(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 116662521 A
(43)申请公布日 2023.08.29

(21)申请号 202310920071.2

(22)申请日 2023.07.26

(71)申请人 广东省建设工程质量安全检测总站有限公司
地址 510000 广东省广州市天河区先烈东路121号之一第三层、第四层、第五层、第九层

(72)发明人 单良 王亚平 路阳 江伟欢 刘伟家 郑楠

(74)专利代理机构 广州渣津专利代理事务所(特殊普通合伙) 44516
专利代理师 申宏辉

(51)Int.Cl.
G06F 16/332(2019.01)
G06F 40/216(2020.01)
G06F 40/284(2020.01)
G06F 40/30(2020.01)
G06F 16/33(2019.01)

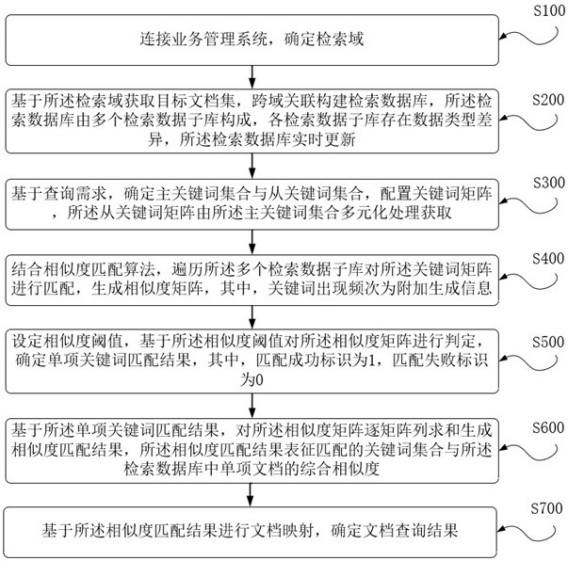
权利要求书3页 说明书11页 附图4页

(54)发明名称

一种电子文档筛选查询方法及系统

(57)摘要

一种电子文档筛选查询方法及系统，属于信息检索领域，方法包括：连接业务管理系统，确定检索域；基于检索域获取目标文档集，跨域关联构建检索数据库；基于查询需求确定主关键词与从关键词集合，配置关键词矩阵；遍历多个检索数据子库对关键词矩阵进行匹配，生成相似度矩阵；设定相似度阈值并对相似度矩阵进行判定，确定单项关键词匹配结果；基于单项关键词匹配结果，对相似度矩阵逐矩阵列求和生成相似度匹配结果；基于相似度匹配结果进行文档映射，确定文档查询结果。本申请解决了现有技术中电子文档筛选查询准确度和效率低的技术问题，实现了电子文档的高精度、动态、多元化查询，达到了提高电子文档筛选准确度和效率的技术效果。



CN 116662521 A

1.一种电子文档筛选查询方法,其特征在于,所述方法包括:

连接业务管理系统,确定检索域;

基于所述检索域获取目标文档集,跨域关联构建检索数据库,所述检索数据库由多个检索数据子库构成,各检索数据子库存在数据类型差异,所述检索数据库实时更新;

基于查询需求,确定主关键词集合与从关键词集合,配置关键词矩阵,所述从关键词集合由所述主关键词集合多元化处理获取;

结合相似度匹配算法,遍历所述多个检索数据子库对所述关键词矩阵进行匹配,生成相似度矩阵,其中,关键词出现频次为附加生成信息;

设定相似度阈值,基于所述相似度阈值对所述相似度矩阵进行判定,确定单项关键词匹配结果,其中,匹配成功标识为1,匹配失败标识为0;

基于所述单项关键词匹配结果,对所述相似度矩阵逐矩阵列求和生成相似度匹配结果,所述相似度匹配结果表征匹配的关键词集合与所述检索数据库中单项文档的综合相似度;

基于所述相似度匹配结果进行文档映射,确定文档查询结果。

2.如权利要求1所述的方法,其特征在于,所述配置关键词矩阵,方法包括:

基于所述查询需求,提炼多个主关键词,作为所述主关键词集合;

配置多元化处理调幅;

基于所述多元化处理调幅,对所述主关键词集合进行上位化处理,确定第一从属关键词集合;

基于所述多元化处理调幅,对所述主关键词集合进行下位化处理,确定第二从属关键词集合;

对所述主关键词集合进行转换处理,确定第三从属关键词集合;

基于所述第一从属关键词集合、所述第二从属关键词集合与所述第三从属关键词集合,确定从关键词集合,所述从关键词集合带有主相关度标识;

将关键词序列作为矩阵行,将关键词类目作为矩阵列,基于所述主关键词集合与所述从关键词集合搭建所述关键词矩阵。

3.如权利要求2所述的方法,其特征在于,所述生成相似度矩阵,方法包括:

基于所述关键词矩阵,提取所述主关键词集合;

遍历所述多个检索数据子库,对所述基于所述主关键词集合进行相似度匹配,确定一项相似度矩阵;

若所述一项相似度矩阵为空,提取所述从关键词集合并遍历所述多个检索数据子库进行相似度匹配,确定二项相似度矩阵;

若所述二项相似度矩阵为空,基于所述主关键词集合,遍历所述多个检索数据子库进行语义识别,获取三项相似度矩阵。

4.如权利要求3所述的方法,其特征在于,获取相似度矩阵计算公式,方法包括:

$$S(A,B)=\alpha*\beta\frac{A*B}{|A||B|}=\begin{vmatrix} S_{11} & S_{12} & \cdots & S_{1N} \\ S_{21} & S_{22} & \cdots & S_{2N} \\ \cdots & \cdots & \cdots & \cdots \\ S_{M1} & S_{M2} & \cdots & S_{MN} \end{vmatrix};$$

其中，α 表征关键词出现频次，β 表征主相关度，A为待进行匹配分析的关键词矩阵，为列矩阵，B为所述检索数据库中的文档矩阵，为行矩阵，S_{MN}表示第M个关键词与第N项文档的相似度，M、N为量值，表征关键词项数与文档项数，针对所述一项相似度矩阵，β =1。

5.如权利要求4所述的方法，其特征在于，获取矩阵列求和公式，方法包括：

$$S_j = \sum_{i=1}^{M} S_{ij};$$

其中，S_j为M个关键词与第j项文档的相似度匹配结果，S_{ij}表示第i个关键词与第j项文档的相似度，$i \in M$，$j \in N$。

6.如权利要求1所述的方法，其特征在于，于所述多个检索数据子库中对所述关键词矩阵进行匹配，之前，方法包括：

配置多元数据处理规则；

基于所述数据处理规则，对所述多个检索数据子库执行规则匹配与数据预处理，确定预处理数据库；

基于所述预处理数据库，进行所述关键词矩阵的匹配执行。

7.如权利要求1所述的方法，其特征在于，所述确定文档查询结果，方法包括：

对所述相似度匹配结果进行正序列化调整，生成相似度序列，所述相似度序列由大到小排列；

获取查询需求项数；

基于所述查询需求项数对所述相似度序列进行截取，反向匹配映射文档，集成作为查询文档集合；

基于所述查询文档集，确定所述文档查询结果，所述文档查询结果具有文档优先级。

8.一种电子文档筛选查询系统，其特征在于，所述系统包括：

确定检索域模块，所述确定检索域模块用于连接业务管理系统，确定检索域；

检索数据库构建模块，所述检索数据库构建模块基于所述检索域获取目标文档集，跨域关联构建检索数据库，所述检索数据库由多个检索数据子库构成，各检索数据子库存在数据类型差异，所述检索数据库实时更新；

关键词矩阵模块，所述关键词矩阵模块用于基于查询需求，确定主关键词集合与从关键词集合，配置关键词矩阵，所述从关键词集合由所述主关键词集合多元化处理获取；

相似度矩阵模块，所述相似度矩阵模块用于结合相似度匹配算法，遍历所述多个检索数据子库对所述关键词矩阵进行匹配，生成相似度矩阵，其中，关键词出现频次为附加生成信息；

关键词匹配结果模块，所述关键词匹配结果模块用于设定相似度阈值，基于所述相似度阈值对所述相似度矩阵进行判定，确定单项关键词匹配结果，其中，匹配成功标识为1，匹配失败标识为0；

相似度匹配结果模块，所述相似度匹配结果模块基于所述单项关键词匹配结果，对所述相似度矩阵逐矩阵列求和生成相似度匹配结果，所述相似度匹配结果表征匹配的关键词集合与所述检索数据库中单项文档的综合相似度；

文档查询结果模块，所述文档查询结果模块基于所述相似度匹配结果进行文档映射，确定文档查询结果。

一种电子文档筛选查询方法及系统

技术领域

[0001] 本发明涉及信息检索领域,具体涉及一种电子文档筛选查询方法及系统。

背景技术

[0002] 当前,随着信息技术的发展,各个组织和企业积累了大量的电子文档数据。这些数据通常分布在不同的业务系统中,存在跨域和异构的特征。目前,实现大规模文档检索的方法主要有:构建集中式索引,采用爬虫技术索引;利用元数据等手段提取文档特征并基于特征索引实现搜索。但是,这些方法在跨域异构数据场景下效果不佳,效率和查询准确度低下。

发明内容

[0003] 本申请通过提供了一种电子文档筛选查询方法及系统,旨在解决现有技术中电子文档筛选查询准确度和效率低的技术问题。

[0004] 鉴于上述问题,本申请提供了一种电子文档筛选查询方法及系统。

[0005] 本申请公开的第一个方面,提供了一种电子文档筛选查询方法,该方法包括:连接业务管理系统,确定检索域;基于检索域获取目标文档集,跨域关联构建检索数据库,检索数据库由多个检索数据子库构成,各检索数据子库存在数据类型差异,检索数据库实时更新;基于查询需求,确定主关键词集合与从关键词集合,配置关键词矩阵,从关键词集合由主关键词集合多元化处理获取;结合相似度匹配算法,遍历多个检索数据子库对关键词矩阵进行匹配,生成相似度矩阵,其中,关键词出现频次为附加生成信息;设定相似度阈值,基于相似度阈值对相似度矩阵进行判定,确定单项关键词匹配结果,其中,匹配成功标识为1,匹配失败标识为0;基于单项关键词匹配结果,对相似度矩阵逐矩阵列求和生成相似度匹配结果,相似度匹配结果表征匹配的关键词集合与检索数据库中单项文档的综合相似度;基于相似度匹配结果进行文档映射,确定文档查询结果。

[0006] 本申请公开的另一个方面,提供了一种电子文档筛选查询系统,该系统包括:确定检索域模块,用于连接业务管理系统,确定检索域;检索数据库构建模块,基于检索域获取目标文档集,跨域关联构建检索数据库,检索数据库由多个检索数据子库构成,各检索数据子库存在数据类型差异,检索数据库实时更新;关键词矩阵模块,基于查询需求,确定主关键词集合与从关键词集合,配置关键词矩阵,从关键词集合由主关键词集合多元化处理获取;相似度矩阵模块,用于结合相似度匹配算法,遍历多个检索数据子库对关键词矩阵进行匹配,生成相似度矩阵,其中,关键词出现频次为附加生成信息;关键词匹配结果模块,用于设定相似度阈值,基于相似度阈值对相似度矩阵进行判定,确定单项关键词匹配结果,其中,匹配成功标识为1,匹配失败标识为0;相似度匹配结果模块,基于单项关键词匹配结果,对相似度矩阵逐矩阵列求和生成相似度匹配结果,相似度匹配结果表征匹配的关键词集合与检索数据库中单项文档的综合相似度;文档查询结果模块,基于相似度匹配结果进行文档映射,确定文档查询结果。

[0007] 本申请中提供的一个或多个技术方案,至少具有如下技术效果或优点:

由于采用了通过连接业务管理系统确定检索域,为后续构建检索数据库和查询提供数据基础;基于检索域获取目标文档集,跨域关联构建检索数据库,实现大规模文档的统一检索平台,从而提高检索效率;基于查询需求获得丰富的主关键词和从关键词,支持多元化查询需求;结合相似度匹配算法,遍历多个检索数据子库对关键词矩阵进行匹配,生成相似度矩阵,结合关键词出现频次,实现高精度匹配;设定相似度阈值,基于相似度阈值对相似度矩阵进行判定,确定单项关键词匹配结果,只保留高度相关的匹配,提高查询准确度;基于单项关键词匹配结果,对相似度矩阵逐矩阵列求和生成相似度匹配结果,从而减少计算量,提高效率;基于相似度匹配结果进行文档映射,确定文档查询结果,确定高精度的查询结果的技术方案,解决了现有技术中电子文档筛选准确度和效率低的技术问题,实现了电子文档的高精度、动态、多元化查询,达到了提高电子文档筛选准确度和效率的技术效果。

[0008] 上述说明仅是本申请技术方案的概述,为了能够更清楚了解本申请的技术手段,而可依照说明书的内容予以实施,并且为了让本申请的上述和其它目的、特征和优点能够更明显易懂,以下特举本申请的具体实施方式。

附图说明

[0009] 图1为本申请实施例提供了一种电子文档筛选查询方法可能的流程示意图;

图2为本申请实施例提供了一种电子文档筛选查询方法中获取相似度矩阵可能的流程示意图;

图3为本申请实施例提供了一种电子文档筛选查询方法中确定文档查询结果可能的流程示意图;

图4为本申请实施例提供了一种电子文档筛选查询系统可能的结构示意图。

[0010] 附图标记说明:确定检索域模块11,检索数据库构建模块12,关键词矩阵模块13,相似度矩阵模块14,关键词匹配结果模块15,相似度匹配结果模块16,文档查询结果模块17。

具体实施方式

[0011] 本申请提供的技术方案总体思路如下:

本申请实施例提供了一种电子文档筛选查询方法及系统,通过构建跨域检索数据库,采用关键词矩阵和相似度算法逐步实现匹配、生成相似度矩阵和查询结果,设置阈值和优先级提高精度与效率,最终达到电子文档高精度、动态、多元化查询和优化文档筛选的技术效果。

[0012] 在介绍了本申请基本原理后,下面将结合说明书附图来具体介绍本申请的各种非限制性的实施方式。

实施例一

[0013] 如图1所示,本申请实施例提供了一种电子文档筛选查询方法,该方法包括:

步骤S100:连接业务管理系统,确定检索域;

具体而言，业务管理系统是指存储和管理电子文档的系统平台。检索域是对电子文档检索的范围，根据查询需求进行设定，如输入需要检索的文档ID段、路径关键字、元数据属性等。连接不同的业务管理系统，如OA系统、ERP系统、CRM系统等，通过系统接口等技术手段获取业务管理系统中的电子文档数据范围信息，确定需要检索和查询的电子文档集合，为后续步骤提供数据源和查询范围基础，系统接口技术包括Web服务、远程调用等。通过连接业务关系系统并确定检索域，实现了电子文档资源获取并确定查询目标，为构建跨域检索数据库和实现高效查询提供了基础。

[0014] 步骤S200：基于所述检索域获取目标文档集，跨域关联构建检索数据库，所述检索数据库由多个检索数据子库构成，各检索数据子库存在数据类型差异，所述检索数据库实时更新；

具体而言，基于确定的检索域，通过与业务系统接口对接或目录浏览等方式获取目标文档集，目标文档集是检索域内需要检索和查询的全部电子文档的集合。跨域关联构建检索数据库是指将不同业务系统和数据源的目标文档集进行关联，构建统一的检索数据库平台。该数据库包括来自不同域的多个检索数据子库，各子库之间存在数据类型差异，如文档格式、元数据表达方式的差异。

[0015] 检索数据库采用文档管理、文档索引引擎等技术构建。需要对目标文档集进行格式转换、元数据提取等预处理，使其在同一检索平台下能被高效查询。检索数据库实时更新是指数据库中的数据随业务系统更新而更新，以保证查询结果的及时性。其中，各检索数据库子库按业务系统、文档类型等建立，在子库构建时采用统一的文档解析、索引建立方式，并支持增量式更新，以及时同步业务系统的文档变更。

[0016] 通过构建跨域关联的检索数据库，实现了异构电子文档的集中管理和统一检索，克服了数据孤岛问题，使数据库逻辑上成为一个整体，增量更新保证了查询结果的实时性，为后续的智能推荐和高效查询提供了技术基础。

[0017] 步骤S300：基于查询需求，确定主关键词集合与从关键词集合，配置关键词矩阵，所述从关键词集合由所述主关键词集合多元化处理获取；

具体而言，基于查询者的具体查询需求，采用查询分析等技术解析查询条件，获得主要概念词和关键词作为主关键词，基于主关键词采用扩展词汇表、语义网络等生成从关键词。主关键词集合表示查询需求的主题和重点，从关键词表示查询需求的相关词或语义词。配置关键词矩阵是指构建二维矩阵，矩阵行表示主关键词，矩阵列表示从关键词，矩阵值可以使用共现次数或相关度进行初始化。矩阵各单元记录主关键词和从关键词在目标文档集中的出现频次或相关度，选择高度相关的主关键词和从关键词，以提高后续的匹配精度。多元化处理是指综合考虑查询需求的多方面语义和相关词，生成丰富的从关键词集合，如基于主关键词扩展词汇表、利用语义识别技术提取相关词、通过运算衍生新的相关词汇作为从关键词等。

[0018] 通过解析查询需求得到主关键词，多元化处理获取从关键词，关键词矩阵以简洁的形态表达主关键词和从关键词间的关联强度，为关键词与文档的相关性判断提供参考依据，实现对查询意图的准确理解和丰富表达，为后续文档智能匹配奠定基础，从而实现精确查询。

[0019] 步骤S400：结合相似度匹配算法，遍历所述多个检索数据子库对所述关键词矩阵

进行匹配，生成相似度矩阵，其中，关键词出现频次为附加生成信息；

具体而言，用相似度匹配算法，在构建的检索数据库的每个检索数据子库内，匹配关键词矩阵的每个主关键词和从关键词，生成相似度矩阵。其中，相似度匹配算法采用向量空间模型、词袋模型等，计算关键词矩阵的主从关键词与每个子库文档之间的相似度；遍历子库是指逐个在每个子库进行匹配运算，由于子库的数据类型不同，需要对不同类型数据选择适用的相似度算法。相似度矩阵是指通过检索得到一个二维矩阵，矩阵行表示主从关键词，矩阵列表示目标文档集的文档，矩阵各单元记录主从关键词与文档的相似度值，表示匹配程度。关键词出现频次表示关键词在文档中的重要性，为后续结果判断提供参考，作为附加信息一并保存。例如，选取主从关键词和文档向量空间模型来计算相似度，可以采用不同子库分别匹配的方式，也可以对文档预处理成统一向量空间后统一匹配，关键词出现频次可以作为相似度的权重因子。

[0020] 通过采用相似度算法实现了关键词矩阵与海量文档的智能匹配，相似度矩阵以数字形式准确表达主从关键词与每个文档的相关程度，为后续查询结果的生成提供支持。采用不同子库分别匹配，可针对其数据特征选择最优算法，提高效率。出现频次的保存为结果判断提供了参考，以提高筛选查询的准确性。

[0021] 步骤S500：设定相似度阈值，基于所述相似度阈值对所述相似度矩阵进行判定，确定单项关键词匹配结果，其中，匹配成功标识为1，匹配失败标识为0；

具体而言，相似度阈值是指主关键词、从关键词与文档匹配的相关度下限。通过分析查询日志、交互反馈、算法优化，根据查询的准确性要求和结果的召回率设定一个相似度阈值，以此作为判断标准。例如，通过数据分析技术分析大量历史查询日志，总结出相似查询条件下用户选择结果的相似度分布区间，选择区间的中上限值作为阈值；初始化一个相对较低的阈值，基于此获得匹配结果，由用户判断和反馈，逐步提高阈值直到用户满意为止；构建目标函数考虑查询精度和召回率，采用机器学习算法反复试算，寻找使目标函数达到最优的阈值。其中，阈值设置过高会导致匹配失败过多，结果准确但召回率低；阈值设置过低召回率高但结果不太准确。

[0022] 获取相似度阈值后，对相似度矩阵的每个单元进行判定，得到单项关键词匹配结果。判定相似度矩阵是指逐个判断矩阵每个单元的相似度值与阈值的大小关系，低于该阈值的匹配被判定为失败，不予考虑；高于或等于该阈值的匹配被判定为成功，对每个结果进行标识，成功标识为1，失败标识为0，从而得到单项关键词匹配结果，为主关键词、从关键词与每个文档的匹配结果，作为查询结果的候选项。

[0023] 通过设置相似度阈值，实现了对相似度矩阵匹配结果的精确判断。判定结果为后续查询结果生成的根据，并为后续的生成相似度结果提供了基础。

[0024] 步骤S600：基于所述单项关键词匹配结果，对所述相似度矩阵逐矩阵列求和生成相似度匹配结果，所述相似度匹配结果表征匹配的关键词集合与所述检索数据库中单项文档的综合相似度；

具体而言，基于得到的主从关键词与每个文档的单项关键词匹配结果，对相似度矩阵的每个矩阵列中的所有单元进行求和运算，生成相似度匹配结果。单项关键词匹配结果以1或0表示主从关键词与每个文档的匹配关系。相似度矩阵记录主从关键词与每个文档的相似度，矩阵列对应目标文档集中的每个文档。逐矩阵列求和是指依次对矩阵的每一列

中的所有单元进行求和，得到矩阵列中的总相似度。例如，可以对单项匹配结果为1的单元的相似度进行求和；也可以所有单元的相似度值进行求和，并乘以匹配成功的主从关键词比例作为权重。相似度匹配结果表示主从关键词集合与每个文档的综合相似度，由相似度矩阵每个矩阵列的总相似度组成，表达主从关键词与该文档的匹配程度。

[0025] 通过对相似度矩阵列中所有单元的相似度进行求和，生成主从关键词与每个文档的综合匹配相似度，以整体数字形式准确表达主从关键词与每个文档的综合相关程度，为后续结果的精确判断和排序提供依据，实现了匹配结果的积聚和综合评价，实现了匹配精度的整体提高。

[0026] 步骤S700：基于所述相似度匹配结果进行文档映射，确定文档查询结果。

[0027] 具体而言，相似度匹配结果记录主从关键词与每个文档的综合相似度，标识匹配的精确程度。文档映射是指根据相似度匹配结果对文档集重新进行排序或筛选，将与查询需求最为相匹配的文档选取出来，可以采用最近邻排序、相似度阈值筛选等方式实现映射，即可以根据相似度从大到小对文档集进行最近邻排序，选择相似度较高的前N个文档作为结果；也可以设定相似度选择阈值，选择相似度超过阈值的文档作为结果。文档查询结果是指从映射结果中选择与查询需求最为相关的前N个文档作为最终结果返回给用户，N的值可以由用户在查询时设定，或根据业务场景经验确定。

[0028] 通过对文档集进行重新映射，实现了基于关键词矩阵匹配结果的文档优先级划分和精确选择。选取的结果文档与查询需求最为匹配和相关，实现了电子文档的高精度、动态、多元化查询，达到了提高电子文档筛选准确度和检索效率的技术效果。

[0029] 进一步的，本申请实施例还包括：

步骤S310：基于所述查询需求，提炼多个主关键词，作为所述主关键词集合；

步骤S320：配置多元化处理调幅；

步骤S330：基于所述多元化处理调幅，对所述主关键词集合进行上位化处理，确定第一从属关键词集合；

步骤S330：基于所述多元化处理调幅，对所述主关键词集合进行下位化处理，确定第二从属关键词集合；

步骤S340：对所述主关键词集合进行转换处理，确定第三从属关键词集合；

步骤S350：基于所述第一从属关键词集合、所述第二从属关键词集合与所述第三从属关键词集合，确定从关键词集合，所述从关键词集合带有主相关度标识；

步骤S360：将关键词序列作为矩阵行，将关键词类目作为矩阵列，基于所述主关键词集合与所述从关键词集合搭建所述关键词矩阵。

[0030] 具体而言，通过解析查询需求，理解查询意图，提炼出表达查询主题和中心词的几个关键词作为主关键词集合。其中，主关键词集合表达查询需求的主要主题和语义，主关键词之间具有相关关联，共同表达查询意图。查询需求可以由用户输入的查询关键词、查询语句或问句确定。提炼主关键词需要对其进行分析和理解，抽取主题词和概念核心词，可以通过词频统计、词频标注、关键词抽取算法等进行提炼。

[0031] 配置多元化处理调幅是指设定一个比率或数值，作为生成从关键词集合的上限，实现主关键词集合的扩充调节。多元化处理调幅控制从关键词集合的生成规模，上调可以增加生成从关键词的数量和种类，下调则减少生成规模，根据具体应用场景和查询类型确

定。在词类或者概念层级上，基于主关键词集合推导和寻找其上位词或高层概念，对主关键词进行上位化处理扩充，生成第一从属关键词集合。在词类或者概念层级上，基于主关键词集合推导和寻找其下位词或低层概念，对主关键词进行下位化处理扩充，生成第二从属关键词集合。基于主关键词的词性转换、同义词替换、相关词扩充等，对主关键词在词汇或语义上转换处理扩充，生成第三从属关键词集合。

[0032] 将生成的三个从属关键词集合组合在一起，构成从关键词集合，并为每个从关键词标注与主关键词的相关度，为主相关度标识。主相关度标识表示从关键词与主关键词集合的关联强度，其中，第一从属关键词集合的主相关度最高；第三从属关键词集合的主相关度最低。

[0033] 关键词序列是将主关键词与从关键词进行合并，表示对查询需求的扩充理解。关键词类目是文档集的分类特征，实现对不同类型数据的区分处理。将主关键词集合与从关键词集合的关键词串联在一起，作为矩阵的行；将文档集中的关键词分类作为矩阵的列，构建关键词矩阵。关键词矩阵是一个二维矩阵，矩阵行为关键词序列，矩阵列为关键词类目，矩阵各单元的交叉点表示关键词序列与文档分类的匹配与关联。

[0034] 通过设置调幅并采用多元处理，实现了对查询需求的深入理解和扩充，实现了对查询与数据特征的匹配建模，为实现高精度智能查询提供了支持。

[0035] 进一步的，如图2所示，本申请实施例还包括：

步骤S410：基于所述关键词矩阵，提取所述主关键词集合；

步骤S420：遍历所述多个检索数据子库，对所述基于所述主关键词集合进行相似度匹配，确定一项相似度矩阵；

步骤S430：若所述一项相似度矩阵为空，提取所述从关键词集合并遍历所述多个检索数据子库进行相似度匹配，确定二项相似度矩阵；

步骤S440：若所述二项相似度矩阵为空，基于所述主关键词集合，遍历所述多个检索数据子库进行语义识别，获取三项相似度矩阵。

[0036] 具体而言，从构建的关键词矩阵中直接提取主关键词集合，主关键词集合为查询需求的核心。将主关键词集合在多个检索数据子库中进行查找和匹配，得到每一主关键词与数据子库中各文档的相似度值，构成一项相似度矩阵，相似度匹配计算主关键词与数据子库文档的语义相似度。可以采用向量空间模型、BM25算法、词嵌入模型等实现，一项相似度矩阵记录主关键词与数据子库中所有文档的相似度匹配值。

[0037] 当存在检索数据子库规模太小、词汇覆盖率低等情况时，一项相似度矩阵为空，则通过列条件过滤、KMeans聚类等方法从关键词矩阵中提取从关键词集合并遍历数据子库进行相似度匹配，得到每一从关键词与各文档的相似度值，构成二项相似度矩阵。二项相似度矩阵记录从关键词与数据子库中所有文档的相似度匹配值。

[0038] 当存在扩展词汇不足、词汇覆盖率低等情况时，二项相似度矩阵为空，基于主关键词集合，采用语义识别技术遍历数据子库进行语义识别，判断主关键词与各文档在语义上是否匹配关联，构成三项相似度矩阵。语义识别通过计算主关键词与文档在语义网络上的邻近程度等实现匹配判断。三项相似度矩阵记录主关键词与数据子库中各文档的语义相关度。

[0039] 通过在多个不同情况下采用不同技术手段建立不同相似度矩阵，实现了主从关键

词与文档集的相似度匹配和语义识别，生成的三个相似度矩阵从数量和质量上为后续结果判断提供了综合判断依据，避免了单一匹配方式的局限，提高了查询精度。

[0040] 进一步的，本申请实施例还包括：

相似度矩阵计算公式为：

$$S(A,B)=\alpha*\beta\frac{A*B}{|A||B|}=\begin{vmatrix} S_{11} & S_{12} & \cdots & S_{1N} \\ S_{21} & S_{22} & \cdots & S_{2N} \\ \cdots & \cdots & \cdots & \cdots \\ S_{M1} & S_{M2} & \cdots & S_{MN} \end{vmatrix};$$

其中，α表征关键词出现频次，β表征主相关度，A为待进行匹配分析的关键词矩阵，为列矩阵，B为所述检索数据库中的文档矩阵，为行矩阵，S_{MN}表示第M个关键词与第N项文档的相似度，M、N为量值，表征关键词项数与文档项数，针对所述一项相似度矩阵，β=1。

[0041] 具体而言，确定好主关键词集合和从关键词集合以及构建各检索数据子库之后，对应不同的数据子库建立文档矩阵B，表示数据子库中分类整理的文档集合，其行数对应文档数量，列数对应文档特征维度。针对不同情况下的相似度匹配，列矩阵A代表的含义不同，构建一项相似度矩阵中，矩阵A为主关键词集合构建的主关键词矩阵；构建二项相似度矩阵中，矩阵A为从关键词集合构建的从关键词矩阵；构建三项相似度矩阵中，矩阵A为基于主从关键词集合的构建的关键词矩阵。M、N分别为关键词项数与文档项数的量值，表示矩阵A和B的规模。

[0042] α表示关键词出现频次，用于衡量关键词的重要性，关键词频次越高，其重要性越大。β表示主相关度，用于衡量从关键词与对应主关键词的关联强度，主相关度越大，关键词与查询需求的关联越紧密。因为一项相似度矩阵是基于主关键词集合构建，主关键词与查询主题的关联强度最大，所以主相关度β=1。α和β的设置根据关键词本身因素和与查询主题的相关性进行设置。

[0043] 相似度矩阵计算公式采用向量相乘的形式，矩阵A的每一列向量与矩阵B的每一行向量进行相乘，得到的相乘结果表示对应关键词和文档的匹配得分。匹配得分除以向量长度的乘积，得到相似度S_{MN}值，构成相似度矩阵。

[0044] 通过相似度矩阵计算公式，直接通过两个待匹配矩阵计算得到匹配结果矩阵，实现高精度查询，提高电子文档筛选查询的效率。

[0045] 进一步的，本申请实施例还包括：

矩阵列求和公式为：

$$S_j=\sum_{i=1}^{M}S_{ij};$$

其中，S_j为M个关键词与第j项文档的相似度匹配结果，S_{ij}表示第i个关键词与第j项文档的相似度，i∈M，j∈N。

[0046] 具体而言，获取相似度矩阵后，对相似度矩阵的每一列的值进行求和，得到矩阵列向量；矩阵列向量表示关键词集合与对应列文档的匹配度，第j列向量的求和结果S_j表示关键词集合与第j项文档的总的匹配度，其中，i∈M，j∈N。矩阵列向量的值越大，表示对应文档

与关键词集合的匹配度越高，该文档的重要性也越大。

[0047] 通过对相似度矩阵每一列的值进行求和，得到关键词集合与每个文档的总匹配度，为实现文档的重要性判定与排序提供支持，为根据匹配度选择与关键词最相匹配的文档提供了依据。

[0048] 进一步的，本申请实施例还包括：

步骤S810：配置多元数据处理规则；

步骤S820：基于所述数据处理规则，对所述多个检索数据子库执行规则匹配与数据预处理，确定预处理数据库；

步骤S830：基于所述预处理数据库，进行所述关键词矩阵的匹配执行。

[0049] 具体而言，根据不同的数据源与特征，综合源数据格式、质量、分类细致度等多方面因素配置对应的数据处理规则，用于指导数据提取、清洗、转换与筛选过程。多源数据处理规则的配置需要从宏观到微观，需要考虑各数据源间与源内的特征要素，设置同源规范化、跨源规范化以及相互配合使用的等方面，需要人工判断与技术算法相结合，例如Apriori算法、Word2Vec算法、贝叶斯分类器等。

[0050] 根据设定的规则，对不同的数据源，即多个检索数据子库执行预处理过程，包括数据清洗、提取、转换与筛选，得到结构化的预处理数据库。预处理数据库提取数据源特征，其数据结构满足后续的矩阵匹配运算要求。通过预处理数据库构建文档矩阵，通过关键词集合构建关键词矩阵，将文档矩阵和关键词矩阵进行相乘运算，得到关键词与数据子库的匹配结果。

[0051] 通过配置数据处理规则，对不同的数据源执行预处理与数据提取，形成满足矩阵匹配要求的结构化数据库，再基于数据库与关键词构建的矩阵进行相乘匹配，实现关键词与海量数据源的快速精准匹配，对数据源进行简化和标准化，并通过矩阵实现文档匹配过程，提高对文档匹配的效率，提高最终结果的准确性。

[0052] 进一步的，如图3所示，本申请实施例还包括：

步骤S710：对所述相似度匹配结果进行正序列化调整，生成相似度序列，所述相似度序列由大到小排列；

步骤S720：获取查询需求项数；

步骤S730：基于所述查询需求项数对所述相似度序列进行截取，反向匹配映射文档，集成作为查询文档集合；

步骤S740：基于所述查询文档集，确定所述文档查询结果，所述文档查询结果具有文档优先级。

[0053] 具体而言，相似度矩阵的匹配结果表示关键词与每个文档的相似度得分，对所有文档的相似度得分进行排序，从大到小排列，生成相似度序列。序列中的每个元素对应一个文档，其值表示关键词与该文档的相似度。查询需求的文档数量N表示用户需要检索的文档篇数，N的值可以由用户在查询时设定，也可以通过系统默认值确定。从相似度序列的头部选取前N个元素，对应的文档构成查询文档集合，包含与关键词最相关的前N篇文档。根据查询文档集合中每个文档在相似度序列中的位置，判定每个文档的优先级，位置越靠前，优先级越高。

[0054] 通过对相似度矩阵匹配结果生成的相似度序列进行截取，获得与关键词匹配度高

的文档集合，再以此确定文档查询结果及其排序，实现根据查询需求提供最相关文档信息的目的，提高电子文档筛选的效率和准确性。

[0055] 综上所述，本申请实施例所提供的一种电子文档筛选查询方法具有如下技术效果：

连接业务管理系统，确定检索域，为后续构建检索数据库和查询提供数据基础；基于检索域获取目标文档集，跨域关联构建检索数据库，检索数据库由多个检索数据子库构成，各检索数据子库存在数据类型差异，检索数据库实时更新，实现大规模文档的统一检索平台，提高检索效率，并且数据库实时更新，支持动态查询；基于查询需求，确定主关键词集合与从关键词集合，配置关键词矩阵，从关键词集合由主关键词集合多元化处理获取，通过多元化处理获得丰富的主关键词和从关键词，支持多元化查询需求；结合相似度匹配算法，遍历多个检索数据子库对关键词矩阵进行匹配，生成相似度矩阵，其中，关键词出现频次为附加生成信息，结合考虑关键词出现频次，实现高精度匹配；设定相似度阈值，基于相似度阈值对相似度矩阵进行判定，确定单项关键词匹配结果，从而确定每个关键词与文档的精确匹配结果，只保留高度相关的匹配，提高查询准确度；基于单项关键词匹配结果，对相似度矩阵逐矩阵列求和生成相似度匹配结果，相似度匹配结果表征匹配的关键词集合与检索数据库中单项文档的综合相似度，减少计算量，提高效率；基于相似度匹配结果进行文档映射，确定文档查询结果，从而实现了电子文档的高精度、动态、多元化查询，达到了提高电子文档筛选准确度和效率的技术效果。

实施例二

[0056] 基于与前述实施例中一种电子文档筛选查询方法相同的发明构思，如图4所示，本申请实施例提供了一种电子文档筛选查询系统，其特征在于，所述系统包括：

确定检索域模块11，用于连接业务管理系统，确定检索域；

检索数据库构建模块12，基于所述检索域获取目标文档集，跨域关联构建检索数据库，所述检索数据库由多个检索数据子库构成，各检索数据子库存在数据类型差异，所述检索数据库实时更新；

关键词矩阵模块13，用于基于查询需求，确定主关键词集合与从关键词集合，配置关键词矩阵，所述从关键词集合由所述主关键词集合多元化处理获取；

相似度矩阵模块14，用于结合相似度匹配算法，遍历所述多个检索数据子库对所述关键词矩阵进行匹配，生成相似度矩阵，其中，关键词出现频次为附加生成信息；

关键词匹配结果模块15，用于设定相似度阈值，基于所述相似度阈值对所述相似度矩阵进行判定，确定单项关键词匹配结果，其中，匹配成功标识为1，匹配失败标识为0；

相似度匹配结果模块16，基于所述单项关键词匹配结果，对所述相似度矩阵逐矩阵列求和生成相似度匹配结果，所述相似度匹配结果表征匹配的关键词集合与所述检索数据库中单项文档的综合相似度；

文档查询结果模块17，基于所述相似度匹配结果进行文档映射，确定文档查询结果。

[0057] 进一步的，本申请实施例还包括：

主关键词集合模块，基于所述查询需求，提炼多个主关键词，作为所述主关键词集

合；

多元化处理调幅模块，用于配置多元化处理调幅；

第一从属关键词集合模块，用于基于所述多元化处理调幅，对所述主关键词集合进行上位化处理，确定第一从属关键词集合；

第二从属关键词集合模块，用于基于所述多元化处理调幅，对所述主关键词集合进行下位化处理，确定第二从属关键词集合；

第三从属关键词集合模块，用于对所述主关键词集合进行转换处理，确定第三从属关键词集合；

从关键词集合模块，用于基于所述第一从属关键词集合、所述第二从属关键词集合与所述第三从属关键词集合，确定从关键词集合，所述从关键词集合带有主相关度标识；

搭建关键词矩阵模块，用于将关键词序列作为矩阵行，将关键词类目作为矩阵列，基于所述主关键词集合与所述从关键词集合搭建所述关键词矩阵。

[0058] 进一步的，本申请实施例还包括：

主关键词集合提取模块，基于所述关键词矩阵，提取所述主关键词集合；

一项相似度矩阵模块，用于遍历所述多个检索数据子库，对所述基于所述主关键词集合进行相似度匹配，确定一项相似度矩阵；

二项相似度矩阵模块，用于若所述一项相似度矩阵为空，提取所述从关键词集合并遍历所述多个检索数据子库进行相似度匹配，确定二项相似度矩阵；

三项相似度矩阵模块，用于若所述二项相似度矩阵为空，基于所述主关键词集合，遍历所述多个检索数据子库进行语义识别，获取三项相似度矩阵。

[0059] 进一步的，本申请实施例还包括：

矩阵计算公式模块，用于进行相似度矩阵技术，公式如下：

$$S(A,B)=\alpha*\beta\frac{A*B}{|A||B|}=\begin{vmatrix}S_{11} & S_{12} & \cdots & S_{1N}\\ S_{21} & S_{22} & \cdots & S_{2N}\\ \cdots & \cdots & \cdots & \cdots\\ S_{M1} & S_{M2} & \cdots & S_{MN}\end{vmatrix};$$

其中，α 表征关键词出现频次，β 表征主相关度，A为待进行匹配分析的关键词矩阵，为列矩阵，B为所述检索数据库中的文档矩阵，为行矩阵，S_{MN}表示第M个关键词与第N项文档的相似度，M、N为量值，表征关键词项数与文档项数，针对所述一项相似度矩阵，β =1。

[0060] 进一步的，本申请实施例还包括：

矩阵列求和模块，用于进行矩阵列求和，公式如下：

$$S_j=\sum_{i=1}^{M}S_{ij};$$

其中，S_j为M个关键词与第j项文档的相似度匹配结果，S_{ij}表示第i个关键词与第j项文档的相似度，$i\in M$，$j\in N$。

[0061] 进一步的，本申请实施例包括：

多元数据处理规则模块，用于配置多元数据处理规则；

预处理数据库模块，用于基于所述数据处理规则，对所述多个检索数据子库执行规则匹配与数据预处理，确定预处理数据库；

匹配执行模块，基于所述预处理数据库，进行所述关键词矩阵的匹配执行。

[0062] 进一步的，本申请实施例还包括：

相似度序列模块，用于对所述相似度匹配结果进行正序列化调整，生成相似度序列，所述相似度序列由大到小排列；

查询需求项数模块，用于获取查询需求项数；

查询文档集合模块，用于基于所述查询需求项数对所述相似度序列进行截取，反向匹配映射文档，集成作为查询文档集合；

文档查询结果模块，基于所述查询文档集，确定所述文档查询结果，所述文档查询结果具有文档优先级。

[0063] 综上所述的方法的任意步骤都可作为计算机指令或者程序存储在不设限制的计算机存储器中，并可以被不设限制的计算机处理器调用识别用以实现本申请实施例中的任一项方法，在此不做多余限制。

[0064] 进一步的，综上所述的第一或第二可能不止代表次序关系，也可能代表某项特指概念，和/或指的是多个元素之间可单独或全部选择。显然，本领域的技术人员可以对本申请进行各种改动和变型而不脱离本申请的范围。这样，倘若本申请的这些修改和变型属于本申请及其等同技术的范围之内，则本申请意图包括这些改动和变型在内。

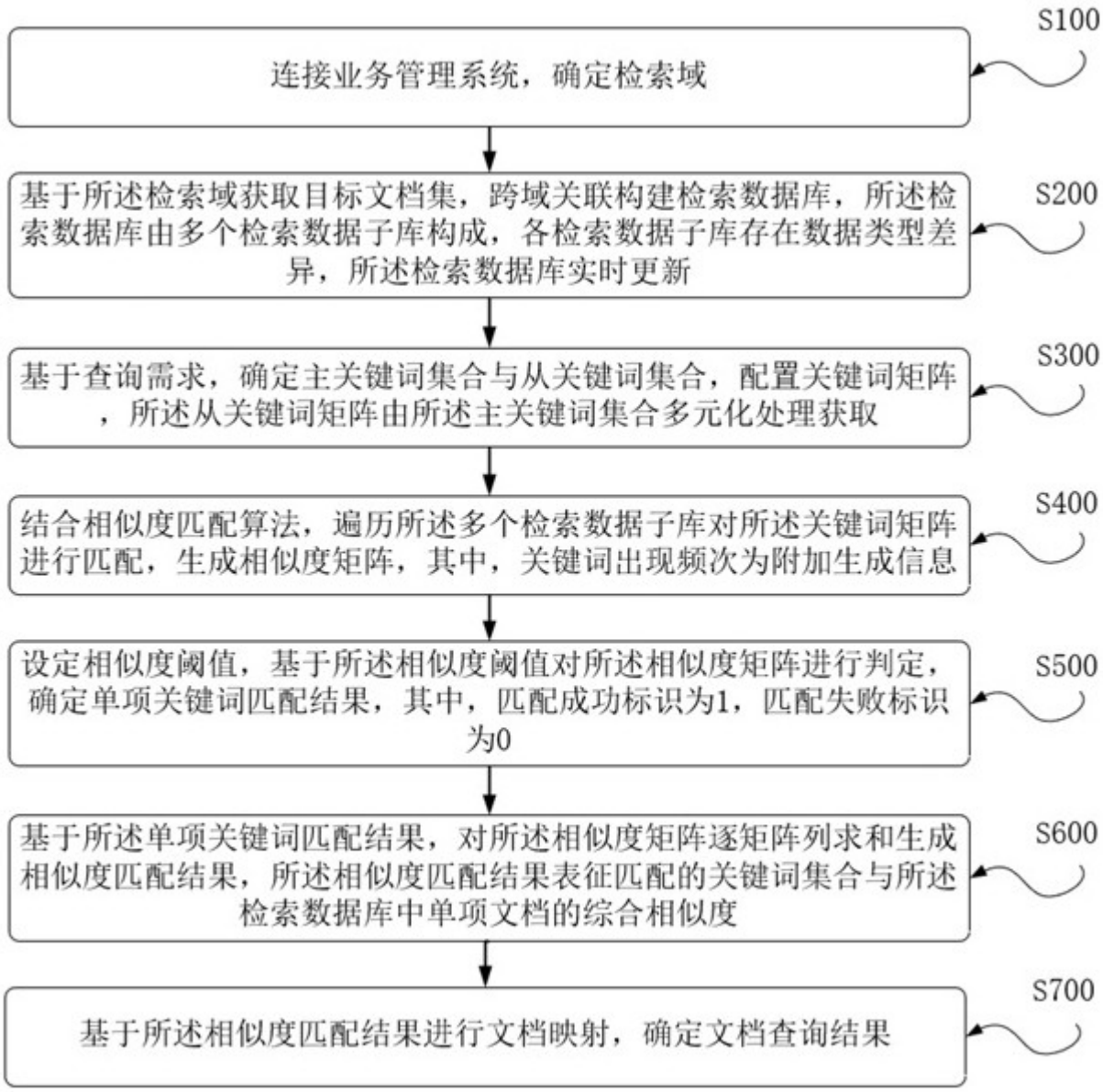

S100
连接业务管理系统，确定检索域
S200
基于所述检索域获取目标文档集，跨域关联构建检索数据库，所述检索数据库由多个检索数据子库构成，各检索数据子库存在数据类型差异，所述检索数据库实时更新
S300
基于查询需求，确定主关键词集合与从关键词集合，配置关键词矩阵，所述从关键词矩阵由所述主关键词集合多元化处理获取
S400
结合相似度匹配算法，遍历所述多个检索数据子库对所述关键词矩阵进行匹配，生成相似度矩阵，其中，关键词出现频次为附加生成信息
S500
设定相似度阈值，基于所述相似度阈值对所述相似度矩阵进行判定，确定单项关键词匹配结果，其中，匹配成功标识为1，匹配失败标识为0
S600
基于所述单项关键词匹配结果，对所述相似度矩阵逐矩阵列求和生成相似度匹配结果，所述相似度匹配结果表征匹配的关键词集合与所述检索数据库中单项文档的综合相似度
S700
基于所述相似度匹配结果进行文档映射，确定文档查询结果

图 1

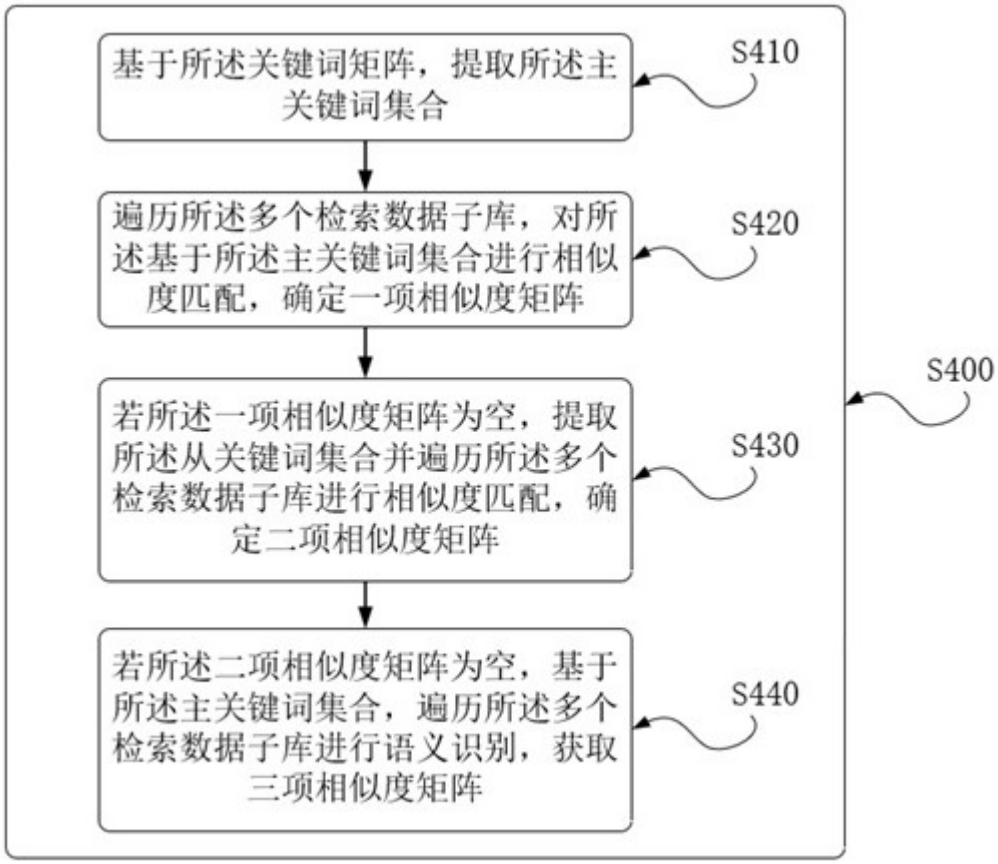

基于所述关键词矩阵，提取所述主关键词集合
S410
遍历所述多个检索数据子库，对所述基于所述主关键词集合进行相似度匹配，确定一项相似度矩阵
S420
若所述一项相似度矩阵为空，提取所述从关键词集合并遍历所述多个检索数据子库进行相似度匹配，确定二项相似度矩阵
S430
若所述二项相似度矩阵为空，基于所述主关键词集合，遍历所述多个检索数据子库进行语义识别，获取三项相似度矩阵
S440
S400

图 2

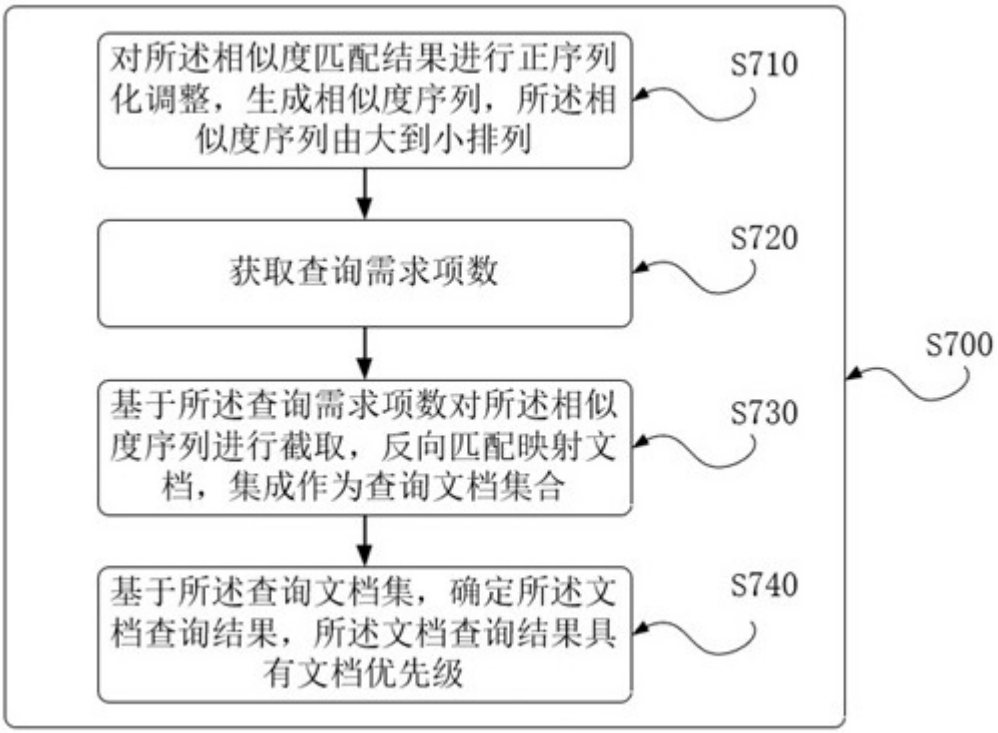

对所述相似度匹配结果进行正序列化调整，生成相似度序列，所述相似度序列由大到小排列
S710
获取查询需求项数
S720
基于所述查询需求项数对所述相似度序列进行截取，反向匹配映射文档，集成作为查询文档集合
S730
基于所述查询文档集，确定所述文档查询结果，所述文档查询结果具有文档优先级
S740
S700

图 3


电子文档筛选查询系统
确定检索域模块11
检索数据库构建模块12
关键词矩阵模块13
相似度矩阵模块14
关键词匹配结果模块15
相似度匹配结果模块16
文档查询结果模块17

图 4