第35卷 第21期 计 算 机 工 程 2009年11月
Vol.35 No.21 Computer Engineering November 2009
·人工智能及识别技术· 文章编号：1000—3428(2009)21—0192—03 文献标识码：A 中图分类号：TP274+.2

基于遗传算法的专业元搜索引擎

赵大明[1]，鱼 滨[2]
(1. 西北大学信息科学与技术学院，西安 710127；2. 西安电子科技大学计算机学院，西安 710071)

摘 要：元搜索引擎返回的查询结果来自独立搜索引擎，要评价此类结果的专业相关性，必须挖掘其位置信息以外的其他信息。研究并实现面向培训领域的元搜索引擎，在充分挖掘网页文本信息的基础上，提取专业网页样本特征，结合遗传算法给出网页专业相关度算法。实验结果表明，该引擎具有较高的专业信息筛选和排序能力。

关键词：元搜索引擎；专业搜索引擎；遗传算法；相关度模型

Professional Meta-search Engine Based on Genetic Algorithm

ZHAO Da-ming[1], YU Bin[2]
(1. College of Information Science and Technology, Northwest University, Xi'an 710127; 2. College of Computer, Xidian University, Xi'an 710071)

【Abstract】 The query result returned by meta-search engine is from the independent search engine. To evaluate the professional relevance of these results, it must exploit more information than the location information. This paper studies and realizes a training field oriented meta-search engine. It fully mines the file information of Web page, extracts swatch characteristics of professional Web page and gives professional relevance degree algorithm for Web page. Experimental results show that this engine has high ability of professional information filtration and taxis.

【Key words】 meta-search engine; professional search engine; Genetic Algorithm(GA); relevance degree model

1 概述

信息搜索的覆盖率和精度是用户使用搜索引擎时最关心的2个指标。由于互联网上的信息量不断增加，使得单个搜索引擎不可能覆盖全部信息。据统计，在现有搜集引擎中，覆盖率最高的AltaVista只能覆盖约40%的相关信息。由于机制、算法不同，同一个检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，集成多个搜索引擎的元搜索引擎成为人们研究的重点。元搜索引擎通过调用多个搜索引擎有效提高了信息覆盖率。如何从多个搜索引擎返回的信息中挑选出精度高的信息，成为元搜索引擎需要解决的关键问题之一。元搜索引擎自身没有数据库，它在多个独立搜索引擎的基础上处理数据，因此，其响应速度低于独立搜索引擎。普通的元搜索引擎无法很好地满足用户要求，对于某些专业领域的用户，使用普通元搜索引擎检索信息时，得到的多数网页是没有用的。因此，有必要研究并实现面向专业领域的专业型搜索引擎。本文出于对教育培训领域用户的信息检索需求的考虑，在研究独立搜索引擎、元搜索引擎和专业搜索引擎各自特点的基础上，实现一个面向培训领域的元搜索引擎系统。

2 相关概念

2.1 元搜索引擎和专业搜索引擎

元搜索引擎通过向多个成员引擎发送请求，调用成员搜索引擎返回的搜索结果，无须自己建立和维护庞大的索引数据库。当一个查询到来时，元搜索引擎自身并不处理，而是按各个成员引擎的查询格式做相应转换后分发到各个成员引擎，有关成员引擎的参数信息可以帮助元搜索引擎进行引擎的选择和协调，各个成员引擎返回结果后，元搜索引擎进行结果合并形成全局按权重排序的序列输出给用户。

专业型搜索引擎是一种以面向某一专业或科学领域的信息服务为目的的搜索引擎，它能满足用户对某一专业信息的需求，通过在一定范围内搜索互联网信息，智能地识别专业信息，为用户提供比通用搜索引擎更多、更精准的专业信息。

2.2 遗传算法

遗传算法(Genetic Algorithms, GA)是一种模仿生物进化过程的随机方法，由J.Hoolland于20世纪60年代提出。它体现了适者生存、优胜劣汰的进化原则，对可能包含解的群体反复使用遗传学的基本操作，不断生成新的群体，使群体不断进化，并以全局并行搜索技术搜索优化群体中的最优个体，以求得满足要求的解。GA具有简单、鲁棒性好、自组织性、自适应性、自学习性和本质并行的突出优点。需要解决的问题越复杂、目标越不明确，GA的优越性越大。

3 元搜索引擎系统架构

培训专业搜索引擎系统框架如图1所示，各模块功能描述如下：

(1)用户接口。为查询用户提供简明易懂的可视化友好查询界面。

(2)管理员接口。管理员进行系统维护的接口，主要负责对培训专业词典的管理与维护。

(3)查询分发器。根据培训专业词典以及管理员对更新频率的配置和各独立搜索引擎查询特点，向各个独立搜索引擎发起查询，并将查询结果保存至URL管理器。

基金项目：国家自然科学基金资助项目(60871097)
作者简介：赵大明(1982－)，男，硕士研究生，主研方向：软件工程；鱼 滨，教授、博士
收稿日期：2009-05-19 **E-mail：**zdm_5240@163.com

(4)检索器。接受用户查询请求，向索引库发起相关检索请求。若检索器找到索引项，则将查询结果交给结果排序模块，并把排序结果返回给用户。若索引库中没有用户查询关键词，则将此次查询关键词添加到培训专业词典，等候管理员对此关键词进行确认。将关键词切分后重新查询索引库，返回相关近似查询结果。如果近似信息为空，就向查询分发器发起检索，查询分发器进行实时查询。此时返回的结果不做任何保存，按记录位置进行简单处理后将信息显示给用户。

(5)结果排序。根据网页的专业相关度，将查询结果排序后返回给用户。

(6)专业字典。按专业信息分类保存的教育培训方面的词汇向各个独立搜索引擎发起检索请求，是培训相关信息的源泉。词典中不包含检索关键词，而是对其进行自主学习。为了保证信息的正确和真实表达专业信息，采用人工方式录入。当用户查询某一未登录词汇相关信息后，未登录词被保存，培训专业词典的维护人员按词汇检索频度，将重要的未登录词按词典领域分类添加到培训专业词典中。

(7)知识库。将系统过滤的网页以一定格式存储在知识库中，方便建立索引和以后系统功能的扩展。

(8)索引库。知识库通常很大，不便查询，因此，需要对知识库建立索引。其功能是理解知识库中的信息，从中抽取出索引项，采用倒排文件的方式生成索引表，方便检索器进行信息检索。

(9)URL 管理器。将原搜索引擎按某一关键词查询到的所有信息中的网页 URL 分析出来，保存到 URL 库中。

(10)结果筛选器。使用本文建立的相关度模型，对 URL 管理器中的网页进行计算，得到其网页相关度值。根据相关度对比并删除重复网页后，保存专业相关度较高的网页到知识库中。

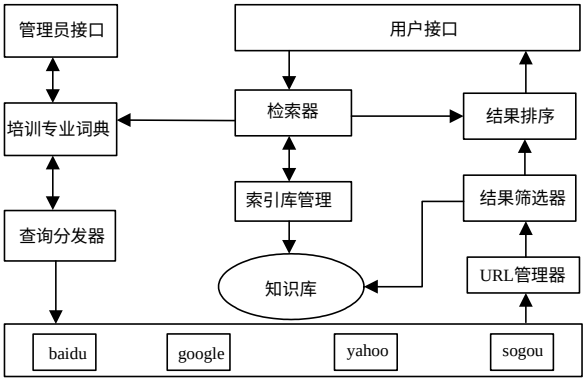

图 1　培训专业搜索引擎系统框架

4　基于遗传算法的相关度模型

4.1　种群编码与初始化

种群编码与初始化过程如下：先从各独立搜索引擎返回的结果中选择若干记录作为专业词汇表训练样本。然后对样本中的每条记录进行中文分词，获得关键词汇。再通过一定过滤算法获得专业词汇表。本文使用遗传算法的目的是从专业词汇表中选出对网页排序影响较大的关键词序列，因此，专业词汇表的长度即遗传算法中个体编码的长度。

遗传算法种群初始化时随机产生个体。种群大小取为个体编码长度的一个线性倍数是实际应用中的常用方法之一。例如，群体规模 m 取为 n 和 $2n$ 之间的一个确定数[1](n 为个体长度)。本文初始群体的选取采用随机选取，即利用随机函数从 n 到 $2n$ 之间产生 m 个染色体组成初始群体。

4.2　适应度函数

适应度函数是评价个体好坏的标准，要得到好的模型必须选择好的适应度函数。在一些优秀的适应度设计方法的基础上[1-2]，本文采用多元线性回归方程作为适应度计算方法：

(1)确定适应度模型的样本网页。选取 2 个独立搜索引擎均出现的记录作为适应度模型训练样本(可以人工挑选，但为了搜索引擎的通用性，本文采用 2 个独立搜索引擎均出现的记录作为训练样本)，以字典中的关键词“培训”为例，具体方法如下：1)以“培训”为关键词向各独立搜索引擎发送查询请求；2)独立搜索引擎返回结果，去除推广链接、无效链接和重复链接；3)随机取上述记录中的 80%作为建模网页集，其余 20%作为测试网页集；4)计算各个独立搜索引擎返回结果中重复 2 次的记录及其在独立搜索引擎中的位置和。

(2)使用获取的专业词汇表，分别对上述训练样本网页进行分析，提取专业关键词汇表中词语在网页文本中出现的频率，如图 2 所示。

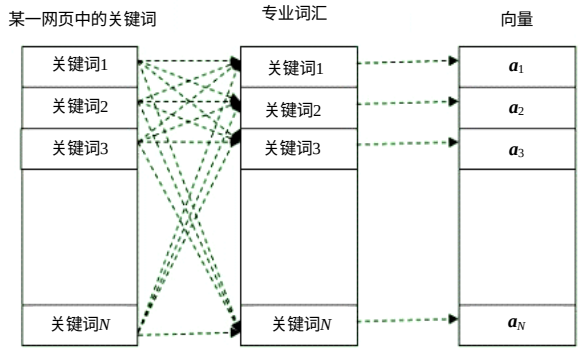

图 2　词频向量获取方法

在图 2 中，只要指向关键词汇表第 i 个关键词的字符串有一个和该关键词匹配，则 $a_i = a_i + 1$，a_i 初始化为 0。

训练样本中的每个网页都存在一个值，该值是每个网页在独立搜索引擎中的位置之和，它间接表明了独立搜索引擎对此条记录相关度的一个评价。本文对位置和进行如下转换，得到的 RankValue 值将被视为此网页专业相关度的评价值：

$$y=\text{RankValue}(\text{totalNumber}, \text{f\_position}_i, \text{s\_position}_i)$$

$$\text{RankValue}(\text{totalNumber}, \text{f\_position}_i, \text{s\_position}_i)=(\text{totalNumber}-(\text{f\_position}_i + \text{s\_position}_i))/\text{totalNumber}*1000$$

其中，totalNumber 是指 2 个独立搜索引擎总共返回的记录数；f_position_i 表示此网页在第 1 个独立搜索引擎中的位置；s_position_i 表示此网页在第 2 个独立搜索引擎中的位置。RankValue 表示此条记录位置与它在独立搜索引擎总记录中的位置比例，其值越高，独立搜索引擎对该条记录打分越高，该条记录的专业相关度越高。

适应度算法的具体流程如下：1)取种群中的一条基因串；2)将基因位为 1 的位所对应训练集矩阵 $\boldsymbol{x}$ 中的列全部取出，组成新的矩阵 $\boldsymbol{x}^{new}$；3)随机选取 $\boldsymbol{x}^{new}$ 中的 80%行数据，组成新的矩阵 $\boldsymbol{x}1$，并与 $\boldsymbol{x}1$ 中每行对应的评价值矩阵 $\boldsymbol{y}$ 中相应的行组成的新矩阵 $\boldsymbol{y}1$，其余 20%组成矩阵 $\boldsymbol{x}2$ 和 $\boldsymbol{y}2$；4)使用最小二乘估计求多元线性回归方程的回归系数，求取矩阵 $\boldsymbol{x}1$ 与 $\boldsymbol{y}1$ 组成的多元一次方程，其系数矩阵记为 $\boldsymbol{z}$；5)得到多元一次方程后，将 $\boldsymbol{x}2$ 代入方程，得到 $\boldsymbol{y}2'$；6)计算 $\boldsymbol{y}2'$与 $\boldsymbol{y}2$ 之间的平均误差值，该误差值就是此条基因的适应度值，其值越小适应度越高。

通过上述适应度算法可以计算出种群中所有个体的适应度值，适应度值是选择下一代个体的关键。

4.3 遗传算子

4.3.1 选择算子

简单遗传算法一般采用轮盘赌选择方法，将个体的适应度值与种群的总适用度相比，得到该个体的相对适应度，并使用个体的相对适应度作为其选择操作中被选中的概率。但该方法存在如下问题：在进化初期，由于适应度高的个体被选择的概率可能很大，因此会复制很多后代，而单一个体无法继续进化，将导致搜索陷入局部最优。

为了解决上述问题，本文采用改进的遗传算法实现选择操作，即在进行选择操作前先保存部分最优个体，然后采用轮盘赌选择算法选择个体。通过改进的遗传算法，可以确保每次将父代种群中适应度较高的个体直接注入下一代种群中，从而加快搜索最优个体的速度，使遗传算法收敛至最优解。

4.3.2 交叉算子

交叉算子的好坏直接影响遗传算法的收敛速度。在传统遗传算法中，配对选择一般是随机进行的。算法缺少对个体间相似度的判断，采用固定交叉率使 2 个个体进行交叉操作，造成父个体中的优良模式不能被遗传到下一代，导致算法收敛速度下降。作为改进，本文先用海明距离侧度[3]判断配对个体的相似度，并由此决定配对与否，然后采用单点交叉方法对配对个体进行交叉操作，交叉点位置采用随机确定方法。

4.3.3 变异算子

变异的目的是使算法避免因局部最优而过早收敛以至找不到满意的解。在变异运算中，每个基因位按变异概率 P_m 进行变异，即该基因取另外一个合理的随机值。本文采用二进制编码，变异即按位取反。变异概率控制了新基因导入种群的比例。变异概率虽然只影响算法的局部搜索能力，但如果变异概率太低，一些有用的基因就不能进人选择。如果变异概率太高，即随机变化太多，那么后代可能失去从双亲中继承的好特性，导致算法失去从过去的搜索中学习的能力。根据经验，变异概率一般取值在 0.000 1~0.1 之间[4]。

4.4 终止规则

终止规则一般分为 2 种：

(1)先对特征群体进行遗传操作得到下一代特征群体，再计算每个个体的适应度并返回到遗传操作，直至运算到指定的最大代数；

(2)当相邻代的平均适应度差值很小时，终止遗传操作，即找到了最优特征组合。本文采用两者相结合的方法，即设定一个最大的遗传代数 maxgen，当相邻数代的平均适应度差值很小时，终止遗传操作，否则算法迭代代数在达到 maxgen 时终止。

5 实验

5.1 模型检测与结果分析

使用遗传算法建模时，样本数据中的 80%作为建模数据，20%作为测试数据，可以得到测试数据中真实值与模型预测值之间的误差，如图 3 所示。

由图 3 可知，误差被控制在很小范围内，证明了建模的可靠性。

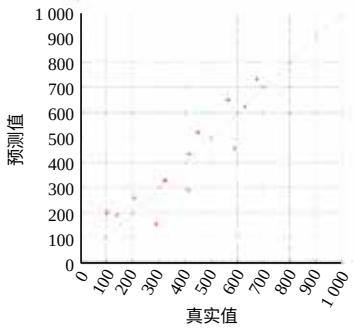

图 3 真实值与模型预测值之间的误差

5.2 实验结果分析

将经过特征选择后获得的专业词汇表作为遗传算法特征选择的基础。把词汇表中的关键词序列视为染色体，对其进行二进制编码，编码长度为词汇表长度。初始群体的选取采用随机选取方法，利用随机函数产生初始群体。设交叉概率为 0.6，变异概率为 0.02，maxgen 为 300。在 Pentium 2.4 CPU，1 GB 内存，Windows XP 环境下，分别使用独立搜索引擎和本系统对关键词“Java”进行搜索，比较返回的前 100 条记录，结果如表 1 所示。

表 1 结果比较

比较项	google	雅虎	百度	搜狗	本系统
时间/s	0.070	0.041	0.001	0.029	0.080
相关率/(%)	5	8	9	7	89

由表 1 可知，在经过本系统过滤和排序后的相关记录中，前 100 条有 89 条是培训、教程、课程、教育等信息，与独立搜索引擎的返回结果相比，本文专业搜索引擎能将更多专业信息优先显示给专业用户。

6 结束语

本文以元搜索引擎作为专业搜索引擎的架构，实现面向培训领域的元搜索引擎。给出采用遗传算法计算专业领域网页相关度的模型。使用该模型对网页进行过滤，可以得到专业相关度高的网页。

参考文献

[1] Schaffer J D. A Study of Control Parameters Affecting On-line Performance of Genetic Algorithms for Function Optimization[C]// Proceedings of the 3rd International Conference on Genetic Algorithms. San Mateo, CA, USA: Morgan Kaufinann Publishers, 1989: 51-60.

[2] Pei M, Goodman E D, Punch W F. Pattern Discovery from Data Using Genetic Algorithms[C]//Proc. of the lst Pacific-Asia Conf. on Knowledge Discovery & Data Mining. Singapore: [s. n.], 1997.

[3] 陈国良, 王煦法. 遗传算法及其应用[M]. 北京: 人民邮电出版社, 1996.

[4] 席裕庚, 柴天佑, 挥为民. 遗传算法综述[J]. 控制理论与应用, 1996, 13(6): 697-708.

编辑 陈 晖