(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 101952824 A
(43)申请公布日 2011.01.19

(21)申请号 200980105767.8
(22)申请日 2009.02.25
(30)优先权数据
12/036,681 2008.02.25 US
(85)PCT申请进入国家阶段日
2010.08.19
(86)PCT申请的申请数据
PCT/JP2009/054009 2009.02.25
(87)PCT申请的公布数据
WO2009/107851 EN 2009.09.03
(71)申请人 三菱电机株式会社
地址 日本东京都
(72)发明人 比克沙·罗摩克里希纳
埃万德罗·B·戈维亚
本特·施密特-尼尔森
加勒特·魏因贝格
布雷特·A·哈沙姆
(74)专利代理机构 北京三友知识产权代理有限公司 11127
代理人 李辉 孙海龙
(51)Int.Cl.
G06F 17/30(2006.01)

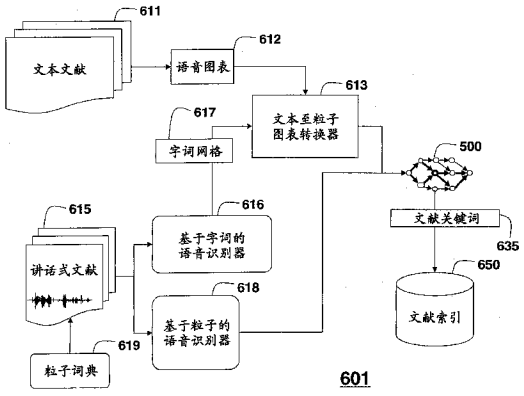
权利要求书 3 页 说明书 10 页 附图 8 页

(54)发明名称

计算机执行的对数据库中的文献进行索引和检索的方法以及信息检索系统

(57)摘要

一种信息检索系统使用粒子和基于粒子的语言模型来存储和检索文献。用于特定语言形式的文献集合的粒子的集合是根据训练文献构建的，使得基于粒子的语言模型的复杂度实质上低于按照相同的训练文献构建的基于字词的语言模型的复杂度。该文献可以随后被转换为文献粒子图表，从该文献粒子图表中提取基于粒子的关键词以形成对文献的索引。用户随后可以使用也为粒子图表形式的查询来检索相关文献。



CN 101952824 A

1. 一种计算机执行的、对数据库中的文献进行索引和检索的方法，该方法包括如下步骤：

将文献集合中的各文献转换为文献粒子图表，所述文献粒子图表包括从粒子集合中选择的粒子；

从相应的粒子图表中为各文献提取文献关键词集合；

将各文献的所述文献关键词存储至对存储所述文献集合的数据库的索引中；

将查询转换为包括查询粒子集合的查询粒子图表，所述查询粒子图表包括从所述粒子集合中选择的粒子；

从所述查询粒子图表中提取查询关键词集合；

根据所述查询关键词和存储在所述索引中的文献关键词从数据库中检索相关文献；以及

向用户输出所述相关文献。

2. 根据权利要求 1 所述的方法，其中所述粒子集合实质性地大于所述文献的语言中音位的数目，并且实质性地小于所述语言中字词的数目。

3. 根据权利要求 1 所述的方法，其中个别粒子跨越字词边界。

4. 根据权利要求 1 所述的方法，其中所述文献和所述查询为文本字词的形式。

5. 根据权利要求 1 所述的方法，其中所述文献为文本字词的形式，所述查询为讲话式字词的形式。

6. 根据权利要求 1 所述的方法，其中所述文献和所述查询都为讲话式字词的形式。

7. 根据权利要求 1 所述的方法，其中所述文献为讲话式字词的形式，所述查询为文本字词的形式。

8. 根据权利要求 1 所述的方法，其中所述查询是被讲出的，所述查询粒子图表为表示讲出的查询中的声音序列的替代的连续分组的网格。

9. 根据权利要求 1 所述的方法，其中所述粒子集合表示能在任一查询中出现的所有可能的声音序列。

10. 根据权利要求 1 所述的方法，其中所述粒子集合从来自所述文献的字词的任何序列的发音得到。

11. 根据权利要求 1 所述的方法，其中所述粒子集合识别任意文献中的将该文献与其它文献区别开的关键词。

12. 根据权利要求 1 所述的方法，其中通过拼写 - 发音机制将所述文献粒子图表和所述查询粒子图表正规化。

13. 根据权利要求 1 所述的方法，其中所述粒子集合中的粒子是在声学上独特的并是独立完整的。

14. 根据权利要求 1 所述的方法，其中粒子出现的可预见性必须高。

15. 根据权利要求 1 所述的方法，其中各粒子具有将本粒子和其它粒子相区分的区别性声学结构，并且相同的粒子的不同实例之间具有相对低的声学可变性。

16. 根据权利要求 1 所述的方法，其中个别粒子出现的可预见性相对较高。

17. 根据权利要求 1 所述的方法，其中所述粒子集合是人工确定的。

18. 根据权利要求 1 所述的方法，其中所述粒子集合是试探性地确定的。

19. 根据权利要求 1 所述的方法，所述方法还包括如下步骤：

使用训练文献构建粒子集合和同时优化的基于粒子的语言模型，其中所述基于粒子的语言模型的复杂度实质性地低于根据相同的训练文献构建的基于字词的语言模型的复杂度。

20. 根据权利要求 19 所述的方法，其中所述粒子集合对目标函数应用期望值最大，其中所述目标函数考虑下面的任意组合：

粒子集合的大小；

在展现文献训练集合和查询训练集合中的所有文献时的错误；

使用粒子集合的检索的准确度；

表示粒子集合的统计模型的熵；以及

从训练集合中的文献和查询得到的粒子级别的语言模型。

21. 根据权利要求 1 所述的方法，其中首先将各文献中的各字词转换为表示该字词的所有可能的发音的语音图表，并接着将所述语音图表转换为所述文献粒子集合。

22. 根据权利要求 1 所述的方法，所述方法还包括：

对所述相关文献进行排位。

23. 根据权利要求 20 所述的方法，其中所述基于粒子的语言模型的复杂度至少比基于字词的语言模型的复杂度低十倍。

24. 一种信息检索系统，该信息检索系统包括：

用于将文献集合中的各文献转换为文献粒子图表的装置，所述文献粒子图表包括从粒子集合中选择的粒子；

用于针对各个文献从相应的粒子图表中提取文献关键词集合的装置；

用于将各文献的所述文献关键词存储在对存储有所述文献集合的数据库的索引中的装置；

用于将查询转换为包括查询粒子集合的查询粒子图表的装置，所述查询图表包括从所述粒子集合中选择的粒子；

用于从所述查询粒子图表中提取查询关键词集合的装置；

用于根据所述查询关键词和存储在所述索引中的文献关键词从数据库中检索相关文献的装置；以及

用于向用户输出相关文献的装置。

25. 一种计算机执行的、对数据库中的文献进行索引和检索的方法，该方法包括如下步骤：

使用基于粒子的语言模型根据训练文献构建粒子集合，其中所述基于粒子的语言模型的复杂度实质性地低于根据相同的训练文献构建的基于字词的语言模型的复杂度；

将文献集合中的各文献转换为文献粒子图表，所述文献粒子图表包括从所述粒子集合中选择的粒子；

针对各个文献从相应的粒子图表中提取文献关键词集合，以形成对所述文献的索引；以及

由用户使用查询粒子图表形式的查询和从所述查询粒子图表中提取的关键词检索相关文献。

26. 一种信息检索系统，该信息检索系统包括：

用于存储文献集合的数据库；

对所述数据库的索引，其中所述索引中的条目是粒子的形式，其中所述粒子选自使用基于粒子的语言模型根据训练文献构建的粒子集合，并且其中所述基于粒子的语言模型的复杂度实质性地低于根据相同的训练文献构建的基于字词的语言模型的复杂度；以及

用于由用户使用所述粒子通过所述索引来访问所述文献的装置。

计算机执行的对数据库中的文献进行索引和检索的方法以及信息检索系统

技术领域

[0001] 本发明总体上涉及信息检索，更具体地涉及对数据库中的文献进行索引和检索。

背景技术

[0002] 检索与文本查询有关的文献的信息检索系统是很普遍的。文献通常为字词的集合，该字词的集合直接由该集合中的字词来索引或通过字词－计数矢量（通常称为文献矢量）的线性变换来索引。查询还可以被表示为用于根据索引检索文献的字词的集合，或被表示为与文献矢量相比较来识别与查询最相关的文献的字词－计数矢量。向用户返回的相关文献通常被称为结果集。

[0003] 自动语音识别 (ASR) 系统的不断增加的可用性允许从基于文本的信息检索系统扩展到说出文献或查询的系统。

[0004] 讲话式文献检索系统可以为广播新闻节目的音频录音、播客、会议记录、演讲、表演等编索引。通常，先人工地或使用 ASR 系统地将讲话式文献转录为文本。将文本中所得到的字词存储在数据库索引中。将查询与字词索引相匹配，并向用户返回文本式抄本或音频记录。

[0005] 讲话式查询系统使用语音来查询文献检索系统。再一次，使用 ASR 系统将查询转换为字词的形式并与索引匹配以进行检索。

[0006] 在上述的全部情况下，索引系统所使用的基本单位为字词。在纯粹的基于文本的系统中，文献和查询都为文本，利用文献中的字词对文献进行索引，并将查询中的字词与索引中的字词进行匹配。在文献或查询为讲话形式时，首先将字词转换为字词序列或字词网格，再将其用于构造字词索引或对照字词索引对查询进行匹配。

[0007] 基于字词的索引方案具有基本的限制，当查询或文献是讲话形式时尤其如此。ASR 系统具有有限的词汇量。系统可以识别的字词的词汇量必需首先被指定。这还意味着只要将包括了当前识别器的词汇量中没有的字词的文献加入到索引，就必需更新识别器的词汇量。

[0008] 在讲话式文献的情况下，由于新的文献的词汇量不能完全地被事先获知，因此会存在问题。对于讲话式查询，这暗示了只要对文献索引进行了更新，就必需对用于输入查询的系统进行更新。在许多应用中这是不切实际的要求。即使文献和查询都是完全基于文本的，基于文本的索引也面临拼错的问题。查询中的字词经常被用户拼写为不同于文献中的字词，当该字词是新词或很复杂时尤其如此。显然，当在文献中拼写的字词和在查询中拼写的字词不匹配时，会对检索产生不利的影响。

[0009] 文献检索系统通常从数据库中返回被认为是与用户查询中的字词相关的一个或更多个文献。术语“文献”的解释是很广义的。例如，对来自网络的文献的检索和对来自个人计算机的文件的检索，或者对来自由元数据所描述的歌曲集合中的音乐的检索都可以被看作是“文献”检索的实例。

[0010] 很明显，并不是文献中的所有信息都适于通过菜单进行遍历的树型结构对话。需要使用通常被称为“信息检索”(IR) 的、不依赖于文献中信息的结构的技术来对信息进行检索。

[0011] 文献并不总是基于文本的。文献还可以包括讲话式数据（如广播新闻节目、讨论会和演讲、公共致辞、会议等）的记录。同样地、用于从数据库中检索文献的查询也不需要一定是文本的。查询也可以被说出。

[0012] 基于文本的检索

[0013] 图 1 示出了常规的基于文本的系统，文献 101 和查询 102 都是文本形式。从所有文献抽出（见 103) 的字词或字词式样的集合被用于构建文献索引 104。还可以从查询中抽出（见 105) 字词或字词式样。该索引具有字词，各字词指向出现了该字词的每一文献，或者该索引具有针对各文献的字词计数矢量。该字词计数矢量具有各字词在文献中出现的次数。

[0014] 于是可以按照与索引的结构相一致的方式对查询进行处理，对文献的结果集合 107 进行评分和排序（见 106)，并返回给用户。

[0015] 讲话式文献检索

[0016] 如图 2 所示，讲话式文献 201 包括语音的音频记录，如上面所述。对该语音进行识别（步骤 202)。有时需要响应于查询 102 对这种文献编索引并进行检索。

[0017] 常规的检索讲话式文献的方法是使用 ASR 系统将文献转换为字词序列。接着按照与文本文献相同的方式对转换后的文献编索引并进行检索。

[0018] 众所周知，ASR 系统本质上是不准确的。由此识别出的针对任何文献的字词可能包含多个错误，该错误将会导致响应于查询而检索到错误的文献。为了解决该问题，通常以字词网格来表示文献，在对文献进行解码时识别器会考虑该字词网格。另选地，可以采用 n- 最好列表（即识别器为文献生成的前 N 个识别假定）来表示文献。接着通过从字词网格 n- 最好列表得到的字词（或字词计数矢量）对文献编索引。其余的索引编排方法和检索过程与文本文献的相同。

[0019] 如图 3 所示，一种另选的方法是将讲话式文献转换为音位的序列或网格 302，或者转换为字词的音节（步骤 301)。按照这些网格来完整地表示文献。然后将查询中的字词与文献中的序列或网格进行匹配，来识别包含了能够与查询中的字词匹配的序列的候选文献。

[0020] 按照讲话式查询的检索

[0021] 例如在使用小型手持设备时或在开车或操作机器时，在查询中输入文本并不总是方便的。文本输入可能是不方便地，或者甚至是不可能的。在这样的情形下，用户可以说出他们的查询。讲话式时查询系统试图使用讲话式查询中的字词来检索文献。

[0022] 如同讲话式文献检索的情况那样，首先由 ASR 系统将讲话式查询转换为字词。再一次地，可以将文献转换为字词的线性序列或网格。查询的文本形式的字词被用于从索引中检索文献，如参见于 2005 年 4 月 5 日向 Wolf 等人签发的美国专利 6,877,001，“Method and system for retrievingdocuments with spoken queries”，以引用的方式将其合并于此。

[0023] 其他系统可以在它们的索引中将文本文献和讲话式文献进行合并，并允许讲话式

查询和基于文本的查询。在所有的情况下，用于将文献与查询相匹配的基本单位为字词。

[0024] 基于字词的匹配的缺陷

[0025] 使用文本查询对文本文献的检索大概是文献检索的所有形式中最可靠的。但是，它有它的限制。在文献中的将该文献与其他的文献相区别的关键词通常为新的字词，具有不常见的拼写。试图对这些文献进行检索的用户经常对这些词条的准确的拼法不确定，并拼错字词。任何基于字词的检索机制都不能够将拼错的字词与相应的文献相匹配。为了解决这样的问题，许多基于字词的系统使用各种拼写校正机制来警告用户可能误拼，但是在用户基本上不能确定拼写的情况下即使是这样也不够。

[0026] 必须首先使用 ASR 系统将讲话式文献转换为字词。ASR 系统具有有限的词汇量，即使是词汇量非常大。超大词汇量系统甚至通常在其识别词汇表中包括最常用的数万个字词，或者在个别情况下，包括数十万个字词。这随即产生了几个问题。首先，在任一文献中的关键区别词条本质上是不常用的，否则它们并不能将该文献与其它文献区别开。结果，恰好这些字词实际上最不可能出现在识别器的词汇表中，由此不太可能被识别出来。为了解决该问题，必须在识别之前将文献中的这些关键词添加到识别器的词汇表中。此处产生了一个必然的问题。在新的文献中，不能够事先得知要被查找的关键词。

[0027] 其次，ASR 系统是事先偏向了的统计机器，使得出现频率高的字词比出现频率低的字词更准确地被识别。结果，即使在某一文献中的关键词实际上已经包含在了 ASR 系统的词汇表中，该关键词还是很有可能被错误识别，由此使得将它们包含在系统的词汇表中的理论变得无效。作为补偿因素，文献中的关键词通常在讲话式文献中被重复多次，则识别器遗漏所有字词实例的可能性大大低于识别器遗漏某单个的实例的可能性。因此，即使在识别器的准确度相对较低的情况下，讲话式文献检索系统也可以合理地运行。

[0028] 即使在讲话式文献被实际上转录为网格以减少词汇表之外的词条的影响的情况下，查询仍然是必须与文献相匹配的整个字词，并且还将遭受上述的误拼的问题。更重要的是，为了对文献进行评分，这将需要对查询中的各字词与各文献的整个粒子网格进行匹配，使整个处理的效率非常低。

[0029] 讲话式查询或许在所有文献检索系统中是最不可靠的。通常如上所述地利用 ASR 系统将查询转换为字词序列或字词网格。查询通常很短。很明显，单个的误识别的代价是非常高的。

[0030] 为了被识别，用户希望在文献中找到的字词必须包含在识别器的词汇表中。这意味着在将文献添加到索引的同时，文献中的关键词必须首先包含在处理查询的识别器的词汇表中。这对于由远程客户端对查询进行初始处理的系统来说尤其是难以负担的。对索引的更新必须迅速地传递到旨在使用该索引的各客户端。这种操作变得非常得费时。

[0031] 即使在搭配有索引的服务器上进行查询处理，时间限制也是一个问题。用户需要迅速地响应查询。ASR 系统操作的速度取决于词汇量，造成识别词汇量增加的文献索引的各个更新将降低 ASR 系统的速度并增加检索的等待时间。ASR 系统所使用的存储器容量也将随词汇量的增加而非线性地增加，限制了可以同时处理的查询的数量。

发明内容

[0032] 常规的信息检索机制按照字词或字词组合来展现文献。不管文献或查询为口语的

还是书面的，这都适用。利用字词组合的索引会造成由拼写或识别的不确定或错误引起的多个限制。由于自动语音识别 (ASR) 系统受到进一步的词汇量的限制，在查询或文献为讲话形式并且必须先于索引而进行识别时，这些限制当然更严重。

[0033] 本发明实施方式提供了一种文献索引和检索系统，该系统以粒子为单位展现文献以采用讲话式查询进行检索。通过适当地选择粒子，可以避免系统的词汇量的限制。此外，该系统可以采用更小的语言模型，在具有比基于字词的索引系统所需的常规的基于字词的信息检索系统更小的存储量和 CPU 要求的情况下运行。

附图说明

[0034] 图 1 为使用文本查询的常规的文献检索系统的框图；

[0035] 图 2- 图 3 为使用讲话式查询的常规的文献检索系统的框图；

[0036] 图 4 为根据本发明的实施方式的语音图表；

[0037] 图 5 为根据本发明实施方式的粒子图表的框图；

[0038] 图 6A 为根据本发明实施方式的文献粒子化器的框图；

[0039] 图 6B 为根据本发明实施方式的查询粒子化器的框图；以及

[0040] 图 6C 为根据本发明实施方式的基于粒子的信息检索系统的框图。

具体实施方式

[0041] 基于粒子的文献索引

[0042] 本发明的实施方式提供了一种基于粒子而不是象现有技术中那样的基于字词为文献编索引并检索文献的方法。

[0043] 粒子本身不是新的，参见 Whittaker, E.W.D., Woodland, P.C. “Particle-based language modeling”，语音语言处理国际会议 (ICSLP)，2000，于 2006 年 8 月 8 日向 Logan 等人签发的美国专利 7,089,188，“Methodto expand inputs for word or document searching”，以及于 2007 年 2 月 20 日向 Thong 等人签发的美国专利 7,181,398，“Vocabulary independentspeech recognition system and method using subword units”。但是，这些粒子被用于识别字词，在文献检索过程中字词被编入索引并被搜索。

[0044] 粒子索引和检索是基于我们的这样的观察，字词的发音可以由一个或更多个声音单位（如音位 (phoneme) 和音节）的序列来描述。从而，任何口语发音都可以基本地被视为是一系列这样的声音单位。字词仅仅被认为是带有语义关系的这样的声音单位的组合。但是，讲话的声音单位可以按照与字词指定的方式不同的任何其它方式顺序地成组。

[0045] 这如表 1 所示。

[0046] 表 1

[0047]

The big dog	dh iy . b ih g . d aa g	/dh iy/	/b ih g/	/d aa g/	
		/dh iy b/	/ih g d/	/aa g/	
		/dh/	/iy b ih/	/g d/	/aa g/

[0048] 表 1 示出了将字词序列“the big dog”（在第一列）表示为最右边的 4 列中的粒子的不同方式。第二列表示讲话中字词的语音发音。在该列中的句点将字词分开。

[0049] 如果我们假设一种讲话的特征在于该讲话中的整个声音序列而不是声音的特定

的连续组合，则表 1 中的所有的粒子分解都是讲话的有效特征。

[0050] 基于粒子的展示的目的则在于提出另选的声音的顺序组合，该声音的顺序组合可以表示推定的和实际的、在文献集合内出现的声音序列。

[0051] 我们将这些组合中的每一个组合称为粒子。例如，在表 1 的示例中，由斜线 (/) 括起来的声音序列中的每一个，如 /dh iy/，/dh iy b/ 和 /dh/，都为粒子。应注意的是，在该表中的一些粒子实际上跨越了字词边界的，这是非常规的。是否接受这些粒子取决于对为了展现文献和查询中的语言而选择的粒子的特定集合的设计。

[0052] 粒子

[0053] 尽管可以以如上所述的多种途径来构建粒子，但并不是所有可能的粒子都可以用于基于粒子的索引。用于展示查询和文献来进行检索的实际的粒子集合是经过仔细挑选的。

[0054] 我们对粒子的集合施加了如下必要条件。

[0055] 1. 粒子必须展现在任意查询中出现的所有可能的声音序列，或者粒子可以由来自文献的任何字词序列的发音来得到。

[0056] 2. 粒子必须使得可识别出任意文献中的能将该文献与其他文献区别开的关键粒子。

[0057] 条件 1 的必要性是不言而喻的。为了准确地展现任意文献或查询，必须能够以粒子的方式完全地表示文献。如果任意句子或讲话不能正确地分解为粒子的序列，它就不能有效地用作索引的关键条目或用作查询中的关键词。

[0058] 但是，在理解了文献（或查询）的未展示部分不可用的情况下，我们可以稍微放松第一条件来规定“粒子必须表示任意查询或文献中的绝大多数声音序列”，如果粒子数目足够小，则不影响系统的总体性能。

[0059] 条件 2 起因于系统目标为信息检索的这样的事实。为了正确地检索到与查询有关的文献，需要能够识别出查询中在相关文献中的比在其它文献中更频繁的模式。

[0060] 在常规的基于字词的对文献和查询的展示中，查询中的字词自身表示用户希望在相关文献中找到的独特样式。

[0061] 当以粒子的方式来展示文献和查询时，同样需要查询中的粒子（或粒子模式）在相关文献中比在其它文献中以更高的频率存在。

[0062] 例如，在文献的语言中的音位集合应满足条件 1 并可以用于展示任何查询或文献。然而，音位出现的相对频率在文献集合中并没有太大的变化（尤其是在该集合很大时（如网络上不计其数的文献）），并展示语言的语音特征而非具体的文献特征。例如，在最常用的语言中的音位数目是很小的，例如约 50。结果，从查询的基于音位的展示中进行检索的任何尝试都有可能返回包含该查询中的音位、但在语义上并不与该查询以任何方式相关的大量文献。很显然，很小的音位集合不是很好的针对 IR 的集合。这样，基于如下原因，与使用字词相比，在检索系统中使用粒子是有优势的。

[0063] 文本正规化和拼写

[0064] 基于字词的检索方案严格地依赖于在文献和查询中字词的正确拼写。基于粒子的索引机制与词典中的粒子序列相匹配，并可以使用自动的拼写 - 到 - 发音机制。发音词典和拼写 - 到 - 发音系统提供了在与实际的字词的发音非常相似的（如果不是完全相同的话）

字词的误拼或不同拼写实例的发音（并由此粒子化），因此使得拼写错误或变型的影响被正规化掉。

[0065] 词汇量大小

[0066] 对于既处理讲话式文献又处理讲话式查询的文献检索系统，必须采用语音识别器来将讲话式音频转换为文本格式。对基于字词的系统，字词级别识别器将音频信号转换为字词序列或图表。基于字词的识别器的性能严格依赖于识别器的词汇量，即识别器必须能够识别的独特字词的总数，词汇量反过来又与文献集合中独特字词的总数有关。随着文献数目的增加，独特字词的数目也不可避免的增加，由此识别器的词汇量也增加。增加的词汇量降低了识别器的准确率，由此大大降低了信息检索的准确率。通常基于字词的识别器可以存储 50,000 到 100,000 个字词。

[0067] 然而，在基于粒子的系统中，由于识别器目前仅识别粒子，并且粒子的集合的大小（如 50）远小于字词级别的词汇量（如 50,000 到 100,000），因此该问题即使不能完全消除，也得到了大大的缓解。理想的粒子集合必须是使得粒子的分布对文献是可以辨别的。

[0068] 词汇表之外的字词

[0069] 没有在识别器的词汇表中的字词是不能被识别的，由此不能用于为文献编索引或检索文献。为了避免该问题，只要将新的文献添加到了索引中就必须对识别器的词汇量进行更新。在每次更新索引时都必须更新信息检索客户端的讲话式查询中，这会变为尤其令人厌烦的问题。对于基于粒子的系统，由于新的字词通常可以被分解为在识别器中存在的粒子集合，该问题大大减轻。通过一个极端示例（其中，粒子为音位）可以对此进行最好的例示。任何新的字词都可以基于其从词典或拼写 - 到 - 发音生成器确定的发音而被表示为音位的序列。更一般地说，在适当地选择识别器的粒子集合的情况下，可以以识别器的粒子集合类似地表示新的字词。

[0070] 除了上述的要求之外，当文献或查询为讲话形式时，我们需要额外的条件，这是由于为了有效的性能，粒子必须由 ASR 系统容易地进行识别。这就导致了如下的要求：

[0071] 3. 粒子的集合必须相对地小；

[0072] 4. 理想地，粒子应当是在声学上独特的以及独立完整的单位 (self-contained unit)；以及

[0073] 5. 粒子的出现的预见性相对较高。

[0074] 条件 3 与识别器的速度、准确率和大小有关。较小的粒子集合导致识别器的较小的识别词汇量，以及对应的仅需要使用较小的粒子词汇量的较小的语法和语言模型。使此平衡的是这样的事实：较小的粒子集合通常包括声学上更短的、不能有效区别文献的粒子。此外，声学上更小的单位具有更少的声学提示 (cue) 并更加难于识别。例如，粒子数目约为 2000。

[0075] 上述条件 4 对于粒子的可识别是重要的。为了可识别，粒子不仅必须具有能够将它们与其它粒子相区分的可区别的声学结构，并且还必须在相同的粒子的不同实例之间展示出相对低的可变性。从这个意义上讲，字词是很好的声学单位，因为它们倾向于具有几个声学提示并且是独立完整的。其它类似的声学区别单位为音节，音节不仅具有可区别的声学结构而且还这样发音以使得在音节边缘由协同发音引起的变化变低，导致在它们表达上的变化减少。但是音节比字词具有更少的声学提示。还可以设计其它类似的粒子集合。理

想的粒子集合不仅会满足条件 4，还会满足其它条件。

[0076] 条件 5 与粒子的语言预测性有关。预测性的一种可能的指标为复杂度 (perplexity)。从统计的观点来看，基于粒子的语言模型的复杂度实际上比利用相同的训练文本构建的基于字词的语言模型的复杂度要低（例如，至少低十倍）。如同本领域中所周知的，复杂度是根据观察到的字词的历史从中选择下一字词的字词集合的大小的指标。我们将复杂度扩展到粒子和基于粒子的语言模型。我们采用该要求是因为语音识别系统的准确度随着语音复杂度的增加而降低。

[0077] 构建粒子集合

[0078] 可以人工地或试探性地构建粒子集合。在具有有限数量的音节的语言（如日语）中，语言中的所有音节的集合形成自然粒子集合。在另一种语言（如英语）中，粒子集合构建起来会更难。

[0079] 在本发明的一种实施方式中，通过对训练文献和同时优化的语言模型的分析来试探性地构建粒子集合。训练文献可以包括文本文献和讲话式文献。尽管没有具体说明实际用于构建粒子集合的方法，我们却描述了一般的指导方针。

[0080] 自动构建粒子集合的一种方法可以使用将前一部分指出的所有要求进行了编码的目标函数，这些要求为：

[0081] 1. 粒子集合大小；

[0082] 2. 在展现给定的训练集合中的所有文献和查询过程中的错误；

[0083] 3. 使用粒子集合的检索的准确度；

[0084] 4. 表示粒子的统计模型的熵 (entropy)；以及

[0085] 5. 可以包括在所述目标函数中的从训练资料库中的所有文献和查询导出的基于粒子的语言模型的复杂度。

[0086] 可以通过仅在目标函数中合并这些条件中的一些来得到粒子集合。例如，在目标函数中嵌入的基于任何熵、复杂度或似然性的标准可以导致基于期望值最大 (EM) 的有效的算法。

[0087] 基于粒子的信息检索 (IR)

[0088] 本发明的主要思路是：基于粒子的信息检索方案与基于字词的方案相比更有可能对拼写、发音或其它错误具有鲁棒性。因此，将基于粒子的 IR 方案应用于所有的场景，即基于文本的对文本文献的检索、基于文本的对讲话式文献的检索、基于讲话式查询的对文本或讲话式文献的检索、以及所有其它这类检索的结合。在最一般的情况下，文献可以为讲话式文献或文本文献。类似地，可以讲出或作为文本来输入查询。下面我们简要地描述如何处理这些情况中的各情况。

[0089] 文本文献

[0090] 文本文献包括字词序列。首先将文献中的文本转换为基于粒子的表示。为此，我们首先将各字词（如“semisoft”）转换为如图 4 中所示的表示该字词的所有可能的发音的语音图表 400。在只有一种字词发音方式的情况下该图表还可以是线性的。

[0091] 通过本发明的定义，粒子可以是音位的短序列或长序列，就如同表 1 的最右侧的四列所示。

[0092] 可以将字词序列的发音分组为表 1 中所示的粒子序列。但是，对于任意给定的粒

子组合，可以存在多个将发音分组为粒子序列的途径。例如，如果我们的粒子组合包括粒子“/dh iy/”，“/b ihg/”，“/d aa g/”，“/dh iyb/”，和“/ih g/”，则字词序列“the big dog”可以被表示为“/dh iy//b ih g//d aag/”或“/dh iy b//ih g//d aa g/”。可以将这些另选的分解表示为图 5 的文献粒子图表 500。

[0093] 基于粒子的信息检索系统

[0094] 图 6A 至图 6C 示出了根据本发明实施方式的基于粒子的信息检索系统的结构。图 6A 示出了文献粒子化器 601。图 6B 示出了查询粒子化器 602。图 6C 示出了使用粒子为文献编索引和检索文献。

[0095] 文献粒子化器

[0096] 文本文献

[0097] 图 6A 示出了本发明的粒子化器 601。文本文献被转换为粒子图表 500。首先通过从发音词典或从音位 - 至 - 语义图变换器得到文本中各字词的发音来将文本转换为语音图表 612。额外的可选的输入可以包括进行各种限制的规则，如对跨越了字词的边界的粒子的限制以及在粒子集合不全（即一些字词序列不能被完全地分解为粒子图表）的情况下的错误最低限度标准。

[0098] 接着，使用语音图表生成粒子图表 (613)。本发明称该过程为文献粒子化。文献粒子图表 500 可以是线性的（即仅单个的粒子序列）或者为图 5 所示的网格。

[0099] 与常规的图表不同，粒子可以跨越字词边界。另选地，可以对文献中的单个字词独立地进行粒子化。从文献粒子图表中抽出文献关键字集合 635。该集合可以包括一个或更多个关键字。将该文献关键字存储在文献索引 650 中。该索引可以直接将图表中选出的粒子用作关键字来引用文献，或者索引可以使用粒子序列。本发明将粒子序列称为 n-gram。另选地，可以使用诸如粒子计数矢量或粒子分布（例如标准化直方图）来展现文献。

[0100] 讲话式文献

[0101] 讲话式文献 615 包含音频信号，例如语音。与文本文献类似，也将讲话式文献转换为文献粒子图表 500 并随后将文献关键字加入索引 650。作为附加特征，讲话式文献的索引可以包括指示何时在文献中出现各种粒子式样的时间戳。回想一下，文本是空间的而语音是随时间变化的，因此基于时间的索引是恰当的。

[0102] 通过使用语音识别器 616 将讲话式文献 615 转换为粒子图表。实现转换的方式可以有多种。在第一种选择中，常规的基于字词的语音识别系统将音频信号转换为序列或字词网格 617。随后将字词网格转换为针对文本文献所描述的粒子图表 (613)。

[0103] 另选地，使用基于粒子的语音识别器 618 直接将讲话式文献 615 转换为粒子图表。该粒子识别器访问将粒子映射到它们的发音的“粒子”词典 619。相应的语法或统计语言模型指定了各种有效粒子序列和它们的概率。粒子识别器输出粒子图表 500，并将从粒子图表中提取出的关键词集合进行存储并用于为文献编索引。

[0104] 在讲话式文献的情况下，还可以得到语音识别器输出的粒子或字词的权重。该权重表示在讲话式数据中实际出现假定的字或粒子的置信度，或者词条（即字词或粒子）在文献中出现的后验概率。在两种情况下，这些权重还可以被因数化在用于展现文献的关键词中。这样，粒子、粒子 n-gram 或粒子直方图都可以通过这些权重以各种方式进行强化。

[0105] 查询粒子化器

[0106] 文本查询

[0107] 如图 6B 所示，文本查询 621 也被转换为如上所述的查询粒子图表 500。使用发音词典或语义图 - 至 - 音位变换器将文本查询中的字词转换为语音图表 612。然后根据发音图表得到粒子图表。还可以从粒子图表得到作为查询关键词 636 的粒子、粒子 n-gram、粒子计数矢量或粒子出现直方图，以从多个文献中检索文献。

[0108] 讲话式查询

[0109] 使用语音识别器 616 和 618 中的任意一个将讲话式查询 625 转换为查询粒子图表 501。与在讲话式文献的情况下一样，可以通过使用基于字词的语音识别器首先将查询转换成字词串或网格，并类似于对文本查询的处理将字词图表进一步转换为语音图表来对查询进行转换，或者可以通过使用基于粒子的识别器直接得到粒子图表来对查询进行转换。与在文本查询的情况下一样，可以从粒子图表得到作为查询关键词 636 的粒子、粒子 n-gram、粒子计数矢量或粒子出现直方图，以使用文献索引进行文献检索。再一次地，可以将从识别器得到的置信度、后验概率或其他权重用于在形成关键词之前对查询中的术语进行加权。

[0110] 基于粒子的文献索引

[0111] 图 6C 示出了基于粒子的 IR 系统 603 的整体。基于粒子的文献索引 650 是存储文献或存储指向文献的指针的数据库。通常，该数据库是为存储器（如磁盘、磁带、RAM 和 ROM 等）的形式。数据库可以集中式的或如因特网一样是广泛分散式的。

[0112] 可以通过各种机制（如粒子、粒子 n-gram、粒子频率直方图或粒子概率直方图）对数据库中的文献编索引。通过从粒子图表 500 中提取 631 粒子或粒子式样作为文献关键词来生成索引。

[0113] 将从查询中得到的粒子图表转换为用于文献索引 650 的（一个或更多个）查询关键词的集合 (632)。

[0114] 粒子 - 图表到查询变换器

[0115] 该模块将从查询得到的粒子图表转换为可以用于从索引 650 中索引文献的关键词 636 的集合。关键词可以为图表自身中的粒子、粒子 n-gram、粒子计数矢量或粒子频率直方图。可以将通过语音识别器确定出的适当的权重用于强化这些关键词。

[0116] 文献记分器

[0117] 文献记分器 650 确定由查询 636 的关键词编入索引的文献的相关度得分。相关度得分可以被确定为根据查询确定的粒子计数矢量或粒子频率直方图与根据文献确定的粒子计数矢量或粒子频率直方图之间的距离（差异）。

[0118] 已知有多种距离指标，如 Kullback-Leibler 距离、余弦距离。另选地，可以以从查询得到的粒子或粒子 n-gram 与文献相匹配的总数的形式来确定相关度。按照相关度下降的顺序将结果集合 637 中的文献返回给用户。

[0119] 本发明的效果

[0120] 本发明提供了一种检索信息的新的方法。文献和查询既可以是文本也可以是语音。与现有技术中使用基于字词的展示不同，本发明将文献和查询分解为比字词更小的小单位，我们称之为粒子。虽然不是必须的，但通常按照发音来定义这些小的粒子，各粒子表示声音的连续的序列。文献无论是讲话式的或文本的，都被转换为这些粒子的序列。按照粒子的形式来编索引。查询也被转换为粒子的序列，这些粒子的序列然后用于从索引中检

索文献。

[0121] 尽管参照优选实施方式的示例描述了本发明，但应当理解，可以在本发明的精神和范围内作出的许多其它的变型和修改。因此，所附的权利要求的目的在于涵盖落入本发明的真正精神和范围内的所有这样的变型和修改。

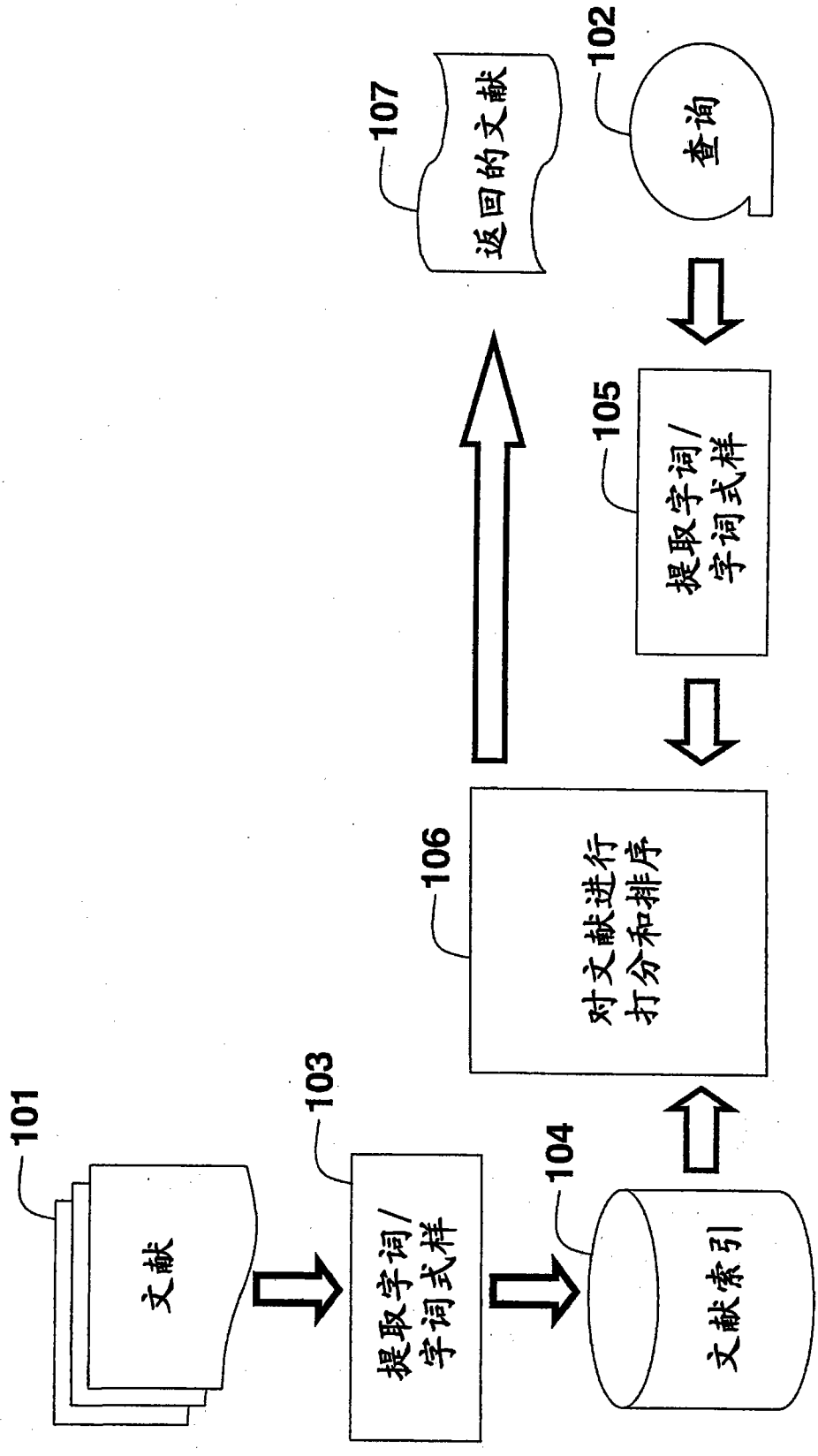

101
文献
103
提取字词/
字词式样
104
文献索引
106
对文献进行
打分和排序
105
提取字词/
字词式样
107
返回的文献
102
查询

图 1 现有技术

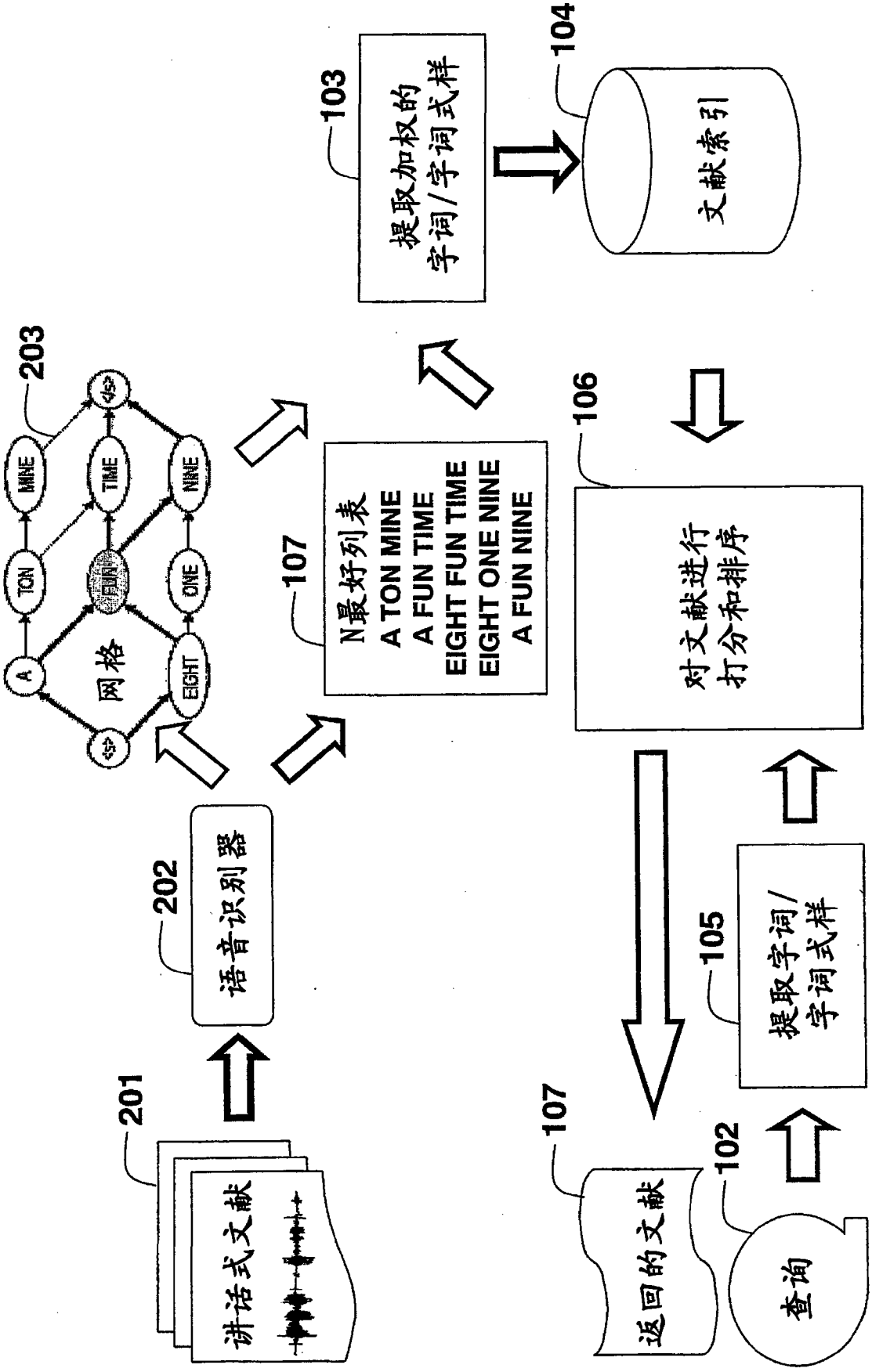

讲话式文献
201
语音识别器
202
203
网格
A
TON
FUN
EIGHT
ONE
MINE
TIME
NINE
N最好列表
A TON MINE
A FUN TIME
EIGHT FUN TIME
EIGHT ONE NINE
A FUN NINE
107
提取加权的
字词/字词式样
103
文献索引
104
对文献进行
打分和排序
106
提取字词/
字词式样
105
返回的文献
107
查询
102

图 2 现有技术

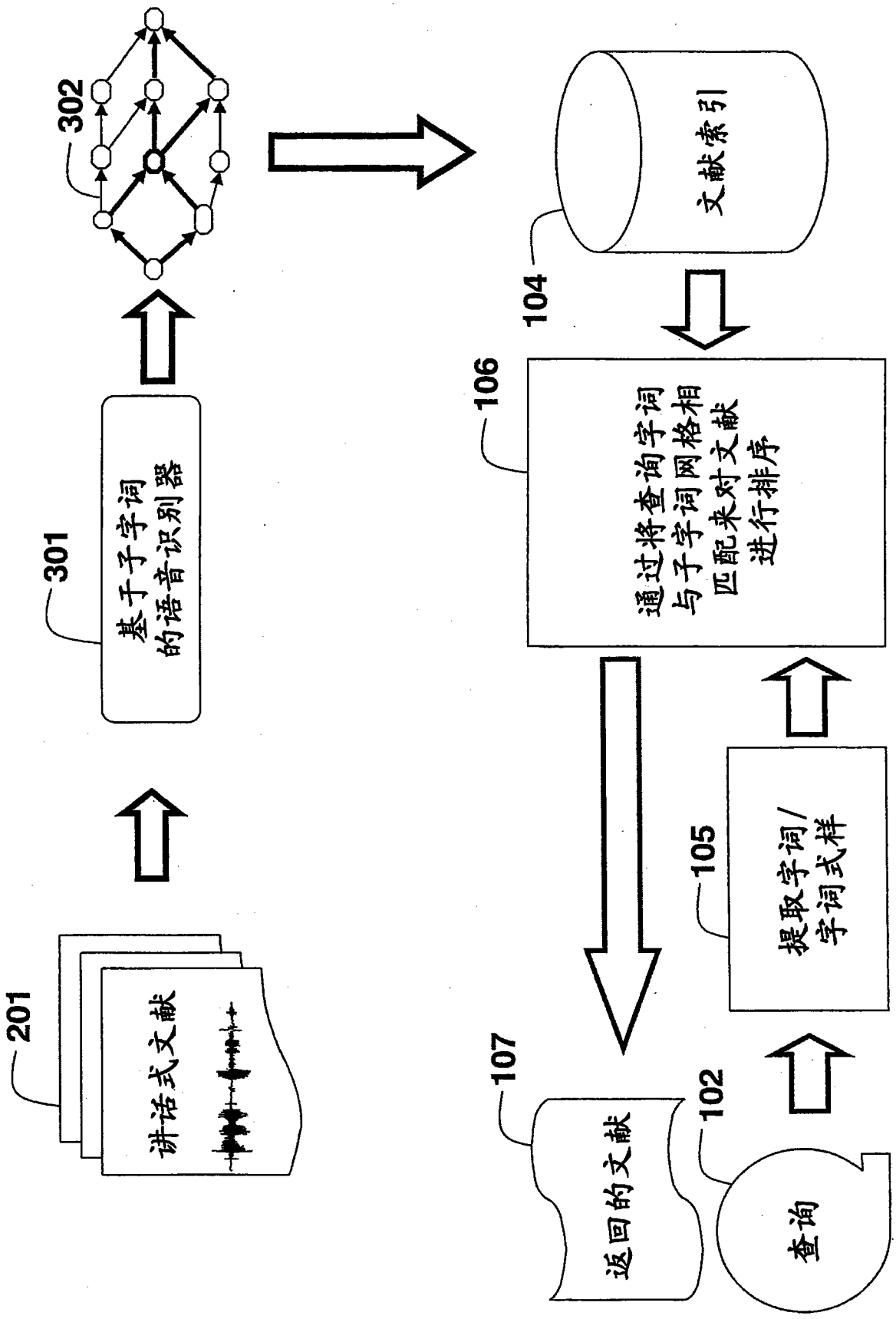

201
讲话式文献
301
基于子字词
的语音识别器
302
104
文献索引
106
通过将查询字词
与子字词网格相
匹配来对文献
进行排序
105
提取字词/
字词式样
107
返回的文献
102
查询

图 3 现有技术

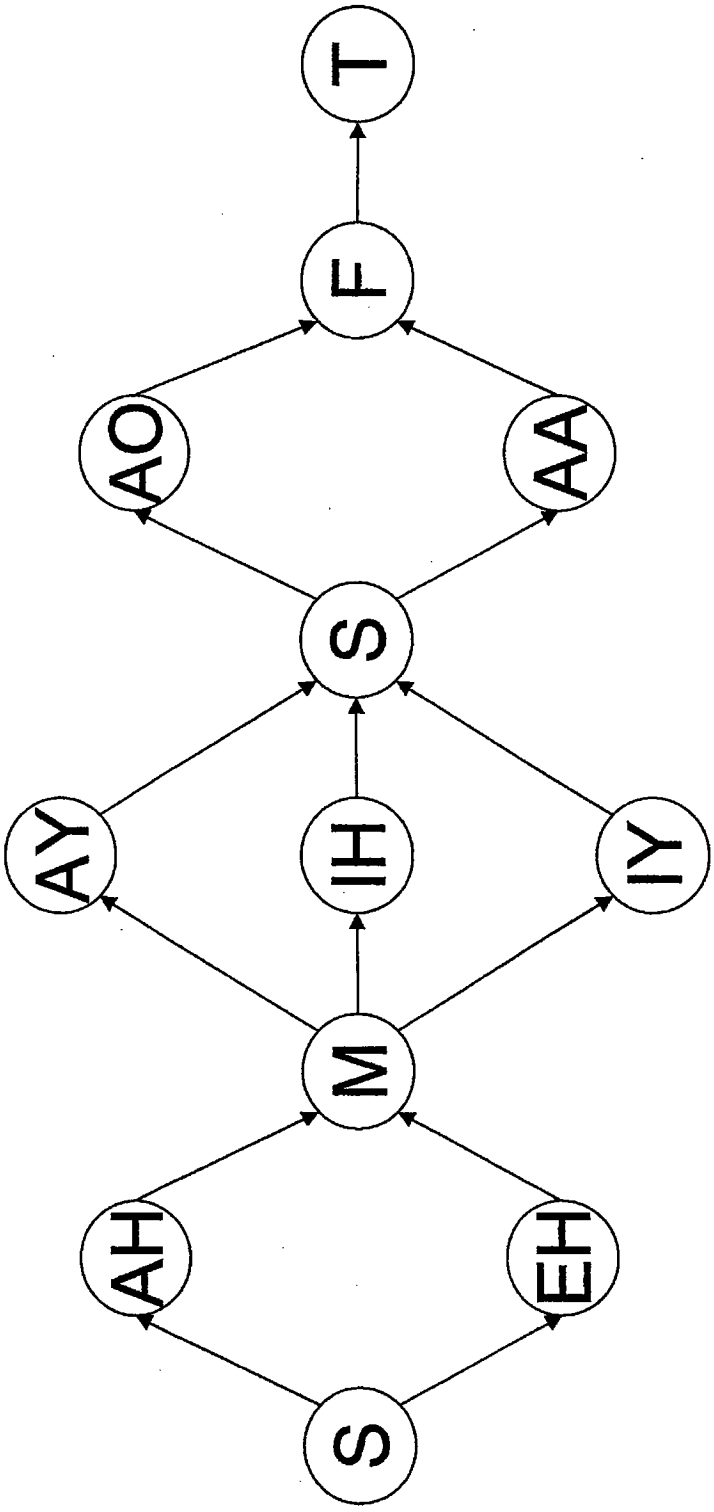


图 4

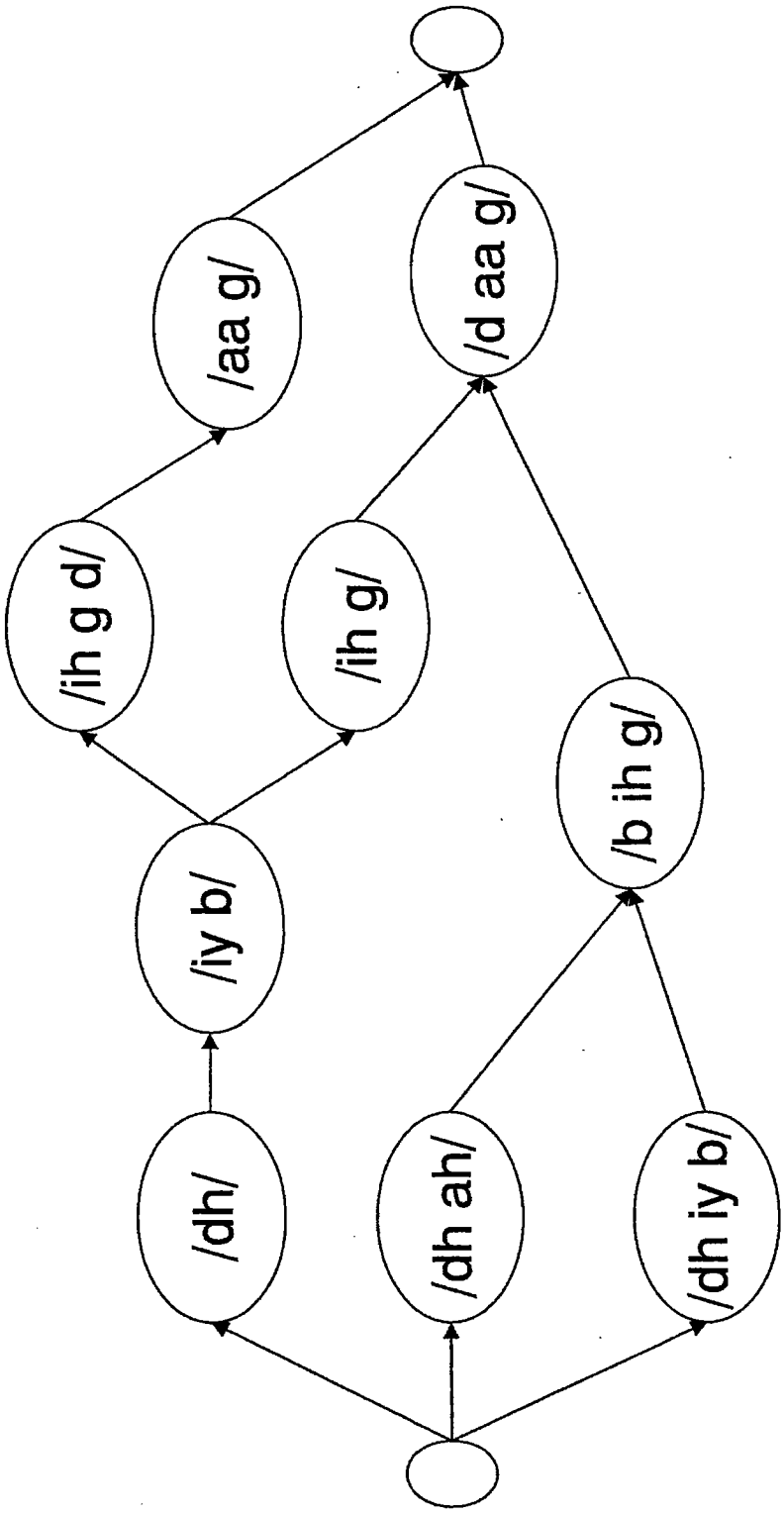


图 5

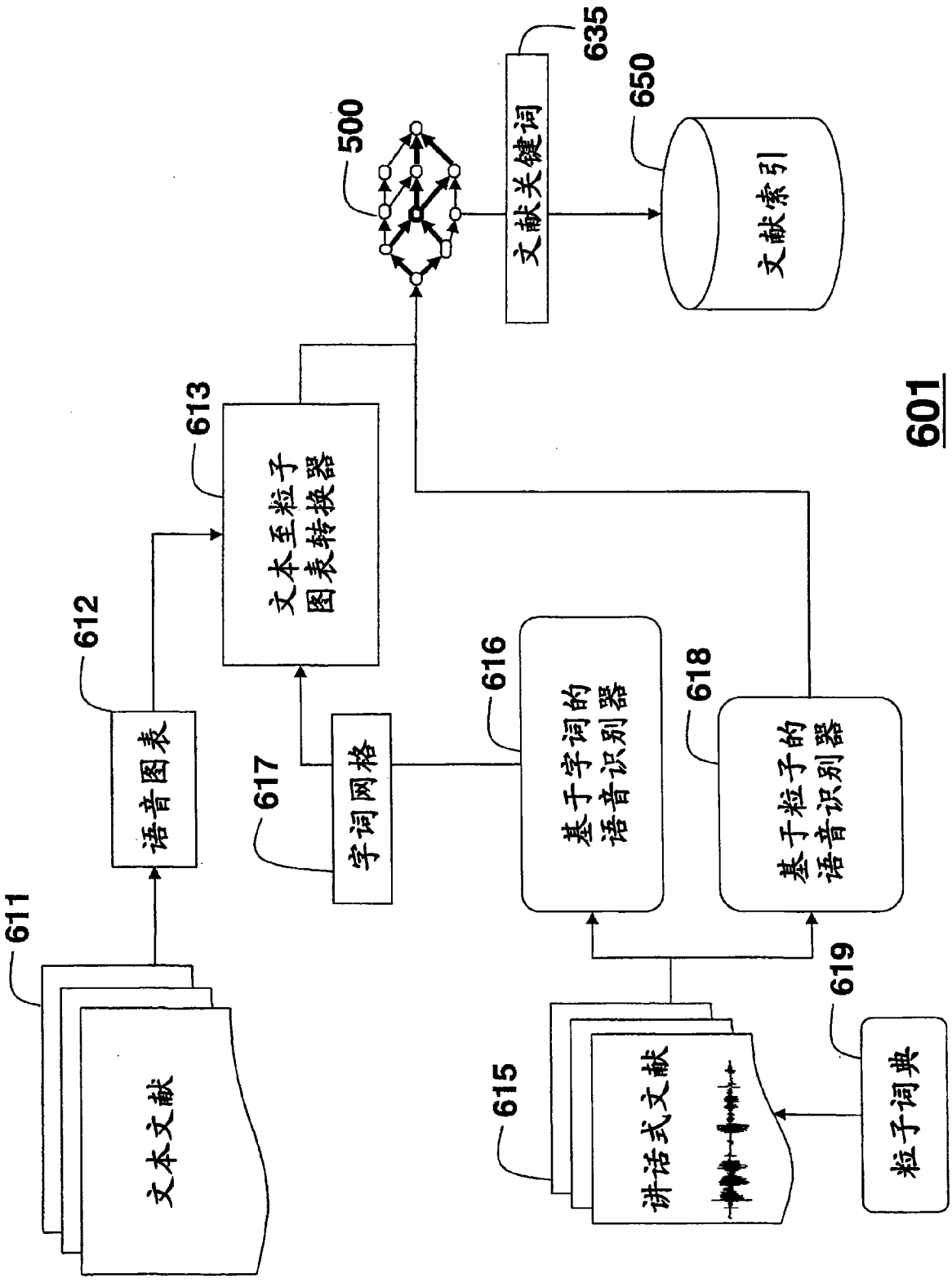

601
611
文本文献
612
语音图表
613
文本至粒子图表转换器
615
讲话式文献
616
基于字词的语音识别器
617
字词网格
618
基于粒子的语音识别器
619
粒子词典
635
文献关键词
650
文献索引
500

图 6A

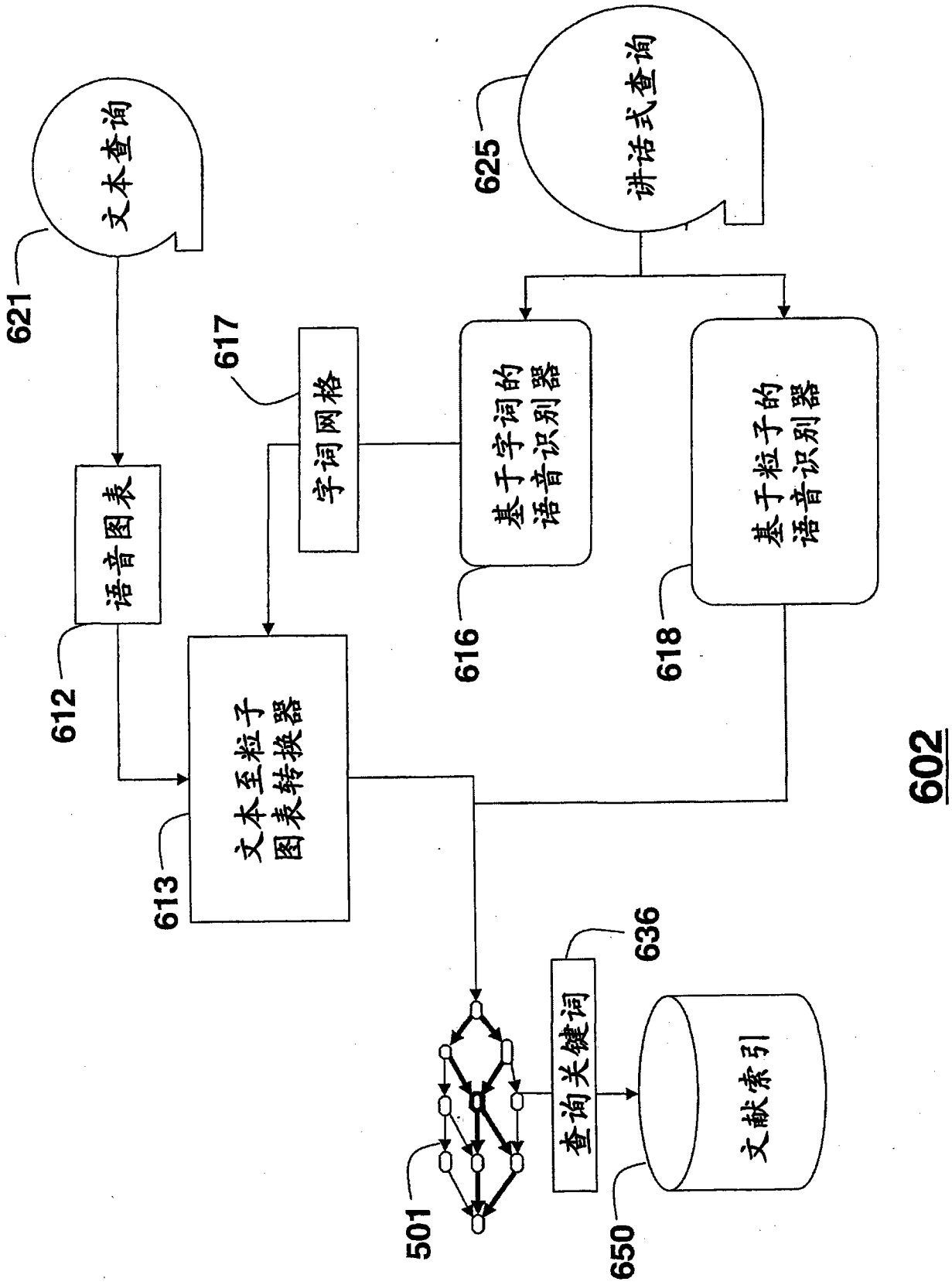

文本查询
621
讲话式查询
625
语音图表
612
字词网格
617
基于字词的语音识别器
616
基于粒子的语音识别器
618
文本至粒子图表转换器
613
602
查询关键词
636
文献索引
650
501

图 6B

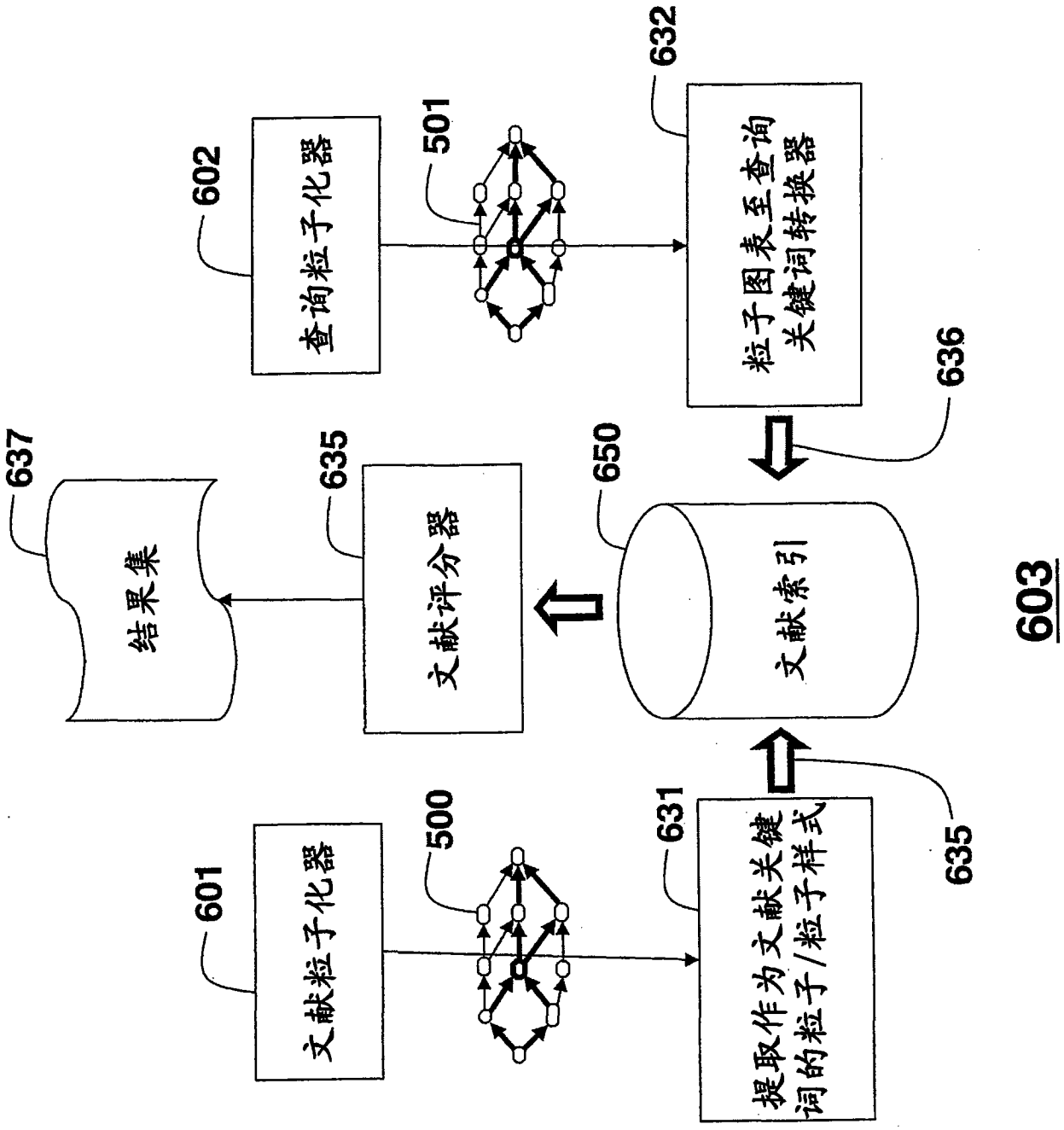

602
查询粒子化器
501
632
粒子图表至查询关键词转换器
636
637
结果集
635
文献评分器
650
文献索引
603
601
文献粒子化器
500
631
提取作为文献关键词的粒子/粒子样式
635

图 6C