一种基于用户属性的搜索算法①

余 坦 王益民 (湖南大学 软件学院 湖南 长沙 410082)

摘 要： 如何提高搜索的准确度是搜索算法改进的一个重要方面。提出了一种基于用户模型的搜索模式相适度算法，首先定义了一系列搜索模式，在搜索模式中有与之匹配的用户模型和相应的信息排序策略。用户检索时，利用空间向量模型计算出与用户最匹配的搜索模式，然后根据所选搜索模式中定义的排序算法，将最切合用户信息需求的数据优先呈现给用户，将用户属性作为信息重排的依据，达到针对用户的具体情况提供个性化搜索服务的目的。实验证明，本算法在可接受的时间消耗下，极大地提高了信息搜索的准确度。

关键词： 搜索模式;用户属性;个性化搜索;相适度算法

User-Attribute Based Search Algorithm

YU Tan, WANG Yi-Min (Software School，Hunan University, Changsha 410082, China)

Abstract: How to improve the accuracy of the search algorithm is important for its search algorithm improvement. This paper researches an algorithm based on the user's model. First, it defines a group of Search-Models.Each Search-Model contains a User-Model and an information resequence tactic. When the user does a searching, Vector Space Model will be used to choose an optimal Search-Model, and then uses the information resequence tactic defined in the Search-Model to resequence the information found by user's input. Finall, it sends the information to the user making user attribute as the premise to filter and re-sequence the information. It provides accurate search service according to the user's specific conditions. Results show that the algorithm can improve the accuracy with an acceptable time consumption.

Keywords: search-model; user attribute;personalized search; compatibility algorithm

1 引言

对搜索算法的改进主要有两个方面，一个是更快，谋求的是在更短的时间内为用户提供所需的信息；一个是更准，谋求的是更精确的满足用户的信息需求。从目前搜索算法的研究情况来看，如何“更快”的问题已经得到了很大程度的解决，但是如何“更准”还存在一定的困难，困难的主要根源在于无法准确掌握用户的信息需求。

当前主流的搜索引擎大多是采用的以关键词索引为基础的检索，总体上来说智能化程度较低，对用户意图的判断主要依赖于用户输入的检索词，而在实际情况中，相同的问题，由不同的人提出来，目的可能是不一样的。目前主流的搜索引擎对不同类型的用户没有任何区分，已经开展的智能搜索研究也主要集中于对语义的理解、自然语言的理解以及用户检索习惯的挖掘上[1-3]。这些努力确实在很大程度上提高了搜索的准确程度，但是仍然很有限，主要是对用户搜索意图的判断没有根本的改善。

如果可以获得用户的背景资料，进行推理，作为判断用户信息需求的一个依据，可以在一定程度上提高对用户意图把握的准确性，进而提高信息搜索的准确性。本文提出了一种以用户属性为过滤条件的搜索算法，通过对用户的各种属性进行分析，计算用户模型并选择匹配度最高的信息组织策略，以求实现因人而异，更准确的为用户提供所需信息。

① 收稿时间:2009-07-24;收到修改稿时间:2010-04-02

2　搜索模式选择

搜索模式的选择以用户模型为基础，搜索模式中包含对应的用户模型和所采用的信息排序策略。在搜索时，首先用纯机械匹配的方法找出所有相关信息，然后根据所能获得的用户属性，利用空间向量模型[4]与搜索模式中的用户模型进行匹配，匹配度越高，则搜索模式越符合用户需求。找到最佳搜索模式后，将所有相关信息按最佳搜索模式所定义的排序策略进行重排，最后将重排后的信息展示给用户。

2.1 相关概念

搜索模式：根据用户属性，为向用户提供差异化的搜索服务而建立的信息搜索和排序策略体系。搜索模式的形式化定义为：S=(SPattern；UFeature；Seq- Tactic)。其中：

SPattern：搜索模式名称。

UFeature：用户特征模型，对搜索者基本信息、使用习惯、背景资料等各种信息的表示，通过将用户的各种属性按一定的权重进行组合而成。形式化表现为一个用户特征向量：UFeature ={(a_1,r_1),(a_2,r_2), · · · ,(a_n,r_n)}，其中 a_i 为用户特征属性，r_i 为对应的权值。

SeqTactic：排序策略，在指定搜索模式下对信息进行过滤和排序的策略。

2.2 基本流程

采用本算法的基本检索流程如图 1 所示：

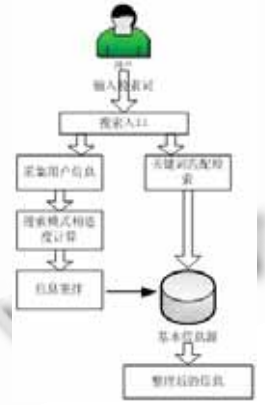

图 1　搜索基本流程

(1) 用户输入检索词开始检索；

(2) 根据用户输入的检索词进行粗查，找到所有的相关信息，按与检索词的匹配度为序作为基本信息源；

(3) 用户模型对搜索模式相适度计算；

(4) 根据最佳搜索模式定义的信息排序策略对找到的信息进行整理；

(5) 将整理后的信息展示给用户。

其中关键步骤为用户模型对搜索模式的相适度计算和信息的重组排序。

2.3 用户模型与搜索模式的相适度算法

(1)首先选择一个要与用户进行匹配的搜索模式，从该模式中抽取 UFeature 定义的用户特征属性向量{a_1,a_2,a_3, · · · ,a_n}(a_i 为特征属性)，a_i 的取值和对应权重分别为 v_i 和 p_i，其中：

$$\sum_{i=1}^{n} p_i = 1$$

则 a_i 的复合权值为 $u_i=v_i*p_i$，所以信息组织策略 S_k 所对应的用户特征属性向量空间为：S_k={(a_{k1}，u_{k1})；(a_{k2}，u_{k2})；(a_{k3}，u_{k3})；· · ·；(a_{kn}，u_{kn})}。

(2)用户模型：从用户数据库中抽取符合 UFeature 定义的用户实际特征属性，并将每个属性按对应的复合权值组合成用户特征属性向量空间：U={(a_1,w_1)；(a_2,w_2)；· · ·；(a_n,w_n)}。

(3) 相适度计算：在空间向量模型中，两个向量的夹角越小表示这两个向量的相适度越高。通常用计算两个向量夹角余弦值的方式来确定夹角的大小，余弦值越大，夹角越小，相适度越高。

记用户模型与信息组织策略之间的夹角余弦值为 Cos(U，S)，则此余弦值就是用户模型与信息组织策略的相适度 α：

$$\alpha(S_k) = \mathrm{Cos}(S,U) = \frac{\sum_{i=1}^{n}(u_i \times w_i)}{\sqrt{\sum_{i=1}^{n} u_i^2} \times \sqrt{\sum_{i=1}^{n} w_i^2}}$$

α 值最大的用户模型就是搜索者所匹配的用户模型，此用户模型对应的信息排序策略就是对于该用户最佳的信息排序策略。

3　信息排序算法

在找到最佳搜索模式后，根据最佳搜索模式所定义的信息排序策略，对基本信息源进行过滤和重排，优化后的信息就是本算法所得到的最符合用户需求的信息。具体的排序算法如下：

(1)搜索模式中的 SeqTactic 中定义了一系列资源特征属性{a_1，a_2，a_3，· · ·，a_n}(a_i 为资源特征属性)，a_i 的权重为 p_i，其中：

$$\sum_{i=1}^{n} p_i = 1$$

如果所选信息 R 具有特征属性 a_i，则令 $a_i=1$，不具备属性 a_i，则令 $a_i=0$，那么 R 的最终特征值为：

$$R=\sum_{i=1}^{n} a_i p_i \quad (0\le R\le 1)。$$

(2)对基本信息源中的信息条目，提取符合 SeqTactic 中定义的属性{a_1，a_2，a_3，…，a_n}，计算信息的 R 值。将基本信息源中的所有信息按 R 值从大到小排列，构成将要向用户推送的信息组合。

4 实验说明

作者参与了某公司承接的大型信息系统项目《某省电力公司技术管理资料支持帮助系统》的研发，并在项目中对算法进行了实验和应用。使用某省电力公司技术管理资料信息库总量大约 28 万条各种类型的技术管理资料作为测试数据。搜索引擎采用目前应用非常广泛的开源项目 lucene 进行改进[5]，实验目的有两个：一是验证算法对搜索准确度的提高程度，二是验证算法为提高准确度而增加的时间开销。

实验中定义了四个搜索模式，请来四位测试人员通过两个搜索入口进行对比测试。

四个搜索模式分别为：S1—研究者模式、S2—管理者摸索、S3—的技工模式、S4—咨询者模式。S1 主推与搜索内容相关的研究信息，S2 主推与搜索内容相关的管理信息，S3 注重为搜索者提供能解决技术工作中所遇问题的信息，S4 为搜索者提供与搜索关键词相关的介绍性信息。

四位测试人员分别为：T1—电力试验研究院从事变压器研究的高级工程师、T2—电力公司科技信息部副主任、T3—某变电站生产班组副组长、T4—电力公司档案科工作人员。

两个搜索入口分别为：E1—采用本文算法的搜索、E2—未采用本文算法的普通搜索。

(1) 在准确度测试中，测试人员使用相同的关键词分别通过两个入口进行搜索，将测试人员认定的第一页 25 条数据中有用数据的比率作为数据有效率。表 1 为数据有效率测试结果：

表 1 准确度对比

测试者	E1	E2
T1	88%	40%
T2	92%	52%
T3	84%	60%
T4	80%	68%

(2) 在时间消耗测试中，记录每位测试人员对每个搜索入口进行同一搜索所耗费的时间。表 2 为时间消耗对比，单位为毫秒(ms)：

表 2 时间消耗对比

测试者	E1（/ms）	E2（/ms）
T1	87	52
T2	92	59
T3	65	42
T4	62	34

通过上述数据可以看出，采用本算法可以大幅提高搜索准确度，但同时也带来了一定的时间损耗。总的来看，准确度的提高非常可观，而时间的损耗也在可接受的范围。

5 总结

搜索准确度提高的关键在于用户意图的判断，在特定的应用环境下如果可以获得用户的各种背景信息，可以利用用户的这些信息来更准确的判断用户的信息需求，进而更准确的为用户提供所需的信息。本文算法的应用场景就是大型企业的内部网络，所有用户都是实名，且所需的用户信息都是可得的。通过对用户属性进行分析，推断用户的信息需求，然后将搜索到的数据按用户感兴趣的程度进行重新排序，将用户最需要的信息优先呈送给用户，以可容忍的时间损耗大幅提高搜索准确度，是提高信息搜索准确度的积极探索。

参考文献

1 柯佳,程显毅.面向用户的智能搜索引擎模型 UOISE 的研究.计算机工程与应用, 2006,35:175－177,232.
2 陈林,杨丹.基于语义理解的智能搜索引擎研究.计算机科学, 2008,30(6):152－154.
3 文振威,秦晓.个性化搜索引擎的研究与设计.计算机工程与设计, 2009,30(2):342－344,394.
4 Salton G, McGill MJ. Introduction to Moderm Information Retrieval. McGraw-Hill, 1983.
5 索红光,孙鑫. 基于 Lucene 的中文全文检索系统的研究与设计.计算机工程与设计, 2008,29(19):5083－5086.