DOI:10.16208/j.issn1000-7024.2013.04.040

基于文本内容的敏感数据识别方法研究与实现

李伟伟，张 涛，林为民，邓 松，时 坚，汪 晨

（中国电力科学研究院南京分院，江苏 南京 211100）

摘 要：为了防止敏感数据的泄露，为数据的访问控制提供依据，提出并实现了一种基于中文文本内容的敏感数据识别方法。通过对敏感数据库和已知分类文档库的学习，完成对文本中敏感数据识别的阈值的确定和未知文档是否敏感数据的判断过程。描述了预处理、文本识别、阈值确定的详细设计和实现过程。通过对搜狗语料库中教育相关部分文本的识别，验证该方法的敏感数据识别过程简单实用并且具有较高的正确率。

关键词：敏感数据；文本识别；内容识别；数据防泄漏；分类算法

中图法分类号：TP393 **文献标识号**：A **文章编号**：1000-7024（2013）04-1202-05

Research and implementation of sensitive data identification method based on text content

LI Wei-wei, ZHANG Tao, LIN Wei-min, DENG Song, SHI Jian, WANG Chen

(China Electric Power Research Institute (Nanjing), Nanjing 211100, China)

Abstract: To prevent the leakage of sensitive data and provide the basis for data access control, a design method of identifying sensitive data based on chinese text content is presented. Through the study of sensitive text library and the text library which included the same number of sensitive text and security text, it can determine the threshold of the sensitive data and judge the unknown classification text whether sensitive data or not. The design and implementation process of pre-processing, text recognition and determination of the threshold is described. Finally, by identifying the education-related text in Sogou corpus, experiments prove that the method is simple and practical and has a high accuracy rate.

Key words: sensitive data; text recognition; content identification; data leakage prevention; classification algorithm

0 引 言

目前，防止数据泄露的方法大致可以分为三大类：安全审计类，安全控制类和文件加密类[1]。其中，敏感数据的识别技术对于数据防泄漏的安全控制起到很好的辅助作用。如果可以智能化识别从内网发往外网的文档哪些是敏感数据并对其加以保护，那么可以很大限度简化人工设置标识或者是访问控制规则的复杂度并且有效防止敏感数据的泄露问题。

目前对敏感数据的识别主要是对文本[2]、web[3]、图像[4]、视频[5]等文件格式的识别。主要广泛应用的是基于文本的数据防泄漏。目前基于文本的识别算法主要有三种：一种是基于概率和信息理论的分类算法，如朴素贝叶斯算法（naïve bayes，NB）[6]，最大熵算法。另一种是基于TFIDF权值计算方法[7]，这类算法主要包括TFIDF算，k邻近算法等。第三类是基于知识学习的算法，如支持向量机（support vector machine，VSM）[8]算法，人工神经网络算法（artificial neural networks，ANN）等。在以往的文本识别算法研究中文本的预处理过程较为复杂，也缺乏灵活的阈值确定机制。

本文所做的主要工作：一是提出了一种简单的特征选择预处理方法，并详细解释了其有效性。二是提出了一种基于学习的阈值确定方法。三是通过实验方法实现了敏感数据的识别并验证其有效性。

收稿日期：2012-07-06；**修订日期**：2012-09-08

基金项目：国家863高技术研究发展计划基金项目（2012AA050802）；国家电网公司科技攻关团队基金项目（SG11034）

作者简介：李伟伟（1985-），女，山东青岛人，硕士，助理工程师，研究方向为信息安全；张涛（1976-），男，陕西榆林人，硕士，高级工程师，研究方向为信息安全；林为民（1964-），男，江苏连云港人，硕士，高级工程师，研究方向为信息安全；邓松（1980-），男，安徽合肥人，博士，工程师，CCF会员，研究方向为信息安全；时坚（1983-），男，湖南岳阳人，中级工程师，研究方向为信息安全；汪晨（1983-），男，安徽广德人，硕士，助理工程师，研究方向为信息安全。E-mail: liweiwei@epri.sgcc.com.cn

1　体系结构

文本分类可以分为下面几个步骤来完成：首先建立数据集，包含训练集和测试集。然后建立文本表示模型并进行文本特征选择。然后在训练集上进行机器学习，建立分类器。最后进行测试和性能的评价.

在本文中，数据集主要包含了训练集和测试集，其中，训练集包含了敏感文本库和已知分类文本库。训练集中的敏感文本库包含了大量敏感数据文档的词库，主要是为了后面进行机器学习并形成分类器。已知分类数据词库，其中由两个小的词库组成，一个是敏感数据，一个是非敏感数据，用来统计学习后生成判断是否敏感数据的阈值。

文本的表示主要采用的是向量空间模型（vector space model，VSM）的方法[9-10]，通过将文档表示为向量的形式来进行描述和计算。在此基础上提出一种基于内容的敏感数据识别的方法，主要架构如图 1 所示。

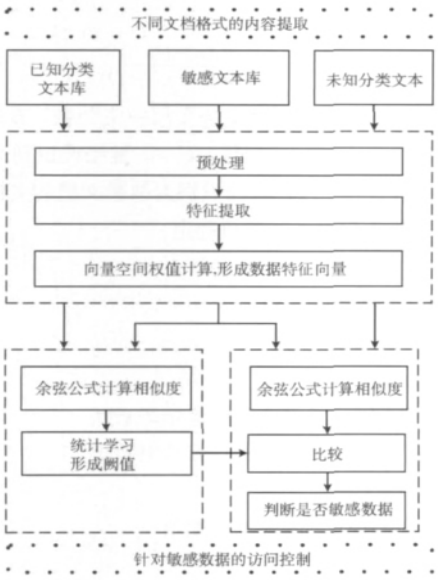

图 1　一种基于内容的敏感数据识别的方法架构

主要实现流程如下：

步骤 1　敏感数据文本库进行预处理和特征提取之后，通过 TFIDF 算法进行向量空间权值计算，形成数据特征向量。

步骤 2　已知分类的文本库进行预处理和特征提取之后，通过 TFIDF 算法计算向量空间权值形成数据特征向量，与将敏感数据形成的特征向量进行余弦计算，并统计学习，根据阈值确定方法确定阈值。

步骤 3　将待判断的未知分类的文档进行预处理和特征提取，通过 TFIDF 算法计算向量空间权值形成数据特征向量后，与敏感数据的特征向量进行余弦计算，将得到的余弦值与阈值进行比较，判断是否是敏感数据。

2　功能组成

2.1　预处理方法

对中文文本的敏感数据识别过程中，首先要进行的就是预处理的阶段，将中文的文本通过中科院计算所汉语此法分析系统 ICTCLAS，划分为单个的词组并标注词性，词长，词频，简便后面的特征提取过程。

文档集合 $T\_pre = \{T_1, T_2, \cdots\cdots, T_i\}$ 通过 ICTCLAS 分词接口，将文本文件进行分词，并在分词的同时统计词长和对词性进行标记，如名词（n）、动词（v）、形容词（a）等。

文本文件 T_i 分词后表示为如下形式

$$T_i = ((a_{i1}, l_{i1}, p_{i1}), (a_{i2}, l_{i2}, p_{i2}), \cdots\cdots, (a_{im}, l_{in}, p_{in}))$$

式中：T_i ——文本 i，a_{in} ——划分出来的词组，l_{in} ——词组的长度，p_{in} ——划分出来的词组的词性。

2.2　特征提取

在文本的学习和识别过程中，如果将所有词性的分词都作为关键词的话，将导致计算量很大，也引入了过多的冗余信息，导致后期的分析存在很大的偏差，因此将预处理之后的分词结果进行提取，降低向量空间的维度，使其更加具有代表性，计算也更加简便有效，如图 2 所示。

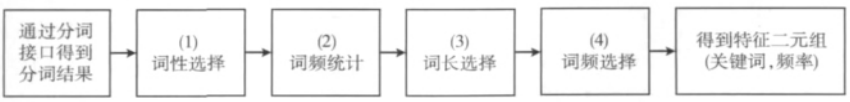

图 2　特征提取流程

（1）词性选择

在中文的文本中，根据词性取其中能够最强烈表达文章内容的关键词，用于后面的特征提取，有助于消除冗余，简便计算过程。因此提取分析后的文本词组中的名词性词组，删除其它词性。文本文件 T_i 经过词性选择后，表示为如下

$$T_{a_i} = ((a_{i1}, l_{i1}), (a_{i2}, l_{i2}), \cdots\cdots, (a_{in'}, l_{in'}))$$

式中：T_{a_i} ——提取名词之后的文本，$(a_{in'}, l_{in'}) \in T_i$ 且 $a_{in'}$ 为名词。

（2）词频统计

统计关键字的出现频率，形成分词三元组，包含词组，词组在本文本中出现的频率和词性。将 T_{a_i} 增加一个词频项，进一步表达为

$$Tb_i = ((a_{i1}, l_{i1}, f_{i1}), (a_{i2}, l_{i2}, f_{i2}), \cdots\cdots, (a_{in'}, l_{in'}, f_{in'}))$$

式中：Tb_i ——统计词频之后的文本，$f_{in'}$ —— $a_{in'}$ 的词频。

(3) 词长选择

在中文的文本中，词比字有着更强的表达能力，计算每个关键字的长度并删除单个字的关键词。进一步表达为

$$Tc_i = ((a_{i1}, f_{i1}), (a_{i2}, f_{i2}), \cdots\cdots, (a_{in''}, f_{in''}))$$

式中：Tc_i ——统计频率之后的文本表示，其中 $a_{in''}$ 为长度大于一个字的关键词。

(4) 词频选择

在中文的文本中，只出现一次的词具有偶然性不具备代表性，因此剔除统计后的文本分词三元组中只出现过一次的词组。得到最终的特征二元组表达为

$$Td_i = ((a_{i1}, f_{i1}), (a_{i2}, f_{i2}), \cdots\cdots, (a_{im}, f_{im}))$$

式中：Td_i ——统计频率之后的文本表示，其中 $f_{im} > 1$ 。

2.3 计算特征向量

2.3.1 计算敏感数据特征向量

经过预处理和特征选择之后的敏感数据文档库表示为

$$T = \{Td_1, Td_2, \cdots\cdots, Td_n\}$$

其中，文本 Td_i 的特征向量表示为

$$Td_i = ((a_{i1}, f_{i1}), (a_{i2}, f_{i2}), \cdots\cdots, (a_{im}, f_{im}))$$

式中：Td_i ——统计频率之后的文本表示，其中 $f_{im} > 1$ 。

对词的权值的计算是衡量特征值的有效方法，目前广泛使用的是基于统计方法的 TF-IDF 公式，这个公式在大量实际使用中被证明是可行的有效的。其核心思想是，认为某个词在其它文本中出现的次数越是少，那么这个词就包含越多的信息，越能够代表文档的类型，相反，如果在其它文档中也是大量的出现，那么这个词就不具有代表性。

目前常用的计 TF-IDF 计算公式表示为

$$d_{ij} = t_{ij} * \log(N/n_j)$$

式中：t_{ij} ——词组 a_{ij} 在文本 T_i 中出现的次数，等于 Td_i 中的 f_{im} ，N——文档的总数，n_j ——文档库中包含词组 a_{ij} 的文档的个数。

由敏感数据组成的特征向量表示为

$$V = ((a_{11}, d_{11}), (a_{12}, d_{12}), \cdots\cdots, (a_{1m}, d_{1m}), \cdots\cdots, (a_{n1}, d_{n1}), (a_{n1}, d_{n1}), \cdots\cdots, (a_{nm}, d_{nm}))$$

简记为

$$V = (d_{11}, d_{12}, \cdots\cdots, d_{1m}, \cdots\cdots, d_{n1}d_{n2}, \cdots\cdots, d_{nm})$$

2.3.2 计算已知分类库和未知分类文档的特征向量

根据敏感数据特征向 V，分别计算对应关键词 a_{ij} 在已知分类的文档库中的权值，得到特征向量如下

$$V' = ((a_{11}, d'_{11}), (a_{12}, d'_{12}), \cdots\cdots, (a_{1m}, d'_{1m}), \cdots\cdots, (a_{n1}, d'_{n1}), (a_{n1}, d'_{n1}), \cdots\cdots, (a_{nm}, d'_{nm}))$$

其中，V' 中的 a_{nm} 等于敏感数据特征向量 V 中的 a_{nm} 。

简记为

$$V' = (d'_{11}, d'_{12}, \cdots\cdots, d'_{1m}, \cdots\cdots, d'_{n1}d'_{n2}, \cdots\cdots, d'_{nm})$$

同样的方法得到未知分类文档的特征向量简记为

$$V'' = (d''_{11}, d''_{12}, \cdots\cdots, d''_{1m}, \cdots\cdots, d''_{n1}d''_{n2}, \cdots\cdots, d''_{nm})$$

2.4 余弦计算

通过余弦公式来计算两个特征向量之间的相似度，余弦相似度计算公式如下

$$\cos\theta = \frac{V_1 \cdot V_2}{\|V_1\| \|V_2\|}$$

式中：V_1 和 V_2 ——两个文档的特征向量，$V_1 \cdot V_2$ ——标准向量点积，定义为 $\sum_{i=1}^{t} V_{1i}V_{2i}$ ，分母中的范数 $\|V_1\|$ 定义为 $\sqrt{V_1 \cdot V_2}$ 。

2.5 阈值确定方法

余弦相似度计算之后，通过对计算结果和阈值的对比，来判断文档是否敏感数据。本文中，采用了已知分类文档学习的方法来确定这个阈值。首先收集安全文件和敏感文件的词库，通过处理与敏感集进行余弦计算，得到值。通过确定相同间隔的阈值，进行是否敏感数据的判断，找到错误率最低的那个阈值作为后面进行未知分类文档判断的阈值。

基于敏感数据的余弦比较中，误判率最低的简单阈值确定方法，误判率如下所示

$$rate = \frac{(B+C)}{(A+B+C+D)}$$

式中：A——被正确识别为安全文档的文档数，B——被错误的识别为敏感文档的文档数，C——被错误识别为安全文档的文档数，D——被正确的识别为敏感文档的文档数。

算法：基于自学习的阈值确定

输入：

i：阈值的初始化值

max_v：最大阈值

$step$：阈值的间隔

$V1[m]$：敏感数据余弦计算结果数组

$V2[m]$：安全数据余弦计算结果数组

输出：

key：错误率最小情况下的阈值

方法：

```
(1) for 每个阈值 i=i+step {
(2)     for V1[m] 数组的每个成员 {
(3)         计算被正确判断的安全文档数 A;
(4)         计算被错误判断的安全文档数 B;}
(5)     for V2[m] 数组的每个成员 {
(6)         计算被错误判断的敏感文档数 C;
(7)         计算被正确判断的敏感文档数 D;}
(8)     //计算当前阈值的错误率
(9)     value_res = (B+C) / (A+B+C+D);
(10)    //判断阈值与最小阈值的大小关系
(11)    if (value_res<value) {
(12)        key=i ;}
(13)}
(14) return key;
```

3　实　验

（1）建立数据集

本系统可以通过更改数据集中训练库的文本类型，来达到不同环境下的敏感数据识别。

在实验中创建一个教育相关的敏感数据识别系统，以 SogouC. reduced. 20061102 语料库的教育部分文档 1600 篇为敏感数据的词库。并随机选取教育相关的文档 100 篇（敏感数据词库之外文档）和非教育相关的文档分别 100 篇，加入到已知分类的文档词库中。随机从其它分类和教育分类中选取文档作为未知分类文档库。

（2）预处理和特征选择

通过分词处理和词性、词频、词长的选择预处理后，经过统计发现，在特征选择过程中，词性选择后符合关键词要求的比例在 30％左右，词长选择后符合关键词要求的比例在 27％左右，词频选择后符合关键词要求的比例在 10％左右。大大的减少了冗余的分词简便了后面的计算。

（3）计算特征向量

根据得到的关键字，经过 TFIDF 算法计算，将敏感数据用向量表示，得到敏感数据的特征向量 V。

计算以敏感数据的关键词为关键词的已知分类的文档 V′和未知分类的特征向量 V″。

（4）计算已知分类与敏感数据余弦值

分别计算包含了相同数目的敏感数据和非敏感数据的已知分类的文档的特征向量 V′与敏感数据特征向量 V 的余弦相似度值。得到余弦相似度值，排序后统计如图 3 所示，其中横坐标表示排序后的文档的编号，纵坐标表示计算得到的余弦相似度值。

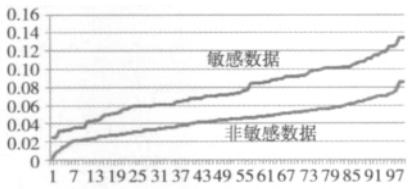

图 3　计算已知分类文档和敏感数据计算得到的余弦统计

由实验数据可知，敏感数据和非敏感数据的余弦相似度是可分的。通过实验方法确定阈值，可以很好的用于未知分类文档的判断。

（5）确定阈值

取一个长度间隔为单位，从取值的底部开始取值，每次增加一个间隔单位。将每次的值设定为阈值，计算这个阈值环境下判断错误率。统计计算后取错误率最低的为实际使用的阈值。实验结果如图 4 所示，其中横坐标表示设定的阈值，纵坐标表示此阈值下的识别错误率。

实验数据可得到，当阈值为 0. 06 的时候，错误率最低为 23％。统计结果图示如下其中横坐标表示阈值，纵坐标表示错误率的值。

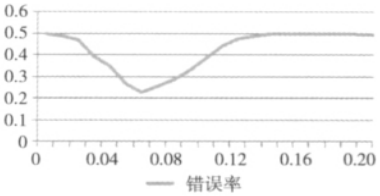

图 4　通过阈值的学习得到错误率最低的阈值

（6）阈值用于敏感文档识别

取未知文档库中的文档，通过上面确定的阈值为标准来进行是否敏感数据的判断，此处取 40 篇，将未知是否敏感数据的文档库中的文档进行预处理和分析，得到以敏感数据关键值为依据的特征向量，与敏感数据特征向量进行计算后，余弦计算得到相似度结果，统计余弦计算结果如下：根据阈值确定的 0. 06，计算未知文档库的识别情况，得到错误率 16. 51％，说明此方法有效可行。

通过多组实验，对比固定阈值和自学习确定阈值对比如图 5 所示，其中横坐标表示实验组，纵坐标表示错误率的值。

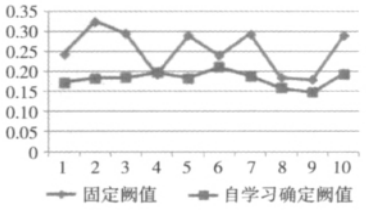

图 5　阈值确定机制错误率对比

通过实验对比发现，自学习确定阈值方法的错较稳定的且基本低于固定的阈值确定机制。

（7）性能测试

根据上述实验验证的阈值，测定敏感数据文档库的数目与错误率的关系如图 6 所示，其中横坐标表示敏感数据文档库的文档数目，纵坐标表示识别的错误率（单位为‰）和初始化时间（单位为秒）。

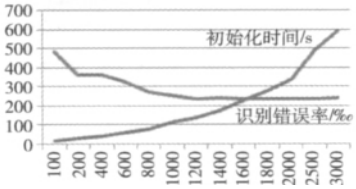

图 6　敏感数据文档库的数目与错误率和初始化时间关系

由实验数据可知，对与学习库的文档数目，并不是越多越好，随着文档数的增多，错误率识别的错误率趋于稳定，而初始化时间递增。

4 结束语

本文设计并实现了一种以词性、词频、词长为参数的简便有效的文本特征提取方法，并通过智能自学习的方式确定阈值，判断文档是否敏感数据法。通过对搜狗语料库中教育相关部分文本的预处理、文本识别、阈值确定的实验结果的分析，较之前通过人工确认阈值的方法具有更高的实用性、准确性和灵活性。此方法可用于在数据防泄漏中对敏感数据的识别和访问控制中。在文本的识别过程中，随着学习库的增加和待处理文本的长度增加，会导致处理效率的降低，还需要进一步的改进。

参考文献：

[1] LIN Zhenbiao. Research on key technologies fo file network-leakage prevention based on data stream analysis [D]. Zhen zhou: PLA Information Engineering University, 2009 (in Chinese). [林臻彪. 基于数据流分析的防文件网络泄露关键技术研究 [D]. 郑州：解放军信息工程大学，2009.]

[2] LI Xiaohong. Feature extraction methods for Chinese text classification [J]. Computer Engineering and Design, 2009, 30 (17): 4127-4129 (in Chinese). [李晓红. 中文文本分类中的特征词抽取方法 [J]. 计算机工程与设计，2009，30 (17)：4127-4129.]

[3] LIU Weiqin. Research on sensitive information monitoring system [D]. Guangzhou: Guangdong University of Technology, 2008 (in Chinese). [刘蔚琴. 网络敏感信息监控系统研究 [D]. 广州：广东工业大学，2008.]

[4] CHEN Liwei, LI Chunyan. Image recognition based on multi-scale semantic analysis [J]. Application Research of Computers, 2009, 26 (2): 799-800 (in Chinese). [陈立伟，李春燕. 一种基于多尺度语义分析的图像识别方法 [J]. 计算机应用研究，2009，26 (2)：799-800.]

[5] PENG Le, XUE Yibo, WANG Chunlu. Survey on recognition and filtering of network video content [J]. Computer Engineering and Design, 2008, 29 (10): 2587-2590 (in Chinese). [彭乐，薛一波，王春露. 网络视频内容的识别和过滤综述 [J]. 计算机工程与设计，2008，29 (10)：2587-2590.]

[6] DING Guanghua, ZHOU Jipeng, ZHOU Min. Design and implementation of parallel bayes classification algorithm using Map Reduce [J]. Microcomputer Information, 2010, 26 (3): 190-192 (in Chinese). [丁光华，周继鹏，周敏. 基于 MapReduce 的并行贝叶斯分类算法的设计与实现 [J]. 微计算机信息，2010，26 (3)：190-192.]

[7] WANG Meifang, LIU Peiyu, ZHU Zhenfang. Feature selection method based on TFIDF [J]. Computer Engineering and Design, 2007, 28 (23): 5795-5796 (in Chinese). [王美方，刘培玉，朱振方. 基于 TFIDF 的特征选择方法 [J]. 计算机工程与设计，2007，28 (23)：5795-5796.]

[8] ZHOU Qifeng, HONG Weicai, SHAO Guifang. Analysis of multi-classification based on SVM in different feature space [J]. Journal of Xiamen University (Natural Science), 2010, 49 (1): 30-33 (in Chinese). [周绮凤，洪文财，邵桂芳. 基于 SVM 的不同特征空间多分类方法研究 [J]. 厦门大学学报(自然科学版)，2010，49 (1)：30-33.]

[9] DUAN Ying. Application of SVM in text categorization [J]. Computer & Digital Engineering, 2012, 40 (7): 87-88 (in Chinese). [段莹. 支持向量机在文本分类中的应用 [J]. 计算机与数字工程，2012，40 (7)：87-88.]

[10] DING Qiong. The research and implement of automatic text classification on system which is based on vector space mode [D]. Shanghai: Tongji University, 2007 (in Chinese). [丁琼. 基于向量空间模型的文本自动分类系统的研究与实现 [D]. 上海：同济大学，2007.]