ISSN 1009-3044
Computer Knowledge and Technology 电脑知识与技术
Vol.16, No.8, March. 2020

E-mail:kfyj@dnzs.net.cn
http://www.dnzs.net.cn
Tel:+86-551-65690963 65690964

网页敏感词过滤与敏感文本分类系统设计

李 伟

(陕西省自强中等专业学校,陕西 宝鸡 721001)

摘要:该文以主动匹配并过滤网页中的敏感词语为目标,研究了文本中敏感词的检索与匹配方法,实现了对网页文本中的敏感词进行主动匹配与过滤处理的过程,设计了一种基于字典树的网页敏感词匹配及过滤方法。并使用决策树方法对含有敏感词的网页文本进行分类。

关键词:网络安全;文本处理;信息过滤;敏感词;决策树

中图分类号:TP391.1 **文献标识码:**A

文章编号:1009-3044(2020)08-0245-03

开放科学(资源服务)标识码(OSID):

Design of Web Sensitive Word Filtering System Based on Decision Tree

LI Wei
(Zi-qiangSecondary Vocational School of Shaanxi Province, Baoji 721000, China)

Abstract: This article aims to actively match and filter sensitive words in web pages, researches the retrieval and matching methods of sensitive words in texts, implements the process of actively matching and filtering sensitive words in web pages, and designs a dictionary-based Web page sensitive word matching and filtering method of tree. And use decision tree method to classify webpage text containing sensitive words.

Key words: network security; text processing; information filtering; sensitive words; decision tree

21世纪以来,随着互联网应用的普及以及应用的日益完善,通过论坛、即时通讯软件以及电子邮件等方式传播敏感信息的网络安全事件发生频繁,在一定程度上严重威胁了社会秩序的稳定和人民群众的正常生活。以合理的技术方式阻止不健康的网络信息在互联网络上肆意传播具有非常积极的意义。

本研究以网页中的文本部分为主要的研究对象,结合对网页中敏感信息的分类,利用自然语言处理对文本的数据挖掘,构建一种基于字典树的网页文本敏感词查找及匹配的模型和算法,并采用决策树的方法对匹配到的敏感文本实现敏感类型的划分。

1 系统结构

本系统主要的处理流程如下:(1)初始化抽取到的网页中的文本数据;(2)查找特定文本中的敏感信息,并做适当处理;(3)构建向量模型表示敏感文本,并得到向量的特征值;(4)构建文本的数据集;(5)构建决策树算法。(6)划分文本的敏感类型。如图1所示。

2 文本内容的获取与分析

敏感词隐藏于文本中,要抽取敏感词,首先要获取网站页面中的文本部分的内容。这个过程就是利用数据挖掘的方式,从页面中获取研究者感兴趣数据的过程。网页上的信息没有严格的格式而言,一般没有结构或者属于半结构化的数据。要在这种无结构化的信息中挖掘其中的文本信息,可以采用DOM遍历的方式来实现。

利用HTML标签的定义可以将标准HTML网页解析为一个树形结构,这个树形结构成为DOM(文档对象模型),采用遍

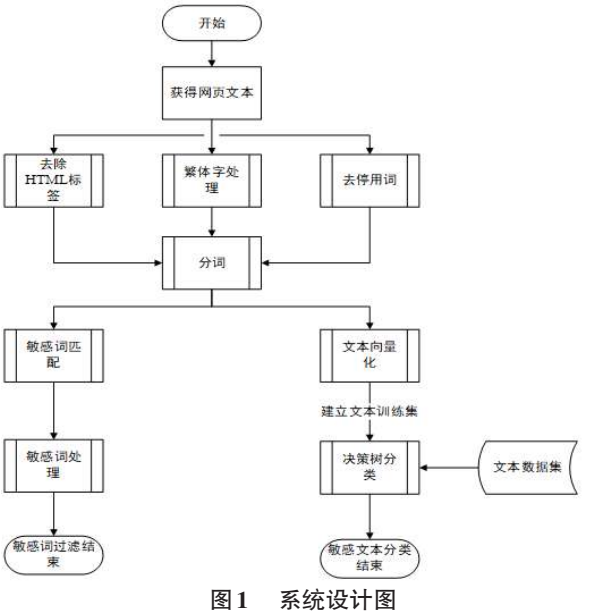

图1 系统设计图

收稿日期:2019-12-21

作者简介:李伟(1977—),男,陕西宝鸡人,讲师,硕士,主要研究方向为计算机网络安全。

历算法，结合文本信息包含在特定HTML标签中的这一特点，确定文本信息所位于的树的节点，并返回这些节点上的数值，即网页中的文本信息内容。DOM文档对象模型可以利用ISO和WC3制定的HTML标准把任何一个网页转变为一棵树型结构。将DOM树的叶子节点上的内容进行提取，再对树进行剪枝处理掉不需要的部分，得到所关心的文本数据。Beautiful Soup作为对中文支持更好的网页解析工具，在文本提取方面更加方便。

3 字典树算法

字典树又称为TIRE树，和哈希表相比，它的查询效率更高，适用在由所有关键词构成的字典汇中查找某些特定关键词语。它的由根节点开始，通过每条有向边分别向下一层节点匹配，最后匹配到底层的叶子结点终止。敏感词字典树型结构如图2所示：

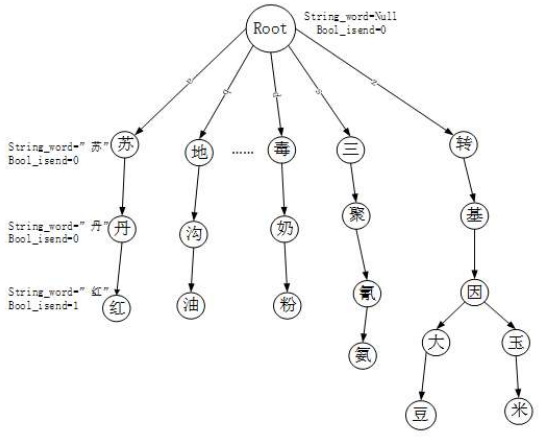

图2 敏感词树示意图

3.1 敏感词匹配过滤的实现

敏感词的匹配实际上要把所有的敏感词构建成一颗敏感词树，并遍历该树形结构，查看是否有从根节点到叶子结点的有效路径，其匹配过程分为建树和查找两个过程。

建树过程：从根节点的第一层节点开始对比要插入的字符，判断当前字符是否存在，若存在则指向该节点。如不存在则创建该字符节点，重复该过程。在遇到字符串末尾结束符时停止该过程，指定末端节点为最后一个非'\0'字符对应的节点。

查找过程：将要插入字符串的第一个字符循环取出，从根节点的第一层开始，查找当前字符是否已经存在，若存在则继续该过程，如不存在则返回False。其算法如下所示。

```
struct Node
Node* child[n]
bool flag;
INSERT( S, Sensitive )
node = Sensitive.root
for i=0 to S.size
c = S[i] - 'p'
ifnode.child[c] == NULL
node.child[c] = new Node
node = node.child[c]
node.flag = true
FIND( S, Trie )
node = Trie.root
for i=0 to S.size
c = S[i] - 'p'
ifnode.child[c] == NULL
return false
node = node.child[c]
returnnode.flag
```

3.2 敏感词预处理及匹配算法设计

网页文本中的敏感信息过滤与匹配步骤如下：

定义敏感词的类别、统计敏感词的数量，，根据敏感信息的类别划分并记录，根据敏感类型设定每种敏感信息的权重。建立敏感词字典树，建立root节点，每颗子树即为敏感词库中的每个敏感词。利用文本分词技术可以将文本中的句子分成若干词语，用这些词语与敏感词字典树进行匹配，判断是否存在敏感信息。为了避免算法过于复杂，结合中文分词工具的一般，规定每个敏感词的汉字数目不超过5个，即构建的字典树的深度小于等于5。

4 敏感文本分类过程

文本分类的过程就是根据文本的特征值将他们划分到规定好的类别中。在敏感词构成的文本分类中，要根据敏感词类别出现的频率和数量来决定敏感文本的类型。

4.1 文本的预处理

分类之前要建立训练和测试用的语料库，如果要对敏感文本进行分类，就要建立不同敏感类型的文本库。建立这里的预处理指的是建立敏感文本语料库，由网站工作人员将敏感页面的文本部分提取出来，根据文本的敏感类型，放入不同的类型目录。可以将这些文本作为文本训练集，并且以随机抽取的方法在这些文件中抽取一定规模的文本作为测试集。

4.2 中文分词

计算机理解不了文本中的句子，因为一般来说，句子是无结构化的数据。如果句子经过分词之后能用若干个关键词来表示句子的意义，那计算机就容易理解。词汇是一种结构化的数据。经过分词操作后，文本可以以词汇的方式简化其意义，变成了一种结构化的数据。

4.3 建立向量空间模型(VSM)

如果文本可以表示为一个空间向量，那么在文本分类时，不同类型文本之间的类别归属就可以抽象为不同向量之间的夹角，以向量的形式表示文本数据，对文本建立向量空间模型，可以将文本分类问题转化为一个数学问题，可以减少处理实际问题的复杂程度。

文本构成的VSM中，一个文本d可以看作由构成该文本的所有单词构成的特征向量。

$$d = \{ (t_1,w_1),(t_2,w_2),\cdots(t_n,w_n) \} \quad (1)$$

其中t_i表示文本d的特征项，w_i为t_i对应的权重，n表示文本的特征维数，如果文本集中的特征维数确定之后，可以只用权重来表示文本d。

$$d = \{ w_1,w_2,\cdots w_n \} \quad (2)$$

4.4 通过计算TF-IDF值得到文本集的权重矩阵

通过计算文本的TF-IDF值来确定文本之间的相似性，进而根据相似程度来进行文本分类。

$$TFIDF(w,d) = TF(w,d) \times IDF(w) \quad (3)$$

上式计算的结果为词w对于文本d的权重。根据这个权重可以建立一个二维矩阵，元素a[i][j]用来表示第j个敏感词在第i个敏感类别中的IF-IDF值。依据同样的方式建立测试及数据的TF-IDF词向量空间模型。测试集和训练集处在同一个词向量空间中，测试集与训练集数据具有不同的敏感词权重矩阵。

4.5 敏感文本分类决策树构建

4.5.1 分词及文本向量的降维处理

在使用向量模型表示文档时，必须要进行语句分词处理。词语本身是有一定意义的，他们组合在一起构成了句子，形式上而言，汉字是组成句子的最小单位，但是单个汉字的意义不明确，所以词从意义表达上来说，是构成语句的最小单位。在敏感文本分类中词语的权重是决定一个文本是否敏感的重要因素。为了降低文本向量的维度，需要去掉其中的语气助词、副词、介词等不具备特定意义的词(除名词与动词之外的词)，进行向量降维，可以加快利用分类算法进行文本分类的速度与精度。

4.5.2 构建敏感词检索模型中的决策树算法

1) 计算每个文本中敏感特征词在由敏感文本构成矩阵中的权重，根据权重的大小确定而对文本进行敏感类别的分类，即预测出这些文本属于哪种具体的敏感类型。利用Python中的scikit-learn工具进行TF-IDF的处理，完成不同类型文本中每个敏感文本的向量化表示。

2) 将多个具有m维特征向量的文本作为样本训练集(训练数据)，利用C4.5算法将其分类到相应的类别中间，以实现敏感文本的分类。

5 实验

敏感文本分类实验采用复旦中文文本分类库，并添加部分其他类别的文本作为训练集。训练集样本点的数目为测试集的3倍左右。经过实验验证，采用决策树对敏感文本的分类准确率能达到80%以上。

6 结束语

本文提出的方案为网页敏感词的主动检索提供了新思路，对敏感文本分类提供了一种新的选择。由于决策树不能很好地处理对连续型属性的量化、编码，文本数据的完整性以及分词算法的优劣都会影响决策树的生成，另外，对决策树合理地进行剪枝能进一步提高决策树分类算法的精确度、简化算法的复杂度。合理的设定阈值并设置剪枝条件从而使建立的学习模型更加简单是日后要关注的一个关键点。

参考文献：

[1] 丁兆云,贾焰,周斌.微博数据挖掘研究综述[J].计算机研究与发展,2014,51(4):691-706.
[2] 袁晓曦.基于机器学习的Web文本自动分类[J].软件导刊,2011,10(1):26-28.
[3] 邓一贵,伍玉英.基于文本内容的敏感词决策树信息过滤算法[J].计算机工程,2014,40(9):300-304.
[4] 李泰.大数据环境下海量多媒体信息过滤技术的改进[J].电子技术与软件工程,2018(4):165.
[5] 李全鑫,魏海平.基于聚类分类法的信息过滤技术研究[J].电子设计工程,2014,22(20):14-16,19.
[6] 宁墨. 信息过滤技术在网站信息监管中的应用与研究[D]. 长春: 吉林大学, 2015.
[7] 李伟. 基于决策树的网页敏感词过滤系统设计[D]. 杨凌: 西北农林科技大学, 2018.

【通联编辑：梁书】