(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 102231168 A
(43)申请公布日 2011.11.02

(21)申请号 201110216654.4

(22)申请日 2011.07.29

(71)申请人 前锦网络信息技术(上海)有限公司
地址 201203 上海市浦东新区张东路 1387 号 3 楼 1F

(72)发明人 俞希林 孔卫东

(74)专利代理机构 上海新天专利代理有限公司 31213
代理人 周涛

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

权利要求书 1 页 说明书 6 页 附图 5 页

(54)发明名称

一种从简历库中快速检索简历的方法

(57)摘要

本发明涉及一种从简历库中快速检索简历的方法,该方法是在简历数据库与 web 服务器之间设置一个简历搜索引擎,利用该简历搜索引擎将简历的全文关键字进行快速索引,通过按字索引的方式使简历数据库中存在的简历数据能够快速有效地被检索出并在 web 服务器中呈现出来。本发明的搜索方法能够实现简历数据库内的所有简历都快速有效地被检索出来,在保证检索准确性的情况下大幅提高了检索速度。

CN 102231168 A

1. 一种从简历库中快速检索简历的方法，其特征在于，该方法是在简历数据库与 web 服务器之间设置一个简历搜索引擎，利用该简历搜索引擎将简历的全文关键字进行快速索引，通过按字索引的方式使简历数据库中存在的简历数据能够快速有效地被检索出并在 web 服务器中呈现出来。

2. 根据权利要求 1 所述的一种从简历库中快速检索简历的方法，其特征在于，该方法包括简历索引生成阶段和简历索引搜索服务阶段：在简历索引生成阶段：

第一步，将简历数据库中的简历按照更新时间进行降序排列，以降序读取新增、修改、逻辑删除的简历数据；

第二步，扫描每份简历在数据库中的索引字段，按照月份生成索引文件，索引文件包括文件头段落、精确搜索段落、字索引段落和详细位置信息段落，每天更新生成当月的索引文件，该索引文件通过复制的方式更新数据至索引服务器上；

简历索引搜索服务阶段：

第三步，所述的索引服务器为多线程模块，其包括有主线程、工作线程和监控线程，主线程通过套接字在指定端口 8454 监听搜索请求，若有搜索请求则将其转给工作线程处理；

第四步，工作线程接收 Web 客户端的搜索请求并将搜索请求信息解析，如果不包含关键字，直接进行精确搜索段落的检索，如果包含关键字，则通过字索引段落，快速找到每个关键字的详细位置信息的起始位置，并判断关键字是否符合搜索请求，如果关键字满足搜索请求，则继续判断精确搜索段落是否满足搜索请求，如果都满足搜索请求，则将简历 ID 放入搜索结果中，搜索完成后返回 Web 客户端；

第五步，监控线程定时扫描文件更新，若索引文件正在更新，则将当前服务器的搜索请求转移到备份服务器上搜索。

3. 根据权利要求 2 所述的一种从简历库中快速检索简历的方法，其特征在于，所述第二步的索引文件中文件头段落包含统计信息，包括精确搜索段落，字索引段落，详细位置信息段落在文件中的起始位置，节点的大小，以及节点的数量；精确搜索段落包括自增长的内部 ID，数据库的简历 ID，居住地，学历，性别，工作年限，简历更新时间以及状态信息；字索引段落包括汉字和英文信息，每个汉字为一个节点，每个英文单词为一个单独节点；详细位置信息段落记录字的内部 ID、字段以及位置信息。

4. 根据权利要求 2 所述的一种从简历库中快速检索简历的方法，其特征在于，所述第四步中查找关键字的详细位置信息时，对比每个关键字的前后位置信息，若所有关键字都满足位置信息，则该条记录满足关键字搜索条件，如果同时满足精确搜索条件，就可以提取该简历 ID 至搜索结果集。

一种从简历库中快速检索简历的方法

技术领域

[0001] 本发明涉及到搜索引擎，特别是一种从简历库中快速检索出简历文件的方法搜索引擎。

背景技术

[0002] 在招聘行业中，大部分网站都采用数据库搜索的方式来实现简历数据的检索。这种数据库搜索技术适用于简历数据量较小的网站，但是一旦简历数量巨大且增长过快，例如达到三千万条的数量，使用普通的数据库技术进行简历搜索时，在巨量的简历数据面前，其检索性能就非常差。特别是在搜索关键字时，处于 web 客户端的用户响应速度很慢，需要几秒到几十秒，甚至出现部分关键字搜索无法获得搜索结果。针对这种情形，作为简历搜索者的企事业单位用户常常投诉简历搜索速度过慢，搜索效果太差。

[0003] 在目前的技术条件下，由于数据库的简历数量非常大，很难直接通过硬件扩展的方式来大幅度提高系统性能，并且对硬件扩展和升级的成本也较高。总结现有技术存在的问题主要在于两点：第一是企业用户简历库搜索简历时搜索速度过慢；第二是搜索结果不完整，很多搜索引擎只返回部分结果，并且搜索引擎本身使用词库的方式来建立索引，不能保证简历库中的所有数据都能够被有效检索。

发明内容

[0004] 本发明的目的在于克服上述现有技术存在的不足，提供一种新的从巨量简历库中快速检索简历的方法。本发明的搜索方法要能够实现简历数据库内的所有简历都能够快速有效地被检索出来，在保证检索准确性的情况下要能大幅提高检索速度。

[0005] 为了达到上述发明目的，本发明提供的技术方案如下：

[0006] 一种从简历库中快速检索简历的方法，其特征在于，该方法是在简历数据库与 web 服务器之间设置一个简历搜索引擎，利用该简历搜索引擎将简历的全文关键字进行快速索引，通过按字索引的方式使简历数据库中存在的简历数据能够快速有效地被检索出并在 web 服务器中呈现出来。

[0007] 上述从简历库中快速检索简历的方法主要包括有简历索引生成阶段和简历索引搜索服务阶段：

[0008] 在简历索引生成阶段：

[0009] 第一步，将简历数据库中的简历按照更新时间进行降序排列，以降序读取新增、修改、逻辑删除的简历数据；

[0010] 第二步，扫描每份简历在数据库中的索引字段，按照月份生成索引文件，索引文件包括文件头段落、精确搜索段落、字索引段落和详细位置信息段落，每天更新生成当月的索引文件，该索引文件通过复制的方式更新数据至索引服务器上；

[0011] 在简历索引搜索服务阶段：

[0012] 第三步，所述的索引服务器为多线程模块，其包括有主线程、工作线程和监控线

程，主线程通过套接字在指定端口 8454 监听搜索请求，若有搜索请求则将其转给工作线程处理；

[0013] 第四步，工作线程接收 Web 客户端的搜索请求并将搜索请求信息解析，如果不包含关键字，直接进行精确搜索段落的判断，如果包含关键字，则通过字索引段落，找到每个关键字的详细位置信息的起始位置，判断是否符合搜索请求，如果关键字满足搜索请求，则继续判断精确搜索段落是否满足搜索请求，如果都满足搜索请求，则将简历 ID 放入搜索结果中，搜索完成后返回客户端；

[0014] 第五步，监控线程定时扫描文件更新，若索引文件正在更新，则将当前服务器的搜索请求转移到备份服务器上搜索。

[0015] 在本发明的从简历库中快速检索简历的方法中，所述第二步的索引文件中文件头段落包含的统计信息包括精确搜索段落，字索引段落，详细位置信息段落在文件中的起始位置，每个节点的大小，以及节点的数量；精确搜索段落包括自增长的内部 ID，数据库的简历 ID，居住地，学历，性别，工作年限，简历更新时间以及状态信息；字索引段落包括汉字和英文信息，每个汉字为一个节点，每个英文单词为一个单独节点；详细位置信息段落记录字的内部 ID、字段以及位置信息。

[0016] 在本发明的从简历库中快速检索简历的方法中，所述第四步中查找关键字的详细位置信息时，对比每个关键字的前后位置信息，若所有关键字都满足位置信息，则该条记录满足关键字搜索条件，如果同时满足精确搜索条件，就可以提取该简历 ID 至搜索结果集。

[0017] 基于上述技术方案，本发明的从简历库中快速检索简历的方法与现有技术相比具有如下技术优点：

[0018] 1. 本发明的简历库搜索方法可以实现简历库内所有简历数据的全文关键字的快速搜索，并且按字索引，从而保证了简历数据库内存在的数据都能够有效地被检索出来。

[0019] 2. 企业用户利用本发明的搜索引擎可以从包含大量简历的数据库中快速准确地进行关键字检索以及部分信息的精确检索，从而解决了搜索速度慢的问题，使得企业用户能够快速的找到需要的简历文件。

附图说明

[0020] 图 1 是本发明从简历库中快速检索简历的方法的总体思路示意图。

[0021] 图 2 是本发明从简历库中快速检索简历的方法中简历索引生成阶段的流程示意图。

[0022] 图 3 是本发明从简历库中快速检索简历的方法中简历索引搜索服务阶段的流程示意图。

[0023] 图 4 是本发明从简历库中快速检索简历的方法中词语搜索过程的示意图。

[0024] 图 5 是本发明从简历库中快速检索简历的方法中字搜索过程的示意图。

具体实施方式

[0025] 下面我们结合附图和具体的实施例来对本发明从简历库中快速检索简历的方法做进一步的详细阐述，以求更为清楚明了地理解其含义和过程算法，但不能以此来限制本发明的保护范围。

[0026] 先请看图 1，图 1 是本发明从简历库中快速检索简历的方法的总体思路示意图。由图可知，本发明从简历库中快速检索简历的方法是在简历数据库与 web 服务器之间设置一个简历搜索引擎，利用该简历搜索引擎将简历的全文关键字进行快速索引，通过按字索引的方式使简历数据库中存在的简历数据能够快速有效地被检索出来，并且在 web 服务器中呈现给用户。而这里的简历搜索引擎主要包括了两个部分，即索引生成器和索引服务器，它们各自相对应的处理阶段为索引生成阶段和索引搜索服务阶段。简历搜索引擎中的索引生成器通过开放式数据库互联 ODBC 连接至简历数据库中，而简历生成器生成的索引文件会复制到索引服务器中，而索引服务器则通过套接字连接到 Web 服务器中。

[0027] 上述的简历搜索引擎中包括的索引生成器用于从简历数据库中读取简历数据，将精确搜索字段和关键字字段信息生成符合搜索规范的数据文件，并分发到多台索引服务器中。这里的索引服务器根据 Web 服务器提供的搜索条件，快速地从索引文件中查询到符合条件的简历数据，然后返回给 Web 服务器进行显示。

[0028] 索引生成器在从简历库中快速检索简历的方法中由索引生成阶段完成其功能，具体流程如图 2 所示，图 2 是本发明从简历库中快速检索简历的方法中简历索引生成阶段的流程示意图。由图可知，该简历索引生成阶段的处理流程如下：

[0029] 这里的简历数据库为 SQL Server 2000 数据库，该阶段就是读取简历数据库中存储的简历数据，按照月份生成索引文件，正式的索引文件结构分为四个段落：

[0030] 1. 文件头段落，记录每个段落的起始位置，数量，节点大小，以及统计信息。

[0031] 2. 精确搜索段落，包括自增长的内部 ID、数据库的简历 ID、居住地、学历、性别、工作年限、简历更新时间、状态需要精确搜索的字段信息，该段落使用顺序结构存储。

[0032] 3. 字索引段落，使用 B+ 树结构存储索引信息，包括关键字、起始位置，记录数量信息。

[0033] 4. 详细位置信息段落，使用顺序结构存储，存放索引的内部 ID，字段编号，关键字位置等信息，包括内部 ID(4 个字节）、字段编号（1 个字节）以及位置（2 个字节）信息，每个字段内超过 30000 的数据将被忽略。

[0034] 所述的索引创建过程如下：

[0035] 1. 预留内存缓冲区，用于存储生成索引过程中产生的临时索引数据，包括关键字、内部 ID、字段编号以及位置信息。

[0036] 2. 连接简历数据库，按照简历更新时间降序，简历 ID 降序获取简历数据结果集，包括精确搜索字段和关键字字段，精确搜索字段包括简历 ID，居住地，学历，性别，工作年限，简历更新时间，状态，关键字字段包括工作经验，教育经验，项目经验。

[0037] 3. 开始循环结果集数据，定义自增长的 ID，从 0 开始计数，获取简历 ID，居住地，学历，性别，工作年限，简历更新时间，状态等精确搜索字段，这些信息保存到精确搜索段落中，判断简历更新时间，如果和前面记录属于不同的月份，则创建新的索引文件。

[0038] 4. 判断简历更新时间是否大于删除日期，如果大于删除日期，则写删除文件。索引生成程序每天凌晨运行一次，删除日期为前一天，删除文件中包含的数据为当天新增或者修改的记录，运行完成后自动更新删除日期为当天。

[0039] 5. 循环每一个关键字字段，循环每一个关键字，填充关键字，内部 ID，字段编号，位置信息到缓冲区中，如果缓冲区满，则调用保存数据模块保存缓冲区数据，否则继续填充

数据，直到所有字段都被处理。

[0040] 6. 保存数据模块将缓冲区数据进行快速排序，排序结果按照关键字 + 内部 ID+ 字段 + 位置顺序有序排列，即相同的字都是在一起的，循环整个缓冲区，将字信息插入到临时索引缓冲区和索引缓冲区，内部 ID+ 字段 + 位置结构顺序写入到临时索引文件中。临时索引缓冲区中关键字 + 起始位置字段唯一，按照顺序以 B+ 树结构存储，索引缓冲区中关键字唯一，以 B+ 树结构存储，保存完成后，清空缓冲区。

[0041] 7. 跳转到第 3 步继续处理下一条简历，直到所有简历都被处理。处理完成后，调用保存数据模块保存最后数据，更新删除文件的开始 4 个字节的内容（long 数据类型）为删除总数。

[0042] 8. 循环读取临时索引缓冲区的节点信息，如果节点的字信息为不同的字，则更新索引缓冲区的起始位置，节点数量信息，保存临时缓冲区的数据到正式索引文件中，如果节点的字相同，继续从临时索引文件中复制数据，直到所有节点的数据都被处理，最后将索引缓冲区的数据写入正式索引文件，写入文件头段落的统计信息，删除临时索引文件。

[0043] 9. 读取分发列表，将正式索引文件和删除文件分发到多台索引服务器上，并将配置文件中的内容从 0 改为 1，表示索引文件已经完成更新。

[0044] 简历搜索引擎中的索引服务器由索引搜索服务阶段完成其功能，这里的索引服务阶段存在多个线程程序，其作用是启动套接字在指定端口监听客户端的搜索请求，如果检测到搜索请求，则进行快速进行解析，数据读取，数据匹配，通过高效的检索，将符合条件的简历 ID 结果集返回到 Web 客户端。其具体的处理流程如图 3 所示，图 3 是本发明从简历库中快速检索简历的方法中简历索引搜索服务阶段的流程示意图。由图可知，简历索引搜索服务阶段的处理流程如下：

[0045] 1. 读取配置文件信息，包括 AWE 内存大小，装载的索引文件时间范围，工作线程数量等配置信息。简历索引服务器使用了地址窗口化扩展插件 (Address Windowing Extensions，简称 AWE)，这样在 32 位的 Windows 操作系统上可以使用 4G 以上的物理内存，将索引文件常驻内存来提高搜索效率。

[0046] 2. 装载索引数据到内存中。将文件头段落，精确搜索段落，字索引段落数据装载到进程空间内的内存中，将索引文件中数据量最大的详细信息段落装载到 AWE 内存中，以提高搜索效率。

[0047] 3. 启动监控线程，用于监控索引文件更新，启动工作线程，用于提供搜索服务。

[0048] 4. 主线程在指定端口 8454 监听客户端搜索请求。

[0049] 5. 检测是否有搜索请求，如果检测到搜索请求，则将套接字接收下来，将主线程阻塞，工作线程将接收套接字复制，并唤醒主线程继续接收其他搜索请求。

[0050] 6. 工作线程判断本机是否出于更新索引状态，如果当前机器处于更新状态，将搜索请求转给备份服务器进行处理，通过套接字连接备份服务器，备份服务器搜索完成后将数据返回给本机，本机将搜索结果转发给 Web 客户端。

[0051] 7. 当前工作线程开始接收客户端详细的搜索请求，并将编码的搜索请求转化为内部的搜索关键字结构，如果传入的数据不符合数据规范，则直接返回 0 给 Web 客户端。

[0052] 索引文件是按照简历更新日期降序排列，搜索从最新的索引文件开始，判断索引文件日期是否小于搜索的开始日期，如果小于，则剩下的索引文件中必然不包含有效数据，

无需继续搜索，否则取出当前索引文件的信息，调用词语搜索算法进行搜索，循环所有的索引文件，最后返回搜索要求的简历 ID 的列表给 Web 客户端。

[0053] 8. 监控线程每隔一段时间，通常为 5 秒扫描一次索引更新目录下的配置文件，判断索引文件是否更新。如果发现索引文件已经更新，则本服务器接收到的搜索请求都转给备份服务器进行搜索，本服务器开始进行数据更新，重新装载最新的索引文件，包括文件头段落，精确搜索段落，字索引段落以及位置索引段落，以前月份的索引文件则只进行精确搜索段落的更新，将在删除文件中存在的简历 ID 的状态从 1 改为 0，这样表示在这个索引文件中这份简历已经无效，在搜索过程中状态为 0 的简历将会被认为不满足精确搜索条件，这样的处理结果保证搜索结果不会出现数据重复。

[0054] 在简历索引搜索服务阶段的处理流程中涉及到了词语搜索，其具体的流程如图 4 所示，图 4 是本发明从简历库中快速检索简历的方法中词语搜索过程的示意图。由图可知，词语搜索算法的处理流程如下：

[0055] 1. 判断是否存在关键字。如果存在关键字，跳转到步骤 3，否则继续执行步骤 2。

[0056] 2. 直接进行精确搜索部分字段的过滤，符合搜索条件的记录，放入结果集中，不符合搜索条件，继续进行搜索，遍历整个精确搜索部分或者搜索结果数量已经达到最大的返回数量。

[0057] 3. 解析关键字，从字索引段落的 B+ 树结构中查找每一个字的起始位置以及数量信息，并从索引文件中或者 AWE 内存中读取详细位置信息的部分或者全部数据到内存中。

[0058] 4. 判断词语的数量，如果包含多于一个词语，则跳转到步骤 5，否则调用字搜索算法搜索当前词语，如果存在满足条件的记录，取出精确搜索部分信息进行过滤，如果满足条件，则将简历 ID 放入结果集中，继续进行搜索，直到遍历所有字的记录或者搜索结果数量达到最大的返回数量。

[0059] 5. 关键字中包含多个词语，比较前后两个词语的内部 ID，如果前一词语的内部 ID 小于后一词语的内部 ID，则移动前一词语，否则移动后一词语，直到前后两个词语的内部 ID 相等或者有词语数据已经全部读取完成。

[0060] 6. 判断在前面移动的过程中，是否移动了前一词语并且前一词语不是第一个词语，如果移动了，则需要回溯词语指针，将前一词语赋给后一词语，前二词语赋给前一词语，重新比较前一词语和后一词语，直到前后两个词语相等并且前一词语没有移动或者前一词语为第一个词语。这样不断循环，直到后一词语为最后一个词语为止或者有词语数据已经全部读取完成，如果有词语数据已经全部读取完成，则置搜索结束标记。

[0061] 7. 判断是否找到搜索结果，如果找到，则进行精确字段信息过滤，如果满足条件，则将简历 ID 放入结果集中，如果搜索的结果数量已经达到最大的返回数量，则置搜索结束标记。

[0062] 8. 判断是否搜索结束，如果有词语的数据已经全部读取完成或者搜索结果数量已经达到最大的返回数量，则搜索结束，继续查找不可能在找到符合条件的结果，执行步骤 9，否则返回步骤 5，继续搜索。

[0063] 9. 将搜索到简历 ID 列表返回，本索引文件搜索结束。

[0064] 在上述的词语搜索中涉及到了字搜索过程，如图 5 所示，图 5 是本发明从简历库中快速检索简历的方法中字搜索过程的示意图。由图可知，其具体的处理流程如下：

[0065] 1. 判断当前词语是否只包含一个字，如果包含多余一个字，执行步骤 3。

[0066] 2. 移动当前字到下一条记录（第一次除外）。如果当前已经读取的内存块中包含记录，直接移动到下一个记录即可，如果当前内存块中已经读取完成，则需要从 AWE 内存中或者从索引文件中读取下一块的信息到内存中，如果已经全部读取完成，则返回 False，否则读取下一记录，记录内部 ID 信息，返回 True。

[0067] 3. 当前词语中如果包含多个字，比较前后字的位置信息，包括内部 ID，字段，以及位置信息，符合条件的词语中前后关键字位置差 1，这样将前一关键字的位置信息加 1 后和后一关键字进行比较，如果前一关键字小于后一关键字，则移动前一关键字，否则移动后一关键字，直到前后两个关键字的内部 ID，字段，位置信息满足要求或者有关键字的数据已经全部读取完成。

[0068] 4. 判断在前面移动的过程中，是否移动了前一关键字并且前一关键字不是第一个关键字，如果移动了，则需要回溯关键字指针，将前一关键字赋给后一关键字，前二关键字赋给前一关键字，重新比较前一关键字和后一关键字，直到前后两个关键字位置信息满足要求并且前一关键字没有移动或者前一关键字为第一个关键字。这样不断循环，直到后一关键字为最后一个关键字为止或者有关键字数据已经全部读取完成，如果有关键字数据已经全部读取完成，返回 False。

[0069] 5. 记录满足条件的内部 ID，返回 true。

[0070] 本发明的搜索方法创造性地提出了在简历数据库和搜索客户端之间设置了搜索引擎，大幅度地提高了从客户端检索简历的速度，同时也将简历检索的精度大大提高。

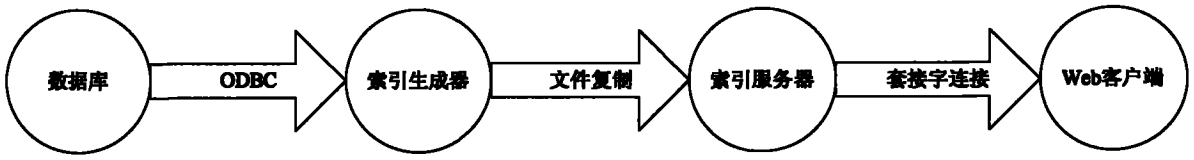

图 1

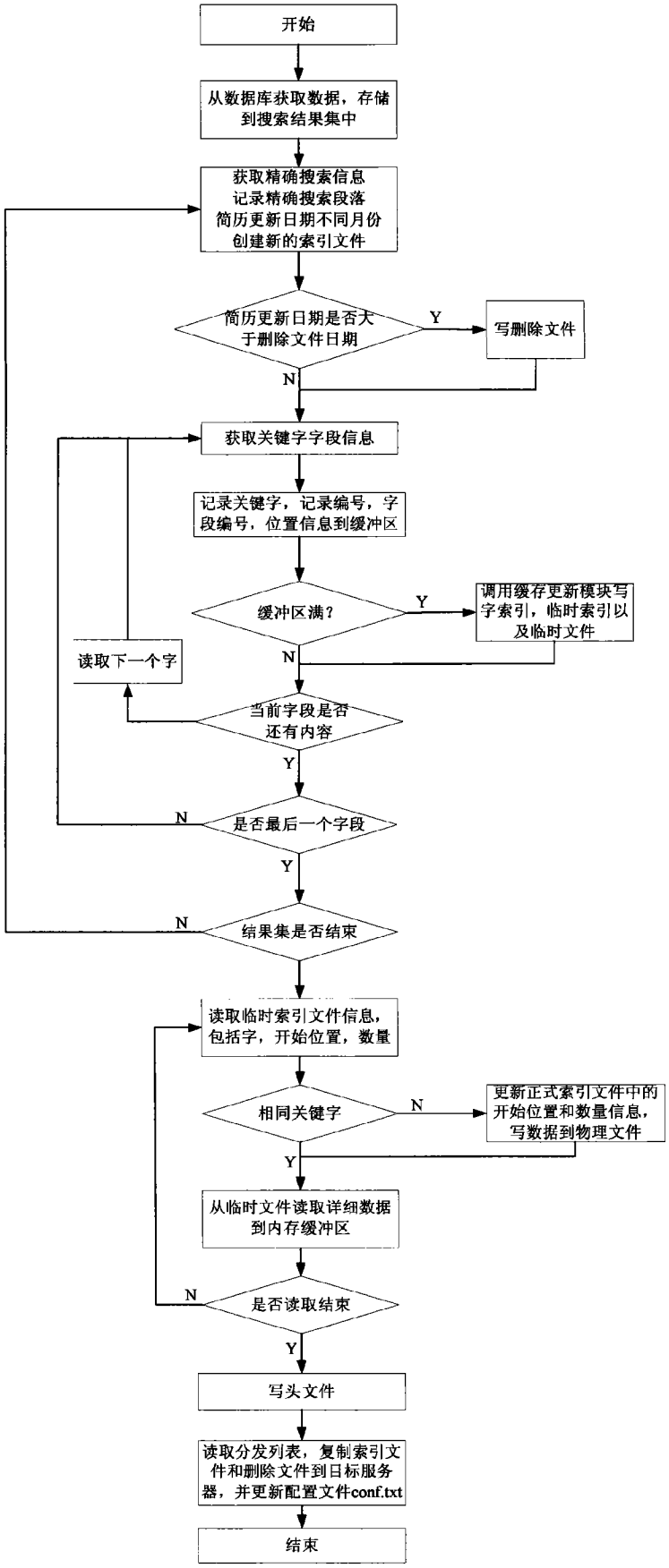
开始
从数据库获取数据，存储到搜索结果集中
获取精确搜索信息
记录精确搜索段落
简历更新日期不同月份
创建新的索引文件
简历更新日期是否大于删除文件日期
Y
写删除文件
N
获取关键字字段信息
记录关键字，记录编号，字段编号，位置信息到缓冲区
缓冲区满？
Y
调用缓存更新模块写字索引，临时索引以及临时文件
N
读取下一个字
当前字段是否还有内容
Y
是否最后一个字段
N
Y
结果集是否结束
N
读取临时索引文件信息，包括字，开始位置，数量
相同关键字
N
更新正式索引文件中的开始位置和数量信息，写数据到物理文件
Y
从临时文件读取详细数据到内存缓冲区
是否读取结束
N
Y
写头文件
读取分发列表，复制索引文件和删除文件到目标服务器，并更新配置文件conf.txt
结束

图 2

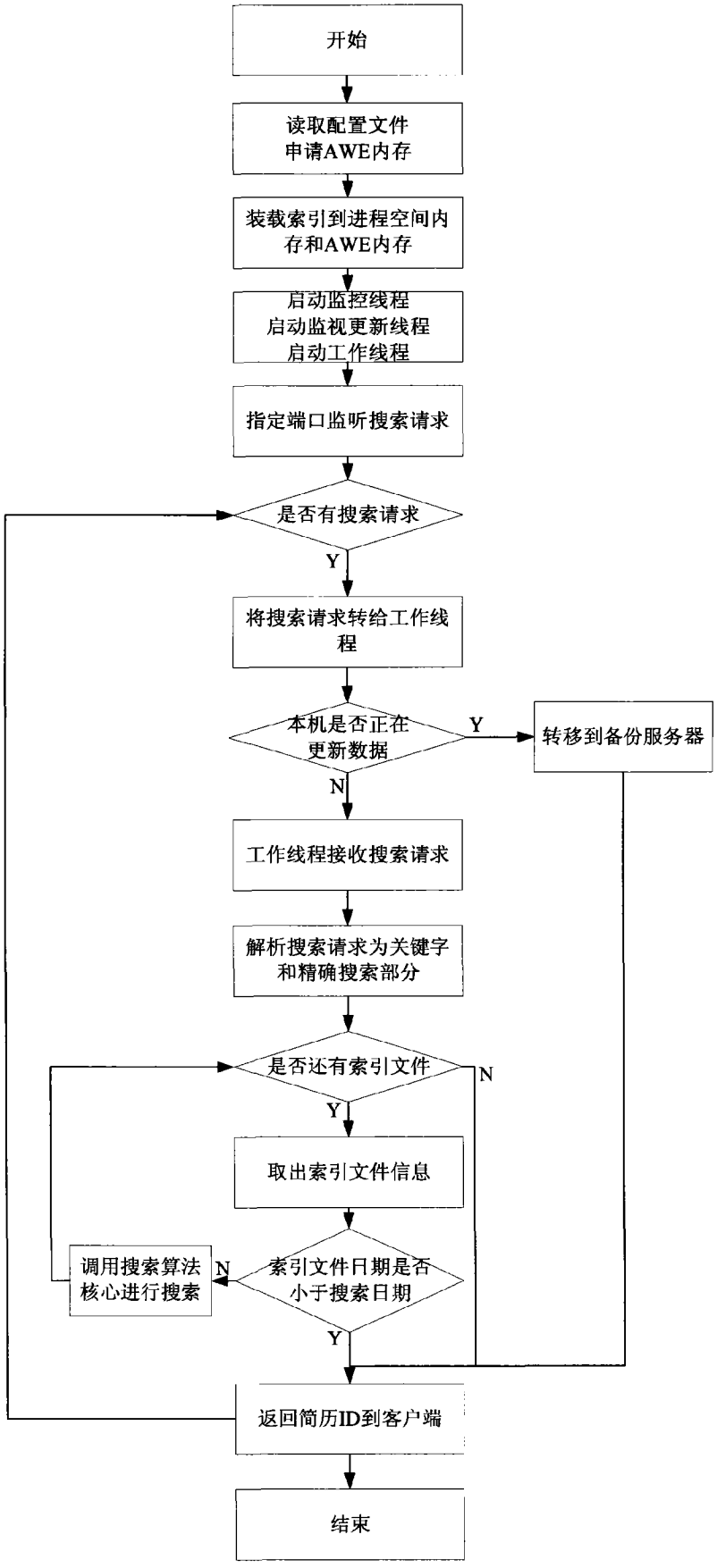
开始
读取配置文件
申请AWE内存
装载索引到进程空间内存和AWE内存
启动监控线程
启动监视更新线程
启动工作线程
指定端口监听搜索请求
是否有搜索请求
Y
将搜索请求转给工作线程
本机是否正在更新数据
Y
转移到备份服务器
N
工作线程接收搜索请求
解析搜索请求为关键字和精确搜索部分
是否还有索引文件
N
Y
取出索引文件信息
索引文件日期是否小于搜索日期
N
调用搜索算法核心进行搜索
Y
返回简历ID到客户端
结束

图 3

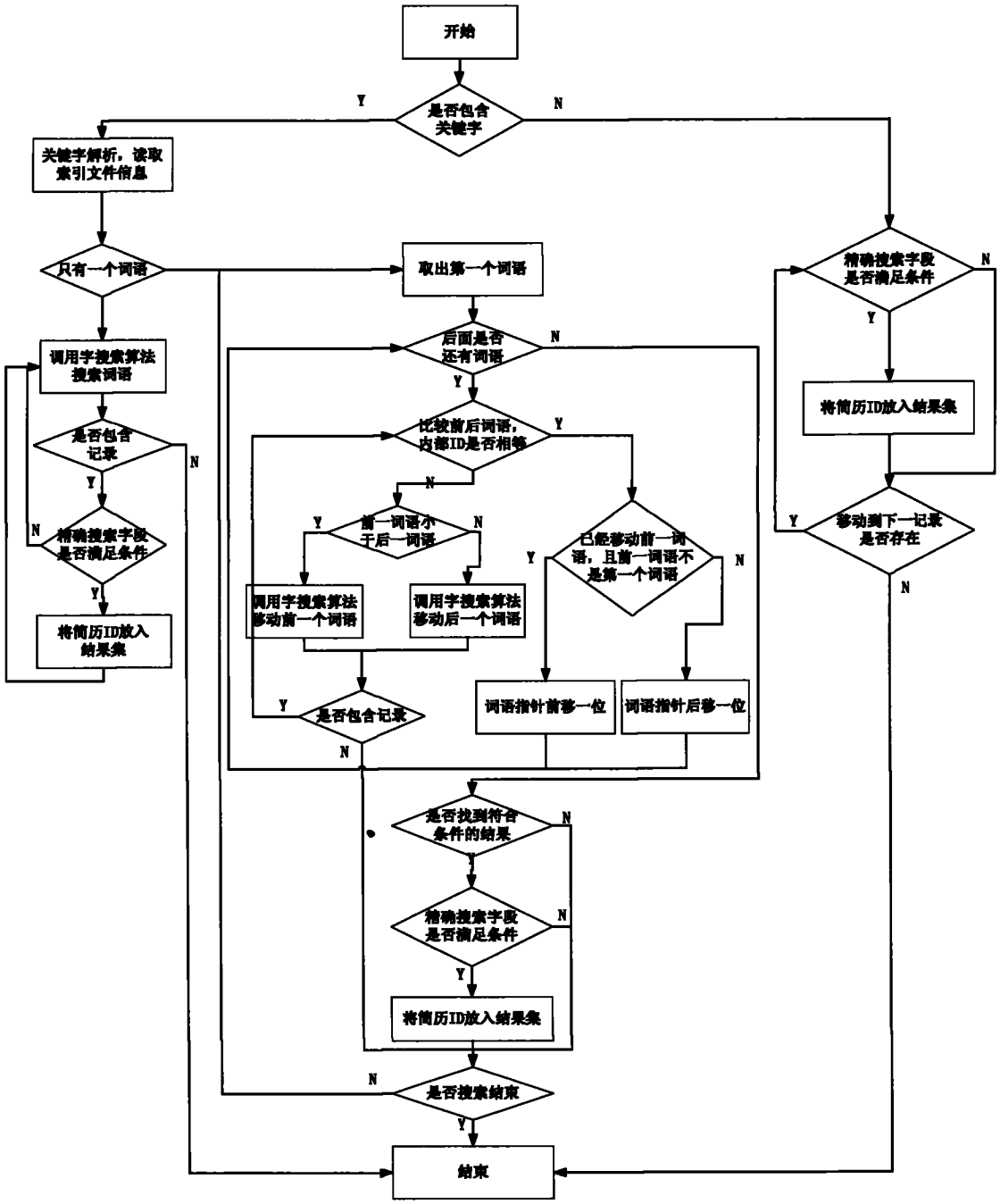
开始
是否包含关键字
Y
N
关键字解析，读取索引文件信息
只有一个词语
取出第一个词语
调用字搜索算法搜索词语
是否包含记录
精确搜索字段是否满足条件
将简历ID放入结果集
后面是否还有词语
比较前后词语，内部ID是否相等
前一词语小于后一词语
调用字搜索算法移动前一个词语
调用字搜索算法移动后一个词语
已经移动前一词语，且前一词语不是第一个词语
词语指针前移一位
词语指针后移一位
是否包含记录
是否找到符合条件的结果
精确搜索字段是否满足条件
将简历ID放入结果集
是否搜索结束
结束
精确搜索字段是否满足条件
将简历ID放入结果集
移动到下一记录是否存在

图 4

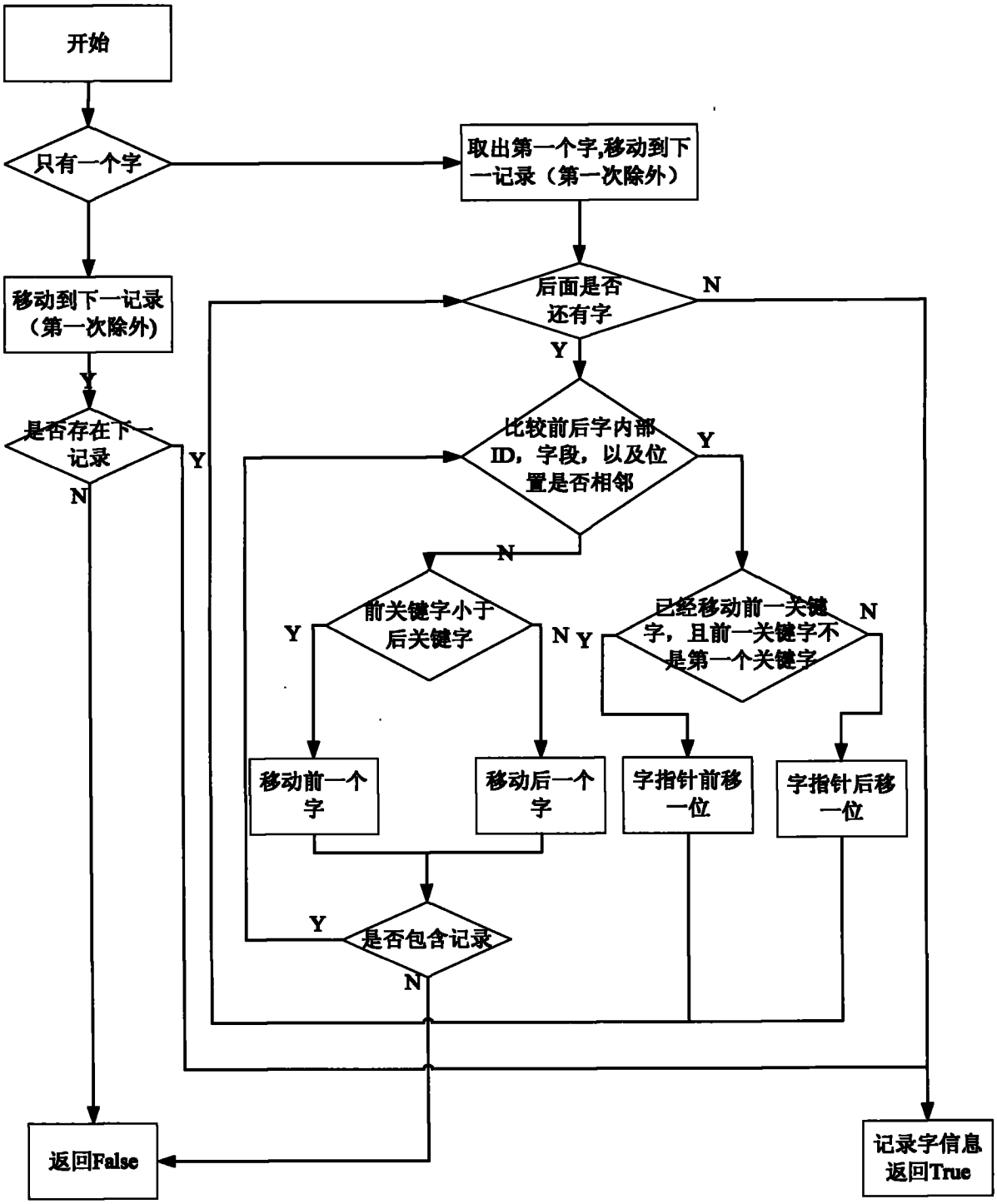
开始
只有一个字
取出第一个字,移动到下一记录（第一次除外）
移动到下一记录（第一次除外）
后面是否还有字
N
Y
Y
是否存在下一记录
Y
N
比较前后字内部ID，字段，以及位置是否相邻
Y
N
前关键字小于后关键字
Y
N
Y
已经移动前一关键字，且前一关键字不是第一个关键字
N
移动前一个字
移动后一个字
字指针前移一位
字指针后移一位
Y
是否包含记录
N
返回False
记录字信息返回True

图 5