华)