(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 102693272 B
(45)授权公告日 2017.04.12

(21)申请号 201210067044.7

(22)申请日 2012.03.14

(65)同一申请的已公布的文献号
申请公布号 CN 102693272 A

(43)申请公布日 2012.09.26

(30)优先权数据
13/048,678 2011.03.15 US

(73)专利权人 微软技术许可有限责任公司
地址 美国华盛顿州

(72)发明人 S·R·维西拉祖 U·R·尤杜帕 A·N·博伊 G·达萨 W·刘 Q·肖

(74)专利代理机构 上海专利商标事务所有限公司 31100
代理人 黄嵩泉

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

(56)对比文件
US 2004030780 A1,2004.02.12,
US 2007048715 A1,2007.03.01,
CN 101154228 A,2008.04.02,
CN 1728134 A,2006.02.01,

审查员 李梦诗

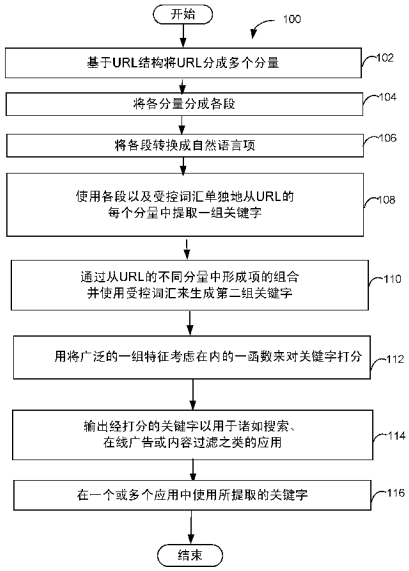
权利要求书2页 说明书8页 附图4页

(54)发明名称

从统一资源定位符(URL)的关键字提取

(57)摘要

本发明涉及从统一资源定位符(URL)中的关键字提取。本文所描述的关键字提取技术从web日志中的统一资源定位符(URL)中提取关键字。该技术充分利用URL的内容和结构来提取相关关键字。首先,URL基于其结构被分成多个分量。在受控词汇的帮助下,单独地从URL的每个分量中提取一组关键字。随后通过从URL的不同段中形成项的组合来生成第二组关键字。仅保留在可控词汇中存在的那些组合作为关键字。最后,用将广泛的一组特征考虑在内的一函数来对这些关键字打分。



CN 102693272 B

1.一种用于从对应于网站的统一资源定位符(URL)中提取关键字的计算机实现的过程,包括:

标识所述URL的分量(102);

基于URL分量的结构将所述URL分成多个段(104);

对所述段执行文本分段以将URL文本转换成自然语言项(106);

基于受控词汇从段项中提取第一组关键字(108);

通过从URL中的与用于生成所述第一组关键字的段不同的段中形成项的组合来生成第二组关键字(110);

基于所述受控词汇验证所述第二组关键字;

从用户从其请求当前页面的网页的引用者URL和相关联的当前URL中提取引用者关键字;

通过组合所述第一组关键字、所述第二组关键字、以及所述引用者关键字来形成最终的一组特征;

基于一组特征来对所述最终的一组关键字的相关性打分(112);以及

按相关性的顺序输出经打分的关键字(114)。

2.如权利要求1所述的计算机实现的过程,其特征在于,基于所述URL的结构将URL分成多个段还包括:

将所述URL分成授权机构、路径、查询和片段分量。

3.如权利要求1所述的计算机实现的过程,其特征在于,提取所述第一组关键字包括:

(a)对照所述受控词汇来比较四个项长度的段短语,

(b)如果在所述受控词汇中找到所述短语,则指派所述短语作为关键字,

(c)如果未在所述受控词汇中找到所述短语,则将段的长度减少1项并对照所述受控词汇再次比较所述短语,

(d)重复(c)直到在所述受控词汇中找到其余的项或者仅留下短语的一个项;以及

(e)如果在所述受控词汇中找到所述短语,则输出所述短语作为关键字,如果未在所述受控词汇中找到所述短语,则忽略所述短语。

4.如权利要求1所述的计算机实现的过程,其特征在于,还包括从所述第二组关键字中删除未在所述受控词汇中找到的项的组合。

5.如权利要求1所述的计算机实现的过程,其特征在于,在提取所述第一组关键字之前将URL文本转换成自然语言文本包括:

用空格来替换URL文本中的每一个定界符来创建项;以及

拆分通常在URL中找到的项。

6.如权利要求1所述的计算机实现的过程,其特征在于,通过从URL的不同分量中形成项的组合来生成第二组关键字还包括:

生成所述第一组关键字;

通过从段对中各取一个关键字并串接来自每个段对中的关键字来从所述URL的各部分中组合段对以生成候选关键字组合;

对照受控词汇来验证所述候选关键字组合;

保留在受控词汇中找到的候选关键字组合作为关键字,并且如果未找到则丢弃所述候

选关键字组合。

7.如权利要求1所述的计算机实现的过程，其特征在于，还包括通过使用外部知识源来扩展从所述URL中提取的关键字。

8.如权利要求1所述的计算机实现的过程，其特征在于，基于一组特征对所述第一组关键字和所述第二组关键字打分还包括：基于每个关键字的父段的位置、关键字的长度以及父段的长度来对每个关键字打分。

9.一种用于从统一资源定位符(URL)地址中提取关键字的计算机实现的过程，包括：

将当前网页的当前URL分成授权机构、路径、查询以及片段四个预定义的URL分量(202)；

基于特定定界符和试探性观察来分开地令牌化各分量以获得段(204)；

对所述段执行文本分段以将URL的文本转换成自然语言项(206)；

基于受控词汇从段项中提取第一组关键字(206)；

通过以下步骤来生成第二组关键字(208)：对所述URL中的每一对段，通过从所述对中的每一个段中选择项并连接所选择的项来生成候选关键字、基于所述受控词汇来验证所述候选关键字、以及仅保留在所述受控词汇中找得到的那些候选关键字作为第二组关键字；

基于所述受控词汇通过从URL中的与所述第一组关键字的段不同的段中形成项的组合；

基于相关性对所述第一组关键字和所述第二组关键字打分以输出经排序的一组经打分的关键字(210)。

10.如权利要求9所述的计算机实现的过程，其特征在于，基于在URL中的从中导出该关键字的段的位置、关键字的长度以及从中导出关键字的段的长度来确定关键字的相关性分数。

从统一资源定位符(URL)的关键字提取

技术领域

[0001] 本发明涉及URL,尤其涉及URL中的关键字提取。

背景技术

[0002] 在计算中,统一资源定位符(URL)是指定所标识的资源在哪里可用并提供一种用于检索该可用资源的机制的统一资源标识符(URI)。例如,URL可以是由主存网页的网站的创建者给予网页的唯一身份。URL以标准格式来定义,该标准格式通常指定方案或协议、域名或网际协议(IP)地址、要取得资源的路径或要运行的程序、查询串以及可任选的片段标识符。URL越来越多地包含与这些URL所对应的网页的话题高度相关的经压缩的文本。在许多应用中,它们可被视为关于网页的话题的有价值的信息源。

发明内容

[0003] 提供本发明内容以便以简化形式介绍将在以下具体实施方式中进一步描述的一些概念。本发明内容并不旨在标识所要求保护主题的关键特征或必要特征,也不旨在用于限制所要求保护主题的范围。

[0004] 本文描述的关键字提取技术从web日志(例如,通常以逆时间顺序包含用户所请求的一系列URL条目的服务器日志)中的URL中提取关键字。该技术充分利用URL的内容和结构来提取相关关键字。在一个实施例中,URL首先基于其结构被分成多个分量。在受控词汇的帮助下,单独地从URL的每个分量中提取一组关键字。通过从URL的不同段中形成项的组合来生成第二组关键字。仅保留在可控词汇中出现的那些组合作为关键字。最后,用将广泛的一组特征考虑在内的一函数来对这些关键字打分。

附图说明

[0005] 参考以下描述、所附权利要求书以及附图,将更好地理解本发明的具体特征、方面和优点,附图中:

[0006] 图1描绘了本文所描述的关键字提取技术的示例性过程的流程图。

[0007] 图2描绘了本文所描述的关键字提取技术的另一示例性过程的流程图。

[0008] 图3是用于实践本文所描述的关键字提取技术的一个示例性实施例的示例性体系结构。

[0009] 图4是可用于实践关键字提取技术的示例性计算环境的示意图。

具体实施方式

[0010] 在以下对关键字提取技术的描述中,对附图作出参考,附图形成了该描述的一部分,且作为可实践本文所描述的关键字提取技术的说明性示例示出。可以理解,可以利用其他实施例,并且可以作出结构上的改变而不背离所要求保护的主题的范围。

[0011] 1.0关键字提取技术

[0012] 以下章节提供了关键字提取技术的概览、以及用于实践本技术的示例性过程和示例性体系结构。还提供了关键字提取技术的各实施例的细节。

[0013] 1.1技术概览

[0014] 本文所描述的关键字提取技术从URL中提取关键字。该技术使用URL的内容和结构来提取相关关键字。这些关键字随后可在各应用中使用，诸如例如在线广告和在线内容过滤。

[0015] 1.2URL结构

[0016] 因为本发明的关键字提取技术在提取关键字时使用URL结构，所以对URL结构的一些解释是有用的。URL的格式基于Unix文件路径句法，其中使用正斜杠来隔开目录或文件夹以及文件或资源名。每一个URL都由以下各项中的某些项组成：scheme name(方案名，通常被称为协议)、之后是冒号、随后取决于该方案是domain name(域名，可另选地，网际协议(IP)地址)、port number(端口号)、要取得资源的path(路径)或要运行的程序、query string(查询串)以及可任选的fragment identifier(片段标识符)。句法是scheme://domain:port/path?query_string#fragment_id。本文所描述的关键字提取技术使用这一URL格式来提取网页的关键字，该关键字可用于各种应用。并不需要下载网页以提取对应于所提取的关键字的网页的关键字。这提供了极高的计算效率。

[0017] 1.3示例性过程

[0018] 图1描绘了用于从URL中提取关键字的示例性计算机实现的过程。如图1所示，框102，标识URL的各分量。更具体地，在关键字提取技术的一个实施例中，URL被分成授权机构(authority)、路径(path)、查询(query)和片段(fragment)分量。

[0019] 所标识的分量随后被拆分成各段，如框104中所示。例如，授权机构分量通过丢弃授权机构分量的协议字段和扩展字段而被拆分成各段；而路径分量通过丢弃与该URL所对应的网页的话题不相关的所有字段而被拆分成各段。查询分量通过提取查询字段中的键-值对而被拆分成各段；以及片段分量通过提取片段字段而被拆分成各段。在本文档的下文中将更详细地讨论关键字的分段。

[0020] 随后通过对各段执行文本分段来将URL文本转换成自然语言项来处理这些段，如框106中所示。例如，在一个实施例中，这通过用空格来替换URL文本中的每个定界符以创建项来完成；并且随后拆分通常在URL中找到的项。

[0021] 随后基于受控词汇从各段项中提取第一组关键字，如框108中所示。各段中匹配受控词汇的项被保留以属于第一组关键字。受控词汇是可从任何URL中提取的有效项和短语的大的列表。基于受控词汇通过从URL中的与用来生成第一组关键字的段不同的段中形成项的组合来生成第二组关键字，如框110中所示。在该技术的一个实施例中，通过以下方式来提取这第二组关键字：通过从URL的段对中各取出一关键字并串接来自每一个段对中的关键字来组合URL的段对以生成候选关键字组合，并且随后对照受控词汇来验证候选关键字组合。在受控词汇中找到的候选关键字组合被提取为关键字而那些未被找到的则被排除。从URL中提取的关键字还可任选地通过使用外部知识源来扩展。例如，通过使用语义映射，“travel(旅游)”可被扩展至“trip(旅行)”和“tour(观光)”。

[0022] 如框112中所示，随后基于一组特征对第一和第二组关键字的相关性打分，并且按相关性的顺序输出经打分的关键字(框114)。在关键字提取技术的一个实施例中，基于每一

关键字的父段的位置、关键字的长度以及父段的长度对关键字打分。

[0023] 输出关键字随后可在各种应用中使用，如框116中所示。例如，所提取的关键字可用于将网页上的关键字与广告客户提供的与广告有关的关键字进行匹配，以便将特定类型的广告定向到特定类型的网站。应该注意，不必下载网页以从给定的网页中提取关键字。可另选地，所提取的关键字可用于内容过滤，例如通过将从网页提取的关键字与令人讨厌的项或短语列表进行匹配以过滤诸如色情之类的内容。所提取的关键字还可用于通过将所提取的网页关键字与搜索查询项进行匹配的搜索应用。

[0024] 图2描绘了根据本发明的技术的用于从URL中提取关键字的另一示例性计算机实现的过程200。图2提供了这一示例性过程的一般过程动作。关于这些过程动作的更多细节将在本文档中的下文中提供。

[0025] 如图2所示，框202，网页的URL被分成授权机构、路径、查询和片段4个预定义URL分量。各分量基于特定定界符和试探性观察被分开地令牌化以获得各段，如框204中所示。如框206中所示，对各段执行文本分段以将URL的文本转换成自然语言项，并基于受控词汇从各段项中提取第一组关键字。如框208中所示，通过从URL中与用于提取第一组关键字的段不同的段中形成项的组合并且提取在受控词汇中的项的组合作为第二组关键字来生成第二组关键字。

[0026] 随后基于相关性对这些第一和第二组关键字打分以输出经排序的一组经打分的关键字，如框210中所示。各种打分技术可用于此目的。该技术还可通过使用外部知识源来通过将关键字映射到其他在语义上等价或相关的字和短语来提供关键字扩展从而生成附加的关键字。

[0027] 1.4示例性体系结构

[0028] 图3示出了用于采用关键字提取技术的示例性体系结构300。如图3所示，该示例性体系结构300包括驻留在通用计算设备400上的关键字提取模块302，这将参照图4更详细地予以讨论。URL 304是输入。分量划分模块306基于URL结构将URL 304分成多个分量308。这一组分量308在分段模块310中被分段，并且各段在语言处理模块312中被转换成自然语言语音项314。随后使用受控词汇(框320)在第一关键字提取模块(框316)中单独地从URL的每个分量中提取第一组关键字318。还在第二关键字提取模块(框322)中通过从URL中的与用于提取第一组关键字的段不同的段中形成项的组合324并且只保留在受控词汇(框320)中存在的关键字来提取第二组关键字(框326)。随后在打分模块(框328)中对第一和第二关键字316、326打分。在关键字提取技术的一个实施例中，基于在URL中的从中提取这些关键字的位置对关键字打分。经打分的关键字330随后被输出以用于一个或多个应用。

[0029] 在下一章节中将讨论这一体系结构的各方面的细节。

[0030] 1.5关键字提取技术的示例性实施例的细节

[0031] 已经讨论了示例性过程和示例性体系结构，以下章节提供关键字提取技术的各实施例的细节。

[0032] 1.5.1URL解析

[0033] URL解析是关键字提取中的第一步骤之一，其中保留URL中含信息量的部分并跳过含噪声的文本。这通过充分利用URL的结构来实现。如前文所讨论的，URL一般包含四个重要的分量：授权机构、路径、查询和片段。在以下段落中更详细地讨论一般的从URL中提取分

量。所提取的分量中的每一个被进一步解析成各段。

[0034] 1.5.1.1授权机构：

[0035] 授权机构是每个URL中的必要分量。它给出了其上主存表示该URL的页面的服务器的名称。授权机构可包含多个部分，诸如由点分开的协议、主机名、域。授权机构总是以诸如“http”、“https”之类的协议开始。同样，授权机构中的最后一个部分采用“com”、“net”、“us”、“org”等值之一，该值广泛地指示网站的种类并且通常在寻找相关关键字时并非是有用的。该技术丢弃URL的协议和最后一个部分，并且保留剩余部分作为来自这一分量的段。例如，http://realestate.msn.com具有段“realestate(房地产)”和“msn”。

[0036] 1.5.1.2路径：

[0037] URL可包含路径字段，该路径字段包含到要取得资源的路径。路径字段在URL中的授权机构之后，并且可包含由“/”分开的目录列表。这些目录可表示对应于该URL的页面所属的类别。有时候，目录可包含如“content(内容)”之类的不含信息量的文本或不与页面的话题相关的一系列数字。这些目录被忽略，而其余目录构成这一分量的段。例如，如果文本太概括(即，“content(内容)”、“file(文件)”)或不具有信息量(即，“123”、“a”)，则这些目录可被忽略。

[0038] 1.5.1.3查询：

[0039] 有时候，URL指向诸如搜索引擎和通用网关接口(CGI)脚本之类的web应用。查询字段是作为输入被发送到这些程序的查询串。查询字段在URL中的路径之后以“?”开始。查询字段包含具有定界符“;”、“&”等的键-值对。键-值对是一组两个链接的数据项：键，是某一数据项的唯一标识符；以及值，或是被标识的数据或是指向该数据的位置的指针。例如，city＝”las vegas”&show＝”cirque du soleil”意思是Cirque du Solei表演在Las Vegas城。查询串中的键-值对被保留为来自这一分量的段。取决于应用，某些键可变得重要而另外一些键可变为噪声。

[0040] 1.5.1.4片段：

[0041] 片段字段是出现在URL末端在井号”#”之后的HTML锚。片段字段被保留为来自这一分量的段。

[0042] 从四个逻辑分量中导出的所有段形成了关键字提取技术对其进行操作的基本单元。

[0043] 1.5.2受控词汇

[0044] 难以从URL中未结构化的文本中找到短语边界，因为不存在关于文本应当如何出现的规则。诸如名称实体识别程序(NER)、部分语音(POS)标签程序之类的用于短语标识的现有的自然语言处理(NLP)工具无法在此处应用，因为它们是在自然语言文本的自由流程上训练的。为克服这一挑战，关键字提取技术利用受控词汇来标识URL中的有效短语。

[0045] 一般地，受控词汇是可从任何URL中提取的有效短语的大的列表。受控词汇的本质和大小可取决于关键字所用于的应用而改变。例如，一般话题标识系统可使用从Wikipedia(维基百科)话题中导出的一般话题列表作为受控词汇。用于广告的关键字提取系统可使用成百万的广告投标短语的列表作为受控词汇。

[0046] 1.5.3文本分段

[0047] 在关键字提取之前，需要附加的过程来将经分段的URL文本转换成自然语言文本。

在一个实施例中，用空格替换诸如“-”或“_”之类的定界符，并且拆分在URL中通常找到的附加项。例如，“savinganddebt”将被拆分成“savings and debt(存款和债务)”。

[0048] 为优化经拆分的项的相关性，首先检查每一个经拆分的项以查看它是否存在于受控词汇中。如果不存在，则该技术试图搜索在受控词汇中存在的有效拆分。如下以迭代的方式来执行项拆分。

[0049] 1)再引入一个空格到项中(例如，这可以通过以迭代方式来反复试用直到在受控词汇中找到匹配来完成)。

[0050] 2)生成具有新的空格的所有可能的字的拆分。

[0051] 3)如果找到一个有效拆分，则返回有效拆分的项。

[0052] 4)如果找到一个以上的有效拆分，则对于每一个有效拆分，计算受控词汇中的个别词的频率总和并返回具有最大总和的有效拆分的项。

[0053] 1.5.4关键字提取

[0054] 在文本分段之后，通过对照受控词汇扫描每一段来从该段中提取关键字。如果来自段的短语出现在受控词汇中，则它被指派为关键字。在关键字提取技术的一个实施例中，最初用最大的可能短语(4个字的长度)从左扫描每一个段。如果找到匹配，则将短语添加到关键字列表。否则，短语长度减少1项至3个字长度，并且该技术重复先前的步骤。这一过程被重复迭代，直到该技术找到受控词汇中的短语，或者该技术留下该段中的第一个字。随后，该技术移动至该段中的下一个字，并重复相同的过程来寻找可能是关键字的短语。

[0055] 在一个实施例中，如果URL是搜索引擎结果页，则提取上述关键字以及附加关键字。从URL的查询分量中提取用户查询并将其作为单独的关键字输出，不管该查询是否存在于受控词汇中。

[0056] 1.5.4关键字组合

[0057] 由于URL中的有限量的文本，从URL中提取关键字并未得到许多关键字。所讨论的关键字提取过程关于提取第一组关键字的一个限制是：该技术仅从连续出现在URL同一段中的字中构建关键字。然而，通过从URL的不同段中组合项来生成相关关键字是可能的。为此，该技术实现以下动作。

[0058] 第一，使用在对第一组关键字的提取步骤中说明的方法来从URL中的每一段中提取一组关键字。对于各段的每一对，通过从两个不同的段中各取一个关键字并串接这些关键字来形成候选关键字组合。对照受控词汇来验证这些候选组合，并且保留在受控词汇中出现的那些候选组合作为关键字并丢弃其他的候选组合。在先前的提取步骤中从各段中提取的最初的一组关键字以及从这一组合步骤中生成的关键字形成了URL的最终的一组关键字。

[0059] 1.5.6智能扩展

[0060] 在一个实施例中，该技术使用智能扩展来扩展从URL中提取的关键字。这一实施例使用外部知识源，该外部知识源提供关键字到相关扩展的映射。例如，专家可创建在语义上相关的项。在这样的映射中，“auto insurance(汽车保险)”可被映射到“car insurance(轿车保险)”。扩展可在以上讨论的关键字组合阶段期间使用。在生成最初的关键字集合之后，使用智能扩展来为每一集合中的所有关键字检索和添加附加的关键字。如在先前的章节中所描述的来对具有扩展的新的集合来执行其余的组合过程。

[0061] 1.5.6相关性打分

[0062] 在本技术的一个实施例中，基于关键字的父段的位置、关键字的长度以及父段的长度计算关键字的相关性分数。第一，基于每个关键字在URL中的位置，每个关键字被分配一个被称为等级的在0到10之间的值。等级值随着在URL中从左到右移动而增加。出现在授权机构中的关键字具有比来自查询的关键字低的等级(片段＞查询＞路径＞授权机构)。使用父段的长度来归一化关键字k的等级。

[0063] $$k.level = \frac{k.level * k.len}{\sum_{i=0}^{n-1} r^i}$$

[0064] 其中k.len是关键字的长度，k.level是关键字的等级，而n是父段的长度。如果关键字是两个关键字k1和k2的组合，则关键字的等级可如下被归一化。

[0065] $$k.level = \frac{k1.level * k1.len + k2.level * k2.len}{\sum_{i=0}^{k1+k2} r^i}$$

[0066] 在0到10000的范围内计算关键字的最终相关性分数(Relevance Score)。它等于该URL可能的最大等级(MaxLevel)所归一化的关键字等级(KeyLevel)的1000倍。关键字的相关性分数由以下给出

[0067] $$RelevanceScore = \frac{log(1+\frac{Keylevel}{10}) * 10000}{log(1+\frac{Maxlevel}{10})}$$

[0068] 取决于所提取的关键字所用于的应用，相关性分数可进一步与关键字的其他度量进行组合。这些度量可在生成受控词汇时获得。例如，在广告应用中，投标广告客户的数量，用户查看、点击的次数，转换或价格都可以是要使用的重要的度量。

[0069] 1.5.6用从引用者URL中提取的关键字来捕捉用户意图

[0070] 在某些应用中，每次用户访问网页时就提取关键字以推断用户意图。在这样的场景中，利用引用者URL以及网页的URL也是可能的。引用者URL是用户从其请求当前页面的先前的网页的URL。它给出了用户在其中访问当前页面的上下文。在关键字提取技术的一个实施例中，当引用者URL也与查询URL一起可用时，使用上文中说明的提取方法单独地从这两个URL中提取关键字。通过组合来自这两个URL的关键字来准备最终的关键字列表。如果关键字源白这两个URL，则具有最高得分的关键字被保留而其他关键字被忽略。

[0071] 2.0示例性操作环境：

[0072] 本文所描述的关键字提取技术可在多种类型的通用或专用计算系统环境或配置内操作。图4示出其上可实现本文所描述的关键字提取技术的各实施例和元素的通用计算机系统的简化示例。应当注意，图4中由折线或虚线所表示的任何框表示简化计算设备的替换实施方式，并且以下描述的这些替换实施方式中的任一个或全部可以结合贯穿本文所描述的其他替换实施方式来使用。

[0073] 例如，图4示出了概括系统图，其示出简化计算设备400。这样的计算设备通常可以在具有至少一些最小计算能力的设备中找到，这些设备包括但不限于个人计算机、服务器计算机、手持式计算设备、膝上型或移动计算机、诸如蜂窝电话和PDA等通信设备、多处理器系统、基于微处理器的系统、机顶盒、可编程消费电子产品、网络PC、小型计算机、大型计算

机、音频或视频媒体播放器等。

[0074] 为允许设备实现关键字提取技术，该设备应当具有足够的计算能力和系统存储器以实现基本的计算操作。具体而言，如图4所示，计算能力一般由一个或多个处理单元410示出，并且还可包括一个或多个GPU 415，这两者中的任一个或全部与系统存储器420通信。注意，通用计算设备的处理单元410可以是专用微处理器，如DSP、VLIW、或其他微控制器、或可以是具有一个或多个处理核的常规CPU，包括多核CPU中的专用的基于GPU核。

[0075] 另外，图4的简化计算设备还可包括其他组件，诸如例如通信接口430。图4的简化计算设备还可包括一个或多个常规计算机输入设备440(例如，定点设备、键盘、音频输入设备、视频输入设备、触觉输入设备、用于接收有线或无线数据传输的设备等)。图4的简化计算设备还可包括其他光学组件，诸如例如一个或多个常规计算机输出设备450(例如，显示设备455、音频输出设备、视频输出设备、用于传送有线或无线数据传输的设备等)。注意，通用计算机的典型的通信接口430、输入设备440、输出设备450、以及存储设备460对本领域技术人员而言是公知的，并且在此不会详细描述。

[0076] 图4的简化计算设备还可包括各种计算机可读介质。计算机可读介质可以是可由计算机400经由存储设备460访问的任何可用介质，并且包括是可移动470和/或不可移动480的易失性和非易失性介质，该介质用于存储诸如计算机可读或计算机可执行指令、数据结构、程序模块或其他数据等信息。作为示例而非限制，计算机可读介质可包括计算机存储介质和通信介质。计算机存储介质包括但不限于：计算机或机器可读介质或存储设备，诸如DVD、CD、软盘、磁带驱动器、硬盘驱动器、光盘驱动器、固态存储器设备、RAM、ROM、EEPROM、闪存或其他存储器技术、磁带盒、磁带、磁盘存储或其他磁存储设备、或可用于存储所需信息并且可由一个或多个计算设备访问的任何其他设备。

[0077] 诸如计算机可读或计算机可执行指令、数据结构、程序模块等信息的存储还可通过使用各种上述通信介质中的任一种来编码一个或多个已调制数据信号或载波或其他传输机制或通信协议来实现，并且包括任何有线或无线信息传递机制。注意，术语“已调制数据信号”或“载波”一般指以对信号中的信息进行编码的方式设置或改变其一个或多个特征的信号。例如，通信介质包括诸如有线网络或直接线连接等携带一个或多个已调制数据信号的有线介质，以及诸如声学、RF、红外线、激光和其他无线介质等用于传送和/或接收一个或多个已调制数据信号或载波的无线介质。上述通信介质的任一组合也应包括在通信介质的范围之内。

[0078] 此外，可以按计算机可执行指令或其他数据结构的形式存储、接收、传送或者从计算机或机器可读介质或存储设备和通信介质的任何所需组合中读取具体化本文所描述的关键字提取技术的各种实施方式中的部分或全部的软件、程序和/或计算机程序产品或其各部分。

[0079] 最终，本文所描述的关键字提取技术还可在由计算设备执行的诸如程序模块等计算机可执行指令的一般上下文中描述。一般而言，程序模块包括执行特定任务或实现特定抽象数据类型的例程、程序、对象、组件、数据结构等。本文描述的各实施例还可以在其中任务由通过一个或多个通信网络链接的一个或多个远程处理设备执行或者在该一个或多个设备的云中执行的分布式计算环境中实现。在分布式计算环境中，程序模块可以位于包括媒体存储设备在内的本地和远程计算机存储介质中。此外，上述指令可以部分地或整体地

作为可以包括或不包括处理器的硬件逻辑电路来实现。

[0080] 还应当注意，可以按所需的任何组合来使用此处所述的上述替换实施例的任一个或全部以形成另外的混合实施例。尽管用结构特征和/或方法动作专用的语言描述了本主题，但可以理解，所附权利要求书中定义的主题不必限于上述具体特征或动作。上述具体特征和动作是作为实现权利要求的示例形式公开的。

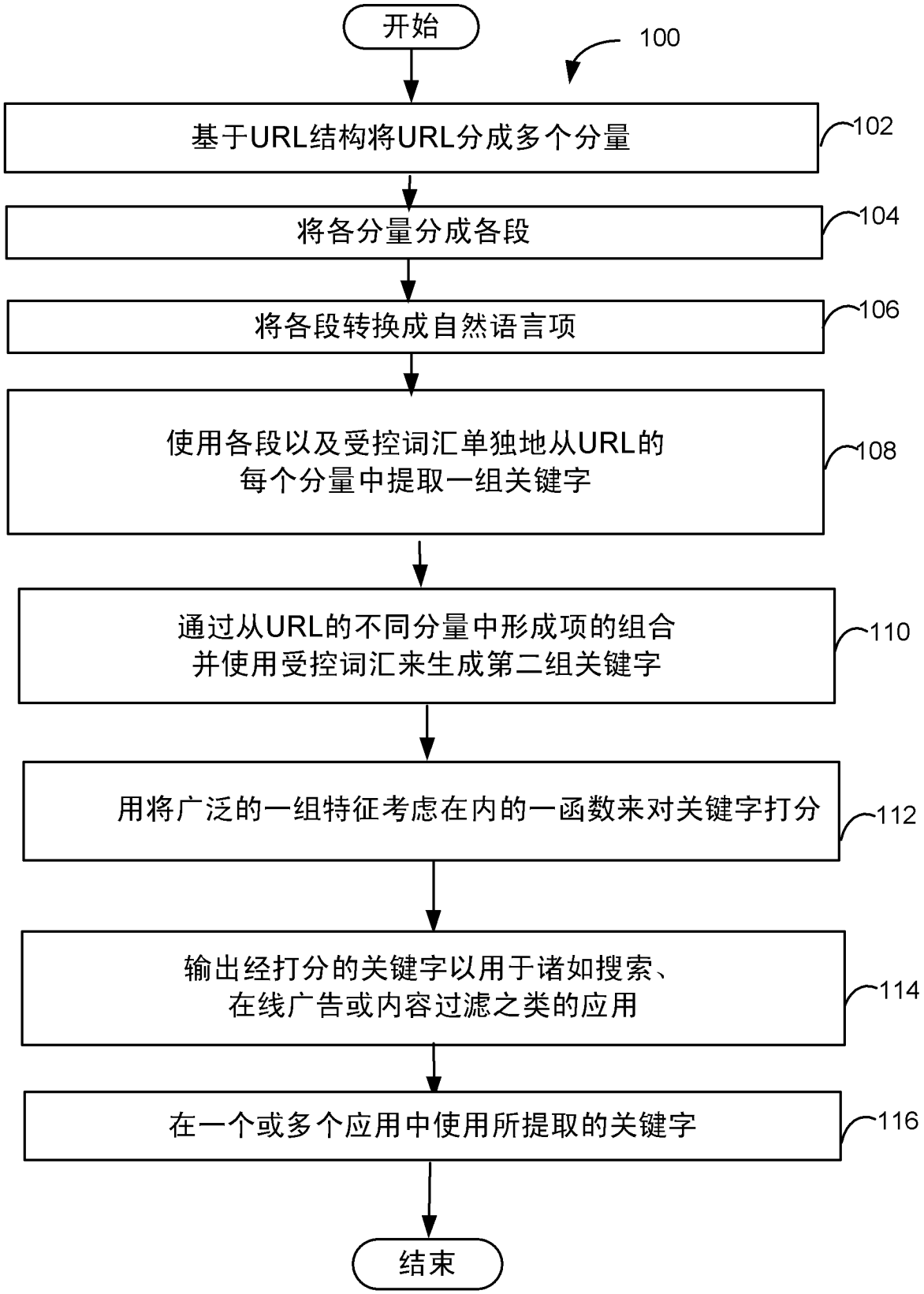

开始
100
基于URL结构将URL分成多个分量
102
将各分量分成各段
104
将各段转换成自然语言项
106
使用各段以及受控词汇单独地从URL的
每个分量中提取一组关键字
108
通过从URL的不同分量中形成项的组合
并使用受控词汇来生成第二组关键字
110
用将广泛的一组特征考虑在内的一函数来对关键字打分
112
输出经打分的关键字以用于诸如搜索、
在线广告或内容过滤之类的应用
114
在一个或多个应用中使用所提取的关键字
116
结束

图1

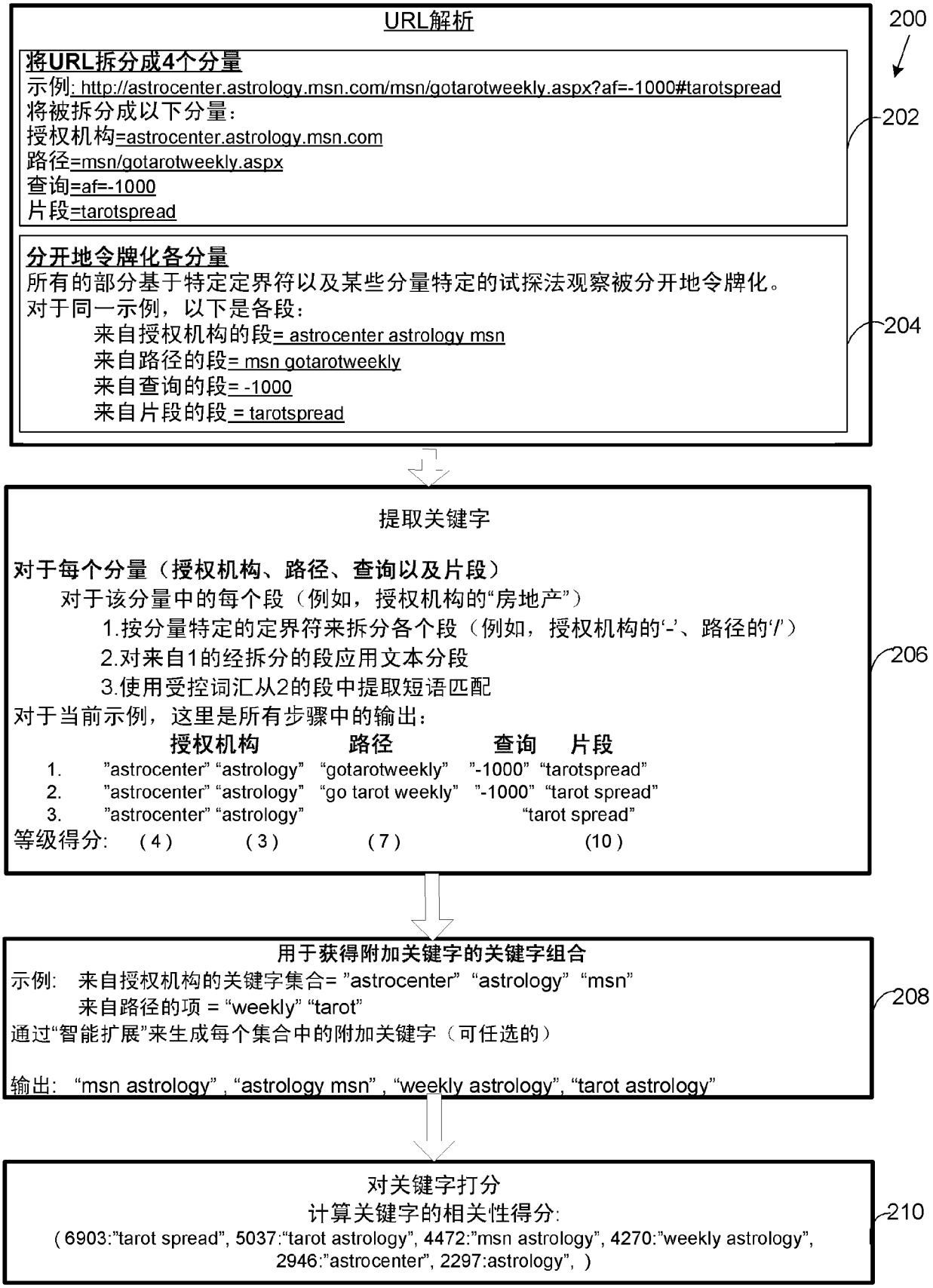

URL解析
200
将URL拆分成4个分量
示例: http://astrocenter.astrology.msn.com/msn/gotarotweekly.aspx?af=-1000#tarotspread
将被拆分成以下分量:
授权机构=astrocenter.astrology.msn.com
路径=msn/gotarotweekly.aspx
查询=af=-1000
片段=tarotspread
202
分开地令牌化各分量
所有的部分基于特定定界符以及某些分量特定的试探法观察被分开地令牌化。
对于同一示例,以下是各段:
来自授权机构的段= astrocenter astrology msn
来自路径的段= msn gotarotweekly
来自查询的段= -1000
来自片段的段 = tarotspread
204
提取关键字
对于每个分量(授权机构、路径、查询以及片段)
对于该分量中的每个段(例如,授权机构的"房地产")
1.按分量特定的定界符来拆分各个段(例如,授权机构的'-'、路径的'/')
2.对来自1的经拆分的段应用文本分段
3.使用受控词汇从2的段中提取短语匹配
对于当前示例,这里是所有步骤中的输出:
授权机构 路径 查询 片段
1. "astrocenter" "astrology" "gotarotweekly" "-1000" "tarotspread"
2. "astrocenter" "astrology" "go tarot weekly" "-1000" "tarot spread"
3. "astrocenter" "astrology" "tarot spread"
等级得分: (4) (3) (7) (10)
206
用于获得附加关键字的关键字组合
示例: 来自授权机构的关键字集合= "astrocenter" "astrology" "msn"
来自路径的项 = "weekly" "tarot"
通过"智能扩展"来生成每个集合中的附加关键字(可任选的)
输出: "msn astrology", "astrology msn", "weekly astrology", "tarot astrology"
208
对关键字打分
计算关键字的相关性得分:
(6903:"tarot spread", 5037:"tarot astrology", 4472:"msn astrology", 4270:"weekly astrology", 2946:"astrocenter", 2297:astrology",)
210

图2

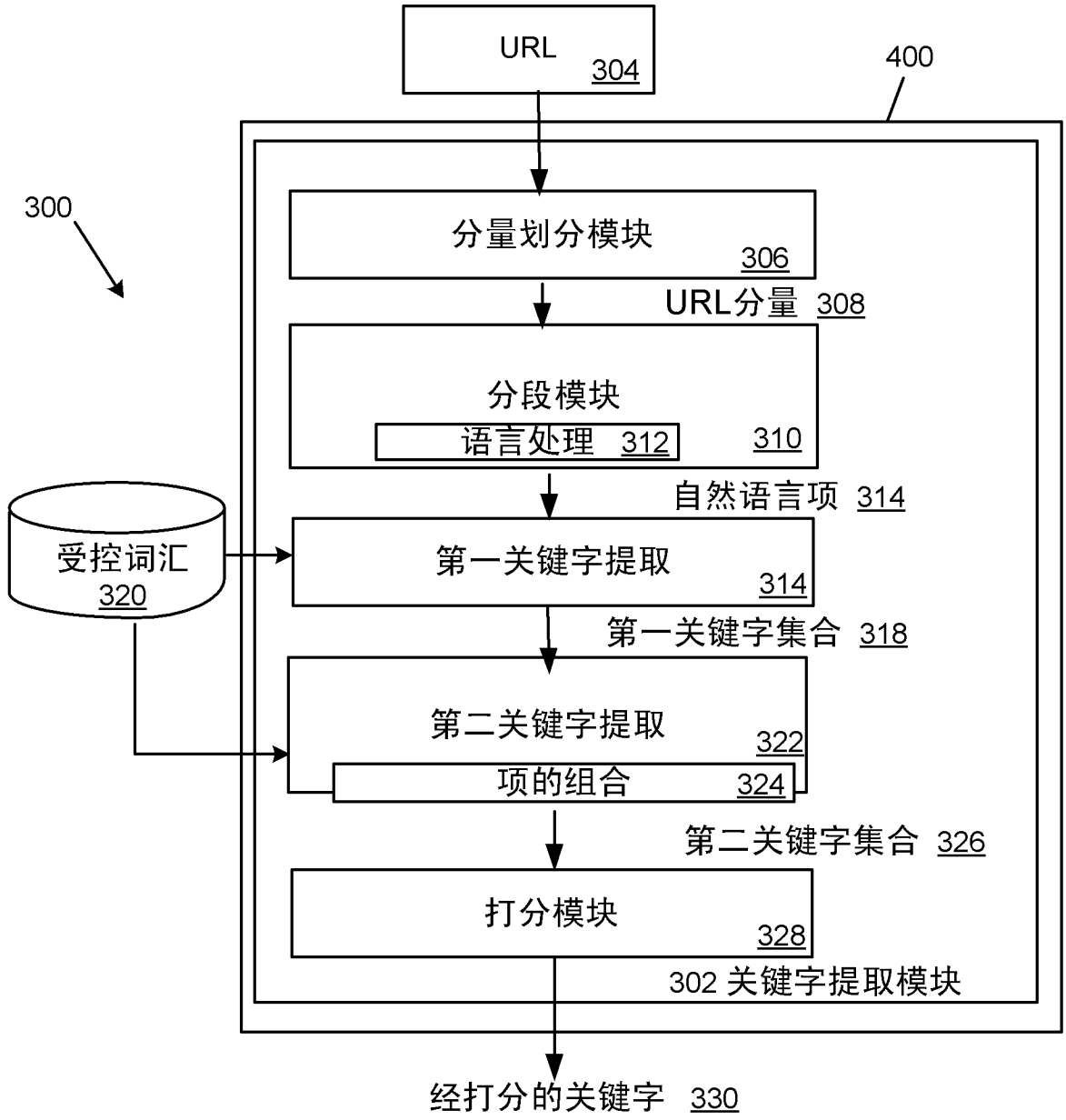

URL
304
400
300
分量划分模块
306
URL分量 308
分段模块
语言处理 312
310
自然语言项 314
受控词汇
320
第一关键字提取
314
第一关键字集合 318
第二关键字提取
322
项的组合 324
第二关键字集合 326
打分模块
328
302 关键字提取模块
经打分的关键字 330

图3

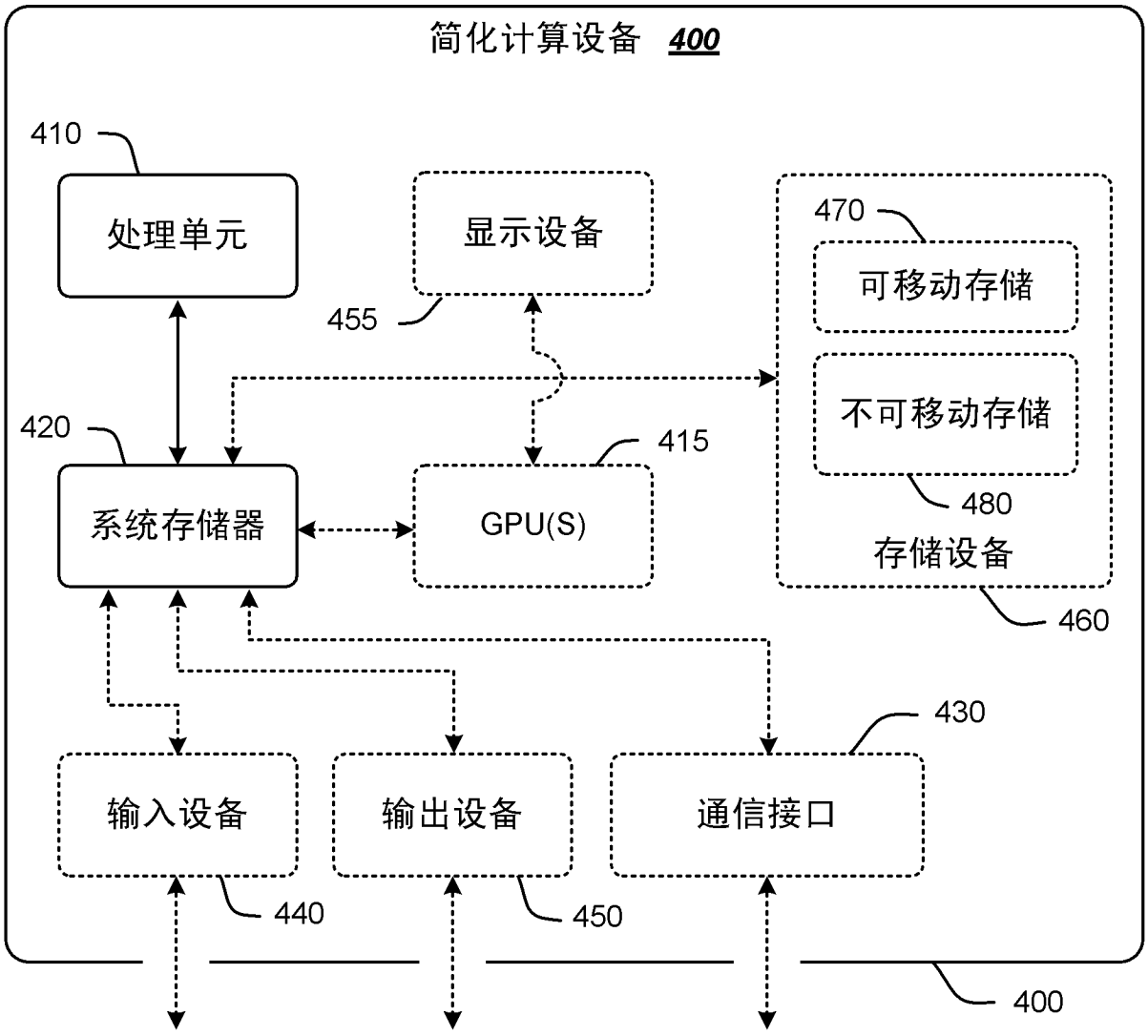

简化计算设备 400
410
处理单元
显示设备
455
470
可移动存储
不可移动存储
480
存储设备
460
420
系统存储器
415
GPU(S)
430
输入设备
输出设备
通信接口
440
450
400

图4