南京大学

研 究 生 毕 业 论 文

（申请硕士学位）

论　文　题　目　基于层次多标签学习的法律适用自动识别

作　者　姓　名　王景峰

学科、专业方向　计算机技术

指　导　教　师　柏文阳 副教授

研　究　方　向　数据挖掘

2016 年 5 月

学　　号 ： MF1333044

论文答辩日期 ： 2016 年 6 月 1 日

指 导 教 师 ：　　（签字）

Automatic Legal Application Recognition Based on Hierarchical Multi-label Learning

by

Jingfeng Wang

Directed by

Associate Professor Wenyang Bai

Department of Computer Science and Technology
Nanjing University

May 2016

Submitted in partial fulfilment of the requirements
for the degree of Master in Computer Technology

南京大学研究生毕业论文中文摘要首页用纸

毕业论文题目：基于层次多标签学习的法律适用自动识别

计算机技术 专业 2013 级硕士生姓名：王景峰

指导教师（姓名、职称）：柏文阳 副教授

摘 要

本文是南京大学学位论文的 LaTeX 模板。目前不支持本科生学位论文格式。

除了介绍 LaTeX 文档类 NJUthesis 的用法外，本文还是一个简要的学位论文写作指南。

关键词： 南京大学; 学位论文; LaTeX 模板

南京大学研究生毕业论文英文摘要首页用纸

THESIS: Automatic Legal Application Recognition Based on Hierarchical Multi-label Learning

SPECIALIZATION: Computer Technology

POSTGRADUATE: Jingfeng Wang

MENTOR: Associate Professor Wenyang Bai

Abstract

This paper is a thesis template of Nanjing University. Besides that the usage of the LaTeX document class `NJUthesis`, a brief guideline for writing the thesis is also included.

Keywords: Nanjing University (NJU), Thesis, LaTeX Template

目　录

目录 ····· iii

第一章 绪论 ····· 1
1.1 研究背景及意义 ····· 1
1.2 研究内容及目标 ····· 2
1.3 论文组织 ····· 3

第二章 相关技术 ····· 5
2.1 文本挖掘 ····· 5
2.1.1 向量空间模型 ····· 6
2.1.2 文本特征选择 ····· 7
2.1.3 特征权重计算 ····· 8
2.2 层次多标签学习 ····· 10
2.2.1 多标签学习 ····· 10
2.2.2 层次多标签学习 ····· 10
2.2.3 层次多标签学习方法 ····· 10

第三章 LocalBalance算法 ····· 11
3.1 算法描述 ····· 11

第四章 案件适用法律自动识别 ····· 12
4.1 实验数据 ····· 12
4.1.1 数据获取 ····· 12
4.1.2 数据清洗 ····· 13
4.1.3 标签提取 ····· 13
4.1.4 特征提取 ····· 13
4.2 实验平台 ····· 13
4.3 评价指标 ····· 14
4.4 实验结果 ····· 14

第五章　总结与展望 **15**
5.1　工作总结 15
5.2　改进方向 15

参考文献 **16**

致谢 **17**

表　　格

插　　图

1.1　法律条文树形结构示例 …… 3

4.1　裁判文书样例 …… 12

第一章　绪论

1.1　研究背景及意义

随着我国法治建设的逐步推进，人民的法律意识日渐提高，人们在遇到争议事件时会更多地选择诉诸法律，以公平公正地解决问题。根据最高人民法院的数据，2015年全国各级法院审结一审民事案件达622.8万件。然而，由于法律的专业性和复杂性，普通民众自身在借助法律维护自身权益的时候往往无所适从，只能求助律师等专业人士；另一方面，法律条文浩如烟海，即便是专业律师也只能专注于某一领域，在面对不熟悉的法律条文或者案例时，也需要一些决策辅助。

信息技术，尤其是信息检索和数据挖掘技术的发展，为法律辅助系统的实现提供了可能。“北大法宝”、“找法网”等一批在线法律信息平台，提供了法规案例检索、律师推荐等功能，在一定程度上为人们诉诸法律解决争端提供了便利。然而，上述平台提供的服务并未直接解决人们的问题。法规案例的检索往往需要用户有明确的搜索目标，甚至需要一定的法律领域知识，而且即便搜索引擎能够给出相应的搜索结果，这些结果通常也无法直接解决用户的问题，需要用户自己的分析和理解。律师推荐能够方便用户找到合适的律师，实际上是连接用户和律师的桥梁，不仅无法提供问题的直接解决方案，还容易受商业化的影响，出现一些律师滥竽充数的情况。

随着我国司法公开改革的推进以及最高人民法院关于人民法院在互联网公布裁判文书的规定的实施，蕴藏了海量信息的裁判文书可以方便地被获取和分析。2014年以来，全国各级法院共在“中国裁判文书网”上传裁判文书六百余万份，最高人民法院和部分省市区法院实现了能够上网的生效裁判文书全部上网的目标。裁判文书记载了人民法院审理案件的过程和结果，是诉讼活动结果的载体，也是人民法院确定和分配当事人实体权利义务的惟一凭证。一份结构完整、要素齐全、逻辑严谨的裁判文书，既是当事人享有权利和负担义务的凭证，也是上级人民法院监督下级人民法院民事审判活动的重要依据。因此，裁判文书中包含的当事人诉求、犯罪行为、行政执法、司法裁判行为和过程、法律的适用等信息，作为重要的历史数据，通过数据挖掘手段进行分析，可以为司法人员、律师和普通民众提供必要的决策支持。[1]实现了一个裁判文书推荐系统，为法官提供与当前裁判文书相似的文书，作为裁判的参考。基于自然语言处理技术提取文书的语义信息，在裁判文书的相似度计算上取得了不错的效

果。裁判文书推荐可以提供决策辅助，但是逐条查阅相似案例需要耗费大量精力，同时由于案例相似度不同，需要用户自行确定各个案例的权重进行综合评判，极大地降低了查询结果的直观性和明确性。

从本质上讲，裁判是法院依照法律，对案件做出决定的过程。“以事实为根据，以法律为准绳”是我国社会主义法律适用遵循的基本原则，司法机关处理一切案件，都是根据客观事实，以国家法律为标准和尺度。因此，根据案件的描述确定适用的法律，是法院判决过程的核心部分，也是律师和普通民众在法律活动中需要解决的首要问题。运用信息技术，根据案件事实描述实现适用法律的自动识别，将在很大程度上为人们的法律活动提供更加直接和明确的帮助。现已公开的裁判文书中包含的案件事实描述以及法律适用信息，为我们提供了大量的带标签数据集，采取合适的数据挖掘手段，可以从中学习得到有效的预测模型，实现对未判案件适用法律的自动识别。

1.2　研究内容及目标

本文希望通过运用数据挖掘方法，从海量的裁判文书中，学习出由案件事实描述到适用法律的预测模型，从而为用户提供直接的法律决策辅助。

一份结构完整的裁判文书包括首部、事实、理由、裁判结果和尾部五个部分，其中首部包括裁判文书的类型、编号、裁判法院，案件当事人、委托代理人等信息，事实部分包含了对案件事实的文字描述，理由部分阐述了法院对于案件的分析以及做出相应裁判结果的理由，裁判结果部分给出了法院对于此次诉讼的判决或裁定结果，尾部则包含了裁判人员、时间等信息。

运用数据挖掘手段进行案件适用法律的自动识别，首先需要提取能够充分描述案件事实的特征，由于裁判文书及其中的事实描述部分主要是以文本形式存在，因此需要运用到文本挖掘技术[2]对裁判文书进行处理，将其转换为结构化数据，包括中文分词，文本表示，特征选择和特征权重计算等。

本文通过监督式学习方法来构建预测模型，样本的标签即为案件适用的法律条文，包含在裁判文书中的裁判理由部分。由于裁判文书格式的不规范性，裁判文书引用法律条文的格式没有统一格式，因此需要对提取的案件适用的法律条文作进一步处理，形成数据集的标签。与传统的分类问题不同的是，一份裁判文书中往往包含多个法律条文的引用，因此法律适用的自动识别问题是一个多标签分类问题[3]。在多标签学习中，每个样本可以对应多个标签，使得学习问题更加复杂。更进一步地，法律条文的组织呈现为树状结构，如图1.1所示。一个案件不仅可能适用多项法律条文，这些法律条文的具体程度也可能不同，即案件适用的法律条文可能位于树结构的叶节点，也可能位于树结构的内

部节点。如果忽略法律条文的树形结构特征，无疑会损失分类标签的重要信息[4]，造成预测模型性能的下降。因此，如何利用标签的结构信息，是本文的重要研究内容。本质上，法律适用自动识别问题是一个层次多标签学习问题[5]，其中样本的特征需要通过文本挖掘手段从文本中提取，而标签的结构呈树形。

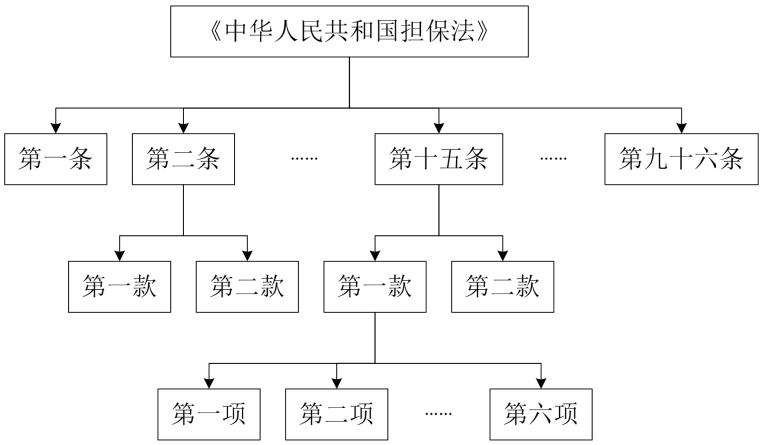

图 1.1: 法律条文树形结构示例

综上，本文研究的目标是解决案件适用法律的自动识别问题，研究的方法是首先利用文本挖掘技术对海量裁判文书进行处理分析，得到案件事实的结构化表示，即样本特征，数据集中的每个样本可以对应于标签空间中的多个标签，标签空间以树状结构组织。在此数据集上通过层次多标签学习构建预测模型，实现对未判案件适用法律的识别。

1.3　论文组织

本文组织如下：

第一章介绍本文研究的背景和意义，阐述法律适用自动识别在当前民众法律活动中的重要辅助作用，并提出研究内容和目标，指出面临的问题及解决方向。

第二章为相关技术部分，主要对本文研究涉及的文本挖掘技术、层次多标签学习技术和非平衡分类问题进行介绍，主要对已有的层次多标签学习的工作进行综述。

第三章详述本文提出的LocalBalance层次多标签学习算法。

第四章为实验部分，包括本文数据集的构建，即利用文本挖掘技术对海量裁判文书进行处理，构建结构化数据集的过程，以及运用本文算法训练得到的

模型的性能分析。

第五章对本文工作进行总结，并提出改进的方向。

第二章　相关技术

本文工作涉及到文本挖掘、层次多标签学习等数据挖掘技术。

2.1　文本挖掘

文本挖掘是一门交叉性学科，涉及数据挖掘、机器学习、模式识别、统计学、计算机语言学、信息学等多个领域。文本挖掘就是从大量的文档中发现隐含知识和模式的一种方法和工具，它从数据挖掘发展而来，但与传统的数据挖掘又有许多不同。文本挖掘的对象是海量、异构、分布的文档，文档内容是人类使用的自然语言，缺乏计算机可理解的语义。传统数据挖掘所处理的数据是结构化的，而文档都是半结构或非结构化的。所以，文本挖掘面临的首要问题是如何在计算机中合理地表示文本，使之既要包含足够的信息以反映文本的特征，又不至于过于复杂使学习算法无法处理。在浩如烟海的网络信息中，80%的信息是以文本的形式存放的。

文本的表示及其特征项的选取是文本挖掘、信息检索的一个基本问题，它把从文本中抽取出的特征词进行量化来表示文本信息。将它们从一个无结构的原始文本转化为结构化的计算机可以识别处理的信息，即对文本进行科学的抽象，建立它的数学模型，用以描述和代替文本。使计算机能够通过对这种模型的计算和操作来实现对文本的识别。由于文本是非结构化的数据，要想从大量的文本中挖掘有用的信息就必须首先将文本转化为可处理的结构化形式。目前人们通常采用向量空间模型来描述文本向量，但是如果直接用分词算法和词频统计方法得到的特征项来表示文本向量中的各个维，那么这个向量的维度将是非常的大。这种未经处理的文本矢量不仅给后续工作带来巨大的计算开销，使整个处理过程的效率非常低下，而且会损害分类、聚类算法的精确性，从而使所得到的结果很难令人满意。因此，必须对文本向量做进一步净化处理，在保证原文含义的基础上，找出对文本特征类别最具代表性的文本特征。为了解决这个问题，最有效的办法就是通过特征选择来降维。

目前有关文本表示的研究主要集中于文本表示模型的选择和特征词选择算法的选取上。用于表示文本的基本单位通常称为文本的特征或特征项。特征项必须具备一定的特性：

- 特征项要能够确实标识文本内容

- 特征项具有将目标文本与其他文本相区分的能力
- 特征项的个数不能太多
- 特征项分离要比较容易实现

在中文文本中可以采用字、词或短语作为表示文本的特征项。相比较而言，词比字具有更强的表达能力，而词和短语相比，词的切分难度比短语的切分难度小得多。因此，目前大多数中文文本分类系统都采用词作为特征项，称作特征词。这些特征词作为文档的中间表示形式，用来实现文档与文档、文档与用户目标之间的相似度计算。如果把所有的词都作为特征项，那么特征向量的维数将过于巨大，从而导致计算量太大，在这样的情况下，要完成文本分类几乎是不可能的。特征选择的主要功能是在不损伤文本核心信息的情况下尽量减少要处理的单词数，以此来降低向量空间维数，从而简化计算，提高文本处理的速度和效率。文本特征选择对文本内容的过滤和分类、聚类处理、自动摘要以及用户兴趣模式发现、知识发现等有关方面的研究都有非常重要的影响。通常根据某个特征评估函数计算各个特征的评分值，然后按评分值对这些特征进行排序，选取若干个评分值最高的作为特征词，这就是特征选择。

特征选择的方式有四种：(1)用映射或变换的方法把原始特征变换为较少的新特征；(2)从原始特征中挑选出一些最具代表性的特征；(3)根据专家的知识挑选最有影响的特征；(4)用数学的方法进行选取，找出最具分类信息的特征，这种方法是一种比较精确的方法，人为因素的干扰较少，尤其适合于文本自动分类挖掘系统的应用。

随着网络知识组织、人工智能等学科的发展，文本特征提取将向着数字化、智能化、语义化的方向深入发展，在社会知识管理方面发挥更大的作用。

2.1.1　向量空间模型

经典的向量空间模型由Salton等人于60年代提出，并成功地应用于著名的SMART文本检索系统。VSM概念简单，把对文本内容的处理简化为向量空间中的向量运算，并且它以空间上的相似度表达语义的相似度，直观易懂。当文档被表示为文档空间的向量，就可以通过计算向量之间的相似性来度量文档间的相似性。文本处理中最常用的相似性度量方式是余弦距离。文本挖掘系统采用向量空间模型，用特征词条(T1 ,T2 ,…Tn) 及其权值Wi 代表目标信息，在进行信息匹配时，使用这些特征项评价未知文本与目标样本的相关程度。特征词条及其权值的选取称为目标样本的特征提取，特征提取算法的优劣将直接影

响到系统的运行效果。设D为一个包含m个文档的文档集合，Di为第i个文档的特征向量，则有D=D1,D2,…,Dm，Di=(di1,di2,…,din)，i=1,2,…,m 其中dij(i=1,2,…,m;j=1,2,…,n)为文档Di中第j个词条tj的权值,它一般被定义为tj在Di中出现的频率tij的函数,例如采用TFIDF函数,即dij=tij*log(N/nj)其中,N是文档数据库中文档总数,nj是文档数据库含有词条tj的文档数目。假设用户给定的文档向量为Di,未知的文档向量为Dj，则两者的相似程度可用两向量的夹角余弦来度量，夹角越小说明相似度越高。相似度的计算公式如下：

通过上述的向量空间模型，文本数据就转换成了计算机可以处理的结构化数据，两个文档之间的相似性问题转变成了两个向量之间的相似性问题。

2.1.2　文本特征选择

机器学习算法的空间、时间复杂度依赖于输入数据的规模，维度规约则是一种被用于降低输入数据维数的方法。维度规约可以分为两类：特征选择(feature selection)，从原始的d维空间中，选择为我们提供信息最多的k个维(这k个维属于原始空间的子集) 特征提取(feature extraction)，将原始的d维空间映射到k维空间中(新的k维空间不属于原始空间的子集) 在文本挖掘与文本分类的有关问题中，常采用特征选择方法。原因是文本的特征一般都是单词，具有语义信息，使用特征选择找出的k维子集，仍然是单词作为特征，保留了语义信息，而特征提取则找k维新空间，将会丧失了语义信息。对于一个语料而言，我们可以统计的信息包括文档频率和文档类比例，所有的特征选择方法均依赖于这两个统计量，目前，文本的特征选择方法主要有：DF, MI, IG, CHI，WLLR,WFO六种。为了方便描述，我们首先一些概率上的定义：p(t):一篇文档x包含特征词t的概率。:文档x不属于C i 的概率。p(C i —t):已知文档x的包括某个特征词t条件下，该文档属于C i 的概率: 已知文档属于C i 条件下，该文档不包括特征词t的概率类似的其他的一些概率如p(C i), ，等，有着类似的定义。为了估计这些概率，我们需要通过统计训练样本的相关频率信息，如下表：

其中：A ij : 包含特征词t i，并且类别属于C j 的文档数量B ij : 包含特征词t i，并且类别属于不C j 的文档数量C ij ：不包含特征词t i，并且类别属于C j 的文档数量D ij ：不包含特征词ti，并且类别属于不C j 的文档数量A ij + B ij : 包含特征词t i 的文档数量C ij + D ij ：不包含特征词t i 的文档数量A ij + C ij ： C j 类的文档数量数据B ij + D ij ： 非C j 类的文档数量数据A ij + B ij + C ij + D ij = N :语料中所有文档数量。有了这些统计量，有关概率的估算就变得容易，如：p(t i) = (A ij + B ij) / N; p(C j) = (A ij + C ij) / N; p(C j —t j) = A ij / (A ij + B ij) .....类似的一些概率计算可以依照上表计算。介绍了事情发展的前因，现在进入

正题：常见的四种特征选择方法如何计算。1）DF(Document Frequency) DF:统计特征词出现的文档数量，用来衡量某个特征词的重要性，DF的定义如下：

DF的动机是，如果某些特征词在文档中经常出现，那么这个词就可能很重要。而对于在文档中出现很少(如仅在语料中出现1次)特征词，携带了很少的信息量，甚至是”噪声”，这些特征词，对分类器学习影响也是很小。DF特征选择方法属于无监督的学习算法(也有将其改成有监督的算法，但是大部分情况都作为无监督算法使用)，仅考虑了频率因素而没有考虑类别因素，因此，DF算法的将会引入一些没有意义的词。如中文的”的”、”是”，”个”等，常常具有很高的DF得分，但是，对分类并没有多大的意义。2）MI(Mutual Information) 互信息法用于衡量特征词与文档类别直接的信息量，互信息法的定义如下：

继续推导MI的定义公式：

从上面的公式上看出：如果某个特征词的频率很低，那么互信息得分就会很大，因此互信息法倾向”低频”的特征词。相对的词频很高的词，得分就会变低，如果这词携带了很高的信息量，互信息法就会变得低效。3）IG(Information Gain) 信息增益法，通过某个特征词的缺失与存在的两种情况下，语料中前后信息的增加，衡量某个特征词的重要性。信息增益的定义如下：

依据IG的定义，每个特征词t i 的IG 得分前面一部分：计算值是一样，可以省略。因此，IG 的计算公式如下：

IG与MI存在关系：

因此，IG方式实际上就是互信息与互信息加权。4）CHI(Chi-square) CHI特征选择算法利用了统计学中的”假设检验”的基本思想：首先假设特征词与类别直接是不相关的，如果利用CHI分布计算出的检验值偏离阈值越大，那么更有信心否定原假设，接受原假设的备则假设：特征词与类别有着很高的关联度。CHI的定义如下：

对于一个给定的语料而言，文档的总数N以及C j 类文档的数量，非C j 类文档的数量，他们都是一个定值，因此CHI的计算公式可以简化为：

CHI特征选择方法，综合考虑文档频率与类别比例两个因素

2.1.3　特征权重计算

经过文本预处理和特征选择，需要计算特征词权重。目前常用的特征词权重计算方法有布尔权重法、对数权重法、TF-IDF权重法、平方根权重法和基于熵的权重法。TF-IDF（term frequency－inverse document frequency）是一种用于资讯检索与资讯探勘的常用加权技术。TF-IDF是一种统计方法，用以评估一字

词对于一个文件集或一个语料库中的其中一份文件的重要程度。字词的重要性随著它在文件中出现的次数成正比增加，但同时会随著它在语料库中出现的频率成反比下降。TF-IDF加权的各种形式常被搜寻引擎应用，作为文件与用户查询之间相关程度的度量或评级。除了TF-IDF以外，因特网上的搜寻引擎还会使用基于连结分析的评级方法，以确定文件在搜寻结果中出现的顺序。

TF-IDF的主要思想是：如果某个词或短语在一篇文章中出现的频率TF高，并且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力，适合用来分类。TFIDF实际上是：TFXIDF，TF词频(Term Frequency)，IDF反文档频率(Inverse Document Frequency)。TF表示词条，在文档d中出现的频率。IDF的主要思想是：如果包含词条t的文档越少，也就是n越小，IDF越大，则说明词条t具有很好的类别区分能力。如果某一类C. 中包含词条t的文档数为m，而其它类包含t的文档总数为k，显然所有包含t的文档数n=m+k，当gfl大的时候，n也大，按照IDF公式得到的IDF的值会小，就说明该词条t类别区分能力不强。但是实际上，如果一个词条在一个类的文档中频繁出现，则说明该词条能够很好代表这个类的文本的特征，这样的词条应该给它们赋予较高的权重，并选来作为该类文本的特征词以区别与其它类文档。这就是IDF的不足之处.　　原理
在一份给定的文件里，词频(term frequency, TF) 指的是某一个给定的词语在该文件中出现的次数。这个数字通常会被正规化，以防止它偏向长的文件。（同一个词语在长文件里可能会比短文件有更高的词频，而不管该词语重要与否。）
逆向文件频率(inverse document frequency, IDF) 是一个词语普遍重要性的度量。某一特定词语的IDF，可以由总文件数目除以包含该词语之文件的数目，再将得到的商取对数得到。　　某一特定文件内的高词语频率，以及该词语在整个文件集合中的低文件频率，可以产生出高权重的TF-IDF。因此，TF-IDF倾向於过滤掉常见的词语，保留重要的词语。　　例子　　有很多不同的数学公式可以用来计算TF- IDF。词频(TF) 是一词语出现的次数除以该文件的总词语数。假如一篇文件的总词语数是100个，而词语「母牛」出现了3次，那麼「母牛」一词在该文件中的词频就是0.03 (3/100)。一个计算文件频率(DF) 的方法是测定有多少份文件出现过「母牛」一词，然後除以文件集里包含的文件总数。所以，如果「母牛」一词在1,000份文件出现过，而文件总数是10,000,000份的话，其文件频率就是0.0001 (1000/10,000,000)。最後，TF-IDF分数就可以由计算词频除以文件频率而得到。以上面的例子来说，「母牛」一词在该文件集的TF- IDF分数会是300 (0.03/0.0001)。这条公式的另一个形式是将文件频率取对数。
在向量空间模型里的应用　　TF-IDF权重计算方法经常会和余弦相似度(cosine similarity)一同使用於向量空间模型中，用以判断两份文件之间的相似性。

TFIDF算法是建立在这样一个假设之上的：对区别文档最有意义的词语应该是那些在文档中出现频率高，而在整个文档集合的其他文档中出现频率少的词语，所以如果特征空间坐标系取TF词频作为测度，就可以体现同类文本的特点。另外考虑到单词区别不同类别的能力，TFIDF法认为一个单词出现的文本频数越小，它区别不同类别文本的能力就越大。因此引入了逆文本频度IDF的概念，以TF和IDF的乘积作为特征空间坐标系的取值测度，并用它完成对权值TF的调整，调整权值的目的在于突出重要单词，抑制次要单词。但是在本质上IDF是一种试图抑制噪音的加权，并且单纯地认为文本频数小的单词就越重要，文本频数大的单词就越无用，显然这并不是完全正确的。IDF的简单结构并不能有效地反映单词的重要程度和特征词的分布情况，使其无法很好地完成对权值调整的功能，所以TFIDF法的精度并不是很高。　　此外，在TFIDF算法中并没有体现出单词的位置信息，对于Web文档而言，权重的计算方法应该体现出HTML的结构特征。特征词在不同的标记符中对文章内容的反映程度不同，其权重的计算方法也应不同。因此应该对于处于网页不同位置的特征词分别赋予不同的系数，然后乘以特征词的词频，以提高文本表示的效果。

2.2　层次多标签学习

层次多标签学习涉及到多标签学习和层次学习两方面内容。本节首先介绍多标签学习的内容，在此基础上引申到层次多标签学习的定义以及相关方法介绍。

2.2.1　多标签学习

2.2.2　层次多标签学习

2.2.3　层次多标签学习方法

第三章　LocalBalance算法

针对案件适用法律的自动识别问题，本文提出了一种局部的层次多标签分类算法。

3.1　算法描述

第四章　案件适用法律自动识别

本章介绍运用本文提出算法进行案件适用法律自动识别的过程，包括数据的获取和处理，实验平台，评价指标以及算法取得的实验效果。

4.1　实验数据

本文实验数据来自浙江法院公开网，包含浙江省各级法院公开的部分裁判文书，处理后的数据集中共包含裁判文书约15万份。

4.1.1　数据获取

浙江省平湖市人民法院
民 事 判 决 书
（2011）嘉平乍商初字第 18 号

原告：XXX。
委托代理人：YYY。
被告：ZZZ。
原告 XXX 为与被告 ZZZ 买卖合同纠纷一案，本院于 2010 年 12 月 31 日立案受理，依法组成合议庭，于 2011 年 7 月 11 日公开开庭进行了审理。原告 XXX 的委托代理人 YYY 到庭参加诉讼，被告 ZZZ 经本院传票合法传唤，无正当理由拒不到庭。本案现已审理终结。
原告诉称，被告于 2007 年在嘉兴港区承接钢贸城水电工程时，向原告赊购管道材料。截止 2008 年 12 月 18 日，被告结欠原告材料款计人民币 62000 元。当时被告约定于 2008 年 12 月 31 日付清，并约定被告如逾期付款由平湖市人民法院审理。但被告未能按时支付，直到 2009 年 1 月 19 日，被告以银行卡汇付了 30000 元，本金 32000 元……
本院认为，合法的买卖关系应受法律保护，被告向原告购买货物后，未及时支付货款，显属欠理，现应承担立即支付货款并负担逾期付款损失的义务。原告的诉讼请求，符合法律规定，本院予以支持。据此，依照《中华人民共和国合同法》第一百六十一条、第一百零七条及《中华人民共和国民事诉讼法》第一百三十条之规定，判决如下：
被告 ZZZ 于本判决生效后十日内支付原告 XXX 货款 32000 元及逾期付款损失……

审判长　AAA
审判员　BBB
审判员　CCC

二〇一一年七月十一日
书记员　DDD

图 4.1: 裁判文书样例

为了获取足量的样本数据用于模型的训练，本文利用基于jsoup的爬虫技术实现裁判文书的快速获取。jsoup 是一款Java 的HTML 解析器，可直接解析某个URL地址、HTML文本内容。它提供了一套非常省力的API，可通过DOM，

CSS以及类似于JQuery的操作方法来取出和操作数据。jsoup的主要功能包括从一个URL、文件或字符串中解析HTML；使用DOM或CSS选择器来查找、取出数据；操作HTML元素、属性、文本等。通过jsoup提供的API，我们首先获取了裁判文书文本所在页面的URL，然后从中解析HTML，并提取出其中的裁判文书文本。

从浙江法院公开网上获取的裁判文书均以文本形式存储，为非结构化数据，而预测模型的训练需要结构化的数据，因此首先需要对数据进行预处理。

一份典型的裁判文书主要由首部、事实、理由、裁判结果和尾部组成。图1.1给出了裁判文书的样例，其中黄色高亮部分为案件事实描述部分，红色高亮部分为案件适用的法律条文。本文数据预处理的第一步就是要从裁判文书中提取案件事实描述和案件适用的法律条文，前者用于构建数据集的特征，而后者则构成了数据集的标签。通过分析裁判文书的行文结构，我们发现案件事实描述的前后段落具有一定的模式，前一段落通常以“本案现已审理终结”或“本案依法缺席审理”结尾，后一段落通常以“本院认为”开头；类似的，裁判理由部分多以“本院认为”开头，而以“判决如下”结尾，其中案件适用的法律条文多以“依照”、“依据”、“根据”、“按照”等开头，以“之规定”、“的规定”等结尾。

4.1.2　数据清洗

介绍对噪声数据进行清洗的过程。

4.1.3　标签提取

介绍从裁判文书中提取法律依据过程，法律依据也是本文实验数据的标签部分，因此，还涉及到标签的表示。

4.1.4　特征提取

介绍使用分词手段进行特征提取以及信息增益方法进行特征选择

4.2　实验平台

介绍本文实验平台Mulan，以及实验中采用的评价指标。

4.3　评价指标

4.4　实验结果

给出算法在实验数据集上的预测表现，与已有算法的性能比较等。

第五章　总结与展望

总结本文的贡献和不足，提出后续的改进方案。

5.1　工作总结

本文首先简述了分布式算法及其原型实现的概念，介绍了论文的背景。然后概括了其他网络模拟平台和离散事件模拟方面的相关工作，分析了其与本文工作的区别和联系。之后，我们阐述了论文的基础理论知识，对分布式计算模型进行了系统分析。论文的主要部分是DAPro 平台的设计与实现，以及基于DAPro 平台两个分布式算法的实现与分析。总的来看，本文的主要工作包括以下几个方面：

- 详细阐述了DAPro 平台的设计实现。总体概述描述了DAPro 平台的模块结构和运行流程，模块设计部分对平台每个模块的功能和实现进行了详细分析，并以一个DFS 生成树的构建算法为例描述了如何使用DAPro 平台实现一个分布式算法，展示了平台运行算法的流程。

- 在原始框架的基础上对DAPro 平台的功能进行了大量扩充。本文在DAPro 平台原始框架的基础上增加了更为丰富的Connector 类型、更多的事件和进程类型、更多的事件动作，并增加了故障生成等模块，提高了系统对于分布式算法的模拟能力。

- 基于DAPro 平台实现了分布式的DFS 生成树构建算法和可容错的协商算法。通过具体的实验和对结果的分析，验证了DAPro 平台的正确性和可用性。

5.2　改进方向

参考文献

[1] 向李兴. 基于自然语义处理的裁判文书推荐系统设计与实现[D]. [S.l.] : 南京大学, 2015.

[2] AGGARWAL C C, ZHAI C. Mining text data[M]. [S.l.] : Springer Science & Business Media, 2012.

[3] TSOUMAKAS G, KATAKIS I. Multi-label classification: An overview[J]. Dept. of Informatics, Aristotle University of Thessaloniki, Greece, 2006.

[4] SILLA JR C N, FREITAS A A. A survey of hierarchical classification across different application domains[J]. Data Mining and Knowledge Discovery, 2011, 22(1-2) : 31 – 72.

[5] BARUTCUOGLU Z, SCHAPIRE R E, TROYANSKAYA O G. Hierarchical multi-label prediction of gene function[J]. Bioinformatics, 2006, 22(7) : 830 – 836.

致　　谢

首先感谢XXX