文本挖掘技术研究

薛为民[1,2],陆玉昌[2]
(1. 北京联合大学 自动化学院,北京 100101;2. 清华大学 计算机科学与技术系,北京 100084)

[摘 要] 文本挖掘是数据挖掘的重要内容之一,其应用十分广泛。对文本挖掘技术的基本概念和理论进行系统地归纳总结,首先给出了数据挖掘、文本挖掘和 Web 文本挖掘的基本概念及主要研究方向,然后分析了文本挖掘的过程和关键技术,最后对文本挖掘技术进行总结和展望。

[关键词] 文本挖掘;数据挖掘;Web 文本挖掘;文本挖掘模型
[中图分类号] TP 391 [文献标识码] A [文章编号] 1005-0310(2005)04-0059-05

文本挖掘是近几年来数据挖掘领域的一个新兴分支,在国际上,文本挖掘是一个非常活跃的研究领域。从技术上说,它实际是数据挖掘和信息检索两门学科的交叉。文本挖掘与传统数据挖掘的差别在于文本数据与一般数据的巨大差异。传统数据挖掘所处理的数据是结构化的,如关系的、事务的、数据仓库的数据,其特征数目通常不超过几百个,而文本数据没有结构,转换为特征矢量后特征数将达到几万甚至几十万。所以,文本挖掘既采用了很多传统数据挖掘的技术,又有自己的特性。

近年来随着 Internet 的大规模普及和企业信息化程度的提高,有越来越多的信息积累,Internet 已经发展为当今世界上最大的信息库。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,传统的信息检索技术已不适应日益增长的大量文本数据处理的需要。如何快速、准确地从来自异构数据源的大规模的文本信息资源中提取符合需要的简洁、精炼、可理解的知识,这就涉及到文本知识挖掘。Internet 的发展,极大地促进了文本挖掘的发展。

1 文本挖掘的基本概念

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、模糊的、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,包括分类、聚类、关联规则挖掘、特征与偏差、时序模式发现、趋势分析等。

传统的数据挖掘技术,主要针对的是结构数据,如关系的、事务的、数据仓库的数据。随着数据处理工具、先进数据库技术以及网络技术迅速发展,大量的形式各异的复杂类型的数据(如结构化与半结构化数据、超文本与多媒体数据)不断涌现。因此数据挖掘面临的一个重要课题就是针对复杂数据类型的挖掘,这包括复杂对象、空间数据、多媒体数据、时间序列数据、文本数据和 Web 数据。

1.2 文本挖掘

文本挖掘(TM,Text Mining)是以计算语言学、统计数理分析为理论基础,结合机器学习和信息检索技术,从文本数据中发现和提取独立于用户信息需求的文档集中的隐含知识。它是一个从文本信息描述到选取提取模式,最终形成用户可理解的信息知识的过程。

Web 文本挖掘就是从 Web 文档和 Web 活动中发现、抽取感兴趣的潜在的有用模式和隐藏的信息的过程。Web 文本挖掘可以对 Web 文档集合的内容进行总结、分类、聚类、关联分析以及趋势预测等。Web 文本挖掘和通常的平面文本挖掘有类似之处,但是,Web 文档中的标记给文档提供了额外的信息,可以借此提高 Web 文本挖掘的性能,Web 文本挖掘是文本挖掘的主要研究内容。

1.3 文本挖掘种类

按照文本挖掘的对象可把文本挖掘分类为:基于单文档的数据挖掘和基于文档集的数据挖掘。

[收稿日期] 2005 - 10 - 08
[基金项目] 国家自然科学基金重大项目(79990584);自然科学基金资助项目(60473115)
[作者简介] 薛为民(1968 —),男,河北邯郸人,清华大学计算机系博士后,副教授,研究方向为数据挖掘、智能计算、人机交互;陆玉昌,男(1937 —),清华大学计算机系教授,博士生导师,研究方向为数据挖掘、知识发现和机器学习。

1) 基于单文档的数据挖掘:基于单文档的数据挖掘中对文档的分析并不涉及其它文档。主要挖掘技术有:文本摘要(Text Summarization)、信息提取(Information Extraction),其中信息提取包括:名字提取(Names of people、organizations and places)、短语提取(Multiword terms)、关系提取等。

2) 基于文档集的数据挖掘:基于文档集的数据挖掘对大规模的文档数据进行模式抽取。主要挖掘技术有:文本分类(Text Categorization)、文本聚类(Document Clustering)、个性化文本过滤(Personalized Content Filtering)、文档作者归属(Authorship Attribution)、因素分析(Factor Analysis)等。

2 文本挖掘的主要研究方向

文本挖掘作为数据挖掘中一个日益流行而重要的研究课题有着广泛的应用前景,主要有网络浏览、文本检索、文本分类、文本聚类、文档总结等。

1) 网络浏览领域:文本挖掘技术可以通过分析用户的网络行为等,帮助用户更好地寻找有用信息,一个典型的例子是 CMU 的 WebWatcher。这是一个在线用户向导,可以根据用户的实际点击行为分析用户的兴趣,预测用户将要选择的链接,从而为用户进行导航。

2) 文本检索领域:文本检索主要研究对整个文档文本信息的表示、存诸、组织和访问,即根据用户的检索要求,从数据库中检索出相关的信息资料。这种检索方法有三种:布尔模型是简单常用的严格匹配模型,如清华大学的《中国学术期刊(光盘版)》;概率模型利用词条间和词条与文档间的概率相关性进行信息检索,如美国马萨诸塞大学开发的 INQRERY 文本检索系统;向量空间模型在于将文档信息的匹配问题转化为向量空间中的矢量匹配问题处理,如美国康乃尔大学基于向量空间模型开发了 SMART 文本检索系统。

3) 文本自动分类:文本分类是指按照预先定义的主题类别,为文档集合中的每个文档确定一个类别。这样用户不仅可以方便地阅读文档,而且可以通过限制搜索范围来使文档查找更容易。近年来涌现出了大量的适合于不同应用的分类算法,如:基于归纳学习的决策树(DT, Decision Tree)、基于向量空间模型的 K-最近邻(KNN, K Nearest Neighbor)、基于概率模型的 Bayes 分类器、神经网络(NN, Neural Network)、基于统计学习理论的支持向量机(SVM, Support Vector Machine)方法等。

4) 文本自动聚类:与文本分类相对应的是文本自动聚类。文本聚类是一种典型的无教师机器学习问题,它与文本分类的不同之处在于,聚类没有预先定义好的主题类别,它的目标是将文档集合分成若干个簇,要求同一簇内文档内容的相似度尽可能大,而不同簇间的相似度尽可能小。

5) 文档总结:文档总结也是 Web 文本挖掘的一个重要内容。它是指从文档中抽取关键信息,用简洁的形式,对文档内容进行摘要和解释,这样用户不需阅读全文就可了解文档或文档集合的总体内容。搜索引擎向用户返回查询结果时,通常需要给出文档摘要,这就是文档总结的一个实例。

3 文本挖掘过程

文本挖掘的过程如图 1 所示,开始是文本信息源,最终结果是用户获得的知识模式。

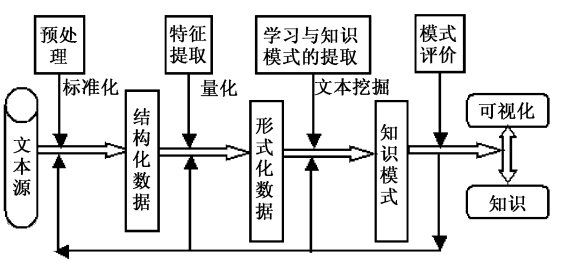

图 1 文本挖掘的过程示意图

文本挖掘一般经过文本预处理、特征提取及约减、学习与知识模式提取、知识模式评价 4 个阶段。

3.1 文本预处理

文本预处理是文本挖掘的第一个步骤,对文本挖掘效果的影响至关重要,文本的预处理过程可能占据整个系统的 80 %的工作量。

与传统的数据库中的结构化数据相比,文档具有有限的结构,或者根本就没有结构,即使具有一些结构,也还是着重于格式,而非文档的内容,且没有统一的结构,因此需要对这些文本数据进行数据挖掘中相应的标准化预处理;此外文档的内容是使用自然语言描述,计算机难以直接处理其语义,所以还需要进行文本数据的信息预处理。信息预处理的主要目的是抽取代表文本特征的元数据(特征项),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。

Internet 上的大部分网页是 HTML 文档或 XML 文档,文本的预处理首先要做的是,利用网页信息抽取模块将网页的内容,去掉跟文本挖掘无关的标

记,转换成统一格式的 TXT 文本存放在文件夹中以备后续处理。中文文本的预处理较英文文本的预处理更为复杂,因为中文的基元是字而不是词,字的信息量比较低,句子中各词语间没有固有的分隔符(如空格),因此对中文文本还需要进行词条切分处理。

3.2 文本的表示

文本的内容是人类所使用的自然语言,表达了丰富的信息,但是要把这些信息编码为一种标准形式是非常困难的。基于自然语言处理和统计数据分析的文本挖掘中的文本特征表示指的是对从文本中抽取出的元数据(特征项)进行量化,以结构化形式描述文档信息。这些特征项作为文档的中间表示形式,在信息挖掘时用以评价未知文档与用户目标的吻合程度,这一步又叫做目标表示。

文本表示的模型常用的有:布尔逻辑模型,向量空间模型(VSM, Vector Space Model),潜在语义索引(LSI, Latent Semantic Indexing)和概率模型(Probablistic Model)。

下面来重点讨论文本挖掘系统中近年来应用较多且效果较好的向量空间模型方法。

向量空间模型的基本思想是使用词袋法(Bag-of-Word)表示文本,这种表示法的一个关键假设,就是文章中词条出现的先后次序是无关紧要的,每个特征词对应特征空间的一维,将文本表示成欧氏空间的一个向量。它的核心概念可以描述如下:

1) 特征项:组成文档的字、词、句子等。$Document = D(t_1, t_2, \cdots t_k, \cdots t_n)$,其中 t_k 表示第 k 个特征项,作为一个维度。

2) 特征项的权重:在一个文本中,每个特征项都被赋予一个权重,以表示特征项在该文本中的重要程度。

3) 向量空间模型(VSM,Vector Space Model):在舍弃了各个特征项之间的顺序信息之后,一个文本就表示成向量,即特征空间的一个点。

如文本 d_i 的表示:

$$V(d_i) = (w_{i1}, w_{i2}, \cdots\cdots w_{ik}, \cdots w_{im})$$

其中,$w_{ik} = f(t_k, c_j)$ 为权值函数,反映特征 t_k 决定文档 d_i 是否属于类 c_j 的重要性。

4) 相似度(similarity):对于所有文档都可映射到此文本向量空间,从而将文档信息的匹配问题转化为向量空间中的矢量匹配问题。n 维空间中点的距离用向量之间的余弦夹角来度量,也即表示了文档间的相似程度。假设目标文档为 U,未知文档为 V_i,夹角越小说明文档的相似度越高。相似度计算公式如下:

$$similarity(V_i, U) = \cos(V_i, U) = \frac{V_i \times U}{V_i \cdot U}$$

$$= \frac{\sum_{k=1}^{m} w_{ik} \cdot w_i}{\sqrt{\sum_{k=1}^{m} w_{ik}^2 \sum_{k=1}^{m} w_k^2}}$$

权重通常是特征项在文档中所出现频率的函数,用 $ft_k(d_i)$ 表示特征 t_k 在文档 d_i 中出现的频率,权重函数有多种:

a. 最简单的布尔型:

$$w_{ik} = \begin{cases} 1, & \text{if } tf_k(d_i) > 0 \\ 0, & otherwise \end{cases}$$

文本向量由 0,1 组成。

b. 词频型:$w_{ik} = tf_k(d_i)$

c. 平方根型:$w_{ik} = tf_k(d_i)^{1/2}$

d. 对数型:$w_{ik} = \lg(tf_k(d_i) + 1)$

e. TF-IDF 公式:

$$w_{ik} = tf_k(d_i) \cdot \lg\left(\frac{N}{N_k} + 0.5\right)$$

比较著名的权值函数是由 Salton 在 1988 年提出的 TF-IDF 公式,N 为训练文本总数,N_k 为训练文本集中出现词条 t_k 的文本数。

归一化后处理后为:

$$w_{ik} = \frac{w_{ik}}{\sqrt{\sum_{j \in d_i} w_{ij}^2}}$$

归一化的目的是使不同的文本具有相同长度。

文本经过分词程序分词后,首先使用停用词表去掉对分类没有贡献的词,还可采取特征词相关性分析、聚类、同义词和近义词归并等策略,最终表示成上面描述的文本向量。

3.3 特征集约减

特征集约减的目的有三个:1) 为了提高程序效率,提高运行速度;2) 数万维的特征对文本分类的意义是不同的,一些通用的、各个类别都普遍存在的特征对分类的贡献小,在某个特定的类中出现的比重大而在其他类中出现比重小的特征对文本的贡献大。3) 防止过拟合(Overfit)。对每一类,去除对分类贡献小的特征,筛选出针对反映该类的特征集合。

一个有效的特征集直观上说必须具备以下两个特点:

1) 完全性:确实体现目标文档的内容;

2) 区分性:能将目标文档同其他文档区分开来。

用向量空间法表示文档时,文本特征向量的维数往往达到数十万维,即使经过删除停用词表中的停用词以及应用 ZIP 法则删除低频词,仍会有数万维特征留下。最后一般只选择一定数量的最佳特征来开展各种文本挖掘工作,所以进一步对特征进行约减就显得异常重要。

通常,特征子集的提取是通过构造一个特征评估函数,对特征集中的每个特征进行评估,每个特征获得一个评估分数,然后对所有的特征按照评估分大小进行排序,选取预定数目的最佳特征作为特征子集。文本特征选择中的评估函数是从信息论中延伸出来的,用于给各个特征词条打分,很好地反映了词条与各类之间的相关程度。常用的评估函数有文档频数(document frequency),信息增益(information gain),期望交叉熵(expected cross entropy),互信息(mutual information),x2 统计(CHI),单词权(term strenth),文本证据权(the weight of evidence for text)和几率比(odd ratio)等。

3.4 文本挖掘方法

文本的特征表示是文本挖掘的基础,而文本分类和聚类是文本挖掘的最重要、最基本的挖掘功能,也是文本挖掘中应用的比较广泛的一个领域。

3.4.1 文本分类

文本分类是一种典型的监督式机器学习方法,一般分为训练(或学习)和分类两个阶段,具体如下:

3.4.1.1 训练阶段

1) 定义类别集合 $C = \{c_1, c_2, \cdots c_m\}$,这些类可以是层次型的,也可以是并列的;

2) 给出训练文档集合 $D = \{s_1, s_2, \cdots, s_n\}$,每个训练文档 s_j 的被标上所属类别标识 c_i;

3) 统计 D 中所有文档的特征矢量 $V(s_j)$,确定代表 C 中每个类别 c_i 的特征矢量 $V(c_j)$。

3.4.1.2 分类阶段

1) 对于测试文档集 $T = \{d_1, d_2, \cdots, d_n\}$ 中的每个待分类文档 d_k,计算其特征矢量 $V(d_k)$ 与每个 $V(c_i)$ 之间的相似度所 $sim(d_k, c_i)$;

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

有时只要 d_k 与这些类别间的相似度超过某个预定阈值,可为 d_k 指定多种类别。但若这种情况发生得太频繁,则说明预定义类别 $C = \{c_1, c_2, \cdots, c_m\}$ 不当,应加以修改。当文档 d 与所有类的相似度都低于该阈值,则将其标注为“其他”类。

衡量两个特征向量的近似程度,通过计算两个特征向量之间的距离,存在三种最通用的距离度量:欧氏距离、余弦距离和内积。因此计算 $sim(d_k, c_i)$ 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

$$sim(d_k, c_k) = \frac{V(d_k) \text{ 与 } V(c_i) \text{ 具有的相同词条数}}{V(d_k) \text{ 与 } V(c_i) \text{ 所有的词条数}}$$

最常用的方法是考虑两个特征矢量之间的夹角余弦,即

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

常用的文本分类方法有基于概率模型的方法,如朴素 Bayes 方法,隐马尔可夫模型等;基于关系学习的决策树方法等;基于统计学习的支持向量机方法等;基于向量空间模型的 K-近邻分类法和神经网络方法等。

3.4.2 文本聚类

文本聚类是一种典型的无监督式的机器学习方法,目前在文献中存在大量的文本聚类方法。聚类方法的选择取决于数据的类型、聚类目的和应用,大致可以分为层次凝聚法和平面划分法两种. 对于给定的文档集合 $D = \{d_1, d_2, \cdots, d_n\}$,层次聚类法的具体过程如下:

1) 将 D 中的每个文件 d_i 看作是一个具有单个成员的簇 $c_i = \{d_i\}$,这些簇构成了 D 的一个群集 $C = \{c_1, c_2, \cdots, c_n\}$;

2) 计算 C 中每对簇的 (c_i, c_j) 之间的相似度 $sim(c_i, c_j)$;

3) 选取具有最大相似度的簇对 $sim(c_i, c_j)$,并将 c_i 和 c_j 合并为一个新的簇 $c_k = c_k \cup c_j$,从而构成了 D 的一个新的群集 $C = \{c_1, c_2, \cdots, c_{n-1}\}$;

4) 重复上述步骤,直至 C 中剩下一个簇为止。

该过程构造出一棵生成树,其中包含了簇的层次信息以及所有簇内和簇之间的相似度。层次群集方法是最常用的群集方法,它能够生成层次化的嵌套簇,且准确度高。但是在每次合并时,需要全局地比较所有簇之间的相似度,并选择出最佳的两个簇,因此执行速度太慢,不适合大量文本聚类。

平面划分法与层次聚类的区别在于,它将文本集合水平地分割为若干个簇,而不是生成层次化的

嵌套,对于给定的文档集合 $D=\{d_1, d_2, \cdots, d_n\}$,平面划分法的具体过程如下:

1) 确定要生成簇的数目 k;

2) 按照某种原则生成 k 个群集中心作为群集的种子 $S=\{s_1, s_2, \cdots, s_n\}$;

3) 对 D 中的每个文档 d_i,依次计算它与各个种子 s_j 的相似度 $sim(d_i, s_j)$;

4) 选取具有最大相似度的种子 $\arg\max_{c_i, c_j \in C} sim(d_k, s_j)$,并将 d_i 化归为以 s_j 为群集中心的簇 c_j,从而构成了 D 的一个新的群集 $C=\{c_1, c_2, \cdots, c_k\}$;

5) 重复此步骤若干次,以得到最为稳定的群集结果。

该方法的执行速度较快,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。

常用的聚类划分方法有 K-平均算法和 K-中心算法。K-平均算法是划分方法中基于质心技术的一种算法,以 K 为参数,把 n 个对象分为 K 个簇,以使簇内具有较高的相似度,而簇间的相似度较低,相似度的计算根据一个簇内对象的平均值(质心)来计算。K-平均算法对于孤立点敏感,为消除这种敏感性,不采用簇中对象平均值作为参考点,而选用簇中位置最中心的对象为参考点,这就是 K-中心算法。

通过上面的分析可知,文本挖掘和数据挖掘在目的上是一致的,都是试图从大量的信息中抽取知识。数据挖掘是从原始数据中抽取,而文本挖掘则是从文本材料中抽取。如果将数据的概念泛化,文本挖掘也就可以看成一种数据挖掘,但是数据挖掘倾向于非常精确和结构化,大多数研究只考虑从数据库中抽取知识,这正是许多数据挖掘技术并不能自如地应用于文本挖掘领域的原因。另外在对文本集进行相关分析时,往往会损失文本中的大量信息,这种信息的遗漏,会影响到挖掘的效果,因此还要探索更高效的文本挖掘新方法。近年来随着 WWW 网的普及,人们越来越迫切需要从浩瀚的 Web 信息资源中发现潜在的、有价值的知识,因此 Web 文本挖掘已经是文本挖掘的主要研究方向。

[参考文献]

[1] 孙建涛, 沈抖, 陆玉昌, 等. 网页分类技术[J]. 清华大学学报(自然科学版), 2004, 44(1): 65 - 68.

[2] 唐焕玲. 文本自动分类方法研究[D]. 北京: 清华大学计算机系, 2003.

[3] 林杰斌, 刘明德, 陈湘. 数据挖掘与 OLAP 理论与务实[M]. 北京: 清华大学出版社, 2003.

[4] Yang Y M. An evaluation of statistical approach to text categorization[R]. In Technical Report CMU-CS-97-127. Computer Science Department, Carnegie Mellon University, 1997.

[5] Han J W, Micheline Kamber M′. 数据挖掘概念与技术[M]. 北京: 机械工业出版社, 2002.

Research On Text Data Mining

XUE Wei-min[1,2], LU Yu-chang[2]

(1. Automation College of Beijing Union University, Beijing 100101, China;
2. Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

Abstract: Text mining is an important part of data mining and it has been widely used in varied fields. A systemic discussion about the text mining is presented. Firstly the definition of data mining, text mining and web mining, and the research directions of the text mining are discussed. Then the process of text mining and its key technologies is introduced in detail. Finally the sum of the text mining and the future of application are represented.

Key words: text mining; data mining; web mining; text mining mode

(责任编辑　彭丹宇)