基于 Web 内容挖掘的信息支持工具 AIS-GAE

张 文 唐锡晋

(中国科学院数学与系统科学研究院,北京 100080)

摘要:对于 Internet 所承载的海量信息,一般需要通过人们主观选择或机器帮助的情况下过滤出为人可利用的有效信息,为此在期望提高信息过滤、使用效率并对人们的各种分析与决策任务支持的驱动下,各种信息挖掘技术的相关研究受到重视。在简要介绍 Web 挖掘技术的基本内容后,本文叙述了中文 Web 内容挖掘的工作过程及技术实现。通过对一个著名的科学论坛"香山科学会议"网站的应用,具体说明 AIS-GAE (Augmented Information Support for Group Argumentation Environment)如何为香山科学会议的各方用户提供有效的信息支持。文章最后指出了一些问题及值得改进和推广的工作。

关键词:中文 Web 内容挖掘;文本挖掘;香山科学会议

引 言

随着互联网络信息承载量以及用户数目的日益膨胀,作为探测有效信息、提高使用效率的有效手段,Web 挖掘的研究已受到关注、成为热点。Web 挖掘旨在使用数据挖掘技术从 Web 资源中发掘出有用的规律和模式[1]。Web 挖掘一般分为三类:Web 内容挖掘、结构挖掘和 Web 使用记录的挖掘,其中 Web 结构挖掘是从 WWW 的组织结构和链接关系中推导知识;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;Web 内容挖掘是从文档内容或其描述中抽取知识的过程,它主要包括直接对 Web 页面文档内容以及对搜索引擎的查询结果进行文本的总结、分类、聚类、关联分析等[2]。

Web 内容挖掘研究目前主要体现在两个方面:一是 Web 内容和关键词抽取和表示技术的研究,如文[3]基于 Web 内容特征提出了一种新的自动生成抽取模式的算法 EGA (EPattern Generation Algorithm),文[4] 根据关键词在文本中的分布和关键词在文本间的协同出现频率来发现知识并开发了 KDT(Knowledge Discovery in Texts)系统,文[5]利用 SOM(Self-Organizing Mapping)方法训练文档表示向量生成超级链接关键词和链接源从而在 Web 页面中自动生成超级链接,等等。另一方面的研究则在抽取与表示技术基础之上进行有关 Web 内容的总结、分类、聚类、观点挖掘等的研究,如文[6]提出了一种新的文本过滤和转换方法,试图缩小学术界和商业界对于同一概念由于称名不同而导致的沟通鸿沟,文[7]通过网页间种子内容的关系,用于在

收稿日期:2006-01-18
基金项目:国家自然科学基金资助项目(70571078;70221001)。
作者简介:张文,中国科学院数学与系统科学研究院硕士研究生;唐锡晋,中国科学院数学与系统科学研究院副研究员。

个人网页间发现个人兴趣相关的子群,文[8]设计和实现了一个 Web 新闻摘要系统 Ai-Times, 并专门针对 Web 新闻设计实现了 Spider 和抽取算法以及摘要生成算法。

本文所介绍的工作综合了以上两方面的研究。首先基于中文 Web 内容挖掘的一般工作过程所包括的最基本的任务介绍其中各个部分的功能及其实现。之后以香山科学会议网站为基础,示范了所集成的 Web 信息挖掘工具 AIS-GAE 的实际应用。最后指出了一些问题、值得的改进和推广。

基于 Web 的内容挖掘

Web 内容挖掘综合运用 Web 技术和文本挖掘技术帮助人们从 Web 页面中更好地发现、组织和表示信息,并通过文本挖掘提取知识,进而实现自动为用户提供相关度较高的文档等个性化服务。内容挖掘是信息使用需求驱动的必然结果,但离不开基础信息抽取技术的支撑。图 1 勾勒了一个包含中文 Web 内容挖掘的最基本任务的工作过程,包括四个部分:网络爬虫、索引程序、文本总结和用户使用接口,每一部分承担 Web 挖掘中的不同任务。下面分别介绍其相关的原理和功能。

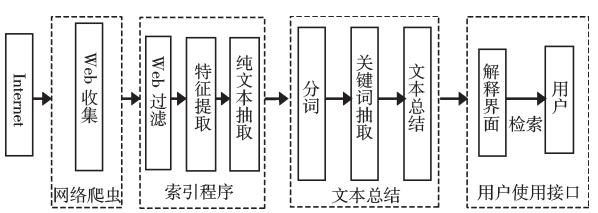

图 1　简化的中文 Web 的内容挖掘的工作过程

1、网络爬虫

网络爬虫的主要功能是从特定的站点下载用户感兴趣的相关信息。以特定的站点为种子站点,保证用户需要的信息来源质量。其原理一般基于传统的 Spider 探测法, 通过解析 Web 页面中的 URL 链接信息逐步下载 Web 页面, 并将 Web 地址和页面分别保存于数据库和文件库。此法的关键在于种子 URL 的选择,一般由用户选定一些感兴趣的相关的 URL,这样可在一定程度上提高获取相关页面的效率。

2、索引程序

索引程序用于整理和过滤网络爬虫收集的 Web 页面, 并将与 Web 页面特定信息存放在数据库中。它包括三个部分:Web 过滤,特征提取和纯文本抽取。Web 过滤用于去除无关的网页,如广告、错误页面等等。特征提取主要根据 Web 页面结构提取特定的信息,如页面的主题、特定位置的信息、页面生成的时间、作者等等与页面密切相关的信息。纯文本抽取则是去除 Web 页面的 HTML 标签和脚本以及页面样式等等信息, 使之生成规范化的纯文本。这种抽取主要依赖于通过对 Web 符号识别,去除 Web 页面中纯文本以外的部分。

3、文本总结

文本总结通过对 Web 页面的内容进行分析,提取出能够代表 Web 页面内容的文本摘要。这是文本挖掘的主要任务之一。文本总结包括三部分:分词(亦称切词),关键词抽取和自动文摘。“词是最小的能够独立活动的有意义的语言成分”[9]。分词是中文文本挖掘的特有的任务,通过分词,纯文本中的语句被切分成词语单元。关键词抽取用于从分词程序处理过后的词语集合中选择能够代表 Web 页面内容的特征词汇,该工作一般与对象密切相关,有关处理对象的深入了解而表现的经验或者知识可以帮助选择代表性显著的特征词汇。自动文摘是按用户需要从文本中抽取出具有代表性的句子,并按句子在原文中的顺序重新组织形成原文所代表的中心内容,以提高读者获取原文信息的效率。目前自动文摘方法有两类:一类是基于语义的文摘,另一类是基于统计的文摘[10],而后者应用更多。

4、用户使用接口

用户使用接口向用户提供了一个经过加工后的信息发布和使用的接口,按照用户的习惯和容易理解的方式将重要的信息表现出来。最朴素的表现为一般的搜索引擎。

可以看到, 通过 Web 技术和文本挖掘技术,实现了对 Internet 上的 Web 做二次的加工、整理、分析,提取出用户感兴趣的信息。当前,中文 Web 内容挖掘的研究大多停留在理论研究阶段或者针对 Web 内容挖掘中的某项具体技术,主要在图 1 所示的“文本总结”中的几项相关技术,且研究成果也因研究对象各有千秋,除了分词技术有一些标准测试以作技术比较外,其它研究一般则根据一些现成算法结合对象而进行。下面介绍我们实现的基于图 1 流程的信息支持工具 AIS-GAE, 并针对香山科学会议这一对象展示了通过 Web 技术和文本挖掘技术,收集、整理、加工、分析有关香山科学会议的 Web 信息。

香山科学会议信息支持工具 AIS-GAE

国内著名的科学论坛香山科学会议自 1993 年创办以来,以紧迫前沿,激励创新,促进交叉,倡导争鸣,发现人才而著称[11]。本节介绍的 AIS-GAE 利用 Web 内容挖掘技术,对香山科学会议网站(www.xssc.ac.cn)实现或集成了图 1 所示的各项功能,为用户提取有关香山科学会议的各种相关信息。同时,AIS-GAE 也是群体研讨环境(Group Argumentation Environment-GAE)电子智暴研讨室信息视图(information viewer)中所展示的积极信息支持功能模块[12]。主控面板如图 2 所示。下面分别介绍 AIS-GAE 的具体实现。

图 2 香山科学会议集成工具主界面

1、香山科学会议 Web 页面采集

网络爬虫是网页采集的工具。由于香山科学会议不断举行,需要定期使用爬虫更新采集内容。图 3 为爬虫程序接口,这里设置香山科学会议网站关于历次会议的 URL 地址为种子站点,爬虫深度为 10。通过爬虫程序,从香山科学会议网站上一次下载了 646 个网页。

图 3 香山科学会议网络爬虫

2、Web 页面索引

对获取的网页进行过滤,只选取有关历次会议相关的网页内容,并将这些网页 URL 和 Web 文件名存储于数据库。通过 Web 过滤,获取了 208 个网页。图 4 为一个典型的香山科学会议的总结网页,可以看到该页面在结构具有一定的特点。通过识别历次会议 Web 页面的结构进行特征抽取,可得到会议题目、会议内容和参加会议的人员等信息。

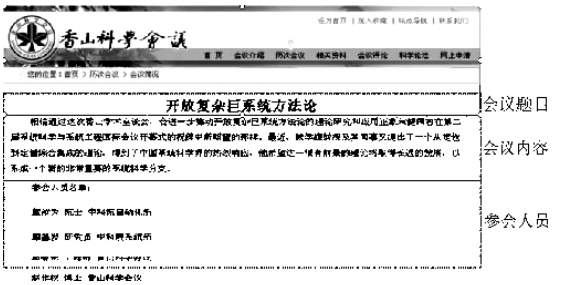

图 4 香山科学会议 Web 页面结构

这些 Web 页面大体结构上一致,而其内容长度以及格式的不一致仍导致实际页面结构上存在具体细节的差异。通过纯文本抽取则识别并删除 Web 页面 HTML 源文件中的 HTML 标签和脚本、样式等特殊字符,获得了该网页上所示的主要内容的纯文本,其中标题存放在数据库中,会议内容介绍、参会人员和 HTML 标签存放在文本文件中。图 5 和 6 分别是识别出来的会议内容介绍、参会人员。

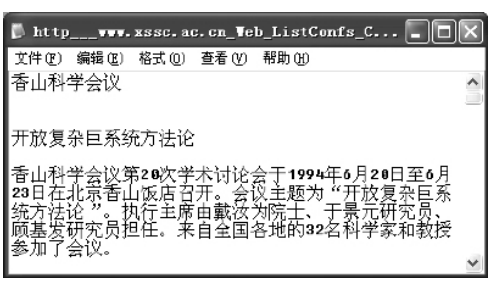

图 5 抽取出的纯文本页面内容

图 6 抽取的参会人员名单

3、页面文本总结

这里,首先采用中科院计算所研制的 ICTCLAS 分词程序[13]对每一份抽取出的纯文本内容切词,并根据词性标注保留其中的名词和名词词组,选取其中 15%的高频词汇进入初选关键词集,从其中又选取 5%的高频词汇加入去除词汇集,以便去除初选关键词中的频繁词语,留下真正能够表征内容的特征词汇。在生成去除词集时,采取了一些人工干预的方法,如,根据香山科学会议更多地表达科技内

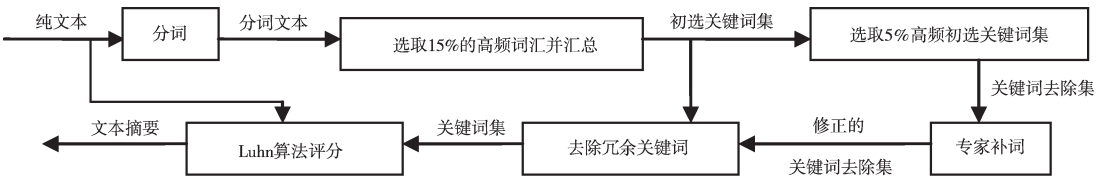

图 7 页面文本总结的详细流程

容的特点,添加某些科技领域频繁词汇(包括简语或者缩略语)或者删除一些具有某种独特意义的词汇。最后,每份纯文本文件的初选关键词集减去关键词去除集即得到了每份纯文本的关键词集。根据所获得的关键词集,采用经典的 Luhn 算法,对该文本中的每个句子进行评分,按照给定的压缩比选取高评分句子,并按它们在原文中的顺序排列而生成该文本的摘要。

图 7 是 Web 页面文本总结的详细流程图。在对特征词汇的选取过程中,我们注意到香山科学会议的中文文本关键词的分布规律并不同于文[14]介绍 Luhn 算法时提到的英文文章的关键词分布规律。经典 Luhn 算法是一种纯统计的方法, 而实际操作中通过人工干预,调整特征词汇集合,使其更具代表性,这样生成的摘要更能概括原文的主要内容。人工干预需要对实际对象进行深入分析,目前所见的通用摘要程序一般仅是基于某个通用统计算法而生成文摘,而针对特定对象显然人类知识的加入能够改善生成文摘的代表能力。

4、用户使用接口

图 8 为 AIS-GAE 的使用界面, 类似于一般的搜索引擎。使用时,用户输入关键词,启动 search,搜索程序对数据库中所有 URL 对应于文件库中的 Web 文件进行全文搜索。若文件包含用户设定的关键词,则该文件被选中。搜索结果显示在搜索栏的下方,包括搜索结果的统计和搜索到的按与设定关键词相关性排列的文本。这里的相关性通过查找到的 Web 文件中设定关键词的出现频率来表示,其中搜索结果显示了查找到的 URL 地址、文件标题、文件中评分最高的句子和查看文摘的链接,该链接指向按照原文 10%的比率生成的原文摘要,并高亮显示了其中的关键词。压缩比可由用户调整。图 9 显示了图 8 所示查找结果的第三条相关记录的摘要。

以上以香山科学会议为例介绍了所实现的基于 Web 文本挖掘技术的信息支持工具 AIS-GAE 的基本功能。而对于香山科学会议本身所独具的特色所反映在网页上的信息内容的挖掘本身也具有多种用途。

关于香山科学会议网页的深入信息挖掘

这里我们以香山科学会议网站的使用人员考察香山科学会议信息支持工具 AIS-GAE 的多种用途。一般来讲,关注香山科学会议乃至其网站内容的主要人员分为一般浏览用户和香山科学会议相关用户。上节所介绍的功能一般能够满足前者的使用需求。香山科学会议相关用户则包括与会专家、会议组织者(会议主席)、会议协调人员(香山科学会议学术秘书)、会议评议人员和理事会单位负责人员(即经费供应方)。这些相关用户通过信息支持工具 AIS-GAE 的使用可以获得下述的帮助。

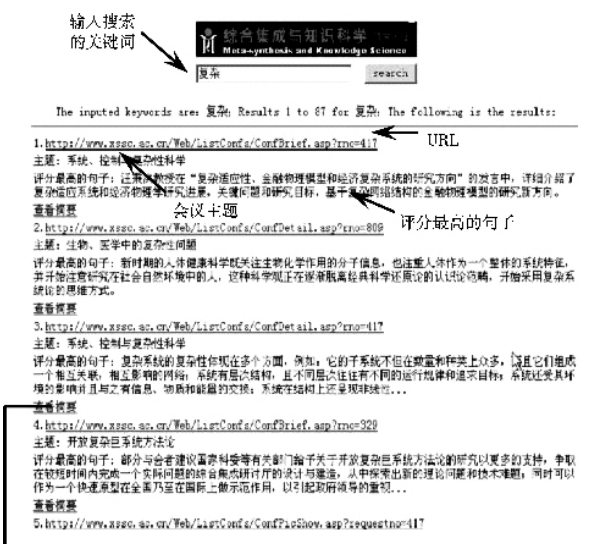

图 8 信息工具界面(信息搜索)

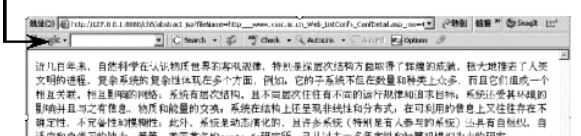

图 9 关键词高亮显示的页面摘要

对于与会专家、会议组织者、会议协调人员、评议专家乃至理事会负责人员:了解以往相关会议的详细信息,一些相关专题在香山科学会议上被讨论的历史记录。图 8 列出了与“复杂”相关的香山科学

会议有关网页的搜索结果。利用 AIS-GAE,在对文件库中所有的文件进行全文搜索过后,总共找到 87 个与“复杂”相关的记录,不同的用户按各自需求与兴趣,根据文摘信息,可以选择是否深入考察下去。这样的搜索对于了解关于“复杂”专题以往的研讨重点,对提交、评价新的会议申请工作提供了及时的帮助。而这样的帮助可以替代对香山科学会议网站的人工搜索,从而帮助专家花费较少的时间掌握大量的信息支持其工作。具体地,对于某一会议评议人员在面对一个新的复杂系统与复杂科学的会议申请时,根据搜索,获得了图 8 所示的结果,共 10 次有关会议的各种 Web 网页。通过考察以往历次会议,勾勒出相关学科前沿研究的国内发展动态,对评价当前申请具有有效的信息支持。如图 8 中排列前面的几项相关条目,第 4 条是 1994 年 6 月香山科学会议召开了一次题为“开放复杂巨系统方法论”会议的简要介绍,第 1 条、第 3 条和第 5 条都是关于 2004 年 5 月召开的题为“系统、控制与复杂性科学”的会议。单从文摘信息所反映的内容,说明了两次会议的侧重很不一样,“开放复杂巨系统方法论”着重点在于研究复杂巨系统问题的方法论综合集成方法论及其具体实践,而“系统、控制与复杂性科学”着重点在复杂系统与复杂性科学的多视角研究。

前文图 6 显示 AIS-GAE 针对香山科学会议网页所抽取的特有信息:与会人员名单。将这样抽取出来的历次会议参与成员名单汇总,形成一个与会人员信息表,也提供了深入的信息支持,如帮助找出可能感兴趣的历次会议的相关专家。若评议一个有关专题为“复杂”的会议,通过查找而得到的以往历次与“复杂”相关的会议参加人员表,对照新申请的会议邀请人员列表,可帮助会议评议人员和协调人员判断会议参与成员知识结构、年龄结构和机构分配的合理性。更进一步,根据信息表中信息统计与会人员对于某类主题相关会议的出现频率,相对高频出现的人员一般可视为有关专题的活跃人员。此外还可通过构造基于各种关系的人际网络,根据社会网络特征来挖掘隐含的关联模式。这些深入挖掘而获取的信息可帮助会议评议人员、协议人员以及理事会负责人员客观审视与会人员组成,了解目前国内有关主题的研究力量的一些变迁和现状,为贯彻香山科学会议的宗旨、拓广科学前沿问题探讨的广度与深度、改进具体会议的人员配置等具有更加有意义的支持与帮助。此项研究单独著文。

结束语

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 挖掘和文本挖掘倍受关注。本文介绍了一种包括基本任务的中文 Web 内容挖掘的工作过程所实现的信息支持工具 AIS-GAE,它主要包括了网络爬虫、索引程序、文本总结和用户使用接口等基本功能。以香山科学会议网站为研究对象,具体介绍了该信息支持工具 AIS-GAE 所实现或者集成的中文 Web 内容挖掘一般工作过程的各项主要功能的原理、技术实现与实例应用。并针对特定对象的领域特色,介绍了更高一级的信息挖掘功能,展现了中文 Web 内容挖掘对香山科学会议各方的信息支持。我们使用 GAE 对香山科学会议作了深入的探索,不断挖掘出一些内隐信息,得到了香山科学会议办公室的肯定[15],而这些挖掘工作的原始资料均来源于公开的香山科学会议网站。

尽管目前 AIS-GAE 中的信息支持功能已完整实现(或者集成了)中文 Web 内容挖掘的一般过程的核心功能,这样的工作仍很初步,尚有许多需要修改和完善的地方。从技术角度,设计效率更高的爬虫算法,如何更准确地选取特征词汇以及摘要的生成算法等都是值得进一步提高和改善的工作。应用方面,就香山科学会议本身而言,因不断分析而展开了更宽的视野,产生了许多值得研究的个性化信息挖掘的需求。此外,扩展的 AIS-GAE 应用领域,如对新闻网站、BBS 的挖掘今后也值得大量的研究投入。

参考文献:

[1] 刘丽珍,宋瀚涛,陆玉昌. Web 使用挖掘的应用研究. 计算机科学, 2003, 30(9): 46-48

[2] 李颖,阎保平. Web 文本挖掘在互联网信息统计中的研究与设计. 微电子学与计算机, 2002, 22(1): 62-69

[3] Li,L.,et al.. EGA: An Algorithm for Automatic Semi-structured Web Documents Extraction. In: Y. Lee, et al. (Eds). Database Systems for Advanced Applications (proceedings of the 9th International Conference, DASFAA 2004), LNCS 2973. Springer, 2004: 787-798

[4] Fieldman,R., I.Dagan. Mining Text Using Keyword Distribution. Journal of Intelligent Information Systems, 1998, 10(3): 281–300
[5] Yang,H.C.,C.H.Lee. A text mining approach for automatic construction of hypertexts. Expert System with Applications. 2005, 29 (4): 723–734
[6] Kanai,T., J.Li, S.Kunifuji. Related Document–based Information Filtering Applied to the Association Model Information Retrieval System. Proceedings of the Fourth International Conference on Knowledge –Based Intelligent Engineering Systems & Allied Technologies, 2000,1:225–228
[7] Nakada,T.,S.Kunifuji. Subgroup Discovery among Personal Homepages. In:G. Grieser, Y.Tanaka & A. Yamamoto (Eds.). Discovery Science (proceedings of the 6th International Conference, DS 2003), LNCS 2843. Springer, 2003:385–392
[8] Liu,N.K.,W.D.Luo,M.C.Chan. Design and Implement a Web News Retrieval System. In: R.Khosla, R.J.Howlett and L. C.Jain(eds.). Knowledge–Based Intelligent Information & Engineering Systems (proceedings of KES'2005, Part III), LNAI 3683. Springer, 2005: 149–156
[9] 朱德熙. 语法讲义. 北京:商务印书馆,1982:11
[10] Ren,F.J.. Automatic abstraction important sentences. International Journal of Information Technology & Decision Making, 2005, 4 (1): 141–152
[11] 香山科学会议网站. http://www.xssc.ac.cn
[12] Tang,X. J., Y.J.Liu, W. Zhang. Computerized Support for Idea Generation during Knowledge Creating Process. In: R.Khosla, R. J. Howlett and L. C.Jain(eds.). Knowledge–Based Intelligent Information & Engineering Systems (proceedings of KES'2005, Part IV), Lecture Notes on Artificial Intelligence, Vol.3684. Berlin Heidelberg:Springer–Verlag, 2005:437–443
[13] ICTCLAS. http://nlp.org.cn/~zhp/ICTCLAS/codes.html
[14] Luhn,H.P.. The Automatic Creation of Literature Abstracts. IBM journal of research and development,1958, 2(2): 159–165
[15] 刘怡君,唐锡晋,李增惠. 对香山科学会议跨学科研讨的一种初步分析. 载:刘思峰等主编.《管理科学与系统科学新进展》(第8届全国青年管理科学与系统科学学术会议论文集). 南京:河海大学出版社,2005年5月:35–40

(上接第55页)

参考文献:

[1] Johnson,H.T., Robert Kaplan. Relevance Lost: The Rise and Fall of Management Accounting. Boston: Harvard Business School Press, 1987
[2] 牧户孝郎. 最近におけるわが国原価管理実践の傾向. 企業會計, 1979, 31(3): 126–132
[3] 日本会計研究学会. 原価企划研究の課題. 东京: 森山書店, 1996
[4] 加登豊.原価企划: 戦略的コストマネジメント. 東京: 日本経済新聞社, 1993
[5] Hiromoto, Toshiro. Another Hidden Edge: Japanese Management Accounting. Harvard Business Review, 1988,64(4): 22–26
[6] Ford, Worthy. Japan´ s Smart Secret Weapon. Fortune, 1991,12
[7] Yutaka, Kato. Target costing support systems: Lessons from leading Japanese companies. Management Accounting Research, 1993, 4(1): 33–47
[8] Chen, Richard C., Chen H. Chung. Cause–Effect Analysis for Target Costing. Management Accounting Quarterly, Winter 2002, 1–9
[9] Kaplan, Robert, Anthony Atkinson. Advanced Management Accounting. New Jersey: Prentice Hall, Upper Saddle River, 1998, 222–239
[10] 陈胜群. 论日本成本管理的代表模式:成本企划. 会计研究, 1997(4): 47–51
[11] 陈胜群. 现代成本管理论. 北京: 中国人民大学出版社, 1998:95–98, 189, 144, 111
[12] 王福胜. 价值链作业产出价值计量的模糊分析. 中国会计学会2nd会计教授会暨2004年学术年会论文集, 2004:795–802
[13] 汪方军, 等. 作业基础成本改善控制研究. 管理评论, 2005, 17(3): 24–27
[14] 汪方军. 基于作业的资源成本模型研究. 系统工程理论与实践, 2004, 24(5): 34–40
[15] 田中一成著, 余纯麟译. 全面成本管理, 上海: 文汇出版社, 2002