中文文档复制检测方法研究

耿 崇 薛德军

(中国学术期刊(光盘版)电子杂志社 北京 100084)

【摘要】 介绍不同的文档复制检测方法,对不同方法的技术特点进行对比,通过实验系统论证不同方法的优缺点,并在 CNKI 海量资源的基础上实现中文文档复制检测系统。最后针对目前文档复制检测存在的问题进行分析并确定后续工作内容。

【关键词】 文档复制检测 抄袭检测 【分类号】 TP391.1

Study on Chinese Document Copy Detection

Geng Chong Xue Dejun

(*China Academic Journal (CD) Publishing House, Beijing 100084, China*)

【**Abstract**】 This paper discusses different methods of Document Copy Detection (DCD), compares with each other in the features of DCD techniques through experiment system, and builds a DCD system based on the CNKI repositories. At last, the paper gives a number of recommendations for further work in the field of DCD.

【**Keywords**】 Document copy detection Plagiarism detection

1 引 言

随着网络数字资源的日益丰富和网络环境对人们存取信息方式的改变,文档复制变得越来越容易,与此同时也带来了文档抄袭、学术成果剽窃等一系列知识产权保护问题。文档复制检测(Copy Detection)或者称为剽窃检测(Plagiarism Detection),就是判断一篇文档是否抄袭了其他文档的内容,是实施知识产权保护的一种重要手段。文档间的抄袭不仅意味着原样复制,还包括词语句子级别的增加、删除、顺序调整和同义改写。

1.1 研究概况

文档复制检测研究最早起源于 20 世纪 70 年代软件程序的非法复制检测。目前,文档复制检测的实现方法主要有基于数字指纹的方法、基于字符串比较的方法、基于 VSM 的方法、基于语义序列模式的方法和基于写作风格的方法等。典型的原型试验系统有 SCOP[1]、SCAM[2]、CHECK[3]、PPChecker[4]、WCopyfind[5] 等,这些研究分别代表了不同的技术实现方法,也有各自的优缺点。SCOP 采用的是数字指纹方法,CHECK 是基于 VSM 的方法,PPChecker 是基于语义序列模式的方法。WCopyfind 是基于字符串比较实现的一个开源项目。除此之外,谢菲尔德大学自然语言处理组的 METER[6] 项目在检测新闻文本重用方面做了很多研究。

目前,国内关于中文文档复制检测的研究也越来越多,西安交大的鲍军鹏、大连理工大学的史彦军等人发表了一系列相关文献并构建了相关试验系统[7-10],但是关于建立在超大规模数据集上的中文文档复制检测系统还没有相关研究。

1.2 应 用

(1)数字资源排重

通常用户在利用搜索引擎进行结果筛选过程中,除了过滤不相关文档,还要对重复文档进行区分过滤,通过文档复制检测技术可以把重复文档提前过滤掉,减小用户负担,帮助用户更快地获得所需信息。不仅是针对搜索引擎,在图书馆和一些公司的内部资源管理中,也应用文档复制检测技术来提高文档资源的利用程度。例如,HP 公司就应用这种方法来实施内容管理[11],从大量技术文档中快速找到相似文档。

(2)知识产权保护

如果不对非法文档复制进行检测干预,就会打击知识创造者的积极性。目前,Turnitin[12] 产品已经应用于 50 多个国家和地区的众多科研机构,帮助这些机构进行文档抄袭检测并取得了较好效果。另外,还有用于网页复制检测的 Copyscape[13] 系统,这是一个通过 Google API 实现的专门发现网页抄袭的平台。

收稿日期:2007-04-24
收修改稿日期:2007-04-28

(3)出版媒体编辑审稿

Turnitin 除了开发用于科研机构的抄袭检测产品外,还和众多报纸电子资源出版商合作开发了相应产品,Turnitin 与 LexisNexis 合作开发了 Copyguard,它可以提取每篇文章的数字特征,并与已存档的文章作数字特征比对。哈特福德报社论专栏编辑也使用了 Turnitin 的同类软件来检测非专业记者为该报社论专栏的投稿[14]。

除此之外,文档复制检测技术还可应用于相似文章自动发现,引文自动挖掘等领域。

2 文档复制检测方法

2.1 基于数字指纹的方法

基于数字指纹的方法是目前应用最多的方法,该方法的基本思想是通过对文档生成数字指纹,或者提取文档特征并映射为数字指纹,然后通过指纹比较实现文档复制检测。指纹的生成方法有很多种,包括数字序列、哈希函数等。基于数字指纹方法的难点在于文档特征的选择,也就是用什么样的、用多少数字指纹表示文档特征。

数字指纹方法的一个显著特点是速度较快,通过笔者对 Turnitin 系统的亲自实验,少量的增加、更改和删除词与短语可以检测出来,显示结果较好。针对较短文献很快返回比较结果,文献较长则耗时较多。由于数字指纹在文档级别对序列的依赖性较强,对于同一篇文章调整了很多词的前后顺序后检测效果较差。

2.2 基于字符串比较与压缩的方法

字符串比较是最直接的文档复制检测方法,重复字符串的多少和长度用来判断文档复制的程度。不同于数字指纹的方法,字符串比较不用考虑文档特征提取和数字指纹长度等问题,如 YAP3 与 MDR 系统,采用的都是直接字符串比较方法[5]。

为了节省空间,通常可以把文档构造成一个后缀树,然后进行长字符串匹配查询。虽然后缀树和特定的字符串查找算法可以节省文档复制检测的时间,但是构造后缀树的开销昂贵,而且不能发现改编、增加、删除等情况,不适于海量信息的实时处理。2003 年 Xin Chen 等人提出的 SID 系统(Software Integrity Diagnosis System)[15]采用压缩的方法进行软件程序抄袭检测,压缩方法可以看成是基于字符串比较方法的一种。

2.3 基于 VSM 的方法

1995 年,Shivakumar 等人提出了 SCAM 原型,SCAM 借鉴了信息检索技术中的向量空间模型,使用基于词频统计的方法来度量文本相似性,此后一些文本复制检测系统大量使用基于词频统计的方法来实现,如 CHECK、CDSDG[16]等。值得一提的是,1997 年提出的 CHECK 系统原型把文档结构引入了文档复制检测,在针对文档提取关键词后,CHECK 系统把整个文档形式化为一个结构树,如果父结点不匹配,就不再比较子结点,从而节省了系统比较时间。

2.4 基于语义序列模式的方法

2006 年,NamOh Kang 等人提出了 PPChecker(Plagiarism Pattern Checker)系统[4],该系统通过不同的参数计算值可以得到详细的复制类型,不仅能判断文档片断的次序颠倒和增加删除词汇方式的复制情况,还通过加入语义词典,在一定程度上实现了区分字面不同,但实际意义相同的改写复制。

2.5 基于写作风格和语言统计的方法

每个人都有自己独特的写作风格,这是其他人所不能替代的,这种检测方法类似于完形填空,通过文档本身风格的一致性来判断是否存在片段的复制。这其中包括分析作者在写作过程中的语法倾向、词汇喜好、词类频率和词语分布规律等。这种方法不依赖于其他文档集合,但需要构建语言统计语料。关于基于写作风格和语言统计的复制检测方法可参阅相关文献[17,18]。

3 试验系统

为了对比分析不同的文档复制检测实现方法,笔者自行设计实现了 4 种试验系统。实验环境为普通 PC(512 内存,P4 2.8GHZ),WindowsXP SP2 操作系统,程序设计采用 C ++ 语言,数据存取采用 CNKI 全文数据库 Kbase5.0。

3.1 基于数字指纹的方法

系统实现如图 1 所示,实现过程为:

(1)利用关键词抽取出 CNKI 期刊库 1.2 万篇与数字图书馆相关的学术文档;

(2)以句号为分割符分割全部文档,提取所有长度大于 10 个汉字的句子;

(3)去掉句子中的虚词、助词、连词,转换全部句子为 MD5 值;

(4)针对用户提交的文档以同样方法分割句子并转化为 MD5 值;

(5)比较步骤(4)与步骤(3)中的 MD5 值,两篇文档间重复的 MD5 越多,文档存在复制的可能性越大。

3.2 基于 VSM 的方法

系统实现如图 2 所示,其中文档特征关键词提取采用 CNKI 关键词提取算法[19],对特征向量进行归一化后应用点积公式计算相似度。

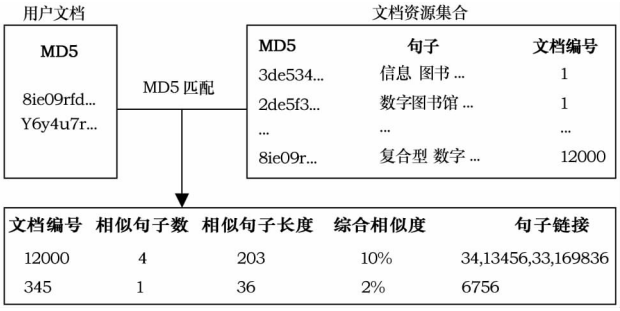

图 1 基于数字指纹方法的复制检测

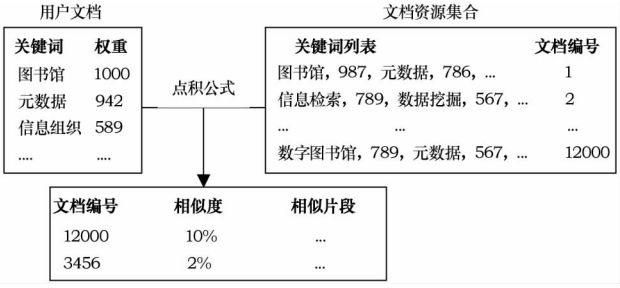

图 2 基于 VSM 方法的文档复制检测

3.3 基于压缩的方法

假设两篇文档合并压缩后的长度变化越大，则表示重复的词汇越多，或者相同词串本身长度越长，那么文档存在复制的可能性也越大。S_o（原文）与 S_c（要检测的文档）的相似程度可以定义为：

$Sim(S_c, S_o) = |(compress(S_c + S_o) - compress(S_c) - compress(S_o))|/compress(S_c)$

$Sim(S_o, S_c) = |(compress(S_c + S_o) - compress(S_c) - compress(S_o))|/compress(S_o)$

```
if ( (Sim(Sc, So) + Sim(So, Sc)) >22% &&( Sim(Sc, So) >1% )
{
  文档 Sc 与 So 相似;分割 2 篇文档为片断:
  Sc = {sega1, sega2, sega3, ⋯, segam}
  So = {segb1, segb2, segb3, ⋯, segbn}
  以相同方法计算 sega(1—m)与 segb(1—n)集合片断之间的相似性
  if( (Sim(segai, segbj) + Sim(segbj, segai) ) >50% )
  {
    两个片断相似,记录两个片断的长度,各自在文档中的起始位置
  }
}
else
{
  文档 Sc 与 So 不相似;
}
```

compress($S_c + S_o$) 表示两篇文档合并后的压缩长度，compress(S_c) 为 S_c 的压缩长度，compress(S_o) 为 S_o 的压缩长度。1%、22%、50% 3 个参数是试验后设定的参考值。通过系统测试，一般情况下如果 $Sim(S_c, S_o) < 11\%$，则基本不存在文档复制情况，考虑两篇文档长度差异造成的压缩值变化，设定选择标准为 $(Sim(S_c, S_o) + Sim(S_o, S_c)) > 22\%$，而不是单独选择 $Sim(S_c, S_o)$ 或者 $Sim(S_o, S_c)$ 值。片断单元是否存在复制情况属于高相似度判断，且片段单元长度较短，实验表明，设定相似阈值为 50% 具有较好的效果。

3.4 基于语义序列模式方法

对文章分词并去掉无意义的虚词和功能词，采用语义系列模式实现相似比较，分词采用最大匹配分词算法，分词词典为 CNKI 关键词词典。系统采用与 PPChecker 相同的实现方法[4]：

$S_o = \{w_1, w_2, w_3, \cdots, w_n\}$，$S_c = \{w_1, w_2, w_3, \cdots, w_m\}$ 分别表示两篇文档的词汇集合，在实验系统中去掉了“的”、“了”这类无意义的虚词。

$Comm(S_o, S_c) = S_o \cap S_c$，表示相同词个数；$Diff(S_o, S_c) = S_o - S_c$，表示不同词的个数；$SynWord(S_o, S_c) = \{w_i | w_i \in Diff(S_o, S_c) \cap Syn(w_i) \in S_o\}$，表示同义词计算，其中，$Syn(w) = \{w$ 的同义词$\}$。

$WordOverlap(S_o, S_c) = |So|/(|Comm(S_o, S_c)| + \alpha * |SynWord(S_o, S_c)|)$，表示加入了同义词的长度重叠参数。$\alpha$ 表示同义词权重。

$SizeOverlap(S_o, S_c) = \sqrt{|Diff(S_o, S_c)| + |Diff(S_c, S_o)|}$，表示不同词长度参数。

$Sim(S_o, S_c) = |S_o|/(e^{WordOverlap(S_o, S_c)-1} + SizeOverlap(S_o, S_c))$

由此可以看出，对于 $Sim(S_o, S_c)$，如果不同的词长度越长，相似度越低；相同词（包括同义词）重叠越多，WordOverlap 越小，相似度越大。

4 试验结果分析

为了对比分析 4 种方法在复制检测过程中的实际效果，笔者随机从 CNKI 中选取 20 篇中文学术论文并去掉其参考文献，从多个文献抽取片段自由组合成新文献 5 篇。实际测试结果分析如下：

4.1 基于数字指纹的方法

系统的速度完全依赖于数字指纹的检索速度。对于完全复制情况的检测速度快速而准确。生成唯一指纹方法的主要缺点是只能检测精确匹配，如果在文档分割阶段对文档片段进行了相应处理，还可以处理一些简单的

词汇增加、删除和更改情况的复制,在系统中是通过保留文档片断的名词、动词、形容词、数量词等有显著意义的词干来实现的。

4.2 基于压缩的方法

该方法能自动发现文章的大部分引用片段,以及部分片段进行了引用但没有注明参考文献的情况,因此,可以作为发现学术剽窃的有效工具。该复制检测方法时间耗费较大,压缩万篇文献(文献长度在 4 000 – 20 000 个汉字左右)的时间在 60 秒以内,这其中包括文档存取时间。系统耗时与要进行复制检测的文档数和文档长度成正比,对于海量数字资源进行分析会造成时间瓶颈。通过压缩比例排序能发现那些有所改动的文档复制情况,但是对于完全改写的文档复制却无能为力。

4.3 基于 VSM 的方法

与压缩的方法相比,有一定比例文章(30%)根据此方法没有找到; 文章级别 VSM 复制检测方法发现的文献都是相关的,但不一定相似,而且大部分都没有复制行为发生;复制与否与 VSM 方法计算后的相似权重不成正比,权重高的不一定是最有复制可能的文章。

4.4 基于语义序列模式的方法

万篇文章比较耗时约 2 分钟,速度变缓至少是压缩方法的 2 倍。如果同义词典记录增大,这个过程还会变得更慢。检测结果发现和压缩方法的前 3 条记录基本吻合。另外一个显著特征是,句子单元级别的语义序列模式方法优于压缩方法,可以检测出加入了同义词替换的片段。如果整个系统从句子单元开始检测会耗时非常大。

语义模式方法最显著的优点是能反馈文档复制的类型,该方法生成的参数值直接代表了不同的复制类型,能给出精确的反馈细节。该方法还依赖于中文分词和词序列的甄选,而基于压缩的检测方法把所有词汇理解为字符编码,可以不用考虑分词问题。

5 问题与展望

5.1 研究基于内容理解的方法

形式上的复制,诸如顺序调整,简单词语的增加删除或更改比较容易被检测出,但是完整内容上的复述改写不容易被检测,虽然 PPchecker 提出了应用 WordNet 语义词典可以检测同义词替换和简单描述更改的复制情形,但其本质上仍然是词语替换,不是文档理解基础上的复制检测。在文档内容理解基础上实现语义内容的复制检测是目前面临的一个难题。

5.2 速度改进

复制检测是基于海量信息进行的,预先过滤不相关文档是改进速度的一个有效方法,是用较小的文档牺牲节约了大量检测时间。文档复制检测本身也可以理解为一个检索过程,只是检索的输入是文档片段,不是单个检索词。笔者应用此种方法改进了 CNKI 跨库文档复制检测系统,通过一次检索后筛选了 1 200 篇左右文档,检测时间平均低于 30 秒,这包括数据分析、检索时间、数据提取和压缩比较时间。此外,采用分布式实现方法也是提高速度的一种途径。

5.3 检测粒度选择

无论是哪一种检测方法,都有一个共同问题,就是文档复制检测粒度的选择。不同的粒度划分是选择复制检测方法和影响检测效果的重要因素。文档复制检测的粒度可以分为:整篇、主题段落、自然段落、句子、定长字符串、词与短语、单个字符或字。目前,主要系统及各种方法的粒度选择是整篇、自然段落、句子、定长字符串,关于主题段落和词与短语级别的粒度比较研究还很少,粒度划分和方法选择的结合是今后要继续研究的内容。

5.4 检测文档中的特殊对象

目前的文档复制检测方法只是针对文档中的文本而言,对于结构化的表格还不能进行针对性的处理,对于文档中的图片、公式也不能进行检测,这也是需要继续深入研究的内容。

参考文献:

1 Brin S, Davis J, Garcia – Molina H. Copy Detection Mechanisms for Digital Documents. In: Proc. SIGMOD 95, New York: ACM Press, 1995: 398 – 409

2 Shivakumar N, Garcia – Molina H. SCAM: A Copy Detection Mechanism for Digital Documents. http://stanford. edu/pub/papers/scam. ps(Accessed Apr. 30,2007)

3 Si A, Leong H V, Lau R W H. CHECK: A Document Plagiarism Detection System. Proceedings of the ACM Symposium on Applied Computing, New York: ACM Press, 1997:70 – 77

4 Kang N, Gelbukh A F, Han S Y. PPChecker: Plagiarism Pattern Checkerin DocumentCopy Detection. Ninth International Conference on TSD. Springer Berlin/Heidelberg,2006: 661 – 667

5 The Plagiarism Resource Site Charlottesville, Virginia. http://plagiarism. phys. virginia. edu/(Accessed Apr. 30,2007)

6 Meter Project. http://www. dcs. shef. ac. uk/nlp/meter/(Accessed Apr. 30,2007)

7 鲍军鹏, 沈钧毅等. 自然语言文档复制检测研究综述. 软件学报, 2003,14(10):1753 – 1761

8 金博, 史彦军等. 中文文档复制检测系统研究 . 计算机工程, 2005,31(19): 79 – 81

9 史彦军, 滕弘飞等. 抄袭论文识别研究与进展. 大连理工大学学报, 2005,45(1):50 – 57

10 鲍军鹏, 沈钧毅等. 一个基于网格的文本复制检测系统. 微电子

学与计算机, 2004,21(9):7 - 10
11 Forman G, Eshghi K, Chiocchetti S. Finding Similar Files in Large Document Repositories . Conference on Knowledge Discovery in Data, New York: ACM Press, 2005: 394 - 400
12 Turnitin. http://www. turnitin. com(Accessed Apr. 30,2007)
13 Copyscape. http://www. copyscape. com(Accessed Apr. 30,2007)
14 稿件检查软件出炉,协助美国新闻界打击内容剽. http://digi. it. sohu. com/20050907/n240352816. shtml (Accessed Apr. 30,2007)
15 Chen X, Francia B, Li M, et al. Shared Information and Program Plagiarism Detection. IEEE Trans. Inform. Theory, 2004,50(7): 1545 - 1551
16 Song Q B, Shen J Y. On Illegal Coping and Distributing Detection Mechanism for Digital Goods. Journal of Computer Research and Development, 2001,38(1):121 - 125
17 Welcome to Glatt Plagiarism Services, Inc. http://www. plagiarism. com(Accessed Apr. 30,2007)
18 Sven Meyer zu Eissen, Benno Stein. Intrinsic Plagiarism Detection, (ECIR - 06), Springer, 2006: 565 - 569
19 张庆国, 薛德军等. 海量数据集上基于特征组合的关键词自动抽取. 情报学报,2006,25(5):587 - 593

(作者 E - mail:gengchong@ gmail. com)

动 态

英国学位论文开放获取的 EThOSnet 项目启动

2007 年 3 月 8 日 EThOSnet 项目宣布启动,两年内该项目将为大英图书馆提供在线学位论文服务,一个完全集成的国家电子学位论文服务正向我们走来。

在参与馆的支持下,英国联合信息系统委员会(Joint Information Systems Committee,JISC)和研究图书馆联盟(Consortium for Research Libraries, CURL)资助该项目拓展获取范围,使研究人员能够获取到原来看不到的、从未使用过的资源。而项目将设立的 EThOS 服务将会开放英国的学位论文为全球使用,同时也是向全世界展示英国最尖端的研究。

2004 年与 2006 年,项目的前期探索为服务开发了一个原型。在此基础上,通过与大英图书馆的合作,EThOSnet 项目将改造英国学位论文的存取方式,提供一个单一入口点来访问学位论文全文。此外,通过与英国机构知识库网络的结合,该项目有望成为国家研究基础设施重要元素。

JISC 项目的主管 Rachel Bruce 说:"该服务将为全国的教育和研究做出重大贡献。JISC 的这一投资将为英国和全世界研究者提供丰富的信息资源"。

CURL 的执行主管 Robin Green 说:"一个全国的电子论文服务将很大程度地为研究者提高资源的可用性,同样加强了在英国进行的研究的质量和范围。CURL 很乐意继续支持这个项目"。

大英图书馆的高等教育主管 Jan Wikinson 说:"在美国有研究显示,电子论文的使用正在急速上升。基于大英图书馆收集和提供印刷版学位论文的经验,EThOS 将使更多的用户可以获取到更丰富的学问论文资源"。

(编译自:Theses unbound: towards a national e - theses service for the UK. http://www. jisc. ac. uk/news/stories/2007/03/news_ethosnet. aspx. [2007 - 3 - 17])

(本刊讯)