



(12) 发明专利

(10) 授权公告号 CN 106716408 B

(45) 授权公告日 2021.09.10

(21) 申请号 201580049293.5

D·叶

(22) 申请日 2015.09.22

(74) 专利代理机构 中国贸促会专利商标事务所
有限公司 11038

(65) 同一申请的已公布的文献号
申请公布号 CN 106716408 A

代理人 李玲

(43) 申请公布日 2017.05.24

(51) Int.Cl.

(30) 优先权数据

G06F 16/33 (2019.01)

62/053,283 2014.09.22 US

G06F 16/338 (2019.01)

14/643,390 2015.03.10 US

G06F 40/211 (2020.01)

G06F 40/30 (2020.01)

(85) PCT国际申请进入国家阶段日
2017.03.14

(56) 对比文件

(86) PCT国际申请的申请数据

US 2013254209 A1, 2013.09.26

PCT/US2015/051403 2015.09.22

CN 103544242 A, 2014.01.29

(87) PCT国际申请的公布数据

CN 101510197 A, 2009.08.19

W02016/048996 EN 2016.03.31

US 6901399 B1, 2005.05.31

(73) 专利权人 甲骨文国际公司
地址 美国加利福尼亚

CN 103530415 A, 2014.01.22

US 2014040288 A1, 2014.02.06

审查员 袁勤

(72) 发明人 V·泽勒温斯基 Y·达塞乌斯凯

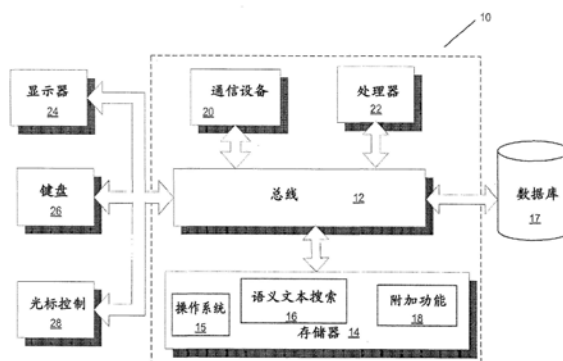
权利要求书3页 说明书9页 附图10页

(54) 发明名称

语义文本搜索

(57) 摘要

一种用于执行语义搜索的系统接收电子文本语料库并且将文本语料库分离成多个语句。该系统解析每一个语句并且将其转换成语句树。该系统接收搜索查询并且将搜索查询与语句树中的一个或多个匹配。



1. 一种计算机可读介质,所述计算机可读介质具有存储在其上的指令,当所述指令被处理器执行时,引起所述处理器执行语义搜索,所述语义搜索包括:

接收电子文本语料库;

将所述文本语料库分离成多个语句;

使用自然语言解析器来解析每一个语句并且将每一个语句转换成语句树;

经由用户界面接收搜索查询;

将所述搜索查询转换成查询树;

将所述查询树与所述语句树中的一个或多个匹配,所述匹配包括确定所述查询树是否是所述语句树中的任何语句树的子树;

响应于所述匹配,提供与匹配的语句树对应的所述搜索查询的细分,其中所述细分包括通过添加一个或多个附加项来增加所述查询树,并且其中所述细分包括展宽的细分和横向的细分,并且所述细分可以根据所述查询树的项之间的依赖关系的类型被分组;以及

响应于所述匹配,提取与匹配的语句树对应的所述搜索查询的情感,其中提取情感包括提取和检索与所述搜索查询的搜索项在语法上联系的修饰语。

2. 如权利要求1所述的计算机可读介质,其中所述一个或多个附加项通过直接边连接到所述搜索查询的搜索项。

3. 如权利要求1所述的计算机可读介质,所述语义搜索还包括:

响应于所述匹配,提供来自所述文本语料库的与匹配的语句树对应的文档;以及

使用指代消解来生成语义搜索结果。

4. 如权利要求1所述的计算机可读介质,其中所述细分被实现为用户界面上的细分链接。

5. 如权利要求1所述的计算机可读介质,所述语义搜索还包括使用所述文本语料库外部的源丰富所述语句树。

6. 如权利要求1所述的计算机可读介质,其中所述用户界面显示所述查询树的上位词匹配。

7. 如权利要求1所述的计算机可读介质,所述语义搜索还包括修改所述语句树中的一个或多个,其中每一个语句树包括节点和至少一个边,并且所述修改包括以下中的至少一个:展开节点、折叠节点、展开边、或折叠边。

8. 如权利要求1所述的计算机可读介质,其中所述查询树包括至少一个节点,并且一个节点与所述搜索查询中的一个搜索项对应。

9. 如权利要求1所述的计算机可读介质,所述语义搜索还包括:在接收所述搜索查询之后,经由所述用户界面提供可选择的细分搜索查询。

10. 一种执行语义搜索的方法,所述方法包括:

接收电子文本语料库;

将所述文本语料库分离成多个语句;

使用自然语言解析器来解析每一个语句并且将每一个语句转换成语句树;

经由用户界面接收搜索查询;

将所述搜索查询转换成查询树;

将所述查询树与所述语句树中的一个或多个匹配,所述匹配包括确定所述查询树是否

是所述语句树中的任何语句树的子树；

响应于所述匹配,提供与匹配的语句树对应的所述搜索查询的细分,其中所述细分包括通过添加一个或多个附加项来增加所述查询树,并且其中所述细分包括展宽的细分和横向的细分,并且所述细分可以根据所述查询树的项之间的依赖关系的类型被分组;以及

响应于所述匹配,提取与匹配的语句树对应的所述搜索查询的情感,其中提取情感包括提取和检索与所述搜索查询的搜索项在语法上联系的修饰语。

11. 如权利要求10所述的方法,其中所述一个或多个附加项通过直接边连接到所述搜索查询的搜索项。

12. 如权利要求10所述的方法,还包括:

响应于所述匹配,提供来自所述文本语料库的与匹配的语句树对应的文档;以及

使用指代消解来生成语义搜索结果。

13. 如权利要求10所述的方法,其中所述细分被实现为用户界面上的细分链接。

14. 如权利要求10所述的方法,还包括使用所述文本语料库外部的源丰富所述语句树。

15. 如权利要求10所述的方法,其中所述用户界面显示所述查询树的上位词匹配。

16. 如权利要求10所述的方法,还包括修改所述语句树中的一个或多个,其中每一个语句树包括节点和至少一个边,并且所述修改包括以下中的至少一个:展开节点、折叠节点、展开边、或折叠边。

17. 如权利要求10所述的方法,其中所述查询树包括至少一个节点,并且一个节点与所述搜索查询中的一个搜索项对应。

18. 如权利要求10所述的方法,还包括:在接收所述搜索查询之后,经由所述用户界面提供可选择的细分搜索查询。

19. 一种语义文本搜索查询系统,包括:

处理器;

耦合到所述处理器的存储设备,其中所述存储设备存储多个语句树,通过将文本语料库分离成多个语句并且使用自然语言解析器来解析每一个语句以及将每一个语句转换成所述语句树中的一个语句树而形成所述多个语句树;

其中所述处理器适于生成接收搜索查询的用户界面并且将所述搜索查询转换成查询树;以及

所述处理器适于将所述查询树与所述语句树中的一个或多个匹配,所述匹配包括确定所述查询树是否是所述语句树中的任何语句树的子树,并且响应于所述匹配,提供与匹配的语句树对应的所述搜索查询的细分,其中所述细分包括通过添加一个或多个附加项来增加所述查询树,并且其中所述细分包括展宽的细分和横向的细分,并且所述细分可以根据所述查询树的项之间的依赖关系的类型被分组,并且响应于所述匹配,提取与匹配的语句树对应的所述搜索查询的情感,其中提取情感包括提取和检索与所述搜索查询的搜索项在语法上联系的修饰语。

20. 如权利要求19所述的语义文本搜索查询系统,其中所述一个或多个附加项通过直接边连接到所述搜索查询的搜索项。

21. 如权利要求19所述的语义文本搜索查询系统,所述处理器还适于:

响应于所述匹配,提供来自所述文本语料库的与匹配的语句树对应的文档;以及

使用指代消解来生成语义搜索结果。

22. 如权利要求19所述的语义文本搜索查询系统,其中所述细分被实现为用户界面上的细分链接。

23. 如权利要求19所述的语义文本搜索查询系统,所述处理器还适于使用所述文本语料库外部的源丰富所述语句树。

24. 如权利要求19所述的语义文本搜索查询系统,所述处理器还适于响应于所述匹配,在所述用户界面上显示所述查询树的上位词匹配。

25. 如权利要求19所述的语义文本搜索查询系统,所述处理器还适于修改所述语句树中的一个或多个,其中每一个语句树包括节点和至少一个边,并且所述修改包括以下中的至少一个:展开节点、折叠节点、展开边或折叠边。

26. 如权利要求19所述的语义文本搜索查询系统,其中所述查询树包括至少一个节点,并且一个节点与所述搜索查询中的一个搜索项对应。

27. 如权利要求19所述的语义文本搜索查询系统,其中所述处理器还适于在接收所述搜索查询之后,经由所述用户界面提供可选择的细分搜索查询。

28. 一种包括用于执行如权利要求10到18中任意一个所述的方法的操作的构件的装置。

语义文本搜索

[0001] 相关申请的交叉引用

[0002] 本申请要求2014年9月22日提交的临时专利申请No.62/053,283的优先权,该专利申请的内容通过引用结合到本文。

技术领域

[0003] 一种实施例一般涉及计算机系统,并且更具体地,涉及提供文本语料库搜索的计算机系统。

背景技术

[0004] 随着可用的基于文本的内容的量持续呈几何级数地增长,在互联网和其它内容储存库(诸如,防火墙后方的数据)二者上,搜索引擎和搜索技术的重要性被加强。实际上每一个用户采用一个或多个搜索引擎基于频率定位相关的内容。随着大量材料可用,已经创建了用于细分搜索引擎结果的具有不同程度成功的各种工具和方法。

[0005] 可用的最普及的搜索引擎主要遵循交互模型,在该交互模型中用户通过搜索引擎界面输入文本搜索项集合,并且该文本搜索项然后用于从由搜索引擎创建或管理的索引中提取结果集合。然而,纯基于文本的搜索的限制之一是,如果使用可以具有多于一个定义或含义的文本搜索项,则检索到的结果集合将不会如可能期望的那样集中于或相关于感兴趣的主体。当用户输入多于一个搜索项时会出现附加的限制。许多搜索引擎将它们对这种多项查询的解释限制为简单请求以定位所有文档,该所有文档包含所有搜索查询项或它们的一些逻辑组合或简单变型(例如,词干提取)。除了最基本的文本文档检索任务外,这种类型的搜索的结果通常并不令人满意。

[0006] 特别地,虽然经由词语来传达含义,但典型的文本或关键词搜索并不搜索含义。待搜索文本的创建者已经在文本内编码了的某些含义。类似地,发起搜索的人在关键词查询中编码期望含义。只有当两个编码相符时,搜索才将返回“正确”结果。

发明内容

[0007] 一种实施例是用于执行语义搜索的系统。该系统接收电子文本语料库并且将文本语料库分离成多个语句。该系统解析每一个语句并且将每一个语句转换成语句树。该系统接收搜索查询并且将搜索查询与语句树中的一个或多个匹配。

附图说明

[0008] 图1是根据本发明的实施例的计算机服务器/系统的框图。

[0009] 图2是根据本发明的一种实施例的图1的语义文本搜索模块和其它元件的功能的高级(high-level)流程图。

[0010] 图3图示根据一种实施例的通过解析语句“The car did not deploy front airbags during the accident”形成的示例树。

[0011] 图4A、图4B和图4C图示根据本发明的实施例的示出已解析的语句和上位词匹配的屏幕截图。

[0012] 图5A和图5B图示根据本发明的实施例的语义搜索用户界面的屏幕截图。

[0013] 图6图示根据一种实施例的经由用户界面的单项修改的屏幕截图。

[0014] 图7A和图7B是图示根据一种实施例的结果集合的细分和总结的示例用户界面。

具体实施方式

[0015] 自计算机初期起存在这样的问题：通过大量电子文档的电子搜索得到对于用户查询而言满意的答案；然而，该问题还未被完全解决。已经存在定位匹配用户查询的文档集合的许多不同方法，包括熟知的搜索引擎，这些搜索引擎包括来自Google公司的“Google”搜索和来自Microsoft公司的“Bing”搜索。

[0016] 总所周知的是，使用无处不在的搜索框的关键词搜索不足以支持许多常见的信息寻找任务。Stewart et al., “Idea Navigation: Structured Browsing for Unstructured Text”, Proceedings of the SIGCHI Conference on Human Factors in Computing System, pp.1789-1792 (2008 ACM) 中公开了一种改进结果的可能的搜索技术，其以引用方式结合到本文。

[0017] 本文中描述的实施例对在大量电子文档中搜索有用信息的问题给出改进的技术解决方案。给出的示例的目的只是用于说明本发明的实施例。本发明的实施例的实际使用情况包括搜索任意大小的文本语料库，该语料库有可能包括数百万或更多的电子文档（例如，电子邮件、文章、书籍、网页、推特(tweet)等），其中，词语的巨大数量使得任何手动信息搜索是不切实际或近乎不可能的，而关键词搜索中固有的查准率/查全率(precision/recall)的权衡使得该方法对于需要高查全率或高查准率的情况没有作用。

[0018] 一种实施例是通过将文本语料库的每一个语句转换成树来执行语义文本搜索的系统。搜索查询然后也被转换成树和/或被解释为树，并且搜索树与文本语料库树中的一个或多个匹配。作为匹配的结果，生成与搜索查询对应的文档的响应。另外，可以生成相关查询的细分(refinement)。树匹配的使用提供基于语义的搜索。本发明的另一种实施例可以定位感兴趣的实体（诸如，品牌或产品名称），并且基于修饰这些感兴趣的实体的其它项来执行高查准率情感提取。

[0019] 通常，实施例不仅基于词语而且基于词语彼此作用和修饰的方式来寻找文本。其它实施例通过用附加的信息（诸如，同义词）丰富文本来应用附加的知识库。实施例用附加的语义知识来扩充文本以允许检索期间的高查全率，并且尽可能多地利用文本的下层结构以获得高查准率。

[0020] 图1是根据本发明的实施例的计算机服务器/系统10的框图。虽然示出为单个系统，但系统10的功能可以实现为分布式系统。另外，可以在可以通过网络耦合在一起的分离的服务器或设备上实现本文中公开的功能。另外，可以不包括系统10的一个或多个组件。例如，对于执行语义文本搜索的服务器的功能，系统10可以不包括外围设备，诸如键盘26和光标控制28。

[0021] 系统10包括用于传达信息的总线12或其它通信机制，并且包括与总线12耦合用于处理信息的处理器22。处理器22可以是任何类型的通用或专用处理器。系统10还包括用于

存储将由处理器22执行的信息和指令的存储器14。存储器14可以包括随机存取存储器(“RAM”)、只读存储器(“ROM”)、静态存储器(诸如,磁盘或光盘)或任何其它类型的计算机可读介质的任何组合。系统10还包括提供对网络的访问的通信设备20,诸如网络接口卡。因此,用户可以通过网络或任何其它方法直接地或远程地与系统10接口连接。

[0022] 计算机可读介质可以是任何可用的介质,该介质可以被处理器22访问并且包括易失性介质和非易失性介质二者、可移除介质和非可移除介质二者以及通信介质。通信介质可以包括计算机可读指令、数据结构、程序模块或调制数据信号(诸如,载波或其它传输机制)中的其它数据以及包括任何信息传递介质。

[0023] 处理器22进一步经由总线12耦合到显示器24,诸如液晶显示器(“LCD”)。键盘26和光标控制设备28(诸如,计算机鼠标)进一步与总线12耦合,以使得用户能够与系统10接口连接。

[0024] 在一种实施例中,存储器14存储当由处理器22执行时提供功能的软件模块。模块包括为系统10提供操作系统功能的操作系统15。模块还包括用于提供语义文本搜索和本文中公开的所有其它功能的语义文本搜索模块16。系统10可以是较大系统的部分。因此,系统10可以包括一个或多个附加的功能模块18以包括附加的功能。数据库17与总线12耦合以提供用于模块16和18的集中式存储并且存储文本语料库、树等。

[0025] 在另一种实施例中,存在从互联网或内联网或它们的任何组合中定位和下载电子文档的第一服务器或第一多个服务器。这些文档然后被存储在数据库(例如,结构化查询语言(“SQL”)或不仅仅是SQL(“NoSQL”)或它们的任何组合)中。第二服务器或第二多个服务器具有语义文本搜索软件,当该语义文本搜索软件被第二服务器处理器执行时,使用数据库中存储的文档执行图2中示出的功能。在一种实施例中,在210处,经由在个人计算机(“PC”)、移动电话或其它移动设备上显示的图形用户界面(“GUI”)接收搜索查询。

[0026] 图2是根据本发明的一种实施例的图1的语义文本搜索模块16和其它元件的功能的高级流程图。

[0027] 在一种实施例中,(一个或多个)电子文档(或本文中的“(一个或多个)文档”)是以需要计算机或其它电子设备来显示、解释和处理它的方式记录的任何信息。这包括由软件生成的并且存储在易失性存储器和/或非易失性存储器上的文档。一些示例包括文章、电子邮件、网页、推特、未结构化文本记录、或它们的任何组合。电子文档包括一些电子可解析文本。

[0028] 文本语料库被理解为一个或多个电子文档的组。文本语料库的示例包括整个互联网、电子图书馆、或文档储存库。

[0029] 在202处,接收文本语料库。文本语料库可以存储在图1的数据库17或者任何远程或本地的易失性存储器或非易失性存储器上。

[0030] 在204处,文本语料库被分离成语句。

[0031] 在206处,每一个语句(或语句片段)被解析并且被转换成树(即,“语句树”)。语句解析可以是语法解析或语句图解(sentence diagram)。为了获得这种解析,不同的实施例可以使用各种可用的由计算机实现的自然语言解析器,包括但不限于“The Stanford Parser:A statistical parser”、“ClearNLP”和其它解析器。每一个树由对应于语句中的每一个项的通过边连接的节点形成。边提供所连接节点的语法关系。例如,边可以指示一个

项是边所连接的另一个项的修饰语。

[0032] 图3图示根据一种实施例的通过解析语句“The car did not deploy front airbags during the accident”形成的示例树。节点包括语句中的词语:“deploy(部署)”、“car(汽车)”、“did”、“not”、“airbags”、“during”、“the”、“front”、“accident(事故)”和“the”。边包括节点的语法关系。例如,“car”是“deploy”的名词主语(“nsubj”边),而“airbag”是“deploy”的直接宾语(“dobj”边)。

[0033] 在另一种实施例中,解析树可以包括包含语句中的词语的节点,并且每一个节点可以包括类型(即,句法功能(例如,主语、动词、宾语)),并且边可以包括节点(例如,解析树的结构中的全部或部分)之间的依赖关系。例如,汽车[主语]依赖于部署,汽车的类型[主语]也是这样,汽车是部署[R00T类型]的主语。边可以可选地包括类型(例如,直接宾语或间接宾语)。例如,语句“Joe passed the salt shaker”和“Joe passed the salt shaker to his father”具有“salt shaker”作为这两个语句中的直接宾语并且具有“his father”作为第二个语句中的间接宾语。第一语句的第一解析树可以具有类型为直接宾语的边,而第二语句的第二解析树还可以包括类型为间接宾语的边。

[0034] 在208处,可选地修改来自206的一个或多个树。例如,树可以被拆分、整理、用附加的边扩充、折叠边类型等。冠词节点可以被丢弃。不同的实施方式可以用不同方式来定义“项”:一种实施例可以将每一个词语解释为分离的项。其它实施例可以执行字典查找或使用统计或自然语言处理技术将多词语序列(诸如,“United States”或“chief of staff”)识别为单个项。

[0035] 在另一种实施例中,可以展开一个或多个节点以包括同义词、上位词、下位词和/或其它相关的词语或短语。这些相关的词语或短语可以在节点内部被组织为平面列表或更复杂的结构(诸如,树)。另外,边也可以被折叠或展开。例如,在语句“Joe passed the salt shaker to his father”中,类型为间接宾语“his father”的边和类型为直接宾语“salt shaker”的边二者都可以被转换成一般的“宾语”类型边。不同的实施例可以例如通过保持原始的边类型(直接宾语、间接宾语)并且在它的上面添加更广义类型的一般“宾语”类型来展开边类型。

[0036] 在210处,接收搜索查询。该查询可以由单个项、数个项组成,或者可以是完整语句的形式。

[0037] 在212处,查询被可选地解释为树和/或被转换成树(即,“查询树”)。在一种实施例中,查询被解释为树,即使该查询只包含单个项(即,一个节点树)。在另一种实施例中,向用户生成/推荐细分以添加更多相关的项,并且细分导致树的创建。查询可以被转换成树(例如,通过解析自动地转换成树),或者通过只允许用户构造连通树(connected tree)的查询的机制来转换查询,该连通树通过所推荐的细分构建。

[0038] 在214处,响应于查询而生成的树(或者如果查询没有被转换成树,则仅仅查询本身)与在206处从文本语料库生成的一个或多个语句树进行匹配。在一种实施例中,匹配确定查询树是否是任何语句树中的子树。可以严格地(即,如果按照完全匹配节点集合以完全相同的方式连接完全匹配节点集合,则认为是匹配的)或近似地(即,节点可以相同而边可以不同,或者相同的边集合可只连接查询节点的子集,等等)执行树匹配。

[0039] 在216处,响应于匹配树,生成对应的匹配文档的响应。具体地,将包括与匹配树对

应的语句的文档选择为匹配文档。

[0040] 在218处,响应于匹配树,基于匹配生成相关查询的细分以将查询树建造成更大的树。作为这样做的结果,用户可以基于实际上作用于彼此的实体来细分用户的搜索。例如,在用户搜索“car”之后,可以提供细分搜索,诸如“drove car”、“crashed car”和“car accident”。“car accident”查询将仅返回事故是汽车事故的文档。因此,将不返回包括“we saw a train accident from our car”的文档。不同的实施例可以推荐展宽的细分(其中推荐的查询是当前查询的子树)或横向的细分(其中项或边被另一个项或边替换)。

[0041] 细分处理可以被迭代如所期望的一样多步。查询“drove car”将返回未来的细分,诸如“John drove car”。这个横向查询不同于常规的文本搜索,因为语义搜索将不会匹配包括“Peter drove the car,while John stayed home”的文档,并且该横向查询不同于短语搜索,因为对于“John drove car”的短语搜索将不会匹配文本“John,my neighbor,drove my car”,而根据实施例的语义搜索将正确匹配它,因为它理解在语句结构中编码的含义。

[0042] 另外,可以由边类型分组细分。例如,如果当前查询是“drove car”,则可以按照主语(“John drove car”、“Peter drove car”);形容词(“drove old car”、“drove new car”);副词(“drove car carelessly”、“drove car carefully”)等分组细分。

[0043] 在220处,响应于匹配树,在一些实施例中执行情感提取。例如,如果搜索项是被称为“Ace”的公司,则检索与搜索项在语法上联系的所有情感,诸如“terrible”、“great”等的修饰语。仅仅包含不修饰目标搜索项的修饰语的语句将不被考虑在内(即,“Used Acme product on my terrible ride home”将不返回项“Acme”的负面情感)。

[0044] 在另一种实施例中,除了吸收的文本语料库之外,其它源也用于丰富在206处生成的树。其它源可以是提供类树(tree-like)结构的外部分类系统(诸如,提供词义的来自普林斯顿大学的“**WordNet®**”)或根据各种分类系统(诸如,地理位置、相关项或类别的分类系统)组织概念的Wikipedia。因此,可以响应于查询项生成上位词、下位词、同义词等。通常,“下位词”是其语义域被包括在另一个词语(“下位词”的“上位词”)的语义域内的词语或短语。换言之,下位词与其上位词共享类型关系(type-of relationship)。例如,鸽子、乌鸦、鹰和海鸥全都是鸟(它们的上位词)的下位词,鸟进而是动物的下位词。

[0045] 例如,如果已知每一个“car”是“vehicle(车辆)”并且每一个“crash(撞毁)”是“accident(事故)”,则可以对“vehicle accident”执行搜索并且可以定位每一个“car crash”实例。反之将不会如此,因为存在除了汽车外的车辆和除了撞毁之外的事件。其它分类系统也可以适用。例如,地理分类系统将允许搜索“crime in Massachusetts”以检索对“jaywalking in Boston”的参考。

[0046] 另外,在一种实施例中,使用指代消解(anaphora resolution)来生成语义搜索结果。例如,如果文本语料库包括文本“John drove the car.He crashed it.”,则实施例可以推断出谁撞毁了什么,并且针对查询返回第二语句“John crashed”和“crashed car”。

[0047] 图4图示根据本发明的实施例的来自示出已解析语句和上位词匹配的GUI的屏幕截图。图4a图示屏幕截图的左侧,图4b图示屏幕截图的右侧,并且图4c图示放大的未结构化文本示例。语句被解析成具有用于动作者(actor) 401、动作(action) 402和对象(object) 403的节点的树。在410和411中示出与查询“actor=/vehicle/car AND object=“/

difficulty/problem”匹配的来自文本语料库的对应的语句和已解析的树,图示包括上位词层次的树的匹配。在这个实施例中,这个查询引起检索所有解析树,其中Actor (动作者)节点与具有上位词“vehicle”的项“car”匹配、并且Object (对象)节点与具有上位词“difficulty (困难)”的项“problem (问题)”匹配。一些实施例只使用从上位词/下位词树中选择的节点,而跳过其它节点。例如,Princeton’s WordNet v3.1包含以下的上位词树:

[0048] S: (n) **car**, auto, automobile, machine, motorcar (a motor vehicle with four wheels; usually propelled by an internal combustion engine) *"he needs a car to get to work"*

...

/ inherited hypernym /

S: (n) motor vehicle, automotive vehicle (a self-propelled wheeled vehicle that does not run on rails)

S: (n) self-propelled vehicle (a wheeled vehicle that carries in itself a means of propulsion)

S: (n) wheeled vehicle (a vehicle that moves on wheels and usually has a container for transporting things or people) *"the oldest known wheeled vehicles were found in Sumer and Syria and date from around 3500 BC"*

[0049] S: (n) **vehicle** (a conveyance that transports people or objects)

S: (n) **problem**, job (a state of difficulty that needs to be resolved) *"she and her husband are having problems"; "it is always a job to contact him"; "urban problems such as traffic congestion and smog"*

...

/ direct hypernym /

S: (n) **difficulty** (a condition or state of affairs almost beyond one's ability to deal with and requiring great effort to bear or overcome) *"grappling with financial difficulties"*

[0050] 某些实施例可以跳过“motor vehicle”、“self-propelled vehicle”和“wheeled vehicle”上位词,将项“car”直接连接到上位词“Vehicle”。

[0051] 图5a图示一种实施例的屏幕截图,在该实施例中用户搜索与查询“ACTOR=/vehicle”匹配的每一个文档,其中,前置反斜杠“/”指示用户对上位词感兴趣。对于此查询,每一个匹配的动作者(汽车、卡车、摩托车等)是车辆的具体类型。词语“vehicle”不需要出现在将匹配的文档的文本中。

[0052] 图5b图示一种实施例的导航的后续步骤,其中用户通过只返回也与“OBJECT=difficulty”查询匹配的文档来过滤“ACTOR=/vehicle”查询,从而进一步细分搜索。因此,

只返回涉及经历任何类型的困难(问题、麻烦等)的任何类型的车辆(汽车、卡车等)的那些文档。

[0053] 本发明的其它实施例不需要用户明确地指定用户感兴趣的是已解析文本的哪个域。在这种实施例中,对于“vehicle”的查询将返回“vehicle”的所有实例,无论它们是动作者、对象或可能是已解析树的其它节点。同样地,在一些实施例中不需要用户明确指定他们对上位词感兴趣。在这中实施例中,查询“vehicle”将返回以下二者:逐字(verbatim)文本提到“vehicle”的文档、以及实体(汽车、卡车等)具有项“vehicle”作为上位词的那些文档。

[0054] 一些实施例允许复杂的查询构造,其中查询不仅指定节点而且指定边。在一种搜索医疗记录的实施例中,用户可以指定检索与项“allergy”或“allergic”匹配的所有节点的查询,通过具体药品的名称细分所得的树,以及然后通过不存在负面边来进一步细分所得的树。这种查询将只检索对具体药品过敏的患者的记录。

[0055] 在一些实施例中,允许用户逐项地修改语义搜索查询。图6图示根据一种实施例的经由用户界面600的单项修改的屏幕截图。对于项“accident”,在601处示出特定的逐字形式,并且在图602处示出下位词/上位词选项。不可用的细分可以被变灰。

[0056] 在一种实施例中,层次的项修改可以用于若干上位词,诸如:

[0057] device>computer>apple

[0058] food>fruit>apple.

[0059] 这种层次还允许用户执行词义消歧。

[0060] 在一种实施例中,面包屑(breadcrumb)可以用于其它被丰富的维度(例如,地理位置、语义作用等),诸如:

[0061] U.S.>Massachusetts>Cambridge

[0062] U.K.>Cambridgeshire>Cambridge

[0063] Person>speaker>Larry Ellison

[0064] Person>investor>Larry Ellison.

[0065] 在一种实施例中,可以诸如下面这样通过将名词短语修饰语暴露为动态维度来修改查询:

[0066] ACCIDENT

[0067] major accident (20)

[0068] serious accident (15)

[0069] first accident (12)

[0070] ...

[0071] 如图6的603处示出的,这种修饰语还可以被包括在查询修改UI600中。

[0072] 标签云(tag cloud)通常在给定查询的情况下提供频繁项的集合。然而,在使用语义的实施例中,标签云可以在给定查询的情况下提供频繁子树的集合。

[0073] 实施例通过将搜索的焦点从令牌转移到文本的图形表示(即,项/实体作为节点以及节点间的连接作为边)来缩小语义和文本(换言之,含义及其表示)之间的差距。现有技术“词袋”(bag of words)模型并不区分诸如[theory group]和[group theory]的搜索,并且将多义项(诸如,“Java”)的所有词义映射到相同令牌。

[0074] 实施例使用被称为“phrases(短语)”的语言构造。在英语中,名词短语可以包括零

个或更多个形容词的序列,跟随着一个或多个名词;其它实施例可以将冠词和介词考虑在内。动词短语是零个或更多个副词,跟随着一个或多个动词。实施例还使用类型化的(typed)实体提取,其中经训练的提取器为了实体(诸如,People、Places、或Organizations)标记语料库文本。

[0075] 经由信息线索丰富界面(information-scent-rich interface)向用户暴露这种实体(相关项、列表、标签云、细分)允许用户不仅仅选择感兴趣的项,而且选择用于传达具体含义的项。

[0076] 含义不仅被编码在实体级上,而且被编码在语句级上。语句结构表达实体作用于彼此并且修饰彼此的方式。这种信息的提取是保持文本的语义所必要的。

[0077] 一种实施例使用动作者/动作/对象三元组的提取(其中,动作是动词短语,并且动作者和对象是名词短语)。语句“The car did not deploy front airbags during the accident”可以因此被解析成一个这样的三元组{car/did not deploy/front airbags}。这种解析是有损失的:它必要地简化语句结构,在这个示例中,包括损失了修饰语“during the accident”。然而,它确实从语句中提取了词袋模型不可访问的丰富的信息。

[0078] 其它实施例没有将所提取的三元组限于上述的格式。“RDF模式”{subject/predicate/object}是上述的超集。这种模式支持提取不同类型的实体和谓语。

[0079] 一种实施例将语句或语句片段解析成图形,项(主语、宾语、动作、修饰语等)被映射到图形中的节点,而谓语形成(有向)边。在一种实施例中,根据语句的语法结构来映射边,包括描述项(诸如,“分句”、“直接”、“介词”等)之间关系的边。在de Marneffe et al., “Stanford typed dependencies manual”(2008年9月,2013年12月Stanford Parser v.3.3修订)中公开了这种边的一种分类系统。

[0080] 实施例提供了引导(guided)(或分面(faceted))导航,该导航包括诸如图7中示出的提供细分链接的用户界面元素(有可能具有在选择时将会获得的记录的计数,或者记录计数的其它指示(诸如,直方图))。

[0081] 在一种实施例中,引导导航界面具有双重目的。一方面,细分提供了将结果集合缩小的场所。另一方面,细分提供了结果集合的总结,该总结向引导导航用户体验提供了语义搜索方面。图7a和图7b是图示根据一种实施例的结果集合的细分和总结的示例界面。

[0082] 具体地,选择具体项可以引起实施例通过查询解析树的数据库并且将细分查询的集合限于包含原始项和通过直接边与原始项连接的一个附加项的查询来显示可能的细分(更详细的)查询。这种细分既向用户展示了当前结果集合的语义内容并且又推荐进一步导航的可能方向。可以从该列表选择进一步细分,或者用户可以经由搜索框来搜索感兴趣的项。图7a示出一个这样的实施例,在该实施例中,对于原始搜索项“color”,连同指示每一个细分的频率的直方图来推荐细分,诸如“find color”、“favorite color”和“color use”。

[0083] 实施例允许在查询处理的最开始时利用相同交互模型,这允许用户起始于经由从已解析语义树的数据库中选择项或项集合来细分整个信息集合,该已解析语义树以与“tag clouds(标记云)”或“word clouds(字词云)”相同的方式显示。

[0084] 如所描述的,实施例使用上下文将令牌分组组成实体和语句结构,以检测实体作用于其它实体的方式。另外,实施例使用语料库外部的信息(诸如,用更一般的层次结构亲值

(parent value) 来扩充已解析语义树的每一个域)。在一些实施例中,可以使用“**WordNet®**”为节点(诸如,名词或动词)定位上位词(扩展同义词)。对于其它实施例,可以使用前置项(在英语中,几乎总是短语的最右项)(诸如,使用“front airbags”中的“airbags”、或“Hillary Clinton”和“Bill Clinton”中的“Clinton”)。

[0085] 图7a图示处理,在该处理中用户以搜索表达搜索意图(“color”)的任何项开始,然后进行选择细分(“favorite color”)来缩小结果集合。实施例然后提供图7b中的其它可能细分。

[0086] 如所公开的,实施例通过将语句解析成树并且提供用于将这些树与用户查询匹配的机制来提供语义文本搜索。匹配树可以生成对匹配对应文档、细分的响应,或者提供语义承载项的集合来推动情感提取。

[0087] 本文中特别地图示和/或描述了数个实施例。然而应该理解的是,在不脱离本发明的精神和所期望的范围的情况下,所公开的实施例的修改形式和变形形式被以上教导覆盖并且在所附权利要求书的范围内。

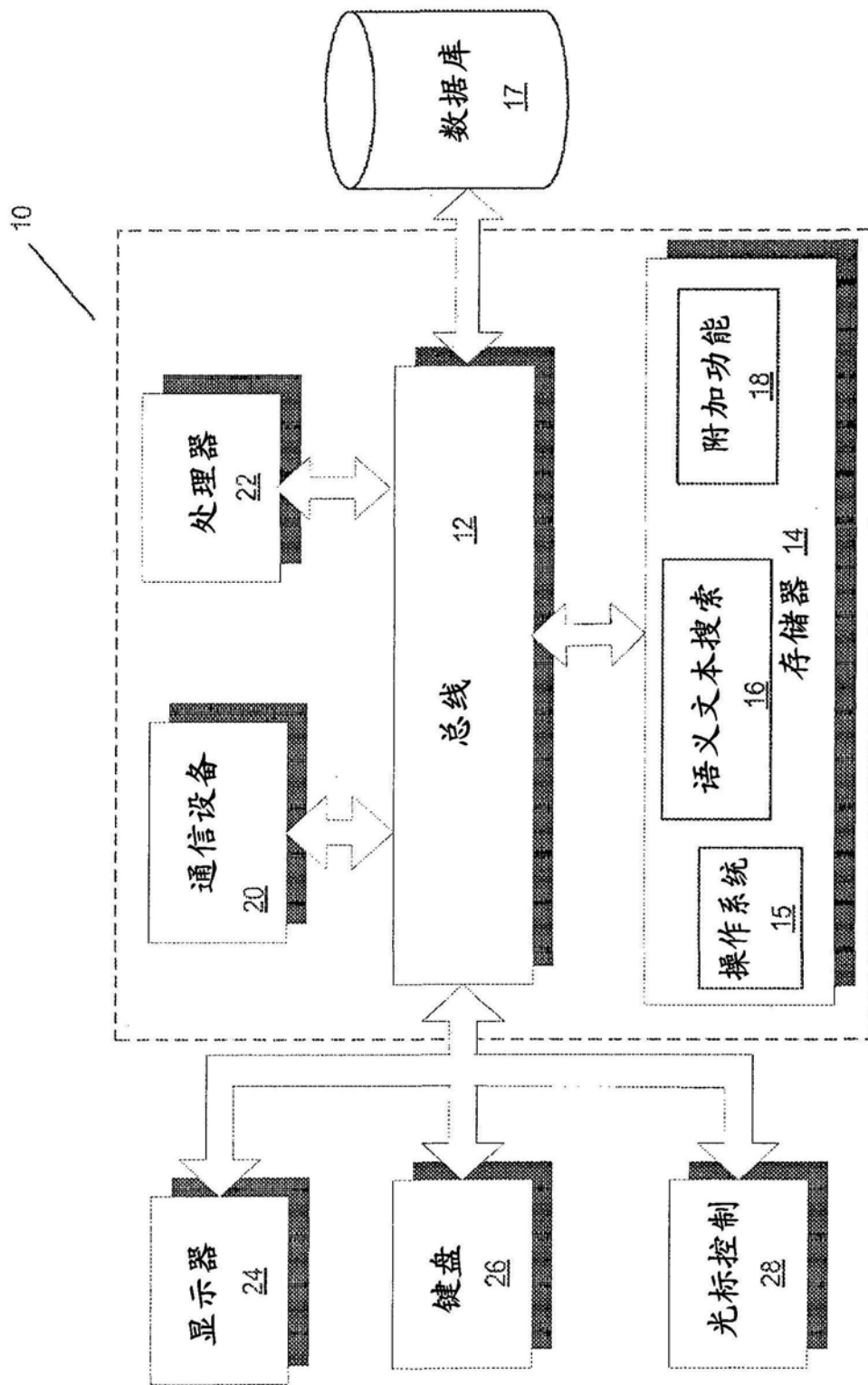


图1

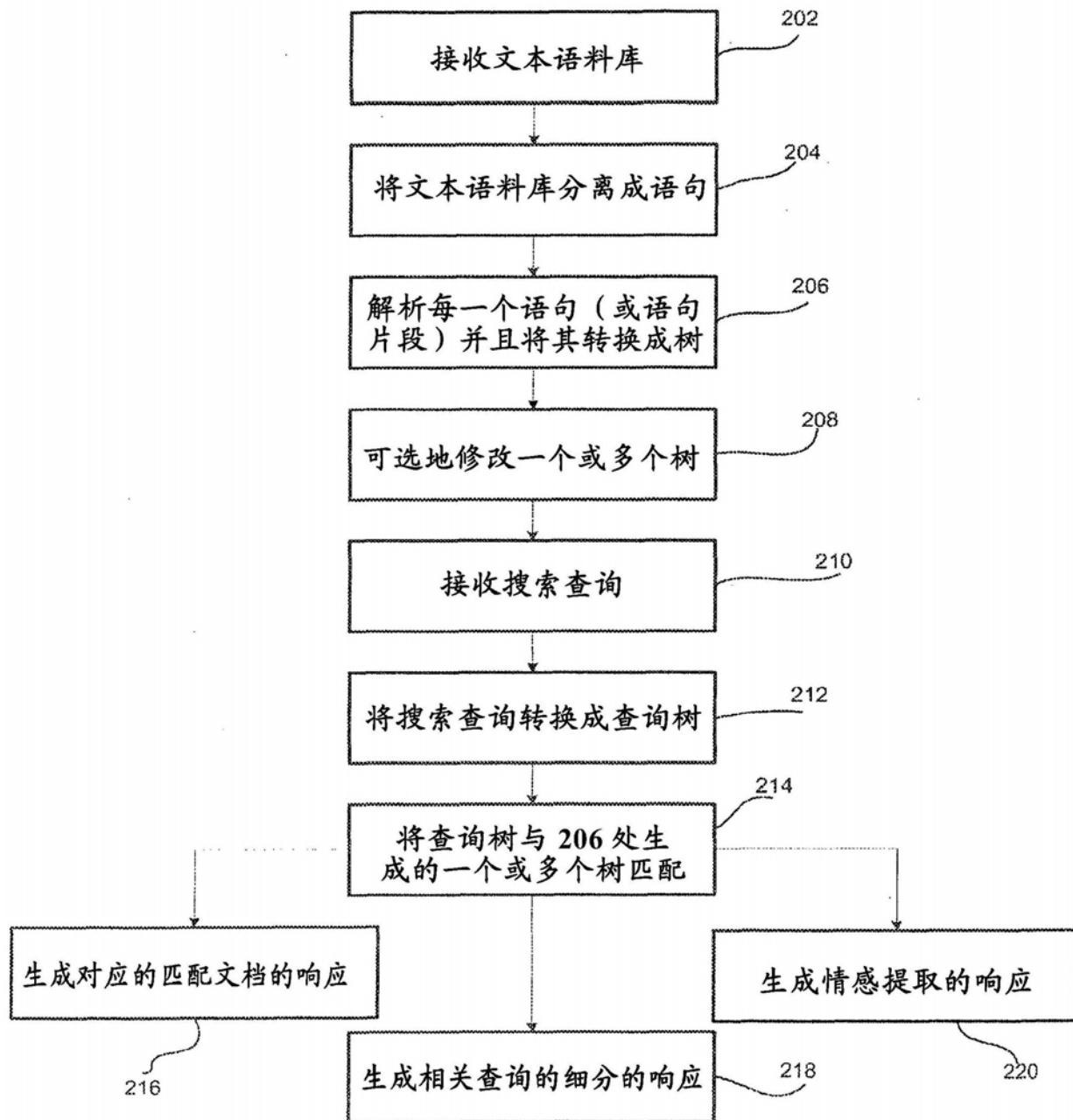


图2

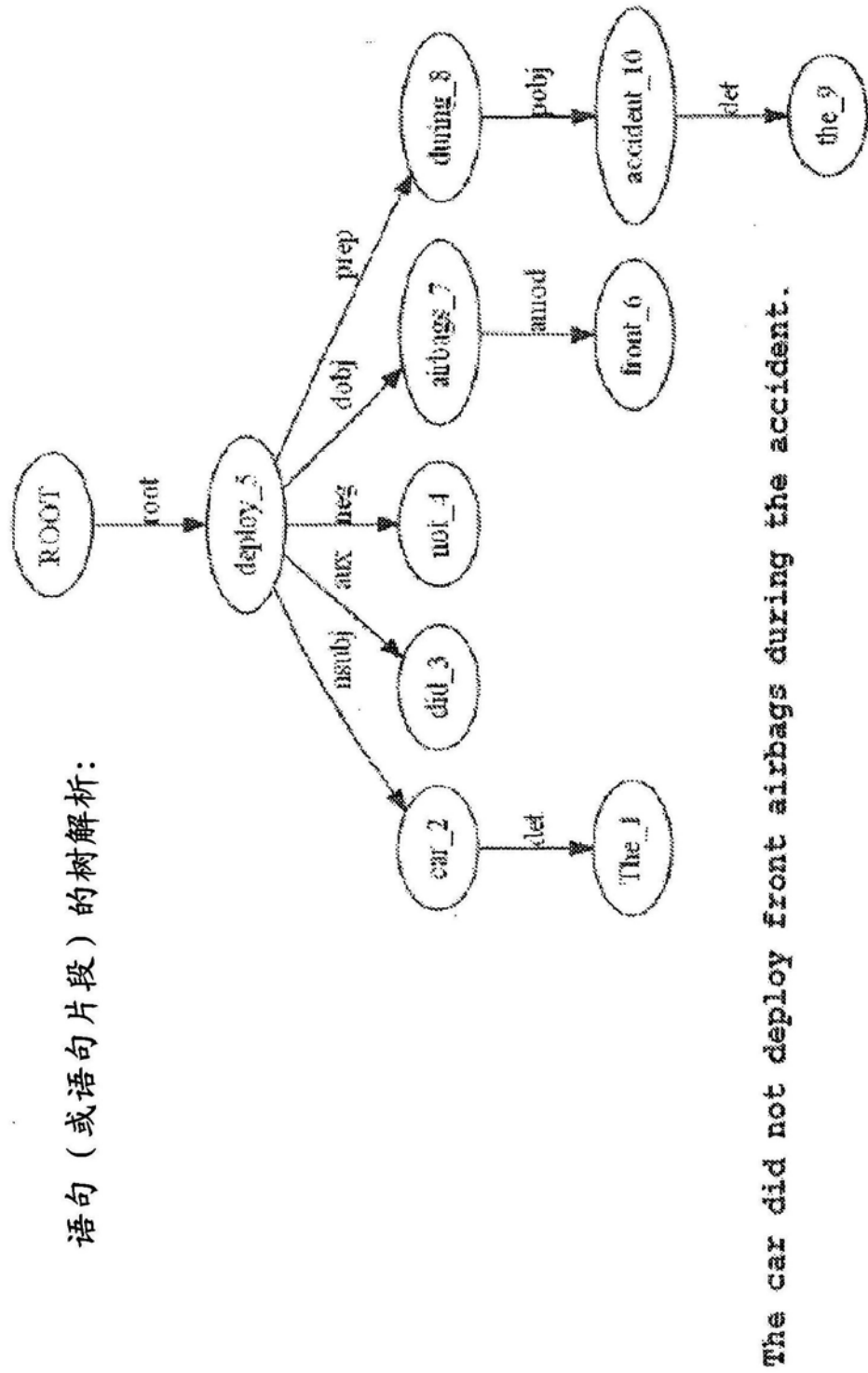


图3

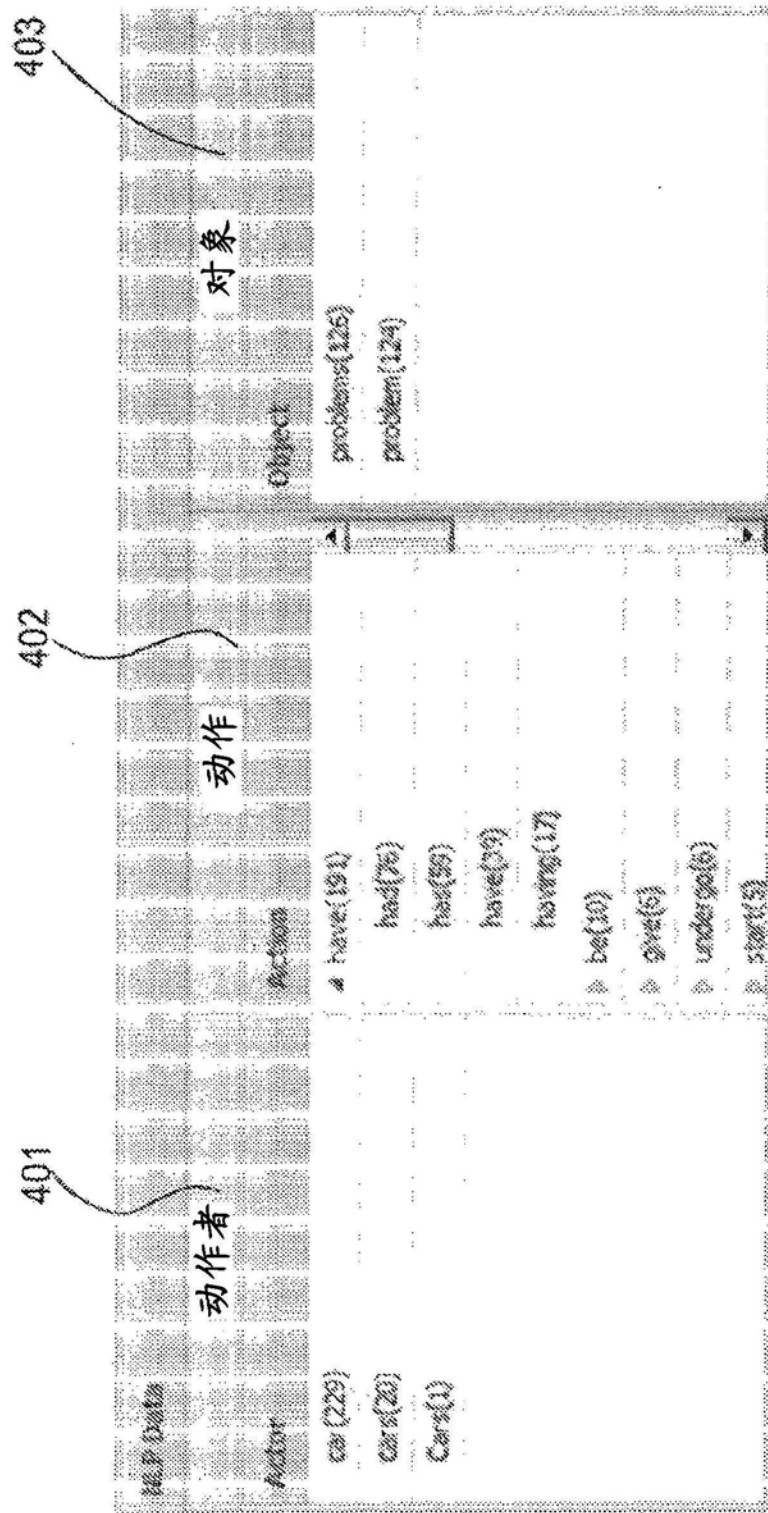


图4A

Breakdown & Search	
Queried Dimension	Value
Acas	/prefix/car
Object	/id/Model/Problem
Assertion	Sentence
(car/SUBJ) has/ROOT ((surging/MOD) problem/OBJ) (during/PREP ((normal/MOD) c... (usually/MOD) by/PREP ((3/MOD time/MOD engine/MOD starts/OBJ) car/SUBJ)P... (car/SUBJ) has/ROOT ((surging/MOD) problem/OBJ) (during/PREP ((cruise/MOD) c... (car/SUBJ) has/ROOT (problem/OBJ) (with/PREP ((cooling/MOD) system/OBJ) is re... (car/SUBJ) had/ROOT (problem/OBJ) starting/MOD) (My/MOD) car/SUBJ also/MOD had/ROOT ((downshifting/MOD) problem/OBJ) (w... (car/SUBJ) had/ROOT (problem/OBJ) (for/PREP ((two/MOD) days/OBJ)) (car/SUBJ) having/ROOT ((sensor/MOD) problem/OBJ) (car/SUBJ) has/ROOT ((surging/MOD) problem/OBJ) (during/PREP ((normal/MOD) c... (Took/MOD) car/SUBJ (dealer/MOD) (toyota/MOD) rep/ROOT) not/NEG find/ROOT ... (car/SUBJ) experiencing/ROOT ((same/MOD) problem/OBJ) (with/PREP ((normal/... (My/MOD) car/SUBJ) had/ROOT (problem/OBJ) starting off/MOD I bought it/MOD) (car/SUBJ) had/ROOT ((same/MOD) problem/OBJ) (My/MOD) car/SUBJ also/MOD had/ROOT ((downshifting/MOD) problem/OBJ) (for... (car/SUBJ) mimic/ROOT (problem/OBJ) at/PREP (time/OBJ) (car/SUBJ) has/ROOT ((surging/MOD) problem/OBJ) (during/PREP ((cruise/MOD) c... (car/SUBJ) still/MOD has/ROOT (problem/OBJ) (car/SUBJ) told/ROOT (them/OBJ) problem/OBJ) (In/PREP (December/OBJ) 14,2006/MOD car/SUBJ) developed/ROOT (problem/OBJ) ... (car/SUBJ) has/ROOT ((mechanical/MOD) problem/OBJ) (new/MOD) car/SUBJ have/ROOT (problem/OBJ) (car/SUBJ) (in/PREP (fact/OBJ)) not/NEG have/ROOT (problem/OBJ) (car/SUBJ) has/ROOT ((jerky/MOD steering/MOD) problem/OBJ) (sever/MOD (tra... (My/MOD) car/SUBJ) experiencing/ROOT ((same/MOD) problem/OBJ) (as/PREP ((... (other/MOD) car/SUBJ) (I owned with abs over past 15 years/MOD) had/ROOT (...	This car has a surging problem during normal or cruise control, occurring between 40 and 50 mph. Usually by the 3 time engine starts and car gets driven the problem of stalling goes away. This car has a surging problem during normal or cruise control, occurring between 40 and 50 mph. The car has a problem with the cooling system, if it really loud an shakes the car really bad. The car has had problems starting. My car also had downshifting problems where the downshift decent control would come on out of now The car has had problems for two days. The car was having sensor problems. This car has a surging problem during normal or cruise control, occurring between 40 and 50 mph. Took car to dealer toyota rep and servicemen could not find problem. There is more, on 6/1/7/03 the car is now at 43,000 miles and experiencing the same problems with My car has had a problem starting off and on the day I bought it and it seems to be a different era. When it was finally done, the car had the same problem. My car also had downshifting problems where the downshift decent control would come on out of now Car not make the car mimic the problems at any time but dealership will not take responsibility because This car has a surging problem during normal or cruise control, occurring between 40 and 50 mph. This car still has the problem. I then decided to take my car to the dealership being that the car was still under warranty and told In December 14,2006 this car developed a problem with the power steering of car. Please advise me how to handle this situation, I feel that the car has a mechanical problem that the When the car stalls it usually starts again immediately however I don't think a new car should have it Saturn, following complaints both to the service department and their corp headquarters has decide The car has jerky steering problems coming out of turns and sever transmission problems. The contact stated that his car was experiencing the same problems as campaign number 99v197. No other car that I have owned with abs over the past 15 years has had this problem.

图4B

[illegible]

图4C

[illegible]

图5A

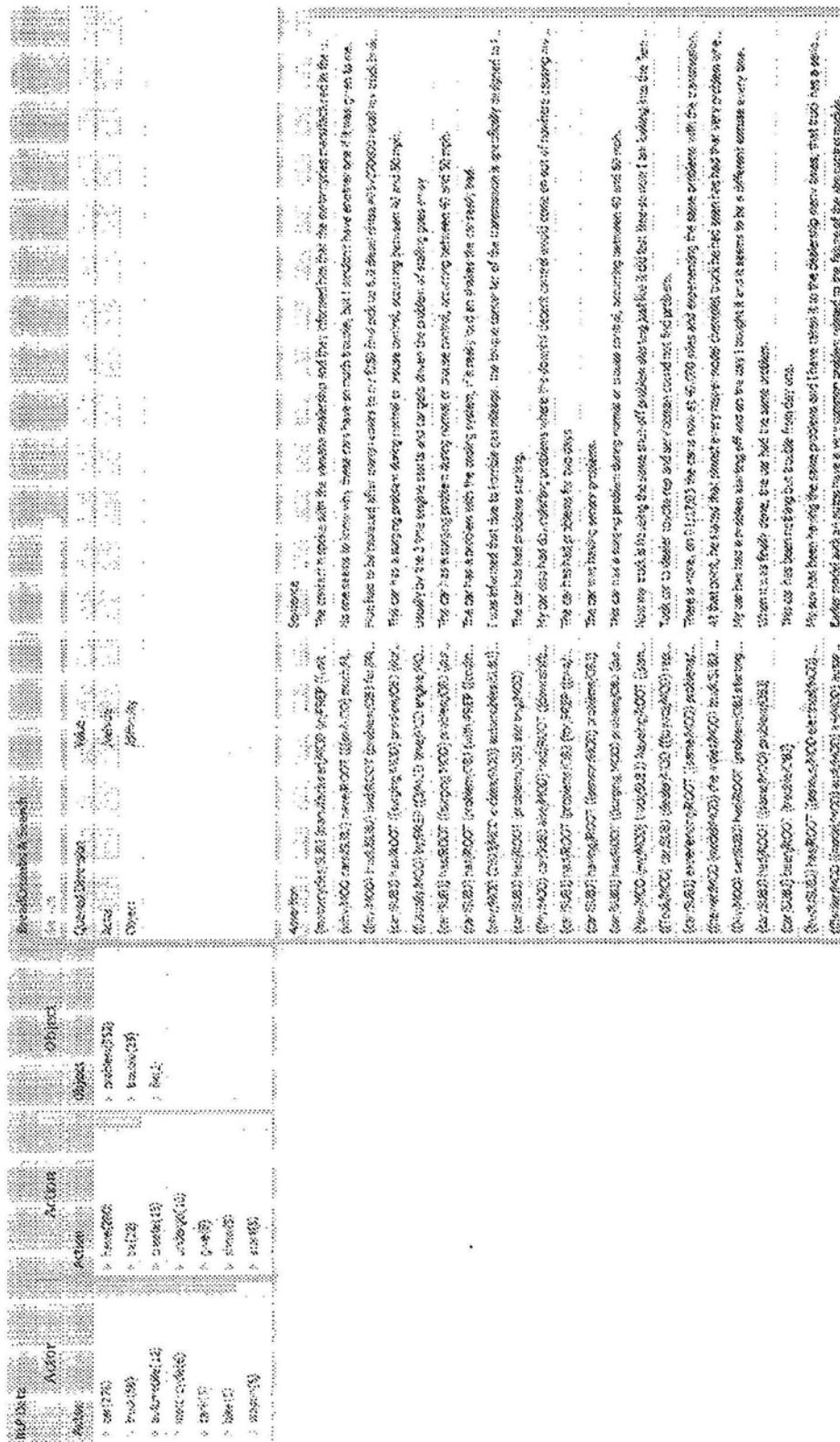


图5B

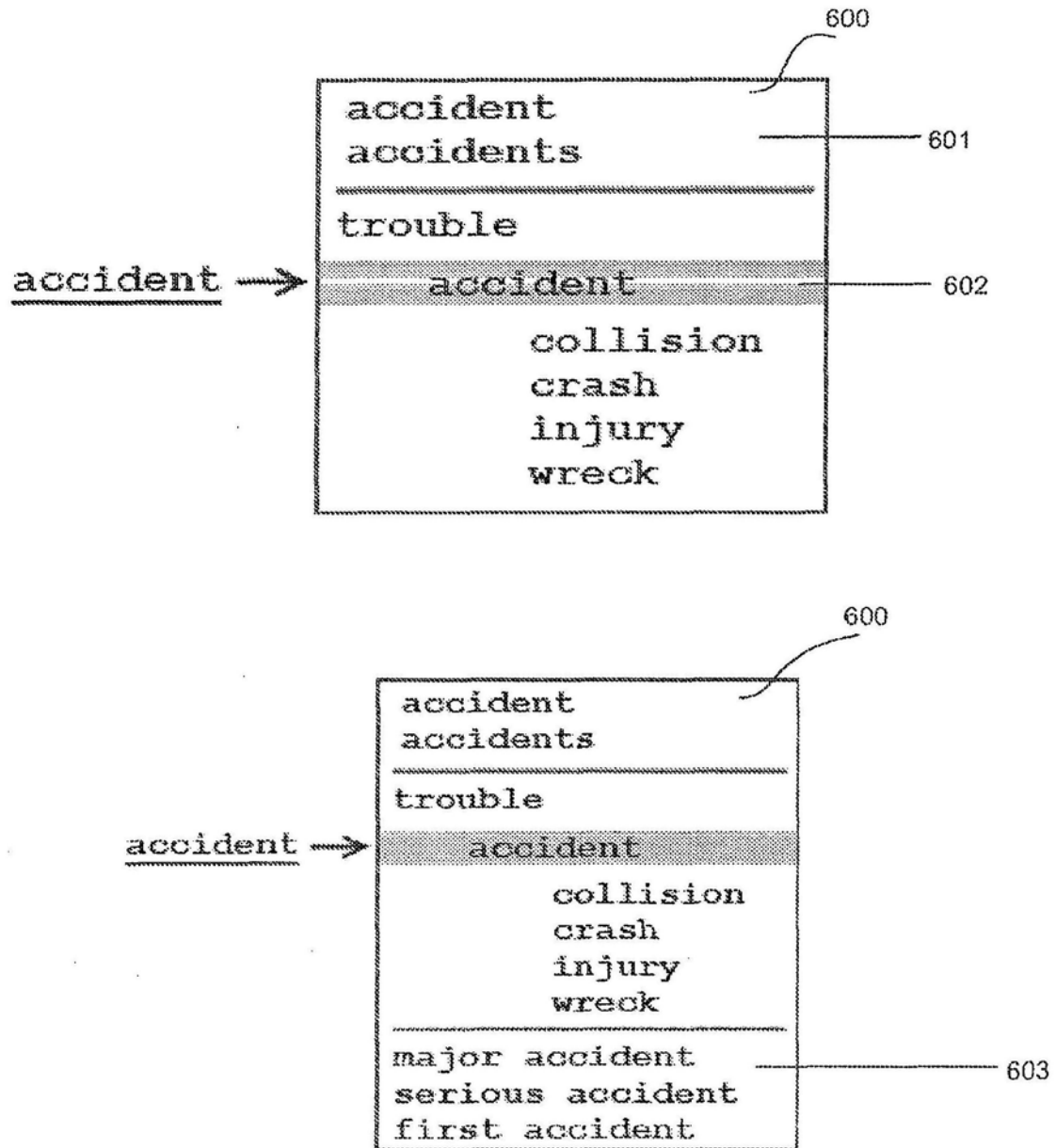


图6

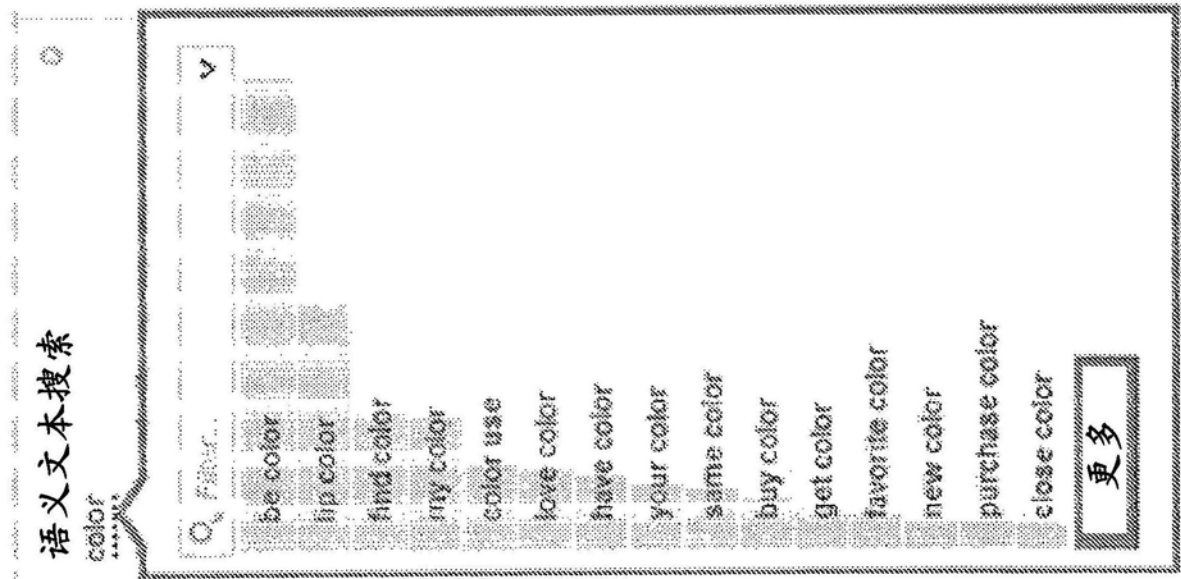


图7A

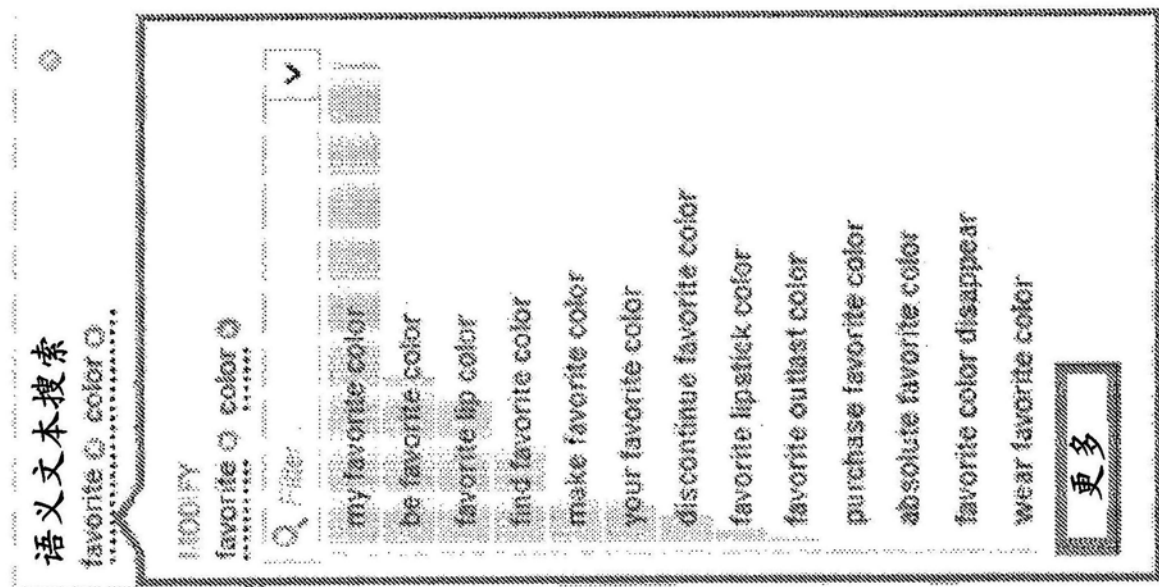


图7B