(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 104424291 A
(43)申请公布日 2015.03.18

(21)申请号 201310392145.6

(22)申请日 2013.09.02

(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资本大厦一座四层847号邮箱

(72)发明人 霍承富 郑伟 朱江涛 林锋

(74)专利代理机构 北京国昊天诚知识产权代理有限公司 11315
代理人 许志勇

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

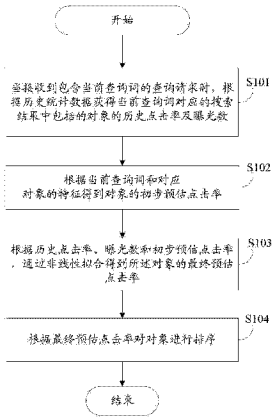
权利要求书2页 说明书12页 附图3页

(54)发明名称

一种对搜索结果进行排序的方法及装置

(57)摘要

本申请提出一种对搜索结果进行排序的方法及装置。该方法包括：当接收到包含当前查询词的查询请求时，根据历史统计数据获得当前查询词对应的搜索结果中包括的对象的历史点击率及曝光数；根据当前查询词和对应对象的特征得到对象的初步预估点击率；以及根据历史点击率、曝光数和初步预估点击率，通过非线性拟合得到对象的最终预估点击率；以及根据最终预估点击率对对象进行排序。本申请所提出的方案有效地达到了准确的点击率预估，进而实现了更准确的搜索结果排序，以满足用户的需求和方便用户的使用。



CN 104424291 A

1. 一种对搜索结果进行排序的方法，其特征在于，包括：

当接收到包含当前查询词的查询请求时，根据历史统计数据获得所述当前查询词对应的搜索结果中包括的对象的历史点击率及曝光数；

根据所述当前查询词和对应对象的特征得到所述对象的初步预估点击率；以及

根据所述历史点击率、曝光数和所述初步预估点击率，通过非线性拟合得到所述对象的最终预估点击率；以及

根据所述最终预估点击率对对象进行排序。

2. 根据权利要求 1 所述的方法，其特征在于，所述非线性拟合是指数型非线性拟合。

3. 根据权利要求 1 所述的方法，其特征在于，还包括：

从网络日志数据中获得曝光数大于或等于预定阈值的历史查询词和所述历史查询词对应的对象；以及

通过统计得到历史统计数据，所述历史统计数据包括所述历史查询词对应的对象的历史点击率和曝光数。

4. 根据权利要求 3 所述的方法，其特征在于，所述根据历史统计数据获得所述当前查询词对应的搜索结果中包括的对象的历史点击率及曝光数的步骤进一步包括：

当所述当前查询词与所述历史查询词中的一个匹配时，将所述历史查询词对应的对象的历史点击率和曝光数，作为所述当前查询词对应的搜索结果中包括的对象的历史点击率及曝光数。

5. 根据权利要求 3 所述的方法，其特征在于，

所述根据历史统计数据获得所述当前查询词对应的搜索结果中包括的对象的历史点击率及曝光数的步骤进一步包括：

当所述当前查询词与所述历史查询词中的任何一个均不匹配时，判断所述当前查询词是否与所述历史查询词语义相关的近义查询词中的任何一个匹配；

如果匹配，则将所述匹配的近义查询词对应的历史查询词对应的对象的历史点击率和曝光数，作为所述当前查询词对应的搜索结果中包括的对象的历史点击率及曝光数；以及

如果不匹配，则根据针对所述历史查询词的平均历史点击率和针对所述历史查询词对应的对象的平均历史点击率，计算所述当前查询词对应的搜索结果中包括的对象的历史点击率，并且将给定预设值作为所述当前查询词对应的搜索结果中包括的对象的曝光数。

6. 根据权利要求 3 所述的方法，其特征在于，所述根据所述当前查询词和所述对象的特征得到所述对象的初步预估点击率的步骤进一步包括：

基于所述历史统计数据提取历史查询词和对应对象的特征；

得到所述历史查询词和对应对象的特征的特征权重；以及

根据所述特征权重建立预估模型，所述预估模型用于根据所述当前查询词和对应对象的特征预估所述当前查询词对应对象的初步预估点击率。

7. 根据权利要求 6 所述的方法，其特征在于，所述根据所述当前查询词和对应对象的特征得到所述对象的初步预估点击率的步骤进一步包括：

提取所述当前查询词和对应对象的特征；以及

根据所述当前查询词和对应对象的特征，通过所述预估模型得到所述当前查询词对应对象的初步预估点击率。

8. 根据权利要求 6-7 中任一项所述的方法，其特征在于，所述特征包括查询词的文本特征、查询词对应对象的标题和 / 或属性特征、以及所述查询词与对应对象的相关性特征。

9. 一种对对搜索结果进行排序的装置，其特征在于，包括 ：

历史预估模块，用于当接收到包含当前查询词的查询请求时，根据历史统计数据获得所述当前查询词对应的搜索结果中包括的对象的历史点击率及曝光数 ；

初步预估模块，用于根据所述当前查询词和对应对象的特征得到所述对象的初步预估点击率 ；

最终预估模块，用于根据所述历史点击率、曝光数和所述初步预估点击率，通过非线性拟合得到所述对象的最终预估点击率 ；以及

排序模块，用于根据所述最终预估点击率对对象进行排序。

10. 根据权利要求 9 所述的装置，其特征在于，所述非线性拟合是指数型非线性拟合。

11. 根据权利要求 9 所述的装置，其特征在于，还包括 ：

获取模块，用于从网络日志数据中获得曝光数大于或等于预定阈值的历史查询词和所述历史查询词对应的对象 ；以及

统计模块，用于通过统计得到历史统计数据，所述历史统计数据包括所述历史查询词对应的对象的历史点击率和曝光数。

12. 根据权利要求 11 所述的装置，其特征在于，所述历史预估模块还包括 ：

第一子模块，用于当所述当前查询词与所述历史查询词中的一个匹配时，将所述历史查询词对应的对象的历史点击率和曝光数，作为所述当前查询词对应的搜索结果中包括的对象的历史点击率及曝光数。

13. 根据权利要求 11 所述的装置，其特征在于，所述历史预估模块还包括 ：

第二子模块，用于当所述当前查询词与所述历史查询词中的任何一个均不匹配时，判断所述当前查询词是否与所述历史查询词语义相关的近义查询词中的任何一个匹配 ；

如果匹配，则将所述匹配的近义查询词对应的历史查询词对应的对象的历史点击率和曝光数，作为所述当前查询词对应的搜索结果中包括的对象的历史点击率及曝光数 ；以及

如果不匹配，则根据针对所述历史查询词的平均历史点击率和针对所述历史查询词对应的对象的平均历史点击率，计算所述当前查询词对应的搜索结果中包括的对象的历史点击率，并且将给定预设值作为所述当前查询词对应的搜索结果中包括的对象的曝光数。

14. 根据权利要求 11 所述的装置，其特征在于，所述初步预估模块还包括 ：

第一提取子模块，用于基于所述历史统计数据提取历史查询词和对应对象的特征 ；

第一计算子模块，用于得到所述历史查询词和对应对象的特征的特征权重 ；以及

第二计算子模块，用于根据所述特征权重建立预估模型，所述预估模型用于根据所述当前查询词和对应对象的特征预估所述当前查询词对应对象的初步预估点击率。

15. 根据权利要求 14 所述的装置，其特征在于，所述初步预估模块还包括 ：

第二提取子模块，用于提取所述当前查询词和对应对象的特征 ；以及

初步预估子模块，用于根据所述当前查询词和对应对象的特征，通过所述预估模型得到所述当前查询词对应对象的初步预估点击率。

16. 根据权利要求 14-15 中任一项所述的装置，其特征在于，所述特征包括查询词的文本特征、查询词对应对象的标题和 / 或属性特征、以及所述查询词与对应对象的相关性特征。

一种对搜索结果进行排序的方法及装置

技术领域

[0001] 本申请涉及计算机网络信息领域，尤其涉及一种对搜索结果进行排序的方法及装置。

背景技术

[0002] 随着互联网业务迅速发展，通过互联网进行多种多样的信息交互成为当今最为广泛的应用。然而，当多种业务同时交互应用时就会产生交互应用堵塞现象。例如，用户想要搜索某一产品时，就会同时出现众多产品列表，但呈现在用户眼前的并不都是用户所需要的，甚至有很多与搜索内容是无关的，这就说明，对搜索结果的排序不够合理化或不够精确。因此，不但给用户应用带来了许多的不便，而且也带来了网络资源的浪费。从而这也体现出合理精确化排序的重要性。

[0003] 当前现有技术中对搜索结果进行排序的方法通常是根据对搜索结果中包括的对象预估的点击率来对对象进行排序，而预估点击率的方法通常是基于特征提取和模型训练的模型预估。具体而言，提取历史查询词和对应的历史对象的文本特征和相关性特征来建立预估模型对对象的点击率进行预估，这些特征一定程度上能够影响用户的关注度，从而描述用户的点击行为。依据此种模型预估的点击率对搜索结果进行排序的方法尽管减少了一些不必要的资源浪费，但是其精确度却不高并且有时还会遗漏相关的重要信息。

[0004] 例如，在上述方法中只关注了用户已经点击过的对象的文本特征和相关性特征，而没有考虑到对象本身对用户的点击行为也起到关键作用的其它重要特征，如图片视觉感知和标签等难以被特征化表示的信息，因而并不能精确地反映出用户的点击行为，也就是预估精确度有待提高。另外，由于只关注了用户点击过的对象，因此忽略了一些暂时没有用户点击或点击量很低的对象，而这种对象往往很可能包含特定用户当前所希望搜索到的信息。当用户搜索这种对象时，由于其点击量过小而无法被搜索到，这就带来资源信息的滞留和浪费。此外，针对拥有大量用户搜索行为、少量用户点击行为的对象来说，由于用户对它的点击行为较少，会造成该对象可能被忽略而不被呈现给用户。因此，现有技术中的方法并没有充分和合理地利用网络日志中的历史数据信息。

发明内容

[0005] 本申请的主要目的在于提供一种对搜索结果进行排序的方法及装置，以解决现有技术存在的上述问题。

[0006] 根据本申请的一个方面的实施例，提出一种对搜索结果进行排序的方法，包括：当接收到包含当前查询词的查询请求时，根据历史统计数据获得当前查询词对应的搜索结果中包括的对象的历史点击率及曝光数；根据当前查询词和对应对象的特征得到对象的初步预估点击率；以及根据历史点击率、曝光数和初步预估点击率，通过非线性拟合得到对象的最终预估点击率；以及根据最终预估点击率对对象进行排序。

[0007] 根据本申请的实施例，在该方法中，非线性拟合是指数型非线性拟合。

[0008] 根据本申请的实施例，在该方法中，还包括：从网络日志数据中获得曝光数大于或等于预定阈值的历史查询词和历史查询词对应的对象；以及通过统计得到历史统计数据，所述历史统计数据包括历史查询词对应的对象的历史点击率和曝光数。

[0009] 根据本申请的实施例，在该方法中，根据历史统计数据获得当前查询词对应的搜索结果中包括的对象的历史点击率及曝光数包括：当当前查询词与历史查询词中的一个匹配时，将历史查询词对应的对象的历史点击率和曝光数，作为当前查询词对应的搜索结果中包括的对象的历史点击率及曝光数。

[0010] 根据本申请的实施例，在该方法中，根据历史统计数据获得当前查询词对应的搜索结果中包括的对象的历史点击率及曝光数包括：当当前查询词与历史查询词中的任何一个均不匹配时，判断当前查询词是否与历史查询词语义相关的近义查询词中的任何一个匹配；如果匹配，则将匹配的近义查询词对应的历史查询词对应的对象的历史点击率和曝光数，作为当前查询词对应的搜索结果中包括的对象的历史点击率及曝光数；以及如果不匹配，则根据针对历史查询词的平均历史点击率和针对历史查询词对应的对象的平均历史点击率，计算当前查询词对应的搜索结果中包括的对象的历史点击率，并且将给定预设值作为当前查询词对应的搜索结果中包括的对象的曝光数。

[0011] 根据本申请的实施例，在该方法中，根据当前查询词和对象的特征得到对象的初步预估点击率包括：基于历史统计数据提取历史查询词和对应对象的特征；得到历史查询词和对应对象的特征的特征权重；以及根据特征权重建立预估模型，所述预估模型用于根据当前查询词和对应对象的特征预估当前查询词对应对象的初步预估点击率。

[0012] 根据本申请的实施例，在该方法中，根据当前查询词和对应对象的特征得到对象的初步预估点击率包括：提取当前查询词和对应对象的特征；以及根据当前查询词和对应对象的特征，通过所述预估模型得到所述当前查询词对应对象的初步预估点击率。

[0013] 根据本申请的实施例，在该方法中，特征包括查询词的文本特征、查询词对应对象的标题和／或属性特征、以及查询词与对应对象的相关性特征。

[0014] 根据本申请的另一方面的实施例，提出一种对搜索结果进行排序的装置，包括：历史预估模块，用于当接收到包含当前查询词的查询请求时，根据历史统计数据获得当前查询词对应的搜索结果中包括的对象的历史点击率及曝光数；初步预估模块，用于根据当前查询词和对应对象的特征得到对象的初步预估点击率；最终预估模块，用于根据历史点击率、曝光数和初步预估点击率，通过非线性拟合得到对象的最终预估点击率；以及排序模块，用于根据最终预估点击率对对象进行排序。

[0015] 与现有技术相比，根据本申请的技术方案，通过针对对象的历史点击率、曝光数和初步预估点击率这三个参数进行非线性拟合来得到该对象的最终预估点击率，可以综合地表征用户的点击行为，实现更准确的点击率预估，进而实现更准确的搜索结果排序，以满足用户的需求和方便用户的使用。

[0016] 进一步而言，根据本申请的技术方案，通过提炼网络日志数据，基于历史查询词对应的对象的曝光数来统计历史数据信息，以得到更合理的历史统计数据，从而提高点击率预估效率。尤其是，针对曝光数较低或为零的对象也提供合理的点击率预估，从而获得与当前查询词对应的所有对象的更准确合理的点击率预估，进而在更加合理地利用历史数据信息的情况下实现更准确且更高效的搜索结果排序。

附图说明

[0017] 此处所说明的附图用来提供对本申请的进一步理解，构成本申请的一部分，本申请的示意性实施例及其说明用于解释本申请，并不构成对本申请的不当限定。在附图中：

[0018] 图 1 是根据本申请实施例的对搜索结果进行排序的方法的流程图。

[0019] 图 2 是根据本申请实施例的获得当前查询词对应对象的历史点击率及曝光数的步骤的具体流程图。

[0020] 图 3 是根据本申请实施例的得到对象的初步预估点击率的步骤的具体流程图。

[0021] 图 4 是根据本申请实施例的对搜索结果进行排序的装置的结构框图。

[0022] 图 5 是根据本申请实施例的求得最终预估点击率过程中的拟合因子曲线图。

具体实施方式

[0023] 本申请的主要思想在于：通过针对对象的历史点击率、曝光数和初步预估点击率这三个参数进行非线性拟合来得到该对象的最终预估点击率，可以综合地表征用户的点击行为，实现更准确的点击率预估，进而实现更准确的搜索结果排序。另外，在预估对象的点击率时，可以基于历史数据的曝光数信息更合理地利用历史数据信息，提高点击率预估效率。尤其是，针对曝光数较低或为零的对象也提供合理的点击率预估值，由此对于当前查询词对应的各个对象来说，不管它是曝光率高的对象，还是曝光率低的对象，都可以得到更合理的点击率预估，由此可以在更合理地利用历史数据资源信息的前提下实现更准确且更高效的点击率预估和搜索结果的排序，从而使得搜索结果更符合用户需求，方便用户的使用。

[0024] 为使本申请的目的、技术方案和优点更加清楚，下面将结合本申请具体实施例及相应的附图对本申请技术方案进行清楚、完整地描述。显然，所描述的实施例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。

[0025] 根据本申请的实施例，提供了一种对搜索结果进行排序的方法。

[0026] 参考图 1，图 1 是根据本申请实施例的对搜索结果进行排序的方法的流程图。

[0027] 如图 1 所示，在步骤 S101 中，当接收到包含当前查询词的查询请求时，根据历史统计数据获得当前查询词对应的搜索结果中包括的对象的历史点击率及曝光数。根据本申请的实施例，对象的历史点击率及曝光数可以是对象在过去一段时间内例如 2-3 个月内的历史点击率及曝光数。

[0028] 当前查询词对应的搜索结果中包括的对象可以理解为产品、服务、网页、新闻、广告等。例如，在某网站上要搜索关于某一对象，用户会先根据自己的需要输入与之相关的查询词，即当前查询词。具体而言，比如，要查询 mp3 播放器，大多数用户可能就会输入查询词“mp3”，但少部分对外形有特定要求的用户可能会输入查询词“love style red mp3”等。

[0029] 历史统计数据可以是从网络日志数据预先得到的。根据本申请的一个实施例，在步骤 S101 之前，可以包括步骤：从网络日志数据中获得曝光数大于或等于预定阈值的历史查询词和历史查询词对应的对象；以及通过统计得到历史统计数据，历史统计数据包括历史查询词对应的对象的历史点击率和曝光数。

[0030] 下面进一步举例说明，从网络日志数据中获得曝光数大于或等于预定阈值的历史

查询词和历史查询词对应的对象、历史查询词对应的对象的历史点击率和曝光数的实际获取过程。

[0031] 根据本申请的一个实施例，由于网络日志中所记录的数据也并非都是真实有效，因此可以对网络日志数据进行预处理，如日志数据反欺诈、反爬虫过滤等相应的安全处理。进而可以在数据安全有效的前提下，统计出历史查询词对应对象的历史点击率和曝光数，提高数据的置信度。

[0032] 如表 1 所示，示出了日志数据的示例。其中，对象的曝光点击日志数据内容格式通常如表 1 所示，包括查询词 Query、曝光对象的 Offer_ID、曝光对象的标题、该曝光对象的展示位置、该曝光对象是否被点击等字段。

[0033]

序号	名称	类型	举例	说明
1	Query	String	Mp3 player	搜索查询词
2	Offer_ID	Bigint	10046592	标识对象的 ID
3	Subject	String	1.3 MP 2GB memory Recordable video pen 80 minutes 30fps video	对象的标题
5	Rank	Int	2	对象曝光的展示位置
6	Is_Click	Int	1	对象是否被点击
.....	.....	.....	.....	.....

[0034] 表 1

[0035] 根据表 1 的网络日志中所获取的相关数据可以整理和统计出历史查询词对应的对象以及相应的曝光数和点击数。

[0036] 根据本申请的实施例，可以基于例如曝光数这个参数来对网络日志数据进行过滤。具体而言，可以将曝光数大于或等于预定阈值的对象（即，常用对象，出现率高、常被搜索或应用的查询词对应的对象）保留，而将曝光数小于预定阈值的对象（即，不常用对象，出现率低、不常被搜索或应用的查询词对应的对象）丢弃，也就是只将曝光数大于或等于预定值的对象用于历史点击率的预估。由此可以避免因对大量不常用对象的历史点击曝光信息创建索引而影响对常用对象的历史统计数据的查询效率，进而可以充分利用所有有效日志资源和提高查询效率。

[0037] 然后，基于上述所保留的对象的曝光数和点击数的统计结果，可以通过本领域已知或未来开发的任意合适方式来得到相应对象的历史点击率。传统上而言，可以通过用对象的点击数除以对象的曝光数来计算对象的历史点击率。在一种更优选实施例中，可以通过对用户点击行为建模，例如，优选地，通过贝叶斯浏览模型 BBM（可参见 C.Liu,F.Guo and C.Faloutsos,KDD2009,“BBM:Bayesian Browsing Model from Petabyte-scale Data”），来计算历史查询词对应对象的标准点击率 sCTR(Standard CTR) 作为该对象的历史点击率。具体而言，设一段时间内日志数据经预处理得到一标准点击率为 sCTR，与历史点击率 hCTR 在本质上是等同的，因此可以直接赋值，即 hCTR=sCTR。相比于上述传统的方法而言，通过

BBM 方法能够消除对象在日志中的位置对其历史点击率的影响，使得计算得到的历史点击率更具可比性。

[0038] 以上描述了如何从网络日志数据中获取历史统计数据的过程，实质为网络线下的预处理操作。

[0039] 当在步骤 S101 中接收到包含当前查询词的查询请求时，可以根据历史统计数据查找与当前查询词对应的历史查询词，从而获取与当前查询词对应的对象以及该对象的历史点击率及曝光数。关于这一点，稍后将结合图 2 进行更具体描述，这里不再赘述。

[0040] 接下来，在步骤 S102 中，根据当前查询词和对应对象的特征得到对象的初步预估点击率。

[0041] 具体而言，步骤 S102 实际可以视为根据当前查询词和对象的特征来为对象的点击率赋予一个初始值，即初步预估点击率。也就是说，当接收到包含当前查询词的查询请求时，是需要对当前查询词有一个初步预估点击率，进而为后续结合历史点击率及曝光得到最终预估点率奠定高精确度基础。

[0042] 应理解到，步骤 S102 可以在步骤 S101 之前或之后执行，也可以与步骤 S101 同时执行。

[0043] 在一个实施例中，可以通过预先建立的预估模型来执行步骤 S102，即，为当前查询词对应对象赋予一个初步预估点击率。具体而言，可以通过特征提取和模型训练来预先建立预估模型(线下操作)。

[0044] 更具体地，可以基于前述从网络日志数据中预先得到的历史统计数据，提取历史查询词和其对应对象的特征。在一个具体实施例中，该特征可以包括历史查询词的文本特征、历史查询词对应对象的标题和 / 或属性等特征、以及历史查询词与对应对象的相关性特征。

[0045] 然后，通过基于上述历史统计数据中的对象作为样本进行模型训练，为上述所提取的各个特征赋予相应的特征权重。

[0046] 针对于特征权重的取得，优选地，如公式(1)，可由最优化目标函数实现。其中，可以将上述通过贝叶斯浏览模型 BBM 得到的 k 个历史查询词及对应对象的集合的标准点击率 $sCTR_k$ 作为目标值，通过给定这 k 个历史查询词及对应对象的集合的模型预估点击率 y_k，训练得到诸如查询词特征、对象特征及相关性特征之类的 i 个特征的特征权重 w_i，用公式表示为

[0047] $$w_i=\min_{w_i}\{\sum_k(y_k-sCTR_k)^2+C\cdot L(w_i)\} \qquad (1)$$

[0048] 其中，$C\cdot L(w_i)$ 为约束项，用于确保特征权重具有稀疏性，减少弱特征，提升预估模型计算效率。

[0049] 接下来，根据特征权重建立预估模型，使预估模型用于根据当前查询词和对应对象的特征预估当前查询词对应对象的初步预估点击率。

[0050] 在一个具体实施例中，可以通过 Logisitc 回归模型得到初步预估点击率，如公式(2)所示。

[0051] $$eCTR=\frac{1}{1+e^{-\sum_i w_i f_i}} \qquad (2)$$

[0052] 其中 f_i 为当前查询词和对应对象的特征。

[0053] 在其它实施例中，还可以采用支持向量机（SVM，Support Vector Machine）模型、决策树模型（GBDT，Gradient Boost Decision Tree）等得到初步预估点击率，本申请对此不作任何限制。下面将结合图 3 以 Logistic 回归模型为例更详细地说明初步预估点击率 eCTR 的获取过程。

[0054] 参考图 3，图 3 是根据本申请实施例的得到对象的初步预估点击率的步骤（步骤 S102）的具体流程图。

[0055] 如图 3 所示，在步骤 S301 中，提取当前查询词和对应对象的特征。

[0056] 具体而言，当接收到包含当前查询词的查询请求时，可以提取当前查询词的文本特征、当前查询词对应对象的标题和 / 或属性等特征以及当前查询词与对应对象的相关性特征，以便于之后将当前查询词和对应对象的特征放入预先建立的预估模型以得到当前查询词和对应对象的初步预估点击率。

[0057] 在步骤 S302 中，根据当前查询词和对应对象的特征，通过预估模型得到当前查询词对应对象的初步预估点击率。

[0058] 具体而言，分别以当前查询词和对应对象的各特征为关键词去预估模型的索引中查找对应的特征权重，然后将特征值与对应的特征权重加权求和，从而可以预估出当前查询词所对应的对象的初步预估点击率。

[0059] 回到图 1，在通过步骤 S101 和 S102 得到当前查询词对应对象的历史点击率及曝光数和初步预估点击率之后，在步骤 S103 中，根据历史点击率、曝光数和初步预估点击率，通过非线性拟合得到对象的最终预估点击率。

[0060] 根据本申请的一个实施例，优选地，拟合函数可以采用指数型非线性拟合。

[0061] 假设当前查询词对应对象的历史点击率为 hCTR、曝光数为 pv、初步预估点击率为 eCTR。以 S 型变换曲线为例，可以定义拟合因子 λ 和最终预估点击率 CTR 的拟合公式如下：

[0062] $$\begin{cases} \lambda = 1 - \dfrac{2}{1 + e^{\alpha \bullet pv}} \\ CTR = (1 - \lambda) \bullet eCTR + \lambda \bullet hCTR \end{cases}$$

[0063] 其中 α 是根据 pv 与 λ 的经验关系预先设定的参数，用于决定 eCTR 和 hCTR 的拟合曲线，设定 α 之后，eCTR 和 hCTR 的拟合权重会随着 pv 的变化而变化。

[0064] 进一步地，由图 5 所示的拟合因子曲线图可见，当 pv 较小时，历史点击率置信度低，拟合因子 λ 随 pv 变化较缓；当 pv 达到一个可信的范围时，拟合因子 λ 随 pv 变化较快，历史点击率对最终预估点击率 CTR 预估值的贡献度快速增加；当 pv 很大时，历史点击率非常可信，拟合因子 λ 逐渐逼近于 1，最终 CTR 预估值依赖于 hCTR。

[0065] 因此，本申请提出的非线性拟合方法能够更合理地均衡模型预估的点击率和历史点击率，使得历史点击率对最终预估点击率的贡献度随着曝光数的变化而与人的意愿感知更一致，从而能够提升点击率预估精度。

[0066] 这里需要说明的是，尽管上述实施例中采用的是指数型非线性拟合，但本申请并不限于此，而是还可以采用其它本领域已知或未来开发的任意合适的非线性拟合方式来根据对象的历史点击率、曝光数和初步预估点击率预估该对象的最终预估点击率。

[0067] 接下来，在步骤 S104 中，根据最终预估点击率对对象进行排序。

[0068] 具体而言，在得到当前查询词对应的搜索结果中包括的所有对象的最终预估点击率之后，可以根据最终预估点击率的高低顺序，对搜索结果中的各对象进行升序或降序排列，从而显示给用户。

[0069] 在一种典型的应用场景中例如广告排序时，可以通过按照广告对象的最终预估点击率与广告对象的优先级系数（竞价系数）的乘积的大小顺序来对广告对象进行排序，其中优先级系数用于体现广告对象的优先级设置。

[0070] 参考图 2，图 2 是根据本申请实施例的获得当前查询词对应对象的历史点击率及曝光数的步骤（步骤 S101）的具体流程图。

[0071] 如图 2 所示，在步骤 S201 中，判断当前查询词是否与历史查询词中的一个匹配。

[0072] 如前面提及的，从网络日志数据可以预先得到历史统计数据，历史统计数据可以包括历史查询词对应的对象的历史点击率和曝光数。在一个实施例中，历史统计数据可以包括历史查询词、与该历史查询词对应的各对象以及各对象对应的历史点击率和曝光数。

[0073] 当接收到包含当前查询词的查询请求时，服务器可以判断当前查询词是否与历史统计数据中的历史查询词中的一个匹配。

[0074] 如果匹配，则前进到步骤 S202。在步骤 S202 中，将匹配的历史查询词对应的对象的历史点击率和曝光数，作为当前查询词对应的搜索结果中包括的对象的历史点击率及曝光数。

[0075] 具体而言，在当前查询词与历史统计数据中的历史查询词中的一个匹配时，可以获取该匹配的历史查询词对应的对象作为与当前查询词对应的对象，进而可以获取到该匹配的历史查询词对应的对象的历史点击率及曝光数作为与当前查询词对应的对象的历史点击率及曝光数。

[0076] 如果不匹配，则前进到步骤 S203。在步骤 S203 中，判断当前查询词是否与历史查询词语义相关的近义查询词中的任何一个匹配。

[0077] 根据本申请的实施例，历史统计数据还可以包括与历史查询词语义相关的近义查询词的集合。由此可以通过体现相似搜索意图的近义查询词集合，来弥补因历史统计数据预备中的曝光数阈值过滤导致的历史点击率覆盖率不足的问题，从而使得历史点击率的预估更合理。

[0078] 进一步而言，在一个具体实施例中，历史查询词语义相关的近义查询词的集合可以是在通过网络日志数据得到历史统计数据时利用语义词库对历史查询词进行相关语义改写例如同义词、近义词替换等来得到的相应扩展结果。在其它具体实施例中，也可以基于网站访问量（Session）分析来实时获取上述与历史查询词语义相关的近义查询词集合。

[0079] 如果当前查询词与历史查询词语义相关的近义查询词中的任何一个匹配，则前进到步骤 S204。在步骤 S204 中，将匹配的近义查询词对应的历史查询词对应的对象的历史点击率和曝光数，作为当前查询词对应的搜索结果中包括的对象的历史点击率及曝光数。

[0080] 具体而言，根据本申请的实施例，近义查询词对应的对象就是与之语义相关的历史查询词对应的对象。因此，可以获取到与当前查询词匹配的近义查询词对应的历史查询词对应的对象以及该对象的历史点击率和曝光数，作为当前查询词对应的对象和该对象的历史点击率和曝光数。

[0081] 如果当前查询词与历史查询词语义相关的近义词查询中的任何一个不匹配，则前进到步骤S205。在步骤S205中，根据针对历史查询词的平均历史点击率和针对历史查询词对应的对象的平均历史点击率，计算当前查询词对应的搜索结果中包括的对象的历史点击率。

[0082] 根据本申请的实施例，历史统计数据还可以包括针对历史查询词的平均历史点击率和针对历史查询词对应的对象的平均历史点击率。进一步地，对于历史查询词的平均历史点击率和历史查询词对应的对象的平均历史点击率的取得，可以是在前述从网络日志数据中得到历史统计数据的过程中完成的，具体地可以针对每个历史查询词分别统计出每个历史查询词与所有历史对象的历史点击率并求平均得到针对历史查询词的平均历史点击率，以及可以针对每个历史对象统计出每个历史对象与所有历史查询词的历史点击率并求平均得到针对历史查询词对应对象的平均历史点击率。

[0083] 对于历史查询词对应对象的平均历史点击率和历史查询词的平均历史点击率，更优选地，可以基于前述贝叶斯浏览模型BBM得到的历史查询词对应对象的标准历史点击率sCTR来得到。

[0084] 具体而言，设历史查询词-对象集合为<Query,Offer>，令 $sCTR_i$ 表示Query与 $Offer_i$ 的标准点击率，则取 $sCTR_i$ 的平均值就可得到当前Query与所有Offer的平均点击率，就可得到相应的历史查询词的平均历史点击率qCTR，即，

[0085] $$qCTR = \frac{\sum_{i=1}^{N} sCTR_i}{N} \qquad (3)$$

[0086] 令 $sCTR_j$ 表示 $Query_j$ 与Offer的标准点击率，则取 $sCTR_j$ 的平均值就可得到当前Offer与所有Query的平均点击率，就可得到相应的历史查询词对应对象的平均历史点击率，即，

[0087] $$oCTR = \frac{\sum_{j=1}^{M} sCTR_j}{M} \qquad (4)$$

[0088] 进一步地，根据上述所取得的历史查询词的平均历史点击率和历史查询词对应对象的平均历史点击率，可以通过基于最小二乘法求解得到的拟合函数来获取针对于不常搜索对象的历史点击率hCTR。具体而言，可以从历史数据中筛选出具有hCTR、qCTR和oCTR的查询词及其对象，然后通过解目标函数公式(5)即，

[0089] $$\beta = \min_{\beta}\{||\Sigma k(hCTR_k - f(qCTR_k, ocTR_k, \beta))||_2\} \qquad (5)$$

[0090] 求解出参数β，那么不常搜索对象的历史点击率hCTR可以由拟合函数公式表示为

[0091] $$hCTR = f(qCTR, oCTR, \beta) \qquad (6)$$

[0092] 进一步地，根据本申请的一个实施例，对于公式(6)结合在实际应用中，还可以具体实施为

[0093] $$hCTR=(1-\beta)\cdot oCTR+\beta\cdot qCTR \qquad (7)$$

[0094] 由此，当出现针对不常用查询词或新出现查询词的查询请求时，可以通过历史数据统计得到的平均历史点击率，来弥补因曝光数阈值过滤而导致的历史点击率覆盖率不足的问题，从而使得历史点击率的预估更合理。

[0095] 接下来，前进到步骤S206。在步骤S206中，将给定预设值作为当前查询词对应的

搜索结果中包括的对象的曝光数。

[0096] 由于当前查询词对应对象的历史点击率，在历史数据中并没有查找到对应的历史点击率及曝光数，而此时的历史点击率是通过历史数据预估出的近似值，那么就意味着，当前查询词对应对象的曝光数在历史数据中也是不存在的，这时就需要将根据经验预设的相应曝光数值赋给当前查询词对应对象，进而为历史点击率和初步预估点击率的非线性均衡拟合做出有效贡献。

[0097] 以上结合图 2 详细描述了获得当前查询词对应对象的最终预估点击率的实现过程。下面结合一更具体实施例来对此进行更详细地说明。

[0098] 在根据本申请的一个具体实施例中(例如，搜索 mp3 播放器)，可以将预先通过基于曝光数的阈值过滤从网络日志数据中所获取的历史统计数据存储于历史点击率索引表中，以备在接收到查询请求时调用。该历史点击率索引表可以进一步包括：查询词和对应对象集合索引子表，其中包含历史查询词、与历史查询词对应的对象、对象的历史点击率、对象的曝光数等项目(例如，查询词为“mp3”)；以及扩展索引子表，是从查询词和对应对象集合索引子表衍生的，其中包括历史查询词、历史查询词的近义查询词集合、历史查询词对应的对象、对象的历史点击率、对象的曝光数等项目(例如，长尾查询词“love style red mp3”的近义查询词为“mp3”)。在一个优选实施例中，扩展索引子表中还可以包括针对历史查询词的平均历史点击率和针对历史查询词对应的对象的平均历史点击率。

[0099] 假设用 Query 表示当前查询词，用 Offer 表示当前查询词对应对象，用 hCTR 表示查询词 - 对象集合 <Query,Offer> 的对象的历史点击率，用 pv 表示查询词 - 对象集合 <Query,Offer> 的对象的曝光数，用 eCTR 表示查询词 - 对象集合 <Query,Offer> 的初步预估点击率。

[0100] 当接收到包含当前查询词的查询请求时，在查询词和对应对象集合索引子表中查找是否有当前查询词所对应的历史查询词(对应于图 2 的步骤 S201)。如果有，则可以确定出所述当前查询词对应的对象以及对象的历史点击率及其曝光数(对应于图 2 的步骤 S202)。

[0101] 如果没有在查询词和对应对象集合索引子表中查找到与当前查询词对应的历史查询词，则在扩展索引子表的近义查询词集合中查找是否有当前查询词(对应于图 2 的步骤 S203)。如果有，则在扩展索引子表中查找与该近义查询词集合对应的历史查询词，从而找到与该历史查询词对应的对象以及对象的历史点击率及曝光数作为当前查询词对应的对象以及该对象的历史点击率及曝光数(对应于图 2 的步骤 S204)。这部分是考虑到当所接收到的查询词新词或是极少被应用的词，即不常搜索对象、新出现对象，那么这时就需要进入扩展索引子表中去查找。

[0102] 当在扩展索引子表中也没有查找到当前查询词对应对象时，此时就基于针对历史查询词的平均历史点击率和针对历史查询词对应的对象的平均历史点击率来为当前查询词对应对象预估出历史点击率 hCTR (对应于图 2 的步骤 S205)，并且将给定预设值作为当前查询词对应对象的曝光数(对应于图 2 的步骤 S206)。

[0103] 在一种极端情况下，当前查询词 - 对象集合 <Query,Offer> 中的 Query 和 Offer 属于新出现对象，那么就无法取得 qCTR 和 oCTR，则就取默认值 $qCTR_0$ 和 $oCTR_0$ 来预估出历史点击率 hCTR，而 $qCTR_0$ 和 $oCTR_0$ 的取值可以是预先设置的平均值。

[0104] 通过上述历史点击率预估，能够弥补因曝光数阈值过滤而导致的历史点击率覆盖率不足，同时对新出现的当前查询词－对象集合<Query,Offer>也能给出近似的历史点击率，从而可以确保多策略拟合均衡对不常用对象或新出现对象都生效。

[0105] 至此结合图1-图3详细描述了根据本申请实施例的用于对搜索结果进行排序的方法，其中通过针对对象的历史点击率、曝光数和初步预估点击率这三个参数进行非线性拟合来得到该对象的最终预估点击率，可以综合地表征用户的点击行为，实现更准确的点击率预估。另外，在预估对象的点击率时，可以基于历史数据的曝光数信息更合理地利用历史数据信息，提高点击率预估效率。尤其是，针对曝光数较低或为零的对象也提供合理的点击率预估值，由此对于当前查询词对应的各个对象来说，不管它是曝光率高的对象，还是曝光率低的对象，都可以得到更合理的点击率预估，由此可以在更合理地利用历史数据资源信息的前提下实现更准确且更高效的点击率预估和搜索结果的排序，从而使得搜索结果更符合用户需求，方便用户的使用。

[0106] 与上述用于对搜索结果进行排序的方法类似，本申请的实施例还提供相应的装置。

[0107] 参考图4，图4是根据本申请实施例的对搜索结果进行排序的装置400的结构框图。如图4所示，装置400可以包括历史预估模块401、初步预估模块402、最终预估模块403和排序模块404。

[0108] 具体而言，历史预估模块401可以用于当接收到包含当前查询词的查询请求时，根据历史统计数据获得当前查询词对应的搜索结果中包括的对象的历史点击率及曝光数。

[0109] 根据本申请的实施例，该历史预估模块401还可以包括：第一子模块(图中未示出)，用于当前查询词与历史查询词中的一个匹配时，将历史查询词对应的对象的历史点击率和曝光数，作为当前查询词对应的搜索结果中包括的对象的历史点击率及曝光数。

[0110] 根据本申请的实施例，该历史预估模块401还可以包括：第二子模块(图中未示出)，用于当前查询词与历史查询词中的任何一个均不匹配时，判断当前查询词是否与历史查询词语义相关的近义查询词中的任何一个匹配；如果匹配，则将匹配的近义查询词对应的历史查询词对应的对象的历史点击率和曝光数，作为当前查询词对应的搜索结果中包括的对象的历史点击率及曝光数；如果不匹配，则根据针对历史查询词的平均历史点击率和针对历史查询词对应的对象的平均历史点击率，计算当前查询词对应的搜索结果中包括的对象的历史点击率，并且将给定预设值作为当前查询词对应的搜索结果中包括的对象的曝光数。

[0111] 另外，初步预估模块402可以用于根据当前查询词和对应对象的特征得到对象的初步预估点击率。

[0112] 根据本申请的实施例，初步预估模块402还可以包括(图中未示出)：第一提取子模块，用于基于历史统计数据提取历史查询词和对应对象的特征；第一计算子模块，用于得到历史查询词和对应对象的特征的特征权重；以及第二计算子模块，用于根据特征权重建立预估模型，预估模型用于根据当前查询词和对应对象的特征预估当前查询词对应对象的初步预估点击率。

[0113] 根据本申请的实施例，初步预估模块402还可以包括(图中未示出)：第二提取子模块，用于提取当前查询词和对应对象的特征；以及初步预估子模块，用于根据当前查询词和

对应对象的特征，通过预估模型得到当前查询词对应对象的初步预估点击率。

[0114] 根据本申请的具体实施例，所述特征包括查询词的文本特征、查询词对应对象的标题和 / 或属性特征、以及查询词与对应对象的相关性特征。

[0115] 此外，最终预估模块 403 可以用于根据历史点击率、曝光数和初步预估点击率，通过非线性拟合得到对象的最终预估点击率。

[0116] 根据本申请的实施例，非线性拟合是指数型非线性拟合。

[0117] 另外，排序模块 404 可以用于根据最终预估点击率对对象进行排序。

[0118] 根据本申请的实施例，装置 400 还可以包括（图中未示出）：获取模块，用于从网络日志数据中获得曝光数大于或等于预定阈值的历史查询词和历史查询词对应的对象；以及统计模块，用于通过统计得到历史统计数据，历史统计数据包括历史查询词对应的对象的历史点击率和曝光数。

[0119] 至此描述了根据本申请实施例的用于对搜索结果进行排序的装置。与上述方法类似，根据该装置，同样可以在更合理地利用历史数据资源信息的前提下实现更准确且更高效的点击率预估和搜索结果的排序，从而使得搜索结果更符合用户需求，方便用户的使用。

[0120] 由于上述用于对搜索结果进行排序的装置的处理与上述结合图 1 至图 3 描述的用于对搜索结果进行排序的方法的处理是对应的，因此关于其具体细节，可以参考之前描述的用于对搜索结果进行排序的方法，这里不再赘述。

[0121] 在一个典型的配置中，计算设备包括一个或多个处理器 (CPU)、输入 / 输出接口、网络接口和内存。

[0122] 内存可能包括计算机可读介质中的非永久性存储器，随机存取存储器 (RAM) 和 / 或非易失性内存等形式，如只读存储器 (ROM) 或闪存 (flash RAM)。内存是计算机可读介质的示例。

[0123] 计算机可读介质包括永久性和非永久性、可移动和非可移动媒体可以由任何方法或技术来实现信息存储。信息可以是计算机可读指令、数据结构、程序的模块或其他数据。计算机的存储介质的例子包括，但不限于相变内存 (PRAM)、静态随机存取存储器 (SRAM)、动态随机存取存储器 (DRAM)、其他类型的随机存取存储器 (RAM)、只读存储器 (ROM)、电可擦除可编程只读存储器 (EEPROM)、快闪记忆体或其他内存技术、只读光盘只读存储器 (CD-ROM)、数字多功能光盘 (DVD) 或其他光学存储、磁盒式磁带，磁带磁磁盘存储或其他磁性存储设备或任何其他非传输介质，可用于存储可以被计算设备访问的信息。按照本文中的界定，计算机可读介质不包括暂存电脑可读媒体 (transitory media)，如调制的数据信号和载波。

[0124] 还需要说明的是，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、商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、商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、商品或者设备中还存在另外的相同要素。

[0125] 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、设备、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。

[0126] 以上所述仅为本申请的实施例而已,并不用于限制本申请,对于本领域的技术人员来说,本申请可以有各种更改和变化。凡在本申请的精神和原则之内,所作的任何修改、等同替换、改进等,均应包含在本申请的权利要求范围之内。

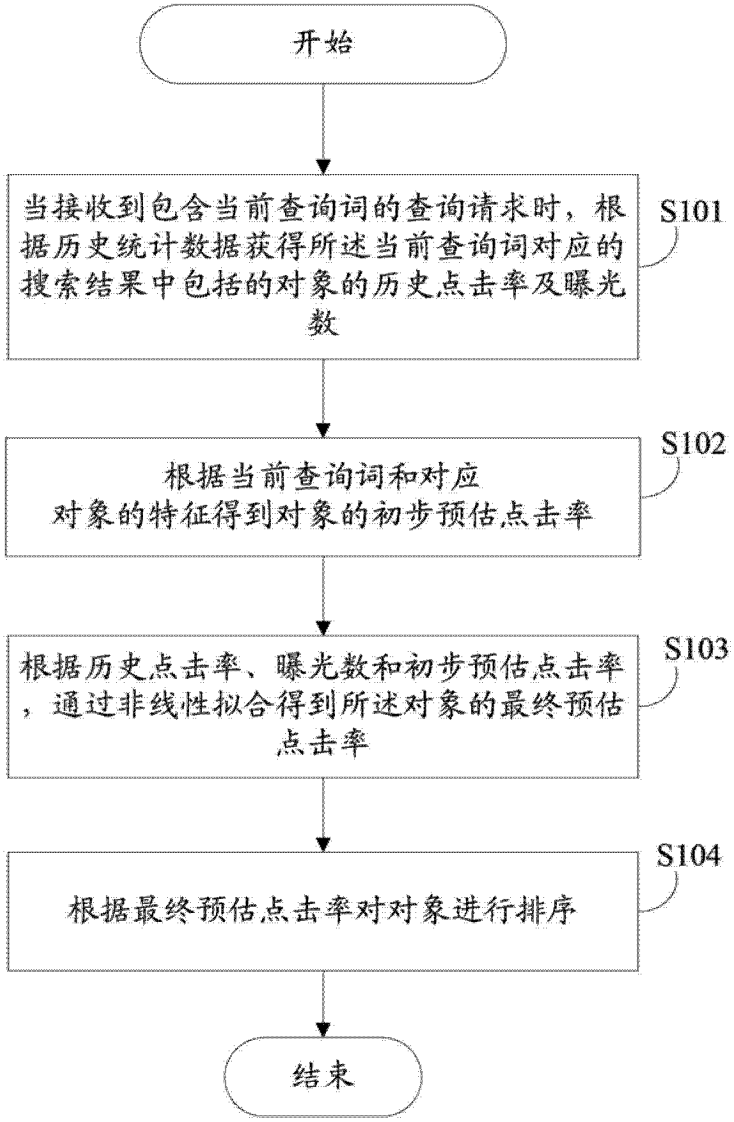

开始
当接收到包含当前查询词的查询请求时，根据历史统计数据获得所述当前查询词对应的搜索结果中包括的对象的历史点击率及曝光数
S101
根据当前查询词和对应对象的特征得到对象的初步预估点击率
S102
根据历史点击率、曝光数和初步预估点击率，通过非线性拟合得到所述对象的最终预估点击率
S103
根据最终预估点击率对对象进行排序
S104
结束

图 1

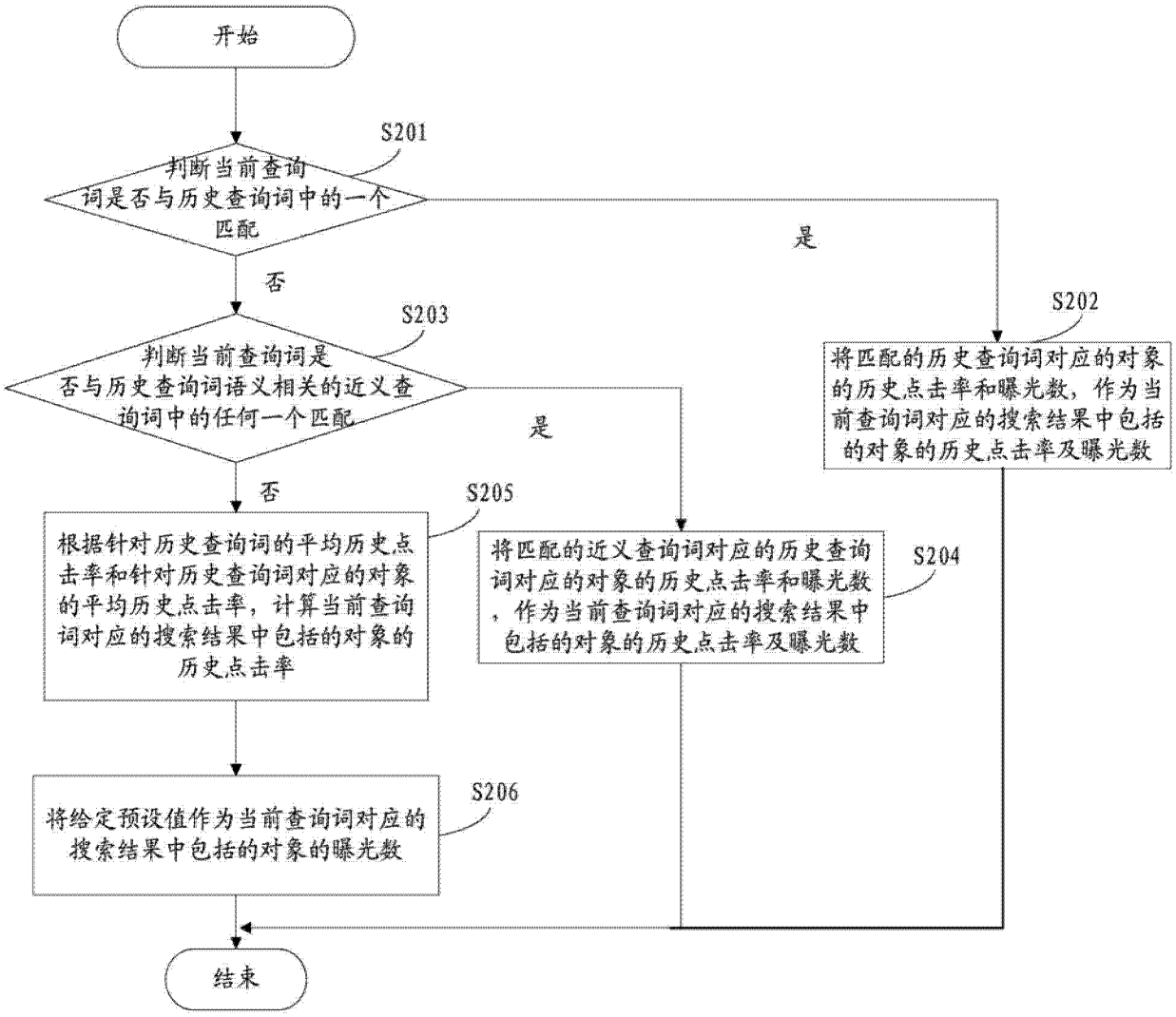

开始
S201
判断当前查询词是否与历史查询词中的一个匹配
是
否
S202
将匹配的历史查询词对应的对象的历史点击率和曝光数，作为当前查询词对应的搜索结果中包括的对象的历史点击率及曝光数
S203
判断当前查询词是否与历史查询词语义相关的近义查询词中的任何一个匹配
是
否
S204
将匹配的近义查询词对应的历史查询词对应的对象的历史点击率和曝光数，作为当前查询词对应的搜索结果中包括的对象的历史点击率及曝光数
S205
根据针对历史查询词的平均历史点击率和针对历史查询词对应的对象的平均历史点击率，计算当前查询词对应的搜索结果中包括的对象的历史点击率
S206
将给定预设值作为当前查询词对应的搜索结果中包括的对象的曝光数
结束

图 2

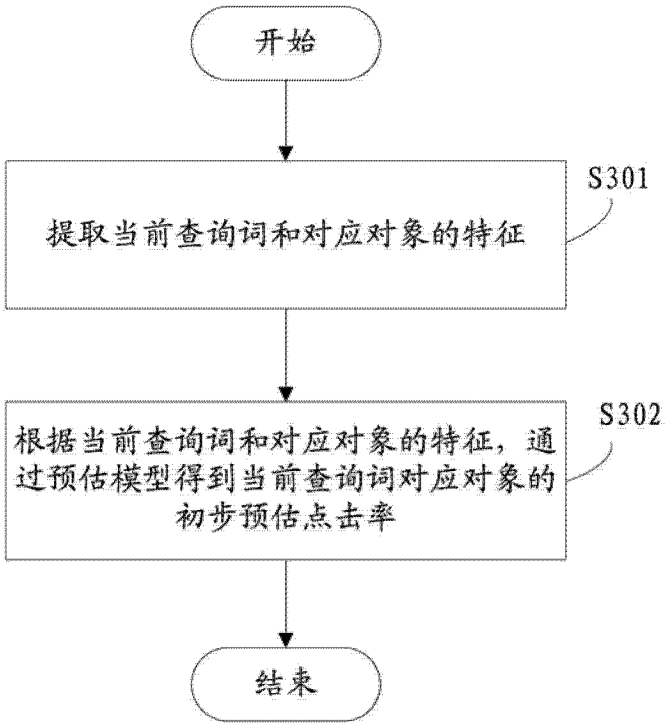

开始
提取当前查询词和对应对象的特征
S301
根据当前查询词和对应对象的特征，通过预估模型得到当前查询词对应对象的初步预估点击率
S302
结束

图 3

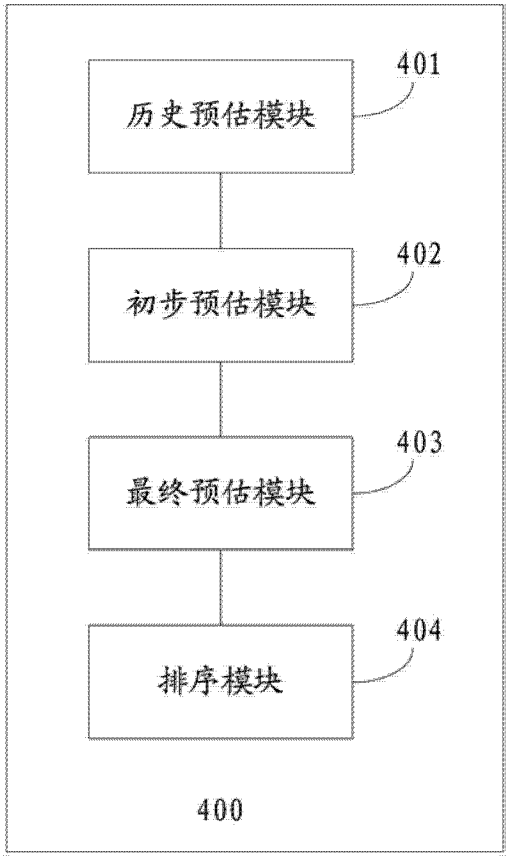

历史预估模块
401
初步预估模块
402
最终预估模块
403
排序模块
404
400

图 4

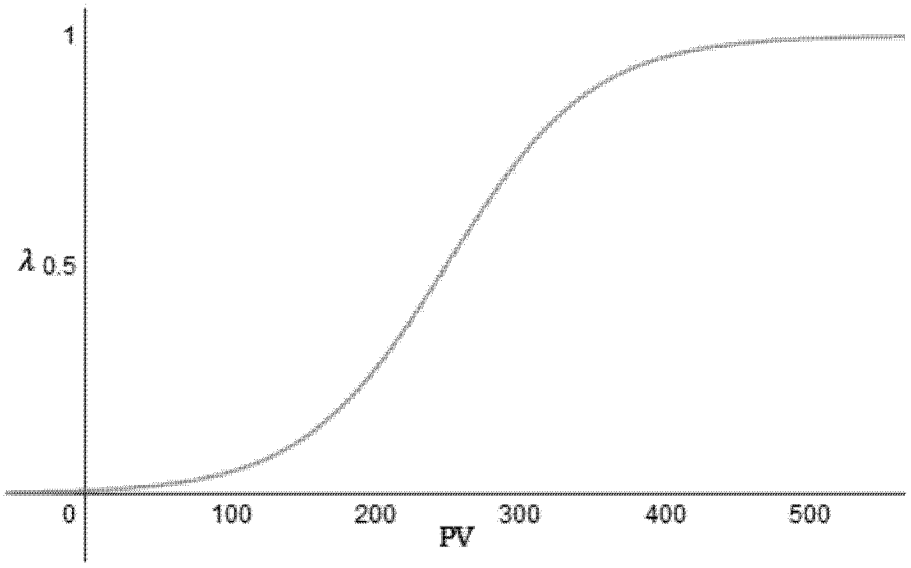

1
λ 0.5
0
100
200
300
400
500
PV

图 5