

基于篇章结构相似度的复制检测算法

金 博¹, 史彦军², 滕弘飞^{*2}

(1. 大连理工大学 计算机科学与工程系, 辽宁 大连 116024)

(2. 大连理工大学 机械工程学院, 辽宁 大连 116024)

摘要: 学术论文的复制检测研究对于知识产权保护和抑制抄袭侵权行为有重要意义. 国内外主要用数字指纹及关键词匹配等技术进行论文的复制检测. 为解决目前中文复制检测难题, 给出了一种基于篇章结构相似度的中文学术论文复制检测算法及其问题的数学模型. 在分析论文篇章结构的基础上, 利用数字指纹和词频统计等技术, 经编程实现, 用于论文的全抄、部分抄袭和拼抄等抄袭现象的初步检测. 与基于全文数字指纹和基于全文词频统计的检测方法相比较, 更适用于要求较准确的论文复制检测.

关键词: 学术论文; 复制检测; 抄袭识别; 数字指纹; 词频统计; 篇章结构

中图分类号: TP391.1 **文献标识码:** A

0 引 言

复制检测又称抄袭识别或副本检测, 是实施知识产权保护和提高信息检索效率的一种手段. 随着互联网的发展, 抄袭或非法复制显得更加便捷和隐蔽, 因此学术论文抄袭识别或复制检测问题已引起国内外学者的关注.

论文抄袭一般有如下几种形式: ① 对他人论文全盘抄袭; ② 将他人论文简单调换段落顺序组成自己的论文; ③ 从他人论文中抄袭部分段落; ④ 将几篇论文拼凑组合成自己的论文; ⑤ 窃取他人论文的核心创新思想或内容, 用抄袭者自己的语言表达. 本文主要研究前 4 种.

关于论文的复制检测或抄袭识别, 宋擒豹等提出了 CSDSG 系统^[1-2], 该系统针对数字商品的非法复制和扩散问题提出基于注册的复制监测机制, 用于解决有关复制检测问题. 较早的英文查重及反抄袭程序有 Austin^[3] 开发的 WordCheck 系统, 该系统采用关键词匹配技术, 用于基金申请的查重. Antonio 等^[4] 针对 Latex 格式开发了论文反抄袭系统 CHECK, 该系统对 Latex 格式的论文进行分解, 再利用向量点积法来比较相似度. Heintze^[5] 研究了基于数字指纹技术的抄袭识别

原型 KOALA, 并发布在其网站上, 以便用户免费试用. 美国 UC Berkeley 大学研究人员利用数字指纹技术进行反抄袭研究, 并创办了英文论文反抄袭网站 <http://www.turnitin.com>, 该网站存储了海量的英文论文数据库, 并通过网络随时更新, 为英文反抄袭提供了较好的服务^[2-6].

上述复制检测或抄袭识别系统具有如下特点: ① 主要采用数字指纹、关键词匹配等识别技术; ② 大部分只能处理全盘抄袭的情况, 并且准确率较低; ③ 适用范围较小, 例如 CHECK^[4] 只能处理 Latex 格式. 由此可以看出, 目前复制检测或抄袭识别技术还不完善. 本文研究的中文学术论文的抄袭识别问题, 相对于英文论文抄袭识别来说, 由于需要额外考虑汉语的词切分、词法及语法特点, 难度更大.

关于篇章结构分析的相关研究主要有: 王继成等^[7] 通过 HTML 标记对 Web 文档进行结构分析; 张益民等^[8] 研究了以 RST 分析为基础的 RS 树篇章结构表示形式. 以上的篇章结构分析都是针对文档原有特定标记进行分析处理, 而不是在篇章理解的基础上进行的篇章结构分析.

本文在学术论文理解的基础上, 针对学术论文的特有结构, 对学术论文进行篇章结构分析, 再

收稿日期: 2005-06-10; 修回日期: 2006-12-10.

基金项目: 国家自然科学基金资助项目 (60674078, 50335040, 50575031); 国家“八六三”高技术研究发展计划资助项目 (2006AA04Z109).

作者简介: 金 博 (1978-), 男, 博士生; 滕弘飞* (1936-), 男, 教授, 博士生导师.

通过数字指纹和词频统计等方法计算出学术论文之间的相似度,从而找出抄袭的现象. 本文的研究只针对书写格式规范的学术论文的抄袭现象.

1 问题的数学模型

如图 1 所示,设某篇论文为 d ,其篇章结构可表示为一八元组,数学表示如下:

$$d = \{M, T, U, A, K, Z, P, R\} \quad (1)$$

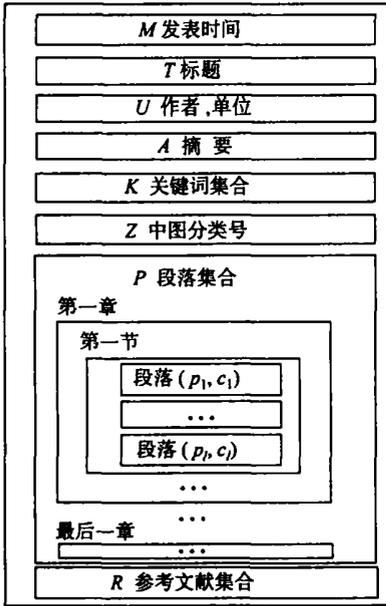


图 1 论文的篇章结构

Fig. 1 Document-structure of a paper

由于一篇文章中关键词、正文段落及参考文献一般不止一个,将这几部分表示为集合形式,数学表示如下:

$$K = \{k_1, k_2, \dots, k_s\} = \{k_i | i \in I, i = 1, 2, \dots, s; 1 \leq s \leq 8\} \quad (2)$$

其中 s 为关键词个数.

$$P = \{(p_1, c_1), (p_2, c_2), \dots, (p_l, c_l)\} = \{(p_i, c_i) | i \in I, i = 1, 2, \dots, l\} \quad (3)$$

其中 l 为段落数, c_1, c_2, \dots, c_l 为段落的特征信息(例如段落 p_i 在 d 中所处位置),两者结合表示一个段落的完整信息.

$$R = \{r_1, r_2, \dots, r_m\} = \{r_i | i \in I, i = 1, 2, \dots, m\} \quad (4)$$

其中 m 为参考文献数量.

对于抄袭识别问题,篇章结构中的 M, T, U, A, K, Z, R 等各部分都具有相应作用,所以在进行抄袭识别时应综合考虑各部分的影响因素.

2 抄袭识别方法

本文所研究的学术论文的全盘抄袭、部分抄袭或拼凑抄袭须同时具备如下几个特征:

- (1) 录用或发表时间不同.
- (2) 标题相类似,其中心词基本相同.
- (3) 作者不同.
- (4) 摘要相同或类似.
- (5) 关键词至少有一个或多个重叠.
- (6) 中图分类号相同或相近.
- (7) 段落中至少有一个或多个相同或类似,甚至全部相同.
- (8) 参考文献互不包括,即排除引用情况.

根据以上特征并结合学术论文的数学模型,本文给出了基于篇章结构和相似度计算的论文抄袭识别算法.

2.1 数据准备

一个分类合理、数据齐全的文档数据库是抄袭识别的前提.按照上述的数学模型,文章的篇章结构用数据库表可以表示为编号(整型);全文特征值(整型);发表时间(时间型);标题(字符串型);作者(字符串型);单位(字符串型);摘要(Memo型);关键词集合(字符串型);中图分类号(字符串型);段落集合(Memo型);参考文献集合(Memo型)等.全文特征值是对某篇论文的全文进行 Hash 处理^[9]得到的整数值.当数据库插入新记录时,首先扫描数据库中所有论文的全文特征值,保证没有重复的论文存入数据库.

2.2 识别函数

根据论文的八元组结构,识别函数设定如下:

$$f(d_1, d_2) = f_m(d_1, d_2) \cdot \lambda_1 f_t(d_1, d_2) \cdot f_u(d_1, d_2) \cdot \lambda_2 f_a(d_1, d_2) \cdot f_k(d_1, d_2) \cdot f_z(d_1, d_2) \cdot \lambda_3 f_p(d_1, d_2) \cdot f_r(d_1, d_2) \quad (5)$$

式中: $f_m()$ 、 $f_t()$ 、 \dots 、 $f_r()$ 分别是八元组中各元素的计算函数.针对文章发表时间 ($f_m()$)、作者 ($f_u()$)、关键词 ($f_k()$)、中图分类号 ($f_z()$) 及参考文献 ($f_r()$) 等一般性特征的计算函数返回值为布尔变量,只要有一项不具有相关性则可认为两篇文章没有抄袭现象,称为否决函数.另外 3 个函数 $f_t()$ 、 $f_a()$ 、 $f_p()$ 分别表示标题、摘要及段落中相关性对抄袭识别的影响,可以用数量值反映两篇论文抄袭程度的不同,称为量化函数,它们对于抄袭识别来说,重要程度不同,所以分别赋予以上函数不同的权值 $\lambda_1, \lambda_2, \lambda_3$.

2.2.1 否决函数 在判定抄袭现象时,否决函数如同掌握一票否决权,其判定方法简单、有效.

计算效率比量化函数好,所以在抄袭识别过程中良好地运用否决函数有助于提高识别效率.

$f_m(d_1, d_2)$ 为发表时间判断函数. 若待测文章 d_1 发表时间晚于 d_2 , 则返回 true, 说明 d_1 可能抄袭 d_2 ; 否则返回 false.

$f_u(d_1, d_2)$ 为作者、单位判断函数. 若待测文章 d_1 的作者没有出现在文章 d_2 中, 则返回 true; 否则返回 false.

$f_k(d_1, d_2)$ 为关键词判断函数. 若 d_1 与 d_2 的关键词集合存在交集, 则返回 true; 否则返回 false.

$f_z(d_1, d_2)$ 为中图分类号判断函数. 中图分类号由两部分组成, 前半部分为字母, 表示学科分类, 后半部分是数字, 包括大类与小类. 若字母部分相同且数字部分之差小于 10, 则返回 true; 若字母部分不同, 或数字部分之差大于 10, 则返回 false.

$f_r(d_1, d_2)$ 为参考文献判断函数. 若待测文章 d_1 的参考文献集合中没有 d_2 出现, 则返回 true; 否则返回 false.

2.2.2 量化函数 在本文的研究中, 抄袭识别的量化函数包括标题、摘要及正文三部分, 从量化的目标来看三者基本相同, 都是找出两段文字之间的相似程度. 这里用到的是向量空间模型^[10](VSM), 在该模型中, 文本被看做是由一组独立词条所组成的向量空间, 每个文本表示为一个特征向量, 通过计算向量夹角余弦来度量两文本的相似程度.

设待计算文本的特征向量为

$$V(p) = (c_1, w_1(p); c_2, w_2(p); \dots; c_n, w_n(p)) \quad (6)$$

其中 c 为独立词条, p 为待计算文本, $w(p)$ 为 c 在 p 中的权重. 权重值一般取决于词条在文本中出现的频率, 目前已经有多种词条权重的计算方法, 式 (7) 是其中的一种^[10]:

$$w_i(p) = \left[\frac{(1 + \ln f_{i,p}) \cdot \ln(1 + N/n_i)}{\sum_{k=1}^K [(1 + \ln f_{k,p}) \cdot \ln(1 + N/n_k)]} \right]^{1/2} \quad (7)$$

其中 $f_{i,p}$ 为第 i 个词条在文本 p 中的出现频率, 即词频; N 为论文库中论文的数量; n_i 为论文库中包含第 i 个词条的论文数量; K 为文本 p 中所有独立词条 c 的数量.

两文本的相似度计算采用余弦法^[10](标准化点积法), 表示如下:

$$sim_a(p_1, p_2) = \frac{\sum_{i=1}^K w_i(p_1) \cdot w_i(p_2)}{\sqrt{\sum_{i=1}^K (w_i(p_1))^2 \cdot \sum_{i=1}^K (w_i(p_2))^2}} \quad (8)$$

上述词频计算公式常应用于信息检索领域, 主要对整段文本进行计算, 而不是针对“段落”. 在计算摘要部分的量化函数时可直接应用式 (9) 来计算.

摘要量化函数 $f_a(d_1, d_2)$ 表示如下:

$$f_a(d_1, d_2) = \frac{\sum_{i=1}^K w_i(a_1) \cdot w_i(a_2)}{\sqrt{\sum_{i=1}^K (w_i(a_1))^2 \cdot \sum_{i=1}^K (w_i(a_2))^2}} \quad (9)$$

式中 a_1 、 a_2 分别为文章 d_1 、 d_2 的摘要部分.

由于文章正文包括多个段落, 而本文定义的抄袭现象又包括只抄袭部分段落的情况, 在计算正文相似度时, 必须进行相应的调整. 设待识别文章 d_1 中的某段落为 d_{1i} , 其中 $1 \leq i \leq l$, 文章 d_2 中的某段落为 d_{2j} , 其中 $1 \leq j \leq m$, 则根据式 (7) 和式 (8), 给出修正后的段落相似度计算公式如下:

$$sim_p(d_{1i}, d_{2j}) = \frac{1}{w_{d_{1i}} w_{d_{2j}}} \sum_{c \in d_{1i} \cap d_{2j}} \left[(1 + \ln f_{c,d_{2j}}) \cdot \ln \left(1 + \frac{N}{n_c} \right) \right] \quad (10)$$

式中: $w_{d_{1i}}$ 和 $w_{d_{2j}}$ 分别为段落 d_{1i} 和 d_{2j} 用式 (7) 计算的权重值, c 为独立词条, $f_{c,d_{2j}}$ 为段落 d_{1i} 和 d_{2j} 交集的某独立词条 c 在段落 d_{2j} 中出现的频率, N 为文章 d_2 中所有段落的总和, n_c 为文章 d_2 中含有 c 的段落数.

利用式 (10) 计算出段落 d_{1i} 和文章 d_2 中所有段落的相似度, 从而得到段落 d_{1i} 对论文 d_2 的相似度为

$$sim(d_{1i}, d_2) = \max \{ sim_p(d_{1i}, d_{21}), sim_p(d_{1i}, d_{22}), \dots, sim_p(d_{1i}, d_{2m}) \} \quad (11)$$

其中 m 为论文 d_2 中段落的数量, d_{2j} ($1 \leq j \leq m$) 为论文 d_2 的第 j 个段落.

依次计算文章 d_1 中各个段落对 d_2 的相似度, 通过加权平均给出正文量化函数 $f_p(d_1, d_2)$:

$$f_p(d_1, d_2) = \sum_{i=1}^l (w_i sim(d_{1i}, d_2)) \Big/ \sum_{i=1}^l w_i \quad (12)$$

其中 w_i 为各段落的权重; l 为论文 d_1 的段落数; $sim(d_{1i}, d_2)$ 为段落 d_{1i} 对 d_2 的相似度.

由于标题内容较少, 词频统计过程中极少出

现词频大于 1 的情况,用向量空间模型计算标题量化函数效果不理想.本文采用数字指纹方法计算标题的量化函数.首先将待计算的两文章标题 t_1, t_2 进行分词处理,并消去虚词及停用词等不能表达实际意义的词,再将余下的词重新生成字符串,分别对两字符串用 Hash 函数生成数字指纹,比较两数字指纹,若相同则返回 1,否则为 0.标题量化函数 $ft(d_1, d_2)$ 的数字表示如下:

$$ft(d_1, d_2) = \begin{cases} 1, & \text{若 } H_{t_1} = H_{t_2} \\ 0, & \text{若 } H_{t_1} \neq H_{t_2} \end{cases} \quad (13)$$

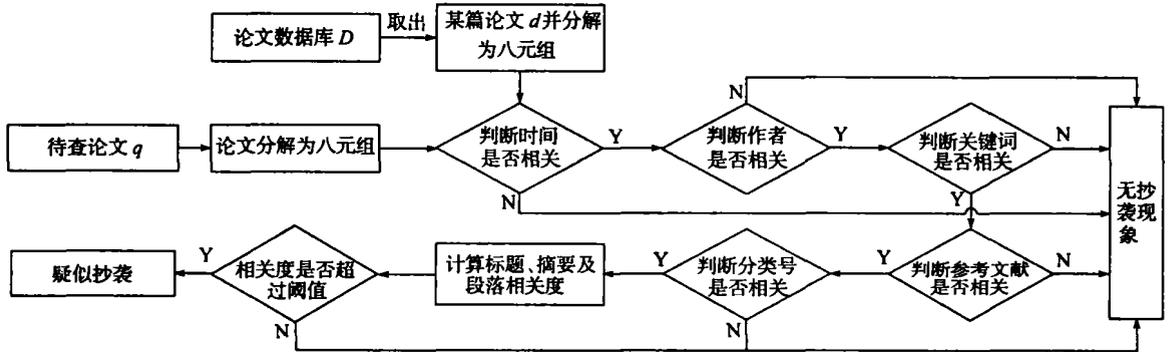


图 2 q 和 d 抄袭识别的流程

Fig. 2 Flow chart of copy detection between q and d

本文研究的算法从本质上说是度量两篇文章的相关程度,而抄袭应该做相似度的判断,但目前由于自然语言理解上的困难,尚无法给出让人满意的相似度算法,本文用相关度来代替相似度.相关度指的是文章之间的相关联程度,而相似度是指文章之间的相似程度.如果论文 q 抄袭了论文 d ,即 q 与 d 相似,那么 q 和 d 肯定是相关的,但如果论文 q 和 d 是相关的,至于是否抄袭,尚需进一步判断.所以本文算法最后判定结果是疑似抄袭,是否是抄袭还须人来做进一步的判断.

本文设定疑似抄袭的阈值为 20%,一般相似度超过这个值其抄袭的可能性就较大,定为疑似抄袭.

3 算例

3.1 抄袭识别算法评价

对于论文抄袭识别算法的评价,可以借鉴信息检索系统中的召回率和查准率.在给定识别论文数量的前提下,召回率是用识别算法识别出的论文抄袭数量与实际论文抄袭数量的百分比;查准率是返回结果中正确结果的百分比.召回率和查准率反映了抄袭识别算法的效果.此外耗时也是一个重要指标,如果耗时太多,就没有实用价

其中 H_{t_1} 为文章 d_1 的标题 t_1 的 Hash 值, H_{t_2} 为文章 d_2 的标题 t_2 的 Hash 值.

2.3 识别流程

设待查论文为 q ,则其进行抄袭识别的主要步骤如下:① 从论文数据库 D 中取出一篇论文 d ;② 根据否决函数判断 q 是否可能抄袭 d ;③ 根据量化函数判断 q 与 d 是否存在抄袭现象;④ 从数据库中取出下一篇文章继续识别,直到遍历 D 中所有论文.抄袭识别流程如图 2 所示.

值.

3.2 测试数据与论文库

为测试算法,本文建立了论文数据库 D ,存入 220 篇科技论文,论文来源于各中文科技期刊的网络版和图书馆,仅供研究使用.论文在数据库中存储格式为预处理后的纯文本格式,并进行了分词处理.同时选定一篇论文 q 作为待识别论文,它有 21 个自然段落.

3.3 计算结果

论文数据库 D 为论文 d 的集合: $D = \{d_i | i \in I, i = 1, 2, \dots, N, N = 220\}$.对于待识别论文 q ,作为考题,事先经人工判断已知 q 为对 d_1 的全盘抄袭; q 对 d_2 为调整顺序抄袭; q 对 d_3, d_4 为部分抄袭,即只抄袭了部分段落;此外 q 又是由 d_5 和 d_6 拼凑而成的; D 中其余的文章与 q 不相关. d_i 的下标 i 为论文在论文库 D 中的序号.

在抄袭识别中,主要的计算集中于量化函数的计算.下面首先给出各量化函数的计算结果并加以分析,最后给出全文抄袭识别的计算结果并与其他算法进行对比.

3.3.1 标题量化函数 标题量化函数计算结果见表 1.

表 1 标题量化函数计算结果 (数字指纹方法)

Tab. 1 Results of title quantifying function (FP, fingerprinting)

论文 ID	相似度	论文 ID	相似度
d_1	1	d_6	0
d_2	0	d_7	0
d_3	1	d_8	0
d_4	1	d_9	0
d_5	0	d_{10}	0

从表 1 中可以看出, q 与 d_1 、 d_3 和 d_4 的标题量化函数计算结果为 1. 除 d_1 的标题与 q 相同外, d_3 和 d_4 的标题与 q 均有差别, 但其所用中心词及语序都相同, 说明本文所用的数字指纹方法对论文标题的抄袭具有识别能力; 但 d_2 和 d_5 与 q 的标题也具有相似性, 其中 d_2 是重新调整了语序, 而 d_5 是换用了一个实词, 标题量化函数没有检测出来, 这与本文采用的数字指纹方法有关, 说明数字指纹方法还具有一定的局限性. 综合来看, 数字指纹方法简单高效, 且由于标题量化函数仅是识别函数的一个部分, 主要的计算在摘要及正文部分, 其计算效果可以满足抄袭识别计算的需求.

3.3.2 摘要量化函数 摘要量化函数是对两段文字计算相似度值, 这里采用向量空间模型的方法, 计算结果见表 2.

表 2 摘要量化函数计算结果 (向量空间模型方法)

Tab. 2 Results of abstract quantifying function (VSM, vector space mode)

论文 ID	$S(B: 0.04899)$	$S\%$
d_1	0.02012	41.07
d_2	0.01465	29.90
d_3	0.01842	37.60
d_4	0.01005	20.52
d_5	0.00908	18.54
d_6	0.00947	19.33
d_7	0.00195	3.98
d_8	0.00127	2.60
d_9	0.00056	1.14
d_{10}	0.00189	3.86

注: 表中 $B: 0.04899$ 为 q 摘要对自身用式 (9) 的计算值, S 为 q 摘要对各文章摘要计算值与 B 值的百分比

从表 2 中可以看出, 虽然本文定义的阈值较低 (20%), 但仍有部分抄袭论文的摘要部分没有从摘要量化函数的计算结果中体现出来, 这与摘要部分篇幅较短有一定关系, 说明对摘要部分采用向量空间模型计算相似度只能粗略地识别抄袭现象, 更准确的结果需要在正文量化函数中得以

体现.

3.3.3 正文量化函数 正文量化函数计算结果见表 3.

表 3 正文量化函数计算结果 (向量空间模型方法)

Tab. 3 Results of context quantifying function (VSM, vector space mode)

论文 ID	$S(B: 0.05438)$	$S\%$
d_1	0.03606	66.32
d_2	0.03101	57.02
d_3	0.02346	43.15
d_4	0.01619	29.78
d_5	0.02166	39.84
d_6	0.01967	36.17
d_7	0.00557	10.24
d_8	0.00614	11.29
d_9	0.00548	10.08
d_{10}	0.00447	8.22

注: 表中 $B: 0.05438$ 为 q 对自身用式 (12) 的计算值, S 为 q 对各文章计算值与 B 值的百分比

从表 3 可以看出, 当对正文量化函数计算时, 相关文章的相似度值大大超过了本文所设定的阈值, 说明存在疑似抄袭. 结果明显优于摘要量化函数的计算结果. 究其原因, 篇幅的大小是其中之一, 但更重要的是这种基于段落的相似度计算方法能查出调换顺序、部分抄袭及拼凑抄袭等情况. 该方法的词频统计是在段落中进行的, 由于抄袭者一般并不抄袭整篇文章, 如果对整篇文章进行词频统计效果必然不好, 所以基于段落的方法对于提高识别效果起了重要作用.

3.3.4 抄袭识别计算结果及几种计算方法的对比 下面用传统的基于全文的数字指纹和基于词频统计的方法与本文的基于篇章结构的方法进行论文抄袭识别对比试验, 从召回率 R 、查准率 P 、耗时 t_r 和适用范围方面进行比较, 用以确定这 3 种方法在抄袭识别中的效果, 见表 4. 用于计算的微机配置为 CPU P4, 主频 1 500 MHz, 内存 256 MB.

从表 4 可以看出, 基于全文的数字指纹方法召回率 R 很低, 主要用来识别重复论文, 不能用于其他类型抄袭识别; 基于全文的词频统计方法尽管查准率较高, 但是过低的召回率和较高的耗时降低了其应用价值; 本文的基于篇章结构和相似度计算的识别方法召回率较高, 能够用于全文抄袭、部分抄袭和拼凑抄袭的论文识别, 虽然对于疑似抄袭论文的计算时间会较长, 但由于使用了否决函数, 检测过程的平均计算时间大为降低. 如

今后使用并行计算等方法将会进一步提高效率。

表 4 三种识别方法的比较

Tab. 4 Comparison of computability by three detection methods

方法	R	P	t_r/s	适用范围
基于全文的数字指纹方法	< 0.000 01	1.00	1	用于检查论文重复
基于全文的词汇统计方法	0.30	0.98	1 300	粗略识别抄袭现象
本文方法	0.86	0.80	700	较准确识别抄袭现象

4 结 语

国外的论文抄袭识别研究已有多年的历史,而中文论文抄袭识别是一个较新的研究领域。本文探讨了基于篇章结构相似度计算的论文抄袭识别技术,给出了相应的模型和计算公式,较好地实现了抄袭识别计算。

需要指出的是,本文只能认定某篇论文为疑似抄袭,即起一个筛选的作用,最终确定是否为抄袭还需要人慎重辨别。

致谢:本文中的中文分词程序采用了中国科学院计算所软件研究室的汉语词法分析系统 ICTCLAS^[11]。

参考文献:

- [1] 宋擒豹,沈钧毅. 数字商品非法复制和扩散的检测机制 [J]. 计算机研究与发展, 2001, 38(1): 121-125
- [2] 鲍军鹏,沈钧毅,刘晓东,等. 自然语言文档复制检测

- 研究综述 [J]. 软件学报, 2003, 14(10): 1753-1760
- [3] AUSTIN R. Word check system [EB/OL]. [2002-12-02] [http // www. wordchecksyste.ms.com](http://www.wordchecksyste.ms.com)
 - [4] ANTONIO S, LEONG H V, RYNSON W H. CHECK: a document plagiarism detection system [C]// **Proceedings of ACM Symposium for Applied Computing**. San Jose [s n], 1997 70-77.
 - [5] HEINTZE N. Scalable document fingerprinting (extended abstract) [C]// **Proceedings of USENIX Workshop on Electronic Commerce**. Oakland [s n], 1996 69-74
 - [6] 史彦军,滕弘飞,金博. 抄袭论文识别研究与发展 [J]. 大连理工大学学报, 2005, 45(1): 50-57 (SHI Yan-jun, TENG Hong-fei, JIN Bo. Study and progress of plagiarism-identification of scientific papers [J]. **J Dalian Univ Technol**, 2005, 45(1): 50-57)
 - [7] 王继成,武港山,周源远,等. 一种篇章结构指导的中文 Web 文档自动摘要方法 [J]. 计算机研究与发展, 2003, 40(3): 398-405.
 - [8] 张益民,陆汝占,沈李斌. 一种混合型的汉语篇章结构自动分析方法 [J]. 软件学报, 2000, 11(11): 1527-1533.
 - [9] UDI M. Finding similar files in a large file system [C]// **1994 Winter USENIX Technical Conference**. San Francisco [s n], 1994 1-10
 - [10] SALTON G, SALTON C. Term-weighting approaches in automatic text retrieval [J]. **Inf Process and Manage**, 1988, 24 513-523
 - [11] ZHANG Hua-ping. HHMM-based Chinese lexical analyzer ICTCLAS [C] // **Second SIGHAN Workshop Affiliated with 41st ACL**. Sapporo [s n], 2003 63-70

Document-structure-based copy detection algorithm

JIN Bo¹, SHI Yan-jun², TENG Hong-fei^{* 2}

(1. Dept. of Comput. Sci. and Eng., Dalian Univ. of Technol., Dalian 116024, China;
2. School of Mech. Eng., Dalian Univ. of Technol., Dalian 116024, China)

Abstract Research on copy detection of academic papers is important in both intellectual property protection and academic plagiarism prevention. Nowadays, researchers mainly use digital fingerprinting technique and keyphrase matching technique on copy detection. To overcome the difficulty of Chinese copy detection, a set of document-structure-based algorithm for identifying Chinese plagiarized papers is presented, and mathematical models on it are given. The plagiarism identification (including full-plagiarism, part-plagiarism and pieced-plagiarism) is realized with the help of document-structure analysis, fingerprinting and word-frequency techniques. Lastly, comparing with two typical identification methods, the effectiveness of accurate paper copy detection of the proposed algorithm is demonstrated.

Key words academic paper; copy detection; plagiarism identification; fingerprinting; word-frequency statistics; document-structure