论文抄袭检测中特征选择①

Feature Selection in Plagiarism Detection of Academic Dissertation

赵俊杰 (安徽财经大学 成人教育学院 安徽 蚌埠 233061)

摘 要: 选取多少个最佳特征以及采用什么评估函数，针对不同的问题选取策略也有所不同。针对论文抄袭检测问题，如何确定特征选择数量和选择评估函数是文章研究的目的。在分析论文抄袭的主要形式和手段基础上，针对文本内容抄袭，阐述了文本特征表示的主要方法和特征选择常用策略，最后对实验结果进行分析，得出基本结论。

关键词: 抄袭检测 特征选择 文本表示 向量空间模型

1 引言

论文抄袭的形式和手段多种多样，具体包括抄袭观点、文字、图像、表格、数据、模型与公式等。在所曝光的案例中，文本抄袭的情况占大多数，因此本文主要研究论文抄袭中文本内容抄袭的检测问题。从一篇文档的语法层次来看，它是由词、短语、句子和段落所构成的。所以，这些要素都可以作为文档的特征。但一般情况下，基于句子和段落层次的文本特征表示应用不多，常用的文档特征有词和短语。由于词和短语的数量太大，直接比较效率太低；且词语之间存在一定的关系，不同的词语所占的权重也不同，因此，文本特征选择的策略显得十分重要。

2 文本的特征表示

文本表示是把半结构化或非结构化的文本数据转换为可供计算机处理的机构化数据[1]。所谓特征表示就是以一定的特征项(如词条或描述)来代表文本信息，特征表示模型有多种，常用的有布尔逻辑型、向量空间型和概率型等。下面简单介绍这三种模型。

2.1 布尔模型

布尔模型[2]是以集合论和布尔代数为理论基础的一个非常简单的检索模型，它基于特征项的严格匹配。它用关键字组合来表示文本信息，关键词的权重为布尔变量，如果某关键字在文本中出现，其取值为 1，否则为 0。用户查询表示为逻辑运算符(与、或、非)连接起来的布尔表达式，用检索状态值(RSV)来度量文档和用户查询之间的相似度，文档与查询的匹配规则遵循布尔运算的法则。如果查询式的值为 1，择 RSV 值为 1，否则为 0。所有 RSV 为 1 的文档与查询式相关，所有 RSV 为 0 的文档则与查询式不相关，因此布尔模型是基于二值评价体系的。

2.2 向量空间模型

向量空间模型(VSM)即使用向量表示文本。在向量空间模型中，文本的内容由一些特征来表达，一般由文本所含有的基本语言单位(字、词、词组或短语)来表示，即文本可以表示为 $Document=D(t_1, t_2, \cdots, t_n)$，其中 t_i 表示各个项，都被赋予一个权重 W，以表示这个特征项在该文本中的重要程度，权重一般都以特征项的频率为基础进行计算的。目前，计算主要采用 TF-IDF 公式，其中 TF 是特征项在文本中的绝对频率，IDF 表示特征项在文本中的文本内频数。这样文本就可以表示为：$(t_1,w_1;t_2,w_2;\cdots;t_i,w_i;\cdots;t_n,w_n)$，可以简记为 $D=D(w_1,w_2,\cdots,w_n)$。两个文本 D_1 和 D_2 之间的相关程度常用它们的相似度 $SIM(D_1, D_2)$来度量。在向量空间模型下，一般借助向量之间的某种距离来表示文本间的相似度。向量空间模型是最简便、最高效的文本表示模型之一，本文的研究即采用向量空间模型。

① 基金项目:教育部社科研究基金青年项目(07JC870006);安徽财经大学教研重点项目(ACJYZD200914)
收稿时间:2009-02-19

2.3 概率模型

布尔模型和向量空间模型都假设关键词之间是相互独立(即相互正交)的，这与实际情况不符。Robertson 和 Spark Jones 提出的概率模型[3]则考虑了关键词之间、关键词和文档之间内在联系，以贝叶斯为理论基础，利用它们的概率相依性进行信息检索。概率模型基于提问关键词在相关和不相关文档中的分布。这是采用关键词的权重来表示的，这样每个查询的文档就按照符合提问的关键词权重之和进行排序。常用的二值独立检索模型是一种实现简单并且效果较好的概率模型。

3 特征选择策略

构成文本的词汇，数量是相当大的，因此表示文本的向量空间的维数也相当大，可以达到几万维，因此我们需要进行维数压缩的工作。目前对文档特征所采用的特征子集选取算法一般是构造一个评价函数，对特征集中的每一个特征进行独立的评估，这样每个特征都获得一个评估分，然后对所有的特征按照评估分的大小排序，选取预定数目的最佳特征作为结果的特征子集。一般采用的评估函数有信息增益、互信息、期望交叉熵、x^2 统计、文本证据权、文档频次和几率比等。

选取多少个最佳特征以及采用什么评估函数，针对不同的问题选取策略也有所不同。本文针对论文抄袭检测问题，讨论特征选择的数量和评估函数的选取策略。下面先给出常见的几种特征选择策略：

3.1 信息增益

信息增益[4]在机器学习中经常被用作特征词评判的标准，它是一个基于熵的评估方法，涉及较多的数学理论和复杂的熵理论公式，定义为某特征在文档中出现前后的信息熵之差。根据训练数据，计算出各个特征词的信息增益，删除信息增益很小的词，其余的按照信息增益从大到小排序。信息增益评估函数被定义为：

$$IG(w)=\sum_{i=1}^{m}p(c_i)\log_2\left(\frac{1}{p(c_i)}\right)-p(w)\sum_{i=1}^{m}p(c_i\wedge w)\log_2\left(\frac{1}{p(c_i\wedge w)}\right)-p(\overline{w})\sum_{i=1}^{m}p(c_i\wedge\overline{w})\log_2\left(\frac{1}{p(c_i\wedge\overline{w})}\right) \quad (1)$$

$\{c_i\}_{i=1}^{m}$ 表示目标空间的类集 c，w 为特征词条，其中 p(w)为词条出现的概率，$\overline{w}$ 表示词条 w 不出现，$p(c_i)$ 为 i 类值的出现概率，$p(c_i\wedge w)$为词条出现时属于第 i 类的条件概率。

3.2 互信息

互信息[5]是普遍应用在相关词统计语言建模中，如果用 A 来表示词条 t 且属于类别 c 的文档频率，B 表示为包含词条 t 但是不属于类别 c 的文档频率，C 表示属于类别 c 但是不含词条 t 的文档频率，N 表示整个训练语料库中的文档总数，词条 t 与类别 c 之间的互信息可以下面公式计算：

$$MI(c,t)=\log\frac{A\times N}{(A+C)\times(A+B)} \quad (2)$$

当 t 与 c 相互独立时，MI(c,t)自然为 0。如果训练集有 m 个类，对于每个词条 W 都有 m 个互信息量，取它们的最大值作为每个词条的全局互信息量，然后将全局互信息值进行排序，将低于设定阈值的词条从原始特征空间中移除，保留高于阈值的词条构成特征空间，从而降低了特征空间的维数。

3.3 期望交叉熵

期望交叉熵[6]和信息增益相似，也是一种基于概率的方法，不同之处在于信息增益中同时考虑了特征在文本中发生与不发生时的两种情况，而期望交叉熵只考虑在文本中发生一种情况。对于特征 f，其期望交叉熵记为 CE(f)，计算公式如下：

$$CE(f)=\sum_{c\in C}p(c,f)\log\left(\frac{p(c,f)}{p(c)p(f)}\right) \quad (3)$$

3.4 x^2 统计

X^2 统计方法[7]度量词条与文档类别之间的相关程度，并假设词条与类别之间符合具有一阶自由度的 X^2 分布。词条对于某个类别的 X^2 统计量越高，表明它与该类之间的相关性越大，所携带的类别信息也就越多。令 A 表示属于类别 c 且包含词条 w 的文档频率，B 表

示不属于类别 c 但包含词条 w 的文档频率，C 表示属于类别 c 但不包含 w 的文档频率，D 表示既不属于类别 c 也不包含词条 w 的文档频率，则词条 W 对于类别 c 的 X^2 统计量由下列式子计算：

$$\chi^2(C,W)=\frac{(AD-CB)^2\times(A+B+C+D)}{(A+C)\times(B+D)\times(A+B)\times(C+D)} \quad (4)$$

3.5 文档频次

文档频次[8]是指有该词条出现的文档数量。在训练文本集中对每个词条计算它的文档频次，并且剔除在特征空间中文档频次小于预先定义的阈值的词条。文档词频是缩减词条的最简单的方法。它通过在训练文档数量中计算线性近似复杂度来衡量巨大的文档集，该方法一般不直接使用，而把它作为评价其他评估函数的标准。

4 实验与分析

4.1 实验设计与结果

本文选取了 2153 个文本文件，共 14 种类别，包括：计算机、医药、经济、环境、军事、艺术、体育、教育、交通、政治、建筑、金融、佛学和电子商务等，作为训练集。本文设计了一篇抄袭文档作为测试样例，它抄袭了类 140 篇中的 3 篇文档，抄袭比例分别为 20%、30%和 50%左右，主要目的是检验抄袭比例对检索结果的影响。其中，特征维数分别选取为 800、1000、1500、2000、2500、3000 六种情况。特征函数选取信息增益、互信息、期望交叉熵和 X^2 统计四种形式。权重的计算采用 TFIDF 公式，其中 TF 是特征项在文本中的绝对频率，而 IDF 表示特征项在文本中的文本内频数。

对于抄袭论文和任一篇待查论文的特征词及权重值，通过余弦公式计算其相似度。余弦公式如下：

$$Sim(Vi,Dj)=\cos\theta=\frac{\sum_{k=1}^{m}v_{ik}\cdot d_{ik}}{\sqrt{\sum_{k=1}^{m}v^2ik}\cdot\sqrt{\sum_{k=1}^{m}d^2{}_{ik}}} \quad (5)$$

下面给出实验的结果，表 1、表 2 和表 3 是抄袭比例在 20%、30%和 50%时，不同特征数与特征函数下检测到得相似文档数。

表 1　抄袭比例在 20%时不同情况下检测到得相似文档数

特征数 / 特征函数	800	1000	1500	2000	2500	3000
信息增益	39	37	39	46	44	40
互信息	27	41	43	48	47	47
期望交叉熵	39	37	39	46	44	40
x^2统计	38	38	43	43	42	43

表 2　抄袭比例在 30%时不同情况下检测到得相似文档数

特征数 / 特征函数	800	1000	1500	2000	2500	3000
信息增益	34	26	30	29	30	32
互信息	21	31	31	33	32	33
期望交叉熵	34	27	32	34	33	32
x^2统计	33	27	31	30	31	31

表 3　抄袭比例在 50%时不同情况下检测到得相似文档数

特征数 / 特征函数	800	1000	1500	2000	2500	3000
信息增益	4	3	2	2	2	2
互信息	1	1	1	3	2	2
期望交叉熵	4	3	2	2	2	2
x^2统计	4	3	2	2	2	2

4.2 基本结论

从以上结果可以看出，特征数在 1000–1500 时，检测效果较好；在实验的特征函数中，信息增益、期

(下转第 126 页)

(上接第 103 页)

望交叉熵和统计效果差不多，其中信息增益稍好一些，互信息效果不太稳定。

当抄袭比例较高时，例如超过 50%，查准率较高，误差较小；而当抄袭比例较低时，查准率也较低，误差较大。因此，这种基于全文特征表示的抄袭检测方法，对于抄袭比例较低的情况，只能粗略的检测。由于检测到的文档较多，所以通常需要进一步精确比较或人工判断。

5 结语

论文抄袭检测的效果和文本特征表示方法以及特征选择策略密切相关，而文本特征表示和特征选择和语料库的选取也有很大关系。目前，国内还没有一个统一的、标准的和开放的中文的语料库可供使用，本文选取的语料库来源于网上收集两千多篇新闻和文章，显然语料库较小，这对于测试结果也有一定的影响。

参考文献

1 郝春风,王忠民.一种用于大规模文本分类的特征表示方法.计算机工程与应用, 2007,(34):170－172.
2 程泽凯,陆小艺.文本分类中的特征选择方法.安徽工业大学学报, 2003,7:220－224.
3 刘丽珍,宋瀚涛.文本分类中的特征选取.计算机工程, 2004,2:14－15,175.
4 王玉玲,王娟.文本分类中的特征选取算法.孝感学院学报, 2003,1:35-37,110－112.
5 谢飞.基于聚类分析的文本分类研究[硕士学位论文].合肥:合肥工业大学, 2007.
6 李荣陆.文本分类及其相关技术研究[博士学位论文].上海:复旦大学, 2005.
7 伍建军,康耀红.文本分类中特征选择方法的比较和改进.郑州大学学报, 2007,39(2):110－113.
8 李凡,鲁明羽,陆玉昌.关于文本特征抽取新方法的研究.清华大学学报(自然科学版), 2001,7:98－101.