●晏创业　张玉峰

智能检索中的网络数据挖掘技术探索 *

摘　要　缺少一种深入信息内容的网络信息搜索工具,是实现智能检索的主要困难。采用网络数据挖掘技术是解决问题的关键。从智能信息检索的角度出发,主要考虑从网络信息内容的关联度来挖掘网络数据。图 1。表 1。参考文献 7。

关键词　智能检索　网络信息检索　数据挖掘
分类号　G252.7

ABSTRACT　The authors think that the major difficulty in realizing intelligent search is the lack of a network information search tool reaching information contents, and the key to solve the problem is the network data mining techniques. They also discuss the relevance of network information contents for the mining of network data. 1 fig. 1 tab. 7 refs.

KEY WORDS　Intelligent search. Network information search. Data mining.
CLASS NUMBER　G252.7

智能检索能帮助人们在开发网络信息资源时做到"取其精华,去其糟粕"。它能摆脱表层信息的干扰,从信息内容的角度出发,搜索出高质量的信息。目前,人们对信息检索过程中的智能化要求主要体现在基于内容的检索、个性化信息检索和知识检索。

因特网上的信息资源不同于一般意义上的数据库,除具有开放性、异构性和分布性等特点外,还具有半结构化、非结构化的动态关联特性。网络信息的特点决定了我们不能像对待静态结构化的数据库信息那样来对待它。然而,当前的一些网络信息搜索工具仍遵循了大型数据库的信息检索思想,即对网页的标题、URL 等表征信息和没有进行深度分析的关键词进行标引,然后建立网络信息的倒排文档,将它们简单地聚合在一起。这种以数据库信息处理方式组织起来的信息源,在检索中主要有 3 个弊端:一是同一关键词检索出来的信息"貌合神离";二是检索结果中出现大量的冗余信息;三是因为信息用户理解差异的存在,在使用某些检索词时根本就检索不到任何信息。

基于内容的检索和个性化的信息检索,都是建立在网络信息内容基础之上的,真正的知识性,是从对信息内容的深层挖掘中体现出来的。面对因特网上源源生成的信息,我们需要一种大批量、深入内容的信息处理技术,使其按照内容特性聚集,并体现一定的知识性。将最初面向数据库的数据挖掘技术引入到因特网中,是解决问题的关键。

1　面向因特网的数据挖掘

1.1　数据挖掘的相关知识

数据挖掘(Data Mining)是从大量的、不完全的、有噪声的、随机的数据中,提取潜在有用的信息和知识的过程。数据挖掘源自人工智能的机器学习领域,是在一个已知状态的数据集上,通过设定一定的学习算法,从数据集中获取所需的知识。这些知识能够用于信息管理、智能查询、决策支持、过程控制以及其他方面。

数据挖掘的最初对象是一些大型的商业数据库,它通过描述数据、计算统计变量(比如平均值、均方差等),并将这些变量用图表直观地表示出来,进而找出数据变量之间的相关性,即发现知识,以提供解决问题的依据。随着数据挖掘技术在商业数据库中的成功应用,它又被迅速移植到电信、医疗保险等领域,因特网的出现为它提供了一个更为广阔的用武空间。借用数据挖掘的原理来实现网络数据的深层挖掘,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。

1.2　网络数据挖掘模型设计

网络数据有不同于一般数据库中数据的特点:异构和半结构化。因特网上的每一个站点都是一个数据源,每一个数据源都有自己的设计风格,即每个站点的信息和组织都不一样,因特网就是一个巨大的异构数据库,不同于传统的关系数据库。因特网上的数据非常复杂,没有统一的模型描述,这些数据虽有一定的结构性,但因自述层次的存在和复杂的相互关联,因而是一种非完全结构化的数据。鉴于网络数据的这些特点,我们在将数据挖掘技术引入因特网的时候,必须要做一定的预处理工作。在此基础上的网络数据挖掘模型如图 1 所示。

*　本文系国家社科基金资助项目(编号:01BTQ011)的研究论文。

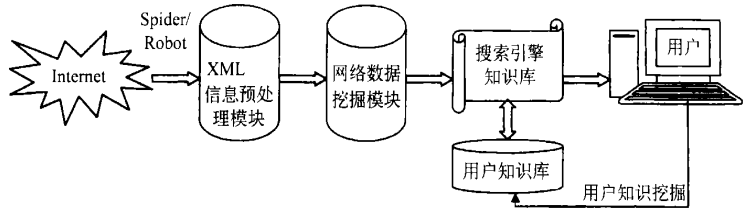

图 1　网络数据挖掘模型

图 1 实际上是一个对传统搜索引擎改进的模型,将数据挖掘技术加载到搜索引擎上,从而实现网络数据的挖掘。

XML 信息预处理模块是实现网络数据挖掘的关键环节,但若遵循常规将 Spider 或 Robot 采集的异构网络数据,按照一种统一的数据结构重新组合,将会因工作量太大而难以实现。XML (Extensible Markup Language) 的出现为解决网络数据挖掘难题带来了希望。XML 是一种简单、开放、高效且可扩充、符合国际标准的网络标记语言,它的扩展性和灵活性使其能描述不同种类应用软件中的数据,从而描述搜集的网页中的数据记录。用 XML 标引的网络数据是一种半结构化的数据模型,但其文档描述可与关系数据库中的属性一一对应起来,从而能实施精确的查询与模型抽取,完成异构网络数据的整合工作。

2　网络数据挖掘的算法及其对智能检索的支持

2.1　算法分析

数据挖掘技术已在许多领域得到成功应用,针对不同领域的数据类型,人们采用了不同的挖掘算法,目前比较成熟的算法有关联规则、聚集、分类、神经网络和决策树等几种。网络信息包罗万象,所有的这些算法在网络数据挖掘中都能派上用场。从智能信息检索的角度出发,主要考虑的是从网络信息内容的关联度来挖掘网络数据,所以这里主要分析关联规则和分类两种挖掘算法。

2.1.1　关联规则

在个性化信息检索中,用户对信息的需求是强调兴趣信息的相关性,而非相异性,根据用户信息需求心理确定适当的规则,可提高数据挖掘的准度和深度。关联规则算法的一般流程是:

(1) 选取合适的元素。要根据不同的统计级别,选择相应的细节程度,细节的颗粒越细,预处理结果的实施性越好,但相应的工作量也就越大。根据选取的元素,建立其属性元组:itemname(Itemid 子类编号,itemname 子类名称,Superitem 父类编号)。

(2) 确立挖掘规则。所谓的规则就是一个条件和一个结论的和,即 If *condition* then *result*。它是组织相关内容的有效方法。衡量一个规则好坏的标准有三个:支持度、可信度和提高率(兴趣度)。下面以一个例子分别对其说明。用户对旅游信息感兴趣的信息有两类:旅游城市信息(C)和航班信息(F),表 1 给出用户关心这两类信息的百分比。

表 1　挖掘规则举例

信息元组	信息点击率(%)
C	45
F	40
C and F	20

支持度是指一个元组被点击的频率,比如 C 的支持度 S(C) = 45 %;可信度是相对于规则而言的,对于一般的规则,它的可信度 = p (condition and result) / p (condition)。比如对于规则 if C then F, 其可信度为 p(if C then F) / p (C) = 20 % ÷45 % = 44.4 %。单纯从支持度和可信度往往还不能判别一个规则的优劣,所以还要引入提高率(兴趣度)的概念,即兴趣度 = p (condition and result) / p (condition) ×p (result)。兴趣度大于 1 的规则为优,而兴趣度小于 1 的规则无实际意义。规则是根据这三个指标综合分析,最终确立的。

(3) 加入分裂规则,以克服现实中数据爆炸的问题。

(4) 加入时标或序列信息,以确立信息的时序性。

(5) 标识信息,用以区别不同的事务。

2.1.2　分类

分类就是要达到"物以类聚"的目的。在智能信息检索中,我们强调的"类"是指信息元组的内容属性。网络数据挖掘达到了从内容属性分析信息元组的深度,其分类模式是一个分类函数(分类器),能够把数据集中的数据项映射到某个给定的类上。

分类的主导思想是通过比较原始数据集与分类器中的训练集的相似度(距离),来确立原始数据集的最后归属。训练集是由历史数据构成的,它是分类的标准,由一组数据库记录或元组构成,每个元组就是一个特征向量,可表示为 (v1, v2, …, vn; c),其中 vi 表示字段值,c 表示类别。训练集必须能很好地覆盖原始数据,即原始类别要尽可能的全。此外,各个类的数量应该大体一致,以便于原始数据集相似度的计算和比较。在确定训练集的基础上,还要确定距离函数,以此计算原始数据与训练集中某一元组的相似度。假设训练集中的某一元组为 A,有一原始数据集 B,那么 A、B 间的距离可用函数 d(A, B) = 1 - score(A, B) / score(A, A) 计算出来。在计算出原始数据集相对于训练集中某一元组的距离之后,我们根据距离的大小给定一个权值,距离越大,表示该数据集与该元组内容的相关性越小,权值也就越小。在计算出一系列的权值之后,就可以根据权值的大小将该数据集归并到相应的类中。

2.2　网络数据挖掘对智能检索的支持

无论是个性化信息检索,还是基于内容的检索,乃至知识检索,网络信息检索系统智能化的关键是知道用户需求什么、提供给用户的东西具有高质量(内容上相关、知识含量高)。数据挖掘对网络智能检索的支持正体现在对用户需求信息和网络源信息的深层分析,以提供智能检索必需的关键知识。

2.2.1　用户知识的挖掘

虽然个体信息用户的信息需求具有特定性,但从用户群整体来看,用户的信息需求又是随机的,这为一般的用户需求信息分析带来很大的困难。数据挖掘从全局出发,以丰富、动态的联机查询和分析来了解用户的信息需求。通过在线提问、调查表等方式,系统可以获取关于用户的用户名、用户访问 IP 地址、用户的职业、年龄、爱好等原始信息。然后,采取一定的挖掘规则(如关联规则、联机分析处理等),对这些数据进行融合分析,其结果是为每个用户建立一个信息需求模型。而且全方位的用户需求信息挖掘,可以将同类信息需求的用户联系起来,从而实施"以一对众"的检索方案。目前用户知识挖掘已步入实用阶段,如 IBM 公司新推出的 DB2 UDB 7.1 就是一种比较理想的用户知识挖掘工具。

2.2.2　网络知识的挖掘

网络知识的挖掘就是要在具有极度不确定性的海量数据中找出信息分布的规律,挖掘隐藏的信息并形成模型,从而发现具有规律性的知识。网络信息分布的规律性就是网络信息内在的关联性。对网络信息这种关联性的挖掘主要体现在两个方面:一是网络内容挖掘。网络内容挖掘是一个从文本、图像、音频、视频、元数据等形式的网络源信息中采用分类、聚类等形式的挖掘方法,发现有用信息,并将这些信息按满足某种检索方式的形式加以组织的过程。二是网络结构挖掘。网络结构挖掘是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。这种模式可以用于网页归类,并且由此可以获得有关不同网页间相似度及关联度的信息。这种链接结构模式有利于实现智能导航。

3　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,智能检索的结果可为数据挖掘提供指南和线索。目前,以数据挖掘技术为主导的网络数据挖掘产品已得到应用。例如,Net Perceotion 公司开发的 Net percerptions,能够挖掘用户信息,从而为实现个性化信息服务打下基础。在开发无尽的网络数据资源时,若能结合机器学习、模式识别等其他人工智能技术,网络数据挖掘技术将在实际应用中更加完善。

参考文献

1　毕强,闫凤英等. Web 信息发布的特征与完全信息结构. 情报学报,2000(6)
2　李爱红. 网络搜索引擎的比较研究. 中国信息导报,1999(1)
3　王军,杨冬青等. 数字图书馆的检索技术. 计算机世界,2000-02-21
4　邹涛,戚广智等. 网络信息挖掘系统 IDGS 的实现. 南京大学学报(自然科学版),2000(2)
5　朱建秋,周皓峰等. 一个基于关联规则的数据采掘工具的设计和实现. http://dmgroup.myetang.com/lw3.htm
6　Hearst. Distinguishing between Web Data Mining and Information Access. Presentation for Web Data Mining Panel at KDD'97
7　Dan R. Greening. Data mining on the web. Web Techniques, 2000(1)

晏创业　武汉大学信息管理学院。通讯地址:湖北武汉。邮编 430072。
张玉峰　武汉大学信息管理学院教授。通讯地址同上。
(来稿时间:2001-10-18)

《永乐大典》编纂 600 周年国际研讨会暨仿真影印出版首发式举行

4 月 17 日,由国家图书馆主办的《永乐大典》编纂 600 周年国际研讨会暨《永乐大典》仿真影印出版首发式在京举行。全国政协副主席罗豪才到会表示祝贺。来自中国、美国、英国、法国、澳大利亚、俄罗斯、日本、韩国等海内外 50 余个研究机构和收藏单位的 90 余位专家学者参加了会议。

《永乐大典》是明成祖朱棣于明永乐元年(公元 1403 年)命太子少师姚广孝和翰林学士解缙主持、3000 多文臣历时 4 年纂修而成。共辑录图书 8000 种,上自先秦,下迄明初,天文地理,人事名物,无所不包。整部典籍共 22877 卷,外加目录等 60 卷,装成 11095 巨册,全部用毛笔工楷书写,是世界上最早、最宏伟的百科全书。目前,存世的《永乐大典》副本零册约 400 册左右,约 800 余卷,不到原书的 4%,分散在 8 个国家和地区。《永乐大典》是中国国家图书馆的四大珍藏之一,从 1912 年第一批《永乐大典》入藏到现在,已拥有 221 册,超过全球藏量的半数,居各处收藏的首位。

在《永乐大典》编纂 600 周年国际研讨会上,来自世界各地的专家、学者就各国收藏、保护、研究《永乐大典》的状况,《永乐大典》的修复、保护、数字化和出版情况进行了广泛而深入的讨论。相信本次会议必将对《永乐大典》的收藏、保护、研究、出版、数字化起到极大的推动作用。

近年来,《永乐大典》在世界各地又续有发现,而这些新发现的《永乐大典》也急需刊布,以嘉惠学林,使其得到充分利用,因此,中国国家图书馆委托北京图书馆出版社将现存我馆的 161 册连同现存国内其他馆收藏的 2 册《永乐大典》率先仿真影印出版。

该社从 2001 年 12 月开始,用特制宣纸,套色印刷,原大仿真分批出版现存于世的《永乐大典》。拟用一年半时间先首批出版收藏于中国大陆的 163 册的大陆珍藏版。待首批出完,再用一年半的时间陆续出版现藏于海外的 200 余册,使之为学界、为大众所共享。