

(12) 发明专利申请

(10) 申请公布号 CN 102033877 A

(43) 申请公布日 2011. 04. 27

(21) 申请号 200910175066. 3

(22) 申请日 2009. 09. 27

(71) 申请人 阿里巴巴集团控股有限公司

地址 英属开曼群岛大开曼岛资本大厦一座
四层 847 号邮箱

(72) 发明人 林锋 康战辉 黄鹏

(74) 专利代理机构 北京集佳知识产权代理有限公司 11227

代理人 逯长明 王宝筠

(51) Int. Cl.

G06F 17/30 (2006. 01)

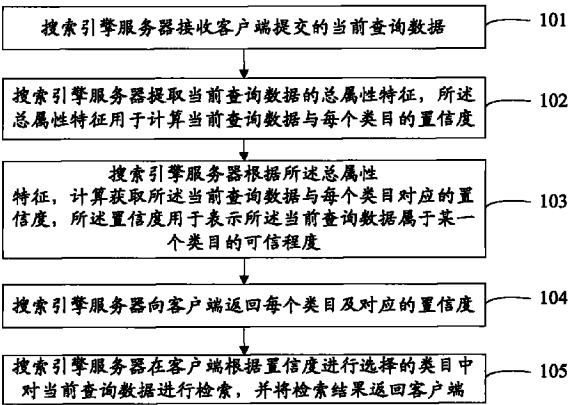
权利要求书 4 页 说明书 17 页 附图 6 页

(54) 发明名称

检索方法和装置

(57) 摘要

本申请提供了一种检索方法、装置和系统,所述方法包括:接收客户端提交的当前查询数据;提取所述当前查询数据的总属性特征,所述总属性特征可以包括:词一级特征、句法特征、语义特征、点击特征、筛选特征和会话特征;根据所述词一级特征、句法特征、语义特征、点击特征、筛选特征和会话特征,计算获取所述当前查询数据与每个类目对应的置信度,所述置信度用于表示所述当前查询数据属于某一个类目的可信程度;向客户端返回所述每个类目及对应的置信度;在客户端根据所述置信度进行选择类目中对所述当前查询数据进行检索,并将检索结果返回给客户端。在本申请中,对用户的查询数据覆盖率可以显著提升,更进一步的,提升搜索引擎服务器的性能。



1. 一种检索方法,其特征在于,该方法包括:

接收客户端提交的当前查询数据;

提取所述当前查询数据的总属性特征,所述属性特征用于计算所述当前查询数据与每个类目的置信度;

根据所述总属性特征,计算获取所述当前查询数据与每个类目对应的置信度,所述置信度用于表示所述当前查询数据属于某一个类目的可信程度;

向客户端返回所述每个类目及对应的置信度;

在客户端根据所述置信度进行选择的类目中对所述当前查询数据进行检索,并将检索结果返回给客户端。

2. 根据权利要求1所述的方法,其特征在于,所述总属性特征具体包括:词一级特征、句法特征、语义特征、点击特征、筛选特征和会话特征;所述词一级特征为所述当前查询数据中的关键词,所述句法特征为所述当前查询数据中的中心词所对应的词根,所述语义特征为所述关键词对应的语义标签,所述点击特征为所述当前查询数据在预置的点击日志中对应的至少一个点击类目,所述筛选特征为所述当前查询数据在预置的筛选日志中对应的至少一个筛选类目,所述会话特征用于表示在同一个会话中,所述当前查询数据对的前一个查询对应的最优类目。

3. 根据权利要求2所述的方法,其特征在于,所述根据所述词一级特征、句法特征、语义特征、点击特征、筛选特征和会话特征,计算获取所述当前查询数据的类目及其对应的置信度,具体包括:

将所述词一级特征、句法特征、语义特征、点击特征、筛选特征和会话特征,作为条件概率模型的给定信息;

采用条件概率模型计算在所述给定信息的情况下,所述当前查询在每个类目下的后验概率;并将所述后验概率作为当前查询数据的在该类目下的置信度。

4. 根据权利要求2所述的方法,其特征在于,所述向用户返回所述每个类目及所述每个类目对应的置信度之前,还包括:

判断所述计算获取的置信度是否大于预置的第一阈值;

如果是,则获取所述大于所述阈值的置信度以及所述置信度对应的类目;

则所述向用户返回所述每个类目及所述每个类目对应的置信度,具体包括:

向用户返回所述大于所述阈值的置信度以及所述置信度对应的类目。

5. 根据权利要求4所述的方法,其特征在于,所述判断所述计算获取的置信度是否大于预置的第一阈值之后,还包括:

当计算获取的所有类目的置信度均小于所述第一阈值,则从所有置信度中直接获取预定个数的相对较大的置信度以及对应的类目。

6. 根据权利要求2所述的方法,其特征在于,所述提取所述当前查询数据的句法特征,具体包括:

对所述当前查询数据中的关键词进行词性标注和短语识别;

从前述标注和识别的结果中获取所述当前查询数据中的名词短语,并抽取出所述名词短语的中心词;

通过词根抽取工具获取所述中心词的词根。

7. 根据权利要求 2 所述的方法,其特征在于,提取所述当前查询数据的语义特征,具体包括:

对所述当前查询数据的关键词进行类型标识;

根据所述类型标识获取所述关键词的意动词标签、产品标签、品牌标签、型号标签和/或其他标签。

8. 根据权利要求 2 所述的方法,其特征在于,提取所述当前查询数据的点击特征,具体包括:

判断在预置的点击日志中是否存在所述当前查询数据,所述点击日志用于保存查询数据与点击特征的对应关系;

如果是,则在预置的点击日志中获取当前查询数据对应的前预定个数的点击类目;所述点击类目为用户在检索该查询数据之后点击的相应产品所属的类目;

如果否,则在点击日志中查询与所述当前查询数据最接近的点击查询数据,所述点击查询数据与所述当前查询数据的编辑距离在点击日志中最小;并将所述点击查询数据的对应的前预定个数的类目作为所述当前查询数据的点击特征。

9. 根据权利要求 2 所述的方法,其特征在于,提取所述当前查询数据的筛选特征,具体包括:

判断在预置的筛选日志中是否存在所述当前查询数据,所述筛选日志用于保存查询数据与筛选特征的对应关系;

如果是,则在预置的筛选日志中获取当前查询数据对应的前预定个数的筛选类目;所述筛选类目表示所述当前查询数据在搜索引擎的映射表中所关联的类目;

如果否,则在筛选日志中查询与所述当前查询数据最接近的筛选查询数据,所述筛选查询数据与所述当前查询数据的编辑距离在筛选日志中最小;并将所述筛选查询数据的对应的前预定个数的类目作为所述当前查询数据的筛选特征。

10. 根据权利要求 2 所述的方法,其特征在于,提取所述当前查询数据的会话特征,具体包括:

在同一个会话中,获取所述当前查询数据的上一个查询数据;

获取到所述上一个查询数据对应的最优分类结果所属的类目,并以所述最优分类结果所属的类目为所述当前查询数据的会话特征。

11. 一种检索方法,其特征在于,该方法包括:

接收客户端提交的当前查询数据;

提取所述当前查询数据的总属性特征,所述属性特征用于计算所述当前查询数据的置信度;

根据所述总属性特征,计算获取所述当前查询数据与每个类目对应的置信度,所述置信度用于表示所述当前查询数据属于某一个类目的可信程度;

在计算得到的最大的置信度对应的类目中,对所述当前查询数据进行检索,并将检索结果返回给客户端。

12. 一种检索装置,其特征在于,该装置包括:

接收模块,用于接收客户端提交的当前查询数据;

提取模块,用于提取所述当前查询数据的总属性特征,所述属性特征用于计算所述当

前查询数据的置信度；

计算获取模块,用于根据所述总属性特征,计算获取所述当前查询数据与每个类目对应的置信度,所述置信度用于表示所述当前查询数据属于某一个类目的可信程度；

返回模块,用于向客户端返回所述每个类目及对应的置信度；

第一检索模块,用于在客户端根据所述置信度进行选择类目中对所述当前查询数据进行检索。

13. 根据权利要求 12 所述的装置,其特征在于,所述提取模块具体用于提取所述当前查询数据的词一级特征、句法特征、语义特征、点击特征、筛选特征和会话特征;所述词一级特征为所述当前查询数据中的关键词,所述句法特征为所述当前查询数据中的中心词所对应的词根,所述语义特征为所述关键词对应的语义标签,所述点击特征为所述当前查询数据在预置的点击日志中对应的至少一个点击类目,所述筛选特征为所述当前查询数据在预置的筛选日志中对应的至少一个筛选类目,所述会话特征用于表示在同一个会话中,所述当前查询数据对的前一个查询对应的最优类目。

14. 根据权利要求 13 所述的装置,其特征在于,所述计算获取模块具体包括:

设置给定信息子模块,用于将所述词一级特征、句法特征、语义特征、点击特征、筛选特征和会话特征,作为条件概率模型的给定信息;

计算置信度子模块,用于采用条件概率模型计算在所述给定信息的情况下,所述当前查询在每个类目下的后验概率;并将所述后验概率作为当前查询数据的在该类目下的置信度。

15. 根据权利要求 13 所述的装置,其特征在于,所述装置还包括:

第一判断模块,用于判断所述计算获取的置信度是否大于预置的第一阈值;

第一获取模块,用于当所述第一判断模块的结果为是时,获取所述大于所述阈值的置信度以及所述置信度对应的类目;

则所述返回模块,具体用于:向用户返回所述大于所述阈值的置信度以及所述置信度对应的类目。

16. 根据权利要求 15 所述的装置,其特征在于,所述装置还包括:

第二获取模块,用于当计算获取的所有类目的置信度均小于所述第一阈值,则从所有置信度中直接获取预定个数的相对较大的置信度以及对应的类目。

17. 根据权利要求 13 所述的装置,其特征在于,所述提取模块具体包括:

标注识别子模块,用于对所述当前查询数据中的关键词进行词性标注和短语识别;

抽取中心词子模块,用于从前述标注和识别的结果中获取所述当前查询数据中的名词短语,并抽取所述名词短语的中心词;

词根抽取工具,用于获取所述中心词的词根。

18. 根据权利要求 13 所述的装置,其特征在于,所述提取模块具体包括:

类型标识子模块,用于对所述当前查询数据的关键词进行类型标识;

获取标签子模块,用于根据所述类型标识获取所述关键词的意动词标签、产品标签、品牌标签、型号标签和/或其他标签。

19. 根据权利要求 13 所述的装置,其特征在于,所述提取模块具体包括:

第一判断子模块,用于判断在预置的点击日志中是否存在所述当前查询数据,所述点

击日志用于保存查询数据与点击特征的对应关系；

第一获取子模块,用于当所述第一判断子模块的结果为是时,在预置的点击日志中获取当前查询数据对应的前预定个数的点击类目;所述点击类目为用户在检索该查询数据之后点击的相应产品所属的类目;

第一查询获取子模块,用于当所述第一判断子模块的结果为否时,在点击日志中查询与所述当前查询数据最接近的点击查询数据,所述点击查询数据与所述当前查询数据的编辑距离在点击日志中最小;并将所述点击查询数据的对应的前预定个数的类目作为所述当前查询数据的点击特征。

20. 根据权利要求 13 所述的装置,其特征在于,所述提取模块具体包括:

第二判断子模块,用于判断在预置的筛选日志中是否存在所述当前查询数据,所述筛选日志用于保存查询数据与筛选特征的对应关系;

第二获取子模块,用于当所述第二判断子模块的结果为是时,在预置的筛选日志中获取当前查询数据对应的前预定个数的筛选类目;所述筛选类目表示所述当前查询数据在搜索引擎的映射表中所关联的类目;

第二查询获取子模块,用于当所述第二判断子模块的结果为否时,在筛选日志中查询与所述当前查询数据最接近的筛选查询数据,所述筛选查询数据与所述当前查询数据的编辑距离在筛选日志中最小;并将所述筛选查询数据的对应的前预定个数的类目作为所述当前查询数据的筛选特征。

21. 根据权利要求 13 所述的装置,其特征在于,所述提取模块具体包括:

第三获取子模块,用于在同一个会话中,获取所述当前查询数据的上一个查询数据;

第四获取子模块,用于获取到所述上一个查询数据对应的最优分类结果所属的类目,并以所述最优分类结果所属的类目为所述当前查询数据的会话特征。

22. 一种检索装置,其特征在于,该装置包括:

接收模块,用于接收客户端提交的当前查询数据;

提取模块,用于提取所述当前查询数据的总属性特征,所述属性特征用于计算所述当前查询数据的置信度;

计算获取模块,用于根据所述总属性特征,计算获取所述当前查询数据与每个类目对应的置信度,所述置信度用于表示所述当前查询数据属于某一个类目的可信程度;

第二检索模块,用于在计算得到的最大的置信度对应的类目中,对所述当前查询数据进行检索。

检索方法和装置

技术领域

[0001] 本申请涉及网络数据处理领域,特别涉及一种检索方法和装置。

背景技术

[0002] 搜索引擎(Search engine)已经成为互联网领域用户获得信息的一种常用手段。从用户的角度看,搜索引擎提供一个页面,用户在页面输入关键词语,提交给搜索引擎后,搜索引擎就会返回跟用户输入的内容相关的信息。用户输入的关键词语可以成为查询数据,即是用户通过查询数据来查询自己感兴趣的相关内容。其中,在电子商务领域,用户输入的查询数据有一个对应的类目,例如,运动、娱乐等。

[0003] 对于搜索引擎来讲,如果能通过用户的查询数据,在进行检索之前,准确识别查询数据对应的类目,并在该类目中检索查询数据,则不仅准确的缩小了搜索范围,还能够更加准确的向用户推荐相关产品。在现有技术中,有实现分类检索方法主要采取查表的方式,即在后台人工构建一个查询数据与行业类目的映射表,例如表 1。

[0004]

查询词	类目1	类目2	类目3
电动自行 车Electric bike	交通工具 Transportation	运动/娱乐 Sports & Entertainment	汽车/摩托车 Automobiles & Motorcycles
...	...	...	...

[0005] 表 1

[0006] 例如,当用户输入的查询数据为“电动自行车”,则搜索引擎首先将该查询数据在该映射表中进行匹配,并将匹配获取到的 3 个类目均返回给客户端,那么用户就可以根据这 3 个类目进行选择,从而由搜索引擎在自己感兴趣的那一个类目中进行检索,并接收搜索引擎返回的检索结果。

[0007] 从上述过程中可以看出,由于映射表中的查询数据和对应的类目都是由人工制定的,因此能处理的查询词的范围就受到映射表中查询数据所属类目多少的限制,只有用户输入的查询词必须与映射表中的查询数据完全匹配,搜索引擎才能将搜索结果返回给用户,因此现有技术中的这种检索方法对于查询数据的覆盖率比较低;更进一步的,由于映射表中的查询数据和对应的类目都是由人工制定的,因此,后续对映射表的维护成本较高,并且,如果行业类目一更改,原有映射表就会无效,可扩展性较差。上述过程还存在一个技术

问题是,搜索引擎在获取到与查询数据匹配的类目后,先将该等类目传输给客户端,然后在客户端对该等类目做出选择后才能进行相应的检索,这无疑增大了搜索引擎服务器与客户端之间的数据传输量,增加了服务器的负担,降低了数据在网络中的传输速率。

[0008] 总之,目前需要本领域技术人员迫切解决的一个技术问题就是:如何能够创新的提出一种检索方法,以解决现有技术中基于映射表进行检索时,对用户的查询数据覆盖率较低的问题,以及服务器负担过重、网络传输速率慢的问题。

发明内容

[0009] 本申请所要解决的技术问题是提供一种检索方法,用以解决现有技术中基于映射表进行检索时,对用户的查询数据覆盖率较低的问题,更进一步的,还可以降低成本,优化可扩展性。

[0010] 本申请还提供了一种检索装置,用以保证上述方法在实际中的实现及应用。

[0011] 为了解决上述问题,本申请公开了一种检索方法,包括:

[0012] 接收客户端提交的当前查询数据;

[0013] 提取所述当前查询数据的总属性特征,所述属性特征用于计算所述当前查询数据与每个类目的置信度;

[0014] 根据所述词一级特征、句法特征、语义特征、点击特征、筛选特征和会话总属性特征,计算获取所述当前查询数据与每个类目对应的置信度,所述置信度用于表示所述当前查询数据属于某一个类目的可信程度;

[0015] 向客户端返回所述每个类目及对应的置信度;

[0016] 在客户端根据所述置信度进行选择的类目中对所述当前查询数据进行检索,并将检索结果返回给客户端。

[0017] 本申请公开了一种检索方法,包括:

[0018] 接收客户端提交的当前查询数据;

[0019] 提取所述当前查询数据的总属性特征,所述属性特征用于计算所述当前查询数据的置信度;

[0020] 根据所述总属性特征,计算获取所述当前查询数据与每个类目对应的置信度,所述置信度用于表示所述当前查询数据属于某一个类目的可信程度;

[0021] 在计算得到的最大的置信度对应的类目中,对所述当前查询数据进行检索,并将检索结果返回给客户端。

[0022] 本申请公开了一种检索装置,包括:

[0023] 接收模块,用于接收客户端提交的当前查询数据;

[0024] 提取模块,用于提取所述当前查询数据的总属性特征,所述属性特征用于计算所述当前查询数据的置信度;

[0025] 计算获取模块,用于根据所述总属性特征,计算获取所述当前查询数据与每个类目对应的置信度,所述置信度用于表示所述当前查询数据属于某一个类目的可信程度;

[0026] 返回模块,用于向客户端返回所述每个类目及对应的置信度;

[0027] 第一检索模块,用于在客户端根据所述置信度进行选择的类目中对所述当前查询数据进行检索。

- [0028] 本申请还提供了一种检索装置,该装置包括:
- [0029] 接收模块,用于接收客户端提交的当前查询数据;
- [0030] 提取模块,用于提取所述当前查询数据的总属性特征,所述属性特征用于计算所述当前查询数据的置信度;
- [0031] 计算获取模块,用于根据所述总属性特征,计算获取所述当前查询数据与每个类目对应的置信度,所述置信度用于表示所述当前查询数据属于某一个类目的可信程度;
- [0032] 第二检索模块,用于在计算得到的最大的置信度对应的类目中,对所述当前查询数据进行检索。
- [0033] 与现有技术相比,本申请包括以下优点:
- [0034] 在本申请中,通过对当前查询数据的总属性特征的提取,可以根据该总属性特征获取到当前查询数据属于某一个类目的置信度,这样对于每
- [0035] 一个类目都有一个置信度与其对应,不管当前查询数据之前有没有出现过,都可以由条件概率模型计算出其归属于某一个类目的置信度。即便后台映射表中没有预先保存某个或某些类目,只要用户根据搜索引擎服务器返回的置信度选择了某个类目,或者搜索引擎服务器直接根据置信度最高的值对应类目对当前查询数据进行检索,就可以在提高当前查询数据的覆盖率的广泛性的同时,在一定程度上减少搜索引擎服务器的工作,提升搜索引擎服务器的工作效率和性能,提高网络中数据传输的速率,同时也能增加用户在数据查询方面的搜索体验。当然,实施本申请的任一产品并不一定需要同时达到以上所述的所有优点。

附图说明

- [0036] 为了更清楚地说明本申请实施例中的技术方案,下面将对实施例描述中所需要使用的附图作简单地介绍,显而易见地,下面描述中的附图仅仅是本申请的一些实施例,对于本领域普通技术人员来讲,在不付出创造性劳动性的前提下,还可以根据这些附图获得其他的附图。
- [0037] 图 1 是本申请的一种检索方法实施例 1 的流程图;
- [0038] 图 2 是本申请中一种产品信息的界面示意图;
- [0039] 图 3 是本申请中搜索引擎的返回的类目的界面示意图;
- [0040] 图 4 是本申请的一种检索方法实施例 2 的流程图;
- [0041] 图 5 是本申请的一种检索方法实施例 3 的流程图;
- [0042] 图 6 是本申请中查询分类器输入和输入内容的结构示意图;
- [0043] 图 7 是本申请的一种检索方法实施例四的流程图;
- [0044] 图 8 是本申请的一种检索装置实施例 1 的结构框图;
- [0045] 图 9 是本申请的一种检索装置实施例 2 的结构框图;
- [0046] 图 10 是本申请的一种检索装置实施例 3 的结构框图;
- [0047] 图 11 是本申请的一种检索装置实施例 4 的结构框图。

具体实施方式

- [0048] 下面将结合本申请实施例中的附图,对本申请实施例中的技术方案进行清楚、完

整地描述,显然,所描述的实施例仅仅是本申请一部分实施例,而不是全部的实施例。基于本申请中的实施例,本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例,都属于本申请保护的范围。

[0049] 本申请可用于众多通用或专用的计算装置环境或配置中。例如:个人计算机、服务器计算机、手持设备或便携式设备、平板型设备、多处理器装置、包括以上任何装置或设备的分布式计算环境等等。

[0050] 本申请可以在由计算机执行的计算机可执行指令的一般上下文中描述,例如程序模块。一般地,程序模块包括执行特定任务或实现特定抽象数据类型的例程、程序、对象、组件、数据结构等等。也可以在分布式计算环境中实践本申请,在这些分布式计算环境中,通过通信网络而被连接的远程处理设备来执行任务。在分布式计算环境中,程序模块可以位于包括存储设备在内的本地和远程计算机存储介质中。

[0051] 本申请的主要思想之一可以包括,在搜索引擎接收到客户端提交的当前查询数据时,首先提取所述当前查询数据的总属性特征,在具体实施例中,所述总属性特征可以包括:词一级特征、句法特征、语义特征、点击特征、筛选特征和会话特征;所述词一级特征为所述当前查询数据中的关键词,所述句法特征为所述当前查询数据中的中心词所对应的词根,所述语义特征为所述关键词对应的语义标签,所述点击特征为所述当前查询数据在预置的点击日志中对应的至少一个点击类目,所述筛选特征为所述当前查询数据在预置的筛选日志中对应的至少一个筛选类目,所述会话特征用于表示在同一个会话中,所述当前查询数据的前一个查询对应的最优类目;根据所述词一级特征、句法特征、语义特征、点击特征、筛选特征和会话特征,计算获取所述当前查询数据与每个类目对应的置信度,所述置信度用于表示所述当前查询数据属于某一个类目的可信程度,并向客户端返回所述每个类目及对应的置信度;最后在客户端根据所述置信度进行选择类目中对所述当前查询数据进行检索。在本申请实施例中,在一定程度上减少了搜索引擎服务器的工作,提升了搜索引擎服务器的工作效率和性能,也能增加用户在数据查询方面的搜索体验。并且,和现有技术相比,也能够具有更广泛的覆盖率,因为不管当前查询数据之前有没有出现过,都可以由条件概率模型计算出其归属于某一个类目的置信度。

[0052] 参考图 1,示出了本申请一种检索方法实施例 1 的流程图,可以包括以下步骤:

[0053] 步骤 101:搜索引擎服务器通过浏览器接收客户端提交的当前查询数据。

[0054] 在本申请实施例中,客户端首先向搜索引擎提交当前输入的查询数据,其中,查询数据可以任意。

[0055] 步骤 102:搜索引擎服务器提取所述当前查询数据的总属性特征,所述总属性特征用于计算当前查询数据与每个类目之间的置信度。

[0056] 在实际中,所述总属性特征具体可包括:词一级特征、句法特征、语义特征、点击特征、筛选特征和会话特征。

[0057] 在本申请实施例中,所述当前查询数据的一些特征被提取出来用于后续条件概率的运算,所述当前查询数据共需要提取六种特征,所述词一级特征为所述当前查询数据中的关键词,即是非停用词对应的词根被抽取为词一级特征;例如,当前查询数据为“find Nokia N95phones”时,所述词一级的特征包括:find、Nokia、N95 和 phone,如果是中文“查询诺基亚手机”,则词一级的特征为:查询、诺基亚和手机,具体实现时可以通过分词器来获

取到当前查询数据的关键词;所述句法特征为所述当前查询数据中的中心词所对应的词根,即是当前查询数据中的句法级别的中心词所对应的词根被抽取为句法特征,中心词可以在一定程度上反映了当前查询数据的意图核心词;例如,对于“find Nokia N95phones”,中心词为 phones,则抽取词根“phon”为句法特征;所述语义特征为所述关键词对应的语义标签,所述当前查询数据中的关键词所对应的词根所对应的语义标签被抽取为语义的特征,在本申请实施例中所采用的语义标签有:意动词标签(Action),用于标识该词为一意愿动词;产品标签(Product),用于标识该词为某一产品词;品牌标签(Brand),用于标识该词为某一品牌词;型号标签(Model),用于标识该词为某一型号词;其他标签(Other),用于其他词所对应的标签;例如,对于“find Nokia N95phones”,对应的语义特征有“Action: find”,“Brand:Nokia”,“Model:N95”,和“Product:phon”。

[0058] 所述点击特征为所述当前查询数据在预置的点击日志中对应的至少一个点击类目,所述预置的点击日志是在搜索引擎服务器上保存的一个数据库,用于保存用户输入查询数据之后,所点击的产品所属的类目。例如,用户输入的查询词为“mp3”,在搜索引擎向用户返回的结果中用户点击了某一个产品,与该产品相关的部分信息有:标题为“Digital Mp3Player”, With Lcd And Fm(IMC278A),可以参见图2所示,为该产品信息在网页上的界面示意图。其中,关键词为“mp3player, music player”,该产品所属的类目为“mp3players”,即是对于查询词“mp3”来说,其点击特征为“mp3players”。所述筛选特征为所述当前查询数据在预置的筛选日志中对应的至少一个筛选类目;在电子商务网站中,例如“www.alibaba.com”,如果用户输入的查询词为 mp3,则搜索引擎在返回搜索到的产品链接的同时,返回这些产品对应的类目(可以参见图3所示,为对应类目的界面示意图),以便用户可以通过点击这些类目的链接来定位所检索的范围;所述筛选日志就用于保存用户输入的查询词和被点击的类目的对应关系。所述会话特征用于表示在同一个会话(Session)中,所述当前查询数据的前一个查询对应的最优类目;在一个有效的时间内,在同一个浏览器内的一系列的查询行为属于一个会话。即是用户打开一个网页之后,在未关这个网页之前,就认为所输入的查询词都属于同一个浏览器。例如,一个会话中包含了三个查询,按照时间先后次序为:“phone”、“Nokia phone”和“Nokia n95”;则“Nokia n95”的 Session 特征为“Nokia phone”所对应的最好的分类类目,而“phone”的 Session 特征为空(NULL)。

[0059] 需要说明的是,不管所述当前查询数据的总属性特征包括了几个具体特征,本领域技术人员都可以实现本发明的目的,因此,所述总属性特征包括的具体特征的个数及其形式和内容不受本发明实施例的限制。

[0060] 步骤103:搜索引擎服务器根据所述总属性特征,计算获取所述当前查询数据与每个类目对应的置信度,所述置信度用于表示所述当前查询数据属于某一个类目的可信程度。

[0061] 其中,根据步骤102中得到的当前查询数据的特征,可以认为是条件概率模型中的给定条件,而所述置信度为一个0~1之间的数值,用于表示在给定条件下某一件事发生的概率。因为条件概率的含义是求在一件事(条件)发生的前提下,另外一件事发生的概率,因此,本实施例中,当前查询数据的词一级特征、句法特征、语义特征、点击特征、筛选特征和会话特征则被认为是条件,而对于某个类目计算出的置信度,则可以理解为当前查询数据属于该某个类目的概率是多少。因此可以看出,置信度越大,即是当前查询数据越有可

能属于对应的某个类目。

[0062] 优选情况下,可以采用最大熵模型来对具体的类目计算对应的置信度。还需要说明的是,本申请实施例在对每个类目计算置信度时,主要采用的是机器学习方法,因为该方法可以通过一定规模的训练查询数据,来模拟用户的查询行为,计算获取到查询数据归类于某一类目的概率。举例来说,对于词一级特征来讲,如果在训练数据中有这样的记录“当前查询数据中包括 phon 时,属于电子类”,那么对于当前查询数据中的词一级特征“phon”来说,当计算置信度的类目为电子类时,则该电子类对应的置信度在计算时对应的词一级特征值就为 1,当计算除了电子类之外的其他类目的置信度时,词一级特征值则为 0。以此类推,可以根据训练数据中的内容分别计算当前查询数据中的词一级特征、句法特征、语义特征、点击特征、筛选特征和会话特征对应的值,并根据条件概率模型计算存在这些特征值的情况下,当前查询数据归属于某一类目的概率值。

[0063] 步骤 104:搜索引擎服务器向客户端返回所述每个类目及对应的置信度。

[0064] 假设实际中的分类有 40 种,则需要对每一个类目都进行计算,在实际中,就计算出对应的 40 个置信度,搜索引擎服务器分别将每个类目及对应的置信度返回给用户,服务器在返回这 40 个置信度的时候,可以优先进行排序,然后将排序后的置信度内容返回给用户,以便于用户根据自己所感兴趣的类目定位检索范围,以更加精准的检索到需要的产品信息。

[0065] 步骤 105:搜索引擎服务器在客户端根据所述置信度进行选择类目中对所述当前查询数据进行检索,并将检索结果返回至客户端。

[0066] 搜索引擎服务器接收用户选择的类目信息,并在后台在所述选择的类目范围内检索和当前查询数据相关的信息。在本实施例中,可以对客户端提交的当前查询数据进行特征的提取,其中特征涉及到当前查询数据的词一级特征、句法特征、语义特征、点击特征、筛选特征和会话特征,根据这些特征在训练数据中进行匹配,从而得到表示匹配结果的特征值,然后利用条件概率模型将所述特征值计算出表示当前查询数据属于某一类目的置信度,从而在能够将置信度和类目的对应关系返回至客户端的基础上,根据客户端所选的类目再对当前查询数据进行检索,这样就在一定程度上减少了搜索引擎服务器的工作,提升了搜索引擎服务器的工作效率和性能,也能增加用户在数据查询方面的搜索体验。

[0067] 参考图 4,示出了本申请一种检索方法实施例 2 的流程图,可以包括以下步骤:

[0068] 步骤 401:查询分类器接收客户端提交的当前查询数据。

[0069] 在本实施例中,可以构建查询分类器,该查询分类器可以对当前查询数据提取相关的特征,并根据提取的特征计算类目对应的置信度,其中,所述查询分类器可以采用软件或硬件实现,可以作为搜索引擎服务器新增的一项功能或者作为实体与搜索引擎服务器相连。

[0070] 步骤 402:查询分类器提取所述当前查询数据的词一级特征。

[0071] 本实施例中提取词一级特征和实施例 1 中的介绍一致。

[0072] 步骤 403:查询分类器提取所述当前查询数据的句法特征。

[0073] 在本实施例中,可以首先采用词性标注工具对当前查询数据标注每一个词的词性,例如:“phones”的词性为 nns,表示该词为复数名词;然后采用短语识别工具对当前查询数据做短语识别,例如,当前查询数据

[0074] “find Nokia N95phones”由两个短语组成,分别是动词短语“find”和名词短语“Nokia N95phones”;再通过规则从名词短语中抽取出中心词“phones”,最后通过词根抽取工具,抽取“phones”的词根为“phon”,其中,所述规则可以是例如“如果一个查询词只有一个名词短语,则该名词短语中最后一个名词为中心词”等预置的中心词抽取方式。其中,词性标注工具、短语识别工具和词根抽取工具是目前已经开发并进行使用的工具。步骤 404: 查询分类器提取所述当前查询数据的语义特征。

[0075] 基于搜索引擎服务器中保存的查询日志,可以将电子商务网上发布的产品信息生成品牌名的表格,例如:品牌“Nokia”,还可以生成型号名的表格,例如:型号“N95”,同时生成基本的产品名的表格,例如:产品名“phone”。因此查询分类器根据上述的各种表格可以对当前查询数据的关键词分别标注出所含有品牌,型号,产品名等等。然后抽取关键词的词根作为语义特征。

[0076] 步骤 405: 查询分类器提取所述当前查询数据的点击特征。

[0077] 查询分类器直接从预置的点击日志中根据当前查询数据和点击特征的对应关系,获取到对应的点击特征即可。在实际应用中,不同的产品可能包含相同的类目信息,对于同一个查询词用户也可能会点击多个产品对应的网页链接,因此在点击日志中同一个查询词可能对应多个类目。而搜索引擎服务器可以记录用户输入的查询词,以及用户在搜索引擎推荐的产品中都点击了哪些产品的网页链接,就记录这些产品所述的不同类目,从而将查询词和类目之间建立对应关系,根据查询词、类目以及两者的对应关系构建点击日志。

[0078] 具体的,在预置的点击日志中,首先统计当前查询数据在点击日志中对应的前 n 个频率最高的类目(n 的取值优选情况可以取为 3),并把这 n 个类目作为点击特征;如果在点击日志中匹配不到当前查询数据,则在点击日志中寻找一个与当前查询数据最相近的查询,并把最相近的查询在点击日志中对应的前 n 个频率最高的类目作为当前查询数据的点击特征。在点击日志中,寻找与当前查询数据的编辑距离最少的查询数据作为最相近的查询。所谓编辑距离是指两个查询数据经过插入、删除或者更改等操作后能够相同所花费的操作的代价,代价越高则编辑距离越大,反之则编辑距离越小。

[0079] 步骤 406: 查询分类器提取所述当前查询数据的筛选特征。

[0080] 查询分类器直接从预置的筛选日志中根据当前查询数据和筛选特征的对应关系,获取到对应的筛选特征即可。在实际应用中,对于同一个查询词搜索引擎也可以根据后台的映射表从而推荐多个类目,因此在筛选日志中同一个查询词可能对应多个类目。搜索引擎服务器可以根据映射表中记录的类目,以及与查询词之间的对应关系,从而生成筛选日志。

[0081] 具体的,在预置的筛选日志中,统计一个查询在筛选日志中对应的前 n (n 的取值优选情况可以取为 3) 个频率最高的类目,并把这 n 个类目作为筛选特征;如果当前查询数据在筛选日志中没有出现,则在筛选日志中寻找一个与当前查询数据最相近的查询,并把最相近的查询在筛选日志中对应的前 n 个频率最高的类目作为该查询的筛选特征。寻找与当前查询数据的编辑距离最少的查询数据作为最相近的查询。所谓编辑距离是指两个查询数据经过插入、删除或者更改等操作后能够相同所花费的操作的代价,代价越高则编辑距离越大,反之则编辑距离越小。

[0082] 步骤 407: 查询分类器提取所述当前查询数据的会话特征。

[0083] 在电子商务领域中,一个 Session 一般围绕同一个主题进行。在同一个 Session 内的查询数据所属的类目一般相同或者具有相关性。例如,在实际中一个 Session 包含了三个查询,按照时间先后次序为:“phone”、“Nokia phone”和“Nokia n95”,这三个查询可以反映用户的查询意图,并且这三个查询对应的所属类目具有相关性。因此,会话特征也可以从一定程度上反映当前查询数据与其属于某一个类别的可信程度。

[0084] 步骤 408:查询分类器将所述词一级特征、句法特征、语义特征、点击特征、筛选特征和会话特征,作为条件概率模型的给定信息。

[0085] 假设某一个类目为 C 类,则当前查询数据被划分为该类目 C 的置信度就可以用如下方式表示:在给定相关信息 x 的条件下,当前查询数据确定属于该类目 C 的后验概率 $p(\text{cis correct} | x)$ 。其中,所述 x 代表的信息即是给定信息,在本实施例中为当前查询数据的词一级特征、句法特征、语义特征、点击特征、筛选特征和会话特征。

[0086] 步骤 409:查询分类器采用条件概率模型计算在所述给定信息的情况下,所述当前查询在每个类目下的后验概率;并将所述后验概率作为当前查询数据的在该类目下的置信度。

[0087] 条件概率模型可以训练当前查询数据所得到的特征的权重,通过训练得到的对应的特征值和权重,来利用条件概率模型计算当前查询数据被确定为某一类目的条件概率,即是置信度。

[0088] 条件概率模型优选情况下可以选择最大熵模型,其公式如下所示:

$$[0089] \quad p(y|x) = \frac{1}{Z(x)} \exp \left(\sum_j \lambda_j f_j(x, y) \right)$$

[0090] 其中 $y \in \{\text{c is correct}, \text{c is incorrect}\}$; x 是当前查询数据所对应的给定信息,在本实施例中包括:词一级特征、句法特征、语义特征、点击特征、筛选特征和会话特征; f_j 是最大熵模型对应的特征值,例如,即是当 x 为词一级特征时,如果在训练数据中能够匹配到“当前查询数据的关键词包括 phon 时,所属类目为 C”,则特征 f_j 在条件“当前查询数据的关键词包括 phon 时”的值就为 1,否则为 0; λ_j 是当前查询数据的各个特征对应的权重,可以通过模型训练得到的; $Z(x)$ 是归一化因子,也通过模型训练得到。

[0091] 在实际中,还可以采用线性回归等机器学习模型来建立条件概率模型。在实际中,还可以采用支持向量机模型,虽然不是条件概率模型,但是计算的分值也可以用来作为置信度。

[0092] 步骤 4010:查询分类器判断所述计算获取的置信度是否大于预置的第一阈值,如果是,则进入步骤 2012,如果不是,则进入步骤 2011。

[0093] 在计算获取了当前查询数据归属于所有类目的置信度之后,查询分类可以预置第一阈值,例如,可以将第一阈值设置为 0.5,当置信度大于 0.5,则认为对应的该类目的可信程度较高。

[0094] 步骤 4011:记录所述小于所述第一阈值的置信度,并当计算获取的所有类目的置信度均小于所述第一阈值,则从所有置信度中直接获取预定个数的相对较大的置信度以及对应的类目,进入步骤 2013。

[0095] 等到所有的置信度都已经判断完毕,如果所有类目对应的置信度都不大于第一阈值,则查询分类器直接可以对所有置信度按照从大到小进行排序,并获取前预定个数的相

对较大的置信度及其对应的类目。

[0096] 步骤 4012: 查询分类器获取所述大于所述阈值的置信度以及所述置信度对应的类目。

[0097] 步骤 4013: 查询分类器向客户端返回所述大于所述阈值的置信度及其对应的类目, 或者返回所述预定个数的置信度及其对应的类目。

[0098] 查询分类器可以将所有大于第一阈值的置信度及其对应类目都返回给客户端, 根据置信度还可以从高至低排序展示; 或者查询分类器还可以会把前预定个数的 (例如 8 个) 置信度及其对应类目在客户端上进行展示。

[0099] 步骤 4014: 查询分类器在客户端根据所述置信度进行选择的类目中对所述当前查询数据进行检索。

[0100] 在本实施例中, 可以利用现有的工具, 例如分词器、词性标注工具等, 这样就降低了实施成本, 并且对于基于条件概率的机器学习方法来讲, 其覆盖率相对较高, 不管当前查询数据之前有没有出现过, 都可以由条件概率模型根据提取出的各种特征得到所有可能相关的类目的置信度。这样就比现有技术中的表格可以应用的更为广泛, 并且准确。

[0101] 参考图 5, 示出了本申请一种检索方法实施例 3 的流程图, 本实施例可以理解为将本申请的检索方法应用于实际中的一个具体例子, 可以包括以下步骤:

[0102] 步骤 501: 查询分类器获取用户提交的当前查询数据中的关键词。

[0103] 本实施例中的查询分类器需要集成分词器, 利用分词器来获取当前查询数据中的关键词。当前查询数据为“find Nokia N95phones”时, 例如对于类目 C 来说, 如果 $f(x, y)$ 为 1 的条件包括: “当前查询词对应的类目为 C, 且查询词包括 find 时”、“当前查询词对应的类目为 C, 且查询词包括 Nokia 时”, 以及“当前查询词对应的类目为 C, 且查询词包括 N95 时”, 其他条件下 $f(x, y)$ 为 0, 那么对于“find Nokia N95phones”提取词一级的特征时, 就认为 $f(x = \text{find}, \text{“find Nokia N95phones”} \in \text{类目 C})$ 的值为 1; 同理, $f(x = \text{Nokia}, \text{“find Nokia N95phones”} \in \text{类目 C})$ 的值也为 1, $f(x = \text{N95}, \text{“find Nokia N95phones”} \in \text{类目 C})$ 的值也为 1, 而 $f(x = \text{phon}, \text{“find Nokia N95phones”} \in \text{类目 C})$ 的值为 0。

[0104] 步骤 502: 查询分类器对所述当前查询数据中的关键词进行词性标注和短语识别。

[0105] 在实际中, 本实施例中的查询分类器需要集成词性标注工具、短语识别工具, 以及词根抽取工具。采用词性标注工具标注每一个关键词的词性, 例如: 当前查询数据为“find Nokia N95phones”时, 关键词“phones”的词性为 nns, 表示该词为复数名词; 并且该当前查询数据包括动词短语“find”和名词短语“Nokia N95phones”。

[0106] 步骤 503: 查询分类器从前述标注和识别的结果中获取所述当前查询数据中的名词短语, 并抽取出所述名词短语的中心词。

[0107] 查询分类器找出名词短语“Nokia N95phones”中的中心词“phones”。

[0108] 步骤 504: 查询分类器通过词根抽取工具获取所述中心词的词根。

[0109] 最后通过词根抽取工具, 抽取“phones”的词根为“phon”。

[0110] 从步骤 502 至 504 为本实施例中获取当前查询数据的句法特征的步骤。则对于“find Nokia N95phones”, 则当训练数据中能够匹配到关于当句法特征为“phon”时, 所述类目为 C 的数据, 则认为“find Nokia N95phones”的句法特征的值为 1, 否则为 0。当前查

询词对应的类目为 C, 且查询词的中心词的词根为 phon, 那么对于类目 C 的置信度计算时, 就认为 $f(x = \text{phon}, "find\ Nokia\ N95\ phones" \in \text{类目 C})$ 的值为 1, 其他情况下为 0。步骤 505: 查询分类器对所述当前查询数据的关键词进行类型标识。

[0111] 即是将当前查询数据的关键词都标注上类型, 例如, 将“find”标注为意愿动词, 将“Nokia”标注为品牌名, 等等。

[0112] 步骤 506: 查询分类器根据所述类型标识获取所述关键词的意动词标签、产品标签、品牌标签、型号标签和 / 或其他标签。

[0113] 则对于“find Nokia N95phones”, 得到的语义特征共有六个, 即是: 语义特征有“Action: find”, “Brand: Nokia”, “Model: N95”, 和“Product: phon”; 其中, Action 表明为意愿动词, 意愿动词的类型标识可以根据后台系统记录的意愿动词表来获取。

[0114] 步骤 505 和 506 是在实际应用中获取当前查询数据的语义特征的过程。

[0115] 步骤 507: 查询分类器判断在预置的点击日志中是否存在所述当前查询数据, 所述点击日志用于保存查询数据与点击特征的对应关系, 如果是, 则进入步骤 508, 如果否, 则进入步骤 509。

[0116] 在检索系统中, 根据用户输入的当前查询数据, 搜索引擎服务器会返回多个与当前查询数据的相关产品链接。其中, 点击日志中记录了用户在输入查询词后, 都点击了哪些产品链接, 以及该产品所属的类目。其中, 所述点击日志中还可以包括产品标题, 产品的关键字等信息。

[0117] 步骤 508: 在预置的点击日志中获取当前查询数据对应的前预定个数的点击类目; 所述点击类目为用户在检索该查询数据之后点击的相应产品所属的类目。

[0118] 例如, 当前查询数据为“find Nokia N95 phones”, 对应的点击特征有“Click_Category_Top1: telecommunications”, “Click_Category_Top2: consumer electronics”, “Click_Category_Top3: computer hardware&software”, 其中 Top i ($i = 1, 2, 3$) 标识了该点击特征是属于前第几个点击类目。所述 Click_Category 为点击类目。

[0119] 步骤 509: 查询分类器在点击日志中查询与所述当前查询数据最接近的点击查询数据, 所述点击查询数据与所述当前查询数据的编辑距离在点击日志中最小; 并将所述点击查询数据的对应的前预定个数的类目作为所述当前查询数据的点击特征。

[0120] 从步骤 507 至 509 为在实际应用中获取当前查询数据的点击特征的具体过程。

[0121] 步骤 5010: 查询分类器判断在预置的筛选日志中是否存在所述当前查询数据, 所述筛选日志用于保存查询数据与筛选特征的对应关系, 如果是, 则进入步骤 5011, 如果否, 则进入步骤 5012。

[0122] 步骤 5011: 查询分类器在预置的筛选日志中获取当前查询数据对应的前预定个数的筛选类目; 所述筛选类目表示所述当前查询数据在搜索引擎的映射表中所关联的类目。

[0123] 步骤 5012: 查询分类器在筛选日志中查询与所述当前查询数据最接近的筛选查询数据, 所述筛选查询数据与所述当前查询数据的编辑距离在筛选日志中最小; 并将所述筛选查询数据的对应的前预定个数的类目作为所述当前查询数据的筛选特征。

[0124] 从步骤 5010 到 5012 是在实际应用中获取当前查询数据的筛选特征的具体过程。

[0125] 步骤 5013: 在同一个会话中, 查询分类器获取所述当前查询数据的上一个查询数

据。

[0126] 在本实施例中,查询分类器则是查询“find Nokia N95phones”的上一个查询数据。

[0127] 步骤 5014:查询分类器获取到所述上一个查询数据对应的最优分类结果所属的类目,并以所述最优分类结果所属的类目为所述当前查询数据的会话特征。

[0128] 例如,用户以此输入三个查询 q_1 、 q_2 、和 q_3 。查询分类器判定 q_1 对应的最优的类目为 c_1 。当查询处理器在处理 q_2 的时候,对应的会话特征即是上一个查询数据对应的最优分类结果所属的类目为 c_1 。

[0129] 步骤 5015:根据所述词一级特征、句法特征、语义特征、点击特征、筛选特征和会话特征,查询分类器计算获取所述当前查询数据与每个类目对应的置信度,所述置信度用于表示所述当前查询数据属于某一个类目的可信程度。

[0130] 其中,假设本实施例中,当前查询数据的词一级的特征值分别为 1、1、1 和 0,句法特征为 1 和 0,语义特征为 1、1、1、0 和 0,点击特征为 0、0 和 1,筛选特征为 1、0 和 1,会话特征为 1,那么根据最大熵模型的公式:

$$[0131] \quad p(y|x) = \frac{1}{Z(x)} \exp \left(\sum_j \lambda_j f_j(x, y) \right)$$

[0132] 可以得到每一个类目对应的置信度的数值。步骤 5016:查询分类器向客户端返回所述每个类目及对应的置信度。

[0133] 查询分类器在计算出置信度之后,将对应的类目和置信度的对应关系按照置信度的大小进行排序,从而返回给客户端。

[0134] 从上述步骤可以看出,所述查询分类器可以独立于搜索引擎服务器存在,当在实际应用中的检索系统需要进行查询分类时,再集成查询分类器到搜索引擎服务器上。参考图 6 所示,为查询分类器输入和输入内容的结构示意图。

[0135] 步骤 5017:搜索引擎服务器在客户端根据所述置信度进行选择的类目中对所述当前查询数据进行检索。

[0136] 对于前述的各方法实施例,为了简单描述,故将其都表述为一系列的动作组合,但是本领域技术人员应该知悉,本申请并不受所描述的动作顺序的限制,因为依据本申请,某些步骤可以采用其他顺序或者同时进行。其次,本领域技术人员也应该知悉,说明书中所描述的实施例均属于优选实施例,所涉及的动作和模块并不一定是本申请所必须的。

[0137] 参考图 7,示出了本申请一种检索方法实施例 4 的流程图,可以包括以下步骤:

[0138] 步骤 701:接收客户端提交的当前查询数据。

[0139] 步骤 702:提取所述当前查询数据的总属性特征,所述属性特征用于计算所述当前查询数据的置信度。

[0140] 步骤 703:根据所述总属性特征,计算获取所述当前查询数据与每个类目对应的置信度,所述置信度用于表示所述当前查询数据属于某一个类目的可信程度。

[0141] 在本实施例中,所述步骤 701 ~ 703 可以和前述的 3 个方法实施例中的执行过程保持一致,并且关于步骤 701 ~ 703 介绍可以参考前述 3 个方法实施例。

[0142] 步骤 704:在计算得到的最大的置信度对应的类目中,对所述当前查询数据进行检索。

[0143] 在本实施例中,与方法实施例 1 不同的是,在计算获取到当前查询数据属于每个类目的各个置信度之后,搜索引擎服务器可以做一个判断,即是在步骤 703 中计算得到的所有置信度中,获取到最大的那个置信度对应的类目,然后在该类目中直接对所述当前查询数据进行检索。更进一步,再将检索结果直接返回给用户。

[0144] 需要说明的是,此处,搜索引擎服务器不仅可以按照最大的置信度进行检索,还可以根据实际情况,或者后台的记录而选择其中一个置信度对应的类目进行选择,或者,满足一定条件的多个置信度对应的多个类目都可以作为检索当前查询数据的范围,本领域技术人员对于本方案作出的一些变形也属于本申请保护的范畴。

[0145] 基于本实施例所述的方法,无需向用户返回置信度及其对应的类目,而是直接根据最优的结果进行检索,这样一般情况下,就能满足用户对于检索结果的要求,同时无需后台保存相应的映射表等来记载类目信息,无论用户的当前查询数据有没有出现过,都能够找出其属于某个类目的最大置信度,并将在该类目中的检索作为最合理的检索操作,在提高了对于查询数据的覆盖率的同时,无需搜索引擎先向客户端传输匹配获取的类目,减少了搜索引擎向客户端传输的数据量,减轻了搜索引擎服务器的负担,提高了网络间数据的传输速率。

[0146] 与上述本申请一种检索方法实施例 1 所提供的方法相对应,参见图 8,本申请还提供了一种检索装置实施例 1,在本实施例中,该装置可以包括:

[0147] 接收模块 801,用于接收客户端提交的当前查询数据。

[0148] 提取模块 802,用于提取所述当前查询数据的总属性特征,所述总属性特征用于计算所述当前查询数据与各个类目之间的置信度。

[0149] 在本申请实施例中,所述当前查询数据的一些特征被提取出来用于后续条件概率的运算,所述当前查询数据共需要提取总属性特征,所述总属性特征在实际应用中优选的可以包括六种特征,所述词一级特征为所述当前查询数据中的关键词,即是非停用词对应的词根被抽取为词一级特征;具体实现时可以通过分词器来获取到当前查询数据的关键词;所述句法特征为所述当前查询数据中的中心词所对应的词根,即是当前查询数据中的句法级别的中心词所对应的词根被抽取为句法特征,中心词可以在一定程度上反映了当前查询数据的意图核心词;所述语义特征为所述关键词对应的语义标签,所述当前查询数据中的关键词所对应的词根所对应的语义标签被抽取为语义的特征,在本申请实施例中所采用的语义标签有:意动词标签(Action),用于标识该词为一意愿动词;产品标签(Product),用于标识该词为某一产品词;品牌标签(Brand),用于标识该词为某一品牌词;型号标签(Model),用于标识该词为某一型号词;其他标签(Other),用于其他词所对应的标签。

[0150] 所述点击特征为所述当前查询数据在预置的点击日志中对应的至少一个点击类目,所述预置的点击日志是在搜索引擎服务器上保存的一个数据库,用于保存用户输入查询数据之后,所点击的产品所属的类目。所述筛选特征为所述当前查询数据在预置的筛选日志中对应的至少一个筛选类目;所述筛选日志就用于保存用户输入的查询词和被点击的类目的对应关系。所述会话特征用于表示在同一个会话(Session)中,所述当前查询数据的前一个查询对应的最优类目;在一个有效的时间内,在同一个浏览器内的一系列的查询行为属于一个会话。即是用户打开一个网页之后,在未关这个网页之前,就认为所输入的查

询词都属于同一个浏览器。

[0151] 计算获取模块 803,用于根据所述总属性特征,计算获取所述当前查询数据与每个类目对应的置信度,所述置信度用于表示所述当前查询数据属于某一个类目的可信程度。

[0152] 其中,根据得到的当前查询数据的特征,可以认为是条件概率模型中的给定条件,而所述置信度为一个 0 ~ 1 之间的数值,用于表示在给定条件下某一件事发生的概率。因为条件概率的含义是求在一件事(条件)发生的前提下,另外一件事发生的概率,因此,本实施例中,当前查询数据的词一级特征、句法特征、语义特征、点击特征、筛选特征和会话特征则被认为是条件,而对于某个类目计算出的置信度,则可以理解为当前查询数据属于该某个类目的概率是多少。因此可以看出,置信度越大,即是当前查询数据越有可能属于对应的某个类目。

[0153] 优选情况下,可以采用最大熵模型来对具体的类目计算对应的置信度。还需要说明的是,本申请实施例在对每个类目计算置信度时,主要采用的是机器学习方法,因为该方法可以通过一定规模的训练查询数据,来模拟用户的查询行为,计算获取到查询数据归类于某一类目的概率。可以根据条件概率模型计算存在这些特征值的情况下,当前查询数据归属于某一类目的概率值。

[0154] 返回模块 804,用于向客户端返回所述每个类目及对应的置信度。

[0155] 假设实际中的分类有 40 种,则需要对每一个类目都进行计算,在实际中,就计算出对应的 40 个置信度,搜索引擎服务器分别将每个类目及对应的置信度返回给用户,以便于用户根据自己所感兴趣的类目定位检索范围,以更加精准的检索到需要的产品信息。

[0156] 第一检索模块 805,用于在客户端根据所述置信度进行选择的类目中对所述当前查询数据进行检索。

[0157] 本实施例所述的装置可以集成到搜索引擎的服务器上,也可以单独作为一个实体与搜索引擎服务器相连,另外,需要说明的是,当本申请所述的方法采用软件实现时,可以作为搜索引擎的服务器新增的一个功能,也可以单独编写相应的程序,本申请不限定所述方法或装置的实现方式。

[0158] 本实施例中,所述装置在能够将置信度和类目的对应关系返回至客户端的基础上,根据客户端所选的类目再对当前查询数据进行检索,这样就在一定程度上减少了搜索引擎服务器的工作,提升了搜索引擎服务器的工作效率和性能,也能增加用户在数据查询方面的搜索体验。

[0159] 与上述本申请一种检索方法实施例 2 所提供的方法相对应,参见图 9,本申请还提供了一种检索装置的优选实施例 2,在本实施例中,该装置具体可以包括:

[0160] 接收模块 801,用于接收客户端提交的当前查询数据。

[0161] 提取模块 802,用于提取所述当前查询数据的词一级特征、句法特征、语义特征、点击特征、筛选特征和会话特征。

[0162] 所述词一级特征为所述当前查询数据中的关键词,所述句法特征为所述当前查询数据中的中心词所对应的词根,所述语义特征为所述关键词对应的语义标签,所述点击特征为所述当前查询数据在预置的点击日志中对应的至少一个点击类目,所述筛选特征为所述当前查询数据在预置的筛选日志中对应的至少一个筛选类目,所述会话特征用于表示在同一个会话中,所述当前查询数据对的前一个查询对应的最优类目。

[0163] 设置给定信息子模块 901,用于将所述词一级特征、句法特征、语义特征、点击特征、筛选特征和会话特征,作为条件概率模型的给定信息。

[0164] 计算置信度子模块 902,用于采用条件概率模型计算在所述给定信息的情况下,所述当前查询在每个类目下的后验概率;并将所述后验概率作为当前查询数据的在该类目下的置信度。

[0165] 条件概率模型可以训练当前查询数据所得到的特征的权重,通过训练得到的对应的特征值和权重,来利用条件概率模型计算当前查询数据被确定为某一类目的条件概率,即是置信度。

[0166] 条件概率模型优选情况下可以选择最大熵模型,其公式如下所示:

$$p(y|x) = \frac{1}{Z(x)} \exp\left(\sum_j \lambda_j f_j(x, y)\right)$$

[0168] 其中 $y \in \{c \text{ is correct}, c \text{ is incorrect}\}$; x 是当前查询数据所对应的给定信息,在本实施例中包括:词一级特征、句法特征、语义特征、点击特征、筛选特征和会话特征; f_j 是最大熵模型对应的特征值,例如,即是当 x 为词一级特征时,如果在训练数据中能够匹配到“当前查询数据的关键词包括 phone 时,所属类目为 C”,则特征 f_j 在条件“当前查询数据的关键词包括 phone 时”的值就为 1,否则为 0; λ_j 是当前查询数据的各个特征对应的权重,可以通过模型训练得到的; $Z(x)$ 是归一化因子,也通过模型训练得到。

[0169] 在实际中,还可以采用线性回归或者支持向量机等方式来建立条件概率模型。

[0170] 第一判断模块 903,用于判断所述计算获取的置信度是否大于预置的第一阈值。

[0171] 在计算获取了当前查询数据归属于所有类目的置信度之后,查询分类可以预置第一阈值,例如,可以将第一阈值设置为 0.5,当置信度大于 0.5,则认为对应的该类目的可信程度较高。

[0172] 第一获取模块 904,用于当所述第一判断模块的结果为是时,获取所述大于所述阈值的置信度以及所述置信度对应的类目。

[0173] 查询分类器可以将所有大于第一阈值的置信度及其对应类目都返回给客户端,根据置信度还可以从高至低排序展示;或者查询分类器还可以会把前预定个数的(例如 8 个)置信度及其对应类目在客户端上进行展示。

[0174] 第二获取模块 905,用于当计算获取的所有类目的置信度均小于所述第一阈值,则从所有置信度中直接获取预定个数的相对较大的置信度以及对应的类目。

[0175] 等到所有的置信度都已经判断完毕,如果所有类目对应的置信度都不大于第一阈值,则查询分类器直接可以对所有置信度按照从大到小进行排序,并获取前预定个数的相对较大的置信度及其对应的类目。

[0176] 则所述返回模块 804,具体用于:向用户返回所述大于所述阈值的置信度以及所述置信度对应的类目;或者向客户端返回所述每个类目及对应的置信度。

[0177] 第一检索模块 805,用于在客户端根据所述置信度进行选择类目中对所述当前查询数据进行检索。

[0178] 与上述本申请一种检索方法实施例 3 所提供的方法相对应,参见图 10,本申请还提供了一种检索装置的优选实施例 3,在本实施例中,该装置具体可以包括:

[0179] 接收模块 801,用于接收客户端提交的当前查询数据。

[0180] 标注识别子模块 1001,用于对所述当前查询数据中的关键词进行词性标注和短语识别。

[0181] 其中,进行词性标注还可以由词性标注工具来实现,短语识别还可以由短语识别工具来实现。

[0182] 抽取中心词子模块 1002,用于从前述标注和识别的结果中获取所述当前查询数据中的名词短语,并抽取出所述名词短语的中心词。

[0183] 词根抽取工具 1003,用于获取所述中心词的词根。

[0184] 类型标识子模块 1004,用于对所述当前查询数据的关键词进行类型标识。

[0185] 获取标签子模块 1005,用于根据所述类型标识获取所述关键词的意动词标签、产品标签、品牌标签、型号标签和 / 或其他标签。

[0186] 第一判断子模块 1006,用于判断在预置的点击日志中是否存在所述

当前查询数据,所述点击日志用于保存查询数据与点击特征的对应关系。

[0188] 第一获取子模块 1007,用于当所述第一判断子模块的结果为是时,在预置的点击日志中获取当前查询数据对应的前预定个数的点击类目;所述点击类目为用户在检索该查询数据之后点击的相应产品所属的类目。

[0189] 第一查询获取子模块 1008,用于当所述第一判断子模块的结果为否时,在点击日志中查询与所述当前查询数据最接近的点击查询数据,所述点击查询数据与所述当前查询数据的编辑距离在点击日志中最小;并将所述点击查询数据的对应的前预定个数的类目作为所述当前查询数据的点击特征。

[0190] 第二判断子模块 1009,用于判断在预置的筛选日志中是否存在所述当前查询数据,所述筛选日志用于保存查询数据与筛选特征的对应关系。

[0191] 第二获取子模块 1010,用于当所述第二判断子模块的结果为是时,在预置的筛选日志中获取当前查询数据对应的前预定个数的筛选类目;所述筛选类目表示所述当前查询数据在搜索引擎的映射表中所关联的类目。

[0192] 第二查询获取子模块 1011,用于当所述第二判断子模块的结果为否时,在筛选日志中查询与所述当前查询数据最接近的筛选查询数据,所述筛选查询数据与所述当前查询数据的编辑距离在筛选日志中最小;并将所述筛选查询数据的对应的前预定个数的类目作为所述当前查询数据的筛选特征。

[0193] 第三获取子模块 1012,用于在同一个会话中,获取所述当前查询数据的上一个查询数据;

[0194] 第四获取子模块 1013,用于获取到所述上一个查询数据对应的最优分类结果所属的类目,并以所述最优分类结果所属的类目为所述当前查询数据的会话特征。

[0195] 计算获取模块 803,用于根据所述词一级特征、句法特征、语义特征、点击特征、筛选特征和会话特征,计算获取所述当前查询数据与每个类目对应的置信度,所述置信度用于表示所述当前查询数据属于某一个类目的可信程度;

[0196] 返回模块 804,用于向客户端返回所述每个类目及对应的置信度;

[0197] 第一检索模块 805,用于在客户端根据所述置信度进行选择类目中对所述当前查询数据进行检索。

[0198] 在本实施例中,可以利用现有的工具,例如分词器、词性标注工具等,这样就降低

了实施成本,并且对于基于条件概率的机器学习方法来讲,其覆盖率相对较高,不管当前查询数据之前有没有出现过,都可以由条件概率模型根据提取出的各种特征得到所有可能相关的类目的置信度。这样就比现有技术中的表格可以应用的更为广泛,并且准确。

[0199] 与上述本申请一种检索方法实施例 4 所提供的方法相对应,参见图 11,本申请还提供了一种检索装置的优选实施例 4,在本实施例中,该装置具体可以包括:

[0200] 接收模块 801,用于接收客户端提交的当前查询数据;

[0201] 提取模块 802,用于提取所述当前查询数据的总属性特征,所述属性特征用于计算所述当前查询数据的置信度;

[0202] 计算获取模块 803,用于根据所述总属性特征,计算获取所述当前查询数据与每个类目对应的置信度,所述置信度用于表示所述当前查询数据属于某一个类目的可信程度;

[0203] 第二检索模块 1101,用于在计算得到的最大的置信度对应的类目中,对所述当前查询数据进行检索。

[0204] 在本实施例中,与装置实施例 1 不同的是,在计算获取模块 803 计算获取到当前查询数据属于每个类目的各个置信度之后,搜索引擎服务器可以做一个判断,即是在计算得到的所有置信度中,获取到最大的那个置信度对应的类目,然后在该类目中直接对所述当前查询数据进行检索。更进一步,再将检索结果直接返回给用户。

[0205] 基于本实施例所述的装置,无需向用户返回置信度及其对应的类目,而是直接根据最优选的结果进行检索,这样一般情况下,就能满足用户对于检索结果的要求,同时无需后台保存相应的映射表等来记载类目信息,无论用户的当前查询数据有没有出现过,都能够找出其属于某个类目的最大置信度,并将在该类目中的检索作为最合理的检索操作,在提高了对于查询数据的覆盖率的同时,优化了搜索引擎服务器的性能,满足了用户不同的检索需求。

[0206] 与上述本申请一种检索方法和装置实施例相对应,本申请还提供了一种检索系统的实施例,在本实施例中,该系统具体可以包括:

[0207] 位于搜索引擎服务器端的:

[0208] 查询分类器,所述查询分类器包括:接收模块、提取模块、计算获取模块和返回模块。

[0209] 所述接收模块,用于接收客户端提交的当前查询数据。

[0210] 所述提取模块,用于提取所述当前查询数据的词一级特征、句法特征、语义特征、点击特征、筛选特征和会话特征。

[0211] 所述计算获取模块,用于根据所述词一级特征、句法特征、语义特征、点击特征、筛选特征和会话特征,计算获取所述当前查询数据与每个类目对应的置信度,所述置信度用于表示所述当前查询数据属于某一个类目的可信程度。

[0212] 所述返回模块,用于向客户端返回所述每个类目及对应的置信度。

[0213] 第一检索模块,用于在客户端根据所述置信度进行选择类目中对所述当前查询数据进行检索。

[0214] 需要说明的是,该系统还可以包括客户端:用于通过浏览器接收用户输入的当前查询数据并提交至搜索引擎服务器。

[0215] 需要说明的是,本说明书中的各个实施例均采用递进的方式描述,每个实施例重

点说明的都是与其他实施例的不同之处,各个实施例之间相同相似的部分互相参见即可。对于装置类实施例而言,由于其与方法实施例基本相似,所以描述的比较简单,相关之处参见方法实施例的部分说明即可。

[0216] 最后,还需要说明的是,在本文中,诸如第一和第二等之类的关系术语仅仅用来将一个实体或者操作与另一个实体或操作区分开来,而不一定要求或者暗示这些实体或操作之间存在任何这种实际的关系或者顺序。而且,术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含,从而使得包括一系列要素的过程、方法、物品或者设备不仅包括那些要素,而且还包括没有明确列出的其他要素,或者是还包括为这种过程、方法、物品或者设备所固有的要素。在没有更多限制的情况下,由语句“包括一个……”限定的要素,并不排除在包括所述要素的过程、方法、物品或者设备中还存在另外的相同要素。

[0217] 以上对本申请所提供的一种检索方法、装置及系统进行了详细介绍,本文中应用了具体个例对本申请的原理及实施方式进行了阐述,以上实施例的说明只是用于帮助理解本申请的方法及其核心思想;同时,对于本领域的一般技术人员,依据本申请的思想,在具体实施方式及应用范围上均会有改变之处,综上所述,本说明书内容不应理解为对本申请的限制。

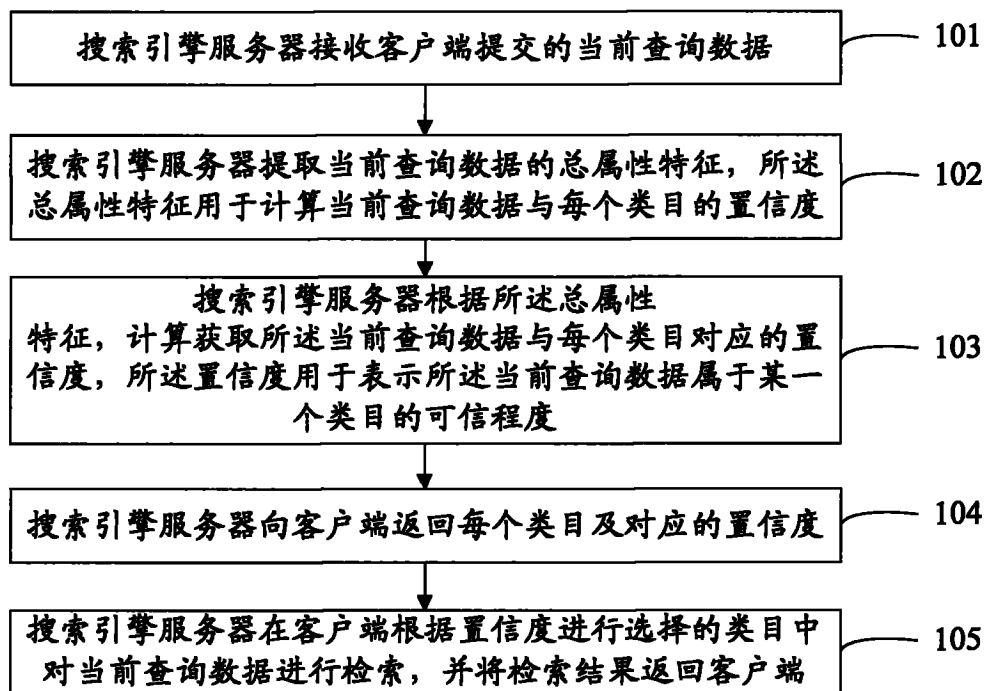


图 1

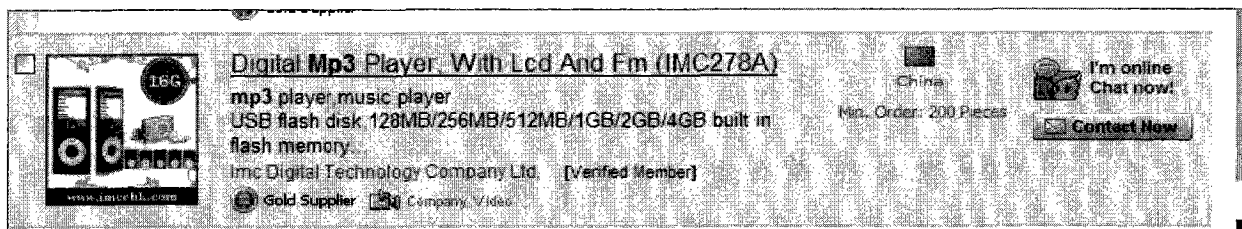


图 2



图 3

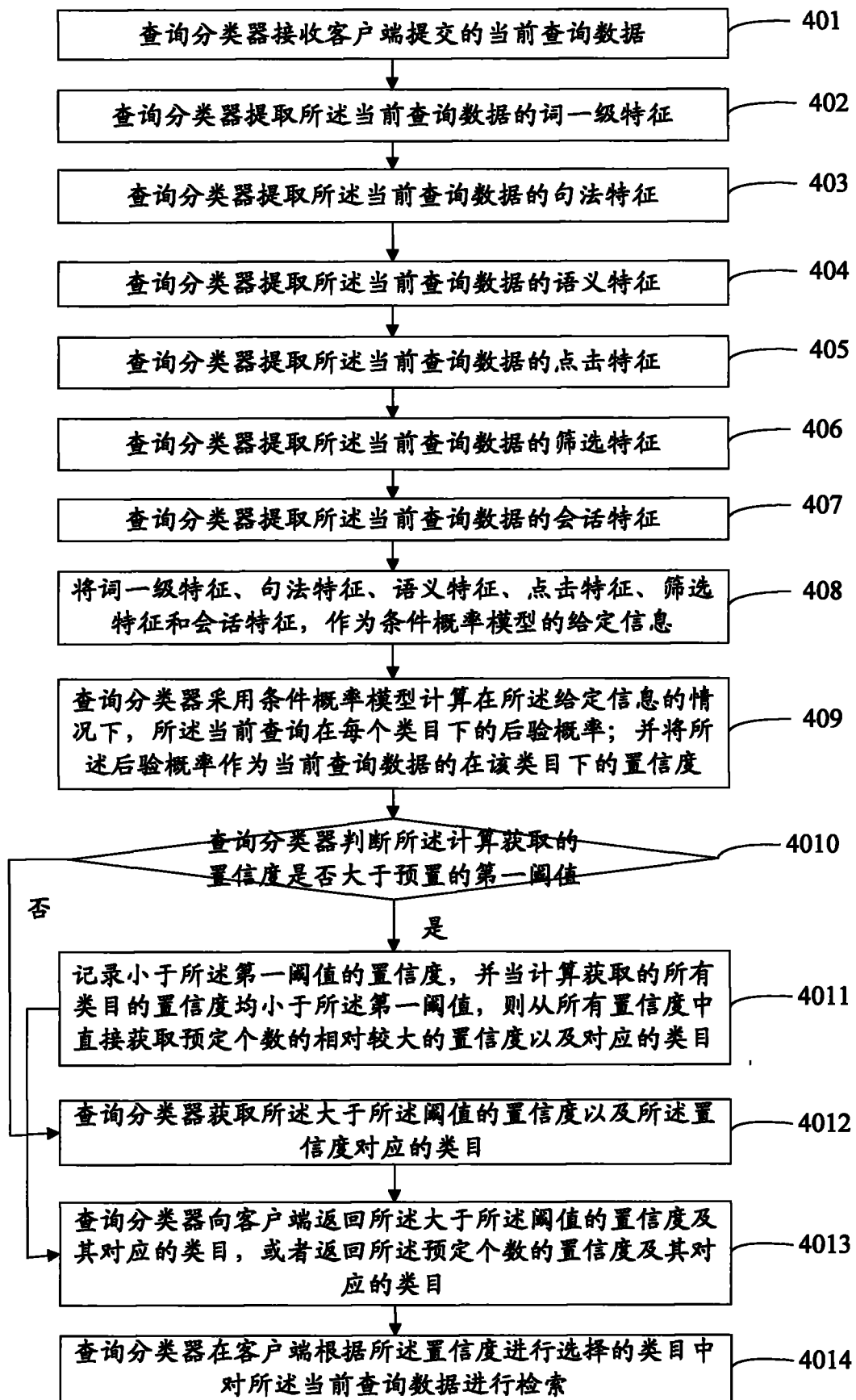


图 4

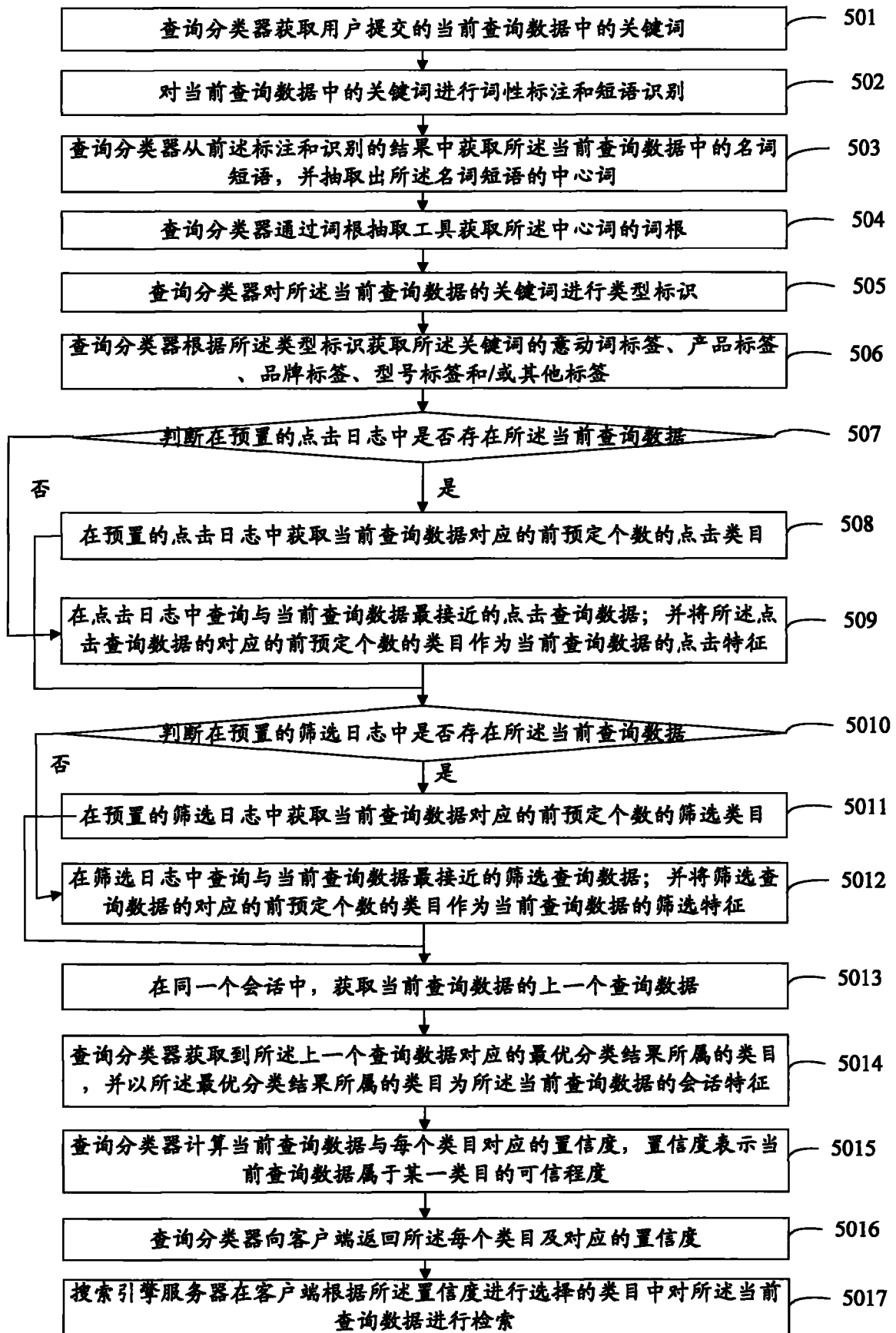


图 5

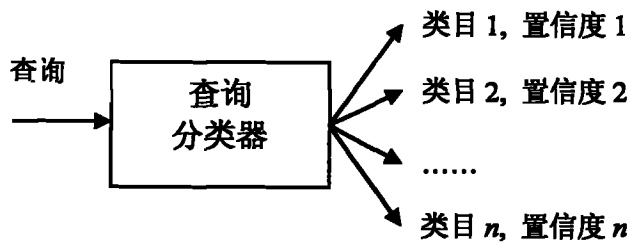


图 6

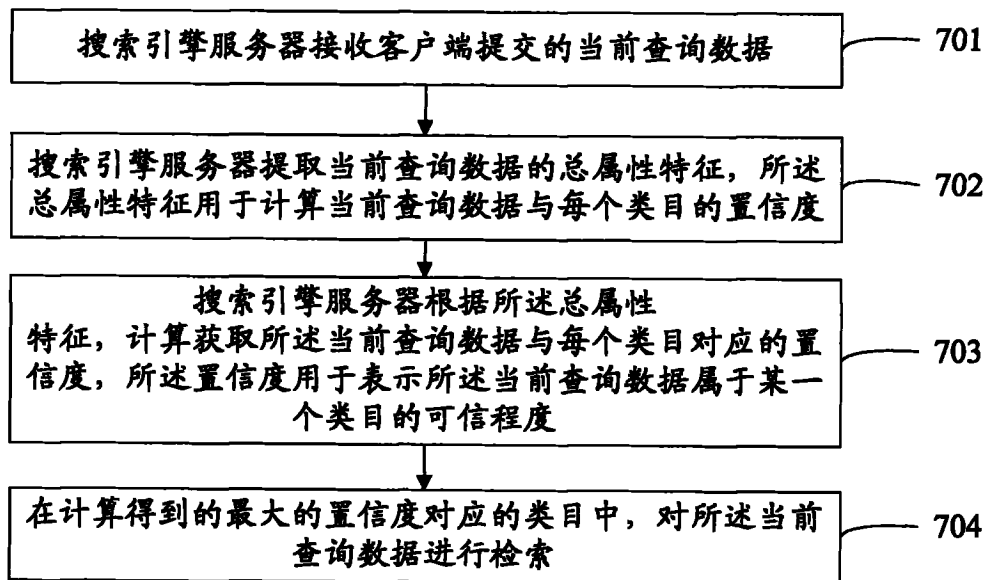


图 7

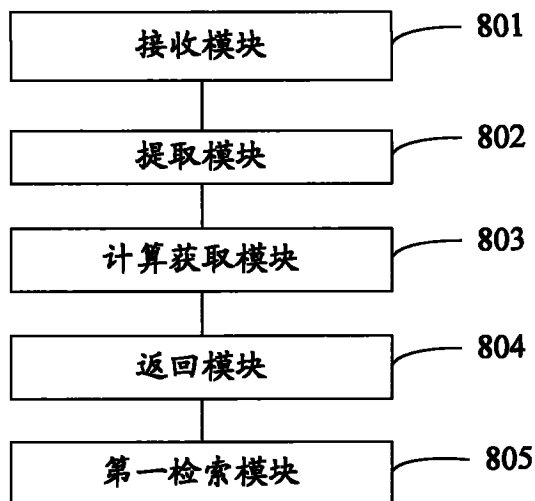


图 8

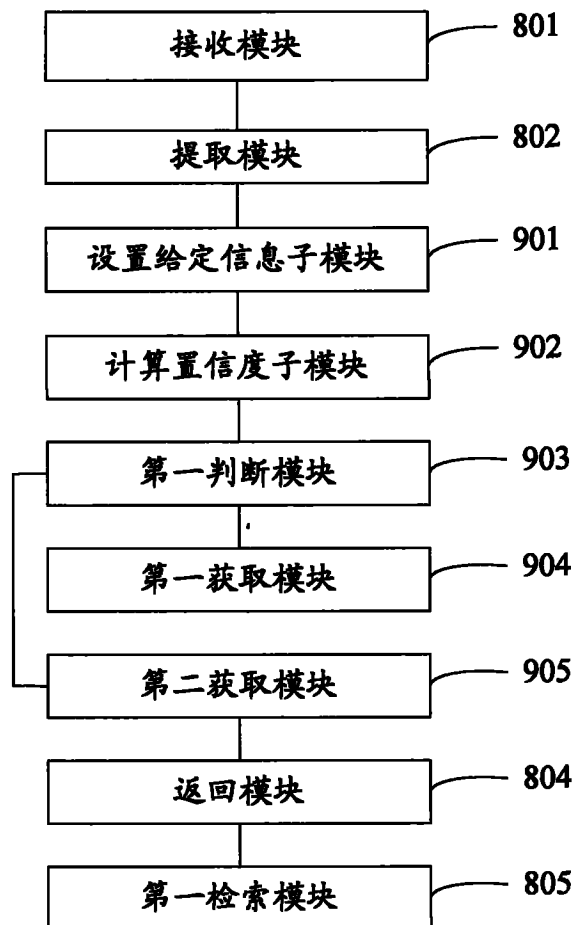


图 9

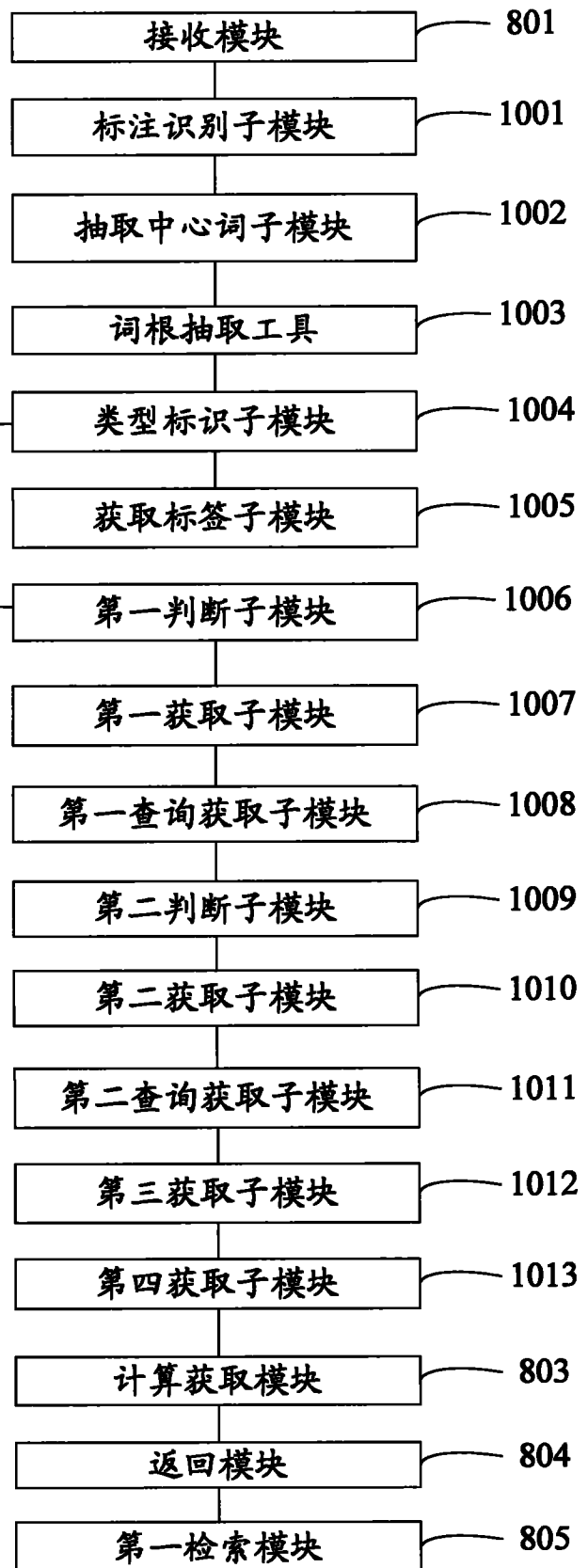


图 10

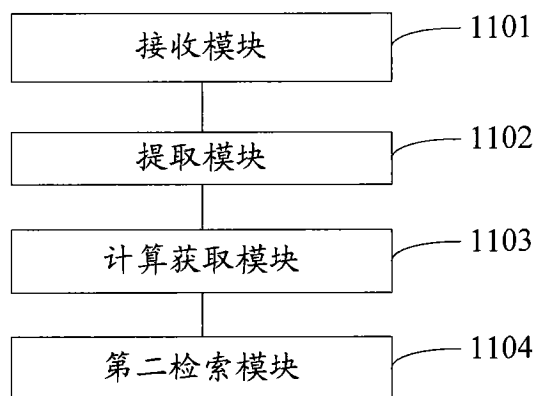


图 11