

# Web 文本挖掘技术研究\*

## Research on Web Text Mining Technology

邹腊梅 肖基毅 龚向坚

(南华大学计算机学院 衡阳 421001)

**摘要** WWW 上的信息量迅猛增长, Internet 成为一个拥有大量 Web 文本资源的巨型数据库。大量异构、非结构化的 Web 文本对数据挖掘技术提出新的挑战。分析了 Web 文本的特点、Web 文本挖掘的一般流程以及 Web 文本挖掘中的关键技术。

**关键词** 数据挖掘 Web 文本 文本特征 分类 聚类

### 1 Web 文本的特点

在浩如烟海的网络信息中, 80% 的信息是以文本的形式存放的, 包括 Web 页面、小说、电子邮件、新闻和广告等, Web 上的文本具有以下一些特点:

**1.1 非结构、半结构性** 与传统数据库、数据仓库中结构化的数据不同, Web 文本大部分是非结构、半结构性的。非结构化文本主要指 Web 上的自由文本, 包括小说、新闻等。此外, Web 上大部分的文本既包含了标题、作者、出版日期、出版刊物名等结构化信息, 又包含了摘要、内容、参考文献等非结构化的信息, 特别在很多文本中又加入了超链接的特殊的成分。这些文本被称为半结构化的文本, 它们构成了 Web 信息源的主体。

非结构、半结构的文本没有特定的模型描述, 没有固定的数据结构。要对这类的信息进行挖掘, 必须进行数据的数据化转换, 并进行结构化数据的存储。

**1.2 自述性、动态可变性** Web 文本具有自我描述的特性。Web 上的数据通常具有一定的结构性, 但因自述层次的存在, 使得 Web 数据成为一种非完全结构化的数据。此外, 各个网站网页的内容是不断更新变化的, 特别是新闻、广告的更新频率更高, 更新周期很短, 具有明显的动态可变性。

**1.3 异构数据库环境** Web 上的数据丰富而复杂, 每一站点的数据库都各自独立设计, 如果将一个站点看作数据源, 那么每个数据源都是异构的。从数据库研究的角度出发, Web 网站上的信息就可以看作一个巨大、复杂的异构数据库环境。因此, 要想对 Web 上的文本信息进行挖掘, 必须解决各个站点之间异构数据的集成问题。

### 2 Web 文本挖掘的定义

Web 文本挖掘是一门交叉性学科, 它涉及到计算机语言学、自然语言处理、数据抽取、信息检索、人工智能、神经网络、统计学、机器学习、数据挖掘等多个领域。由于 Web 文本挖掘涉及的领域广泛, 挖掘的内容丰富、复杂, 目前对 Web 文本挖掘国内外还

没有统一、准确的定义。根据研究的侧重点不同, 研究人员提出了一些定义:

Oren Etzioni 定义为: 使用数据挖掘技术自动地从 Web 文档和服务中发现和提取信息和知识的技术<sup>[1]</sup>。该定义偏重于挖掘技术和挖掘的目的研究。王继成、潘金贵等定义为: 从大量 Web 文档的集合  $C$  中发现隐含的模式  $P$ 。如果将  $C$  看作输入, 将  $P$  看作输出, 那么挖掘的过程就是从输入到输出的一个映射  $\zeta: C \rightarrow P$ <sup>[5]</sup>。该定义偏重于 Web 文本挖掘过程。

### 3 Web 文本挖掘的一般流程

Web 文本挖掘可以分为这样几个步骤: Web 文本的收集和预处理、特征的表示和特征的提取、数据挖掘、挖掘结果评价、信息表示和信息导航。挖掘流程图如图 1 所示:

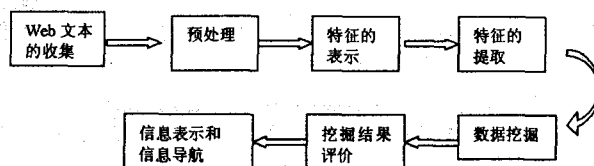


图 1 Web 文本挖掘流程图

Web 文本的收集和预处理: 程序 (Robot) 能自动利用网页中超链接来收集相关主题的网页。为了提高数据的质量, 可以对文本作一些预处理, 如清除图像文件、脚本程序等。

特征的表示和提取: 从 Web 文本中抽取代表其特征的元数据, 这些特征可以用结构化的形式保存, 作为文档的中间表示形式。特征的提取是为了减少特征向量的维度。

Web 文本挖掘: Web 文本挖掘是对大量 Web 文档进行分类、聚类、关联分析以及对 Web 文档进行自动文摘的过程。

挖掘结果评价: 对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。

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浏览和获取信息。

基金项目: 湖南省自然科学基金“网络环境下信息资源共建共享、机、环境研究”(编号: 04JJ40051) 研究成果之一。

作者简介: 邹腊梅, 女, 1977 年生, 硕士研究生, 讲师, 研究方向为数据挖掘; 肖基毅, 男, 1962 年生, 副教授, 硕士生导师, 研究方向为数据挖掘; 龚向坚, 男, 1977 年, 硕士研究生, 助教, 研究方向为网络搜索引擎。

## 4 Web文本挖掘的关键技术

4.1 文本特征表示 文本特征指的是代表 Web 文本的元数据,分为描述性和语义性特征。描述性特征是指 Web 文本的名称、日期、大小、类型;语义性特征主要指文本的作者、机构、标题、内容等;在这两种特征中,对于名字、日期、大小、类型、作者、机构等具有明显标志的专有特征项,我们可以利用其特点将其提取出来,比如名字有名字识别算法,日期有日期识别算法等。对 Web 文本的内容特征的提取主要有布尔模型、概率模型以及近年来常用的向量空间模型(VSM)。在 VSM 模型中,文本被看作由一组正交词条所生成的向量空间,如果将  $t_i$  看作词条项,  $w_i(d)$  看作  $t$  在文本  $d$  中的权值,每个文档  $d$  可以看成是一个规范化向量  $V(d) = (t_1, w_1(d), \dots, t_i, w_i(d), \dots, t_n, w_n(d))$ 。通常,将  $d$  中出现的所有单词作为  $t_i$ ,为了提高内容特征表示的准确性,也可以将词  $t_i$  作为  $d$  中出现的所有短语。 $w_i(d)$  一般定义为  $t_i$  在  $d$  中出现频率  $tf_i(d)$  的函数,即  $w_i(d) = \theta(tf_i(d))$ 。常用的函数有布尔函数、平方根函数、对数函数以及常用的 TFIDF 函数。TFIDF 函数如下所示:

$$W_{ik} = \frac{tf_{ik} \log(\frac{N}{n_k} + 0.01)}{\sqrt{\sum_{i=1}^n (tf_{ik})^2 \times \log^2(\frac{N}{n_k} + 0.01)}}$$

其中,  $tf_{ik}$  表示词条  $t_k$  在文档  $d_i$  中的出现频数,  $N$  表示全部样本文档总数,  $n_k$  表示词条  $t_k$  的文档频数。

4.2 文本特征提取 Web 文本的数据量非常大,用来表示文本的特征向量的维数很大,可能会达到几万维,因此我们需要从中提取权值较高的词条作为文档的特征项,来达到对特征向量降维的目的。特征提取的方式主要有:a.从原始特征中挑选出一些最具代表性的特征;b.根据专家的知识挑选最有影响的特征;c.用映射或变换的方法把原始特征变换为较少的新特征;d.评估函数法,对特征集中的每个特征进行独立的评估并给定一个评估分值,选取预定数目的最佳特征作为特征子集。本文将介绍矩阵奇异值分解(SVD)和评估函数法方式。

4.2.1 矩阵奇异值分解(SVD)。矩阵的奇异值分解基本原理:任意矩阵都可以进行奇异值分解,设  $A_{m \times n}$  是任意一个  $m \times n$  实矩阵,  $A^T$  表示  $A$  的转置矩阵,  $r(A)$  表示  $A$  的秩,则存在一个  $m$  阶正交阵  $U$ ,  $m \times n$  广义对角阵  $D$ ,  $n$  阶正交阵  $V$ ,使  $A = UDV^T$ 。

利用训练文本集建立  $m \times n$  的词-文本矩阵  $A_{m \times n} = [b_{ij}]$ ,其中  $b_{ij} = L(i, j) \times G(i)$ ,  $L(i, j)$  是词  $i$  在文本  $j$  中局部权重,  $G(i)$  是单词  $i$  在文本集中的全局权重,  $m$  为提取单词数,  $n$  为文本数。对  $A$  进行 SVD 分解(设  $m > n$ ,  $\text{rank}(A) = r$ , 存在  $k, k < r, k < \min(m, n)$ ),则在 2-范数意义下,  $A$  的秩- $k$  近似矩阵  $A_k$  为:  $A \approx A_k = U_k \sum_{k=1}^r k V_k^T$ 。其中,  $U_k$  和  $V_k$  的列分别被称为矩阵  $A_k$  的左右奇异向量,  $\sum_{k=1}^r k$  是对角矩阵,对角元素被称为矩阵  $A_k$  的奇异值。

通过对文本集的词-文本矩阵的奇异值分解,提取  $k$  个最大的奇异值及其对应的奇异向量构成新矩阵来近似表示原文本集的词条-文本矩阵,达到减少矩阵维度的目的。

## 4.2.2 评估函数法。

a. 信息增益<sup>[3,6]</sup>(Information Gain)IG

信息增益通过文本特征项在文本中出现与不出现的情况来推算该特征项的信息量。定义如下:

$$IG(t) =$$

$$P(t) \sum_i P(c_i/t) \log \frac{P(c_i/t)}{P(c_i)} + P(\bar{t}) \sum_i P(c_i/\bar{t}) \log \frac{P(c_i/\bar{t})}{P(c_i)}$$

$c_i$  表示目标文本的类集,  $\bar{t}$  表示特征项在文本中不出现,  $P(c_i)$  表示第  $i$  类出现的概率。

## b. 期望交叉信息熵(Expected Cross Entropy)

$$ECN(t) = p(t) \sum_i P(c_i/t) \log \frac{P(c_i/t)}{P(c_i)}$$

$P(c_i/t)$  表示文本中出现词条  $t$  时,文本属于  $c_i$  的概率,  $P(c_i)$  是类别出现的概率。此外还有互信息(Mutual Information)MI,  $X^2$  统计法(Chi)等。

4.3 Web文本挖掘 Web文本挖掘是根据用户的需要,在大量的 Web 文档中将有价值的、用户以前未曾注意的有用信息挖掘出来。近年来研究和应用最多的 Web 文本挖掘技术有关联规则、文档分类、文档聚类、自动文摘,本文将介绍文档分类、文档聚类和自动文摘技术。

4.3.1 分类(Classification)。分类是典型的有教师的机器学习方法,也是数据挖掘中一个重要的研究课题。分类的目的是通过对训练数据的不断学习得到分类函数或分类器,该函数或分类器可以将测试数据映射到给定类别中的某一个。数据分类过程分为两步:第一步,通过分析属性描述的数据库元组(训练数据集)来建立模型,描述预定的数据类集或概念集。第二步,评估模型的预测准确率(正确被模型分类的测试样本的百分比),评估预测准确率的方法与策略主要有  $k$ -折交叉确认( $k$ -fold cross-validation)、装袋(bagging)、推进(boosting)。如果认为预测准确率可以接受,就可以使用模型对未知标号的文本进行分类。常用的分类方法有决策树分类、贝叶斯分类法、神经网络等。

a. 决策树分类算法。决策树分类起源于概念学习系统,为了判断决策树中属性的相对重要性,Quinlan 提出了 ID3 算法;为了处理连续属性的分类,Quinlan 又提出了 C4.5 算法;因为决策树算法要求数据驻留在内存中,因此该算法能处理的数据有限,其后提出了可伸缩的算法,如 SLIQ、SPRINT 和“雨林”算法。

b. 贝叶斯分类法。贝叶斯分类是统计学分类方法,它是一类利用概率统计知识进行分类的算法。其中,朴素贝叶斯分类算法得到广泛的应用,该算法能运用到大型数据库中,且方法简单、分类准确率高、速度快。由于贝叶斯定理假设一个属性值对给定类的影响独立于其它属性的值,而此假设在实际情况中经常是不成立的,因此其分类准确率可能会下降。为此,就出现了许多降低独立性假设的贝叶斯分类算法,如 TAN(Tree Augmented Bayes Network)。

c. 神经网络分类法。神经网络概念是由心理学家和神经生物学家提出的,主要是模拟人脑的结构,将文本看成一组相连的输入/输出单元,其中每个单元连接于一个相连,通过调整单元的权值,来进行训练样本的学习。常用的神经网络算法有:后向传播算法、前向传播算法、自组织网络(SOM)。神经网络的优点是对噪声数据的高承受能力,对未经训练的数据分类能力。缺点是需要的训练时间,可解释性差。

4.3.2 聚类(Clustering)。聚类是典型的无教师的机器学习方法,也是数据挖掘中一个重要手段。聚类的目的是将物理或抽象对象的集合分组成为由类似的对象组成的多个类。与分类不同,聚类要划分的类是未知的。常用的聚类方法有:划分方法,层次的方法,基于密度的方法,基于网格的方法。

a. 划分方法。给定一个  $n$  个对象或元组的数据库, 划分方法将构建  $k$  个划分,  $k \leq n$ , 并且每个组至少包含一个对象; 每个对象必须属于且只属于一个组。常用的算法有  $K$ -平均算法、 $K$ -中心点算法。

b. 层次的方法。层次的方法是创建给定数据对象集合的一个层次的分解, 该方法分为凝聚方法(自底向上)和分裂的方法(自顶向下)。常用算法有 AGNE(Agglomerative NESting)、DIANA(Divisive ANALysis)、URE(Clustering Using Representatives)。

c. 基于密度的方法。为了发现任意形状的聚类结果, 提出基于密度的聚类方法。该方法将簇看作是数据空间中低密度域分割开的高密度对象区域。其主要思想是: 只要临近区域的密度超过某个阈值, 就继续聚类。常用的算法有 DBSCAN 和 OPTICS。

d. 基于网格的方法。基于网格的聚类方法采用一个多分辨率的网格数据结构。把对象空间量化为有限数目的单元, 形成了一个网格结构。所有的聚类操作都在这个网格结构上进行。这种方法的主要优点是它的处理速度很快, 其处理时间独立于数据对象的数目, 只与量化空间中每一维的单元数目有关。常用的算法有 STING 和 WaveCluster。

4.3.3 自动文摘(Automatic Summarization)。自动文摘技术的作用是生成给定原文的中心内容, 或把所需要的内容从文章中自动抽取出来, 并用同于或不同于原文的句子表示出来<sup>[11]</sup>。自动文摘主要分为文本分析、文本转换、文摘生成三个步骤: 文本分析是寻找最能代表原文内容的成分; 文本转换过程是通过摘录或概括的方法压缩文本; 文摘生成是重组原文内容, 生成文摘。

目前研制开发的自动文摘系统主要采用自动摘录(机械摘录)、理解文摘技术。自动摘录是根据外在的特征抽取原文中的部分句子作为摘要。主要原理是分析原文内容、找出反映文章主题的词(关键词)、将关键词出现频率较高的句子作为关键句, 构成文章的摘要。在自动文摘处理过程中主要采用统计的方法来计算词权、句权。向量空间模型(VSM)是自动文摘中的基本方法。利用自动摘要技术开发的系统有 James A. Rush 开发的 ADAM 系统, 复旦大学研制的复旦中文自动摘要系统等。

理解文摘是利用自然语言理解技术来获取语言结构、更重要的是利用领域知识进行判断、推理, 得到文摘的意义表示, 最后从意义表示中生成摘要。理解文摘主要采用的方法有脚本、概念从属结构、框架、一阶谓词、关联网络、修辞结构以及语用功能等。利用理解文摘技术开发的系统有美国耶鲁大学的 Schank 研制开

发的 SAM 系统, 德国康斯坦茨大学的 Kuhlen 等人研制开发的 TOPIC 系统, 哈尔滨工业大学研制的基于理解的军事领域自动文摘使用系统等。

## 5 结束语

由于 Web 信息量的快速增长, 人们急需新的技术来处理大量的、异构的、半结构化的数据, Web 文本挖掘就是将 WWW 和数据挖掘结合的新技术, 也是数据挖掘研究领域中的一个重要的课题。近年来研究人员提出许多的理论和具体的挖掘算法, 但还没有形成统一的理论体系, 大部分的算法还不成熟, 存在一定的缺陷, 需要进一步的研究和完善, 这也成为促使文本挖掘技术发展的动力。随着文本挖掘技术的不断完善, 其应用领域也会不断增长, 应用前景会越来越好。

## 参考文献

- Oren Etzioni. The World Wide Web: Quagmire or Gold Mine Communication of the ACM, 1996; 39(11)
- Freitag D, McCallum A. Information Extraction with HMMs and Shrinkage. Inc: Proc. Workshop on ML and IE, AAAI-99, 1999
- Jiawei Han, Data Mining. Concepts and Techniques. 北京: 机械工业出版社, 2005
- D Smith, M Lopez. Information Extraction for Semi-structured Documents in: Proc. of 1st Workshop on Management of Semi-structured Data. Arizona, 1997
- 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; (5)
- 李 琦, 杨 峰. 基于增益的隐马尔可夫模型的文本组块分析. 计算机科学, 2004; (2)
- Kin Keuny. Multi-agent Web Text Mining on the Grid for Enterprise Decision Support 2006/Volume 2437
- Ronen Feldman 等. Text Mining via Information Extraction 2004/Volume
- Laurence A. F. Park A Novel Web Text Mining Method Using the Discrete Cosine Transform 2002/Volume 2437
- Nahm U Y Mooney R J. Text Mining with Information Extraction To Appear in the AAAI 2002 Spring Symposium on Mining Answers from Texts and Knowledge Bases 2002
- 杨建林. 自动文摘的逻辑解释. 情报理论与实践, 2002; (25)
- 许建潮, 胡 明. 中文 Web 文本的特征获取与分类. 计算机工程, 2005; (31)
- 王光宏, 蒋 平. 数据挖掘综述. 同济大学学报 2004; (2)
- 金 博, 史彦军, 滕弘飞等. 自动文摘技术及应用. 计算机应用研究, 2004; (12)

(责编: 阳王京)

(上接第 52 页)后接收消费者的订单请求。从该网上商店的整个业务处理流程看, 该网上商店实现了自身系统与商品供应商等业务伙伴之间的自主式的极少人工干预的自动化系统集成, 而无需考虑商品供应商等业务伙伴的内部具体实现细节, 能松散地耦合商品供应商的订单服务, 而将自身的主要精力放在面向消费者的用户界面上。

## 4 小 结

随着企业竞争全球化的发展, 企业在提高其信息化水平的前提下, 整合和优化企业运行的业务流程显得尤为重要。基于松散耦合的 Web 服务技术为整合和优化业务流程提供了有效的解决方案, 企业依靠 Web 服务技术在原有业务流程应用最

小改动和尽可能通用的前提下, 可以有效地解决企业内部或企业间异构平台上独立业务流程应用的有效整合和优化, 从而完成构建复杂的业务流程, 扩大业务流程的应用范围。

## 参考文献

- 蔡 斌, 赵明剑, 黄丽华. 业务流程管理(BPM)技术演进及新动态. 科技导报, 2004; (11)
- 柴晓路, 梁宇奇. Web Services 技术、架构和应用. 北京: 电子工业出版社, 2003
- 彭敦陆, 杜雪峰. 基于 Web 服务的组件集成技术在客户关系管理中的应用. 上海理工大学学报, 2004; (1)
- 李秀军, 林丁禹, 房丽娜. 自适应 Web 服务工作流体系结构的研究. 计算机工程与应用, 2005; (21)
- 刘绍华, 魏 峻, 黄 涛. 基于服务协作中间件的动态流程模型. 软件学报, 2004; (10)

(责编: 阳梅)