[19] 中华人民共和国国家知识产权局

[51] Int. Cl.
G06F 17/30 (2006.01)
G06F 17/27 (2006.01)

[12] 发明专利申请公布说明书

[21] 申请号 200910001555.7

[43] 公开日 2009 年 6 月 24 日

[11] 公开号 CN 101464897A

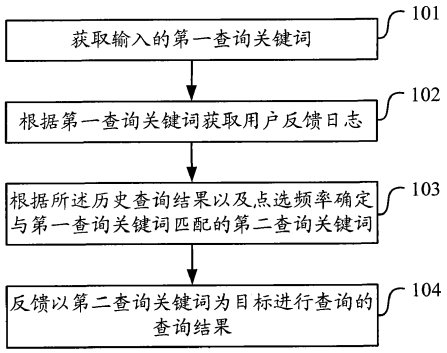
[22] 申请日 2009.1.12
[21] 申请号 200910001555.7
[71] 申请人 阿里巴巴集团控股有限公司
地址 英属开曼群岛大开曼岛
[72] 发明人 谢宇恒 欧文武
[74] 专利代理机构 北京同达信恒知识产权代理有限公司
代理人 魏 杉
权利要求书 8 页 说明书 25 页 附图 3 页

[54] 发明名称

一种词匹配及信息查询方法及装置

[57] 摘要

本申请公开了一种词匹配及信息查询方法及装置，包括：获取输入的第一查询关键词；根据第一查询关键词获取用户反馈日志，所述用户反馈日志包括历次以所述第一查询关键词为目标进行查询的查询结果，以及历次用户对查询结果的点选频率；根据所述查询结果以及点选频率确定与第一查询关键词匹配的第二查询关键词；反馈以第二查询关键词为目标进行查询的查询结果。 由于在本申请实施中采用了用户反馈日志作为发现用户查询信息潜在词义的基础，因此在拥有大量的数据情况下，能够利用以往的用户反馈信息准确的确定出查询信息的潜在词义，从而提高了信息查询的准确性。



1、一种计算机执行的词匹配方法，其特征在于，包括如下步骤：

获取待匹配词；

根据待匹配词获取用户反馈日志；

根据所述用户反馈日志确定与待匹配词匹配的词。

2、如权利要求 1 所述的方法，其特征在于，所述用户反馈日志包括以所述待匹配词为目标进行查询的历史查询结果，以及用户对历史查询结果的点选频率。

3、如权利要求 2 所述的方法，其特征在于，根据所述用户反馈日志中的历史查询结果以及点选频率确定与待匹配词匹配的词。

4、如权利要求 2 或 3 所述的方法，其特征在于，所述点选频率包括：对历史查询结果的点选频率和/或对历史查询结果的内容的点选频率。

5、如权利要求 4 所述的方法，其特征在于，所述根据对历史查询结果的内容的点选频率确定与待匹配词匹配的词，包括：

获取待匹配词的历史查询结果的内容；

对历史查询结果的内容进行分词处理获得分词后的词；

根据分词后的词的点选频率确定与待匹配词匹配的词。

6、如权利要求 5 所述的方法，其特征在于，所述分词后的词包括下述方式的词或者其组合：

分词后与待匹配词相邻的词；

分词后包含待匹配词的词；

分词后包括待匹配词组成部分的词。

7、如权利要求 1 至 6 任一所述的方法，其特征在于， 在根据所述历史查询结果以及点选频率确定与待匹配词匹配的词时，所述点选频率大于设定阈值。

8、如权利要求 1 至 7 任一所述的方法，其特征在于，所述获取待匹配词，

包括:

获取用户输入的信息内容;

对所述信息内容进行分词处理后获得分词后的词，和/或，将所述信息内容分解为字;

将分词后的词和/或字作为待匹配词。

9、如权利要求 1 至 8 任一所述的方法，其特征在于，所述点选频率包括历史查询结果的点击频率、历史查询结果的曝光频率、对历史查询结果的阅读时间、历史查询结果的重要度其中之一或者其组合。

10、如权利要求 1 至 9 任一所述的方法，其特征在于，进一步包括:

在用户输入待匹配词时，获取该用户的用户特征;

所述获取用户反馈日志时，根据该用户特征获取用户反馈日志。

11、如权利要求 1 至 9 任一所述的方法，其特征在于，进一步包括:

在用户输入待匹配词时，获取该用户的用户特征;

所述获取用户反馈日志时，获取用户反馈日志中包括以所述待匹配词为目标进行查询的历史查询结果，以及用户对历史查询结果的点选频率，所述历史查询结果包括所述用户特征。

12、如权利要求 1 至 9 任一所述的方法，其特征在于，进一步包括:

在用户输入待匹配词时，获取该用户的用户特征;

所述根据所述用户反馈日志确定与待匹配词匹配的词时，根据所述用户特征确定与待匹配词匹配的词。

13、一种词匹配装置，其特征在于，包括:

待匹配词获取模块，用于获取待匹配词;

用户反馈日志获取模块，用于根据待匹配词获取用户反馈日志;

匹配模块，用于根据所述用户反馈日志以及点选频率确定与待匹配词匹配的词。

14、如权利要求 13 所述的装置，其特征在于，所述用户反馈日志获取模

块进一步用于获取包括以所述待匹配词为目标进行查询的历史查询结果，以及用户对历史查询结果的点选频率的用户反馈日志。

15、如权利要求 14 所述的装置，其特征在于，所述匹配模块进一步用于根据所述用户反馈日志中的历史查询结果以及点选频率确定与待匹配词匹配的词。

16、如权利要求 14 所述的装置，其特征在于，所述用户反馈日志获取模块进一步用于获取包括:对历史查询结果的点选频率和/或对历史查询结果的内容的点选频率作为所述点选频率。

17、如权利要求 15 所述的装置，其特征在于，所述匹配模块包括:

内容获取单元，用于获取待匹配词的历史查询结果的内容;

分词单元，用于对历史查询结果的内容进行分词处理获得分词后的词;

匹配单元，用于根据分词后的词的点选频率确定与待匹配词匹配的词。

18、如权利要求 17 所述的装置，其特征在于，所述分词单元进一步用于在分词后获得下述方式的词或者其组合:

分词后与待匹配词相邻的词;

分词后包含待匹配词的词;

分词后包括待匹配词组成部分的词。

19、如权利要求 14 至 18 所述的装置，其特征在于，所述匹配模块进一步用于在根据所述历史查询结果以及点选频率确定与待匹配词匹配的词时，所述点选频率大于设定阈值。

20、如权利要求 13 至 19 任一所述的装置，其特征在于，所述待匹配词获取模块包括:

信息内容获取单元，用于获取用户输入的信息内容;

分词/分解单元，用于对所述信息内容进行分词处理后获得分词后的词，和/或，将所述信息内容分解为字;

待匹配词确定单元，用于将分词后的词和/或字作为待匹配词。

21、如权利要求 13 至 20 任一所述的装置，其特征在于，所述用户反馈日志获取模块进一步用于获取包括历史查询结果的点击频率、历史查询结果的曝光频率、对历史查询结果的阅读时间、历史查询结果的重要度其中之一或者其组合的参数作为点选频率。

22、如权利要求 13 至 21 任一所述的装置，其特征在于，所述待匹配词获取模块进一步用于在用户输入待匹配词时，获取该用户的用户特征;

所述用户反馈日志获取模块进一步用于根据用户特征获取用户反馈日志。

23、如权利要求 13 至 21 任一所述的装置，其特征在于，所述待匹配词获取模块进一步用于在用户输入待匹配词时，获取该用户的用户特征;

所述用户反馈日志获取模块进一步用于在获取用户反馈日志时，获取用户反馈日志中包括以所述待匹配词为目标进行查询的历史查询结果，以及用户对历史查询结果的点选频率，所述历史查询结果包括所述用户特征。

24、如权利要求 13 至 21 任一所述的装置，其特征在于，所述待匹配词获取模块进一步用于在用户输入待匹配词时，获取该用户的用户特征;

所述匹配模块进一步用于在根据所述用户反馈日志确定与待匹配词匹配的词时，根据所述用户特征确定与待匹配词匹配的词。

25、一种信息查询方法，其特征在于，包括如下步骤:

获取输入的第一查询关键词;

根据第一查询关键词获取用户反馈日志;

根据所述用户发馈日志确定与第一查询关键词匹配的第二查询关键词;

反馈以第二查询关键词为目标进行查询的查询结果。

26、如权利要求 25 所述的方法，其特征在于，所述用户反馈日志包括以所述第一查询关键词为目标进行查询的历史查询结果，以及用户对历史查询结果的点选频率。

27、如权利要求 26 所述的方法，其特征在于，根据所述用户反馈日志中的历史查询结果以及点选频率确定与第一查询关键词匹配的第二查询关键词。

28、如权利要求 26 或 27 所述的方法，其特征在于，所述点选频率包括：对历史查询结果的点选频率和/或对历史查询结果的内容的点选频率。

29、如权利要求 28 所述的方法，其特征在于，所述根据对历史查询结果的内容的点选频率确定与第一查询关键词匹配的第二查询关键词，包括：

获取第一查询关键词的历史查询结果的内容；

对历史查询结果的内容进行分词处理获得分词后的词；

根据分词后的词的点选频率确定与第一查询关键词匹配的第二查询关键词。

30、如权利要求 29 所述的方法，其特征在于，所述分词后的词是指下述方式的词或者其组合：

分词后与第一查询关键词相邻的词；

分词后包含第一查询关键词的词；

分词后包括第一查询关键词组成部分的词。

31、如权利要求 25 至 30 任一所述的方法，其特征在于， 在根据所述历史查询结果以及点选频率确定与第一查询关键词匹配的第二查询关键词时，所述点选频率大于设定阈值。

32、如权利要求 25 至 31 任一所述的方法，其特征在于，所述获取输入的第一查询关键词，包括：

获取用户输入的信息内容；

对所述信息内容进行分词处理后获得分词后的词，和/或，将所述信息内容分解为字；

将分词后的词和/或字作为第一查询关键词。

33、如权利要求 25 至 32 任一所述的方法，其特征在于，所述点选频率包括历史查询结果的点击频率、历史查询结果的曝光频率、对历史查询结果的阅读时间、历史查询结果的重要度其中之一或者其组合。

34、如权利要求 25 至 33 任一所述的方法，其特征在于，进一步包括：

在用户输入第一查询关键词词时，获取该用户的用户特征;

所述获取用户反馈日志时，根据该用户的用户特征获取用户反馈日志。

35、如权利要求 25 至 33 任一所述的方法，其特征在于，进一步包括:

在用户输入第一查询关键词词时，获取该用户的用户特征;

所述获取用户反馈日志时，获取用户反馈日志中包括以所述第一查询关键词为目标进行查询的历史查询结果，以及用户对历史查询结果的点选频率，所述历史查询结果包括所述用户特征。

36、如权利要求 25 至 33 任一所述的方法，其特征在于，进一步包括:

在用户输入第一查询关键词词时，获取该用户的用户特征;

所述根据所述用户反馈日志确定第二查询关键词时，根据所述用户特征确定第二查询关键词。

37、一种信息查询装置，其特征在于，包括:

第一查询关键词获取模块，用于获取输入的第一查询关键词;

用户反馈日志获取模块，用于根据第一查询关键词获取用户反馈日志;

匹配模块，用于根据所述用户反馈日志确定与第一查询关键词匹配的第二查询关键词;

查询结果反馈模块，用于反馈以第二查询关键词为目标进行查询的查询结果。

38、如权利要求 37 所述的装置，其特征在于，用户反馈日志获取模块进一步用于获取包括以所述第一查询关键词为目标进行查询的历史查询结果，以及用户对历史查询结果的点选频率的用户反馈日志。

39、如权利要求 38 所述的装置，其特征在于，所述匹配模块进一步用于根据所述用户反馈日志中的历史查询结果以及点选频率确定与第一查询关键词匹配的第二查询关键词。

40、如权利要求 38 所述的装置，其特征在于，所述用户反馈日志获取模块进一步用于获取包括:对历史查询结果的点选频率和/或对历史查询结果的内

容的点选频率作为所述点选频率。

41、如权利要求 40 所述的装置，其特征在于，所述匹配模块包括:

内容获取单元，用于获取第一查询关键词的历史查询结果的内容;

分词单元，用于对历史查询结果的内容进行分词处理获得分词后的词;

匹配单元，用于根据分词后的词的点选频率确定与第一查询关键词匹配的第二查询关键词。

42、如权利要求 41 所述的装置，其特征在于，所述分词单元进一步用于在分词后获得下述方式的词或者其组合:

分词后与第一查询关键词相邻的词;

分词后包含第一查询关键词的词;

分词后包括第一查询关键词组成部分的词。

43、如权利要求 37 至 40 任一所述的装置，其特征在于， 所述匹配模块进一步用于在根据所述历史查询结果以及点选频率确定与第一查询关键词匹配的第二查询关键词时，所述点选频率大于设定阈值。

44、如权利要求 37 至 43 任一所述的装置，其特征在于，所述第一查询关键词获取模块包括:

信息内容获取单元，用于获取用户输入的信息内容;

分词/分解单元，用于对所述信息内容进行分词处理后获得分词后的词，和/或，将所述信息内容分解为字;

第一查询关键词确定单元，用于将分词后的词和/或字作为第一查询关键词。

45、如权利要求 37 至 44 任一所述的装置，其特征在于，所述用户反馈日志获取模块进一步用于获取包括历史查询结果的点击频率、历史查询结果的曝光频率、对历史查询结果的阅读时间、历史查询结果的重要度其中之一或者其组合的参数作为点选频率。

46、如权利要求 37 至 45 任一所述的装置，其特征在于，所述第一查询关

键词获取模块进一步用于在用户输入第一查询关键词词时，获取该用户特征；

所述用户反馈日志获取模块进一步用于根据用户特征获取用户反馈日志。

47、如权利要求 37 至 45 任一所述的装置，其特征在于，所述第一查询关键词获取模块进一步用于在用户输入第一查询关键词词时，获取该用户特征；

所述用户反馈日志获取模块进一步用于在获取用户反馈日志时，获取用户反馈日志中包括以所述待匹配词为目标进行查询的历史查询结果，以及用户对历史查询结果的点选频率，所述历史查询结果包括所述用户特征。

48、如权利要求 37 至 45 任一所述的装置，其特征在于，所述第一查询关键词获取模块进一步用于在用户输入第一查询关键词词时，获取该用户的用户特征；

所述匹配模块进一步用于在根据所述用户反馈日志确定第二查询关键词时，根据所述用户特征确定第二查询关键词。

一种词匹配及信息查询方法及装置

技术领域

本申请涉及数据处理技术，特别涉及一种词匹配及信息查询方法及装置。

背景技术

潜在词义通常是指一个词（包括短语）潜在的意义，通常可以通过另外一个或多个词（包括短语）来表达，比如通常所称的“冰箱”其一般情况下潜在的词义是指“电冰箱”，而“棉拖”其一般情况下潜在的词义是指“全棉拖鞋”等。

自动发现潜在词义是自然语言处理的一个基本问题，它的解决可以提高文档理解、机器翻译和搜索引擎的效果和性能。

分词技术是自然语言处理中常用的技术，分词是将一个输入字符串分成若干个词或短语，比如“曾经有一段诚挚的感情摆在我的面前”，经过分词处理后，通常情况下得到的分词结果为“曾经|有|一段|诚挚|的|感情|摆在|我|的|面前”。

用户反馈日志记录了查询词对应的查询结果（文档或网页 ID 等）和查询结果点击频率、曝光频率等。点击频率、曝光频率等信息反应了用户对该查询结果的认同程度，一般意义上符合用户需求的文档点击率比不符合用户意图的点击率要高，比如查“西药”，结果“批发西药”和“江西药厂”的单字的匹配程度是一样的，但是通常第一个结果的点击率会比第二个结果要高。

通过分析用户反馈日志可以发现与查询词字符匹配程度较高，同时表达方式不同的词，比如搜索“冰箱”一词时，会发现很多带“电冰箱”的结果，比如“双开门电冰箱”、“发明了冰箱”、“电冰箱厂家”、“销售电冰箱”、“存冰箱子”等，收集点击率相对较高的结果，并且对出现冰箱的句子分词，统计每个分词的频率，如果某个或多个分词结果大于设定的阈值，则做下面处理：查询

词包含在一个高频分词结果中，比如“冰箱”包含在“电冰箱”中，则认为“电冰箱”是“冰箱”的潜在词义；查询词包含在相邻的两个高频分词中，例如：查询词“玻璃瓶”包含在“玻璃”和“瓶子”这两个高频分词中，这也通常被认为“玻璃瓶子”是“玻璃瓶”的潜在词义。

目前潜在语意的自动发现上已经有过不少的研究，大多是通过词语的共现或链接关系来发现近义词。例如陆勇、侯汉清在文章“基于 PageRank 算法的汉语同义词自动识别”中介绍了一种同义词的自动发现方法，该文章将词汇之间解释与被解释的关系看成是一种链接，把 PageRank 值看成是体现词汇之间语义相似性的衡量指标，然后根据语义相似度的大小识别同义词。这个方法的缺点是：基于人工标注的语料，挖掘得到的词条数量会比较有限。如果改成基于互联网网页之间的链接关系，这种链接关系有时又很不可靠，同义词自动发现的效果很难得到保障。

搜索引擎的索引方式包括单字搜索、分词索引和混合索引。单字索引需要计算文档内单字之间的距离，效率不高，并且精确率低，比如搜索“农药”时，单字索引无法区分“神农药厂”和“神农农药厂”的区别；而分词搜索精确率高，速度快，但是分词索引召回率有时比较低，比如搜“冰箱”时，分词索引方法只能找到“冰箱”的结果，而找不到“电冰箱”的结果；单字索引和分词索引结合的混合索引方法通常是先根据分词索引查询，然后再根据单字索引查询，比如查“玻璃瓶”时，先按分词索引找到“玻璃瓶”的结果，再按单字索引找出其他结果，这种弥补了两种方法的缺点，但是“玻璃瓶子”是根据单字索引的方式找到的，搜索引擎不能区分“玻璃瓶子”和“生产玻璃瓶颈在于”的差异，影响搜索的准确性；

前面的方法缺少足够的数据量，或者缺少用户的反馈，抽取出来的潜在语意太少或很有可能是错误的。

如陆勇、侯汉清提到的词义自动发现方法主要是通过已有的词典数据作为抽取来源，样本量在几千条左右。如果是以互联网网页等大数据量为基础的挖

掘方法，又缺乏准确性。

因此现有技术的不足在于：当面临如互联网等存在着大数据量的情况时，尚没有一种好的查询方案能够准确的预知用户真正所需查询的内容，也因此不能向用户反馈用户真正所需的查询结果。

发明内容

本申请提供了一种词匹配方法及装置，用以提供一种在存在海量数据的情况下，准确判断词与词之间的内在联系，并将其匹配的方案。

本申请实施例提供了一种词匹配方法，包括如下步骤：

获取待匹配词；

根据待匹配词获取用户反馈日志；

根据所述用户反馈日志确定与待匹配词匹配的词。

较佳地，所述用户反馈日志包括以所述待匹配词为目标进行查询的历史查询结果，以及用户对历史查询结果的点选频率。

较佳地，根据所述用户反馈日志中的历史查询结果以及点选频率确定与待匹配词匹配的词。

较佳地，所述点选频率包括：对历史查询结果的点选频率和/或对历史查询结果的内容的点选频率。

较佳地，所述根据对历史查询结果的内容的点选频率确定与待匹配词匹配的词，包括：

获取待匹配词的历史查询结果的内容；

对历史查询结果的内容进行分词处理获得分词后的词；

根据分词后的词的点选频率确定与待匹配词匹配的词。

较佳地，所述分词后的词包括下述方式的词或者其组合：

分词后与待匹配词相邻的词；

分词后包含待匹配词的词；

分词后包括待匹配词组成部分的词。

较佳地， 在根据所述查询结果以及点选频率确定与待匹配词匹配的词时，所述点选频率大于设定阈值。

较佳地，所述获取待匹配词，包括:

获取用户输入的信息内容;

对所述信息内容进行分词处理后获得分词后的词，和/或，将所述信息内容分解为字;

将分词后的词和/或字作为待匹配词。

较佳地，所述点选频率包括历史查询结果的点击频率、历史查询结果的曝光频率、对历史查询结果的阅读时间、历史查询结果的重要度其中之一或者其组合。

较佳地，进一步包括:

在用户输入待匹配词时，获取该用户的用户特征;

所述获取用户反馈日志时，根据该用户的用户特征获取用户反馈日志。

较佳地，进一步包括:

在用户输入待匹配词时，获取该用户的用户特征;

所述获取用户反馈日志时，获取用户反馈日志中包括以所述待匹配词为目标进行查询的历史查询结果，以及用户对历史查询结果的点选频率，所述历史查询结果包括所述用户特征。

较佳地，进一步包括:

在用户输入待匹配词时，获取该用户的用户特征;

所述根据所述用户反馈日志确定与待匹配词匹配的词时，根据所述用户特征确定与待匹配词匹配的词。

本申请实施例还提供了一种词匹配装置，包括:

待匹配词获取模块，用于获取待匹配词;

用户反馈日志获取模块，用于根据待匹配词获取用户反馈日志;

匹配模块，用于根据所述用户反馈日志以及点选频率确定与待匹配词匹配的词。

较佳地，所述用户反馈日志获取模块进一步用于获取包括以所述待匹配词为目标进行查询的历史查询结果，以及用户对历史查询结果的点选频率的用户反馈日志。

较佳地，匹配模块进一步用于根据所述用户反馈日志中的历史查询结果以及点选频率确定与待匹配词匹配的词。

较佳地，所述用户反馈日志获取模块进一步用于获取包括：对历史查询结果的点选频率和/或对历史查询结果的内容的点选频率作为所述点选频率。

较佳地，所述匹配模块包括：

内容获取单元，用于获取待匹配词的历史查询结果的内容；

分词单元，用于对历史查询结果的内容进行分词处理获得分词后的词；

匹配单元，用于根据分词后的词的点选频率确定与待匹配词匹配的词。

较佳地，所述分词单元进一步用于在分词后获得下述方式的词或者其组合：

分词后与待匹配词相邻的词；

分词后包含待匹配词的词；

分词后包括待匹配词组成部分的词。

较佳地，所述匹配模块进一步用于在根据所述历史查询结果以及点选频率确定与待匹配词匹配的词时，所述点选频率大于设定阈值。

较佳地，所述待匹配词获取模块包括：

信息内容获取单元，用于获取用户输入的信息内容；

分词/分解单元，用于对所述信息内容进行分词处理后获得分词后的词，和/或，将所述信息内容分解为字；

待匹配词确定单元，用于将分词后的词和/或字作为待匹配词。

较佳地，所述用户反馈日志获取模块进一步用于获取包括历史查询结果的

点击频率、历史查询结果的曝光频率、对历史查询结果的阅读时间、历史查询结果的重要度其中之一或者其组合的参数作为点选频率。

较佳地，所述待匹配词获取模块进一步用于在用户输入待匹配词时，获取该用户的用户特征；

所述用户反馈日志获取模块进一步用于根据用户特征获取用户反馈日志。

较佳地，所述待匹配词获取模块进一步用于在用户输入待匹配词时，获取该用户的用户特征；

所述用户反馈日志获取模块进一步用于在获取用户反馈日志时，获取用户反馈日志中包括以所述待匹配词为目标进行查询的历史查询结果，以及用户对历史查询结果的点选频率，所述历史查询结果包括所述用户特征。

较佳地，所述待匹配词获取模块进一步用于在用户输入待匹配词时，获取该用户的用户特征；

所述匹配模块进一步用于在根据所述用户反馈日志确定与待匹配词匹配的词时，根据所述用户特征确定与待匹配词匹配的词。

基于同一构思，本申请提供一种信息查询方法及装置，用以提供一种在存在海量数据的情况下，利用前述的词与词之间匹配关系，准确判断用户查询信息的真实需要，并反馈用户真正所需的查询结果。

本申请实施例中提供了一种信息查询方法，包括如下步骤：

获取输入的第一查询关键词；

根据第一查询关键词获取用户反馈日志；

根据所述用户发馈日志确定与第一查询关键词匹配的第二查询关键词；

反馈以第二查询关键词为目标进行查询的查询结果。

较佳地，所述用户反馈日志包括以所述第一查询关键词为目标进行查询的历史查询结果，以及用户对历史查询结果的点选频率。

较佳地，根据所述用户反馈日志中的历史查询结果以及点选频率确定与第一查询关键词匹配的第二查询关键词。

较佳地，所述点选频率包括：对历史查询结果的点选频率和/或对历史查询结果的内容的点选频率。

较佳地，所述根据对历史查询结果的内容的点选频率确定与第一查询关键词匹配的第二查询关键词，包括：

获取第一关键词的历史查询结果的内容；

对历史查询结果的内容进行分词处理获得分词后的词；

根据分词后的词的点选频率确定与第一查询关键词匹配的第二查询关键词。

较佳地，所述分词后的词是指下述方式的词或者其组合：

分词后与第一查询关键词相邻的词；

分词后包含第一查询关键词的词；

分词后包括第一查询关键词组成部分的词。

较佳地，在根据所述历史查询结果以及点选频率确定与第一查询关键词匹配的第二查询关键词时，所述点选频率大于设定阈值。

较佳地，所述获取输入的第一查询关键词，包括：

获取用户输入的信息内容；

对所述信息内容进行分词处理后获得分词后的词，和/或，将所述信息内容分解为字；

将分词后的词和/或字作为第一查询关键词。

较佳地，所述点选频率包括历史查询结果的点击频率、历史查询结果的曝光频率、对历史查询结果的阅读时间、历史查询结果的重要度其中之一或者其组合。

较佳地，进一步包括：

在用户输入第一查询关键词时，获取该用户的用户特征；

所述获取用户反馈日志时，根据该用户的用户特征获取用户反馈日志。

较佳地，进一步包括：

在用户输入第一查询关键词时，获取该用户的用户特征;

所述获取用户反馈日志时，获取用户反馈日志中包括以所述第一查询关键词为目标进行查询的历史查询结果，以及用户对历史查询结果的点选频率，所述历史查询结果包括所述用户特征。

较佳地，进一步包括:

在用户输入第一查询关键词时，获取该用户的用户特征;

所述根据所述用户反馈日志确定第二查询关键词时，根据所述用户特征确定第二查询关键词。

本申请实施例中还提供了一种信息查询装置，包括:

第一查询关键词获取模块，用于获取输入的第一查询关键词;

用户反馈日志获取模块，用于根据第一查询关键词获取用户反馈日志;

匹配模块，用于根据所述用户反馈日志确定与第一查询关键词匹配的第二查询关键词;

查询结果反馈模块，用于反馈以第二查询关键词为目标进行查询的查询结果。

较佳地，用户反馈日志获取模块进一步用于获取包括以所述第一查询关键词为目标进行查询的历史查询结果，以及用户对历史查询结果的点选频率的用户反馈日志。

较佳地，匹配模块进一步用于根据所述用户反馈日志中的历史查询结果以及点选频率确定与第一查询关键词匹配的第二查询关键词。

较佳地，所述用户反馈日志获取模块进一步用于获取包括: 对历史查询结果的点选频率和/或对历史查询结果的内容的点选频率作为所述点选频率。

较佳地，所述匹配模块包括:

内容获取单元，用于获取第一关键词的历史查询结果的内容;

分词单元，用于对历史查询结果的内容进行分词处理获得分词后的词;

匹配单元，用于根据分词后的词的点选频率确定与第一查询关键词匹配的

第二查询关键词。

较佳地，所述分词单元进一步用于在分词后获得下述方式的词或者其组合:

分词后与第一查询关键词相邻的词;

分词后包含第一查询关键词的词;

分词后包括第一查询关键词组成部分的词。

较佳地，所述匹配模块进一步用于在根据所述历史查询结果以及点选频率确定与第一查询关键词匹配的第二查询关键词时，所述点选频率大于设定阈值。

较佳地，所述第一查询关键词获取模块包括:

信息内容获取单元，用于获取用户输入的信息内容;

分词/分解单元，用于对所述信息内容进行分词处理后获得分词后的词，和/或，将所述信息内容分解为字;

第一查询关键词确定单元，用于将分词后的词和/或字作为第一查询关键词。

较佳地，所述用户反馈日志获取模块进一步用于获取包括历史查询结果的点击频率、历史查询结果的曝光频率、对历史查询结果的阅读时间、历史查询结果的重要度其中之一或者其组合的参数作为点选频率。

较佳地，所述第一查询关键词获取模块进一步用于在用户输入第一查询关键词时，获取该用户的用户特征;

所述用户反馈日志获取模块进一步用于根据用户特征获取用户反馈日志。

较佳地，所述第一查询关键词获取模块进一步用于在用户输入第一查询关键词时，获取该用户的用户特征;

所述用户反馈日志获取模块进一步用于在获取用户反馈日志时，获取用户反馈日志中包括以所述待匹配词为目标进行查询的历史查询结果，以及用户对历史查询结果的点选频率，所述历史查询结果包括所述用户特征。

较佳地，所述第一查询关键词获取模块进一步用于在用户输入第一查询关键词时，获取该用户的用户特征；

所述匹配模块进一步用于在根据所述用户反馈日志确定第二查询关键词时，根据所述用户特征确定第二查询关键词。

本申请有益效果如下：

本申请实施中，在获取输入的第一查询关键词后，就去获取第一查询关键词的用户反馈日志，而用户反馈日志中包括了以所述第一查询关键词为目标进行查询的历史查询结果，以及用户对历史查询结果的点选频率；然后根据历史查询结果以及点选频率来确定与第一查询关键词匹配的第二查询关键词；最后反馈的是以匹配后的第二查询关键词为目标进行查询的查询结果。由于在此过程中采用了用户反馈日志作为发现用户查询信息潜在词义的基础，因此在拥有大量的数据情况下，能够利用以往的用户反馈信息准确的确定出查询信息的潜在词义，从而提高了信息查询的准确性。

附图说明

图 1 为本申请实施例中信息查询方法实施流程示意图；

图 2 为本申请实施例中信息查询装置结构示意图；

图 3 为本申请实施例中匹配模块结构示意图；

图 4 为本申请实施例中第一查询关键词获取模块结构示意图；

图 5 为本申请实施例中词匹配方法实施流程示意图；

图 6 为本申请实施例中词匹配装置结构示意图。

具体实施方式

下面结合附图对本申请的具体实施方式进行说明。

图 1 为信息查询方法实施流程示意图，如图所示，可以包括如下步骤：

步骤 101、获取输入的第一查询关键词；

步骤 102、根据第一查询关键词获取用户反馈日志;

用户反馈日志包括以所述第一查询关键词为目标进行查询的历史查询结果，以及用户对历史查询结果的点选频率;

步骤 103、根据所述历史查询结果以及点选频率确定与第一查询关键词匹配的第二查询关键词;

步骤 104、反馈以第二查询关键词为目标进行查询的查询结果。

下面对各步骤的具体实施进行说明。

步骤 101 中，对于第一查询关键词，可以是:

获取用户输入的信息内容;

对所述信息内容进行分词处理后获得分词后的词，和/或，将所述信息内容分解为字;

将分词后的词和/或字作为第一查询关键词。

可以看出，本申请实施过程中用于查询的关键词可以是词也可以是字，当是字时，可以视为通常所指的单字查询，通过对用户输入的需要查询的信息内容来说，以各种查询单位，如字或词来查询，或者结合起来查询显然可以使查询结果的精度更高、更准确。

步骤 102 中，用户反馈日志通常是指搜索引擎用来收集用户输入的关键词和历史查询结果（通常是网页文档 ID 等）和历史查询结果的点击频率、曝光率等。

实施中，用户反馈日志可以包括的是历次以第一查询关键词为目标进行查询的历史查询结果，以及历次用户对历史查询结果的点选频率，用户反馈日志作为建立潜在词义的样本，可以采用历次的记录，但是，用户反馈日志的目的在于通过以往的记录来确定词与词之间的内在关系，从而建立潜在词义，只要能实现该目的，显然也可以选取部分历史查询结果，或者是随机选取等等方式来采集确定潜在词义的样本。同样道理，用户反馈日志在选取时，并不是以用户为对象来进行选取，而是以历史上进行查询的词为目标来进行选取，例如需

要获取第一查询关键词为“西药”的用户反馈日志时，获取的是历史上用“西药”为查询词的所有或者部分用户的用户反馈日志。

潜在词义的自动发现特指找出一个词（短语）和另外词义相关或相近的一个词（短语）或多个词（短语）。本申请实施例的本质在于通过利用用户参与的用户反馈日志以便能够非常可靠的自动发现查询词和历史查询结果之间体现用户意图的潜在词义关系，并利用该关系来提高搜索引擎的准确率和智能。因此，用户反馈日志中可以包括历次以所述第一查询关键词为目标进行查询的历史查询结果，以及历次用户对历史查询结果的点选频率。并在步骤 103 中基于历史查询结果以及点选频率来寻找第一查询关键词的潜在词义。即，在步骤 102 中获取的是用户反馈日志，并利用用户反馈日志来确定第一查询关键词的潜在词义，从而能够通过步骤 103 输出和步骤 101 中第一查询关键词之间存在潜在词义关系的第二查询关键词。

其中，点选频率可以包括：对历史查询结果的点选频率和/或对历史查询结果的内容的点选频率。

下面对步骤 103 的具体实施进行说明。

首先对根据对历史查询结果的内容的点选频率确定与第一查询关键词匹配的第二查询关键词进行说明。

获取第一关键词的历史查询结果的内容；

对历史查询结果的内容进行分词处理获得分词后的词；

根据分词后的词的点选频率确定与第一查询关键词匹配的第二查询关键词。

实施中，分词后的词是指下述方式的词或者其组合：

第一种词：分词后与第一查询关键词相邻的词，为描述方便，实施例中将该种情况下的点选频率相关的统计结果记为 P1；

第二种词：分词后包括第一查询关键词组成部分的词，为描述方便，实施例中将该种情况下的点选频率相关的统计结果记为 P2；

第三种词：分词后包含第一查询关键词的词，为描述方便，实施例中将该种情况下的点选频率相关的统计结果记为 P3。

下面先对步骤 103 的实施原理进行说明。

用户反馈日志是用来记录查询词对应的历史查询结果和历史查询结果的点击率、曝光频率等信息的，如查询结果为网页等；发明人在发明过程中注意到：对于某个查询词点击率越高的网页与查询词越相关。一个词的潜在词义是指和它同义、近义或者部分同义的词，比如“玻璃瓶”和“玻璃瓶子”，又如“双人床”、“单人床”、“弹簧床”等词都潜在“床”的词义，而“机床”等则不潜在“床”的词义。在本申请实施例中定义了三种潜在词义：第一种词是经常成对出现的词，比如“摩托罗拉”和“公司”，“摩托罗拉”和“手机”，这种关系通常是一个词和另外一个词密切相关，即，分词后的有些词与查询词相邻；第二种词是一个词和另外多个并按一定顺序出现的词，比如“玻璃瓶”和“玻璃”“瓶子”，“美女”和“美丽的”“女人”，即分词后其包含了查询词的组成部分；第三种词是一个词是一个词组成部分，比如“虾”和“对虾”，“酒”和“啤酒”，即，分词后的词包含了查询词。这些通过点击率等用户反馈自动发现的潜在词义往往代表了用户输入的搜索关键字的潜在意图，可以用来提高搜索引擎的准确率，比如用户搜索“床”时大部分用户的实际意图是睡觉的床比如“单人床”、“双人床”、“木板床”等，而不是机械设备比如“机床”或“车床”。通过用户点击等反馈就能知道前者有“床”的潜在词义，而后者（机床等）没有。

本申请在具体实施中，首先输入第一查询关键词、历史查询结果（网页，文档 ID 等）和历史查询结果的点击率、曝光率等信息或其中之一，即输入步骤 101 中的第一查询关键词以及步骤 102 获取用户反馈日志的执行结果；然后对第一查询关键词进行分词，如果第一查询关键词包括多个词，则将这条查询词的用户反馈日志中对应的历史查询结果和相关信息添加到这条查询词中相应的每个分词中去，即，使这条查询词在分词后的每个词都有自己的历史查询

结果，这样处理后，用户反馈日志的每个 query（查询）都是一个单独的分词；然后对每个分词后得到词或其中部分分别做上述与 P1、P2、P3 有关的处理，直到所有或部分分词后的词处理完毕，历史查询结果的选取可以根据历史查询结果总的查询次数、点击次数、曝光次数等信息或其中之一确定；对分词后的词对应的历史查询结果分别做处理直到所有历史查询结果处理完毕；从用户反馈日志中的历史查询结果中找出所有与分词后的词完全匹配的字符串（这里完全匹配是指分词后的词是字符串的一个子串），字符串的尺度可以是包含分词后的词的句子长度，或包含分词后的词长度的 M 倍，M 可以是大于 1 的任何数，然后对字符串分词后做上述与 P1、P2、P3 有关的处理，需要说明的是，为便于描述，下述实施例中以文档为查询结果，实施时，同时考虑了对查询结果的点选频率和对查询结果的内容的点选频率，显然，只考虑其中一个同样能实现申请目的。

具体实施中，当在输入第一查询关键词、历史查询结果（网页，文档 ID 等）和历史查询结果的点击率、曝光率等信息或其中之一时，可以设置一个查询词典，提前输入历史查询结果（网页，文档 ID 等）和历史查询结果的点击率、曝光率等信息或其中之一，这样当输入第一查询关键词时，通过查询词典便可以快捷的获得第二查询关键词。也就是将以往的用户反馈日志的内容预先存储用于查询，也可以根据新的用户反馈日志随时对查询词典进行更新；当然也可以在输入第一查询关键字后再调用用户反馈日志。

第一种：分词后与第一查询关键词相邻的词的实施。

如果第一查询关键词是字符串的一个分词，比如第一查询关键词是“美女”，用户反馈日志中的历史查询结果是“中国|古代|美女|西施|名|夷光|，|春秋|战国|时期|出生”（这里“|”表示分词结果），这时将查询词前后的 T 个分词在字符串中出现的次数乘以该文档的点击频率和曝光频率（或其中之一）作为权重的一个系数，记为次数加权（1），加到总的查询结果的统计 P1，P1 中包含了第一查询关键词前后出现的每个词的次数加权（1），例如本例中，如果文档

的权重为 0.5，则 P1 中“古代”和“西施”（这只是 T 等于 1 的情况）对应的结果会相应加 0.5。

第二种：分词后包括第一查询关键词组成部分的词。

如果第一查询关键词包含在字符串相邻的多个分词结果中，比如第一查询关键词是“美女”，用户反馈日志中的历史查询结果是“西施|是|个|美丽的|女人|”（这里“|”表示分词结果），这时将包含第一查询关键词的分词出现次数并乘以该文档的点击频率/曝光频率（或其中之一）作为权重的一个系数，记为次数加权（2），加到总的查询结果的统计 P2，P2 中是包括第一查询关键词的多个分词按照相同顺序出现的次数加权(2)，例如本例中，如果文档的权重为 0.3，则将 P2 中“|美丽的|女人|” 对应的结果加 0.3。

第三种：分词后包含第一查询关键词的词。

如果第一查询关键词是字符串一个分词的字串，比如查询词是“冰箱”，用户反馈日志中的历史查询结果是“电冰箱|空调器|原理|与|维修”（这里“|”表示分词结果），这时将包含第一查询关键词的分词出现次数并乘以该文档的点击频率和曝光频率（或其中之一）作为权重的一个系数，记为次数加权（3），加到总的查询结果的统计 P3，P3 是包括第一查询关键词的分词出现的次数加权（3），例如本例中，如果文档的权重为 0.8，则将 P3 中“电冰箱”对应的结果加 0.8。

不断重复直到对于单个分词后的词所有的用户反馈日志中的历史查询结果全部处理完毕；按照 P1 中分词出现的次数加权和，取次数加权和大于设定的第一阈值的分词，将这些分词作为该查询词的第一种潜在词义关系，同样，按照 P2，P3 中分词出现的次数加权和，并取次数加权和大于设定的第二、第三阈值的分词，将这些分词作为该词的第二种潜在词义和第三种潜在词义关系。

本领域技术人员容易知道，实施中可以选用三种选择潜在词义中的一种，也可以任意两种组合或三种组合；

同样，实施中，第一、第二、第三阈值可以是固定阈值，也可以根据查询词总体查询结果动态设定，比如将所有包含了匹配字符串的文档权重求和，然后再乘以一个系数，该系数便可根据查询结果动态设定；阈值设置的目的在于有选择的确定一部分查询词的潜在词义的词，并非将所有的词都无条件反馈。

具体实施中，在根据所述历史查询结果以及点选频率确定与第一查询关键词匹配的第二查询关键词时，可以要求点选频率大于设定阈值，其中，点选频率可以是用户对历史查询结果的点选频率，也可以是用户对历史查询结果的内容的点选频率。其目的在于将文档或者其内容的点击频率和曝光频率（或其中之一）作为权重的一个系数，该系数可以与点击率和曝光率二者之一或两者的组合，系数大小和点击和曝光频率可以是线性或非线性的关系，比如（不限于）两者频率高于某一设定阈值的全部为 1，其他为 0；或者点击率和曝光率最高的为 1，其他的除以最大值归一化到[0,1]。点选频率的选取目的在于通过它来发现潜在词义，因而可以通过设定阈值来过滤一些点选频率较低的信息，从而提高发现潜在词义的速度，同时也可以避免一些信息的干扰。

实施中，点选频率包括历史查询结果的点击频率、历史查询结果的曝光频率、对历史查询结果的阅读时间、历史查询结果的重要度其中之一或者其组合。本领域技术人员容易理解，该文档的点击频率和曝光频率（或其中之一）作为权重的一个系数，系数也可以是文档的其他信息，比如阅读时间，重要程度等或其中之一或与点击率曝光率的结合。

实施中，潜在词义不但是查询词与潜在词义的关系，反过来也成立。例如“玻璃瓶”潜在词义“玻璃|瓶子”，等价于“玻璃|瓶子”潜在“玻璃瓶”，或者“冰箱”潜在词义“电冰箱”，等价于“电冰箱”潜在词义“冰箱”。

在确定了第一查询关键词的潜在词义后，便可以执行步骤 104，步骤 104、反馈以潜在词义，即第二查询关键词为目标进行查询的查询结果了。

实施中，在步骤 101 的获取输入的第一查询关键词时，可以进行如下处理：

获取用户输入的信息内容；

对所述信息内容进行分词处理后获得分词后的词，和/或，将所述信息内容分解为字；

将分词后的词和/或字作为第一查询关键词。

在确定第一查询关键词时，可以采用两种来源，一种是对用户输入的信息内容先进行分词，然后用分词后的结果进行查询，或者将该信息内容以字为单位分解后进行单字查询。显然这两种方式可以同时进行也可以组合进行，在组合时可以是：先对用户输入的查询词分词，再根据分词结果做查询，然后再根据查询词分词的潜在语意做查询，最后做单字查询。分词结果做查询是指根据查询词的分词结果从分词索引中查询相关结果；单字查询是指从单字索引中查询结果；潜在语意查询是指利用查询词的潜在意义得到查询结果，对于在上述实施例中提到的三种语意（或其中任意一种）分别（或单独）做如下处理：

对于第一种潜在词义的词，通过“查询词+第一种潜在词义的词”查询得到相关结果，如查询词是“摩托罗拉”，那么相应的第一种潜在词义的词查询为“摩托罗拉公司”、“摩托罗拉手机”，这里假定“摩托罗拉”的第一种潜在词义的词是“公司”和“手机”；对于第二种潜在词义的词，通过第二种潜在词义的“相邻查询词”得到查询结果，比如“玻璃瓶”相应的第二种潜在词义的词为“玻璃|瓶子”；对于第三种潜在词义的词，是通过第三种潜在词义的词得到的查询结果，例如查询“电冰箱”，第三潜在词义的词是“冰箱”。

显然，基于潜在词义查询的查询结果在计算查询词与文档的相关程度时，应该比单字查询得到结果的相关程度高，这个相关程度的分值会影响查询结果的排序（根据相关程度和网页重要程度等，如 pageRank）。

进一步的，实施中还可以在步骤 101 获取第一查询关键词时，还获取输入第一查询关键词的用户的用户特征；即，可以在用户输入第一查询关键词时，获取该用户的用户特征。

这样，在步骤 102 获取用户反馈日志时，还可以根据用户特征获取用户反馈日志。

或者，在获取用户反馈日志时，获取用户反馈日志中包括以所述第一查询关键词为目标进行查询的历史查询结果，以及用户对历史查询结果的点选频率，而在这些历史查询结果中则包括了这些用户特征。

或者，在根据用户反馈日志确定第二查询关键词时，根据用户特征确定第二查询关键词。

即：在根据用户反馈日志匹配第二查询关键词时，还可以根据输入第一查询关键词的用户特征匹配不同的第二查询关键词。采用用户特征来对用户反馈日志进行甄选，有利于更进一步的发现第一查询关键词的潜在词义。比如：按前述实施例，用户在搜索“床”时，大部分用户的实际意图是睡觉的床，比如“单人床”、“双人床”、“木板床”等，而不是机械设备比如“机床”或“车床”。这时通过用户点击等反馈就能知道前者有“床”的潜在词义，而潜在词义中则不包含“机床”等；然而，同样的查询关键词“床”，如果用户是机械设备领域的技术人员，则其潜在词义则应当是“机床”，而非“单人床”、“双人床”、“木板床”等，本实施例中，“机械设备领域的技术人员”便是用户特征，其作用在于对用户反馈日志进行分类，以便更好的发现词的潜在词义。

再例如：用户输入的第一查询关键词是“苹果”，如果用户特征是计算机工作者，则匹配电脑类的第二查询关键词；如果用户特征是农业科学工作者，则匹配水果类的第二关键词。具体实施中，用户特征可以包括用户所在区域（例如所在国家、地区、城镇）、用户以前频繁浏览的网页、用户不久前浏览的网页、用户以前输入的搜索关键词、用户的性别、年龄、职业、爱好等等。对用户特征的分析归类上，可以根据需要使用分析 IP 地址、分析用户端浏览器历史数据、分析用户端 COOKIE 数据、分析用户网上注册信息等技术手段，这对本领域技术人员来说是容易了解的。

基于同一发明构思，本申请还提供了一种词匹配方法及装置、一种信息查询装置，由于词匹配方法及装置、信息查询装置与信息查询方法是基于同一发明构思，它们具有相似的原理，因此在词匹配方法及装置、信息查询装置实施

中可以参考信息查询方法的实施，重复之处不再赘述。

图 2 为信息查询装置结构示意图，如图所示，装置中可以包括:

第一查询关键词获取模块 201，用于获取输入的第一查询关键词;

用户反馈日志获取模块 202，用于获取第一查询关键词的用户反馈日志;

匹配模块 203，用于根据所述用户反馈日志确定与第一查询关键词匹配的第二查询关键词;

查询结果反馈模块 204，用于反馈以第二查询关键词为目标进行查询的查询结果。

实施中，用户反馈日志获取模块可以进一步用于获取包括历次以所述第一查询关键词为目标进行查询的历史查询结果，以及历次用户对历史查询结果的点选频率的用户反馈日志;

匹配模块则可以进一步用于根据所述用户反馈日志中的历史查询结果以及点选频率确定与第一查询关键词匹配的第二查询关键词。

实施中，用户反馈日志获取模块可以进一步用于获取包括: 对历史查询结果的点选频率和/或对历史查询结果的内容的点选频率作为所述点选频率。

图 3 为匹配模块结构示意图，如图所示，匹配模块可以包括:

内容获取单元 2031，用于获取第一关键词的历史查询结果的内容;

分词单元 2032，用于对历史查询结果的内容进行分词处理获得分词后的词;

匹配单元 2033，用于根据分词后的词的点选频率确定与第一查询关键词匹配的第二查询关键词。

在实施中，分词单元还可以进一步用于在分词后获得下述方式的词或者其组合:

分词后与第一查询关键词相邻的词;

分词后包含第一查询关键词的词;

分词后包括第一查询关键词组成部分的词。

实施中，匹配模块可以进一步用于在根据所述历史查询结果以及点选频率确定与第一查询关键词匹配的第二查询关键词时，所述点选频率大于设定阈值。

图 4 为第一查询关键词获取模块结构示意图，如图所示，第一查询关键词获取模块中可以包括:

信息内容获取单元 2011，用于获取用户输入的信息内容;

分词/分解单元 2012，用于对所述信息内容进行分词处理后获得分词后的词，和/或，将所述信息内容分解为字;

第一查询关键词确定单元 2013，用于将分词后的词和/或字作为第一查询关键词。

实施中，用户反馈日志获取模块可以进一步用于获取包括历史查询结果的点击频率、历史查询结果的曝光频率、对历史查询结果的阅读时间、历史查询结果的重要度其中之一或者其组合的参数作为点选频率。

实施中，第一查询关键词获取模块可以进一步用于在用户输入第一查询关键词时，获取该用户的用户特征用户特征; 用户反馈日志获取模块可以进一步用于根据用户特征获取用户反馈日志。

实施中，第一查询关键词获取模块可以进一步用于在用户输入第一查询关键词时，获取该用户的用户特征;

用户反馈日志获取模块还可以进一步用于在获取用户反馈日志时，获取用户反馈日志中包括以所述待匹配词为目标进行查询的历史查询结果，以及用户对历史查询结果的点选频率，所述历史查询结果包括所述用户特征。

实施中，第一查询关键词获取模块还可以进一步用于在用户输入第一查询关键词时，获取该用户的用户特征;

匹配模块可以进一步用于在根据所述用户反馈日志确定第二查询关键词时，根据用户特征确定第二查询关键词。

图 5 为词匹配方法实施流程示意图，如图所示，在进行词匹配时可以包括

如下步骤:

步骤 501、获取待匹配词;

步骤 502、根据待匹配词获取用户反馈日志，所述用户反馈日志包括历次以所述待匹配词为目标进行查询的历史查询结果，以及历次用户对历史查询结果的点选频率;

步骤 503、根据所述历史查询结果以及点选频率确定与待匹配词匹配的词。

实施中，点选频率可以包括: 对历史查询结果的点选频率和/或对历史查询结果的内容的点选频率。

实施中，根据对历史查询结果的内容的点选频率确定与待匹配词匹配的词，可以为:

获取待匹配词的历史查询结果的内容;

对历史查询结果的内容进行分词处理获得分词后的词;

根据分词后的词的点选频率确定与待匹配词匹配的词。

实施中，分词后的词是指下述方式的词或者其组合:

分词后与待匹配词相邻的词;

分词后包含待匹配词的词;

分词后包括待匹配词组成部分的词。

实施中，在根据所述历史查询结果以及点选频率确定与待匹配词匹配的词时，所述点选频率大于设定阈值。

获取待匹配关键词时，可以为:

获取用户输入的信息内容;

对所述信息内容进行分词处理后获得分词后的词，和/或，将所述信息内容分解为字;

将分词后的词和/或字作为待匹配词。

实施中，点选频率可以包括历史查询结果的点击频率、历史查询结果的曝光频率、对历史查询结果的阅读时间、历史查询结果的重要度其中之一或者其

组合。

实施中，还可以进一步包括:

在用户输入待匹配词时，获取该用户的用户特征;

获取用户反馈日志时，根据用户特征获取用户反馈日志。

实施中，还可以进一步包括:

在用户输入待匹配词时，获取该用户的用户特征;

获取用户反馈日志时，获取用户反馈日志中包括以所述待匹配词为目标进行查询的历史查询结果，以及用户对历史查询结果的点选频率，所述历史查询结果包括所述用户特征。

实施中，还可以进一步包括:

在用户输入待匹配词时，获取该用户的用户特征;

根据用户反馈日志确定与待匹配词匹配的词时，根据所述用户特征确定与待匹配词匹配的词。

图 6 为词匹配装置结构示意图，如图所示，可以包括:

待匹配词获取模块 601，用于获取待匹配词;

用户反馈日志获取模块 602，用于根据待匹配词获取用户反馈日志;

匹配模块 603，用于根据所述用户反馈日志确定与待匹配词匹配的词。

实施中，用户反馈日志获取模块可以进一步用于获取包括历次以所述待匹配词为目标进行查询的历史查询结果，以及历次用户对历史查询结果的点选频率的用户反馈日志;

匹配模块可以进一步用于根据所述用户反馈日志中的历史查询结果以及点选频率确定与待匹配词匹配的词。

用户反馈日志获取模块可以进一步用于获取包括: 对历史查询结果的点选频率和/或对历史查询结果的内容的点选频率作为所述点选频率。

实施中，匹配模块可以包括:

内容获取单元，用于获取待匹配词的历史查询结果的内容;

分词单元，用于对历史查询结果的内容进行分词处理获得分词后的词；

匹配单元，用于根据分词后的词的点选频率确定与待匹配词匹配的词。

分词单元可以进一步用于在分词后获得下述方式的词或者其组合：

分词后与待匹配词相邻的词；

分词后包含待匹配词的词；

分词后包括待匹配词组成部分的词。

匹配模块可以进一步用于在根据所述历史查询结果以及点选频率确定与待匹配词匹配的词时，所述点选频率大于设定阈值。

待匹配词获取模块可以包括：

信息内容获取单元，用于获取用户输入的信息内容；

分词/分解单元，用于对所述信息内容进行分词处理后获得分词后的词，和/或，将所述信息内容分解为字；

待匹配词确定单元，用于将分词后的词和/或字作为待匹配词。

用户反馈日志获取模块可以进一步用于获取包括历史查询结果的点击频率、历史查询结果的曝光频率、对历史查询结果的阅读时间、历史查询结果的重要度其中之一或者其组合的参数作为点选频率。

实施中，待匹配词获取模块进一步用于在用户输入待匹配词时，获取该用户的用户特征；用户反馈日志获取模块进一步用于根据用户特征获取用户反馈日志。

实施中，待匹配词获取模块可以进一步用于在用户输入待匹配词时，获取该用户的用户特征；

用户反馈日志获取模块还可以进一步用于在获取用户反馈日志时，获取用户反馈日志中包括以所述待匹配词为目标进行查询的历史查询结果，以及用户对历史查询结果的点选频率，所述历史查询结果包括所述用户特征。

实施中，待匹配词获取模块还可以进一步用于在用户输入待匹配词时，获取该用户的用户特征；

匹配模块可以进一步用于在根据所述用户反馈日志确定与待匹配词匹配的词时，根据所述用户特征确定与待匹配词匹配的词。

由上述实施例可知，本申请实施中基于对用户反馈日志分析，因而能够自动发现词语的潜在语意，从而能够准确发现词之间的内在联系；进一步的，还利用自动发现词语的潜在语意和将查询词的相关语意用来提高搜索引擎的效果；进一步的，在自动发现查询词的潜在词义时，还可以根据查询词前后单字的词频，而不是仅用分词结果来达到类似的效果。因此，在本申请实施例中通过自动发现词的潜在词义提高搜索引擎的性能，与传统方式相比，能够提高搜索的精确度和效率；

例如与现有技术中陆勇、侯汉清提到的词义自动发现方法相比，其主要是通过已有的词典数据作为抽取来源，样本量在几千条左右。如果它是以互联网网页等大数据量为基础来抽取，就会缺乏准确性。而本申请实施中通过用户参与的用户反馈日志，就可以非常可靠的自动发现查询词和查询结果之间体现用户意图的潜在词义关系，特别适合原来提高搜索引擎的准确率和智能。

为了描述的方便，描述以上系统时以功能分为各种模块或单元分别描述。当然，在实施本发明时可以把各模块或单元的功能在同一个或多个软件和/或硬件中实现。

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，本申请可采用完全硬件实施例、完全软件实施例、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、CD-ROM、光学存储器等）上实施的计算机程序产品的形式。

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装置。

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装置的制造品，该指令装置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。

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。

尽管已描述了本申请的优选实施例，但本领域内的技术人员一旦得知了基本创造性概念，则可对这些实施例作出另外的变更和修改。所以，所附权利要求意欲解释为包括优选实施例以及落入本申请范围的所有变更和修改。

显然，本领域的技术人员可以对本申请进行各种改动和变型而不脱离本申请的精神和范围。这样，倘若本申请的这些修改和变型属于本申请权利要求及其等同技术的范围之内，则本申请也意图包含这些改动和变型在内。

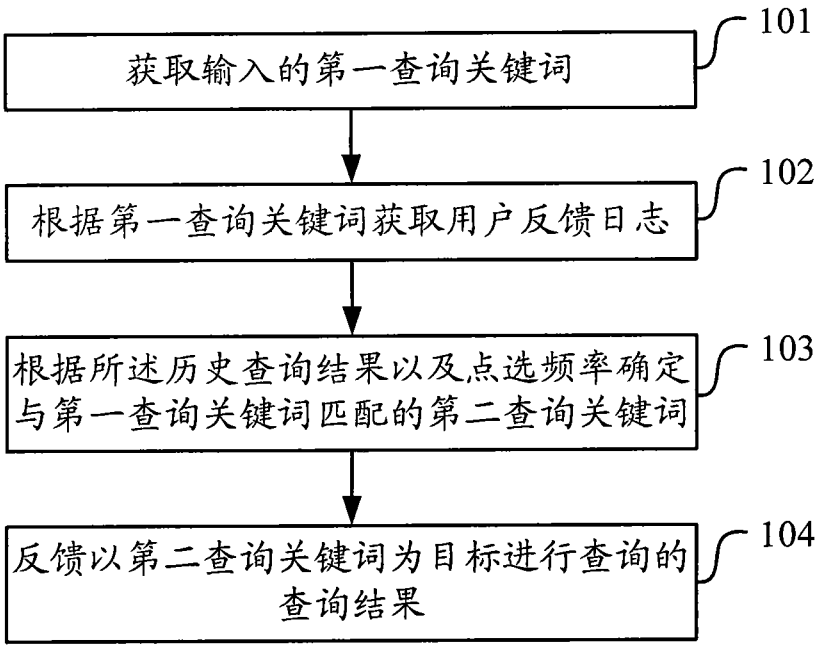

101
获取输入的第一查询关键词
102
根据第一查询关键词获取用户反馈日志
103
根据所述历史查询结果以及点选频率确定与第一查询关键词匹配的第二查询关键词
104
反馈以第二查询关键词为目标进行查询的查询结果

图 1

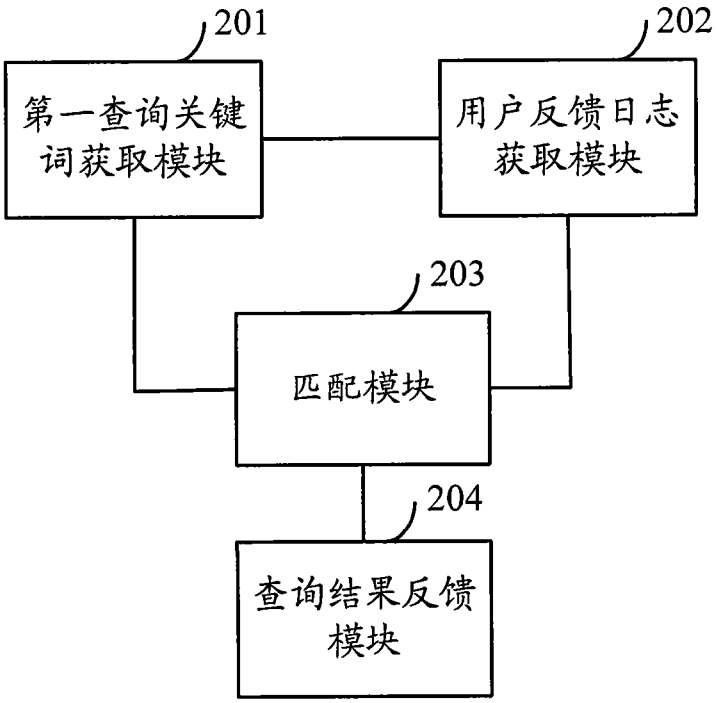

201
第一查询关键词获取模块
202
用户反馈日志获取模块
203
匹配模块
204
查询结果反馈模块

图 2

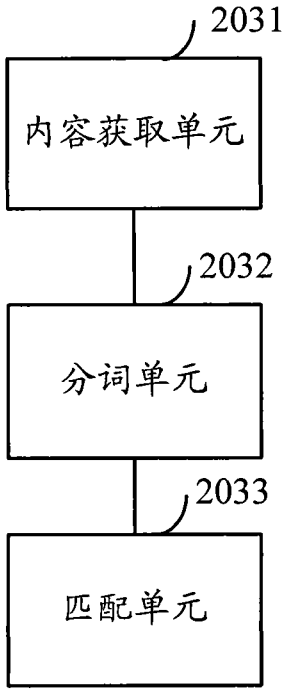

2031
内容获取单元
2032
分词单元
2033
匹配单元

图 3

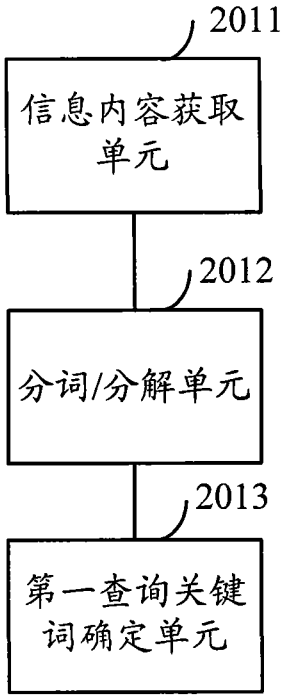

2011
信息内容获取单元
2012
分词/分解单元
2013
第一查询关键词确定单元

图 4

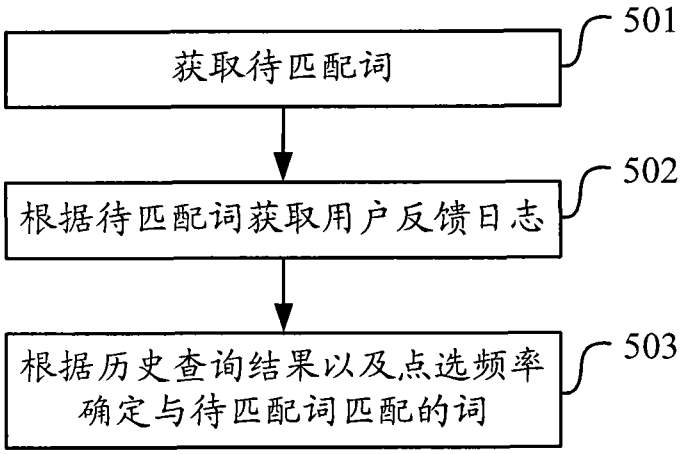

501
获取待匹配词
502
根据待匹配词获取用户反馈日志
503
根据历史查询结果以及点选频率确定与待匹配词匹配的词

图 5

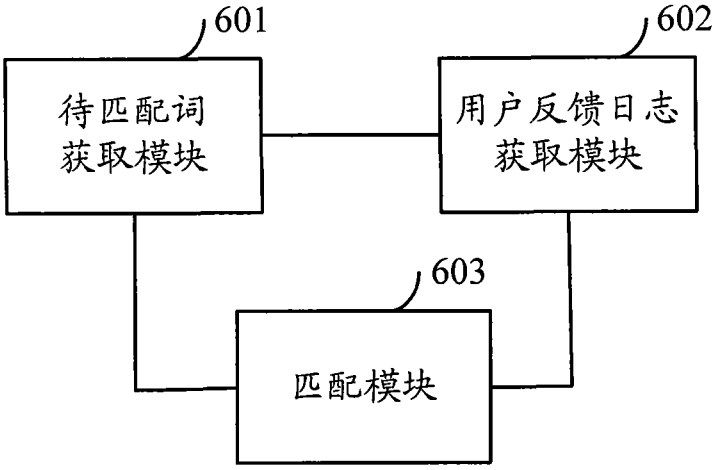

601
待匹配词获取模块
602
用户反馈日志获取模块
603
匹配模块

图 6