

[19] 中华人民共和国国家知识产权局



[12] 发明专利说明书

专利号 ZL 200480023961.9

[51] Int. Cl.

G06F 17/30 (2006.01)

G06F 17/27 (2006.01)

G06F 17/28 (2006.01)

[45] 授权公告日 2010 年 1 月 13 日

[11] 授权公告号 CN 100580666C

[22] 申请日 2004.8.20

US6038560A 2000.3.14

[21] 申请号 200480023961.9

审查员 蓝娟

[30] 优先权

[74] 专利代理机构 北京邦信阳专利商标代理有限公司

[32] 2003.8.21 [33] US [31] 60/496,681

代理人 黄泽雄 崔华

[86] 国际申请 PCT/CA2004/001530 2004.8.20

[87] 国际公布 WO2005/020093 英 2005.3.3

[85] 进入国家阶段日期 2006.2.21

[73] 专利权人 伊迪利亚公司

地址 加拿大魁北克省

[72] 发明人 马修·科来奇 杰里米·巴恩斯

权利要求书 3 页 说明书 12 页 附图 5 页

[56] 参考文献

US6453315B1 2002.9.17

US5873056A 1999.2.16

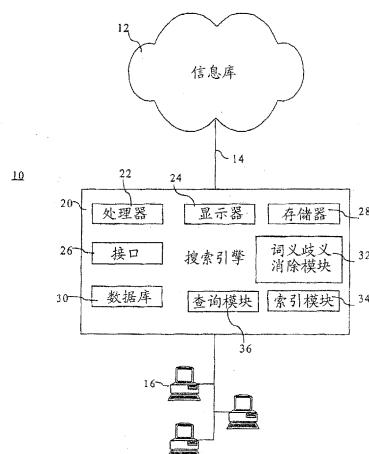
EP0597630B1 2002.7.31

WO0142984A1 2001.6.14

使用消除歧义的查询搜索消除歧义信息的方法和系统

[57] 摘要

本发明涉及一种使用查询在数据库中搜索信息的系统和方法。在该方法中，包括以下步骤：消除该查询的歧义以确定与该查询有关的关键词义；根据关键词义在数据库中消除信息的歧义；根据关键词义在数据库中索引所述信息；扩展该关键词义以包含关于该关键词义的相应语义同义词来创建一个扩展的关键词义列表；使用扩展的关键词义搜索数据库以找到关于该查询的相应信息；以及提供所述包括信息的搜索结果，该信息包含关键词义和其他语义相关的词义。该系统包括消除查询与信息的歧义的模块和在词义数据库中索引该信息的模块。



1. 一种使用查询在数据库中搜索信息的方法，所述方法包括以下步骤：

a) 消除信息库中的信息的歧义，以确定其中所含的词的关键词义；

根据所述信息的关键词义在所述数据库中索引所述信息；

以及

b) 消除所述查询的歧义以确定包含在所述查询中的词的关键词义；

使用所述查询的关键词义与其它词义的相关语义关系，扩展所述查询的关键词义来创建扩展的查询的关键词义列表；

搜索所述数据库以确定所扩展的查询的关键词义和所索引的信息的关键词义之间的匹配，所述匹配包括有关所述查询的信息；以及

提供搜索结果，所述搜索结果包括与所匹配的信息的关键词义相关的信息。

2. 按照权利要求 1 所述使用查询在数据库中搜索信息的方法，其中：

在执行步骤 b) 之前执行步骤 a) 作为预算步骤，其中，能够偶尔或定期执行步骤 a) 来保持所述数据库的流通，还能够通过对所述信息的子集合执行步骤 a) 来增量地更新所述数据库；以及，能够多次执行步骤 b) 而无需重复执行步骤 a)。

3. 按照权利要求 2 所述使用查询在数据库中搜索信息的方法，其中消除查询的歧义包括给所述查询的关键词义分配概率。

4. 按照权利要求 2 所述使用查询在数据库中搜索信息的方法，其中在所述信息库中消除所述信息的歧义包括为所述信息的关键词义添加概率。

5. 按照权利要求 3 所述使用查询在数据库中搜索信息的方法，其中所述搜索所述数据库的步骤还包括：在确定所述查询的关键词义和所述信息的关键词义之间的匹配时使用所述查询的关键词义的概率作为排列手段。

6. 按照权利要求 2 所述使用查询在数据库中搜索信息的方法，其中所述扩展所述查询的关键词义的步骤还包括通过解析查询的关键词义的语法结构来解释所述查询，并使用原始查询的关键词义或所扩展的查询的关键词义将其转换为其他语义相等的查询。

7. 按照权利要求 6 所述使用查询在数据库中搜索信息的方法，其中所述查询的关键词义和所述信息的关键词义分别表示各自精细关键词义的粗略分组。

8. 按照权利要求 2 所述使用查询在数据库中搜索信息的方法，其中所述查询的关键词义和所述信息的关键词义分别表示各自精细关键词义的粗略分组。

9. 按照权利要求 2 所述使用查询在数据库中搜索信息的方法，其中所述歧义消除的步骤包括使用词义间语义关系知识库，所述知识库与所述数据库关联。

10. 按照权利要求 2 所述使用查询在数据库中搜索信息的方法，其中根据所述查询的关键词义和所匹配的信息的关键词义之间的语义关系对所述结果进行加权。

11. 按照权利要求 2 所述使用查询在数据库中搜索信息的方法，其中根据分配给所述查询的关键词义的概率和分配给所匹配的信息的关键词义间的这些相同含义的概率对所述结果进行加权。

12. 按照权利要求 4 所述使用查询在数据库中搜索信息的方法，其中所述搜索所述数据库的步骤还包括：在确定所述查询的关键词义和所述信息的关键词义之间的匹配时使用所述查询的关键词义的概率作为排列手段。

13. 一种响应查询从数据库提供信息的系统，所述系统包括：

一个数据库，包含将被所述查询搜索的信息；

一个索引模块，为将被所述查询所使用的信息创建一个参照索引；

一个查询处理模块，将所述查询应用于所述数据库；

一个歧义消除模块，用于消除包含在所述信息和所述查询中的词的歧义以分别确定信息的关键词义和查询的关键词义，

其中

所述索引模块根据所述信息的关键词义在所述数据库中索引所述信息；

所述查询处理模块：

使用所述查询的关键词义与其它词义的相关语义关系，扩展所述查询的关键词义来创建扩展的查询的关键词义列表；通过将所述扩展的查询的关键

词义与在所述数据库中的索引的信息的关键词义匹配而搜索所述数据库从而为所述查询找到相关信息；

以及提供包含所述相关信息的搜索结果。

14. 按照权利要求 13 所述响应查询从数据库提供信息的系统，其中所述歧义消除模块根据所述查询的关键词义和所匹配的信息的关键词义之间的语义关系来排列所述查询的关键词义。

15. 按照权利要求 14 所述响应查询从数据库提供信息的系统，其中所述查询的关键词义和所述信息的关键词义分别表示各自精细关键词义的粗略分组。

16. 按照权利要求 13 所述响应查询从数据库提供信息的系统，其还包括词义间语义关系知识库，所述知识库与所述数据库关联并与所述歧义消除模块关联。

使用消除歧义的查询搜索消除歧义信息的方法和系统

相关申请

本申请要求 2003 年 8 月 21 日提交的申请号为 60/496,681 的美国临时专利申请的优先权。

技术领域

本发明涉及因特网搜索，尤其涉及使用语义歧义消除与扩展的因特网搜索。

背景技术

在处理庞大数据集时，例如文档数据库或因特网上的网页，可用数据的容量会造成难以找到有关信息。尝试使用了各种各样的搜索方法来在这样的信息库中找到有关信息。其中一些最好的公知系统是网络搜索引擎，例如允许用户执行基于关键词搜索的 Yahoo (商标) 和 Google (商标)。这些搜索通常涉及将用户输入的关键词与网页索引中的关键词进行匹配。

然而，现有的因特网搜索方法经常产生不是特别有用的结果。该搜索会返回很多结果，但仅有少量与用户的查询有关或根本没有与用户查询有关的。另一方面，该搜索会返回少量的结果，其中没有一个是用户正在搜寻的精确结果，同时无法返回潜在的相关结果。

在执行这样的搜索中遭遇一些困难的原因是在自然语言中使用的单词的歧义。特别是，由于一个单词会具有多种含义，所以经常遭遇困难。过去通过使用被称为词义歧义消除的技术处理该难题，该技术包括将单词转换成具有特定的语义的词义。例如，单词 “bank” 可以具有所附的 “金融机构” 的含义或其他定义。

第 6,453,315 号美国专利公开了基于含义的信息组织和检取 (retrieval)。该专利公开了通过概念词典 (lexicon of concepts) 与概念之间的关系来创建一个语义空间。将查询映射到表示该查询的位置与语义空间的含义区分器。通

过确定区分器之间的语义差别以确定接近 (closeness) 和含义来完成搜索。该系统依赖用户以该系统所确定的含义为基础来精炼该搜索或另外通过在搜索结果中所找到的节点来导航。

如现有技术中已知的，通过“精度”和“回想”来对信息检取的有效性评价进行量化。通过用在一个搜索中找到的正确结果的数量除以结果的整体数量可以量化精度。通过用在一个搜索中找到的正确结果的数量除以可能正确的结果的整体数量可以量化回想。可以简单地通过返回所有可能的结果来获得完美的回想 (即 100%)，但是这会导致非常差的精度。大多数现有的系统力争平衡精度与回想的标准。例如通过使用同义词来提供更多的可能结果而提高回想必然会降低精度。另一方面，通过缩小搜索结果，例如通过选择与查询中的单词的精确顺序相匹配的结果将降低回想。

因此需要一种解决现有技术的缺陷的查询处理系统和方法。

发明内容

根据本发明的一个方面，提供了一种信息搜索方法，其包括以下步骤：消除查询的歧义，根据关键词义消除信息的歧义并索引信息，通过使用该查询中的关键词义以及与该查询中的关键词义语义相关的其他词义搜索该已索引的信息以找到与该查询相关的信息，以及返回搜索结果，所述搜索结果包括包含关键词义和其他语义相关的词义的信息。

该方法可以应用于任何使用关键词进行索引的数据库。优选地，该方法被应用于因特网搜索。

语义关系可以是任何逻辑上或语法上定义的两个单词之间的关系类型。这种关系的实例为同义关系、下义关系等。

消除查询的歧义的步骤可以包括给关键词义分配概率。类似地，消除信息的歧义的步骤可以包括为关键词义添加概率。

在本方法中使用的关键词义是比较精细的关键词义的粗略分组。

在另一方面，提供一种使用查询在数据库中搜索信息的方法。该方法包括以下步骤：根据关键词义在数据库中消除信息的歧义；根据关键词义在数据库中索引信息；消除该查询的歧义以确定与该查询有关的关键词义；扩展

该关键词义以包含该关键词义的相关语义关系从而创建一个扩展的关键词义列表；使用扩展的关键词义搜索数据库以找到关于该查询的相关信息；以及提供所包括的信息的搜索结果，该信息包含关键词义和其他语义相关的词义。

在该方法中，在数据库中消除信息的歧义可以包括为关键词义添加概率。可以利用多个含义索引该信息中的单词，并且该含义的概率可以与该含义一起被存储在索引中。

在该方法中，消除查询的歧义可以包括给所述关键词义分配概率。

在该方法中，消除该查询的歧义以确定特定的关键词义还可以包括使用每一个所述特定关键词义的概率。

在该方法中，扩展特定的关键词义还包括通过解析特定的关键词义的语法结构来解释该查询并确定其他语义相等的查询。

在该方法中，该关键词义可以表示精细的关键词义的粗略分组。

在另一方面，提供一种响应查询从数据库提供信息的系统。该系统包括：一个包含将被查询搜索的数据的数据库；一个为将被该查询所使用的数据创建一个参照索引的索引模块；一个将所述查询应用于数据库的查询处理模块；以及一个用于消除查询的歧义以确定与该查询有关的关键词义的歧义消除模块。尤其对于该系统：歧义消除模块根据关键词义在数据库中消除信息的歧义；索引模块根据关键词义在数据库中索引信息；查询处理模块扩展关键词义以包含关于所述关键词义的相应语义同义词来创建一个扩展的关键词义列表，通过使用扩展的关键词义开始数据库搜索以便为该查询找到相关的信息；以及提供包括信息的搜索结果，该信息包含关键词义和其他语义相关的词义。

在该系统中，歧义消除模块给关键词义分配概率来排列所述关键词义。可以利用多个含义索引该信息中的单词，并且该含义的概率可以与该含义一起被存储在索引中。在该系统中，该关键词义表示精细的关键词义的粗略分组。

该系统还可以结合前面所描述的方法所提到的其他方面的功能。

另一方面，提供以上方面的集合或子集合的各种组合。

附图说明

本发明上述和其他方面将会根据以下其特定实施例和仅借助于实例说明本发明原理的附图的说明而变得更为明显。在附图中，同一附图标记表示相同的元素特征（并且其中单个元素带有唯一的字母后缀）：

图1是一个提供与本发明实施例相关的词义歧义消除的信息检索系统的示意图；

图2是与图1的系统有关的单词与词义的示意图；

图3A是用于图1的系统的代表性的语义关系或单词的示意图；

图3B是用来表示用于图1的系统的图3A的语义关系的数据结构图；以及

图4是由图1的系统使用图2的词义以及图3A的语义关系所执行的方法的流程图。

具体实施方式

借助于对一个实例或多个实例，尤其是本发明原理内特定实施例的举例说明来提供以下说明及其中所描述的实施例。提供这些实例的目的在于解释而非限制本发明的原理。在以下的说明中，在整个说明书与附图中用相同的各个附图标记标注相同的部件。

以下术语将在后续的说明中被使用并具有以下所示的含义：

计算机可读存储介质：用于存储关于计算机的指令或数据的硬件。例如，磁盘、磁带、诸如CDROM那样的光学可读介质，以及诸如PCMCIA（个人计算机存储卡国际协会）卡那样的半导体存储器。在每一种情况下，该介质可以采取诸如小型磁盘、软盘、盒式磁带那样的便携物件的形式，或采取诸如硬盘驱动器、固态存储卡或RAM那样的相对较大或固定物件的形式。

信息：包含可搜索的、用户感兴趣的内容的文档、网页、电子邮件、图像描述、抄本、存储文本等，例如，与新闻文章、新闻组消息、网络日志等有关的内容。

模块：执行特定步骤和/或处理过程的软件或硬件组件；可以在运行在通用处理器上的软件中实现。

自然语言：希望被人而非机器或计算机所理解的单词表达。

网络: 配置成通过使用特定协议在通信信道上进行通信的设备的互连系统。它可以是一个局域网、广域网，因特网或在通信线路上或通过无线传输工作的类似网络。

查询: 表示要求的搜索结果的一列关键词；可以使用布尔运算符(例如“与”、“或”); 可以用自然语言表示。

查询模块: 处理一个查询的硬件或软件组件。

搜索引擎: 响应来自用户的查询来提供涉及该用户感兴趣的信息的搜索结果的硬件或软件组件。可以根据关联性排列和/或分类查询结果。

参照图 1，用数字 10 整体指示与一个实施例有关的信息检取系统。该系统包括一个可以通过网络 14 访问的信息库 12。还可以使用其他已知的访问方法。信息库 12 可以包括文档、网页、数据库等。优选地，该网络 14 是因特网，而且信息库 12 包含网页。当网络 14 是因特网时，该协议包括 TCP/IP (传输控制协议/网络协议)。各种客户机 16 通过在物理网络情况下的线路或通过无线发射器和接收器连接到网络 14。每一个客户机 16 包括一个为本领域技术人员所理解的网络接口。网络 14 为客户机 16 提供对信息库 12 中内容的访问。为了使客户机 16 能够在信息库 12 内找到特定的信息、文档、网页等，系统 10 被配置成允许客户机 16 通过提交查询来搜索信息。该查询包括至少一列关键词，而且还具有采取诸如“与”和“或”那样的布尔关系的形式的结构。还可以在自然语言中构成作为句子 (sentence) 或提问 (question) 的所述查询。

该系统包括搜索引擎 20，其连接到网络 14 来接收来自客户机 16 的查询以将它们导向信息库 12 内的单个文档。搜索引擎 20 可以被实现为专用硬件，或运行在通用处理器上的软件。该搜索引擎运行以定位信息库 12 中与来自客户机的查询相关的文档。

搜索引擎 20 通常包括一个处理器 22。该引擎还可以直接连接到或通过网络或其他诸如此类的通信方式间接地连接到显示器 24、接口 26 和计算机可读存储介质 28。处理器 22 与显示器 24 和接口 26 耦合，所述接口 26 可以包括诸如键盘、鼠标或其他适合的设备那样的用户输入设备。如果显示器 24 是对触摸敏感的，则可以将显示器 24 本身用作接

口 26。计算机可读存储介质 28 与处理器 22 耦合用以向处理器 22 提供指令来指示和/或设置处理器 22 来执行与搜索引擎 20 的操作相关的步骤或算法，如接下来所解释的。计算机可读存储介质 28 的一部分或全部都可以在物理上被置于搜索引擎 28 之外以容纳例如非常大的存储量。本领域技术人员将理解可以在本发明中使用多种形式的搜索引擎。

可选地并且为了更高的计算速度，搜索引擎 20 可以包括多个并行工作的处理器或任何其他的多处理结构。这样使用多个处理器可以使搜索引擎 20 在多个处理器中划分任务。此外，如本领域技术人员所理解的，所述多个处理器不必在物理上被置于相同位置，而可以将它们地理上分离地设置并在网络上互连。

优选地，搜索引擎 20 包括数据库 30，所述数据库 30 用于存储词义的索引并存储由搜索引擎 20 所使用的知识库。如本领域技术人员所理解的，数据库 30 以结构化方式存储索引以实现计算地有效的存储和检索。可以通过添加另外的关键词义或对另外的文档引用现有的关键词义来更新数据库 30。数据库 30 还提供一种能够确定哪一个文档包含特定的关键词义的检索能力。还可以为了更高的效率而划分数据库 30 并在多个位置中进行存储。

根据一个实施例，搜索引擎 20 包括一个用于处理输入文档中的单词或者对词义的查询的词义歧义消除模块 32。词义是考虑到一个单词使用的上下文 (context) 及其相邻单词而赋予该单词的特定解释。例如，句子 “为我预定到纽约的航班 (Book me a flight to New York)” 中的单词 “book” 是歧义的，因为 “book” 可以是一个名词或动词，该名词或动词的每一个都具有多个潜在的含义。由歧义消除模块 32 进行单词处理的结果是包含词义的一个已消除歧义的文档或是一个已消除歧义的查询，而不是歧义的或未解释的单词。该输入文档可以是信息库中的任何信息单元或从客户机接收的查询中的一个。词义歧义消除模块 32 为文档或查询中的每个单词辨别词义。词义歧义消除模块 32 通过使用广泛的互连语言技术 (interlinked linguistic technique) 来确定单词的哪一个特定含义是所期望的含义以分析上下文中的语法 (例如词

性、语法关系)和语义(例如逻辑关系)。词义歧义消除模块32在执行歧义消除时,可以使用表示词义之间明确的语义关系的词义知识库来加以辅助。该知识库可以包括以下参照图3A和3B所描述的关系。

搜索引擎20包括一个索引模块34,该索引模块用于处理一个已消除歧义的文档来创建关键词义的索引并在数据库30中存储该索引。所述索引包括用于与文档相关的每个关键词义的一个条目,在文档中可以找到该关键词义。该索引最好被分类并包括每一个已索引的关键词义的位置指示。索引模块34通过处理已消除歧义的文档并将每个关键词义添加到索引来创建该索引。某些关键词会出现太多次而无用和/或几乎不包含语义信息,诸如“a”或“the”。对这些关键词将不进行索引。

搜索引擎20还包括一个用于处理来自客户机16的查询的查询模块36。查询模块36被配置成接收查询并将它们传送到消除歧义模块32用于处理。如接下来所描述的,查询模块36随后在索引中找到与已消除歧义的查询有关的结果。该结果包括与已消除歧义的查询中的词义语义相关的关键词义。查询模块36将结果提供给客户机。可以根据关联性对该结果进行排列和/或打分以帮助客户理解它们。

参见图2,用附图标记100整体指示单词与词义之间的关系。从该实例可见,某些单词具有多个含义。在多个其他可能性中,单词“bank”可以表示:(i)涉及金融机构的名词;(ii)涉及河岸的名词;或者(iii)涉及一种攒钱行为的动词。词义歧义消除模块32将带有歧义的单词“bank”分成几个具有较轻歧义的词义用于存储在索引中。类似地,单词“interest”具有多个含义,包括:(i)表示涉及一种未偿还的投资或贷款的应支付的金钱数额的名词;(ii)表示对某事/某物特别注意的名词;或者(iii)表示在某事/某物中的合法权利的名词。

参照图3A和图3B,显示了词义之间的实例语义关系。这些语义关系是基于含义所精确定义的两个单词之间的关系类型。此关系是在词义之间的,即单词的特定含义。

尤其是在图3A中,例如,单词“bank”(取河岸的含义时)是一种地形而单词“bluff”(取意味着一种陆地构造(land formation)的名词

时)也是一种地形。单词“bank”(取河岸的含义时)是一种斜坡(取地面坡度的含义)。单词“bank”取金融机构的含义时与“银行公司”或“银行中心(banking concern)”同义。单词“bank”还是一种金融机构,所述金融机构也是一种商业类型。根据通常所理解的银行在存款上支付利息并在贷款上收取利息的事实,单词“bank”(取金融机构的含义)涉及单词“interest”(取为投资支付的钱的含义)并且也涉及单词“loan”(取贷款的含义时)。

应当理解存在很多其他类型的可使用的语义关系。尽管在现有技术中已知,以下是一些单词之间的语义关系的实例:处于同义词中的单词就是彼此同义的词。上义词是一种关系,其中一个词表示整个一类的特定例子。例如“运输工具”是用于包括“火车”、“战车(chariot)”、“狗拉的雪橇”和“汽车”的一类词的上义词,因为这些词提供该类别的特定例子。同时,下义词是一种关系,其中一个词是一类例子中的一个成员。根据之前的列表,“火车”是“运输工具”类别的下义词。局部词是一种关系,其中一个词是某事物的一个组成部分、一个成分(substance)或一个成员。例如,关于“腿”与“膝盖”之间的关系,“膝盖”是“腿”的局部词,因为膝盖是腿的一个组成部分。同时,整体词是一种关系,其中一个词是被称为一部分的局部词的全部。根据之前的例子,“腿”是“膝盖”的整体词。可以使用落入这些分类的任何语义关系。另外,可以使用任何公知的指出词义之间的特定语义和语法关系的语义关系。

已知当提供关键词的字符串作为查询时在解释上存在歧义,以及在查询中带有扩展的关键词列表增加了在搜索中找到的结果的数量。该实施例提供了一种系统和方法来为查询确定关联的、已消除歧义的关键词列表。提供这样一个按照词义所描绘的列表减少了检取到的无关信息的数量。该实施例扩展了查询语言而不会由于一个单词的附加含义而获得无关结果。例如,扩展单词“bank”的“金融机构”的含义不会同时扩展诸如“河岸”或“存钱”的其他含义。这允许信息管理软件更精确地确定客户正在查找的信息。

扩展一个查询涉及使用以下一个或两个步骤：

1. 向一个已消除歧义的查询关键词义添加其他单词以及该单词相关的含义，所述该单词相关的含义是指与已消除歧义的关键词的含义语义上相关的含义。
2. 通过解析其语法结构来解释该查询并将其转换成其他语义相等的查询。索引包含确定从该信息的语法结构中衍生的关键词义对之间的语义从属性的字段。解释是本领域公知的术语和概念。

还应当认识到在搜索中使用词义歧义消除解决了检取关联性的问题。此外，用户经常如同表达语言一样表达查询。然而，由于可以以多种不同的方式描述相同的含义，所以用户就在他们不以同一特定方式表达一个查询时遭遇困难，其中以该特定方式对关联信息进行初始分类。

例如，如果用户正在查找有关岛屿“爪哇（Java）”的信息，并对在爪哇（岛屿）上的“假日（holidays）”感兴趣，那么用户就不会检取到已经通过使用关键词“爪哇（Java）”和“休假（vacation）”进行分类的有用的文档。应当认识到，根据实施例的语义扩展特性解决了这个问题。已经认识到在自然表达的查询中为每一个关键术语衍生精确的同义词和子概念（sub-concept）增加了关联性检取的容量。如果通过使用词表（thesaurus）来执行检取且不执行词义歧义消除就会恶化该结果。例如，语义上扩展单词“Java”而没有首先确定其精确含义将产生大规模且难于处理的结果集合，该集合带有潜在地基于不同的词义选定的结果，所述不同的词义例如为印度尼西亚”和“计算机程序设计”。还将理解所描述的解释每一个单词的含义然后语义上扩展该含义的方法返回一个更全面同时具有更多目标的结果集合。

参照图 3B，为了帮助消除这种词义的歧义，该实施例使用如以上对于图 3A 所描述的获得单词关系的词义知识库 400。知识库 400 与数据库 300 相关联并通过访问以帮助 WSD 模块 32 执行词义歧义消除。知识库 400 包含对于一个单词的每个词义的词的定义，还包含词义对之间的关系的信息。这些关系包括词义和相关词性（名词、动词等）的定义、精细词义同义词、反义词、下义词、局部词、与名词相关的形容词

(pertainym)、类似的形容词关系以及现有技术中已知的其他关系。当在系统中使用了现有技术的电子词典和词汇数据库时，例如 WordNet(商标)，知识库 400 提供增强的单词与关系的目录。知识库 400 包括：(i) 词义之间的附加关系，例如将精细的含义归合到粗略的含义，新型的屈折 (inflectional) 和派生 (derivational) 的词素 (morphological) 关系，以及其他特殊用途的语义关系；(ii) 对来自出版源 (published source) 的数据中的错误的大规模校正；以及 (iii) 在其他现有技术知识库中不存在的其他的单词、词义以及相关关系。

在该实施例中，知识库 400 是一种概括的图形数据结构并作为节点表 402 和有关连接两个节点的边缘关系表 404 来实现。每一个都依次被描述。在其他实施例中，还可以使用其他诸如链接列表那样的数据结构来实现知识库 400。

在表 402 中，每一个节点是表 402 一个行元素。每一个节点的记录可以具有多至以下的字段：ID 字段 406，类型字段 408 和注释字段 410。在表 402 中存在两种类型的条目：单词与词义定义。例如，通过类型字段 408A 中的条目“单词”确定 ID 字段 406A 中的单词“bank”为一个单词。此外，示范性的表 402 提供单词的多个定义。为了对所述定义进行分类并区分表 402 中的单词条目与定义条目，可以使用标签来确定定义条目。例如，将 ID 字段 406B 中的条目标记为“标签 001”。类型字段 408B 中的一个相应的定义将该标签标记为“精细的含义”单词关系。注释字段 410B 中的一个相应的条目将该标签标记为“名词，金融机构”。这样，现在可以将单词“bank”连接到该词义定义。此外，还可以将单词“经纪行 (brokerage)”的条目连接到该词义定义。另一个实施例可以使用带有附加后缀的常用单词，以便辅助识别该词义定义。例如，另一种标签可以为“银行/n1”，其中后缀“/n1”表明该标签为名词并且是该名词的第一含义。应当理解可以使用其他形式的标签。可以使用其他标识符来确定形容词、副词和其他词性。在类型字段 408 中的条目确定了与单词相关的类型。一个单词存在多种有效的类型，包括：单词，精细的含义和粗略的含义。还可以提供其他类型。在本实施例中，当一

一个单词实例具有一个精细的含义时，该实例还具有注释字段 410 中的一个条目来提供关于该单词实例的更多细节。

边缘/关系表 404 包含表示节点表 402 中两个条目之间关系的记录。表 404 具有以下条目：源节点 ID 栏 412、目的节点 ID 栏 414、类型栏 416 和注释栏 418。栏 412 与栏 414 一起用来连接表 402 中的条目。栏 416 确定连接两个条目的关系类型。记录具有源节点和目的节点的 ID、关系的类型并且可能具有基于该类型的注释。关系的类型包括“根单词到单词”、“单词到精细含义”、“单词到粗略含义”、“粗略含义到精细含义”、“衍生”、“下义词”、“类别”、“与名词相关的形容词”、“类似”、“具有部分”。还可以在其中记录其他关系。注释栏 418 中的条目提供一个（数字）键来为一给定的词性确定一种从一单词节点到粗略的节点或精细的节点的边缘类型。

现在提供对有关该实施例通过利用根据消除与一个查询有关的一个单词的歧义的结果来执行一个搜索的执行步骤更详细的描述。参照图 4，通过附图标记 300 整体地指示执行这样一个搜索的过程。可以将该过程划分为两个大致阶段。第一阶段包括预先处理信息（或信息的子集合）来辅助响应查询的第二阶段。在第一阶段预处理，概括信息库中的每一个文档（或信息库的子集合）来创建数据库中的索引。在步骤 302，词义歧义消除模块 32 对每个文档中的每一个单词进行词义之间的区分。所述词义歧义消除模块 32 在前面已经定义过了。

然后在步骤 304 中，搜索引擎将索引模块应用到已消除歧义的信息从而获得关键词义的索引。索引模块 34 通过处理已消除歧义的文档并将每一个关键词义添加到索引来创建该索引。某些关键词会出现太多次而无用，例如“a”或“the”。优选地，对这些关键词不进行索引。应当理解，该步骤有效地索引一个单词当作几个不同的词义。在步骤 306 中，将词义的索引存储到数据库中。

在该处理的第二阶段，在步骤 308 中，搜索引擎接收来自客户机中的一个客户机的查询。将该查询解析成单词组件，然后可以对每个单词单独分析其上下文以及与其邻近的单词一起分析其上下文。对单词串的

解析技术是本领域已知的，在这里不再重复。在步骤 310 中，词义歧义消除模块 32 为该查询中每个单词的含义进行区分。

在优选实施例中，如步骤 312 所示通过使用知识库 400 (图 3B)，搜索引擎扩展并解释该已消除歧义的查询以包括与该查询中特定的关键词义语义相关的关键词义。在词义基础上执行该扩展并相应地产生与该查询的含义有关的一列词义。该语义关系可以是以上参照图 3A 和图 3B 所描述的那些。

然后在步骤 314 中，搜索引擎将已消除歧义并已扩展的查询与数据库中的词义信息进行比较。选择知识库中其词义与该查询中的关键词义匹配的条目作为结果。如上所述，该知识库包括已索引的文档的数据库。然后在步骤 316 中，搜索引擎将结果返回到客户机。在一个实施例中，可以根据该结果中所发现的词义与该查询中关键词的词义之间的语义关系对该结果进行加权。因此，例如，相比于包含具有一种下义关系的词义的结果而言，一个包含与该查询中的关键词义具有同义关系的词义的结果可以被赋予更高的加权值。还可以通过概率对该结果进行加权，此概率为已消除歧义的查询和/或已消除歧义的文档中的一个关键词义为正确的概率。还可以通过与该结果对应的诸如关联词义频率或彼此位置那样的文档或网页的其他特性来对该结果进行加权，或者通过其他为本领域技术人员所理解的用于排列结果的技术来对该结果进行加权。

应当认识到，可以在与客户机交互之前执行该处理的第一阶段作为预计算步骤。可以多次执行第二阶段而无需重复第一阶段。可以偶尔或定期执行第一阶段来保持数据库的流通 (currency)。还可以通过选择对该信息的子集合执行第一阶段来增量地更新该数据库。

尽管参照了一些特定实施例来描述本发明，但是本领域技术人员很清楚在不脱离在后附权利要求中所概括的本发明范围的情况下可以进行多种变化。本领域技术人员对下一个或更多的专业都具有充分的了解：计算机编程，机器学习和计算机语言学。

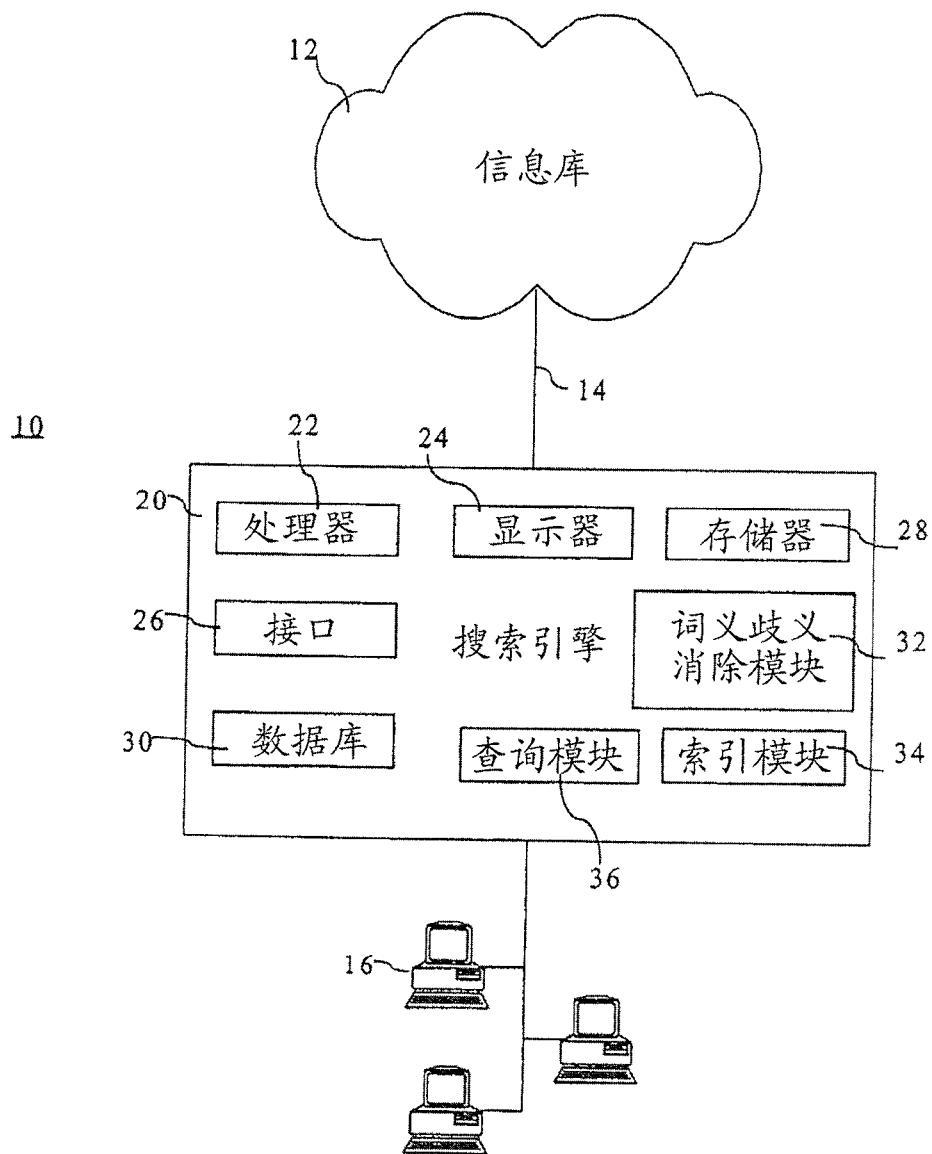


图1

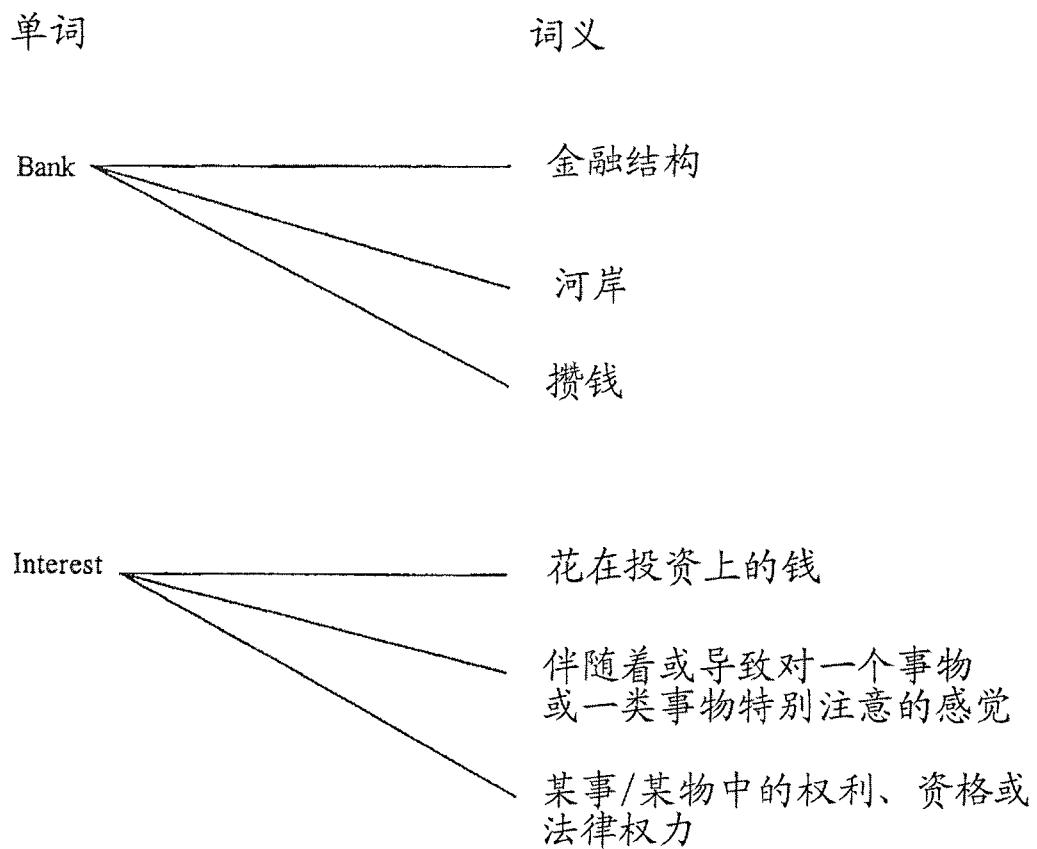


图2

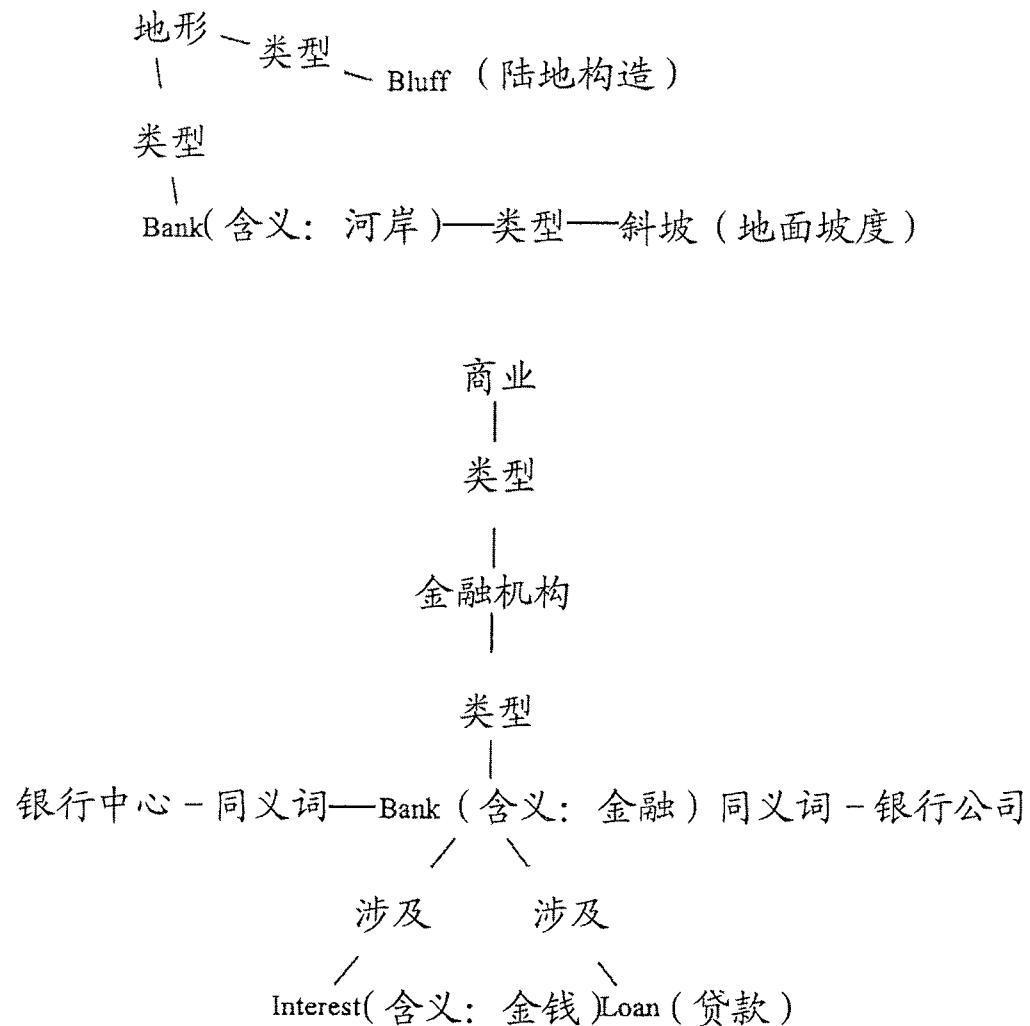
200

图 3A

知识库 400

节点表 402

ID 406	类型 408	注释 410
bank 406A	单词408A	
标签001 408B	精细含义 408B	名词， 金融机构 410B
标签002	精细含义	名词， 坡地或海岸线
标签003	精细含义	名词， 商业银行在其中进行交易的建筑
标签004	精细含义	动词， 与银行进行交易
标签005	粗略含义	
deposit	单词	
标签006	精细含义	名词， 一种设备， 可以为存储或安全保护而将物品存放在该设备中

边缘/关系表 404

源节点ID 412	目的节点ID 414	类型 416	注释 418
bank	标签001	单词到精细含义	1
bank	标签002	单词到精细含义	2
bank	标签003	单词到精细含义	4
bank	标签004	单词到精细含义	3
bank	标签005	单词到粗略含义	
标签005	标签001	粗略含义到精细含义	
标签005	标签003	粗略含义到精细含义	
标签005	标签004	粗略含义到精细含义	
deposit	标签006	单词到精细含义	8
标签006	标签003	下义词	

图 3B

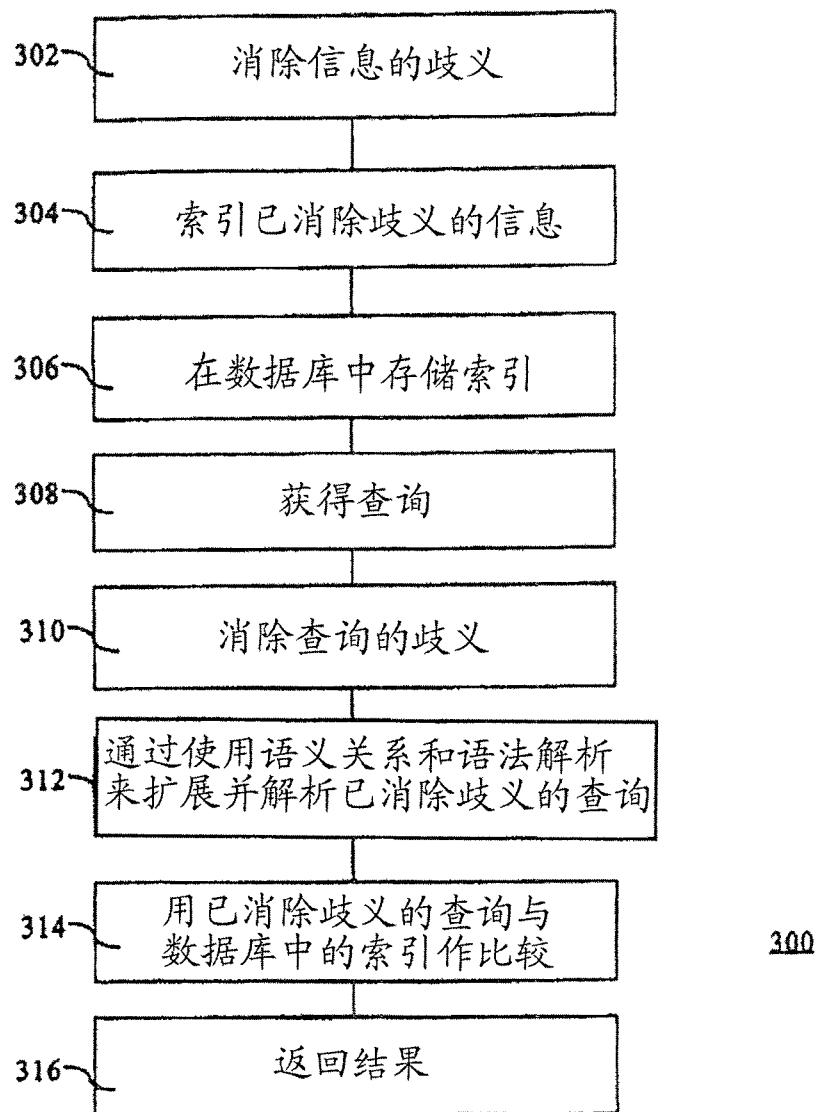


图 4