(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国 际 局

(43) 国际公布日
2015 年 4 月 30 日 (30.04.2015)

WIPO | PCT

(10) 国际公布号
WO 2015/058604 A1

(51) 国际专利分类号:
G06F 15/18 (2006.01)

(21) 国际申请号: PCT/CN2014/086838

(22) 国际申请日: 2014 年 9 月 18 日 (18.09.2014)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201310495856.6 2013 年 10 月 21 日 (21.10.2013) CN
201310495881.4 2013 年 10 月 21 日 (21.10.2013) CN
201310495641.4 2013 年 10 月 21 日 (21.10.2013) CN

(71) 申请人: **北京奇虎科技有限公司 (BEIJING QIHOO TECHNOLOGY COMPANY LIMITED)** [CN/CN]; 中国北京市西城区新街口外大街 28 号 D 座 112 室（德胜园区）, Beijing 100088 (CN)。 **奇智软件（北京）有限公司 (QIZHI SOFTWARE (BEIJING) COMPANY LIMITED)** [CN/CN]; 中国北京市朝阳区酒仙桥路 6 号院 2 号楼 B 座 2 层、3 层 301-306 室, Beijing 100015 (CN)。

(72) 发明人: **孙林 (SUN, Lin)**; 中国北京市朝阳区酒仙桥路 6 号院 2 号楼, Beijing 100015 (CN)。 **陈培军 (CHEN, Peijun)**; 中国北京市朝阳区酒仙桥路 6 号院 2 号楼, Beijing 100015 (CN)。 **秦吉胜 (QIN, Jisheng)**; 中国北京市朝阳区酒仙桥路 6 号院 2 号楼, Beijing 100015 (CN)。

(74) 代理人: **北京市隆安律师事务所 (BEIJING LONGAN LAW FIRM)**; 中国北京市朝阳区建国门外大街 21 号北京国际俱乐部 188 室, Beijing 100020 (CN)。

(81) 指定国 (除另有指明，要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

(84) 指定国 (除另有指明，要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

[见续页]

(54) **Title**: APPARATUS AND METHOD FOR OBTAINING DEGREE OF ASSOCIATION OF QUESTION AND ANSWER PAIR AND FOR SEARCH RANKING OPTIMIZATION

(54) **发明名称**: 获取问答对相关联程度、优化搜索排名的装置和方法

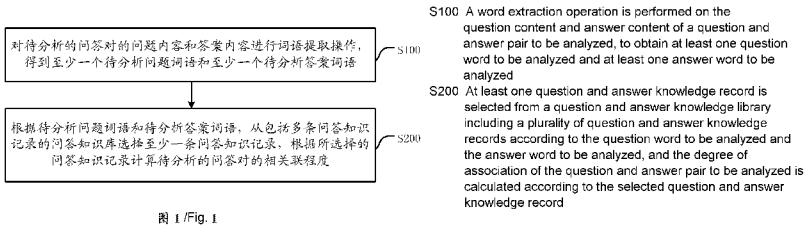


图 1 /Fig. 1

(57) **Abstract**: An apparatus and method for obtaining the degree of association of a question and answer pair, a method for the search ranking optimization of the question and answer pair, and an apparatus and method for determining the crawling frequency of a network resource point. The method for obtaining the degree of association of the question and answer pair comprises the following steps: performing a word extraction operation on the question content and answer content of a question and answer pair to be analyzed, to obtain at least one question word to be analyzed and at least one answer word to be analyzed; selecting at least one question and answer knowledge record from a question and answer knowledge library including a plurality of question and answer knowledge records according to the question word to be analyzed and the answer word to be analyzed, and calculating the degree of association of the question and answer pair to be analyzed according to the selected question and answer knowledge record. With the apparatus and method for obtaining the degree of association of the question and answer pair, the quality of the question and answer pair can be evaluated semantically, and the evaluation effect is better; in addition, the apparatus and method are easy to implement and excellent in universality.

(57) **摘要**:

[见续页]

WO 2015/058604 A1

本国际公布:

— 包括国际检索报告(条约第 21 条(3))。

一种获取问答对的相关联程度的装置和方法，一种优化问答对的搜索排名的装置和方法，以及一种确定网络资源点的抓取频率的装置和方法，其中，获取问答对的相关联程度的方法包括如下步骤：对待分析的问答对的问题内容和答案内容进行词语提取操作，得到至少一个待分析问题词语和至少一个待分析答案词语；根据待分析问题词语和待分析答案词语，从包括多条问答知识记录的问答知识库选择至少一条问答知识记录，根据所选择的问答知识记录计算待分析的问答对的相关联程度。该获取问答对的相关联程度的装置和方法，可以从语义方面评价问答对的质量，评价效果更好，而且容易实现、通用性强。

获取问答对相关联程度、优化搜索排名的装置和方法

技术领域

本发明涉及网络数据通信技术领域，具体涉及一种获取问答对的相关联程度的装置和方法，一种优化问答对的搜索排名的装置和方法，以及一种确定网络资源点的抓取频率的装置和方法。

背景技术

问答社区是一种用户产生内容的网络应用，基本形式是由用户根据自己的需求提出问题，并由其他的用户来给出回答。这种形式为用户在网络上获取信息提供了新的渠道。然而由于任何用户都可以随意地创建内容，导致了问答社区中的信息质量差异非常大，以至于问答社区中出现了大量的低质量问答对。这不但给用户查找信息带来了诸多不便，同时也降低了问答社区的质量。同时，现有技术的判断问答对质量的方法，更多地依赖于问答对的非文本特征来评价问答对质量，会影响其通用性。

另外，使用现有的搜索技术进行问答搜索时，获取的搜索结果中存在部分低质量的问答对而现有技术的对搜索结果进行排序的方法，更多地依赖于问答对所属的网站和问答对的非文本特征来对问答对进行排序，会影响搜索结果的精确性和通用性。

同时地，使用现有的搜索技术进行问答搜索时，难以判断问答社区作为网络资源点的质量而现有技术（例如，爬虫蜘蛛）的对网络资源点设置抓取频率方法，更多地依赖于问答对网站的链接的分析，这样的方法用于问答搜索，不能从语义上分析问答对也不能根据网络资源点的质量调整抓取频率（或，爬取细度、爬取频率），会影响搜索结果的精确性和通用性。

发明内容

鉴于上述问题，提出了本发明以便提供一种克服上述问题或者至少部分地解决上述问题的一种获取问答对的相关联程度的装置和方法，一种优化问答对的搜索排名的装置和方法，以及一种确定网络资源点的抓取频率的装置和方法。

依据本发明的一个方面，提供了一种获取问答对的相关联程度的装置，该装置包括：问答知识库，适于存储多条问答知识记录；词语提取单元，适于对待分析的问答对的问题内容和答案内容进行词语提取操作，得到至少一个待分析问题词语和至少一个待分析答案词语；相关联程度计算单元，适于根据待分析问题词语和待分析答案词语，从问答知识库选择至少一条问答知识记录，根据所选择的问答知识记录计算待分析的问答对的相关联程度。

根据本发明的另一方面，提供了一种优化问答对的搜索排名的装置，该装置包括：问答知识库，适于存储多条问答知识记录；搜索单元，适于接收使用者的搜索请求，根据使用者的搜索请求，获取与搜索请求匹配的多个待分析问答对；计算单元，适于根据问答知识库获取每个待分析问答对的相关联程度；搜索排名单元，适于根据所述待分析问答对的相关联程度优化所述待分析问答对的搜索排名。

根据本发明的又一方面，提供了一种确定网络资源点的抓取频率的装置，该装置包括：问答知识库，适于存储多条问答知识记录；资源分析单元，适于由网络资源点抓取多个待分析问答对；计算单元，适于根据问答知识库获取每个待分析问答对的相关联程度；抓取频率确定单元，根据所述待分析问答对的相关联程度确定所述网络资源点的抓取频率。

根据本发明的另一方面，提供了一种获取问答对的相关联程度的方法，该方法包括如下步骤：对待分析的问答对的问题内容和答案内容进行词语提取操作，得到至少一个待分析问题词语和至少一个待分析答案词语；根据待分析问题词语和待分析答案词语，从包括多条问答知识记录的问答知识库选择至少一条问答知识记录，根据所选择的问答知识记录计算待分析的问答对的相关联程度。

根据本发明的又一方面，提供了一种优化问答对的搜索排名的方法，该方法包括如下步骤：接收使用者的搜索请求，根据使用者的搜索请求，获取与搜索请求匹配的多个待分析问答对；根据包括多条问答知识记录的问答知识库，获取每个待分析问答对的相关联程度；根据所述待分析问答对的相关联程度优化所述待分析问答对的搜索排名。

根据本发明的再一方面，提供了一种确定网络资源点的抓取频率的方法，该方法包括如下步骤：由网络资源点抓取多个待分析问答对；根据包括多条问答知识记录的问答知识库，获取每个待分析问答对的相关联程度；根据所述待分析问答对的相关联程度确定所述网络资源点的抓取频率。

根据本发明的技术方案，从含有问答对的网页提取出多个问答对并根据提取的问答对构建包括多条

问答知识记录的问答知识库，对待分析的问答对的问题内容和答案内容进行词语提取操作而得到至少一个待分析问题词语和至少一个待分析答案词语，进而根据待分析问题词语和待分析答案词语从问答知识库选择至少一条问答知识记录并根据所选择的问答知识记录计算待分析的问答对的相关联程度，可以从语义方面评价问答对的质量，解决了现有技术仅在词法层面上评价问答对的质量而导致的评价效果不佳的问题，同时，在根据使用者的搜索请求获取的与搜索请求匹配的多个待分析问答对的情况下，根据问答知识库获取每个待分析问答对的相关联程度并根据待分析问答对的相关联程度优化待分析问答对的搜索排名，可以从语义方面评价待分析问答对的质量，解决了现有技术依赖于问答对所属的网页和问答对的非文本特征来对问答对进行排序而导致的排序效果不佳的问题；进一步地，借助由网络资源点抓取多个待分析问答对，根据问答知识库获取每个待分析问答对的相关联程度并根据待分析问答对的相关联程度确定所述网络资源点的抓取频率，可以通过评价网络资源点的质量确定抓取频率，解决了现有技术不能根据网络资源点的质量调整抓取频率而导致的搜索效果不佳的问题。而且本申请的方案容易实现、通用性强。

上述说明仅是本发明技术方案的概述，为了能够更清楚了解本发明的技术手段，而可依照说明书的内容予以实施，并且为了让本发明的上述和其它目的、特征和优点能够更明显易懂，以下特举本发明的具体实施方式。

附图说明

通过阅读下文优选实施方式的详细描述，各种其他的优点和益处对于本领域普通技术人员将变得清楚明了。附图仅用于示出优选实施方式的目的，而并不认为是对本发明的限制。而且在整个附图中，用相同的参考符号表示相同的部件。在附图中：

图 1 示出了根据本发明一个实施例的获取问答对的相关联程度的方法的流程图；

图 2 示出了构建问答知识库的详细的流程图；

图 3 示出了使用如图 2 所示的步骤而得到的问答知识库的一个解释模型示意图；

图 4 示出了图 1 中步骤 S200 的详细的流程图；以及

图 5 示出了根据本发明一个实施例的获取问答对的相关联程度的装置的框图；

图 6 示出了根据本发明一个实施例的优化问答对的搜索排名的方法的流程图；

图 7 示出了根据本发明一个实施例的优化问答对的搜索排名的装置的框图；

图 8 示出了根据本发明一个实施例的确定网络资源点的抓取频率的方法的流程图；

图 9 示出了根据本发明一个实施例的确定网络资源点的抓取频率的装置的框图；

图 10 示出了用于执行根据本发明的方法的应用服务器的框图；以及

图 11 示出了用于保持或者携带实现根据本发明的方法的程序代码的存储单元。

附图实施例

现有的获取问答对的相关联程度的方法，是使用文本特征和非文本特征来描述问答对的问题和答案。类似地，现有的获取问答对的搜索排名的方法，是使用文本特征和非文本特征来描述问答对的问题和答案从而对问答对进行排名，或根据问答对所属的网站的排名对问答对进行排名。文本特征主要包括文本视觉特征（例如标点符号密度，平均词长，文本熵等）和文本内容特征（例如文本内容词比例，疑问词密度，相关词覆盖等），并提取中文自动差错广泛采用的特征（例如单字密度特征等）；非文本特征包含用户的权威度指标，答案问题状态，答案回答时间，用户关系交互特征等。在对问题和答案分别提取出特征后，在训练集上分别学习出一个问题质量预测模型和答案质量预测模型，并使用两个模型的输出结果来评价问答对质量。然而，使用现有的获取问答对的相关联程度的方法对于答案质量进行评价时，仅仅使用了相关词覆盖特征来描述问题和答案间的语义匹配度，这不但仅仅是停留在词法层面上的，而且没有考虑问题和答案间的语义匹配度。然而问题和答案间的语义匹配度恰恰是问答对质量的核心，比如问题为“中国的首都是哪里？”，答案 1 为“北京”，答案 2 为“中国的首都是上海”。那么问题经过分词及丢弃停用词处理后，为“中国 首都 哪里”，答案 1 分词结果为“北京”，答案 2 分词结果为“中国 首都 上海”。现有技术中，语义匹配度可以定义为：问题和答案中共同出现的词语个数除以问题和答案中所有词语的个数。则问题和答案 1 的语义匹配度为：0/4=0。问题和答案 2 的语义匹配度为：2/4=0.5。使用现有技术，就会认为答案 2 和问题较为匹配。而我们知道这显然是不当的。

下面将参照附图更详细地描述本公开的示例性实施例。虽然附图中显示了本公开的示例性实施例，然而应当理解，可以以各种形式实现本公开而不应被这里阐述的实施例所限制。相反，提供这些实施例是为了能够更透彻地理解本公开，并且能够将本公开的范围完整的传达给本领域的技术人员。

图 1 示出了根据本发明一个实施例的获取问答对的相关联程度的方法的流程图。根据本发明的另一方面，提供了一种获取问答对的相关联程度的方法，该方法包括如下步骤 S100 和步骤 S200：

S100、对待分析的问答对的问题内容和答案内容进行词语提取操作，得到至少一个待分析问题词语和至少一个待分析答案词语。

在本发明的一个实施例中，对待分析的问答对的问题内容和答案内容进行词语提取操作具体包括：对待分析的问答对的问题内容和答案内容进行分词、去除停用词、词合并（word join），和提取实体词（例如名词、动词等）的操作。则由待分析的问答对的问题内容得到至少一个待分析问题词语，由待分析的问答对的答案内容得到至少一个待分析答案词语。

S200、根据待分析问题词语和待分析答案词语，从包括多条问答知识记录的问答知识库选择至少一条问答知识记录，根据所选择的问答知识记录计算待分析的问答对的相关联程度。

本实施例的步骤 S200，可以通过利用问答知识库从语义方面对待分析问答对的问题内容和答案内容进行分析以获得待分析问答对的相关联程度，评价效果更好而且容易实现。

进一步地，所述包括多条问答知识记录的问答知识库，是通过预先从含有问答对的网页提取出多个问答对，根据提取的问答对构建而得到的。在本发明的一个实施例中，在从含有问答对的网页提取出多个问答对时，抓取与所述问答对对应的类别。则在根据提取的问答对构建问答知识库时，根据问答对和与所述问答对对应的类别构建问答知识记录。得到的问答知识库之中的每个问答知识记录对应于一个类别，分别包括一个问题词语（QW）、一个答案词语（AW），以及所述问题词语和所述答案词语之间的语义相关度。

通过利用由网页提取的海量的、高质量的问答对构建包括多条问答知识记录的问答知识库，可以基于对海量信息的学习而获取多条问答知识记录的问题词语和答案词语之间的语义相关度；而且通过利用从网页提取得到的信息构建问答知识库，适用的范围更广，方法的通用性更强。

图 2 示出了构建问答知识库的详细的流程图。具体包括以下步骤 S310、步骤 S320 和步骤 S330:

S310、预先从含有问答对的网页提取出多个问答对，抓取与所述问答对对应的类别。

本实施例中，可以通过使用网络爬虫，从互联网上含有高质量问答对的网页抓取数据并提取出问答对，以保证所提取的问答对的质量；所述含有高质量问答对的网页包括 cQA（Customer Quality Assurance 客户品质保证）社区、各大专业论坛等，则可以使用楼层识别技术，根据楼主（即针对一个问题首个发出帖子的使用者）提问题，1 楼 2 楼（即依序回复帖子的使用者）等回复的内容为答案的方式，来提取问答对。由于所述含有高质量问答对的网页中包括对应于每个问答对的类别信息，所以可以在抓取问答对的同时一并抓取与所述问答对对应的类别。

S320、对每个问答对，对该问答对的问题内容和答案内容进行词语提取操作，得到问题词语集合和答案词语集合；令问题词语集合中的每个问题词语与答案词语集合中的每个答案词语分别在与该问答对对应的每个类别上形成一条信息记录。

在本发明的一个实施例中，对步骤 S310 中提取得到的所述问答对中的每一个问答对的问题内容和答案内容进行词语提取操作，具体包括，对问答对的问题内容和答案内容进行分词、去除停用词、词合并，和提取实体词的操作。

则由每一个问答对的问题内容得到至少一个问题词语，由每一个问答对的答案内容得到至少一个答案词语，则可以得到针对该问答对的类别集合<C_1，…，C_k，…，C_p>、问题词语集合<QW_1，…，QW_i，…，QW_m>和答案词语集合<AW_1，…，AW_j，…，AW_n>。

通过令问题词语集合中的每个问题词语（QW_i）与答案词语集合中的每个答案词语（AW_j）分别在与该问答对对应的每个类别（C_k）上形成一条信息记录，例如<QW_i，AW_j，C_k>，则可以形成 m*n*p 条信息记录。

S330、对每一条信息记录，执行以下操作：计算该答案词语属于该类别的概率，计算在该类别上该答案词语对该问题词语的解释的专一程度，计算在该类别上该问题词语用该答案词语进行解释的强度；将上述概率、专一程度和强度相乘，所得到的乘积是该答案词语和该问题词语的语义相关度；令该问题词语、该答案词语和其语义相关度形成一条对应于该类别的问答知识记录<QW_i，AW_j，weight（QW_i，AW_j）>或<QW_i，AW_j，C_k，weight（QW_i，AW_j）>。本实施例中的步骤 S330，可以是在对从网页抓取的海量的问答对进行了如步骤 S320 所述的词语提取操作而得到海量的信息记录之后基于所述海量的信息记录进行的，则基于海量的信息记录而获取的语义相关度更准确。

较佳地，所述计算该答案词语属于该类别的概率，具体包括：

$$P(Ck|AWj)=P(AWj|Ck)*\frac{P(Ck)}{P(AWj)};$$

所述计算在该类别上各个答案词语对该问题词语的解释的专一程度，具体包括：

$$specific(QWi,AWj|C=Ck)=P(QWi|AWj,C=Ck)=\frac{\#(QWi,AWj)}{\#(AWj)}|C=Ck;$$

所述计算在该类别上该问题词语用各个答案词语进行解释的强度，具体包括：

$$interpret(QWi, AWj|C = Ck) = P(AWj|QWi, C = Ck) = \frac{\#(QWi, AWj)}{\sum_{j=1}^{x} \#(QWi, AWj)}|C = Ck;$$

将上述概率、专一程度和强度相乘，具体包括：

$$weight(QWi, AWj|C = Ck) = P(Ck|AWj) * specific(QWi, AWj|C = Ck) * interpret(QWi, AWj|C = Ck);$$

其中，P（C_k）表示类别 C_k 出现的概率；P（AW_j）表示答案为 AW_j 的概率；P（AW_j | C_k）表示 C_k 类别属于 AW_j 的概率；

#（QW_i，AW_j）表示问题词语为 QW_i 且答案词语为 AW_j 的次数；

#（AW_j）表示答案词语为 AW_j 的次数。

由步骤 S310、步骤 S320 和步骤 S330，可以得到问答知识记录而构建问答知识库。图 3 示出了使用如图 2 所示的步骤而得到的问答知识库的一个解释模型示意图。可知，对于每一问题词语 QW_i，可以针对类别集合<C_1，…，C_k，…，C_p>中的每一类别，获得 n 条问答知识记录。当然，本领域技术人员可以了解的是，若计算得到的语义相关度为 0，则可以删除相应的问答知识记录；再者，如果问答知识库中问答知识记录的数量过大而使得存储问答知识记录和计算待分析问答对的相关联程度的开销过大，可以预设一个阈值，将语义相关度小于阈值的问答知识记录删除以减小开销。

图 4 示出了图 1 中步骤 S200 的详细的流程图。在通过步骤 S100 得到至少一个待分析问题词语和至少一个待分析答案词语后，步骤 S200 具体包括以下步骤 S210、步骤 S220 和步骤 S230：

S210、选取其包括的问题词语与待分析问题词语匹配且包括的答案词语与待分析答案词语匹配的问答知识记录。本实施例中，问题词语与待分析问题词语匹配是指待分析问题词语与问题词语相同或待分析问题词语是问题词语的子串；答案词语与待分析答案词语匹配是指待分析答案词语与答案词语相同或待分析答案词语是答案词语的子串，本实施例通过步骤 S210，使用字段匹配或字段搜索的方法，从问答知识库中选出部分与待分析的问答对相关的问答知识记录。

S220、根据所述选取的问答知识记录中对应于相同类别的问答知识记录，得到该待分析的问答对分别针对各个类别的相关联程度，具体包括：将选取的问答知识记录中对应于相同类别的问答知识记录的语义相关度加权相加，得到该待分析的问答对分别针对各个类别的相关联程度。

本实施例，将通过步骤 S210 选出的问答知识记录根据其所对应的类别进行分组，对应于相同类别的问答知识记录为一组；将每一组的问答知识记录的语义相关度加权（例如，权值为 1 或 100）相加，得到该待分析的问答对针对该类别的相关联程度；由此得到至少一个（本实施例中的相关联程度的数目即待分析问答对对应的类别的数目）相关联程度。

S230、选取上述该待分析的问答对针对各个类别的相关联程度的最大值，以该最大值作为待分析的问答对的相关联程度。

图 5 示出了根据本发明一个实施例的获取问答对的相关联程度的装置的框图。该装置包括问答知识库 100、词语提取单元 200 和相关联程度计算单元 300。

问答知识库 100，适于存储多条问答知识记录；本实施例的问答知识库 100 能够通过抓取网页中的海量问答对构建得到。

词语提取单元 200，适于对待分析的问答对的问题内容和答案内容进行词语提取操作，得到至少一个待分析问题词语和至少一个待分析答案词语。

在本发明的一个实施例中，词语提取单元 200，适于对待分析的问答对的问题内容和答案内容进行分词、去除停用词、词合并（word join），和提取实体词（例如名词、动词等）的操作，以得到至少一个待分析问题词语和至少一个待分析答案词语。

相关联程度计算单元 300，适于根据待分析问题词语和待分析答案词语，从问答知识库选择至少一条问答知识记录，根据所选择的问答知识记录计算待分析的问答对的相关联程度。

在本发明的一个实施例中，相关联程度计算单元 300，适于选取其包括的问题词语与待分析问题词语匹配且包括的答案词语与待分析答案词语匹配的问答知识记录。本实施例中，问题词语与待分析问题词语匹配是指待分析问题词语与问题词语相同或待分析问题词语是问题词语的子串；答案词语与待分析答案词语匹配是指待分析答案词语与答案词语相同或待分析答案词语是答案词语的子串；根据所述选取的问答知识记录中对应于相同类别的问答知识记录，得到该待分析的问答对针对各个类别的相关联程度，更具体地，是将选取的问答知识记录中对应于相同类别的问答知识记录的语义相关度加权（例如，权值为 1 或 100）相加而得到该待分析的问答对分别针对各个类别的相关联程度，由此得到至少一个（本实施例中的相关联程度的数目即待分析问答对对应的类别的数目）相关联程度；选取上述该待分析的问答对针对各个类别的

相关联程度的最大值，以该最大值作为待分析的问答对的相关联程度。

利用问答知识库 100、词语提取单元 200 和相关联程度计算单元 300，通过利用待分析问题词语和待分析答案词语，从问答知识库选择至少一条问答知识记录，以及根据所选择的问答知识记录计算待分析的问答对的相关联程度，可以从语义方面对待分析问答对进行分析，评价效果更好而且容易实现，通过利用从网页提取得到的信息构建问答知识库，适用的范围更广，通用性更强。

在本实施例中，该装置还包括问答知识库构建单元 400，问答知识库构建单元 400 适于预先从含有问答对的网页提取出多个问答对，根据提取的问答对构建包括多条问答知识记录的问答知识库。在图 5 所示的装置中，问答知识库是已有的，由于实际网络的信息量不断增加，信息内容的变化速度快，问答知识库的内容往往需要更新，通过增设问答知识库构建单元 400 构建（或者说更新）问答知识库，可以保证问答知识库的内容的即时性和可靠性。

较佳地，在从含有问答对的网页提取出多个问答对时，问答知识库构建单元 400 抓取与所述问答对对应的类别。本实施例中，可以通过使用网络爬虫，从互联网上含有高质量问答对的网页抓取数据并提取出问答对，以保证所提取的问答对的质量；所述含有高质量问答对的网页包括 cQA 社区、各大专业论坛等。由于所述含有高质量问答对的网页中包括对应于每个问答对的类别信息，所以问答知识库构建单元 400 可以在抓取问答对的同时一并抓取与所述问答对对应的类别。

在本实施例中，问答知识库构建单元 400，适于对每个问答对执行以下操作：对该问答对的问题内容和答案内容进行词语提取操作，得到问题词语集合和答案词语集合，具体地，问答知识库构建单元 400 对提取得到的所述问答对中的每一个问答对的问题内容和答案内容进行分词、去除停用词、词合并，和提取实体词的操作而得到问题词语和答案词语；令问题词语集合中的每个问题词语与答案词语集合中的每个答案词语分别在与该问答对对应的每个类别上形成一条信息记录。问答知识库构建单元 400，适于对每一条信息记录，执行以下操作：计算该答案词语属于该类别的概率，计算在该类别上该答案词语对该问题词语的解释的专一程度，计算在该类别上该问题词语用该答案词语进行解释的强度；将上述概率、专一程度和强度相乘，所得到的乘积是该答案词语和该问题词语的语义相关度；令该问题词语、该答案词语和其语义相关度形成一条对应于该类别的问答知识记录。

更具体地，问答知识库构建单元 400，适于按照如下的方法计算该答案词语属于该类别的概率：

$$P(Ck|AWj)=P(AWj|Ck)*\frac{P(Ck)}{P(AWj)};$$

更具体地，问答知识库构建单元 400，适于按照如下的方法计算在该类别上各个答案词语对该问题词语的解释的专一程度：

$$specific(QWi,AWj|C=Ck)=P(QWi|AWj,C=Ck)=\frac{\#(QWi,AWj)}{\#(AWj)}|C=Ck;$$

更具体地，问答知识库构建单元 400，适于按照如下的方法计算在该类别上该问题词语用各个答案词语进行解释的强度：

$$interpret(QWi,AWj|C=Ck)=P(AWj|QWi,C=Ck)=\frac{\#(QWi,AWj)}{\sum_{j=1}^{x}\#(QWi,AWj)}|C=Ck;$$

更具体地，问答知识库构建单元 400，适于按照如下的方法将上述概率、专一程度和强度相乘：

$$weight(QWi,AWj|C=Ck)=P(Ck|AWj)*specific(QWi,AWj|C=Ck)*interpret(QWi,AWj|C=Ck);$$

其中，P（C_k）表示类别 C_k 出现的概率；P（AW_j）表示答案为 AW_j 的概率；P（AW_j | C_k）表示 C_k 类别属于 AW_j 的概率；

#（QW_i，AW_j）表示问题词语为 QW_i 且答案词语为 AW_j 的次数；

#（AW_j）表示答案词语为 AW_j 的次数。

以下通过一个例子说明使用本发明的实施例所能达到的效果，比如有如下问答对，类别为“医疗健康”：

问题	孩子咳嗽流鼻涕怎么办
答案	这是小儿上感的症状,可先给予药物治疗： 1、抗病毒：小儿感冒颗粒，按说明剂量,日服 3 次； 2、止咳：小儿咳喘灵（中药冲剂）,按说明剂量,日服 3 次； 必要时加用抗生素：阿莫西林颗粒（再林）

	口服。或用安克 口服。也可用太儿欣（罗红霉素）口服。如药物口服疗效不佳时，应及时检查，或抗生素静滴治疗。

通过分词技术处理，得到待分析问题词语和待分析答案词语如下：

待分析问题词语	孩子，咳嗽，鼻涕
待分析答案词语	症状，药物,治疗，抗病毒，小儿感冒颗粒，说明，剂量，止咳，中药，冲剂，抗生素，阿莫西林，阿莫西林颗粒，颗粒,口服，罗红霉素，疗效

从分词结果可以看出，问题和答案中没有相关词覆盖，因此如果使用现有技术则容易认为该问答对相关联程度低，质量不高。但是实际上使用人工判断明显可知该问答对是一个高质量的问答对。

若使用本发明的方法和装置处理上述问答对，首先，可以调取已有的问答知识库，或者通过抓取 cQA 社区、各大专业论坛的问答对，构建问答知识库；

第二步，对上述待分析的问答对，经过词语提取操作得到待分析问题词语集合<孩子，咳嗽，鼻涕>、待分析答案词语集合<症状，药物,治疗，抗病毒，小儿感冒颗粒，说明，剂量，止咳，中药，冲剂，抗生素，阿莫西林，阿莫西林颗粒，颗粒,口服，罗红霉素，疗效>，并且得到待分析的问答对的类别为“医疗健康”；

第三步，根据各个待分析问题词语以及该类别，从问答知识库中选择得到问题词语与待分析问题词语匹配的若干问答知识记录，从而得到如下答案词语及语义相关度（为了方便阅读，下表中的语义相关度的数值是进行了适当的归一化处理后的数值）：

问题词语	类别	答案词语及语义相关度
孩子	医疗健康	快克:32，治疗:11，感冒药:7，检查:7,月经:5,病情:5,妊娠:4,患儿:4,面诊患者:3，排卵:3，发烧:3，剖腹产:3，药流:2，精子:2，抽动症:2，输卵管:2，家长:2，保胎:2，发生:1，休息:1，了解:1，恢复:1，锌硒宝:1，癫痫:1，对症:1 病因:1，筛查:1，生命阳光:1，上海康尧:1，儿科:1，复查:1，哺乳:1，调理:1 抵抗力:1，食物:1，人工流产:1，脑瘫:1，呕吐:1，叶酸:1，康尧:1，苯丙醇:1，神经:1
咳嗽	医疗健康	支气管炎:40，婴儿久:19，阴功效:18，咽炎:18，剑花猪肺汤:18，塘婴儿:18，慢性气管炎支气管哮喘:18，丝瓜粥:17，痰热:15，猪肺:15，南杏:15，款冬花:13，果猪肺汤:12，杏仁热水提取物粗制剂:12，润肺:12，噬菌体作用:11，白萝卜性味辛甘:11，肺凉心:11，苡仁米:11，小鼠肉瘤:11，碎小块:11，橄榄萝卜:11，稀稠粥:10，银耳锅:10，气管炎:10，玉竹汁液:10，快克:10，汤肴:9，脾虚湿盛:8，马兜铃:8 瓷杯:8，治疗:8，咯血:7，白萝卜:7，鼻涕:7，肺片:7，急支糖浆:7，蛇胆川贝:6，宣肺:6，普通人流感:6，肺血管破裂:6，抑制率:6，气道:5，热感冒:5，白猪:5，炎症:5，温热服:5，流感病毒:5，绿脓杆菌:5，阴虚咳喘:4，恶阴:4，猪肉杏仁汤治:4，防尘口罩:4，豆腐糖:4，康尧:4，熟羊脂:4，喉癌:3，山药杏仁粥:3，痰量:3，内加白糖:3，痉挛性:3，弱虚:3，黄精冰糖:3，黄河利亚:3，细支气管炎:3，麻醉性强效:3，抵抗力:3 盖盘:3，白痰:3，散酒毒:3，支气管粘膜:3，乐信药业:3，结核:3，空腹开水:3，生梨:3，右美沙芬:3，化痰药:3，猪油蜜膏:3，荞面蛋清:3，开水碗:3，组织胺:2，机体损耗:2，对症:2，患儿:2，患者:2，微寒:2，检查:2，鲜良:2，反射性:2，理脾:2，养阴清肺丸:2，吸入:2，含片:2，排痰:2，宰杀:2，粳米:2，大蒜水:2，蜂蜜:2，血痰:2，甲型:2，喉咙痒:2，呼吸衰竭:2，上海康尧:2，沙氏鹿茸草:2，枇杷:2，中枢性因素:2，马乳:2，银耳:2，保护性:2，博列康尼:2，病毒性感冒:2，瑜伽洗鼻壶:2，鼻咽处:2，胸部胀满:2，肺阴虚:2，抗衰:2，服法:2，老年人正气:2，交感神经药:1，海石粉:1，白开水:1，糖水:1，蛇胆陈皮:1，麦冬:1，肾咳喘:1，瓜蒌:1，子宫颈癌:1，虚汗多:1，合剂:1，低热:1，CT:1，例男性:1，发汗:1，阴润肺:1，大汤碗:1，口罩:1，白前:1，食物:1，配伍:1，本方清肝火:1，肺泡壁弹性组织:1，痰涎:1，上火:1，浮沫:1，无花果:1，保护性作用:1，大蒜:1，旋复花:1，解毒片:1，多发性炎症:1，呼吸道疾病:1，萝卜水:1，肺汤加:1，大肠杆菌:1，气管黏膜:1，儿科:1，绿色部份:1，支原体感染:1，肝阴亏虚:1，过敏:1，止咳露:1，患者食道 24 小时 PH:1，气痛:1，鼠曲草:1，咽部黏膜:1，组织表面鼻涕:1，中医顾铭浩验方:1，休息:1，标本虚实:1，鼻窦炎:1，调味:1，传染源:1，病因:1，柚肉:1，生姜红糖水:1，神经末梢感受器:1，晾温:1，调理:1，气管异物:1，桑菊饮:1，解

		毒:1，呕吐:1，贝母瓜蒌散:1，淡豆豉:1，鲜竹沥口服液:1，小青龙:1，白稀痰:1，本方益肾:1，病菌:1，鲜藕:1，咸易:1，胸膜疾病:1，气虚:1，散寒:1，头孢类:1 耳鼻科医生:1，果菜汁:1，纱布包:1，生姜切片:1，鸡屎藤:1，主治:1，补肾:1，鼻粘膜表面:1
鼻涕	医疗健康	鼻涕:108，鼻窦炎:29，脑脊梁:19，快克:17，热感冒:15，感冒药:15，鼻漏:15，鼻窦:14，鼻腔粘膜:12，洗鼻器:6，鼻腔膜:5，瑜伽洗鼻壶:5，鼻咽处:5，解毒片:4，药物性鼻炎:4 病毒性感冒:4，治疗:4，鼻中隔:3，咽喉:3，脓鼻涕:3，组织表面鼻涕:3，感冒颗粒:3，粘液性:3，冷空气:3，口罩:2，多为上颌窦:2，普通人流感:2，急性鼻窦炎:2，咽炎:2，抵抗力:2，鼻腔分泌物:2，多可少:2 代茶频:2，呋麻滴鼻液:2，季节性:2，清毒作用:2，维 C 银翘片:2，鼻窦部位:2，低温等离子:2，鼻腔间歇性:2，类型鼻炎:2，鼻粘膜表面:2，小儿感冒颗粒:1，对症:1，寿司芥末酱:1，新康泰克:1，痒感:1，氨酚黄那敏颗粒:1，姐姐症状差:1，RHINITIS:1，水样:1，吸入:1，开水:1，脱敏:1，鼻腔粘液腺:1，热伤风:1，千柏:1，上火:1，息斯敏:1，鼻腔黏膜:1，病毒大本营:1，流感症状:1，鼻腔分泌物涂片:1，支气管炎:1，清热:1，净鼻腔:1，生理因素:1，口腔病菌:1，刺激物:1，考特:1，发作:1，甲型:1，儿童感冒药效果:1，多鼻涕:1，季节性鼻炎:1，流脓:1，全身病因:1，通窍:1，着凉:1，宝宝:1，病菌:1，快克感冒药:1，老干丝瓜:1，生姜:1，苏理肺:1，臭鼻症:1，阵发性喷嚏:1

第四步，根据待分析答案词语集合中的待分析答案词语，在第三步所选择得到的问答知识记录的基础上筛选出其包括的答案词语与待分析答案词语匹配的问答知识记录，进而得到所筛选出的问答知识记录的语义相关度。经分析可知，本例中与问答知识记录中的答案词语匹配的待分析答案词语包括：<口服，咳喘，小儿感冒颗粒，检查，止咳，治疗，流感症状，感冒颗粒>。

再计算上述待分析的问答对的相关联程度可以得出，该待分析的问答对的相关联程度达到了 0.9（在相关联程度取值范围为 0～1 的条件下）。

图 6 示出了根据本发明一个实施例的优化问答对的搜索排名的方法的流程图。该方法包括如下步骤 S610、步骤 S620 和步骤 S630:

S610、接收使用者的搜索请求，根据使用者的搜索请求，获取与搜索请求匹配的多个待分析问答对。

在本发明的一个实施例中，可以是使用网络搜索技术，例如使用问答对搜索引擎，根据使用者的搜索请求获取待分析问答对。

S620、根据包括多条问答知识记录的问答知识库，获取每个待分析问答对的相关联程度。

本实施例的步骤 S620，可以通过利用问答知识库从语义方面对待分析问答对的问题内容和答案内容进行分析以获得待分析问答对的相关联程度，评价效果更好而且容易实现。

更具体地，本实施例的步骤 S620 的获得待分析问答对的相关联程度的具体实施方式，与如图 1、4 所示的获取问答对的相关联程度的方法大致相同，此处不再重复。

进一步地，所述包括多条问答知识记录的问答知识库，是通过预先从含有问答对的网页提取出多个问答对，根据提取的问答对构建而得到的。在本发明的一个实施例中，在从含有问答对的网页提取出多个问答对时，抓取与所述问答对对应的类别。则在根据提取的问答对构建问答知识库时，根据问答对和与所述问答对对应的类别构建问答知识记录。得到的问答知识库之中的每个问答知识记录对应于一个类别，分别包括一个问题词语（QW）、一个答案词语（AW），以及所述问题词语和所述答案词语之间的语义相关度。通过利用由网页提取的海量的、高质量的问答对构建包括多条问答知识记录的问答知识库，可以基于对海量信息的学习而获取多条问答知识记录的问题词语和答案词语之间的语义相关度；通过利用从网页提取得到的信息构建问答知识库，适用的范围更广，方法的通用性更强。

更具体地，本实施例的方法还包括构建问答知识库的步骤，构建问答知识库的流程与图 2 所示的流程大致相同；本实施例的问答知识库的解释模型与如图 3 所示的解释模型大致相同。此处不再重复。

S630、根据所述待分析问答对的相关联程度优化所述待分析问答对的搜索排名。

由于待分析问答对的相关联程度反映了质量，所以可以利用相关联程度优化所述待分析问答对的搜索排名，排名效果更好。

具体的方法，可以是以所述待分析问答对的相关联程度的次序作为所述待分析问答对的搜索排名，即相关联程度高的问答对的搜索排名靠前；也可以是先根据搜索排列技术初步排列所述待分析问答对所属的网站，根据该初步排列的次序号与所述待分析问答对的相关联程度计算所述待分析问答对的搜索排名，例如，可以将所述待分析问答对所属的网站的初步排列的次序号与所述待分析问答对的相关联程度相乘，以相乘运算的结果的次序作为所述待分析问答对的搜索排名；通过将待分析问答对的质量和其所属网站的排

名结合，以对待分析问答对进行排序，使用者使用问答对搜索时，能够获得更好的结果排序的质量。

图 7 示出了根据本发明一个实施例的优化问答对的搜索排名的装置的框图。该装置包括问答知识库 710、搜索单元 720、计算单元 730 和搜索排名单元 740。

问答知识库 710，适于存储多条问答知识记录。本实施例的问答知识库 710 能够通过抓取网页中的海量问答对构建得到。

搜索单元 720，适于接收使用者的搜索请求，根据使用者的搜索请求，获取与搜索请求匹配的多个待分析问答对。

在本发明的一个实施例中，搜索单元 720 可以是问答对搜索引擎，根据使用者的搜索请求获取待分析问答对；例如搜索单元 720 是用于问答对搜索的网络搜索引擎，接收使用者通过浏览器输入的搜索请求并获取待分析问答对。

计算单元 730，适于根据问答知识库 710 获取每个待分析问答对的相关联程度。

本发明的计算单元 730 可以通过利用问答知识库从语义方面对待分析问答对的问题内容和答案内容进行分析以获得待分析问答对的相关联程度，评价效果更好而且容易实现。问答知识库 710 利用由网页提取的海量的、高质量的问答对构建并且包括多条问答知识记录，可以基于对海量信息的学习而获取多条问答知识记录的问题词语和答案词语之间的语义相关度。

搜索排名单元 740，适于根据所述待分析问答对的相关联程度优化所述待分析问答对的搜索排名。

由于待分析问答对的相关联程度反映了质量，所以可以利用相关联程度优化所述待分析问答对的搜索排名，排名效果更好。具体的方法，可以是以所述待分析问答对的相关联程度的次序作为所述待分析问答对的搜索排名，即相关联程度高的问答对的搜索排名靠前；也可以是先根据搜索排列技术初步排列所述待分析问答对所属的网站，根据该初步排列的次序号与所述待分析问答对的相关联程度计算所述待分析问答对的搜索排名，例如，可以将所述待分析问答对所属的网站的初步排列的次序号与所述待分析问答对的相关联程度相乘，以相乘运算的结果的次序作为所述待分析问答对的搜索排名。

在本实施例中，该装置还包括问答知识库构建单元 750，问答知识库构建单元 750 适于预先从含有问答对的网页提取出多个问答对，根据提取的问答对构建包括多条问答知识记录的问答知识库。在图 7 所示的装置中，问答知识库 710 是已有的，由于实际网络的信息量不断增加，信息内容的变化速度快，问答知识库 710 的内容往往需要更新，本实施例通过增设问答知识库构建单元 750 构建（或者说更新）问答知识库 710，可以保证问答知识库 710 的内容的即时性和可靠性。本实施例的问答知识库构建单元 750 与如图 5 所示的问答知识库构建单元 400 相同，此处不再重复说明。

图 7 中的计算单元 630 具体包括词语提取子单元和相关联程度计算子单元（图未示）。

词语提取子单元，适于对待分析的问答对的问题内容和答案内容进行词语提取操作，得到至少一个待分析问题词语和至少一个待分析答案词语。

在本发明的一个实施例中，词语提取子单元，适于对待分析的问答对的问题内容和答案内容进行分词、去除停用词、词合并（word join），和提取实体词（例如名词、动词等）的操作，以得到至少一个待分析问题词语和至少一个待分析答案词语。

相关联程度计算子单元，适于根据待分析问题词语和待分析答案词语，从问答知识库选择至少一条问答知识记录，根据所选择的问答知识记录计算待分析的问答对的相关联程度。

在本发明的一个实施例中，相关联程度计算子单元，适于选取其包括的问题词语与待分析问题词语匹配且包括的答案词语与待分析答案词语匹配的问答知识记录。本实施例中，问题词语与待分析问题词语匹配是指待分析问题词语与问题词语相同或待分析问题词语是问题词语的子串；答案词语与待分析答案词语匹配是指待分析答案词语与答案词语相同或待分析答案词语是答案词语的子串；根据所述选取的问答知识记录中对应于相同类别的问答知识记录，得到该待分析的问答对针对各个类别的相关联程度，更具体地，是将选取的问答知识记录中对应于相同类别的问答知识记录的语义相关度加权（例如，权值为 1 或 100）相加而得到该待分析的问答对分别针对各个类别的相关联程度，由此得到至少一个（本实施例中的相关联程度的数目即待分析问答对对应的类别的数目）相关联程度；选取上述该待分析的问答对针对各个类别的相关联程度的最大值，以该最大值作为待分析的问答对的相关联程度。

图 8 示出了根据本发明一个实施例的确定网络资源点的抓取频率的方法的流程图。该方法包括如下步骤 S810、步骤 S820 和步骤 S830：

S810、由网络资源点抓取多个待分析问答对。

在本发明的一个实施例中，可以是对于特定的需要确定抓取频率的网络资源点，例如需要确定抓取频率的问答社区，使用楼层识别技术，根据楼主（即针对一个问题首个发出帖子的使用者）提问题，1 楼 2 楼（即依序回复帖子的使用者）等回复的内容为答案的方式，来提取待分析问答对。

S820、根据包括多条问答知识记录的问答知识库，获取每个待分析问答对的相关联程度。

本实施例的步骤 S820，可以通过利用问答知识库从语义方面对待分析问答对的问题内容和答案内容进

行分析以获得待分析问答对的相关联程度，评价效果更好而且容易实现。

更具体地，本实施例的步骤 S820 的获得待分析问答对的相关联程度的具体实施方式，与如图 1、4 所示的获取问答对的相关联程度的方法大致相同，此处不再重复。

进一步地，所述包括多条问答知识记录的问答知识库，是通过预先从含有问答对的网页提取出多个问答对，根据提取的问答对构建而得到的。在本发明的一个实施例中，在从含有问答对的网页提取出多个问答对时，抓取与所述问答对对应的类别。则在根据提取的问答对构建问答知识库时，根据问答对和与所述问答对对应的类别构建问答知识记录。得到的问答知识库之中的每个问答知识记录对应于一个类别，分别包括一个问题词语（QW）、一个答案词语（AW），以及所述问题词语和所述答案词语之间的语义相关度。通过利用由网页提取的海量的、高质量的问答对构建包括多条问答知识记录的问答知识库，可以基于对海量信息的学习而获取多条问答知识记录的问题词语和答案词语之间的语义相关度；通过利用从网页提取得到的信息构建问答知识库，适用的范围更广，方法的通用性更强。

更具体地，本实施例的方法还包括构建问答知识库的步骤，其中构建问答知识库的流程与图 2 所示的流程大致相同；本实施例的问答知识库的解释模型与如图 3 所示的解释模型大致相同。此处不再重复。

S830、根据所述待分析问答对的相关联程度确定所述网络资源点的抓取频率。

由于待分析问答对的相关联程度反映了质量，所以可以利用多个待分析问答对的相关联程度确定网络资源点的质量，进而确定网络资源点的抓取频率。

具体的方法，可以是以所述待分析问答对的相关联程度的平均值作为所述网络资源点的抓取频率，即相关联程度的平均值大（即质量好）的网络资源点的抓取频率越高（例如，蜘蛛爬虫爬取该网络资源点的频率高）；也可以是使用蜘蛛爬虫获取所述网络资源点的初始抓取频率，计算所述待分析问答对的相关联程度的平均值，使用该平均值调整所述初始抓取频率而确定所述网络资源点的抓取频率，例如，可以使用蜘蛛爬虫获取所述网络资源点的初始抓取频率，使用所述相关联程度的平均值对上述初始抓取频率进行加权（包括相乘、归一化等）而确定所述网络资源点的抓取频率，使得质量高的网络资源点的抓取频率得到提高，从而可以优化搜索质量。

本实施例通过分析由网络资源点抓取待分析问答对的相关联程度，并根据相关联程度确定网络资源点的抓取频率，可以提高抓取结果的准确性。

图 9 示出了根据本发明一个实施例的确定网络资源点的抓取频率的装置的框图。该装置包括问答知识库 91、资源分析单元 920、计算单元 930 和抓取频率获取单元 940。

问答知识库 910，适于存储多条问答知识记录。本实施例的问答知识库 910 能够通过抓取网页中的海量问答对构建得到。

资源分析单元 920，适于由网络资源点抓取多个待分析问答对。

在本发明的一个实施例中，资源分析单元 920 可以对于特定的需要确定抓取频率的网络资源点，例如需要确定抓取频率的问答社区，使用楼层识别技术，根据楼主（即针对一个问题首个发出帖子的使用者）提问题，1 楼 2 楼（即依序回复帖子的使用者）等回复的内容为答案的方式，来提取待分析问答对。

计算单元 930，适于根据问答知识库获取每个待分析问答对的相关联程度。

本发明的计算单元 930 可以通过利用问答知识库从语义方面对待分析问答对的问题内容和答案内容进行分析以获得待分析问答对的相关联程度，评价效果更好而且容易实现。问答知识库 910 利用由网页提取的海量的、高质量的问答对构建并且包括多条问答知识记录，可以基于对海量信息的学习而获取多条问答知识记录的问题词语和答案词语之间的语义相关度。

抓取频率确定单元 940，适于根据所述待分析问答对的相关联程度确定所述网络资源点的抓取频率。

由于待分析问答对的相关联程度反映了质量，所以可以利用多个待分析问答对的相关联程度确定网络资源点的质量，进而确定网络资源点的抓取频率。具体的方法，可以是以所述待分析问答对的相关联程度的平均值作为所述网络资源点的抓取频率，即相关联程度的平均值大（即质量好）的网络资源点的抓取频率越高（例如，蜘蛛爬虫爬取该网络资源点的频率高）；也可以是使用蜘蛛爬虫获取所述网络资源点的初始抓取频率，计算所述待分析问答对的相关联程度的平均值，使用该平均值调整所述初始抓取频率而确定所述网络资源点的抓取频率，例如，可以使用蜘蛛爬虫获取所述网络资源点的初始抓取频率，使用所述相关联程度的平均值对上述初始抓取频率进行加权（包括相乘、归一化等）而确定所述网络资源点的抓取频率，使得质量高的网络资源点的抓取频率得到提高，从而可以优化搜索质量。

在本实施例中，该装置还包括问答知识库构建单元 950，问答知识库构建单元 950 适于预先从含有问答对的网页提取出多个问答对，根据提取的问答对构建包括多条问答知识记录的问答知识库。在图 9 所示的装置中，问答知识库 910 是已有的，由于实际网络的信息量不断增加，信息内容的变化速度快，问答知识库 910 的内容往往需要更新，本实施例通过增设问答知识库构建单元 950 构建（或者说更新）问答知识库，可以保证问答知识库的内容的即时性和可靠性。本实施例的问答知识库构建单元 950 与如图 5 所示的问答知识库构建单元 400 相同，此处不再重复说明。

图 9 中计算单元 930 具体包括词语提取子单元和相关联程度计算子单元（图未示）。

词语提取子单元，适于对待分析的问答对的问题内容和答案内容进行词语提取操作，得到至少一个待分析问题词语和至少一个待分析答案词语。

在本发明的一个实施例中，词语提取子单元，适于对待分析的问答对的问题内容和答案内容进行分词、去除停用词、词合并（word join），和提取实体词（例如名词、动词等）的操作，以得到至少一个待分析问题词语和至少一个待分析答案词语。

相关联程度计算子单元，适于根据待分析问题词语和待分析答案词语，从问答知识库选择至少一条问答知识记录，根据所选择的问答知识记录计算待分析的问答对的相关联程度。

在本发明的一个实施例中，相关联程度计算子单元，适于选取其包括的问题词语与待分析问题词语匹配且包括的答案词语与待分析答案词语匹配的问答知识记录。本实施例中，问题词语与待分析问题词语匹配是指待分析问题词语与问题词语相同或待分析问题词语是问题词语的子串；答案词语与待分析答案词语匹配是指待分析答案词语与答案词语相同或待分析答案词语是答案词语的子串；根据所述选取的问答知识记录中对应于相同类别的问答知识记录，得到该待分析的问答对针对各个类别的相关联程度，更具体地，是将选取的问答知识记录中对应于相同类别的问答知识记录的语义相关度加权（例如，权值为 1 或 100）相加而得到该待分析的问答对分别针对各个类别的相关联程度，由此得到至少一个（本实施例中的相关联程度的数目即待分析问答对对应的类别的数目）相关联程度；选取上述该待分析的问答对针对各个类别的相关联程度的最大值，以该最大值作为待分析的问答对的相关联程度。

本发明的各个部件实施例可以以硬件实现，或者以在一个或者多个处理器上运行的软件模块实现，或者以它们的组合实现。本领域的技术人员应当理解，可以在实践中使用微处理器或者数字信号处理器（DSP）来实现根据本发明实施例的获取问答对的相关联程度的装置，优化问答对的搜索排名的装置，以及确定网络资源点的抓取频率的装置中的一些或者全部部件的一些或者全部功能。本发明还可以实现为用于执行这里所描述的方法的一部分或者全部的设备或者装置程序（例如，计算机程序和计算机程序产品）。这样的实现本发明的程序可以存储在计算机可读介质上，或者可以具有一个或者多个信号的形式。这样的信号可以从因特网网站上下载得到，或者在载体信号上提供，或者以任何其他形式提供。

例如，图 10 示出了用于执行根据本发明的获取问答对的相关联程度的方法，优化问答对的搜索排名的方法，以及确定网络资源点的抓取频率的方法的服务器，例如应用服务器的框图。该应用服务器传统上包括处理器 1010 和以存储器 1020 形式的计算机程序产品或者计算机可读介质。存储器 1020 可以是诸如闪存、EEPROM（电可擦除可编程只读存储器）、EPROM、硬盘或者 ROM 之类的电子存储器。存储器 1020 具有用于执行上述方法中的任何方法步骤的程序代码 1031 的存储空间 1030。例如，用于程序代码的存储空间 1030 可以包括分别用于实现上面的方法中的各种步骤的各个程序代码 1031。这些程序代码可以从一个或者多个计算机程序产品中读出或者写入到这一个或者多个计算机程序产品中。这些计算机程序产品包括诸如硬盘，紧致盘（CD）、存储卡或者软盘之类的程序代码载体。这样的计算机程序产品通常为如参考图 11 所述的便携式或者固定存储单元。该存储单元可以具有与图 10 的应用服务器中的存储器 1020 类似布置的存储段、存储空间等。程序代码可以例如以适当形式进行压缩。通常，存储单元包括计算机可读代码 1131'，即可以由例如诸如处理器 1010 之类的处理器读取的代码，这些代码当由服务器运行时，导致该服务器执行上面所描述的方法中的各个步骤。

本文中所称的“一个实施例”、“实施例”或者“一个或者多个实施例”意味着，结合实施例描述的特定特征、结构或者特性包括在本发明的至少一个实施例中。此外，请注意，这里“在一个实施例中”的词语例子不一定全指同一个实施例。

在此处所提供的说明书中，说明了大量具体细节。然而，能够理解，本发明的实施例可以在没有这些具体细节的情况下被实践。在一些实例中，并未详细示出公知的方法、结构和技术，以便不模糊对本说明书的理解。

应该注意的是上述实施例对本发明进行说明而不是对本发明进行限制，并且本领域技术人员在不脱离所附权利要求的范围的情况下可设计出替换实施例。在权利要求中，不应将位于括号之间的任何参考符号构造成对权利要求的限制。单词“包含”不排除存在未列在权利要求中的元件或步骤。位于元件之前的单词“一”或“一个”不排除存在多个这样的元件。本发明可以借助于包括有若干不同元件的硬件以及借助于适当编程的计算机来实现。在列举了若干装置的单元权利要求中，这些装置中的若干个可以是通过同一个硬件项来具体体现。单词第一、第二、以及第三等的使用不表示任何顺序。可将这些单词解释为名称。

此外，还应当注意，本说明书中使用的语言主要是为了可读性和教导的目的而选择的，而不是为了解释或者限定本发明的主题而选择的。因此，在不偏离所附权利要求书的范围和精神的情况下，对于本技术领域的普通技术人员来说许多修改和变更都是显而易见的。对于本发明的范围，对本发明所做的公开是说明性的，而非限制性的，本发明的范围由所附权利要求书限定。

权利要求书

1、一种获取问答对的相关联程度的装置，该装置包括：

问答知识库，适于存储多条问答知识记录；

词语提取单元，适于对待分析的问答对的问题内容和答案内容进行词语提取操作，得到至少一个待分析问题词语和至少一个待分析答案词语；

相关联程度计算单元，适于根据待分析问题词语和待分析答案词语，从问答知识库选择至少一条问答知识记录，根据所选择的问答知识记录计算待分析的问答对的相关联程度。

2、根据权利要求 1 所述的装置，其中，该装置进一步包括问答知识库构建单元，

所述问答知识库构建单元，适于预先从含有问答对的网页提取出多个问答对，根据提取的问答对构建包括多条问答知识记录的问答知识库；

所述问答知识库构建单元，进一步适于在从含有问答对的网页提取出多个问答对时，抓取与所述问答对对应的类别；

所述问答知识库构建单元，进一步适于在根据提取的问答对构建问答知识库时，根据问答对和与所述问答对对应的类别构建问答知识记录；每个问答知识记录对应于一个类别，分别包括一个问题词语、一个答案词语，以及所述问题词语和所述答案词语之间的语义相关度。

3、根据权利要求 1 或 2 所述的装置，其中，

所述相关联程度计算单元，适于选取其包括的问题词语与待分析问题词语匹配且包括的答案词语与待分析答案词语匹配的问答知识记录；根据所述选取的问答知识记录中对应于相同类别的问答知识记录，得到该待分析的问答对针对各个类别的相关联程度；选取上述该待分析的问答对针对各个类别的相关联程度的最大值，以该最大值作为待分析的问答对的相关联程度。

4、根据权利要求 2 所述的装置，其中，

所述问答知识库构建单元，适于对每个问答对执行以下操作：

对该问答对的问题内容和答案内容进行词语提取操作，得到问题词语集合和答案词语集合；令问题词语集合中的每个问题词语与答案词语集合中的每个答案词语分别在与该问答对对应的每个类别上形成一条信息记录；

所述问答知识库构建单元，适于对每一条信息记录，执行以下操作：

计算该答案词语属于该类别的概率，计算在该类别上该答案词语对该问题词语的解释的专一程度，计算在该类别上该问题词语用该答案词语进行解释的强度；将上述概率、专一程度和强度相乘，所得到的乘积是该答案词语和该问题词语的语义相关度；令该问题词语、该答案词语和其语义相关度形成一条对应于该类别的问答知识记录。

5、根据权利要求 1 至 4 任一权利要求所述的装置，其中，

所述相关联程度计算单元，适于将选取的问答知识记录中对应于相同类别的问答知识记录的语义相关度加权相加，得到该待分析的问答对分别针对各个类别的相关联程度。

6、根据权利要求 1 至 5 任一权利要求所述的装置，其中，

可选地，所述词语提取单元，适于对待分析的问答对的问题内容和答案内容进行分词、去除停用词、词合并，和提取实体词的操作。

7、根据权利要求 1 至 6 任一权利要求所述的装置，其中，

所述问答知识库构建单元，适于按照如下的方法计算该答案词语属于该类别的概率：

$$P(Ck|AWj)=P(AWj|Ck)*\frac{P(Ck)}{P(AWj)};$$

所述问答知识库构建单元，适于按照如下的方法计算在该类别上各个答案词语对该问题词语的解释的专一程度：

$$specific(QWi,AWj|C=Ck)=P(QWi|AWj,C=Ck)=\frac{\#(QWi,AWj)}{\#(AWj)}|C=Ck;$$

所述问答知识库构建单元，适于按照如下的方法计算在该类别上该问题词语用各个答案词语进行解释的强度：

$$interpret(QWi,AWj|C=Ck)=P(AWj|QWi,C=Ck)=\frac{\#(QWi,AWj)}{\sum_{j=1}^{x}\#(QWi,AWj)}|C=Ck;$$

所述问答知识库构建单元，适于按照如下的方法将上述概率、专一程度和强度相乘：

$$weight(QWi, AWj|C=Ck) = P(Ck|AWj) * specific(QWi, AWj|C=Ck) * interpret(QWi, AWj|C=Ck)$$；

其中，P（C_k）表示类别 C_k 出现的概率；P（AW_j）表示答案为 AW_j 的概率；P（AW_j | C_k）表示 C_k 类别属于 AW_j 的概率；

#（QW_i，AW_j）表示问题词语为 QW_i 且答案词语为 AW_j 的次数；

#（AW_j）表示答案词语为 AW_j 的次数。

8、一种优化问答对的搜索排名的装置，该装置包括：

问答知识库，适于存储多条问答知识记录；

搜索单元，适于接收使用者的搜索请求，根据使用者的搜索请求，获取与搜索请求匹配的多个待分析问答对；

计算单元，适于根据问答知识库获取每个待分析问答对的相关联程度；

搜索排名单元，适于根据所述待分析问答对的相关联程度优化所述待分析问答对的搜索排名。

9、根据权利要求 8 所述的装置，其中，所述计算单元包括：

词语提取子单元，适于对待分析的问答对的问题内容和答案内容进行词语提取操作，得到至少一个待分析问题词语和至少一个待分析答案词语；

相关联程度计算子单元，适于根据待分析问题词语和待分析答案词语，从问答知识库选择至少一条问答知识记录，根据所选择的问答知识记录计算待分析的问答对的相关联程度。

10、根据权利要求 8 或 9 所述的装置，其中，

所述搜索排名单元，适于以所述待分析问答对的相关联程度的次序作为所述待分析问答对的搜索排名。

11、根据权利要求 8 至 10 任一项所述的装置，其中，该装置还包括问答知识库构建单元，

所述问答知识库构建单元，适于预先从含有问答对的网页提取出多个问答对，根据提取的问答对构建包括多条问答知识记录的问答知识库；

所述问答知识库构建单元，进一步适于在从含有问答对的网页提取出多个问答对时，抓取与所述问答对对应的类别；

所述问答知识库构建单元，进一步适于在根据提取的问答对构建问答知识库时，根据问答对和与所述问答对对应的类别构建问答知识记录；每个问答知识记录对应于一个类别，分别包括一个问题词语、一个答案词语，以及所述问题词语和所述答案词语之间的语义相关度。

12、根据权利要求 8 至 11 任一项所述的装置，其中，

所述相关联程度计算子单元，适于选取其包括的问题词语与待分析问题词语匹配且包括的答案词语与待分析答案词语匹配的问答知识记录；根据所述选取的问答知识记录中对应于相同类别的问答知识记录，得到该待分析的问答对针对各个类别的相关联程度；选取上述该待分析的问答对针对各个类别的相关联程度的最大值，以该最大值作为待分析的问答对的相关联程度。

13、根据权利要求 8 至 12 任一项所述的装置，其中，

所述相关联程度计算子单元，适于将选取的问答知识记录中对应于相同类别的问答知识记录的语义相关度加权相加，得到该待分析的问答对分别针对各个类别的相关联程度。

14、根据权利要求 8 至 13 任一项所述的装置，其中，

所述词语提取子单元，适于对待分析的问答对的问题内容和答案内容进行分词、去除停用词、词合并，和提取实体词的操作。

15、根据权利要求 8 至 14 任一项所述的装置，其中，

所述问答知识库构建单元，适于对每个问答对执行以下操作：对该问答对的问题内容和答案内容进行词语提取操作，得到问题词语集合和答案词语集合；令问题词语集合中的每个问题词语与答案词语集合中的每个答案词语分别在与该问答对对应的每个类别上形成一条信息记录；

所述问答知识库构建单元，适于对每一条信息记录，执行以下操作：计算该答案词语属于该类别的概率，计算在该类别上该答案词语对该问题词语的解释的专一程度，计算在该类别上该问题词语用该答案词语进行解释的强度；将上述概率、专一程度和强度相乘，所得到的乘积是该答案词语和该问题词语的语义相关度；令该问题词语、该答案词语和其语义相关度形成一条对应于该类别的问答知识记录。

16、根据权利要求 8 至 15 任一项所述的装置，其中，

所述问答知识库构建单元，适于按照如下的方法计算该答案词语属于该类别的概率：

$$P(Ck|AWj) = P(AWj|Ck) * \frac{P(Ck)}{P(AWj)};$$

所述问答知识库构建单元，适于按照如下的方法计算在该类别上各个答案词语对该问题词语的解释的专一程度：

$$specific(QWi, AWj|C=Ck)=P(QWi|AWj, C=Ck)=\frac{\#(QWi, AWj)}{\#(AWj)}|C=Ck;$$

所述问答知识库构建单元，适于按照如下的方法计算在该类别上该问题词语用各个答案词语进行解释的强度：

$$interpret(QWi, AWj|C=Ck)=P(AWj|QWi, C=Ck)=\frac{\#(QWi, AWj)}{\sum_{j=1}^{x}\#(QWi, AWj)}|C=Ck;$$

所述问答知识库构建单元，适于按照如下的方法将上述概率、专一程度和强度相乘：

$$weight(QWi, AWj|C=Ck)=P(Ck|AWj)*specific(QWi, AWj|C=Ck)*interpret(QWi, AWj|C=Ck);$$

其中，P（C_k）表示类别 C_k 出现的概率；P（AW_j）表示答案为 AW_j 的概率；P（$AW_j|C_k$）表示 C_k 类别属于 AW_j 的概率；

#（QW_i，AW_j）表示问题词语为 QW_i 且答案词语为 AW_j 的次数；

#（AW_j）表示答案词语为 AW_j 的次数。

17、一种确定网络资源点的抓取频率的装置，该装置包括：

问答知识库，适于存储多条问答知识记录；

资源分析单元，适于由网络资源点抓取多个待分析问答对；

计算单元，适于根据问答知识库获取每个待分析问答对的相关联程度；

抓取频率确定单元，根据所述待分析问答对的相关联程度确定所述网络资源点的抓取频率。

18、根据权利要求 17 所述的装置，其中，所述计算单元包括：

词语提取子单元，适于对待分析的问答对的问题内容和答案内容进行词语提取操作，得到至少一个待分析问题词语和至少一个待分析答案词语；

相关联程度计算子单元，适于根据待分析问题词语和待分析答案词语，从问答知识库选择至少一条问答知识记录，根据所选择的问答知识记录计算待分析的问答对的相关联程度。

19、根据权利要求 17 或 18 所述的装置，其中，

所述抓取频率确定单元，适于以所述待分析问答对的相关联程度的平均值作为所述网络资源点的抓取频率；或，使用蜘蛛爬虫获取所述网络资源点的初始抓取频率，计算所述待分析问答对的相关联程度的平均值，使用该平均值调整所述初始抓取频率而确定所述网络资源点的抓取频率。

20、根据权利要求 17 至 19 任一项所述的装置，其中，该装置还包括问答知识库构建单元，

所述问答知识库构建单元，适于预先从含有问答对的网页提取出多个问答对，根据提取的问答对构建包括多条问答知识记录的问答知识库；

所述问答知识库构建单元，进一步适于在从含有问答对的网页提取出多个问答对时，抓取与所述问答对对应的类别；

所述问答知识库构建单元，进一步适于在根据提取的问答对构建问答知识库时，根据问答对和与所述问答对对应的类别构建问答知识记录；每个问答知识记录对应于一个类别，分别包括一个问题词语、一个答案词语，以及所述问题词语和所述答案词语之间的语义相关度。

21、根据权利要求 17 至 20 任一项所述的装置，其中，

所述相关联程度计算子单元，适于选取其包括的问题词语与待分析问题词语匹配且包括的答案词语与待分析答案词语匹配的问答知识记录；根据所述选取的问答知识记录中对应于相同类别的问答知识记录，得到该待分析的问答对针对各个类别的相关联程度；选取上述该待分析的问答对针对各个类别的相关联程度的最大值，以该最大值作为待分析的问答对的相关联程度。

22、根据权利要求 17 至 21 任一项所述的装置，其中，

所述相关联程度计算子单元，适于将选取的问答知识记录中对应于相同类别的问答知识记录的语义相关度加权相加，得到该待分析的问答对分别针对各个类别的相关联程度。

23、根据权利要求 17 至 22 任一项所述的装置，其中，

所述词语提取子单元，适于对待分析的问答对的问题内容和答案内容进行分词、去除停用词、词合并，和提取实体词的操作。

24、根据权利要求 17 至 23 任一项所述的装置，其中，

所述问答知识库构建单元，适于对每个问答对执行以下操作：对该问答对的问题内容和答案内容进行词语提取操作，得到问题词语集合和答案词语集合；令问题词语集合中的每个问题词语与答案词语集合中的每个答案词语分别在与该问答对对应的每个类别上形成一条信息记录；

所述问答知识库构建单元，适于对每一条信息记录，执行以下操作：计算该答案词语属于该类别的概率，计算在该类别上该答案词语对该问题词语的解释的专一程度，计算在该类别上该问题词语用该答案词

语进行解释的强度；将上述概率、专一程度和强度相乘，所得到的乘积是该答案词语和该问题词语的语义相关度；令该问题词语、该答案词语和其语义相关度形成一条对应于该类别的问答知识记录。

25、根据权利要求 17 至 24 任一项所述的装置，其中，

所述问答知识库构建单元，适于按照如下的方法计算该答案词语属于该类别的概率：

$$P(Ck|AWj)=P(AWj|Ck)*\frac{P(Ck)}{P(AWj)};$$

所述问答知识库构建单元，适于按照如下的方法计算在该类别上各个答案词语对该问题词语的解释的专一程度：

$$specific(QWi,AWj|C=Ck)=P(QWi|AWj,C=Ck)=\frac{\#(QWi,AWj)}{\#(AWj)}|C=Ck;$$

所述问答知识库构建单元，适于按照如下的方法计算在该类别上该问题词语用各个答案词语进行解释的强度：

$$interpret(QWi,AWj|C=Ck)=P(AWj|QWi,C=Ck)=\frac{\#(QWi,AWj)}{\sum_{j=1}^{x}\#(QWi,AWj)}|C=Ck;$$

所述问答知识库构建单元，适于按照如下的方法将上述概率、专一程度和强度相乘：

$$weight(QWi,AWj|C=Ck)=P(Ck|AWj)*specific(QWi,AWj|C=Ck)*interpret(QWi,AWj|C=Ck);$$

其中，P（C_k）表示类别 C_k 出现的概率；P（AW_j）表示答案为 AW_j 的概率；P（$AW_j|C_k$）表示 C_k 类别属于 AW_j 的概率；

#（QW_i，AW_j）表示问题词语为 QW_i 且答案词语为 AW_j 的次数；

#（AW_j）表示答案词语为 AW_j 的次数。

26、一种获取问答对的相关联程度的方法，该方法包括如下步骤：

对待分析的问答对的问题内容和答案内容进行词语提取操作，得到至少一个待分析问题词语和至少一个待分析答案词语；

根据待分析问题词语和待分析答案词语，从包括多条问答知识记录的问答知识库选择至少一条问答知识记录，根据所选择的问答知识记录计算待分析的问答对的相关联程度。

27、根据权利要求 26 所述的方法，其中，该方法进一步包括：

预先从含有问答对的网页提取出多个问答对，根据提取的问答对构建包括多条问答知识记录的问答知识库；

在从含有问答对的网页提取出多个问答对时，抓取与所述问答对对应的类别；

在根据提取的问答对构建问答知识库时，根据问答对和与所述问答对对应的类别构建问答知识记录；

每个问答知识记录对应于一个类别，分别包括一个问题词语、一个答案词语，以及所述问题词语和所述答案词语之间的语义相关度。

28、根据权利要求 26 或 27 所述的方法，其中，

所述根据待分析问题词语和待分析答案词语，从问答知识库选择至少一条问答知识记录，根据所选择的问答知识记录计算待分析的问答对的相关联程度，具体包括：

选取其包括的问题词语与待分析问题词语匹配且包括的答案词语与待分析答案词语匹配的问答知识记录；

根据所述选取的问答知识记录中对应于相同类别的问答知识记录，得到该待分析的问答对针对各个类别的相关联程度；

选取上述该待分析的问答对针对各个类别的相关联程度的最大值，以该最大值作为待分析的问答对的相关联程度。

29、根据权利要求 26 至 28 任一权利要求所述的方法所述的方法，其中，

根据所述选取的问答知识记录中对应于相同类别的问答知识记录，得到该待分析的问答对分别针对各个类别的相关联程度，具体包括：

将选取的问答知识记录中对应于相同类别的问答知识记录的语义相关度加权相加，得到该待分析的问答对分别针对各个类别的相关联程度。

30、根据权利要求 26 至 29 任一权利要求所述的方法所述的方法，其中，所述根据问答对和与所述问答对对应的类别构建问答知识库，具体包括：

对每个问答对，对该问答对的问题内容和答案内容进行词语提取操作，得到问题词语集合和答案词语集合；

令问题词语集合中的每个问题词语与答案词语集合中的每个答案词语分别在与该问答对对应的每个

类别上形成一条信息记录;

对每一条信息记录，执行以下操作:

计算该答案词语属于该类别的概率，计算在该类别上该答案词语对该问题词语的解释的专一程度，计算在该类别上该问题词语用该答案词语进行解释的强度;

将上述概率、专一程度和强度相乘，所得到的乘积是该答案词语和该问题词语的语义相关度;

令该问题词语、该答案词语和其语义相关度形成一条对应于该类别的问答知识记录。

31、根据权利要求 26 至 30 任一权利要求所述的方法所述的方法，其中，

所述计算该答案词语属于该类别的概率，具体包括:

$$P(Ck|AWj)=P(AWj|Ck)*\frac{P(Ck)}{P(AWj)};$$

所述计算在该类别上各个答案词语对该问题词语的解释的专一程度，具体包括:

$$specific(QWi,AWj|C=Ck)=P(QWi|AWj,C=Ck)=\frac{\#(QWi,AWj)}{\#(AWj)}|C=Ck;$$

所述计算在该类别上该问题词语用各个答案词语进行解释的强度，具体包括:

$$interpret(QWi,AWj|C=Ck)=P(AWj|QWi,C=Ck)=\frac{\#(QWi,AWj)}{\sum_{j=1}^{x}\#(QWi,AWj)}|C=Ck;$$

将上述概率、专一程度和强度相乘，具体包括:

$$weight(QWi,AWj|C=Ck)=P(Ck|AWj)*specific(QWi,AWj|C=Ck)*interpret(QWi,AWj|C=Ck);$$

其中，P（C_k）表示类别 C_k 出现的概率；P（AW_j）表示答案为 AW_j 的概率；P（$AW_j|C_k$）表示 C_k 类别属于 AW_j 的概率;

#（QW_i，AW_j）表示问题词语为 QW_i 且答案词语为 AW_j 的次数;

#（AW_j）表示答案词语为 AW_j 的次数。

32、根据权利要求 26 至 31 任一权利要求所述的方法，其中，

所述对所述待分析的问答对的问题内容和答案内容进行词语提取操作，具体包括：对待分析的问答对的问题内容和答案内容进行分词、去除停用词、词合并，和提取实体词的操作。

33、一种优化问答对的搜索排名的方法，该方法包括如下步骤:

接收使用者的搜索请求，根据使用者的搜索请求，获取与搜索请求匹配的多个待分析问答对;

根据包括多条问答知识记录的问答知识库，获取每个待分析问答对的相关联程度;

根据所述待分析问答对的相关联程度优化所述待分析问答对的搜索排名。

34、根据权利要求 33 所述的方法，其中，所述根据包括多条问答知识记录的问答知识库获取每个待分析问答对的相关联程度，包括对每个待分析问答对执行以下操作:

对该待分析的问答对的问题内容和答案内容进行词语提取操作，得到至少一个待分析问题词语和至少一个待分析答案词语;

根据待分析问题词语和待分析答案词语，从问答知识库选择至少一条问答知识记录，根据所选择的问答知识记录计算该待分析的问答对的相关联程度。

35、根据权利要求 33 或 34 所述的方法，其中，所述根据所述待分析问答对的相关联程度调整所述待分析问答对的搜索排名，具体包括:

以所述待分析问答对的相关联程度的次序作为所述待分析问答对的搜索排名。

36、根据权利要求 33 至 35 任一项所述的方法，其中，该方法进一步包括:

预先从含有问答对的网页提取出多个问答对，根据提取的问答对构建包括多条问答知识记录的问答知识库;

在从含有问答对的网页提取出多个问答对时，抓取与所述问答对对应的类别;

在根据提取的问答对构建问答知识库时，根据问答对和与所述问答对对应的类别构建问答知识记录;

每个问答知识记录对应于一个类别，分别包括一个问题词语、一个答案词语，以及所述问题词语和所述答案词语之间的语义相关度。

37、根据权利要求 33 至 36 任一项 所述的方法，其中，

所述根据待分析问题词语和待分析答案词语，从问答知识库选择至少一条问答知识记录，根据所选择的问答知识记录计算待分析的问答对的相关联程度，具体包括:

选取其包括的问题词语与待分析问题词语匹配且包括的答案词语与待分析答案词语匹配的问答知识记录;

根据所述选取的问答知识记录中对应于相同类别的问答知识记录，得到该待分析的问答对针对各个类别的相关联程度；

选取上述该待分析的问答对针对各个类别的相关联程度的最大值，以该最大值作为待分析的问答对的相关联程度。

38、根据权利要求 33 至 37 任一项所述的方法，其中，

根据所述选取的问答知识记录中对应于相同类别的问答知识记录，得到该待分析的问答对分别针对各个类别的相关联程度，具体包括：

将选取的问答知识记录中对应于相同类别的问答知识记录的语义相关度加权相加，得到该待分析的问答对分别针对各个类别的相关联程度。

39、根据权利要求 33 至 38 任一项所述的方法，其中，

所述对所述待分析的问答对的问题内容和答案内容进行词语提取操作，具体包括：

对待分析的问答对的问题内容和答案内容进行分词、去除停用词、词合并，和提取实体词的操作。

40、根据权利要求 33 至 39 任一项所述的方法，其中，

所述根据问答对和与所述问答对对应的类别构建问答知识库，具体包括：

对每个问答对，对该问答对的问题内容和答案内容进行词语提取操作，得到问题词语集合和答案词语集合；

令问题词语集合中的每个问题词语与答案词语集合中的每个答案词语分别在与该问答对对应的每个类别上形成一条信息记录；

对每一条信息记录，执行以下操作：

计算该答案词语属于该类别的概率，计算在该类别上该答案词语对该问题词语的解释的专一程度，计算在该类别上该问题词语用该答案词语进行解释的强度；

将上述概率、专一程度和强度相乘，所得到的乘积是该答案词语和该问题词语的语义相关度；

令该问题词语、该答案词语和其语义相关度形成一条对应于该类别的问答知识记录。

41、根据权利要求 33 至 40 任一项所述的方法，其中，

所述计算该答案词语属于该类别的概率，具体包括：

$$P(Ck|AWj)=P(AWj|Ck)*\frac{P(Ck)}{P(AWj)};$$

所述计算在该类别上各个答案词语对该问题词语的解释的专一程度，具体包括：

$$specific(QWi,AWj|C=Ck)=P(QWi|AWj,C=Ck)=\frac{\#(QWi,AWj)}{\#(AWj)}|C=Ck;$$

所述计算在该类别上该问题词语用各个答案词语进行解释的强度，具体包括：

$$interpret(QWi,AWj|C=Ck)=P(AWj|QWi,C=Ck)=\frac{\#(QWi,AWj)}{\sum_{j=1}^{x}\#(QWi,AWj)}|C=Ck;$$

将上述概率、专一程度和强度相乘，具体包括：

$$weight(QWi,AWj|C=Ck)=P(Ck|AWj)*specific(QWi,AWj|C=Ck)*interpret(QWi,AWj|C=Ck);$$

其中，$P(C_k)$ 表示类别 C_k 出现的概率；$P(AW_j)$ 表示答案为 AW_j 的概率；$P(AW_j|C_k)$ 表示 C_k 类别属于 AW_j 的概率；

$\#(QW_i, AW_j)$ 表示问题词语为 QW_i 且答案词语为 AW_j 的次数；

$\#(AW_j)$ 表示答案词语为 AW_j 的次数。

42、一种确定网络资源点的抓取频率的方法，该方法包括如下步骤：

由网络资源点抓取多个待分析问答对；

根据包括多条问答知识记录的问答知识库，获取每个待分析问答对的相关联程度；

根据所述待分析问答对的相关联程度确定所述网络资源点的抓取频率。

43、根据权利要求 42 所述的方法，其中，所述根据包括多条问答知识记录的问答知识库获取每个待分析问答对的相关联程度，包括对每个待分析问答对执行以下操作：

对该待分析的问答对的问题内容和答案内容进行词语提取操作，得到至少一个待分析问题词语和至少一个待分析答案词语；

根据待分析问题词语和待分析答案词语，从问答知识库选择至少一条问答知识记录，根据所选择的问答知识记录计算该待分析的问答对的相关联程度。

44、根据权利要求 42 或 43 所述的方法，其中，所述根据所述待分析问答对的相关联程度确定所述网

络资源点的抓取频率，具体包括：

以所述待分析问答对的相关联程度的平均值作为所述网络资源点的抓取频率；

或，

使用蜘蛛爬虫获取所述网络资源点的初始抓取频率，计算所述待分析问答对的相关联程度的平均值，使用该平均值调整所述初始抓取频率而确定所述网络资源点的抓取频率。

45、根据权利要求 42 至 44 任一项所述的方法，其中，该方法进一步包括：

预先从含有问答对的网页提取出多个问答对，根据提取的问答对构建包括多条问答知识记录的问答知识库；

在从含有问答对的网页提取出多个问答对时，抓取与所述问答对对应的类别；

在根据提取的问答对构建问答知识库时，根据问答对和与所述问答对对应的类别构建问答知识记录；

每个问答知识记录对应于一个类别，分别包括一个问题词语、一个答案词语，以及所述问题词语和所述答案词语之间的语义相关度。

46、根据权利要求 42 至 45 任一项所述的方法，其中，

所述根据待分析问题词语和待分析答案词语，从问答知识库选择至少一条问答知识记录，根据所选择的问答知识记录计算待分析的问答对的相关联程度，具体包括：

选取其包括的问题词语与待分析问题词语匹配且包括的答案词语与待分析答案词语匹配的问答知识记录；

根据所述选取的问答知识记录中对应于相同类别的问答知识记录，得到该待分析的问答对针对各个类别的相关联程度；

选取上述该待分析的问答对针对各个类别的相关联程度的最大值，以该最大值作为待分析的问答对的相关联程度。

47、根据权利要求 42 至 46 任一项所述的方法，其中，

根据所述选取的问答知识记录中对应于相同类别的问答知识记录，得到该待分析的问答对分别针对各个类别的相关联程度，具体包括：

将选取的问答知识记录中对应于相同类别的问答知识记录的语义相关度加权相加，得到该待分析的问答对分别针对各个类别的相关联程度。

48、根据权利要求 42 至 47 任一项所述的方法，其中，

所述对所述待分析的问答对的问题内容和答案内容进行词语提取操作，具体包括：

对待分析的问答对的问题内容和答案内容进行分词、去除停用词、词合并，和提取实体词的操作。

49、根据权利要求 42 至 48 任一项所述的方法，其中，

所述根据问答对和与所述问答对对应的类别构建问答知识库，具体包括：

对每个问答对，对该问答对的问题内容和答案内容进行词语提取操作，得到问题词语集合和答案词语集合；

令问题词语集合中的每个问题词语与答案词语集合中的每个答案词语分别在与该问答对对应的每个类别上形成一条信息记录；

对每一条信息记录，执行以下操作：

计算该答案词语属于该类别的概率，计算在该类别上该答案词语对该问题词语的解释的专一程度，计算在该类别上该问题词语用该答案词语进行解释的强度；

将上述概率、专一程度和强度相乘，所得到的乘积是该答案词语和该问题词语的语义相关度；

令该问题词语、该答案词语和其语义相关度形成一条对应于该类别的问答知识记录。

50、根据权利要求 42 至 49 任一项所述的方法，其中，

所述计算该答案词语属于该类别的概率，具体包括：

$$P(Ck|AWj)=P(AWj|Ck)*\frac{P(Ck)}{P(AWj)};$$

所述计算在该类别上各个答案词语对该问题词语的解释的专一程度，具体包括：

$$specific(QWi,AWj|C=Ck)=P(QWi|AWj,C=Ck)=\frac{\#(QWi,AWj)}{\#(AWj)}|C=Ck;$$

所述计算在该类别上该问题词语用各个答案词语进行解释的强度，具体包括：

$$interpret(QWi,AWj|C=Ck)=P(AWj|QWi,C=Ck)=\frac{\#(QWi,AWj)}{\sum_{j=1}^{x}\#(QWi,AWj)}|C=Ck;$$

将上述概率、专一程度和强度相乘，具体包括：

$$weight(QWi, AWj|C=Ck) = P(Ck|AWj) * specific(QWi, AWj|C=Ck) * interpret(QWi, AWj|C=Ck)$$ ；

其中，P（C_k）表示类别 C_k 出现的概率；P（AW_j）表示答案为 AW_j 的概率；P（AW_j | C_k）表示 C_k 类别属于 AW_j 的概率；

#（QW_i，AW_j）表示问题词语为 QW_i 且答案词语为 AW_j 的次数；

#（AW_j）表示答案词语为 AW_j 的次数。

51、一种计算机程序，包括计算机可读代码，当所述计算机可读代码在计算设备上运行时，导致所述计算设备执行根据权利要求 26 至 50 中的任一个所述的方法。

52、一种计算机可读介质，其中存储了如权利要求 51 所述的计算机程序。

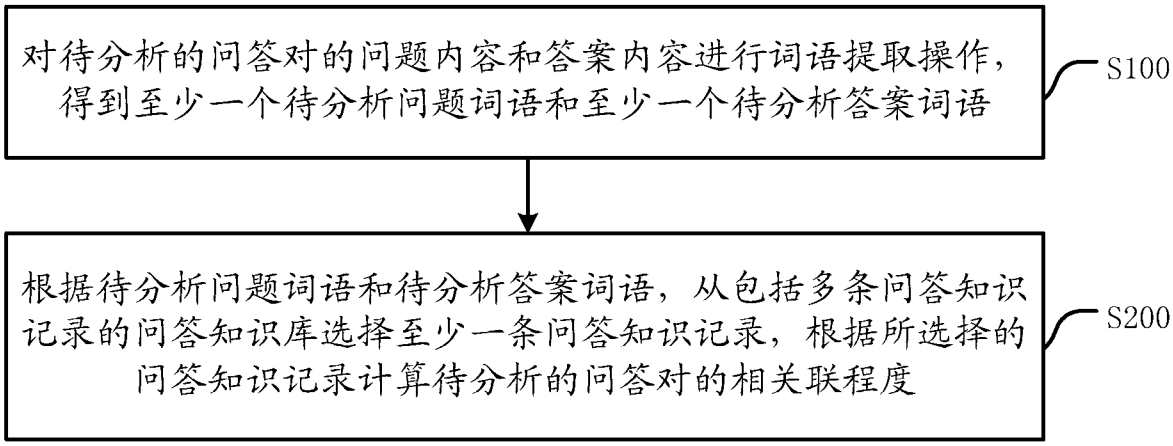

对待分析的问答对的问题内容和答案内容进行词语提取操作，得到至少一个待分析问题词语和至少一个待分析答案词语
S100
根据待分析问题词语和待分析答案词语，从包括多条问答知识记录的问答知识库选择至少一条问答知识记录，根据所选择的问答知识记录计算待分析的问答对的相关联程度
S200

图 1

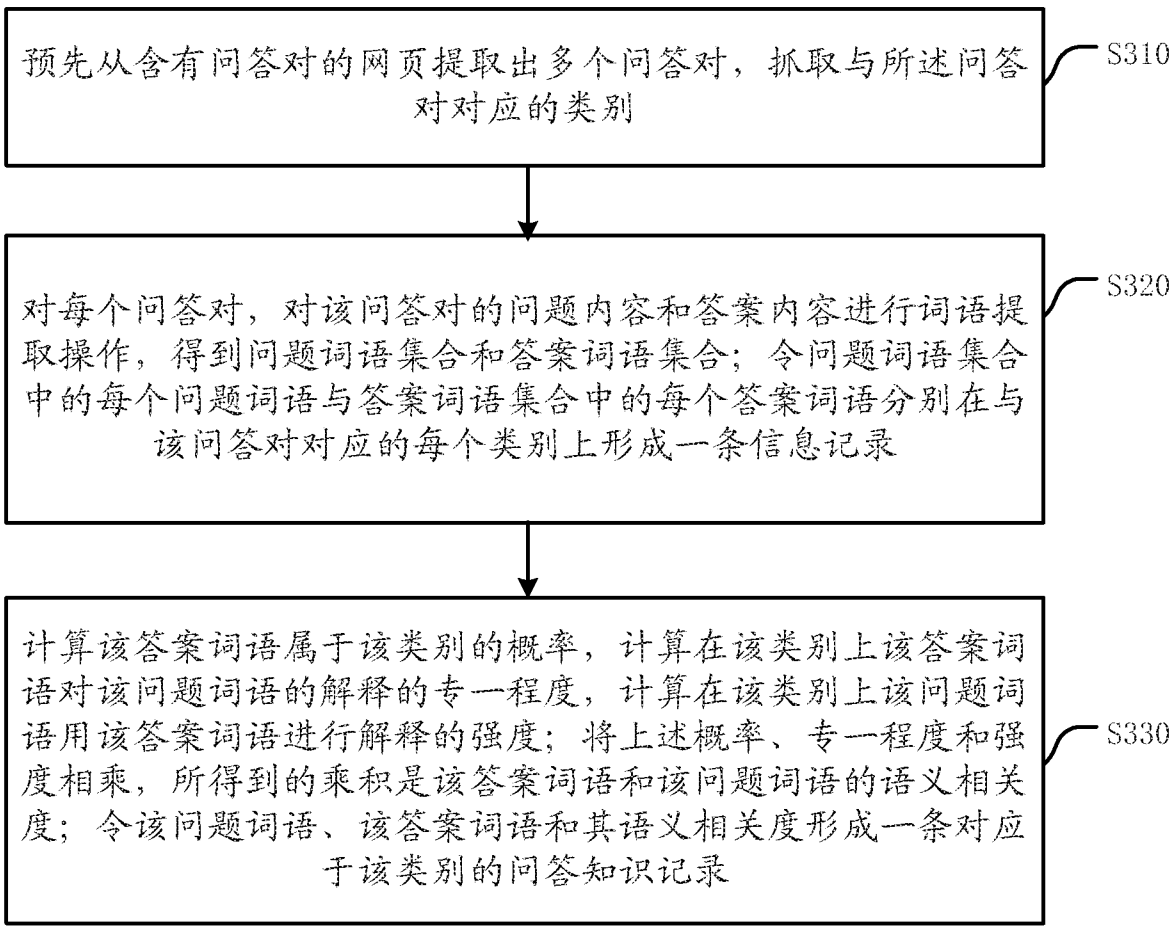

预先从含有问答对的网页提取出多个问答对，抓取与所述问答对对应的类别
S310
对每个问答对，对该问答对的问题内容和答案内容进行词语提取操作，得到问题词语集合和答案词语集合；令问题词语集合中的每个问题词语与答案词语集合中的每个答案词语分别在与该问答对对应的每个类别上形成一条信息记录
S320
计算该答案词语属于该类别的概率，计算在该类别上该答案词语对该问题词语的解释的专一程度，计算在该类别上该问题词语用该答案词语进行解释的强度；将上述概率、专一程度和强度相乘，所得到的乘积是该答案词语和该问题词语的语义相关度；令该问题词语、该答案词语和其语义相关度形成一条对应于该类别的问答知识记录
S330

图 2

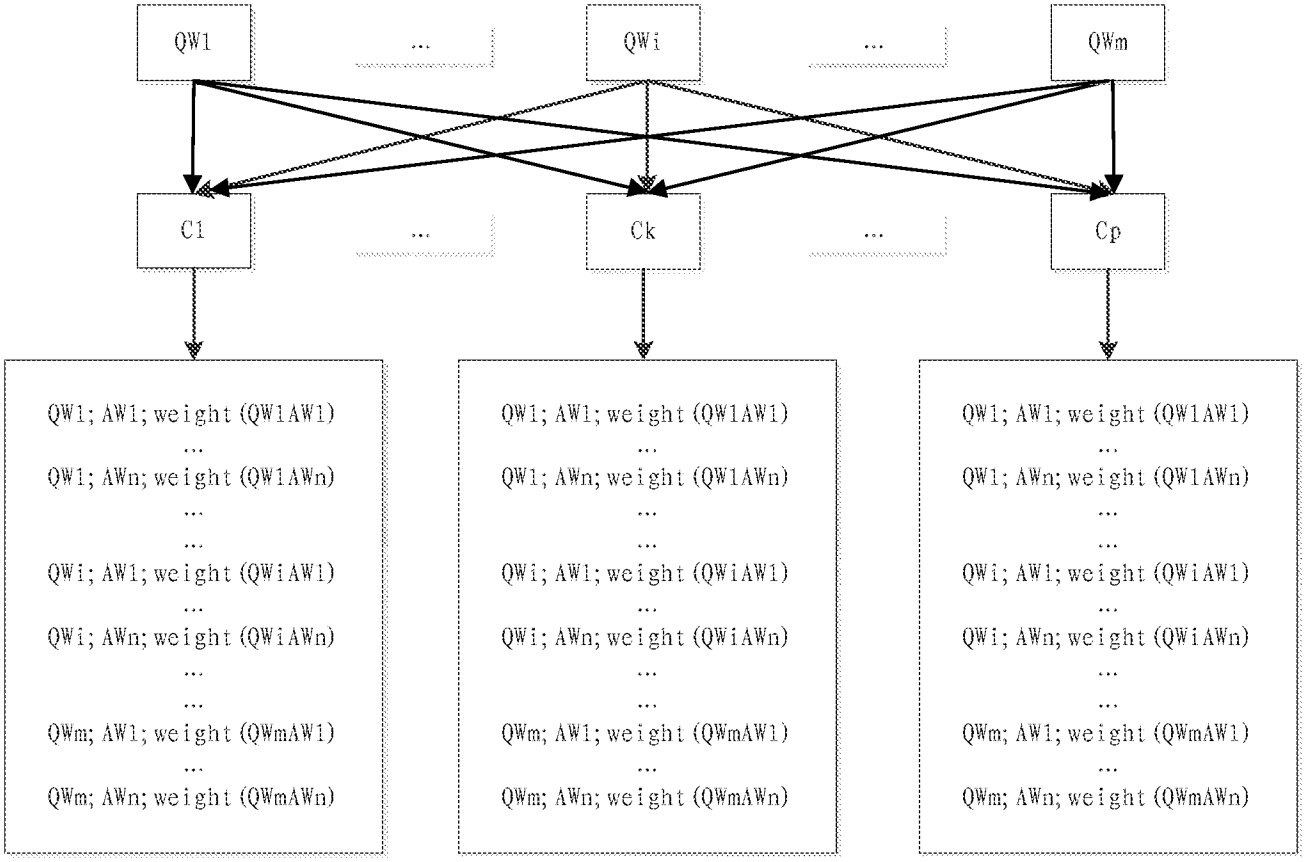

QW1
...
QWi
...
QWm
C1
...
Ck
...
Cp
QW1; AW1; weight (QW1AW1)
...
QW1; AWn; weight (QW1AWn)
...
...
QWi; AW1; weight (QWiAW1)
...
QWi; AWn; weight (QWiAWn)
...
...
QWm; AW1; weight (QWmAW1)
...
QWm; AWn; weight (QWmAWn)
QW1; AW1; weight (QW1AW1)
...
QW1; AWn; weight (QW1AWn)
...
...
QWi; AW1; weight (QWiAW1)
...
QWi; AWn; weight (QWiAWn)
...
...
QWm; AW1; weight (QWmAW1)
...
QWm; AWn; weight (QWmAWn)
QW1; AW1; weight (QW1AW1)
...
QW1; AWn; weight (QW1AWn)
...
...
QWi; AW1; weight (QWiAW1)
...
QWi; AWn; weight (QWiAWn)
...
...
QWm; AW1; weight (QWmAW1)
...
QWm; AWn; weight (QWmAWn)

图 3

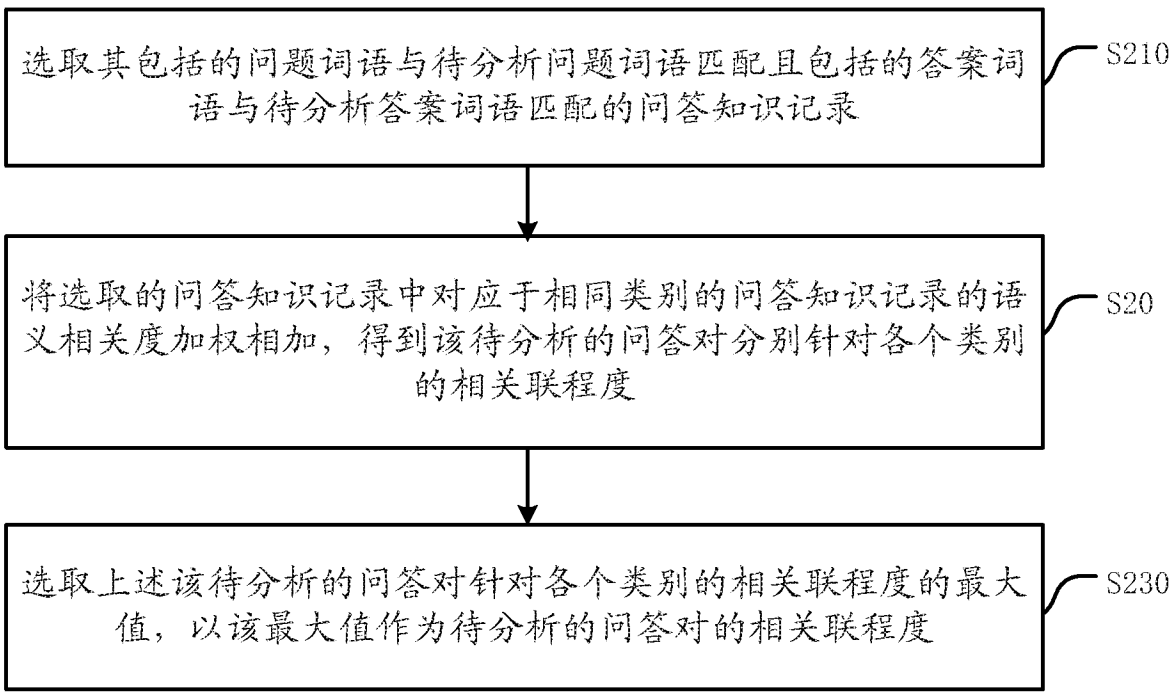

选取其包括的问题词语与待分析问题词语匹配且包括的答案词语与待分析答案词语匹配的问答知识记录
S210
将选取的问答知识记录中对应于相同类别的问答知识记录的语义相关度加权相加，得到该待分析的问答对分别针对各个类别的相关联程度
S20
选取上述该待分析的问答对针对各个类别的相关联程度的最大值，以该最大值作为待分析的问答对的相关联程度
S230

图 4

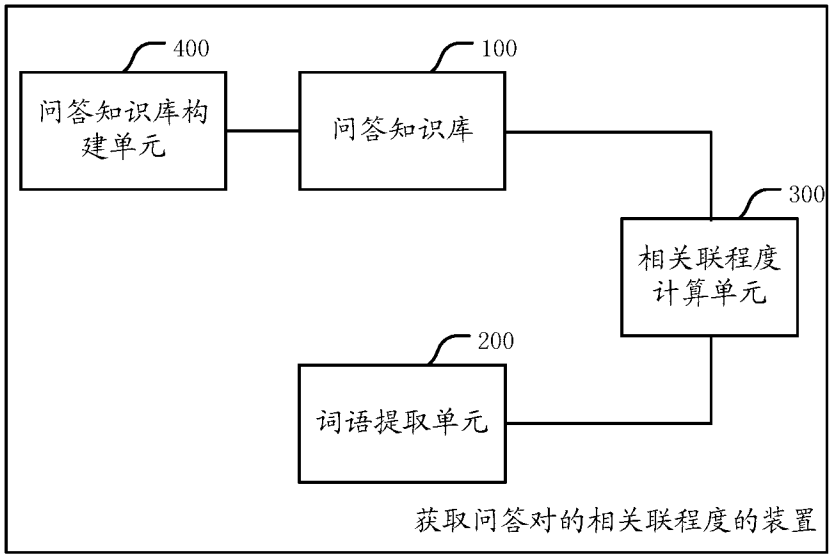

400
问答知识库构建单元
100
问答知识库
300
相关联程度计算单元
200
词语提取单元
获取问答对的相关联程度的装置

图 5

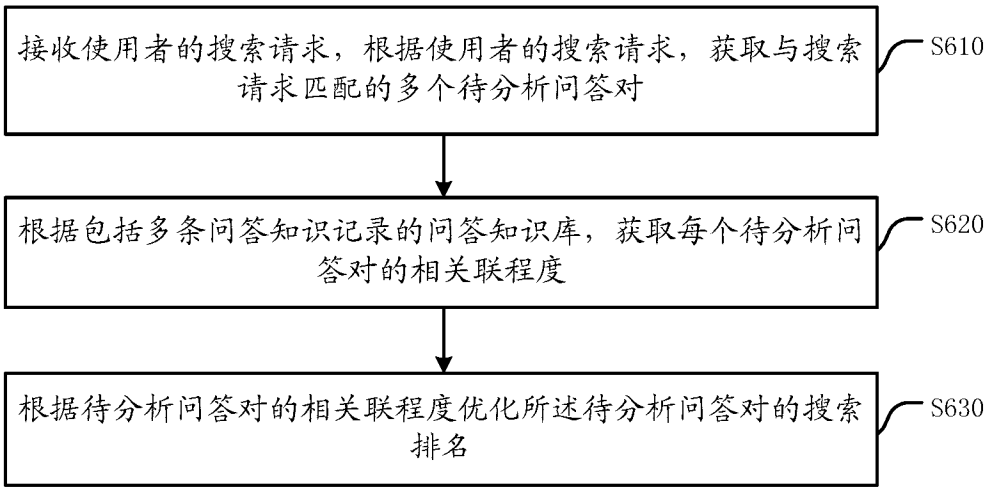

接收使用者的搜索请求，根据使用者的搜索请求，获取与搜索请求匹配的多个待分析问答对
S610
根据包括多条问答知识记录的问答知识库，获取每个待分析问答对的相关联程度
S620
根据待分析问答对的相关联程度优化所述待分析问答对的搜索排名
S630

图 6

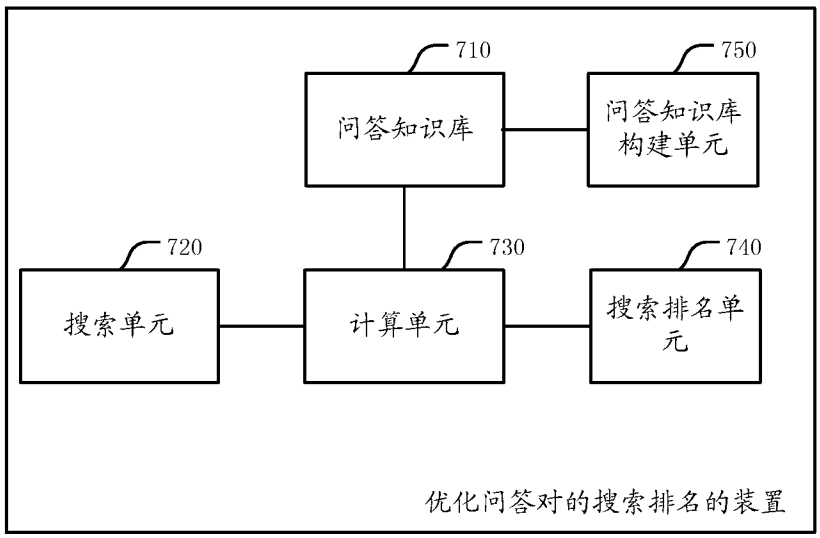

710
问答知识库
750
问答知识库构建单元
720
搜索单元
730
计算单元
740
搜索排名单元
优化问答对的搜索排名的装置

图 7

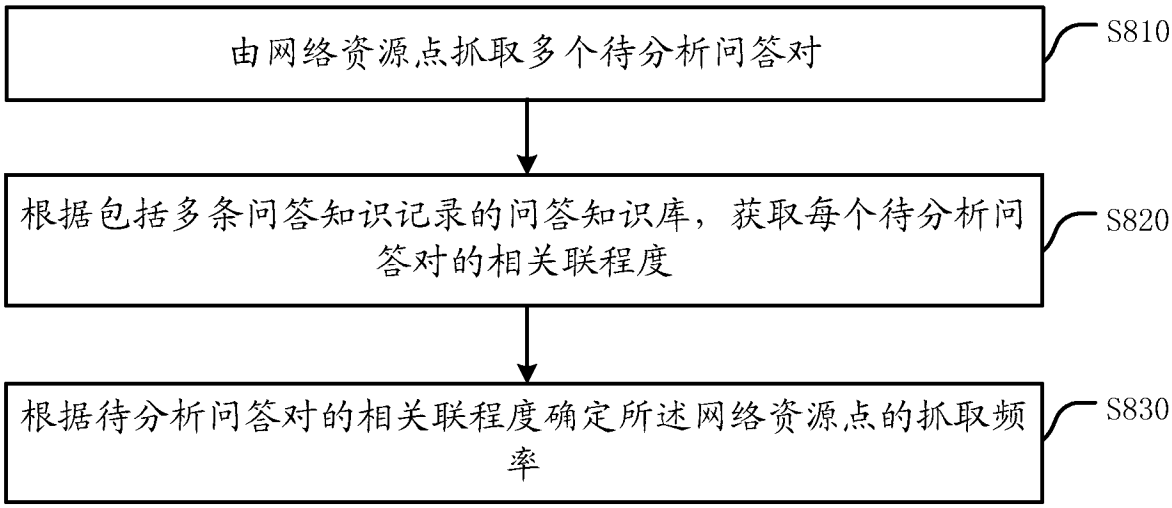

由网络资源点抓取多个待分析问答对
S810
根据包括多条问答知识记录的问答知识库，获取每个待分析问答对的相关联程度
S820
根据待分析问答对的相关联程度确定所述网络资源点的抓取频率
S830

图 8

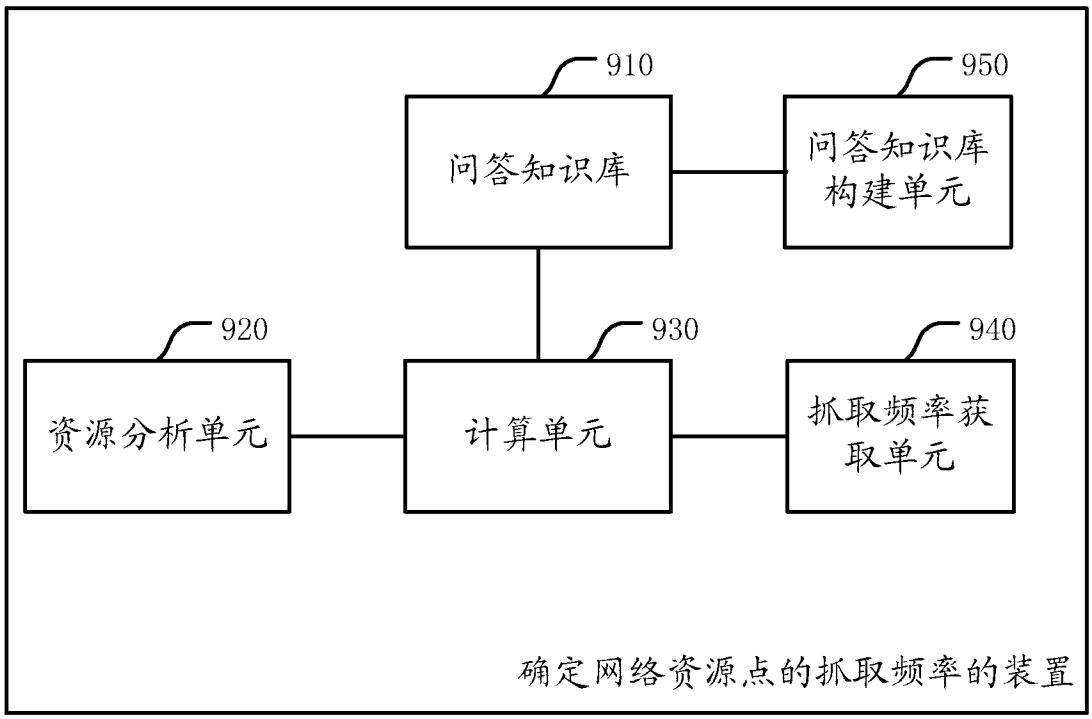

910
问答知识库
950
问答知识库构建单元
920
资源分析单元
930
计算单元
940
抓取频率获取单元
确定网络资源点的抓取频率的装置

图 9

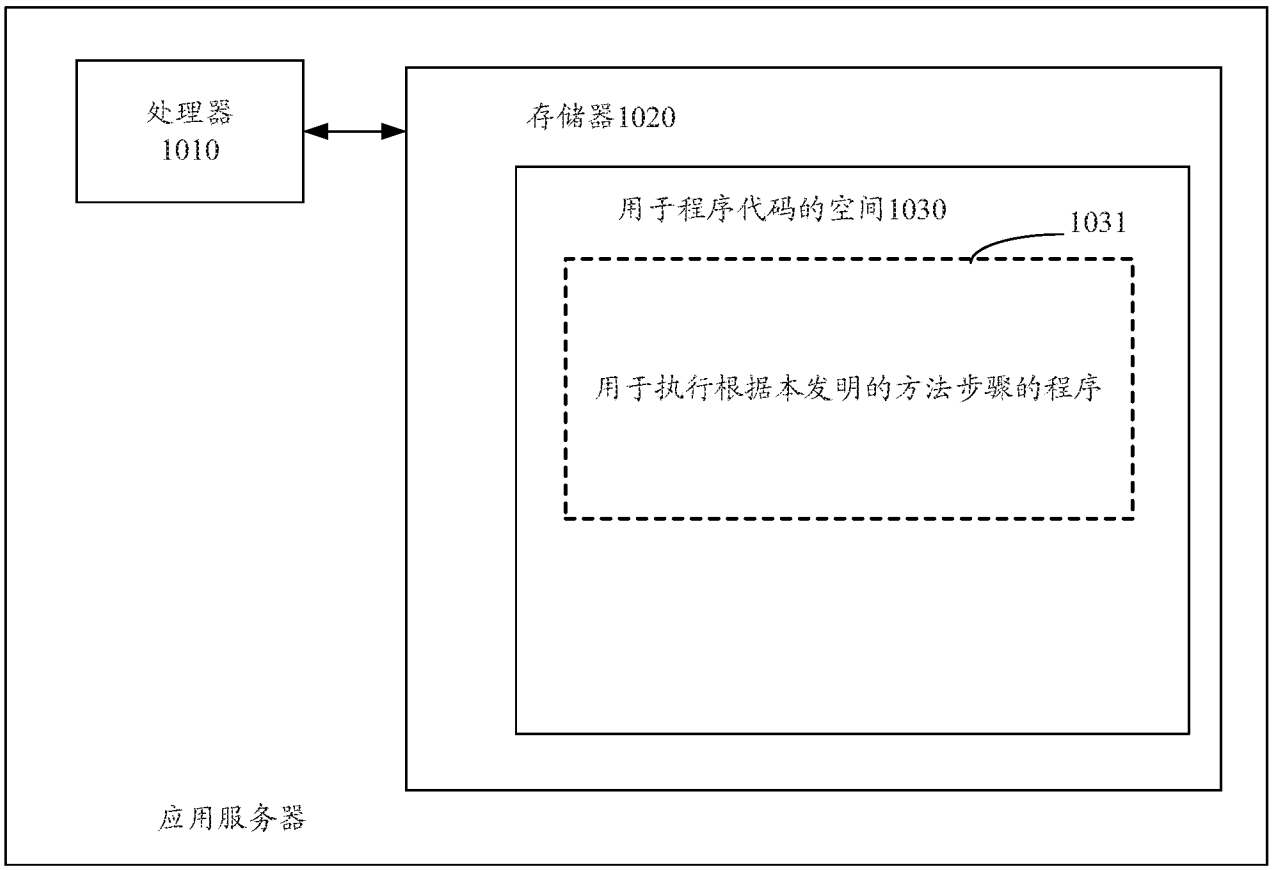

处理器
1010
存储器1020
用于程序代码的空间1030
1031
用于执行根据本发明的方法步骤的程序
应用服务器

图 10

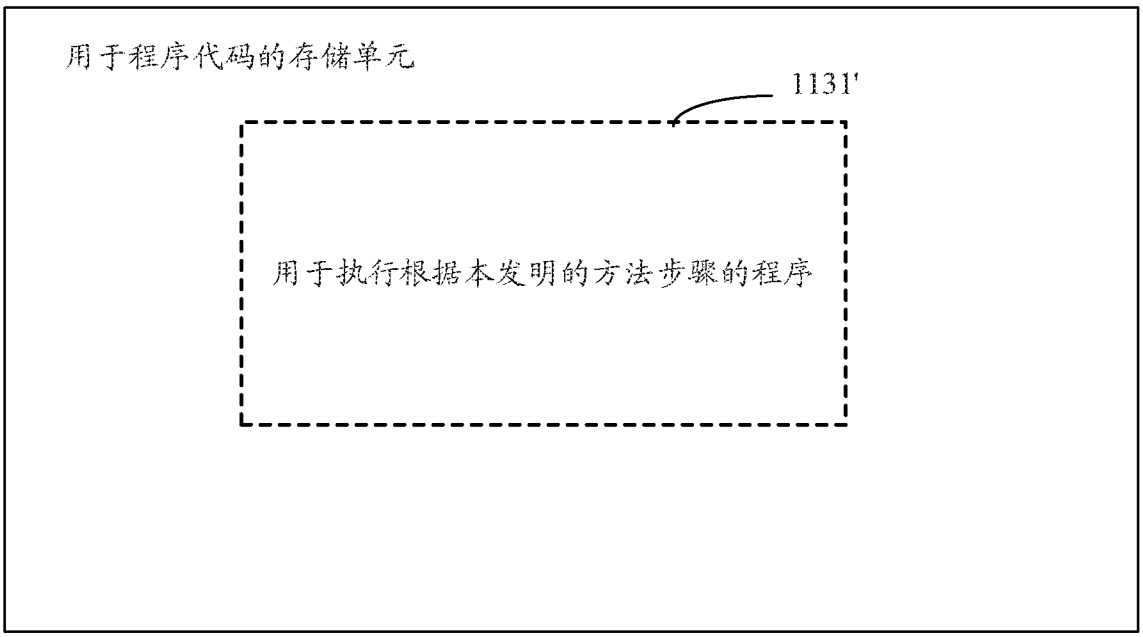

用于程序代码的存储单元
1131'
用于执行根据本发明的方法步骤的程序

图 11

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2014/086838

A. CLASSIFICATION OF SUBJECT MATTER

G06F 15/18 (2006.01) i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F; G06N

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNPAT; WPI; EPODOC: questions and answers, question and answer pair, question and answer matching pair, similarity, relevancy, matching degree, quality, question+, pair?, answer+, value, score, evaluat+

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 8346701 B2 (MICROSOFT CORP.), 01 January 2013 (01.01.2013), description, paragraphs [0030]-[0053] and [0068]-[0069], and figures 2 and 4	1-2, 6, 8-10, 14, 17-19, 23, 26-27, 32-35, 39, 42-44, 48, 51-52
A	US 2012078826 A1 (INT. BUSINESS MACHINES CORP.), 29 March 2012 (29.03.2012), the whole document	1-52
A	CN 101520802 A (TENCENT TECHNOLOGY (SHENZHEN) CO., LTD.), 02 September 2009 (02.09.2009), the whole document	1-52
A	CN 101986293 A (BAIDU ONLINE NETWORK TECHNOLOGY (BEIJING) CO., LTD.), 16 March 2011 (16.03.2011), the whole document	1-52
A	CN 102884527 A (NATIONAL UNIVERSITY OF SINGAPORE), 16 January 2013 (16.01.2013), the whole document	1-52

☐ Further documents are listed in the continuation of Box C. ☒ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
04 December 2014 (04.12.2014)	**26 December 2014 (26.12.2014)**
Name and mailing address of the ISA/CN: State Intellectual Property Office of the P. R. China No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088, China Facsimile No.: (86-10) 62019451	Authorized officer **TANG, Yan** Telephone No.: (86-10) **62413899**

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2014/086838

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
US 8346701 B2	01 January 2013	None	
US 2012078826 A1	29 March 2012	US 2012330882 A1	27 December 2012
CN 101520802 A	02 September 2009	None	
CN 101986293 A	16 March 2011	None	
CN 102884527 A	16 January 2013	SG 184445 A1	29 November 2012
		US 2013024457 A1	24 January 2013
		WO 2011126458 A1	13 October 2011

国际检索报告

国际申请号
PCT/CN2014/086838

A. 主题的分类

G06F 15/18(2006.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06F;G06N

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称，和使用的检索词（如使用）)

CNPAT;WPI;EPODOC:答案，答复，问答，问题，问答对，问答匹配对，相似度，关联度，匹配度，质量，评价，question+，pair?，answer+，value,score,evaluat+

C. 相关文件

类 型*	引用文件，必要时，指明相关段落	相关的权利要求
X	US 8346701 B2 (MICROSOFT CORP.) 2013年 1月 01日 (2013-01-01) 说明书第[0030]-[0053],[0068]-[0069]段，图2，4	1-2，6，8-10，14，17-19，23，26-27，32-35，39，42-44，48，51-52
A	US 2012078826 A1 (INT. BUSINESS MACHINES CORP.) 2012年 3月 29日 (2012-03-29) 全文	1-52
A	CN 101520802 A (腾讯科技深圳有限公司) 2009年 9月 02日 (2009-09-02) 全文	1-52
A	CN 101986293 A (百度在线网络技术北京有限公司) 2011年 3月 16日 (2011-03-16) 全文	1-52
A	CN 102884527 A (新加坡国立大学) 2013年 1月 16日 (2013-01-16) 全文	1-52

☐ 其余文件在C栏的续页中列出。

☑ 见同族专利附件。

* 引用文件的具体类型：

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件，或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件（如具体说明的）

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布，与申请不相抵触，但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件，单独考虑该文件，认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件，当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时，要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期	国际检索报告邮寄日期
2014年 12月 04日	2014年 12月 26日

ISA/CN的名称和邮寄地址	受权官员
中华人民共和国国家知识产权局(ISA/CN) 北京市海淀区蓟门桥西土城路6号 100088 中国 传真号 (86-10)62019451	唐嫣 电话号码 (86-10)010-62413899

表 PCT/ISA/210 (第2页) (2009年7月)

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2014/086838

检索报告引用的专利文件			公布日 (年/月/日)	同族专利			公布日 (年/月/日)
US	8346701	B2	2013年 1月 01日		无		
US	2012078826	A1	2012年 3月 29日	US	2012330882	A1	2012年 12月 27日
CN	101520802	A	2009年 9月 02日		无		
CN	101986293	A	2011年 3月 16日		无		
CN	102884527	A	2013年 1月 16日	SG	184445	A1	2012年 11月 29日
				US	2013024457	A1	2013年 1月 24日
				WO	2011126458	A1	2011年 10月 13日