

一种新的 Web 搜索优化模型及其应用

詹 芹

(九江学院 信息科学与技术学院, 江西 九江 332005)

摘 要: 随着 Web 上数据不断增加以及人们对查询质量的要求不断提高, 传统的 Web 搜索技术已经很难满足高质量的检索任务. 针对向量空间模型以及查询结果的精确率和查全率等方面进行了深入研究, 并提出一种新的 Web 搜索优化模型和算法. 仿真实验有效提高了 Web 搜索的质量和效率.

关键词: 搜索引擎; Web 搜索; 精确率; 查全率

中图分类号: TP312

文献标识码: A

文章编号: 1000-7180(2011)08-0184-03

A New Web Search Optimization Model and Its Application

ZHAN Qin

(School of Information Science and Technology, Jiujiang University, Jiujiang 332005, China)

Abstract: With the increasing accumulation of data on the Web and the rising demand of high quality of the retrieval results, it has been very difficult for the traditional Web information retrieval technologies to satisfy the task of highly desirable results. This paper mainly discusses how to improve the precision and recall of the retrieval the information retrieval model, and propose a new Web search optimization models and algorithms. Experimental results indicate that this adaptive algorithm enhances the quality and efficiency.

Key words: search engine; web search; precision; recall

1 引言

近年来, 人们从生物进化的机理和一些物理现象中受到启发, 提出了许多用以解决复杂优化问题的新方法, 因其高效的优化性能、无需问题特殊信息等优点, 得到各领域广泛的关注和应用. 已有研究包括遗传算法(GA)、免疫克隆算法(ICA)、模拟退火(SA)、禁忌搜索(TS)、进化规则(EP)、进化策略(ES)等. 这些算法具有全局优化的性能, 同时具有区别于常规算法的搜索机制和特点, 被称为现代启发式算法. 因此, 本文将免疫克隆算法(Immune Clonal Algorithm, ICA)应用到 Web 搜索领域, 提出一种新的免疫克隆信息检索算法(Immune Clonal Information Retrieval Algorithm, ICIRA). ICIRA 算法利用人类在人工免疫系统的最新研究成果, 充分利用克隆和高斯变异等操作, 希望能对 Web 搜索引擎进行优化, 从而有效提高 Web 搜索引擎质量和

效率, 尤其是精确率和查全率. 仿真实验取得了较理想的实验结果, 从而表明该方法的可行性和有效性.

2 改进的 VSM 优化模型

我们首先对 Salton 等人提出的 VSM 模型^[1-3]进行改进, 并采用如下形式化描述:

定义 1 信息检索模型 IR

$$IR = \{D, K, IDF, TF, W\}$$

$$D = \{d_1, d_2, \dots, d_N\}$$

$$K = \{k_1, k_2, \dots, k_t\}$$

$$idf_i = \log(N/n_i)$$

$$tf_{ij} = f_{ij} / \max\{tf_{1j}, tf_{2j}, \dots, tf_{Nj}\}$$

$$W = \{w_{11}, \dots, w_{ij}, \dots, w_{tN}\}$$

$$w_{ij} = tf_{ij} \times idf_i$$

其中, D 是一组 Web 文档集, K 是一组用户查询的索引词集, IDF 表示一个索引词在整个文档集中出现的倒频率, TF 表示索引词在 Web 文档中出现

的频率^[4], W 表示 Web 文档的权重. TF 归一化后频率为 tf_{ij} , 它是 k_i 在文档 d_j 中出现的频率. N 为文档集合中文档总数, t 是索引词的个数. idf_i 是索引词 k_i 在 Web 文档 d_j 中倒频率, n_i 是 k_i 出现过的文档数.

w_{ij} 是 k_i 在 d_j 中的权重, 它表示了 k_i 对 d_j 语义内容的重要性^[5]. 如果一个索引词在一个文档中没有出现, 那么 $w_{ij} = 0$.

定义2 查询 q

$$q = (w_{1q}, w_{2q}, \dots, w_{iq}, \dots, w_{nq})$$

式中, w_{iq} 表示为索引词 k_i 在用户查询 q 中的权重, t 为查询中索引词的个数.

定义3 相关度 δ

$$\delta = \frac{\sum_{i=1}^t w_{ij} \times w_{iq}}{\sqrt{\sum_{i=1}^t (w_{iq})^2 \times \sum_{i=1}^t (w_{ij})^2}} - \varphi$$

式中, δ 表示为用户查询 q 与文档 d_j 之间的相关度, φ 为我们预先设置的阈值, 当 $\delta \geq 0$ 时, 表示查询与目标文档相关, 否则无关.

3 免疫克隆信息检索算法

我们将免疫克隆算法与 IR 模型相结合, 提出一种新的 Web 搜索方法 - 免疫克隆信息检索算法 ICIRA. ICIRA 算法描述如下:

(1) 获取初始抗体群 $A(0)$ (Web 页面集) 和用户查询 q ; 并对 $A(0)$ 进行抗体编码.

$A(0) = \{a_1(0), a_2(0), \dots, a_j(0), \dots, a_N(0)\}$ 表示由 N 个抗体 (Web 页面) 组成, $a_j(0) = w_{1q}w_{2q} \dots w_{iq} \dots w_{nq}$, w_{iq} 表示为索引词 k_i 在用户查询 q 中的权重.

(2) 初始化种群规模 N (Web 页面数目), 变异概率 P_m , 相关度阈值 φ , 最大进化代数 MaxGen , 进化代数 $k = 0$;

(3) 计算抗体群中每个抗体的亲和度 $f(a_i(k))$, $f(a_i(k)) = \sum_{i=1}^N \text{sim } i(q, d_i) / N, i = 1, 2, \dots, n$;

(4) 对抗体群 $A(k)$ 进行克隆操作, 产生一个临时克隆群体 $B(k)$, 每个抗体克隆数目与它的亲和度函数成正比; 抗体 $a_i(k)$ 的克隆数目为 $100 \times a_i(k) / \sum f(a_i(k))$.

(5) 对抗体群 $B(k)$ 进行高斯变异操作, 变异后产生临时抗体群为 $C(k)$;

$$C(k) = B(k) + P_m \times \exp(-f(*)) \times N(0, 1).$$

$N(0, 1)$ 是均值为 0, 方差 $\sigma = 1$ 的高斯变量; P_m 是变异概率; $f(*)$ 是父代抗体的亲和度值.

(6) 从抗体群 $C(k)$ 中, 选择亲和度最大的 100 个抗体组成新的抗体群 $A(k+1)$;

(7) $k = k + 1$; 删除亲和度小于阈值 φ 的抗体 (保留相关的 Web 页面), 并返回到步骤 (3). 当 $k > \text{MaxGen}$ 时, 算法搜索结束, 并将搜索结果进行排序显示给用户.

4 仿真实验与结果分析

为了测试本文提出的 ICIRA 算法性能, 我们将 ICIRA 算法与基于 GA 的 Web 搜索算法和基于 KDD 的 Web 搜索算法进行了对比实验. 算法中用到的主要参数有: 种群规模 $N = 1\,000$, 变异概率 $P_m = 0.15$, 相关度阈值 $\varphi = 0.82$, 最大进化代数 $\text{MaxGen} = 200$.

常用搜索引擎评价标准是精确率和查全率^[6-7]. 本文的精确率和查全率分别用 Precision 和 Recall 表示:

$$\text{Precision} = \frac{R'}{S}$$

$$\text{Recall} = \frac{R'}{R}$$

其中, R 为查询 q 在整个文档库中的相关文档集合, S 为用 q 对 IR 进行查询得到的文档集, S 与 R 的交集为 R' .

图 1、图 2 分别显示的是 ICIRA 检索算法, GA 检索算法和 KDD 检索算法的精确率和查全率比较. 在图 1、图 2 中, ICIRA 检索算法的振幅最小, 曲线最平滑, 表明该算法稳定性最好. 图 1 中, GA, KDD 和 ICIRA 三者的平均精确率分别为 76.8%,

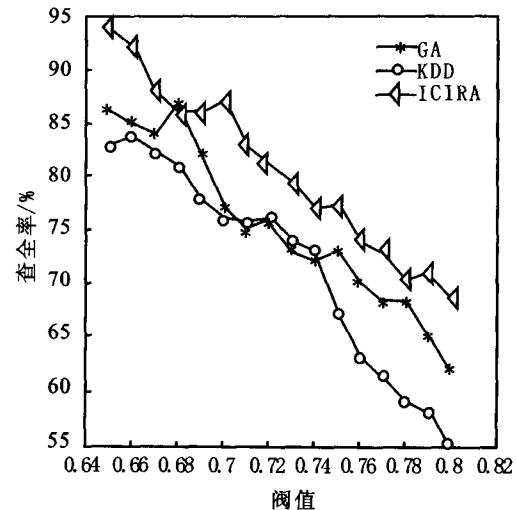


图1 三种检索算法的精确率比较

75.6%, 82.3%; 图 2 中 GA, KDD 和 ICIRA 三者的平均查全率分别为 72.6%, 72.1%, 82.8%。两次对比实验均表明 ICIRA 检索算法的性能最好, 同时能获取比较高质量的检索文档。

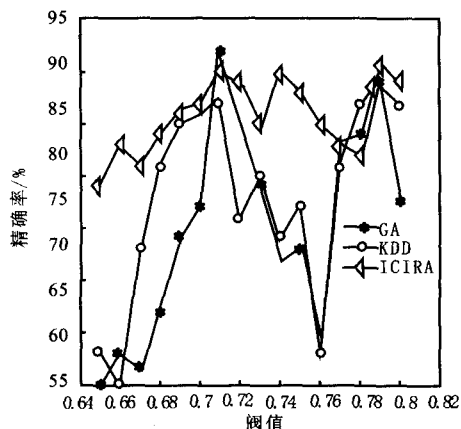


图 2 三种检索算法的查全率比较

5 结束语

随着 Internet 的飞速发展及其应用的日益普及, 网上的信息急剧膨胀, 要在这样一个浩如烟海的信息空间里查找所需的信息, 人们觉得力不从心, 往往花费了很多时间和精力却收获甚少。各种搜索引擎的出现是解决上述问题的主要途径。但是目前的搜索引擎检索出的信息量庞大, 常常混杂相当数量用户不感兴趣的, 甚至无关的信息。因此, 如何在 Internet 上获得高检索精确度和高检索率, 是当前 In-

ternet 资源发现的重要研究课题之一。本文将克隆的思想和方法引入 Web 搜索引擎, 能有效提高 Web 搜索引擎质量和效率, 具有很强的理论意义和实际价值。

参考文献:

- [1] 曾致远, 张莉. 基于向量空间模型的网页文本表示改进算法[J]. 计算机工程, 2006, 32(3): 134-136.
- [2] 凌波, 周水庚, 周傲英. P2P 信息检索系统的查询结果排序与合并策略[J]. 计算机学报, 2007, 30(3): 405-414.
- [3] 陈福集, 杨善林. 一种基于 KDD 的 Web 搜索引擎框架[J]. 情报学报, 2002, 21(3): 264-268.
- [4] 李新安, 石冰. 基于决策树方法的特定主题 Web 搜索策略[J]. 计算机应用, 2006, 23(1): 235-268.
- [5] 张卫丰, 徐宝文, 周晓宇. Web 搜索引擎综述[J]. 计算机科学, 2001, 28(9): 24-28.
- [6] 王晓峰, 刘惟一. 通过网页集团对 Web 信息搜索进行优化[J]. 云南大学学报: 自然科学版, 2003, 25(6A): 75-78.
- [7] Yang J R R. Effects of query term weights modification in document retrieval: A study base on a genetic algorithm[C]// proceedings of the second annual symposium on Document Analysis and Information Retrieval, New York, NY, USA: ACM, 1993: 271-185.

作者简介:

詹 芹 女, (1977-), 硕士, 讲师. 研究方向为数据挖掘、嵌入式系统。

(上接第 183 页)

- [2] Strong D, Lee Y, Wang R. Data quality in context[J]. Commun ACM, 1997, 40(5): 103-110.
- [3] Daniel Aebi, Louis Perrochon. Towards improving data quality[C]//Proc of the international conference on information systems and management of data. Delhi: Louis Perrochon, 1993: 273-281.
- [4] 张昌年. 一种基于 VSM 的检测相似重复记录的方法[J]. 微电子学与计算机, 2008, 25(8): 184-187.
- [5] 韩京宇, 徐立臻, 董逸生. 数据质量研究综述[J]. 计算机科学, 2008, 35(2): 1-5.
- [6] Breunig M M, Kriegel H P, Ng R, et al. LOF: Id-entifying density-based local outliers[C]// Proc SIGMOD Conf on Management of Data, NY, USA: ACM, 2000.
- [7] Hipp J, Guntzer U, Grimmer U. Data quality mining-making a virtue of necessity[C]// Proc of the 6th ACM SIGMOD Workshop on DMICD. [s. l.]: ACM, 2001: 52-57.
- [8] Dominik Luebbbers, Udo Grimmer, Matthias Jarke. Systematic development of data mining-based data quality Tools proceedings of the 29th international conference on very Large data. Berlin, Germany: ACM, 2003.
- [9] 金瑛. 广义货币供应量统计数据质量评估研究[D]. 湖南大学, 2010.
- [10] Nigrini M J, Mittermaier L J. The use of benford's law as an aid in analytical procedures[J]. Auditing: A Journal of Practice & Theory, 1997, 16(2): 52-67.

作者简介:

余宇峰 男, (1979-), 硕士, 讲师. 研究方向为数据管理与数据挖掘。

万定生 男, (1963-), 教授. 研究方向为数据管理与数据挖掘。