

说明书

基于大数据的人力资源文本信息管理方法、装置和介质

技术领域

本申请涉及大数据技术领域，具体地涉及基于大数据的人力资源文本信息管理方法、装置和介质。

背景技术

传统的人力资源管理模式主要依赖于纸质文件和简单的数据库系统，这种依赖使得信息处理和分析的效率严重低下，无法充分挖掘和利用企业内部宝贵的人力资源数据。许多企业在关键的招聘、绩效评估、员工培训及离职管理等过程中，往往仅依靠个人经验和直觉来做出决策。这种方法不仅缺乏科学依据，决策的准确性和有效性也受到很大影响。与此同时，这种传统管理方式还导致了严重的信息孤岛现象，部门之间信息共享不足，无法实现综合分析和协同工作，进一步增加了管理的复杂性和成本。随着商业环境变化，企业不断积累了大量的非结构化文本数据，这些数据包括简历、绩效评估报告、培训记录、员工反馈以及离职面谈记录等，蕴含了丰富的信息和洞察。然而，由于这些数据通常以非结构化的形式存在，传统的数据处理方式难以高效提取其中的有效信息，导致企业无法真正利用这些数据来优化管理流程和决策。

因此，迫切需要一种基于大数据技术的解决方案，以便有效管理和利用这些丰富的文本数据，从而提升人力资源管理的质量和效率。

发明内容

为了解决上述技术问题，本申请实施例的目的是提供一种基于大数据的人力资源文本信息管理方法、装置和机器可读存储介质。

第一方面，本申请提供了一种基于大数据的人力资源文本信息管理方法，包括：

获取从多渠道采集的待分析对象的文本数据；

提取所述文本数据的代表性特征，所述代表性特征基于所述文本数据中的每个特征的基本信息得到，所述基本信息包括每个特征的编码、频率和长度；

对所述代表性特征进行处理以将所述代表性特征转化为另一维度的特征；

对处理后的代表性特征进行特征分类和特征分组，所述特征分类用于评价所述代表性特征的位阶，所述特征分组用于聚合属于相近特征的代表性特征；

根据所述代表性特征的特征分组和特征分类的结果，计算待分析对象的文本数据的特征得分。

优选地，提取所述文本数据的代表性特征，所述代表性特征基于所述文本数据中的每个特征的基本信息得到，所述基本信息包括每个特征的编码、频率和长度，包括：

采用公式（1）提取所述文本数据的代表性特征：

$$D = \sum \frac{1}{N} [f_1(w_k, T) * \log(\delta \|w_k\|^2 + \gamma f_2(w_k)) + \frac{f_1(w_k, T) * \sqrt{\gamma f_2(w_k) + 5 \|w_k\|^2}}{L(w_k, T) + K(w_k, T)}]$$

公式（1）

其中， D 表示代表性特征， T 表示文本数据， w_k 表示文本数据 T 中的第 k 个分词， f_1 表示分词编码函数， f_2 表示分词概率密度函数， $L(w_k, T)$ 表示文本数据 T 中的第 k 个分词 w_k 的长度， $K(w_k, T)$ 表示第 k 个分词 w_k 在文本数据 T 中出现的次数， γ 表示分词概率影响因子， δ 表示第 k 个分词 w_k 的缩放因子， N 表示文本数据 T 中的分词的总数， $\| \cdot \|$ 表示范数符号。

优选地，对所述代表性特征进行处理以将所述代表性特征转化为另一维

度的特征，包括：

对所述代表性特征 D ，采用公式（2）进行处理：

$$D_1 = \sum D * \frac{\log \frac{D + \sqrt{D^2 - 4\mu c}}{2\sigma} - a_1 \log \sqrt{D^2 + 5Db^2}}{e^{\sqrt{D^2 - 4\mu c}} + \alpha e^{-\beta D^2}} \quad \text{公式（2）}$$

其中， D 表示代表性特征， D_1 表示预处理后的代表性特征，即，另一维度的特征， μ 表示代表性特征 D 的均值， σ 表示代表性特征 D 的方差， a_1 表示代表性特征 D 的平滑因子， α 表示代表性特征 D 的正则化因子， β 表示代表性特征 D 的惩罚因子， b 表示代表性特征 D 的缩放因子， c 表示代表性特征 D 的均值平滑因子。

优选地，对处理后的代表性特征进行特征分类和特征分组中的特征分组，包括：

$$R_1 = \sum \| D_1 - V_j \|^2 * \log \frac{D + \sqrt{D^2 - 4\mu c}}{2\sigma} \quad \text{公式（3）}$$

其中， R_1 表示特征分组的结果， D_1 表示预处理后的代表性特征，即，另一维度的特征， μ 表示代表性特征 D 的均值， V_j 表示预处理后的代表性特征 D_1 中的第 j 个中心特征， σ 表示代表性特征 D 的方差， D 表示代表性特征， $\| \cdot \|$ 表示范数符号， c 表示代表性特征 D 的均值平滑因子。

优选地，对处理后的代表性特征进行特征分类和特征分组中的特征分类，包括：

$$R_2 = \sum \frac{D_1 w_k}{\sqrt{\| w_k \|^2 + \varepsilon}} + p_i \frac{D_1 \theta_i}{\| D_1 \| * \| \theta_i \|} \log \sqrt{D_1^2 - Db^2} \quad \text{公式（4）}$$

其中， R_2 表示特征分类的结果， D_1 表示预处理后的代表性特征，即，另

一维度的特征， b 表示代表性特征 D 的缩放因子， $|||$ 表示范数符号， w_k 表示文本数据 T 中的第 k 个分词， p_i 表示预处理后的代表性特征 D_1 中的第 i 个特征的权重因子， ε 表示预处理后的代表性特征 D_1 的惩罚因子， θ_i 表示预处理后的代表性特征 D_1 中的第 i 个特征的缩放因子。

优选地，根据所述代表性特征的特征分组和特征分类的结果，计算待分析对象的文本数据的特征得分，包括：

采用公式（5）计算所述待分析对象的特征得分：

$$R = \frac{\tau_1 e^{X_1 R_1} + \tau_2 e^{X_2 R_2}}{\tau_3 + \tau_4 (|| R_1 ||^2 + || R_2 ||^2)}$$

公式（5）

其中， R 表示特征得分， R_1 表示特征分组的结果， R_2 表示特征分类的结果， X_1 表示特征得分的结果 R_1 的权重影响因子， X_2 表示特征分组的结果 R_2 的权重影响因子， τ_1 表示特征得分的结果 R_1 的特征缩放系数， τ_2 表示特征分组的结果 R_2 的特征缩放系数， τ_3, τ_4 表示特征调节因子。

优选地，在根据所述代表性特征的特征分组和特征分类的结果，计算待分析对象的文本数据的特征得分之后，基于大数据的人力资源文本信息管理方法还包括：

将所述待分析对象的特征得分推送至分析用户。

优选地，在提取所述文本数据的代表性特征之前，基于大数据的人力资源文本信息管理方法还包括：

对所述文本数据进行数据清洗。

第二方面，本申请提供了一种基于大数据的人力资源文本信息管理装置，包括：

存储器，被配置成存储指令；以及

处理器，被配置成从存储器调用指令以执行上述的基于大数据的人力资

源文本信息管理方法。

第三方面，本申请提供了一种机器可读存储介质，该机器可读存储介质上存储有指令，该指令用于使得机器执行上述的基于大数据的人力资源文本信息管理方法。

本发明的方法包括：通过获取从多渠道采集的待分析对象的文本数据；提取所述文本数据的代表性特征，所述代表性特征基于所述文本数据中每个特征的基本信息得到，所述基本信息包括每个特征的编码、频率和长度；对所述代表性特征进行处理以将所述代表性特征转化为另一维度的特征；对处理后的代表性特征进行特征分类和特征分组，所述特征分类用于评价所述代表性特征的位阶，所述特征分组用于聚合属于相近特征的代表性特征；根据所述代表性特征的特征分组和特征分类的结果，计算所述待分析对象的文本数据的特征得分。在本发明中，通过采用大数据技术和智能化的数据分析方法，企业能够高效处理和分析大量的非结构化文本数据，显著提升人力资源管理的效率和准确性。数据采集、存储及分析过程的自动化减少了人工操作和人为错误，节省了时间和成本。本发明减弱了传统人力资源管理模式中的信息孤岛现象，促进了跨部门之间的信息共享与协作。各部门能够实时访问相关数据，进行综合分析，从而更好地协同工作，形成合力，有助于提高整体管理水平。本发明借助数据特征分析，企业可以构建更科学的决策模型，为招聘、绩效评估、培训与发展等提供数据驱动的支持。这种基于数据的决策方式能够提高决策的准确性，推动人力资源管理向智能化、数字化的方向发展。

本申请实施例的其它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

附图是用来提供对本申请实施例的进一步理解，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申请实施例，但并不构成对本申请实施例的限制。在附图中：

图1为本申请的实施例提供的一种基于大数据的人力资源文本信息管理方法的流程示意图；

图2为本申请的实施例提供的一种基于大数据的人力资源文本信息管理装置的结构框图。

具体实施方式

为使本申请实施例的目的、技术方案和优点更加清楚，下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行清楚、完整地描述，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申请实施例，并不用于限制本申请实施例。基于本申请中的实施例，本领域普通技术人员在没有做出创造性劳动的前提下所获得的所有其他实施例，都属于本申请保护的范围。

需要说明，若本申请实施例中有涉及方向性指示（诸如上、下、左、右、前、后……），则该方向性指示仅用于解释在某一特定姿态（如附图所示）下各部件之间的相对位置关系、运动情况等，如果该特定姿态发生改变时，则该方向性指示也相应地随之改变。

另外，若本申请实施例中有涉及“第一”、“第二”等的描述，则该“第一”、“第二”等的描述仅用于描述目的，而不能理解为指示或暗示其相对重要性或者隐含指明所指示的技术特征的数量。由此，限定有“第一”、“第二”的特征可以明示或者隐含地包括至少一个该特征。另外，各个实施例之间的技术方案可以相互结合，但是必须是以本领域普通技术人员能够实现为基础，当技术方案的结合出现相互矛盾或无法实现时应当认为这种技术方案的结合不存在，也不在本申请要求的基于大数据的人力资源文本信息管理方法的流程示意图。

在本申请中，文本信息可以等同于文本数据。数据在某些情况下可以等同于数据特征。本文中某些地方只针对一种情况（比如，招聘或者企业内部在职员工历史数据的分析评价）进行说明，这并不是对其余情况的限制，而是可以包括任何等同情况（比如，文本分析、基于员工数据的员工分析等），在此不做赘述。本申请不仅限于人力资源领域，而是可以用于任何采用本申请中的技术方案的数据分析方法。

如图 1 所示，本发明的实施例的基于大数据的人力资源文本信息管理方法可以包括下列步骤：

步骤 S10：获取从多渠道采集的待分析对象的文本数据；

在本发明实施例中，数据采集模块：数据采集从多个来源如招聘网站、内部管理系统、员工反馈平台等自动获取相关的结构化文本数据和非结构化文本数据，包括简历、绩效评估报告、培训记录、员工反馈和离职面谈记录。

首先，招聘网站是获取潜在人才信息的重要渠道。在符合法律法规的情况下，通过与主流招聘平台（如智联招聘、前程无忧、BOSS 直聘等）建立数据接口，实现简历数据的自动采集。这些简历包含了求职者的基本信息（姓名、性别、年龄、联系方式等）、教育背景（学历、专业、毕业院校、毕业时间等）、工作经历（工作单位、职位、工作时间、工作职责和业绩等）、技能证书以及职业规划等内容。采集过程中，运用网络爬虫技术和数据接口技术，按照预设的规则和频率，定期从招聘网站上抓取最新的简历数据，并将其存储到数据仓库中。

其次，企业内部管理系统涵盖了丰富的人力资源相关信息。企业内部管理系统可能包括人力资源信息系统（HRIS）、绩效管理系统、培训管理系统等。人力资源信息系统：该系统记录了员工的基本档案信息、考勤记录、薪酬福利数据、劳动合同信息等。通过与 HRIS 系统的对接，实现这些数据的自动抽取和整合。例如，从员工基本档案中获取员工的入职时间、所在部门、

岗位等信息；从考勤记录中提取员工的出勤天数、迟到早退次数等数据，为后续的绩效评估和人力资源分析提供基础支持。在绩效管理系统中，绩效评估报告是了解员工工作表现的关键依据。从绩效管理系统中自动采集员工的绩效评估结果、上级评价、自我评估以及绩效改进计划等信息。这些数据能够反映员工在一定时期内的工作目标完成情况、工作能力和工作态度，有助于企业识别高绩效和低绩效员工，为员工的晋升、调薪和培训发展提供参考。

在数据采集过程中，首先需要分析招聘网站的页面结构和数据格式，确定爬虫的抓取规则和路径。然后，使用专业的爬虫工具（如 **Scrapy**、**BeautifulSoup** 等）编写爬虫程序，模拟浏览器访问招聘网站，按照预设的规则提取所需的简历信息。并对采集到的数据进行合法性和完整性检查。

与企业内部管理系统和员工反馈平台的数据对接，主要通过数据接口技术实现。企业内部管理系统通常提供了标准的数据接口（如 **API** 接口），可以通过调用这些接口，按照约定的数据格式和协议，将系统中的数据自动传输到数据仓库中。同时，为了保证数据的安全性和稳定性，需要对数据接口进行身份验证和加密处理，防止数据泄露和非法访问。

在本实施例中，为了确保采集到的数据质量，采用严格的数据质量控制机制。如对数据进行验证、对数据进行审核等，以此保证采集的数据的质量，为后续对决策提供支撑提供高质量数据。具体地，在数据采集过程中，对采集到的数据进行实时验证，检查数据的完整性、准确性和合法性。例如，检查简历中的必填字段是否填写完整，绩效评估报告中的评分是否在合理范围内等。对于不符合要求的数据，及时进行标记和处理，确保数据的质量。此外，定期对采集到的数据进行审核，特别是对于一些关键信息和重要数据，审核过程中发现的数据问题，及时反馈给系统进行数据标记。再者，需要对来自不同数据源的数据进行一致性检查，确保同一员工的信息在不同系统中保持一致。例如，员工的基本信息在 **HRIS** 系统和招聘系统中的记录应该一

致，如果发现不一致的情况，需要及时进行排查和纠正，避免数据冲突和错误。

步骤 S20：提取所述文本数据的代表性特征，所述代表性特征基于所述文本数据中的每个特征的基本信息得到，所述基本信息包括每个特征的编码、频率和长度；

在对文本数据或文本信息进行特征提取之前，进一步需要对文本数据或文本信息进行必要的预处理操作。具体包括对文本数据或文本信息进行数据清洗、进行分词处理、进行词性标注与命名实体识别等操作。数据清洗中，对采集到的文本数据或文本信息进行清洗，去除其中的噪声数据，如 HTML 标签、JavaScript 代码、多余的空白字符、特殊符号等。同时，对文本中的错别字、缩写词进行纠正和扩展，提高文本的可读性和准确性。根据文本的语言类型，选择合适的分词工具（如中文的结巴分词、LTP，英文的 NLTK、Stanford CoreNLP 等）将文本分割成一个个独立的词语或短语。分词是后续文本分析的基础，能够将文本转化为计算机可以处理的基本单元。对分词后的文本进行词性标注，确定每个词语的词性（如名词、动词、形容词等），同时使用命名实体识别技术识别文本中的人名、地名、组织名、产品名等实体信息。这些信息有助于进一步理解文本的语义和结构。

根据待分析对象的特点和分析目的，确定需要提取的特征类型。常见的特征包括词汇特征（如关键词、高频词、主题词等）、语义特征（如情感倾向、语义相似度等）、结构特征（如句子长度、段落结构等）。例如，在分析产品用户反馈时，产品的功能名称、用户评价的情感词汇（如“满意”、“失望”）等可以作为重要的特征；在分析行业报告时，行业关键词、政策法规提及次数等可能是关键特征。

对经过分词处理后的文本进行遍历，识别出不同的特征。这些特征可以是与人力资源管理相关的特定概念，如“招聘”、“绩效评估”、“员工培

训”、“离职管理”等；也可以是描述管理方式、数据状态等方面的词汇，如“传统管理模式”、“信息孤岛”、“非结构化文本数据”等。

对文本中每个特征出现的次数进行统计。可以通过建立一个字典数据结构，以特征出现的次数为值，来记录每个特征的频率。例如，特征“招聘”在文本中出现了3次，那么在字典中对应频率值为{X: 3}，X可以为根据需要设定的编码值或者其他用于表征的值。为了使不同文本之间的特征频率具有可比性，可以对特征频率进行归一化处理。

此外，对于每个特征，可以计算其包含的字符数或词语数，以此作为特征的长度。例如，特征“传统管理模式”包含5个字符，那么其长度为5；特征“非结构化文本数据”包含7个字符，长度即为7。更全面地描述文本特征的长度信息，可以计算所有特征的平均长度。将所有特征的长度总和除以特征的总数，得到平均特征长度。这个指标可以反映文本中特征的整体长度水平。

根据实际需求和业务背景，为特征的特征属性、频率和长度分配不同的权重。例如，如果认为特征频率在反映文本信息中更为重要，可以为频率分配较高的权重；如果对特征的特定标识较为关注，可适当提高编码的权重；若希望突出特征的描述详细程度，可加大长度的权重。

最后，计算综合得分，对于每个特征，将其特征属性值、频率和长度分别乘以对应的权重，然后相加，得到每个特征的综合得分。计算公式可以表示为：综合得分 = 特征属性权重 × 特征属性值 + 频率权重 × 频率值 + 长度权重 × 长度值。

当然，在其他实施例中，还可以进一步考虑文本信息的每个特征的其他基本信息，例如文本的关键度、文本与分析目标的关联度等等基本信息来进行代表性特征提取的考虑因素。

最后，筛选代表性特征，根据综合得分对所有特征进行排序，选取得分

较高的若干个特征作为代表性特征。这些代表性特征能够在一定程度上概括文本的核心信息，反映文本所讨论的主要内容和关键要点。

通过以上步骤，处理器可以基于文本信息每个特征的特征属性、频率和长度，提取出具有代表性的特征，从而更深入地理解文本内容，为后续的数据分析和决策提供有力支持

步骤 S30：对所述代表性特征进行处理以将所述代表性特征转化为另一维度的特征；

对收集到的非结构化文本数据或数据特征进行整理和清洗，将其转化为可分析的格式的数据或数据特征（或者，另一维度的特征）。通过去除数据中的干扰因素，提高数据质量和分析的准确性，使得后续分析阶段的结果更加可靠。

非结构化文本数据的来源极为丰富多样，包括但不限于招聘媒体平台、企业内部文档、客户反馈记录等，不同来源的数据具有不同的特点和格式。这些数据没有固定的格式和结构，可能包含各种特殊字符、HTML 标签、排版格式等。例如，网页上的文本可能夹杂着大量的 HTML 标签用于控制页面布局和样式，这些标签对于文本内容的分析并无实际意义，反而会干扰数据处理。数据中往往存在着大量的噪声和冗余信息，如广告内容、重复的文本片段、无关的链接等。使用正则表达式或文本处理工具，去除文本中的特殊字符（如标点符号、表情符号、数学符号等）和 HTML 标签。这一步骤可以将文本简化为纯文本形式，便于后续处理。例如，对于包含 HTML 标签的文本“<p>这是一段 加粗 的文本</p>”，经过处理后可以得到“这是一段 加粗 的文本”。采用文本格式转换方法将文本转换为统一的大小写格式，以消除大小写差异对分析的影响。同时，对文本中的缩写、简写进行扩展，使其具有完整的语义。例如，将“don't”扩展为“do not”，“it's”扩展为“it is”等。此外，通过比较文本的相似度，识别并去除数据中的重

复文本片段。同时，根据业务需求和数据特点，制定规则去除噪声信息，如广告内容、无关的链接等。例如，可以通过匹配特定的关键词或正则表达式来识别广告文本，并将其从数据中删除。

在完成数据清洗后，对处理后的数据进行验证和质量检查，确保数据的准确性和完整性。这可以通过统计数据的基本特征（如文本长度分布、词汇频率等）、人工抽样检查等方式来实现。如果发现数据中仍然存在问题，需要返回前面的步骤进行进一步的处理。

经过上述整理和清洗步骤后，将非结构化文本数据或数据特征转化为可分析的格式的数据或数据特征（或者，另一维度的特征）。

通过对非结构化文本数据或数据特征进行整理和清洗，并将其转化为可分析的格式的数据或数据特征（或者，另一维度的特征），可以有效提高数据质量，为后续的数据分析和挖掘工作奠定坚实的基础。

步骤 S40：对处理后的代表性特征进行特征分类和特征分组，所述特征分类用于评价所述代表性特征的位阶（即，特征重要性的等级），所述特征分组用于聚类属于相近特征的代表性特征；

特征分组的目的是通过对简历、绩效评估以及员工反馈的特征进行分析和区分，聚合具有相似属性的特征，如技能、工作经历和反馈类型。具体来说，这一过程是通过最小化各特征与中心特征之间的平方距离来实现的。通过这种方法，系统可以高效地识别和归纳出那些在属性或语义上相似的文本特征，从而使其聚集在一起。

特征分组的优势在于减少数据冗余，优化数据结构，使得特征表示更加紧凑。这种紧凑的特征表示不仅能够提高数据存储的效率，还能显著加快后续的数据处理速度。通过聚合相似特征，系统可以更容易地洞察数据的整体趋势，并提供更加一致和精准的分析结果。

特征分类的目的是将简历和绩效评估等文档中提取到的特征进行系统

化的区分和标识。例如，这一过程可以包括将简历中的求职者按照技能类型（如技术类、管理类）进行分类，以及对绩效评估结果进行等级划分，以明确高绩效与低绩效员工的标准。通过特征分类，系统不仅能够有效识别和标记每个特征所对应的类别，还能帮助人力资源管理者在评估员工表现时，快速而准确地获取必要的信息。这种分类结果为后续的决策提供了依据，使得在进行人才招聘、绩效分析和员工发展时，能够基于数据做出更为科学和合理的判断。

步骤 S50：根据所述代表性特征的特征分组和特征分类的结果，计算待分析对象的文本数据的特征得分。

处理器通过结合特征分组和特征分类的结果，计算出特征得分，从而为人力资源管理提供灵活、动态和数据驱动的支持。不仅提高了决策的效率与准确性，还促进了人才管理的整体优化，为企业在人力资源方面提供了坚实的技术基础。处理器利用特征分组和特征分类的结果来计算特征得分，进而为人力资源分析提供支持。通过结合特征分组的结果和特征分类的结果，处理器能够计算出特征得分，人力资源部门可以使用特征得分对员工状况进行分析，帮助识别高潜力员工、评估培训需求或制定改善措施。例如，通过对绩效得分的分析，可以发现工作表现出色的员工，并为他们提供更多职业发展的机会，或者在招聘过程中，通过计算求职者简历的特征得分，可以快速筛选出最符合职位要求的人选，从而提高招聘效率。基于特征得分进行的分析可以帮助管理层制定更具针对性的人员配置、培训与发展策略。通过识别员工特征与业务成绩之间的关联，可以推动组织整体绩效的提升。处理器能够将来自不同来源的数据特征进行整合，形成一个统一的特征得分，避免信息孤岛，从而提升数据使用效率。

本发明的方法包括：通过获取从多渠道采集的待分析对象的文本数据或文本信息；提取所述文本数据或文本信息的代表性特征，所述代表性特征基

于所述文本数据或文本信息中的每个特征的基本信息得到，所述基本信息包括每个特征的编码、频率和长度；对所述代表性特征进行处理以将所述代表性特征转化为另一维度的特征；对处理后的代表性特征进行特征分类和特征分组，所述特征分类用于评价所述代表性特征的位阶，所述特征分组用于聚合属于相近特征的代表性特征；根据所述代表性特征的特征分组和特征分类的结果，计算所述待分析对象的文本数据或文本信息的特征得分。在本发明中，通过采用大数据技术和智能化的数据分析方法，企业能够高效处理和分析大量的非结构化文本数据，显著提升人力资源管理的效率和准确性。数据采集、存储及分析过程的自动化减少了人工操作和人为错误，节省了时间和成本。本发明减弱了传统人力资源管理模式中的信息孤岛现象，促进了跨部门之间的信息共享与协作。各部门能够实时访问相关数据，进行综合分析，从而更好地协同工作，形成合力，有助于提高整体管理水平。本发明借助数据特征分析，企业可以构建更科学的决策模型，为招聘、绩效评估、培训与发展等提供数据驱动的支持。这种基于数据的决策方式能够提高决策的准确性，推动人力资源管理向智能化、数字化的方向发展。

在本发明的另一实施例中，提取所述文本数据的代表性特征，所述代表性特征基于所述文本数据中的每个特征的基本信息得到，所述基本信息包括每个特征的编码、频率和长度，包括：

采用公式（1）提取所述文本信息的代表性特征：

$$D = \sum \frac{1}{N} [f_1(w_k, T) * \log(\delta \| w_k \|^2 + \gamma f_2(w_k)) + \frac{f_1(w_k, T) * \sqrt{\gamma f_2(w_k) + 5 \| w_k \|^2}}{L(w_k, T) + K(w_k, T)}]$$

公式（1）

其中， D 表示代表性特征， T 表示文本数据， w_k 表示文本数据 T 中的第 k 个分词， f_1 表示分词编码函数， f_2 表示分词概率密度函数， $L(w_k, T)$ 表示文

本数据 T 中的第 k 个分词 w_k 的长度, $K(w_k, T)$ 表示第 k 个分词 w_k 在文本数据 T 中出现的次数, γ 表示分词概率影响因子, δ 表示第 k 个分词 w_k 的缩放因子, N 表示文本数据 T 中的分词的总数, $|||$ 表示范数符号。

具体地, 在本实施例中, 采取公式 (1) 的方法提取文本数据的代表性特征。除了上述实施例中的采用权重值的计算方式计算得到文本数据的每个特征的分值, 根据分值确定代表性特征的方法之外, 本实施例采用一种复杂的线性计算方式来提取文本的代表性特征, 并且综合考虑了文本的长度、频率以及分词概率等信息来进行代表性特征提取。

具体地, 特征提取模块接受到文本数据之后会进行特征提取, 将上述文本特征应用公式 (1) 进行计算, 采用公式 (1) 的计算方法可以量化每个特征的价值。通过结合各个特征的编码、频率和长度等因素, 使得最终得到的代表性特征具备代表性和可靠性。

采用本实施例的计算方法, 结合词的频率和重要性进行综合评估, 确保提取出具有较高区分度和代表性的特征。通过引入词的长度、出现频率和分词概率等因素, 使得特征提取更为准确和全面。选择这种复杂的线性关系公式来确保不同的文本特征能被充分量化, 以便为后续分析提供精确的数值支持。

在本发明的另一实施例中, 对所述代表性特征进行处理以将所述代表性特征转化为另一维度的特征, 包括:

对所述代表性特征 D , 采用公式 (2) 进行处理:

$$D_1 = \sum D * \frac{\log \frac{D + \sqrt{D^2 - 4\mu\sigma}}{2\sigma} - a_1 \log \sqrt{D^2 + 5Db^2}}{e^{\sqrt{D^2 - 4\mu\sigma}} + \alpha e^{-\beta D^2}}$$

公式 (2)

其中, D 表示代表性特征, D_1 表示预处理后的代表性特征, 即, 另一维度的特征, μ 表示代表性特征 D 的均值, σ 表示代表性特征 D 的方差, a_1 表示

代表性特征 D 的平滑因子， α 表示代表性特征 D 的正则化因子， β 表示代表性特征 D 的惩罚因子， b 表示代表性特征 D 的缩放因子， c 表示代表性特征 D 的均值平滑因子。

采用本实施例的方法对文本数据特征进行处理，通过去除干扰因素和应用正则化，能够显著提高数据的清晰度和准确性。此外，综合考虑均值、方差和惩罚因子，有助于避免过拟合，提升模型的泛化能力。通过此处理公式，确保得到的预处理后的代表性特征更具代表性，并适合后续分析中的各类算法。本实施例的方法能够帮助修正潜在的偏差及噪音影响。

在本发明的另一实施例中，对处理后的代表性特征进行特征分类和特征分组中的特征分组，包括：

$$R_1 = \sum \| D_1 - V_j \|^2 * \log \frac{D + \sqrt{D^2 - 4\mu c}}{2\sigma}$$

公式(3)

其中， R_1 表示特征分组的结果， D_1 表示预处理后的代表性特征，即，另一维度的特征， μ 表示代表性特征 D 的均值， V_j 表示预处理后的代表性特征 D_1 中的第 j 个中心特征， σ 表示代表性特征 D 的方差， D 表示代表性特征， $\| \cdot \|$ 表示范数符号， c 表示代表性特征 D 的均值平滑因子。

减少数据冗余：通过将相似特征聚集在一起，可以突出有价值的信息，减少处理时的复杂度。

高效特征表示：紧凑的特征表示加速后续分析和计算。

选择理由：

选择这样的分组方法可以更好地为后续的分类和分析提供清晰的数据结构，使得数据处理变得更高效且一致。

在本发明的另一实施例中，对处理后的代表性特征进行特征分类和特征分组中的特征分类，包括：

$$R_2 = \sum \frac{D_1 w_k}{\sqrt{\|w_k\|^2 + \varepsilon}} + p_i \frac{D_1 \theta_i}{\|D_1\| * \|\theta_i\|} \log \sqrt{D_1^2 - D b^2}$$

公式 (4)

其中, R_2 表示特征分类的结果, D_1 表示预处理后的代表性特征, 即, 另一维度的特征, b 表示代表性特征 D 的缩放因子, $\| \cdot \|$ 表示范数符号, w_k 表示文本数据 T 中的第 k 个分词, p_i 表示预处理后的代表性特征 D_1 中的第 i 个特征的权重因子, ε 表示预处理后的代表性特征 D_1 的惩罚因子, θ_i 表示预处理后的代表性特征 D_1 中的第 i 个特征的缩放因子。

采用本实施例的方法, 通过明确分类, 提高了特征的可解释性, 便于 HR 进行决策。确保高效员工与低效员工的标准系统化, 帮助及时采取适当的管理措施。这样的分类方法旨在清晰区分特征属性, 使得在评估和决策时更为高效, 符合业务需求。

在本发明的另一实施例中, 根据所述代表性特征的特征分组和特征分类的结果, 计算待分析对象的文本数据的特征得分, 包括:

采用公式 (5) 计算所述待分析对象的特征得分:

$$R = \frac{\tau_1 e^{X_1 R_1} + \tau_2 e^{X_2 R_2}}{\tau_3 + \tau_4 (\|R_1\|^2 + \|R_2\|^2)}$$

公式 (5)

其中, R 表示特征得分, R_1 表示特征分组的结果, R_2 表示特征分类的结果, X_1 表示特征得分的结果 R_1 的权重影响因子, X_2 表示特征分组的结果 R_2 的权重影响因子, τ_1 表示特征得分的结果 R_1 的特征缩放系数, τ_2 表示特征分组的结果 R_2 的特征缩放系数, τ_3, τ_4 表示特征调节因子。

在复杂多变的商业环境中, 人力资源管理对于企业的成功至关重要。为了更全面、精准地评估员工, 本发明引入了一种创新的综合评估方法, 该方法融合了多维特征、指数增长平滑性、动态调整能力、避免信息孤岛以及支持决策制定等关键特性。

在传统的员工评估中，往往只关注单一或少数几个方面的特征，这使得对员工的了解不够全面和深入。而本发明的综合评估方法充分考虑了多维特征，涵盖了员工在工作中的各个层面，包括但不限于工作绩效、专业技能、团队协作能力、创新思维、沟通能力以及学习能力等。

为了更有效地处理这些丰富的特征信息，本发明运用先进的数据分析算法对特征进行分组和分类。例如，将与工作成果直接相关的特征归为一组，如项目完成率、工作质量等；将反映员工个人能力和素质的特征归为另一组，如专业知识水平、解决问题的能力等；把体现员工社交和团队协作方面的特征归为一组，像团队合作默契度、沟通效果等；此外，对这些特征再进行级别高低的分类，即等级划分。通过这样的分组和分类，能够更清晰地了解每个特征在整体评估中的位置和作用。

然后，本发明运用算法公式将这些分组和分类的结果结合在一起。这个公式并非简单的线性相加，而是综合考虑了各个特征之间的相互关系和权重。例如，对于不同类型的工作岗位，某些特征可能具有更高的权重。通过这种方式，最终得分能够全面反映员工在各个维度上的表现，为我们呈现出一个完整、立体的员工画像。

在综合评估中，考虑到不同特征对于员工表现和潜力的重要性是不同的。为了更准确地体现这种差异，本发明引入了指数增长平滑性这一特性。

具体而言，本发明使用指数函数来处理特征得分。指数函数的特性使得特征得分对于分组和分类的重要性具有更敏感的反应。对于那些低得分的特征，其在对最终得分的贡献上不会太大。这是因为指数函数的曲线在低得分区域较为平缓，随着得分的降低，其对最终得分的影响会迅速减小。这也是本发明采用此计算方法来计算最终的特征得分的目的之一。

相反，对于高得分的特征，指数函数的曲线在高得分区域变得陡峭，这意味着高得分特征的影响将被显著放大。例如，一个在创新思维方面表现卓

越的员工，其在这一特征上的高分通过指数函数的作用，会在最终得分中得到充分体现，从而与其他在该方面表现一般的员工拉开差距。这种方式大大提升了得分的区分度，使分析者能够更准确地识别出那些在某些关键方面具有突出优势的员工。

通过调整权重因子以及缩放系数，可以根据实际情况对特征得分的计算进行动态调整。例如，在企业处于业务扩张阶段时，可能更注重员工的市场开拓能力和团队协作能力，此时可以适当提高这两个方面特征的权重；而当企业进入技术创新阶段，对员工的专业技能和创新思维的要求更高，我们就可以相应地调整这些特征的权重。

为了避免信息孤岛对员工评估造成的不利影响，我们的综合评估模块通过整合来自不同来源的特征得分，形成了一个统一的视图。这个过程不仅仅是数据的简单汇总，更是对不同来源信息的深度融合和分析。

例如，我们可以将人力资源管理系统中记录的员工基本信息和培训记录，与项目管理系统中反映的员工在项目中的实际表现相结合，再融入绩效考核系统中的量化数据，从而全面、准确地评估员工的工作能力和潜力。通过消除数据孤岛，我们能够在决策时拥有更全面的信息基础，避免因信息不完整而导致的决策失误。

最终的特征得分不仅仅是一个数字，它具有重要的实际应用价值，能够直接支持人力资源管理中的一系列决策。

在识别高潜力员工方面，通过综合评估得到的特征得分可以帮助我们准确地筛选出那些在多个维度上表现优秀且具有较大发展潜力的员工。这些员工可能是企业未来的核心力量，为他们提供更多的发展机会和资源，有助于提升企业的整体竞争力。

在精准招聘过程中，我们可以将招聘岗位所需的关键特征与求职者的特征得分进行匹配，从而更准确地筛选出符合岗位要求的人才。这种基于数据

的招聘方式能够提高招聘效率和质量，减少招聘风险。

对于优化培训需求，特征得分可以帮助我们发现员工在哪些方面存在不足，从而有针对性地制定培训计划。例如，如果某个员工在沟通能力方面得分较低，我们可以为其安排相关的沟通技巧培训课程，帮助员工提升能力，实现个人和企业的共同发展。

综上所述，本发明实施例的公式将特征分组和分类的结果结合在一起，使得最终得分能够反映出更全面的信息。这种综合评估有助于更好理解员工的表现和潜力。此外，使用指数函数使得特征得分对于分组和分类的重要性有更敏感的反应，低得分的特征在对最终得分的贡献上不会太大，而高得分特征的影响将被放大，从而提升了得分的区分度。通过调整权重因子以及缩放系数能够根据需要动态调整特征得分的计算，以适应不同的业务场景和管理目标。通过整合来自不同来源的特征得分，模块能够形成统一的视图，消除数据孤岛，有助于在决策时形成全面的信息基础。最终的特征得分可以直接支持人力资源管理中的一系列决策，帮助识别高潜力员工、精准招聘、优化培训需求等。

在本发明的另一实施例中，在根据所述代表性特征的特征分组和特征分类的结果计算所述待分析对象的特征得分之后，基于大数据的人力资源文本信息管理方法还包括：

将所述待分析对象的特征得分推送至分析用户。

本发明实施例中最终将得到的特征得分推送至需要此数据的用户，例如人力资源分析部门。通过此特征得分，人力资源分析部门可以根据自身需要做出决策。

下面将以具体的实施例完整地描述本方案的内容：

假设需要分析一批软件工程师的简历，以识别高潜力的候选人并进行有效的人员配置。处理器将按照以下步骤进行数据处理：

1. 数据采集

处理器首先获得了一组简历数据，每份简历包含以下信息：

姓名

联系方式

教育背景

专业技能（如编程语言、工具等）

项目经验

工作经历

个人介绍

其中，上述简历数据的源数据示例：

```
{  
  "姓名": "张三",  
  "联系方式": "123456789",  
  "教育背景": "计算机科学与技术",  
  "专业技能": ["Python", "Java", "机器学习"],  
  "项目经验": ["智能推荐系统开发", "大数据处理平台"],  
  "工作经历": ["A 公司(2019-2021)", "B 公司(2022 至今)"],  
  "个人介绍": "对技术充满热情，擅长学习新技术。"  
}
```

2. 特征提取

假设需要关心的是以下特征：

技能（Python、Java、机器学习）

项目数量

工作经验年限

特征提取的结果将如下所示：

特征	Python	Hava	机器学习	项目数量	工作经验年限
张三	1	1	1	2	5

解释：

Python、Java 和机器学习等技能存在则编码为 1，不存在则为 0。

项目数量和工作经验年限以整数字段表示。

3.特征预处理

在特征预处理阶段，处理器将检查并清洗数据。假设没有缺失值，此时将保持数据形式的二进制表示。

4. 特征分组

将特征分组，以便聚合成几个主要类别。处理器将技能、项目经验和工作经验聚集到以下类别中：

编程技能： 使用二进制编码表明数量

项目经验： 使用项目数量表示

工作经验： 使用经经验年限表示

特征分组的结构化结果（假设编程技能数量为 1-3 的间隔）：

分组	编程技能数量	项目经验	工作经验
张三	3	2	5

5. 特征分类

在特征分类阶段，我们将编码分为基础技能与高级技能。基于技能评估：

基础技能

Java

高级技能

Python

机器学习

计算后的特征分类结果为：

分类	基础技能数量	高级技能数量
张三	1	2

6. 特征得分计算

根据特征分组的结果 R_1 和特征分类的结果 R_2 来计算特征得分 R 。假设使用以下权重及调节因子：

根据特征分组的结果 R_1 和特征分类的结果 R_2 来计算特征得分 R 。假设使用以下权重及调节因子：

- $X_1=0.5$ （特征分组的权重）
- $X_2=1.5$ （特征分类的权重）
- $\tau_1=1$
- $\tau_2=1$
- $\tau_3=1$
- $\tau_4=1$

应用公式计算：

$$R = \frac{\tau_1 e^{X_1 R_1} + \tau_2 e^{X_2 R_2}}{\tau_3 + \tau_4 (\|R_1\|^2 + \|R_2\|^2)}$$

公式（5）

代入数值计算：

假设 R_1 的总值=3+2+5=10

假设 R_2 的总值=1+2=3

最终可以得出 R 的值：

假设计算结果为 $R=85$

7. 结果分析

通过特征得分 R ：

识别高潜力员工：根据特征得分高低，张三拥有较高的得分（85），可

以被视为高潜力员工。

制定招聘策略：人力资源团队可以根据得分快速决定该候选人的面试阶段或进一步筛选。

通过特征得分及对应的特点分析，完成人员选择及决策支持，使招聘过程更高效和具有针对性。

图2为本申请的实施例提供的一种基于大数据的人力资源文本信息管理装置的结构框图。如图2所示，基于大数据的人力资源文本信息管理装置可以包括：

存储器 210，被配置成存储指令；以及

处理器 220，被配置成从存储器 210 调用指令以及在执行指令时能够实现上述的基于大数据的人力资源文本信息管理方法。

本发明的实施例还提供一种基于大数据的人力资源文本信息管理系统，所述系统可以包括：

数据采集模块，用于获取从多渠道采集的待分析对象的文本数据；

特征提取模块，用于提取所述文本数据的代表性特征，所述代表性特征基于所述文本数据中的每个特征的基本信息得到，所述基本信息包括每个特征的编码、频率和长度；

特征预处理模块，用于对所述代表性特征进行处理以将所述代表性特征转化为另一维度的特征；

文本数据管理模块，用于对处理后的代表性特征进行特征分类和特征分组，所述特征分类用于评价所述代表性特征的位阶，所述特征分组用于聚类属于相近特征的代表性特征；

文本数据分析模块，用于根据所述代表性特征的特征分组和特征分类的结果，计算所述待分析对象的文本数据的特征得分。

本发明实施例提供了基于大数据的人力资源文本信息管理系统（即，基

于数据分析的决策支持系统），针对人力资源管理需求所设计的特定机器学习模型算法，以及用于获取、清洗、分析非结构化文本数据的具体方法和技术细节等，以帮助人力资源管理人员做出更科学的决策。

本申请实施例还提供一种机器可读存储介质，该机器可读存储介质上存储有指令，该指令用于使得机器执行上述的基于大数据的人力资源文本信息管理方法。

本领域内的技术人员应明白，本申请的实施例可提供为方法、系统或计算机程序产品。因此，本申请可采用完全硬件实施例、完全软件实施例、或结合软件和硬件方面的实施例的形式。而且，本申请可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质（包括但不限于磁盘存储器、**CD-ROM**、光学存储器等）上实施的计算机程序产品的形式。

本申请是参照根据本申请实施例的方法、设备（系统）和计算机程序产品的流程图和/或方框图来描述的。应理解可由计算机程序指令实现流程图和/或方框图中的每一流程和/或方框、以及流程图和/或方框图中的流程和/或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理设备的处理器以产生一个机器，使得通过计算机或其他可编程数据处理设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的装置。

这些计算机程序指令也可存储在能引导计算机或其他可编程数据处理设备以特定方式工作的计算机可读存储器中，使得存储在该计算机可读存储器中的指令产生包括指令装置的制造品，该指令装置实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能。

这些计算机程序指令也可装载到计算机或其他可编程数据处理设备上，使得在计算机或其他可编程设备上执行一系列操作步骤以产生计算机实现的处理，从而在计算机或其他可编程设备上执行的指令提供用于实现在流程

图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的步骤。

在一个典型的配置中，计算设备包括一个或多个处理器（CPU）、输入/输出接口、网络接口和内存。

存储器可能包括计算机可读介质中的非永久性存储器，随机存取存储器（RAM）和/或非易失性内存等形式，如只读存储器（ROM）或闪存（flash RAM）。存储器是计算机可读介质的示例。

计算机可读介质包括永久性和非永久性、可移动和非可移动媒体可以由任何方法或技术来实现信息存储。信息可以是计算机可读指令、数据结构、程序的模块或其他数据。计算机的存储介质的例子包括，但不限于相变内存（PRAM）、静态随机存取存储器（SRAM）、动态随机存取存储器（DRAM）、其他类型的随机存取存储器（RAM）、只读存储器（ROM）、电可擦除可编程只读存储器（EEPROM）、快闪记忆体或其他内存技术、只读光盘只读存储器（CD-ROM）、数字多功能光盘（DVD）或其他光学存储、磁盒式磁带，磁带磁盘存储或其他磁性存储设备或任何其他非传输介质，可用于存储可以被计算设备访问的信息。按照本文中的界定，计算机可读介质不包括暂存电脑可读媒体（transitory media），如调制的数据信号和载波。

还需要说明的是，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、商品或者设备不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、方法、商品或者设备所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括要素的过程、方法、商品或者设备中还存在另外的相同要素。

以上仅为本申请的实施例而已，并不用于限制本申请。对于本领域技术人员来说，本申请可以有各种更改和变化。凡在本申请的精神和原理之内所

作的任何修改、等同替换、改进等，均应包含在本申请的权利要求范围之内。