(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 113723085 B
(45)授权公告日 2024.05.24

(21)申请号 202110985907.8
(22)申请日 2021.08.26
(65)同一申请的已公布的文献号
申请公布号 CN 113723085 A
(43)申请公布日 2021.11.30
(73)专利权人 北京航空航天大学
地址 100191 北京市海淀区学院路37号
(72)发明人 连小利 吕鹤阳 黄丹 张莉
(74)专利代理机构 工业和信息化部电子专利中心 11010
专利代理师 罗丹
(51)Int.Cl.
G06F 40/279(2020.01)
G06F 40/237(2020.01)
G06N 3/04(2023.01)

(56)对比文件
CN 102970652 A,2013.03.13
CN 112364165 A,2021.02.12
CN 113282955 A,2021.08.20
JP 2013109475 A,2013.06.06
RU 2662688 C1,2018.07.26
WO 0179957 A2,2001.10.25

审查员 郭王欢

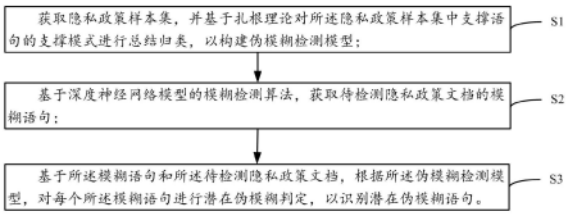
权利要求书2页 说明书8页 附图2页

(54)发明名称

一种隐私政策文档中伪模糊检测方法

(57)摘要

本发明公开了一种隐私政策文档中伪模糊检测方法,所述方法包括:获取隐私政策样本集,并基于扎根理论对所述隐私政策样本集中支撑语句的支撑模式进行总结归类,以构建伪模糊检测模型;基于深度神经网络模型的模糊检测算法,获取待检测隐私政策文档的模糊语句;基于所述模糊语句和所述待检测隐私政策文档,根据所述伪模糊检测模型,对每个所述模糊语句进行潜在伪模糊判定,以识别潜在伪模糊语句。本发明增加了对模糊语句的二次检测,可筛除第一次检测中出现的错误结果,提高了检测的准确性。



CN 113723085 B

1.一种隐私政策文档中伪模糊检测方法,其特征在于,包括:

获取隐私政策样本集,并基于扎根理论对所述隐私政策样本集中支撑语句的支撑模式进行总结归类,以构建伪模糊检测模型;

基于深度神经网络模型的模糊检测算法,获取待检测隐私政策文档的模糊语句;

基于所述模糊语句和所述待检测隐私政策文档,根据所述伪模糊检测模型,对每个所述模糊语句进行潜在伪模糊判定,以识别潜在伪模糊语句;

所述基于扎根理论对所述隐私政策样本集中支撑语句的支撑模式进行总结归类,以构建伪模糊检测模型,包括:

对所述隐私政策样本集中每个所述隐私政策文档的模糊词进行标注,并确定具有模糊词的模糊语句的模糊程度;

判断模糊程度大于阈值的模糊语句在对应的隐私政策文档中是否具有支撑语句,以识别出潜在伪模糊语句;

分析所述潜在伪模糊语句与其支撑语句的特征和关联关系,以对支撑语句的支撑模式进行归类,并确定各个支撑模式的识别算法,以构建伪模糊检测模型。

2.如权利要求1所述的方法,其特征在于,所述支撑模式包括:补充支撑模式;

对于所述补充支撑模式,设计基于关键词匹配和段落结构匹配的识别算法。

3.如权利要求2所述的方法,其特征在于,所述基于所述模糊语句和所述待检测隐私政策文档,根据所述伪模糊检测模型,对每个所述模糊语句进行潜在伪模糊判定,以识别潜在伪模糊语句,包括:

对所述待检测隐私政策文档进行分句分段;

对分句分段后的待检测隐私政策文档进行不完整语句识别,以识别出起始语句和枚举项语句;

将所述模糊语句与所述起始语句和枚举项语句进行相似性检测,输出相似性检测结果大于第一设定值的模糊语句为潜在伪模糊语句。

4.如权利要求1所述的方法,其特征在于,所述支撑模式包括:举例支撑模式;

对于所述举例支撑模式,设计基于关键字匹配的识别算法。

5.如权利要求4所述的方法,其特征在于,所述基于所述模糊语句和所述待检测隐私政策文档,根据所述伪模糊检测模型,对每个所述模糊语句进行潜在伪模糊判定,以识别潜在伪模糊语句,包括:

基于所述待检测隐私政策文档,判断所述模糊语句的下一句是否是以for example/for instance开头的语句,若是,则输出该模糊语句为潜在伪模糊语句。

6.如权利要求1所述的方法,其特征在于,所述支撑模式包括:解释支撑模式;

对于所述解释支撑模式,设计基于关键词特征识别解释型候选语句和识别候选语句中被解释词的识别算法。

7.如权利要求6所述的方法,其特征在于,所述基于所述模糊语句和所述待检测隐私政策文档,根据所述伪模糊检测模型,对每个所述模糊语句进行潜在伪模糊判定,以识别潜在伪模糊语句,包括:

利用关键字匹配获取所述待检测隐私政策文档中的解释型语句;

根据所述待检测隐私政策文档中语句的文本内容、句法结构树和语义依存关系,基于

启发式规则,提取所述解释型语句中的被解释词;

将所述解释型语句中的被解释词与所述模糊语句中的模糊词进行相似性检测,输出相似性检测结果大于第二设定值的模糊语句为潜在伪模糊语句。

8.如权利要求7所述的方法,其特征在于,所述相似性检测包括同义术语判断和基于LCS的词组相似性检测。

9.如权利要求1所述的方法,其特征在于,所述基于深度神经网络模型的模糊检测算法,获取待检测隐私政策文档的模糊语句,包括:

采用Stanford NLP Group提供的分词工具对待检测隐私政策文档进行分句处理;

将分句处理后的待检测隐私政策文档输入到基于深度神经网络模型的模糊检测算法中,以获取所述待检测隐私政策文档的模糊语句。

一种隐私政策文档中伪模糊检测方法

技术领域

[0001] 本发明涉及信息技术处理领域,尤其涉及一种隐私政策文档中伪模糊检测方法。

背景技术

[0002] 近年来,个人与国家都越来越重视用户的隐私问题。隐私政策作为企业与用户之间有约束力的协议,是用户问责和法律监管的依据,必须确保其描述准确无二义。而大量的企业案例以及学术研究证明,隐私政策中存在大量的模糊之处。

[0003] 现有的研究只关注到隐私政策中的模糊词语或者孤立语句,而没有考虑隐私政策中上下文之间的关联。这将导致模糊性检测不够准确,部分检测到的模糊性在隐私政策上下文中存在着对其进行解释支撑的内容。

发明内容

[0004] 本发明实施例提供一种隐私政策文档中伪模糊检测方法,用以解决现有技术检测过程中未考虑隐私政策上下的关联导致模糊性检测不够准确问题。

[0005] 根据本发明实施例的隐私政策文档中伪模糊检测方法,包括:

[0006] 获取隐私政策样本集,并基于扎根理论对所述隐私政策样本集中支撑语句的支撑模式进行总结归类,以构建伪模糊检测模型;

[0007] 基于深度神经网络模型的模糊检测算法,获取待检测隐私政策文档的模糊语句;

[0008] 基于所述模糊语句和所述待检测隐私政策文档,根据所述伪模糊检测模型,对每个所述模糊语句进行潜在伪模糊判定,以识别潜在伪模糊语句。

[0009] 根据本发明的一些实施例,所述基于扎根理论对所述隐私政策样本集中支撑语句的支撑模式进行总结归类,以构建伪模糊检测模型,包括:

[0010] 对所述隐私政策样本集中每个所述隐私政策文档的模糊词进行标注,并确定具有模糊词的模糊语句的模糊程度;

[0011] 判断模糊程度大于阈值的模糊语句在对应的隐私政策文档中是否具有支撑语句,以识别出潜在伪模糊语句;

[0012] 分析所述潜在伪模糊语句与其支撑语句的特征和关联关系,以对支撑语句的支撑模式进行归类,并确定各个支撑模式的识别算法,以构建伪模糊检测模型。

[0013] 根据本发明的一些实施例,所述支撑模式包括:补充支撑模式;

[0014] 对于所述补充支撑模式,设计基于关键词匹配和段落结构匹配的识别算法。

[0015] 根据本发明的一些实施例,所述基于所述模糊语句和所述待检测隐私政策文档,根据所述伪模糊检测模型,对每个所述模糊语句进行潜在伪模糊判定,以识别潜在伪模糊语句,包括:

[0016] 对所述待检测隐私政策文档进行分句分段;

[0017] 对分句分段后的待检测隐私政策文档进行不完整语句识别,以识别出起始语句和枚举项语句;

[0018] 将所述模糊语句与所述起始语句和枚举项语句进行相似性检测,输出相似性检测结果大于第一设定值的模糊语句为潜在伪模糊语句。

[0019] 根据本发明的一些实施例,所述支撑模式包括:举例支撑模式;

[0020] 对于所述举例支撑模式,设计基于关键字匹配的识别算法。

[0021] 根据本发明的一些实施例,所述基于所述模糊语句和所述待检测隐私政策文档,根据所述伪模糊检测模型,对每个所述模糊语句进行潜在伪模糊判定,以识别潜在伪模糊语句,包括:

[0022] 基于所述待检测隐私政策文档,判断所述模糊语句的下一句是否是以for example/forinstance开头的语句,若是,则输出该模糊语句为潜在伪模糊语句。

[0023] 根据本发明的一些实施例,所述支撑模式包括:解释支撑模式;

[0024] 对于所述解释支撑模式,设计基于关键词特征识别解释型候选语句和识别候选语句中被解释词的识别算法。

[0025] 根据本发明的一些实施例,所述基于所述模糊语句和所述待检测隐私政策文档,根据所述伪模糊检测模型,对每个所述模糊语句进行潜在伪模糊判定,以识别潜在伪模糊语句,包括:

[0026] 利用关键字匹配获取所述待检测隐私政策文档中的解释型语句;

[0027] 根据所述待检测隐私政策文档中语句的文本内容、句法结构树和语义依存关系,基于启发式规则,提取所述解释型语句中的被解释词;

[0028] 将所述解释型语句中的被解释词与所述模糊语句中的模糊词进行相似性检测,输出相似性检测结果大于第二设定值的模糊语句为潜在伪模糊语句。

[0029] 根据本发明的一些实施例,所述相似性检测包括同义术语判断和基于LCS的词组相似性检测。

[0030] 根据本发明的一些实施例,所述基于深度神经网络模型的模糊检测算法,获取待检测隐私政策文档的模糊语句,包括:

[0031] 采用StanfordNLPGroup提供的分词工具对待检测隐私政策文档进行分句处理;

[0032] 将分句处理后的待检测隐私政策文档输入到基于深度神经网络模型的模糊检测算法中,以获取所述待检测隐私政策文档的模糊语句。

[0033] 采用本发明实施例,利用结合隐私政策文档上下文的检测方法,对基于深度神经网络模型的模糊检测算法获取的待检测隐私政策文档中的模糊语句进行二次检测,有效的过滤了潜在的伪模糊语句,提高了现有模糊性检测方法的准确性。

[0034] 上述说明仅是本发明技术方案的概述,为了能够更清楚了解本发明的技术手段,而可依照说明书的内容予以实施,并且为了让本发明的上述和其它目的、特征和优点能够更明显易懂,以下特举本发明的具体实施方式。

附图说明

[0035] 通过阅读下文实施方式的详细描述,各种其他的优点和益处对于本领域普通技术人员将变得清楚明了。附图仅用于示出优选实施方式的目的,而并不认为是对本发明的限制。在附图中:

[0036] 图1是本发明实施例中伪模糊检测方法流程图;

[0037] 图2是本发明实施例中补充支撑模式检测流程图；

[0038] 图3是本发明实施例中解释支撑模式检测流程图。

具体实施方式

[0039] 下面将参照附图更详细地描述本发明的示例性实施例。虽然附图中显示了本发明的示例性实施例，然而应当理解，可以以各种形式实现本发明而不应被这里阐述的实施例所限制。相反，提供这些实施例是为了能够更透彻地理解本发明，并且能够将本发明的范围完整的传达给本领域的技术人员。

[0040] 本发明实施例提出一种隐私政策文档中伪模糊检测方法，如图1所示，包括：

[0041] S1，获取隐私政策样本集，并基于扎根理论对所述隐私政策样本集中支撑语句的支撑模式进行总结归类，以构建伪模糊检测模型；

[0042] 这里所述隐私政策样本集为隐私政策文档的集合，其中包括了若干隐私政策文档。

[0043] S2，基于深度神经网络模型的模糊检测算法，获取待检测隐私政策文档的模糊语句；

[0044] S3，基于所述模糊语句和所述待检测隐私政策文档，根据所述伪模糊检测模型，对每个所述模糊语句进行潜在伪模糊判定，以识别潜在伪模糊语句。

[0045] 这里的潜在伪模糊语句可以理解为在隐私政策文档中存在支撑语句对其进行解释说明的语句。

[0046] 本发明实施例通过提前构建的伪模糊检测模型，对基于深度神经网络模型的模糊检测算法所检测出的模糊语句，再次结合待检测的隐私政策文档进行为伪模糊检测，进一步避免了错误检测结果的出现，提高了检测的准确性。

[0047] 在上述实施例的基础上，进一步提出各变型实施例，在此需要说明的是，为了使描述简要，在各变型实施例中仅描述与上述实施例的不同之处。

[0048] 在本发明的一些实施例中，所述基于扎根理论对所述隐私政策样本集中支撑语句的支撑模式进行总结归类，以构建伪模糊检测模型，包括：

[0049] 对所述隐私政策样本集中每个所述隐私政策文档的模糊词进行标注，并确定具有模糊词的模糊语句的模糊程度；

[0050] 在本发明的一些示例中，可以设定多个反映模糊程度的区间，每个区间对应不同的模糊程度。例如，可以设定[1,2]，(2,3]，(3,4]，(4,5]四个区间，分别对应“清晰”，“有点模糊”，“模糊”，“极其模糊”四个类别。

[0051] 分析所述潜在伪模糊语句与其支撑语句的特征和关联关系，以对支撑语句的支撑模式进行归类，并确定各个支撑模式的识别算法，以构建伪模糊检测模型。

[0052] 例如，在第一个周期中使用属性编码(Attributecoding)数据处理策略，专注于分析隐私政策中的模糊语句是否具有“潜在伪模糊”属性，即判断模糊语句在隐私政策全文中是否有支撑语句。本阶段让两名标注者A和B独立阅读15篇隐私政策的全文，并判断模糊语句集在隐私政策全文中是否有对其本身或对其某个模糊词进行支撑的语句。如果有，标注为<潜在伪模糊语句，支撑语句>语句对。在第二周期中本文应用模式编码(Patterncoding)数据处理策略，对支撑语句的支撑模式进行归类。本阶段标注者A和B首先对第一周期中标

注出的<潜在伪模糊语句,支撑语句>语句对进行讨论分析,保留二者均认为支撑语句对模糊语句有支撑效果的语句对,保证标注数据的准确性和一致性。接着对支撑语句对潜在伪模糊语句的支撑关系进行归类,制订归类指南。然后让第三个标注者C独立阅读15篇隐私政策,标注出其中的潜在伪模糊语句及其支撑语句,并根据归类指南对语句对进行分类。最后标注者ABC进行共同讨论,将C的标注结果与AB的标注结果进行对比分析,并合理的改进标注样本,达到最终一致。并对归类指南提出改进意见,精化支撑模式归类。

[0053] 在双周期编码过程中,标注出的潜在伪模糊语句及其支撑语句经过了充分讨论,因此最终结果是准确一致的。本方法也将其作为分析潜在伪模糊语句及其支撑语句识别规则的样本。

[0054] 在本发明的一些实施例中,所述支撑模式包括:补充支撑模式;对于所述补充支撑模式,设计基于关键词匹配和段落结构匹配的识别算法。

[0055] 在此需要说明的是,在阅读隐私政策文档的过程中,发现经常出现在介绍复杂概念或事实时对其进行分条陈述解释的情况。在人工阅读隐私政策时,这是很清晰的一种表达方式。但是自然语言分句时却往往会将这些语句分割开来。在没有上下文语境的情况下,造成了在当前深度学习算法识别过程中被误判成模糊的。不完整语句有两种情况:起始语句和枚举项语句,起始语句与枚举陈述语句互为补充。起始语句是对枚举项语句的概述,对其陈述目标的说明,而枚举项语句是对起始语句的逐条细化。本发明实施例将这种起始语句对枚举陈述语句的目标说明,以及枚举语句对于起始语句的细则陈述定义为补充支撑模式。

[0056] 根据本发明的一些实施例,所述基于所述模糊语句和所述待检测隐私政策文档,根据所述伪模糊检测模型,对每个所述模糊语句进行潜在伪模糊判定,以识别潜在伪模糊语句,包括:

[0057] 对所述待检测隐私政策文档进行分句分段;例如,可以对待检测隐私政策文档进行语句分句的同时保留其段落结构信息,将属于同一段的语句放在一个列表中。

[0058] 对分句分段后的待检测隐私政策文档进行不完整语句识别,以识别出起始语句和枚举项语句;其中,起始语句是对枚举项语句的概述,对其所陈述目标的说明;起始语句往往会明显的以冒号结尾,代表以下的内容是对本句的分条陈述。枚举项语句是对起始语句的逐条细化。枚举项语句的特征较多,包括i)标点符号特征:单个枚举项语句以“;”结尾,所有枚举以“.”结束;ii)顺序特征:语句以数字、字母或罗马数字等开头来组织各枚举项;iii)段落特征:枚举项陈述是多个以主题词开头的段落,每个主题属于所要表达的复杂概念的一方面;iv)特殊表达特征:如今的信息系统都不是孤立存在的,大多会使用一些第三方服务。一般不会对第三方服务进行说明时,而直接给出第三方服务的网址索引。

[0059] 基于以上总结出的五条启发式规则,可以采用正则匹配算法和段落结构匹配算法,从而实现了对补充支撑模式(起始语句及枚举项语句)的自动识别。由于这两种语句在隐私政策中的位置是紧邻的,因此可以先识别起始语句,再判断紧邻在起始语句后的语句是否符合枚举项语句特征。对于补充支撑模式的语句识别过程如图2所示。

[0060] 将所述模糊语句与所述起始语句和枚举项语句进行相似性检测,输出相似性检测结果大于第一设定值的模糊语句为潜在伪模糊语句。

[0061] 所述第一设定值可以基于检测的灵敏度需求以及对检测的要求进行灵活设置。

[0062] 根据本发明的一些实施例,所述支撑模式包括:举例支撑模式;对于所述举例支撑模式,设计基于关键字匹配的识别算法。

[0063] 在陈述一个重要的事实,或较难理解的事务时,人们总爱举例说明。举例说明语句在一定程度上会帮助用户理解该模糊语句。本发明实施例将隐私政策中对模糊语句进行举例说明的语句归为举例支撑模式。

[0064] 本发明的一些实施例,所述基于所述模糊语句和所述待检测隐私政策文档,根据所述伪模糊检测模型,对每个所述模糊语句进行潜在伪模糊判定,以识别潜在伪模糊语句,包括:

[0065] 基于所述待检测隐私政策文档,判断所述模糊语句的下一句是否是以for example/forinstance开头的语句,若是,则输出该模糊语句为潜在伪模糊语句。

[0066] 本发明的一些实施例,所述支撑模式包括:解释支撑模式;对于所述解释支撑模式,设计基于关键词特征识别解释型候选语句和识别候选语句中被解释词的识别算法。

[0067] 解释支撑模式的语句是隐私政策中对模糊语句的某个模糊词进行解释的语句。

[0068] 本发明的一些实施例,所述基于所述模糊语句和所述待检测隐私政策文档,根据所述伪模糊检测模型,对每个所述模糊语句进行潜在伪模糊判定,以识别潜在伪模糊语句,包括:

[0069] 提取隐私政策样本集中的解释型语句,对样本集中的解释型语句的特征进行分析,从而获得解释型语句的识别规则,例如:可以利用关键字匹配获取所述待检测隐私政策文档中的解释型语句;

[0070] 根据所述待检测隐私政策文档中语句的文本内容、句法结构树和语义依存关系,基于启发式规则,提取所述解释型语句中的被解释词;

[0071] 将所述解释型语句中的被解释词与所述模糊语句中的模糊词进行相似性检测,输出相似性检测结果大于第二设定值的模糊语句为潜在伪模糊语句。

[0072] 所述第二设定值可以基于检测的灵敏度需求以及对检测的要求进行灵活设置。

[0073] 本发明的一些实施例,所述相似性判断包括针对模糊词的同义术语判断和基于短语匹配(LCS)的词组相似性检测。

[0074] 本发明的一些实施例,所述基于深度神经网络模型的模糊检测算法,获取待检测隐私政策文档的模糊语句,包括:

[0075] 采用StanfordNLPGroup提供的分词工具对待检测隐私政策文档进行分句处理;

[0076] 将分句处理后的待检测隐私政策文档输入到基于深度神经网络模型的模糊检测算法中,以获取所述待检测隐私政策文档的模糊语句。

[0077] 下面参照图2-图3以一个具体的实施例详细描述根据本发明实施例的隐私政策文档中伪模糊检测方法。值得理解的是,下述描述仅是示例性说明,而不是对本发明的具体限制。凡是采用本发明的相似结构及其相似变化,均应列入本发明的保护范围。

[0078] 首先随机选取了Logan语料库中的15个公司的隐私政策文档来标注和分析潜在伪模糊语句和支撑语句。LoganLebanoff提供的网站隐私政策语料库,该语料库包括100篇网站隐私政策。这些隐私政策是通过亚马逊土耳其机器人网(AmazonMechanicalTurk)收集的,来源于15个类别(从艺术、商业、计算机到科学、购物、体育等)中最常访问的网站。隐私政策共计133K个单词和4.5K个语句。

[0079] 该语句库使用众包的方式标注隐私政策语句的模糊词和模糊程度。对于每个隐私政策语句都招募了五名人员来标注,标注人员需要标注出语句中的模糊词,并对语句的模糊程度进行打分。语句模糊程度的分值为从1到5。接着取五个标注者评分的平均值,语句模糊程度的平均值分布在[1,2],(2,3],(3,4],(4,5]四个区间内,分别对应“清晰”,“有点模糊”,“模糊”,“极其模糊”四个类别。

[0080] 由于本方法是对隐私政策中的模糊语句进行研究,因此首先对隐私政策样本集中的非模糊语句进行了过滤,去除那些分类为清晰的语句,即模糊程度平均分低于2分的语句。最终用于支撑模式归类分析的隐私政策样本集包括(1)15篇隐私政策的原文:以XML格式表示,将隐私政策分为若干段落,每个段落都有一个标题。(2)人工标注的标准答案:以json格式表示,包括模糊语句,语句中的模糊词和语句的模糊程度分值。

[0081] 在第一个周期中使用属性编码(Attributecoding)数据处理策略,专注于分析隐私政策中的模糊语句是否具有“潜在伪模糊”属性,即判断模糊语句在隐私政策全文中是否有支撑语句。本阶段让两名标注者A和B独立阅读15篇隐私政策的全文,并判断模糊语句集在隐私政策全文中是否有对其本身或对其某个模糊词进行支撑的语句。如果有,标注为<潜在伪模糊语句,支撑语句>语句对。在第二周期中本文应用模式编码(Patterncoding)数据处理策略,对支撑语句的支撑模式进行归类。本阶段标注者A和B首先对第一周期中标注出的<潜在伪模糊语句,支撑语句>语句对进行讨论分析,保留二者均认为支撑语句对模糊语句有支撑效果的语句对,保证标注数据的准确性和一致性。接着对支撑语句对潜在伪模糊语句的支撑关系进行归类,制订归类指南。然后让第三个标注者C独立阅读15篇隐私政策,标注出其中的潜在伪模糊语句及其支撑语句,并根据归类指南对语句对进行分类。最后标注者ABC进行共同讨论,将C的标注结果与AB的标注结果进行对比分析,并合理的改进标注样本,达到最终一致。并对归类指南提出改进意见,精化支撑模式归类。

[0082] 在双周期编码过程中,标注出的潜在伪模糊语句及其支撑语句经过了充分讨论,因此最终结果是准确一致的。本方法也将其作为分析潜在伪模糊语句及其支撑语句识别规则的样本。

[0083] 基于扎根理论,本方法根据支撑语句对潜在伪模糊的支撑关系,将潜在伪模糊语句分为了四类:描述现象的潜在伪模糊语句,被补充支撑的潜在伪模糊语句,被举例支撑的潜在伪模糊语句和被解释支撑的潜在伪模糊语句。其中,描述现象的潜在伪模糊语句是没有支撑语句的。这类语句所描述的是其他事物的特征,而与隐私政策讨论的核心内容弱相关。本方法对于这种支撑模式暂不处理,因为涉及到的概念较为宽泛,关乎到具体应用与产品的领域知识,难以统一识别。

[0084] 根据以上对原始数据集的标注与分析,本文将支撑语句归类为以下三种支撑模式:

[0085] 1、基于补充支撑模式

[0086] 由于某些语句所涉及的内容较为复杂,可能会进行分条陈述。这种类型的语句在隐私政策原文中通常包括一个起始语句和几条对其细化的枚举项语句,其中起始语句是对枚举项语句的概述,对其陈述目标的说明,而枚举项语句是对起始语句的逐条细化。在分句的时候,这些语句往往被分开,造成了在当前深度学习算法识别过程中起始语句和枚举陈述语句的不完整。而在此类句子中,起始语句与枚举陈述语句互为补充。因此,本方法将这

种起始语句对枚举陈述语句的目标说明,以及枚举语句对于起始语句的细则陈述定义为补充支撑模式。

[0087] 在阅读隐私政策文档的过程中,发现经常出现在介绍复杂概念或事实时对其进行分条陈述解释的情况。在人工阅读隐私政策时,这是很清晰的一种表达方式。但是自然语言分句时却往往会将这些语句分割开来。在没有上下文语境的情况下,这些语句将被误判成模糊的。不完整语句有两种情况:起始语句和枚举项语句。其中起始语句是对枚举项语句的概述,对其陈述目标的说明,而枚举项语句是对起始语句的逐条细化。

[0088] 本发明将15篇隐私政策中所有的不完整陈述语句提取出来,并从文本内容及段落结构上进行特征分析,总结了补充支撑模式特征。起始语句往往会明显的以冒号结尾,代表以下的内容是对本句的分条陈述。枚举项语句的特征较多,包括i)标点符号特征:单个枚举项语句以“;”结尾,所有枚举以“.”结束;ii)顺序特征:语句以数字、字母或罗马数字等开头来组织各枚举项;iii)段落特征:枚举项陈述是多个以主题词开头的段落,每个主题属于所要表达的复杂概念的一方面;iv)特殊表达特征:如今的信息系统都不是孤立存在的,大多会使用一些第三方服务。一般不会对第三方服务进行说明时,而直接给出第三方服务的网址索引。

[0089] 基于以上总结出的五条启发式规则,本方法采用了正则匹配算法和段落结构匹配算法,从而实现了对补充支撑模式(起始语句及枚举项语句)的自动识别。由于这两种语句在隐私政策中的位置是紧邻的,因此本文先识别起始语句,再判断紧邻在起始语句后的语句是否符合枚举项语句特征。对于补充支撑模式的语句识别过程如图2所示。首先对XML隐私政策原文进行语句分句的同时保留其段落结构信息,将属于同一段的语句放在一个列表中。再对分句分段的隐私政策进行不完整语句识别,识别到起始语句后在从下一句进行枚举项语句识别。接着判断识别到的<起始句,枚举项>语句集中是否有模糊语句,如果有,输出潜在伪模糊语句及其补充型支撑语句。

[0090] 2、举例支撑模式

[0091] 在陈述一个重要的事实,或较难理解的事务时,人们总爱举例说明。举例说明语句在一定程度上会帮助用户理解该模糊语句。本文将隐私政策中对模糊语句进行举例说明的语句归为举例支撑模式。

[0092] 通过对隐私政策文本分析,发现对原文中的前一句进行举例的支撑语句绝大多数都以明显的关键词forexample/forinstance开头。不过也有极少部分的举例型支撑语句不以forexample/forinstance开头。对于该种语句的判断要结合对语句语义的深度理解,极为困难。对于没有特征词的举例语句,本方法暂不识别。

[0093] 对于举例支撑模式,本文的匹配规则直接以判断当前模糊句的下一句是否以forexample/forinstance开头。如果是,则当前句是潜在伪模糊语句,下一句是支撑语句。

[0094] 3、解释支撑模式

[0095] 解释支撑模式的语句是隐私政策中对模糊语句的某个模糊词进行解释的语句。

[0096] 本方法将隐私政策原文中对模糊语句中的模糊词进行解释的语句归类到解释支撑模式。解释型支撑语句及其潜在伪模糊语句一般分布在文档的不同章节,识别起来较为困难,因此对这种模式的识别是本文的研究重点,其流程图如图3所示,主要包括以下三点工作:

[0097] (1)识别解释型语句

[0098] 本阶段对解释型支撑语句的样本进行特征分析,定义解释型语句的识别规则,并实现从隐私政策中识别出解释型语句候选集的识别算法。

[0099] (2)抽取解释型语句的被解释词语

[0100] 本阶段从文本内容、句法解析结构树和语义依存关系三个角度对候选解释语句进行特征分析以定义提取语句中被解释词的启发式规则。接着根据规则实现被解释词提取算法,输出候选解释型语句中的被解释词。

[0101] (3)匹配模糊语句和解释型支撑语句

[0102] 关联模糊语句和解释型支撑语句的桥梁是术语,即模糊语句中的模糊词,也是解释型语句的被解释词。本文对隐私政策中的所有模糊语句的模糊词与候选解释型语句的被解释词进行匹配。模糊语句的模糊词如果与解释型语句的被解释词相似,那么该模糊语句属于潜在伪模糊语句,解释型语句即是其支撑语句。

[0103] 根据本发明的一些实施例,将基于扎根理论分析和标注了15篇隐私政策作为训练数据集,根据支撑型语句对潜在伪模糊语句的支撑关系,将支撑语句归类为“补充支撑模式”,“举例支撑模式”,“解释支撑模式”三种支撑模式。接下来手工分析不同模式下支撑语句的文本特征,定义模式识别的启发式规则。对补充支撑模式提出了5条识别规则,对举例支撑模式提出了1条识别规则。解释支撑模式的识别较为复杂,包括三个步骤。(i)利用关键字匹配获取解释型候选语句。(ii)手工分析语句的文本内容,句法结构树和语义依存关系,定义了提取被解释词的5条启发式规则。并根据启发式规则提取隐私政策中所有解释型语句的被解释词。(iii)将解释型语句的被解释词与隐私政策模糊语句的模糊词进行相似性检测,识别出解释支撑模式的潜在伪模糊语句及支撑语句。其中相似性检测包括同义术语判断和基于LCS的词组相似性检测。

[0104] 对于上述三种支撑模式分别定义识别支撑语句的启发式规则和匹配模糊语句和支撑语句的启发式规则,基于启发式规则给出潜在伪模糊及其支撑语句的识别算法。

[0105] 本领域技术人员可以理解,实现上述实施例方法的全部或部分流程,可以通过计算机程序来指令相关的硬件来完成,所述的程序可存储于计算机可读存储介质中。其中,所述计算机可读存储介质为磁盘、光盘、只读存储记忆体或随机存储记忆体等。

[0106] 与现有技术相比,在进行模糊检测时,本发明采用了一种结合隐私上下文的模糊性检测方法:首先基于现有的模糊性检测算法识别隐私政策中的模糊语句及模糊词。再通过识别模糊语句是否具有支撑语句来过滤掉这些潜在伪模糊语句,从而提高现有的模糊性检测方法的准确性。

[0107] 需要说明的是,以上所述仅为本发明的优选实施例而已,并不用于限制本发明,且本发明的各实施例可进行自由组合实施,对于本领域的技术人员来说,本发明可以有各种更改和变化。凡在本发明的精神和原则之内,所作的任何修改、等同替换、改进等,均应包含在本发明的保护范围之内。

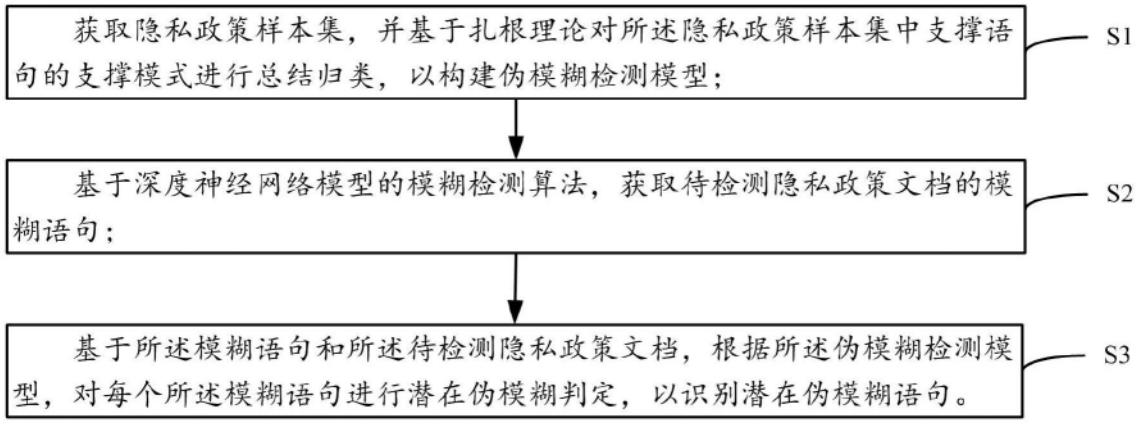

获取隐私政策样本集，并基于扎根理论对所述隐私政策样本集中支撑语句的支撑模式进行总结归类，以构建伪模糊检测模型；
S1
基于深度神经网络模型的模糊检测算法，获取待检测隐私政策文档的模糊语句；
S2
基于所述模糊语句和所述待检测隐私政策文档，根据所述伪模糊检测模型，对每个所述模糊语句进行潜在伪模糊判定，以识别潜在伪模糊语句。
S3

图1

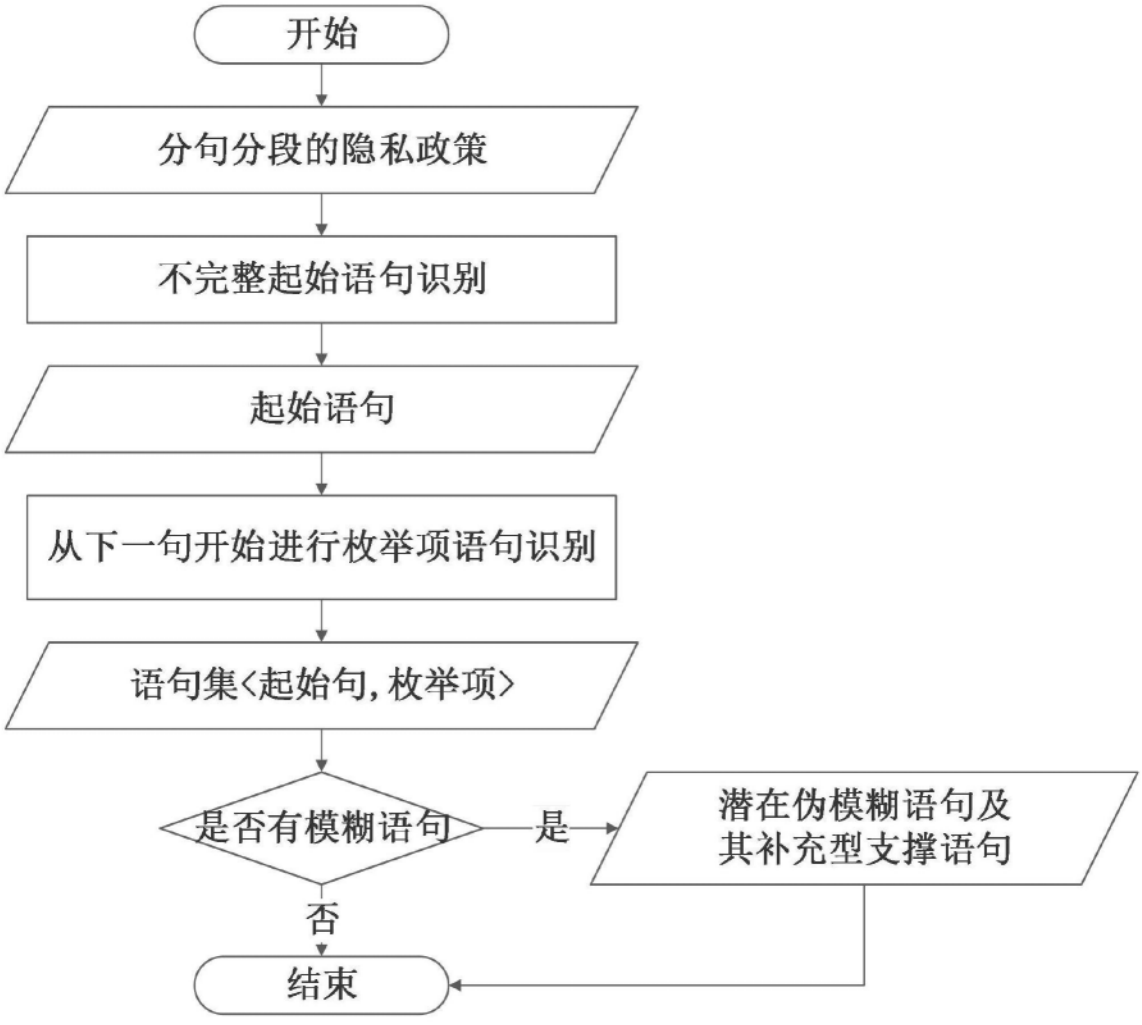

开始
分句分段的隐私政策
不完整起始语句识别
起始语句
从下一句开始进行枚举项语句识别
语句集<起始句,枚举项>
是否有模糊语句
是
潜在伪模糊语句及
其补充型支撑语句
否
结束

图2

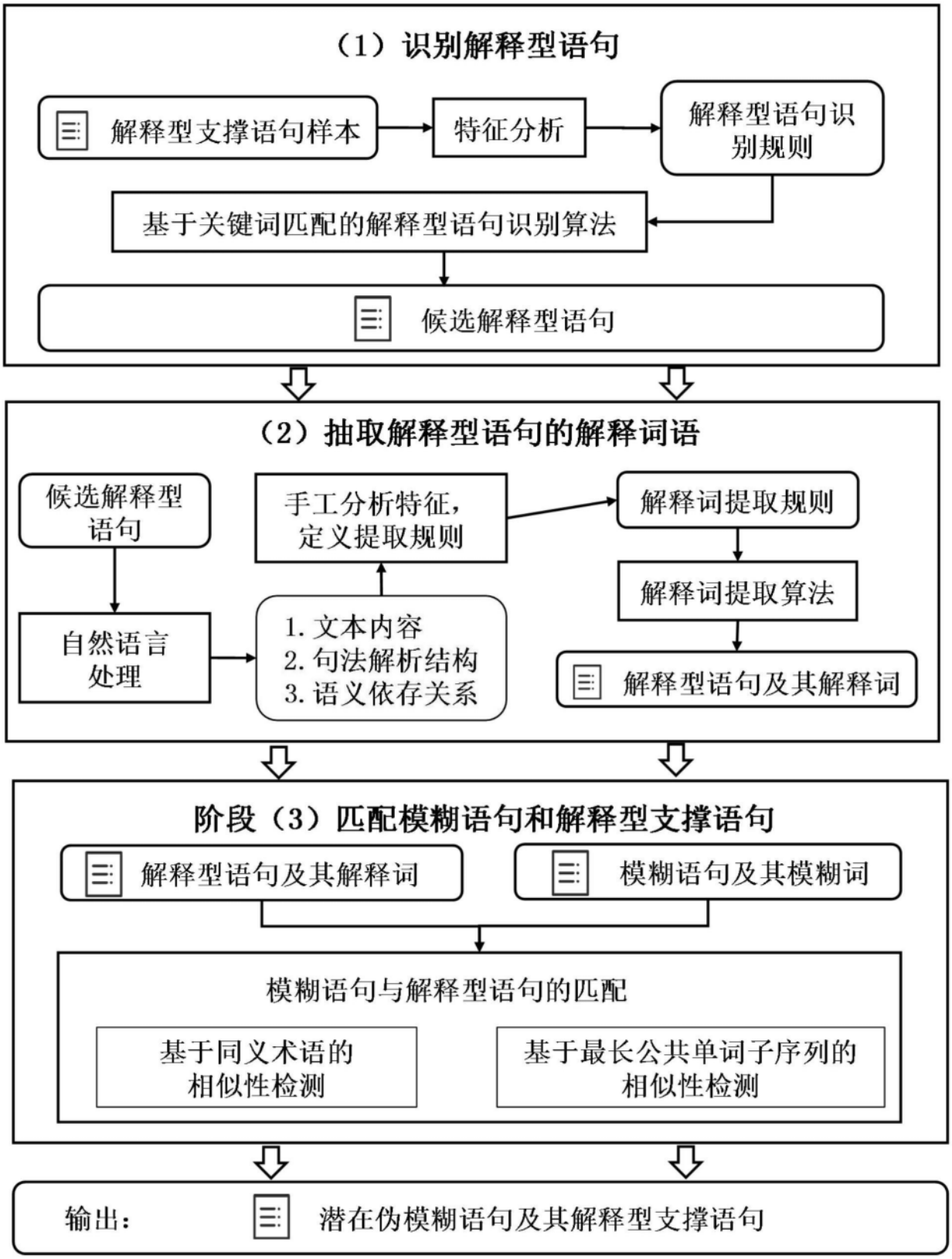

（1）识别解释型语句
解释型支撑语句样本
特征分析
解释型语句识别规则
基于关键词匹配的解释型语句识别算法
候选解释型语句
（2）抽取解释型语句的解释词语
候选解释型语句
手工分析特征，定义提取规则
解释词提取规则
解释词提取算法
自然语言处理
1. 文本内容
2. 句法解析结构
3. 语义依存关系
解释型语句及其解释词
阶段（3）匹配模糊语句和解释型支撑语句
解释型语句及其解释词
模糊语句及其模糊词
模糊语句与解释型语句的匹配
基于同义术语的相似性检测
基于最长公共单词子序列的相似性检测
输出：
潜在伪模糊语句及其解释型支撑语句

图3