第36卷 第19期 计 算 机 工 程 2010年10月
Vol.36 No.19 Computer Engineering October 2010
·人工智能及识别技术· 文章编号：1000—3428(2010)19—0203—02 文献标识码：A 中图分类号：TP311

基于GA的专业元搜索引擎结果排序算法

孔德镛，张建军
(西北大学信息科学与技术学院，西安 710127)

摘 要：针对元搜索引擎的查准率等问题，分析专业元搜索引擎返回结果的专业相关度问题，提出一种基于遗传算法的网页排序算法，讨论算法的原理及实现过程。在独立搜索引擎返回结果中，挖掘除位置信息以外的其他信息，利用遗传算法建立专业网页相关度模型，对网页的专业相关度进行计算。实验结果表明该算法效果较好。

关键词：搜索引擎；专业元搜索引擎；遗传算法

Professional Meta Search Engine Results Ranking Algorithm Based on Genetic Algorithm

KONG De-yong, ZHANG Jian-jun
(Institute of Information Science and Technology, Northwest University, Xi'an 710127, China)

【Abstract】 Aiming at the problems of meta search engine such as precision, this paper analyzes professional search engine returns results professional correlation in-depth exploration of a given Web page based on genetic algorithm sorting algorithm. In mining to return to the results of independent search engine, in addition to location information of the other information based on the use of genetic algorithm to build a professional Web site correlation model for Web pages related to the degree of professional computing. The algorithm discusses the implementation of the principle and provides programs to improve the algorithm. Experimental results show that the algorithm is effective.

【Key words】 search engine; professional meta search engine; genetic algorithm

1 概述

随着Web信息的日益增长，人们不得不花费大量的时间搜索浏览自己需要的信息。搜索引擎随之成为人们最常用的信息检索工具。但是，目前大多数搜索引擎提供的服务还不能令用户满意，尤其是个性化的查询请求。本文针对目前搜索引擎存在的不足，研究专业元搜索引擎的发展现状，讨论其结果排序的算法，提出一种优化策略。

2 专业元搜索引擎

专业搜索引擎是一种以某一专题或科学领域的信息服务为目的的搜索引擎[1]，通过一定范围内互联网信息的搜索，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。基于元搜索的专业搜索引擎虽然可以提高检索的查全 率[2]，但是如何在独立搜索引擎的返回结果中找到属于某一专题的网页关系到搜索引擎的查准率问题。本文采用改进的遗传算法(Genetic Algorithm, GA)计算专业领域网页的专业相关度模型，使用模型对网页进行过滤，得到专业相关度高的网页，然后根据网页的相关度对结果进行排序。

3 基于遗传算法的专业元搜索引擎结果排序算法

3.1 基于遗传算法的专业相关度模型建立

3.1.1 编码

使用遗传算法解决问题的第1步是对问题进行编码[3]。首先从独立搜索引擎的返回结果中选择若干条记录作为专业词汇表训练样本。然后对样本记录进行中文分词，获得关键词汇。下面以“培训”领域为例，描述如何抽取专业相关词汇建立专业词汇表：

(1)配置baidu、google、yahoo、sogou 4个独立搜索引擎。

(2)以“培训”为关健词向独立搜索引擎发起查询请求，要求每个独立搜索引擎返回前20条记录。

(3)对独立搜索引擎返回的前 20 条记录进行过滤，将推广连接和无效连接剔除后取有效连接10条。

(4)分别对 10 条连接进行文本提取、分词、噪音词汇过滤等操作，将其余词汇保存在一个数组中，记为wordsList[]。

(5)以“培训”为关健词再次向独立搜索引擎发起查询请求，要求每个独立搜索引擎返回关于培训的前50页~100页记录。

(6)将每个独立搜索引擎返回的记录去掉重复连接、推广连接和无效连接后，对记录中重复的记录进行统计，按照重复的次数统计重复2次、重复3次和重复4次的连接，并对应每条记录登记它们的位置和，将重复的记录保存到4个list列表中，分别记为：list1, list2, list3, list4。

(7)依次取list n(n初始化为1)中的记录，按照wordsList[]中词汇的顺序分别统计list n中网页中专业词汇的出现频率。每一个网页都得到一个词频向量，记作 ***vector***[i]，一个list n对应一个词频矩阵，记作 ***matrix*** n，如图1所示。

(8)统计 ***matrix*** n 中每列为0的个数，然后按照80%的比例淘汰0较多的列。同时剔除wordsList[]中这些列对应的关键词。

(9)$n=n+1$，如果 $n<5$，转到步骤(7)，否则，转到(10)。

(10)将上述过程得到的一些关键词记录到wordsResult数组中。

作者简介：孔德镛(1982－)，男，硕士研究生，主研方向：语义网，人工智能；张建军，副教授

收稿日期：2010-03-10 **E-mail：**kondeyong2009@yahoo.cn

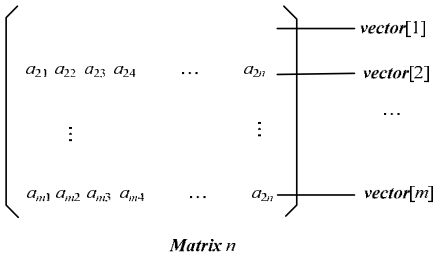


图 1　词频矩阵 matrix 的构成

使用上述方法，得到培训专业的关键词汇表部分如下：

| 认证 | 绩效 | 英语 | 测试 | 应用 |管理 | 培训班 | 职称 | 训 | 证 | 网页 | 投资 | 工作 | 新 | 设计 | 行业 | 初级 | 大学 | 教 | 销售 | 推荐 | 报 | 管理 | 等级 | 动画 | 辅导 | 电脑 | 学历 | 系统 | 全国 | 学员 | 会员 | 留学 | 联系 | 案例 | 生产 | 分析 | 人才 | 招聘 | 项目 | 查询 | 网络 | 员工 | 现代 | 创业 | 考 | 成功 | 职 | 计算机 | 技术 | 实务 | 证书 | 维修 |…

遗传算法种群初始化时随机产生个体。个体的长度与上述关键词汇表的长度相同。关于种群大小，专家建议群体的个数取 20~100 之间。本文初始化种群的大小取值 100。

3.1.2　适应度函数

本文将多元线性回归方程作为适应度计算的方法。建立适应度模型必须先确定用来训练适应度模型的样本网页。专业样本的选取来自各个独立搜索引擎返回的结果中具有代表性的网页作为适应度模型训练的样本。这里选取 2 个独立搜索引擎均出现的记录作为适应度模型训练的样本，如图 2 所示。

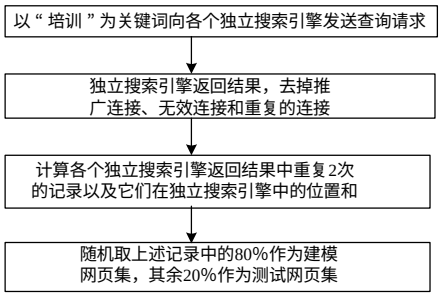


图 2　样本获取方法

其次是对样本进行分析。用上一节中获取的专业关键词汇表，分别对上述训练样本网页进行分析，具体处理方法为：对应每一个网页中的文本信息，产生一个相应的词频向量。将此关键词和关键词汇表中的词汇相比较。如果当前的关键词出现在关键词汇表中，向量对应的位置加 1，如图 3 所示。

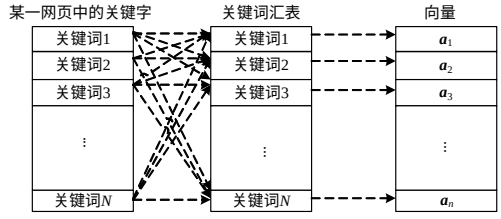


图 3　词频向量获取方法

在图 3 中，只要指向关键词汇表第 i 个关健词的字符串有一个和该关键词匹配，那么 $\boldsymbol{a}_i=\boldsymbol{a}_i+1$，$\boldsymbol{a}_i$ 初始化为 0。

训练样本中的每一个网页都存在着一个值，这个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，本文将位置和进行下述转换，得到的 *RankValue* 值将被看作此网页专业相关度的评价值。

$$y=RankValue(totalNumber,f_position_i,s_position_i) \tag{1}$$

$$RankValue(totalNumber,f_position_i,s_position_i)=(totalNumber-(f_position_i+s_position_i))/totalNumber\times 1\,000 \tag{2}$$

其中，*totalNumber* 是指 2 个独立搜索引擎总共返回的记录数；*f_positioni* 是指此网页在第 1 个独立搜索引擎中的位置；*s_positioni* 是指此网页在第 2 个独立搜索引擎中的位置；*RankValue* 是指此条记录位置和在独立搜索引擎中总共记录中的位置比例，值越大，这条记录的专业相关度越高，反之，说明该条记录专业相关度很低。

综上分析可以得到训练样本特征的 2 个矩阵，分别是词频矩阵 $\boldsymbol{x}$ 和评价值矩阵 $\boldsymbol{y}$，如式(3)、式(4)所示。

$$\boldsymbol{x}=\begin{pmatrix} a_{11},a_{12},\cdots,a_{1m} \\ a_{21},a_{22},\cdots,a_{2m} \\ \vdots\quad\vdots\qquad\vdots \\ a_{n1},a_{nj},\cdots,a_{nm} \end{pmatrix} \tag{3}$$

$$\boldsymbol{y}=\begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_n \end{pmatrix} \tag{4}$$

其中，a_{ij} 表示训练样本中关键词在关键词汇表中出现与否以及关键词出现的频数；n 为训练样本的个数；m 为关健词词汇表中关键词的个数。如果 a_{ij} 位置对应的关健词没有出现，则 a_{ij} 值为 0；如果出现过，则 a_{ij} 的值为关键词出现的频数。

3.2　网页相关度判断

使用上述方法从关键词汇表中分析出对排序影响较大的一些关键词，并得到这些关键词与其评价值的如下线性关系：

$$\begin{aligned} y=&-0.196\,18x_7+0.134x_{13}-0.283x_{18}-0.324\,88x_{23}+0.529\,06x_{27}- \\ &0.112\,58x_{35}-0.111\,58x_{36}-0.595\,35x_{38}+0.348\,26x_{48}+0.109\,45x_{56}+ \\ &0.640\,41x_{58}+0.014\,546x_{60}+0.313\,75x_{65}-0.0469\,18x_{67}+ \\ &0.461\,34x_{76}-0.443\,97x_{77}-0.105\,22x_{83}-0.383\,2x_{87}- \\ &0.010\,079x_{89}+0.151\,97x_{108}-0.268\,06x_{112}+0.153\,28x_{119}+ \\ &0.443\,86x_{122}-0.348\,41x_{124}-0.300\,56x_{127}+0.373\,72x_{131} \end{aligned}$$

其中，x_i 为专业词汇表第 i 个词汇所对应的词频；y 为网页相关性的评价。将网页下载到本地后，提取网页特征信息后即可得到上节描述的 $\boldsymbol{x}$ 矩阵，使用上述方程可对网页的评价值进行预测，得到 $\boldsymbol{y}$，并将此预测值作为网页排序的依据。

4　改进的相关度计算模型实验分析

简单遗传算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[4]，但是这种方法存在 2 个问题：(1)进化初期，有可能适应度高的个体复制出很多后代，个体单一无法继续进化，使搜索陷入局部最优。(2)进化后期，当各个个体的适应度差距不大时，该方法已经不再具有选择能力，体现不出个体的优劣。

为了避免遗传算法陷入局部最优解以及算法丧失选择能力的问题，本文的改进措施为：(1)保存最优解。简单遗传算法对种群采用轮盘赌选择，而改进的遗传算法是每次将父代种群中适应度较高的个体直接注入到下一代种群中，删除父代中适应度交叉的部分个体，然后对父代中删除最差个体后的个体进行交叉和变异操作。这样可以加快搜索最优个体的速度，防止适应度较差个体进入下一代种群。(2)防止进化过程中的振荡现象。在父代个体进行交叉操作时，将交叉后产生的新的个体与父代个体进行比较，选择适应度大的 2 个个体进入下一代。产生的新个体不一定比父代的个体优秀，因

（下转第 207 页）