计算机科学 2002Vol. 29№. 12

文本挖掘及其关键技术与方法*)

The Text Mining and its Key Technigues and Methods

王丽坤 王 宏 陆玉昌

(智能技术与系统国家重点实验室 清华大学计算机科学与技术系 北京 100084)

Abstract With the dramatically development of Internet, the information processing and management technology on WWW have become a great important branch of data mining and data warehouse. Especially, nowadays, Text Mining is marvelously emerging and plays an important role in interrelated fields. So it is worth summarizing the content about text mining from its definition to relational methods and techniques. In this paper, combined to comparatively mature data mining technology, we present the definition of text mining and the multi-stage text mining process model. Moreover, this paper roundly introduces the key areas of text mining and some of the powerful text analysis techniques, including: Word Automatic Segmenting, Feature Representation, Feature Extraction, Text Categorization, Text Clustering, Text Summarization, Information Extraction, Pattern Quality Evaluation, etc. These techniques cover the whole process from information preprocessing to knowledge obtaining.

Keywords Text mining, Knowledge discovery in database, Data mining, Word automatic segmenting, Feature representation, Feature extraction, Text categorization, Text clustering

从1969年美国国防部的计算机网络ARPANET起步，至今已有32年历史的Internet，已经发展成为包含多种信息资源、站点遍布全球的巨大信息服务系统，为其用户提供了极具价值的、巨大的数据资料。在数字图书馆和Internet上，在线可获得的信息量呈指数级增长，导致了信息爆炸。WWW以超文本的形式呈现给用户，一个网页里包含了多种不同的数据类型，其中最主要的信息源就是文本数据。文本表达了大量的、丰富的信息，同时包含了许多未被所有者发现的潜在知识。面对浩瀚的文本资源，传统的文档和文本处理工具已经不能满足用户的需求。于是在人工智能研究领域结合结构化数据库中的数据挖掘技术，提出了一种有效的、可以充分利用这些文本数据的新的信息处理技术——文本挖掘(Text Mining)。

文本挖掘涉及多个学科领域：数据库、信息检索、信息提取、机器学习、自然语言处理、计算语言学、统计数据分析、线性几何、概率理论，甚至还有图论。本文以结构化数据库中知识发现的数据挖掘模型为依托，提出了对应的文本挖掘模型，总结和介绍了模型中涉及的关键方法和技术。在论文的第一部分介绍了几种文本挖掘的定义，并对其加以总结给出了一个较为全面的定义。按挖掘对象和内容，提出了文本挖掘的分类。结合文本挖掘的定义，从数据挖掘模型出发，提出了文本挖掘的处理模型，以期给出综合的文本挖掘概貌。模型中主要涉及五个方面：信息预处理、特征表示、特征提取、文本挖掘技术、模型质量评价，因此论文在后面的阐述中也将按照这一文本挖掘框架进行组织。

1 文本挖掘的定义及其处理模型

1.1 文本挖掘的定义

国际上第一次关于数据挖掘与知识发现的研讨会于1989年6月在美国底特律召开。当时提出的数据挖掘技术基于结构化数据库中的海量数据处理。近年来，Data Mining在研究和应用方面发展迅速，尤其是在商业和银行领域的应用比研究的发展速度还要快。

根据W. J. Frawley 和 G. P. Shapiro[1]等人提出的定义：

知识发现KDD(knowledge discovery in database)：是指从大量数据中提取出可信的、新颖的、有效的，并能被人理解的模式的高级处理过程。

数据挖掘(Data Mining)：是指从大型数据库的数据中提取出人们感兴趣的知识，这些知识是隐含的、事先未知的、潜在的有用信息。

随着网络时代的到来，用户可获得的信息包含了从技术资料、商业信息到新闻报道、娱乐资讯等多种类别和形式的文档，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。而这个数据库中存放的是非结构化的文本数据。结合人工智能研究领域中的自然语言理解和计算语言学，从数据挖掘中派生出了两类新兴的数据挖掘研究领域：网络挖掘和文本挖掘。网络挖掘侧重于分析和挖掘网页相关的数据，包括文本、链接结构和访问统计(最终形成用户网络导航)。一个网页里包含了多种不同的数据类型。因此网络挖掘就包含了文本挖掘、数据库中数据挖掘、图像挖掘等。文本挖掘作为一个新的数据挖掘研究领域，目前并没有给出统一的、确切的定义，但是文本挖掘的目的就是把文本信息转化为人可利用的知识。从这一目的出发，有如下5种关于文本挖掘的定义。

定义1[13] 文本挖掘是指对大规模文档集的处理和从文本数据集中提取隐含的知识。

定义2[14] 文本挖掘是信息处理和信息管理中的重要研究问题。它基于语义学，使用贝叶斯模型、概率理论、向量空间模型、统计模型甚至是图论，从文档中挖掘出知识模式，及短语结构。

定义3[15] 文本挖掘是使用计算语言学规则从文本中提取信息的研究和应用方法。文本挖掘的关键领域包括：特征提取、主题索引、聚类、摘要。

定义4[16] 文本挖掘是指从自然语言文本中提取模式，可以定义为按照特定目标从文本中提取信息的分析过程。

定义5[17] 文本挖掘结合数据挖掘的规则、信息提取、信息检索、文本分类、概率模型、线性几何、机器学习、计算语言

*)得到国家基础研究项目(973)(G1998030414)和清华大学信息学院基础研究的(985)的资助。王丽坤 硕士研究生。王 宏 副教授。陆玉昌 教授。

学去发现文本集中的结构、模式、知识。

根据上述描述并结合数据挖掘的定义，总结出文本挖掘的定义为：文本挖掘(Text Mining)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。

1.2 文本挖掘的分类

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。

基于单文档的数据挖掘 对文档的分析并不涉及其它文档。主要的文本挖掘技术有：文本摘要(TextSummarization)、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。

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。主要的技术有：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。

1.3 文本挖掘处理模型

本文文本挖掘的处理模型是在知识发现的基础上提出的，所以先介绍多阶段的数据库中知识发现的处理模型。

数据挖掘是KDD中最重要的处理阶段，因此人们往往不加区别地使用两者[2]。图1所示是Usama M. Fayyad等人给出的KDD处理模型[3]，这是一个多阶段的数据库中知识发现的处理模型。

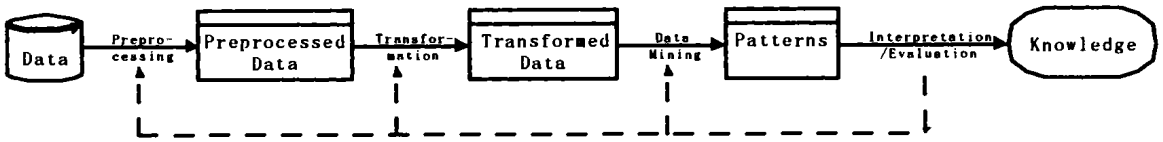

图1 多阶段的数据库中知识发现的处理模型

该处理模型中，主要有以下五个关键阶段：数据准备；数据挖掘；评估、解释模式模型；巩固知识；运用知识。其中，数据挖掘是KDD最关键的步骤，也是技术难点所在。数据挖掘根据模型中提出的目标，选取相应算法并设置参数，经过分析数据，最终得到可能形成的知识模式。

在整个KDD过程中，可能需要多次的循环反复，每一个步骤一旦与预期目标不符，都要回到前面的步骤，重新调整，重新执行。

根据这个框架结合文本挖掘的定义和特点，给出如图2所示的基于KDD的多阶段文本挖掘处理模型。开始处是文本信息源，最终结果是用户获得的知识模式。经历了从信息预处理→文本挖掘→质量评价三个主要阶段。

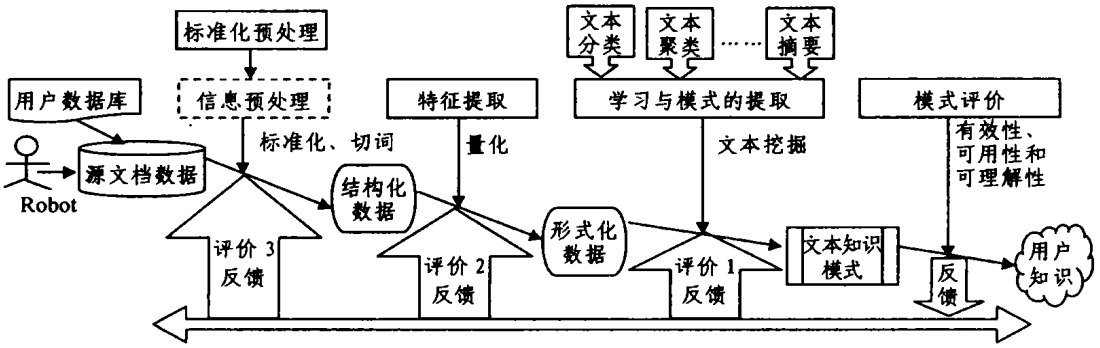

图2 基于KDD的多阶段文本挖掘处理模型

2 信息预处理

预处理是文本挖掘的第一个步骤，也是比较重要的一个步骤。数据准备是否做好将直接影响到文本挖掘的效率和准确度以及最终模式的有效性。

文本挖掘处理的是大量的、非结构化的文本数据，这些数据一般是长期积累的结果，且没有统一的结构。因此不仅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，如：数据的选择(选择相关的数据内容)、净化(消除噪音、冗余数据)、推测(推算缺失数据)、数据缩减(减少数据量)，而且文本使用自然语言描述，计算机所能理解的只是它的ASCII码，根本就是没有意义的，所以还需要进行文本数据的信息预处理。

在这里信息预处理指的是抽取代表文本特征的元数据(特征项)，对元数据进行标记、语形学分析、词性标注、短语边界辨认等。一般"词"能表达完整的语义对象，所以通常选用词作为文本特征的元数据。而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，字的信息价值比较低，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。汉语语义及结构上的复杂性和多边性给中文自动分词带来极大困难，也成为中文文本信息处理中的技术难点之一。下面就集中介绍中文分词的技术方法。

汉语分词问题的一般描述[4] 从统计学的角度，汉语分词是一个求最大概率问题。设有一含有n个汉字的待切分字串 $C=c_1c_2\cdots c_n$ 和相应的切分词典 $D=\{d_1d_2\cdots d_k\}$，其切分结果可能有 $M(M\geqslant 1)$ 个，记作：$W=\{W_i|W_i=w_{i,1}w_{i,2}\cdots w_{i,mi}$ 其中 $w_{ij}\in D, 1\leqslant i\leqslant M, 1\leqslant j\leqslant m_i\}$。但其中只有某个 $W_i\in W$ 是正确的。汉语分词的任务就是从可能的切分集W中挑选这样一个最满足语言学的切分串 W_i。于是，分词问题转化为：在所有可能的切分序列W中，选取具有最高评分 $P(W_i|C)$ 的切分串作为最佳切分串 W'，即：$W'=\arg\max_{1\leqslant i\leqslant M}P(W_i|C)$。

分词方法　目前，汉语分词主要有两大类方法：基于词典与规则的方法和基于统计的方法。基于词典与规则的方法应用词典匹配、汉语词法或其它汉语语言知识进行分词，如：最大匹配法(Maximum Matching)[5]、最小分词方法[6]等。这类方法简单、分词效率较高，但对词典的完备性、规则的一致性等要求比较高。基于统计的分词方法则将汉语基于字和词的统计信息，如相邻字间互信息、词频及相应的贡献信息等应用于分词，由于这些信息是通过训练集动态获得，因而具有较好的鲁棒性能，但是完备性相对比较差。

在这两大类方法基础上又可将分词的基本方法归纳为如下几种[7]：

· 词典匹配法：如最大匹配法、逆向匹配法、增字或减字匹配法、双向扫描法、二次扫描法、逐词遍历法、部件词典法。

· 设立标志法：如切分标志法、统计标引法、多层次列举法。

· 词频统计法：如高频优先法、基于期望法、最少分词词频法。

· 联想词群法：如联想回溯 AB 法、词链法、多遍扫描联想法、联想树分析法、无词库法。

· 语义语用法：如邻接约束法、扩充转移网络法、综合匹配法、后缀分词法。

· 知识与规则法：如切此规则法、切分与语义校正法、规则描述切词法、生成-测试法、语境相关法、短语结构法、词语结构类比法。

· 人工智能法：如专家系统法、神经网络方法等。专家系统分词法：将自动分词过程看作是知识推理过程，力求从结构与功能上分离分词过程和实现分词所依赖的汉语词法知识、句法知识以及部分语义知识。把知识的表示、知识库的逻辑结构与知识库的维护放在系统设计的首位考虑。其知识库按常识性知识与启发性知识分别进行组织。对于常识性分词知识采用"语义网络"表示，对于启发性分词知识采用"产生式规则"表示。知识库是使专家系统具有"智能"的关键性部件。基于神经网络的分词方法：以模拟人脑运行，分布处理和建立数值计算模型工作的。它将分词知识所分散隐式的方法存入神经网络内部，通过自学习和训练修改内部权值，以达到正确的分词结果。这种方法的关键在于知识库(权重链表)的组织和网络推理机制的建立。

在诸多的分词方法中，最大匹配法虽然处理精度不高但是简单高效，因此被广泛运用。

总之，汉语自动分词是中文信息处理的"瓶颈"问题，它的最终解决依赖于汉语的分词结构、句法结构、语义等语言知识的深入系统的研究；依赖于对语言与思维的本质的揭示；同时，在很大程度上还依赖于神经网络、专家系统、知识工程等人工智能技术的研究进展。

3　特征表示

文本表达了巨大的、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。

下面介绍两种主要的文本特征表示法：矢量空间模型(VSM)、布尔模型。其中 VSM 是近年来应用较多且效果较好的方法之一[8]。

矢量空间模型(Vector Space Model)　它把文档看作是由一组正交词条矢量所张成的矢量空间，每个文档 Doc 表示为其中的一个范化特征矢量：$\vec{V}(Doc)=(val(t_1),\cdots,val(t_i),\cdots,val(t_m))$，其中 t_i 是词条项，$val(t_i)$是 t_i 在 Doc 中在文档中的重要程度。即将文档看作为是由一组相互独立的词条组 $\{t_1,t_2,\cdots,t_m\}$构成，把 $t_1,t_2,\cdots,t_m$ 看成一个 m 维坐标系中的坐标轴，对于每一词条 t_i 根据其重要程度赋以一定的权值 $0\leqslant val(t_i)\leqslant 1$，$val(t_i)$就作为对应坐标轴的坐标值。这样由$\{t_1,t_2,\cdots,t_m\}$分解而得的正交词条矢量组就构成了一个文档向量空间，每篇文档则映射成为这个空间中的一个点。对于所有文档和用户目标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。m 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。假设用户目标为 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：

$$similarity(V_i,U)=\cos(V_i,U)$$

$$=\sum_{k=1}^{m} val_{V_i}(t_k)\cdot val_U(t_k)\Big/\sum_{k=1}^{m} val_{V_i}(t_k)^2\sum_{k=1}^{m} val_U(t_k)^2$$

可将 VSM 法形式化描述为：映射函数 $f:T\to[0,1]$，其中 $T=\{t_1,t_2,\cdots,t_m\}$是文档元数据的集合，$0\leqslant val(t_i)\leqslant 1$，$val(t_i)$将描述 t_i 相对于文档和文档集的重要性。

布尔模型　是 VSM 模型的一种简化。它是一种简单的严格匹配向量模型(Exact Match Model)，定义了一个二值映射函数 $f:T\to\{0,1\}$，元数据 t_i 的值不再是权值，而是一个布尔值。文本表示的结果是 0-1 向量，即：

$$Doc=(val(t_1),val(t_2),\cdots,val(t_m))$$

其中 $\begin{cases} val(t_i)=0 & t_i \text{ 在 } Doc \text{ 中没有出现；} \\ val(t_i)=1 & t_i \text{ 在 } Dos \text{ 中出现。} \end{cases}$

布尔模型实现简单，用于检索速度很快。它只需要进行简单的 0-1 匹配就能判断检索条件同文档的关系，从而将检索文档分为两个集合：匹配集和非匹配集。但因为布尔模型忽略了元数据的文档项频，所以无法在匹配结果集中进行相关性大小的排序。且逻辑表达式过于严格，往往会因为一个条件不满足而忽略了其它的重要特征，造成大量的遗漏。但速度快是它的优势，所以在许多检索系统中得到应用，如 Yahoo、搜狐等诸多著名网络检索站点均采用了布尔模型。

上述的文本表示方法基于词条正交，没有对此条间的相关性信息进行描述。所以如果要对文本进行更有效、更智能的数据挖掘，这种表示法还需要改进，或需要寻求更好的表示模型。这必然是目前及今后研究的重点、难点之一。

4　特征提取(缩减)

同数据挖掘使用固定的属性特征不同，用于文本挖掘的特征属性是灵活的、多变的。抽象概念难于表示、难于形式化，文本特征往往是高维的，根据 Dunja Mladenic 和 Marko Grobelnik 在著名的搜索引擎 YAHOO 上的实验结果表明，对全体文档集进行特征表示时，其维数将高达 69,280～255,602[9]维。而另一方面文档的许多信息又是高冗余的，所以文本特征的提取(缩减)是相当重要的，这往往决定了文本挖掘的效率。

对目标表示中词条 t_i 的选取被称为特征提取。主要有两大类方法：独立评估方法和综合评估方法(代表方法：主成分分析方法[27~29])。前者的基本思想是对特征集中的每个特征进行独立的评估，让每个特征都获得一个权值，然后按权值大小排序，根据权阈值或预定的特征数目选取最佳特征子集作为特征提取的结果。后者则是从高维的、彼此间不独立的原始特征集中找出较少的描述这些特征的综合指标，且这些综合

指标之间相互独立，然后又用得到的综合指标对特征集进行特征选择。

独立评估方法 对元数据的权值评价有多种标准：文本权重(Text Weight)[25]、信息收益(Information Gain)[30]、期望交叉信息熵(Expected Cross Entropy)[31]、互信息(Mutual Information)[32]、文本证据权(Weight of Evidence for Text)[33]、奇率(Odds Radio)[34]、词频(Word Frequency)[32]等。这些评价标准来自：神经元网络法、决策树法、遗传算法、集合论法、统计法等。下面就逐一介绍。

文本权重(Text Weight) 在自然语言文档中，各词条在不同内容的文档中所呈现的频率分布是不同的，因此可根据词条的频率特性用统计的方法进行特征权值评价。

一个有效的特征项集，必须具备两个特性：I. 完全性：特征项能够概括目标内容；II. 区分性：根据特征项集，可以对不同目标加以区分。

根据以上两个特性可得：在文档中，词条的重要性与词条的项频(某一文档内的频数)成正比，而与外频(训练文档集中出现该词条的文档频数)成反比，所以可构造文档 Doc_i 的词条权值评价函数为：$val_i(t_k)=tf_{ik}\cdot\log(\frac{N}{n_k}+\alpha)$，其中 tf_{ik} 表示词条 t_k 的项频；N 是训练文档集的文档数目；n_k 表示词条 t_k 的外频；α 是一个常数。为避免因文档长度引起的频数变化，还应对其作规范化处理，即：

$$val_i(t_k)=\frac{tf_{ik}\cdot\log(\frac{N}{n_k}+\alpha)}{\sqrt{\sum_{k=1}^{m}tf_{ik}^2\cdot\log^2(\frac{N}{n_k}+\alpha)}}。$$

对 WWW 页面中的文档，在计算特征项权值时还可考虑 HTML 语言提供的结构特征，如〈A〉域、〈Title〉域、〈Meta〉域、〈H〉域等的标记信息。这些信息通常对其页面内容有着很高的概括性，在计算权值时对出现在上述域中的词条可以赋以较高的权重。

信息收益(Information Gain) 信息收益多在决策树中使用。

$$InfGain(t_k)=P(t_k)\sum_i P(C_i|t_k)\log\frac{P(C_i|t_k)}{P(C_i)}+P(\bar{t}_k)\sum_i P(C_i|\bar{t}_k)\log\frac{P(C_i|\bar{t}_k)}{P(C_i)}。$$

期望交叉信息熵(Expected Cross Entropy)

$$CrossEntropyTxt(t_k)=P(t_k)\sum_i P(C_i|t_k)\log\frac{P(C_i|t_k)}{P(C_i)}。$$

与信息收益的区别是期望交叉信息熵不考虑文档中未出现的词条的信息。

互信息(Mutual Information)

$$MutualInfoTxt(t_k)=\sum_i P(C_i)\log\frac{P(t_k|C_i)}{P(t_k)}。$$

文本证据权(Weight of Evidence for Text)

$$WeightOfEvidTxt(t_k)=\sum_i P(C_i)\times P(t_k)\times\left|\log\frac{P(C_i|t_k)(1-P(C_i))}{P(C_i)(1-P(C_i|t_k))}\right|。$$

奇率(Odds Radio) 按照特征词与类的相关性给出特征的相关度：

$$OddsRatio(t_k)=\log\frac{odds(t_k|C_i)}{odds(t_k|\bar{C}_i)}=\log\frac{P(w_k|C_i)(1-P(t_k|\bar{C}_i))}{(1-P(t_k|C_i))P(t_k|\bar{C}_i)}。$$

有三种扩展形式[35]：

I. $FreqOddsRatio(t)=Freq(t)\times OddsRatio(t)$；

II. $FreqLogP(t)=Freq(t)\times\log\frac{P(t|C_1)}{P(t|C_2)}$；

III. $ExpP(t)=e^{P(t|C_1)-P(t|C_2)}$。

词频(Word Frequency)

$Freq(t_k)=TF(t_k)$。

其中，t_k 表示词集 V 中的第 k 个词；$P(t_k)$是词 t_k 发生的概率；$\bar{t}_k$ 表示词 t_k 未出现；$P(C_i)$表示第 i 类的概率值；$P(C_i|t_k)$是给定词 t_k 类 C_i 发生的条件概率；$P(t_k|C_i)$是给定类 C_i 词 t_k 发生的条件概率；$P(t_k|pos)$是给定正例时，词 t_k 发生的条件概率；$P(t_k|neg)$是给定反例时，词 t_k 发生的条件概率；$TF(t_k)$是词 t_k 的出现频率。

Dunja Mladenic[35]提供的实验数据表明奇率的特征评价效果是优于其它方法的。

主成分分析方法 原始特征和主成分之间的映射关系可解释为每个主成分是原始特征的线性组合。一般设文档 D 表示为一个 p 维向量：$D=(t_1,t_2,\cdots,t_p)'$，主成分 Y 表示为一个 m 维向量：$Y=(y_1,y_2,\cdots,y_m)'$。进行变换：$Y=AD$，$A=\begin{bmatrix}a_{11}&a_{12}&\cdots&a_{1p}\\a_{21}&a_{22}&\cdots&a_{2p}\\\vdots&\vdots&&\cdots\\a_{m1}&a_{m2}&\cdots&a_{mp}\end{bmatrix}=\begin{bmatrix}a_1'\\a_2'\\\vdots\\a_m'\end{bmatrix}$。根据如下两条原则进行系数 a_{ij}的选择：(1)y_i 和 $y_j(i\neq j,i,j=1,2,\cdots,m)$互不相关；(2)$y_1$ 是 $t_1,t_2,\cdots,t_p$ 的一切线性组合中方差最大的，即使得 $a_1'D$ 具有最大的方差；y_2 是与 y_1 不相关的在 $t_1,t_2,\cdots,t_p$ 的一切线性组合中方差最大的，y_m 是与 $y_1,y_2,\cdots,y_{m-1}$都不相关的，在 $t_1,t_2,\cdots,t_p$ 的一切线性组合中方差最大的。分别称 $y_1,y_2,\cdots,y_m$ 为原始随机变量的一个，第二个，…，第 m 个主成分。即满足 $a_1'a_1=1$ 的条件下，使得主成分 y_1 的方差 $Var(y_1)=a_1'\Sigma a_1$ 达到最大，其它主成分可依次求出。

因此可令 $a_i=(a_{i1},a_{i2},\cdots,a_{ip})'$等于 $u_i=(u_{1i},u_{2i},\cdots,u_{pi})$，$u_i$ 是 D 的协方差矩阵 Σ 的特征根 $\lambda(\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_p)$所对应的特征向量。所以 $Y=U'X$，则有 $X=UY$。此时称矩阵 $U_{p\times m}$ 为主成分的载荷矩阵，u_{ij}为 D 的第 i 个变量 t_i 在第 j 个主成分 y_j 上的载荷。U 的第 i 列反映的是主成分 y_i 对原始指标 $D=(t_1,t_2,\cdots,t_p)'$的各个分量的作用。通过分析载荷矩阵的具体形式，可以得到各个主成分对每个原始指标的贡献，因而可以利用载荷矩阵的分析进行特征选择，保留那些有多个主成分对其起作用的原始指标，而剔除只有少数主成分对其起作用的原始特征，达到特征选择的目的。

5 文本挖掘技术

进行文本挖掘的主要目标有：文本分类、文本聚类、文本总结、信息提取等。其中文本分类问题是最重要的，也是应用较多的文本挖掘技术。

5.1 文本分类

文本分类是指按照预先定义的主题类别，为文档集合中的每篇文档确定一个类别。这样用户不仅可方便地阅读文档，而且可以通过限制搜索范围来使文档查找更容易。目前，YAHOO 已实现用文本自动分类技术对 Web 文档进行自动分类，这大大提高了效率。目前文本分类方法很多，下面将介绍几种较为典型且分类正确率较高的方法：简单贝叶斯分类法(Naive Bayesian Classifier)[11,12]、矩阵变换法、K 最近邻参照分类算法(K-Nearest Neighbor)[9]，及利用 Boosting 方法解决兼类问题的技术[36,37]。

Naive Bayes 分类法 是一种利用贝叶斯网络(Bayesian Network)进行文本分类的概率模型。常用的 Naive Bayes 分类法主要有两种模型：多变量贝努利模型(Multi-variate Bernoulli Model)和多事件模型(Multinomial Model)。之所以

称为 Naive，是因为它基于贝叶斯假设(Naive Bayes Assumption)：对给定的类内容，样本中的所有属性相互独立。

贝叶斯分类法假设文本数据为一个参数模型，使用训练样本进行贝叶斯最小错率估计(Bayes-optimal estimates)。对新的测试文档使用贝叶斯规则计算文档的后验概率，进行分类。其框架描述如下：

类 c_j 的概率表示为：$P(c_j|\theta)=(\sum_{i=1}^{|D|}P(c_j|d_i))/|D|$；每一个文档 d_i 的概率表示为：$P(d_i|\theta)=\sum_{j=1}^{|C|}P(c_j|\theta)P(d_i|c_j;\theta)$。其中，$\theta$ 是模型参数；C 是类别集合；$c_j\in C$，代表第 j 类。

判断一个文档 d_i 是否属于某个类 c_j，可通过计算 $P(c_j|d_i;\theta)=\frac{P(c_j|\theta)P(d_i|c_j;\theta)}{P(d_i|\theta)}$ 的概率完成，即给定文档 d_i，它属于文档类 c_j 的概率是多少。而 Naive Bayes 规则就是把文档 d_i 指定到使 $P(c_j|d_i)$ 达到最大概率的 c_j 类中，即求解 $\arg\max P(c_j|d_i)$。

下面两种模型的不同之处表现在对参数的估计使用不同的规则。

①多变量贝努利模型(Multi-variate Bernoulli Model)。是基于多变量贝努利事件模型，以文档为“事件(event)”，一个词元的出现与否看作事件的属性。所以使用布尔模型表示文本，不考虑一个词元在文档中出现的次数。使用所有属性变量(包括文档中未出现的词)来计算文档概率。该方法适用于属性比较固定的情况。在该模型中：$P(d_i|c_j;\theta)=\prod_{k=1}^{|V|}B_{ik}P(t_k|c_j;\theta)+(1-B_{ik})(1-P(t_k|c_j;\theta)))$，其中 B_{ik} 表示文本 d_i 的布尔向量表示中的第 k 维的布尔值；t_k 表示词集 V 中的第 k 个词；$P(t_k|c_j;\theta)$ 表示类 c_j 中出现词 t_k 的概率。而 $\hat{\theta}_{t_k|c_j}=P(t_k|c_j;\theta)=\frac{1+\sum_{i=1}^{|D|}B_{ik}P(c_j|d_i)}{2+\sum_{i=1}^{|D|}P(c_j|d_i)}$，其中 $P(c_j|d_i)\in\{0,1\}$ 表示文档 d_i 属于类 c_j 的概率。

②多项模型(Multinomial Model)。以词元为“事件(event)”，把文档看作词元事件的集合。记录一个词元在文档中的出现频率。计算文档概率时仅使用文档中出现的词元。该方法常用于计算语言模型中，又被称为单元语言模型(Unigram Language Model)。

在该模型中：$P(d_i|c_j;\theta)=P(|d_i|)|d_i|!\prod_{k=1}^{|V|}\frac{P(t_k|c_j;\theta)^{N_{ik}}}{N_{ik}!}$，其中 N_{ik} 表示文本 d_i 中词 t_k 出现的次数。$\hat{\theta}_{t_k|c_j}=P(t_k|c_j;\hat{\theta}_j)=\frac{1+\sum_{i=1}^{|D|}N_{ik}P(c_j|d_i)}{|V|+\sum_{s=1}^{|V|}\sum_{i=1}^{|D|}N_{si}P(c_j|d_i)}$，其中 $0\leqslant\theta_{t_k|c_j}\leqslant 1$ 且 $\sum_k\theta_{t_k|c_j}=1$。

实验表明[11]在词集规模较小的情况下，多变量贝努利模型优于多项模型。但对大规模的文档集而言，情况正好相反，多项模型平均可使使用多变量贝努利模型时的错误率减少 27%。

虽然对文本的分类处理作了很多理想化假设，但 Naive Bayes 法仍然能得到较高的分类正确率。这是因为最终得到的贝叶斯分类器仅只是一个符号函数，该函数的近似程度可能很差，但是仍然能得到很高的分类精度。

矩阵变换分类法 其主要思想是为文档集和文档类分别建立向量空间，通过矩阵变换找到文档与类之间的映射关系，从而把分类问题转化为矩阵变换的数学问题来解决。文档空间的项表示文档中的词，而文档类空间的项则根据不同情况，分别选用文档表示符或描述词。

下面用一个例子说明矩阵变换分类法的分类过程。假设文档集中共 6 篇训练文档 D_1、D_2、D_3、D_4、D_5、D_6，文档分别属于 4 个文档类 C_1、C_2、C_3、C_4，其中 $D_1\in\{C_1,C_2\}$、$D_2\in\{C_3\}$、$D_3\in\{C_1,C_3,C_4\}$、$D_4\in\{C_1,C_2,C_3,C_4\}$、$D_5\in\{C_2\}$、$D_6\in\{C_3,C_4\}$。矩阵 *Doc* 表示文档空间，矩阵 *Doc* 的行表示一篇文档，列表示文档集中出现的词，*Doc* 中的一个元素 doc_{ij} 表示词条 t_j 在文档 D_i 中的权重。矩阵 *Class* 表示文档类空间，矩阵 *Class* 中的行表示一篇文档 D_i 所属的类别，列表示文档类标识符，*Class* 中的元素 $class_{ij}$ 表示文档 D_i 是否属于类 C_j，若 $D_i\in C_j$，则 $class_{ij}=1$；若 $D_i\notin C_j$ 则 $class_{ij}=0$。*Doc*、*Class* 矩阵如下所示：

$$Doc:=\begin{array}{c}\begin{matrix}t_1 & t_2 & t_3 & t_4 & t_5 & t_6\end{matrix}\\\begin{bmatrix}8.5 & 7.2 & 5.0 & 0. & 1.0 & 0.\\3.0 & 0. & 1.0 & 8.5 & 7.2 & 4.2\\7.2 & 0. & 3.0 & 5.0 & 3.0 & 7.2\\5.0 & 2.2 & 7.2 & 3.0 & 5.0 & 7.2\\3.0 & 8.5 & 8.5 & 0. & 0. & 0.\\5.0 & 0. & 3.0 & 7.2 & 8.1 & 5.0\end{bmatrix}\end{array}$$

$$Class:=\begin{array}{c}\begin{matrix}C_1 & C_2 & C_3 & C_4\end{matrix}\\\begin{bmatrix}1 & 1 & 0 & 0\\0 & 0 & 1 & 0\\1 & 0 & 1 & 1\\1 & 1 & 1 & 1\\0 & 1 & 0 & 0\\0 & 0 & 1 & 1\end{bmatrix}\end{array}$$

矩阵 *Dos*、*Class* 之间的关系用矩阵 F 表示：$F*A^T=B^T$，即 $F=B^T*(A^T)^{-1}$(对矩阵 *Doc* 不是 n 阶方阵的情况，可用奇异值分解法求得广义矩阵的逆矩阵)。经过计算得到 F 为：

$$F:=\begin{bmatrix}-.0077 & .422 & -.419 & -.283 & .124 & .466\\-.1181 & .451 & -.292 & -.273 & .214 & .340\\-.017 & .038 & -.032 & .0312 & .0362 & .133\\.217 & -.617 & .540 & .115 & -.1028 & -.340\end{bmatrix}$$

矩阵 F 反映了文档词条和文档类之间的相关性。F 中的行对应于文档类，列对应于文档词条，f_{ij} 反映了文档类和词条间的相关度。求得转换矩阵 F 后，即可用它来进行文档分类。对任意的文档向量 $D_i=(t_1,t_2,\cdots,t_n)$，它在文档类空间的投影 $C=(c_1,c_2,\cdots,c_m)$，可由 F 导出：$C=(F*D_i^T)^T$。C 中的各分量 c_1，$c_2,\cdots,c_m$ 反映了 D_i 与各文档类 $C_1,C_2,\cdots,C_m$ 的相关度，把 $c_1,c_2,\cdots,c_m$ 从大到小排序，就可得到文档的所属类别。

K 最近邻参照分类算法 把对一个文档的所属类别范畴的预测建立在对与之最为相似的 k 个文档所属类别的概率分布上。文档 *Doc* 属于 C_j 类的概率为：

$$P(C_j|Doc)=\frac{\sum_{t=1}^{k}similarity(Doc,D_t)P(C_j|D_t)}{\sum_{i=1}^{|C|}\sum_{t=1}^{k}similarity(Doc,D_t)P(C_i|D_t)}$$

其中 D_t 为与文档 *Doc* 最近邻的 k 个文档之一。

它既可以按不同的概率属于不同的类别，也可以属于唯一的一个类别 C_h，此时

$$\begin{cases}P(C_j|D_t)=1 & \text{当 } j=h \text{ 时；}\\P(C_j|D_t)=0 & \text{其它。}\end{cases}$$

两个文档间的相似程度通常使用向量夹角的余弦来度量，即：

$$similarity(Doc,D_t)=cos(\vec{Doc},\vec{D_t})=\frac{\sum_i doc_i\cdot d_i}{\sqrt{\sum_i doc_i^2\sum_i d_i^2}}$$

AdaBoost.MH 和 AdaBoost.MR 上述方法都只考虑文档的单标签情况，即一个文档只属于一个类别。然而就实际问

题而言，一个文档往往可以属于多类。对此通常采用的解决方案是把任务分解成不相交的二值分类问题，每个类对应这样的一个二值问题。分类时，所有分类器都对未知类文档进行预测并给出单个决策，结果是文档可能属于的类的列表，或是对可能主题的排序。这种方法的缺点是它忽略了这些类别间的任何相关性。对训练样本来说，如果把样本打散成为多个单标签样本，它的弊端是使本来就较难刻画的问题空间决策面变得更加模糊，对学习任务可能产生更大的干扰。对这方面的研究不多，但值得注意的是 Boosting 组合学习方法中的 AdaBoost.MH 和 AdaBoost.MR 可用于解决多类多标签问题。

AdaBoost.MH 和 AdaBoost.MR 是 1998 年 Schapire 和 Singer 提出的基于汉明距离和损失排序的多类多标签 Boosting 算法。

Boosting 中的"兼类"问题，即每个待分类对象可以属于多个类别中的一个或多个，描述如下：

设 C 为标签或类的有限集，且令 $k=|C|$。在多标签情况下，每个实例 $x_i \in X$ 可能属于 C 中的多个标签。因此，训练样本是 (x_i, C_i) 对，其中 $C_i \subseteq C$ 是 x_i 的标签集。分类任务是在样本空间中寻找最小化 $H(x_i) \notin C_i$ 的概率的假设 $H: X \to C$。因为它度量的是连一个标签都不正确的概率，所以称为假设 H 的 *one-error*。用 $one\text{-}error_D(H)$ 表示假设 H 在样本集上关于分布 D 的 *one-error*，即：$one\text{-}error_D(H)=Pr_{(x,C)\sim D}[H(x) \notin C]$。

AdaBoost.MH：基于汉明损失，它直接以兼类样本为训练数据：$\{(x_1, C_1), \cdots, (x_m, C_m)\}$，其中 $x_i \in X, C_i \subseteq C$。其思想是对每个样本 x_i 和每个标签 y，提问："对样本 x_i，正确标签是 y，还是其他？"，从而把问题分解成 k 个正交的二值分类问题，也就是说把 Y 看作是 k 个二值标签的特定值。由此通过学习产生的预测标签集假设，将最小化汉明损失（预测集和观察结果的差距）。

AdaBoost.MR 算法：基于排序损失，其目的是确切地找到与每个实例相联系的标签集的假设，而 AdaBoost.MR 的目的是寻找一个对标签进行排序的假设，希望正确的标签能得到最高的级别。算法的结果是寻找仅有少量标签被错排的某函数 f，从而最小化预期的错排关键对，即排序损失。

5.2 文档聚类

文档聚类同文档分类相比，最主要的区别就是分类学习的样本或数据对象有类别标记，而要聚类的样本则没有标记，需要由聚类学习算法来自动确定。因此，在机器学习中聚类又称作无监督归纳（Unsupervised Induction）。

聚类是指把一组对象集合按照相似性归成若干类别。它的目的是使得属于同一类别的对象之间的相似度最大，而不同类别的对象间的相似度最小。一般情况下，对某一问题没有唯一的或是最好的解决方案。在文本挖掘中利用聚类可以进行诸如来自客户 email 邮件的主题分析。聚类的效果使文本集分割成为若干子集，子集内部具有某种特征的相关性。聚类可以按照文档内容聚类，也可以按照文档属性聚类（如：日期、长度、价钱等）。文本挖掘中的聚类能被用于：提供大规模文档集内容的总括；识别隐藏的文档间的相似度；减轻浏览相关、相似信息的过程。

聚类方法通常有如下四种[38~40]：分割聚类、层次聚类、自组织映射和平衡迭代消减聚类法（BIRCH）。

分割聚类（Partition Clusters） 通过优化一个评估函数把数据集分割成 k 个部分。分割聚类中最著名的算法是经典 K-Means 算法，它要求用户预先给定聚类的类数目 k，然后任意选取 k 个文档作为类中心开始迭代：□对每篇文档 d_i，把它分配到与其相似性最高的类 c_j 中；□类 c_j 由于 d_i 的加入使得类中心发生变化，因此重新调整 c_j 的类中心。以上两步反复进行，直到没有对象被重新分配，且满足判别函数（聚类标准）。

其中，文本间的相似性可采用距离来度量。如欧氏距离（Euclidean）、明氏距离（Minkowshi）、马氏距离（Mahalanobis）、Cambera 距离、LP 距等。

层次聚类方法（Hierarchical Clusters） 在不同层次上对数据进行分割，具有明显的层次性，算法的执行过程可以用一棵层次树（多为二叉树）来描述。

层次聚类法具体分为两种：聚合聚类（Agglomerative）和分裂聚类（Splitting）。

聚合聚类方法的主要思想是把问题空间看作 n 个对象划分成 k 个聚类的划分序列。第一个划分包括 n 个类，每个类只有一个对象；第二个划分包括 n-1 个类，……，直到最终形成一个类或者已经满足聚类要求。在此过程中一颗二叉树被构建，称为层次树[41]。它包含了全部的聚类信息：对象的内外相似度。实际层次树时，往往采用结构分割技术，各个集合各生成一颗树，最终再合并为一棵大树，因而不是严格意义上的二叉树。

分裂聚类方法的主要思想是连续地把 n 个对象构成的实体集划分成更小的簇。从层次树的产生来看，前者是从叶结点开始，逐步聚合，最终形成根结点；后者则由根依次分裂，直到叶结点。

自组织映射法（Self-organizing maps，SOMs） 使用神经元网络映射稀疏的高维空间到二维空间。相似的文档趋向于映射到二维空间中相同位置的格点。

平衡迭代消减聚类法（BIRCH） 是针对大数据库的聚类方法。它以一种较为灵活的方式递增（incremental）聚类，根据内存的配置大小而自动调整程序对内存的需要。

BIRCH 涉及两个概念：聚类特征（Clustering Feature）和聚类特征树（Clustering Feature-Tree）。聚类特征是一个三元组，它总结了一个类的有关信息。聚类特征树是一个满足节点分枝限制（每个节点的最大子女个数）和类直径限制（类中对象集的直径大小）的平衡树。树中的非叶子节点总结其子女节点的聚类特征信息。

聚类特征树动态构造，因此不要求所有数据一次读入内存。新的数据项总是插入到树中与该数据相似度最大的节点上。如果插入后使得该节点的直径大小超过了类直径限制，就对其进行分裂，直到满足类直径限制为止。新的数据项插入后，需要重新计算从此节点到根的各个祖先节点的聚类特征值。

5.3 文档总结

文档总结[15]也是文本挖掘的一个重要内容。文档总结就是指从文档中抽取关键信息，用简洁的形式来描述文档的关键（主题）内容，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，一般摘要的内容能减到原文内容的 20%。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。目前绝大多数搜索引擎采用的方法是简单截取文档前几行，显然有很大缺陷。

有多种摘要算法，但关键技术都采用词性标注，进行语义分析；用统计方法提取高频词（去除停用词之后），以确定摘要。有些算法对开始句和结束句中出现的短语给予较高的权重。还有一些方法通过寻找关键短语，确定重要的句子，例如结论句等。

总之，目前对文档总结的处理技术大多还基于经验理解，

也是文本挖掘中的难题之一。

5.4 信息提取

文本挖掘中的信息提取[42],不是简单地进行文本数据的顺序分析或是从文本中简单提取一些高频词,而是通过挖掘从文本中获得更多隐含信息,如短语间的关系、规则、典型的框架等。这些信息将体现主题、意图、期望及要求等。信息提取有很好的商业价值,对用户需求、市场预测、趋向分析等都很有帮助。

目前,信息提取主要征对如下 3 个方面:名字提取、缩写识别、关系提取。主要的技术是基于语言学的激发启发式规则[41],利用自然语言处理技术提取文本中的信息。通过建立各种词表,如同义词表、蕴含词表等解决一词多义及一义多词的语言复杂性。把文档中出现的单词分成不同的类并且度量它们对文档内容的重要性。充分利用文本中有限的结构信息,如明显的排版式样和其它语言规律性。

名字提取 一篇文档中可能混杂着人名、地名、组织名、事件名等诸多名字。通过名字提取要做的就是正确识别文本中的这些名字,把它们看作一个整体,并有一定的语义信息,如:正确识别"Washington"是人名还是地名。根据语言固有的结构模糊通过使用语言上的激发启发式规则可以得到很好的处理。

缩写识别 用于发现和匹配文本中短语和名字的其他变量。例如 EEPROM and electrical erasable PROM 被识别为 Electrical Erasable Programmable Read-Only Memory 的简写形式。

关系提取 根据类型模式,通过使用基于语言的启发式规则去识别发生的某种关系。如某人是某公司的客户等。因为是基于语言结构的关系提取,因而不能发现非邻近的关系列表。

信息提取技术对动态建立词典库亦有很大的帮助,但对文档集而言,解决效率问题是很关键的。

上述的文本分类和文本聚类是以文档集为对象的挖掘方法,而文本总结和信息提取的对象则是单文档。

6 模型质量评价

后期处理中一个重要的环节是对模型进行质量评价。

在机器学习基础上进行的数据挖掘使我们得到了隐含的、先前未知的、潜在的知识、规则和信息。但这些信息是否是有价值的或是在某种意义下满足用户目标,这就需要通过模型质量评价来作出评价。

6.1 学习器评估

用于对学习器进行评估的常用方法有两种:预留法(Hold-out)和交叉验证法(Cross- validation)。它们都假定待预测数据和训练集具有相同的分布。

预留法 把数据集分成训练集和测试集两部分。学习器使用训练集数据来构造分类器,然后使用这个分类器对测试集进行分类,得出的错误率就是评估错误率。

这种方法的优点是速度快,但由于仅使用训练集的数据来构造分类器,因此它没有充分利用所有的数据来进行学习。如果使用所有的数据,那么可能构造出更精确的分类器。

交叉验证法 把数据集被分成 k 个互不相交的数据子集,大小大致相同。学习器进行 k 次训练和测试;每一次,学习器使用去除一个子集的剩余数据作为训练集,然后用被去除的子集作为测试集进行测试。取 k 次错误率的平均值作为评估错误率。这种评估方法准确度的提高是以运行时间为代价的。

通常预留评估法被用在最初试验性的场合,或者多于 5000 条记录的数据集;交叉验证法被用于建立最终的分类器,或者很小的数据集。

6.2 评价指标

常用的评价指标有[24]:分类正确率、查准率与查全率[20]、收益、支持度、置信度等来衡量所发现知识的有效性,可用性和可理解性。下面简要描述其定义。

分类正确率

$$Accuracy(M)=\sum_{ex}P(ex)Accuracy(M,ex)=P(\hat{C}(ex)=C(ex))$$

$$Accuracy(M,ex)=\begin{cases}1 & \hat{C}(ex)=C(ex)\\ 0 & \text{otherwise}\end{cases}$$

其中 $C(ex)$ 为样本 ex 的实际类值,$\hat{C}(ex)$ 为通过模型 M 对样本 ex 的预测类值,$P(ex)$ 为样本 ex 的概率。

查准率与查全率 查准率是指正确分类的对象所占对象集的大小。对目标类 C,模型 M 的精度用如下公式估计:

$precision(M,C)=P(C|\hat{C})$。

查全率是指集合中所含指定类的对象数占实际目标类中对象数的比例。查全率用如下公式估计:

$recall(M,C)=P(\hat{C}|C)$。

收益 收益定义为:

$$M\ arg\ in=(y\sum_i a_i h_i(x))/\sum_i a_i$$

其中,h_i 是一个弱假设 $h_i:X\to Y$;y 是标记;a_i 是 h_i 的权重。

支持度与置信度 支持度定义为:

$\delta(A)=D$ 中包含 A 的事务数量/D 的总事务量

其中 A 是数据项集,D 是全体事务集。

规则 $A\cup B$ 的置信度定义为:

$Confidence(A\cup B)=\delta(A\cup B)/\delta(A)$

表示 S 中包含 A 事务同时也包含 B 的可能性。

置信度表示规则的强度,支持度表示规则的频度。

结束语 文本挖掘是一个崭新的人工智能研究方向,本文根据所阅读的大量现有有关论文对文本挖掘技术作了详细的综述,还有许多技术是有待于进一步研究和改进的,比如文本特征表示、特征提取、文本摘要等。此外,本文对各个技术的介绍是孤立的,但在实际应用中各种方法之间是相互关联的。比如:如果确定了提取模式为文本分类而且选用多变量贝努利模型进行分类,则文本的特征表示方法就已经确定,即用布尔模型表示文本。对模型质量的评价可以选用错误率来评估。所以针对具体问题,应该综合考虑选用的方法和技术。

最后提供一些含有相关信息的 WWW 站点,以供参考:http://www.kdnuggets.com/;http://www.sinokdd.163.net/;http://www.almaden.ibm.com/cs/people/ragrawal/pubs.html#txt;http://www.cs.cmu.edu/afs/cs/project/theo-4/text-learning/www/;http://www-ai.ijs.si/DunjaMladenic/papers/pww/。

参考文献

1 Fayyad U M,Piatetsky-Shapiro G,Smyth P. Adavance in Knowledge Discovery and Data Mining. Cambridge MA: AAAI/MIT Press,1996

2 John George H. Enhancements to the data mining process:[Ph. D. Thesis]. Stanford University, 1997

3 Rao A S. AgentSpeak(L):BDI Agents Speak Out in a Logical Computable Language. In:Proc. Eur. Workshop Model. Auto. Agents Multi-Agent World(MAAMAW-96,7th),1996. 42～55

4 付国宏,王晓龙. 基于词形的汉语文本切分方法. 情报学报

(JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATION),1999,18(3)
5 梁南元,郑延斌. 一个汉语自动分词模型 CWSM 及自动分词系统 PC-CWSS. Communications of COLIPS,1991,1(1):51～55
6 Wang XiaoLong,et al. The Problem of Separating Characters into Fewest Words and Its Algorithms. Chinese Science Bulletin, 1989,34(22):1924～1928
7 尹锋. 汉语自动分词研究的现状与新思维. 现代图书情报技术, 1998(4)
8 Salton G,Wong A,Yang C S. A Vector Space Model for Automatic Indexing. Communication of the ACM 1995,18:613～620
9 Mladenic D. Machine Learning on non-homogeneous, distributed text data. Doctoral Dissertation, University of Ljubljana,1998
10 王继成,潘金贵,张福炎. Web 挖掘技术研究. 南京大学多媒体计算机研究所,1999
11 McCallum A,Nigam K. A Comparison of Event Models for Naive Bayes Text Classification. Just Research 4616 Henry Street Pittsburgh,PA 15213
12 McCallum A,Nigam K. Text Classification by Bootstrapping with Keywords, EM and Shrinkage. Just Research 4616 Henry Street Pittsburgh,PA 15213
13 The International Journal of Artificial Intelligence. Neural Networks, and Complex Problem-Solving Technologies. http://textmining.krdl.org.sg/APIN/TWMcfp.html,2001
14 First SIAM International Conference on Data Mining. http://www.cs.utk.edu/tmw01,2001
15 Sullivan D. The Need for Text Mining in Business Intelligence. Published in DM Review in Dec. 2000
16 http://www.cs.waikato.ac.nz/～nzdl/textmining,Sep. 2000
17 IMA "HOT TOPICS" Workshop,Apr. 2000
18 "What is Text Mining?". http://www.ikdi.com.au/text-mining.html,2000
19 Helena Ahonen Mika Klemettinen,Oskari Heinonen A. Inkeri Verkamo. Applying Data Mining Techniques in Text Analysis. University of Iielsinki, Department of Computer Science P.O. Box 26, FIN-00014 University of Iielsinki,Finland
20 Lewis D D. Evaluating and Optimizating Autonomous Text Classification Systems. In:Proceedings of the 7th Annual International ACM-SIGIR Conf. on Research can Development in Information Retrieval,Dubline
21 Joachims T. A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization. ICML97,1997. 143～151
22 Mladenic D. Personal Web Watcher : Implementation and Design:[Technical Report, US-DP-7472]. Oct. 1996
23 邹涛,黄源,张福炎. 基于 WWW 的文本信息挖掘. 情报学报, 1999,18(4)
24 王伟强,高文,段立鹤. Internet 上的文本数据挖掘. 中国科学院计算技术研究所
25 邹涛,王继成,黄源,张福炎. 中文文档自动分类系统的设计与实现. 中文信息学报,1999,13(3)
26 朱廷劭,高文,Ling C X. 数据库中知识发现的处理过程模型的研究. 计算机科学,1999,26(2)
27 王玲玲, 周纪芗. 常用统计方法. 华东师范大学出版社,1994
28 项静恬,史久恩. 非线性系统中数据处理的统计方法. 科学出版社,1997
29 肖云茹. 概率统计计算方法. 南开大学出版社,1994
30 Quinlan J R. Constructing Decision Tree in C4. 5. Morgan Kaufman Publishers,Programs for Machine Learning,1993. 17～26
31 Koller D,Sahami M. Hierarchically classifying documents using very few words. In: Proc. of the 14th Intl. Conf. on Machine Learning ICML97,1997. 170～178
32 Yang Y,Pedersen J O. A Comparative Study on Feature Selection in Text Categorization. In: Proc. of the 14th Intl. Conf. on Machine Learning ICML97,1997. 412～420
33 Kononenko I. On biases estimating multi-valued attributes. In: Proc. of the 14th Intl. Joint Conf. on Artificial Intelligence IJCAI-95,1995. 1034～1040
34 Rijsbergen V,et al. The selection of good search terms. Information Processing and Management 17th,1981. 77～91
35 Mladenic D. Feature subset selection in text-learning. Department of Intelligent Systems, J. Stefan Institute. http://www-ai.ijs.si/DunjaMladenic
36 Schapire R E,Singer Y. BoostTexer: A system for multiclass multi-label text categorization. Machine Learning,1998
37 Schapire R E,Singer Y,Singhal A. Boosting and Rocchio Applied to Text Filtering. SIGIR'1998
38 Spath H. Cluster Dissection and Analysis: theory, fortran program examples. Ellis Horwood Limited ,Chichester,1980
39 Everitt B S. Cluster Analysis. Copublished in the Americas by Halsted Press,1993
40 边肇祺,张学工. 模式识别. 清华大学出版社,2000
41 Tkach D. Turning Information Into Knowledge A White Paper from IBM. IBM Software Solutions,Feb. 1998
42 http://www-4.ibm.com/software/data/iminer/fortext/