doi 10.3969/j.issn.1671-1122.2013.11.008

基于感知哈希的在线发表论文版权保护系统

李娇，王健，曹继文，陈彤

（北京交通大学计算机与信息技术学院，北京 100044）

摘 要：针对语义相似词语替换形成的论文抄袭，文章提出了基于主题词语义感知的论文分块感知哈希算法。算法首先使用交叉熵准确提取主题词感知论文内容主题，根据论文结构生成分块感知哈希定位修改位置；然后基于主题词语义相似度计算比较论文相似度，根据注册时间戳、作者 ID 判断版权归属，保护论文原创作者知识版权；最后采用 ECDH 和 Rijndael 算法加密感知哈希防止明文泄露，完成作者、出版发行商双向身份认证。测试表明算法具有鲁棒性、可区分性。

关键词：感知哈希；语义相似；版权保护；主题词提取；内容安全

中图分类号：TP309 文献标识码：A 文章编号：1671-1122（2013）11-0031-04

Thesis Copyright Protection based on Topic Words Semantic Perception Hash

LI Jiao,WANG Jian,CAO Ji-wen, CHEN Tong

(*School of Computer and Information Technology, Beijing Jiaotong University, Beijing* 100044*, China*)

Abstract: In allusion to paper plagiarism replaced with semantic similar words, a thesis segment perceptual hash algorithm based on the topic words semantic perception is proposed in the paper. Firstly, the algorithm extracted topic words from the thesis according to Kullback-Leibler Divergence, sensed the theme of content, built segments perceptual hash on the structure to locate modify. Then compared thesis similarity computed topic words semantic similarity, estimated ownership of copyright used author ID and registered timestamp, protected the author original copyright. Finally it encrypted hash with ECDH and Rijndael algorithm to prevent plaintext leaked and complete mutual authentication between author and publisher. Experimental results show the approach has robustness and discrimination.

Key words: perceptual hash; semantic similarity; copyright protection; extraction of topic words; content security

0 引言

如何有效保护论文原创作者知识版权是网络时代科技论文快速共享发展急需解决的关键问题之一，教育部科技发展中心已将其列入 2013 年研究项目指南 [1]。

以“ 中国知网学术不端文献检测系统 ” 为代表的论文检测系统有效打击了论文抄袭行为，保护了论文原创作者知识版权，在防治学术不端行为中发挥了重要作用。其原理是查找文字最长公共子序列进行自适应多阶指纹对比 [2]，根据文字复制比度量论文相似度，支持篇章、段落多粒度检测定位和改写、组合等论文变形检测，准确率高。

目前论文查重的主要指标是文字复制比，仅从文字形式反映了文字重复程度，缺乏对文字构成词语语义的深层次反映，因此对文字复制、段落重排等形成的抄袭论文效果明显，但对相似词语替换形成的抄袭论文效果就稍逊一筹。提高论文语义层面查重效果是论文检测系统的发展方向。

感知哈希 (Perceptual Hash) 鲁棒性地感知多媒体信息内容特征生成摘要，基于摘要匹配实现多媒体信息内容识别、检索、认证，是多媒体内容相似比较的新方法，已成功应用于谷歌图像搜索、重复网页检测等大数据搜索业务，具有速度快、准确性高的优点。

收稿日期：2013-10-18

基金项目：教育部人文社会科学研究规划基金 [13YJA870013]、2013 年北京交通大学大学生创新训练计划项目

作者简介：李娇 (1992-)，女，浙江，本科，主要研究方向：信息安全；王健 (1976-)，男，浙江，讲师，博士，主要研究方向：网络信息安全；曹继文 (1992-)，女，湖南，本科，主要研究方向：信息安全；陈彤 (1992-)，女，天津，本科，主要研究方向：信息安全。

目前感知哈希研究集中在图像、音视频等多媒体方向，文本感知哈希研究成果仅有胡东辉等将文本感知哈希应用于文本来源可信性检测[3]。

本文创新性地将感知哈希应用于发现语义相似词语替换形成的抄袭论文，提出了一种基于论文主题词语义的分块感知哈希算法，探索从语义层次检测论文相似度，进一步保护论文原创作者知识版权。

1 论文感知哈希算法

1.1 基于交叉熵的论文主题词提取

交叉熵计算公式为[4]：

$$f(w) = \sum_{left} -p_{left} \ln p_{left} - \sum_{right} -p_{right} \ln p_{right} \quad \cdots\cdots (1)$$

词语 w 每出现一次，分上文（left）和下文（right）分别计算熵 p_{left} 和 p_{right}。所有上下文熵之和定义该词权重。

两篇论文比较时，常见词（停用词）不代表论文内容主题，不应考虑；出现一次的词，语义不确定，不需考虑。名称性词组语义确定、鉴别能力强，权重大的词语鉴别能力强，因此选取每篇论文中权重 Top N 的若干名词性词组就能完整地反映该论文内容主题。

本文使用 NLPIR 汉语分词系统[5]进行汉语分词、词性标注、新词发现，然后提取名词性词语按照公式（1）计算权重，选取权重 TopN 的词语作为主题词生成论文指纹，实现从语义上准确感知论文主题内容。

1.2 论文感知哈希生成算法

算法原理是将一篇论文按照结构特征切分成若干片段，对片段用 1.1 节中方法计算对应指纹，通过比较指纹找出大致相同的文字序列，定位修改位置。

如图 1 所示，本文算法将论文分成摘要和正文两块文本分别生成感知哈希。进一步减小粒度，可提高定位精度。

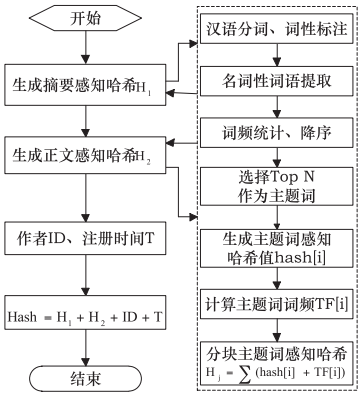

图1 论文感知哈希生成算法

每块论文文本提取论文主题词，取词频 Top N 主题词构成 < 主题词，词频 > 对。主题词从语义上感知论文内容主题，词频从主题词对内容主题的贡献度感知论文的内容主题。两者组合生成的 < 主题词，词频 > 能够充分有效感知论文内容主题，体现论文主题内容。< 主题词，词频 > 数值化生成对应主题词感知哈希 H，每个主题词感知哈希值由 32 位词语哈希值和 8 位词频构成。将作者 ID、论文发表时间 T 和 H 组合构成完整论文感知哈希。

测试论文感知哈希如表 1 所示。表中斜体部分是 Top 5 摘要主题词哈希值及词频构成的摘要感知哈希，非斜体部分是 Top 10 主题词哈希值及词频构成的正文感知哈希。

表1 感知哈希结构

0b5a8f0904 fc373e3d02 ead1e6c902 94e67af802 6c9ed85b02
c44fcd0c32 0b5a8f0928 6c9ed85b21 0324f0131c efadb8c1a
f195d02918 0a4fc46c16 ead1e6c916 fce5eae415 e94d87015

1.3 论文感知哈希模型

如图 2 所示，模型由感知哈希注册、感知哈希库、感知哈希识别三大模块组成。

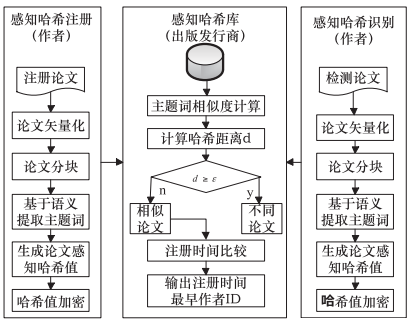

图2 论文感知哈希模型

在感知哈希注册阶段，论文作者提交论文和作者 ID、论文发表时间 T 等相关信息生成感知哈希值，加密后存储于感知哈希库。

为防止感知哈希库内容明文泄露，本文采用 ECDH 算法协商 Rijndael 对称密钥，使用 Rijndael 算法对感知哈希进行加密[6]，实现作者和出版发行商双向身份认证，建立在线发表论文版权全程追溯机制；为支持移动终端进行在线发表论文版权查询和论文版权盲检，可将加密感知哈希值转换为二维码嵌入论文[7]。

在感知哈希识别阶段，检测者提交待检论文，模型生成感知哈希值或提取待检论文中嵌入的感知哈希值，然后计算与感知哈希数据库中存储感知哈希的哈希距离，判断论文相似度，确定匹配论文、显示相似主题词，根据注册时间先后判断原创的和转载（抄袭）的论文，实现论文内容认证、论文抄袭鉴别、拷贝检测，确定在线发表论文版权归属，从而有效解决版权纠纷问题，实现论文版权保护。

综上所述，模型使用感知哈希生成算法，通过提取论文主题词从语义上感知论文主题，计算相应词频感知主题词对论文主题的贡献度，生成分块感知哈希定位相似位置，计算感知哈希距离比较论文相似度，根据注册时间戳、作者 ID 判

断版权归属，实现在线发表论文版权保护。

1.4 基于词汇语义的论文相似度计算

假定分块论文 Pi 包含 k 个主题词 KW_{ik}，$TF(KW_{ik}[K])$ 为 KW_{ik} 对应词频，KW_{1k} 和 KW_{2k} 语义相似度为wordsim(kw_{1k},kw_{2k})。兼顾主题词语义和词频对相似度的影响，定义论文相似度计算公式为：

$$\mathrm{sim}(p_1,p_2)=\sum_{k=1}^{n}\{|TF(kw_{1k}\text{-}kw_{2k})|+(1-\mathrm{wordsim}(kw_{1k},kw_{2k}))\} \quad \cdots\cdots (2)$$

本文提取的主题词是复合词组，多不属于已有基本词（义原），因此不能直接使用基于 CNKI 的词汇语义相似度计算方法计算出词语相似度[8]。

本文将主题词进一步分词为基本词，构成语义计算矩阵，定义公式（3）计算主题词语义相似度：

$$\mathrm{wordsim}(kw_1,kw_2)=\sum_{i=1}^{m}\sum_{j=1}^{n}\mathrm{sim}(s_i,s_j)/(mn) \quad \cdots\cdots (3)$$

公式（3）中 KW_1、KW_2 为进行相似度比较的两个关键词，KW_1 分词为 m 个已知义原 S_i，KW_2 分词为 n 个已知义原 S_j。

“版权图像”和“版权水印”相似计算矩阵，两主题词相似度为 0.581315。

2 测试

为了测试本文算法性能，进行了鲁棒性测试、可区分性测试和基于语义相似度的主题词替换，并与文献 [5] 提供的指纹哈希生成算法进行比较。

2.1 鲁棒性测试

鲁棒性测试从格式攻击、文字重排（乱序）和正常论文编辑（小于 40% 的增、删、改操作）两方面进行，具体测试结果如表 3 所示。表中 *Hnew* 是使用本文感知哈希计算的论文相似度，*HA* 是摘要部分相似度，*HC* 是正文部分相似度，*Hold* 是使用对比哈希计算的论文相似度。

分析表 2 可得以下结论：

1）乱序测试表明 *Hnew* 根据论文语义生成，与论文文字次序无关，对论文内容乱序具有鲁棒性，实现了根据论文语义进行相似度比较，可有效检测出论文格式攻击、文字重排形成的相似论文。

这是因为论文内容乱序后，结构、形式发生虽然发生了变化但内容保持不变，故提取的主题词和词频与原文完全相同，感知哈希值不变，相应相似度仍保持为 100%。

Hold 根据论文词语生成，与论文文字次序相关，具有明显的脆弱性。

2）正常论文编辑测试表明 *Hnew* 同时具有鲁棒性、可区分性。无论对摘要内容还是正文内容进行正常论文编辑，*Hnew* 都敏锐地感知到了内容变动，检测为不同论文，具有明确的区分性，减少了误判率；同时 *Hnew* 值保持在较高水平，具有鲁棒性，保证了论文正常使用。

表2 鲁棒性测试结果

操作	*Hnew*		*HA*	*HC*	*Hold*
乱序	正文	100	100	100	50
	摘要	100	100	100	50
删除	正文	98.88	97.76	100	50
	摘要	95.29	100	90.58	43.75
增加	正文	99.93	99.85	100	50
	摘要	94.78	100	89.56	59.38
修改	正文	90.80	81.60	100	50
	摘要	94.77	100	89.54	43.75

Hold 具有明显的脆弱性，不能准确感知出论文内容变化，缺乏区分性。

3）*HA* 仅在摘要内容变化时变化，*HC* 仅在正文内容变化时变化，说明 *Hnew* 具有分块感知定位功能，能够区分数据修改位于摘要部分还是正文部分。进一步减小检测粒度，可以实现更准确的定位。

Hold 无感知定位功能。

2.2 可区分性测试

可区分性指论文内容发生改变时哈希值相应明显变化，具有明显差异性；完全不同论文，哈希值不相似概率应等于或接近1。

对论文正文内容进行增、删、改操作，计算相应相似度结果如图 3 至图 5 所示。

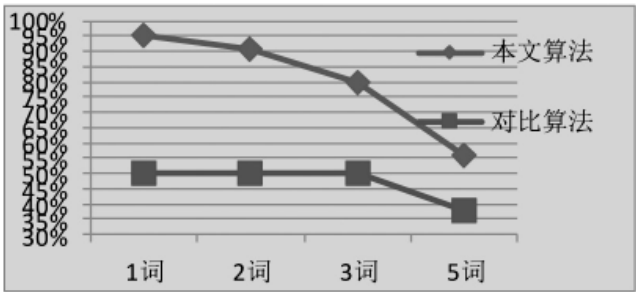

图3 正文内容增加相似度（共10词）

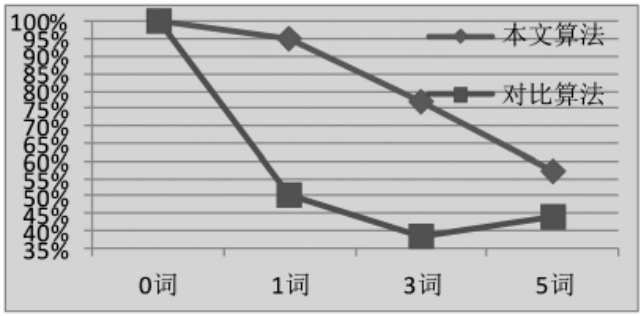

图4 正文内容删除相似度

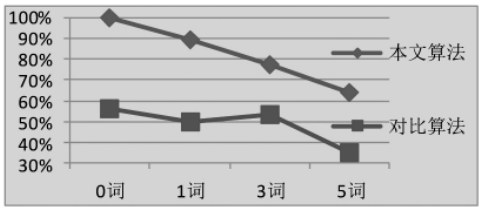

图5 正文内容修改相似度

测试表明随着论文内容修改程度不断增加，Hnew 线性递减，可以明确判别出其与原文的差异程度，取值变化范围较大，具有较强的可区分性。

Hold 可区分性不明显。

2.3 基于语义的相似主题词替换测试

将正文部分的主题词“ 算法 ”依次替换为“ 腾讯 ”、“ 算盘 ”、“ 算术 ”，对应主题词语义相似度依次为 0、0.149、0.896。测试结果如图 6 所示。

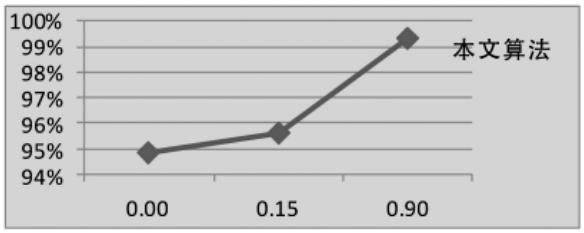

图6 语义相似主题词替换

测试表明本文感知哈希基于主题词语义生成，随着主题词相似度增大，对应论文相似度同步增大，相似度之间存在明显区分度，可以有效防范语义相似词替换形成的论文抄袭。

2.4 测试结论

综上所述，Hnew 根据论文语义生成，与论文文字次序无关，能够准确感知论文内容的变化，具有良好的鲁棒性、可区分性 , 实现了根据论文语义进行相似度比较 ；Hnew 具有分块感知功能，能够定位修改位于论文摘要部分还是正文部分 ；第三方检测、感知哈希库加密存储保证了检测结果的公平性和可信度、避免了感知哈希库内容明文泄露。

3 结束语

本文创新性地将感知哈希应用于论文版权保护，设计了基于论文主题词语义的分块感知哈希算法，提出了兼顾论文语义和词频统计的论文相似度计算新方法，实现了相似内容定位，使用作者 ID、注册时间戳判断版权归属，进行了基于语义保护论文知识版权的有益探索，

本文定义了复合词语义相似度计算公式，在计算论文相似度时通过计算词语相似度提高了检测准确性，能够有效发现语义相似词语替换形成的抄袭论文，减小误判率。

论文相似度比较使用感知哈希实现，运算速度快，能够快速在感知哈希库中搜索出相似论文，适合进行大数据量论文比较，算法具有实用性。

测试表明算法具有很强的鲁棒性、可区分性，在论文语义层次上能够有效区分论文相似度。下一步工作方向是使用 LDA 模型进一步优化论文主题词提取[9]。（责编 吴晶）

参考文献
[1] 教育部科技发展中心 . 网络时代的科技论文快速共享研究项目指南（2013）[EB/OL]. http://www.cutech.edu.cn/cn/zxgz/2013/03/1354173588707562.htm , 2013 - 10 - 08.
[2] 胡东辉 , 王丽娜 , 张娟 , 等 . 基于认知 Hash 的文本来源可信性检测 [C]. 中国电子学会 . 第九届全国信息隐藏暨多媒体信息安全学术大会 CIHW2010 论文集 . 成都 :2010, 175 - 177.
[3] 张华平 . 微博博主特征与行为大数据挖掘分析报告 [EB/OL]. http://blog.sciencenet.cn/blog - 72577 - 684681.html, 2013 - 10 - 31.
[4] 张华平 .NLPIR 汉语分词系统 [EB/OL]. http://ictclas.nlpir.org, 2013 - 10 - 31.
[5] 李冠朋 , 田振川 , 朱贵良 . 基于 ECDH 与 Rijndael 的数据库加密系统 [J]. 计算机工程 ,2013,39(4) :174 - 176.
[6] 李启南 , 李娇 , 武让 . 基于双重零水印的数据库版权保护 [J]. 计算机工程 ,2012,38(08) :107 - 110.
[7] 葛斌 , 李芳芳 , 郭丝路 , 等 . 基于知网的词汇语义相似度计算方法研究 [J]. 计算机应用研究 ,2010,27(09):3329 - 3333.
[8] 闫光辉 , 舒昕 , 马志程 , 等 . 基于主题和链接分析的微博社区发现算法 [J]. 计算机应用研究 ,2013,30(07) :1953 - 1957.

资 讯

北京网络行业协会 2013 年“ 中海浩通杯 ”足球友谊赛成功举办

2013 年 10 月 13 日，2013 年“ 中海浩通杯 ” 足球友谊赛在北京信息科技大学体育场拉开帷幕。本届足球赛事由北京网络行业协会主办，中海浩通（北京）艺术品投资管理有限公司赞助，北京信息科技大学计算机学院协办。本届足球赛事的发起是为促进北京网络行业的有序发展，增进各信息安全企业之间的互动与交流，努力在工作之余为各信息安全企业之间搭建和谐、健康、团结、友好的交流平台。

来自北京市公安局网络安全保卫总队、海淀分局网安大队、朝阳分局网安大队、中海浩通（北京）艺术品投资管理有限公司、北京信息科技大学计算机学院 、杭州安恒信息技术有限公司等 12 支球队参加了比赛，最终北京电信通电信工程有限公司的鹏博士队、中科信息安全共性技术国家工程研究中心有限公司的中科信安队、北京信息科技大学计算机学院队分别获得冠、亚、季军。

本次足球友谊赛的成功举办开创了监管单位与行业单位之间沟通与协作的新形式，为监管单位与行业单位之间的合作与交流提供了新的思路。（记者 马珂）