(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 103455620 B
(45)授权公告日 2017.05.03

(21)申请号 201310414978.8

(22)申请日 2013.09.12

(65)同一申请的已公布的文献号
申请公布号 CN 103455620 A

(43)申请公布日 2013.12.18

(73)专利权人 百度在线网络技术(北京)有限公司
地址 100085 北京市海淀区上地十街10号百度大厦三层

(72)发明人 王维维

(74)专利代理机构 北京铭硕知识产权代理有限公司 11286
代理人 张川绪 薛义丹

(51)Int.Cl.
G06F 17/30(2006.01)

(56)对比文件
US 2012/0239761 A1,2012.09.20,
CN 102360383 A,2012.02.22,
王兴义.基于模式匹配的中文专有名词识别.《中国优秀硕士学位论文全文数据库 信息科技辑》.2005,(第7期),

审查员 侯德芹

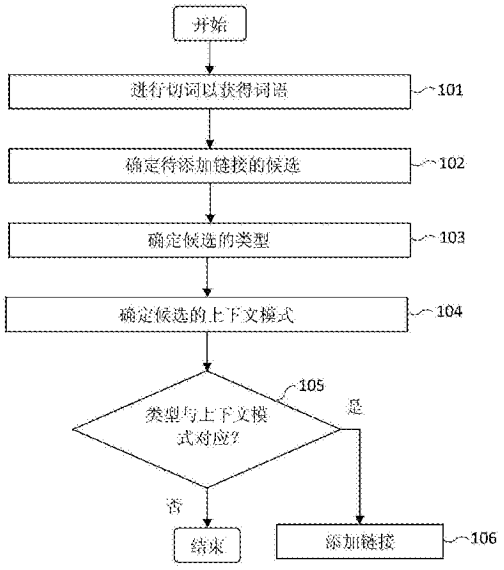
权利要求书2页 说明书6页 附图3页

(54)发明名称

在内容中添加链接的方法和设备

(57)摘要

本发明提供一种在内容中添加链接的方法和设备。所述方法包括:对内容进行切词,以获得词语;从获得的词语确定待添加链接的候选;确定候选的类别;确定候选的上下文模式;当候选的类别与候选的上下文模式对应时,向候选添加链接。根据本发明的在内容中添加链接的方法和设备,可以实现自动在内容中添加链接,从而可以避免了人工进行链接的添加,提高了效率。



CN 103455620 B

1.一种在内容中添加链接的方法,包括:

对内容进行切词,以获得词语;

从获得的词语之中确定待添加链接的候选;

确定候选的类别;

确定候选的上下文模式,其中,上下文模式是指候选与候选在内容中的上下文之间的语法结构和/或语意结构;

当候选的类别与候选的上下文模式对应时,向候选添加链接,

其中,从获得的词语确定待添加链接的候选的步骤包括:

根据在内容中的原始次序对获得的词语进行组合,以得到至少一个第一组合,并且将获得的各个词语分别作为一个第一组合;

从第一组合之中选择存在于预定数据库中的第一组合作为候选,

其中,当候选的类别与候选的上下文模式对应时,向候选添加链接的步骤包括:

当候选的类别与候选的上下文模式对应时,确定候选在内容中的上下文与候选在其他内容中的上下文的相似度;

当确定的相似度大于预定阈值时,向候选添加链接。

2.根据权利要求1所述的方法,其中,从得到的第一组合之中选择存在于预定数据库中的第一组合作为候选的步骤还包括:

当在选择的第一组合之中存在位置上重叠的第一组合时,从重叠的第一组合之中选择最长的第一组合作为候选。

3.根据权利要求1所述的方法,其中,向候选添加链接的步骤还包括:

当确定的相似度不大于预定阈值时,不向候选添加链接。

4.根据权利要求1所述的方法,其中,确定候选在内容中的上下文与候选在其他内容中的上下文的相似度的步骤包括:

获取候选在内容中的上下文与候选在其他内容中的上下文的重复词语;

确定每个重复词语的语意表达能力;

确定的语意表达能力之中最高的语意表达能力作为所述相似度。

5.根据权利要求1所述的方法,其中,所述其他内容具有所述候选并且所述候选已经在所述其他内容中被添加了链接。

6.一种在内容中添加链接的设备,包括:

切词单元,对内容进行切词,以获得词语;

候选确定单元,从获得的词语之中确定待添加链接的候选;

类别分析单元,确定候选的类别;

上下文模式确定单元,确定候选的上下文模式,其中,上下文模式是指候选与候选在内容中的上下文之间的语法结构和/或语意结构;

链接添加单元,当确定的类别与确定的上下文模式对应时,向候选添加链接,

其中,候选确定单元包括:

组合单元,根据在内容中的原始次序对获得的词语进行组合,以得到至少一个第一组合,并且将切词单元获得的各个词语分别作为一个第一组合;

选择单元,从第一组合之中选择存在于预定数据库中的第一组合作为候选,

其中，还包括：相似度确定单元，当候选的类别与候选的上下文模式对应时，确定候选在内容中的上下文与候选在其他内容中的上下文的相似度，

其中，当确定的相似度大于预定阈值时，链接添加单元向候选添加链接。

7.根据权利要求6所述的设备，其中，当在选择的第一组合之中存在位置上重叠的第一组合时，选择单元从重叠的第一组合之中选择最长的第一组合作为候选。

8.根据权利要求6所述的设备，其中，当确定的相似度不大于预定阈值时，链接添加单元不向候选添加链接。

9.根据权利要求6所述的设备，其中，所述相似度表示候选在内容中的上下文与候选在其他内容中的上下文之间的重复词语的语意表达能力之中的最高语意表达能力。

10.根据权利要求6所述的设备，其中，所述其他内容具有所述候选并且所述候选已经在所述其他内容中被添加了链接。

在内容中添加链接的方法和设备

技术领域

[0001] 本发明涉及计算机网络领域。更具体地讲，涉及一种在内容中添加链接的方法和设备。

背景技术

[0002] 随着信息技术的发展，越来越多的内容出现在网络中，供网络用户进行浏览。然而，用户在浏览内容时，可能遇到一些不熟悉的信息，或者希望对一些信息做进一步了解。这时，通常需要将这些信息复制或输入到搜索引擎，然后通过搜索引擎来检索这些信息。

[0003] 解决上述问题的一个方案是在内容中对这些信息添加链接(例如，超级链接)，从而用户可以通过这些链接来访问相应的信息，而不需要另外进行搜索。

[0004] 然而，在现有技术中，为了在内容中添加链接，通常需要人工来完成链接的添加，成本较高，并且效率较低。另外，人工添加链接受添加操作的执行人员的主观因素和背景知识的影响也较大，导致添加的链接的质量参差不齐，并且无法准确反映网络用户对链接的一般需要。因此，需要一种能够在内容中自动添加链接并且添加的链接能够反映网络用户对链接的需要的链接添加技术。

发明内容

[0005] 本发明的目的在于提供一种能够在内容中自动添加链接的技术，从而不要用人工进行链接的添加，并且能够反映网络用户对链接的需要。

[0006] 本发明的一方面提供一种在内容中添加链接的方法，所述方法包括：对内容进行切词，以获得词语；从获得的词语确定待添加链接的候选；确定候选的类别；确定候选的上下文模式；当候选的类别与候选的上下文模式对应时，向候选添加链接。

[0007] 可选地，上下文模式是指候选与候选在内容中的上下文之间的语法结构和/或语意结构。

[0008] 可选地，从获得的词语确定待添加链接的候选的步骤包括：根据在内容中的原始次序对获得的词语进行组合，以得到至少一个第一组合，并且将获得的各个词语分别作为一个第一组合；从得到的第一组合之中选择存在于预定数据库中的第一组合作为候选。

[0009] 可选地，从得到的第一组合之中选择存在于预定数据库中的第一组合作为候选的步骤还包括：当在选择的第一组合之中存在位置上重叠的第一组合时，从重叠的第一组合之中选择最长的第一组合作为候选。

[0010] 可选地，向候选添加链接的步骤包括：当候选的类别与候选的上下文模式对应时，确定候选在内容中的上下文与候选在其他内容中的上下文的相似度；当确定的相似度大于预定阈值时，向候选添加链接。

[0011] 可选地，向候选添加链接的步骤还包括：当确定的相似度不大于预定阈值时，不向候选添加链接。

[0012] 可选地，确定候选在内容中的上下文与候选在其他内容中的上下文的相似度的步

骤包括：获取候选在内容中的上下文与候选在其他内容中的上下文的重复词语；确定每个重复词语的语意表达能力；确定的语意表达能力之中最高的语意表达能力作为所述相似度。

[0013] 可选地，所述其他内容具有所述候选并且所述候选已经在所述其他内容中被添加了链接。

[0014] 本发明的另一方面提供一种在内容中添加链接的设备，包括：切词单元，对内容进行切词，以获得词语；候选确定单元，从获得的词语确定待添加链接的候选；类别分析单元，确定候选的类别；上下文模式确定单元，确定候选的上下文模式；链接添加单元，当确定的类别与确定的上下文模式对应时，向候选添加链接。

[0015] 可选地，上下文模式是指候选与候选在内容中的上下文之间的语法结构和/或语意结构。

[0016] 可选地，候选确定单元包括：组合单元，根据在内容中的原始次序对获得的词语进行组合，以得到至少一个第一组合，并且将获得的各个词语分别作为一个第一组合；选择单元，从得到的第一组合之中选择存在于预定数据库中的第一组合作为候选。

[0017] 可选地，当在选择的第一组合之中存在位置上重叠的第一组合时，选择单元从重叠的第一组合之中选择最长的第一组合作为候选。

[0018] 可选地，所述设备还包括：相似度确定单元，当候选的类别与候选的上下文模式对应时，确定候选在内容中的上下文与候选在其他内容中的上下文的相似度，其中，当确定的相似度大于预定阈值时，链接添加单元向候选添加链接。

[0019] 可选地，当确定的相似度不大于预定阈值时，链接添加单元不向候选添加链接。

[0020] 可选地，所述相似度表示候选在内容中的上下文与候选在其他内容中的上下文之间的重复词语的语意表达能力之中的最高语意表达能力。

[0021] 可选地，所述其他内容具有所述候选并且所述候选已经在所述其他内容中被添加了链接。

[0022] 根据本发明的在内容中添加链接的方法和设备，可以实现自动在内容中添加链接，从而可以避免了人工进行链接的添加，提高了效率。此外，根据本发明的在内容中添加链接的方法和设备在内容中所添加的链接能够反映网络用户对链接的需要。

[0023] 将在接下来的描述中部分阐述本发明另外的方面和/或优点，还有一部分通过描述将是清楚的，或者可以经过本发明的实施而得知。

附图说明

[0024] 通过下面结合附图进行的详细描述，本发明的上述和其它目的、特点和优点将会变得更加清楚，其中：

[0025] 图1是示出根据本发明的实施例的在内容中自动添加链接的方法的流程图；

[0026] 图2是示出根据本发明的另一实施例的在内容中自动添加链接的方法的流程图；

[0027] 图3是示出根据本发明的实施例的在内容中自动添加链接的设备的框图；

[0028] 图4是示出根据本发明的另一实施例的在内容中自动添加链接的设备的框图。

具体实施方式

[0029] 下面参照附图详细描述本发明的实施例。

[0030] 图1是示出根据本发明的实施例的在内容中自动添加链接的方法的流程图。

[0031] 如图1所示，在步骤101，对内容进行切词，以获得词语。可利用现有的切词技术对内容进行切词。此外，这里的内容可以是具有文字的各种内容。

[0032] 例如，通过对“东北财经大学会计学院任教”进行切词，可以得到下面的多个词语：东北、财经、大学、会计、学院、任教。

[0033] 在步骤102，从在步骤101获得的多个词语之中确定待添加链接(例如，超级链接)的候选。换言之，从获得的多个词语之中确定需要添加链接的词语。例如，可预先建立候选数据库，将存在于候选数据库之中的词语确定为候选。

[0034] 在一个优选实施例中，为了从获得的多个词语之中确定候选，首先根据获得的多个词语在内容中的原始次序对获得的多个词语进行组合，以得到至少一个第一组合。

[0035] 例如，仍以上面的“东北财经大学会计学院任教”为例，第一组合可以是东北财经、财经大学、财经大学会计、会计学院等，但不能是财经东北、东北大学等。

[0036] 此外，获得的各个词语也可分别作为一个第一组合。

[0037] 随后，从得到的第一组合之中选择存在于候选数据库中的第一组合作为候选。

[0038] 优选地，当选择出的多个第一组合在内容中的原始位置上存在重叠时，从重叠的第一组合之中选择最长的第一组合作为候选。换言之，如果最终得到多个候选，则最终得到的多个候选不会在内容中的原始位置上存在重叠。

[0039] 例如，仍以上面的“东北财经大学会计学院任教”为例，第一组合可以是东北财经大学、东北财经大学会计学院，这两个词在内容的原始位置“东北财经大学”处存在重叠，因此选择“东北财经大学会计学院”作为候选。

[0040] 在步骤103，确定每个候选的类别。例如，可预先对候选数据库中的不同词语进行分类，从而可以确定出不同候选的类别。

[0041] 例如，类别可以是自然、文化、地理、历史、经济、人物等。然而，本发明不限于于此，也可以以其他方式进行分类。

[0042] 在步骤104，确定每个候选的上下文模式。在本发明中，上下文模式是指候选与候选在内容中的上下文之间的语法结构和/或语意结构。例如，语法结构可以是动宾结构、并列结构、偏正结构、主谓结构等。语意结构是指候选与候选的上下文中的特定词语的组合，也即，存在于候选的上下文中的特定词语。例如，特定词语可以是：位于、坐落于、作词、作曲等。

[0043] 在步骤105，分别针对每个候选确定候选的类别与候选的上下文模式是否对应。可预先设置类别与上下文模式之间的对应关系。例如，可预先形成类别与上下文模式的映射表，通过该映射表来确定候选的类别与候选的上下文模式是否对应。

[0044] 当在步骤105确定某个候选的类别与该候选的上下文模式对应时，在步骤106向该候选添加链接。换言之，对于在步骤105中被确定类别与上下文模式对应的候选，在步骤106向该候选添加链接。

[0045] 可预先为不同的候选设置链接。例如，候选数据库中的词语可与对应的链接关联地存储在候选数据库中，从而可基于候选从候选数据库得到与该候选对应的链接。

[0046] 当在步骤105确定某个候选的类别与该候选的上下文模式不对应时，不向该候选

添加链接，并结束该方法。换言之，对于在步骤105中被确定类别与上下文模式不对应的候选，不向该候选添加链接。

[0047] 图2是示出根据本发明的另一实施例的在内容中自动添加链接的方法的流程图。

[0048] 如图2所示，步骤201-205分别与图1所示的步骤101-105相同，不再赘述。

[0049] 图2的方法与图1的方法相比，不同在于，当在步骤205确定某个候选的类别与该候选的上下文模式对应时，在步骤206确定候选在当前内容中的上下文与候选在其他内容中的上下文的相似度。

[0050] 这里，其他内容是指具有所述候选并且所述候选在其中已经被添加了链接的内容。可预先获取不同候选在其他内容中的上下文。例如，可预先收集候选数据库中的每个词语在其他内容中的上下文，并与其对应的词语关联地存储。

[0051] 由于一些词语(例如，虚词)基本没有语意，导致确定的相似度不准确。为此，在本发明的另一个实施例中，候选在当前内容中的上下文与候选在其他内容中的上下文的相似度可被定义为候选在当前内容中的上下文与候选在其他内容中的上下文之间的重复词语的语意表达能力之中的最高语意表达能力。

[0052] 这样，为了确定候选在内容中的上下文与候选在其他内容中的上下文的相似度，首先获取候选在内容中的上下文与候选在其他内容中的上下文的重复词语，确定每个重复词语的语意表达能力，然后确定语意表达能力之中最高的语意表达能力作为所述相似度。不同词语的语意表达能力可被预先确定，从而可以在使用时进行查询。此外，应该理解，候选在内容中的上下文是指在候选附近预定范围内的内容。

[0053] 随后，在步骤207，确定在步骤206中得到的相似度是否大于预定阈值。

[0054] 当在步骤207确定相似度大于预定阈值时，在步骤208向候选添加链接。

[0055] 当在步骤207确定相似度不大于预定阈值时，不向候选添加链接，所述方法结束。

[0056] 图3是示出根据本发明的实施例的在内容中自动添加链接的设备的框图。

[0057] 如图3所示，根据本发明的实施例的在内容中自动添加链接的设备300包括：切词单元310、候选确定单元320、类别分析单元330、上下文模式确定单元340、链接添加单元350。

[0058] 切词单元310对内容进行切词，以获得词语。可利用现有的切词技术对内容进行切词。

[0059] 候选确定单元320从获得的词语确定待添加链接的候选。换言之，候选确定单元320从获得的词语之中确定需要添加链接的词语。例如，可预先建立候选数据库，候选确定单元320将存在于候选数据库之中的词语确定为候选。

[0060] 在一个优选实施例中，候选确定单元320包括组合单元和选择单元。

[0061] 组合单元根据在内容中的原始次序对获得的多个词语进行组合，以得到至少一个第一组合，并且将切词单元获得的各个词语也分别作为一个第一组合。选择单元从得到的第一组合之中选择存在于预定数据库中的第一组合作为候选。

[0062] 优选地，当选择出的多个第一组合在内容中的原始位置上存在重叠时，选择单元从重叠的第一组合之中选择最长的第一组合作为候选。换言之，如果最终得到多个候选，则最终得到的多个候选不会在内容中的原始位置上存在重叠。

[0063] 类别分析单元330确定候选的类别。例如，可预先对候选数据库中的不同词语进行

分类，从而类别分析单元330可以根据分类结果确定出不同候选的类别。

[0064] 例如，类别可以是自然、文化、地理、历史、经济、人物等。然而，本发明不限于于此，也可以以其他方式进行分类。

[0065] 上下文模式确定单元340确定每个候选的上下文模式。前面已经描述了上下文模式的含义，不再赘述。

[0066] 链接添加单元350可分别针对每个候选确定候选的类别与候选的上下文模式是否对应。可预先设置类别与上下文模式之间的对应关系。例如，可预先形成类别与上下文模式的映射表，链接添加单元350通过该映射表来确定候选的类别与候选的上下文模式是否对应。

[0067] 链接添加单元350在确定的类别与确定的上下文模式对应时向候选添加链接。链接添加单元350在确定的类别与确定的上下文模式不对应时不向候选添加链接。

[0068] 图4是示出根据本发明的另一实施例的在内容中自动添加链接的设备的框图。

[0069] 如图4所示，根据本发明的实施例的在内容中自动添加链接的设备400包括：切词单元310、候选确定单元320、类别分析单元330、上下文模式确定单元340、相似度确定单元410、链接添加单元420。

[0070] 图4所示的切词单元310、候选确定单元320、类别分析单元330、上下文模式确定单元340已经在前面进行了描述，不再赘述。

[0071] 相似度确定单元410可分别针对每个候选确定候选的类别与候选的上下文模式是否对应。在候选的类别与候选的上下文模式对应时，相似度确定单元410确定候选在内容中的上下文与候选在其他内容中的上下文的相似度。在候选的类别与候选的上下文模式不对应时，相似度确定单元410停止进行操作。

[0072] 这里的相似度可以是通过现有的相似度技术确定的相似度，也可以是前面描述的本发明所定义的相似度。

[0073] 链接添加单元420在由相似度确定单元410确定的相似度大于预定阈值时，向候选添加链接。否则，在由相似度确定单元410确定的相似度不大于预定阈值时，链接添加单元420不向候选添加链接。

[0074] 根据本发明的在内容中添加链接的方法和设备，可以实现自动在内容中添加链接，从而可以避免了人工进行链接的添加，提高了效率。此外，根据本发明的在内容中添加链接的方法和设备在内容中所添加的链接能够反映网络用户对链接的需要。

[0075] 此外，应该理解，根据本发明的在内容中自动添加链接的方法也可实现为计算机可读记录介质上的计算机可读代码。计算机可读记录介质是可存储其后可由计算机系统读出的数据的任意数据存储装置。计算机可读记录介质的示例包括：只读存储器(ROM)、随机存取存储器(RAM)、CD-ROM、磁带、软盘、光数据存储装置和载波(诸如经有线或无线传输路径通过互联网的数据传输)。计算机可读记录介质也可分布于连接网络的计算机系统，从而计算机可读代码以分布式存储和执行。此外，完成本发明的功能程序、代码和代码段可容易地被与本发明相关的领域的普通程序员在本发明的范围之内解释。

[0076] 此外，根据本发明的示例性实施例的在内容中自动添加链接的设备中的各个单元可被实现硬件组件。本领域技术人员根据限定的各个单元所执行的处理，可以例如使用现场可编程门阵列(FPGA)或专用集成电路(ASIC)来实现各个单元。

[0077] 尽管已经参照其示例性实施例具体显示和描述了本发明，但是本领域的技术人员应该理解，在不脱离权利要求所限定的本发明的精神和范围的情况下，可以对其进行形式和细节上的各种改变。

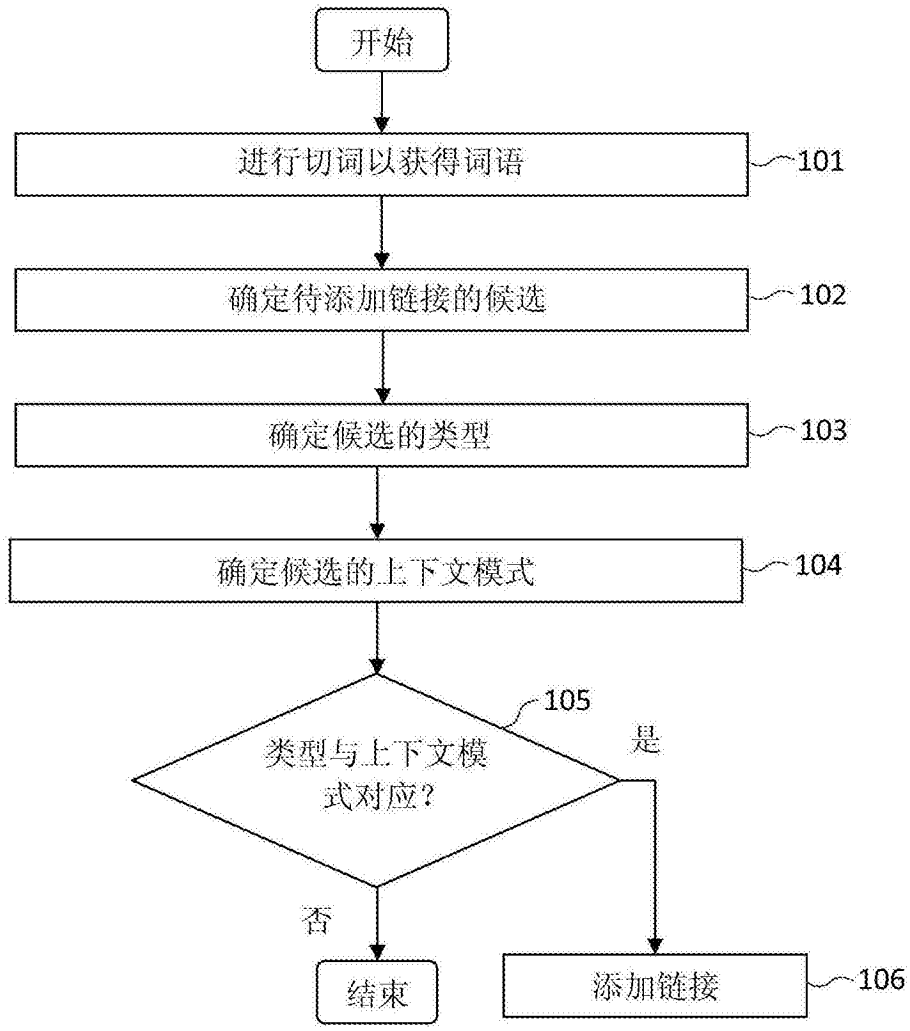

开始
进行切词以获得词语
101
确定待添加链接的候选
102
确定候选的类型
103
确定候选的上下文模式
104
105
类型与上下文模式对应？
是
否
结束
添加链接
106

图1

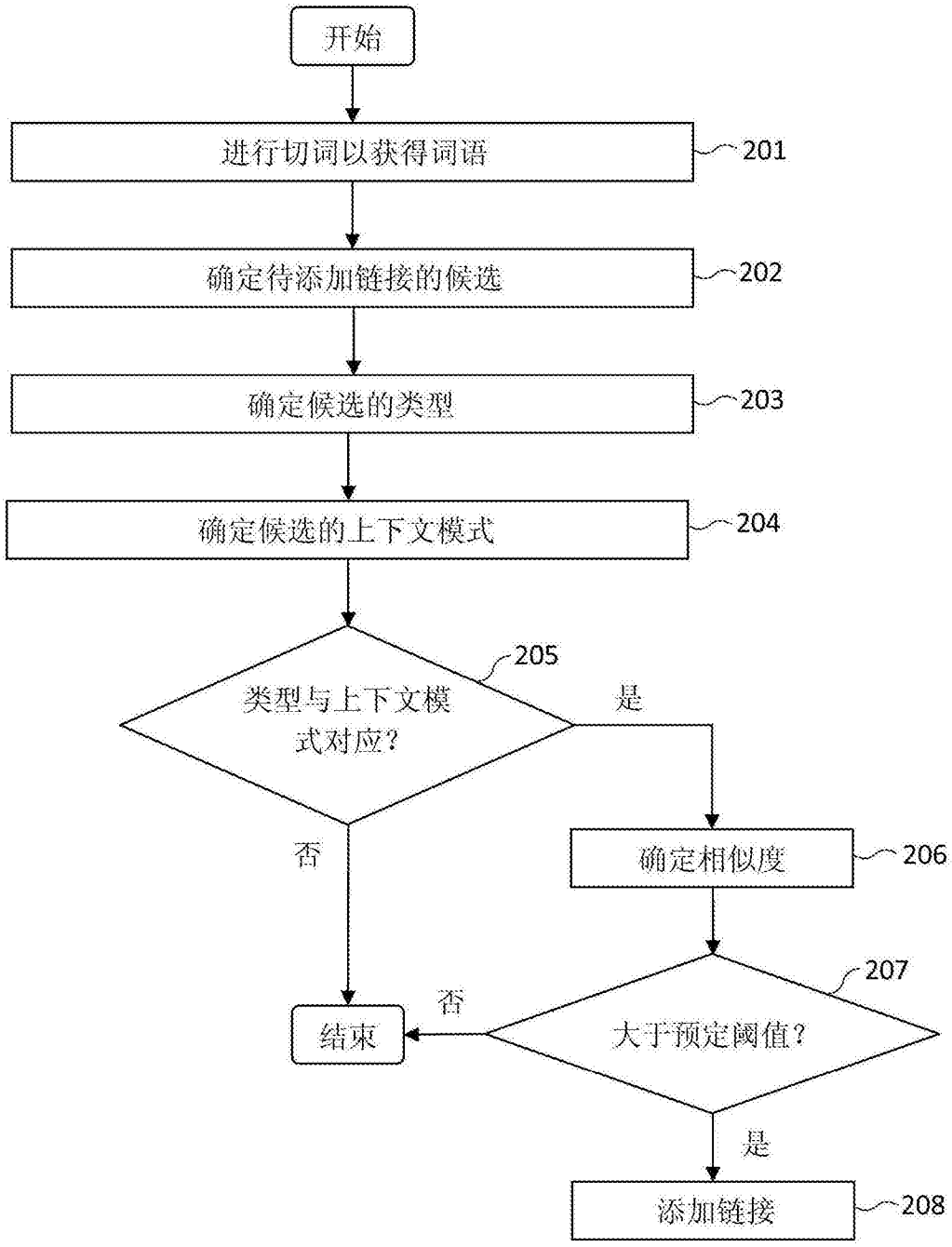

开始
进行切词以获得词语
201
确定待添加链接的候选
202
确定候选的类型
203
确定候选的上下文模式
204
类型与上下文模式对应?
205
是
否
确定相似度
206
结束
否
大于预定阈值?
207
是
添加链接
208

图2

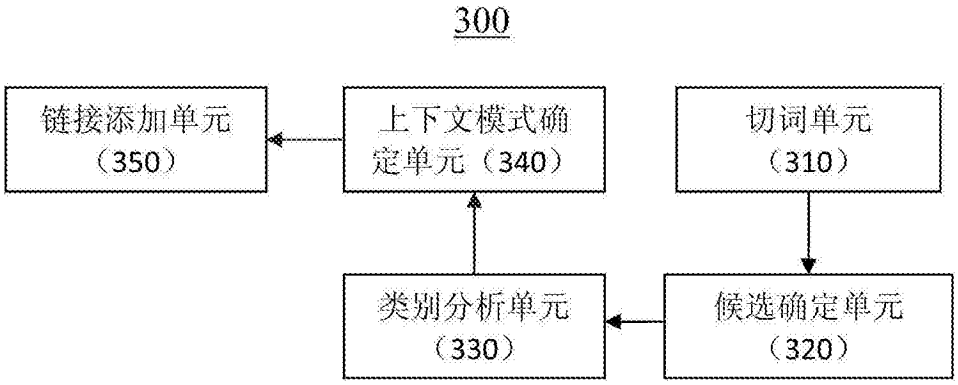

300
链接添加单元（350）
上下文模式确定单元（340）
切词单元（310）
类别分析单元（330）
候选确定单元（320）

图3

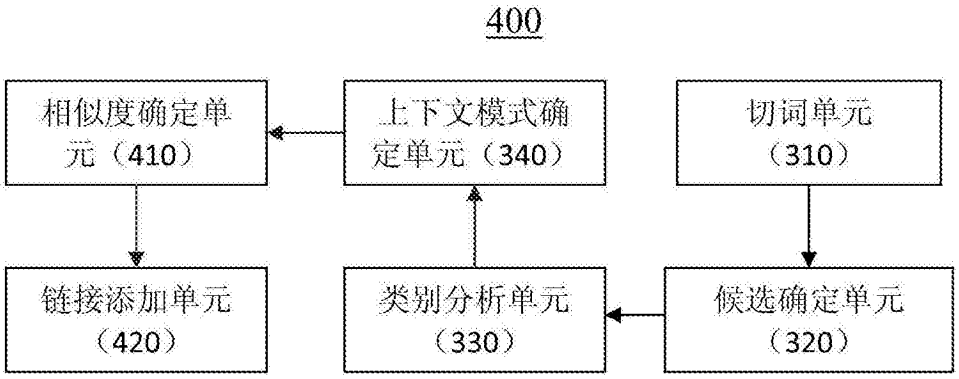

400
相似度确定单元（410）
上下文模式确定单元（340）
切词单元（310）
链接添加单元（420）
类别分析单元（330）
候选确定单元（320）

图4