·人工智能及识别技术· 文章编号：1000—3428(2013)05—0230—05 文献标识码：A 中图分类号：TP311

基于语言模型和特征分类的抄袭判定

李 惠，刘 颖
(清华大学中国语言文学系，北京 100084)

摘 要：信息时代作者版权的保护问题已受到越来越多的关注。针对部分小说存在的文本大面积相似问题，提出基于语言模型和特征分类的方法。统计文本二元~六元的语言模型并且绘制拓扑图，通过计算重合概率和词性比来分析词语的重合程度和语法信息，在此基础上利用主成分分析和随机森林的方法，进行分类判别。机器学习的结果表明，该方法能够有效地鉴别小说是否存在抄袭现象。

关键词：抄袭判定；语言模型；语法信息；主成分分析；随机森林；分类

Plagiarism Judgment Based on Language Model and Feature Classification

LI Hui, LIU Ying
(Department of Chinese Language and Literature, Tsinghua University, Beijing 100084, China)

【Abstract】The protection of copyright property arouses much attention in the present information age. Aiming at the dispute problem caused by the text similarity between some novels, this paper proposes a method based on language model and feature classification, with statistics of coincidences and the proportion of pos to analyze the grammatical collocations and the coincidences. The methods of Principal Component Analysis(PCA) and Random Forest(RF) used to extract characteristics for automatic classification are added into experiments. The result of machine learning shows that the method can effectively identify whether novels exist plagiarism phenomenon.

【Key words】plagiarism judgment; language model; grammatical information; Principal Component Analysis(PCA); random forest; classification
DOI: 10.3969/j.issn.1000-3428.2013.05.051

1 概述

2003 年 12 月，庄羽向北京市第一中级人民法院起诉，称郭敬明所著《梦里花落知多少》一书剽窃了自己的作品《圈里圈外》。随后，法院做出一审判决，认定《梦里花落知多少》剽窃了《圈里圈外》中“具有独创性的人物关系”，造成《梦里花落知多少》与《圈里圈外》“整体上构成实质性相似”。郭敬明不满上诉。北京市高级人民法院审理后认定，“《梦里花落知多少》中有 12 个主要情节与《圈里圈外》雷同，在一般情节和语句上共有 57 处雷同，侵犯了原告的著作权”。

关于文本抄袭的检测，现在常用的有 2 种方法：(1)数字指纹[1]，针对待比较数据，提供多个级别的索引，每个级别的索引采用不同的指纹特征，通过对比样本库匹配指纹进行抄袭检测，比如知网的学术不端文献检测系统[2]，适合大规模文本计算；(2)词频统计，主要是依靠向量空间模型来实现[3-4]。在此模型中，每个文本表示为一个特征向量(字、词、句组成)计算特征向量间的相似度，此方法比较适用于英文，中文文本的识别率并不高。

文献[5]对句子提取关键词，并排序重构，提出编码和词频结合的方法获取句子指纹，以此计算文本间的相似度，检测速度较快，但对于拼凑成的文章识别精度不高。文献[6-7]在分析论文篇章结构的基础上，结合了数字指纹和词频统计，用相关度代表相似度，筛选论文，召回率较高，但识别时间较长。文献[8-9]基于文本分类的思想，进行全篇相似度计算，对结果进行精确比较，再进行段落相似度计算和语句相似度的计算，较为实用，但效率还待进一步提高。

本文结合语言学和文本分类的相关知识，提出一种基于语言模型和特征分类的方法。

基金项目：国家自然科学基金资助项目“基于语用信息的交互行为与语言特征的建模研究”(61171114)
作者简介：李 惠(1987－)，女，硕士研究生，主研方向：计算语言学；刘 颖，副教授
收稿日期：2012-05-09 **修回日期：**2012-08-06 **E-mail：**lh9743@126.com

2　语料预处理及模型

2.1　语料预处理

选取郭敬明的《梦里花落知多少》和庄羽的《圈里圈外》作为主要研究对象，同时选取了郭敬明的另一篇和《梦里花落知多少》风格相似的《悲伤逆流成河》，以及庄羽的《地久天长》这 2 篇小说作为辅助判定，如表 1 所示。

表 1　选取的语料

小说	作者	出版社	出版时间	字数	来源
梦里花落知多少	郭敬明	春风文艺出版社	2003 年 11 月	150 002	
悲伤逆流成河	郭敬明	长江文艺出版社	2007 年 5 月	115 833	新浪爱问电子书
地久天长	庄羽	文汇出版社	2009 年	60 123	
圈里圈外	庄羽	中国文联出版社	2003 年 2 月	180 109	

同时选取了 R 语言、MLTP、Text Forever 这些语料处理工具来实现语料分词、批量文本选取、语言模型统计、词语搭配分析，以及文本分类。

R 语言:R 是用于统计分析绘图的语言和操作环境的软件(http://www.r-project.org)。

MLTP：多语言文本统计分析工具(Multi-lingual Text Processor)，可分析 3 种语言(中文、英文、日语)的句长、词长、词频、词性等，与 R 语言结合使用。

TextForever：TXT 处理软件。在这里运用该软件将每篇文章分成 2 000 字~3 000 字的小文本(http://www.duote.com/soft/7626.html)。

2.2　语言模型

本文选择 *N* 元语言模型(*N*-Gram Model)。*N*-Gram 指的是由 *N* 个词组成的序列。当 *N*=1 时，称为 1 元语法，相当于词频表，给出所有词出现的频率；当 *N*=2 时，称为 2 元语法，相当于一个转移矩阵，给出每一个词后面出现另一个词的概率；当 *N*=3 时，称为 3 元语法，相当于一个三维转移矩阵，给出每一个词对(连续 2 个词)后面出现另一个词的概率。*N*-Gram 的目的就是建立一个给定词序列在语言中的出现概率分布。

基于 *N* 元模型，进行了 2 组实验。第 1 组，分别统计《圈里圈外》和《梦里花落知多少》二元~六元的语言模型，取每个语言模型前 100 个出现的相同词对制成如图 1 所示的模型。其中，横轴自左至右代表了这 100 个词对，纵轴代表了它们在 2 篇文章中分别出现的次数，左边对应的是《圈里圈外》的数量刻度，右边倒转对应的是《梦里花落知多少》的数量刻度。

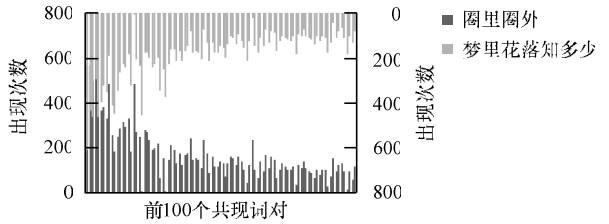

图 1　二元语法

以二元模型做例子，本文一共抽取了 2 篇小说二元词对 54 444 个,其中频率至少为 1 的相同词对共有 16 130 个。图 1 表示 2 篇文章中前 100 个共同出现且完全重合的词对的概率分布。可以发现，100 个词对在 2 部小说中的出现频率总体变化规律非常相近，大部分词对频率比较接近。比如,“小北”是这 100 个词对中出现次数由高至低的第 3 个，作为人名，在《圈里圈外》中出现了 496 次，《梦里花落知多少》中出现了 461 次，数量非常接近。“北京”是这 100 个词对中的第 75 个，《圈里圈外》出现 105 次，《梦里花落知多少》出现 119 次。于是又统计了三元~六元模型前 100 个出现的相同词串，进一步佐证，如图 2~图 5 所示。

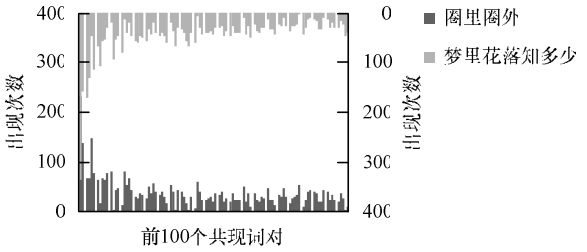

图 2　三元模型

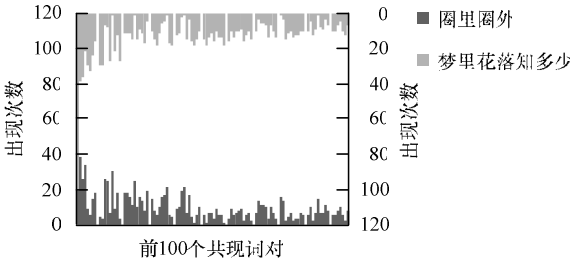

图 3　四元模型

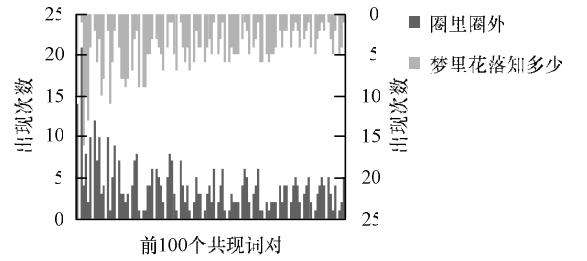

图 4　五元模型

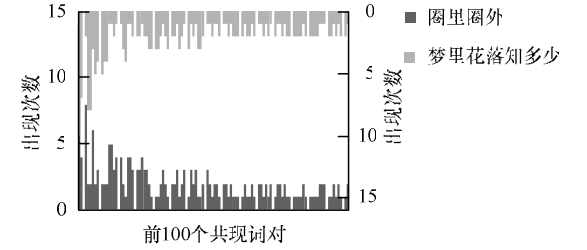

图 5　六元模型

随着元数的增加，2 篇文章的相同词对越来越少，二元模型为 16 130 个，三元模型为 14 507 个，四元模型为 5 845 个，五元模型为 1 726 个，六元模型为 481 个，从万级别递减到百位级，因此，统计到六元模型，已经足够本文进行 2 篇文章的词语和词频相似程度的分析。

图 2~图 5 中词对的频率变化趋势也非常相似，统计的词串中也出现了许多雷同的带有作者风格和主观色彩的词语序列。例如：三元模型中出现的“狗脾气”、“撒丫子”、“胡汉三”，都是带有俚语色彩的词语；四元模型中的“大尾巴狼”、“吧嗒吧嗒”、“祸害人民”，都是体现作者风格的调侃性词语；五元模型中的“纯洁的男女，跟朵花似的”，“防盗防记者”；六元模型中的“怕什么来什么”，“电梯直接入户”，“怎么着怎么着”，这些带着京味的词语，能够运用自如的话，应该要有北方生活的阅历。通过对特色词语的分析，2 篇小说的重叠语对用“巧合”来说比较牵强，因此，第 2 组实验引入了重合概率，目的是为了计算《梦里花落知多少》和《圈里圈外》所有二元词串到六元词串的重合程度：

重合概率=重合语对/所在 N 元模型中出现的所有语对

同时也分别计算了《梦里花落知多少》和庄羽的《地久天长》的重合概率、《圈里圈外》和郭敬明的《悲伤逆流成河》的重合概率，以及这 2 位作者各自的 2 篇文章的重合概率。图 6 清晰地反映出它们的关系并且列举了重合概率的具体数值。

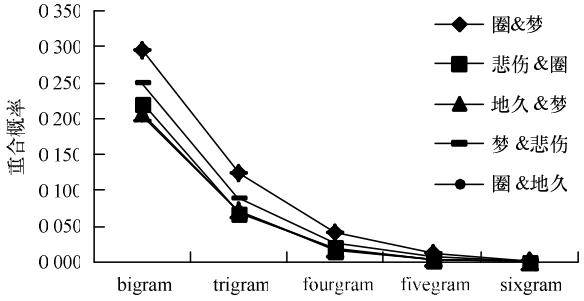

图 6 语言模型的重合概率

郭敬明的《梦里花落知多少》和庄羽的《圈里圈外》的二元词对到六元词串的重合概率最高，比郭敬明本人写的 2 篇小说都高(二元高 5%，三元高将近 4%，四元高 2%)，比庄羽本人写的 2 篇小说的重合概率高(二元高 9%，三元高 5%，四元高 2%)。

而郭敬明的另一篇小说《悲伤逆流成河》和庄羽的《圈里圈外》的二元词对到六元词串重合概率远低于《梦里花落知多少》和庄羽的《圈里圈外》的重合概率。

任何 2 部小说二元到六元的重合概率逐渐降低，说明字符串越长，重复率越低。

3 语法分析

抽取了《圈里圈外》和《梦里花落知多少》中的所有名词，统计二元语言模型，进行文本的结构特征分析，利用 R 语言的程序包，制成网络拓扑图。将小说中具体出现的名词抽象成“结点”，将它们的连接关系抽象成“边”，就构建了一张如图 7 所示的“网络”。如果 2 个词在二元模型中以词对出现，则产生一条有向边相连。边的指向代表了词对的先后顺序。

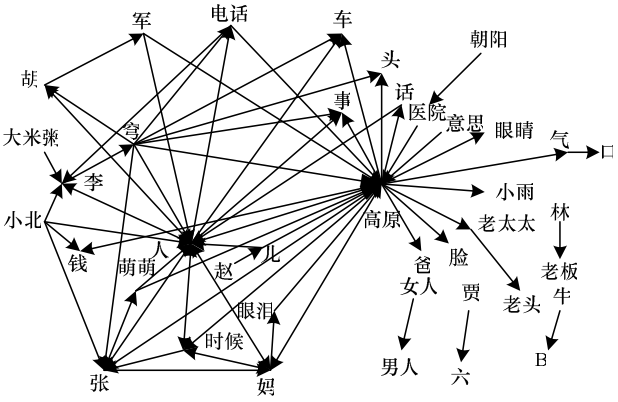

图 7 《圈里圈外》的网络拓扑图

根据拓扑图分别选取了两文中主要出现人物的姓名搭配，进行对比分析，发现 2 篇小说在人物姓名选择上有很多雷同：

(1)“张萌萌”和“姚姗姗”，名字的结构相似，在两篇小说中都是插足女主角爱情最后成为明星的人物。“高原”和“顾小北”同样也存在相似：“萌萌”、“高原”在《圈里圈外》的搭配次数和“姗姗”、“顾”在《梦里花落知多少》里的一样多，出现 15 次。

(2)“小北”这个名字在 2 部小说中都出现了，只不过《圈里圈外》是“张小北”，《梦里花落知多少》是“顾小北”。“小北”、“时候”在《圈里圈外》出现 15 次，《梦里花落知多少》出现 12 次。“小北”、“人”在《圈里圈外》出现 10 次，《梦里花落知多少》出现 12 次，“小北”、“张”在《圈里圈外》出现 17 次，“小北”、“顾”在《梦里花落知多少》出现 17 次。

对每篇文本计算特定语言结构在文本中的频率，可以帮助辨别不同作者的语言风格。由此，统计了这 4 篇小说中出现次数最多的 10 个词性(动词/v，名词/n，代词/r，助词/u，副词/d，介词/p，形容词/a，人名/nr，数词/m，方位词/f)占总词性的百分比以及其中的名词出现次数分别和 9 个词性出现次数的比例，绘制了如图 8、图 9 所示的频率曲线。

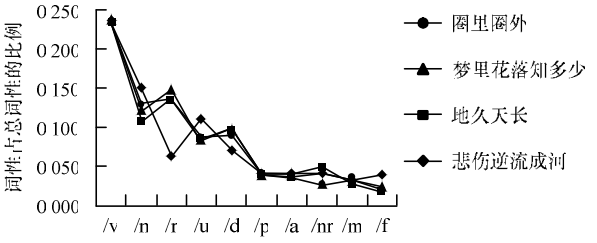

图 8 语法分析

在图 8 中，代词占总词数之比(/r)在郭敬明的《悲伤逆流成河》中最低(比《梦里花落知多少》低近 8%，比庄羽的 2 篇文章都低近 6%)。人名(/nr)占总词数之比，庄羽的《地久天长》比其他 3 篇文章都高。综合看这 10 个词性占总词性的比例，只有《圈里圈外》和《梦里花落知多少》一直保持着非常相近的百分比和图线变化趋势。

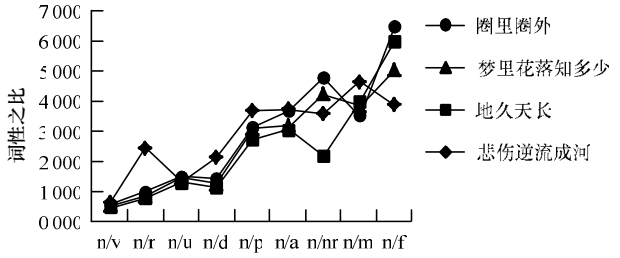

图 9　名词与 9 大词性之比

在图 9 中，名词和副词之比(n/d)，郭敬明的《悲伤逆流成河》比其他 3 篇文章都高(比《圈里圈外》高近 0.7，比《梦里花落知多少》高近 0.9，比《地久天长》高 1.03)。名词和介词(n/p)之比，庄羽的《地久天长》最低。也只有《圈里圈外》和《梦里花落知多少》一直保持着非常相似的比率和曲线变化规律。

4　主成分原理及分析

主成分分析是一种寻找综合变量的方法。对于很多变量的数据(高维)，利用变量之间的相关关系，通过线性变换，尽量多保持原来信息的条件下压缩(降维)到较少的无相关变量的数据降维方法。

4.1　主成分原理

假设有 n 个随机变量，$X_1,X_2,\cdots,X_n$，样本均数为 $\bar{X}_1,\bar{X}_2,\cdots,\bar{X}_n$，样本标准差为 $S_1,S_2,\cdots,S_n$，首先做标准化变换 $X_i=(X_i-\bar{X}_i)S_i,\ i=1,2,\cdots,n$。

所有线性组合 $X_1,X_2,\cdots,X_n$ 的相关矩阵为 $\boldsymbol{R}$，$(a_{i1},a_{i2},\cdots,a_{in})$则是相关矩阵 $\boldsymbol{R}$ 的第 i 个特征向量[10]。而且，每个特征值就是每个主成分的方差，即 $Var(C_i)$。

在 $C_1=a_{11}x_1+a_{12}x_2+\cdots+a_{1n}x_n$ 中，组合系数$(a_{i1},a_{i2},\cdots,a_{in})$构成的向量是单位向量，限定 $a^2_{11}+a^2_{12}+\cdots+a^2_{1n}=1$，使得 $Var(C_1)$最大，C_1 便称为第 1 主成分。

类似地，可以定义第 2、第 3……直到第 n 个主成分。

利用主成分分析的方法，对原始的文本变量计算主成分，提取前若干个主成分，计算每个主成分的方差和贡献率(该主成分方差/所有主成分的方差之和)，构造评价函数 F，计算每个主成分的综合得分，代表原始的文本。$F=\alpha_1y_1+\alpha_2y_2+\cdots+\alpha_ny_n$。$\alpha_n$ 代表每个主成分的贡献率；n 代表选取的主成分的数量；y_n 代表第 n 个主成分。

4.2　主成分分析

对选取的 4 篇文章《圈里圈外》和《梦里花落知多少》，《悲伤逆流成河》和《地久天长》都进行了预处理，将它们分别分成 2 000 字~3 000 字的小文本:《梦里花落知多少》51 篇，《圈里圈外》51 篇，《地久天长》31 篇，《悲伤逆流成河》43 篇。

首先对《圈里圈外》和《梦里花落知多少》一共 102 个文本中的所有词性进行统计和主成分分析，并基于此进行分类。取了 2 个主成分 PC_1 和 PC_2，它们的方差贡献率分别为 79.4%和 4.6%，累积贡献率为 84.0%，代表了原样本 84%的信息，基本覆盖了原来 102 个变量的大部分信息，达到了降维的要求。因此，选取这 2 个主成分来计算主成分得分，进而分析 2 位作者的文本分类。

▲和○分别代表 2 类文本，每一个具体的点分别表示由主成分 PC_1 和主成分 PC_2 反映的 102 个文本变量，每个变量对应的 PC_1 和 PC_2 的数值作为横纵坐标的数值。又分别选取了 2 位作者的另外 2 篇文章加入，共对语料进行 5 次实验(包括 2 位作者自身 2 篇文章的分类)，进行两两比较的主成分分析。实验结果清晰表示《圈里圈外》和《地久天长》、《梦里花落知多少》和《悲伤逆流成河》、《梦里花落知多少》和《地久天长》(见图 10)等分类都很清晰，选取《悲伤逆流成河》和《圈里圈外》作为示例，如图 11 所示，符号混杂，无法分类；符号分散，并无混杂。由此有理由怀疑郭敬明的《梦里花落知多少》的原创性。

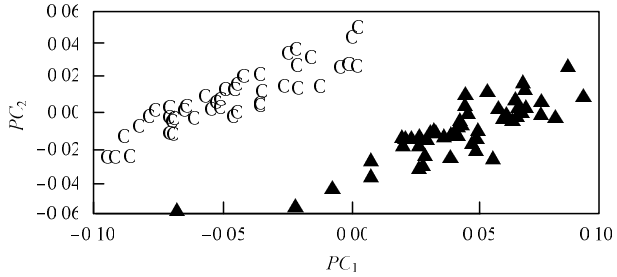

图 10　《悲伤逆流成河》和《圈里圈外》的主成分得分

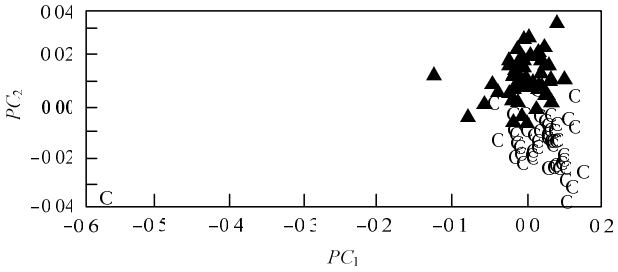

图 11　《梦里花落知多少》和《圈里圈外》的主成分得分

5　随机森林

随机森林通过 Bagging 抽样的方法生成训练样本，由训练样本生成多个决策树，组成森林，决策树之间互不相关。随机森林是一个树型分类器$\{h(\boldsymbol{x},\boldsymbol{\beta k}),\ k=1,2,\cdots\}$的集合[11]。其中，元分类器 $h(\boldsymbol{x},\boldsymbol{\beta k})$是用 CART 算法构建的没有剪枝的分类回归树；$\boldsymbol{x}$ 是输入向量；$\boldsymbol{\beta k}$ 是独立同分布的随机向量，决定了单棵树的生长过程；假设共有 M 个属性，指定一个属性数 $F\leqslant M$，在每个内部结点，从 M 个属性中随机抽取 F 个属性作分裂属性集，以这 F 个属性上最好的分裂方式对结点进行分裂(在整个森林的生长过程中，F 的值一般维持不变)，每棵树任其生长，不进行剪枝。

森林的输出采用简单多数投票法。每一棵决策树就是一个精通于某一个窄领域的专家，这样在随机森林中就有了很多个精通不同领域的专家，对一个新的输入样本，可以用不同的角度去看待它，最终由各个专家，投票得到结

果。每次抽样生成样本集，全体样本中不在样本集的剩余样本称为 OOB(Out of Bag)数据，每次的预测结果进行汇总得到错误率的 OOB 估计，用于评估组合分类器的正确率。所有树的误分率取平均得到随机森林的 OOB 误分率。OOB 误分率是随机森林泛化误差的一个无偏估计，其结果近似于需要大量计算的 k 折交叉验证[12]。

本文将《梦里花落知多少》单独作为测试语料，郭敬明的另外一篇小说《悲伤逆流成河》和庄羽的《圈里圈外》《地久天长》作为训练语料，制成语言模型。为了保证测试的公平性，将《圈里圈外》作为测试语料，其余 3 篇作为训练语料，同样也进行了测试。

图 12 所示的是构建 500 棵树的随机森林。用《梦里花落知多少》作为测试语料。可以看出，当树的数目达到 200 棵后，分类错误率已基本稳定。

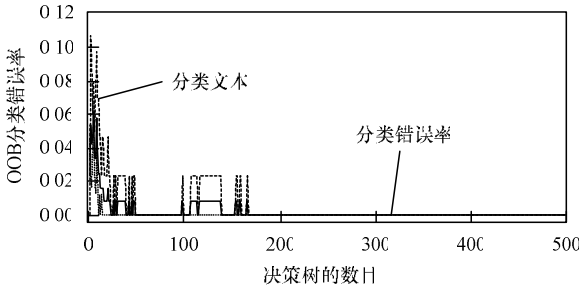

图 12 除《梦里花落知多少》以外的 3 篇文章的分类树描述

这 2 条线最后稳定的纵轴数值为 0，表示样本语料分类全正确。在得到森林之后，用郭敬明的《梦里花落知多少》的 51 个文本进行测试，让森林中的每一棵决策树分别进行判断，看看这每一个文本应该属于哪一类，然后看看哪一类被选择最多，就预测这个样本为哪一类。测试语料的分类结果为，郭敬明的 51 篇文本，1 篇判为郭敬明类，50 篇错判为庄羽类。正确率为 1.961%。

再用庄羽的《圈里圈外》作为测试语料，同样其余 3 篇小说分类完全正确。让机器用分类的结果来判断《圈里圈外》的归属，庄羽的 51 篇文本，41 篇判为庄羽类，10 篇错判为郭敬明类。正确率为 80.392%。《梦里花落知多少》的分类效果比《圈里圈外》差了很多，可见《梦里花落知多少》确实和《圈里圈外》存在相当数量的相似性。

6 结束语

目前已有的抄袭识别方法，错查、漏查的事故屡屡发生。考虑到计算机不能真正理解文本，而很多抄袭并不是原封不动的照抄，准确的抄袭识别很难实现。本文结合语言学和文本分类的相关知识，从多个角度考察判定《梦里花落知多少》和《圈里圈外》是否抄袭。

首先基于语言模型，对选取的语料进行二元到六元的词串重合的分析，并分别给出了高频重合词语序列的统计、频率分布的规律和重合概率的计算。在此基础上，制成了名词的网络拓扑图，分析词语搭配和姓名选择上的相似性；同时结合语法信息，说明了 2 篇文章在词性选择上的变化规律和雷同之处。还引入了特征提取和文本分类的方法：主成分分析，提取特征向量，选取累积贡献率较大的主成分进行得分计算，通过多次实验，判断并验证 2 类文本的特征分布是否清晰；随机森林的方法从决策树判断的角度进行甄别，让计算机自动判断测试语料的类别归属。综合以上的各种方法，实验的结论也证实，《梦里花落知多少》一文确实和《圈里圈外》从用词到文本特征上存在着大量的相似，有理由判断为抄袭。

下一步工作可以将本文的分析方法用于大规模语料的风格分析上。

参考文献

[1] Schleimer S, Wilkersond S. Winnowing: Local Algorithms for Document Fingerprinting[C]//Proc. of International Conference on Management of Data. New York, USA: ACM Press, 2003: 204-212.

[2] 知网. CNKI 学术不端文献检测系统宣传册[EB/OL]. (2012-07-08). http://www.cnki.net.

[3] Monostori K, Zaslavsky A. Match Detect Reveal: Finding Overlapping and Similar Digital Documents[C]//Proc. of International Conference on Information Resources Management Association. [S. l.]: ACM Press, 2000: 955-957.

[4] Finkel R A, Zaslavsky A. Signature Extraction for Overlap Detection in Documents[C]//Proc. of the 25th Australian Computer Science Conference. Melbourne, Australia: Australian Computer Society Inc., 2002: 59-64.

[5] 秦玉平, 冷强奎, 王秀坤, 等. 基于局部词频指纹的论文抄袭检测算法[J]. 计算机工程, 2011, 37(6): 193-194.

[6] 金 博, 史彦军, 滕弘飞. 基于篇章结构相似度的复制检测算法[J]. 大连理工大学学报, 2007, 47(1): 125-130.

[7] 史彦军, 滕弘飞, 金 博. 抄袭论文识别研究与进展[J]. 大连理工大学学报, 2005, 45(1): 50-56.

[8] 赵俊杰. 基于分类思想的论文抄袭判定系统的设计与实现[J]. 数字图书馆论坛, 2008, (11): 73-75.

[9] 赵俊杰, 胡学钢. 一种基于段落词频统计的论文抄袭判定算法[J]. 计算机技术与发展, 2009, 19(4): 231-238.

[10] 李靖华, 郭耀煌. 主成分分析用于多指标评价的方法研究——主成分评价[J]. 管理工程学报, 2002, 16(1): 39-43.

[11] 陆 秋, 程小辉. 基于属性相似度的决策树算法[J]. 计算机工程, 2009, 35(6): 82-84.

[12] 汪 伟, 华 琳, 郑卫英, 等. 基于独立成分分析和随机森林判别法的 Microarray 分析及在分子生物学中的应用[J]. 中国优生与遗传杂志, 2009, 17(8): 8-10.

编辑 顾逸斐