[19] 中华人民共和国国家知识产权局

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

[12] 发明专利说明书

专利号 ZL 03145727.4

[45] 授权公告日 2008 年 1 月 9 日

[11] 授权公告号 CN 100361125C

[22] 申请日 2003.6.30 [21] 申请号 03145727.4

[30] 优先权

[32] 2002. 6.28 [33] US [31] 10/186,174

[73] 专利权人 微软公司

地址 美国华盛顿州

[72] 发明人 周 明

[56] 参考文献

CN1131302A 1996.9.18

CN1302412A 2001.7.4

US6006221A 1999.12.21

审查员 高 可

[74] 专利代理机构 上海专利商标事务所有限公司

代理人 张政权

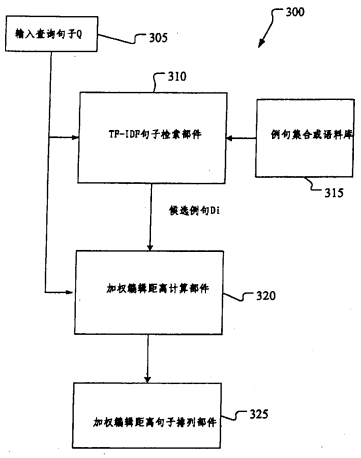
权利要求书 2 页 说明书 11 页 附图 5 页

[54] 发明名称

基于加权编辑距离的自动例句检索的系统和方法

[57] 摘要

提供一种方法和计算机可读媒体从句子集合中检索例句。 收到输入查询句子，用检索词频率倒置文件频率算法从句子集合中为输入查询句子选择候选例句。 基于所选候选例句与输入查询句子之间的加权编辑距离所选候选例句被重新排列。 也提供了实施该方法的系统。



1. 从句子集合中检索例句的方法，其特征在于：

接收输入查询句子；

用检索词频率倒置文件频率算法从句子集合中为输入查询句子选择候选例句；以及

基于所选候选例句和输入查询句子之间的编辑距离重新排列所选候选例句，其中基于所述编辑距离重新排列所选候选例句还包括：

为每个候选例句计算独立的加权编辑距离作为所述候选例句中检索词的函数，并作为对应于所述候选例句中的检索词的加权分数的函数，其中基于与候选例句中对应的检索词相关的语言的组成部分，所述加权分数具有不同的值；以及

基于为每个候选例句计算所得的独立的加权编辑距离重新排列所选候选例句。

2. 如权利要求 1 所述的方法，其特征在于，在重新排列所选候选例句的步骤中的编辑距离定义为将每个候选例句转变为输入查询句子所需的最小运算数。

3. 如权利要求 1 所述的方法，其特征在于，在重新排列所选候选例句的步骤中的编辑距离定义为将输入查询句子转变为每个候选例句所需的最小运算数。

4. 如权利要求 1 所述的方法，其特征在于，用检索词频率倒置文件频率算法从句子集合中为输入查询句子选择候选例句，还包括：

标记与句子集合中的句子中的对应的检索词相关的语言的组成部分；

从输入查询句子中去除停止词；以及

为句子集合中每个句子计算检索词频率倒置文件频率分数。

5. 如权利要求 4 所述的方法，其特征在于，用检索词频率倒置文件频率算法从句子集合中为输入查询句子选择候选例句，还包括选择句子集合中的那些检索词频率倒置文件频率分数高于门限的句子作为候选例句。

6. 从句子集合中检索例句的系统，其特征在于：

接收查询句子的输入；

检索词频率倒置文件频率句子检索部件，与输入耦合的，该部件用检索词频率倒置文件频率算法为输入查询句子从句子集合中选择候选例句；

加权编辑距离计算部件，与检索词频率倒置文件频率句子检索部件耦合的，该部件为每个侯选句子计算独立的加权编辑距离作为侯选例句中检索词的函数，以及作为与侯选例句中检索词对应的加权分数的函数，其中基于与侯选例句中检索词相关的语言的组成部分，所述加权分数具有不同的值；以及

排列部件，与加权编辑距离计算部件耦合的，该部件基于为每个侯选例句计算所得的独立的加权编辑距离重新排列所选侯选例句。

基于加权编辑距离的自动例句检索的系统和方法

技术领域

本发明涉及机器辅助书写系统和方法，尤其涉及自动例句检索的系统和方法以辅助书写或翻译处理。

背景技术

在各种应用中自动例句检索是必须的或有用的。例如，基于例子的机器翻译中，有必要检索语法上与被翻译的句子相似的句子。然后通过激活或选择一条检索的句子得到翻译。

在机器辅助翻译系统中，例如一个翻译存储器系统，需要一个检索方法以获得关联句子。然而许多检索算法存在不同类型的缺点，其中一些是无效的。例如，被检索出的句子常常和输入的句子几乎没有关系。许多检索算法的其他问题事实上包括其中的一些是无效的、一些需要可观的存储器和处理资源、一些需要句子语料库的预自动标注，预自动标注是个非常费时的负荷。

自动例句检索也可用作书写辅助，例如作为文字处理器的一种帮助功能。不管用户用他或她的母语书写或用非母语书写这可以是真的。例如，随着全球经济的持续增长和因特网的迅速发展，全世界的人们越来越熟悉用非他们的母语书写。不幸的是，对于一些拥有完全不同文化和书写风格的社会，用非母语书写的能力是永远存在的障碍。当用非母语书写（例如英语）时，非当地的说话者经常会犯语言使用错误（例如说汉语、日语、韩语或其他非英语的语言）。例句检索向书写者提供具有相似内容、相似语法结构的例句，或都为帮助修饰由书写者写出的例句。

因此，一种提供有效的例句检索的改进的方法或算法是非常重要的。

发明内容

提供一种从一个句子集合中检索例句的方法、计算机可读媒体和系统。收到一个查询句子，用检索词频率倒置文件频率（TF-IDF）算法从句子集合中为输入查询句子选择候选例句。然后基于所选候选例句和输入查询句子之间的加权编辑距离所选候选例句被重新排列。

在一些实施方式中，按将每个候选例句转变为输入查询句子所需的最小运算数的函数所选候选例句被重新排列。在另外实施方式中，按将输入查询句子转变为每个候选例句所需的最小运算数的函数所选候选例句被重新排列。

在各种实施方式中，基于所选候选例句和输入查询句子之间的加权编辑距离所选候选例句被重新排列。在一些实施方式中，基于加权编辑距离重新排列所选候选例句还包括为每个候选例句计算独立的加权编辑距离作为候选例句中检索词的函数，以及作为与候选例句中检索词对应的加权分数的函数。基于与候选例句中对应的检索词相关的语言组成部分加权分数具有不同的值。然后基于为每个候选例句计算所得的独立的加权编辑距离所选候选例句被重新排列。

附图说明

图 1 是本发明可实践的一个计算环境的框图，。

图 2 是本发明可实践的另一个可选用的计算环境的框图。

图 3 说明一个系统的框图，该系统可在图 1 和 2 中所示的计算环境中实现，该系统依据本发明的实施方式检索例句和基于编辑距离排列例句。

图 4 说明依据本发明的检索例句和基于编辑距离排列例句的方法的框图。

图 5 说明依据本发明的进一步实施方式的检索例句和基于编辑距离排列例句的方法的框图。

具体实施方式

图 1 说明了一个本发明可实现的合适的计算系统环境 100 的例子。计算系统环境 100 只是合适的计算环境的例子之一，也不试图建议任何应用范围或发明的功能的限制。计算系统环境 100 既不被解释为有任何独立性，也不被解释为与在用作范例的操作环境 100 中说明的任何一个部件或部件的组合有关的要求。

本发明可以和大量其他通用或专用计算系统环境或配置一起运算。适合与本发明一起使用的众所周知的计算系统、环境、和/或配置的例子包括，但不限于，个人计算机、服务器计算机、手持或膝上型设备、多处理器系统、基于微处理机的系统、机顶盒、可编程消费电子产品、网络 PC、小型计算机、主计算机、电话系统、包括以上系统或设备中的任何的分布式计算环境，等等之

类的。

本发明可用计算机可执行指令的通用语镜进行描述，如程序模块，由计算机执行。一般来说，程序模块包括例程、程序、对象、分量、数据结构等等，它们完成特定任务或实现特定抽象数据类型。本发明也可能在分布式计算环境中运行，在那里任务通过通信网络链接的远程处理设备来完成。在分布式计算机环境中，程序模块可位于包括存储器存储设备的本地和远程计算机存储媒体中。

关于图 1，为实现本发明用作范例的系统包括以计算机 110 为形式的通用计算设备。计算机 110 的部件可包括，但不限于，处理单元 120、系统存储器 130、以及系统总线 121，系统总线将包括系统存储器在内的各种系统部件耦合到处理单元 120 上。系统总线 121 可为总线结构的任何几种类型，包括存储器总线或存储器控制器、外围总线、以及采用任何各种总线体系结构的局部总线。作为例子，而非限制，这样的体系结构包括工业标准体系结构（ISA）总线、微通道体系结构（MCA）总线、增强型工业标准体系结构（EISA）总线、视频电子标准协会(VESA)局部总线、以及称为 Mezzanine 总线的外围部件互连(PCI)总线。

计算机 110 一般包括各种计算机可读媒体。计算机可读媒体可以是能被计算机 110 访问的任何可获得的媒体，它包括易失性和非易失性媒体、可移动和不可移动的媒体。作为例子，而非限制，计算机可读媒体可包括计算机存储媒体和通信媒体。计算机存储媒体包括易失性和非易失性媒体、可移动和不可移动的媒体，这些媒体用任何方法和技术来实现作为信息的存储，这些信息如计算机可读指令、数据结构、程序模块或其他数据的信息。计算机存储器媒体包括，而非限制，RAM、ROM、EEPROM、闪烁存储器或其它存储器技术、CD-ROM、数字多用光盘或其它光盘存储、磁带盒、磁带、磁盘存储或其它磁性存储设备、或任何其它可用于存储所需信息并能被计算机 110 访问的媒体。通信媒体一般具备计算机可读指令、数据结构、程序模块、或其他包含在诸如载波或其它传输机械的调制数据信号中的数据，通信媒体还包括任何信息传递媒体。术语“调制数据信号”是指具有一个或多个特征集合的信号或转变成这样一个将信息编码在信号中的方式。作为例子，而非限制，通信媒体包括诸如有线网络或直接有线连接的有线媒体，诸如声、RF、红外和其它无线媒体的无线媒体。以上任何的结合也可包含在计算机可读媒体的范围内。

计算机存储器130包括以易失性和/或非易失性存储器形式存在的计算机存储媒体，如只读存储器131和随机存取存储器132。基本输入/输出系统133（BIOS）一般存储于RO M131，它包括帮助传送计算机110中各单元之间的信息的基本例程，如在启动期间。RAM132一般包括数据和/或程序模块，它们被处理单元120立即访问和/或现行操作的。作为例子，而非限制，图1说明操作系统134、应用程序135、其它程序模块136和程序数据137。

计算机110可包括其它可移动/不可移动的、易失性/非易失性存储媒体。仅仅作为例子，图1说明了从不可移动的非易失性的磁性媒体读取或写入的硬盘驱动器141、从可移动的非易失性的磁盘152读取或写入的磁盘驱动器151、从诸如CD ROM或其它光媒体的可移动的非易失性的光盘156读取或写入的光盘驱动器155。其它可在用作范例的操作环境中使用的可移动/不可移动的、易失性/非易失性的计算机存储媒体包括，而非限制，磁带（盒）、闪烁存储器卡、数字多用磁盘、数字录像磁带、固态RAM、固态ROM，等等之类的。硬盘驱动器141一般通过如接口140的不可移动存储器接口与系统总线121相连。磁盘驱动器151和光盘驱动器155一般通过如接口150的可移动存储器接口与系统总线121相连。

以上讨论和图1中说明的设备及其相关计算机存储媒体，为计算机可读指令、数据结构、程序模块和计算机110的其它数据提供了存储。在图1中，例如，说明硬盘驱动器141存储操作系统144、应用程序145、其它程序模块141、以及程序数据147。注意这些部件可以与操作系统134、应用程序135、其它程序模块136和程序数据137相同或不同。操作系统144、应用程序145、其它程序模块141和程序数据147在此给出不同的数字说明，至少，它们是不同的复制。

用户可通过输入设备将命令和信息键入计算机110，输入设备如键盘162、话筒163、诸如鼠标、跟踪球或触摸板的指点器161。其他输入设备的（未展示）可包括操纵杆、游戏板、卫星抛物面、扫描仪、等等之类的。这些和其它输入设备通常通过与系统总线耦合用户输入接口160与处理单元120相连，也可以由其它接口和总线结构与处理单元相连，其它接口和总线结构如平行端口、游戏端口或通用串行总线（USB）。监视器191或其它形式的显示设备也通过如视频接口190的接口与系统总线相连。除了监视器，计算机还可包括其它外围输出设备，如扬声器197和打印机196，这些设备可通过输出外围设备

接口 195 相连。

计算机 110 可在联网的环境中运作，该环境采用逻辑连接至一个或多个如远程计算机 180 的远程计算机。远程计算机 180 可以是个人计算机、手持式设备、服务器、路由器、网络 PC、同级设备或其它普通网络节点，以及一般包括以上描述的与计算机 110 有关的大多或所有部件。图 1 中描绘的逻辑连接包括局域网（LAN）171 和广域网（WAN）173，但也可包括其它网络。如此网络环境在办公室、企业级计算机网络、内联网和因特网中是很平常的。

当用于 LAN 网络环境中，计算机 110 通过网络接口或适配器 170 与 LAN 相连。当用于 WAN 网络环境中，计算机 110 一般包括调制解调器 172 或其它在如因特网的 WAN 173 上建立通信的设备。调制解调器 172，可以是内置的或外置的，可通过用户输入接口 160 或其它合适的机械与系统总线 120 相连。在联网环境中，所描述的与计算机 110 或其中部分有关的程序模块可存储于远程存储器设备。作为例子，而非限制，图 1 说明了远程应用程序 185 常驻在远程计算机 180 上。能理解的是所示的网络连接是用作范例的，并且建立计算机之间的通信连接的其他方法是可用的。

图 2 是移动设备 200 的框图，它是一个用作范例的计算环境。移动设备 200 包括微处理机 202、存储器 204、输入/输出（I/O）部件 206、以及用作与远程计算机或其它移动设备通信的通信接口 208。在实施方式中，前面所述的部件通过合适的总线 210 耦合起来相互通信。

存储器 204 以非易失性电子存储器来实现，如带有电池备份模块（未展示）的随机存取存储器，如此使得当移动设备 200 的主电源关闭时在存储器 204 中存储的信息不会遗失。存储器 204 的一部分较佳地被分配为程序执行用的可编址的存储器，而存储器 204 的另一部分较佳地被用作存储，如模拟磁盘驱动器上的存储。

存储器 204 包括操作系统 212、应用程序 214 以及目标存储 216。在操作中，操作系统 212 较佳地被处理器 202 从存储器 204 中执行。在较佳的实施方式中，操作系统 212 是从微软公司购买的 WINDOWS® CE 注册操作系统。操作系统 212 较佳地为移动设备设计，并且通过一组所述的应用程序接口和方法实现被应用 214 利用的数据库特性。在目标存储 216 中目标由应用 214 和操作系统 212 维护，至少部分可响应传至所述的应用程序接口和方法的呼叫。

通信接口 208 代表使得移动设备 200 能发送和接收信息的大量设备和技

术。设备包括有线和无线的调制解调器、卫星接受机和以此为例的广播调谐。移动设备 200 在此可以和计算机直接相连交换数据。在这种情况下，通信接口可以为红外的收发器、或串行或并行的通信连接，所有这些可以发射信息流。

输入输出部件 206 包括各种输入设备，如触敏屏幕、按钮、滚筒和话筒，以及各种输出设备，包括音频发生器、振动设备和显示器。以上所列的设备是作为例子，并且不需要所有都出现在移动设备 200 上。另外，在本发明的范围内其它输入/输出设备可随移动设备 200 附带或发现。

依据本发明各个方面，建议自动检索例句的系统和方法以辅助书写和翻译处理。本发明的系统和方法能在图 1 和图 2 所示的计算环境中实现，也可以在其它计算环境中实现。依据本发明的例句检索算法包括两个步骤：用加权检索词频率倒置文件频率（TF-IDF）方法选择候选句子，然后通过加权编辑距离排列候选句子。图 3 是说明实现这个方法的系统 300 的框图。图 4 是说明通用方法的框图。

如图 3 所示，查询句子 Q，在 305 处所示的，是系统的输入。基于查询句子 305，句子检索部件 310 用常规 TF-IDF 算法或方法从 315 处所示的例句集合 D 中选择候选例句 D_i。输入查询句子的相关步骤 405，以及从集合 D 中选择候选例句 D_i 的相关步骤 406 在图 4 中展示。尽管 TF-IDF 方法在传统的信息检索（IR）系统中广泛应用，用作范例的实施方式中检索部件使用的 TF-IDF 算法的讨论在下文中给出。

在句子检索部件 310 从集合 315 中选出候选例句后，加权编辑距离计算部件 320 为每个候选例句生成加权编辑距离。如以下更详细描述，输入查询句子和候选例句之一之间的编辑距离被定义为将候选例句转变为查询句子所需的最小运算数。依据发明，在编辑距离的计算中不同的语言组成部分（POS）被分配不同的加权或分数。排列部件 325 按编辑距离的次序重新排列候选例句。具有最低编辑距离值的例句被排列到最高。根据加权距离重新排列所选的或候选的例句的相关步骤在图 4 的 415 处表示。该步骤可包括生成或计算加权编辑距离的子步骤。

1．用 TF-IDF 方法选择候选句子

如上关于图 3 和 4 的描述，用在 IR 系统中普遍的 TF-IDF 方法从句子集合中选择候选句子。下面的讨论提供 TF-IDF 方法的一个例子，该方法可以由图 3 中所示的部件 310 按图 4 所示的步骤 410 使用。其它 TF-IDF 方法也可使

用。

表示为 D 的例句的整个集合 315 由一些“文件”组成，每个文件实际上就是一个例句。采用常规的 IR 索引方法的一个文件（只包含仅仅一个句子）索引结果可以为表示如方程 1 所示的加权的向量。

方程 1

$$D_i \rightarrow (d_{i1}, d_{i2}, \ldots, d_{im})$$

其中 $d_{ik}(1 \le k \le m)$ 是文件 D_i 的中检索词 t_k 加权，m 是向量空间的大小，由集合中发现的不同检索词的数目决定。在例子实施方式中，检索词为英语词语。一个文件中一个检索词的加权 d_{ik} 根据该检索词在文件中出现的频率（tf——检索词频率）以及其在整个集合中的分布（idf——倒置文件频率）来计算。有多种计算和定义检索词加权的方法。在此，作为例子，我们采用方程 2 所示的关系

方程 2

$$d_{ik} = \frac{[\log(f_{ik}) + 1.0] * \log(N / n_k)}{\sqrt{\sum_j [(\log(f_{jk}) + 1.0) * \log(N / n_k)]^2}}$$

其中 f_{ik} 是在文件 D_i 中检索词 t_k 的出现频率，N 是集合中文件的总数，n_k 是包含检索词 t_k 的文件数。这是在 IR 中最普遍的 TF-IDF 加权方案。

在 TF-IDF 加权方案中也是普遍的，查询 Q，即用户输入句子，也由类似的方法索引，为查询获得一个向量，如方程 3 所示。

方程 3

$$Q_j \rightarrow (q_{j1}, q_{j2}, \ldots, q_{jm})$$

其中查询 Q_j 的向量加权 $q_{jm}(1 \le k \le m)$ 由方程 2 的关系类型决定。

文件集合中文件 D_i 和查询句子 Q_j 之间的相似性 $Sim(D_i, Q_j)$ 由他们的向量内乘计算而得，如方程 4 所示。

方程 4

$$Sim(D_i, Q_j) = \sum_k (d_{ik} * q_{jk})$$

输出是一组句子 S，S 如方程 5 所示定义为：

方程 5

$$S=\{D_i \mid Sim(D_i, Q_j) \geq \delta\}$$

2．根据加权编辑距离重新排列句子集合 S

如上关于图 3 和 4 的描述，从集合中所选候选例句集合 S 被从最短编辑距离到最长编辑距离重新排列，编辑距离与输入查询句子 Q 有关。下面的讨论提供了一个编辑距离的计算算法的例子，该算法可以由在图 3 所示的部件 320 中按图 4 所示的步骤 415 使用。其他编辑距离计算方法也可使用。

如所描述的，加权编辑距离方法被用于重新排列所选句子集合 S。在句子集合 S 中给定一所选句子 $D_i \rightarrow (d_{i1}, d_{i2}, ..., d_{im})$，在 D_i 和 Q_j 之间的编辑距离，表示为 $ED(D_i, Q_j)$，被定义为使列 A 和 B 两串相等所需的检索词插入、删减和替代的最小数。编辑距离，有时也指 Levenshtein 距离(LD)，是两串，源串和目标串之间相似性的测量。距离代表将源串变换为目标串所需的删减、插入和替代的数。

$ED(D_i, Q_j)$ 特别定义为将 D_i 转变为 Q_j 的最小运算数，在此运算为其中之一：

1. 改变一个检索词
2. 插入一个检索词，或
3. 删除一个检索词

然而，依照本发明可使用的编辑距离的另一个定义是将 Q_j 转变为 D_i 的最小运算数。

一个动态的程序算法被用来计算两串的编辑距离。用动态程序算法，一个两维的矩阵，m[i,j]，被用来保持编辑距离数值，i 从 0 至|S1|（其中|S1|是第一个侯选句子的检索词的数）j 从 0 至|S2|（其中|S2|是查询句子的检索词的数）。该两维矩阵也可表示[0...|S1|,0...|S2|]。动态程序算法使用如下列伪码中所描述的方法定义所包含的编辑距离值 m[i,j]。

```
m[i,j]=ED(S1[1...i],S2[1...j])
m[0,0]=0
m[i,0]=i,  i =1...|S1|
m[0,j]=j,  j =1...|S2|
m[i,j]=min(m[i-1,j-1]
          +if S1[i]=S2[j] then 0 else 1,
```

m[i-1, j]+1,

m[i, j-1], +1),

i=1...|S1|, j=1...|S2|

编辑距离值 m[,]可以逐行计算。行 m[i,]只依赖于行 m[i-1,]。该算法的时间复杂度是 O(|S1|*|S2|)。如果 S1 和 S2 根据检索词的数有相似的长度，例如 n，复杂度为 $Q(n^2)$。依据本发明使用的加权编辑距离是指每个运算（插入、删除或替代）的补偿不总是等于 1，如同在常规编辑距离计算技术的情况下，而是补偿可以基于检索词的显著性设置成不同的分数。例如，上面的算法可以根据如表格 1 所示的语言的组成部分调整使用分数表。

表 1

语言	分数
名词	0.6
动词	1.0
形容词	0.8
副词	0.8
介词	0.8
其它	0.4

因此，在下列问题中算法可以被修订以考虑检索词的语言组成部分。

m[i, j]=ED(S1[1...i], S2[1...j])

m[0, 0]=0

m[i, 0]=i, i =1...|S1|

m[0, j]=j, j =1...|S2|

m[i, j]=min(m[i-1, j-1]

+if S1[i]=S2[j] then 0 else [score],

m[i-1, j]+ [score], m[i, j-1]+ [score]),

i=1...|S1|, j=1...|S2|

例如，在算法的某些状态，对一个名词来说如果需要做任何运算（插入、删除），那么分数为 0.6。

S1 和 S2 的编辑距离计算是个递归的过程。为计算

ED(S1[1...i],s2[1...j]), 我们需要从下列三种情形中的最小。

1) S1 和 S2 都去掉尾词(或其它编辑单元)——在矩阵中表示为 m[i-1,j-1]+分数;

2) 只有 S1 去掉尾词，S2 保持——表示为 m[i-1,j]+分数;

3) 只有 S2 去掉尾词，S1 保持——表示为 m[i,j-1]+分数;

对于情形 1，分数可以如此计算：

如果 S1 和 S2 的尾词是相同的，那么分数=0;

否则，分数为 1;（成本是一个运算）//在加权的 ED 中，分数是可变的。看上面提到的表格，例如名词是 0.6。

如提到的，为了计算递归过程，被称为动态程序的方法可使用。

尽管展示了特殊的 POS 分数，在其它实施方式中语言不同组成部分的分数在不同应用中可以从表格 1 所示的那些值上被改变。因此，由 TF-IDF 方法所选的句子 $S=\{D_i|Sim(D_i,Q_j) \geq \delta\}$ 按加权编辑距离 ED 被排列，并且一个有序列表 T 可以获得

$$T=\{T_1,T_2,T_3,\ldots,T_n\},\text{其中，}ED(T_1,Q_j) \geq ED(T_{i+1},Q_j) \quad 1\leq i\leq n$$

其中 T_1 至 T_n 是候选例句（也是以前所指的 D_1 至 D_n），且 $ED(T_i,Q_j)$ 是句子 T_i 和输入查询句子 Q_j 之间的计算所得的编辑距离。

图 4 所示的通用系统和方法的另一个实施方式在图 5 框图中表示。如图 5 中 505 处所示的，输入句子 Q_j 作为查询提供给系统，用在本领域所知的类型的 POS 标记给查询句子 Q_j 的语言的组成部分标上标记，在 515 处停止词被从 Q_j 中去除。在信息检索领域中停止词被认为是不包含许多以信息检索为目的信息的词。这些词一般是高频率出现的词，如“is”、“he”、“you”、“a”、“the”、“an” 等等。去除它们可以提高程序的空间需求和效率。

如在 520 处所示，句子集合中每个句子的 TF-IDF 分数由以上所描述或相似的方法获得。具有超过门限δ的 TF-IDF 分数的句子被选为候选例句用作提炼或修饰输入查询句子 Q，或用作机器辅助翻译处理。这在方框 525 处展示。然后，所选的候选例句如前面讨论的那样被重新排列。在图 5 中，在 530 处说明计算每个所选句子和输入句子之间的编辑距离“ED”，并在 535 处说明根据“ED”分数排列候选句子。

尽管描述了本发明关于特殊的实施方式，本领域的专业人员会认识到在不脱离本发明的精神和范围下在形式和细节上可以做出改变。例如，在本应用中

作为例子的表示的特别的 TF-IDF 算法可以用本领域知晓的类型的算法改变或替换。同样，在基于加权编辑距离重新排列的候选句子中，作为例子提供的算法之外的算法可以被采用。

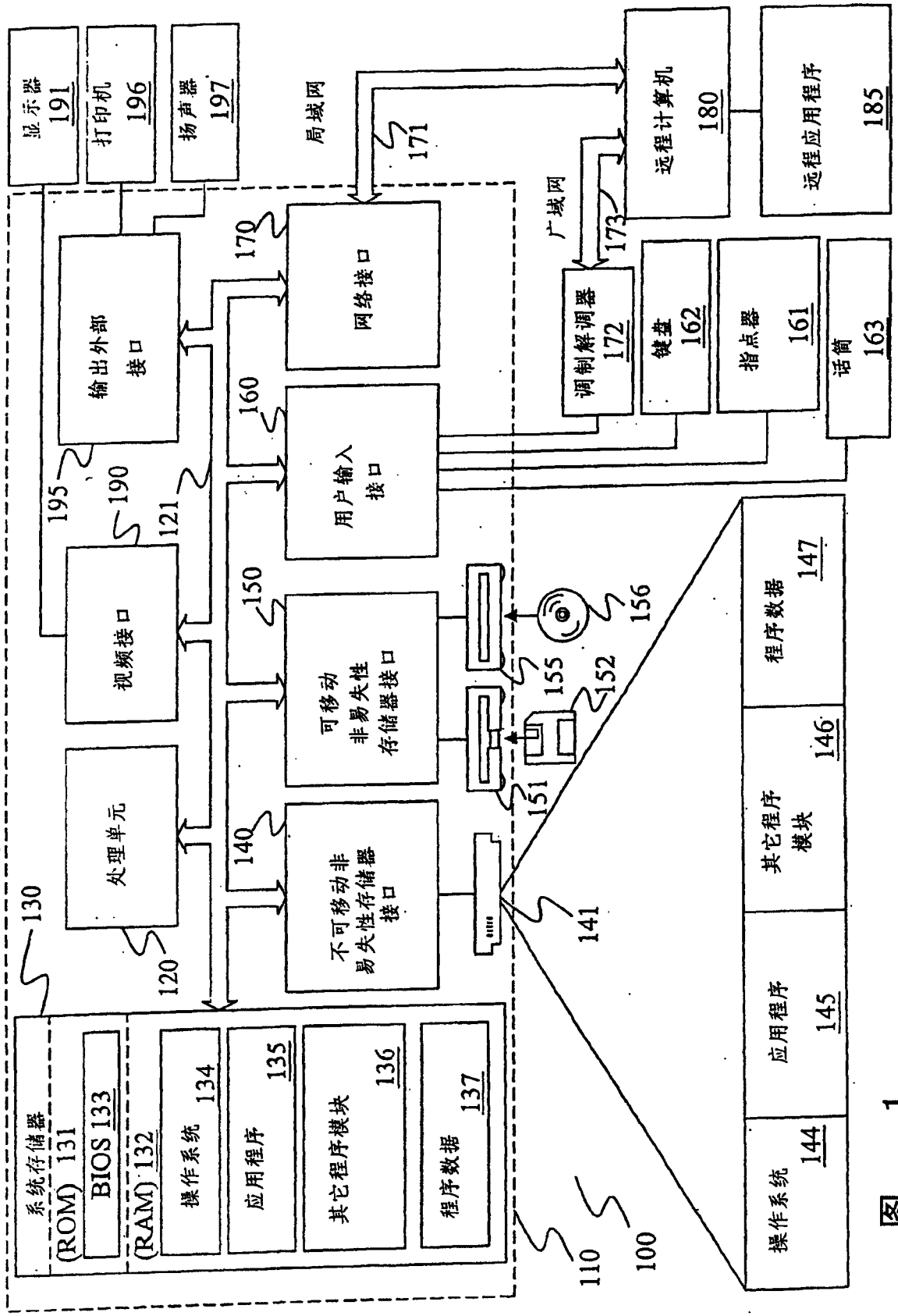

显示器 191
打印机 196
扬声器 197
局域网
171
远程计算机 180
远程应用程序 185
广域网
173
输出外部接口
网络接口
170
调制解调器 172
键盘 162
指点器 161
话筒 163
160
用户输入接口
195
190
121
视频接口
150
可移动非易失性存储器接口
156
155
152
151
处理单元
140
不可移动非易失性存储器接口
141
130
120
系统存储器
(ROM) 131
BIOS 133
(RAM) 132
操作系统 134
应用程序 135
其它程序模块 136
程序数据 137
程序数据 147
其它程序模块 146
应用程序 145
操作系统 144
110
100

图 1

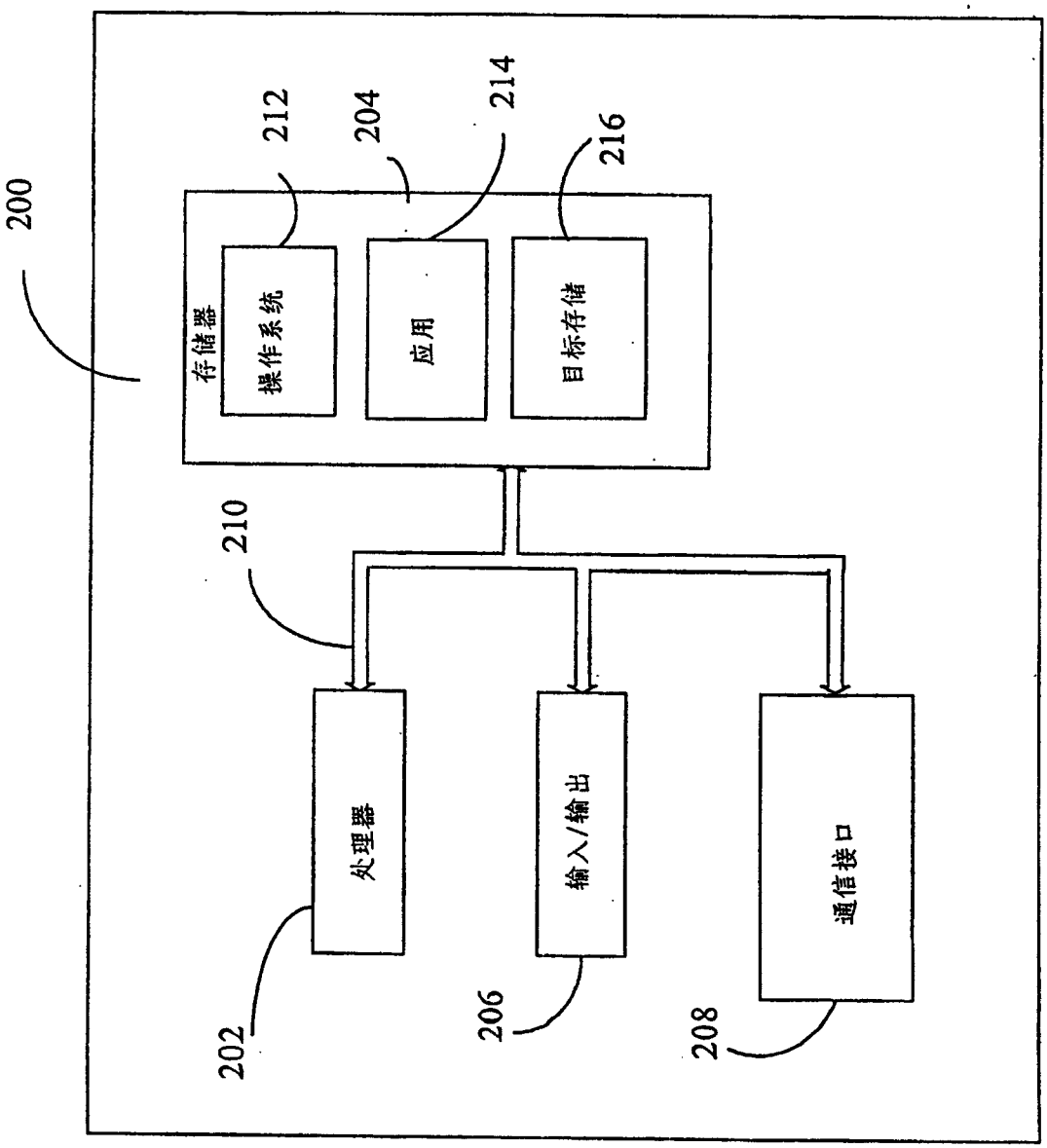

200
存储器
操作系统
212
204
应用
214
目标存储
216
210
处理器
202
输入/输出
206
通信接口
208

图 2

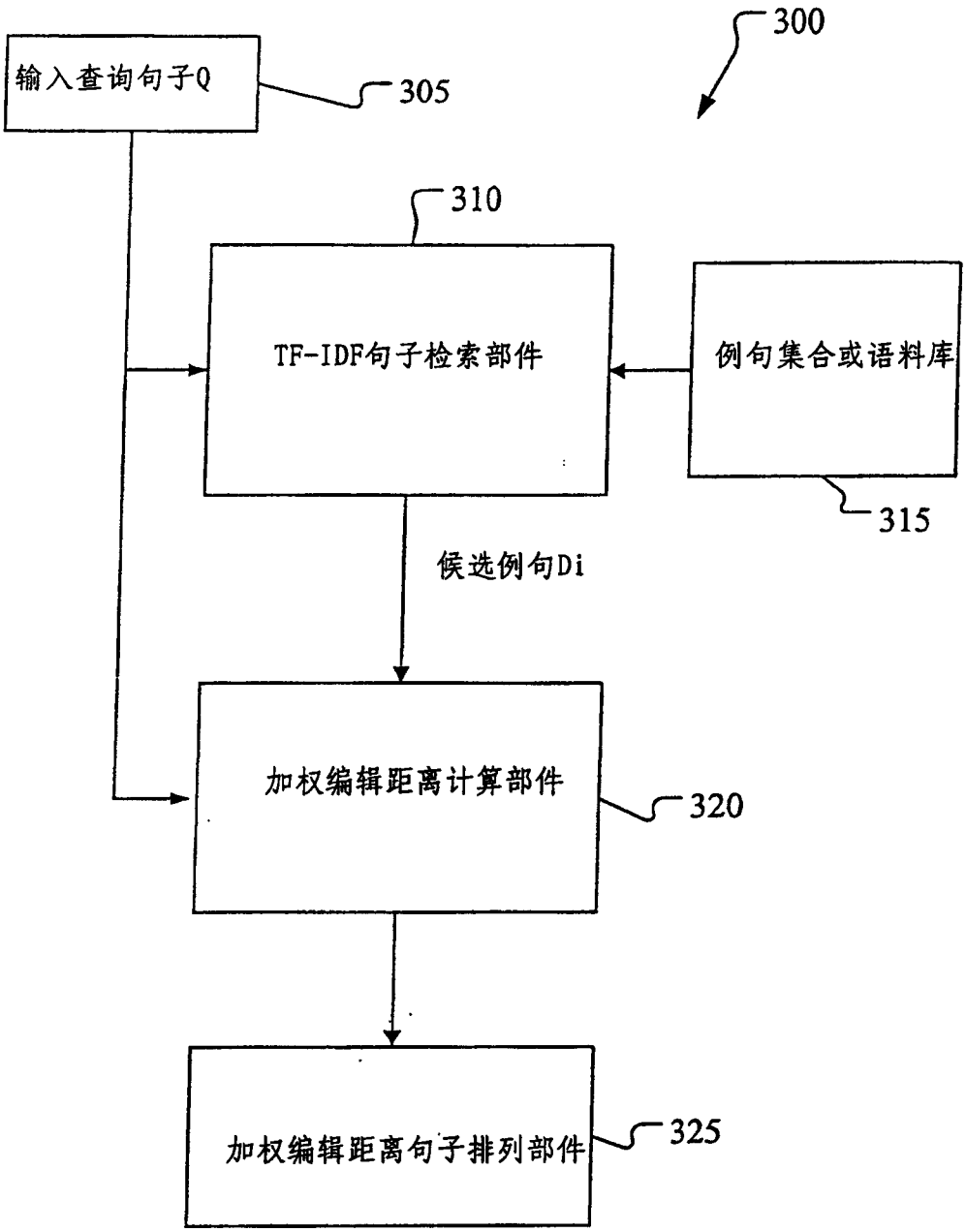

300
输入查询句子Q
305
310
TF-IDF句子检索部件
例句集合或语料库
315
候选例句Di
加权编辑距离计算部件
320
加权编辑距离句子排列部件
325

图 3

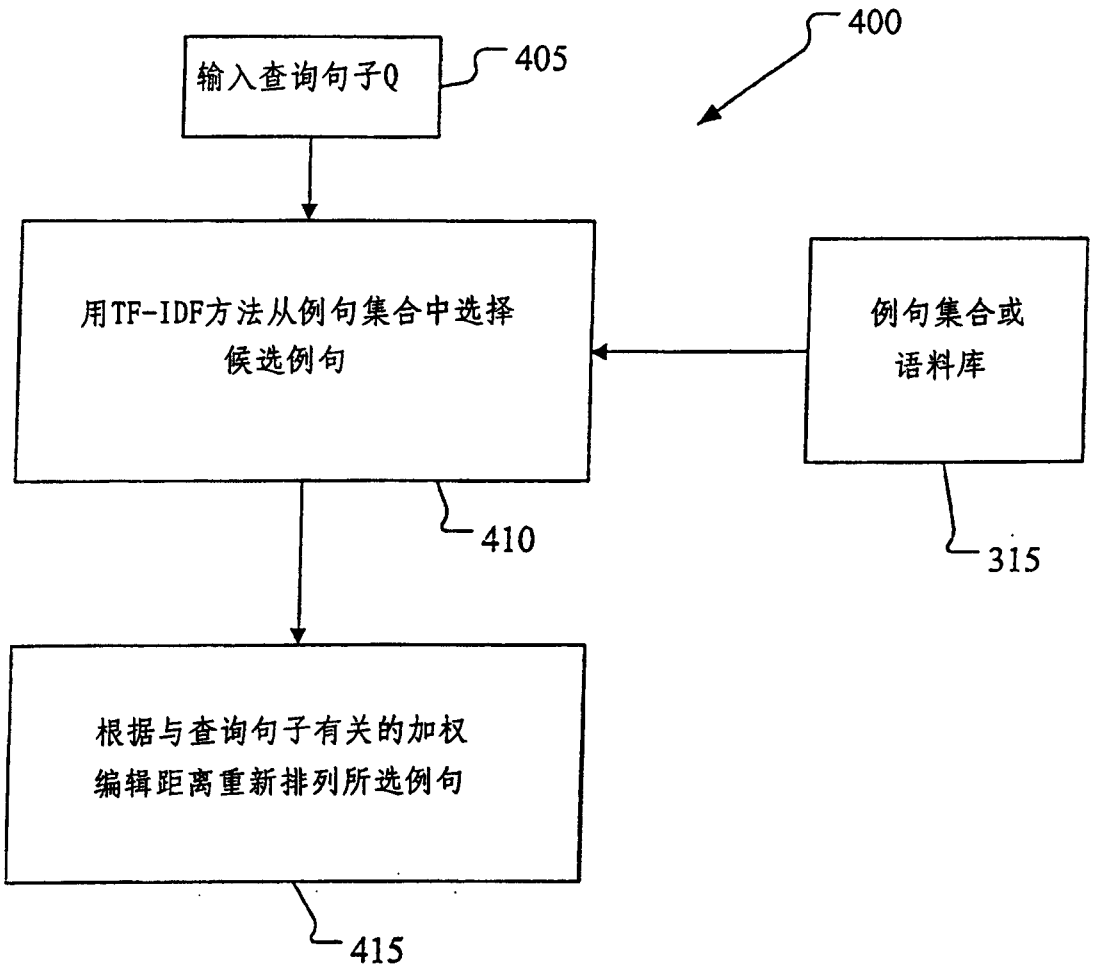

400
输入查询句子Q
405
用TF-IDF方法从例句集合中选择候选例句
例句集合或语料库
410
315
根据与查询句子有关的加权编辑距离重新排列所选例句
415

图 4

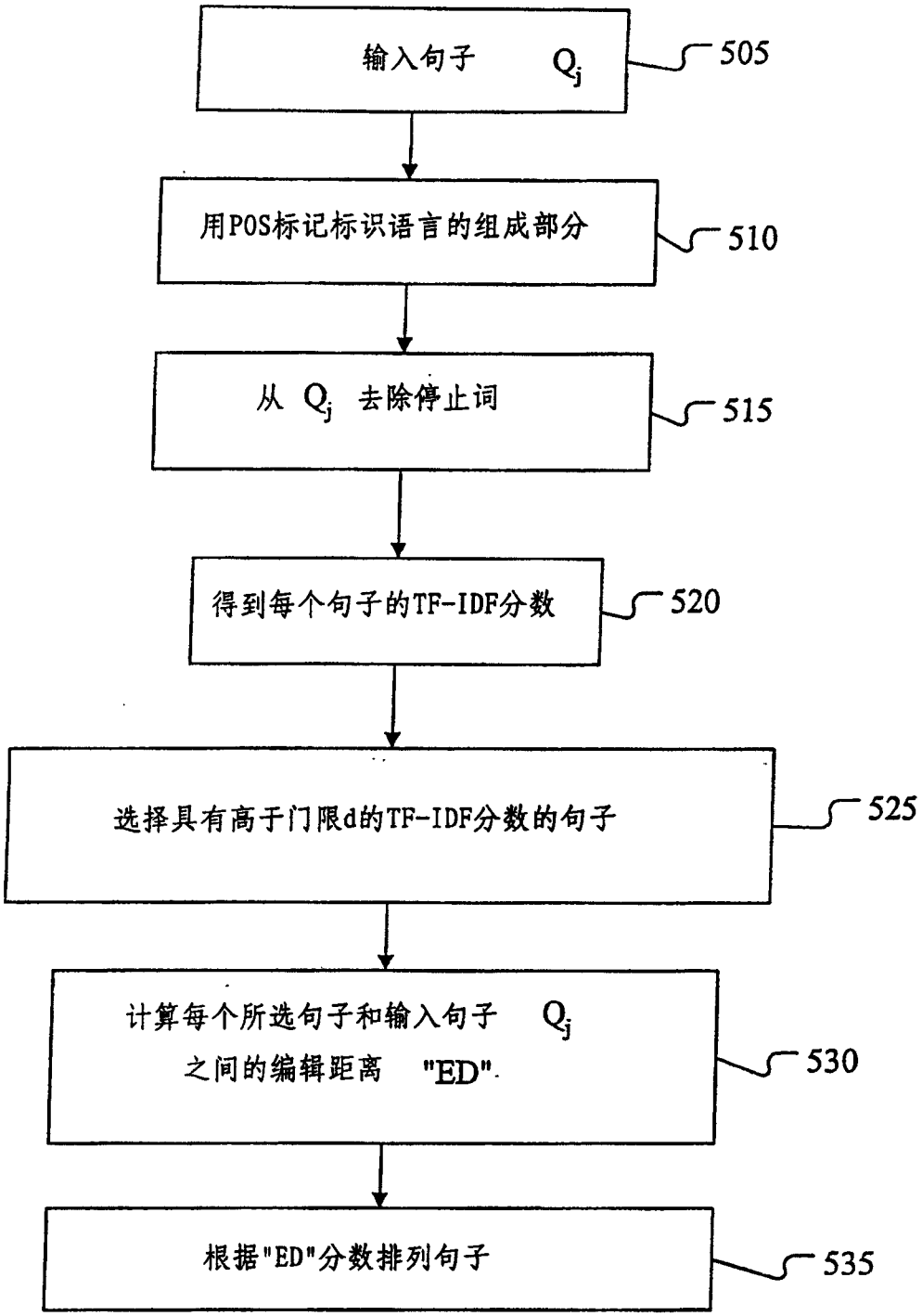

输入句子 Q_j
505
用POS标记标识语言的组成部分
510
从 Q_j 去除停止词
515
得到每个句子的TF-IDF分数
520
选择具有高于门限d的TF-IDF分数的句子
525
计算每个所选句子和输入句子 Q_j 之间的编辑距离 "ED".
530
根据"ED"分数排列句子
535

图 5