

基于文本特征的文本水印算法

斯 琴,张 力,廉德亮

(深圳大学 信息工程学院,广东 深圳 518060)

(clover.0404@hotmail.com)

摘 要:基于格式的文本水印算法对格式攻击的鲁棒性比较差,而基于自然语言的文本水印算法相对难以实现,因此提出一种基于词频的文本零水印算法。对文本内容进行分词并计算每个分词的词频,根据设定的词频阈值范围依次提取分词序列作为文本特征,将文本特征、水印和密钥注册于版权保护(IPR)信息库。水印检测可实现盲检测。将该算法用于含有图像等多媒体信息的中英文文档,试验结果证明,该算法对剪切、粘贴、内容顺序颠倒等攻击有较强的鲁棒性。

关键词:文本水印;文本特征;特征提取;词频;分词

中图分类号: TP391 **文献标志码:** A

Text watermarking based on text feature

SI Qin, ZHANG Li, LIAN De-liang

(College of Information Engineering, Shenzhen University, Shenzhen Guangdong 518060, China)

Abstract: The format-based text watermarking algorithm has poor robustness against format attacks, and the natural-language-based text watermarking algorithm is difficult to realize. A text zero-watermarking based on word frequency was proposed. Words were segmented and word frequency was computed. The words were sequentially extracted in threshold range of word frequency to be text feature. Text feature, watermark and secret key were registered to the information database. Watermarking detection was blind. Both Chinese and English documents with multimedia information were tested in the experiments. Experimental results demonstrate that the technique has good robustness against attacks, such as cutting, pasting and reversing.

Key words: text watermarking; text feature; feature extraction; word frequency; word segmentation

0 引言

数字水印技术主要解决网络环境中的版权保护和信息安全的防护,嵌入水印之后的宿主媒体不能在视觉上有所降质或可察觉性,但又要求无意或恶意攻击之后,水印难以被去除,且能检测出水印。文本不像图像、音频、视频有较多的冗余量嵌入数字水印。现有的文本水印算法主要有两种:一种是在文本的格式特征如行移、字移、字符特征等中嵌入水印^[1],由于依赖文本格式特征,因此对格式转换的抗攻击性较差;另一种文本水印算法是利用自然语言技术,通过转换句法结构,在语义理解的基础上构造 TMR (Text Meaning Representation) 树,采用嫁接 (Grafting)、剪枝 (Pruning)、等价信息替换 (Adding/Substitution) 等方法对 TMR 树进行操作来嵌入水印^[2-3]。由于汉字本身的普遍多义性以及中文自然语言技术自身不成熟,中文文本中基于自然语言技术的水印算法实现起来难度较大。

本文采用了不修改任何宿主信息嵌入水印信息的方法——零水印技术^[4],有效解决了鲁棒性和隐蔽性之间的矛盾。本文提出的文本水印算法,通过解析得到文本内容,利用正向最大匹配法 (Forward Maximum Matching Method) 对文本内容进行分词。对得到的每个分词计算词频,根据设定的词频阈值范围,顺序提取文本内容对应的阈值范围内的所有分词,作为该文本的特征。将文本特征和水印存入于版权保护 (Intellectual Property Right, IPR) 信息库中,并产生相应的密

钥以保证水印的安全性。该算法适应于有多媒体信息的常用格式和无格式的中英文文档,实验结果证明,该算法有很好的鲁棒性、隐蔽性和安全性。

1 文本特征提取

目前文本分类中主要使用向量空间模型 (Vector Space Model, VSM) 来表示文本^[5]。这种方法的缺点在于文本向量的维数很高,需要对高维特征向量进行降维处理。特征降维方法主要包括特征选择和特征抽取两种方法:特征抽取是把特征集映射到较低维的空间中,是原始特征的某种组合;特征选择是通过构造一个评估函数,对每个特征进行评估,并将按评估值进行降序排序,从中选取若干个评估值最高的特征项。特征选择方法相对简单、易于理解和计算复杂度低而被广泛采用。常用的文本特征评估函数有交叉熵、信息增益 (Information Gain)、互信息 (Mutual Information)、基于词频方法^[6-7]等。本文提取的文本特征是能表示文本主要内容的特征,而不是对文本进行分类,所以本文采用简单的词频方法提取文本的特征。

英文文本单词用空格分隔,可直接进行统计分析。而中文是以汉字为基础的,词与词之间没有明显的分割符号,必须把词与词分割开来,即分词。目前的中文分词方法大体上分为以下三类:基于理解的分词方法、基于词典的分词方法和基于统计的分词方法^[8]。

本文采用基于词典的分词方法中的比较成熟的方法——

收稿日期:2009-03-24;修回日期:2009-05-15。 基金项目:国家自然科学基金资助项目(60502027)。

作者简介:斯琴(1985-),女(蒙古族),内蒙古乌兰浩特人,硕士研究生,主要研究方向:文本数字水印、文本分类; 张力(1973-),女,山东莱西人,教授,博士,主要研究方向:数字水印、信息隐藏; 廉德亮(1965-),男,江苏徐州人,副教授,博士,主要研究方向:信号处理、信息安全。

正向最大匹配法,它的基本思想是将待处理的字符串从左到右与词典中的词进行比较,如果在词典中找到某个字符串,则表示匹配成功^[9]。

2 算法实现

传统零水印算法的重点在于如何根据作品特征来构造零水印,但零水印不像常规的水印一样具有特定的内容。在本算法中,允许用户嵌入自己的特定文字水印信息,并将特定位置和水印容量等信息作为密钥来保证水印的安全性。将密钥、水印和特征存储于 IPR 信息库中,作为版权归属的凭证。

文本分类中提取文本特征的目的是进行分类,而零水印是反映作品的主要信息。简单地按词频阈值范围内提取分词作为文本的特征可能导致不同但大概相似的文本内容的特征相同,这可能导致版权归属的纠纷。为了有效鉴别不同的文本,采用顺序提取文本内容对应的阈值范围内的所有分词,作为该文本的特征,从而有效地反映和区分了文本内容。

但由于文本受到剪切、复制、粘贴等攻击之后对文本内容的顺序有或多或少的影响,从而可能会影响文本特征的特征。由于一个分词的顺序的改变,可能使待检文本的特征和 IPR 信息库中的特征不能完全匹配,而判断为不是同一个文本。所以本文中引入了编辑距离(Levenshtein Distance)^[10-11]的计算方法计算数据库中的所有特征和待检测文本特征的相似度。相似度高于阈值则认为属于同一个文本,当相似度低于阈值则认为不是不同的文本,从而提高了特征的鲁棒性,与此时也提高了水印的鲁棒性。 $LD(s_1, s_2)$ 表示两个字符串(s_1, s_2)间的编辑距离, $MaxLen(s_1, s_2)$ 表示(s_1, s_2)的最长字符串长度, $Sim(s_1, s_2)$ 表示(s_1, s_2)的相似度,文本中采用的相似度计算式为:

$$Sim(s_1, s_2) = 1 - LD(s_1, s_2) / MaxLen(s_1, s_2) \quad (1)$$

2.1 水印嵌入

设待嵌文本为 $D, D = \{t_1, t_2, \dots, t_N\}$ 。文本中的任意分词记为 t_i, N 为文本内容的分词数目。 tv 表示阈值范围,水印嵌入过程如图1所示,水印嵌入的算法步骤为:

1) 水印信号的产生。嵌入的水印信息采用文本形式,目前还未实现多类水印。水印是直接嵌入到文本特征的相应位置,所以对水印信息不进行任何转换。

2) 文本预处理及分词处理。解析文本得到文本内容。由于只是对文本内容的中英文进行分词处理,所以在解析过程中去除视频、音频、图片等多媒体信息,而对于表格、公式等附带文字信息的多媒体载体,抽取文字信息。因此本文提出的水印算法对那些针对多媒体信息的攻击具有很好的鲁棒性。然后对解析之后的文本内容采用正向最大匹配法进行分词处理,并提供停用词表来去除对文本特征没有价值的分词,得到去除停用词之后的文本所有分词。

3) 特征选择。对以每一个 $t_i \in D$, 计算文档 D 中的词频 tf_i 。根据文本内容的分词分布情况,选择一个词频阈值范围 tv , 查询 $tf_i \in tv$ 的所有分词。根据文本中出现的顺序抽取阈值范围内的所有分词,将得到的有序分词序列作为文本特征。

4) 计算相似度。待嵌文本的特征与 IPR 信息库中的特征进行相似度计算,相似度高于阈值 v , 认为该文档已在 IPR 信息库中存在。相似度低于 v 阈值,视该文档未嵌入过水印,该用户为待嵌文档的版权所有者。

5) 产生密钥。产生两位随机整数,作为用户密钥的前两位。用户选择的词频阈值范围 tv 作为密钥的 3 ~ 6 位,水印的长度作为密钥的后几位。用户提取水印时必须提供有效的密钥,否则不能成功提取水印。

6) 水印的嵌入。将用户的水印、密码、特征存入于 IPR

信息库中,作为版权归属的凭证。

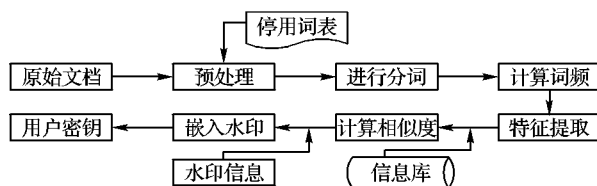


图1 水印嵌入过程

2.2 水印提取

本文所提出的水印算法在发生版权纠纷时可以完全提取有意义的水印信息,而不仅仅是检测到水印信息。水印提取过程如图2所示,水印的提取步骤如下。

1) 文本预处理及分词处理,并产生特征。处理过程如水印嵌入过程的步骤2) ~ 步骤3)。

2) 相似度比较。待检文档的特征和 IPR 信息库中的特征进行比较,假如大于阈值,则认为该文档已嵌入过水印,否则未嵌入过水印,不能进行检测。

3) 水印信息提取。用户所持密钥和 IPR 信息库中的密钥进行匹配,检测用户密码的合法性,合法则从 IPR 信息库中提取用户的水印信息。

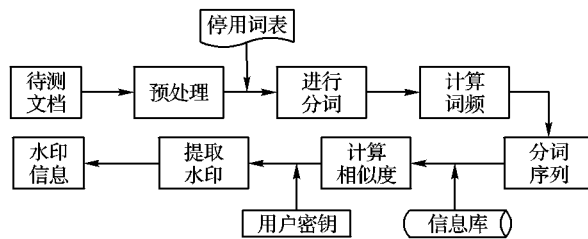


图2 水印提取过程

3 实验结果及分析

为了验证本文提出的算法的性能,选取大小为4851字符数(计空格)的带有图片、公式和表格的中英文 PDF 文档。水印信息采用的是事先产生的文本,水印信息为“深圳大学”。词频阈值范围 tv 设为 $[05, 10]$, 相似度阈值 v 设为0.90。在Java平台上实现了该系统。对该算法进行了粘贴、剪切和文档内容顺序颠倒等攻击之后检测水印,验证该算法的性能。

3.1 原始文档嵌入和检测水印

图3(a)给出了原始的 PDF 文档的部分内容,在未受任何攻击的情况下,原始文档的检测的结果如图3(b)所示。未受攻击时精确地提取到了水印信息“深圳大学”。

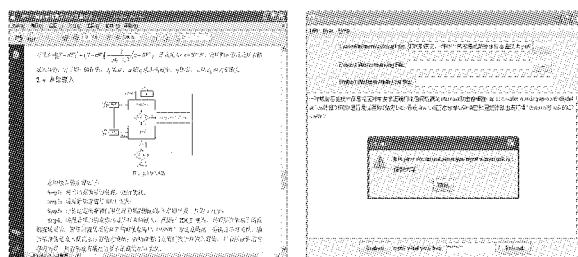


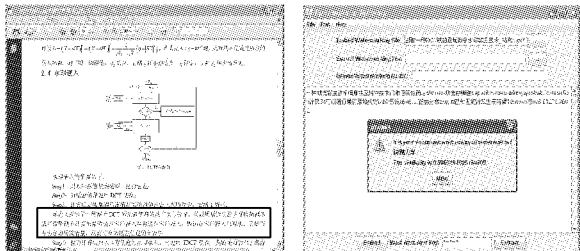
图3 原始文档嵌入和检测水印

3.2 粘贴

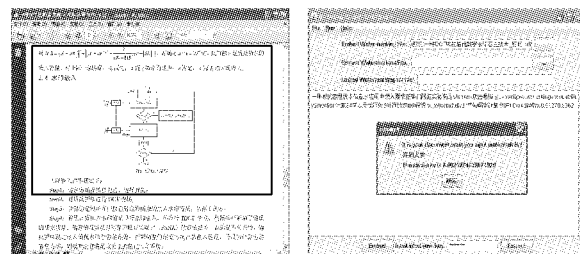
在对原文的“Step3”和“Step4”中间粘贴了三行文字信息,如图4(a)所示。图4(b)为粘贴内容之后提取水印的过程。特征相似度大概为0.9805,大于相似度阈值 v , 判断为同一个文档。粘贴自然段之后,虽然文本的特征有了稍微的改变,但是高于设定的相似度阈值 v , 所以成功地提取了水印信息。实验结果表明,本文算法对粘贴攻击的抵抗比较有效。

3.3 剪切

对原文的“可设”行到“Step2”行之间的所有文字和图片信息进行了剪切,如图5(a)所示。图5(b)为提取水印的结果图。剪切攻击之后提取到了水印信息,且相似度大约为0.902,大于我们预先设置的相似度阈值0.90。实验结果证明本文算法的抗剪切攻击能力强。



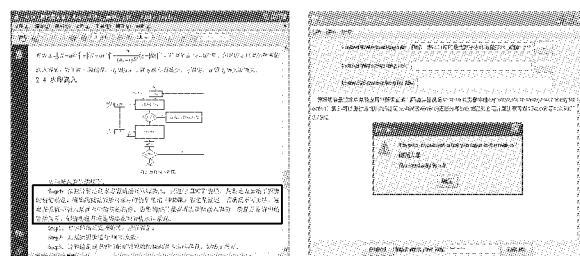
(a) 原始文档上粘贴内容 (b) 原始文档中粘贴内容之后提取水印
图4 原始文档上粘贴内容和提取水印



(a) 原文档上剪切内容 (b) 原文档剪切内容之后提取水印
图5 原文档上剪切内容和提取水印

3.4 颠倒内容

把原文的“Step4”自然段提前到了“Step1”自然段之前,如图6(a)所示。图6(b)为颠倒攻击之后提取水印信息的过程。颠倒攻击之后提取到了水印信息,且相似度为1.0。颠倒的内容中的分词也许不属于阈值范围内的分词,或该分词在所颠倒的内容之前已出现,所以对待检文档的特征没有任何的影响。从实验结果看出,本算法对文档内容顺序的颠倒的攻击的鲁棒性强。



(a) 原文档上颠倒内容 (b) 原文档颠倒内容之后提取水印
图6 原文档上颠倒内容和提取水印

4 结语

针对现有文本水印算法存在的不足,结合文本特征提取技术和零水印技术思想,提出了一种基于文本特征的文本水印算法。采用计算词频来抽取文本特征,并将文本特征、水印和密钥存入于IPR信息库中,作为数据版权的凭证。提取水印时必须提供密钥,有效地保证了用户密钥的安全性。该算法由于实现了对常用和无格式文档内容的解析功能,所以不是针对单一格式或纯文本的文档。该算法是基于文本内容,所以对格式转换攻击的鲁棒性比较强。实验结果证明,基于文本特征的文本水印算法的鲁棒性强,能抵抗常用文本攻击,并有效地解决了不可见性与鲁棒性之间存在的矛盾。

参考文献:

- [1] BRASSIL J T, LOW S, MAXEMCHUK N F. Copyright protection for the electronic distribution of text documents [J]. *Proceedings of the IEEE*, 1999, 87(7): 1181–1196.
- [2] ATALLAH M J, RASKIN V, HEMPELMANN C F, *et al.* Natural language watermarking and tamperproofing [C]// *Proceedings of the 5th International Workshop on Information Hiding*. London: Springer-Verlag, 2002: 196–212.
- [3] ATALLAH M J, RASKIN V, CROGAN M, *et al.* Natural language watermarking: Design, analysis, and a proof-of-concept implementation [C]// *Proceedings of the 4th Information Hiding Workshop*. Pittsburgh, PA: [s. n.], 2001: 185–199.
- [4] 温泉, 孙铤锋, 王树勋. 零水印的概念与应用[J]. *电子学报*, 2003, 31(2): 214–216.
- [5] 苏金树, 张博锋, 徐昕. 基于机器学习的文本分类技术研究进展[J]. *软件学报*, 2006, 17(9): 1848–1859.
- [6] SEBASTIANI F. Machine learning in automated text categorization [J]. *ACM Computing Surveys*, 2002, 34(1): 1–47.
- [7] LEWIS D D. Feature selection and feature extraction for text categorization [C]// *Proceedings of the Workshop on Speech and Natural Language*. New York: [s. n.], 1992: 212–217.
- [8] 马玉春, 宋涛瀚. Web 中中文文本分词技术研究[J]. *计算机应用*, 2004, 24(4): 134–136.
- [9] 吴胜远. 一种汉语分词方法[J]. *计算机研究与发展*, 1996, 33(4): 306–311.
- [10] MONGE A E, ELKAN C P. The field-matching problem: algorithm and applications [C]// *Proceedings of the 2th Internet Conference on Knowledge Discovery and Data Mining*. Menlo Park, CA, AAAI Press, 1996: 267–270.
- [11] NAVARRO G. A guided tour to approximate string matching [J]. *ACM Computing Surveys*, 2001, 33(1): 31–88.

(上接第2347页)

- [6] HARMSSEN J J, PEARLMAN W A. Steganalysis of additive noise modelable information hiding [C]// *Proceedings of SPIE: Watermarking Multimedia Contents*. New York: [s. n.], 2003: 131–142.
- [7] XUAN GUO-RONG, SHI YUN-QING, GAO JIAN-JIONG, *et al.* Steganalysis based on multiple features formed by statistical moments of wavelet characteristic functions [C]// *Proceedings of the 7th International Information Hiding Workshop*, LNCS 3727. Berlin: Springer, 2005: 262–277.
- [8] FRIDRICH J, GOLJAN M, HOGEA D. Steganalysis of JPEG image: Breaking the F5 algorithm[C]// *Proceedings of the 5th International Workshop on Information Hiding*, LNCS 2578. Berlin: Springer, 2002: 310–323.
- [9] CHANG C C, LIN C J. LIBSVM[EB/OL]. (2008–10–30) [2009–01–05]. <http://www.csie.ntu.edu.tw/~cjlin/Hlibsvm/>.
- [10] FRIDRICH J. Feature-based steganalysis for JPEG images and its implications for future design of steganographic schemes [C]// *Proceedings of the 6th Information Hiding Workshop*, LNCS 3200. Berlin: Springer, 2004: 67–81.
- [11] RENCHER A C. Methods of multivariate analysis [M]. New York: John Wiley, 2002.
- [12] United States Department of Agriculture. NRCS photo gallery [EB/OL]. [2009–01–10]. <http://photogallery.ThresT.usda.gov/>.