(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 113254659 A
(43)申请公布日 2021.08.13

(21)申请号 202110153678.3

(22)申请日 2021.02.04

(71)申请人 天津德尔塔科技有限公司
地址 300384 天津市滨海新区高新区华苑产业区工华道2号8号楼-1-3

(72)发明人 衣秀 张成 苏卫卫 黄瑞 杨文起

(74)专利代理机构 天津市尚仪知识产权代理事务所(普通合伙) 12217
代理人 邓琳

(51)Int.Cl.
G06F 16/36(2019.01)
G06F 16/31(2019.01)
G06F 16/33(2019.01)
G06F 16/335(2019.01)
G06F 16/35(2019.01)
G06F 40/295(2020.01)
G06F 40/289(2020.01)

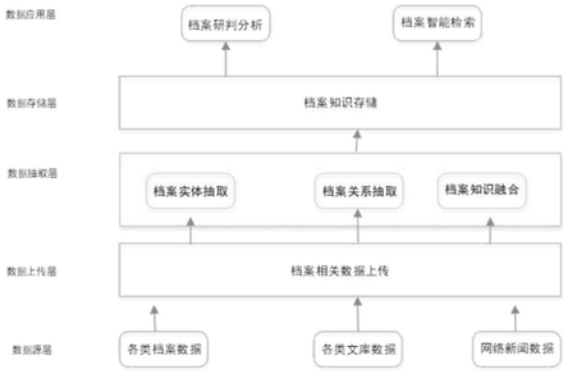
权利要求书2页 说明书9页 附图3页

(54)发明名称

一种基于知识图谱技术的档案研判方法及系统

(57)摘要

本发明提供一种基于知识图谱技术的档案研判方法及系统，所述档案研判方法包括以下步骤：档案上传；档案实体抽取：抽取人名、地名、机构名、高校、著作、事件、建筑和资料实体；档案关系抽取：抽取两个实体之间的关联关系，构建档案知识图谱的边；档案知识融合：将抽取的不同来源的档案实体进行融合，产出一个唯一的属性全面的档案实体；档案知识存储：对抽取的档案实体和关系数据进行存储，可以存储在图数据库中；档案研判分析：档案智能检索。本发明在知识图谱、机器学习技术的支撑下，充分利用创建的档案知识图谱，实现各类相关档案实体的关联推荐，以及相关档案的精准推荐，帮助档案研判人员扩展阅读，并拓展档案研判的深度和广度。



CN 113254659 A

1.一种基于知识图谱技术的档案研判方法及系统,其特征在于,所述档案研判方法包括以下步骤:

步骤1:档案上传;

步骤2:档案实体抽取:抽取人名、地名、机构名、高校、著作、事件、建筑和资料实体;

步骤3:档案关系抽取:抽取两个实体之间的关联关系,构建档案知识图谱的边;

步骤4:档案知识融合:将抽取的不同来源的档案实体进行融合,产出一个唯一的属性全面的档案实体;

步骤5:档案知识存储:对抽取的档案实体和关系数据进行存储,可以存储在图数据库中;

步骤6:档案研判分析:包括档案热词分析、档案敏感词分析、档案专有词识别、档案自动分类、档案聚类,基于利用自然语言处理技术和深度学习技术,辅助档案分析研判人员对档案数据进行高效研判分析;

步骤7:档案智能检索:档案智能检索包括检索语义关键词推荐、相关人物推荐、相关机构推荐、相关事件推荐、相关著作推荐、相关政策推荐,通过语义关联推荐,帮助用户扩展阅读,提升信息查找与分析的深度和广度,按照各章节权重对相似度进行加权平均,得到文档整体相似度。

2.如权利要求1所述的一种基于知识图谱技术的档案研判方法及系统,其特征在于,所述步骤2包括以下步骤:

步骤21:将语料向量化后输入到网络;

步骤22:利用双向LSTM自动提取输入的特征;

步骤23:利用CRF层对上层结果进行句子级别的序列标注。

3.如权利要求1所述的一种基于知识图谱技术的档案研判方法及系统,其特征在于,所述步骤3包括以下步骤:

步骤31:对输入的档案文本进行句子内容解析,并进行向量化,再输入到下一层网络;

步骤32:使用双向LSTM进行正向和反向学习上下文信息;

步骤33:使用Attention机制选择一组异常候选集进行分析,生成一个权重向量,通过与这个权重向量相乘,使每一次迭代中的词汇级的特征合并为句子级的特征;

步骤34:将Attention层产出的向量送入softmax分类器来预测标签值,选择最大的概率的标签作为预测标签。

4.如权利要求1所述的一种基于知识图谱技术的档案研判方法及系统,其特征在于,所述步骤6包括如下步骤:

步骤41:对档案热词进行分析;

步骤42:档案关键词抽取:抽取标识输入档案的一组词;

步骤43:档案自动分类:对位置的编研档案自动分类;

步骤44:档案人物关联分析:产出与该人物相相关联的事件、机构。

5.如权利要求1所述的一种基于知识图谱技术的档案研判方法及系统,其特征在于,所述档案研判系统包括:档案上传模块、档案实体抽取模块、档案关系抽取模块、档案知识融合模块、档案知识存储模块、档案研判分析模块。

6.如权利要求5所述的一种基于知识图谱技术的档案研判方法及系统,其特征在于,所

述的档案上传模块用于上传采集档案文档，并利用自然语言处理技术实现档案文档标题、作者、关键字、分类、摘要这些元数据进行自动标引；所述的档案实体抽取模块用于从档案中自动抽取人名、地名、机构名、事件、建筑、资料这些档案实体；所述的档案关系抽取模块用于从档案中抽取档案实体之间的关联关系从而形成档案知识图谱；所述档案知识融合模块用于对从各个档案文献中抽取的档案实体属性进行歧义消除和指代消解，实现档案实体属性的合并与融合；所述档案知识存储模块用于将抽取构建的档案知识图谱进行存储，并提供档案知识图谱的数据查询；所述档案研判分析模块用于给档案研判分析人员提供档案热词分析、档案敏感词分析、档案专有词识别、档案自动分类、档案聚类、档案人物关系抽取等智能化工具，同时提供档案智能检索，将档案知识图谱和档案全文检索进行整合，实现档案检索关键词、档案实体、相关档案的精准推荐。

一种基于知识图谱技术的档案研判方法及系统

技术领域

[0001] 本发明属于档案管理与研判技术领域，尤其涉及一种基于知识图谱技术的档案研判方法及系统。

背景技术

[0002] 档案馆是国家发展历史记录的珍藏地，档案是过去工作和历史情况的记录，是历史的真实凭证，随着信息化的发展，数字档案馆已经是各级档案馆的必建之物。然而，随着国民和社会经济的发展，传统的档案信息化服务已经不能满足对档案日益增长的现实需求，大力推进智慧档案建设已经成为当今社会发展的必然需求。

[0003] 以往对档案的研判分析，主要通过多维数据分析工具、数据挖掘工具等进行数据统计分析层面的分析，但是智慧档案建设需要我们能够挖掘档案更深层的含义，如某一人和其它人的关联关系、和相关机构的关系、和某些事件的关系，从而实现档案研判对象相关人物、城市、事件、著作等的关联分析与推荐。

[0004] 因此，需要一种基于知识图谱技术的档案研判方法及系统，结合深度学习技术，构建档案知识图谱，辅助用户进行档案关联分析和智能检索，为档案研判决策提供数据支撑。

发明内容

[0005] 为了解决上述技术问题，本发明提供一种基于知识图谱技术的档案研判方法及系统，所述档案研判方法包括以下步骤：

[0006] 步骤1:档案上传；

[0007] 步骤2:档案实体抽取:抽取人名、地名、机构名、高校、著作、事件、建筑和资料实体；

[0008] 步骤3:档案关系抽取:抽取两个实体之间的关联关系，构建档案知识图谱的边；

[0009] 步骤4:档案知识融合:将抽取的不同来源的档案实体进行融合，产出一个唯一的属性全面的档案实体；

[0010] 步骤5:档案知识存储:对抽取的档案实体和关系数据进行存储，可以存储在图数据库中；

[0011] 步骤6:档案研判分析:包括档案热词分析、档案敏感词分析、档案专有词识别、档案自动分类、档案聚类，基于利用自然语言处理技术和深度学习技术，辅助档案分析研判人员对档案数据进行高效研判分析。

[0012] 步骤7:档案智能检索:档案智能检索包括检索语义关键词推荐、相关人物推荐、相关机构推荐、相关事件推荐、相关著作推荐、相关政策推荐，通过语义关联推荐，帮助用户扩展阅读，提升信息查找与分析的深度和广度，按照各章节权重对相似度进行加权平均，得到文档整体相似度。

[0013] 优选的，所述步骤2包括以下步骤：

[0014] 步骤21:将语料向量化后输入到网络；

[0015] 步骤22:利用双向LSTM自动提取输入的特征;

[0016] 步骤23:利用CRF层对上层结果进行句子级别的序列标注。

[0017] 优选的,所述步骤3包括以下步骤:

[0018] 步骤31:对输入的档案文本进行句子内容解析,并进行向量化,再输入到下一层网络;

[0019] 步骤32:使用双向LSTM进行正向和反向学习上下文信息;

[0020] 步骤33:使用Attention机制选择一组异常候选集进行分析,生成一个权重向量,通过与这个权重向量相乘,使每一次迭代中的词汇级的特征合并为句子级的特征;

[0021] 步骤34:将Attention层产出的向量送入softmax分类器来预测标签值,选择最大的概率的标签作为预测标签。

[0022] 优选的,所述步骤6包括如下步骤:

[0023] 步骤61:对档案热词进行分析;

[0024] 步骤62:档案关键词抽取:抽取标识输入档案的一组词;

[0025] 步骤63:档案自动分类:对位置的编研档案自动分类;

[0026] 步骤64:档案人物关联分析:产出与该人物相相关联的事件、机构;

[0027] 优选的,所述档案研判系统包括:档案上传模块、档案实体抽取模块、档案关系抽取模块、档案知识融合模块、档案知识存储模块、档案研判分析模块。

[0028] 优选的,所述的档案上传模块用于上传采集档案文档,并利用自然语言处理技术实现档案文档标题、作者、关键字、分类、摘要这些元数据进行自动标引;所述的档案实体抽取模块用于从档案中自动抽取人名、地名、机构名、事件、建筑、资料这些档案实体;所述的档案关系抽取模块用于从档案中抽取档案实体之间的关联关系从而形成档案知识图谱;所述档案知识融合模块用于对从各个档案文献中抽取的档案实体属性进行歧义消除和指代消解,实现档案实体属性的合并与融合;所述档案知识存储模块用于将抽取构建的档案知识图谱进行存储,并提供档案知识图谱的数据查询;所述档案研判分析模块用于给档案研判分析人员提供档案热词分析、档案敏感词分析、档案专有词识别、档案自动分类、档案聚类、档案人物关系抽取等智能化工具,同时提供档案智能检索,将档案知识图谱和档案全文检索进行整合,实现档案检索关键词、档案实体、相关档案的精准推荐。

[0029] 与现有技术相比,本发明的有益效果为:

[0030] 1、本发明在知识图谱、机器学习技术的支撑下,在档案搜索中充分利用创建的档案知识图谱,实现各类相关档案实体的关联推荐,以及相关档案的精准推荐,从而帮助档案研判人员扩展阅读,并拓展档案研判的深度和广度。

[0031] 2、本发明通过提供档案智能编研方法包括档案热词分析、档案敏感词分析、档案专有词识别、档案自动分类、档案聚类、档案实体关联分析多种自动化方法,辅助编研人员提取高质量档案信息,大幅提高档案编研工作效率。

附图说明

[0032] 图1为本发明的总体流程图;

[0033] 图2为本发明的BILSTM-CRF架构图;

[0034] 图3为本发明的基于Attention机制的双向LSTM架构图;

[0035] 图4为本发明的基于知识图谱的智能搜索的流程图；

[0036] 图5为本发明的档案编研系统总体结构示意图。

具体实施方式

[0037] 以下结合附图对本发明做进一步描述：

[0038] 实施例：

[0039] 如附图1所示，一种基于知识图谱技术的档案编研方法，具体步骤如下所示：

[0040] 步骤1:档案上传：

[0041] 档案数据主要包括档案目录数据、档案全文数据、档案照片数据、档案多媒体数据、档案专业档案数据、档案历史文献数据和档案管理数据。

[0042] 用户可以通过上传按钮或者拖拽等方式上传档案，利用自然语言处理等技术自动形成档案元数据，如档案名称、档案作者、档案时间、档案分类、档案关键字、档案摘要、档案密级等信息，同时记录该份数据的上传、查询、修改和删除操作记录，大幅提高档案元数据录入的效率。

[0043] 步骤2:档案实体抽取：

[0044] 档案实体主要包括人名、地名、机构名、高校、著作、事件、建筑和资料这些实体。

[0045] 档案实体抽取方法主要采用基于模式或规则的档案关系抽取方法、基于序列标注监督学习技术的档案关系抽取方法、基于文本分类监督学习的档案关系抽取方法，本文以基于文本分类等监督学习的档案关系抽取方法为例：

[0046] 基于文本分类监督学习的档案关系抽取方法，首先使用BIO标注法对一部分档案数据进行实体标注，然后利用BILSTM-CRF、SVM、贝叶斯等机器学习和深度学习方法构建模型。

[0047] 以BILSTM-CRF方法为例，档案命名实体识别的输入是档案中一个句子对应的词序列$s=<w_1,w_2\cdots w_n>$，输出是档案实体及对应的起止位置，当使用监督学习的方法时，NER问题是序列标注问题，使用BIO标注法处理档案语料，其中B 表示档案实体的起始位置，I表示档案实体的中间或者结束位置，O表示不是档案实体；如果标注人名，人名起始标注B-PERSON，I-PERSON表示人名实体的中间或者结束位置，地名的起始标注B-LOC，I-LOC表示地名实体的中间或者结束位置。

[0048] 在NER识别任务中，常用的深度网络有CNN卷积神经网络和RNN循环神经网络，CNN卷积神经网络用于向量特征学习，RNN循环神经网络同时学习向量特征和序列标注，以一种RNN的长短时记忆网络LSTM建模为例：

[0049] BILSTM-CRF是基于深度学习的NER识别方法，如图2所示，主要由输入层、BILSTM层、CRF层和输出层构成。

[0050] 每层的功能如下：

[0051] 步骤21:输入层：

[0052] 第一层是输入层，输入档案文本分词之后产出一组词，每个词表示成一个向量，使用预训练好的word2vec模型，从模型中查询每个词的向量，传入下一层。

[0053] 步骤22:BILSTM层：

[0054] 第二层是双向LSTM层，自动提取输入的特征，包括两个LSTM，一个正向输入序列，

一个反向输入序列,使模型能够同时考虑前向过程提取的特征和后向过程提取的特征,即考虑到过去的特征和未来的特征,将输入的词向量序列(w0, w1,w2,…wn)作为双向LSTM各个时间步的输入,将正向LSTM输出的隐状态序列(h1,h2,h3,…hn)和反向LSTM的隐状态序列(r1,r2,r3…..rn)在各个位置的输出按照顺序进行拼接,得到完整的隐含状态序列;

[0055] 再接入一个线性层,将隐含状态序列向量进行降维,降到k维,k是档案语料集合的标签数,进而得到自动提取的句子特征,记为C=(c1,c2,c3…ck)。每一项是词分类到第j个标签的打分值,如果后面进行softmax,相当于对各个位置独立进行k分类,这样各个位置的标注无法利用已经标注过的信息,所以将该结果输入到CRF层进行标注;

[0056] 步骤23:CRF层

[0057] 第三层是CRF层,对上层结果进行句子级别的序列标注,通过BILSTM层学习到了上下文信息,将结果经过一个隐含层输出到CRF,CRF层输入每个词属于每个每类的打分值,通过序列标注,选择预测得分最高的序列作为最佳答案;

[0058] CRF层的参数是(k+2)×(k+2)的矩阵,矩阵的每一项表示从第i个标签到第j个标签的转移分值,k+2的含义是在句首添加了一个起始状态,在句尾添加了一个终止状态,如果一个句子的标签序列为$y=(y_1,y_2,....,y_n)$,模型预测的标签打分结果为

[0059] $$\mathrm{score}(x,y)=\sum_{i=1}^{n}P_{i,y_i}+\sum_{i=1}^{n+1}A_{y_{i-1}y_i}$$

[0060] 整个序列的打分结果是各个位置的打分结果累加和,使用softmax进行归一化的概率为:

[0061] $$P(y|x)=\frac{\exp(score(x,y))}{\sum_{y'}\exp(score(x,y'))}$$

[0062] 模型预测时,使用动态规划的维特比(Viterbi)算法求解最优路径

[0063] $$y^{*}=arg_{y'}^{maxscore(x,y')}$$

[0064] 最优求解之后,模型预测结果为<字,标签>,如

[0065] 王B-PERSON

[0066] 贺I-PERSON

[0067] 毕O

[0068] 业O

[0069] 于O

[0070] 同B-ORG

[0071] 济I-ORG

[0072] 大I-ORG

[0073] 学I-ORG

[0074] 步骤3:档案关系抽取;

[0075] 档案关系抽取是抽取两个实体之间的关联关系,构建档案知识图谱的边,一般时三元组<主体,谓词,客体>,即SPO结构;

[0076] 档案关系抽取方法主要采用基于模式或规则的抽取方法、基于序列标注的监督学习方法和基于文本分类的监督学习方法，我们以基于文本分类的监督学习方法为例；

[0077] 使用基于Attention机制的双向LSTM神经网络模型进行档案关系抽取，Attention机制能够自动发现对档案关系分类重要的词，使模型可以从句子中捕获最重要的语义信息。

[0078] 如附图3所示，主要包括输入层、BILSTM层、Attention层和输出层四部分，每层的功能如下：

[0079] 步骤31：输入层

[0080] 第一层是输入层，对输入的档案文本进行句子内容解析，产出一组词，将词进行词嵌入，产出向量，并传递给模型下一层。

[0081] 步骤32：BILSTM层：

[0082] 双向LSTM包含两个LSTM网络，可以正向和反向学习上下文信息，最终输出的当前隐藏状态是由当前细胞状态和输出门权重矩阵，第i个词的输出如下所示；

[0083] $h_i = [\overrightarrow{h_i} \oplus \overleftarrow{h_i}]$

[0084] 将正向和反向的结果进行加和产出该层的输出；

[0085] 步骤33：Attention层

[0086] 生成一个权重向量，通过与这个权重向量相乘，使每一次迭代中的词汇级的特征合并为句子级的特征。

[0087] 将LSTM层的输入向量表示为$h:[h_1, h_2, \ldots h_N]$，Attention层的权重矩阵由下面公式得到：

[0088] $M = \tanh(H)$

[0089] $a = \mathrm{softmax}(w^N M)$

[0090] $r = Ha^N$

[0091] 其中，$H \in R^{d^w \times N}$，d^w，d^w是词向量的维度，w是训练过程中的变量矩阵，w^N是训练学习到的参数向量的转置矩阵，我们通过下面公式计算句子的最终分类。

[0092] $h^* = \tanh(r)$

[0093] 使用attention机制，encoder层的每一步输出都会和当前的输出进行联立计算，后面用softmx函数生成概率值。通过保留LSTM编码器对输入序列的中间输出结果，然后训练模型来对这些输入进行选择性的学习并且在模型输出时将输出序列与之进行关联。

[0094] 步骤34：输出层

[0095] 将Attention层产出的向量送入softmax分类器来预测标签值，我们选择最大的概率的标签作为预测标签，最终产出三元组；

[0096] 为了避免过拟合，我们在网络前向计算时使用了dropout。Droupout是指在前向传播计算时，让某个神经元的激活值以一定的概率p停止工作，这样两个神经元不一定每次都在一个dropout网络中出现，权值的更新不再依赖固有关系的隐含结点，迫使网络学习更加鲁棒的特征，模型泛化能力更强。

[0097] 步骤4：档案知识融合

[0098] 将抽取的不同来源的档案实体进行融合，产出一个唯一的属性全面的档案实体；

[0099] 档案知识融合，主要分为两类，一类是基于多个档案知识图谱的融合，每组档案数据源单独构建一个档案知识图谱，再将多个档案知识图谱融合；第二类是基于多个不同数据源的融合，档案知识通过解析每个数据源得到，再将所有档案知识合并成一个知识图谱。

[0100] 我们以多个不同数据源的融合为例，将档案知识融合过程分为档案实体链接和档案知识合并两部分。档案实体链接时将抽取到的档案实体链接到已有的档案实体对象上，档案知识合并是将新链接的档案实体融入到已有的档案实体中。

[0101] 步骤41：档案实体链接

[0102] 档案实体链接是从档案文本中抽取得到档案实体对象，将其链接到档案知识库中已有的档案实体对象上，处理流程为：

[0103] 从档案文本中抽取档案实体对象；

[0104] 进行指代消解、实体消歧操作，判断档案知识库中的同名档案实体和当前抽取的档案实体的含义是否一致；

[0105] 如果和档案知识库中的档案实体一致，则将抽取的档案实体链接到档案知识库中的档案实体。

[0106] 指代消解，解决多个名称对应同一档案实体对象的问题，利用指代消解，可以将其关联到正确的档案实体上。

[0107] 实体消歧，是解决同名档案实体产生歧义的问题。假设两个档案实体的属性记录为x和y，在第i个属性上的值是x_i和y_i，那么通过计算累加单个属性的相似度可以得到档案实体的相似度。

[0108] $[sim(x_1,y_1),sim(x_2,y_2),\cdots,sim(x_n,y_n)]$

[0109] 我们使用余弦夹相似性计算属性相似度。

[0110] $$cos\theta = \frac{\sum_{i=1}^{n} x_i \times y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \times \sqrt{\sum_{i=1}^{n} y_i^2}}$$

[0111] 产出的夹角余弦值越大，两个档案实体的相似度越高；

[0112] 如我们抽取了一个档案人物实体，包含姓名、身份证号、手机号、年龄信息，首先去档案知识库中检索和该名一致的档案实体列表，通过使用夹角余弦计算相似度，选择相似值最大且大于阈值的档案实体，将其链接到该档案实体上，否则，是一个新档案实体。

[0113] 步骤42：档案知识合并

[0114] 将链接到一起的新抽取的档案实体和档案知识库中的档案实体进行合并，将档案知识图谱构建成一个更完整的图谱，如将档案人物实体链接在一起的实体属性进行融合，将抽取到的新属性、描述更完善的属性，添加到档案知识库的档案实体属性中。

[0115] 步骤5：档案知识存储；

[0116] 对抽取的档案实体和关系数据进行存储，档案知识图谱数据存储采用两种方式：基于关系模型的存储方式和基于图模型的存储方式，我们以基于图模型的存储方式为例，将知识数据存储在图数据库中。

[0117] 我们以使用Neo4j图数据库存储抽取的档案实体和关系数据为例，Neo4j是一个原生的图数据库引擎，它有独特的存储结构免索引邻居节点存储方法，且有相应的图遍历算

法,具有非常高的查询性能;图数据结构自然伸展特性及其非结构化的数据格式,让Neo4j的数据库设计可以具有很大的伸缩性和灵活性。

[0118] 步骤6:档案研判分析;

[0119] 依托机器学习技术、自然语言处理技术和构建的档案知识图谱,对海量档案数据进行挖掘,针对不同编研用户的信息需求,利用文本信息抽取、文本分类、文本聚类、自动摘要抽取、档案人物关系和智能检索等智能分析技术,自动化提取档案文本关键信息,产生高价值的档案信息,有效提升编研人员工作效率,为编研人员编写高质量编研报告提供数据和工具支撑。

[0120] 步骤61:档案热词分析

[0121] 我们提供一种档案热词分析的方法,以档案的文本文件数据为基础,通过对档案全文数据挖掘和分析,产出当前输入档案数据对应的热词列表,并以列表或者词云形式展现。

[0122] 档案热词分析使用tf-idf方法实现。

[0123] tf公式如下所示。

[0124] $$tf_{ij} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

[0125] 其中,分子表示输入档案文本中某一个词出现的次数,分母表示输入档案文本中所有词的数量。

[0126] idf公式如下所示。

[0127] $$idf_i = log\frac{|D|}{1 + |\{d \in D: t \in d\}|}$$

[0128] 其中,分子表示输入档案文本中文档总数,分母表示包含词的文档数量,如果词语不在档案文本中,就会导致分母为零,所以我们给分母加1。

[0129] 计算tf与idf的乘积的公式如下所示。

[0130] $tfidf_{i,j} = t_{fi,j} \times id_{fi}$

[0131] 如果一个词的词频很高,且在其它文档很少出现,则这个词是一个区分度比较高的词,依次计算每个词,最终产出topN的热词列表,可以使用词云呈现给用户;

[0132] 步骤62:档案关键词抽取

[0133] 对输入的一段档案文本,抽取能标识该档案文本的一组词,广泛用于档案关键词标引和档案信息检索中,我们采用的关键词抽取包括两种方法,tf-idf 和textrank。

[0134] 1.tf-idf方法

[0135] tf-idf是常用的对句子中词语打分的方法,一个词的tf-idf值取决于两个因素:词频和该词的重要程度。通过计算产出top词语构成关键词集合。

[0136] 2.textrank方法

[0137] textrank方法源于page-rank,它认为文档或句子中相邻词的重要性是相互影响的,引入了词的顺序信息。

[0138] $$A(V_i) = (1-d) + d * \sum_{j \in In(V_i)} \frac{1}{|Out(V_j)|} S(V_j)$$

[0139] 其中，V_i标识要计算权重的词，$S(V_j)$表示词的权重，d是阻尼系数，$In(V_i)$表示和V_i在同一窗口的词集合，$Out(V_j)$表示与V_j在同一窗口的词的集合，加上绝对值表示词集合的个数；

[0140] Textrank会先对每个词的权重进行初始化，然后根据公式进行更新直到收敛，筛选出的词集合最能反应整个档案文档或句子。

[0141] 步骤63：档案自动分类

[0142] 提供档案自动分类的功能，可以对未知分类的编研档案自动分类，提供基于规则的自动分类和基于机器学习的自动分类两种方法。

[0143] 1.基于规则的自动分类

[0144] 根据实际分析场景，由用户提供如下形式的规则：每一种分类使用哪些关键词来描述，以及各个关键词的权重是什么，系统根据用户设置的规则，对未知分类的档案进行自动分类，规则文件包括分类类别及该类别下对应的词列表和权重，输入的档案数据进行分词处理及去停用词之后，依次与规则文件的每一类下的词列表进行位置关联并累积加权计算权重，最终给出整段输入数据所属的类别及概率，用户可以指定返回topN类概率最高的类别及概率。

[0145] 2.基于机器学习的自动分类

[0146] 用户提供训练语料，系统利用机器学习算法自动学习语料中的分类规律，建立档案自动分类模型，实现对未知分类档案的自动分类，提供一系列的机器学习档案分类方法，如贝叶斯分类基于贝叶斯定理，在已知某条件下的概率，计算两条件交换后的概率；贝叶斯分类是采用属性条件独立性假设的条件下的生成模型分类；支持向量机是通过核函数将低位空间线性不可分的样本映射到高维线性可分空间的方法。

[0147] 提供机器学习分类模型训练功能，输入指定的档案主题数据集，并制定好分类标签，进行分类模型训练。

[0148] 提供机器学习分类模型评估功能，输入指定的主题数据集(打好分类标签)，生成模型评估结果。

[0149] 分类模型的效果评估使用准确率、召回率和$F_{1\text{-score}}$，如下所示：

[0150] $$精确率=\frac{tp}{tp+fp}$$

[0151] $$召回率=\frac{tp}{tp+fn}$$

[0152] $$F_{1-score=2} \cdot \frac{准确率 \cdot 召回率}{准确率 + 召回率}$$

[0153] 其中，tp：预测是正确的正样本；fp：预测是错误的正样本；fn：预测是错误的负样本；

[0154] 分类模型的效果评估达标之后，可以对档案文本进行自动打分类标签，支持标签分布统计；

[0155] 步骤64:档案人物关系分析

[0156] 在档案实体抽取中,我们已经抽取了人、机构和事件等档案实体信息,在档案人物关系分析中,输入待分析的目标人物,从档案知识图谱中查询出与该人相关联的事件、机构,并通过语义网络的形式展现给用户。

[0157] 步骤7:档案智能检索。

[0158] 搜索是用户发起的信息检索的行为,用户提交查询请求,在系统收到查询请求后,查找匹配的内容,查询结果通过排序后,返回给用户,基于知识图谱的档案智能检索,在传统全文检索的基础之上,通过语义关联推荐,能够帮助用户扩展阅读,提升信息查找与分析的深度和广度。

[0159] 如附图4所示,基于档案知识图谱的智能搜索主要包括如下几步:

[0160] 1、用户意图分析:从用户提交的查询中识别出用户的目标实体,并为查找生成目标实体的查询条件;

[0161] 2、目标查询:对目标实体的查询条件,使用查询语句在知识图谱中查找目标实体及与其关联的相关内容;

[0162] 3、结果展现:如果目标实体不唯一,则需要对结果进行排序;

[0163] 4、目标实体探索:产出目标实体后,对与目标实体具有关联关系的相关实体展现到扩展的搜索结果处。

[0164] 具体的,如附图5所示,所述档案研判系统包括:档案上传模块、档案实体抽取模块、档案关系抽取模块、档案知识融合模块、档案知识存储模块、档案研判分析模块,所述的档案上传模块用于上传采集档案文档,并利用自然语言处理技术实现档案文档标题、作者、关键字、分类、摘要这些元数据进行自动标引;所述的档案实体抽取模块用于从档案中自动抽取人名、地名、机构名、事件、建筑、资料这些档案实体;所述的档案关系抽取模块用于从档案中抽取档案实体之间的关联关系从而形成档案知识图谱;所述档案知识融合模块用于对从各个档案文献中抽取的档案实体属性进行歧义消除和指代消解,实现档案实体属性的合并与融合;所述档案知识存储模块用于将抽取构建的档案知识图谱进行存储,并提供档案知识图谱的数据查询;所述档案研判分析模块用于给档案研判分析人员提供档案热词分析、档案敏感词分析、档案专有词识别、档案自动分类、档案聚类、档案人物关系抽取等智能化工具,同时提供档案智能检索,将档案知识图谱和档案全文检索进行整合,实现档案检索关键词、档案实体、相关档案的精准推荐。

[0165] 利用本发明所述的技术方案,或本领域的技术人员在本发明技术方案的启发下,设计出类似的技术方案,而达到上述技术效果的,均是落入本发明的保护范围。

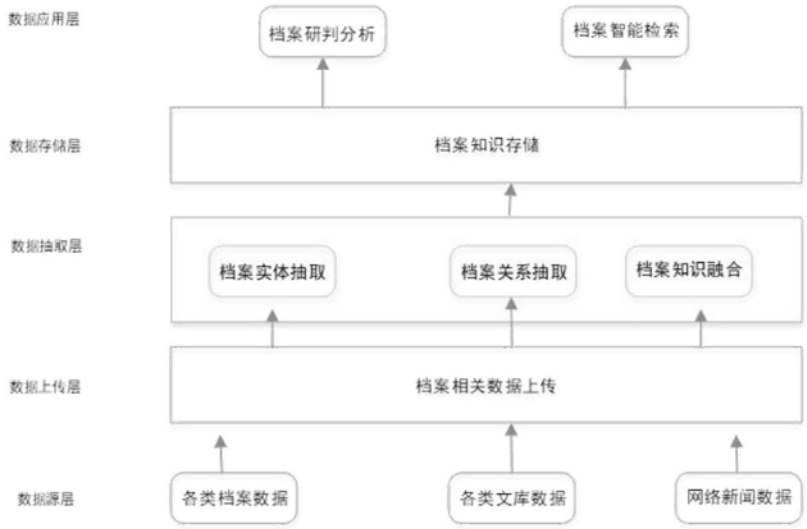

数据应用层
档案研判分析
档案智能检索
数据存储层
档案知识存储
数据抽取层
档案实体抽取
档案关系抽取
档案知识融合
数据上传层
档案相关数据上传
数据源层
各类档案数据
各类文库数据
网络新闻数据

图1

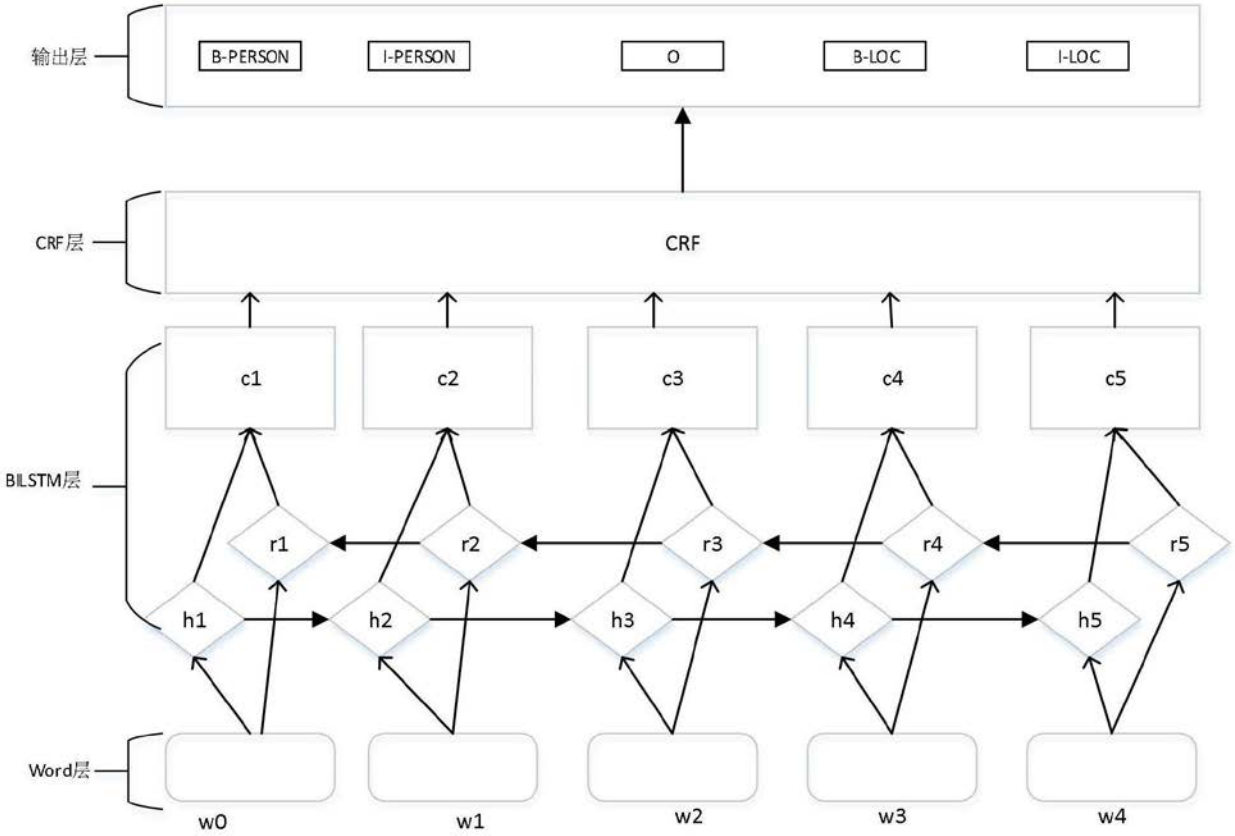

输出层
B-PERSON
I-PERSON
O
B-LOC
I-LOC
CRF层
CRF
c1
c2
c3
c4
c5
BILSTM层
r1
r2
r3
r4
r5
h1
h2
h3
h4
h5
Word层
w0
w1
w2
w3
w4

图2

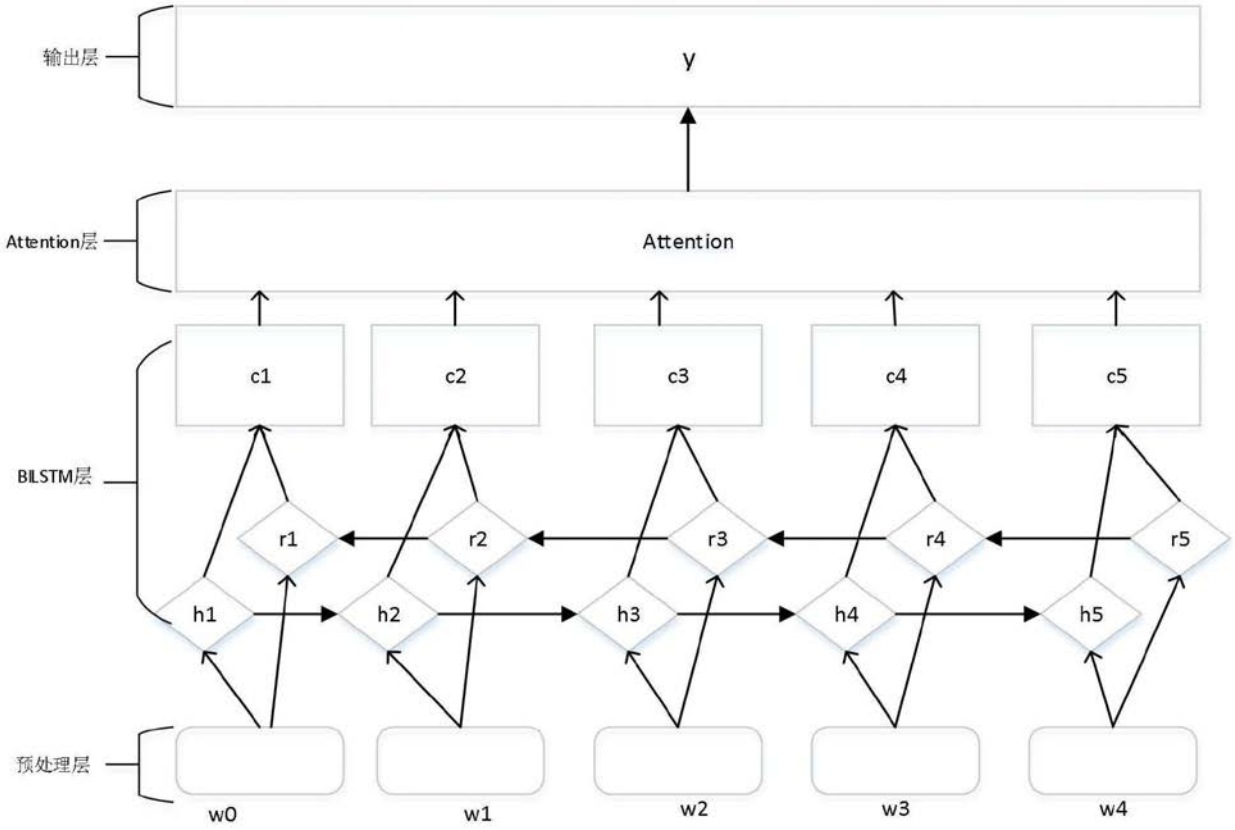

输出层
y
Attention层
Attention
c1
c2
c3
c4
c5
BILSTM层
r1
r2
r3
r4
r5
h1
h2
h3
h4
h5
预处理层
w0
w1
w2
w3
w4

图3


搜索意图分析
目标查询
结果展现
目标探索
词法分析
语法分析
语义理解
图数据库查询
结果排序
结果布局
实体信息
相关实体

图4

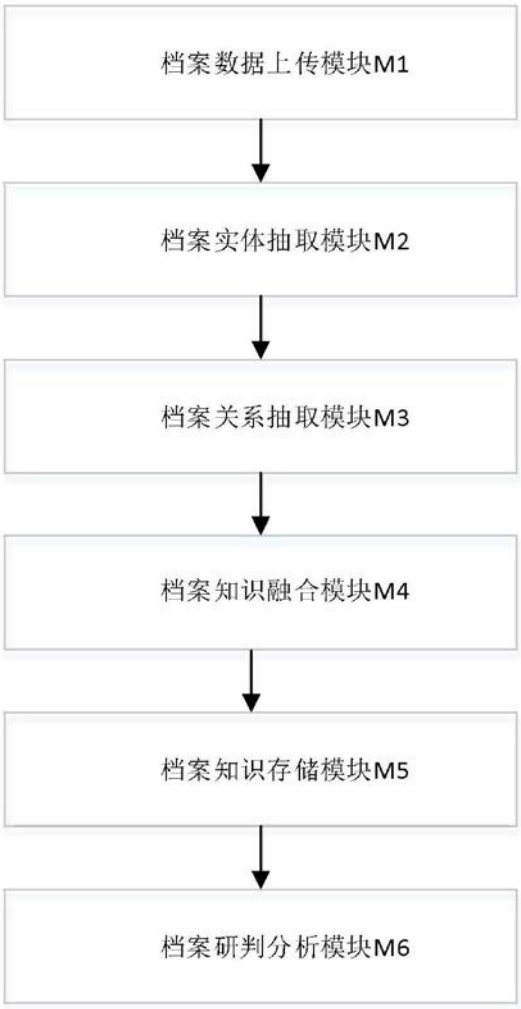

档案数据上传模块M1
档案实体抽取模块M2
档案关系抽取模块M3
档案知识融合模块M4
档案知识存储模块M5
档案研判分析模块M6

图5