计算机科学2003Vol.30№.1

基于 Web 的文本挖掘系统的研究与实现*)

唐 菁[1] 沈记全[2] 杨炳儒[1]

(北京科技大学信息工程学院 北京100083)[1] (河南省焦作工学院计算机科学系 河南焦作454000)[2]

The Research and Development of Text Mining System Based on Web

TANG Jing[1] SHEN Ji-Quan[2] YANG Bing-Ru[1]

(The University of Science and Technology Beijing, Information Engineering School, Beijing 100083)[1]

(Jiaozuo Institute of Technology, Computer Science Department, Henan Jiaozuo 454000)[2]

E-mail:Bradtang@263.net

Abstract With the development of network technology, the spread of information on Internet becomes more and more quick. There are many types of complicated data in the information ocean. How to acquire useful knowledge quickly from the information ocean is the very difficult. The Text Mining based on Web is the new research field which can solve the problem effectively. In this paper, we present a structure model of Text Mining and research the core arithmetic —— Classification arithmetic. We have developed the Text Mining system based on Web and applied it in the modern long-distance education. This system can automatically classify the text information of education field which is collected from education site on Internet and help people to browser the important information quickly and acquire knowledge.

Keywords Text mining based on Web, Text classification, Information navigation

1. 引言

60年代,大的物理流伴随着大信息流。传统的文件方式不能适应信息处理的需求,因此出现了数据库技术。90年代,人类积累的数据量以高于每月15%(或每年5.3倍)的速度增加,但是数据海洋不能产生决策意志,为了进行决策,人们不断地扩大数据库能力,搜集海量数据,但这使得决策者更难于决策,因此出现了数据挖掘技术,以便从数据库中发现知识。数据挖掘技术包括特征、分类、关联、聚类、偏差、时间序列、趋势分析等。

近年来,Internet 正以令人难以置信的速度在飞速发展,在信息时代最显著的特征之一就是信息的产生、传播速度更加迅速,信息的交流量日益增加。目前网络上普遍存在着"信息爆炸"的问题,即信息极大丰富而知识相对匮乏。据美国基础软件开发商英克托米(inktomi.com)今年1月18日公布的意向研究结果表明,因特网上可编索引的网页已超过10亿,如果加上大量无法索引的页面,因特网网页的总数则更多。所以 Web 已经发展成为拥有10多亿页面的分布式信息空间。现在越来越多的机构、团体和个人在 Internet 上发布信息、查找信息。

为了从大量数据的集合中发现有效、新颖、潜在有用的、可理解的模式。在数据库领域采用了数据挖掘技术。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。Forrest Research 的统计资料指出,在 Internet 和 Intranet 中80 %以上的数据都是以非结构化的形式存在,如技术报告、技术文档、Email、专家陈述等。因此,对这些非结构化的信息进行知识发现,难度将会更大,但其意义也将更加重大[1~4]。而对于这些文本信息的分析和处理主要是结合了文本挖掘(Text Mining)的方法和技术。近年来,文本挖掘技术应用越来越广泛,它可以同搜索引擎、信息推送、信息过滤等信息处理技术相结合,有效地提高了信息服务的质量,即提高了信息的含金量。

本论文将深入地探讨文本挖掘系统的结构模型,并研究文本挖掘的核心算法——分类算法,同时结合现代远程教育这个应用背景实现了基于 Web 的文本挖掘系统。该系统可以对各类远程教育站点中收集的文本资料信息自动进行分类,从而更好地帮助人们进行文本信息导航,获取重要的知识。

2. Web 文本挖掘系统的研究

1. Web 文本挖掘的定义

Web 文本挖掘是一门交叉性学科,涉及数据挖掘、机器学习、模式识别、人工智能、统计学、计算机语言学、计算机网络技术、信息学等多个领域。

Web 文本挖掘是指借鉴数据挖掘的基本思想和理论方法,从大量非结构化、异构的 Web 文档的集合 D 中发现有效的、新颖的、潜在可用的及最终可理解的知识 K(包括概念(Concepts)、模式(Patterns)、规则(Rules)、规律(Regularities)、约束(Constraints)及可视化(Visualizations)等形式)的非平凡过程。如果将 D 看作输入,将 K 看作输出的话,那么 Web 文本挖掘的过程就是从输入到输出的一个映射 ξ:D→K。

这个描述性的定义是借鉴了数据挖掘的定义而给出的,在这里我们将解释几个概念:过程通常是指多阶段的一个过程,涉及数据预处理、学习与知识模式的生成、模型质量的评价以及反复的修改求精;该过程要求是非平凡的,意思是要有一定程度的智能化、自主性(仅仅给出所有数据的总和不能算作是一个发现过程)。而以上所提及的有效性、新颖性、潜在有用性和最终可理解性综合在一起可称为兴趣性(Interestingness)。

2. Web 文本挖掘系统的结构模型

*)获国家教育部科技重点项目(教技司[2000]175)资助。杨炳儒 教授,博士生导师,主要研究方向:推理机制与知识发现,柔性建模与集成技术。唐 菁 博士,主要研究方向:Web 挖掘,智能信息检索。

下面我们将给出 Web 文本挖掘系统的整体结构模型，如图1所示。

该文本挖掘系统的工作流程为：

1)特征提取：对 Web 上采集到的挖掘目标样本进行特征提取，采用空间向量模型、潜在语义索引和小波分析方法生成挖掘目标的特征矢量；同时应根据特征项集选取的两个基本原则即完全性和区分性原则进行特征项集的选取，并将提取得到的特征矢量经过特征子集的选取后存放到文本特征库中形成文本中间表示形式。

2)文本挖掘过程：对于 Web 文本的中间表示形式采用向量空间的距离测度分类算法进行分类挖掘处理，也可以结合聚类和关联挖掘算法。最终得到潜在的知识或者模式。

3)模型质量评价：将挖掘得到知识或者模式进行评价，将符合一定标准的知识或者模式呈现给用户。其中使用的评价指标主要是：查全率(Recall)和查准率(Precision)。

4)信息表示和信息导航：将反馈的结果用可视化的方式进行显示，同时对用户提供信息导航功能，从而在极大的程度上方便用户有效地浏览和获取信息。

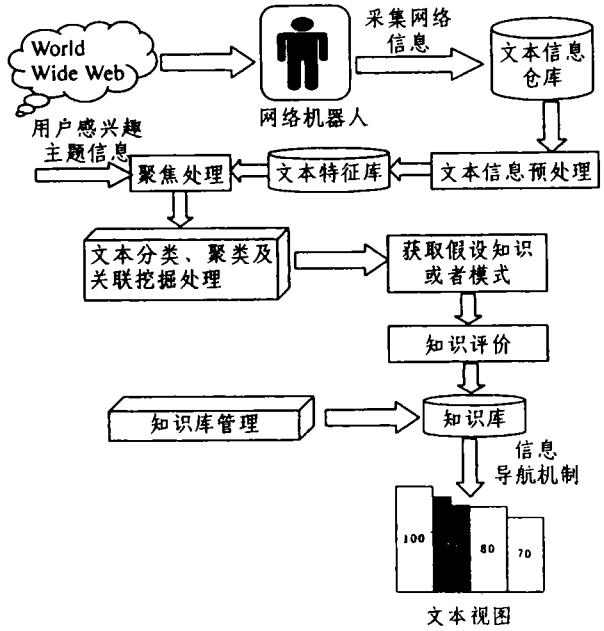

图1

3. 关键技术研究

1)文本的特征表示　与数据库中的结构化数据相比，Web 文档具有有限的结构，或者根本就没有结构。文本信息源的这些特征使得现有的数据挖掘技术与方法无法直接应用于其上。我们需要对文本进行预处理，抽取其特征并用结构化的形式保存，作为文档的中间表示形式。目前，结构化标记语言 XML 能够对 Web 文档资源进行描述。这将有利于 Web 文档的信息抽取。

特征表示是指以一定的特征项(如词条或描述)来代表文档信息，特征表示模型有多种，常用的有布尔逻辑型、向量空间型、概率型等。近年来应用较多且效果较好的特征表示法是向量空间模型(Vector Space Model，VSM)法[5]。

在 VSM 中，将文本文档看成是一组词条$(T_1,T_2,\cdots,T_n)$构成，对于每一词条 T_i，都根据其在文档中的重要程度赋以一定的权值 W_i，我们可以将其看成一个 n 维坐标系，$W_1,W_2,\cdots,W_n$ 为对应的坐标值，因此每一篇文档都可以映射为由一组词条矢量张成的向量空间中的一点，对于所有待挖掘的文档都用词条特征矢量$(T_1,W_1,T_2,W_2,\cdots,T_n,W_n)$表示。这种向量空间模型的表示方法优点在于将非结构化的文本表示为向量形式，使得各种数学处理成为可能。

2)文本的特征子集的选取　目前对 WWW 文档特征所采用的特征子集选取算法一般是构造一个评价函数，对特征集中的每一个特征进行独立的评估，这样每个特征都获得一个评估分，然后对所有的特征按照其评估分的大小进行排序，选取预定数目的最佳特征作为结果的特征子集。特征子集选取的好坏将直接影响后续的分类挖掘处理。所以该过程一直来都是文本挖掘系统的关键所在。

一般采用的评估函数有词频(Word Frequency)、信息增益(InformationGain)、期望交叉熵(ExpectedCross Entropy)、互信息(Mutual Information)、文本证据权(the Weight of Evidence for Text)、奇率比(Odds ratio)等[6]。在实际开发过程，使用比较广泛的是互信息和信息增益等评估函数。

3)文本的挖掘过程　目前文本的分类算法有很多种，例如，朴素贝叶斯方法(Naive Bayesian)、K-最近邻参照分类算法(K-Nearest Neighbor)、向量空间距离测度分类法(Distance Measure)、支持向量机分类法(Support Vector Machine)和神经网络分类法(Neural Network)等[7,8]。下面我们将介绍本系统实现的一种具有良好分类效果的算法——向量空间距离测度分类法。

我们知道文本分类是一种典型的有教师的机器学习问题，一般分为训练和分类两个阶段，该算法的具体过程如下：

训练阶段：

(1)首先定义类别集合 $C=\{c_1,\cdots,c_i,\cdots,c_m\}$这些类别可以是层次式的，也可以是并列式的；

(2)然后给出训练文档集合 $S=\{s_1,\cdots,s_i,\cdots,s_n\}$，每一个训练文档都被标上所属的类别标识 c_i；

(3)提取训练文档集合 S 中所有文档的特征矢量 $V(s_j)$，并采用一定的原则来确定代表 C 中每一个类别的特征矢量 $V(c_j)$；

分类阶段：

(1)对于测试文档集合 $T=\{d_1,\cdots,d_k,\cdots,d_r\}$中的每一个待分文档 d_k，计算其特征矢量 $V(d_k)$与每一个 $V(c_j)$之间的相似度 $sim(d_k,c_j)$。最常用的方法就是考虑两个特征矢量之间的夹角的余弦，即$=sim(d_k,c_j)=\dfrac{V(d_k)\cdot V(c_i)}{|V(d_k)|\times|V(c_i)|}$；

(2)选取相似度最大的一个类别 $\arg\max_{c_i\in C} sim(d_k,c_i)$作为 d_k 的类别。

如果 d_k 与所有的类别的相似度均低于阈值，那么通常将该文档放在一边，由用户来做最终的决定。当经常出现类别与预定义类别不匹配的文档时，则说明需要修改预定义类别，然后再重新进行上述训练与分类过程[4]。

4)模型质量的评价　对 Internet 上的文本数据进行文本挖掘可以看作是一种机器学习的过程。在机器学习中的学习的结果是某种知识模型 M，机器学习的一个重要组成部分便是对产生的模型 M 进行评估。在机器学习中常用的模型质量评估指标有分类正确率(Classification Accuracy)，查准率(Precision)与查全率(Recall)[7]，查准率与查全率的几何平均数，信息估值(Information Score)，兴趣性(Interestingness)。其中兴趣性是一个主客观结合的评价指标。

查全率$=\dfrac{\text{分类的正确文本数}}{\text{应有的文本数}}$，它是人工分类结果应有的文本中分类系统吻合的文本所占的比率。

查准率$=\dfrac{\text{分类的正确文本数}}{\text{实际分类的文本数}}$，它是所有判断的文本中与

人工分类结果吻合的文本所占的比率。

5)知识模式的信息导航机制　基于 Web 的文本挖掘系统最终挖掘出来的知识或者模式信息如果能够用可视化的方式进行显示，同时对用户提供信息导航的功能，那么将在极大的程度上方便用户有效、快速地浏览和获取信息。鉴于该目的的考虑，该文本挖掘系统中将提供信息表示和信息导航功能。信息导航的原则是提供给用户简明、多视角的方法。

通过使用可视化图形界面的信息表示技术和信息导航技术，用户将能够更快地接受信息并根据自己的兴趣度对所反馈的挖掘结果进行有目的的查询和浏览。本系统中主要使用了树状结构信息导航机制和图形结构信息导航机制。

3. Web 文本挖掘系统的实现

1. Web 文本挖掘系统的应用

我们承担了国家教育部重点项目——智能化、个性化的现代远程教育系统中关键技术研究，因此我们在以上提出的 Web 文本挖掘的结构模型及关键技术研究的基础上构建了一个适用于现代远程教育的文本挖掘系统。它能够充分利用 Web 站点(远程教育站点)上积累的丰富文本信息，更好地服务于远程教育。

该系统的主界面如图2所示，它主要包括文本预处理，文本挖掘处理，文本信息导航三个组成部分。

图2　Web 文本挖掘系统的主界面

经过文本预处理、文本分类挖掘处理后，对文本信息的分类结果进行导航，其挖掘结果显示如图3所示。

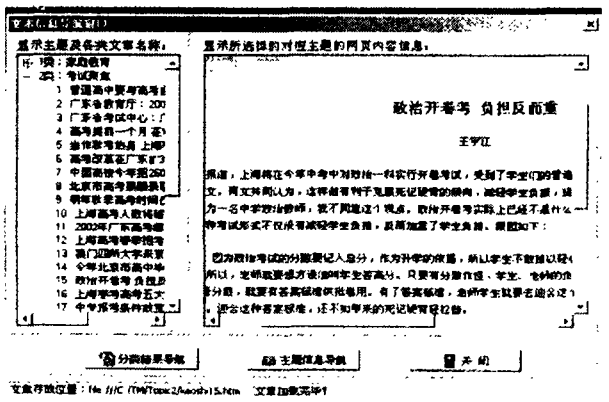

图3　分类挖掘结果导航界面

2. 测试数据和实验结果

本系统的测试数据集是一个具有500篇中文文本的语料库，语料库中的文本绝大多数采自远程教育站点的教育类热点新闻，为了结合我们所承担的国家教育部重点项目的研究，我们将教育类新闻分为四个类别，即：家庭教育、考试聚焦、素质教育、教育热点。利用测试数据集中的文档测试自动文本分类模型，从而获取了其性能测试结果，如表1所示。

整个测试语料库集合中没有被分类的文本为未知类别的文本，由于这类文本的数量不多，可以通过人工方式进行判断，以确定其类别。

本系统采用的分类挖掘算法是向量空间的距离测度分类法，该算法与朴素贝叶斯分类算法、K-最近邻参照分类算法相比而言，主要是分类的算法容易实现，分类速度也比较迅速。其时间复杂度为 O(cn)，其中 c 是类别数目，n 是抽取的特征项维数。从该算法分类结果来看，向量空间的距离测度分类法的查准率和查全率都比较高，表现出来的算法性能比较好。

表1　测试数据集的实验结果

类别	家庭教育	考试聚焦	素质教育	教育热点	合计
正确分类数	101	104	97	102	404
实际分类数	115	123	124	130	492
应有文本数	118	125	120	132	495
查准率	87.8%	84.6%	78.2%	78.5%	82.1%
查全率	85.6%	83.2%	80.8%	77.3%	81.6%

结束语　本文主要介绍了 Web 文本挖掘的研究与实现，提出了一个基于 Web 的文本挖掘系统的结构模型；并结合现代远程教育中关键技术研究的应用背景构建了一个 Web 文本挖掘系统原型。它主要实现了对于教育类信息进行分类挖掘，同时对分类结果进行信息导航，指导用户快速地浏览信息，以获取其所需要的知识或者模式。

随着网络技术的迅猛发展，Web 文本挖掘在今后无疑将具有广泛的应用前景。但是目前文本挖掘技术尚处于初期发展阶段，IBM 公司曾在其一份技术白皮书中指出[1]：目前的大多数文本挖掘工具只实现了对于信息进行抽取和分类等基础工作，离其真正的目标还有一段距离。因此，我们将在今后的研究工作中更加注重对于系统整体结构模型、挖掘算法的完善和对于挖掘所形成知识的可视化工作。

参 考 文 献

1　韩客松，王永成. 文本挖掘、数据挖掘和知识管理——二十一世纪的智能信息处理. 情报学报，2001，20(1)：100～104

2　Feldman R，Dagan I. Knowledge discovery in textual database (KDT). In : Proc. of 1st Intl'l Conf. on Knowledge Discovery and Data Mining. Montreal，Canada，1995

3　Salton G. Automatic Text Processing. Reading，MA : Addison-Wesley，1989

4　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(5)：513～520

5　邹涛，王继成，朱华宇，金翔宇，张福炎. WWW 上的信息挖掘技术及实现. 计算机研究与发展，1999，36(8)：1019～1024

6　王实，高文，段立鹃. Internet 上的文本数据挖掘. 计算机科学，2000，27(4)：32～36

7　庞剑锋、卜东波，白硕. 基于向量空间模型的文本自动分类系统的研究与实现. 计算机应用研究，2001，10(9)：23～26

8　张月杰、姚天顺. 基于特征相关性的汉语文本自动分类模型的研究. 小型微型计算机系统，1998，19(8)：49～55