



(12) 发明专利申请

(10) 申请公布号 CN 112380318 A

(43) 申请公布日 2021.02.19

(21) 申请号 202011260570.6

G06F 40/30 (2020.01)

(22) 申请日 2020.11.12

G06K 9/62 (2006.01)

G06Q 50/26 (2012.01)

(71) 申请人 中国科学技术大学智慧城市研究院
(芜湖)

地址 241000 安徽省芜湖市智慧协同创新中心

(72) 发明人 陈恩红 陈钢

(74) 专利代理机构 北京润平知识产权代理有限公司 11283

代理人 董杰

(51) Int.Cl.

G06F 16/31 (2019.01)

G06F 16/33 (2019.01)

G06F 16/35 (2019.01)

G06F 40/289 (2020.01)

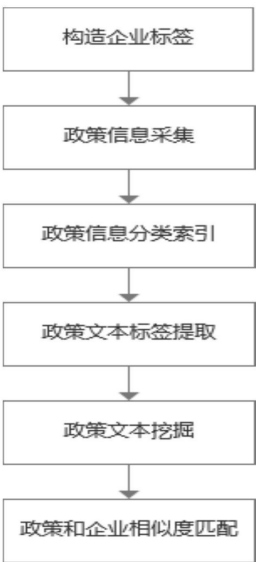
权利要求书3页 说明书6页 附图1页

(54) 发明名称

基于标签相似度的企业政策匹配方法

(57) 摘要

本发明公开了一种基于标签相似度的企业政策匹配方法,包括:步骤1、自动化构建企业标签;步骤2、自动化构建政策标签;步骤3、基于标签相似度的政策匹配;步骤1中包括:A1、构造企业基本属性标签;B1、构造企业经营状况标签;C1、构造企业风险信息标签;步骤2中包括:A2、政策信息自动化采集;B2、政策信息自动化分类索引;C2、政策文本标签提取;D2、政策文本挖掘。该方法能够快速高效地获取政策信息并将政策供给侧和需求侧精准匹配,保证政策落实效果,提高政府政务服务质量。



1. 一种基于标签相似度的企业政策匹配方法,其特征在于,包括:

步骤1、自动化构建企业标签;

步骤2、自动化构建政策标签;

步骤3、基于标签相似度的政策匹配;其中,

步骤1中包括:

A1、构造企业基本属性标签,包括经营范围标签、行业标签、司龄标签、地域标签和规模标签;

B1、构造企业经营状况标签,包括企业创新力标签、企业竞争力标签、企业发展潜力标签和企业发展动力标签;

C1、构造企业风险信息标签,包括企业自身风险标签、企业周边风险标签、预警提醒标签和经营风险标签;

步骤2中包括:

A2、政策信息自动化采集;

B2、政策信息自动化分类索引;

C2、政策文本标签提取;

D2、政策文本挖掘。

2. 根据权利要求1所述的基于标签相似度的企业政策匹配方法,其特征在于,步骤A1中包括:

经营范围标签,采用TextRank算法对企业经营范围文本内容做关键词提取得到经营范围标签;

行业标签,通过TF-IDF算法生成每个短文本的特征向量,使用多项朴素贝叶斯、逻辑回归、随机梯度下降分类算法训练模型,通过模型预测企业行业,得到行业标签;

司龄标签,当前日期减去企业的注册日期得到企业司龄标签;

地域标签,编辑地名自定义词典,在地名自定义词典基础上对地址做分词,将分词后的地址和地名列表匹配,匹配成功取出地名作为地域标签;

规模标签,根据各个行业的企业的注册资金,将企业分为大型、中型、小型、微型企业。

3. 根据权利要求1所述的基于标签相似度的企业政策匹配方法,其特征在于,步骤B1中包括:

企业创新力标签,如果企业具有软件著作权、专利、对外网站和商标信息,显示创新力标签;

企业竞争力标签,如果企业有成功投标的记录,显示竞争力标签;

发展潜力标签,如果企业有分支机构和对外投资,显示发展潜力标签;

发展动力标签,如果企业有资质证书和行政许可,显示发展动力标签。

4. 根据权利要求1所述的基于标签相似度的企业政策匹配方法,其特征在于,步骤C1中包括:

自身风险标签,如果企业法定代表人、企业股东和企业自身具有法律诉讼和法院公告,显示自身风险标签;

周边风险标签,若企业管理人员被列入失信人、被执行人时,显示周边风险标签;

预警提醒标签,出于增加新业务、业务减少或公司转行,企业产生地址变更、经营范围

变更或合伙人变更,显示预警提醒标签;

经营风险标签,根据企业是否被列入“经营异常”名录,是否存在行政处罚、是否存在股权出质信息,来判断企业的经营风险,显示经营风险标签。

5. 根据权利要求1所述的基于标签相似度的企业政策匹配方法,其特征在于,在步骤A2中,实时监控政策发布源头目标网站时,把最新的网页及时采集到本地,进行内容分析和信息过滤等流程,完成政策信息本地存储;数据采集过程中,不仅将网页的非结构化数据转变成半结构化数据,同时自动提取政策名称、发布时间、政策文本内容,以及发文单位名称、发布网站名称、频道名称和发文链接地址,以此构建政策元数据数据库。

6. 根据权利要求1所述的基于标签相似度的企业政策匹配方法,其特征在于,在步骤B2中,采用自动分类和规则分类技术,对政策做多维度分类标引,用以系统针对不同企业在不同需求场景下更加快速、有针对性地查找到所需类目和对应的政策信息;其中,分类标引包括政策所属行业领域、所属地域名称、发布单位名称、所属主题名称、发文形式和所属年份;

在前端应用功能中,利用这些政策索引,采用细分导航的方式,进行政策列表展示;

通过组合式的检索功能对政策进行搜索,让企业及政府服务部门可以通过自定义关键词的方式获取个性化的检索结果,达到快速、全面了解信息的目的;

同时,对政策标题、正文和主题提供全文检索功能,对政策的发布单位名称、发文形式、所属行业领域、所属地域、发布年份等字段,提供筛选功能。

7. 根据权利要求1所述的基于标签相似度的企业政策匹配方法,其特征在于,步骤C2包括:

a、将抓取到的政策文本信息进行预处理,包括分词和过滤掉停用词;每个事件文本摘要 T 被分割成 m 个句子 S ,即 $T = [S_1, S_2, \dots, S_m]$;每个句子 S_i 再被分词成词语 t ,即 $S_i = [t_{i1}, t_{i2}, \dots, t_{in}]$,其中 $t_{ij} \in S_i$ 是保留后的候选关键词;

b、构建候选关键词图 $G = (V, E)$,其中 V 为节点集,是由步骤a生成的候选关键词组成,使用一个大小为10的窗口依次滑过这些关键词,当任意两个节点在这个窗口中共同出现的时候,在这两点间连接一条边;

c、迭代传播各节点的权重,直至收敛;将得到的各节点权重值进行从高到低排序,取Top10词语作为政策文本标签。

8. 根据权利要求1所述的基于标签相似度的企业政策匹配方法,其特征在于,步骤D2包括:

采用文本自动分词和词性标注等自然语言处理技术,基于规则与统计相结合的方式,将政策文本进行中文分词以及政策信息提取,包括政策主题关键词、相关人物、机构和地区名称的结构化提取,完成政策的关键词和实体标引;

利用语义分析技术,把多政策之间的相关度超过一定阈值的文章关联到一起,实现复杂语义关系的深度挖掘,实现往年政策相互关联分析。

9. 根据权利要求1所述的基于标签相似度的企业政策匹配方法,其特征在于,步骤3中包括对上面得到的企业标签和政策标签数据,运用词向量计算文本相似度,运用Glove模型、word2vec模型、Bert模型训练生成词向量,计算标签词向量的相似度,设定指定阈值,对政策标签和企业标签进行融合;其中,

融合度 = 企业标签和政策标签匹配的数量 ÷ 政策标签总数量 × 100%;当融合度高于

80%，则认为企业和该政策相匹配。

基于标签相似度的企业政策匹配方法

技术领域

[0001] 本发明涉及一种基于标签相似度的企业政策匹配方法。

背景技术

[0002] 近年来,各级政府在融资、科技、产业、人才等方面出台各种地方性政策。对于企业而言,查询政策信息的难度在于来源广泛的问题。每年发布的政策中,只有很少一部分是在固定时间段内发布,另外的大部分都是根据社会发展需要而实时推出的,企业无法提前准备政策信息获取和申报工作。

[0003] 目前,许多企业对政策还一知半解,缺乏足够人力和精力去解读,并同时存在一些政策宣传不到位、未落地等现象,影响了政策的有效性。

[0004] 因此,在信息过载的时代,如何快速高效获取政策信息,如何借助数据挖掘和机器分析能力汇聚分析信息为政府部门和企业提供服务,成为政策供给侧和需求侧精准匹配需要解决的关键问题。

发明内容

[0005] 本发明的目的是提供一种基于标签相似度的企业政策匹配方法,该方法能够快速高效地获取政策信息并将政策供给侧和需求侧精准匹配,保证政策落实效果,提高政府政务服务质量。

[0006] 为了实现上述目的,本发明提供了一种基于标签相似度的企业政策匹配方法,包括:

[0007] 步骤1、自动化构建企业标签;

[0008] 步骤2、自动化构建政策标签;

[0009] 步骤3、基于标签相似度的政策匹配;其中,

[0010] 步骤1中包括:

[0011] A1、构造企业基本属性标签,包括经营范围标签、行业标签、司龄标签、地域标签和规模标签;

[0012] B1、构造企业经营状况标签,包括企业创新力标签、企业竞争力标签、企业发展潜力标签和企业发展动力标签;

[0013] C1、构造企业风险信息标签,包括企业自身风险标签、企业周边风险标签、预警提醒标签和经营风险标签;

[0014] 步骤2中包括:

[0015] A2、政策信息自动化采集;

[0016] B2、政策信息自动化分类索引;

[0017] C2、政策文本标签提取;

[0018] D2、政策文本挖掘。

[0019] 优选地,步骤A1中包括:

[0020] 经营范围标签,采用TextRank算法对企业经营范围文本内容做关键词提取得到经营范围标签;

[0021] 行业标签,通过TF-IDF算法生成每个短文本的特征向量,使用多项朴素贝叶斯、逻辑回归、随机梯度下降分类算法训练模型,通过模型预测企业行业,得到行业标签;

[0022] 司龄标签,当前日期减去企业的注册日期得到企业司龄标签;

[0023] 地域标签,编辑地名自定义词典,在地名自定义词典基础上对地址做分词,将分词后的地址和地名列表匹配,匹配成功取出地名作为地域标签;

[0024] 规模标签,根据各个行业的企业的注册资金,将企业分为大型、中型、小型、微型企业。

[0025] 优选地,步骤B1中包括:

[0026] 企业创新力标签,如果企业具有软件著作权、专利、对外网站和商标信息,显示创新力标签;

[0027] 企业竞争力标签,如果企业有成功投标的记录,显示竞争力标签;

[0028] 发展潜力标签,如果企业有分支机构和对外投资,显示发展潜力标签;

[0029] 发展动力标签,如果企业有资质证书和行政许可,显示发展动力标签。

[0030] 优选地,步骤C1中包括:

[0031] 自身风险标签,如果企业法定代表人、企业股东和企业自身具有法律诉讼和法院公告,显示自身风险标签;

[0032] 周边风险标签,若企业管理人员被列入失信人、被执行人时,显示周边风险标签;

[0033] 预警提醒标签,出于增加新业务、业务减少或公司转行,企业产生地址变更、经营范围变更或合伙人变更,显示预警提醒标签;

[0034] 经营风险标签,根据企业是否被列入“经营异常”名录,是否存在行政处罚、是否存在股权出质信息,来判断企业的经营风险,显示经营风险标签。

[0035] 优选地,在步骤A2中,实时监控政策发布源头目标网站时,把最新的网页及时采集到本地,进行内容分析和信息过滤等流程,完成政策信息本地存储;数据采集过程中,不仅将网页的非结构化数据转变成半结构化数据,同时自动提取政策名称、发布时间、政策文本内容,以及发文单位名称、发布网站名称、频道名称和发文链接地址,以此构建政策元数据库。

[0036] 优选地,在步骤B2中,采用自动分类和规则分类技术,对政策做多维度分类标引,用以系统针对不同企业在不同需求场景下更加快速、有针对性地查找到所需类目和对应的政策信息;其中,分类标引包括政策所属行业领域、所属地域名称、发布单位名称、所属主题名称、发文形式和所属年份;

[0037] 在前端应用功能中,利用这些政策索引,采用细分导航的方式,进行政策列表展示;

[0038] 通过组合式的检索功能对政策进行搜索,让企业及政府服务部门可以通过自定义关键词的方式获取个性化的检索结果,达到快速、全面了解信息的目的;

[0039] 同时,对政策标题、正文和主题提供全文检索功能,对政策的发布单位名称、发文形式、所属行业领域、所属地域、发布年份等字段,提供筛选功能。

[0040] 优选地,步骤C2包括:

[0041] a、将抓取到的政策文本信息进行预处理,包括分词和过滤掉停用词;每个事件文本摘要T被分割成m个句子S,即 $T=[S_1, S_2, \dots, S_m]$;每个句子 S_i 再被分词成词语t,即 $S_i=[t_{i1}, t_{i2}, \dots, t_{in}]$,其中 $t_{ij} \in S_i$ 是保留后的候选关键词;

[0042] b、构建候选关键词图 $G=(V, E)$,其中V为节点集,是由步骤a生成的候选关键词组成,使用一个大小为10的窗口依次滑过这些关键词,当任意两个节点在这个窗口中共同出现的时候,在这两点间连接一条边;

[0043] c、迭代传播各节点的权重,直至收敛;将得到的各节点权重值进行从高到低排序,取Top10词语作为政策文本标签。

[0044] 优选地,步骤D2包括:

[0045] 采用文本自动分词和词性标注等自然语言处理技术,基于规则与统计相结合的方式,将政策文本进行中文分词以及政策信息提取,包括政策主题关键词、相关人物、机构和地区名称的结构化提取,完成政策的关键词和实体标引;

[0046] 利用语义分析技术,把多政策之间的相关度超过一定阈值的文章关联到一起,实现复杂语义关系的深度挖掘,实现往年政策相互关联分析。

[0047] 优选地,步骤3中包括对上面得到的企业标签和政策标签数据,运用词向量计算文本相似度,运用Glove模型、word2vec模型、Bert模型训练生成词向量,计算标签词向量的相似度,设定指定阈值,对政策标签和企业标签进行融合;其中,

[0048] 融合度=企业标签和政策标签匹配的数量÷政策标签总数量×100%;当融合度高于80%,则认为企业和该政策相匹配。

[0049] 根据上述技术方案,本发明从涉及企业的政务数据和涉及政策的互联网文本数据出发,利用数据挖掘、自然语义处理等技术构造企业标签和政策标签并计算两者相似度,以此完成企业和政策的匹配,让政府在政务服务提供上更加有效地发掘企业需要,解决了供给侧与需求侧之间的精准匹配问题。

[0050] 本发明的其他特征和优点将在随后的具体实施方式部分予以详细说明。

附图说明

[0051] 附图是用来提供对本发明的进一步理解,并且构成说明书的一部分,与下面的具体实施方式一起用于解释本发明,但并不构成对本发明的限制。在附图中:

[0052] 图1是本发明中基于标签相似度的企业政策匹配方法的流程图。

具体实施方式

[0053] 以下结合附图对本发明的具体实施方式进行详细说明。应当理解的是,此处所描述的具体实施方式仅用于说明和解释本发明,并不用于限制本发明。

[0054] 在本发明中,在未作相反说明的情况下,包含在术语中的方位词仅代表该术语在常规使用状态下的方位,或为本领域技术人员理解的俗称,而不应视为对该术语的限制。

[0055] 参见图1,本发明提供一种基于标签相似度的企业政策匹配方法,包括:

[0056] 步骤1、自动化构建企业标签;

[0057] 步骤2、自动化构建政策标签;

[0058] 步骤3、基于标签相似度的政策匹配;其中,

[0059] 步骤1中包括：

[0060] A1、构造企业基本属性标签，包括经营范围标签、行业标签、司龄标签、地域标签和规模标签；

[0061] 标签具体提取方法如下：

[0062] 经营范围标签，采用TextRank算法对企业经营范围文本内容做关键词提取得到经营范围标签；

[0063] 行业标签，对于行业信息缺失的企业，根据企业的经营范围来预测其行业。通过TF-IDF算法生成每个短文本的特征向量，使用多项朴素贝叶斯、逻辑回归、随机梯度下降分类算法训练模型，通过模型预测企业行业，得到行业标签；

[0064] 司龄标签，当前日期减去企业的注册日期得到企业司龄标签；

[0065] 地域标签，编辑地名自定义词典，在地名自定义词典基础上对地址做分词，将分词后的地址和地名列表匹配，匹配成功取出地名作为地域标签；

[0066] 规模标签，参照国家统计局2017年发布的《统计上中小微企业划分办法》的通知，根据各个行业的企业的注册资金，将企业分为大型、中型、小型、微型企业。

[0067] B1、构造企业经营状况标签，包括企业创新力标签、企业竞争力标签、企业发展潜力标签和企业发展动力标签；具体操作为：

[0068] 企业创新力标签，如果企业具有软件著作权、专利、对外网站、商标信息，说明企业具有一定的创新力，相应地显示创新力标签。

[0069] 企业竞争力标签，如果企业有成功投标的记录，说明企业具有一定的竞争力，相应地显示竞争力标签。

[0070] 发展潜力标签，如果企业有分支机构和对外投资，相应地显示发展潜力标签。

[0071] 发展动力标签，如果企业有资质证书和行政许可，相应地显示发展动力标签。

[0072] C1、构造企业风险信息标签，包括企业自身风险标签、企业周边风险标签、预警提醒标签和经营风险标签；具体操作为：

[0073] 自身风险标签，如果企业法定代表人、企业股东和企业自身具有法律诉讼和法院公告，说明企业具有一定的自身风险，显示自身风险标签。

[0074] 周边风险标签，若企业管理人员被列入失信人、被执行人时，说明企业具有一定的周边风险，显示周边风险标签。

[0075] 预警提醒标签，出于增加新业务、业务减少、公司转行等原因，企业会产生地址变更、经营范围变更、合伙人变更等变更信息，因此，这些变更可作为预警提醒维度的参考。当企业发生变更信息时，显示预警提醒标签。

[0076] 经营风险标签，根据企业是否被列入“经营异常”名录，是否存在行政处罚、是否存在股权出质信息，来判断企业的经营风险，显示经营风险标签。

[0077] 步骤2中包括：

[0078] A2、政策信息自动化采集；

[0079] 实时监控政策发布源头目标网站时，把最新的网页及时采集到本地，进行内容分析和信息过滤等流程，完成政策信息本地存储。数据采集过程中，不仅将网页的非结构化数据转变成半结构化数据，同时自动提取政策名称、发布时间、政策文本内容，以及发文单位名称、发布网站名称、频道名称、发文链接地址等政策相关数据，以此构建政策元数据数据

库。

[0080] B2、政策信息自动化分类索引；

[0081] 采用自动分类和规则分类技术，对政策做多维度分类标引，用以系统针对不同企业在不同需求场景下更加快速、有针对性地查找到所需类目和对应的政策信息。分类标引包括政策所属行业领域、所属地域名称、发布单位名称、所属主题名称、发文形式、所属年份等分类索引。在前端应用功能中，利用这些政策索引，采用细分导航的方式，进行政策列表展示。通过组合式的检索功能对政策进行搜索，让企业及政府服务部门可以通过自定义关键词的方式获取个性化的检索结果，达到快速、全面了解信息的目的。同时，对政策标题、正文和主题提供全文检索功能。对政策的发布单位名称、发文形式、所属行业领域、所属地域、发布年份等字段，提供筛选功能。

[0082] C2、政策文本标签提取；

[0083] a、将抓取到的政策文本信息进行预处理，包括分词和过滤掉停用词；每个事件文本摘要T被分割成m个句子S，即 $T = [S_1, S_2, \dots, S_m]$ ；每个句子 S_i 再被分词成词语t，即 $S_i = [t_{i1}, t_{i2}, \dots, t_{in}]$ ，其中 $t_{ij} \in S_i$ 是保留后的候选关键词；

[0084] b、构建候选关键词图 $G = (V, E)$ ，其中V为节点集，是由步骤a生成的候选关键词组成，使用一个大小为10的窗口依次滑过这些关键词，当任意两个节点在这个窗口中共同出现的时候，在这两点间连接一条边；

[0085] c、迭代传播各节点的权重，直至收敛；将得到的各节点权重值进行从高到低排序，取Top10词语作为政策文本标签。

[0086] D2、政策文本挖掘；

[0087] 采用文本自动分词和词性标注等自然语言处理技术，基于规则与统计相结合的方式，将政策文本进行中文分词以及政策信息提取，包括政策主题关键词、相关人物、机构、地区名称等信息的结构化提取，完成政策的关键词和实体标引。利用语义分析技术，把多政策之间的相关度超过一定阈值的文章关联到一起，实现复杂语义关系的深度挖掘，实现往年政策相互关联分析。

[0088] 步骤3中包括：对上面得到的企业标签和政策标签数据，运用词向量计算文本相似度，运用Glove模型、word2vec模型、Bert模型训练生成词向量，计算标签词向量的相似度，设定指定阈值，对政策标签和企业标签进行融合；其中，

[0089] 融合度 = 企业标签和政策标签匹配的数量 ÷ 政策标签总数量 × 100%；当融合度高于80%，则认为企业和该政策相匹配。

[0090] 本申请针对“政府在建设传统电子政务平台时通常从其政务处理需求出发，将所有自己可以提供的服务不加区分全部堆叠在门户网站上，让企业难以寻找到自己需要的服务”这一问题，从涉及企业的政务数据和涉及政策的互联网文本数据出发，利用数据挖掘、自然语义处理等技术构造企业标签和政策标签并计算两者相似度，以此完成企业和政策的匹配，让政府在政务服务提供上更加有效地发掘企业需要，解决了供给侧与需求侧之间的精准匹配问题。保证政策落实效果，提高政府政务服务质量。

[0091] 以上结合附图详细描述了本发明的优选实施方式，但是，本发明并不限于上述实施方式中的具体细节，在本发明的技术构思范围内，可以对本发明的技术方案进行多种简单变型，这些简单变型均属于本发明的保护范围。

[0092] 另外需要说明的是,在上述具体实施方式中所描述的各个具体技术特征,在不矛盾的情况下,可以通过任何合适的方式进行组合,为了避免不必要的重复,本发明对各种可能的组合方式不再另行说明。

[0093] 此外,本发明的各种不同的实施方式之间也可以进行任意组合,只要其不违背本发明的思想,其同样应当视为本发明所公开的内容。



图1