第33卷 第24期 计 算 机 工 程 2007年12月
Vol.33 No.24 Computer Engineering December 2007
·安全技术· 文章编号：1000—3428(2007)24—0133—03 文献标识码：A 中图分类号：TP309

基于胎记技术的自然语言文本版权保护方案

杨建龙，王建民，王朝坤，李德毅
(清华大学软件学院，北京 100084)

摘 要：将胎记技术应用于文本，提出了一种全新的自然语言文本版权保护方案。从文本中提取代表特征信息的胎记，并由此识别文本的副本。在与数字水印相似的应用场景下，可用于解释版权归属，为文本提供有效版权保护。实验结果表明该胎记技术具有良好的置信性和鲁棒性。

关键词：胎记技术；数字水印；版权保护；信息隐藏

Copyright Protection Scheme of Natural Language Text Using Birthmark

YANG Jian-long, WANG Jian-min, WANG Chao-kun, LI De-yi
(School of Software, Tsinghua University, Beijing 100084)

【Abstract】 To identify the copyright of natural language text effectively, this paper presents a novel scheme to derive birthmark from a text. Since the birthmark is a unique and native characteristic of every text, a text with the same birthmark of another can be easily suspected as a copy. It proposes an application scenario for birthmark, in which birthmark can be used as an effective method to identify the copyright of text. The experiments show that birthmark successfully distinguishes non-copied files and has quite a good tolerance against meaning-preserving attacks.

【Key words】 birthmark technology; digital watermarking; copyright protection; information hiding

数字水印技术作为版权保护的有效手段，已经成为信息安全领域的热点研究方向。该技术利用数字作品的冗余性和随机性将版权信息以水印形式嵌入其中，当发生版权纠纷时，可以提取出水印以标识和验证版权。文献[1-4]通过对文本内容进行语法和语义变换嵌入水印信息。这些变换均在不改变文本原意的前提下，在自然语言层面操作文本内容，因此被称为自然语言文本水印技术。但是，由于水印技术均需要对文本内容进行不同程度的修改，实际应用中有很大局限性。胎记技术是从数字水印技术中衍生出的一种版权保护技术。它在不修改数字作品的前提下，从数字作品中提取特征信息用于版权标识。该技术最早由Tamada等人提出，他们定义了4种静态的软件胎记[5-6]和1种动态的软件胎记[7]，用于检测Java程序的副本。随后，Myles等人也提出了两种新的软件胎记[8-9]。到目前为止，胎记技术主要应用于软件版权保护领域，且国内鲜有此类研究成果。本文首次将胎记技术应用于文本版权保护，提出一种全新的自然语言文本版权保护 方案。

1 胎记

1.1 概念和定义

定义1(副本关系) 设 *Texts* 是自然语言文本的集合，p 和 q 是自然语言文本。用≌表示 p, q 在 *Texts* 上的等价关系，使其满足以下条件：对于 $p, q \in Texts$，$p \cong q$ 当且仅当 q 是 p 的副本，则称≌为副本关系。

准确地说，这种副本关系不是一成不变的，而应该根据用户的要求而规定。为了方便讨论，本文给出两个评价标准：

(1) q 与 p 完全相同。

(2)q 是 p 经过变换后的语义近似产物，q 与 p 意思基本相同。

若 q 与 p 之间的关系满足以上两个评价标准，则称 $q \cong p$。

定义2(胎记) 设 p 和 q 是自然语言文本，k 是密钥，f 是从文本中提取特征信息的函数。如果 $f(p,k)$ 满足如下两个条件：

(1) $f(p,k)$ 是在 k 的协助下，仅从 p 中即可提取出来的。

(2)$q \cong p \Rightarrow f(p,k)=f(q,k)$。

则称 $f(p,k)$ 是 p 的胎记。

条件(1)的意思是，胎记是文本本身固有的、本质的属性，而不是额外添加的信息。条件(2)的意思是，如果 q 是 p 的副本，则 q 与 p 的胎记相同，也就是说，如果 q 的胎记与 p 不同，则 q 不是 p 的副本。

条件(2)要求胎记必须具备如下两个性质：

(1)置信性：设 p 和 q 是相互独立的两个自然语言文本，k 是密钥。如果 $f(p,k) \neq f(q,k)$，则称 f 是可置信的。

(2)鲁棒性：设 q 是 p 经过变换后的语义近似产物，k 是密钥。如果 $f(p,k)=f(q,k)$，则称 f 是鲁棒的。

置信性要求从相互独立的两个文本中提取出的胎记不同，它关心的是 q 被误判成 p 的副本的可能性。鲁棒性要求

基金项目：国家自然科学基金资助项目(60473077)；2005 年教育部新世纪人才支持计划基金资助项目

作者简介：杨建龙(1981－)，男，硕士研究生，主研方向：数字水印；王建民，教授、博士、博士生导师；王朝坤，讲师、博士；李德毅，教授、博士、博士生导师、中国工程院院士

收稿日期：2007-02-05 **E-mail：**yang-jl04@mails.tsinghua.edu.cn

胎记必须能抵御变换攻击，它关心的是在 p 被变换攻击后还能成功提取出胎记的可能性。

1.2 算法

本文的胎记算法基于矢量空间模型(vector space model)，应用于英文文本。将自然语言文本看成对词的统计对象，并表示成 n 维矢量：

$$\boldsymbol{c}=(c_1,c_2,\cdots,c_n) \tag{1}$$

其中，$\boldsymbol{c}$是文本的矢量；c_i是第i维的单词总数。

对英文文章而言，名词、动词、形容词和副词是表征文章内容的主要词汇。因此定义 4 种胎记：名词胎记(BN)，动词胎记(BV)，形容词胎记(BADJ)和副词胎记(BADV)。对于文本 p 和 q，如果至少一种胎记不同，则 $q\subseteq p$ 不成立；如果 4 种胎记都相同，则 $q\subseteq p$。

以 BN 为例，阐述胎记算法。主要思想是：从文本中提取出所有名词，并将其分成 n 组，每组的词数代表 $\boldsymbol{c}$ 中该维的频率， $\boldsymbol{c}$ 表示该文本的胎记。以下是分组的细节描述：

分组的过程需要用到“代理词汇表”(word-book)、英语 WordNet[1]、单向哈希函数(one-way hash function)和密钥。分组之前，先定义一个“代理词汇表”，表中的每个单词称为代理词，分别代表了一个同义/近义词集合，这是分组的依据。在定义该表的过程中，必须保证所有相关联的同义/近义词将被一个代理词代表，由此保证了所有相关联的同义/近义词在分组时将被分在同一组。分组时：首先，对于每一个从文本中提取出的名词，查询英语WordNet找出它的同义/近义词集合，并查找“代理词汇表”得到其代理词。其次，利用单向哈希函数(例如SHA-1，MD5 等)和密钥便可以计算出该词的组别。以SHA-1 为例，它可以将代理词在密钥的作用下映射成固定长度的大整数，再将该大整数除以分组数n的余数作为该词的组别。下面给出BN的胎记提取算法。

算法 Birthmark Extraction Algorithm－for BN

//k is the private key known only to the owner of the text

initiate n (n>0), the dimension of birthmark vector

initiate $\boldsymbol{c}=(c_1,c_2,\cdots,c_n)=(0,0,\cdots,0)$

extract all words from the text T

for each word w∈T do

  if (w is a noun) then

    use WordNet to find out its synonym set S

    lookup word-book to find out its delegate-word w_d, which is contained in S

    groupIndex $i=Hash(w_d\circ k)\bmod n$

    c_i++

return **c**

“代理词汇表”、英语 WordNet 和单向哈希函数都是预先定义且公开的，因此，密钥 k 就动态决定了分组的结果。单向哈希函数的安全性保证了攻击者没有 k 就不可能计算出分组的结果，也就不能推算出胎记，由此保证了算法的安全性。

1.3 相似度

设$f(p,k)=(p_1,p_2,\cdots,p_n)$和$f(q,k)=(q_1,q_2,\cdots,q_n)$分别是文本$p$和$q$的的$n$维胎记。严格地说，胎记$f(p,k)=f(q,k)$，当且仅当$p_i=q_i$对于所有的$i(1\le i\le n)$成立。但考虑到$q$可能不是$p$的精确副本，则定义相似度以表征$p$和$q$的相似性。本文借用文本挖掘领域广泛应用的余弦公式来衡量其相似度，给出定义 3。

定义 3(相似度) 设$f(p,k)=(p_1,p_2,\cdots,p_n)$和$f(q,k)=(q_1,q_2,\cdots,q_n)$分别是文本$p$和$q$的$n$维胎记。则$f(p,k)$和$f(q,k)$的相似度定义如下：

$$Q=\frac{\langle f(p,k),f(q,k)\rangle}{\sqrt{\langle f(p,k)f(p,k)\rangle\langle f(q,k)f(q,k)\rangle}}=\frac{\sum_{i=1}^{n}p_iq_i}{\sqrt{\sum_{i=1}^{n}p_i^2\sum_{i=1}^{n}q_i^2}} \tag{2}$$

给定检测阈值 $\sigma(0<\sigma<1)$，若 $Q>\sigma$，则 $q\subseteq p$。

2 利用胎记解释版权归属

2.1 解释版权归属

水印 W 和密钥 k 是水印技术用于检测数字作品 q 是 p 的副本，并申明版权所有者对 q 的版权的两个关键要素。W 是 p 的标志性特征，当 p 和 q 中都提取出 W 时，就证明 $p\subseteq q$。k 是版权所有者独有的隐私信息，只要能提供用于从 p 和 q 中提取 W 的 k，就能证明其版权。

与水印技术相似，胎记 B 虽不是版权所有者添加到 p 中的信息，但它是 p 的固有的、本质的属性，也能作为 p 的标志性特征。密钥 k 是版权所有者独有的隐私信息，它与水印技术中的 k 一样，也能作为版权申明的工具。

如果知识产权局或著名门户网站在实际应用中能扮演一个权威仲裁机构的角色，提供胎记的注册，则能为胎记技术的应用提供法律保障。数字作品 p 发布前，版权所有者先将 p 的胎记 B 在仲裁机构中注册，加上时间戳 T。若随后 p 的副本 q 也到该仲裁机构注册，其时间戳 T'一定在 T 之后。时间戳的先后在法律上就证明了 B 和 k 的有效性，从而使 B 和 k 充分地起到版权保护的作用。

2.2 胎记的应用场景

Alice 利用密钥从她的数字作品 p 中提取出胎记 B，并将 B 到仲裁机构注册，由仲裁机构为其加上时间戳 T。随后，Alice 将 p 卖给 Bob，两种盗版情况发生了：

(1)Bob 在没有经过 Alice 授权的情况下，将 p(Bob 可能称其为 q)未作任何修改即卖给其他人。Alice 发现后利用密钥 k 从 q 中提取出水印 B，从而证明 q 是 p 的副本。同时仲裁机构注册过的时间戳和 k 证明了 Alice 对 q 的版权。

(2)Bob 先对 p 作变换后生成 q，企图去除或破坏胎记 B，然后卖给其他人。如果胎记 B 的鲁棒性很高，能够抵御住 Bob 的变换，则 Alice 仍可以利用密钥 k 从 q 中提取出 B，证明 q 是 p 的副本，进而用情况(1)的方法证明其对 q 的版权；如果 B 已经被完全去除或被严重破坏，则 Alice 无法提取出 B，版权验证失败。

3 实验结果与分析

3.1 置信性

本实验验证胎记的置信性。笔者选择文本挖掘领域广泛应用的 20 Newsgroups 数据集(class atheism)来构建测试集。该测试集共有 799 篇两两相互独立的文本，依次选取两篇文本一一配对，测试胎记能否有效地区分这些文本对。测试文本对总数为 318 801，所用胎记分别为 32 和 16 维矢量。密钥为“Identifying the Copyright of Natural Language Text Using Birthmark”。实验结果如表 1。

表 1 胎记的置信性(σ=0.8)

胎记种类	胎记成功区分出的文本对占全部文本对的比例/(%)	
	胎记矢量维度 32	胎记矢量维度 16
BN	97.099 4	65.452 4
BV	99.153 1	86.317 8
BADJ	99.842 5	97.130 5
BADV	99.578 1	99.021 0
BN+BV+BADJ+BADV	99.992 2	99.987 1

结果表明，如果同时使用 4 种胎记，区分度可达到 99.992 2%，该胎记成功地区分了不同的文本，具有良好的置信性。

3.2 鲁棒性

本实验验证胎记的鲁棒性。攻击者企图通过变换攻击文本去除或破坏胎记，胎记技术应该在一定程度上成功地抵御这些攻击。本文考虑的变换均指在不改变文本原意的前提下，在自然语言层面操作文本内容的变换，即语义保持变换，主要包括语法变换和语义变换。语法变换将改变文本的标点符号、句型和句子结构等语法要素，但文本中的词却保持不变，例如Atallah方法[2, 4, 5]。语义变换用同义词和近义词替换文本中具体的词，但文本的语法结构基本保持不变，例如Jensen方法[3]。

笔者从 20 Newsgroups 数据集(class atheism)选取最长的 20 篇文章和最短的 20 篇文章来构建测试集；利用 Atallah 和 Jensen 的方法来模拟对测试集的攻击；用从攻击后的文本中提取的胎记和原文本胎记的相似度来衡量胎记技术的抗攻击能力。

实验结果如表 2 所示，表中列出了针对最长的 20 篇文本和最短的 20 篇文本，从攻击后的文本中提取的胎记和原文本胎记的相似度。分析如下：

(1)大部分的胎记都能抵御语义保持变换，该胎记技术具有相对较高的鲁棒性。

(2)语法变换对胎记的影响很小。因为语法变换仅仅改变文本的语法结构，但文本中的词却保持不变。

(3)语义变换对胎记的影响不大。如果对文本作同义/近义词替换，因为所有相关联的同义/近义词均被分在同一组，影响最大程度地限制在同组内，几乎不会改变胎记矢量。但如果对文本中的词作随机替换，将会破坏文本的语义；且增大某些维度频率的同时也等概率地减小该维度的频率，总体看来，对胎记矢量影响有限。

表 2　胎记的鲁棒性(n=32)

比较项目	相似度/(%)			
	最长 20 篇		最短 20 篇	
	语法变换	语义变换	语法变换	语义变换
平均	99.814 2	98.717 8	97.761 5	96.538 3
最小	99.761 5	98.315 1	97.662 5	95.452 4
最大	99.864 6	99.081 5	98.735 3	97.343 0

4　结束语

本文对自然语言文本胎记作了精确定义，提出了 4 种文本胎记，并给出了其算法与实现。在本文提出的版权保护场景下，它可解释版权归属，是一种全新的自然语言文本版权保护方案。实验验证本文提出的胎记技术具有良好的置信性和鲁棒性。

胎记是一种新的版权保护手段，针对具体应用，其解决方案还有待进一步研究和完善。今后，笔者将继续改进文本胎记技术，提高其抗语义保持变换的能力。同时，开发成熟的文本胎记系统也是一项重要任务。

参考文献

[1] Atallah M J, McDonough C J, Raskin V. Natural Language Processing for Information Assurance and Security: An Overview and Implementations[C]//Proc. of NSPW'00. New York: ACM Press, 2001: 51-65.

[2] Jensen C D. Fingerprinting Text in Logical Markup Languages [C]//Proc of ISC'01. Berlin: Springer-Verlag, 2001: 433-445.

[3] Atallah M J, Raskin V, Crogan M, et al. Natural Language Watermarking: Design, Analysis, and a Proof-of-Concept Implementation[C]//Proc. of IH'01. Berlin: Springer-Verlag, 2001: 185-199.

[4] Atallah M J, Raskin V, Hempelmann C F, et al. Natural Language Watermarking and Tamperproofing[C]//Proc. of IH'02. Berlin: Springer-Verlag, 2003: 196-212.

[5] Tamada H, Nakamura M, Monden A, et al. Design and Evaluation of Birthmarks for Detecting Theft of Java Programs[C]//Proc. of IASTED'04. Spain: ACTA Press, 2004: 569-575.

[6] Tamada H, Nakamura M, Monden A, et al. Java Birthmarks——Detecting the Software Theft[J]. IEICE Transactions on Information and Systems, 2005, E88-D(9): 2148-2158.

[7] Tamada H, Okamoto K, Nakamura M, et al. Dynamic Software Birthmarks to Detect the Theft of Windows Applications[C]//Proc. of ISFST'04. Xi'an, China: [s.n.], 2004.

[8] Myles G, Collberg C. Detecting Software Theft via Whole Program Path Birthmarks[C]//Proc. of ISC'04. Berlin: Springer-Verlag, 2004: 404-415.

[9] Myles G, Collberg C. K-gram Based Software Birthmarks[C]//Proc. of SAC'05. New York: ACM Press, 2005: 314-318.

(上接第 132 页)

Signaling[J]. Proceedings of the IEEE, 1987, 75(1): 56-73.

[7] Dimitar T, Sonja F, Marija E, et al. Ad hoc Networks Connection Availability Modeling[C]//Proceedings of the 1st ACM International Workshop on Performance Evaluation of Wireless Ad Hoc, Sensor, and Ubiquitous Networks. Venezia, Italy: [s. n.], 2004: 56-60.

[8] 胡光明, 蒋　杰, 龚正虎. 移动自组网络分簇算法综述[J]. 计算机工程与科学, 2005, 27(1): 48-50.

[9] 彭　伟, 卢锡城. 一个新的分布式最小连通支配集近似算法[J]. 计算机学报, 2001, 24(3): 254-258.

[10] Chang C, Chang C, Huang P. Dynamic Channel Assignment and Reassignment for Exploiting Channel Reuse Opportunities in Ad Hoc Wireless Networks[C]//Proc. of the 8th International Conference on Communication Systems. [S. l.]: IEEE, press, 2002: 1053-1057.