**VSM模型在文档结构识别中的应用**

宋昊苏，李 宁，张 伟

（北京信息科技大学 计算机学院，北京100101）

摘 要**：**借鉴向量空间模型（Vector Space Model，VSM），将文档的单元结构映射到特征向量，并对其进行归一化和加权处理，再计算标准文档结构分类特征向量和待查文档单元结构特征向量的相似度，实现对文档结构的识别。测试结果表明，该方法实现了文档格式的向量表示，易于计算，具有较好的扩展性，可在此基础上实现其他特定结构文档格式的自动检查或文档理解。

关 键 词**：**文档结构识别；向量空间模型；文档格式校验；文档理解

中图分类号：　TP317.2 文献标识码　A

Application of VSM model to document structure identification

SONG Hao-su, LI Ning, ZHANG Wei

(School of Computer Science， Beijing Information Science and Technology University, Beijing 100101，China)

Abstract: By referencing VSM model, the basic structure of document component is mapped into the characteristic vector. After normalization and weight assignment, the vector similarity between the standard document component structure and the structure of the document component to be identified can be calculated, thus the identification of whole document structure can be implemented. The method achieves vector formalization of document format. It is easily calculated, and has good expansibility. Applications such as automatic format checking or document understanding can be built on that basis.

Key words: document structure identification，VSM(Vector Space Model), document format checking；document understanding

0 引言

随着开放文档格式的广泛应用[1-3]，越来越多的文档采用开放的标准格式进行记录，使通过文档的存储格式来分析、理解文档的内容成为可能。人们也可以在理解文档结构的基础上进一步实现对文档格式的校验[4-7]。例如，如何让学生提交规范格式的论文是教师经常面对的问题，而通过人工的方式检查每篇论文的排版是否符合要求却十分困难。例如，学生是通过普通段落的方式还是用题注的方式来设置图题，很难用肉眼从显现式样中判断，而这却会影响日后对文档内容的检索利用。如果能有自动化的工具对文档的结构进行识别，例如从文档格式上判断插图后面的文字是普通段落还是通过题注添加的图题，并随时提示用户改正不规范的排版方式，将大大减轻工作量，并有助于保证最终提交的文档符合统一的排版规范。

显而易见，文档结构理解是格式检查的前提条件。而与文档的逻辑内容不同，文档格式主要记录的是文档的显现式样。例如，可以很容易地判断文档中存在某些段落，它们的字体大小超过一般的正文，但是却不能简单地由此认为这些字体较大的段落就是文章的标题，更难以判断是第几级标题。只有我们明确了某部分的内容属于哪一级标题时，才有可能依据排版规则（如1级标题用2号黑体）来检验其排版的方式是否符合规范。因此，文档的结构识别是文档理解和自动化格式检查的基础。

一般来说，学位论文等结构性较强的文档往往由十几种基本单元组成，例如，标题、摘要、关键字、章节、图表、索引等，文档结构识别的任务就是通过分析文档格式记录的式样信息和上下文信息，判断出文档各部分的结构。除了可能使用的多媒体内容之外，带格式的文本（或段落）是文档的基本组成单元，因此，文档结构识别的基本任务就是确定段落的分类或角色。可以把文本的式样表示为文本的一系列特征。而排版规则可以看作一些特定类别的文本（如标题、正文）所应具有的标准分类特征。对于待识别的文本，可以首先按照分类特征对其分类，确定其可能的角色，再计算该文本的特征空间与标准分类特征空间的差异。如果存在差异，便可指出其排版上可能存在的错误，从而实现文档格式的自动检查。

将文本对象转换为方便计算机处理的数值模型可以借助机器学习的一些成熟方法。本文采用的是VSM[8]，它是由Salton等提出的一种较为经典的文本表示模型，该模型在文本检索和文本自动分类系统中得到广泛应用。在文本检索[9]和文本分类系统[10]中，通常是将文本表示为向量形式，根据向量的相似度对返回的文本结果进行排序、分类等操作。笔者在文档结构识别中引入向量空间模型，借助文献[11]提出“平均值”阈值匹配方法，将文档的组成单元表示为由式样特征组成的特征向量，并对待查文本特征向量和标准分类特征向量进行归一化，经过加权后计算特征向量的夹角，最后根据预先给定的阈值对文档结构进行识别，取得了较好的识别效果。

1 特征向量空间模型的构建

在检索系统或文档分类系统的应用中，文章或者文章片段的特征项一般为词或短语。本文的重点是对文档结构进行识别，识别的对象是组成特定文档类型的基本单元，例如标题、中英文摘要、章节段落等。因此，可以对文档作如下定义。

定义1 文档结构。组成文档的基本单元。例如，对于学位论文，可能有标题、中英文摘要、章节内容和文献索引等，记为***S***。

定义2 特征项。指论文组成单元的式样特征，如字体、字形、字号、文本位置和特征字串等，记为***C***。

定义3 特征项权重。表示特征项在文档结构中的重要程度，记为***W***。

一篇文档*D*如果包含*n*个基本单元，那么该文档可以用*D*（*S*1，*S*2，*S*3，….，*S*n）表示，对于每个基本单元***S***，如果该单元包含*m*个格式特征，则该单元可以用*S*（*C*1，*C*2，….，Cm）描述。由于不同式样特征在某一种文档结构中的重要程度可能不同，需要对不同的文档格式特征分配一定的权重，权重的取值不同会使得比较向量所用的夹角阈值也不同，本文所用的特征权重取值基于统计实验，目的是获得较为合理的夹角阈值，使文档结构的识别结果更加准确。

2 特征项的量化

文档的基本组成单元是带格式的文本（段落），因此，识别时依次读取文档的各个段落，识别该段落属于哪种分类（角色）。本文将读取的段落特征项表示为：字体、字号、字形、对齐方式、首行缩进、行距、段前、段后、关键字等9个特征项，记为***C***1，***C***2.，…，***C***9。VSM模型中常用的向量权值计算方法主要有布尔函数法，词频权重法和TD-IDF权重法等[12]，其中布尔权重法较为简单，且应用广泛。针对段落特征项中字体、字形、对齐方式、关键字等较难量化的特征项，作如下处理：如果读取的特征符合标准的分类特征，则赋值为1；否则赋值为0，其他特征值为该特征项对应的实际取值。例如，某校的学位论文规范中对中文摘要正文的要求是：1）字体：宋体；2）字号：小四号；3）字形：不加粗；4）对齐方式：两端对齐；5）首行缩进：24磅；6）行距：20磅；7）段前：0磅；8）段后：0磅；9）段首关键字：“摘要”，则标准中文摘要的分类特征向量为（1,12,1,1,24,20,0,0,1），记为***S****a*；若某实际读取的段落的特征依次为：1）非宋体；2）9号字体；3）不加粗；4）两端对齐；5）无缩进；6）两倍行距；7）段前1行；8）段后0行；9）段首包含“摘要”，则该段落的特征向量为（0,9,1,1,0,24,15.6,0,1），记为***S****b*。

由于特征向量中各特征项的值在数量级上有较大差别，需要对段落特征向量统一数量级，即归一化处理。在模式识别中，向量归一化往往作为特征提取前的预处理技术，对提取特征后的特征向量进行特征转换[13]。

对段落特征向量*S*归一化的目的是使每个特征项***C***的值介于0~1之间，因此，对布尔值类型的特征项不作归一化处理，而对字号、首行缩进、行距等特征项***C****i*（*i*=2,5,6,7,8）应用公式（1）：

***C****i*=|***C****i*-***C***min|/|***C***max-***C***min| （1）

其中，特征项***C****i*的样本值集合为某类文档（如学位论文）规范中涉及所有段落的样本值，***C***max和***C***min分别为样本值中的最大值和最小值，利用式（1）得到归一化后某种分类的标准特征向量和待查段落的特征向量，分别记为***S’****a*，***S’****b*。其中每个特征项的取值都在0~1之间。考虑到每个特征项对分类的重要程度不同，例如，中文摘要中段首包含“摘要”关键词这一特征比对齐方式、首行缩进等特征更加重要，因此需要对特征向量进行加权处理。

**4 特征项加权**

特征向量的加权即依据一定的准则对特征项赋予一定权重，基于机器学习中加权的处理方法[15-16]，对文本特征向量作如下处理：

=***S****’a* （2）

其中，***S***’*a*是归一化处理后的标准分类特征向量； 是该分类的特征权重向量；***S***’*a*和作点乘运算后得到加权后的标准特征向量，将待查文本向量***S***’*b*进行加权后得到向量。特征权重向量的每个分量的取值范围是0~1，且各分量的总和为1，每个分量的取值反映的是该特征项在该分类中的重要程度，并且会影响下文中提到的向量比较所用的夹角阈值，进而影响文档结构的识别。

下面举例说明。某论文中文摘要标准格式是：字体：宋体；字号：小四；文本位置：两端对齐；首行缩进：0字符；行距：20磅；关键字串：段首有“关键词”；表示为特征向量：（1,12,1,0,20,1）；得到归一化后的标准特征向量为（1,0.155,1,0,1,1），权重特征向量取为（0.1,0.2,0.1,0,0.1,0.4），则加权后的标准特征向量为（0.1,0.03,0.1,0,0.06,0.4）。若实际出现的一段文本格式为字体：仿宋；字号：9号；文本位置：两端对齐；首行缩进：2字符；行距：两倍行距；关键字串：段首有“关键词”；那么经过归一化、加权后得到的待查特征向量为（0,0.02,0.1,0.05,0.1,0.4）；应用向量夹角余弦公式（3）计算得到向量夹角值为0.932（无单位）。

本文采取统计方法确定特征项的权重，对文档结构识别影响较小的特征分量权重较低，反之相反。这里还需要考虑文档编辑过程中用户习惯性排版错误导致的视觉误差，在此基础上可不断调整的分量取值，并确定标准向量和实际向量间的夹角阈值。目前，该夹角阈值的取值范围设定为0.85~0.95，例如，中文摘要标准特征向量和待查特征向量在视觉误差范围内的向量夹角阈值取为0.9，则在识别中文摘要标题后读取下一段时，将该段特征向量与中文摘要标准特征向量进行归一化、加权后应用公式（3）得到两向量的夹角值，记为*X*，若*X*0.9，则认为待查段落有较大可能是中文摘要，否则，认为该段可能是中文摘要的下一个文档结构，继续应用此法对当前段落进行识别。

**5文档结构相似度的计算**

在文本分类领域，基于VSM的分类方法主要有神经网络算法、贝叶斯算法、简单向量距离法[17]等，其中简单向量距离法较容易实现，效率高，分类精度能满足大多文本分类系统的需求，且应用较广。文本采用简单向量距离法将文档组成单元映射为向量空间中的点特征项，不同向量特征项之间的距离用向量间的余弦夹角度量，该余弦夹角反映了某文档单元在角色分类上的相似度。此外，本文所用的方法还将对特征项组成的特征向量进行归一化和加权处理，最终得到标准论文结构特征向量。假设待查论文结构特征向量为，则和的相似度为：

sim（******,）=cos（,）= （3）

sim（,）表示某类标准的文档单元结构和待查文档单元结构的相似度，在（3）中，和作点乘运算后的值除以2个向量分别作模运算后的乘积得到相似度的值，该值越大，说明某文档结构与某标准分类的相似程度越高，最大值为1；反之说明该结构的相似度越小，如果小于相应的阈值，说明待查的段落不符合某标准的结构分类，需要继续判断是否符合其他的结构分类，从而实现完整文档结构的识别功能。

**6实验结果分析**

本文的实验环境为：系统:windows XP，开发工具:Visual Studio 2008，所用的开发语言为C#；首先，需要定义文档结构的标准格式、对各文档结构分配向量权重、向量夹角阈值，并对各文档结构进行归一化、加权处理，得到标准结构向量，然后，向系统提交编辑好的硕士学位论文，通过程序自动读取该论文各论文结构的格式信息并映射为向量，对该向量进行归一化、加权处理后得到待查结构向量；最后，计算标准结构向量和待查结构向量的夹角并和夹角阈值进行比较，并将识别结果输出到系统界面。

本文对文档结构的识别结果作如下统计，具体见表1，在该表中文档结构向量权重值对

表 1 文档结构识别结果统计表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **文档结构** | **文档结构向量权重** | **向量夹角阈值** | **正确识别文档数** | **识别率** |
| 摘要标题 | （0.1,0.1,0.1,0.1,0.6） | 0.87 | 360 | 97.2% |
| 摘要正文 | （0.1,0.2,0.1,0.1,0.5） | 0.92 | 352 | 95.1% |
| 关键词 | （0.2,0.3,0.05,0.05,0.4） | 0.90 | 348 | 94.1% |
| 正文章标题 | （0.1,0.2,0.1,0.1,0.1,0.1,0.3） | 0.94 | 350 | 94.5% |
| 论文正文 | （0.15,0.2,0.05,0.05,0.2,0.1,0.25） | 0.89 | 356 | 96.2% |
| …. | … | … | … | … |

应每个文档结构的特征项，本文并未对所有文档结构一一列举，选取五个有代表性的文档结构来说明问题，其中摘要标题对应的特征项为（字体，字号，对齐方式，首行缩进值，关键字），论文正文章标题对应的特征项为（字体，字号，字形，对齐方式，首行缩进值，段后值，关键字），其他文档结构与此类似。通过计算待检查文档结构和标准文档结构向量的夹角值，统计大于向量夹角阈值的文档数，并计算出识别率。可以看出影响识别率的因素主要是文档结构向量权重和向量夹角阈值的取值，本文目前采用的是统计方法分配文档结构向量权重值，并在较差编辑习惯情况下计算出向量夹角阈值。

**7 结束语**

一篇学位论文通常包含论文标题、中文摘要、章节正文等20个左右的单元结构，准确识别这些单元结构，才能在此基础上进行论文格式检查等工作。衡量识别成功与否的指标是识别的准确率，即所识别的结构与正确的结构分类相符的比例。本文利用上述方法对北京信息科技大学硕士学位论文中的370篇样本进行了测试，统计结果如上表1所示，通过计算文档结构识别率的平均值，可以得到本方法对学位论文单元结构识别的准确率达到了95.3%，达到了实验结果预期目的。

准确识别文档结构能够为进一步检查论文的格式正确与否提供基础，此外，本文的方法也可以应用到其他需要判定文档结构的领域（如文档理解），具有较好的应用前景。

本文的方法也存在不足之处，如为确定特征向量权重而采用的统计方法计算量大、与文档类型有较强的依赖性，在今后的研究中可尝试通过样本学习，自动计算权重，减轻工作量，并进一步提高结构识别的准确率。此外，本文并未深入讨论上下文结构对文档结构识别的贡献，将在今后进一步探讨。

**参考文献：**

[1]ISO/IEC 26300:2006 Information Technology -Open Document Format for Office Applications(OpenDocument)v1.0[S].German,OASIS,2006

[2]GB/T 20916-2007 中文办公软件文档格式规范[S].北京:中国标准出版社, 2007

[3] ISO/IEC 29500:2006 Information Technology --Document description and processing languages-Office Open XML File Formats--Parl:4[S],UK, ISO/IEC JTC1,2008

[4]徐东风,彭红星,廖俊杰. 基于Java的文档格式检查技术的研究及其应用[J].计算机工程与设计,2010,31(19):4309-4315

[5]陈国胜,何宗明.基于XML技术的Word文档录入及格式检测系统设计[J].计算机时代，2009(4)：35-37

[6]张卫丰,周国强. 毕业设计文档管理技术探索及实践[J].教育与教学研究,2009(11):58-67.

[7]林雪云. 基于论文格式智能检查系统的研究[J].电脑知识与技术,2009(11):9433-9434

[8]Salton G, Wong A, Yang CS. A vector space model for automatic indexing [J]. Communications of the ACM ,1995, 18( 11) : 613-620.

[9]Sahon G. The SMART retrieval system-experiments in automatic document progressing[M].Englewood Clifs, NJ :Prentice Hall Inc, 1971:313-323

[10]庞剑锋,卜东波,白硕.基于向量空间模的文本自动分类系统的研究与实现[J].计算机应用研究,2001 (9):23 -26

[11]武旭,须德.基于向量空间模型的文本自动分类系统的研究与实现[J].北方交通大学学报，2003(4):38-41

[12]孙挺,耿国华,周明全. 一种有效的特征权重计算方法[J].郑州大学学报,2008(12):48-51

[13]Xiao HG, Cai CZ. Comparison study of normalization of feature vector[J]. Computer Engineering and Applications, 2009(22):117-119

[14]Zhan Y, Chen H, Hang GC. An optimization algorithm of K-NN classifier[C]// FITME.Proceedings of the Fifth International Conference on Machine Learning and Cybernetics. Dalian, China：2006(8):2246-2251.

[15]Wang XZ, Wang YD,Wang LJ.Improving fuzzy c-meg,n8 clustering based on feature-weight learning[J]. Pattern Recognition Letters,2004(10): 1123-1132

[16]李洁，高新波，焦李成．基于特征加权的模糊聚类新算法[J]．电子学报,2006,34(1)：89-92．

[17]Salton G. Automatic Text Processing : The transformation , analysis , and retrieval of information by computer[M] . Boston USA ,Addison-Wesley, 1989.