(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 109844737 A
(43)申请公布日 2019.06.04

(21)申请号 201780065526.X

(22)申请日 2017.08.10

(30)优先权数据
15/245265 2016.08.24 US

(85)PCT国际申请进入国家阶段日
2019.04.23

(86)PCT国际申请的申请数据
PCT/EP2017/070271 2017.08.10

(87)PCT国际申请的公布数据
WO2018/036827 EN 2018.03.01

(71)申请人 罗伯特·博世有限公司
地址 德国斯图加特

(72)发明人 D.S.金 C.陈

(74)专利代理机构 中国专利代理(香港)有限公司 72001
代理人 毕铮 申屠伟进

(51)Int.Cl.
G06F 16/958(2019.01)
G06F 16/80(2019.01)

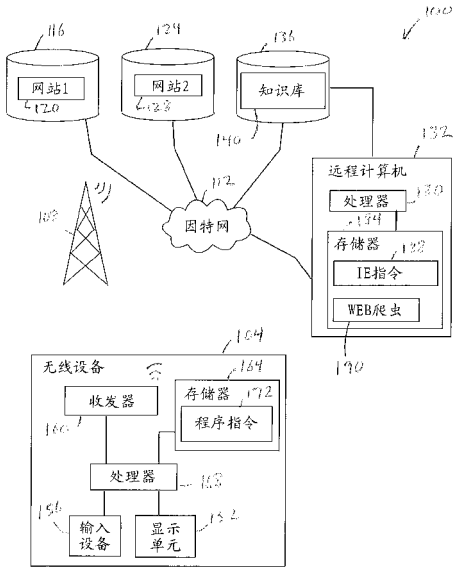
权利要求书3页 说明书12页 附图9页

(54)发明名称

用于非监督式信息提取的方法和设备

(57)摘要

一种将有信息量内容添加到电子知识库的方法包括根据多个网站来生成多个数据库表,每个数据库表基于对应的网站;以及单独地处理每个数据库表以从每个数据库表中标识潜在有信息量内容。所述方法此外包括将被包括在所述多个数据库表中的第一预定数目的数据库表中的潜在有信息量内容标识为有信息量内容,所述第一预定数目大于一;以及将所述有信息量内容添加到电子知识库。



CN 109844737 A

1.一种将有信息量内容添加到电子知识库的方法,包括:

根据多个网站来生成多个数据库表,每个数据库表基于对应的网站;

单独地处理每个数据库表以从每个数据库表中标识潜在有信息量内容;

将被包括在所述多个数据库表中的第一预定数目的数据库表中的潜在有信息量内容标识为有信息量内容,所述第一预定数目大于一;以及

将所述有信息量内容添加到电子知识库。

2.根据权利要求1所述的方法,其中单独地处理每个数据库表包括:

将潜在有信息量内容标识为具有小于数据库表中对应一个数据库表中的预定出现数目的内容。

3.根据权利要求1所述的方法,其中:

所述多个数据库表中的第一数据库表根据具有多个网页的所述多个网站中的第一网站而被生成;

每个网页包括多个内容项;并且

单独地处理每个数据库表以从每个数据库表中标识潜在有信息量内容包括处理第一数据库表来将所述多个内容项中出现在第一网站的仅一个网页上的内容项标识为潜在有信息量内容。

4.根据权利要求1所述的方法,其中标识为有信息量内容包括:

将评分指派给来自每个数据库表的潜在有信息量内容;

比较来自第一数据库表的潜在有信息量内容的第一评分与来自第二数据库表的潜在有信息量内容的第二评分;以及

如果第一评分与第二评分的比较指示了来自第一数据库表的潜在有信息量内容被包括在第一数据库表和第二数据库表中,则确定来自第一数据库表的潜在有信息量内容是有信息量内容。

5.根据权利要求4所述的方法,其中:

第一数据库表包括与第一数据库表的潜在有信息量内容相关联的第一多个数据项;

第二数据库表包括与第二数据库表的潜在有信息量内容相关联的第二多个数据项;

将评分指派给来自每个数据库表的潜在有信息量内容包括

将数值指派给所述第一多个数据项中的数据项以形成第一评分向量,以及

将数值指派给所述第二多个数据项中的数据项以形成第二评分向量;

比较第一评分与第二评分包括确定在第一评分向量与第二评分向量之间的余弦相似性;以及

标识为有信息量内容包括如果所述余弦相似性大于或等于预定阈值,则确定来自第一数据库表的潜在有信息量内容被包括在第一数据库表和第二数据库表中。

6.根据权利要求5所述的方法,其中将评分指派给来自每个数据库表的潜在有信息量内容此外包括:

使被指派到所述第一多个数据项中在所述第一多个数据项中最频繁出现的数据项的数值加倍;以及

使被指派到所述第二多个数据项中在所述第二多个数据项中最频繁出现的数据项的数值加倍。

7.根据权利要求5所述的方法,此外包括:

通过分析与第一数据库表相关联的第一网站的HTML标签序列来确定所述第一多个数据项;以及

通过分析与第二数据库表相关联的第二网站的HTML标签序列来确定所述第二多个数据项。

8.根据权利要求7所述的方法,其中第一网站和第二网站具有不同的网际协议地址。

9.一种信息提取系统,包括:

远程计算机,其可操作地被连接到因特网并且包括处理器,所述处理器被配置成:

根据经由因特网可访问的多个网站来生成多个数据库表,每个数据库表基于对应的网站;

单独地处理每个数据库表以从每个数据库表中标识潜在有信息量内容;

将被包括在所述多个数据库表中的第一预定数目的数据库表中的潜在有信息量内容标识为有信息量内容,所述第一预定数目大于一;以及

将所标识的有信息量内容添加到电子知识库。

10.根据权利要求9所述的信息提取系统,其中所述处理器此外被配置成:

将潜在有信息量内容标识为具有小于数据库表中对应一个数据库表中的预定出现数目的内容。

11.根据权利要求9所述的信息提取系统,其中:

所述多个数据库表中的第一数据库表根据具有多个网页的所述多个网站中的第一网站而被生成;

每个网页包括多个内容项;并且

所述处理器此外被配置成将所述多个内容项中出现在第一网站的仅一个网页上的内容项标识为潜在有信息量的。

12.根据权利要求9所述的信息提取系统,其中所述处理器此外被配置成:

通过如下操作来标识有信息量内容

将评分指派给来自每个数据库表的潜在有信息量内容;

比较来自第一数据库表的潜在有信息量内容的第一评分与来自第二数据库表的潜在有信息量内容的第二评分;以及

如果第一评分与第二评分的比较指示了来自第一数据库表的潜在有信息量内容被包括在第一数据库表和第二数据库表中,则确定来自第一数据库表的潜在有信息量内容是有信息量内容。

13.根据权利要求12所述的信息提取系统,其中:

第一数据库表包括与第一数据库表的潜在有信息量内容相关联的第一多个数据项;

第二数据库表包括与第二数据库表的潜在有信息量内容相关联的第二多个数据项;

所述处理器此外被配置成通过如下操作来将评分指派给来自每个数据库表的潜在有信息量内容

将数值指派给所述第一多个数据项中的数据项以形成第一评分向量,以及

将数值指派给所述第二多个数据项中的数据项以形成第二评分向量;

比较第一评分与第二评分包括确定在第一评分向量与第二评分向量之间的余弦相似

性;以及

标识为有信息量内容包括如果所述余弦相似性大于或等于预定阈值,则确定来自第一数据库表的潜在有信息量内容被包括在第一数据库表和第二数据库表中。

14.根据权利要求13所述的信息提取系统,其中所述处理器此外被配置成通过如下操作来指派评分:(i)使被指派到所述第一多个数据项中在所述第一多个数据项中最频繁出现的数据项的数值加倍,以及(ii)使被指派到所述第二多个数据项中在所述第二多个数据项中最频繁出现的数据项的数值加倍。

15.根据权利要求13所述的信息提取系统,其中所述处理器此外被配置成:

通过分析与第一数据库表相关联的第一网站的HTML标签序列来确定所述第一多个数据项中的数据项;以及

通过分析与第二数据库表相关联的第二网站的HTML标签序列来确定所述第二多个数据项中的数据项。

16.根据权利要求15所述的信息提取系统,其中第一网站和第二网站具有不同的网际协议地址。

17.根据权利要求9所述的信息提取系统,此外包括:

可操作地被连接到因特网和远程计算机的数据服务器,所述数据服务器被配置成存储电子知识库。

18.根据权利要求9所述的信息提取系统,此外包括:

客户端,其可操作地被连接到因特网并且被配置成访问电子知识库。

用于非监督式信息提取的方法和设备

技术领域

[0001] 本公开内容涉及信息提取的领域，并且特别地涉及改善通过非监督式信息提取所生成的数据的有用性。

背景技术

[0002] 大量数据作为HTML网站被存储在因特网上。典型地，以人类读者容易理解的方式组织和呈现数据。例如，在查看包含对合适的停车场区域进行描述的信息的网站的网页中，人类读者典型地能够定位并且理解相关信息。人类读者针对上下文线索或标注而略读网页，所述上下文线索或标注指向或建议相关信息的位置。即使信息被包含在表格中或被混合在文本框内，人类读者也定位相关信息。在定位了相关信息之后，人类读者记录信息或对其做出心记使得相关信息可以在适当的时间、诸如当导航到所期望的停车场区域的时候被使用。如以上所阐明的，人类读者能够略读多段网页并且定位相关信息是几秒钟的事情；然而，对于计算机而言难以从网页中高效地标识相关信息。

[0003] 信息提取(“IE”)是指使用计算机来从网站提取相关信息的过程。所提取的信息然后被存储到通过其它计算机容易可访问且可搜索的经组织的相关信息的数据库。IE的已知方式是监督式或非监督式的。监督式的IE需要工程师或技术员回顾从网站所提取的信息并且手动地确定信息是否是合期望的。也就是说，工程师或技术员手动地在无用或无兴趣的信息与有用或引起兴趣的信息之间进行辨别。工程师使得计算机通过如下操作来将有用、引起兴趣和/或有信息量的信息(在下文中统称为“有信息量内容”)存储到数据库并且丢弃无用或无兴趣的信息：创建规则集或训练示例以用于计算机遵循。规则和训练示例中的一些可特定于仅仅单个网站或网页的所提取的信息；然而，其它规则可具有更全局的使用，使得随着时间计算机在标识有信息量内容方面可变得更高效。非监督式的IE不需要工程师或技术员创建规则来用于确定所提取的信息是否有用或引起兴趣。代替地，参与非监督式IE的计算机在所提取的信息之上执行统计分析以标识有信息量内容并且以数据库表的形式输出所期望的数据。由于非监督式IE需要很少至无人类干预，所以它通常比监督式IE更快且更高效。

[0004] 非监督式IE与监督式IE相比典型地不太准确。典型地，执行非监督式IE的已知系统生成“假阳性”，其是如下数据：人类将会认为所述数据无用或无兴趣，但是计算机确定为是有用或引起兴趣的。当系统将假阳性存储到有信息量内容的知识库的时候，系统的准确性降低。此外，如果不以符合计算机所应用的统计分析途径的方式呈现网站的信息内容，那么计算机可能不恰当地提取并且组织有信息量内容。

[0005] 非监督式IE具有更高效地将有信息量内容添加到知识库的潜能。然而，存在增大非监督式IE的精度和查全率的持续的需要。因而，在非监督式IE的领域中的进一步的发展是合期望的。

发明内容

[0006] 根据本公开内容的示例性实施例，一种用于将有信息量内容添加到电子知识库的方法包括：根据多个网站来生成多个数据库表，每个数据库表基于对应的网站；以及单独地处理每个数据库表以从每个数据库表中标识潜在有信息量内容。所述方法此外包括将被包括在所述多个数据库表中的第一预定数目的数据库表中的潜在有信息量内容标识为有信息量内容，所述第一预定数目大于一；以及将所述有信息量内容添加到电子知识库。

[0007] 根据本公开内容的另一示例性实施例，一种信息提取系统包括可操作地连接到因特网并且包括处理器的远程计算机，所述处理器被配置成：根据经由因特网可访问的多个网站来生成多个数据库表，每个数据库表基于对应的网站；单独地处理每个数据库表以从每个数据库表中标识潜在有信息量内容；将被包括在所述多个数据库表中的第一预定数目的数据库表中的潜在有信息量内容标识为有信息量内容，所述第一预定数目大于一；以及将所标识的有信息量内容添加到电子知识库。

附图说明

[0008] 通过参考以下详细描述以及附图，上述特征和优点以及其它应当变得对于本领域普通技术人员而言更加容易清楚，在所述附图中：

图1是如本文中所公开的非监督式信息提取系统的框图；

图2是具有将由图1的非监督式信息提取系统来提取的数据的示例性第一网站的框图；

图3是具有将由图1的非监督式信息提取系统来提取的数据的示例性第二网站的框图；

图4是一流程图，其图示了操作图1的信息提取系统的示例性方法；

图5图示了与图2的网站相关联的示例性数据库表；

图6图示了与图3的网站相关联的示例性数据库表；

图7图示了来自图5的数据库表的所选数据群组，来自图6的数据库表的另一所选数据群组，以及基于来自所选数据群组的数据项的评分向量的表；

图8图示了来自图5的数据库表的所选数据群组，来自图6的数据库表的另一所选数据群组，以及基于来自所选数据群组的数据项的评分向量的表；

图9A是针对包括来自机场网站的数据的语料库的精度相对于语料库大小的图；

图9B是针对包括来自机场网站的数据的语料库的查全率相对于语料库大小的图；

图9C是针对包括来自机场网站的数据的语料库的F-评分相对于语料库大小的图；

图9D是针对包括来自医院网站的数据的语料库的精度相对于语料库大小的图；

图9E是针对包括来自医院网站的数据的语料库的查全率相对于语料库大小的图；并且

图9F是针对包括来自医院网站的数据的语料库的f-评分相对于语料库大小的图。

具体实施方式

[0009] 为了促进理解本公开内容的原理的目的，现在将参考在附图中所图示的以及在以下所撰写的说明书中所描述的实施例。理解到，从而不意图对公开内容范围的任何限制。此外理解到，本公开内容包括对所图示的实施例的任何变更和修改，并且包括如本公开内容所关于的领域中的技术人员通常将会想到的本公开内容原理的另外的应用。

[0010] 在随附的描述中公开本公开内容的各方面。可以设计本公开内容及其等同物的可

替换实施例，而不离开本公开内容的精神或范围。应当注意到，本文中关于“一个实施例”、“实施例”、“示例性实施例”等等的任何讨论指示：所述的实施例可以包括特定特征、结构或特性，并且这样的特定特征、结构或特性可能不一定被包括在每一个实施例中。另外，对前述内容的提及不一定包括对相同实施例的提及。最后，无论是否被明确地描述，本领域普通技术人员都将会容易地领会到给定实施例的特定特征、结构或特性中的每一个可以结合或组合本文中所讨论的任何其它实施例中的那些被利用。

[0011] 为了本公开内容的目的，短语“A和/或B”意指(A)、(B)或(A和B)。为了本公开内容的目的，短语“A、B和/或C”意指(A)、(B)、(C)、(A和B)、(A和C)、(B和C)、或(A、B和C)。

[0012] 如关于本公开内容的实施例所使用的术语“包括”、“包括有”、“具有”等等是同义的。

[0013] 如图1中所示，信息提取系统100包括被配置用于与蜂窝网络108和因特网112通信的无线通信设备104，被配置用以服务并且存储第一网站120的第一数据服务器116，被配置用以服务并且存储第二网站128的第二数据服务器124，被配置用于与因特网112通信的远程计算机132，以及电连接到远程计算机132并且被配置用以服务并且存储电子知识库140的第三数据服务器136。IE系统100被配置成通过如下操作来将有信息量内容(即有用、引起兴趣和/或所期望的信息)添加到知识库140：在没有通过利用两种类型的冗余处理的非监督式IE的过程的监督的情况下从至少网站120、128提取信息。特别地，由IE系统100所使用的过程通过如下操作来准确地区分有信息量内容与非有信息量内容(即无用、无兴趣和/或非期望的信息)：通过使用内部冗余途径和之间冗余途径来处理从至少网站120、128所提取的数据。所述两种类型的冗余处理大大增大IE系统100的精度和查全率，使得较少的假阳性被生成并且被存储到知识库140。以下描述IE系统100的组件，继之以对由IE系统100使用来将有信息量内容添加到知识库140的方法400(图4)的描述。

[0014] 无线通信设备104是一种示例性的客户端设备，所述示例性的客户端设备被配置成经由因特网112来向知识库140发送查询，并且经由因特网112来接收被存储在知识库140中的数据。设备104包括显示单元152、输入设备156、收发器160以及存储器164，其各自可操作地被连接到处理器168。无线设备104典型地是蜂窝电话、移动电话、智能电话、平板计算机、或任何其它合适的设备。

[0015] 显示单元152是一种液晶显示(LCD)面板，其被配置成显示文本、图像、和其它视觉上可理解的数据。在另一实施例中，显示单元152是如本领域普通技术人员所期望的任何显示器，包括但不限于有源矩阵有机发光二极管显示器。

[0016] 输入设备156被配置成使得用户能够录入数据并且操纵被示出在显示单元152上的对象。例如，输入设备156被配置成生成输入数据，所述输入数据对应于将被发送到知识库140的质询。在另一实施例中，输入设备156是被施加在显示单元152之上的触摸屏，所述显示单元152被配置成响应于手指或触笔的触摸。在又一实施例中，输入设备156是被配置成生成输入信号的任何设备，如本领域普通技术人员所期望的那样。

[0017] 收发器160，其还被称为无线发射器和接收器，可操作地连接到处理器168并且被配置成与蜂窝网络108、无线局域网(“Wi-Fi”)、个域网和/或任何其它无线网络无线地通信。因此，收发器160与任何所期望的无线通信标准或协议兼容，所述无线通信标准或协议包括但不限于近场通信(“NFC”)、IEEE 802.11、IEEE 802.15.1(“Bluetooth®”)、全球移动

系统(“GSM”)、以及码分多址(“CDMA”)。

[0018] 处理器168被配置成执行存储器164中所存储的程序指令172,以用于操作被连接到那里的组件,诸如显示单元152、输入设备156以及收发器160。处理器168被提供为微处理器、控制器或任何其它类型的电子控制芯片。在一个实施例中,处理器168执行程序指令172(即软件、应用、或“app”),其下载自因特网112。例如,用于向知识库140发送质询并且用于从知识库140接收所选数据的程序指令172可以从因特网112被下载、被存储在存储器164中,并且由处理器168执行。

[0019] 存储器164是电子数据存储单元,其在本文中还被称为非暂时性计算机可读介质。如本文中所阐明的,存储器164被配置成存储程序指令172以用于操作无线设备104。任何其它电子数据也可以被存储在存储器164中,诸如经由因特网112从知识库140被下载的数据。

[0020] 第一数据服务器116可操作地被连接到因特网112并且被配置成服务至少第一网站120并且存储与至少第一网站120相关联的数据。更具体地,响应于经由因特网112来访问网站120,数据服务器116从因特网112接收针对第一网站120的数据,并且将数据发送到客户端设备或服务客户端设备,所述客户端设备诸如无线设备104和远程计算机132。

[0021] 第二数据服务器124可操作地被连接到因特网112并且被配置成服务至少第二网站128并且存储与至少第二网站128相关联的数据。数据服务器124被配置成响应于经由因特网112访问网站128而从因特网112接收针对第二网站128的数据并且将数据发送到客户端设备或服务客户端设备。

[0022] 第三数据服务器136可操作地被连接到因特网112以及远程计算机132。第三数据服务器136被配置成服务至少知识库140并且存储与至少知识库140相关联的数据。数据服务器136从远程计算机132接收有信息量内容数据,所述有信息量内容数据将被存储或添加到知识库140。数据服务器136还从客户端设备接收针对知识库140的质询数据,并且响应于所述质询数据而将数据发送到客户端设备或服务客户端设备。如图1中所示,数据服务器136直接电连接到远程计算机132。在其它实施例中,数据服务器136可以通过因特网112或另一计算机网络而间接地电连接到远程计算机132。

[0023] 知识库140是至少部分地由远程计算机132所构建的有信息量内容的数据库。知识库140包括例如从网站120、128所提取的有信息量内容。在一个实施例中,知识库140是专门化的知识库,其包括与特定研究领域或感兴趣的区域相关的有信息量内容。在其它实施例中,知识库140是一般性知识库,其包括所有类型的有信息量内容。在一个特定的示例中,知识库140包括与某个地理区域的机场中和周围的汽车停车信息有关的有信息量内容。知识库140被配置成接收并且处理与停车信息有关的质询数据,并且向对应的客户端设备提供与质询有关的有信息量内容。例如,客户端设备可以向知识库140发送质询,询问“在米内塔圣荷西国际机场附近最不昂贵的停车场是什么?”知识库140可以利用有信息量内容数据来进行响应,所述有信息量内容数据包括停车场的地址以及停在那里每天或每小时的花费。

[0024] 远程计算机132包括可操作地被连接到存储器184的处理器180,所述存储器184被配置成存储程序指令188以及web爬虫程序190。程序指令188和web爬虫程序190与非监督式信息提取方法400相关联。处理器180被提供为微处理器、控制器或任何其它类型的电子控制芯片。存储器184是电子数据存储单元,其在本文中还被称为非暂时性计算机可读介质。如本文中所阐明的,存储器184被配置成存储程序指令188。任何其它电子数据也可以被存

储在存储器184中，诸如web爬虫190数据、web抓取器数据、以及经由因特网112从知识库140和无线设备104所下载的其它数据。

[0025] 为了辅助描述方法400，在图2和3中分别示出了网站120、128的示例性配置。第一网站120包括网际协议地址(“IP地址”)，使得网站120在因特网112上是可寻址的。第一网站120还包括与IP地址相关联的统一资源定位符(“URL”)。向域名服务器(没有被示出)注册第一网站120的URL和IP地址，使得通过远程计算机132和无线设备104可访问网站120。第二网站128也包括IP地址，使得网站128在因特网112上是可寻址的。第二网站128包括与IP地址相关联的URL。向域名服务器注册第二网站128的URL和IP地址，使得通过远程计算机132和无线设备104可访问第二网站128。第二网站128的URL和IP地址的主机名不同于第一网站120的URL和IP地址的主机名。

[0026] 如图2中所示，第一网站120包括多个网页196A、196B，其中的两个被图示。网站120可以具有例如从两个到五十个网页196A、196B。每个网页196A、196B包括多个内容项，所述多个内容项包括菜单202、侧边栏204、页脚208以及内容区域212A、212B。在每个网页196A、196B上，菜单202、侧边栏204和页脚208是相同的、大体上相同的、或等同的。菜单202包括例如去往网页196A、196B的链接。在网页196A、196B中的每一个上，菜单202包括相同的链接。侧边栏204包括例如被认为可适用于网页196A、196B中每一个的文本。在网页196A、196B中的每一个上，侧边栏204包括相同的文本。页脚208包括例如联系信息以及去往其它网站的链接。在网页196A、196B中的每一个上，页脚208包括相同的联系信息和链接。第一网站120的内容212A、212B包括以表、段落或任何其它布置所组织的文本、数字和/或其它符号。在每个网页196A、196B上，内容212A、212B是不同的。这样，内容212A不同于内容212B。网站120的每个网页196A、196B在本文中还被称为HTML文档。

[0027] 如图3中所示，第二网站128包括多个网页218A、218B，其中的两个被图示。每个网页218A、218B包括菜单222、侧边栏226、页脚230以及内容区域234C、234D。在每个网页218A、218B上，菜单222、侧边栏226、和页脚230是相同的、大体上相同的或等同的，并且不同于网站120的菜单202、侧边栏204和页脚208。菜单222包括例如去往网页218A、218B的链接。在网页218A、218B中的每一个上，菜单222包括相同的链接。侧边栏226包括例如被认为可适用于网页218A、218B中每一个的文本。在网页218A、218B中的每一个上，侧边栏226包括相同的文本。页脚230包括例如联系信息以及去往其它网站的链接。在网页218A、218B中的每一个上，页脚230包括相同的联系信息和链接。网站128的内容234C、234D包括以表、段落或任何其它布置所组织的文本。在每个网页218A、218B上，内容234C、234D是不同的，并且至少以某些方式不同于网站120的内容212A、212B。每个网页218A、218B在本文中还被称为HTML文档。

[0028] 系统的操作

在操作中，远程计算机132的处理器180执行程序指令188，用于从至少网站120、128执行非监督式IE的方法400。在方法400的描述中，方法400正执行某个任务或功能的陈述是指：控制器或通用处理器执行在被操作地连接到控制器或处理器的非暂时性计算机可读存储介质中所存储的经编程的指令，以操纵数据或操作IE系统100中的一个或多个组件来执行该任务或功能。特别地，远程计算机132的处理器180和/或设备104的处理器168的处理器电路/逻辑可以是这样的控制器或处理器。可替换地，处理器168、180可以利用多于一个处理器以及相关联的电路和组件来被实现，其中的每一个被配置成形成本文中所述的一个或

多个任务或功能。另外，方法400可以按任何可行的时间次序来被执行，而无论图中所示的次序或用来描述方法400的次序如何。

[0029] 方法400使得远程计算机132将来自网站、诸如网站120、128的有信息量内容添加到知识库、诸如知识库140。方法400确定网站120、128的内容的“兴趣性”，并且在一个实施例中，仅仅将有信息量内容添加到知识库140。方法400通过利用两种类型的冗余处理来标识在网站120、128中所包含的有信息量内容——如果有的话。第一类型的冗余处理是跨一个特定网站120、128的网页的冗余(即内部冗余)，并且第二类型的冗余处理是跨不同网站120、128的冗余(即之间冗余)。简言之，有信息量内容典型地不是内部冗余的，典型地是之间冗余的。利用这两种类型的冗余处理，方法400执行多个网站的非监督式IE。方法400标识有信息量内容并且将有信息量内容存储在知识库140中，使得客户端计算机、诸如无线设备104可以用自动化方式访问并且操纵所提取的信息。

[0030] 从输入网站提取数据

如图4的框404中所示，方法400包括通过使用远程计算机132来从输入网站、诸如网站120、128提取数据。提取数据包括将网站120、128的电子数据从数据服务器116、124传递到远程计算机132的存储器184。在一个实施例中，处理器180执行web爬虫程序190来定位用于IE的合适网站120、128。在已经标识了合适的网站之后，方法400使用例如被称为web抓取的技术，用于下载输入网站的电子数据并且将所述电子数据组织在数据语料库中以用于处理。方法400还包括使用任何过程、方法或应用以便抵达数据语料库以用于处理。

[0031] 通过使用非监督式IE途径来处理所提取的数据

方法400的框404此外包括通过使用非监督式IE途径来处理来自网站120、128的电子数据，在所述非监督式IE途径中，内容项被组织到多个数据库表240、244中，如图5和6中所示。远程计算机132使用任何类型的非监督式IE途径来预备数据库表240、244，诸如包装器归纳和自动提取。在示例性的实施例中，远程计算机132使用非监督式IE途径，所述非监督式IE途径被称为使用部分树对准的数据提取(下文中“DEPTA”)。在以下论文中描述了DEPTA：YanHong Zhai和Bing Liu (2005)的《Web Data Extraction based on Partial Tree Alignment》，其公开内容通过引用以其全部被并入本文中。DEPTA是两部分过程，其包括标识网页中单独的数据记录，以及从所标识的数据记录中提取数据项。

[0032] 如图5和6中所示，已经根据DEPTA被处理的网站数据被组织到多个示例性的数据库表240、244中。远程计算机132通过使用DEPTA来处理第一网站120的内容项，用于生成数据库表240，所述数据库表240被存储在存储器184中。数据库表240包括针对网站120的每个内容项的数据群组248、250、252、254、256、258、260、262(即网站120的每个数据记录)。数据群组248对应于网页196A的菜单202，数据群组250对应于网页196A的侧边栏204，数据群组252对应于网页196A的页脚208，数据群组254对应于网页196A的内容212A，数据群组256对应于网页196B的菜单202，数据群组258对应于网页196B的侧边栏204，数据群组260对应于网页196B的页脚208，数据群组262对应于网页196B的内容212B。

[0033] 远程计算机132还通过使用DEPTA来处理网站128的内容项，并且生成数据库表244，所述数据库表244被存储在存储器184中。数据库表244包括针对网站128的每个内容项的数据群组266、268、270、272、274、276、278、280(即网站128的每个数据记录)。数据群组266对应于网页218A的菜单222，数据群组268对应于网页218A的侧边栏226，数据群组270对

应于网页218A的页脚230，数据群组272对应于网页218A的内容234C，数据群组274对应于网页218B的菜单222，数据群组276对应于网页218B的侧边栏226，数据群组278对应于网页218B的页脚230，数据群组280对应于网页218B的内容234D。

[0034] 每个数据群组248、250、252、254、256、258、260、262、266、268、270、272、274、276、278、280包括数据项286的一列(图5和6的左列)以及评注288的一列(图5和6的右列)。数据项286例如是用户当使用无线设备104来访问知识库140的时候可搜索的项或关键词。数据项286还可以例如是一个或多个句子、短语、段落或文本页。在图5和6中，代替于示出针对每个数据项286的整个文本段落，代替地仅仅示出数据项286的最频繁使用的中心名词。因此，中心名词“出租车”是在数据群组254的第一数据项286中最频繁使用的中心名词。评注288是与数据项286相关联的文本、数字或其它数据。在一个示例中，数据项286可以包括描述出租车公司的文本的段落或行，并且对应的评注288包括出租车公司的电话号码。在另一示例中，数据项286可以包括描述医院的段落，并且对应的评注288包括医院的地址。数据库表240、244包括主要占位符数据项286和评注288。为与内容212A、212B、234C、234D相关联的数据群组254、262、272、280提供实际的数据项286。在一个实施例中，处理器180通过分析网站120、128的HTML标签序列来确定数据项286和评注288。然而，任何其它方法也可以被用于标识数据项286和评注288。在所图示的实施例中，每个数据群组248、250、252、254、256、258、260、262、266、268、270、272、274、276、278、280包括两列数据；然而，在其它实施例中，数据群组可以具有多于两列的数据。在一个示例中，每个数据群组248、250、252、254、256、258、260、262、266、268、270、272、274、276、278、280包括两列或更多列的评注288。在另一示例中，一列或多列评注288中的评注是对应数据项286的属性。

[0035] 关于网站120，网页196A的菜单202与网页196B的菜单202等同。作为结果，对应的数据群组248、256也是等同的。此外，由于网站120的侧边栏204和页脚208是等同的，所以结果得到的数据群组250、258以及数据群组252、260也是等同的。内容212A不同于内容212B；因而，数据群组254不同于数据群组262，如通过不同的数据项286所示的。网站128的数据库表244是类似的。特别地，由于每个菜单222是等同的，每个侧边栏226是等同的，并且每个页脚230是等同的，所以数据群组266、274是等同的，数据群组268、276以及数据群组270、278是等同的。内容234C不同于内容234D；因而，数据群组272不同于数据群组280，如通过不同的数据项286所示的。

[0036] 利用内部冗余途径来处理数据库表

接下来，如框408中所示，方法400包括通过使用内部冗余途径来处理数据库表240、244，以便标识网站120、128的潜在有信息量内容。换言之，方法400包括处理数据库表240、244来滤出网站120、128的非有信息量内容。该过程被称为内部冗余途径，因为：为每个数据群组、如相比于对应数据库表240、244的仅仅其它数据群组来确定冗余。例如，数据库表240的每个数据群组248、250、252、254、256、258、260、262与数据库表240的每个其它数据群组248、250、252、254、256、258、260、262相比，以确定内部冗余。如果数据库表240包括冗余数据群组(即重复的数据群组)，那么那些数据群组被视为内部冗余的。如果数据库表240包括不重复(即唯一或大体上唯一的)的数据群组248、250、252、254、256、258、260、262，那么该数据群组248、250、252、254、256、258、260、262不是内部冗余的。因此，内部冗余处理是用于从网站120、128过滤非有信息量内容。

[0037] 参考图5，处理器180处理数据群组248、250、252、254、256、258、260、262来确定数据库表240内的每个数据群组出现的数目。具有大于或等于预定出现数目的出现数目的任何数据群组248、250、252、254、256、258、260、262被认为是内部冗余的。具有小于预定出现数目的出现数目的任何数据群组248、250、252、254、256、258、260、262不被认为是内部冗余的并且被认为包含潜在有信息量内容。在一个示例中，预定出现数目是二。在其它实施例中，预定出现数目是三到五。

[0038] 此处的概念是如果数据跨特定网站的网页重复地出现，那么重复的数据很可能来自模板，并且因此不是有信息量内容。作为示例，考虑跨网站的每个网页的顶部出现的典型菜单栏(即菜单202、222)。菜单栏在网站的每一个网页上，并且在每个网页上典型地是等同的。在对网站进行导航之外，菜单栏中的数据典型地是无兴趣的或非有信息量的。然而，如果数据出现仅仅一次(或小于预定出现数目)，于是该数据是潜在有用、重要和/或引起兴趣的，并且是潜在有信息量内容。对此的示例是内容212A、212B、234C、234D。

[0039] 为了确定内部冗余，处理器180处理数据库表240、244，并且标识具有少于预定出现数目的任何数据群组248、250、252、254、256、258、260、262、266、268、270、272、274、276、278、280，以用于通过处理器180进一步处理。内部冗余(即具有大于或等于预定出现数目)的数据群组248、250、252、254、256、258、260、262、266、268、270、272、274、276、278、280不被进一步处理，并且不被添加到知识库140。对数据群组248、250、252、254、256、258、260、262、266、268、270、272、274、276、278、280的该过滤是在使用仅仅DEPTA的IE系统之上的改进。因此，之间冗余处理用于标识或确定有信息量内容。

[0040] 例如，预定出现数目是二，并且处理器180处理数据库表240的数据群组248、250、252、254、256、258、260、262。处理器180确定数据群组248、250、252、256、258、260各自出现两次。因而，数据群组248、250、252、256、258、260是内部冗余的，并且不被进一步处理或被添加到知识库140。因而，在内部冗余处理步骤中，方法400将内部冗余数据排除在进一步处理之外。处理器180还确定了数据群组254、252各自出现仅仅一次。因而，数据群组254、262不是内部冗余的。数据群组254、262适合用于进一步的处理并且包含潜在有信息量内容。内容是潜在有信息量的，因为方法400在确定内容是否是有信息量内容中应用另一水平的冗余处理。

[0041] 利用之间冗余途径来处理数据库表

接下来，如在图4的框412中所示，处理器180使用之间冗余途径来确定非内部冗余数据群组254、262、272、280(即包括潜在有信息量内容的数据群组)中的任何数据群组在相比于其它网站的时候是否是冗余的。该过程被称为之间冗余途径，因为跨至少两个网站来确定冗余。此处的概念是如果数据群组中的信息是有信息量的，那么信息很可能出现在其它网站中，因为它是引起兴趣的、重要的和/或有信息量的。例如，当考虑机场网站的时候，出租车服务的电话号码很可能出现在网页之一上，作为潜在有信息量内容。当扩充到机场网站的语料库的时候，电话号码还很可能出现在每一个网站上，因为出租车服务对于每个机场的顾客而言是重要的。根据本文中所阐明的方法400，出租车服务的电话号码典型地不是内部冗余的并且典型地是之间冗余的，并且因此典型地是有信息量内容。作为另一示例，考虑具有描述公司历史的网页的公司网站。该信息很可能不是内部冗余的，并且还很可能不是之间冗余的。所述信息因此是潜在有信息量内容，但是处理器180确定所述信息不是有信息

量内容，因为所述信息仅仅出现在公司网站上并且没有出现在任何其它网站上。因而，当处理从网站所提取的数据的时候，方法400搜索非内部冗余数据并且是之间冗余的数据，用于添加到知识库140。

[0042] 在一个实施例中，处理器180确定：如果具有潜在有信息量内容的即时数据群组或充分类似于该即时数据群组的数据群组被包括在至少预定数目的数据库表中，则该即时数据群组包括有信息量内容。数据库表的预定数目大于一。数据库表的典型预定数目是二。在其它实施例中，数据库表的预定数目是从三到六的整数。

[0043] 处理器180通过如下操作来确定具有潜在有信息量内容的数据群组254、262、272、280是否出现在多于一个数据库表240、244中：将评分指派给数据群组254、262、272、280的数据项286，并然后以使得处理器180能够确定数据群组254、262、272、282的相似性的方式来比较结果得到的评分向量。如图7中所示，基于来自网站120的内容212A的数据群组254与基于来自网站128的内容234C的数据群组272被比较。评注列不被包括在图7中。处理器180已经修正了数据群组254、272，用以包括两个附加的数据列，包括频率列(FREQ，频率)和评分列(SCORE，评分)。频率列指示某些词语或术语(例如中心名词)出现在对应的数据项286中的频率(即次数)。例如，在数据群组254的第一数据项286中，中心名词“出租车”是最频繁使用的中心名词。中心名词“出租车”在与数据群组254的第一数据项286相关联的文本中出现十次。在一个实施例中，评分列基于频率列。特别地，评分列使对应的数据群组254、272中最频繁使用的中心名词(或另一所选词语、术语、短语或句子)的频率加倍。数据项286的其它中心名词的评分等于评率。因而，在一个示例中，对于数据群组254，第一数据项286包括中心名词“出租车”十次，使得“出租车”是在数据群组254中最频繁出现的中心名词。因此，与第一数据项286相关联的评分从十被加倍到二十。对于使最频繁引用的中心名词的频率计数加倍的示例性原因是：在确定数据群组254、272的主题中，中心名词典型地比其它词语更重要。在其它实施例中，处理器180可以对频率做出任何其它所期望的调整以便获得评分。例如，评分列可以等于频率列。相同的过程被用于将评分指派给数据群组272的数据项286。

[0044] 接下来，处理器180基于评分来为具有潜在有信息量内容的每个数据群组254、262、272、280形成归一化向量290。如图7中所示，归一化向量290包括针对在将被比较的数据群组254、272二者中所使用的数据项286中的每一个的评分条目。由于数据群组254不包括数据项286“停车”，所以“零”评分已经被添加到向量A以使向量归一化。针对数据群组254的经归一化的评分向量290是[20,7,8,5,6,0,0]，并且针对数据群组272的经归一化的评分向量290是[16,7,5,0,0,4,3]。

[0045] 在使评分向量归一化之后，处理器180比较经归一化的评分向量290来确定数据群组254、272的内容是否包含类似的内容。在一个实施例中，这通过如下操作来进行：比较两个经归一化的向量290的余弦相似性与预定阈值相似性水平。向量290的余弦相似性通过等式(1)被给出，并且等于向量290的点积除以向量290的范数的乘积。

[0046] 余弦相似性 $= \dfrac{\vec{a} \cdot \vec{b}}{\|\vec{a}\| \|\vec{b}\|}$ (1)

如果所述余弦相似性小于阈值相似性水平，那么在所比较的数据群组254、272中的数

据不相似，并且不出现在网站120、128二者上。如果所述余弦相似性大于或等于相似性阈值，那么在所比较的数据群组254、272中的数据相似，并且出现在网站120、128二者上。具有大于或等于相似性阈值的余弦相似性的数据群组是之间冗余的。关于经归一化的向量A和B(向量290)，点积等于409 (20*16+7*7+8*5+5*0+6*0+0*4+0*3)，向量A的范数是24.0(sqrt($20^2+7^2+8^2+5^2+6^2$))，并且向量B的范数是18.8(sqrt($16^2+7^2+5^2+4^2+3^2$))。因而，经归一化的向量A和B的余弦相似性是0.91(409/(24.0*18.8)。

[0047] 余弦相似性范围从0到1。如果余弦相似性是0，则两个向量不相似，并且如果余弦相似性是1，则两个向量是等同的。处理器180比较所计算的余弦相似性与预定的阈值相似性水平。示例性的预定阈值相似性水平是0.8。如果处理器180确定了两个经归一化的向量具有大于或等于0.8的余弦相似性，那么处理器180确定两个评分向量对应于之间冗余的数据群组。也就是说，处理器180确定：相同的信息存在于所比较的数据群组二者中，并且数据群组的潜在有信息量内容是有信息量内容。在以上示例中，与数据群组254、272相关联的经归一化的向量290的余弦相似性大于预定阈值相似性水平(0.91＞0.8)，并且处理器180确定了数据群组254、272是之间冗余的并且包括有信息量内容。作为结果，数据群组254、272的内容适合于包括在知识库140中。

[0048] 作为另一示例，考虑在内容212B与内容234C和234D之间的比较。如图8中所示，在内容212B(图2)的数据群组262的数据项286与内容234C(图3)的数据群组278的数据项286之间不存在重叠，因而，经归一化的评分向量的点积是零，并且在向量之间的余弦相似性是零。由于零在预定阈值相似性水平0.8以下，所以处理器180确定内容212B和内容234C与彼此不同并且不是之间冗余的。考虑在内容212B与内容234D之间的比较，仅一个数据项286重叠；即数据项286“机场”。因而，在这两个内容212B、234D之间存在某种相似性。特别地，通过使用上述的评分途径，在与内容212B、234D相关联的经归一化的向量(未被示出)之间的余弦相似性是0.04。因而，尽管内容212B、234D共享数据项286，但是该数据项286不是高评分的数据项286。作为结果，在内容212B、234D之间的余弦相似性在预定阈值相似性水平0.8以下。因此，处理器180确定内容212B仅仅出现在网站120上，并且不出现在网站128上，因为内容212B与内容234C、234D不相似。处理器180此外确定内容212B不针对有信息量内容，并且知识库140不应当基于内容212B被更新。

[0049] 总而言之，处理器180确定内容212A、234C包括有信息量内容，因为对应的数据群组254、272包括之间冗余信息。处理器180确定内容212B不包括有信息量内容，因为对应的数据群组262不包括之间冗余信息。方法400不需要人类干预来标识有信息量内容。

[0050] 利用有信息量内容来更新知识库

接下来，如在图4的框416中所示，方法400包括处理器180将有信息量内容212A、234C添加到知识库140。当附加的有信息量内容212A、234C被保存到知识库140的时候，知识库140甚至更好地被装备以向来自客户端设备、诸如无线设备104的质询提供准确的信息。

[0051] 示例性的测试结果

如在以下插入的表1中所示，IE系统100的评估结果被提供，其比较三个IE系统的精度(“P”)、查全率(“R”)以及f评分(“F”)。第一IE系统是仅仅使用DEPTA的非监督式IE系统。第二个是使用DEPTA和内部冗余处理的非监督式IE系统。第三个是使用DEPTA、内部冗余和之间冗余处理的非监督式IE系统100。

[0052] 对照与美国机场相关联的网站的语料库以及与美国医院相关联的网站的语料库来测试这三个IE系统。针对机场语料库的测试聚焦于同停车场、汽车租赁、穿梭车、公交车和出租车相关联的信息。针对医院语料库的测试聚焦于服务区域、疾病、症状、电话号码和工作小时数。

	机场语料库						医院语料库					
	确切匹配			部分匹配			确切匹配			部分匹配		
系统	P	R	F	P	R	F	P	R	F	P	R	F
DEPTA	0.12	**0.65**	0.20	0.15	**0.80**	0.25	0.19	**0.79**	0.31	0.43	**0.92**	0.58
+内部-R	0.29	0.65	0.40	0.37	0.80	0.51	0.52	0.68	0.59	0.71	0.86	0.77
+内部-R 和之间-R	**0.42**	0.61	**0.50**	**0.52**	0.74	**0.61**	**0.67**	0.60	**0.63**	**0.86**	0.80	**0.83**

[0053] 表1

通过查询由上述三个IE系统所形成的知识库来确定测试结果。如在本文中以及在下述等式中所使用的，“系统评注”是根据这三个IE系统之一而从网站所提取的评注。知识库140例如包括多个系统评注。而且，如本文中所使用的，“地面实况评注”是由技术员或工程师为了生成以上在表1中所呈现的测试结果的目的而从网站手动提取的评注。

[0054] 如典型地在IE上下文中所进行的，为部分匹配和确切匹配生成测试结果。部分匹配是指利用搜索项来查询知识库，并且使知识库返回仅仅部分地匹配搜索项的结果。例如，查询可以包括项“出租车提供商”，并且知识库可以返回评注，所述评注包括“出租车”、“提供商”、“出租车服务”和“服务提供商”，其中的每一个都是部分匹配，因为评注包括搜索项中的至少一个。确切匹配是指利用与系统评注确切匹配的搜索项来查询知识库。例如，查询可以包括项“出租车提供商”，并且知识库可以返回评注“出租车提供商”，其是确切的匹配。

[0055] 根据以下呈现的等式(2)和(3)来计算部分匹配列。等式(2)通过将部分匹配的系统评注的数目除以系统评注的数目来计算部分匹配的精度。

$$P = \frac{\#(\text{部分匹配的系统评注})}{\#(\text{系统评注})} \quad (2)$$

[0056] 等式(3)确定部分匹配的查全率。查全率对系统所发现的地面实况评注的百分比进行度量。等式(3)通过将部分匹配的地面实况评注的数目除以地面实况评注的数目来确定查全率。

$$R = \frac{\#(\text{部分匹配的地面实况评注})}{\#(\text{地面实况评注})} \quad (3)$$

[0057] 根据以下呈现的等式(4)和(5)来计算确切匹配列。等式(4)通过将确切匹配的系统评注的数目除以系统评注的数目来确定精度。

$$P = \frac{\#(\text{精确匹配的系统评注})}{\#(\text{系统评注})} \quad (4)$$

[0058] 等式(5)通过将确切匹配的地面实况评注的数目除以地面实况评注的数目来确定查全率。

$$R = \frac{\#(\quad \text{确切匹配的关键评注} \quad)}{\#(\quad \text{地面实况评注} \quad)} \tag{5}$$

[0059] F评分是精度和查全率的调和平均值。通过使用如以下所呈现的等式(6)来计算F评分。

$$\text{F评分} = 2 * \frac{\text{精度} \quad * \text{查全率}}{\text{精度} \quad + \text{查全率}} \tag{6}$$

[0060] 在表1中,更高的数对应于更好执行的IE系统。表1示出了,对于机场语料库和医院语料库二者,非监督式IE系统100胜过仅仅使用DEPTA的IE系统以及使用DEPTA和仅仅内部冗余处理的IE系统。比其它IE系统的f评分更高的系统100的f评分使得这是明显的。详细地,如相比与其它两个IE系统,IE系统100为部分匹配和确切匹配二者生成精度方面的巨大增益。如相比于其它两个IE系统,IE系统100仅仅遭受查全率方面的微小损失。在精度方面的巨大增益弥补查全率方面的小损失,并且使得IE系统100的f评分远远超过其它IE系统的f评分。

[0061] 在图9A直到9F中为不同大小的语料库绘制IE系统100的精度、查全率和性能。如图9A和9D中所示,IE系统100的精度相对于语料库的大小是大体上恒定的。也就是说,将附加的网站添加到语料库不会大幅改善IE系统100的精度。如图9B和9E中所示,IE系统100的查全率一般响应于增大语料库的大小而增大,如所预期的那样。如图9C和9F中所示,IE系统100的f评分也一般响应于增大语料库的大小而增大。

[0062] 虽然已经在附图和前述描述中详细图示和描述了本公开内容,但是本公开内容在特性方面应当被视为是说明性的而不是限制性的。理解到,已经呈现了仅仅优选的实施例,并且期望保护在本公开内容的精神内的所有改变、修改和另外的应用。

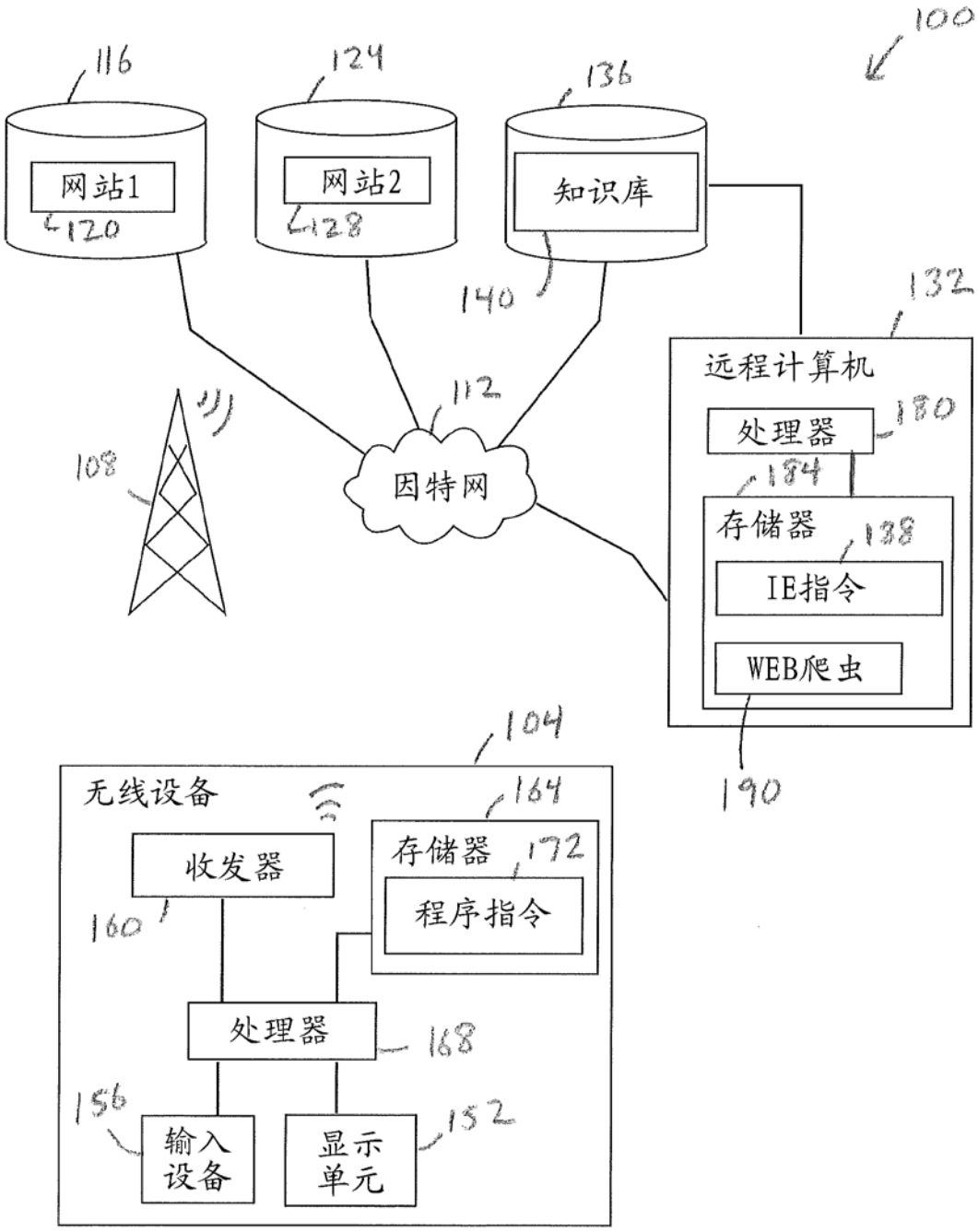

100
116
124
136
网站1
120
网站2
128
知识库
140
132
远程计算机
处理器
180
184
存储器
138
IE指令
WEB爬虫
190
112
因特网
108
104
无线设备
164
存储器
172
程序指令
收发器
160
处理器
168
156
输入设备
显示单元
152

图 1

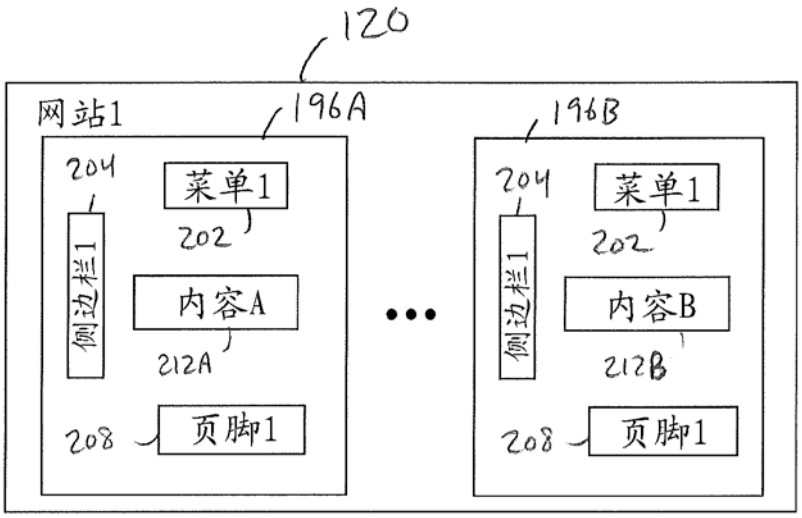

120
网站1
196A
196B
204
菜单1
202
侧边栏1
内容A
212A
208
页脚1
204
菜单1
202
侧边栏1
内容B
212B
208
页脚1

图 2

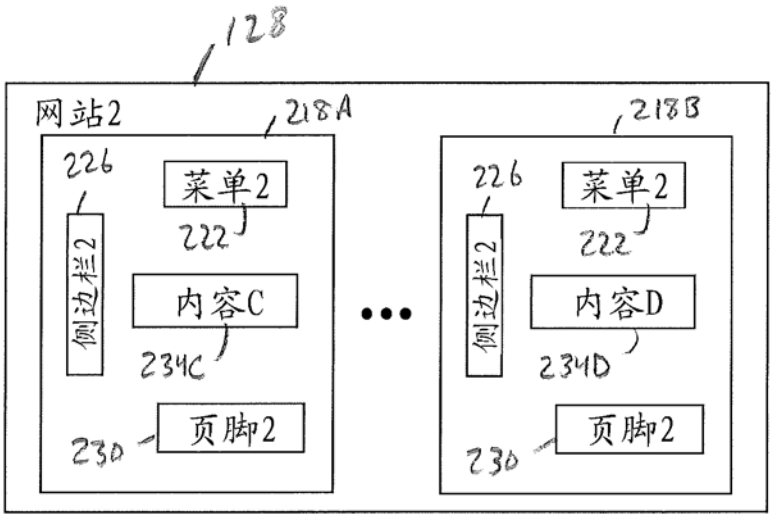

128
网站2
218A
218B
226
菜单2
222
侧边栏2
内容C
234C
230
页脚2
226
菜单2
222
侧边栏2
内容D
234D
230
页脚2

图 3

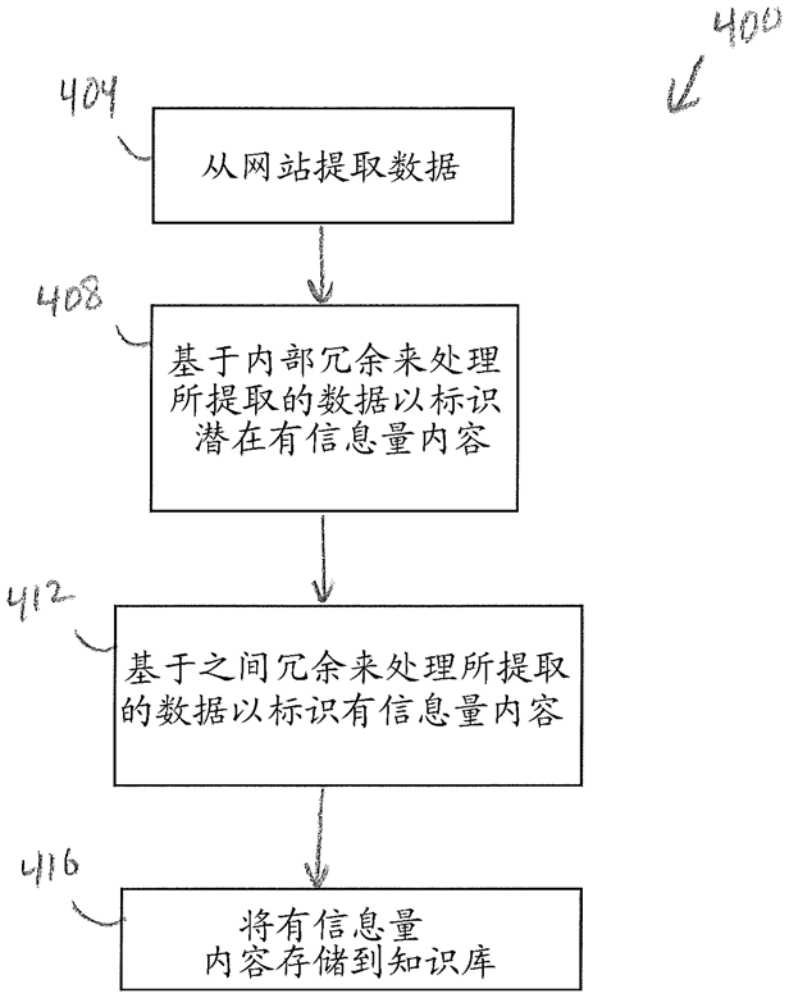

400
404
从网站提取数据
408
基于内部冗余来处理所提取的数据以标识潜在有信息量内容
412
基于之间冗余来处理所提取的数据以标识有信息量内容
416
将有信息量内容存储到知识库

图 4

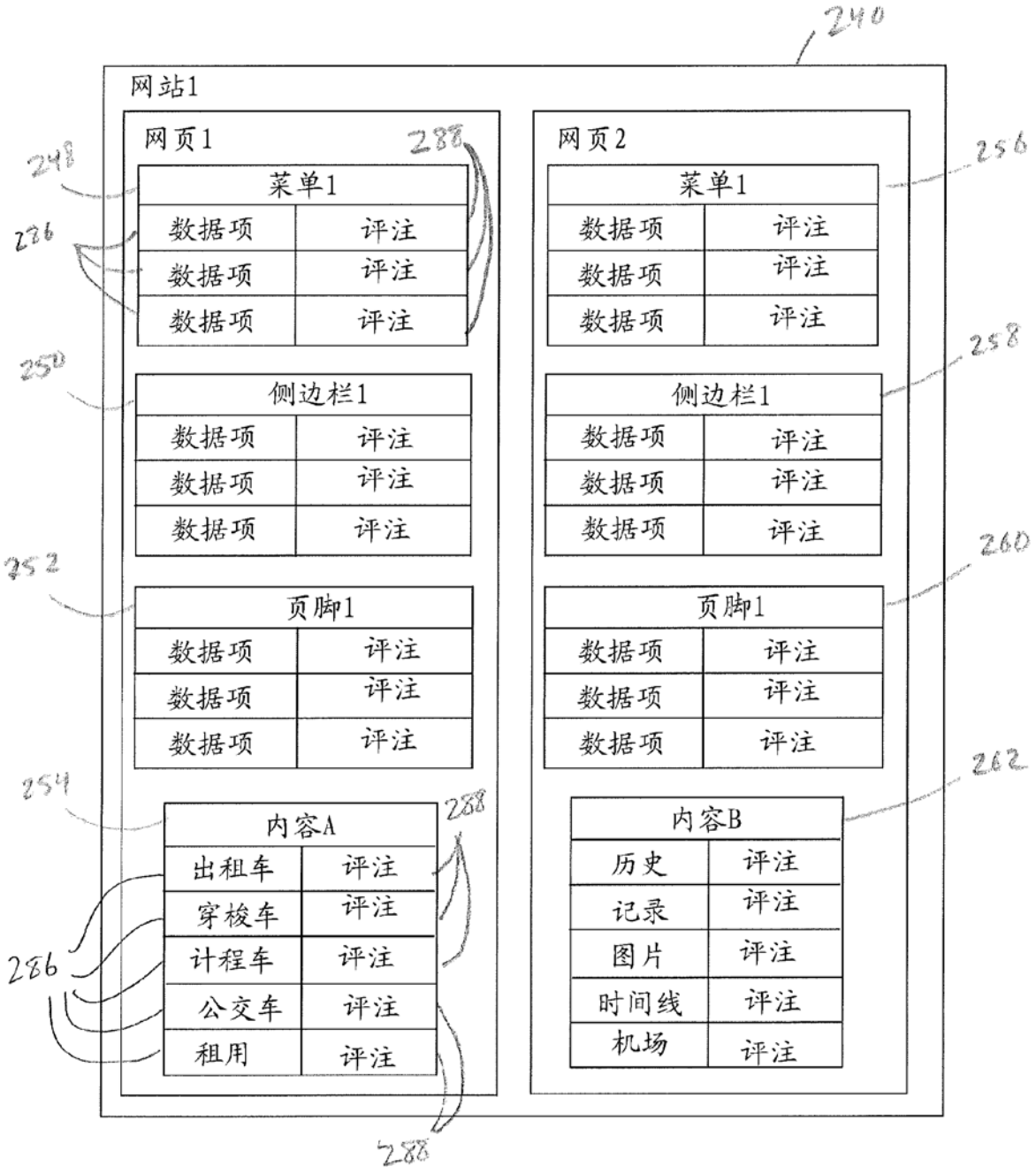

240
网站1
网页1
288
248
菜单1
数据项 评注
数据项 评注
数据项 评注
246
250
侧边栏1
数据项 评注
数据项 评注
数据项 评注
252
页脚1
数据项 评注
数据项 评注
数据项 评注
254
内容A
出租车 评注
穿梭车 评注
计程车 评注
公交车 评注
租用 评注
286
288
网页2
256
菜单1
数据项 评注
数据项 评注
数据项 评注
258
侧边栏1
数据项 评注
数据项 评注
数据项 评注
260
页脚1
数据项 评注
数据项 评注
数据项 评注
262
内容B
历史 评注
记录 评注
图片 评注
时间线 评注
机场 评注

图 5

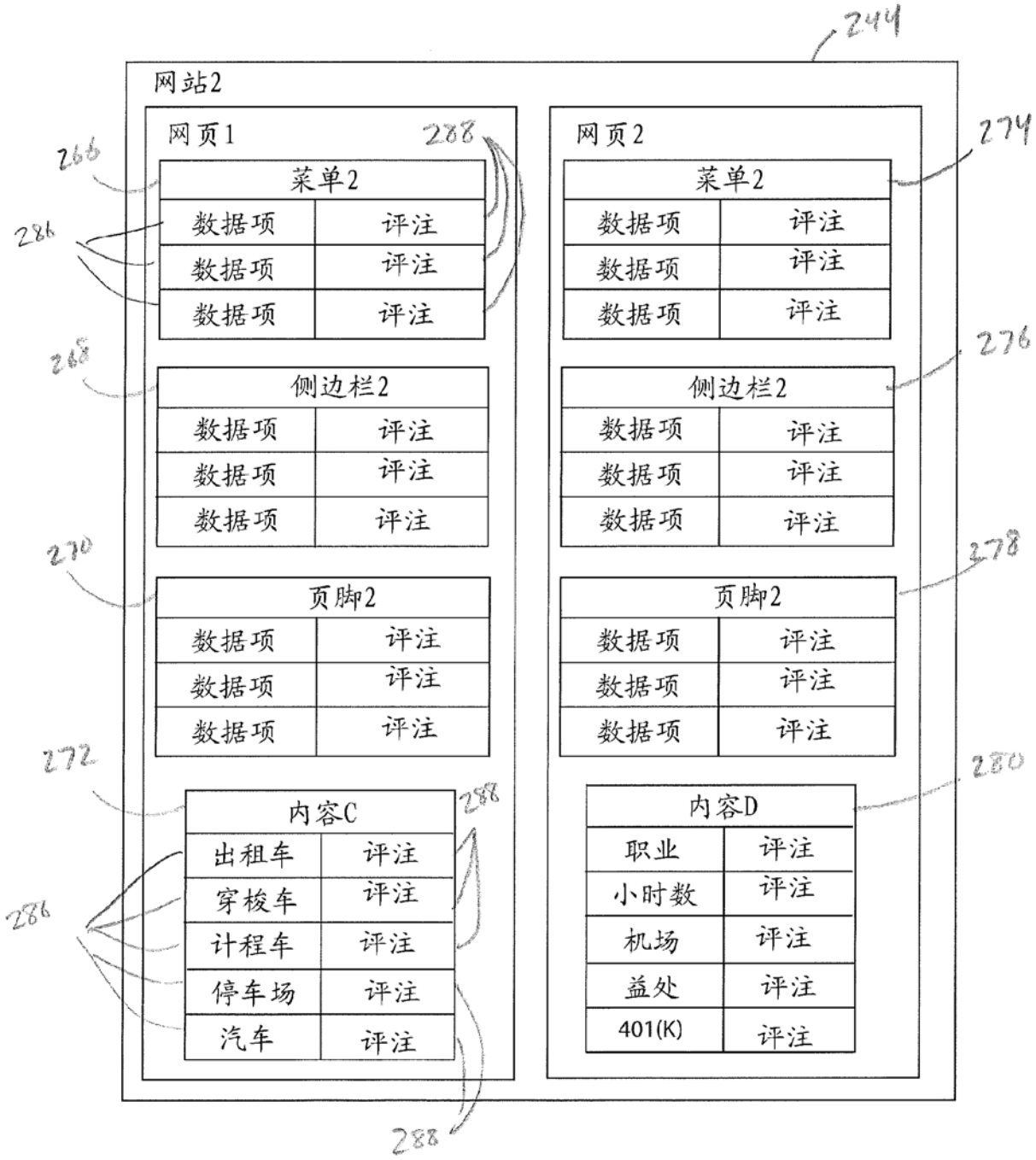

244
网站2
网页1
288
266
286
菜单2
数据项 评注
数据项 评注
数据项 评注
268
侧边栏2
数据项 评注
数据项 评注
数据项 评注
270
页脚2
数据项 评注
数据项 评注
数据项 评注
272
内容C
出租车 评注
穿梭车 评注
计程车 评注
停车场 评注
汽车 评注
286
288
288
网页2
274
菜单2
数据项 评注
数据项 评注
数据项 评注
276
侧边栏2
数据项 评注
数据项 评注
数据项 评注
278
页脚2
数据项 评注
数据项 评注
数据项 评注
280
内容D
职业 评注
小时数 评注
机场 评注
益处 评注
401(K) 评注

图 6

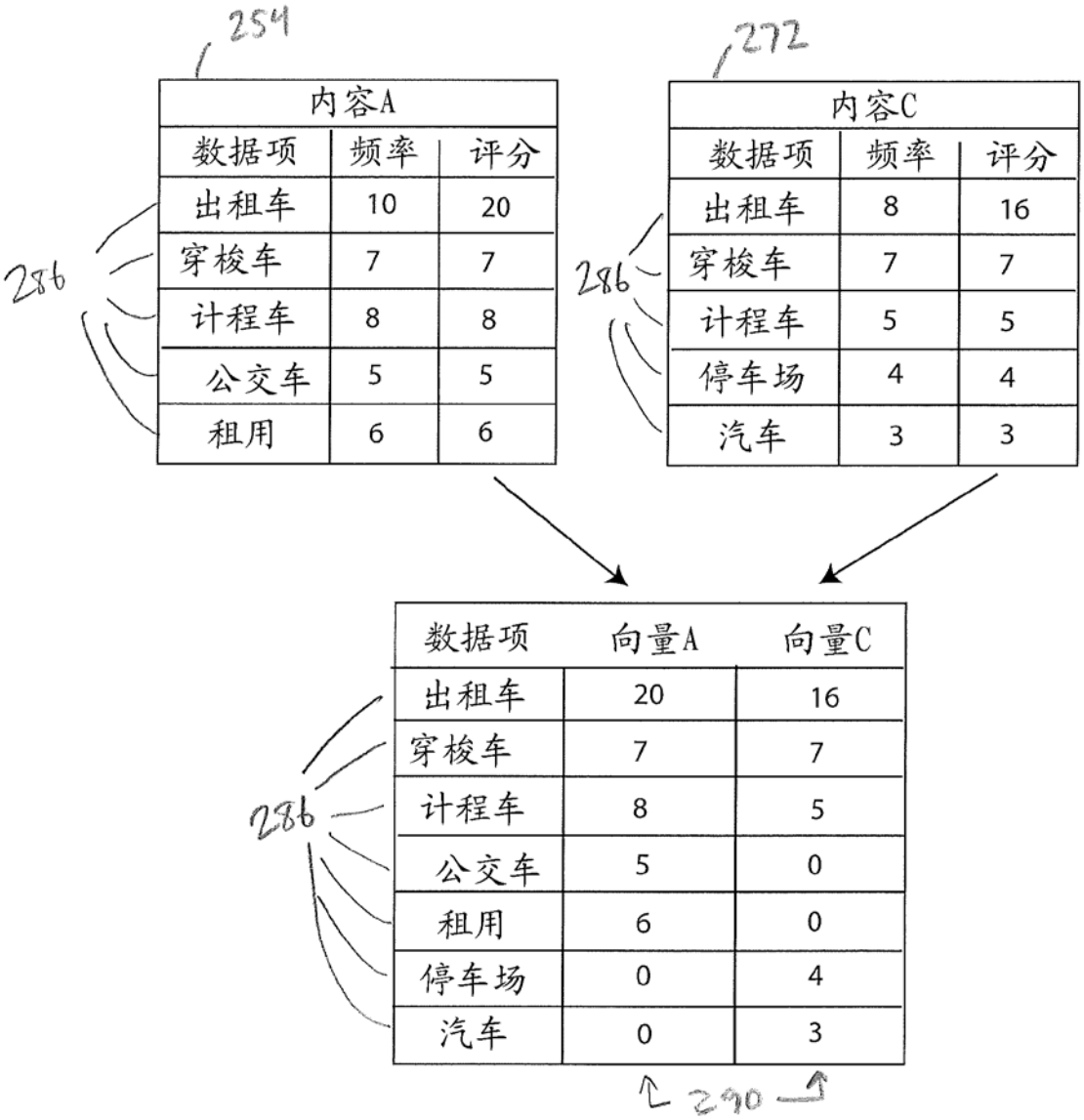
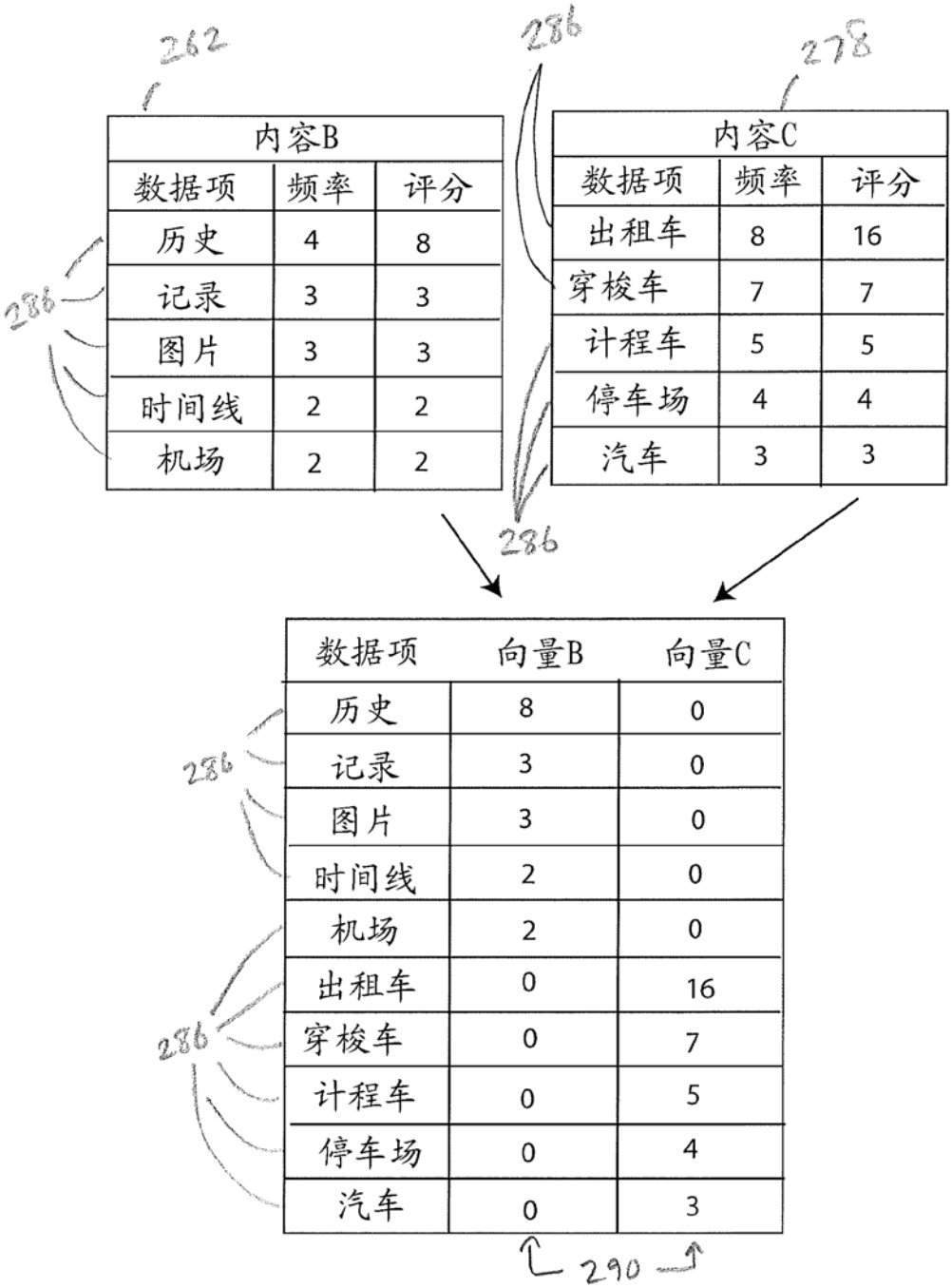

254
内容A
数据项 频率 评分
出租车 10 20
穿梭车 7 7
计程车 8 8
公交车 5 5
租用 6 6
286
272
内容C
数据项 频率 评分
出租车 8 16
穿梭车 7 7
计程车 5 5
停车场 4 4
汽车 3 3
286
数据项 向量A 向量C
出租车 20 16
穿梭车 7 7
计程车 8 5
公交车 5 0
租用 6 0
停车场 0 4
汽车 0 3
286
290

图 7



内容B

数据项	频率	评分
历史	4	8
记录	3	3
图片	3	3
时间线	2	2
机场	2	2

内容C

数据项	频率	评分
出租车	8	16
穿梭车	7	7
计程车	5	5
停车场	4	4
汽车	3	3

数据项	向量B	向量C
历史	8	0
记录	3	0
图片	3	0
时间线	2	0
机场	2	0
出租车	0	16
穿梭车	0	7
计程车	0	5
停车场	0	4
汽车	0	3

图 8

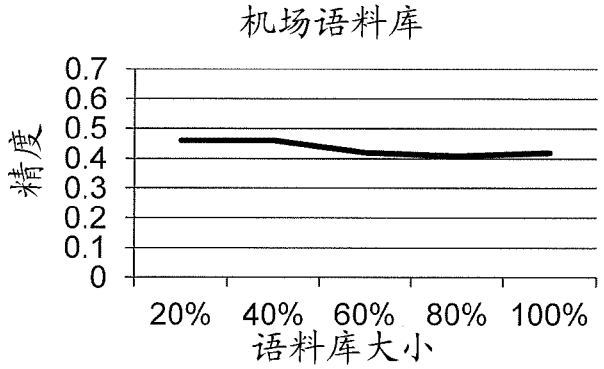

机场语料库
精度
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
20% 40% 60% 80% 100%
语料库大小

图 9A

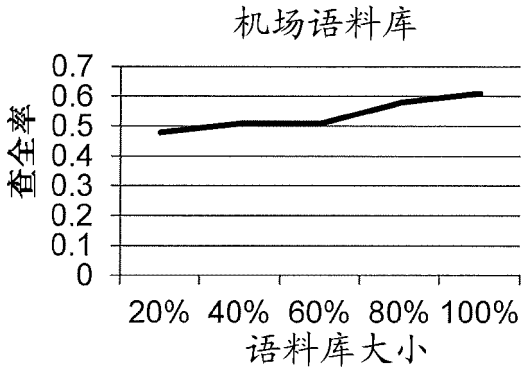

机场语料库
查全率
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
20% 40% 60% 80% 100%
语料库大小

图 9B

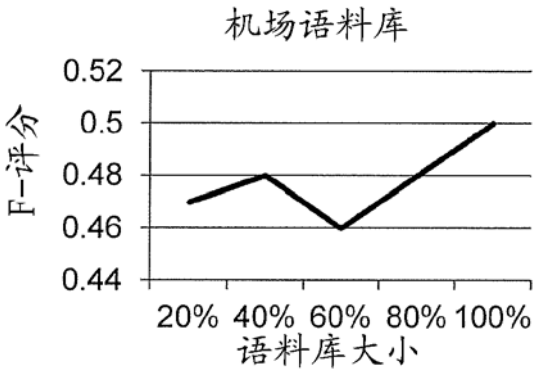

机场语料库
F-评分
0.52
0.5
0.48
0.46
0.44
20% 40% 60% 80% 100%
语料库大小

图 9C

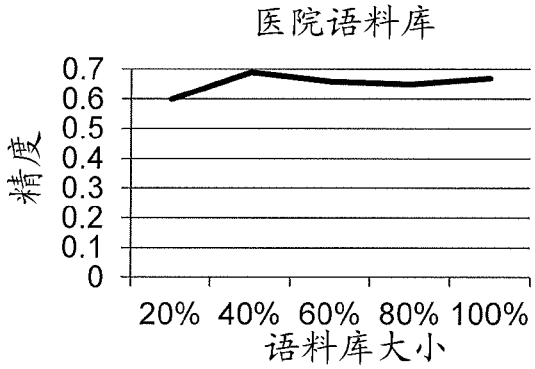

医院语料库
精度
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
20% 40% 60% 80% 100%
语料库大小

图 9D

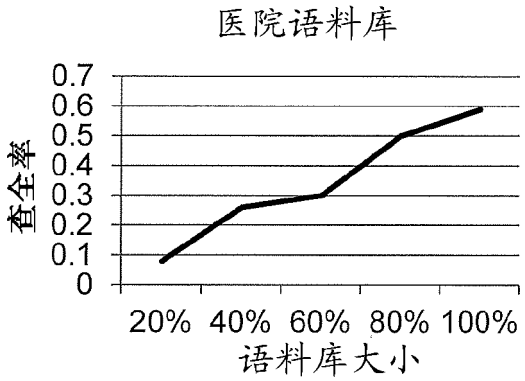

医院语料库
查全率
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
20% 40% 60% 80% 100%
语料库大小

图 9E

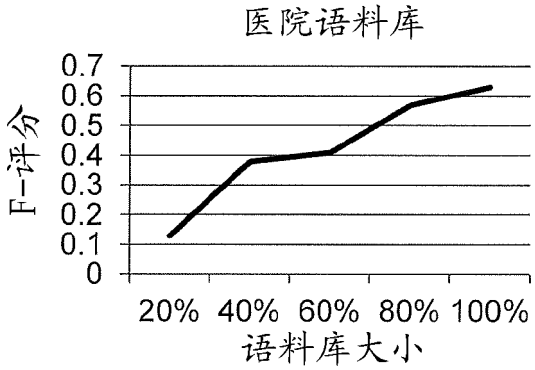

医院语料库
F-评分
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
20% 40% 60% 80% 100%
语料库大小

图 9F