

# 基于VSM的文本分类挖掘算法综述

夏火松 刘 建

(武汉科技学院经济管理学院 湖北 430073)

**摘 要:**简要介绍了VSM和文本分类挖掘的流程,分析了基于统计方法和基于机器学习的6种常用构造文本分类挖掘分类器的算法,指出了利用各种算法构造的分类器的特点,同时给出了这些算法的优化方向,为使用者选择、学习、改进算法提供依据。

**关键词:**VSM 文本分类 分类算法 分类器

**中图分类号:**G350

**文献标识码:**A

**文章编号:**1005-8095(2010)09-0018-04

## 1 引言

目前大多数信息是以非结构化的文本形式来保存的,给信息的查询和检索带来了极大的麻烦。在过去的十年中,技术的进步已经使这一领域迅速取得进展,文本挖掘已经是信息检索、数据挖掘、机器学习、统计以及计算语言学等学科中的重要领域。典型的文本挖掘方法包括文本分类、文本聚类、概念/实体挖掘、生产精确分类、观点分析、文档摘要和实体关系模型(即学习已命名实体之间的关系)。

文本分类就是在给定的分类模型下,由计算机根据文本内容自动判别文本类别的过程<sup>[1]</sup>。随着文本分类技术的发展,不同的文本表示模型逐渐出现多种文本分类算法,使得文本挖掘领域道路越来越宽。目前已经出现多种中文文本表示方法,如布尔模型、向量空间模型、潜在语义模型和概率模型等。所以在构造自动文本分类器时,面临的选择也越来越多,如何有效的基于某种模型选择一种文本分类算法来构造分类器已经成为一个重要的课题。空间向量模型是一种出现较早的文本表示模型,但现在仍然在广泛的使用。本文的重点是对已经出现的基于向量空间模型的文本分类算法进行研究分析。

## 2 文本的表示

中文文本信息多数是无结构化的,并且使用自然语言,很难被计算机处理。因此,如何准确地表示中文文本是影响文本分类性能的主要因素。经过多年发展(图1),研究人员提出了布尔模型<sup>[2]</sup>、向量空间模型<sup>[3]</sup>、潜在语义模型和概率模型等文本表示模型<sup>[4]</sup>(见表1),用某种特定结构去表达文本的语义。

基于分类速度的考虑,目前的文本分类挖掘系统主要采用向量空间模型来表示文本。

VSM<sup>[5]</sup>由哈佛大学的G Salton提出,这一模型将给定的文本转换成一个维数很高的向量,并以特征项作为文本表示的基本单位,向量的各维对应文

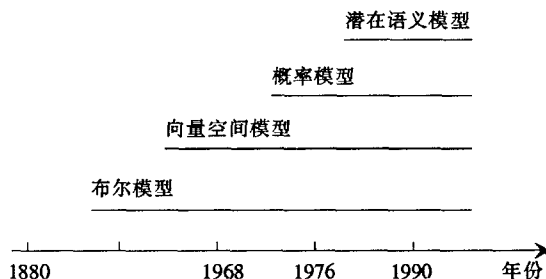


图1 文本表示模型的发展

本中的一个特征项,而每一维本身则表示了其对应的特征项在该文本中的权值。权值代表了特征项对于所在文本的重要程度,即该特征项能够多大程度上反映它所在文档的类别。

对于特征项权值的计算方法有:tf算法、idf算法、tf\*idf算法<sup>[6]</sup>和mutual information算法等。其中tf\*idf在文本检索和机器学习中被频繁使用,目前存在多种tf\*idf公式,比较普遍的tf\*idf公式如下:

$$w(t,d) = \frac{tf(t,d) \times \log_2 \left( \frac{N}{N_t} + 0.01 \right)}{\sqrt{\sum_{t \in d} [tf(t,d) \times \log_2 \left( \frac{N}{N_t} + 0.01 \right)]^2}} \quad (1)$$

## 3 文本分类流程

文本分类系统是在给定的分类体系下,根据文本的内容自动地确定所属类别。从数学上看,文本分类是一个映射的过程,它将未标明类别的文本映射到已知类别中,这种类别可以是一对一也可以是一对多。用数学语言来描述,文本分类可表示为:

对于每个文档与类别的二元组 $(d_i, c_i) \in D \times C$ ,判断其值,如果值为1,则表示文档 $d_i$ 属于类别 $c_i$ ;如果值为0,则表示文档 $d_i$ 不属于类别 $c_i$ (其中, $d_i$ 属于文档集合 $D$ 中的一个文档, $C = \{c_1, c_2, c_3, \dots, c_n\}$ 是预先定义类别集)。用数学公式表达:

$$F: D \times C \rightarrow \{1, 0\} \quad (2)$$

收稿日期:2010-03-01

作者简介:夏火松(1964—),男,博士,教授,武汉科技学院院长,研究方向为知识管理、数据挖掘;刘建(1985—),男,2009级硕士研究生,研究方向为数据挖掘。

表1 文本表示的典型模型

类别	特征	优点	局限性
布尔模型 (Boolean Model)	第一个被提出的模型。特征项在文本中只有两种状态:出现或者不出现,相应地每个特征项的权值为1或者0	模型简单、容易理解	表示能力非常刚性,不能反映出特征项对文本语义的重要程度
向量空间模型 (Vector Space Model)	基于这种模型每篇文档都形式化为高维向量空间中的一个向量,向量中的每个分量对应文档的一个特征词条的权重。词的权重一般采用TF*IDF的计算方法	把对文本的处理转化成向量空间中的向量运算,使得问题的复杂度大为降低,提高了文本处理的速度	该模型假设文本向量中的特征词是相互独立的,这一假设在自然语言文本中是不成立的
潜在语义索引模型 (Latent Semantic Indexing Model)	可以看作是向量空间模型的一种改进,其基本思想是利用文本中词与词之间存在的某种潜在的语义结构来表示词和文本之间的内在关系	克服因词语多义性和同义性带来的问题。该模型在尽量保证特征信息量一致的基础上,向量空间大大缩减,简化计算的复杂性	增加了文本预处理的复杂度,需时较多
概率模型 (Probabilistic Model)	该模型综合考虑了词频、文档频率和文档长度等因素,把文档和用户兴趣(查询)按照一定的概率关系融合,形成了著名的OKAPI公式	概率模型更为准确地描述关键词和文档之间的相关关系	由于需要事先确定关键词和文档之间相关概率,限制了其广泛使用,但对检索系统的理论研究提供了依据

这个函数又称为分类器。基本流程见图2。

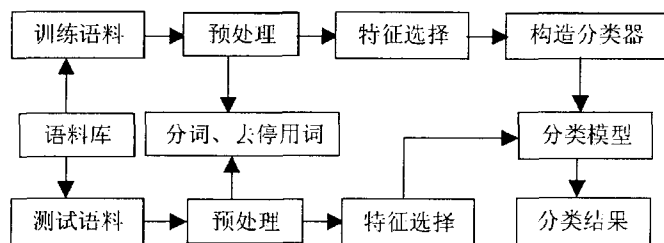


图2 文本分类挖掘一般流程

以下对文本分类流程中的重要步骤进行说明:

(1)语料库:用于分类的文本的集合,它是能够代表同类某一领域的语言现象的真实语言材料的集合。语料库被分为训练语料和测试语料。

(2)预处理:预处理通常包括如下环节:去除标记;去停用词、进行数字合并、词根还原;分词<sup>[7]</sup>、词性标注、短语识别;词频统计;进行数据清洗,除去不适合的噪音文档或文档中的垃圾数据。分词后的特征项有冗余,需要建立合适的特征评价函数,对特征项进行选择,达到降低计算复杂度和提高分类准确率的目的,并为以后的分类器设计提供参数。

(3)特征选择:常用的选择算法有文档频率、信息增益、互信息、 $\chi^2$ 统计量和期望交叉熵等,其中以信息增益和 $\chi^2$ 统计量的效果为最好<sup>[8]</sup>。特征空间的维度确定问题非常重要,但是目前对此问题尚无理想的方法,基本上是根据实验效果而定。

(4)构造分类器:一个优秀的分类器是在分类的准确度和效率上都是优秀的,而分类算法的选择是决定分类器好坏的关键<sup>[9]</sup>。

#### 4 基于VSM的文本分类挖掘算法综述

中文文本分类的一个关键问题是分类器的构建。文本分类通过分类器将待分类文本划分到相应的类别空间中,而分类器是通过文本分类算法实现的。文本分类算法就是根据一个文档的特征向量,计算该文档的类别,是设计实现分类器的理论基础。文

本分类算法实现形式和算法的性能各异。

目前基于VSM的常用文本分类的方法有以下几种:Rocchio算法、贝叶斯算法、K-近邻算法、判定树算法、支持向量机算法、神经网络算法等。

##### 4.1 Rocchio 算法

根据待分类文本向量与每个类别中心向量的相似度来判断文本的类别。相似度是以待分类文本向量与中心类别向量之间的距离来计算的。距离越近它们就越相似,最后将待分类文本分到最大相似度所对应的文本类别中去。中心向量代表某一类别的文本向量,通过每个类别中所有的训练文本向量的算术平均得到。待分类文本的特征向量与每一类别文本中心向量间的相似度计算公式如下:

$$SIM(d_i, d_j) = \frac{\sum_{k=1}^n w_{ki} \times w_{kj}}{\sqrt{(\sum_{k=1}^n w_{ki}^2)(\sum_{k=1}^n w_{kj}^2)}} \quad (3)$$

$w_{ki}$  和  $w_{kj}$  分别表示文档  $d_i$  和  $d_j$  的第  $k$  个特征项的权值,  $n$  为文本的特征项数。

Rocchio 算法<sup>[10]</sup>思想比较简单,分类复杂度不高。其缺点是比较明显的:受文本中的噪音的影响较大,对非线性可分的情况分类能力较差。所以 Rocchio 算法在现在的文本分类系统中使用较少。

##### 4.2 贝叶斯算法

贝叶斯算法<sup>[11]</sup>(Bayes 算法)是一种统计学分类算法,可以预测类成员关系的可能性。它利用贝叶斯公式通过类别的先验概率和词的分布来计算未知文本属于某一类别的概率,最终将新样本分配到概率最大的类别中去。贝叶斯分类模型将训练文档分解成特征向量和类别决策变量,假设特征向量各分量间相对于类别决策变量是相对独立的,即各分量独立的作用于类别决策变量。

贝叶斯文本分类器<sup>[12]</sup>基于一个“独立性假定”:给定一个文本的类标签,文本中每个属性的出现独立于文本中其他属性的出现。理论上讲,贝叶斯分类

具有最小的出错率。然而,实践中并非总是如此。这是由于对其应用的假定(如类条件独立性)的不正确性,以及缺乏可用的概率数据造成的。该方法有以下不足:有个很强的独立性假设,而这个假设一般是不符合实际情况的,从而缩小了其使用范围。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算法<sup>[13]</sup>。

文本数据的高维性和稀疏性导致该方法存在分类精度不高且模型规模较大等问题。针对以上问题,出现大量的基于贝叶斯算法的改进模型<sup>[14-16]</sup>。

#### 4.3 K-近邻算法

K-近邻算法<sup>[17]</sup>(K-Nearest Neighbours, KNN算法)是模式识别中一种重要的非参数法。其基本思想为:对于一个测试文本,计算它与训练样本集中每个文本的相似度,找出K个最相似的文本,根据加权距离和来判断测试文本所属的类别。K值的确定一般先采用一个初始值,然后根据实验测试的结果调整K值,一般初值定为几百到几千。

KNN算法在20世纪60年代就已经成为了非常重要的分类方法<sup>[18]</sup>。KNN算法的优点:该方法简单、有效,重新训练的代价较低,可以利用增量式更新的方式,非常方便地持续更新训练集;计算复杂度不高<sup>[19]</sup>。此外,如果训练集足够大,它还能对噪音数据表现出非常好的鲁棒性<sup>[20]</sup>。同时这种算法的缺点也是非常明显的<sup>[21]</sup>:KNN算法是懒散的分类算法,所有的计算都推迟到了分类时,所以训练时快,但分类时慢,并且其分类时间是非线性的,当训练样本数增加时,其分类时间将急剧增加;计算相似度时,特征向量维数高,没有考虑特征词间的关联关系;样本距离计算时,各维权值相同,使得特征向量之间的距离计算不够准确,影响分类精度。针对以上的缺点,很多学者已从多个角度改善KNN算法<sup>[22-25]</sup>。

#### 4.4 判定树算法

判定树算法<sup>[24]</sup>(Decision Tree, DT算法)是类似于流程图的树结构;每个内部结点表示在一个属性上的测试,每个分枝代表一个测试输出,每个树叶结点代表类或类分布。树的最顶层结点是根结点。

判定树归纳的基本算法是贪心算法,是一种比较成熟的算法,其具有强的伸缩性。判定树系统不使用领域知识,判定树归纳的学习和分类步骤通常很快。但其缺点也是比较明显的:特征间的相关性强调不够,联系还是松散的;对噪声较为敏感;当训练集增加时,决策树也会随之变化。在建树过程中各特征的互信息会随例子的增加而改变,从而使决策树也变化,这对渐进学习不够方便<sup>[1,25]</sup>。

判定树分类算法有ID3、C4.5和C5.0等<sup>[26]</sup>。它们的基本思想是相同的。不同在于选择属性时所用的度量不同。ID3算法用的是信息增益作为度量。C4.

5算法采用的是增益比率,C5.0是C4.5的商业版本。为了适应处理大规模数据集的需要,后来又提出了若干改进的算法,其中SLIQ<sup>[27]</sup>和SPRINT<sup>[28]</sup>是比较有代表性的两个算法。

#### 4.5 支持向量机算法

支持向量机(Support Vector Machine, SVM)是Vapnik<sup>[29]</sup>等人在统计学习理论的VC维及结构风险最小化原理的基础上提出的一种机器学习新方法。它能够有效地解决高维、非线性以及有限样本下的模式识别问题。SVM算法的主要思想是在高维空间中寻找一个超平面作为两类样本的分割,以保证最小的分类错误率。它通过分线性转换,将输入向量映射到高维空间H,并在H中构造最优分类超平面,从而达到最好的泛化能力。

SVM在文本挖掘、人脸识别、基因检测和手写数字识别等领域得到了广泛的应用。在文本分类挖掘中,SVM也存在着一些缺点:结构风险最小原理并不能严格证明好的推广能力;学习机的VC维的分析尚没有通用方法<sup>[30]</sup>。

#### 4.6 神经网络算法

神经网络算法(Neural Networks, NN)旨在寻求开发和测试神经的计算模拟<sup>[31]</sup>。神经网络是一组连接的输入/输出单元,其中每个连接都与一个权相相联。通过调整权,使得能够预测输入样本的正确类标号来学习。

BP神经网络是应用最为广泛的一种网络模型,具有逼近任意连续函数和非线性映射的能力。BP神经网络属于前馈型神经网络,具有很强的自学习性、自组织性、容错性、高度非线性、高的鲁棒性、联想记忆功能和推理意识功能等。但它仍然存在易陷入局部极小、收敛速度慢、全局搜索能力弱、网络结构没有成型的理论指导和可解释性差等缺点。针对神经网络的缺点,很多的学者提出混合算法将其它的算法和神经网络结合起来达到较好的效果<sup>[32-33]</sup>。

但神经网络的优点也是明显的,其对噪音数据的高承受能力,以及高度非线性的分类能力受到大多数学者的青睐。此外,最近已提出了一些由训练过的神经网络提取规则的算法,如[SN88, Gal93, TS93, Fu94, Avn95, LSL95, CS96b, LGT97]<sup>[24]</sup>。

此外还有一些其他分类算法,比如基于案例的推理、遗传算法、粗糙集和模糊集方法、基于中心点的分类方法及线性最小二乘(LLSF)等。

#### 5 总结

本文对目前比较优秀的各种分类算法进行了介绍、分析和比较。文本挖掘是一个非常活跃的研究领域,尤其以文本分类应用十分广泛。采用了自动分类技术,可将不同的内容分到不同的类目中去,从而大大优化搜索引擎性能,节省用户的判断时间,提高检

索效率。文本分类算法是自动文本分类系统的关键组成部分之一,由于文本所具有的多意性、模糊性等特点使得自动文本分类在许多方面的表现难以令人满意,而选择不同的算法对分类的准确性和效率产生很大的影响,所以在自动文本分类系统中应按照对待分类文本的要求和算法的特点来选择恰当的算法。同时,根据现实的要求,对各种算法进行优化是必须的,以使分类的结果满足我们的需求。

同时,从文献中可以看出文本分类算法朝着混合算法的趋势发展,充分利用了各种算法的优点,扬长避短。例如基于遗传算法的文本分类方法等<sup>[1]</sup>,在构造分类器时KNN、SVM等算法被频繁地使用,而SVM逐渐地成为文本分类挖掘的主流算法。

#### 参考文献

- [1] 戴文华. 基于遗传算法的文本分类及聚类研究[M]. 北京:科学出版社,2008
- [2] Zorkadis V, Karras D A, Panayotou M. Efficient information theoretic strategies for classifier combination: feature extraction and performance evaluation in improving false positives and false negatives for spam e-mail filtering[J]. Neural Networks, 2005(18): 799-807
- [3] Nigam K, McCallum A K, Thrun S, et al. Text Classification from Labeled and Unlabeled Documents Using EM[J]. Machine Learning, 2000(39): 103-134
- [4] 苏新宁. 信息检索理论与技术[M]. 北京: 科学技术文献出版社, 2004
- [5] Salton G, Wang A, Yang C S. A vector space model for automatic indexing[J]. Communication of the ACM, 1975, 18(11): 613-620
- [6] Salton G, Buckley B. Term-Weighting approaches in automatic text retrieval[J]. Information Processing and Management, 1998, 24(5): 513-523
- [7] Kim H, Howland P, Park H. Dimension Reduction in Text Classification with Support Vector Machines[J]. The Journal of Machine Learning Research, 2005, 6(1): 37-53
- [8] QU Jun, LIN Xu. Comparison and Analysis of Feature Extraction Methods for Text Categorization[J]. Modern Computer, 2007(4): 10
- [9] Qu K S, Liang J Y, Wang J H, et al. The algebraic properties of concept lattice[J]. Journal of Systems Science and Information, Research Information Ltd UK, 2004, 2(2): 271-277
- [10] TANG Pei li, WANG Shu ming, HU Ming. Algorithm of Thematic Words Extraction from Chinese Texts Based on Semantic[J]. Journal of Jilin University: Information Science Edition, 2005, 23(5): 535-540
- [11] 曾砺锋. 基于 Rocchio 方法和 k 均值聚类的支持向量机文本分类方法[J]. 软件导刊, 2008, 7(6): 37-39
- [12] McCallum A, Nigam K. A Comparison of Event Models for Naive Bayes Text Classification[J]. AAAI-98 Workshop on Learning for Text Categorization, Madison, Wisconsin: AAAI Press, 1998
- [13] Baker L D, McCallum A. Distributional Clustering of Words for Text Classification[R]. Proceedings of the 21th Annual International ACM SIGIR, 1998: 96-103
- [14] Mitchell T M. Machine learning[M]. Beijing: China Machine Press, 2003: 165-171
- [15] Witten I H, Frank E. Data Mining: Practical Machine Learning Tools and Techniques[M]. 2nd ed. Beijing: China Machine Press, 2006
- [16] 杨延娇, 王治和. 基于树桩网络的贝叶斯文本分类算法[J]. 计算机工程, 2009, 35(16): 201-205
- [17] 白莉媛, 黄晖. 基于自助平均的朴素贝叶斯文本分类器[J]. 计算机工程, 2007, 33(12): 190-192
- [18] 高影繁, 马润波, 刘玉树. 文本分类中影响因素的定量分析[J]. 计算机工程, 2008, 34(9): 222-224
- [19] 王潇. 基于向量空间模型的文本自动分类算法的研究与改进[D]. 兰州: 西北师范大学, 2006
- [20] Yang Y, Pedersen J. A comparative study on feature selection in text categorization[M]. San Francisco: Morgan Kaufmann Publishers, 1997
- [21] Ganter B, Wille R. Formal Concept Analysis: Mathematical Foundations[M]. Berlin: Springer-Verlag, 1999
- [22] 闫鹏, 郑雪峰, 等. 一种优化的 k-NN 文本分类算法[J]. 计算机科学, 2009, 36(10): 217-221
- [23] Delany S J, Cunningham P, Tsybalb A, et al. A case-based technique for tracking concept drift in spam filtering[J]. Knowledge-based Systems, 2005(18): 187-195
- [24] Han Jia wei, Micheline Kamber. 数据挖掘概念与技术[M]. 北京: 机械工业出版社, 2001: 3-6
- [25] Quinlan J R. C4. 5: Programs for Machine learning[J]. Morgan Kaufman, 1993, 2(3): 89-94
- [26] 张桂杰, 王帅. 决策树分类 ID3 算法研究[J]. 吉林师范大学学报: 自然科学版, 2008, 3(11): 135-137
- [27] Mehta M, Agrawal R, Rissanen J. SLIQ: A Fast Scalable Classifier for Data Mining [A]. Lecture Notes in Computer Sci. Proc. of the 5th Int. Conf. on Extending Database Tech[C]. 1996: 18-33
- [28] Vapnik V. 统计学习理论的本质[M]. 张学工, 译. 北京: 清华大学出版社, 2000
- [29] 张学工. 关于统计学习理论与支持向量机[J]. 自动化学报, 2000, 26(1): 32-42
- [30] Simon Haykin. Neural Networks: A comprehensive Foundation, Second Edition[M]. 叶世伟, 译. 北京: 机械工业出版社, 2004
- [31] Hecht Nielsen R. Theory of the Back Propagation Neural Network [J]. Proceeding of IJCNN, 1989, 1(1): 593-603
- [32] 陈世立, 高野军. 基于神经网络与贝叶斯的混合文本分类研究[J]. 情报杂志, 2007(5): 34-36
- [33] Han E H, Karypis G. Centroid-based Document Classification: Analysis and Experimental Results[A]. University of Minnesota, 2000