



(12)发明专利申请

(10)申请公布号 CN 109635107 A

(43)申请公布日 2019.04.16

(21)申请号 201811378557.3

(22)申请日 2018.11.19

(71)申请人 北京亚鸿世纪科技发展有限公司
地址 100095 北京市海淀区高里掌路3号院
2号楼2层201-1至201-8号

(72)发明人 王娜 陈维 林飞 古元 毛华阳
华仲锋

(51)Int.Cl.

G06F 16/35(2019.01)

G06F 16/9535(2019.01)

G06F 17/27(2006.01)

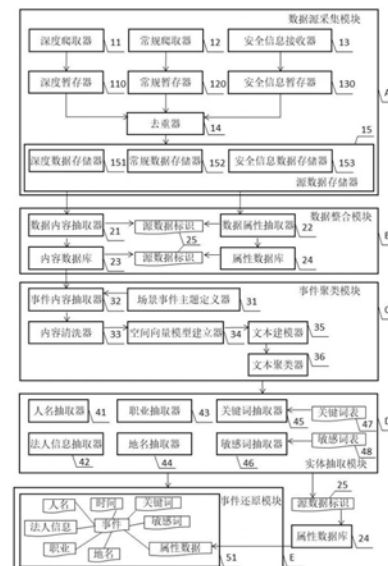
权利要求书3页 说明书11页 附图1页

(54)发明名称

多数据源的语义智能分析及事件场景还原的方法及装置

(57)摘要

多数据源的语义智能分析及事件场景还原的装置涉及信息技术领域,尤其是大数据分析及语义识别技术领域的场景还原。本发明由数据源采集模块、数据整合模块、事件聚类模块、实体抽取模块、事件还原模块组成;本专利解决了监管机构对于大型信息安全事件或者舆情热点信息收集后人工分析工作量大等问题。通过趋近场景还原的方式有效降低了人工量,解决了目前的场景还原存在人工维护量较大的问题。



1. 多数据源的语义智能分析及事件场景还原的装置由数据源采集模块、数据整合模块、事件聚类模块、实体抽取模块、事件还原模块组成；数据采集模块由深度爬取器、常规爬取器、安全信息接收器、深度暂存器、常规暂存器、安全信息暂存器、去重器、源数据存储器组成，其中源数据存储器由深度数据存储器、常规数据存储器、安全信息数据存储器组成；数据整合模块由数据内容抽取器、数据属性抽取器、内容数据库、属性数据库组成；事件聚类模块由场景主题定义器、事件内容抽取器、内容清洗器、空间向量模型建立器、文本建模器、文本聚类器组成；实体抽取模块由人名抽取器、法人信息抽取器、职业抽取器、地名抽取器、关键词抽取器、敏感词抽取器、关键词表、敏感词表组成；

实现多数据源的语义智能分析及事件场景还原的装置的主要步骤包括：

1) 由数据源采集模块进行数据采集

① 深度网页爬取数据：由深度爬取器对已收录监管表单的新闻网站、博客、论坛、微博、微信公众号、社交网站、音视频网站评论进行文本采集并将采集到的文本记录到深度暂存器；

② 常规网页爬虫爬取数据：由常规爬取器爬取非论坛类网站一级域名下的网页内容生成文本并记录到常规暂存器；论坛类网站包括：新闻网站、博客、论坛、微博、微信公众号、社交网站、音视频网站评论；

③ 接收安全信息数据：安全信息接收器作为与信息安全系统的接口，将信息安全系统下达的监测指令所返回的结果数据从信息安全系统备份到安全信息暂存器；

④ 去除重复数据：由去重器将深度暂存器中的数据进行去除重复数据的操作后存储于深度数据存储器；由去重器将常规暂存器中的数据进行去除重复数据的操作后存储于常规数据存储器；由去重器将安全信息暂存器中的数据进行去除重复数据的操作后存储于安全信息数据存储器；

⑤ 由深度数据存储器、常规数据存储器、安全信息数据存储器组成源数据存储器；源数据存储器对所存储的数据按照数据来源生成源数据标识，并将具备源数据标识的深度数据存储器中的数据和具备源数据标识的常规数据存储器中的数据及具备源数据标识的安全信息数据存储器中的数据作为源数据存储于源数据存储器；

2) 由数据整合模块进行数据整合

① 数据的内容抽取：由数据内容抽取器读取源数据存储器中的源数据生成带源数据标识的内容数据并将带源数据标识的内容数据存储于内容数据库，带源数据标识的内容数据包含：源数据标识、标题、作者、文本、音频、视频、图片；

② 数据的属性抽取：由数据属性抽取器读取源数据存储器中的源数据生成带源数据标识的属性数据并将带源数据标识的属性数据存储于属性数据库，带源数据标识的属性数据包含：数据来源URL、内容发表时间、内容浏览量、内容评论量、内容转发量、域名、源ip、目的ip、端口号、机房、首次发现时间、最后发现时间、24小时累计访问量在内的信息安全监测信息；

3) 由事件聚类模块进行主题确认和事件聚类

① 由场景事件主题定义器完成所需还原的事件主题的定义和确认，即完成事件主题的关键词表的内容输入；

② 由事件内容抽取器对存储于内容数据库的全部带源数据标识的内容数据根据事件

主题的关键词表进行抽取,生成主题抽取完成的内容数据;主题抽取完成的内容数据是包含至少一个关键词的带源数据标识的内容数据,关键词是事件主题的关键词表中的关键词;

③由内容清洗器对主题抽取完成的内容数据进行数据清洗生成清洗完成的内容数据,清洗过程首先通过拨测清除无效链接、清除重复无关数据,其次使用jieba分词组件进行分词特征提取,剔除停顿词,语义贡献极小词,无意义词;

④由空间向量模型建立器对清洗完成的内容数据进行空间向量模型建立,建立逻辑为:将一片文本视为若干特征词的序列,该序列可视为一个多维的向量,维度是特征项数量,每个维度大小对应其出现频度以及权重;抽象成公式:文本集合为D,由n个文档组成: $D = \{d_1, d_2, \dots, d_n\}$,其中包含M个特征项 $\{t_1, t_2, \dots, t_n\}$,其中每个文档的都可以使用向量的方法进行抽象化: $d_i = \{(t_{i1}, w_{i1}), (t_{i2}, w_{i2}), \dots, (t_{iM}, w_{iM})\}$,其中 w_{ij} 是第i篇文献中特征项 t_j 的权重;

⑤由文本建模器将清洗完成的内容数据向量化,生成向量化文档集合,具体方法是:

a)对各个特征词进行向量化,使用Word2vec模型利用上下文信息,将每一个特征词转化为固定维度的实数向量,且相似的词在向量空间中也临近,Word2vec模型的skip-gram框架定义的词向量 $v(w)$ 的公式定义为: $v(w) \leftarrow v(w) + \eta \sum_{j=2}^{l(w)} \frac{\partial l(w, u, j)}{\partial x_w}$, η 为学习效率, $\sum_{w \in W} v(w)$ 是内容中的词的向量累加;

b)使用目前最成熟的TF-IDF技术进行文本特征权重赋予,并为之后的文本聚类做基础:假定特征词为t,出现的文本为 \bar{d} 中,如果t出现的频率较高,用TF因子表示;如果t在本文中出现的频率低,但是在全部事件中出现频率较高,用IDF因子表示;

TF*IDF为文档本身特征,基于TF-IDF可以如下表示:
$$\omega(t, \bar{d}) = \frac{tf(t, \bar{d}) \times \log(N/n_t + 0.01)}{\sqrt{\sum_{t \in \bar{d}} [tf(t, \bar{d}) \times \log(N/n_t + 0.01)]^2}}$$

$\omega(t, \bar{d})$ 为特征词t在文本 \bar{d} 中的权重, $tf(t, \bar{d})$ 为词t在 \bar{d} 中的词频,N为训练的文本总数, n_t 为N中出现特征词t的数量;TF-IDF方法能够给在当前文档中出现次数较高在其他文档中出现次数低的特征较高的权重,这样能够增强文档之间的区分度;对于相对应的两个文档 \bar{d}_1 和 \bar{d}_2 ,

其关联度可用其余弦表示:
$$Sim(\bar{d}_1, \bar{d}_2) = \frac{\sum_{k=1}^M \omega_{1k} \times \omega_{2k}}{\sqrt{(\sum_{k=1}^M \omega_{1k}^2)(\sum_{k=1}^M \omega_{2k}^2)}}$$
,其中,M是维度, ω_{ik} 是 \bar{d}_i 的第k维的权重;

c)将获取的词向量和特征词的词权重结合,用以获得整个文档的向量化:通过TF-IDF获取的特征项 t_{ij} 在文档 \bar{d}_i 的中权重为 $\omega(t_{ij})$,特征项 t_{ij} 使用word2vec模型skip-gram框架获得的固定维度的词向量 $v(t_{ij})$;依据上述方法获得参量,可将当前文本转化成特征词和特征权重的序列 $d_i = \{(t_{i1}, w_{i1}), (t_{i2}, w_{i2}), \dots, (t_{iM}, w_{iM})\}$;最终利用该公式向量化所有清洗完成的内容数据,根据源数据标识的差异生成向量化的文档,一个源数据标识生成一份向量化的文档,形成向量化文档集合;

⑥由文本聚类器对向量化文档集合中的向量化的文档进行聚类,使用k-means的算法设定近似文档数量参数K,从而对相似内容的向量化文档进行汇聚,生成已汇聚的向量化文档集;已汇聚的向量化文档集与场景事件主题定义器所定义的事件主题相对应;

4)由实体抽取模块进行特征实体的抽取,包括:

① 由人名抽取器对已汇聚的向量化文档集采用基于角色标注的中国人名抽取方法进行人名实体的抽取,首先使用语料库自动抽取角色信息,并采取Viterbi算法对抽词结果进行角色标注,最终在角色序列的基础上进行最大匹配,实现对于人名的抽取,生成人名信息;人名抽取器将抽取到的人名信息及对应的事件主题发送给事件还原模块;

②由法人信息抽取器对已汇聚的向量化文档集通过第三方工商信息库比对抽取法人信息;法人信息抽取器将抽取到的法人信息及对应的事件主题发送给事件还原模块;

③由职业抽取器对已汇聚的向量化文档集通过常用职业库比对抽取职业信息;职业抽取器将抽取到的职业信息及对应的事件主题发送给事件还原模块;

④由地名抽取器对已汇聚的向量化文档集通过国家、省、市、县名称比对抽取地名信息;地名抽取器将抽取到的地名信息及对应的事件主题发送给事件还原模块;

⑤ 由关键词抽取器对已汇聚的向量化文档集通过关键词表比对抽取关键词信息;关键词抽取器将抽取到的关键词信息及对应的事件主题发送给事件还原模块;关键词表由场景事件主题定义器在定义事件主题时生成并发送给关键词抽取器;

⑥由敏感词抽取器对已汇聚的向量化文档集通过敏感词表比对抽取敏感词信息;敏感词抽取器将抽取到的敏感词信息及对应的事件主题发送给事件还原模块;敏感词表由实体抽取模块根据互联网管理部门统一要求的敏感词内容生成;由敏感词抽取器对已汇聚的向量化文档集通过常用日期、时间,格式匹配抽取时间信息,并将时间信息及对应的事件主题发送给事件还原模块;

5)由事件还原模块完成事件还原,生成关联图谱:

① 由事件还原模块根据收到的事件主题确定对应的已汇聚的向量化文档集,并抽取已汇聚的向量化文档集所对应的源数据标识,根据源数据标识从数据整合模块的属性数据库抽取属性数据;

②由事件还原模块根据事件主题将收到的人名信息、法人信息、职业信息、地名信息、关键词信息、敏感词信息、时间信息及属性数据组合生成关联图谱。

多数据源的语义智能分析及事件场景还原的方法及装置

技术领域

[0001] 本专利涉及信息技术领域,尤其是大数据分析和语言语义分析方面的事件还原应用领域。

背景技术

[0002] 在当今网络化大时代下,各个网络监管机构对于网络上的各种网络大型信息安全事件、舆情热点事件都非常关注。当一个大型信息安全事件或者舆情热点事件发生后,相关的人员都希望能够全面的对于已发生的事件进行全面的了解。然而,现今各个监管机构都缺乏一个能够对于各类型事件进行全面场景还原的手段。目前,相关机构对于各类型事件往往只能通过关键词关联方式,找到相关的事件集合,缺乏对于事件关联性的聚合,并无法进行相应的统计用以进行场景还原。

[0003] 仅依赖爬虫及传统舆情系统缺陷:

首先,舆情热点事件指的是较多数群众所关注的社会现象问题以及表现出来的态度。因此,针对互联网上的舆情态势分析研判需要大量的舆情源数据作为分析对象,数据采集获取的来源越广泛,采集的数据越全面,信息数据越多(如页面内容被反爬),则获得的热点、或舆情分析结果越准确。目前常用的互联网信内容信息数据源仅仅为网络爬虫爬取等方式,由于反爬技术的普及(及需要注册等拦截),使得爬虫爬取的有价值的舆情数据较少,无法准确有效的进行数据分析。

[0004] 其次,原有的关键词搜索方式,获得的舆情分析结果数据无法直观的进行场景还原;在进行初步的情感分析后,展示的内容仅为基于当前采集数据源情况下的舆情正负情况,对于人们所关注的舆情发展情况,舆情严重情况,舆情灾区分布情况等舆情场景都未进行有效的场景还原。

[0005] 仅依赖信息安全采集及传统信息安全系统缺陷:

首先,数据源存在问题,由于采用流量方式采集,在弱过滤情况下信息安全数据采集到的数据量巨大,而微调过滤条件,采集信息游又极少。另外有用流量信息的特质采集的信息中存在大量的垃圾数据,由于大量垃圾信息对正常语义的干扰,无法或非常困难的直接对于采集到的数据进行聚类或者语义分析。

[0006] 最后,监管机构对于各大型安全事件缺少有效的信息归类、筛选、及全面的场景还原手段。目前大部分监管机构仅能通过事件通报的手段对于安全事件进行描述,这在对于大型安全事件的场景及时通过互联网信息直接还原方面是极为欠缺的,对于大型事件的发生发展无法全面的进行还原。

[0007] 总而言之,各个网络监管机构都无法对于网络上突发的大型信息安全事件或者舆情热点事件进行有效的场景还原,而究其根本则在于数据源的获取以及最终分析以及展示手段的缺陷上。因此,本专利将着重对于这些缺陷进行研究,提供一个能够有效的对于监管机构所关注的问题事件进行有效场景还原的方法。

[0008] 共有技术

深度网页爬取数据:相对较为成熟的技术,通过新闻网站、博客、论坛、微博、微信公众号、社交网站、音视频网站评论等地方进行文本采集,获得相应的事件数据源。深度网页爬取,爬取对象为已知的论坛类网站,针对特定内容进行爬取,相比一般网页爬虫爬取内容,噪音少,可用数据量大。

[0009] 常规网页爬虫爬取数据:常用数据采集手段,爬取一级域名下的网页内容。

[0010] 信息安全系统下达的监测指令所返回的结果数据:信息安全系统日常对于可能存在的事件下达监测指令,指令内容包含关键词、出现时间等一系列有可能存在事件问题的指令,返回结果为包含监测指令的网页内容数据以及信息安全属性数据。各类型信息安全系统指令监控数据包括:1、IDC/ISP信安监测数据,数据表包含机房ID、源IP、目的IP、源端口、目的端口、域名、累积访问量、代理类型、代理IP、代理端口、标题、内容、URL、附件、采集时间,原始格式xml,获取周期即时;2、IRCS信安监测数据,数据表包含源IP、目的IP、源端口、目的端口、域名、累积访问量、代理类型、代理IP、代理端口、标题、内容、URL、首次触发时间,原始格式xml,获取周期即时;3、CDN信安监测数据,数据表包含源IP、目的IP、源端口、目的端口、域名、累积访问量、代理类型、代理IP、代理端口、标题、内容、URL、首次触发时间,原始格式xml,获取周期即时。

[0011] word2vec

Word2vec,是为一群用来产生词向量的相关模型。这些模型为浅而双层的神经网络,用来训练以重新建构语言学之词文本。网络以词表现,并且需猜测相邻位置的输入词,在word2vec中词袋模型假设下,词的顺序是不重要的。训练完成之后,word2vec模型可用来映射每个词到一个向量,可用来表示词对词之间的关系,该向量为神经网络之隐藏层。

[0012] 随着计算机应用领域的不断扩大,自然语言处理受到了人们的高度重视。机器翻译、语音识别以及信息检索等应用需求对计算机的自然语言处理能力提出了越来越高的要求。为了使计算机能够处理自然语言,首先需要对自然语言进行建模。自然语言建模方法经历了从基于规则的方法到基于统计方法的转变。从基于统计的建模方法得到的自然语言模型称为统计语言模型。有许多统计语言建模技术,包括n-gram、神经网络以及 log_linear 模型等。在对自然语言进行建模的过程中,会出现维数灾难、词语相似性、模型泛化能力以及模型性能等问题。寻找上述问题的解决方案是推动统计语言模型不断发展的内在动力。在对统计语言模型进行研究的背景下,Google 公司在 2013年开放了 Word2vec这一款用于训练词向量的软件工具。Word2vec 可以根据给定的语料库,通过优化后的训练模型快速有效地将一个词语表达成向量形式,为自然语言处理领域的应用研究提供了新的工具。Word2vec依赖skip-grams或连续词袋(CBOW)来建立神经词嵌入。Word2vec为托马斯·米科洛夫(Tomas Mikolov)在Google带领的研究团队创造。该算法渐渐被其他人所分析和解释。

[0013] 词袋模型

词袋模型(Bag-of-words model)是个在自然语言处理和信息检索(IR)下被简化的表达模型。此模型下,像是句子或是文件这样的文字可以用一个袋子装着这些词的方式表现,这种表现方式不考虑语法以及词的顺序。最近词袋模型也被应用在计算机视觉领域。词袋模型被广泛应用在文件分类,词出现的频率可以用来当作训练分类器的特征。关于“词袋”这个用字的由来可追溯到泽里格·哈里斯于1954年在Distributional Structure的文章。

[0014] Skip-gram 模型

Skip-gram 模型是一个简单但却非常实用的模型。在自然语言处理中,语料的选取是一个相当重要的问题:第一,语料必须充分。一方面词典的词量要足够大,另一方面要尽可能地包含反映词语之间关系的句子,例如,只有“鱼在水中游”这种句式在语料中尽可能地多,模型才能够学习到该句中的语义和语法关系,这和人类学习自然语言一个道理,重复的次数多了,也就会模仿了;第二,语料必须准确。也就是说所选取的语料能够正确反映该语言的语义和语法关系,这一点似乎不难做到,例如中文里,《人民日报》的语料比较准确。但是,更多的时候,并不是语料的选取引发了对准确性问题的担忧,而是处理的方法。 n 元模型中,因为窗口大小的限制,导致超出窗口范围的词语与当前词之间的关系不能被正确地反映到模型之中,如果单纯扩大窗口大小又会增加训练的复杂度。Skip-gram 模型的提出很好地解决了这些问题。顾名思义,Skip-gram 就是“跳过某些符号”,例如,句子“中国足球踢得真是太烂了”有4个3元词组,分别是“中国足球踢得”、“足球踢得真是”、“踢得真是太烂”、“真是太烂了”,可是我们发现,这个句子的本意就是“中国足球太烂”可是上述 4个3元词组并不能反映出这个信息。Skip-gram 模型却允许某些词被跳过,因此可以组成“中国足球太烂”这个3元词组。如果允许跳过2个词,即 2-Skip-gram。

[0015] word2vec的应用

Word2vec用来建构整份文件(而非独立的词)的延伸应用已被提出,该延伸称为 paragraph2vec或doc2vec,并且用C、Python和 Java/Scala实做成工具。Java和Python也支援推断文件嵌入于未观测的文件。对word2vec框架为何做词嵌入如此成功知之甚少,约阿夫·哥德堡(Yoav Goldberg)和欧莫·列维(Omer Levy)指出word2vec的功能导致相似文本拥有相似的嵌入(用余弦相似性计算)并且和约翰·鲁伯特·弗斯的分布假说有关。词嵌入是自然语言处理(NLP)中语言模型与表征学习技术的统称。概念上而言,它是指把一个维数为所有词的数量的高维空间嵌入到一个维数低得多的连续向量空间中,每个单词或词组被映射为实数域上的向量。词嵌入的方法包括人工神经网络、对词语同现矩阵降维、概率模型以及单词所在上下文的显式表示等。在底层输入中,使用词嵌入来表示词组的方法极大提升了NLP中语法分析器和文本情感分析等的效果。词嵌入技术起源于2000年。约书亚·本希奥等人在一系列论文中使用了神经概率语言模型(Neural probabilistic language models)使机器“习得词语的分布式表示(learning a distributed representation for words)”,从而达到将词语空间降维的目的。罗维斯(Roweis)与索尔(Saul)在《科学》上发表了用局部线性嵌入(LLE)来学习高维数据结构的低维表示方法。这个领域开始时稳步发展,在2010年后突飞猛进;一定程度上而言,这是因为这段时间里向量的质量与模型的训练速度有极大的突破。词嵌入领域的分支繁多,有许多学者致力于其研究。2013年,谷歌一个托马斯·米科洛维(Tomas Mikolov)领导的团队发明了一套工具word2vec来进行词嵌入,训练向量空间模型的速度比以往的方法都快。许多新兴的词嵌入基于人工神经网络,而不是过去的 n 元语法模型和非监督式学习。

[0016] 词向量

词向量具有良好的语义特性,是表示词语特征的常用方式。词向量每一维的值代表一个具有一定的语义和语法上解释的特征。所以,可以将词向量的每一维称为一个词语特征。词向量具有多种形式,distributed representation 是其中一种。一个 distributed representation 是一个稠密、低维的实值向量。distributed representation 的每一维

表示词语的一个潜在特征,该特征捕获了有用的句法和语义特性。可见,distributed representation 中的 distributed 一词体现了词向量这样一个特点:将词语的不同句法和语义特征分布到它的每一个维度去表示。

[0017] K-means

K-means算法是硬聚类算法,是典型的基于原型的目标函数聚类方法的代表,它是数据点到原型的某种距离作为优化的目标函数,利用函数求极值的方法得到迭代运算的调整规则。K-means算法以欧式距离作为相似度测度,它是求对应某一初始聚类中心向量V最优分类,使得评价指标J最小。算法采用误差平方和准则函数作为聚类准则函数。

[0018] K-means算法是很典型的基于距离的聚类算法,采用距离作为相似性的评价指标,即认为两个对象的距离越近,其相似度就越大。该算法认为簇是由距离靠近的对象组成的,因此把得到紧凑且独立的簇作为最终目标。

[0019] k个初始类聚类中心点的选取对聚类结果具有较大的影响,因为在该算法第一步中是随机的选取任意k个对象作为初始聚类的中心,初始地代表一个簇。该算法在每次迭代中对数据集中剩余的每个对象,根据其于各个簇中心的距离将每个对象重新赋给最近的簇。当考察完所有数据对象后,一次迭代运算完成,新的聚类中心被计算出来。如果在一次迭代前后,J的值没有发生变化,说明算法已经收敛。

[0020] 算法过程如下:

- 1) 从N个文档随机选取K个文档作为质心;
- 2) 对剩余的每个文档测量其到每个质心的距离,并把它归到最近的质心的类;
- 3) 重新计算已经得到的各个类的质心;
- 4) 迭代2~3步直至新的质心与原质心相等或小于指定阈值,算法结束。

[0021] jieba分词组件:结巴分词组件是开源软件中最好的Python中文分词组件。

[0022] 支持三种分词模式:

精确模式,试图将句子最精确地切开,适合文本分析;

全模式,把句子中所有的可以成词的词语都扫描出来,速度非常快,但是不能解决歧义;

搜索引擎模式,在精确模式的基础上,对长词再次切分,提高召回率,适合用于搜索引擎分词。支持繁体分词,支持自定义词典。可以免费下载安装。

[0023] distribution representation分布表现方法

分布表示(distributional representation):是基于分布假设理论,利用共生矩阵来获取词的语义表示,可以看成是一类获取词表示的方法。

[0024] TF-IDF (Term Frequency - Inverse Document Frequency)

这个算法用来评价一个词(Term)对整个文档的重要程度,它只考虑了两个因素:(1)这个词条在本文档中出现的次数是否高(2)这个词在所有文档中出现的次数是否高。算法的思想很容易搞懂:在本文档中出现次数多的词儿自然是重要的,但是得惩罚那些常用词汇,也就是所有文档中出现的次数都很高词。TF-IDF经常用在搜索引擎,用来计算query与document的相关度。公式看维基百科:<http://en.wikipedia.org/wiki/Tf%E2%80%93idf>

Viterbi算法:维特比算法

维特比算法是一种动态规划算法用于寻找最有可能产生观测事件序列的-维特比路

径-隐含状态序列,特别是在马尔可夫信息源上下文和隐马尔可夫模型中。术语“维特比路径”和“维特比算法”也被用于寻找观察结果最有可能解释相关的动态规划算法。例如在统计句法分析中动态规划算法可以被用于发现最可能的上下文无关的派生(解析)的字符串,有时被称为“维特比分析”。

[0025] 维特比算法由安德鲁·维特比(Andrew Viterbi)于1967年提出,用于在数字通信链路中解卷积以消除噪音。此算法被广泛应用于CDMA和GSM数字蜂窝网络、拨号调制解调器、卫星、深空通信和802.11无线网络中解卷积码。现今也被常常用于语音识别、关键字识别、计算语言学和生物信息学中。例如在语音(语音识别)中,声音信号作为观察到的事件序列,而文本字符串,被看作是隐含的产生声音信号的原因,因此可对声音信号应用维特比算法寻找最有可能的文本字符串。

[0026] 维特比算法的基础可以概括成下面三点:

1. 如果概率最大的路径 p (或者说最短路径)经过某个点,比如途中的 X_{22} ,那么这条路径上的起始点 S 到 X_{22} 的这段子路径 Q ,一定是 S 到 X_{22} 之间的最短路径。否则,用 S 到 X_{22} 的最短路径 R 替代 Q ,便构成一条比 P 更短的路径,这显然是矛盾的。证明了满足最优性原理;

2. 从 S 到 E 的路径必定经过第 i 个时刻的某个状态,假定第 i 个时刻有 k 个状态,那么如果记录了从 S 到第 i 个状态的所有 k 个节点的最短路径,最终的最短路径必经过其中一条,这样,在任意时刻,只要考虑非常有限的最短路即可;

3. 结合以上两点,假定当我们从状态 i 进入状态 $i+1$ 时,从 S 到状态 i 上各个节点的最短路径已经找到,并且记录在这些节点上,那么在计算从起点 S 到第 $i+1$ 状态的某个节点 X_{i+1} 的最短路径时,只要考虑从 S 到前一个状态 i 所有的 k 个节点的最短路径,以及从这个节点到 X_{i+1} , j 的距离即可。

[0027] 语料库

指经科学取样和加工的大规模电子文本库。借助计算机分析工具,研究者可开展相关的语言理论及应用研究。我国有国家语委现代汉语平衡语料库和古籍语料库。

[0028]

发明内容

[0029] 多数据源的语义智能分析及事件场景还原的装置由数据源采集模块、数据整合模块、事件聚类模块、实体抽取模块、事件还原模块组成;数据采集模块由深度爬取器、常规爬取器、安全信息接收器、深度暂存器、常规暂存器、安全信息暂存器、去重器、源数据存储器和源数据存储器组成,其中源数据存储器由深度数据存储器、常规数据存储器、安全信息数据存储器组成;数据整合模块由数据内容抽取器、数据属性抽取器、内容数据库、属性数据库组成;事件聚类模块由场景主题定义器、事件内容抽取器、内容清洗器、空间向量模型建立器、文本建模器、文本聚类器组成;实体抽取模块由人名抽取器、法人信息抽取器、职业抽取器、地名抽取器、关键词抽取器、敏感词抽取器、关键词表、敏感词表组成;

实现多数据源的语义智能分析及事件场景还原的装置的主要步骤包括:

1) 由数据源采集模块进行数据采集

① 深度网页爬取数据:由深度爬取器对已收录监管表单的新闻网站、博客、论坛、微博、微信公众号、社交网站、音视频网站评论进行文本采集并将采集到的文本记录到深度暂

存器；

② 常规网页爬虫爬取数据：由常规爬取器爬取非论坛类网站一级域名下的网页内容生成文本并记录到常规暂存器；论坛类网站包括：新闻网站、博客、论坛、微博、微信公众号、社交网站、音视频网站评论；

③接收安全信息数据：安全信息接收器作为与信息安全系统的接口，将信息安全系统下达的监测指令所返回的结果数据从信息安全系统备份到安全信息暂存器；

④去除重复数据：由去重器将深度暂存器中的数据进行去除重复数据的操作后存储于深度数据存储器；由去重器将常规暂存器中的数据进行去除重复数据的操作后存储于常规数据存储器；由去重器将安全信息暂存器中的数据进行去除重复数据的操作后存储于安全信息数据存储器；

⑤由深度数据存储器、常规数据存储器、安全信息数据存储组成源数据存储；源数据存储对所存储的数据按照数据来源生成源数据标识，并将具备源数据标识的深度数据存储器中的数据和具备源数据标识的常规数据存储器中的数据及具备源数据标识的安全信息数据存储器中的数据作为源数据存储于源数据存储；

2) 由数据整合模块进行数据整合

①数据的内容抽取：由数据内容抽取器读取源数据存储中的源数据生成带源数据标识的内容数据并将带源数据标识的内容数据存储于内容数据库，带源数据标识的内容数据包含：源数据标识、标题、作者、文本、音频、视频、图片；

② 数据的属性抽取：由数据属性抽取器读取源数据存储中的源数据生成带源数据标识的属性数据并将带源数据标识的属性数据存储于属性数据库，带源数据标识的属性数据包含：数据来源URL、内容发表时间，内容浏览量、内容评论量、内容转发量、域名、源ip、目的ip、端口号、机房、首次发现时间、最后发现时间、24小时累计访问量在内的信息安全监测信息；

3) 由事件聚类模块进行主题确认和事件聚类

①由场景事件主题定义器完成所需还原的事件主题的定义和确认，即完成事件主题的关键词表的内容输入；

② 由事件内容抽取器对存储于内容数据库的全部带源数据标识的内容数据根据事件主题的关键词表进行抽取，生成主题抽取完成的内容数据；主题抽取完成的内容数据是包含至少一个关键词的带源数据标识的内容数据，关键词是事件主题的关键词表中的关键词；

③由内容清洗器对主题抽取完成的内容数据进行数据清洗生成清洗完成的内容数据，清洗过程首先通过拨测清除无效链接、清除重复无关数据，其次使用jieba分词组件进行分词特征提取，剔除停顿词，语义贡献极小词，无意义词；

④ 由空间向量模型建立器对清洗完成的内容数据进行空间向量模型建立，建立逻辑为：将一片文本视为若干特征词的序列，该序列可视为一个多维的向量，维度是特征项数量，每个维度大小对应其出现频度以及权重；抽象成公式：文本集合为D，由n个文档组成： $D = \{d_1, d_2, \dots, d_n\}$ ，其中包含M个特征项 $\{t_1, t_2, \dots, t_M\}$ ，其中每个文档的都可以使用向量化的方法进行抽象化： $d_i = \{(t_{i1}, w_{i1}), (t_{i2}, w_{i2}), \dots, (t_{iM}, w_{iM})\}$ ，其中 w_{ij} 是第i篇文献中特征项 t_j 的权重；

⑤由文本建模器将清洗完成的内容数据向量化，生成向量化文档集合，具体方法是：

a) 对各个特征词进行向量化,使用Word2vec模型利用上下文信息,将每一个特征词转化为固定维度的实数向量,且相似的词在向量空间中也临近,Word2vec模型的skip-gram框架定义的词向量 $v(w)$ 的公式定义为:
$$v(w) \leftarrow v(w) + \eta \sum_{j=2}^{l(w)} \frac{\partial l(w;u,j)}{\partial x_w}$$
, η 为学习效率, X_w 是内容中的词的向量累加;

b) 使用目前最成熟的TF-IDF技术进行文本特征权重赋予,并为之后的文本聚类做基础:假定特征词为 t ,出现的文本为 \bar{d} 中,如果 t 出现的频率较高,用TF因子表示;如果 t 在本文中出现的频率低,但是在全部事件中出现频率较高,用IDF因子表示。TF*IDF为文档本身特征,

基于TF-IDF可以如下表示:
$$\omega(t, \bar{d}) = \frac{tf(t, \bar{d}) \times \log(N/n_t + 0.01)}{\sqrt{\sum_{t \in \bar{d}} [tf(t, \bar{d}) \times \log(N/n_t + 0.01)]^2}}$$
, $\omega(t, \bar{d})$ 为特征词 t 在文本 \bar{d} 中的权重, $tf(t, \bar{d})$ 为词 t 在 \bar{d} 中的词频, N 为训练的文本总数, n_t 为 N 中出现特征词 t 的数量;TF-IDF方法能够给在当前文档中出现次数较高在其他文档中出现次数低的特征较高的权重,这样能够增强文档之间的区分度;对于相对应的两个文档 \bar{d}_1 和 \bar{d}_2 ,其关联度可用其余弦表示:

$$\text{Sim}(\bar{d}_1, \bar{d}_2) = \frac{\sum_{k=1}^M \omega_{1k} \times \omega_{2k}}{\sqrt{(\sum_{k=1}^M \omega_{1k}^2)(\sum_{k=1}^M \omega_{2k}^2)}}$$
,其中, M 是维度, ω_{ik} 是 \bar{d}_i 的第 k 维的权重;

c) 将获取的词向量和特征词的词权重结合,用以获得整个文档的向量化:通过TF-IDF获取的特征项 t_{ij} 在文档 \bar{d}_i 的中权重为 $\omega(t_{ij})$,特征项 t_{ij} 使用word2vec模型skip-gram框架获得的固定维度的词向量 $v(t_{ij})$;依据上述方法获得参量,可将当前文本转化成特征词和特征权重的序列 $\bar{d}_i = \{(t_{i1}, w_{i1}), (t_{i2}, w_{i2}), \dots, (t_{iM}, w_{iM})\}$;最终利用该公式向量化所有清洗完成的内容数据,根据源数据标识的差异生成向量化的文档,一个源数据标识生成一份向量化的文档,形成向量化文档集合;

⑥由文本聚类器对向量化文档集合中的向量化的文档进行聚类,使用k-means的算法设定近似文档数量参数 K ,从而对相似内容的向量化文档进行汇聚,生成已汇聚的向量化文档集;已汇聚的向量化文档集与场景事件主题定义器所定义的事件主题相对应;

4) 由实体抽取模块进行特征实体的抽取,包括:

① 由人名抽取器对已汇聚的向量化文档集采用基于角色标注的中国人名抽取方法进行人名实体的抽取,首先使用语料库自动抽取角色信息,并采取Viterbi算法对抽词结果进行角色标注,最终在角色序列的基础上进行最大匹配,实现对于人名的抽取,生成人名信息;人名抽取器将抽取到的人名信息及对应的事件主题发送给事件还原模块;

②由法人信息抽取器对已汇聚的向量化文档集通过第三方工商信息库比对抽取法人信息;法人信息抽取器将抽取到的法人信息及对应的事件主题发送给事件还原模块;

③由职业抽取器对已汇聚的向量化文档集通过常用职业库比对抽取职业信息;职业抽取器将抽取到的职业信息及对应的事件主题发送给事件还原模块;

④由地名抽取器对已汇聚的向量化文档集通过国家、省、市、县名称比对抽取地名信息;地名抽取器将抽取到的地名信息及对应的事件主题发送给事件还原模块;

⑤ 由关键词抽取器对已汇聚的向量化文档集通过关键词表比对抽取关键词信息;关键词抽取器将抽取到的关键词信息及对应的事件主题发送给事件还原模块;关键词表由场

景事件主题定义器在定义事件主题时生成并发送给关键词抽取器；

⑥由敏感词抽取器对已汇聚的向量化文档集通过敏感词表比对抽取敏感词信息；敏感词抽取器将抽取到的敏感词信息及对应的事件主题发送给事件还原模块；敏感词表由实体抽取模块根据互联网管理部门统一要求的敏感词内容生成；由敏感词抽取器对已汇聚的向量化文档集通过常用日期、时间，格式匹配抽取时间信息，并将时间信息及对应的事件主题发送给事件还原模块；

5)由事件还原模块完成事件还原，生成关联图谱：

①由事件还原模块根据收到的事件主题确定对应的已汇聚的向量化文档集，并抽取已汇聚的向量化文档集所对应的源数据标识，根据源数据标识从数据整合模块的属性数据库抽取属性数据；

②由事件还原模块根据事件主题将收到的人名信息、法人信息、职业信息、地名信息、关键词信息、敏感词信息、时间信息及属性数据组合生成关联图谱。

[0030] 有益效果

本专利解决了监管机构对于大型信息安全事件或者舆情热点信息收集后人工分析工作量大等问题。通过趋近场景还原的方式有效降低了人工量，解决了目前的场景还原存在人工维护量较大的问题。

附图说明

[0031] 图1是本发明的组成结构图。

具体实施方式

[0032] 参看图1实现本发明的多数据源的语义智能分析及事件场景还原的装置，包括：数据源采集模块A、数据整合模块B、事件聚类模块C、实体抽取模块D、事件还原模块E；数据源采集模块A由深度爬取器11、常规爬取器12、安全信息接收器13、深度暂存器110、常规暂存器120、安全信息暂存器130、去重器14、源数据存储器组成15，其中源数据存储器15由深度数据存储器151、常规数据存储器152、安全信息数据存储器153组成；数据整合模块B由数据内容抽取器21、数据属性抽取器22、内容数据库23、属性数据库24组成；事件聚类模块C由场景主题定义器31、事件内容抽取器32、内容清洗器33、空间向量模型建立器34、文本建模器35、文本聚类器36组成；实体抽取模块D由人名抽取器41、法人信息抽取器42、职业抽取器43、地名抽取器44、关键词抽取器45、敏感词抽取器46、关键词表47、敏感词表48组成；

实现多数据源的语义智能分析及事件场景还原的装置的主要步骤包括：

1)由数据源采集模块A进行数据采集

①深度网页爬取数据：由深度爬取器11对已收录监管表单的新闻网站、博客、论坛、微博、微信公众号、社交网站、音视频网站评论进行文本采集并将采集到的文本记录到深度暂存器110；

②常规网页爬虫爬取数据：由常规爬取器12爬取非论坛类网站一级域名下的网页内容生成文本并记录到常规暂存器120；论坛类网站包括：新闻网站、博客、论坛、微博、微信公众号、社交网站、音视频网站评论；

③接收安全信息数据：安全信息接收器130作为与信息安全系统的接口，将信息安全系

统下达的监测指令所返回的结果数据从信息安全系统备份到安全信息暂存器130；

④去除重复数据：由去重器14将深度暂存器110中的数据进行去除重复数据的操作后存储于深度数据存储单元151；由去重器14将常规暂存器120中的数据进行去除重复数据的操作后存储于常规数据存储单元152；由去重器14将安全信息暂存器130中的数据进行去除重复数据的操作后存储于安全信息数据存储单元153；

⑤由深度数据存储单元151、常规数据存储单元152、安全信息数据存储单元153组成源数据存储单元15；源数据存储单元15对所存储的数据按照数据来源生成源数据标识25，并将具备源数据标识25的深度数据存储单元151中的数据和具备源数据标识25的常规数据存储单元152中的数据及具备源数据标识25的安全信息数据存储单元153中的数据作为源数据存储于源数据存储单元15；

2) 由数据整合模块B进行数据整合

①数据的内容抽取：由数据内容抽取器21读取源数据存储单元15中的源数据生成带源数据标识的内容数据并将带源数据标识的内容数据存储于内容数据库23，带源数据标识的内容数据包含：源数据标识、标题、作者、文本、音频、视频、图片；

②数据的属性抽取：由数据属性抽取器22读取源数据存储单元15中的源数据生成带源数据标识的属性数据并将带源数据标识的属性数据存储于属性数据库24，带源数据标识的属性数据包含：数据来源URL、内容发表时间、内容浏览量、内容评论量、内容转发量、域名、源ip、目的ip、端口号、机房、首次发现时间、最后发现时间、24小时累计访问量在内的信息安全监测信息；

3) 由事件聚类模块C进行主题确认和事件聚类

①由场景事件主题定义器31完成所需还原的事件主题的定义和确认，即完成事件主题的关键词表47的内容输入；

②由事件内容抽取器32对存储于内容数据库23的全部带源数据标识的内容数据根据事件主题的关键词表35进行抽取，生成主题抽取完成的内容数据；主题抽取完成的内容数据是包含至少一个关键词的带源数据标识的内容数据，关键词是事件主题的关键词表47中的关键词；

③由内容清洗器23对主题抽取完成的内容数据进行数据清洗生成清洗完成的内容数据，清洗过程首先通过拨测清除无效链接、清除重复无关数据，其次使用jieba分词组件进行分词特征提取，剔除停顿词，语义贡献极小词，无意义词；

④由空间向量模型建立器34对清洗完成的内容数据进行空间向量模型建立，建立逻辑为：将一片文本视为若干特征词的序列，该序列可视为一个多维的向量，维度是特征项数量，每个维度大小对应其出现频度以及权重；抽象成公式：文本集合为D，由n个文档组成： $D = \{d_1, d_2, \dots, d_n\}$ ，其中包含M个特征项 $\{t_1, t_2, \dots, t_n\}$ ，其中每个文档的都可以使用向量化的方法进行抽象化： $d_i = \{(t_{i1}, w_{i1}), (t_{i2}, w_{i2}), \dots, (t_{iM}, w_{iM})\}$ ，其中 w_{ij} 是第i篇文献中特征项 t_j 的权重；

⑤由文本建模器35将清洗完成的内容数据向量化，生成向量化文档集合，具体方法是：

a) 对各个特征词进行向量化，使用Word2vec模型利用上下文信息，将每一个特征词转化为固定维度的实数向量，且相似的词在向量空间中也临近，Word2vec模型的skip-gram框架

定义的词向量 $v(w)$ 的公式定义为： $v(w) \leftarrow v(w) + \eta \sum_{j=2}^{l(w)} \frac{\partial l(w, u, j)}{\partial x_w}$ ， η 为学习效率， X_w 是内容

中的词的向量累加；

b) 使用目前最成熟的TF-IDF技术进行文本特征权重赋予,并为之后的文本聚类做基础:假定特征词为 t ,出现的文本为 \bar{d} 中,如果 t 出现的频率较高,用TF因子表示;如果 t 在本文中出现的频率低,但是在全部事件中出现频率较高,用IDF因子表示。TF*IDF为文档本身特征,

基于TF-IDF可以如下表示: $\omega(t, \bar{d}) = \frac{\text{tf}(t, \bar{d}) \times \log(N/n_t + 0.01)}{\sqrt{\sum_{t \in \bar{d}} [\text{tf}(t, \bar{d}) \times \log(N/n_t + 0.01)]^2}}$, $\omega(t, \bar{d})$ 为特征词 t 在文本 \bar{d} 中

的权重, $\text{tf}(t, \bar{d})$ 为词 t 在 \bar{d} 中的词频, N 为训练的文本总数, n_t 为 N 中出现特征词 t 的数量;TF-IDF方法能够给在当前文档中出现次数较高在其他文档中出现次数低的特征较高的权重,这样能够增强文档之间的区分度;对于相对应的两个文档 \bar{d}_1 和 \bar{d}_2 ,其关联度可用其余弦表示:

$\text{Sim}(\bar{d}_1, \bar{d}_2) = \frac{\sum_{k=1}^M \omega_{1k} \times \omega_{2k}}{\sqrt{(\sum_{k=1}^M \omega_{1k}^2)(\sum_{k=1}^M \omega_{2k}^2)}}$,其中, M 是维度, ω_{ik} 是 \bar{d}_i 的第 k 维的权重;

c) 将获取的词向量和特征词的词权重结合,用以获得整个文档的向量化:通过TF-IDF获取的特征项 t_{ij} 在文档 \bar{d}_i 的中权重为 $\omega(t_{ij})$,特征项 t_{ij} 使用word2vec模型skip-gram框架获得的固定维度的词向量 $v(t_{ij})$;依据上述方法获得参量,可将当前文本转化成特征词和特征权重的序列 $\bar{d}_i = \{(t_{i1}, w_{i1}), (t_{i2}, w_{i2}), \dots, (t_{iM}, w_{iM})\}$;最终利用该公式向量化所有清洗完成的内容数据,根据源数据标识25的差异生成向量化的文档,一个源数据标识25生成一份向量化的文档,形成向量化文档集合;

⑥由文本聚类器36对向量化文档集合中的向量化的文档进行聚类,使用k-means的算法设定近似文档数量参数 K ,从而对相似内容的向量化文档进行汇聚,生成已汇聚的向量化文档集;已汇聚的向量化文档集与场景事件主题定义器31所定义的事件主题相对应;

4)由实体抽取模块D进行特征实体的抽取,包括:

①由人名抽取器41对已汇聚的向量化文档集采用基于角色标注的中国人名抽取方法进行人名实体的抽取,首先使用语料库自动抽取角色信息,并采取Viterbi算法对抽词结果进行角色标注,最终在角色序列的基础上进行最大匹配,实现对于人名的抽取,生成人名信息;人名抽取器41将抽取到的人名信息及对应的事件主题发送给事件还原模块E;

②由法人信息抽取器42对已汇聚的向量化文档集通过第三方工商信息库比对抽取法人信息;法人信息抽取器42将抽取到的法人信息及对应的事件主题发送给事件还原模块E;

③由职业抽取器43对已汇聚的向量化文档集通过常用职业库比对抽取职业信息;职业抽取器43将抽取到的职业信息及对应的事件主题发送给事件还原模块E;

④由地名抽取器44对已汇聚的向量化文档集通过国家、省、市、县名称比对抽取地名信息;地名抽取器44将抽取到的地名信息及对应的事件主题发送给事件还原模块E;

⑤由关键词抽取器45对已汇聚的向量化文档集通过关键词表47比对抽取关键词信息;关键词抽取器45将抽取到的关键词信息及对应的事件主题发送给事件还原模块E;关键词表47由场景事件主题定义器31在定义事件主题时生成并发送给关键词抽取器45;

⑥由敏感词抽取器46对已汇聚的向量化文档集通过敏感词表48比对抽取敏感词信息;敏感词抽取器46将抽取到的敏感词信息及对应的事件主题发送给事件还原模块E;敏感词表48由实体抽取模块D根据互联网管理部门统一要求的敏感词内容生成;由敏感词抽取器46对已汇聚的向量化文档集通过常用日期、时间,格式匹配抽取时间信息,并将时间信息及

对应的事件主题发送给事件还原模块E；

5) 由事件还原模块E完成事件还原,生成关联图谱51:

① 由事件还原模块E根据收到的事件主题确定对应的已汇聚的向量化文档集,并抽取已汇聚的向量化文档集所对应的源数据标识25,根据源数据标识25从数据整合模块B的属性数据库24抽取属性数据;

② 由事件还原模块E根据事件主题将收到的人名信息、法人信息、职业信息、地名信息、关键词信息、敏感词信息、时间信息及属性数据组合生成关联图谱51。

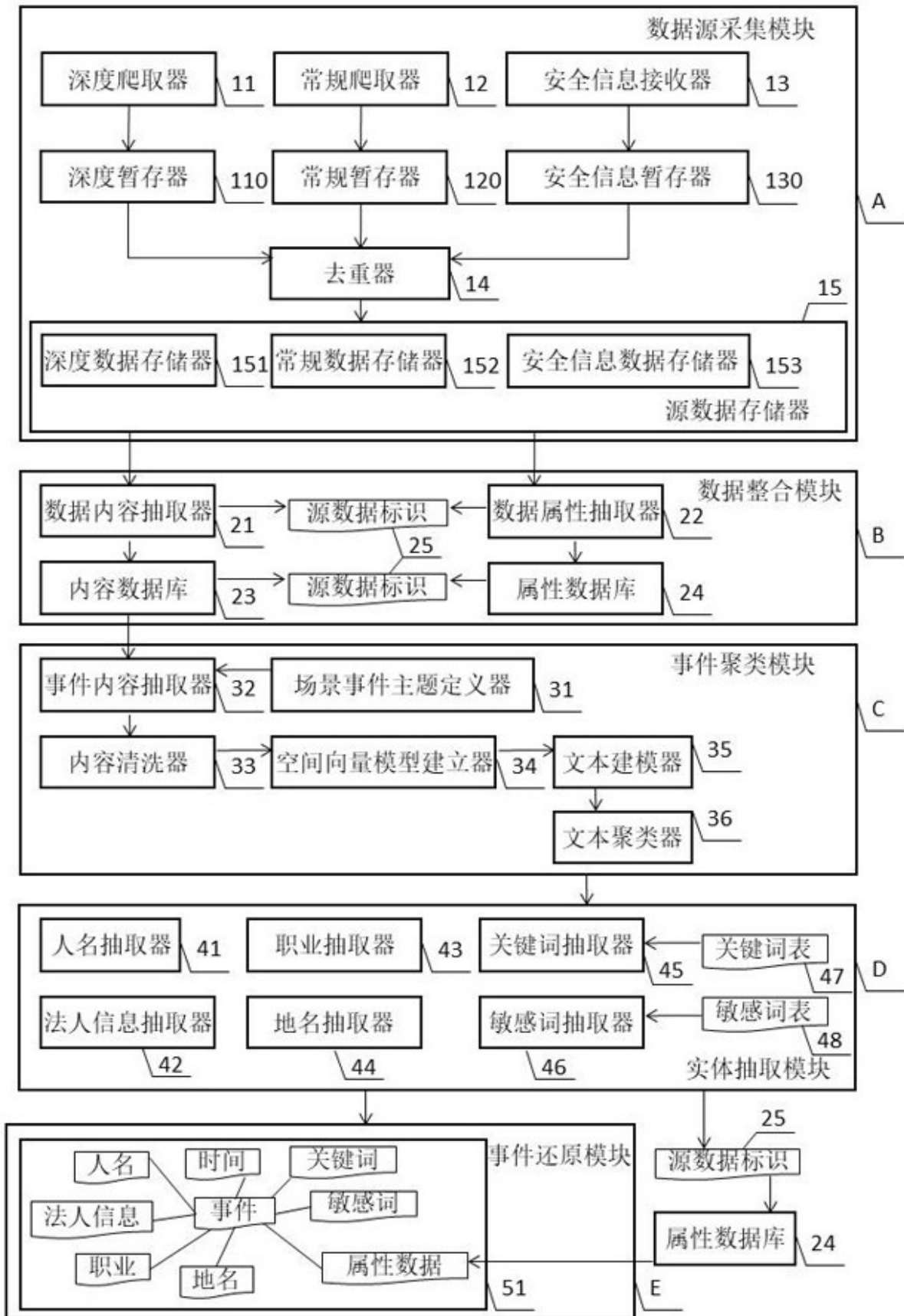


图1