



(12) 发明专利申请

(10) 申请公布号 CN 102160066 A

(43) 申请公布日 2011.08.17

(21) 申请号 200980133803.1

(51) Int. Cl.

(22) 申请日 2009.06.23

G06N 7/00 (2006.01)

G06N 1/00 (2006.01)

(30) 优先权数据

61/129408 2008.06.24 US

61/136414 2008.09.03 US

(85) PCT申请进入国家阶段日

2011.02.24

(86) PCT申请的申请数据

PCT/IL2009/000621 2009.06.23

(87) PCT申请的公布数据

W02009/156987 EN 2009.12.30

(71) 申请人 沙伦·贝伦宗

地址 以色列特拉维夫

申请人 利亚特·贝林森

(72) 发明人 沙伦·贝伦宗 利亚特·贝林森

(74) 专利代理机构 中国专利代理(香港)有限公

司 72001

代理人 杜娟娟 高为

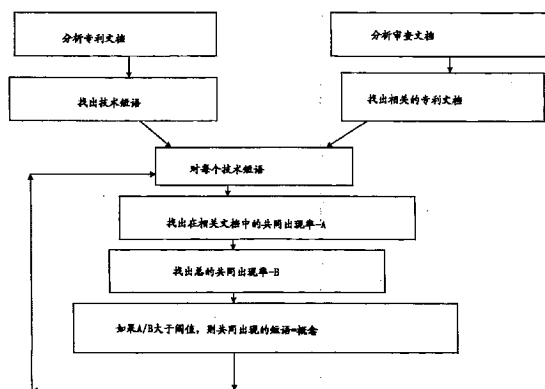
权利要求书 2 页 说明书 30 页 附图 18 页

(54) 发明名称

特别适用于专利文献的搜索引擎和方法

(57) 摘要

一种用于基于相关的科学或技术概念进行搜索的搜索引擎,包括:学习模块,其用于基于技术短语在相关文档中的出现率来了解所述技术短语之间的关系,从其由相关短语的组合形成概念;以及搜索模块,其用于基于存在于查询文档中的概念在所述相关文档的出现来搜索查询文档的相关文档,所述学习模块基于文档的训练集和文档之间的关系来进行所述了解。



1. 一种用于基于相关的科学或技术概念进行搜索的搜索引擎,包括:
学习模块,其用于基于技术术语在相关文档中的出现率来了解所述技术短语之间的关系,从其中由相关短语的组合形成概念;以及
搜索模块,其用于基于存在于查询文档中的概念在所述相关文档中的出现来搜索所述查询文档的相关文档,
其中,所述学习模块基于文档的训练集和文档之间的关系来进行所述了解。
2. 如权利要求 1 所述的搜索引擎,其中,所述训练集包括专家提供的指示相关文档的联系。
3. 如权利要求 1 所述的搜索引擎,其中,所述学习模块配置为基于在概念内组合的技术短语在单个文档内的出现,使所述概念与所述单个文档的相关度量化。
4. 如权利要求 3 所述的搜索引擎,其中,所述学习模块配置为基于在两个文档之间具有共同地相对较高相关度的大量概念使所述两个文档之间的关系量化,所述量化能够与指示相关文档的阈值一起使用。
5. 如权利要求 2 所述的搜索引擎,其中,所述学习模块配置为识别所述训练集的所有文档中的所述技术短语,并且对照总的共同出现率计算在相关文档中的共同出现率,从而形成所述组合。
6. 如权利要求 4 所述的搜索引擎,其中,所述学习模块配置为识别所述训练集的所有文档中的所述技术短语,并且对照总的共同出现率计算在相应的相关文档中的共同出现率,从而形成所述组合。
7. 如权利要求 2 所述的搜索引擎,其中,所述训练集包括一组专利文档,并且所述文档之间的关系包括从专利审查文件的衍生。
8. 如权利要求 7 所述的搜索引擎,其中,所述学习模块包括基于记录在所述专利审查文件中的专利性结果使两个专利文档相关的文档分析能力。
9. 如权利要求 8 所述的搜索引擎,其中,所述专利审查文件包括对所述文档中的第一个的专利审查报告,并且其中所述技术短语的识别是基于所述专利审查报告所涉及的所述第一个文档的版本。
10. 如权利要求 9 所述的搜索引擎,其中,所述专利审查报告被分析以识别驳回或反对,并识别提供所述驳回或反对的基础的其他专利申请的引用。
11. 如权利要求 1 所述的搜索引擎,其中,所述学习模块包括用于识别出现在所述训练集中的所述技术短语的技术短语识别单元。
12. 如权利要求 11 所述的搜索引擎,其中,所述搜索模块配置为基于技术短语在各个目标文档中的出现来对照查询文档排列所述目标文档,所述技术短语属于在所述查询文档中发现的所述概念中的一些。
13. 一种搜索专利文献的方法,包括:
获得包括专利申请和相应的专利局报告的专利文献的训练集;
使用所述专利局报告将特定的专利申请组合在一起作为相关的专利申请;
如果技术短语在相关的专利申请中比在所述训练集整体中更经常出现,则将所述技术术语组合在一起;以及
利用相应的相关术语搜索技术文献。

14. 一种用于基于相关的科学或技术概念进行搜索的搜索方法,包括:

基于技术短语在训练集的相关文档中的出现率来了解所述技术短语之间的关系,从而形成相关短语的概念;以及

基于存在于查询文档中的概念的出现搜索所述查询文档的相关文档。

15. 如权利要求 14 所述的搜索方法,其中,所述训练集包括专家提供的指示相关文档的联系。

16. 如权利要求 15 所述的搜索方法,其中,所述了解包括识别所述训练集的所有文档中的所述技术短语,以及对照总的共同出现率计算在相关文档中的共同出现率,从而形成所述组合。

17. 如权利要求 15 所述的搜索方法,其中,所述训练集包括一组专利文档,并且所述文档之间的关系包括从专利审查文件的衍生。

18. 如权利要求 17 所述的搜索方法,其中,所述了解包括基于记录在所述专利审查文件中的专利性结果使两个专利文档相关的文档分析。

19. 如权利要求 14 所述的搜索方法,其中,所述了解包括用于识别出现在训练集中的所述技术短语的技术短语识别阶段。

20. 如权利要求 19 所述的搜索方法,其中,所述搜索包括基于技术短语在相应目标文档中的出现来对照查询文档排列所述目标文档,所述技术短语属于在所述查询文档中发现的所述概念中的一些。

21. 如权利要求 14 所述的搜索方法,还包括计算查询文档与目标文档之间的度量,所述度量在识别应用风险中可使用。

22. 如权利要求 21 所述的搜索方法,其中,对相关文档的所述搜索在可获得的技术文献上进行。

特别适用于专利文献的搜索引擎和方法

技术领域

[0001] 本发明在其一些实施例中涉及一种搜索引擎和方法,更具体地但非唯一地,本发明涉及适用于专利文献的这样一种用于进行专利搜索的搜索引擎和方法。

背景技术

[0002] 企业的成长和繁荣的核心是使知识资产专利化。在过去的十年,专利申请的数量已经急剧地上升,单单在美国和欧洲就达到每年大约一百万件新申请。

[0003] 对现有技术的搜索是专利申请过程的关键部分,并且是专利范围的决定因素。在专利申请没有能够识别出所有的相关现有技术的情况下,专利申请的权利要求可能被审查员驳回,或者在专利已授权的情况下受到昂贵的诉讼。

[0004] 为了获得授权,专利申请的权利要求不得不满足关于现有技术的两个条件:其必须是新颖且非显而易见的。新颖性意指权利要求必须唯一地限定出一种新的知识,该新的知识在过去没有被授予过专利,并且没有公开地出版过。显而易见性意指创造性,即对现有技术的技术优点,必须不仅仅是简单直接的改变。为了判断一件新的专利申请是否确实是新颖的和非显而易见的,专利审查员对其他专利文档和公开出版物中的相关现有技术进行搜索。

[0005] 对现有技术进行搜索这一市场在专利提交呈指数增长之后已经在迅猛地发展。图1和图2示出了在世界范围内,在主要的专利局的专利提交数量的增长。在2005年,在世界范围内提交了大约1,660,000件专利申请。自1995年来,专利申请的提交已经以4.7%的年增长率增加。

[0006] 对现有技术的搜索发生在创新过程的每个阶段。发明人进行对现有技术的搜索以对所属领域进行研究,并审查他的想法的新颖性和专利性,风险投资家对现有技术进行搜索以评估商业价值,专利律师在提交专利申请时对现有技术进行搜索,并且专利审查员对现有技术进行搜索以判断专利性和专利范围。保守地估计,专利现有技术搜索的年市场规模为四十亿美元(该数量反映出以每次搜索的费用为2,000美元的两百万次现有技术搜索)。

[0007] 对现有技术的搜索对技术许可的广阔市场也是重要的。技术许可的市场正在迅速地增长,估计每年为数十亿美元。

[0008] 最后,现有技术的搜索对专利诉讼,尤其是侵权和无效诉讼是必需的。在美国,每年提交大约1000件专利诉讼案。对与这些诉讼相关的资金转让,没有明确的估计,可能在零(交叉许可协议的结果)到亿万美元之间(例如,在黑莓诉讼案中,RIM公司付了NTP公司612,5百万美元)。

[0009] 对现有技术的搜索的跨度在数百万个专利文档之间。自动化的现有技术搜索的主要挑战是对专利文档的大规模数据集,如何基于文字特征识别科学方面的关系。现有搜索引擎的普遍假设是,专利文档的语义相似性反映出科学方面的相关性。该假设在实际中的可执行性较差,因为科学相关性通常与语义相似性没有关系。在实际中,相关的科学构想通

常包括不同的科学概念。判断文字和技术用语的概念相关性需要特定的专业知识和对千百种相关技术的评价。直至最近,这种系统知识还几乎是不可能获得。在技术使用广泛不同的软件领域,该问题特别突出。

[0010] 当前的市场方案

[0011] 在过去一些年,已经出现了多种盈利性和非盈利性的专利搜索引擎。主要的现有技术搜索引擎是:USPTO, EPO, Google 专利、Dialog 和 Delphion。

[0012] 这些搜索引擎大多数基于语义相似性分析,也称为词袋 (bag-of-words) 方法 (BOW)。搜索过程基于每个文档或查询中词的文字重叠的衡量来计算文档的关联程度。本质上,主要的假设是代表相关科学构想的专利共有共同或相似的语义。就违反该假设来说,现有搜索引擎的性能总是不令人满意。

[0013] 包括其扩展(见下文)在内的语义相似性方法的主要缺点是,其并不提供任何关于词和技术短语的概念含义的信息。例如,词 x 可与词 y 表示完全相同的含义。在没有告知 x 和 y 表示相同含义的外界信息或科学“词典”的情况下,基于语义相似性的信息搜索将失败。

[0014] 专利现有技术搜索领域的语义搜索引擎的性能差有四个主要原因。首先,发明者具有以与最相关的现有技术尽可能远的方式表述他的发明的动机,希望减小申请被审查员驳回的风险。

[0015] 第二,用来描述科学构想的文字域 (textual domain) 一般较大。

[0016] 第三,在很多情况下,与申请本身相比,专利审查员引用的现有技术来自不同的技术领域,其中在现有技术与申请之间存在非常小的文字重叠。例如,发明名称为“车辆部件的鉴定 (Authentication of Vehicle Components)”的专利 No. 7137001 (IPCH04L 数字信息的传输) 与发明名称为“执行分层组结构中的组排除的方法 (Method for Performing Group Exclusion in Hierarchical Group Structures)”的专利 No. 5220604 (IPC G06F 电数字数据处理) 具有非常小的语义相似性。然而,在专利 No. 7137001 的申请过程中,专利审查员引用专利 No. 5220604 作为相关的现有技术并作为以显而易见性驳回原始申请的理由。另一个例子是发明名称为“System and Method for Monitoring a Pressurized System”的专利 No. 7051570 (IPC G01L 测量力,应力,扭矩,功,机械动力,机械效率,或流体压力),该专利被专利审查员以发明名称为“Cellular Digital Packet Data Network Transmission System Incorporating Cellular Link Integrity Monitoring”的专利 No. 7051570 (IPC G08B 发射信号) 驳回。

[0017] 第四,专利文档通常包括技术术语(例如,CMOS——互补金属氧化物半导体和 PMOS——阳极金属氧化物半导体或便携式媒体操作系统)。语义相似性将不能识别不同技术短语之间的关系,因为这些术语很可能几乎没有文字相似性。例如,基于专利审查员的评价,发现技术短语 PMG (永磁电机) 和 BLDC (无刷直流控制器) 尽管在文字上不同,但是在科学含义上是相关的。

[0018] 其他的背景技术包括 US 专利 4839853 “使用潜在语义结构的计算机信息搜索 (Computer information retrieval using latent semantic structure)”。公开了检索文字数据对象的方法。通过假设在数据对象中字的用法中存在隐含的、潜在的语义结构,在统计域中对信息进行处理。对该潜在结构的估计被用来表示和搜索对象。用户查询在新的

统计域中被重新表达,然后在计算机系统中被处理,以提取隐含的含义,从而响应于查询。

[0019] 美国专利 5297039——基于关键字匹配和关键字进行定位的文本搜索系统 (Text search system for locating on the basis of keyword matching and keyword)——教导了一种文本信息提取装置,其从文本提取分析网络并将其存储在数据库中。分析网络由行构成,每个行包括从文本提起的要素和关系。分析网络经由同义词 / 相近词 / 近义词并经由互补模板而补码,并且其行经由概念模板而被加权。文本相似性匹配装置基于词、词对和行的协议来判断输入和数据库分析网络的相似性。文本搜索系统将文本和从中准备的互补术语列表存储在相应的数据库中。查询以分析网络的形式被输入,从分析网络中提取关系和关键词集。在关于从每个输入查询中提取出的关键词而搜索存储在数据库中的文本和互补术语列表后,便确定了关系和关键词集的协议。

[0020] 美国专利 5963965——文本处理和搜索系统和方法 (Text processing and retrieval system and method)——教导了一种基于内容的用于文本处理和搜索的系统和方法,其中基于内容来处理多份文本,从而为每份文本产生索引,所述索引包括表示该份文本的内容的短语列表。所述短语被集成成组在一起,以产生基于短语的关系度的群集,并且产生分层结构,该分层结构包括多个图 (map),每个图对应于预定的关系度,所述图以图表形式描绘预定关系度的群集,并且包括多个节点 (每个节点表示群集),以及连接相关的节点的多个联系。所述图被显示给用户,用户选择图上的特定群集,并且基于用户选择的群集从该份文本中提取一部分文本。

[0021] 美国专利 No. 5991751——用于专利中心的系统、方法和计算机程序产品 (System, method, and computer program product for patent-centric)——教导了一种用于处理数据的系统、方法和计算机程序产品,用于处理数据的系统、方法和计算机程序产品在此被描述。该系统保持第一专利数据库,以及公司实体所关注的非专利信息的第二数据库。该系统还保持一个或多个组。每个组包括来自第一数据库的任何数量的专利。在接收到适当的操作人员命令后,该系统自动地结合来自第二数据库的非专利信息来处理所述组之一中的专利。因此,所述系统执行以专利为中心和以组为导向的数据处理。组还可以包括任何数量的非专利文档。组可以是基于产品的、基于人的、基于公司实体的,或者用户定义的。其他类型的组也是可被包括,例如临时的组。

[0022] 美国专利 No. 6298327——用于授权发明公开的专家支持系统 (Expert support system for authoring invention disclosures)——教导了一种计算机执行的专家支持系统,用于授权发明公开,并用于评价所公开的发明的可能的专利性和适销性。该系统至少包括计算机、输入设备、输出设备以及软件程序。该软件程序利用以面向对象的设计过程来开发,并以诸如 C++ 等面向对象的计算机语言来实现。该系统促进了发明特性的交流并以多种格式实现了发明公开的输出,包括专利申请的格式。

[0023] 美国专利 No. 6363378——对信息搜索系统中的查询反馈项的进行排列 (Ranking of query feedback terms in an information)——教导了一种信息搜索系统,其对用户输入查询进行处理,并识别查询反馈,包括对查询反馈的进行排列,以促进用户重新编排新的查询的格式。包括多个描绘术语概念的节点的知识库被设置成反映节点之间的概念相近性。该信息搜索系统处理查询,识别与查询相关的标题以及查询反馈项,然后利用对应的术语概念将标题和反馈项两者联系到知识库的节点。基于标题从知识库中选择至少一个焦点

节点,以确定焦点节点与查询反馈节点之间的概念相近性。基于与焦点节点的概念相近性对查询反馈项进行排列。还公开了识别来自在查询反馈处理中使用的大量文档的主题的内容处理系统。

[0024] 美国专利 No. 6452613——评估新技术的自动打分工具用系统和方法 (System and method for an automated scoring tool for assessing new technologies)——教导了一种用于评价发明提交的自动发明提交和打分工具的装置和方法。该系统包括服务器系统和多个服务器系统。该服务器系统通过网络连接将提交问卷呈送给用户系统处的提交者。用户完成问卷,问卷被返回到服务器系统以便进行处理。服务器系统对答案进行处理,以基于专利性和至少一个其他参数(例如影响或价值)提供对该提交的定量评价。评价器系统处的评价器能够看到发明提交的定量评估的显示。评价器还能够在状态综述页面上看到多个发明提交的结果。提供了状态综述页面、单个问卷与单个评估显示之间的联系。

[0025] 美国专利 No. 6542889——基于概念进行索引的用于相似性文本搜索的方法和装置 (Methods and apparatus for similarity text search based on conceptual indexing)——教导了一种进行概念相似性搜索的方法,该方法包括以下步骤:从要被使用在概念相似性搜索中的一个或多个文档中产生一个或多个概念字链;利用所述一个或多个概念字链构建文档的概念索引;以及利用该概念索引评价相似性查询。评价步骤优选地将从搜索产生的最接近的文档中的一个或多个、所述一个或多个文档中的一个或多个匹配的字链、以及所述一个或多个文档的一个或多个匹配的主题词返回。

[0026] 美国专利 No. 7054856——利用技术领域词汇绘制专利地图的系统和方法 (System for drawing patent map using technical field word and method)——公开了一种利用技术领域词汇绘制专利地图的系统和方法。在该系统和方法中,通过计算重要词的权重值提取将要用于绘制专利地图的词,然后将该提取出的词与绘制专利地图的专利相匹配,所述重要词是通过将不必要的词从专利数据中除去而得到的。

[0027] 美国专利申请 No. 11/697,447——增强的专利现有技术搜索引擎 (Enhanced Patent Prior Art Search Engine)——教导了一种搜索引擎,其被配置成搜索文档的数据库并将搜索结果提供给终端用户。该搜索引擎可以配置为向终端用户提供终端用户提交的搜索查询中的术语的同义词列表,并允许终端用户识别这些应被包括在搜索引擎中的同义词。作为替代设置或作为补充,该搜索引擎可以配置成向终端用户提供调查问题,对这些问题的答案可以被用来进一步地限定搜索查询。数据库可包括与数据库中的特定文档相关的注解和/或广告。

[0028] 美国专利申请 No. 11/745,549——用于通过网络分析语义文档的系统和方法 (Systems and Methods for Analyzing Semantic Documents Over a Network)——教导了一种通过如下方式来处理知识产权 (IP) 的系统和方法:提供自动代理来为用户执行一个或多个搜索,以定位与知识产权 (IP) 权益相关的一个或多个文档,所述自动代理评估用户属性以确定用户的 IP 权益并识别一个或多个 IP 文档,每个 IP 文档都具有响应于 IP 权益的标记;将自动代理定位的一个或多个文档排列;以及显示由自动代理所定位的所述一个或多个文档。

[0029] 美国专利申请 No. 11/809,455——基于概念的跨媒体索引和语言搜索 (Concept based cross media indexing and retrieval of speech)——教导了语言文档(包括但不

限于记录书、声频广播、记录谈话) 的内容的索引、搜索和搜索, 这是通过发现并搜索在概念水平与查询术语相关的语言文档而实现的, 即使语言文档不含有口语的(或文字上的) 的查询术语也是如此。使用基于概念的跨媒体信息搜索。从文档的训练集来构建术语-音素/文档矩阵。然后, 文档被添加到从训练数据构建的矩阵。使用奇异值分解法来从术语-音素/文档矩阵计算向量空间。结果是较低维度的数字空间, 在该空间下, 术语-音素和文档向量在概念上作为最近邻相关联。查询引擎计算该空间中查询向量和所有其他向量之间的余弦值, 并返回具有最高余弦值的术语-音素和/或文档的列表。

[0030] 美国专利申请 No. 11/812, 135——用于分析专利价值的系统和方法(System and method for analyzing patent value)——教导了至少一个示例性实施例, 并公开了用于评价法律文档(例如, 与专利相关的文档) 的价值的系统、计算机程序产品和方法。根据至少一个示例性实施例, 潜在语义分析(“LSA”) 搜索引擎能够搜索专利相关文档的数据库, 以识别“N”个因为与目标文档相关而因此被认为是专利相关文档的专利相关文档, 并且能够按照相关的所识别的专利相关文档的指示器(indic) 来比较并打分目标专利相关文档的指示器。至少一个示例性实施例利用法律、商业和/或技术因素来评价专利相关文档值的多个指示器。

发明内容

[0031] 根据本发明的一些实施例的一个方面, 提供了一种方法, 其使用专利审查机构所提供的搜索和审查报告作为学习数据库来训练搜索引擎, 从而了解科学和技术概念之间的关系。使用这种学习数据库进行训练后, 搜索引擎然后能够执行有意义的搜索, 不仅仅是专利文献的搜索, 而是普遍的技术和科学文献的搜索。

[0032] 根据本发明的一些实施例的一个方面, 提供了一种用于基于相关的科学或技术概念进行搜索的搜索引擎, 包括:

[0033] 学习模块, 其用于基于技术短语在相关文档中的出现率来了解技术短语之间的关系, 从其中由相关短语的组合形成概念; 和

[0034] 搜索模块, 其用于基于存在于查询文档中的概念在相关文档中的出现来搜索与查询文档相关的文档, 所述学习模块基于文档的训练集和文档之间的关系来进行学习。

[0035] 在一个实施例中, 训练集包括专家提供的指出相关的文档的联系。

[0036] 在一个实施例中, 学习模块配置为基于在概念内所组合的技术短语的各个文档中的出现来量化所述概念与各个文档的相关度。

[0037] 在一个实施例中, 学习模块配置为基于在两个文档之间共同具有相对较高的相关度的多个概念来量化两个文档之间的关系, 所述量化能够与指定相关文档的阈值一起使用。

[0038] 在一个实施例中, 学习模块配置为识别训练集的所有文档中的技术短语, 并按照总的共同出现率来计算在相关文档中的共同出现率, 从而形成组合。

[0039] 在一个实施例中, 学习模块配置为识别训练集的所有文档中的技术短语, 并按照总的共同出现率来计算在相应的相关文档中的共同出现率, 从而形成组合。

[0040] 在一个实施例中, 训练集包括一组专利文档, 并且文档之间的关系包括从专利审查文件的衍生。

[0041] 在一个实施例中,学习模块包括基于专利审查文件中记录的专利性结果使两个专利文档相关的文档分析能力。

[0042] 在一个实施例中,专利审查文件包括关于第一篇文档的专利审查报告,并且其中对技术短语的识别基于专利审查报告所涉及的所述第一篇文档的版本。

[0043] 在一个实施例中,专利审查报告被分析,以识别驳回(rejection)或反对(objection),并识别对提供驳回或反对的基础的其他专利申请的参考。

[0044] 在一个实施例中,学习模块包括用于识别出现在训练集中的技术短语的技术短语识别单元。

[0045] 在一个实施例中,搜索模块配置为基于技术短语在相应的目标文档中的出现来针对查询文档对目标文档进行排列,所述技术短语属于在查询文档中发现的概念中的一些。

[0046] 根据本发明的第二方面,提供了一种搜索专利文献的方法,包括:

[0047] 获取包括专利申请和相应的专利局报告的专利文献的训练集;

[0048] 使用专利局报告将某些专利申请组合在一起,作为相关的专利申请;

[0049] 如果技术术语在相关的专利申请中比在作为整体的训练集中出现得多,则将技术术语组合在一起,作为相关的术语;以及

[0050] 分别使用相关的术语来搜索技术文献。

[0051] 根据本发明的第三方面,提供了一种基于相关的科学或技术概念进行搜索的搜索方法,包括:

[0052] 基于技术短语在训练集的相关文档中的出现率来了解技术短语之间的关系,从而形成相关短语的概念;以及

[0053] 基于存在于查询文档中的概念的出现搜索查询文档的相关文档。

[0054] 在一个实施例中,训练集包括专家提供的指定相关的文档的联系。

[0055] 在一个实施例中,所述了解包括识别训练集的所有文档中的技术短语以及针对总的共同出现率来计算在相关文档中的共同出现率,从其形成组合。

[0056] 在一个实施例中,训练集包括一组专利文档,并且文档之间的关系包括从专利审查文件的衍生。

[0057] 在一个实施例中,所述了解包括基于专利审查文件中记录的专利性结果使两个专利文档相关的文档分析。

[0058] 在一个实施例中,所述了解包括用于识别出现在训练集中的技术短语的技术短语识别阶段。

[0059] 在一个实施例中,所述搜索包括基于技术短语在相应的目标文档中的出现来对照查询文档对目标文档进行排列,所述技术短语属于在查询文档中发现的概念中的一些。

[0060] 所述搜索方法可包括计算查询文档与目标文档之间的度量,所述度量能够在识别应用风险中使用。

[0061] 在一个实施例中,对相关文档的搜索在可获得的技术文献中进行。

[0062] 除非另外限定,否则,此处所使用的所有技术和/或科学术语都具有如本发明所属领域的普通技术人员所普遍理解的相同的含义。尽管与此处所描述的方法和材料相似或等同的方法和材料能够在本发明的实践中或本发明实施例的测试中使用,但是下面仅仅描述示例性方法和/或材料。在相冲突的情况下,包括限定在内的专利说明书将占主导地位。

此外,材料、方法和示例仅仅是示例性的,并不意在必然是限定性的。

[0063] 本发明的实施例的方法和 / 或系统的实施可以涉及手动地、自动地、或者手动自动相结合地执行或完成所选择的任务。另外,根据本发明的方法和 / 或系统的实施例的实际仪器和设备,能够利用操作系统通过硬件、通过软件、或通过固件、或者通过硬件、软件和固件的组合来执行多个所选任务。

[0064] 例如,根据本发明的实施例的用于执行所选任务的硬件可以实施为芯片或电路。作为软件,根据本发明的实施例的所选任务可以实施为多个软件指令,所述软件指令由计算机利用任何适当的操作系统来执行。在本发明的示例性实施例中,根据此处描述的方法和 / 或系统的示例性实施例的一个或多个任务由数据处理器执行,例如用于执行多个指令的计算平台。可选地,所述数据处理器包括用于存储指令和 / 或数据的易失性存储器,和 / 或用于存储指令和 / 或数据的非易失性存储器,例如磁性硬盘和 / 或可移动媒介。可选地,还提供网络连接。可选地,还提供显示器和 / 或诸如键盘或鼠标的用户输入设备。

附图说明

[0065] 此处参照附图通过仅为示例的方式描述了本发明的一些实施例。现在具体参照附图,需要强调的是,特定的细节是通过示例的方式示出的,用于本发明的实施例的示例性论述的目的。关于这一点,结合了附图的描述将使可以如何实施本发明的实施例对本领域普通技术人员显而易见。

[0066] 在附图中:

[0067] 图 1 是从 1985 至 1986 年,每年的专利提交的图表,显示了长期以来一贯的增长态势;

[0068] 图 2 是一百多年以来,不同专利局的专利提交的图表;

[0069] 图 3 是简化的框图,示出了根据本发明的实施例的具有学习模块和搜索模块的搜索引擎的概念细节;

[0070] 图 4 是示出了图 3 的学习模块的部件的简化框图;

[0071] 图 5 是示出图 3 的搜索引擎的学习阶段的流程图的简化图;

[0072] 图 6 是示出图 3 的搜索引擎的搜索阶段的简化流程图;

[0073] 图 7 是示出所分析的文档之间的关系如何可导致概念的训练集的图;

[0074] 图 8 是示出根据本发明的实施例,文档之间的相关度及其量化的概念的图;

[0075] 图 9 以更多的细节示出了图 5 的流程;

[0076] 图 10 以更多的细节示出了图 9 的流程的一部分;

[0077] 图 11 是流程图,示出了使用在图 9 和图 10 中确定的量来进行搜索;

[0078] 图 12 示出了图 1 的搜索引擎的用户界面;

[0079] 图 13-17 是试验结果图,在所述试验中,根据本发明实施例的概念搜索被与现有技术的文字搜索作比较;

[0080] 图 18A 是流程图,示出了根据本发明的优选实施例,用于为待搜索的专利文档提供概念空间并将包括概念词的同义词的词映射到该空间的程序;

[0081] 图 18B 示出了用于提取术语以构建概念空间的一系列模型;以及

[0082] 图 19 是图 18A 的细节图,示出了用于同义词选择的可能方法。

具体实施方式

[0083] 本发明在其一些实施例中涉及搜索引擎和方法,更具体地但非唯一地,涉及特别适用于专利文献的搜索引擎和方法。

[0084] 本方法可使用专利审查机构提供的搜索和审查报告作为学习数据库以训练搜索引擎,从而了解科学和技术概念之间的关系。使用这种学习数据库进行训练后,搜索引擎然后能够执行有意义的搜索,不仅仅是对专利文献的搜索,而且是普遍的技术和科学文献的搜索。

[0085] 所述方法涉及对诸如机构所提供的搜索和审查报告等文档的结构进行理解,从而能够假定该报告的主题的专利或申请的权利要求与在该报告中所引用的对比专利或申请的文本之间的关系。该假定的关系然后被用来指定出现在该文档两者中的短语的相关性。随着上述过程在成百上千个文档上被执行,便建立起两个短语相关的概率的数据库。

[0086] 在学习阶段之后,数据库变能够被用在搜索中,以排列文档之间的关系,使得具有紧密相关主题的文档在搜索结果具有高的排名。因此,专利搜索阶段指定了语义关系,并与用于相同关键词的需要没有依附关系。

[0087] 在详细说明本发明的至少一个实施例之前,需要理解的是,本发明在其应用中并不一定限于下面的描述中阐述的和/或附图和/或示例中示出的部件和/或方法的构造和设置的细节。本发明能够以其他实施例或以多种方式来实施或执行。

[0088] 现在参照附图,图1和图2作为背景来参照。

[0089] 图3是简化的框图,示出了本发明的概念性实施例。搜索引擎10包括学习模块12和搜索模块14,搜索引擎10用于准备数据库以及随后基于相关的科学或技术概念进行搜索。

[0090] 学习模块12基于技术短语在相关文档中的出现率去了解技术短语之间的关系,并基于此将概念组合在一起。更具体地,当将相关文档中的共同出现率与背景共同出现率相比较时,便获得了比值,这将在下文更详细地论述。如果在相关文档中的出现率高于背景出现率,则说明比值是正的,两个技术短语可以被联系在一起以形成概念,即对相关短语进行组合。在一个实施例中,联系可以是绝对的有/没有联系,其中无论何时比值超过预定的阈值,便进行联系。可替代地,可以基于获得的实际比例来定义相对联系。所述比值可以从技术文档的训练文档数据库16以及哪些文档与哪些其他文档相关的列表来计算。

[0091] 如将在下面更详细地论述地,哪些文档与哪些其他文档相关的列表可以由本领域的专家来提供。如将要注意到的,专利文献在这方面具有特别的意义,因为专利要被审查,并且专利审查员(他们是相关领域的专家)所提供的审查文档可以用作专家输入以限定初始关系。因此,训练集16包括两个部分:专利申请文献18和专利审查文档20,专利申请文献18提供技术文献本身,专利审查文档20包括专家提供的联系,表明哪些专利申请与哪些其他专利申请相关。

[0092] 搜索模块14然后使用学习模块发现的概念——训练集22。当被提供查询文档时,该训练集22查询技术短语,然后搜索短语本身或已经在学习模式中以概念的形式与该技术短语联合的其他短语。然后,根据属于该概念的技术短语的出现率来排列目标文档,并且将所排列的文档中的一些或全部呈送给用户。

[0093] 现在参照图 4,其更详细地示出了学习模块 12。学习模块 12 包括文档分析单元 30,其分析专利审查文献和专利申请文献两者。

[0094] 文档分析单元包括光学字符识别 (OCR) 单元 32,其从互联网上一般能够获得的图像来读取文档文本。关系发现单元 34 然后分析专利审查文献,以从审查文档的文本识别审查员指出的相关文档。在识别这些文档时,专利审查文献趋于遵从特定的形式,使得学习模块所需的分析能力变的更容易。将在后面更详细地论述所述分析。

[0095] 文档分析单元可以进一步包括技术短语识别单元 36,其用于识别出现在专利申请文档的训练集中的技术短语。

[0096] OCR 之后的文档分析过程示出在图 5 的流程图中,现在参照图 5。如图 5 所示,专利申请文档的分析发现技术短语,使得每个专利申请文档被化为一系列的技术短语。专利审查文档然后被分析,从而发现相关的专利文档。

[0097] 然后,每个技术短语被依次提取,并测试其与其他文档中的其他技术短语的共同出现。在相关文档中的共同出现被标注为 A,并且在全部文档上的共同出现被标注为 B。比值 A/B 被对照例如阈值测试,如果测试证明是正,则两个技术短语被交付给单个概念。

[0098] 现在参照图 6,图 6 是简化图,示出了使用图 5 中获得的概念的搜索程序的流程。文档被用户作为初始查询输入。该文档具有其自己的被提取的技术短语。所提取的技术短语可以属于或者不属于在训练阶段所识别的概念。在属于所述概念的地方,该概念的所有技术短语被用来制定用于实际使用的搜索查询。然后,搜索查询从所识别的概念中,根据技术短语的出现频率来排列数据库中的文档。

[0099] 对于不构成延伸概念的一部分的技术短语,可以根据熟练的用户的喜好,简单地添加各个技术短语(因为这些技术短语将搜索查询),或者可以忽略这些技术短语。

[0100] 因此,搜索模块基于各种技术短语在不同目标文档中的出现来对照查询文档排列目标文档进行,所述各种技术短语属于在查询文档中发现的不同概念。

[0101] 如所论述的,本发明的实施例提供评价专利文档的科学相关度的方法。该方法使用相关科学概念的训练集 22 来计算不同专利文档的科学相似性。该训练集通过关于不同科学概念的科学相关度的外部信息,将不同的科学概念联系在一起,并参照上文在训练过程中形成。

[0102] 训练集包括专利审查机构(例如,美国专利和商标局以及欧专利局)的专利审查员的数千个搜索和审查报告。这些报告被自动地和手动地分析,以搜索关于由专利审查员所确定的科学文档的相关度的信息,这些专利审查员是特定技术领域的有经验的且训练过的专家。除了审查员搜索报告数据集之外,还可以使用搜索报告所涉及的原始的驳回前(pre-rejection)专利申请。该组合允许以不仅仅基于关键词相似性,还基于高度训练过的科学家的专业意见中所阐释的含义的方式,将来自专利申请的科学概念与现有技术联系在一起。

[0103] 在过去的十年,基于内容的自动搜索引擎(被称为信息搜索)已经在学术和商业应用中深受欢迎。信息搜索(IR)的模型检查文档之间的关系,如从每个文档的文字特征中所推断的。一种常用的技术是向量空间模型,也称为词袋表示。根据该技术,每个文档由特征的索引向量表示,所述特征通常为所出现的或与文档相关的词。特征通常被加权,其中权重随着特征在该文档中出现的频率与特征在所有其他文档中出现的频率之比而增加。然后

基于每个文档的加权的词袋表示的相似性来计算文档的相关度。

[0104] 了解文档的概念含义也称为文本分类 (TC)——基于它们的内容的自然语言文本到一个或多个预定类别的分配。文本分类旨在将文档归类到预定的标签集,其中每个标签与文档中使用的词集相关联。每个文档表现为能以数学方式操控的代表术语的集合。然后,到类别和概念中的分类基于代表向量的相似度。

[0105] 信息搜索领域的专家所面临的问题是如何基于文档的文字特性了解文档的内容相关度。语义相似性是用来测量文档的相关度的最简单的技术 (Baeza-Yates 和 Ribeiro-Neto,1999):用来评价文档相关度的唯一标准是每个文档中使用的词之间的重合程度(每个词通常通过该词在该文档和所有其他文档中使用的相对频率来加权)。词袋方法存在两个主要缺点:(1) 在文字特征方面的高维数;和(2) 对精确文字重合的完全依赖。

[0106] 最先进的算法属于 LSA 家族。LSA 方法通过利用来自线性代数的技术形成代表概念的向量来减小词袋的维数。其他通过降低它的维数来改进词袋的尝试包括 n-gram 方法和统计及语言模型。LS 方法注意到,在专利文档中,词以技术短语群出现是非常普遍的(见上面的例子)。因此,除了出现在专利文档中的常规词之外,还可以根据 LSA 方法来审查词群,而在评价概念相关度时它们的各个术语无关。

[0107] 与降低词的维数对比,使用来自外部源的背景词的特征生成技术扩展了词袋。例如,Sahami 和 Heilman(2006) 提出了通过将基于文档的文字特征的查询发送到网页搜索引擎来比较短的文档,并且然后计算对于每个查询从网页检索到的词的向量之间的相似性。按照这条思路,Strube 和 Ponzetto(2006) 以及 Gabrilovich 和 Markovitch(2007) 对出现在查询中的词集,导出了来自维基百科 (Wikipedia) 的信息,旨在引入附加的背景词,这些背景词将为原始文档中的词的概念含义提供线索。与 LSA 相似,主要局限在于缺乏关于词和技术短语的概念相关度的信息。

[0108] 不同的技术使用训练集和机器学习算法来检查文档之间的关系。机器学习技术已经得到了发展,并已执行在下列信息搜索任务上:迁移学习 (transfer learning) 和半监督学习,EM 算法,潜在语义核和广义向量空间模型。机器学习算法被用来近似将文档的文字特征映射到文档的类别的功能。这种近似是利用训练集来执行的,所述训练集包括其类别已预知的文档。通常的情况是,训练集由专业的个人(专家)来集合。然后,机器学习技术被执行,以了解已分类文档之间的文字联系,从而形成文档和类别之间的映射函数,从而被用来自动地对训练集以外的文档进行分类。

[0109] 形成系统地覆盖自然语言的概念含义的训练集是极为复杂的任务。这种训练集的一个例子是 WordNet,其是将词组合成概念的特殊的词典。除了将词组合成概念之外(同义词集 (Synsets)),WordNet 还提供关于概念之间的等级关系的信息。基于该等级关系,开发了多种技术以测量词的概念相关度。WordNet 的发展多半是手工的,并且耗费了大约 20 年。其主要缺点仍然是词的覆盖相对较低,另一个缺点是分配给单个词的概念的多重性。

[0110] 本发明的实施例基于专家训练集机器学习系统 (ETSM),其中在受监督的学习过程中从相关专利文档的成千个专家评价报告来了解文档的科学相关度。

[0111] 本实施例提供技术文档搜索引擎,更具体地但非唯一地,提供专利文档搜索引擎。它们建立在新颖的数据集之上,所述数据集提供关于专利文档中的词和技术短语之间的科学关系的信息。与语义相似性引擎不同,本实施例不比较专利文档的文字重合,而是比较可

能与它们相关的科学概念的重合。

[0112] 本实施例开发了一种基于从专利审查机构的专利审查员的搜索报告提取的显著相关的文档的训练集的机器学习系统,该专利审查机构例如为美国专利商标局 (USPTO) 和欧洲专利局 (EPO)。搜索报告包括关于代表相关或相同技术的不同文档的独特信息。当专利审查员驳回具体的权利要求时,她需要提供对其驳回决定的解释,其中该解释通常引用相关的现有技术文档。一旦审查员能够将来自专利申请的权利要求与现有技术文档相关,则可以假定所涉及的两个文档之间的概念关系。该概念关系此时包括在被驳回的专利文档的说明书中使用的词集和在引用的现有技术的说明书中使用的词集。

[0113] 本实施例因此使用专利审查员提供的搜索和审查报告作为概念联系知识库。此处的例子中使用的审查员报告(也称为非最终驳回)可以直接在网络上从 USPTO 的文件袋 (File Wrapper) 部分获得,该文件袋部分追踪并记录专利申请过程。仅仅来自原始提交的专利申请的信息被认为是对提供训练集有效的。由于权利要求在后来被修改以区别于现有技术,所以授权的专利不再表现出相同的关系。因此,从授权专利的文本中了解用来描述同一发明的不同语义的可能性较小。另一方面,在专利申请过程期间,专利审查员具体地判断哪些权利要求在概念上与相关的现有技术相似,并且正是该审查员的判断构成系统的专家输入,从而提供科学相关度的识别源。

[0114] 在驳回的情况下,专利审查员指出被驳回的权利要求的编号和驳回的理由。USPTO 审查员最常用的驳回理由是新颖性 (102) 和显而易见性 (103) :

[0115] 根据 U. S. C. 第 35 段 102(b) 或 102(e) 的驳回

[0116] (b) 在美国的专利申请日之前一年以上,本发明已经在本国或外国被授予专利权,或被描述在本国或外国的出版公开物上,或者在本国公开使用或销售。

[0117] (e) 本发明已经描述在 (1) 由他人在专利申请人的发明之前在美国提交,在 122(b) 节下公开的专利申请中,或者 (2) 在由他人在专利申请人的发明之前在美国提交的专利申请上授权的专利,不同的情况是:在 351(a) 节中定义的条约下提交的国际申请只有在指定了美国并且在该条约的 21(2) 条下以英语语言公开的情况下,应具有在美国提交的申请的该子节的效力。

[0118] 根据 U. S. C. 第 35 段 103(a) 的驳回

[0119] 虽然发明没有如该标题的 102 节中所阐述地那样被等同地公开或描述,但如果寻求专利保护的主体与现有技术之间的差别使得所述主题整体上在实现发明时对所述主题所属领域的普通技术人员是显而易见的,则不能获得专利权。专利性不应以实现本发明的方式而否定。

[0120] 除了所参照的现有技术之外,专利审查员还可以提供关于该申请和可能导致驳回决定的现有技术中的不同技术方面的详细论述。

[0121] 被驳回申请和专利审查员所引用的现有技术的组合形成概念关系。遵照词袋方法,该概念关系通过其文字特征来表示。然后,执行学习算法来估算观察到隐含的概念的概率,作为基于文字表示的条件概率。

[0122] 现在参照图 7,图 7 是简化的框图,其示出了如何结合被驳回申请的文本和引用的现有技术的文本从搜索或审查报告来得到概念的训练集。用于专利文档之间的相关度的信息源是专利审查员提供的搜索或审查报告。从每个报告来识别被专利审查员驳回的原始权

利要求和引用的现有技术。被驳回申请和现有技术的组合形成概念。

[0123] 非最终驳回：示例

[0124] 下面是对于专利申请 No. 10/204, 119 和专利申请 No. 09/451, 127, 来自专利审查员的搜索报告的权利要求驳回的两个示例。第一个示例示出了基于新颖性的两个权利要求的驳回：这些权利要求已经公开在专利 No. 5406044 中。这意味着被驳回的权利要求中公开的科学与现有技术高度相关。第二个示例示出了基于显而易见性的驳回：在驳回中公开的创造性步骤相对专利 No. 5764485 不足以显著地证明其专利性。因此，同样在此，被驳回的权利要求的技术内容与其现有技术太相似。

[0125]

美国专利申请 10/204,119 的搜索报告的一部分

权利要求驳回 35 USC § 102

下面是对 35 U.S.C 102 的适用段落的引用，其构成该审查意见通知书中所做的根据该条款驳回的基础：专利权应当被授予，除非：
(b) 在美国的专利申请日之前一年以上，本发明已经在本国或外国被授予专利权，或被描述在本国或外国的出版公开物上，或者在本国公开使用或销售。

根据 35 U.S.C 102 (b)，权利要求 1 和 2 因被 Killian 等人（5, 406, 044）预见到而驳回。Killian 等人的专利公开了螺栓焊接头、线性装置以及电磁长度测量系统。关于权利要求 2 中的用语“渐增长度测量系统”，Killian 等人的系统明显能够测量渐增长度，并因此能够“定义”为渐增测量系统。如果申请人意在使权利要求 2 针对的是一种具有渐增输出的长度测量系统，应当注意，Killian 等人的系统因其能够连接到计算机而公开了该特征。这种数字连接将固有地涉及渐增长度测量输出，因为来自元件 34 的模拟输出必须被转化为数位，由此满足权利要求用语的解释。

[0126] 表 1：美国专利申请 No. 10/204, 119 的审查报告（非最终驳回）的摘录。该申请的权利要求被专利审查员以缺乏新颖性为由驳回——该权利要求中公开的发明已经被授予了专利权（专利 No. 5406044）。对于该示例的概念关系包括原始被驳权利要求的文字特征和现有技术的文字特征。

[0127]

美国专利申请 No. 09/451,127 的搜索报告的一部分

权利要求驳回 35 USC § 103

下面是 35 U.S.C 103(a)的引用，其构成该审查意见通知书中阐述的所有显而易见性驳回的基础：

(a) 虽然发明没有如该标题的 102 节中所阐述地那样被等同地公开或描述，但如果寻求专利保护的主题与现有技术之间的差别使得所述主题整体上在实现发明时对所述主题所属领域的普通技术人员是显而易见的，则不能获得专利权。专利性不应以实现本发明的方式而否定。

根据 35 U.S.C 103(a)，权利要求 1 因其相对于 Lebaschi (US 5,764,485)不能被授予专利权而被驳回。

Lebaschi 公开了一种印刷布线板，其包括：多个间隔开的电路层，所述电路层通过盲孔/通孔电连接，其中在导电金属层的盲孔/通孔上方形成表面接触垫板。尽管 Lebaschi 没有具体地表述通过在盲孔/通孔上方沉积来形成接触垫板，但是这对本领域普通技术人员在通过沉积材料形成垫板而实施本发明时是显而易见的，因为这是本领域的公知技术。

[0128]

[0129] 表 2：该图来自专利申请 No. 09/4581, 127 的审查报告（非最终驳回）。该权利要求被专利审查员以显而易见性为由驳回——该权利要求中公开的发明与另一个已经授权的发明（专利 No. 5764485）太相似。对于该示例的概念关系包括原始被驳权利要求的文字特征和现有技术的文字特征。

[0130] 词和技术短语之间的概念关系

[0131] 本方法基于词和技术短语在文档中出现的频率在词和技术短语之间形成联系，所述文档已经被高度专业的该领域专业人员（专利审查员）归类成在技术上相关。将申请 - 现有技术对限定为概念，并且应当注意，如果使用相同的现有技术来驳回不止一件申请，那么这种概念可以包括不止一件申请。对于训练集外的每个专利文档，构建概念的加权向量，其中更可能与专利相关的概念接收较高的权重。基于可能与它们相关的关系概念的向量的相似度来进行专利文档之间的比较。因此，如果预测的概念向量表现出较大的相似性，则认为不同的专利在技术上更加相关。

[0132] 在详细描述根据本实施例的算法的执行和性能之前，先提供核心程序的综述，如图 9 所总结。将在下面更详细地论述图 9。如图 8 所示，训练集包括所有的关系概念。训

训练集提供词的集合和该词出现在其中的技术上相关的文档的集合。问题是在观察出现在专利文档 j 中的词 k 的条件下, 计算概念 i 与新的专利文档 j (位于数据集外) 相关的条件概率。程序在图 10 中概略示出。在对每个概念-词对计算了条件概率之后, 转为将每个专利文档表示为概念袋 (bag-of-concept)。用 c_i 来表示发明性概念 i 。记住, c_i 的词袋表示包括所有出现在被驳回的权利要求和现有技术中的词。该向量中的每个词被加权, 其中权重表示所述词关于它与哪个关系概念相关而提供的信息内容。基于表示每个专利和这些词接收的概念权重的词的集合, 概念被分配分类状态值 (CSV), 如示出的。CSV 表示来自训练集的文档的实际相关度到每个文档中的文字特征的预测 (projection)。较高的 CSV 意味着该概念更可能与专利相关。更正式地, CSV 能够定义为对每个对分配值, $\langle d_j, c \rangle \in D \times C$, 其中 d_j 是专利 (位于训练集外), c_i 是概念, D 是数据库中的专利的总数, 而 C 是训练集中的概念的总数。

[0133] 如图 10 所示, 训练集包括成对的被驳回权利要求和相关现有技术。对于每个文档 (位于训练集外) 和概念, 分配分类状态值 (CSV)。CSV 表示对来自训练集的文档的实际相关度在专利文档中的文字特征上的预测 (projection)。假定如果专利对的 CSV 向量更为相似, 则它们具有较强的科学关系。

[0134] 现在更详细地考虑各个构成部分。

[0135] 构建训练集

[0136] 本实施例的搜索算法可使用一组技术相关的文档 (从这些文档了解词和技术短语的科学相关度), 并且形成概念的训练集。该训练集从专利申请本身和由专业的专利审查员所评价的它们之间的相互联系来构建。该部分更详细地论述构建训练集的步骤。

[0137] 阶段 1: 原文档的准备

[0138] 现在参照图 9, 并如所提及的, 所述算法使用来自提交给美国专利商标局 (USPTO) 的专利的申请过程的信息。能够对提交给欧洲专利局 (EPO) 的专利执行类似的方法。

[0139] I. 下载申请文档

[0140] a. 从 USPTO 的文件袋部分下载名为“非最终驳回”的审查报告——阶段 1。如果存在不止一个的非最终驳回, 则第一个是优选的。

[0141] b. 还从 USPTO 的文件袋部分下载审查报告所涉及的原始申请权利要求。原始权利要求不包括对反对意见进行的修改, 这些修改可能涉及对关键联系材料的移除。

[0142] II. 将文档从图像格式转化为文本格式

[0143] a. 审查报告由 USPTO 扫描, 并作为图像文件上传到网页上。因此, 文档分析的先决条件是使用 OCR 软件将所有文档转化为文本格式, 如 PDF。

[0144] b. 然后可将申请文档从 PDF 格式转化为 TXT 格式或类似格式。

[0145] c. 然后可将文档分成两个单独的部分: 包括审查报告的驳回决定和包括原始申请的权利要求的权利要求。

[0146] III. 自动的文档分析

[0147] a. 驳回决定部分被扫描, 以识别驳回决定的样式。这些样式包括驳回声明、驳回理由和专利编号或构成驳回基础的文档的其他标识。

[0148] b. 阶段 (a) 的输出被组织到 100 个驳回的列表中 (该列表中的每个行包括申请号、参考的专利号、驳回权利要求的编号和驳回理由的代码)。每个列表被添加到单独的文

文件夹中,该文件夹还包括从 USPTO 下载的用于手动检查的原始图像文件。

[0149] IV. 手动文档分析

[0150] (i) 自动文档分析(阶段 2)之后是手动分析(阶段 3)。手动分析的主要目的是识别自动过程中的错误。常见的错误包括:

[0151] a) 来自文档的标题和脚注的不相关的文本被包括进来(例如,序列号、文件名等)。

[0152] b) 扫描文档上的手写的涂划或线在文本文件中被转化成混乱信息。

[0153] c) 文档可能被 USPTO 以低质量扫描,这可能导致文档文字特征的失真。

[0154] d) 数学公式以及分子结构可能显示得不好。

[0155] e) 确保审查报告指的是原始提交的权利要求,而不是初始修改的权利要求等,在后者的情况下,概念的形成易于失真。

[0156] V. 词干化

[0157] 在我们的搜索算法中,我们关注技术短语具有两个主要的理由。首先,技术短语对专利文档是重要的,因此技术短语的正确识别是文档本身的正确分类的一部分。其次,与依赖文档之间的关键词相似性的文字搜索的传统模型不同,本算法意在识别不同技术短语的相关度。

[0158] (i) 自动的词干:根据所接受的词干化规则(例如,丢弃停止词,去除 ing 等)来使词词干化。

[0159] 如果(a)第一个字符是数字,(b)词含有诸如“~”、“#”、“±”的字符,以及(c)词仅含数字,则丢弃词。

[0160] (ii) 技术短语的手动词干:专利文档通常包括技术短语的缩写。这些短语很可能描述专利的关键科学概念。技术短语可以例如通过使用大写字母来区分。因此,如果词的第一个字母属于上面的情况,则特征被标记为可能的技术短语。此外,所有的词被手动地扫描,以识别其他的技术短语(即,“JavaBean”)。属于上述情况的词不被词干化,除非它们是复数形式。例如,词“ETCs”被词干化为“ETC”,而词“OLED”则不被词干化为“OL”。化学成分(即, Li 或 CO₂)也通过其第一个大写字母来识别;它们被标记但不被词干化。

[0161] (iii) 丢弃不能找到其科学含义的缩写(在大多数情况下,这些缩写表示将文档从图像格式向文本格式转化时的错误)。

[0162] VI. 词索引和频率

[0163] 对每个词干化的词进行索引,并计算其频率。在两个层次计算词的频率:

[0164] (i) 专利层次:在专利层次的词的频率是词干化的词在专利权利要求中出现的次数。

[0165] (ii) 概念层次:在概念层次的词的频率是词干化的词在概念中出现的次数,其中概念包括所有出现在被驳回的申请的原始权利要求中和现有技术的权利要求中的词。对于每个词干化的词出现在其中的概念,计算该频率。

[0166] 阶段 2:计算概念和语义关系

[0167] I. 概念关系

[0168] 基于在阶段 1 构建的训练集,在词和技术短语之间形成概念联系。这些联系形成告知我们哪些词被用来描述相似的发明的“词典”。专利对之间的概念关系在两个步骤被计

算。在第一个步骤中,对每个专利形成概率性的概念向量。该向量中的每个概念被分配零和表示概念与专利相关的概率 (CSV) 的值之间的值。在第二步骤中,计算专利对的概念相关度作为其概率概念向量重合的程度。该过程意味着不是基于语义相似性,而是基于每个专利含有的期望科学构想的相似性来确定相关度。

[0169] 采用概率性的 Bayesian 模型作为学习算法。定义 $P(c_i | \vec{d}_j) = \frac{P(c_i) \cdot P(\vec{d}_j | c_i)}{P(\vec{d}_j)}$ 作为

用于每个专利 (训练集外) 和概念 CSV 函数,即 $P(c_i | \vec{d}_j)$ 是在词袋 $\vec{d}_j$ 的条件上观察概念 c_i 的概率,其中 $\vec{d}_j = (\omega_{1j}, \dots, \omega_{\tau j})$, $\omega \in (0, 1)$ 和 τ 是文档集合中的特征总数。执行朴素 Bayesian 分类器方案并假定词在文档中的出现是彼此独立的:

$$[0170] \quad P(\vec{d}_j | c_i) = \prod_{k=1}^{\tau} P(\omega_{kj} | c_i)$$

[0171] 由于二元假设, $P(\omega_{kj} | c_i)$ 能够写成 $p_{ki}^{\omega_{kj}} \cdot (1 - p_{ki})^{1 - \omega_{kj}}$, 其中 p_{ki} 是在观察概念 i 的条件下词 k 出现在文档 j 中的概率。简单的数学运算提供了用于 CSV 的估算等式:

$$[0172] \quad CSV_{ij} = \sum_{k=1}^{\tau} \omega_{kj} \log \frac{p_{ki} \cdot (1 - p_{k\bar{i}})}{p_{k\bar{i}} \cdot (1 - p_{ki})}$$

[0173] 其中, p_{ki} 是词 k 出现在概念 i 中的概率, $p_{k\bar{i}}$ 是词 k 出现在除 i 外的概念中的概率。这些概率从训练集计算。一共对每个概念计算 2τ 个参数 $\{p_{1i}, \dots, p_{\tau i}\}$ 和全部 $2\tau \times C$ 个参数, 其中 C 是训练集中的概念的总数。利用来自训练集的相对频率来近似这些参数。

[0174] 最后,在已经对每个专利计算了概率性概念向量后,计算专利对的概念相似性作为概念向量的重合度。为了减小概念向量的维数,仅包括由其 CSV 排列的最高的 100 个概念。与公共的较大数量的概念相关的专利对被分类为在科学上更为相关。其他的度量也可以使用。为了细化该测量,还根据概念的出现向量中的排列对概念的出现进行加权。因此,如果概念向量包括相同概念中的更多一些并且具有较大的概念排名相似性,则认为概念向量更为相似。

[0175] 现在参照图 11,其示出了搜索过程。每个技术短语属于包括其他技术短语的概念,因此技术短语此时能够在查询文档中被识别。相同的诸如词干化、词典等的前面使用的技术可以被用来以更统一的方式获得短语,使得可以对查询 (公开) 计算 CSV。然后可以对目标文档计算 CSV,并且然后可以根据目标文档与查询文档的关系对目标文档进行排列,最接近的文档被输出给用户。

[0176] II. 语义关系

[0177] 作为概念关系算法的基准,还对数据集中的专利对计算语义关系。根据 ltc (对出现计数、倒 (inverse) 文档频率和余弦标准化 (cosine normalization) 进行对数加权) 方案对文字特征进行加权:

$$[0178] \quad ltc(t_k, d_j) = (1 + \log count(t_k, d_j)) \cdot \log \frac{N}{df_k}, \text{ 对于 } count(t_k, d_j) > 0$$

[0179] 以及,

$$[0180] \quad ltc(t_k, d_j) = 0, \text{ 对于 } count(t_k, d_j) = 0$$

[0181] 其中, N 是专利文档的总数, df_k 是含有词 k 的专利文档的数量。将权重项正态化,

以控制文档的长度,如下:

$$[0182] \quad \omega_{kj} = \frac{ltc(t_k, d_j)}{\sqrt{\sum_{s=1}^r ltc(t_s, d_j)^2}}$$

[0183] 其中, r 是文字特征的总数。最后,在给定词的加权向量的情况下,将专利对之间的语义相似性定义为:

$$[0184] \quad Text_{ij} = \frac{1}{K} \sum_{\omega} (\omega_{kj} - \omega_{ki})^2$$

[0185] 其中, K 是出现在文档 i 和 j 中的不同的词的数量。应当注意,技术短语被集成成组,并且在概念算法或语义算法中都不是作为单独的特征来计数的。

[0186] 搜索结果界面

[0187] 对于每个发明公开,搜索算法的输出是所暗示的专利现有技术的列表。图 12 示出了公开和搜索结果的主界面布局。用户将公开文本形式的查询插入在上部的框中,在下部的框中提供有相关的现有技术。

[0188] 公开

[0189] 用于搜索查询的公开能够以两种形式来描述:自由文本和关键词。自由文本部分允许用户以自然语言描述她的发明,没有特定的结构或长度限制。关键词部分旨在限定发明的技术领域,并细化搜索结果。搜索可以基于自由文本、关键词或该两者进行。基于自由文本的搜索得益于较丰富的公开,其可能导致对更多精确搜索结果的搜索。然而,自由文本搜索的缺点是存在缺乏技术要点的风险,这可能同样导致干扰性更强的搜索结果。

[0190] 搜索结果

[0191] I. 现有技术选择

[0192] 在进行搜索之前,用户选择所检索的现有技术的特征。当点击“搜索现有技术”按钮时,选择标准示出在单独的“弹出”屏幕中。选择标准包括:出现在搜索结果中的授权年份(范围或具体的年份)、申请年份(范围或具体的年份)、技术领域(多达六位的 IPC)、受让人姓名和现有技术的数量。默认的设置是不选择现有技术特征并且执行组合的自由文本和关键词搜索(如果由用户实现可用的话)。

[0193] II. 搜索结果布局

[0194] 搜索结果中的每个现有技术都可包括某些屏显信息:

[0195] (i) 相关度量度,此处以现有技术文档标题上方的横条上的星形图标表示。横条的右端被着以绿色,而横条的左端被着以红色。相关度随着从色谱的绿色调向红色调移动而升高。替代性的方案包括将相关度显示为数字或比值。

[0196] (ii) 标题—现有技术的完整标题可以列在相关度横条的下方。该标题被联系到含有详细专利信息(如全文、技术领域、引述等)的单个页面。

[0197] (iii) 主要专利特征:专利号、提交日和受让人姓名出现在标题线下方。

[0198] 试验模型允许用户将公开和相应的搜索结果导入到办公室应用,如 Word 或 Excel,以便进一步分析。

[0199] 等级评价

[0200] 对搜索引擎的试验模型的测试使用具有大约 8,000 个申请和 16,500 个申请—现

有技术联系的样本,以检查根据本实施例的算法对由专利审查员对于 500 个专利申请所引用的现有技术正确地进行预测的能力。

[0201] 对于每个申请-现有技术对(位于训练集外),选择 10 个控制对。所述控制对包括与实际对相同的申请,其中将该申请与对照的(placebo)现有技术相联系。对照现有技术与实际现有技术来自同一技术领域(4 位 IPC)。基于本发明的算法,针对每个申请,对现有技术进行排列,并检查实际现有技术收到高排名的程度。下面论述更多的试验细节。

[0202] 已经发现,在多于 80% 的申请中,实际现有技术(即专利审查员引用的现有技术)处于最高的两个选择,其中在多于 66% 的申请中,实际现有技术出现在第一选择中。将该性能与语义相似性算法作比较。在仅仅 41% 的申请中,语义算法将实际现有技术排在最高的两个选择中,并且在仅仅大约 23% 的申请中,排名为第一选择。

[0203] 例如,发现发明名称为“多米诺多路连接的方法、装置和系统(Method, apparatus and system of domino multiplexing)”(IPCH03K/脉冲技术)的专利申请 No. 7183807 与发明名称为“具有快速预充电的或型 CMOS 逻辑电路(OR-type CMOS logic circuit with fast precharging)”(IPC G11C/静态存储器)的现有技术 No. 4899066 相关。申请的权利要求中使用最频繁的词为:[设置、块、连接、多米诺、输出、P-沟道、预处理、N-沟道、状态]。现有技术中使用最频繁的词为:[CMOS、连接、数据、线路、MOS、预充电、单个、部分、晶体管]。这两个专利共有的词非常少;因此,它们的语义相似性低。此外,搜索引擎还从训练集发现技术短语 CMOS(互补金属氧化物半导体,晶体管类型)与 P-沟道(另一种晶体管类型)之间的关系,并且识别出强烈的概念相似性。CMOS 和 P-沟道的相关度的了解是通过专利号 7161387 的专利申请和现有技术 6774696(审查员所引用)之间的联系进行的。该申请包含短语 P-沟道,其中现有技术含有短语 CMOS。

[0204] 专利信息

[0205] 试验中使用的数据来自美国专利商标局(USPTO)。对从 1975 至 2007 年在 USPTO 授权的所有专利提取所有特征和文字信息。专利数据库包括两个部分:(1)用于每个专利权利要求的词,和(2)专利特征,如:授权日和申请日、发明人、受让人、地址和国籍、对其他专利的引用、对非专利文献的引用、技术领域类别(国际专利分类和美国分类)。专利数据库一共包括横跨主要技术领域分布的 4,069,412 个专利,如下:19%为化学;14%为计算机和通信,8%为药物和医疗,18%为电气和电子,22%为机械,剩下的专利被划分到其他技术领域。

[0206] 审查报告和原始权利要求

[0207] 审查报告和申请原始权利要求来自 USPTO 的文件袋部分。该文件袋是追踪专利申请过程的文档的集合。使用两个文档:(1)非最终驳回,和(2)原始权利要求。非最终驳回是专利审查员对专利申请的最初应答。一般地,非最终驳回指的是被专利审查员驳回的一组权利要求和构成驳回基础的相关现有技术。申请文档包括非最终驳回文档所涉及的提交到 USPTO 的原始权利要求。排除化学和药物领域的专利申请,因为这些申请通常包括分子结构和其它非文字特征。

[0208] 如表 3 所示,依照上面描述的手动和自动过程,训练集数据库包括 7,992 个非最终驳回和相同数量的原始权利要求。从非最终驳回中,识别出原始申请的特征和现有技术的特征之间的 16,444 个概念关系。平均地,被驳回的申请被 2.06 个现有技术参照。在仅仅

大约 38.3% 的审查员驳回决定中,所参照的现有技术在与该申请相同的 4 位 IPC。对于 6 位 IPC,该百分比降到大约 17%。这意味着,由于现有技术和驳回来自不同的技术领域,因此它们很可能具有不同的文字表示。平均地,申请的提交年份与现有技术的授权年份之间的滞后为 4.82 年。

[0209] 表 1

[0210]

专利变量的描述性统计						
	专利总数	收到引用的 专利的百分比	引用总数	公司百分比	权利要求平均数	平均授权年份
申请						
所有技术领域	7,992	4.6%	705	40.7%	22.9	2006
计算机和通信	2,991	3.2%	131	47.3%	25.6	2006
电气和电子	2,219	7.8%	419	36.9%	20.3	2006
机械	1,214	4.8%	105	29.8%	19.7	2006
现有技术						
所有技术领域	1,6126	99.6%	274,171	37.1%	18.9	1998
计算机和通信	6,154	99.7%	133,528	46.4%	21.9	2000
电气和电子	4,674	99.7%	71,894	35.5%	17.2	1998
机械	2,680	99.3%	33,931	25.1%	16.0	1996

[0211] 表 3 专利变量的描述性统计

[0212] 表 2

[0213]

驳回的描述性统计							
	申请数量	现有技术数量	每个申请的现有技术	每个现有技术的申请	相同主要技术领域的百分比	相同 4 位 IPC 的百分比	相同 6 位 IPC 的百分比
所有驳回							
所有技术领域	7,992	16,126	2.06	1.02	76.5%	38.3%	17.0%
计算机和通信	2,991	6,154	2.14	1.03	82.4%	37.2%	14.2%
电气和电子	2,219	4,674	2.07	1.02	80.8%	40.2%	18.4%
机械	1,214	2,680	2.00	1.01	75.8%	38.7%	21.9%

[0214] 表 4 :驳回的描绘统计

[0215] 技术短语

[0216] 专利文档通常包括描述其科学内容的技术短语。本实施例的学习算法在识别技术短语之间的科学关系时特别重要,这些技术短语在语义上不同,但是可表示相关的科学概念。例如,技术短语 CDPD(其代表蜂窝数字分组数据)对专利 No. 5805997 中描述的构想是重要的。

[0217] 平均地,概念包括 116 个词干化的词,其中 68 个词出现在现有技术中,而 62 个词出现在申请的原始权利要求中。训练集中的词和技术短语的总数为 28,227。这些词出现了 10,085,912 次。

[0218] 表 6 总结了出现在训练集中的若干常见技术短语的例子。

[0219] 表 3

[0220]

文字特征的描述性统计						
	总的特征出现	特征总数	平均特征频率	总的短语出现	技术短语总数	技术短语平均数
申请						
所有技术领域	3,054,542	19,628	61.91	47,962	2,471	1.08
计算机和通信	1,316,250	10,974	61.95	23,277	1,457	1.21
电气和电子	742,746	8,637	58.50	14,155	884	1.27
机械	409,790	6,238	63.09	3,606	358	0.69
现有技术						
所有技术领域	7,031,370	18,394	67.83	102,339	2,670	0.95
计算机和通信	3,158,589	11,545	68.48	53,005	1,849	1.11
电气和电子	1,853,808	9,491	65.36	32,436	949	1.08
机械	1,034,988	7,349	69.80	7,919	412	0.58

[0221] 表5:文字特征的描绘统计

[0222] 表4

[0223]

常见技术短语的例子				
技术短语	缩写	现有技术频率	申请频率	含义
(1) 射频识别	RFID	577	1411	自动识别方法, 依赖利用称作 RFID 标签或收发机存储和远程搜索数据
(2) 金属氧化物半导体场效应晶体管	MOSFET	576	587	用来放大或转换电子信号的设备。是目前为止在数字和模拟电路中最常用的场效应晶体管
(3) 通用串行总线	USB	515	490	接口设备的串行总线标准
(4) 全球定位系统	GPS	524	1365	基于卫星的导航系统, 由美国国防部设置在轨道中的 24 颗卫星的网络构成
(5) 可扩展标记语言	XML	273	334	用于创建自定义标记语言的通用规格

[0224]

(6) 动态随机存取存储器	DRAM	391	304	一类随机存取存储器, 其将数据的每个位存储在集成电路内的单独电容中
(7) 微机电系统	MEMS	377	170	微观技术, 以纳米等级融合到纳米机电系统中, 以及纳米技术
(8) 印刷电路板	PCB	268	253	薄板, 芯片和其他电子器件放置在其上
(9) 电压控制振荡器	VCO	194	235	电子振荡器, 设计成由电压输入在振荡频率中进行控制
(10) 快速傅立叶变换	FFT	41	386	计算离散傅立叶变换及其逆的高效算法

[0225] 表 7- 常见技术短语的例子

[0226] 手动识别技术短语及其含义是可能的。首先, 识别以大写字母开头的词。其次, 对于每个词和缩写, 利用各种计算源 (如维基百科和 Abbreviations.com), 手动地检查其是否表示技术短语。该过程留给我们 4,699 个技术短语。表 7 总结了相关技术短语的一些例子。相关度基于不同技术短语在相同概念中共同出现的程度。

[0227] 例如, 短语 SCSI 在专利 7181553 的专利申请的原始权利要求中出现了 6 次。这些权利要求被相对专利号为 7080140 的专利而驳回, 在专利号 7080140 的专利中, 短语 SAN 出现了 3 次。通过将被驳回的申请和相关的现有技术联系在一起识别两个短语之间的科学关系。另一个例子是短语 DMD 和 CRT, 短语 DMD 在专利号 7180554 的申请中出现了 11 次, 而短语 CRT 在现有技术 6456432 的权利要求中出现了 8 次。

[0228] 评价

[0229] 试验设计

[0230] 使用在前一节中描述的数据来测试本发明的概念关系算法的性能, 并将其与当前市场方案中使用的语义相似性算法的性能作比较。检查概念算法和语义算法预测对于我们的样本中的申请, 被专利审查员引用的现有技术的能力。

[0231]

表 5

相关技术短语的例子				
第一短语			第二短语	
技术短语 1	缩写 1	含义	技术短语 2	缩写 2
(1) 存储区域网	SAN	设计成附接计算机存储设备的网络	小型计算机系统接口	SCSI
(2) 阴极射线管	CRT	含有电子枪和荧光屏的真空玻璃封套	数字微镜设备	DMD
(3) 节目时钟基准	PCR	用在数字视频压缩中指示时间标记信息包离开编码器的时刻的系统时钟值的时间戳	打包基本流	PES
(4) 时分多路复用	TDM	沿单个电信传输路径发送多个数字信号的方法	局域网	LAN
(5) 协议数据单元	PDU	在网络的个体之间作为单元输送的信息和可含有控制信息、地址信息或数据的信息	混合自动重复查询	H-ARQ
(6) 随机存取存储器	RAM	一类计算机数据存储设备	操作系统	OS
(7) 有机发光设备	OLED	电致发光层由有机化合物薄膜构成的任何发光二极管	有机金属半导体场效应晶体管	OMESFET
(8) 波分多路复用	WDM	一种技术, 通过使用不同波长的激光携带不同的信号在单个光纤上多路传输多个光学载体信号	梯度指数	GRIN
(9) 场效应晶体管	FET	一种晶体管, 其依赖电场控制半导体材料中的“沟道”的形状和因此的导电性	直接空间—时间脉冲发生器	DSTPG
(10) 通用异步接收器/发射器	UART	一种计算机硬件, 在并行和串行形式之间转换数据	数据载体检测	DCD

[0232] 表 8 相关技术短语的例子

[0233] 对于每个申请 - 现有技术队 (位于训练集外), 选择 10 个控制对。控制对包括与

实际对中相同的申请,其中将该申请与对照的现有技术相联系。对照现有技术与实际现有技术来自同一技术领域(4位IPC)。基于本发明的算法,针对每个申请,对现有技术进行排列,并检查实际现有技术收到高排名的程度。

[0234]

专利申请号	排名	现有技术	概念量度	现有技术	语义距离
7143262	1	5390315	0.213	5113041	0.0096
	2	5111384	0.039	5111384	0.0125
	3	5111424	0.025	4945505	0.0132
	4	4959782	0.020	4959782	0.0142
	5	5111421	0.017	5390315	0.0144
	6	5111431	0.013	5113042	0.0145
	7	4945505	0.012	5113363	0.0151
	8	5113363	0.007	5111431	0.0152
	9	5113361	0.005	5111424	0.0153
	10	5113041	0.003	5113361	0.0170
	11	5113042	0.002	5111421	0.0170

[0235] 表9:对专利号7143262的美国专利和相关现有技术5390315进行的排名。其他用于该申请的现有技术是控制。

[0236] 结果

[0237] 下面的表10报告了对实际驳回及其控制之间的概念和语义量度之差的平均比较测试。对于概念量度(面板A),实际驳回对的平均概念量度是0.38,其中对控制对的平均概念量度是0.12。这意味着,平均地,实际驳回的概念量度比控制的概念量度大两倍。对于语义量度(面板B),实际驳回的平均语义距离是0.015,而对于控制是0.018。尽管均值方面的差值在1%的水平是显著的(因此,实际驳回比控制具有明显更高的相似性),但是该差值在数量上并不大,因为与概念量度的204%相比,其仅占18%的平均控制语义量度。这种形式的结果在主要技术领域是稳健的。

[0238] 表6

[0239]

面板 A: 平均值比较: 概念量度						
	实际驳回			控制		
申请	对的数量	平均值	标准偏差 (Std.dev)	对的数量	平均值	标准偏差
所有技术领域	500	0.38	0.344	5000	0.12	0.189
平均值的差别		0.254***				
t-统计		26.05				
计算机和通信	169	0.32	0.322	1690	0.09	0.172
平均值的差别		0.229***				
t-统计		14.859				
电气和电子	191	0.47	0.401	1,910	0.15	0.226
平均值的差别		0.315***				
t-统计		16.852				
机械	74	0.35	0.270	740	0.12	0.156
平均值的差别		0.227***				
t-统计		10.997				
B 区: 平均值比较: 语义距离						
	实际驳回			控制		
申请	对的数量	平均值	标准偏差	对的数量	平均值	标准偏差
所有技术领域	500	0.015	0.005	5000	0.018	0.005
平均值的差别		-0.002***				
t-统计		-8.907				

[0240]

计算机和通信	169	0.015	0.005	1690	0.017	0.005
平均值的差别		-0.002***				
t-统计		-4.563				
电气和电子	191	0.016	0.005	1910	0.018	0.005
平均值的差别		-0.002***				
t-统计		-5.721				
机械	74	0.015	0.005	740	0.017	0.017
平均值的差别		-0.002***				
t-统计		-3.472				

[0241] ***表示 1% 的显著水平

[0242] 表 10

[0243] 利用已接受的来自文本分类文献的标准来限定本发明的概念搜索引擎的性能。对于每个专利对,使用本发明的算法来确定现有技术是否被专利审查员引用,作为驳回申请的理由,或者现有技术是否为控制。对于每个对,已经知道其是否表示驳回关系或者其是否为控制。然后,检查本发明的算法能够正确地从所述控制识别实际驳回的程度。用 a 来表示本发明的算法正确地识别驳回的次数,而用 b 作为本发明的算法没有识别正确的驳回的次数。表 11 的矩阵总结了所有可能的结果:

[0244] 实际算法 / 驳回 非 - 驳回

[0245] 现有技术 A B

[0246] 控制 C D

[0247] 表 11 :所有可能的结果的矩阵

[0248] 在信息搜索研究中常用的量度包括“撤回”和“错误”。“撤回”定义为 $a/(a+b)$, 并且“错误”定义为 $(b+c)/(a+b+c+d)$ 。“撤回”测量所述算法能够正确地识别驳回的程度,其中错误测量不正确的分类。

[0249] 在论述概念和语义量度的性能之前,先论述分类规则。当前的试验试图将专利申请分类到 11 个类别之一中,其中每个类别都是现有技术文档。本发明的分类规则如下。对于每个申请,根据与申请的(语义的和概念的)距离来对现有技术文档进行排列。然后,仅将第一 γ 个专利划分为相关的现有技术。算法的性能在多数情况下由正确的驳回分类来确定,换言之,由实际驳回在最上面的 γ 个所选现有技术中出现的次数来确定。

[0250] 概念和语义算法的“撤回”和“错误

[0251] 表 12 以及图 13 和 14 报告了对于 $\gamma = 1$, 概念和语义算法的性能,换言之,当仅将具有最高排名的现有技术归为驳回并且将所有其他现有技术归为非驳回时实现的性能。一共对 500 个申请进行分类。概念算法正确地分类 331 个现有技术,即对于 331 个申请,实际驳回收到最高的概念排名。相反,语义算法仅正确地分类 115 个现有技术,即对于仅 115 个申请,实际现有技术收到最高的语义排名。这些图意味着,概念算法的撤回量度是 0.66 (331/500),而语义算法的撤回量度仅是 0.23 (115/500)。相似地,概念算法的错误量度

是 $0.061(192+192)/5,500$ ，而语义算法的错误量度是 $0.14(388+388)/5,500$ 。

[0252] 表 7

[0253]

概念和文字分类							
所有技术领域							
概念				文字			
	驳回	非驳回	总数		驳回	非驳回	总数
驳回	331	169	500	驳回	115	385	500
控制	169	4831	5000	控制	385	4615	5000
总数	500	5000	5500	总数	500	5000	5500
计算机和通信							
概念				文字			
	驳回	非驳回	总数		驳回	非驳回	总数
驳回	121	48	169	驳回	38	131	169
控制	48	1642	1690	控制	131	1559	1690
总数	169	1690	1859	总数	169	1690	1859
电气和电子							
概念				文字			
	驳回	非驳回	总数		驳回	非驳回	总数
驳回	131	60	191	驳回	46	145	191
控制	60	1850	1910	控制	145	1765	1910
总数	191	1910	2101	总数	191	1910	2101
机械							
概念				文字			
	驳回	非驳回	总数		驳回	非驳回	总数
驳回	46	28	74	驳回	18	56	74
控制	28	712	740	控制	56	684	740
总数	74	740	824	总数	74	740	814

[0254]

[0255] 表 12 概念和语义分类

[0256] 现在参照图 15, 其示出了对于不同的 γ 个削减 (cut-off) 的现有技术, 概念和语义算法的正确分类的累积分布。与语义算法的仅仅大约 41% 相比, 多于 80% 的实际驳回被

概念算法排列为最高的两个选择。

[0257] 算法性能和训练集大小

[0258] 上面描述的性能结果基于数据集中的所有技术领域。期望概念算法的性能随着训练集的大小增加而提高。大的训练集使得概念算法能够了解词和短语之间更多的科学关系,并改善相关现有技术的识别。

[0259] 现在参照图 16 和 17,其表示每个技术领域的概念算法的性能。对于所有技术中的每一个,电气和电子、计算机和通信、以及机械,图 16 描绘了概念和语义方法之间的正确分类,而图 17 示出了两个方法的相应错误率。因此,考虑具有最大训练集的技术领域(“电气数字数据处理”)。该技术领域包括 1,916 个文档(1,209 个申请和 1,501 个现有技术)。概念和语义算法的撤回率分别为 78 和 18.2——图 16,其中错误率分别为 4 和 14.9——图 17。

[0260] 搜索模型的进一步细节

[0261] 在下文中,参照所附的图 18A、18B 和 19 给出了可能的搜索模型的进一步的细节。

[0262] 如所论述的,执行与专利审查过程相关的完整的文档集的提取。这些文档被用来将术语组合成具有相似的技术含义的类别。基于这些组合和打分算法,对在用户搜索查询中使用的每个词创建同义词列表,全部如图 18A 所示。图 18A 是示例性流程图,示出了根据本发明的优选实施例,提供用于要被搜索的专利文档的概念空间并将包括概念词的同义词的词映射到该空间的过程。现有技术引用部分、原始申请以及审查文献全部被用来限定该概念空间。然后可基于如上所述的词的频率和位置因素将词映射到该概念空间,以提供加权。然后可以计算词之间的评分,并可使用同义词数据库来添加词。

[0263] 利用若干模型中的一个或多个,根据这些文档所含有的关于专利申请和专利审查员引用的现有技术之间的技术相似性的信息,对审查文档进行分类。图 18B 示出了 5 个可能的模型,用于获得建立概念空间的初始阶段的搜索术语。下面的模型通常单独地应用于 USPTO 和 EPO 审查记录。

[0264] 1. 使用用于仅根据初始审查报告进行组合的术语的信息。第一个非最终驳回通常被认为在相关现有技术方面是最一般的。

[0265] 2. 使用用于根据最后审查报告进行组合的术语的信息。最后的非最终驳回可能比初始的非最终驳回包括更相关的现有技术。这之后可以是授权通知。

[0266] 3. 进一步的搜索模型包括放弃之前的非最终驳回。导致放弃专利申请的现有技术可能比导致授权的现有技术与专利申请更为相关。

[0267] 4. 还从文档“审查员的搜索策略和结果”来识别审查员在审查过程期间所使用的实际搜索词。

[0268] 5. 最后的模型从 1-4 中的两个或更多个源集合组合信息。

[0269] 现在参照图 19,其是简化图,示出了与本发明的实施例使用的同义词选择的过程。用户输入搜索查询并被提供来自同义词数据库的同义词以供选择。系统查询能够从所输入的词拆开的词和所输入的词的词干,然后基于结果中的查询短语的权重执行搜索。更相似的同义词能够在结果中被更高地加权,或者完全匹配能够比同义词匹配更高地加权。无论是哪种方式,都基于找到搜索查询或由同义词修改的搜索查询来向每个文档提供总的分数。

[0270] 用来对搜索结果和最终结果进行排列的术语然后被作为搜索结果提供给用户,排名较高并且因此被推断为更接近的相关现有技术被列在前面。

[0271] 总结

[0272] 现有技术搜索对创新过程和确定专利范围是重要的。自动现有技术搜索的主要挑战是如何基于用于专利文档的大规模数据集的文字特征来识别科学关系。

[0273] 已有的现有技术搜索引擎在来确定科学相关度时依赖语义相似性。在实际中,这种引擎的表现差,因为概念相关度通常不与语义相似性相联系。本发明的实施例公开了一种搜索引擎,其基于相关技术的专业评价(来自专利申请过程的专利审查报告)的全面的训练集来度量专利文档的科学相关度。这些报告提供了关于专利审查员对被驳回的专利申请引用的现有技术的信息。利用机器学习技术,基于每个专利含有的期望科学构想来度量专利文档的相关度。

[0274] 为了测试本算法的性能,检查该算法能够正确地将被驳回的申请分类到其实际现有技术的程度。本发明的算法产生了对语义相似性算法的充分提高。发现对于多于 80% 的申请,本发明的算法将实际现有技术排在最高的两个选择,其中对于多于 66% 的申请,实际现有技术被排为第一选择。相反,语义相似性算法仅对 42% 的申请将实际现有技术排在最高的两个选择,并且仅对大约 23% 的申请将实际现有技术排为第一选择。

[0275] 应用

[0276] 基本的应用是对专利文献进行搜索以发现对查询文档的匹配。查询文档可以是对还没有提交的新的专利申请的建议,或者是对一般的新技术投资的建议,用于确定提交专利申请的可行性或确定投资新技术的明智性。可替代地,查询文档可以是实际专利申请,对于该实际专利申请,用户希望预测专利局由于任何理由的结果。作为进一步的替代,用户实际上可以是希望找到好的相关现有技术的专利局审查员。

[0277] 在进一步的申请中,一旦能够得到每个查询专利对的记分,则记分可以用来计算专利申请的可行性的风险量度,例如驳回和诉讼的风险。更具体地,商业应用中的一种是帮助企业知识产权专家管理他们的专利布局,并帮助他们通过识别市场上可能侵权他们的专利的技术和产品而带来许可收入。在这种情况下,尽管专利文献被用来发现技术术语的组,但是最终的搜索可以在一般的技术文献(包括公司网址和商业报告)上进行。

[0278] 术语“包括”、“包含”、“具有”及其同根的变形形式意指“包括但不限于”。该术语包括术语“由……构成”和“基本上由……构成”。

[0279] 如此处所使用的,单数形式“一个”和“所述”、“该”包括复数形式,除非上下文中明确地另有所指。

[0280] 应当注意到,本发明的某些为了清楚起见而在分开的实施例的背景下描述的特征也可以在单个实施例中以组合的形式被提供。相反,本发明的各种为了简便起见而在单个实施例的背景下描述的特征也可以分开地提供,或者以任何适当的子组合提供,或者在本发明的任何其他描述的实施例中适当地提供。各种实施例的背景下描述的某些特征不应当被认为是这些实施例的必要特征,除非该实施例没有这些要素则无法实施。

[0281] 尽管已经结合具体的实施例描述了本发明,但是显然许多替代设置、修改和变型对本领域普通技术人员将是显而易见的。因此,意在包括所有落在所附的权利要求的精神和宽泛范围内的这些替代设置、修改和变型。

[0282] 本说明书中提及的所有公开、专利和专利申请在本文都通过引用的方式全部并入本说明书中,如同每个单独的公开、专利或专利申请都具体并单独地指定为通过引用的方式并入此处。此外,本申请中对任何参考文献的引用或识别都不应当理解为承认该参考文献可用作对本发明的现有技术。就使用章节标题而言,其不应当理解为一定是限制性的。

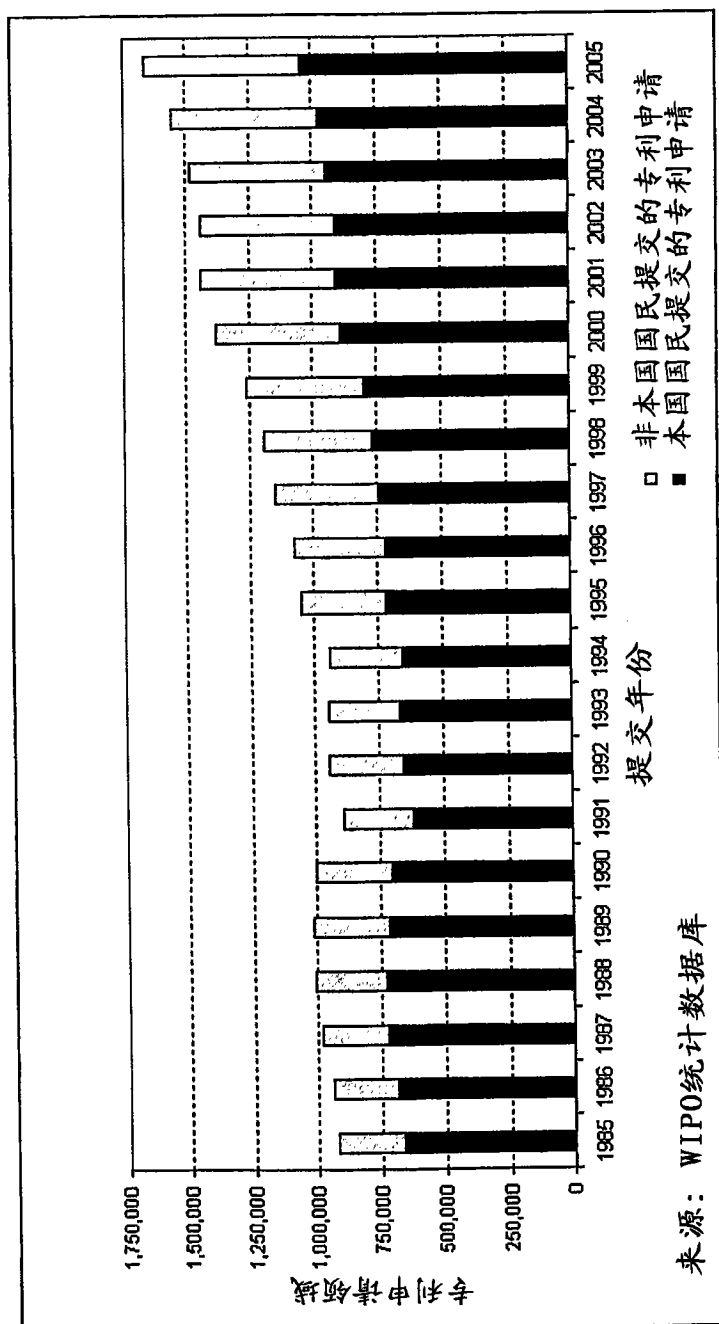


图 1

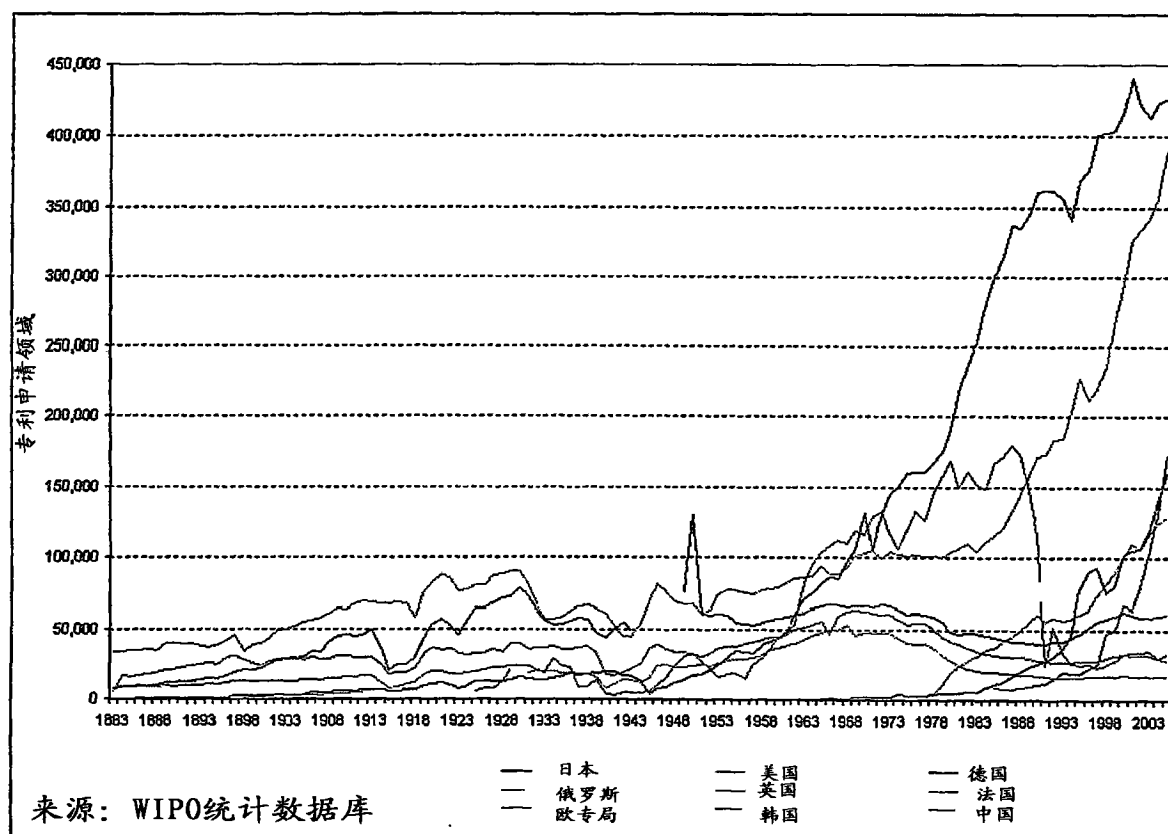


图 2

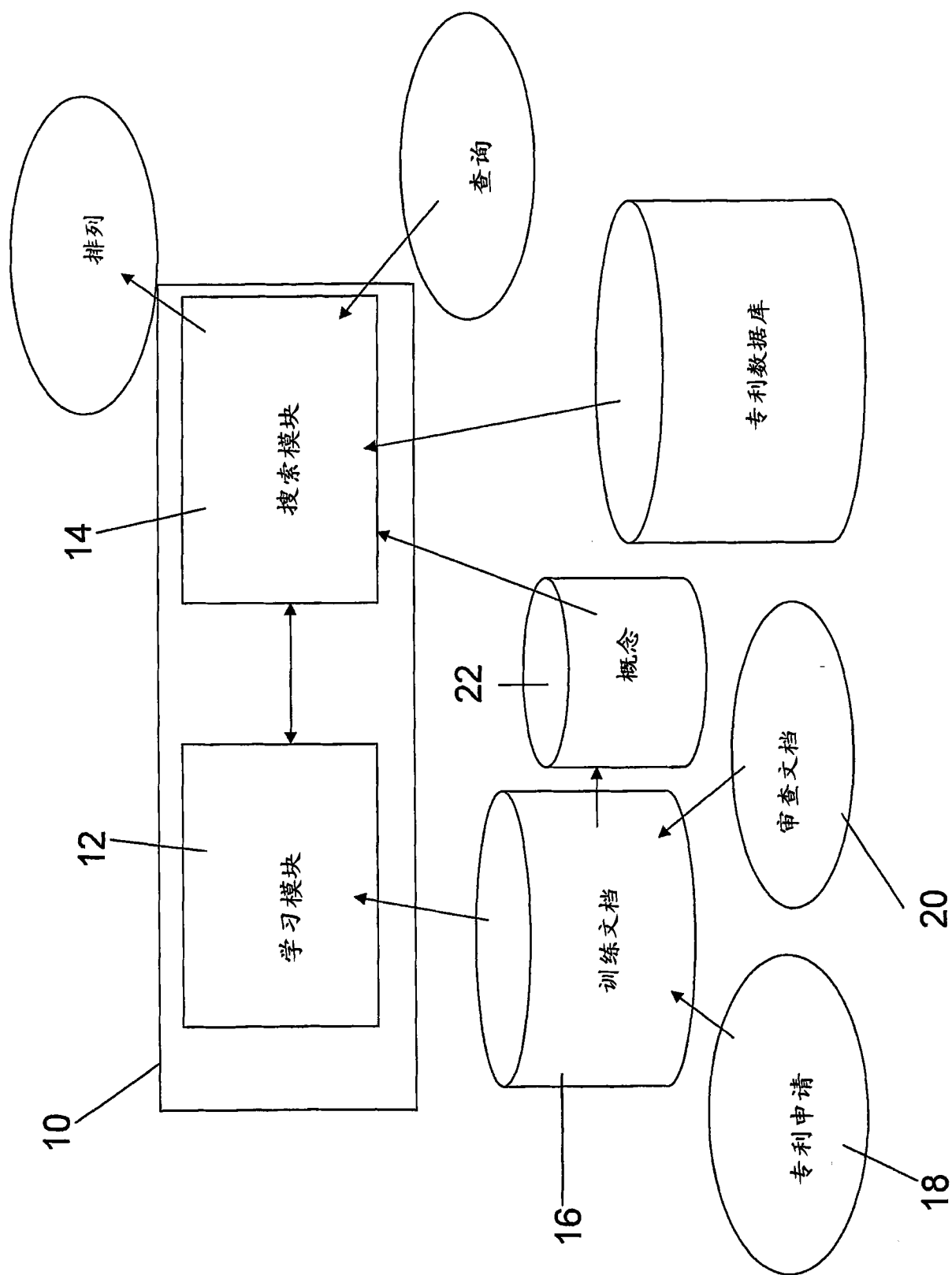


图 3

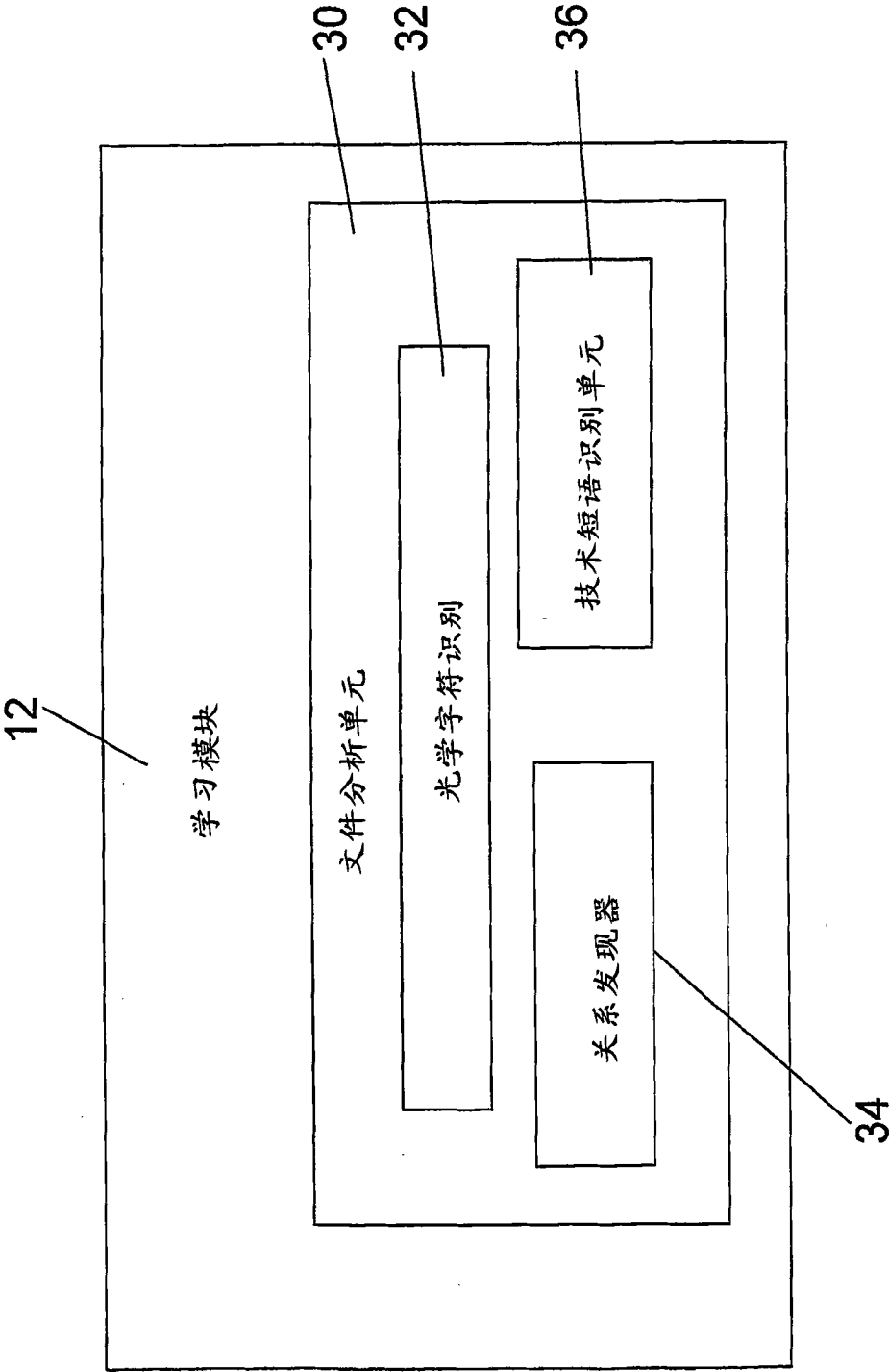


图 4

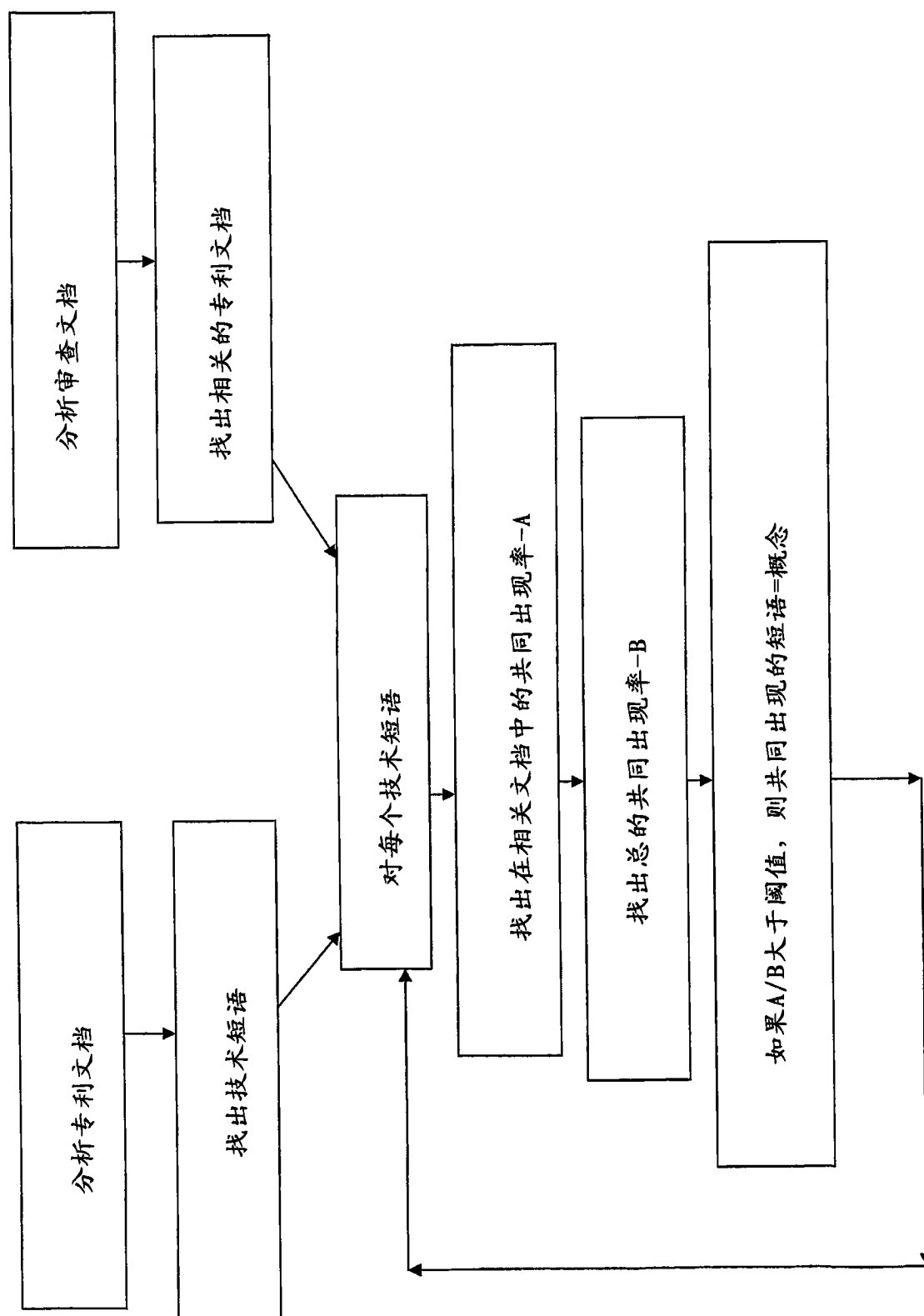


图 5

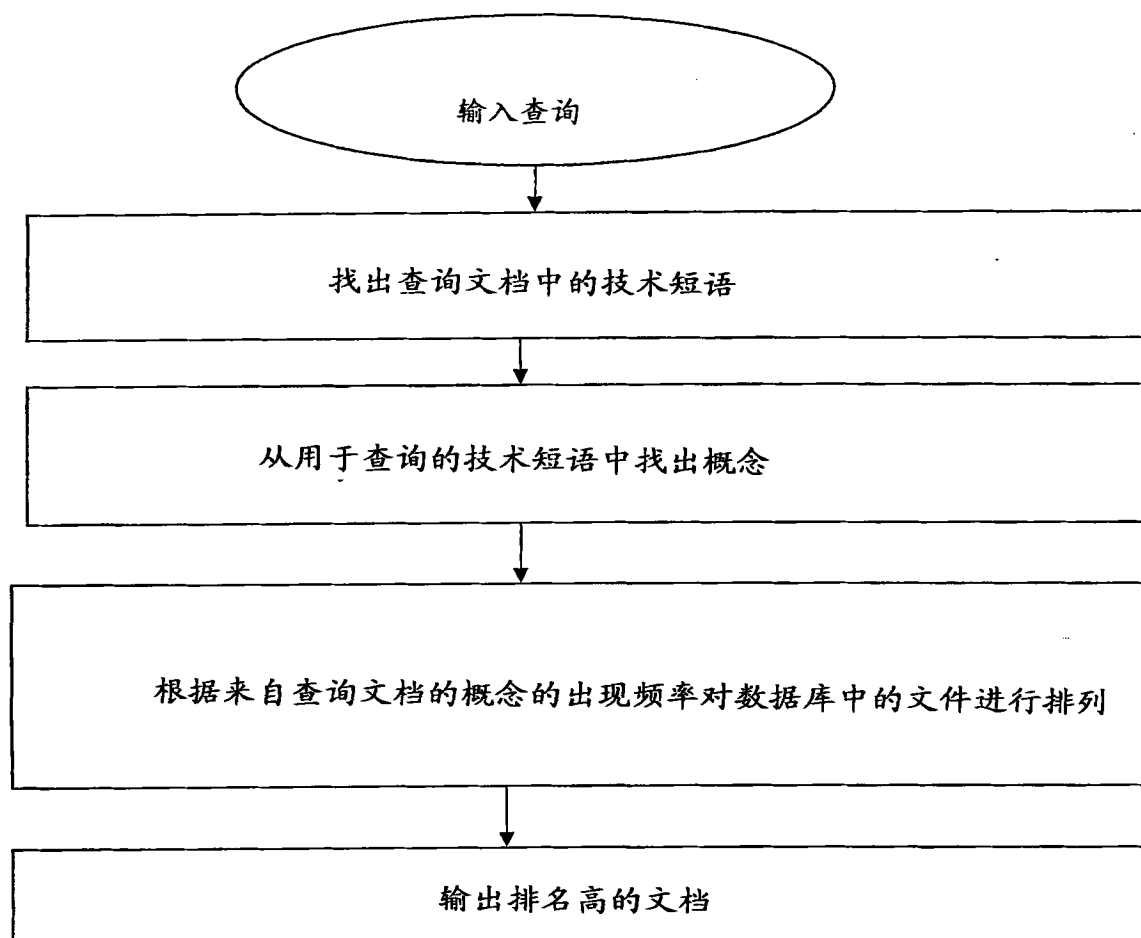


图 6

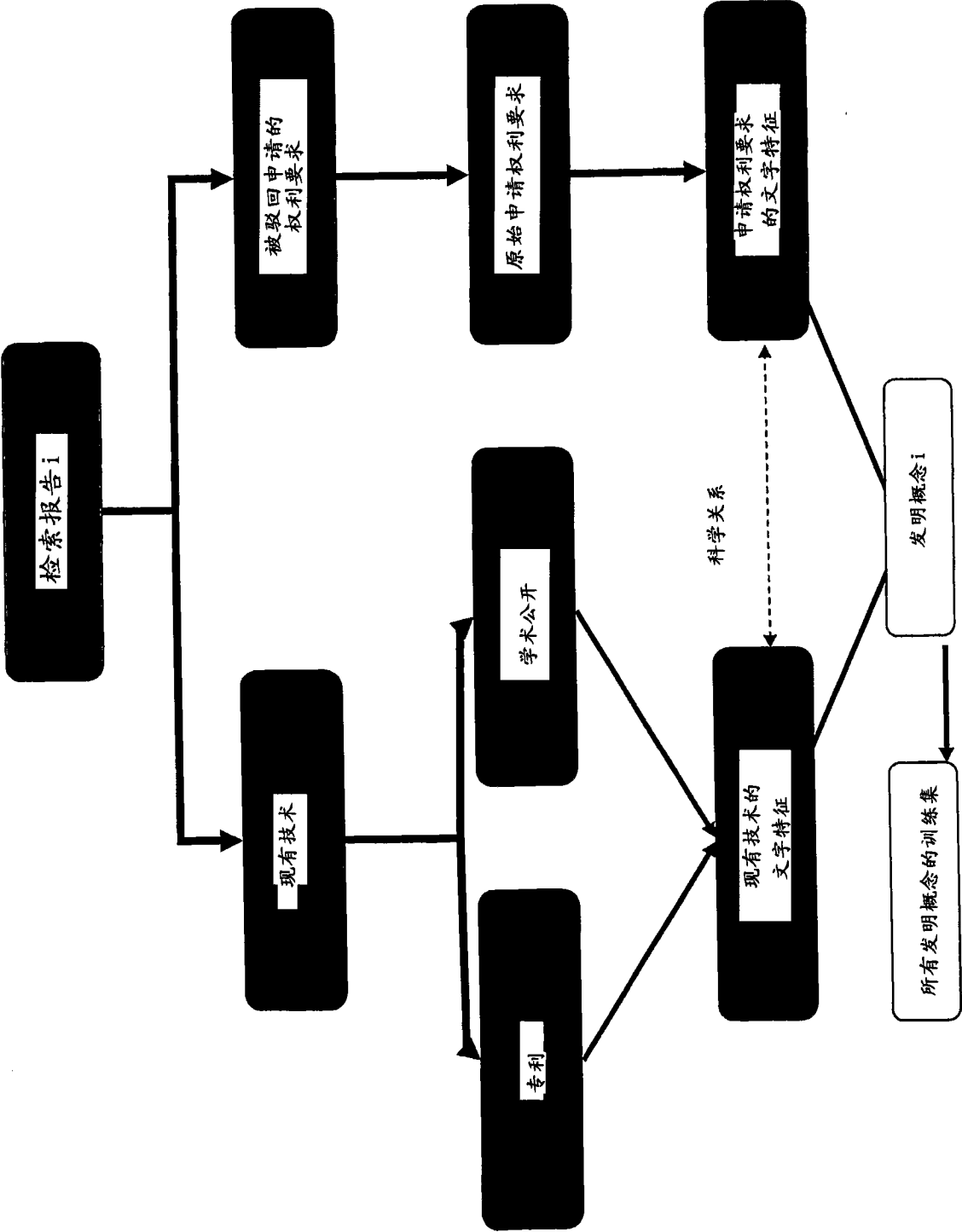


图 7

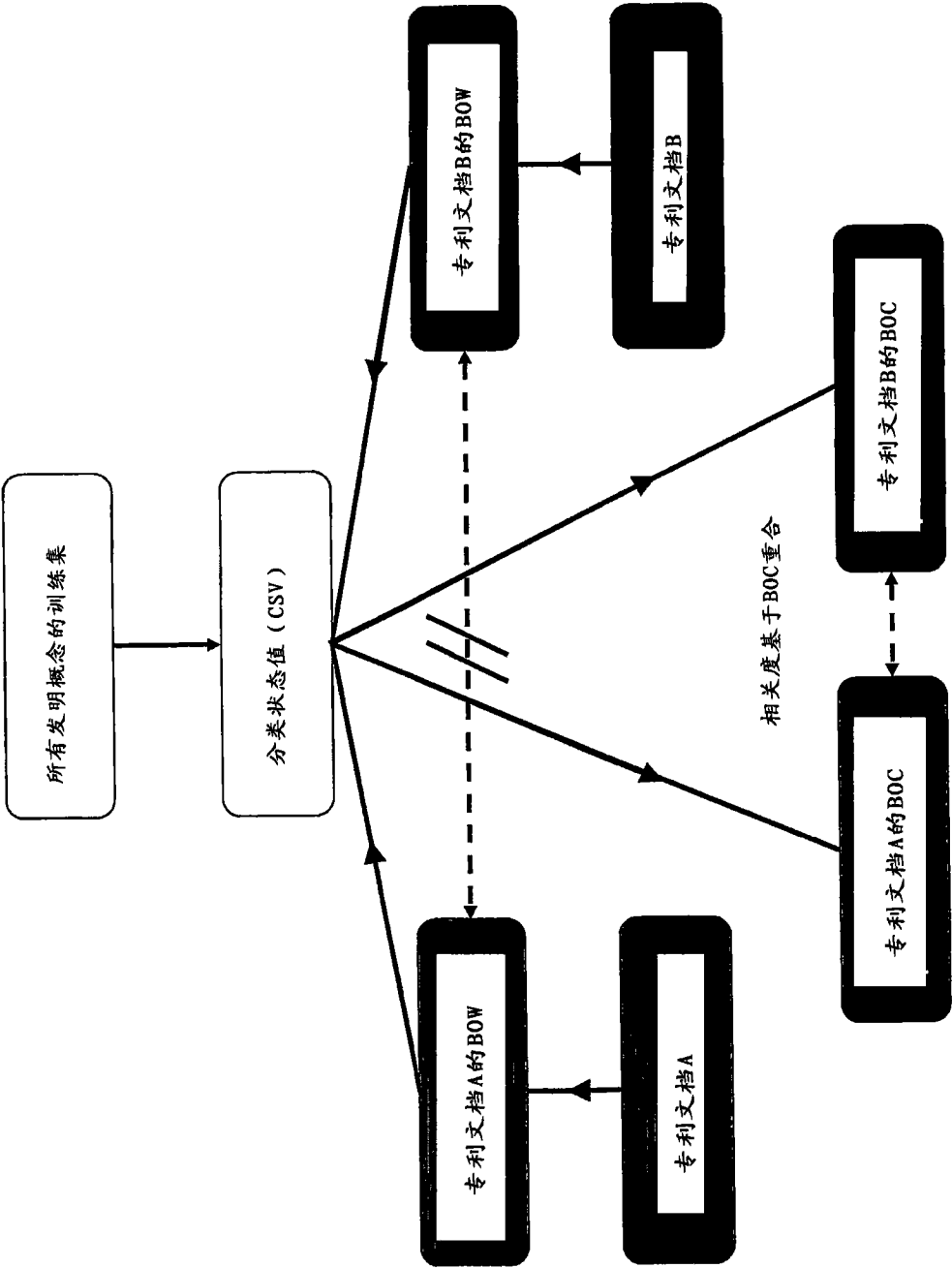


图 8

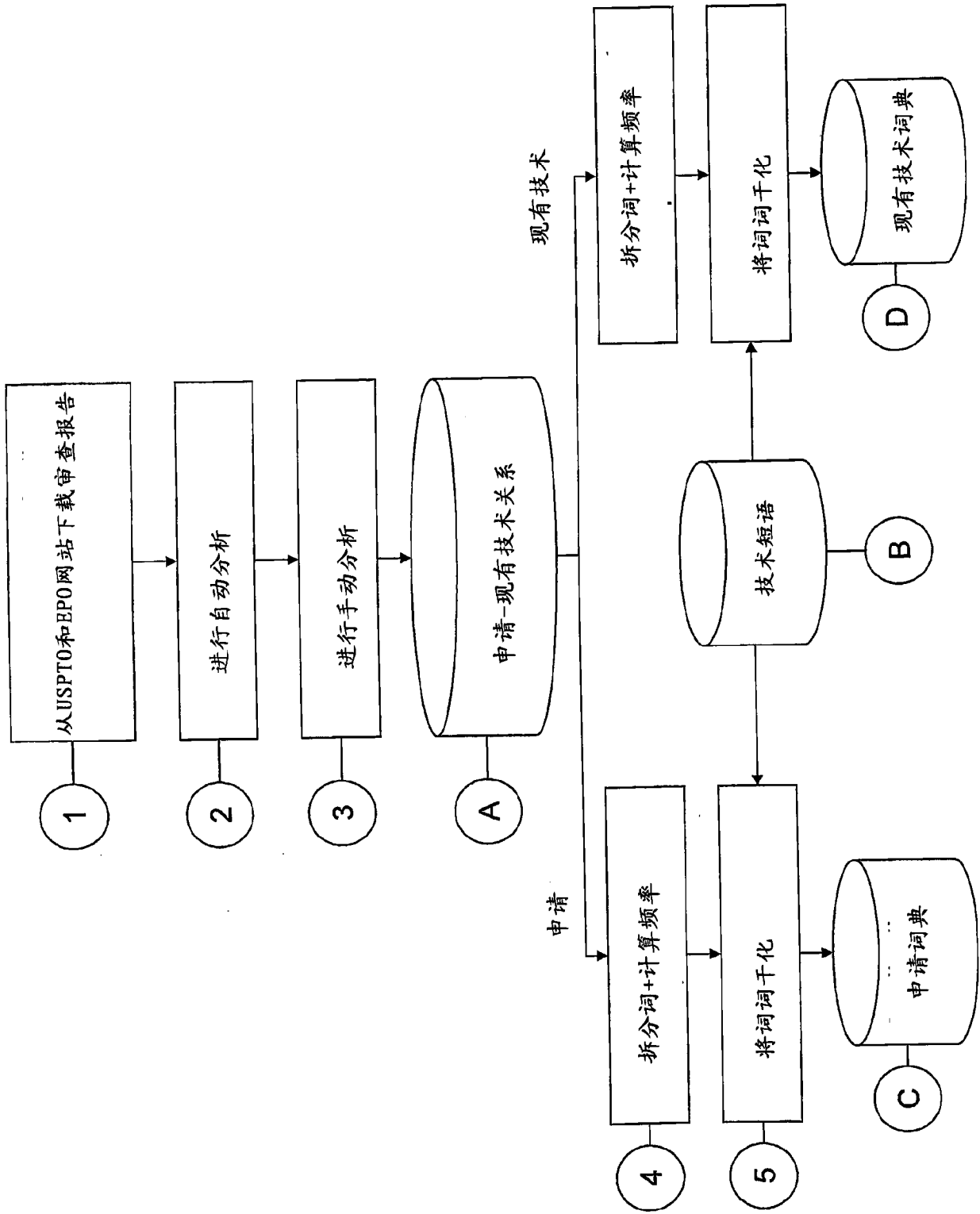


图 9

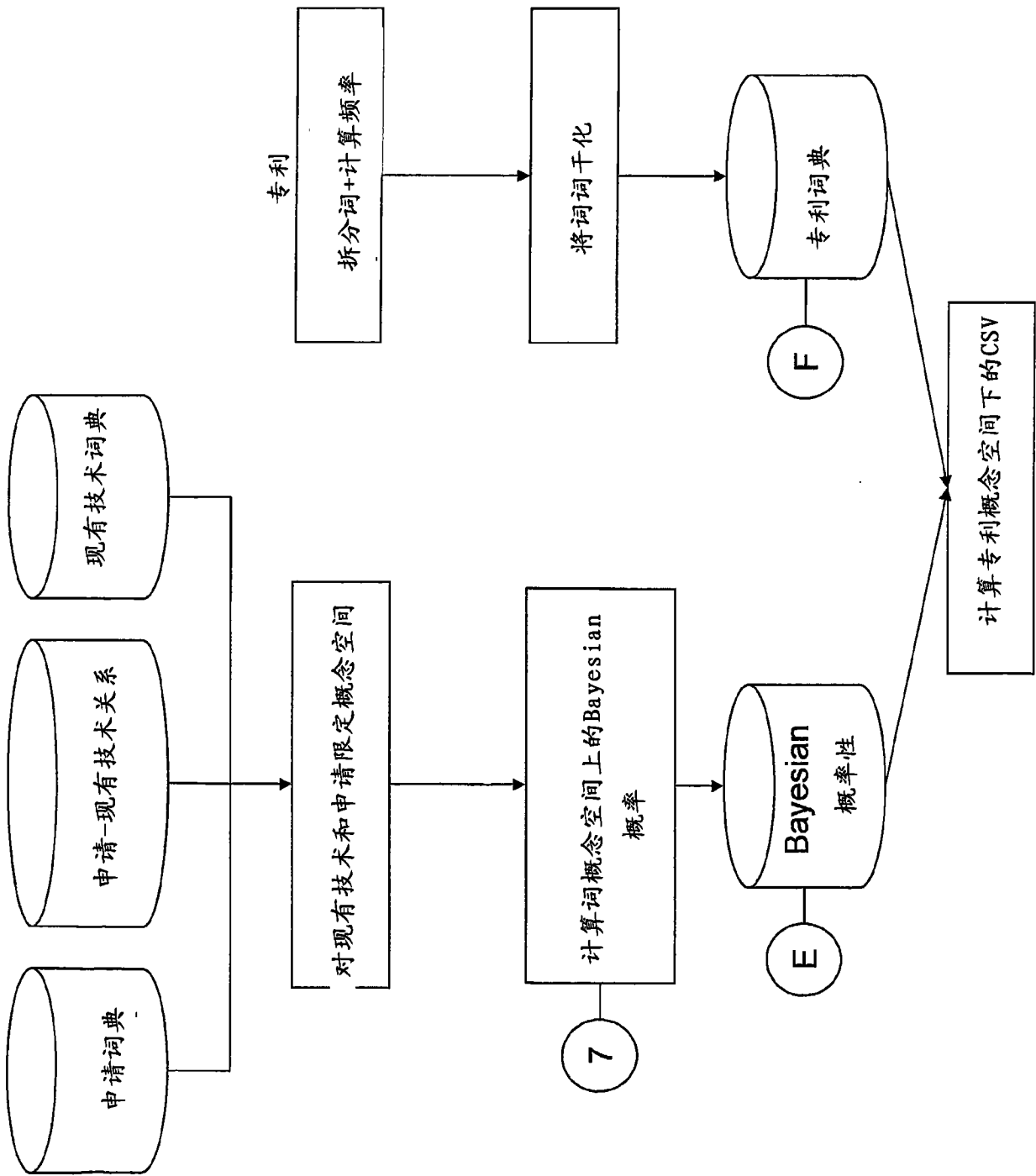


图 10

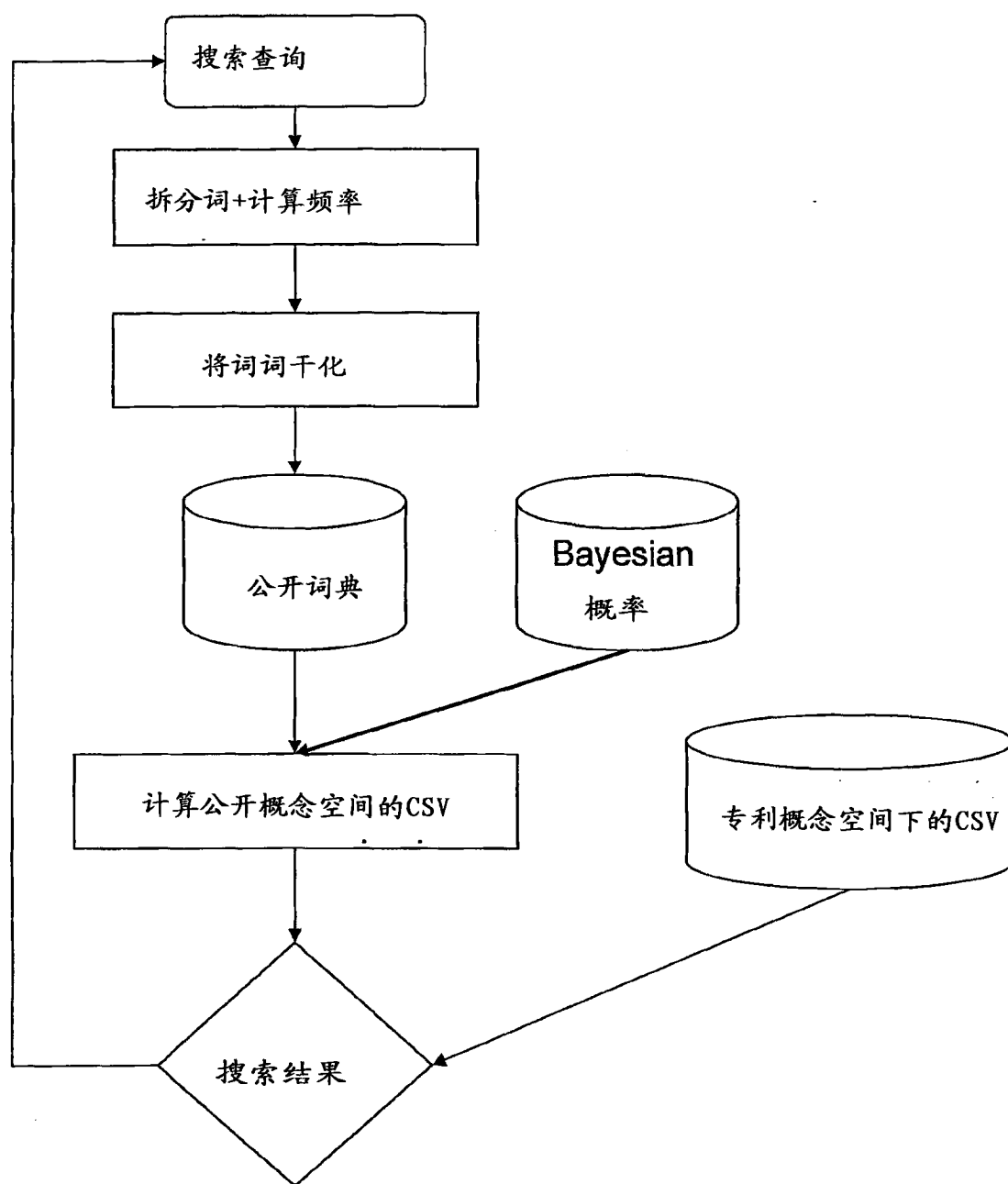


图 11

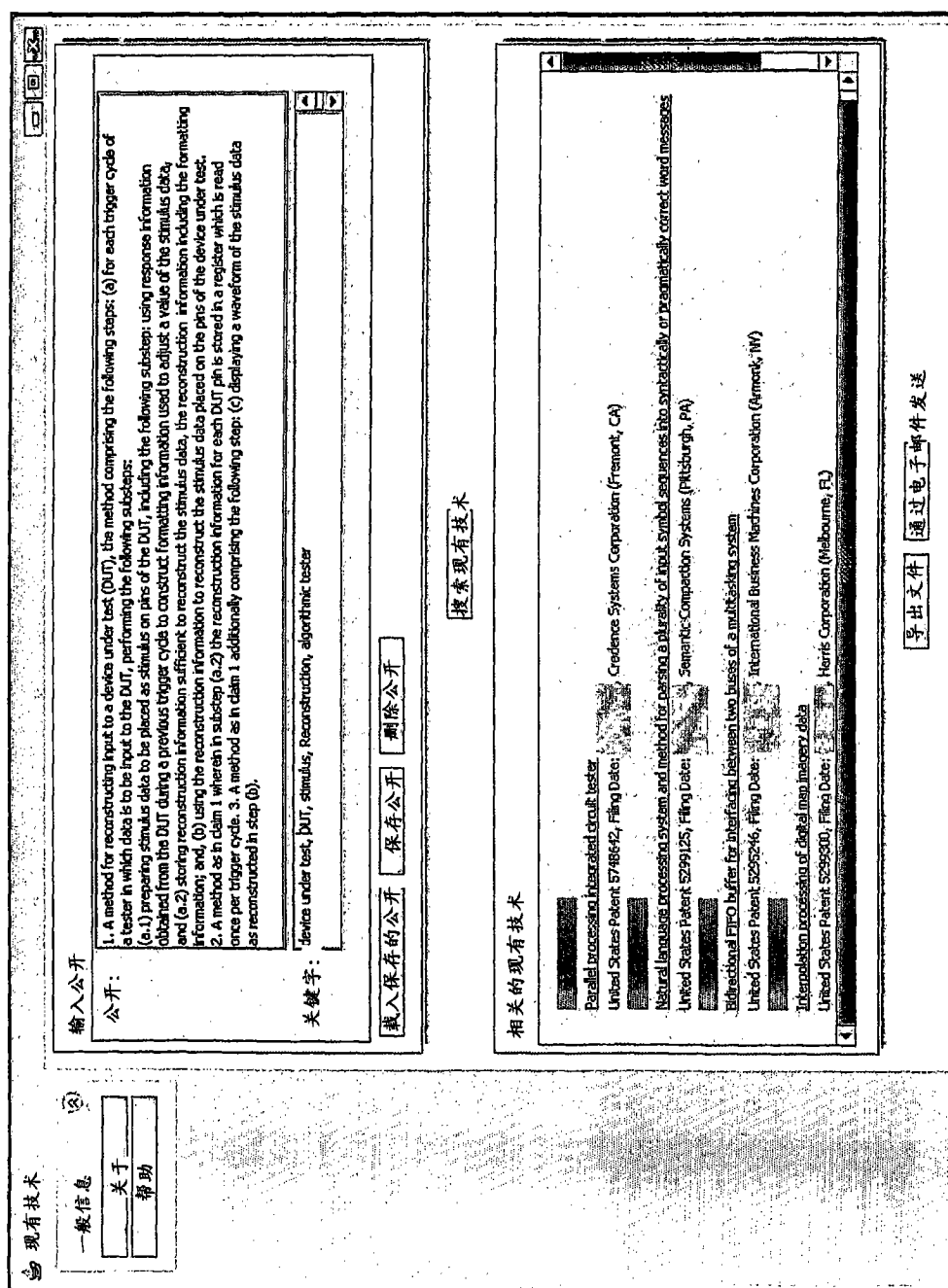


图 12

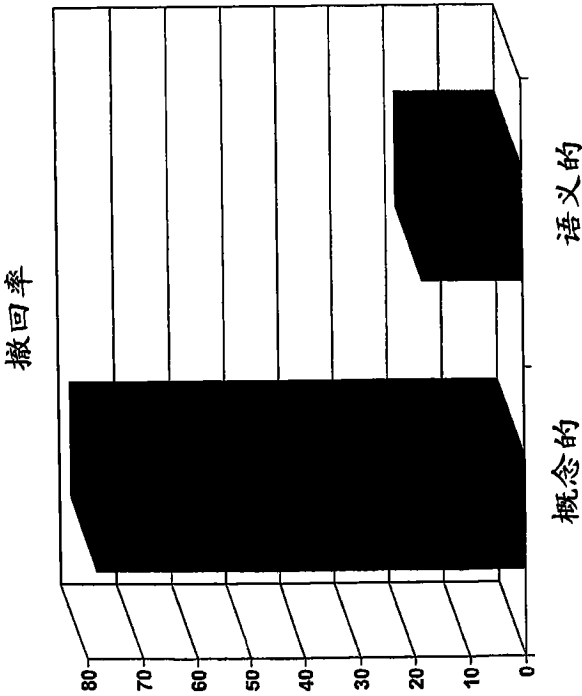


图 13

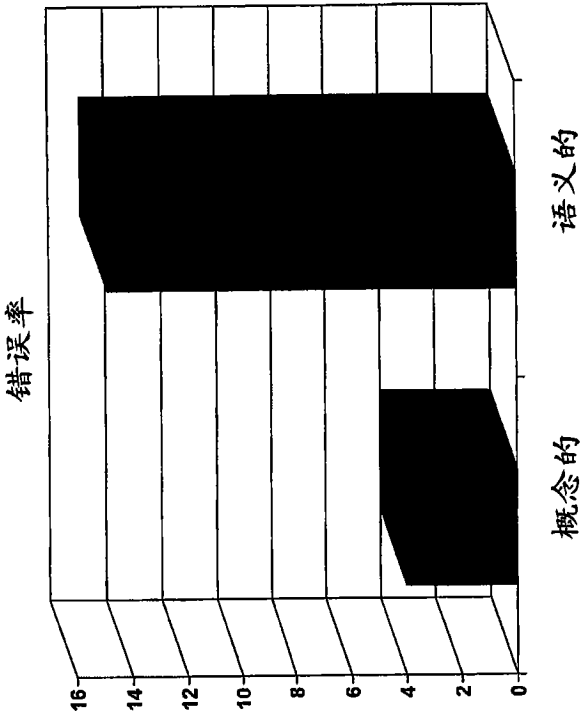


图 14

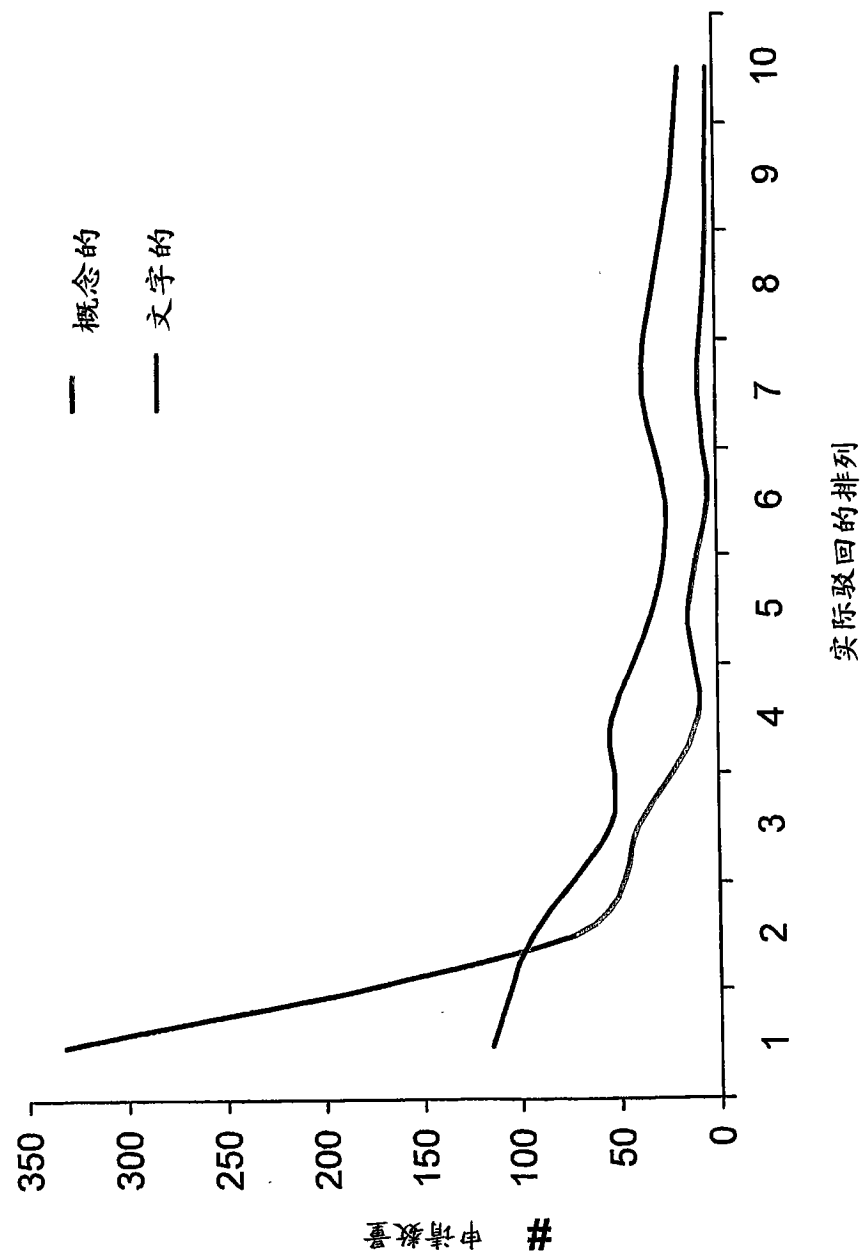


图 15

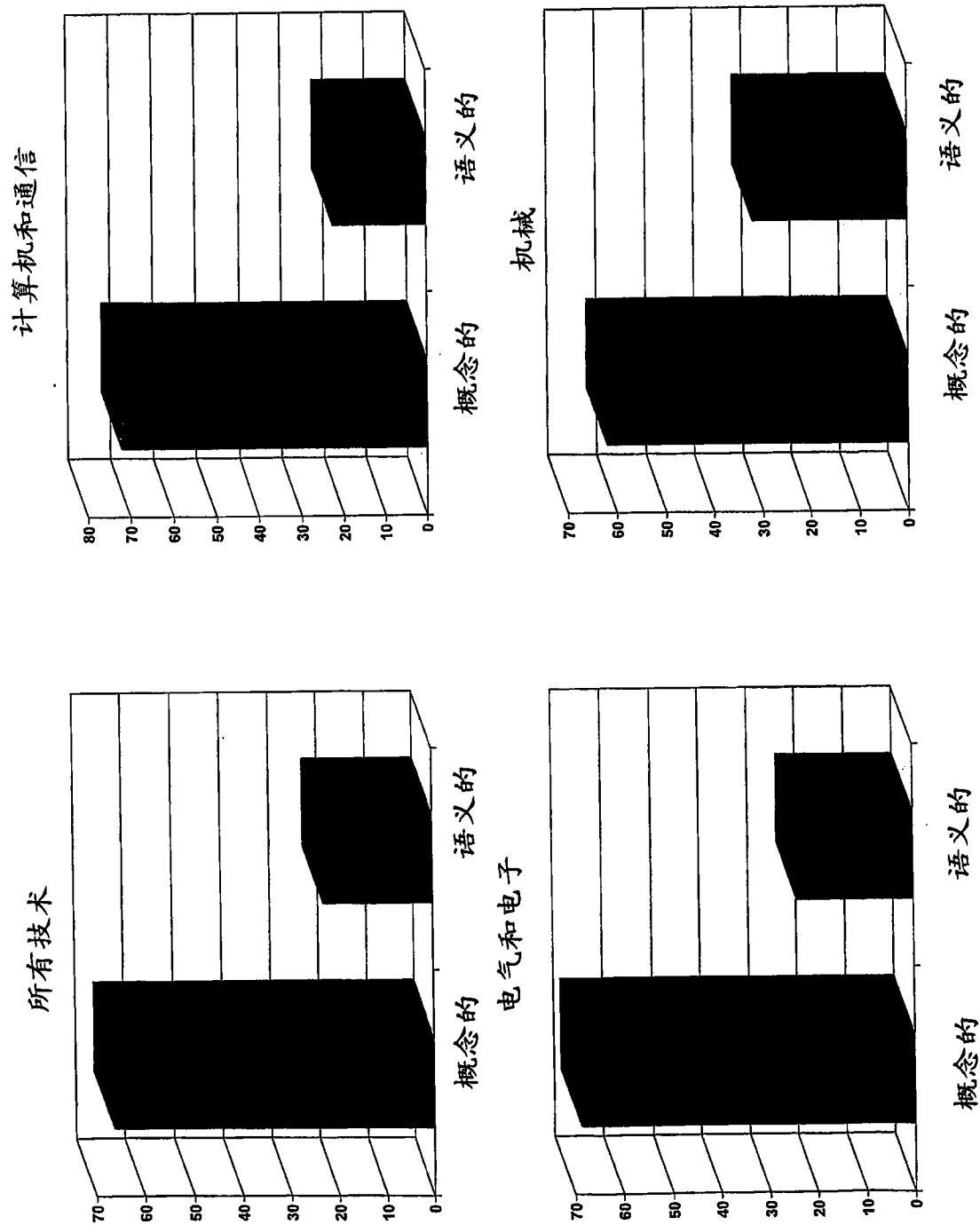


图 16

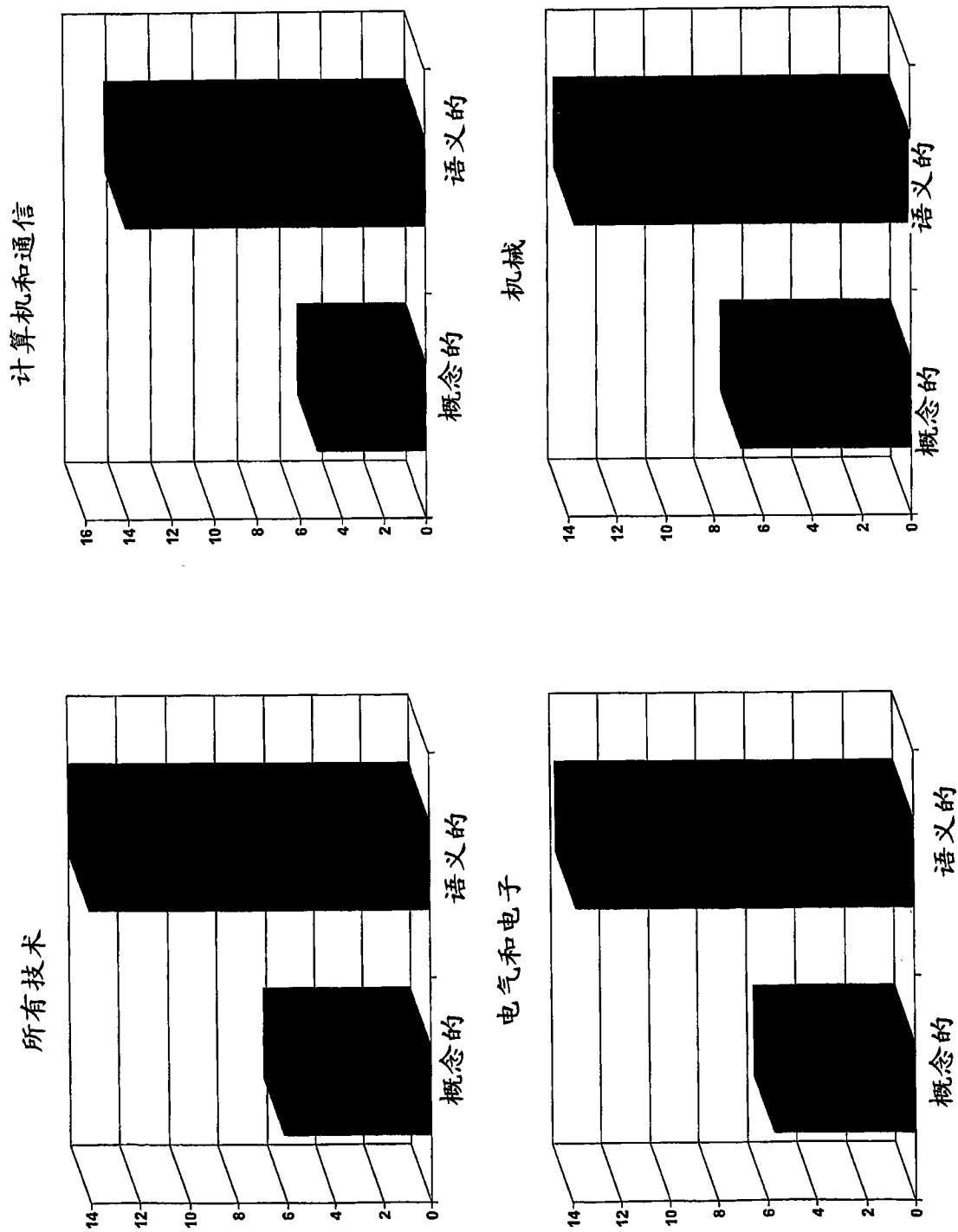


图 17

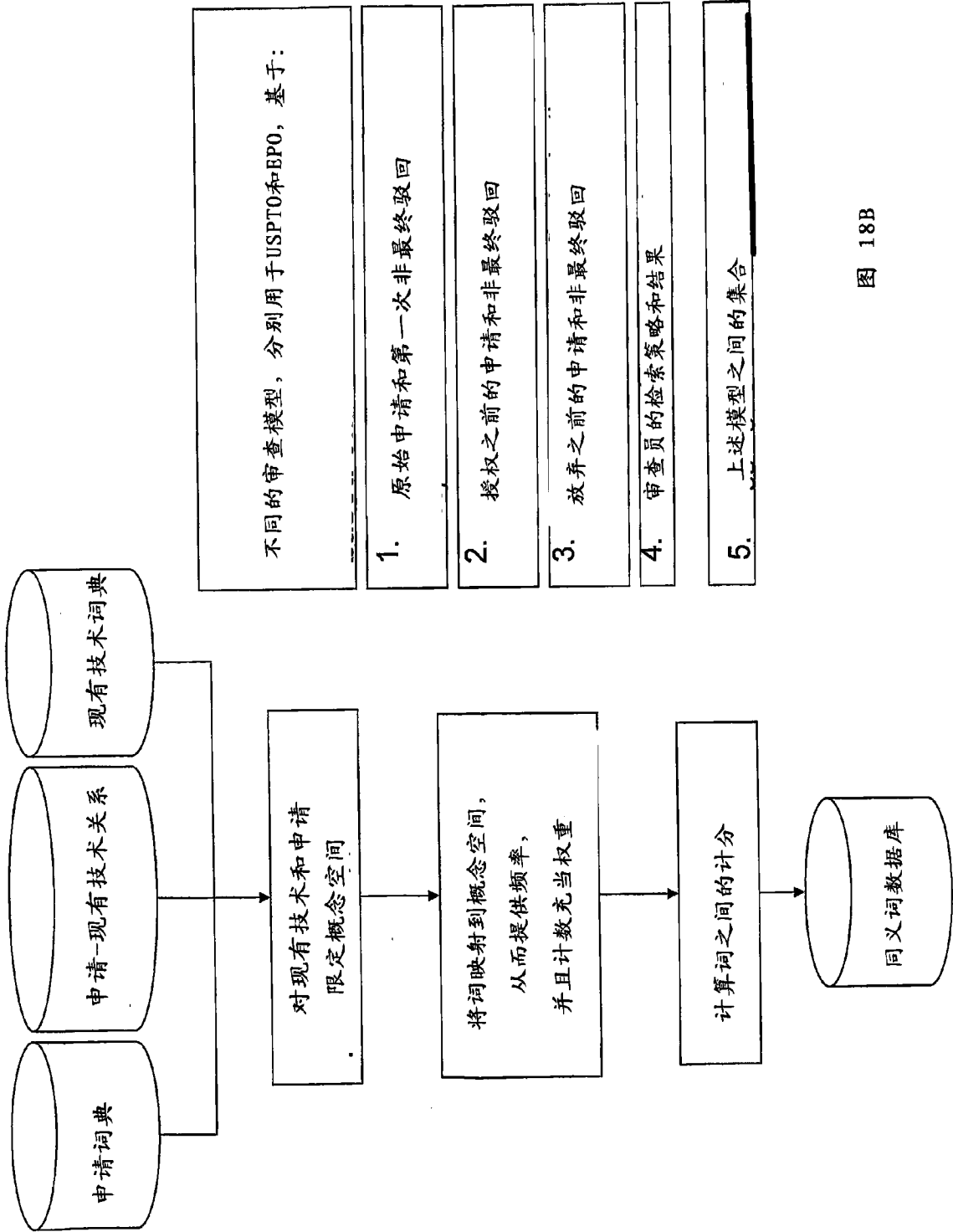


图 18A

图 18B

基于同义词选择的查询扩展搜索

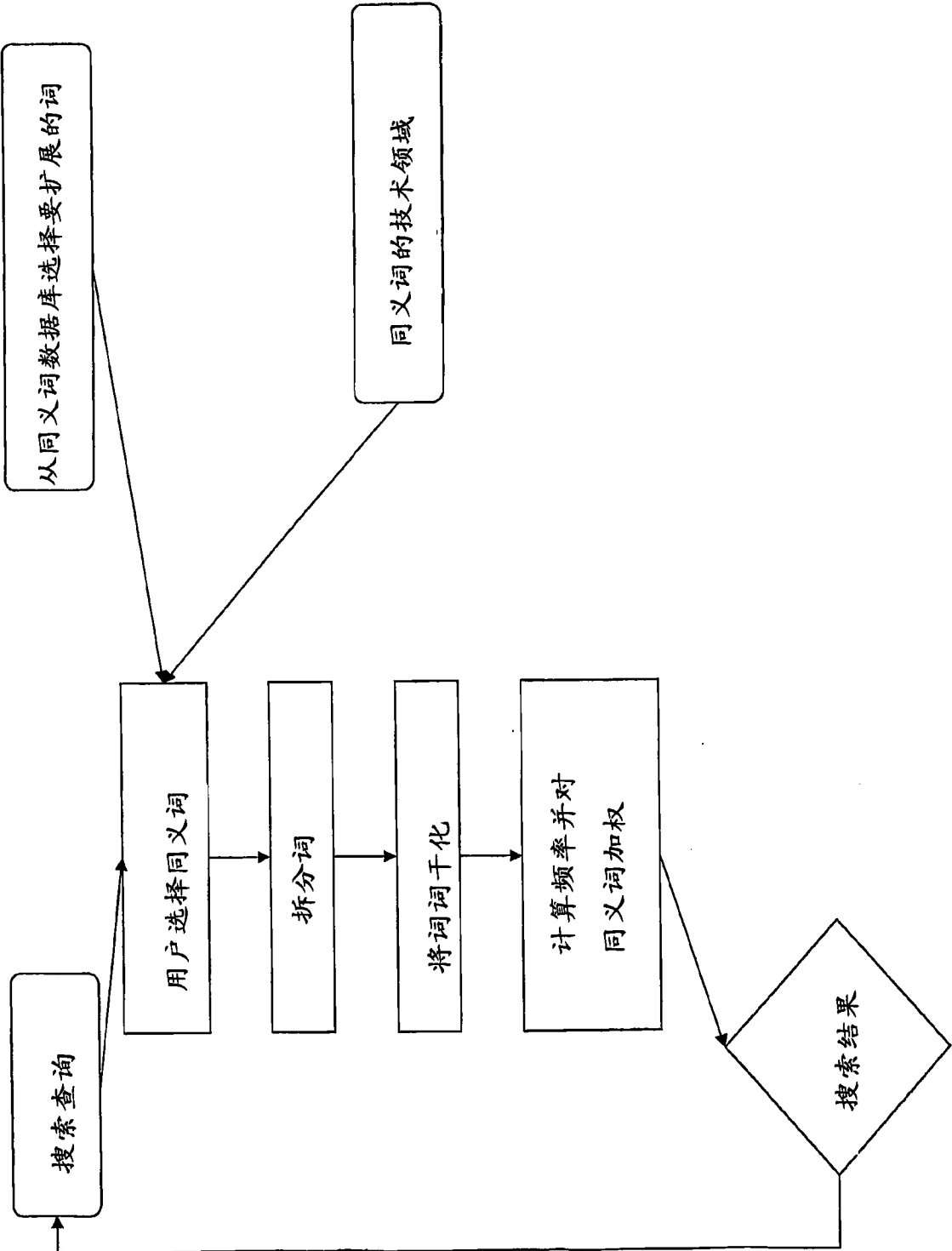


图 19