(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 102945237 B
(45)授权公告日 2016.08.10

(21)申请号 201210199408.7

(22)申请日 2006.05.04

(30)优先权数据
11/122,873 2005.05.04 US

(62)分案原申请数据
200680021940.2 2006.05.04

(73)专利权人 谷歌公司
地址 美国加利福尼亚州

(72)发明人 吴军 林德康 钱哲 周杰

(74)专利代理机构 中原信达知识产权代理有限责任公司 11219
代理人 周亚荣 安翔

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

(56)对比文件
US 6006225 A,1999.12.21,对比文件1说明书第4栏第9-12行,第6栏第65一第7栏第3行,第10栏8-33行,第12栏第25一第13栏第10行、说明书附图5A, 7, 8A, 8B.
CN 1460948 A,2003.12.10,全文.
WO 2004086192 A2,2004.10.07,全文.
WO 2005026989 A1,2005.03.24,全文.
李文翔等.基于差异相似矩阵算法的词语用法知识发现.《计算机应用与软件》.2005,第22卷(第1期),92-94.

审查员 董洪梅

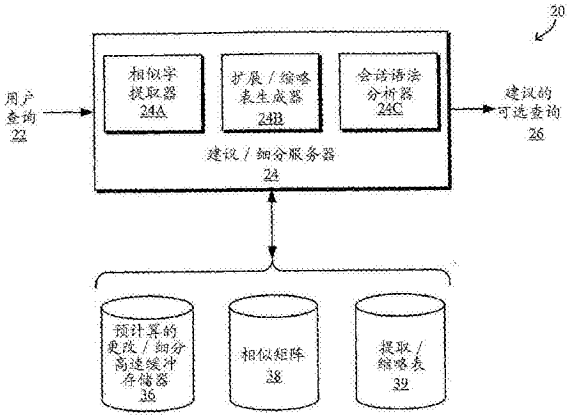
权利要求书2页 说明书10页 附图10页

(54)发明名称

基于原始用户输入建议和细分用户输入的系统和方法

(57)摘要

提供了一种基于原始用户输入建议和细分用户输入的系统和方法。该方法可以针对基于罗马语系的语言和/或诸如汉语的基于非罗马语系的语言来实现。该方法通常可以包括:接收原始用户输入并识别其中的核心词语;通过根据相似矩阵用另一词语替代原始输入中的核心词语和/或通过根据扩展/缩略表用另一个字序列代替原始输入中的字序列来确定潜在可选输入,其中,一个字序列是另一个字序列的子串;计算每个潜在可选输入的似然;以及根据预定标准(例如,可选输入的似然至少是原始输入的似然)来选择最有可能的可选输入。可以提供包含预计算的原始用户输入和相应的可选输入的高速缓冲存储器。



CN 102945237 B

1.一种用于生成替代用户输入的计算机实现的方法,包括:

接收原始用户输入,所述原始用户输入具有至少一个核心词语;

生成一个或多个替代用户输入,包括使用相似词语的相似性矩阵将所述原始用户输入中的核心词语替换为第二词语,所述相似词语的相似性矩阵包括在所述核心词语和所述第二词语之间的相似性值,所述相似性矩阵是通过以下操作来构建的:

构建出现在语料库中的一个或多个词语中的每个词语的特征矢量,其中所述一个或多个词语包括所述核心词语,并且其中该特征矢量包括该特征矢量中的每个特征的计数;

针对所述一个或多个词语中的每个词语的特征矢量,将该特征矢量中的每个特征的值确定为该特征和该词语之间的点态交互信息;

将在所述核心词语的第一特征矢量和所述语料库中出现的所述一个或多个词语的每个相应特征矢量之间的相应相似性测度确定为在所述第一特征矢量和该相应特征矢量之间的角的余弦;以及

使用所确定的在所述核心词语的第一特征矢量和所述语料库中出现的所述一个或多个词语的相应特征矢量之间的所述相应相似性测度来构建所述相似性矩阵;

确定所述一个或多个替代用户输入中的每一个替代用户输入的相应的分数,其中该分数基于该替代用户输入与所述原始用户输入之间的相关性;

选择其分数至少是所述原始用户输入的分数的一个或多个替代用户输入;以及

将所选择的所述一个或多个替代用户输入作为所述原始用户输入的替选存储在查询细化高速缓冲存储器中。

2.根据权利要求1所述的方法,还包括:

确定所述原始用户输入出现在所述查询细化高速缓冲存储器中;以及

响应于所述原始用户输入提供所述一个或多个相关的替代用户输入作为查询建议。

3.根据权利要求1所述的方法,还包括:

对于所述一个或多个词语中的每个词语的特征矢量的每个特征,计算所述核心词语和该特征之间的点态交互信息并且将所述核心词语和该特征之间的所述点态交互信息用作所述特征的值。

4.根据权利要求1所述的方法,还包括:

从用户输入日志或用户输入数据库或两者生成扩展/缩略表,其中所述扩展/缩略表包括表示字序列出现的频率值。

5.根据权利要求4所述的方法,其中生成所述扩展/缩略表包括确定频繁出现的字序列,滤出非短语字序列,以及将计数与词语序列相关联作为所述频率值。

6.一种用于生成替代用户输入的系统,包括:

查询细化高速缓冲存储器,其将替代用户输入存储为原始用户输入的替选;以及

建议/细分服务器设备,其被配置为接收原始用户输入,所述原始用户输入具有至少一个核心词语,并且执行如下操作:

生成一个或多个替代用户输入,包括使用相似词语的相似性矩阵将所述原始用户输入中的核心词语替换为第二词语,所述相似词语的相似性矩阵包括在所述核心词语和所述第二词语之间的相似性值,所述相似性矩阵是通过以下操作来构建的:

构建出现在语料库中的一个或多个词语中的每个词语的特征矢量,其中所述一个或多

个词语包括所述核心词语，并且其中所述特征矢量包括所述特征矢量中的每个特征的计数；

针对所述一个或多个词语中的每个词语的特征矢量，将该特征矢量中的每个特征的值确定为该特征和该词语之间的点态交互信息；

将在所述核心词语的第一特征矢量和所述语料库中出现的所述一个或多个词语的每个相应特征矢量之间的相应相似性测度确定为在所述特征矢量之间的角的余弦；以及

使用所确定的在所述核心词语的第一特征矢量和所述语料库中出现的所述一个或多个词语的相应特征矢量之间的所述相应相似性测度来构建所述相似性矩阵；

确定所述一个或多个替代用户输入中的每一个替代用户输入的相应的分数，其中该分数基于该替代用户输入与所述原始用户输入之间的相关性；

选择其分数至少是所述原始用户输入的分数的一个或多个替代用户输入；以及

将所选择的所述一个或多个替代用户输入作为所述原始用户输入的替选存储在所述查询细化高速缓冲存储器中。

7.根据权利要求6所述的系统，其中所述建议/细分服务器设备进一步被配置为执行如下操作：

确定所述原始用户输入出现在所述查询细化高速缓冲存储器中；以及

响应于所述原始用户输入提供所述一个或多个相关的替代用户输入作为查询建议。

8.根据权利要求6所述的系统，其中所述建议/细分服务器设备进一步被配置为执行如下操作：

对于所述一个或多个词语中的每个词语的特征矢量的每个特征，计算所述核心词语和该特征之间的点态交互信息并且将所述核心词语和该特征之间的所述点态交互信息用作该特征的值。

9.根据权利要求6所述的系统，其中所述建议/细分服务器设备进一步被配置为执行如下操作：

从用户输入日志或用户输入数据库或两者生成扩展/缩略表，其中所述扩展/缩略表包括表示字序列出现的频率值。

10.根据权利要求9所述的系统，其中生成所述扩展/缩略表包括确定频繁出现的字序列，滤出非短语字序列，以及将计数与词语序列相关联作为所述频率值。

基于原始用户输入建议和细分用户输入的系统和方法

[0001] 本申请是分案申请，其原案申请的申请号为200680021940.2，申请日为2006年5月4日，发明名称为“基于原始用户输入建议和细分用户输入”。

技术领域

[0002] 本发明总的来说涉及生成可选用户输入。更具体地，公开了基于诸如搜索查询的原始用户输入生成更改和细分的用户输入的系统和方法。

背景技术

[0003] 在给定的搜索会话(session)期间，许多用户常常，有时重复地，更改或细分其原始搜索查询。例如，用户可以将原始搜索查询更改为更具体的搜索查询、更宽泛的搜索查询、和/或使用可选的查询词语(term)的搜索查询，直到生成期望的搜索结果。用户搜索查询细分通过基于罗马语系语言(例如，英语)的查询、以及通过基于非罗马语系语言(例如，汉语、日语、韩语(CJK)、泰国语等)来产生。当原始搜索查询未产生一组好的搜索结果时(例如，如果搜索查询太具体或太宽泛，或者如果搜索查询使用不合适的词语)，用户通常更改或细分他们的搜索查询。例如，当一个或多个搜索词语有多种意思且一些返回文档涉及不同于用户想要的多义搜索词语的一个意思时、和/或用户仅对搜索词语的许多方面中给定搜索词语的一个方面感兴趣时，原始用户搜索查询可能产生许多不相关的结果。当用户仅探究关于所指定搜索词语的概念时，原始用户搜索查询也可能产生很多不相关的结果。

[0004] 许多搜索引擎提供与用户原始搜索查询相关的一系列建议的搜索查询。例如，如果用户的原始搜索查询是“Amazon”，则搜索引擎可以建议其它相关的搜索查询，例如，“Amazon.com”、“Amazon Rainforest”、以及“Amazon River”。搜索查询建议对于基于非罗马语系语言用户(例如，CJK用户)特别有用。具体地，因为基于非罗马语系语言通常具有一组大量的字符且每个字符可能都需要多次按下使用传统基于罗马语系的键盘的按键，所以基于非罗马语系语言用户可以优选单击或选择全部键入的更改搜索查询中的一个建议的搜索查询。例如，许多汉语用户使用拼音(语音拼写法)来输入汉语字符。典型地，传统拼音输入系统转换拼音输入，并提供用户可以从中选择期望的汉语字符集的一组候选的汉语字符集。明显的是，多步输入处理将是繁重且耗时的。

[0005] 搜索查询建议也将对基于罗马语系的语言用户有用。许多搜索引擎(例如，Yahoo、Teoma、Alta Vista、Askjeeves、AllTheWeb以及Baidu)都提供例如以相关搜索、查询细分、或查询分簇形式的特征。

发明内容

[0006] 公开了一种基于原始用户输入(例如，搜索查询)来生成更改或细分的用户输入的系统和方法。应当理解，本发明可以多种方式来实施，这些方式包括诸如处理、设备、系统、装置、方法、或其中的程序指令通过光通信线路或电子通信线路来发送的计算机可读介质(例如，计算机可读存储介质或计算机网络)。术语计算机通常指具有计算能力的任何装置，

例如，个人数字助理(PDA)、蜂窝式电话、和网络交换机。以下将描述本发明的几个创造性实施例。

[0007] 该方法可以被应用于基于非罗马语系语言(例如，汉语)的查询。该方法通常可以包括接收和识别原始用户输入中的核心词语；通过根据相似矩阵用另一词语代替原始输入中的核心词语、和/或根据扩展/缩略表用另一个字序列替代原始输入中的字序列来确定潜在可选的用户输入，其中，一个序列是另一个序列的子串；计算潜在可选的用户输入的似然；以及根据预定标准(例如，每个所选的可选用户输入的似然至少为原始用户输入的似然)来选择最可能的可选用户输入。该方法还可以包括确定原始用户输入是否在所建议的可选用户输入的预计算高速缓冲存储器中，如果在，则输出存储在预计算高速缓冲存储器中的预计算的最可能的可选用户输入。

[0008] 相似矩阵可以利用语料库生成，且可以具有两个相似词语(包括例如“New York”和“Los Angeles”的短语词)之间的相似值，尽管每个对应词语对(New和Los和York和Angeles)不具有高相似性，但这些短语词可以具有非常高的相似性。在一个实施例中，可以通过构建对于语料库中的字的特征向量并利用他们的特征向量来确定两个字/短语之间的相似值来生成相似矩阵。

[0009] 可以从用户输入数据库中生成扩展/缩略表，且可以具有与每对词语序列相关的频率值。在一个实施例中，可以通过确定常用的字序列、滤出非短语字序列、以及使计数与每个术语序列相关作为频率值来生成扩展/缩略表。仅为了示出，扩展/缩略表中的项的实例可以为“The United State of America”和“United States”。

[0010] 可以通过确定以下的至少一项来计算潜在可选用户输入的似然：(a)原始用户输入和潜在可选用户输入之间的关联性，(b)用户将选择潜在可选用户输入的概率，以及(c)潜在可选用户输入的位置分数(score)。具体地，可以使用原始输入和潜在可选用户输入的调整(align)词语之间的相关值来确定原始用户输入和潜在可选用户输入之间的关联性。

[0011] 在另一实施例中，用于建议可选用户输入的系统通常包括建议/细分服务器，被配置为接收具有至少一个核心词语的原始用户输入；从原始用户输入中识别出核心词语；通过执行以下中的至少一项来确定潜在可选用户输入：(a)根据相似矩阵用另一词语来代替原始用户输入中的至少一个核心词语，该相似矩阵具有两个词语之间的相似值，(b)根据扩展/缩略表用另一个字序列来替代原始用户输入中的字序列，其中，一个序列是另一个序列的子串，该扩展/缩略表具有与每个词语序列相关联的频率值，计算潜在可选用户输入的似然；以及根据预定标准来选择和输出最可能的可选用户输入。

[0012] 在又一实施例中，用于建议可选用户输入的计算机程序产品和计算机系统一起使用，该计算机程序产品包括在其上存储有在计算机处理器上可执行的指令的计算机可读存储介质。该指令通常可以包括接收和识别原始用户输入中的核心词语；通过根据相似矩阵用另一个词语代替在原始输入中的核心词语、和/或根据扩展/缩略表用另一个字序列代替原始输入中的字序列来确定潜在可选用户输入，一个序列是另一个序列的子串；用潜在可选用户输入计算潜在可选用户输入的似然和可选地计算预测用户满意度；以及根据预定标准(例如，每个所选的可选用户输入的似然至少为原始用户输入的似然)来选择最可能的可选用户输入。

[0013] 实现系统和方法的应用可以在服务器站点(例如，在搜索引擎上)上实现，或者可

以在客户端站点上(例如,用户计算机)被实现(例如,下载),以提出建议的可选输入或与远程服务器(例如,搜索引擎)连接。

[0014] 下面将通过下列详细描述和附图更详细地呈现本发明的这些和其它特征以及优点,其中,以本发明的实例原理的方式示出了附图。

附图说明

[0015] 结合附图,通过下面详细的描述,将更好地理解本发明,其中,相同的参考标号表示相同的结构元件。

[0016] 图1A是用于生成诸如用户搜索查询的建议的更改/细分的用户输入的示例性系统的框图。

[0017] 图1B是示出用于通过建议－细分服务器的相似字提取器生成相似矩阵的过程的框图。

[0018] 图1C是示出用于通过建议－细分服务器的扩展/缩略表生成器生成扩展/缩略表的过程的框图。

[0019] 图1D是示出用于通过建议－细分服务器的会话语法分析器生成初始的更改/细分高速缓冲存储器的过程的框图。

[0020] 图2A是示出用于生成如可以由图1A所示的系统实现的更改/细分用户输入高速缓冲存储器的示例性过程的流程图。

[0021] 图2B是示出用于生成如可以由图1A所示的系统实现的诸如用户查询的建议的更改/细分用户输入的示例性过程的流程图。

[0022] 图3示出了通过对原始用户查询进行语法分析而生成的示例性查询网格。

[0023] 图4是示出了用于通过代替查询词语来构建在生成建议的更改/细分查询中使用的相似矩阵的示例性过程的流程图。

[0024] 图5是列出了对于从示例性文本中生成的词语“communities”的特征和对应计数的表。

[0025] 图6是列出了对于从语料库中生成的词语“communities”的示例性特征和对应计数的表。

[0026] 图7是用于代替词语以生成建议的更改/细分查询的示例性相似矩阵。

[0027] 图8是示出用于通过替代查询中的复合字来构建在生成建议的更改/细分查询中使用的复合字对的提取/缩略表的示例性过程的流程图。

[0028] 图9是示出用于代替查询中的复合字来生成建议的更改/细分查询的扩展/缩略表中的一些示例性项的表格。

[0029] 图10是示出用于确定建议的更改/细分查询分数的示例性过程的流程图。

[0030] 图11示出两个查询Q和Q'的词语对准映射的实例。

[0031] 图12是示出用于生成用于检测的新项的相关值的示例性过程的流程图。

具体实施方式

[0032] 披露了一种基于原始用户输入(例如,搜索查询)来生成更改或细分的用户输入的系统和方法。应当注意,仅为了清楚,通常以汉语查询输入的词语来表示文中表示的实例。

然而，用于建议细分/更改的用户输入的系统和方法可以类似地应用于其它基于非罗马语系的语言（例如，日语、韩语、泰国语等）以及基于罗马语系的语言。此外，用于建议的细分/更改用户输入的系统和方法可以类似地应用于其它非查询用户输入。表示下列描述，以使任何本领域技术人员能够制造和使用本发明。仅提供了特定实施例和应用的描述作为实例，并且各种更改对于本领域技术人员来说是显而易见的。在不背离本发明的精神和范围的情况下，可将文中限定的主要原理应用于其它实施例和应用。因此，本发明应与包括与文中公开的原理和特征相一致的各种改变、更改和等同替换的最宽范围相一致。为了清楚，未对关于与本发明相关的的技术领域中已知的技术资料的细节进行详细描述，以避免对本发明产生不必要的混淆。

[0033] 该系统和方法基于原始用户输入（例如，查询）、基于用户的查询历史和用户查询词语之间的关系来生成更改或细分的用户输入。该系统和方法可以包括用于提取包括新名称项（例如，适当名称、电影名、歌曲和产品等）的新词语以及词语之间关系的系统和方法。文中描述的系统和方法适用于生成查询（或其它用户输入）细分，并且还将适用于诸如新文章分类、拼写校正、媒体搜索和分段的许多其它应用。对于许多用户，初始搜索查询通常不是最佳搜索查询，因此在给定搜索会话期间，用户有时多次更改或细分搜索查询。

[0034] 图1A是用于从诸如用户搜索查询22的原始用户输入中生成建议更改/细分的输入26的示例性系统20的框图。系统20通常包括建议/细分服务器24，其使用可以从多个数据源中推导出的概率来生成建议的更改/细分查询26。多个数据源的实例包括可选的建议/细分高速缓冲存储器36，其存储预先计算的查询建议或细分的高速缓冲存储器。首先，可以通过建议/细分服务器24的会话语法分析器24C生成建议/细分高速缓冲存储器36。另一数据源可以是可由建议/细分服务器24的相似字提取器24A生成的相似矩阵38，以及由扩展/缩略表生成器24B生成的扩展/缩略表39。相似矩阵38和扩展/缩略表39通常接近于词语和/或词语序列之间的关系。系统20可以周期性地更新和/或再生相似矩阵38和/或扩展/缩略表39。以下将参照图1B至图1D分别详细描述建议/细分服务器24的相似字提取器24A、扩展/缩略表生成器24B、和会话语法分析器24C。

[0035] 图1B是示出用于通过相似字提取器24A生成相似矩阵38的过程的框图。如图所示，相似字提取器24A可以利用用于生成相似矩阵38的多种数据源。由相似字提取器24A利用的数据源的实例包括诸如网络语料库30的语料库（例如，新闻、网页、和链接锚文本信息）、查询和相关用户选择（例如，存储在查询日志32中的查询和相关用户选择）、和/或可以包括在每个给定会话中的查询历史的会话数据34。网络语料库30还可以包括链接锚文本信息。例如，查询日志32不仅可以包括用户查询日志，还可以包括由用户进行的搜索结果选择以及在返回搜索结果之前用户停留在所选搜索结果处的持续时间。

[0036] 图1C是示出用于通过扩展/缩略表生成器24B生成提取/缩略表39的过程的框图。如图所示，扩展/缩略表生成器24B可以将查询日志32和/或会话数据34用作用于生成提取/缩略表39的数据源。图1D是示出用于通过会话语法分析器24C生成初始更改/细分高速缓冲存储器36a的处理的框图。如图所示，会话语法分析器24C将会话数据34用作其用于生成初始更改/细分高速缓冲存储器36a的数据源。

[0037] 图2A和图2B是示出可以通过建议/细分服务器24来执行示例性过程的流程图。具体地，图2A是示出了用于生成如可以通过图1A示出的系统20实现的更改/细分用户输入高

速缓冲存储器的示例性处理40的流程图。在方框41处，可以使用会话语法分析器从会话数据中生成初始更改/细分高速缓冲存储器。注意，会话数据可以包括在每个给定用户输入或查询会话中的查询历史。然后，过程40进入包括方框43～48的循环，该循环针对在方框42中的预定数量的最普通用户输入中的每一个，例如，查询。具体地，在方框43处，过程对高速缓冲存储器中的建议的更改/细分查询执行查找。在方框43处的查找可以生成建议1、2、…M。

[0038] 更改/细分高速缓冲存储器中的每个用户输入或查询项可以包含一列预定数量N个的建议查询。因此，为生成建议M＋1、M＋2…N，即，为填补每个查询的建议查询列，还可以执行方框44～47。具体地，在方框44和45处，可以(理论上地)构建扩展查询网格，以生成附加的建议的更改/细分(可选)查询。方框44通常表示词语替代查询更改/细分方法，而方框45通常表示扩展/缩略查询更改/细分方法。具体地，在方框44处，可以使用相似词语的相似矩阵用相似词语替代原始查询中的词语来创建扩展查询网格。词语替代用相似字或词语替代原始查询中的字或词语(包括短语词)。相似词语可包括同义字或近义字(例如，社区和相邻区域)、缩写词、和/或在相同语法/语义类别中的词语(例如，Toyota和Honda、Dell和HP、DVD和数码相机、以及Nokia和Motorola)。

[0039] 在方框45处，可以通过使用复合字对的扩展/缩略表在原始查询中添加/删除词语来附加地和/或选择性地构建扩展查询网格。具体地，扩展/缩略表中的每一项都是一个复合字对，其中，一个复合字是另一个的子串，例如，$T_1T_2<=>T_1T_2T_3$，以及$T_4T_5T_6<=>T_4T_5$。汉语复合字对的实例包括上海和上海市以及电视和电视机。复合字对可以包括多义词语和它们明确的上下文(例如，Amazon和Amazon rain forest和/或Amazon.com)、概念及其细分(例如，cell和stem cell和/或cell phone)、词语及其属性(例如，计算机和存储器、硬盘驱动器、和/或DVD驱动器)、以及名称(例如，人名、公司名称等)和他们对应的活动性、工作、产品等(例如，诸如Tom Hanks和Forrest Gump的演员-电影、诸如Apple和iPod的公司-产品、如Bill Gates和Microsoft或CEO的法人-公司或头衔、作者-书、歌手-歌曲等)。

[0040] 在构建了包含多种可选路径的扩展查询网格之后，在方框46处，对于扩展查询网格中预定数量的最佳查询的路径和分数被识别为潜在建议的查询。在方框47处，计算原始普通用户查询的分数，使得仅提供其分数至少是那个原始普通用户查询分数的潜在建议的查询，作为建议的更改/细分查询。分数可以表示作为由用户选择或期望的查询的给定查询(原始或潜在建议的查询)的似然。可以仅提供其分数至少是原始普通用户查询分数的查询，作为建议的更改/细分查询来填充更改/细分高速缓冲存储器的建议列表项。可以将生成的建议的查询存储到预计算的更改/细分查询高速缓冲存储器。注意，过程40或包括方框42～49的循环可以周期性重复，以更新更改/细分高速缓冲存储器。

[0041] 图2B是示出用于生成如可以由图1A所示的系统实现的诸如用户查询的建议的更改/细分用户输入的示例性过程50的流程图。在方框51处，接收诸如用户查询的用户输入。在判断框52处，可以将在方框51接收的原始用户输入与在可选的预计算的更改/细分高速缓冲存储器中的项进行比较。如果在判断框52处确定原始用户查询处于建议/细分高速缓冲存储器中，则在方框53，由来自预计算的更改/细分高速缓冲存储器的多至N个预计算查询建议至少部分地填充大小为N的查询建议列表。如果如在判断框54处确定的建议列表被填满，则过程50完成。注意，建议列表可以具有预定大小N，例如，10个建议或单个最佳建议。可选地，如果如在判断框54处确定的建议列表未被填满，则过程50继续执行方框55和56。类

似地，如果在判断框52处原始用户查询确定未处于建议/细分高速缓冲存储器中，则过程也继续执行方框55和56。注意，参照图2A，方框55～58类似于如上所述的过程40的方框44～47。因此，为了清楚，某种程度上它们是相似的描述将不在这里重复。

[0042] 在方框55和56处，(概念上)构建扩展查询网格来生成建议的更改/细分(可选)查询。在构建包含多个可选路径的扩展查询网格之后，在方框57处，对于扩展查询网格中预定数量的最佳查询的路径和分数被识别为潜在建议查询。在方框58处，计算原始用户查询的分数，使得仅提供其分数至少是原始用户查询的分数的潜在建议查询，作为建议的更改/细分查询。其分数至少是原始用户查询的分数的查询可以被提供给用户，作为建议的更改/细分查询来填充建议列表或建议列表的剩余部分。尽管未示出，可以可选地提供单个最佳查询。此外，可以将原始用户查询和生成的建议查询附加地存储到预计算的更改/细分查询高速缓冲存储器。

[0043] 下面将更详细地描述如上面参照图2A和图2B示出并描述的用于生成建议的更改/细分用户输入的过程40和50的各个方框。

[0044] 图3示出示例性的扩展查询网格示意图。如图所示，原始查询可以包括多个核心字或词语T_1、T_2、T_3、T_4，以及非核心字或词语s_1、s_2、s_3。例如，在汉语查询“新浪的URL”中，核心词语或项是“新浪”而不是“URL”。非核心术语通常还包括无用字。无用字通常被定义为在诸如网络语料库的语料库中的30个最频繁出现的汉语字或100个最频繁出现的英文字。

[0045] 在识别原始查询的核心项之后，可以应用一个或多个查询更改或细分方法(例如，词语替代和/或扩展/缩略)来构建扩展查询网格。如上所述注意，词语替代是指替代与核心项类似(例如，同义字或近义字)并可以使用诸如相似矩阵来识别的字和/或词语。仅为了示例性的目的，图3示出可以通过用词语T_1’或T_1”来替代词语T_1和/或用词语T_4’替代词语T_4来构建扩展查询网格。

[0046] 同样如上所注意的，扩展/缩略是指利用诸如复合字的扩展/缩略表来将核心项添加到原始查询中和/或从原始查询中删除一些核心项。仅为示例性的目的，复合字的扩展/缩略表可以包括复合字对T_1T_2和$T_1T_2T_5$的表项，以使图3的原始查询中的复合字T_1T_2可以用复合字$T_1T_2T_5$(即，添加新词语T_5)替代，从而进一步构建扩展查询网格。类似地，复合字的扩展/缩略表还可以包括对于复合字对$T_2T_3T_4$和T_3T_4的表项，以使图3的原始查询中的复合字$T_2T_3T_4$可以用复合字T_3T_4(即，删除核心项T_2)替代，从而进一步构建扩展查询网格。

[0047] 参照图4至图7，更详细地描述用于生成相似词语的相似矩阵的一种示例性的方法。图4是示出用于通过替代查询词语来构建在生成建议的更改/细分查询中使用的相似矩阵的示例性过程60的流程图。相似矩阵可以是在每对字或术语w之间分布字相似性的矩阵。可以通过构建对于在诸如网页的语料库中的每个字w的特征向量、并确定在每个对字的每个字之间的相似性作为其对应特征向量之间角的余弦来获得分布字相似性。字或词语的特征向量可以包括所有出现的字或词语的周围字(surrounding word)。

[0048] 尽管可以类似地采用各种其它特征向量和相似矩阵构建方法，但参照图4示出了构建特征向量和相似矩阵的一个实例。具体地，在方框62处，语料库(例如，网络语料库)中的每个字/词语的特征向量与特征向量中对于每个特征f的计数一起被构建。字/词语w的特征可以包括在字/词语w之前和之后直到第一个无用字出现的多个字。仅为示例性目的，对于给出的句子“Because communities assess at different percentages of fair

market value,the only way to improve tax rates among communities is by using equalized rates,”,在图5的表中列出了特征字communities及其对应的共同出现计数。注意,在具有不同形式的给定字(例如,诸如“community”和“communities”的单数或复数、或诸如“walk”、“walking”、和“walked”的不同时态)的语言中,系统可以将不同形式的字视为单个字,但通常是相似词语。这种不同形式的给定字的处理将与不具有这种区别的语言无关(例如,通常在例如汉语的情况下)。此外,还注意到,具有前缀“L:”或“R:”的特征分别是表示字w的左侧或右侧的字。在该实施例中,在字w的给定情况的左侧和右侧中的每一个上的一个或多个特征的计数和是1。例如,在字“communities”的第一种情况中,左侧和右侧特征中的每一个都被分配为计数1。此外,当在字“communities”附近存在一个或多个无用字(例如,“between”、“is”、和“by”)时,字“communities”的给定情况的每一侧的特征计数可被平分,从而,被计算为小数。在字“communities”的第二种情况中,存在两个左侧特征,使得为每个左侧特征分配计数0.5。类似地,在字“communities”的第二种情况中,存在两个右侧特征,使得为每个右侧特征分配计数0.33。

[0049] 再次参照图4,在方框64处,特征向量中的每个特征f的值可以被确定为在字w和特征f之间的点态(point-wise)交互信息MI。使用点态交互信息MI的值,这是因为虽然频繁出现的字(例如,无用字)容易具有较高计数,但这种字常常是无意义的。因此,因为特征计数可能不是重要特征较好的指示符,所以字w和特征f之间的点态交互信息MI(w,f)可被用作特征f的值。点态交互信息MI(w,f)可以被定义为w和f的有效联合概率P(w,f)、以及w的期望概率P(w)、和f的期望概率P(f)之间的对数比,如果他们是相互独立的,则共同存在:

[0050] $$MI(w,f)=P(w\cdot f)\cdot\log\frac{P(w\cdot f)}{P(w)\cdot P(f)}$$

[0051] 其中,可以使用诸如语料库中的其各自概率来确定特征概率P(f)和字概率P(w)(例如,相对频率)。仅作为实例,图6是列出了对于从网络语料库中生成的词语“communities”的示例性特征和相应概率的表。图6中示出的特征向量表列出了字“communities”的特征子集以及特征与字“communities”之间的概率和交互信息。注意,特征向量可能相当大。例如,从语料库中提取的字communities的全部特征集包括大约2000项。

[0052] 再次参照图4,在方框66处,将使用特征向量中的特征值,将两个字或短语w_1和w_2之间的相似性测量值或值sim确定为其特征向量之间角的余弦。具体地,两个术语或字w_1和w_2之间的相似性sim可以被定义为:

[0053] $$sim(w_1,w_2)=\frac{\sum_i f_{1i}f_{2i}}{\sqrt{\sum_i f_{1i}^2}\sqrt{\sum_i f_{2i}^2}}$$

[0054] 其中,w_1和w_2的特征向量分别由(f_{11},f_{12}…,f_{1n})和(f_{21},f_{22}…,f_{2n})表示。

[0055] 然后,在方框68处,根据对于词语的每个字对的相似值构建相似矩阵,并可以通过用相似词语替代查询词语来生成建议的更改/细分查询。具体地,例如,相似值可用于确定潜在建议查询的分数。注意,可周期性地重新计算相似矩阵,和/或可以将词语的相似值(例如,新近识别的那些词语)添加到矩阵。图7是用于代替词语以生成建议的更改/细分查询的示例性的相似矩阵。

[0056] 现在，将参照图8至图11更详细地描述用于生成在应用现存的词语替代查询更改/细分方法中使用的相似矩阵的示例性方法、用于生成在应用扩展/缩略查询更改/细分方法中使用的复合字对的扩展/缩略表的示例性方法。图8是示出用于构建复合字对的提取/缩略表的示例性过程70的流程图。如上所述注意，扩展/缩略表中的每项都是一对复合字，该复合字对中，一个复合字是另一个的子串，使得如果查询包含扩展/缩略表的项中复合字对中的一个复合字，则该复合字可以被在延伸的网格中的复合字对项中的其它复合字替代。理想地，提取/缩略表中的每个复合字都应当是有意义的短语。仅作为实例，复合字对可以是上海和上海市，或者电视和电视机。如上所注意的，复合字对可以包括，例如，多义词语及其清晰的上下文(例如，Amazon和Amazon rain forest)、人名及其对应的活动、词语的属性、概念的细分、演员、作者、产品、法人地位等。

[0057] 在方框71处，查询日志(或用户输入的另一数据库)中的查询可以被分割成使查询的全部概率最大化的字序列。具体地，因为汉字并不需要用空格或其它分隔符明确地描述，使得查询可以是没有分隔符的汉语字符串，所以分割器可以用于将字符序列分割成字序列。字序列可以使字概率的积是所有可能字符序列段中的最大序列。显然，不需要对在相邻字之间存在清晰描述的某种语言(例如，英语)执行方框71。

[0058] 在方框72处，为了识别复合字/短语，识别常用字序列或n元(gram)(n个序列的序列)。同样在方框72处，对字序列中所有相邻字对是常用n元的字序列进行计数，以识别任意长度的常用字序列。注意，常用字序列可以是或者不是复合字。例如，某些常用字序列可以是复合字，而其它字序列可以是非短语或非复合字序列。

[0059] 在方框73处，通过需要复合字/短语在最少数量查询的开始以及结尾处出现，来识别非短语序列(但在相同的查询中不是必须的)。查询的最小数量可以是大于或等于1的任意数，但通常远大于1，例如，50或100。

[0060] 在方框74处，对于语料库(例如，网络语料库)中的每个n元特征向量，与对于特征向量中的每个特征f的计数一起构建。在方框75处，特征向量中的每个特征f的值可以被确定为在n元和特征f之间的点态交互信息MI。在方框76处，可以利用特征向量中的特征值，将两个n元之间的相似测量值或值sim确定为其特征向量之间角的余弦。注意，方框74、75、和76分别类似于如参照图4描述的过程60的方框62、64、和66。因此，为了清楚目的，这里将不被重复描述在某种程度上与它们类似的描述。

[0061] 然后，在方框77处，扩展/缩略表可被构建为复合字对，在该复合字对中，一个复合字是另一个复合字的子串。此外，还可以确定多个复合字的计数并将其存储在扩展/缩略表中。

[0062] 图9是示出在替代查询中的复合字以生成建议的更改/细分查询中的扩展/缩略表中的一些示例性项的表格。如图所示，扩展/缩略表的每一行都包含两个复合字或字序列，其中，一个复合字是另一个复合字的子串。例如，每个复合字还与一个计数(或其它频率值)相关，该计数可以根据查询日志或一些其它用户输入数据库来确定。计数可被用作取舍点(cut off)以降低数据库的大小，和/或例如，通过使用log(计数)可以至少部分地用来确定词语或复合字的权重。如上所述，参照图2A和2B，一旦通过替换原始查询中的词语和/或通过添加/删除原始查询中的词语来延伸查询网格，则根据扩充网格来确定N个最佳查询的路径和分数，作为潜在建议查询。图10是示出了用于确定建议的更改/细分查询的分数(例如，

扩充查询网格中的路径)的示例性过程80的流程图。

[0063] 查询建议的确定可被视为在当前查询会话中基于在先查询的预测问题。给定当前搜索会话中的查询历史Q_1、Q_2,…,Q_{n-1},可以进行关于用户最可能选择的下一查询Q_n是什么的预测。建议或预测的下一查询Q_n应当不仅与当前会话中的查询历史Q_1、Q_2,…,Q_{n-1}关联,还应当产生好的搜索结果。例如,搜索结果有多好的测量值可以是单击位置(用户选择的搜索结果的位置)和单击持续时间(用户停留在所选的搜索结果页面多长时间)的函数。

[0064] 在一个实施例中,每个潜在建议查询的分数可以被确定为目标函数F的值:

[0065] $F(Q,Q_1,\cdots,Q_{n-1})=Rel(Q,Q_1,...,Q_{n-1})*Click(Q)*Position(Q)$;

[0066] 其中,

[0067] $Rel(Q,Q_1,...,Q_{n-1})$是查询历史Q_1、Q_2,…,Q_{n-1}和候选建议查询Q之间的关联性;

[0068] Click(Q)是用户将选择候选建议查询Q的概率;以及

[0069] Position(Q)是将要被单击的候选建议查询Q的搜索结果的位置。

[0070] 如上参照图2所述,一个或多个建议或预测的下一查询Q可以被提供给用户。因此,最佳的N个建议下一查询(例如,扩充查询网格中的路径)是具有最高目标函数值的N个查询,且最佳(例如,最可能的)的建议下一查询可以被表示为使目标函数F的值最大化的查询:

[0071] $Q_n=ArgMax\_Q\{F(Q,Q_1,...,Q_{n-1})\}$

[0072] 在图10的流程图中示出了每个潜在建议或预测的下一查询Q 80的分数的确定。在方框82中,利用查询中的调整词语的相关性来确定当前会话中的用户查询历史Q_1、Q_2,…,Q_{n-1}和预测的查询Q之间的关联性$Rel(Q,Q_1,...,Q_{n-1})$。具体地,为了估计关联函数Rel,识别原始查询Q中的词语或核心项。利用核心项之间的相关性,能够从其核心项的相关性中推导出两个查询Q和Q'之间的关联性Rel(Q,Q')。具体地,关联性Rel(Q,Q')可以被表示为:

[0073] $Rel(Q,Q')=Max\_f\ Prod\_\{i=1\}\hat{}k\ Cor(T_i,T_i')*w(T_i)$

[0074] 其中:

[0075] 调整函数(alignment function)$f=f(T_1,T_2,\cdots T_k,T_1',T_2',\cdots,T_k')$进行与相关查询Q和Q'的词语映射,例如,图11中示出的实例,在$\{T_1,\cdots,T_k,e\}$和$\{T_1',\cdots,T_k',e\}$之间的映射;

[0076] $Cor(T_i,T_i')$是词语T_i,T_i'之间的相关性,且是实数向量;

[0077] $Q=T_1,T_2,\cdots,T_k$(在任意词语T_i都可以是空词语e的查询Q中的核心项);

[0078] $Q'=T_1',T_2',\cdots T_k'$(在任意词语T_i'都可以是空词语e的查询Q'中的核心项);以及

[0079] $w(T_i)$是词语T_i的重要度,例如,T_i的TF/IDF,其中,TF表示词语频率(例如,词语的计数)以及IDF表示反向(inverted)文档频率。

[0080] 接下来,在方框84处,例如,根据单击持续时间或归一化的单击持续时间来确定用户将选择查询Q的概率Click(Q)。在方框86处,例如,根据单击位置、归一化的单击位置、或反向单击位置来确定预测查询Q的位置的分数Position(Q)。最后,在方框88处,根据如上所述的方框82、方框84、和方框86的结果来确定潜在建议或预测的下一查询Q的目标函数F的值。

[0081] 参照图12,更详细地描述在确定两个查询之间的关联性中使用的相关值$Cor(T_i,T_i')$的确定。具体地,图12是示出用于生成词语对或核心项T、T'之间的相关值的示例性过

程90的流程图。在方框92处，可以使用交互信息从语料库(例如，网络语料库)和用户查询中识别出新的核心项。在方框92的一个示意性实施中，如果Motorola是一个项，且“Motorola Announced”、“Motorola cell phone”、和“buy Motorola”以及“Nokia Announced”、“Nokia cell phone”、和“buy Nokia”处于语料库中，则Nokia也被识别为一个项。注意，尽管现用的字典能够提供传统的核心项，但许多新的核心项常常被引入到词汇表中。新核心项的实例包括恰当的名称(例如，人名和公司名)，以及多种其它新词和短语(例如，产品模型、电影、和歌曲名等)。

[0082] 在方框94处，例如，可以使用查询日志、网页和链接锚文本来确定核心项对T、T'之间的相关值。两个核心项T_1和T_2之间的相关性可以被定义为实数向量的函数：

[0083] $Cor(T_1,T_2)=f(w_1,w_2,\cdots,w_n)$

[0084] 其中，$w_1,w_2,\cdots,w_n$是某些预定关系的权重。预定关系的实例包括(1)同义词、缩写词和反义词，(2)复合短语，例如，上海对应上海市、电视对应电视机，(3)相同语法/语义类中的词语，例如，Toyota和Honda，(4)多义词语及其清楚的上下文，(5)人名及其相应的活动，例如，Oprah和现场访谈主持人，(6)词语的属性，例如，计算机和存储器，(7)概念的细分，例如，Amazon和Amazon River、Amazon Rain Forrest、和Amazon.com，(8)电影-演员、书-作者、公司-产品、人-职位等，例如，Tom Hanks和Forrest Gump、以及Bill Gates和CEO。

[0085] 在方框96中，相关向量Cor(T1,T2)的值可以被归一化为[0-1]。

[0086] 用于生成更改或细分用户输入的系统和方法可以建议多个查询，该查询可能被用户使用和/或生成用户可能选择的最佳结果。该系统和方法定量地测量两个查询之间的相关性。显然，两个查询不需要具有任何共同的词语或一致的同义词。例如，与原始查询(例如，汉语)有关的查询，对于歌曲“Now and Forever”的mp3文件“‘Now and Forever’mp3”，可以包括例如“CoCo Lee”(歌手)以及相同艺术家的其它歌曲或唱片集。因此，建议的查询可以不是简单的原始查询的扩展而是具有更好搜索结果(例如，用户最可能选择的搜索结果)的查询。在一个实例中，建议的查询可以包括实现消除了查询意义多义性的查询，其中，原始查询是简短且多义的。作为另一个实例，建议的查询可以包括将原始查询分成多个较短查询的查询，其中，原始查询可以是较长和/或包含彼此独立的词语。

[0087] 虽然本文描述和示出了本发明的多个示例性实施例，但应当理解，仅在不背离本发明的精神和范围内，对其进行示出和更改。因此，本发明的范围将仅根据下列可被修改的权利要求来限定，其中，每个权利要求都与作为本发明的一个实施例的本“具体实施方式”明确相结合。

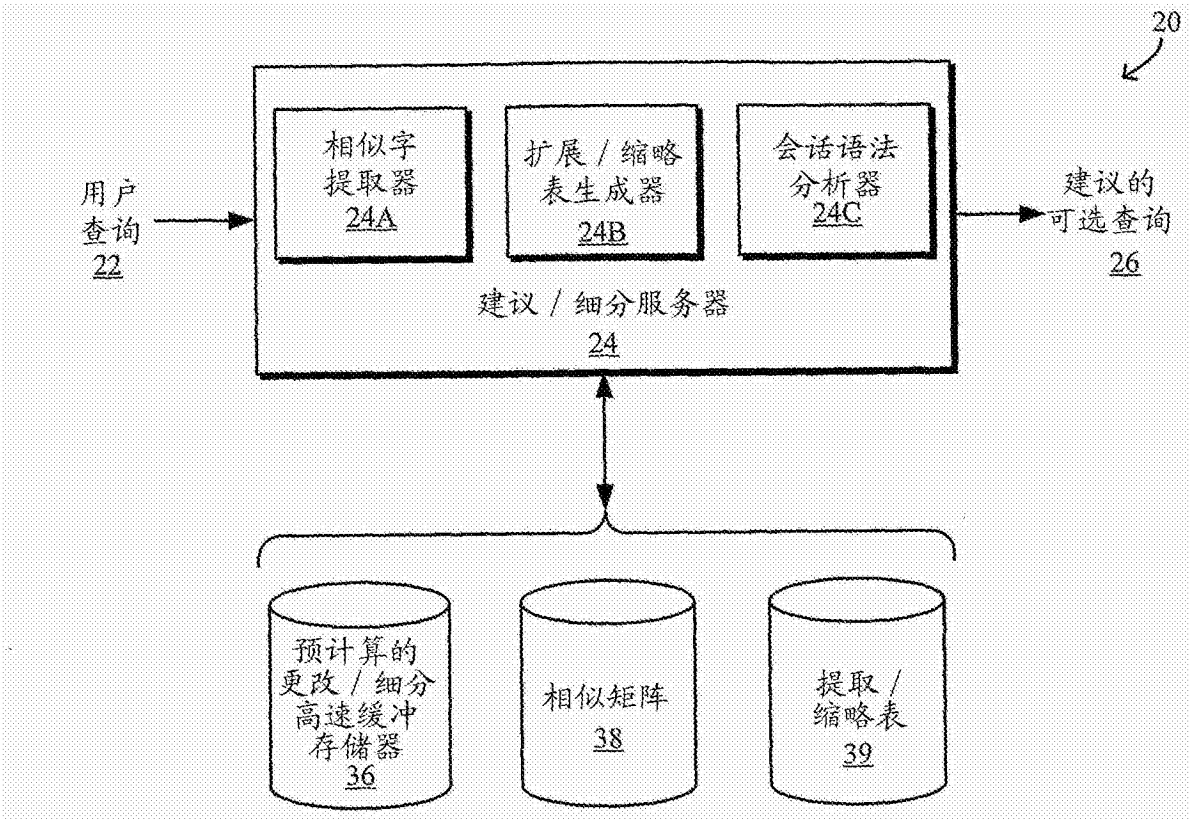

20
用户查询 22
相似字提取器 24A
扩展 / 缩略表生成器 24B
会话语法分析器 24C
建议的可选查询 26
建议 / 细分服务器 24
预计算的更改 / 细分高速缓冲存储器 36
相似矩阵 38
提取 / 缩略表 39

图1A

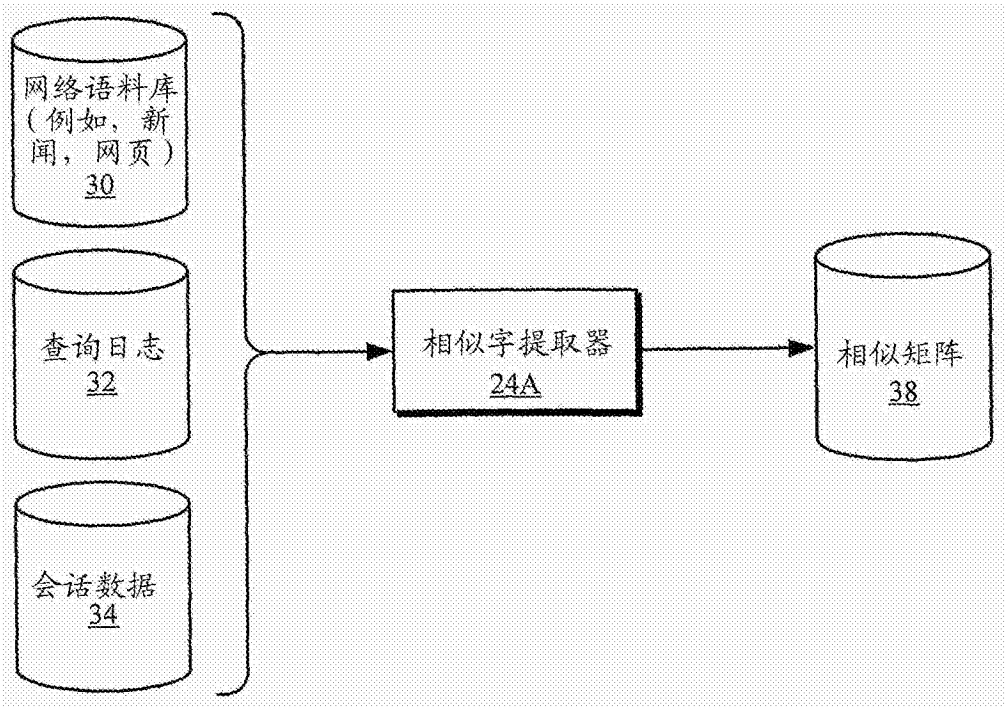

网络语料库（例如，新闻，网页）30
查询日志 32
会话数据 34
相似字提取器 24A
相似矩阵 38

图1B

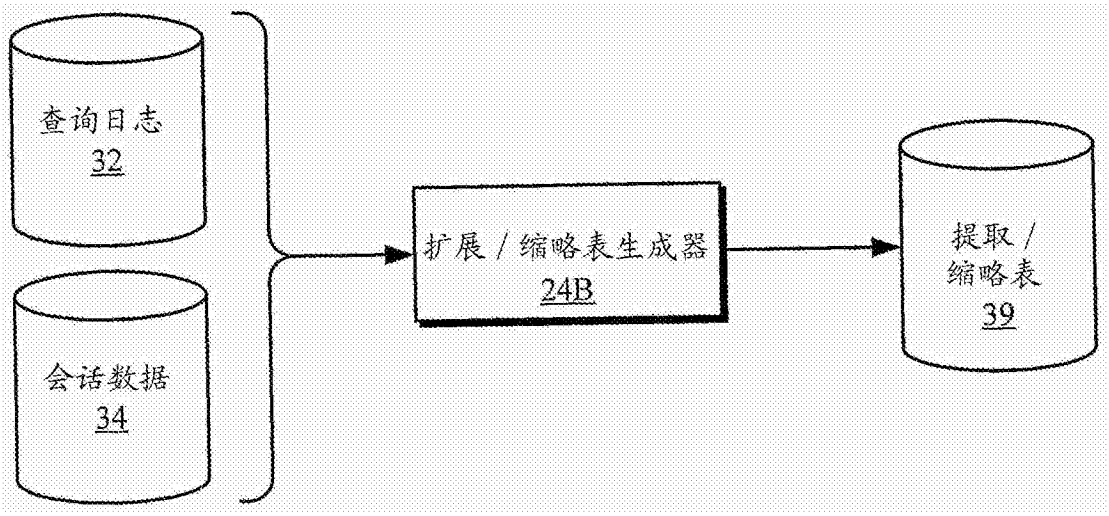

查询日志
32
会话数据
34
扩展 / 缩略表生成器
24B
提取 /
缩略表
39

图1C

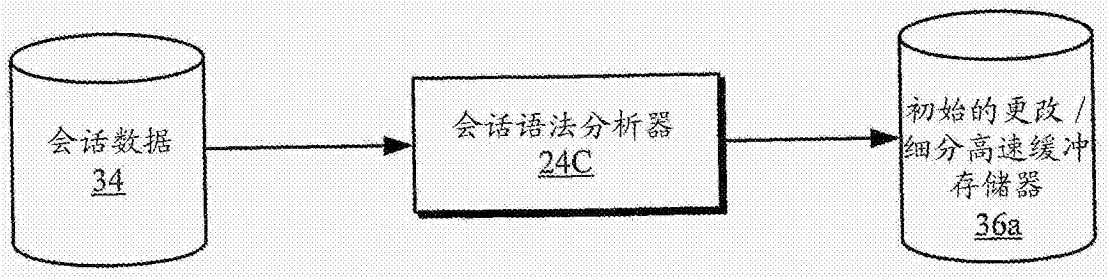

会话数据
34
会话语法分析器
24C
初始的更改 /
细分高速缓冲
存储器
36a

图1D

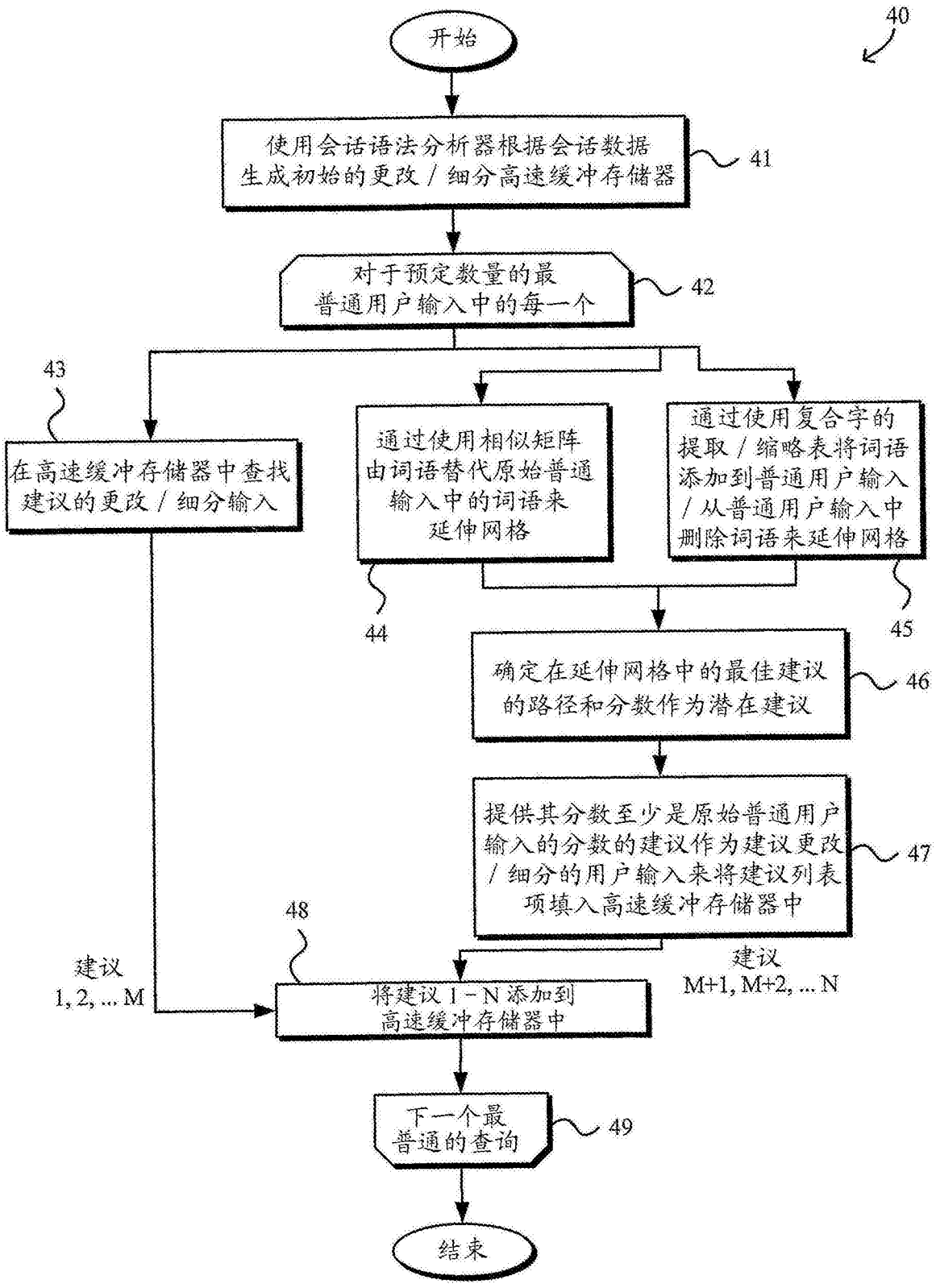

40
开始
使用会话语法分析器根据会话数据生成初始的更改 / 细分高速缓冲存储器
41
对于预定数量的最普通用户输入中的每一个
42
43
在高速缓冲存储器中查找建议的更改 / 细分输入
通过使用相似矩阵由词语替代原始普通输入中的词语来延伸网格
通过使用复合字的提取 / 缩略表将词语添加到普通用户输入 / 从普通用户输入中删除词语来延伸网格
44
45
确定在延伸网格中的最佳建议的路径和分数作为潜在建议
46
提供其分数至少是原始普通用户输入的分数的建议作为建议更改 / 细分的用户输入来将建议列表项填入高速缓冲存储器中
47
48
建议 1, 2, ... M
建议 M+1, M+2, ... N
将建议 1 - N 添加到高速缓冲存储器中
下一个最普通的查询
49
结束

图2A

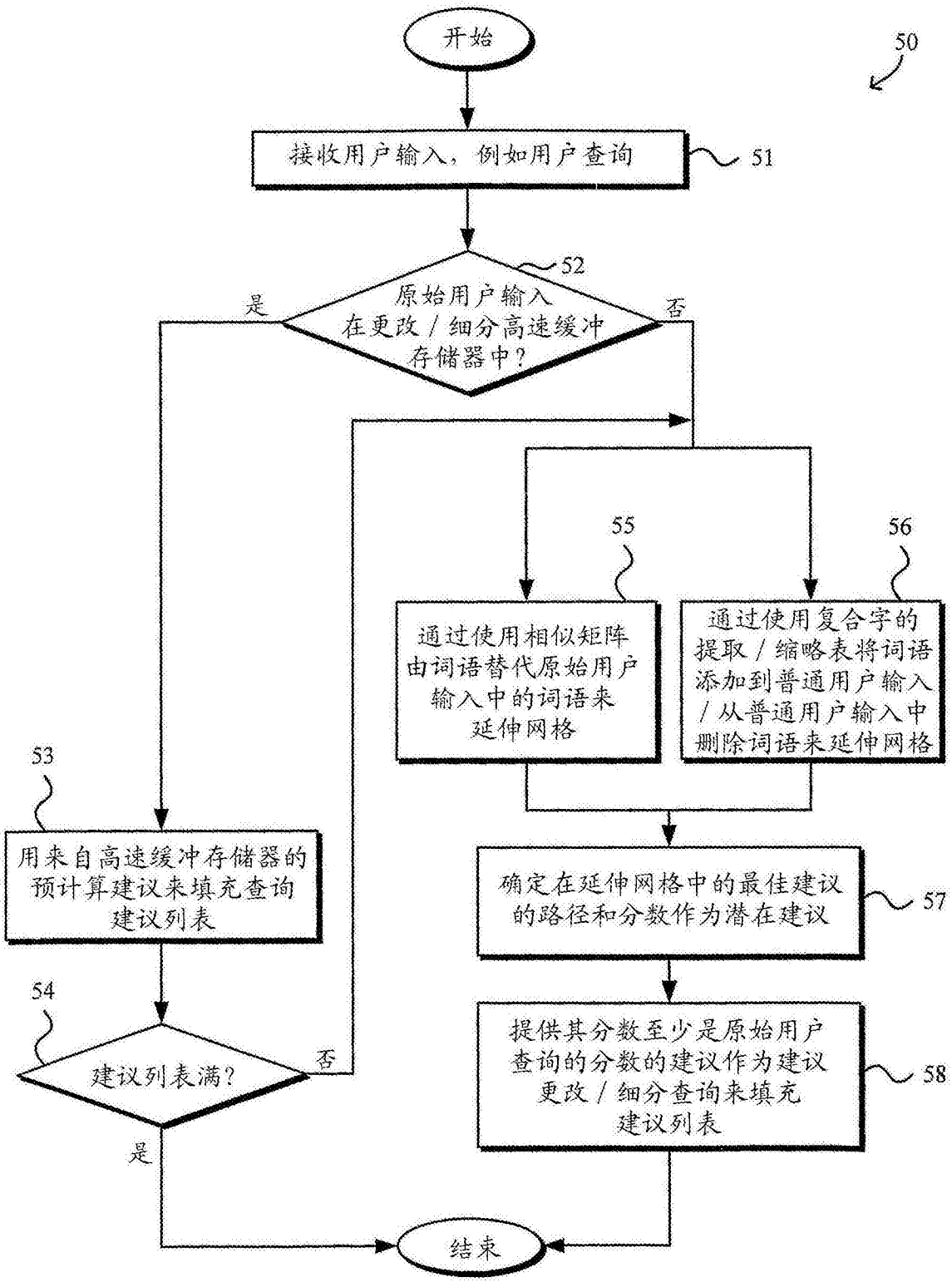

开始
50
接收用户输入，例如用户查询
51
52
原始用户输入在更改/细分高速缓冲存储器中？
是
否
55
56
通过使用相似矩阵由词语替代原始用户输入中的词语来延伸网格
通过使用复合字的提取/缩略表将词语添加到普通用户输入/从普通用户输入中删除词语来延伸网格
53
用来自高速缓冲存储器的预计算建议来填充查询建议列表
确定在延伸网格中的最佳建议的路径和分数作为潜在建议
57
54
建议列表满？
否
是
提供其分数至少是原始用户查询的分数的建议作为建议更改/细分查询来填充建议列表
58
结束

图2B

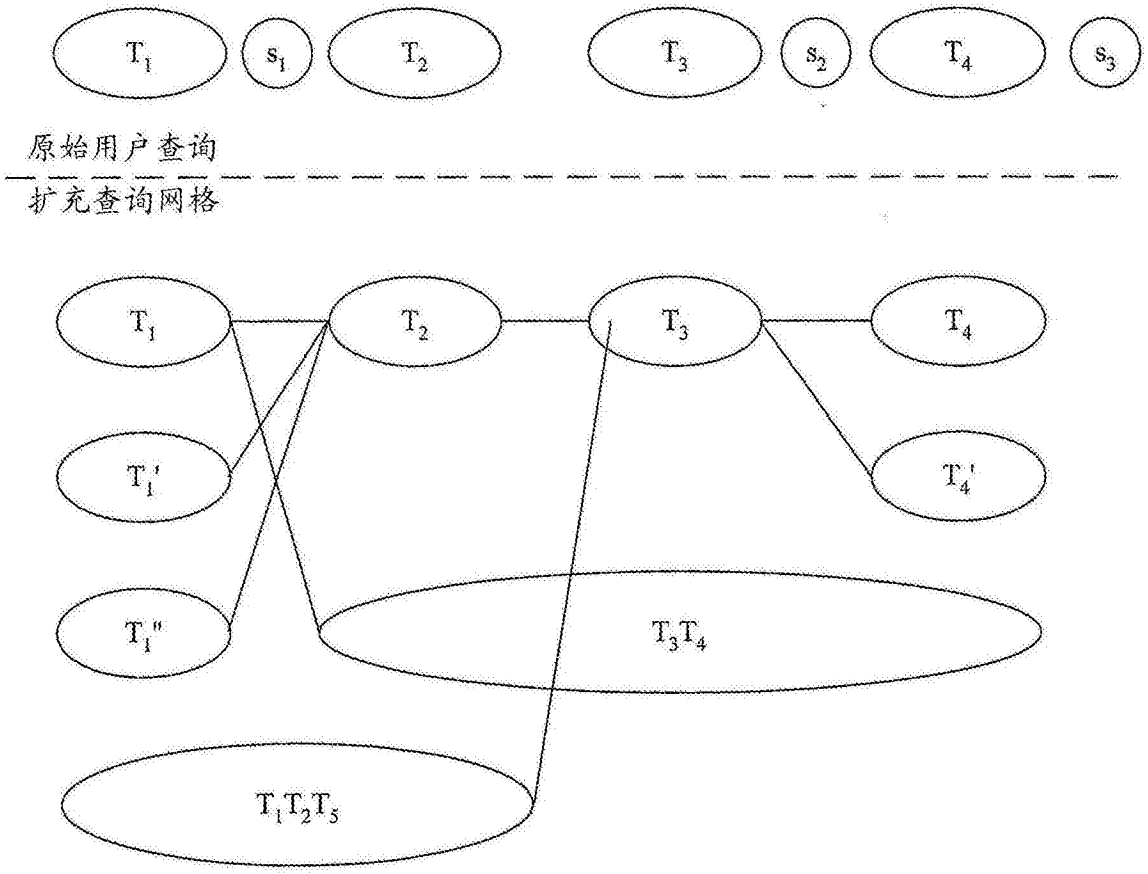

T1
s1
T2
T3
s2
T4
s3
原始用户查询
扩充查询网格
T1
T2
T3
T4
T1'
T4'
T1"
T3T4
T1T2T5

图3

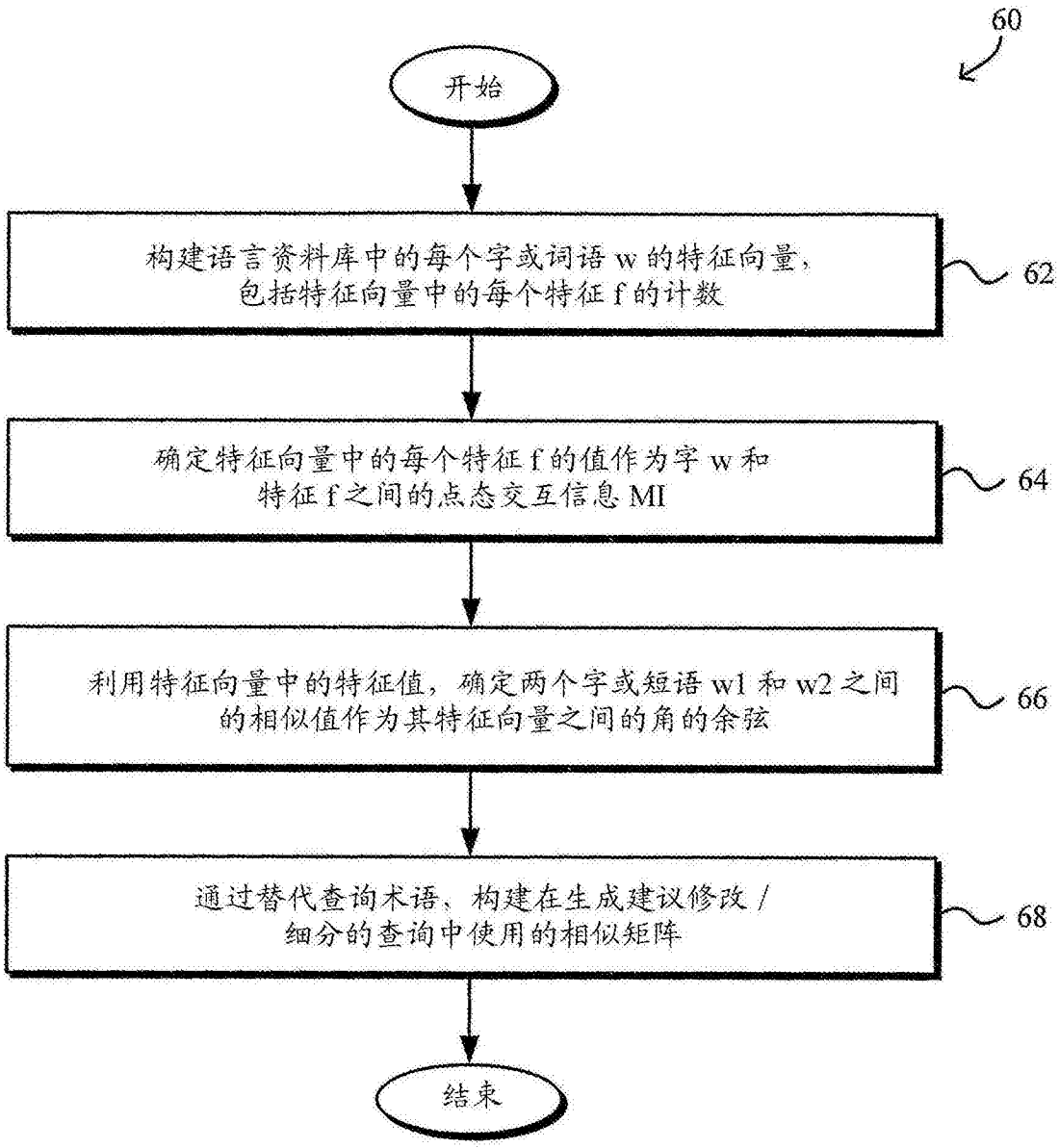

60
开始
构建语言资料库中的每个字或词语 w 的特征向量，包括特征向量中的每个特征 f 的计数
62
确定特征向量中的每个特征 f 的值作为字 w 和特征 f 之间的点态交互信息 MI
64
利用特征向量中的特征值，确定两个字或短语 w1 和 w2 之间的相似值作为其特征向量之间的角的余弦
66
通过替代查询术语、构建在生成建议修改 / 细分的查询中使用的相似矩阵
68
结束

图4

特征	计数
L:because	1
R:assess	1
L:rates	0.5
L:between	0.5
R:is	0.33
R:by	0.33
R:using	0.33

图5

特征	计数	MI
L:outlying	530	7.16619
L:disadvantaged	682.49	6.61504
L:marginalized	75.66	5.84836
L:close-knit	97.5	5.79445
L:enriching	123.5	5.78579
L:Mountain	2856	5.74945
L:Appalachian	63	5.29764
L:bedroom	498.5	5.08576
R:across	6285.07	4.83219
R:bordering	60	4.51492
R:clashed	84.5	4.25515
R:prosper	47.32	4.17204
R:impacted	70.57	3.64408
R:grapple	13.75	3.38906
R:namely	19	3.31465
R:encompassing	11	3.30112

图6

相似矩阵							
	W_1	W_2	W_3	W_4	W_5	…	W_n
W_1	---	S_{12}	S_{13}	S_{14}	S_{15}		S_{1n}
W_2		---	S_{23}	S_{24}	S_{25}		S_{2n}
W_3			---	S_{34}	S_{35}		S_{3n}
W_4				---	S_{45}		S_{4n}
W_5					---		S_{5n}
…						---	…
W_n							---

图7

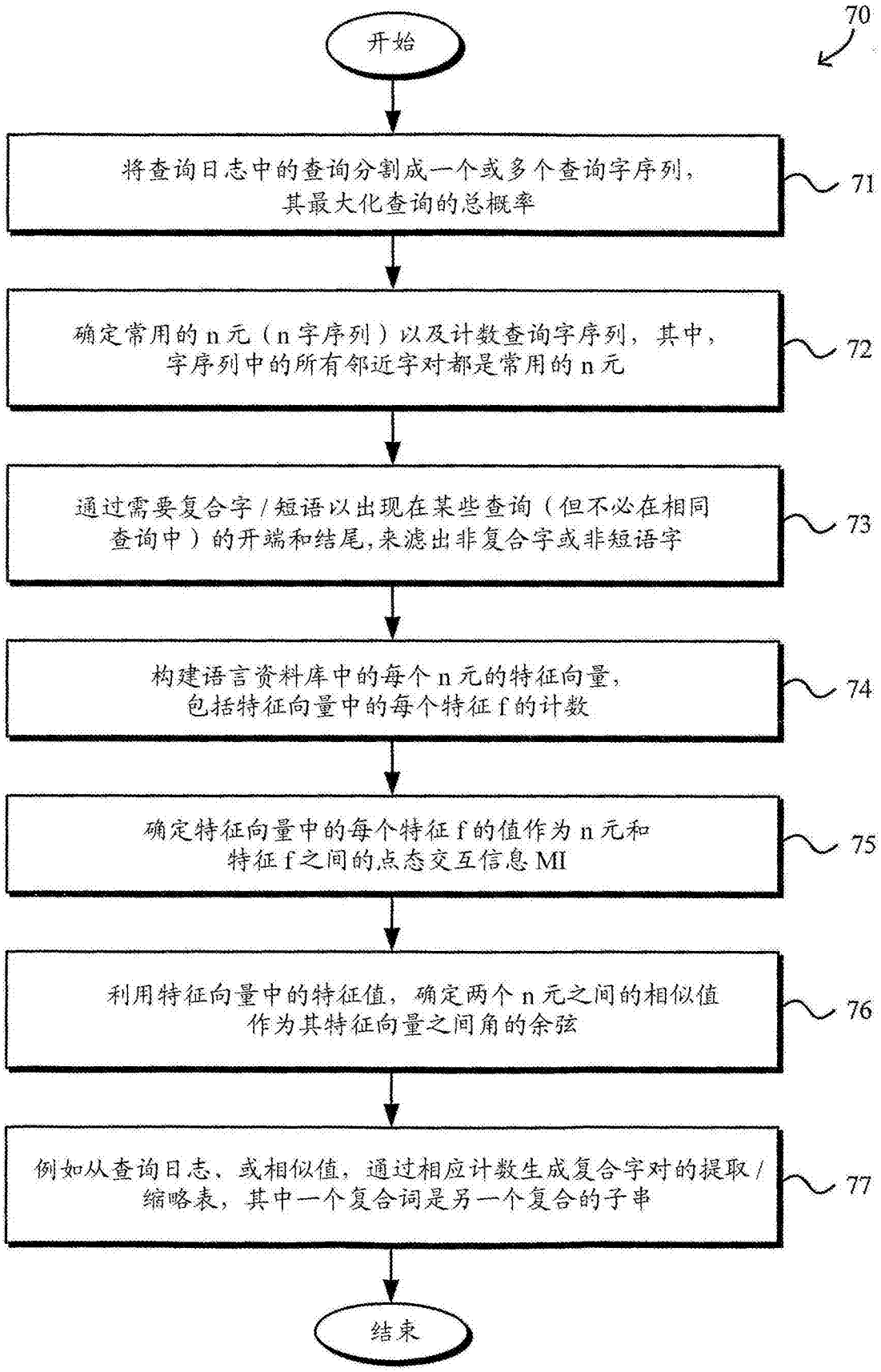

70
开始
将查询日志中的查询分割成一个或多个查询字序列，其最大化查询的总概率
71
确定常用的 n 元（n 字序列）以及计数查询字序列，其中，字序列中的所有邻近字对都是常用的 n 元
72
通过需要复合字 / 短语以出现在某些查询（但不必在相同查询中）的开端和结尾，来滤出非复合字或非短语字
73
构建语言资料库中的每个 n 元的特征向量，包括特征向量中的每个特征 f 的计数
74
确定特征向量中的每个特征 f 的值作为 n 元和特征 f 之间的点态交互信息 MI
75
利用特征向量中的特征值，确定两个 n 元之间的相似值作为其特征向量之间角的余弦
76
例如从查询日志，或相似值，通过相应计数生成复合字对的提取 / 缩略表，其中一个复合词是另一个复合的子串
77
结束

图8

扩展 / 缩略表			
复合字	计数	复合字	计数
c_1 c_2 c_3 c_4 c_5 c_6	56	c_1 c_2 c_3	6849
c_1 c_2 c_3 c_4 c_5 c_6	56	c_1 c_2 c_3 c_4	58
c_1 c_2 c_3 c_4 c_5 c_6	56	c_1 c_2 c_3 c_4 c_5 c_6 c_7 c_8	2
c_1 c_2 c_3 c_4 c_5 c_6	56	c_1 c_2 c_3 c_4 c_5 c_6 c_9 c_{10} c_{11}	3
c_1 c_2 c_3 c_4 c_5 c_6	56	c_4 c_5 c_6	2336

图9

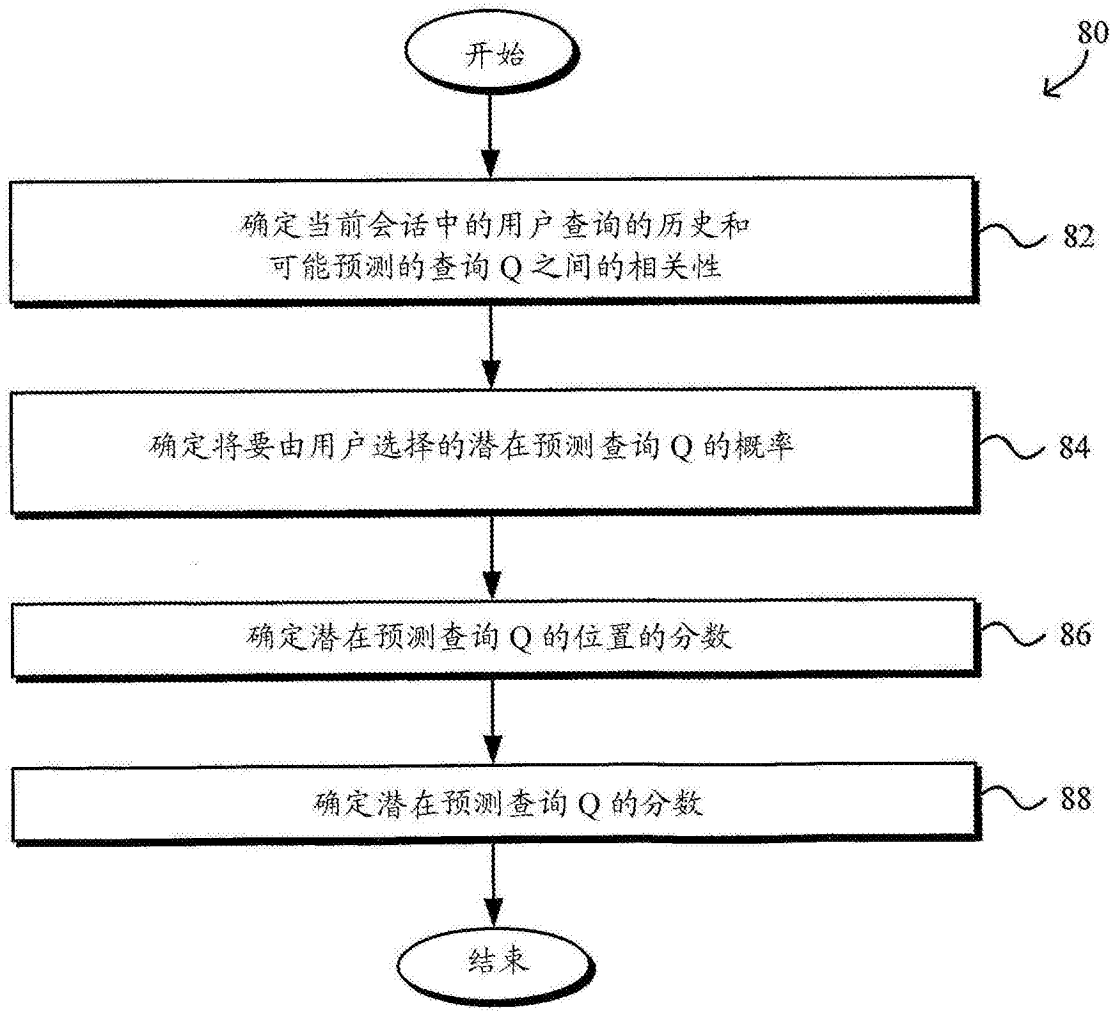


图10

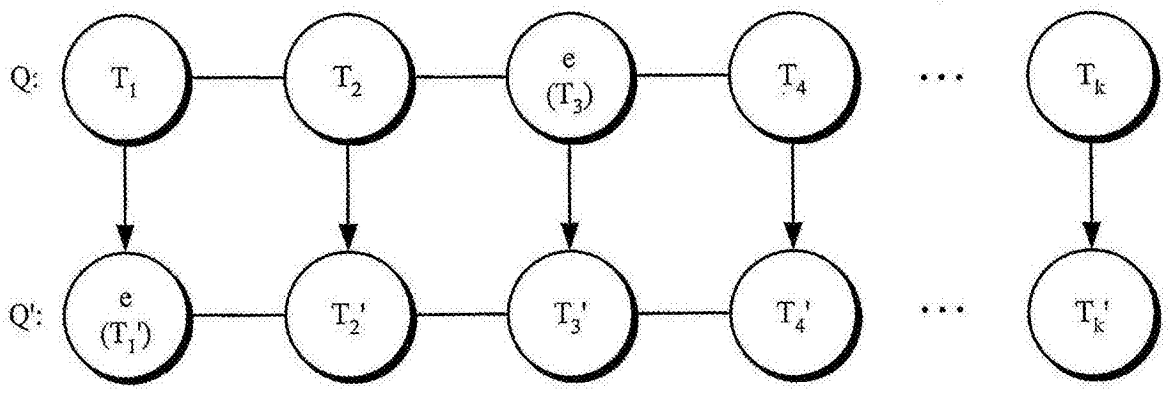

Q:
T1
T2
e
(T3)
T4
...
Tk
Q':
e
(T1')
T2'
T3'
T4'
...
Tk'

图11

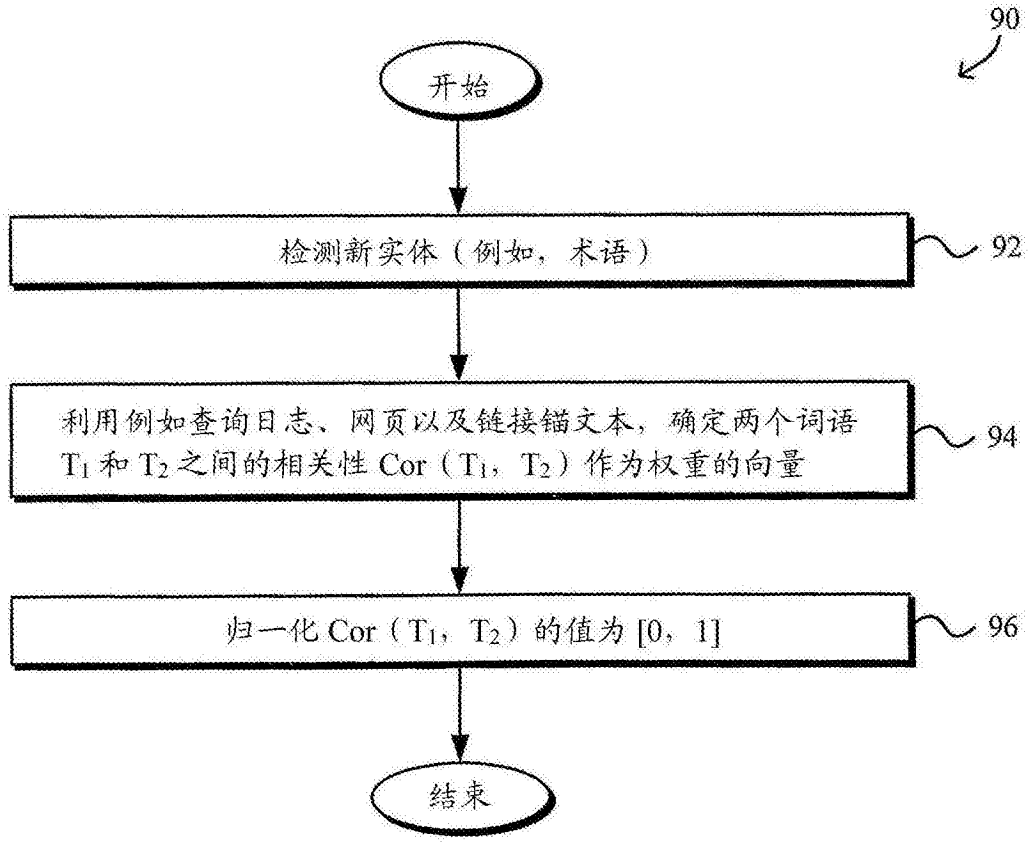

90
开始
检测新实体（例如，术语）
92
利用例如查询日志、网页以及链接锚文本，确定两个词语 T1 和 T2 之间的相关性 Cor（T1，T2）作为权重的向量
94
归一化 Cor（T1，T2）的值为 [0，1]
96
结束

图12