[19] 中华人民共和国国家知识产权局

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

[12] 发明专利申请公布说明书

[21] 申请号 200910020951.4

[43] 公开日 2009 年 7 月 8 日

[11] 公开号 CN 101477554A

[22] 申请日 2009.1.16
[21] 申请号 200910020951.4
[71] 申请人 西安电子科技大学
地址 710071 陕西省西安市太白路 2 号
[72] 发明人 杜晨光 颜 涛 邓双成 李晓辉

[74] 专利代理机构 陕西电子工业专利中心
代理人 王品华 黎汉华

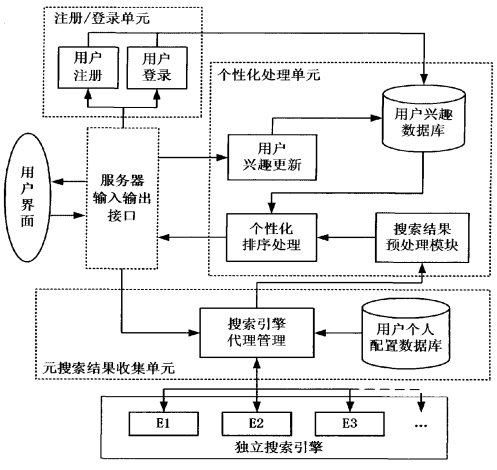
权利要求书 3 页 说明书 12 页 附图 5 页

[54] 发明名称

基于用户兴趣的个性化元搜索引擎及搜索结果处理方法

[57] 摘要

本发明公开了一种基于用户兴趣的个性化元搜索引擎及搜索结果处理方法。 其搜索引擎包括用户注册/登录单元、元搜索结果收集单元和个性化处理单元，该个性化处理单元通过搜索结果预处理模块、个性化排序处理模块和用户兴趣更新模块，完成对搜索结果的筛选、个性化排序及对用户兴趣模型的建立和更新。 其对搜索结果的处理为：建立并初始化用户兴趣模型，存入服务器；将用户输入的搜索词语，按各搜索引擎要求的格式分发；将收集的返回结果转换成统一的格式，依次进行去除重复处理和个性化排序，并提交给用户；捕获用户点击的搜索结果，改变兴趣种类和权值大小，更新用户兴趣模型数据。 本发明具有查询覆盖度和准确度高的优点，可用于互联网的搜索引擎。



1. 一种基于用户兴趣的个性化元搜索引擎，主要由用户注册/登录单元、元搜索结果收集单元和个性化处理单元组成，其特征在于个性化处理单元，包括：

搜索结果预处理模块，用于接收元搜索结果收集单元送入的搜索结果原始数据，进行基于网址和基于摘要内容的去除重复处理，并将处理后的搜索结果送入个性化排序处理模块；

个性化排序处理模块，用于接收搜索结果预处理模块送入的搜索结果数据，并进行基于用户兴趣和搜索意图的个性化排序处理，将排序后的搜索结果提交给用户；

用户兴趣更新模块，用于捕获用户对搜索结果的点击行为，对用户点击过的搜索结果进行兴趣分析，并对用户兴趣数据库中存储的用户兴趣模型数据进行更新；

用户兴趣数据库，用于存储网站注册用户的兴趣模型数据，并将这些数据提供给个性化排序处理模块进行个性化排序。

2. 一种基于用户兴趣的个性化元搜索引擎结果处理方法，包括：

步骤 1. 建立并初始化用户兴趣模型数据，保存在服务器的用户兴趣数据库中；

步骤 2. 将用户输入的搜索词语，按各独立搜索引擎要求的格式分发出去，并收集它们返回的结果，将收集的返回结果转换成统一的格式；

步骤 3. 将转换成统一格式的搜索结果进行去除重复处理；

步骤 4. 根据用户兴趣模型和本次搜索词语对去除重复处理后的搜索结果进行个性化排序，并将该排序结果提交给用户；

步骤 5. 捕获用户点击的搜索结果，对其进行兴趣分析，改变兴趣种类和权值大小，并根据改变的结果更新用户兴趣数据库中用户的兴趣模型数据。

3. 根据权利要求 2 所述的基于用户兴趣的个性化元搜索引擎结果处理方法，其中步骤 1 所述的建立用户兴趣模型数据，包括两种方式：一是用户直接通过主动选择网站提供的兴趣类别并设置权值大小，建立该用户的用户兴趣模型数据；二是搜索引擎系统通过兴趣更新模块对用户搜索行为的兴趣分析，自动建立用户兴趣模型数据。

4. 根据权利要求 2 或 3 所述的基于用户兴趣的个性化元搜索引擎结果处理方法，其中所述的用户兴趣模型数据，用 $I(C)=\{(c_1,w_1),(c_2,w_2),\ldots,(c_m,w_m)\}$ 表示，其中 (c_i,w_i) 是用户的一个兴趣分量，c_i 为兴趣类别，w_i 为 c_i 对应的归一化权值，该用户兴趣模型数据是用于定量描述用户兴趣喜好及其喜好程度的数据集。

5. 根据权利要求 2 所述的基于用户兴趣的个性化元搜索引擎结果处理方法，其中步骤 3 所述的将转换成统一格式的搜索结果进行去除重复处理，是先将统一格式的搜

索结果按照网址是否相同进行过滤，只保留网址不同的搜索结果；再将各条搜索结果的摘要内容与其它搜索结果的摘要内容进行文本比较，滤除摘要内容相似的搜索结果。

6. 根据权利要求 2 所述的基于用户兴趣的个性化元搜索引擎结果处理方法，其中步骤4所述的根据用户兴趣模型和本次搜索词语对去除重复处理后的搜索结果的个性化排序方法，按如下步骤进行：

（6a） 分别计算搜索词语与用户兴趣的相关度向量和搜索结果与用户兴趣的相关度向量；

（6b） 根据步骤（6a）得到的相关度向量，计算搜索结果与用户搜索词语的相关度向量；

（6c） 计算搜索结果在各独立搜索引擎上的排名得分向量；

（6d） 将步骤（6b）和（6c）得到的数值进行加权综合，得到每条搜索结果基于用户兴趣的权值，并按权值大小排序，得到个性化排序结果。

7. 根据权利要求 6 所述的基于用户兴趣的个性化元搜索引擎结果处理方法，其中步骤（6a）所述的分别计算搜索词语与用户兴趣的相关度向量和搜索结果与用户兴趣的相关度向量，按照如下步骤进行：

（7a） 按照公式 $sim(Q,c_i)=\sum_{j=1}^{h}(w_j*x_j)\Big/\sqrt{\sum_{j=1}^{h}w_j^2*\sum_{j=1}^{h}x_j^2}$，计算搜索词语 Q 与一个兴趣类别 c_i 的相关度，

式中，w_j 是 Q 经分词处理后的一个关键词对应在用户模型中的兴趣类别 c_i 上的归一化权值，x_j 是该关键词在 Q 中的归一化重要度，当所有 w_j 都为零，$sim(Q,c_i)=0$；

（7b） 对用户兴趣模型中的所有兴趣类别进行相关度计算，得到搜索词语 Q 与用户兴趣的相关度向量 $Sim(Q,C)=(sim(Q,c_1),\ldots,sim(Q,c_m))$；

（7c） 按照公式 $sim(r_i,c_j)=\sum_{i=1}^{n}(w_i*x_i)\Big/\sqrt{\sum_{i=1}^{n}w_i^2*\sum_{i=1}^{n}x_i^2}$，计算一条搜索结果 r_i 与一个兴趣类别 c_j 的相关度，

式中，w_i 是 r_i 经分词处理后的一个关键词对应在用户模型中的兴趣类别 c_j 上的归一化权值，x_i 是该关键词在 r_i 中的归一化重要度，当所有 w_i 都为零，$sim(r_i,c_j)=0$；

（7d） 对用户兴趣模型中的所有兴趣类别进行相关度计算，得到一条搜索结果 r_i 与用户兴趣的相关度向量 $Sim(r_i,C)=(sim(r_i,c_1),\ldots,sim(r_i,c_n))$；

（7e） 对所有搜索结果与所有用户兴趣类别的相关度进行计算，得到搜索结果

集 R 与用户兴趣的相关度向量 $Sim(R, C) = (Sim(r_i, C),..., Sim(r_i, C))$。

8. 根据权利要求 6 所述的基于用户兴趣的个性化元搜索引擎结果处理方法，其中步骤（6b）所述的计算搜索结果与用户搜索词语的相关度向量，按照如下步骤进行：

（8a） 计算 r_i 与 Q 在兴趣类别 c_j 上的相关度 $sim(r_i, Q, c_j)$：当 $Sim(Q,C)$ 中所有分量全为 0 时，$sim(r_i, Q, c_j) = sim(r_i, c_j)$，否则 $sim(r_i, Q, c_j) = sim(r_i, c_j) \times sim(Q, c_j)$；

（8b） 计算所有用户兴趣类别与搜索结果 r_i 的相关度向量，得到：

$Sim(r_i, Q, C) = (sim(r_i, Q, c_1) ,..., sim(r_i, Q, c_n))$，并计算 Q 与 r_i 的相关度：

$$sim(r_i,Q) = \frac{1}{n}\sum_{c_j=1}^{n} sim(r_i,Q,c_j)$$；

（8c） 计算所有搜索结果与搜索词语的相关度，得到搜索结果集 R 与搜索词语的相关度向量 $Sim(R,Q) = (sim(r_1,Q),..., sim(r_n,Q))$。

9. 根据权利要求 6 所述的基于用户兴趣的个性化元搜索引擎结果处理方法，其中步骤（6d）所述的计算每条搜索结果在各独立搜索引擎上的排名得分，按照如下步骤进行：

（9a） 按照公式 $weight_{SE}(r_i) = 1 - \prod_{i=1}^{k}(1 - \frac{1}{k \cdot n_i})$，计算搜索结果 r_i 在各独立搜索引擎上的排名得分，

式中，k 是包含 r_i 的独立搜索引擎的个数，n_i 是在相应搜索引擎上的排名名次；

（9b） 计算所有搜索结果在各独立搜索引擎上的排名得分，得到搜索结果集 R 在各独立搜索引擎上的排名得分向量 $Weight_{SE}(R)=(weight_{SE}(r_1) ,..., weight_{SE}(r_n))$。

10. 根据权利要求 2 所述的基于用户兴趣的个性化元搜索引擎结果处理方法，其中步骤 5 所述的更新用户兴趣数据库中用户的兴趣模型数据，按如下步骤进行：

（10a） 捕获用户在客户端点击的搜索结果，并传回服务器端；

（10b）对该搜索结果的标题和摘要进行分词处理，得到该搜索结果的关键词集；

（10c） 依据关键词集进行兴趣分析，得到该用户最新的兴趣类别及相应的权值大小；

（10d） 根据最新的兴趣类别及相应的权值大小对用户兴趣数据库中的用户兴趣模型数据进行更新。

基于用户兴趣的个性化元搜索引擎及搜索结果处理方法

技术领域

本发明属于互联网信息处理技术领域，涉及搜索引擎、Web 数据挖掘和知识发现技术，特别是涉及基于用户兴趣的个性化元搜索系统及方法，用于互联网的搜索引擎。

背景技术

搜索引擎的出现，大大提高了人们对互联网信息检索的能力和效率，已经成为互联网的基础应用之一。据中国互联网络信息中心在 2008 年中期的统计，中国网民搜索引擎的使用率为 69.2%，并处在高速增长之中，而在互联网高度普及的美国，网民对搜索引擎的使用率已达 91%。可见，上网用户对搜索引擎已经产生了强烈的依赖。

目前，搜索引擎领域主要有以下几种技术：

（1）传统搜索引擎：这种搜索引擎目前应用最广泛且用户数量最多，主要代表有谷歌（www.google.com）、百度（www.baidu.com）、雅虎（cn.yahoo.com）等。

这种搜索引擎虽然给人们带来了便利，但是它们却存在着本身无法克服的缺陷。根据专业评测，目前主流搜索引擎的网络资源覆盖面加在一起只占整个网络的约 42%，返回的结果相关度不足 45%，而且由于对网页的索引和排序机制互不相同，导致同样一个搜索请求在不同搜索引擎中的查询结果的重复率不足 34%。因此，单个这样的搜索引擎是无法满足用户搜索需求的，要想获得一个比较全面、准确的搜索结果，用户就必须反复调用多个搜索引擎，这大大降低了用户的检索效率，提高了信息检索的难度。

（2）元搜索引擎（Meta-Search Engine）：元搜索引擎的出现，在一定程度上弥补了传统搜索引擎的不足，其主要代表有国外的 MetaCrawler（www.metacrawler.com）、Dogpile（www.dogpile.com）和国内的比比猫（www.bbmao.com）等。元搜索是一种将用户检索请求同时发送给多个独立搜索引擎，并将它们的搜索结果汇集在一起返回给用户的搜索技术。它的优点是综合了多个独立搜索引擎的搜索结果，从而提高了搜索结果在整个网络资源上的覆盖率，省去了用户自己逐个调用不同搜索引擎进行查询的麻烦。

但是，目前已投入实用的元搜索引擎的搜索结果排序方式仅仅是以各独立搜索引擎返回结果的排序或某种统一的排序原则为依据的，所以对与不同用户的搜索请求不能做到根据用户的兴趣喜好和搜索意图返回与之相适应的排序结果，即搜索的准确度并未得到有效提高。因此，在信息量巨大的互联网世界里用户想要找到自己需要的信息的难易程度并未得到有效改善。

（3）个性化搜索引擎（Personalized Search Engine）：为了满足用户的个性化搜索需求，弥补传统搜索引擎和元搜索引擎的不足，给用户提供更为精准的搜索服务，人们提出了个性化搜索引擎的思想，这种搜索引擎目前还处于技术研究和初步应用阶

段。在这方面的研究中，具有代表性的方法一个是通过用户对搜索结果进行打分来调节搜索结果的排列次序，一个是将用户的搜索历史存放在用户计算机的 cookie 文件中，作为以后用户进行搜索的参考来影响搜索结果的次序。

但是这些方法仍存在缺陷。对于依靠用户打分来说，大量用户对搜索结果的评价并不能准确刻划某个特定用户的兴趣喜好，无法实现针对每个用户的个性化服务；对于在用户计算机上记录用户搜索历史来说，这种方法实际上只是记录了这台计算机上进行过的搜索历史，如果使用该计算机的用户更换或者用户在别的计算机上进行搜索，则这种个性化搜索的作用就失效了。

从上面介绍的目前存在的三种搜索引擎技术来看，个性化搜索技术无疑是搜索引擎进一步发展的方向，但这个领域的技术研究还远未达到成熟阶段，需要有更加有效和实用的个性化搜索技术来改善用户的搜索体验。

发明内容

本发明的目的在于避免上述已有搜索引擎的缺陷，提供一种基于用户兴趣的个性化元搜索引擎及其搜索结果处理方法，以准确确定用户兴趣和搜索意图，在服务器上长期保存和及时更新用户兴趣，并利用用户兴趣和搜索意图对元搜索的搜索结果进行个性化排序，提高搜索结果的覆盖度和搜索的准确度。

本发明的目的是这样实现的：

本发明的搜索系统主要由用户注册/登录单元、元搜索结果收集单元和个性化处理单元组成，其中个性化处理单元，包括：

搜索结果预处理模块，用于接收元搜索结果收集单元送入的搜索结果原始数据，进行基于网址和基于摘要内容的去除重复处理，并将处理后的搜索结果送入个性化排序处理模块；

个性化排序处理模块，用于接收搜索结果预处理模块送入的搜索结果数据，并进行基于用户兴趣和搜索意图的个性化排序处理，将排序后的搜索结果提交给用户；

用户兴趣更新模块，用于捕获用户对搜索结果的点击行为，对用户点击过的搜索结果进行兴趣分析，并对用户兴趣数据库中存储的用户兴趣模型数据进行更新；

用户兴趣数据库，用于存储网站注册用户的兴趣模型数据，并将这些数据提供给个性化排序处理模块进行个性化排序。

所述的用户兴趣模型数据用 $I(C)=\{(c_1,w_1),(c_2,w_2),\ldots,(c_m,w_m)\}$ 表示，其中 (c_i,w_i) 是用户的一个兴趣分量，c_i 为兴趣类别，w_i 为 c_i 对应的归一化权值，该用户兴趣模型数据是用于定量描述用户兴趣喜好及其喜好程度的数据集。

本发明的引擎搜索结果处理方法，包括：

步骤 1. 建立并初始化用户兴趣模型，保存在服务器的用户兴趣数据库中；

步骤 2. 将用户输入的搜索词语，按各独立搜索引擎要求的格式分发出去，并收集它们返回的结果，将收集的返回结果转换成统一的格式。

步骤 3. 将转换成统一格式的搜索结果进行去除重复处理；

步骤 4. 根据用户兴趣模型和本次搜索词语对去除重复处理后的搜索结果进行个性化排序，并将该排序结果提交给用户；

步骤 5. 捕获用户点击的搜索结果，对其进行兴趣分析，改变兴趣种类和权值大小，并根据改变的结果更新用户兴趣数据库中用户的兴趣模型数据。

上述引擎结果处理方法，其中步骤 1 所述的建立用户兴趣模型，包括两种方式：一是用户直接通过主动选择网站提供的兴趣类别并设置权值大小，建立该用户的初始兴趣模型；二是搜索引擎系统通过兴趣更新模块对用户搜索行为的兴趣分析，自动建立用户兴趣模型。

上述引擎结果处理方法，其中步骤 3 所述的将转换成统一格式的搜索结果进行去除重复处理，是先将统一格式的搜索结果按照网址是否相同进行过滤，只保留网址不同的搜索结果；再将各条搜索结果的摘要内容与其它搜索结果的摘要内容进行文本比较，滤除摘要内容相似的搜索结果。

上述引擎结果处理方法，其中步骤 4 所述的根据用户兴趣模型和本次搜索词语对去除重复处理后的搜索结果的个性化排序方法，按如下步骤进行：

1) 分别计算搜索词语与用户兴趣的相关度向量和搜索结果与用户兴趣的相关度向量；

2) 根据步骤 1) 得到的相关度向量，计算搜索结果与用户搜索词语的相关度向量；

3) 计算搜索结果在各独立搜索引擎上的排名得分向量；

4) 将步骤 2) 和 3) 得到的数值进行加权综合，得到每条搜索结果基于用户兴趣的权值，并按权值大小排序，得到个性化排序结果。

上述引擎结果处理方法，其中步骤 5 所述的更新用户兴趣数据库中用户的兴趣模型数据，按如下步骤进行：

a) 捕获用户在客户端点击的搜索结果，并传回服务器端；

b) 对该搜索结果的标题和摘要进行分词处理，得到该搜索结果的关键词集；

c) 依据关键词集进行兴趣分析，得到该用户最新的兴趣类别及相应的权值大小；

d) 根据最新的兴趣类别及相应的权值大小对用户兴趣数据库中的用户兴趣模型数据进行更新。

本发明与背景技术相比具有的优势在于：

本发明是一种个性化元搜索引擎技术，适用于建立互联网上的个性化元搜索引擎；

本发明通过元搜索技术同时抓取多个独立搜索引擎的搜索结果，提高了搜索结果的覆盖度，克服了单个独立搜索引擎搜索结果覆盖度低的问题；

本发明通过为每个用户建立各自的用户兴趣模型，并将其长期保存在服务器数据库中，而且随着用户的搜索过程对用户兴趣数据不断更新，使得用户不论身处何时何地，本发明的搜索系统均能准确定位用户兴趣，为其提供个性化搜索服务，不仅克服了一般元搜索引擎不能提供个性化服务的缺点，而且克服了现有个性化搜索技术不能长期保存用户兴趣和不能精准定位个人兴趣的缺点；

本发明通过独创的引擎搜索结果处理机制将多个独立搜索引擎的搜索结果进行去除重复处理，并计算每条搜索结果的个性化权值 PersonalRank，为用户提供最适合其搜索意图和兴趣喜好的搜索结果排列方式，使得搜索结果的准确度得到显著提高，用户的搜索需求得到最大程度的满足，用户通过本发明的搜索系统找到自己需要的搜索结果的难度大大降低。

附图说明

图 1 是本发明搜索引擎系统结构框图；

图 2 是本发明搜索结果处理流程图；

图 3 是本发明用户兴趣模型示例图；

图 4 是本发明去除重复搜索结果流程图；

图 5 是本发明基于用户兴趣的个性化排序流程图；

图 6 是本发明用户兴趣更新流程图。

具体实施方式

参照图 1，本发明的搜索引擎系统主要由用户注册/登录单元，元搜索结果收集单元，个性化处理单元，服务器输入输出接口和外部独立搜索引擎资源组成，其中：

所述的用户注册/登录单元，由注册模块和登录模块组成。注册模块负责接收新用户通过服务器输入输出接口发来的注册请求，通过收集和向数据库中保存必要的用户信息，使其成为网站注册用户；登录模块负责利用存储的用户信息验证请求登录的用户的合法性，使合法用户登录进网站中进行搜索活动。

所述的元搜索结果收集单元，由搜索引擎代理管理模块和用户个人配置数据库组成。用户个人配置数据库负责存储用户的搜索配置数据，如选择的独立搜索引擎种类、每个独立搜索引擎抓取的搜索结果数量和搜索结果的显示效果；搜索引擎代理管理模块负责在用户通过服务器输入输出接口向网站发出搜索请求时，根据用户个人配置数

据库中存储的用户配置信息，为用户选择相应的独立搜索引擎，按照各个独立搜索引擎的搜索格式向外部独立搜索引擎资源发出搜索请求，并收集它们返回的搜索结果，把它们转换成统一的格式。

所述的个性化处理单元，由搜索结果预处理模块、个性化排序处理模块、用户兴趣更新模块和用户兴趣数据库组成。该搜索结果预处理模块，用于接收元搜索结果收集单元送入的搜索结果原始数据，进行基于网址和基于摘要内容的去除重复处理，其中网址去重和摘要去重依次进行：首先将统一格式的搜索结果按照网址是否相同进行过滤，只保留网址不同的搜索结果，再将网址去重后的各条搜索结果的摘要内容与其它搜索结果的摘要内容进行文本比较，滤除摘要内容相似的搜索结果，最后将处理后的搜索结果送入个性化排序处理模块；该个性化排序处理模块，用于接收搜索结果预处理模块送入的搜索结果数据，并进行基于用户兴趣和搜索意图的个性化排序处理，在处理过程中，综合考虑搜索词语与搜索结果基于用户兴趣模型的相关度以及搜索结果在独立搜索引擎的排名得分，计算出个性化权值 PersonalRank，并以此为依据进行排序，将排序后的搜索结果通过服务器输入输出接口提交给用户界面；该用户兴趣更新模块，用于捕获用户对搜索结果的点击行为，对用户点击过的搜索结果进行兴趣分析，并对用户兴趣数据库中存储的用户兴趣模型数据进行更新，其中捕获用户对搜索结果的点击行为是通过在搜索结果显示页面上设置特定代码实现，并由服务器输入输出接口传回服务器进行兴趣分析，从而更新用户兴趣数据库中的用户兴趣模型数据；该用户兴趣数据库，用于存储网站注册用户的兴趣模型数据，这些数据是个性化排序处理模块进行个性化排序的依据，并由兴趣更新模块进行更新。

所述的服务器输入输出接口，是网站服务器用于服务器端与用户端进行数据交互的接口，将需要经过接口交互的数据送到相应的模块中。

所述的外部独立搜索引擎资源是互联网中各种提供搜索服务的独立搜索引擎，是本发明的搜索系统获取搜索结果数据的来源，由搜索结果收集单元通过发出搜索命令与外部独立搜索引擎资源 E_i 进行连接。

参照图 2，本发明的对搜索引擎结果的处理步骤如下：

步骤一，建立并初始化用户兴趣模型数据，保存在服务器的用户兴趣数据库中。

参照图 3，本发明中的用户兴趣模型是用户兴趣类别及其权值的数据记录集，其中包含若干个兴趣类别分量，用 $I(C)=\{(c_1,w_1),(c_2,w_2),\ldots,(c_m,w_m)\}$ 表示。其中 (c_i,w_i) 是用户的一个兴趣分量，c_i 为一个兴趣类别，w_i 为对应的归一化权值，即所有 w_i 之和为 1，w_i 越大说明兴趣类别 c_i 在该用户兴趣中的比重越大，也就是该用户在兴趣类别 c_i 方面

的喜好程度越大。对于用户兴趣模型的建立，包括两种方式：一是用户直接通过主动选择网站提供的兴趣类别并设置权值大小，建立该用户的用户兴趣模型数据；二是搜索引擎系统通过兴趣更新模块对用户搜索行为的兴趣分析，为用户自动建立用户兴趣模型数据。将按照以上方式建立的用户兴趣模型数据保存入服务器的用户兴趣数据库中，作为后续对搜索结果进行个性化排序的依据。

步骤二，将用户输入的搜索词语，按各独立搜索引擎要求的格式分发出去。

对于用户输入的搜索词语，首先由搜索引擎代理管理模块从用户个人配置数据库中取出该用户选定的独立搜索引擎种类，以及需要抓取的搜索结果数目这些必要的配置数据；然后按照各个独立搜索引擎的链接格式，将用户配置数据组合成相应的搜索链接；最后将这些组合好的搜索链接通过网络命令向独立搜索引擎资源分发出去。

步骤三，收集各独立搜索引擎返回的结果，将收集的返回结果转换成统一的格式。

搜索引擎代理管理模块接收到相应独立搜索引擎返回的搜索结果数据流，对这些数据流进行格式分析，分割出这些数据流中搜索结果的网址、标题、内容摘要以及在相应搜索结果中的原始排名名次信息，并将每组这样的信息作为本搜索引擎系统的一条统一格式的搜索结果数据。

步骤四，在搜索结果预处理模块中将转换成统一格式的搜索结果进行去除重复处理。

首先，进行基于网址的搜索结果去除重复处理。将统一格式的搜索结果按照网址是否相同进行过滤，只保留网址不同的搜索结果。在处理过程中，对于网址相同的搜索结果，优先保留在独立搜索引擎原始排名中名次靠前的那条搜索结果，将相对靠后的其它重复搜索结果删除。

然后，将各条搜索结果的摘要内容与其它搜索结果的摘要内容进行文本比较，滤除摘要内容相似的搜索结果，具体步骤如图 4 所示：

（4.1）设置有关参数

将用户搜索词语 S 由元搜索结果收集单元得到的独立搜索引擎返回的搜索结果集设为：$R_0(s) = \{r_{1,1}(1), r_{1,2}(2), ..., r_{i,j}(n), ...\}$，其中 $R_0(s).sum$ 表示查询结果总数，$r_{i,j}(n)$ 表示第 i 个独立搜索引擎的第 n 条搜索结果且在整个集合中排在第 j 位，$r_{i,j}(n).summary$ 表示该条的摘要，$r_{i,j}(n).length$ 为摘要的长度，$r_{i,j}(n).flag$ 为去重标志位；

将经过去除重复处理后的搜索结果集设为：$R(s) = \{r_1(x_1,y_1,...), r_2(x_2,y_2,...), ..., r_n(x_n,y_n,...)\}$，其中 $r_i(x_i,y_i,...)$表示 $R(s)$中的第 i 条搜索结果，且在包含该结果的独立搜索引擎上的排名分别为 $x_i,y_i,...$；

（4.2）将 $R_0(s)$中所有 $r_{i,j}(n).flag$ 置为 0，表示相应的 $r_{i,j}(n)$未进行过去除重复处理；

（4.3）从第一条搜索结果 $r_{1,1}(1)$开始，对于 $r_{i,k}(n)$和 $r_{j,t}(m)$，其中 $k<t$，若 $r_{j,t}(m).flag = 1$，表示已进行过去除重复处理，或 $r_{i,k}(n).length$ 和 $r_{j,t}(m).length$ 相差大于 50%，表

示两者摘要长度相差太大，不做处理，否则，从 $r_{i,k}$(n).*summary* 的前中后部分别截取长为 $0.6\times r_{i,k}$(n).*length* 的三个子串与 $r_{j,t}$(m).*summary* 进行比较，若 $r_{j,t}$(m).*summary* 包含子串，则认为两者摘要相似，将两者合并为 $r_{i,k}(n,\ m)$，并置 $r_{j,t}$(m).*flag*=1，若不包含，则不做处理；

若 $t<R_0(s).sum$，令 $t=t+1$，转向下一条结果，重做步骤（4.3）；若 $t=R_0(s).sum$，说明 $r_{i,k}$(n)与其后的所有结果均已比较完毕，则将 $r_{i,k}(n, m,...)$归入 $R(s)$中，并令 $i=i+1$，若 $i=R_0(s).sum$，转向步骤（4.4），否则重做步骤（4.3）；

（4.4）当 $i=R_0(s).sum$ 时，说明 $R_0(s)$中除最后一项 $r_{x,R0(s).sum}(y)$的所有条目均已进行过去除重复处理，若 $r_{x,R0(s).sum}(y).flag=1$，说明与前面的条目重复，不计入 $R(s)$中，否则将它归入 $R(s)$中；

（4.5）$R(s)$已包含所有去除重复处理后的搜索结果，由搜索结果预处理模块将这些搜索结果传给个性化排序模块进行后续处理。

用基于摘要内容的搜索结果去除重复方法对搜索结果进行处理的必要性在于：

对于经过基于网址的去除重复处理后的搜索结果，虽然它们的网址不同，但有些页面上的实际内容还是有可能很相似甚至完全相同，对于用户而言也属于重复结果，应该予以滤除。因此，经过网址去除重复处理之后，还要对搜索结果进行内容去除重复处理。而且利用元搜索技术可以得到搜索结果网页的标题的摘要，其中对于标题而言，相似与否并不能说明其内容是否相似，例如标题为“山西省人民政府网站”和“陕西省人民政府网站”的两个网页，它们的内容其实完全不同，而网页摘要虽然简短，但它是页面中与用户查询最相关的一部分信息，这些信息可以很好的反应网页的内容。而且往往是用户搜索到的许多网页虽然它们来源不同标题不同，但它们的内容很相似甚至完全相同，都是对一些已有信息的简单复制，这些网页对用户来说没有更多价值，在用户查找有用信息时还会造成干扰。所以通过分析网页摘要内容来判断内容相似度从而进行去除重复处理是一个提高用户搜索体验的必要过程。

步骤五，根据用户兴趣模型和本次搜索词语对去除重复处理后的搜索结果进行个性化排序，并将该排序结果提交给用户，具体步骤如图 5 所示：

（5.1）设置有关参数

将某用户的搜索词语 S 经过分词处理后得到的关键词集设为：$Q=\{key_1,key_2,...,key_h\}$，其中 key_i 表示第 i 个关键词，共有 h 个，且它们在查询语句中相对应的归一化重要度向量为 $X(Q)=(x_1, x_2,...,x_h)$，其中各分量之和为 1；

用户兴趣数据库的特征词基础数据表是各种特征词与兴趣类别的对应关系表，将 $K(c_i)$设为属于兴趣类别 c_i 的特征词集合；在特征词基础数据表中逐一查找 Q 中的关键词，得到分别所属的兴趣类别，再与用户的 $I(C)$对照，将用户本次查询的兴趣类别集合设为：$I(Q)=\{(c_1,w_1),(c_2,w_2),...,(c_m,w_m)\}\subseteq I(C)$；

（5.2）对于 $I(Q)$中的每个兴趣类别 c_i，分别计算 Q 中各个关键词权重向量

$W_Q(c_i)=(w_1,w_2,\ldots,w_h)$，其中 $w_j=\begin{cases} w_i, & t_j \in K(c_i) \\ 0, & t_j \notin K(c_i) \end{cases}$；

若 $W_Q(c_i)$中存在 w_j 不为零，则对 $X(Q)$和 $W_Q(c_i)$进行基于向量空间模型的相关度计算，得到搜索词语 Q 与兴趣类别 c_i 的相关度：$sim(Q,c_i)=\sum_{j=1}^{h}(w_j * x_j) \Big/ \sqrt{\sum_{j=1}^{h} w_j^2 * \sum_{j=1}^{h} x_j^2}$，表示 Q 与兴趣类别 c_i 的相关程度；若 w_j 全为零，则 $sim(Q,c_i)=0$；

（5.3）对用户兴趣模型中的所有兴趣类别进行相关度计算，得到搜索词语 Q 与用户兴趣的相关度向量 $Sim(Q,C)=(sim(Q,c_1),\ldots,sim(Q,c_m))$；

（5.4）对于搜索结果集合 $R(s)$中的每条记录 r_i，将 r_i 的标题和摘要分别进行分词处理，得到若干关键词，在特征词库中找出其中归属于 $I(Q)$中各兴趣类别的关键词集，表示为 $K_{title}(r_i)=\{key_1, key_2,\ldots, key_k\}$和 $K_{summary}(r_i)=\{key_1, key_2,\ldots, key_p\}$；

对于 $I(Q)$中的每个兴趣类别 c_j，逐个计算 $K_{title}(r_i)$ 和 $K_{summary}(r_i)$的权重向量 $W_{title}(r_i,c_j)=(w_1,w_2,\ldots,w_k)$和 $W_{summary}(r_i,c_j)=(w_1,w_2,\ldots,w_p)$，

式中，$w_i=\begin{cases} 0.6\times y_j, & t_i \in K(c_j) \\ 0, & t_i \notin K(c_j) \end{cases} (t_i \in K_{title}(r_i))$

$w_k=\begin{cases} 0.4\times y_j, & t_k \in K(c_j) \\ 0, & t_k \notin K(c_j) \end{cases} (t_k \in K_{summary}(r_i))$；

将 $K_{title}(r_i)$、$K_{summary}(r_i)$以及 $W_{title}(r_i,c_j)$、$W_{summary}(r_i,c_j)$分别合并为 $K(r_i)=(t_1,t_2,\ldots,t_n)$和 $W(r_i,c_j)=(w_1,w_2,\ldots,w_n)$，

式中，$K(r_i)$包含 $K_{title}(r_i)$ 和 $K_{summary}(r_i)$中的所有关键词，$W(r_i,c_j)$中的权重为 $W_{title}(r_i,c_j)$和 $W_{summary}(r_i,c_j)$中相应权重之和；

经过分词处理后，$K(r_i)$中包含的关键词在 r_i 中的归一化重要度向量为：

$X(r_i)=(x_1, x_2, \ldots, x_n)$；

若 $W(r_i,c_j)$中存在 w_i 不为零，则将 $X(r_i)$和 $W(r_i,c_j)$进行基于向量空间模型的相关度计算，得到搜索结果 r_i 与兴趣类别 c_j 的相关度 $sim(r_i,c_j)=\sum_{i=1}^{n}(w_i * x_i) \Big/ \sqrt{\sum_{i=1}^{n} w_i^2 * \sum_{i=1}^{n} x_i^2}$，表示搜索结果 r_i 与兴趣类别 c_j 的相似程度，若 w_i 全为零，则 $sim(r_i,c_j)=0$；

（5.5）对用户兴趣模型中的所有兴趣类别进行相关度计算，得到搜索结果 r_i 与用户兴趣的相关度向量 $Sim(r_i, C)=(sim(r_i, c_1),\ldots, sim(r_i, c_n))$；

（5.6）对所有搜索结果与所有用户兴趣类别的相关度进行计算，得到搜索结果集 R 与用户兴趣的相关度向量 $Sim(R, C)=(Sim(r_i, C),\ldots, Sim(r_i, C))$；

（5.7）计算 r_i 与 Q 在兴趣类别 c_j 上的相关度 $sim(r_i, Q, c_j)$：当 $Sim(Q,C)$中所有分量全为 0 时，$sim(r_i, Q, c_j)=sim(r_i, c_j)$，否则 $sim(r_i, Q, c_j)=sim(r_i, c_j)\times sim(Q, c_j)$；

（5.8）计算所有的用户兴趣类别与一条搜索结果 r_i 的相关度向量，得到相关度向

量 $Sim(r_i, Q, C) = (sim(r_i, Q, c_1), \ldots, sim(r_i, Q, c_n))$；

（5.9）将相关度向量 $Sim(r_i, Q, C)$进行综合处理，得到 Q 与 r_i 的相关度 $sim(r_i, Q) = \frac{1}{n}\sum_{c_j=1}^{n} sim(r_i, Q, c_j)$；

（5.10）计算所有搜索结果与搜索词语的相关度，得到搜索结果集 R 与搜索词语的相关度向量 $Sim(R,Q) = (sim(r_1,Q), \ldots, sim(r_n,Q))$；

（5.11）对于搜索结果 r_i，可按该式计算它在独立搜索引擎上的排名得分：$weight_{SE}(r_i) = 1 - \prod_{i=1}^{k}(1 - \frac{1}{k \cdot n_i})$，

式中，k 表示搜索结果包含 r_i 的独立搜索引擎的个数，n_i 表示在相应搜索引擎上的排名，该式表明 r_i 被越多的搜索引擎索引且在搜索引擎上排名越靠前则其得分较高；

（5.12）由于 $sim(r_i,Q)$和 $weight_{SE}(r_i)$均为归一化的数值，所以将两者按一定比例综合即可得到 r_i 的权值 $weight(r_i) = 0.6 \times sim(r_i,Q) + 0.4 \times weight_{SE}(r_i)$，该权值是该搜索结果的个性化权值 PersonalRank；

（5.13）按照 PersonalRank 的数值，由大到小对搜索结果进行排序，得到符合用户兴趣和搜索意图的排序方式，并按照此排序方式将搜索结果提交给用户。

步骤六，捕获用户点击的搜索结果，对其进行兴趣分析，改变兴趣种类和权值大小，并根据改变的结果更新用户兴趣数据库中用户的兴趣模型数据，其步骤如图 6 所示：

（6.1）通过在搜索结果显示页面设置特定代码，捕获用户在客户端点击的搜索结果，并传回服务器端；

（6.2）对传回的搜索结果 r_i 的标题和摘要分别进行分词，得到该搜索结果的标题和摘要关键词集 $K_{Title}(r_i) = \{key_{T1}, key_{T2}, \ldots, key_{Tk}\}$和 $K_{summary}(r_i) = \{key_{S1}, key_{S2}, \ldots, key_{Sm}\}$；

（6.3）对于 $K_{Title}(r_i)$和 $K_{summary}(r_i)$中的每个关键词 key_{Ti}和 key_{Si}，进行如下兴趣分析步骤：

（6.3a）在用户兴趣数据库的特征词基础数据表中查找 key_{Ti} 所属的兴趣类别，对找到的每个兴趣类别 c_i，若该用户兴趣模型中存在该兴趣类别，且其被涉及次数 $Count_{Ci} = m$，则将其更新为 $Count_{Ci} = Count_{Ci} + 1.2$，相应的权值更新为 $Weight_{Ci} = 0.1 \times \sqrt{\frac{(m+1.2+10)^2}{100}} - 1$；若找不到兴趣类别，则将这个兴趣分量加入用户兴趣模型中，且 $Count_{Ci} = 1.2$，$Weight_{Ci} = 0.1 \times \sqrt{\frac{(1.2+10)^2}{100}} - 1$；

（6.3b）在用户兴趣数据库的特征词基础数据表中查找 key_{Si} 所属的兴趣类别，对找到的每个兴趣类别 c_i，若该用户兴趣模型中存在该兴趣类别，且其被涉及次数 $Count_{Ci} = m$，则将其更新为 $Count_{Ci} = Count_{Ci} + 0.8$，相应的权值更新为 $Weight_{Ci} = 0.1 \times \sqrt{\frac{(m+0.8+10)^2}{100}} - 1$，若找不到兴趣类别，则将这个兴趣分量加入用户兴趣模型中，且 $Count_{Ci} = 0.8$，$Weight_{Ci} = 0.1 \times \sqrt{\frac{(0.8+10)^2}{100}} - 1$；

所述步骤（6.3a）和步骤（6.3b）中的权值计算公式的意义在于：被用户点击浏览涉及次数越多的兴趣类别的兴趣权值越大，且随着涉及次数的增多，这种权值增加的趋势会逐渐减缓，即用户的兴趣喜好逐渐趋于稳定；

（6.4）对于经过一定的更新次数门限后没有被更新过的兴趣分量，说明该用户对这些兴趣领域已经不再关注，将它们从用户兴趣模型中删除；

（6.5）对搜索结果 r_i 的标题和摘要中所有关键词处理结束之后，将该用户的所有兴趣权重进行归一化处理，变为取值在 0 和 1 之间且总和为 1 的数值，作为更新后的用户兴趣权值，对用户兴趣数据库中的相应兴趣权值进行更新。

本发明的效果通过以下仿真实例进一步说明：

1．去除重复搜索结果实例

在一个利用本发明所涉及的搜索引擎系统及其搜索方法实现的实例搜索引擎系统中，预先设置 79 个兴趣类别，并为每个兴趣类别设置若干能够代表该类别特征的特征词；设置每次搜索请求通过搜索引擎代理管理模块向百度抓取 50 条、谷歌抓取 50 条、有道抓取 10 条、搜狗抓取 20 条共 130 条搜索结果。

本实例中分别用 5 个不同的搜索词语在该搜索系统上进行搜索，经过本发明中的去除网址重复和基于摘要内容的搜索结果去除重复方法处理后，得到的统计结果如表 1 所示。

表 1 搜索结果去除重复数据

搜索次数	去重前总条数	网址去重后总条数	内容去重后总条数	内容去重覆盖度	内容去重准确度
1	130 条	120 条	93 条	27/31	27/27
2		122 条	93 条	29/34	29/29
3		111 条	105 条	6/6	5/6
4		121 条	111 条	10/12	10/10
5		128 条	92 条	36/43	36/36
平均		120.4 条	98.8 条	88.9%	96.67%

表 1 中，内容去重覆盖度为实际去重条目与应去重总条目之比；内容去重准确度为去重条目中正确去重条目与去重总条目之比。

2．个性化排序实例

在实例搜索系统中，分别设置一个兴趣爱好分布在“信息技术”这个类别的用户 A 和一个兴趣爱好分布在“个人电子产品”这个类别的用户 B，首先对这两个用户以基本的元搜索技术进行搜索，再对这两个用户以登录状态进行本发明的个性化搜索，得到表 2 的统计结果：

表 2 搜索结果排序数据

用户	兴趣类别	搜索词语	元搜索状态下前 30 条中相关条数	个性化搜索前 30 条中相关条数	个性化搜索响应时间
A	信息技术	MIMO	13 条	28 条	1.52s
		CDMA	9 条	16 条	1.373s
		无线	14 条	25 条	1.584s
B	个人电子产品	MP3	10 条	21 条	1.551s
		照相机	13 条	20 条	1.325s
		DVD	8 条	15 条	1.585s
平均值			11.2 条	20.8 条	1.49s

表 2 中，个性化搜索响应时间为服务器从接收搜索请求到向用户返回搜索结果间经历的时间间隔。

3．实例系统性能分析

从表 1 中的数据可以看出，对从四个独立搜索引擎抓取的 130 条搜索结果经过本发明的基于摘要内容的去除重复方法处理后，得到的搜索结果条数相比仅进行网址去重处理有了显著减少，内容去重覆盖度平均为 88.9%，内容去重准确度平均为 96.67%。

这说明本发明的内容去除重复技术可以准确的识别和去除重复的搜索结果，使搜索结果数量得到大幅度精简，从而免去了用户在大量重复的搜索结果中寻找有用信息的烦恼。

从表 2 中的数据可以看出，对于具有一定兴趣爱好的用户，在基本的元搜索情况下，得到的搜索结果排在前 3 页的 30 条搜索结果中满足其搜索需求的平均不足 12 条，而经过本发明的个性化搜索系统的处理之后，符合用户兴趣的搜索结果平均达到了 20.8 条。

这说明利用本发明中的基于用户兴趣的个性化排序技术实现的搜索引擎系统能够准确的识别用户兴趣，并能根据用户喜好为用户返回合适的搜索结果排序方式，这样使得用户在最靠前的搜索结果中找到感兴趣的内容的几率大大增加，从而提高了用

户信息检索的效率。

从搜索系统的响应时间上看，用户从提交搜索请求到服务器为用户返回搜索结果之间的时延平均约为 1.49 秒。据有关调查数据显示，中国网民认为打开网页的最佳速度应在 5 秒之内，而本发明的搜索系统的响应时间即使考虑服务器与用户端的通信时延，也完全可以满足用户这一要求，这说明利用本发明所涉及的技术实现的搜索系统具有实际可行性。

综合以上的性能分析，本发明包括的基于用户兴趣的个性化元搜索引擎及搜索结果处理方法，与传统搜索引擎相比，提高了搜索结果的覆盖度，克服了单个独立搜索引擎搜索结果覆盖度低的问题；与一般的元搜索引擎以及现有的个性化搜索技术相比，通过为每个用户建立各自的用户兴趣模型，并将其长期保存在服务器数据库中，而且随着用户的搜索过程对用户兴趣数据不断更新，使得用户不论身处何时何地，均能准确定位用户兴趣，为其提供个性化搜索服务，不仅克服了一般元搜索引擎不能提供个性化服务的缺点，而且克服了现有个性化搜索技术不能长期保存用户兴趣和不能精准定位个人兴趣的缺点。

本发明通过独创的引擎搜索结果处理机制将多个独立搜索引擎的搜索结果进行去除重复处理，并计算每条搜索结果的个性化权值 PersonalRank，为用户提供最适合其搜索意图和兴趣喜好的搜索结果排列方式，使得搜索结果的准确度得到显著提高，用户的搜索需求得到最大程度的满足，用户找到自己需要的搜索结果的难度大大降低。本发明的搜索系统是一种性能优越、实际可行的互联网信息检索领域的新技术。

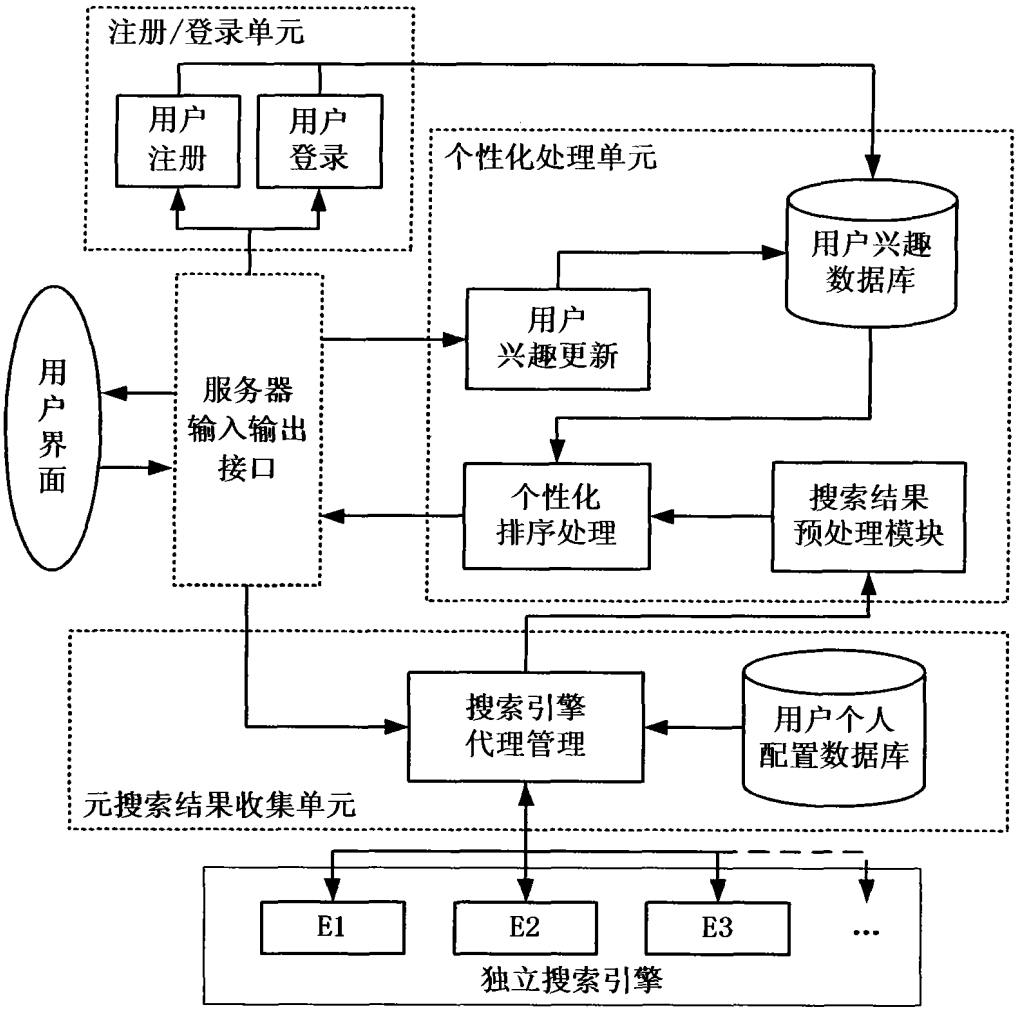

注册/登录单元
用户注册
用户登录
个性化处理单元
用户兴趣数据库
用户兴趣更新
用户界面
服务器输入输出接口
个性化排序处理
搜索结果预处理模块
搜索引擎代理管理
用户个人配置数据库
元搜索结果收集单元
E1
E2
E3
...
独立搜索引擎

图 1

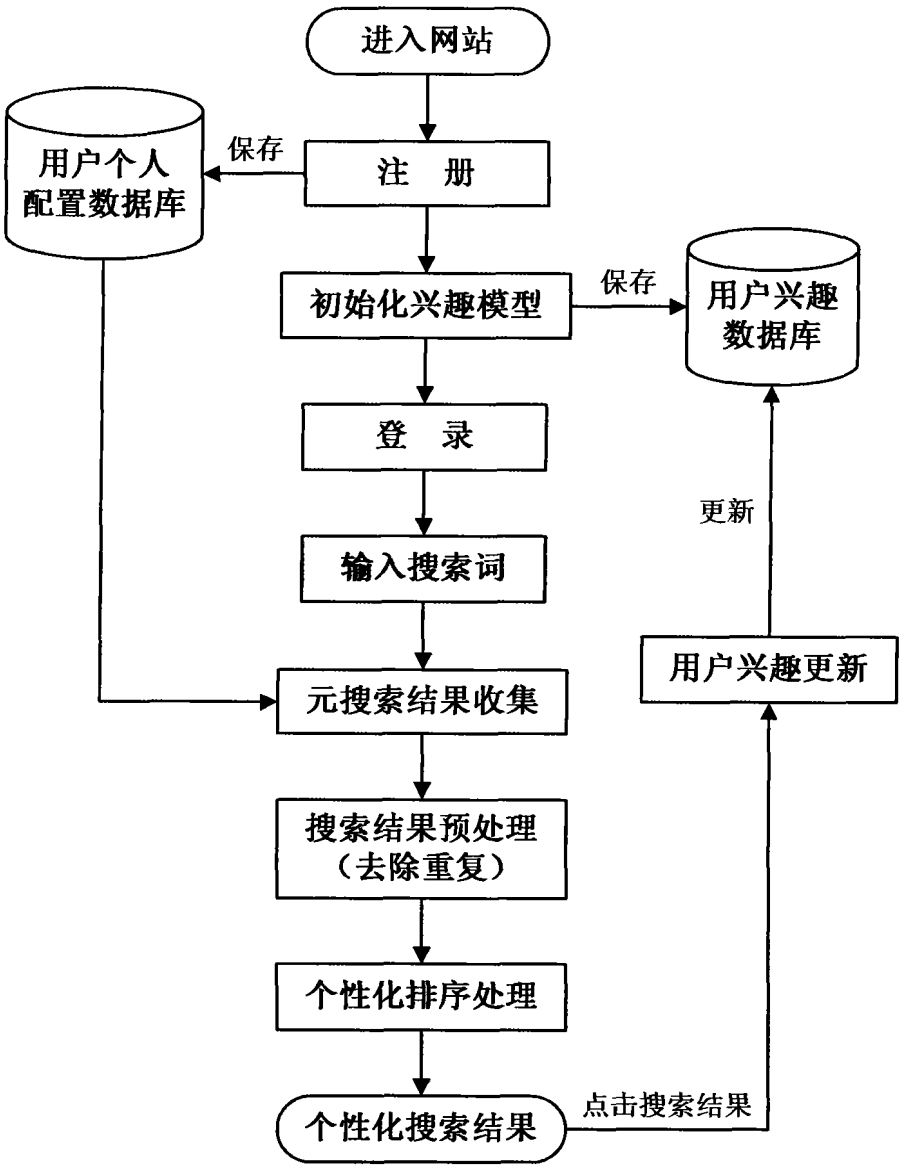

进入网站
用户个人配置数据库
保存
注 册
初始化兴趣模型
保存
用户兴趣数据库
登 录
更新
输入搜索词
用户兴趣更新
元搜索结果收集
搜索结果预处理（去除重复）
个性化排序处理
个性化搜索结果
点击搜索结果

图 2

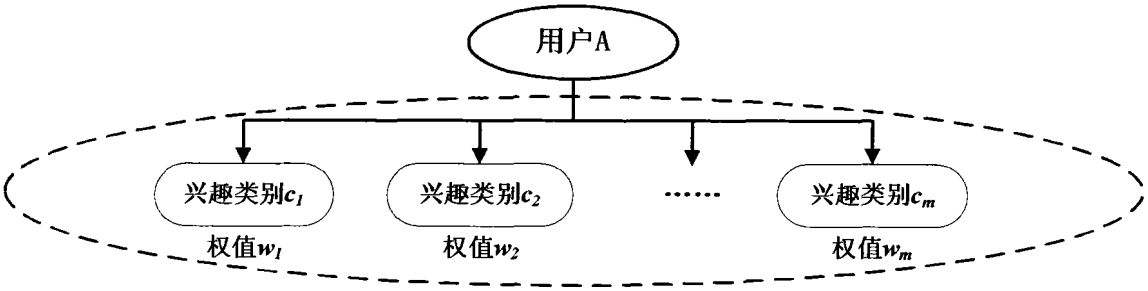

用户A
兴趣类别c_1
权值w_1
兴趣类别c_2
权值w_2
……
兴趣类别c_m
权值w_m

图 3

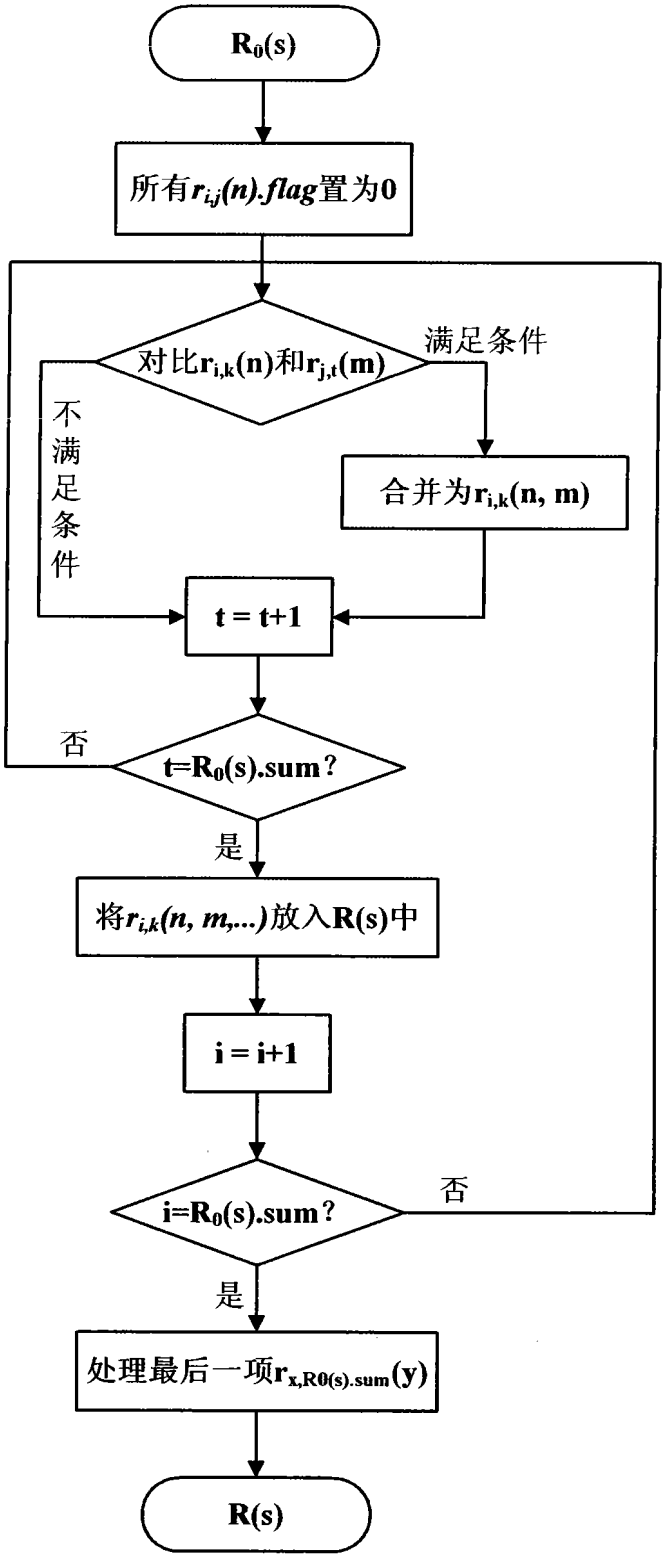

R0(s)
所有ri,j(n).flag置为0
对比ri,k(n)和rj,t(m)
满足条件
不满足条件
合并为ri,k(n, m)
t = t+1
否
t=R0(s).sum?
是
将ri,k(n, m,...)放入R(s)中
i = i+1
i=R0(s).sum?
否
是
处理最后一项rx,R0(s).sum(y)
R(s)

图 4

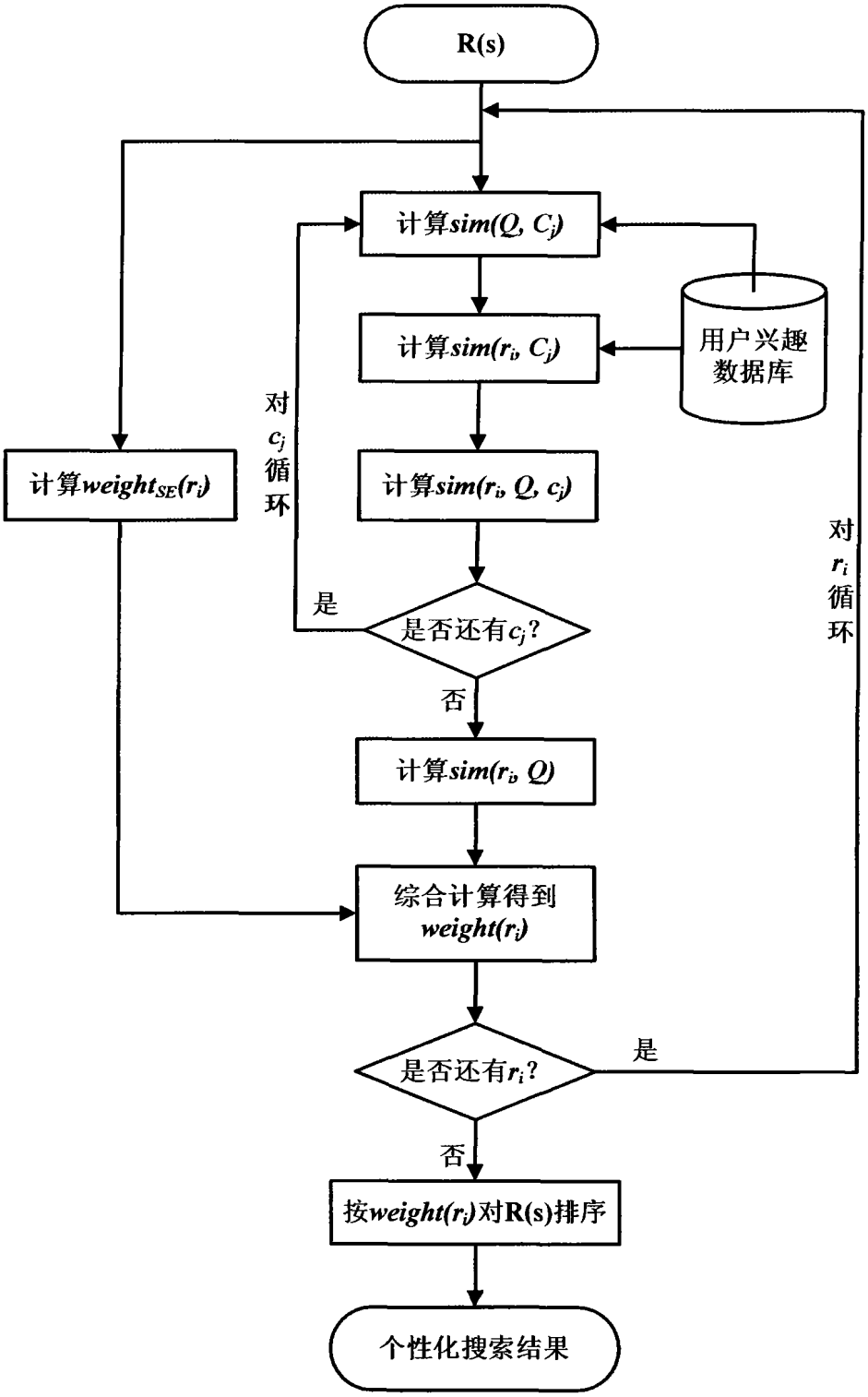

R(s)
计算sim(Q, Cj)
计算sim(ri, Cj)
用户兴趣数据库
计算sim(ri, Q, cj)
对cj循环
是
是否还有cj?
否
计算weightSE(ri)
计算sim(ri, Q)
综合计算得到weight(ri)
对ri循环
是
是否还有ri?
否
按weight(ri)对R(s)排序
个性化搜索结果

图 5

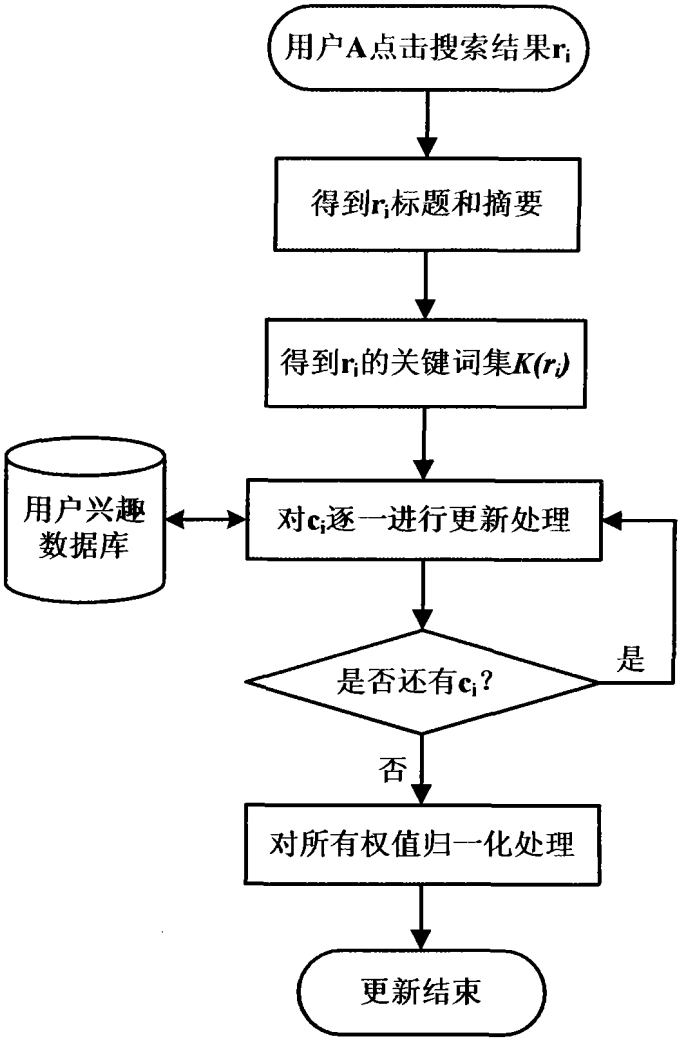

用户A点击搜索结果r_i
得到r_i标题和摘要
得到r_i的关键词集$K(r_i)$
用户兴趣数据库
对c_i逐一进行更新处理
是否还有c_i?
是
否
对所有权值归一化处理
更新结束

图 6