面向 Web 的文本信息挖掘研究

Study of Text Mining Technology Oriented Web

张宏松 刘建辉 (辽宁工程技术大学研究生学院 阜新 123000)

摘要:万维网是一个巨大的、分布广泛的、全球性的信息服务中心,它包含了丰富的信息资源。Web 挖掘可以快速有效地获取所需要的信息。基于 Web 的文本挖掘是数据挖掘的重要组成部分,探讨了 Web 文本特征提取、文本分类、文本聚类等 Web 文本挖掘关键实现技术,最后讨论了 Web 文本挖掘的价值及其对 Web 发展的重要性。

关键词: Web 挖掘 文本挖掘 文本分类 文本聚类

1 Web 文本挖掘技术

Web 挖掘一门交叉性学科,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 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、模式、规则、规律、约束及可视化等形式)的过程[1]。

当前研究的 Web 挖掘一般可分为三类:

(1) Web 内容挖掘。它是从 Web 文档内容或其描述的挖掘获取知识的过程。

(2) Web 结构挖掘。它是从 WWW 的组织结构和链接关系的挖掘获取知识。

(3) Web 访问信息挖掘。它是通过从 Web 的访问信息的挖掘获取知识。Web 挖掘分类结构图如图 1 所示。

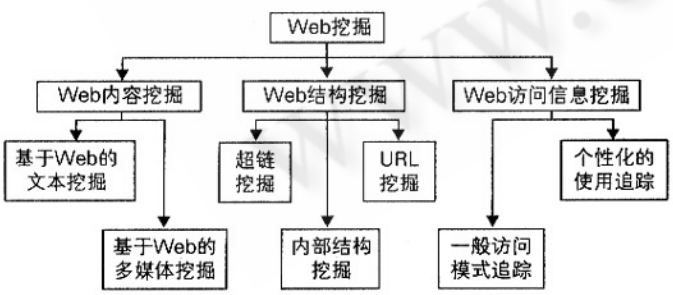

图 1 Web 挖掘分类

Web 挖掘对象分为资源发现和信息获取。资源发现就是定位文本的位置,并自动生成文档的索引。Web 上的资源一般分为两类:文档和服务。目前 Web 上的资源发现主要集中于 Web 内容的挖掘。文本挖掘是指将数据挖掘技术应用在大量的文本集合上,发现其中隐含的知识的过程。大多数基于数据库的数据挖掘方法均可作用于文本挖掘,如数据归纳、分类、聚类、关联规则挖掘等。文本挖掘的结果既可以是对某个文本内容的概括,也可以是对整个文本集合的分类结果或聚类结果等[2]。

Web 上的数据与传统的数据库中的数据不同,传统的数据库都有一定的数据模型,可以根据此模型来具体描述特定的数据。而 Web 上的数据非常复杂,没有特定的模型描述,每一站点的数据都各自独立设计,并且数据本身具有自述性和动态可变性。因而 Web 上的数据具有一定的结构性。但因自述层次的存在,从而是一种非完全结构化的数据,半结构化是形成了 Web 文本挖掘的特色。

2 Web 文本挖掘过程

Web 挖掘过程一般包括相关网页采集、文本的预处理、文本模型表示、信息或文本特征性抽取、文本分类(聚类)或结果集的数据挖掘等步骤,以得到结果,从而极大程度的方便用户有效地浏览和获取信息[3]。Web 挖掘过程如图 2 所示。

2.1 Web 文本抽取及预处理

Web 页面是通过 HTML 语言来定义的,Web 页面通过多用途 Internet 邮件 MIME 来标识不同类型的内容。该系统直接挂接在 Internet 上,数据来源和

用户界面主要都通过 Web 实现。由一个 Robot 程序自动通过 Web 进行用户主题信息的文本的自动搜集。为了提高数据挖掘的效率和有效性，将高速缓存中的一些无用数据清除，如清除 GIF 和 JPEG 格式的图像文件、清除 Web 页面上中的脚本程序等。

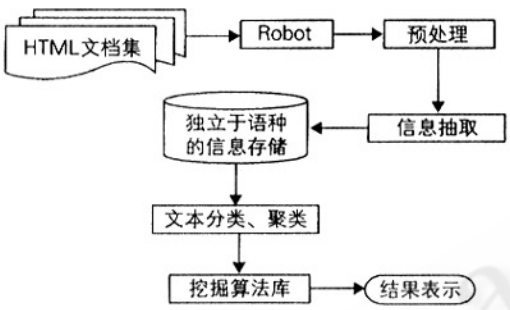

图 2 Web 文本挖掘过程

2.2 特征选取和信息抽取

对经过预处理的样本进行特征提取，利用浅层的自然语言处理技术可以实现高效率的自然语言处理；将非结构化的信息改变为利于计算机存储、处理的结构化形式，采用独立于语种的方式将信息存储于信息库。用户就不必关心原始文本的语种，可以用所熟悉的语种进行抽取请求，并得到希望语种表示的信息抽取结果。同时信息抽取技术能够自动地从庞大的文本库中，动态地根据用户感兴趣的主题内容提取文本蕴含的信息。

将 Web 页面上的数据按不同的数据类型、出现的时间顺序、不同的分割方式和实现方法收集并重组，便产生了用于不同挖掘任务的数据集。将到特征向量 V 中，$V=\{(t1,w1),(t2,w2),(t3,w3),\cdots\}$。利用特征选择方法计算其每一项的权值。在 Web 文档特征所采用的特征选择算法中，一般是构造一个评价函数，对特征集中的每个特征进行独立的评估，这样每个特征都获得一个评估分即权值，然后对所有的特征按照其权值大小排序，选取预定数目的最佳特征作为结果的特征子集[4]。所以，选取多少个最佳特征以及采用什么评价函数都要针对一个具体的问题通过实验来决定。特征选择主要用于排除那些被认为无关或关联性不大的特征即术语，依据文档集统计数据，这些特征处于无信息量的状态；并自动将那些低频的特征用正交方法合并成高频特征。

词、词组和短语组成文档的基本元素，并且在不同内容的文档中，各词条出现频率有一定的规律性，不同的特征词条就可以区分不同内容的文本。因此可以抽取一些特征词条构成特征矢量，用这个特征矢量来表示 Web 文本，一个有效的特征词条集，必须具备以下三个特征：完全性，特征词条能够确实表示目标内容；区分性，根据特征矢量，能将目标同其它文档相区分；精练性，特征矢量的维数应该尽可能的小。

目前文本特征矢量的获取一般是通过一些分词算法和词频统计方法，从文档中选出尽可能多的词、词组和短语，由它们来构成文档矢量。但是由这种方法来表示文档，矢量的维数非常巨大。这种未经处理的文档矢量给后续的处理工作带来巨大的计算开销，使整个处理过程的效率非常低下。因此必须对文档矢量作进一步精化处理，在保证原文含义的基础上，找出最能反映文本内容，又比较简洁的特征矢量。Web 文本实际上可以看作是由众多的特征词条构成的多维信息空间，特征矢量的选择实际上就是多维信息空间中的寻优过程。因此处理特征矢量的选择问题很自然地就想到使用高效的寻优算法。

3 Web 文本挖掘

Web 文本挖掘的处理结果是结构化的数据。存放于后台数据仓库中，再根据不同的文本挖掘目标如关联知识发现、趋势预测、序列知识发现等到采用不同的数据挖掘算法。

3.1 文本关联分析

文本关联分析主要是实现 Web 页面信息的概念提升及多层关联规则的挖掘功能。在 Web 文本内容挖掘的过程中，主要是利用向量空间模型法（VSM）。它的主要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。但是，向量空间模型关于词间关系相互独立的基本假设（正交假设）在实际环境中很难满足，文本中出现的词往往存在一定的相关性，即出现斜交情况，在某种程度上会影响计算的结果。同时词汇具有的同义或者多义现象也会影响文本信息的分析和处理。在这种情况下，为了更加有效地解决文本信息处理时存在的各种歧义问题，可以结合潜在语义索引的概念，对于词进行概念上的提升，使其包含更加丰富的语义信息，并根据需要形成多个抽

象级。在不同的抽象级上,分别对应相应的具体含义。因此在进行文本信息处理时,需要构建一个概念词典。概念词典主要反应了层次结构的语义组织,不同的层次表明了其抽象的程度不同。层次越高,则概括性就越强,包含的下位概念可能就越多。在不同概念级别上将进行多层关联规则的挖掘;在页面集合中寻找不同词语之间的关系[5]。可以结合 Apriori 算法并经过适当地改进处理后,应用到文本数据的挖掘上来。最后,以可视化的信息导航结构图的形式表示发现的知识(模式)及它们之间可能存在的关系。

3.2 文本分类

文本分类是一种典型的有教师的机器学习问题,一般可分为训练和分类两个阶段,其中训练阶段过程如下:首先定义类别集合 $C=(c_1,c_2,\cdots,c_l,\cdots,c_m)$,这些类别可以是层次式的也可以是并列式的;然后给出训练文档集合 $S=(s_1,s_2,\cdots,s_l,\cdots,s_n)$,每一个训练文档都被标上所属的类别标识 c_l;最后提取训练文档集合 S 中所有文档的特征矢量 $V(s_l)$,并采用一定某原则来确定代表 C 中每一个类别的特征矢量 $V(c_l)$。分类阶段过程如下:首先对于测试文档集合 $T=(d_1,d_2,\cdots,d_k,\cdots,d_t)$ 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 $V(c_l)$ 之间的相似度 $sim(d_k,c_l)$。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 $sim(d_k,c_l)=\frac{V(d_k)\cdot V(c_l)}{|V(d_k)\times V(c_l)|}$;然后选取相似度最大的一个类别 $\underset{c_l\in C}{\operatorname{argmax}} sim(d_k,c_l)$ 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。除了上述经典方法外,目前使用很广泛的还有朴素贝叶斯分类算法和 K-最近邻参照分类算法。

3.3 文本聚类

文本分类可实现将 Web 文本归类,以便于用户在搜索时可以快速的找到相关的 Web 文档,文本分类是将文档归入到已经存在的类中;文本聚类的目标和文本分类是一样的,只是实现的方法不同,文本聚类是无教师的机器学习,在文档归类之前没有定义好的类可供选择,在文本聚类时,将所有类型接近的文档归为一类,使类型相同的文档尽量归为一类,类型不相同的尽量隔离开来,聚类的标准可以是 Web 文本的属性,也可以是 Web 文本的内容[6]。

常用的文本聚类算法可以归为两类:分割式的聚类和分层式的聚类。分割聚类算法通过优化一个评价函数把数据集分割为 k 个部分。分层聚类是由不同层次的分割聚类组成,层次之间的分割具有嵌套的关系。分层聚类法的构造过程类似于构造哈夫曼树的过程。使用层次凝聚法聚类的过程如下,假设有 n 篇文档 $D=(d_1,d_2,\cdots,d_n)$,首先把每篇文档都看成是单独的一类,有 $(c_1,c_2,\cdots,c_n)$ n 类,每个类之间的相似度构成一个矩阵:$\begin{bmatrix} sim_{11} & sim_{12} & \cdots & sim_{1n} \\ sim_{21} & sim_{22} & \cdots & sim_{2n} \\ \cdots & \cdots & sim_{ll} & \cdots \\ sim_{n1} & sim_{n2} & \cdots & sim_{nn} \end{bmatrix}$。其中 sim_{ll} 是 c_l,c_l 之间的相似度,在此矩阵中选取最大值 sim_{ll},所对应的文档类分别为 c_l,c_l 将相似性最大的两类合并为一个新的类 sim_{uv}。重复以上过程,直到只剩下一个类为止,最后构成一颗二叉树。如果将 Web 文本分为 K 类,根据构造的二叉树,每层都有 $l(1\leqslant l\leqslant n)$ 棵树,采用一定的搜索策略找到具有 K 棵树的层,这 K 棵树的根分别记为 $CT_1,CT_2,\cdots CT_k$,以 $CT_l(1\leqslant l\leqslant K)$ 为根的树上的所有叶子节点归为 CT_l 类,由于聚类的过程是构造一个二叉树,所以效率不是很高。K-means、K-median 算法则是一种平面式的聚类方法,在一定程度上提高了效率,适合于处理 Web 文本这种具有大量数据的对象。

3.4 Web 文本挖掘应用

一个文本检索系统按一定查询格式的输入检索出了一组文档,文本分类系统的评估指标根据文本检索的度量来定义,即查准率(precision)和查全率(recall)。查准率 p 是指分类器判定的属于类别 ci 的所有文档中与实际相符的文档所占的比例(即反映正确性)。查全率 r 是指专家判定的属于类别 ci 的文档中,分类器做出同样判定的文档所占比例。

本系统针对计算机控制类技术文档进行分类,对从 670 篇文档进行训练和测试。所有这些文档分为 4 个类别:组态软件、工控机、现场总线、PLC,取出 170 篇作为测试集,另外 500 篇又分为两部分:360 篇作为训练文档集;余下 140 篇作为 Validation set,用于调整矢量维度。测试结果如表 1 所示。

表 1 测试结果 %

类别	组态软件	工控机	CAN 总线	PLC	平均
p	91.5	100.0	77.0	92.2	90.2
r	82.0	83.5	85.0	76.2	81.7

由测试结果看出，本系统达到了较好的分类效果。另外，从算法的时间复杂度考虑，若训练文档集有 m 篇，矢量维数为 n，类别数为 k，则训练算法复杂度为 O(mn)，分类算法复杂度为 O(kn)。

4 结语

web 上具有丰富的资源，怎样能把这些资源充分挖掘出来为用户所使用，是近几年研究的重点之一。但目前文本挖掘技术还存在着一些问题有待解决，例如多语种的问题，根据信息抽取技术的特征，构建跨语种的信息抽取系统是可能的，可以构建中间语汇，将抽取后的信息以独立于语种的方式表述。各种算法都存在一定的限制，大部分聚类算法由于都是基于欧式距离的、所以只能处理数值属性。因此许多算法还需要不断的改进使之不仅能处理数值属性，还能够处理符号属性。自动分词的完善，可以通过添加一些附加属性来提高分词的质量。自动标注的实现，如实体的识别、指代分析等等。可通过将数据的显示和数据的外观分离，有利于实施精确的数据查询与数据挖掘。由于处理的对象是海量的 web 文本，所以效率并不高，需要进行改进，增强聚类、分类的能力，争取做到在扫描大量数据时，只需扫描一次。由于 web 上的资源大都是以文本形式提供的，所以基于 web 的文本挖掘是 web 知识发现的一个重要组成部分，对于推动基于 Web 的数据挖掘和知识获取以及实现电子数据交换、电子商务等具有重要的意义。

参考文献

1 Chakrabarti S, Dom B E, Kumar S R, et al. Mining the Web's Link Structure. Computer, 1999, 32(8): 60 -67.

2 McHugh J, Abiteboul S, Goldman R, et al. Lore: A Database Management System for Semistructured Data. SIGMOD Record, 1997, 26(3): 54 -66.

3 邹涛、王继成、朱华宇等，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。

4 陈莉、焦李成，Internet/Web 数据挖掘研究现状及最新进展[J]，西安电子科技大学学报(自然科学版)，2001；28(1)：114 -119。

5 陈滢、王能斌，半结构化数据查询的处理和优化[J]，软件学报，1999；10(8)：883 -890。

6 唐青、沈记全、杨炳儒，基于 web 的文本挖掘系统的研究与实现[J]，计算机科学，2003；30(1)：60 -62。