(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 102236655 A
(43)申请公布日 2011.11.09

(21)申请号 201010155971.5

(22)申请日 2010.04.26

(71)申请人 北京大学
地址 100871 北京市海淀区颐和园路 5 号

(72)发明人 张岩

(74)专利代理机构 北京北新智诚知识产权代理有限公司 11100
代理人 赵郁军

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

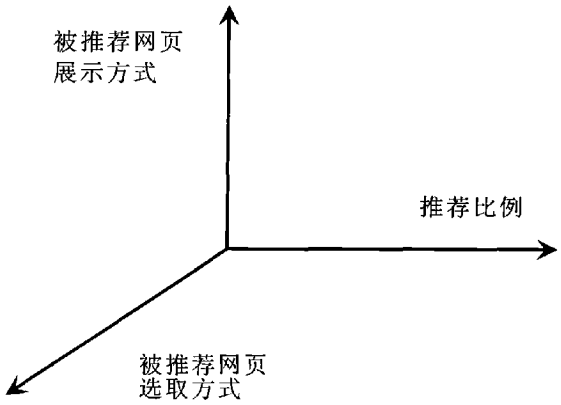
权利要求书 1 页 说明书 7 页 附图 2 页

(54)发明名称

一种 Web 新网页推荐方法

(57)摘要

本发明公开了一种 Web 新网页推荐方法。在该方法中，对于新网页，首先确定新网页出现在搜索引擎返回结果中的推荐比例；然后估测新页面的潜在质量，根据潜在质量的高低以预定的概率选取被推荐的新网页；将被推荐的新网页和搜索引擎返回结果一起展示给用户。本发明可以在尽量不影响搜索结果质量的前提下，向搜索引擎用户推荐一些高质量的新页面，使得它们获得被关注和点击的机会，在较短的时间内获得较高的认知度，通过用户的点击对新页面的质量进行自动评判，降低搜索引擎对新页面流行度演化过程的阻滞作用，使有价值的新页面可以“健康”地成长。



CN 102236655 A

1. 一种 Web 新网页推荐方法，其特征在于包括如下步骤：

对于新网页，首先确定所述新网页出现在搜索引擎返回结果中的推荐比例；然后估测新页面的潜在质量，根据潜在质量的高低以预定的概率选取被推荐的新网页；将被推荐的新网页和搜索引擎返回结果一起展示给用户。

2. 如权利要求 1 所述的 Web 新网页推荐方法，其特征在于：

所述推荐比例为 15％～ 20％。

3. 如权利要求 1 所述的 Web 新网页推荐方法，其特征在于：

使用新网页当前的 PageRank 结果来估测新页面的潜在质量。

4. 如权利要求 1 所述的 Web 新网页推荐方法，其特征在于：

使用与某一个新网页具有相同父网页的所有兄弟网页 PageRank 的平均值来估测新页面的潜在质量。

5. 如权利要求 4 所述的 Web 新网页推荐方法，其特征在于：

所述平均值通过如下步骤获得：首先对每个网页计算其子网页 PageRank 的平均值，再通过子网页 PageRank 的均值计算兄弟页面 PageRank 的平均值。

6. 如权利要求 1 所述的 Web 新网页推荐方法，其特征在于：

使用与某一个新网页具有相同父网页的所有兄弟网页 PageRank 的中值来估测新页面的潜在质量。

7. 如权利要求 1 所述的 Web 新网页推荐方法，其特征在于：

在将被推荐的新网页和搜索引擎返回结果一起展示给用户时，将被推荐的新页面与原来搜索引擎返回结果混合在一起，没有区分地展示给用户。

8. 如权利要求 1 所述的 Web 新网页推荐方法，其特征在于：

在将被推荐的新网页和搜索引擎返回结果一起展示给用户时，将被推荐的新页面以显著不同于原有搜索引擎返回结果的方式展示给用户。

一种 Web 新网页推荐方法

技术领域

[0001] 本发明涉及一种推荐 Web 新网页的方法，尤其涉及一种针对搜索引擎排名算法的弱点，在尽量不影响搜索结果质量的前提下，向用户推荐高质量的新网页的方法，属于互联网搜索技术领域。

背景技术

[0002] 当前，搜索引擎已经渐渐成为人们获取信息的主要方式之一。当搜索引擎把查询结果返回给用户时，查询结果的排列方式对结果被用户关注和点击的概率具有绝对的影响。那么应该如何对结果合理排名呢？在结果与查询关键词的匹配度一样的情况下，最理想的排名方案应该是按照结果页面质量从高到低排列。然而页面质量 (page quality) 是一个相当主观的概念。它需要通过网络用户的主观判断而得到，但是首先人工的评判需要大量的人力物力财力，耗费大量的时间。其次，由于不同的个体对于同一个 Web 页面的质量可能会得出不同的评价，使得人工网页评价成为相当不实际的做法。

[0003] 在此背景下，搜索引擎排名算法的设计者转而考虑用一些客观的易于被观察和计算的特征值来替代和近似估计一个网页的重要性程度。因而，Web 页面流行度 (page popularity) 的概念被提出了。Web 页面流行度代表一个网页被用户喜欢的程度，可以用页面获得的 inlink 数或者点击次数来计算。1998 年 S. Brin 和 L. Page 等人提出了利用 Web 链接结构图来获得页面重要性的算法，即 PageRank 算法。PageRank 算法很好地利用了“群体智慧”，使搜索引擎的性能大为提高，是当前排名算法中的佼佼者。因此，PageRank 算法目前仍然是绝大多数商业搜索引擎的排名算法的基础。

[0004] 基于网页流行度的排名算法虽然能够帮助网络用户搜寻有用的信息，但它同时也引发了一些弊端。Web 是一个规模巨大、新旧页面不断更替的动态环境，其中每个新页面都会经历一个流行度从小到大，直到与其本身质量相一致的成长过程。当用网页流行度来近似获得网页质量时，Web 新页面的质量往往被低估了。特别是当搜索引擎开始主导用户的浏览模式时，新页面被访问的概率大大下降，流行度的成长过程被大大的延长了。

[0005] 为了解决新网页容易被搜索引擎所忽视的问题，S. Pandey 和 S. Roy 等人提出了一种 Shuffling 方法。该方法的本质是剥夺一个成熟页面被展示的机会，而把这个机会让给某个被随机推荐出来的新页面。由于 Shuffling 方法在选取被推荐的新页面时采用的是随机法，因此被推荐的新页面的质量可能参差不齐，所以很可能造成一个质量不佳的新页面占据了良好的展示位置出现在结果集合中，却没有能为用户带来有用信息的局面。虽然 Shuffling 方法在理论上对于新页面提升认知度有一定作用，但实际应用的效果难以得到保障。

[0006] 在提高搜索结果时效性方面也有很多类似的技术。例如在专门针对新闻时事的搜索引擎中，搜索结果通常都是按照页面发布的时间或者时新度 (freshness) 排名的，因为新闻搜索引擎的用户更关注的是最新报道，根据结果时效性来对结果排名是最符合用户需求的。页面时效性对于新闻搜索排名是有实际意义的，但无法用于评价新网页的质量，因此

并不适合作为推荐普遍意义上的新页面的衡量标准。

发明内容

[0007] 本发明所要解决的技术问题在于提供一种 Web 新网页推荐方法。该方法可以使新页面有更多机会被用户认知，有助于优化搜索引擎的排名，提高用户满意度。

[0008] 为了实现上述的发明目的，本发明采用下述的技术方案：

[0009] 一种 Web 新网页推荐方法，其特征在于包括如下步骤：

[0010] 对于新网页，首先确定所述新网页出现在搜索引擎返回结果中的推荐比例；然后估测新页面的潜在质量，根据潜在质量的高低以预定的概率选取被推荐的新网页；将被推荐的新网页和搜索引擎返回结果一起展示给用户。

[0011] 其中，所述推荐比例为 15％～ 20％。

[0012] 在估测新页面的潜在质量时，使用的方法包括以下三种：

[0013] (1) 使用新网页当前的 PageRank 结果（简称 Naive 方法）

[0014] (2) 使用兄弟页面 PageRank 的均值（简称 ASP 方法）

[0015] (3) 使用兄弟页面 PageRank 的中值（简称 MSP 方法）

[0016] 将被推荐的新网页和搜索引擎返回结果一起展示给用户时，可以采用的页面展示方法包括显式展示法和隐式展示法，其中隐式展示法是将被推荐的新页面与原来的 10 个结果混合在一起，没有区分的展示给用户；显式展示法是将被推荐的新页面以某种显著的方式展示给用户。

[0017] 本 Web 新网页推荐方法可以在尽量不影响搜索结果质量的前提下，向搜索引擎用户推荐一些高质量的新页面，使得它们获得被关注和点击的机会，在较短的时间内获得较高的认知度，通过用户的点击对新页面的质量进行自动评判，降低搜索引擎对新页面流行度演化过程的阻滞作用，使有价值的新页面可以“健康”地成长。

附图说明

[0018] 下面结合附图和具体实施方式对本发明作进一步的详细说明。

[0019] 图 1 为三种估测网页潜在质量的方法的对比试验结果示意图；

[0020] 图 2 为本发明所提供的 RankPro 推荐方法从三个维度确定新网页推荐方式的示意图；

[0021] 图 3 为摄影图片网站实验中，三种推荐方法在每个类别的 POH 值；

[0022] 图 4 为摄影图片网站实验中，三种推荐方法在每个类别的 AOR 值。

具体实施方式

[0023] 现有研究表明，页面流行度并不总是与页面本身的质量成正比，而是在开始时总是较低，随着页面的慢慢成熟，获得较多的用户关注和点击之后逐渐成长，直到最终与页面本身的质量一致。对于 Web 上每一个页面，都存在这样一个流行度成长的过程，这个过程同时也是页面本身从一个新生页面过渡到成熟页面的过程。

[0024] 因此，用当前的网页流行度对网页进行排序，即使是高质量的新网页，也会由于其流行度火候不足而受到压制，排名靠后。把流行度位于不同起点上的新旧网页混在一起进

行排名，对新网页是不公平的。同时，人们往往倾向于搜索较新的信息，这会降低搜索结果的用户满意度。再则，新页面由于排名靠后，被访问的机会很低，被认可的机会也就低了，其流行度的成长就会受到严重阻碍。

[0025] 从搜索引擎的角度分析基于网页流行度的排名算法存在的问题，需要考虑以下两点：

[0026] (1)Web 的动态性

[0027] Web 是一个动态的环境，不断有网页产生、更新和消亡，其链接结构也随着不断变化。这是毋庸置疑的，问题在于变化的速度和程度如何。如果 Web 上页面更替很小很慢，那么由于新网页流行度和质量不匹配造成的搜索性能降低很小，没有必要花很大的代价去寻找新的排名技术并整合到排名策略中；如果页面更替频繁，新页面出现的速率很高，那么搜索性能可提高的空间就很大。所以，Web 的动态程度决定了是否有必要寻找新的排名方法。

[0028] (2) 搜索引擎排名的影响

[0029] 当前主要的搜索引擎回应用户查询的方式是返回一个 URL 的有序列表。如果用户对返回结果排序的依赖比较小，那么新网页即使排得稍微靠后些，被访问的可能性也不会降低太多，对搜索性能和新网页流行度的成长不会造成太大影响；反之，如果用户严重依赖结果的排名，那么搜索引擎就应该考虑对新网页的排名进行一定的提升。

[0030] 现有研究和发明人所做的观察实验都表明：Web 页面演化的速度相当快；搜索引擎用户对搜索结果的依赖很强，特别对于返回结果中排名比较靠前的，特别是前二、三十个结果最为关注，而排名靠后的页面很难获得点击。

[0031] 鉴于 Web 页面和搜索引擎的上述特点，向用户推荐高质量的新网页需要考虑如下的三个问题：

[0032] ●如何决定被推荐新页面个数同原来搜索结果之间的推荐比例？

[0033] ●如何决定哪些新页面应该被推荐？

[0034] ●被推荐页面应以怎样的方式与原排名结合并展示给搜索引擎用户？

[0035] 下面分别进行详细的说明。

[0036] 1. 推荐比例的确定

[0037] 推荐比例指的就是每十个返回结果要搭配多少个被推荐的新网页一同展示给用户。例如在 Shuffling 方法中，每 10 个返回结果，推荐一个新网页展示给用户，那么 Shuffling 方法的推荐比例为 10％。

[0038] 我们认为推荐比例是影响推荐技术效果的重要因素之一。因为推荐比例直接影响被推荐页面被用户关注的概率。推荐比例越小，每次返回结果时被一起展示给用户的新网页的个数就越少。因此推荐比例至少应该大于 10％，使得每次返回结果时，至少有一个 Web 新页面能够被推荐给用户。

[0039] 然而推荐比例并不是越大越好的。首先，当推荐比例过大时，会对原有搜索结果造成较大的影响，尤其是当被推荐页面的质量不能得到保证时（如 Shuffling 方法采用随机法选择被推荐页面），过大的推荐比例会使得整体的搜索质量下降。其次，当推荐比例过大时，页面上呈现的结果个数过多，会造成用户浏览页面的负担，继而产生对推荐页面的反感。

[0040] 通过实验和计算，发明人认为 15％～ 20％的推荐比例是比较适宜的。

[0041] 2. 被推荐页面的选取

[0042] 被推荐页面的选取是推荐方法的一个重要步骤。在本发明中，选取被推荐页面的方法有以下两种：

[0043] (1) 随机选取

[0044] 在候选集合中随机挑选某些新页面进行推荐。其优点在于，操作简单，算法复杂度低。但是缺点显而易见，那就是被推荐页面的质量很不稳定，使得返回结果整体质量会受到影响。

[0045] (2) 根据潜在质量按概率挑选

[0046] 首先根据一定的方法估计新页面的潜在质量（或者给出新页面按质量从高到低的一个排序），然后根据潜在质量的高低以预先确定的概率进行选取。新页面潜在质量估计方法的好坏对性能会有很大的影响。但无论如何，总有一些的信息（比如 Web 链接结构）可以或多或少地提示一个新页面的潜在质量的高低，总比随机选取强。这种推荐方法的优点在于，被推荐页面的质量相对有所保证，搜索结果的整体质量也相应得到保障，推荐页面不会产生很大的噪音。但缺点在于需要较大的计算量，如果是需要实时推荐的话，可能会增加响应时间。

[0047] 在估测新页面的潜在质量时，可用的方法包括以下三种：

[0048] (1) 直接使用新网页当前的 PageRank 结果（简称 Naive 方法）

[0049] 这种估计方法假设：那些新网页当前的 PageRank 就能够比较好地反映新网页将来的质量，至少能够反映新网页之间的相对质量。

[0050] (2) 兄弟页面 PageRank 的均值（简称 ASP 方法）

[0051] 首先假设页面的质量分布具有一定的局部性，具有同一父页面或者祖先页面的若干个页面的质量存在一定的关联，即页面倾向于链向具有相似质量的页面。当一个新的页面被发布出来，如果它的兄弟网页质量普遍高，那么它很可能是一个高质量的网页。这样可以通过新页面的兄弟页面的质量来预测它本身的质量。

[0052] 基于以上假设，可以把与某一个新网页具有相同父网页的所有网页的 PageRank 的平均值作为它的质量的估计值。但是，如果个别父网页有非常多的链出网页，那么这些网页会对 ASP 值产生很大的偏移。为了消除这个影响，首先对每个网页计算其子网页 PageRank 的均值 (theAverage of Children PageRank，简称 ACP 值)，再通过 ACP 值计算 ASP 值，以使得每个父网页的权重等价。

[0053] $$ACP(q)=\frac{\sum_{q\to p}PR(p)}{outdegree(q)},\ ASP(p)=\frac{\sum_{q\to p}ACP(q)}{indegree(p)} \qquad (1)$$

[0054] (3) 兄弟页面 PageRank 的中值（简称 MSP 方法）

[0055] 与 ASP 方法类似，把与某一个新网页具有相同父网页的所有网页的 PageRank 的中值作为它的 PageRank 估计值。

[0056] 同样地，可以通过计算 MCP 来求 MSP，

[0057] MCP(q) = Median of{PageRank(p)|q → p}

[0058] (2)

[0059] MSP(p) = Median of{PageRank(q)|q → p}

[0060] 为了比较上述三种估测方法的实际使用效果，发明人进行了对比实验，结果如图 1

所示。从图 1 可以看出，普通的 PageRank 值基本上不能体现出新网页的潜在质量，而 MSP 方法和 ASP 方法则可以在一定程度上体现新网页的质量，特别是 ASP 的效果更好。

[0061] 3. 如何展示被推荐的新页面

[0062] 在确定了推荐比例和被推荐页面选取方法之后，接下来要选择合适的展示方法来向用户推荐新页面。

[0063] 本发明中可以采用的页面展示方法包括显式展示法和隐式展示法：

[0064] 隐式展示法是将被推荐的新页面与原来的 10 个结果混合在一起，没有区分地展示给用户。这种方式对于用户搜索行为没有影响，用户会自然的点击访问被推荐页面，但是如果被推荐页面的质量不高，有时甚至是 Spam 页面时，由于用户对其的期待较高，反而会导致用户对搜索结果质量的不满。

[0065] 显式展示法是将被推荐的新页面以显著不同于原有搜索引擎返回结果的方式展示给用户，譬如用不同的颜色标注结果及其摘要；或者在被推荐结果边上添加“推荐新页面”标识符；抑或在结果页面下方设置一个“新发现的相关结果”栏，将被推荐的新页面放置其中。

[0066] 综上所述，本发明所提供的 Web 新页面推荐方法是通过在三个维度上选取不同的方法进行组合而来的，即选取一定的推荐比例、一定的选择被推荐页面的方法和一定的被推荐页面的展示方法，如图 2 所示。为了便于与已有技术相区别，发明人将其命名为 RankPro 推荐方法（也称概率显示推荐法）。

[0067] 在本发明的一个具体实施例中，RankPro 推荐方法的详细内容如下：

[0068] ●推荐比例：每页原有 10 个结果加上推荐 2 个新页面，共 12 个结果；

[0069] ●被推荐页面选取方式：根据新网页相对排名按照一定概率选取；

[0070] ●被推荐页面展示方式：把新页面列在原 10 个结果之后显示展示给用户。

[0071] 与现有的 Shuffling 方法相比较，本发明所提出的 RankPro 推荐方法不会剥夺成熟页面被展示的机会，只是在原有结果基础之上，在同一返回页面中以适当的方式向用户若干符合需求的但较新的页面，这样的推荐方法在总体上不会影响返回结果的质量，因此不会存在 Shuffling 方法中存在的问题。

[0072] 为了验证 RankPro 推荐方法的有效性，发明人进行了一系列的实验。通过与基准方法以及另一种推荐方法的比较，结果显示 RankPro 推荐方法能显著提高新页面被关注的程度。

[0073] 以摄影图片网站实验 (Photograph Website Experiment，简称 PWE) 为例，发明人模拟并对比了三种推荐方法（参见表 1），分析在这些推荐方法下，新网页被访问的概率以及被访问的新网页的平均质量。

[0074]

推荐方法	选取策略	展示策略
基准方法	无	无
随机隐含推荐法	随机选取	隐式补充插入
RankPro 方法	按照新页面相对排名以概率选取	显示补充续接

[0075] 表 1

[0076] 发明人从一个摄影爱好者网站 (www.altphotos.com) 上下载了 6912 张摄影图片，以及每张图片的缩略图、原图和简要的说明信息。这 6912 张图片分属于 6 个类别，每个类别大约 1000 张。从每个类别中，随机抽取出 600 张标记为新图片，并将其分值降为原始的 1/5。这样做的目的是来模拟新页面刚刚产生时较低的流行度即 PageRank 值。

[0077] 在 PWE 实验中，共有三种推荐方法参与评价 (参见表 1)。这三种推荐方法的具体实现细节如下 :

[0078] (1) 基准方法 :

[0079] 推荐方式 :不对新网页进行推荐，作为比较实验结果的基准。

[0080] 排序方式 :所有图片以评分值的大小从高到低排列。

[0081] 展示方式 :每页 6 行，每行 2 个显示单元。

[0082] (2) 随机隐含推荐法 :

[0083] 推荐方式 :从所有的新图片中随机选择 10 个来推荐。

[0084] 排序方式 :旧图片以分值的大小从高到低排序，新图片以被选择的次序排序。

[0085] 展示方式 :前 5 页结果中，每页 6 行，每行 2 个显示单元，其中包含 10 个旧图片和 2 个被推荐新图片 ;从第 6 页开始，每页 5 行，每行 2 个显示单元，10 个均为旧图片。在前 5 页中，10 个旧图片的显示单元按其评分值大小排序，2 个被推荐的新图片随机选择除第 1 个位置以外的其他位置插入到旧图片序列中。

[0086] (3)RankPro 推荐方法

[0087] 推荐方式 :所有的新图片以其评分值的大小从高到低排序，以 $P(r) = c*r^{-3/2}$(r 为排序位置) 为概率选择第 r 个图片 ;总共选择 10 个。

[0088] 排序方式 :旧图片以分值的大小从高到低排序，新图片以被选择的次序排序。

[0089] 展示方式 :前 5 页结果中，每页先列出旧图片，共 5 行，每行有 2 个显示单元 ;随后再添加一行，在其中列出被推荐新图片的 2 个显示单元，同时注明是“新发现的一些作品”。推荐完所有 10 个新页面后从第 6 页起，每页 5 行，每行 2 个显示单元，10 个均为旧图片。

[0090] 在为期 47 天的实验中，来自 455 个 IP 的用户访问了 PWE 实验网站，共计 2572 次的访问量和 3734 个点击记录。相应的实验结果如图 3 所示。从图 3 可以看出，进行推荐后，新图片被访问的可能性大大地提高了，随机隐含推荐法和 RankPro 推荐方法的效果都比较好。其中 RankPro 推荐方法更好一些。这可能有两个原因，一是因为显示地标注了“新发现的图片”，更容易吸引用户的注意力 ;二是因为通过概率选取的图片的质量较高，容易收到用户的关注。

[0091] 以下进一步对随机隐含推荐法和 RankPro 推荐方法进行对比分析。假设已经获得了新图片的排名，一个最简单也很直观的衡量推荐算法质量的方法是以被点击的新图片排名的平均值作为比较。AOR 定义如下 :$AOR_i = average(the\ rank\ of\ P_i)$，其中 P_i 是被 Groupi 中用户点击过的任意新页面

[0092] 同样的，我们分析 Average-of-Ranking(AOR) 在总体和每个类别上的情况，

[0093] $AOR_{i,k} = average(the\ rank\ of\ P_{i,k})$，

[0094] 其中 $P_{i,k}$ 是类别 k 中被 Group i 用户点击过的任意新页面

[0095] 从图 4 中显示的结果数据可以看出，在 RankPro 推荐方法中被点击的新图片的排名远比随机隐含推荐法的靠前。这点很容易解释，当使用 RankPro 推荐方法时，那些本身排

名较高的页面会以较高的概率被推荐给用户，所以被访问新页面的平均质量就相应较高。而这正是我们的出发点：如果我们可以间接或者直接地获得或者估计新图片或者新页面的质量，为何不利用它来提高推荐方法的质量？如果仅是从数以千／万计的新网页中随机选取提升的对象，新网页被认知的可能性确实提高了，却也会在一定的程度上降低搜索引擎返回结果的整体质量；同时，由于所有的新网页站在同一起跑线上，反而降低了质量高的网页被认知的速度。因而，利用已有的信息较好地估计新网页的质量然后以一定概率地进行推荐是更好的选择。

[0096] 以上对本发明所提供的 Web 新网页推荐方法进行了详细的说明。对本领域的一般技术人员而言，在不背离本发明实质精神的前提下对它所做的任何显而易见的改动，都将构成对本发明专利权的侵犯，将承担相应的法律责任。

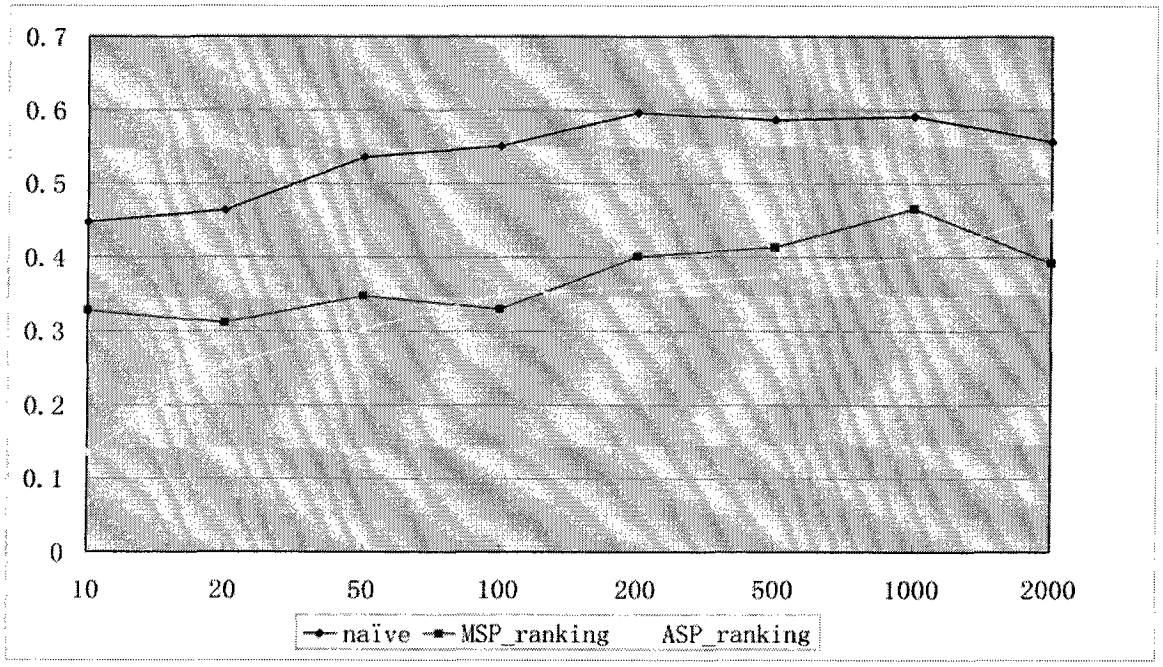

0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
10
20
50
100
200
500
1000
2000
naïve
MSP_ranking
ASP_ranking

图 1

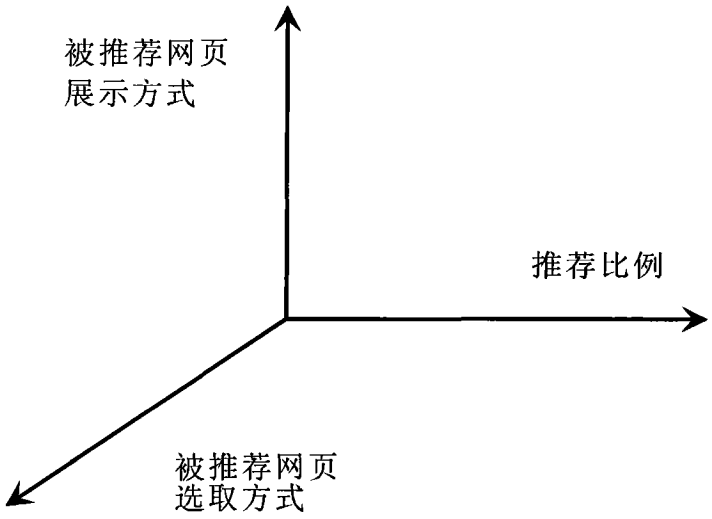

被推荐网页
展示方式
推荐比例
被推荐网页
选取方式

图 2

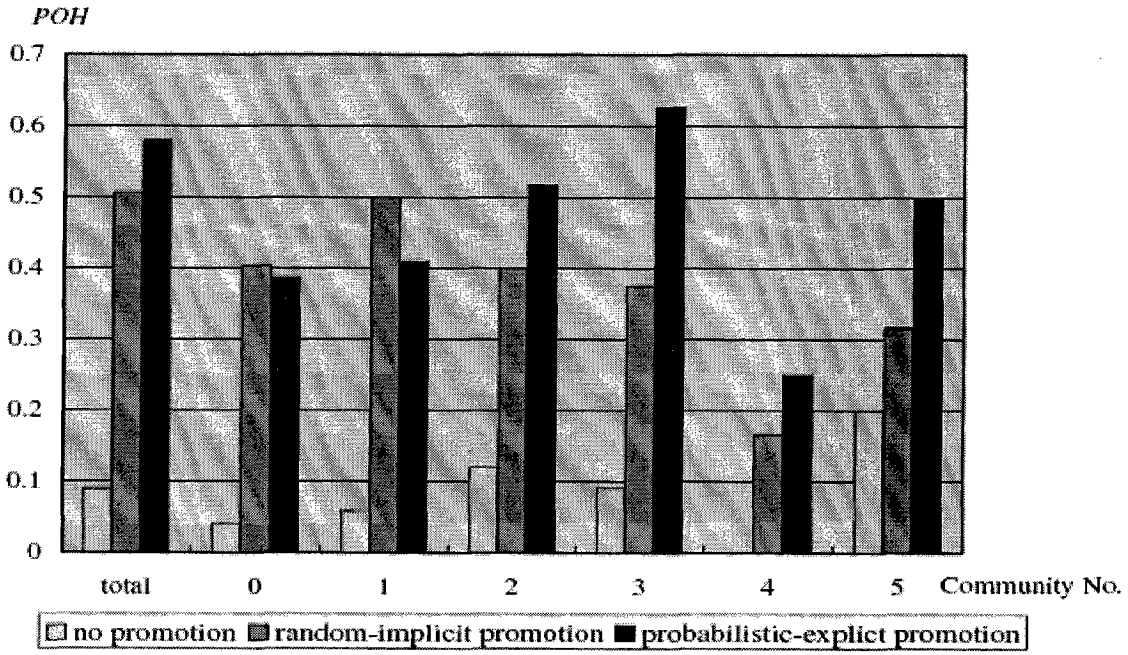

POH
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
total
0
1
2
3
4
5
Community No.
no promotion
random-implicit promotion
probabilistic-explict promotion

图 3

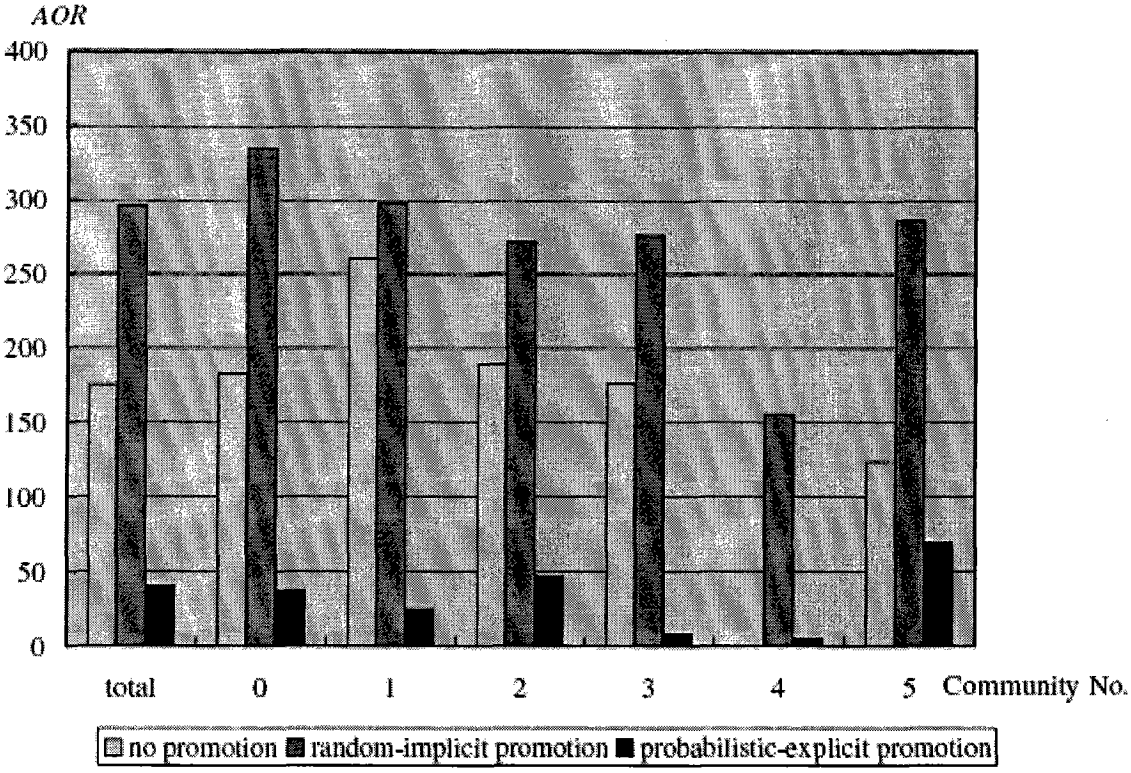

AOR
400
350
300
250
200
150
100
50
0
total
0
1
2
3
4
5
Community No.
no promotion
random-implicit promotion
probabilistic-explicit promotion

图 4