(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 110245197 A
(43)申请公布日 2019.09.17

(21)申请号 201910419656.X

(22)申请日 2019.05.20

(71)申请人 北京百度网讯科技有限公司
地址 100085 北京市海淀区上地十街10号百度大厦2层

(72)发明人 任可欣 冯知凡 汪琦 张强 张扬

(74)专利代理机构 北京鸿德海业知识产权代理事务所(普通合伙) 11412
代理人 田宏宾

(51)Int.Cl.
G06F 16/28(2019.01)
G06F 16/951(2019.01)

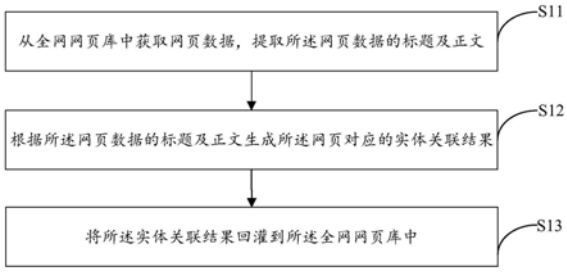
权利要求书2页 说明书12页 附图2页

(54)发明名称

一种全网实体关联方法及系统

(57)摘要

本发明公开了一种全网实体关联方法及系统,其中所述方法包括从全网网页库中获取网页数据,提取所述网页数据的标题及正文;根据所述网页数据的标题及正文生成所述网页对应的实体关联结果;将所述实体关联结果回灌到所述全网网页库中。应用本发明所述方案,能够对全网实体数据进行解析,并将解析得到的实体同知识库进行关联。支持大规模的网页库,提高了实体关联的准确率和召回率。



CN 110245197 A

1.一种全网实体关联方法,其特征在于,包括以下步骤:

从全网网页库中获取网页数据,提取所述网页数据的标题及正文;

根据所述网页数据的标题及正文生成所述网页对应的实体关联结果;

将所述实体关联结果回灌到所述全网网页库中。

2.根据权利要求1所述的方法,其特征在于,所述根据所述网页数据的标题及正文生成所述网页对应的实体关联结果包括:

确定所述标题中的实体;从所述正文中提取所述实体的上下文信息;

从知识库中确定所述实体对应的实体描述信息;

计算所述实体的上下文信息与所述实体对应的实体描述信息之间的相似度;

基于所述相似度,生成所述网页对应的实体关联结果。

3.根据权利要求2所述的方法,其特征在于,所述将所述实体关联结果回灌到所述全网网页库中包括:

利用所述实体对应的实体描述信息对所述正文中的实体进行关联。

4.根据权利要求1所述的方法,其特征在于,所述根据所述网页数据的标题及正文生成所述网页对应的实体关联结果包括:

判断所述网页的实时性;

对实时性低于或等于阈值的网页,进行批量刷库;

对于实时性高于阈值的网页,进行流式刷库。

5.根据权利要求4所述的方法,其特征在于,所述批量刷库包括:

采用Hadoop机制,根据所述网页数据的标题及正文生成与所述网页对应的实体关联结果。

6.根据权利要求4所述的方法,其特征在于,所述流式刷库包括:

采用网格计算系统,根据所述网页数据的标题及正文生成所述网页对应的实体关联结果。

7.一种全网实体关联系统,其特征在于,包括:

提取单元,用于从全网网页库中获取网页数据,提取所述网页数据的标题及正文;

生成单元,用于根据所述网页数据的标题及正文生成所述网页对应的实体关联结果;

回灌单元,用于将所述实体关联结果回灌到所述全网网页库中。

8.根据权利要求7所述的系统,其特征在于,所述生成单元具体用于:

确定所述标题中的实体;从所述正文中提取所述实体的上下文信息;

从知识库中确定所述实体对应的实体描述信息;

计算所述实体的上下文信息与所述实体对应的实体描述信息之间的相似度;

基于所述相似度,生成所述网页对应的实体关联结果。

9.根据权利要求8所述的系统,其特征在于,所述回灌单元具体用于:

利用所述实体对应的实体描述信息对所述正文中的实体进行关联。

10.根据权利要求7所述的系统,其特征在于,所述生成单元包括:

判断子模块,用于判断所述网页的实时性;

批量刷库子模块,用于对实时性低于或等于阈值的网页,进行批量刷库;

流式刷库子模块,用于对于实时性高于阈值的网页,进行流式刷库。

11.根据权利要求10所述的系统，其特征在于，所述批量刷库子模块具体用于：

采用Hadoop机制，根据所述网页数据的标题及正文生成与所述网页对应的实体关联结果。

12.根据权利要求10所述的系统，其特征在于，所述流式刷库子模块具体用于：

采用网格计算系统，根据所述网页数据的标题及正文生成所述网页对应的实体关联结果。

13.一种计算机设备，包括存储器、处理器及存储在所述存储器上并可在所述处理器上运行的计算机程序，其特征在于，所述处理器执行所述程序时实现如权利要求1～6中任一项所述的方法。

14.一种计算机可读存储介质，其上存储有计算机程序，其特征在于，所述程序被处理器执行时实现如权利要求1～6中任一项所述的方法。

一种全网实体关联方法及系统

【技术领域】

[0001] 本发明涉及计算机应用技术，特别涉及全网实体关联方法及系统。

【背景技术】

[0002] 互联网网页中存在大量的实体，大部分网页本身并没有这些实体的说明，为了帮助人们更好的理解网页内容，很多网站往往会把网页中的实体链接到相应知识库上，为读者提供更详尽的背景材料，这种做法实际上将网页和知识库建立了链接关系。这种链接技术一般称为实体关联。

[0003] 这样将网页和知识库建立链接关系，一方面可以辅助知识库的构建，比如在实体链接的基础上从网页中挖掘实体间的关系用来构建知识库，另一方面，也可以支持网页搜索等相关应用。

[0004] 目前，在网页的基础上对实体进行扩展的方式通常是对网页文本中的实体直接匹配来获取对应的实体解释信息，其准确率及召回率较低，无法达到全网实体解析及关联的要求。同时，也无法对大规模的网页库(百亿级别)进行全网实体解析及关联。

【发明内容】

[0005] 本申请的多个方面提供了全网实体关联方法、系统、设备及存储介质，能够支持大规模的网页库，提高了实体关联的准确率和召回率。

[0006] 本申请的一方面，提供一种全网实体关联方法，包括以下步骤：

[0007] 从全网网页库中获取网页数据，提取所述网页数据的标题及正文；

[0008] 根据所述网页数据的标题及正文生成所述网页对应的实体关联结果；

[0009] 将所述实体关联结果回灌到所述全网网页库中。

[0010] 如上所述的方面和任一可能的实现方式，进一步提供一种实现方式，所述根据所述网页数据的标题及正文生成所述网页对应的实体关联结果包括：

[0011] 确定所述标题中的实体；从所述正文中提取所述实体的上下文信息；

[0012] 从知识库中确定所述实体对应的实体描述信息；

[0013] 计算所述实体的上下文信息与所述实体对应的实体描述信息之间的相似度；

[0014] 基于所述相似度，生成所述网页对应的实体关联结果。

[0015] 如上所述的方面和任一可能的实现方式，进一步提供一种实现方式，所述将所述实体关联结果回灌到所述全网网页库中包括：

[0016] 利用所述实体对应的实体描述信息对所述正文中的实体进行关联。

[0017] 如上所述的方面和任一可能的实现方式，进一步提供一种实现方式，所述根据所述网页数据的标题及正文生成所述网页对应的实体关联结果包括：

[0018] 判断所述网页的实时性；

[0019] 对实时性低于或等于阈值的网页，进行批量刷库；

[0020] 对于实时性高于阈值的网页，进行流式刷库。

[0021] 如上所述的方面和任一可能的实现方式，进一步提供一种实现方式，所述批量刷库包括：

[0022] 采用Hadoop机制，根据所述网页数据的标题及正文生成与所述网页对应的实体关联结果。

[0023] 如上所述的方面和任一可能的实现方式，进一步提供一种实现方式，所述流式刷库包括：

[0024] 采用网格计算系统，根据所述网页数据的标题及正文生成所述网页对应的实体关联结果.

[0025] 本申请的另一方面，提供了一种全网实体关联系统，包括：

[0026] 提取单元，用于从全网网页库中获取网页数据，提取所述网页数据的标题及正文；

[0027] 生成单元，用于根据所述网页数据的标题及正文生成所述网页对应的实体关联结果；

[0028] 回灌单元，用于将所述实体关联结果回灌到所述全网网页库中。

[0029] 如上所述的方面和任一可能的实现方式，进一步提供一种实现方式，所述生成单元具体用于：

[0030] 确定所述标题中的实体；从所述正文中提取所述实体的上下文信息；

[0031] 从知识库中确定所述实体对应的实体描述信息；

[0032] 计算所述实体的上下文信息与所述实体对应的实体描述信息之间的相似度；

[0033] 基于所述相似度，生成所述网页对应的实体关联结果。

[0034] 如上所述的方面和任一可能的实现方式，进一步提供一种实现方式，所述回灌单元具体用于：

[0035] 利用所述实体对应的实体描述信息对所述正文中的实体进行关联。

[0036] 如上所述的方面和任一可能的实现方式，进一步提供一种实现方式，所述生成单元包括：

[0037] 判断子模块，用于判断所述网页的实时性；

[0038] 批量刷库子模块，用于对实时性低于或等于阈值的网页，进行批量刷库；

[0039] 流式刷库子模块，用于对于实时性高于阈值的网页，进行流式刷库。

[0040] 如上所述的方面和任一可能的实现方式，进一步提供一种实现方式，所述批量刷库子模块具体用于：

[0041] 采用Hadoop机制，根据所述网页数据的标题及正文生成与所述网页对应的实体关联结果。

[0042] 如上所述的方面和任一可能的实现方式，进一步提供一种实现方式，所述流式刷库子模块具体用于：

[0043] 采用网格计算系统，根据所述网页数据的标题及正文生成所述网页对应的实体关联结果。

[0044] 本发明的另一方面，提供一种计算机设备，包括存储器、处理器及存储在所述存储器上并可在所述处理器上运行的计算机程序，所述处理器执行所述程序时实现如以上所述的方法。

[0045] 本发明的另一方面，提供一种计算机可读存储介质，其上存储有计算机程序，所述

程序被处理器执行时实现如以上所述的方法。

[0046] 基于上述介绍可以看出，采用本发明所述方案，能够支持大规模的网页库，提高了实体关联的准确率和召回率。

【附图说明】

[0047] 图1为本发明所述全网实体关联方法的流程图；

[0048] 图2为本发明所述全网实体关联系统的结构图；

[0049] 图3示出了适于用来实现本发明实施方式的示例性计算机系统/服务器012的框图。

【具体实施方式】

[0050] 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，显然，所描述的实施例是本申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的全部其他实施例，都属于本申请保护的范围。

[0051] 图1为本发明所述全网实体关联方法实施例的流程图，如图1所示，包括以下步骤：

[0052] 步骤S11、从全网网页库中获取网页数据，提取所述网页数据的标题及正文；

[0053] 步骤S12、根据所述网页数据的标题及正文生成所述网页对应的实体关联结果；

[0054] 步骤S13、将所述实体关联结果回灌到所述全网网页库中。

[0055] 在步骤S11的一种优选实现方式中；

[0056] 所述全网网页库为大规模的网页库(中文网页至少百亿级别以上)，例如百度搜索引擎从网络中所爬取的中文网页页面数据。

[0057] 优选地，从所述全网网页库中获取网页数据，提取所述网页数据的标题及正文。

[0058] 在步骤S12的一种优选实现方式中，

[0059] 所述根据所述网页数据的标题及正文生成所述网页对应的实体关联结果包括以下子步骤：

[0060] 子步骤S121、确定所述标题中的实体；从所述正文中提取所述实体的上下文信息；

[0061] 子步骤S122、从知识库中确定所述实体对应的实体描述信息；

[0062] 子步骤S123、计算所述实体的上下文信息与所述实体对应的实体描述信息之间的相似度；

[0063] 子步骤S124、基于所述相似度，生成所述网页对应的实体关联结果。

[0064] 在子步骤S121的一种优选实现方式中，

[0065] 优选地，对一个给定的网页，对其HTML代码进行解析，然后采用基于规则的方法从标题标签中提取实体；从网页正文中提取所述实体的上下文信息。

[0066] 在本实施例中，实体可以是预设类型的词语，例如术语、专有名词等等。实体描述信息集合中的实体描述信息与实体集合中的实体一一对应。实体集合中的实体可以是百科词条，也可称为百科条目，是词条的一种特定表现形式，用以指百科全书中的词条，是构成百科全书的基本单元，这里的百科全书可以使用纸质和网络等不同的载体。与实体对应的实体描述信息可以是对一个词条所对内容的概括性描述。通常，实体描述信息可以包括但

不限于以下至少一项:文本信息、图片信息、音频信息、视频信息等等。

[0067] 优选地,从网页正文中提取所述实体的上下文信息。其中,实体的上下文信息可以表征实体在网页正文中的含义。在一些实施例中,上述执行主体可以从网页正文中提取出包含该实体的语句,作为该实体的上下文信息。在另一些实施例中,上述执行主体可以从网页正文中提取出包含该实体的段落,作为该实体的上下文信息。

[0068] 在本实施例的另一种优选实现方式中,对网页数据的标题及正文进行解析,确定所述网页数据的标题及正文中的实体,并从中提取实体的上下文信息。

[0069] 优选地,可以通过多种方式确定所述网页数据的标题及正文中的实体。例如,对所述网页数据的标题及正文进行分词,得到关键词,并将得到的全部或部分关键词作为所述网页数据的标题及正文中的实体。例如,首先对所述网页数据的标题及正文进行分词,得到关键词;然后将关键词在实体描述信息集合对应的实体集合中匹配,得到匹配结果;最后基于匹配结果,确定所述网页数据的标题及正文中的实体。

[0070] 在本实施例的另一种优选实现方式中,对所述网页数据的标题即正文进行实体识别,识别出待关联的实体和概念集合。

[0071] 在子步骤S122的一种优选实现方式中,

[0072] 优选地,从实体描述信息集合中确定出所述网页正文中的实体对应的实体描述信息。具体地,首先将所述网页正文中的实体在实体描述信息集合对应的实体集合中匹配,确定出与所述网页正文中的实体匹配的实体;然后从实体描述信息集合中查找出匹配的实体对应的实体描述信息,作为所述网页正文中的实体对应的实体描述信息。

[0073] 优选地,从实体描述信息集合中确定出所述网页正文中的实体对应的所有实体描述信息。

[0074] 在子步骤S123的一种优选实现方式中,

[0075] 在本实施例的一个优选实施例中,

[0076] 优选地,基于相似度,利用所述实体对应的实体描述信息对网页正文中的实体进行处理。可以将相似度与预先设定的相似度阈值(例如0.8)进行比较,若大于相似度阈值,那么认为实体对应的实体描述信息与网页正文中的实体关联,反之,则不进行关联。通常,相似度越高,说明网页正文中的实体与实体对应的实体描述信息越匹配,反之,说明网页正文中的实体与实体对应的实体描述信息越不匹配。

[0077] 优选地,利用dssm深度语言匹配模型对所述实体的上下文信息的特征向量及所述实体对应的所有实体描述信息进行rank排序,获得rank得分。

[0078] 在本实施例的另一个优选实施例中,

[0079] 优选地,计算所述实体的上下文信息的特征向量及所述实体对应的实体描述信息的特征向量之间的相似度。

[0080] 优选地,将实体的上下文信息输入至预先训练的第一特征提取模型,得到实体的上下文信息的特征向量。其中,实体的上下文信息的特征向量可以用于表征实体的上下文信息的主要内容。

[0081] 所述第一特征提取模型用于提取实体的上下文信息的特征向量,表征实体的上下文信息与实体的上下文信息的特征向量之间的对应关系。第一特征提取模型可以是对大量样本实体的上下文信息和对应的特征向量进行统计分析,而得到的存储有多个样本实体的

上下文信息与对应的特征向量的对应关系表。

[0082] 优选地，将实体对应的实体描述信息输入至预先训练的第二特征提取模型，得到实体对应的实体描述信息的特征向量。其中，实体对应的实体描述信息的特征向量可以用于表征实体对应的实体描述信息的主要内容。

[0083] 所述第二特征提取模型用于提取实体对应的实体描述信息的特征向量，表征实体对应的实体描述信息与实体对应的实体描述信息的特征向量之间的对应关系。第二特征提取模型可以是对大量样本实体的实体描述信息和对应的特征向量进行统计分析，而得到的存储有多个样本实体的实体描述信息与对应的特征向量的对应关系表。

[0084] 优选地，计算实体的上下文信息的特征向量与实体对应的实体描述信息的特征向量之间的余弦相似度。

[0085] 所述，余弦相似度是通过测量两个向量的夹角的余弦值来度量它们之间的相似度。

[0086] 在子步骤S124的一种优选实现方式中，

[0087] 优选地，基于所述相似度，生成所述网页对应的实体关联结果。

[0088] 优选地，输出所述网页中的实体以及对所述实体的实体关联结果。

[0089] 优选地，将相似度与预先设定的相似度阈值(例如0.8)进行比较，若大于相似度阈值，那么认为实体对应的实体描述信息与网页正文中的实体关联，反之，则不进行关联。

[0090] 优选地，对rank排序的top1结果与网页正文中的实体关联。

[0091] 优选地，对排序第一的实体关联结果进行关联决策，例如，进行神经-免疫-学习NIL判别，以对关联结果进行有效性确认，规避掉关联错误或实体不在库中的情况。

[0092] 在本实施例的一个优选实施例中，

[0093] 由于全网网页库的量级问题(中文网页至少百亿以上)，现有计算方式无法满足对上述量级的数据的处理需求。

[0094] 优选地，判断所述网页的实时性。全网网页库中的网页，其实时性存在差异，大批量的网页实时性不高，例如读书、服务等板块，其更新较慢；而另外一些小批量的网页实时性较高，例如新闻、娱乐版块的网页，其更新较快。因此，针对其实时性的高低，采取不同的处理机制。

[0095] 优选地，对实时性低于或等于阈值的网页，进行批量刷库；对于实时性高于阈值的网页，进行流式刷库。其中，所述批量刷库包括：通过接口调用Hadoop机制，根据所述网页数据的标题及正文生成与所述网页对应的实体关联结果。所述流式刷库包括：通过接口调用网格计算系统，根据所述网页数据的标题及正文生成所述网页对应的实体关联结果。

[0096] Hadoop，是一个分布式系统基础架构，由Apache基金会开发。用户可以在不了解分布式底层细节的情况下，开发分布式程序。充分利用集群的威力高速运算和存储。简单地说来，Hadoop是一个可以更容易开发和运行处理大规模数据的软件平台。该平台使用的是面向对象编程语言Java实现的，具有良好的可移植性。Hadoop的核心组件主要是由HDFS、MapReduce和Hbase组成。HDFS是Google File System(GFS)的开源实现。MapReduce是Google MapReduce的开源实现。HBase是Google BigTable的开源实现。

[0097] 本实施例中，采用Hadoop机制实现了一个分布式文件系统，将大批量的实时性不高的网页数据发送到Hadoop集群中，由Hadoop集群根据所述网页数据的标题及正文生成所

述网页对应的实体关联结果。其中，所述Hadoop集群已经预置了相应的实体关联代码包。

[0098] 优选地，通过接口获取Hadoop机制生成的与所述网页对应的实体关联结果。

[0099] 所述网格计算系统，例如BVC(Baidu Volunteer Computing)百度网格计算系统，通过对线上、线下机器的接入管理，将闲散的时间和闲散的资源有效的组织成一个海量的计算资源池，并支持丰富的计算模型。简单来说，利用百度公司各个产品线"空闲资源"满足离线业务的计算需求。"空闲资源"是服务器的多个维度，包括但不限于:CPU,内存,磁盘,IO。其中，所述BVC中已经推送了相应的实体关联代码包到BVC框架中，通过建立远程server的方式，对小批量的实时性较高的网页进行实体关联计算。这是因为，实时性较高的网页，其更新较快，无法将其统一发送至Hadoop集群中建立并行计算任务，需要实时的对每一条网页数据进行处理。BVC可以满足上述时效性需求，以秒、分钟、小时、半天、添、周、月、季度等时间周期进行控制。并且，BVC还实现了负载均衡。

[0100] 优选地，通过接口获取BVC生成的与所述网页对应的实体关联结果。

[0101] 优选地，随着BVC计算能力的增长，可以将部分实时性不高的网页数据也发送到BVC中进行实体关联计算。

[0102] 在步骤S13的一种优选实现方式中，

[0103] 优选地，将所述实体关联结果回灌到所述全网网页库中。

[0104] 优选地，若实体对应的实体描述信息与网页正文中的实体关联，则为所述网页正文中的实体创建锚点，将所述实体对应的实体描述信息创建为到锚点的链接。即，将网页中的实体链接到相应知识库上。

[0105] 应用本发明所述方案，提高了实体关联的准确率和召回率，并且可以对大量级的全网网页数据进行实体关联，可以进一步地辅助知识库的构建，比如在实体链接的基础上从网页中挖掘实体间的关系用来构建知识库;还可以支持网页搜索等相关应用。

[0106] 需要说明的是，对于前述的各方法实施例，为了简单描述，故将其都表述为一系列的动作组合，但是本领域技术人员应该知悉，本申请并不受所描述的动作顺序的限制，因为依据本申请，某些步骤可以采用其他顺序或者同时进行。其次，本领域技术人员也应该知悉，说明书中所描述的实施例均属于优选实施例，所涉及的动作和模块并不一定是本申请所必须的。

[0107] 以上是关于方法实施例的介绍，以下通过装置实施例，对本发明所述方案进行进一步说明。

[0108] 图2为本发明所述全网实体关联系统实施例的流程图，如图2所示，包括：

[0109] 提取单元21，用于从全网网页库中获取网页数据，提取所述网页数据的标题及正文；

[0110] 生成单元22，用于根据所述网页数据的标题及正文生成所述网页对应的实体关联结果；

[0111] 回灌单元23，用于将所述实体关联结果回灌到所述全网网页库中。

[0112] 在提取单元21的一种优选实现方式中；

[0113] 所述全网网页库为大规模的网页库(中文网页至少百亿级别以上)，例如百度搜索引擎从网络中所爬取的中文网页页面数据。

[0114] 优选地，所述提取单元21从所述全网网页库中获取网页数据，提取所述网页数据

的标题及正文。

[0115] 在生成单元22的一种优选实现方式中，

[0116] 所述生成单元22用于根据所述网页数据的标题及正文生成所述网页对应的实体关联结果，包括：

[0117] 提取子模块，用于确定所述标题中的实体；从所述正文中提取所述实体的上下文信息；

[0118] 确定子模块，用于从知识库中确定所述实体对应的实体描述信息；

[0119] 计算子模块，用于计算所述实体的上下文信息与所述实体对应的实体描述信息之间的相似度；

[0120] 生成子模块，用于基于所述相似度，生成所述网页对应的实体关联结果。

[0121] 在提取子模块的一种优选实现方式中，

[0122] 优选地，对一个给定的网页，对其HTML代码进行解析，然后采用基于规则的方法从标题标签中提取实体；从网页正文中提取所述实体的上下文信息。

[0123] 在本实施例中，实体可以是预设类型的词语，例如术语、专有名词等等。实体描述信息集合中的实体描述信息与实体集合中的实体一一对应。实体集合中的实体可以是百科词条，也可称为百科条目，是词条的一种特定表现形式，用以指百科全书中的词条，是构成百科全书的基本单元，这里的百科全书可以使用纸质和网络等不同的载体。与实体对应的实体描述信息可以是对一个词条所对内容的概括性描述。通常，实体描述信息可以包括但不限于以下至少一项：文本信息、图片信息、音频信息、视频信息等等。

[0124] 优选地，从网页正文中提取所述实体的上下文信息。其中，实体的上下文信息可以表征实体在网页正文中的含义。在一些实施例中，上述执行主体可以从网页正文中提取出包含该实体的语句，作为该实体的上下文信息。在另一些实施例中，上述执行主体可以从网页正文中提取出包含该实体的段落，作为该实体的上下文信息。

[0125] 在本实施例的另一种优选实现方式中，对网页数据的标题及正文进行解析，确定所述网页数据的标题及正文中的实体，并从中提取实体的上下文信息。

[0126] 优选地，可以通过多种方式确定所述网页数据的标题及正文中的实体。例如，对所述网页数据的标题及正文进行分词，得到关键词，并将得到的全部或部分关键词作为所述网页数据的标题及正文中的实体。例如，首先对所述网页数据的标题及正文进行分词，得到关键词；然后将关键词在实体描述信息集合对应的实体集合中匹配，得到匹配结果；最后基于匹配结果，确定所述网页数据的标题及正文中的实体。

[0127] 在本实施例的另一种优选实现方式中，对所述网页数据的标题即正文进行实体识别，识别出待关联的实体和概念集合。

[0128] 在确定子模块的一种优选实现方式中，

[0129] 优选地，从实体描述信息集合中确定出所述网页正文中的实体对应的实体描述信息。具体地，首先将所述网页正文中的实体在实体描述信息集合对应的实体集合中匹配，确定出与所述网页正文中的实体匹配的实体；然后从实体描述信息集合中查找出匹配的实体对应的实体描述信息，作为所述网页正文中的实体对应的实体描述信息。

[0130] 优选地，从实体描述信息集合中确定出所述网页正文中的实体对应的所有实体描述信息。

[0131] 在处理子模块的一种优选实现方式中，

[0132] 在本实施例的一个优选实施例中，

[0133] 优选地，基于相似度，利用所述实体对应的实体描述信息对网页正文中的实体进行处理。可以将相似度与预先设定的相似度阈值(例如0.8)进行比较，若大于相似度阈值，那么认为实体对应的实体描述信息与网页正文中的实体关联，反之，则不进行关联。通常，相似度越高，说明网页正文中的实体与实体对应的实体描述信息越匹配，反之，说明网页正文中的实体与实体对应的实体描述信息越不匹配。

[0134] 优选地，利用dssm深度语言匹配模型对所述实体的上下文信息的特征向量及所述实体对应的所有实体描述信息进行rank排序，获得rank得分。

[0135] 在本实施例的另一个优选实施例中，

[0136] 优选地，计算所述实体的上下文信息的特征向量及所述实体对应的实体描述信息的特征向量之间的相似度。

[0137] 优选地，将实体的上下文信息输入至预先训练的第一特征提取模型，得到实体的上下文信息的特征向量。其中，实体的上下文信息的特征向量可以用于表征实体的上下文信息的主要内容。

[0138] 所述第一特征提取模型用于提取实体的上下文信息的特征向量，表征实体的上下文信息与实体的上下文信息的特征向量之间的对应关系。第一特征提取模型可以是对大量样本实体的上下文信息和对应的特征向量进行统计分析，而得到的存储有多个样本实体的上下文信息与对应的特征向量的对应关系表。

[0139] 优选地，将实体对应的实体描述信息输入至预先训练的第二特征提取模型，得到实体对应的实体描述信息的特征向量。其中，实体对应的实体描述信息的特征向量可以用于表征实体对应的实体描述信息的主要内容。

[0140] 所述第二特征提取模型用于提取实体对应的实体描述信息的特征向量，表征实体对应的实体描述信息与实体对应的实体描述信息的特征向量之间的对应关系。第二特征提取模型可以是对大量样本实体的实体描述信息和对应的特征向量进行统计分析，而得到的存储有多个样本实体的实体描述信息与对应的特征向量的对应关系表。

[0141] 优选地，计算实体的上下文信息的特征向量与实体对应的实体描述信息的特征向量之间的余弦相似度。

[0142] 所述，余弦相似度是通过测量两个向量的夹角的余弦值来度量它们之间的相似度。

[0143] 在生成子模块的一种优选实现方式中，

[0144] 优选地，基于所述相似度，生成所述网页对应的实体关联结果。

[0145] 优选地，输出所述网页中的实体以及对所述实体的实体关联结果。

[0146] 优选地，将相似度与预先设定的相似度阈值(例如0.8)进行比较，若大于相似度阈值，那么认为实体对应的实体描述信息与网页正文中的实体关联，反之，则不进行关联。

[0147] 优选地，对rank排序的top1结果与网页正文中的实体关联。

[0148] 优选地，对排序第一的实体关联结果进行关联决策，例如，进行神经-免疫-学习NIL判别，以对关联结果进行有效性确认，规避掉关联错误或实体不在库中的情况。

[0149] 在本实施例的一个优选实施例中，

[0150] 由于全网网页库的量级问题(中文网页至少百亿以上),现有计算方式无法满足对上述量级的数据的处理需求。

[0151] 优选地,所述生成单元还包括判断子模块,用于判断所述网页的实时性。全网网页库中的网页,其实时性存在差异,大批量的网页实时性不高,例如读书、服务等板块,其更新较慢;而另外一些小批量的网页实时性较高,例如新闻、娱乐版块的网页,其更新较快。因此,针对其实时性的高低,采取不同的处理机制。

[0152] 优选地,所述生成单元还包括批量刷库子模块,用于对实时性低于或等于阈值的网页,进行批量刷库;流式刷库子模块,用于对于实时性高于阈值的网页,进行流式刷库。其中,所述批量刷库子模块具体用于,通过接口调用Hadoop机制,根据所述网页数据的标题及正文生成与所述网页对应的实体关联结果。所述流式刷库子模块具体用于,通过接口调用网格计算计算系统,根据所述网页数据的标题及正文生成所述网页对应的实体关联结果。

[0153] Hadoop,是一个分布式系统基础架构,由Apache基金会开发。用户可以在不了解分布式底层细节的情况下,开发分布式程序。充分利用集群的威力高速运算和存储。简单地说来,Hadoop是一个可以更容易开发和运行处理大规模数据的软件平台。该平台使用的是面向对象编程语言Java实现的,具有良好的可移植性。Hadoop的核心组件主要是由HDFS、MapReduce和Hbase组成。HDFS是Google File System(GFS)的开源实现。MapReduce是Google MapReduce的开源实现。HBase是Google BigTable的开源实现。

[0154] 本实施例中,采用Hadoop机制实现了一个分布式文件系统,将大批量的实时性不高的网页数据发送到Hadoop集群中,由Hadoop集群根据所述网页数据的标题及正文进行分布式计算,生成所述网页对应的实体关联结果。其中,所述Hadoop集群已经预置了相应的实体关联代码包。

[0155] 优选地,所述批量刷库子模块具体还用于,通过接口获取Hadoop机制生成的与所述网页对应的实体关联结果。

[0156] 所述网格计算系统,例如BVC(Baidu Volunteer Computing)百度网格计算系统,通过对线上、线下机器的接入管理,将闲散的时间和闲散的资源有效的组织成一个海量的计算资源池,并支持丰富的计算模型。简单来说,利用百度公司各个产品线"空闲资源"满足离线业务的计算需求。"空闲资源"是服务器的多个维度,包括但不限于:CPU,内存,磁盘,IO。其中,所述BVC中已经推送了相应的实体关联代码包到BVC框架中,通过建立远程server的方式,对小批量的实时性较高的网页进行实体关联计算。这是因为,实时性较高的网页,其更新较快,无法将其统一发送至Hadoop集群中建立并行计算任务,需要实时的对每一条网页数据进行处理。BVC可以满足上述时效性需求,以秒、分钟、小时、半天、添、周、月、季度等时间周期进行控制。并且,BVC还实现了负载均衡。

[0157] 优选地,所述流式刷库子模块具体还用于,通过接口获取BVC生成的与所述网页对应的实体关联结果。

[0158] 优选地,随着BVC计算能力的增长,可以将部分实时性不高的网页数据也发送到BVC中进行实体关联计算。

[0159] 在步骤S13的一种优选实现方式中,

[0160] 优选地,将所述实体关联结果回灌到所述全网网页库中。

[0161] 优选地,若实体对应的实体描述信息与网页正文中的实体关联,则为所述网页正

文中的实体创建锚点，将所述实体对应的实体描述信息创建为到锚点的链接。即，将网页中的实体链接到相应知识库上。

[0162] 应用本发明所述方案，提高了实体关联的准确率和召回率，并且可以对大量级的全网网页数据进行实体关联，可以进一步地辅助知识库的构建，比如在实体链接的基础上从网页中挖掘实体间的关系用来构建知识库；还可以支持网页搜索等相关应用。

[0163] 所属领域的技术人员可以清楚地了解到，为描述的方便和简洁，所述描述的终端和服务器的具体工作过程，可以参考前述方法实施例中的对应过程，在此不再赘述。

[0164] 在本申请所提供的几个实施例中，应该理解到，所揭露的方法和装置，可以通过其它的方式实现。例如，以上所描述的装置实施例仅仅是示意性的，例如，所述单元的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式，例如多个单元或组件可以结合或者可以集成到另一个系统，或一些特征可以忽略，或不执行。另一点，所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口，装置或单元的间接耦合或通信连接，可以是电性，机械或其它的形式。

[0165] 所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。

[0166] 另外，在本申请各个实施例中的各功能单元可以集成在一个处理器中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个单元中。所述集成的单元既可以采用硬件的形式实现，也可以采用硬件加软件功能单元的形式实现。

[0167] 图3示出了适于用来实现本发明实施方式的示例性计算机系统/服务器012的框图。图3显示的计算机系统/服务器012仅仅是一个示例，不应对本发明实施例的功能和使用范围带来任何限制。

[0168] 如图3所示，计算机系统/服务器012以通用计算设备的形式表现。计算机系统/服务器012的组件可以包括但不限于：一个或者多个处理器或者处理器016，系统存储器028，连接不同系统组件(包括系统存储器028和处理器016)的总线018。

[0169] 总线018表示几类总线结构中的一种或多种，包括存储器总线或者存储器控制器，外围总线，图形加速端口，处理器或者使用多种总线结构中的任意总线结构的局域总线。举例来说，这些体系结构包括但不限于工业标准体系结构(ISA)总线，微通道体系结构(MAC)总线，增强型ISA总线、视频电子标准协会(VESA)局域总线以及外围组件互连(PCI)总线。

[0170] 计算机系统/服务器012典型地包括多种计算机系统可读介质。这些介质可以是任何能够被计算机系统/服务器012访问的可用介质，包括易失性和非易失性介质，可移动的和不可移动的介质。

[0171] 系统存储器028可以包括易失性存储器形式的计算机系统可读介质，例如随机存取存储器(RAM)030和/或高速缓存存储器032。计算机系统/服务器012可以进一步包括其它可移动/不可移动的、易失性/非易失性计算机系统存储介质。仅作为举例，存储系统034可以用于读写不可移动的、非易失性磁介质(图3未显示，通常称为“硬盘驱动器”)。尽管图3中未示出，可以提供用于对可移动非易失性磁盘(例如“软盘”)读写的磁盘驱动器，以及对可移动非易失性光盘(例如CD-ROM,DVD-ROM或者其它光介质)读写的光盘驱动器。在这些情况

下,每个驱动器可以通过一个或者多个数据介质接口与总线018相连。存储器028可以包括至少一个程序产品,该程序产品具有一组(例如至少一个)程序模块,这些程序模块被配置以执行本发明各实施例的功能。

[0172] 具有一组(至少一个)程序模块042的程序/实用工具040,可以存储在例如存储器028中,这样的程序模块042包括——但不限于——操作系统、一个或者多个应用程序、其它程序模块以及程序数据,这些示例中的每一个或某种组合中可能包括网络环境的实现。程序模块042通常执行本发明所描述的实施例中的功能和/或方法。

[0173] 计算机系统/服务器012也可以与一个或多个外部设备014(例如键盘、指向设备、显示器024等)通信,在本发明中,计算机系统/服务器012与外部雷达设备进行通信,还可与一个或者多个使得用户能与该计算机系统/服务器012交互的设备通信,和/或与使得该计算机系统/服务器012能与一个或多个其它计算设备进行通信的任何设备(例如网卡,调制解调器等等)通信。这种通信可以通过输入/输出(I/O)接口022进行。并且,计算机系统/服务器012还可以通过网络适配器020与一个或者多个网络(例如局域网(LAN),广域网(WAN)和/或公共网络,例如因特网)通信。如图3所示,网络适配器020通过总线018与计算机系统/服务器012的其它模块通信。应当明白,尽管图3中未示出,可以结合计算机系统/服务器012使用其它硬件和/或软件模块,包括但不限于:微代码、设备驱动器、冗余处理器、外部磁盘驱动阵列、RAID系统、磁带驱动器以及数据备份存储系统等。

[0174] 处理器016通过运行存储在系统存储器028中的程序,从而执行本发明所描述的实施例中的功能和/或方法。

[0175] 上述的计算机程序可以设置于计算机存储介质中,即该计算机存储介质被编码有计算机程序,该程序在被一个或多个计算机执行时,使得一个或多个计算机执行本发明上述实施例中所示的方法流程和/或装置操作。

[0176] 随着时间、技术的发展,介质含义越来越广泛,计算机程序的传播途径不再受限于有形介质,还可以直接从网络下载等。可以采用一个或多个计算机可读的介质的任意组合。计算机可读介质可以是计算机可读信号介质或者计算机可读存储介质。计算机可读存储介质例如可以是——但不限于——电、磁、光、电磁、红外线、或半导体的系统、装置或器件,或者任意以上的组合。计算机可读存储介质的更具体的例子(非穷举的列表)包括:具有一个或多个导线的电连接、便携式计算机磁盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦式可编程只读存储器(EPROM或闪存)、光纤、便携式紧凑磁盘只读存储器(CD-ROM)、光存储器件、磁存储器件、或者上述的任意合适的组合。在本文件中,计算机可读存储介质可以是任何包含或存储程序的有形介质,该程序可以被指令执行系统、装置或者器件使用或者与其结合使用。

[0177] 计算机可读的信号介质可以包括在基带中或者作为载波一部分传播的数据信号,其中承载了计算机可读的程序代码。这种传播的数据信号可以采用多种形式,包括——但不限于——电磁信号、光信号或上述的任意合适的组合。计算机可读的信号介质还可以是计算机可读存储介质以外的任何计算机可读介质,该计算机可读介质可以发送、传播或者传输用于由指令执行系统、装置或者器件使用或者与其结合使用的程序。

[0178] 计算机可读介质上包含的程序代码可以用任何适当的介质传输,包括——但不限于——无线、电线、光缆、RF等等,或者上述的任意合适的组合。

[0179] 可以以一种或多种程序设计语言或其组合来编写用于执行本发明操作的计算机程序代码，所述程序设计语言包括面向对象的程序设计语言—诸如Java、Smalltalk、C++，还包括常规的过程式程序设计语言—诸如“C”语言或类似的程序设计语言。程序代码可以完全地在用户计算机上执行、部分地在用户计算机上执行、作为一个独立的软件包执行、部分在用户计算机上部分在远程计算机上执行、或者完全在远程计算机或服务器上执行。在涉及远程计算机的情形中，远程计算机可以通过任意种类的网络——包括局域网(LAN)或广域网(WAN)连接到用户计算机，或者，可以连接到外部计算机(例如利用因特网服务提供商来通过因特网连接)。

[0180] 所属领域的技术人员可以清楚地了解到，为描述的方便和简洁，所述描述的系统，装置和单元的具体工作过程，可以参考前述方法实施例中的对应过程，在此不再赘述。

[0181] 在本申请所提供的几个实施例中，应该理解到，所揭露的方法和装置，可以通过其它的方式实现。例如，以上所描述的装置实施例仅仅是示意性的，例如，所述单元的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式，例如多个单元或组件可以结合或者可以集成到另一个系统，或一些特征可以忽略，或不执行。另一点，所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口，装置或单元的间接耦合或通信连接，可以是电性，机械或其它的形式。

[0182] 所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。

[0183] 另外，在本申请各个实施例中的各功能单元可以集成在一个处理器中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个单元中。所述集成的单元既可以采用硬件的形式实现，也可以采用硬件加软件功能单元的形式实现。

[0184] 最后应说明的是：以上实施例仅用以说明本申请的技术方案，而非对其限制；尽管参照前述实施例对本申请进行了详细的说明，本领域的普通技术人员应当理解：其依然可以对前述各实施例所记载的技术方案进行修改，或者对其中部分技术特征进行等同替换；而这些修改或者替换，并不使相应技术方案的本质脱离本申请各实施例技术方案的精神和范围。

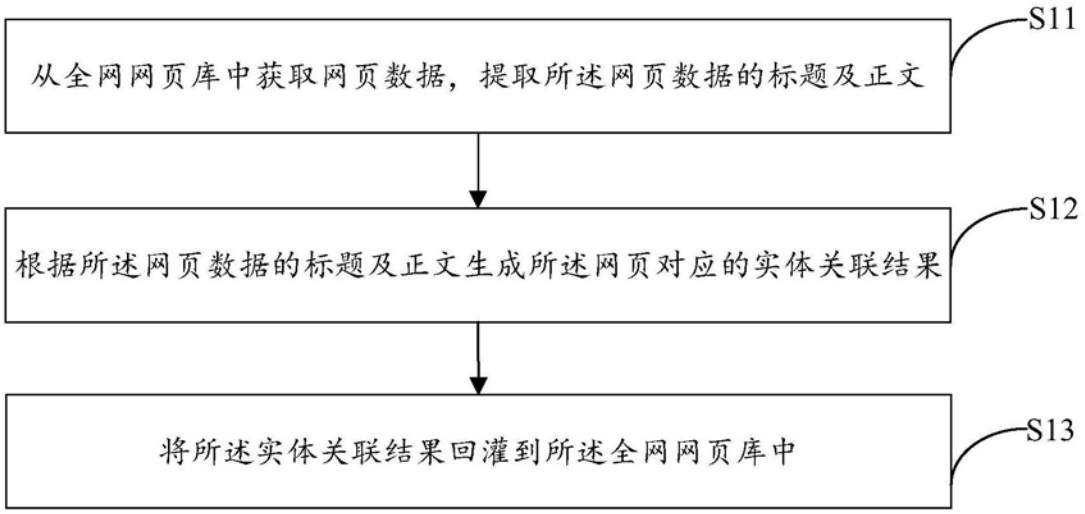

从全网网页库中获取网页数据，提取所述网页数据的标题及正文
S11
根据所述网页数据的标题及正文生成所述网页对应的实体关联结果
S12
将所述实体关联结果回灌到所述全网网页库中
S13

图1

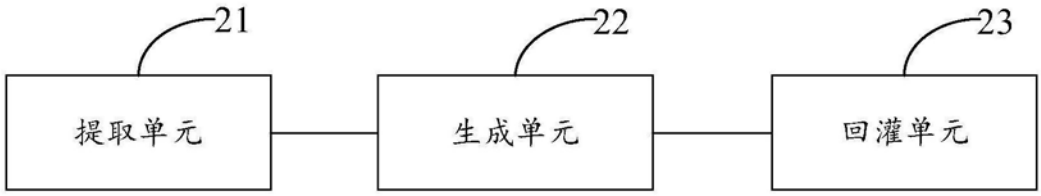

21
提取单元
22
生成单元
23
回灌单元

图2

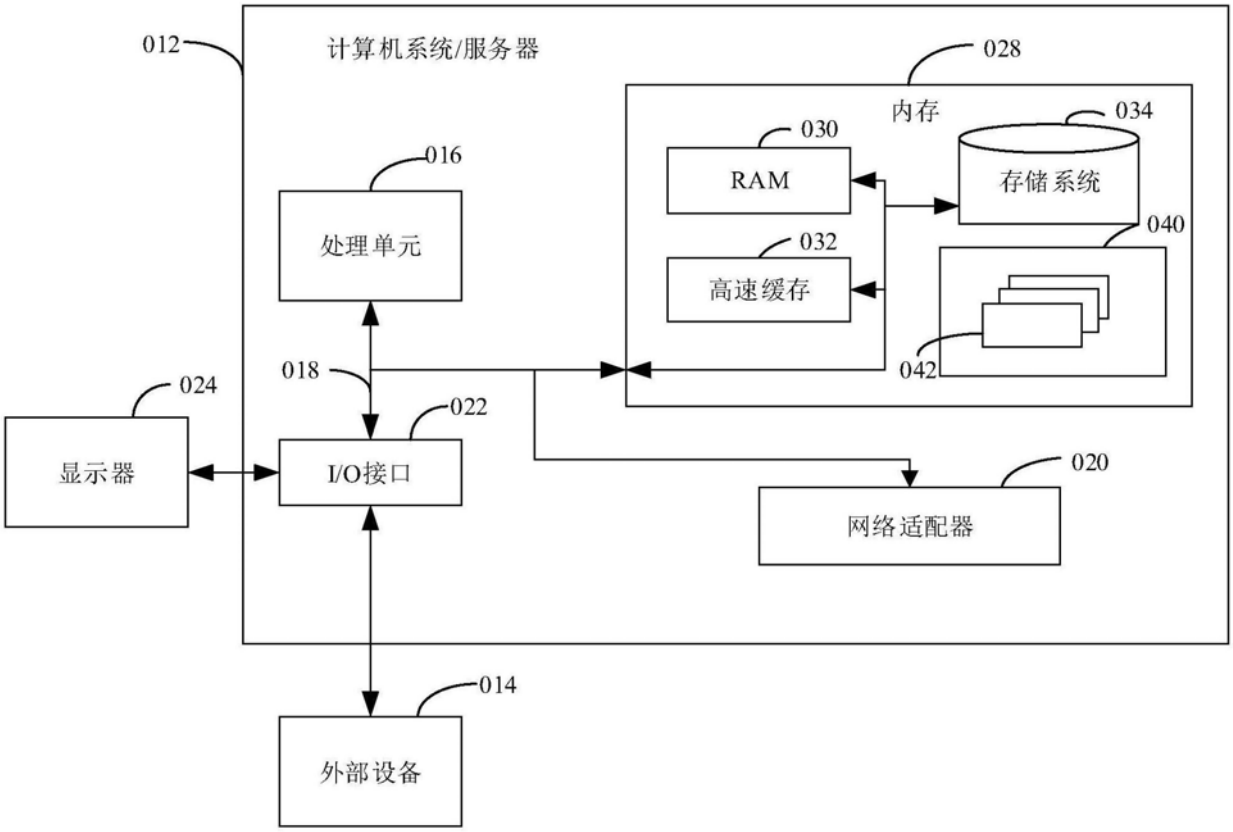

012
计算机系统/服务器
028
内存
030
RAM
034
存储系统
040
016
处理单元
032
高速缓存
042
018
024
022
显示器
I/O接口
020
网络适配器
014
外部设备

图3