(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 103678481 A
(43)申请公布日 2014.03.26

(21)申请号 201310495397.1

(22)申请日 2004.09.30

(30)优先权数据
10/675,234 2003.09.30 US

(62)分案原申请数据
200480030053.2 2004.09.30

(71)申请人 雅虎公司
地址 美国加利福尼亚州

(72)发明人 王学军 布赖恩·埃克坦
文卡特·潘查帕克森

(74)专利代理机构 北京东方亿思知识产权代理有限责任公司 11258
代理人 李晓冬

(51)Int.Cl.
G06F 17/30(2006.01)
G06Q 30/06(2012.01)

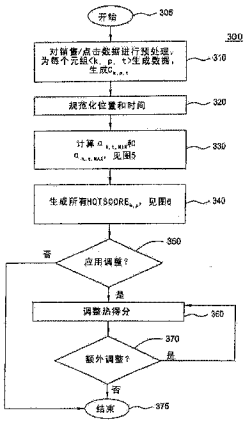
权利要求书2页 说明书12页 附图8页

(54)发明名称

用于搜索记分的方法和设备

(57)摘要

本发明涉及用于搜索记分的方法和设备。用于生成具有较高相关性的搜索结果的方法和设备。本发明利用了以下事实：用户对每个给定搜索词语的选择往往会覆盖来自若干个商家的若干个产品，并且所有结果都是与搜索词语非常相关的。在一个实施例中，这些结果被用于判定每个搜索词语的商家的顺序。通过获知用户的选择，尤其是从购买和/或点击信息(310)中获知用户的选择，比起仅限文本相关的产品来，高度相关并且最受欢迎的产品可以被分配以更高的得分或等级(350)。



CN 103678481 A

1. 一种搜索结果处理方法，包括：

收集与文档相关联的销售信息，其中所述文档被列在响应搜索词语的搜索结果集合内

根据所述销售信息并考虑所述文档在所述搜索结果集合的列出顺序内的位置，将所述文档与搜索结果集合内的其他文档进行比较，确定相关影响；

根据所述相关影响，生成所述文档的得分；

使用该得分影响之后的搜索的响应。

2. 如权利要求 1 所述的方法，其特征在于：

所述之后的搜索使用所述搜索词语，所述之后的搜索的响应包括一个搜索结果集合，所述搜索结果集合中文档的顺序被所述得分影响。

3. 如权利要求 1 所述的方法，还包括以下步骤：

调整所述得分以考虑到时间的流逝。

4. 如权利要求 1 所述的方法，还包括以下步骤：

调整所述得分以考虑到关于所述文档的特定知识。

5. 如权利要求 1 所述的方法，还包括以下步骤：

调整所述得分以考虑到关于所述搜索词语的特定知识。

6. 如权利要求 1 所述的方法，还包括以下步骤：

结合文本相关性得分、付费收录得分或付费赞助得分来应用所述得分。

7. 如权利要求 1 所述的方法，其中，所述文档是产品。

8. 如权利要求 1 所述的方法，其中，所述文档是目录页面。

9. 如权利要求 8 所述的方法，其中，所述目录页面代表产品显示页面，该产品显示页面显示许诺销售所述产品的多个商家。

10. 如权利要求 9 所述的方法，其中，所述目录页面还显示所述多个商家关于所述产品的价格信息。

11. 如权利要求 1 所述的方法，其中，所述生成所述文档的得分的步骤中，根据至少一个销售类型生成所述文档的得分。

12. 如权利要求 11 所述的方法，其中，所述至少一个销售类型包括代表通过优选商家进行的销售的优选商家销售类型。

13. 如权利要求 11 所述的方法，其中，所述至少一个销售类型包括代表通过非优选商家进行的销售的非优选商家销售类型。

14. 如权利要求 11 所述的方法，其中，所述至少一个销售类型包括代表来自有关搜索的通过优选商家进行的销售的有关搜索优选商家销售类型。

15. 如权利要求 11 所述的方法，其中，所述至少一个销售类型包括代表利用目录页面进行的销售的目录销售类型。

16. 如权利要求 15 所述的方法，其中，所述目录页面代表产品显示页面，该产品显示页面显示许诺销售所述产品的多个商家。

17. 如权利要求 11 所述的方法，其中，所述至少一个销售类型包括代表来自有关搜索的利用目录页面进行的销售的有关搜索目录销售类型。

18. 如权利要求 11 所述的方法，其中，所述至少一个销售类型包括代表与目录页面相关联的产品的销售的映射目录销售类型。

19. 如权利要求 11 所述的方法，其中，所述至少一个销售类型包括代表来自有关搜索的与目录页面相关联的产品的销售的有关搜索映射目录销售类型。

20. 如权利要求 11 所述的方法，还包括以下步骤：

为所述至少一个销售类型中的每一个计算配置参数，其中所述得分是根据所述配置参数和所述至少一个销售类型生成的。

21. 如权利要求 20 所述的方法，其中，所述得分是根据下式生成的：

$Hotscorek,p = \Sigma (\alpha k,t,T(t)Ck,p,t)$

其中，$C_{k,p,t}$ 是关于所述文档 p、针对所述搜索词语 k 的所述至少一个销售类型 t 的发生的数目，$\alpha_{k,t,T(t)}$ 是所述配置参数。

22. 如权利要求 1 所述的方法，其中，所述生成所述文档的得分，还包括配置选择步骤，通过选择的配置生成所述文档的得分。

23. 如权利要求 1 所述的方法，其中，所述销售信息包括至少一个与所述搜索词语相关的商家 / 商品标识对，返回的搜索结果集合中的每个商家 / 商品标识与商品被购买的结果相关。

24. 如权利要求 23 所述的方法，还包括以下步骤：

将所述商家 / 商品标识对分类为至少一个类型，并且清除低信用的商家 / 商品标识对。

25. 如权利要求 1 所述的方法，其中，所述生成所述文档的得分包括从多个公式中选择一个得分策略的公式计算得分，所述多个公式中的每个公式彻重于不同的得分策略。

26. 一种用于生成文档的得分的设备，所述设备用于：

收集与文档相关联的销售信息，其中所述文档被列在响应搜索词语的搜索结果集合内；

根据所述销售信息并考虑所述文档在所述搜索结果集合的列出顺序内的位置，将所述文档与搜索结果集合内的其他文档进行比较，确定相关影响；

根据相关影响，生成所述文档的得分。

27. 如权利要求 26 所述的搜索结果处理设备，所述设备还用于：

使用该得分影响之后的搜索的响应。

28. 如权利要求 27 所述的搜索结果处理设备，其中，

所述之后的搜索使用所述搜索词语，所述之后的搜索的响应包括一个搜索结果集合，所述搜索结果集合中文档的顺序被所述得分影响。

用于搜索记分的方法和设备

[0001] 本申请是申请日为 2004 年 9 月 30 日，申请号为 200480030053.2，名称为“用于搜索记分的方法和设备”的发明专利申请的分案申请。

技术领域

[0002] 本发明涉及用于对搜索结果进行记分或分级的方法和设备。更具体而言，本发明涉及基于事务和 / 或点击记录的记分方法。

背景技术

[0003] 随着因特网上大量信息的增殖，通常，如果不首先花大量时间来仔细察看许多不相关搜索结果就很难搜索和定位相关信息。根据所寻求的材料，用户常常由于必须查看许多无关紧要的搜索结果而感到受挫。

[0004] 记分或分级是搜索中的核心问题之一，例如在购物 / 产品搜索中尤其如此。如果搜索不能在搜索结果列表的顶部处提供最相关的文档，则这通常被称为不相关 (irrelevant)。比起常规 web 搜索来，对于诸如购物 / 产品搜索这样的搜索，用户往往具有更高的相关性 (relevancy) 要求，因为他们的目标不仅仅是找到一个相关结果。它们常常希望看到最相关的产品，并且希望能够在不同产品和不同商家之家进行比较。

[0005] 基于纯文本相关性的记分是若干搜索技术的基础。基本思想是找到匹配文档标题、描述和其他字段的文本。可以添加额外的细化，例如向某些字段 (比如标题) 提供更高的权重、向短语匹配提供更高的权重等等。但是，所有这些纯文本相关性记分方法都有生成最相关的搜索结果的问题，因为它们不能精确地确定用户想要搜索什么。

[0006] 例如，在纯文本相关性搜索中，当搜索词语“computer (计算机)”时，具有像“Sony VAIO FX340”这样的标题的文档不会被视为良好的文本匹配，因为标题不包含词语“computer”，而具有像“computer case (计算机壳)”这样的标题的文档却会被视为良好的匹配。这个示例证明了对 computer 的搜索很可能会产生具有许多不相关项目的搜索结果。

[0007] 即使在所有结果都被认为是相关的时，仍然优选向更受欢迎的产品提供更高的得分或等级。但是，纯文本相关性搜索将不能提供这种重要区别。

[0008] 因此，本领域中需要一种提供具有更高相关性的搜索结果的方法和设备。

发明内容

[0009] 在一个实施例中，本发明提供了一种用于生成具有更高相关性的搜索结果的方法和设备。例如，本发明提供了一种为购物 / 产品搜索生成具有更高相关性的搜索结果的方法和设备。

[0010] 本发明的一个前提是 :用户通过购买和 / 或点击其所喜爱的产品，从而针对受欢迎的搜索词语广播了其关于最喜爱的产品的偏好。当用户在购买 / 产品搜索站点中搜索一个词语时，虽然该站点可能返回许多不相关的结果，但是许多用户可以就通过选择其所感兴趣的结果 (即相关结果) 来过滤掉不相关的结果。这在用户确实从搜索结果列表中购买产

品时尤其精确，从而不仅指示了搜索词语的结果的相关性，还指示了所购买的产品的价格的相关性和 / 或销售所购买的产品的商家的相关性。

[0011] 本发明利用了以下事实：用户对每个给定搜索词语的选择往往会覆盖来自若干个商家的若干个产品，并且所有结果都是与搜索词语非常相关的。在一个实施例中，这些结果被用于判定每个搜索词语的商家的顺序。通过获知用户的选择，尤其是从购买和 / 或点击信息中获知用户的选择，比起仅限文本相关的产品来，高度相关并且最受欢迎的产品可以被分配以更高的得分或等级。

附图说明

[0012] 通过参考附图，从以下对本发明的优选实施例的详细描述中更好地理解前述和其他方面和优点，附图中：

[0013] 图 1 是示出本发明的记分系统的框图；

[0014] 图 2 示出应用本记分方法来影响搜索结果中的文档的列出顺序的关系；

[0015] 图 3 示出用于生成多个产品的热得分（hotscore）的方法的流程图；

[0016] 图 4 示出用于对销售和点击数据进行预处理的方法的流程图；

[0017] 图 5 示出用于计算配置参数 α 的方法的流程图；

[0018] 图 6 示出本发明的用于生成热得分的方法的流程图；

[0019] 图 7 示出本发明的用于调整热得分的方法的流程图；以及

[0020] 图 8 示出本发明的用于调整热得分的第二方法的流程图。

具体实施方式

[0021] 图 1 是示出本发明的记分系统 100 的框图。记分系统 100 的任务是为根据搜索词语生成的搜索结果集合内的文档（例如产品）记分。

[0022] 更具体而言，图 1 示出与网络（例如因特网 102）交互的记分系统 100，在该因特网 102 中，多个用户 105 被允许进行搜索。搜索通常由输入一个或多个搜索词语的用户所触发，所述搜索词语例如是“laptop computer”、“DVD”、“gas grill”等等。搜索可以包括对用户所需的产品和服务的搜索。产品和服务可以由维护记分系统 100 的实体提供，所述实体例如是操作一个提供大量产品和服务的网站的公司，例如 Walmart 之类的。或者，产品和服务可以由多个商家 107 提供，其中记分系统 100 是由第三方部署的，并且其任务只是生成与用户所提供的搜索词语相关联的搜索结果，例如搜索引擎应用。总之，本发明的记分系统 100 并不局限于其部署方式。

[0023] 在一个实施例中，记分系统 100 是用通用计算机或任何其他硬件等同物来实现的。更具体而言，记分系统 100 包括处理器（CPU）110、存储器 120（例如随机访问存储器（RAM）和 / 或只读存储器（ROM））、记分引擎或应用 122、搜索引擎或应用 124、跟踪引擎或应用 126 以及各种输入 / 输出设备 130（例如：存储设备（包括但不限于磁带驱动器、软盘驱动器、硬盘驱动器或紧致盘驱动器）、接收器、发送器、扬声器、显示器、输出端口、用户输入设备（例如键盘、小键盘、鼠标等等）或者用于捕捉语音命令的麦克风）。

[0024] 应当理解，记分引擎或应用 122、搜索引擎或应用 124 和跟踪引擎或应用 126 可以被实现为经由通信信道耦合到 CPU110 的物理设备或系统。或者，记分引擎或应用 122、搜索

引擎或应用 124 和跟踪引擎或应用 126 可以由一个或多个软件应用(或者甚至软件和硬件的组合,例如利用专用集成电路(ASIC))代表,其中软件从存储介质(例如磁或光驱动器或盘)被加载到计算机的存储器 120 中并被 CPU 所操作。这样,本发明的记分引擎或应用 122、搜索引擎或应用 124 和跟踪引擎或应用 126 (包括相关联的数据结构)可以被存储在计算机可读介质上,例如 RAM 存储器、磁或光驱动器或盘等等。

[0025] 总之,记分系统被设计以解决对提高搜索相关性的紧迫需求。本发明利用了以下事实:用户通过购买或点击其所喜爱的产品,从而针对受欢迎的搜索词语公开了其关于最喜爱的产品的偏好。当用户在购物 / 产品搜索站点搜索词语时,站点常会返回许多不相关的结果,而且这些不相关的结果甚至处于顶部结果位置。通常,用户只是过滤掉错误结果,而只选择其所感兴趣的结果,即相关的结果。当用户实际购买了从搜索结果中选择出的产品时,搜索结果的相关性被有效地证实。即,当用户决定购买产品时,则在产品的价格和 / 或销售产品的商家的上下文内,他或她选择的产品必然是与搜索词语高度相关的。

[0026] 已经确定,如果跟踪数据量充分大,则用户关于每个给定的搜索词语的选择往往会覆盖来自若干个商家的若干个产品,并且所有结果都与搜索词语非常相关。通过获知和应用用户的选择,尤其是来自购买和 / 或点击的选择,比起仅限文本相关的产品来,高度相关的产品可以被分配以更高的得分 / 等级。这种新颖的方法将会对搜索词语产生高度相关的搜索结果。实际上,可以应用额外的细化或规范化(normalization),例如针对每个搜索词语的商家排序。这些可选的调整在下文中进一步描述。

[0027] 在本发明的一个实施例中,响应于搜索词语分配给产品的基于用户购买和 / 或点击信息的得分被称为“热得分”。这个热得分可以被搜索引擎用于响应于搜索词语产生搜索结果。应当注意,当前的热得分可以被用作生成搜索结果时的主导(权重更重)参数,或者用于为当前采用其他参数作为主导参数的搜索引擎提供补充,所述其他参数例如是付费收录(paid inclusion)、付费赞助、文本相关性。

[0028] 图 2 示出应用本记分方法来以较大的相关性影响搜索结果集合中的文档列表的关系。图 2 示出响应于特定搜索词语而生成并提供给用户的第一结果集合 220。在该示例中,搜索结果集合中的项目被广泛地定义为文档,其中在购物的场景内,文档应当是产品或产品 - 商家对。但是,文档想要广泛地包括网站、文本文档、图像等等。

[0029] 图 2 示出通过跟踪第一搜索结果集合内的各种文档的购买和 / 或点击 210 来跟踪用户对第一结果集合 220 的响应。该购买和 / 或点击信息被跟踪,然后被记分过程 230 用来生成多个得分(热得分)240,其中每个得分与文档之一相关联。热得分 240 又可选地被另一记分系统 250 用来响应于生成第一结果集合的同一搜索词语生成第二搜索结果集合 260,所述记分系统 250 可以结合文本得分 252 和其他得分 254 (例如付费收录得分)来应用热得分。图 2 示出热得分的应用现在已经影响了文档的排序,并且还可能影响第二结果集合中文档的添加或删除,从而在第二搜索结果集合中提供更好的相关性。

[0030] 在一个实施例中,对于每个搜索词语,本发明跟踪每个用户点击并最终购买的商家 / 产品对。更详细的信息也被跟踪,其中包括当点击 / 购买发生时产品在搜索结果中的位置、这种行为发生的时间以及当这种行为发生时产品被分配的部门。

[0031] 图 3 示出用于生成多个产品的热得分的示例性方法 300 的流程图。方法 300 开始于步骤 305 中,并且进行到步骤 310。

[0032] 在步骤 310 中，方法 300 根据特定搜索词语对每个产品的销售和 / 或点击数据进行预处理。例如，本发明为每个元组 <k,p,t> 生成数据，其中 k 是搜索词语，p 是产品，t 是类型。即，方法 300 将会生成 $C_{k,p,t}$，该 $C_{k,p,t}$ 是在“tp”时间段中发生的针对搜索词语 k 的 t 类事件的计数或数目。t 类事件可以定义特定类型的购买事件和 / 或点击事件（例如，对来自优选卖方的产品的购买或对搜索结果上的文档的点击）。多个示例性类型事件在下文中公开。

[0033] 具体而言，对于可以在配置文件中定义和调节的给定时间范围，每个搜索词语的所有商家 / 产品 -id 对被分类成不同类型，并且被基于 $C_{k,p,t}$ 计数。此外，消除低置信度的结果。低置信度的结果可以包括兜售信息（spamming）结果和分散的结果。分散的结果是在给定阈值下重复的结果，例如偶然被访问而实质上并不指示链接的相关性的链接。

[0034] 在步骤 320 中，方法 300 可选地对数据进行规范化以考虑到时间和 / 或位置。具体而言，已观察到，产品在搜索结果集合中的位置“越高”，它被用户点击 / 购买的概率就越高。更具体而言，还观察到，点击很受位置影响（例如位置较高的产品常被“点击”），而购买则略受位置影响（例如，购买者只被相关产品的位置略微影响）。从而，用户可以点击位置较高的产品，但是最后却可能由于相关性而购买被列在低得多的位置处的产品。

[0035] 搜索结果集合中的第一顶部位置被视为位于搜索结果集合内的最高位置。为了找到更相关的结果，基于点击 / 购买发生时的位置来规范化商家 / 产品 -id 对的置信度。例如，对结果集合内位置非常低的文档的购买或点击将会指示该文档与搜索词语的高度相关性。

[0036] 可选地，可以对数据进行规范化以考虑到时间（“出现时间”或“发生时间”）。即，文档的销售和 / 或点击离目前有多长时间。虽然商家 / 产品 -id 对的“发生时间”不应当影响该对的相关性，但是它确实有可能或者潜在地影响市场中的新趋势。捕捉这个趋势并且总将最受欢迎的结果显示在第一位是本记分发明的目标之一。换言之，可以按考虑到产品的受欢迎度或“时间相关性”的顺序来列出相关产品。可以部署用于位置和时间规范化的各种规范化函数。

[0037] 在步骤 330 中，方法 300 计算配置参数 α。更具体而言，方法 300 为每个 <k,t> 对计算 $\alpha_{k,p,MAX}$ 和 $\alpha_{k,p,MIN}$。配置参数被用于定义不同类型的购买和 / 或点击的影响。例如，通过商店（例如被视为非优选的小型商家）进行的购买不同于通过目录（例如被视为优选的大型商家）进行的购买。类似地，通过“优选商家”进行的购买不同于向“一般商家”进行的购买。这些区别对于本记分系统的操作者是很重要的，因为这种与购买和点击类型有关的信息可以被用于进一步细化搜索结果的相关性，如下所述。

[0038] 在步骤 340 中，方法 300 基于购买和 / 或点击信息为针对每个搜索词语的每个产品生成得分（热得分）。这个得分可以以下文进一步公开的多种不同方法来生成。即，可以应用不同的规则以与公司的策略相对应。从而，在一个规则中计算出的商家 / 产品 -id 对的热得分可能不同于在第二规则中计算出的。

[0039] 在步骤 350 中，方法 300 查询是否有必要调整热得分。具体而言，可以任选地应用调整，以考虑到不同的知识，例如关于搜索词语的特定知识、关于商家 - 产品对的性能的知识、关于购买者行为的知识、关于购买者年龄的知识、关于购买者性别的知识等等。如果这种知识可用，则可以相应的调整热得分。

[0040] 例如，可以基于受欢迎的搜索词语来对热得分做出调整。对于包含在知识库中的某些受欢迎的搜索词语，本发明可以向搜索词语添加销售信息。例如，在一个实施例中，搜索词语“dell”可以被翻译成“manufacturer=Dell”，其中本发明可以将关于“manufacturer=Dell”的所有销售信息应用到搜索词语“dell”。

[0041] 或者，可以基于用户对有关搜索词语的行为来对热得分做出调整。用户对有关搜索的行为可以帮助创建一般搜索词语与和它有关的较窄的搜索词语之间的实际联系。即，这将会帮助用户使其搜索缩小到一般搜索词语上。在一个实施例中，本发明将商家 / 产品对的有关搜索词语的热得分添加到一般搜索词语，从而扩展了覆盖范围。

[0042] 或者，如果数据指示正在执行商家 - 产品对的匹配，则可以对热得分做出调整，即调整热得分以降低不正确的或不受欢迎的文档的得分的影响。例如，在热得分被分配给商家 - 产品对之后，本发明继续对结果进行估计。性能不佳的对被假定为是搜索结果集合的错误选择的文档或不受欢迎的文档，从而其热得分将会被降低。例如，搜索结果可以提供多个相关文档(例如与搜索词语高度相关的商家 - 产品对)，但是由于种种原因，购买者对商家 - 产品对中的特定子集不感兴趣。在这种情形下，这种相关的、但却不受欢迎的产品对被“惩罚”，从而使它们将会具有较低的、甚至负的热得分。

[0043] 返回步骤 350，如果对查询的回答是否定的，则方法 300 在步骤 375 中结束。如果对查询的回答是肯定的，则方法 300 进行到步骤 360，在该步骤中热得分被调整。

[0044] 在步骤 370 中，方法 300 查询是否有必要对热得分进行额外的调整。如果对查询的回答是肯定的，则方法 300 进行到步骤 360，在该步骤中热得分再次被调整。如果对查询的回答是否定的，则方法 300 在步骤 375 中结束。

[0045] 一旦热得分被生成，搜索引擎 124 就可以立即应用热得分以影响购物 / 产品搜索。在一个实施例中，利用当前的热得分实时地(on the fly)调整基于任何搜索方法的搜索记分。例如，当用户键入搜索词语时，购物 / 产品搜索系统将会向搜索引擎发出具有热得分提高比率的搜索。这个比率可以非常高，这意味着所有具有热得分的产品都将会在那些没有热得分的产品的前面。它也可以非常低，这意味着热得分只会最低限度地影响搜索结果的顺序。

[0046] 图 4 示出用于对销售和点击数据进行预处理的方法 400 的流程图。方法 400 开始于步骤 405 中，并且进行到步骤 410。

[0047] 在步骤 410 中，方法 400 查询点击信息是否是关于产品的实际销售的。如果对查询的回答是肯定的，则方法 400 进行到步骤 492，在该步骤中原始点击信息被使用。即，产品的销售就搜索结果的相关性而言提供了最高的置信度。从而，与销售相关联的点击信息被保留并使用。如果对查询的回答是否定的，则方法 400 进行到步骤 420。

[0048] 在步骤 420 中，方法 400 查询点击信息是否低于预定阈值。如果对查询的回答是肯定的，则方法 400 进行到步骤 430。如果对查询的回答是否定的，则方法 400 进行到步骤 494，在该步骤中点击信息被丢弃。即，步骤 420 的意图是去除错误的点击数据，例如人为地增加对搜索结果内的特定文档的访问的泛滥式攻击。

[0049] 在步骤 430 中，方法 400 查询点击信息是否来自受信站点。如果对查询的回答是肯定的，则方法 400 进行到步骤 492，在该步骤中原始点击信息被使用。即，来自受信站点的产品的点击信息就搜索结果的相关性而言提供了一些置信度。从而，点击信息被保留和使

用。如果对查询的回答是否定的，则方法 400 进行到步骤 440。

[0050] 在步骤 440 中，方法 400 查询来自特定 IP 地址的点击信息是否多于来自其他 IP 地址的点击信息。换言之，从统计上而言，与特定 IP 地址相关联的点击信息与来自其他 IP 地址的点击信息相比是否反常地高。如果对查询的回答是肯定的，则方法 400 进行到步骤 450，在该步骤中来自特定 IP 地址的点击信息被丢弃。即，来自特定 IP 地址的点击信息是可疑的。如果对查询的回答是否定的，则方法 400 进行到步骤 460。

[0051] 在步骤 460 中，方法 400 查询点击和页面查看速率是否远高于平均比率。如果对查询的回答是肯定的，则方法 400 进行到步骤 470，在该步骤中点击信息被丢弃。即，如果点击和页面查看的速率或频率非常高，即用户点击一个文档然后立即点击另一个文档，而花在查看最初点击的页面的时间却非常少，则点击信息是可疑的。如果对查询的回答是否定的，则方法 400 进行到步骤 480。

[0052] 在步骤 480 中，方法 400 查询搜索结果集合内的文档的点击数目是否远高于关于同一搜索词语的同一搜索结果集合中的其他文档的点击数目。例如，如果搜索结果集合内的一个特定文档被重复访问的次数远高于同一搜索结果集合中的其他文档，则点击信息是可疑的。前提是以下情况是非常反常的：用户重复点击某个文档的频率远高于点击同一搜索结果中的其他文档的频率。如果对查询的回答是否定的，则方法 400 进行到步骤 492，在该步骤中原始点击信息被使用。

[0053] 如果对查询的回答是肯定的，则方法 400 进行到步骤 490，在该步骤中点击信息的平均被使用。方法 400 在步骤 495 中结束。

[0054] 图 5 示出用于计算类型的配置参数 α 的方法 500 的流程图。更具体而言，方法 500 为每个 <k,t> 对计算 $\alpha_{k,p,MAX}$ 和 $\alpha_{k,p,MIN}$。配置参数被用于描述不同类型的购买和 / 或点击的影响。方法 500 开始于步骤 505 中，并且进行到步骤 510。

[0055] 在步骤 510 中，方法 500 选择元组 <k,t>，其中 k 是搜索词语，t 是类型。然后在步骤 520 中，方法 500 为 <k,t> 选择 $C_{k,p,t}$，其中 k 是搜索词语，p 是产品，t 是类型。即，$C_{k,p,t}$ 是在某个时间段中发生的关于产品 p、针对搜索词语 k 的 t 类事件的计数或数目。

[0056] 在步骤 530 中，方法 500 计算配置参数 α。更具体而言，α 可以被表达为：

[0057] $\alpha_{k,t,MIN}=m_t$ （方程 1）

[0058] $\alpha_{k,t,MAX}=m_t/MAX(C_{k,1,t},C_{k,2,t},\ldots,C_{k,n,t})$ （方程 2）其中 m_t 是 t 类事件的基本得分，如下表 1 和 2 所示，这两个表是基于两个不同的业务要求来定义的。应当注意，对于每个 t 类事件，可以采用方程 1 和方程 2 中的“min（最小）”或“max（最大）”函数中的任何一个，如下所示。

[0059]

类型	m_t
最小优选商家销售：	150
最小有关搜索优选商家销售：	120
最大优选商家点击：	100

最大非优选(商店)销售：	80
最小目录销售：	600
最小有关搜索目录销售：	500
最小映射目录销售：	550
最小有关搜索映射目录销售：	450
最大映射目录点击：	160
最小基于知识的销售：	580

[0060] 表 1

[0061]

类型	m_t
最小优选商家销售：	110

[0062]

最小有关搜索优选商家销售：	105
最大优选商家点击：	100
最大非优选(商店)销售：	105
最小目录销售：	600
最小有关搜索目录销售：	500
最小映射目录销售：	550
最小有关搜索映射目录销售：	450
最大映射目录点击：	160
最小基于知识的销售：	550

[0063] 表 2

[0064] 应当注意，分配给各种类型的销售和点击的值 m_t 可以被调整以针对特定实现方式。以下类型定义如下：

[0065] 优选商家销售被定义为通过优选商家进行的销售。将商家定义为优选商家的标准是应用特定的，例如，向搜索实体付费的商家可以被视为优选商家。

[0066] 有关搜索优选商家销售被定义为这样的销售：该销售是利用与搜索词语有关但包括优选商家的名称的搜索词语进行的。为了说明，假设有两个搜索词语：“digital camera”

和“Sony digital camera”。对来自从搜索词语“Sony digital camera”生成的搜索结果的产品“A”的购买将会导致表 1 所示的 120 的 m_t 被添加到产品“A”的得分，而对来自从搜索词语“digital camera”生成的搜索结果的产品“A”的购买将会导致表 1 所示的 150 的 m_t 被添加到产品“A”的得分。这种方法将较窄的搜索“Sony digital camera”与更宽、更一般化的搜索词语“digital camera”联系起来。

[0067] 优选商家点击被定义为对与优选商家相关联的搜索结果集合内的文档的点击。

[0068] 非优选销售被定义为通过非优选商家(例如小型商家)进行的销售。将商家定义为非优选商家的标准是应用特定的，例如，向搜索实体提供很少费用或者不提供费用的小型商家可以被视为非优选商家。

[0069] 目录销售被定义为利用目录页面或产品指南页面进行的销售。目录页面被定义为特定产品的显示页面，其显示以下信息中的一种或多种：商家列表、商家 - 价格对(例如以特定价格许诺销售产品的商家)的列表、产品评论列表、产品描述等等。从该目录页面进行的购买被假定为与搜索词语高度相关。

[0070] 有关目录销售被定义为利用有关目录页面或产品指南页面进行的销售。为了说明，假设有两个两个搜索词语：“digital camera”和“Sony digital camera”。对来自从搜索词语“Sony digital camera”生成的目录页面的产品“A”的购买将会导致表 1 所示的 500 的 m_t 被添加到针对搜索词语“digital camera”的产品“A”的得分，而对来自从搜索词语“digital camera”生成的目录页面的产品“A”的购买将会导致表 1 所示的 600 的 m_t 被添加到产品“A”的得分。

[0071] 映射目录销售被定义为与映射的目录页面或产品指南页面相关联的销售。即，购买不是从目录页面进行的，而是直接经由商家的页面进行的。例如，特定搜索词语的搜索结果包含多个目录页面和多个商家页面。用户随后选择访问特定商家页面，于是直接通过商家进行产品购买。从而，检测到产品购买是直接从特定商家进行的，并且如果系统还检测到所购买的产品被“映射”到特定目录页面或产品指南页面，则购买信息将会导致表 1 所示的 550 的 m_t 被添加到目录页面得分。应当注意，为文档广泛生成热得分，其中文档可以包括产品、商家 - 产品对或目录页面。向相关目录页面分配高得分是合乎需要的，这是因为用户被提供以许诺销售同一产品的商家的比较。换言之，在目录页面中购买产品是理想的购物环境，其中高热得分的分配将会导致目录页面被频繁地提供给用户。

[0072] 有关搜索映射目录销售被定义为与有关映射目录页面或有关映射产品指南页面相关联的销售。

[0073] 映射目录点击被定义为对可被映射到目录页面或产品指南页面的商家页面的点击。即，该点击不是对目录页面做出的，而是直接对商家的页面做出的。例如，特定搜索词语的搜索结果包含多个目录页面和多个商家页面。用户随后选择为了某个产品而点击特定商家页面。如果系统还检测到被点击的产品被“映射”到特定目录页面或产品指南页面，则点击信息将会导致表 1 所示的 160 的 m_t 被添加到目录页面的得分。

[0074] 基于知识的销售被定义为利用基于关于搜索词语的某些知识而调整的结果进行的销售。例如，如果搜索词语是“sony”，则搜索词语被调整为“brand=Sony”。来自这种搜索结果的产品的销售将会导致被购买的产品接收表 1 所示的 580 的 m_t。

[0075] 返回图 5，在步骤 540 中，方法 500 查询是否已经例如根据以上所示的方程 2 计算

了所有 $C_{k,p,t}$。如果对查询的回答是否定的，则方法 500 返回步骤 520。如果对查询的回答是肯定的，则方法 500 进行到步骤 550。

[0076] 在步骤 550 中，方法 500 查询是否所有元组 <k,t> 都已经被总结。如果对查询的回答是否定的，则方法 500 返回步骤 510。如果对查询的回答是肯定的，则在步骤 555 中方法 500 结束。

[0077] 图 6 示出了本发明的用于生成热得分的方法 600 的流程图。方法 600 开始于步骤 605 中，并且进行到步骤 610。

[0078] 在步骤 610 中，方法 600 任选地查询特定配置是否已被选择用于生成热得分。即在一个实施例中，可以部署多个配置或规则来针对不同的系统需求。例如，某些系统可能赞成热得分的使用，从而导致 MAX 配置被选择，其中热得分将会对搜索结果集合中列出的文档有很大影响。或者，某些系统可能希望减轻热得分的使用，从而导致 MIN 配置被选择，其中热得分对搜索结果集合中列出的文档的影响将会较小。

[0079] 但是，如果没有设想多个配置，则可以省略步骤610，并且选择标准配置。如果对查询的回答是否定的，则方法 600 进行到步骤 615，在该步骤中选择配置。如果对查询的回答是肯定的，则方法 600 进行到步骤 620。

[0080] 在步骤 620 中，方法 600 选择元组 <k,p>，其中 k 是搜索词语，p 是产品。然后，在步骤 630 中，方法 600 选择类型 t。

[0081] 在步骤 640 中，方法 600 查询 <k,p,t> 的 $C_{k,p,t}$ 是否存在，其中 k 是搜索词语，p 是产品，t 是类型。$C_{k,p,t}$ 是在某个时间段中发生的关于产品 p、针对搜索词语 k 的 t 类事件的计数或数目。如果对查询的回答是否定的，则方法 600 返回步骤 630，在该步骤中另一类型被选择。如果对查询的回答是肯定的，则方法 600 进行到步骤 650。

[0082] 在步骤 650 中，方法 600 根据所选择的配置计算配置因子 α。在一个实施例中，对于搜索词语 k，商家 / 产品对 p 的热得分被定义为：

[0083] $Hotscore_{k,p}=\sum(\alpha_{k,t,T(t)}C_{k,p,t})$ （方程 3）其中 $C_{k,p,t}$ 是关于产品 p、针对搜索词语 k 的 t 类事件的发生次数。$\alpha_{k,t,T(t)}$ 是在以上方程 2 和方程 3 中定义的配置因子。

[0084] 在一个实施例中，可以定义 T(t) 函数，例如，其中 T(t) 可以是 MAX 函数或 MIN 函数。这些函数的值的示例在以上表 1 或表 2 中示出。T(t) 函数的值可以在记分系统的配置中预先定义。虽然本发明公开了两个配置函数 MAX 和 MIN，但是本发明并不局限于此。即，可以部署任何数目的配置以针对特定记分系统的需求。

[0085] 在步骤 660 中，方法 600 查询是否所有类型 t 都已经被处理。如果对查询的回答是否定的，则方法 600 返回步骤 630，在该步骤中另一个类型被选择。如果对查询的回答是肯定的，则方法 600 进行到步骤 670，在该步骤中方程 3 被用于生成所选择的元组 <k,p> 的热得分。

[0086] 在步骤 680 中，方法 600 查询是否所有元组 <k,p> 都已经被处理。如果对查询的回答是否定的，则方法 600 返回步骤 620，在该步骤中另一个元组被选择。如果对查询的回答是肯定的，则在步骤 685 中方法 600 结束。

[0087] 在一个实施例中，当前热得分被用于现有搜索记分系统中。为了说明，对于搜索词语 t、商家 / 产品对 p 按下式获得 $score_{k,p}$：

[0088] $Score_{k,p}=BT_{k,p}+H(hotscore_{k,p})+OB_{k,p}$ （方程 4）其中 $BT_{k,p}$ 是产品 p 针对搜索词语

k 获得的基本文本相关性得分，$hotscore_{k,p}$ 是 p 针对搜索词语 k 的热得分，H 是用于调整搜索记分系统的热得分的使用函数(如果必要的话)，$OB_{k,p}$ 是搜索词语 k 的其他可选提高性得分之和。应当注意，H 是描述热得分应当如何被用于整个得分中的函数，如下所述。

[0089] 可以采用许多种规范化函数。以下给出各种类型的函数。

[0090] 在一个实施例中，利用如下“影响因子”来规范化原始热得分：

[0091] $H(hotscore_{k,p})=hotscore_{k,p}*af$ (方程 5) 其中 af 被称为影响因子，它可以被定义如下：

[0092] af=standard_hotscore/standard_score_for_hotscore_in_whole score

[0093] (方程 6)

[0094] 该函数选择热得分中的得分作为标准，并且选择整体得分中的得分作为热得分部分的标准得分。然后通过使用影响因子将热得分应用到整体记分中。在这种方法中，对于热得分的使用没有上限或下限。从而，置信度非常高的产品可以被保证拥有高等级。

[0095] 在第二实施例中，可以按下式来规范化热得分：

[0096] 如果 $hotscore_{k,p}=0$，则 $H(hotscore_{k,p})=0$；

[0097] 否则， (方程 7)

[0098] $H(h_{k,p})=H_L+(H_U-H_L)*(h_{k,p}-MIN(h_{k,1},h_{k,2},\ldots,h_{k,n})/MAX(h_{k,1},h_{k,2},\ldots,h_{k,n})-MIN(h_{k,1},h_{k,2},\ldots,h_{k,n}))$

[0099] 其中 H_L 是总得分中热得分的下限，H_U 是总得分中热得分的上限。函数 H 判定热得分在搜索记分中的作用有多重要。H_U 定义热得分在得分中的最大影响，H_L 定义热得分在得分中的最小影响。

[0100] 一种极端的方案是为 H_U 和 H_L 赋予非常大的值，从而使得热得分将会主导整个得分。或者，另一个极端是为 H_U 和 H_L 赋予非常小的值，从而使得热得分只影响具有相同的方程 4 的 $BT_{k,p}$ 和 $OB_{k,p}$ 的产品的等级。前一种方法适用于闭合系统，其中所有事务信息都可用。对于其中只有某些销售信息可用的开放系统，仅向 H_U 赋予较高的值以使得置信度高的热得分主导得分，而置信度低的热得分只起非常有限的作用，并且与其他记分影响相混合，则将会是更加适当的。

[0101] 在第三实施例中，可以对热得分进行位置规范化。具体而言，令 AC_i 为位置 i 处的所有点击数目，$C_{k,p,i}$ 为位置 i 处针对搜索词语 k 的产品 p 的点击数目，$NC_{k,p,i}$ 为规范化后的位置 i 处针对搜索词语 k 的产品 p 的点击数目，从而：

[0102] $NC_{k,p,i}=C_{k,p,i}*AC_0/AC_i$ (方程 8) 其中 AC_o/AC_i 被称为位置 i 的常规提高因子。为了抑制对搜索结果集合内的位置很高的文档的点击的影响，本方法可以将 AC_i 限制到某个数字，例如 AC_{30}，从而使得对高位置一次错误点击不会不成比例地影响整个记分系统。

[0103] 此外，由于在不同的时间 <k,p> 对的点击位置可能不同，因此通过计算给定时间段中 <k,p> 的平均点击位置来确定 i。

[0104] 该函数将一个 <k,p> 对的一个位置上的点击数目与平均点击数目相比较。只有那些优于正常点击率的才能在规范化后拥有较高的数字，即它实际上将 $C_{k,p,0}/C_{k,p,i}$ 与 AC_0/AC_i 相比较。从而，这种方法将会使自提高(self-boosting)的概率最小化。应当注意，同样的函数也可以被应用到销售位置规范化。

[0105] 在第四实施例中，可以对热得分进行时间规范化。具体而言，令 E 为事件发生的次

数，NE 为事件的规范化次数，age 为事件发生距离当前时间的天数，ff 为“遗忘因子”，即系统倾向于遗忘某个事件的比率。遗忘因子被定义在配置文件中，以便本系统可以相应的调节它。E 的规范化如下：

[0106] $NE=E*(1-ff)^{age}, (0<=age<=n)$ （方程 9）方程 9 中的“age”的上限（n）可以被调整以满足特定应用或不同产品的需求。

[0107] 图 7 示出本发明的用于基于知识参数调整热得分的方法 700 的流程图。方法 700 开始于步骤 705 中，并且进行到步骤 710。

[0108] 在步骤 710 中，方法 700 从知识库中选择搜索词语 k。即，取得知识 KN_k。例如，如果搜索词语是“dell”，则知识 KN_k 可以被表达为“manufacturer=Dell”。

[0109] 在步骤 720 中，方法 700 查询是否存在关于知识 KN_k 的应用的配置因子或规则。例如，配置因子可以规定所 Dell 产品的热得分都被调整以考虑到所有 Dell 产品的销售。或者，配置因子可以规定所有 Dell 计算机产品的热得分都被调整以考虑到所有 Dell 计算机产品的销售，等等。如果对查询的回答是否定的，则方法 700 返回步骤 710，并且另一个搜索词语被选择。如果对查询的回答是肯定的，则方法 700 进行到步骤 730。

[0110] 在步骤 730 中，方法 700 取得与每个产品的知识 KN_k 有关的所有销售信息 $(P_{KNk}1),...,(P_{KNk}n)$。例如，收集关于桌面型计算机、笔记本电脑、PDA、打印机、监视器、扬声器等的销售信息。下面可以应用这种信息。

[0111] 在步骤 740 中，方法 700 可以任选应用如上所述的时间和位置规范化。

[0112] 在步骤 750 中，方法 700 从步骤 730 中所述的产品中选择产品 p。例如，Dell 桌面型计算机被选择。

[0113] 在步骤 760 中，方法 700 基于步骤 720 中所述的配置因子或规则调整 $hotscore_{k,p}$。例如，Dell 桌面型计算机的热得分被调整，从而使得 Dell 笔记本电脑的销售信息被用于提高 Dell 桌面型计算机的热得分。这种调整之所以合理可能因为是 Dell 是优选商家，或者存在关于优选 Dell 笔记本电脑的购买者也会优选 Dell 桌面型计算机的知识。这样一来，可以利用特定的知识来进一步细化热得分。

[0114] 在步骤 770 中，方法 700 查询是否已经调整了所有有关产品。如果对查询的回答是否定的，则方法 700 返回步骤 750，并且另一个产品被选择。如果对查询的回答是肯定的，则方法 700 进行到步骤 780。

[0115] 在步骤 780 中，方法 700 查询是否所有有关知识都已经被处理。如果对查询的回答是否定的，则方法 700 返回步骤 710，并且另一个搜索词语被选择。如果对查询的回答是肯定的，则在步骤 785 中方法 700 结束。

[0116] 图 8 示出本发明的用于基于有关较窄搜索来调整热得分的方法 800 的流程图。方法 800 开始于步骤 805 中，并且进行到步骤 810。

[0117] 在步骤 810 中，方法 800 查询是否存在关于有关较窄搜索的应用的配置因子或规则。例如，搜索词语“computer with SDRAM”将会被视为“computer”的较窄的搜索词语。如果对查询的回答是否定的，则在步骤 890 中方法 800 结束。如果对查询的回答是肯定的，则方法 800 进行到步骤 820。

[0118] 在步骤 820 中，方法 800 选择搜索词语 k。在步骤 830 中，方法 800 又选择有关较窄搜索词语 k_1。

[0119] 在步骤 840 中，方法 800 查询是否存在与有关较窄搜索词语 k_1 相关联的销售和 / 或点击信息。例如，方法 800 可以确定是否存在任何与搜索词语“computer with SDRAM”相关联的销售信息。如果对查询的回答是否定的，则方法 800 返回步骤 830，并且另一个有关搜索词语 k_n 被选择。如果对查询的回答是肯定的，则方法 800 进行到步骤 850。

[0120] 在步骤 850 中，方法 800 查询有关搜索词语的销售信息是否大于某个阈值。换言之，方法 800 确定销售信息是否可以可靠地用于调整搜索词语 k 的热得分。在一个实施例中，在实际应用销售信息以影响更宽、更一般化的搜索词语之前，验证存在针对有关较窄搜索词语的大量销售可能是比较谨慎的。从而，如果对查询的回答是否定的，则方法 800 返回步骤 830，并且另一个有关搜索词语 k_n 被选择。如果对查询的回答是肯定的，则方法 800 进行到步骤 860。

[0121] 在步骤 860 中，方法 800 从得自搜索词语 k 的搜索结果集合中列出的产品中选择热得分。接下来，根据与搜索词语 k_i 相关联的销售和 / 或点击信息调整 $hotscore_{k,p}$。实际上，可以直接根据 $hotscore_{ki,p}$ 来调整 $hotscore_{k,p}$。

[0122] 在步骤 870 中，方法 800 查询是否来自从搜索词语 k 得出的搜索结果集合的产品的所有热得分都已经被调整。如果对查询的回答是否定的，则方法 800 返回步骤 860，并且另一个产品被选择。如果对查询的回答是肯定的，则方法 800 进行到步骤 880。

[0123] 在步骤 880 中，方法 800 查询是否所有有关较窄搜索词语都已经被处理。如果对查询的回答是否定的，则方法 800 返回步骤 830，并且另一个搜索词语被选择。如果对查询的回答是肯定的，则方法 800 进行到步骤 885。

[0124] 在步骤 885 中，方法 800 查询是否所有一般搜索词语都已经被处理。如果对查询的回答是否定的，则方法 800 返回步骤 820，并且另一个一般搜索词语被选择。如果对查询的回答是肯定的，则在步骤 890 中方法 800 结束。

[0125] 应当注意，上述公开内容在购物场境中描述了本发明。但是，本领域的技术人员将会意识到，本发明并不局限于此。即，在一个实施例中，本发明可以被实现为用于一般搜索，例如根据点击信息生成得分。

[0126] 虽然以上已经描述了各种实施例，但是应当理解，它们只是示例性而不是限制性的。从而，优选实施例的广度和范围不应当由任何上述示例性实施例所限，而应当仅根据以下权利要求及其等同物来限定。

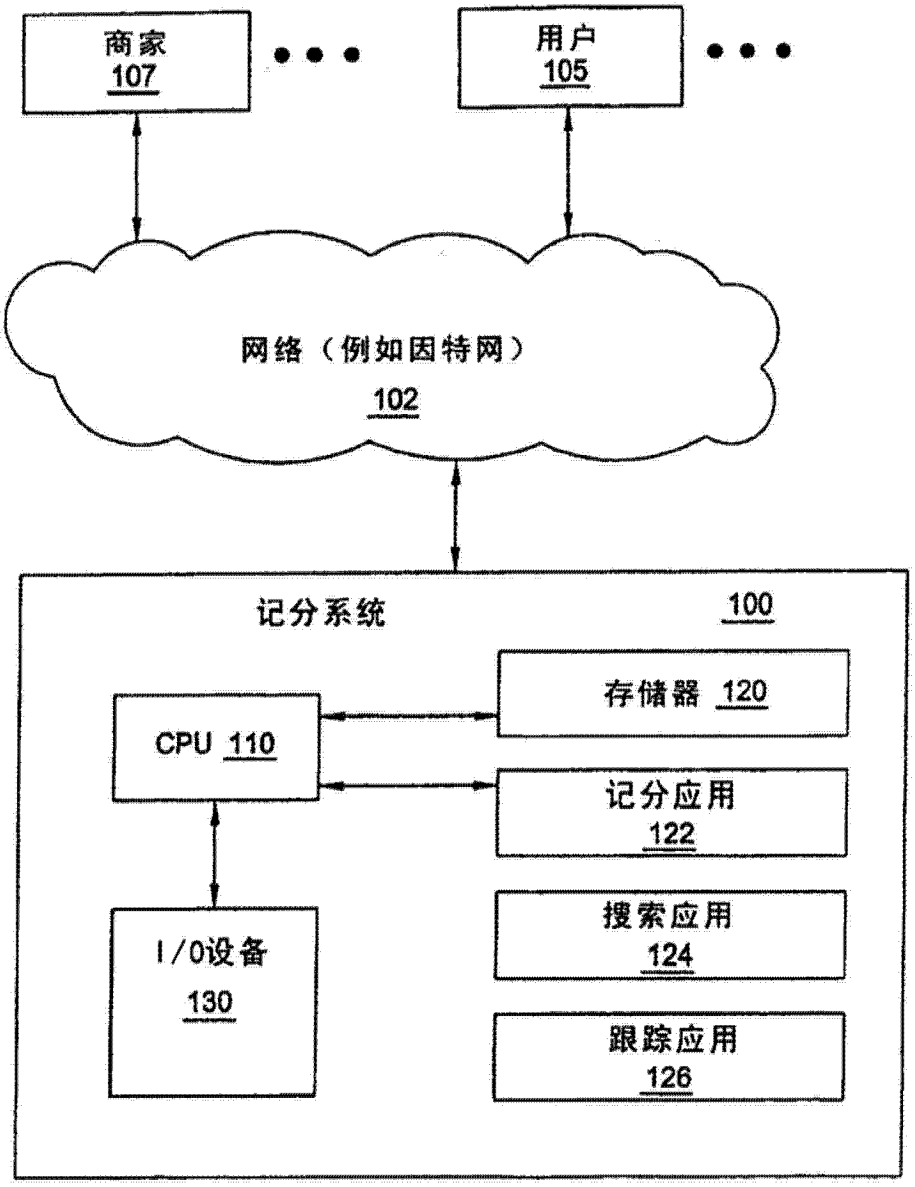

商家
107
用户
105
网络（例如因特网）
102
记分系统
100
存储器 120
CPU 110
记分应用
122
搜索应用
124
I/O设备
130
跟踪应用
126

图 1

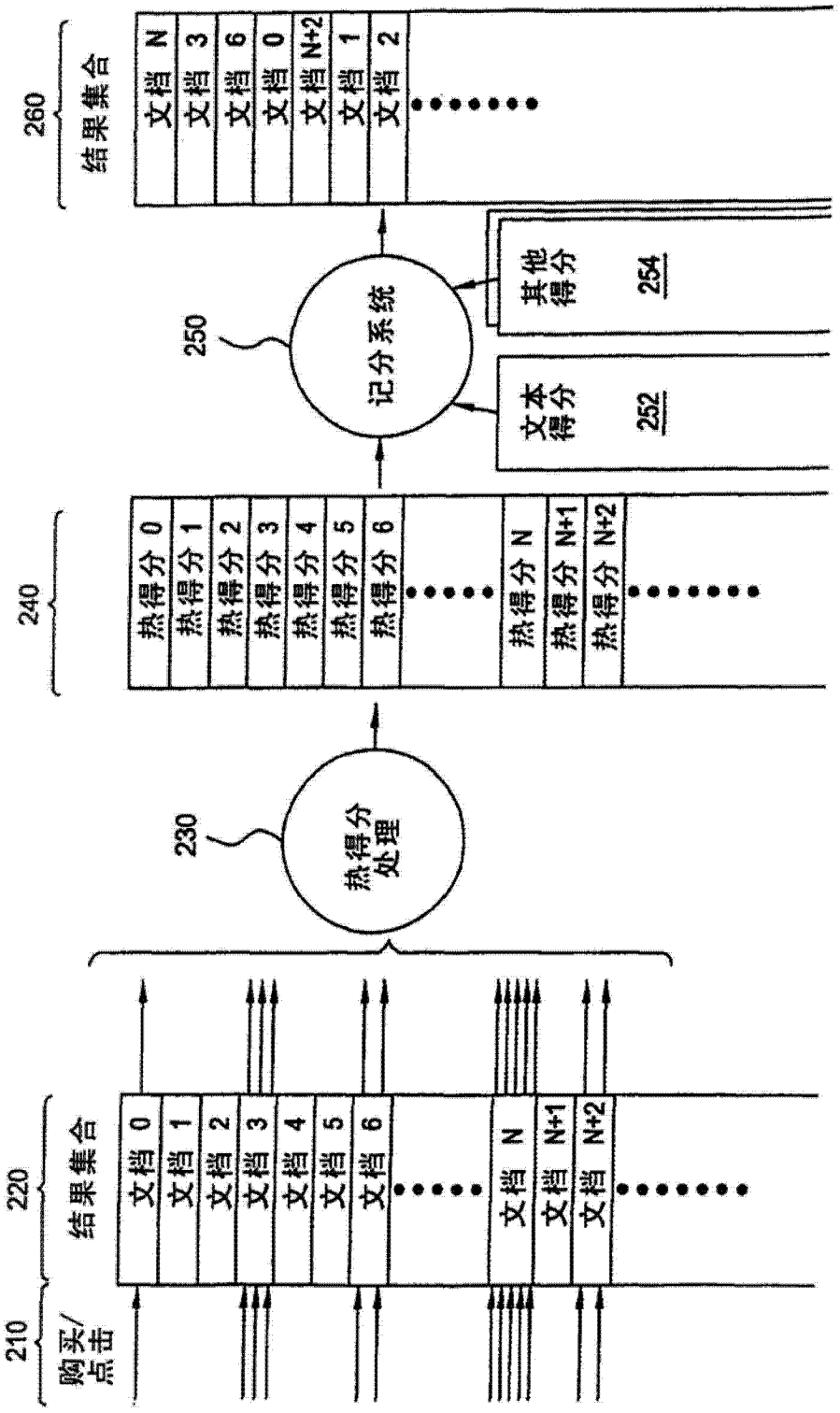

210
购买/点击
220
结果集合
文档 0
文档 1
文档 2
文档 3
文档 4
文档 5
文档 6
文档 N
文档 N+1
文档 N+2
230
热得分处理
240
热得分 0
热得分 1
热得分 2
热得分 3
热得分 4
热得分 5
热得分 6
热得分 N
热得分 N+1
热得分 N+2
250
记分系统
其他得分 254
文本得分 252
260
结果集合
文档 N
文档 3
文档 6
文档 0
文档 N+2
文档 1
文档 2

图 2

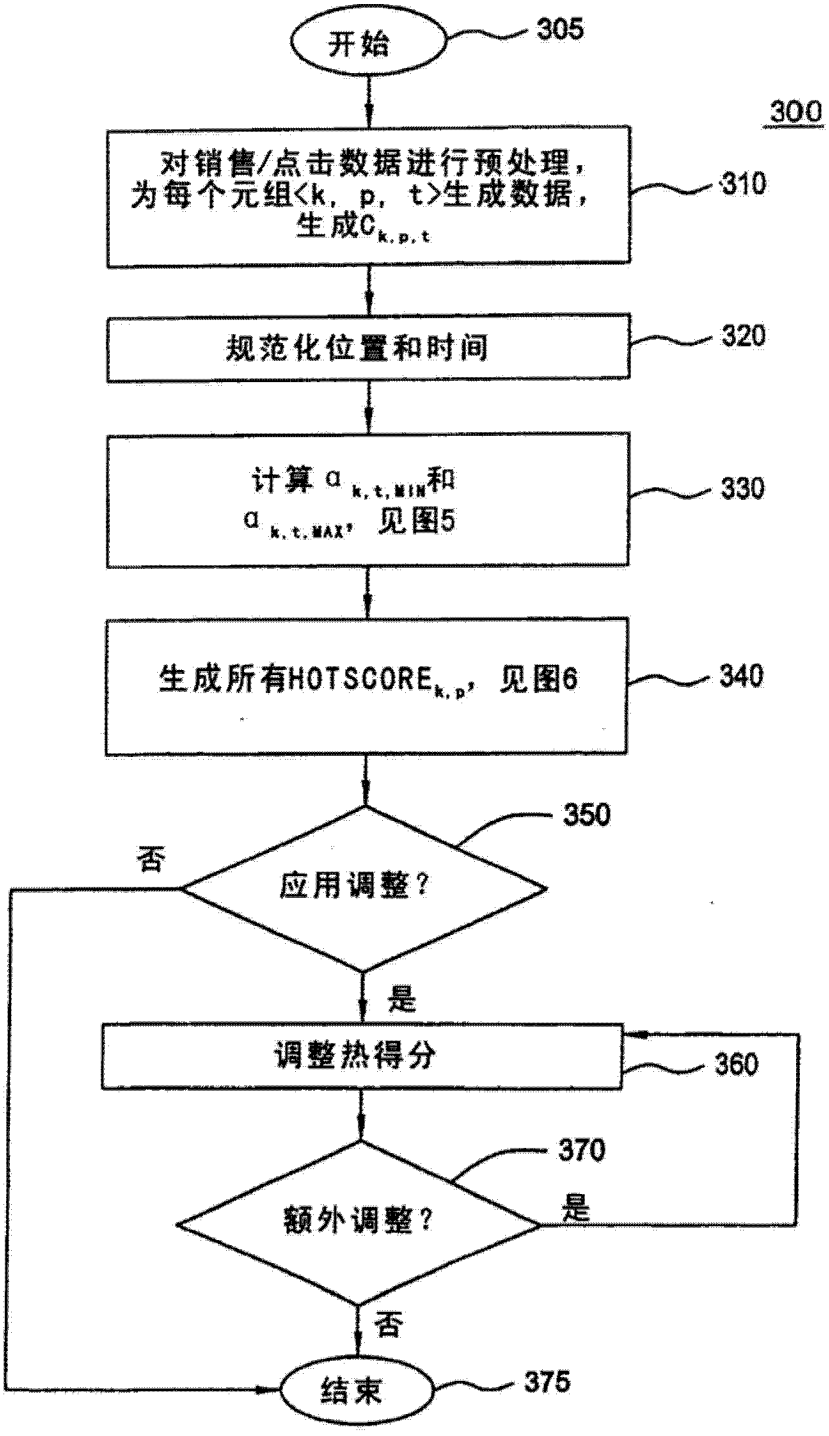

开始
305
300
对销售/点击数据进行预处理，为每个元组<k，p，t>生成数据，生成$C_{k,p,t}$
310
规范化位置和时间
320
计算$\alpha_{k,t,MIN}$和$\alpha_{k,t,MAX}$，见图5
330
生成所有$HOTSCORE_{k,p}$，见图6
340
应用调整？
350
否
是
调整热得分
360
额外调整？
370
是
否
结束
375

图 3

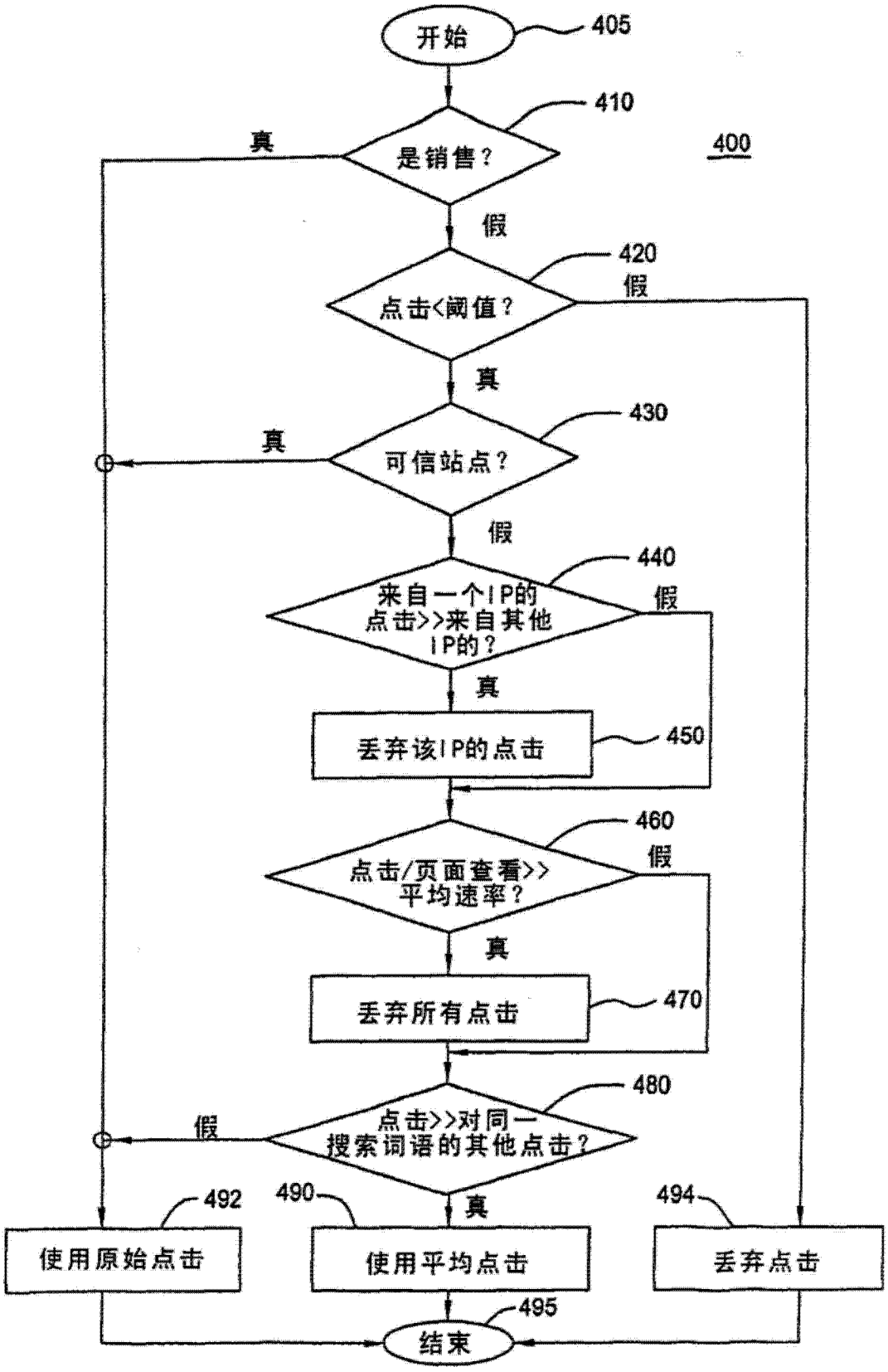

开始
405
410
是销售？
真
400
假
420
点击<阈值？
假
真
430
可信站点？
真
假
440
来自一个IP的点击>>来自其他IP的？
假
真
丢弃该IP的点击
450
460
点击/页面查看>>平均速率？
假
真
丢弃所有点击
470
480
点击>>对同一搜索词语的其他点击？
假
真
492
使用原始点击
490
使用平均点击
494
丢弃点击
495
结束

图 4

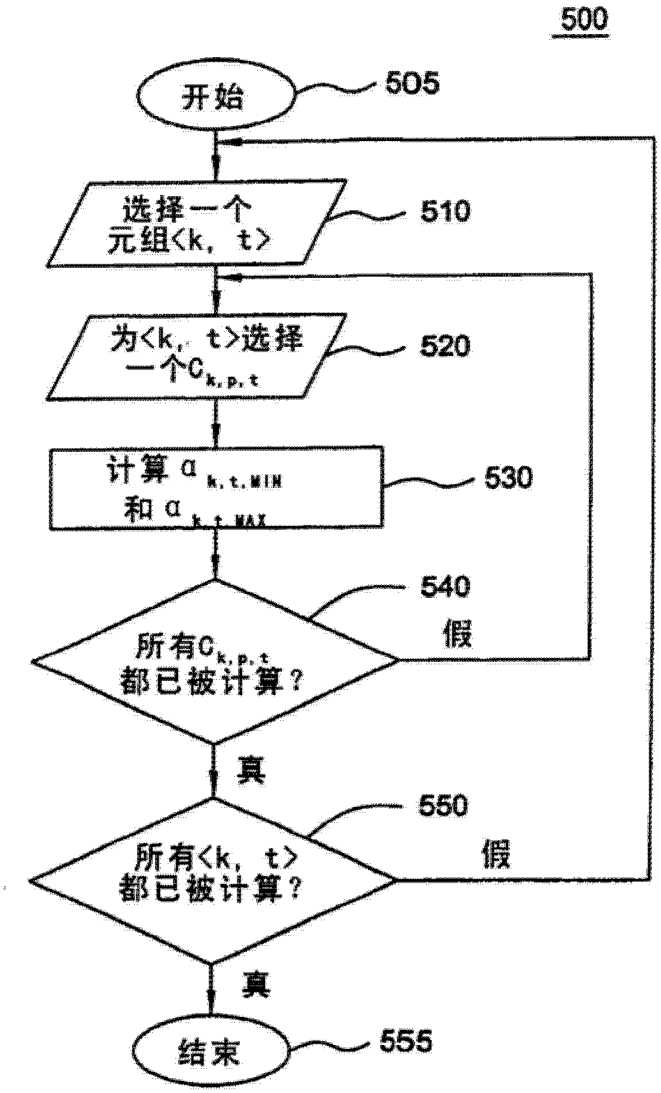

500
开始
505
选择一个元组<k, t>
510
为<k, t>选择一个$C_{k,p,t}$
520
计算$\alpha_{k,t,MIN}$和$\alpha_{k,t,MAX}$
530
所有$C_{k,p,t}$都已被计算?
540
假
真
所有<k, t>都已被计算?
550
假
真
结束
555

图 5

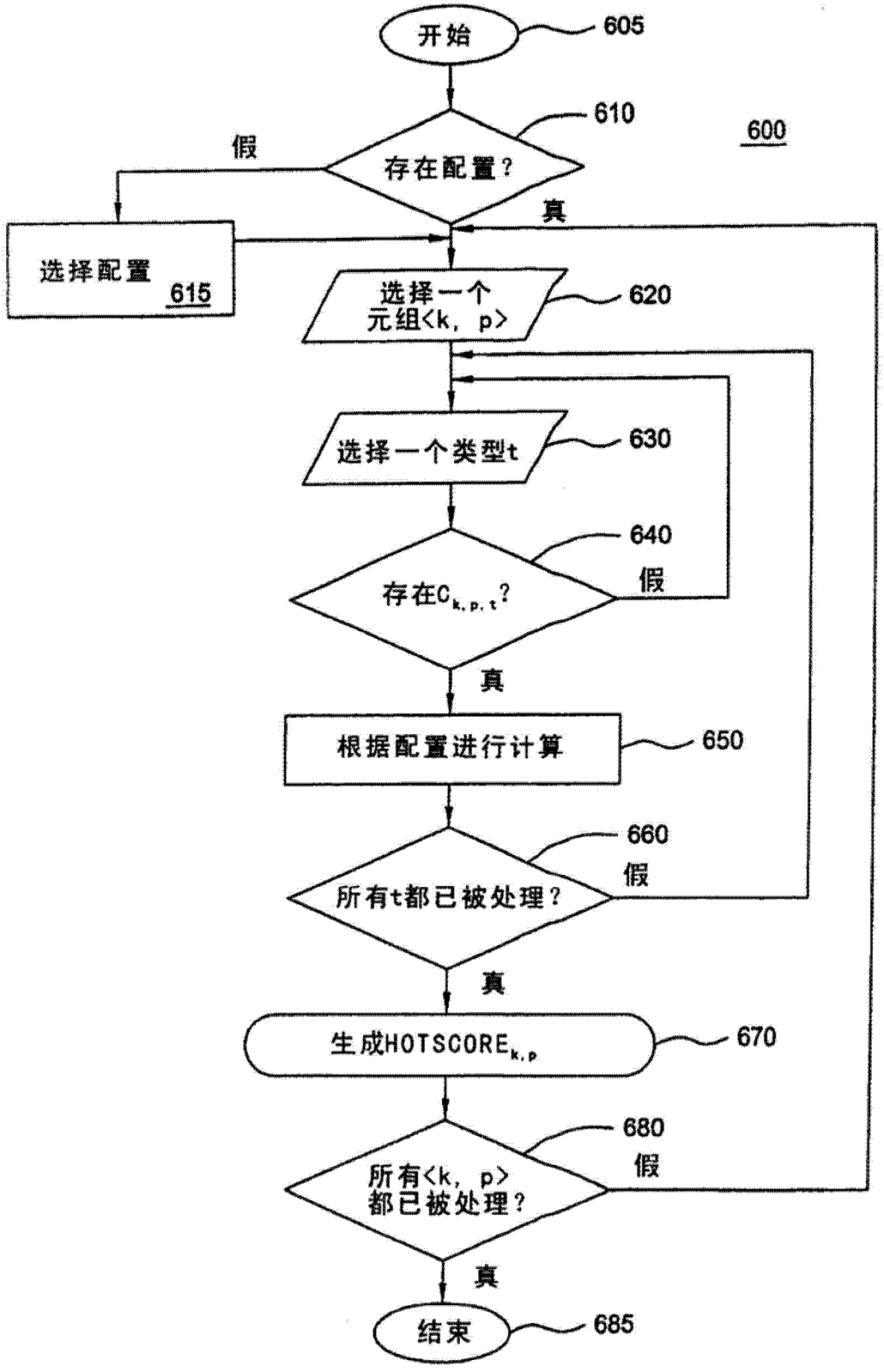

开始
605
610
600
假
存在配置？
真
选择配置
615
选择一个
元组<k，p>
620
选择一个类型t
630
640
假
存在$C_{k,p,t}$？
真
根据配置进行计算
650
660
假
所有t都已被处理？
真
生成$HOTSCORE_{k,p}$
670
680
假
所有<k，p>
都已被处理？
真
结束
685

图 6

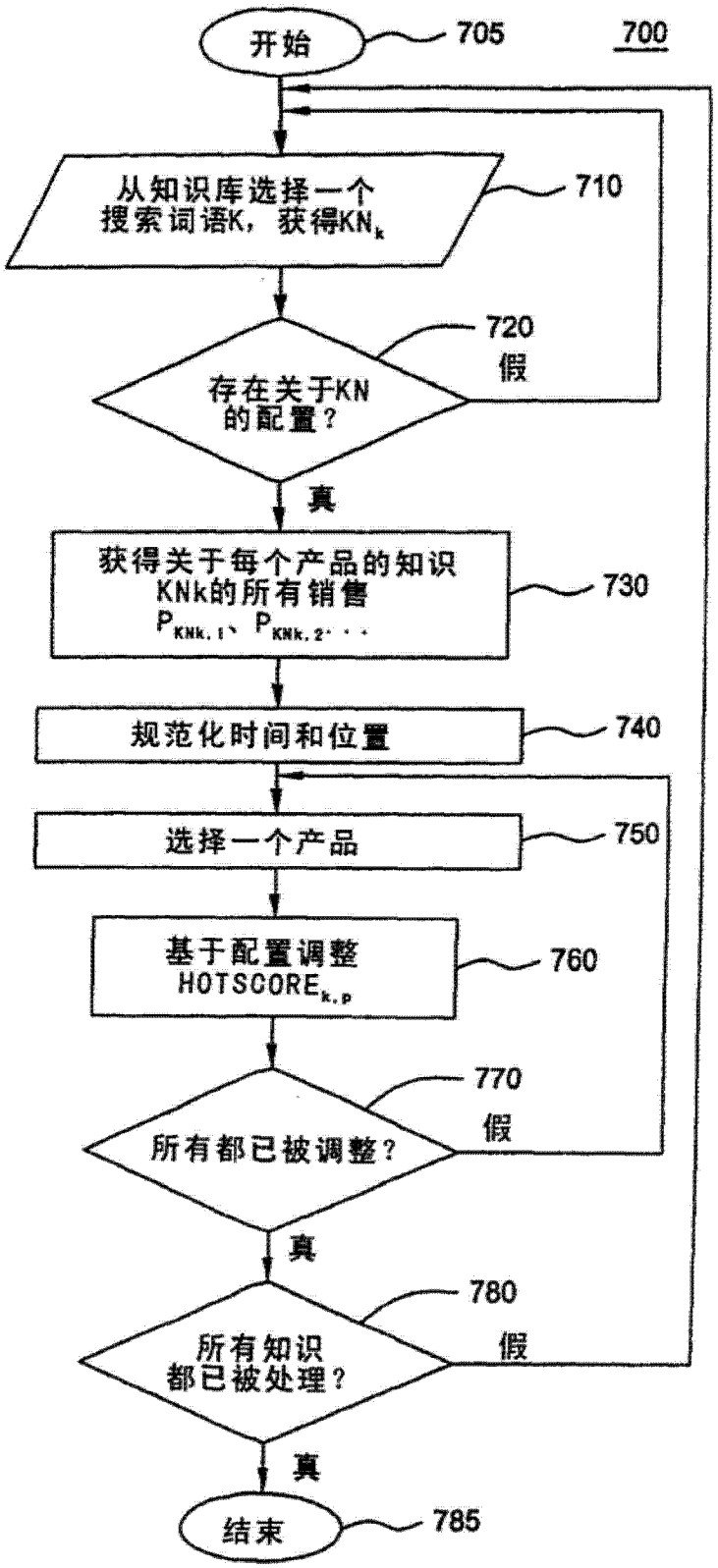

开始
705
700
从知识库选择一个搜索词语K，获得KN_k
710
存在关于KN的配置？
720
假
真
获得关于每个产品的知识KNk的所有销售 P_KNk.1、P_KNk.2···
730
规范化时间和位置
740
选择一个产品
750
基于配置调整 HOTSCORE_k,p
760
所有都已被调整？
770
假
真
所有知识都已被处理？
780
假
真
结束
785

图 7

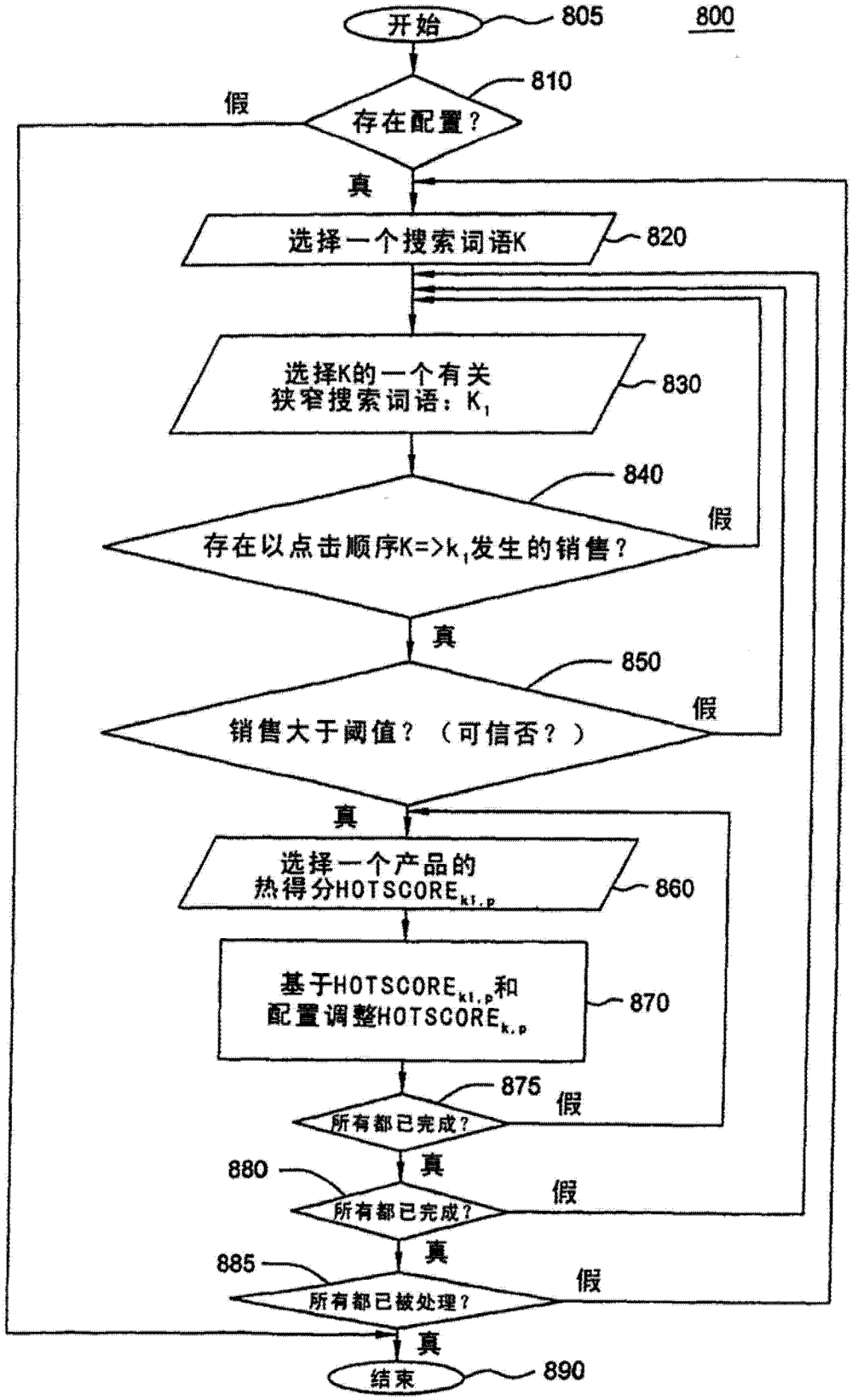

开始
805
800
810
假
存在配置？
真
选择一个搜索词语K
820
选择K的一个有关
狭窄搜索词语：K_1
830
840
假
存在以点击顺序K=>k_1发生的销售？
真
850
假
销售大于阈值？（可信否？）
真
选择一个产品的
热得分$HOTSCORE_{k1,p}$
860
基于$HOTSCORE_{k1,p}$和
配置调整$HOTSCORE_{k,p}$
870
875
假
所有都已完成？
真
880
所有都已完成？
假
真
885
所有都已被处理？
假
真
结束
890

图 8