一种改进的基于句子相似度的检测算法①

邢长征　孙　伟 (辽宁工程技术大学 电信学院 辽宁 葫芦岛 125105)

摘　要： 基于句子相似度的文档复制检测算法在抓住了文档的全局特征的同时又兼顾文档的结构信息，在该算法的基础上对相似度算法进行了改进，解决了人工设定阈值的问题，并提高了检测精度。实验证明，该算法是可行的，并减少了响应时间。

关键词： 文档复制检测；句子相似度；指纹；词频统计

An Improved Detection Algorithm for Document Reproduction

XING Chang-Zheng, SUN Wei

(Liaoning Technical University, Huludao 125105, China)

Abstract: The document copy detection algorithm based on the similarity of the sentences cannot only emphasize on the whole document, but also on the structure of the document. This paper improves the similarity algorithm based on it, solves the artificial problem of threshold setting and improves the detection accuracy. The result of experiments shows that it is feasible and the running time is reduced.

Keywords: document copy detection; sentence similarity; fingerprints; frequency statistics

文档复制检测(Document Copy Detection)就是判断一个文件的内容是否抄袭、剽窃或者复制于另外一个或者多个文件，剽窃不仅仅是原封不动地照搬，还包括对原文内容的移位交换、同义词替换、说法重述等。文档复制检测技术可以应用在数字图书馆、互联网、网上论文提交系统等来发现重复文本。文本、图像、音频、视频等是数字产品的表现形式，其中文本是互联网信息中最常用的组成部分。从结构上看，文本便于操作，可以进行添加、删除、修改、复制、粘贴等操作，所以文本是被复制得最多的数字产品之一。为了加强知识产权保护，急需建立起一套文档复制检测机制和方法来有效地保护知识产权。目前数字知识产权主要有两种保护措施：一种是“阻止”法；另一种是“检测”法。“阻止”法就是对文本进行加密、嵌入水印或其它载体以防止被复制，“检测”法的思想就是复制检测技术，自从 20 世纪 90 年代此项技术的研究已开始兴起，并且很多相关产品相继问世。像 COPS、SCAM、CHECK、SSK、MDR 等[1]。至今可用于文档复制检测的方法大致分为两种:基于词频统计的方法和基于字符串比较的方法[2]。

1　现有的面向文档的检测算法

1.1 基于字符串比较的方法

这一类方法的主要思想是：从文档中选取一些字符串,这些字符串被称为“指纹”.然后把指纹映射到 Hash 表中,一个指纹对应一个数字.最后统计 Hash 表中相同的指纹数目或者比率,作为依据来衡量文章的相似度。

例如：1993 年,ARIZONA 大学的 Manber 提出的 sif[3]工具, 1995 年美国斯坦福大学数字实验室发明了用于文档复制检测的 COPS 原型系统,贝尔实验室的 Heintze 开发的用于剽窃检测的 KOALA 系统[4], 还有 Broder 等人提出的“shingling”方法[5].另外还有可以发现句子的部分重叠的 K-words 算法等。

1.2 基于词频统计的方法

这一类方法的主要思想是：首先统计文档中各个单

① 收稿时间:2009-05-22

词出现的次数，然后根据单词频度构成文档的特征向量，再采用点积、余弦或类似的方法度量文档相似度。

例如：Garcia-Molina 和 Shivakumar 等人提出的 SCAM(Stanford copy analysis method)原型改进 COPS 系统，以及后来的 DSCAM 模型[6],用于发现知识产权冲突.这种检测方法无关文档的匹配机会随着参数值的变大也越大，即正误差越大。香港理工大学的 Si 和 Leong 等人建立的 CHECK 原型[6]采用统计关键词的方法来度量文本相似性.

基于句子相似度的文档复制检测算法(BSP)结合了基于字符串比较的方法和词频统计的方法的优点[7]对 COPS 算法进行了改进,克服了 COPS 对于局部修改过于敏感的缺点,其算法的核心思想是:以句子为单位,对文档进行指纹提取,在文档相似度计算的过程中,不再仅仅关注文档中精确相等的句子,还包括那些大部分内容重叠的句子。对内容相似的句子进行加权,相似度越高,权值越大,相似度越低,权值越低。同时,设定句子相似度阈值,过滤那些相似度较低的句子。最后,统计相似句子的数量,并把它与两篇文档共有的句子数量的比值作为文档的相似度。BSP 算法与基于句子指纹选取的检测算法(COPS) 、空间向量模型(VSM) 、基于连续 k 个关键词的指纹选取检测算法(K_WORDS)3 个 有代表性的检测算法进行了比较实验，证明了该算法的有效性，尤其在同义替换和综合复制的检测上优于其他三种算法，但是在同等条件下耗时要多于其它三种算法。本文针对相似度计算策略提出了一种新的方案，实验表明，该方案是有效的。

2 本文提出的检测算法

本文提及的是改进后的基于句子相似度的文档复制检测算法,它是将整篇文档以句子标点符号(“，”、“。”、“;”、“?”)为界将文档分解成句子序列，再将这些句子通过中文分词，去掉连词、助词等无意义的词后组成关键词序列称之为“有效句子”。用求最长公共子序列的算法来计算两文档中有效句子间的最大相似度，每篇文档可以用向量空间模型 VSM 表示，向量元素是由相似句子的最大相似度。再利用改进的余弦公式计算两篇文档的相似度。这种算法与基于句子相似度的文档复制检测算法相比不用人工设定阈值，避免了人为因素，提高检测的精确度，并减少了检测响应时间。一篇文档用向量空间模型来表示，就涉及到文本特征的问题：文本块的选择、文本块的表示及相似性度量。

2.1 句子相似度的计算

本文采用求两个句子的最大公共子序列(LCS)的方法来计算句子间的相似度。

设有两个句子 S1,S2,它们的公共子序列集合为 C,其中长度最长的元素设为 c,设 c 的长度为 k,S1,S2 的长度分别为 L1,L2,则 S1,S2 的相似度定义为:

$$sim(S1,S2)=2k/(L1+L2) \quad (1)$$

之所以选择这个公式，是基于以下常识和假设：

① 两个句子的相似度满足：$0\leqslant sim(S1,S2)\leqslant 1$

② 当且仅当 S1=S2 时，sim(S1,S2)=1

③ sim(S1,S2)=sim(S2,S1)

容易证出公式(1)是满足以上三点的。

2.2 文档相似度计算

文档相似度的计算是建立在句子相似度计算的基础上的，方法如下：

$$\mathrm{sim}(VA,VB)=\frac{\sum_{i=1}^{|R|}X_{A,i}\bullet X_{B,i}}{\sqrt{\sum_{i=1}^{|R|}X_{A,i}^2\bullet\sum_{i=1}^{|R|}X_{B,i}^2}}\bullet\frac{\sum X_{1,i}+\sum X_{2,i}}{L(X_1)+L(X_2)} \quad (2)$$

其中 VA、VB 分别表示用 VSM 表示时文档 A、B 中有效句子的最大相似度向量。R 定义如下：

$$R=VA\cup VB=\{a_{R,1},a_{R,2}\ldots a_{R,k}\} \quad (3)$$

i 表示文档 A、B 中相似句子的数量,、表示 A、B 中相似句子经归一化后的向量。归一化公式如下：

$$X_{A,i}=\begin{cases}0 & 若a_{R,i}\notin V_A\\ w_{A,i} & 若a_{R,i}\in V_A\end{cases} \quad (4)$$

其中，表示有效句子中第 i 个句子 aRi 的权重，即该有效句子的最大相似度。此公式是改进的余弦公式，在以前余弦的基础上乘上一因子，是为了解决这样的问题[8]：如果两篇文档最后用向量表示的结果为 VA=(0.5,0.3,0.2,0.1),VB=(0.5,0.3,0.2,0.1)，把向量 VA,VB 同时扩大 2 倍即 VC=(1,0.6,0.4, 0.2), VD=(1,0.6,0.4,0.2)，如果只用余弦法来计算相似度的话，会得出 sim(VA,VB)=sim(VC,VD)=1，即文档 A 与文档 B 完全相同，文档 C 与文档 D 也完全相同，显然这种方法计算的结果不准确。所以在余弦的基础上乘上一因子，分母是两向量的维数之和，分子是各向量的元素之和，因为每一项元素都是在[0,1]之间，所以此因子也是在[0,1]的一个数，当且仅当每个向量

全为 1 的时候，此公式的结果才为 1，全为 0 的时候，此公式的结果才为 0。还是拿上面的向量组为例，计算的结果 sim(VA,VB)= 0.275，sim(VC,VD)= 0.550。

3 实验及相关分析

为了验证本算法的有效性,笔者从知网上下载了 10 篇不同领域的期刊论文,处理后平均大小为 25k。针对文献[6]中指出的 4 种文档复制类型,以及多种复制行为并存的综合复制类型,在原文档的基础上分别对文档集进行相关修改:

(1) 完全复制文档.仅仅修改了源文件的名字,原封不动地复制了源文档的内容;

(2) 对每一篇分别进行部分章节及其标题的选取，形成新的文档进行部分抄袭的检测;

(3) 小部分复制文档.对每一篇进行语句和词语的更改和换位，形成新的文章进行整体变化过后的抄袭检测:文档的小部分内容复制于源文档;

(4) 对每一篇进行 1:n 的文本复制验证，从多篇文章中抽取部分章节交错组合，形成新的文档利用其进行一篇文档抄袭多篇的复制检测;

(5) 完全非复制文档.文档没有复制行为发生。

前四种情况通过修改得到了 40 篇测试用例，第五种情况的实验用例可以在增加 10 篇完全不同的文档，这样共得到 50 篇文档测试用例。实验中利用了自然语言处理开发平台和中科院 ICTCLAS 进行中文分词等文本预处理。并对信息检索领域的查全率和查准率的概念做了针对性的更改以便进行评价实验结果。

实验结果:

表 1　相似度比较统计结果

类型	平均最大相似度
完全复制	1.000
移位	1.000
部分复制	0.734
交叉组合	0.381
无复制	0.037

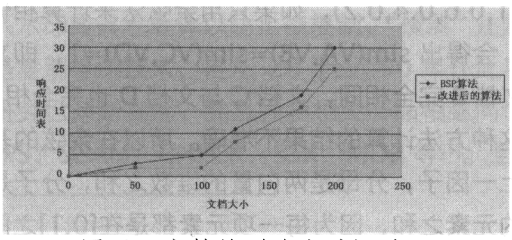

图 1　文档检测响应时间对照

由图 1 可知,耗时改进后的算法在同等条件下要少于 BSP 种算法,BSP 算法是在与 3 个具有代表性的算法基于句子指纹选取的检测算法(COPS)、空间向量模型(VSM)、基于连续 k 个关键词的指纹。

选取检测算法(K_WORDS)进行比较后确实在检测精度上有所提高，但耗时要高于三种算法，但是，我们认为，用更多的计算时间来提高相对精确度在现在的计算机硬件发展水平条件下是值得的。

本文在句子相似性计算策略上做了改进，实验结果证明了该算法的有效性，并且对完全复制和完全非复制的查全率和查准率为 100%。在其他几项测评中，本文的算法也表现良好。

4 结语

本文介绍了一种新的基于句子相似度的文档复制检测算法,本算法选择句子作为文档的特征,引入了句子相似度的思想,较好地抓住了文档的全局信息和文档的结构特征,克服了以往算法中两者不能兼顾的缺陷,提高了检测的精度。实验证明,本算法在文档复制检测中比较有效。

参考文献

1 史彦军,滕弘飞,金博.抄袭论文识别研究与进展.大连理工大学学报, 2005,45(1):50－571.

2 宋擒豹,杨向荣,沈钧义,等.数字商品非法复制的检测算法.计算机学报, 2002,25(11):1206－12111.

3 Kang NO, Gelbukh A, et al. PPCheck: Plagiarism Pat2-tern Checkerin Document CopyDetection. http://www.gelbukh.com/CV/Publications/2006/TSD-2006-Plagia2rism.pdf.

4 Andrei ZB. On the Resemblance and Containment of Documents.Compression and Complexity of SEQU-ENCES 11997.Saler 2no.Italy.1997.21－291.

5 Shiva kumar N, Molina HG. SCAM: A Copy Detection Mechanism f or Digital Documents. The 2nd Interna-tional Conferencein. Theory and Practice of Digital Libraries.Austin.Texas.USA.1995.9－171.

6 Manber U. Finding Similar File sina Large File System.USE 2NIX Conference. San Francisco. CA. 1994. 1－ 101.

7 鲍军鹏,沈钧毅,刘晓东,等.自然语言文档复制检测研究综述.软件学报, 2003,14(10):1753－17601.

8 何明,胡彩霞.一种文本相似性的度量方法和计算方法.黄山学院学报, 2005,7(6):71－721.