

(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
2018 年 9 月 20 日 (20.09.2018)



(10) 国际公布号
WO 2018/166115 A1

(51) 国际专利分类号:
G06F 17/30 (2006.01) **G06F 17/27** (2006.01)

(21) 国际申请号: PCT/CN2017/091372

(22) 国际申请日: 2017 年 6 月 30 日 (30.06.2017)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201710147702.6 2017 年 3 月 13 日 (13.03.2017) CN

(71) 申请人: 平安科技 (深圳) 有限公司 (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) [CN/CN]; 中国广东省深圳福田区八卦岭八卦三路平安大厦吴东勤, Guangdong 518000 (CN)。

(72) 发明人: 王健宗 (WANG, Jianzong); 中国广东省深圳福田区八卦岭八卦三路平安大厦吴东勤, Guangdong 518000 (CN)。袁威强 (YUAN, Weiqiang); 中国广东省深圳福田区八卦岭八卦三路平安大厦吴东勤, Guangdong 518000

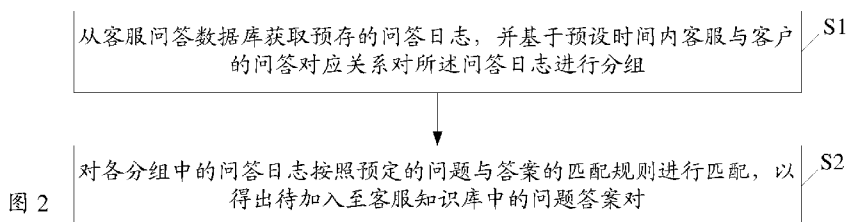
(CN)。韩茂琨 (HAN, Maokun); 中国广东省深圳福田区八卦岭八卦三路平安大厦吴东勤, Guangdong 518000 (CN)。肖京 (XIAO, Jing); 中国广东省深圳福田区八卦岭八卦三路平安大厦吴东勤, Guangdong 518000 (CN)。

(74) 代理人: 深圳市沃德知识产权代理事务所 (普通合伙) (SHENZHEN WORLD INTELLECTUAL PROPERTY AGENCY (GENERAL PARTNERSHIP)); 中国广东省深圳福田区园岭街道八卦四路 10 号中浩大厦 1528-1530 室于志光, Guangdong 518000 (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, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, 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,

(54) Title: METHOD FOR PROCESSING CUSTOMER SERVICE QUESTION-ANSWER DATA, ELECTRONIC DEVICE, AND COMPUTER READABLE STORAGE MEDIUM

(54) 发明名称: 客服问答数据的处理方法、电子装置及计算机可读存储介质



S1 OBTAIN PRE-STORED QUESTION-ANSWER LOGS FROM A CUSTOMER SERVICE QUESTION-ANSWER DATABASE AND GROUP THE QUESTION-ANSWER LOGS ON THE BASIS OF THE QUESTION-ANSWER CORRESPONDENCE BETWEEN CUSTOMER SERVICE STAFFS AND CUSTOMERS WITHIN A PRESET TIME PERIOD
S2 MATCH THE QUESTION-ANSWER LOGS IN EACH GROUP ACCORDING TO A PREDETERMINED QUESTION-ANSWER MATCHING RULE TO OBTAIN QUESTION-ANSWER PAIRS TO BE ADDED INTO A CUSTOMER SERVICE KNOWLEDGE LIBRARY

(57) Abstract: A method for processing customer service question-answer data, an electronic device, and a computer readable storage medium. The method for processing customer service question-answer data comprises: obtaining pre-stored question-answer logs from a customer service question-answer database and grouping the question-answer logs on the basis of the question-answer correspondence between customer service staffs and customers within a preset time period (S1); and matching the question-answer logs in each group according to a predetermined question-answer matching rule to obtain question-answer pairs to be added into a customer service knowledge base (S2). The question-answer pairs can be obtained by automatically matching on the basis of question-answer logs in the customer service question-answer database without compilation of the customer service staffs; the coverage of the question-answer pairs is wide, the construction efficiency and the update rate are high.

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)。

本国际公布:

- 包括国际检索报告 (条约第21条(3))。

(57) 摘要: 一种客服问答数据的处理方法、电子装置及存储介质, 所述客服问答数据的处理方法包括: 从客服问答数据库获取预存的问答日志, 并基于预设时间内客服与客户的问答对应关系对所述问答日志进行分组 (S1); 对各分组中的问答日志按照预定的问题与答案的匹配规则进行匹配, 以得出待加入至客服知识库中的问题答案对 (S2)。上述基于客服问答数据库的问答日志自动匹配得出问题答案对, 不需要客服人员进行编写, 问题答案对的覆盖面广、构建效率高且更新速度快。

客服问答数据的处理方法、电子装置及计算机可读存储介质

优先权申明

5 本申请基于巴黎公约申明享有 2017 年 3 月 13 日递交的申请号为 CN201710147702.6、名称为“客服问答数据的处理方法及系统”中国专利申请

的优先权，该中国专利申请的整体内容以参考的方式结合在本申请中。

技术领域

10 本发明涉及计算机软件技术领域，尤其涉及一种客服问答数据的处理方法、电子装置及计算机可读存储介质。

背景技术

15 目前，客服自动问答系统一般是基于问题答案对作为知识库进行设计的，根据用户的问题检索到知识库中最相似的问题，再根据知识库中已有的问题答案对（即问题与答案的映射对），直接给出回答。在这种系统设计框架下，有三个主要核心：问题理解、检索算法以及知识库构建。对于知识库，传统的构建方法一般是由具备多年专业经验的客服人员人工进行编写，这种构建知识库的方式存在的不足之处为：问题答案对的覆盖面窄、构建效率低且更新速度慢。

发明内容

20 本发明的目的在于提供一种客服问答数据的处理方法、电子装置及计算机可读存储介质，旨在基于客服问答数据库的问答日志自动匹配得出问题答案对，且问题答案对的覆盖面广、构建效率高且更新速度快。

25 本发明第一方面提供一种客服问答数据的处理方法，所述客服问答数据的处理方法包括：

S1，从客服问答数据库获取预存的问答日志，并基于预设时间内客服与客户的问答对应关系对所述问答日志进行分组；

30 S2，对各分组中的问答日志按照预定的问题与答案的匹配规则进行匹配，以得出待加入至客服知识库中的问题答案对。

优选地，所述步骤 S1 包括：

将预设时间内的同一客户与同一客服的问答日志分为同一分组；或

将预设时间内的同一客户与各客服的问答日志分为同一分组；或

将预设时间内的各客户与同一客服的问答日志分为同一分组。

35 本发明第二方面提供一种电子装置，包括处理设备、存储设备及客服问答数据的处理系统，该客服问答数据的处理系统存储于该存储设备中，包括至少一个计算机可读指令，该至少一个计算机可读指令可被所述处理设备执行，以实现以下操作：

S1，从客服问答数据库获取预存的问答日志，并基于预设时间内客服与

客户的问答对应关系对所述问答日志进行分组；

S2，对各分组中的问答日志按照预定的问题与答案的匹配规则进行匹配，以得出待加入至客服知识库中的问题答案对。

5 本发明第三方面提供一种计算机可读存储介质，其上存储有至少一个可被处理设备执行以实现以下操作的计算机可读指令：

S1，从客服问答数据库获取预存的问答日志，并基于预设时间内客服与客户的问答对应关系对所述问答日志进行分组；

S2，对各分组中的问答日志按照预定的问题与答案的匹配规则进行匹配，以得出待加入至客服知识库中的问题答案对。

10 本发明的有益效果是：本发明在构建客服知识库时，由客服问答数据的处理系统自动从客服问答数据库中获取问答日志，对问答日志进行分组，使得同一分组中的问答日志具有某些相同的属性，对于每一分组中具有某些相同属性的问答日志按照预定的问题与答案的匹配规则进行匹配，以匹配得出待加入至客服知识库问题答案对，本发明基于客服问答数据库的问答日志自
15 动匹配得出问题答案对，不需要客服人员进行编写，问题答案对的覆盖面广、构建效率高且更新速度快。

附图说明

20 图 1 是本发明实现客服问答数据的处理方法的一实施例的应用环境示意图；

图 2 为本发明客服问答数据的处理方法一实施例的流程示意图；

图 3 为图 2 所示步骤 S2 的细化流程示意图；

图 4 为本发明客服问答数据的处理系统一实施例的结构示意图；

图 5 为图 4 所示匹配模块的结构示意图。

25

具体实施方式

以下结合附图对本发明的原理和特征进行描述，所举实例只用于解释本发明，并非用于限定本发明的范围。

30 参阅图 1 所示，是本发明实现客服问答数据的处理方法的一实施例的应用环境示意图。所述应用环境示意图包括电子装置 1、终端设备 2。所述电子装置 1 可以通过网络、近场通信技术等技术与所述终端设备 2 进行数据交互。

35 终端设备 2 包括，但不限于，任何一种可与用户通过键盘、鼠标、遥控器、触摸板或者声控设备等方式进行人机交互的电子产品，例如，个人计算机、平板电脑、智能手机、个人数字助理（Personal Digital Assistant, PDA）、游戏机、交互式网络电视（Internet Protocol Television, IPTV）、智能式穿戴设备等。

电子装置 1 是一种能够按照事先设定或者存储的指令，自动进行数值计算和/或信息处理的设备。电子装置 1 可以是计算机、也可以是单个网络服务

器、多个网络服务器组成的服务器组或者基于云计算的由大量主机或者网络服务器构成的云，其中云计算是分布式计算的一种，由一群松散耦合的计算机集组成的一个超级虚拟计算机。

5 在本实施例中，电子装置 1 包括，但不限于，可通过系统总线相互通信连接的存储设备 11、处理设备 12、及网络接口 13。需要指出的是，图 1 仅示出了具有组件 11-13 的电子装置 1，但是应理解的是，并不要求实施所有示出的组件，可以替代的实施更多或者更少的组件。

其中，存储设备 11 包括内存及至少一种类型的可读存储介质。内存为电子装置 1 的运行提供缓存；可读存储介质可为如闪存、硬盘、多媒体卡、卡型存储器等的非易失性存储介质。在一些实施例中，可读存储介质可以是电子装置 1 的内部存储单元，例如该电子装置 1 的硬盘；在另一些实施例中，该非易失性存储介质也可以是电子装置 1 的外部存储设备，例如电子装置 1 上配备的插接式硬盘，智能存储卡（Smart Media Card, SMC），安全数字（Secure Digital, SD）卡，闪存卡（Flash Card）等。本实施例中，存储设备 11 的可读存储介质通常用于存储安装于电子装置 1 的操作系统和各类应用软件，例如本申请一实施例中的客服问答数据的处理系统 10 的程序代码等。此外，存储设备 11 还可以用于暂时地存储已经输出或者将要输出的各类数据。

20 处理设备 12 在一些实施例中可以包括一个或者多个微处理器、微控制器、数字处理器等。该处理设备 12 通常用于控制电子装置 1 的运行，例如执行与终端设备 2 进行数据交互或者通信相关的控制和处理等。在本实施例中，处理设备 12 用于运行存储设备 11 中存储的程序代码或者处理数据，例如运行客服问答数据的处理系统 10 等。

25 网络接口 13 可包括无线网络接口或有线网络接口，该网络接口 13 通常用于在电子装置 1 与其他电子设备之间建立通信连接。本实施例中，网络接口 13 主要用于将电子装置 1 与一个或多个终端设备 2 相连，在电子装置 1 与一个或多个终端设备 2 之间建立数据传输通道和通信连接。

30 客服问答数据的处理系统 10 包括至少一个存储在存储设备 11 中的计算机可读指令，该至少一个计算机可读指令可被处理设备 12 执行，以实现本申请各实施例的控件响应区域显示控制方法。如后续所述，该至少一个计算机可读指令依据其各部分所实现的功能不同，可被划为不同的逻辑模块。

35 在一实施例中，客服问答数据的处理系统 10 被处理设备 12 执行时，实现以下操作：从客服问答数据库获取预存的问答日志，并基于预设时间内客服与客户的问题对应关系对所述问答日志进行分组；对各分组中的问答日志按照预定的问题与答案的匹配规则进行匹配，以得出待加入至客服知识库中的问题答案对。

如图 2 所示，图 2 为本发明客服问答数据的处理方法一实施例的流程示意图，该客服问答数据的处理方法包括以下步骤：

步骤 S1, 从客服问答数据库获取预存的问答日志, 并基于预设时间内客服与客户的问答对应关系对所述问答日志进行分组;

本实施例中, 客服问答数据库中存储有所有的人工客服与客户对话过程中产生的问答日志, 客服问答数据的处理系统与客服问答数据库通信连接, 能够自动从客服问答数据库中获取问答日志, 以进行进一步的处理。

本实施例中, 可以实时地从客服问答数据库中获取问答日志, 例如实时地获取新增的客服与客户的问答日志; 也可以定时地从客服问答数据库中获取问答日志, 例如每隔 6 小时从客服问答数据库中获取问答日志。通过实时或定时地获取客服问答数据库中的问答日志, 以获取最新的问答日志。

然后, 对获取的问答日志基于预设时间内客服与客户的问答对应关系进行分组, 例如, 一天内某一客服当天内与 k 个客户进行问答对话, 则将该客服当天内与 k 个客户进行问答对话对应的问答日志分为一组。对问答日志进行分组的主要目的在于, 将具有共同属性的问答日志归为一组, 以方便后续进行处理分析。

步骤 S2, 对各分组中的问答日志按照预定的问题与答案的匹配规则进行匹配, 以得出待加入至客服知识库中的问题答案对。

本实施例中, 对各分组中的问答日志按照预定的问题与答案的匹配规则进行匹配, 例如按照提问的时间及回答的时间的先后顺序先提取出问答日志中的问题语句及回答语句, 分析问题语句中的长度及所包括的业务关键词等, 分析回答语句的长度及业务关键词是否与问题语句中的业务关键词对应等等, 根据上述的分析结果将问题语句及回答语句进行配对; 又如, 按照提问的时间及回答的时间的先后顺序先提取出问答日志中的问题语句及回答语句, 分析问题语句中的业务关键词及对业务关键词进行扩展, 分析回答语句是否与该业务关键词对应或者与该业务关键词进行扩展后的词语对应, 根据上述的分析结果将问题语句及回答语句进行配对, 等等。当然, 本实施例按照预定的问题与答案的匹配规则进行匹配不限定上述的这两种方式, 其他能够将问题语句及回答语句进行匹配以完成配对的方式也在本实施例的保护范围内。

本实施例中, 在将问题语句与回答语句进行匹配以配对出问题答案对之后, 将问题答案对加入至客服知识库中, 以不断地增加或丰富客服知识库中的知识。

与现有技术相比, 本实施例在构建客服知识库时, 由客服问答数据的处理系统自动从客服问答数据库中获取问答日志, 对问答日志进行分组, 使得同一分组中的问答日志具有某些相同的属性, 对于每一分组中具有某些相同属性的问答日志按照预定的问题与答案的匹配规则进行匹配, 以匹配得出待加入至客服知识库问题答案对, 本实施例基于客服问答数据库的问答日志自动匹配得出问题答案对, 不需要客服人员进行编写, 问题答案对的覆盖面广、构建效率高且更新速度快。

在一优选的实施例中，在上述图 2 的实施例的基础上，上述步骤 S1 包括：将同一预设时间内的同一客户与同一客服的问答日志分为同一分组，例如以预设时间（例如一天或者一周）为划分依据，将同一天内的同一客户与同一客服的问答日志分为同一分组，例如问答日志包括：客户 A 与客服 a 的问答日志 1、客户 A 与客服 a 的问答日志 2、客户 A 与客服 b 的问答日志 3、/客服 B 与客户 a 的问答日志 4、客服 B 与客户 b 的问答日志 5，则将问答日志 1 及 2 分为一组，其他的问答日志 3、4 及 5 各自为一组；或

将同一预设时间内的同一客户与各客服的问答日志分为同一分组，例如将上述的问答日志 1、2 及 3 分为一组，将问答日志 4 及 5 分为一组；或

将预设时间内的各客户与同一客服的问答日志分为同一分组，例如将上述的问答日志 1、2 及 4 分为一组，将问答日志 3 及 5 分为一组。当然本实施例不限于上述的分组方式。

本实施例经过分组后的问答日志具有某些相同的属性，例如同一客户可能提问相似的问题，其提问语句具有相似的特点，或者同一客服可能回答类似的问题，其回答语句具有相似的特点，具有相同的属性的问答日志能够为后续进一步的分析提供便利。

在一优选的实施例中，如图 3 所示，在上述图 2 的实施例的基础上，上述步骤 S2 包括：

S21，对每一分组中的问答日志进行数据预处理；

S22，对数据预处理后的问答日志中的客户的问题语句和客服的回答语句按照预设的分析规则进行分析，以分析出待匹配的问题语句与对应的回答语句；

S23，抽取待匹配的问题语句与对应的回答语句，将抽取出的问题语句与对应的回答语句关联为问题答案对并将问题答案对加入至客服知识库中。

本实施例中，对每一分组中的问答日志进行数据预处理，优选地，进行数据预处理包括：将每一分组中的问答日志中的字体统一转换为繁体字或者简体字、将每一分组中的问答日志中的字母统一转换为大写字母或者小写字母、去除问答日志中的特定语句，以及对每一分组中的问答日志中的语句进行分词处理等等。特定语句包括重复的语句、常用寒暄语（例如经常使用的一些礼貌用语“你好”、“谢谢”等）；对问答日志中的语句进行分词处理可以使用以下三大类的分词方法执行：基于字符串匹配的分词方法、基于理解的分词方法和基于统计的分词方法。

然后，对数据预处理后的问答日志中的客户的问题语句和客服的回答语句按照预设的分析规则进行分析，优选地，预设的分析规则包括问题语句的分析规则及回答语句的分析规则，所述问题语句的分析规则为：

计算问题语句的关键指标值 S：

$$S=S_a+S_b+S_c, \quad S_a=\begin{cases} -\frac{a-l}{a}, l < a \\ 0, a \leq l \leq b \\ -\frac{l-b}{l}, l > b \end{cases}, \quad S_b=\frac{m}{l}, \quad S_c=-\frac{n}{l}, \quad \text{其中, } l \text{ 为问题语句的}$$

长度, 长度 l 可以以分词后的词语个数进行度量, 即词语个数越多则长度 l 越大, m 为问题语句所含的业务关键词的数目, m 越大则关键指标值 S 越大, n 为问题语句的停用词的数目, n 越大则关键指标值 S 越小, 问题质量得分越低, a 为问题语句的第一长度预设值, b 为问题语句的第二长度预设值, 长度 l 在 $[a, b]$ 之内时, 该问题语句的长度适中, 问题质量得分高; 长度 l 在 $[a, b]$ 之外时, 相差越多, 问题质量得分越低。

若关键指标值 S 大于预设指标阈值, 例如 $S > 0$, 则保留所述问题语句作为待匹配的问题语句, 或者, 若关键指标值 S 小于或者等于预设指标阈值, 例如 $S < 0$, 则去除所述问题语句。

优选地, 回答语句的分析规则为: 若回答语句不是陈述性语句, 则去除该回答语句及其对应的问题语句。

另外, 在分析出待匹配的问题语句与对应的回答语句后, 在抽取待匹配的问题语句与对应的回答语句时, 如果有重复的问题语句与对应的回答语句, 则仅抽取重复的问题语句与对应的回答语句中的一个问题语句与对应的回答语句即可; 以及, 将待匹配的问题语句与对应的回答语句按照其所包含的关键业务词权重由大至小进行排列, 抽取队列中关键业务词权重大于某一预设阈值的待匹配的问题语句与对应的回答语句, 将所抽取出来的问题语句与对应的回答语句关联为问题答案对, 以将问题答案对加入至客服知识库中。

如图 4 所示, 图 4 为本发明客服问答数据的处理系统 10 一实施例的结构示意图, 该客服问答数据的处理系统 10 包括:

分组模块 101, 用于从客服问答数据库获取预存的问答日志, 并基于预设时间内客服与客户的问答对应关系对所述问答日志进行分组;

本实施例中, 客服问答数据库中存储有所有的人工客服与客户对话过程中产生的问答日志, 客服问答数据的处理系统 10 与客服问答数据库通信连接, 客服问答数据的处理系统 10 能够自动从客服问答数据库中获取问答日志, 以进行进一步的处理。

本实施例中, 客服问答数据的处理系统 10 可以实时地从客服问答数据库中获取问答日志, 例如实时地获取新增的客服与客户的问答日志; 也可以定时地从客服问答数据库中获取问答日志, 例如每隔 6 小时从客服问答数据库中获取问答日志。通过实时或定时地获取客服问答数据库中的问答日志, 以获取最新的问答日志。

然后, 对获取的问答日志基于预设时间内客服与客户的问答对应关系进

行分组，例如，一天内某一客服当天内与 k 个客户进行问答对话，则将该客服当天内与 k 个客户进行问答对话对应的问答日志分为一组。对问答日志进行分组的主要目的在于，将具有共同属性的问答日志归为一组，以方便后续进行处理分析。

5 匹配模块 102，用于对各分组中的问答日志按照预定的问题与答案的匹配规则进行匹配，以得出待加入至客服知识库中的问题答案对。

本实施例中，对各分组中的问答日志按照预定的问题与答案的匹配规则进行匹配，例如按照提问的时间及回答的时间的先后顺序先提取出问答日志中的问题语句及回答语句，分析问题语句中的长度及所包括的业务关键词等，分析回答语句的长度及业务关键词是否与问题语句中的业务关键词对应等等，根据上述的分析结果将问题语句及回答语句进行配对；又如，按照提问的时间及回答的时间的先后顺序先提取出问答日志中的问题语句及回答语句，分析问题语句中的业务关键词及对业务关键词进行扩展，分析回答语句是否与该业务关键词对应或者与该业务关键词进行扩展后的词语对应，根据上述的分析结果将问题语句及回答语句进行配对，等等。当然，本实施例按照预定的问题与答案的匹配规则进行匹配不限定上述的这两种方式，其他能够将问题语句及回答语句进行匹配以完成配对的方式也在本实施例的保护范围内。

本实施例中，在将问题语句与回答语句进行匹配以配对出问题答案对之后，将问题答案对加入至客服知识库中，以不断地增加或丰富客服知识库中的知识。

在一优选的实施例中，在上述图 4 的实施例的基础上，所述分组模块 101 具体用于将同一预设时间内的同一客户与同一客服的问答日志分为同一分组，例如以预设时间（例如一天或者一周）为划分依据，将同一天内的同一客户与同一客服的问答日志分为同一分组，例如问答日志包括：客户 A 与客服 a 的问答日志 1、客户 A 与客服 a 的问答日志 2、客户 A 与客服 b 的问答日志 3、/客服 B 与客户 a 的问答日志 4、客服 B 与客户 b 的问答日志 5，则将问答日志 1 及 2 分为一组，其他的问答日志 3、4 及 5 各自为一组；或

30 将同一预设时间内的同一客户与各客服的问答日志分为同一分组，例如将上述的问答日志 1、2 及 3 分为一组，将问答日志 4 及 5 分为一组；或

将预设时间内的各客户与同一客服的问答日志分为同一分组，例如将上述的问答日志 1、2 及 4 分为一组，将问答日志 3 及 5 分为一组。当然本实施例不限定于上述的分组方式。

35 本实施例经过分组后的问答日志具有某些相同的属性，例如同一客户可能提问相似的问题，其提问语句具有相似的特点，或者同一客服可能回答类似的问题，其回答语句具有相似的特点，具有相同的属性的问答日志能够为后续进一步的分析提供便利。

在一优选的实施例中,如图5所示,在上述图4的实施例的基础上,匹配模块102包括:

预处理单元1021,用于对每一分组中的问答日志进行数据预处理;

5 分析单元1022,用于对数据预处理后的问答日志中的客户的问题语句和客服的回答语句按照预设的分析规则进行分析,以分析出待匹配的问题语句与对应的回答语句;

关联单元1023,用于抽取待匹配的问题语句与对应的回答语句,将抽取出的问题语句与对应的回答语句关联为问题答案对并将问题答案对加入至客服知识库中。

10 本实施例中,对每一分组中的问答日志进行数据预处理,优选地,预处理单元1021具体用于将每一分组中的问答日志中的字体统一转换为繁体字或者简体字、将每一分组中的问答日志中的字母统一转换为大写字母或者小写字母、去除问答日志中的特定语句,以及对每一分组中的问答日志中的语句进行分词处理等等。特定语句包括重复的语句、常用寒暄语(例如经常使用的一些礼貌用语“你好”、“谢谢”等);对问答日志中的语句进行分词处理可以使用以下三大类的分词方法执行:基于字符串匹配的分词方法、基于理解的分词方法和基于统计的分词方法。

20 然后,对数据预处理后的问答日志中的客户的问题语句和客服的回答语句按照预设的分析规则进行分析,优选地,预设的分析规则包括问题语句的分析规则及回答语句的分析规则,所述问题语句的分析规则为:

计算问题语句的关键指标值S:

$$S=S_a+S_b+S_c, \quad S_a=\begin{cases} -\frac{a-l}{a}, & l < a \\ 0, & a \leq l \leq b \\ -\frac{l-b}{l}, & l > b \end{cases}, \quad S_b=\frac{m}{l}, \quad S_c=-\frac{n}{l}, \quad \text{其中, } l \text{ 为问题语句的}$$

25 长度,长度 l 可以以分词后的词语个数进行度量,即词语个数越多则长度 l 越大, m 为问题语句所含的业务关键词的数目, m 越大则关键指标值 S 越大, n 为问题语句的停用词的数目, n 越大则关键指标值 S 越小,问题质量得分越低, a 为问题语句的第一长度预设值, b 为问题语句的第二长度预设值,长度 l 在 $[a, b]$ 之内时,该问题语句的长度适中,问题质量得分高;长度 l 在 $[a, b]$ 之外时,相差越多,问题质量得分越低。

30 若关键指标值 S 大于预设指标阈值,例如 $S>0$,则保留所述问题语句作为待匹配的问题语句,或者,若关键指标值 S 小于或者等于预设指标阈值,例如 $S<0$,则去除所述问题语句。

优选地,回答语句的分析规则为:若回答语句不是陈述性语句,则去除该回答语句及其对应的问题语句。

35 另外,在分析出待匹配的问题语句与对应的回答语句后,在抽取待匹配的问题语句与对应的回答语句时,如果有重复的问题语句与对应的回答语

句，则仅抽取重复的问题语句与对应的回答语句中的一个问题语句与对应的回答语句即可；以及，将待匹配的问题语句与对应的回答语句按照其所包含的关键业务词权重由大至小进行排列，抽取队列中关键业务词权重大于某一预设阈值的待匹配的问题语句与对应的回答语句，将所抽取出来的问题语句与对应的回答语句关联为问题答案对，以将问题答案对加入至客服知识库中。

以上所述仅为本发明的较佳实施例，并不用以限制本发明，凡在本发明的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本发明的保护范围之内。

权利要求书

1. 一种客服问答数据的处理方法, 其特征在于, 所述客服问答数据的处理方法包括:

5 S1, 从客服问答数据库获取预存的问答日志, 并基于预设时间内客服与客户的问答对应关系对所述问答日志进行分组;

S2, 对各分组中的问答日志按照预定的问题与答案的匹配规则进行匹配, 以得出待加入至客服知识库中的问题答案对。

10 2. 根据权利要求 1 所述的客服问答数据的处理方法, 其特征在于, 所述步骤 S2 包括:

S21, 对每一分组中的问答日志进行数据预处理;

S22, 对数据预处理后的问答日志中的客户的问题语句和客服的回答语句按照预设的分析规则进行分析, 以分析出待匹配的问题语句与对应的回答语句;

15 S23, 抽取待匹配的问题语句与对应的回答语句, 将抽取出的问题语句与对应的回答语句关联为问题答案对并将问题答案对加入至客服知识库中。

3. 根据权利要求 2 所述的客服问答数据的处理方法, 其特征在于, 所述步骤 S21 包括:

20 将每一分组中的问答日志中的字体统一转换为繁体字或者简体字、将每一分组中的问答日志中的字母统一转换为大写字母或者小写字母、去除问答日志中的特定语句, 以及对每一分组中的问答日志中的语句进行分词处理。

4. 根据权利要求 2 所述的客服问答数据的处理方法, 其特征在于, 所述预设的分析规则包括问题语句的分析规则及回答语句的分析规则, 所述问题语句的分析规则为:

25 计算问题语句的关键指标值 S:

$$S = S_a + S_b + S_c, \quad S_a = \begin{cases} -\frac{a-l}{a}, & l < a \\ 0, & a \leq l \leq b \\ -\frac{l-b}{l}, & l > b \end{cases}, \quad S_b = \frac{m}{l}, \quad S_c = -\frac{n}{l}, \quad \text{其中, } l \text{ 为问题语句的}$$

长度, m 为问题语句所含的业务关键词的数目, n 为问题语句的停用词的数目, a 为问题语句的第一长度预设值, b 为问题语句的第二长度预设值;

30 若关键指标值 S 大于预设指标阈值, 则保留所述问题语句作为待匹配的问题语句, 或者, 若关键指标值 S 小于或者等于预设指标阈值, 则去除所述问题语句;

所述回答语句的分析规则为: 若回答语句不是陈述性语句, 则去除所述回答语句及其对应的问题语句。

35 5. 根据权利要求 1 所述的客服问答数据的处理方法, 其特征在于, 所述步骤 S1 包括:

将预设时间内的同一客户与同一客服的问答日志分为同一分组；或
将预设时间内的同一客户与各客服的问答日志分为同一分组；或
将预设时间内的各客户与同一客服的问答日志分为同一分组。

6. 根据权利要求 5 所述的客服问答数据的处理方法，其特征在于，所述步骤 S2 包括：

S21，对每一分组中的问答日志进行数据预处理；

S22，对数据预处理后的问答日志中的客户的问题语句和客服的回答语句按照预设的分析规则进行分析，以分析出待匹配的问题语句与对应的回答语句；

10 S23，抽取待匹配的问题语句与对应的回答语句，将抽取出的问题语句与对应的回答语句关联为问题答案对并将问题答案对加入至客服知识库中。

7. 根据权利要求 6 所述的客服问答数据的处理方法，其特征在于，所述步骤 S21 包括：

15 将每一分组中的问答日志中的字体统一转换为繁体字或者简体字、将每一分组中的问答日志中的字母统一转换为大写字母或者小写字母、去除问答日志中的特定语句，以及对每一分组中的问答日志中的语句进行分词处理。

8. 根据权利要求 6 所述的客服问答数据的处理方法，其特征在于，所述预设的分析规则包括问题语句的分析规则及回答语句的分析规则，所述问题语句的分析规则为：

20 计算问题语句的关键指标值 S：

$$S = S_a + S_b + S_c, \quad S_a = \begin{cases} -\frac{a-l}{a}, & l < a \\ 0, & a \leq l \leq b \\ -\frac{l-b}{l}, & l > b \end{cases}, \quad S_b = \frac{m}{l}, \quad S_c = -\frac{n}{l}, \quad \text{其中, } l \text{ 为问题语句的}$$

长度， m 为问题语句所含的业务关键词的数目， n 为问题语句的停用词的数目， a 为问题语句的第一长度预设值， b 为问题语句的第二长度预设值；

25 若关键指标值 S 大于预设指标阈值，则保留所述问题语句作为待匹配的问题语句，或者，若关键指标值 S 小于或者等于预设指标阈值，则去除所述问题语句；

所述回答语句的分析规则为：若回答语句不是陈述性语句，则去除所述回答语句及其对应的问题语句。

30 9. 一种电子装置，包括处理设备、存储设备及客服问答数据的处理系统，该客服问答数据的处理系统存储于该存储设备中，包括至少一个计算机可读指令，该至少一个计算机可读指令可被所述处理设备执行，以实现以下操作：

S1，从客服问答数据库获取预存的问答日志，并基于预设时间内客服与客户的问答对应关系对所述问答日志进行分组；

35 S2，对各分组中的问答日志按照预定的问题与答案的匹配规则进行匹

配，以得出待加入至客服知识库中的问题答案对。

10. 根据权利要求 9 所述的电子装置，其特征在于，所述步骤 S2 包括：

S21，对每一分组中的问答日志进行数据预处理；

5 S22，对数据预处理后的问答日志中的客户的问题语句和客服的回答语句按照预设的分析规则进行分析，以分析出待匹配的问题语句与对应的回答语句；

S23，抽取待匹配的问题语句与对应的回答语句，将抽取出的问题语句与对应的回答语句关联为问题答案对并将问题答案对加入至客服知识库中。

11. 根据权利要求 10 所述的电子装置，其特征在于，所述步骤 S21 包括：

10 将每一分组中的问答日志中的字体统一转换为繁体字或者简体字、将每一分组中的问答日志中的字母统一转换为大写字母或者小写字母、去除问答日志中的特定语句，以及对每一分组中的问答日志中的语句进行分词处理。

12. 根据权利要求 10 所述的电子装置，其特征在于，所述预设的分析规则包括问题语句的分析规则及回答语句的分析规则，所述问题语句的分析规则为：

15 计算问题语句的关键指标值 S：

$$S = S_a + S_b + S_c, \quad S_a = \begin{cases} -\frac{a-l}{a}, & l < a \\ 0, & a \leq l \leq b \\ -\frac{l-b}{l}, & l > b \end{cases}, \quad S_b = \frac{m}{l}, \quad S_c = -\frac{n}{l}, \quad \text{其中, } l \text{ 为问题语句的}$$

长度， m 为问题语句所含的业务关键词的数目， n 为问题语句的停用词的数目， a 为问题语句的第一长度预设值， b 为问题语句的第二长度预设值；

20 若关键指标值 S 大于预设指标阈值，则保留所述问题语句作为待匹配的问题语句，或者，若关键指标值 S 小于或者等于预设指标阈值，则去除所述问题语句；

所述回答语句的分析规则为：若回答语句不是陈述性语句，则去除所述回答语句及其对应的问题语句。

25 13. 根据权利要求 9 所述的电子装置，其特征在于，所述步骤 S1 包括：将预设时间内的同一客户与同一客服的问答日志分为同一分组；或
将预设时间内的同一客户与各客服的问答日志分为同一分组；或
将预设时间内的各客户与同一客服的问答日志分为同一分组。

14. 根据权利要求 13 所述的电子装置，其特征在于，所述步骤 S2 包括：

30 S21，对每一分组中的问答日志进行数据预处理；

S22，对数据预处理后的问答日志中的客户的问题语句和客服的回答语句按照预设的分析规则进行分析，以分析出待匹配的问题语句与对应的回答语句；

35 S23，抽取待匹配的问题语句与对应的回答语句，将抽取出的问题语句与对应的回答语句关联为问题答案对并将问题答案对加入至客服知识库中。

15. 根据权利要求 14 所述的电子装置, 其特征在于, 所述步骤 S21 包括:
将每一分组中的问答日志中的字体统一转换为繁体字或者简体字、将每一分组中的问答日志中的字母统一转换为大写字母或者小写字母、去除问答日志中的特定语句, 以及对每一分组中的问答日志中的语句进行分词处理。

5 16. 根据权利要求 14 所述的电子装置, 其特征在于, 所述预设的分析规则包括问题语句的分析规则及回答语句的分析规则, 所述问题语句的分析规则为:

计算问题语句的关键指标值 S:

$$S = S_a + S_b + S_c, \quad S_a = \begin{cases} -\frac{a-l}{a}, & l < a \\ 0, & a \leq l \leq b \\ -\frac{l-b}{l}, & l > b \end{cases}, \quad S_b = \frac{m}{l}, \quad S_c = -\frac{n}{l}, \quad \text{其中, } l \text{ 为问题语句的}$$

10 长度, m 为问题语句所含的业务关键词的数目, n 为问题语句的停用词的数目, a 为问题语句的第一长度预设值, b 为问题语句的第二长度预设值;

若关键指标值 S 大于预设指标阈值, 则保留所述问题语句作为待匹配的问题语句, 或者, 若关键指标值 S 小于或者等于预设指标阈值, 则去除所述问题语句;

15 所述回答语句的分析规则为: 若回答语句不是陈述性语句, 则去除所述回答语句及其对应的问题语句。

17. 一种计算机可读存储介质, 其上存储有至少一个可被处理设备执行以实现以下操作的计算机可读指令:

20 S1, 从客服问答数据库获取预存的问答日志, 并基于预设时间内客服与客户的问答对应关系对所述问答日志进行分组;

S2, 对各分组中的问答日志按照预定的问题与答案的匹配规则进行匹配, 以得出待加入至客服知识库中的问题答案对。

18. 根据权利要求 17 所述的存储介质, 其特征在于, 所述步骤 S2 包括:

S21, 对每一分组中的问答日志进行数据预处理;

25 S22, 对数据预处理后的问答日志中的客户的问题语句和客服的回答语句按照预设的分析规则进行分析, 以分析出待匹配的问题语句与对应的回答语句;

S23, 抽取待匹配的问题语句与对应的回答语句, 将抽取出的问题语句与对应的回答语句关联为问题答案对并将问题答案对加入至客服知识库中。

30 19. 根据权利要求 18 所述的存储介质, 其特征在于, 所述步骤 S21 包括:
将每一分组中的问答日志中的字体统一转换为繁体字或者简体字、将每一分组中的问答日志中的字母统一转换为大写字母或者小写字母、去除问答日志中的特定语句, 以及对每一分组中的问答日志中的语句进行分词处理。

20. 根据权利要求 18 所述的存储介质, 其特征在于, 所述预设的分析规则包括问题语句的分析规则及回答语句的分析规则, 所述问题语句的分析规

则为:

计算问题语句的关键指标值 S :

$$S = S_a + S_b + S_c, \quad S_a = \begin{cases} -\frac{a-l}{a}, & l < a \\ 0, & a \leq l \leq b \\ -\frac{l-b}{l}, & l > b \end{cases}, \quad S_b = \frac{m}{l}, \quad S_c = -\frac{n}{l}, \quad \text{其中, } l \text{ 为问题语句的}$$

长度, m 为问题语句所含的业务关键词的数目, n 为问题语句的停用词的数目, a 为问题语句的第一长度预设值, b 为问题语句的第二长度预设值;

若关键指标值 S 大于预设指标阈值, 则保留所述问题语句作为待匹配的问题语句, 或者, 若关键指标值 S 小于或者等于预设指标阈值, 则去除所述问题语句;

所述回答语句的分析规则为: 若回答语句不是陈述性语句, 则去除所述回答语句及其对应的问题语句。

1/3

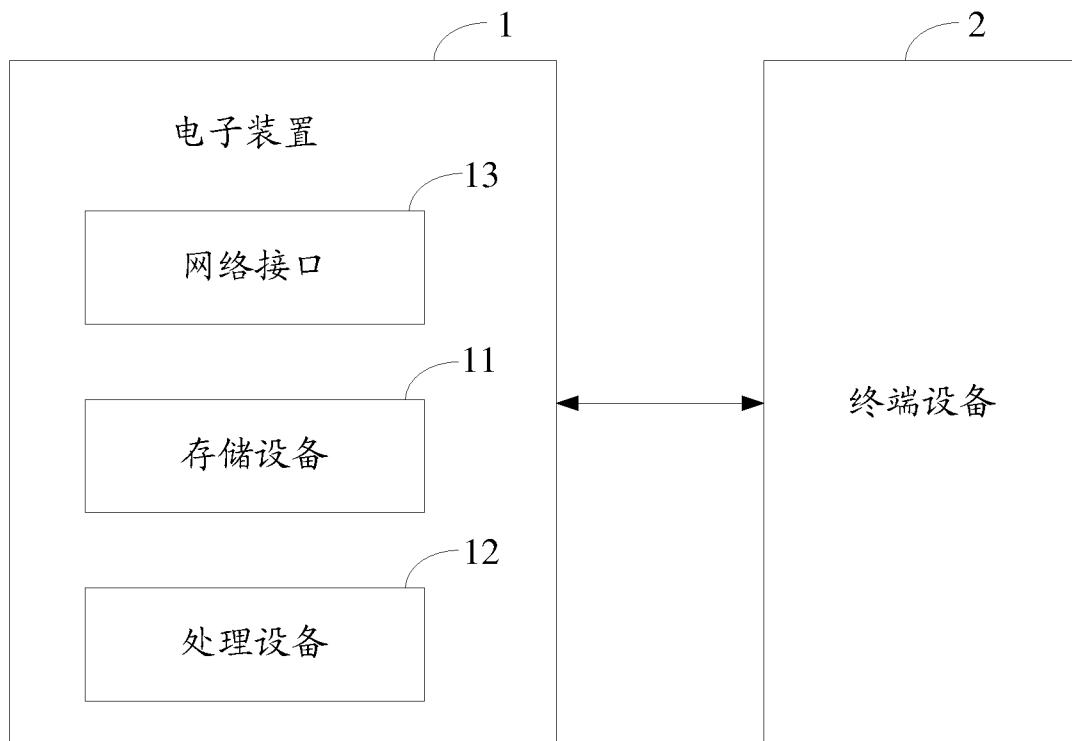


图 1

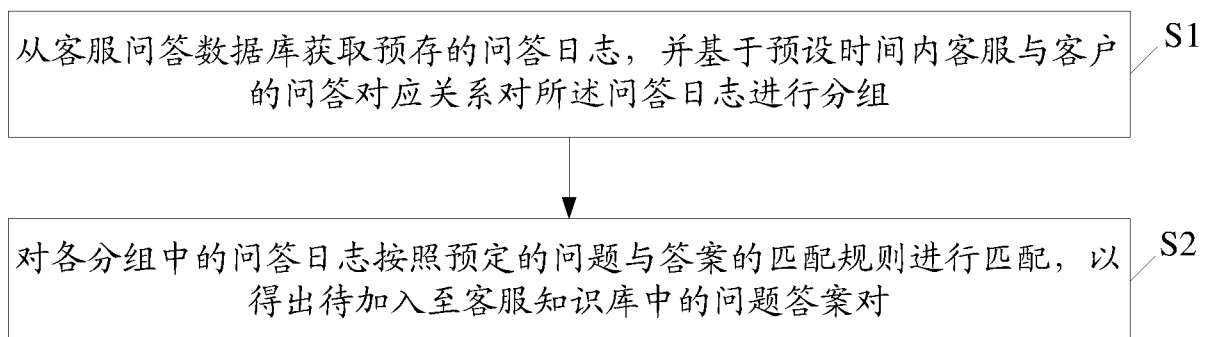


图 2

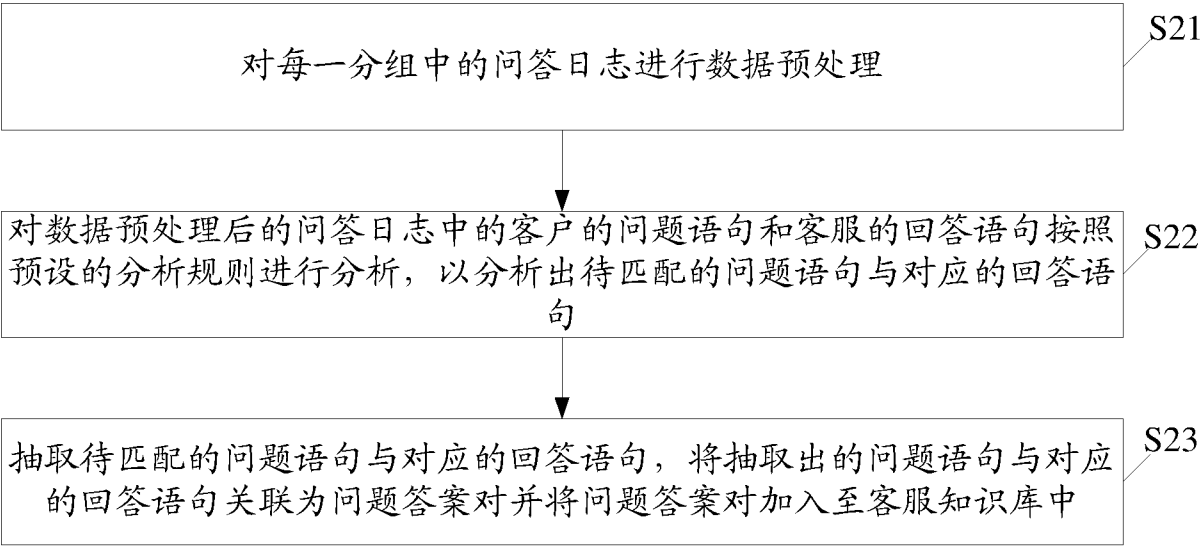


图 3

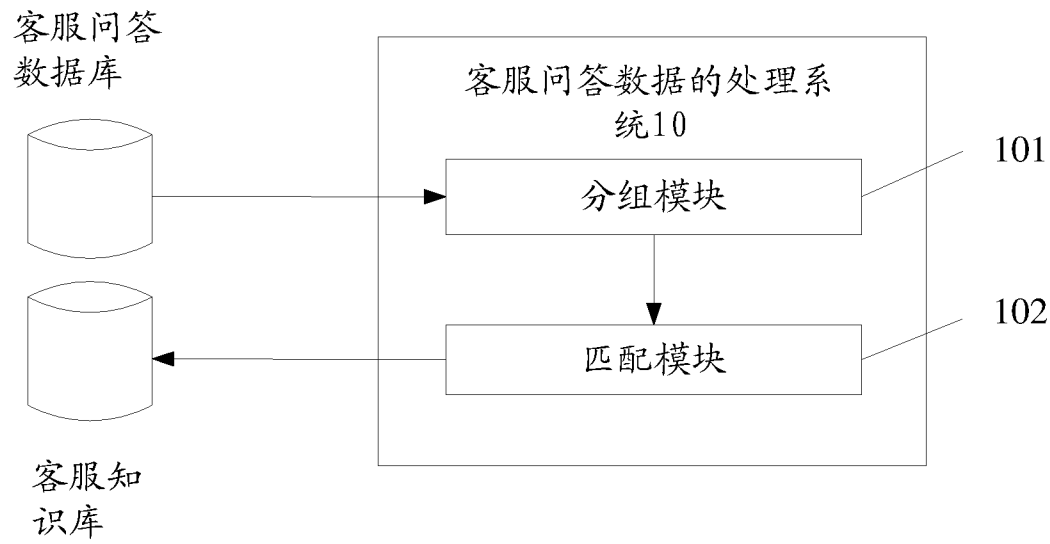


图 4

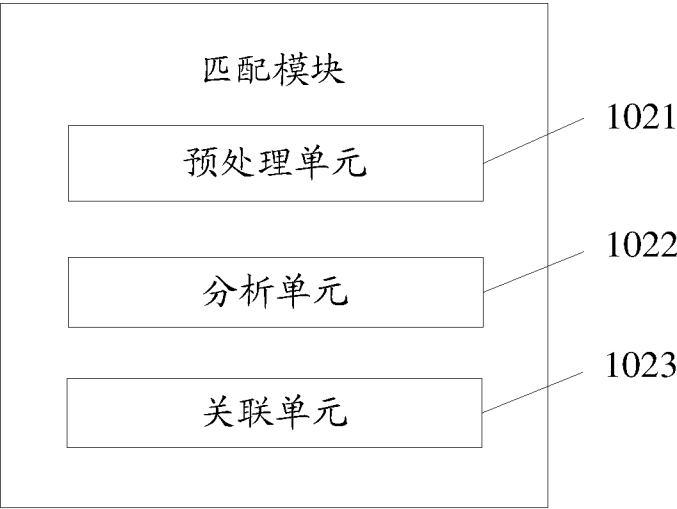


图 5

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2017/091372

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/30 (2006.01) i; G06F 17/27 (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

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)

EPODOC, WPI, CNPAT, CNKI, IEEE: 提问, 问题, 答案, 回答, 对, 抽取, 库, 创建, 更新, 客服, 聊天, 日志, 分组, 匹配, 阈值, question, answer, pair, extract, database, create, update, chat, log, divide, match, threshold

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 104679815 A (BEIJING UNISOUND INFORMATION TECHNOLOGY CO., LTD.), 03 June 2015 (03.06.2015), abstract, description, paragraphs 4 and 33-71, and figure 1	1-20
A	CN 105786851 A (BEIJING QIHOO TECHNOLOGY CO., LTD. et al.), 20 July 2016 (20.07.2016), entire document	1-20
A	CN 105608221 A (BEIJING GUANGNIAN WUXIAN SCIENCE AND TECHNOLOGY CO., LTD.), 25 May 2016 (25.05.2016), entire document	1-20
A	CN 102099803 A (RESULT LIMITED), 15 June 2011 (15.06.2011), entire document	1-20
A	US 2014122619 A1 (DUAN, Xiaojiang), 01 May 2014 (01.05.2014), entire document	1-20

☐ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents:	“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
“A” document defining the general state of the art which is not considered to be of particular relevance	“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
“E” earlier application or patent but published on or after the international filing date	“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
“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)	“&” document member of the same patent family
“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	

Date of the actual completion of the international search 19 September 2017	Date of mailing of the international search report 29 September 2017
Name and mailing address of the ISA 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 JIANG, Xiaoqing Telephone No. (86-10) 010-62413801

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2017/091372

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 104679815 A	03 June 2015	None	
CN 105786851 A	20 July 2016	None	
CN 105608221 A	25 May 2016	None	
CN 102099803 A	15 June 2011	EP 2220576 A2	25 August 2010
		AU 2008315748 A1	30 April 2009
		WO 2009053739 A2	30 April 2009
		US 2011153312 A1	23 June 2011
		IN- 201002776 E	29 October 2010
		CHENP	
US 2014122619 A1	01 May 2014	None	

国际检索报告

国际申请号

PCT/CN2017/091372

A. 主题的分类 G06F 17/30(2006.01)i; G06F 17/27(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 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)) EPODOC, WPI, CNPAT, CNKI, IEEE: 提问, 问题, 答案, 回答, 对, 抽取, 库, 创建, 更新, 客服, 聊天, 日志, 分组, 匹配, 阈值, question, answer, pair, extract, database, create, update, chat, log, divide, match, threshold																				
C. 相关文件 <table border="1"> <thead> <tr> <th>类 型*</th> <th>引用文件, 必要时, 指明相关段落</th> <th>相关的权利要求</th> </tr> </thead> <tbody> <tr> <td>X</td> <td>CN 104679815 A (北京云知声信息技术有限公司) 2015年 6月 3日 (2015 - 06 - 03) 摘要, 说明书第4、33-71段, 图1</td> <td>1-20</td> </tr> <tr> <td>A</td> <td>CN 105786851 A (北京奇虎科技有限公司 等) 2016年 7月 20日 (2016 - 07 - 20) 全文</td> <td>1-20</td> </tr> <tr> <td>A</td> <td>CN 105608221 A (北京光年无限科技有限公司) 2016年 5月 25日 (2016 - 05 - 25) 全文</td> <td>1-20</td> </tr> <tr> <td>A</td> <td>CN 102099803 A (瑞5优特有限公司) 2011年 6月 15日 (2011 - 06 - 15) 全文</td> <td>1-20</td> </tr> <tr> <td>A</td> <td>US 2014122619 A1 (DUAN, XIAOJIANG) 2014年 5月 1日 (2014 - 05 - 01) 全文</td> <td>1-20</td> </tr> </tbody> </table>			类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求	X	CN 104679815 A (北京云知声信息技术有限公司) 2015年 6月 3日 (2015 - 06 - 03) 摘要, 说明书第4、33-71段, 图1	1-20	A	CN 105786851 A (北京奇虎科技有限公司 等) 2016年 7月 20日 (2016 - 07 - 20) 全文	1-20	A	CN 105608221 A (北京光年无限科技有限公司) 2016年 5月 25日 (2016 - 05 - 25) 全文	1-20	A	CN 102099803 A (瑞5优特有限公司) 2011年 6月 15日 (2011 - 06 - 15) 全文	1-20	A	US 2014122619 A1 (DUAN, XIAOJIANG) 2014年 5月 1日 (2014 - 05 - 01) 全文	1-20
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求																		
X	CN 104679815 A (北京云知声信息技术有限公司) 2015年 6月 3日 (2015 - 06 - 03) 摘要, 说明书第4、33-71段, 图1	1-20																		
A	CN 105786851 A (北京奇虎科技有限公司 等) 2016年 7月 20日 (2016 - 07 - 20) 全文	1-20																		
A	CN 105608221 A (北京光年无限科技有限公司) 2016年 5月 25日 (2016 - 05 - 25) 全文	1-20																		
A	CN 102099803 A (瑞5优特有限公司) 2011年 6月 15日 (2011 - 06 - 15) 全文	1-20																		
A	US 2014122619 A1 (DUAN, XIAOJIANG) 2014年 5月 1日 (2014 - 05 - 01) 全文	1-20																		
☐ 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 2017年 9月 19日		国际检索报告邮寄日期 2017年 9月 29日																		
ISA/CN的名称和邮寄地址 中华人民共和国国家知识产权局(ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10)62019451		授权官员 姜晓庆 电话号码 (86-10)010-62413801																		

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2017/091372

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)
CN	104679815	A	2015年 6月 3日	无			
CN	105786851	A	2016年 7月 20日	无			
CN	105608221	A	2016年 5月 25日	无			
CN	102099803	A	2011年 6月 15日	EP	2220576	A2	2010年 8月 25日
				AU	2008315748	A1	2009年 4月 30日
				WO	2009053739	A2	2009年 4月 30日
				US	2011153312	A1	2011年 6月 23日
				IN- CHENP	201002776	E	2010年 10月 29日
US	2014122619	A1	2014年 5月 1日	无			