

基于 Web 的文本挖掘技术研究

许高建^{1,2}

(1. 合肥工业大学 计算机与信息学院, 安徽 合肥 230009;

2. 安徽农业大学 信息与计算机学院, 安徽 合肥 230036)

摘要: Internet 上大多数信息的表现形式为文本, 如何在浩瀚的文本信息中挖掘到潜在的知识是一个有待解决的问题。文本挖掘的目的是从不同格式的文本中发现有用的知识, 这是一个分析文本并从中抽取特定信息的过程。系统地介绍了文本挖掘的含义, 并对文本挖掘过程的各个方面进行了进一步的探讨, 包括文本特征的建立、特征的提取技术、文本的分类、文本的聚类等相关技术。同时提出了一种基于 Web 的文本信息挖掘的模型, 将以高校 BBS 论坛为信息源, 利用高级语言开发技术来构建一个自动的文本分类器。

关键词: Web 挖掘; 文本挖掘; 文本分类; 文本聚类

中图分类号: TP393

文献标识码: A

文章编号: 1673-629X(2007)06-0187-04

Research on Text Mining Techniques Web - Based

XU Gao-jian^{1,2}

(1. School of Computer & Information, Hefei University of Technology, Hefei 230009, China;

2. School of Information & Computer, Anhui Agricultural University, Hefei 230036, China)

Abstract: Most information on Internet are text formatting. How to find the potential knowledge from the immensity text information is an awaiting to be settled question, which is the purpose of text mining. This is a process for analyzing text and getting the customizing messages from them. This paper introduces what is the text mining systemically, and it also further discusses the aspects involved in text mining process, including text architecture construction, feature mining, text categorization, text clustering etc. And a text mining model based on Web is presented. And will build an automatic text classification system on BBS by using programming language.

Key words: Web mining; text mining; text categorization; text clustering

0 引言

近年来, 互联网技术飞速发展, Internet 已经成为世界上最大的信息积聚地。这些巨量的 Web 信息数据中, 蕴涵着巨大潜在价值的知识。Internet 上的信息, 是以网页形式存放的, 而网页的内容又多以文本方式来表示, 但它们的结构更加复杂, 风格多样, 构成了一个异常庞大的具有异构性、开放性的分布式数据库^[1]。如何快速地、有效地从 Web 上获取有用的知识, 已经成为当今热门的研究方向。文本挖掘^[2]是一个非常活跃的研究领域, 是近几年来数据挖掘领域的一个分支。所以, 文本挖掘既采用了很多传统的数据挖掘技术^[3], 又有自己的特性。

1 文本挖掘技术的含义

1.1 数据挖掘

数据挖掘(DM, Data Mining)是从大量的、不完全的、有噪声的、随机的实际应用数据中采掘出隐含的、先前未知的、对决策有潜在价值的知识和规则的过程, 是知识发现最关键的步骤。数据挖掘的第一步是要确定挖掘的任务, 如进行数据总结、分类、聚类、关联规则发现、特征与偏差、时序模式发现、趋势分析等, 然后才能决定使用何种挖掘算法。选择合适的挖掘算法^[1]包括选取合适的模型和参数, 并使得知识发现算法与整个知识发现的评判标准相一致。

但是数据挖掘的主要对象是结构化的数据仓库^[4](Data Warehouse), 对于 Web 上的异质、非结构化信息, 并不能直接应用数据挖掘的技术。

1.2 文本挖掘

文本挖掘(TM, Text Mining)是近几年来数据挖掘领域的一个新兴分支。其基本思想是: 首先利用文

收稿日期: 2006-08-25

基金项目: 安徽省高校省级自然科学基金项目(2006KJ168B)

作者简介: 许高建(1974-), 男, 安徽肥东人, 讲师, 研究方向为计算机应用、文本挖掘。

本切分技术,抽取文本特征,将文本数据转化为能描述文本内容的结构化数据,然后利用聚类、分类技术和关联分析等数据挖掘技术,形成结构化文本,并根据该结构发现新的概念和获取相应的关系。

1.3 Web 文本挖掘

Web 文本挖掘是以 Web 文本文档为对象的一种数据挖掘技术,是一门交叉性学科^[5],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 挖掘是指从大量非结构化、异构的 Web 信息资源中发现有效的、新颖的、潜在可用的及最终可理解的知识(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^[6]。

2 Web 文本挖掘的方法

2.1 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无法直接应用于其上。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^[7]。这将有利于 Web 文档的信息抽取。

特征表示^[8]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。在 VSM 中,将每个文本文档 d 看成是一组词条(T_1, T_2, \dots, T_n)构成,对于每一词条 T_i ,都根据其在文档 d 中的重要程度赋予一定的权值 W_i ,可以将其看成一个 n 维坐标系, W_1, W_2, \dots, 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($T_1, W_1(d), T_2, W_2(d), \dots, T_n, W_n(d)$)表示。这种向量空间模型的表示方法,可以将 d 中出现的所有单词作为 T_i ,也可以将 d 中出现的所有短语作为 T_i ,从而提高特征表示的准确性。 $W_i(d)$ 一般被定义为 T_i 在 d 中出现率 $tf_i(d)$ 的函数,即 $W_i(d) = \Psi(tf_i(d))$ 。常用的 Ψ 有:

$$* \text{ 布尔函数: } \Psi = \begin{cases} 1, & tf_i(d) \geq 1 \\ 0, & tf_i(d) = 0 \end{cases}$$

$$* \text{ 平方根函数: } \Psi = \sqrt{tf_i(d)}$$

$$* \text{ 对数函数: } \Psi = \lg(tf_i(d) + 1)$$

$$* \text{ TFIDF 函数: } \Psi = tf_i(d) \times \lg \frac{N}{n_i},$$

N 为所有文件的数目, n_i 为含有词条 T_i 的文件数目。

2.2 文本的特征子集的选取

构成文本的词汇数量是相当大的,因此表示文本的向量空间的维数也相当大,可以达到几万维,因此需要进行维数压缩的工作。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。

一般用的评估函数^[9]有几率比(Odds ratio)、信息增益(Information Gain)、期望交叉熵(Expected Cross Entropy)、互信息(Mutual Information)、词频(Word Frequency)等。

2.3 Web 文本分类

试图对 Web 上的所有文档进行分类是不可行的,这里提供的分类方法更适用于对特定的专业领域的 Web 文档进行分类。文本分类是一种典型的有指导机器学习问题,一般分为训练和分类两个阶段^[10],具体过程如下:

●训练阶段:

(1) 根据该专业领域已有的分类体系,事先确定类别的集合 $C = \{c_1, \dots, c_i, \dots, c_m\}$,这些类别可以是层次式的,也可以是并列式的;

(2) 选择适量具有代表性的 Web 文档,给出训练文档集合 $S = \{s_1, \dots, s_j, \dots, s_n\}$;

(3) 对于 S 中的每个训练文档 s_j ,确定其所属的类别 c_i ;

(4) 抽取训练文档 s_j 的特征,得到特征向量 $V(s_j)$;

(5) 统计 S 中所有文档的特征矢量 $V(s_j)$,以此确定代表 C 中每个类别的特征矢量 $V(c_i)$;

●分类阶段:

(1) 对于测试文档集合 $T = \{d_1, \dots, d_k, \dots, d_r\}$ 中的每个待分类文档 d_k ,计算其特征矢量 $V(d_k)$ 与每个 $V(c_i)$ 之间的相似度 $\text{sim}(d_k, c_i)$;

(2) 选取相似度最大的一个类别作为 d_k 的类别。有时也可以为 d_k 指定多个类别,只要 d_k 与这些类别之间的相似度超过某个预定的阈值。如果 d_k 与所有类别的相似度均低于阈值,那么通常将该文档放在一边,由用户来做最终决定。对于类别与预定义类别不匹配的文档而言,这是合理的,也是必需的。如果这种情况经常发生,则说明需要修改预定义类别,然后

重新进行上述训练与分类过程。在计算 $\text{sim}(d_k, c_i)$ 时,有多种方法可供选择。最简单的方法是仅考虑两个特征矢量中所包含的词条的重叠程度。即

$$\text{sim}(d_k, c_i) = \frac{n_1(d_k, c_i)}{n_y(d_k, c_i)}$$

其中, $n_1(d_k, c_i)$ 是 $V(d_k)$ 和 $V(c_i)$ 具有的共同词条数目, $n_y(d_k, c_i)$ 是 $V(d_k)$ 和 $V(c_i)$ 具有的所有词条数目。

最常用的方法是考虑两个特征矢量之间的夹角余弦,即

$$\text{sim}(d_k, c_i) = \frac{V(d_k) \cdot V(c_i)}{|V(d_k)| \times |V(c_i)|}$$

训练方法和分类算法是分类系统的核心部分,目前存在多种基于向量空间模型的训练算法和分类算法^[12],例如,支持向量机算法、神经网络方法、最大平均熵方法、最近 K -邻居方法和贝叶斯方法等等。

2.4 文本聚类

文本聚类^[13]是从给定的文档本身出发,根据文档特征词矢量,将相关者聚成一类。根据文本聚类的结果不同,可以将聚类方法分为层次聚类法和平面聚类法两种类型。

●对于给定的文档集合 $D = \{d_1, \dots, d_i, \dots, d_n\}$, 层次聚类的过程如下:

(1) 将 D 中的每一个文档 d_i 作为一个聚类中心 $c_i = \{d_i\}$, 形成 D 的一个聚类集合 $C = \{c_1, \dots, c_i, \dots, c_n\}$;

(2) 计算 C 中每个聚类对 (c_i, c_j) 之间的相似度 $\text{sim}(c_i, c_j)$;

(3) 选取具有最大相似度的两个聚类 (c_i, c_j) —— $\max \text{sim}(c_i, c_j)$, 将合并成一个新的聚类 $c_k = c_i \cup c_j$, 同时合并 c_i 和 c_j 的特征矢量,从而构成了 D 的一个新聚类集合 $C = \{c_1, \dots, c_k, \dots, c_{n-1}\}$;

(4) 重复上述步骤,根据所要产生聚类的数目和相似度阈值限制,得到最终聚类结果。

平面划分法与层次凝聚法的区别在于,它将文档集合水平地分割为若干个聚类,而不是生成层次化的嵌套聚类。

●对于给定的文档集合 $D = \{d_1, \dots, d_i, \dots, d_n\}$, 平面划分法的具体过程如下:

(1) 确定要生成的聚类的数目 k ;

(2) 抽取 D 中每个文档的特征矢量 $V(d_i)$;

(3) 从 D 中抽取 k 个文档形成聚类的中心 $S = \{s_1, \dots, s_j, \dots, s_k\}$ 。为了提高聚类的准确度,在确定聚类中心时应该依据一定的原则。常用的确定聚类中心的方法有逆中心距法和密度测试法等。

(4) 对 D 中剩下的文档,依次计算它们与各个聚类中心的相似度 $\text{sim}(d_i, s_j)$ 。根据预定的相似度阈值,将文档聚集在聚类中心的周围,形成稳定的聚类结果。

从上面的聚类过程可以看出,层次聚类对文档集合 D 中的每一个文档进行了多次遍历,其结果实质上构造出了一个生成树,其中包含了聚类的动态过程和层次信息。层次聚类方法是最为常用的聚类方法,因为它能够产生层次化的嵌套聚类,所以有很高的准确度。另外,在层次聚类过程中,最大相似度呈递减趋势,因此必须确定适当的相似度阈值,保证同一个聚类中文档的紧密相关。而平面划分法的运行速度较快,但是必须事先确定 k 的取值,且种子选取的好坏对聚类结果有较大影响。

3 基于 Web 的文本挖掘模型

基于 Web 的文本挖掘系统^[14]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。鉴于该目的考虑,文中在设计信息挖掘过程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、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^[15],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。如图 1 所示。

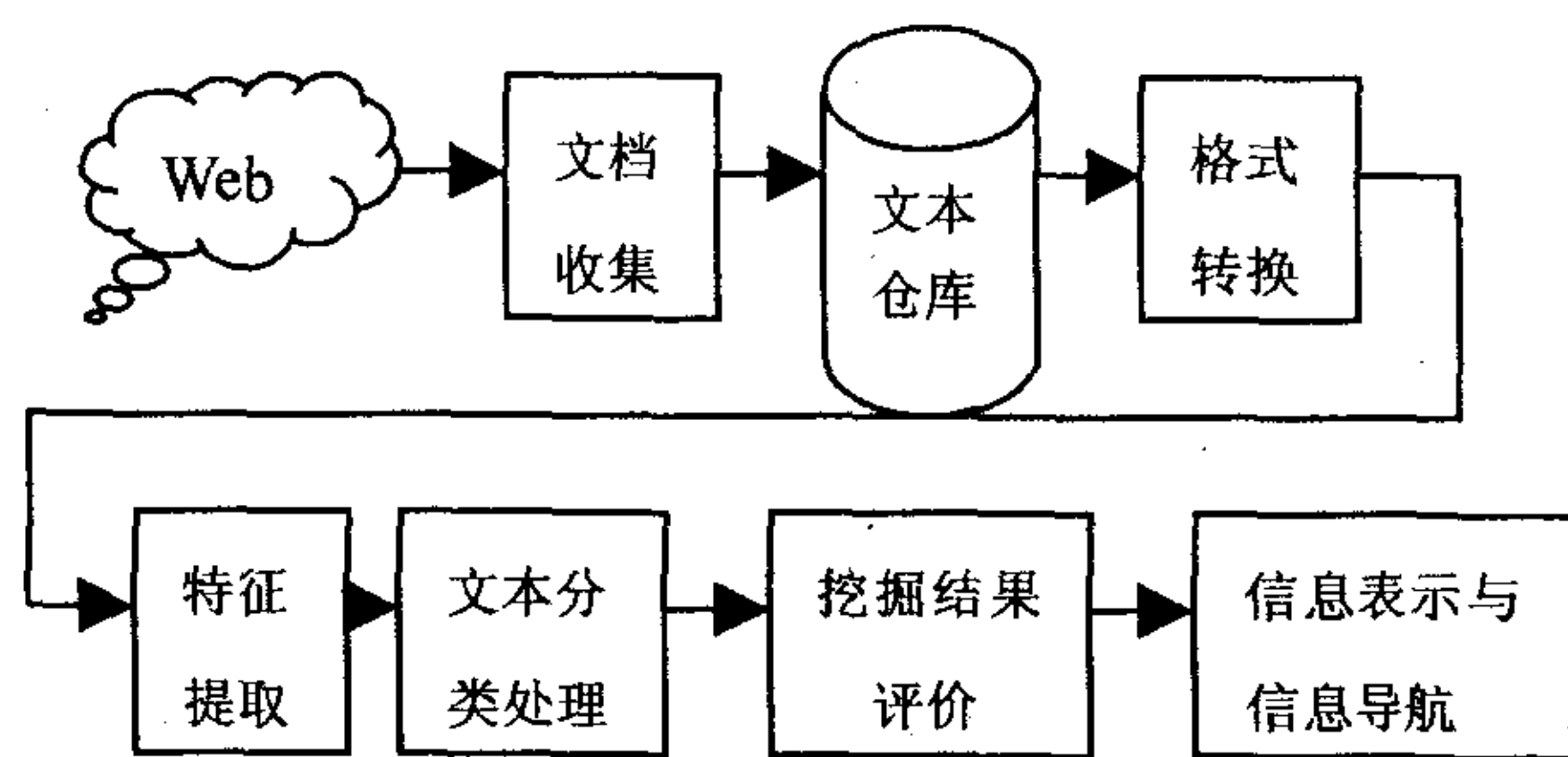


图 1 基于 Web 的文本挖掘模型的结构图

该文本挖掘结构模型的工作流程安排如下:

(1) 特征提取:对 Web 上收集到的挖掘目标样本进行特征提取,生成挖掘目标的特征矢量;特征项集选取应该根据两个基本原则即完全性和区分性原则来进行,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:将数据挖掘中的若干算法进行适当改进后,对于 Web 文本的中间表示形式进行挖掘处理,得到潜在的知识或模式。

(3) 挖掘结果评价:将挖掘得到的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。

(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。

4 Web 挖掘的应用前景

随着 Internet 技术的迅速发展和不断的普及应用,网络信息资源越来越丰富,如何分析和利用这些海量的数据,是当前比较突出的一个问题。网络信息挖掘在实际工作中具有重要的实践意义和广阔的应用前景。

在电子商务领域^[16],网络信息挖掘可以提供不同用户的特定信息,有的放矢地传播网络广告,可以建立客户关系管理系统,极大地提升企业的竞争优势;在电子政务领域,通过对政务数据进行定性和定量分析,可为高层管理者提供决策参考;可以提高搜索引擎获取信息的准确性,并可以对用户搜索结果进行相关处理,可以提高查准率和查全率。

目前,各种应用服务越来越多,电子邮件、BBS 等成为人们普遍采用的信息传播手段,网络信息的管理工作成为大家越来越关注的问题。

5 结束语

Web 挖掘是 Web 技术中一个重要的研究领域,Web 文本挖掘是将数据挖掘技术应用于互联网的知识发现过程,它同时具有自身的特点。Web 文本挖掘是 Web 挖掘的重要代表,可以使用户比较准确找到需要的资料,可以帮助用户节约检索时间,它使充分利用 Web 大量的真正有价值的信息成为可能,为智能化 Web 奠定了基础。Web 文本挖掘技术也将随着人工智能等学科的发展而发挥更大的作用,是人类在信息

社会中应用互联网面临的一个新的挑战。

参考文献:

- [1] 韩家炜,孟小峰,王 静. Web 挖掘研究[J]. 计算机研究与发展,2001(4):405-414.
- [2] 薛为民,陆玉昌. 文本挖掘技术研究[J]. 北京联合大学学报,2005(4):59-63.
- [3] 王 实,高 文,段立鹏. Internet 上的文本挖掘[J]. 计算机科学,2000,27(4):32-36.
- [4] 郭庚麒. Web 文本挖掘技术[J]. 计算机与网络,2004(1-2):114-116.
- [5] 王继成,潘金贵,张福炎. Web 文本挖掘技术研究[J]. 计算机研究与发展,2000,37(5):513-520.
- [6] Han Jiawei, Kamber M. Data Mining: Concept and Techniques [M]. [s. l.]:Morgan Kaufmans Publishers, Inc,2001:274-277.
- [7] 邹 涛,王继成,朱华宇. WWW 上的信息挖掘技术及实现[J]. 计算机研究与发展,1999,36(8):1019-1024.
- [8] 张卫丰,徐宝文,周晓宇. Web 搜索引擎综述[J]. 计算机科学,2001,28(9):24-28.
- [9] 杨炳儒. 知识工程与知识发现[M]. 北京:冶金工业出版社,2000:5-20.
- [10] 唐 菁,张 前,陈泓婕,等. 基于 Web 的文本挖掘[J]. 计算机工程与应用,2002(21):198-201.
- [11] 王连军. Web 文本挖掘浅析[J]. 现代图书情报技术,2002(6):38-40.
- [12] 林士敏,田风占. 用于数据采掘的贝叶斯分类器研究[J]. 计算机科学,2000,27(10):73-76.
- [13] 易高翔,程耕国. Web 文本挖掘研究[J]. 武汉科技大学学报:自然科学版,2005(1):72-74.
- [14] 汪启军,申瑞民. 基于 Web Mining 的智能化、个性化的远程教育模型研究[J]. 计算机工程与应用,2000,36(12):157-159.
- [15] 林鸿飞,姚天顺. 基于潜在语义索引的文本浏览机制[J]. 中文信息学报,2000,14(5):49-56.
- [16] 黄晓斌. 网络信息挖掘[M]. 北京:电子工业出版社,2005:160-167.

(上接第 186 页)

SP Proxy 在客户端与服务器端的 RTSP 协商过程中对媒体传输信息进行修改,客户端可以直接在 RTSP 的交互中获取媒体传输信息。

参考文献:

- [1] Gilligan R, Thomson S, Bound J, et al. Basic Socket Interface Extensions for IPv6[S]. RFC2133. 1997.
- [2] 张 雪,董永强,黄一鸣. 支持 IPv4/IPv6 的 RTSP 流媒体应用代理的设计与实现[J]. 计算机科学,2006,33(3):140-144.
- [3] Schulzrinne H, Rao A, Lanphier R. Real Time Streaming Pro-

ocol (RTSP)[S]. RFC 2326. 1998.

- [4] Schulzrinne H, Frederick R, Jacobson V. RTP: A Transport Protocol for Real-Time Applications[S]. RFC1889. 1996.
- [5] Handley M, Jacobson V. SDP: Session description protocol [S]. RFC 2327. 1998.
- [6] Olson S, Camarillo G, Roach A B. Support for IPv6 in Session Description Protocol (SDP)[S]. RFC3266. 2002.
- [7] Handley M. Session Announcement Protocol[S]. RFC2974. 2000.
- [8] Deutsch P, Gailly J L. Zlib compressed data format specification version 3.3[S]. RFC1950. 1996.