(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 110347804 B
(45)授权公告日 2023.05.12

(21)申请号 201910659196.8
(22)申请日 2019.07.22
(65)同一申请的已公布的文献号
申请公布号 CN 110347804 A
(43)申请公布日 2019.10.18
(73)专利权人 同方知网数字出版技术股份有限公司
地址 100084 北京市海淀区清华大学华业大厦
(72)发明人 张庆国 刘嘉 赵正青
(74)专利代理机构 北京天奇智新知识产权代理有限公司 11340
专利代理师 陈新胜
(51)Int.Cl.
G06F 16/33(2019.01)
G06F 16/31(2019.01)
(56)对比文件
CN 102591941 A,2012.07.18
CN 103619064 A,2014.03.05
CN 104199966 A,2014.12.10
CN 109446198 A,2019.03.08
审查员 郑骏

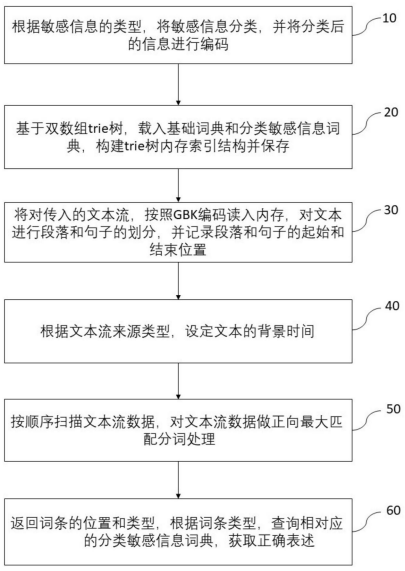
权利要求书1页 说明书5页 附图2页

(54)发明名称

一种线性时间复杂度的敏感信息检测方法

(57)摘要

本发明公开了一种线性时间复杂度的敏感信息检测方法，包括根据敏感信息的类型，将敏感信息分类，并将分类后的信息进行编码；基于双数组trie树，载入基础词典和分类敏感信息词典，构建trie树内存索引结构并保存；对传入的文本流，按照GBK编码读入内存，对文本进行段落和句子的划分，并记录段落和句子的起始和结束位置；根据文本流来源类型，设定文本的背景时间；按顺序扫描文本流数据，对文本流数据做正向最大匹配分词处理；返回词条的位置和类型，根据词条类型，查询相对应的分类敏感信息词典，获取正确表述。本发明支持中文、英文以及中英文和数字混合的敏感信息的检测。由于加入了自然语言处理技术，能够在高速扫描的同时，减少正常信息的误报。



CN 110347804 B

1.一种线性时间复杂度的敏感信息检测方法,其特征在于,所述方法包括:

A根据敏感信息的类型,将敏感信息分类,并将分类后的信息进行编码;

B基于双数组trie树,载入基础词典和分类敏感信息词典,构建trie树内存索引结构并保存;

C对传入的文本流,按照GBK编码读入内存,对文本进行段落和句子的划分,并记录段落和句子的起始和结束位置;

D根据文本流来源类型,设定文本的背景时间;

E按顺序扫描文本流数据,对文本流数据做正向最大匹配分词处理;

F返回词条的位置和类型,根据词条类型,查询相对应的分类敏感信息词典,获取正确表述;

上述步骤B中将trie树内存索引结构保存到本地磁盘,以提高下次加载词典的速度;同时,根据类型载入各分类敏感信息词典,建立分类敏感信息双数组trie树索引结构;

基于双数组trie树,载入基础词典,构建基础trie树内存索引结构,并将trie树内存索引结构保存到本地磁盘,以提高下次加载词典的速度;同时,根据类型载入各分类敏感信息词典,建立分类双数组trie树索引结构;具体地,双数组trie树内存索引包括等长的base数组和check数组,base数组用于存储trie树根节点及所有后裔节点的基地址,check数组用于存储父节点在base数组中的位置;对于一个接收字符c从状态s移动到t的转移,在双数组中保存的条件是:check[base[s]+c]＝s且base[s]+c＝t,其中s和t均为状态;在对双数组中一个节点插入一个新的分支节点时,如果base数组中满足新分支保存条件的位置已经分配给其他节点,需要对该节点的所有分支节点进行重新定位,重新定位的方法是:利用check数组中的空闲单元构建空闲单元链表,每个空闲单元中保存指向下一个空闲单元的指针或单元索引;当需要对节点的所有分支节点进行重新定位时,遍历该链表,查找供节点重新定位的空闲位置并分配给包括新分支节点在内的所有分支节点,并将相应的单元从空闲单元链表中删除。

2.如权利要求1所述的线性时间复杂度的敏感信息检测方法,其特征在于,所述步骤A还包括:整理常见中英文词汇,将中英文词汇类型编码设定为零,并将敏感信息和常见中英文词汇去重后合并,称为基础词典;将敏感信息按照分类整理,形成分类敏感信息词典。

3.如权利要求1所述的线性时间复杂度的敏感信息检测方法,其特征在于,所述步骤D中若无法根据文本流类型确定背景时间,则需要对段落和句子进行时间识别,记录句子和段落背景时间。

4.如权利要求1所述的线性时间复杂度的敏感信息检测方法,其特征在于,所述步骤E中若当前指针开始不是基础词典中词条,根据当前指针指向的内容的ASCII编码值,判断指针移动步长;若为词典中词条,则词条长度作为指针移动的步长,记录词条的前后位置和词条类型,同时,根据词条类型,若为与原则性相违背的相关的词条,则根据背景时间,若背景时间在1950年之前或无法确认背景时间,则重置词条类型为零。

一种线性时间复杂度的敏感信息检测方法

技术领域

[0001] 本发明涉及中英文自然语言处理、信息检索技术领域，尤其涉及一种基于双数组trie树索引结构的中英文敏感信息检测方法。

背景技术

[0002] 中文作为一种复杂的象形文字，存在较多形似、音似字，且由于历史文化的发展，还出现了一些异体字。在信息技术普及之前，属于传统的书籍出版时代，各类信息由具备较高文化素养的知识分子撰写，并经过出版社或编辑部编辑的审校，较少出现错别字和政治敏感信息。但随着信息技术的发展，人们日益使用计算机编写各类文档，信息传播加速，各类信息呈爆炸性增长。在互联网时代，尤其是自媒体爆炸发展的当今，越来越多的人成为互联网信息的提供者，这些人的文化素养良莠不齐，且不再有专业文字编辑的审校，错误信息的出现日渐频繁。

[0003] 信息的爆炸使得人工即时检查变得困难。大量的已经产生的信息，随着时间的推移，也可能出现新的与当前经济和政治形势相关的敏感问题。目前，已经出现了多种技术手段解决上述棘手的问题，最常见的技术手段是基于关键词的信息过滤技术。由于常见错误用词和敏感信息词条数量众多，对单个词条的逐一过滤会导致整个系统的效率低下，难以应对大量并发出现的情况。基于搜索引擎文件索引的方法是另一种常见技术手段，该方法适合对大量文件的批量处理，在应对需要实时处理的突发信息流的情况的时候，该方法会面临窘境。

[0004] 因此，如何快速检测文本信息中的敏感信息，包括错别字词、政治敏感、民族宗教、低俗、色情、暴恐等各类敏感信息，是摆在各出版社、报刊杂志，尤其是政府网站、新闻网站、论坛网站、网络游戏、微博、微信、客户端等信息传播源管理人员面前的亟待解决的问题，也是净化网络环境，营造晴朗网络空间的内在要求。

发明内容

[0005] 为解决上述技术问题，本发明的目的是提供一种线性时间复杂度的敏感信息检测方法。

[0006] 本发明的目的通过以下的技术方案来实现：

[0007] 一种线性时间复杂度的敏感信息检测方法，包括

[0008] A根据敏感信息的类型，将敏感信息分类，并将分类后的信息进行编码；

[0009] B基于双数组trie树，载入基础词典和分类敏感信息词典，构建trie树内存索引结构并保存；

[0010] C对传入的文本流，按照GBK编码读入内存，对文本进行段落和句子的划分，并记录段落和句子的起始和结束位置；

[0011] D根据文本流来源类型，设定文本的背景时间；

[0012] E按顺序扫描文本流数据，对文本流数据做正向最大匹配分词处理；

[0013] F返回词条的位置和类型，根据词条类型，查询相对应的分类敏感信息词典，获取正确表述。

[0014] 与现有技术相比，本发明的一个或多个实施例可以具有如下优点：

[0015] 通过自然语言处理技术和基于双数组trie树的检索技术，可以在O(n)时间复杂度内完成所有类型敏感信息的实时检测。支持中文、英文以及中英文和数字混合的敏感信息的检测。由于加入了自然语言处理技术，能够在高速扫描的同时，减少正常信息的误报。

附图说明

[0016] 图1是线性时间复杂度的敏感信息检测方法流程图；

[0017] 图2是基于双数组trie树进行敏感信息检测的示意图。

具体实施方式

[0018] 为使本发明的目的、技术方案和优点更加清楚，下面将结合实施例及附图对本发明作进一步详细的描述。

[0019] 如图1所示，为线性时间复杂度的敏感信息检测方法流程，包括以下步骤：

[0020] 步骤10根据敏感信息的类型，将敏感信息分类，并将分类后的信息进行编码；

[0021] 步骤20基于双数组trie树，载入基础词典和分类敏感信息词典，构建trie树内存索引结构并保存；

[0022] 步骤30对传入的文本流，按照GBK编码读入内存，对文本进行段落和句子的划分，并记录段落和句子的起始和结束位置；

[0023] 步骤40根据文本流来源类型，设定文本的背景时间；

[0024] 步骤50按顺序扫描文本流数据，对文本流数据做正向最大匹配分词处理；

[0025] 步骤60返回词条的位置和类型，根据词条类型，查询相对应的分类敏感信息词典，获取正确表述。

[0026] 上述步骤10中整理常见中英文词汇，将常见中英文词汇类型编码设定为零；将上述敏感信息和常见中英文词汇去重后合并，称为基础词典。

[0027] 根据敏感信息的类型，将敏感信息按照如下表1分类：

[0028] 表1

[0029]

序号	敏感信息类型	类型编码
1	犯罪、暴恐	101
2	政治错误	102
3	原则性相违背的词条	103
4	淫秽色情	104
5	生僻字	105
6	违禁繁体字	106
7	违禁拼音	107
8	宗教	108
9	异形字（错别字）	109
10	领导人拼写错误	110

[0030]

11	政府公文常见拼写错误	111
12	常见词汇	0

[0031] 为减少敏感信息误识，特加入10万规模的常见词汇，其类型设为0。上述12个类型的词汇排重后合并，建立基础词典文件，文件格式为(key,value)，每对key与value占用一行，value代表词汇类型。上表序号1-11分别建立独立的专项词典文件，格式仍为(key,value)，只是其value对应的是key的正确表述。需要说明的是，部分敏感信息出现即为错误，可能没有对应的正确表述。

[0032] 上述步骤20中将trie树内存索引结构保存到本地磁盘，以提高下次加载词典的速度。同时，根据类型载入各分类敏感信息词典，建立分类敏感信息双数组trie树索引结构。

[0033] 基于双数组trie树，载入基础词典，构建基础trie树内存索引结构，并将trie树内存索引结构保存到本地磁盘，以提高下次加载词典的速度。同时，根据类型载入各分类敏感信息词典，建立分类双数组trie树索引结构。具体地，双数组trie树内存索引包括等长的base数组和check数组，base数组用于存储trie树根节点及所有后裔节点的基地址，check数组用于存储父节点在base数组中的位置。对于一个接收字符c从状态s移动到t的转移，在双数组中保存的条件是:check[base[s]+c]＝s且base[s]+c＝t，其中s和t均为状态。在对双数组中一个节点插入一个新的分支节点时，如果base数组中满足新分支保存条件的位置已经分配给其他节点，在这种情况下，需要对该节点的所有分支节点进行重新定位，方法是:利用check数组中的空闲单元构建空闲单元链表，每个空闲单元中保存指向下一个空闲单元的指针或单元索引。当需要对节点的所有分支节点进行重新定位时，遍历该链表，查找可供这些节点重新定位的空闲位置并分配给包括新分支节点在内的所有分支节点，并将相应的单元从空闲单元链表中删除。。可选的，在构建双数组trie树内存索引时，可以另建立一数组，保存各词条字符串的非公共后缀，以减少双数组的存储空间，并加快词条查找速

度。可选的，对于大型词典，可以将其拆分成若干个较小的子词典，并通过哈希函数将各词条分配到唯一对应的小词典，以加快大型词典的构造速度。

[0034] 上述步骤30中对传入的文本流，按照GBK编码读入内存。对文本进行段落和句子的划分，记录段落和句子的起始和结束位置。句子标识符包括中文句号、问号和感叹号。如果文本结构清晰，还可以记录是否标题段落，段落在文本中的层级，是否带有脚注或引用等结构化数据。

[0035] 上述步骤40中若无法根据文本流类型确定背景时间(精确到年即可)，则需要对段落和句子进行时间识别，记录句子和段落背景时间。

[0036] 传入的文本可以附带背景时间和来源类型两个参数，这两个参数的作用都是用于确定文本的背景时间。文本的来源类型分为新闻报道、政府报告(含政府网站)、科普文章、学术报告、文学作品等。若设定了背景时间，则不需要考虑来源类型参数；若背景时间为空，则需要根据来源类型确定背景时间，其中，新闻报道、政府报告(含政府网站)、科普文章其背景时间均设定为当前年份，其他类型文本背景时间无法确定；若背景时间为空，根据来源类型无法确定背景时间，则需要对段落和句子进行时间识别，记录句子和段落的背景时间。

[0037] 上述步骤50中若当前指针开始不是基础词典中词条，根据当前指针指向的内容的ASCII编码值，判断指针移动步长；若为词典中词条，则词条长度作为指针移动的步长，记录词条的前后位置和词条类型，同时，根据词条类型，若为与原则性相违背的相关的词条，则根据背景时间，若背景时间在1950年之前或无法确认背景时间，则重置词条类型为零。

[0038] 顺序扫描文本流数据，对文本流数据做正向最大匹配分词处理，具体操作步骤如下：

[0039] (1)动态分配词位置数组EndPos[ ]和词类型数组Pos[ ]，数组最大长度为文本长度加一，初始化词位置数组和词类型数组为字符0，初始化词数量计数器nIndex为1。

[0040] (2)声明char*类型指针p_Src指向文本数据，在双数组trie树中查询以*p_Src为开头的词条，查询过程是一个DFA的状态转移过程，返回长度为wordLen的词条。

[0041] (3)若wordLen<1，则该字符构不成词典中词条，若*p_Src<0x80，指针p_Src移动1个位置，否移动2个位置，指针p_Src所处当前位置为EndPos[nIndex]的值，0为Pos[nIndex]的值，词数量计数器自增1。

[0042] (4)若wordLen>0，指针p_Src移动wordLen个位置，指针p_Src所处当前位置为EndPos[nIndex]的值，查询双数组trie树返回的类型值为Pos[nIndex]的值，词数量计数器自增1。若背景时间小于1950年，与相关称谓相关的词条(类型值为103)，Pos[nIndex]的值重置为0。

[0043] (5)循环执行上述2-4步骤，直至指针p_Src指向文本数据结束位置。设EndPos[nIndex]＝-1，Pos[nIndex]＝-1，返回nIndex-1处理结束。

[0044] 表2展示了上述实施例技术方法的有效性和快速性，表2使用7671篇文本进行测试，共计215965350字符(205.96MB)，总共耗时22秒353毫秒，平均速度9.214MB/s。

[0045] 表2

[0046]

字数区间	篇数	敏感信息数	用时
2万字以内	4209	8286	7秒119毫秒
20001字-40000字	1840	22087	5秒857毫秒

40001字-60000字	916	23697	4秒395毫秒
60001字-80000字	306	4866	2秒107毫秒
80001字-100000字	148	5963	942毫秒
100001字-120000字	74	1541	416毫秒
120001字-140000字	51	3387	480毫秒
140001字-160000字	40	1831	420毫秒
160001字-180000字	17	1981	138毫秒
180001字-200000字	23	2175	124毫秒
200001字-220000字	6	105	69毫秒
220001字-240000字	8	2967	127毫秒
240001字-260000字	5	213	37毫秒
260001字-280000字	3	101	10毫秒
30万字以上	10	18751	26毫秒

[0047] 上述实施例支持中文、英文以及中英文和数字混合的敏感信息的检测。由于加入了自然语言处理技术，能够在高速扫描的同时，减少正常信息的误报。

[0048] 虽然本发明所揭露的实施方式如上，但所述的内容只是为了便于理解本发明而采用的实施方式，并非用以限定本发明。任何本发明所属技术领域内的技术人员，在不脱离本发明所揭露的精神和范围的前提下，可以在实施的形式上及细节上作任何的修改与变化，但本发明的专利保护范围，仍须以所附的权利要求书所界定的范围为准。

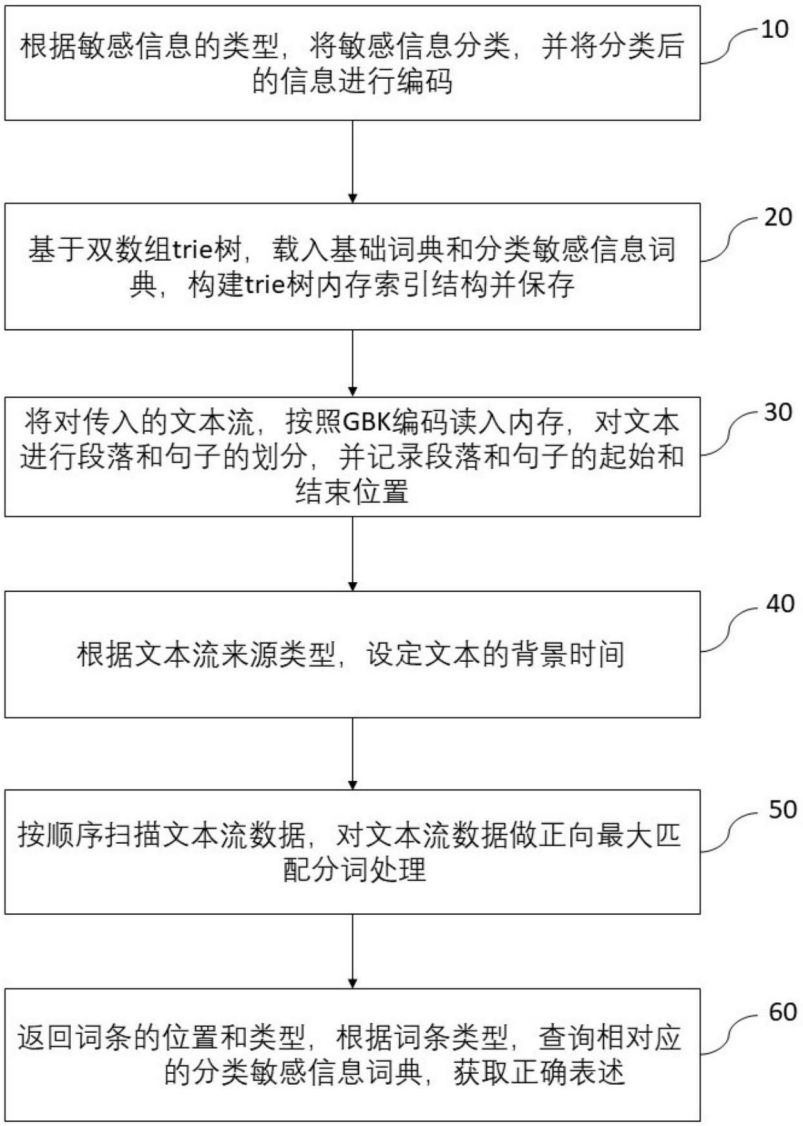

根据敏感信息的类型，将敏感信息分类，并将分类后的信息进行编码
10
基于双数组trie树，载入基础词典和分类敏感信息词典，构建trie树内存索引结构并保存
20
将对传入的文本流，按照GBK编码读入内存，对文本进行段落和句子的划分，并记录段落和句子的起始和结束位置
30
根据文本流来源类型，设定文本的背景时间
40
按顺序扫描文本流数据，对文本流数据做正向最大匹配分词处理
50
返回词条的位置和类型，根据词条类型，查询相对应的分类敏感信息词典，获取正确表述
60

图1

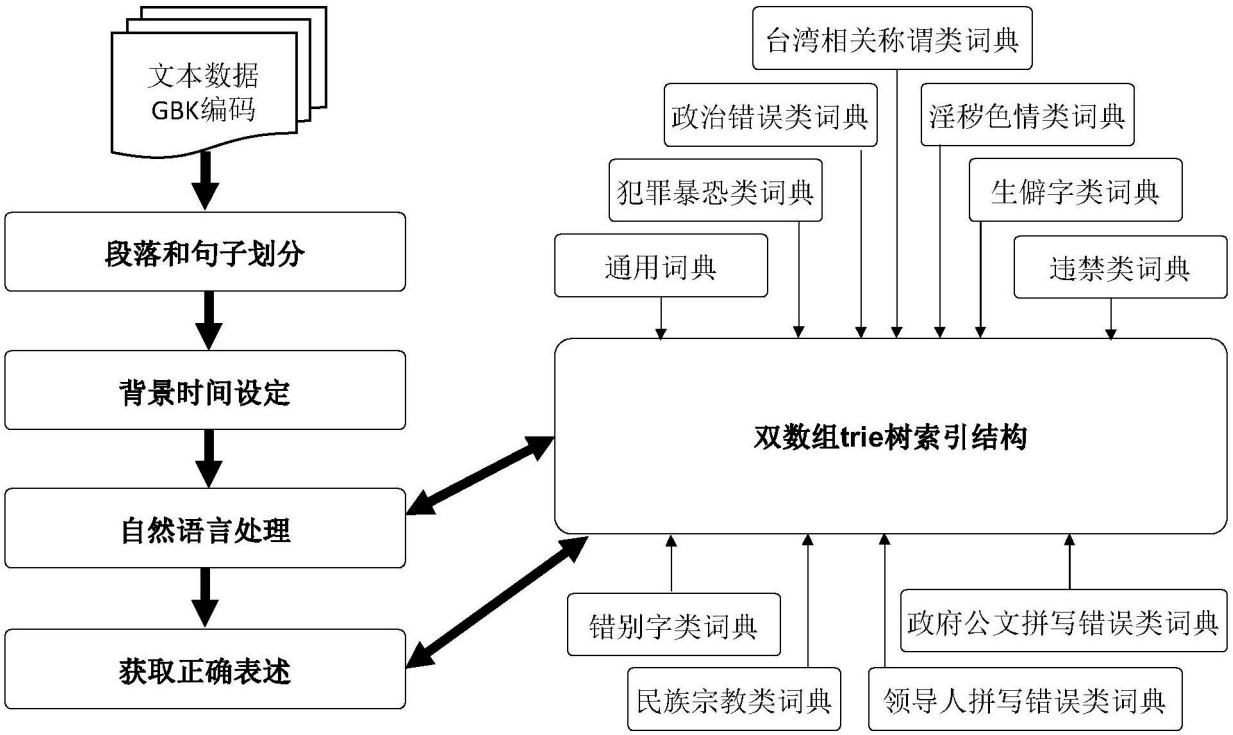

文本数据
GBK编码
段落和句子划分
背景时间设定
自然语言处理
获取正确表述
台湾相关称谓类词典
政治错误类词典
淫秽色情类词典
犯罪暴恐类词典
生僻字类词典
通用词典
违禁类词典
双数组trie树索引结构
错别字类词典
政府公文拼写错误类词典
民族宗教类词典
领导人拼写错误类词典

图2