(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 112585596 B
(45)授权公告日 2024.11.12

(21)申请号 201980055147.1

(22)申请日 2019.06.25

(65)同一申请的已公布的文献号
申请公布号 CN 112585596 A

(43)申请公布日 2021.03.30

(30)优先权数据
62/689,737 2018.06.25 US
62/794,177 2019.01.18 US
62/832,085 2019.04.10 US

(85)PCT国际申请进入国家阶段日
2021.02.22

(86)PCT国际申请的申请数据
PCT/US2019/039051 2019.06.25

(87)PCT国际申请的公布数据
WO2020/005986 EN 2020.01.02

(73)专利权人 硕动力公司
地址 美国加利福尼亚州

(72)发明人 E·B·帕夫里尼 J·R·布里格斯 M·克莱曼维纳 J·R·弗兰克 T·巴兰斯弗尔 C·D·卡尔弗 K·J·道尔 T·M·杜伯依斯 K·M·加布雷尔斯基 A·R·加兰特 A·W·哈斯克尔 阿卜迪-哈金·迪里 D·约翰逊 G·I·米尔斯坦 D·A·罗伯茨 A·M·泰勒 H·F·L·华莱士 L·E·佐尔纳

(74)专利代理机构 北京市联德律师事务所 11361
专利代理师 黄大正 张来光

(51)Int.Cl.
G06F 16/36(2019.01)
G06F 16/335(2019.01)

(56)对比文件
US 2006116994 A1,2006.06.01
US 2014040275 A1,2014.02.06
US 2017017708 A1,2017.01.19

审查员 向苗

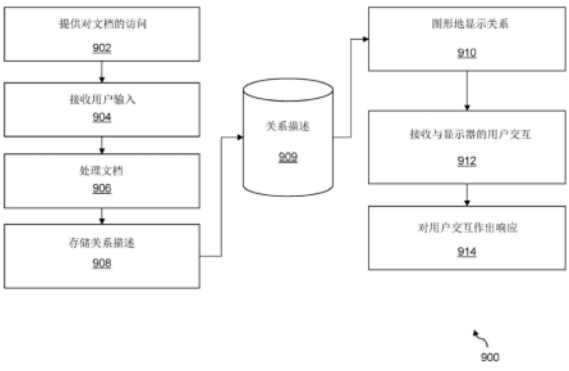
权利要求书4页 说明书43页 附图16页

(54)发明名称

用于调查实体之间的关系的系统和方法

(57)摘要

通过处理文档集合以识别包含对两个不同实体中的每个实体的共同出现提及的一个或多个文档来识别所述两个不同实体之间的联系。这种关系可以在用户界面中用对于所述两个实体中的每个实体的图标等,连同将所述实体互连的图形链接一起图形地显示。所述图形链接可以是用户界面的活动元素,其通过提供对所述文档集合内的证实所述两个实体之间的联系的证据的访问来响应用户交互。在一个方面中,用户界面中的搜索输入表单栏可以用于明确地请求证实两个实体之间的关系的文档。在另一方面中,用户可以通过明确地选择提及感兴趣实体的文档来基础化实体提及。



CN 112585596 B

1.一种系统,包括:

搜索引擎,所述搜索引擎被配置成接收对第一实体和第二实体的识别,并且处理文档集合以识别证实所述第一实体与所述第二实体之间的联系的证据段落,所述证据段落包括在所述文档集合中的至少一个文档内的对所述第一实体和所述第二实体的共同出现提及;以及

计算机,所述计算机具有处理器、存储器、显示器和用户输入设备,所述处理器由计算机可执行代码配置以执行以下步骤:识别所述第一实体和所述第二实体;将所述第一实体和所述第二实体提交给所述搜索引擎;以及响应于来自所述搜索引擎的结果,在用户界面内显示对于所述第一实体的第一图标、对于所述第二实体的第二图标、以及连接器符号,所述连接器符号将所述第一图标视觉地联接到所述第二图标并且表示其间的所述联系,并且所述连接器符号通过所述用户界面可操作以在所述用户界面内提供对所述证据段落的访问。

2.根据权利要求1所述的系统,其中所述搜索引擎在所述计算机上本地执行。

3.根据权利要求1所述的系统,其中所述搜索引擎是由所述计算机通过数据网络可访问的远程搜索引擎。

4.一种用于与实体之间的关系交互的计算机程序产品,所述计算机程序产品包括在非暂时性计算机可读介质中体现的计算机可执行代码,当在一个或多个计算设备上执行时,所述计算机可执行代码执行以下步骤:

经由用于搜索引擎的用户界面的搜索输入表单栏接收来自用户的关键字搜索;

将共指算法应用于所述关键字搜索,以预测由所述用户预期的并且由一个或多个实体相关的提及组表征的实体;

在所述用户界面中向所述用户呈现第一搜索结果,所述第一搜索结果包括第一多个实体标签和来自文档集合的多个文档,所述多个文档中的每个文档包含由所述搜索引擎在所述一个或多个实体相关的提及组中的至少一个提及组中定位的提及,并且所述多个实体标签中的每个实体标签对应于所述提及中的至少一个提及;

接收来自所述用户的从所述第一多个实体标签中对于第一实体的第一实体标签的第一选择,所述第一实体由第一实体相关的提及组表征,所述第一选择指示用户确认所述关键字搜索曾预期引用所述第一实体;

将所述第一实体标签添加到所述搜索输入表单栏;

应用相关性算法以识别第二多个文档中的提供所述第一实体与一个或多个其他实体之间的关系的证据的文本,所述一个或多个其他实体由一个或多个其他实体相关的提及组表征,所述文本包含对所述第一实体和所述一个或多个其他实体的共同出现提及;

在所述用户界面中呈现第二多个实体标签,所述第二多个实体标签各自对应于所述一个或多个其他实体中的一个实体;

接收来自所述用户的从所述第二多个实体标签中对第二实体标签的第二选择;以及

呈现包括所述第二多个文档中的一个或多个文档的第二搜索结果,所述一个或多个文档证实所述第一实体与所述第二实体之间的关系。

5.根据权利要求4所述的计算机程序产品,进一步包括执行以下步骤的代码:在所述用户界面中将所述关系图形地显示为对于所述第一实体的第一图标、对于所述第二实体的第

二图标、以及将所述第一图标视觉地联接到所述第二图标的连接器。

6.根据权利要求4所述的计算机程序产品,进一步包括执行以下步骤的代码:在所述搜索输入表单栏内将所述关系显示为所述第一实体标签由关系运算符联接到所述第二实体标签。

7.根据权利要求6所述的计算机程序产品,其中所述关系运算符包括“<>”符号。

8.根据权利要求4所述的计算机程序产品,其中所述文档集合包括在接收所述关键字搜索的用户设备上本地托管的文档。

9.根据权利要求4所述的计算机程序产品,其中所述文档集合包括分布在广域网上的文档。

10.根据权利要求4所述的计算机程序产品,其中所述多个文档包括在远程云存储设施上托管的文档。

11.根据权利要求4所述的计算机程序产品,其中呈现所述第二搜索结果包括图形地呈现多个实体标签,所述多个实体标签各自通过图谱中的边连接到所述第一实体。

12.一种方法,包括:

在用于搜索引擎的用户界面的搜索输入表单栏中接收来自用户的关键字搜索;

预测由所述用户预期的一个或多个实体;

在所述用户界面中向所述用户呈现第一搜索结果,所述第一搜索结果包括第一多个实体标签和多个文档,所述多个文档中的每个文档包含由所述搜索引擎定位的对所述一个或多个实体中的至少一个实体的提及,并且所述多个实体标签中的每个实体标签对应于所述提及中的至少一个提及;

接收来自所述用户的从所述第一多个实体标签中对于第一实体的第一实体标签的第一选择,所述第一选择指示用户确认所述关键字搜索曾预期引用所述第一实体;

将所述第一实体标签添加到所述搜索输入表单栏;

识别第二多个文档中的文本,所述文本提供所述第一实体与一个或多个其他实体之间的关系的证据,所述文本包含对所述第一实体和所述一个或多个其他实体的共同出现提及;

在所述用户界面中呈现对于所述一个或多个其他实体的第二多个实体标签;

接收来自所述用户的从所述第二多个实体标签中对于第二实体的第二实体标签的第二选择;以及

呈现包括所述第二多个文档中的一个或多个文档的第二搜索结果,所述一个或多个文档证实所述第一实体与所述第二实体之间的关系。

13.根据权利要求12所述的方法,进一步包括在所述用户界面中将所述关系图形地显示为对于所述第一实体的第一图标、对于所述第二实体的第二图标、以及将所述第一图标视觉地联接到所述第二图标的连接器。

14.根据权利要求12所述的方法,进一步包括在所述搜索输入表单栏内将所述关系显示为所述第一实体标签由关系运算符联接到所述第二实体标签。

15.根据权利要求14所述的方法,其中所述关系运算符包括“<>”符号。

16.根据权利要求12所述的方法,其中在所述关键字搜索中预测由所述用户预期的一个或多个实体包括将共指算法应用于所述关键字搜索以预测由所述用户预期的一个或多

个实体。

17.根据权利要求12所述的方法,其中识别所述多个文档中的提供所述第一实体与一个或多个其他实体之间的关系的证据的文本包括应用相关性算法以识别所述多个文档中的提供所述第一实体与一个或多个其他实体之间的关系的证据的文本。

18.根据权利要求12所述的方法,其中所述多个文档包括从接收所述关键字搜索的用户设备上本地托管的文档集合中获得的文档。

19.根据权利要求12所述的方法,其中所述多个文档包括从分布在广域网上的文档集合中获得的文档。

20.根据权利要求12所述的方法,其中所述多个文档包括从远程云存储设施上托管的文档集合中获得的文档。

21.根据权利要求12所述的方法,进一步包括存储表征所述第一实体与所述第二实体之间的关系的关系描述。

22.根据权利要求12所述的方法,进一步包括图形地显示对于所述第一实体的第一图标、对于所述第二实体的第二图标、以及将所述第一图标视觉地联接到所述第二图标以说明所述关系的连接器,其中所述连接器包括用户界面元素,所述用户界面元素提供由用户对所述第二多个文档内的所述关系的证据的访问。

23.一种系统,包括:

搜索引擎;

数据网络;

计算设备,所述计算设备通过所述数据网络耦接到所述搜索引擎,所述计算设备包括处理器、存储器和显示器,所述存储器存储由所述处理器可执行以执行以下步骤的代码:在用户界面的搜索输入表单栏中接收来自用户的关键字搜索,预测由所述用户预期的一个或多个实体;在所述用户界面中向所述用户呈现第一搜索结果,所述第一搜索结果包括第一多个实体标签和多个文档,所述多个文档中的每个文档包含由所述搜索引擎定位的对所述一个或多个实体中的至少一个实体的提及,并且所述多个实体标签中的每个实体标签对应于所述提及中的至少一个提及;接收来自所述用户的从所述第一多个实体标签中对于第一实体的第一实体标签的第一选择,所述第一选择指示用户确认所述关键字搜索曾预期引用所述第一实体;将所述第一实体标签添加到所述搜索输入表单栏;识别第二多个文档中的文本,所述文本提供所述第一实体与一个或多个其他实体之间的关系的证据,所述文本包含对所述第一实体和所述一个或多个其他实体的共同出现提及;在所述用户界面中呈现对于所述一个或多个其他实体的第二多个实体标签;接收来自所述用户的从所述第二多个实体标签中对于第二实体的第二实体标签的第二选择;以及呈现包括所述第二多个文档中的一个或多个文档的第二搜索结果,所述一个或多个文档证实所述第一实体与所述第二实体之间的关系。

24.一种用于与实体之间的关系交互的计算机程序产品,所述计算机程序产品包括在非暂时性计算机可读介质中体现的计算机可执行代码,当在一个或多个计算设备上执行时,所述计算机可执行代码执行以下步骤:

显示用于搜索引擎的用户界面,所述用户界面包括搜索输入表单栏;

接收对所述搜索输入表单栏的用户输入;

解析所述用户输入以识别标识第一实体的第一文本串、标识第二实体的第二文本串、以及在所述第一文本串与所述第二文本串之间的运算符,所述运算符向所述搜索引擎指定搜索所述第一实体与所述第二实体之间的关系的证据的请求;

响应于所述运算符,在文档集合中搜索一个或多个文档,所述一个或多个文档包含共指算法预测的提及作者预期引用第一实体的提及以及共指算法预测的提及作者预期引用第二实体的其他提及,并且提供相关性算法预测的提供所述第一实体与所述第二实体之间的关系的证据的文本,所述文本包含对所述第一实体和所述第二实体的共同出现提及;以及

在所述用户界面中呈现所述一个或多个文档中的多个文档。

用于调查实体之间的关系的系统和方法

[0001] 相关申请的交叉引用

[0002] 本申请要求于2018年6月25日提交的美国临时专利申请第62/689,737号、于2019年1月18日提交的美国临时专利申请第62/794,177和于2019年4月10日提交的美国临时专利申请第62,832,085号的优先权。这些申请中的每个申请的全部内容通过引用结合于此。

[0003] 本申请还涉及于2018年6月6日提交的美国专利申请第16/001,874号和于2015年5月12日提交的美国专利申请第14/710,342号(现在为美国专利第9,275,132号)。这些申请中的每个申请的全部内容通过引用结合于此。

技术领域

[0004] 本公开文本总体上涉及对基于实体的关系的调查,包括用于查询或以其他方式调查和管理实体之间的关系的系统、方法和用户界面。

背景技术

[0005] 已经开发了各种工具来帮助收集和分类信息。然而,仍然需要改进的工具来发现和管理基于实体的数据和关系。

发明内容

[0006] 通过处理文档集合以识别包含对两个不同实体中的每个实体的共同出现提及的一个或多个文档来识别所述两个不同实体之间的联系。这种关系可以在用户界面中用对于所述两个实体中的每个实体的图标等,连同将所述实体互连的图形链接一起图形地显示。所述图形链接可以是用户界面的活动元素,其通过提供对所述文档集合内的证实所述两个实体之间的联系的证据的访问来响应用户交互。在一个方面中,用户界面中的搜索输入表单栏可以用于明确地请求证实两个实体之间的关系的文档。在另一方面中,用户可以通过明确地选择提及感兴趣实体的文档来基础化(ground)实体提及。

附图说明

[0007] 如附图中所展示的,本文中所描述的设备、系统和方法的前述及其他目的、特征和优点将从其具体实施方案的以下描述中显而易见。附图不一定按比例绘制,而是强调展示本文中所描述的设备、系统和方法的原理。

[0008] 图1示出了用于以实体为中心的信息取得和聚合的联网环境。

[0009] 图2示出了用于推荐内容的系统。

[0010] 图3示出了用于推荐内容的方法的流程图。

[0011] 图4示出了用于推荐内容的系统。

[0012] 图5示出了用于推荐内容的方法的流程图。

[0013] 图6示出了用于在计算机辅助的知识发现过程中对活动进行日志记录的方法的流程图。

[0014] 图7示出了用于计算机辅助的研究和知识发现的系统。

[0015] 图8展示了计算机系统。

[0016] 图9示出了用于调查实体关系的方法的流程图。

[0017] 图10示出了用于调查实体之间的关系的方法的流程图。

[0018] 图11展示了用于调查实体之间的关系的用户界面。

[0019] 图12展示了用于调查实体之间的关系的用户界面。

[0020] 图13展示了用于调查实体之间的关系的用户界面。

[0021] 图14展示了用于调查实体之间的关系的用户界面。

[0022] 图15展示了用于调查实体之间的关系的用户界面。

[0023] 图16展示了用于调查实体之间的关系的用户界面。

具体实施方式

[0024] 现在将参考附图来描述实施方案,在附图中示出了优选实施方案。然而,前述内容可以用许多不同的形式来体现并且不应被解释为局限于本文中所阐述的说明性实施方案。

[0025] 本文中所提及的所有文档均通过引用以其全文结合于此。对单数形式的项目的引用应被理解为包括复数形式的项目,并且反之亦然,除非另有明确说明或从上下文中明显的。语法连词旨在表达连结的子句、句子、词语等的任何和所有的反意连接词和连接词的组合,除非另有说明或者从上下文中明显的。因此,术语“或”一般应被理解成意指“和/或”等。

[0026] 本文中的值的范围的详述不旨在是限制性的,而是单独地涉及落在所述范围内的任何和所有值,除非在本文中另有指示。此外,此类范围内的每个单独的值被结合到本说明书中,如同其在本文中被单独地叙述一样。单词“大约”、“近似”等当伴随数值时要被解释为指示如由本领域普通技术人员将理解的为了预期目的而令人满意地操作的偏差。值和/或数值的范围在本文中仅被提供作为实施例,并且不构成对所描述的实施方案的范围的限制。在本文中所提供的任何和所有实施例或示例性语言(“例如”、“诸如”等)的使用仅旨在更好地阐明实施方案并且不对实施方案或权利要求的范围造成限制。不应将本说明书中的语言解释为将任何未要求保护的元素指示为对实施方案的实践必不可少。

[0027] 在以下描述中,应理解,诸如“第一”、“第二”、“第三”、“上方”、“下方”等术语是方便性的词语并且不应被解释为限制性术语,除非另外明确陈述。

[0028] 图1示出了用于以实体为中心的信息取得和聚合的联网环境。一般地,环境100可以包括以通信关系将多个参与设备互连的数据网络102。参与设备可以例如包括任何数量的客户端设备104、服务器106、内容源108及其他资源110。

[0029] 数据网络102可以是适合于在环境100中的参与者之间传送数据和信息的任何一个或多个网络或一个或多个互联网络。这可以包括诸如互联网等公共网络、私有网络、诸如公共交换电话网或使用第三代(例如,3G或IMT-2000)、第四代(例如,LTE(E-UTRA)或高级WiMax(IEEE 802.16m))和/或其他技术的蜂窝网络等电信网络,以及可能用来在环境100中的参与者之间载送数据的多种企业区域网或局域网及其他交换机、路由器、集线器、网关等中的任何。

[0030] 数据网络102的每个参与者可以包括适当的网络接口,其包括例如网络接口卡,所述术语在本文中被宽泛地用来包括适合于建立和维持有线和/或无线通信的任何硬件(连

同用以控制相同操作的软件、固件等)。网络接口卡可以包括但不限于有线以太网接口卡(“NIC”)、无线802.11联网卡、无线802.11USB设备或用于有线或无线局域联网的其他硬件。网络接口还可以或替代地包括蜂窝网络硬件、广域无线网络硬件或用于可以用来连接到网络并载送数据的集中式、自组织(ad hoc)、对等(peer-to-peer)或其他无线电通信的任何其他硬件。在另一方面中,网络接口可以包括直接连接到诸如台式计算机等本地计算设备的串行或USB端口,所述本地计算设备进而提供到数据网络102的更一般的网络连接性。

[0031] 客户端设备104可以包括环境100内的由用户操作用于实践如在本文中设想的以实体为中心的信息取得和聚合技术的任何设备。具体地,客户端设备104可以包括用于发起和进行搜索、聚集信息、草拟实体简档、执行其他研究任务等以及管理、监测在本文中设想的系统和方法中包括的工具、平台和设备或者以其他方式与所述工具、平台和设备交互的任何设备。举例而言,客户端设备104可以包括一个或多个台式计算机、膝上型计算机、网络计算机、平板计算机、移动设备、便携式数字助理、消息传送设备、蜂窝电话、智能电话、便携式媒体或娱乐设备,或者如本文中设想的可以参与环境100的任何其他计算设备。如以上所讨论的,客户端设备104可以包括可以用于与联网环境100交互的任何形式的移动设备,诸如,任何无线的、电池供电的设备。还将理解的是,客户端设备104中的一个可以在相关功能(例如,搜索、存储实体简档等)被另一实体(诸如,服务器106、内容源108或其他资源110中的一个)执行时协调所述相关功能。

[0032] 每个客户端设备104一般地可以提供用户界面,诸如,本文中所描述的任何用户界面。用户界面可以由在从例如服务器106和内容源108接收关于实体的数据的客户端设备104中的一个上的本地执行应用来维持。在其他实施方案中,诸如在服务器106或其他资源110中的一个包括web服务器的情况下,所述web服务器通过可以在客户端设备104中的一个上执行的web浏览器或类似客户端内显示的一个或多个web页面等来提供信息,可以远程地提供并在客户端设备104中的一个上呈现用户界面。用户界面一般地可以创建用于客户端设备104中的一个的显示设备上的用户交互的适当视觉呈现,并且提供接收任何适当形式的用户输入,包括例如来自键盘、鼠标、触控板、触摸屏、手势或一个或多个其他用户输入设备的输入。

[0033] 服务器106可以包括数据储存器、网络接口以及处理器和/或其他处理电路。在以下描述中,其中描述了服务器106的功能或配置,这旨在包括服务器106的处理器的(例如,通过编程的)对应功能或配置。一般地,服务器106(或其处理器)可以执行与本文中所讨论的以实体为中心的信息取得和聚合技术相关的多种处理任务。例如,服务器106可以管理从客户端设备104中的一个或多个接收的信息,并且提供相关的支持功能,诸如,数据的搜索和管理。服务器106还可以或替代地包括对由用户在客户端设备104中的一个或多个处执行的动作作出反应的后端算法。所述后端算法还可以或替代地位于环境100中的别处。

[0034] 服务器106还可以包括促进由客户端设备104对服务器106的能力的基于web的访问的web服务器或类似前端。服务器106还可以或替代地与内容源108和其他资源110通信,以便获得用于通过客户端设备104上的用户界面而提供给用户的信息。在用户指定诸如搜索类型、语言过滤器、相关性准则(例如,用于确定搜索结果与实体的相关性)、置信度准则(例如,用于确定结果是否与特定实体相关)等的搜索准则或者例如通过对在客户端设备104上包括的实体简档上执行的动作而另外指定搜索准则的情况下,此信息可以被服务器

106(和任何关联算法)用于访问其他资源,诸如,内容源108或其他资源110,以取得相关或新的信息并对搜索结果重新排位。在此上下文中,可以有用地执行附加处理,诸如,向用户推荐新的搜索策略或者向用户推荐潜在地新的信息以用于添加到实体简档。

[0035] 服务器106还可以维持内容的数据库112连同用于使用户在客户端设备104处使用本文中所提供的任何技术(例如,自动地通过对实体简档执行的动作)来执行对数据库内容的搜索和取得的接口。因此,在一个方面中,服务器106(或包括服务器106的任何系统)可以包括实体信息的数据库112,并且服务器106可以充当提供搜索引擎的服务器,所述搜索引擎用于在数据库112中定位特定属性并且提供支持服务(诸如,对被定位的文档的取得)。

[0036] 在另一方面中,服务器106可以通过周期性地搜索数据网络102上的远程位置处的内容并索引任何得到的内容以用于客户端104的后续搜索来支持搜索活动。这可以包括存储特定文档的位置或地址信息以及以任何适当方式来解析文档以识别词语、图像、媒体、元数据等,以及特征向量或其他衍生数据的创建以帮助类似类型比较、相异比较或其他分析。在一个方面中,可以根据任何期望的标准来人工地管理(curated)数据库112。服务器106可以提供或者以其他方式支持诸如本文中所描述的任何接口之类的接口,其可以在客户端104处被提供给用户。

[0037] 服务器106还可以或替代地被配置成诸如通过提供用于管理对内容源的订阅的接口来跟踪整合内容(syndicated content)等。这可以包括用于搜索现有订阅、定位或指定新的源、订阅内容源等的工具。在一个方面中,服务器106可以管理订阅并根据来自用户的输入而自动地将新内容从这些订阅引导至客户端设备104。因此,虽然设想客户端设备104可以通过网络接口来自主地订阅内容源并直接地从此类源接收新内容,但还设想可以通过诸如服务器106等远程资源来维持此特征。在一个方面中,服务器106可以包括搜索引擎或用于本文所描述的搜索引擎算法、搜索技术、数据存储或其他算法、处理等中的任一个的其他主机(host),包括但不限于基于实体的搜索工具、基础化实体提及工具、推荐引擎、消歧过程、机器学习或机器分析平台等。

[0038] 内容源108可以包括能够被本文中所描述的技术利用例如用以更新或细化由用户创建的实体简档的呈任何结构化、半结构化或非结构化格式的任何数据或信息源。例如,内容源108可以包括但不限于Web页面(例如,公共或私有页面)、搜索引擎或搜索服务、到各种搜索服务的接口、到远程数据源的应用程序接口(API)、本地或远程数据库(例如,私有数据库、企业数据库、政府数据库、机构数据库、教育数据库等)、库(libraries)、其他在线资源、社交网络、计算机程序和应用、其他实体简档等。内容源108可以包括各种类型的信息和数据,其包括但不限于文本信息(例如,出版或未出版的信息,诸如,书、刊物、期刊、杂志、报纸、论文、报告、法律文档、报导、字典、百科全书、博客、维基等)、图形信息(例如,图表、图谱、表格等)、图像或其他视觉数据(例如,照片、图画、油画、平面图、透视图、模型、草图、图解、计算机辅助设计等)、音频数据、数值数据、地理数据、科学数据(例如,化学组成、科学配方等)、数学数据等。

[0039] 其他资源110可以包括可以在如本文中所描述的设备、系统和方法中有用地采用的任何资源。例如,其他资源110可以包括但不限于其他数据网络、人类行动者(例如,程序员、研究员、注释者、编辑等)、传感器(例如,音频或视觉传感器)、文本挖掘工具、web爬行器(web crawlers)、知识库加速(KBA)工具或其他内容监测工具等。其他资源110还可以或替

代地包括可以在如本文中设想的联网应用中有用地采用的任何其他软件或硬件资源。例如,其他资源110可以包括用于授权对内容订阅、内容购买或其他的支付的支付处理服务器或平台。作为另一实施例,其他资源110可以包括可以例如用于共享实体简档或由用户进行的其他研究或者作为实体信息的附加源的社交联网平台。在另一方面中,其他资源110可以包括用于身份的第三方验证、内容的加密或解密等的证书服务器或其他安全资源。在另一方面中,其他资源110可以包括与客户端设备104中的一个共同定位(例如,在与其相同的局域网上或者通过串行或USB电缆与其直接耦接)的台式计算机等。在此情况下,其他资源110可以为客户端设备104提供补充功能。其他资源110还包括诸如扫描仪、相机、打印机等补充资源。

[0040] 环境100可以包括一个或多个web服务器114,其向和从环境100中的任何其他参与者提供基于web的访问。虽然被描绘为单独的网络实体,但将容易理解的是,web服务器114可以与本文中所描述的其他设备中的一个逻辑地或物理地相关联,并且可以例如以准许通过数据网络102的例如来自客户端设备104的用户交互的方式而包括或提供用于对服务器106(或与其耦接的数据库112)中的一个、内容源108中的一个或其他资源110中的任一个的web访问的用户界面。

[0041] 将理解的是,环境100中的参与者可以包括用于执行如本文中所描述的各种功能的任何硬件或软件。例如,客户端设备104和服务器106中的一个或多个可以包括存储器和处理器。

[0042] 上述联网环境100的各种组件可以被布置和配置成以多种方式支持本文中所描述的技术。例如,在一个方面中,客户端设备104通过数据网络102连接到服务器106,所述服务器执行与以实体为中心的信息取得和聚合相关的多种处理任务。例如,服务器106可以托管运行以实体为中心的信息取得和聚合程序的网站,其中用户构建被用作对与实体相关的信息进行搜索、取得和排位的查询的实体简档。以此方式,当用户在客户端设备104上显示的界面上构建实体简档时,服务器106可以使用内容源108、其他资源110或数据库112来更新针对与实体简档相关的新且相关的信息的搜索。如下面更详细地讨论的,服务器106(或环境100中的另一参与者)可以包括一个或多个算法,其定义搜索并且允许服务器106对已对以实体为中心的信息取得和聚合程序采取的动作(诸如,对实体简档作出的修正或信息的选择)作出反应。更一般地,如本文中所描述的信息的搜索、处理和呈现的方面可以任何适当的方式分布。例如,搜索引擎的例如搜索功能可以在定位和处理文档的一个或多个服务器之间以及在客户端设备之间分布,所述客户端设备可以例如本地解析用户界面内的查询、向远程搜索引擎传输对信息的请求、呈现来自搜索引擎的结果、或以其他方式参与本文中所描述的信息和其他功能的各种搜索、处理和显示。类似地,在搜索内包括本地设备或云存储设施上的文档的情况下,搜索引擎可以分布式方式部署或以其他方式提供有对感兴趣的文档储存库的访问。

[0043] 图2示出了用于推荐内容的系统。具体地,推荐引擎202可以基于用户交互在用户界面206中向人类用户提供关于例如来自文档语料库204的相关文档和概念的推荐。

[0044] 用户界面206可以例如包括适合于向人类用户显示诸如知识图谱208等内容并且接收来自人类用户的诸如图形界面输入、文本输入等输入的任何显示器、界面设备、界面组件等。这可以例如包括用于上述任何客户端设备的显示器。一般地,知识图谱208可以包括

在用户界面206中由文档图标表示的一个或多个文档,连同在用户界面206中由概念图标表示的一个或多个概念。为了形成知识图谱208,这些文档和概念可以通过关系(诸如,被表示为用户界面206内的提及的视觉指示符的图谱边(graph edges))相关联。如所描绘的,在显示器中呈现的用户界面206可以包括文档图标210、第一概念图标212和第一视觉指示符216。文档图标210可以例如与文档语料库204中的第一文档218相关联,所述文档语料库可以被存储在例如单个数据库或数据存储装置中或者跨数据网络或其他分布式环境分布,并且被索引为适合于通过搜索引擎等进行识别。将理解的是,虽然与用户界面206分开地描绘,但是文档语料库204中的第一文档218和一个或多个其他文档可以位于托管和控制用户界面206的设备上。文档还可以或替代地远离设备,或其某种组合。

[0045] 应当理解的是,虽然知识图谱208在本文中被称为视觉地呈现信息的适当方式,但是还可以或替代地使用对应数据的任何其他视觉表达。例如,如本文中所描述的,与诸如知识图谱208等视觉表达的用户和机器交互被记录在用作用于相关活动的数据储存库的操作日志中。虽然可以聚集操作日志中的记录以创建用于呈现给用户的知识图谱208,但记录还可以或替代地被聚集以创建由操作日志表示的项目的任何其他适当的视觉表达。因此,例如,为了方便起见,在文本引用知识图谱208的情况下,应当理解的是,所述文本还旨在引用诸如以下描述的概要卡(summary cards)和提及突出显示等其他视觉表达,以及适合于向用户呈现信息组织和/或接收与此类信息相关的用户输入的任何其他视觉表达。

[0046] 一般地,第一概念可以在第一文档218中提及,如由在用户界面206中将第一概念图标212与文档图标210相关联的第一视觉指示符216所表示的。用户界面206可以被配置成自动地或者响应于用户请求来识别文档语料库204中也提及第一概念(在图2中一般指定为(a))的其他文档。推荐引擎202可以自动地或者响应于明确的用户请求来识别文档语料库中的标识第一概念(a)和第二概念(b)的文档213,并且然后推荐第二概念(b)以用来包括在知识图谱208中。一般地,推荐引擎202可以在用户界面206中自动地创建标识第二概念的第二概念图标214以及用于使第一概念图标212与第二概念图标214视觉地相关联的第二视觉指示符222。推荐引擎202还可以或替代地识别候选关系并且将所述关系传达给在托管用户界面206的设备上本地执行的代理,并且所述代理可以确定如何最好地配置和显示视觉指示符和/或图标。还将理解的是,推荐引擎202可以在用户界面206内自动地填充知识图谱208,或者可以通过用户界面206向用户呈现候选文档、概念或关系,使得用户可以接受、拒绝、修改或请求对所提议的添加的阐明。

[0047] 推荐引擎202还可以为任何所提议的添加提供多种支持信息。例如,推荐引擎202可以创建(包含对(a)和(b)的提及的)多个其他文档213的排位列表224,并且将排位列表224传输到设备以供在显示器中(例如,在用户界面206中)呈现。这可以包括例如证实(a)与(b)之间的关系的片段、内容摘录等,或者在所述一个或多个文档内对于对应概念中的任一个或两个的提及。可以使用多种排位技术来选择和对排位列表224中的文档进行排序,如例如以下所讨论的。

[0048] 图3示出了用于推荐内容的方法的流程图。一般地,用户可以与用户界面中显示的知识图谱交互,并且推荐引擎可以基于特定用户请求和其他上下文信息来响应地生成对添加到知识图谱的推荐。此过程可以迭代地继续以通过例如下面所描述的用户发起的计算机辅助的推荐和选择的序列来支持知识图谱的创建。

[0049] 如步骤302中所示,方法300可以开始于提供用于在显示器中呈现的文档图标、第一概念图标和第一视觉指示符。文档图标可以例如与诸如本地文档储存库、远程文档储存库、由搜索引擎索引的文档集合或其某种组合等文档语料库中的第一文档相关联。第一概念图标可以与第一文档中提及的并且也在文档语料库中的多个其他文档中提及的第一概念相关联。如本文进一步描述的,视觉指示符可以在显示器中视觉地将文档图标与第一概念图标相关联。这些显示元素可以由远程源提供、响应于用于知识图谱的远程存储模型而在本地生成、在本地生成以显示操作日志的各方面(如本文所描述的)、或以其他方式提供或创建以在显示器中的用户界面内视觉地呈现知识图谱。如上所述,虽然知识图谱是用于呈现来自操作日志的信息的一种便利技术,但是还可以或替代地使用适合于传达信息组织和/或接收与其相关的用户输入的操作日志的任何其他视觉表达。

[0050] 如步骤304中所示,方法300可以包括显示知识图谱,所述知识图谱例如视觉地示出和关联文档、概念和其间的提及,所有这些都如本文中一般设想的。一般地,知识图谱可以是用户界面内的交互式对象,并且可以诸如通过促进对附加支持信息的取得和显示或对附加推荐的请求等来支持与各个顶点(例如,文档或概念)和边(例如,在文档/概念之间的提及)的用户交互。知识图谱可以与任何其他有用的上下文信息一起显示。例如,方法300可以包括呈现来自多个其他文档的摘录以供在显示器中呈现,所述摘录证实知识图谱中的第一概念与知识图谱中的第二概念的关系。

[0051] 如步骤306中所示,方法300可以包括创建对第二概念的推荐,诸如,与在来自提及第一概念的文档语料库的多个其他文档中提及的第一概念不同的第二概念。这可以包括使用推荐引擎等创建推荐。在一个方面中,可以响应于对推荐的明确用户请求(诸如,与用户界面中的知识图谱的交互)来创建推荐。在另一方面中,可以例如响应于导航(navigation)通过知识图谱或其他用户上下文、用户活动等来自动生成推荐。

[0052] 如步骤308中所示,方法300可以包括提供用于推荐的视觉元素。例如,这可以包括:生成标识第二概念的第二概念图标和用于在显示器中将第一概念图标与第二概念图标视觉地相关联的第二视觉指示符,以及提供用于在显示器中呈现的第二概念图标和第二视觉指示符。如上所述,这可以在托管用户界面的客户端处本地执行,或者例如由提供对知识图谱的推荐的推荐引擎或由托管知识图谱呈现平台的服务器或其他云服务等远程执行。在一个方面中,第一视觉指示符可以具有与第二视觉指示符视觉地可区分的外观,例如使得用户可以在用户选择的关系与机器选择的关系之间进行区分,从而使得用户可以区分关系的创建顺序,或者以便提供对于在知识图谱内包含的关系的不同源和类型之间进行区分有用的任何其他信息。

[0053] 如步骤310中所示,方法300可以包括创建提及用于在显示器中呈现的第一概念的多个其他文档的排位列表。所述文档可以例如基于知识图谱的上下文(诸如,通向当前图谱的用户选择的历史、机器生成的推荐等)来有利地排位。所述列表还可以或替代地针对多种潜在用途进行排位,诸如,基于文档如何良好地支持两个概念图标之间的关系或者每个文档是否包含对于添加到知识图谱可能有用的新信息。通过非限制性实施例的方式,排位列表可以包括根据与第一概念的相似性、与第一文档的差异、或者第二概念与第一概念的证实化强度中的至少一个的排位。创建排位列表还可以包括响应于第二视觉指示符的用户选择而例如利用本地程序/资源或从远程资源来呈现用于在显示器中呈现的排位列表。

[0054] 如步骤312中所示,方法300可以包括接收用户输入。这可以包括键盘输入、鼠标操作或指示用户在知识图谱和相关联信息的上下文中所期望的动作的其他用户输入等。

[0055] 如步骤314中所示,方法300可以包括生成附加推荐。例如,在用户输入包括对第二文档的用户选择的情况下,响应于用户选择生成附加推荐可以包括提供用于在所述显示器中呈现的第二文档图标、第三概念图标以及第三视觉指示符,第二文档图标与第二文档相关联,第三概念图标与第二文档中的概念相关联,并且第三视觉指示符在显示器中将第二文档图标与第三概念图标视觉地相关联。这可以促进附加推荐在显示器中的呈现。将注意到,虽然附加推荐可以响应于明确用户输入而生成,但例如作为基于知识图谱的状态的后台任务,或响应于暗示用户将请求附加推荐的用户动作或者用户可能对已经可用的附加推荐感兴趣,附加推荐还可以或替代地在没有用户输入的情况下生成。

[0056] 在一个方面中,附加推荐可以例如响应于用户界面中对第一概念图标和/或第二概念图标的用户选择而包括对第三概念的推荐。进一步地,所述过程可以是迭代的,并且可以包括识别除了潜在感兴趣的新文档或概念之外的知识图谱的元素之间的关系。因此,例如,附加推荐可以包括:响应于接收对第二文档的用户选择来创建第四视觉指示符,以在显示器上将第一概念图标和第二概念图标中的至少一个与第三概念图标视觉地相关联。

[0057] 用于生成此类推荐的推荐引擎可以在呈现显示器的客户端上本地执行。推荐引擎还可以或替代地从呈现显示器的客户端远程地执行。一般地,推荐引擎可以由多种数据库和其他数据源、内容索引、处理资源、服务等中的任一种支持。因此,尽管图3中未展示,但将理解的是,方法300可以包括将文档语料库存储在推荐引擎可访问的一个或多个位置中。在另一方面中,方法300可以包括存储用于文档语料库的索引,所述索引标识文档语料库中的每个文档的至少一个概念。

[0058] 如步骤316中所示,方法300可以包括更新知识图谱或操作日志的其他视觉表达。例如,这可以包括更新包含关于知识图谱的信息(例如,文档和概念之间的关系连同此类关系的证实化信息)的数据结构。这还可以或替代地包括更新数据结构,诸如,包括与知识图谱和/或知识图谱的元素的用户交互和机器交互的历史记录的操作日志。例如,方法300可以包括存储包括与文档图标、第一概念图标和第二概念图标中的至少一个的一个或多个用户交互的操作日志以及来自推荐引擎的至少一个推荐。存储操作日志可以包括将操作日志存储在独立于托管显示器的设备的持久性存储器中。存储操作日志还可以或替代地包括将操作日志存储在独立于托管推荐引擎的设备的持久性存储器中。

[0059] 更新知识图谱还可以或替代地包括更新知识图谱的显示,这可以包括本地更新设备上的显示或从远程资源远程地更新显示。在一个方面中,所述显示可以与具有由服务器或其他远程资源控制的用户界面的远程客户端相关联,并且提供文档图标、第一概念图标和第一视觉指示符可以包括将文档图标、第一概念图标和第一视觉指示符传输至远程客户端以供在远程客户端的用户界面中呈现。

[0060] 将理解的是,方法300的一些或全部步骤可以被迭代地重复以开发知识图谱。例如,如在步骤312中可以接收附加输入,并且可以生成附加推荐并将其用于进一步更新知识图谱。

[0061] 根据上文,本文中还公开了一种用于推荐内容的系统。所述系统可以包括设备,所述设备包括显示器、所述设备的处理器和以通信关系与所述设备耦接的推荐引擎。处理器

可以被配置成提供用于在显示器中呈现的文档图标、第一概念图标和第一视觉指示符，其中文档图标与第一文档相关联，其中第一概念图标与在第一文档中提及的并且也在多个其他文档中提及的第一概念相关联，并且其中第一视觉指示符在显示器中将文档图标与第一概念图标视觉地相关联。可以远离设备或在设备本地或其某种组合的推荐引擎可以被配置成从设备接收对与第一概念相关的概念的请求并且创建对在多个其他文档中提及的第二概念的推荐，第二概念不同于第一概念并且与第一概念相关，推荐引擎进一步被配置成将包括证实第一概念与第二概念之间的关系的、来自所述多个其他文档的摘录的推荐传送给设备。

[0062] 图4示出了用于推荐内容的系统。一般地，系统400被配置成呈现诸如本文中所描述的那些的自动内容推荐的人类可读的证实。

[0063] 系统400可以包括具有用户界面404的显示器402，所述用户界面显示数个文档406的表示，诸如，文件列表、图标组或文档语料库408中的文件的任何其他适当的表示。响应于在用户界面404中对文档403的选择，系统400可以通过使用推荐引擎412从文档语料库408推荐一个或多个其他文档410来发起搜索，如本文中一般描述的。具体地，推荐引擎412可以搜索提及所选择的文档403中的第一概念(a)的文档，以及提及潜在地与第一概念(a)相关的第二概念(b)的文档。如本文中进一步描述的，推荐引擎412可以返回任何文档和概念的图形表示，连同其间的关系。推荐引擎412还可以返回信息(诸如，来自文档410的证实与所选择的文档403的关系的片段或摘录)，所述信息可以呈现在列表414(诸如，排位列表)等中。

[0064] 列表414可以进一步包含交互式内容，诸如，到支持文档410的链接、到知识图谱416内的概念的位置的链接等。

[0065] 推荐引擎412或系统400内的某个其他适当的服务或实体还可以维持如本文中所述描述的操作日志418，其通常存储与知识图谱416的人类用户交互和机器用户交互的记录，所述记录可以被共享、编辑、用作进一步推荐的上下文等。还将理解的是，系统400可以被配置成支持知识图谱416的迭代创建。例如，列表414可以用作文档406的表示，使得随着每个新的推荐列表被创建，其可以被用户用来搜索附加概念和推荐。

[0066] 因此，知识图谱416通常可以是可扩展的，并且可以例如作为操作日志418进一步与其他用户共享，以提供用于协同知识发现的平台。

[0067] 图5示出了用于推荐内容的方法的流程图。

[0068] 如步骤502中所示，方法500可以开始于诸如通过在如上所述的设备的用户界面中显示来自文档语料库的第一多个文档的表示来显示文档。这可以包括文件列表、具有对于文档的图标的窗口、或用于在用户界面内显示和操纵的文档的任何其他适当的表示。

[0069] 如步骤504中所示，方法500可以包括诸如通过在用户界面中接收从第一多个文档中对第一文档中的第一概念的用户选择来接收用户选择。这可以包括键盘操作、鼠标点击或任何其他适当的用户界面交互。

[0070] 如步骤506中所示，方法500可以包括自动地执行数个步骤以创建知识图谱的视觉元素。例如，这可以包括选择与第一文档中提及的第一概念相关联的第一概念图标，以及呈现第一概念图标以供在用户界面中显示。这还可以包括在用户界面中呈现第一视觉指示符，所述第一视觉指示符将对于第一文档的文档图标与第一概念图标视觉地连接。这还可

以包括(例如,自动地响应于用户选择)利用远离设备的推荐引擎来创建与第一概念不同的第二概念的推荐。例如,第二概念可以是在来自文档语料库的第二多个文档中提及的概念,所述第二多个文档均包括对(由用户选择的)第一概念和第二概念的提及。这还可以包括(还可选地自动响应于用户选择)呈现与第二概念相关联的第二概念图标以供在用户界面中显示,以及在用户界面中呈现第二视觉指示符,所述第二视觉指示符将第一概念图标与第二概念图标视觉地连接。

[0071] 如步骤508中所示,方法500可以包括创建从包含第一概念和第二概念的提及的第二多个文档中的一个或多个内容选择的排位列表。所述列表可以包含例如文档标题、文件名、创建日期等,以及与概念有关的信息片段以促进人工审阅。排位列表可以用于编辑、更新、验证或以其他方式审阅或修改知识图谱的任何适当的方式来排位。例如,可以根据两个相关概念之间的图谱距离对排位列表进行排位。排位列表可以根据第一概念与第二概念之间的关系的证实的估计来进行排位,例如,可以使用机器学习、人工智能、语义处理、或者用于对文档内容进行自动评估和比较的任何其他工具来对其进行评估。排位列表还可以或替代地根据第二多个文档中的每个文档内对第一概念与第二概念的提及之间的距离进行排位,其可以用作文档内的两个概念的关系的代理。排位列表还可以或替代地根据第一概念和第二概念的提及的数量、或用于评估或估计概念之间的关系或与知识图谱的相关性的任何其他适当的度量或度量组合来进行排位。

[0072] 如步骤510中所示,方法500可以包括呈现排位列表的至少一部分以供在用户界面中显示。以这种方式,可以自动地生成通过第二概念和由用户选择的第一概念而彼此相关的一组文档,并且将其作为列表呈现在用户界面中以用于进一步的用户交互。例如,一个或多个内容选择的排位列表可以包括支持第一概念图标与第二概念图标之间的关系的内容。

[0073] 如步骤512中所示,方法500可以包括为用户创建推荐。例如,在呈现排位列表之后,可以接收对附加相关概念推荐的用户请求。响应于此类用户请求,方法500可以包括将与来自文档语料库的文档中提及的概念相关联的一个或多个附加概念图标添加到用户界面,以及针对每个附加概念图标添加至少一个视觉指示符,所述视觉指示符将附加概念图标与在用户界面中显示的概念图标中的一个或多个其他概念图标视觉地相关联。

[0074] 如步骤514中所示,方法500可以包括接收和处理任何数量的附加用户请求以迭代地探索、扩展和细化知识图谱。例如,方法500可以包括接收第二用户对证实两个概念图标之间的关系的内容的请求。响应于这个第二请求,所述方法可以包括呈现来自文档语料库的文档内的一个或多个内容项目,其描述两个概念图标之间的联系。

[0075] 这还可以或替代地包括更新用户界面中的显示,诸如通过更新文档的呈现、知识图谱、文档的排位列表和用户界面的任何其他部分。这还可以或替代地包括更新存储与知识图谱的人类和机器交互的操作日志。

[0076] 图6展示了用于在计算机辅助的知识发现过程中对活动进行日志记录的方法。

[0077] 如步骤602中所示,方法600可以包括显示信息。一般地,这可以包括知识、信息、数据源、关系等的任何结构化表示。这可以例如包括如本文中所描述的知识图谱,或者适合于在设备的用户界面中呈现的任何其他知识表示等。因此,在一个方面中,这可以包括在第一表面上的视觉显示元素中呈现图谱。这可以包括例如计算机、平板计算机、智能电话等的显示,以及此类物理显示介质内的应用、过程等的窗口或其他活动图形部分。对于知识图谱

等,所述图谱可以包括多个图谱元素,所述图谱元素包括呈现为图谱的顶点的一个或多个文档和概念,以及呈现为图谱的边的在一个或多个文档和概念之间的一个或多个关系。

[0078] 如步骤604中所示,所述方法可以包括接收用户动作。这可以例如包括从人类用户接收使用视觉显示元素中的多个图元(graph elements)之一进行的用户输入,或者从人类用户接收使用呈现知识图谱或其他知识表示的用户界面或设备进行的任何其他输入。在一个方面中,这可以包括对推荐的请求、对知识图谱中关系的证实的请求、对由知识图谱中的文档图标标识的文档的请求、或与知识图谱的内容相关或基于知识图谱的内容的任何其他信息。在另一方面中,用户动作可以是对知识图谱的操纵。例如,第一动作可以包括将第一文档作为顶点添加到图谱、从图谱中移除顶点、从图谱中移除边、或者向图谱中添加边。

[0079] 在一个方面中,在用户界面显示概念的提及(例如,作为概念图标或作为文档内的文本)的情况下,由用户进行的第一动作可以包括与第一提及的用户交互,诸如,对第一提及的选择、对相关文档的请求、对相关概念的请求等。

[0080] 如步骤606中所示,方法600可以包括存储第一动作的记录。所述记录可以有用地包括关于知识图谱的状态、用户交互的性质的任何信息、或对识别或解读所述动作或采取所述动作的上下文有用的其他信息。在一个方面中,记录可以包括第一文档与一个或多个其他文档(例如,在显示内或在提供给用户的搜索或推荐内的其他文档)的关系。在另一方面中,记录可以包括第一提及与一个或多个其他文档中的一个或多个其他提及之间的关系。更一般地,记录可以包括关于知识图谱或其他知识表示、用户与用户界面交互的方式、由用户提供的任何请求或指令的特定或一般性质等的任何信息。在一个方面中,记录可以包括与第一动作相关联的第一文档的识别信息以及与第一动作相关联的第一文档内的第一概念的第一提及。

[0081] 存储记录还可以或替代地包括将第一动作的记录存储在操作日志中。如本文中所描述的,操作日志可以持久地存储在数据存储装置中,所述数据存储装置可由第一表面访问并且还可由在独立于托管第一表面的设备执行的计算平台上操作的基于机器的算法(诸如用于生成推荐的基于机器的算法等)访问。

[0082] 如步骤608中所示,方法600可以包括接收计算机动作,诸如对来自用户的第一动作作出响应的来自基于机器的算法的第二动作。第二动作可以是对用户动作的任何适当的响应。例如,第二动作可以识别由用户在第二文档中选择的概念的第二提及。在另一方面中,这可以包括可由本文中设想的推荐引擎等提供的任何推荐,包括由机器学习算法、神经网络或其他模式匹配算法等识别的相关或潜在相关的文档或概念。在另一方面中,这可以包括数据处理或操纵。例如,计算机动作可以包括在知识图谱内、在添加到知识图谱的可能推荐的列表内、或其某种组合对来自第二文档的一个或多个摘录的自动选择,所述一个或多个摘录证实第一概念的第一提及和第二概念的第二提及之间的关系。这还可以或替代地包括对知识图谱的推荐改变,诸如,添加顶点、边等。例如,基于机器的算法可以识别已经在知识图谱内的两个概念或与知识图谱相关的新概念之间的关系,并且第二动作可以包括将边添加到图谱以指示所述关系。

[0083] 如步骤610中所示,方法600可以包括存储计算机动作的记录。这可以例如包括将第二动作(或第二动作的描述)存储在操作日志或其他适当的储存库中。这还可以包括记录适合于记录的随后使用的任何上下文信息等。

[0084] 如步骤612中所示,方法600可以包括更新例如用户界面中的知识图谱或其他知识表示、或相关项(诸如,文档的排位列表、推荐等)的显示。在一个方面中,这可以包括在第一表面上的视觉显示元素中呈现在步骤608中推荐的第二文档。

[0085] 如步骤614中所示,方法600可以包括可以基于操作日志中的累积记录或者人和机器操作对知识图谱等的其他类似记录累积来执行的其他活动。如上所述,可以将操作日志存储在网络可访问的位置中,其允许例如在多个用户之间共享和协同。知识图谱的当前版本可以从操作日志中导出并根据需要显示在任何数量的设备上。更一般地,操作日志可以任何适当的方式应用于显示、共享、编辑或以其他方式操纵知识图谱或操作日志的其他视觉表达。

[0086] 例如,这可以包括应用操作日志以在第二表面上显示图谱。在一个方面中,这允许用户将包含在知识图谱内的特定研究项目移植(port)到第二表面,诸如,由用户使用的另一设备。在另一方面中,这允许用户通过应用操作日志在用于其他用户的第二表面上显示图谱来与其他用户共享知识图谱。

[0087] 在另一方面中,这可以包括促进与一个或多个其他用户共享操作日志,使得这些其他用户可以查看、修改、复制图谱或以其他方式与图谱交互。这可以例如包括通过数据网络与一个或多个其他用户共享操作日志的数据结构,或者基于操作日志发布交互式知识图谱以供其他用户使用。

[0088] 在另一方面中,单个用户可以具有多个研究项目,每个研究项目由单独的操作日志来表示。因此,方法600可以包括存储数个操作日志,其中一些可以与人类用户具有写入权限的不同项目相关联。为了确保用户动作正确地分布在此类日志中,用户界面可以采用多种机制中的任何机制来管理特定用户动作被记录在何处。例如,在一个方面中,用户界面可以请求对特定项目或操作日志的明确选择,并且此项目可以用于捕捉例如跨不同的设备、文档、应用等的所有用户交互,直到用户指定不同的项目或操作日志。在另一方面中,系统可以基于应用的改变、设备的改变等来推断改变。因此,方法600可以包括监测人类用户在一个或多个表面上的活动,并且自动选择多个操作日志中的一个以记录当前活动,或者在各种条件下或者响应于各种事件来请求对多个操作日志中的一个的用户选择。在另一方面中,系统可以在多个日志中记录操作。例如,用户可以具有允许记录其所有操作的长期运行的历史日志,以及针对也接收用户操作的子集的某些探索或任务的多个其他日志。

[0089] 在另一方面中,操作日志还可以用于其他类型的处理。例如,方法600可以包括基于操作日志中的数据的聚合来对提及概念和第二概念的文档列表进行排位。这包括明确地指定例如文档内容或知识图谱的结构的数据。这还可以或替代地包括暗示关系的数据。例如,用户动作的上下文可以暗示相关性,诸如,其中作出推荐请求的上下文。在另一方面中,用户动作的顺序可以暗示相关性,诸如,用户向知识图谱添加概念或请求对图谱的边进行证实的顺序。在另一方面中,用户动作的频率可以暗示相关性,诸如,用户请求对图谱的特定边进行证实的频率或用户请求与由图谱中的顶点表示的特定概念相关的概念的频率。

[0090] 根据前述内容,本文中还公开了一种用于在计算机辅助的知识发现过程中对活动进行日志记录的系统。一般地,所述系统可以包括数据存储装置、存储在数据存储装置上的操作日志、以及具有显示器和处理器的设备。

[0091] 数据存储装置可以例如是本文中所描述的数据库、数据存储装置、数据储存库或

其他存储器等以及前述的组合中的任一个。一般地,操作日志可以包括与项目的人类和机器交互的累积记录,所述累积记录进而可以被呈现为知识图谱或存储在操作日志中的信息的其他视觉表达。例如,操作日志可以包括对与显示器中的视觉表达(诸如,一个或多个文档、一个或多个概念等)的一个或多个人类交互和一个或多个机器交互的记录。数据存储装置可以是可由设备通过数据网络访问的远程数据存储装置,并且可以是共享的、私有的(例如,被保护以免被其他用户访问)或其某种组合。在另一方面中,数据存储装置可以是与显示器相关联的设备上的本地数据存储装置。

[0092] 设备的处理器可以被配置成支持日志的创建和使用以及相关功能,诸如,生成或显示操作日志的视觉表达。例如,处理器可以被配置成基于操作日志在显示器中呈现项目的视觉表达,并且处理器可以被配置成接收与显示器中的视觉表达的用户交互。为了维持操作日志,处理器可以被配置成将用户交互添加到操作日志。例如,在数据存储装置在本地的情况下,处理器可以直接执行将记录添加到日志的过程等。另一方面,在数据存储这种是远程数据存储装置等的情况下,处理器可以使设备向数据存储装置传输适当的指令和其他信息以在远程位置处在操作日志中创建用户交互的记录。处理器还可以执行其他相关功能,诸如,从推荐引擎请求与用户交互相关联的推荐(所述推荐引擎可以是在设备上执行的本地推荐引擎或可通过数据网络访问的远程推荐引擎)、以及由推荐引擎向操作日志添加响应。

[0093] 在显示器中呈现的项目的视觉表达可以采取任何适当的形式。这可以例如包括形成知识图谱的图标或顶点。这还可以或替代地包括提供如下所述的文档的摘录的概要卡、文档内的潜在相关内容的提及突出显示或其他视觉指示符、以及关于提及、概念等的概要信息。更一般地,在本文中所设想的项目的视觉表达内可以有用地采用操作日志中有用地向用户呈现信息和/或促进与项目相关的用户输入的记录的任何视觉表达。

[0094] 可以采用本文中所描述的系统和方法来支持计算机辅助的研究和知识发现。一般地,这由包括本文中所描述的各种架构特征的系统700支持。例如,知识图谱或类似构造通过用户设备704的用户界面702内的相关概念的明确视觉表示连同对支持文档和文本的便利访问来促进人类对文档语料库710的理解和交互。同时,包括例如推荐引擎、搜索引擎等的机器分析系统706基于可用文档语料库710内的实体和概念的关系来促进对新关系和洞察的发现。与此知识图谱的人类和机器交互可以由操作日志708或保存知识图谱的上下文和历史的类似数据结构支持,其可以一方面以支持人类用户之间的共享和协同的方式来聚集,以及另一方面通过机器算法改进上下文分析。

[0095] 以下描述了使用这种架构的机器辅助的知识发现平台的进一步特征和方面,其开始于对本文中所使用的术语的数个代表性描述。

[0096] 如本文中所使用的,推荐引擎可以是解读来自一个或多个人类用户的上下文信息并通过基于用户将选择作用于推荐的算法预测(例如,通过打开和读取推荐的文档)从文档语料库推荐内容来进行响应的任何系统或方法。

[0097] “用户”是与工具(诸如,程序、web服务或显示器)交互的人类或计算机过程。

[0098] “协同代理”(也被称为“智能助理”或“智能虚拟助理”)可以是通过一个或多个通信信道(诸如,电话上的语音对话、数据馈送、电子邮件或聊天应用中的文本消息)与用户交互的任何软件系统。典型的协同代理集中于单个信道的有限范围。在本文档中,与“协同代

理”可互换地使用“基于机器的用户”,因为人类和协同代理都是本文中设想的知识操作系统(KOS)的用户。

[0099] 如本文中所使用的表面(surface)可以包括用于与用户通信的任何媒介。表面可以包括用于与用户交互的任何设备、屏幕、应用窗口、声学环境、振动致动器和任何其他传感器-致动器机构。一种广泛使用的表面是计算机中的应用窗口,诸如,web浏览器、Microsoft Word应用、PDF查看器或人可通过其来阅读本文档的其他程序。“视觉表面”可以包括具有视觉组件的任何此类表面。

[0100] 通常,人类用户在一天内使用若干表面。例如,人们常常在膝上型计算机上同时打开电子邮件程序和web浏览器,并且还同时打开着具有正在进行的聊天对话的移动电话。这许多表面对于机器学习算法理解用户的行为具有巨大的价值。根据此上下文,协同代理可以尝试推断用户接下来需要什么。

[0101] 例如,通过推荐引擎“推荐内容”是指尝试帮助用户访问用户可能先前未意识到或认知到将对用户有帮助的内容。内容推荐的一个典型目标是在推荐显示中提供足够的信息以表征被推荐的一条内容的潜在效用。围绕推荐的这种上下文可以帮助人类用户认知到接受或使用推荐的益处。

[0102] 如本文中预期的“知识”至少包括概念以及在概念之间的关系。这可以包括实体之间的关系,这是一种特殊的概念。“实体”可以通过强类型化属性(诸如,家乡、电话号码、楼层数(number of stories)和DNA)来区分。命名实体具有给定名称,通常不是唯一的。实体的代词性提及是指先行词;例如,他、你。实体的名词性提及描述没有提供名称的特定实体,例如“将军”或“这三位新闻记者”。实体类别的实施例包括:个人、公司或组织、设施或建筑物、载具、设备或网络标识符(例如,电话号码、skype句柄、电子邮件地址、IP地址)、化合物、行星物体和蛋白质。

[0103] “文档”是包含概念的提及的数字媒体对象。非数字记录(诸如,文档的印刷形式)仅是人工制品,而不是如本文中一般描述的词语意义上的文档。文档出现在许多上下文中。这可以例如包括用户可用作参考材料的源文档和用户正在创建的工作笔记,诸如,电子邮件线程中的回复。

[0104] 软件系统经常以各种形式与提及交互,并且因此本文中所描述的是用于识别提及的一些概念和术语。“表面形式”是图像或另一原始数据表示中的字符串或声音序列或像素集,至少一些用户将其认知为引用概念。如上所述,采取以下立场:概念在上下文中仅由其表面形式限定。因此,为了识别概念,必须识别文档和作为提及的文档的一部分。许多文档包含丰富的结构(诸如,层级标题和表格),并且这些可以用于识别文档的部分。通常,表面形式提及是文档中的字符或像素的特定子范围,并且这些范围通常被称为“跨度(spans)”。已经开发了用于从文档中自动选择提及跨度的许多算法。

[0105] “文档语料库(corpus of documents)”是数据文件的集合,诸如,便携网络图形(PNG)格式的图像、或便携文档格式(PDF)或.docx格式的Microsoft Word文件、或以超文本标记语言(HTML)编写的web页面、或许多其他数据格式中的任何数据格式。语料库中的文档可以包含各种各样的数字媒体,包括图像、视频、音频、文本、时间序列、数值矩阵或可以二进制形式表达以使其可以在计算机系统之间传输的任何其他信息。甚至可以通过扫描纸张制作图像来将纸张文档表示为语料库,然后可以通过光学字符识别等将其转换为文本。文

档语料库的大小可以不受限制(诸如,其中非常频繁地创建新文档的公共Web)。

[0106] 文本文档通常包含可在音频文档中说出的词语。此类可说出的信息通常被称为自然语言,因为其结构是在计算机时代之前通过人类会话出现的自然现象。文档中的其他数据描述了用于由机器解读的结构化信息,诸如,超文本标记语言(HTML),其包含告知web浏览器如何显示信息的标签。典型的web页面包含结构化HTML以及自然语言两者。

[0107] 广泛使用的用于识别文本中的提及跨度的技术是将文档的整个文本描述为字符或Unicode代码点的单个连续阵列。然后,可以通过此代码点阵列中的开始和结束索引来唯一识别提及跨度。甚至可以将具有丰富结构的文档序列化为单个串。例如,web页面可以表示为HTML,并且来自HTML文档的串可以用作文档的单个连续串。为了在同一文档上运行多个提及跨度选择算法,通常有用的是确保其标识符使用相同的代码点阵列。这使得容易检测不同算法何时选择相同或重叠的提及跨度,并且其使得即使当不同算法选择跨度时也能够保留提及跨度的排序。例如,可以运行用于检测电子邮件地址的正则表达式和统计序列标记模型(诸如,跳链条件随机字段模型)两者,并且通过在相同的底层代码点阵列上运行它们,可以将通过这两个算法选择的提及跨度组合成单个集。如果特定算法无法正确地处理HTML标记标签,则那些标签可由空格代码点替换,使得后续自然语言文本的阵列位置不变。本公开文本可以将此类阵列中的索引位置称为“字符偏移”。例如,如果本公开文本将前一句子中的字母“W”视为位置零,则单词“refer”的字符偏移为[9,13](含)或[9,14)(其中结尾索引位置不包括在提及跨度内)。

[0108] 统一资源定位符(URL)广泛用于标识网络连接系统(诸如,万维网(WWW)和其他内容管理系统(CMS))中的文档。如各种标准文档(诸如,IETF RFC(https://tools.ietf.org/html/rfc3986 and https://tools.ietf.org/html/rfc1738))中所定义的,URL提供用于访问文档的通信“方案”、“主机名”和文件“路径”。URL可以包括附加数据元素,诸如,查询参数和哈希片段。实际上,URL字符串通常是可变长度的,并且因此可能引起软件系统中的操作问题。用于处理这些问题的常见技术是应用具有低冲突率的单向哈希函数(诸如,MD5、SHA1或SHA256),其生成可容易地以十六进制表示的固定长度字符串。由于哈希极不频繁地冲突,因此经常将它们当作文档的唯一标识符来使用。本公开文本将此类基于哈希的标识符HashDocID以及文档的通常任何种类的基于字符串的标识符称为“docID”。

[0109] 因此,可以通过docID和标识所述文档内的提及跨度的开始和结束的字符偏移的组合来标识文本提及。此类组合被称为“提及标识符”或“mentionID”。例如,再次考虑开始于“We often refer…”的示例句子。想象仅包含这一个句子的文档,并且假设所述文档具有URL“file://server/doc.txt”,其中MD5哈希为da98292ac64ea61dc4706dac2a48881a。因此,所述文档中的单词“refer”的mentionID为“da98292ac64ea61dc4706dac2a48881a:9,14”,其中冒号用于从字符偏移开始到结束对docID进行定界。

[0110] 类似的考虑为图片和视频中的子图像以及音频轨道中的话语提供有用的具体标识符。这些提及的标识符被称为mentionID。实体是概念的子集,并且因此实体通过其mentionID来标识。出于描述本文中的实施方案的目的,提及是将需要被描述或解决的实体的唯一实体化或表征化。为了清楚起见,为了引用特定实体的mentionID(诸如,歌手Black Frances),系统可以写入“MentionID(Black Francis)”。这可以使得构造使用有关mentionID的阐明实施例更加简单。

[0111] 关系可以是实体的一种属性。关系可以由作为概念的两个实体标识,并且因此通过提及来证实(对于这两个实体中的每一个的至少一个提及)。即,一对mentionID可以是用于标识关系的必要信息。

[0112] 关系可以被描述为具有由通常称为本体(ontology)的特殊种类的文档定义的“类型”。对本体中的各种概念的提及(诸如,“…的成员”)可以用于描述实体之间的关系,诸如,“Black Francis曾是Pixies的成员”。(MentionID(Black Francis),MentionID(The Pixies))的二元组可以是更基本的对象。为了引用具有类型的关系,可以同时识别标识所述一对实体的二元组以及用于定义该关系类型的本体部分和本体文档的标识符。

[0113] 证实两个概念之间的关系的文档可以提供关系类型的证据。例如,考虑描述维生素C可以如何帮助人体对抗引起普通感冒(疾病)的病毒的期刊文章。此类文档提供了概念“维生素C”与概念“普通感冒”之间的关系的证据。所述文档中的其他概念的提及提供表征关系的上下文。

[0114] 共指消解或“coref”是将含义分派给提及的过程。一般地,具有相同表面形式名称的两个提及是可能引用同一概念的候选者。如果围绕这两个提及的上下文相似,则人类更有可能将所述提及感知为具有相同的含义,即,感知为是共指的。算法可以将围绕这两个提及的上下文中的数据进行比较,并且估计人类将所述提及感知为是共指的可能性。

[0115] 这个过程可以在解读提及的基于人或机器的代理中发生。所述过程可以通过断言某些提及引用相同的概念(即,共同引用)来减少歧义。行动者在为每一个提及分派含义时都经历了这种歧义的减少。这很微妙,因为没有人可以直接触及含义或事实。相反,可以从证据推断事实。因此,可能无法获得分派含义的字面含义。相反,消歧的可操作性概念是共指消解。如果行动者相信特定提及与另一提及引用同一概念,则行动者相信这两个提及是共同引用或共指的,即,具有相同的含义。在已经通过共指消解和其一些关系的理解将X与其他概念相关联时,人们说“我理解X”或“我知道X的含义”。将提及与同一概念的其他提及联系起来是为了解决其含义或为了将其“coref”,在此之后它已经被“coref”。

[0116] 代理的coref带有额外的权重,因为它识别行动者。身份对于真实世界中的行动者可以具有许多实际意义。由于许多在线网络行动者隐匿其身份,所以特定术语“角色”可用于指代另一行动者可能尚未准确地融入其他角色的在线行动者。

[0117] 在发现关于概念和实体的信息时,可能遇到若干类型的文档。例如,这可以包括“一级源”文档,其通常是与各种事件同时代的为行动者之间的通信而创建的人工制品(例如,事件的照片或采访的转录本)。这还可以包括“二级源”,其通常在取决于或聚焦于先前事件的后续事件期间分析一级源文档,例如,典型的智能公告,诸如,记录单个一级源文档的单个源报告文档(诸如,对人的采访或对纸质人工制品的扫描)。报纸文章是二级源材料的常见实施例。“三级源”通常编译来自二级源文档的信息以提供那些分析的概要或简缩,例如,正确编写的百科全书文章(诸如,许多维基百科文章)。

[0118] 如本文中所使用的,上下文(context)可以是文档的一部分,其使得能够或帮助人理解该上下文内提及的预期通信。上下文是可变的。通常,包括文档的较大部分作为上下文将使得读者能够理解得更多。另一方面,较大的上下文需要更多的时间来由人类读者消化和由机器进行处理。例如,搜索结果列表中的文本的片段旨在给读者足够的上下文来有效地决定是否打开链接以访问完整的文档。将理解的是,如本文中所使用的,上下文还可以包

括在其中采取动作的用户或计算机上下文。这可以例如包括作出请求的设备或应用以及任何其他可观察的物理上下文(一天中的时间)、计算上下文(网络地址、执行过程等)、研究上下文(知识图谱中的位置、信息请求的类型等)等。

[0119] 引证(citation)是对文献的参考。URL是一种引证。然而,引证还可以或替代地为读者提供任何信息以对文档的特定版本消岐。当文档包含对另一文档的引证时,引证的上下文可以指示在这两者中提及的一些概念。

[0120] 本文中设想的这个系统可以采用效用函数,所述效用函数一般可以适于模仿用户评估潜在有用信息的源的隐式效用函数。在一个方面中,这可以动态地模仿特定用户的研究活动,使得机器可以通过动态地优化跟踪用户效用函数的效用函数来与用户协同。

[0121] 这可以包括捕捉(文档内)提及周围的上下文,并且进一步捕捉来自用户的关于特定提及是否推动了用户的兴趣的反馈。这意味着软件代理的效用函数可以由用户的当前工作上下文来定义。具体地,软件代理的目标函数可以是找到在用户的工作上下文中提及的实体与尚未在用户的工作上下文中的其他实体之间的关系的证据。通过动态地对此类相关实体和相关内容进行重新排位,所述系统可以鼓励用户将这些文档添加(引用)到用户的当前项目中。如本文中设想的推荐引擎或其他计算机辅助的发现工具可以因此模仿查询远程系统、取得文档、概要内容、核对实体(corefing entities)、以及寻找填补人类协同者工作上下文之间的空白的人工过程。

[0122] 知识图谱或其他视觉表示可以有用地采用多种视觉显示元素。在一个方面中,这可以包括表示文档或概念的顶点(例如,作为图标),以及表示文档和概念之中的提及(诸如由各种文档的内容证实的提及)的边。以下讨论多种附加的、有用的视觉显示元素或视觉知识元素。

[0123] 1.上下文中的提及突出显示:最接近文本的可以是视觉突出显示,其在文本显示系统领域中被广泛使用。视觉突出显示可以改变构成出于某种目的感兴趣的提及跨度的字符跨度中的字符的颜色或色彩对比度或其他视觉方面。突出显示还可以或替代地用在图像上,其中可以在图像的顶部上呈现形状或多边形的轮廓以标识感兴趣的子图像或其他图像内容。音频轨道可以用由任何适当的视觉、音频或文本(例如,通过注释)装置标识的子片段来分段。视频中的概念的提及的突出显示也可以是直白的,例如,利用视觉和音频指示符。对于上下文中的提及的突出显示,可以将完整的源文档加载到查看器或编辑器中。这通常是文档的原生形式。例如,web浏览器可以显示从某个URL加载的HTML web页面或PDF文档,并且然后可以将突出显示插入文档的文档对象模型(DOM)中。例如,在作为DARPA Memex的一部分创建并且在GitHub上以“dossier/html-highlighter.”开源的HTML突出显示项目中描述了一种适当的突出显示工具。这是将突出显示插入文档的上下文中的工具的实施例。其他突出显示工具是可公开获得的,并且本领域普通技术人员将容易理解适用于突出显示文档内的概念提及的各种突出显示的资源和技术。

[0124] 2.概要卡:搜索结果片段的列表是所谓的“队列中的概要卡”的熟悉形式。其他熟悉形式的概要卡是Trello或Jira或Zenhub的Kanban式工作板中的可拖动卡,以及GitHub上的插件。此类基于卡的显示可以示出内容的一部分,所述部分是从文档中提取的或者由概要生成算法创建为摘要并且被显示为使得用户可以看到文档的表示,其中在文档中提及的概念的子集以紧凑形式向用户突出显示。通常,这种紧凑形式使得用户能够在视觉显示器

上在一个视图中同时查看许多文档的表示。例如,概要卡可以示出在文档或若干文档中提及的一个或多个概念的表面形式。概要卡可以代表多个文档;例如,卡可以示出若干文档均一起提及相同的两个概念,因而提供概念之间的关系的证据。概要卡还可以示出图像或子图像或者视频或音轨的部分。短语“片段(snippet)”可以用于指代从文档中提取或摘要并且在概要卡中显示的内容部分。可以在列表中、或选项卡中、或多列显示中显示概要卡的视觉显示。概要卡的列表可以被称为“队列”,因为卡上的用户动作可以通过从队列顶部附近的位置移除卡来对列表进行分类。

[0125] 3.图谱中的顶点:可以用由线连接的图标或符号集合来呈现来自文档的信息的视觉表示。此类图谱显示中的图标可以表示特定概念或概念集合。图标可以表示文档或文档的特定版次或版本、特定请求者在特定时间对来自URL的内容的快照、或任何其他文档或文档部分。图标还可以或替代地表示基于来自一个或多个文档的提及集合的概念。例如,本公开文本提及概念“图谱”,并且可以通过用线将其连接到用表面形式“图形”标记的概念顶点来呈现文档图标,从而将该关系的视觉显示描绘为图谱中的顶点。图谱上的标记可以与上下文中突出显示和概要卡为用户提供上下文的方式大致相同的方式为用户提供视觉上下文。此类图谱显示中的线可以具有各种视觉质量、属性、标记和其他符号。例如,将文档图标连接到文档中提及的概念的线可以是虚线,其中箭头指向提及的概念。连接到在多个文档中一起提及的概念的线可以被描绘为具有分派以不同颜色的各种含义的实线或有色线。

[0126] 上下文中突出显示可以标识上下文中的提及跨度。概要卡可以标识在描绘多个文档的视图中感兴趣(或潜在感兴趣)的提及。图谱可以使用顶点来表示概念/提及和文档,其中边描绘这些项之间的关系。全部三个级别的视觉抽象在本文中可以被称为视觉知识元素,并且可以支持知识/信息以及用户输入(诸如,识别感兴趣的提及、消除用户决定不感兴趣的提及、请求推荐等的输入等)的显示。这些相同的视觉知识元素还可以实现与知识结构的自动化或基于计算机的交互,并且支持操作(诸如,对用户可能感兴趣的提及的基于机器的选择和呈现、以及对用户界面或其他视觉上下文内的提及的呈现的动态更新)。一般地,系统可以向用户显示的这些不同的视觉知识元素是由处理来自日志的知识操作记录(KOR)的客户端应用生成的聚合。例如,概要卡队列可以由处理来自日志的操作的用户界面显示应用来构造。因此,更一般地,如本文中设想的知识操作系统可以支持将相关动作和文档组织到操作日志中,以及使用日志中的数据聚合的所得知识结构的不同表达。

[0127] 前述视觉元素可以支持知识结构中的信息的呈现。所述系统还可以包括控件以支持此上下文中的不同动作。例如,按钮或其他控件(诸如,下拉菜单、拖放操作、刻度盘、滑块等)可以用于使得用户能够请求动作。例如,电子邮件程序可以呈现标记为打开新的草稿电子邮件文档的“撰写”或用于发送电子邮件的“发送”的按钮。此类视觉动作按钮可以不同于视觉知识元素,因为前者由它们提供给用户以发起的动作来标记,而后者由用户可以读取或访问的源数据来标记。

[0128] 在一个示例性实现方式中,当用户将鼠标指针带到具有上下文中突出显示的提及跨度附近时,可以显示诸如图标等视觉动作按钮,并且如果用户点击图标,则系统可以激活对所提及的概念的附加处理。用户还可以选择字符、图像、音频或其他数字媒体的跨度,即使其没有被视觉知识元素具体识别。如果用户选择此类自定义提及跨度,则系统还可以显示视觉动作按钮,使得用户可以基于用户选择的跨度来发起动作。

[0129] 在一个方面中,上下文跟踪可以用于促进知识发现。基于机器的协同代理可以通过从多个文档接收用户动作来跟踪用户的工作上下文。这种跟踪可以被记录在如本文中所描述的操作日志中,所述操作日志可以例如被实现为知识操作的仅追加记录,其中知识操作是基于人或机器的代理在语料库内可能采取的任何动作。在以下讨论中,这种操作日志也被称为“知识操作日志”或“KOJ”。此类日志可以通过定义已知操作的列表并且然后将带有时间戳的观察到的操作记录在数据库中来实现。如上所述,每个记录还可以或替代地包括用户上下文和与所记录的操作相关的其他信息。因此,带有时间戳的操作的数据库定义项目,所述项目有时被等效地称为“知识项目”。

[0130] 作为显著的优点,以这种方式实现的知识项目可以允许与知识项目相关的用户活动跨多个文档、应用、设备等一起被跟踪。这还可以使本文中所描述的不同类型的视觉知识元素能够被同步和协调。例如,考虑某个用户可能在类似Firefox等网络浏览器中打开了一个或多个文档,并且还在Microsoft Word中打开了一个或多个文档,并且还在Outlook中打开了一个或多个电子邮件,而且还在Gmail客户端中打开了其他电子邮件,而且还在其他工具中打开了其他文件。如果用户邀请协同代理(诸如,推荐引擎或其他机器辅助工具)进入这些文档中的若干文档中,则知识项目可以跟踪这些文档中提及的概念并可以对用户操作作出反应。这还可以结合其他设备(诸如,平板计算机、智能电话等)上的用户动作。用户还可以发起多个知识项目并且将各种文档添加到各种项目中。文档可以在一个或多个项目中,并且多个用户和/或基于机器的协同代理可以共享项目。

[0131] 将文档添加到项目可以是对相应的知识操作日志的操作。在一个示例性实现方式中,OpDocumentAdd操作触发协同代理的算法,以自动分析文档以选择提及跨度,并且通过OpConceptAdd操作将概念添加到项目。

[0132] 用户动作

[0133] 对视觉知识元素的用户动作的实施例可以包括以下。

[0134] 1.肯定动作:用户可以选择用于跟踪的提及。此类动作可以通过若干不同的视觉知识元素来接收。用户可以点击上下文中突出显示或在突出显示附近显示的视觉动作按钮。类似地,用户可以点击提及在概要文档中提及的概念的概要卡。类似地,用户可以点击视觉图谱中的顶点。

[0135] 2.否定动作:用户可以通过点击概要卡上的视觉动作按钮来拒绝来自协同代理的提议,所述视觉动作按钮可以用所述动作的描述来标记,诸如“错误实体”或“不感兴趣”。类似地,用户可以拒绝红色的上下文中突出显示以告诉算法特定提议没有帮助。类似地,用户可以丢弃视觉图谱显示中的顶点。具体实现方式可以提供多个丢弃动作,从而允许用户在给定的顶点或边或证据上传达更多有关其感受的信息。

[0136] 3.中性丢弃动作:用户可以修剪自动生成的视觉知识元素,而不必告诉系统它们是错误的。这些非肯定且非否定的动作可以允许用户通过去激活突出显示、移除概要卡或从图谱显示移除顶点或边来清理视野。

[0137] 4.过程发起动作:可以在视觉知识元素上向用户曝光特定命令和控制过程。例如,示例性实现方式可以提供“自动构建器”按钮,所述按钮允许用户告诉系统他们对一组特定的多个概念感兴趣,并且因此请求协同代理对这些概念进行所有可用的探索动作。

[0138] 5.探索动作:基于机器的用户可以通过自动制定使用提及的表面形式或表面形式

的变异形式或从提及派生的替代词的查询,然后将这些查询发送到远程搜索引擎以取得文档来作用于提及,它可以贯穿下面描述的各个索引层,并最终建议将这些文档中的某些提及内容添加到项目中。

[0139] 6.推荐提及:基于机器的用户可以通过推荐其估计很可能共指的其他提及来作用于提及。

[0140] 7.推荐相关实体:基于机器的用户可以通过推荐与很可能与所述提及共指的提及共同出现的其他实体的提及来作用于提及。

[0141] 8.推荐动作:基于人类或基于机器的协同代理可以推荐和采取非常宽范围的动作。在一个方面中,由基于机器的用户提议或采取的动作响应于提及而发生。

[0142] 一般地,可以将用户动作添加到知识操作日志。这可以使得基于机器的协同代理能够应用算法来维持与人类用户关于他们寻求的知识或动作的对话的历史。日志可以使得基于机器的工具能够以推进用户的一个或多个研究目标的方式更好地推断用户的意图。

[0143] 肯定动作和否定动作的集可以包括指示两个提及跨度是或者不是同一实体。例如,当概要卡从用户的文档中识别出提及跨度(诸如,名为John Smith的人的提及)时,则协同代理算法可以在一些其他文档中找到“John Smith”的提及并且呈现关于该文档的概要卡。通过读取在概要卡中提供的或以其他文档的全视图形式提供的上下文信息,用户可以判定在其他文档中的提及与在用户的文档中提及的John Smith不是指同一个人。为了指示这种“错误实体”断言,用户可以点击显示负号符号的按钮。机器可以使用这种反馈来改进其在未来处理中的共指消解处理。相反,如果用户判定其是同一人,则用户可以通过点击示出正号符号的视觉动作按钮来将此作为反馈提供给机器。此类动作可以创建存储在知识日志中的操作中的肯定和否定共指断言。

[0144] 系统可以从其他用户动作推断coref断言。例如,继续上述John Smith的实施例,考虑用户正在编辑正在进行的文档(诸如,草稿电子邮件),并且使用关于其他文档的概要卡来创建引证。此类引证动作可以通过操作来记录在知识日志中,所述操作可以被命名为OpCitationAdd。在随后的处理中,系统可以将该操作解读为意味着用户也相信在其他文档中提及的John Smith与用户的文档也涉及的是同一个John Smith,即,这两个提及是共指的。

[0145] 索引层

[0146] 若干数据结构和索引可以组合使用,以如本文中设想的知识操作系统内提供文档搜索、概念搜索和关联、对实况知识结构的跟踪等。知识操作系统的其他方面(诸如,知识图谱、推荐列表、边或顶点信息等)可以通过聚集来自这些索引的记录来创建,并且具体地,聚集来自以下描述的存储跨多个表面的人类和计算机动作的日志的tier3索引的记录来创建。

[0147] “Tier0索引”可以是语料库中每个文档的全部内容的标准索引。在一个方面中,这个索引(或多个索引)可以提供用于查找和取得文档的快速方法。例如,反向索引可以使得支持关键字和短语查询的所谓的“全文”搜索能够找到提及由用户输入的词语或短语的文档。基于关键字的索引可以基于与文档相关联的特定值(诸如,元数据)的确切匹配查找来实现查找。与文档和知识操作日志一起使用这些技术可以实现如下所述的知识操作日志的有利能力。如本文中所使用的,“标准索引”是指基于全文和基于关键字的取得技术中的任

一者或两者。

[0148] “Tier1索引”是每个提及链子文档的标准索引。可以通过NER和其他提及选择算法来计算附加特征数据,并且这些提及跨度被用作用每个提及链子文档索引的上下文的一部分,因此使得能够进行更复杂的取得查询,诸如,高效地取得提及彼此接近的两个名称的文档。命名实体识别(NER)和其他自然语言处理(NLP)工具在本领域广泛用于生成关于所有类型的文档的元数据。算法可以用于将数据流分段和标记文本、图像、音频和视频中的概念和实体的提及。系统可以使用这些丰富算法来识别对文档中感兴趣的提及集,所述提及集在本文中被称为“提及链”。这些可以被构建到具体提及周围的MC或“提及链子文档”。这些子文档包含围绕提及的上下文,使得用户(基于人或机器)可以理解提及的含义。

[0149] “Tier2索引”由被称为层级凝聚群集(HAC)的算法生成,所述算法将提及链子文档一起分组成嵌套的提及集。更接近层级中的叶级的集可以包含算法估计更可能共指的元素。例如,具有相同表面形式名称的两个提及是可能引用同一概念的候选者。如果围绕这两个提及的上下文相似,则人类更有可能将所述提及感知为具有相同的含义,即,感知为是共指的。算法可以将围绕这两个提及的上下文中的数据进行比较,并且估计人类将所述提及感知为是共指的可能性。由HAC生成的结构化树可以提供快速查找数据结构,以用于找到对同一概念的提及。树结构可以被存储为提及链子文档上的基于关键字的元数据。例如,如果将MC分派给包含在子集97内的叶集34,所述子集本身包含在本身包含在子集14内的子集104内,则所述MC的树地址是[14,104,97,34],其是有序关键字集。这些关键字作为用于快速查找的取得关键字来存储在搜索索引中的MC记录上。

[0150] 模型可以定义每个嵌套集的共指可能性的阈值。例如,被定义为最接近叶级的集很可能是90%、接下来是80%、接下来是70%等。远离叶级的集可能更不确定。参考上面的实施例,具有树地址[14,104,3,402]的另一叶在集104中具有最小公共祖先(LCA),并且因此将有70%的可能是共指的。

[0151] 这可以实现如下高效取得:在HAC已对包括来自用户文档的至少一个的提及链子文档的集合进行操作之后,则可通过请求共享祖先关键字的文档来取得可能引用在用户文档中提及的同一实体的其他提及。

[0152] Tier3索引可以实现跨用户的许多窗口、应用、设备和同类知识工作者(fellow knowledge workers)的协同同步。一般地,Tier3索引可以包含用户动作的记录的序列,包括如例如存储在操作日志或类似数据结构中的跨多个表面的一个或多个人类用户和一个或多个机器用户。可以使用Tier3索引的聚合来导出知识操作系统的各种特征(诸如,知识图谱的显示或相关联的视觉显示元素(诸如,卡或列表))。

[0153] 致动器框架

[0154] 致动器框架可以使得系统能够找到数据并且将其拉入项目中。

[0155] 元搜索是致动器框架的一种特定实现方式。元搜索通过将查询推送到支持数据库的网站上的搜索表单栏中并且从由支持数据库的索引返回的搜索结果中的链接中取得文档来爬取所谓的深度web。元搜索处理操作日志的历史以找到提及串并且根据这些串来制定查询。元搜索致动器可以使协同代理算法能够将那些查询串推送到搜索引擎(诸如,用于联邦机构(例如,证券交易委员会或美国专利局)的搜索引擎)、商业数据库等中以发现文档。这个过程可以是自动的以消除对人类动作和监督的需要。通过自动地跟踪在搜索引擎

结果页面(SERP)中返回的链接,元搜索致动器框架可以获得它可以添加到项目的附加文档。

[0156] 致动器框架的另一个实施例是主动wikifier系统。wikifier是将实体的提及链接到参考知识库(诸如,维基)的算法。如本文中设想的系统可以通过元搜索众所周知的参考源和社交媒体中的简档来动态地响应于用户添加到项目的提及,使得其可将用户的提及链接到那些外部源。所述系统可以使用公共维基化目标的缓存,使得它不需要针对最常见的概念进行元搜索。然而,动态wikifier执行器框架可以使得协同代理能够跟上遍及web的维基资源的演进状态。

[0157] 知识操作系统

[0158] 知识操作系统(KOS)是促进KOJ(本文中也被称为“项目”)上的各种动作的内容管理系统(CMS)。用户可以创建新的项目,给它名称,并且将来自文档的提及添加到项目。每个项目在可以持久地存储在盘上的可串行化数据结构中具有单独的表征化。这样,项目就像新类型的文档。以与传统文档可以经由引证或结构化引用(诸如,URL)来引用其他文档相同的方式,KOJ还可以包含对许多其他文档的引用。与大多数其他类型的文档不同,KOJ可以跟踪与其引用的文档相关的用户动作。与任何其他种类的文档不同,KOJ可以使得算法能够使用那些跟踪的动作来自动扩展KOJ本身。

[0159] KOS可以提供编程接口(诸如,传输JavaScript对象符号的代表性状态转移web服务),其使得客户端程序能够与表面交互,从而可以跟踪其他文档上的用户动作。KOS还可以提供致动器框架,所述致动器框架使得自动算法能够从不同的数据源取得文档和数据,使得可以将其添加到KOJ。KOS web服务可以使得多个表面能够显示呈现来自同一项目的信息的视觉知识元素。各种表面可以重复地轮询KOS接口,或者KOS可以使用使其能够将改变推送到日志的协议。

[0160] KOS可以有利地使得用户能够跨设备和应用来控制和管理项目,并且能够在知识发现方面与人和机器资源进行协同。与传统搜索引擎中的搜索历史不同,KOS可以使得用户能够进行多个同时的项目。用户可以选择在哪里记录探索动作,使得每一项目准确地跟踪特定活动或努力区域的用户BDI。虽然可以尝试通过仅查看用户的点击和类似动作来推断用户正在从事哪个项目,但这并不是人与另一个人共事的方式。如果两个人正在不同项目上一起工作,则一个人通常告诉另一个人他/她正在考虑哪个项目。人们经常听到队友或同事说“现在让我们将话题切换到另一个项目…”或“正如你将回想到的,对于这次即将到来的会议,重要因素是…”和其他上下文设置语句。因此,用户一般可以指定项目。在另一个方面中,KOS可以透明地(例如,没有用户通知)或交互地(例如,具有明确的用户通知和请求)或其某种组合来推断项目和/或在多个话题之间进行切换。

[0161] 如今,计算机的大多数人类用户习惯于操纵计算机软件,就好像它是没有思想的无生命工具一样。随着启用智能虚拟助理的算法的出现,软件可能需要从人类协同者征求并接收此类项目上下文设置信号的能力。KOS可以启用此类上下文设置。

[0162] 在一个实现方式中,在KOJ中记录每个用户动作。用户动作发生在诸如MS Word或web浏览器中的选项卡的表面上。用户动作可以由在这些应用内部运行的软件插件或由在父操作系统中运行的屏幕图像捕捉软件来捕捉。下文描述用于设置上下文的方法的一些示例性实现方式。

[0163] 为了使得用户能够指定哪个KOJ将接收动作,KOS可以显示用于选择KOJ的视觉动作按钮。在一个实现方式中,显示单独的此类视觉动作按钮以用于在每个web浏览器选项卡中(例如,在Chrome或Firefox中)选择项目,或者显示用于表面或设备的单个视觉动作按钮。KOS可以作为桌面应用等部署在特定用户设备上,并且可以使得用户能够将基于机器的协同代理邀请到每个文档、应用、窗口等中。在一个方面中,用户可以将表示每个新文档的图标拖放到不同的项目中,或者将表示项目的图标拖放到用户希望与所述项目相关联的每个新文档、窗口或应用中。

[0164] 在一个方面中,用户可以例如在KOS桌面应用的软件偏好面板中激活自动项目切换。在这个开关被设置为“开”时,则当用户在一个窗口或应用中选择项目时,则在别处打开的与KOS项目相关联的所有其他文档或窗口也可以自动地切换到所述项目。这可以节省用户手动地将每个窗口切换到新项目的麻烦。这种自动项目切换可能具有的不利之处在于,其使得用户更难以在同一时刻使多个项目活动,因此可能容易禁用自动切换特征或者可能提供条件切换特征(其中在检测到新项目或项目切换时询问用户是否将其他窗口或文档切换到新项目)。

[0165] 在另一个示例性实现方式中,用户可以从搜索结果列表中选择文档并且点击用于“添加到项目”的按钮。所述系统可以显示示出现有项目的名称的选择器,因此用户可以选择哪个项目将接收来自所选文档的提及。所述系统还可以使得用户能够开始新项目。在这个工作流中,当用户点击以确认哪个项目或命名新项目时,系统可以自动采取两个动作:(1)其打开草稿电子邮件以帮助用户与同事共享项目;以及(2)其打开项目的图谱视图中的顶点,包括文档和其他顶点。在打开图谱视图时,如果用户将提及或串识别为其搜索过程的一部分,则那些概念的顶点可以自动扩展以示出相关概念。

[0166] 所述系统可以提供示出关于每个项目的概要视图的项目仪表板(dashboard)。概要视图可以列出项目中的一些文档和概念,并且提供用于打开项目的图谱视图中的顶点的链接。在一个方面中,可以将项目共享,使得多个用户参与单个项目。例如,两个人类用户和两个不同的软件系统可以向单个日志贡献操作。

[0167] 如本文中设想的日志可以是活跃文档,因为它可以继续累积新操作。与其他活跃文档一样,日志可以分叉以创建随后与原稿不同的副本。日志也可以被锁定,使得不能添加新操作。日志可以具有若干形式的访问控制权限。可以通过一个访问控制权限集来定义可以用只读权限访问日志的用户,并且可以通过不同的访问控制权限集来定义可以用写入权限访问日志的用户。可以将“所有者”用户集定义为具有改变只读用户列表和写入允许用户列表的权利。

[0168] 为了授权协同代理算法代表用户行动,可以给日志授予访问权限以供代理使用。这些访问权限可以被称为“日志的访问权利”。此类权限可能要求它们来自具有对日志的控制的用户,因为日志不将成为恶意用户获取对所述用户他/她自己无法访问的数据的访问的攻击手段。日志可以允许用户将特定访问权利分派给在日志上操作的若干协同代理。

[0169] 例如,当用户从在搜索结果列表中找到的文档开始项目时,用户可能已经利用访问权限集来执行所述搜索。默认地,可以向日志授予那些相同的访问权限作为其访问权利,以使得对日志中的数据作出反应的协同代理可以继续搜索相关内容以代表用户扩展项目。由于用户可能希望与其他用户共享项目,所以KOS可以使得用户能够对于日志的访问权利

选择更受限的访问权限集。以此方式,被邀请到所述项目的其他人可能不能获得对所述项目的原始创建者可以访问的所有相同数据的访问。

[0170] 例如,日志的原始创建者可能已经邀请协同代理读取和索引用户的所有电子邮件。当用户决定围绕特定问题制作另一日志时,则用户可以限制所述日志可以使用的权限,使得被邀请到所述项目中的其他人不能通过所述日志来访问用户的电子邮件。

[0171] 在另一方面中,日志可以被配置成支持多个用户的异步修改。以下实施例提供了关于使用实体和关系的图谱的用于协同式机器辅助知识发现的用户界面的示例实施方案的细节。在用户双击概念顶点以请求与所述顶点相关的更多概念时,可以发生以下一系列事件。首先,用户界面可以例如通过桌面应用或支持知识图谱的显示的其他本地软件代理从机器分析系统706请求顶点的相关概念。机器分析系统706(或支持KOS的其他适当的后端)可以生产工作单元,所述工作单元旨在完成找到相关概念并且将其添加到图谱的请求。这个工作单元异步地执行。机器分析系统706可以包括指示用户何时发起对相关概念的请求的时间戳的方式将请求提交到其操作记录。所述提交还可以包括对于工作单元的标识符。机器分析系统706可以利用自从用户界面702的最后呈现以来提交给记录的所有操作(包括上述提交)来响应于请求。客户端可以使用新操作来在特定顶点上呈现加载图标。注意,由于"加载"状态是图谱记录的一部分,因此即使用户关闭图谱并将其重新打开,这个加载图标也可以有效保留。由于所提交的操作包含时间戳,因此用户界面702可以在预定的超时之后省略加载图标。

[0172] 异步地,工作单元可以开始执行,并且机器分析系统706可以执行对相关概念的搜索。然后可以获取对操作记录的写入锁。当这个锁被保持时,不能将其他过程直接写入记录。注意,这不影响用户界面或图谱对于任何实际或潜在用户的可用性。这个写入锁可以在执行全图谱去重复操作以防止数据竞争的任何时候是必要的。在锁就位的情况下,可以针对当前图谱对任何相关概念进行去重复。在一个方面中,这个过程可以通过使用名称相似性试探法等来实现。任何新的概念顶点和证据边可以通过写入操作记录而在图谱中创建。默认情况下,可以隐藏所有新边。然后可以释放写入锁。

[0173] 异步地,用户界面下一次从后端接收更新(通过轮询或通过类似于1-4的另一操作)时,它可以包括在操作记录中添加的附加顶点和边。由于它们被默认隐藏,因此用户界面可以在顶点上显示表明存在多少隐藏边的指示符。当用户要求对其进行查看时,用户界面(或更精确地,控制用户界面的过程)可以执行示出其的操作。

[0174] 写入锁的简单实现方式可以是获取操作日志中的表级ROW EXCLUSIVEPostgreSQL锁。然而,这可以阻止对所有图谱而非一个特定图谱的写入请求。在另一方面中,此问题可以通过将隔离模式设定为SERIALIZABLE来解决,但是这可能会导致重复处理,例如,最终用户快速连续地双击两个不同的顶点。在另一方面中,所述系统可以使用PostgreSQL事务级咨询锁来在日志上实现写入锁,以支持多个用户的并发使用(concurrent use)。

[0175] 知识操作日志

[0176] 知识操作日志(KOJ)可以是知识操作记录(KOR)的集合。所述集合可以被排序到仅追加记录中。这个操作记录可以形成由所述日志的用户采取的探索事件或面向知识的动作的记录。

[0177] 操作转换日志是本领域中的标准。诸如Etherpad、Google Wave、Google Docs、

Office 365和Dropbox Paper等协同编辑工具使用此类日志结构来跟踪多人可编辑文档中的编辑事件。为了使得此类系统能够起作用,计算机科学的子领域已经合并以研究无冲突的复制数据类型(CRDT)。如同CRDT和相关操作日志领域中的标准一样,可以通过形成检查点操作来简化日志中的数据,所述检查点操作总结直到所述点的先前操作的聚集效果。这可以使得日志的客户端能够不必重新处理整个操作历史;相反,客户端可以从最近的检查点开始,并且通过进行后续操作来更新其对日志状态的视图。

[0178] KOR可以有用地包括用于每个动作记录的字段,诸如,用户的用于采取动作的标识符、用于来自语料库的文档中的提及的标识符、以及在适当或有帮助的情况下关于引用所识别的提及的视觉知识元素的动作。KOR的各方面(诸如,用户标识符)可以用于以示出特定用户的特定贡献的方式来呈现项目的知识图谱或其他视觉呈现。因此,例如,在多个人类用户添加到项目的情况下,由每个用户添加的顶点或边可以被颜色编码、被用户图标或化身(avatar)标记、或以其他方式被视觉地编码以标识特定添加的源。机器用户的活动还可以或者替代地以这种方式被视觉地编码。类似地,修改权限可以取决于添加的用户,例如,用户请求对顶点进行改变的所述顶点。因此,例如,用户可以禁止、允许或有条件地允许改变对所述用户条目的修改。

[0179] 一般地,本文中设想的用户可以包括人类用户和机器用户两者。例如,机器用户可以包括执行呈现提及“John Smith”的概要卡的动作的基于机器的算法,并且人类用户可以执行点击所述概要卡的动作,例如,以请求关于其的更多信息。

[0180] 通过收集KOR的集合,基于人和机器的用户可以跨许多表面查看所有用户的过去动作。进一步地,动作和跨多个KOR的动作可以促进用户意图的算法推断,使得机器用户可以采取进一步动作,这些进一步动作推进人类研究兴趣和/或在所述日志上产生更多KOR。

[0181] 在示例性实现方式中,被称为“OpEdgeSubstantiateStart”和“OpEdgeSubstantiateEnd”的操作实现关系证据包(relation evid ence bundles)。将表达对关系感兴趣的用户动作记录为OpEdgeSub stantiateStart操作。这种操作可以携带用户想要看到证据的关系中的两个概念的标识符。协同代理通过运行算法以找到证据来对这种操作作出反应。当其找到证据时,其将证据包存储在存储系统中。然后,其利用指向存储位置的指针将OpEdgeSubstantiateEnd操作添加到日志。处理日志的客户端程序然后可以通过使用指针访问所存储的证据来显示证据。这两个操作可以在日志中形成组。撤销操作可以携带指向操作或操作组的指针,使得操作的分组便于撤销。引述操作使得用户能够将来自文档的段落附加到日志。这使得用户能够对包含许多提及的数据的较长部分采取动作。

[0182] 通过处理KOR的日志,过程可以创建表示机器和人类用户的动作历史的聚合。为了构造聚合,系统可以处理日志中的操作序列。然后,检查点操作可以将所述聚合的结果存储在日志上作为过去操作的某些方面的概要。存在许多可以由单个日志形成的可能的聚合。例如,可以将一个用户的所有动作收集到所述一个用户的活动和/或意图的描述中。替代性地,可以使用多个用户的操作来捕捉指示例如一组用户的想法过程或研究目标的一组意图。

[0183] 在一个示例性实现方式中,用于聚合操作以对一个或多个用户的意图进行建模的机制如下。系统可以生成描述围绕每个提及的上下文的特征向量。向量可以是高维向量空

间的一部分,其中在语料库中提及的每个词语、短语或概念被视为向量空间的维度。为了形成对于给定用户集的聚合,系统可以创建概要向量(summary vector)。最初,概要向量是零向量。系统然后对日志中的操作进行迭代,从而过滤由给定集中的用户进行的操作。如果通过过滤器的KOR携带用于在语料库中的提及的标识符,则系统可以使用所述提及的特征向量来细化概要向量。如果操作携带肯定的用户动作,则提及的特征向量将添加到概要向量。如果操作携带否定的用户动作,则将从概要向量中减去提及的特征向量。如果提及携带中性或模糊的用户动作或没有用户动作,则系统可以使用以加权系数将提及的特征向量添加到概要向量,所述加权系数改变概要向量上的改变的幅度和符号。

[0184] 概要向量可以提供对用户的当前知识状态进行建模的数据结构。在概要向量中具有非零分量的向量空间的维度可以描述用户的当前知识状态。具有正值的维度可以描述用户期望的知识。系统可以使用从日志聚集的此概要向量来对提及推荐和相关概念推荐进行排位。系统可以通过使用在概要特征向量中具有正值的不同串查询上述Tier0和Tier1和Tier2索引来生成候选提及和候选相关概念。如果候选提及的特征向量与概要向量具有强重叠,则系统可以断言提及的上下文很可能“切题”,即,关于用户正在研究的内容。类似地,对于关系推荐,如果用于共同出现的概念的提及的特征向量的组合与概要向量具有强重叠,则系统可以断言它很可能是切题的。

[0185] 共同出现是生成候选关系的简单方式。如果两个概念的提及出现在同一文档中,则这两个概念被称为“共同出现”。此类共同出现暗示这两个概念之间的有意义关系的可能性随着提及的邻近度的增加而增加。即,在文本或图像或其他媒体中更靠近在一起的提及可能更有可能实际上相关。

[0186] 系统可以使用简单的点积、马氏距离或另一成对核函数(pairwise kernel function)或由应用于向量的部分的核组成的潜在函数的加权和来测量这些向量的重叠。在一个方面中,特征向量可以是来自若干较小向量空间的向量的乘积。例如,一个子空间是出现在提及的上下文中的人名,而另一子空间是出现在提及的上下文中的组织名称。

[0187] 除了提及或相关概念关于用户的项目切题的可能性之外,概要向量可以用于估计提及将对用户有用的可能性。如果用户正在寻求用户尚未知道的新材料,则系统可以对提及的特征向量(或提及的向量的相关概念组合)中的维度的数量进行计数,并且然后还对未在概要向量中存在的那些维度的数量进行计数。新颖性分数可以被计算为不在聚合概要向量中的概念的数量与在候选的向量中的概念的数量的比率。在考虑许多候选推荐的列表时,系统还可以寻求使候选多样化。这可以例如使用多种多样化模型中的任何模型(诸如,最大边际相关性)来实现到此专用的上下文,其中来自KOJ的聚合概要向量可以表示用户意图。

[0188] KOJ的一个优点是,不同的视觉显示元素可以使其对用户的知识呈现同步。例如,用户可以正在与KOS一起工作,所述KOS在用户已经打开的每个文档中显示概要卡队列。从一个卡队列中丢弃卡的用户动作可以使得所述卡从示出相同项目的其他队列中移除。类似地,用户可以采取确认对自动系统认为提及感兴趣实体的所提议的文档感兴趣。所述动作可以通过一些视觉知识元素(例如,文档中的上下文中的突出显示)来接收。通过KOJ,可以响应于此来更新其他视觉显示。例如,如果用户点击上下文中的提及的突出显示上的动作按钮,则所述项目的图谱内顶点聚合可以示出所述概念的顶点,所述顶点具有将其连接到

其他顶点的适当的线。

[0189] KOJ结构还可以有利地促进机器学习算法的使用。通过记录肯定用户动作和否定用户动作,KOJ可以为训练分类器和其他可训练算法提供真肯定和假肯定的源。因为KOJ动态地跟踪用户和用户焦点,所以KOJ可以为宽范围的机器学习模型等提供高质量的训练标记源。

[0190] 引述注释可以是特殊类型的操作,其中用户捕捉文本的片段并且将其作为观察到的数据的快照存储在日志中。引述注释可以提及图谱中的其他概念顶点,其可以表征为从引述注释到概念的提及边。引述注释还可以在用户想要引用提及链时创建。在用户引用提及链时,新的引述注释可以与引用一起创建,并且它可以具有到包含相应的提及链的概念顶点的引用边。

[0191] 可能有助于在KOS中解决的一个问题是对文档或相同概念的引用的去重复。例如,考虑用户可以在大约相同的时间请求两个不同概念的相关概念。这种请求可以由后端异步地处理。在发现新的相关概念时,可以将其作为单独的操作添加到操作日志。如果这些概念中的一些相同(即,coref算法可以检测的提及可能正引用相同的概念),则可以如下所述使用几种不同的解决方案。

[0192] 对于前述问题,存在几种不同的高级方法。一种可以基于身份,这表示为顶点选择标识符的过程为其提供了身份,并且因此,去重复在用户图谱的上下文中自动发生。这种方法的问题可能在于,所有去重复逻辑被推入为定义概念的身份,这通常意味着引入定义概念的规范词典并且要求所有文档遵从同一词典,这对于实施或强加于真实数据可能是不切实际的。

[0193] 另一种方法可以是迎面直对去重复问题。即,每当顶点被自动添加到知识图谱时,可以首先将其与日志中的其他概念进行比较以确定其是否引用已经在所述项目中的其他地方提及的概念。如果是,则系统可以决定完全丢弃它还是将其与所述顶点合并。丢弃或合并的动作可以由术语“去重复”概括。

[0194] 图8展示了计算机系统。一般地,计算机系统800可以包括例如通过外部设备804连接到网络802的计算设备810。计算设备810可以是或包括本文中所描述的任何类型的计算设备,诸如,以上参考图1描述的任何计算设备。例如,计算设备810可以包括台式计算机工作站。计算设备810还可以或替代地是具有处理器并且通过网络802进行通信的任何其他设备,包括但不限于膝上型计算机、台式计算机、个人数字助理、平板计算机、移动电话等。计算设备810还可以或替代地包括服务器、搜索引擎、云计算资源等,或者计算设备810可以设置在服务器上或设置在虚拟或物理服务器群内,或呈任何其他物理或虚拟化环境或形式。

[0195] 计算设备810可以是以上参考图1描述的任何计算设备。例如,计算设备810可以是服务器、客户端、数据库、搜索引擎、或本文中所描述的任何其他设施或计算设备。在某些方面中,计算设备810可以使用硬件(例如,在台式计算机中)、软件(例如,在虚拟机等中)或软件与硬件的组合(例如,具有在台式计算机上执行的程序)来实现。计算设备810可以是独立设备、集成到另一实体或设备中的设备、跨多个实体分布的平台、在虚拟化环境中执行的虚拟化设备或其某种组合。

[0196] 网络802可以包括任何网络或网络组合,诸如,适于以通信关系耦接其他实体并在计算机系统800中的参与者之间传送数据和控制信息的一个或多个数据网络或互联网络。

网络802可以包括诸如互联网等公共网络、私有网络、以及诸如公共交换电话网或使用第三代蜂窝技术(例如,3G或IMT-2000)、第四代蜂窝技术(例如,4G、LTE.MT-Advanced、E-UTRA等)或高级WiMAX(IEEE 802.16m))和/或其他技术的蜂窝网络等电信网络,以及可以用来在计算机系统800中的参与者之间载送数据的多种企业区域网、城域网、校园网或其他局域网或企业网以及任何交换机、路由器、集线器、网关等中的任何。网络802还可以包括数据网络的组合,并且不必局限于严格公共或私有的网络。

[0197] 外部设备804可以是通过网络802连接到计算设备810的任何计算机或其他远程资源。这可以包括服务器、搜索引擎、客户端、数据库、网络存储设备、托管内容的设备、或可以通过网络802连接到计算设备810的任何其他资源或设备。

[0198] 计算设备810可以包括处理器812、存储器814、网络接口816、数据存储装置818以及一个或多个输入/输出设备820。计算设备810可以进一步包括一个或多个外围设备822和其他外部输入/输出设备224,或者与其通信。

[0199] 处理器812可以是如本文中所描述的任何处理器,并且一般可以能够处理用于在计算设备810或计算机系统800内执行的指令。处理器812可以包括单线程处理器、多线程处理器、多核处理器、或任何其他处理器、处理电路、或适用于处理本文中设想的数据和指令的前述的组合。处理器812可能够处理存储在存储器814中或数据存储装置818上的指令。处理器812还可以或替代地包括模拟物理处理器的虚拟化机器。

[0200] 存储器814可以将信息存储在计算设备810或计算机系统800内。存储器814可以包括任何易失性或非易失性存储器或其他计算机可读介质,包括但不限于随机存取存储器(RAM)、闪存、只读存储器(ROM)、可编程只读存储器(PROM)、可擦除PROM(EPROM)、寄存器等。存储器814可以存储程序指令、程序数据、可执行文件和可用于控制计算设备810的操作和配置计算设备810以执行用于用户的功能的其他软件和数据。存储器814可以包括用于计算设备810的操作的不同方面的数个不同的阶段和类型。例如,处理器可以包括板上存储器和/或缓存以用于对某些数据或指令的更快访问,并且可以包括单独的主存储器等来按需扩展存储器容量。

[0201] 一般地,存储器814可以包括含有计算机代码的非易失性计算机可读介质,当由计算设备810执行时,所述计算机代码创建用于所讨论的计算机程序的执行环境,例如,构成处理器固件、协议栈、数据库管理系统、操作系统、或前述各项的组合的代码和/或执行在各种流程图中阐述的步骤中的一些或全部以及本文中阐述的其他算法描述的代码。虽然描绘了单个存储器814,但将理解的是,可以有用地将任何数量的存储器结合到计算设备810中。例如,第一存储器可以提供非易失性存储器,诸如,即使当计算设备810断电时也用于文件和代码的永久或长期存储的盘驱动器。诸如随机存取存储器的第二存储器可以提供用于存储用于执行过程的指令和数据的易失性(但更高速度的)存储器。第三存储器可以用于通过为寄存器、缓存等提供物理上与处理器812相邻的甚至更高速度的存储器来改善性能。在另一方面中,存储器814可以包括模拟物理存储器资源的虚拟存储器。

[0202] 网络接口816可以包括用于通过网络802与其他资源以通信关系连接计算设备810的任何硬件和/或软件。这可以包括使用例如物理连接(例如,以太网)、射频通信(例如,WiFi)、光学通信(例如,光纤、红外等)、超声通信或这些的任何组合或者可以用于在计算设备810与其他设备之间载送数据的其他介质到资源(诸如,可通过互联网访问的远程资源以

及使用短距离通信协议可用的本地资源)的连接。网络接口816可以例如包括路由器、调制解调器、网卡、红外收发器、射频(RF)收发器、近场通信接口、射频识别(RFID)标签读取器或任何其他数据读取或写入资源等。

[0203] 数据存储装置818可以是提供计算机可读介质的任何内部存储器存储装置,诸如,盘驱动器、光学驱动器、磁驱动器、闪存驱动器或能够为计算设备810提供大容量存储的其他设备。数据存储装置818可以非易失性形式存储计算机可读指令、数据结构、程序模块和用于计算设备810或计算机系统800的其他数据以供后续取得和使用。数据存储装置818可以存储用于操作系统、应用程序和其他程序模块、软件对象、库、可执行文件等的计算机可执行代码。数据存储装置818还可以存储程序数据、数据库、文件、媒体等。

[0204] 输入/输出接口820可以支持来自可以耦接到计算设备810的其他设备的输入和向其的输出。外围设备822可以包括用于向计算设备810提供信息或从其接收信息的任何设备或设备组合。这可以包括人类输入/输出(I/O)设备,诸如,键盘、鼠标、鼠标垫、跟踪球、操纵杆、麦克风、脚踏板、相机、触摸屏、扫描仪或可以被用户830采用以向计算设备810提供输入的其他设备。这还可以或替代地包括显示器、扬声器、打印机、投影仪、耳机或用于向用户呈现信息或以其他方式提供来自计算设备810的机器可用或人类可用输出的任何其他视听设备。其他硬件826可以结合到计算设备810中,诸如,协处理器、数字信号处理系统、数学协处理器、图形引擎、视频驱动器等。其他硬件826还可以或替代地包括扩展的输入/输出端口、额外存储器、附加驱动器(例如,DVD驱动器或其他附件)等。总线832或总线组合可以用作用于互连计算设备810的组件(诸如,处理器812、存储器814、网络接口816、其他硬件826、数据存储装置818和输入/输出接口)的机电平台。如图中所示,计算设备810的各个组件可以使用系统总线832或用于传达信息的其他通信机构来互连。

[0205] 本文中所描述的方法和系统可以使用计算机系统800的处理器812来实现,以执行包含在存储器814中的一个或多个指令序列来执行预定任务。在实施方案中,计算设备810可以被部署为数个并行处理器,所述并行处理器被同步以一起执行代码从而改善性能,或者计算设备810可以在虚拟化环境中实现,其中管理程序或其他虚拟化管理设施上的软件在适当时模拟计算设备810的组件以再现计算设备810的硬件实例化的功能中的一些或全部。

[0206] 图9示出了用于调查实体之间的关系的方法的流程图。一般地,搜索引擎可以搜索标识实体的提及或更具体地标识实体相关提及组的信息,以及在文档语料库内的此类实体相关提及组之间的关系。相关联的用户界面可以交互地向用户呈现信息,以便用户引导对感兴趣的实体(或实体相关提及组)和关系的选择和呈现。

[0207] 如步骤902中所示,方法900可以开始于提供对文档集合的访问。这可以包括用于将搜索引擎和用户显示联系到感兴趣的文档的多种技术中的任何技术。例如,这可以包括提供到搜索引擎的连接,所述搜索引擎已经搜索、分析和索引可用文档(诸如,本地目录、云存储设施或公共网络(诸如,万维网)或这些的任何组合)。这还可以或替代地包括例如通过支持托管用户界面的设备与文档源之间的连接来提供直接访问,其可以用于在接收到用户查询时动态地搜索和处理文档。更一般地,用于支持用户搜索、编程搜索和取得、分析(例如,包括实体提及的证据,在本文中也被称为“证据段落”,其实质化实体之间的关系或以其他方式支持实体相关的提及分组)、用户显示和查看以及与底层内容的用户交互的任何技

术可以被用来提供对文档的访问,如步骤902中所示。将理解的是,宽范围的索引技术、网络搜索引擎等是可用的。类似地,数据托管和数据储存库系统涉及宽范围的访问控制技术和访问集成技术,包括Kerberos、活动目录(Active Directory)、Oauth、公钥基础设施(PKI)和其他数据传输、登录技术和/或基于云的存储技术。可以单独地或者与推荐引擎、机器分析系统等组合使用前述中的任何,以提供对文档的访问并且支持如本文中所描述的此类文档内的实体相关提及组的用户调查。

[0208] 如步骤904中所示,方法900可以包括接收用户输入。这可以是发起由本文中所描述的实体调查平台的进一步处理的任何用户输入。如步骤906中所示,方法900可以包括处理文档。在此上下文中,处理文档可以包括诸如搜索和分析的一般后台处理以创建实体数据库、索引内容等,或者处理文档可以包括响应于具体用户请求或其他用户输入的特定处理。一般地,取决于具体用户交互,接收用户输入可以发生在文档处理之前、期间或之后,或者这些的任何组合。

[0209] 例如,在一个方面中,接收用户输入可以包括向搜索输入表单栏提供文本输入以发起新的搜索或项目。然后,文档的处理可以包括在可用文档语料库内搜索要呈现给用户的候选实体提及。在一个方面中,用户可以直接在搜索输入表单栏内识别两个或更多个实体,并且方法900可以包括在用户界面内显示数个候选提及,例如作为标签等。在另一方面中,包含各种实体的提及的文档或来自此类文档的摘录可以显示在用户界面内,并且每个文档可以与标识文档内的提及的一个或多个标签一起显示。用户然后可以具体地选择感兴趣的提及。

[0210] 将理解的是,如本文中所使用的术语“实体(entity)”不暗示物理世界中的特定的已知实体。通常,本文中设想的数据处理得出有关真实世界中的实体的结论是毫无意义的。而是,本文中所描述的实体是对位于文档中的数个提及之间的关系的推论,其中可以推断或预测一组共指提及来引用称为实体的单个事物。因此,本文中设想的“实体”是由于其中出现提及的文档的周围上下文而看起来是共指的一组提及,或者表征或存留此类基于实体的提及组的数据结构,而不是真实世界实体的基础事实识别。因此,除非明确说明或以其他方式从上下文中清楚,如在所附说明和权利要求中使用的术语“实体”应被理解为用作在文档内出现的引用单个实体的实体相关提及组、或者表征或反映实体相关组的任何数据结构等的简写。一般地,这种实体相关组中的相应提及是看上去共指的提及,而不暗示有关共指或此类共指所暗示的单个实体在真实世界中的重要性的任何结论。

[0211] 在另一方面中,方法900的步骤904和906可以包括诸如通过在文档集合中选择引用第一实体和第二实体的文本跨度来识别第一实体和第二实体,每一个在文档集合中被提及。这还可以或替代地包括用于指定实体或从实体中进行选择的数个不同的用户交互。例如,用户可以明确地从由平台提供的实体列表中选择两个实体,或者用户可以输入两个实体描述并且平台可以建议在文档内可能对应于由用户描述的实体的提及。在另一方面中,用户可以识别一个实体,并且平台可以基于例如在文档内的共指来搜索可能感兴趣的其他相关实体。用户然后可以从由平台识别并在用户界面中呈现给用户的实体标签和提及中选择第二实体。

[0212] 在其中实体被识别并且用户输入包括对支持文档内的关系的证据的请求的一个方面中,处理文档可以包括处理文档集合以识别证实第一实体与第二实体之间的联系的证

据(例如,证据段落),所述证据包括第一实体和第二实体在文档集合的至少一个文档内的共同出现提及。这还可以或替代地包括处理文档集合以识别文档集合中的一个或多个文档中的证实第一实体与第二实体之间的联系的证据段落,诸如,包括对第一实体的第一提及和对第二实体的第二提及的证据段落。

[0213] 其他交互还可以或替代地用于基于文档集合中可用的证据来识别和选择实体之间的关系。例如,所述方法可以包括基于多个实体之间的共指来向用户呈现数个候选联系,以及响应于由用户对所述数个联系中的一个联系(例如,如下述)的选择在用户界面中创建联系以供显示。在另一方面中,方法900可以包括接收由用户对第一实体的选择以及自动建议与第二实体的联系,因而向用户提供识别文档集合中的信息所建议的可能关系以及从所述可能关系中进行选择的辅助。在另一方面中,方法900可以包括接收由用户对第一实体和第二实体的选择,以及在文档集合内搜索证据段落以证实联系。

[0214] 如步骤908中所示,方法900可以包括存储关系描述,所述关系描述表征来自文档集合内的证实第一实体与第二实体之间的联系的证据,或者以其他方式基于文档集合中的一个或多个文档内的信息来提供对两个实体之间的可能关系的支持。例如,这可以包括包含两个实体的共指的文档的标题和/或位置,连同文档的全文,或者提及两个实体的一个或多个摘录。这还可以或替代地包括相关的感兴趣信息,诸如,提及涉及感兴趣实体的可能性、两个实体的提及之间的距离(例如,词语的数量)、实体之间的关系的明确陈述的识别(例如,基于对文档的语义分析等)、包含共指的其他文档的数量、或对基于集合中的一个或多个文档来证实或表征实体之间的关系有用的任何其他信息。

[0215] 此信息可以作为关系描述909存储在日志、数据库或可由支持用户界面的平台访问的其他数据储存库中。可以识别底层实体(例如,实体相关提及组)和一个或多个关系,并且可以自动地、响应于用户请求手动地或其某种组合来创建关系描述909。例如,如上所述接收用户输入可以包括基于对多个实体的提及的共同出现向用户呈现作为候选的数个联系,以及响应于由用户对所述数个联系中的一个联系的选择来创建关系描述909。在另一方面中,接收用户输入可以包括接收由用户对第一实体的选择以及自动建议与第二实体的联系。在另一方面中,接收用户输入可以包括接收由用户对第一实体和第二实体的选择以及在文档集合内搜索证据(例如,证据段落)以证实联系。接收用户输入还可以或替代地包括接收由用户对多于两个实体的选择以及在文档集合内搜索对联系的证实。更一般地,关系描述909可以使用本文中所描述的技术中的任一种来自动地和/或交互地创建。

[0216] 如步骤910中所示,方法900可以包括图形地显示第一实体与第二实体之间的关系。一般地,一旦已经识别或选择了两个实体之间的关系,就可以使用数种技术在图形用户界面内支持进一步的用户交互。例如,这可以包括在用户界面内显示对于第一实体的第一图标和对于第二实体的第二图标,以及在用户界面内显示联接第一图标和第二图标的连接器符号等。连接器符号可以是线、箭头或视觉地将第一图标联接到第二图标并且表示第一实体与第二实体之间的联系的其他视觉特征等。为了支持用户交互,连接器符号(本文中也简称为“连接器”)还可以包括用户界面元素,用户可以激活所述用户界面元素来访问支持所述联系的证据的关系描述909。

[0217] 如步骤912中所示,方法900可以包括接收与显示器的用户交互。例如,这可以包括用户点击或悬停在用户界面内的连接器符号上、选择第一图标和第二图标、或以其他方式

以预定方式与图形显示器交互,从而请求附加信息、提供附加输入、或以其他方式与关系描述909或支持证据交互。

[0218] 如步骤914中所示,方法900可以包括对用户交互作出响应。如上所述,方法900可以包括以下步骤:处理数个文档,识别其中的实体的提及,以及进一步识别此类文档内的证明两个实体之间的关系的证实化信息,所述证实化信息可以被存储为关系描述。利用此信息,用户界面可以支持多种有用的交互以促进调查特定关系的基础,以及搜索由文档集合建议的附加的相关实体和关系。

[0219] 在一个方面中,对用户交互作出响应可以包括通过提供对用户界面中的证据的访问来对与连接器的用户交互作出响应。例如,响应于与视觉地连接两个图标的连接的用户交互,这可以包括诸如通过显示先前存储的关系描述的一部分来提供对支持用户界面内所描绘的关系的证据的访问。在此上下文中,提供访问可以包括将证据的至少一部分显示为文本等。例如,证据可以包括来自文档集合中的包含第一实体和第二实体的共同出现提及的一个或多个文档的文本,并且提供对证据的访问可以包括在文本内突出显示共同出现提及。

[0220] 在另一方面中,响应于与用户界面中显示的连接器符号的用户交互(诸如,鼠标悬停或鼠标选择)可以包括更新用户界面的搜索输入表单栏以包括对证实第一实体与第二实体之间的联系的文档的搜索请求。

[0221] 根据前述内容,本文中还描述了包括搜索引擎和计算机的系统。搜索引擎可以托管在上述任何设备上,并且可以被配置成接收对第一实体和第二实体的识别,并且处理文档集合以识别证实第一实体与第二实体之间的联系的证据,所述证据包括在文档集合中的至少一个文档内的对第一实体和第二实体的共同出现提及。计算机可以包括处理器、存储器、显示器和用户输入设备,并且计算机可以被配置成支持与关系描述909和文档集合的用户交互以调查实体关系,如本文中所描述的。例如,处理器可以由计算机可执行代码配置以执行以下步骤:识别第一实体和第二实体;将第一实体和第二实体提交给搜索引擎;以及响应于来自搜索引擎的结果,在用户界面内显示对于第一实体的第一图标、对于第二实体的第二图标、以及连接器符号,连接器符号将第一图标视觉地联接到第二图标并且表示其间的联系,并且连接器符号通过用户界面可操作以在用户界面内提供对证据的访问。在一个方面中,搜索引擎在计算机上本地执行。在另一方面中,搜索引擎是可由计算机通过数据网络访问的远程搜索引擎。搜索引擎还可以数种方式在这些设备之间分布。例如,在将查询转发到搜索引擎之前,计算机可以本地解析查询的文本(例如,以识别明确搜索运算符)或以其他方式预处理文本。在另一方面中,计算机上的本地搜索引擎可以支持对本地托管的文档的调查,并且可以与远程搜索引擎协作,所述远程搜索引擎处理分布在广域网上或以其他方式在除所述计算机之外的计算机资源上托管或可用的文档集合。

[0222] 图10示出了用于调查实体之间的关系的方法的流程图。

[0223] 如步骤1002中所示,方法1000可以开始于经由来自用于搜索引擎的用户界面的搜索输入表单栏接收来自用户的关键字搜索,所述搜索引擎诸如本文中所描述的任何搜索引擎或其他基于实体的工具。关键字搜索可以包括任何文本输入以及任何运算符或形成用户的搜索意图的明确陈述的其他文本等。

[0224] 如步骤1004中所示,方法1000可以包括基于关键字搜索来预测用户意图。例如,这

可以包括通过将共指算法应用于关键字搜索或以其他方式分析关键字搜索的文本以识别其中标识的可能实体,来预测用户在关键字搜索中预期引用的一个或多个实体。如所指出的,在此上下文中,术语“实体”应被理解为与已被识别为共指的实体相关提及组的简写。因此,步骤1004可以更具体地被理解为包括将共指算法应用于关键字搜索,以预测用户预期并由一个或多个实体相关提及组表征的实体,或换句话说,以预测用户预期引用的实体相关提及组。

[0225] 如步骤1006中所示,方法1000可以包括呈现第一搜索结果。这可以例如包括在用户界面中显示搜索结果,所述用户界面包括第一多个实体标签(本文中也简称为“标签”)和例如从文档集合所获得的多个文档。一般地,每个文档将不以全部内容进行显示。而是,文档在用户界面中的显示将包括以下显示:元数据(诸如,标题、日期、源和/或其他信息)连同代表性内容(诸如,证据段落或其他摘录等)以及可选地证据段落内对实体提及的突出显示。所述多个文档中的每个文档可以包含由搜索引擎定位的对所述一个或多个实体中的至少一个实体的提及(或由搜索引擎定位在所述一个或多个实体相关提及组中的至少一个实体相关提及组中的提及),并且所述多个实体标签中的每个实体标签可以对应于所述提及中的至少一个提及。还将理解的是,每个实体标签因此表示来自已经被识别为共指或以其他方式相关的实体相关提及组的具体提及。一般地,文档集合可以本地托管在接收关键字搜索的用户设备上、分布在广域网上、托管在远程云存储设施上或其某种组合。

[0226] 如步骤1008中所示,方法1000可以包括例如通过接收来自用户的从第一多个实体标签中对于第一实体的第一实体标签的选择来接收对实体标签中的一个实体标签的用户选择。如所指出的,第一实体可以由第一实体相关提及组来表征,或者可以是对应的实体相关提及组。选择可以机械地包括对用户界面内的实体标签的点击或其他操作。一般地,这种第一选择可以指示关键字搜索曾预期引用第一实体的用户确认,从而将在用户选择的特定文档中对对应的实体提及基础化。

[0227] 如步骤1010中所示,方法1000可以包括将第一实体标签添加到搜索输入表单栏。因此,文本的一部分可以由对应的实体和/或在由用户选择的文档中为所述实体提供的特定上下文来替换。在一个方面中,这个动作(将实体标签放置在搜索输入表单栏中)是向用户提供已经参考一个或多个特定文档选择了特定实体的视觉反馈的用户界面功能。在另一方面中,这通过提供可用于代替用户初始提供的一个或多个关键字的信息来提供基于实体的搜索功能,以支持围绕实体的改进的基于实体的搜索。

[0228] 如步骤1012中所示,方法1000可以包括找到与其他实体的关系的证据。在此上下文中,证据可以是根据相关性算法或其他工具或度量(metric)倾向于证实实体之间的关系的任何证据段落等。同样如本文中所指出的,每个实体可以由实体相关提及组来表征或者可以包括所述实体相关提及组。因此,方法1000可以包括应用相关性算法或其他工具或度量来识别第二多个文档中提供第一实体与一个或多个其他实体之间的关系的证据(诸如,证据段落)的文本,或者以其他方式定位证实与其他实体的可能关系的存在的信息。

[0229] 如步骤1014中所示,方法1000可以包括在用户界面中呈现对于在搜索与其他实体的关系的证据期间识别的一个或多个其他实体(或一个或多个其他实体相关提及组)的第二多个实体标签。例如,可以在搜索输入表单栏的正下方或者在一些其他视觉上相邻的位置中显示这些标签以供用户方便地进行视觉分析和选择。

[0230] 如步骤1016中所示,方法1000可以包括接收来自用户的从第二多个实体标签中对于第二实体(或实体相关提及组)的第二实体标签的第二选择,诸如通过针对用户界面中第二多个实体标签中的标签中的一个的指向和点击。

[0231] 如步骤1018中所示,响应于对第二标签的用户选择,方法1000可以包括基于第二标签执行搜索。具体地,可以对证实第一实体与第二实体之间的关系的文档执行新的搜索,例如通过证据段落等。这可以包括对任何适当的文档集合的全新搜索,或者这可以包括在先前搜索关系的证据时(例如,在以上步骤1012中)获得的结果的取得。

[0232] 如步骤1020中所示,方法1000可以包括例如通过呈现包括证实第一实体与第二实体之间的关系的第二多个文档中的一个或多个文档的第二搜索结果来显示新的搜索的结果。这些结果可以多种方式中的任一种显示。例如,这可以包括在用户界面中将第一实体与第二实体之间的关系图形地显示为对于第一实体的第一图标、对于第二实体的第二图标、以及将第一图标视觉地联接到第二图标的连接器符号,例如从而视觉地展示所述关系。连接器符号可以包括用户界面元素,所述用户界面元素提供用户对第二多个文档内的关系的证据的访问。显示结果还可以或替代地包括在搜索输入表单栏内将关系显示为第一实体标签通过关系运算符(诸如,“<>”符号)联接到第二实体标签。将理解的是,还可以或替代地执行附加的相关处理。例如,可以任何适当的方式和位置存储表征第一实体与第二实体之间的关系的关系描述,例如以供用户在用户界面内进一步使用。

[0233] 在另一方面中,可以图形地显示已经使用本文中的技术识别的数个实体标签和关系。因此,例如,呈现第二搜索结果可以包括图形地呈现各自通过图谱中的边连接到第一实体的多个实体标签,其中每个边在所述用户界面内可操作,以供用户取得支持证据段落,并且每个实体标签可由用户操作,以从由图形显示反映的知识图谱添加或移除对应的实体相关提及组。

[0234] 在另一方面中,本文中公开了支持上述方法1000的系统。例如,本文中公开的系统可以包括搜索引擎、数据网络和通过数据网络耦接到搜索引擎的计算设备。计算设备可以包括处理器、存储器和显示器,其中存储器存储可由处理器执行以执行以下步骤的代码:在用户界面的搜索输入表单栏中接收来自用户的关键字搜索,预测由用户预期的一个或多个实体;在用户界面中向用户呈现第一搜索结果,所述第一搜索结果包括第一多个实体标签和多个文档,所述多个文档中的每个文档包含由搜索引擎定位的对所述一个或多个实体中的至少一个实体的提及,并且所述多个实体标签中的每个实体标签对应于提及中的至少一个提及;接收来自用户的从第一多个实体标签中对于第一实体的第一实体标签的第一选择,所述第一选择指示用户确认关键字搜索曾预期引用第一实体;将第一实体标签添加到搜索输入表单栏;识别第二多个文档中的文本,所述文本提供第一实体与一个或多个其他实体之间的关系的证据;在用户界面中呈现对于所述一个或多个其他实体的第二多个实体标签;接收来自用户的从第二多个实体标签中对于第二实体的第二实体标签的第二选择;并且呈现包括第二多个文档中的一个或多个文档的第二搜索结果,所述一个或多个文档证实第一实体与第二实体之间的关系。

[0235] 本文中还公开了支持基于实体的关系的发现和导航的另外的用户界面技术。例如,在一个方面中,本文中设想的方法包括:显示用于搜索引擎的用户界面,所述用户界面包括搜索输入表单栏;接收对搜索输入表单栏的用户输入;解析用户输入以识别标识第一

实体的第一文本串、标识第二实体的第二文本串、以及在第一文本串与第二文本串之间的运算符,所述运算符向搜索引擎指定搜索第一实体与第二实体之间的关系的证据的请求;响应于运算符,在文档集合中搜索一个或多个文档,所述一个或多个文档包含共指算法预测的提及作者预期引用第一实体的提及,以及共指算法预测的提及作者预期引用第二实体,和提供相关性算法预测的提供第一实体与第二实体之间的关系的证据的文本的其他提及;以及在用户界面中呈现所述一个或多个文档中的多个文档。

[0236] 图11展示了用于调查实体之间的关系的用户界面。一般地,用户界面1100可以被呈现在计算设备的显示器上,并且可以包括被配置成从用户接收文本输入的搜索输入表单栏1102。如图11所展示的,例如来自搜索结果的感兴趣文档的概要1104可以与引用可能由文档描述的实体的一个或多个实体标签1106一起显示。实体标签1106可以是用户可选择的,并且如上所述,可以通过(a)将所选择的标签视觉地放置在搜索输入表单栏1102内,和/或(b)将搜索结果中所示的特定文档中对实体的提及基础化来对用户选择作出响应。如本文中进一步描述的,随着在特定文档中基础化的实体提及,用户界面1100然后可以呈现可以与搜索输入表单栏1102中识别的实体相关的数个附加实体的标签。用户可以例如通过与实体标签1106交互来选择对应的实体,并且对这些关系的用户指示可以用于搜索证实所述关系的适当的证据段落,以创建证实的关系的图形可视化等,所有都如本文中所描述的。

[0237] 图12展示了用于调查实体之间的关系的用户界面。在图12的用户界面1200中,在搜索输入表单栏1202中呈现两个实体,以及指示对证实这两个实体之间的关系的文档的搜索的运算符符号1204(在此实施例中为“<>”),其中结果1206通常呈现在用户界面1200内,例如,其中对应实体的提及被突出显示以用于快速视觉识别。还可以基于特定用户搜索或者基于包含对应关系描述和其他信息的预先存在的项目等来提供实体关系的图形呈现1208。将注意到,诸如线或箭头等连接器符号1210可以被用来描绘关系,并且可以是用户可选择的以执行诸如搜索包含支持关系的存在或性质的证据段落的文档的功能。

[0238] 现在借助于上述方法和系统的非限制性示例实现方式的方式提供附加细节和解释。非限制性举例而言,基础化实体提及和相关技术的进一步细节也可以在2019年4月10日提交的美国临时申请第62/832,085号中找到,并且所述申请通过引用以其全文结合于此。

[0239] 如本文中所使用的,概念的提及可以是人类读者可能认知为引用共享概念的字符串。提及当包括读者可以解读提及集合的含义以减少不确定性的其他概念的充分附加提及时被基础化。这种消歧过程与基础化并进。概念在其含义的丰富讨论中越彻底地基础化,就越彻底地消除其歧义,并且读者就可以越精确地对其进行理解。概念是可以在人之间传达的一般性想法,使得讨论者相信他们具有共享的概念。事件是在空间和时间中锚定的概念,使得讨论此类共享想法的人们具有对事件发生在何处和何时的共享理解。事件通常具有发生或正在发生的空间范围和时间跨度。“实体”是特殊的概念子集,所述概念也是通过区分属性来识别的,从而使得讨论实体的人具有对他们正在讨论哪个实体的共享理解。人和公司是熟悉的实体类型。实体通常具有名称,并且当人们讨论给定实体时,其名称通常被用作引用真实世界实体的共享概念。虽然前面的描述强调实体,但将理解的是,本文中的系统和方法可以替代或附加于实体来与任何其他概念一起使用。

[0240] 当文档一起提及两个概念时,所述概念“共同出现”。引用两个不同实体的提及的共同出现不一定暗示这两个实体具有直接关系。例如,web页面可以包含出售鼓风机的公司

的列表,并且仅因为两个不同的公司的名称出现在所述列表中,所述共同出现不一定意味着这两个公司具有直接关系(诸如,合作关系)。此类共同出现仅可以证实间接关系,诸如两者都是类似或相关产品的制造商。然而,相反的情况通常是事实。即,如果两个实体具有直接关系,则所述直接关系经常在文本段落或引用两个实体的其他媒体对象中表达或提及,即,包含对所述两个实体的共同出现引用。因此,寻找包含共同出现引用的文档通常是有用的,因为这些文档是用于证实或提供两个实体或概念之间的关系的证据的候选文档。

[0241] 如本文中设想的搜索引擎界面可以使得用户能够表达对可以证实关系的此类候选文档的请求。视觉界面可以包括至少三个元素:(i)第一基础化提及的第一视觉图标,其对第一实体消歧,(ii)第二基础化提及的视觉图标,其对第二实体消歧,以及(ii i)将第一视觉图标和第二视觉图标互连的图形链接。此类输入可以被搜索引擎查询解析器解读为搜索证实第一实体与第二实体之间的可能关系的文档的请求。图标表示基础化提及,例如,人类将可能认知为所述文档的作者旨在引用某个真实世界实体的引用的特定文档中的字符跨度。

[0242] 在优选实施方案中,用户可以输入对于不同实体的一个、两个、三个或更多个此类视觉图标。通过输入描绘基础化提及的一个此类视觉图标,用户可以请求提及所述特定实体的文档,而不是可能提及与所描绘的基础化提及具有相似或串相同(string-identical)名称的不同实体的其他文档。通过输入描绘对各种实体的基础化提及的两个或更多个视觉图标,用户请求提及这多个实体的文档。此类查询可以被称为“多实体查询”。

[0243] 在实施方案中,搜索引擎算法通过对提及列表中更高的更多所识别实体的文档进行排位来对此类多实体查询作出响应。此类查询可以被称为SHOULD BOOSTING查询,而不是STRICT AND。如果搜索引擎算法将STRICT AND语义应用于查询,则它将仅用提及由用户的多实体查询输入所识别的所有实体的文档来作出响应。相反,SHOULD BOOSTING语义允许排位算法用提及比由多实体查询输入所识别的所有实体少的实体的文档来作出响应。虽然提及全部或大部分实体的文档被给予较高排位,但是结果列表可以包括提及少于全部实体(可能仅包括实体之一)的文档。如果排位算法确定那些结果中的内容与围绕用户的多实体查询输入中的各个实体的视觉图标中描绘的基础化提及的上下文相关,则结果列表甚至可以包括没有提及输入查询中的实体的结果。

[0244] 围绕基础化提及的上下文可以使得人类读者能够认知含义并且推断源/作者所预期的共指。这种上下文可以与本文中所描述的系统和方法一起使用,以改进搜索结果和锚定搜索活动,特别是用户标记的实体的提及。

[0245] 在一个优选实施方案中,用户界面允许用户为查询中的每个视觉图标标识多个基础化提及。例如,John Smith的视觉图标可以表示由该名字命名的特定人的若干提及的集合。这些文档可以出现在多个文档中,其各种上下文段落围绕每个文档中的提及。给定的文档可以多次提及实体,并且围绕每个此类提及的词语提供上下文。在优选实施方案中,搜索引擎算法建立围绕这一个或多个提及的上下文的维度缩减的向量表示,使得每个视觉图标对应于围绕这些提及的聚合或凝聚上下文的机器表示。此类向量表示在本领域中是常见的并且可以采取许多形式。例如,一个表示可以使用词包,所述词包计数来自每个提及周围的词窗口(诸如,三句窗口)的词和短语。在另一个实施例中,在训练神经网络的内部活动以基于天然出现的词语序列预测短语之后,从神经网络的内部活动构建嵌入向量。在另一示例

方法中,通过将共同出现统计的矩阵因式分解来学习嵌入向量。其他方法可以结合这些和其他向量化方法。所有这些方法的一个显著特征是搜索查询输入收集用户自己的对搜索的用户预期目标消歧的基础化提及,并且搜索引擎算法可以进而基于这些改进的输入来改进结果的排位。

[0246] 短语“实体参考集”或ERS在本文中用于指对实体的提及的集合,并且更具体地指由用户(利用本文中描述的界面)收集到基础化查询的一部分的集中的基于实体的提及组。虽然此提及组在本文中有时也被简称为“实体”,但是实体参考集更精确地是文档和每个文档内的特定子范围的标识符的集合,其中文档内容引用感兴趣的实体。ERS还可以包含“否定提及”,所述否定提及是用户已经标记为未引用所讨论的实体的实体提及。用户已经标记为正确引用所讨论的实体的提及被称为“肯定提及”。例如,在探索关于General Motors的董事长John Smith的内容时,如果用户找到提及正确的人的文档,则用户可以在ERS中将所述文档标记为对John Smith的肯定提及。如果用户发现对足球教练John Smith的提及,则用户可以将后者提及标记为没有引用预期的人,并且因此标记为否定提及。以此方式,用户可以具有查询,所述查询具有描绘两个不同的ERS的视觉图标,其具有相似或甚至相同的表面形式名称串。一个ERS可以引用董事长,而另一个可以引用教练,并且每个ERS中的基础化提及对预期含义进行消岐,即使表面名称串相同。

[0247] 收集实体提及的过程可以被称为“提及级加书签”,因为由系统存储的标识符(例如,在关系描述或其他数据结构中)标识文档还有具体提及文档。对于一个实体的此类书签的集合是ERS。可以一起使用若干ERS来指定针对实体之间的一种或多种关系的证据的多实体查询。实体参考集的集合也可以作为“项目”进行跟踪。如上所述,知识操作日志或“KOJ”可以实现承载一个或多个基础化实体(即,对感兴趣实体的消歧引用)的项目。此类项目是用户可以用项目名称保存并在将来返回的知识人工制品(knowledge artifact),就像诸如演示文件和备忘录等其他类型的办公文件人工制品一样。搜索引擎算法可以允许用户将特定项目标识为查询的上下文,并且然后搜索引擎可以基于项目不同地对文档进行排位。将项目用于搜索上下文的优选实施方案是从项目中的所有实体参考集中的所有基础化提及的所有上下文构建聚合向量,并且将所述向量用作对搜索结果的多样化影响。多样化影响将轻微地抑制看起来对于项目是冗余的文档,使得用户具有更高的可能性从搜索结果命中中学习新信息。

[0248] 针对每个ERS构造的向量可以不同地使用肯定提及和否定提及。肯定提及可以提供扩充搜索引擎取得到的文档的提升短语,而否定提及可以提供拒绝不正确文档的抑制短语。例如,如果用户仅寻找关于董事长的文档,则肯定提及可以提供“汽车”作为上下文词,搜索排位算法可以使用所述上下文词来提升关于正确的John Smith的排位位置文档。相反,否定提及可以提供上下文词“体育场”,搜索排位算法可以使用所述上下文词将关于错误的John Smith的文档降低排位。

[0249] 搜索引擎排位算法可以使用此类基础化提及上下文信息来以若干方式改进排位。作为排位改进的另一实施例,搜索引擎可以抑制与肯定提及的上下文类似的文档,因为这提供了帮助用户了解目标实体的新的和更可能新颖的信息。替代性地,搜索引擎可以提升类似于肯定提及上下文的文档,因为这提供了更可能印证用户已经从已经收集的对文档的提及收集到的理解的文档。

[0250] 可以在使用Lucene索引引擎的标准关键字搜索引擎(诸如,Elasticsearch或SOLR)的顶上实现此类排位技术。ERS的集合可以出现在用户界面中,并且查询处理模块可以解读实体参考集的多实体查询以构建针对低级搜索索引的查询,所述查询被配置成接收包含肯定和否定提升上下文的复杂查询。否定提升可以被实现为SHOULD NOT子句上的肯定提升。

[0251] 当查询从此类搜索引擎取得文档集合时,其还可以请求引擎生成其他共同出现术语的聚合,诸如其他实体的提及或为消歧引擎已经评估为共指的提及组设置标识符。用户界面可以呈现来自此类聚合的术语作为推荐的其他感兴趣实体。在由查询取得到的文档子集内接收频繁提及的其他实体是“相关实体”。系统可以收集相关实体并且将其呈现给用户作为使用多实体查询进行进一步探索的鼓励。

[0252] 例如,多实体查询可以具有由三个ERS中的基础化提及标识的三个实体。第一个引用GM董事长John Smith,第二个引用底特律,第三个引用艾里逊变速箱(Allison Transmission)。响应于此查询,搜索引擎可以找到提及实体的文档,然后聚合查询寻找也在这些文档中频繁提及的其他实体(诸如,达美航空(Delta Airlines)),其中Smith先生也是董事会成员。系统可以使用这种方法来呈现与用户可能还没有考虑探索的其他实体的许多候选联系。通过提出此类相关实体,系统邀请用户探索更多的多实体查询。例如,用户可以移除底特律和Allison Transmission,并添加Delta以作出新的双实体查询来研究Smith先生与Delta之间的关系的证据。

[0253] 用于生成相关实体建议的此类聚合查询可以产生表面形式名称串,这可能不足以形成基础化提及的ERS,特别是在多个实体共享相似或相同的表面形式名称串的情况下。因此,系统可以进一步例如基于用户选择的基础(user-selected groundings)来将相关实体与提及所提议的相关实体的特定文档对准,使得当用户选择此类建议时,其作为由实体视觉图标描绘的ERS进入搜索查询表单栏中。

[0254] 当文档提及一个特定实体一次或多次时,用户界面可以通过视觉线来描绘这一点,所述视觉线将文档图标连接到概念图标,其箭头指向概念图标。来自特定文档的此类箭头的集合是可视化ERS的另一种方式。

[0255] 可以从文档集合中收集关系的证据。证实两个概念之间的关系的文档可以进一步提供关系类型的证据。例如,考虑描述维生素C可以如何帮助人体对抗引起普通感冒(疾病)的病毒的期刊文章。此类文档提供了概念“维生素C”与概念“普通感冒”之间的关系的证据。所述文档中的其他概念的提及提供表征关系的上下文。

[0256] 如上所述,共指消解或“coref”是将含义分派给提及的过程。为了帮助共指消解,用户动作的上下文可以暗示相关性,诸如,其中作出推荐请求的上下文。在另一方面中,用户动作的顺序可以暗示相关性,诸如,用户向知识图谱添加概念或请求对图谱的边进行证实的顺序。在另一方面中,用户动作的频率可以暗示相关性,诸如,用户请求对图谱的特定边进行证实的频率或用户请求与由图谱中的顶点表示的特定概念相关的概念的频率。

[0257] 阅读或理解口头语言的过程是解读作者或说话者的预期含义的过程。人类想法不断地推断或猜测作者预期或意指什么。如果读者或收听者从未观察或设想传达者正尝试引用的事物,则读者或收听者难以进行理解。相反,当读者或收听者先前已经有想法时,则通过检测这种等同性(equivalence)大大地简化了理解动作。以类似的方式,代替向搜索引擎

提交关键字搜索,本文中所描述的技术在某些时间请求明确人类输入,以通过识别由预期引用搜索用户的感兴趣实体的一些人类撰写的特定段落来更精确地表达预期的搜索查询。例如,代替键入“john smith”的关键字搜索并且通过引用具有该名字的许多不同人的搜索结果命中进行除杂,本文中设想的搜索引擎的用户可以通过识别提及特定John Smith的段落来细化他/她的查询。在这样做时,搜索用户利用围绕那些提及的丰富的上下文,其中书写该上下文的作者正在由作者预期引用那个特定的John Smith为条件的思想框架中书写。

[0258] 通过所描述的技术的进一步实施例和本文中所使用的术语,考虑以下文本:“自1992年以来作为Zeel公司CEO的John Smith在47岁时看到他的薪酬跃升了21%,达到120万美元,还接替Sandra Jones成为金融服务公司的董事长主席,其曾任职3年”。

[0259] 所述段落包括引用特定实体的提及:John Smith、Zeel公司和Sandra Jones。这些实体共同出现在此具体提及串中。然而,还存在一些名词性提及,诸如“他的(his)”和“其(who)”,人类读者自然地解析为提及与段落中的其他提及字符串相同的真实世界实体。在这种情况下,“他的”指代“John Smith”所指代的同一实体,并且“其”指代与“Sandra Jones”相同的实体。即,“他的”与“John Smith”共指,并且“其”与“Sandra Jones”共指。混淆这些关联导致共指错误,诸如假设“其”指代John Smith或“他的”Sandra Jones。其他文档中的其他段落也可以参考这些实体,并且人类读者在确定另一个文档中的另一个提及字符串“John Smith”是否引用与这个段落中的“John Smith”相同的实体时执行共指过程。

[0260] 例如,具有相同表面形式名称的两个提及是可能引用同一概念的候选者。如果围绕这两个提及的上下文相似,则人类更有可能将所述提及感知为具有相同的含义,即,感知为是共指的。算法可以将围绕这两个提及的上下文中的数据进行比较,并且估计人类将所述提及感知为是共指的可能性。

[0261] 所述段落还包含共同出现的实体之间的关系。例如,John Smith、Zeel Corp和Santra Jones通过这个段落的内容而相关。跨文档处理这些提及可以包括提取对应于每个实体的提及串,同时保持对相关的共同出现的实体的跟踪。实体之间的关系可以使用将实体互连的图形链接来可视化。当文档多次提及一个特定实体时,它可以通过具有指向文档图标的箭头的边来描绘。然后,这个段落可以成为证实两个或更多个共同出现的实体之间的关系的一份证据,例如,证据段落。

[0262] 关系及其证据是用户进行搜索以单独地了解关于实体本身的更多内容、而且还理解它们之间的联系的重要目标。如上所述,用户在例如图形显示中或经由对两个实体标签的选择来选择特定连接可以改变搜索输入表单栏中的查询。在此搜索输入表单栏中,“<>”符号(或另一符号)可以充当搜索运算符以明确地指示对证实由所述符号连结的两个实体之间的关系的文档的查询。关键字搜索输入可以首先显示第一搜索串和响应于第一搜索串而列出的搜索结果。知识对象概要卡(或知识操作日志的其他表征化)可以显示相关实体推荐,并且响应于用户点击相关实体推荐,用户界面可以改变关键字输入以显示表示相关实体的查询串和用于找到相关实体之间的关系的证据的搜索运算符。另外,用户界面可以改变搜索结果列表以示出提及相关实体的文档,并且所列出的结果中的至少一个可以示出对相关实体的突出显示提及。例如,用户可以搜索“John Smith<>Zeel公司”,这对这两个实体之间的关系的搜索。在此搜索过程中,可以提供附加的用户界面控件,以便用户控制新颖性和相关性以及肯定和否定共指等,从而在看到证实关系的更多证据与用户尚未看到的新关

系的证据之间有区别地折衷。

[0263] 图13至图16通过本教导的实施例和进一步说明展示了用于调查实体之间的关系的各种用户界面。在图13的用户界面1300中,关键字搜索1302在查询输入表单栏中。响应于输入的关键字搜索1302,搜索引擎已经显示搜索结果命中列表1304。通常,结果示出文本片段1306。在本教导的系统中,结果还可以显示一个或多个标签1308,每个标签描绘对文档内的实体的一个或多个提及。标签1308可以示出来自文档的示例表面形式名称串。在实施方案中,可以将着色、加粗或其他视觉强调置于标签1308,其中共指算法预测所述标签可能引用用户由关键字搜索1302所预期的实体。由于关键字搜索1302尚未基础化,因此用户界面1300使得用户能够通过选择标签1308来将他/她的查询基础化。在选择标签1308时,系统将所述标签的视觉表示插入搜索框中,如图14的用户界面中所示。具体地,如图14的用户界面1400中所示,响应于从图13的结果列表中的特定文档中选择特定标签的用户动作,在表单栏1402中显示表单栏1402中的标签1416。此动作将具体提及识别为引用用户由图13的关键字搜索1302预期的相同实体。通过将查询升级到基础化实体提及搜索查询(在表单栏1402中显示),所述交互允许系统消歧并且适当地抑制图13的底部搜索结果命中1304,因为不太可能引用用户的预期感兴趣实体,因此,图13的底部搜索结果命中1304已经被抑制,并且在图14中不显示。值得注意的是,图14具有空的结果1414,在该处本将显示图13的底部搜索结果命中1304。图13的底部搜索结果命中1304以相同的名称提及不同的实体,并且具有标签1416的改进的查询使得搜索引擎排位算法能够消除不正确的结果,从而导致空结果1414。

[0264] 转回到图13的用户界面1300,系统还向用户呈现多个相关实体1310,例如,在此特定实施例中,来自语料库1314中的文档的对实体Jack Wang 1312的提及。搜索引擎算法识别Jack Wang在与用户的关键字搜索1302查询相关的各种文档中被提及,并且使用统计排位算法来将“Jack Wang”的提及排位高于其他实体的提及,并且因此决定在图13的用户界面1300中将其进行显示。

[0265] 在图13的用户界面1300中,用户正在查看针对John Smith的标签1416的ERS查询的结果。所述ERS包括由1408描绘的提及,因为用户先前点击了所述标签。用户可以选择一个或多个附加标签以将它们添加到ERS。例如,用户可以点击董事长Smith 1422以在查询表单栏中将所述附加提及添加到ERS。在实施方案中,用户界面允许用户打开列出ERS中的提及的第二窗口,并且允许用户从ERS移除提及或将它们拆分成单独的ERS。在实施方案中,用户界面跟踪用户在一段时间的过程中已聚集的各个实体引用集(ERS)。

[0266] 在图14的用户界面1400中,系统正在响应于对John Smith的标签1416的ERS的基础提及搜索查询而显示结果。用户可以点击标签,诸如,QingChun公司1418。标签表示由该结果表示的文档中的特定的一组提及,因此当用户选择所述标签时,用户正在选择一个或多个文档1420中的提及集,其提供关于用户预期哪个QingChun公司的消歧上下文。通过点击标签QingChun公司1418,用户向系统指示用户希望看到可以证实由标签1416所描绘的提及标识的两个实体与QingChun公司1418的标签之间的关系的证据段落。在图15的用户界面1500中描绘了此类多实体查询,其中关系运算符1524(即,<>)被搜索引擎的查询解析器解析为找到以下文档的指令:所述文档可以证实由1528描绘的ERS描绘的提及所标识的实体与由1526描绘的ERS之间的关系。搜索引擎用户界面1500为算法推断的文本子串提供可能是对两个感兴趣实体的提及的视觉强调。例如,由1530描述的结果具有对董事长Smith

1516和QingChun公司1518的提及。

[0267] 图14中的用户界面1400还在特定结果上示出标签。结果1404提及名为John Smith的人,并且所述提及由1408描绘。用户界面1400还示出了与由ERS在搜索查询输入表单栏1402中识别的实体可能具有关系的提议实体1410的集合。例如,视觉图标1412描绘了名为Jack Wang的人的提及的集合,并且用户可以点击这个视觉图标1412以将其放入搜索框中,作为对于由这个视觉图标1412的提及识别的实体与由John Smith的提及(标签1416)识别的实体之间的关系证据的新查询。

[0268] 图16的用户界面1600示出显示两个标签图标之间的关系连接器符号1608的方法,所述标签图标描绘来自文档语料库的实体的提及。在图16中,用户界面1600正在示出搜索输入工具1602中的查询的结果。所述结果标识包含各种实体的提及的文档,并且所述结果显示可以示出描绘那些提及中的一些的标签,诸如,描绘Jack Wang的提及的标签1610。由标签标识的此类实体可以在或者可以不在搜索输入工具1602的查询中。在这个实施例中,在第一结果1609中描绘的文档可以是来自语料库的文档之一,其由被显示给用户作为可能感兴趣的关系的视觉图标1604描绘。描绘Jack Wang的ERS的视觉图标1604通过连接器符号1608连接到描绘Sally Smith的ERS的视觉图标1606。这两个视觉图标1604和1606从文档语料库中识别出共指算法已分组在一起作为同一实体的可能提及的一个或多个提及的集合。系统可以提出这两个实体之间的关系,因为关系推荐器算法预测用户可能对阅读证实此类关系的文档感兴趣。当用户例如通过点击连接器符号1608来与其交互时,用户界面1600可以通过改变搜索输入工具1602中的查询来作出反应。这个经更新的查询在第二搜索输入工具1611中示出,其中图标1614描绘Jack Wang的ERS,并且视觉图标1616描绘Sally Smith的ERS,并且关系搜索运算符1612向搜索排位算法指示其应当取得可能证实这两个实体之间的关系的文档。

[0269] 图16的用户界面1600示出了响应于这个经更新的查询的结果。结果1618示出了对来自查询的两个实体的提及1620的视觉强调。结果1622还示出了诸如在实体的提及上的加粗的视觉强调。肯定和否定共指输入指示符1624可以使用户能够同意或不同意给定提及与第二搜索输入工具1611查询中的ERS图标1614中的其他提及共指的共指算法预测。在实施方案中,颜色编码或其他指示符帮助用户有效地认知用户向哪个实体参考集添加系统提出的哪个提及。例如,Sally提及可以是紫色的,并且Jack提及是绿色的。

[0270] 图16中的用户界面1600还可以示出具有第二连接器符号1626的附加关系提议。这鼓励用户反复地探索相关实体的网络。

[0271] 在优选实施方案中,用户界面允许用户在小数量的点击(诸如仅一次点击)中输入肯定或否定共指标记的批量分派。例如,考虑用户正在研究特定的Jack Wang。当用户发起针对与Jack Wang相关的实体(诸如,Sally Smith)的关系查询时,可能是所有结果指代与用户的探索目标不同的Jack Wang的情况。例如,当用户看到Jack Wang与Sally Smith之间的所提议的关系并且点击连接器符号1608时,用户正在探索,并且可能发现这种关系属于与他们的感兴趣人具有相同姓名(“Jack Wang”)的不同人。在这种情况下,批量标记能力使得用户能够在单个快速动作中丢弃所有结果。此类丢弃在查询中的ERS与显示给用户的结果中的各个Jack Wang提及(诸如前十个)之间分派否定共指标记。此类丢弃动作还可以从推荐关系中去除对Sally Smith的提及,使得用户不会被所述关系的任何进一步的提议所

打扰。

[0272] 以上系统、设备、方法、过程等可以用适用于特定应用的硬件、软件或其任何组合来实现。硬件可以包括通用计算机和/或专用计算设备。这包括在一个或多个微处理器、微控制器、嵌入式微控制器、可编程数字信号处理器或其他可编程设备或处理电路以及内部和/或外部存储器中实现。这还可以或替代地包括一个或多个专用集成电路、可编程门阵列、可编程阵列逻辑组件或可以被配置成处理电子信号的一个或多个任何其他设备。将进一步理解的是,上述过程或设备的实现可以包括使用结构化编程语言(诸如C)、面向对象编程语言(诸如C++)或可以被存储、编译或解读以在上述设备中的一个上运行的任何其他高级或低级编程语言(包括汇编语言、硬件描述语言以及数据库编程语言和技术)以及处理器的异构组合、处理器架构或不同硬件和软件的组合创建的计算机可执行代码。在另一方面中,方法可以在执行其步骤的系统中体现,并且可以用数种方式跨设备分布。同时,处理可以跨设备(诸如,上述各种系统)分布,或者可以将所有功能集成到专用的独立设备或其他硬件中。在另一方面中,用于执行与上述过程相关联的步骤的装置可以包括上述任何硬件和/或软件。所有此类置换和组合旨在落入本公开文本的范围内。

[0273] 本文中公开的实施方案可以包括计算机程序产品,其包括当在一个或多个计算设备上执行时执行其任何和/或所有步骤的计算机可执行代码或计算机可用代码。所述代码可以被以非暂时性方式存储在计算机存储器中,其可以是程序从其执行的存储器(诸如,与处理器相关联的随机存取存储器)或者存储设备,诸如,盘驱动、闪存或任何其他的光学、电磁、磁性、红外或其他设备或设备组合。在另一方面中,可以用承载计算机可执行代码和/或其任何输入或输出的任何适当传输或传播介质中体现上述任何系统和方法。

[0274] 本文中所描述和描绘的元件(包括贯穿附图的流程图和框图)暗示元件之间的逻辑边界。然而,根据软件或硬件工程实践,所描述的元件及其功能可以通过计算机可执行介质实现于机器上,所述计算机可执行介质具有处理器,所述处理器能够执行存储于其上的作为单片软件结构、作为独立软件模块、作为采用外部例程、代码、服务等的模块、或其任何组合的程序指令,并且所有此类实现方式都可以在本公开文本的范围内。此类机器的实施例可以包括但不限于个人数字助理、膝上型计算机、个人计算机、移动电话、其他手持式计算设备、医疗装备、有线或无线通信设备、换能器、芯片、计算器、卫星、平板PC、电子书、小组件、电子设备、具有人工智能的设备、计算设备、网络装备、服务器、路由器等。此外,在流程图和框图中所描绘的元素或任何其他逻辑部件可以在能够执行程序指令的机器上实现。因此,虽然前述附图和描述阐述了所公开系统的各功能方面,但不应从这些描述中推断用于实现这些功能方面的软件的具体布置,除非明确说明或以其他方式从上下文中清楚。类似地,可以理解的是,上文所标识和描述的各种步骤都是可以变化的,并且这些步骤的顺序可以适于本文中所公开的技术的具体应用。所有此类变化和修改均旨在落入本公开文本的范围之内。这样,对各个步骤的顺序的描绘和/或描述不应被理解为要求对这些步骤的具体执行顺序,除非具体应用要求或者明确陈述或以其他方式从上下文中清楚。在没有相反的明确指示的情况下,在不脱离本公开文本的范围的情况下,可以对公开的步骤进行修改、补充、省略和/或重排序。

[0275] 本文中所描述的实现方式的方法步骤旨在包括与以下权利要求的专利性一致的致使执行此类方法步骤的任何适当方法,除非明确地提供了不同的含义或以其他方式从上

下文中清楚。因此,例如执行X的步骤包括用于致使另一方(诸如远程用户、远程处理资源(例如,服务器或云计算机)或机器执行X的步骤的任何适当方法。类似地,执行步骤X、Y和Z可以包括引导或控制此类其他个体或资源的任何组合执行步骤X、Y和Z以获得此类步骤的益处的任何方法。因此,本文中所描述的实现方式的方法步骤旨在包括与以下权利要求的专利性一致的致使一个或多个其他参与方或实体执行步骤的任何适当方法,除非明确地提供了不同的含义或以其他方式从上下文中清楚。此类各方或实体不需要在任何另一方或实体的引导或控制下,并且不需要位于特定管辖区域内。

[0276] 将理解的是,上述方法和系统是通过举例而非限制的方式阐述的。许多变化、添加、省略及其他修改对于本领域技术人员而言将是显而易见的。另外,以上描述和附图中的方法步骤的顺序或呈现并不旨在要求执行所叙述步骤的此顺序,除非明确地要求特定顺序或以其他方式从上下文中清楚。因此,虽然已经示出并描述了特定实施方案,但对于本领域技术人员而言将显而易见的是在不脱离本公开文本的精神和范围的情况下可以进行形式和细节方面的各种改变和修改,并且其旨在构成如将在法律允许的最宽泛意义上解读的以下权利要求定义的本发明的一部分。

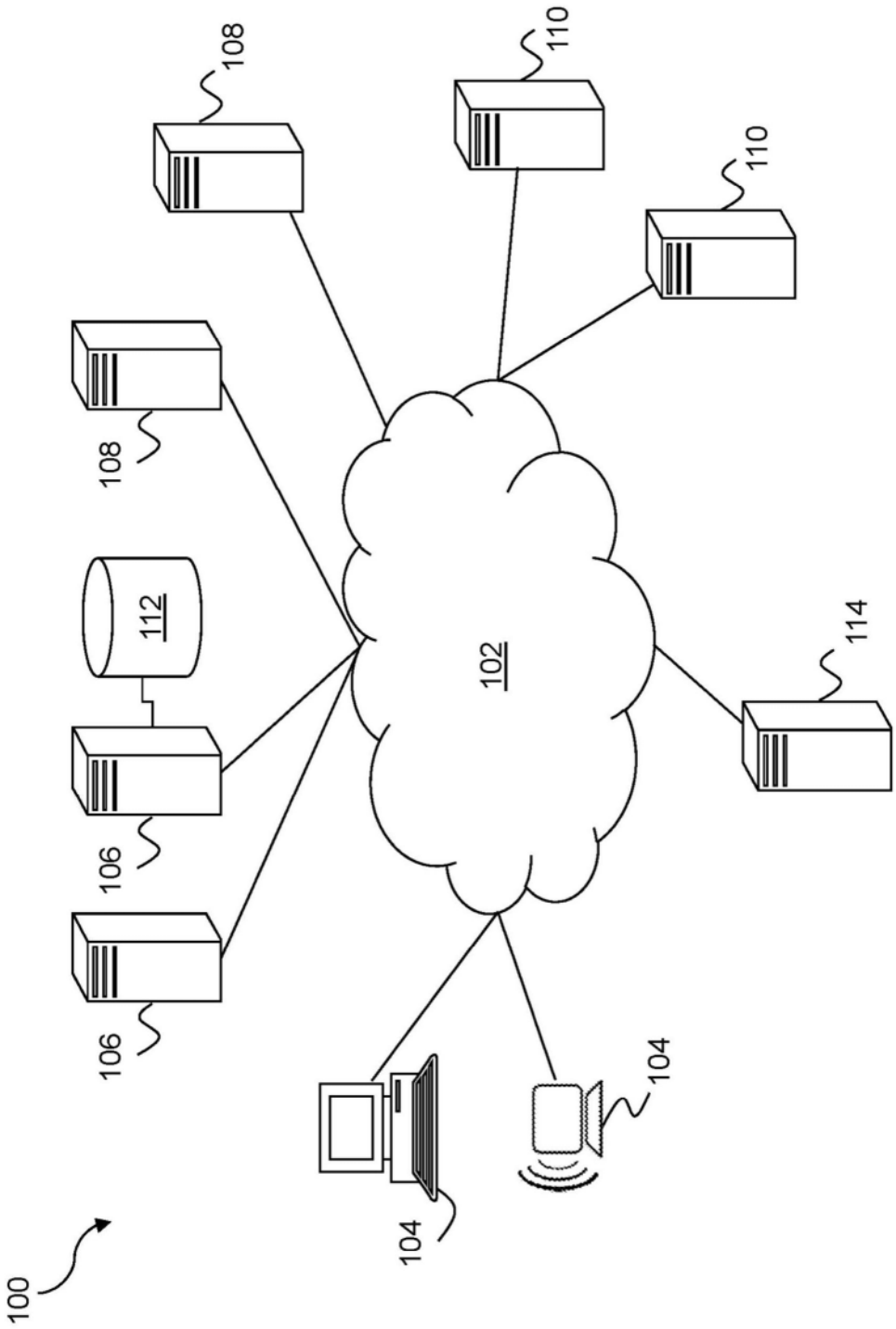


图1

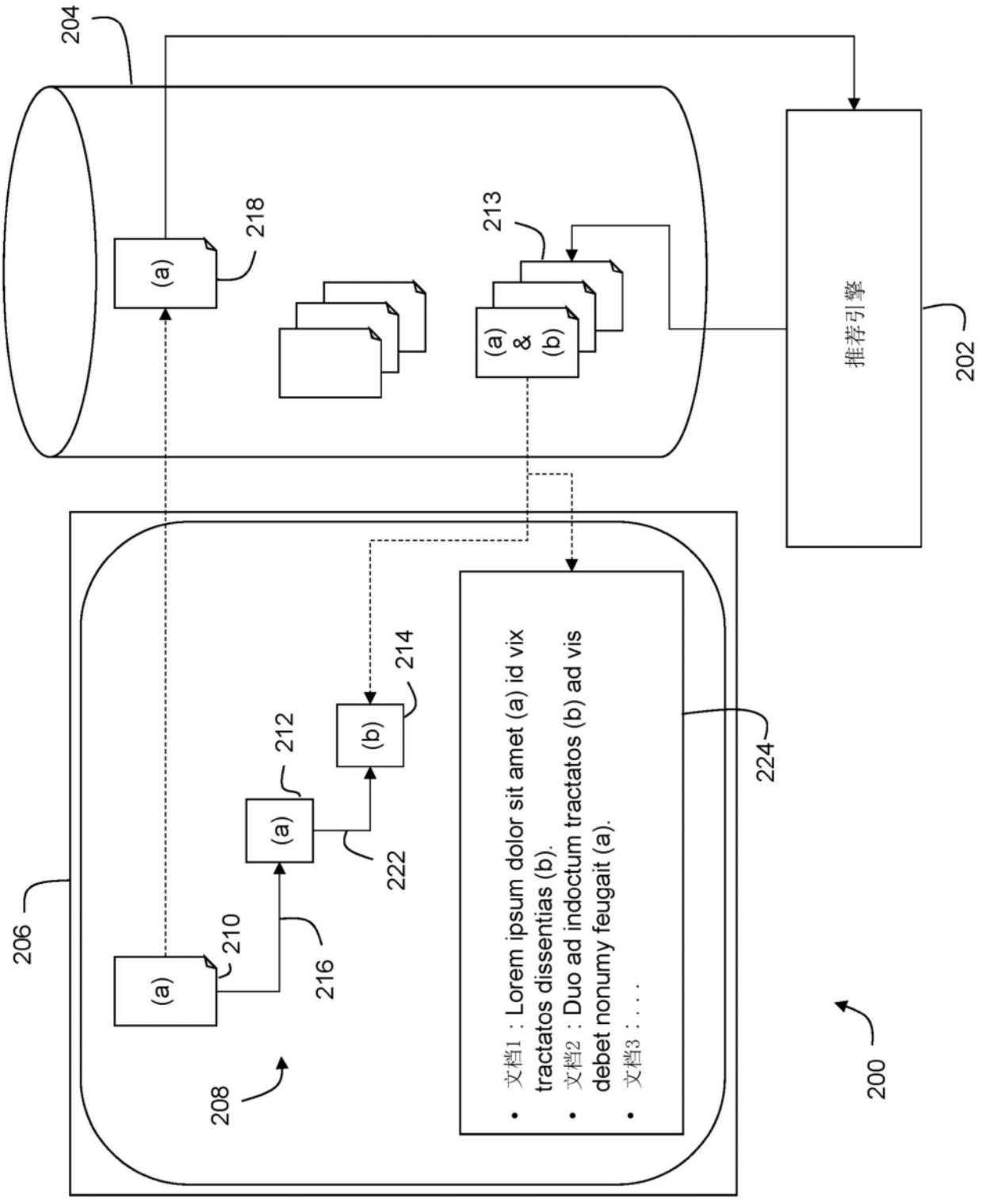

204
218
(a)
213
(a)
&
(b)
推荐引擎
202
206
(a)
210
216
(a)
212
222
(b)
214
208
224
• 文档1：Lorem ipsum dolor sit amet (a) id vix tractatos dissentias (b).
• 文档2：Duo ad indoctum tractatos (b) ad vis debet nonumy feugait (a).
• 文档3：. . .
200

图2

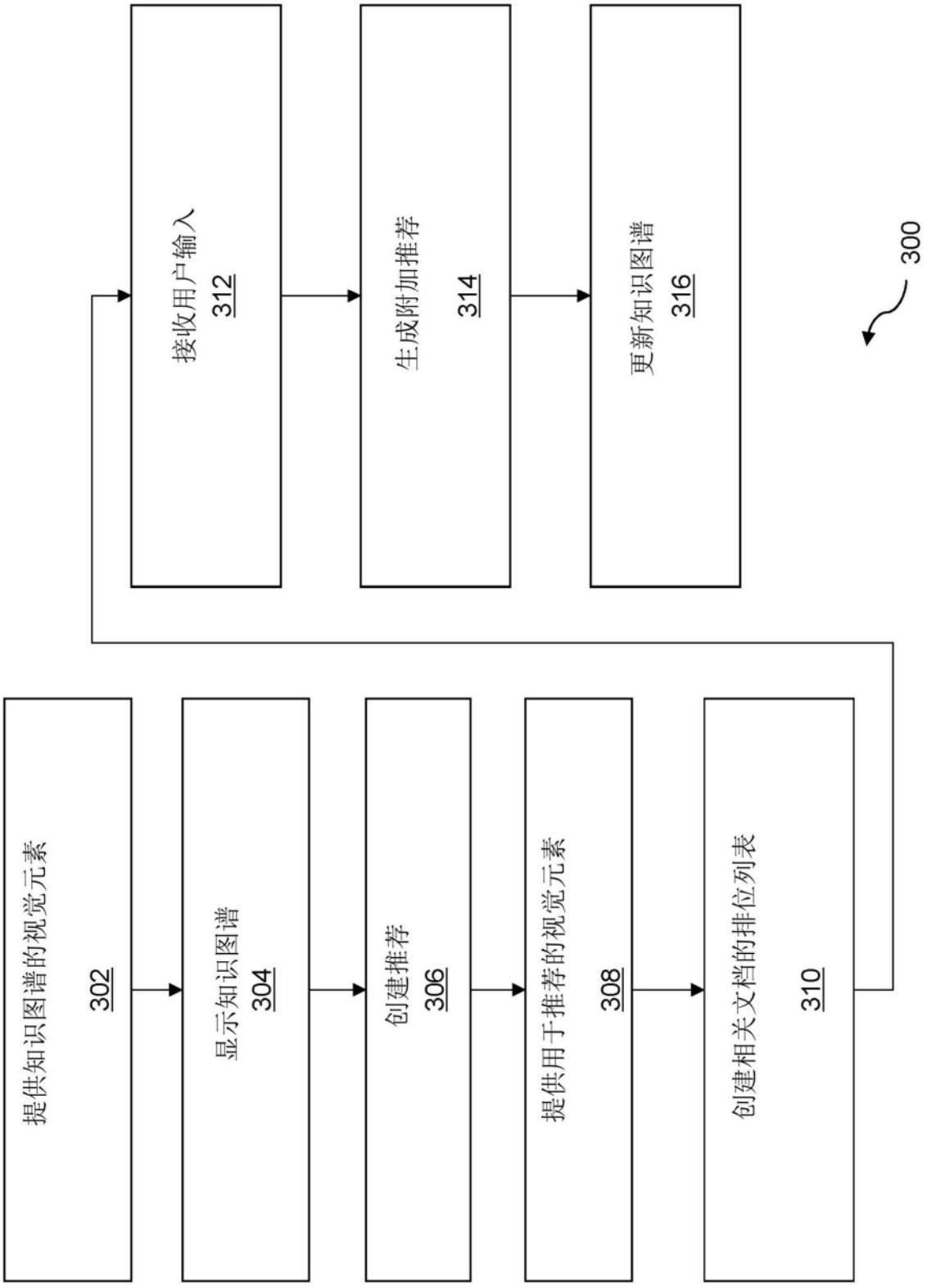

提供知识图谱的视觉元素
302
显示知识图谱
304
创建推荐
306
提供用于推荐的视觉元素
308
创建相关文档的排位列表
310
接收用户输入
312
生成附加推荐
314
更新知识图谱
316
300

图3

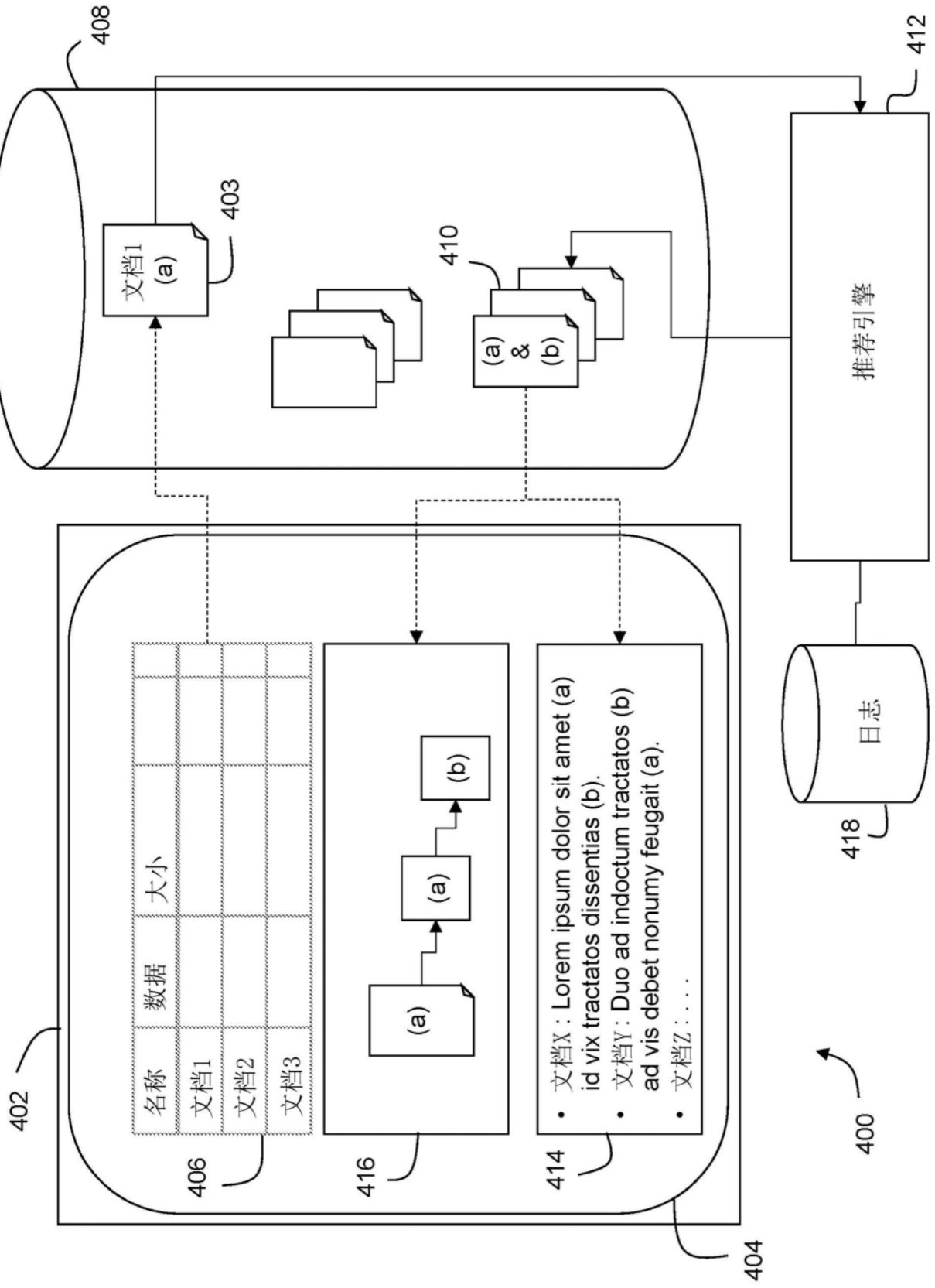

408
412
403
文档1
(a)
410
(a)
&
(b)
推荐引擎
402
名称
数据
大小
文档1
文档2
文档3
406
(a)
(a)
(b)
416
• 文档X : Lorem ipsum dolor sit amet (a)
id vix tractatos dissentias (b).
• 文档Y : Duo ad indoctum tractatos (b)
ad vis debet nonumy feugait (a).
• 文档Z : . . .
414
日志
418
400
404

图4

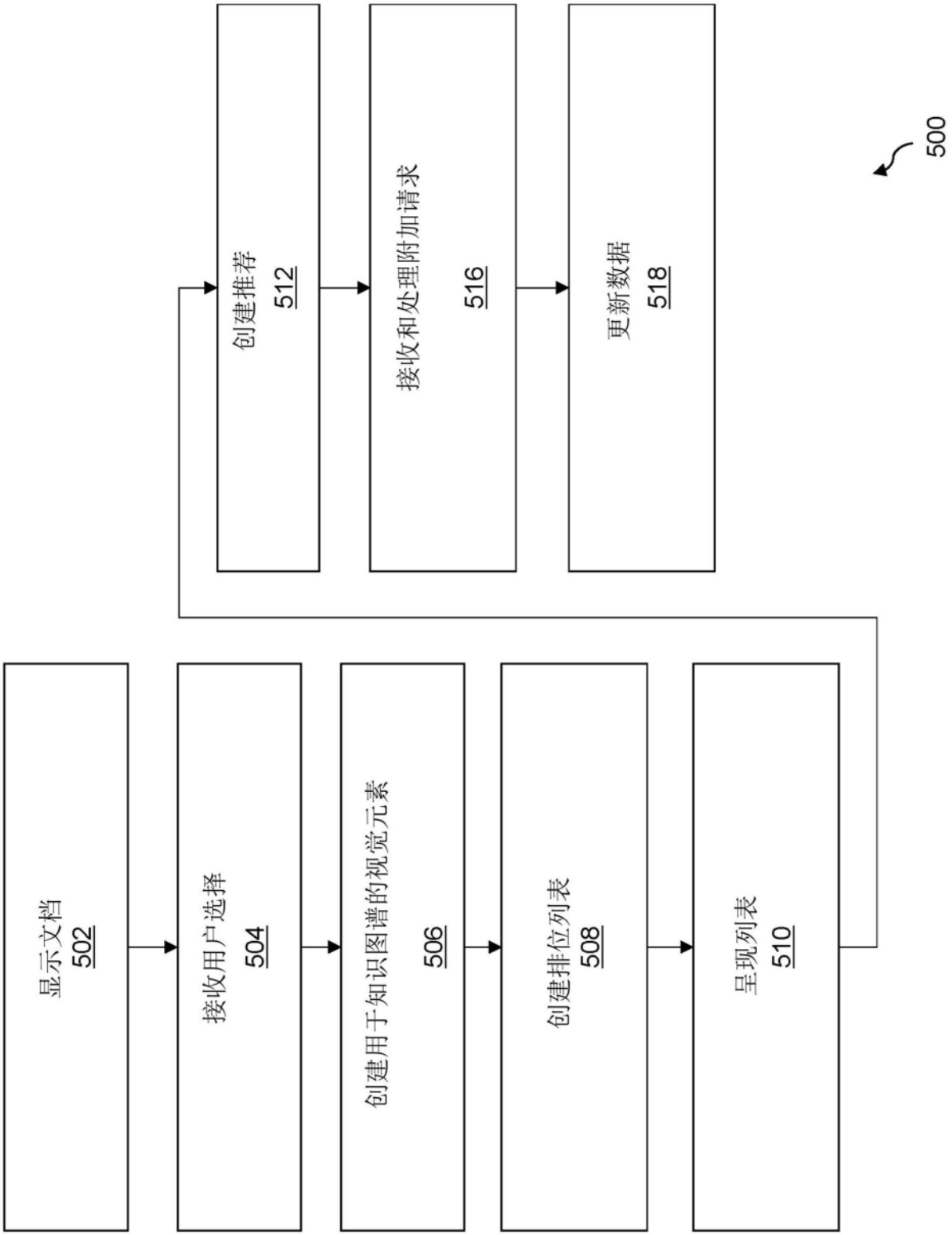

500
显示文档
502
接收用户选择
504
创建用于知识图谱的视觉元素
506
创建排位列表
508
呈现列表
510
创建推荐
512
接收和处理附加请求
516
更新数据
518

图5

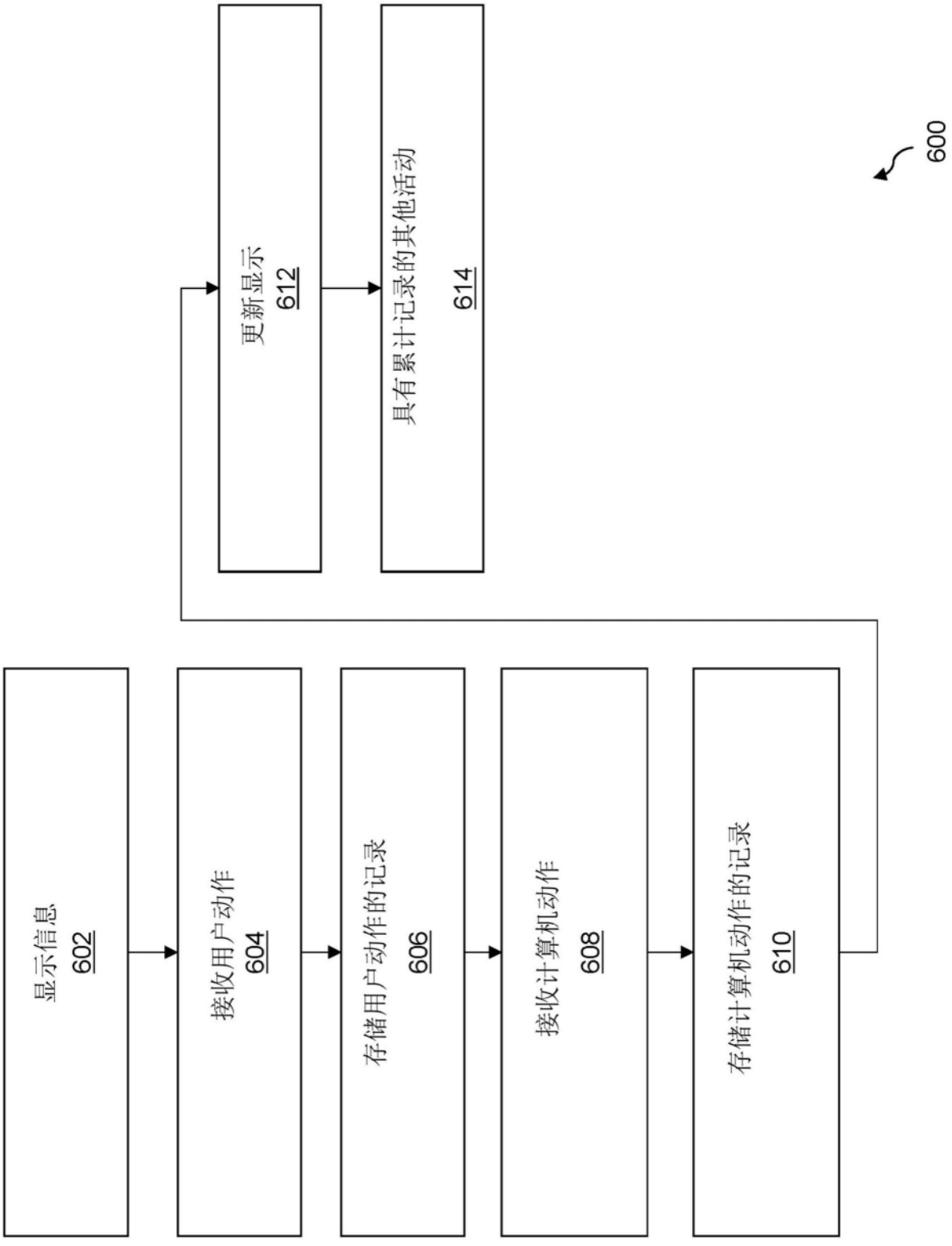

600
显示信息
602
接收用户动作
604
存储用户动作的记录
606
接收计算机动作
608
存储计算机动作的记录
610
更新显示
612
具有累计记录的其他活动
614

图6

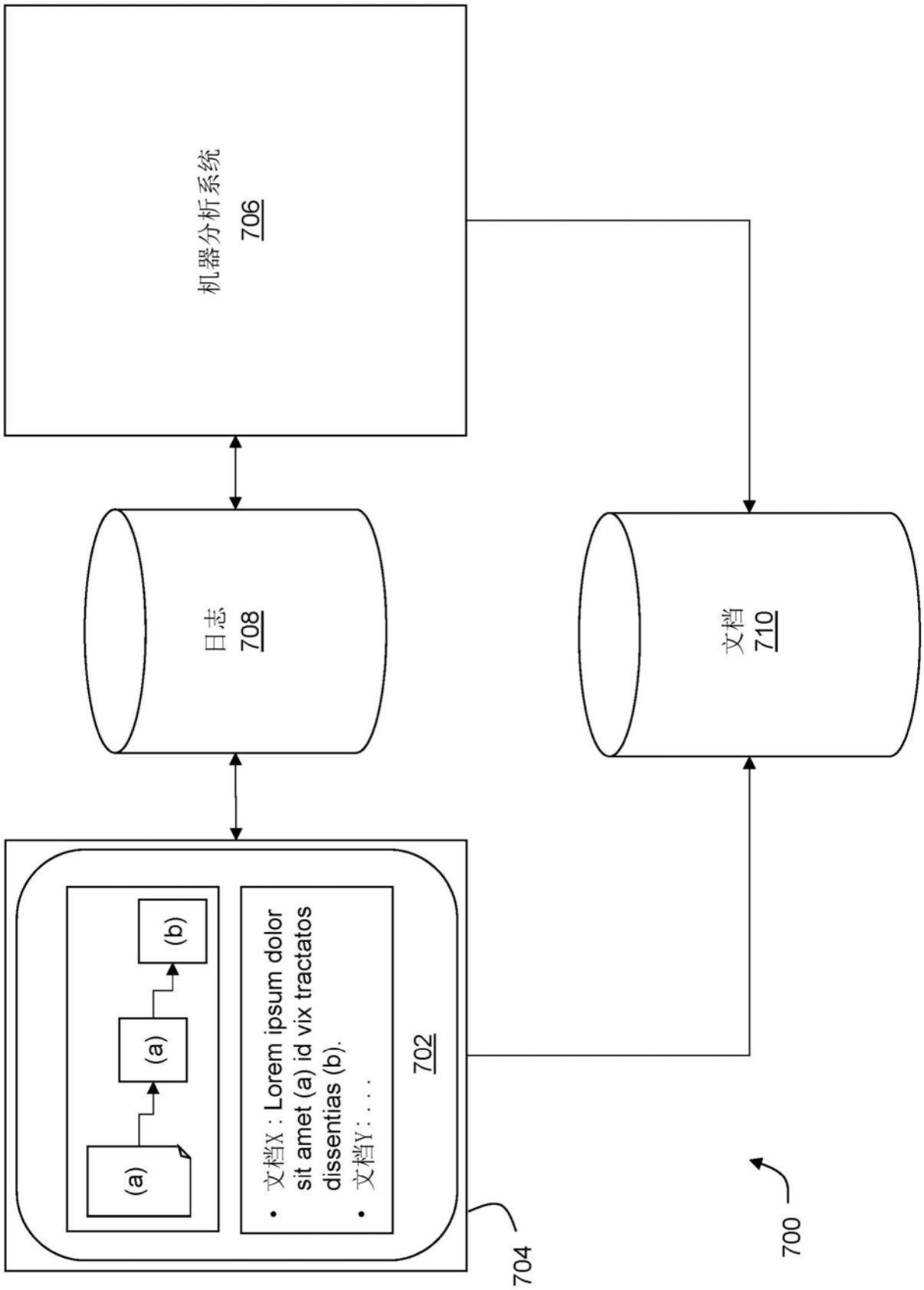

机器分析系统
706
日志
708
文档
710
(a)
(a)
(b)
• 文档X : Lorem ipsum dolor sit amet (a) id vix tractatos dissentias (b).
• 文档Y : . . .
702
704
700

图7

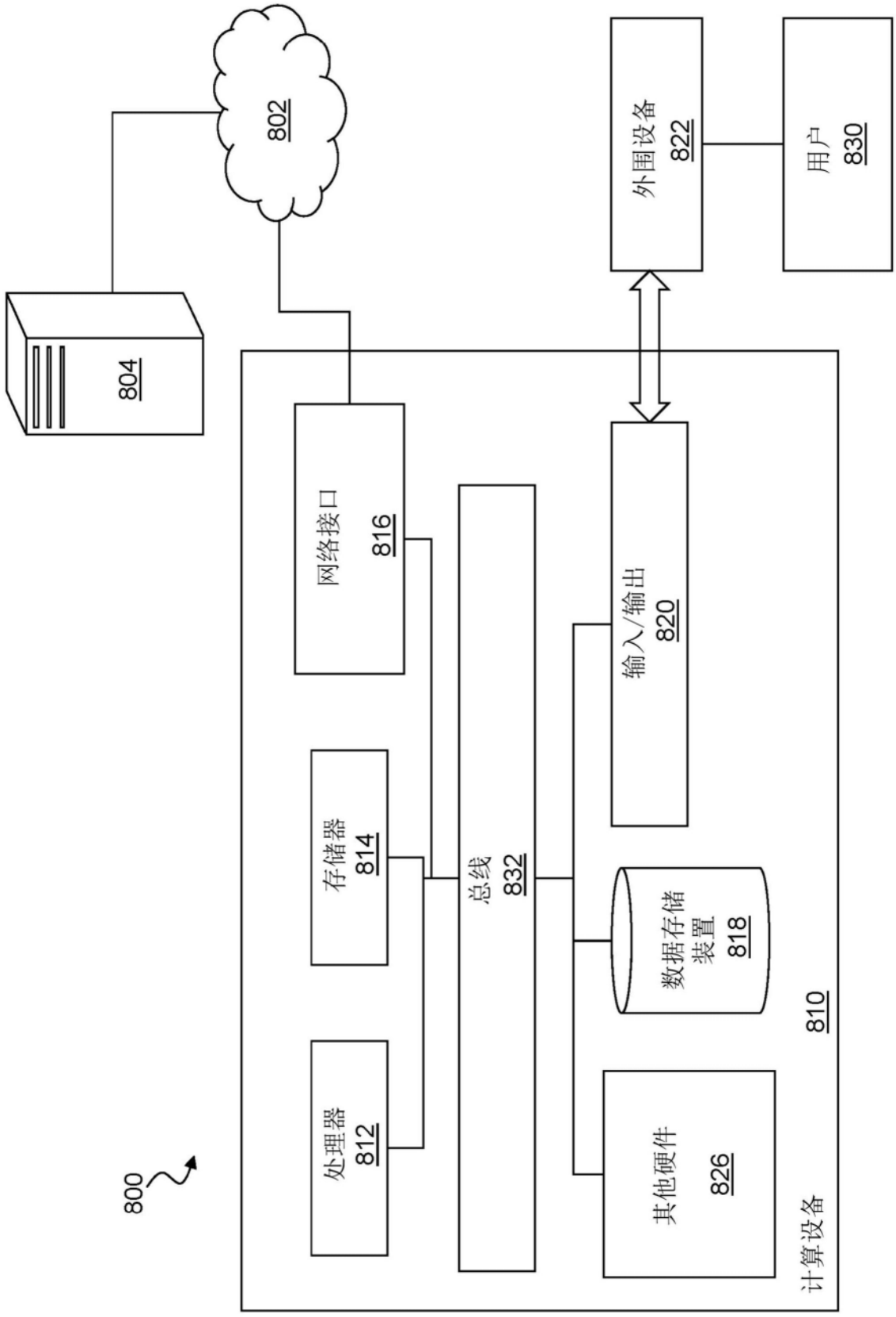

802
外围设备
822
用户
830
804
网络接口
816
输入/输出
820
存储器
814
总线
832
数据存储
装置
818
处理器
812
其他硬件
826
计算设备
810
800

图8

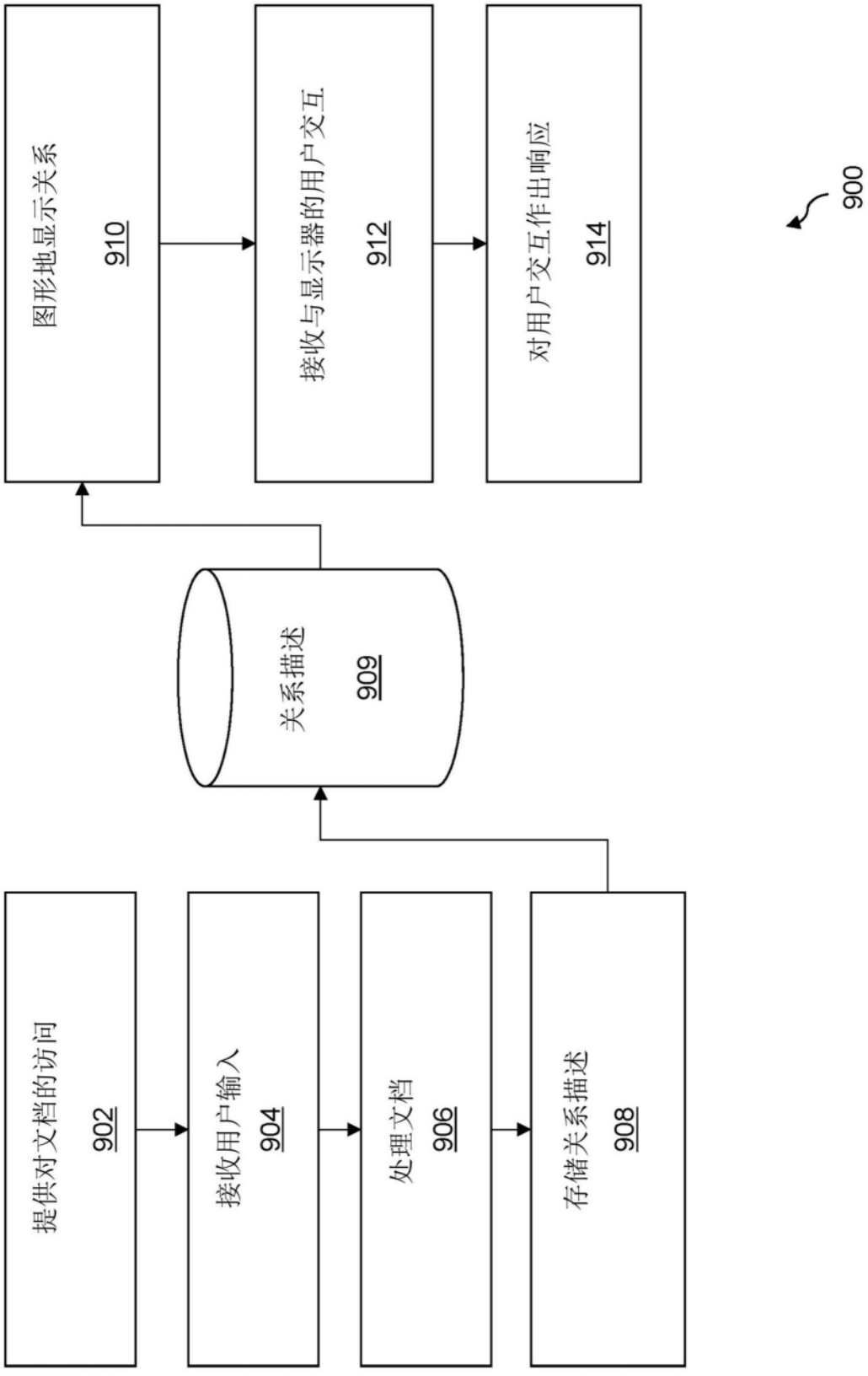

提供对文档的访问
902
接收用户输入
904
处理文档
906
存储关系描述
908
关系描述
909
图形地显示关系
910
接收与显示器的用户交互
912
对用户交互作出响应
914
900

图9

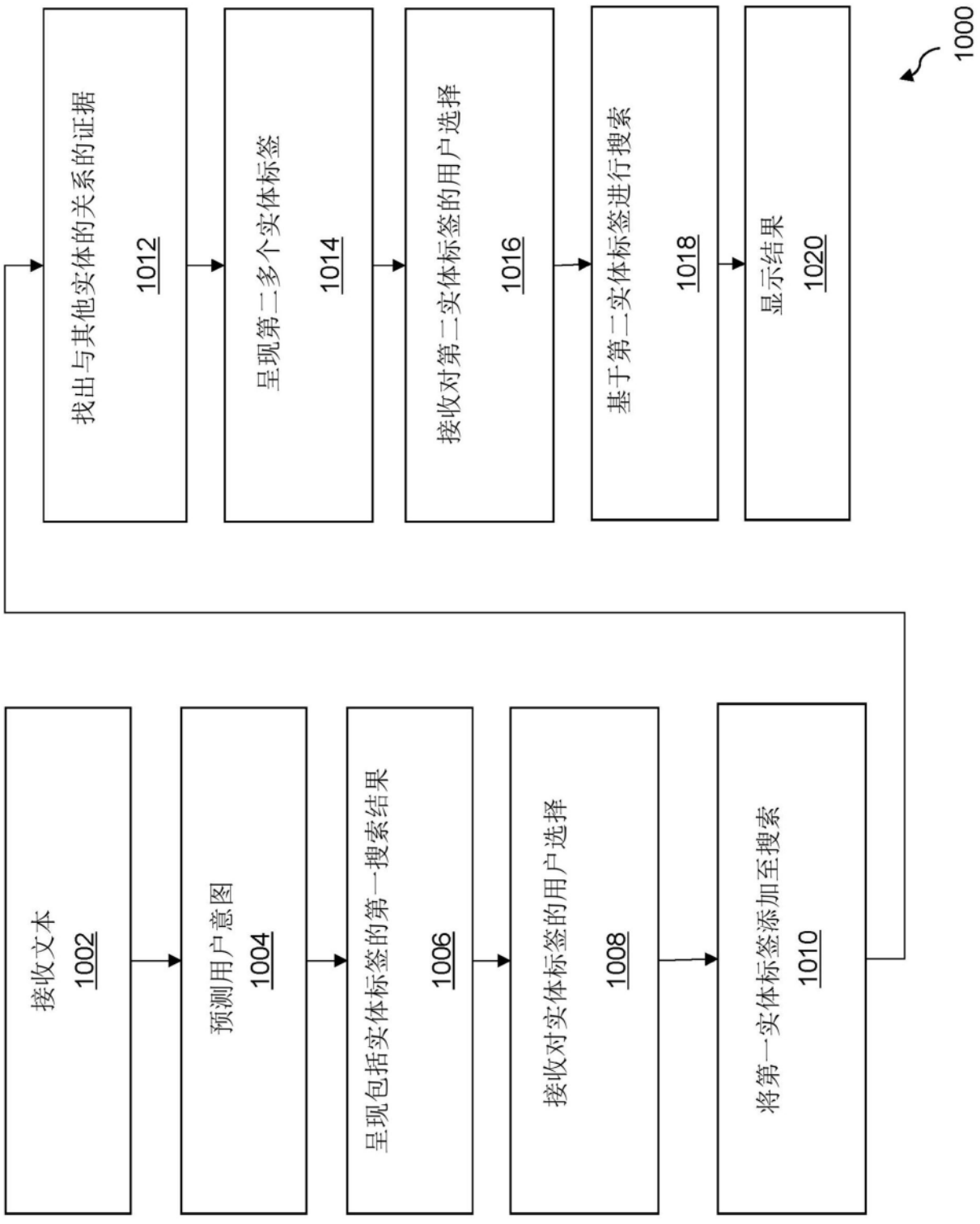

1000
找出与其他实体的关系的证据
1012
呈现第二多个实体标签
1014
接收对第二实体标签的用户选择
1016
基于第二实体标签进行搜索
1018
显示结果
1020
接收文本
1002
预测用户意图
1004
呈现包括实体标签的第一搜索结果
1006
接收对实体标签的用户选择
1008
将第一实体标签添加至搜索
1010

图10

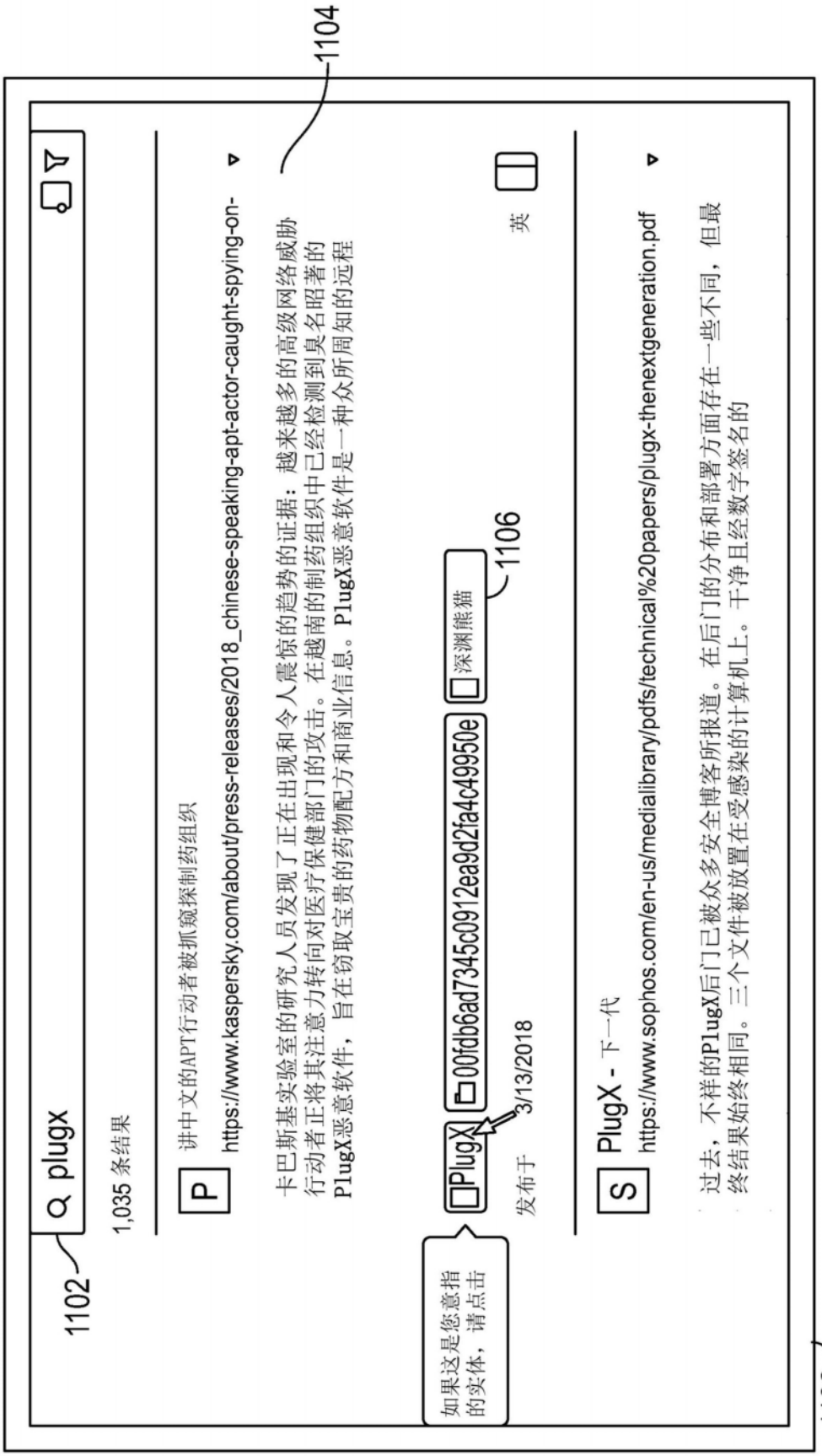

1102
plugx
1,035 条结果
讲中文的APT行动者被抓窥探制药组织
https://www.kaspersky.com/about/press-releases/2018_chinese-speaking-apt-actor-caught-spying-on-
1104
卡巴斯基实验室的研究人员发现了正在出现和令人震惊的趋势的证据：越来越多的高级网络威胁行动者正将其注意力转向对医疗保健部门的攻击。在越南的制药组织中已经检测到臭名昭著的P1ugX恶意软件，旨在窃取宝贵的药物配方和商业信息。P1ugX恶意软件是一种众所周知的远程
如果这是您意指的实体，请点击
PlugX
00fdb6ad7345c0912ea9d2fa4c49950e
深渊熊猫
1106
发布于 3/13/2018
英
PlugX - 下一代
https://www.sophos.com/en-us/medialibrary/pdfs/technical%20papers/plugx-thenextgeneration.pdf
过去，不祥的P1ugX后门已被众多安全博客所报道。在后门的分布和部署方面存在一些不同，但最终结果始终相同。三个文件被放置在受感染的计算机上。干净且经数字签名的
1100

图11

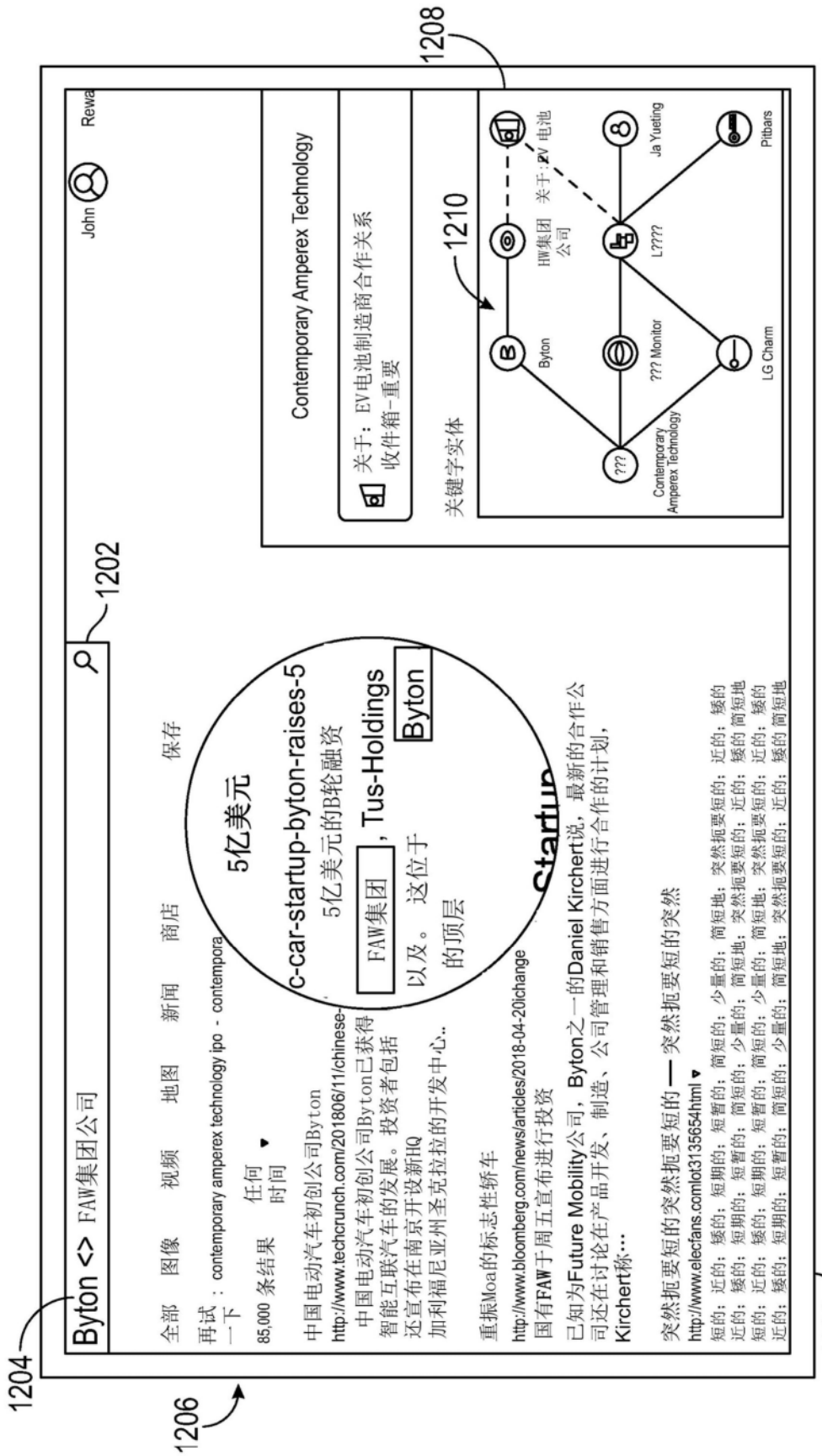

1200
1202
1204
1206
1208
1210
Byton <> FAW集团公司
John
Rewa
全部 图像 视频 地图 新闻 商店 保存
再试一下：contemporary amperex technology ipo - contempora
85,000 条结果 任何时间
中国电动汽车初创公司Byton
http://www.techcrunch.com/201806/11/chinese-
中国电动汽车初创公司Byton已获得智能互联汽车的发展。投资者包括还宣布在南京开设新HQ加利福尼亚州圣克拉拉的开发中心..
重振Moa的标志性轿车
http://www.bloomberg.com/news/articles/2018-04-20ichange
国有FAW于周五宣布进行投资
已知为Future Mobility公司，Byton之一的Daniel Kirchert说，最新的合作公司还在讨论在产品开发、制造、公司管理和销售方面进行合作的计划，Kirchert称…
突然扼要短的突然扼要短的 — 突然扼要短的突然
http://www.elecfans.comlot3135654html
短的；近的；矮的；短期的；短暂的；简短的；少量的；简短地；突然扼要短的；近的；矮的
近的；矮的；短期的；短暂的；简短的；少量的；简短地；突然扼要短的；近的；矮的简短地
短的；近的；矮的；短期的；短暂的；简短的；少量的；简短地；突然扼要短的；近的；矮的
近的；矮的；短期的；短暂的；简短的；少量的；简短地；突然扼要短的；近的；矮的简短地
5亿美元
c-car-startup-byton-raises-5
5亿美元的B轮融资
FAW集团 , Tus-Holdings
以及。 这位于
的顶层
Byton
Startup
Contemporary Amperex Technology
关于：EV电池制造商合作关系
收件箱-重要
关键字实体
Byton
??? Monitor
LG Charm
HW集团公司
关于：EV 电池
L????
Ja Yueting
Pitbars
???
Contemporary Amperex Technology

图12

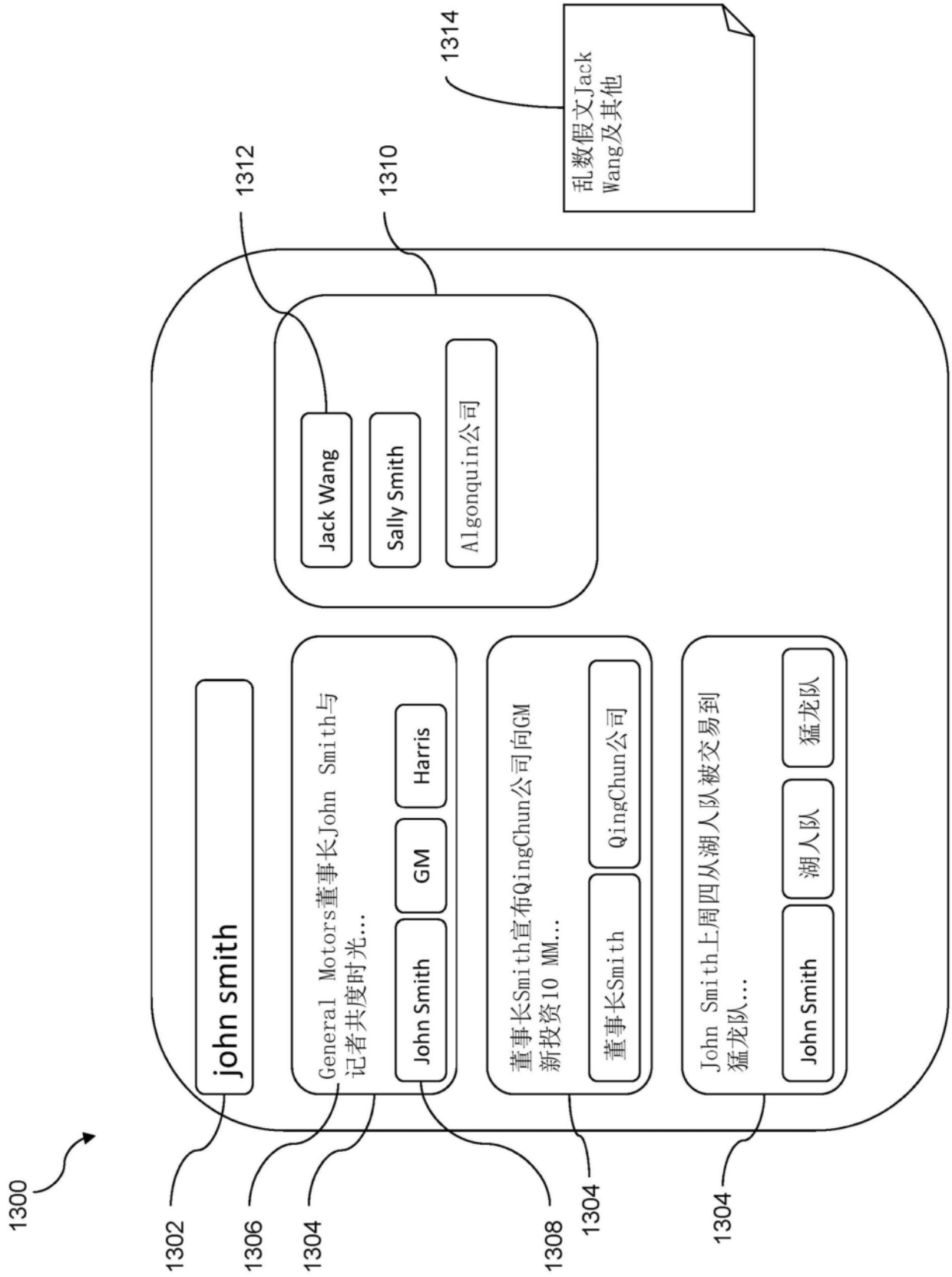

1300
1302
john smith
1306
General Motors董事长John Smith与记者共度时光...
John Smith
GM
Harris
1304
1308
董事长Smith宣布QingChun公司向GM新投资10 MM...
董事长Smith
QingChun公司
1304
John Smith上周四从湖人队被交易到猛龙队...
John Smith
湖人队
猛龙队
1310
Jack Wang
Sally Smith
Algonquin公司
1312
1314
乱数假文Jack Wang及其他

图13

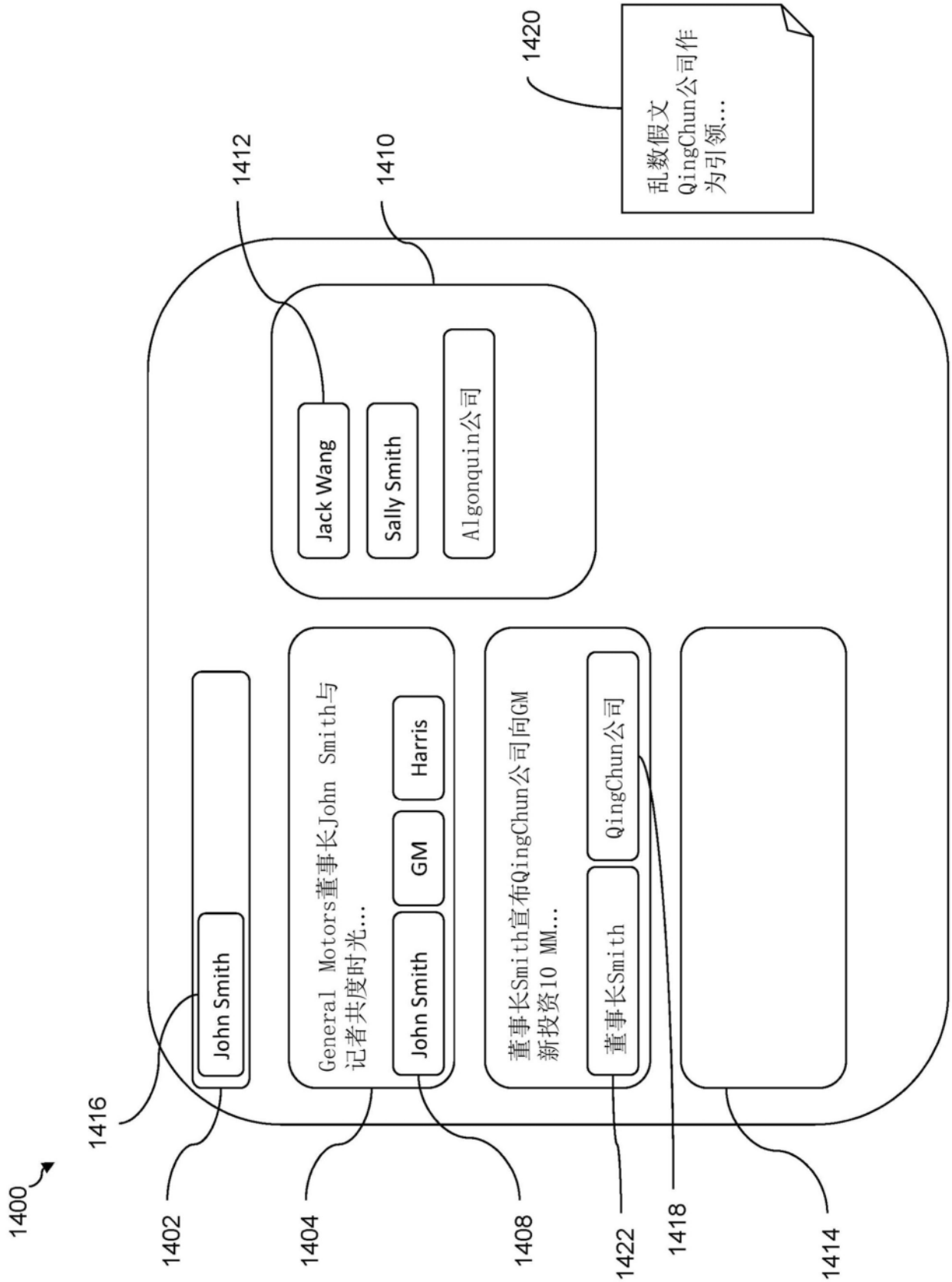

1400
1416
1402
John Smith
1404
General Motors董事长John Smith与记者共度时光...
John Smith
GM
Harris
1408
董事长Smith宣布QingChun公司向GM新投资10 MM...
董事长Smith
QingChun公司
1422
1418
1414
1412
Jack Wang
Sally Smith
Algonquin公司
1410
1420
乱数假文
QingChun公司作为引领...

图14

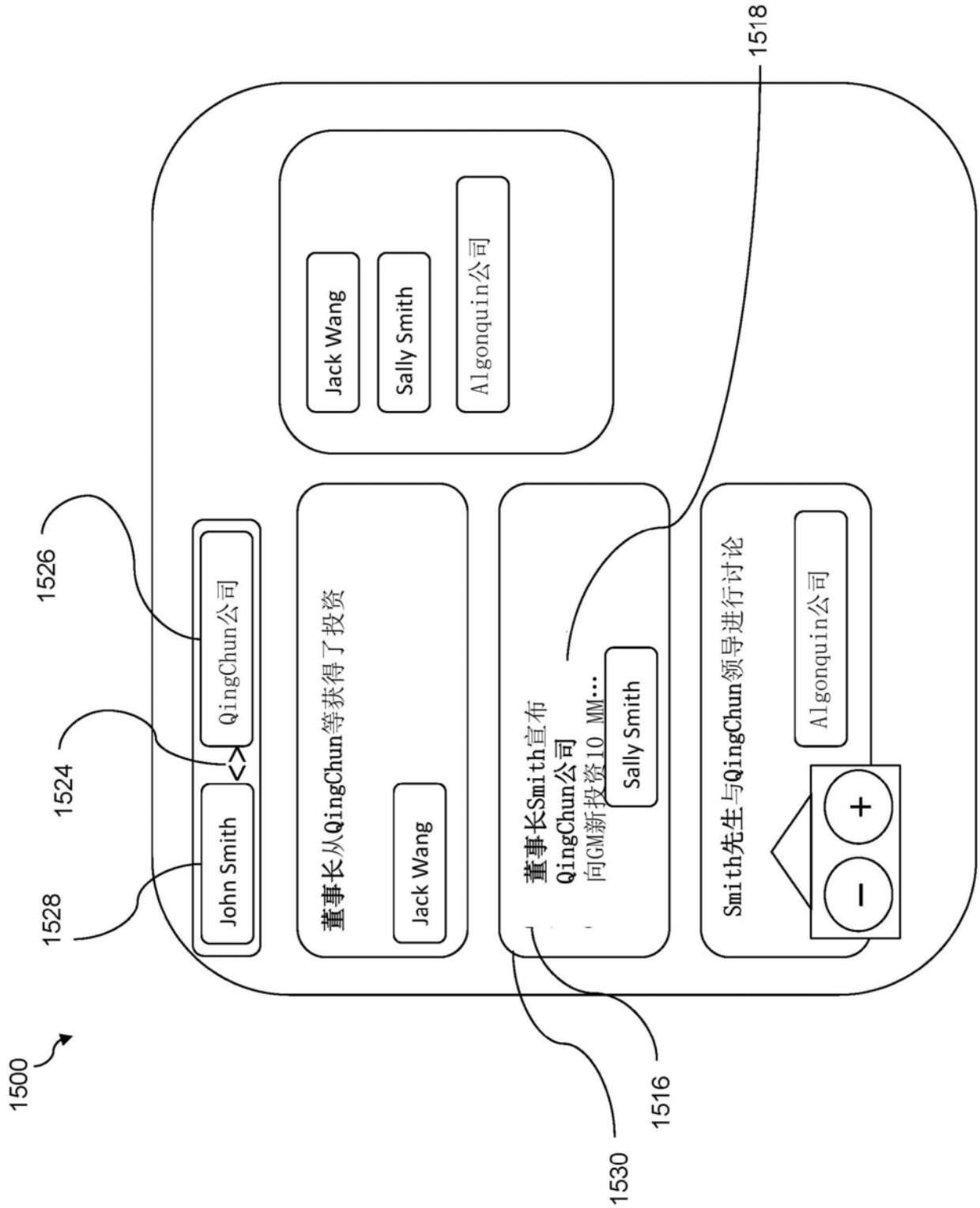

1500
1518
1526
1524
1528
1516
1530
John Smith
QingChun公司
董事长从QingChun等获得了投资
Jack Wang
董事长Smith宣布
QingChun公司
向GM新投资10 MM…
Sally Smith
Smith先生与QingChun领导进行讨论
Algonquin公司
Jack Wang
Sally Smith
Algonquin公司

图15

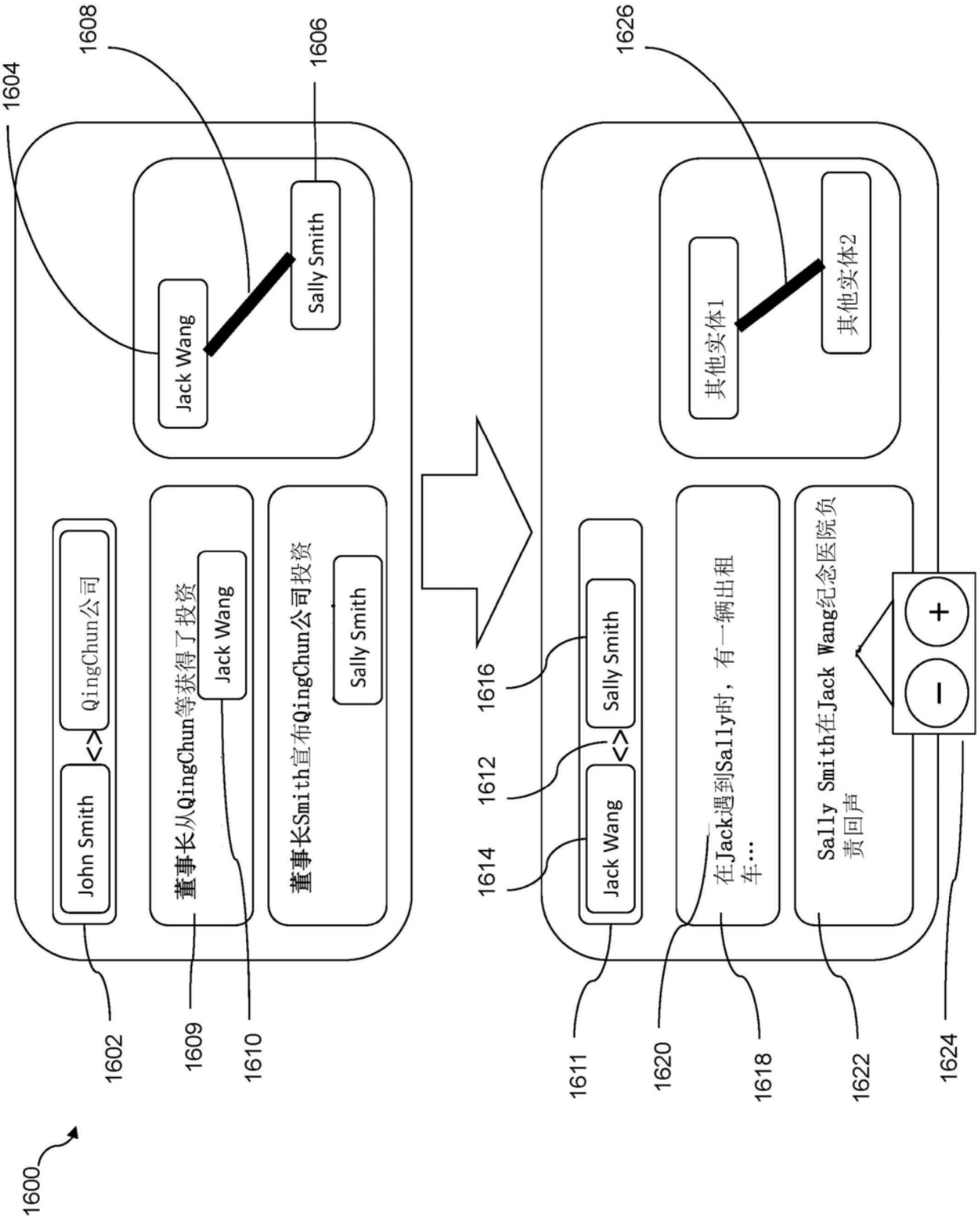

1600
1602
John Smith
<>
QingChun公司
1604
1606
Sally Smith
1608
Jack Wang
1609
董事长从QingChun等获得了投资
Jack Wang
1610
董事长Smith宣布QingChun公司投资
Sally Smith
1611
Jack Wang
1612
<>
1614
1616
Sally Smith
1618
在Jack遇到Sally时，有一辆出租车…
1620
1622
Sally Smith在Jack Wang纪念医院负责回声
1624
+
−
1626
其他实体1
其他实体2

图16