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

表　　格

插　　图

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，特征权重计算和特征选择等。

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。如果忽略标签的树形结构特征，无疑会损失分类标签的重要信息[4]，模型无法达到满意的预测性能。因此，如何利用法律适用识别问题中标签的结构信息，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。

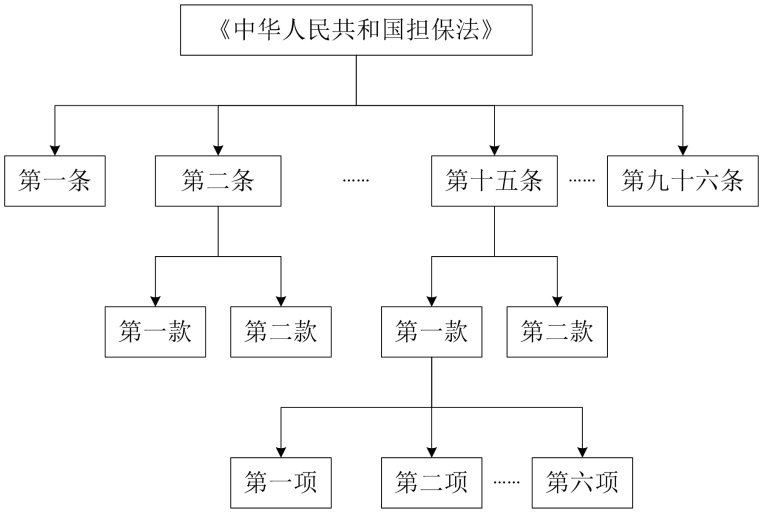

图 1.1: 法律条文树形结构示例

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社会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。

第二章主要对本文研究涉及的文本挖掘相关技术进行介绍，包括文本的表示模型，特征词权重计算和特征选择等。

第三章介绍层次多标签学习的概念、现有方法、常用评价指标及其应用。

第四章详述本文提出的$LocalBalance$层次多标签学习算法。

第五章为实验部分，包括裁判文书文本的处理过程，以及对实验结果的分析。

第六章对本文工作进行总结，并提出改进的方向。

第二章　文本挖掘技术

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、计算机语言学等多个领域。文本挖掘旨在从大量文本中发现隐含的知识和模式，它从数据挖掘发展而来，但又与传统的数据挖掘有许多不同。文本挖掘的对象是海量、异构、分布的文本，文本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，而文本挖掘所处理的文本都是非结构化或半结构化的。所以，文本挖掘面临的首要问题是如何合理地表示文本，使之既能包含足够的信息以充分反映文本的特征，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在的。

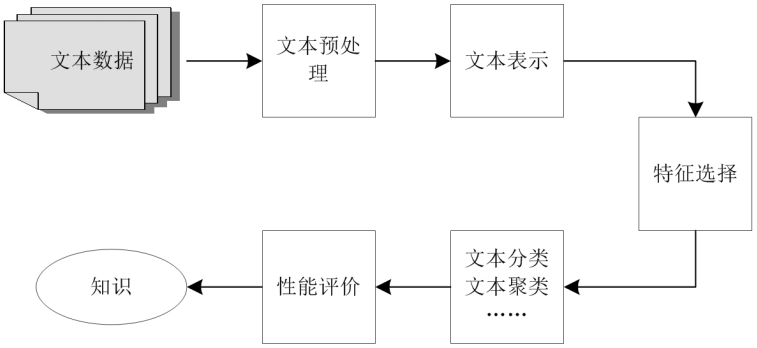

图 2.1: 文本挖掘流程

文本挖掘的一般流程如图2.1所示，其中文本的表示及其特征项的选取是文本挖掘、信息检索的核心问题，其目的是将非结构化的文本数据转化为结构化的计算机可以识别处理的数据。目前有关文本挖掘的研究主要集中于文本表示模型和特征选择算法的研究上。用于表示文本的特征项必须具备一定的特性：(1)特征项要能够确切标识文本的内容；(2) 特征项具有区分不同文本的能力；(3)特征项的个数不能过多；(4)特征项的提取要比较容易实现。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，用来实现文档之间的相似度计算。如果把切分得到的所有词都作为特征项，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。特征选择的目的是在

保留文本重要信息的情况下尽量减少特征词的数量，以此来降低特征向量空间维数，从而简化计算，提高文本挖掘的速度和效率。文本特征选择对文本内容的过滤和分类、聚类、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词的评分值，然后按评分值对这些特征词进行排序，选取若干个评分值最高的特征词作为特征。

2.1　文本表示

2.1.1　文本表示模型

目前，最常用的文本表示模型有两种，一种是布尔模型，另一种是向量空间模型(Vector Space Model, VSM)。经典的向量空间模型由Salton等人[6]于上世纪七十年代提出，并成功地应用于著名的SMART文本检索系统。向量空间模型概念简单，将文档d_i表示成m维特征空间中的一个向量，即$d_i = (w_{i1}, w_{i2}, ..., w_{im})$，特征空间$(t_1, t_2, , , t_m)$由从文档集合中得到的$m$个特征词组成，$w_{ij}(1 \leq j \leq m)$表示特征$t_j$ 在文档d_i中的权重。布尔表示模型可以看做向量空间模型的一个特例，区别在于布尔表示模型中特征词的权重由该特征词在文档中是否出现，若出现则权重为1，否则为0；而向量空间模型中，特征词的权重通常为该特征词在文档中出现频率的函数。

利用布尔模型和向量空间模型表示文本，假定了文本中的字或词的出现是相互独立的，即文本中字或词出现的次序对文本的分类没有影响。在实际情况中，这一假设往往是不成立的，忽略字词出现的顺序将丢失大量的文本内容信息，造成模型的表示能力不足。尽管存在表示能力不足的问题，向量空间模型的简洁实用，以及在实际应用中取得的较好效果，使得其成为了应用最为广泛的文本表示模型。

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，可以利用欧氏距离、余弦相似度等进行衡量。

2.1.2　特征权重计算

向量空间模型中，特征的权重计算有不同的方法，特征词的权重取值，会在较大程度上影响文本分类算法的性能。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等，其中TF-IDF 算法是应用最为普遍的方法。

TF-IDF(Term Frequency-Inverse Document Frequency)是一种应用于信息检索与文本挖掘的常用加权技术。TF-IDF 是一种统计方法，用以评估字或词对于一个文档集合或文档集合中某个文档的重要程度。TF-IDF的主要思想是：一个字或词在特定文档中的重要性随着其在文档中出现的频率成正比提高，同时会随着其在文档集合中出现的频率成反比降低，即如果某个字词或短语在一份文档中出现的频率高，并且在文档集合中的其他文档中很少出现，则认为该字词或者短语具有很好的类别区分能力，适合用来分类，应该具有更高的权重。

利用TF-IDF计算特征词的权重需要考虑三方面的因素：

- 词频tf(Term Frequency)；特征词在文档中出现的频率。
- 倒排文档频率idf(Inversed Document Frequency)：特征词在文档集合中分布情况的量化，度量特征词在文本集中出现的频繁程度。常用的计算方法是$\log(N/n_j+0.01)$，其中N是文档集合中文档的总数，n_j是文档集合中出现特征词t_j的文档数量。
- 归一化因子(Normalization Factor)：对文本向量中的各个分量进行归一化，归一化后的文本向量为单位向量，从而减轻文本长度对于文本特征权重的影响。

综上，TF-IDF的计算公式如下：

$$tf\text{-}idf_{ij}=\frac{tf_{ij}\times\log(N/n_j+0.01)}{\sqrt{\sum_{t_j\in d_i}[tf_{ij}\times\log(N/n_j+0.01)]^2}} \tag{2.1}$$

其中，tf_{ij}表示特征词t_j在文档d_i中出现的频率，idf_j表示特征词t_j在文档集合中的反文档频率，N表示文档集合中的文档总数，n_j表示特征词t_j在文档集合中的文档频率，即文档集合中出现特征词t_j的文档数目，分母为归一化因子。

2.2　文本特征选择

目前人们通常采用向量空间模型来表示文本，通过把从文本中抽取出的特征词进行量化来表示文本信息，但是如果直接用分词算法和词频统计方法得到的特征词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向量表示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。

特征降维方法可以分为两类：特征选择(Feature Selection)和特征提取(Feature Extraction)。特征选择是从原始的d维特征空间中，选择为我们提供信息最多的k个维，这k个维是原始特征空间的子集；特征提取则是将原始的d维特征空间映射到k维空间中，新的k维空间不属于原始特征空间。在文本挖掘与文本分类中，通常采用特征选择方法进行维度约简，原因在于文本的特征一般都是词，具有语义信息，使用特征选择得到的k维特征，仍然是以单词作为特征，保留了语义信息，而特征提取得到的k维特征，会失去原有的语义信息，影响模型的可解释性。

常用的文本特征选择方法主要基于文档频率(DF)，信息增益(IG)，互信息(MI)，卡方统计(CHI)等衡量指标。

2.2.1　文档频率

文档频率(Document Frequency, DF)通过统计特征词在文档集合中出现的文档数量，来衡量某个特征词的重要性。对于文档集合中的每个特征词，计算其DF值，DF值低于某个设定的阈值的词将从特征空间中移除。使用文档频率进行特征选择的原理是，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的特征词，携带的信息量很少，甚至可能是“噪声”，这些特征词，对分类器学习影响很小。文档频率特征选择方法属于无监督的学习算法，仅考虑了频率因素而没有考虑类别因素，因此，文档频率算法将会考虑一些频率较高但没有分类意义的词，如中文里的“的”、“是”，“个”等。这类词常常具有很高的文档频率，但是对分类影响很小。

基于文档频率的特征选择是最为简单的方法，往往无法选取最具分类信息的特征词，但是由于其计算复杂度低，随文本集合中文档的数量线性增长，因此适合海量数据的处理。

2.2.2　信息增益

信息增益(Information Gain, IG)方法是通过比较当得知一个文档中某个特征词存在或缺失时，所获得的用于标签预测的信息量，来衡量某个特征词的重要性。令$C_{i}{}_{i=1}^{m}$表示目标空间中的标签集合，信息增益的定义为：

$$\begin{aligned} IG(t) = & -\sum_{i=1}^{m} P_r(C_i) \log P_r(C_i) \\ & + P_r(t) \sum_{i=1}^{m} P_r(C_i|t) \log P_r(C_i|t) \\ & + P_r(\bar{t}) \sum_{i=1}^{m} P_r(C_i|\bar{t}) \log P_r(C_i|\bar{t}) \end{aligned} \tag{2.2}$$

上述定义是信息增益的一种通用形式，可以应用到多分类情形中，考虑到文本分类往往不是简单的二分类问题，通用形式的信息增益可以在衡量特征词的重要性时考虑到所有类别。对于文档集合中的每个特征词，我们计算其信息增益IG，IG值低于某个设定的阈值的词将从特征空间中移除。信息增益的计算包括条件概率的估计和熵的计算，其中概率估计的时间复杂度为$O(N)$，空间复杂度为$O(VN)$，其中N为文档集合的大小，V为特征词的数量；熵的计算时间复杂度为$O(Vm)$。

2.2.3 互信息

互信息法用于表征特征词与文档类别之间的相关性。考虑特征词t和文档类别c，令A为包含特征词t且属于类别c的文档数量，B为包含特征词t而不属于类别c的文档数量，C 为不包含特征词t而属于类别c的文档数量，N为总文档数量，则特征词t和类别c的互信息定义为

$$MI(t,c) = \log \frac{P_r(t \wedge c)}{P_r(t) \times P_r(c)} \tag{2.3}$$

可以由上述定义的A, B, C, N进行估算：

$$MI(t,c) \approx \log \frac{A \times N}{(A+C) \times (A+B)} \tag{2.4}$$

当特征词t和类别c独立时，其互信息为0。为了衡量特征词的重要性，需要综合各个类别的互信息，综合的方法有取均值和取最大值两种：

$$MI_{avg}(t) = \sum_{i=1}^{m} P_r(c_i) MI(t,c_i) \tag{2.5}$$

$$MI_{max}(t) = \max_{i=1}^{m} MI(t,c_i) \tag{2.6}$$

互信息的缺点在于容易受到特征词的边缘概率影响，即如果某个特征词的频率很低，那么其互信息值就会很大，因此互信息倾向于选择“低频”的特征词，而词频很高的词互信息值就会变低，如果这词携带了很高的信息量，互信息法就会变得低效。互信息的计算复杂度与信息增益类似。

2.2.4 卡方统计

卡方统计(Chi-square)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：即特征词与类别有着很高的关联度。考虑特征词t和文档类别c，令A为包含特征词t且属于类别c的文档数量，B为包含特征词t而不属于类别c的文档数量，C为不包含特征词t而属于类别c的文档数量，D为不包含特征词t且不属于类别c的文档数量，N为总文档数量，则特征词t和类别c的卡方统计量定义为：

$$\chi^2(t,c) = \frac{N \times (AD - BC)^2}{(A+C) \times (B+D) \times (A+B) \times (C+D)} \tag{2.7}$$

当特征词t和类别c独立时，其卡方统计量为0。针对某个特征词，我们计算其关于各个类别的卡方统计量，然后分别用取均值和取最大值两种方式进行综合：

$$\chi^2_{avg}(t) = \sum_{i=1}^{m} P_r(c_i)\chi^2(t,c_i) \tag{2.8}$$

$$\chi^2_{max}(t) = \max_{i=1}^{m} \chi^2(t,c_i) \tag{2.9}$$

第三章　层次多标签学习

层次多标签学习涉及到层次和多标签两方面特性。本章首先介绍层次多标签学习的定义，然后对现有的层次多标签学习方法进行综述。

3.1　问题定义

层次多标签问题是

3.2　学习方法

3.2.1　局部学习方法

3.2.2　全局学习方法

第四章　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。

4.1　算法描述

第五章　法律适用自动识别

本章介绍利用本文提出的层次多标签学习算法进行案件适用法律自动识别的过程，包括裁判文书文本的获取和处理，使用的实验平台，采用的评价指标以及算法取得的实验效果等。

5.1　实验数据

5.1.1　数据获取

本文实验数据取自浙江法院公开网公开的浙江省各级人民法院裁判文书，所有裁判文书均以文本形式存在。为了获取足量的裁判文书样本用于预测模型的学习，本文利用基于jsoup 的爬虫技术实现裁判文书的快速自动获取。jsoup是一款用于处理实际HTML的Java库，它提供了非常方便的API，能够充分利用DOM，CSS以及类似JQuery的方式来提取和操作数据。jsoup 实现了WHATWG HTML 规范，可以将HTML解析成与现代浏览器一样的DOM。jsoup的主要功能包括：

- 从URL、文件或字符串中解析HTML
- 使用DOM遍历器或CSS选择器来查找和提取数据
- 操作HTML元素、属性、文本
- 基于安全的白名单对用户提交的内容进行清理来预防跨站脚本攻击
- 输出整洁的HTML

通过jsoup提供的API，我们首先从浙江法院公开网获取了裁判文书文本所在页面的URL，然后从URL中解析HTML，提取出其中的裁判文书文本，并以文本文件的形式保存至本地。

5.1.2　数据预处理

从浙江法院公开网获取的裁判文书均为文本形式，为非结构化数据，因此需要进行数据的预处理。数据预处理主要完成以下几项工作：

- 提取案件事实描述和适用法律，将裁判文书文本转化为半结构化数据

- 对案件适用法律中的错误和格式不一致进行修正
- 对案件事实描述进行分词和词性标注

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 5.1: 裁判文书样例

一份典型的裁判文书主要由首部、案件事实描述、裁判理由、裁判结果和尾部组成。图5.1 给出了裁判文书的样例，其中直线下划线标注部分为案件事实描述部分，曲线下划线标注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，前者用于生成样本的特征向量，而后者则为样本的标签。通过分析裁判文书的行文结构，我们发现案件事实描述在裁判文书中的位置是固定的，其所在段落的前一段落通常以“本案现已审理终结”或“本案依法缺席审理”结尾，而后一段落通常以“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部分多以“依照”、“依据”、“根据”、“按照”等开头，以“之规定”、“的规定”等结尾。根据以上规律，我们可以从裁判文书文本中提取出案件事实描述和适用法律两部分内容，将原始的裁判文书文本初步转化为半结构化数据。

由于裁判文书的书写、电子化等过程基本是人为实现，因此难免会出现笔误、格式错误等现象，在对上述半结构化数据进行进一步处理之前，需要对数据进行清理，减少数据中的错误。与案件事实描述不同，案件适用法律来自既已成文的法律条文，有规范可循，因此对数据的清理主要是对案件适用法律的校验。一种常见错误是法律名称错误或格式不一致，比如“《中华人民共和国民事诉讼法》”被错误写成“《中华人民共和国民诉法》”，“最高人民法院《关于适用〈中华人民共和国婚姻法〉若干问题的解释（二）》”与“《最高人民法院关于适用〈中华人民共和国婚姻法〉若干问题的解释（二）》”格式不一致等；另一种常见错误是法条编号的错误或格式不一致，比如编号中出现非数字、文字编号和数字编号混乱等；第三种错误是间隔符错误，由于案件适用的法律法条通常有多个，一般不同法律之间以逗号间隔，同一法律不同法条之间以顿号间隔，这在后续的标签处理过程中非常重要。由于本文需要处理的裁判文书数量较大，因此我们只对出现次数较多的错误进行了处理，而对出现较少的错误出于时间成本的关系选择了忽略。

本文采用向量空间模型对案件事实描述文本进行表示，首先对案件事实描述的内容进行分词，然后根据特征词将案件事实描述结构化地表示为特征向量。本文选用了哈工大的语言技术平台（Language Technology Platform, LTP）作为语言处理工具。LTP 是一整套中文语言处理系统，制定了基于XML的语言处理结果表示，并在此基础上提供了一整套自底向上的丰富而且高效的中文语言处理模块（包括词法、句法、语义等六项中文处理核心技术），以及基于动态链接库（Dynamic Link Library, DLL）的应用程序接口、可视化工具，并且能够以网络服务（Web Service）的形式进行使用。本文主要利用LTP语言处理系统进行案件事实描述的分词和词性标注，原始的文本内容经过处理转化为一系列词及其词性的列表。分词产生的词即特征词，用于结构化地表示案件事实，词性标注则有助于我们去除一些无意义的词，减少特征词的数量，减少后续的计算量。

5.1.3　标签处理

案件适用法律的自动识别作为一个层次多标签分类问题，其标签，即案件适用的法律以标签向量的形式表示，向量的每一维代表一项具体的法律或法条，如“《中华人民共和国民事诉讼法》”、“《中华人民共和国民事诉讼法》第六十四条”、“《中华人民共和国民事诉讼法》第六十四条第一款”等，标签对应法律条文的名称也直接反映了标签的树形层次结构。

在数据预处理的过程中，我们对案件适用法律中的几种错误和不一致情况

进行了处理，然而处理后的结果无法直接表示为标签向量，需要进一步的提取。如前文所述，在裁判文书中，一个案件可能适用多项法律条文，不同法律之间以逗号间隔，同一法律不同法条之间以顿号间隔，因此标签的提取主要是将案件适用法律切分为一项或多项完整的法律条文。通过对裁判文书中法律适用格式的分析，我们总结了案件适用法律书写的几种格式：

- 单项法律条文：如“《中华人民共和国合同法》第二百零六条”
- 多项完整法律条文：如“《中华人民共和国合同法》第二百零六条，《中华人民共和国民事诉讼法》第一百四十四条”
- 带非完整法律条文：如“《中华人民共和国合同法》第二百零五条、第二百零六条、第二百零七条，《中华人民共和国民事诉讼法》第一百四十四条”
- 带文字省略的法律条文：如“《中华人民共和国刑法》第七十二条第一、二、三款”

上述几种法律适用的书写格式中，前三种情况比较容易处理，也最为常见，而第四种情况形式更为多样，处理也较为复杂。经过标签的提取，原本书写格式混乱的法律适用中每一项都转换为完整法律条文，如“《中华人民共和国合同法》第二百零五条、第二百零六条、第二百零七条”将转换为“《中华人民共和国合同法》第二百零五条；《中华人民共和国合同法》第二百零六条、《中华人民共和国合同法》第二百零七条”；“《中华人民共和国刑法》第七十二条第一、二、三款”将转换为“《中华人民共和国刑法》第七十二条第一款；《中华人民共和国刑法》第七十二条第二款；《中华人民共和国刑法》第七十二条第三款”，从而可以直接对应于标签空间的某一维。

理想地，标签空间由所有既已成文的法律条文构成，然而一方面这会造成标签空间维度过高，使得学习算法在时间和空间上耗费太多资源；另一方面大多数法律条文在数据集中没有出现或者出现极少，无法得到有效的预测模型。因此，需要对标签空间进行约简，选择合适的标签组成标签空间。本文基于标签频率进行标签的选择，即选择出现频率达到一定阈值的标签，构成标签空间。这样处理的原因还在于，上一节中对案件适用法律中的错误和格式不一致无法做到完全修正，基于阈值的标签选择可以很容易地将噪声数据过滤。

经过上述过程，预处理之后的标签部分，即案件适用法律表示为标签的向量，向量的每一维代表一项具体的法律或法条。对于每一个案件样本的标签向

量，适用的法律条文对应的条目值为1，没有适用的法律条文对应的条目值为0，考虑标签的树形层次结构，

5.1.4　特征处理

本文采用向量空间模型作为文本表示模型，特征的权重采用TF-IDF方法计算。

5.2　实验平台

Mulan[7]

5.3　评价指标

5.4　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。

第六章　总结与展望

总结本文的贡献和不足，提出后续的改进方案。

6.1　工作总结

6.2　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.

[6] SALTON G, WONG A, YANG C-S. A vector space model for automatic indexing[J]. Communications of the ACM, 1975, 18(11) : 613 – 620.

[7] TSOUMAKAS G, SPYROMITROS-XIOUFIS E, VILCEK J, et al. Mulan: A java library for multi-label learning[J]. The Journal of Machine Learning Research, 2011, 12 : 2411 – 2414.

致　　谢

首先感谢XXX