(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 116932487 B
(45)授权公告日 2023.11.28

(21)申请号 202311189545.7

(22)申请日 2023.09.15

(65)同一申请的已公布的文献号
申请公布号 CN 116932487 A

(43)申请公布日 2023.10.24

(73)专利权人 北京安联通科技有限公司
地址 100000 北京市海淀区蓝靛厂东路2号院2号楼(金源时代商务中心2号楼)15层1单元(A座)18B

(72)发明人 杨桢

(74)专利代理机构 北京国源中科知识产权代理事务所(普通合伙) 16179
专利代理师 胡勋勋

(51)Int.Cl.
G06F 16/16(2019.01)
G06F 16/17(2019.01)
G06F 40/279(2020.01)
G06F 21/62(2013.01)
G06N 3/0442(2023.01)
G06N 3/08(2023.01)

(56)对比文件
CN 112464281 A,2021.03.09
CN 113962364 A,2022.01.21
CN 115392252 A,2022.11.25
CN 115718792 A,2023.02.28
CN 115952291 A,2023.04.11
CN 116484848 A,2023.07.25
WO 2021218322 A1,2021.11.04

审查员 王宇莉

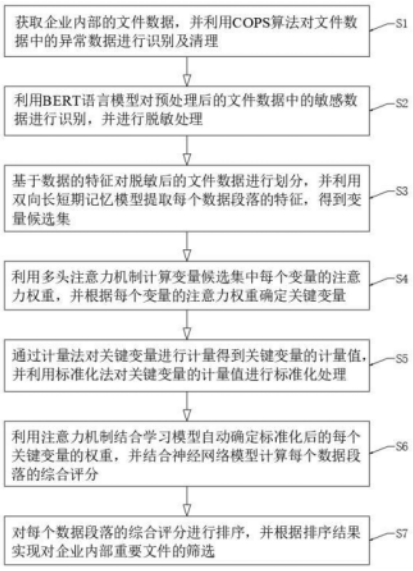
权利要求书3页 说明书10页 附图1页

(54)发明名称

一种基于数据段落划分的量化式数据分析方法及系统

(57)摘要

本发明涉及数据处理技术领域,公开了一种基于数据段落划分的量化式数据分析方法及系统,该方法包括以下步骤:获取企业内部的文件数据,并进行异常数据的识别及清理;对预处理后的文件数据中的敏感数据进行脱敏处理;提取每个数据段落的特征,得到变量候选集;计算变量候选集中每个变量的注意力权重,并确定关键变量;计量得到关键变量的计量值并进行标准化处理;确定标准化后的每个关键变量的权重,并计算每个数据段落的综合评分;根据综合评分的排序结果实现对企业内部重要文件的筛选。本发明不仅可以更高效地管理文件数据,节省存储空间,降低存储成本,而且还可以有效地提升决策效率和精准性,更好地满足于企业的使用需求。



CN 116932487 B

1.一种基于数据段落划分的量化式数据分析方法,其特征在于,该方法包括以下步骤:

S1、获取企业内部的文件数据,并利用COPS算法对文件数据中的异常数据进行识别及清理;

S2、利用BERT语言模型对预处理后的文件数据中的敏感数据进行识别,并进行脱敏处理;

S3、基于数据的特征对脱敏后的文件数据进行划分,并利用双向长短期记忆模型提取每个数据段落的特征,得到变量候选集;

S4、利用多头注意力机制计算变量候选集中每个变量的注意力权重,并根据每个变量的注意力权重确定关键变量;

S5、通过计量法对关键变量进行计量得到关键变量的计量值,并利用标准化法对关键变量的计量值进行标准化处理;

S6、利用注意力机制结合学习模型自动确定标准化后的每个关键变量的权重,并结合神经网络模型计算每个数据段落的综合评分;

S7、对每个数据段落的综合评分进行排序,并根据排序结果实现对企业内部重要文件的筛选;

所述获取企业内部的文件数据,并利用COPS算法对文件数据中的异常数据进行识别及清理包括以下步骤:

S11、获取企业内部的文件数据,得到初始文件数据集;

S12、采用COPS算法对初始文件数据集中的数据进行清洗,识别和移除不良数据和异常值,获得文件数据集;

S13、对清洗后的文件数据集进行分词、去噪及缺失值填充处理;

所述采用COPS算法对初始文件数据集中的数据进行清洗,识别和移除不良数据和异常值,获得文件数据集包括以下步骤:

S121、选取与文件内容和主题相关的特征,并从初始文件数据集的每个文件中提取选定的特征,得到文件的特征向量;

S122、对每个文件的特征向量进行归一化处理,并初始化聚类次数k=n和阈值向量$T=T_0$,且$T_0=0$;

S123、基于增量△增大阈值向量T得到不同的聚类划分$C^k=\{C_1,C_2,\cdots,C_k\}$,并计算相应的聚类有效性指标Q;

S124、重复执行S123,直至k=1,得到一系列的有效性指标Q,并选取有效性指标Q最小的聚类划分作为最佳聚类结果;

S125、计算每个聚类的聚类中心o_i及其模值$|o_i|$,则具有最小模值$|o_i|$的聚类为正常数据件聚类,其余聚类中的数据为异常数据;

S126、移除识别出的异常数据,得到文件数据集;

所述增量△的计算公式为:

$$\Delta=\varepsilon\times\frac{\max\{\sigma_1,\sigma_2,\ldots,\sigma_s\}}{\sigma_m},(m=1,2,\ldots,s);$$

所述聚类有效性指标Q的计算公式为:

$$Q=\frac{1}{M}\left(\alpha\cdot Scat\left(C^{k}\right)+\beta\cdot Sep\left(C^{k}\right)\right);$$

式中,ε表示COPS算法精度的参数;

σ_m表示归一化数据的标准偏差,m表示维度;

M表示初始状态时的类间分离度;

α和β表示组合参数;

$Scat\left(C^{k}\right)$表示类内紧凑度;

$Sep\left(C^{k}\right)$表示类间分离度。

2.根据权利要求1所述的一种基于数据段落划分的量化式数据分析方法,其特征在于,所述利用BERT语言模型对预处理后的文件数据中的敏感数据进行识别,并进行脱敏处理包括以下步骤:

S21、采集包含敏感信息的文件并进行敏感信息的标签标注,并利用标注的数据对BERT模型进行训练;

S22、利用训练后的BERT模型对预处理后的文件数据进行预测,得到每个词语被标注为敏感信息的概率;

S23、根据得到的概率确定文件数据中涉及敏感信息的位置和内容,并对确定的敏感信息进行脱敏处理。

3.根据权利要求1所述的一种基于数据段落划分的量化式数据分析方法,其特征在于,所述基于数据的特征对脱敏后的文件数据进行划分,并利用双向长短期记忆模型提取每个数据段落的特征,得到变量候选集包括以下步骤:

S31、利用脱敏后的文件数据中的关键词和主题词构建文件的特征向量;

S32、通过聚类算法基于文件的特征向量进行聚类,并将每个聚类作为一个数据段落;

S33、利用训练好的双向长短期记忆模型提取每个数据段落中语句的特征向量,得到变量候选集。

4.根据权利要求1所述的一种基于数据段落划分的量化式数据分析方法,其特征在于,所述利用多头注意力机制计算变量候选集中每个变量的注意力权重,并根据每个变量的注意力权重确定关键变量包括以下步骤:

S41、对变量候选集中的每个变量进行标准化处理,并将标准化后的变量候选集输入多头注意力机制中;

S42、利用每个注意力头对变量候选集中变量之间的关系进行建模,产生对应的注意力权重;

S43、将各个注意力头产生的权重进行加权平均得到最终的注意力权重;

S44、按照最终注意力权重值由高至低的顺序对每个变量进行排序,并选取前L个变量作为关键变量。

5.根据权利要求1所述的一种基于数据段落划分的量化式数据分析方法,其特征在于,所述利用注意力机制结合学习模型自动确定标准化后的每个关键变量的权重,并结合神经网络模型计算每个数据段落的综合评分包括以下步骤:

S61、利用注意力机制结合学习模型自动确定标准化后的每个关键变量的权重；

S62、利用训练后的神经网络模型输出每个关键变量的评分，并利用加权求和法结合关键变量的权重和评分计算每个数据段落的综合评分。

6.根据权利要求1所述的一种基于数据段落划分的量化式数据分析方法，其特征在于，所述对每个数据段落的综合评分进行排序，并根据排序结果实现对企业内部重要文件的筛选包括以下步骤：

S71、获取每个数据段落的综合评分，并按照评分由高至低的顺序进行排序，得到排序结果；

S72、根据评分排序结果结合预设的评分阈值，选取评分高于阈值的前N个数据段落作为重要数据段落；

S73、统计每个文件数据中重要数据段落的比例，当该比例高于预设的比例阈值时，则确定该文件数据为重要文件。

7.一种基于数据段落划分的量化式数据分析系统，用于实现权利要求1-6中任一项所述的基于数据段落划分的量化式数据分析方法的步骤，其特征在于，该系统包括数据清洗模块、数据脱敏模块、变量候选集确定模块、关键变量确定模块、变量计量值计算模块、综合评分模块及数据筛选模块；

其中，所述数据清洗模块用于获取企业内部的文件数据，并利用COPS算法对文件数据中的异常数据进行识别及清理；

所述数据脱敏模块用于利用BERT语言模型对预处理后的文件数据中的敏感数据进行识别，并进行脱敏处理；

所述变量候选集确定模块用于基于数据的特征对脱敏后的文件数据进行划分，并利用双向长短期记忆模型提取每个数据段落的特征，得到变量候选集；

所述关键变量确定模块用于利用多头注意力机制计算变量候选集中每个变量的注意力权重，并根据每个变量的注意力权重确定关键变量；

所述变量计量值计算模块用于通过计量法对关键变量进行计量得到关键变量的计量值，并利用标准化法对关键变量的计量值进行标准化处理；

所述综合评分模块用于利用注意力机制结合学习模型自动确定标准化后的每个关键变量的权重，并结合双向长短期记忆模型计算每个数据段落的综合评分；

所述数据筛选模块用于对每个数据段落的综合评分进行排序，并根据排序结果实现对企业内部重要文件的筛选。

一种基于数据段落划分的量化式数据分析方法及系统

技术领域

[0001] 本发明涉及数据处理技术领域,具体来说,涉及一种基于数据段落划分的量化式数据分析方法及系统。

背景技术

[0002] 随着企业发展规模的不断扩大,企业内部的文件数据也在持续增加。同时,伴随着无纸化技术的逐步成熟和广泛应用,大多数现代企业已经转向电子文档方式来存储和处理企业内部的文件数据。然而,传统的文件存储方法主要是依据文件的类型进行分类存储,并在每个类别的文件夹中按照名称、大小、项目类型、修改日期等因素对文件进行排列和存储。这种方法,固然在一定程度上方便了文件的检索,但却无法体现每个类别中不同文件的重要性。

[0003] 由于缺乏对文件重要性的有效评判机制,企业管理者在查找和分析数据文件时,往往无法准确快速地找到各类别中的重要文件。在实际操作中,他们可能需要反复打开和阅览多个文件,通过逐一分析才能判断文件的重要程度。这种方式不仅效率低下,而且可能因为人为因素导致重要文件被遗漏。这种情况严重浪费了资料查询人员的时间,也可能影响到企业决策的效率和准确性。

[0004] 基于这样的背景,企业对于一种能够进行快速、准确的文件重要性判断的技术需求日益强烈。需要一种能够将大量的文件数据进行有效处理和量化分析的技术,以此提高文件检索的准确性和效率,实现对企业内部重要文件的筛选,为企业决策提供更加精准、高效的支持。

[0005] 因此,本发明提出来一种基于数据段落划分的量化式数据分析方法及系统。

发明内容

[0006] 针对相关技术中的问题,本发明提出一种基于数据段落划分的量化式数据分析方法及系统,以克服现有相关技术所存在的上述技术问题。

[0007] 为此,本发明采用的具体技术方案如下:

[0008] 根据本发明的一个方面,提供了一种基于数据段落划分的量化式数据分析方法,该方法包括以下步骤:

[0009] S1、获取企业内部的文件数据,并利用COPS算法对文件数据中的异常数据进行识别及清理;

[0010] S2、利用BERT语言模型对预处理后的文件数据中的敏感数据进行识别,并进行脱敏处理;

[0011] S3、基于数据的特征对脱敏后的文件数据进行划分,并利用双向长短期记忆模型提取每个数据段落的特征,得到变量候选集;

[0012] S4、利用多头注意力机制计算变量候选集中每个变量的注意力权重,并根据每个变量的注意力权重确定关键变量;

[0013] S5、通过计量法对关键变量进行计量得到关键变量的计量值,并利用标准化法对关键变量的计量值进行标准化处理;

[0014] S6、利用注意力机制结合学习模型自动确定标准化后的每个关键变量的权重,并结合神经网络模型计算每个数据段落的综合评分;

[0015] S7、对每个数据段落的综合评分进行排序,并根据排序结果实现对企业内部重要文件的筛选。

[0016] 作为优选地,所述获取企业内部的文件数据,并利用COPS算法对文件数据中的异常数据进行识别及清理包括以下步骤:

[0017] S11、获取企业内部的文件数据,得到初始文件数据集;

[0018] S12、采用COPS算法对初始文件数据集中的数据进行清洗,识别和移除不良数据和异常值,获得文件数据集;

[0019] S13、对清洗后的文件数据集进行分词、去噪及缺失值填充处理。

[0020] 作为优选地,所述采用COPS算法对初始文件数据集中的数据进行清洗,识别和移除不良数据和异常值,获得文件数据集包括以下步骤:

[0021] S121、选取与文件内容和主题相关的特征,并从初始文件数据集的每个文件中提取选定的特征,得到文件的特征向量;

[0022] S122、对每个文件的特征向量进行归一化处理,并初始化聚类次数k=n和阈值向量$T=T_0$,且$T_0=0$;

[0023] S123、基于增量△增大阈值向量T得到不同的聚类划分$C^k=\{C_1,C_2,\cdots,C_k\}$,并计算相应的聚类有效性指标Q;

[0024] S124、重复执行S123,令k=k-1,直至k=1,得到一系列的有效性指标Q,并选取有效性指标Q最小的聚类划分作为最佳聚类结果;

[0025] S125、计算每个聚类的聚类中心o_i及其模值$|o_i|$,则具有最小模值$|o_i|$的聚类为正常数据件聚类,其余聚类中的数据为异常数据;

[0026] S126、移除识别出的异常数据,得到文件数据集。

[0027] 作为优选地,所述增量△的计算公式为:

[0028] $$\Delta=\varepsilon\times\frac{\max\{\sigma_1,\sigma_2,...,\sigma_s\}}{\sigma_m},(m=1,2,...,s)$$

[0029] 所述聚类有效性指标Q的计算公式为:

[0030] $$Q=\frac{1}{M}\left(\alpha\cdot Scat\left(C^k\right)+\beta\cdot Sep\left(C^k\right)\right)$$

[0031] 式中,ε表示COPS算法精度的参数;σ_m表示归一化数据的标准偏差,m表示维度;M表示初始状态时的类间分离度;α和β表示组合参数;$Scat\left(C^k\right)$表示类内紧凑度;$Sep\left(C^k\right)$表示类间分离度。

[0032] 作为优选地,所述利用BERT语言模型对预处理后的文件数据中的敏感数据进行识别,并进行脱敏处理包括以下步骤:

[0033] S21、采集包含敏感信息的文件并进行敏感信息的标签标注,并利用标注的数据对BERT模型进行训练;

[0034] S22、利用训练后的BERT模型对预处理后的文件数据进行预测,得到每个词语被标注为敏感信息的概率;

[0035] S23、根据得到的概率确定文件数据中涉及敏感信息的位置和内容,并对确定的敏感信息进行脱敏处理。

[0036] 作为优选地,所述基于数据的特征对脱敏后的文件数据进行划分,并利用双向长短期记忆模型提取每个数据段落的特征,得到变量候选集包括以下步骤:

[0037] S31、利用脱敏后的文件数据中的关键词和主题词构建文件的特征向量;

[0038] S32、通过聚类算法基于文件的特征向量进行聚类,并将每个聚类作为一个数据段落;

[0039] S33、利用训练好的双向长短期记忆模型提取每个数据段落中语句的特征向量,得到变量候选集。

[0040] 作为优选地,所述利用多头注意力机制计算变量候选集中每个变量的注意力权重,并根据每个变量的注意力权重确定关键变量包括以下步骤:

[0041] S41、对变量候选集中的每个变量进行标准化处理,并将标准化后的变量候选集输入多头注意力机制中;

[0042] S42、利用每个注意力头对变量候选集中变量之间的关系进行建模,产生对应的注意力权重;

[0043] S43、将各个注意力头产生的权重进行加权平均得到最终的注意力权重;

[0044] S44、按照最终注意力权重值由高至低的顺序对每个变量进行排序,并选取前L个变量作为关键变量。

[0045] 作为优选地,所述利用注意力机制结合学习模型自动确定标准化后的每个关键变量的权重,并结合神经网络模型计算每个数据段落的综合评分包括以下步骤:

[0046] S61、利用注意力机制结合学习模型自动确定标准化后的每个关键变量的权重;

[0047] S62、利用训练后的神经网络模型输出每个关键变量的评分,并利用加权求和法结合关键变量的权重和评分计算每个数据段落的综合评分。

[0048] 作为优选地,所述对每个数据段落的综合评分进行排序,并根据排序结果实现对企业内部重要文件的筛选包括以下步骤:

[0049] S71、获取每个数据段落的综合评分,并按照评分由高至低的顺序进行排序,得到排序结果;

[0050] S72、根据评分排序结果结合预设的评分阈值,选取评分高于阈值的前N个数据段落作为重要数据段落;

[0051] S73、统计每个文件数据中重要数据段落的比例,当该比例高于预设的比例阈值时,则确定该文件数据为重要文件。

[0052] 根据本发明的另一个方面,提供了一种基于数据段落划分的量化式数据分析系统,该系统包括数据清洗模块、数据脱敏模块、变量候选集确定模块、关键变量确定模块、变量计量值计算模块、综合评分模块及数据筛选模块;

[0053] 其中,所述数据清洗模块用于获取企业内部的文件数据,并利用COPS算法对文件数据中的异常数据进行识别及清理;

[0054] 所述数据脱敏模块用于利用BERT语言模型对预处理后的文件数据中的敏感数据

进行识别,并进行脱敏处理;

[0055] 所述变量候选集确定模块用于基于数据的特征对脱敏后的文件数据进行划分,并利用双向长短期记忆模型提取每个数据段落的特征,得到变量候选集;

[0056] 所述关键变量确定模块用于利用多头注意力机制计算变量候选集中每个变量的注意力权重,并根据每个变量的注意力权重确定关键变量;

[0057] 所述变量计量值计算模块用于通过计量法对关键变量进行计量得到关键变量的计量值,并利用标准化法对关键变量的计量值进行标准化处理;

[0058] 所述综合评分模块用于利用注意力机制结合学习模型自动确定标准化后的每个关键变量的权重,并结合双向长短期记忆模型计算每个数据段落的综合评分;

[0059] 所述数据筛选模块用于对每个数据段落的综合评分进行排序,并根据排序结果实现对企业内部重要文件的筛选。

[0060] 与现有技术相比,本发明提供了基于数据段落划分的量化式数据分析方法及系统,具备以下有益效果:

[0061] 本发明不仅可以利用COPS算法实现对企业内部文件数据中异常数据的识别和清除,而且还可以基于BERT语言模型识别和脱敏处理敏感信息,有效地保护数据隐私,满足数据安全的要求,从而使得企业可以更高效地管理文件数据,节省存储空间,降低存储成本,同时,本发明还可以利用双向长短期记忆模型、多头注意力机制及神经网络模型来实现对企业内部重要文件的筛选,从而可以根据文件的重要程度对企业内部的文件数据进行存储,使得企业管理者在决策时可以快速准确地找到重要的文件数据,即快速准确的找到决策依据,从而可以有效地提升决策效率和精准性,进而可以更好地满足于企业的使用需求。

附图说明

[0062] 此处所说明的附图用来提供对本申请的进一步理解,构成本申请的一部分,本申请的示意性实施例及其说明用于解释本申请,并不构成对本申请的不当限定。在附图中:

[0063] 图1是根据本发明实施例的一种基于数据段落划分的量化式数据分析方法的流程图。

具体实施方式

[0064] 下面将结合本发明实施例中的附图,对本发明实施例中的技术方案进行清楚、完整地描述,显然,所描述的实施例仅仅是本发明一部分实施例,而不是全部的实施例。基于本发明中的实施例,本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例,都属于本发明保护的范围。

[0065] 根据本发明的实施例,提供了一种基于数据段落划分的量化式数据分析方法及系统。

[0066] 现结合附图和具体实施方式对本发明进一步说明,如图1所示,根据本发明的一个实施例,提供了一种基于数据段落划分的量化式数据分析方法,该方法包括以下步骤:

[0067] S1、获取企业内部的文件数据,并利用COPS算法对文件数据中的异常数据进行识别及清理;

[0068] 其中,COPS算法是一种改进的DBSCAN算法,它通过迭代的方式确定最佳的聚类结

果和聚类数。这种方法可以有效识别出离群点和异常数据,清除这些不良数据后,可以显著提高数据集的质量。而高质量的数据是进行准确的数据分析的基础。所以采用COPS算法清洗数据可以为后续的变量选择、数据段落划分、特征提取等步骤提供更加可靠的数据基础。这有助于产生更加准确和高质量的分析结果。

[0069] COPS算法是一个很好的数据清洗方法,它可以有效提高数据的质量,为数据分析方法提供更加准确和可靠的数据基础。采用COPS算法清洗数据可以显著提高基于数据段落划分的量化式数据分析方法的效果和精度。

[0070] 具体的,所述获取企业内部的文件数据,并利用COPS算法对文件数据中的异常数据进行识别及清理包括以下步骤:

[0071] S11、获取企业内部的文件数据,得到初始文件数据集;

[0072] S12、采用COPS算法对初始文件数据集中的数据进行清洗,识别和移除不良数据和异常值,获得文件数据集;

[0073] 所述采用COPS算法对初始文件数据集中的数据进行清洗,识别和移除不良数据和异常值,获得文件数据集包括以下步骤:

[0074] S121、选取与文件内容和主题相关的特征,这些特征应能够区分正常数据和异常数据。这些特征可以是文件中的关键词、主题词、命名实体等,并从初始文件数据集的每个文件中提取选定的特征,得到文件的特征向量,对每个文件的特征向量进行归一化处理以便比较;

[0075] S122、对每个文件的特征向量进行归一化处理并初始化聚类次数k=n(文件总数)和阈值向量$T=T_0$(T_0为初始值);

[0076] S123、基于增量△增大阈值向量T(即之后每步给T一个增量△)得到不同的聚类划分$C^k=\{C_1,C_2,\cdots,C_k\}$,并计算相应的聚类有效性指标Q;

[0077] S124、重复执行S123,直至k=1,得到一系列的有效性指标Q,并选取有效性指标Q最小的聚类划分作为最佳聚类结果;

[0078] 步骤S123和S124实现了COPS算法的主要功能。其算法通过迭代的方式确定最佳的聚类数和聚类结果,这可以有效识别出离群点。但算法结果的准确性还是依赖于特征选择和参数调整。

[0079] S125、计算每个聚类的聚类中心o_i及其模值$|o_i|$,则具有最小模值$|o_i|$的聚类为正常数据件聚类,其余聚类中的数据为异常数据;

[0080] S126、移除识别出的异常数据,得到文件数据集。在步骤S126中还可以人工对识别结果进行校验,确认结果的准确性。如果有错误识别,则需要返回步骤S122和S123,重新调整算法参数。只有结合人工判断,才能达到较高的识别准确率。

[0081] 所述增量△的计算公式为:

[0082] $$\Delta=\varepsilon\times\frac{\max\{\sigma_1,\sigma_2,...,\sigma_s\}}{\sigma_m},(m=1,2,...,s)$$

[0083] 所述聚类有效性指标Q的计算公式为:

[0084] $$Q=\frac{1}{M}\left(\alpha\cdot Scat\left(C^k\right)+\beta\cdot Sep\left(C^k\right)\right)$$

[0085] $$Scat\left(C^k\right)=\sum_{i=1}^{k}\sum_{x,y\in C_i}\left\|x-y\right\|^2$$

[0086] $$Sep\left(C^k\right)=\sum_{i=1}^{k}\left(\sum_{j=1,j\neq i}^{k}\frac{1}{\left|C_i\right|\cdot\left|C_j\right|}\sum_{x\in C_i,y\in C_j}\left\|x-y\right\|^2\right)$$

[0087] 式中，ε表示COPS算法精度的参数，ε越大，COPS算法的搜索步数就会越少，反之ε越小，算法搜索步数就会越多，结果就更有可能趋于最优结果，但相应时间耗费也会越多；

[0088] σ_m表示归一化数据的标准偏差，m表示维度；

[0089] M表示初始状态时的类间分离度；

[0090] α和β表示组合参数，α=0.4，β=1.6，用于平衡$Scat\left(C^k\right)$和$Sep\left(C^k\right)$；

[0091] $Scat\left(C^k\right)$表示类内紧凑度；

[0092] $Sep\left(C^k\right)$表示类间分离度；

[0093] $\left\|x-y\right\|$表示数据x和数据y之间的欧式距离，$Scat\left(C^k\right)$值越小，表明类间越紧凑，$Sep\left(C^k\right)$值越大，表明类间分离性越强；

[0094] $\left|C_i\right|$表示聚类C_i中包含的数据点个数。

[0095] S13、对清洗后的文件数据集进行分词、去噪及缺失值填充处理，填充缺失值的方法通常包括平均数填充、中位数填充、众数填充、使用模型预测填充等。

[0096] 具体的，在S13完成后还需要再次进行数据质量检查，确保清洗过程没有引入新的问题。

[0097] S2、利用BERT语言模型对预处理后的文件数据中的敏感数据进行识别，并进行脱敏处理；

[0098] 其中，所述利用BERT语言模型对预处理后的文件数据中的敏感数据进行识别，并进行脱敏处理包括以下步骤：

[0099] S21、采集包含敏感信息的文件并进行敏感信息的标签标注，并利用标注的数据对BERT模型进行训练；

[0100] 具体的，选取包含敏感信息的文件，通过人工标注的方式获得敏感信息对应的标签。这些文件和标签会作为模型的训练数据。采用BERT模型的架构，使用标注的数据进行模型的预训练。这一步骤的目的是让BERT模型学会识别不同类型的敏感信息。

[0101] S22、利用训练后的BERT模型对预处理后的文件数据进行预测，得到每个词语被标注为敏感信息的概率；

[0102] 具体的，对预处理后的文件数据进行分批，每批数据输入到BERT模型中进行预测。BERT模型会对每个输入的文本进行预测，给出每个词被标注为敏感信息的概率。根据这些概率可以确定文本中涉及的敏感信息的位置和内容。

[0103] S23、根据得到的概率确定文件数据中涉及敏感信息的位置和内容，并对确定的敏感信息进行脱敏处理。

[0104] 具体的,根据BERT模型的预测结果对文件中的敏感信息进行脱敏处理,如用“***”替换敏感词或删除敏感词等。经过脱敏处理后,文件中的敏感信息已经被隐藏,这有助于保护数据的隐私和安全。此外,本实施例中还可以人工检查BERT模型的预测结果和脱敏效果,确认敏感信息已经被正确识别和处理。如果有不满意的地方需要对模型进行再训练和优化。

[0105] S3、基于数据的特征对脱敏后的文件数据进行划分,并利用双向长短期记忆模型(Bi-LSTM模型)提取每个数据段落的特征,得到变量候选集;

[0106] 其中,所述基于数据的特征对脱敏后的文件数据进行划分,并利用双向长短期记忆模型提取每个数据段落的特征,得到变量候选集包括以下步骤:

[0107] S31、利用脱敏后的文件数据中的关键词和主题词构建文件的特征向量;

[0108] S32、通过聚类算法基于文件的特征向量进行聚类,并将每个聚类作为一个数据段落,文件中的句子被划分到不同的聚类中;

[0109] 上述步骤的目的是基于文件内在的语义特征自动发现数据段落的划分结构。

[0110] S33、利用训练好的双向长短期记忆模型提取每个数据段落中语句的特征向量,得到变量候选集,具体包括:

[0111] Bi-LSTM模型训练:采用Bi-LSTM的网络结构,使用文件中的句子作为模型的训练数据。Bi-LSTM模型会学习文件中的时序特征和长期依赖关系,为每个句子产生一个特征向量。这一步骤的目的是获得每个句子的语义特征表示,为后续的变量选择提供信息。

[0112] 特征提取:将文件中的每个句子输入到Bi-LSTM模型中,获得其特征向量。这些特征向量构成了变量候选集,代表了文件中每个句子的语义信息。

[0113] 该方法利用无监督学习的聚类分析法发现数据的内在结构,实现自动的数据段落划分。同时,采用Bi-LSTM模型提取每个句子的语义特征,为变量选择提供信息。

[0114] S4、利用多头注意力机制计算变量候选集中每个变量的注意力权重,并根据每个变量的注意力权重确定关键变量;

[0115] 其中,所述利用多头注意力机制计算变量候选集中每个变量的注意力权重,并根据每个变量的注意力权重确定关键变量包括以下步骤:

[0116] S40、定义注意力机制,采用多头注意力机制,它包含多个注意力头,每个头都是一个注意力机制,每个注意力头会学习变量候选集中变量之间的不同关系,产生注意力权重,这些注意力权重通过加权平均得到最终的注意力权重,这种机制可以从多个角度理解变量之间的关系,提高注意力权重的准确性;

[0117] S41、对变量候选集中的每个变量进行标准化处理,并将标准化后的变量候选集输入多头注意力机制中;

[0118] S42、利用每个注意力头对变量候选集中变量之间的关系进行建模,产生一组对应的注意力权重;

[0119] S43、将各个注意力头产生的权重进行加权平均得到最终的注意力权重,注意力权重越高,表示对应的变量对目标的影响越大;

[0120] S44、按照最终注意力权重值由高至低的顺序对每个变量进行排序,并选取前L个变量作为关键变量。

[0121] 具体的,根据注意力权重,选择前L个权重最大的变量,作为关键变量。L是一个预

设的阈值,可以根据实际情况进行设置。这些关键变量代表数据中最重要和相关的信息,将用于后续的计量、标准化和评分计算。

[0122] S5、通过计量法对关键变量进行计量得到关键变量的计量值,并利用标准化法对关键变量的计量值进行标准化处理;

[0123] 具体的,通过计量法和标准化法对关键变量进行处理的步骤如下:

[0124] 确定变量类型:判断关键变量的类型,是定性变量还是定量变量。定性变量需要进行编码,定量变量可以直接使用其原值。

[0125] 定性变量编码:对于定性变量,需要进行编码以获得其计量值。常用的编码方法有:

[0126] 计数法:对于二分类变量,可以设置1表示有,0表示无。

[0127] 哑变量法:对于多分类变量,设置多个虚拟变量,每个变量代表一种类别,有该属性则为1,否则为0。

[0128] 独热编码:也是为多分类变量设置多个虚拟变量,但每个变量只有一个类别为1,其余为0。

[0129] 定量变量计量:对于定量变量,可以直接使用其原始值作为计量值。也可以根据需要进行一定的转换,如按区间分组等。

[0130] 变量标准化:使用标准化方法,将不同变量的计量值转换到同一量纲下,方便比较和加权求和。常用的标准化方法有:

[0131] 最小-最大标准化:将变量值转换到[0,1]区间。

[0132] Z-score标准化:将变量值转换到均值为0,标准差为1的分布下。

[0133] 小数定标标准化:保留变量原来的量纲,但将绝对值调整到小于1。

[0134] 上述步骤可以实现对不同类型变量的计量和标准化,使其可以进行比较和加权求和。

[0135] S6、利用注意力机制结合学习模型自动确定标准化后的每个关键变量的权重,并结合神经网络模型计算每个数据段落的综合评分;

[0136] 其中,所述利用注意力机制结合学习模型自动确定标准化后的每个关键变量的权重,并结合神经网络模型计算每个数据段落的综合评分包括以下步骤:

[0137] S60、选择注意力机制,采用门控注意力机制或多头注意力机制等。注意力机制可以自动学习每个关键变量的权重,代表其对评分的影响程度;选择学习算法,选择一个学习算法,如神经网络、随机森林、GBDT等。这个学习算法将用于建立评分模型,产生每个数据段落的评分;

[0138] S61、利用注意力机制结合学习模型自动确定标准化后的每个关键变量的权重;具体包括:

[0139] 构建模型:在学习算法的基础上添加注意力层。注意力层可以自动学习每个关键变量的权重。

[0140] 模型训练:将标准化后的关键变量作为特征输入学习算法模型。在训练过程中,注意力层会同时学习每个特征的权重,代表其对研究目标的影响程度。学习算法模型和注意力层会相互提高,最终产生学习模型和注意力权重。

[0141] S62、利用训练后的神经网络模型输出每个关键变量的评分,并利用加权求和法结

合关键变量的权重和评分计算每个数据段落的综合评分,具有包括:

[0142] 选择神经网络模型:根据数据的特征选择一个神经网络模型,如多层感知机、CNN或RNN等。该模型将用于计算每个关键变量的评分。

[0143] 模型训练:将关键变量的特征作为输入,神经网络模型的输出作为关键变量的评分。使用标注数据训练神经网络模型,最小化评分的预测误差。训练过程中,模型会学习特征与评分之间的映射关系。

[0144] 评分预测:将新的数据段落的关键变量特征输入神经网络模型,获得每个变量的评分。

[0145] 加权求和:将每个关键变量的评分与其权重相乘,得到加权评分。然后对所有关键变量的加权评分进行求和,得到数据段落的综合评分。

[0146] S7、对每个数据段落的综合评分进行排序,并根据排序结果实现对企业内部重要文件的筛选。

[0147] 其中,所述对每个数据段落的综合评分进行排序,并根据排序结果实现对企业内部重要文件的筛选包括以下步骤:

[0148] S71、获取每个数据段落的综合评分,并按照评分由高至低的顺序进行排序,得到排序结果;

[0149] S72、根据评分排序结果结合预设的评分阈值,选取评分高于阈值的前N个数据段落作为重要数据段落;

[0150] S73、统计每个文件数据中重要数据段落的比例,当该比例高于预设的比例阈值时,则确定该文件数据为重要文件。

[0151] 根据本发明的另一个实施例,提供了一种基于数据段落划分的量化式数据分析系统,该系统包括数据清洗模块、数据脱敏模块、变量候选集确定模块、关键变量确定模块、变量计量值计算模块、综合评分模块及数据筛选模块;

[0152] 其中,所述数据清洗模块用于获取企业内部的文件数据,并利用COPS算法对文件数据中的异常数据进行识别及清理;

[0153] 所述数据脱敏模块用于利用BERT语言模型对预处理后的文件数据中的敏感数据进行识别,并进行脱敏处理;

[0154] 所述变量候选集确定模块用于基于数据的特征对脱敏后的文件数据进行划分,并利用双向长短期记忆模型提取每个数据段落的特征,得到变量候选集;

[0155] 所述关键变量确定模块用于利用多头注意力机制计算变量候选集中每个变量的注意力权重,并根据每个变量的注意力权重确定关键变量;

[0156] 所述变量计量值计算模块用于通过计量法对关键变量进行计量得到关键变量的计量值,并利用标准化法对关键变量的计量值进行标准化处理;

[0157] 所述综合评分模块用于利用注意力机制结合学习模型自动确定标准化后的每个关键变量的权重,并结合双向长短期记忆模型计算每个数据段落的综合评分;

[0158] 所述数据筛选模块用于对每个数据段落的综合评分进行排序,并根据排序结果实现对企业内部重要文件的筛选。

[0159] 综上所述,借助于本发明的上述技术方案,本发明不仅可以利用COPS算法实现对企业内部文件数据中异常数据的识别和清除,而且还可以基于BERT语言模型识别和脱敏处

理敏感信息,有效地保护数据隐私,满足数据安全的要求,从而使得企业可以更高效地管理文件数据,节省存储空间,降低存储成本,同时,本发明还可以利用双向长短期记忆模型、多头注意力机制及神经网络模型来实现对企业内部重要文件的筛选,从而可以根据文件的重要程度对企业内部的文件数据进行存储,使得企业管理者在决策时可以快速准确地找到重要的文件数据,即快速准确的找到决策依据,从而可以有效地提升决策效率和精准性,进而可以更好地满足于企业的使用需求。

[0160] 以上所述实施例的各技术特征可以进行任意的组合,为使描述简洁,未对上述实施例中的各个技术特征所有可能的组合都进行描述,然而,只要这些技术特征的组合不存在矛盾,都应当认为是本说明书记载的范围。本领域普通技术人员可以理解实现上述实施例方法中的全部或部分步骤是可以通过程序来指令相关的硬件来完成,所述的程序可以存储于一计算机可读取存储介质中,该程序在执行时,包括以上方法所述的步骤,所述的存储介质,如:ROM/RAM、磁碟、光盘等。

[0161] 以上所述实施例仅表达了本发明的几种实施方式,其描述较为具体和详细,但并不能因此而理解为对发明专利范围的限制。应当指出的是,对于本领域的普通技术人员来说,在不脱离本发明构思的前提下,还可以做出若干变形和改进,这些都属于本发明的保护范围。因此,本发明专利的保护范围应以所附权利要求为准。

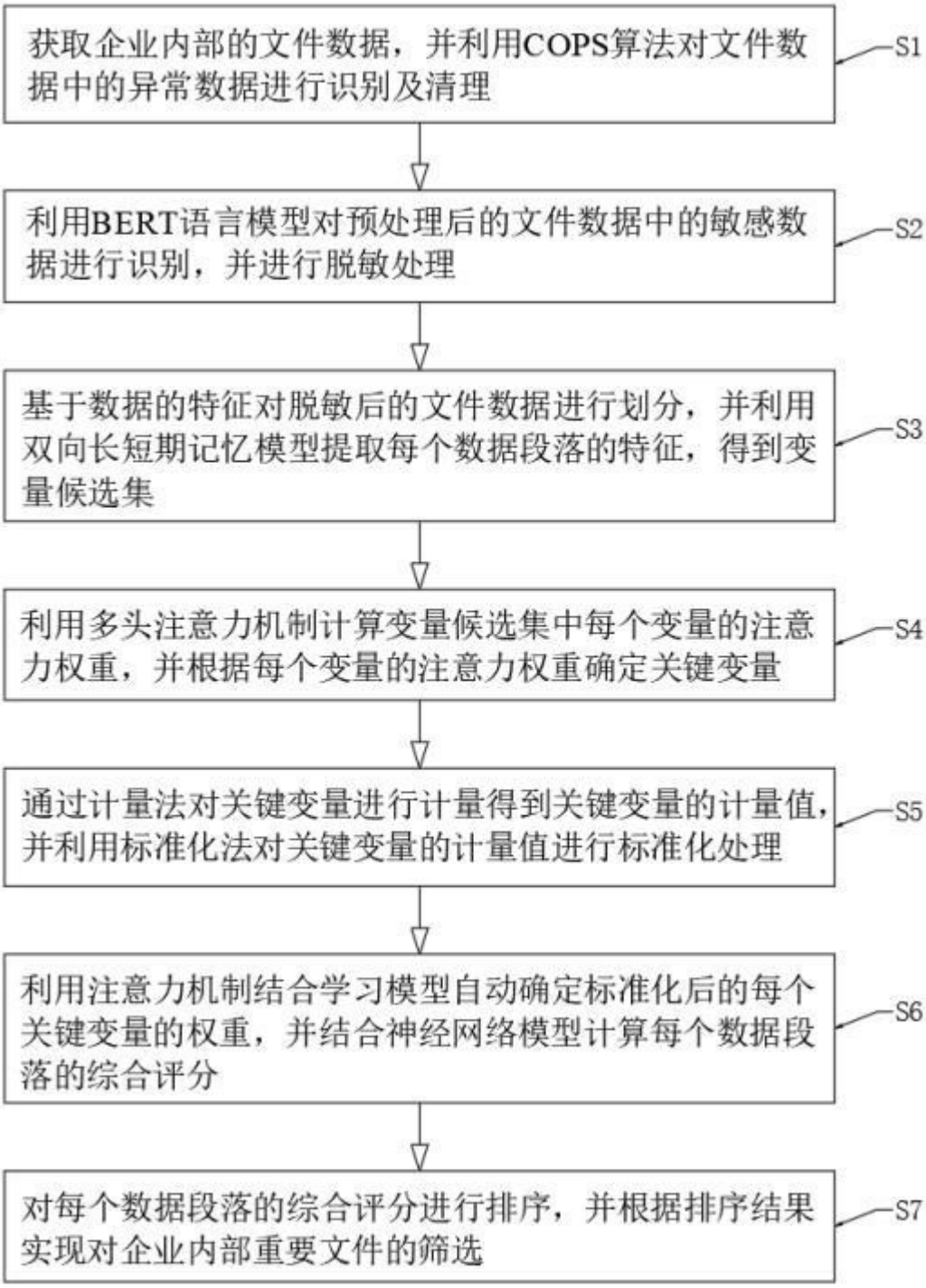

获取企业内部的文件数据，并利用COPS算法对文件数据中的异常数据进行识别及清理
S1
利用BERT语言模型对预处理后的文件数据中的敏感数据进行识别，并进行脱敏处理
S2
基于数据的特征对脱敏后的文件数据进行划分，并利用双向长短期记忆模型提取每个数据段落的特征，得到变量候选集
S3
利用多头注意力机制计算变量候选集中每个变量的注意力权重，并根据每个变量的注意力权重确定关键变量
S4
通过计量法对关键变量进行计量得到关键变量的计量值，并利用标准化法对关键变量的计量值进行标准化处理
S5
利用注意力机制结合学习模型自动确定标准化后的每个关键变量的权重，并结合神经网络模型计算每个数据段落的综合评分
S6
对每个数据段落的综合评分进行排序，并根据排序结果实现对企业内部重要文件的筛选
S7

图 1