申请上海交通大硕士学位论文

中文多标签文本分类算法研究

硕士研究生：周　浩

学　　号：1110369059

导　　师：宦飞副教授
　　刘功申副教授

申请学位：工学硕士

学　　科：计算机技术

所在单位：信息安全工程学院

答辩日期：2013 年 12 月

授予学位单位：上海交通大学

万方数据

Dissertation Submitted to Shanghai Jiao Tong University
for the Degree of Master

Arithmetic Research for Multi-label Chinese Text Classification

Candidate：	Hao Zhou
Student ID:	1110369059
Supervisor：	Associate Prof. Huan fei Associate Prof. Liu GongShen
Academic Degree Applied for：	Master of Engineering
Speciality：	Computer Technology
Affiliation：	School of Information Security Engineering
Date of Defence：	Dec, 2013
Degree-Conferring-Institution：	Shanghai Jiao Tong University

万方数据

中文多标签分类算法研究

摘 要

伴随着海量互联网信息的爆炸式扩展，文本分类已经逐渐成为数据挖掘研究领域内的热点核心技术之一，分类算法的性能评估方法也渐渐成为研究热点。信息通过多标签文本分类可以准确快速地定位到其相关的话题与类别。目前，对于多标签分类的研究主要针对于多标签分类前的特征选择以及分类算法的研究与发展。本文首先提出文本的特征选择方法相比于现有最常用的特征选择算法，更注重分类区分度高的强特征纹理，删除稀疏特征、保留有利于分类的特征维度。而后基于此特征选择的结果，提出的是一种相关信息加权的自适应多标签分类算法, 该算法具有相关信息加权、自适应阈值调整、权重投票相结合的特点。

对中文多标签文本特征选择算法，本文寻找一种有效的特征选择方法, 降低特征空间维数, 提高分类精度和效率。鉴于样本特征在不同类别文档中出现的频率影响，分布得越离散对类别判定越重要，基于这一特点可以来考查特征选择空间在分类结果的影响程度。分布离散度往往通过标准差或协方差计算，本文使用的基于强特征选择以及在文档类中的改进的特征选择权重计算方法，将此特征化概率标准差作为基本权重。实验结果表明该算法在性能上的指标优于现有一些常用的多标签文本分类特征选择方法，综合了多种特征评估函数选择特征子集，不受具体文本语料的限制，降低了“噪音”影响。

对中文多标签文本分类算法，本文结合了问题转化和多标签算法改进的思想，提出的是一种在各类特征选择基准调整后，基于已有单标签分类结果进行加权、自适应阈值设定，不同权重投票相结合的方法，对待分类实例进行多标签分类，能提高多标签文本分类的分类准

万方数据

确度与精度。实验结果表明，本文算法提供了一种更为有效，分类可靠性更高的多标签分类算法，在某些性能指标优于现有一些常用的多标签分类方法。

关键词： 多标签分类，特征选择，自适应回归，相关信息加权投票，强类别纹理

万方数据

ARITHMETIC RESEARCH FOR MULTI-LABEL CHINESE TEXT CLASSIFICATIONS

ABSTRACT

With the explosion of massive Internet information,text classification technology has gradulally become an emerging core technology in data mining. Therefore, the study of classification algorithm performance evaluation has also become an important issue. Information through the label text classification can be quickly and accurately positioning to its related topics and categories.At present,this field mainly focuses on the former feature selection and the research and development of the algorithm. This paper pays more attention to the classification characteristics of high degree of differentiation of texture,deletes the sparse feature,and reserves for classification feature dimension.Then, based on the results of the feature selection,put forwards a Adaptive Algorithm for Multi-Label Classification Based on Related Information Weighting,which featured as the Single-Label classification result weighting, adaptive threshold adjustment,related information noting.

In the Chinese label text feature selection algorithms, this paper goals to find an effective feature selection method, which reduces the dimension of feature space, and improves the classification accuracy and efficiency.Due to the characteristics of the frequency of the uneven in the document class, namely the discrete characteristics of distribution, tend to judge the more important characteristics of category, make use of this

万方数据

nature can be important degree of examination features in classification.Discrete degree usually can be used to calculate the standard deviation or variance, this article USES the characteristics of probability standard deviation in the document class to quantitative description of characteristics of importance, this features probability standard deviation will be used as the basic weight in text categorization.Experimental results show that the algorithm on the performance indexes of some commonly used tabbed text categorization is superior to the existing feature selection methods, a combination of a variety of feature evaluation function to choose feature subset, not limited by the specific text corpus, to reduce the effect of "noise".

For multi-label in Chinese text classification algorithm, this paper combines the problem and improvement of tabbed algorithm, proposed is a kind of feature selection benchmark adjusted, based on the existing single tag classification results are weighted, adaptive threshold setting, the combination of different weighted voting method, classification instance treated with more tags, can improve the classification accuracy and precision of the tabbed text classification. Experimental results show that the algorithm provides a more effective and classification reliability higher multiple tags classification algorithm, in some performance index is superior to the existing tabbed classification of some commonly used methods.

Keywords: Multi-Label, Feature Selection, Adaptive Regression , Related Information Weighted Noting , Strong Category Texture

万方数据

目 录

第一章 绪论 ……1
1.1 研究背景和意义 ……1
1.2 多标签分类 ……2
1.3 国内外研究现状 ……3
1.4 论文的结构安排 ……6
第二章 多标签文本分类相关技术 ……7
2.1 文本分类的定义及过程 ……7
2.1.1 文本分类的定义 ……7
2.1.2 文本分类的一般过程 ……7
2.1.3 多标签文本分类评估方法 ……8
2.2 文本分类的分类方法 ……9
2.2.1 决策树算法 ……9
2.2.2 Ricchio 算法 ……10
2.2.3 KNN 算法 ……11
2.2.4 神经网络算法 ……12
2.2.5 朴素贝叶斯算法 ……13
2.2.6 支持向量机 ……14
2.3 本章小结 ……14
第三章 基于中文多标签分类的特征选择 ……16
3.1 文本特征选择 ……16
3.2 特征选择方法 ……17
3.2.1 过滤无意义信息 ……17
3.2.2 汉语文本自动分词 ……17
3.2.3 汉语文本粗降维 ……18
3.2.4 文本表示模型 ……18
3.2.5 常用特征选择方法 ……19
3.3 改进的特征选择方法 ……20
3.3.1 强类别纹理挖掘算法 ……20
3.3.2 常用权重计算方法 ……21

万方数据

3.3.3 改进的特征选择和加权抽取 ……22
3.4 多标签分类特征选择算法的框架 ……24
3.5 本章小结 ……25
第四章 相关信息加权的自适应多标签分类算法 ……26
4.1 常用多标签分类算法 ……26
4.1.1 Navie-Bayes 算法 ……26
4.1.2 ML-Knn 算法 ……27
4.1.3 RAkEL 算法 ……27
4.2 信息加权模型算法 ……28
4.3 WeightedLabelPower 投票预测 ……29
4.4 多标签分类算法的框架 ……29
4.5 本章小结 ……31
第五章 实验及结果分析 ……32
5.1 多标签文本分类数据集 ……32
5.2 多标签文本分类特征选择实验 ……33
5.2.1 强特征挖掘实验 ……33
5.2.2 改进的特征选择和加权抽取实验 ……34
5.3 相关信息加权的自适应多标签分类实验 ……37
5.3.1 实验环境 ……37
5.3.2 实验数据 ……37
5.3.3 结果分析 ……41
5.4 本章小结 ……45
第六章 总结与展望 ……46
6.1 本文工作总结 ……46
6.2 研究展望 ……46
参 考 文 献 ……48
致 谢 ……53
攻读硕士期间发表的论文以及专利 ……54

万方数据

图 录

图 2-1 文本分类一般过程……8
图 2-2 决策树算法表示图……10
图 2-3 KNN 算法表示图……12
图 2-4 神经网络算法表示图……12
图 2-5 支持向量机模型示意图……14
图 3-1 系统整体模型图……16
图 3-2 文档的向量空间模型示意图……18
图 5-1 强特征挖掘流程图……34
图 5-2 用于类别描述的特征选择方法流程图……35
图 5-3 多标签文本分类系统流程图……38
图 5-4 分词器初始化……40
图 5-5 强特征和多标签结合一……40
图 5-6 强特征和多标签结合二……40
图 5-7 强特征和多标签结合三……41
图 5-8 EMOTIONS 数据集评价指标图表趋势……43
图 5-9 SCENE 数据集评价指标图表趋势……43
图 5-10 YEAST 数据集评价指标图表趋势……44
图 5-11 同济新闻 数据集评价指标图表趋势……44

万方数据

表 录

表 2-1 多标签性能指标评估公式……9
表 4-1 测试集训练实例……28
表 4-2 类标签间的权重……29
表 4-3 WEIGHTEDLABELPOWER 投票预测……29
表 5-1 数据实例集描述……33
表 5-2 数据实例集强特征纹理描述……33
表 5-3 同济新闻语料库部分强特征输出……39
表 5-4 EMOTIONS 数据集性能比较……41
表 5-5 SCENE 数据集性能比较……42
表 5-6 YEAST 数据集性能比较……42
表 5-7 同济新闻数据集性能比较……42

万方数据

第一章 绪论

1.1 研究背景和意义

伴随着互联网信息技术的发展，世界正处于信息爆炸却缺乏知识汲取的时代。在中文信息领域里，我国网民已超六亿人，而且互联网普及率也达 45%以上。在移动互联网中，手机网民规模为超过 4.5 亿，随着日新月异的智能手机技术，使用手机上网的网民也被证实为是所有网名中的主力军。与此同时，移动互联网将进一步发展，崭新的服务形态、商业模式、高端技术也将不断涌现，在与大数据、云计算应用的深度融汇的过程中，催生出新兴的电子商业模式，将在不久的将来逐渐对互联网产业产生巨大影响，甚至很大程度上影响到社会经济的未来发展。不止是“双十一”,“双十二”等电子购物节，网络经济日趋明显地突显出自己在中国经济中饰演的角色。即使世界处于经济增长乏力的大环境，互联网产业却如朝阳般展现出前所未有的发展潜力。

英特网信息资源中存在着海量诸如文本、图像和音乐等各语种数据等的多标签分类问题。然而如何在信息中迅速且又高效地挖掘有用信息，准确过滤并定位出可用信息，已日渐成为数据挖掘领域的主流方向。信息化时代越来越迫切地需求自动快速且精度准确的文本分类，基于机器学习的自动文本分类方法正在成为当今领域内的重要研究课题。

数据分类是指前期利用训练样本集所构建的模型体系将测试样本集划分到不定项个类别的方法。传统单标签分类假设类别间关系是相互独立，单个样本只能确定地归于其中某一个类别，现有的算法可支持部分语料库分类精确度高达 90%以上。多标签分类问题指的是，由于数据样本的复杂性分布，在实际应用中，分类样本往往会和多个类别相关联，同样，在互联网信息日趋人工智能及个性化定制的特定环境下，需要将样本准确同时定位到多个类别中。例如一张偏重于描绘大海也有星空的图片，单标签分类只能将其分为海图或天图，多标签分类便可将其同时识别为海图和天图，能更全面的反映该图片的实际特性；又如在文本中的新闻内容，其既包括了教育又包含着经济，赋予它两个标签后，就可以在搜索两个类别时都能检索出它。现实生活中的信息往往如上述样本一样，拥有多个标签的分类问题即为本文所研究的多标签分类问题。

在现有所有的多标签学习任务模型框架中，每个待测样本都会与一个判定类

万方数据

标签集合相关联，多标签学习过程就是要为待测样本预测未知大小的标签结果集。半监督学习的方法已经在分类领域内得到了较为广泛的应用，此方法需要大量的数据集来训练得到较为满意的分类模型。构造训练样本集和相关语料库需要该领域专家花费大量的精力与时间。训练样本过多不仅会使得学习过程变得极为缓慢，甚至有可能造成训练模型数据过度拟合。因此，提出了基于机器主动自学习过程，从而可以有效地克服训练中的瓶颈。自学习过程中，根据现有分类模型来采用的样本选择策略，选择、增加或修改一些有价值的样本从而进行标记，能更好地修正偏差，改进分类性能与效果。

分类基础体系一般由人工构造而成。现有的模式有三种:二类分类问题，表示为属于或不属于；多类分类问题，多个类别，可拆分成多个二类分类问题；多标签分类问题，一个文本可以属于多个类别。目前现有的文本分类算法相对集中于以某种统计模型，将文本用某些特征向量来表示；特征的选择可通过机器学习或者人为数据挖掘得到，从而将文本特征与类别特征进行比较、筛选，得到类别区分度较高的特征向量，构建特征向量空间。

1.2 多标签分类

数据分类是数据挖掘的一个重要研究方向，受到了人们的广泛关注和研究。数据分类问题的研究目标是如何将每条数据准确地划分到某些类别中，其分类目标是找出能描述实验数据集典型化特征的函数或模型的集合，从而用以识别未知数据的类别。构造分类器有很多种方法，例如分而治之算法和分离治之算法。

目前基于机器学习的自动分类方法有贝叶斯分类、决策树、最近邻分类、神经网络和支持向最机等。有一种将多个分类器的判定合并为一个分类器结果的方法叫做“投票”。“投票”方法基于一个思想：对于需要专家系统判定的任务，N个相互独立的分类器判断经过适当调整归并，得出符合要求的的判断结果。如今对于多标签分类问题的主流研究内容分为以下三大方向:

(1)在分类器构建时考虑类间相互关系，基于文本由分属各个类别相关主题特征词混合而成[3]，建立分类函数模型。此研究方向形象地勾画出现实世界的具体内容，实际中很难在父子关系间抽取主题建模。

(2)在判定是初判为相互独立的类别，忽略项目关联。可以将多标签分类问题综合各个二元分类依据的结果，训练出一个排序函数或者阈值判定从而进行配对评分，从而将文本划到评分较高的不定项个类别中。当类别之间存在关系时，这类算法的分类性能明显下降。

万方数据

(3)第三大方向是采用半监督学习方法，构建分类模型时综合考虑层次类别和各个类别文本间的相互映射联系，在自学习的过程同时有效利用类别修正映射偏差，提升分类效果。由于其自学习训练过程的复杂度在达到饱和状态下呈现指数级增加，在构建海量文本分类的快速分类中难以构建出层次关系逻辑图。

分类利用了训练样本建立的函数模型将测试样本集划分到多个类别中。在实际应用中，样本和多个类别相关联，需归到多个类，即为多标签分类问题。多标签分类算法的研究大致可分为整体算法优化和基于分类数据划分两种方法。整体算法优化方法对所有样本和标签构建一个优化问题，例如 Rank-SVM 算法、BoosTexter 算法、多标签 KNN 算法以及最大化熵算法等等。Rank-SVM 其优化形式其实是一个二次动态规划问题，以排序函数进行阈值判定从而选出类标签子集作为预测标签；BoosTexter 算法是依照弱分类器的结果，调整错误分类样本权重，加权组合成强分类器的方法；ML-kNN 以 k 个最近邻居为基础，在使用朴素贝叶斯算法计算测试样本的后验概率，从而确定该样本属于哪些标签；最大化熵算法考虑了样本与标签之间的关系，以及标签与标签之间的关系，使用了一个正则化参数离散调整经验风险概率和实际的分布，避免产生过度学习。整体算法优化方法的优点是没有改变数据的结构，没有破坏类间关系，却需要花费大量时间优化问题，无法将其应用到大规模数据集。因此，目前专家们正在转向更易于实现的基于分类数据划分的多标签分类算法。

在现有的模式识别问题中，实例往往简单地对应唯一类标签。可鉴于客观事物复杂性与抽象模糊性，单实例对象通常从属于多个类标签结果项。例如在文本分类中，一篇文本可以同时属于新闻和经济两类。除此之外，还有多实例单标签（MISL）问题和多实例多标签（MIML）问题。前者指的是单个对象以各个属性组成对应同一类标签的多个实例，后者则是用不同组类属性综合构成的实例，来相对应分析不同的类标签。在图片模式分类中，图片中的各个主要部分都可以代表为实例，对应于不同的类别标签。例如一张主要有大海、沙滩和人群的风景图片，就包含了三大实例类别。现有的这些算法如 AdaBoost.MM, Ad-aboost.MR[35]等能较为有效地解决多实例多标签分类问题，在评估此类算法时，针对特定多标签分类提出了一些特殊的问题的评估标准，如 Hamming-Loss, One-error, Ranking-loss,Coverage 和 Average precision 等。

1.3 国内外研究现状

传统文本分类的研究最早起源于上世纪 50 年，但直到 90 年代，才逐渐形成

万方数据

对机器学习理论的研究。现在已经提出了许多文本分类相关的算法中[1,2]，最著名的有支持向量机模型(SVM)、k 近邻模型(KNN)，贝叶斯模型(NB)等。经过基于知名的英语数据集 Reuters 21578 和标准数据集 RCV1 的系统测试实验，在传统的文本分类的实际应用中这些分类算法是十分有效的。

随着电子文本和知识更新的涌现，只依靠人工制定规则的系统在规模和可移植性上都已经无法分类应用的需求，所以机器学习的方法已经取代知识工程，成为了发展至今的主流分类技术[4]。基于机器学习的文本分类只需在初期进行人工参与算法构造，即构造训练文本和选择算法，其他步骤则由系统自动学习规则后，训练出分类模型，最终应用进行自动分类。Sebastiani 在文本分类技术的基础知识方面，如文本表示、特征降维、分类器算法、性能评估指标等，与 Reuters 语料库上做了实验并详实分析、总结各个步骤方法。但以往的论述在类别体系的多样性、数据集的倾斜、文本类别标注不完整等方面的论述还不足[5]。

Rogati 和 Yang[6]对降低空间维数的特征抽取算法进行了详尽的介绍和实验，结果表明信息增益（IG）、卡方统计或统计量组合的算法效果更佳，针对不同的特征降维方法所适合的分类算法也不同。在人脸识别和手写体字符识别领域，降低维数的特征抽取算法应用得多一些[7]。Lan[8]等人研究了向量空间模型，采用了几种特征权重计算方法，在常用的有监督权重法中，只考虑特征在整个样本空间中的分布，未考虑到出现频率较高的特征具有更强的特征区分能力，由此提出了一种改进的 TF-IDF 特征加权法，使用该方法赋权的分类效果要好于现有常用的权重计算方法。

分类算法的目标是为了提高分类精度，但同时也应该顾及到系统开销。一些构造算法简单的模型，如 Rocchio、Naive Bayes、决策树及 K-NN 算法等，在合适的数据集下表现尚可，却对数据集分布倾斜的情况适用性较差，因此不断有学者提出对这些方法的进一步改善。K-NN 算法最主要的两个不足之处在于首先它是一种懒惰学习法，在动态网页挖掘应用中效率较低；其分类效果依赖于 K 值的选择。对 k-NN 算法的改进主要集中点在于吞吐量和最佳 K 值的自适应选取中[9]。

Joachims 最早提出在文本分类领域应用支持向量机[10]。在相同的数据集上用改进的朴素贝叶斯算法与 K-NN 算法对比实验后[5]，发现支持向量机更稳定且分类性能更佳。支持向量机目前主要用于单类赋值[11]。按策略集成分类方法也可以改善单个分类器存在的不足，使分类性能达到最佳。用 Rocchio 算法[13]进行初步分类后，再用支持向量机对重要类别局部调整，以较低的额外复杂度换取较好的分类效果。大多数投票或按权重计算，输出最终类别以覆盖最优的手段，如

万方数据

Boosting 方法以同一分类方法划分出文本子集，样本训练后，不同的分类器以用于测试文本的分类。

多标签分类问题比二类分类或多类分类等单标签分类问题要复杂得多，因其需要考虑类间的关系[15]。多标签分类问题的算法通常分为两种：一是如同 k-NN 算法、C4.5 决策树算法、贝叶斯算法等将其拆分为多个单标签分类问题[16]。二是如 Rank-SVM 算法[16]，多标签 k-NN 算法 MLkNN[18]，反向传播多标签学习算法 BP-MLL[19]直接改进分类算法使其能处理多个类标签。

传统文本分类算法仍有如下不足之处：1、当类别的规模在一定节点上增加时精确度急速下降，以至分类结果缺乏实际意义。2、分类算法的训练时间通常随着样本规模的变大而以系统无法负荷的规模级增加。3、现实生活中的分类数据集类标签结果往往是层次结构，甚至于不仅仅在父子关系，更复杂的诸如兄弟关系等难以得到实际有效的系统分类。

多标签分类问题按策略来划分，主要有两种方法：big-bang[21]和 top-down。前者更多地考虑类别分层以结构单个分类器。分类器将待分类文档指定到不定项个类别中，具体采用基于支持向量机分类器[22]，相关反馈分类器[23]，基于规则[19]，关联规则[20]的分类器。其时间复杂度大于后者。在基于搜索的分类方法中[27]，采用源搜索出 K 个最相似类别，然后进行小范围更精确地过滤这 K 个候选类别的结果。基于 shrinkage 层次分类方法中，同时考虑父类结点以及稀疏子类结点的平滑参数。Chen 实现了 top-down 自顶向下的层次分类，其分类基础基于多贝叶斯分类器，也有学者是基于 SVM 分类器[28,29]。Xing D K 是针对源搜索搜索结果的分类，以关键词检索小范围候选，然后划分至深层类别集合。

Crammer 和 Singer[39]提出对分解后的数据分类的主题排序算法。当类间关系相对独立时，效果很好，但当类间关系明显时下降。Schapire 和 Singer[30]利用 Boosting 学习排序函数后进行阈值判定，得出结果集。Elisseeff 和 Weston 利用核 SUM 作排序函数。Crammer 和 Singer [31]以文本特征向量内积和类权重向量的内积来排序。Clare 和 King 突破了 C4.5 算法，修正方法使得允许多个类标签，用样本重复采集判定来处理特定少样本类别的修正性问题。Comite et al 扩展了 ADTree 算法[35]，每个节点有一组实数值对应一个类标签。Har-Peled et al.对分类结构做限制[36]，在高维空间里把分类问题转换成二元分类问题。这些方法都没有将类别间关联考虑在内。Ueda 和 Saito [33]提出的产生式体系，虽然包含了任意两个类间的关系，但实际应用中问题较多。Zhu 等人获取类配对间关系是以最大熵模型来限制的。Mc-Callum, Yu 和 Tresp 采取的确是基于潜在变量[34]。Zhang 和 zhou 则采用双层主

万方数据

题模型，基于实例的差异性构建模型。半监督方法在某些情况下相比于基于主题的方法更为高效，Liu 等人分解获得最优的样本标签是通过解带约束条件的非负矩阵。构建两层结构是 Chen 和 Song[37]基于解出 sylvester 方程。随机游走的多标签学习算法是 Jiang 等人[38]提出的。这些方法适合待测文本远大于分类样本的情况，非显示第有效利用了类间的关系。

1.4 论文的结构安排

本论文的结构安排主要分为六个章节，具体内容如下：

第一章的绪论部分主要阐述了文本分类以及多标签分类的在数据挖掘领域内的重要背景和意义、给出了国内外相关工作的研究现状以及本文的主要工作内容及结构安排。

第二章预处理及基础知识介绍。

第三章研究了特征选择及是本文的重点之一。

第四章给出的算法是本文的另一个重点。

第五章分别对算法进行了实验，实验与结果分析针对本文提出的模型和实现的系统进行了实验，并对结果进行了讨论和分析。

第六章对论文的研究工作做出了总结，并给出了本文课题相关的后续研究的方向。

万方数据

第二章 多标签文本分类相关技术

本文讨论的中文文本分类是从由待分类输入文本经过训练后的分类器判定，最终生成更符合文本特征的类别判定，其合理性包括两方面，即文本本体上的意义合理性和正确度上的合理性。本章是文本分类技术的基础，主要讨论了文本分类的定义及过程，介绍了多标签分类所采用的性能评价指，本文实验所采用的的数据集，以及各类目前流行的文本分类基础算法。例如：决策树算法，Rocchio算法，KNN，神经网络，朴素贝叶斯，支持向量机等。

2.1 文本分类的定义及过程

2.1.1 文本分类的定义

文本分类问题即将文档归入不定数个预先定义的几个类别中，自动分类则为以计算机程序经过训练实现这样的分类过程。计算机会自动地分类一篇文章，说的究竟是经济，体育还是科技。 首先，用于分类判定的类别项是预先确定的，其次，文档的判定即使在人为条件下也具有主观能动性，找 10 个人判断一篇文章主题是属于生物，医药还是自然，可能会给出许多不同的答案，一篇文章往往被分配到多个类别中，某些让人信服，某些让人模棱两可，即置信度不一样。文本分类技术其实还能用来判定文章的写作风格，作者主观态度、作者真伪等等，例如判断《红楼梦》到底是否全由曹雪芹所著。目前据此技术构建最多为搜索引擎系统，它不完全等于网页分类。必须使用文本分类技术对数字图书馆，档案和海量大规模文字信息相关的系统进行管理。

2.1.2 文本分类的一般过程

文本分类过程中主要包含三大部分：一、训练及测试文本集构建。建立起人工判定的较为具代表性的已分类文档数据集，获取深度与广度取决于应用分类系统的目标划分。二、文本表示，将文本内容转换成为被计算机分类方法所识别且能分类的形式，比如将对样本类别区分度高的字、词、甚至于句子结构都用数字向量即特征向量来表示。特征选择即在特定的特征空间范围内获取区分度高的特征信息，主要方法包括了特征削减与权重计算。三、构建针对性更高的分类器和

万方数据

方法。依据所提取的文本特征向量，经过不断人工反馈、机器自学习迭代抽取提高该分类器的特征性能，从而将待测文本映射到相应类别中。现有流行的多标签文本分类方法主要有 Naive-Bayes 算法、KNN 算法、SVM 等。四、分类性能评估。目前，有已被广泛认可的衡量分类结果的正确性与精确度多标签测试指标，例如 Hamming Loss、One-Error、Ranking Loss、Coverage、Average Precision 等等。用图 2.1 来表述文本分类的一般过程。

文本训练过程：训练文本集⇒特征选择⇒分类器训练⇒建立分类模型

文本分类过程：待测数据集⇒特征抽取⇒分类器分类⇒性能评估

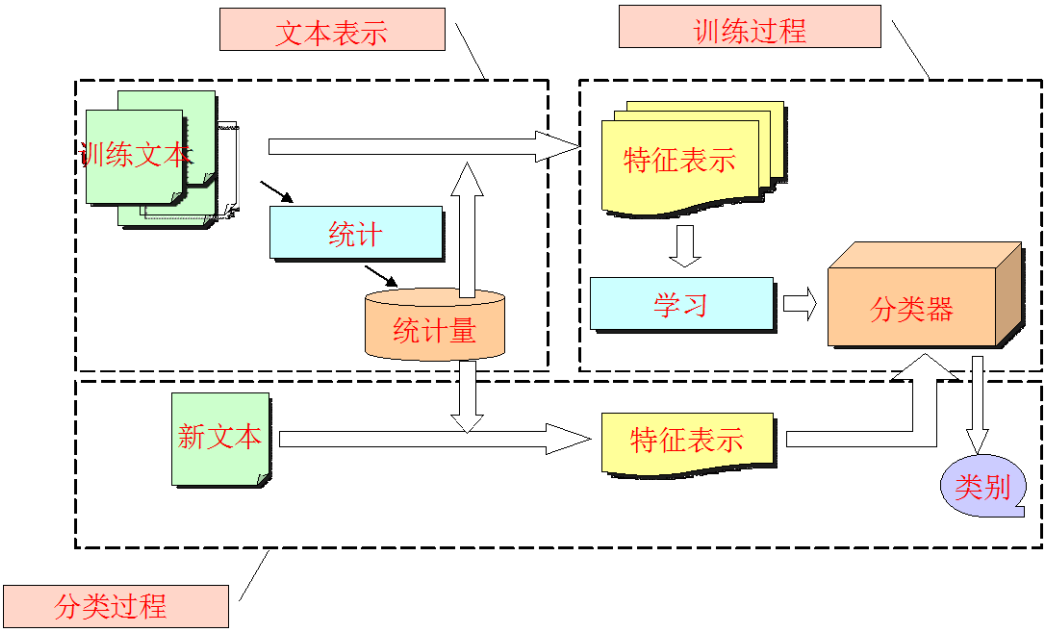

图 2-1 文本分类一般过程

Figure 2-1 The Text Classification General Process

2.1.3 多标签文本分类评估方法

本文选取的多标签性能指标为 Hamming Loss、One-Error、Ranking Loss、Coverage、Average Precision，如表 2-1 所示。Hammingloss 指的是实例真实结果与实例预测结果集间的异或，此评价代表了实例标签对错分类的次数；One-error 是指该预测实例类别相关度最高的类与实际结果的异或，此评价代表了最高排名的标签不在例子实际分类中的次数；Coverage 指的是正确结果的长度，此评估代表了平均每个预测实例需要降低多少格才能找到精确的标签；Ranki-loss 指的是评估了平均标签对的局部排序错误，该评估反应了预测结果在排名上的错误，

万方数据

Average-percision 评估了预测出的标签平均精确程度。前四个方面评估值越小越好，但最后的 Average-percision 值是越大表现越好。

表 2-1 多标签性能指标评估公式

评价指标	评估公式
Hamming-loss ↓	$\frac{1}{m}\sum_{i=1}^{m}\frac{1}{Q}\lvert h(x_i)\Delta Y_i\rvert$ 其中 $h(X_i)\Delta Y_i$ 分类结果为分类结果和实际样本的差集
Ranking-loss ↓	$\frac{1}{m}\sum_{i=1}^{m}\frac{\lvert R(x_i)\rvert}{\lvert Y_i\rVert\overline{Y_i}\rvert}$ $R(x_i)=\{(l_0,l_1)\lvert \frac{f(x_i,l_1)}{f(x_i,l_0)}\le 1,(l_0,l_1)\in Y_i*\overline{Y}_i\}$
Coverage ↓	$\frac{1}{m}\sum_{1}^{m}\lvert C(x_i)-1\rvert$ $C(x_i)=\{1\lvert f(x_i,1)\ge f(x_i,l_i'),l\in y\}$ $l_i'=\arg\min_{k\in Y_i} f(x_i,k)$
One-error ↓	$\frac{1}{m}\sum_{i=1}^{m}H(x_i)$ 结果中第一名在实际标签中 $H(x_i)=1$
Average Percision ↑	$\frac{1}{m}\sum_{i=1}^{m}\frac{1}{Y_i}P(x_i)P(x_i)=\sum_{k\in Yi}\frac{\lvert\{l\lvert f(x_i,l)\}\ge f(x_i,k),l\in Y_i\}\rvert}{\lvert\{l\lvert f(x_i,l)\}\ge f(x_i,k),l\in y\}\rvert}$

注：↑表示值越大效果越好，↓表示值越小效果越好。

2.2 文本分类的分类方法

目前，机器学习与概率统计的方法日趋成熟。本节中将主要介绍一些基于概率方法类、基于文本实例类、支持向量机等方法，简要说明决策树算法、Rocchio 算法、神经网络、朴素贝叶斯算法等。

2.2.1 决策树算法

决策树通过构造决策规则，是以样本的属性作根节点，以取值作为分支树结构，归纳和分析样本信息论熵。样本判定决策树取决于层次化结构，root 点在样

万方数据

本中占最大比重的信息量，其子节点为其子树中依层次类推，其最底层叶节点是样本类别。对新样本分类时，决策树对从根节点开始，根据样本属性的取值，自顶向下直到树的叶节点，即对新样本分类。常用的决策树算法是 CART 算法、ID3 算法、C4.5、C5.0 等。自顶向下构造决策树的主要步骤如下:

1)随机在训练集中选择包含正反例的候选子集;

2)构建决策树;

3)子集外样例以决策树进行类别判定，去除错误例子集

4)将错误例子集插入候选例子集，检查是否为空，重复步骤 2)，否则结束。

构建决策树过程：

1)对集合计算各特征的互信息后去除最大熵特征;

2)归一化取值相同子集;

3)若子集依旧包含正反例，递归调用构建算法;

图 2-2 决策树算法表示图

Figure 2-2 Diagram of decision tree algorithm

2.2.2 Ricchio 算法

向量的相似度度量方法有两种：欧几里德距离和夹角余弦。判定距离最小为其所属类别。Rocchio 算法是基于类中心向量距离分类速度较快的简单向量距离算法。文本类别相似度即为在训练时采取中心特征向量，分类时通过计算待分类文本特征向量与各个类别的欧式距离差。Rocchio 算法以向量空间模型理论，向量空间模型（VSM）采用向量将文本处理后转化为空间中向量的运算。

类别 C 的中心向量所对应特征 tl 的权重可由下式计算：

万方数据

$$w_{Ci} = \beta \cdot \frac{\sum_{d_j \in C} w_{ji}}{|C|} - \gamma \cdot \frac{\sum_{d_j \notin C} w_{ji}}{N - |C|} \tag{2-3}$$

其中 W_{ji} 是文本向量 d_j 中第 i 维特征向量的权重，N 是总实例数，|C|是类文本数。β 与 γ 分别是调整正、负例样本影响的参数。

文本类别相似度计算公式为：

$$Sim(d, C) = \frac{\sum_{k=1}^{m} w_{dk} \times w_{Ck}}{\sqrt{(\sum_{k=1}^{m} w_{dk}^2)(\sum_{k=1}^{m} w_{Ck}^2)}} \tag{2-4}$$

W_{dk} 为文本 d 中特征 t_k 的权值，W_{ck} 为类中特征 t_k 的权值。最终分类判定方法有两种：一是相似度，并将这些相似度从大到小进行排序，最后把文本归到与其相似度最大的类中。二是首先为个每个类别确定一个文本向量相似度阈值，若其与类标准向量的相似度比较高于给定值，那么文本属于此类，这样文本会被归到不同的类别中。

2.2.3 KNN 算法

zkq 20150910

KNN 算法即 K 最近邻算法，找出文本实例与特征空间中的 k 个特征空间中最邻近的实例子集，将其归属于最大类别。算法只依据最邻近的 K 个实例的类别来决定待分样本所属的类别。该方法从原理上依赖于极限定理，可以很好地避免样本的不平衡问题，也更适合对于类域的交叉或重叠较多的待分样本集。

KNN 算法主要的瓶颈在于时间复杂度，全体训练样本需要和每个待检测分类样本都必须计算其到的距离从而求设定的最近 K 个相邻节点。为解决大规模数据问题，解决方案通常是对已知样本集事先去除对分类作用不大的样本或随机抽取。

KNN 算法是一种懒惰学习法，时间复杂度为 O(L*N*m)，L 为测试实例总数，N 为训练样本总数，m 为特征数。分类样本时，必须对整个训练样本进行计算其各个类别权重以及相似度排序。K 值的适当选取是决定分类性能的关键，根据实验观察或经验取值。

万方数据

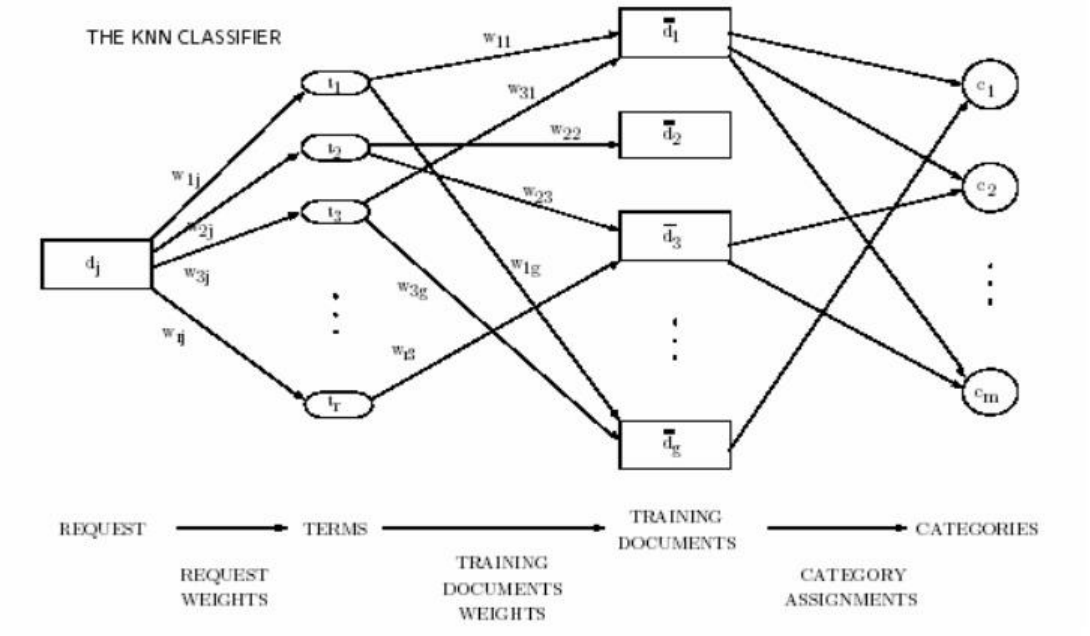

图 2-3 KNN 算法表示图

Figure 2-3 Diagram of K-Nearest Neighbour algorithm

2.2.4 神经网络算法

zkq 20150910

神经网络算法是构造逻辑单元，加权系数求和若超过了阈值，则输出一个量。如输入 a_1，a_2.....a_n 和其权重系数:w_1, w_2, ...,w_n，求和计算出的 ai*wi，其中 a_i 是各条记录出现频率或其他特征参数，w_i 是特征评估模型中的权重系数。神经网络分类方法是经验风险最小化原则，依旧有难以确定神经网络层数和神经元个数等现象，常常陷入局部极小化以及过学习情况。

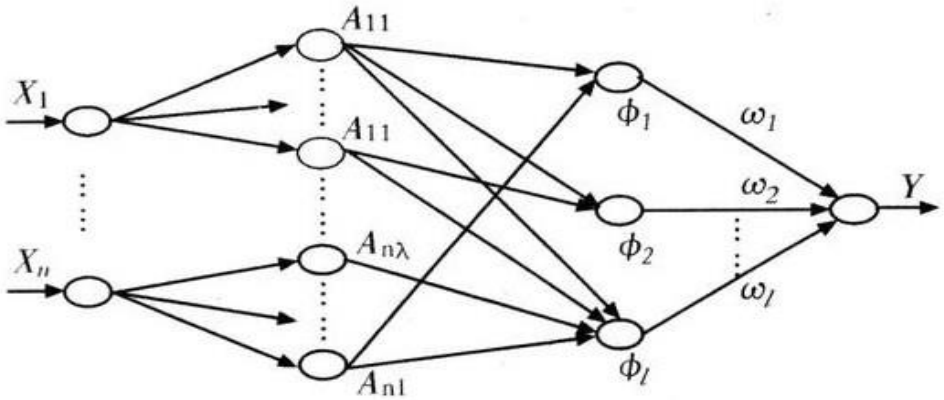

图 2-4 神经网络算法表示图

Figure 2-4 Diagram of Neural Network Algorithm

万方数据

2.2.5 朴素贝叶斯算法

NB 分类器，是一种基于贝叶斯定理，统计学理论的分类方法，采用已知类别的训练集数据来测试得出待测样本属于各类的先验概率。因此前提条件假设是做出类别特征各不相关的独立。朴素贝叶斯分类模型是一种典型的基于统计方法的分类模型，它利用先验信息和样本数据信息来确定事件的后验概率。根据贝叶斯公式 2-5：

$$P(C_i \mid D_j) = \frac{P(D_j \mid C_i)P(C_i)}{P(D_j)} \tag{2-5}$$

可知贝叶斯文本分类的任务是将表示成为向量的待分类文本 D（d_1，d_2,...,dn)归类到与其关联最为紧密的类别 C(C_1，C_2，.....,C_n)，因为 P(D_j)不会影响改变结果，所以可以被忽略。P（D_j|C_i）可以被下列公式 2-6 得到

$$P(D_j \mid C_i) = \prod_{i=1}^{m} P(A_k \mid C_i) \tag{2-6}$$

A_1、A_2、A_3…..A_k即为文章中的特征

$$\hat{P}(C = C_i) = \frac{N_i}{N} \tag{2-7}$$

$$\hat{P}(A_k \mid C_i) = \frac{1 + N_{ki}}{m + \sum_{k=1}^{m} N_{ki}} \tag{2-8}$$

公式 2-7 中 N_i 表示在 C_i 中的文本数量

公式 2-8 中 N_{ki} 表示特征词 A_k 在 C_i 中出现的总频率，m 表示为特征数量

如果是单标签分类器：定义如果

$$P(C_i \mid D_j) = \max_{y=1}^{k}\{P(C_y|D_j)\} \quad \text{那么 } D_j \in C_i \tag{2-9}$$

即确定最大值为其文章的分类

如是多标签分类器：定义阈值

$$P_{yu} = \frac{\sum_{i=1}^{n} P(C_i|D_j)}{n} \tag{2-10}$$

当 P（C_i|D_j）≥P_{yu}时，则可以判定文章 D_j 属于 C_i，只有当所有 P（C_i|D_j）相等时会出现最巧合的属于所有分类的情况。朴素贝叶斯分类算法以最大后验概率的类别来预测样本。其较快且效果尚优的学习过程得到各方向分类领域的了广泛应用。

万方数据

2.2.6 支持向量机

支持向量机模型(SVM)由于其提供支持向量回归和支持向量聚类，准确度较高，从而适用于文本分类。支持向量机采用的是非线性特征的变换空间，将目标向量映射到类别分向量，从而得到了线性可分的高维空间，然后是选择最优超平面对其进行分类，得到分类结果。

SVM 分类器将其转化为二类分类来解决多标签文本分类，如图 2-5，然后以投票机制处理，重点在于核函数以及相关参数的调整。

SVM 的主体思想概括为两点:

(1)采用非线性映射算法在线性不可分时，将低维输入空间转化为高维特征空间使其线性可分，采用线性算法对样本的非线性特征进行高纬度线性分析。

(2)基于结构风险最小化理论，建构最优分割超平面，全局最优化分类器，以某个概率在整个样本空间的期望风险上满足上界。

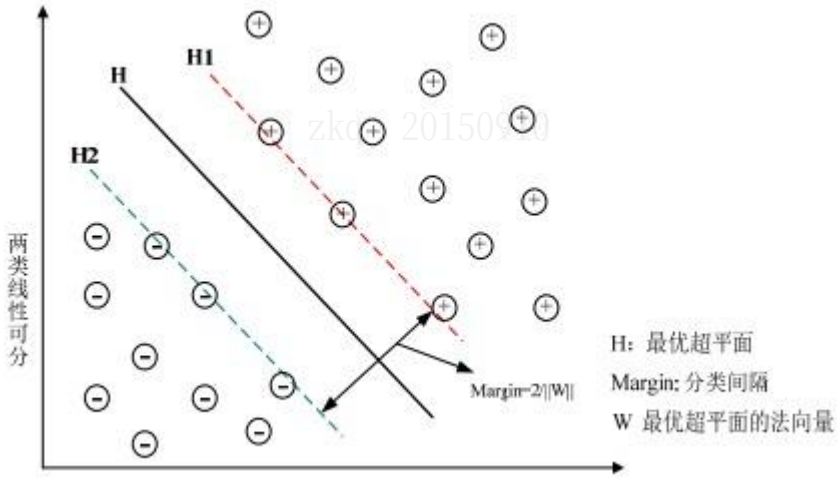

图 2-5SVM 超平面划分模型示意图
Figure 2-5 Hyper Plane Classification Diagram of Support Vector Machine

2.3 本章小结

本章第一部分主要叙述数据挖掘领域内文本分类技术及其一般特征，并且对目前主流的主要包括主流的多标签分类系统的性能评估指标进行了介绍。对文本分类的概念理论做了基础介绍，叙述了一些常用文本分类的理论依据和实际应用情况。本章介绍的分类性能判定标准，分类模型的一般定义与基本方法，是下一

万方数据

步工作的基础。

zkq 20150910

万方数据

第三章 基于中文多标签分类的特征选择

本章给出了文本分类特征选择过程的整体模型图，并详细介绍了在特征选择前的基础工作及步骤，主要包括：语料库 Dataset 预处理、构建文本特征向量过程、强特征纹理统计，多标签特征选择加权计算等过程。对特征选择模型，本章通过对于文本语料的强特征纹理筛选，特征向量加权计算，分析了特征选择中的关键技术，并给出了特征选择的整体流程图。

3.1 文本特征选择

目前常用特征选择方法主要有文档频数 DF、互信息 MI、信息增益 IG 和 CHI 方法等。基本思路是对通过文本分词后计算每一个特征向量项的某种统计度量值 ，在经过某种特定算法后为每一个类别或综合层次类别的划分，设定一个阈值 T,过滤度量值小于 T 的特征，保留有效特征。文档频率(DF)基于频率文档类别统计量，是其他特征评估标准的前置条件与基础；信息增益(IG)，同时考虑文本出现特征与文本未出现特征；互信息(MI)，不仅考虑到高频特征词，同时考虑了低频词信息量，并且赋予了低频词更高的互信息权重，因为可能过于偏重低频词的影响，分类效果时常不好。

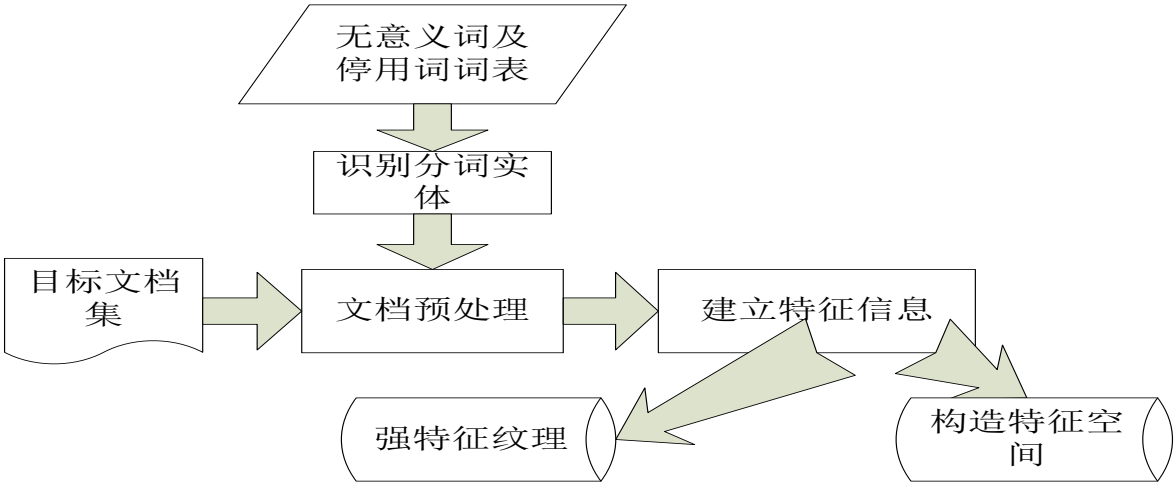

图 3-1 系统整体模型图
Figure 3-1 Procedure of the entire system

本文提出的是将强特征纹理及改进的特征选择权重相结合的多特征选择方

万方数据

法，在不同的语料库实验条件下，为了减弱噪音量对判定结果的影响从而综合评估特征对于文本分类。

在对待分类样本进行实验分类时，即特征分布得越离散，越不均匀，那么该特征的区分度越大，也就是说属于高强度类别判定依据，合适高效的特征选择方法就是为了找到样本类别区分度大，且能较好地权衡各分类类别的算法。通常用标准差或方差来计算特征离散度，本文使用的特征选择总体流程系统整体模型图如图3-1所示，对区分度高的特征赋予定量的权重值，标准化文档概率，特征选择的概率标准差将在性能和速度上对后期多标签文本分类有着深远的影响。

3.2 特征选择方法

3.2.1 过滤无意义信息

过滤无意义信息指的是通过已知的特征筛选，删除文本中的一些无意义的符号、用户其他信息、特殊表情符号等，目的是将与主题相关内容提取出来，以某些社交网络的文本为例，各个论坛或微博等往往在发帖和文字的同时带有固定格式标识，需要去除如“@某某某”、“/微笑”等与主题无关项，但诸如“#相关话题描述#”等与主题相符合的内容却要保存。除此之外，URL 地址、Html 类标签信息等与主题无关的内容可能会引入了较高的噪声，必须预处理阶段删选。例如：微博文本为“#请停止破坏环境，还一个洁净的地球# 地球太累了[愤怒] @环境保护协会 http://tencent.cn/tmaczhouGttf”，经过过滤无用信息后，文本内容被处理为“请停止破坏环境 还一个洁净的地球 地球太累了”。

3.2.2 汉语文本自动分词

本文采用的是最大正向匹配的中文分词算法，相当于分词粒度等于0。假设在分词词典中的最长词有k个汉字字符 ，用被处理文本的目标字符串中的前 i个字作为匹配字段查找字典 。若字典中存在这样 的一个K字词，即为匹配成功，作为一个词切分出来。如果词典中找不到这样 的一个K字词，即为匹配失败，将匹配字段中的最后一个字去掉 ，对剩下的字符串重新进行匹配处理……如此迭代进行下去，直到匹配成功，切分出一个词或剩余字串的长度为零为止。然后取下一个K字字符串进行匹配处理，直至扫描完文本。文本向量空间是由正交向量组成。训练语料库文本首先过滤无意义信息，经过分词后去除停用词，然后统计词频，最

万方数据

终表示为特征文本向量。

3.2.3 汉语文本粗降维

粗降维指的是训练文本经分词后首先经过去掉停用词的处理，即为去掉一些没有实际分类意义的高频词、稀有词。高频词会多次出现在各种类别的文本中，稀有词属于偶尔出现在各个类别中，没有实际检索意义，因些在分词之后，需要清除停用词，同时清除些多余的符号等冗余。本文中采用了建立停用词表，通过词表法去掉高频词和稀有词。

3.2.4 文本表示模型

向量空间模型是一个用户表示文本的代数模型，是目前应用最多且效果较好的文本表示法之一。将文本抽象表示成为高纬度空间向量，每个特征值维度值就是对应特征的相对类别权重，权重度量了相应特征对应文档的类特性。文本 D 的向量空间模型表示 $D = D(t_1, w_1; t_2, w_2; .....; t_n, w_n;)$，其中 t_i 是对应特征项， w_i 是 t_i 特征的权重，$1 \le i \le n$。

实验文本D需要严格满足：1. 各个特征项 t_i $(1 \le i \le n)$ 独立不重复；2. 各个特征项 t_i 无指定先后顺序，故将特征项是n维坐标系 $t_1, t_2, ..., t_n$，对应权重 $w_1, w_2, ..., w_n$ 为相应的维度坐标值，如图3-2所示，可以用n维空间中的向量来表示实验文本。称 $D = D(t_1, w_1; t_2, w_2; .....; t_n, w_n;)$ 为文本D的向量空间模型。

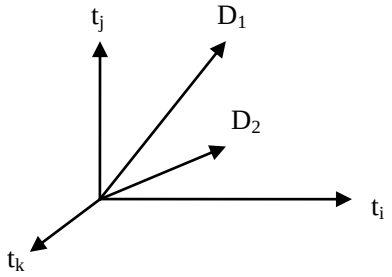

图 3-2 文档的向量空间模型示意图
Figure 3-2 Text's Vector Space Model Figure

将文本数据以向量空间模型表示成结构化数据，或者以向量矢量距离计算的得到两个判定文本间的联系。向量空间模型普遍在文本检索、文本过滤、文本摘要、文本分类、关键字词提取等领域中应用。

万方数据

综上，对实验文本D，假没所有训练实验集的特征数是n，那么每一个文本d都可以构成一个n维特征向量空间：文本训练向量的表示为$X_i=(x_1,x_2...x_n),x_i\in R^n$，其对应的标签集表示为$Y_i=(y_1,y_2...y_m),y_j\in\{0,1\}$。当样本属于第J类时，$y_j=1$，不属于第j类时，$y_j=0$。单标签分类问题即为多标签分类的一个当Y向量的值中只有一个1时的特例。它的向量空间模型可表示为$D=D(t_1,w_1;t_2,w_2;.....;t_n,w_n;)$，其中$t_i$是特征项，$w_i$是$t_i$对应的权重，$1\le i\le n$。

3.2.5 常用特征选择方法

目前主流的特征选择方法有如下几种：

1）信息增益

信息增益（IG）是基于信息熵的特征选择方法，在信息论中，信息熵是用于度量信息量的一个概念，也可以说是系统有序化程度的一个度量。文本特征越是有序，信息熵就越低；反之，一个系统越是混乱，信息熵就越高。所以，信息熵可以更好地区分特征对于类别区分度的贡献。对于特征t_i，其对类别C_j的信息增益统是其于类别C_j中是否存在来计算，即为度量前后信息熵差[1]，IG 的计算公式如式3-1：

$$IG(t_i)=p(t_i)\sum_{j=1}^{m}p(C_j\mid t_i)\log\frac{p(C_j\mid t_i)}{p(C_j)}+p(\bar{t_i})\sum_{j=1}^{m}p(C_j\mid\bar{t_i})\log\frac{p(C_j\mid\bar{t_i})}{p(C_j)} \quad (3\text{-}1)$$

m 为类别个数。贡献度越大的特征项的信息增益值越大，也更有区分度。进行特征选取时，设定判定阈值，保留信息增益值大的特征项，可以较为妥善地选择出分类目标特征。

2）文档频率

特征选择项的文档频率（DF）是样本特征在训练集中的出现次数[2]。DF 特征选择前提主要基于假设：DF 值低于阈值即为低频特征，对分类有较少的特征信息量，为提高分类速度，避免维度爆炸，直接从特征空间去除，简单有效降低了特征空间维度。

3）互信息法

语言统计模型中的互信息（MI）越大，其特征类别相关度越高。特征项与类别互信息量计算公式如式 3-2 以及式 3-3：

万方数据

$$MI(t_i,C_j)=\log\frac{p(t_i\mid C_j)}{p(t_i)} \tag{3-2}$$

$$MI_{\max}(t_i)=\max_{j=1}^{m} MI(t,C_j) \tag{3-3}$$

m 为类别个数。贡献度越大的特征项的互信息值越大，也更有区分度。进行特征互信息量选取最大的前 n 个保留特征项，可以快速准确地选择出待分类目标的特征项。

4）χ^2 统计量法

χ^2 统计值高低[3]判定了特征项与类别间的关联程度，关联度越大，信息量也越多。其计算公式为 3-4：

$$\chi^2(t_i,C_j)=\frac{N\times(A\times D-B\times C)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)} \tag{3-4}$$

N 是训练数据集文本总篇幅数，A 是包含特征项t_i且属于类别C_j文本个数，B 是包含特征项t_i不属于C_j的文本个数，C 是不包含特征项t_i的但属于C_j文本个数，D 即不包含特征项t_i又不属于C_j且的文本个数。

$$\chi^2(t_i)=\max_{j=1}^{m}\chi^2(t,C_j) \tag{3-5}$$

通过计算最大 χ^2 值，m 为类别个数。贡献度越大的特征项的 χ^2 值越大，也更有区分度。进行特征 χ^2 值选取最大的前 n 个保留特征项，可以选择出待分类目标的特征项。

以上这些特征选择方法主要是构造阈值评估体系,对训练样本特征集合中评估每个特征，为每个特征选择项赋予权值。排序后提取可调节的预定数目最优特征集。其效果取决于特征评估算法的质量以及实验语料库偏差性。

信息增益的对于计算信息熵差会造成高时间复杂度，人们普遍采用的是交叉熵和互信息量的方法。由于中文分词的特性不如英文空格那么简单明显，上述这些方法,对于不同语料库有各自的长处,但效率不高。中文分词计算量巨大，提取效率低,从而影响到文本分类系统的整体速度与效率。另外，计算特征词的权重也是一个极为复杂需要维度缩减的计算方法。

3.3 改进的特征选择方法

3.3.1 强类别纹理挖掘算法

万方数据

强类别纹理指的是文本中的某些特征强烈代表了其属于某一类别，其覆盖了某各类中尽量多的文档数目；强类别纹理具有最强的区分类的能力。在分词特征字抽取的基础上，对每个文档进行特征分词的自动抽取，并最终形成“D_x：W_1，W_2，W_3，…，W_n”，等“文档号—概念”链表。链表的排序规则是特征字串在文档中的重要程度。根据“D(文档）——C(类别)”之间的人工标注信息，用并行操作形成：C_y：W_1，W_2，W_3，…，W_m链表，m>>n。链表的排序规则是各个特征字串出现的文档数量递减。

强类别纹理选择过程如下：

（1）去掉泛滥纹理：对于每个类号，观察C-W链表中的每个W，如果W出现在每个类号中，则去掉这个W。

（2）标记强纹理：对于C-W链中的每个W，如果出现在主类别的概率大于次类别概率Ω%，则标记为强纹理。Ω的值随着各个语料库文本偏向不同而改变，一般取值范围为30%-50%。

（3）特征迭代：

1）结束条件：对于每个类，被标记的强特征和所有D-W链的交集非空，则算法结束。

2）去掉任意X篇奇异文档：如果D-W链表与所在类的C-W交集为空，说明该文档无分类代表性，去除该文档。满足结束条件，则算法结束。

在数据库中增加列：交际是否为空、是否被当作奇异文档或者增加了强特征；

选择奇异文档或增加强特征的顺序：概念的权重、doccount、classcount等。

3）根据任意Y篇文档增加强特征：从D-W和C-W交集为空的文档中提取新特征并加入到C-W中，同时标记为强类别纹理。满足结束条件，则算法结束。

迭代结束时，给出去掉的奇异文档比例。最终的得到的记为各个分类的强类别特征纹理。

3.3.2 常用权重计算方法

目前在文本分类中流行的权重计算法是 TF-IDF 和布尔权重法。

1）布尔权重法

对于特征权重进行 0-1 值的赋予常被用在两类分类问题上，但需要对特征维度数有很好的掌控，计算如公式 3-6：

万方数据

$$w_{ij}=\begin{cases}1, & 若tf_{ij}>0\\ 0, & 否则\end{cases} \tag{3-6}$$

其中tf_{ij}表示特征项t_i在D_j中出现过即权重值为1，否则为0。

2）TF-IDF权重法

相对于 0-1 布尔权重忽视了特征项对文本的类别相关联度，在多标签文本分类的模型中，TF-IDF 函数更为普遍地用来权衡每个特征项权重。

TF-IDF 主要思想是：倘若字、词或短语在某一篇文档中出现的频率很高，但在其他类别文档中出现的频率很低，即认为该特征区分度高。TF-IDF 的计算如公式 3-7：[4]

$$Weight = TF \times IDF \tag{3-7}$$

TF 指的是词频，IDF 指的是倒排文档频率，前者指特征在待测单个文本中出现频数，后者指特征项在整个文本集中出现的频数的倒数，如公式 3-8。如果特征项在文本中的出现次数很多，便对于其权重的赋予可以很好地代表该文本所相关类别。

$$IDF=\log\frac{N}{n} \tag{3-8}$$

n 表示有该特征项的文本篇数，N 为整个训练数据集的总文本篇数，用来修正改特征项的偏差。

可以得出的结论是，若特征项属于训练文本集的所有文本中，那么其 IDF 值便为 0，则 Weight 为 0，此时该特征项完全不具有任何区分性。IDF 的可以更好地削减多频简单词、以及各类相关普遍的权重。

3.3.3 改进的特征选择和加权抽取

无论是字词、还是字串、以及特征项的词性，在中文领域都会在切分后呈现较高的数量级。为了计算时间与空间的合理，特征选择的目的是只保留区分度较高的特征项，并且在避免特征过度拟合后的失真。

特征簇的概念对某类具有高代表性的集合，属于所有特征集的子集。某个类C的特征簇用符号SofC(C)。若T为特征选择结果集合，T与各个特征簇的关系为：T=SofC(C_1)∪SofC(C_2)∪…∪SofC(C_K)。特征选择的最终目的是为了特征集能够满足：1.特征簇间的交集越小越好，这样可以更好地避免类之间干扰信息的产生，特征分类结果的准确性越高。理想结果是各个交集皆为Null，即θ= SofC(C_1)∩

万方数据

SofC(C_2)∩…∩SofC（C_K)。即特征簇间无干扰信息，该集合属于最优情况。2.对应包含类分布相对均匀或为其加入类别调整参数，从而尽量避免样本倾斜度所带来的偏差。特征簇内特征项在其相应类内部的分布权重反映了该特征项之于类的重要程度。若某特征项平均地出现在各个该类样本中，说明其对于此类有很强的代表性，若只特定地急剧出现在某个样本中，说明该特征项属于特殊文本的特殊情况，代表性不强。

综上理论指导，本文在改进了传统的特征选择和加权方法的理论基础后，提出了新的特征选择方法。

设 $f(w,c_j)$ 为特征项w和C_j类的相关程度，特征选择函数即为S(w)，如公式3-9所示：

$$s(w)=\max_{1<i<k}[f(w,c_i)-\sum_{1<j<k,j\neq i}f(w,c_j)] \tag{3-9}$$

S(W)的值反映了特征项w对C_j类的代表性，值越大越明显。倘若把C_j类的的所有样本视为文档数据集，即假定C_j类由一组文档Dx（$0<x\leq D_j$，D_j是C_j类中包含的文档篇数)形成,则 $f(w,c_j)$ 的值应该正比于特征项w在C_j类出现的频率，与其在Cj中分布均匀度呈正比。fincij是特征项i在类j中出现的频数。dofwij为类j中含有特征项i的文档的个数。不失一般性，令w为T中的特征项i，则 $f(w,c_j)$ 定义如公式3-10：

$$f(w,c_j)=\frac{\log(finc_{ij}+1)\log(\frac{dofw_{ij}}{D_j}+1)}{\sqrt{\sum_{i=1}^{|T|}\log(finc_{tj}+1)\log(\frac{dofw_{tj}}{D_j}+1)}} \tag{3-10}$$

相比于传统的特征加权算法往往没有在函数中启用调节因子，或其调节因子只考虑了整个样本集对该特征的影响偏差。文本采用的调节因子更偏重于类相关信息对特征项的影响，采用的特征加权函数定义如公式3-11：

$$a_{ik}=\log(f_{ik}+1.0)*(1+\frac{1}{\log(K)}\sum_{j=1}^{K}\left[\frac{finc_{ij}}{n_i}\log(\frac{finc_{ij}}{n_i})\right]) \tag{3-11}$$

上式中，K为样本集中的类别总个数；$finc_{ij}$表示特征项W_i在个C_j类中的出现频率；$\frac{1}{\log(K)}\sum_{j=1}^{K}\left[\frac{finc_{ij}}{n_i}\log(\frac{finc_{ij}}{n_i})\right]$为调解因子。如特征项$W_i$均匀分布于各个类，可以很明显地看出，$a_{ik}$为最小值0；当特征项$W_i$只出现某个类中，调节因子为0，$a_{ik}$

万方数据

为最大值 $\log(f_{ik}+1.0)$ 。

3.4 多标签分类特征选择算法的框架

Input： D:Multi-label trainset

output：Y:feature selevtion result

C-W :texture features for each category

Process:

Step1：//文本预处理

For i=0 to n

Take the word segmentation and computer the word frequence saved in array DWFi

End for

Step2：//计算强类别纹理

For j=0 to n

Arrange the (Dj：W1，W2，W3，… ，Wn)

To (Cy：W1，W2，W3，…) with count ；

For k=0 to m

Cut down the Wk which belongs to all category

For h=0 to m

Pick the Wh which the main category frequency-secondary category frequency>Ω

Get the strong texture features array C-W

End for

Step3： //计算所有特征的权重函数s(w)

For i=0 to |n|

$$f(w,c_j)=\frac{\log(finc_{ij}+1)\log(\frac{dofw_{ij}}{D_j}+1)}{\sqrt{\sum_{i=1}^{|T|}\log(finc_{tj}+1)\log(\frac{dofw_{tj}}{D_j}+1)}}$$

$$s(w)=\max_{1<i<k}[f(w,c_i)-\sum_{1<j<k,j\neq i}f(w,c_j)]$$

Get the Sij[][] who holds the feature and weight

万方数据

End for

Step4： //按照特征选择偏差要求，选择特征个数

For i=0 to |S|

If($S_i(w) > \frac{\sqrt{5}-1}{2} \sum_{i=0}^{n} S(w)$)

{Add Si to array Result

}

End for

Step5: //计算每个特征在每个类中的权值

For i=0 to |Result|

$$a_{ik} = \log(f_{ik} + 1.0) * (1 + \frac{1}{\log(K)} \sum_{j=1}^{K} \left[\frac{finc_{ij}}{n_i} \log(\frac{finc_{ij}}{n_i}) \right])$$

End for

3.5 本章小结

本章首先给出了多标签文本分类特征选择的整体模型图，以中文文本数据集为输入，以强特征纹理集以及特征选择空间向量为输出，系统整体模型图包括文本预处理表示及改进的文本多标签特征选择模型。3.2 节主要描述了文本特征选择的一般基础工作，包括过滤无意义信息、汉语文本自动分类、文本粗降维、文本表示四个关键步骤。3.3 节在描述了常用的特征选择算法之后，以强特征纹理提取、改进的特征选择加权函数为重点，提出了改进的文本特征选择算法。3.4 节给出了该特征选择算法的具体流程描述。

万方数据

第四章 相关信息加权的自适应多标签分类算法

本章简单介绍了目前常用的文本数据挖掘中的多标签分类法，包括：Navie-Bayes 算法，MLkNN 算法及 Rakel 算法。本章结合了问题转化和多标签算法改进的思想，提出的是一种在各类特征选择基准调整后，基于已有单标签分类结果进行加权、自适应阈值设定，不同权重投票相结合的方法，对待分类实例进行多标签分类，能提高多标签文本分类的分类准确度与精度。

4.1 常用多标签分类算法

针对多标签文本分类目前有两个主要方法，分类问题转化法以及分类算法改进法。前者是将一个多标签问题转化成一组单标签问题后运用已有的单标签分类方法解决,其最大的优势在于灵活性,通过从现有的单标签分类器直接抽象成一个特定的分类器来适应需求。常见的有 BR(BinaryRelevance)、基于标签对比 PW (pairwise comparison)、LP(LabelPowerset)等算法。BR 算法的优势在于概念上的简单和相对快速，但却被认为其脱离了标签间的相关信息，PW 算法的缺点在于其时间复杂度过大，LP 算法的缺点在于其只能对新例子进行分类，而对训练集中的例子过度拟合。后者则是通过改变已有的单标签分类算法，从而使其能够处理多标签数据，如 AdaBoost.MH 算法，其对由简单决策树算法产生的弱规则进行加强，经若干次迭代后，得到一个准确度更高的规则，但训练速度慢，难以处理大文本量信息、ML-kNN 算法、贝叶斯算法等等，他们训练速度快，但若原始语料出现较大的类别偏差，会降低效率。

4.1.1 Navie-Bayes 算法

在文本分类中，Naive-Bayes 分类器凭虽然有独特的简易性，但往往能处理极为复杂的，属性个数较多的分类问题，改进的多标签朴素贝叶斯分类方法综合了朴素贝叶斯的简易性以及贝叶斯网表示依赖关系的能力,使其能容纳属性间存在的某种依赖关系。

不同类别相应条件特征下耦合分离表示特征项的分布假定为相互独立的估量维度。在独立条件假设下维度爆炸的问题很大程度上得到了有效的控制,即使样本

万方数据

的特征个数在训练语料中急剧增长，也只需要以较小的成本与代价就能满足算法计算的需要。但由于实际样本的复杂与不平衡程度，朴素贝叶斯通常无法精确估算出各类概率，也就是说，只能满足类别差异度较大的分类要求。朴素贝叶斯分类器中，由于分类结果项只需与同文本其他类结果项做比较，加上阈值的判定设定可以使得后验概率允许偏差，在对分类影响性不大的情况中，少量类特征项的偏差不会影响分类结果的准确度。Navie-Bayes 分类器拥有足够的鲁棒性来允许其概率模型上存在的偏差。

4.1.2 ML-Knn 算法

Multi-Label k-Nearest Neighbor 简称 MLkNN 是从熟悉的 KNN 算法派生而来。由于针对每个测试样本，它的 KNN 都已经在学习样本中确定，所以根据这些已经获取的近邻节点的前期训练统计数据，其训练样本分布与特征项的选择决定了分类的准确度，用最大后验概率原则(MAP)去决定测试样本的标签集合，最大后验概率是基于 KNN 对每个标签的前验和后验概率。ML-KNN 是在投票机制与贝叶斯方法融合的分类算法，该方法集成了贝叶斯算法与思路简单的优势，并且更进一步地降低了错误率、提高了分类性能，但其大计算量与、低效率的瓶颈，使其无法应用与海量文本数据集或者对网络实时性有要求的情况。

多标签 K 近邻算法与 KNN 算法一样，忽略了类标签之间的相关层次性，同样只适应于类别独立的情况。但往往现实生活中的分类应用场景中，标签集之间大多有着内在的层次联系。

4.1.3 RAkEL 算法

随机游走来源于物理学中分子“布朗运动”,它是微观粒子的运动形成的一种模型。将其应用与文本分类算法中，主要基于其在合理范围区间内的遍历过程能带来更好的准确度。当训练文本数据线性映射成为多标签随机游走图后，每当输入一个待分类文本，便会随机建立一个多标签随机概率分布的游走系列。其算法复杂度换来的是其经过便利后可以得到图中每个节点的权重概率分布，再将这个点权重概率分布转化成每个标签的概率分布后，随机游走模型变建立完成，根据随机游走权重计算，可以得出该待测样本多标签文本分类结果集。

万方数据

4.2 信息加权模型算法

信息相关模型加权的基本思想是，从一个文本出发，随机找到其相邻文本，并计算出文本间的距离作为权重。遍历其在一定距离范围内的邻居文本，反复迭代后得到一个与初始文本相关度最大的各个文本并得到距离概率分布。

首先将训练集合 D 映射成模型图中的一个点集合 V，对于待处理点计算其 vi 相邻点的欧氏距离并且将其相连，基于欧氏距离的相似概率可定义为：

$$SIM_E(d_u,d_i)=\sqrt{\sum_{j=1}^{|C|}|c_{uj}-c_{ij}|^2} \tag{4-1}$$

模型图可表示为：图 G 中有点集合 V，其包含的边为距离在一定范围内的相邻点

$$\begin{aligned} &G=(V,E)\\ &V=\{v_i|x_i\in X,1\le i\le m\}\\ &E=\{(v_i,v_j)|v_i,v_j\in V,Y_i\cap Y_j\neq\Phi,i\neq j\} \end{aligned} \tag{4-2}$$

其权重值表示为 Wij

$$W_{ij}=\begin{cases}0, & v_i=v_j\\ \infty, & v_i\neq v_j,\ (v_i,v_j)\notin E\\ dis(i_i,v_j), & v_i\neq v_j,\ (v_i,v_j)\in E\end{cases} \tag{4-3}$$

例如：根据一个四类标签集合语料，$Y=\{y_1,y_2,y_3,y_4\}$，训练数据集中包含四个文本实例，文本类标签为表 4-1，模型计算的新加入实例与各个类标签间的权重为 4-2

表 4-1 测试集训练实例

实例序号	标签集合
1	$Y_1=\{y_1, y_2, y_4\}$
2	$Y_2=\{y_1, y_4\}$
3	$Y_3=\{y_2, y_3\}$
4	$Y_4=\{y_2, y_3, y_4\}$

万方数据

表 4-2 类标签间的权重

测试实例	1	2	3	4
S	0.95	0.84	0.73	0.68

将测试实例在样本特征项上的置信度赋予为该样本相对类别的权重，基于多类标数据集的类结果数大于等于一,相对主题类似相关，内容相近的文本在类特征表现中会呈现较高的共性与相关联系。因此提出基于类置信度赋予的属性加权调节权值的算法将会分别对训练数据集特征空间的每个特征分量进行分析,计算每个待测节点的 K 个相邻节点,得到预测的类标签集合。

4.3 WeightedLabelPower 投票预测

基于特征投票机制设计一种线性文本分类方法，运用信任机制理论分析文档类别对特征的信任关系，给出具体特征信任度的模型。以每个基础分类模型的置信度结果作为权重，计算入类标签结果评定后，相应的特征项也就因此非线性地调整了样本偏差。这个方法即将多标签问题转换为单标签多类分类问题，转换类的属性值与训练样本实例的标签集相关，基于投票机制，对所属文本进行类标签判断，如表 4-3 显示，总计大于阈值 K*0.5 的即为预测标签，此例为 K=4。

表 4-3 WeightedLabelPower 投票预测

实例序号	Y1	Y2	Y3	Y4
1	0.95	0.95	0	0.95
2	0.84	0	0	0.84
3	0	0.73	0.73	0
4	0	0.68	0.68	0.68
总计	1.79	2.35	1.41	2.47
标签预测	0	1	0	1

4.4 多标签分类算法的框架

对于训练集的样本特征进行统计后得到每个特征的权重调整，从而使特征更

万方数据

能反应其类别特性。为每个测试实例通过调整后的权重特征，找到其在训练集中相应的 K 个邻居实例，将它们与其 K 个邻居节点间的距离作为类别实例权重，通过 WeightedLabelPower 投票策略，预测出分类结果，对于总体结果进行统计性能测试,基于 Hamming Loss、Ranking Loss、Coverage、Average Precision、One-Error 的总体评价，调整邻居节点 K 的数目。

Algorithm：IWLC

Input： D:Multi-label trainset、S:test example set

output：Y:predicted label set for S

Process:

Step1：//计算每个权重的特征调整

For i=0 to n

Computer the WCi for feature

End for

Step2：//计算并选择根据欧氏距离找出样本 s 的邻居节点

$N(d_u) \Leftarrow \Phi$

For j=0 to k

$n_i \Leftarrow \underset{dj \in \Omega - N(du)}{\arg\min} SIM(d_u, d_j)$

//标准化 Sk 与样本间各距离差值

$D(s,di) = \sqrt{\sum_{i=1}^{n} (Sc_i - Xc_i)^2}$

$D_j = D(s,di)/maxD(s,di)$

End for

Step3：//找出 K 个距离最小的实例和其标签集作为点的邻居图，将距离值作为其权重值

For i=0 to |S|,j=0 to k

Get the $S_{[i][j]}$ who holds the neighbour and weight

End for

Step4：//按照 WeightedLabelPower 投票机制找出初步预测的标签集

For i=0 to |S|,j=0 to k

Get the $T_{[i][j]}$ who holds the preliminary forecasting

万方数据

End for

Step5: //对于预测出的各个样本标签集做各方法的性能测试

For i=0 to |S|,j=0 to k

$Ranking_{[i][j]}$,$Confidence_{[i][j]}$,$Truelabels_{[i][j]}$

End for

Test the Hamming Loss、Ranking Loss、Coverage、Average Precision、One-Error

Step5：//反复根据 step4 的结果做更改每个类的 K 值的 Step2-Stpe4 的迭代，找出最佳方案

4.5 本章小结

本章在介绍了常用的一些多标签分类算法后，采用的一种相关信息加权的自适应多标签分类算法，相对于现有的一些多标签分类方法在大部分性能指标上有所提高。自适应选择的过程会帮助算法在针对不同领域的的语料库有更好的效果，将经典线性回归体系扩展到多标签分类。

万方数据

第五章 实验及结果分析

本章以同济新闻语料集为实验数据，共包含有效文本 5820 篇，对上文提出的多标签文本分类特征选择算法、信息加权的自适应多标签分类算法，分别进行了实验。将实验结果与现如今流行的特征选择算法与分类算法做比较，都较为理想。

5.1 多标签文本分类数据集

本文采用的是酵母[12]、景象[13]和情感[14]英文数据集和一个来自同济大学卫志华老师提供的中文新闻文本语料库[41]。如表 5-1 所示．Yeast 数据集是基因功能分类的数据集，样本信息代表了基因在各种条件及情况下发生的变化及组成架构；Scene 数据集是由图片成像特征构成，其特征来源于图片内部的组成架构；Emotions 数据集来源于人类情感特征的分析树据，主要基于人类对于音乐变化及语言所产生的自然反应。中文文本语料库的样本是取自教育，经济，军事，科技，商务，社会，体育，娱乐，政治共九大类的中文文本新闻数据集。现实的新闻语料的多标签情况受到许多因素的影响，如人为主观影响、概念划分模糊、标签从属关系不明确等。数据本身就存在大量噪声。 此外，在多标签数据中各类样本分布很不均匀，所以要尽量选取较为平均分布的语料。

表 5-1 数据实例集描述

数据集	Emotions	Scene	Yeast	同济新闻语料库
所属领域	音乐	图像	生物	新闻
文档数	593	2407	2417	5894
特征数目	72	294	103	2344
标签数	6	14	14	9
标签势	1.869	1.074	4.237	1.197
标签密度	0.311	0.179	0.303	0.199
不同标签集数	27	15	198	125

表五中的标签势 LC(D)是指训练数据集中实例的平均标签数目，而标签密度 LD(D)指的是标签势数除以标签数。

万方数据

$$LC(D)=\frac{1}{|\Omega|}\sum_{i=1}^{|\Omega|}\sum_{j=1}^{|L|}l_{ij} \quad (2\text{-}1)$$

$$LD(D)=\frac{1}{|\Omega \| L|}\sum_{i=1}^{|\Omega|}\sum_{j=1}^{|L|}l_{ij} \quad (2\text{-}2)$$

5.2 多标签文本分类特征选择实验

5.2.1 强特征挖掘实验

经过强类别纹理挖掘算法的测试，经过 1%递增的迭代测试，使用三种不同的多标签分类方法下，采用了各测试性能加权平均的衡量标准，选取了针对此语料库纹理主次类别频率差最佳阈值Ω为 42%，得到的各个类别纹理如表 5-2：

表 5-1 数据实例集强特征纹理描述

类别	个数	纹理特征举例
经济	54	个股，牛市，基本面等
政治	13	送检，副处长，总召集人等
军事	95	美军，司令，国防部等
体育	83	比分，进球，后卫等
娱乐	58	银幕，任天堂，闪存等
科技	41	性激素，微血管，基因组等
社会	12	英才网，咨询师，面试官等
商务	7	店址， 跨零点，加航等
教育	69	升学率，题海，考点等

在上述实验数据集中，把每篇文档都表示成由特征字符串组成的序列后，通过迭代算法挖掘各个类别的强特征（字符串）集合。强特征具有的特点是：某个类的强特征覆盖该类中尽量多的文档，强特征对所有类别的区分度高。已知类别的强特征自动挖掘算法采用递归算法，逐步逼近特征的最优化状态。

万方数据

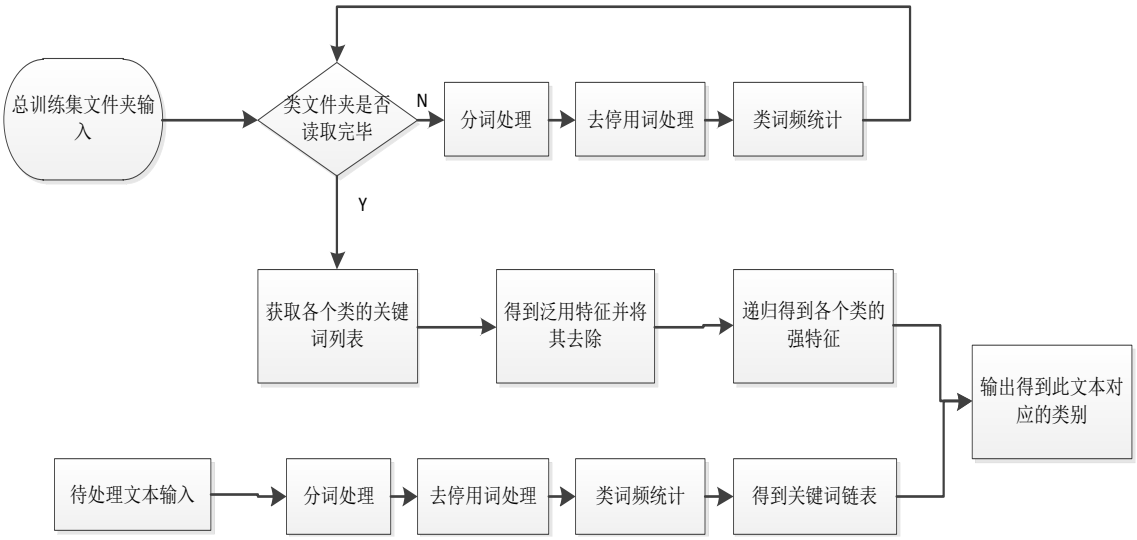

图 5-1 强特征挖掘流程图

Figure 5-1 Flow charts of Strong feature mining

5.2.2 改进的特征选择和加权抽取实验

文本向量化，输入文件名，计算文本向量，19647毫秒，约合每篇3.37毫秒。

数据库写入，输入文件名，计算文本向量后写入DB，22198毫秒，相当于写入数据库时间为2551毫秒。

篇章检索，输入文件名，读取DB并计算向量距离，36569毫秒。

分词与词性标注处理，输入待处理文本，分词并标注词性，加载词典19505毫秒，分词、标注词性37秒。

文本特征选择的主要目标是选出能够很好反映文本内容的词，及除去特征集有效信息量较低的特征项，从而改进分类精度并减少时空复杂度。在第三章介绍的目前流行的四种文本特征选择方法：文档频度（DF），信息增益（IG），χ2统计，互信息(MI)。通过大量实验表明，前三者更有高效，效果也更好。对于特征选择结果的特征加权算法，除了第三章介绍的布尔权重TFIDF权重之外，还有TFC、LTC以及信息熵权重等方法。

目前这些通常采取的特征选择和加权方法都得到了广泛的应用，在该研究领域取得了一定的认可度，然而针对本项目的特殊需求，这些方法并不能应用到本项目中。以上方法中的关键因素是对于调节因子的使用与否，对于文本分了过程来说，调节因子的重点应放在类信息对特征项的相对影响因素。

万方数据

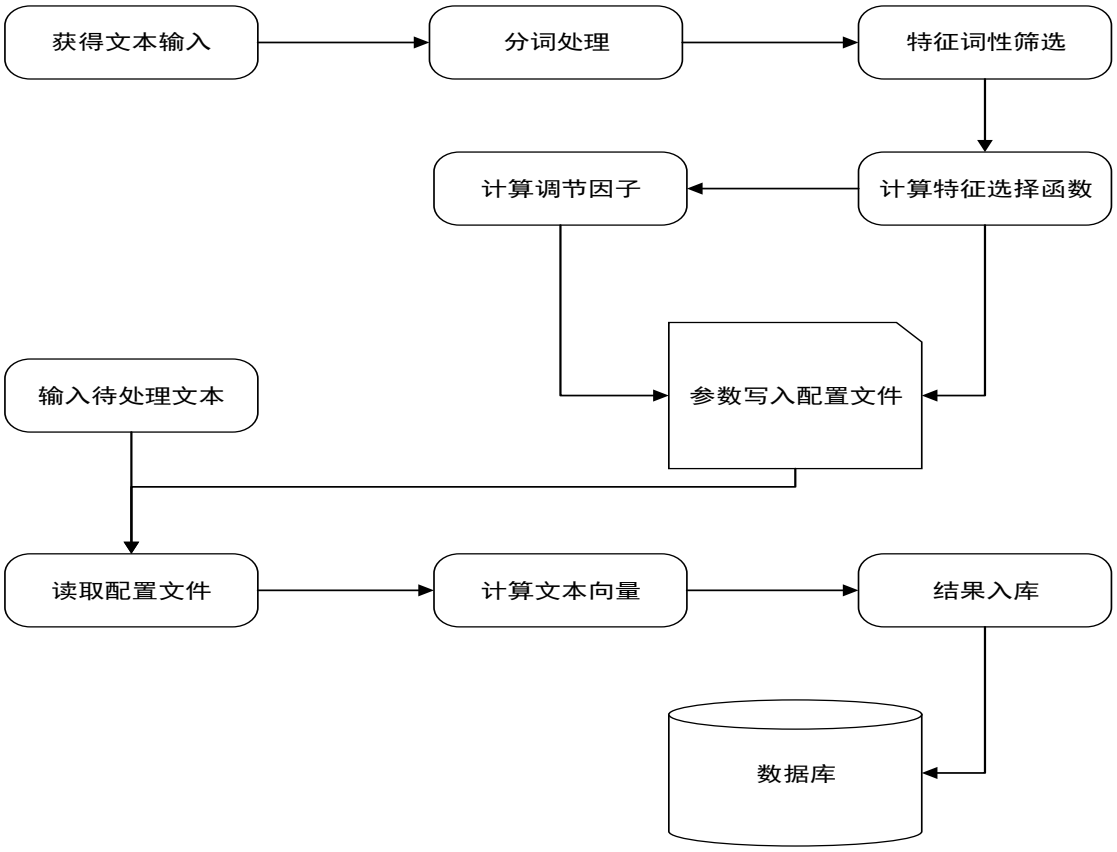

图 5-2 用于类别描述的特征选择方法流程图
Figure 5-2 Flow charts of Feature Selection in Category Description

通过文本采集，可以采集到初始语料内容为：

阳朔住宿选择在西街附近是很不错的，但是有一个问题是，不能离西街酒吧街太近，因为那样子晚上很吵，我们就是选择离酒吧街太近了，虽然出行非常方便，但是晚上会很吵，一直会吵闹到晚上 11 点左右，好在那几天白天玩得比较累，比较容易入睡，所以感觉就还好一些。如果选择阳朔西街附近住宿的话，可以选择比较有特色的客栈入住，价格不算贵，淡季的时候一个房间几十块到两百块就可以了，旺季的话一百多到两百多也可以租到比较不错的客栈了。建议大家在选择客栈的时候，多参考一下晚上大家对于那个酒店的意见评价，有一些客栈老板人非常好，会让入住的客人有种回到家里的感觉，我出发之前就在网上找寻了很久，最后看到有一个阳光客栈的评价很高，大家都说老板人非常非常

万方数据

好，我觉得出去玩心情很重要，老板有时候提供的一些建议和服务会让我们的心情变得更好，于是我就选择了一间三人房，因为只是提前一两天预定，价格要168一晚，在西街这个价格不算便宜了。好在客栈老板真的非常好，客栈附近没有停车位，一直很热情地帮我们找，虽然后来是自己去了收费停车场，但是让我对于阳朔的第一感觉非常好，到了客栈又端茶递水果，吃到了那儿的野枣，非常好吃，虽然住宿条件一般般，但是我们入住的那三天过得非常开心。

第一步，经过预处理步骤，得到如下的结果：

阳朔#NR 住宿#NN 选择#VV 在#P 西街#NR 附近#NN 是#VC 很#AD 不错#VA 的#SP ，#PU 但是#AD 有#VE 一#CD 个#M 问题#NN 是#VC ，#PU 不#AD 能#VV 离#P 西街#NR 酒吧#NN 街#NN 太#AD 近#VA ，#PU 因为#P 那#DT 样子#NN 晚上#NT 很#AD 吵#VV ， #PU 我们#PN 就#AD 是#VC 选择#VV 离#P 酒吧#NN 街#NN 太#AD 近#VA 了#SP ， #PU 虽然#CS 出行#NN 非常#AD 方便#VA ， #PU 但是#AD 晚上#NT 会#VV 很#AD 吵#VV ， #PU 一直#AD 会#VV 吵闹#VV 到#VV 晚上#NT 1 1 #CD 点#M 左右#LC ， #PU 好在#AD 那#DT 几#CD 天#M 白天#NT 玩#VV 得#DER 比较#AD 累#VA ，#PU 比较#AD 容易#AD 入睡#VV ，#PU 所以#AD 感觉#VV 就#AD 还#AD 好#VA 一些#AD 。#PU 如果#CS 选择#VV 阳朔西街#NR 附近#NN 住宿#NN 的话#SP ， #PU 可以#VV 选择#VV 比较#AD 有#VE 特色#NN 的#DEC 客栈#NN 入住#VV ，#PU 价格#NN 不#AD 算#VV 贵#VA ，#PU 淡季#NN 的#DEG 时候#NN 一#CD 个#M 房间#NN 几十#CD 块#M 到#CC 两百#CD 块#M 就#AD 可以#VV 了#AS ， #PU 旺季#NN 的#DEG 话#NN 一百多#CD 到#CC 两百多#CD 也#AD 可以#VV 租#VV 到#VV 比较#AD 不错#VA 的#DEC 客栈#NN 了#SP 。#PU 建议#VV 大家#PN 在#P 选择#VV 客栈#NN 的#DEC 时候#NN ，#PU 多#AD 参考#VV 一下#AD 晚上#NT 大家#PN 对于#P 那个#DT 酒店#NN 的#DEG 意见#NN 评价#NN ， #PU 有#VE 一些#CD 客栈#NN 老板#NN 人#NN 非常#AD 好#VA ， #PU 会#VV 让#VV 入住#VV 的#DEC 客人#NN 有#VE 种#M 回到#VV 家里#NN 的#DEC 感觉#NN ，

万方数据

#PU 我#PN 出发#VV 之前#LC 就#AD 在#P 网#NN 上#LC 找寻#VV 了#AS 很#AD 久#AD ，#PU 最后#AD 看到#VV 有#VE 一#CD 个#M 阳光#NN 客栈#NN 的#DEG 评价#NN 很#AD 高#VA ，#PU 大家#PN 都#AD 说#VV 老板#NN 人#NN 非常#AD 非常#AD 好#VA ，#PU 我#PN 觉得#VV 出去#VV 玩#VV 心情#NN 很#AD 重要#VA ,#PU 老板#NN 有时候#AD 提供#VV 的#DEC 一些#CD 建议#NN 和#CC 服务#NN 会#VV 让#VV 我们#PN 的#DEG 心情#NN 变得#VV 更#AD 好#VA ,#PU 于是#AD 我#PN 就#AD 选择#VV 了#AS 一#CD 间#M 三#CD 人#NN 房#NN ,#PU 因为#P 只是#AD 提前#VV 一#CD 两#CD 天#M 预定#VV ，#PU 价格#NN 要#VV １６８一#CD 晚#NN ，#PU 在#P 西街#NN 这个#DT 价格#NN 不#AD 算#VV 便宜#VA 了#AS 。#PU 好在#AD 客栈#NN 老板#NN 真的#AD 非常#AD 好#VA ,#PU 客栈#NN 附近#NN 没有#VE 停车位#NN ,#PU 一直#AD 很#AD 热情#VA 地#DEV 帮#VV 我们#PN 找#VV ，#PU 虽然#CS 后来#AD 是#VC 自己#PN 去#VV 了#AS 收费#NN 停车场#NN ,#PU 但是#AD 让#VV 我#PN 对于#P 阳朔#NR 的#DEG 第一#OD 感觉#NN 非常#AD 好#VA ,#PU 到#VV 了#AS 客栈#NN 又#AD 端#VV 茶递#NN 水果#NN ，#PU 吃到#VV 了#AS 那儿#PN 的#DEG 野枣#NN ，#PU 非常#AD 好吃#VA ,#PU 虽然#CS 住宿#NN 条件#NN 一般#AD 般#VA ,#PU 但是#AD 我们#PN 入住#VV 的#DEC 那#DT 三#CD 天#M 过#VV 得#DER 非常#AD 开心#VA 。#PU

第二步，根据预处理后的文本内容，进行特征选择，得到初始特征词项。后记录其在文本中出现的频率，根据选择特征项的统计常量，生成向量arff。

经过测试，输入1700篇文章进行测试，对于词典加载预处理1min40s，运行25s后出结果。

5.3 相关信息加权的自适应多标签分类实验

5.3.1 实验环境

本文实验环境为Intel（R）Xero CPU E5620@ 2.40GHz，15.9GB内存，1T硬盘的华为服务器，操作系统为Winserver2003，Java版本Sun JDK1.7.0。

5.3.2 实验数据

万方数据

（1）改进的特征选择和加权方法实验过程

文本向量化

测试内容	输入文件名，计算文本向量
测试结果	正常结束
时间消耗	19647 毫秒，约合每篇 3.37 毫秒
安全相关	未发生内存泄露

分词与词性标注处理

测试内容	输入待处理文本，分词并标注词性
测试结果	正常结束
时间消耗	加载词典 19505 毫秒，分词、标注词性 37 秒
安全相关	未发生内存泄露

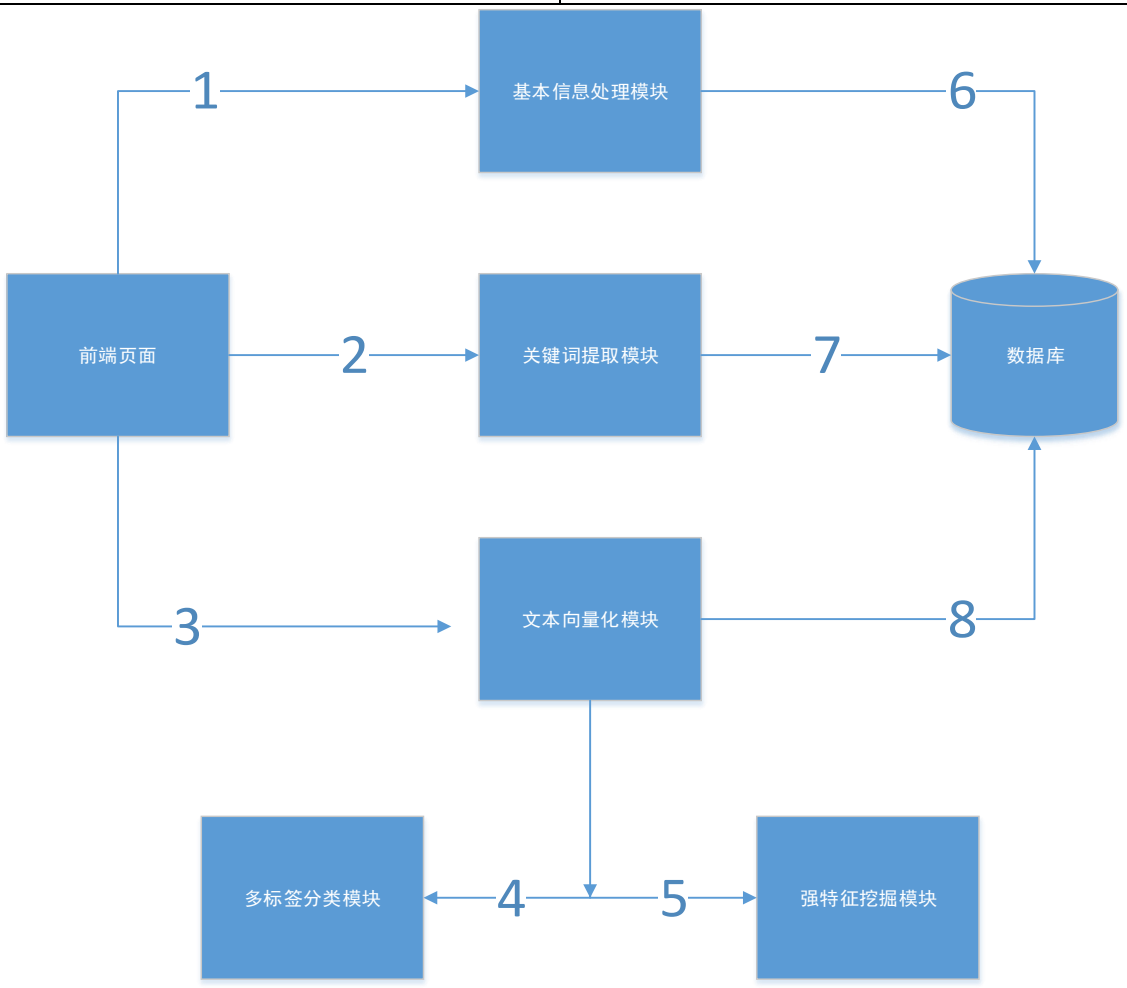

图 5-3 多标签文本分类系统流程图

Figure 5-3 flow chart of Multi label text classificationsystem

万方数据

（2）训练过程

强特征程序运行过程：

1. 斯坦福分词器初始化

2. 遍历处理每一篇文章

3.生成各个类的强特征

将强特征输出至txt文件，为强特征获取做预备工作

根据集合测试实验，强特征训练结果如表5-3所示：

表 5-2 同济新闻语料库部分强特征输出

（经济）	(政治)	（军事）	（体育）	（娱乐）	（科技）	（社会）	（商务）	(教育）
个股	送检	座舱	比分	忌食休克	性激素	羊羊	加航	升学率
股指	副总长	发射器	进球	周绮思	超声	编后语	零点	解题
牛市	罗本立	气动	客场	家畜	乳剂	投递	跨零点	做题
估值	促进法	美制	夺冠	性兴奋	颅内	英才网	店址	考点
跳空	永固	责编	中国队	食俗	XP	底薪	索菲特	题海
基本面	输血科	近程	主教练	擦去	带氧	领带	王振堂	加分
千五	外请	国防军	后卫	蔡惟	橙子	洗头妹	舒泽	招考
权证	总召集人	续航	赛前	白嫩	腰肌劳损	中华英		阅卷
增持	多收费	敌机	射门	圣何塞	外科学	舞员		望子成龙
上证	报告	舰长	主帅	狗肉	小鼠	闲聊		教育界
港股	出库单	猎鹰	半场	侵权案	体位	咨询师		招生办
盘面	记账单	战斧	赛后	甜味	蛋类	面试官		咨询会
市盈率	顾大局	JSF	半决赛	乳白色	皮质醇	踏步		助学金
券商	九二	C-130	下半场	人好	人参	隔壁		考分
加息		进气道	申花	银幕	内分泌学	大忌		教职工
			summer					
			攻门					
.....	.....	.....	.....	.....	.....	.....	.....	.....
共61	共14	共65	共117	共59	共41	共22	共7	共107

（2）强特征效果提升

强特征覆盖了某各类中尽量多的文档数目；同时，强特征也具有最强的区分类的能力。已知类别的强特征自动挖掘算法采用递归算法，逐步逼近特征的最优化状态。

强特征类别测试（与多标签类别融合）

测试多篇文章的类别，先用强特征方法测，如果有则输出此类别，否则使用

万方数据

多标签测试类别，并且进行输出。

1、多标签分类器和斯坦福分词器初始化

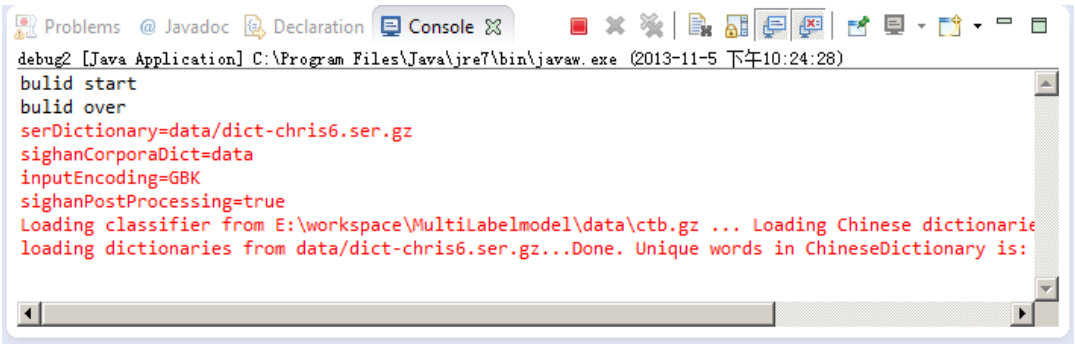

图 5-4 分词器初始化

Figure 5-4 Initiation of Word Segment

2、使用强特征和多标签逐篇文章进行类别提取。如强特征无法提取类别则输出类别 0，使用多标签进行类别提取。

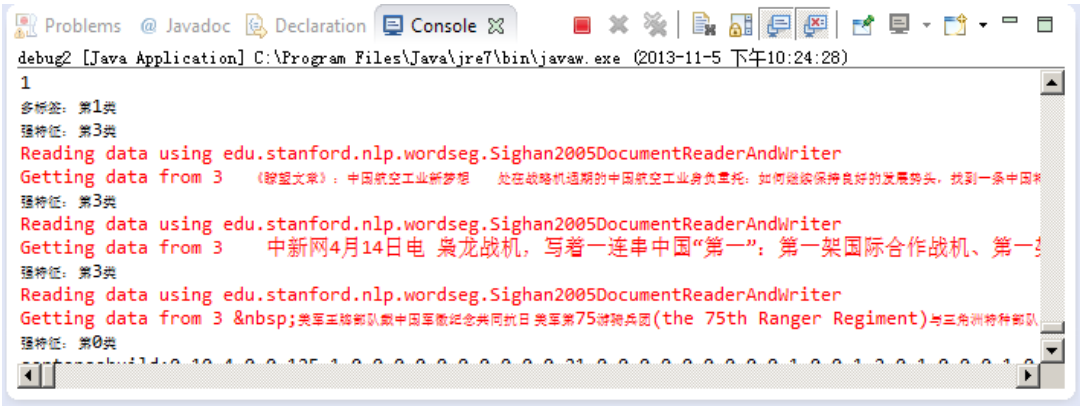

图 5-5 强特征和多标签结合一

Figure 5-5 Combinition of Strong feature and Multi-labels No.1

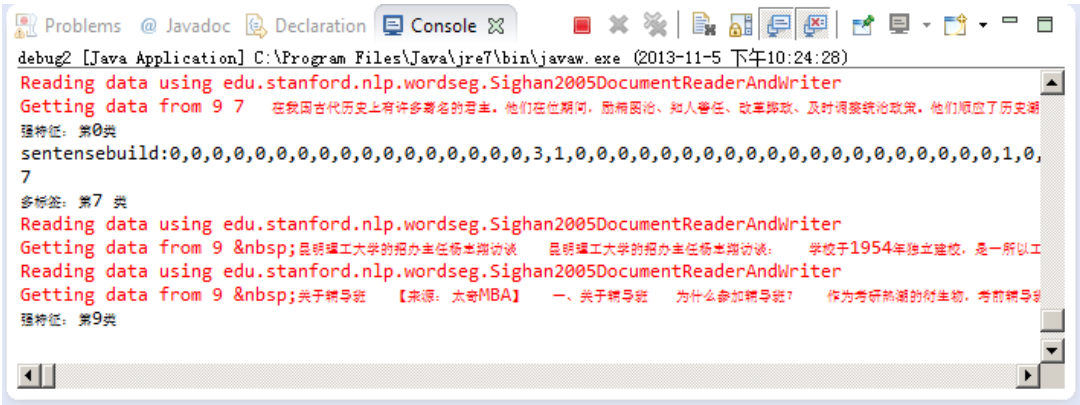

图 5-6 强特征和多标签结合二

Figure 5-6 Combinition of Strong feature and Multi-labels No.2

万方数据

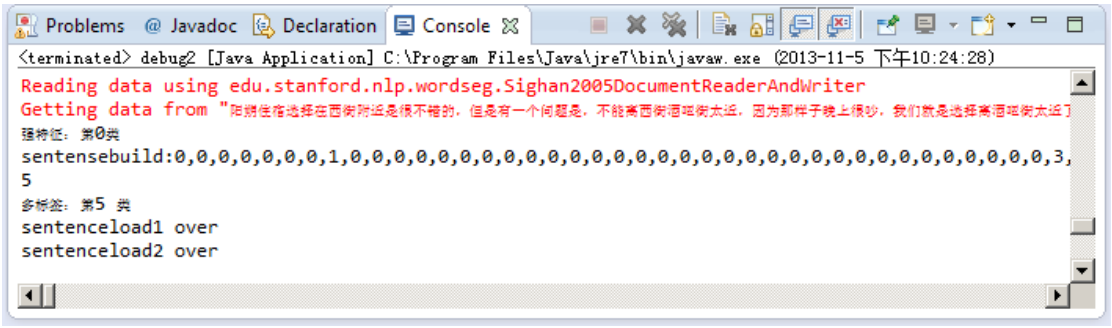

图 5-7 强特征和多标签结合三

Figure 5-7 Combinition of Strong feature and Multi-labels No.3

5.3.3 结果分析

对于相关信息加权的多标签分类算法，我们采用 10 倍交叉验证(10—fold Cross—validation)策略对四个数据集进行了仿真实验。根据自适应迭代测试，情感、景象和酵母数据集和同济新闻语料库的初始 K 值分别选定为 10、10、10、15。实验中与 IWLC 算法采用的对比算法有 MLkNN[16](Multi-Label k-nearest neighbor)、BRkNN[6](Binary Relevance k-nearest neighbor)、RAkEL[17](Random k-Labelsets)、NB[6](Naive Bayes)。在对比实验中，将原有数据集和测试集混合，随机平衡采样各类并排序。

表 5-3 Emotions 数据集性能比较

性能对比	Hamming-loss ↓	Ranking Loss ↓	Coverage ↓	Average Precision ↑	One-Error ↓
ML-kNN	0.2100±0.0175	0.2621±0.0183	2.3627±0.0832	0.7356±0.0289	0.3391±0.0630
BR-k NN	0.2049±0.0147	0.2494±0.0214	2.2919±0.1000	0.7497±0.0256	0.3188±0.0461
RAkEl	0.2555±0.0001	0.2589±0.0036	2.2125±0.0030	0.7083±0.0105	0.4469±0.0227
Naive Bayes	0.2518±0.0010	0.2208±0.0128	2.1063±0.0322	0.7580±0.0195	0.3306±0.0377
IWLC	0.1520±0.0018	0.0968±0.0102	1.4527±0.0508	0.8715±0.0206	0.1500±0.0172

万方数据

表 5-4 Scene 数据集性能比较

性能对比	Hamming-loss ↓	Ranking Loss ↓	Coverage ↓	Average Precision ↑	One-Error ↓
ML-kNN	0.0946±0.0041	0.1535±0.0102	0.8674±0.0487	0.8044±0.0148	0.2833±0.0223
BR-k NN	0.0996±0.0073	0.1856±0.0062	1.0336±0.0287	0.7819±0.0120	0.3070±0.0219
RAkEl	0.2417±0.0002	0.2141±0.0030	1.1595±0.0132	0.6530±0.0048	0.5808±0.0097
Naive Bayes	0.1601±0.0008	0.1468±0.0009	0.8214±0.0049	0.7710±0.0057	0.3781±0.0110
IWLC	0.11390±0.0015	0.0691±0.0021	0.4577±0.0019	0.8580±0.0139	0.2534±0.0114

表 5-5 Yeast 数据集性能比较

性能对比	Hamming-loss ↓	Ranking Loss ↓	Coverage ↓	Average Precision ↑	One-Error ↓
ML-kNN	0.1982±0.0020	0.3312±0.0108	9.1345±0.1949	0.6641±0.0065	0.2565±0.0049
BR-k NN	0.2077±0.0018	0.2977±0.0055	8.5229±0.0436	0.6831±0.0020	0.2681±0.0032
RAkEl	0.2857±0.0001	0.3278±0.0060	8.8445±0.1903	0.6093±0.0044	0.4390±0.0052
Naive Bayes	0.2682±0.0002	0.2652±0.0027	7.9243±0.1171	0.6698±0.0003	0.3471±0.0135
IWLC	0.1556±0.0013	0.1033±0.0005	5.7339±0,2012	0.8393±0.0041	0.1266±0.0178

表 5-6 同济新闻数据集性能比较

性能对比	Hamming-loss ↓	Ranking Loss ↓	Coverage ↓	Average Precision ↑	One-Error ↓
ML-kNN	0.1509±0.0047	0.2900±0.0067	3.6860±0.0884	0.6186±0.0133	0.4508±0.0297
BR-k NN	0.1605±0.0082	0.2596±0.0156	3.4267±0.1257	0.6502±0.0203	0.4141±0.0351

万方数据

RAkEl	0.1712±0.0053	0.1612±0.0013	2.4136±0.0299	0.7049±0.0017	0.4440±0.0090
Naive Bayes	0.1810±0.0107	0.1218±0.0097	2.0604±0.1132	0.7856±0.0127	0.2984±0.0181
IWLC	0.1796±0.0075	0.1277±0.0014	2.3352±0.0917	0.8344±0.0121	0.1591±0.0120

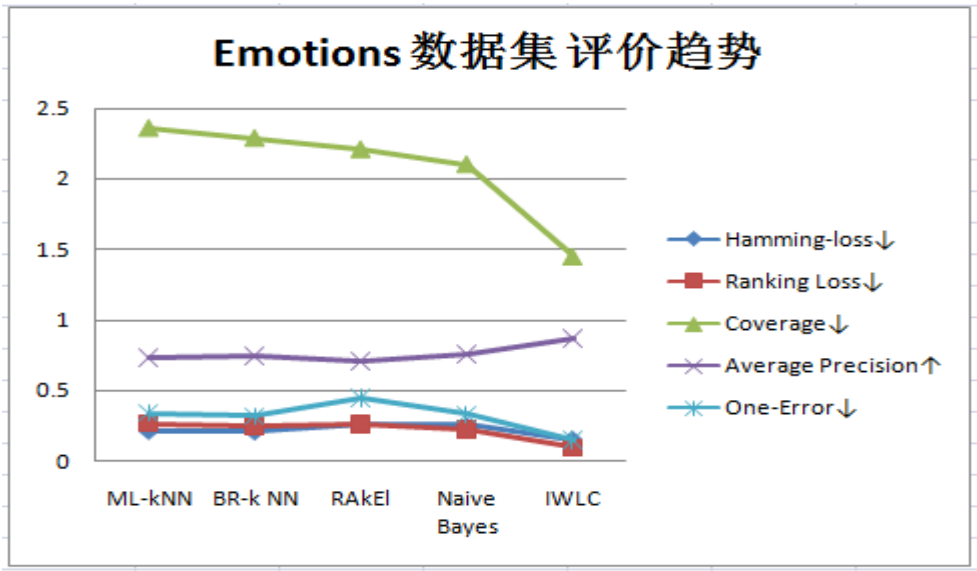

图 5-8 Emotions 数据集评价指标图表趋势

Figure 5-8 Emotions Data set evaluation index trend chart

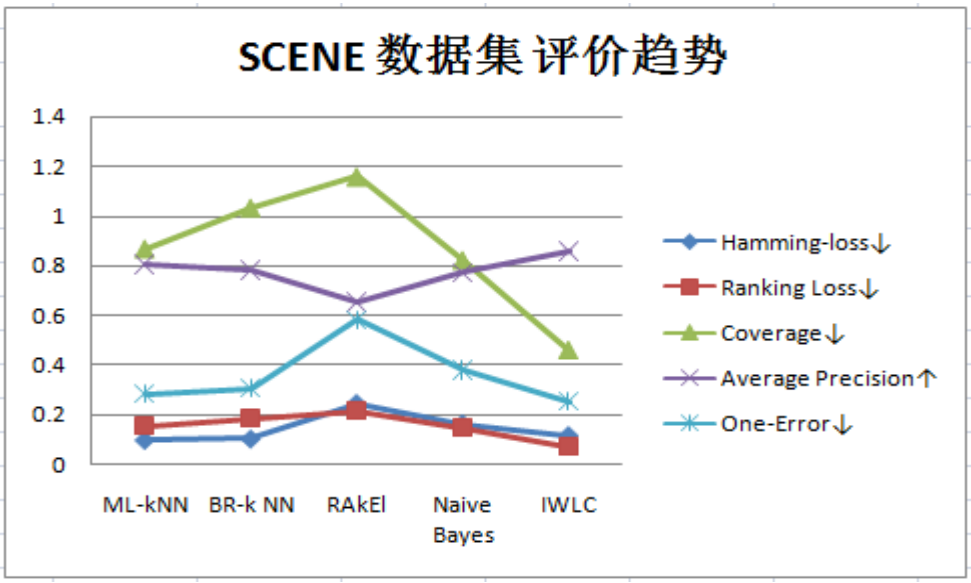

图 5-9 Scene 数据集评价指标图表趋势

Figure 5-9 Scene Data set evaluation index trend chart

万方数据

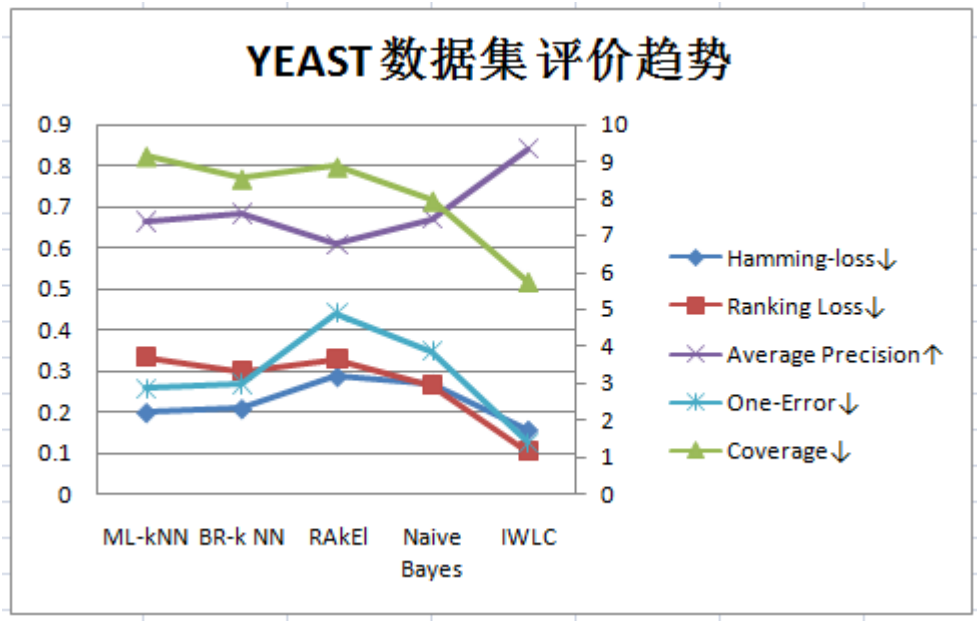

图 5-10 Yeast 数据集评价指标图表趋势

Figure 5-10 Yeast Data set evaluation index trend chart

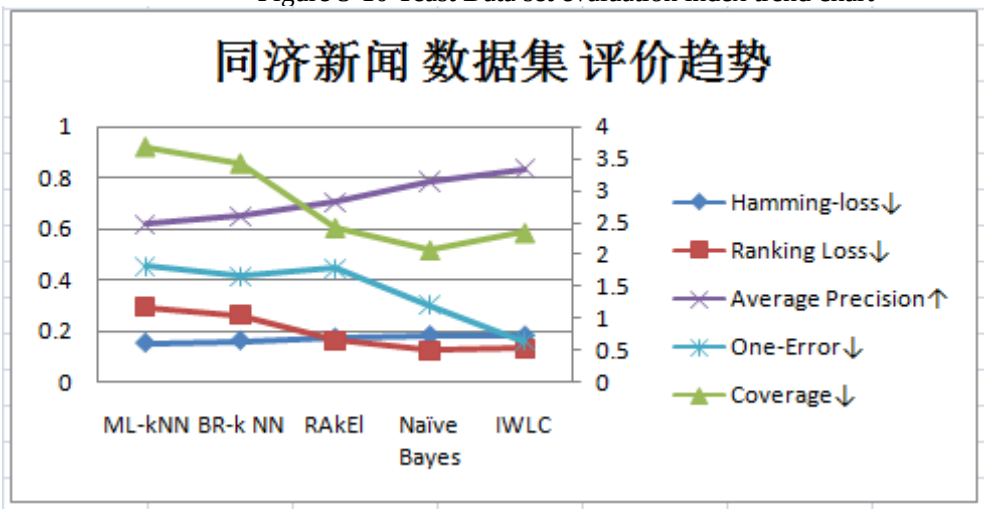

图 5-11 同济新闻 数据集评价指标图表趋势

Figure 5-11 Tongji news Data set evaluation index trend chart

从表 6 到表 9 中的 5 个评价指标中可以看出，在 Emotions 和 Scene 以及 yeast 的大部分指标上 IWLC 均好于 Naive Bayes 方法，说明了在小数据集分类方面 IWLC 有着明显的分类准确提高；但在 Scene 上的 Hamming-loss 分类效果上略逊于 ML-kNN 和 BR-kNN，这很有可能是因为其语料为图像数据且特征选取代表性不均衡且标签势太小。在非海量数据实例集上测试，IWLC 的分类效率都显著高于其余各种测试方法。在对同济大学提供的大信息新闻语料库的测试中，由于数据分布的

万方数据

复杂性和分类算法达到效果的侧重点不同，ML-kNN 和在 Hamming-loss 分类效果上稍优于 IWLC，但其他方面 IWLC 方法例如 Rankingloss、One-Error 尤为突出，在 Coverage，Average Percision 上也优于其他方法，综合比较还是一种较为可行且有效的多标签分类算法。

5.4 本章小结

本章分别对上文提出算法进行实验。首先，以同济大学新闻语料库作为研究对象，对强特征文本纹理挖掘进行实验提取，对语料库进行改进的特征选择和加权抽取，实验数据结果可以看出，改进的中文多标签选择算法相比于传统的多标签分类算法，具有更高的效率及更好的效果，在多标签特定判别指标上可以保证较低的漏检率及误检率。而后，本章对相关信息加权的自适应多标签分类算法进行了实验，实验结果在四个不同类别的语料库上分别进行了测试，与现有的知名多标签算法诸如 ML-KNN，Naive-Bayes，BR-KNN，RAKEL 等做比较，表明新算法在某些多标签分类准确度评估指标上有所提高。

万方数据

第六章 总结与展望

6.1 本文工作总结

本文首先研究了中文多标签分类领域的相关技术与系统整体框架图，并分别给出了其中的多标签特征选择模型及多标签分类模型。介绍了相应的现有多标签分类方法以及其优势与缺陷。对于多标签文本分类相关技术，也详细描述了分类结果的测试性能指标。

而后，首先研究了中文多标签的特征选择，发现强特征纹理在文本分类领域不仅能够快速节省复杂算法带来的消耗，而且可以得到精度高，第一类别强度的分类信息，因此也可以相应得出结论，部分文本可以通过少数几个信息量明显且巨大的强特征纹理而判定分类类别。本文还采取了改进的特征选择和加权抽取，较之于现有的特征选择算法有更高的类别区分度。

对中文多标签分类，本文详细描述了一种相关信息加权的自适应多标签分类算法, 该算法具有相关信息加权、自适应阈值调整、权重投票相结合的特点，其中在前期的预处理中包括无意义文本过滤、特征分词、强特征选择、改进的特征权重抽取，该算法可以较为准确、高效找到文本分类结果，并将文本分类至正确的类别簇。本算法结合了问题转化和多标签算法改进的思想，提出的是一种在各类特征选择基准调整后，基于已有单标签分类结果进行加权、自适应阈值设定，不同权重投票相结合的方法，对待分类实例进行多标签分类，能提高多标签文本分类的分类准确度与精度。自适应选择的过程也会帮助算法在针对不同领域的的语料库有更好的效果，将经典线性回归体系扩展到多标签分类。

第五章的多个实验表明，本文提出的中文多标签文本分类模型有较好的准确度以及可行性，自适应调整也是机器自学习不断提高其分类准确效果。

6.2 研究展望

未来的研究中本文将进一步的改进并优化以下几点：

强特征纹理选择优化

本文目前只能对整个成型语料库进行强特征纹理判定，语料库间信息尚未通用识别，据中文文本研究量的增加，可以挖掘语料间的共同性，从而对未来的强

万方数据

特征挖掘产生的学习效果提供成型的快速挖掘提取依据，对于包含偏离文本点的语料库强特征也可识别调整。

文本分类算法优化

本文的转发链关键点预测实验中只选择了同济中文新闻多标签语料库，后续研究可以在赝本语料库丰富的同时可以进一步扩大中文实验数据的范围，增强分类算法的适用性以及预测实验的精准度。在不同领域分类测试中逐步得到各分类领域的特性以及相应最佳阈值信息。

万方数据

参 考 文 献

[1] Yang Y, Pedersen J O. A comparative study on feature selection in text categorization[C]//ICML. 1997, 97: 412-420.

[2] Weston J, Watkins C. Multi-class support vector machines[R]. Technical Report CSD-TR-98-04, Department of Computer Science, Royal Holloway, University of London, May, 1998.

[3] Mache J and Apon A. Deep Classification in Large-scale Text Hierarchies. In: Proc. of ACM SIGIR conference on Research and development in information retrieval. ACM, 2008: 619-626.

[4] 崔晓源. 词间语义关系的研究及其在文本分类中的应用[D].天津大学 , 2006.

[5] Sebastiani F. Machine learning in automated text categorization[J]. ACM computing surveys (CSUR) , 2002, 34(1): 1-47.

[6] 苏金树, 张博锋, 徐昕. 基于机器学习的文本分类技术研究进展 [J][J]. 软件学报, 2006, 17(9): 1848-1859.

[7] Rogati M, Yang Y. High-performing feature selection for text classification[C]//Proceedings of the eleventh international conference on Information and knowledge management. ACM, 2002: 659-661.

[8] 宋枫溪, 高秀梅, 刘树海, 等. 统计模式识别中的维数削减与低损降维[J]. 计算机学报, 2005, 28(11): 1915-1922.

[9] Lan M, Tan C L, Su J, et al. Supervised and traditional term weighting methods for automatic text categorization[J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2009, 31(4): 721-735.

万方数据

[10]Guo G, Wang H, Bell D, et al. KNN model-based approach in classification[M]//On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE. Springer Berlin Heidelberg, 2003: 986-996.

[11]Joachims T. Text categorization with support vector machines: Learning with many relevant features[M]. Springer Berlin Heidelberg, 1998.

[12] Kumar M A, Gopal M. A comparison study on multiple binary-class SVM methods for unilabel text categorization[J]. Pattern Recognition Letters, 2010, 31(11): 1437-1444.

[13] 梁英毅. 集成学习综述 [EB/OL][D]. , 2006.

[14] Erenel Z, Altınçay H. Improving the precision-recall trade-off in undersampling-based binary text categorization using unanimity rule[J]. Neural Computing and Applications, 2013: 1-18.

[15] Stevens J, Williams M. uBoost: A boosting method for producing uniform selection efficiencies from multivariate classifiers[J]. arXiv preprint arXiv:1305.7248, 2013.

[16] Yoo H Y, Park N W, Hong S, et al. Feature Extraction and Classification of Multi-temporal SAR Data Using 3DWavelet Transform[J]. Korean Journal of Remote Sensing, 2013, 29(5).

[17] Reynolds C F, O'Hara R. DSM-5 Sleep-Wake Disorders Classification: Overview for Use in Clinical Practice[J]. American Journal of Psychiatry, 2013, 170(10): 1099-1101.

[18] Elisseeff A, Weston J. A kernel method for multi-labelled classification[C]//Advances in neural information processing systems. 2001: 681-687.

[19] Jiang L, Cai Z, Zhang H, et al. Naive Bayes text classifiers: a locally weighted learning approach[J]. Journal of Experimental & Theoretical Artificial Intelligence, 2013, 25(2): 273-286.

万方数据

[20] Szatmary B, Izhikevich E M. Tag-based apparatus and methods for neural networks: U.S. Patent 20,130,073,496[P]. 2013-3-21.

[21] Fürber C, Hepp M. Using Semantic Web Technologies for Data Quality Management[M]//Handbook of Data Quality. Springer Berlin Heidelberg, 2013: 141-161.

[22]Liu D, Tu B, Qian H, et al. Large-Scale Hierarchical Classification via Stochastic Perceptron[J]. 2013.

[23]Ju Q, Moschitti A, Johansson R. Learning to rank from structures in hierarchical text classification[M]//Advances in Information Retrieval. Springer Berlin Heidelberg, 2013: 183-194.

[24] Taglino F, Taglialatela A. A semantic platform to support software artifacts reuse[J].

[25] Lee Y, Chang Y, Kim N, et al. Corrigendum to “Fast and efficient lung disease classification using hierarchical one-against-all support vector machine and cost-sensitive feature selection”[Comput. Biol. Med. 42 (2012) 1157–1164][J]. Computers in Biology and Medicine, 2013.

[26] Fraternali F, Rofouei M, Alshurafa N, et al. Opportunistic hierarchical classification for power optimization in wearable movement monitoring systems[C]//SIES. 2012: 102-111.

[27] Dabirmoghaddam A, Garcia-Luna-Aceves J J. Opportunistic walks on Random Geometric Networks and their application in scalability analysis[C]//Sensor, Mesh and Ad Hoc Communications and Networks (SECON), 2013 10th Annual IEEE Communications Society Conference on. IEEE, 2013: 559-567.

[28] McCreadie R, Macdonald C, Ounis I, et al. An examination of content farms in web search using crowdsourcing[C]//Proceedings of the 21st

万方数据

ACM international conference on Information and knowledge management. ACM, 2012: 2551-2554.

[29] Kristanto W, van Ooijen P M A, Jansen-van der Weide M C, et al. A systematic review and hierarchical classification of HU-based atherosclerotic plaque characterization criteria[J]. Visualization, Classification and Quantification of Coronary Atherosclerotic Plaque using CT Soft-and Hardware Phantom Models, 2012: 49.

[30] Shen D, Ruvini J D, Mukherjee R, et al. A study of smoothing algorithms for item categorization on e-commerce sites[J]. Neurocomputing, 2012, 92: 54-60.

[31]Sun S, Guo Q, Dong F, et al. On-line boosting based real-time tracking with efficient HOG[C]//Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013: 2297-2301.

[32]Xia H, Wu P, Hoi S C H. Online multi-modal distance learning for scalable multimedia retrieval[C]//Proceedings of the sixth ACM international conference on Web search and data mining. ACM, 2013: 455-464.

[33]Yang T, Wu L, Bonissone P P. A Directed Inference Approach towards Multi-class Multi-model Fusion[M]//Multiple Classifier Systems. Springer Berlin Heidelberg, 2013: 352-363.

[34]Di C, Chan G K, Liang K Y. Supremum test statistics for a semi-parametric mixture case-control model[J]. 2012.

[35]Xia X, Yang X, Li S, et al. A Bayesian Network Nearest K-labels Method for Multi-label Classification[J]. Advances in Information Sciences & Service Sciences, 2012, 4(8).

[36]Gjorgjevikj D, Madjarov G, D ž eroski S. Hybrid Decision Tree

万方数据

Architecture utilizing Local SVMs for Efficient Multi-Label Learning[J]. International Journal of Pattern Recognition and Artificial Intelligence, 2013.

[37]Cissé M, Artières T, Gallinari P. Learning compact class codes for fast inference in large multi class classification[M]//Machine Learning and Knowledge Discovery in Databases. Springer Berlin Heidelberg, 2012: 506-520.

[38]Luo Y, Tao D, Xu C, et al. Vector-Valued Multi-View Semi-Supervised Learning for Multi-Label Image Classification[C]//Twenty-Seventh AAAI Conference on Artificial Intelligence. 2013.

[39]姜远, 余俏俏, 黎铭, 等. 一种直推式多标记文档分类方法[J]. 计算机研究与发展, 2008, 45(11): 1817-1823.

[40]Ho K I J, Leung C S, Sum J. Convergence and objective functions of some fault/noise-injection-based online learning algorithms for RBF networks[J]. Neural Networks, IEEE Transactions on, 2010, 21(6): 938-947.

[41]苗夺谦, 卫志华. 中文文本信息处理的原理与应用[M]. 清华大学出版社, 2007.

万方数据

致 谢

两年半的研究生生活即将画上一个句号。在这两年半的时间内，我收获了很多，在这里，我要感谢所有给我带来过指导与帮助的老师以及同学们。

首先，我要感谢我的导师宦飞老师与刘功申老师。两位老师严肃的科学态度，严谨的治学精神，精益求精的工作作风，深深地感染和激励着我。同时，在对我的学习与工作能力的培养上帮助很大，更锻炼了我提出问题、分析问题以及解决问题的能力，相信这些都将为我今后的求学以及工作道路带来深远的影响。

另外，我要感谢同组的许阳、张昊、许歆艺、丁霄云、胡琼同学，以及更多我无法逐一列出名字的朋友，我们经常就学习问题、科研问题进行讨论，在整个研究生阶段，我们互相帮助，在学习中都取得了不错的成绩。

我还要感谢两年半来所有的老师们对我的专业培养，我的研究课题离不开这两年半所学到的各种理论知识，衷心地感谢上海交通大学信息安全工程学院的全体老师。

我更要感谢我所有的同学以及朋友们，两年半来，我们和睦相处，在学习上，他们给我带来了很多的帮助，在生活上，他们给我带来了很多的快乐。在我完成毕业论文的过程中，他们也给予了我很多的指导以及支持。

最后，我要感谢我的父母，感谢你们为我所付出的一切！没有你们，就没有今天的我！我爱你们！多年辛苦的的培养和教育，才能让我的人生如此的精彩和完美，让我在漫长的人生旅途中使心灵有了虔敬的归依，而且也为我能够顺利的完成毕业论文提供了巨大的支持与帮助。在未来的日子里，我会更加努力的学习和工作，不辜负父母对我的殷殷期望！

在这即将毕业之际，我要感谢所有指导、帮助、支持过我的人，也祝你们能一路平安、健康，一帆风顺!

万方数据

攻读硕士期间发表的论文以及专利

已发表论文：

[1] 第一作者，相关信息加权的多标签分类算法，计算机软件与应用，已录用

[2] 第一作者，基于中文多标签文本分类的特征选择，信息安全与技术，已录用

参与专利项目：

[1]面向网络舆情的定领域情感分析研究，国家自然基金项目，课题编号：61272441

[2]海量网络舆情信息获取、分析及表达关键技术研究，国家自然基金项目，课题编号：61171173

万方数据

附件五

上海交通大学

学位论文版权使用授权书

本学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权上海交通大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

本学位论文属于

保密□，在___年解密后适用本授权书。

不保密☑。

（请在以上方框内打“√”）

学位论文作者签名：周浩　　指导教师签名：宣丞

日期：2014年 1 月 2日　　日期：2014年 1 月 2日

万方数据

附件四

上海交通大学

学位论文原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

学位论文作者签名：

日期：2013 年 12 月 25 日

万方数据