文章编号:1008-0570(2006)04-3-0244-03

# 数据挖掘在智能搜索引擎中的应用

The Application of Data Mining on Intelligent Search Engine

(成都西南交通大学)杨占华　杨　燕
Yang,Zhanhua　Yang,Yan

摘要:随着互联网的迅速发展,WWW 上信息增长越来越快,传统搜索引擎已经不能满足用户的需求。将数据挖掘技术应用到搜索引擎领域,从而产生智能搜索引擎,将会给用户提供一个高效、准确的 Web 检索工具。文章首先介绍了搜索引擎的工作原理和相关概念,然后介绍了数据挖掘的概念。最后,详细讨论了数据挖掘技术在智能搜索引擎中的重要应用。

关键词:数据挖掘;搜索引擎;Web 挖掘
中图分类号:TP311　　文献标示码:A

Abstract:With the rapid development of the Internet and valuable information, the history search engine can't satisfy people's requirements. Combining the technology of data mining and search engine, intelligent search engine is formed,which will provide users an effective and accurate web information search tool.This paper introduces the theory and correlative conception of search engine first,then introduces the conception of data mining.In the end,the applications of data mining in intelligent search engine are discussed in detail.

Keywords:Data mining;Search engine;Web mining

技术创新

## 1 引言

随着 Internet 和 Web 技术的飞速发展和普及,信息获取已经从手工获取,到计算机获取,以及到现在的通过网络进行信息获取。要想在这浩如烟海的网络世界中找到所需信息,作为现代信息获取技术的主要应用-搜索引擎是必不可少的。据 CNNIC 于 2005 年 1 月 19 日发布的统计资料显示,有 86.6%的用户是通过搜索引擎获得新网站的;搜索引擎的应用占到网络应用的 65.0%,成为第二大互联网应用,它的应用广泛度仅次于电子邮件。

搜索引擎的出现极大的方便了用户,但是由于搜索引擎是由传统检索技术发展而来, 它存在很大缺陷,例如:返回结果表示方法简单,逻辑运算符组合受限,不能利用检索的历史信息等。在当前用户要求不断提高的情况下,单单依靠传统搜索引擎已经不能够满足人们的需要。例如,当我们进行搜索时,搜索结果中存在大量的无用信息,其比例高达 80%以上,搜索引擎通常会返回成千上万条结果,而这些结果只是按照与查询的相关度从大到小排列成一个线性列表,缺乏清晰明了的结构,这些结果中相关信息和无关信息掺杂在一起,这就使得我们要想找到所需的信息仍旧存在很大困难。于是,人们总结出了所谓的"因特网信息检索定律":在因特网上总能找到(甚至只能找到)不需要的东西为了解决这些问题,进一步利用 Web 上的信息资源,人们开始研究信息获取的方法,提出了一些新的信息管理手段。Web 挖掘就是其中之一,Web 挖掘是指将传统的数据挖掘技术和 Web 结合起来,既在 WWW 上挖掘有趣的、潜在的、蕴藏的信息以及有用的模式这样一个过程。将 Web 挖掘应用在搜索引擎中,可以改善检索结果的组织,提高查准率和查全率,增强检索用户的模式研究,对搜索引擎进行优化。

## 2 搜索引擎概述

搜索引擎可分为四个模块:搜索器,索引器,检索器和用户接口。搜索器根据一定的网页搜集策略和规划,调度运行网页自动搜索软件(如 Crawl、Spider、pursuit、harvest 等), 对互联网上的网页进行快速有效的搜集,并将它们存入搜索引擎的网页数据库中。常用的有三种搜索策略:根据所提供的"种子 URL"开始搜索;根据网站受欢迎程度,规划一组 URLs,然后开始搜索;根据网址名称或国家编码,将 Web 空间划分为若干块开始进行搜索。索引器的功能是理解搜索器所搜索的信息,从中抽取出索引项,用于表示文档以及生成文档库的索引表。索引器可以使用集中式索引算法或分布式索引算法。检索器的功能是根据用户的查询在索引库中快速检出文档,进行文档与查询的相关度评价,对将要输出的结果进行排序,并实现某种用户相关性反馈机制。常用的信息检索模型有集合理论模型、代数模型、概率模型和混合模型四种。用户接口

杨占华:硕士研究生
四川省重大基础研究项目 (04JY029-001-4)
西南交通大学科技发展基金(A2004015)

的作用是输入用户查询、显示查询结果、提供用户相关性反馈机制。分为简单接口和复杂接口两种。简单接口只提供用户输入查询串的文本框，复杂接口可以让用户对查询进行限制。

一个搜索引擎的好坏与以下几个因素有关：网页覆盖率、网页更新率、网页检索速度、网页检索质量。

## 3 数据挖掘概述

我们现在已经生活在一个网络化的时代，信息变化异常快速。面对信息爆炸的时代，人们开始考虑：“如何才能不被信息淹没，而是从中及时发现有用的知识、提高信息利用率？”。面对这一挑战，数据挖掘（也称知识发现）技术应运而生，并显示出强大的生命力。

数据挖掘技术已被应用在商业数、电信和医疗保险等领域，Internet 的出现使它的应用更广阔，用数据挖掘的原理来对网络数据进行深层挖掘，发现并组织网络知识，是将网络信息检索技术推向智能化高度的有效手段。于是，Web 挖掘应运而生，Web 挖掘是指将传统的数据挖掘技术和 Web 结合起来，既在 WWW 上挖掘有趣的、潜在的、蕴藏的信息以及有用的模式这样一个过程。将 Web 挖掘应用在搜索引擎中，可以改善检索结果的组织，提高查准率和查全率，增强检索用户的模式研究，对搜索引擎进行优化。

Web 数据挖掘一般可分为三类：Web 内容挖掘（Web content mining），Web 结构挖掘（Web Construct mining）和 Web 使用记录挖掘（Web usage mining）。Web 内容挖掘是从 Web 文档内容及其描述中获取有用知识，是对网页数据进行挖掘，包括文档挖掘和多媒体挖掘。We 文档通常包含以下几种数据类型：文本、图像、音频、视频、元数据和超链接，主要挖掘的对象是 HTML 文档所包含的半结构化数据和无结构的文本数据。Web 结构挖掘是从网页的超级链接中发现其结构及其相互关系。通过找到隐藏在一个个页面之后的链接结构模型，就可以利用这个模型对 Web 页面重新分类，也可以用于寻找相似的网站。Web 结构挖掘可以进行网页分类，总结网页和网站的结构，生成诸如网站间相似性、网站间关系的信息。Web 使用记录挖掘是从用户“访问痕迹”中获取有价值的信息，是对 Web 上日志数据及相关数据的挖掘。这些数据包括：客户端数据、服务器端数据和代理端数据。Web 使用记录挖掘可分为一般存取路径追踪和专用化追踪。前者是用 KDD（Knowledge Discovery in Database，从数据库中获取知识）技术理解一般访问模式和趋势，如 Web 日志挖掘；后者是分析某一时刻每一个用户的访问模式，网站将根据这些模式自动重建结构，如自适应站点。Web 使用记录的挖掘的目的是预测用户网上的行为，比较网站的实际使用与期望的差别，根据用户的兴趣调整网站结构。

## 4 数据挖掘在搜索引擎中的应用

利用 Web 挖掘技术可以对搜索引擎中的 Web 文档处理部分进一步完善。当对搜索引擎数据库中的网页进行索引时，可以利用数据挖掘中的文本总结技术从文档中抽取出关键信息，然后以简洁的形式对 Web 文档的信息进行摘要或表示；同时利用数据挖掘中的文本分类技术把网页按照预先定义的主题类别进行分类，等等。

在搜索引擎中建立词典库，当用户给出搜索条件时，用人工智能中的自然语言处理技术对搜索条件进行分析，搜索引擎查找相应的同义词典、相关词词典等辅助词典，在数据库里进行匹配查找，以提高信息搜索的查全率。例如用户想查找有关“计算机”的信息，则搜索引擎通过查找词典，会扩展此搜索条件，把有关（计算机 or 电脑 or 微机）的信息都返回给用户。可见，加入同义词的概念，使得用户的兴趣容易得到表达，这样不仅表达准确，返回的结果比较集中，也不易漏检。

为了使搜索结果更符合用户的要求，在搜索引擎中建立用户个性化信息表。可以利用个性化页面服务的方式收集用户信息，并且个性化的服务也要求让用户可以编辑自己的显示界面，主动服务器对这些设定信息的进行分析加入到用户信息库，可以从一定程度上反映用户的偏好，将其作为个性化服务的基础。用户信息库中放置了社会时尚信息、职业与兴趣的关联规则，年龄与兴趣爱好的关联规则，等等。然后采用一定的挖掘规则（如关联规则、聚类分析、联机分析挖掘等）对这些数据进行分析，预测用户的兴趣、访问偏好。利用用户个性化信息库进行信息过滤可以提高检索精度。当关键词送给搜索引擎后，搜索引擎返回原始搜索结果，然后借助用户个性化信息表，挖掘出用户的兴趣，对原始搜索结果进行过滤，然后将用户感兴趣的信息发送给用户。

当搜索引擎在数据库中进行搜索，寻找相匹配的网页时，利用 web 内容挖掘对网页的标题、关键词、URL 和其它标记进行分析，对文档进行自动分类，提炼出重要信息形成文档摘要，使用户能够快速、方便的了解搜索出的信息。搜索到的网页进行计算相关度的大小，同时利用 web 结构挖掘找出权威页，综合对搜索结果进行排序。假设要搜索某一给定话题的 Web 页面，例如金融投资方面的页面，这时我们希望得到与之相关的 Web 页面外，还希望所检索到的页面具有高质量，即针对该话题具有权威性。权威性（authority）隐藏在 Web 页面链接中。Web 不仅由页面组成，而且还包含了从一个页面指向另一个页面的超链接、超链接包含了大量人类潜在的注释，它有助于自动推断权威性概念。当一个 Web 页面的作者建立指向另一个页面的指针时，这可以看作是作者对另一页面的认可。

把一个页面的来自不同作者的注释收集起来,就可以反映该页面的重要性,并可以很自然地用于权威 Web 页面的发现。因此,大量的 Web 链接信息提供了丰富的关于 Web 内容相关性、质量和结构方面的信息,这时利用 Web 结构挖掘可以找出权威页。

尽管如此,有时当用户进行搜索时,最后返回的结果也有很多与用户无关的内容,因为用户每次想搜索的资料毕竟也不完全相同,比如:用户上次想搜索数据挖掘技术方面的文章,而这次他想搜索数据挖掘应用方面的文章,或者下次想搜索数据挖掘软件方面的资料等等。因此,要对搜索引擎返回的结果进行聚类,从而使得在搜索引擎返回的非常大的文档列表中的过滤操作变得十分方便,这些聚类是搜索引擎返回的文档集合上的高层视图,使用户对搜索引擎结果有个一目了然的感觉,从而方便了用户浏览。搜索引擎结果聚类技术实质上就是为了方便用户的浏览,将聚类技术用于信息检索结果的可视化输出。

技术创新

## 5 结束语

将数据挖掘技术引入到网络资源的开发中来,能加快智能检索的发展,数据挖掘的结果是实现智能检索的基础。在最近的一次高级技术调查中,数据挖掘和人工智能被认为是"未来三到五年内将对工业产生深远影响的五大关键技术"之首。在学习用户的兴趣时,结合机器学习和模式识别等其它人工智能技术,研究更高效的学习算法就是深入研究的重点。

参考文献:
[1]徐宝文 张卫丰. 搜索引擎与信息获取技术[M].北京:清华大学出版社,2003.
[2]朱明.数据挖掘 [M].合肥:中国科学技术大学出版社,2002.
[3]韩家炜, Kamber M.数据挖掘:概念与技术[M].北京:机械工业出版社, 2001.
[4]陈旭春,赵明生. 分布式多搜索引擎系统的研究与实现[J]微计算机信息,2005,10:37-39
[5]Jon M Kleinberg. Authoritative Sources in a Hyperlinked Environment[Z] the Proceedings of the ACM-SIAM Symposiumon Discrete Algorithms . 1999.
[6] 杨思洛. 搜索引擎的排序技术研究 [J]. 现代图书情报技术, 2005,1:43-47.

作者简介:杨占华(1981-),男(汉族),安徽阜阳人,硕士研究生,主要研究方向:数据挖掘、计算智能、信息检索;E-mail:yzhcomp@yahoo.com.cn;杨燕(1964-),女(汉族),安徽合肥人,副教授,硕士生导师,主要研究方向:数据挖掘、模式识别。
(610031 四川成都西南交通大学信息科学与技术学院)杨占华 杨 燕
(School of Information Science & Technology, Southwest Jiaotong University, Chengdu, Sichuan, 610031) Yang,Zhanhua Yang,Yan
通讯地址:(610031 四川成都西南交通大学 320# 研(六)班) 杨占华

(投稿日期:2005.7.11) (修稿日期:2005.8.26)

(接 145 页)

```
{string UserName = FormsAuthentication.FormsCookieName; // 提取窗体身份验证 cookie
    HttpCookie authCookie = Context.Request.Cookies[UserName];
    if(null == authCookie)
    { return; }
    FormsAuthenticationTicket authTicket = null;Try
    {authTicket = FormsAuthentication.Decrypt(authCookie.Value); }
    catch(Exception ex)
    {return;}
    if (null == authTicket)
    {return; }
    string[] roles = authTicket.UserData.Split(new char[] {'|'});//提取角色
    FormsIdentity id = new FormsIdentity( authTicket ); // 创建 Identity object
    GenericPrincipal principal = new GenericPrincipal(id, roles);
    Context.User = principal;
    }
```

参考文献:
[1] 陈湘.ASP.NET 与网站开发编程实战. 清华大学出版社. 2002
[2] 龙银香. 基于移动计算的数据挖掘研究[J]微计算机信息 2005,4:216-217
[3] Jason Bell 等.ASP.NET 程序员参考手册. 清华大学出版社. 2002 年 5 月
[4] Scott Worley 著.《ASP.NET 技术内幕》.王文龙 刘湘宁译.人民邮电出版社,2002
[5] 王保健.ASP.NET 网站建设专家. 清华大学出版社. 2005 年 7 月
[6] 李兰友,杨晓光.ASP.NET 实用程序设计.清华大学出版社. 2005 年 2 月

作者简介:郭长金(1971-),男,汉族,重庆市忠县人,讲师,主要研究方向计算机软件理论。Email:cqgcj168@163.com 崔轩辉(1964-),男,汉族,陕西人,硕士,副教授,主要研究方向计算机网络。
Author brief introduction:Guo changjin (1971-),male, Han ethnic group,zhongxian of chongqing, lecturer, main research direction: software theories.Cui xuanhui(1964-), male,Han ethnic group,shanxi,master, associate professor, main research dirction:network.
(400050 重庆科技学院)郭长金 崔轩辉
(Chongqing University of Science and Technology 400050) Guo,changjin cui,xuanhui
通讯地址:(400050 重庆科技学院电子信息工程学院)郭长金

(投稿日期:2005.8.5) (修稿日期:2005.9.16)