(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 102419755 B
(45)授权公告日 2013.04.24

(21)申请号 201010299100.0
(22)申请日 2010.09.28
(73)专利权人 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 金华兴 郑伟 黄鹏 杨旭 林锋 冯炯 张勤
(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)

(56)对比文件
CN 101334773 A,2008.12.31,
CN 1211769 A,1999.03.24,

审查员 高霞

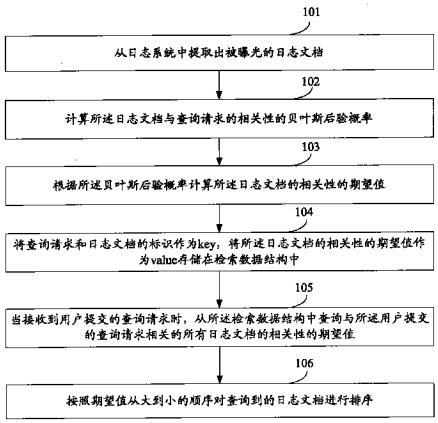
权利要求书3页 说明书9页 附图7页

(54)发明名称

一种搜索结果的排序方法和装置

(57)摘要

本申请实施例公开了一种搜索结果的排序方法和装置。其中，所述方法包括：从日志系统中提取出被曝光的日志文档；计算所述日志文档与查询请求的相关性的贝叶斯后验概率；根据所述贝叶斯后验概率计算所述日志文档与查询请求的相关性的期望值；将查询请求和日志文档的标识作为键，将所述日志文档与查询请求的相关性的期望值作为值存储在检索数据结构中；当接收到用户提交的查询请求时，从所述检索数据结构中查询与所述用户提交的查询请求相关的所有日志文档与查询请求的相关性的期望值；按照期望值从大到小的顺序对查询到的日志文档进行排序。根据本申请实施例，可以减少搜索过程对于搜索引擎服务器的消耗，并节省搜索引擎服务器的系统资源。



CN 102419755 B

1. 一种搜索结果的排序方法，其特征在于，包括：

从日志系统中提取出被曝光的日志文档；

计算所述日志文档与查询请求的相关性的贝叶斯后验概率，所述计算日志文档与查询请求的相关性的贝叶斯后验概率具体为：根据用户是否点击对应日志文档的曝光搜索结果与用户是否浏览到对应该日志文档的曝光搜索结果，以及该日志文档与用户查询请求的相关性程度有关，用户是否浏览对应下面的日志文档的曝光搜索结果与对应之前日志文档的曝光搜索结果的点击情况有关，当先验分布在 [0,1] 上服从均匀分布时，计算用户点击对应日志文档的曝光搜索结果后，该日志文档与查询请求的相关性的联合后验分布；

根据所述贝叶斯后验概率计算所述日志文档与查询请求的相关性的期望值；

将查询请求和日志文档的标识作为键，将所述日志文档与查询请求的相关性的期望值作为值存储在检索数据结构中；

当接收到用户提交的查询请求时，从所述检索数据结构中查询与所述用户提交的查询请求相关的所有日志文档与查询请求的相关性的期望值；

按照期望值从大到小的顺序对查询到的日志文档进行排序。

2. 根据权利要求 1 所述的排序方法，其特征在于，在所述根据贝叶斯后验概率计算日志文档与查询请求的相关性的期望值之前，还包括：

根据全局参数对日志文档进行过滤，使全局参数小于对应的预设阈值的日志文档被过滤；

则所述根据贝叶斯后验概率计算日志文档与查询请求的相关性的期望值为：根据贝叶斯后验概率计算过滤后的日志文档与查询请求的相关性的期望值。

3. 根据权利要求 2 所述的排序方法，其特征在于，所述根据全局参数对日志文档进行过滤，使全局参数小于对应的预设阈值的日志文档被过滤包括：

从提取出的被曝光的日志文档中筛选出被曝光一次且没有被点击的日志文档；

从筛选出的日志文档中，按照过滤条件公式 $\beta_{r,d} < \frac{1/E_{th}}{\frac{1}{3}/\frac{1}{2}E_{th}}$ 过滤全局参数小于对应的预设阈值的日志文档，其中，$\beta_{r,d}$ 为全局参数，$\hat{\beta}_{r,d} = \min\{1, \frac{2N_{r,d}}{N_{r,d} + \tilde{N}_{r,d}}\}$，$N_{r,d}$ 为在被筛选出的日志文档所在的同一个点击序列中，位置 r 处的日志文档和位置 r+d 处的日志文档都被点击，位置 r 到 r+d 之间的日志文档没有被点击在所有情况下发生的次数；$\tilde{N}_{r,d}$ 为在被筛选出的日志文档所在的同一个点击序列中，位置 r 处的日志文档被点击，位置 r+d 处的日志文档没有被点击，位置 r 到 r+d 之间的日志文档没有被点击在所有情况下发生的次数；r 的取值范围为小于或等于 M-1 的所有自然数，d 的取值范围为小于或等于 M-r 的所有整数，M 表示被筛选出的日志文档所在的同一点击序列中所有日志文档的总数，E_{th} 为与相关性的期望值对应的预设阈值。

4. 根据权利要求 1 所述的排序方法，其特征在于，在所述将查询请求和日志文档的标识作为键，将所述日志文档的期望值作为值存储在检索数据结构中之前，还包括：

根据日志文档与查询请求的相关性的期望值或者方差对日志文档进行过滤，使期望值

或者方差等于对应的预设阈值的日志文档被过滤；

则所述将查询请求和日志文档的标识作为键，将所述日志文档的期望值作为值存储在检索数据结构中为：将查询请求和日志文档的标识作为键，将过滤后的日志文档的期望值作为值存储在检索数据结构中。

5. 根据权利要求 1 所述的排序方法，其特征在于，在所述将查询请求和日志文档的标识作为键，将所述日志文档与查询请求的相关性的期望值作为值存储在检索数据结构中之后，还包括：

对所述检索数据结构进行校验；

则所述从检索数据结构中查询与所述用户提交的查询请求相关的所有日志文档与查询请求的相关性的期望值为：从通过校验的检索数据结构中查询与所述用户提交的查询请求相关的所有日志文档与查询请求的相关性的期望值。

6. 一种搜索结果的排序装置，其特征在于，包括：

提取模块，用于从日志系统中提取出被曝光的日志文档；

概率计算模块，用于计算所述日志文档与查询请求的相关性的贝叶斯后 验概率，所述计算日志文档与查询请求的相关性的贝叶斯后验概率具体为：根据用户是否点击对应日志文档的曝光搜索结果与用户是否浏览到对应该日志文档的曝光搜索结果，以及该日志文档与用户查询请求的相关性程度有关，用户是否浏览对应下面的日志文档的曝光搜索结果与对应之前日志文档的曝光搜索结果的点击情况有关，当先验分布在 [0,1] 上服从均匀分布时，计算用户点击对应日志文档的曝光搜索结果后，该日志文档与查询请求的相关性的联合后验分布；

期望值计算模块，用于根据所述贝叶斯后验概率计算所述日志文档与查询请求的相关性的期望值；

索引建立模块，用于将查询请求和日志文档的标识作为键，将所述日志文档与查询请求的相关性的期望值作为值存储在检索数据结构中；

检索模块，用于当接收到用户提交的查询请求时，从所述检索数据结构中查询与所述用户提交的查询请求相关的所有日志文档与查询请求的相关性的期望值；

排序模块，用于按照期望值从大到小的顺序对查询到的日志文档进行排序。

7. 根据权利要求 6 所述的排序装置，其特征在于，还包括第一过滤模块，用于在根据所述贝叶斯后验概率计算日志文档与查询请求的相关性的期望值之前，根据全局参数对日志文档进行过滤，使全局参数小于对应的预设阈值的日志文档被过滤；

则所述期望值计算模块，用于根据贝叶斯后验概率计算过滤后的日志文档与查询请求的相关性的期望值。

8. 根据权利要求 7 所述的排序装置，其特征在于，所述第一过滤模块包括：

筛选子模块，用于从提取出的被曝光的日志文档中筛选出被曝光一次且没有被点击的日志文档；

过滤子模块，从筛选出的日志文档中，按照过滤条件公式 $\beta_{r,d} < \frac{1/E_{th}}{\frac{1}{3}/\frac{1}{2}E_{th}}$ 过滤掉全局参

数小于对应的预设阈值的日志文档，其中，$\beta_{r,d}$ 为全局参数，$\hat{\beta}_{r,d}=\min\{1,\frac{2N_{r,d}}{N_{r,d}+\tilde{N}_{r,d}}\}$，$N_{r,d}$ 为在被筛选出的日志文档所在的同一个点击序列中，位置 r 处的日志文档和位置 r+d 处的日志文档都被点击，位置 r 到 r+d 之间的日志文档没有被点击在所有情况下发生的次数；$\tilde{N}_{r,d}$ 为在被筛选出的日志文档所在的同一个点击序列中，位置 r 处的日志文档被点击，位置 r+d 处的日志文档没有被点击，位置 r 到 r+d 之间的日志文档没有被点击在所有情况下发生的次数；r 的取值范围为小于或等于 M-1 的所有自然数，d 的取值范围为小于或等于 M-r 的所有整数，M 表示被筛选出的日志文档所在的同一点击序列中所有日志文档的总数，E_{th} 为与相关性的期望值对应的预设阈值。

9. 根据权利要求 6 所述的排序装置，其特征在于，还包括第二过滤模块，用于在所述将查询请求和日志文档的标识作为键，将所述日志文档的期望值作为值存储在检索数据结构中之前，根据日志文档与查询请求的相关性的期望值或者方差对日志文档进行过滤，使期望值或者方差等于对应的预设阈值的日志文档被过滤，

则所述索引建立模块，用于将查询请求和日志文档的标识作为键，将过滤后的日志文档的期望值作为值存储在检索数据结构中。

10. 根据权利要求 6 所述的排序装置，其特征在于，还包括校验模块，用于在所述将查询请求和日志文档的标识作为键，将所述日志文档与查询请求的相关性的期望值作为值存储在检索数据结构中之后，对所述检索数据结构进行校验，

则所述检索模块，用于从通过校验的检索数据结构中查询与所述用户提交的查询请求相关的所有日志文档与查询请求的相关性的期望值。

一种搜索结果的排序方法和装置

技术领域

[0001] 本申请涉及通信和计算机技术领域，特别是涉及一种搜索结果的排序方法和装置。

背景技术

[0002] 当用户向搜索引擎提交一个查询请求后，搜索引擎会检索到大量与用户的查询请求相关的信息。同时，搜索引擎会根据每个信息与查询请求的相关性程度，对信息进行排序，以便用户可以快速地通过搜索引擎查找到最想要的信息。

[0003] 目前，搜索引擎大多利用 CTR(Click-Through-Rate，点击到达率）反映每个信息与用户查询请求的相关性程度，其中，CTR 为信息被点击的次数与信息被曝光次数的商。当搜索引擎计算得到每个信息的 CTR 后，在搜索结果列表中，按照 CTR 从大到小的顺序对信息进行排序。

[0004] 但是，发明人在研究中发现，一个信息与用户查询请求的相关性程度往往与该信息在搜索结果列表中的位置和在搜索结果列表中的点击顺序有关。例如，在同一个搜索结果展现页面中，即使相关性相同，不同位置的信息的 CTR 也会不一样。或者，在先被用户点击的信息，会影响位于其后面的信息的被点击概率。

[0005] 然而，现有技术中在对搜索结果进行排序的过程中一方面只考虑到了信息被点击的次数和信息被曝光的次数，而没有考虑到在整个排序过程中，信息在搜索结果列表中的位置因素和在搜索结果列表中的被点击顺序的因素，使排序时所依据的相关性分数存在较大偏差，在多数情况下把用户想要获得的信息排在了搜索结果列表的后面，最终导致对搜索结果的排序效果差。另一方面，当对搜索结果的排序效果较差的时，用户通常需要进一步浏览和点击更多的信息才能获得最想要的信息，而用户在网站上“盲目地”进行大范围的浏览和点击的过程时，势必会增加网络系统，特别是搜索引擎服务器的负载，降低了网络系统的利用率。从而增加了搜索过程对于搜索引擎服务器的消耗，同时，也浪费了搜索引擎服务器的系统资源。

发明内容

[0006] 为了解决上述技术问题，本申请实施例提供了一种搜索排序方法和装置，以减少搜索过程对于搜索引擎服务器的消耗，并节省搜索引擎服务器的系统资源。

[0007] 本申请实施例公开了如下技术方案：

[0008] 一种搜索结果的排序方法，包括：从日志系统中提取出被曝光的日志文档；计算所述日志文档与查询请求的相关性的贝叶斯后验概率；根据所述贝叶斯后验概率计算所述日志文档与查询请求的相关性的期望值；将查询请求和日志文档的标识作为键，将所述日志文档与查询请求的相关性的期望值作为值存储在检索数据结构中；当接收到用户提交的查询请求时，从所述检索数据结构中查询与所述用户提交的查询请求相关的所有日志文档与查询请求的相关性的期望值；按照期望值从大到小的顺序对查询到的日志文档进行排

序。

[0009] 一种搜索结果的排序装置，包括：提取模块，用于从日志系统中提取出被曝光的日志文档；概率计算模块，用于计算所述日志文档与查询请求的相关性的贝叶斯后验概率；期望值计算模块，用于根据所述贝叶斯后验概率计算所述日志文档与查询请求的相关性的期望值；索引建立模块，用于将查询请求和日志文档的标识作为键，将所述日志文档与查询请求的相关性的期望值作为值存储在检索数据结构中；检索模块，用于当接收到用户提交的查询请求时，从所述检索数据结构中查询与所述用户提交的查询请求相关的所有日志文档与查询请求的相关性的期望值；排序模块，用于按照期望值从大到小的顺序对查询到的日志文档进行排序。

[0010] 由上述实施例可以看出，本申请在整个排序过程中，考虑到了信息在搜索结果列表中的位置因素和在搜索结果列表中的被点击顺序的因素，即，基于贝叶斯后验概率计算日志文档与查询请求的相关性的期望值，当从检索数据结构中查询到与用户提交的查询请求相关的所有日志文档与查询请求的相关性的期望值后，按照相关性的期望值从大到小的顺序对日志文档进行排序，从而使排序时所依据的相关性更好。同时，也使用户减少浏览的时间和点击的次数，快速地获得最想要的信息，减少搜索过程对于搜索引擎服务器的消耗，并节省搜索引擎服务器的系统资源。

附图说明

[0011] 为了更清楚地说明本申请实施例或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍，显而易见地，对于本领域普通技术人员而言，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。

[0012] 图 1 为本申请一种搜索结果的排序方法的一个实施例的流程图；

[0013] 图 2 为本申请一种构建的概率模型结构示意图；

[0014] 图 3 为本申请一种搜索结构的排序方法的另一个实施例的流程图；

[0015] 图 4 为本申请一种搜索排序系统的结构示意图；

[0016] 图 5 为本申请一种搜索结果的排序装置的一个实施例的结构示意图；

[0017] 图 6 为本申请一种搜索结果的排序装置的另一个实施例的结构示意图；

[0018] 图 7 为本申请一种搜索结果的排序装置的另一个实施例的结构示意图；

[0019] 图 8 为本申请一种搜索结果的排序装置的另一个实施例的结构示意图。

具体实施方式

[0020] 为使本申请的上述目的、特征和优点能够更加明显易懂，下面结合附图对本申请实施例进行详细描述。

[0021] 实施例一

[0022] 请参阅图 1，其为本申请一种搜索结果的排序方法的一个实施例的流程图，该方法包括以下步骤：

[0023] 步骤 101：从日志系统中提取出被曝光的日志文档；

[0024] 其中，在日志系统中以增量更新的方式保存有历史被曝光的日志文档和当天被曝光的日志文档。从日志系统中提取出在日志系统保存的所有被曝光的日志文档。用户通过

搜索引擎进行查询时，与查询请求相关的搜索结果会展示给用户，该展示给用户的搜索结果即为被曝光的搜索结果，该被曝光的搜索结果以日志文档的形式保存在日志系统中。

[0025] 步骤 102：计算所述日志文档与查询请求的相关性的贝叶斯后验概率；

[0026] 其中，发明人在研究中发现，一个信息与用户查询请求的相关性程度往往与该信息在搜索结果列表中的位置和在搜索结果列表中的被点击顺序有关。例如，当用户面对一个搜索结果展示页面时，一般会从上至下逐一浏览日志文档在网页中所展示的展示信息，如果发现某一个日志文档的展示信息符合自身的搜索意图，就会点击该展示信息并查看详细内容。当查看完该展示信息的详细内容后，可能会继续浏览下面的日志文档的展示信息，也有可能因为查看到了需要的内容而结束浏览。由此可见，用户是否点击某个日志文档主要取决于用户是否浏览到该日志文档，以及该日志文档与用户查询请求的相关性程度。而用户是否继续浏览下面的日志文档主要取决于之前日志文档的点击情况。

[0027] 基于上述情况，建立一个数学模型。如图 2 所示，其为本申请一种构建的概率图模型结构示意图。图 2 中的每个节点代表一个随机变量，S 表示日志文档与用户查询请求的相关性，E 表示用户是否看到日志文档，C 表示用户是否点击日志文档，下标表示日志文档在一个搜索结果展示页面中的位置，M 表示一个搜索结果展示页面中的日志文档总数目。根据前述分析可知，用户是否点击某个日志文档与用户是否浏览到该日志文档，以及该日志文档与用户查询请求的相关性程度有关，用户是否浏览下面的日志文档与之前日志文档的点击情况有关，因此，从图 2 中可以看出，在该模型中，例如，S1 和 E1 分别指向 C1，表示用户是否点击一个日志文档 C1 与用户是否浏览到该日志文档 E1，以及该日志文档与用户查询请求的相关性程度 S1 有关，而 C1 指向 E2，表示用户是否继续浏览下面的日志文档 E2 与之前日志文档 C1 的点击情况有关。

[0028] 根据如图 2 所示的数学模型的概率推论，当先验分布在 [0，1] 上服从均匀分布时，其中，0 和 1 表示实数值区间的两个端点，即，相关性变量的先验分布是从 0 到 1 的实数值区间上的均匀分布，在用户点击日志文档 C1、C2... 和 CN 的情况下，日志文档与查询请求的相关性的联合后验分布计算公式为：

[0029] $$p(R \mid c^{1:N}) = \frac{1}{z}\prod_{j=1}^{N} R_j^{N_j} \prod_{(r,d)\in T} (1-\beta_{r,d}R_j)^{\tilde{N}_{j,r,d}}$$

[0030] 其中，上述公式中的 R_j 表示日志文档 j 与用户查询请求的相关性随机变量，N_j 表示日志文档 j 被点击的总次数，$\tilde{N}_{j,r,d}$ 表示日志文档 j 位于 r+d 处且没有被点击，位置 r 处的日志文档被点击，位置 r 到 r+d 之间的日志文档没有被点击在所有情况下的发生次数，T 表示所有 (r，d) 的可能取值，$\beta_{r,d}$ 为一个全局参数，N 表示从日志系统中提取的日志文档的总数目，z 表示归一化系数。

[0031] 从上述联合分布的形式可以看出，联合分布可以分解为单个文档分布的乘积。因此，单个文档 j 的相关性后验分布计算公式为：

[0032] $$p(R_j \mid c^{1:N}) = \frac{1}{z} R_j^{N_j} \prod_{0\le r\le M-1, 1\le d\le M-r} (1-\beta_{r,d}R_j)^{\tilde{N}_{j,r,d}}$$

[0033] $$z = \int_0^1 R_j^{N_j} \prod_{0\le r\le M-1, 1\le d\le M-r} (1-\beta_{r,d}R_j)^{\tilde{N}_{j,r,d}} dR_j$$

[0034] 其中，$\beta_{r,d}$为一个全局参数，其估计值$\hat{\beta}_{r,d}=\min\{1,\frac{2N_{r,d}}{N_{r,d}+\tilde{N}_{r,d}}\}$，$N_{r,d}$为在日志文档j所在的同一个点击序列中，位置r处的日志文档和位置r+d处的日志文档都被点击，位置r到r+d之间的日志文档没有被点击在所有情况下的发生次数；$\tilde{N}_{r,d}$为在日志文档j所在的同一个点击序列中，位置r处的日志文档被点击，位置r+d处的日志文档没有被点击，位置r到r+d之间的日志文档没有被点击在所有情况下的发生次数；r的取值范围为小于或等于M-1的所有自然数，d的取值范围为小于或等于M-r的所有整数，M表示日志文档j所在的同一点击序列中所有日志文档的总数。其中，所述同一个点击序列即为针对同一个用户查询请求而获得的所有查询结果构成的序列。例如，针对用户查询请求“MP3”，有100个日志文档为该用户查询请求的所有查询结果，则100个日志文档构成针对“MP3”的同一个点击序列。

[0035] 步骤103：根据所述贝叶斯后验概率计算所述日志文档与查询请求的相关性的期望值；

[0036] 其中，根据上述步骤得到的贝叶斯后验概率计算日志文档与查询请求的相关性的期望值，期望值的计算公式为：

[0037] $$E(R_j)=\int_0^1 R_j p(R_j \mid C^{1:N})dR_j$$

[0038] 需要说明的是，由于计算期望值的开销比较大，会消耗较大的系统资源。为了避免计算期望值所带来的资源消耗，在本步骤计算日志文档的相关性的期望值之前，对日志文档进行过滤。其中，有一些日志文档与用户查询请求的相关性不好也不坏。在实际应用中，为了节省空间和时间，需要对这种相关性不好也不坏的日志文档进行过滤。

[0039] 通常，当日志文档与用户查询请求的相关性的期望值为0.5时，表示相关性不好也不坏，因此，可以过滤掉与用于查询请求的相关性的期望值为0.5的日志文档。而本申请需要提供一种在计算期望值之前，就可以过滤掉相关性不好也不坏的日志文档。

[0040] 优选的，在所述根据贝叶斯后验概率计算日志文档与查询请求的相关性的期望值之前，还包括：根据全局参数对日志文档进行过滤，使全局参数小于对应的预设阈值的日志文档被过滤。其中，全局参数指的是和用户查询请求无关的一个参数，反应的是用户对一个搜索引擎的评价的一组指标参数。在给定的情况下，全局参数与日志文档与查询请求的相关性的期望值通过概率分布函数建立起了对应关系，按照期望值的阈值过滤等价于全局参数的阈值过滤，这种关系的推导需要预先做出解析，然后在系统初始化阶段根据期望值的阈值计算全局参数的阈值，在日志处理阶段就可以根据全局参数的阈值进行过滤。

[0041] 例如，从提取出的被曝光的日志文档中筛选出被曝光一次且没有被点击的日志文档；从筛选出的日志文档中按照过滤条件公式$\beta_{r,d}<\frac{1/E_{th}}{\frac{1}{3}/\frac{1}{2}E_{th}}$过滤全局参数小于对应的预设阈值的日志文档，其中，$\beta_{r,d}$为全局参数，$\hat{\beta}_{r,d}=\min\{1,\frac{2N_{r,d}}{N_{r,d}+\tilde{N}_{r,d}}\}$，$N_{r,d}$为在被筛选出的日志文档所在的同一个点击序列中，位置r处的日志文档和位置r+d处的日志文档都被点击，

位置 r 到 r+d 之间的日志文档没有被点击在所有情况下的发生次数；例如，为了便于描述，以一个包含 5 个日志文档的同一个点击序列为例来说明，已知位置排列第一、第三和第五的日志文档被点击，位置排列第二和第四的日志文档没有被点击。则位置排列第一和第三的日志文档都被点击，而位置排列位于第二和第二之间，即位置排列位于第二的日志文档没有被点击，该情况的发生次数为 1 次，同时，位置排列第三和第五的日志文档都被点击，而位置排列位于第三和第五之间，即位置排列位于第四的日志文档没有被点击，该情况的发生次数为 1 次。因此，在以上的同一个点击序列中，$N_{r,d}$ 为 2。

[0042] $\tilde{N}_{r,d}$ 为在被筛选出的日志文档所在的同一个点击序列中，位置 r 处的日志文档被点击，位置 r+d 处的日志文档没有被点击，位置 r 到 r+d 之间的日志文档没有被点击在所有情况下的发生次数；r 的取值范围为小于或等于 M-1 的所有自然数，d 的取值范围为小于或等于 M-r 的所有整数，M 表示被筛选出的日志文档所在的同一点击序列中所有日志文档的总数，E_{th} 为与日志文档与查询请求的相关性的期望值对应的预设阈值。

[0043] 需要说明的是，上述 E_{th} 为与相关性的期望值对应的预设阈值。通常，对日志文档与查询请求的相关性的期望值设定一个阈值，如设定阈值为 0.5，则可以过滤掉相关性的期望值为 0.5 的日志文档。当然，可以根据用户的使用需求和应用场景任意设定与相关性的期望值对应的预设阈值，本申请实施例对此并不限定。

[0044] 当对日志文档进行过滤后，在本步骤中，计算过滤后的日志文档的期望值。

[0045] 另外，当计算了日志文档与查询请求的相关性的期望值后，且在将日志文档与查询请求的相关性的期望值作为 value 存储在检索数据结构之前，优选的，还可以再进行一次日志文档的过滤，以保证经过二次过滤后，检索数据结构中保存的日志文档与查询请求的相关性更高，搜索引擎可以快速地从检索数据结构中检索到与用户提交的查询请求相关的日志文档和其期望值。此处，由于已经计算得到了日志文档与查询请求的相关性的期望值，因此，可以直接利用日志文档与查询请求的相关性的期望值进行过滤，即，当日志文档与查询请求的相关性的期望值等于预设数值时，过滤掉该日志文档。

[0046] 此外，还可以根据贝叶斯后验概率计算日志文档与查询请求的相关性的方差，可以直接利用日志文档与查询请求的相关性的方差进行过滤，即，当日志文档与查询请求的相关性的方差等于预设数值时，过滤掉该日志文档。其中，方差的计算公式为：

$$D(R_j)=\int_0^1 (R_j-E(R_j))^2 p(R_j \mid C^{LN}) dR_j \text{。}$$

[0047] 还需要说明的是，可以根据用户的使用需求和应用场景任意设定与日志文档与查询请求的相关性的期望值或者方差对应的预设阈值，本申请实施例对此并不限定。

[0048] 步骤 104：将查询请求和日志文档的标识作为 key，将所述日志文档与查询请求的相关性的期望值作为 value 存储在检索数据结构中；

[0049] 例如，key ＝查询请求和日志文档的标识所占内存的连续块，其中，查询请求的一个字符占一个内存字节，日志文档的标识用 4 个字节的内存表示；value ＝期望值乘以 10000 的整数部分所占内存。key 和 value 在检索数据结构中的索引可以采用常用的 trie 树建立，本申请实施例对此不再做详细说明。

[0050] 优选的，为了保证检索数据结构的准确性，在将查询请求和日志文档的标识作为键，将所述日志文档与查询请求的相关性的期望值作为值存储在检索数据结构中之后，还

包括：对所述检索数据结构进行校验；则所述从检索数据结构中查询与所述用户提交的查询请求相关的所有日志文档与查询请求的相关性的期望值为：从通过校验的检索数据结构中查询与所述用户提交的查询请求相关的所有日志文档与查询请求的相关性的期望值。

[0051] 其中，将检索数据结构中的键作为输入数据，经过搜索引擎在检索数据结构的检索后，如果输出的结果是与作为输入数据的键对应的值，则通过验证。例如，在检索数据结构中有一组键值对，该键值对中键对应的查询请求为“MP3”，对应的日志文档标识为 ID1、ID2 和 ID3，该键值对中值对应的日志文档与查询请求的相关性的期望值为 0.5、0.8 和 0.7。分别将“MP3 和 ID1”、“MP3 和 ID2”和“MP3 和 ID3”作为输入数据，经过搜索引擎在检索数据结构中检索后，如果输出的结果分别为 0.5、0.8 和 0.7，则通过检验，否则，没有通过检验。

[0052] 将检索数据结构中的所有键按照上述方式逐一地校验，当所有键都通过校验后，则该检索数据结构通过校验。

[0053] 步骤 105：当接收到用户提交的查询请求时，从所述检索数据结构中查询与所述用户提交的查询请求相关的所有日志文档与查询请求的相关性的期望值；

[0054] 步骤 106：按照期望值从大到小的顺序对查询到的日志文档进行排序。

[0055] 由上述实施例可以看出，本申请基于贝叶斯后验概率计算日志文档与查询请求的相关性的期望值，当从检索数据结构中查询到与用户提交的查询请求相关的所有日志文档与查询请求的相关性的期望值后，按照期望值从大到小的顺序对日志文档进行排序，考虑到了信息的位置因素和点击顺序的因素，使用户快速获得最想要的信息。减少搜索过程对于搜索引擎服务器的消耗，并节省搜索引擎服务器的系统资源。

[0056] 实施例二

[0057] 下面详细说明一种搜索结果的排序方法的优选实施方式。请参阅图 3，其为本申请一种搜索结果的排序方法的另一个实施例的流程图，所述方法包括以下步骤：

[0058] 步骤 301：从日志系统中提取出当天被曝光的日志文档和历史被曝光的日志文档；

[0059] 其中，还可以分别保留一定时间段内当天被曝光的日志文档和历史被曝光的日志文档，例如，保留一个滑动时间窗口内的日志文档，作为一种备份，一旦系统运行过程中发现异常情况，可以用来排查问题和恢复数据。

[0060] 步骤 302：根据日志系统中提取出的当天被曝光的日志文档和历史被曝光的日志文档，分别计算当天局部统计量和历史局部统计量；

[0061] 其中，局部统计量包括 N_j 和 $\tilde{N}_{j,r,d}$，N_j 表示日志文档 j 被点击的总次数，$\tilde{N}_{j,r,d}$ 表示日志文档 j 位于 r+d 处且没有被点击，位置 r 处的日志文档被点击，位置 r 到 r+d 之间的日志文档没有被点击在所有情况下的发生次数，T 表示所有 (r，d) 的可能取值。

[0062] 步骤 303：将当天局部统计量和历史局部统计量进行合并；

[0063] 其中，还可以分别保存一段时间内的当天局部统计量和历史局部统计量，以支持增量更新和排查运行中可能出现的问题，以及恢复数据。

[0064] 步骤 304：根据日志系统中提取的当天被曝光的日志文档和历史被曝光的日志文档，分别计算当天全局统计量和历史全局统计量；

[0065] 其中，全局统计量包括 $N_{r,d}$ 和 $\tilde{N}_{r,d}$，$N_{r,d}$ 为在日志文档 j 所在的同一个点击序列中，位置 r 处的日志文档和位置 r+d 处的日志文档都被点击，位置 r 到 r+d 之间的日志文档没

有被点击在所有情况下的发生次数；$\tilde{N}_{r,d}$为在日志文档 j 所在的同一个点击序列中，位置 r 处的日志文档被点击，位置 r+d 处的日志文档没有被点击，位置 r 到 r+d 之间的日志文档没有被点击在所有情况下发生的次数，r 的取值范围为小于或等于 M-1 的所有自然数，d 的取值范围为小于或等于 M-r 的所有整数，M 表示日志文档 j 所在的同一点击序列中所有日志文档的总数。

[0066] 步骤 305：将当天局部统计量和历史局部统计量进行合并；

[0067] 其中，还可以分别保存一段时间内的当天全局统计量和历史全局统计量，以支持增量更新和排查运行中可能出现的问题，以及恢复数据。

[0068] 步骤 306：根据合并后的全局统计量，计算全局参数；

[0069] 其中，全局参数为$\hat{\beta}_{r,d}$，$\hat{\beta}_{r,d}=\min\{1,\frac{2N_{r,d}}{N_{r,d}+\tilde{N}_{r,d}}\}$。

[0070] 步骤 307：根据全局参数对日志文档进行过滤，使全局参数小于对应的预设阈值的日志文档被过滤；

[0071] 步骤 308：计算过滤后的日志文档与查询请求的相关性的贝叶斯后验概率；

[0072] 其中，计算日志文档与查询请求的相关性的贝叶斯后验概率的过程已经在实施例一中进行了详细地说明，故此处不再赘述，相关计算过程可以参见实施例一。

[0073] 步骤 309：根据贝叶斯后验概率计算过滤后的日志文档与查询请求的相关性的期望值；

[0074] 其中，计算日志文档与查询请求的相关性的期望值的过程已经在实施例一中进行了详细地说明，故此处不再赘述，相关计算过程可以参见实施例一。

[0075] 步骤 310：根据日志文档与查询请求的相关性的期望值对日志文档进行过滤，使相关性的期望值等于预设预置的日志文档被过滤掉；

[0076] 步骤 311：将查询请求和日志文档的标识作为 key，将所述日志文档与查询请求的相关性的期望值作为 value 存储在检索数据结构中；

[0077] 其中，还可以对检索数据结构进行校验，得到通过校验的检索数据结构。

[0078] 步骤 312：当接收到用户提交的查询请求时，从所述检索数据结构中查询与所述用户提交的查询请求相关的所有日志文档与查询请求的相关性的期望值；

[0079] 步骤 313：按照期望值从大到小的顺序对查询到的日志文档进行排序。

[0080] 本申请中的搜索结果的排序方法可以应用在搜索领域，如图 4 所示，其为本申请一种搜索排序系统的结构示意图。每天提取新的搜索日志，每天增量更新，输出相关性的索引数据，更新到搜索排序系统中，作为排序的一个重要因素。

[0081] 另外，本申请中的搜索结果的排序方法还可以应用在排序的评价系统。例如，给定一个查询请求，通过本申请中的排序方法得到日志文档序列中的每个日志文档与查询请求的相关性分值，这种分值是用户对排序的一种隐式评价，可以归一化该相关性分值序列，形成一个概率分布函数 p(x)。同时，用待评价的排序方法对同一个日志文档序列中的每个日志文档计算相关性分值并进行归一化，形成一个概率分布函数 g(x)。将 p(x) 和 g(x) 的距离作为对待评价的排序方法的评估，差距越小，则待评价的排序方法的评价越高。距离计算公式可以为：$\sum_x \frac{p(x)\log(p(x))}{g(x)}$。

[0082] 另外，本申请中的搜索结果的排序方法还可以应用在排序的训练系统。例如，通过本申请中的排序方法得到（查询，文档）对的相关性分值，用 Y 表示，然后抽取（查询，文档）对的特征，用 X 表示，如文本特征和图像特征。然后，用于机器学习方法训练得到相关性算法 Y = f(X)。

[0083] 由上述实施例可以看出，本申请基于贝叶斯后验概率计算日志文档与查询请求的相关性的期望值，当从检索数据结构中查询到与用户提交的查询请求相关的所有日志文档与查询请求的相关性的期望值后，按照期望值从大到小的顺序对日志文档进行排序，考虑到了信息的位置因素和点击顺序的因素，使用户快速获得最想要的信息。减少搜索过程对于搜索引擎服务器的消耗，并节省搜索引擎服务器的系统资源。

[0084] 实施例三

[0085] 与上述一种搜索结果的排序方法相对应，本申请实施例还提供了一种搜索结果的排序装置。请参阅图 5，其为本申请一种搜索结果的排序装置的一个实施例的结构示意图，包括：提取模块 501、概率计算模块 502、期望值计算模块 503、索引建立模块 504、检索模块 505 和排序模块 506。下面结合该装置的工作原理进一步介绍其内部结构以及连接关系。

[0086] 提取模块 501，用于从日志系统中提取出被曝光的日志文档；

[0087] 概率计算模块 502，用于计算所述日志文档与查询请求的相关性的贝叶斯后验概率；

[0088] 期望值计算模块 503，用于根据所述贝叶斯后验概率计算所述日志文档与查询请求的相关性的期望值；

[0089] 索引建立模块 504，用于将查询请求和日志文档的标识作为键，将所述日志文档与查询请求的相关性的期望值作为值存储在检索数据结构中；

[0090] 检索模块 505，用于当接收到用户提交的查询请求时，从所述检索数据结构中查询与所述用户提交的查询请求相关的所有日志文档与查询请求的相关性的期望值；

[0091] 排序模块 506，用于按照期望值从大到小的顺序对查询到的日志文档进行排序。

[0092] 优选的，请参阅图 6，其为本申请一种搜索结果的排序装置的另一个实施例的结构示意图。除了包括有提取模块 501、概率计算模块 502、期望值计算模块 503、索引建立模块 504、检索模块 505 和排序模块 506 之外，所述装置还包括：第一过滤模块 507，用于在根据所述贝叶斯后验概率计算日志文档与查询请求的相关性的期望值之前，根据全局参数对日志文档进行过滤，使全局参数小于对应的预设阈值的日志文档被过滤；

[0093] 则期望值计算模块 503，用于根据贝叶斯后验概率计算过滤后的日志文档与查询请求的相关性的期望值。

[0094] 其中，第一过滤模块 507 进一步包括：筛选子模块 5071 和过滤子模块 5072，

[0095] 筛选子模块 5071，用于从提取出的被曝光的日志文档中筛选出被曝光一次且没有被点击的日志文档；

[0096] 过滤子模块 5072，从筛选出的日志文档中，按照过滤条件公式 $\beta_{r,d} < \frac{1/E_{th}}{\frac{1}{3}/\frac{1}{2}E_{th}}$ 过滤掉全局参数小于对应的预设阈值的日志文档，其中，$\beta_{r,d}$ 为全局参数，

$\hat{\beta}_{r,d}=\min\{1,\frac{2N_{r,d}}{N_{r,d}+\tilde{N}_{r,d}}\}$，$N_{r,d}$为在被筛选出的日志文档所在的同一个点击序列中，位置 r 处的日志文档和位置 r+d 处的日志文档都被点击，位置 r 到 r+d 之间的日志文档没有被点击在所有情况下的发生次数；$\tilde{N}_{r,d}$为在被筛选出的日志文档所在的同一个点击序列中，位置 r 处的日志文档被点击，位置 r+d 处的日志文档没有被点击，位置 r 到 r+d 之间的日志文档没有被点击在所有情况下发生的次数；r 的取值范围为小于或等于 M-1 的所有自然数，d 的取值范围为小于或等于 M-r 的所有整数，M 表示被筛选出的日志文档所在的同一点击序列中所有日志文档的总数，E_{th}为与相关性的期望值对应的预设阈值。

[0097] 优选的，请参阅图 7，其为本申请一种搜索结果的排序装置的另一个实施例的结构示意图。所述装置还包括：第二过滤模块 508，用于在所述将查询请求和日志文档的标识作为 key，将所述日志文档的期望值作为 value 存储在检索数据结构中之前，根据日志文档与查询请求的相关性的期望值或者方差对日志文档进行过滤，使期望值或者方差等于对应的预设阈值的日志文档被过滤，

[0098] 则索引建立模块 504，用于将查询请求和日志文档的标识作为 key，将过滤后的日志文档的期望值作为 value 存储在检索数据结构中。

[0099] 优选的，请参阅图 8，其为本申请一种搜索结果的排序装置的另一个实施例的结构示意图。除了包括有提取模块 501、概率计算模块 502、期望值计算模块 503、索引建立模块 504、检索模块 505 和排序模块 506 之外，所述装置还包括：校验模块 509，用于在所述将查询请求和日志文档的标识作为键，将所述日志文档与查询请求的相关性的期望值作为值存储在检索数据结构中之后，对所述检索数据结构进行校验，

[0100] 则检索模块 505，用于从通过校验的检索数据结构中查询与所述用户提交的查询请求相关的所有日志文档与查询请求的相关性的期望值。

[0101] 由上述实施例可以看出，本申请基于贝叶斯后验概率计算日志文档与查询请求的相关性的期望值，当从检索数据结构中查询到与用户提交的查询请求相关的所有日志文档与查询请求的相关性的期望值后，按照期望值从大到小的顺序对日志文档进行排序，考虑到了信息的位置因素和点击顺序的因素，使用户快速获得最想要的信息。减少搜索过程对于搜索引擎服务器的消耗，并节省搜索引擎服务器的系统资源。

[0102] 需要说明的是，本领域普通技术人员可以理解实现上述实施例方法中的全部或部分流程，是可以通过计算机程序来指令相关的硬件来完成，所述的程序可存储于一计算机可读取存储介质中，该程序在执行时，可包括如上述各方法的实施例的流程。其中，所述的存储介质可为磁碟、光盘、只读存储记忆体 (Read-Only Memory，ROM) 或随机存储记忆体 (Random AccessMemory，RAM) 等。

[0103] 以上对本申请所提供的一种搜索结果的排序方法和装置进行了详细介绍，本文中应用了具体实施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。

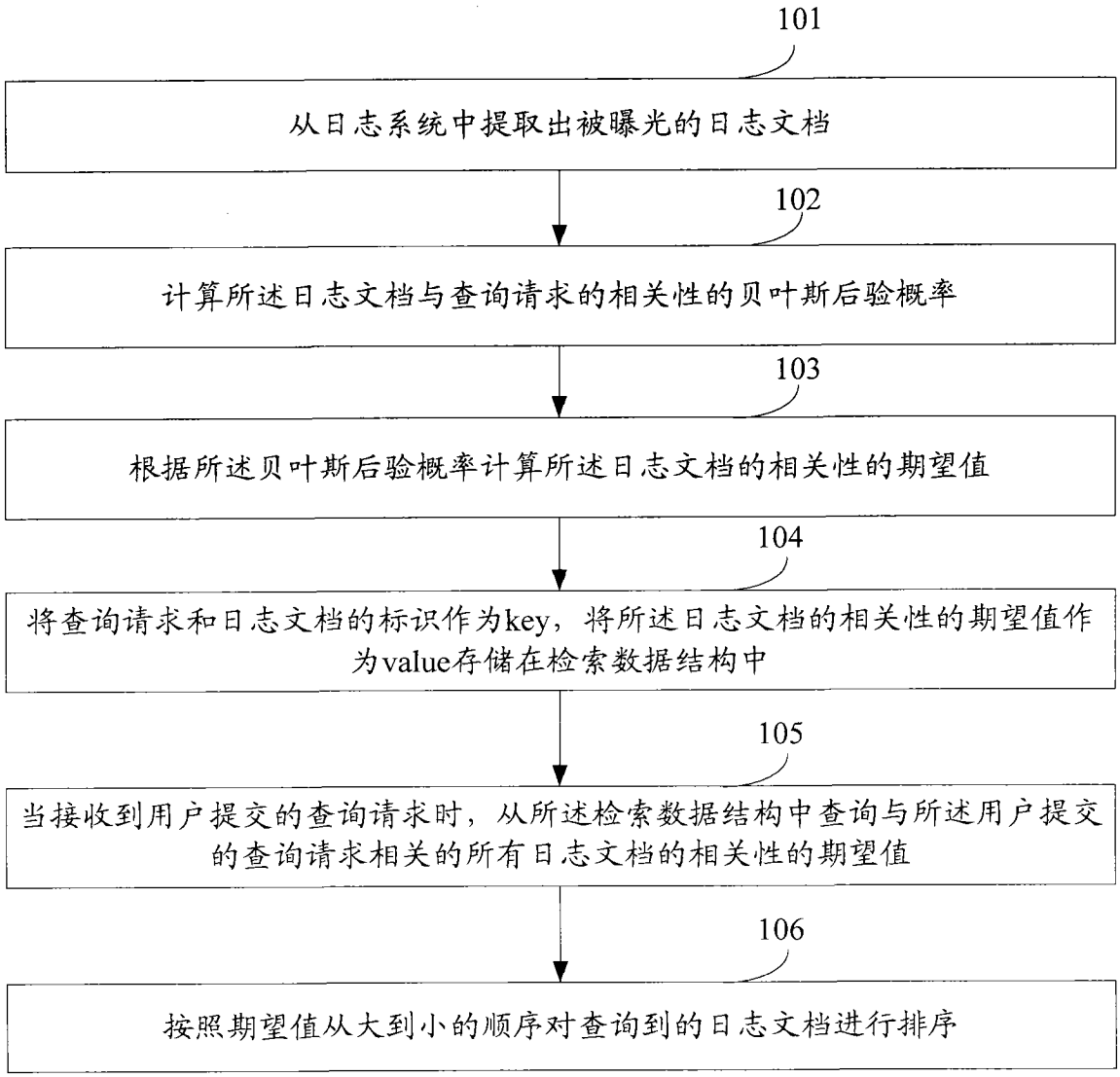

101
从日志系统中提取出被曝光的日志文档
102
计算所述日志文档与查询请求的相关性的贝叶斯后验概率
103
根据所述贝叶斯后验概率计算所述日志文档的相关性的期望值
104
将查询请求和日志文档的标识作为key，将所述日志文档的相关性的期望值作为value存储在检索数据结构中
105
当接收到用户提交的查询请求时，从所述检索数据结构中查询与所述用户提交的查询请求相关的所有日志文档的相关性的期望值
106
按照期望值从大到小的顺序对查询到的日志文档进行排序

图 1

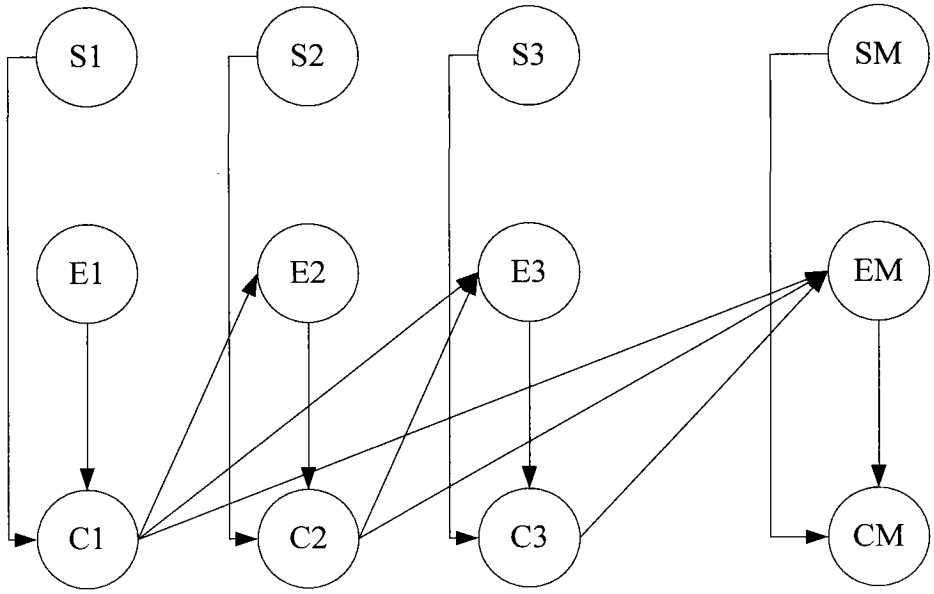


图 2

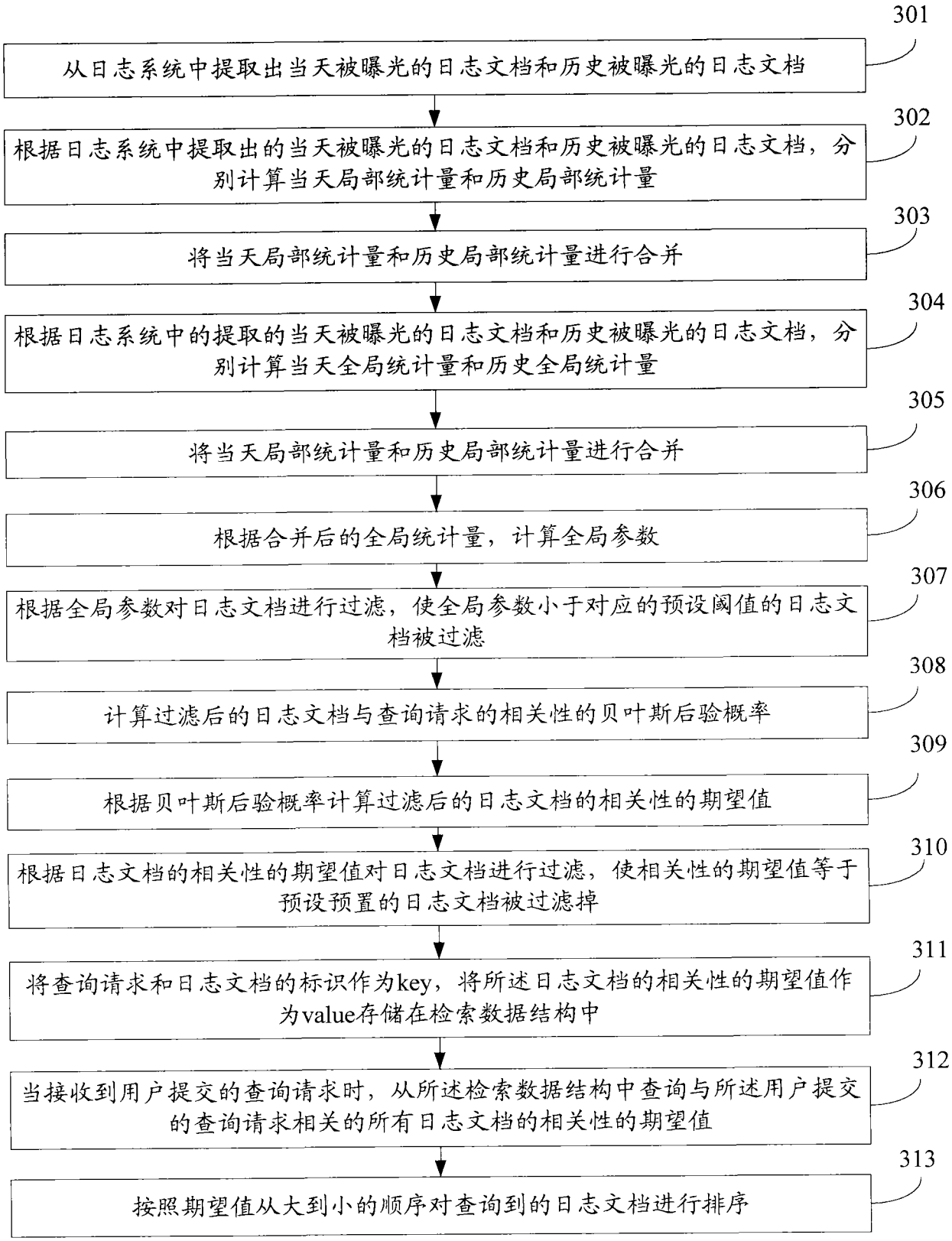

301
从日志系统中提取出当天被曝光的日志文档和历史被曝光的日志文档
302
根据日志系统中提取出的当天被曝光的日志文档和历史被曝光的日志文档，分别计算当天局部统计量和历史局部统计量
303
将当天局部统计量和历史局部统计量进行合并
304
根据日志系统中的提取的当天被曝光的日志文档和历史被曝光的日志文档，分别计算当天全局统计量和历史全局统计量
305
将当天局部统计量和历史局部统计量进行合并
306
根据合并后的全局统计量，计算全局参数
307
根据全局参数对日志文档进行过滤，使全局参数小于对应的预设阈值的日志文档被过滤
308
计算过滤后的日志文档与查询请求的相关性的贝叶斯后验概率
309
根据贝叶斯后验概率计算过滤后的日志文档的相关性的期望值
310
根据日志文档的相关性的期望值对日志文档进行过滤，使相关性的期望值等于预设预置的日志文档被过滤掉
311
将查询请求和日志文档的标识作为key，将所述日志文档的相关性的期望值作为value存储在检索数据结构中
312
当接收到用户提交的查询请求时，从所述检索数据结构中查询与所述用户提交的查询请求相关的所有日志文档的相关性的期望值
313
按照期望值从大到小的顺序对查询到的日志文档进行排序

图 3

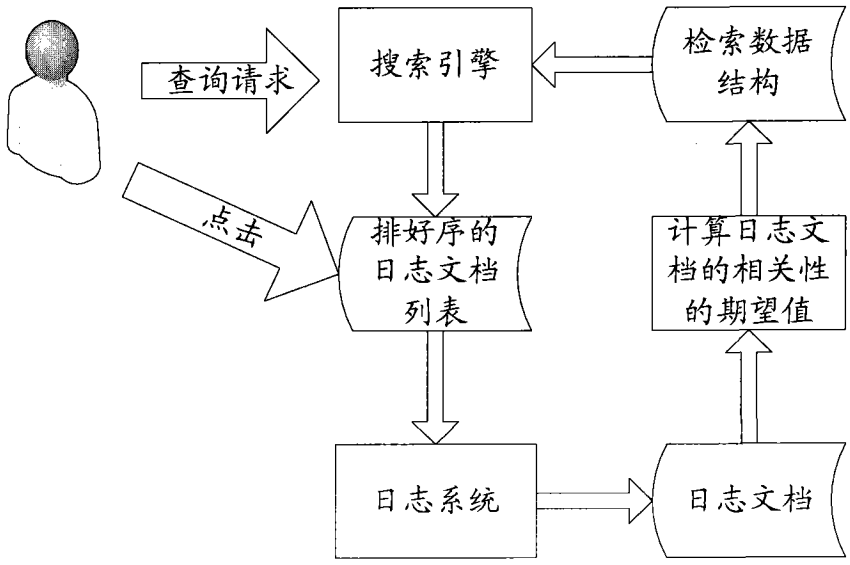


图 4

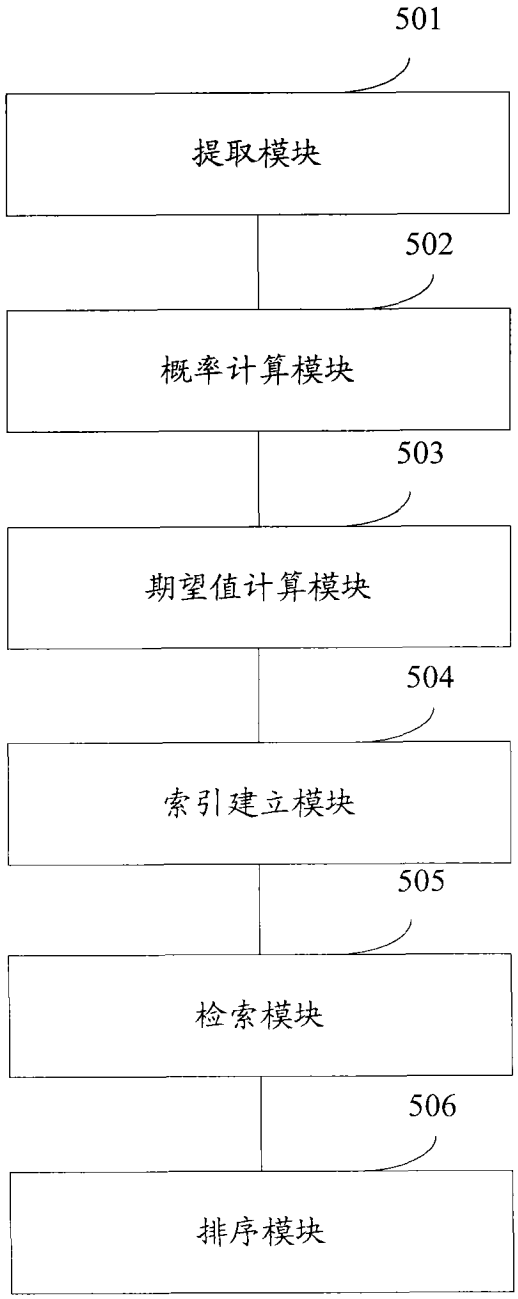

501
提取模块
502
概率计算模块
503
期望值计算模块
504
索引建立模块
505
检索模块
506
排序模块

图 5

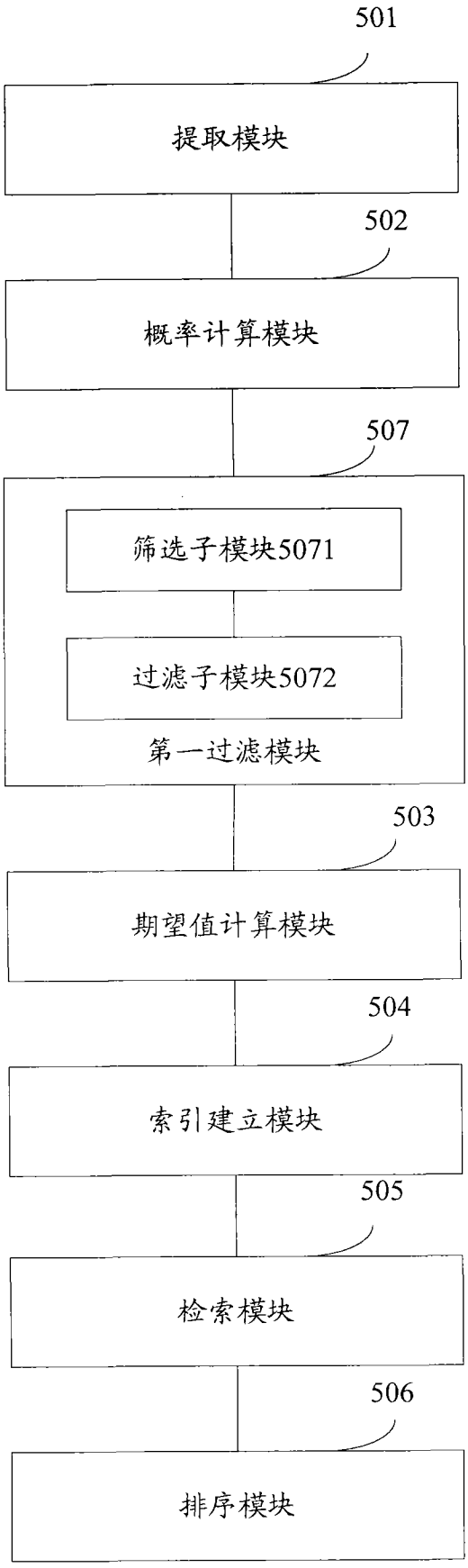

501
提取模块
502
概率计算模块
507
筛选子模块5071
过滤子模块5072
第一过滤模块
503
期望值计算模块
504
索引建立模块
505
检索模块
506
排序模块

图 6

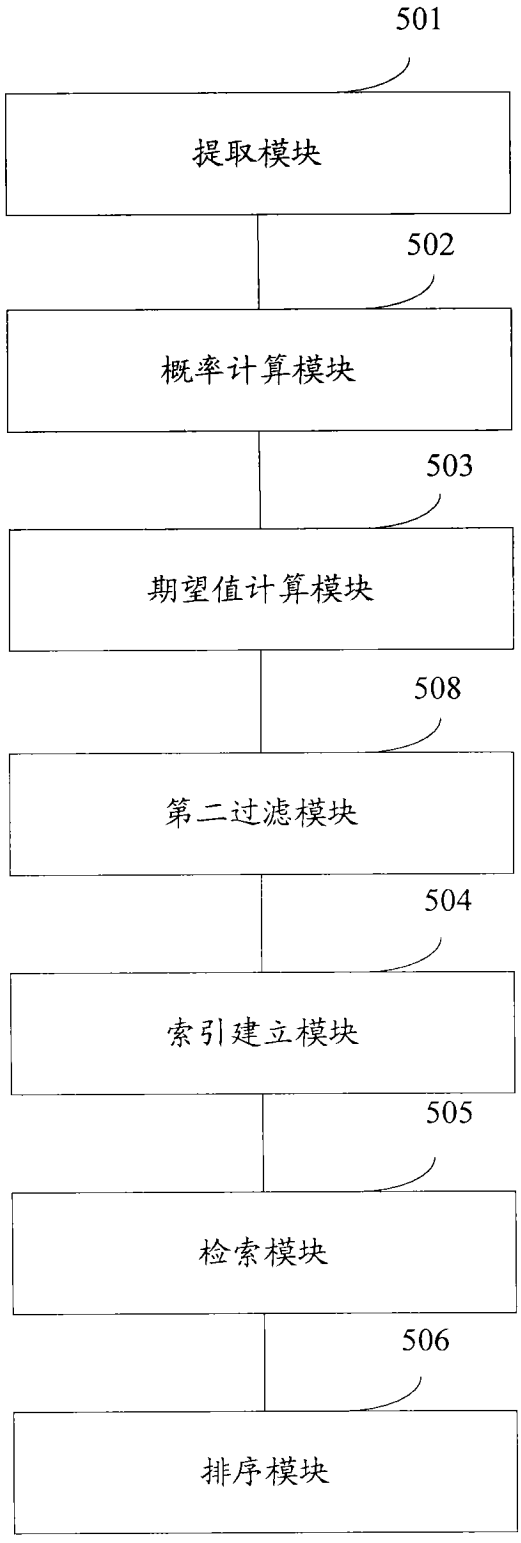

501
提取模块
502
概率计算模块
503
期望值计算模块
508
第二过滤模块
504
索引建立模块
505
检索模块
506
排序模块

图 7

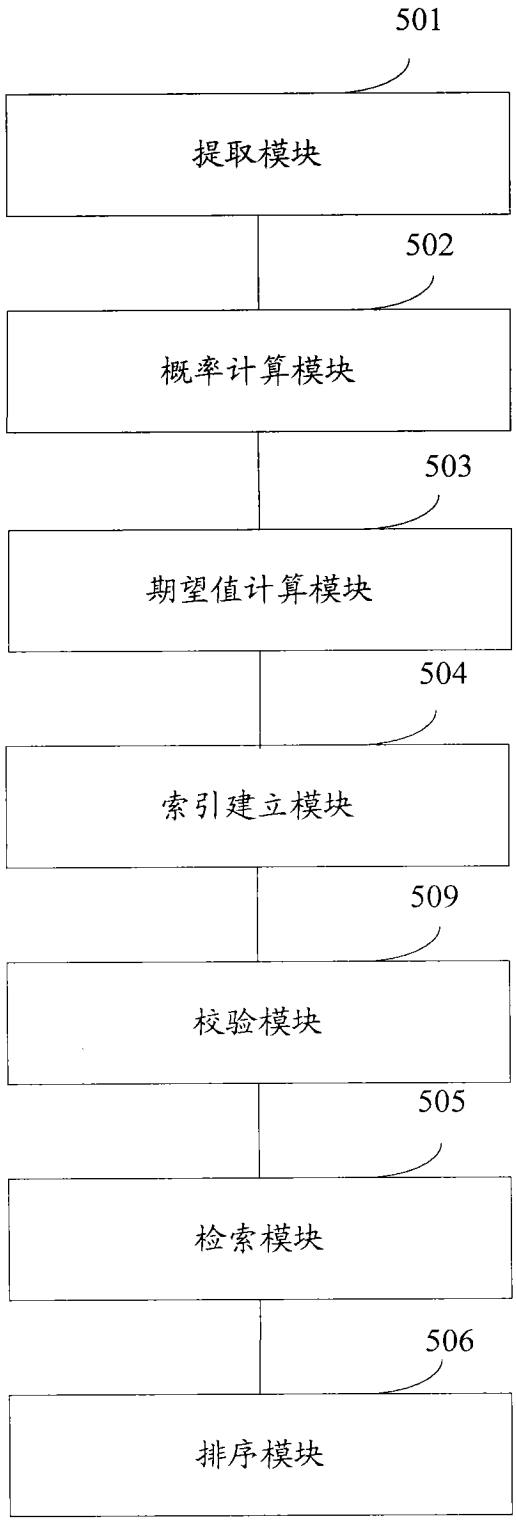

501
提取模块
502
概率计算模块
503
期望值计算模块
504
索引建立模块
509
校验模块
505
检索模块
506
排序模块

图 8