(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 101467125 B
(45)授权公告日 2010.12.22

(21)申请号 200780021902.1
(22)申请日 2007.04.19
(30)优先权数据
11/408,245 2006.04.19 US
11/408,242 2006.04.19 US
11/408,243 2006.04.19 US
11/407,860 2006.04.19 US
(85)PCT申请进入国家阶段日
2008.12.12
(86)PCT申请的申请数据
PCT/US2007/067014 2007.04.19
(87)PCT申请的公布数据
WO2007/124385 EN 2007.11.01
(73)专利权人 谷歌公司
地址 美国加利福尼亚州
(72)发明人 鲁齐拉·S·达特
法比奥·洛皮亚诺
(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 张焕生 安翔
(51)Int.Cl.
G06F 7/00(2006.01)
G06F 17/30(2006.01)
(56)对比文件
US 6167369 A,2000.12.26, 说明书第2栏第44-50行,第17栏第52-56、61-64行.
US 6519585 B1,2003.02.11, 权利要求9,24,28.
审查员 杨洁

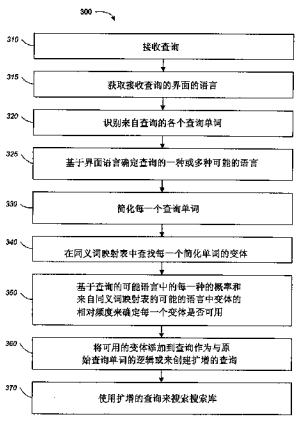
权利要求书 1 页 说明书 15 页 附图 20 页

(54)发明名称

用于处理查询词语的方法和系统

(57)摘要

用于执行与处理提供给搜索引擎的搜索查询中的查询词语有关的操作的方法、系统和包括计算机程序产品的设备。在一个方面,一种方法包括从查询词语和用户界面的语言确定查询语言。在另一个方面中,一种方法包括使用界面语言来选择一个或多个映射并且使用所述映射来简化每一个查询词语;以及将每一个简化查询词语应用于同义词映射表以识别可能的同义词,用所述可能的同义词来扩增搜索查询。在另一个方面中,从文档库生成同义词映射表。在另一个方面中,一种方法包括通过在同义词映射表中查找简化查询词语来为查询词语识别一个或多个潜在同义词,同义词映射表将多个键中的每一个映射到一个或多个变体,每一个变体是与一种或多种文档语言相关联的单词。



CN 101467125 B

1. 一种计算机实现的处理查询词语的方法，包括：

通过用户界面从用户接收包括一个或多个查询词语的搜索查询，所述用户界面具有界面语言，所述界面语言是自然语言，所述用户界面以所述自然语言来展示信息；

从所述查询词语和所述界面语言为所述搜索查询确定查询语言，所述查询语言是多种自然语言中的一种；

使用所确定的查询语言来选择一个或多个映射并且使用所选择的映射来将所述查询词语中的一个或多个简化为相对应的简化查询词语；

使用所简化的查询词语来识别可能的同义词；以及

基于所述搜索查询存在于所述多种自然语言的每一种中的概率以及基于所识别的同义词在所述多种自然语言的每一种中出现的频率，从所识别的同义词中选择用来扩增所述搜索查询的一个或多个同义词。

2. 如权利要求 1 所述的方法，其中为所述搜索查询确定查询语言进一步包括：

为所述多种自然语言的每一种确定分值，所述分值指示所述查询语言是所述多种自然语言中的所述每一种的可能性。

3. 如权利要求 2 所述的方法，进一步包括：

使用用于所述多种自然语言的所述分值的每一个来识别用来扩增所述搜索查询的同义词。

4. 一种处理查询词语的系统，包括：

用于通过用户界面从用户接收包括一个或多个查询词语的搜索查询的装置，所述用户界面具有界面语言，所述界面语言是自然语言，所述用户界面以所述自然语言来展示信息；

用于从所述查询词语和所述界面语言为所述搜索查询确定查询语言的装置，所述查询语言是多种自然语言中的一种；

使用所确定的查询语言来选择一个或多个映射并且使用所选择的映射来将所述查询词语中的一个或多个简化为相对应的简化查询词语的装置；

使用所简化的查询词语来识别可能的同义词的装置；以及

基于所述搜索查询存在于所述多种自然语言的每一种中的概率以及基于所识别的同义词在所述多种自然语言的每一种中出现的频率，从所识别的同义词中选择用来扩增所述搜索查询的一个或多个同义词的装置。

5. 如权利要求 4 所述的系统，其中为所述搜索查询确定查询语言的装置进一步包括：

用于为所述多种自然语言的每一种确定分值的装置，所述分值指示所述查询语言是所述多种自然语言中的所述每一种的可能性。

6. 如权利要求 5 所述的系统，进一步包括：

用于使用用于所述多种自然语言的所述分值的每一个来识别用来扩增所述搜索查询的同义词的装置。

用于处理查询词语的方法和系统

背景技术

[0001] 本发明涉及在处理搜索查询中以及在包括文档和其它可搜索资源的库上的搜索中处理语言不确定性，其中查询和资源可以以多种不同语言中的任何一种来表示。

[0002] 搜索引擎对文档进行索引并且提供方法来搜索其内容由搜索引擎进行索引的文档。文档以许多不同的语言书写；一些文档具有用多种语言的内容。各种字符被用来表示这些语言的单词：拉丁字母（即，从 A 到 Z 的 26 个非重读字符，大小写体）、区别音符（即，重读字符）、连字（例如，Æ、β、Œ）、西里尔字符以及其它。

[0003] 遗憾的是，产生这些字符的能力和简便性在装置与装置之间差别极大。内容的作者和搜索引擎的用户可能都不能够便利地产生其更喜欢的字符。反而，这样的装置的用户将经常提供作为相近替代物的字符或字符序列。例如，AE 可以被提供来替代 Æ。而且，这样的替代的惯例在语言和用户之间不同。例如，搜索 AE 的某些用户可能更喜欢看见也包括 Æ 的结果。

[0004] 用于解决在搜索引擎中的该问题的一种方法是处理索引内容以移除重音并将特殊字符转换为一组标准字符。该方法从索引移除信息，使得不可能仅检索单词的特定重读实例。该方法也因语言不可知论 (agnosticism) 而受损，其中所述语言不可知论不受这样的用户影响：所述用户的预期由所述用户的特定语言的惯例所形成。

[0005] 发明内容

[0006] 本说明书公开了用于使用搜索查询的词语的技术的各种实施例。 实施例表征为 (feature) 方法、系统、设备，包括计算机程序产品设备。在本发明内容中将参考方法描述这些中的每一个，对于所述方法存在相对应的系统和设备。

[0007] 一般而言，在一个方面中，方法具有以下特征：通过用户界面从用户接收包括一个或多个查询词语的搜索查询，所述用户界面具有界面语言，所述界面语言是自然语言；以及从查询词语和界面语言为查询确定查询语言，所述查询语言是自然语言。这些和其它的实施例可以可选地包括下列特征中的一个或多个。所述方法包括为多种语言的每一种确定分值，所述分值指示查询语言是多种语言中的一种的可能性。所述方法包括使用查询语言来选择一个或多个映射并且使用所选择的一个或多个映射来将每一个查询词语简化为相对应的简化查询词语；以及将每一个简化查询词语应用于同义词映射表以识别扩增 (augment) 搜索查询的可能的同义词。所述方法包括为多种语言的每一种确定分值，所述分值指示查询语言是多种语言中的一种的可能性。

[0008] 一般而言，在另一个方面，方法具有以下特征：通过用户界面从用户接收由一个或多个查询词语组成的搜索查询，所述用户界面具有界面语言，所述界面语言是自然语言；使用界面语言来选择一个或多个映射并且使用所选择的一个或多个映射来将每一个查询词语简化为相对应的简化查询词语；以及将每一个简化查询词语应用于同义词映射表以识别扩增搜索查询的可能的同义词。

[0009] 一般而言，在另一个方面，方法具有以下特征：从文档库生成同义词映射表，每一个文档具有归属 (attribute) 于该文档的文档语言，所述文档语言每一种都是自然语言；

其中同义词映射表将多个键中的每一个映射到一个或多个相对应的变体；以及每一个变体与文档语言中的一种或多种相关联。这些和其它的实施例可以可选地包括下列特征中的一个或多个。所述方法包括：对于每一种相关联的语言，每一个变体与指示该变体在用于相同键的相关联的语言的所有变体中的相 对频度的分值相关联。自动确定每一个文档的文档语言归属。

[0010] 一般而言，在另一个方面，方法具有以下特征：通过将依赖于语言的映射的第一集合应用于库中的单词以为映射表生成键来从文档库生成同义词映射表，每一个文档具有归属于该文档的文档语言，归属于每一个文档的文档语言被用来确定应用于文档中的单词的依赖于语言的映射。这些和其它的实施例可以可选地包括下列特征中的一个或多个。所述方法包括通过将依赖于语言的映射的第二集合应用于每一个查询词语来从搜索查询中的每一个查询词语生成简化查询词语，所述搜索查询具有归属于该搜索查询的查询语言，归属于该搜索查询的查询语言被用来确定应用于每一个查询词语的依赖于语言的映射。依赖于语言的映射的第一集合与依赖于语言的映射的第二集合不同。

[0011] 一般而言，在另一个方面，方法具有以下特征：通过将依赖于语言的映射的第一集合应用于库中的单词以为映射表生成键来从文档库生成同义词映射表，每一个文档具有归属于该文档的文档语言，归属于每一个文档的文档语言被用来确定应用于文档中的单词的依赖于语言的映射；通过将依赖于语言的映射的第二集合应用于搜索查询中的查询词语来从搜索查询生成简化查询词语，所述搜索查询具有归属于该搜索查询的查询语言，归属于该搜索查询的查询语言被用来确定应用于查询词语的依赖于语言的映射；其中所述搜索查询包括第一查询词语，通过来自查询语言所确定的依赖于语言的映射的第二集合的所应用的依赖于语言的映射将第一查询词语映射到第一简化查询词语，通过查询语言所确定的依赖于语言的映射的第一集合中的依赖于语言的映射将第一查询词语映射到第一键，并且第一简化查询词语与第一键不同。这些和其它的实施例可以可选地包括下列特征中的一个或多个。所述方法包括将界面语言归属于查询作为查询语言。

[0012] 一般而言，在另一个方面中，方法具有以下特征：通过用户界面从用户接收包括查询词语的搜索查询，所述搜索查询具有归属于该搜 索查询的查询语言；从查询词语获得简化查询词语；以及通过在同义词映射表中查找简化查询词语为查询词语识别一个或多个潜在同义词，所述同义词映射表将多个键中的每一个映射到一个或多个相对应的变体，每一个变体是与一种或多种文档语言相关联的单词，并且每一个变体对于每一种相关联的语言与指示该变体在用于相同键的相关联的语言的所有变体中的相对频度的变体－语言分值相关联。这些和其它的实施例可以可选地包括下列特征中的一个或多个。所述方法包括使用所归属的查询语言和用于简化查询词语的一个或多个变体的变体－语言分值来选择变体以在扩增搜索查询中使用。所述方法包括将界面语言归属于查询作为查询语言。在搜索查询具有归属于该搜索查询的多种查询语言的情况下，每一种具有各自的查询－语言分值，所述方法进一步包括使用 (a) 查询－语言分值以及 (b) 用于简化查询词语的一个或多个变体的变体－语言分值来选择变体以在扩增搜索查询中使用。使用查询－语言分值和变体－语言分值包括对所有语言的以下乘积求和：对于每一种语言，用于该语言的查询－语言分值和用于该语言的变体－语言分值的乘积。

[0013] 一般而言，在另一个方面中，方法具有以下特征：通过用户界面从用户接收由一

个或多个查询词语组成的搜索查询；以及接收在简化搜索查询的查询词语中应用标音(transliteration)的用户偏好的指示。这些和其它的实施例可以可选地包括下列特征中的一个或多个。所述方法包括：如果用户偏好是应用标音则在简化搜索查询的查询词语中应用标音来生成简化查询词语，否则在简化搜索查询的查询词语中不应用标音来生成简化查询词语；以及使用简化查询词语来识别同义词以在扩增搜索查询中使用。在简化搜索查询中应用标音的用户偏好的指示是对多种特定界面语言中的一种的用户选择。所述方法包括通过用户界面从用户接收由一个或多个查询词语组成的搜索查询；在简化搜索查询的查询词语中应用标音来生成简化查询词语；以及使用简化查询词语来识别同义词以在扩增搜索查询中使用。

[0014] 一般而言，在另一个方面中，方法具有以下特征：通过用户界面从用户接收由一个或多个原始查询词语组成的搜索查询用于搜索文档的集合，所述用户界面具有用户界面语言；将用户界面语言识别为小规模语言或非小规模语言，小规模语言是在文档的集合中具有相对较少的表现的自然语言；将每一个查询词语简化为简化形式；以及如果用户界面语言是小规模语言，则对于具有与原始词语不同的简化形式的每一个原始查询词语，使用原始查询词语本身并且不为查询词语提供任何同义词，而对于与其简化形式相同的每一个原始查询词语，使用简化形式来为原始查询词语识别同义词用于在扩增搜索查询中使用。这些和其它的实施例可以可选地包括下列特征中的一个或多个。简化每一个查询词语包括标音。

[0015] 可以实现本发明的特定实施例以实现下列优势中的一个或多个。系统可以正确地将适当的重音添加到用西班牙语或葡萄牙语的单词，其中重音在每一种语言中不同。系统可以正确地将重音添加到用与用户正与之交互的用户界面的语言不同的语言的单词。系统可以在适当的情况下标音。系统可以避免将不必要的可区别变体添加到搜索查询，增加搜索结果将用用户所希望的语言的可能性。

[0016] 在附图和下面的描述中阐述了本发明的一个或多个实施例的细节。本发明的其它特征、方面和优势从描述和附图以及从权利要求中将是显而易见的。

附图说明

[0017] 图 1 是用于建立同义词映射表的过程的流程图。

[0018] 图 2 是用于从普通形式条目创建同义词映射表的过程的流程图。

[0019] 图 3 是重写查询的过程的流程图。

[0020] 图 4 是同义词映射表的图示。

[0021] 图 5A、5B 和 5C 以及 6-34 示出了转换映射表组。

[0022] 图 35 是搜索引擎的框图。

[0023] 在各个附图中相同的引用数字和标记指示相同的元素。

具体实施方式

[0024] 如图 1 中所示，过程 100 从文档库创建同义词映射表。文档可以是 HTML（超文本标记语言）文档、PDF（便携式文档格式）文档、文本文档、字处理文档（例如，Microsoft® Word 文档）、用户网文章或具有文本内容（包括元数据内容）的任何其它种类的文档。过

程 100 也可以应用于其它种类的文本可搜索的资源，例如通过元数据识别的媒体资源。

[0025] 同义词映射表包含作为键的普通形式的单词，所述普通形式的单词中的每一个与一个或多个变体相关联。例如，考虑在其中仅找到两种语言 :法语和英语的简单库。如果“elephant”是同义词映射表中的普通形式的条目，则如果在库中找到变体“elephant”、“éléphant”和“eléphant”，这些变体将作为值与该条目相关联。每一个值也包括附加信息 :变体的实例在其中出现的文档的语言，以及变体以该语言出现的次数。继续该示例，在库中，“eléphant”可能在被认为是英语的文档中被找到 90 次，并且在被认为是法语的文档中被找到 300 次。

[0026] 过程 100 在文档的训练库上操作（步骤 110）。文档的训练库理想地是代表包含在搜索库中的文档的文档的集合。替选地，训练库和搜索库可以是相同的库，或者训练库可以是搜索库的快照或来自搜索库的提取部分。训练库应当包含来自在搜索库中表现的所有语言的文档。训练库应当包含用每一种语言的足够数量的文档，以使文档包含在搜索库中该语言的所有文档内找到的单词的重要部分。

[0027] 在一个实施方式中，以已知并且一致的字符编码对训练和搜索库中的每一个文档编码，所述字符编码诸如 8 位统一转换格式 (UTF-8)，其可以以 Unicode 标准（即，大部分已知的字符和表意文字）来对任何字符编码。不一致或未知编码的文档须经编码转换。在一个实施方式 中，库是 web 爬行器从 Web 发现的文档的集合。

[0028] 识别训练库中的每一个文档的语言。确定每一个文档的语言可以明确地是过程 100 的一部分（步骤 120）。替选地，文档的语言可以是包含在训练库中的信息的一部分。文档或单词的语言不一定简单地对应于自然语言。语言可以包括由其拼写、语法、词汇或词法定义的任何可区别的语言系统。例如，罗马印度语言，一组语言（例如孟加拉语和印地语）的罗马化标音的等价体，可被看作是在传统拼写字体中独立于孟加拉语和印地语两者的语言。

[0029] 文档语言检测过程使用统计学习理论。在一个实施方式中，其使用朴素贝叶斯（**Naïve** Bayes）分类模型来计算可能的种类的可能性并且预测具有最大可能性的种类。种类是语言 / 编码对，例如英语 /ASCII、日语 /Shift-JIS 或俄语 /UTF8，文档可以用所述语言 / 编码对来表示。某些语言与多个种类相对应，因为可以用多种编码对所述语言编码，而某些编码与多个种类相对应，因为所述编码可以被用来表示多种语言。

[0030] 朴素贝叶斯模型被用来基于文本页的文本和（可选地）统一资源定位符 (URL) 为文本页确定最可能的种类。

[0031] 使用朴素贝叶斯模型来确定文本页的编码，所述朴素贝叶斯模型基于表现文本的字节的配对来预测最大可能性的编码。如果文本页的 URL 是可用的，假设文本来自某一顶级域（即，因特网域名的最后部分）则该模型也将特定编码的概率计算在内。

[0032] 在执行语言检测时将文本从其原始编码转换为 Unicode，并且使用特征来执行该语言检测。典型地，自然语言单词是要用的最佳特征，因此将文本分割成单词。给定了语言，朴素贝叶斯模型计算各个单词的概率并且基于该概率来为文本预测最大可能性语言。

[0033] 可以使用以各种编码和语言的大量电子文档样本来训练并测试朴 素贝叶斯模型。训练朴素贝叶斯模型实质上是计算特征对于给定语言的概率。

[0034] 过程 100 创建包含在训练库中的所有文档中找到的每个唯一单词的字典（步骤

125)。根据在其中找到该单词的文档的所识别的语言来对在库中找到的给定单词的每一个实例计数。将以每一种文档语言的每一个单词的频度记录在字典中。例如,如果遇到 200 次 hello—在被识别为英语文档的文档内 150 次以及在被识别为德语文档的文档内 50 次—则 hello 字典条目记录在英语和德语文档中找到了 hello 并且分别找到 150 和 50 次。

[0035] 对于每一种语言,可以定义预定的字符黑名单。字符的黑名单是在该语言的文档中通常不会出现的字符的列表。字符的黑名单不一定反映语言的严格固有特点。例如,'w' 不在法语单词中出现,因此可以将其添加到法语黑名单。然而,包含 'w' 的借用的和外来的单词在法语文档中出现足够多次,则可以从法语黑名单中排除 'w'。可以全部地或部分地人工确定列表。替选地,可以统计地分析在已知为特定语言的文档中字符的出现次数,以告知人工过程或自动产生字符的黑名单。

[0036] 过程 100 可以使用字符的黑名单来确定在训练库中找到的单词是否看起来违反语言的常规规则。忽略这样的单词,即不将这样的单词插入字典中。例如,如果"QqWwXxYy"是用于匈牙利语的字符的黑名单,则当在匈牙利语文档中发现"xylophone"时将其忽略。

[0037] 过程 100 将字典中的每一个单词条目映射到用于单词看起来是的每一种语言的普通形式(步骤 130)。通常,普通形式是符合简化的、标准的、规范的或其它一致的拼写的单词,例如没有使用重读字符来表示的单词。过程 100 根据预定义和特定于语言的映射来映射每一个单词。例如,映射将在识别为法语的文档中找到的"éléphant"转换为"elephant"。

[0038] 根据特定于语言的映射将单词映射到普通形式。每一个特定于语言的映射是一个或多个字符转换映射表的集合。每一个转换映射表指定一个或多个输入字符和一个或多个输入字符被映射到的一个或多个输出字符。过程 100 以映射表的一个或多个输出字符来替代与转换映射表的输入相匹配的字符的最大序列(或前缀)。其它字符复制不变。对于任何给定的单词,该字符转换过程的结果生成该单词的普通形式。设计来帮助最长前缀匹配的数据结构可以被用来存储特定于语言的映射(例如,查找树 (trie) 或前缀树)。

[0039] 例如,来自俄语文档的"в о д к а"被映射到"в д к а"(未改变),而在塞尔维亚语文档中的"в о д к а"被映射到"vodka"。特定于语言的转换旨在捕捉那些语言的作者的预期。这反映了虽然俄语作家可能提供"в о д к а",但是塞尔维亚语习惯暗示在搜索查询中西里尔语单词更常作为罗马化标音的等价体给出。

[0040] 指定多于一个输入字符的转换映射表是用于映射包含可叠缩连字的单词的转换的特殊情况。可叠缩连字是两个字符组合,在某些语言中其可被表现为单个、通常重读的字符。例如,德语转换暗示如果 'Ü' 不能被排版,则 'Ue' 或 'UE' 是适当的替代体。因此德语文档可以将单词"über"拼作"ueber"。在映射到普通形式期间,两个字符转换映射表将经常叠缩可叠缩的连字并且将结果去重音。例如,在一个实施方式中,德语转换映射表将"ueber"和"über"都转换为"uber"。

[0041] 过程 100 从普通形式映射、字典条目以及条目的相关联的语言统计来创建同义词映射表(步骤 150)。如上所获得的每一个不同的普通形式成为同义词映射表中的键。映射到给定键的字典条目使用用于条目的语言的每一种的映射成为键的值。在同义词映射表中,字典条目将被称为变体。通常,每一个键与多个变体相关联,变体中的每一个与变体的语言统计相关联。倘若是在上述示例中的映射,"в о д к а"是 一个键,其值指的是与俄

语（而非塞尔维亚语）相关联的至少一个变体“в о д к а”。此外，“vodka”是另一个键，其值指的是与塞尔维亚语（而非俄语）相关联的至少一个变体“в о д к а”。

[0042] 图 2 示出了用于创建同义词映射表（图 1 的步骤 150）的过程 200 的一个实施方式。过程 200 包括接收普通形式条目，如上所述（步骤 210）。从同义词映射表中略去仅包含与其普通形式相同的一个变体的任何普通形式条目（步骤 220）。这样的条目不为普通形式提供同义词。

[0043] 过程 200 也移除与具有未超过预定义的绝对阈值的频度的变体相关联的任何语言（步骤 230）。绝对阈值是预先确定的并且以每一种语言为基础来指定。这种阈值被用来移除在训练库中可能被拼错或弄错的变体。对于在训练库中被充分表现的语言，大的阈值（例如，用于英语是 40）将通常略去微弱的拼错。用于未被充分表现的小规模语言的阈值将被设置为较低（例如 10）以保留合法但罕见的单词。对于在库中被不足地表现的语言，阈值可以被关闭（或被设置为 0）。

[0044] 在特定语言内，如果变体包含叠缩连字并且其重读等价体也不是用于键的变体，则过程 200 略去用于该键的该变体（步骤 240）。

[0045] 某些变体仅依赖于其重音就可能具有不同的含意。为了避免这样的变体对同义词映射表的不希望的污染，可以定义特定于语言的单词黑名单。每一个黑名单包含应当不是与给定语言相关联的变体的单词列表。如果变体在语言的黑名单上，则该语言被从变体解除关联。例如，如果“the”在法语黑名单上，则其普通形式是“the”的变体不能与法语相关联。这防止了在英语“the”和法语“thé”之间的混淆。

[0046] 对于每一个键，计算每一个变体在用于特定语言的所有变体中的相对频度（步骤 250）。为了计算在给定语言中任何给定变体的相对频度，对于相同的键，将该变体在该语言中出现的次数除以在相同语言 中所有变体的出现的总数。例如，如果键是“elephant”，并且“éléphant”在英语和法语中分别出现了 100 和 1000 次；以及“eléphant”在英语和法语中分别出现了 90 和 300 次，则在英语中“éléphant”的相对频度是 52％（即，100/(100+90)）。在一个实施方式中，对于每一种语言每一个变体的相对频度被存储在同义词映射表中。

[0047] 如果语言的相对频度不满足预定义的相对阈值（例如 10％）则过程 200 从同义词映射表的每一个变体移除该任何语言（步骤 260）。相同的阈值应用于所有变体和所有语言。也从同义词映射表移除不与至少一种语言相关联的任何变体（步骤 270）。

[0048] 为了说明性的目的，过程 200 已被描述为例如通过从现有同义词映射表移除条目或变体来改变该现有同义词映射表的过程。替选地，在同义词映射表的初始构造期间通过首先不包括某些条目或变体可以获得相同的效果。

[0049] 在图 4 中示出了说明性的示例同义词映射表。该图示假设库由四种语言表现：英语、法语、罗马印度语和孟加拉语。该映射表包含三个键：“elephant”、“liberte”和“nityananda”。每一个键与多个变体相关联。具体地，变体“nity.a-nanda”(410) 在来自库的被识别为罗马印度语和孟加拉语的文档中出现。然而，该变体在每一种语言中仅出现 6 次。如果为每一种语言指定了大于 6 的绝对阈值，则将从同义词映射表中移除这些语言和变体。

[0050] 变体“**nityānānda**”在三种语言中出现 (430)，根据语言的相对频度，与每一种语言

中的其它变体相比较该变体相对较小。如果应用 10％的相对频度阈值，则这些语言和整个变体将被从同义词映射表移除。假设相同的相对阈值用于“nityAnanda”变体，与孟加拉语(420)的关联也将被移除。该变体和其余的该变体的语言关联将保留，因为这些其它语言每一种都频繁出现足以超过假设的相对和绝对阈值。

[0051] 可以利用同义词映射表进行的有用的事之一是使用该同义词映射表来扩增对搜索引擎的查询。

[0052] 如图 3 中所示，过程 300 可以被用来扩增查询以合并来自同义词映射表的同义词。实际上，接收（步骤 310）的查询通常未完美描述用户的想要的查询。用户受输入装置的局限性和精确指示查询的语言的不便所约束。理想的同义词是反映用户在理想的环境下将提供的内容的那些词。过程 300 旨在通过对相对于查询中的单词和用户意指的语言的同义词映射表中的变体评分来逼近理想的同义词，所述用户意指的语言由查询的语言逼近。

[0053] 过程 300 确定接收了查询的界面的语言（步骤 315）。用户将查询提供给界面。该界面将具有界面语言，即界面向用户展示信息所用的语言，例如英语、法语或世界语。然而，在查询内的单词不一定用与查询被提供到的界面相同的语言。

[0054] 过程 300 识别来自查询的各个单词（步骤 320）。单词的识别依赖于查询语言的特定惯例。例如，在拉丁字体语言中，单词通过空格或其它标点（例如‘-’）分割。

[0055] 过程 300 确定查询可能是用的何种语言（步骤 325）。在一个实施方式中，以两部分来确定查询语言：确定查询是用界面的语言的可能性，例如概率；以及对于查询中的每一个词语确定该词语是用某种特定语言的可能性，例如概率。

[0056] 确定查询是否可能是用与界面语言相同的语言可以使用过去的查询来进行。如果过去的查询递送了搜索结果，则过去的查询可以基于用户随后选择的结果的语言被自动分类为用特定语言。以下假设是合理的：查询的语言与用户选择查看的文档的语言相同，尤其是如果选 择的展现包括来自搜索结果文档的摘录。也可以人工检查过去的查询来确定其语言。自动和人工技术可以被组合：已被人工分类的查询用作在自动确定期间使用的种子以提高精确度。自动分类器的结果可以告知分类器的后继调整。人工确定种子和查询分类器的调整可以被反复重复以进一步提高精确度。将当前查询整体与相同界面接收的过去查询相匹配，生成查询是用与界面语言相同的语言的可能性分值或概率。

[0057] 过程 300 确定来自查询的词语在库中在用于每一种语言的文档内出现的频度。从频度计数生成向量，该向量对于每一种语言给出词语用该语言的在 0 到 1 的范围内的可能性分值。为查询中的每一个词语生成分值向量，例如概率向量。

[0058] 例如专有名称（例如因特网）的以许多不同的语言出现的单词可能过度影响用于查询的分值向量。如果在查询词语中找到这样的单词，则所述单词的分值可被任意设置以表明该单词可能是用界面语言。替选地，这样的单词可以被忽略。

[0059] 过程 300 可以通过平滑每一个向量来进一步处理每一个向量。在一个实施方式中，在计算向量时，添加小的平滑值 s 以减少噪声。例如，如果词语 t 在语言 L 中出现 n 次并且在整个 k 种语言中出现 N 次，则该词语是用该语言的概率被平滑为 $P(L|t) = (n+s)/(k\times s+N)$，而非 $P(L|t) = n/N$。平滑值可以根据 N 和 k 的大小来选择。例如，s 可被选择以随着 N 增大而增大并且随着 k 增大而减小。

[0060] 将来自先前步骤的所有向量相乘。合成向量与查询是用界面的语言的概率（或分

值）相乘，产生查询概率（或分值）向量。该查询概率向量包含对于每一种语言，查询是用该语言的概率（或分值）。将具有最大概率（或分值）的语言选择为归属于该查询的查询语言。

[0061] 过程 300 简化查询中的每一个单词（步骤 330）。在简化每一个单词中，过程叠缩连字、移除重音以及对每一个单词中的字符标音。这以与如上所述的从训练库获取普通形式完全相同的方式来完成。然而，在此使用来简化查询单词的特定转换映射表在某些方面与在创建同义词映射表中使用的转换映射表不同。具体地，简化每一个单词通常独立于语言。

[0062] 然而，在特定情况中，所识别的查询语言可以影响如何简化查询单词。当单词简化的结果在查询语言中无意义时这尤其重要。例如，在土耳其语中‘ue’是用于‘ü’的无意义替代物，与德语中不同。对于土耳其语用户将“Türk”简化为“Tuerk”将是不希望的。

[0063] 通常，来自查询的简化单词被用来使用作为键的每一个简化单词从同义词映射表查找和检索变体（步骤 340）。每一个变体都是原始查询单词的潜在同义词。在每一种语言内的每一个变体的键下的相对频度被用来估计该变体是否被期望作为用于每一种语言中的键的同义词（步骤 350）。该估计通过对以下乘积求和来计算：对于每一种语言，查询是用该语言的概率与在该语言中的变体的相对频度相乘。例如，考虑当“éléphant”在英语中 52％的次数是变体而在法语中 77％的次数是变体时。然后对于查询，被确定可能是用英语具有 70％的概率而被确定可能是用法语具有 30％的概率，用于“éléphant”的合成估计为：52％ ×70％ +77％ ×30％＝ 59.5％。如果所计算的估计超过同义词概率阈值（例如 50％），则该变体被选择来扩增查询。给定同义词映射表中的语言统计和查询语言分类器提供的概率，选择特定同义词概率阈值来提供优良结果。在变体在给定语言中是可叠缩连字的结果的特殊情况中，则在计算该变体的估计时降低该变体的相对频度（例如变为四分之一）。对变体的相对频度的这种惩罚反映不恰当地叠缩了变体的连字的潜在风险。

[0064] 将每一个所选择的变体添加到查询（步骤 360），除非变体是无 用词以及变体在可能的查询语言中未出现：这样的变体被忽略。用来自查询的每一个原始词语的每一个所选择的变体来扩增该原始词语。每一个变体作为与原始词语的联合被附加。例如，查询“eléphant trunk”被扩增为“(eléphant or elephant or éléphant)trunk”，其中假设 elephant 和 éléphant 都被选择为用于 eléphant 的变体。

[0065] 过程使用已扩增的查询来搜索搜索库（步骤 370）。搜索库包含处于其原始、未改变的形式的文档。除扩增查询的影响外，从库搜索并提供结果不会另外受影响。

[0066] 如果可能的查询语言是搜索库中未被充分表现的语言（即，全部文档的很小比例），则可能不希望包括来自同义词映射表的变体。将变体添加到搜索查询增加了与来自所希望的语言外的文档相匹配的风险，潜在地使结果中充斥了大量其它语言的文档。然而，当原始查询单词仅包含非重读的字母并且没有包含可叠缩的连字（例如，“ueber”，被简化为“uber”）时，则应当不考虑可能的查询语言来寻找变体。在一个实施方式中，包括变体的决定取决于界面语言而非查询语言。

[0067] 图 5A 至图 34 示出了用来映射训练库中的单词或用来简化搜索查询中的单词的转换映射表的一个实施方式。每一个图示出了一个或多个转换映射表的命名组。每一个转换映射表被示出为图中一列中的一行。转换映射表被示为至少具有与如上所述的输入字符和

输出字符。另外，标记为“UCS”的列根据通用字符集 (UCS) 示出了字符的编码的十六进制值。当未给出 UCS 值时，每一个字符是 95 个可印刷的 ASCII 字符中的一个。

[0068] 根据便捷或惯例而非必要来管制转换映射表的分组：一个或多个转换映射表组可以构成用于特定语言的特定于语言的映射。用于特定语言的组的组合可以取决于所述组是否被用来映射训练库中的单词或用来简化查询中的单词。

[0069] 图 5A、5B 和 5C 示出了通用转换映射表组。通常，这些是不可能与关于特定语言的转换映射表相冲突的安全转换映射表。

[0070] 图 6 示出了俄语转换映射表组。该组被用来在同义词映射表的生成期间映射来自俄语文档的单词。

[0071] 图 7 示出了马其顿语转换映射表组。该组被用来在同义词映射表的生成期间映射来自马其顿语文档的单词。

[0072] 图 8 示出了乌克兰语转换映射表组。该组被用来在同义词映射表的生成期间映射来自乌克兰语文档的单词。

[0073] 图 9 示出了希腊语转换映射表组。该组被用来在同义词映射表的生成期间映射来自希腊语文档的单词。

[0074] 如图 10 和图 11 中所示，某些转换映射表也指定叠缩的连字的重读等价体（在图中加标题“A.E.”的列）。这些映射表具有两个字符输入（即可叠缩的连字）和一个输出（叠缩的连字）。该信息可以被用来确定两个字符（输入）是否是可叠缩的连字。替选地，该信息也指示特定字符（输出）是否可能是可叠缩的连字的结果。

[0075] 图 10 示出了世界语 H/X- 体系转换映射表组。该组被用来在同义词映射表的生成期间映射来自世界语文档的单词。

[0076] 图 11 示出了 Ch 和 ShZh 转换映射表组。该组在同义词映射表的生成和查询词语简化期间与其它组相组合。

[0077] 图 12 示出了克罗地亚语转换映射表组。该组被用来在同义词映射表的生成期间映射来自克罗地亚语文档的单词。通用、Ch、ShZh、A- 元音变音、O- 元音变音、U- 元音变音和 Y- 元音变音组被组合并且被用来简化被识别为克罗地亚语的查询词语。A- 元音变音、O- 元音变音、U- 元音变音和 Y- 元音变音组将在下面参考图 23 描述。

[0078] 图 13 示出了加泰罗尼亚语转换映射表组。该组被用来在同义词映射表的生成期间映射来自加泰罗尼亚语文档的单词。

[0079] 图 14 示出了塞尔维亚语转换映射表组。该组与克罗地亚语组相组合并且被用来在同义词映射表的生成期间映射来自塞尔维亚语文档的单词。通用、A- 元音变音、O- 元音变音、U- 元音变音、Y- 元音变音、Ch、ShZh 和塞尔维亚语组被组合并且被用来简化被识别为塞尔维亚语的查询词语。

[0080] 图 15 示出了法语转换映射表组。该组被用来在同义词映射表的生成期间映射来自法语文档的单词。

[0081] 图 16 示出了意大利语转换映射表组。该组被用来在同义词映射表的生成期间映射来自意大利语文档的单词。

[0082] 图 17 示出了葡萄牙语转换映射表组。该组被用来在同义词映射表的生成期间映射来自葡萄牙语文档的单词。

[0083] 图 18 示出了罗马尼亚语转换映射表组。该组被用来在同义词映射表的生成期间映射来自罗马尼亚语文档的单词。

[0084] 图 19 示出了西班牙语转换映射表组。该组被用来在同义词映射表的生成期间映射来自西班牙语文档的单词。

[0085] 图 20 示出了荷兰语转换映射表组。该组被用来在同义词映射表的生成期间映射来自荷兰语文档的单词。通用、A- 元音变音、O- 元音变 音、U- 元音变音和荷兰语 -Y 组被组合并且被用来简化被识别为荷兰语的查询词语。

[0086] 图 21 示出了丹麦语转换映射表组。该组被用来在同义词映射表的生成期间映射来自丹麦语文档的单词。

[0087] 图 22 示出了英语转换映射表组。该组被用来在同义词映射表的生成期间映射来自英语文档的单词。

[0088] 图 22 还示出了德语转换映射表组。该组被用来在同义词映射表的生成期间映射来自德语文档的单词。通用、Y- 元音变音和德语元音变音组被用来简化被识别为德语的查询词语。

[0089] 图 22 还示出了荷兰语 -Y 转换映射表组。该组与其它组相组合来简化被识别为荷兰语的查询词语。

[0090] 图 22 还示出了德语元音变音转换映射表组。该组与其它组相组合来简化被识别为德语的查询词语。

[0091] 图 22 还示出了瑞典语转换映射表组。该组被用来在同义词映射表的生成期间映射来自瑞典语文档的单词。通用、U- 元音变音和 Y- 元音变音组被用来简化被识别为瑞典语或芬兰语的查询词语。

[0092] 图 23 示出了四个组 :A- 元音变音、O- 元音变音、U- 元音变音和 Y- 元音变音组。这些组被用来与其它组相组合以简化查询词语。

[0093] 图 24 示出了冰岛语转换映射表组。该组被用来在同义词映射表的生成期间映射来自冰岛语文档的单词。

[0094] 图 25 示出了捷克语转换映射表组。该组与 ShZh 组相组合并且被 用来在同义词映射表的生成期间映射来自捷克语文档的单词。通用、A- 元音变音、O- 元音变音、U- 元音变音、Y- 元音变音和 ShZh 组被用来简化被识别为捷克语的查询词语。

[0095] 图 26 示出了拉脱维亚语转换映射表组。该组与 Ch 和 ShZh 组相组合并且被用来在同义词映射表的生成期间映射来自拉脱维亚语文档的单词。通用、A- 元音变音、O- 元音变音、U- 元音变音、Y- 元音变音、Ch 和 ShZh 组被用来简化被识别为拉脱维亚语的查询词语。

[0096] 图 27 示出了立陶宛语转换映射表组。该组与 Ch 和 ShZh 组相组合并且被用来在同义词映射表的生成期间映射来自立陶宛语文档的单词。通用、A- 元音变音、O- 元音变音、U- 元音变音、Y- 元音变音、Ch 和 ShZh 组被用来简化被识别为立陶宛语的查询词语。

[0097] 图 28 示出了波兰语转换映射表组。该组被用来在同义词映射表的生成期间映射来自波兰语文档的单词。

[0098] 图 29 示出了斯洛伐克语转换映射表组。该组与 ShZh 组相组合并且被用来在同义词映射表的生成期间映射来自斯洛伐克语文档的单词。通用、A- 元音变音、O- 元音变音、

U- 元音变音、Y- 元音变音和 ShZh 组被组合并且被用来简化被识别为斯洛伐克语的查询词语。

[0099] 图 30 示出了斯洛文尼亚语转换映射表组。该组与 Ch 和 ShZh 组相组合并且被用来在同义词映射表的生成期间映射来自斯洛文尼亚语文档的单词。

[0100] 图 31 示出了爱沙尼亚语转换映射表组。该组与 Ch 和 ShZh 组相组合并且被用来在同义词映射表的生成期间映射来自爱沙尼亚语文档的单词。通用、A- 元音变音、O- 元音变音、U- 元音变音、Y- 元音变音、Ch 和 ShZh 组被组合并且被用来简化被识别为爱沙尼亚语的查询词语。

[0101] 图 32 示出了匈牙利语转换映射表组。该组被用来在同义词映射表的生成期间映射来自匈牙利语文档的单词。

[0102] 图 33 示出了世界语转换映射表组。该组与世界语 HX- 体系组相组合并且被用来在同义词映射表的生成期间映射来自世界语文档的单词。通用、A- 元音变音、O- 元音变音、U- 元音变音、Y- 元音变音和世界语 HX- 体系组被组合并且被用来简化被识别为世界语的查询词语。

[0103] 图 34 示出了土耳其语转换映射表组。该组被用来在同义词映射表的生成期间映射来自土耳其语文档的单词。

[0104] 下面的表示出了哪些转换映射表组可以被用来在同义词映射表的生成期间映射单词。每一种语言被指定了其字符黑名单（如上所述）和一个或多个转换映射表组，所述转换映射表组一起构成在从训练库中的单词获得普通形式时使用的一套转换映射表。

[0105]

语言	字符黑名单	转换映射表
加泰罗尼亚语	kw	加泰罗尼亚语
法语		法语
意大利语	jkwxy	意大利语
葡萄牙语	kw	葡萄牙语
罗马尼亚语	kqwy	罗马尼亚语
西班牙语	w	西班牙语
丹麦语		丹麦语
荷兰语		荷兰语
英语		英语
德语		德语
冰岛语	cqw	冰岛语
瑞典语		瑞典语
爱沙尼亚语	qwxy	Ch、ShZh、爱沙尼亚语
芬兰语	bcfqwxz	
匈牙利语	qwxy	匈牙利语
希腊语		希腊语
土耳其语	qwx	土耳其语
克罗地亚语	qwxy	克罗地亚语
捷克语	qwx	ShZh、捷克语
拉脱维亚语	qwxy	Ch、ShZh、拉脱维亚语
立陶宛语	qwxy	Ch、ShZh、立陶宛语
马其顿语		马其顿语
挪威语		丹麦语
波兰语	qvx	波兰语
俄语		俄语
塞尔维亚语	qwxy	克罗地亚语、塞尔维亚语
斯洛伐克语	qw	ShZh、斯洛伐克语
斯洛文尼亚语	qwxy	Ch、ShZh、斯洛文尼亚语
乌克兰语		乌克兰语
世界语	qwxy	世界语（hx-体系）、世界语

[0106] 图 35 是接收多语言查询并且作为响应提供多语言结果的搜索引擎 3550 的示意图。系统 3550 通常被配置来从各种源获取与词语的出现和频度有关的信息，并且基于在这

样的源中的单词使用的分析响应于查询生成搜索结果。这样的源可以包括例如在因特网上找到的多语言文档和文件。

[0107] 系统 3550 包括一个或多个界面 3552，其中每一个用不同的语言。界面允许用户使用搜索引擎的服务并且允许用户与搜索引擎的服务相交互。具体地，界面从用户接收查询。查询包括一列单词，其中每一个单词可以用任何的语言。查询中的单词不需要用界面的语言。接收用户的查询的特定界面 3552 取决于对界面的用户的选择。

[0108] 系统 3550 可以被通信地连接到诸如因特网 3558 的网络，并且因此可以与连接到因特网的各种装置通信，所述装置诸如无线通信装置 3562 和个人计算机 3564。用于任何装置的通信流可以是双向的，以使系统 3550 从装置接收信息（例如，查询或文档的内容）并且也可以将信息（例如结果）发送到装置。

[0109] 界面 3552 接收的查询被提供给查询处理器 3566。查询处理器 3566 处理查询、可选地扩增查询并且将查询传递给系统 3550 的另一个组件。例如，查询处理器 3566 可以促使检索系统 3570 生成与查询相对应的搜索结果。这样的检索系统 3570 可以使用如 Google PageRank™ 系统使用的数据检索和搜索技术。检索系统 3570 生成的结果然后可以被提供回原始查询装置。

[0110] 系统 3550 为了其适当的操作可以依靠多个其它的组件。例如，每当发出请求时系统 3550 参考文档的搜索库 3572。搜索库可以被索引以使搜索更有效。使用从在 Web 上找到的文档（例如，通过 web 爬行器）收集的信息可以填增搜索库。文档也可以被存储在训练库 3574 中用于以后处理。

[0111] 训练库 3574 可以由同义词处理器 3580 处理。同义词处理器 3580 可以从训练库 3574 生成同义词映射表 3585。同义词映射表 3585 可以由查询处理器 3566 使用来用同义词扩增搜索查询。

[0112] 在本说明书中描述的本发明的实施例和所有功能性操作可以在数字电子电路中、或在计算机软件、固件或硬件（包括在本说明书中公开的结构以及其结构等价体中）或在上述中的一个或多个的组合中来实现。本发明的实施例可以作为一个或多个计算机程序产品来实现，所述计算机程序产品即用于被数据处理设备执行或控制数据处理设备的操作的编码在计算机可读介质上的计算机程序指令的一个或多个模块。计算机可读介质可以是机器可读存储装置、机器可读存储基片、存储装置、实现机器可读传播信号的物质的合成物或上述中的一个或多个的组合。术语“数据处理设备”涵盖用于处理数据的所有设备、装置和机器，以示例的方式包括可编程处理器、计算机、或多处理器或计算机。除硬件外，设备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，例如构成处理器固件、协议栈、数据库管理系统、操作系统或上述中的一个或多个的组合的代码。传播信号是人为生成的信号，例如机器生成的电、光或电磁信号，其被生成来对信息编码用以传输到适当的接收者设备。

[0113] 计算机程序（也被称为程序、软件、软件应用、脚本或代码）可以以任何形式的编程语言来编写，包括编译或解释语言，并且其可以以任何形式来部署，包括作为单机程序或作为适合于在计算环境中使用的模块、组件、子程序或其它单元。计算机程序不一定与文件系统中的文件相对应。程序可以被存储在保存其它程序或数据的文件的一部分中（例如，存储在标记语言文档中的一个或多个脚本）、被存储在专用于正讨论的程序的单个文件中、或被存储在多个同等文件（例如，存储一个或多个模块、子程序或代码部分的文件）中。计

算机程序可以被部署来在一个计算机上或在多个计算机上执行，所述多个计 算机位于一个地点或分布在多个地点并且通过通信网络相互连接。

[0114] 在本说明书中描述的过程和逻辑流程可以由一个或多个可编程的处理器来执行，所述一个或多个可编程的处理器执行一个或多个计算机程序以通过操作输入数据并且生成输出来执行功能。过程和逻辑流程也可以由专用逻辑电路来执行，并且设备也可以被实现为专用逻辑电路，所述专用逻辑电路例如 FPGA（现场可编程门阵列）或 ASIC（专用集成电路）。

[0115] 适合于执行计算机程序的处理器包括，以示例的方式，通用和专用微处理器以及任何种类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其两者接收指令和数据。计算机的基本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装置。通常，计算机也将包括用于存储数据的一个或多个海量存储装置，或操作地连接以从所述一个或多个海量存储装置接收数据或将数据转送到所述一个或多个海量存储装置，或两者，所述海量存储装置例如是磁、磁光盘或光盘。然而，计算机不需要具有这样的装置。此外，计算机可以被嵌入另一个装置中，所述装置例如是移动电话、个人数字助理 (PDA)、移动音频播放器、全球定位系统 (GPS) 接收器，仅指出了一些。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、介质和存储装置，以示例的方式包括例如 EPROM、EEPROM 和闪存装置的半导体存储装置；磁盘，例如内部硬盘或可移动盘；磁光盘；以及 CD-ROM 和 DVD-ROM 盘。处理器和存储器可以由专用逻辑电路增补，或并入到专用逻辑电路中。

[0116] 为了提供与用户的交互，本发明的实施例可以在具有以下装置的计算机上实现：用于向用户显示信息的显示装置，例如 CRT（阴极射线管）或 LCD（液晶显示）监视器，和通过其用户可以向计算机提供输入的键盘和例如鼠标或跟踪球的指向装置。其它种类的装置也可以 被用来提供与用户的交互；例如，提供给用户的反馈可以是任何形式的感官反馈，例如视觉反馈、听觉反馈或触觉反馈；以及可以以任何形式接收来自用户的输入，包括声音、语音或触觉输入。

[0117] 本发明的实施例可以在计算系统中实现，所述计算系统包括例如作为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，或包括前端组件，例如通过其用户可以与本发明的实施方式相交互的具有图形用户界面或 Web 浏览器的客户端计算机，或一个或多个这样的后端、中间件或前端组件的任何组合。系统的组件可以通过任何形式或介质的数字数据通信来相互连接，所述数字数据通信例如是通信网络。通信网络的示例包括局域网（“LAN”）和例如因特网的广域网（“WAN”）。

[0118] 计算系统可以包括客户端和服务器。客户端和服务器通常相互远离并且典型地通过通信网络相交互。客户端和服务器的关系由在各个计算机上运行并且相互间具有客户端 - 服务器关系的计算机程序产生。

[0119] 虽然本发明包含许多细节，但是这些不应当被解释为对本发明或对可能主张的权利的范围的限制，而是作为具体到本发明的特定实施例的特征的描述。在本说明书中在不同实施例的上下文中描述的某些特征也可以组合到单个实施例中实现。反之，在单个实施例的上下文中描述的各种特征也可以在多个实施例中分别实现或在任何适当的子组合中实现。此外，尽管特征在上面可能被描述为在某些组合中起作用并且甚至最初主张为如此，

但是来自所主张的组合的一个或多个特征在某些情况中可以从组合中删除，并且所主张的组合可以被导向到子组合或子组合的变形。

[0120] 类似地，虽然在附图中以特定的次序来描述操作，但是不应当理解为需要按示出的特定次序或按顺序次序来执行这样的操作、或需要执行所有示出的操作来实现希望的结果。在某些情况下，多任务和并 行处理可以是有利的。此外，如上所述的实施例中的各种系统组件的分离不应当理解为在所有的实施例中都需要这样的分离，并且应当理解，所描述的程序组件和系统通常可以被共同集成在单个软件产品中或被封装入多个软件产品中。

[0121] 因此，描述了本发明的特定实施例。其它的实施例在下面的权利要求的范围内。例如，在权利要求中陈述的行为可以以不同的次序执行并且仍实现希望的结果。

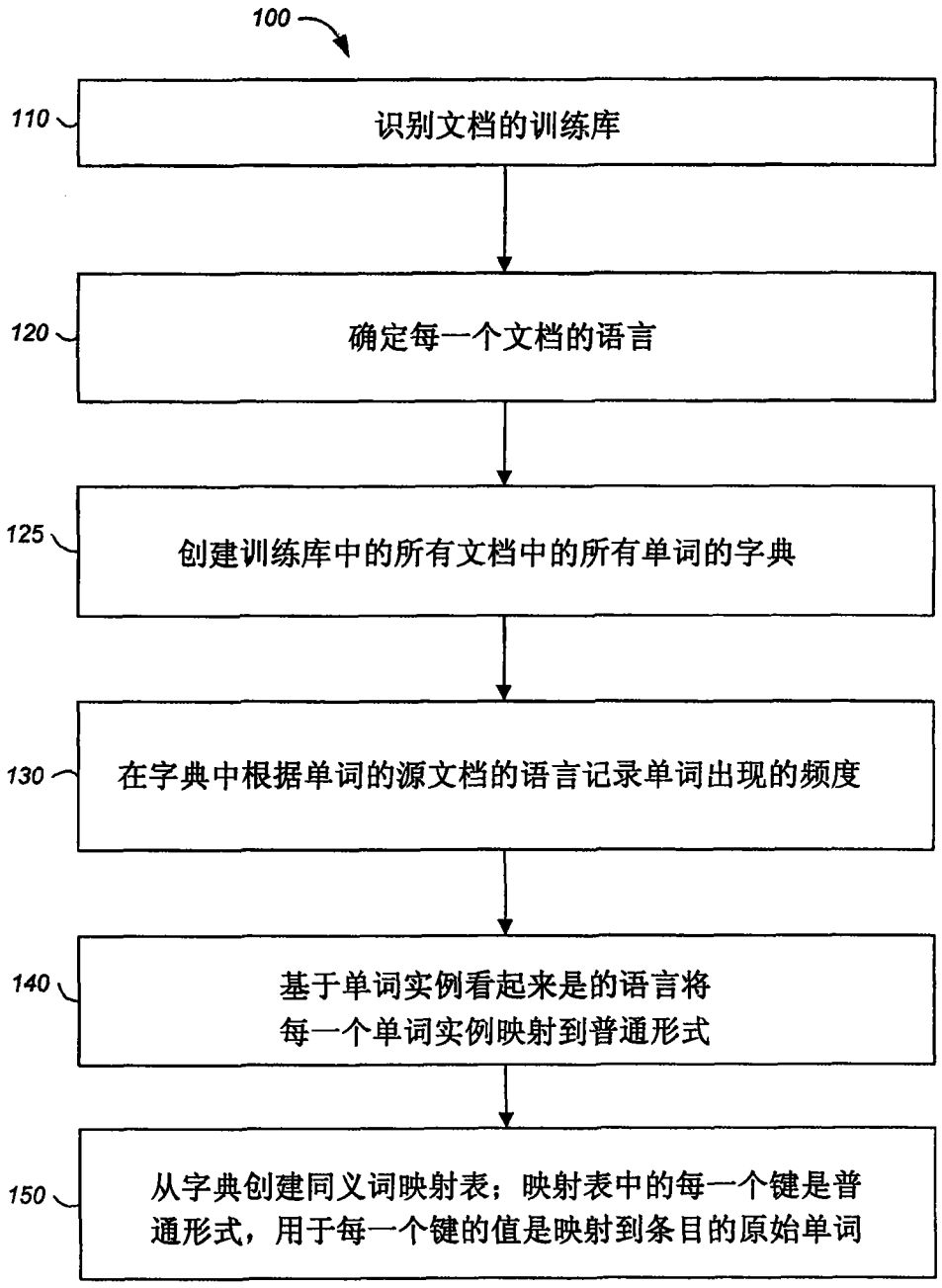

100
110
识别文档的训练库
120
确定每一个文档的语言
125
创建训练库中的所有文档中的所有单词的字典
130
在字典中根据单词的源文档的语言记录单词出现的频度
140
基于单词实例看起来是的语言将
每一个单词实例映射到普通形式
150
从字典创建同义词映射表；映射表中的每一个键是普
通形式，用于每一个键的值是映射到条目的原始单词

图 1

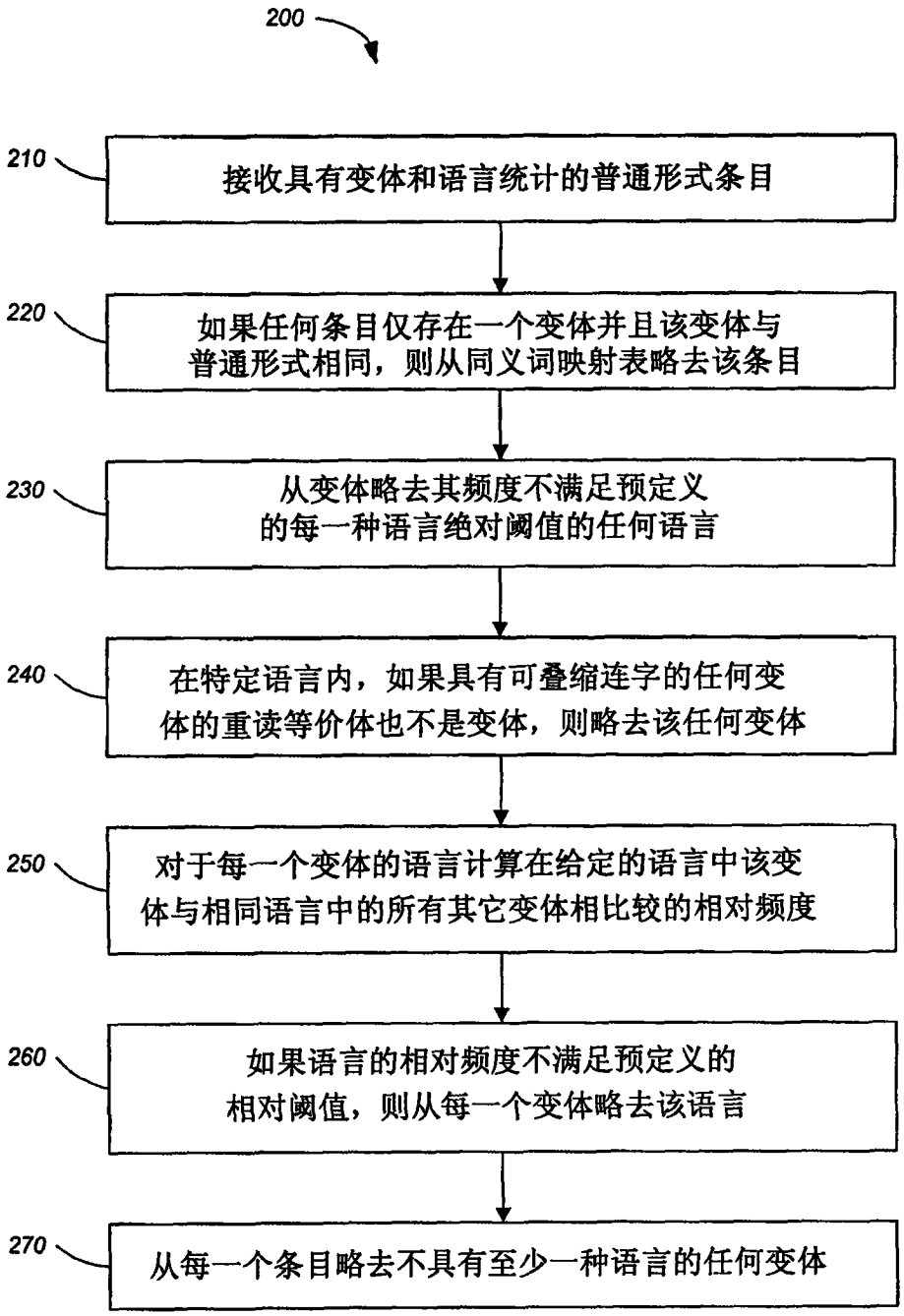

200
210
接收具有变体和语言统计的普通形式条目
220
如果任何条目仅存在一个变体并且该变体与普通形式相同，则从同义词映射表略去该条目
230
从变体略去其频度不满足预定义的每一种语言绝对阈值的任何语言
240
在特定语言内，如果具有可叠缩连字的任何变体的重读等价体也不是变体，则略去该任何变体
250
对于每一个变体的语言计算在给定的语言中该变体与相同语言中的所有其它变体相比较的相对频度
260
如果语言的相对频度不满足预定义的相对阈值，则从每一个变体略去该语言
270
从每一个条目略去不具有至少一种语言的任何变体

图 2

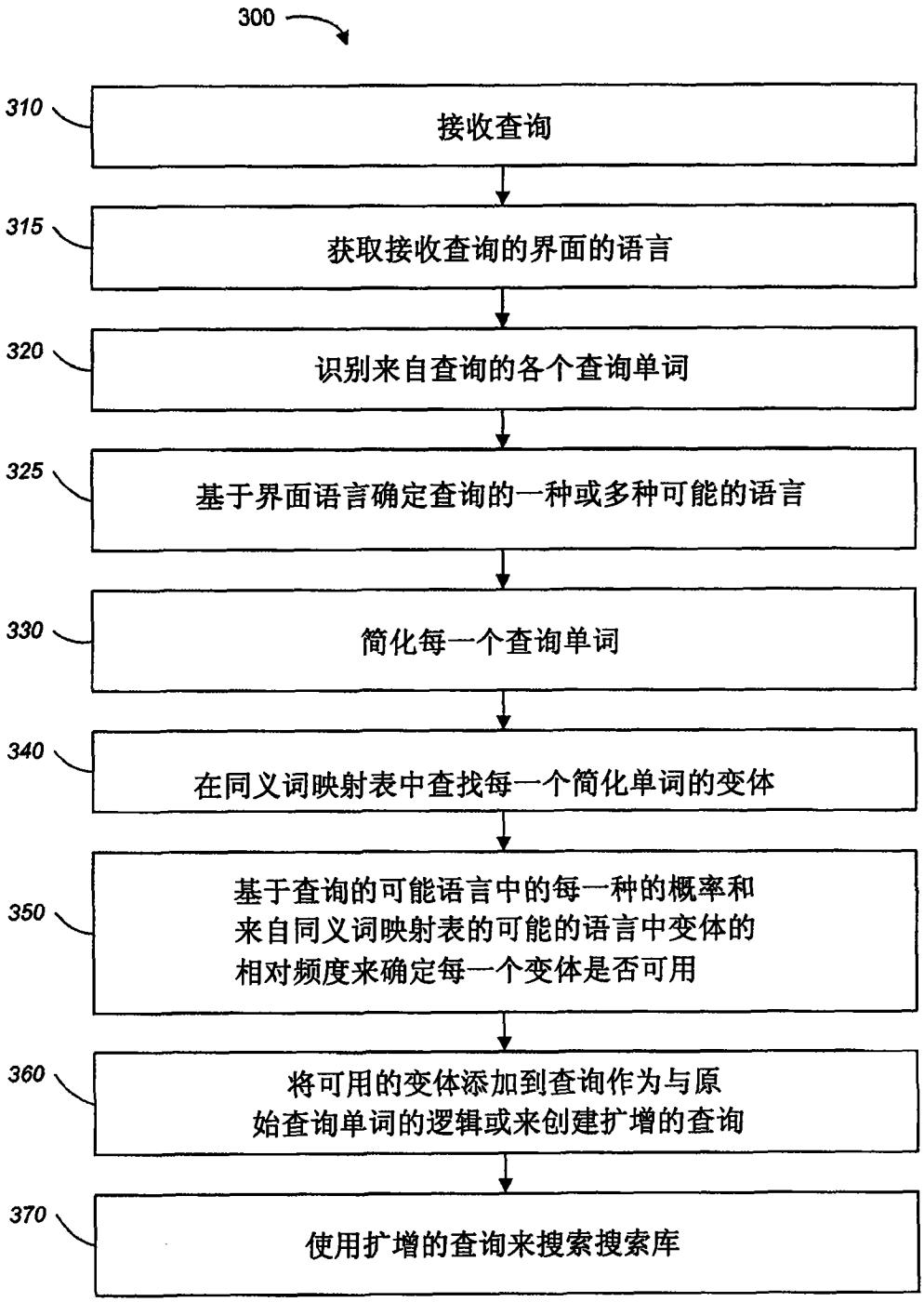

300
310
接收查询
315
获取接收查询的界面的语言
320
识别来自查询的各个查询单词
325
基于界面语言确定查询的一种或多种可能的语言
330
简化每一个查询单词
340
在同义词映射表中查找每一个简化单词的变体
350
基于查询的可能语言中的每一种的概率和
来自同义词映射表的可能的语言中变体的
相对频度来确定每一个变体是否可用
360
将可用的变体添加到查询作为与原
始查询单词的逻辑或来创建扩增的查询
370
使用扩增的查询来搜索搜索库

图 3

键（普通形式）	变体	语言	相对频度	频度
elephant	éléphant	英语	52%	100
		法语	77%	1000
	eléphant	英语	48%	90
		法语	23%	300
liberte	liberte	英语	58%	550
		法语	29%	650
	liberté	英语	42%	400
		法语	71%	1600
nityananda	nityananda	英语	51%	200
		孟加拉语	10%	200
		[罗马]印度语	39%	600
410	nity.a-nanda	[罗马]印度语	1%	6
		孟加拉语	1%	6
	nityAnanda	英语	13%	50
		[罗马]印度语	53%	800
420		孟加拉语	7%	150
430	nityānanda	英语	8%	30
		[罗马]印度语	8%	120
		孟加拉语	9%	190
	নিত্যানন্দ	孟加拉语	74%	1550
		英语	28%	110

图 4

通用输入	UCS	输出											
À	00C0	A	Ê	00CA	E	İ	0130	I	Ñ	00D1	N		
Á	00C1	A	Ë	00CB	E	I	0406	I	Ń	0143	N		
Â	00C2	A	Ē	0112	E	Ï	0407	I	Ņ	0145	N		
Ã	00C3	A	Ė	0116	E	ì	00EC	i	Ň	0147	N		
Ā	0100	A	Ę	0118	E	í	00ED	i	ñ	00F1	n		
Ă	0102	A	Ě	011A	E	î	00EE	i	ń	0144	n		
Ą	0104	A	è	00E8	e	ï	00EF	i	ņ	0146	n		
à	00E0	a	é	00E9	e	ī	012B	i	ň	0148	n		
á	00E1	a	ê	00EA	e	į	012F	i	Ò	00D2	O		
â	00E2	a	ë	00EB	e	ı	0131	i	Ó	00D3	O		
ã	00E3	a	ē	0113	e	i	0456	i	Ô	00D4	O		
ā	0101	a	ė	0117	e	ï	0457	i	Õ	00D5	O		
ă	0103	a	ę	0119	e	Ĵ	0134	J	Ő	0150	O		
ą	0105	a	ě	011B	e	J	0408	J	ò	00F2	o		
Ç	00C7	C	Ĝ	011C	G	ĵ	0135	j	ó	00F3	o		
Ć	0106	C	Ğ	011E	G	j	0458	j	ô	00F4	o		
Ĉ	0108	C	Ģ	0122	G	Ķ	0136	K	õ	00F5	o		
Č	010C	C	ĝ	011D	g	ķ	0137	k	ő	0151	o		
ç	00E7	c	ğ	011F	g	ĸ	0139	L	Ŕ	0154	R		
ć	0107	c	ģ	0123	g	Ļ	013B	L	Ŗ	0156	R		
ĉ	0109	c	Ĥ	0124	H	Ľ	013D	L	Ř	0158	R		
č	010D	c	ĥ	0125	h	L·	013F	L	ŕ	0155	r		
Ð	00D0	D	Ì	00CC	I	Ł	0141	L	ŗ	0157	r		
Ď	010E	D	Í	00CD	I	Í	013A	l	ř	0159	r		
ð	00F0	d	Î	00CE	I	ļ	013C	l	Ś	015A	S		
đ	010F	d	Ï	00CF	I	Ľ	013E	l	Ŝ	015C	S		
È	00C8	E	Ī	012A	I	l·	0140	l	Ş	015E	S		
É	00C9	E	Į	012E	I	ł	0142	l	Š	0160	S		
									Ş	0218	S		

图 5A

通用（续上页）

输入	UCS	输出						
ś	015B	s	Ý	00DD	Y	ǌ	01CC	nj
ŝ	015D	s	Ĳ	0132	Y	Ø	00D8	Oe
ş	015F	s	ý	00FD	y	Œ	0152	Oe
š	0161	s	ĳ	0133	y	ø	00F8	oe
ș	0219	s	Ź	0179	Z	œ	0153	oe
Ţ	0162	T	Ż	017B	Z	ß	00DF	ss
Ť	0164	T	Ž	017D	Z	Þ	00DE	Th
Ț	021A	T	ź	017A	z	þ	00FE	th
ţ	0163	t	ż	017C	z			
ť	0165	t	ž	017E	z			
ț	021B	t	Å	00C5	Aa			
Ù	00D9	U	Æ	00C6	Ae			
Ú	00DA	U	å	00E5	aa			
Û	00DB	U	æ	00E6	ae			
Ū	016A	U	Đ	0110	Dj			
Ŭ	016C	U	đ	0111	dj			
Ů	016E	U	Ǆ	01C4	Dz			
Ű	0170	U	ǅ	01F1	Dz			
Ų	0172	U	ǆ	01C5	Dz			
ù	00F9	u	ǲ	01F2	Dz			
ú	00FA	u	ǆ	01C6	dz			
û	00FB	u	ǳ	01F3	dz			
ū	016B	u	Ǉ	01C7	Lj			
ŭ	016D	u	ǉ	01C8	Lj			
ů	016F	u	ǉ	01C9	lj			
ű	0171	u	Ǌ	01CA	Nj			
ų	0173	u	ǋ	01CB	Nj			

图 5B

通用（续上页）

输入 UCS		输出 UCS	
Ά	0386	Α	0391
ά	03AC	α	03B1
ϐ	03D0	β	03B2
Έ	0388	Ε	0395
έ	03AD	ε	03B5
ϵ	03F5	ε	03B5
Ή	0389	Η	0397
ή	03AE	η	03B7
ϴ	03F4	Θ	0398
ϑ	03D1	θ	03B8
Ί	038A	Ι	0399
Ϊ	03AA	Ι	0399
ΐ	0390	ι	03B9
ί	03AF	ι	03B9
ϊ	03CA	ι	03B9
ϰ	03F0	κ	03BA
Ό	038C	Ο	039F
ό	03CC	ο	03BF
ϖ	03D6	π	03C0
ϱ	03F1	ρ	03C1
ς	03F2	ς	03C2
Ϲ	03F9	Σ	03A3
Ύ	038E	Υ	03A5
Ϋ	03AB	Υ	03A5
ϒ	03D2	Υ	03A5
ϓ	03D3	Υ	03A5
ϔ	03D4	Υ	03A5
ΰ	03B0	υ	03C5
ϋ	03CB	υ	03C5
ύ	03CD	υ	03C5
ϕ	03D5	φ	03C6
Ώ	038F	Ω	03A9
ώ	03CE	ω	03C9
Ѓ	0403	Г	0413
Ґ	0490	Г	0413
ѓ	0453	г	0433
ґ	0491	г	0433
Ё	0401	Е	0415
Є	0404	Е	0415
ё	0451	е	0435
є	0454	е	0435
Ќ	040C	К	041A
ќ	045C	к	043A
Љ	0409	Л	041B
љ	0459	л	043B
Ц	040F	Ц	0426
ц	045F	ц	0446

图 5C

俄语

输入 UCS		输出 UCS	
Ё	0401	Е	0415
ё	0451	е	0435

图 6

马其顿语

输入	UCS	输出	UCS
Ј	0408	J	004A
ј	0458	j	006A
Ѓ	0403	Г	0413
ѓ	0453	г	0433
Ќ	040C	К	041A
ќ	045C	к	043A
Љ	0409	Л	041B
љ	0459	л	043B
Џ	040F	Ц	0426
џ	045F	ц	0446

图 7

乌克兰语

输入	UCS	输出	UCS
Ї	00CF	I	0049
І	0406	I	0049
Ї	0407	I	0049
ї	00EF	I	0069
і	0456	I	0069
ї	0457	I	0069
Ґ	0490	Г	0413
ґ	0491	г	0433
Є	0404	Е	0415
є	0454	е	0435

图 8

希腊语

输入	UCS	输出	UCS
Ά	0386	Α	0391
ά	03AC	α	03B1
ϐ	03D0	β	03B2
Έ	0388	Ε	0395
έ	03AD	ε	03B5
ϵ	03F5	ε	03B5
Ή	0389	Η	0397
ή	03AE	η	03B7
Θ	03F4	Θ	0398
ϑ	03D1	θ	03B8
Ί	038A	Ι	0399
Ϊ	03AA	Ι	0399
ΐ	0390	ι	03B9
ί	03AF	ι	03B9
ϊ	03CA	ι	03B9
ϰ	03F0	κ	03BA
Ό	038C	Ο	039F
ό	03CC	ο	03BF
ϖ	03D6	π	03C0
ϱ	03F1	ρ	03C1
ς	03F2	ς	03C2
ϲ	03F9	Σ	03A3
Ύ	038E	Υ	03A5
Ϋ	03AB	Υ	03A5
ϒ	03D2	Υ	03A5
ϓ	03D3	Υ	03A5
ϔ	03D4	Υ	03A5
ΰ	03B0	υ	03C5
ϋ	03CB	υ	03C5
ύ	03CD	υ	03C5
ϕ	03D5	φ	03C6
Ώ	038F	Ω	03A9
ώ	03CE	ω	03C9

图 9

世界语（hx-体系）

输入	输出	A.E.	UCS				
CH	C	Ĉ	0108	CX	C	Ĉ	0108
Ch	C	Ĉ	0108	Cx	C	Ĉ	0108
ch	c	ĉ	0109	cx	c	ĉ	0109
GH	G	Ĝ	011C	GX	G	Ĝ	011C
Gh	G	Ĝ	011C	Gx	G	Ĝ	011C
gh	g	ĝ	011D	gx	g	ĝ	011D
HH	H	Ĥ	0124	HX	H	Ĥ	0124
Hh	H	Ĥ	0124	Hx	H	Ĥ	0124
hh	h	ĥ	0125	hx	x	ĥ	0125
JH	J	Ĵ	0134	JX	J	Ĵ	0134
Jh	J	Ĵ	0134	Jx	J	Ĵ	0134
jh	j	ĵ	0135	jx	j	ĵ	0135
SH	S	Ŝ	015C	SX	S	Ŝ	015C
Sh	S	Ŝ	015C	Sx	S	Ŝ	015C
sh	s	ŝ	015D	sx	s	ŝ	015D

图 10

Ch

输入	输出	A.E.	UCS
CH	C	Č	010C
Ch	C	Č	010C
Ch	c	č	010D

ShZh

输入	输出	A.E.	UCS
SH	S	Š	0160
Sh	S	Š	0160
sh	s	š	0161
ZH	Z	Ž	017D
Zh	Z	Ž	017D
zh	z	ž	017E

图 11

克罗地亚语

输入	UCS	输出
Ć	0106	C
Č	010C	C
ć	0107	c
č	010D	c
Š	0160	S
š	0161	s
Ž	017D	Z
ž	017E	z
Đ	0110	Dj
đ	0111	dj
DŽ	01C4	Dz
Dž	01C5	Dz
dž	01C6	dz
LJ	01C7	Lj
Lj	01C8	Lj
lj	01C9	lj
NJ	01CA	Nj
Nj	01CB	Nj
nj	01CC	nj

图 12

加泰罗尼亚语

输入	UCS	输出
À	00C0	A
à	00E0	a
Ç	00C7	C
ç	00E7	c
È	00C8	E
É	00C9	E
è	00E8	e
é	00E9	e
Í	00CD	I
Ï	00CF	I
í	00ED	i
ï	00EF	i
L·	013F	L
l·	0140	l
Ò	00D2	O
Ó	00D3	O
ò	00F2	o
ó	00F3	o
Ù	00D9	U
Ú	00DA	U
Ü	00DC	U
ù	00F9	u
ú	00FA	u
ü	00FC	u

图 13

塞尔维亚语

输入	UCS	输出	输入	UCS	输出	输入	UCS	输出
А	0410	A	К	041A	K	Ж	0416	Z
а	0430	a	к	043A	k	З	0417	Z
Б	0411	B	Л	041B	L	ж	0436	z
б	0431	b	л	043B	l	з	0437	z
Ћ	040B	C	М	041C	M	Ђ	0402	Dj
Ц	0426	C	м	043C	m	ђ	0452	dj
Ч	0427	C	Н	041D	N	Џ	040F	Dz
ћ	045B	c	н	043D	n	џ	045F	dz
ц	0446	c	О	041E	O	Љ	0409	Lj
ч	0447	c	о	043E	o	љ	0459	lj
Д	0414	D	П	041F	P	Њ	040A	Nj
д	0434	d	п	043F	p	њ	045A	nj
Е	0415	E	Р	0420	R			
е	0435	e	р	0440	r			
Ф	0424	F	С	0421	S			
ф	0444	f	Ш	0428	S			
Г	0413	G	с	0441	s			
г	0433	g	ш	0448	s			
Х	0425	H	Т	0422	T			
х	0445	h	т	0442	t			
И	0418	I	У	0423	U			
и	0438	i	у	0443	u			
Ј	0408	J	В	0412	V			
ј	0458	j	в	0432	v			

图 14

法语

输入	UCS	输出
À	00C0	A
Á	00C1	A
Â	00C2	A
à	00E0	a
á	00E1	a
â	00E2	a
Ç	00C7	C
ç	00E7	c
È	00C8	E
É	00C9	E
Ê	00CA	E
Ë	00CB	E
è	00E8	e
é	00E9	e
ê	00EA	e
ë	00EB	e
Ì	00CC	I
Í	00CD	I
Î	00CE	I
Ï	00CF	I
ì	00EC	i
í	00ED	i
î	00EE	i
ï	00EF	i
Ò	00D2	O
Ó	00D3	O
Ô	00D4	O
ò	00F2	o
ó	00F3	o
ô	00F4	o
Ù	00D9	U
Ú	00DA	U
Û	00DB	U
Ü	00DC	U
ù	00F9	u
ú	00FA	u
û	00FB	u
ü	00FC	u
Æ	00C6	A
æ	00E6	a
Œ	0152	O
œ	0153	o

图 15

意大利语

输入	UCS	输出
À	00C0	A
à	00E0	a
È	00C8	E
è	00E8	e
Í	00CD	I
ì	00EC	i
Ò	00D2	O
ò	00F2	o
Ù	00D9	U
ù	00F9	u

图 16

葡萄牙语

输入	UCS	输出
À	00C0	A
Á	00C1	A
Â	00C2	A
Ã	00C3	A
à	00E0	a
á	00E1	a
â	00E2	a
ã	00E3	a
Ç	00C7	C
ç	00E7	c
É	00C9	E
Ê	00CA	E
é	00E9	e
ê	00EA	e
Í	00CD	I
í	00ED	i
Ó	00D3	O
Ô	00D4	O
Õ	00D5	O
ó	00F3	o
ô	00F4	o
õ	00F5	o
Ú	00DA	U
Ü	00DC	U
ú	00FA	u
ü	00FC	u

图 17

罗马尼亚语

输入	UCS	输出
Â	00C2	A
Ă	0102	A
â	00E2	a
ă	0103	a
Î	00CE	I
î	00EE	i
Ş	015E	S
Ș	0218	S
ş	015F	s
ș	0219	s
Ţ	0162	T
Ț	021A	T
ţ	0163	t
ț	021B	t

图 18

西班牙语

输入	UCS	输出
À	00C0	A
Á	00C1	A
à	00E0	a
á	00E1	a
È	00C8	E
É	00C9	E
è	00E8	e
é	00E9	e
Ì	00CC	I
Í	00CD	I
ì	00EC	i
í	00ED	i
Ñ	00D1	N
ñ	00F1	n
Ò	00D2	O
Ó	00D3	O
ò	00F2	o
ó	00F3	o
Ù	00D9	U
Ú	00DA	U
ù	00F9	u
ú	00FA	u

图 19

荷兰语

输入	UCS	输出
À	00C0	A
Á	00C1	A
Ä	00C4	A
à	00E0	a
á	00E1	a
ä	00E4	a
È	00C8	E
É	00C9	E
Ê	00CA	E
Ë	00CB	E
è	00E8	e
é	00E9	e
ê	00EA	e
ë	00EB	e
Í	00CD	I
Ï	00CF	I
í	00ED	i
ï	00EF	i
Ó	00D3	O
Ö	00D6	O
ó	00F3	o
ö	00F6	o
Ú	00DA	U
Ü	00DC	U
ú	00FA	u
ü	00FC	u
Ÿ	0178	Ij
Ĳ	0132	Ij
ÿ	00FF	ij
ĳ	0133	ij

图 20

丹麦语

输入	UCS	输出
É	00C9	E
é	00E9	e
Å	00C5	Aa
Æ	00C6	Ae
Ø	00D8	Oe
å	00E5	aa
æ	00E6	ae
ø	00F8	oe
Œ	0152	Oe
œ	0153	oe

图 21

英语

输入	UCS	输出
Æ	00C6	Ae
æ	00E6	ae
Œ	0152	Oe
œ	0153	oe

德语

输入	UCS	输出
ß	00DF	ss

荷兰语-Y

输入	UCS	输出
Ÿ	0178	Ij
ÿ	00FF	ij

德语元音变音

输入	UCS	输出
Ä	00C4	Ae
ä	00E4	ae
Ö	00D6	Oe
Ö	00F6	oe
Ü	00DC	Ue
ü	00FC	ue

瑞典语

输入	UCS	输出
É	00C9	E
é	00E9	e
Ü	00DC	U
ü	00FC	u

图 22

A-元音变音

输入	UCS	输出
Ä	00C4	A
ä	00E4	a

O-元音变音

输入	UCS	输出
Ö	00D6	O
Ö	00F6	o

U-元音变音

输入	UCS	输出
Ü	00DC	U
ü	00FC	u

Y-元音变音

输入	UCS	输出
Ÿ	0178	Y
ÿ	00FF	y

图 23

冰岛语

输入	UCS	输出
Á	00C1	A
á	00E1	a
Ð	00D0	D
ð	00F0	d
É	00C9	E
é	00E9	e
Í	00CD	I
í	00ED	i
Ó	00D3	O
Ö	00D6	O
ó	00F3	o
ö	00F6	o
Ú	00DA	U
ú	00FA	u
Ý	00DD	Y
ý	00FD	y
Æ	00C6	Ae
æ	00E6	ae
Œ	0152	Oe
œ	0153	oe
Þ	00DE	Th
þ	00FE	th

图 24

捷克语

输入	UCS	输出
Á	00C1	A
á	00E1	a
Č	010C	C
č	010D	c
Ď	010E	D
ď	010F	d
É	00C9	E
Ě	011A	E
é	00E9	e
ě	011B	e
Í	00CD	I
í	00ED	i
Ň	0147	N
ň	0148	n
Ó	00D3	O
ó	00F3	o
Ř	0158	R
ř	0159	r
Š	0160	S
š	0161	s
Ť	0164	T
ť	0165	t
Ú	00DA	U
Ů	016E	U
ú	00FA	u
ů	016F	u
Ý	00DD	Y
ý	00FD	y
Ž	017D	Z
ž	017E	z

图 25

拉脱维亚语

输入	UCS	输出
Ā	0100	A
ā	0101	a
Č	010C	C
č	010D	c
Ē	0112	E
ē	0113	e
Ģ	0122	G
ģ	0123	g
Ī	012A	I
ī	012B	i
Ķ	0136	K
ķ	0137	k
Ļ	013B	L
ļ	013C	l
Ņ	0145	N
ņ	0146	n
Ŗ	0156	R
ŗ	0157	r
Š	0160	S
š	0161	s
Ū	016A	U
ū	016B	u
Ž	017D	Z
ž	017E	z
Dž	01C4	Dz
DŽ	01F1	Dz
dž	01C5	Dz
Dz	01F2	Dz
dž	01C6	dz
dz	01F3	dz

图 26

立陶宛语

输入	UCS	输出
Ą	0104	A
ą	0105	a
Č	010C	C
č	010D	c
Ė	0116	E
Ę	0118	E
ė	0117	e
ę	0119	e
Į	012E	I
į	012F	i
Š	0160	S
š	0161	s
Ū	016A	U
Ų	0172	U
ū	016B	u
ų	0173	u
Ž	017D	Z
ž	017E	z

图 27

波兰语

输入	UCS	输出
Ą	0104	A
ą	0105	a
Ć	0106	C
ć	0107	c
Ę	0118	E
ę	0119	e
Ł	0141	L
ł	0142	l
Ń	0143	N
ń	0144	n
Ó	00D3	O
ó	00F3	o
Ś	015A	S
ś	015B	s
Ź	0179	Z
Ż	017B	Z
ź	017A	z
ż	017C	z

图 28

斯洛伐克语

输入	UCS	输出
Á	00C1	A
Ä	00C4	A
á	00E1	a
ä	00E4	a
Č	010C	C
č	010D	c
Ď	010E	D
ď	010F	d
É	00C9	E
é	00E9	e
Í	00CD	I
í	00ED	i
Ĺ	0139	L
ľ	013D	L
Í	013A	l
Ľ	013E	l
Ñ	00D1	N
Ň	0147	N
ñ	00F1	n
ň	0148	n
Ó	00D3	O
Ô	00D4	O
ó	00F3	o
ô	00F4	o
Ŕ	0154	R
ŕ	0155	r
Š	0160	S
š	0161	s
Ť	0164	T
ť	0165	t
Ú	00DA	U
ú	00FA	u
Ý	00DD	Y
ý	00FD	y
Ž	017D	Z
ž	017E	z
Ǆ	01C4	Dz
ǲ	01F1	Dz
ǅ	01C5	Dz
ʣ	01F2	Dz
ǆ	01C6	dz
ʤ	01F3	dz

图 29

斯洛文尼亚语

输入	UCS	输出
Č	010C	C
č	010D	c
Š	0160	S
š	0161	s
Ž	017D	Z
ž	017E	z

图 30

爱沙尼亚语

输入	UCS	输出
Ä	00C4	A
ä	00E4	a
Č	010C	C
č	010D	c
Õ	00D5	O
Ö	00D6	O
õ	00F5	o
ö	00F6	o
Š	0160	S
š	0161	s
Ü	00DC	U
ü	00FC	u
Ž	017D	Z
ž	017E	z

图 31

匈牙利语

输入	UCS	输出
Á	00C1	A
á	00E1	a
É	00C9	E
é	00E9	e
Í	00CD	I
í	00ED	i
Ó	00D3	O
Ô	00D4	O
Ö	00D6	O
Ő	0150	O
ó	00F3	o
ô	00F4	o
ö	00F6	o
ő	0151	o
Ú	00DA	U
Û	00DB	U
Ü	00DC	U
Ű	0170	U
ú	00FA	u
û	00FB	u
ü	00FC	u
ű	0171	u

图 32

世界语

输入	UCS 输出	
Ĉ	0108	C
ĉ	0109	c
Ĝ	011C	G
ĝ	011D	g
Ĥ	0124	H
ĥ	0125	h
Ĵ	0134	J
ĵ	0135	j
Ŝ	015C	S
ŝ	015D	s
Ŭ	016C	U
ŭ	016D	u

图 33

土耳其语

输入	UCS 输出	
Ç	00C7	C
ç	00E7	c
Ö	00D6	O
ö	00F6	o
Ğ	011E	G
ğ	011F	g
İ	0130	I
ı	0131	i
Ş	015E	S
ş	015F	s
Û	00DB	U
Ü	00DC	U
û	00FB	u
ü	00FC	u

图 34

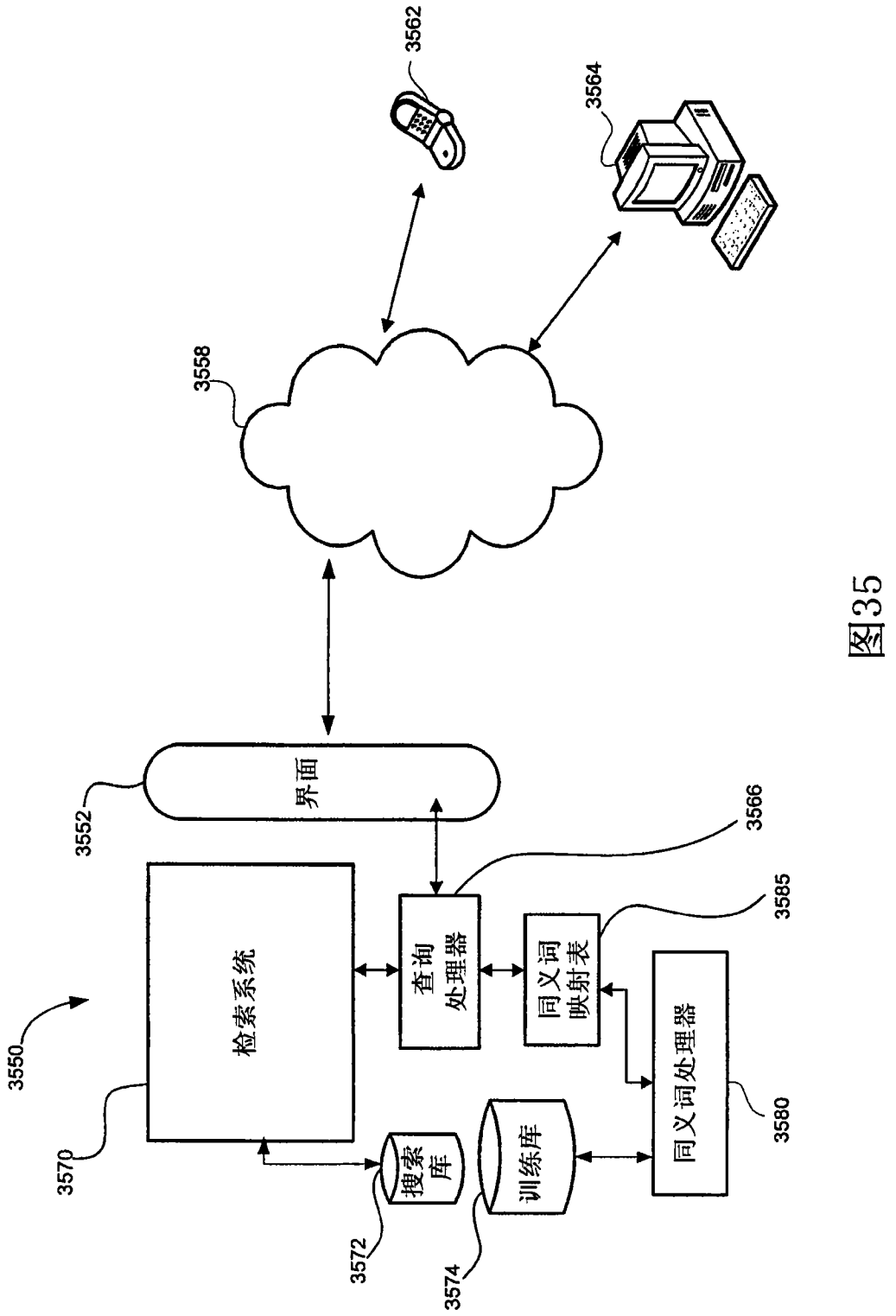

3562
3564
3558
界面
3552
3566
3585
检索系统
查询处理器
同义词映射表
3550
3570
搜索库
训练库
同义词处理器
3580
3572
3574

图35