基于智能 Agent 的多维权值信息检索模型

徐小龙[①]　王汝传[①②]
[①](南京邮电学院计算机科学与技术系　南京　210003)
[②](南京大学计算机软件新技术国家重点实验室　南京　210093)

摘　要：为了弥补目前信息检索系统中存在的资源耗费大、实效性不高、用户满意度低等缺点，该文提出并实现一种新型的基于智能 Agent 的多维权值信息检索模型。该模型主要思想是在信息检索系统中应用了智能 Agent 技术，通过将检索任务分担到用户客户机、检索服务器和被检索主机的方式以达到提高检索实效性和节省系统和网络资源的目的；该文还通过综合考虑用户检索偏好等特征以及信息本身的重要度和检索匹配程度，提出了一种多维权值排序算法 MWRA，提高检索的排序能力，从而为用户提供符合其个性化特征的检索结果。该文首先描述了基于智能 Agent 的多维权值信息检索模型，然后深入分析了多维权值排序算法 MWRA，最后对模型的性能进行了比较和分析。实验结果和性能分析结论表明，基于智能 Agent 的多维权值信息检索模型的性能尤其在排序能力等方面有明显的提高。

关键词：信息检索；搜索引擎；Agent；多维权值排序算法
中图分类号：TP391　　**文献标识码**：A　　**文章编号**：1009-5896(2008)02-0482-04

The Agent-Based Information Retrieval Model with Multi-weight Ranking Algorithm

Xu Xiao-long[①]　Wang Ru-chuan[①②]
[①](*Department of Computer Science and Technology, Nanjing University of Posts and Telecommunications, Nanjing* 210003, *China*)
[②](*State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing* 210093, *China*)

Abstract: In this article, a brand-new information retrieval model based on Agent technology is proposed, which is to counteract some significant deficiencies existing in current information retrieval systems, such as resource consuming much, information updating delayed and so on. The main idea of this model is to apply Agent technology into the information retrieval system in order to provide users a information retrieval model of new pattern, which could update on time, save resource through distributing the information retrieval task among clients, retrieval severs and information owners. And MWRA(Multi-weight Ranking Algorithm) is also proposed in this article to improve the ranking capability of the information retrieval system, which is based on several facts, including the inclination of users, the importance of information and the matching of query. In this article, we firstly introduced the agent-based information retrieval model. Then the multi-weight ranking algorithm, which included in the model, is analyzed. Finally, the performance of the model is discussed and the prototype system is tested through a series of examinations, the result of which addresses the Agent-based information retrieval model with MWRA is better in ranking and other capabilities.

Key words: Information retrieval; Search engine; Agent; Multi-Weight Ranking Algorithm(MWRA)

1　引言

目前的信息检索系统(如搜索引擎)大都是将远程站点上的内容全部或者部分下载到本地，然后进行处理，其中存在大量无用的信息，而且难以保证搜索结果的实效性；目前的搜索引擎(如 Google 等)采用的排序算法主要是通过对 Web 网页的链接情况进行分析，从而对网页的质量和重要度进行评价，这种机械、单维的排序方法忽略了用户本身的个性化特征，常常将与用户完全无关的信息排在前面，使得用户满意度不高。

本文正是针对目前信息检索系统中不足之处，将智能 Agent[1]技术应用于信息检索系统中，并通过综合考虑用户检

2006-09-07 收到，2007-08-13 改回
国家自然科学基金(60573141，60773041)、江苏省自然科学基金(BK2005146)、江苏省高技术研究计划(BG2005037、BG2005038、BG2006001)、国家高科技 863 项目(2006AA01Z201、2006AA01Z219、2007AA01Z404)、南京市高科技项目(2007 软资 106，2007 软资 127) 、现代通信国家重点实验室基金(9140C1101010603)和江苏省计算机信息处理技术重点实验室基金(kjs06006)资助课题

索偏好特征以及信息本身的重要度和检索匹配程度，提出多维权值排序算法 MWRA，并构建了基于智能 Agent 的多维权值信息检索模型，该模型的目标是为用户提供符合其个性化特征的新型检索模型，并通过 Agent 技术达到提高检索实效性和节省系统和网络资源的目标。

2　基于智能 Agent 的多维权值信息检索模型

2.1 模型的体系架构

基于智能 Agent 的信息检索系统的运行平台分为 3 个区域：检索用户域、检索服务器域、被检索主机域，其上运行着基本相同的 Agent 服务器和不同功能的 Agent，系统的整体模型如图 1 所示。

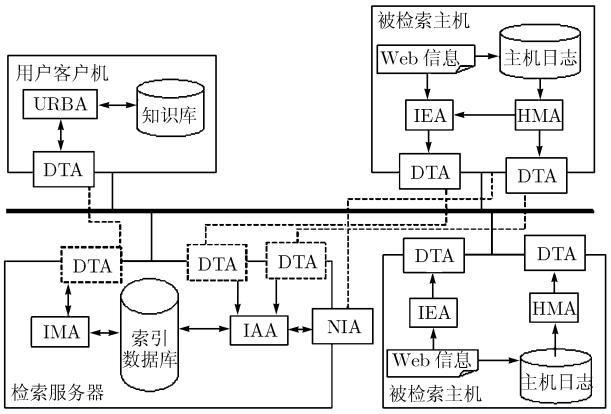

图 1 基于智能 Agent 的信息检索系统模型

根据 Agent 间数据流和指令流的内在层次关系，系统中设计了以下 7 种 Agent：(1)用户检索代理 Agent(User Retrieval Broker Agent，URBA) 驻留于检索用户客户机上，负责管理用户的身份和检索偏好等个人信息，代替用户提交能尽量反映用户真实需要的检索请求，并将检索结果个性化的呈现给用户；(2)数据传递 Agent(Data Transfer Agent，DTA) 负责在检索用户客户机和检索服务器之间传递检索请求、检索结果，在检索服务器和被检索主机之间传递关键信息等数据；(3)信息抽取 Agent(Information Extract Agent，IEA) 驻留于被检索主机上，能够主动分析整理主机的信息资源，从中提取资源的关键信息，并通过派生出的 DTA，传递到检索服务器上；(4)信息分析 Agent(Information Analysis Agent，IAA) 驻留于检索服务器上，将由 IEA 派来的 DTA 所携带的数据，再进行分类整理，并按照一定的方式存储至索引数据库中，或是更新索引数据库中内容；(5)主机监控 Agent(Host Monitoring Agent，HMA) 驻留于被检索主机上，负责监视主机运行和信息资源的变动情况，并将主机的性能等运行情况通过 DTA 主动报告给检索服务器上的 IAA，将信息变动情况即时通知本地的 IEA；(6)信息匹配 Agent(Information Matching Agent，IMA)驻留于检索服务器上，负责接收由 URBA 派来的 DTA 所携带的检索请求，从索引数据库中搜索出匹配的信息，再按照检索请求的限定的排序策略，对结果进行排序，反馈给 DTA；(7)网络巡视 Agent(Network Inspecting Agent，NIA) 由检索服务器派发，在网络各主机节点间漫游，主动发现各被检索主机的当前情况，防止由于主机宕机造成索引数据库中存在无效信息，或是无法及时新的资源节点。

2.2 与现有检索模型的比较

根据该模型，本文实现了一个原型系统 AIRS(Agent-based Information Retrieval System)，采用 JAVA 作为开发语言，并采用 IKV++的 Grasshopper 作为 Agent 的开发和运行平台。基于智能 Agent 的多维权值信息检索模型采用驻留于被检索主机上的信息抽取 Agent IEA 和主机监控 Agent HMA 来替代目前的搜索引擎常常采用的网络蜘蛛(Crawler)[2]。这样，一方面检索服务器获得的都是当前的最新信息，达到很高的实效性；另一方面，无用的信息和旧信息无需传递，降低了网络的负载。同时系统设计了网络巡视 Agent NIA，NIA 主动发现各被检索主机的当前情况，进一步保证了系统的实效性。该模型采用 Agent 技术将原来集中于检索服务器上的功能模块分布于检索用户机、检索服务器和被检索主机之上，这样就避免检索服务器承担所有的计算成为性能的瓶颈。

3　多维权值排序算法(MWRA)

3.1 MWRA 和 Mw 的提出

搜索引擎进行信息检索时，需要将最符合用户检索需求的网页排在最前面。目前已经有了一些网页排序算法来满足上述需求，如 Brin 和 Page 提出的 PageRank 算法[3]，J. Kleinberg 提出的 HITS 算法，Lempel 和 Moran 提出了 SALSA 算法，以及 Borodin 等提出了完全的贝叶斯统计方法来确定 Hub 和 Authoritive 网页[4]。这些算法基本是针对网页的超链接情况进行分析，从普遍意义的角度来推测用户对网页的需要从而进行排序；还有一些算法利用信息检索模型中检索结果的自然匹配度来对网页进行排序，如向量检索模型[5]。但这些都没有真正从检索用户个体需要的角度来考虑。在基于智能 Agent 的信息检索系统模型中，本文提出多维权值排序算法 MWRA 对检索结果进行排序。首先对多维检索衡量权值作以下的形式化描述：

$$\mathrm{Mw} \rightarrow \langle R,\ S,\ I,\ T\rangle \tag{1}$$

其中 Mw 为多维检索衡量权值(Multi-weights)，其中 R 为信息本身的客观重要度权值，通过分析 Web 超链接的引用(reference)情况获得；S 为用户输入的检索关键词与网页的匹配度(similarity)；I 为用户自身的检索偏好(inclination)与信息的归属度值；T 为信息更新的时间(time)。下面将对每个参数的确定方法进行描述。

3.2 R 值的确定

在本文的 MWRA 算法中吸取了 PageRank 算法。设 d

是一个 Web 页面，$F(d)$ 是 d 引用的页面集合($|F(d)|$为集合 $F(d)$ 的大小)，$B(d)$是引用 d 的页面集合，由此确定表示页面 d 客观重要度的 R 值：

$$R(d)=\sum_{d'\in B(d)}(R(d')/|F(d)|) \tag{2}$$

3.3 S 值的确定

用户输入的检索关键词是最能反映用户的查询愿望的因素。在本文提出的检索模型中，信息获取模块采用向量空间模型。该模型将文档映射为一个特征向量：

$$\boldsymbol{V}(\boldsymbol{d})=(t_1,\omega_1(d);\cdots;t_n,\omega_n(d)) \tag{3}$$

式(3)中 $t_i(i=1,2,\cdots,n)$ 为一列独立词条项，$\omega_i(d)$ 为 t_i 在 d 中的权值，一般被定义为 t_i 在 d 中的词频 f_{id} 的函数[6]。由 TF-IDF 函数可得：

$$\omega_i(d)=\frac{f_{id}\log\left(\dfrac{N}{n_i}\right)}{\sqrt{\sum_{k=1}^{n}(f_{kd})^2\times\log^2\left(\dfrac{N}{n_k}\right)}} \tag{4}$$

式(4)中的 N 为信息库中文档的数目，n_i 为含有词条 t_i 的文档数目。由于两文档之间的相似度可以用其对应的向量之间的夹角余弦来表示，在将关键词查询串 Q 进行向量化后，可得到包含 n 个关键词 q_i 的查询向量 Q 与页面 d 之间的匹配程度 S 值：

$$S(Q,d)=\frac{\sum_{i=1}^{n}\omega_i(d)\times q_i}{\sqrt{\left(\sum_{i=1}^{n}\omega_i^2(d)\right)\left(\sum_{i=1}^{n}q_i^2\right)}} \tag{5}$$

式(5)中，关键词 q_i 出现在页面 d 中则值为 1，否则为 0。

3.4 I 值的确定

接下来需要确定的是Mw中标识用户检索偏好与信息的归属度的 I 值。假定用户的使用偏好的总集合为 Interest，其中包含了若干偏好：

$$\text{Interest}=\text{In}_1\cup\text{In}_2\cup\text{In}_3\cup\cdots\cup\text{In}_i\cup\cdots \tag{6}$$

$$\text{In}_i=\{k_1,k_2,k_3,\cdots,k_r,\cdots\} \tag{7}$$

式(6)，式(7)中的 In_i 为每一类偏好集合，集合中的 k_r 元素是包含于 In_i 类偏好集合的单词。k_r 在 In_i 中的权值为 $\omega_i(\text{In}_i)$，权值大小的差异是由于 k_r 在各个偏好集合中分布的程度决定：如果 k_r 仅存在于其中某一个 In_i 中，就表明由 k_r 可完全确定该偏好 In_i，权值自然最大；如果 k_r 分布于每个偏好集合，则 k_r 不具备确认任何偏好的能力，权值自然最小。$\omega_i(\text{In}_i)$ 是事先从语料库中的文档中计算出来。文档的种类由 Interest 指定，这样文档对于 In_i 就可分为相关文档和无关文档：

$$\frac{P(\text{Rel}|\text{Doc})}{P(\text{Notrel}\,|\,\text{Doc})}\geq 1 \tag{8}$$

式(8)中的 $P(\text{Rel}|\text{Doc})$表示文档 Doc 与 In_i 有关的条件概率，$P(\text{Notrel}|\text{Doc})$表示文档 Doc 与 In_i 不相关的条件概率。URBA 通过与用户交互方式，获取用户选定的偏好子集合 SubInterest，由 m 个单类偏好集合组成，代表该用户感兴趣的 m 个类别：

$$\text{SubInterest}=\text{In}_1\cup\text{In}_2\cup\text{In}_3\cup\cdots\cup\text{In}_m \tag{9}$$

式(9)中，SubInterest $\subset$ Interest 。URBA 根据用户的偏好子集合 SubInterest，利用概率统计模型[7]，可得 k_r 在 In_i 中的 $\omega_i(\text{In}_i)$ 的值：

$$w_r(\text{In}_i)=\frac{t/(T-t)}{(n-t)/(N-n-(T-t))} \tag{10}$$

式(10)中的 N 是语料库中与偏好子集合 SubInterest 相关的文档的数量，n 是语料库中与偏好集合 SubInterest 相关且包含单词 k_r 的文档的数量，T 是与 In_i 相关的文档数目，t 是与 In_i 相关而且包含 k_r 的文档数目。URBA 根据用户当次搜索输入可得到包含 n 个关键词 q_i 的查询串 Q 和偏好子集合 SubInterest 来确定该用户当次检索偏好。Q 与 In_i 之间的相关度值为

$$\text{Sim}(Q,\text{In}_i)=\frac{\sum_{r=1}^{n}\omega_r(\text{In}_i)\times q_r}{\sqrt{\left(\sum_{r=1}^{n}\omega_r^2(\text{In}_i)\right)\left(\sum_{r=1}^{n}q_r^2\right)}} \tag{11}$$

从 SubInterest 中选出与 Q 相关度最高的 t 个 In_i，查询偏好名 Lln_i 是 In_i 的名称。将当次检索偏好及其与 Q 的相关度组合，用集合方式描述如下：

$$\begin{aligned}\text{CurIn}=\{&(\text{Lln}_1,\text{Sim}(Q,\text{ln}_1)),(\text{Lln}_2,\text{Sim}\,(Q,\text{ln}_2)),\cdots,\text{Lln}_t,\\&\text{Sim}\,(Q,\text{ln}_t))\}\end{aligned} \tag{12}$$

由式(12)，URBA 获得可通过 DTA 传递给 IMA 的检索偏好参数 CurIn。由于 IEA 从被检索主机的文档中抽取的关键信息中与检索偏好直接对应的为类别，但主题、摘要和主要内容等里面也常常包含了查询偏好名，由此得到用户的当次的检索偏好与页面 d 的归属度值 I：

$$\begin{aligned}I(\text{CurIn},d)=&\alpha\sum_{i=1}^{t}\text{Sim}(Q,\text{In}_i)\times s\text{LIn}_i\\&+\beta\frac{\sum_{i=1}^{t}\omega_i(b)\times b\text{LIn}_i\times\text{Sim}(Q,\text{In}_i)}{\sqrt{\left(\sum_{i=1}^{t}\omega_i^2(b)\right)\left(\sum_{i=1}^{t}(b\text{LIn}_i\times\text{Sim}(Q,\text{In}_i))^2\right)}}\end{aligned} \tag{13}$$

式(13)中，查询偏好名 Lln_i 出现在页面类别中则 $s\text{Lln}_i$ 为 1，否则 $s\text{Lln}_i$ 为 0；查询偏好名 Lln_i 出现在页面的内容 b 中(包括主题、摘要和主要内容)则 $b\text{Lln}_i$ 为 1，否则为 $b\text{Lln}_i$ 为 0。α 、β 为调节查询偏好名 Lln_i 与类别的匹配程度和与内容的归属程度主次关系的参数，应以与类别的匹配程度为主。$\omega_i(b)$ 为 Lln_i 在 b 中的权值，这依据式(4)计算得出。

3.5 Mw 值的确定

最后讨论信息更新的时间 T 值的问题。由于用户有指定搜寻某个时间段信息的需求，即以时间为参数提高查精率，因此这个 T 值由用户自行指定，在搜索的结果中由系统简单的进行判断是否选取即可；或是用户希望将信息按更新时间的先后顺序进行排序。在这里，本模型中采取的策略是在利

用 R、S 和 I 值得出 Mw 值并综合排序后，再利用 T 值对于相同的 Mw 值的页面按时间先后进行排序。因此，T 值虽作为排序的权值参数之一，但不参与 Mw 值的计算。根据式(2)，式(5)和式(13)，得出 Mw 值的计算公式：

$$\mathrm{Mw}(Q,d) = R(d)\times(1+\nu S(Q,d)+\lambda I(\mathrm{CurIn},d)) \qquad (14)$$

式(14)中的 υ，λ 为调节查询内容匹配度和检索偏好两者间重要性程度关系的参数，这两个参数需要通过实验来不断的优化确定。

4　实验与性能分析

4.1 实验过程与实验结果

本文重点是对系统多维权值排序算法 MWRA 的排序能力进行了实验比较。在实验中反复检索 150 次，分为 5 大组，每大组让用户选择一次检索偏好，每大组分为 10 小组，每小组以相同关键词查询串检索 3 次，对应基于单独 R 值(情况 1，单纯基于网页链接的网页重要度，即令式(14)中的 ν，λ 均为 0)、基于 R 值和 S 值(情况 2，基于网页链接的网页重要度结合基于向量空间模型的检索匹配度算法，即令式(14)中的 λ 为 0)和基于 R 值、S 值和 I 值(情况 3，基于多维权值排序算法 MWRA)这 3 种情况下排序结果的变化。首先，设 B_{ti} 为第 t 小组 (t=1,2,…,50) 第 i 种(i=1,2,3)情况下检索出的所有文档中用户真正需要的文档的集合，集合中的元素个数用 $|B_{ti}|$ 表示。设 C_{ti} 为第 t 小组第 i 种情况检索出的前 20 个文档中，用户真正需要的文档的集合，集合中的元素个数用 $|C_{ti}|$ 表示。

$G_i(t)$ 是指第 t 小组第 i 种情况下检索时，前 20 个文档中用户真正需要的文档个数和检索出的所有文档中用户真正需要的文档的个数之间的比值，由此评价算法的排序能力。当 $|B_{ti}|=0$ 时，$|C_{ti}|$ 必然为 0，令此时 $G_i(t)=1$。结果基于单独 R 值(情况 1)的情况下，$G_1(t)$ 的平均值为 60.6%；基于 R 值和 S 值(情况 2)的情况下，$G_2(t)$ 的平均值为 71.24%；基于 R 值，S 值和 I 值(情况 3)的情况下，$G_3(t)$ 的平均值为 75.88%。，这 3 种情况下平均查精 $P(t)$结果测得均为 32.6%。

4.2 性能分析

下面从查全率、查精率、排序能力、检索实效性和检索速度这 5 个方面对基于智能 Agent 的多维权值信息检索模型的性能进行分析。上述实验结果表明基于多维权值排序算法 MWRA 的排序能力得到了明显的提高。而且，由于检索偏好参数单独提交，避免了通过简单的查询扩展的方式带来的查全率或查精率的下降。本模型采用了多维权值排序算法 MWRA，增加了结果排序的计算量。但是检索任务已经被负载分担了：Mw 中标识用户检索偏好与信息的归属度的 I 值的大部分计算量都交由用户客户机来完成，检索服务器只确定了 I 最终值；同时，模型中信息过滤等繁重工作都由被检索主机自行完成，所以检索服务器总的负担降低，由此即使文档信息数量增加，其检索速度不至于明显下降。

5　结束语

综上所述，基于智能 Agent 的多维权值信息检索模型可以弥补目前信息检索系统中存在的种种不足之处。而需要进一步研究的是：如何充分体现 Agent 的智能性特征，如用户检索代理 Agent URBA 不但能获取用户的当前检索偏好，应能够通过用户的检索经验的反馈信息，不断训练，自动修正用户真正的检索偏好和已获得检索偏好参数之间的误差；如何保障信息交互过程的安全，由于该模型将功能分布于检索用户机、检索服务器和被检索主机之上，信息通过数据传递 Agent DTA 进行交流，而 DTA 是移动 Agent，这就涉及到移动 Agent 安全问题；如何保证信息真实性，由于信息的过滤功能交给了被检索主机，这就要防止实际应用时被检索主机为了获取较高的排名次序从而捏造提交给检索服务器的关键信息。

参 考 文 献

[1] 史忠植. 智能主体及其应用. 北京：科学出版社，2000 年 12 月, 第 3 章-第六章.

[2] Henzinger M. Link analysis in Web information retrieval. *IEEE Data Engineering Bulletin*, September 2000: 3-8.

[3] Brin S and Page L. The anatomy of a large-scale hypertexual web search engine. In Proc. of the WWW Conference, Brisbane, Australia, April 1998: 107-117.

[4] Heydon A and Najork M. Mercator: A scalable, extensible Web crawler. *World Wide Web*, 1999, 2(4): 219-229.

[5] Wong S K M, Ziarko W and Raghavan V V. On modeling of information retrieval concepts in vector spaces. *ACM Trans. on Database Systems*, 1987, 12(2), 299-321.

[6] Salton G and Buckley C. Term-weighting approaches in automatic text retrieval, *Inf. Process. Manage.*, 1988, 24(5): 513-523.

[7] Savoy J. Searching information in legal hypertext systems. *Artificial Intelligence & Law*, 1993, 2(3): 205-232.

徐小龙：男，1977 年生，讲师，博士生，研究方向为计算机软件、分布式计算、信息安全、移动 Agent 和移动数据库等.

王汝传：男，1943 年生，教授，博士生导师，主要研究方向为计算机软件、计算机网络和网格、信息安全、移动 Agent 和虚拟现实技术等.