(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 109582847 B
(45)授权公告日 2021.08.24

(21)申请号 201811410496.4
(22)申请日 2018.11.23
(65)同一申请的已公布的文献号
申请公布号 CN 109582847 A
(43)申请公布日 2019.04.05
(73)专利权人 咪咕视讯科技有限公司
地址 201206 上海市浦东新区自由贸易试验区云桥路636号1幢
专利权人 咪咕文化科技有限公司
中国移动通信集团有限公司
(72)发明人 桑永嘉
(74)专利代理机构 北京派特恩知识产权代理有限公司 11270
代理人 张振伟 张颖玲
(51)Int.Cl.
G06F 16/951(2019.01)

(56)对比文件
CN 104915458 A,2015.09.16
CN 104915458 A,2015.09.16
CN 108829267 A,2018.11.16
CN 107273537 A,2017.10.20
US 2017/0308522 A1,2017.10.26
CN 107329583 A,2017.11.07
CN 108062373 A,2018.05.22
CN 103258023 A,2013.08.21
CN 108319376 A,2018.07.24
CN 108241740 A,2018.07.03
CN 108170293 A,2018.06.15
CN 106919682 A,2017.07.04
CN 104965826 A,2015.10.07
CN 108227954 A,2018.06.29

审查员 刘瑞

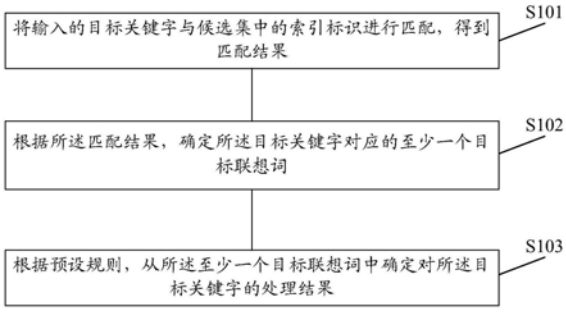
权利要求书2页 说明书19页 附图11页

(54)发明名称

一种信息处理方法及装置、存储介质

(57)摘要

本申请实施例公开了一种信息处理方法及装置、存储介质，其中，所述方法包括：将输入的目标关键字与候选集中的索引标识进行匹配，得到匹配结果，其中，所述候选集用于表征目标联想词和索引标识的对应关系；根据所述匹配结果，确定所述目标关键字对应的至少一个目标联想词；根据预设规则，从所述至少一个目标联想词中确定对所述目标关键字的处理结果。



CN 109582847 B

1.一种信息处理方法,其特征在于,所述方法包括:

将输入的目标关键字与候选集中的索引标识进行匹配,得到匹配结果,其中,所述候选集用于表征目标联想词和索引标识的对应关系;

根据所述匹配结果,确定所述目标关键字对应的至少一个目标联想词;

根据预设规则,从所述至少一个目标联想词中确定对所述目标关键字的处理结果;

所述根据所述匹配结果,确定所述目标关键字对应的至少一个目标联想词,包括:

如果匹配成功,从所述候选集中确定所述目标关键字对应的至少一个目标联想词;

如果匹配不成功,将所述目标关键字与目标联想词列表中的索引标识进行匹配,得到第二匹配结果;

根据所述第二匹配结果,确定所述目标关键字对应的至少一个目标联想词。

2.根据权利要求1所述的方法,其特征在于,所述方法还包括:

根据获取的至少两类源数据,确定所述目标联想词列表;

根据所述目标联想词列表中联想词的属性信息,确定所述联想词的索引标识,所述索引标识用于对所述联想词进行标记;

根据所述索引标识从所述目标联想词列表中,确定相应的目标联想词;

根据确定的目标联想词,形成相应的候选集。

3.根据权利要求2所述的方法,其特征在于,所述根据所述确定的目标联想词,形成相应的候选集,包括:

将所述目标联想词列表中与所述索引标识对应的联想词,确定为目标联想词;

获取所述目标联想词对应的权重值;

根据所述目标联想词和所述目标联想词对应的权重值,形成候选集。

4.根据权利要求2所述的方法,其特征在于,所述根据获取的至少两类源数据,确定目标联想词列表,包括:

获取至少两类源数据,每一类所述源数据包括搜索词;

将被搜索的频率和/或次数大于预设阈值的搜索词,确定为联想词;

根据所述联想词被搜索的频率和/或次数,确定所述联想词对应的权重值;

根据确定的联想词和所述联想词对应的权重值,形成每一类源数据对应的联想词列表;

将所述每一类源数据对应的联想词列表进行合并,形成目标联想词列表。

5.根据权利要求4所述的方法,其特征在于,所述将所述每一类源数据对应的联想词列表进行合并,形成目标联想词列表,包括:

将所有类别的源数据所对应的联想词列表中的联想词和权重值分别进行合并,生成合并联想词和所述合并联想词对应的权重值;

根据所述合并联想词和所述合并联想词对应的权重值,形成目标联想词列表。

6.根据权利要求4所述的方法,其特征在于,所述源数据包括文本数据、视频数据和自定义数据;相应地,所述将所述每一类源数据对应的联想词列表进行合并,形成目标联想词列表,包括:

将所述文本数据、所述视频数据和所述自定义数据所对应的联想词列表中的联想词和权重值分别进行合并,生成合并联想词和所述合并联想词对应的权重值;

根据所述合并联想词和所述合并联想词对应的权重值，形成目标联想词列表。

7.根据权利要求4所述的方法，其特征在于，如果所述源数据为文本数据，所述根据所述联想词被搜索的频率和/或次数，确定所述联想词对应的权重值，包括：

获取至少一个单位时间内的联想词和所述联想词的被搜索频率和/或次数；

将每一所述单位时间内的被搜索频率和/或次数进行合并，得到第一合并结果；

将所述第一合并结果进行标准化处理，得到每一所述联想词的权重值。

8.根据权利要求4所述的方法，其特征在于，如果所述源数据为视频数据，所述根据所述联想词被搜索的频率和/或次数，确定所述联想词对应的权重值，包括：

获取至少一个单位时间内的所述视频数据被搜索的频率和/或次数，建立所述联想词与所述视频被搜索的频率和/或次数的关联关系；

将每一所述单位时间内的被搜索的频率和/或次数进行合并，得到第二合并结果；

将所述第二合并结果进行标准化处理，得到每一所述联想词的权重值。

9.根据权利要求1所述的方法，其特征在于，所述根据预设规则，从所述至少一个目标联想词中确定对所述目标关键字的处理结果，包括：

根据所述至少一个目标联想词，形成目标联想词集合；

根据目标用户输入的历史内容，构建目标用户信息集合；

获取所述目标联想词集合和目标用户信息集合的相似度；

根据所述相似度对所述目标联想词集合中的目标联想词进行排序，得到排序结果；

根据所述排序结果，从所述目标联想词集合中确定对所述目标关键字的处理结果。

10.一种信息处理装置，其特征在于，所述装置包括：

匹配单元，配置为将输入的目标关键字与候选集中的索引标识进行匹配，得到匹配结果，其中，所述候选集用于表征目标联想词和索引标识的对应关系；

第一确定单元，配置为根据所述匹配结果，确定所述目标关键字对应的至少一个目标联想词；如果匹配成功，从所述候选集中确定所述目标关键字对应的至少一个目标联想词；如果匹配不成功，将所述目标关键字与目标联想词列表中的索引标识进行匹配，得到第二匹配结果；根据所述第二匹配结果，确定所述目标关键字对应的至少一个目标联想词；

第二确定单元，配置为根据预设规则，从所述至少一个目标联想词中确定对所述目标关键字的处理结果。

11.一种计算机可读存储介质，其特征在于，所述计算机可读存储介质中存储有计算机可执行指令，该计算机可执行指令配置为执行上述权利要求1至9任一项提供的信息处理方法。

一种信息处理方法及装置、存储介质

技术领域

[0001] 本申请实施例涉及计算机信息技术，涉及但不限于一种信息处理方法及装置、存储介质。

背景技术

[0002] 搜索引擎已经成为众多信息服务类产品的重要入口，当用户通过搜索引擎输入某个查询信息后，搜索框下方会自动联想推荐的备选词，并提示给用户，帮助用户快速进入需要搜索的地址，并找到需要搜索的内容。

[0003] 在视频搜索领域，相关技术中沿用了传统搜索引擎的搜索词联想方法，通过字典树模型，采用以用户疑问与字典树模型匹配，字典树的数据源主要来自于历史用户疑问，存在覆盖率不足、效果不佳的问题。

发明内容

[0004] 有鉴于此，本申请实施例为解决现有技术中存在的数据源的覆盖率不足，推荐的联想词不够精确，导致用户体验效果不佳的问题，而提供一种信息处理方法及装置、存储介质。

[0005] 本申请实施例的技术方案是这样实现的：

[0006] 第一方面，本申请实施例提供了一种信息处理方法，所述方法包括：

[0007] 将输入的目标关键字与候选集中的索引标识进行匹配，得到匹配结果，其中，所述候选集用于表征目标联想词和索引标识的对应关系；

[0008] 根据所述匹配结果，确定所述目标关键字对应的至少一个目标联想词；

[0009] 根据预设规则，从所述至少一个目标联想词中确定对所述目标关键字的处理结果。

[0010] 第二方面，本申请实施例提供了一种信息处理装置，所述装置包括：

[0011] 匹配单元，配置为将输入的目标关键字与候选集中的索引标识进行匹配，得到匹配结果，其中，所述候选集用于表征目标联想词和索引标识的对应关系；

[0012] 第一确定单元，配置为根据所述匹配结果，确定所述目标关键字对应的至少一个目标联想词；

[0013] 第二确定单元，配置为根据预设规则，从所述至少一个目标联想词中确定对所述目标关键字的处理结果。

[0014] 第三方面，本申请实施例提供了一种计算机可读存储介质，所述计算机可读存储介质中存储有计算机可执行指令，该计算机可执行指令配置为执行上述实施例提供的信息处理方法。

[0015] 本申请实施例提供了一种信息处理方法及装置、存储介质，所述方法包括：将输入的目标关键字与候选集中的索引标识进行匹配，得到匹配结果，其中，所述候选集用于表征目标联想词和索引标识的对应关系；根据所述匹配结果，确定所述目标关键字对应的至少

一个目标联想词;根据预设规则,从所述至少一个目标联想词中确定对所述目标关键字的处理结果。如此,能够提高数据源的覆盖率,融合用户行为数据和外部数据,根据用户的行为数据确定用户的喜好,并结合用户的喜好和外部数据确定对应的联想词,以优先推荐用户兴趣度高的联想词。

附图说明

[0016] 图1为本申请实施例提供的信息处理方法流程示意图一;

[0017] 图2为本申请实施例提供的信息处理方法流程示意图二;

[0018] 图3为本申请实施例提供的信息处理方法流程示意图三;

[0019] 图4为本申请实施例提供的信息处理方法流程示意图四;

[0020] 图5为本申请实施例提供的信息处理方法流程示意图五;

[0021] 图6为本申请实施例提供的信息处理方法流程示意图六;

[0022] 图7为本申请实施例提供的信息处理方法流程示意图七;

[0023] 图8为本申请实施例提供的一种目标联想词列表的生成流程示意图;

[0024] 图8A为本申请实施例提供的一种通过热搜词日志获得热搜词权重表的流程示意图;

[0025] 图8B为本申请实施例提供的一种通过视频库片名和视频点击量获得片名权重表的流程示意图;

[0026] 图8C为本申请实施例提供的一种联想词在线服务处理流程的示意图;

[0027] 图8D为本申请实施例提供的一种联想词权重表的生成流程示意图;

[0028] 图9为本申请实施例提供的信息处理装置的组成结构示意图;

[0029] 图10为本申请实施例提供的一种计算机设备结构示意图。

具体实施方式

[0030] 相关技术中,通过字典树模型,采用以用户输入的内容与字典树模型匹配,来得到对应的查询结果,由于字典树的数据源主要来自于历史用户输入的内容,且未能与视频业务特点、用户行为结合,存在如下缺点:

[0031] (1)数据源单一或不全面:现有搜索联想词数据源使用的是自身站点用户历史输入的内容,部分能结合到搜索内容库,并未考虑到单用户的喜好行为、外部数据。

[0032] (2)数据融合排序方法单一:现有的搜索联想词一般采用以用户输入的搜索词的数量等单一指标作为排序因素,未能将多因素进行融合。

[0033] 针对相关技术的不足,本申请实施例提出一种信息处理的方法。本技术方案分为两部分:目标联想词列表生成,目的是根据多方数据源生成一份带权重的目标联想词列表;联想词在线服务处理流程则是以目标联想词列表作为输入,提供对于用户输入的内容的在线响应。

[0034] 为使本申请实施例的目的、技术方案和优点更加清楚,下面将结合本申请实施例中的附图,对发明的具体技术方案做进一步详细描述。以下实施例用于说明本申请,但不用来限制本申请的范围。

[0035] 本申请实施例提供一种信息处理方法,该方法应用于计算机设备,该方法所实现

的功能可以通过计算机设备中的处理器调用程序代码来实现，当然程序代码可以保存在计算机存储介质中，可见，该计算机设备至少包括处理器和存储介质。

[0036] 图1为本申请实施例提供的信息处理方法流程示意图一，如图1所示，该方法包括：

[0037] 步骤S101，将输入的目标关键字与候选集中的索引标识进行匹配，得到匹配结果。

[0038] 这里，当检测到用户输入需要搜索的内容时，可以将用户输入的内容进行解析，并从需要搜索的内容中提取目标关键字，其中，需要搜索的内容可以为文本数据、图片或者其他多媒体信息，只要搜索引擎能够识别，并且从中获取到相应的目标关键字即可。该目标关键字可以为数字、词语、单词或者单个英文字母等。当从需要搜索的内容中解析出目标关键字之后，将目标关键字与候选集中的索引标识进行匹配，并获得匹配结果，其中，候选集用于表征目标联想词和索引标识的对应关系。

[0039] 例如，从用户搜索的内容中解析出来的目标关键字为“o”时，则将“o”与候选集中的索引标识进行匹配，如果候选集包括有“o”的索引标识和该索引标识对应的联想词“你好，好”，则获取“o”所对应的联想词为“你好，好”。

[0040] 步骤S102，根据所述匹配结果，确定所述目标关键字对应的至少一个目标联想词。

[0041] 这里，候选集中存储有目标联想词和索引标识，可以通过轮循的方式将获取到的目标关键字和候选集中的索引标识一一匹配，获取到相应的匹配结果，并根据匹配结果确定目标关键字对应的至少一个目标联想词。

[0042] 步骤S103，根据预设规则，从所述至少一个目标联想词中确定对所述目标关键字的处理结果。

[0043] 这里，预设规则可以是将确定的至少一个目标联想词进行排序，并根据排序结果从所有的目标联想词中确定对目标关键字的处理结果，对目标关键字的处理结果可以为获取与该目标关键字关联关系最密切的内容。

[0044] 本申请实施例中，将输入的目标关键字与候选集中的索引标识进行匹配，得到匹配结果，根据匹配结果，确定目标关键字对应的至少一个目标联想词，根据预设规则，从至少一个目标联想词中确定对目标关键字的处理结果，可以精确定位到目标关键字所对应的目标联想词。

[0045] 图2为本申请实施例提供的信息处理方法流程示意图二，如图2所示，该方法包括：

[0046] 步骤S201，根据获取的至少两类源数据，确定目标联想词列表。

[0047] 这里，每一类源数据的来源和/或数据属性不同，例如，源数据的来源可以为当前所访问的网站内的数据或者外网的访问数据，获取的数据的属性可以为文本数据、视频数据或自定义数据。本申请实施例中，可以根据不同源数据确定不同的源数据对应的联想词列表，然后通过合并不同的源数据对应的联想词列表的方式，确定目标联想词列表。

[0048] 在其他的实施例中，不同属性的源数据可以为文本数据、视频数据或自定义数据，其中，文本数据可以为热搜词，热搜词即为当前网站或者其他网站的点击量超过一定数值的关键词；视频数据可以为预存在视频库中的视频数据，可以从视频库中提取名称库，名称库泛指需要被搜索到的字段，包括不限于视频标题、关联的明星等；自定义数据可以为通过人工干预获取到的外部热词数据或者人工输入的数据。

[0049] 步骤S202，根据所述目标联想词列表中联想词的属性信息，确定所述联想词的索引标识，所述索引标识用于对所述联想词进行标记。

[0050] 这里，可以将目标联想词列表中的联想词进行解析，并根据一定的规则从联想词中提取属性信息，并根据该属性信息确定该联想词的索引标识，例如，对该联想词进行解析，获取该联想词的中文拼音，则可以将该中文拼音作为该联想词的索引标识，例如，如果目标联想词列表为：

[0051] 你好 100

[0052] 好 90

[0053] 则该目标联想词列表中的联想词和其对应的索引标识可以表示为：

[0054] n:你好

[0055] i:你好

[0056] h:你好,好

[0057] a:你好,好

[0058] o:你好,好

[0059] ni:你好

[0060] nih:你好

[0061] niha:你好

[0062] nihao:你好

[0063] nh:你好

[0064] ha:你好,好

[0065] hao:你好,好

[0066] 你:你好

[0067] 你好:你好

[0068] 好:你好,好

[0069] 其中，第一列中的“n;i;h;a;o;ni;nih;niha;nihao;nh;ha;hao;你;你好;好”表示的是索引标识;第二列中的“你好;你好;你好,好;你好,好;你好,好;你好;你好;你好;你好;你好,好;你好,好;你好;你好;你好,好”表示的是第一列的索引标识对应的联想词。

[0070] 步骤S203，根据所述索引标识从所述目标联想词列表中，确定相应的目标联想词。

[0071] 这里，当从联想词中提取属性信息，并根据该属性信息确定该联想词的索引标识，就可以根据索引标识确定对应的目标联想词，例如，以上述联想词和其对应的索引标识为例，当确定的索引标识为“o”，那么就能够确定其所对应的联想词为“你好,好”，即“你好,好”为根据索引标识“o”确定的目标联想词。

[0072] 步骤S204，根据所述确定的目标联想词，形成相应的候选集。

[0073] 这里，可以根据用户输入的历史内容和索引标识，确定需要存储至候选集的目标联想词，然后根据确定的目标联想词，形成相应的候选集。本申请实施例中，候选集可以为存储在独立设置的缓存区域，通过单独设置候选集，并通过候选集中转传输数据，相较于直接从目标联想词列表获取联想词信息，能够节省数据传输的路径。

[0074] 步骤S205，将输入的目标关键字与候选集中的索引标识进行匹配，得到匹配结果。

[0075] 步骤S206，根据所述匹配结果，确定所述目标关键字对应的至少一个目标联想词。

[0076] 步骤S207，根据预设规则，从所述至少一个目标联想词中确定对所述目标关键字

的处理结果。

[0077] 本申请实施例中，通过根据获取的至少两类源数据，确定目标联想词列表；根据目标联想词列表中联想词的属性信息，确定联想词的索引标识；根据索引标识从目标联想词列表中，确定相应的目标联想词；根据确定的目标联想词，形成相应的候选集，再从候选集中确定目标关键字的处理结果。通过将多个数据元结合，能够解决相关技术中覆盖率不足；效果不佳的问题。

[0078] 图3为本申请实施例提供的信息处理方法流程示意图三，如图3所示，该方法包括：

[0079] 步骤S301，根据获取的至少两类源数据，确定目标联想词列表。

[0080] 步骤S302，根据所述目标联想词列表中联想词的属性信息，确定所述联想词的索引标识，所述索引标识用于对所述联想词进行标记。

[0081] 步骤S303，根据所述索引标识从所述目标联想词列表中，确定相应的目标联想词。

[0082] 步骤S304，将所述目标联想词列表中与所述索引标识对应的联想词，确定为目标联想词。

[0083] 这里，当从联想词中提取属性信息，并根据该属性信息确定该联想词的索引标识，就可以根据索引标识确定对应的目标联想词，例如，以上述联想词和其对应的索引标识为例，当确定的索引标识为“o”，那么就能够确定其所对应的联想词为“你好，好”，并将“你好，好”确定为目标联想词。

[0084] 步骤S305，获取所述目标联想词对应的权重值。

[0085] 这里，目标联想词中还存储有目标联想词的权重值，例如，目标联想词和目标联想词对应的权重值为：

[0086] 西游记 100

[0087] 水浒传 100

[0088] 红楼梦 50

[0089] 这里，以目标联想词为“西游记”为例，则可以根据“西游记”确定的权重值为“100”。

[0090] 步骤S306，根据所述目标联想词和所述目标联想词对应的权重值，形成候选集。

[0091] 这里，可以根据用户输入的历史内容，确定需要存储至候选集的目标联想词，然后根据确定的目标联想词和该目标联想词对应的权重值，形成相应的候选集，其中，权重值能够表征联想词被访问的频率和/或次数，是通过相应的算法，根据联想词被搜索的频率和/或次数计算得到的。

[0092] 本申请实施例中，候选集可以为存储在独立设置的缓存区域，通过单独设置候选集，并通过候选集中转传输数据，相较于直接从目标联想词列表获取联想词信息，能够节省数据传输的路径。

[0093] 上述的步骤S304至步骤S306提供了一种实现步骤“根据所述确定的目标联想词，形成相应的候选集”的方式。该方式中，通过将目标联想词列表中与索引标识对应的联想词，确定为目标联想词，获取目标联想词对应的权重值；根据目标联想词和目标联想词对应的权重值，形成候选集，形成一个覆盖多种源数据的候选集，能够解决相关技术中覆盖率不足；效果不佳的问题。

[0094] 步骤S307，将输入的目标关键字与候选集中的索引标识进行匹配，得到匹配结果。

[0095] 步骤S308，根据所述匹配结果，确定所述目标关键字对应的至少一个目标联想词。

[0096] 步骤S309，根据预设规则，从所述至少一个目标联想词中确定对所述目标关键字的处理结果。

[0097] 图4为本申请实施例提供的信息处理方法流程示意图四，如图4所示，该方法包括：

[0098] 步骤S401，获取至少两类源数据，每一类所述源数据包括搜索词。

[0099] 这里，可以从获取的至少两类源数据中解析出每一类源数据所包括的搜索词。搜索词可以为关键字、词语或者英文字母等。

[0100] 步骤S402，将被搜索的频率和/或次数大于预设阈值的搜索词，确定为联想词。

[0101] 这里，统计站内的搜索词的频率和/或次数，并进行频率和/或次数过滤。这里，频率和/或次数过滤的目的是过滤掉搜索频率和/或次数过低的搜索词，即，搜索频率和/或次数过低的搜索词无需存入联想词库。过滤方式可以包含：

[0102] (1)绝对值频率和/或次数过滤：设置频率和/或次数低于N(如100)的搜索词过滤掉。

[0103] (2)按比例动态过滤：设置频率和/或次数在排序后N％(如后30％)的搜索词过滤掉。这里，N值为示例性的值，具体的阈值或比例值可以根据业务规模配置。

[0104] 步骤S403，根据所述联想词被搜索的频率和/或次数，确定所述联想词对应的权重值。

[0105] 这里，可以根据预设算法对获取待的联想词被搜索的频率和/或次数进行相应的处理和计算，得到相应的权重值。例如，根据半衰期算法对获取到的联想词被搜索的频率和/或次数进行计算，得到相应的数值，并根据标准化算法对该数值进行标准化处理，得到联想词最终的权重值。

[0106] 在其他的实施例中，如果所述源数据为文本数据，所述根据所述联想词被搜索的频率和/或次数，确定所述联想词对应的权重值，包括：获取至少一个单位时间内的联想词和所述联想词的被搜索频率和/或次数；将每一所述单位时间内的被搜索频率和/或次数进行合并，得到第一合并结果；将所述第一合并结果进行标准化处理，得到每一所述联想词的权重值。

[0107] 这里，以联想词为热搜词为例，可以获取至少一个单位时间内的联想和所述联想词的被搜索频率和/或次数，采用半衰期算法将每一所述单位时间内的被搜索频率和/或次数进行合并，半衰期算法的计算公式为公式(1)：

[0108] $$N(t)=N_0*\left(\frac{1}{2}\right)^{t/T} \quad (1);$$

[0109] 公式(1)中，N_0为获取的当天的词频，N(t)为根据当天的频值N_0进行衰减完之后的值；t为从计算词频开始到当天的总天数；T为半衰周期，可以根据业务特点进行设置，如T＝15(半个月半衰期)，而最终的点击量计算公式为公式(2)：

[0110] $$C=\sum_{t=1}^{T_1}N(t) \quad (2);$$

[0111] 公式(2)中，将t至当前周期T_1内的点击量进行求和，T_1可以根据业务特点进行设置，如T_1＝30；其中，t为从计算词频开始到当天的总天数，T_1为当前周期。经过上述步骤输出

为合并热搜词词频表。

[0112] 本申请实施例中，合并热搜词词频表可存储为文件，每行一个视频统计，格式为："query\t score"，示例如下：

[0113] 西游记 99990

[0114] 水浒传 90000

[0115] 其中，"西游记"和"水浒传"即为query对应的一列，表示热搜词；"99990"和"99990"即为t score对应的一列，表示"西游记"和"水浒传"的词频的合并值。

[0116] 然后利用权重算法得到热搜词权重表，即将上述经过半衰期算法计算出来的离散数值进行标准化，标准化结果为[0-100]之间的一个权重分值。此处的标准化可以适用于标准化方法，如"min-max法"、"Z-score法"等。就可以得到每一联想词的权重值。

[0117] 在其他的实施例中，如果所述源数据为视频数据，所述根据所述联想词被搜索的频率和/或次数，确定所述联想词对应的权重值，包括：获取至少一个单位时间内的所述视频数据被搜索的频率和/或次数，建立所述联想词与所述视频被搜索的频率和/或次数的关联关系；将每一所述单位时间内的被搜索的频率和/或次数进行合并，得到第二合并结果；将所述第二合并结果进行标准化处理，得到每一所述联想词的权重值。

[0118] 步骤S404，根据确定的联想词和所述联想词对应的权重值，形成每一类源数据对应的联想词列表。

[0119] 步骤S405，将所述每一类源数据对应的联想词列表进行合并，形成目标联想词列表。

[0120] 由于上述源数据已经进行过标准化，合并时可采用三种方法，根据业务特点选其一：

[0121] (1)"并集取max法"：即将多个数据源的数据进行取并集，相同词取权重分数高值作为权重值。

[0122] (2)"并集求和限制max法"：即将多个数据源的数据进行取并集，相同词将多个源的分值相加，相加后分值超过100的则设置权重值为100。

[0123] (3)"并集求平均值限制max法"：即将多个数据源的数据进行取并集，相同词将多个源的分值相加，相加后分值取平均值作为该词的权重值。

[0124] 权重表在基于上述处理后，还需要进行一个过滤步骤，过滤使用黑名单机制，对于一些无意义的词(标点符号、敏感词等)使用黑名单进行过滤。黑名单来源根据数据的特点总结规律进行维护过程的动态补充，以及支持人工干预。

[0125] 在其他的实施例中，所述源数据包括文本数据、视频数据和自定义数据；相应地，所述将所述每一类源数据对应的联想词列表进行合并，形成目标联想词列表，包括：将所述文本数据、所述视频数据和所述自定义数据所对应的联想词列表中的联想词和权重值分别进行合并，生成合并联想词和所述合并联想词对应的权重值；根据所述合并联想词和所述合并联想词对应的权重值，形成目标联想词列表。

[0126] 上述的步骤S401至步骤S405提供了一种实现步骤"根据获取的至少两类源数据，确定目标联想词列表"的方式。该方式中，通过获取不同的源数据，并根据不同的源数据形成目标联想词列表，能够覆盖更广的资源，为用户搜索的词语提供更全面的响应数据。通过将多个数据元结合，能够解决相关技术中覆盖率不足、效果不佳的问题。

[0127] 步骤S406，根据所述目标联想词列表中联想词的属性信息，确定所述联想词的索引标识，所述索引标识用于对所述联想词进行标记。

[0128] 步骤S407，根据所述索引标识从所述目标联想词列表中，确定相应的目标联想词。

[0129] 步骤S408，根据所述确定的目标联想词，形成相应的候选集。

[0130] 步骤S409，将输入的目标关键字与候选集中的索引标识进行匹配，得到匹配结果。

[0131] 步骤S410，根据所述匹配结果，确定所述目标关键字对应的至少一个目标联想词。

[0132] 步骤S411，根据预设规则，从所述至少一个目标联想词中确定对所述目标关键字的处理结果。

[0133] 图5为本申请实施例提供的信息处理方法流程示意图五，如图5所示，该方法包括：

[0134] 步骤S501，获取至少两类源数据，每一类所述源数据包括搜索词。

[0135] 步骤S502，将被搜索的频率和/或次数大于预设阈值的搜索词，确定为联想词。

[0136] 步骤S503，根据所述联想词被搜索的频率和/或次数，确定所述联想词对应的权重值。

[0137] 步骤S504，根据确定的联想词和所述联想词对应的权重值，形成每一类源数据对应的联想词列表。

[0138] 步骤S505，将所有类别的源数据所对应的联想词列表中的联想词和权重值分别进行合并，生成合并联想词和所述合并联想词对应的权重值。

[0139] 这里，还可以包括文本数据、视频数据、外部热词数据和人工干预数据，通过将这些数据对应的联想词列表进行合并处理，能够得到合并联想词和该合并联想词对应的权重值。本申请实施例中，可以通过以下方式进行合并：

[0140] (1)“并集取max法”：即将多个数据源的数据进行取并集，相同词取权重分数高值作为权重值。

[0141] (2)“并集求和限制max法”：即将多个数据源的数据进行取并集，相同词将多个源的分值相加，相加后分值超过100的则设置权重值为100。

[0142] (3)“并集求平均值限制max法”：即将多个数据源的数据进行取并集，相同词将多个源的分值相加，相加后分值取平均值作为该词的权重值。

[0143] 步骤S506，根据所述合并联想词和所述合并联想词对应的权重值，形成目标联想词列表。

[0144] 这里，例如，生成的目标联想词列表为：

[0145] 西游记 100

[0146] 水浒传 100

[0147] 红楼梦 90

[0148] 则其中“西游记”、“水浒传”和“红楼梦”表示为合并联想词，“100”、“100”和“90”为合并联想词对应的权重值。

[0149] 上述的步骤S505至步骤S506提供了一种实现步骤“将所述每一类源数据对应的联想词列表进行合并，形成目标联想词列表”的方式。该方式中，通过将所有类别的源数据所对应的联想词列表中的联想词和权重值分别进行合并，生成合并联想词和合并联想词对应的权重值，并根据合并联想词和合并联想词对应的权重值，形成目标联想词列表，能够实现数据的全面覆盖，并且得到的权重值更加精确。

[0150] 步骤S507，根据所述目标联想词列表中联想词的属性信息，确定所述联想词的索引标识，所述索引标识用于对所述联想词进行标记。

[0151] 步骤S508，根据所述索引标识从所述目标联想词列表中，确定相应的目标联想词。

[0152] 这里，当从联想词中提取属性信息，并根据该属性信息确定该联想词的索引标识，就可以根据索引标识确定对应的目标联想词，例如，以上述联想词和其对应的索引标识为例，当确定的索引标识为“o”，那么就能够确定其所对应的联想词为“你好，好”，即“你好，好”为根据索引标识“o”确定的目标联想词。

[0153] 步骤S509，根据所述确定的目标联想词，形成相应的候选集。

[0154] 步骤S510，将输入的目标关键字与候选集中的索引标识进行匹配，得到匹配结果。

[0155] 步骤S511，根据所述匹配结果，确定所述目标关键字对应的至少一个目标联想词。

[0156] 步骤S512，根据预设规则，从所述至少一个目标联想词中确定对所述目标关键字的处理结果。

[0157] 图6为本申请实施例提供的信息处理方法流程示意图六，如图6所示，该方法包括：

[0158] 步骤S601，将输入的目标关键字与候选集中的索引标识进行匹配，得到匹配结果。

[0159] 步骤S602，如果匹配成功，从所述候选集中确定所述目标关键字对应的至少一个目标联想词。

[0160] 这里，查询候选集缓存，若缓存命中则返回获取结果，否则执行步骤S603。

[0161] 步骤S603，如果匹配不成功，将所述目标关键字与所述目标联想词列表中的索引标识进行匹配，得到第二匹配结果；根据所述第二匹配结果，确定所述目标关键字对应的至少一个目标联想词。

[0162] 这里，从目标联想词列表中获取候选集，若获取到，则返回结果，并将结果写入缓存库，未获取到则返回空。

[0163] 上述的步骤S602至步骤S603提供了一种实现步骤“根据所述匹配结果，确定所述目标关键字对应的至少一个目标联想词”的方式。该方式中，通过将输入的目标关键字与候选集中的索引标识进行匹配，得到匹配结果，根据匹配结果，确定目标关键字对应的至少一个目标联想词，根据预设规则，从至少一个目标联想词中确定对目标关键字的处理结果，可以精确定位到目标关键字所对应的目标联想词。

[0164] 步骤S604，据预设规则，从所述至少一个目标联想词中确定对所述目标关键字的处理结果。

[0165] 图7为本申请实施例提供的信息处理方法流程示意图七，如图7示，该方法包括：

[0166] 步骤S701，将输入的目标关键字与候选集中的索引标识进行匹配，得到匹配结果。

[0167] 步骤S702，根据所述匹配结果，确定所述目标关键字对应的至少一个目标联想词。

[0168] 步骤S703，根据所述至少一个目标联想词，形成目标联想词集合。

[0169] 这里，以目标关联词是标题为例，则取对应视频的标签列表，若为非标题，则直接使用该词，形成列表list1，这里，标签列表为该视频预存在视频库中的标题列表。

[0170] 步骤S704，根据目标用户输入的历史内容，构建目标用户信息集合。

[0171] 这里，可以根据预设的用户画像中获取到用户的标签列表list2，用户画像即为预存的或者预先获取的用户相关的参数信息，例如，用户所访问的网站，以及用户浏览的数据等。

[0172] 步骤S705,获取所述目标联想词集合和目标用户信息集合的相似度。

[0173] 这里,根据相似度算法公式计算步骤S703和S704中的list1和list2的相似度。相似度的计算公式为公式(3):

[0174] $$\text{smilarity}(list1,\ list2)=\frac{\prod(list1)\cap\prod(list2)}{\prod(list1)\cup\prod(list2)} \quad (3);$$

[0175] 公式(3)中,smilarity(list1,list2)表示list1和list2的相似度值;∏为求乘积符号;∩为求交集;∪为求并集。

[0176] 步骤S706,根据所述相似度对所述目标联想词集合中的目标联想词进行排序,得到排序结果。

[0177] 这里,根据步骤S705中的公式(3)计算出list1和list2的相似度值之后,可以将计算出来的相似度值进行排序。在其他的实施例中,也可以根据其他的排序算法进行排序,例如,LR排序模型算法。

[0178] 在其他的实施例中,如果计算得到的相似值为零,则直接获取目标联想词对应的权重值,并根据获取的权重值进行排序,根据权重值的排序结果,从目标联想词集合中确定对所述目标关键字的处理结果。

[0179] 步骤S707,根据所述排序结果,从所述目标联想词集合中确定对所述目标关键字的处理结果。

[0180] 上述的步骤S703至步骤S707提供了一种实现步骤“根据预设规则,从所述至少一个目标联想词中确定对所述目标关键字的处理结果”的方式。该方式中,将输入的目标关键字与候选集中的索引标识进行匹配,得到匹配结果,根据匹配结果,确定目标关键字对应的至少一个目标联想词,根据预设规则,从至少一个目标联想词中确定对目标关键字的处理结果,可以精确定位到目标关键字所对应的目标联想词。并且将多因素进行融合,结合用户的喜好,能够克服现有技术中排序方法单一的问题。

[0181] 本申请实施例中,提供了一种信息处理的方法,图8为本申请实施例提供的一种目标联想词列表的生成流程示意图,这里,目标联想词列表也可称为联想词权重表。如图8所示,生成目标联想词列表的步骤包括:

[0182] 步骤S801,通过热搜词日志,获得热搜词权重表。

[0183] 这里,热搜词日志为记录热搜词的文档,本申请实施例中,可以通过从网站获取到相应的热搜词,并根据相应的规则获取到热搜词的权重表。

[0184] 图8A为本申请实施例提供的一种通过热搜词日志获得热搜词权重表的流程示意图,如图8A所示,通过热搜词日志获得热搜词权重表主要包括如下步骤:

[0185] 步骤S801a,获取单位时间的热搜词日志,统计站内的搜索词频率和/或次数,并进行频率和/或次数过滤。

[0186] 这里,频率和/或次数过滤的目的是过滤掉搜索频率和/或次数过低的搜索词,即,搜索频率和/或次数过低的搜索词无需存入联想词库。过滤方式可以包含:

[0187] (1)绝对值频率和/或次数过滤;设置频率和/或次数低于N(如100)的搜索词过滤掉。

[0188] (2)按比例动态过滤;设置频率和/或次数在排序后N%(如后30%)的搜索词过滤掉。

[0189] 这里,N值为示例性的值,具体的阈值或比例值可以根据业务规模配置。本步骤的输出是热搜词词频表,这里,热搜词词频表即为热搜词对应的联想词列表。

[0190] 本申请实施例中,热搜词对应的联想词列表可存储为文件,每行一个视频统计,格式为:“query\t count”,示例如下:

[0191] 西游记 9999

[0192] 水浒传 9990

[0193] 其中,“西游记”和“水浒传”即为query对应的一列,表示热搜词;“9999”和“9990”即为t count对应的一列,表示“西游记”和“水浒传”被搜索的次数。

[0194] 步骤S802a,根据半衰期算法对单位时间词频进行合并,获得合并热搜词词频表。

[0195] 本申请实施例中,基于前述步骤S801a中的计算仅针对一个单位时间(例如1天、2天等)的数据,每天凌晨计算前一天的数据。而热搜词的被搜索的次数(词频)是一个随时间累积的过程,并且也是时间敏感的,即最近的词频对于用户输入联想的影响大于历史词频。因此需要将单位时间的数据合并至历史数据中,形成合并热搜词词频表。合并算法采用半衰期算法,半衰期算法的计算公式为公式(4):

[0196] $$N(t)=N_0*\left(\frac{1}{2}\right)^{t/T} \quad (4);$$

[0197] 公式(4)中,N_0为获取的当天的词频,N(t)为根据当天的频值N_0进行衰减完之后的值;t为从计算词频开始到当天的总天数;T为半衰周期,可以根据业务特点进行设置,如T=15(半个月半衰期),而最终的点击量计算方式为公式(5):

[0198] $$C=\sum_{t=1}^{T_1}N(t) \quad (5);$$

[0199] 公式(5)中,将t至当前周期T_1内的点击量进行求和,T_1可以根据业务特点进行设置,如T_1=30;其中,t为从计算词频开始到当天的总天数,T_1为当前周期。

[0200] 经过上述步骤输出为合并热搜词词频表。

[0201] 本申请实施例中,合并热搜词词频表可存储为文件,每行一个视频统计,格式为:“query\t score”,示例如下:

[0202] 西游记 99990

[0203] 水浒传 90000

[0204] 其中,“西游记”和“水浒传”即为query对应的一列,表示热搜词;“99990”和“99990”即为t score对应的一列,表示“西游记”和“水浒传”的词频的合并值。

[0205] 步骤S803a,根据合并热搜词词频表,利用权重算法得到热搜词权重表。

[0206] 这里,将合并的离散数值进行标准化,标准化结果为[0-100]之间的一个权重分值。此处的标准化可以适用于业界常见的标准化方法,如“min-max法”、“Z-score法”等。

[0207] 以标准化方法为“min-max法”为例,依据视频业务的特点,本申请实施例中采用一种改进的“min-max”优化方法。标准“min-max”将最大值max作为100,对于异常值效果不好,如某一搜索词的搜索量明显高于其它搜索词一个量级,则标准化后会导致分数都太低,区分度小,影响效果。改进的“min-max”法则在max的计算上采取“95值”方式,即按词频表,取词频从高到低的第5%(即高于95%的视频)的点击量作为max,高于max的点击量都统一记

录为100分，其余视频则继续按“min-max”法将词频表标准化到[0-100]之间。经过上述步骤之后输出为热搜词权重表。

[0208] 本申请实施例中，热搜词权重表可存储为文件，每行一个视频统计，格式为：“query\t score”，示例如下：

[0209] 西游记 100

[0210] 水浒传 100

[0211] ……

[0212] 红楼梦 50

[0213] 其中，“西游记”、“水浒传”和“红楼梦”即为query对应的一列，表示热搜词；“100”、“100”和“50”即为t score对应的一列，表示“西游记”、“水浒传”和“红楼梦”的权重值。

[0214] 步骤S802，通过视频库片名和视频点击量，获得片名权重表。

[0215] 图8B为本申请实施例提供的一种通过视频库片名和视频点击量获得片名权重表的流程示意图，如图8B所示，通过热搜词日志获得热搜词权重表主要包括如下步骤：

[0216] 步骤S801b，提取单位时间的片名词频。

[0217] 本申请实施例中，通过从视频库中提取名称库，并且关联点击量，即词频。实施中名称库泛指需要被搜索到的字段，包括不限于视频标题；关联的明星等。提取规则是将视频库中的所有符合业务规则的视频标题均提取出来，再将点击量文件与之匹配，得到视频点击量列表。对于无点击量的视频，点击量设置为0。

[0218] 步骤S802b，根据半衰期对单位时间片名词频进行合并，获得片名词频表。

[0219] 本申请实施例中，基于前述步骤S801b中的计算仅针对一个单位时间(例如1天、2天等)的数据，每天凌晨计算前一天的数据。而热搜词的被搜索的次数(词频)是一个随时间累积的过程，并且也是时间敏感的，即最近的词频对于用户输入联想的影响大于历史词频。因此需要将单位时间的数据合并至历史数据中，形成合并热搜词词频表。合并算法采用半衰期算法，半衰期算法的计算公式为公式(6)：

[0220] $$N(t)=N_0*\left(\frac{1}{2}\right)^{t/T} \quad (6);$$

[0221] 公式(6)中，N_0为获取的当天的词频，N(t)为根据当天的频值N_0进行衰减完之后的值；t为从计算词频开始到当天的总天数；T为半衰周期，可以根据业务特点进行设置，如T＝15(半个月半衰期)，而最终的点击量计算方式为公式(7)：

[0222] $$C=\sum_{t=1}^{T_1}N(t) \quad (7);$$

[0223] 公式(7)中，将t至当前周期T_1内的点击量进行求和，T_1可以根据业务特点进行设置，如T_1＝30；其中，t为从计算词频开始到当天的总天数，T_1为当前周期。

[0224] 经过上述步骤输出为合并片名词频表。

[0225] 本申请实施例中，合并片名词频表可存储为文件，每行一个视频统计，格式为：“query\t score”，示例如下：

[0226] 西游记 2333

[0227] 水浒传 2200

[0228] 其中，“西游记”和“水浒传”即为query对应的一列，表示热搜词；“2333”和“2200”即为t score对应的一列，表示“西游记”和“水浒传”的词频的合并值。

[0229] 步骤S803b，根据合并片名词频表，利用权重算法得到片名权重表。

[0230] 这里，根据步骤S802b得到的合并片名词频表，需要根据权重算法将其进行标准化至[M,100]之间的一个权重分数。

[0231] 此处标准化最小值不取0，而取M(0<M<100)，原因是视频片名为强匹配需求，即片名应该尽可能被联想出来，因此需要默认将片名的分值提升。M的取值根据业务特点进行设置，设置在[60,85]之间会取得比较好的效果。

[0232] 本申请实施例中，依据视频业务的特点，采用一种改进的“min-max”优化方法。标准“min-max”将最大值max作为100，对于异常值效果不好，如某一部热播视频点击量明显高于其它视频一个量级，则标准化后会导致分数都太低，区分度小，影响效果。改进的“min-max”法则在max的计算上采取“95值”方式，即按词频表，取词频从高到低的第5％(即高于95％的视频)的点击量作为max，高于max的点击量都统一记录为100分。其余视频则继续按“min-max”法将词频表标准化到[M-100]之间。基于上述步骤输出为片名权重表。

[0233] 本申请实施例中，片名权重表可存储为文件，每行一个视频统计，格式为：“video_name\t score”，示例如下：

[0234] 西游记 95

[0235] 水浒传 90

[0236] ……

[0237] 红楼梦 80

[0238] 其中，“西游记”、“水浒传”和“红楼梦”即为video_name对应的一列片名；“95”、“90”和“80”即为t score对应的一列，表示“西游记”、“水浒传”和“红楼梦”的权重值。

[0239] 步骤S803，通过外部热词数据，获得外部热词权重表。

[0240] 本申请实施例中，通过外部热词数据获得外部热词权重表主要包括如下步骤：

[0241] 步骤1，获得外部热词数据。

[0242] 这里，由于视频站点内部热点并不同等于全网网民关心的热点，因此对于外部的热词需要做监测补充。数据源可通过外部公开热词数据爬取、数据合作或人工途径获取。

[0243] 步骤2，根据外部热词数据，生成外部热词权重表。

[0244] 这里，因外部热词通常比较少能有定量的区分，权重的设置需要根据具体数据源，设置为[M1,M2]之间(0<M1<M2<100)。实施中M1与M2差距较小，如设置为[80,90]之间。

[0245] 步骤S804，结合干预数据权重表，对以上权重表进行合并和过滤，获得联想词权重表。

[0246] 基于步骤S801至步骤S803，可以得到了热搜词权重表、片名权重表、外部热词权重表，同时，可以加入人工干预机制，该人工干预机制支持配置权重干预得到干预数据权重表，并将四种数据源合并成最终的联想词权重表。

[0247] 本申请实施例中，由于四种数据源已经进行过标准化，合并时可采用三种方法，根据业务特点选其一：

[0248] (1)“并集取max法”：即将多个数据源的数据进行取并集，如果有相同的词，取权重分数高值作为权重值。

[0249] (2)“并集求和限制max法”:即将多个数据源的数据进行取并集,相同词将多个源的分值相加,相加后分值超过100的则设置权重值为100。

[0250] (3)“并集求平均值限制max法”:即将多个数据源的数据进行取并集,相同词将多个源的分值相加,相加后分值取平均值作为该词的权重值。

[0251] 权重表在基于上述处理后,还需要进行一个过滤步骤,过滤使用黑名单机制,对于一些无意义的词(标点符号、敏感词等)使用黑名单进行过滤。黑名单来源根据数据的特点总结规律进行维护过程的动态补充,以及支持人工干预。

[0252] 本申请实施例中,输出的联想词权重表可存储为文件,每行一个视频统计,格式为:“word\t score”,示例如下:

[0253] 西游记 100

[0254] 水浒传 100

[0255] ……

[0256] 红楼梦 90

[0257] 则其中“西游记”、“水浒传”和“红楼梦”表示为联想词,“100”、“100”和“90”为该联想词对应的权重值。

[0258] 图8D为本申请实施例提供的一种联想词权重表的生成流程示意图,如图8D所示,通过对搜热词日志800d进行统计和/或筛选,得到热搜词词频表801d,根据半衰期算法得到合并热搜词词频表802d;通过权重算法对合并热搜词词频表802d进行处理,得到热搜词权重表803d。通过片名权重提取算法,根据视频点击量804d和视频库805d中的视频数据得到片名词频表806d;根据半衰期算法得到合并片名词频表807d,通过权重算法对合并片名词频表807d进行处理,得到片名权重表808d。对得到的热搜词权重表803d、片名权重表808d、外部热词数据809d和人工干预数据810d进行合并和/或过滤处理,得到联想词权重表。

[0259] 图8C为本申请实施例提供的一种联想词在线服务处理流程的示意图,如图8C所示,联想词在线服务处理流程主要包括如下步骤:

[0260] 步骤S801c,联想词内存索引构建。

[0261] 本申请实施例中,该步骤为离线过程,仅服务启动或词库更新时执行。内存索引构建模块将上一部分生成的联想词权重表生成中英文全局的带权重索引,索引采用“key:[value1,value2]”形式,其中,key为根据联想词权重表生成的每个可能的输入搜索词,含中文及拼音,value1和value2则是按照权重值由高到低排序的匹配词列表,也会考虑匹配度等其它因素value1和value2可预先设置个数上限,如10,以加快计算效率和节省存储空间。

[0262] 本申请实施例中,例如,如果表联想词列表为:

[0263] 你好 100

[0264] 好 90

[0265] 则联想词和其对应的索引标识可以表示为:

[0266] n:你好

[0267] i:你好

[0268] h:你好,好

[0269] a:你好,好

[0270] o:你好,好

[0271] ni:你好

[0272] nih:你好

[0273] niha:你好

[0274] nihao:你好

[0275] nh:你好

[0276] ha:你好,好

[0277] hao:你好,好

[0278] 你:你好

[0279] 你好:你好

[0280] 好:你好,好

[0281] 以上示例仅用于直观展示,实际实现时可采用hash_map或其他更合适的结构。

[0282] 步骤S802c,请求搜索词输入,联想词模块接收搜索词,根据搜索词获取候选集。

[0283] 这里获取候选集的过程为:

[0284] (1)查询候选集缓存,若缓存命中则返回获取结果,否则进入下述(2)。

[0285] (2)从内存索引服务中获取候选集,若获取到,则返回结果,并将结果写入缓存库,未获取到则返回空。

[0286] 步骤S803c,候选集重排序。

[0287] 这里,根据候选集结果,结合到用户画像数据,对获取的结果进行重排序。对于该过程,需要将获取到的结果与用户画像标签库进行匹配,匹配方法可以为相似度计算,或进行LR等排序模型算法。

[0288] 这里,以相似度计算方法为例,通过上一步返回的联想词候选集,候选集中的每个词进行相似度计算。词若为标题,则取对应视频的标签列表,若为非标题,则直接使用该词,为list1。从用户画像中获取到用户的标签列表list2。对list1与list2进行相似度计算。相似度计算计算公式为公式(8):

[0289] $$\mathrm{smilarity}(list1,\ list2)=\frac{\prod(list1)\cap\prod(list2)}{\prod(list1)\cup\prod(list2)} \qquad (8);$$

[0290] 公式(8)中,smilarity(list1,list2)表示list1和list2的相似度值;∏为求乘积符号;∩为求交集;∪为求并集。

[0291] 计算出来相似度后,可根据相似度对结果进行重排序。对于该过程实施过程需要考虑一下因素:

[0292] (1)性能优先:因画像引入主要用排序,若耗时过长,可能影响服务,因此考虑排序算法时性能因素更重要。

[0293] (2)基于性能原因,对于该服务使用的画像应做简化,使用较大颗粒的画像标签,类似于给用户进行分类,则最终的联想词结果可以加入缓存复用,极大降低后端排序压力和提升访问速度。

[0294] (3)基于相似度的排序开启时,可以配置从内存索引获取的候选集加大条数,高于最终需要的联想词条数(如2倍),以便在加入排序后能更好筛选出用户兴趣的候选词。

[0295] (4)相似度排序结果可考虑与原有分数做一个加权,而非仅使用相似度结果,加权

权重可根据不同业务进行配置。

[0296] 本申请实施例中，目标联想词列表生成规则，考虑了热搜词日志、视频点击量、视频库、外部热词数据、人工干预等多种数据源进行融合；热搜词的频次统计，使用了半衰期算法融入历史频次；热搜词进行标准化，采用了改进的“min-max”标准化方法，引入了“95值”方法；片名的频次统计，使用了半衰期算法融入历史频次；片名频次进行标准化，采用了改进的“min-max”标准化方法，引入了“95值”方法；权重表的合并方法，采用“并集取max法”或“并集求和限制max法”，将多种数据源进行融合处理；联想词在线服务引入基于用户画像的排序过程，使得联想词与用户兴趣进行结合，优先推荐用户兴趣度高的联想词，能够达到更好的效果。

[0297] 基于前述的实施例，本申请实施例提供一种信息处理装置，该装置所包括的各单元：以及各单元所包括的各子单元，都可以通过服务器中的处理器来实现：当然也可通过的逻辑电路实现：在实施的过程中，处理器可以为中央处理器(CPU)：微处理器(MPU)：数字信号处理器(DSP)或现场可编程门阵列(FPGA)等。

[0298] 图9为本申请实施例提供的信息处理装置的组成结构示意图，如图9所示，所述装置包括：

[0299] 匹配单元901，配置为将输入的目标关键字与候选集中的索引标识进行匹配，得到匹配结果，其中，所述候选集用于表征目标联想词和索引标识的对应关系；

[0300] 第一确定单元902，配置为根据所述匹配结果，确定所述目标关键字对应的至少一个目标联想词；

[0301] 第二确定单元903，配置为根据预设规则，从所述至少一个目标联想词中确定对所述目标关键字的处理结果。

[0302] 在其他的实施例中，所述装置还包括：第三确定单元，配置为根据获取的至少两类源数据，确定目标联想词列表；第四确定单元，配置为根据所述目标联想词列表中联想词的属性信息，确定所述联想词的索引标识，所述索引标识用于对所述联想词进行标记；第五确定单元，配置为根据所述索引标识从所述目标联想词列表中，确定相应的目标联想词；生成单元，配置为根据所述确定的目标联想词，形成相应的候选集。

[0303] 在其他的实施例中，所述生成单元，还配置为：将所述目标联想词列表中与所述索引标识对应的联想词，确定为目标联想词；获取所述目标联想词对应的权重值；根据所述目标联想词和所述目标联想词对应的权重值，形成候选集。

[0304] 在其他的实施例中，所述第三确定单元，还配置为：获取至少两类源数据，每一类所述源数据包括搜索词；将被搜索的频率和/或次数大于预设阈值的搜索词，确定为联想词；根据所述联想词被搜索的频率和/或次数，确定所述联想词对应的权重值；根据确定的联想词和所述联想词对应的权重值，形成每一类源数据对应的联想词列表；将所述每一类源数据对应的联想词列表进行合并，形成目标联想词列表。

[0305] 在其他的实施例中，所述第三确定单元，还配置为：将所有类别的源数据所对应的联想词列表中的联想词和权重值分别进行合并，生成合并联想词和所述合并联想词对应的权重值；根据所述合并联想词和所述合并联想词对应的权重值，形成目标联想词列表。

[0306] 在其他的实施例中，所述源数据包括文本数据、视频数据和自定义数据；相应地，所述第三确定单元，还配置为：将所述文本数据、所述视频数据和所述自定义数据所对应的

联想词列表中的联想词和权重值分别进行合并，生成合并联想词和所述合并联想词对应的权重值；根据所述合并联想词和所述合并联想词对应的权重值，形成目标联想词列表。

[0307] 在其他的实施例中，如果所述源数据为文本数据，所述第三确定单元，还配置为：获取至少一个单位时间内的联想词和所述联想词的被搜索频率和/或次数；将每一所述单位时间内的被搜索频率和/或次数进行合并，得到第一合并结果；将所述第一合并结果进行标准化处理，得到每一所述联想词的权重值。

[0308] 在其他的实施例中，如果所述源数据为视频数据，所述第三确定单元，还配置为：获取至少一个单位时间内的所述视频数据被搜索的频率和/或次数，建立所述联想词与所述视频被搜索的频率和/或次数的关联关系；将每一所述单位时间内的被搜索的频率和/或次数进行合并，得到第二合并结果；将所述第二合并结果进行标准化处理，得到每一所述联想词的权重值。

[0309] 在其他的实施例中，所述第一确定单元，还配置为：如果匹配成功，从所述候选集中确定所述目标关键字对应的至少一个目标联想词；如果匹配不成功，将所述目标关键字与所述目标联想词列表中的索引标识进行匹配，得到第二匹配结果；根据所述第二匹配结果，确定所述目标关键字对应的至少一个目标联想词。

[0310] 在其他的实施例中，所述第二确定单元，还配置为：根据所述至少一个目标联想词，形成目标联想词集合；根据目标用户输入的历史内容，构建目标用户信息集合；获取所述目标联想词集合和目标用户信息集合的相似度；根据所述相似度对所述目标联想词集合中的目标联想词进行排序，得到排序结果；根据所述排序结果，从所述目标联想词集合中确定对所述目标关键字的处理结果。

[0311] 需要说明的是，本申请实施例中，如果以软件功能模块的形式实现上述信息处理的方法，并作为独立的产品销售或使用时，也可以存储在一个计算机可读取存储介质中。基于这样的理解，本申请实施例的技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质中，包括若干指令用以使得一台服务器执行本申请各个实施例所述方法的全部或部分。而前述的存储介质包括：U盘、移动硬盘、只读存储器(Read Only Memory，ROM)、磁碟或者光盘等各种可以存储程序代码的介质。这样，本申请实施例不限制于任何特定的硬件和软件结合。

[0312] 对应地，本申请实施例提供一种计算机可读存储介质，其上存储有计算机程序，该计算机程序被处理器执行时实现上述实施例提供的信息处理的方法中的步骤。

[0313] 本申请装置实施例的描述，与上述方法实施例的描述是类似的，具有同方法实施例相似的有益效果。对于本申请装置实施例中未披露的技术细节，请参照本申请方法实施例的描述而理解。

[0314] 这里需要指出的是：以上存储介质和设备实施例的描述，与上述方法实施例的描述是类似的，具有同方法实施例相似的有益效果。对于本申请存储介质和设备实施例中未披露的技术细节，请参照本申请方法实施例的描述而理解。

[0315] 需要说明的是，图10为本申请实施例提供的一种计算机设备结构示意图，如图10所示，该计算机设备1000至少包括：处理器1001、通信接口1002和存储器1003，其中

[0316] 处理器1001通常控制计算机设备1000的总体操作。

[0317] 通信接口1002可以使计算机设备通过网络与其他计算机设备或服务器通信。

[0318] 存储器1003配置为存储由处理器1001可执行的指令和应用，还可以缓存待处理器1001以及计算机设备1000中各模块待处理或已经处理的数据(例如，图像数据、音频数据、语音通信数据和视频通信数据)，可以通过闪存(FLASH)或随机访问存储器(Random Access Memory，RAM)实现。

[0319] 当然，本申请实施例中的装置还可有其他类似的协议交互实现案例，在不背离本申请精神及其实质的情况下，本领域的技术人员当可根据本申请实施例做出各种相应的改变和变形，但这些相应的改变和变形都应属于本申请方法所附的权利要求的保护范围。

[0320] 本领域内的技术人员应明白，本申请的实施例可提供为方法、系统、或计算机程序产品。因此，本申请可采用硬件实施例、软件实施例、或结合软件和硬件方面的实施例的形式。而且，本申请可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质(包括但不限于磁盘存储器和光学存储器等)上实施的计算机程序产品的形式。

[0321] 本申请是参照根据本申请实施例的方法、设备(系统)、和计算机程序产品的流程图和/或方框图来描述的。应理解可由计算机程序指令实现流程图和/或方框图中的每一流程和/或方框、以及流程图和/或方框图中的流程和/或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理设备的处理器以产生一个机器，使得通过计算机或其他可编程数据处理设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的设备。

[0322] 这些计算机程序指令也可存储在能引导计算机或其他可编程数据处理设备以特定方式工作的计算机可读存储器中，使得存储在该计算机可读存储器中的指令产生包括指令设备的制造品，该指令设备实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能。

[0323] 这些计算机程序指令也可装载到计算机或其他可编程数据处理设备上，使得在计算机或其他可编程设备上执行一系列操作步骤以产生计算机实现的处理，从而在计算机或其他可编程设备上执行的指令提供用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的步骤。

[0324] 应理解，说明书通篇中提到的“一个实施例”或“一实施例”意味着与实施例有关的特定特征、结构或特性包括在本申请的至少一个实施例中。因此，在整个说明书各处出现的“在一个实施例中”或“在一实施例中”未必一定指相同的实施例。此外，这些特定的特征、结构或特性可以任意适合的方式结合在一个或多个实施例中。应理解，在本申请的各种实施例中，上述各过程的序号的大小并不意味着执行顺序的先后，各过程的执行顺序应以其功能和内在逻辑确定，而不应对本申请实施例的实施过程构成任何限定。上述本申请实施例序号仅仅为了描述，不代表实施例的优劣。

[0325] 需要说明的是，在本文中，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、物品或者装置不仅包括那些要素，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、物品或者装置所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括该要素的过程、方法、物品或者装置中还存在另外的相同要素。

[0326] 在本申请所提供的几个实施例中，应该理解到，所揭露的设备和方法，可以通过其它的方式实现。以上所描述的设备实施例仅仅是示意性的，例如，所述模块的划分，仅仅为

一种逻辑功能划分，实际实现时可以有另外的划分方式，如：多个模块或组件可以结合，或可以集成到另一个系统，或一些特征可以忽略，或不执行。另外，所显示或讨论的各组成部分相互之间的耦合、或直接耦合、或通信连接可以是通过一些接口，设备或模块的间接耦合或通信连接，可以是电性的、机械的或其它形式的。

[0327] 上述作为分离部件说明的模块可以是、或也可以不是物理上分开的，作为模块显示的部件可以是、或也可以不是物理模块；既可以位于一个地方，也可以分布到多个网络模块上；可以根据实际的需要选择其中的部分或全部模块来实现本实施例方案的目的。

[0328] 另外，在本申请各实施例中的各功能模块可以全部集成在一个处理模块中，也可以是各模块分别单独作为一个模块，也可以两个或两个以上模块集成在一个模块中；上述集成的模块既可以采用硬件的形式实现，也可以采用硬件加软件功能模块的形式实现。

[0329] 本领域普通技术人员可以理解：实现上述方法实施例的全部或部分步骤可以通过程序指令相关的硬件来完成，前述的程序可以存储于计算机可读取存储介质中，该程序在执行时，执行包括上述方法实施例的步骤；而前述的存储介质包括：移动存储设备、只读存储器(Read Only Memory，ROM)、磁碟或者光盘等各种可以存储程序代码的介质。

[0330] 或者，本申请上述集成的模块如果以软件功能模块的形式实现并作为独立的产品销售或使用时，也可以存储在一个计算机可读取存储介质中。基于这样的理解，本申请实施例的技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质中，包括若干指令用以使得一台服务器执行本申请各个实施例所述方法的全部或部分。而前述的存储介质包括：移动存储设备、ROM、磁碟或者光盘等各种可以存储程序代码的介质。

[0331] 以上所述，仅为本申请的实施方式，但本申请的保护范围并不局限于此，任何熟悉本技术领域的技术人员在本申请揭露的技术范围内，可轻易想到变化或替换，都应涵盖在本申请的保护范围之内。因此，本申请的保护范围应以所述权利要求的保护范围为准。

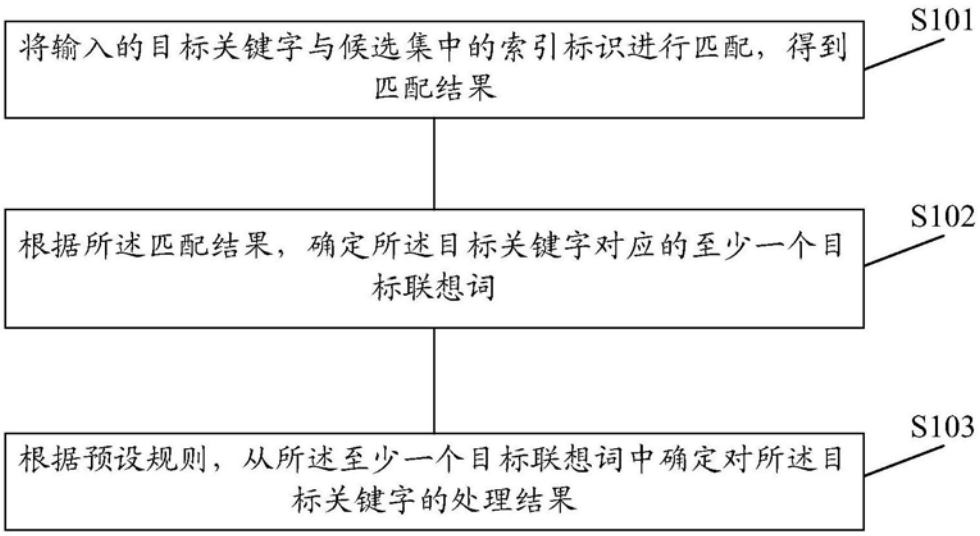

S101
将输入的目标关键字与候选集中的索引标识进行匹配，得到匹配结果
S102
根据所述匹配结果，确定所述目标关键字对应的至少一个目标联想词
S103
根据预设规则，从所述至少一个目标联想词中确定对所述目标关键字的处理结果

图1

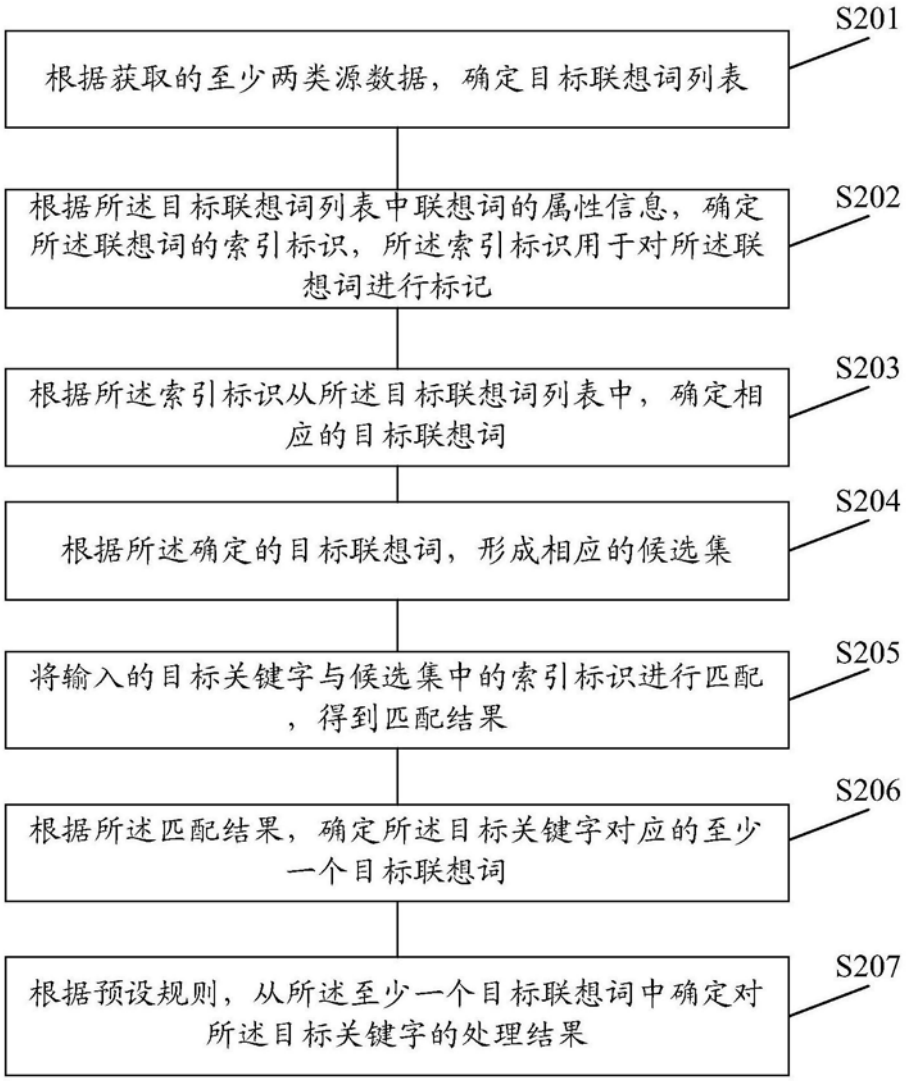

S201
根据获取的至少两类源数据，确定目标联想词列表
S202
根据所述目标联想词列表中联想词的属性信息，确定所述联想词的索引标识，所述索引标识用于对所述联想词进行标记
S203
根据所述索引标识从所述目标联想词列表中，确定相应的目标联想词
S204
根据所述确定的目标联想词，形成相应的候选集
S205
将输入的目标关键字与候选集中的索引标识进行匹配，得到匹配结果
S206
根据所述匹配结果，确定所述目标关键字对应的至少一个目标联想词
S207
根据预设规则，从所述至少一个目标联想词中确定对所述目标关键字的处理结果

图2


S301
根据获取的至少两类源数据，确定目标联想词列表
S302
根据所述目标联想词列表中联想词的属性信息，确定所述联想词的索引标识，所述索引标识用于对所述联想词进行标记
S303
根据所述索引标识从所述目标联想词列表中，确定相应的目标联想词
S304
将所述目标联想词列表中与所述索引标识对应的联想词，确定为目标联想词
S305
获取所述目标联想词对应的权重值
S306
根据所述目标联想词和所述目标联想词对应的权重值，形成候选集
S307
将输入的目标关键字与候选集中的索引标识进行匹配，得到匹配结果
S308
根据所述匹配结果，确定所述目标关键字对应的至少一个目标联想词
S309
根据预设规则，从所述至少一个目标联想词中确定对所述目标关键字的处理结果

图3

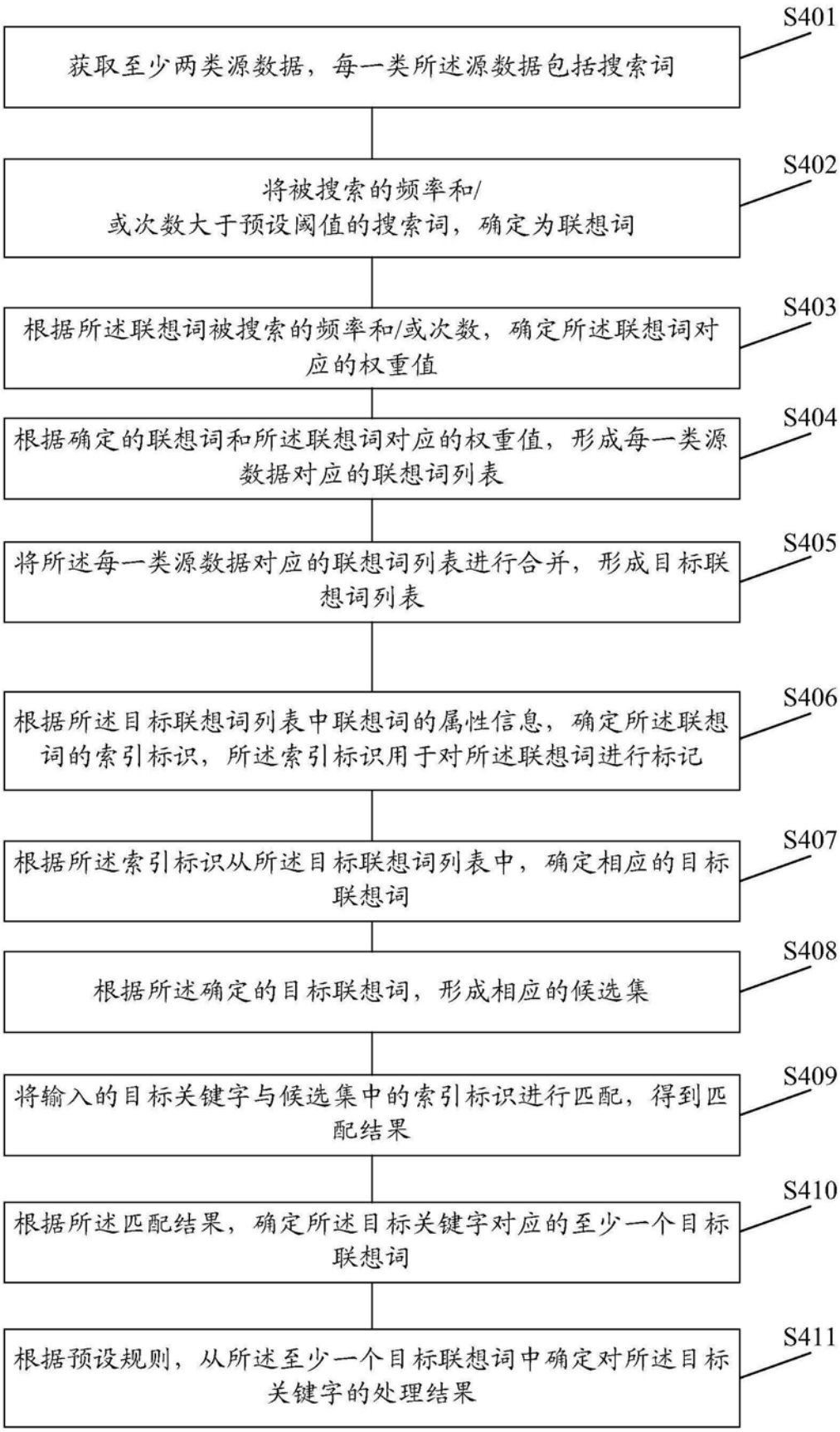

S401
获取至少两类源数据，每一类所述源数据包括搜索词
S402
将被搜索的频率和/
或次数大于预设阈值的搜索词，确定为联想词
S403
根据所述联想词被搜索的频率和/或次数，确定所述联想词对应的权重值
S404
根据确定的联想词和所述联想词对应的权重值，形成每一类源数据对应的联想词列表
S405
将所述每一类源数据对应的联想词列表进行合并，形成目标联想词列表
S406
根据所述目标联想词列表中联想词的属性信息，确定所述联想词的索引标识，所述索引标识用于对所述联想词进行标记
S407
根据所述索引标识从所述目标联想词列表中，确定相应的目标联想词
S408
根据所述确定的目标联想词，形成相应的候选集
S409
将输入的目标关键字与候选集中的索引标识进行匹配，得到匹配结果
S410
根据所述匹配结果，确定所述目标关键字对应的至少一个目标联想词
S411
根据预设规则，从所述至少一个目标联想词中确定对所述目标关键字的处理结果

图4

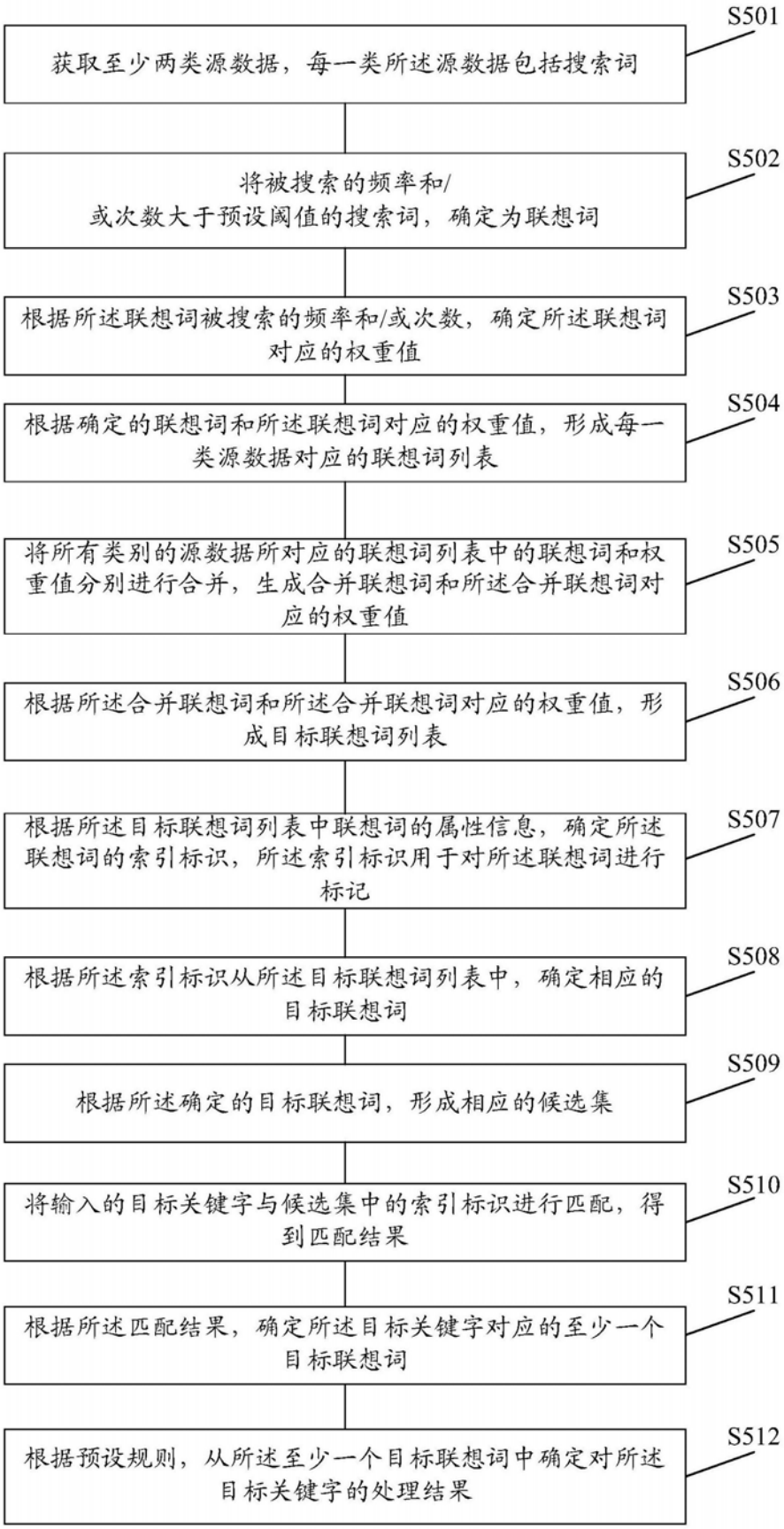

S501
获取至少两类源数据，每一类所述源数据包括搜索词
S502
将被搜索的频率和/
或次数大于预设阈值的搜索词，确定为联想词
S503
根据所述联想词被搜索的频率和/或次数，确定所述联想词对应的权重值
S504
根据确定的联想词和所述联想词对应的权重值，形成每一类源数据对应的联想词列表
S505
将所有类别的源数据所对应的联想词列表中的联想词和权重值分别进行合并，生成合并联想词和所述合并联想词对应的权重值
S506
根据所述合并联想词和所述合并联想词对应的权重值，形成目标联想词列表
S507
根据所述目标联想词列表中联想词的属性信息，确定所述联想词的索引标识，所述索引标识用于对所述联想词进行标记
S508
根据所述索引标识从所述目标联想词列表中，确定相应的目标联想词
S509
根据所述确定的目标联想词，形成相应的候选集
S510
将输入的目标关键字与候选集中的索引标识进行匹配，得到匹配结果
S511
根据所述匹配结果，确定所述目标关键字对应的至少一个目标联想词
S512
根据预设规则，从所述至少一个目标联想词中确定对所述目标关键字的处理结果

图5

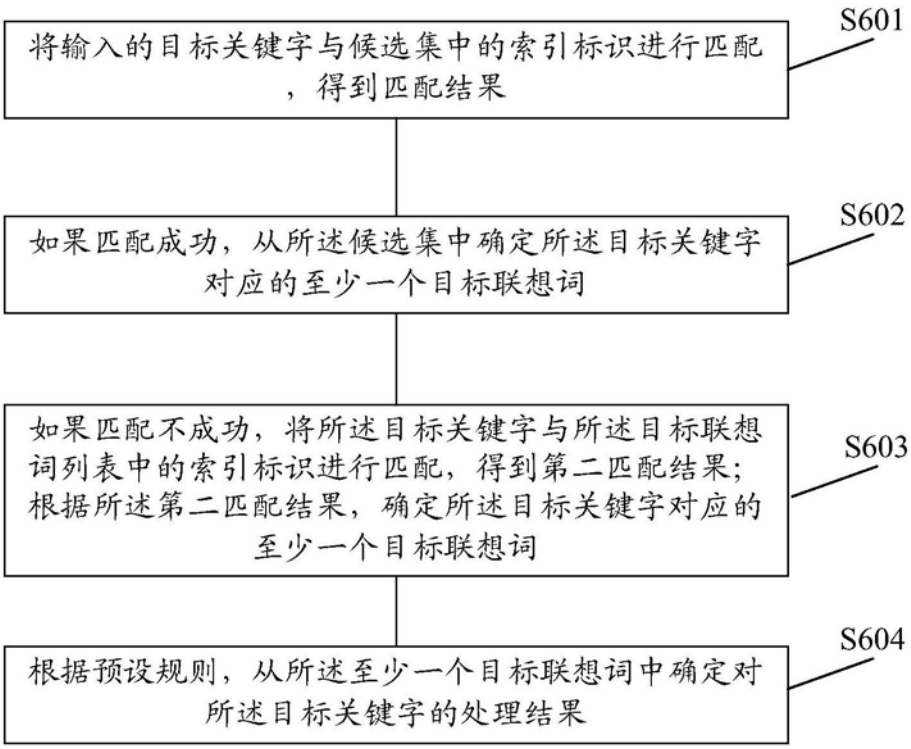

S601
将输入的目标关键字与候选集中的索引标识进行匹配，得到匹配结果
S602
如果匹配成功，从所述候选集中确定所述目标关键字对应的至少一个目标联想词
S603
如果匹配不成功，将所述目标关键字与所述目标联想词列表中的索引标识进行匹配，得到第二匹配结果；根据所述第二匹配结果，确定所述目标关键字对应的至少一个目标联想词
S604
根据预设规则，从所述至少一个目标联想词中确定对所述目标关键字的处理结果

图6

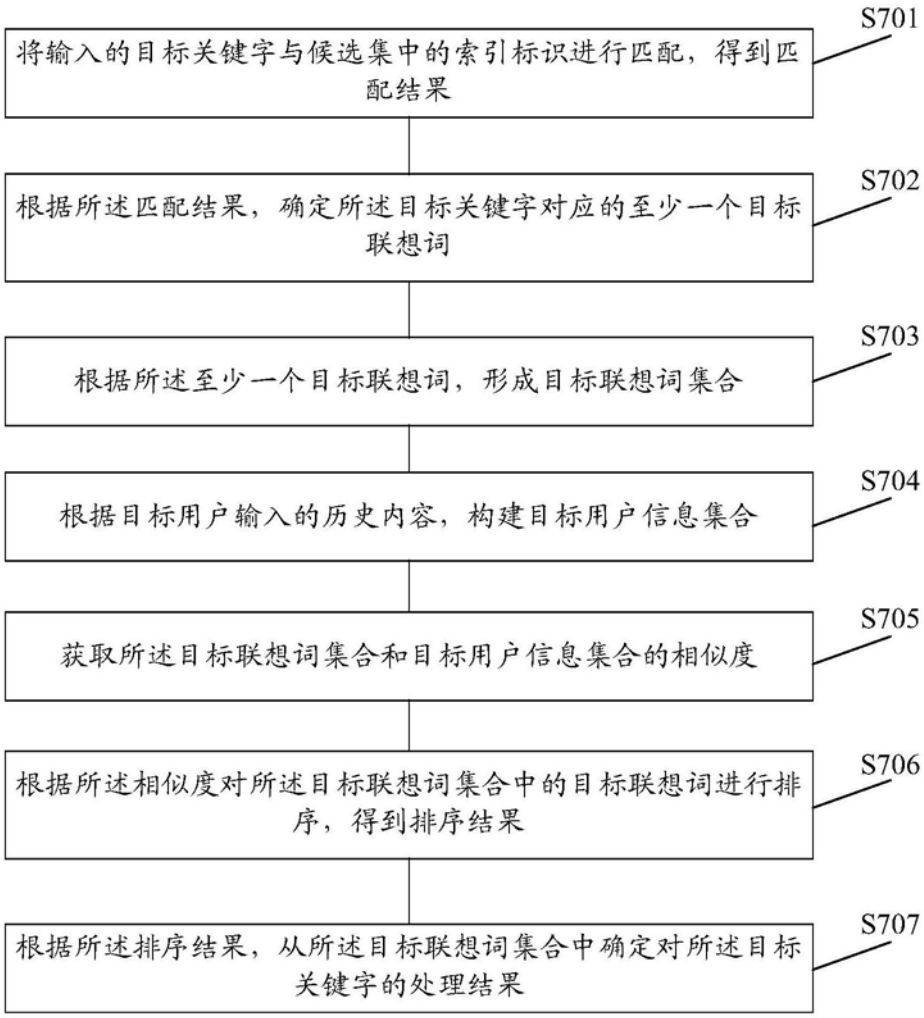

S701
将输入的目标关键字与候选集中的索引标识进行匹配，得到匹配结果
S702
根据所述匹配结果，确定所述目标关键字对应的至少一个目标联想词
S703
根据所述至少一个目标联想词，形成目标联想词集合
S704
根据目标用户输入的历史内容，构建目标用户信息集合
S705
获取所述目标联想词集合和目标用户信息集合的相似度
S706
根据所述相似度对所述目标联想词集合中的目标联想词进行排序，得到排序结果
S707
根据所述排序结果，从所述目标联想词集合中确定对所述目标关键字的处理结果

图7

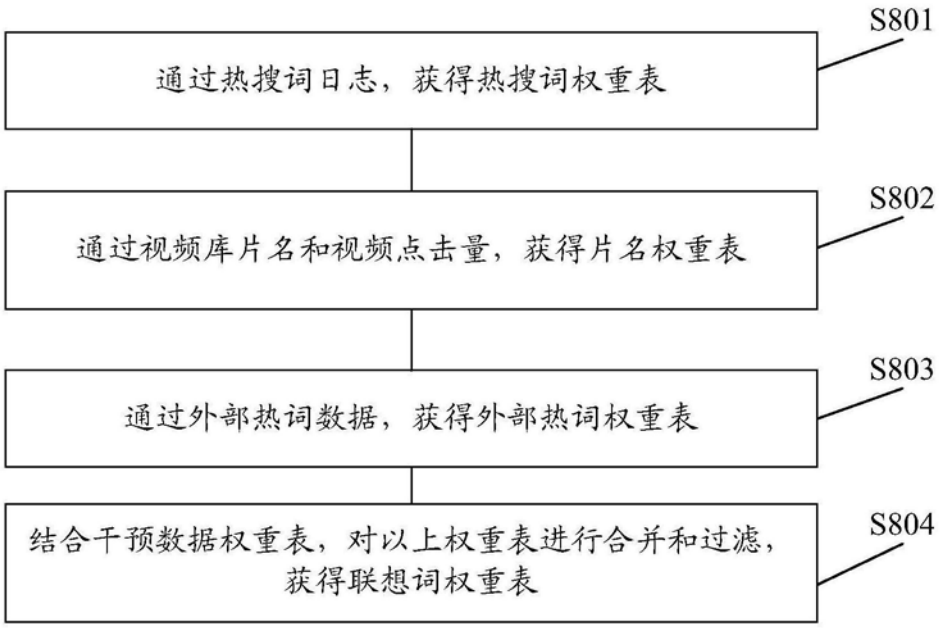

S801
通过热搜词日志，获得热搜词权重表
S802
通过视频库片名和视频点击量，获得片名权重表
S803
通过外部热词数据，获得外部热词权重表
S804
结合干预数据权重表，对以上权重表进行合并和过滤，获得联想词权重表

图8

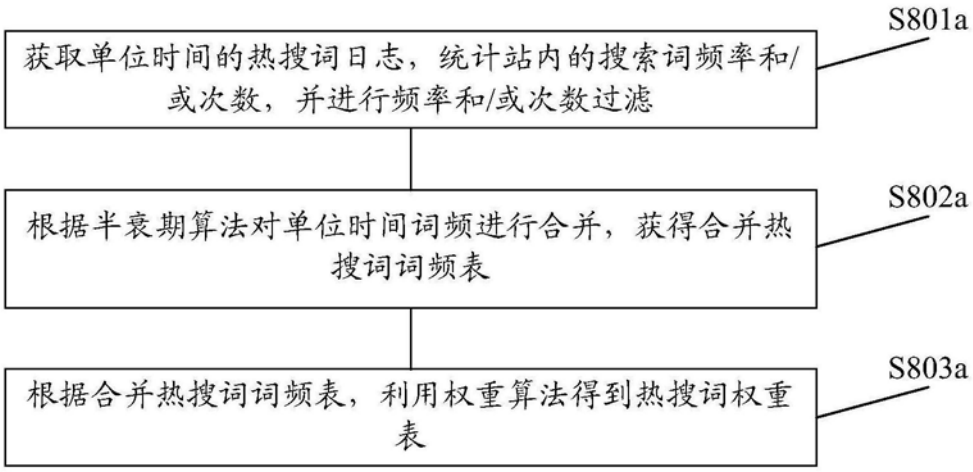

S801a
获取单位时间的热搜词日志，统计站内的搜索词频率和/或次数，并进行频率和/或次数过滤
S802a
根据半衰期算法对单位时间词频进行合并，获得合并热搜词词频表
S803a
根据合并热搜词词频表，利用权重算法得到热搜词权重表

图8A

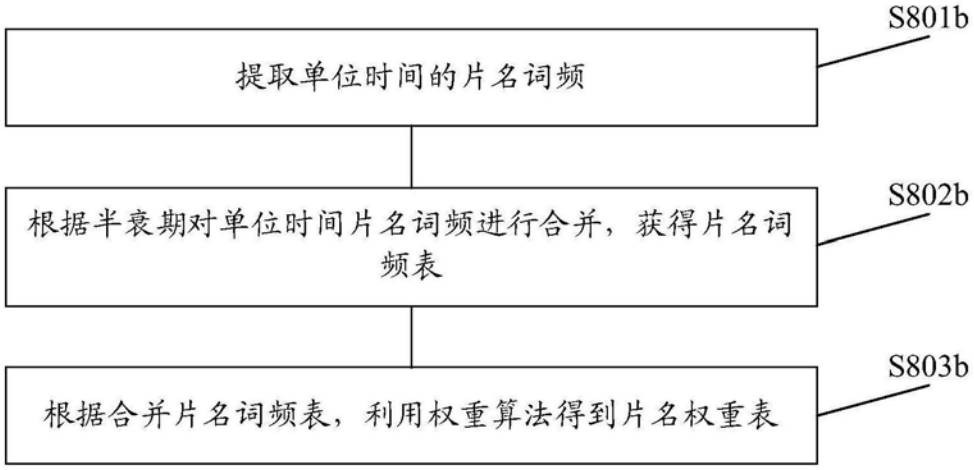

S801b
提取单位时间的片名词频
S802b
根据半衰期对单位时间片名词频进行合并，获得片名词频表
S803b
根据合并片名词频表，利用权重算法得到片名权重表

图8B

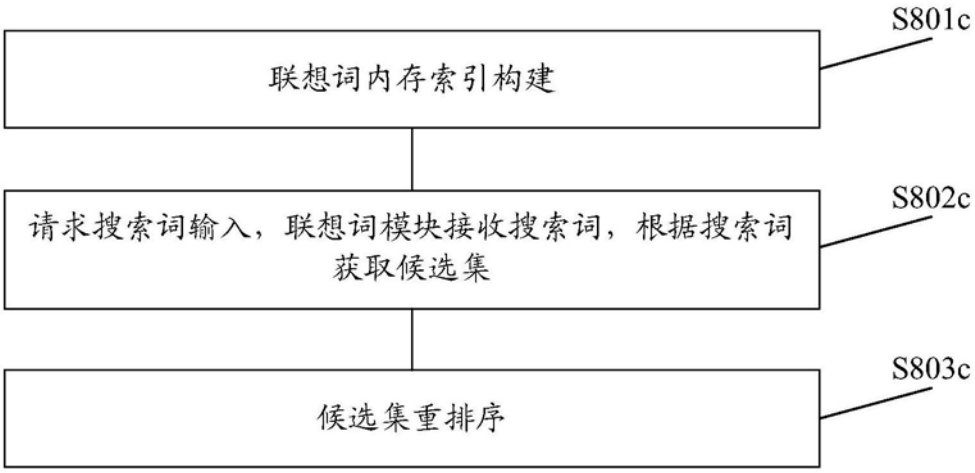

S801c
联想词内存索引构建
S802c
请求搜索词输入，联想词模块接收搜索词，根据搜索词获取候选集
S803c
候选集重排序

图8C

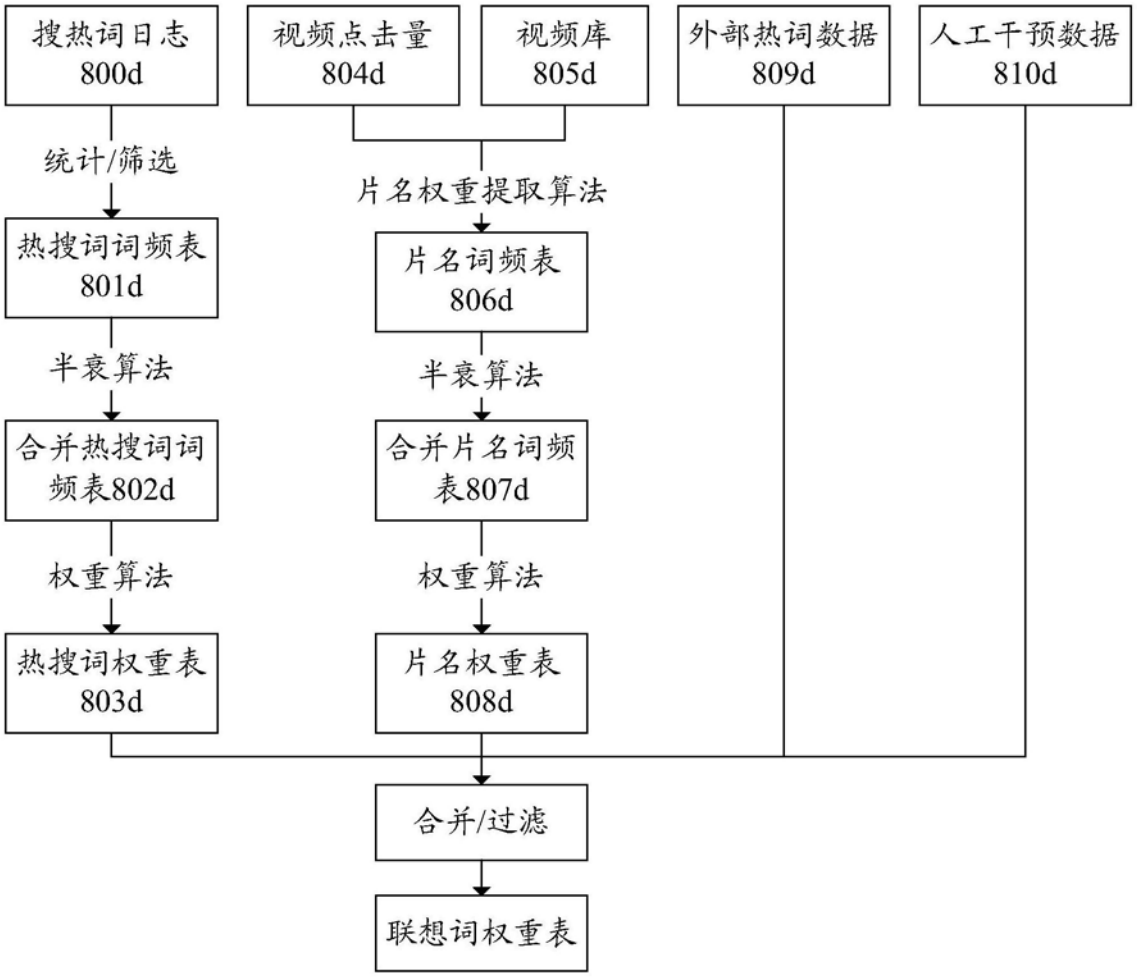

搜热词日志
800d
视频点击量
804d
视频库
805d
外部热词数据
809d
人工干预数据
810d
统计/筛选
片名权重提取算法
热搜词词频表
801d
片名词频表
806d
半衰算法
半衰算法
合并热搜词词频表802d
合并片名词频表807d
权重算法
权重算法
热搜词权重表
803d
片名权重表
808d
合并/过滤
联想词权重表

图8D

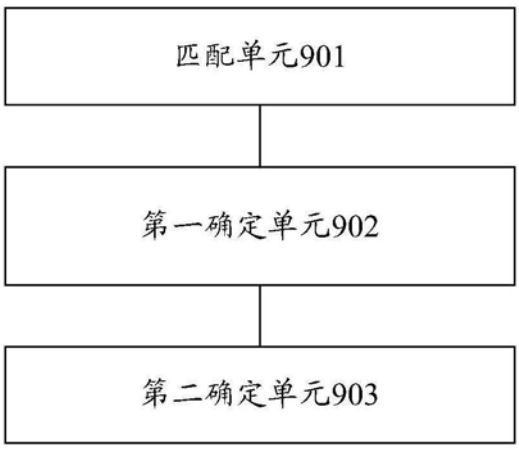

匹配单元901
第一确定单元902
第二确定单元903

图9

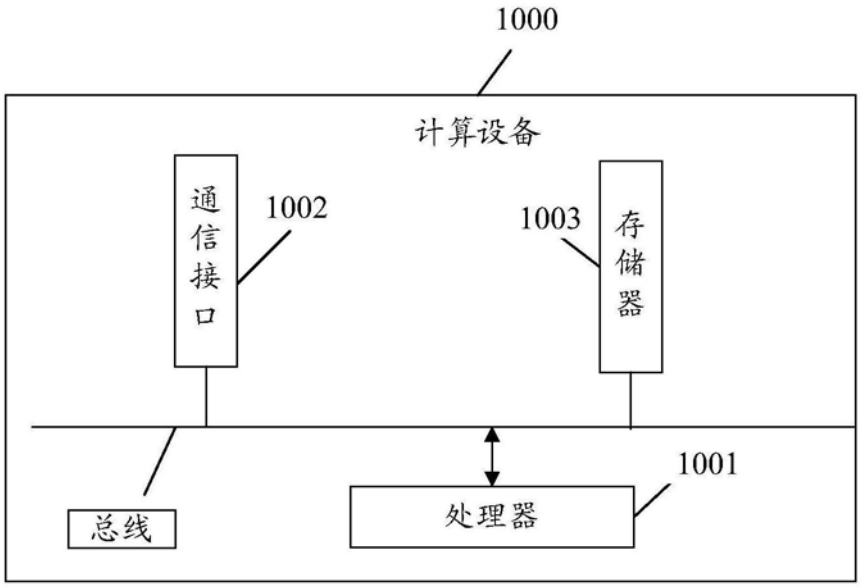


图10