

[19] 中华人民共和国国家知识产权局

[51] Int. Cl⁷

G06F 17/30
G06F 17/21

[12] 发明专利申请公开说明书

[21] 申请号 00122279.1

[43] 公开日 2001 年 2 月 7 日

[11] 公开号 CN 1282934A

[22] 申请日 2000.7.28 [21] 申请号 00122279.1
[30] 优先权
[32] 1999.7.30 [33] JP [31] 216617/1999
[71] 申请人 松下电器产业株式会社
地址 日本国大阪府
[72] 发明人 佐藤光弘 伊藤快

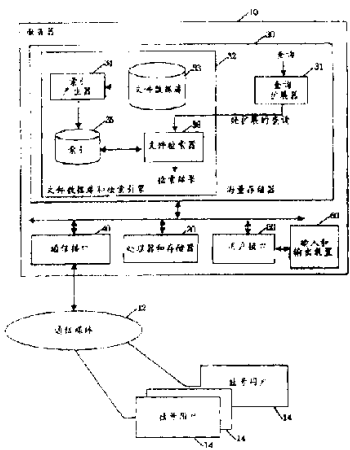
[74] 专利代理机构 上海专利商标事务所
代理人 李 玲

权利要求书 4 页 说明书 14 页 附图页数 10 页

[54] 发明名称 相似字选取和文件检索的方法和系统

[57] 摘要

提供从文件的文件数据库选取给定字母序列的相似字。相似字的出现模式与给定序列的出现模式相似。准备一个特征矢量表,它包括文件数据库中每个字的第 一特征矢量。通过检索相应于给定序列的文件数据库而得到有关给定序列的统计信息。从统计信息,计算相应于给定序列的第二特征矢量。计算第二特征矢量与每个第一特征矢量之间的相似度。选择产生相似度高于预 定值的这种字。



ISSN 1008-4274

1. 一种从文件的文件数据库选取给定查询的相似字的方法，相似字的出现模式和给定查询的出现模式相似，所述方法包括下列步骤：

准备特征矢量表，它包含所述文件数据库中每个字的第一特征矢量；

通过检索相应于所述给定查询的所述文件数据库得到有关所述给定查询的统计信息；

从所述统计信息计算相应于所述给定查询的第二特征矢量；

计算所述第二特征矢量与每个所述第一特征矢量之间的相似度；以及
选择产生的相似度高于预定值的这种字。

2. 如权利要求 1 所述的方法，其特征在于，准备特征矢量表的所述步骤包括下列步骤：

从所述文件数据库产生索引文件，所述索引文件是这样的，可以从所述索引文件得到在所述文件数据库中使用的每个字的统计信息；

从所述索引文件计算每个第一特征矢量，以及

其中所述得到统计信息的步骤包括搜索所述给定查询的所述索引文件的步骤。

3. 一种从文件的文件数据库选取给定字母序列的相似字的方法，相似字的出现模式和给定序列的出现模式相似，所述方法包括下列步骤：

准备特征矢量表，它包含所述文件数据库中每个字的第一特征矢量；

通过检索相应于所述给定序列的所述文件数据库得到有关所述给定序列的统计信息；

从所述统计信息计算相应于所述给定序列的第二特征矢量；

计算所述第二特征矢量与每个所述第一特征矢量之间的相似度；以及
选择产生的相似度高于预定值的这种字。

4. 在包括第一文件数据库和检索器的文件检索系统中，一种扩展给定查询以将经扩展的查询提供给检索器的方法，该方法包括下列步骤：

从文件的第二文件数据库选取所述给定查询的相似字或者所述给定查询中的查询字，所述相似字的出现模式与所述给定查询或所述查询字的出现模式是相似的；及



利用所述给定查询的所述相似字或所述给定查询的所述查询字；

其中选取所述相似字的所述步骤包括下列步骤：

准备特征矢量表，它包含所述第二文件数据库中每个字的第一特征矢量；

通过检索相应于所述查询或所述查询字的所述第二文件数据库得到有关所述查询或所述查询字的统计信息；

从所述统计信息计算相应于所述查询或所述查询字的第二特征矢量；

计算所述第二特征矢量与每个所述第一特征矢量之间的相似度；及

选择产生的相似度高于预定值的这种字作为所述相似字。

5. 如权利要求 4 所述的方法，其特征在于，所述第一和第二文件数据库是相同的東西。

6. 一种接收源语言中的任何给定字母序列和提供目标语言中的译文，致使在源语言文件数据库中所述给定序列的出现模式与在目标语言文件数据库中译文的出现模式相同的方法，其中所述两个文件数据库之一是另一个的译文，所述方法包括下列步骤：

准备特征矢量表，它包含所述目标语言文件数据库中每个字的第一特征矢量；

通过检索相应于所述给定序列的所述源语言文件数据库得到有关所述给定序列的统计信息；

从所述统计信息计算相应于所述给定序列的第二特征矢量；

计算所述第二特征矢量与每个所述第一特征矢量之间的相似度；及

选择产生相似度高于预定值的这种目标语言字作为所述译文。

7. 如权利要求 6 所述的方法，其特征在于，准备特征矢量表的所述步骤包括下列步骤：

从所述目标语言文件数据库产生索引文件，所述索引文件是这样的，可以从所述索引文件得到所述目标语言文件数据库中使用的每个字的统计信息；

从所述索引文件计算每个第一特征矢量，以及

其中得到统计信息的所述步骤包括搜索所述给定序列的所述索引文件的步骤。

8. 一种在包括第一文件数据库和检索器的文件检索系统中将源语言查询转换成目标语言查询以馈送给检索器的查询扩展方法，该方法包括下列步骤：

从目标语言文件的第二文件数据库选取所述源语言查询中每个查询字的译文，所述译文的出现模式与每个查询字的出现模式相似；

用所述查询字的译文代替所述源语言查询中的每个查询字；及

其中选取所述译文的所述步骤包括下列步骤：

准备特征矢量表，它包含所述目标语言文件数据库中每个字的第一特征矢量；

通过检索相应于所述查询字的所述源语言文件数据库得到有关所述查询字的统计信息；

从所述统计信息计算相应于所述查询字的第二特征矢量；

计算所述第二特征矢量与每个所述第一特征矢量之间的相似度；及

选择产生相似度高于预定值的这种目标语言字作为所述译文。

9. 一种接收给定字母序列和提供具有足够划分适宜性的平分字母序列的方法，所述方法包括下列步骤：

形成用于检索文件的检索表达式，所述检索文件包含所述字母序列的所有字母但不包含字母序列本身；

根据所述检索表达式执行检索而得到统计信息；

从所述统计信息产生所述检索表达式的特征矢量，把所述特征矢量称为“掩蔽特征矢量”；

把所述字母序列划分成两个子序列，逐个地移动划分点以得到 N-1 组平分字母序列，这里 N 是在所述字母序列中的字母数；

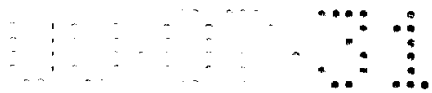
计算每个所述 N-1 组的每个所述平分字母序列的特征矢量；

用掩蔽特征矢量掩蔽每个所述经计算的特征矢量并使经掩蔽经计算的特征矢量归一化；

计算每个所述 N-1 组的相似度，定义所述相似度为每个所述 N-1 组的所述平分字母序列的内积；以及

输出相似度大于阈值的这种平分字母序列。

10. 一种配备接收第一语言的给定字和提供第二语言的译文的翻译装置的系统，所述给定字在第一语言文件数据库中的出现模式与译文在第二语言文件数据库中的出现模式相同，其中所述两个文件数据库之一是另一个的译文，其特征在于，所述翻译装置包括特征矢量表，所述特征矢量表包括所述第二语言



文件数据库中每个字的第一特征矢量；通过检索相应于所述查询字的所述第一语言文件数据库而得到有关所述查询字的统计信息的装置；由所述统计信息计算相应于所述查询字的第二特征矢量的装置；计算所述第二特征矢量与每个所述第一特征矢量之间的相似度的装置；以及选择产生相似度高于预定值的这种第二语言字作为所述译文的装置，一种对给定字寻找至少一个包括给定字的变量的方法，所述方法包括下列步骤：

利用所述翻译装置得到所述给定字的所述译文；

通过用包含所述第一语言文件数据库中每个字的第三特征矢量的特征矢量表代替所述特征矢量表以及通过交换所述第一语言文件数据库和所述第二语言文件数据库来重新配置所述翻译装置；以及

通过利用重新配置的翻译装置得到每个所述译文的第二译文作为所述至少一个变量。

说明书

相似字选取和文件检索的方法和系统

本发明一般地涉及文件检索系统，尤其涉及一种出现基于模式的相似字选取器以及其在查询扩展和多语言文件检索的应用。

由于因特网和个人计算机的扩展，当大量文件已经计算机化和分发时，对文件进行电子搜索变得很普遍。通常，传统的文件检索系统参考用户输入的查询关键字并提供包含关键字的文件作为检索的结果。然而，由于可能存在许多用不同的表达方法描述相同主题事件的文件，对用户来说仅使用用户碰到的关键字来寻找所有这种符合他或她的要求的文件是很困难的。为了这个目的，到目前为止已经提议了各种查询扩展技术，这些技术为扩展检索的给出查询。

由于情况的改善，能通过因特网的扩散而访问世界上的信息，所以以用户母语以外的语言作出检索信息的请求正在日益增加。为了符合这种请求，近来，在多语言文件检索方法方面的研究正在增长。

在日本信息处理协会(IPSJ)的第 52 届全国会议的刊物中，“具有使用字网的查询扩展的信息检索系统”(1996, pp. 4-201-4-202) 是由 Akamine 等人提出的查询扩展技术的一个例子。在该系统中，通过使用固定的同义词词汇寻找查询项的同义名，并用找到的同义名来扩展查询。然而，由于找到的同义名和固定同义词词汇的词汇表有关，它不适用于特殊字段中的检索。

另一个查询扩展技术是由 K.Saito 等人在 IPSJ 研究组报告(信息研究基础 47-10, 1997, pp.67-74)的“基于概念的查询扩展”中提出的。在该系统中，使出现在文件数据库中的待检索的每个字在矢量空间上映射，它的大小相应于构成文件数据库的文件。在矢量空间中的两个字定义的矢量的内积确定两个字之间的相似度。得到与所要求的文件有关的相似字并使用于查询的扩展。

然而，对于未包含在字典词汇表中的字用刚才所述的查询扩展技术不能获得相似程度。为此，该技术不能应用于含有任何未包含在字典中的查询项的查询。

L.Ballesteros 等人的“解决交叉语言检索的意义不明确”(ACM-SIGIR98,1998,pp.64-71)描述了多语言文件检索系统。在参考 3 中，通过使用包

含一组文件和它们的一种或多种语言译文的平行文集(corpora)，对于所给的源语言查询列出所有可能的目标语言译文。通过使用共同出现统计使列出的译文范围变窄而产生目标语言查询。

因为在这个多语言系统中假设使用一种或多种翻译词典，所选择的翻译和翻译词典的词汇有关，这意味着不能得到未包括在词汇中的项目的翻译。如果将系统用于象日文情况那样字之间未留间隔的一种语言的文件数据库，而且如果所给出的查询包括未包含在词汇表中的字母序列(通常是中文字符)，则得不到字母序列的翻译。

本发明试图解决上述的和其它的问题，并使之成为一个目的，提供能够从预定文件数据库选取尚未被广泛接受的给定查询项目的相似字的相似字选取器。

本发明的另一个目的是提供装有相似字选取器的查询扩展技术。

本发明的再一个目的是提供装有相似字选取器的多语言文件检索系统。

根据本发明的一个方面，提供一种方法和一种系统，用于从文件的文件库选取字母的给定序列的相似字。相似字的出现模式与给定序列的出现模式相似。准备了一个特征矢量表，该表包含文件数据库中的每一个字的第一特征矢量。相应于给定序列，通过文件数据库的检索而得到在给定序列上的统计信息。从该统计信息计算相应于给定序列的第二特征矢量。计算第二特征矢量和每个第一特征矢量之间的相似度。选择产生的相似度大于预定值的这种字。

通过产生包括在文件数据库中的字的字组；通过从文件数据库产生索引文件(索引文件是这样的，可以从索引文件得到在文件数据库中使用的每个字的统计信息)；以及通过计算每个来自索引文件的第一特征矢量来准备特征矢量表。为了得到统计信息，搜索给定序列的索引文件。

根据本发明的一个方面，提供一种查询扩展方法和系统，在包括第一文件数据库和检索器的文件检索装置中使用。扩展给定查询，馈送至检索器。为了这个目的，利用在给定查询中的查询字的相似字。从文件的第二文件数据库选取查询字的相似字。相似字的出现模式与查询字的出现模式相似。通过准备包含第二文件数据库中每个字的第一特征矢量的特征矢量表；通过相应于查询字的第二文件数据库的检索而得到查询字上的统计信息；通过对相应于来自统计信息的查询字的第二特征矢量的计算；通过计算第二特征矢量和每个第一特征

两个子序列，逐个地移动划分点以得到 N-1 组平分字母序列；对每个 N-1 组的平分字母序列的每一个计算特征矢量；用掩蔽特征矢量掩蔽每一个经计算的特征矢量并使经掩蔽经计算的特征矢量归一化；对每个 N-1 组计算相似度，把相似度定义为每个 N-1 组的平分字母序列的内积；以及输出相似度大于一个阈值的这种平分字母序列。

根据本发明的另一个方面，对于给定字，通过使用上述译文选取系统至少可以得到一个包括给定字的变量。为了达到这一点，通过使用译文选取系统首先得到关于给定字的译文。通过用含有第一语言文件数据库中每一个字的第三特征矢量的特征矢量表代替特征矢量表以及交换第一语言文件数据库和第二语言文件数据库来配置译文选取系统。通过使用所配置的译文选取系统，得到对于每一个译文的第二译文作为变量。

从下面对本发明的示例实施例和附图的描述，本发明的特征和优点将显而易见，其中：

图 1 是示意方框图，示出可以实施本发明的文件检索计算机系统的示例布局；

图 2 是示出本发明的相似字选取器的示例布局的图；

图 3 是流程图，示出在根据本发明的相似字选取器 130 的控制下处理器 20 执行的操作；

图 4 是示出特征矢量表 160 的示例结构的图；

图 5 是流程图，示出使用图 3 的相似字选取器 130 的查询扩展器的操作；

图 6 是示出译文选取系统 1300a 的布局的图，所述译文选取系统用于交叉语言检索的查询扩展器；

图 7 是结合选取器 130 执行的查询扩展程序 31b 的流程图；

图 8 是流程图，示出根据本发明的字母序列划分器的示例操作；

图 9 是示出关于字母序列“自律移动ロボット”的检索表达式的例子的图；

图 10 是示出关于字母序列“自律移动ロボット”的七组划分序列的图；

图 11 是示出步骤 212 的处理例子的图

图 12 是示出判定每个划分序列组的划分是否足以适宜的图；以及

图 13 是流程图，示出根据本发明的变量寻找器 300 的示例操作。

对于所有的图纸，当在多张图纸中示出同一个元件时，以同一个参考号来指定。

图 1 是示意方框图，示出可以实施本发明的文件检索计算机系统的示例布局。文件检索计算机系统 1 一般包括服务器 10、通信媒体 12 和多个用户 14。在系统 1 中，服务器接收来自(一般是远程的)用户 14 的查询并响应地将检索结果返回给用户。通信媒体 12 通常包括诸如因特网、局域网等的各种通信网络。虽然在图 1 中示出一种用户-服务器配置，但是本发明可以作为独立计算机系统实施，在这种情况下，将从图纸中删除通信媒体 12 和用户 14。

服务器 10 一般是计算机，具有同时向几个用户提供文件检索服务所必需的容量和速度，它包括：一个或多个作为计算机核心的处理器和存储器 20；存储各种程序和数据的海量存储器 30；与用户 14 通信的通信接口 40；用户接口 50；以及允许管理器对服务器 10 进行操作的输入和输出装置 60。海量存储器 30 最好包括硬盘(未示出)，用于存储文件检索所必需的程序和数据。海量存储器 30 可以进一步包括，例如一个或多个光盘(未示出)，用于存储文件数据库 33。上述服务器 10 和/或元件 20、30 和 40 到 60 可以是任何合适的标准元件。

每个用户 14 可以是任何合适的终端或带诸如调制解调器或网络接口之类的通信接口(未示出)的个人计算机。

海量存储器 30 一般存储文件数据库/检索核心程序 32 和查询扩展器 31，用于接收来自用户 14 的查询并提供根据下面描述的本发明的原理的查询的扩展版本。文件数据库/检索核心程序 32 和查询扩展器 31 可以是任何合适的文件数据库和任何合适的检索核心程序的组合，一般包括包含大量文件的文件数据库 33、索引产生器 34、来自数据库 33 的由索引产生器 34 产生的索引文件 35 以及通过使用索引文件 35 检索大多数和查询有关的文件的文件检索器 36。

在说明下列解说性的实施例之前先定义几个所用的术语，这对较好地理解本发明是有帮助的。

除非另行表明，应把术语“文件”表示为完整的文件或它的任何部分，诸如标题、摘要、或一个或多个从句、句子或段落。

除非另行表明，应把术语“查询”表示为为了从文件数据库 33 选择文件子集的目的而输入的文本。查询一般包括多个逻辑表达式或自然语言形式的查询项目。每个查询项目是一种支持语言的字母序列，这种序列一般包括支持语

言的一个或多个字。即，构成查询的每个字母序列可以是单个的字、复合的字、短语和其它多-字结构。根据所用的语言，单个字之间可以留有间隔也可以不留间隔。

除非另行表明，应把术语“字”包括单个字、复合字、短语和其它多-字结构。此外，术语“字”和“项目”通常可互换地使用。例如，项目和字包括名词、代名词、复合的名词性的词、名词短语、动词、副词、数字表达式和形容词。

下面，我们揭示：

- (1)相似字选取器；
- (2)使用相似字选取器的查询扩展器(31a)；
- (3)查询扩展器(31b)，它同支持与查询使用不同语言的文件数据库/检索核心程序 32 相适配(例如，语言交叉语言检索的查询扩展器)；
- (4)字母序列划分器；
- (5)变量(或等效字)寻找器；以及
- (6)根据相似字选取器和结合字母序列划分器和变量寻找器的较佳查询扩展器(31c)。

相似字选取器、字母序列划分器和变量寻找器是用于查询扩展器 31 中的有力工具。从下面的说明可以看到，用相似字选取器来实现字母序列划分器和变量寻找器。

相似字选取器

图 2 是示出根据本发明的相似字选取器系统 1300 的示例布局图。在图 2 中，从源到目的地画出的一路箭头表示目的地的数据是从源得到的，而在表示程序的方框和表示数据结构的方框之间画出的两路箭头表示程序指数据结构。

在相似字选取的实际操作之前，最好从第二文件数据库 33a 准备在相似字选取操作中使用的第二索引文件 134 和特征矢量表 160。

虽然第二文件数据库 33a 可能与文件数据库/检索核心程序 32 的文件数据库 33 不同，但是强烈地建议第二文件数据库 33a 应与文件数据库 33 相同。即使第二文件数据库 33a 和文件数据库 33 不同，第二文件数据库 33a 最好具有与文件数据库 33 的特征相同的特征，以提高选取字的正确度(即，相关性

(relevance))。(为了这个原因,此后把第二文件数据库 33a 称为“文件数据库 33a”或简称为“数据库 33a”)。

另一方面,第二索引文件 134 没有必要和文件数据库/检索核心程序 32 的索引文件 34 相同。然而,要注意,最好这样安排第二索引文件 134,当文件检索器 136 为了包含在来自用户 14 的查询中的查询项目而搜索第二索引文件 134 时,文件检索器 136 可以得到诸如在数据库 33a 中每个文件中的查询项目的出现频率(称之为“基于文件的项目频率”)和包含查询项目的文件数(即使查询项目是任何字母序列,即即使查询项目是字典中找不到的这种字)之类的统计信息。根据日本未审查专利申请第 Hei-249354,本申请人可以实现这种索引文件 134。

产生特征矢量表 160 如下。假设在数据库 33a 中的文件 $\{D_i \mid i=1,2,\dots,M\}$ 数是 M ; 以及给出合适的字组 (WORD SET)150, 包括项目(或字) $\{T_j \mid j=1,2,\dots,N\}$, 其中 N 是在数据库 33a 中的字种类的总数。对于在字组 150 中的每个字 T_j , 计算 M -维特征矢量 t_j 。定义每个特征矢量 t_j 为

$$t_j = (w(j,1), w(j,2), \dots, w(j,i), \dots, w(j,M)) \quad (1)$$

如从表达式(1)所看到,特征矢量 t_j 的单元 $w(j,1), w(j,2), \dots, w(j,i), \dots, w(j,M)$ 相应于文件 $D_1, D_2, \dots, D_i, \dots, D_M$ 。根据本领域所公知的 $tf(j,i) * idf(j)$ 计算(score)来计算特征矢量 t_j 的第 i 个单元 $w(j,i)$ 作为权重。“ $tf(j,i)$ ”是在给定的文件中项目 T_j 的出现次数,而且被称为在文件 D_i 中的项目 T_j 的“项目频率”。把“ $idf(j)$ ”称为单元 T_j 的“逆文件频率”,并定义为:

$$idf(i,j) = \log(M/d_j)$$

其中 d_j 是在文件中有项目 T_j 出现的文件数。(对于 $tf(j,i) * idf(j)$ 计算的一步说明参考 1983 年 McGraw-Hill 出版公司的 G.Saltion 和 M.J.McGilld 的“现代信息检索介绍”)。在这种情况下,可以如此地归一化特征矢量 t_j , 使之幅度或长度为 1。这样,产生了特征矢量表 160, 而且现在已准备好相似字选取操作。

图 3 是流程图, 示出处理器 20 在根据本发明的相似字选取器 130 的控制下执行的操作。响应于接收到来自远程用户 14 的查询, 调用相似字选取器 130。尽管在以下特定的例子中可获得给定查询中每个查询项目的相似字, 应当注意, 用与以下描述相同的方式可以获得给定查询自身的相似字。

如果收到所接收查询中的第 k 个查询项目 q_k , 则处理器 20 根据接收到的

查询项目 q_k 搜索第二索引文件 134，并在步骤 136 中输出每个文件 D_i 的查询项目 $tf(q_k, i)$ 的项目频率和查询项目 $idf(q_k)$ 的逆文件频率。在步骤 138 中，处理器 20 从逆文件频率 $idf(q_k)$ 计算给定查询项目 q_k 的特征矢量 q_k 以及查询项目的项目频率 $tf(q_k, 1), tf(q_k, 2), \dots, tf(q_k, M)$ 。

步骤 40 对在特征矢量表 160 中应经受下列处理的记录作出标志。为了简化该步骤，最好如图 4 那样配置特征矢量表 160 的结构。特征矢量表 160 的每个记录可以包括：包括分类代码、子分类代码等的一个或多个分类字段(CATE 1, CATE 2, ...)168；以及包括表示对记录作出标志的特殊代码的标志字段 166；以及字字段 162 和特征矢量字段 164。如果在下一个步骤中要求仅对一个或多个特殊字段的字计算相似度，则在记录(在所述记录中分类字段 168 的值落在特殊字段上)的标志字段 166 中写入特殊代码。还有，如果要求从下面的处理排除查询项目本身，如果项目 q_k 存在于特征矢量表 160 中，即在字组 150 中，则只要从查询项目 q_k 的记录的标志字段 166 中除去特殊代码。

然而，应注意，经常不能在字组 150 或特征矢量表 160 中找到查询项目 q_k ，只因为通过文件检索器 136 已经对于查询项目 q_k 成功地得到逆文件频率和项目频率。这是因为如此地安排第二索引文件 134，以致造成对于如上所述的任何查询项目 q_k 得到逆文件频率和项目频率。换句话说，即使对于字组 150 中找不到的查询项目（或者即使对于查询本身），通过利用第二索引文件 134 的检索能够获得这种统计信息。

然后，步骤 142 计算算出的特征矢量 q_k 和在特征矢量表 160 中的每个带标志的记录特征矢量之间的相似度。计算特征矢量之间的相似度作为特征矢量的内积。例如，查询项目“www”和字“internet”之间的相似度，即， $SIM(www, internet)$ 计算如下：

$$\begin{aligned} SIM(www, internet) \\ &= 1.10 \cdot 0.15 + 0.00 \cdot 0.00 + 0.12 \cdot 0.01 + \dots + 0.07 \cdot 0.10 \\ &= 0.9 \end{aligned}$$

另一方面，使用诸如交互信息、t-计算(t-score)等统计信息可以计算矢量之间的相似度。对于交互信息和 t-计算(t-score)的进一步的说明参阅 K.W.Church 和 R.L.Mercer 的“使用大文集的计算机化语言学的特殊版本的介绍”，(Computational Linguistics, Vol. 19, No. 1, 1993, pp. 1-24)。

步骤 144 将带标志记录字以计算得到的相似度为次序而分级。步骤 146 将相似性大于预定值的字作为相似字输出，并结束操作。在图 2 中，对于查询项目“www”，字“internet”和“HTML”作为相似字输出。

在上述实施例中，已经使用了在文件数据库 33a 中的所有的 M 个文件。然而，可以通过字段将数据库 33a 划分成相关字段的文件子集；为每个文件子集准备第二索引文件 134 和特征矢量表 160；以及对于给定的字段，使用和给定字段相关的第二索引文件 134 和特征矢量表 160。另一方面，可以对每个所需字段准备文件数据库 33a、第二索引文件 134 和特征矢量表 160。

在上述实施例中，在实际选取相似字之前已经准备了第二索引文件 134 和特征矢量表 160。然而，如果字组 150 包含极少的字，则在步骤 138 中，可以在计算特征矢量之后产生第二索引文件 134 和特征矢量表 160。

虽然已经使用字组 150 来产生特征矢量表 160，但是也可以从数据库 33a 直接产生特征矢量表 160。

字组 150 可以包含所有出现在文件数据库 33a 中的字或仅在一个或多个特殊字段中的字。通过限制字组 150 的词汇，相应地可以限制在特征矢量表 160 中的记录。

使用相似字选取器的查询扩展器

图 5 是流程图，说明使用图 3 的相似字选取器 130 的查询扩展器的操作，查询扩展器 31a 是图 1 的查询扩展器 31 的说明实施例。在图 5 中，步骤 102 从给定的查询中选择第一查询项目。步骤 104 通过使用相似字选取器 130 寻找所选择查询项目的相似字。步骤 106 以例如所找到的相似字的逻辑和来代替在查询中的所选择的查询项目。判定步骤 108 进行测试以观察在查询中的的项目是否已经被提取(或代替)。如果没有，则将控制传到步骤 110，在该步骤中选择在给定的查询中的下一个项目并使控制返回步骤 104。如果在步骤 108 中的测试结果是“是”，则将控制进到步骤 112，在该步骤中将经扩展的查询传到文件检索器 36 并使操作终止。

在接收到经扩展的查询时，文件检索器 36 以传统的方法根据所接收到的经扩展的查询搜索索引文件 35。

例如，为了知道有关在 WW0W(万维网)上的搜索引擎，在客户处用户输入(www<and> “search engine(搜索引擎)”)作为查询，然后相似字选取器 130 将

提供“internet(因特网)”作为查询项目“www”的相似字以及“retrieval(检索)”作为查询项目“search engine”的相似字，查询扩展器 31 相应地扩展查询(www<and> “search engine”)以提供经扩展的查询((www<or> internet)<and> (“search engine”<or> retrieval)。这使包含表达式“retrieval services on the internet(在因特网上检索服务)”的文件得到检索的结果。

在查询扩展器 31a 中，已经用相似字选取器 130 扩展了所有的查询项目。另一方面，只有在字组 150 中找不到的查询项目的项目才可以用字选取器 130 扩展。

已经把在步骤 104 中对所选查询项目找到的相似字代替在步骤 106 查询中所选的(或原始的)项目。另一方面，可以把所找到的相似字增加到原始项目中。

尽管查询扩展器 31a 已经扩展了在给定查询中知道的每个查询字，但是可以如此配置查询扩展器 31a，即对相似字选取操作测试单个检索条件；将所产生的检索条件传送到相似字选取器 130 以获得相似字；并把所获得的相似字传送到文件检索器 36。

交叉语言检索的查询扩展器

图 6 示出译文选取系统 1300a 的布局图，该系统在交叉语言检索的查询扩展器中使用。如从图 6 所见，译文选取系统 1300a 与图 2 的相似字选取系统 1300 极相似。

除了译文选取系统 1300a 采用源语言索引文件 235 而不是第二索引文件 134 之外，两个选取系统 1300a 和 1300 在实际选取操作方面是相同的。从 SL 文件数据库 233 传送源语言(SL)索引文件 235。由于这个原因，以 SL 索引文件 235 代替第二索引文件 134，就可以使图 3 的流程所示的相似字选取器 130 用于译文选取操作。还有，在该情况下，如果在步骤 146 中要输出多于一个的译文，则最好以逻辑积的形式来输出译文。

还有，与相似字选取系统 1300 的不同之处在于，特征矢量表 160 是由从 TL 文件数据库 333 产生的给定目标语言(TL)字组 350 和 TL 索引文件 335 两者产生的。图 6 的 TL 文件数据库 333 最好和图 1 的文件数据库 33 相同。然而，然而，文件数据库 333 可以不同于图 1 的文件数据库 33，如果这样，则两个数据库 333 和 33 应具有相同的语言和相同的字段。

SL 文件数据库 233 和 TL 文件数据库 333 必须互相作为译文。如果不能

获得一个文件数据库的译文文本，通过另一个文件数据库 333 或 233 的机器译文分别可以得到一个文件数据库 233 或 333。

图 7 是程序 31b 的流程图，当与在译文选取系统 1300a 中的译文选取器 130 结合而执行时，该程序作为交叉语言检索的查询扩展器而操作。根据收到的查询而执行查询扩展器程序 31b。在图 7 的步骤 122 中，处理器 20 把给定的查询转换成“积的和”的形式。步骤 124 把在和中的每个积传送到选取器 130，以得到结果(在该情况下的译文)。然后，步骤 126 把结果的逻辑和传送到文件检索器 36。

这样，使查询扩展器程序 31b 适应于文件数据库/检索引擎 32，它支持与查询所使用的语言(即源语言)不同的语言(即目标语言)。要注意，可以互换地使用查询扩展器 31a 和 31b 并可以被用作单种语言检索的查询扩展和交叉语言检索的查询扩展两种情况。尽管在步骤 122 中将给定查询转换为积的和的形式，但是可以将给定查询传送给查询扩展器 130。

正如从图 6 可见，如果在客户处用户 14 输入日文(情报检索<or>情报抽出)，它相应于英文的(“information retrieval” <or> “information extraction”)，则在译文选取系统 1300a 中的选取器 130 对查询项目“情报检索”提供译文“information”和“retrieval”，而对查询项目“情报抽出”提供译文“information”和“extraction”。如果对查询项目选取多于一个的译文，则译文选取系统 1300a 的选取器 130 输出如上所述的逻辑积形式的译文。相应地，把检索情况

((information<and>retrieval)<or>(information<and>extraction))传送到文件检索器 36。

如从图 6 可见，一旦已经产生特征矢量表 160 和 SL 索引文件 235，就不再需要 SL 和 TL 数据库 233 和 333、TL 字组 350 和 TL 索引文件 335。通过机器翻译在一种语言中的文件数据库来准备各种语言的特征矢量表 160，一种查询扩展器适应于多语言文件检索。

字母序列划分器

图 8 是流程图，示出根据本发明的字母序列划分器 200 的示例操作。如果字母序列划分器 200 接收查询项目，即字母序列(例如，“L1L2L3…Ls，其中 s 是字母数目)，则步骤 202 形成检索表达式

$$(L1<and>L2<and>L3\cdots<and>Ls)<not> L1L2L3\cdots Ls。$$

图 9 示出对字母序列“自律移动ロボット”的这种检索表达式的例子。这是一个日文项目，意思是“自主移动机器人”。

步骤 204 检索满足检索表达式的文件并得到诸如在步骤 136 中描述的统计信息。步骤 206 产生用于检索表达式的特征矢量(称之为“掩蔽特征矢量”)。满足检索表达式的文件是诸如包括任何字母序列的字母但不包括字母序列本身之类的文件。例如，选择仅包含一个表达式“自律的に移動するロボット”的文件，而不选择仅包含一个表达式“自律移動ロボットについて”的文件。

步骤 208 把字母序列划分成两部分，逐个地移动划分点以得到 s-1 组经划分的序列。可以通过词态分析的方法得到划分，如果它产生经划分的子序列。图 10 示出对于字母序列“自律移動ロボット”的七组经划分的序列。步骤 210 对于在每组中每个经划分的序列计算特征矢量。例如对于一组经划分的序列(自、律移動ロボット)，步骤 210 计算对子序列“自”的特征矢量和对子序列“律移動ロボット”的特征矢量。在相应于包括子序列“自”的文件的大小的元件中，对于子序列“自”的特征矢量具有正值。例如，在仅包含“自由の追求”的文件的大小中，对于子序列“自”的特征矢量具有正值。

然后，步骤 212 用在步骤 206 中得到的掩蔽特征矢量掩蔽每个经计算的特征矢量，并使经掩蔽的特征矢量归一化，以致长度为 1。用掩蔽特征矢量对特征矢量进行掩蔽意味着留下相应于这种具有正值的特征矢量的元(如它们是的话)的特征矢量元件，而使其它特征矢量元为零。图 11 是示出步骤 212 的处理的例子。在图 11 中，用掩蔽特征矢量掩蔽一组(自、律移動ロボット)的特征矢量。在图中，具有 0 值的元件或大小用“o”来表示，而具有正值的元或大小用“x”来表示。在该例子中的组合特征矢量中仅包含“自由の追求”的文件的大小、仅包含“自律移動ロボット”的文件的大小等具有 0 值；而只有两个字母子序列“自”和“律移動ロボット”独立地出现而彼此不连接的文件的大小才具有正值。

通过计算一组特征矢量的内积，步骤 214 寻找每一组的相似度。图 12 是示出经划分的序列组的图，由 0.35 的阈值来判断相应的相似度和相应的划分的适宜性。根据本实施例，如果在文件中更频繁地共同出现字母序列组的经划分的序列，则字母序列组的经划分的序列之间的相似度变成较高。由于在步骤 202 和 212 中排除包含如那样的字母序列的文件，经划分的序列频繁独立地出现产

生较高的相似度。认为独立出现字母序列是它们自己有意义的那些。

步骤 216 输出相似度等于或大于阈值的字母序列的经划分的序列。由于在图 12 的例子中将阈值设置为 0.35, 判断(自律、移动ロボット)和(自律移动、ロボット)两个经划分的序列组有较佳较适宜的划分, 并相应地输出。

如上所述, 根据本发明的字母序列划分器 200 尝试对给定字母序列的所有可能的划分, 以产生具有足够适宜度的平分字母序列。

在图 5 和 7 中具有标为“A”的点的路径上, 可以利用字母序列划分器 200。例如, 在图 5 中, 如果在字组 150 中没有找到项目, 则从使用字母序列划分器 200 的项目得到平分字母序列, 并代替项目而使用。

变量寻找器

图 13 是流程图, 示出根据本发明的变量寻找器 300 的示例操作。在图 13 中, 步骤 302 把第一语言索引文件和第二语言特征矢量表 160a 附到图 3 的译文选取器 130 上。通过使用译文选取器 130, 步骤 304 寻找输入第一语言查询项目的第二语言译文。如果输入第一语言查询项目是, 例如日文项目“ギリツァ”, 它相应于英文字“Greek”, 则假设得到第二语言(假定英文)翻译成“GREECE”作为译文。

然后, 在步骤 306, 分别用译文选取器 130 中的第二语言索引文件和第一语言特征矢量表代替第一语言索引文件和第二语言特征矢量表。步骤 308 通过使用译文选取器 130 寻找每个第二语言译文的第一语言译文。在该步骤中, 输入项目“GREECE”将产生三个译文“ギリツァ”、“ギリツヤ”和“ギリツア”输出。在该情况下, 对于输入项目“ギリツァ”得到这三个变量。

如果一个字有变量, 则在文件中通常使用相同的表达式。由于在单个文件中使用变量较为困难, 虽然这些变量彼此极相似, 但是相似字选取器 130 通常不能选取这种变量。

根据本发明的变量寻找器 300, 可以得到与输入项目的第二语言译文相同的字母序列第二语言译文作为输入项目的变量。换言之, 如果一个字有一个或多个第一语言的变量而第二语言只有一个相应于该字的表达式, 则作为这种字群的成员, 通过使用变量寻找器 300 可以得到其它的变量。

在图 5 和 7 中在标为“B”的点的路径上, 可以利用变量寻找器 300。特别, 对以日文“katakana”字母写的字搜索所得到的查询, 这些字趋向于具有

变量。对于每个所找到的字，通过使用变量寻找器 300 来寻找变量，而使用所找到的变量的逻辑和以及原始字来代替原始字。这样做进一步扩展了查询，造成有效地检索文件的结果。

在本申请中提到的所有文章和参考文献(包括专利文献)的揭示内容在此整体地作为参考。

可以构成许多不同的本发明的实施例而不偏离本发明的精神和范围。不用说，本发明不限于本说明描述的特殊实施例，而是由所附的权利要求书所定义。

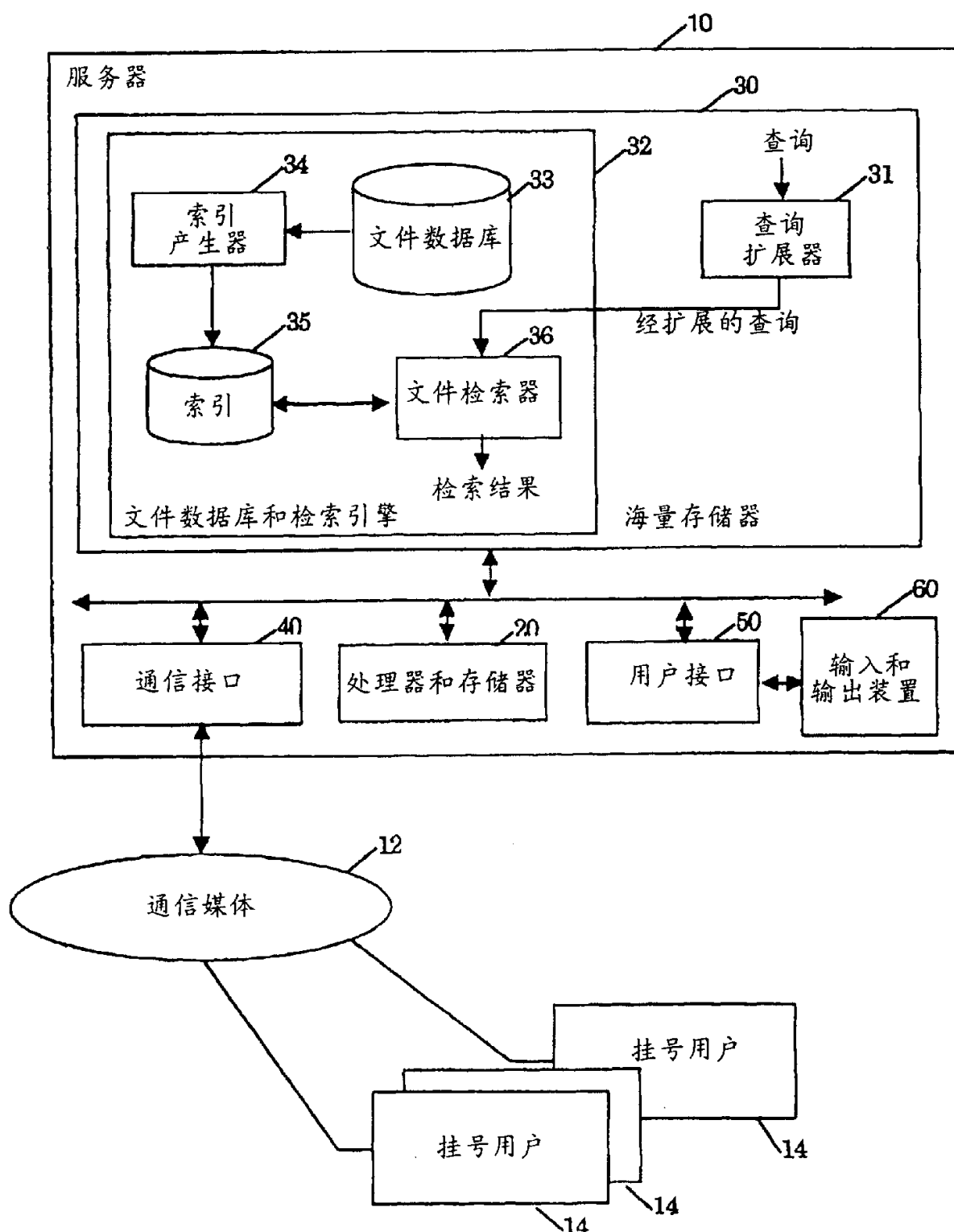


图 1

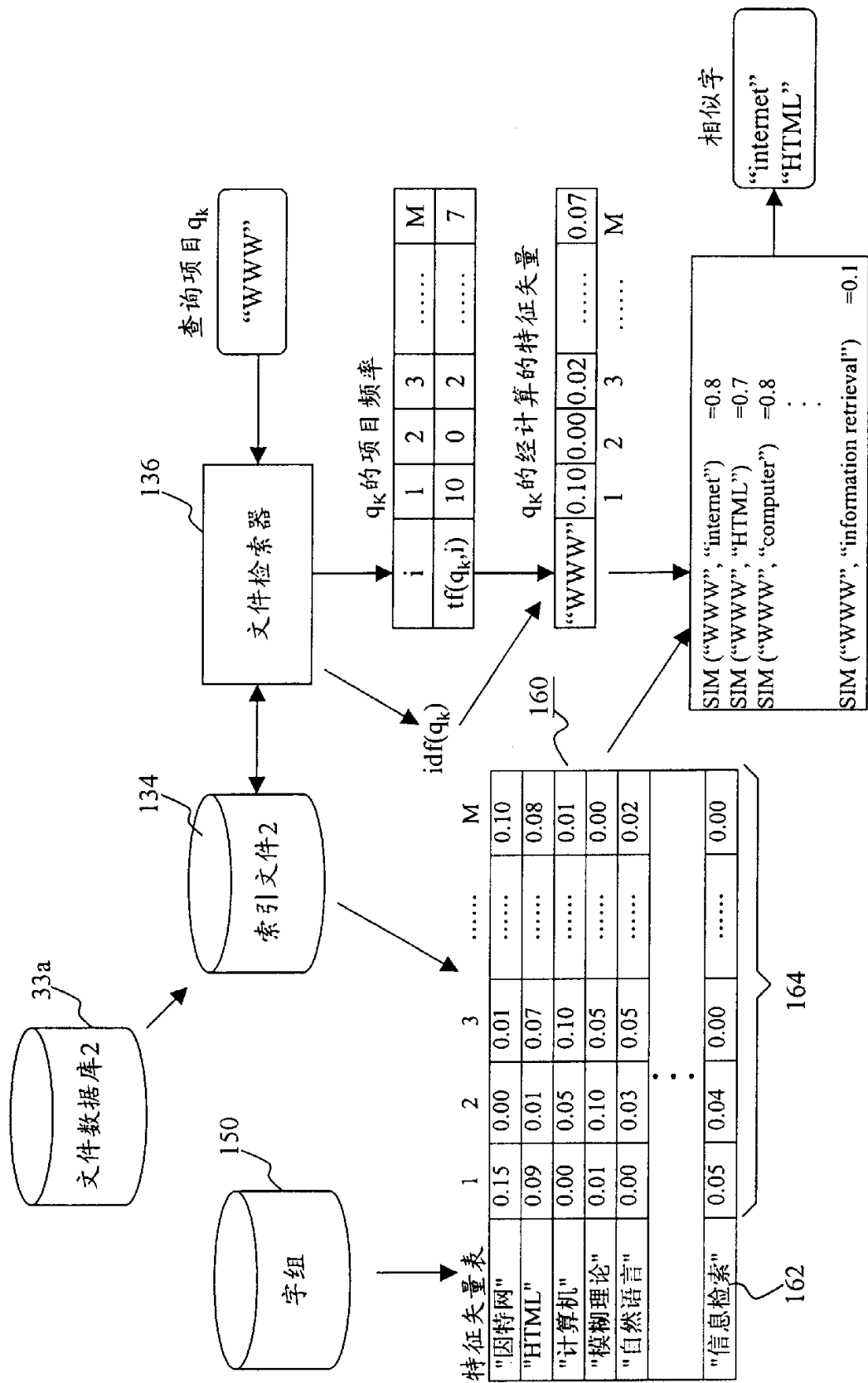


图 2

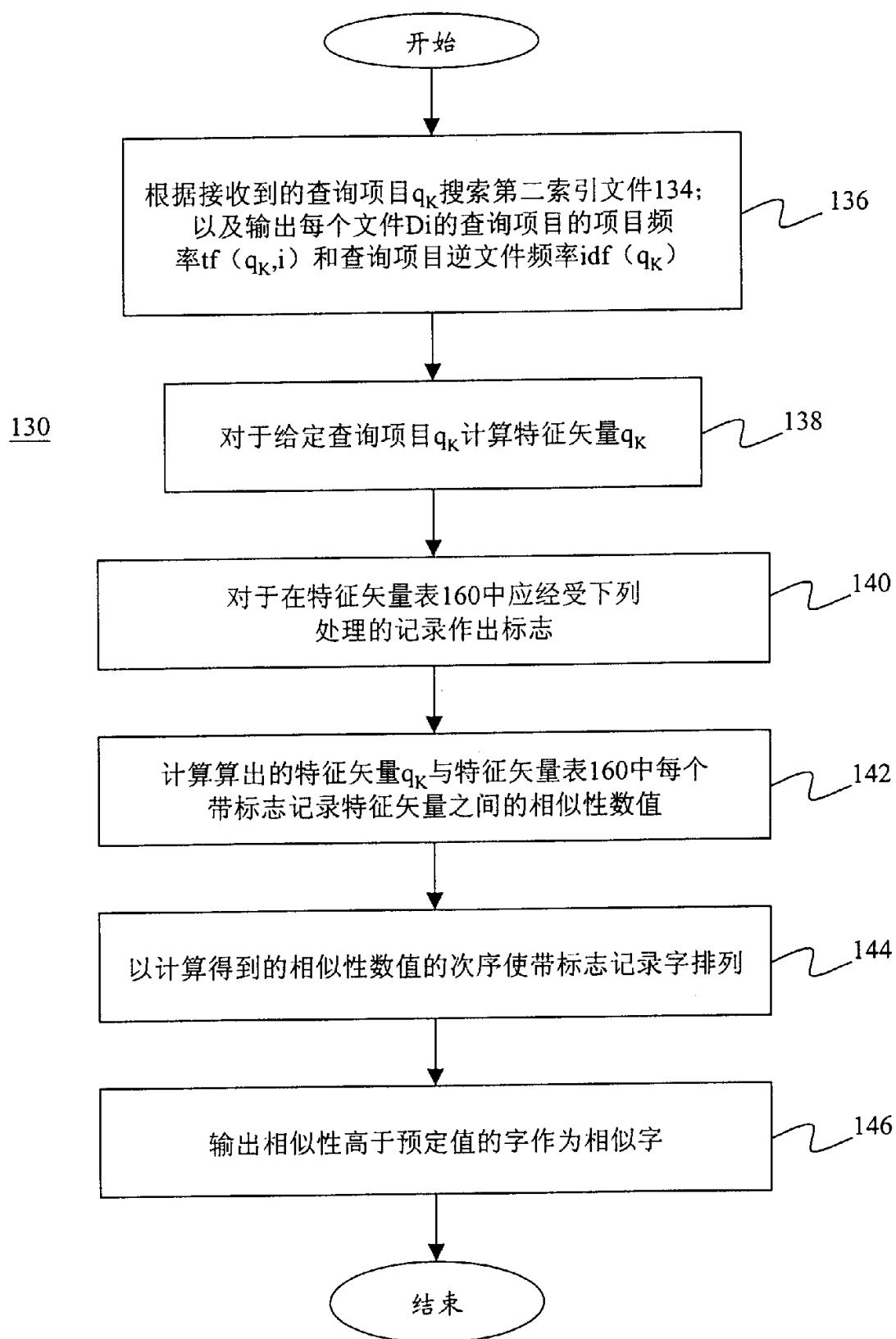


图 3

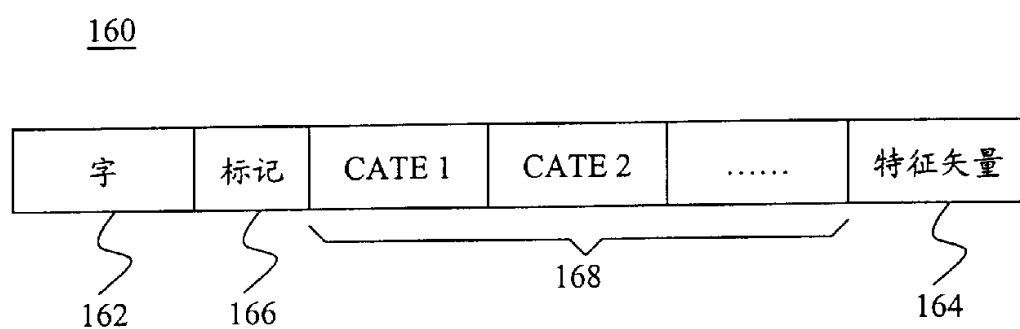


图 4

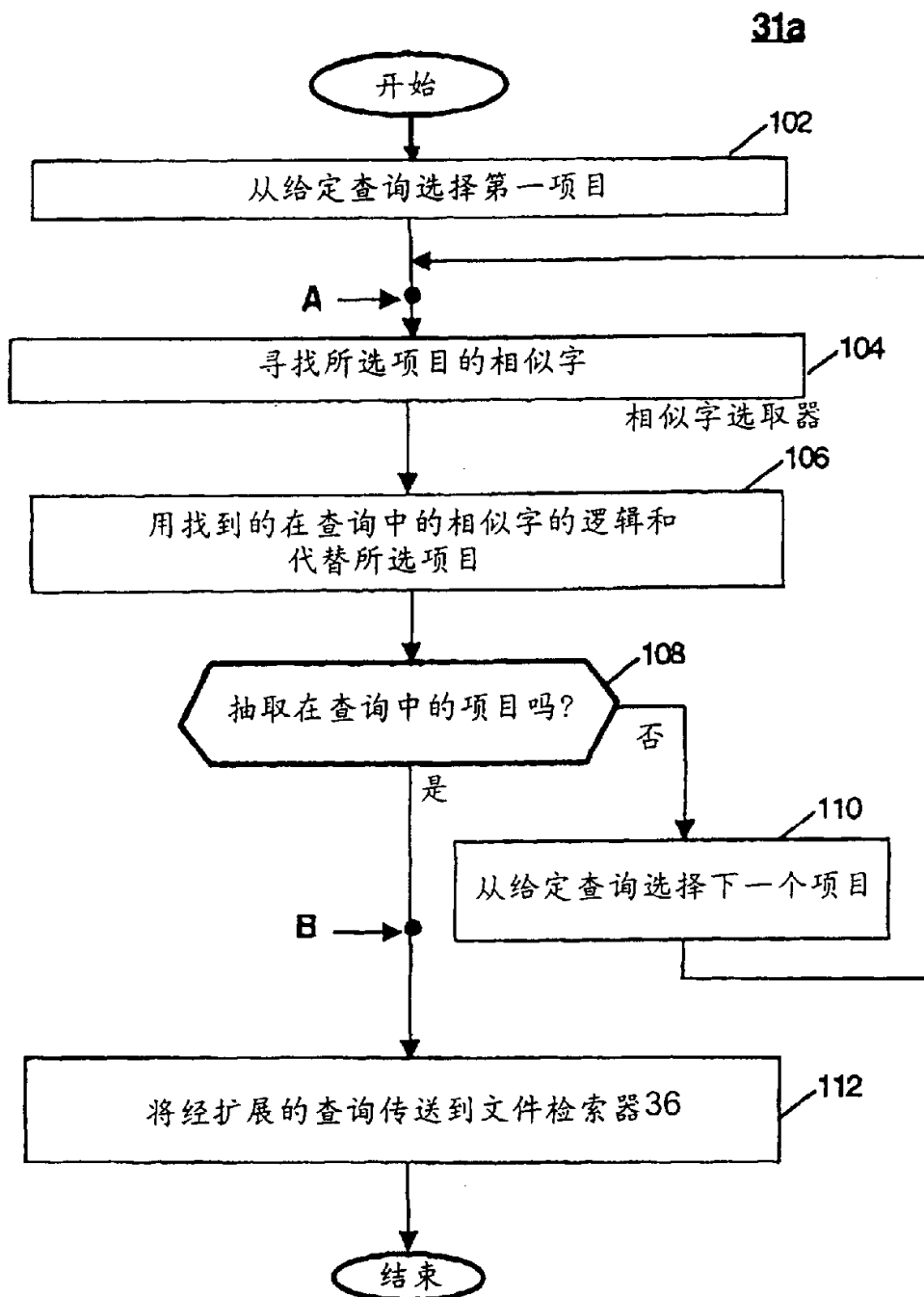


图 5

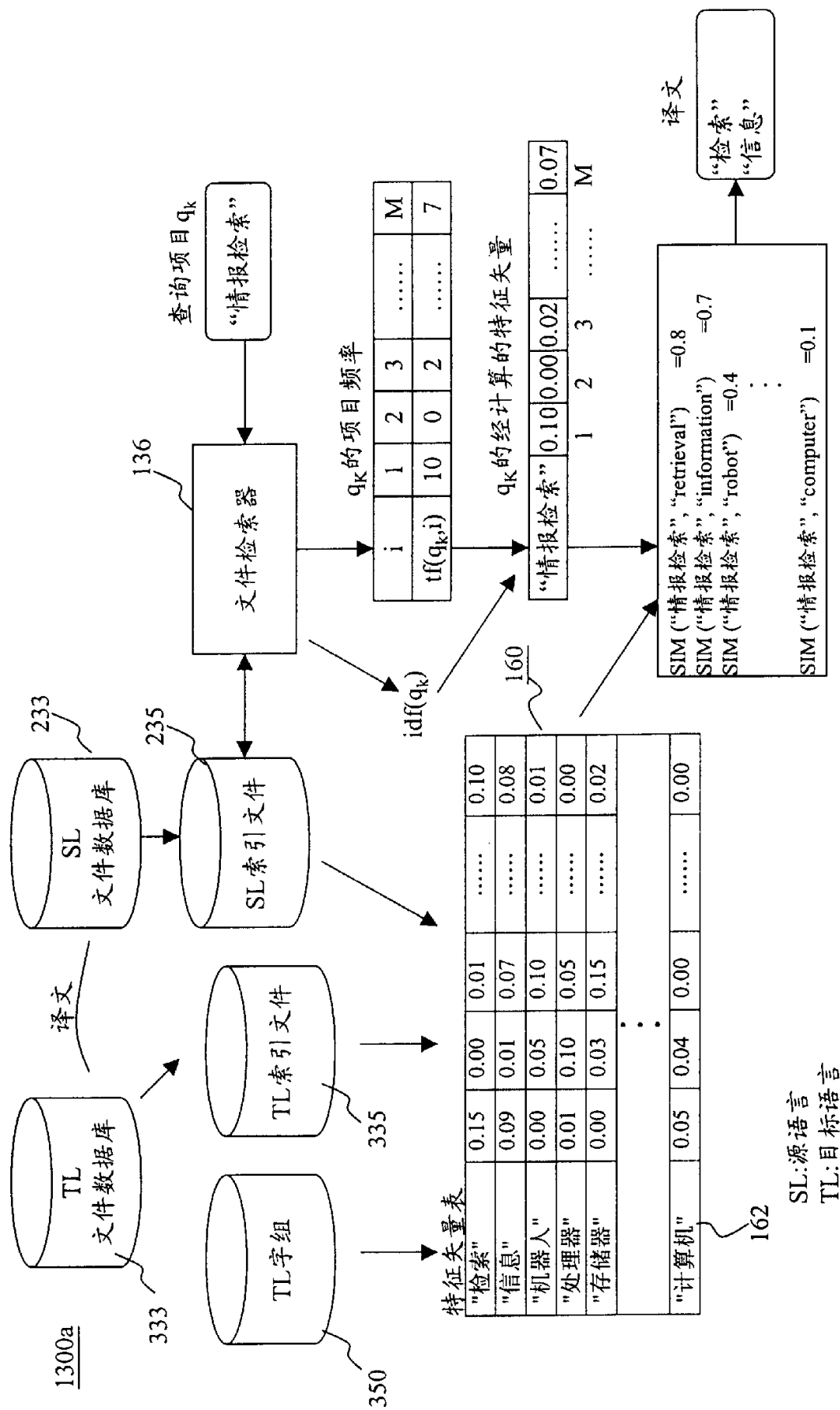


图 6

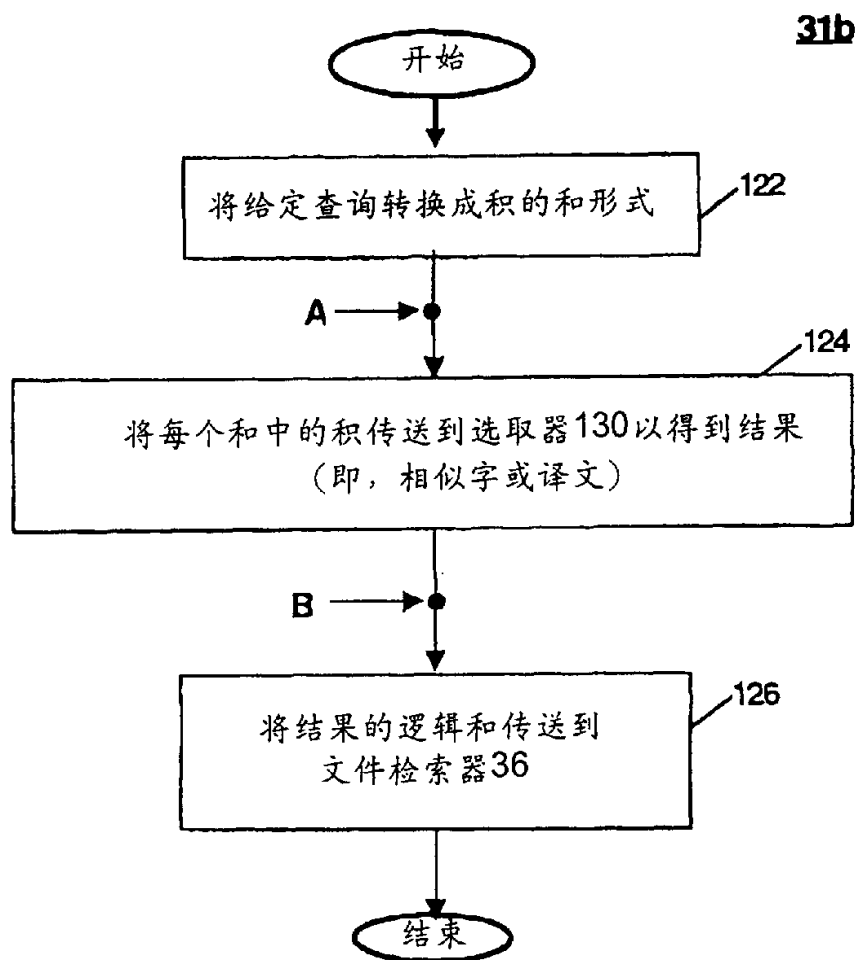
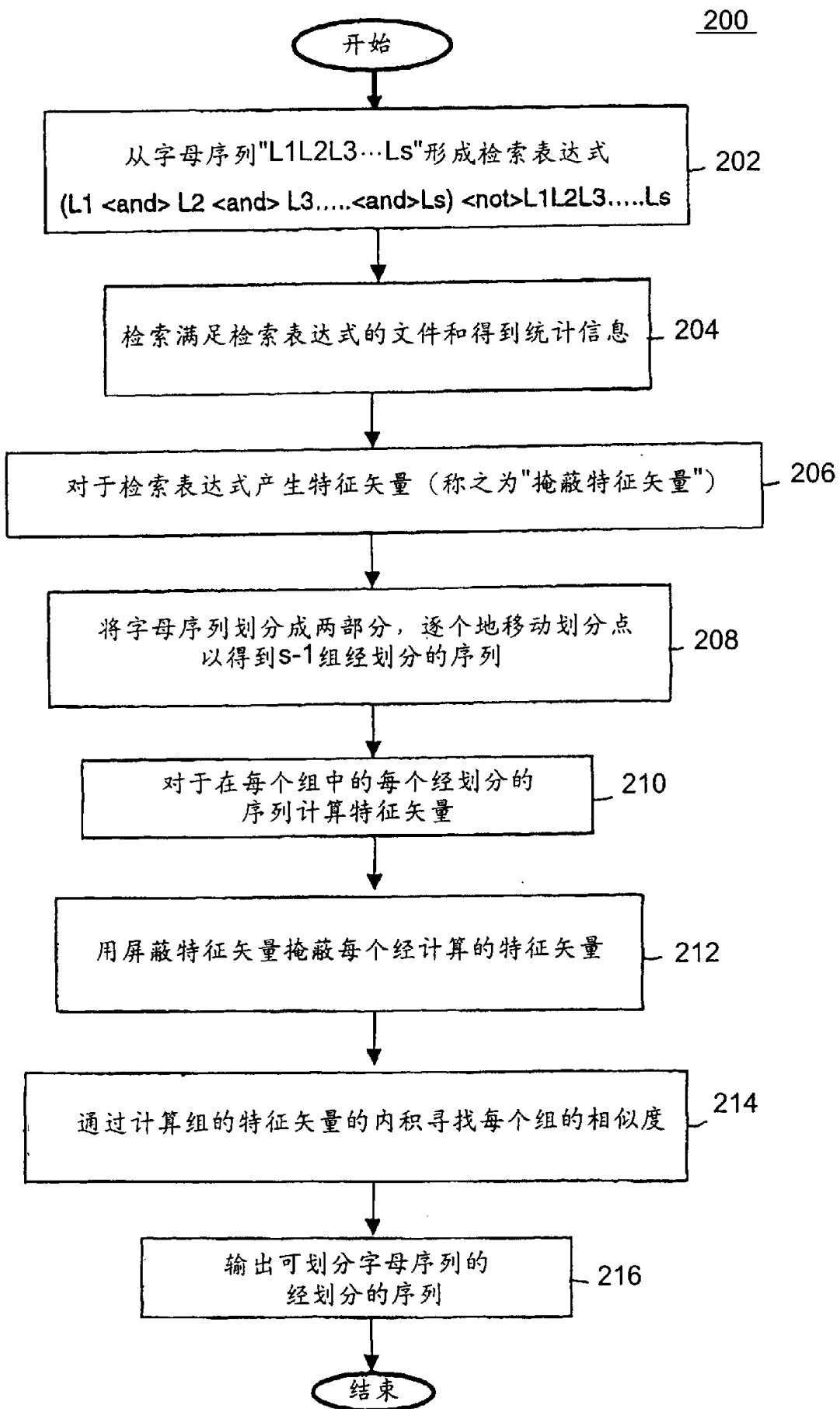


图 7



图

8

七组经划分的序列

输入字母序列

七个划分

"自立移動ロボット"

↑↑↑↑↑↑↑↑
/ 2 3 4 5 6 7

划分点

10

冬

相同

掩蔽特征向量 (0, X, X, X, 0, X, X, X, 0, 0, 0, 0, 0, ..., 0, 0, 0, X) = (0, X, X, X, 0, X, X, X, 0, 0, 0, 0, 0, ..., 0, 0, 0, X)

对于序列"自律移動ロボット"的特征向量

(0, X, 0, X, 0, X, X, X, 0, X, X, 0, ..., X, X, X, X)

对于序列"自"的特征向量

(0, X, 0, X, 0, X, X, X, 0, 0, 0, 0, ..., 0, 0, 0, X)

结果特征向量

0: 一个零值
X: 一个正值

11

冬

相似度 划分适宜性

(自、自律移動ロボット)
(自律、移動ロボット)
(自律移、動ロボット)
(自律移動、ロボット)
(自律移動ロ、ボット)
(自律移動ロボ、ット)
(自律移動ロボッ、ト)

0. 0 1 NO-
0. 4 0 YES
0. 0 2 NO
0. 5 0 YES
0. 0 2 NO
0. 0 1 NO
0. 0 2 NO

冬

12

阈值 = 0.35

输入字母序列

“自立移動ロボット”

检索表达式

(自<and>立<and>移<and>動<and>ロ<and>ボ<and>ツ<and>ト)<not>自立移動ロボット

图 9

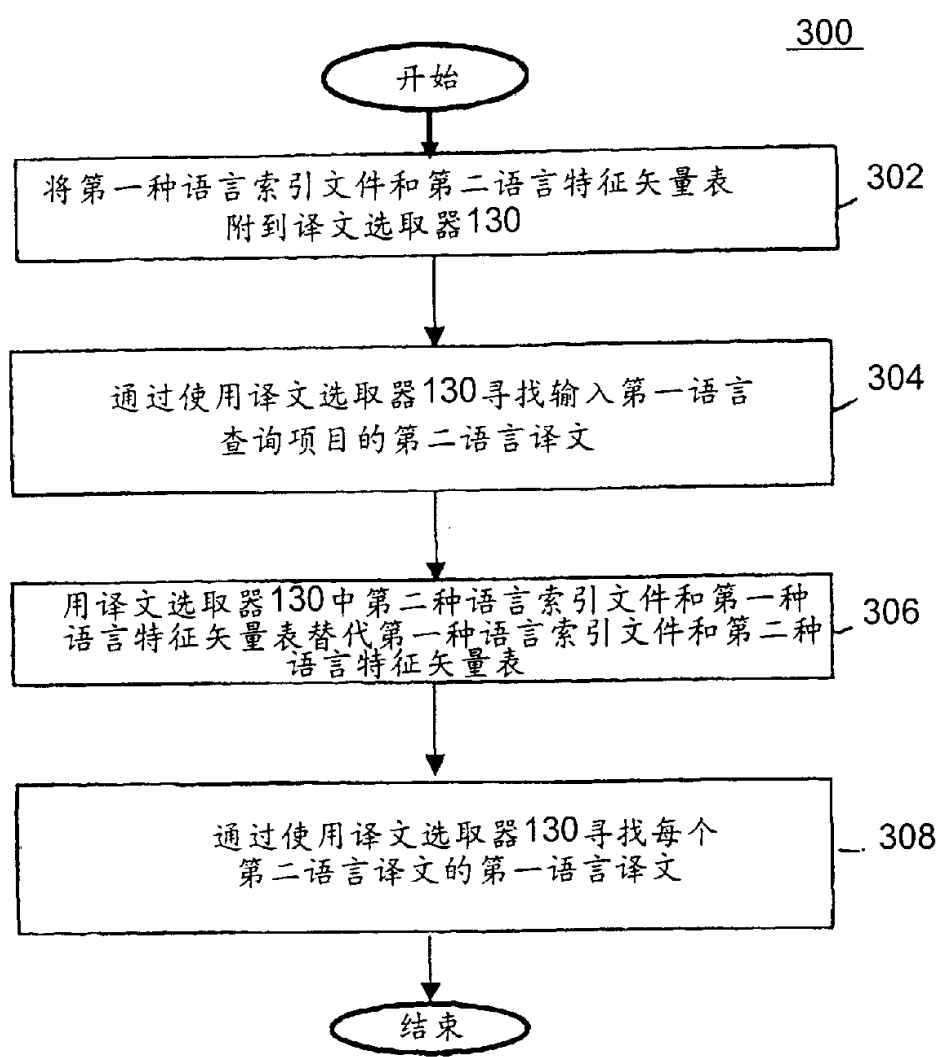


图 13