(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 102760153 A
(43)申请公布日 2012.10.31

(21)申请号 201210130002.3

(22)申请日 2012.04.20

(30)优先权数据
13/091405 2011.04.21 US

(71)申请人 帕洛阿尔托研究中心公司
地址 美国加利福尼亚州

(72)发明人 J·方 B·陈

(74)专利代理机构 中国专利代理(香港)有限公司 72001
代理人 方世栋 王忠忠

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

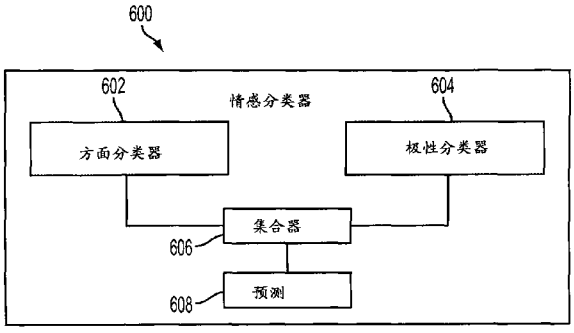
权利要求书 2 页 说明书 12 页 附图 8 页

(54)发明名称

将词典知识合并入SVM学习以改进情感分类

(57)摘要

用于内容的情感分类的情感分类器。方面分类器被配置为将内容分类为与信息的特定方面相关,所述方面分类器结合所述领域特定情感词典的至少一部分。极性分类器随后被配置为将由所述方面分类器分类的所述内容分类为具有下列之一:信息的所述特定方面的正面的情感、信息的所述特定方面的负面的情感,或者将由所述方面分类器分类的所述内容分类为不具有关于信息的所述特定方面的情感。所述极性分类器也结合所述领域特定情感词典的至少一部分。



CN 102760153 A

1. 一种用于内容的情感分类的情感分类器，包括：

方面分类器，所述方面分类器被配置为将内容分类为与信息的特定方面相关，所述方面分类器结合领域特定情感词典的至少一部分；以及

极性分类器，所述极性分类器被配置为将由所述方面分类器分类的所述内容分类为具有下列之一：信息的所述特定方面的正面的情感、信息的所述特定方面的负面的情感，或者将由所述方面分类器分类的所述内容分类为不具有信息的所述特定方面的情感，所述极性分类器结合所述领域特定情感词典的至少一部分。

2. 如权利要求 1 所述的情感分类器，其中所述方面分类器进一步结合通用情感词典。

3. 如权利要求 1 所述的情感分类器，其中所述极性分类器进一步结合通用情感词典。

4. 如权利要求 1 所述的情感分类器，所述情感分类器被实现为支持向量机。

5. 如权利要求 1 所述的情感分类器，其中所述内容是以句子的形式被配置的文本，所述句子具有单词和 / 或短语。

6. 如权利要求 5 所述的情感分类器，进一步被配置为在所述句子级别将所述内容分类为关于信息的所述特定方面是正面的、负面的或者不具有情感。

7. 如权利要求 6 所述的情感分类器，进一步被配置为在所述句子级别预测与所述情感相关联的主要主题。

8. 如权利要求 1 所述的情感分类器，其中所述领域特定情感词典由下列配置：(i) 来自已被过滤的注释的语料库的领域特定单词和 / 或短语，(ii) 通过使用预定的语言模式搜索万维网并过滤所返回的搜索结果而获得的领域特定单词和 / 或短语，以及 (iii) 通过字典扩展技术而获得的领域特定单词和 / 或短语。

9. 如权利要求 8 所述的情感分类器，其中所述领域特定词典包括将所述领域特定单词和 / 或短语分类为与信息的所述方面中的一个相关联的子词典，和将情感关联到信息的所述分类的一个方面的另一子词典。

10. 一种执行内容的情感分类的方法，包括：

通过方面分类器将内容分类为与信息的特定方面相关，其中所述方面分类器结合领域特定情感词典的至少一部分；以及

通过使用极性分类器将由所述方面分类器分类的所述内容分类为具有下列之一：信息的所述特定方面的正面的情感、信息的所述特定方面的负面的情感，或者将由所述方面分类器分类的所述内容分类为不具有信息的所述特定方面的情感，其中所述极性分类器结合所述领域特定情感词典的至少一部分。

11. 如权利要求 10 所述的方法，其中所述方面分类器进一步结合通用情感词典。

12. 如权利要求 10 所述的方法，其中所述极性分类器进一步结合通用情感词典。

13. 如权利要求 10 所述的方法，其中所述方面分类器和所述极性分类器被集合在一起以形成被实现为支持向量机的情感分类器。

14. 如权利要求 10 所述的方法，其中所述内容是以句子的形式而被配置的文本，所述句子具有单词和短语。

15. 如权利要求 14 所述的方法，进一步包括在所述句子级别将所述内容分类为是信息的所述特定方面的正面的情感、负面的情感或不具有信息的所述特定方面的情感。

16. 如权利要求 15 所述的方法，进一步包括在所述句子级别预测与所述情感相关联的

主要主题。

17. 如权利要求 10 所述的方法，其中所述领域特定情感词典由下列配置 :(i) 通过过滤注释的语料库获得领域特定单词和 / 或短语，(ii) 通过使用预定语言模式经由因特网搜索万维网并且过滤返回的搜索结果而获得领域特定单词和 / 或短语，以及 (iii) 在通过 (i) 和 (ii) 获得的领域特定单词和 / 或短语上执行字典扩展操作。

18. 如权利要求 17 所述的方法，其中已从所述注释的语料库被过滤的单词和 / 或短语已经通过下列而被过滤 :

识别信息的方面 ;

从所述注释的语料库提取被标记为表达信息的所述识别的方面的句子中的单词和 / 或短语 ;

从所述提取的单词和 / 或短语形成对应于信息的所述识别的方面的词典的初始列表 ;

对照于针对信息的其它方面的词典的其它初始列表，检查来自针对信息的所述识别的方面的词典的所述初始列表的单词和 / 或短语 ;

过滤出匹配来自针对信息的其它方面的词典的任何其它初始列表的单词和 / 或短语的任何来自针对信息的所述方面的词典的所述初始列表的单词和 / 或短语 ;以及

针对信息的所述方面产生过滤的词典的列表，表示领域特定词典。

19. 如权利要求 17 所述的方法，其中通过搜索万维网或因特网而获得的内容通过下列而被获得 :

产生对应于选择的信息的方面的语言模式 ;

选择所述产生的语言模式中的一个 ;

通过搜索引擎将所述选择的语言模式发送到万维网或因特网以获得基于所述选择的语言模式的搜索结果 ;

从所述从所述搜索返回的结果提取单词和 / 或短语 ;

通过下列之一过滤出噪声单词和 / 或短语 :(i) 去除预定的非用单词和 / 或短语或 (ii) 使用候选内容 (单词) 中的每个作为新的种子内容 (单词) 并且重新进行搜索，并且当所述原始的种子内容 (单词) 被重新调整时保持所述内容 (单词)，否则过滤出所述内容 (单词) ;以及

添加保留到所述领域特定词典的单词和 / 或短语。

20. 如权利要求 17 所述的方法，其中所述字典扩展包括 :发现针对所述列表中的单词和 / 或短语的同义词和反义词中的至少一个，并且将所述同义词和反义词中的至少一个添加到所述列表。

将词典知识合并入 SVM 学习以改进情感分类

技术领域

[0001] 本申请目的在于自动的分类，并且更特别地目的在于自动的情感分类，其中情感分类被理解为是特定类型的文本分类，其用作分类信息（诸如以文本的形式）的意见或情感，当其涉及特定的论题或主题时。

背景技术

[0002] 两种典型的用于情感分析的方法是词典查找和机器学习。词典查找方法通常从正面的和负面的单词的词典开始。例如，“漂亮的”被确认为正面的单词并且“丑陋的”被确认为负面的单词。文本的总的情感由一组单词的情感和在所述文本中出现的表达确定。

[0003] 综合性的情感词典可以提供简单然而有效的用于情感分析的解决方案，因为其是普通的并且不需要预先的训练。因此，已经花费关注和努力用于构建这样的词典。然而，对该方法的重大的挑战是：许多单词的极性依赖于领域和上下文。例如，“长”在“长的电池寿命”中是正面的并且在“长的快门迟滞”中是负面的。当前的情感词典不捕获情感表达的这样的领域和上下文敏感性。它们排除这样的领域和上下文依赖的情感表达或者基于从某个语料库 (corpus)（诸如通过因特网而被访问的万维网）收集的统计资料而用总的极性趋势标记它们。虽然排除这样的表达导致差的覆盖范围，用极性趋势简单地标记它们导致差的精度。

[0004] 由于这些限制，机器学习方法已经正在情感分析的领域中得到日益增加的普及。诸如使用支持向量机 (SVM) 的那些机器学习方法不依靠情感词典以确定单词和表达的极性，并且可以自动地学习一些在训练数据中示出的上下文相关性。例如，如果“长的电池寿命”和“长的快门迟滞”在所述训练数据中分别被标记为正面的和负面的，学习算法可以学会：当其与短语“电池寿命”相关联时“长”是正面的，而当与短语‘快门迟滞’相关联时其是负面的。

[0005] 然而，这样的方法的成功严重地依赖所述训练数据。对于情感分析的任务，由于自然语言的丰富，数据不足是不能被容易地解决的固有问题。特别地，人们倾向于使用不同的表达来表示相同的情感，并且也倾向于在相同的句子或文件中不重复他们的情感。因此，收集足够表示人们如何对于各种主题表达情感的训练数据是非常困难的。与一些其它文本分类任务相比，该数据不足问题已经导致了对情感分类的相对低的准确度。

[0006] 因此，尽管最近的研究已经显示对于情感分析的任务，机器学习方法通常优于所述词典查找方法，忽视由情感词典提供的优势和知识可能不是最优的。

[0007] 然而，少数研究已经致力于将这两种方法相结合以改进情感分类。一些已经探索使用通用情感词典以改进短语的上下文极性的识别。一些其它最近的研究已显示：将通用情感词典合并到机器学习算法中可以改进在文件级别上的情感分类的准确度。在所有这些工作中，通用情感词典包含具有独立于上下文 / 领域的极性的单词。本情感分类器系统和方法不同于这些以前的方法。

发明内容

[0008] 用于内容的情感分类的情感分类器。方面分类器被配置为将内容分类为与信息的特定方面相关，所述方面分类器合并所述领域特定情感词典的至少一部分。极性分类器随后被配置为将由所述方面分类器分类的内容分类为具有下列之一：信息的特定方面的正面的情感、信息的特定方面的负面的情感，或者将其分类为不具有关于信息的特定方面的情感。所述极性分类器也合并所述领域特定情感词典的至少一部分。

附图说明

[0009] 图 1 示出了具有两个子词典的名称词典，包括主题子词典和情感子词典；

[0010] 图 2 示出了根据本申请的呈现语料库过滤的方法的流程图；

[0011] 图 3 示出了显示用于使用语言模式进行网络 (web) 搜索和过滤的处理的流程图；

[0012] 图 4 是显示由图 3 的方法识别的一些噪声单词的图；

[0013] 图 5 是示出了将极性提供给领域特定词典的单词和 / 或短语的方法的流程图；

[0014] 图 6 提供了示出具有方面分类器和极性分类器的情感分类器的框图；

[0015] 图 7 是描绘了图 6 的所述情感分类器的操作的流程图；

[0016] 图 8 是显示了属于类的点之间的距离的例图；

[0017] 图 9 描绘了在领域特定词典的创建的过程期间本申请的系统；

[0018] 图 10 描绘了在当所述情感分类器正在被训练时的过程期间的本系统的结构；以及

[0019] 图 11 描绘了当所述情感分类器正在工作时本系统的结构。

具体实施方式

[0020] 情感分类系统和方法被公开，其将情感词典作为先验知识与机器学习方法（诸如支持向量机 (SVM)）结合以改进情感分析的准确度。所描述的系统和方法为该学习目的产生领域特定情感词典。所采取的实现上面的概念的实验的结果显示：与通用的领域独立的情感词典相比，被结合进机器学习方法中的领域特定词典导致在所述情感分类过程中的更显著的准确度改进。

[0021] 此处所描述的情感分类系统（在此处也被称为情感分类器或二级情感分类器）和方法提供了信息的方面的精细粒度的情感分析。

[0022] 注意到的是：在本公开中，信息的方面（在此处也被称为信息方面）是普通术语，其在其它使用中包括产品的方面（例如，产品方面 -- 诸如照相机的方面），主题（例如，主题方面 -- 诸如天气），等等。以包含单词和 / 或短语的文本（即，内容）的形式提供所述信息。

[0023] 在下文中，为了解释的目的，所述情感分类任务主要针对分类照相机评论。即，对于照相机评论中的每个句子，所公开的情感分类器被配置为预测在研究中的句子是否讨论任何照相机方面（例如，所述照相机的电池寿命；由照相机拍摄的图片的质量，等等），并且如果该句子讨论被考虑的照相机方面，则所述情感分类器识别相关联的情感（例如，是意见正面的或负面的）。此处所描述的实验结果显示：通过结合由本方法产生的领域特定情感词典，所述情感分类任务的准确度被显著地改进。

[0024] 如所讨论的，仅少数研究已致力于合并词典查找和机器学习方法以改进情感分类。不像之前的工作（在其中仅通用情感词典被使用），本情感分类器不仅将通用情感词典而且将领域特定情感词典结合到所述学习机（例如，SVM 学习）中，以改进情感分类的准确度。所述领域特定情感词典包括指示各种主题或领域的词典以及由具有与特定的主题或领域相关联的极性的单词或短语组成的词典。

[0025] 例如，在被进行的实验中，关于“电池寿命”建立领域特定词典，其包括诸如“电池”的单词的第一词典和诸如“快速地 :负面的”和“长的 :正面的”的单词或短语的第二词典。所述第一词典由对“照相机电池寿命”的主题而言是好的指示符的单词或短语组成，而所述第二词典由具有对“电池寿命”的主题而言的特定的极性的单词或短语组成。例如，“快速地”和“长的”可能在不同的领域中不携带负面的和正面的情感。如果所述领域是不同的，则它们也可以携带相反的情感。更重要地，所述实验结果显示 :虽然通用情感词典仅提供了较小的准确度改进，结合领域特定词典（字典）导致了对所述情感分类任务的更显著的改进。

[0026] 第二，所述之前的工作探索了结合词典知识以改进在文件级别的情感分类（即，将整个文件分类为是正面的或者负面的）的优势。与这些工作相比，本情感分类器是精细粒度的。特别地，情感分类在句子级别被执行，并且对于每个句子，所述情感分类器不仅预测句子是否是正面的、负面的或者客观的，而且其也预测与该情感相关联的主要主题。所述实验表明 :由发明者建立的领域特定词典（字典）导致对这些任务的两者的改进。

[0027] 关于情感词典的构建，之前的研究已集中在产生通用字典。这些方法范围从手动的方法到半自动化的和自动化的方法。在本公开中，使用下列的结合建立所述领域特定情感词典 :(i) 语料库过滤，(ii) 使用语言模式的网络搜索和 (iii) 字典扩展技术。下面详细地描述了该构建。

[0028] 1. 产生领域特定词典

[0029] 下面使用数字照相机的主题作为其例子描述了产生领域特定词典的方法。然而，将被理解的是 :该方法也适用于其它的领域并且所述照相机的主题仅仅作为方便的例子而被提供。

[0030] 如上面所讨论的，许多单词或短语的情感是依赖于上下文或领域的。例如，如果其与“电池寿命”的照相机方面相关联，则“长的”是正面的。然而，当其与“快门迟滞”的照相机方面相关联时，相同的单词携载负面的情感。因此，当试图确定所述相关联的情感时，知道正被讨论的主题 / 领域是至关重要的。

[0031] 基于该观察，领域 / 主题特定词典被建立，覆盖指示特定领域的表达和指示与该特定领域相关联的不同的情感的表达两者。例如，如在图 1 中所显示的，关于领域 / 主题“照相机图片质量”100 的词典由两个子词典组成。第一子词典 102 包括在数字照相机的领域中对“图片质量”的领域 / 主题而言是好的指示符的单词和 / 或短语，诸如图片、图像、照片、关闭等等。如果所述相关联的领域 / 主题是照相机图片质量 100，则另一个子词典 104 包括携载正面的或负面的情感的单词和 / 或短语。例如，该第二子词典 104 会指示 :当它们与图片的质量相关联（即，领域 / 主题 :照相机图片质量 100）时，虽然“锐利的”和“清晰的”是正面的，“模糊的”是负面的。通过使用所述的下列的组合，该目标被实现 :(i) 语料库过滤，(ii) 用语言模式的网络搜索和 (iii) 字典扩展。在下面的小节中详细地描述了这

些技术中的每个。

[0032] 语料库过滤

[0033] 语料库过滤方法 200 在图 2 中被示出。最初，带注释的训练语料库被提供 (202)。如果针对所关心的领域 / 主题训练语料库不存在，则将需要以本领域已知的方式构建一个。例如，在考虑照相机评论时，通过注释每个评论（其是所述训练语料库的一部分）而构建训练语料库。更特别地，用照相机方面以及在该句子中被表达的相关联的情感注释要被包括在所述训练语料库中的每个包括的照相机评论的每个句子。一旦被构造（或者如果合适的训练语料库已经存在），其直接使用该资源以建立用于领域特定词典的构建的基础。

[0034] 接着，对于每个信息方面（例如，诸如“耐久性”、‘图片质量’等等的照相机方面），在被标记为表达该方面的训练句子中存在的所有内容单词和 / 或短语被提取 (204)。被提取的所述内容单词和 / 或短语包括名词、动词、形容词、副词以及它们的否定形式。根据该提取的内容，针对每个信息方面的初始的词典列表被形成 (206)。

[0035] 随后，对于在针对所述照相机方面的每个的所述列表中的每个单词和 / 或短语，检查被进行以查看该单词或短语是否也存在在任何其它的照相机方面的词典列表中 (208)。如果是，则从所述词典去除该单词和 / 或短语 (210)。如果该单词和 / 或短语不在任何其它列表上，则该单词和 / 或短语被保持在所述列表上 (212)。这些步骤被重复直到没有额外的单词和 / 或短语被留下 (214)。在该过滤的步骤之后，对于每个照相机方面获得词典的列表，其在所述训练语料库中仅包含对该照相机方面而言是唯一的单词和 / 或短语 (216)。

[0036] 使用该方法产生的词典的质量通常是非常高的。例如，基于具有覆盖 23 类（即，22 个照相机方面和“无”的类，意思是所述 22 个照相机方面中一个也没有被讨论）的 2131 句子的相对小的训练语料库产生下面的关于照相机方面“耐久性”的词典。

[0037] 耐久性词典 :[刮痕、构造、建造、摇动 (rock)、修理、损害、易坏的、不易坏的、垃圾 (junk)、坚固的、较坚固的、坚硬的、耐用的、坚韧的、弯曲的、牢固的、不值得的、稳固的、废料 (rug)、破产的 (broke)、防弹的]

[0038] 然而，该方法的缺点是 :所述词典的覆盖范围将完全地依靠所述语料库的覆盖范围，并且注释宽的覆盖范围的训练语料库是费时的、昂贵的并且有时由于自然语言的丰富而对诸如情感分析的任务而言是非常困难的。

[0039] 通过经由网络搜索和使用语言模式的过滤以及字典扩展而扩增从所述训练语料库获得的所述初始的领域特定词典，该缺点被克服。在接着的两个小节中示出了这两个方法。

[0040] 1.2 使用语言模式的网络搜索和过滤

[0041] 转向图 3，流程图 300 被显示，用于使用语言模式的网络搜索和过滤以改进从所述训练语料库获得的所述领域特定词典的覆盖范围。最初，语言模式被设计（在该例子中，两个这样的语言模式被设计）并且被用作搜索查询以发现概念上与所关心的信息方面相关联的更多的单词和短语（例如，所述照相机方面）(302)。在所述照相机评论例子中使用的两个语言模式是 :

[0042] 模式 1 :‘照相机方面包括 *’

[0043] 模式 2 :照相机方面 + ‘种子单词和 *’

[0044] 在这两个模式中，“照相机方面”指诸如“照相机附件”和“照相机价格”的表达。“种子单词”指用于特定的照相机方面的种子单词。例如，“便宜的”和“昂贵的”可以用作用于照相机方面价格的种子单词。注意：在模式 1 中，所述照相机方面名称被包括作为精确的搜索查询的一部分，而在模式 2 中，所述照相机方面名称用作用于所述搜索查询的上下文。

[0045] 取决于所述信息方面的语义特性，选择特定的模式（例如，在所述照相机方面的例子中，所述两个模式中的一个被选择以发现概念上与该方面相关的表达 (304)。例如，虽然‘照相机附件包括 *’对于发现附件表达是非常有效的，‘照相机图片 + ‘清晰的和 *’’对于发现与照相机图片相关的表达是更好的。

[0046] 所述选择的语言模式被提供给搜索引擎，所述搜索引擎将所述模式作为查询发送到因特网，导致搜索结果被返回 (306)。例如，当模式 1 被使用时，其作为查询被发送到搜索引擎。在该实验集中，搜索引擎 Bing（来自微软公司）被使用，尽管将被理解的是：其它的搜索引擎也可以被使用（例如，Google、Yahoo、等等）。接着，从所述返回的搜索结果提取相关的单词 (308)。例如，当模式 1 被使用时，提取由所述搜索引擎返回的前 50 个结果中在“包括 (include)”或“包括 (includes)”之后的单词或短语。在每个返回的结果中，跟随在“包括 (include)”或“包括 (includes)”之后的单词被提取直到“包括 (include)”或“包括 (includes)”之后的第一句子边界被到达。接下来的步骤是从所述提取的单词中去除诸如“该 (the)”（除了别的以外）的普通非用词 (stop words) 和诸如“具有 (with)”和“的 (of)”（除了别的以外）的功能单词 (310)。最后，所剩余的单词被添加到由图 2 的过程形成的所述合适的领域特定词典的列表中 (312)。使用该方法，随后的用于照相机附件的词典在所述照相机例子中被产生。

[0047] 附件词典：[芯片、多个芯片、壳、袋、卡片、软件、三脚架、条带、电缆、适配 (adapt)、充电器、端口、存储器、罩、连接器、工具 (kit)、附件、手套、带子、上边带、麦克风 (mic)、束带圈 (beltloop)、闪存、程序、皮革、包装、连接、非带子、非条带、拉链]

[0048] 作为进一步的例子，当模式 2 被使用时，前 50 个返回的结果中的单词被提取。然而，不同的算法被用于滤出这些返回的结果中的噪声。例如，为了发现概念上与照相机的图片质量相关的表达，“照相机图片”被用作上下文单词并且“清晰的”被用作种子单词。该模式将匹配“清晰的和锐利的”和“清晰的和正常的”两者。然而，虽然“锐利的”通常被用于描述图片质量，“正常的”不是。为了过滤诸如“正常的”的噪声单词，候选单词的每个被用作模式 2 中的新的种子单词，并且如果由所述新的查询返回的前 50 个结果包括原始的种子单词“清晰的”，则所述候选单词被保留。否则，其被丢弃。例如，在所述实验中，虽然‘照相机图片 + ‘锐利的和 *’’将返回匹配“锐利的和清晰的”的结果，‘照相机图片 + “正常的和 *”’将不会返回匹配“正常的和清晰的”的结果。通过该方法，“锐利的”可以区别于“正常的”，并且“正常的”被识别为噪声单词。图 4 显示了当概念上与照相机图片相关的表达在所述的实验期间被提取时一些由该方法识别的噪声单词 (400)。在该图中，由空的圆形表示的单词被识别为噪声并且从所述照相机图片质量词典中被去除。相反，由实心圆形表示的单词被保留在所述词典中。

[0049] 在一个实施例中，当使用模式 2 时，被用于构建领域特定词典的算法被如下识别：算法 1：FindingRelatedWords，其依次使用被识别为如下的算法：算法 2：HarvestByBing 和算法 3：isReversible。

[0050] 使用该方法，通过使用模式 2 作为具有两个种子单词“清晰的”和“模糊的”的搜索查询而建立下面的用于照相机图片质量的词典。

[0051] 图片质量词典：[清晰的、锐利的、颜色、明亮的、京瓷 (Kyocera)、响应、适度的 (sober)、稳定的、整齐的、鲜艳的、分解、细节、纹理、安全的、流动的、黑暗的、阳光充足的、暗淡的、清新的 (crisp)、焦点、图案、曲线、蓝色、潮湿的、不清楚的 (fuzzy)、橙色、黄色、灰色、模糊的、模糊、青色、不清楚的 (indistinct)、粒状的、雾浊的、模糊的 (blurred)]

[0052]

```
Algorithm: FindingRelatedWords
Input: seedword, contextword, depth
Output: relatedwordset
unprocessed = [seedword] ;
relatedwords = [seedword] ;
foreach Depth in [1...N] do
    tempset = [ ] ;
    foreach word in unprocessed do
        newwords = HarvestByBing(word,
        contextword) ;
        foreach newword in newwords do
            if isReversable(word, newword,
            contextword) then
                Add newword to tempset ;
    foreach newword in tempset do
        Add newword to relatedwords
    unprocessed = tempset ;
return relatedwords
```

[0053] 算法 1 :FindingRelatedWords

[0054]

```
Algorithm: HarvestByBing
Input: word, contextword
Output: newwords
LPattern = contextword + “word and *” ;
newwords = words matchig * in texts of top 50
results returned from Bing using LPattern as a
query ;
return newwords
```

[0055] 算法 2 :HarvestByBing

[0056]

```
Algorithm: isReversable
Input: word, newword, contextword
Output: True or False
newwords = HarvestThroughBing(newword,
contextword) ;
if word in newwords then
    return True
else
    return False
```

[0057] 算法 3 :isReversible

[0058] 1.3 字典扩展

[0059] 尽管在建立通用情感词典时通过查找被记录在字典中的同义词和反义词的扩展是通常使用的方法,该方法被发现不总是适合于建立领域特定词典。原因在于 :建立领域特定词典要求发现概念上相关的表达 ;然而,概念上相关的表达不必要是同义词或反义词。例如,“锐利的”和“清晰的”概念上与照相机图片质量相关,但是它们不是从语言观点上看的真正的同义词。

[0060] 然而,有时,使用字典仍然可以是非常有效的。例如,使用模式 2 通过网络搜索和过滤建立下面的用于照相机价格的词典。

[0061] 价格词典 :[便宜的、最低的、折扣、广告的 (promo)、票证 (coupon)、促销 (promote)、昂贵的、有价值的、价值]

[0062] 通过包括如在下面所显示的在 WordNet(Fellbaum,1998) 中的“便宜的”和“昂贵的”的同义词,进一步扩展所述价格词典是可能的。

[0063] WordNet 中的“昂贵的”的同义词 :[昂贵的、大价 (big-ticket)、高价 (high-ticket)、贵重的、高价的 (high-priced)、价格高的、昂贵的 (pricy)、昂贵的 (dearly-won)、费用大的 (costly)、定价过高的]

[0064] Wordnet 中的“便宜的”的同义词 :[便宜的、不昂贵的、廉价的、减价的、削价、不值钱的、非常便宜的、低预算的、低成本的、低价的、买得起的、便宜的 (dime)、花费极少的 (penny)、便宜的 (halfpenny)]

[0065] 1.4 领域特定极性词典

[0066] 到现在为止已经公开了领域特定词典的结构,例如,已经描述了如何已为不同的照相机方面建立领域特定词典。接下来的步骤是在每个领域词典中将携载正面的情感的表达从那些携载负面的情感的表达分离。

[0067] 例如,能够建立下面的用于“图片质量”的子词典是所期望的。

[0068] 图片质量正面的词典 :[清晰的、锐利的、明亮的、适度的、稳定的、整齐的、鲜艳的、阳光充足的、清新的]

[0069] 图片质量负面的词典 :[黑暗的、暗淡的、潮湿的、不清楚的 (fuzzy)、灰色的、模糊

的、模糊、不清楚的 (indistinct)、粒状的、雾浊的、模糊的 (blurred)]

[0070] 转向图 5，描述向如上面所描述的领域特定词典中的单词和 / 或短语提供极性的方法的流程图 500 被示出 (502)。对于通过语料库过滤、网络搜索和字典扩展的组合而被构建的所述产生的词典（例如，所述图片质量词典）中的每个表达（例如，单词或短语），检查被进行以查看正被检查的单词或短语是否仅出现在被标记为表达正面的意见或负面的意见（例如，关于所述照相机的图片质量）的所述训练数据中 (504)。如果其是正面的意见，则该表达被包括到所述图片质量正面的词典中 (506)，而如果其是负面的意见，则该表达被包括到所述图片质量负面的词典中 (508)。

[0071] 已经示出了用于构建领域特定情感词典的本方法，接着描述如何将词典知识结合到 SVM 学习中以改进情感分类。

[0072] 2. 将词典知识结合到 SVM 学习中以改进情感分类

[0073] 已经产生了包含正面的领域特定子词典和负面的领域特定子词典的领域特定词典，本公开现在描述将所述领域特定词典中的单词和表达结合到机器学习系统中以便执行如下的情感分类任务。对于关于照相机的每个评论句子，所述情感分类器需要预测在该句子中讨论的照相机方面以及关于该照相机方面的相关联的情感两者。例如，对于下面的评论句子：

[0074] (1) 其使用两个 (2) 电池并且所述电池比我的使用四个 (4) 电池维持的上一个照相机持续更长。

[0075] 所述情感分类器将识别该句子表达关于所述照相机的电池寿命的正面的意见。

[0076] 通过采用如在图 6 中所示的两步情感分类器 600（具有方面分类器 602 和极性分类器 604）以执行两步分类，该目标被实现。在步骤 1，情感分类器 600 的方面分类器 602 被训练以预测正被讨论的方面（例如，所述照相机方面）。在步骤 2，情感分类器 600 的极性分类器 604 被训练以预测与该方面相关联的情感。最后，在集合器 606 中将所述两步预测结果集合在一起以产生最后的预测。

[0077] 在该两步中，所述词典知识被结合到常规的机器学习系统（例如，SVM 学习）中。为了示出该方法，下面的句子 (2) 被用作与图 7 的流程图 700 相结合的例子，其中句子 (2) 被呈现到所述 SVM(702)。

[0078] (2) 外壳是坚硬的因此其给所述照相机额外的好的保护。

[0079] 在常规的 SVM 学习中使用名词、动词、形容词和副词作为特征单词，该句子被表示为下面的单词矢量 (704)。

[0080] [外壳、坚硬的、给、照相机、额外的、好的、保护]

[0081] 所述产生的词典被结合到所述 SVM 中 (706)。通过该结合在所述词典中被编码的知识，附加的特征被自动地产生并且被插入到上面的表示中。

[0082] 例如，当执行步骤 1 方面分类时 (708)，由于上面的表示中的特征单词“外壳”被列在关于照相机附件的领域特定词典中，附加的特征单词“附件”被插入，并且下面的新的表示被产生。

[0083] [外壳、坚硬的、给、照相机、额外的、好的、保护、附件]。

[0084] 通过这样做，如果所述句子中存在照相机附件的表达则所述照相机方面是“附件”的可能性被提升。

[0085] 在极性预测的下一步 (710) 中，从多视角问题回答 (MPQA) 意见语料库 (例如，参见 Wiebe 等人，2005) 提取的领域特定情感词典和通用的独立于领域的情感词典两者被结合。仅提取被指示为来自所述 MPQA 意见语料库的脱离上下文的强主观的单词。

[0086] 例如，因为“好的”被指示为所述 MPQA 词典中的正面的单词，特征单词“正面的”将被插入。此外，如果用于句子 (2) 的所述第一步预测结果是“附件”，并且“坚硬的”在关于照相机附件的领域特定词典中也是正面的单词，则额外的特征单词“正面的”将在如在下面所示的针对所述第二步极性预测的用于句子 (2) 的最后的表示中被产生。

[0087] [外壳、坚硬的、给、照相机、额外的、好的、保护、正面的、正面的]。

[0088] 因此，关于“附件”的方面，“正面的”预测被提升 (例如，因此当附加的单词被识别为正面的单词时，对应的附加的额外的特征“正面的”将被加在所述最后的表示中。

[0089] 所述实验显示 :将词典知识结合到 SVM 学习中显著地改进了所述分类任务的准确度 ;与通用 MPQA 情感词典相比，所述构建的领域特定词典是更有效的。在接下来的小节中，实验设置和结果被报告。

[0090] 3. 实验设置和结果

[0091] 在所述实验中执行的情感分析任务是组合的 45- 方式 (45-way) 的情感分类任务。这 45 个类源自与照相机购买相关的 22 个方面 (诸如，“图片质量”、“LCD 屏幕”、“电池寿命”和“客户支持”) 和它们的相关联的极性值“正面的”和“负面的”，以及关于任何所述 22 个方面的无意见的类。这样的类的例子是“图片质量 :正面的”。目标是将每个输入句子映射到所述 45 个类中的一个。

[0092] 如在前面的小节中所述的，对所述任务执行两步分类。即，最后的组合的分类器由两个分类器组成。第一个是方面分类器，其执行 23- 方式的照相机方面分类。第二个是极性分类器，其执行 3- 方式 (正面的、负面的和无) 的分类。根据这两个分类器产生的预测集合最后的预测。

[0093] 所述分类准确度被定义如下。

[0094] $$Accuracy = \frac{NumberofSentencesCorrectlyClassified}{TotalNumberofSentences}. \quad (1)$$

[0095] 在所述实验中，发明人使用 2718 个手工标记的句子。从由 Blitzer 等人在 " Bollywood, Boom-Boxes, and Blenders " (用于情感分类的领域自适应 (Domain adaptation for sentiment classification)，计算语言学联合会 (ACL) 会议论文集 (Proc.of the Assoc.for Computational Linguistics(ACL))，2007) 中创建的多领域情感数据集随机地选择所有的句子。

[0096] 本公开的系统和方法已经被与将词典知识与 SVM 学习相结合的思想以及与仅仅常规的 SVM 学习相比较，因为后者是在用于情感分析的文献中被报告的当前技术现状的算法。

[0097] 名词、动词、形容词和副词被选择作为来自用于训练和测试的句子的一元文法 (unigram) 单词特征。使用如由 Rijsbergen 等人在 " 概率信息检索中的新的模型 (New Models in Probabilistic Information Retrieval)" (技术报告 (Technical Report)，英国图书馆研究和发展报告 (British Library Research and Development Report)，No.5587,1980) 中所描述的 Porter Stemmer，它们中的所有被取词干 (stemmed)。非门

(Negators) 被附接到接下来选择的特征单词。同样使用的是非用词的小的集合以排除诸如“take”的系词和单词。所使用的非用词包括系词和下面的单词：“take、takes、make、makes、just、still、even、too、much、enough、back、again、far，same”。这些单词被选择作为非用词的原因是：因为它们不但是时常使用的而且是不明确的并且因此趋向于对所述分类器具有负面的影响。所述被采用的 SVM 算法由 Chang 等人在“LIBSVM：用于支持向量机的程序库 (A Library for Support Vector Machines)”(软件在 http://www.csie.ntu.edu.tw/cjlin/libsvm 可获得，2001) 中实现。线性的内核类型被使用并且对所有其他参数的缺省设置被使用。

[0098] 四个 (4) 实验被进行。在实验 1 中，常规的 SVM 算法被使用，在其中没有词典知识被结合；该实验被称为 SVM。在实验 2 中，仅在所述独立于领域的 MPQA 意见字典中被编码的知识被结合到 SVM 学习中；该实验被称为“MPQA+SVM”。在实验 3 中，仅在所述构建的领域特定词典中被编码的知识被结合到 SVM 学习中；该实验被称为“领域词典 +SVM”。在实验 4 中，在所述 MPQA 中被编码的知识和所述构建的领域特定词典两者被结合到 SVM 学习中；该实验被称为“领域词典 +MPQA+SVM”。所有的实验结果是基于 10- 折叠 (10-fold) 的交叉验证，并且它们在表 1 中被概括。

[0099] 表 1 显示了：结合所述独立于领域的 MPQA 词典和所述构建的领域特定词典两者实现了最佳的总性能。在这两种类型的词典中，结合所述领域特定词典是更有效的，因为它们对所述分类准确度的改进贡献最大。根据成对的 t- 测试，由我们的方法实现的改进是统计上显著的，具有 $\rho < 0.000001$。

[0100]

学习方法	准确度
SVM	41.7%
MPQA + SVM	44.3%
领域词典 + SVM	46.2%
领域词典 +MPQA + SVM	47.4%

[0101] 表 1：总性能比较

[0102] 表 2 进一步示出了：将词典知识与 SVM 学习相结合显著地改进了针对照相机方面分类的准确度和针对极性分类的准确度两者。分别根据成对的 t- 测试，该两种改进是统计上显著的，具有 $\rho < 0.000001$ 和 $\rho < 0.05$。

[0103]

学习方法	方面准确度	极性准确度
SVM	47.1%	65.6%
领域词典 +MPQA+SVM	56.2%	66.8%

[0104] 表 2：细分性能比较

[0105] 4. 所公开的系统和方法为什么工作

[0106] 小节 3 提供了经验证据：将词典知识结合到 SVM 学习改进了情感分类的准确度。下面提供了关于这为什么是真的的理论证据。

[0107] 在支持向量机的情况下，数据点被视为p-维矢量，并且SVM的策略是：发现产生任何两类之间的最大间隔或者余量的(p-1)-维超平面。两类之间的余量(margin)越大，这两类是越可分离的。所描述的系统和方法改进分类的准确度的原因是：因为基于所述情感词典而被插入的所述额外的特征扩大了属于不同的类的点之间的距离，同时保持属于相同的类的点的距离不变。下面示出了证据。

[0108] 假定点$\vec{a}=(x_1^a, x_2^a, \ldots, x_p^a)$和点$\vec{b}=(x_1^b, x_2^b, \ldots, x_p^b)$属于类A，并且点$\vec{c}=(x_1^c, x_2^c, \ldots, x_p^c)$和点$\vec{d}=(x_1^d, x_2^d, \ldots, x_p^d)$属于类A15和类B

[0109] 在所述实验中，SVM和线性的内核一起使用，在其中数据点之间的距离由这些点之间的欧几里德(Euclidean)距离测量。例如，$\vec{a}$和$\vec{c}$之间的距离和$\vec{c}$和$\vec{d}$之间的距离等于下面的$\text{Distance}_{\text{old}}(\vec{a}, \vec{c})$和

[0110] $\text{Distance}_{\text{old}}(\vec{c}, \vec{d})$ 。

[0111] $$Distance_{old}(\vec{a},\vec{c})$$

[0112] $$=\sqrt{(x_1^a-x_1^c)^2+\cdots+(x_p^a-x_p^c)^2}$$

[0113] $$Distance_{old}(\vec{c},\vec{d})$$

[0114] $$=\sqrt{(x_1^c-x_1^d)^2+\cdots+(x_p^c-x_p^d)^2}$$

[0115] 当根据所述构建的类/领域特定词典将额外的特征μ添加到所有属于类A的点并且将不同的额外的特征v添加到所有属于类B的点时，额外的维度被添加到所述数据点的每个。随后，$\vec{a}$和$\vec{c}$之间的新的距离和$\vec{c}$和$\vec{d}$之间的新的距离可以如下被计算。

[0116] $$Distance_{new}(\vec{a},\vec{c})$$

[0117] $$=\sqrt{(x_1^a-x_1^c)^2+\cdots+(x_p^a-x_p^c)^2+(\mu-\mu)^2}$$

[0118] $$=Distance_{old}(\vec{a},\vec{c})$$

[0119] $$Distance_{new}(\vec{c},\vec{d})$$

[0120] $$=\sqrt{(x_1^c-x_1^d)^2+\cdots+(x_p^c-x_p^d)^2+\mu^2+\nu^2}$$

[0121] $$>Distance_{old}(\vec{c},\vec{d})$$

[0122] 从上面的计算如下是清楚的：在将所述额外的特征单词μ添加到类A中的所有点并且将v添加到类B中的所有点之后，$\vec{a}$和$\vec{c}$之间的距离保持相同而$\vec{c}$和$\vec{d}$之间的距离将被扩大。

[0123] 总而言之，如在图8中所示的，虽然属于相同的类(即，类A(800)或类B(802))的点之间的距离保持不变，在根据我们的类/领域特定词典插入所述额外的特征单词之后，属于不同的类的点之间的距离将被扩大。

[0124] 这也意味着：在添加所述额外的特征之后，SVM可以发现具有更大余量或者具有相同的余量长度但较少支持矢量的超平面，其可以更有效地分离类。这而又导致针对分类的更高的准确度。

[0125] 5. 系统配置

[0126] 根据图9-11，系统配置被示出，在其中本公开的方法被实现。特别地，图9-11示出了情感分类器的开发的各个阶段，包括：(i) 产生被结合到所述情感分类器的领域特定词典，(ii) 训练所述情感分类器，以及 (iii) 操作所述情感分类器，包括对各种类型的文本文件（包括但不限于文章、报纸、博客 (blogs)、杂志，其可以诸如在因特网上以电子形式正目前存在，或者在硬拷贝中，其随后被扫描成电子的格式或可由所述系统电子地阅读的别的形式）的情感分析。

[0127] 在实施例中，所述情感分类系统被适当地具体化为计算机900或运行合适的软件的其他数字处理设备以配置、训练所述情感分类器并且执行情感分析。所述示出的计算机900包括用于显示结果的显示器902和用于帮助用户控制所述情感分类器的输入设备（键盘、鼠标等等）904。所述计算机900也被配置有合适的互连以访问所述因特网906以及局域网908、打印机910和扫描器912，所述打印机能够打印出文件的硬拷贝，其可以被认为是从所述情感分类器的输出，并且所述扫描器被配置为扫描输入 (scan in) 硬拷贝文件以便在电子版本中操作所述硬拷贝文件用于情感分析。

[0128] 所述软件被适当地具体化为被存储在磁盘、光盘、磁带、FLASH存储器、随机存取存储器 (RAM)、只读存储器 (ROM)、或其它存储介质上的指令，这样的指令可由计算机900或其它数字设备执行以执行所公开的能够执行情感分析的情感分类器的构建、训练和操作。

[0129] 继续关注图9，本申请的系统在该阶段操作以通过使用注释的数据库916、过滤的因特网搜索结果918和字典920而开发所述领域特定词典914。一旦所述领域特定词典已被开发，所述情感分类器被训练。特别地，如在图10中所示的，所述两级情感分类器设计1000包括方面分类器1002和极性分类器1004。如所示的，这些分类器已结合情感标记的领域特定词典1006和情感标记的通用词典1008的至少一部分。在所述训练操作期间，来自训练数据库1010的数据被提供给所述情感分类器1000，其对所述数据进行操作。所述方面分类器1002和所述极性分类器1004的输出进一步被配置为向集合器1012提供它们的输出以产生预测输出1014。一旦所述情感分类器1000已被训练，情感分析被开始进行。例如，如在图11中所示的，所述训练的情感分类器1100接收内容1102，其被分析以产生输出1104。特别地，在逐个句子 (sentence-by-sentence) 的基础上分析所述内容并且关于所述正被分析的文本的情感的预测被产生作为所述输出1104。

[0130] 6. 结论

[0131] 已经显示了：结合在情感词典（尤其是领域特定词典）中被编码的知识可以显著地改进针对精细粒度的情感分析任务的准确度。也描述了构建领域特定情感词典的方法，其中特定的例子是：针对照相机评论的领域构建领域特定情感词典，其中这样的构建包括语料库过滤、网络搜索和过滤以及字典扩展的组合。此外，已经开发和显示了将所述词典知识结合到机器学习算法（诸如SVM）以改进情感学习的方法。

[0132] 将被理解的是：上面所公开的和其它特征和功能的变体或者其替代物可以被组合成许多其它不同的系统或应用。各种目前未预见到的或未预料到的其中的替代物、修改、变化或改进可以由本领域技术人员随后进行，其也意在被下面的权利要求包括。

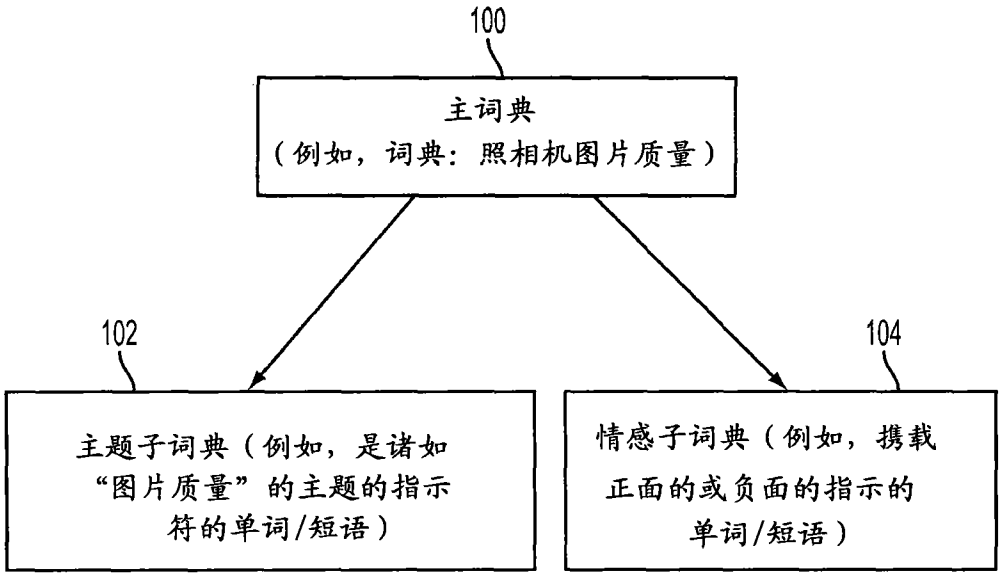

100
主词典
（例如，词典：照相机图片质量）
102
主题子词典（例如，是诸如“图片质量”的主题的指示符的单词/短语）
104
情感子词典（例如，携载正面的或负面的指示的单词/短语）

图 1

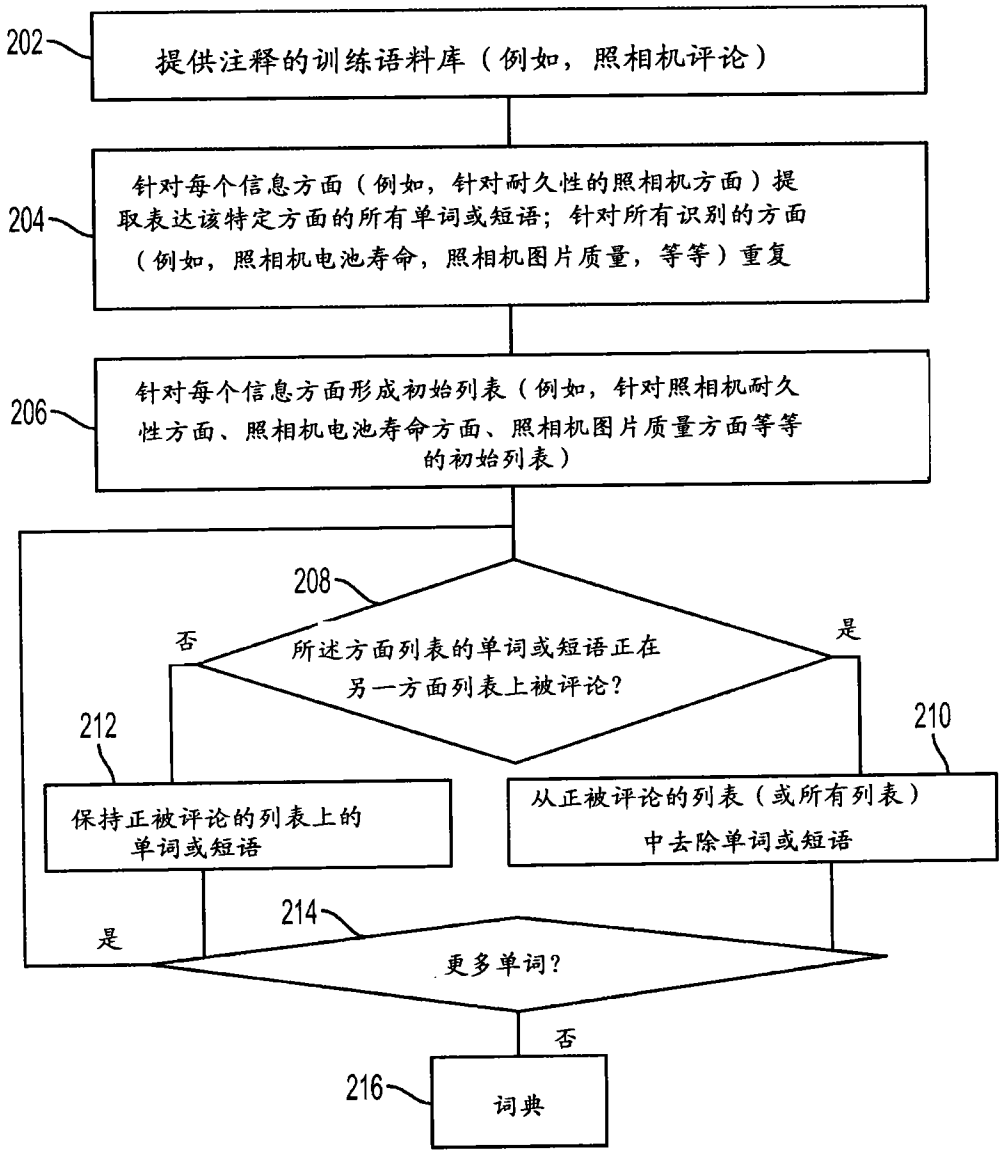

202
提供注释的训练语料库（例如，照相机评论）
204
针对每个信息方面（例如，针对耐久性的照相机方面）提取表达该特定方面的所有单词或短语；针对所有识别的方面（例如，照相机电池寿命，照相机图片质量，等等）重复
206
针对每个信息方面形成初始列表（例如，针对照相机耐久性方面、照相机电池寿命方面、照相机图片质量方面等等的初始列表）
208
所述方面列表的单词或短语正在另一方面列表上被评论？
否
是
212
保持正被评论的列表上的单词或短语
210
从正被评论的列表（或所有列表）中去除单词或短语
214
更多单词？
是
否
216
词典

图 2

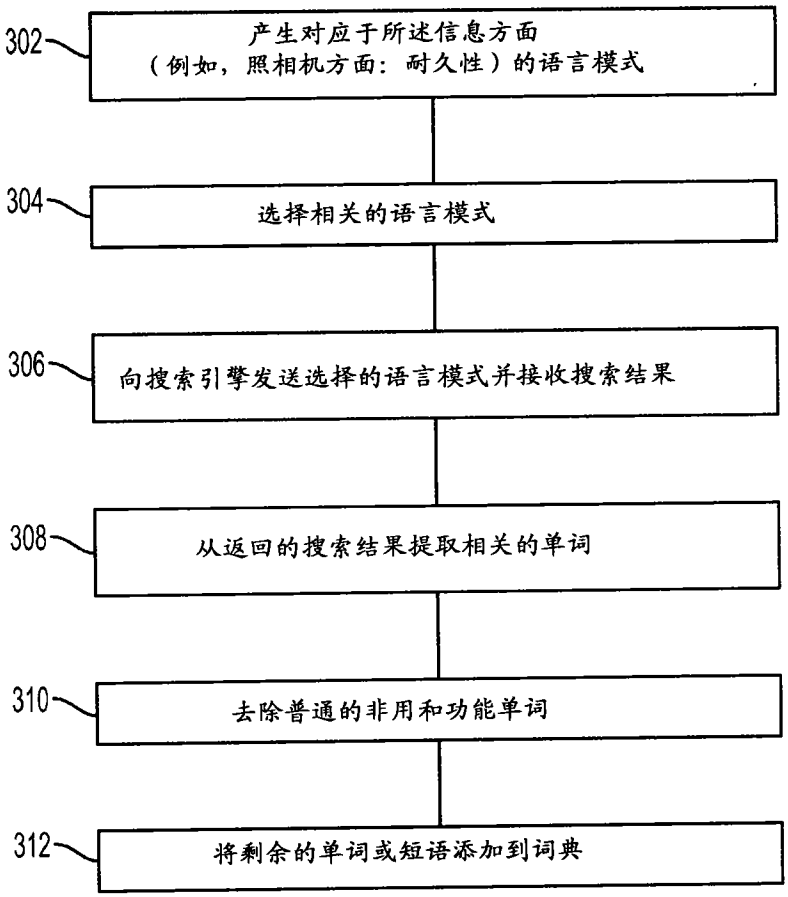

302
产生对应于所述信息方面
（例如，照相机方面：耐久性）的语言模式
304
选择相关的语言模式
306
向搜索引擎发送选择的语言模式并接收搜索结果
308
从返回的搜索结果提取相关的单词
310
去除普通的非用和功能单词
312
将剩余的单词或短语添加到词典

图 3

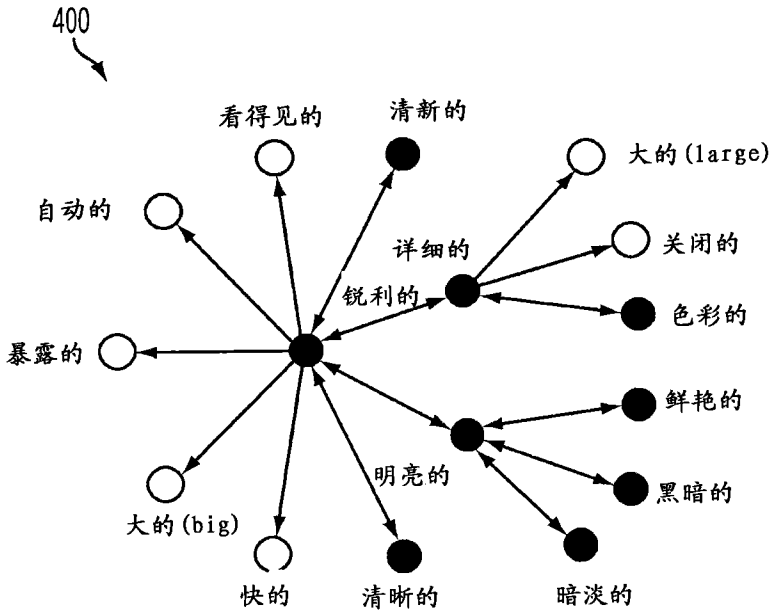

400
看得见的
清新的
大的(large)
自动的
详细的
关闭的
锐利的
色彩的
暴露的
鲜艳的
明亮的
黑暗的
大的(big)
快的
清晰的
暗淡的

图 4

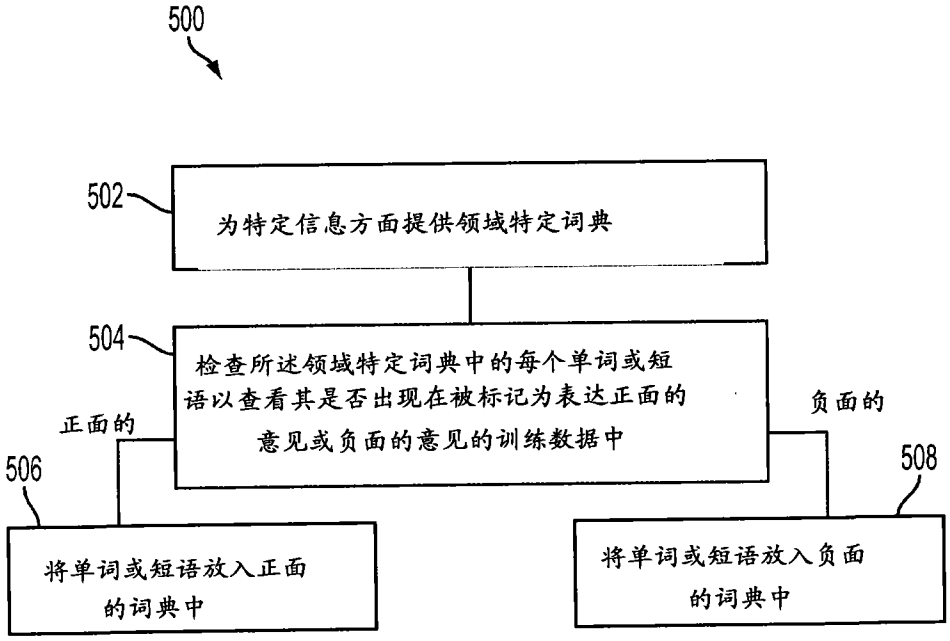

500
502
为特定信息方面提供领域特定词典
504
检查所述领域特定词典中的每个单词或短语以查看其是否出现在被标记为表达正面的意见或负面的意见的训练数据中
正面的
负面的
506
508
将单词或短语放入正面的词典中
将单词或短语放入负面的词典中

图 5

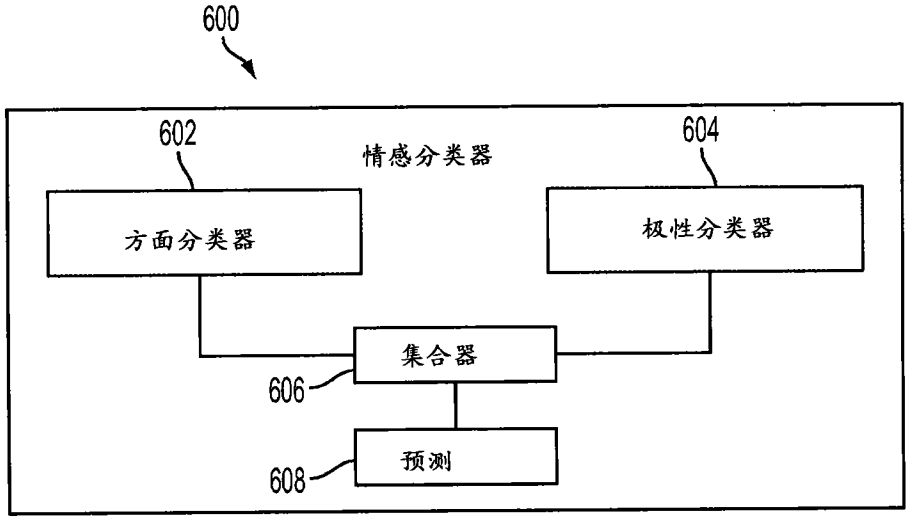

600
602
情感分类器
604
方面分类器
极性分类器
集合器
606
预测
608

图 6

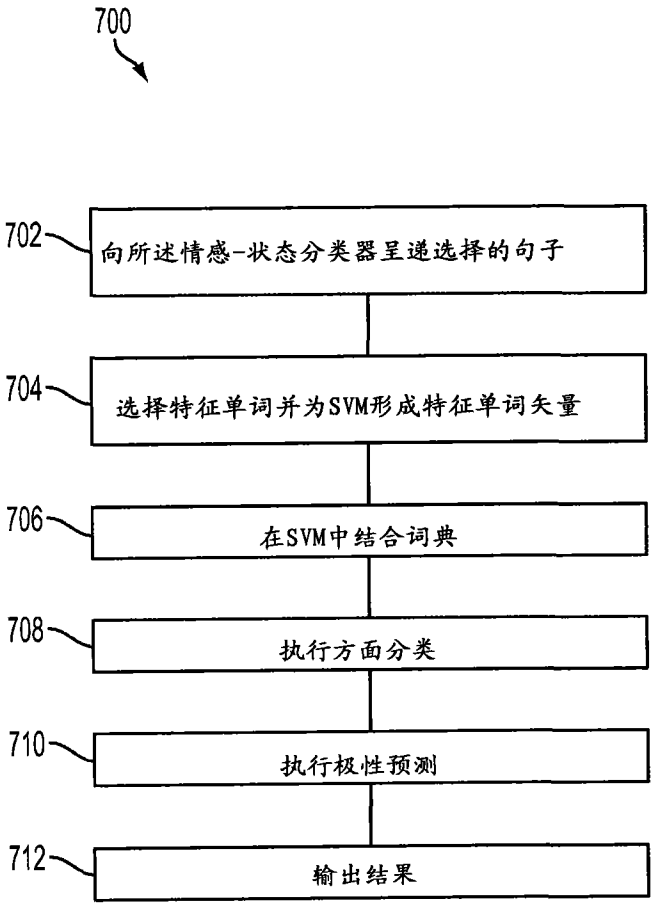

700
702
向所述情感-状态分类器呈递选择的句子
704
选择特征单词并为SVM形成特征单词矢量
706
在SVM中结合词典
708
执行方面分类
710
执行极性预测
712
输出结果

图 7

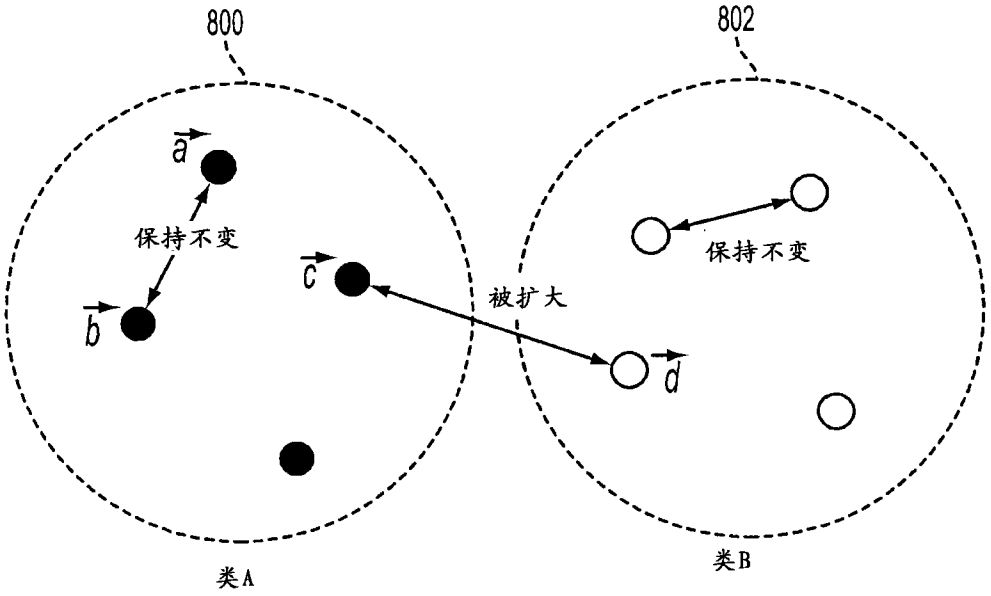

800
802
a
保持不变
b
c
被扩大
d
保持不变
类A
类B

图 8

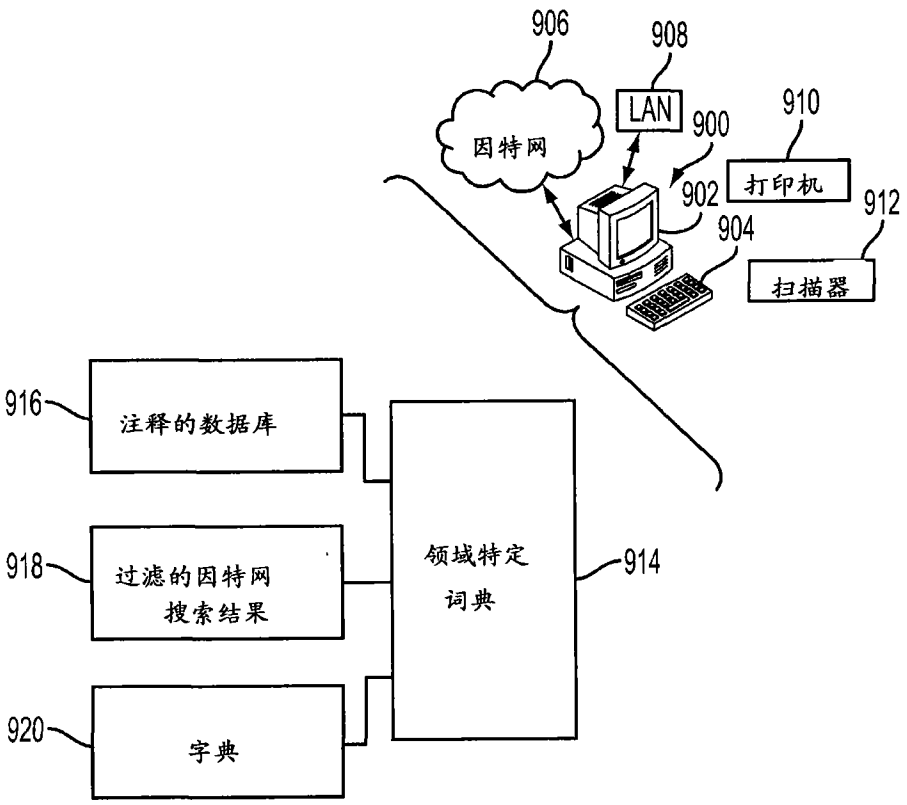

906
908
LAN
900
910
因特网
打印机
902
912
904
扫描器
916
注释的数据库
领域特定
词典
914
918
过滤的因特网
搜索结果
920
字典

图 9

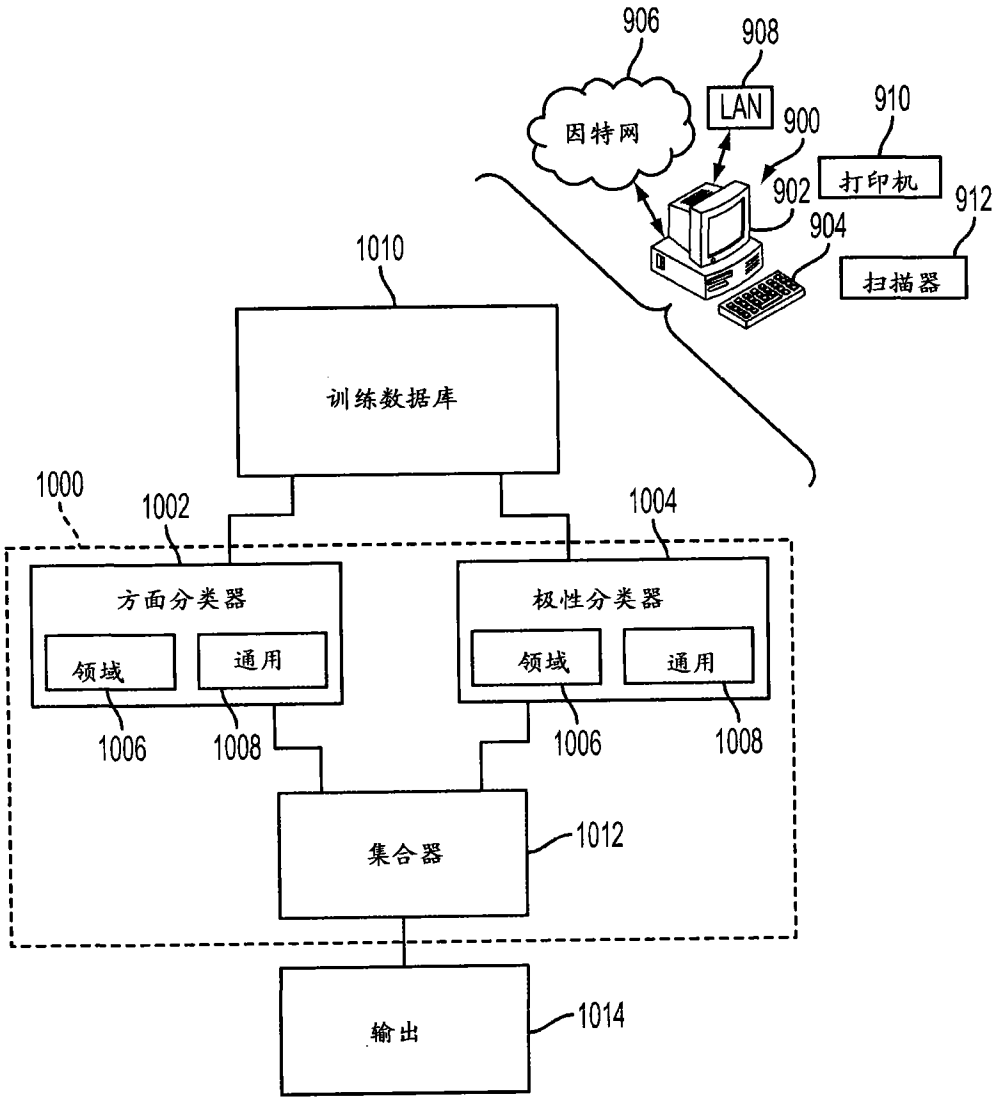

906
908
因特网
LAN
900
910
打印机
902
912
904
扫描器
1010
训练数据库
1000
1002
1004
方面分类器
领域
通用
极性分类器
领域
通用
1006
1008
1006
1008
集合器
1012
输出
1014

图 10

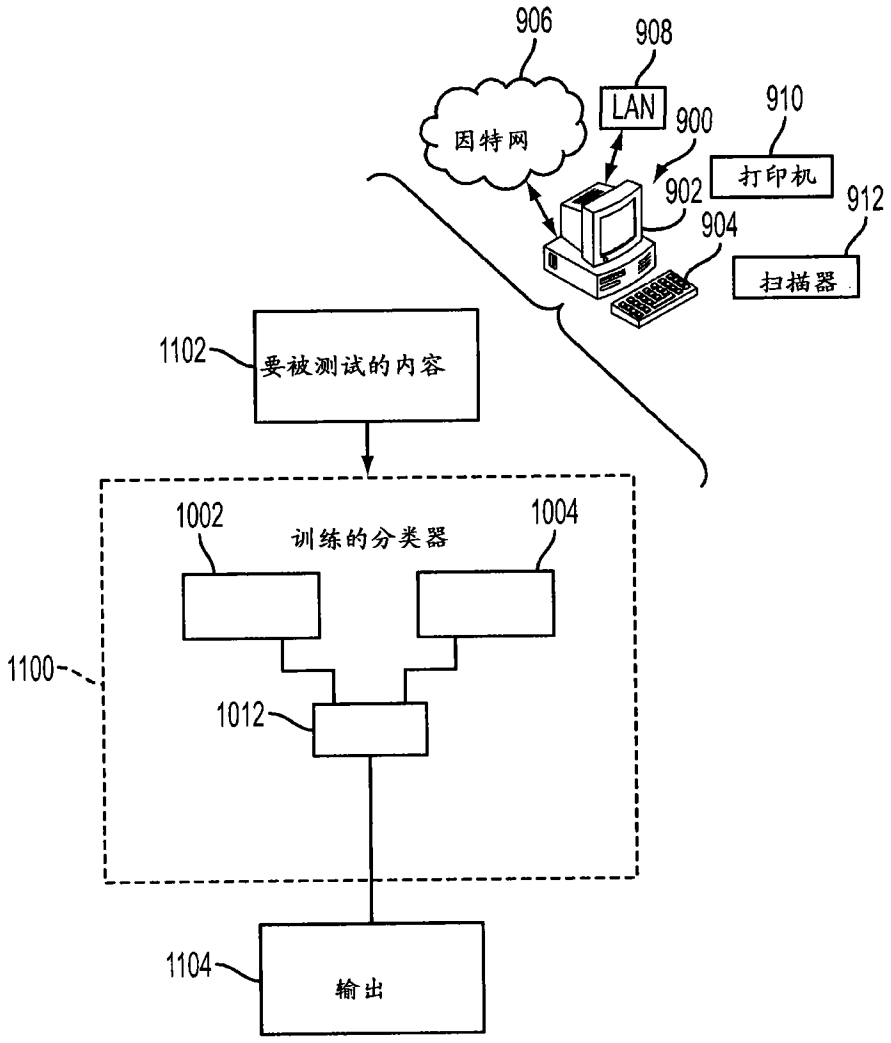

906
908
LAN
910
因特网
900
打印机
902
912
904
扫描器
1102
要被测试的内容
1002
训练的分类器
1004
1100
1012
1104
输出

图 11