



(12) 发明专利申请

(10) 申请公布号 CN 104111972 A

(43) 申请公布日 2014. 10. 22

(21) 申请号 201410266074. X

(51) Int. Cl.

(22) 申请日 2009. 07. 20

G06F 17/30 (2006. 01)

(30) 优先权数据

61/082, 165 2008. 07. 18 US

12/503, 806 2009. 07. 15 US

(62) 分案原申请数据

200910164542. 1 2009. 07. 20

(71) 申请人 谷歌公司

地址 美国加利福尼亚州

(72) 发明人 皮尤什·普拉拉德卡

拉利特什·卡特拉嘎达

维内特·古普塔

(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

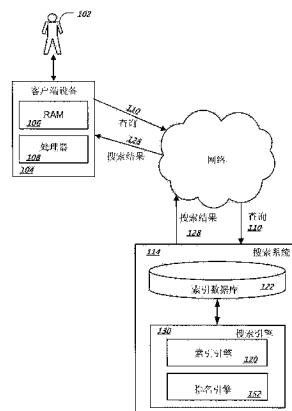
权利要求书3页 说明书11页 附图7页

(54) 发明名称

用于查询扩展的音译

(57) 摘要

本发明涉及用于识别用于查询扩展的音译词语的候选同义词的方法和系统。在一方面中，方法包括识别目标语言的多个音译词语。对于目标语言的多个音译词语中的每个音译词语，所述音译词语被映射到源语言的一个或多个词语。对于目标语言的多个音译词语中的第一音译词语，目标语言的多个音译词语中的一个或多个第二音译词语被识别为第一音译词语的候选同义词，其中所述一个或多个第二音译词语中的每一个被映射到也从所述第一音译词语映射的源语言的至少一个词语。



1. 一种识别用于查询扩展的音译词语的候选同义词的方法，包括：

使用一个或多个计算机识别目标语言的多个音译词语，所述多个音译词语中的每一个音译词语表示源语言的一个或多个相应源词语在所述目标语言中的转换；

对于所述目标语言的所述多个音译词语中的每个音译词语，将该音译词语映射到所述源语言的所述一个或多个相应源词语；

确定所述多个音译词语中的第一音译词语被映射到所述源语言的一个或多个特定源词语并且所述多个音译词语中的一个或多个第二音译词语也被映射到所述源语言的所述一个或多个特定源词语；以及

基于所述第一音译词语和所述一个或多个第二音译词语都被映射到所述源语言的所述一个或多个特定源词语，将所述一个或多个第二音译词语识别为所述第一音译词语的候选同义词。

2. 根据权利要求 1 所述的方法，其中将该音译词语映射到所述源语言的所述一个或多个相应源词语进一步包括：

将所述目标语言的该音译词语音译为所述源语言的所述一个或多个相应源词语。

3. 根据权利要求 2 所述的方法，其中被识别为所述第一音译词语的候选同义词的所述一个或多个第二音译词语中的每一个第二音译词语具有超过指定阈值的相对于所述第一音译词语的置信值。

4. 根据权利要求 3 所述的方法，其中所述一个或多个第二音译词语中的每一个第二音译词语的置信值是从所述第一音译词语和该第二音译词语两者映射的所述源语言的所述一个或多个特定源词语的数目的函数。

5. 根据权利要求 3 所述的方法，其中将所述目标语言的该音译词语音译为所述源语言的所述一个或多个相应源词语进一步包括：

产生用于所述目标语言的该音译词语到所述源语言的所述一个或多个相应源词语的每一个音译的音译分值。

6. 根据权利要求 5 所述的方法，其中所述一个或多个第二音译词语中的每一个第二音译词语的置信值是 web 资源中该第二音译词语的出现概率、用于该第二音译词语到所述一个或多个特定源词语的音译的音译分值、以及用于所述第一音译词语到所述一个或多个特定源词语的音译的音译分值中的一个或多个的函数。

7. 根据权利要求 1 所述的方法，其中识别所述目标语言的所述多个音译词语进一步包括：

识别仅包含所述目标语言的字符的词语。

8. 根据权利要求 7 所述的方法，进一步包括：

计算用于每一个所识别的仅包含所述目标语言的字符的词语的统计量；

将用于每一个所识别的词语的所述统计量与指定的阈值进行比较；以及

如果用于特定的所识别的词语的统计量超过所指定的阈值，则将所述特定的所识别的词语包括在所述目标语言的所述多个音译词语中。

9. 根据权利要求 1 所述的方法，进一步包括：

对于所述目标语言的所述多个音译词语中的所述第一音译词语，将从所述第一音译词语映射的和从所述一个或多个第二音译词语映射的所述源语言的所述一个或多个特定源

词语识别为所述第一音译词语的候选同义词。

10. 根据权利要求 1 所述的方法,进一步包括:

接收包括所述第一音译词语的查询;

利用所述第一音译词语的所述候选同义词中的一个或多个扩展所述查询;

将所述扩展的查询提供到搜索引擎;以及

接收用于所述扩展的查询的搜索结果。

11. 根据权利要求 1 所述的方法,进一步包括:

接收包括所述第一音译词语的查询;以及

提供一个或多个扩展的查询供用户选择,每一个扩展的查询包括所述查询以及所述第一音译词语的所述候选同义词中的一个或多个。

12. 根据权利要求 1 所述的方法,进一步包括:

接收包括所述第一音译词语的查询;

将所述查询提供到搜索引擎,其中所述搜索引擎识别以下的 web 资源作为用于所述查询的可能的搜索结果,所述 web 资源包括所述第一音译词语的所述候选同义词中的至少一个但是不包括所述查询中的任何词语;以及

修改与所述 web 资源相关联的分值,所述分值用于在对用于所述查询的可能的搜索结果进行排名中使用。

13. 根据权利要求 1 所述的方法,进一步包括:

接收包括所述第一音译词语的查询;

将所述查询提供到搜索引擎,其中所述搜索引擎识别以下的 web 资源作为用于所述查询的可能的搜索结果,所述 web 资源包括从所述第一音译词语映射的和从所述一个或多个第二音译词语映射的所述源语言的所述一个或多个特定源词语中的至少一个但是不包括所述查询中的任何词语;以及

修改与所述 web 资源相关联的信息检索分值,所述信息检索分值用于在对用于所述查询的可能的搜索结果进行排名中使用。

14. 一种识别用于查询扩展的音译词语的候选同义词的系统,包括:

用于使用一个或多个计算机识别目标语言的多个音译词语的装置,所述多个音译词语中的每一个音译词语表示源语言的一个或多个相应源词语在所述目标语言中的转换;

用于对于所述目标语言的所述多个音译词语中的每个音译词语,将该音译词语映射到所述源语言的所述一个或多个相应源词语的装置;

用于确定所述多个音译词语中的第一音译词语被映射到所述源语言的一个或多个特定源词语并且所述多个音译词语中的一个或多个第二音译词语也被映射到所述源语言的所述一个或多个特定源词语的装置;以及

用于基于所述第一音译词语和所述一个或多个第二音译词语都被映射到所述源语言的所述一个或多个特定源词语,将所述一个或多个第二音译词语识别为所述第一音译词语的候选同义词的装置。

15. 根据权利要求 14 所述的系统,其中用于将该音译词语映射到所述源语言的所述一个或多个相应源词语的所述装置进一步包括:

用于将所述目标语言的该音译词语音译为所述源语言的所述一个或多个相应源词语

的装置。

16. 根据权利要求 15 所述的系统, 其中被识别为所述第一音译词语的候选同义词的所述一个或多个第二音译词语中的每一个第二音译词语具有超过指定阈值的相对于所述第一音译词语的置信值。

17. 根据权利要求 16 所述的系统, 其中所述一个或多个第二音译词语中的每一个第二音译词语的置信值是从所述第一音译词语和该第二音译词语两者映射的所述源语言的所述一个或多个特定源词语的数目的函数。

18. 根据权利要求 16 所述的系统, 其中用于将所述目标语言的该音译词语音译为所述源语言的所述一个或多个相应源词语的所述装置进一步包括 :

用于产生用于所述目标语言的该音译词语到所述源语言的所述一个或多个相应源词语的每一个音译的音译分值的装置。

19. 根据权利要求 18 所述的系统, 其中所述一个或多个第二音译词语中的每一个第二音译词语的置信值是 web 资源中该第二音译词语的出现概率、用于该第二音译词语到所述一个或多个特定源词语的音译的音译分值、以及用于所述第一音译词语到所述一个或多个特定源词语的音译的音译分值中的一个或多个的函数。

20. 根据权利要求 14 所述的系统, 其中用于识别所述目标语言的所述多个音译词语的所述装置进一步包括 :

用于识别仅包含所述目标语言的字符的词语的装置。

21. 根据权利要求 20 所述的系统, 进一步包括 :

用于计算用于每一个所识别的仅包含所述目标语言的字符的词语的统计量的装置 ;

用于将用于每一个所识别的词语的所述统计量与指定的阈值进行比较的装置 ; 以及

用于如果用于特定的所识别的词语的统计量超过所指定的阈值, 则将所述特定的所识别的词语包括在所述目标语言的所述多个音译词语中的装置。

22. 根据权利要求 14 所述的系统, 进一步包括 :

用于对于所述目标语言的所述多个音译词语中的所述第一音译词语, 将从所述第一音译词语映射的和从所述一个或多个第二音译词语映射的所述源语言的所述一个或多个特定源词语识别为所述第一音译词语的候选同义词的装置。

## 用于查询扩展的音译

[0001] 分案说明

[0002] 本申请属于申请日为 2009 年 7 月 20 日的中国专利申请 200910164542.1 的分案申请。

### 技术领域

[0003] 本说明涉及用于用户向搜索引擎提交查询的查询扩展。

### 背景技术

[0004] 搜索引擎 – 以及,特别地,因特网搜索引擎 – 的目标在于识别与用户的需求相关的资源(例如,网页、图像、文本文档、多媒体内容(context))以及将与资源有关的信息以对用户最有用的方式进行呈现。因特网搜索引擎响应于用户提交的查询返回搜索结果。如果用户对于为查询返回的搜索结果不满意,那么用户能够尝试精化所述查询以更好地匹配用户的需求。

[0005] 一些搜索引擎为用户提供搜索引擎识别与用户的查询相关的建议的替选查询,例如扩展的查询。用于找到用于查询扩展的查询词的同义词的技术通常依赖于自然语言模型或者用户搜索日志数据。识别出的查询词的同义词能够在识别附加的或者更相关的资源的尝试中用于扩展查询以改进用户搜索体验。

[0006] 电子文档通常用多种不同语言书写。通常在特定的书写系统(即文字(script))中表达每一种语言,所述书写系统的特征通常在于特定的字母表。例如,使用拉丁字母表来表达英语语言,而使用梵文字母表来表达印度语语言。一些语言所使用的文字包括已经被扩展为包括附加的标记或者字符的特定字母表。在音译(transliteration)中,一种语言的文字被用于表示通常以另一种语言的文字书写的词。例如,音译词语能够是从一种文字转换成另一种文字的词语或者以一种文字的词语的另一种文字的语音表示。用于找到用于查询扩展的查询词的同义词的技术可能对于找到音译词语的查询词语的同义词不能很好的工作。例如,当前的自然语言技术对于音译数据不能很好的工作,并且搜寻日志数据通常不能很好地覆盖大多数音译的变体。

### 发明内容

[0007] 本发明描述了涉及识别用于查询扩展的音译词语的候选同义词的技术。

[0008] 一般来说,在本说明中描述的主题的一方面能够被具体化为计算机实现的方法,该方法包括下述动作:使用一个或者多个计算机识别目标语言的多个音译词语;对于目标语言的多个音译词语中的每一个音译词语,将音译词语映射到源语言的一个或者多个词语;以及对于目标语言的多个音译词语中的第一音译词语,识别目标语言的多个音译词语的一个或多个第二音译词语作为第一音译词语的候选同义词,其中所述一个或者多个第二音译词语中的每一个被映射到也从第一音译词语映射的源语言的至少一个词语。该方面的其它实施例包括对应的系统、装置以及计算机程序产品。

[0009] 这些和其它实施例能够可选地包括下述特征中的一个或多个。识别目标语言的多个音译词语能够进一步包括从 web 资源识别只包含目标语言的字符的词语。该方面能够进一步包括计算用于仅包含目标语言的字符的每个识别的词语的统计量, 将用于每个识别的词语的统计量与指定的阈值进行比较, 并且如果用于特定的识别词语的统计量超过指定的阈值, 则将特定的识别的词语包括在目标语言的多个音译词语中。

[0010] 用于每个识别的词语的统计量能够是与讲源语言的一个或多个地区 (locale) 相关联的顶级域的 web 资源中识别的词语的出现概率相对于与任何地区相关联的顶级域的 web 资源中识别的词语的出现概率的比率。用于每个识别的词语的统计量能够是与讲源语言的一个或多个地区相关联的 web 资源中识别的词语的出现概率相对于与任何地区相关联的 web 资源中识别的词语的出现概率的比率。web 资源与讲源语言的地区的关联能够通过 web 资源的顶级域来确定。

[0011] 将音译词语映射到源语言的一个或多个词语能够进一步包括将目标语言的音译词语音译为源语言的一个或多个词语。被识别为第一音译词语的候选同义词的一个或多个第二音译词语中的每一个能够具有超过指定的阈值的相对于第一音译词语的置信 (confidence) 值。第二音译词语的置信值能够是从第一音译词语和第二音译词语两者映射的源语言的词语的数目的函数。将目标语言的音译词语音译为源语言的词语能够进一步包括产生用于目标语言的音译词语到源语言的词语的音译的音译分值。第二音译词语的置信值能够是 web 资源中第二音译词语的出现概率、用于第二音译词语到也被从第一音译词语映射的源语言的词语的音译的音译分值、以及用于第一音译词语到源语言的词语的音译的音译分值中的一个或多个的函数。

[0012] 所述方面能够进一步包括, 对于目标语言的多个音译词语的第一音译词语, 识别从第一音译词语映射的以及从一个或多个第二音译词语中的至少一个映射的源语言的一个或多个词语作为第一音译词语的候选同义词。该方面能够进一步包括接收包括第一音译词语的查询, 用第一音译词语的候选同义词中的一个或多个扩展该查询, 将扩展的查询提供给搜索引擎, 并且接收用于扩展的查询的搜索结果。该方面能够进一步包括接收包括第一音译词语的查询, 以及提供一个或多个扩展的查询用于供用户选择, 每个扩展的查询包括所述查询以及第一音译词语的候选同义词中的一个或多个。

[0013] 该方面能够进一步包括接收包括第一音译词语的查询; 将该查询提供给搜索引擎, 其中所述搜索引擎识别以下的 web 资源作为用于该查询的可能的搜索结果, 所述 web 资源包括第一音译词语的候选同义词中的至少一个但是不包括查询中的任何词语; 以及修改与所述 web 资源相关联的分值, 所述分值用于在排名用于所述查询的可能的搜索结果中使用。该方面能够进一步包括接收包括第一音译词语的查询; 将该查询提供给搜索引擎, 其中所述搜索引擎识别以下的 web 资源作为用于该查询的可能的搜索结果, 所述 web 资源包括从第一音译词语映射的以及从一个或多个第二音译词语中的至少一个映射的源语言的词语中的至少一个, 但是不包括查询中的任何词语; 以及修改与该 web 资源相关联的信息检索分值, 所述信息检索分值用于在排名用于该查询的可能的搜索结果中使用。

[0014] 在本说明中描述的主题的另一方面能够具体化为计算机实现的方法, 所述方法包括下述动作: 使用一个或多个计算机产生用于目标语言的可能的音译同义词的训练组; 使用训练组来训练概率模型以学习音译同义词在目标语言中的拼写变体的概率; 以及将概率

模型应用于目标语言的特定音译词语以识别特定音译词语的一个或多个候选同义词。该方面的其它实施例包括对应的系统、装置以及计算机程序产品。

[0015] 本说明中描述的主题的另一方面能够被具体化为计算机实现的方法，所述方法包括下述动作：使用一个或多个计算机识别目标语言的多个音译词语；对于目标语言的多个音译词语的第一音译词语，识别目标语言的多个音译词语的一个或多个第二音译词语作为第一音译词语的候选同义词；以及使用第一音译词语的候选同义词来扩展包括第一音译词语的查询。该方面的其它实施例包括对应的系统、装置以及计算机程序产品。

[0016] 本说明中描述的主题的特定实施例能够被实现为实现下面优点中的一个或多个。音译词语被识别为用于特定音译词语的候选同义词，其中所述候选同义词能够被用于扩展包括特定音译词语的查询。能够为较新的音译词语（例如从源语言的词语音译词语、从当前新闻故事或当前文化参考而音译词语）识别音译目标语言的同义词，这可能在用户搜索日志数据中具有较差的覆盖。能够将用户的查询扩展为包括用于给定的音译词语的候选音译同义词的系统可以返回比不具有这样的查询扩展能力的搜索系统更好的搜索结果。

[0017] 在附图以及下面的描述中阐述了本说明中描述的主题的一个或多个实施例的细节。根据说明书、附图以及权利要求，主题的其它特征、目标以及优点将更加明显。

## 附图说明

[0018] 图 1 是示例搜索系统的框图。

[0019] 图 2A 至 2C 示出用于识别音译词语的候选同义词的示例技术。

[0020] 图 3 是用于识别音译词语的候选同义词的示例过程的流程图。

[0021] 图 4 是用于提供用于包括音译词语和候选同义词的扩展的查询的搜索结果的示例过程的流程图。

[0022] 图 5 是用于识别音译词语的候选同义词的示例过程的流程图。

[0023] 各附图中的相同的附图符号和标记表示相同的元素。

## 具体实施方式

[0024] 图 1 是如能够在因特网、内联网 (intranet)、或者另外的客户端与服务器环境中实现的示例搜索系统 114 的框图，该搜索系统 114 能够用于提供与提交的查询相关的搜索结果。搜索系统 114 是信息检索系统的示例，在该信息检索系统中能够实现下面所述的系统、组件以及技术。

[0025] 用户 102 能够通过客户端设备 104 与搜索系统 114 交互。例如，客户端 104 能够是通过局域网 (LAN) 或者例如因特网的广域网 (WAN) 镶接到搜索系统 114 的计算机。在一些实现方式中，搜索系统 114 和客户端设备 104 能够是一台机器。例如，用户能够将桌面搜索应用安装在客户端设备 104 上。客户端设备 104 通常将包括随机存取存储器 (RAM) 106 和处理器 108。

[0026] 用户 102 能够将查询 110 提交到搜索系统 114 中的搜索引擎 130。当用户 102 提交查询 110 时，查询 110 被通过网络发送到搜索系统 114。搜索系统 114 能够被实现为例如运行在一个或多个位置中通过网络彼此连接的一个或多个计算机上的计算机程序。搜索系统 114 包括索引数据库 122 以及搜索引擎 130。搜索系统 114 通过生成搜索结果 128 来响

应查询 110，该搜索结果 128 被通过网络以能够呈现给用户 102 的形式（例如，作为要在客户端 104 上运行的 web 浏览器中显示的搜索结果网页）发送到客户端设备 104。

[0027] 当搜索引擎 130 接收到查询 110 时，搜索引擎 130 识别匹配查询 110 的资源。搜索引擎 130 通常包括索引资源（例如，因特网上的网页、图像、或者新闻文章）的索引引擎 120、存储索引信息的索引数据库 122、以及排名匹配查询 110 的资源的排名引擎 152（或其他软件）。搜索引擎 130 能够通过网络将搜索结果 128 发送到客户端设备 104 以呈现给用户 102。

[0028] 在一些方案中，查询包括是音译词语的一个或多个词语。音译将源语言的词语转换为目标语言的音译词语。在转换之后，通过目标语言的字母或字符表示源语言的词语的字母或者字符。例如，在标题为“Machine Learning for Transliteration（用于音译的机器学习）”、于 2008 年 3 月 6 日提交的美国专利申请 No. 12/043854 中描述了用于音译的机器学习技术。

[0029] 从一种语言音译为另一种语言的词语能够在因特网资源中使用。例如，在因特网资源（例如，印度博客或者电子印度技术教科书）上像印度语、泰米尔语、泰卢固语、埃纳德语和马拉雅拉姆语的印度语言有时被音译为英语。这些语言和一些非印度语言（例如，中文或者其他语标书写系统）常常没有开发得很好的替选输入机制，从而输入这些语言的字符是很麻烦的。

[0030] 音译没有校正拼写的概念。结果，对于源语言的词的音译常常存在目标语言的多个拼写。对于具有目标语言的多个音译源语言的特定词语，从目标语言的给定的音译词语变化的目标语言的音译词语能够被处理为给定的音译词语的候选同义词。这些候选的音译同义词是源语言的同一词语的不同可能的音译。

[0031] 作为示例，印度语词 **চক্ৰবৰ্তী** 能够被音译为英语“chakrabarti”或者“chakrabarty”。因此，音译词语“chakrabarty”能够被识别为给定的音译词语“chakrabarti”的候选同义词。

[0032] 对给定的音译词语识别的候选同义词能够用于扩展包括给定的音译词语的查询。例如，如果在因特网上若干网站上存在可用的流行的新的印度语歌曲，那么如果网站将歌曲标题的印度语词音译为第一音译词语而用户输入带有用于同一印度语词的第二音译词语的查询，那么用户会发现很难查找到该歌曲。能够将用户的查询扩展为包括第二音译词语的候选音译同义词的搜索系统可以返回比不具有相同查询扩展能力的搜索系统更好的搜索结果。

[0033] 图 2A 至 2C 示出用于识别音译词语的候选同义词的示例技术。为了方便，将参考执行该技术的系统描述该示例技术。该示例技术能够用于将包括音译词语的查询扩展以在尝试改进为该查询返回的搜索结果中包括音译词语的同义词。该示例技术使用音译技术来确定目标语言（例如，英语）的哪些词语是从源语言（例如，印度语）的同一词语音译的。若干技术能够被实现为增加候选同义词的精度或者质量。

[0034] 图 2A 示出目标语言英语的可能的音译词语的列表 210，其中源语言是印度语。系统能够以任何数目的不同方式产生或者识别可能的音译词语的列表 210。

[0035] 例如，系统能够识别来自 web 资源的列表 210 的可能的音译词语作为仅包含目标语言的字符（例如，拉丁字符）的词语。仅包含目标语言的字符的识别的词语包括在目标

语言中有意义的词和在目标语言中没有意义的可能的音译词语。

[0036] 为了从非音译词语（例如有意义的词）分离可能的音译词语，系统能够计算仅包含目标语言的字符的识别的词语的统计量并且能够将该统计量与指定的阈值比较。即，对于每个识别的词语，计算统计量并且将该统计量与阈值比较，其中如果识别的词语的统计量超过了指定的阈值则该系统将识别的词语包括在可能的音译词语的列表 210 中。

[0037] 在英语是目标语言并且印度语是源语言的一个示例中，英语的音译词语在印度语的 web 资源上比在非印度语的 web 资源上具有更高的出现概率。在该示例中，仅包含拉丁字符的每个识别的词语的统计量能够是印度语的 web 资源上的出现概率的函数。

[0038] 在一些实现方式中，用于每个识别的词语的统计量是与讲源语言的一个或多个地区（例如，国家或区域）相关联的顶级域的 web 资源中识别的词语的出现概率相对于与任何地区相关联的顶级域的 web 资源中识别的词语的出现概率的比率。例如，统计量能够是印度语网页上出现识别的词语的概率相对于任何网页上出现识别的词语的概率的比率。如果计算的用于特定识别的词语的统计量超过指定的阈值，则特定的识别词语能够被包括在可能的音译词语的列表 210 中。

[0039] 在一些其他的实现方式中，用于每个识别的词语的统计量是与讲源语言的一个或多个地区（例如，国家或者区域）相关联的 web 资源中识别的词语的出现概率相对于与任何地区相关联的 web 资源中识别的词语的出现概率的比率。web 资源与讲源语言的地区的关联能够通过 web 资源的顶级域来确定。例如，统计量能够是在印度语的 web 域上出现的识别的词语的概率相对于在任何 web 域上出现的识别的词语的概率的比率。如果为特定的识别的词语计算的统计量超过了指定的阈值，那么特定的识别的词语能够被包括在可能的音译词语的列表 210 中。

[0040] 在一些方案中，特定的网页或者特定的 web 域可能使用特定的识别的词语非常多次，这可能使用于特定的识别的词语的统计量歪斜 (skew)。在一些实施方式中，系统以指定的限制为用于每个识别的词语的统计量或者用于每个识别的词语的统计量的分量设定上限以防止使统计量歪斜。例如，系统能够给印度语的网页上的识别的词语的每页的贡献或者印度语的域上识别的词语的每域的贡献设定上限。

[0041] 在一些实施方式中，用于每个识别的词语的统计量是被包括在提交到具有源语言的界面的搜索引擎的查询中的识别的词语的概率相对于被包括在提交到具有任何语言的界面的搜索引擎的查询中的识别的词语的概率的比率。例如，系统能够使用印度语和非印度语搜索日志来计算统计量。

[0042] 在一些实施方式中，为了从非音译词语（例如，目标语言中有意义的词）分离可能的音译词语，系统计算用于每个仅包含目标语言的字符的识别的词语的多个统计量并且将该多个统计量与各阈值进行比较。如果用于特定的识别的词语的多个统计量每个都超过了相应阈值，那么系统能够将特定的识别的词语包括在可能的音译词语的列表 210 中。

[0043] 列表 210 的可能的音译词语能够替选地通过只爬行 (crawl) 已知的与源语言相关联的 web 资源来进行识别。对于源语言是印度语的示例，系统能够通过爬行已知的印度语网站，例如印度语博客站点或者翻译印度语歌曲或者印度语技术教科书的网站，来识别可能的音译词语。

[0044] 图 2B 示出列表 210 的每个可能的音译词语与源语言印度语的一个或多个词语 220

之间的关系 215。每个关系 215 是将第一组（即目标语言的可能的音译词语）中的元素映射到第二组（即源语言的词语 220）的一个或多个元素的结果。即，映射形成可能的目标语言的音译词语与源语言的一个或多个词语 220 之间的单向关系。在图 2B 的示例技术中，关系 215 是通过例如由被实现为系统的元素的英语到印度语机器音译器执行的音译映射的结果。

[0045] 在一些实施方式中，映射包括为每个从目标语言的可能的音译词语到源语言的词语 220 的音译产生音译分值 225。例如，图 2B 示出用于每个音译的音译分值 225，包括从“sreeram”到 H2 的分值（例如， $score_{E1 \text{ 至 } H2}$ ），从“shriram”到 H2 的分值（例如， $score_{E3 \text{ 至 } H2}$ ），以及从“shreeram”至 H6 的分值（例如， $score_{E4 \text{ 至 } H6}$ ）。

[0046] 如果通过映射产生音译分值 225，那么列表 210 的给定的可能的音译词语的音译分值 225 能够是相对于另一可能的音译词语的给定的可能的音译词语的置信值。系统能够在识别应被认为用于特定的音译词语的候选同义词的可能的音译词语中使用这些置信值。相对于图 2C 更详细地描述音译分值 225 和置信值。

[0047] 图 2C 示出为第一可能的音译词语 230 识别一个或多个第二可能的音译词语 240 作为第一可能的音译词语 230 的候选同义词。

[0048] 如果音译器从目标语言的两个或更多可能的音译词语映射源语言的词语 220，则这暗示了在目标语言的两个或更多可能的音译词语之间的同义词关系。例如，H2 是通过音译器从三个可能的音译词语：“sreeram”、“shriram”以及“shreeram”映射的源语言的印度语词，暗示了这三个音译词语是同义词。

[0049] 在图 2C 的示例技术中，系统通过识别被映射到源语言的至少一个词语 220 的列表 210 的可能的音译词语来识别第二可能的音译词语 240 作为第一可能的音译词语 230 的候选同义词，其中所述至少一个词语 220 也被从第一可能的音译词语 230 映射。源语言的词语 220 的交集给出了用于音译同义词的候选组。若干技术能够被实现以增加用于音译同义词的候选组的可靠性。

[0050] 在一些实施方式中，除了第一可能的音译词语 230 之外，列表 210 的可能的音译词语的每一个具有相对于第一可能的音译词语 230 的置信值。在这些实施方式中，如果特定的可能的音译词语具有超过指定的阈值的相对于第一可能的音译词语 230 的置信值，那么特定的可能的音译词语是被识别为第一可能的音译词语 230 的候选同义词的第二可能的音译词语 240。如果映射不为每个音译产生音译分值 225，则用于给定的第二可能的音译词语 240 的置信值能够是从第一可能的音译词语 230 和给定的第二可能的音译词语 240 两者映射的源语言的词语 220 的数目的函数。

[0051] 例如，“shriram”和“sriraam”每个映射到也是从第一可能的音译词语 230 “sreeram”映射的仅仅一个词语 220（即，分别是 H2 和 H6）。音译词语“shreeram”映射到也从第一可能的音译词语 230 “sreeram”映射的两个词语 220（即，H2 和 H6）。与源语言的映射的词语 220 的“sreeram”的重叠对于“shreeram”情况大于对于“shriram”和“sriraam”的情况，暗示了“shreeram”可能是比“shriram”或者“sriraam”更可靠的用于“sreeram”的候选同义词。该增加的可靠性能够被反映在相对于“sreeram”更高的用于“shreeram”的置信值中。

[0052] 如果映射为每个音译产生音译分值 225，那么用于给定的第二可能的音译词语

240 的置信值能够是第一可能的音译词语 230 和给定的第二可能的音译词语 240 的音译分值 225 的函数。例如,第二可能的音译词语 240 “shriram”的相对于第一可能的音译词语 230 “sreeram”的置信值能够是音译分值  $225score_{E1 \text{ 至 } H2}$  和  $score_{E3 \text{ 至 } H2}$  的函数,其中两个音译词语映射到 H2。

[0053] 在一些实现方式中,用于给定的第二可能的音译词语 240 的置信值是 web 资源中给定的第二可能的音译词语 240 的出现概率的函数。例如,出现概率能够是给定的第二可能的音译词语 240 的 web 资源中每页的贡献或者 web 资源中每域的贡献。一般来说,较高的出现概率表示给定的第二可能的音译词语 240 是从源语言的词语音译更常见的形式。较高的概率表示常见的音译词语中较高的置信,这能够被反映在用于音译词语的较高的置信值中。

[0054] 在一些实现方式中,用于给定的第二可能的音译词语 240 的置信值是例如音译分值 225 和出现概率的多个分量的函数。尽管图 2C 包括所有映射到也从第一可能的音译词语 230 映射的源语言的词语 220 的可能的音译词语作为第二可能的音译词语 240,用于增加候选组的可靠性的上述技术中的任何一种的实现能够将候选同义词组减少到图 2C 中示出的第二可能的音译词语 240 的子组 (subgroup)。

[0055] 在一些实现方式中,系统识别从第一可能的音译词语 230 和从至少一个第二可能的音译词语 240 映射的源语言的词语 220 中的一个或多个作为除了第二可能的音译词语 240 之外的或者代替第二可能的音译词语 240 的第一可能的音译词语 230 的候选同义词。例如,对于第一可能的音译词语 230“sreeram”,系统能够识别词语 H2 和 H6 作为“sreeram”的候选同义词。在一些实现方式中,系统识别从目标语言的同一音译词语映射的源语言的词语 220 作为候选同义词组。对于图 2C 的示例,系统能够识别从相同的音译词语“sreeram”和“shreeram”映射的词语 H2 和 H6 作为候选同义词。

[0056] 系统能够使用候选音译同义词 (即,第二可能的音译词语 240) 用于查询扩展。例如,当搜索系统 (例如,图 1 的搜索系统 114) 接收包括第一可能的音译词语 230 的查询时,搜索系统能够识别第一可能的音译词语 230 的一个或多个候选音译同义词。该查询能够利用第一可能的音译词语 230 的所识别的候选音译同义词中的一个或多个来扩展。在图 2C 的示例中,该查询能够扩展包括“sreeram”的查询以包括“shriram”、“shreeram”以及“sriraam”中的一个或多个。在一些实现方式中,系统按置信值排名候选同义词,并且系统选择带有 N 个最高置信值的仅仅 N 个候选同义词以包括在扩展的查询中。该系统将扩展的查询提供给搜索引擎 (例如,图 1 的搜索引擎 130),并且接收用于该扩展的查询的搜索结果。

[0057] 在一些实施方式中,如果系统选择可能的音译词语作为用于给定的音译词语的候选音译同义词,则系统还选择给定的音译词语作为用于可能的音译词语的候选音译同义词。在其他的实现方式中,如果系统选择可能的音译词语作为用于给定的音译词语的候选音译同义词,则系统不选择给定的音译词语作为用于可能的音译词语的候选的音译同义词。即,可能存在或者可能不存在音译同义词的逆映射。例如,如果第一音译词语“a”很少被使用并且第二音译词语“b”经常被使用,则带有“b”的“a”的查询扩展通常很有效,因为该扩展将导致返回更多的搜索结果。然而,自动地扩展带有“a”的“b”的查询可能不是很有效的,因为该扩展可能返回不相关的搜索结果。

[0058] 在一些实现方式中,在查询搜索的文档侧发生将候选的音译同义词映射到给定的音译词语,代替利用一个或多个候选音译同义词扩展的查询。对于以上示例,如果用户提交包括音译词语“b”而不是音译词语“a”的查询并且如果 web 文档包含“a”而不包含“b”,那么搜索系统(例如,图 1 的搜索系统 114)能够像 web 文档也包含“b”一样处理该 web 文档,从而该 web 文档是用于包括“b”的搜索的候选搜索结果。然而,由于 web 文档实际上不包括“b”,因此搜索系统能够减少与该 web 文档相关联的分值(例如,用于排名该 web 文档作为候选的搜索结果的信息检索分值),这因此能够减少该 web 文档对于该搜索返回的机会。

[0059] 在一些实现方式中,候选同义词的文档级映射包括源语言的一个或多个词语 220。对于图 2C 的示例,搜索系统能够像 web 文档也包含印度语 H2 或 H6 一样处理包含“sreeram”的 web 文档。该搜索系统也能够相应地减少与该 web 文档相关联的分值。

[0060] 图 3 是用于识别用于音译词语的候选同义词的示例过程 300 的流程图。为了方便,示例过程 300 将参考图 2A 至 2C 的示例技术以及执行过程 300 的系统来进行描述。

[0061] 系统识别目标语言的多个音译词语(步骤 310)。例如,系统识别图 2A 中的列表 210 的可能的音译词语。

[0062] 对于目标语言的多个音译词语中的每个音译词语,系统将音译词语映射到源语言的一个或多个词语(步骤 320)。图 2B 示出使用英语到印度语音译器的映射的示例。

[0063] 对于目标语言的多个音译词语的第一音译词语,系统识别多个音译词语的一个或多个第二音译词语作为第一音译词语的候选同义词(步骤 330)。一个或多个第二音译词语中的每一个被映射到也被从第一音译词语映射的源语言的至少一个词语。例如,图 2C 示出被识别为第一可能的音译词语 230(即“sreeram”)的候选同义词的第二可能的音译词语 240(即,“shriram”、“shreeram”以及“sriraam”)。候选同义词能够被用于查询扩展,例如,如参考图 4 所描述的。

[0064] 图 4 是用于为包括音译词语和候选同义词的扩展的查询提供搜索结果的示例过程 400 的流程图。为了方便,示例过程 400 将参考图 2A 至 2C 的示例技术以及执行该过程 400 的系统来进行描述。

[0065] 系统接收包括第一音译词语的查询(步骤 410)。例如,该查询能够包括图 2C 中使出的音译词语“sreeram”。

[0066] 系统提供一个或多个扩展的查询用于供用户选择,其中每个扩展的查询包括该查询以及第一音译词语的一个或多个候选同义词(步骤 420)。例如,候选同义词能够使用图 3 的示例过程 300 来进行识别。对于包括音译词语“sreeram”的查询,系统能够提供还包括“shriram”、“shreeram”以及“sriraam”中的一个或多个的扩展的查询,如图 2C 中所示。

[0067] 系统从用户接收对扩展的查询的选择(步骤 430)。例如,扩展的查询能够呈现给用户作为在运行在客户端设备(例如,图 1 的客户端设备 104)上的 web 浏览器的界面上的可选择的超链接。系统能够作为用户对用于选择的扩展的查询的超链接做出的选择接收对扩展的查询的选择。在一些实现方式中,系统产生具有一个或多个候选同义词的扩展的查询并且前进到步骤 440 而不执行步骤 420 和 430。

[0068] 系统将扩展的查询提供给搜索引擎(步骤 440)。例如,系统能够将扩展的查询提交到图 1 的搜索引擎 130。搜索引擎执行搜索,将用于扩展的查询的搜索结果发送到系统。

系统接收用于扩展的查询的搜索结果（步骤 450）。

[0069] 在一些实现方式中，系统将步骤 410 接收的查询提供给搜索引擎而不扩展该查询。相反地，系统如以上参考图 2C 描述地执行文档级映射。例如，搜索引擎能够识别包括第一音译词语的候选同义词中的至少一个但是不包括查询中的任何词语（例如，第一音译词语）的 web 资源作为用于查询的可能的搜索结果。或者，搜索引擎能够识别不包括查询中的任何词语（例如第一音译词语）但是包括从第一音译词语映射的和从候选同义词中的至少一个映射的源语言的词语中的至少一个的 web 资源作为用于查询的可能的搜索结果。当实现文档级映射时，系统能够修改（例如，减少）用于排名中使用的与被识别为可能的搜索结果的 web 资源相关联的分值。

[0070] 图 5 是用于识别用于音译词语的候选同义词的示例过程 500 的流程图。为了方便，将参考执行过程 500 的系统描述示例过程 500。总的来说，过程 500 直接学习用于目标语言的音译词语的拼写中的可能变体。由于音译同义词通常在发音上相似，因此在音译同义词之间的变化是特定于语言的。

[0071] 系统产生目标语言的可能的音译同义词的训练组（步骤 510）。系统使用训练组来训练概率模型以学习目标语言的音译同义词中的拼写变化的概率（步骤 520）。系统将概率模型应用于目标语言的特定的音译词语以识别特定的音译词语的一个或多个候选同义词（步骤 530）。系统能够使用候选同义词用于如上所述的查询扩展。

[0072] 在本说明中描述的主题以及功能操作的实施例可以实现在数字电子电路中，或实现在计算机软件、固件、或硬件中，包括在本说明中公开的结构和他们的结构等价物，或者实现在他们中的一个或多个的组合中。本说明中描述的主题的实施例可以被实现为一个或多个计算机程序产品，即用于由数据处理装置执行的或控制数据处理装置的操作的编码在有形程序载体上的计算机程序指令的一个或多个模块。有形程序载体能够是传播的信号或者计算机可读介质。传播信号是人工生成的信号，例如，机器生成的电的、光学或电磁的信号，其被生成以编码用于传送到适当接收器装置的信息以供计算机执行。计算机可读介质可以是机器可读的存储装置、机器可读的存储基片、存储器设备、实现机器可读的传播信号的物质成分或他们中的一个或多个的组合。

[0073] 术语“数据处理装置”涵盖用于处理数据的所有装置、设备以及机器，例如包括可编程处理器、计算机或多处理器或计算机。除硬件之外，所述装置可以包括创建用于正讨论的计算机程序的执行环境的代码，例如组成处理器固件、协议栈、数据库管理系统、操作系统或他们中的一个或多个的组合的代码。

[0074] 计算机程序（也称作程序、软件、软件应用、脚本或代码）可以以任何形式的编程语言编写，包括编译或解释语言，或者声明或过程性语言，以及它可以以任何形式部署，包括作为独立程序或模块、组件、子程序或适合在计算环境中使用的其它单元。计算机程序没有必要对应于文件系统中的文件。程序可以被存储在保持其它程序或数据的文件（例如，存储在标记语言文档中的一个或多个脚本）的一部分、专用于正被讨论的程序的单个文件或者多个协调文件（例如，存储一个或多个模块、子程序或部分代码的文件）中。计算机程序可以被部署为在一个计算机上或者在位于一个地点或跨多个地点分布并且由通信网络互连的多个计算机上执行。

[0075] 在本说明中描述的过程和逻辑流可以由执行一个或多个计算机程序的一个或多

个可编程处理器执行以通过操作输入数据和生成输出来执行功能。过程和逻辑流还可以通过专用的逻辑电路执行,以及装置还可以被实现为专用的逻辑电路,专用的逻辑电路例如FPGA(场可编程门阵列)或ASIC(专用集成电路)。

[0076] 适合于计算机程序的执行的处理器包括例如通用和专用的微处理器,以及任何类型的数字计算机的任何一个或多个处理器。通常,处理器将从只读存储器或随机存取存储器或两者接收指令和数据。计算机的主要元件是用于执行指令的处理器和用于存储指令和数据的一个或多个存储设备。通常,计算机还包括用于存储数据的一个或多个海量储存设备,例如磁盘、磁光盘或光盘,或可操作地耦接到所述一个或多个海量储存设备以从其接收数据或者向其传输数据,或者其两者。然而,计算机不必具有这样的设备。此外,计算机能够被嵌入在另外的设备中,举几个例子,例如,移动电话、个人数字助理(PDA)、移动音频或视频播放器、游戏控制台、全球定位系统(GPS)接收器、带有说话语言输入的设备。智能电话是带有说话语言输入的设备的示例,其能够接受语音输入(例如,说话输入到设备上的麦克风中的用户查询)。

[0077] 适合于存储计算机程序指令和数据的计算机可读介质包括所有形式的非易失性存储器、媒体和存储设备,例如包括:半导体存储器设备,例如EPROM、EEPROM和闪存设备;磁盘,例如内部硬盘或活动盘;磁光盘;以及CD-ROM和DVD-ROM盘。处理器和存储器可以由专用的逻辑电路补充,或并入专用的逻辑电路。

[0078] 为了提供与用户的交互,本说明中描述的主题的实施例可以在具有以下的计算机上实现:用于向用户显示信息的显示设备,例如,CRT(阴极射线管)或LCD(液晶显示)监视器,和用户通过其可以提供输入给计算机的键盘和指示设备,例如,鼠标或轨迹球。也可以使用其他类型的设备提供与用户的交互;例如,提供给用户的反馈可以是任何形式的传感反馈,例如视觉反馈、听觉反馈或触觉反馈;以及可以任何形式,包括声学的、话语或触觉的输入,接收来自用户的输入。

[0079] 本说明中描述的主题的实施例可以在包括例如数据服务器的后端组件、或包括例如应用服务器的中间件组件或包括例如具有图形用户界面或用户通过其可以与在本说明中描述的主题的实施方式交互的Web浏览器的客户端计算机的前端组件、或这样的后端、中间件、或前端组件中的一个或多个的组合的计算系统中实现。系统的组件可以通过任何形式或介质的数字数据通信互连,例如通信网络。通信网络的示例包括局域网("LAN")和广域网("WAN"),例如因特网。

[0080] 计算系统可以包括客户端和服务器。客户端和服务器通常彼此远离并且通常通过通信网络进行交互。客户端和服务器的关系依靠在各自的计算机上运行并且彼此具有客户端-服务器关系的计算机程序产生。

[0081] 尽管本说明包括许多特定实施方式细节,但是这些细节不应该被看作是对任何发明或者所要求的范围的限定,而应该看作针对特定发明的特定实施例的特征的描述。在本说明书中分立实施例的上下文中描述的某些特征还可以在单独实施例的组合中实现。相反地,在单个实施例的上下文中描述的各种特征还可以分立地在多个实施例中实现或者在任何适当的子组合中实现。此外,虽然特征可能在上面被描述为在某些组合中起作用,甚至最初要求这样,但是在一些情况下来自所要求的组合的一个或多个特征可以从组合中删去,并且所要求的组合可以指向子组合或者子组合的变体。

[0082] 同样地,虽然在附图中以特定的顺序描述了操作,但是不应该理解为这样的操作需要以所示的特定顺序被执行或者以连续的顺序被执行、或者全部图示的操作要被执行以实现所希望的结果。在某些环境中,多任务并且并行处理可以是有利的。此外,在如上所述实施例中的各种系统组件的分离不应该被理解为在全部实施例中都需要这样的分离,并且应当理解的是描述的程序组件和系统通常可以被集成到一起成为单个软件产品或封装为多个软件产品。

[0083] 已经描述了本说明中描述的主题的特定的实施例。其它实施例在所附权利要求的范围内。例如,权利要求中记载的动作能够以不同的顺序来执行并且仍然实现所希望的结果。作为一个示例,在附图中描绘的过程不必需要所示出的特定顺序,或者连续的顺序以实现所希望的结果。在某些实现方式中,多任务并且并行处理可以是有利的。

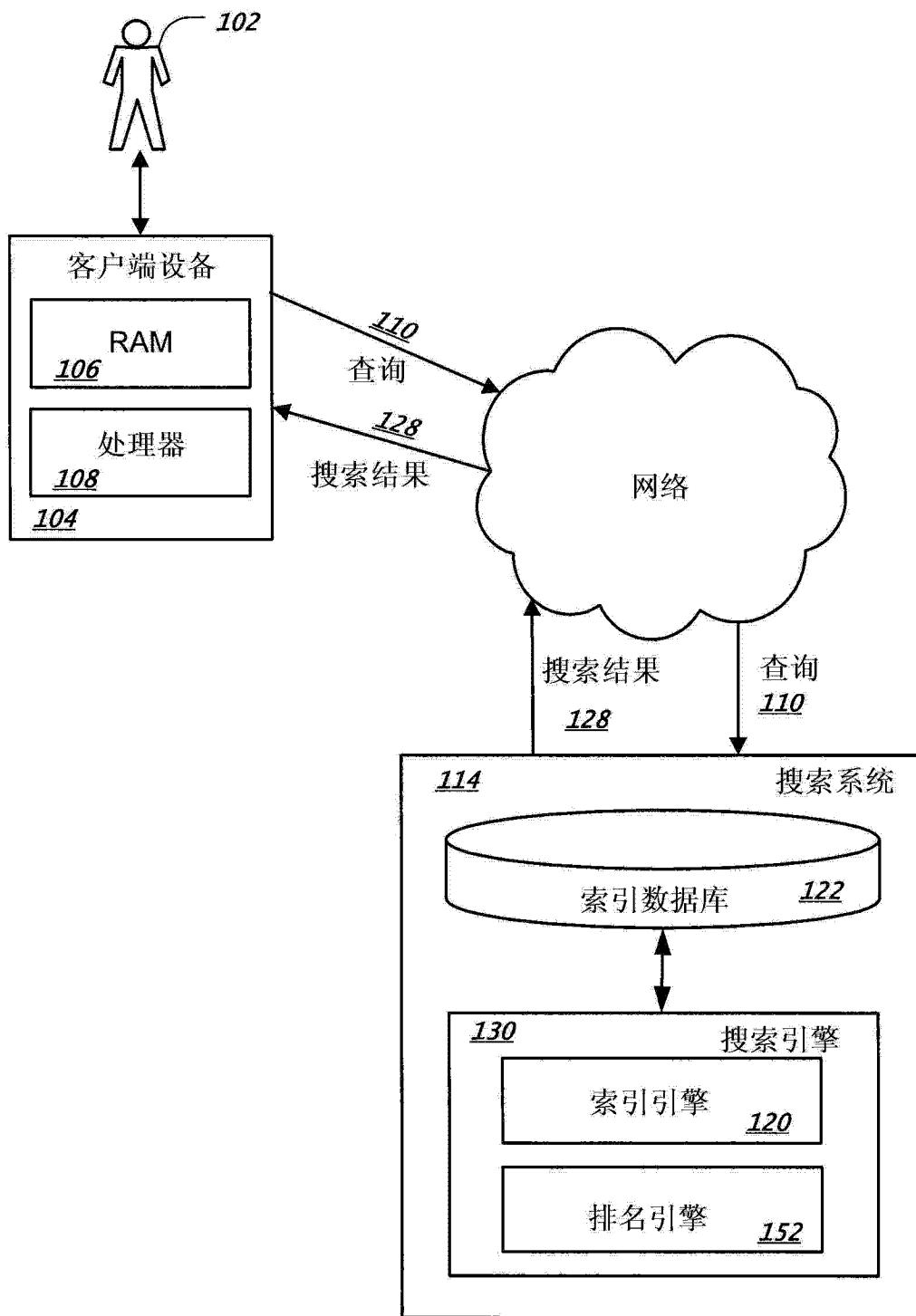


图 1

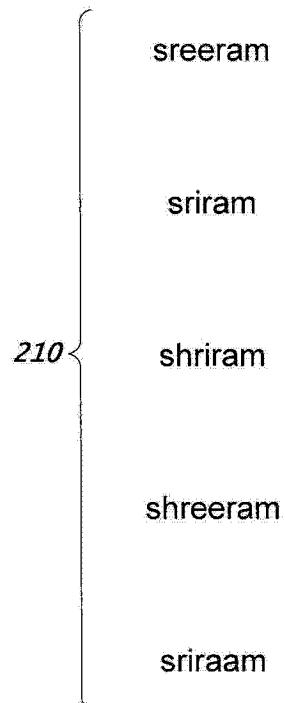
印度语词的音译（以英语）

图 2A

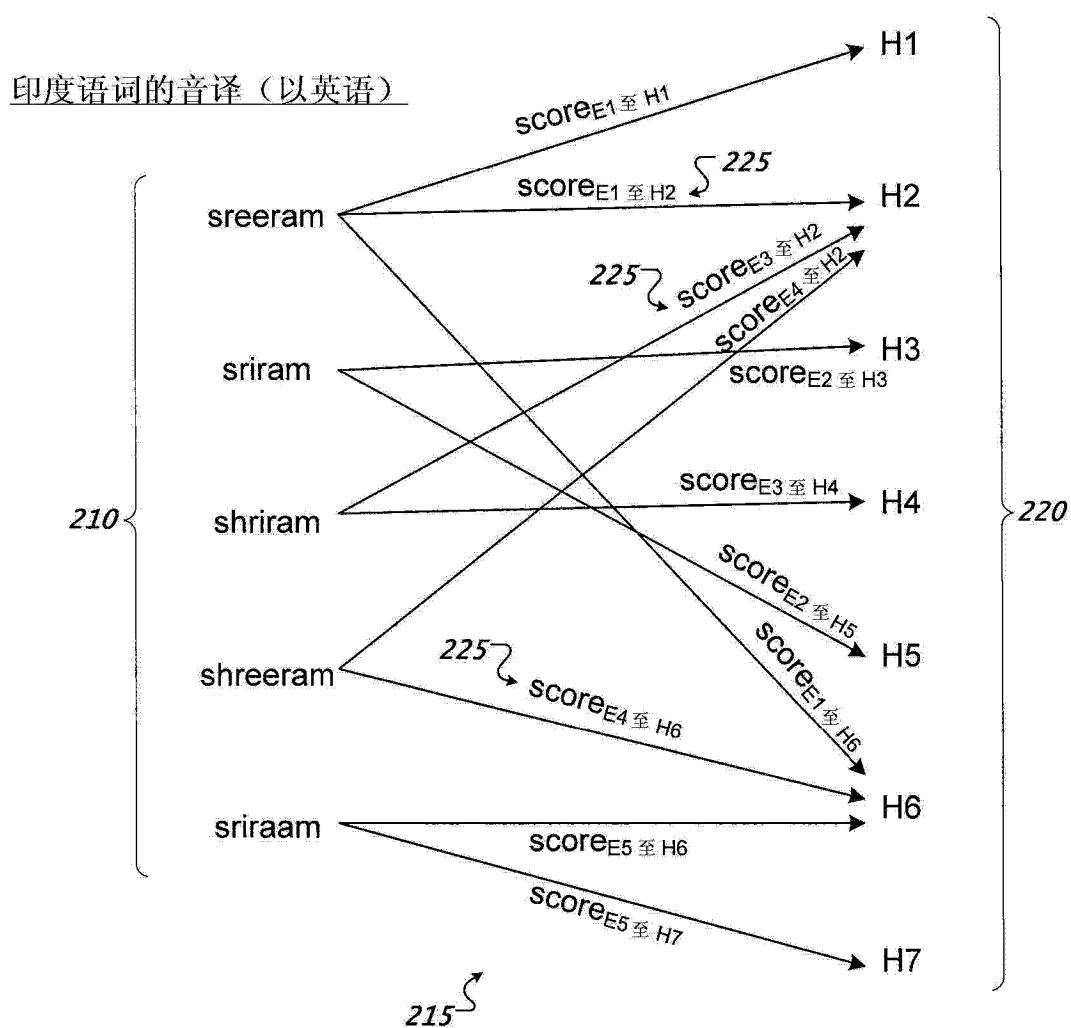
印度语词

图 2B

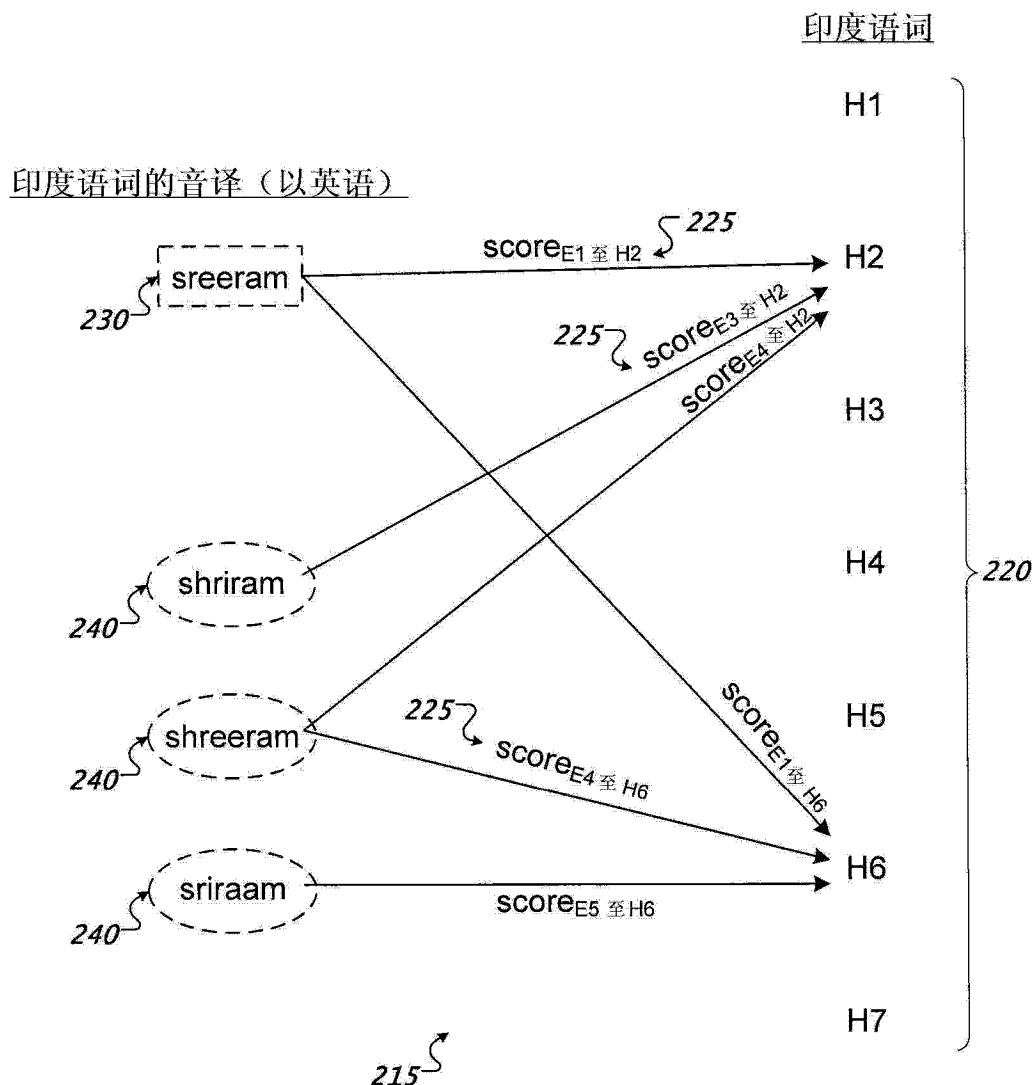


图 2C

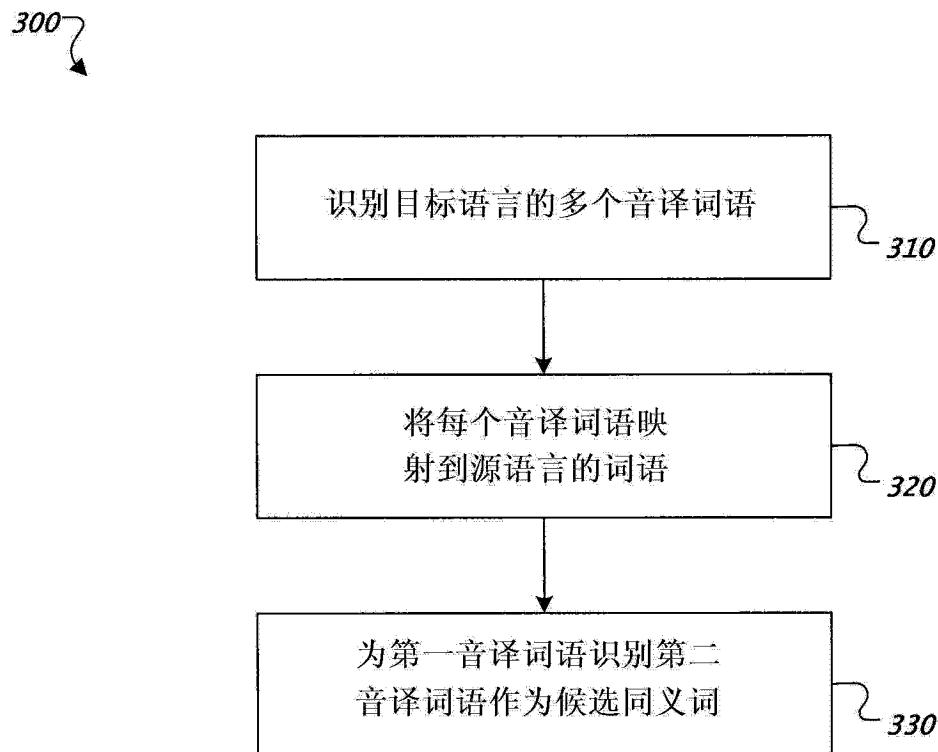


图 3

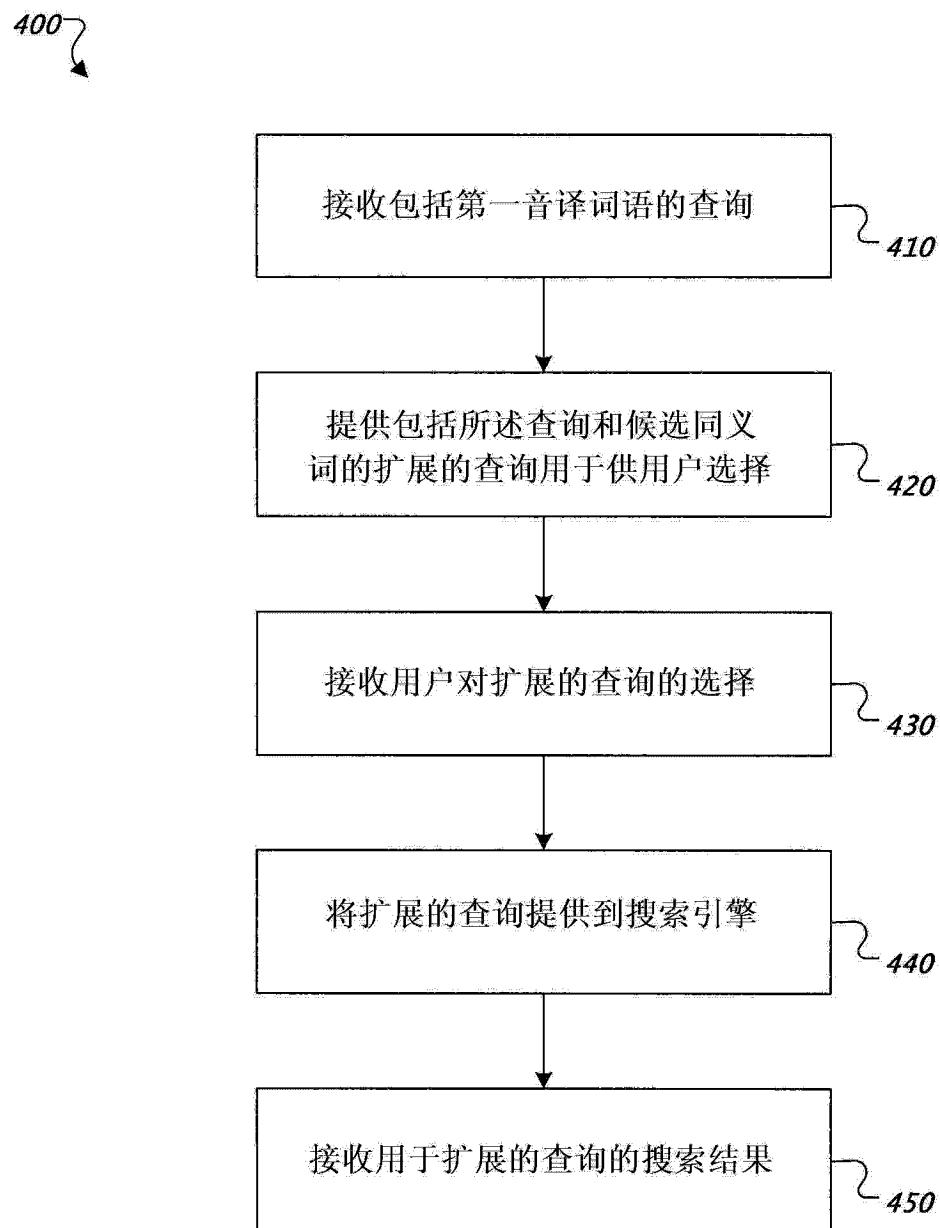


图 4

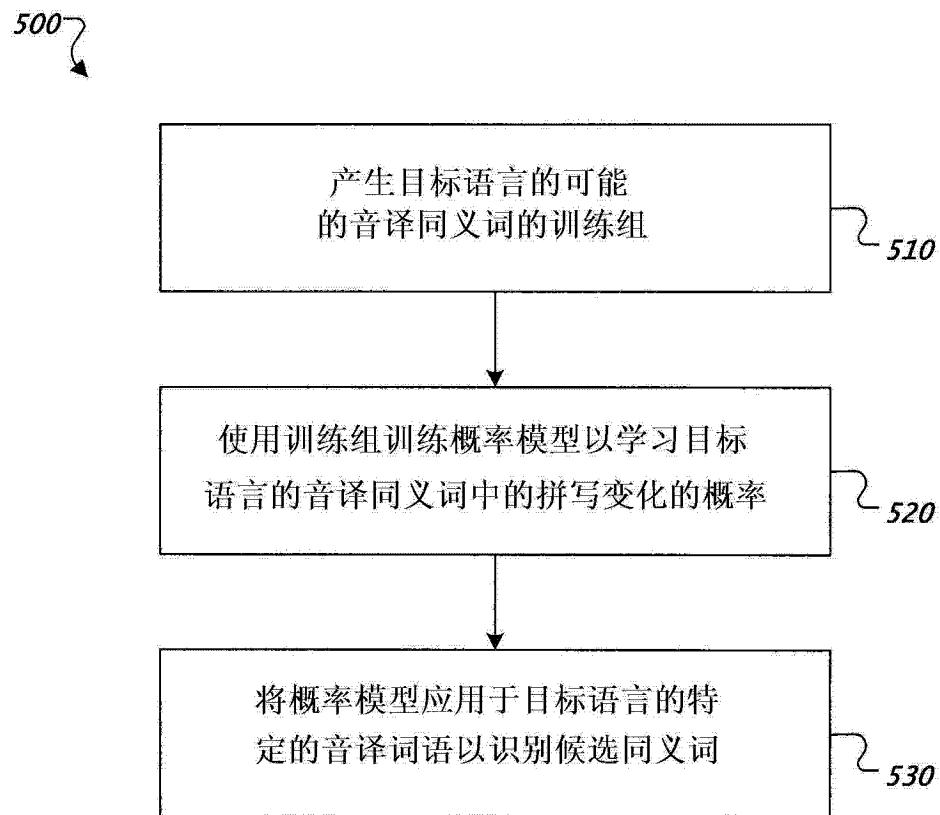


图 5