(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 102368260 A
(43)申请公布日 2012.03.07

(21)申请号 201110308830.7

(22)申请日 2011.10.12

(71)申请人 北京百度网讯科技有限公司
地址 100085 北京市海淀区上地十街 10 号百度大厦 2 层

(72)发明人 柴春光 黄际洲 时迎超

(74)专利代理机构 深圳市威世博知识产权代理事务所（普通合伙） 44280
代理人 何青瓦 李庆波

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

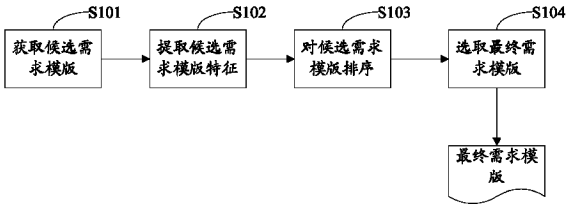
权利要求书 5 页 说明书 17 页 附图 5 页

(54)发明名称

一种生成领域需求模版的方法及其装置

(57)摘要

本发明提供了一种生成领域需求模版的方法及其装置，其中生成领域需求模版的方法包括：A. 获取特定领域的候选需求模版；B. 提取候选需求模版的特征；C. 利用提取的特征对候选需求模版进行排序；D. 根据排序的结果从候选需求模版中选择最终需求模版作为特定领域的需求模版。通过上述方式，实现了一种通用的生成高质量的领域需求模版的方法，为搜索引擎理解用户的行为目的提供了保障。



CN 102368260 A

1. 一种生成领域需求模版的方法，其特征在于，所述方法包括：

A. 获取特定领域的候选需求模版；

B. 提取候选需求模版的特征，所述特征至少包括：表征候选需求模板与所述特定领域之间紧密度的相似度特征、表征候选需求模板覆盖用户搜索请求 query 能力的泛化能力特征以及表征候选需求模板中未泛化的词语对候选需求模板正确性所产生影响的边界词特征中的至少一种；

C. 利用提取的特征对候选需求模版进行排序；

D. 根据排序的结果从候选需求模版中选择最终需求模版作为特定领域的需求模版。

2. 根据权利要求 1 所述的方法，其特征在于，所述步骤 A 包括：

A1. 从搜索日志中选取用户 query 中与预设的所述特定领域的限定词匹配的 query；

A2. 将选取的 query 中与预设的所述特定领域的槽关键词匹配的部分替换为通配符，得到候选需求模版。

3. 根据权利要求 2 所述的方法，其特征在于，在所述步骤 A2 之后还包括：根据预设的对所述特定领域的槽位数量要求，从所述步骤 A2 得到的候选需求模版中过滤掉不满足槽位数量要求的候选需求模版。

4. 根据权利要求 1 所述的方法，其特征在于，提取候选需求模版 W 的相似度特征的步骤包括：

获取所述 W 的核心词向量及所述特定领域的核心词向量；

计算所述 W 的核心词向量与所述特定领域的核心词向量之间的相似度，并将该相似度作为所述 W 的相似度特征。

5. 根据权利要求 4 所述的方法，其特征在于，获取所述 W 的核心词向量的步骤包括：

从所述 W 在搜索日志中覆盖的 query 里选取查询次数最多的 N_1 个 query，并在所述 N_1 个 query 从搜索引擎返回的搜索结果中确定核心词及核心词的权重，以形成所述 W 的核心词向量，其中 N_1 为正整数。

6. 根据权利要求 4 所述的方法，其特征在于，获取所述特定领域的核心词向量的步骤包括：

利用所述特定领域的种子 query 获取搜索引擎返回的搜索结果，并在该搜索结果中确定核心词及核心词的权重，以形成所述特定领域的核心词向量。

7. 根据权利要求 6 所述的方法，其特征在于，所述特定领域的种子 query 的获取方式包括：

方式一、从所述特定领域包含的所有候选需求模版中选取在搜索日志中覆盖 query 数最多的 N_2 个候选需求模版，并针对所述 N_2 个候选需求模版，从每个候选需求模版覆盖的 query 中选择查询次数最多的 M_1 个 query 作为种子 query，其中 N_2 及 M_1 为正整数；或者，

方式二、将预设的所述特定领域的槽关键词与预设的所述特定领域的限定词进行组合生成所述特定领域的种子 query；或者，

方式三、利用所述方式一选择出部分种子 query 后，利用预设的所述特定领域的槽关键词词典将所述方式一选择出的种子 query 中的槽关键词替换为所述槽关键词词典中的其他槽关键词得到扩展的种子 query；所述部分种子 query 和所述扩展的种子 query 构成所述特定领域的种子 query。

8. 根据权利要求 1 所述的方法，其特征在于，提取候选需求模版 W 的泛化能力特征的步骤包括：

确定所述 W 对应的槽关键词序列，统计所述 W 对应的槽关键词序列中互异的槽关键词序列的数量并依据该数量计算所述 W 的泛化能力特征，其中所述 W 对应的一个槽关键词序列是由所述 W 在搜索日志中覆盖的一个 query 中的槽关键词组成的序列。

9. 根据权利要求 1 所述的方法，其特征在于，提取候选需求模版 W 的边界词特征的步骤包括：

将所述特定领域包含的所有候选需求模版切分为片段，从得到的各切分片段中选取正片段并确定各正片段的权重以生成所述特定领域的正向量，从得到的各切分片段中选取负片段并确定各负片段的权重以生成所述特定领域的负向量；

确定所述 W 的切分片段的权重并使用所述 W 的切分片段及切分片段的权重构成所述 W 的向量；

计算所述 W 的向量与所述正向量的相似度 S_1，以及，所述 W 与所述负向量的相似度 S_2，并根据所述 S_1 与所述 S_2 的差值得到所述 W 的边界词特征。

10. 根据权利要求 9 所述的方法，其特征在于，所述特定领域的正向量和负向量的生成过程具体包括：

确定各切分片段对应的槽关键词序列，其中一个切分片段对应的一个槽关键词序列是包含该切分片段的一个候选需求模版所覆盖的一个 query 中的槽关键词组成的序列；

T1. 如果一个切分片段对应的所有槽关键词序列相同，则将该切分片段作为负片段，且该负片段的权重为 1；

T2. 如果一个切分片段对应的所有槽关键词序列不完全相同，但存在一个槽关键词序列在该切分片段的所有槽关键词序列中占的比例 P 大于预设的第一阈值，则将该切分片段作为负片段，且该负片段的权重为所述比例 P；

T3. 确定特定领域包含的每个候选需求模版对应的互异的槽关键词序列的数量，得到该数量中的最大值 Z_1，如果一个切分片段不满足所述 T1 及所述 T2 中的条件，且该切分片段对应的互异的槽关键词序列的数量 Z_2 与所述 Z_1 的比值大于预设的第二阈值，则将该切分片段作为正片段，且该正片段的权重为 Z_2 与 Z_1 的比值。

11. 根据权利要求 9 所述的方法，其特征在于，确定所述 W 的切分片段的权重的步骤包括：

统计所述 W 的切分片段在所述 W 中出现的次数并将该次数作为对应切分片段的权重。

12. 根据权利要求 1 所述的方法，其特征在于，所述步骤 C 包括：

从候选需求模版中选取标准模版集；

利用所述标准模版集训练提取的各特征对应的参数，将训练中使得所述标准模版集中的模版在所有候选需求模版中的排名无法更靠前时的参数值作为对应特征的权重；

使用提取的各特征及特征的权重计算候选需求模版的得分，并根据该得分对各候选需求模版进行排序。

13. 根据权利要求 12 所述的方法，其特征在于，从候选需求模版中选取标准模版集的步骤包括：

针对提取的每个特征分别基于特征值对候选需求模版进行排序，分别针对每个特征取

排列在前 N_3 位的候选需求模版作为对应特征的模版集合，其中 N_3 为正整数；

取各特征的模版集合之间的交集作为标准模版集。

14. 根据权利要求 1 所述的方法，其特征在于，所述步骤 D 包括：

将排序位于前 N_4 位的候选需求模版选取为最终需求模版，其中 N_4 为正整数；

利用排序位于前 M_2 位的候选需求模版的边界词获取关键词集合，并将排序位于前 N_4 位之后的候选需求模版中的边界词均属于所述关键词集合的候选需求模版选取为最终需求模版，其中所述边界词为候选需求模版中未被泛化的词，所述关键词是与所述边界词同义的词或与所述边界词之间的互信息满足要求的词，M_2 为正整数且 M_2 小于或等于 N_4。

15. 一种生成领域需求模版的装置，其特征在于，所述装置包括：

候选模版获取单元，用于获取特定领域的候选需求模版；

特征提取单元，用于提取候选需求模版的特征，其中所述特征提取单元至少包括相似度特征提取单元、泛化能力特征提取单元或边界词特征提取单元中的一个，所述相似度特征提取单元用于提取表征候选需求模板与所述特定领域之间紧密度的相似度特征，所述泛化能力特征提取单元用于提取表征候选需求模板覆盖用户搜索请求 query 能力的泛化能力特征，所述边界词特征提取单元用于提取表征候选需求模板中未泛化的词语对候选需求模板正确性所产生影响的边界词特征；

排序单元，用于利用所述特征提取单元提取的特征对候选需求模版进行排序；

选取单元，用于根据所述排序单元排序的结果从候选需求模版中选择最终需求模版作为特定领域的需求模版。

16. 根据权利要求 15 所述的装置，其特征在于，所述候选模版获取单元包括：

限定单元，用于从搜索日志中选取用户 query 中与预设的所述特定领域的限定词匹配的 query；

泛化单元，用于将所述限定单元选取的 query 中与预设的所述特定领域的槽关键词匹配的部分替换为通配符，得到候选需求模版。

17. 根据权利要求 16 所述的装置，其特征在于，所述候选模版获取单元进一步包括过滤单元，用于根据预设的对所述特定领域的槽位数量要求，从所述泛化单元得到的候选需求模版中过滤掉不满足槽位数量要求的候选需求模版。

18. 根据权利要求 15 所述的装置，其特征在于，所述相似度提取单元包括：

模版词向量生成单元，用于在提取候选需求模版 W 的相似度特征时，获取所述 W 的核心词向量；

领域词向量生成单元，用于获取所述特定领域的核心词向量；

计算单元，用于计算所述 W 的核心词向量与所述特定领域的核心词向量之间的相似度，并将该相似度作为所述 W 的相似度特征。

19. 根据权利要求 18 所述的装置，其特征在于，所述模版词向量生成单元从所述 W 在搜索日志中覆盖的 query 里选取查询次数最多的 N_1 个 query，并在所述 N_1 个 query 从搜索引擎返回的搜索结果中确定核心词及核心词的权重，以形成所述 W 的核心词向量，其中所述 N_1 为正整数。

20. 根据权利要求 18 所述的装置，其特征在于，所述领域词向量生成单元利用所述特定领域的种子 query 获取搜索引擎返回的搜索结果，并在该搜索结果中确定核心词及核心

词的权重，以形成所述特定领域的核心词向量。

21. 根据权利要求 20 所述的装置，其特征在于，所述领域词向量生成单元获取所述特定领域的种子 query 的方式包括：

方式一、从所述特定领域包含的所有候选需求模版中选取在搜索日志中覆盖 query 数最多的 N_2 个候选需求模版，并针对所述 N_2 个候选需求模版，从每个候选需求模版覆盖的 query 中选择查询次数最多的 M_1 个 query 作为种子 query，其中 N_2 及 M_1 为正整数；或者，

方式二、将预设的所述特定领域的槽关键词与预设的所述特定领域的限定词进行组合生成所述特定领域的种子 query；或者，

方式三、利用所述方式一选择出部分种子 query 后，利用预设的所述特定领域的槽关键词词典将所述方式一选择出的种子 query 中的槽关键词替换为所述槽关键词词典中的其他槽关键词得到扩展的种子 query；所述部分种子 query 和所述扩展的种子 query 构成所述特定领域的种子 query。

22. 根据权利要求 15 所述的装置，其特征在于，所述泛化能力特征提取单元在提取候选需求模版 W 的泛化能力特征时，确定所述 W 对应的槽关键词序列，统计所述 W 对应的槽关键词序列中互异的槽关键词序列的数量并依据该数量计算所述 W 的泛化能力特征，其中所述 W 的一个槽关键词序列是由所述 W 在搜索日志中覆盖的一个 query 中的槽关键词组成的序列。

23. 根据权利要求 15 所述的装置，其特征在于，所述边界词特征提取单元包括：

切分单元，用于将特定领域包含的所有候选需求模版切分为片段；

正负向量生成单元，用于从所述切分单元得到的各切分片段中选取正片段并确定正片段的权重以生成所述特定领域的正向量，从得到的各切分片段中选取负片段并确定各负片段的权重以生成所述特定领域的负向量；

模版向量生成单元，用于在提取候选需求模版 W 的边界词特征时，确定所述 W 的切分片段的权重并使用所述 W 的切分片段及切分片段的权重构成所述 W 的向量；

相似度计算单元，用于计算所述 W 的向量与所述正向量的相似度 S_1，以及，所述 W 的向量与所述负向量的相似度 S_2，并根据所述 S_1 与所述 S_2 的差值得到所述 W 的边界词特征。

24. 根据权利要求 23 所述的装置，其特征在于，所述正负向量生成单元包括：

槽关键词序列确定单元，用于确定各切分片段对应的槽关键词序列，其中一个切分片段对应的一个槽关键词序列是包含该切分片段的一个候选需求模版所覆盖的一个 query 中的槽关键词组成的序列；

正负片段选取单元，用于按照下列方式从各切分片段中选取正片段和负片段以及确定正片段和负片段的权重：

T1. 如果一个切分片段对应的所有槽关键词序列相同，则将该切分片段作为负片段，且该负片段的权重为 1；

T2. 如果一个切分片段对应的所有槽关键词序列不完全相同，但存在一个槽关键词序列在该切分片段的所有槽关键词序列中占的比例 P 大于预设的第一阈值，则将该切分片段作为负片段，且该负片段的权重为所述比例 P；

T3. 确定特定领域包含的每个候选需求模版对应的互异的槽关键词序列的数量，得到该数量中的最大值 Z_1，如果一个切分片段不满足所述 T1 及所述 T2 中的条件，且该切分片段

对应的互异的槽关键词序列的数量 Z_2 与所述 Z_1 的比值大于预设的第二阈值，则将该切分片段作为正片段，且该正片段的权重为 Z_2 与 Z_1 的比值。

25. 根据权利要求 23 所述的装置，其特征在于，所述模版向量特征生成单元在确定所述 W 的切分片段的权重时，统计所述 W 的切分片段在所述 W 中出现的次数并将该次数作为对应切分片段的权重。

26. 根据权利要求 15 所述的装置，其特征在于，所述排序单元包括：

标准模版集选取单元，用于从候选需求模版中选取标准模版集；

训练单元，用于利用所述标准模版集训练提取的各特征对应的参数，将训练中使得所述标准模版集中的模版在所有候选需求模版中的排名无法更靠前时的参数值作为对应特征的权重；

计算与排序单元，用于使用所述特征提取单元提取的各特征及所述训练单元得到的各特征的权重计算候选需求模版的得分，并根据该得分对候选需求模版进行排序。

27. 根据权利要求 26 所述的装置，其特征在于，所述标准模版集选取单元包括：

模版集合确定单元，用于针对提取的每个特征基于特征值对候选需求模版进行排序，分别针对每个特征取排列在前 N_3 位的候选需求模版作为对应特征的模版集合，其中 N_3 为正整数；

交集单元，用于取各特征的模版集合之间的交集作为标准模版集。

28. 根据权利要求 15 所述的装置，其特征在于，所述选取单元包括：

第一选取单元，用于将排序位于前 N_4 位的候选需求模版选取为最终需求模版，其中 N_4 为正整数；

第二选取单元，用于利用排序位于前 M_2 位的候选需求模版的边界词获取关键词集合，并将排序位于前 N_4 位之后的候选需求模版中的边界词均属于所述关键词集合的候选需求模版选取为最终需求模版，其中所述边界词为候选需求模版中未被泛化的词，所述关键词是与所述边界词同义的词或与所述边界词之间的互信息满足要求的词，M_2 为正整数且 M_2 小于或等于 N_4。

一种生成领域需求模版的方法及其装置

【技术领域】

[0001] 本发明涉及自然语言处理技术，特别涉及一种生成领域需求模版的方法及其装置。

【背景技术】

[0002] 搜索引擎为人们找到所需信息提供了极大的便利。在传统的搜索引擎为用户提供信息的方式中，是通过查找包含用户搜索关键字的索引，为用户返回与关键字匹配的相关页面来实现的。例如，用户的搜索请求 (query) 为“北京汽车 4S 店招聘销售主管”，这时会得到招聘网站的搜索结果页面，用户可以通过点击该页面进入招聘网站，然后在该招聘网站内填写相关信息并在站内进行检索，得到自己真正需要的信息。如果搜索引擎能够更好地理解用户在检索时的真正目的，那么搜索引擎就能够更准确地向用户返回真正符合其需求的信息。因此，自然语言处理对搜索引擎而言非常重要。在自然语言处理中，可以采用基于领域的需求模版对用户的搜索目的进行识别。例如，用户的 query 为“大钟寺到西单怎么走”，如果该 query 与交通领域的需求模版相匹配，就可以得知该用户有交通领域的需求，因此可以直接向该用户返回与交通领域相关的应用。可见，是否能够产生高质量的领域需求模版，对搜索引擎正确理解用户的搜索意图而言，非常重要。

[0003] 在以往生成领域需求模版时，针对不同的应用，通常采用不同的挖掘方法进行，这不仅浪费了大量的人力物力，而且这种生成领域需求模版的方法，适应性差，难以随着应用的变化而做出相应的改变。

【发明内容】

[0004] 本发明所要解决的技术问题是提供一种生成领域需求模版的方法及装置，以解决采用现有技术生成的领域需求模版适应性差的缺陷。

[0005] 本发明为解决技术问题而采用的技术方案是提供一种生成领域需求模版的方法，包括 :A. 获取特定领域的候选需求模版 ;B. 提取候选需求模版的特征，所述特征至少包括 : 表征候选需求模板与所述特定领域之间紧密度的相似度特征、表征候选需求模板覆盖用户搜索请求 query 能力的泛化能力特征以及表征候选需求模板中未泛化的词语对候选需求模板正确性所产生影响的边界词特征中的至少一种 ;C. 利用提取的特征对候选需求模版进行排序 ;D. 根据排序的结果从候选需求模版中选择最终需求模版作为特定领域的需求模版。

[0006] 根据本发明之一优选实施例，所述步骤 A 包括 :A1. 从搜索日志中选取用户 query 中与预设的所述特定领域的限定词匹配的 query ;A2. 将选取的 query 中与预设的所述特定领域的槽关键词匹配的部分替换为通配符，得到候选需求模版。

[0007] 根据本发明之一优选实施例，在所述步骤 A2 之后还包括 :根据预设的对所述特定领域的槽位数量要求，从所述步骤 A2 得到的候选需求模版中过滤掉不满足槽位数量要求的候选需求模版。

[0008] 根据本发明之一优选实施例，提取候选需求模版 W 的相似度特征的步骤包括：获取所述 W 的核心词向量及所述特定领域的核心词向量；计算所述 W 的核心词向量与所述特定领域的核心词向量之间的相似度，并将该相似度作为所述 W 的相似度特征。

[0009] 根据本发明之一优选实施例，获取所述 W 的核心词向量的步骤包括：从所述 W 在搜索日志中覆盖的 query 里选取查询次数最多的 N_1 个 query，并在所述 N_1 个 query 从搜索引擎返回的搜索结果中确定核心词及核心词的权重，以形成所述 W 的核心词向量，其中 N_1 为正整数。

[0010] 根据本发明之一优选实施例，获取所述特定领域的核心词向量的步骤包括：利用所述特定领域的种子 query 获取搜索引擎返回的搜索结果，并在该搜索结果中确定核心词及核心词的权重，以形成所述特定领域的核心词向量。

[0011] 根据本发明之一优选实施例，所述特定领域的种子 query 的获取方式包括：方式一、从所述特定领域包含的所有候选需求模版中选取在搜索日志中覆盖 query 数最多的 N_2 个候选需求模版，并针对所述 N_2 个候选需求模版，从每个候选需求模版覆盖的 query 中选择查询次数最多的 M_1 个 query 作为种子 query，其中 N_2 及 M_1 为正整数；或者，方式二、将预设的所述特定领域的槽关键词与预设的所述特定领域的限定词进行组合生成所述特定领域的种子 query；或者，方式三、利用所述方式一选择出部分种子 query 后，利用预设的所述特定领域的槽关键词词典将所述方式一选择出的种子 query 中的槽关键词替换为所述槽关键词词典中的其他槽关键词得到扩展的种子 query；所述部分种子 query 和所述扩展的种子 query 构成所述特定领域的种子 query。

[0012] 根据本发明之一优选实施例，提取候选需求模版 W 的泛化能力特征的步骤包括：确定所述 W 对应的槽关键词序列，统计所述 W 对应的槽关键词序列中互异的槽关键词序列的数量并依据该数量计算所述 W 的泛化能力特征，其中所述 W 对应的一个槽关键词序列是由所述 W 在搜索日志中覆盖的一个 query 中的槽关键词组成的序列。

[0013] 根据本发明之一优选实施例，提取候选需求模版 W 的边界词特征的步骤包括：将所述特定领域包含的所有候选需求模版切分为片段，从得到的各切分片段中选取正片段并确定各正片段的权重以生成所述特定领域的正向量，从得到的各切分片段中选取负片段并确定各负片段的权重以生成所述特定领域的负向量；确定所述 W 的切分片段的权重并使用所述 W 的切分片段及切分片段的权重构成所述 W 的向量；计算所述 W 的向量与所述正向量的相似度 S_1，以及，所述 W 与所述负向量的相似度 S_2，并根据所述 S_1 与所述 S_2 的差值得到所述 W 的边界词特征。

[0014] 根据本发明之一优选实施例，所述特定领域的正向量和负向量的生成过程具体包括：确定各切分片段对应的槽关键词序列，其中一个切分片段对应的一个槽关键词序列是包含该切分片段的一个候选需求模版所覆盖的一个 query 中的槽关键词组成的序列；T1. 如果一个切分片段对应的所有槽关键词序列相同，则将该切分片段作为负片段，且该负片段的权重为 1；T2. 如果一个切分片段对应的所有槽关键词序列不完全相同，但存在一个槽关键词序列在该切分片段的所有槽关键词序列中占的比例 P 大于预设的第一阈值，则将该切分片段作为负片段，且该负片段的权重为所述比例 P；T3. 确定特定领域包含的每个候选需求模版对应的互异的槽关键词序列的数量，得到该数量中的最大值 Z_1，如果一个切分片段不满足所述 T1 及所述 T2 中的条件，且该切分片段对应的互异的槽关键词序列的数量

Z_2 与所述 Z_1 的比值大于预设的第二阈值，则将该切分片段作为正片段，且该正片段的权重为 Z_2 与 Z_1 的比值。

[0015] 根据本发明之一优选实施例，确定所述 W 的切分片段的权重的步骤包括：统计所述 W 的切分片段在所述 W 中出现的次数并将该次数作为对应切分片段的权重。

[0016] 根据本发明之一优选实施例，所述步骤 C 包括：从候选需求模版中选取标准模版集；利用所述标准模版集训练提取的各特征对应的参数，将训练中使得所述标准模版集中的模版在所有候选需求模版中的排名无法更靠前时的参数值作为对应特征的权重；使用提取的各特征及特征的权重计算候选需求模版的得分，并根据该得分对各候选需求模版进行排序。

[0017] 根据本发明之一优选实施例，从候选需求模版中选取标准模版集的步骤包括：针对提取的每个特征分别基于特征值对候选需求模版进行排序，分别针对每个特征取排列在前 N_3 位的候选需求模版作为对应特征的模版集合，其中 N_3 为正整数；取各特征的模版集合之间的交集作为标准模版集。

[0018] 根据本发明之一优选实施例，所述步骤 D 包括：将排序位于前 N_4 位的候选需求模版选取为最终需求模版，其中 N_4 为正整数；利用排序位于前 M_2 位的候选需求模版的边界词获取关键词集合，并将排序位于前 N_4 位之后的候选需求模版中的边界词均属于所述关键词集合的候选需求模版选取为最终需求模版，其中所述边界词为候选需求模版中未被泛化的词，所述关键词是与所述边界词同义的词或与所述边界词之间的互信息满足要求的词，M_2 为正整数且 M_2 小于或等于 N_4。

[0019] 本发明还提供了一种生成领域需求模版的装置，包括：候选模版获取单元，用于获取特定领域的候选需求模版；特征提取单元，用于提取候选需求模版的特征，其中所述特征提取单元至少包括相似度特征提取单元、泛化能力特征提取单元或边界词特征提取单元中的一个，所述相似度特征提取单元用于提取表征候选需求模板与所述特定领域之间紧密度的相似度特征，所述泛化能力特征提取单元用于提取表征候选需求模板覆盖用户搜索请求 query 能力的泛化能力特征，所述边界词特征提取单元用于提取表征候选需求模板中未泛化的词语对候选需求模板正确性所产生影响的边界词特征；排序单元，用于利用所述特征提取单元提取的特征对候选需求模版进行排序；选取单元，用于根据所述排序单元排序的结果从候选需求模版中选择最终需求模版作为特定领域的需求模版。

[0020] 根据本发明之一优选实施例，所述候选模版获取单元包括：限定单元，用于从搜索日志中选取用户 query 中与预设的所述特定领域的限定词匹配的 query；泛化单元，用于将所述限定单元选取的 query 中与预设的所述特定领域的槽关键词匹配的部分替换为通配符，得到候选需求模版。

[0021] 根据本发明之一优选实施例，所述候选模版获取单元进一步包括过滤单元，用于根据预设的对所述特定领域的槽位数量要求，从所述泛化单元得到的候选需求模版中过滤掉不满足槽位数量要求的候选需求模版。

[0022] 根据本发明之一优选实施例，所述相似度提取单元包括：模版词向量生成单元，用于在提取候选需求模版 W 的相似度特征时，获取所述 W 的核心词向量；领域词向量生成单元，用于获取所述特定领域的核心词向量；计算单元，用于计算所述 W 的核心词向量与所述特定领域的核心词向量之间的相似度，并将该相似度作为所述 W 的相似度特征。

[0023] 根据本发明之一优选实施例，所述模版词向量生成单元从所述 W 在搜索日志中覆盖的 query 里选取查询次数最多的 N_1 个 query，并在所述 N_1 个 query 从搜索引擎返回的搜索结果中确定核心词及核心词的权重，以形成所述 W 的核心词向量，其中所述 N_1 为正整数。

[0024] 根据本发明之一优选实施例，所述领域词向量生成单元利用所述特定领域的种子 query 获取搜索引擎返回的搜索结果，并在该搜索结果中确定核心词及核心词的权重，以形成所述特定领域的核心词向量。

[0025] 根据本发明之一优选实施例，所述领域词向量生成单元获取所述特定领域的种子 query 的方式包括：方式一、从所述特定领域包含的所有候选需求模版中选取在搜索日志中覆盖 query 数最多的 N_2 个候选需求模版，并针对所述 N_2 个候选需求模版，从每个候选需求模版覆盖的 query 中选择查询次数最多的 M_1 个 query 作为种子 query，其中 N_2 及 M_1 为正整数；或者，方式二、将预设的所述特定领域的槽关键词与预设的所述特定领域的限定词进行组合生成所述特定领域的种子 query；或者，方式三、利用所述方式一选择出部分种子 query 后，利用预设的所述特定领域的槽关键词词典将所述方式一选择出的种子 query 中的槽关键词替换为所述槽关键词词典中的其他槽关键词得到扩展的种子 query；所述部分种子 query 和所述扩展的种子 query 构成所述特定领域的种子 query。

[0026] 根据本发明之一优选实施例，所述泛化能力特征提取单元在提取候选需求模版 W 的泛化能力特征时，确定所述 W 对应的槽关键词序列，统计所述 W 对应的槽关键词序列中互异的槽关键词序列的数量并依据该数量计算所述 W 的泛化能力特征，其中所述 W 的一个槽关键词序列是由所述 W 在搜索日志中覆盖的一个 query 中的槽关键词组成的序列。

[0027] 根据本发明之一优选实施例，所述边界词特征提取单元包括：切分单元，用于将特定领域包含的所有候选需求模版切分为片段；正负向量生成单元，用于从所述切分单元得到的各切分片段中选取正片段并确定正片段的权重以生成所述特定领域的正向量，从得到的各切分片段中选取负片段并确定各负片段的权重以生成所述特定领域的负向量；模版向量生成单元，用于在提取候选需求模版 W 的边界词特征时，确定所述 W 的切分片段的权重并使用所述 W 的切分片段及切分片段的权重构成所述 W 的向量；相似度计算单元，用于计算所述 W 的向量与所述正向量的相似度 S_1，以及，所述 W 的向量与所述负向量的相似度 S_2，并根据所述 S_1 与所述 S_2 的差值得到所述 W 的边界词特征。

[0028] 根据本发明之一优选实施例，所述正负向量生成单元包括：槽关键词序列确定单元，用于确定各切分片段对应的槽关键词序列，其中一个切分片段对应的一个槽关键词序列是包含该切分片段的一个候选需求模版所覆盖的一个 query 中的槽关键词组成的序列；正负片段选取单元，用于按照下列方式从各切分片段中选取正片段和负片段以及确定正片段和负片段的权重：T1. 如果一个切分片段对应的所有槽关键词序列相同，则将该切分片段作为负片段，且该负片段的权重为 1；T2. 如果一个切分片段对应的所有槽关键词序列不完全相同，但存在一个槽关键词序列在该切分片段的所有槽关键词序列中占的比例 P 大于预设的第一阈值，则将该切分片段作为负片段，且该负片段的权重为所述比例 P；T3. 确定特定领域包含的每个候选需求模版对应的互异的槽关键词序列的数量，得到该数量中的最大值 Z_1，如果一个切分片段不满足所述 T1 及所述 T2 中的条件，且该切分片段对应的互异的槽关键词序列的数量 Z_2 与所述 Z_1 的比值大于预设的第二阈值，则将该切分片段作为正片段，且该正片段的权重为 Z_2 与 Z_1 的比值。

[0029] 根据本发明之一优选实施例，所述模版向量特征生成单元在确定所述 W 的切分片段的权重时，统计所述 W 的切分片段在所述 W 中出现的次数并将该次数作为对应切分片段的权重。

[0030] 根据本发明之一优选实施例，所述排序单元包括：标准模版集选取单元，用于从候选需求模版中选取标准模版集；训练单元，用于利用所述标准模版集训练提取的各特征对应的参数，将训练中使得所述标准模版集中的模版在所有候选需求模版中的排名无法更靠前时的参数值作为对应特征的权重；计算与排序单元，用于使用所述特征提取单元提取的各特征及所述训练单元得到的各特征的权重计算候选需求模版的得分，并根据该得分对候选需求模版进行排序。

[0031] 根据本发明之一优选实施例，所述标准模版集选取单元包括：模版集合确定单元，用于针对提取的每个特征基于特征值对候选需求模版进行排序，分别针对每个特征取排列在前 N_3 位的候选需求模版作为对应特征的模版集合，其中 N_3 为正整数；交集单元，用于取各特征的模版集合之间的交集作为标准模版集。

[0032] 根据本发明之一优选实施例，所述选取单元包括：第一选取单元，用于将排序位于前 N_4 位的候选需求模版选取为最终需求模版，其中 N_4 为正整数；第二选取单元，用于利用排序位于前 M_2 位的候选需求模版的边界词获取关键词集合，并将排序位于前 N_4 位之后的候选需求模版中的边界词均属于所述关键词集合的候选需求模版选取为最终需求模版，其中所述边界词为候选需求模版中未被泛化的词，所述关键词是与所述边界词同义的词或与所述边界词之间的互信息满足要求的词，M_2 为正整数且 M_2 小于或等于 N_4。

[0033] 由以上技术方案可以看出，本发明提供了一种通用性的领域需求模版的生成方法，针对不同的领域，均可通过本方法自动挖掘候选需求模版，并提取候选需求模版的特征对候选需求模版的质量进行评定，从而能够在候选需求模版中得到高质量的需求模版。本发明得到的高质量的各个领域的需求模版为搜索引擎理解用户的行为目的提供了保障。

【附图说明】

[0034] 图 1 为本发明中生成领域的需求模版的方法的流程示意图；

[0035] 图 2 为本发明中获取候选需求模版的实施例的流程示意图；

[0036] 图 3 为本发明中利用种子 query 获取搜索引擎返回数据的示意图；

[0037] 图 4 为本发明中生成领域需求模版的装置的实施例的结构示意框图；

[0038] 图 5 为本发明中相似度特征提取单元的实施例的结构示意框图；

[0039] 图 6 为本发明中边界词特征提取单元的实施例的结构示意框图；

[0040] 图 7 为本发明中标准模版集选取单元的实施例的结构示意框图。

【具体实施方式】

[0041] 为了使本发明的目的、技术方案和优点更加清楚，下面结合附图和具体实施例对本发明进行详细描述。

[0042] 请参考图 1，图 1 为本发明中生成领域的需求模版的方法的流程示意图。如图 1 所示，该方法包括：

[0043] 步骤 S101：获取特定领域的候选需求模版。

[0044] 步骤 S102 :提取候选需求模版的特征。

[0045] 步骤 S103 :利用提取的特征对候选需求模版进行排序。

[0046] 步骤 S104 :根据排序的结果从候选需求模版中选取最终的需求模版作为特定领域的需求模版。

[0047] 下面通过具体的实施例对上述方法进行详细介绍。

[0048] 本发明中,特定领域是反映用户搜索目的的一个范围,如公交领域、天气领域等等,这些领域反映了用户搜索信息时的搜索目的。

[0049] 请参考图 2,图 2 为本发明中获取候选需求模版的实施例的流程示意图。在本实施例中,利用了领域限定词词典与槽关键词词典对用户搜索日志 (querylog) 中的用户搜索请求 query 进行处理,从而生成候选需求模版。

[0050] 领域限定词词典包含了与各个领域相关的词语,其中特定领域的限定词是与特定领域相关的词语,在本实施例中,特定领域的限定词用于在选取 query 时,对 query 进行过滤。只有包含特定领域的限定词的 query,才会进行泛化,泛化生成的候选需求模版,就属于特定领域的候选需求模版。领域限定词词典中的词语可以通过下列途径收集得到 :

[0051] 首先可以从用户的 query 中挖掘领域种子词作为领域限定词,其中领域种子词可以通过人工的方式配置,或者采用人工的方式在搜索日志中标注。

[0052] 然后通过查找同义词词典,得到与领域种子词同义的词语作为领域限定词,此外,还可以通过使用度量两个词紧密程度的互信息选取搜索日志中与种子词关联程度高的词语一并作为领域限定词。词语之间的互信息可通过对大规模语料进行统计得到,由于属于现有技术,在此不再赘述。以公交领域为例,表 1 给出了部分领域限定词的示例 :

[0053] 表 1

[0054]

领域种子词	领域限定词
公交	工交/工交车/公车/公共交通/公共交通线路/公共汽车/公交/公交车/公交联营车/公交路线/公交汽车/公交线/公交线路/公汽/共交/市区公交/公交车线路
怎么走	如何去/怎么去/怎么走/怎样走/

[0055] 生成候选需求模版的过程,就是对 query 进行泛化的过程,所谓泛化,指的是将用户 query 中与特定领域的槽关键词匹配的部分替换为通配符。槽关键词是用于泛化的词语,通过查找槽关键词词典确定,该词典可通过收集各种专有名词得到。

[0056] 例如“北京 15 路公交车路线”这样的 query,在泛化以后,可以得到“[城市名][公交路线] 公交车路线”这样的需求模版。每一个“[]”符号代表模版的一个槽位,表示该位置在满足通配符属性要求的情况下可进行替换,例如上面这个模版与“上海郊 14 路公共车路线”也匹配。

[0057] 在得到上述候选需求模版后,还可以根据对候选需求模版所属的特定领域预设的槽位数量要求决定是否对这些候选需求模版进行过滤处理。例如在火车信息查询领域,query 中的可变信息一般仅涉及起点和终点,因此可以将火车信息查询领域的模版预定槽

数设置为 2，凡是不符合预定槽数要求的模版都会被过滤掉，以降低后续对候选需求模版进行处理的复杂度。

[0058] 本实施例中，步骤 S102 中提取的特征，至少包括以下特征中的一种：

[0059] 相似度特征，用于描述候选需求模版与特定领域联系的紧密度；泛化能力特征，用于描述候选需求模版覆盖用户搜索请求 query 的能力；边界词特征，用于描述候选需求模版中未被泛化的词语对候选需求模版的正确性产生的影响。

[0060] 下面对上述三个特征的计算方式的实施例进行具体介绍。

[0061] 1、相似度特征

[0062] 一个候选需求模版 W 的相似度特征可以通过计算候选需求模板 W 的核心词向量与该候选需求模板 W 所属特定领域的核心词向量之间的余弦距离得到，具体可以采用下列公式 (1) 进行计算：

[0063] sim_score = CossSimilarity(pattern_vector, seed_query_centroid) (1)

[0064] 其中，sim_score 表示候选需求模版 W 的相似度特征值，pattern_vector 表示候选需求模板 W 的核心词向量，seed_query_centriod 表示特定领域的核心词向量，CossSimilarity 表示余弦相似度函数。

[0065] 核心词向量，是由核心词为向量特征形成的向量。因此，在计算相似度特征时，首先要确定如何选取核心词。

[0066] 在确定特定领域的核心词时，可以利用该特定领域的种子 query 获取搜索引擎返回的数据，并利用搜索引擎返回的数据确定核心词。请参考图 3，图 3 为本发明中利用种子 query 获取搜索引擎返回数据的示意图。如图 3 所示，种子 query 为“北京 15 路公交车路线”，该种子 query 可以从搜索引擎得到多个搜索结果。将这些搜索结果的标题 (title) 和内容 (text) 进行预处理 (包括分句、分词、去除停用词等) 后，得到统计语料。针对统计语料中的每个词，统计该词出现的句子数及该词与检索词共同出现的句子数，并统计包含检索词的句子数，其中检索词是种子 query 分词后得到的词语。

[0067] 得到上述信息后，可采用下列公式 (2) 计算每个词的权重，并将权值大于设定阈值的词语作为核心词，这些核心词的权重相应地构成了对应向量特征的权重。

[0068]

$$Centrality_{sch_term}(w)=\frac{\log(Co(w,sch_term)+1)}{\log(sf(w)+1)+\log(sf(sch_term)+1)}\times\log(idf(w)+1) \qquad (2)$$

[0069] 其中，$Centrality_{sch_term}(w)$ 表示词 w 的权值，Co(w, sch_term) 表示词 w 与检索词 sch_term 共同出现的句子个数；sf(sch_term) 表示含有检索词 sch_term 的句子个数；sf(w) 表示包含词 w 句子个数；idf(w) 表示词 w 的逆向文档频率，可通过查找利用大规模语料统计得来的逆文档频率表得到。

[0070] 在获取特定领域的种子 query 时，可采用下列几种实施方式：

[0071] 实施方式一：

[0072] 在特定领域包含的候选需求模版中选取在搜索日志中覆盖的 query 数最多的 N_2 个候选需求模版，并针对这 N_2 个候选需求模版，从每个候选需求模版覆盖的 query 中选择查询次数最多的 M_1 个 query 作为种子 query，其中 N_2、M_1 为正整数，优选地，M_1 等于 1。例如下面表 2 为公交领域的候选需求模版：

[0073] 表 2

[0074]

候选需求模板	覆盖的 query	查询次数
[城市名][公交路线]公交车路线	北京 15 路公交车路线	239
	上海郊 14 路公交车路线	141
	沈阳铁西 2 线公交车路线	11
[城市名]公交[公交路线]	北京公交 14 线	23
	北京公交 23 路	834
[地点名]到[地点名]的公交车	故宫到奥运村的公交车路线	2
	西二旗到西直门的公交车路线	5
到[地点名]公交车路线	到西二旗公交车路线	4
公交月票 [城市名]	公交月票 上海	1
公交卡[地点名]充值点	公交卡西二旗充值点	1
公交 [城市名] 电话	公交 北京 电话	53
公交 [地点名] 抓贼	公交 西直门 抓贼	11
公交卡断了怎么办[城市名]	公交卡断了怎么办 北京	1

[0075] 假设 $N_2 = 2, M_1 = 1$，则表 3 示出了针对表 2 中的候选需求模版采用实施方式一得到的种子 query 及其相应的候选需求模版。

[0076] 表 3

[0077]

种子 query	对应模板
北京 15 路公交车路线	[城市名][公交路线]公交车路线
北京公交 23 路	[城市名]公交[公交路线]

[0078] 在这种实施方式下，种子 query 来源于用户的真实 query，能够更好地代表用户的习惯。

[0079] 实施方式二：

[0080] 将特定领域的槽关键词与特定领域限定词进行组合生成种子 query。

[0081] 以生成公交领域的种子 query 为例，请参考表 4：

[0082] 表 4

[0083]

生成的种子 query	对应的槽关键词	对应的领域限定词
北京 15 路 公交车路线	北京 15 路	公交车路线
上海 公交	上海	公交

[0084] 这种方式下，生成的种子 query 结构简单。

[0085] 优选地，可采用实施方式三来获取种子 query。

[0086] 实施方式三：

[0087] 采用实施方式一的方法选出部分种子 query，然后利用槽关键词词典将选取的种子 query 中的槽关键词替换为特定领域的其他槽关键词以得到扩展的种子 query。

[0088] 例如表 5 所示为采用实施方式三得到的种子 query。

[0089] 表 5

[0090]

选取的种子 query	扩展的种子 query
北京 15 路公交车路线	沈阳 15 路公交车路线
北京公交 23 路	济南公交 12 路

[0091] 上述过程可得到特定领域的核心词向量，下面将描述获取候选需求模版的核心词向量的过程。

[0092] 首先，与获取特定领域的核心词向量类似的，需要先获取统计语料。在获取统计语料时，首先从候选需求模版在搜索日志中覆盖的 query 里，选取查询次数最多的 N_1 个 query 作为待搜索 query，然后使用这些待搜索 query 从搜索引擎中获取搜索结果，对这些搜索结果的 title 和 text 进行预处理，就可以得到统计语料了，其中 N_1 为正整数。

[0093] 在得到的统计语料中，统计每个词的在语料中出现的频率，并按照下列公式 (3) 计算每个词的权重，权重大于设定阈值的词就可作为候选需求模版的核心词，核心词的权重即为对应的向量特征的权重。

[0094] $$Weight(w) = \log(tf(w)+1) \times \log(idf(w)+1) \quad (3)$$

[0095] 其中，Weight(w) 表示词 w 的权值，tf(w) 表示词 w 的频率，idf(w) 表示词 w 的逆向文档频率，可通过查找利用大规模语料统计得来的逆文档频率表得到。

[0096] 在得到候选需求模版的核心词向量与特定领域的核心词向量后，就可按照公式 (1) 计算候选需求模版的相似度特征了。

[0097] 2、泛化能力特征

[0098] 泛化能力特征可用候选需求模版对应的槽关键词序列中互异的槽关键词序列的数量来衡量，其中候选需求模版对应的一个槽关键词序列是由候选需求模版在搜索日志中覆盖的一个 query 中的槽关键词组成的序列。

[0099] 例如对模版“[城市名][公交路线]公交车路线”，其覆盖的 query 有“北京 15 路公交车路线”、“上海郊 14 路公交车路线”、“沈阳铁西 2 线公交车路线”、“北京 15 路公交车

路线图查询”，则槽关键词序列有“北京 15 路”、“上海郊 14 路”、“沈阳铁西 2 线”和“北京 15 路”，互异的槽关键词序列为“北京 15 路”、“上海 郊 14 路”和“沈阳 铁西 2 线”，因此对模版“[城市名][公交路线] 公交车路线”而言，它的泛化能力特征值就是 3。

[0100] 优选的，泛化能力特征采用下列方式进行计算。首先确定特定领域包含的每个候选需求模版对应的互异的槽关键词序列的数量及该数量中的最大值，然后按照下列公式 (4) 计算每个候选需求模版的泛化能力特征值 ：

[0101] $general_score_i = \log(pattern_dif_query_i+1)/\log(max_dif_query+1)$ (4)

[0102] 其中，$general_score_i$ 表示候选需求模版 i 的泛化能力特征值，$pattern_dif_query_i$ 表示候选需求模版 i 对应的互异的槽关键词序列的数量，max_dif_query 表示该候选需求模板 i 所属特定领域包含的每个候选需求模版对应的互异的槽关键词序列的数量中的最大值。

[0103] 3、边界词特征

[0104] 边界词是候选需求模版中未被泛化的词语。候选需求模版中未被泛化的词语对最终生成的模版的正确性产生影响。例如在公交领域，“[城市名][公交路线] 公交车路线”这样的需求模版，显然比“公交卡断了怎么办 [城市名]”这样的模版更能反映公交领域的需求。

[0105] 在本发明中，候选需求模版 W 的边界词特征通过下面的公式 (5) 来计算。

[0106] boundary_word_score

[0107] = CosSimilarity(pattern_centroid, positive_centroid) (5)

[0108] -CosSimilarity(pattern_centroid, negative_centroid)

[0109] 其中，boundary_word_score 为候选需求模版 W 的边界词特征，CosSimilarity 为余弦相似度函数，pattern_centroid 为候选需求模版 W 形成的向量，positive_centroid 为特定领域的正向量，negative_centroid 为特定领域的负向量。

[0110] 下面分别介绍如何获取公式中的各个变量值。

[0111] 生成特定领域的正负向量的过程包括 ：

[0112] 将特定领域包含的所有候选需求模版按照 n 元词组 (n-gram)(n > 1) 的方式进行切分，优选地，取 n = 2，可得到各个切分片段，其中所谓 n-gram 就是能够进行语义表达的最小粒度的 n 个词语按顺序出现的组合，其中 n 为预设的正整数。例如对“[城市名][公交路线] 公交车路线”这个模版，假设其能够进行语义表达的最小粒度的词语分别为“[城市名]”、“[公交路线]”和“公交车路线”，则该模版的 2-gram 的切分片段分别是“[城市名][公交路线]”、“[公交路线] 公交车路线”，或者对“公交卡断了怎么办 [城市名]”这个模版，假设其能够进行语义表达的最小粒度的词语分别为“公交卡”、“断了”、“怎么办”和“[城市名]”，则该模版的 2-gram 的切分片段分别是“公交卡断了”、“断了怎么办”、“怎么办 [城市名]”。

[0113] 从各切分片段中选取正片段和负片段，其中一个正片段就是正向量的一个向量特征，一个负片段就是负向量中的一个向量特征，并确定各个向量特征的权重。该过程包括 ：

[0114] A. 确定各切分片段对应的槽关键词序列，其中一个切分片段的一个槽关键词序列是包含该切分片段的一个候选需求模版所覆盖的一个 query 中的槽关键词组成的序列。

[0115] 例如，对切分片段“[城市名] 公交”来说，包含该切分片段的候选需求模版及其覆

盖的 query 如表 6 所示：

[0116] 表 6

[0117]

候选需求模版	覆盖的 query
[城市名]公交[公交路线]	北京公交 15 路
	上海公交 36 路
[城市名]公交车[公交路线]路线查询	北京公交车 15 路路线查询
[城市名]公交罢工了	杭州公交罢工了

[0118] 则对切分片段“[城市名]公交”而言，它的槽关键词序列包括“北京 15 路”、“上海 36 路”、“北京 15 路”、“杭州”。

[0119] B. 按照下列方式确定从各切分片段中选取正向量特征和负向量特征并确定各向量特征的权重：

[0120] (1) 如果一个切分片段的所有槽关键词序列相同，则该切分片段作为负向量特征，且该负向量特征的权重为 1。

[0121] (2) 如果一个切分片段的所有槽关键词序列不完全相同，但存在一个槽关键词序列在该切分片段的所有槽关键词序列中占的比例 P 大于预设的第一阈值时，则将该切分片段作为负向量特征，且该向量特征的权重为比例 P，优选地，第一阈值为 90%。

[0122] (3) 确定特定领域包含的每个候选需求模版对应的互异的槽关键词序列的数量，得到该数量中的最大值 Z_1，如果一个切分片段不符合上述两种情况，且该切分片段的互异的槽关键词序列的数量 Z_2 与 Z_1 的比值大于预设的第二阈值时，则将该切分片段作为正向量特征，且该正向量特征的权重为 Z_2 与 Z_1 的比值，优选地，第二阈值为 1%。

[0123] 例如上面的切分片段“[城市名]公交”，互异的槽关键词序列分别为“北京 15 路”、“上海 36 路”、“杭州”，互异的槽关键词序列的数目为 3，其中“北京 15 路”在所有槽关键词序列中的比例为 2/4，“上海 /36 路”在所有槽关键词序列中的比例为 1/4，“杭州”在所有槽关键词序列中的比例为 1/4，因此该切分片段不符合 (1) 或 (2) 中情况，所以该切分片段不属于负向量特征，假设特定领域包含的每个候选需求模版对应的互异的槽关键词序列的数量中的最大值为 10 且第二阈值为 1%，则由于 3/10 大于 1%，所以该切分片段应该作为正向量特征。

[0124] 以表 2 所示的模版为例，采用上述方式得到的正向量与负向量分别如表 7 和表 8 所示：

[0125] 表 7

[0126]

正向量中的向量特征	特征权重
[城市名][公交路线]	1.000000

[公交路线]公交车路线	1.000000
[城市名]公交	0.666667
公交[公交路线]	0.666667
[地点名]到	0.666667
到[地点名]	1.000000
[地点名]的	0.666667
的公交车	0.666667

[0127] 表 8

[0128]

负向量中的向量特征	特征权重
[地点名]公交车路线	1.000000
公交月票	1.000000
月票[城市名]	1.000000
公交卡[地点名]	1.000000
[地点名]充值点	1.000000
公交[城市名]	1.000000
[城市名]电话	1.000000
公交[地点名]	1.000000
[地点名]抓贼	1.000000
公交卡断了	1.000000
断了怎么办	1.000000
怎么办[城市名]	1.000000

[0129] 候选需求模版 W 形成的向量中的向量特征是候选需求模版 W 的切分片段，其中切分的方式与正负向量中描述的类似，而特征权重可由对应的切分片段在候选需求模版 W 中出现的次数来确定。

[0130] 例如“[城市名][公交路线]公交车路线”这个模版包含的切分片段分别为“[城市名][公交路线]”和“[公交路线]公交车路线”，由于这两个切分片段在该模版中出现

的次数都是 1，所以模版“[城市名][公交路线] 公交车路线”对应的向量特征“[城市名][公交路线]”和“[公交路线] 公交车路线”的特征权重分别都是 1。如果一个模版为“[城市名][公交路线][城市名][公交路线]”，那么对这个模版的向量特征“[城市名][公交路线]”而言，特征权重就是 2。

[0131] 候选需求模版的向量特征的特征权重的确定方式不唯一，除了以切分片段在模版中出现的次数作为对应的向量特征的特征权重，还可以采用布尔值的形式来确定对应的向量特征的特征权重，在此不对特征权重的计算方式进行限定。

[0132] 以表 2 所示的候选需求模版为例，各个候选需求模版的边界词特征如表 9 所示：

[0133] 表 9

[0134]

候选需求模板	与正向量的余弦相似度	与负向量的余弦相似度	边界词特征
[城市名][公交路线]公交车路线	0.618853	0	0.618853
[城市名]公交[公交路线]	0.412569	0	0.412569
[地点名]到[地点名]的公交车	0.656393	0	0.656393
到[地点名]公交车路线	0.309426	0.204124	0.105302
公交月票 [城市名]	0	0.408248	-0.408248
公交卡[地点名]充值点	0	0.408248	-0.408248
公交 [城市名] 电话	0	0.408248	-0.408248
公交 [地点名] 抓贼	0	0.408248	-0.408248
公交卡断了怎么办 [城市名]	0	0.5	-0.5

[0135] 在步骤 S103 中，排序的过程包括：

[0136] 1、从候选需求模版中选取标准模版集，包括：

[0137] 针对提取的每个特征分别基于特征值对候选需求模版进行排序，分别针对每个特征取排列在前 N_3 位的候选需求模版作为对应特征的模版集合，其中 N_3 为正整数。

[0138] 取各特征的模版集合之间的交集，并将该交集作为标准模版集。

[0139] 例如：针对特征 1、2、3 对候选需求模版 S1-S10 进行排序，得到表 10：

[0140] 表 10

[0141]

特征 1		特征 2		特征 3	
模版	排名	模版	排名	模版	排名
S5	1	S4	1	S2	1
S6	2	S5	2	S10	2
S4	3	S2	3	S5	3
S2	4	S8	4	S6	4
S1	5	S1	5	S1	5
…	…	…	…	…	…

[0142] 如果 $N_3 = 5$，则特征 1 的模版集合为 {S5 S6 S4 S2 S1}，特征 2 的模版集合为 {S4 S5 S2 S8 S1}，特征 3 的模版集合为 {S2 S10 S5 S6 S1}，则各特征的模版集合的交集就是 {S1 S2 S5}。

[0143] 2、利用标准模版集训练提取的各特征对应的参数，将训练中使得标准模版集中的模版在所有候选需求模版中的排名无法更靠前时的参数值作为对应特征的权重。

[0144] 公式 (6) 是基于提取的全部特征对所有候选需求模版进行排序时，各候选需求模版的得分，得分越高说明该候选需求模版的质量越好，因此排名就越靠前。

[0145] $$total_score = \lambda_1 sim_score + \lambda_2 general_score + \lambda_3 boundary_word_score \quad (6)$$

[0146] 其中，sim_score、general_score 和 boundary_word_score 分别是相似度特征、泛化能力特征及边界词特征的值，λ_1、λ_2 及 λ_3 是待训练的参数，代表了各个特征的权重。

[0147] 训练参数采用的方法是梯度下降，通过连续迭代，不停调整参数的值，以使得标准模版集中的模版的排名尽可能地靠前，直到标准模版集中的模版在所有候选需求模版中的排序不再提前，这时的各参数值即为对应特征的权重。

[0148] 3、使用提取的各特征及其权重计算候选需求模版的得分，并根据该得分对候选需求模版进行排序，即采用下列公式 (6) 计算候选需求模版的得分，其中公式 (6) 中的 λ_1、λ_2 及 λ_3 为训练得到的各个特征的权重。

[0149] 通过上述方式计算出候选需求模版的得分，便可以按照得分从高到低的顺序对候选需求模版进行排序。

[0150] 步骤 S104 中选取最终的需求模版时，除了会将排序位于前 N_4 位的候选需求模版作为最终需求模版以外，还会利用排序位于前 M_2 位的候选需求模版的边界词从排序位于前 N_4 位之后的候选需求模版中选取最终需求模版，其中 M_2 与 N_4 均为正整数且 $M_2 \leqslant N_4$。

[0151] 具体做法为：

[0152] 利用关键词词典，获取与排序位于前 M_2 位的候选需求模版的边界词对应的关键词集合，其中所述关键词是与所述边界词同义的词或与所述边界词之间的互信息满足要求的词；

[0153] 将排序位于前 N_4 位之后的候选需求模版中的边界词均属于关键词集合的候选需求模版作为最终需求模版。

[0154] 假设排名在前 M_2 位以内的模版有：[城市名][公交路线]公交车路线、[地点名]到[地点名]的公交车、[城市名]公交[公交路线]，其中边界词有“公交车路线”、“到”、“公交车”、“的”，通过关键词词典，可以得到与上述边界词对应的关键词集合为“公交 / 工交 / 工交车 / 公车 / 公共交通 / 公共交通线路 / 公共汽车 / 公交 / 公交车 / 公交联营车 /

公交路线 / 公交汽车 / 公交线 / 公交线路 / 公汽 / 共交 / 市区公交 / 公交车线路 / 的 / 到 / 到达”，则对于排名在前 N_4 位之后的模版“到［地点名］公交车路线”而言，由于这个模版的边界词“到”与“公交车路线”均在关键词集合里，因此这个模版也可以被选取为最终模版。上述关键词词典中的关键词可通过各种现有技术得到，如挖掘同义词或互信息计算等，在此不再详述。

[0155] 请参考图 4，图 4 为本发明中生成领域模版的装置的实施例的结构示意框图。如图 4 所示，该装置包括：候选需求模版获取单元 201、特征提取单元 202、排序单元 203 及选取单元 204。

[0156] 其中候选需求模版获取单元 201 用于获取特定领域的候选需求模版。优选地，候选需求模版获取单元 201 包括限定单元 2011 和泛化单元 2012。

[0157] 其中限定单元 2011 用于从搜索日志中选取用户搜索请求 query 中与预设的特定领域的限定词匹配的 query，其中特定领域限定词是与特定领域相关的词语。泛化单元 2012 用于将选取的 query 中与预设的特定领域的槽关键词匹配的部分替换为通配符，以得到候选需求模版，其中特定领域的槽关键词是特定领域用于泛化的词语。

[0158] 进一步地，所述候选需求模版获取单元 201 还可包括一过滤单元，用于根据预设的对所述特定领域的槽位数量要求，从泛化单元得到的候选需求模版中过滤掉不满足槽位数量要求的候选需求模版。

[0159] 特征提取单元 202 用于提取候选需求模版的特征。优选地，特征提取单元 202 包括相似度特征提取单元 2021、泛化能力特征提取单元 2022 及边界词特征提取单元 2023 中的至少一种。

[0160] 其中，相似度特征提取单元 2021 用于提取候选需求模版的相似度特征，所述相似度特征用于描述候选需求模版与特定领域联系的紧密度。请参考图 5，图 5 为本发明中相似度特征提取单元的实施例的结构示意框图。如图 5 所示，相似度特征提取单元 2021 包括模版词向量生成单元 2021_1、领域词向量生成单元 2021_2 和计算单元 2021_3。

[0161] 其中模版词向量生成单元 2021_1 用于在提取候选需求模版 W 的相似度特征时，获取 W 的核心词向量。

[0162] 领域词向量生成单元 2021_2 用于获取特定领域的核心词向量。

[0163] 计算单元 2021_2 用于计算该候选需求模版的核心词向量与特定领域的核心词向量之间的相似度，并将该相似度作为 W 的相似度特征。

[0164] 优选地，模版词向量生成单元 2021_1 在获取 W 的核心词向量时，从 W 在搜索日志中覆盖的 query 中选取查询次数最多的 N_1 个 query，并在这 N_1 个 query 从搜索引擎返回的搜索结果中确定核心词及核心词的权重，以形成 W 的核心词向量，其中所述 N_1 为任意正整数。

[0165] 领域词向量生成单元 2021_2 获取特定领域的种子 query 的方式包括：

[0166] 方式一、从特定领域包含的所有候选需求模版中选取在搜索日志中覆盖 query 数最多的 N_2 个候选需求模版，并针对这 N_2 个候选需求模版，从每个候选需求模版覆盖的 query 中选择查询次数最多的 M_1 个 query 作为种子 query，其中 N_2 及 M_1 为正整数。

[0167] 方式二、将预设的特定领域的槽关键词与预设的特定领域的限定词进行组合生成所述特定领域的种子 query。

[0168] 方式三、利用方式一选择出部分种子 query 后，利用预设的特定领域的槽关键词词典将方式一选择出的种子 query 中的槽关键词替换为槽关键词词典中的其他槽关键词得到扩展的种子 query ；所述部分种子 query 和所述扩展的种子 query 构成特定领域的种子 query。

[0169] 优选地，领域词向量生成单元 2021_2 可采用方式三获取特定领域的种子 query。

[0170] 请继续参考图 4。泛化能力特征提取单元 2022，用于提取候选需求模版的泛化能力特征。所述泛化能力特征用于描述候选需求模版覆盖用户搜索请求 query 的能力。

[0171] 优选地，泛化能力特征提取单元 2022 在提取候选需求模版 W 的泛化能力特征时，确定 W 对应的槽关键词序列，统计 W 对应的槽关键词序列中互异的槽关键词序列的数量并依据该数量计算 W 的泛化能力特征，其中 W 对应的一个槽关键词序列是由 W 在搜索日志中覆盖的一个 query 中的槽关键词组成的序列。

[0172] 边界词特征提取单元 2023，用于提取候选需求模版的边界词特征。所述边界词特征用于描述候选需求模版中未被泛化的词语对候选需求模版的正确性产生的影响。

[0173] 请参考图 6，图 6 为本发明中边界词特征提取单元的实施例的结构示意框图。如图 6 所示，该实施例包括 ：切分单元 2023_1、正负向量生成单元 2023_2、模版向量生成单元 2023_3 及相似度计算单元 2023_4。

[0174] 其中切分单元 2023_1 用于将特定领域包含的所有候选需求模版切分为片段。

[0175] 正负向量生成单元 2023_2 用于从切分单元 2023_1 得到的各切分片段中选取正片段并确定正片段的权重以生成特定领域的正向量，从得到的各切分片段中选取负片段并确定负片段的权重以生成特定领域的负向量。优选地，正负向量生成单元 2023_3 包括槽关键词序列确定单元 2023_21 及正负片段选取单元 2023_22。

[0176] 其中槽序列词确定单元 2023_21 用于确定各切分片段对应的槽关键词序列，其中一个切分片段对应的一个槽关键词序列是包含该切分片段的一个候选需求模版所覆盖的一个 query 中的槽关键词组成的序列。

[0177] 正负片段选取单元 2023_22 用于按照下列方式从各切分片段中选取正片段和负片段并确定正片段和负片段的权重 ：

[0178] (1) 如果一个切分片段对应的所有槽关键词序列相同，则将该切分片段作为负片段，且该负片段的权重为 1 ；

[0179] (2) 如果一个切分片段对应的所有槽关键词序列不完全相同，但存在一个槽关键词序列在该切分片段的所有槽关键词序列中占的比例 P 大于预设的第一阈值，则将该切分片段作为负片段，且该负片段的权重为所述比例 P ；

[0180] (3) 确定特定领域包含的每个候选需求模版对应的互异的槽关键词序列的数量，得到该数量中的最大值 Z_1，如果一个切分片段不满足所述 T1 及所述 T2 中的条件，且该切分片段对应的互异的槽关键词序列的数量 Z_2 与所述 Z_1 的比值大于预设的第二阈值，则将该切分片段作为正片段，且该正片段的权重为 Z_2 与 Z_1 的比值。

[0181] 模版向量生成单元 2023_3 用于在提取候选需求模版 W 的边界词特征时，确定 W 的切分片段的权重并使用 W 的切分片段及切分片段的权重构成 W 的向量。优选地，模版向量生成单元 2023_3 在确定 W 的切分片段的权重时，统计 W 的切分片段在 W 中出现的次数，并将该次数作为对应切分片段的权重。

[0182] 相似度计算单元 2023_4 用于计算 W 的向量与正向量的相似度 S_1 以及 W 的向量与负向量的相似度 S_2，并根据 S_1 与 S_2 的差值得到 W 的边界词特征。

[0183] 请继续参考图 4。排序单元 203 用于利用特征提取单元 202 提取的特征对候选需求模版进行排序。排序单元 203 包括标准模版集选取单元 2031、训练单元 2032 及计算与排序单元 2033。

[0184] 其中，标准模版集选取单元 2031 用于从候选需求模版中选取标准模版集。请参考图 7，图 7 为本发明中标准模版集选取单元的实施例的结构示意框图。如图 7 所示，标准模版集选取单元 2031 包括模版集合确定单元 2031_1 和交集单元 2031_2。其中模版集合确定单元 2031_1，用于针对提取的每个特征基于特征值对各候选需求模版进行排序，分别针对每个特征取排列在前 N_3 位的候选需求模版作为对应特征的模版集合，其中 N_3 为正整数。交集单元 2031_2，用于取各特征的模版集合之间的交集作为标准模版集。

[0185] 请继续参考图 4。训练单元 2032 用于使用标准模版集训练提取的各特征对应的参数，将训练中使得标准模版集中的模版在所有候选需求模版中的排名无法更靠前时的参数值作为对应特征的权重。

[0186] 计算与排序单元 2033 用于使用特征提取单元 202 提取的各特征及训练单元 2032 得到的各特征的权重计算候选需求模版的得分，并根据该得分对各候选需求模版进行排序。优选地，按照得分从高到低对各候选需求模版进行排序。

[0187] 选取单元 204 用于根据排序单元 203 排序的结果从候选需求模版中选取最终需求模版作为特定领域的需求模版。优选地，选取单元 204 包括第一选取单元 2041 和第二选取单元 2042。其中第一选取单元 2041 用于将排序位于前 N_4 位的候选需求模版选取为最终需求模版，其中 N_4 为正整数。第二选取单元 2042 用于利用排序位于前 M_2 位的候选需求模版的边界词获取关键词集合，并将排序位于前 N_4 位之后的候选需求模版中的边界词均属于关键词集合的候选需求模版选取为最终需求模版，其中所述边界词为候选需求模版中未被泛化的词，所述关键词是与所述边界词同义的词或与所述边界词之间的互信息满足要求的词，M_2 为正整数且 M_2 小于或等于 N_4。

[0188] 以上所述仅为本发明的较佳实施例而已，并不用以限制本发明，凡在本发明的精神和原则之内，所做的任何修改、等同替换、改进等，均应包含在本发明保护的范围之内。

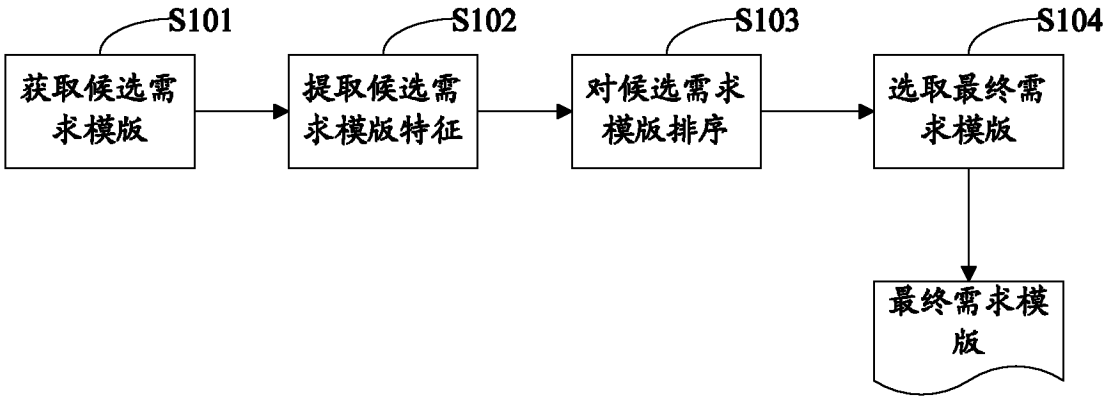

S101
S102
S103
S104
获取候选需求模版
提取候选需求模版特征
对候选需求模版排序
选取最终需求模版
最终需求模版

图 1

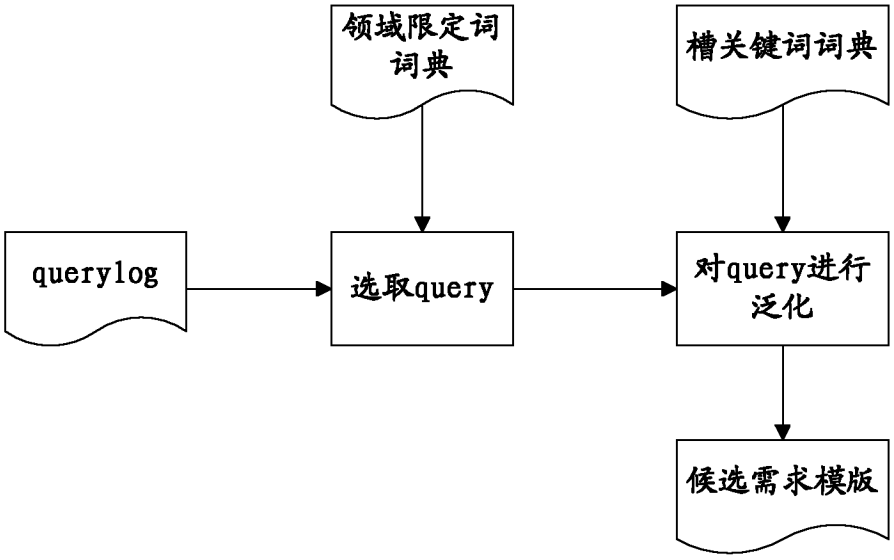

领域限定词词典
槽关键词词典
querylog
选取query
对query进行泛化
候选需求模版

图 2

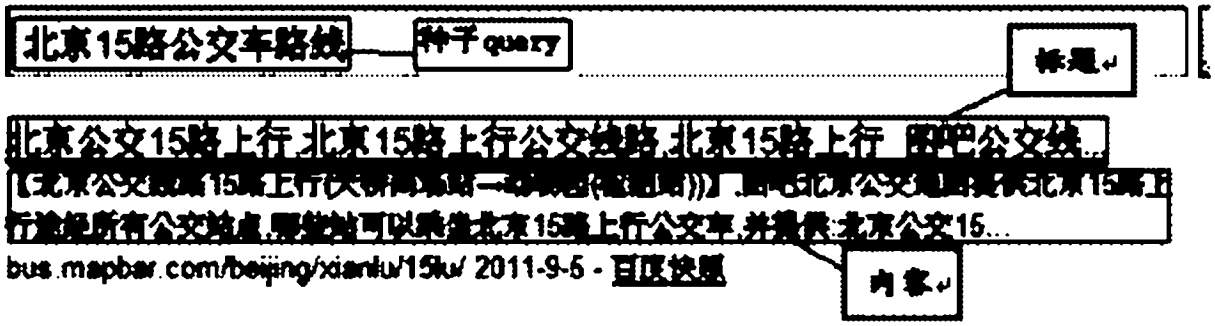

北京15路公交车路线
种子query
标题
北京公交15路上行 北京15路上行公交线路 北京15路上行 图吧公交线...
内容
bus.mapbar.com/beijing/xianlu/15lu/ 2011-9-5 - 百度快照
北京15路 北京15路公交 北京公交15路 北京15路公交车
15路线路信息：1-599路 动物园枢纽站5:00-23:00...15路站牌信息 共24站 全程...北京所有公交线路一览 北京所有公交站点一览表 关线路 地铁1号线1换车1路专线1路...
bus.aibang.com/beijing/line-15路(动物园 ... 2011-9-5 - 百度快照
北京15路 北京15路公交车 北京15路线路详情 -- 丁丁地图
丁丁地图提供北京15路公交线路各个站点名称、所在位置，并且用地图显示北京15路公交车行车路线，使你对北京15路公交线路一目了然。
www.ddmap.com/mstmap50007/g_busdetail.jsp ... 2011-8-28 - 百度快照
北京昌平15路公交车路线 | 北京公交昌平15路线路地图查询 - 北京...
北京经典网提供北京昌平15路公交路线查询服务，包括昌平15路公交停靠站点、北京昌平15路公交线路地图等。
bj.ganji.com/bus/line/110100013036.html 2011-9-9 - 百度快照

图 3

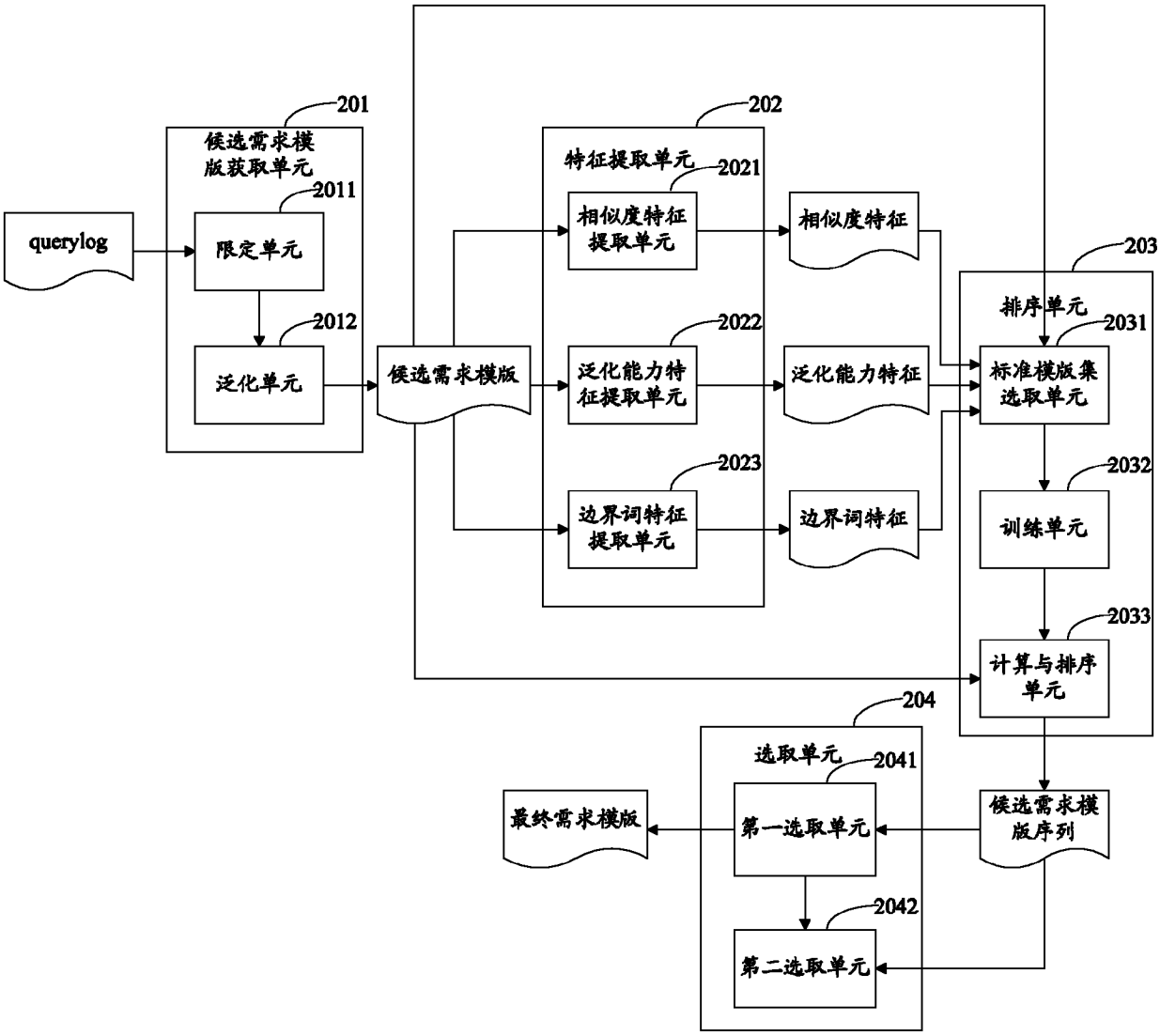

201
候选需求模版获取单元
2011
querylog
限定单元
2012
泛化单元
候选需求模版
202
特征提取单元
2021
相似度特征提取单元
相似度特征
2022
泛化能力特征提取单元
泛化能力特征
2023
边界词特征提取单元
边界词特征
203
排序单元
2031
标准模版集选取单元
2032
训练单元
2033
计算与排序单元
204
选取单元
2041
第一选取单元
2042
第二选取单元
候选需求模版序列
最终需求模版

图 4

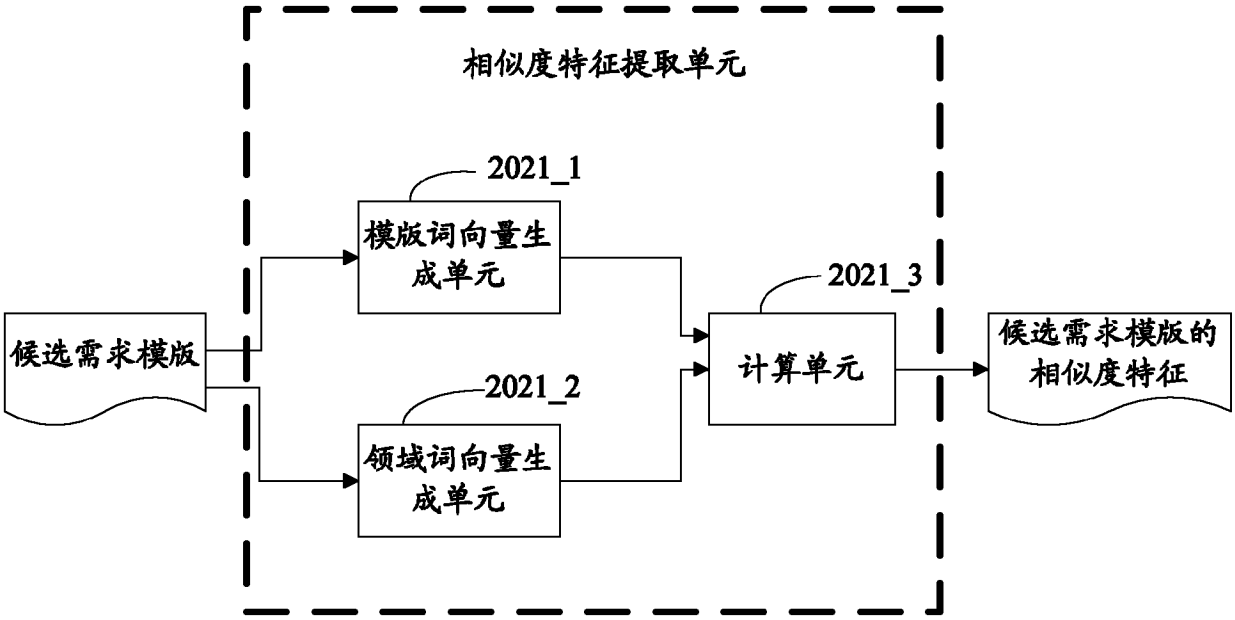

相似度特征提取单元
2021_1
模版词向量生成单元
2021_3
候选需求模版
计算单元
候选需求模版的相似度特征
2021_2
领域词向量生成单元

图 5

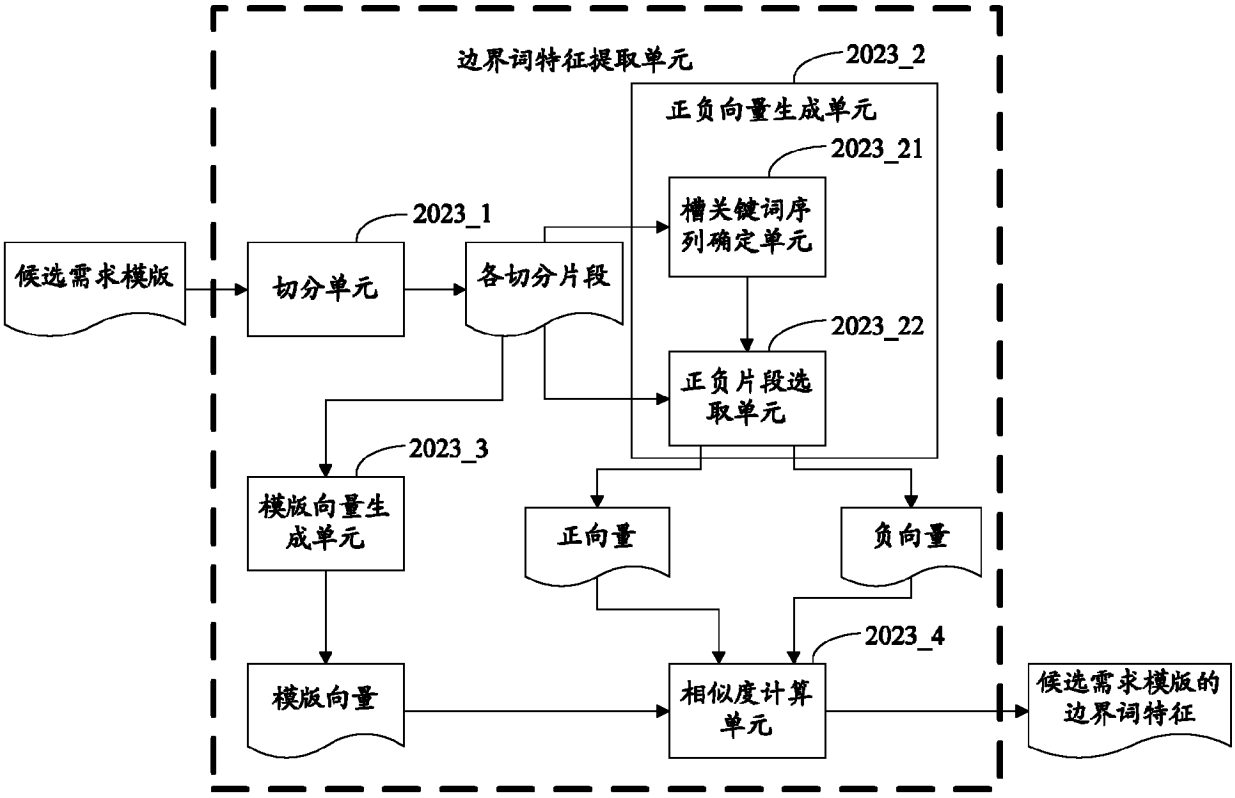

边界词特征提取单元
2023_2
正负向量生成单元
2023_21
槽关键词序列确定单元
2023_1
候选需求模版
切分单元
各切分片段
2023_22
正负片段选取单元
2023_3
模版向量生成单元
正向量
负向量
2023_4
模版向量
相似度计算单元
候选需求模版的边界词特征

图 6

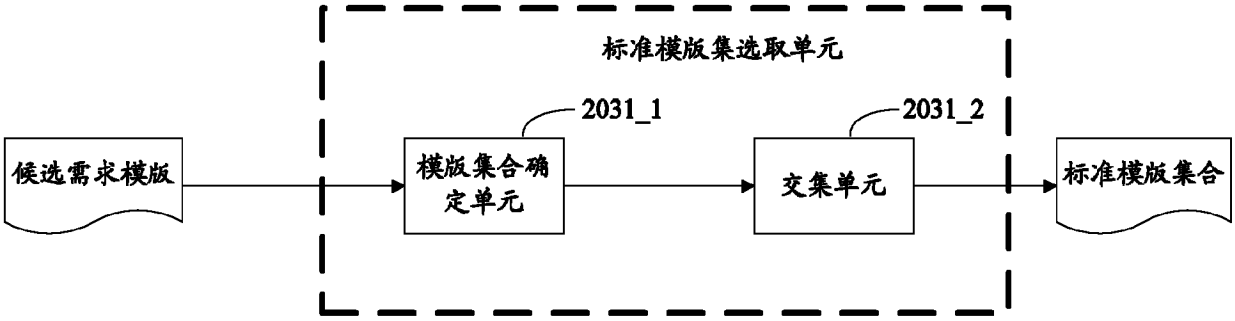


图 7