Computer Science and Application 计算机科学与应用, 2022, 12(8), 1895-1903
Published Online August 2022 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2022.128190

Hans汉斯

加权判定遗传算法在数据采集中的研究

陈卓聪，孙　杰

天津工业大学，天津

收稿日期：2022年7月1日；录用日期：2022年7月30日；发布日期：2022年8月5日

摘　要

随着互联网的快速发展，使得如何从海量的网络资源中快速准确地获取用户所需的信息成为一个关键问题。通用搜索引擎通过网页采集和索引为用户提供检索服务，但这种基于关键词匹配的检索方式，往往忽略用户真实查询意图的识别和匹配。垂直搜索引擎则通过缩小采集范围为特定领域和背景的用户提供专业化、定制化信息检索服务，是当前搜索领域研究的热点。主题爬虫是垂直搜索引擎的网页采集模块，在搜索路径上只保留与主题相关的网页，本文主要围绕主题爬虫的网页分析方法和搜索策略，探讨如何提高爬虫的指标性能。在以往的研究中，针对于链接结构评价和网页内容评价相结合的爬虫策略取得了较好的效果。但这种方法一般是将链接评价问题作为单目标问题处理，难以适应网页的多样性，同时全局搜索能力不强，容易陷入局部最优。经过对以上情况的分析，本文提出了一种加权判定遗传算法的主题爬虫策略，该策略在现有遗传算法爬行策略基础上新引入改进的TrustRank算法来提高反作弊能力和计算的网页的重要程度，采用多项网页内容信息来判断网页与主题的相关性，并通过选择遗传因子和设置适应度函数赋予这两项指标相应的权重来判定待下载网页的价值，在保证了利用遗传算法增强整体搜索性能的前提下，增强了爬取页面的重要性和主题相关性。相比于已有遗传算法，加权判定遗传算法的搜索策略能在一定程度上提高主题爬虫的查准和查全率，扩大爬虫的搜索范围，更符合用户的主题检索需求。

关键词

主题爬虫，数据采集，遗传算法

Research on Weighted Decision Genetic Algorithm in Data Acquisition

Zhuocong Chen, Jie Sun

Tiangong University, Tianjin

Received: Jul. 1st, 2022; accepted: Jul. 30th, 2022; published: Aug. 5th, 2022

文章引用: 陈卓聪, 孙杰. 加权判定遗传算法在数据采集中的研究[J]. 计算机科学与应用, 2022, 12(8): 1895-1903.
DOI: 10.12677/csa.2022.128190

Abstract

With the rapid development of the Internet, how to quickly and accurately obtain the information required by users from the massive network resources has become a key issue. General search engines provide users with retrieval services through web page collection and indexing, but this retrieval method based on keyword matching often ignores the identification and matching of users' real query intentions. Vertical search engines provide specialized and customized information retrieval services for users in specific fields and backgrounds by narrowing the collection range, which is a hot research topic in the current search field. The topic crawler is a web page collection module of a vertical search engine. Only the topic-related web pages are kept on the search path. This paper mainly focuses on the webpage analysis method and search strategy of the topic crawler, and discusses how to improve the index performance of the crawler. In previous studies, the crawler strategy combining link structure evaluation and web content evaluation has achieved good results. However, this method generally treats the link evaluation problem as a single-objective problem, which is difficult to adapt to the diversity of web pages. At the same time, the global search ability is not strong, and it is easy to fall into local optimum. After analyzing the above situation, this paper proposes a topic crawling strategy based on weighted decision genetic algorithm. This strategy introduces an improved TrustRank algorithm based on the existing genetic algorithm crawling strategy to improve the anti-cheating ability and the importance of the calculated webpage, using a number of webpage content information to judge the relevance of web pages and themes, and by selecting genetic factors and setting fitness functions to give these two indicators the corresponding weights to judge the value of the webpage to be downloaded, which ensures the use of genetic algorithms to enhance the overall. On the premise of search performance, the importance and topic relevance of crawling pages are enhanced. Compared with the existing genetic algorithm, the search strategy of the weighted decision genetic algorithm can improve the precision and recall rate of the subject crawler to a certain extent, expand the search scope of the crawler, and better meet the user's subject retrieval needs.

Keywords

Topic Crawler, Data Acquisition, Genetic Algorithm

1. 引言

通用搜索引擎能够提供大量的互联网信息资源给用户，同时也将用户真正感兴趣的资源淹没其中，如何从中快速、准确地找到用户所需的资源是当前要解决的主要问题。就这个问题而言，针对特定领域、特定人群和个性化需求的垂直搜索引擎提供了一种很好的解决方案。具有“专、精、深”特点的垂直搜索引擎为用户提供具有一定价值的信息和服务，受到广大用户的欢迎和认可，在搜索引擎的发展史中写下了光辉的一页。目前来看，搜索由通用到专业已是一种发展趋势。近年来，研究人员对于垂直搜索引擎不管是在理论方面，还是在技术方面都做了许多研究。

对于各行各业的各类信息，几乎都可以应用垂直搜索技术，从而实现便捷的检索，像物流、购物、餐饮、旅游等许多领域也都有相应的垂直搜索系统。由于各个行业的信息都很复杂，如果在搜集相关信

息并对其分析的过程中，不注重行业的特点，那么很难对网页的重要性做出合理的分析，这也恰恰体现出了研究垂直搜索技术的价值所在。垂直搜索引擎对于构建政府门户网站和行业门户网站都有着很高的应用价值，为企业整合内外信息资源提供了一种很好的方法。对于垂直搜索引擎的研究，它的意义不言而喻，必将翻开搜索引擎应用发展的新篇章。

网络爬虫如同人的心脏一样，是搜索引擎的重要组成部分，起着非常关键的作用。它是一个自动下载网页的程序，但必须遵循一定的规则，将万维网上的网页下载下来交给搜索引擎。与通用搜索引擎的网络爬虫不同，主题爬虫是根据目标主题，搜集主题相关页面，它是垂直搜索引擎的核心。故研究主题爬虫的相关技术，在垂直搜索引擎的发展、应用及推广方面，都有着重大意义。

2. 主题爬虫的现有关键技术以及不足

2.1. 基于内容评价的主题爬虫策略

网页内容主要包括网页标题、文本内容、锚文本以及上下文等，通过分析网页内容来判断网页或预测链接的主题相关性，并进行排序，返回搜索结果，是非常有效的爬虫策略。

Fish-Search 算法是一种典型的基于内容评价的主题爬虫策略。该算法形象的将主题爬虫在 Web 中爬取网页比喻成鱼在海洋中觅食，这里将互联网信息比作海洋，将网页资源比作食物，将网页中的 URLs (统一资源定位符)比作鱼的后代。然后用一组关键词和少量文本信息计算网页的主题相关性，最后利用深度优先搜索策略采集有效的网页信息，该算法对网页主题相关性的判断采用的是离散的二值判断(1 或 0)方法，这使得其无法对候选链接进行优先级划分。

在 Fish-Search 算法的基础上，1998 年 Michael 等人提出了 Shark-Search 算法，它是对 Fish-Search 算法的改进，主要改进有两方面，一方面是在评价文档和主题之间相关度时，使用[0, 1]之间的连续变量即相似性度量，取代 Fish-Search 算法中的 0、0.5、1 三种数值；使用这个连续变量确定待爬取队列中 URL 的优先级；另一方面，使用锚文本(锚文本是链接的一种形式，将关键词作为链接指向其他网站)与主题之间的相关度，以及从父网页继承过来的递减相关度，共同确定爬取队列中 URL 的优先级。

Shark-Search 算法比 Fish-Search 算法能返回更多查询主题相关的数据。然而 Shark-Search 算法只考虑了文本内容与主题的相关性，不仅忽略了网页结构而且只采用了相似性度量作为判定主题相关度的依据，很容易导致网页的误选。

Best-First 算法是在主题搜索引擎出现时就被使用的算法，其基本思想是：当某个网页有很多入链时，Best-First 算法会假设网页的入链个数越多，则该网页越重要，因此 Best-First 算法会早早的抓取到该网页。虽然 Best-First 算法使用机械的抓取策略，计算量较小，易于实现，但在确定主题相关度时，仅使用了锚文本的链接信息，不够准确，容易陷入局部最优。

网页文本内容作为表达网页内容最重要的载体，基于文本内容的主题爬虫根据网页文本中的正文、锚文本等文字内容为判断依据，很好地把握了网页主题信息，对网页的主题相关性判断有重要意义，但是该方法由于只关注网页内容信息，忽略了网页与网页通过链接相连接的信息，没有考虑链接结构对主题相关性的影响，因此该种方法缺乏全局性，有“近视”的缺点。

2.2. 基于链接结构的主题爬虫策略

基于内容评价的算法主要是利用 Web 页面的标题文字、网页正文、链接字符串、锚文本等文字内容信息预测待爬行 URL 的主题相关性，但是忽略了 Web 中站点、网页之间的链接关系信息[1]。页面中包含的超链接在一定程度上反映了网页之间内容或结构上的关系，而这些关系间接地反映了页面与链接的重要性，同样是一种非常有价值的信息。

PageRank 算法由 Google 的创始人 Larry Page 和 Sergey Brin 提出。该算法借鉴了学术界论文重要性的评估方法，认为一个网页如果被很多网页链接指向就说明这个网页的重要性很高，且一个网页的重要性会随着被重要性较高的网页链接到而提高。PageRank 算法用这种思想迭代计算网页的重要性。网页的 PageRank 值 $PR(a)$计算公式如下：

$$PR(a)_{i+1} = \sum_{i=1}^{n} \frac{PR(Ti)_i}{L(Ti)}$$

$PR(Ti)$代表的是其他节点(指向节点 a 的节点)的 PR 值；

$L(Ti)$代表的是其他节点(指向节点 a 的节点)的出链数；i 代表循环次数。当 $i = 0$ 时所有节点的初始值初始化为 $1/N$ (N 为所有节点数目，及网页数目)。

PR 值需要通过多次循环迭代才能达到一个稳定值。

搜索引擎在使用 PageRank 算法计算网页排名的时候，非常依赖网络页面之间的链接关系，因此，链接的质量计算排名变得越来越重要。但是，有些网页采用作弊的行为来提升自己的排名，因此，一个良好的检测作弊网页的算法变得越来越重要。Google 为了提高网站的检索质量，设计出了 TrustRank 算法来检测垃圾作弊网站。

TrustRank 算法基于了一个重要的观察的经验：好的页面很少指向坏的页面。这个概念是相当直观的，作弊页面是为了误导搜索引擎而被建立的，不能提供有效地信息。因此，人们创建可信赖的页面很少有原因指向作弊页面。

基于以上假设，挑选完全可以信赖的网站，将网站的 TrustRank 值设为最高，通过迭代运算将可信任值传播出去。也有一些可信赖的网站被欺骗链接到作弊网站，不过距离第一级网站越远信任值指数便会逐渐下降。这样通过 TrustRank 算法就可以对所有网站计算相应的信任值，信任值越高的网站可信赖信就越大。

基于链接结构的主题判定策略往往过分关注网页的权威度，而对主题相关度关注不足，容易在爬行过程中出现“主题偏移”现象，这样可能导致爬虫爬取到很多主题无关网页，对爬虫的效率有影响。

2.3. 链接结构评价和网页内容评价相结合的爬虫策略

基于文本内容的爬行策略以及链接结构的爬行策略都有自己的缺陷，因此之后的研究中往往在爬行策略中混合考虑文本内容和链接结构。

有研究者将 Shark-Search 与 PageRank 算法合并起来：采用 Shark-Search 算法计算网页得分，在用 PageRank 计算页面之间 URL 链接的权重值定义页面的重要性，同时弥补了两个传统算法的缺陷[2]。

同时有学者通过分析基于内容的链接选择 Best-First 算法，引入能够体现链接价值的 HITS 算法，提出了新的链接选择策略，并证明了比单一的 Best-First 算法具有更好的性能表现。

链接结构评价和网页内容评价相结合的爬虫策略取得了较好的效果。但这种方法一般是将链接评价问题作为单目标问题处理，难以适应网页的多样性，同时全局搜索能力不强，容易陷入局部最优。

3. 加权判定遗传算法

3.1. 遗传算法

遗传算法(Genetic Algorithm)是由 Michigan 大学的 J. Holland 教授于 1975 年提出的一种模拟进化随机搜索算法，受达尔文生物进化理论和孟德尔遗传学说的启发，模拟生物进化过程中出现的自然选择和遗传现象，理论上能够在有限时间内搜寻到目标解，一般应用于多目标搜索和组合优化。遗传算法借鉴生

物进化过程中发生的种群繁殖、多个体基因交叉和单个体基因变异现象，通过适应度函数从候选解中选取较优个体，同时利用选择、交叉和变异操作不断产生新的候选解，直至找到目标解[3]。

适应度函数作为评价个体优劣的指标，应结合求解问题进行设计。

基本遗传算法的基本步骤是：

1) 随机产生种群；

2) 用轮盘赌策略确定个体的适应度，判断是否符合优化准则，若符合，输出最佳个体及其最优解，结束，否则，进行下一步；

3) 依据适应度选择再生个体，适应度高的个体被选中的概率高，适应度低的个体被淘汰；

4) 按照一定的交叉概率和交叉方法，生成新的个体；

5) 按照一定的变异概率和变异方法，生成新的个体；

6) 由交叉和变异产生新一代种群，返回步骤 2。

选择操作按照适者生存的规则筛选出种群中适应能力较强的个体，即个体适应度越高，其生存能力越强，被保留到下代种群的概率越高。

交叉操作则对两个匹配的个体依据交叉概率，相互交换其部分基因，形成两个新个体，使大量优质基因得以保留，是遗传算法中产生新个体的主要操作。

变异操作则依据变异概率对个体编码串上的某些基因用其它基因替代，形成一个新的个体，提高了群体多样性。

3.2. 加权判定遗传算法

3.2.1. 关键思路

本文在已有遗传算法的基础上，对网页评价方法和搜索策略进行改进，提出了一种加权判定遗传算法，重新设计了遗传算法中的适应度函数和遗传操作，采用向量空间模型(VSM)分析网页主题相关度，改进 TrustRank 计算网页主题重要性，适应度函数综合考虑网页内文本内容和网页间链接结构，选择操作筛选出适应度较高的网页个体，交叉操作则针对网页中包含的超链接，按照链接主题重要性进行排序，变异操作则通过搜索引擎查询拼接的主题关键词，选出靠前的结果，产生的新链接用于扩充待爬取 URL 队列[4]。最后，主题爬虫分别以基于最佳优先、已有遗传算法和改进遗传算法的搜索策略进行爬取，实验结果表明，基于改进遗传算法的搜索策略能提高主题爬虫的指标性能，扩大爬虫的搜索范围，符合用户的主题检索需求。

如图 1 是设计的主要思路：

改进遗传算法策略的基本思想是：综合遗传算法、基于网页链接结构和内容评价的爬行策略的优点，以遗传算法的全局寻优特性来保证爬行的全局性，以网页的相互链接关系来确定网页重要性，以网页文本信息来判断网页与主题的相关性[5]。把重要性和相关性信息作为遗传基因、设置适应度函数和对遗传基因进行排序的依据，使优势基因优先被选择，通过遗传、交叉操作产生新的个体，从而减少遗传基因在传递过程中出现主题漂移的情况。策略的步骤和流程如下图 2 所示。

3.2.2. 向量空间模型(VSM)

向量空间模型在自然语言处理中是最常用的文档相似度计算模型，它假设文章中的词组对文档所表征含义的作用是相互独立的，即词组出现的顺序对含义无影响[6]，因此，文档可表示为对文档分词后提取出的特征关键词及关键词对应权重组成的向量，不同的权重表示特征关键词对文档的影响程度不同。首先，给定主题可由若干主题关键词组成的集合表征，将下载网页的文本内容分词后，统计其包含的关键词的绝对词频[7]，然后将主题和网页文本内容映射到 n 维向量空间，n 是主题关键词集所包含的关键

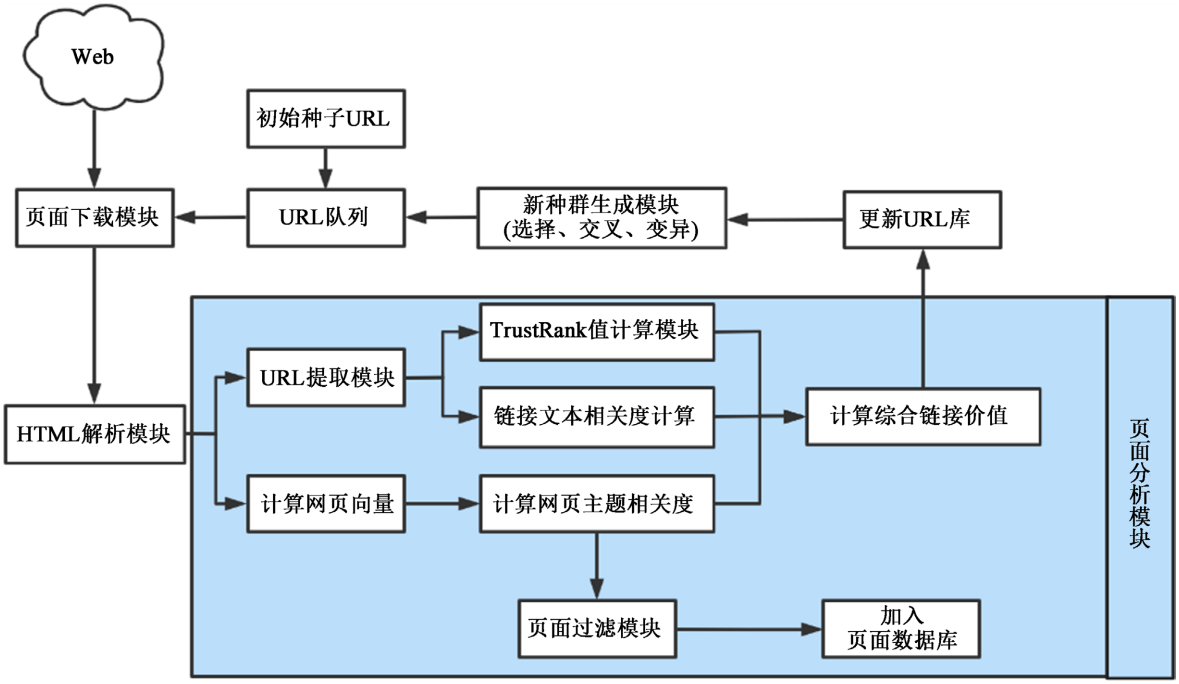

Figure 1. Experimental design diagram
图 1. 实验设计图

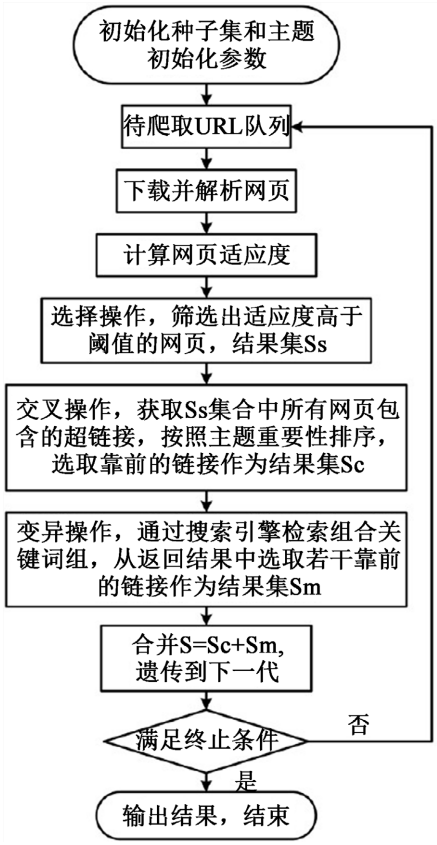

Figure 2. Improved genetic algorithm flow chart
图 2. 改进遗传算法流程图

词数量。

主题可用向量表示为：

$$\text{topic}=(\alpha_1,\alpha_2,\alpha_3,\cdots,\alpha_n)$$

其中 α_n 为第 n 维关键词的权重。

网页可用向量表示为：

$$p=(\eta_1\alpha_1,\eta_2\alpha_2,\eta_3\alpha_3,\cdots,\eta_n\alpha_n)$$

其中 η_n 为第 n 维关键词在网页中出现的绝对词频。

VSM 用二个向量夹角的余弦值表示网页主题相关度：

$$\text{Sim}(p)=\frac{\eta_1\alpha_1^2+\eta_2\alpha_2^2+\eta_3\alpha_3^2+\cdots\eta_n\alpha_n^2}{\sqrt{\alpha_1^2+\alpha_2^2+\cdots+\alpha_n^2}\sqrt{\eta_1^2\alpha_1^2+\eta_2^2\alpha_2^2+\cdots+\eta_n^2\alpha_n^2}}$$

其中，$\text{Sim}(p)$ 表示页面 p 的主题相关度。当 $\text{Sim}(p)\geq S_0$ 时，表示该页面与主题相关，S_0 为自定义的临界值[8]。

3.2.3. 改进 TrustRank 算法

TrustRank 算法是 TrustRank 应该是截至目前最出名的一种 PageRank 变体算法。相较于 PageRank，TrustRank 的最大不同在于它使用了部分有标签(labelled)的数据，因此它可以看作是“有监督的”，而 PageRank 则是“无监督的”[9]。

传统的 PageRank 算法的基础是随机游走模型，但是它没有考虑用户访问具有主题性，在实际浏览网页过程中，用户会根据出链的锚文本提示信息来决定是否点击，通常，出链锚文本与查询主题相关性越高，用户点击该出链的概率越大。

因此，需要对已有的 TrustRank 公式进行调整，采用主题访问模型，网页根据出链的锚文本主题相关度来分配 TR 值，改进后的 TrustRank 公式为：

$$TR(A)=(1-\beta)/N+\beta\sum_{i\in B_A}TR_A(P_i)$$

$$TR_A(P_i)=\frac{\text{Sim}(A)}{\sum_{j\in F_i}\text{Sim}(j)}TR(P_i)$$

3.2.4. 适应度函数

在加权判定遗传算法中，网页分析方法即作为遗传算法的适应度指标，综合考虑网页的内容价值和链接价值，作为衡量网页的综合价值并指引主题爬虫的搜索方向。

内容价值采用 VSM 计算，将主题和网页映射到向量空间，两个向量之间的夹角的余弦值用来表征网页与主题之间的相关性[10]；链接价值利用改进 TrustRank 算法进行衡量，网页在互联网上的重要性越高，网页可信度也随之增加，同时加入了防作弊因素，通过网页间的链接结构来衡量网页的重要程度，即 Tr 值，改进之处在于网页根据链接锚文本主题相关度传递，Tr 值给其包含的子链接。

因此，适应度函数综合考虑网页内文本内容和网页间链接结构，可表示为：

$$\text{Fitness}(A)=\alpha\text{Sim}(A)+(1-\alpha)TR(A)$$

4. 实验结果分析

4.1. 评价指标

查准率：被正确检索的样本数与被检索到样本总数之比。即：

$$TP/(TP+FP).$$

查全率：被正确检索的样本数与应当被检索到的样本数之比。即：

$$TP/(TP+FN).$$

4.2. 实验结果比较

本次实验以“财经”、“经济”、“金融”为主题，腾讯新闻网和搜狐新闻网作为初始种子 URL，进行采集数据，以查准率和查全率为评价指标，分别对 Best-First 算法、PageRank 算法和加权判定遗传算法进行统计结果进行比较，如图 3、图 4 所示，其中蓝色的线条为加权判定遗传算法，黑色的为 Best-First 算法，红色的为 PageRank 算法，横轴为采集深度，纵轴为查准率和查全率：

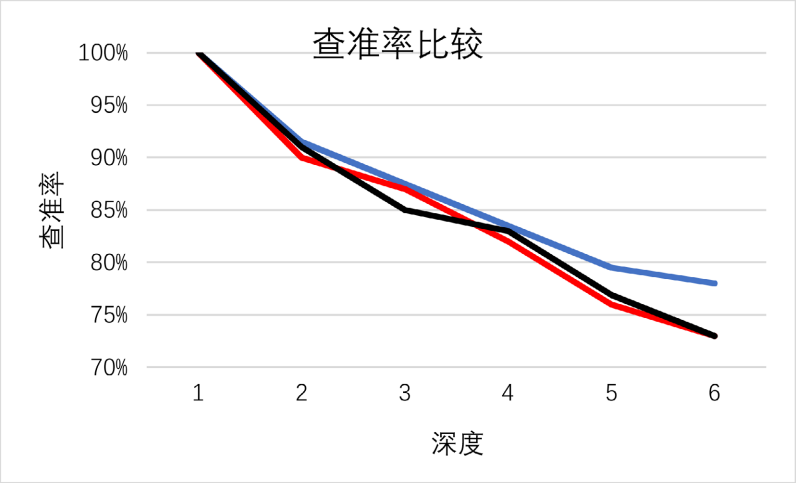

Figure 3. Precision comparison chart
图 3. 查准率比较图

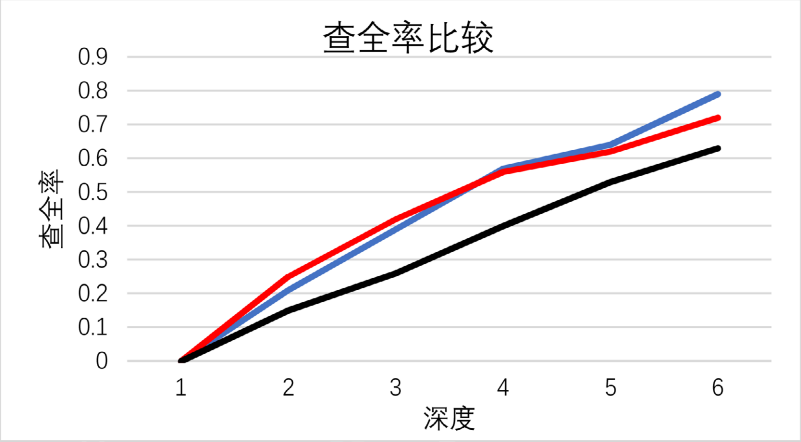

Figure 4. Recall comparison chart
图 4. 查全率比较图

根据图中所示，加权判定遗传算法在采集深度小的时候对比其他两种算法的优势并不明显，但随着爬取深度的增加，加权判定遗传算法的查全率和查准率明显要优于 Best-First 算法和 PageRank 算法。

同时为了测试 TrustRank 算法的防作弊效果，在第二次和第四次爬取深度的时候人为添加了一些相

关广告页面，实验结果表明加权判定遗传算法能够有效地识别出作弊页面，并在遗传算法的筛选过程中降低作弊页面的优先级。

5. 结论

根据上述分析，加权判定遗传算法能够提高主题爬虫采集的效率，扩大搜索范围，提升容错率，防止作弊页面的侵入，综合权衡页面的相关度和重要程度，更符合用户的检索需求。

参考文献

[1] 左薇, 张熹, 董红娟, 于梦君. 主题网络爬虫研究综述[J]. 软件导刊, 2020, 19(2): 278-281.

[2] 安子建. 基于 Scrapy 框架的网络爬虫实现与数据抓取分析[D]: [硕士学位论文]. 长春: 吉林大学, 2017.

[3] 徐璐. 遗传算法模式识别的机理与意义[D]: [硕士学位论文]. 哈尔滨: 黑龙江大学, 2019.

[4] Liu, J.F., Li, X., Zhang, Q.S. and Zhong, G. (2022) A Novel Focused Crawler Combining Web Space Evolution and Domain Ontology. *Knowledge-Based Systems*, **243**, Article No. 108495. https://doi.org/10.1016/j.knosys.2022.108495

[5] Cheok, S.M., Hoi, L.M., Tang, S.-K. and Tse, R. (2022) Crawling Parallel Data for Bilingual Corpus Using Hybrid Crawling Architecture. *Procedia Computer Science*, **198**, 122-127. https://doi.org/10.1016/j.procs.2021.12.218

[6] 萧婧婕, 陈志云. 基于灰狼算法的主题爬虫[J]. 计算机科学, 2018, 45(S2): 146-148+166.

[7] 范会联, 李献礼, 曾广朴. 基于改进遗传算法的聚焦爬虫设计[J]. 计算机工程与科学, 2010, 32(5): 126-129.

[8] 刘成军. 基于查询扩展和多目标优化的主题爬虫系统的研究和实现[D]: [硕士学位论文]. 北京: 北京邮电大学, 2020.

[9] 白江伟. 改进的遗传算法在兰州自助终端巡检系统中的研究与运用[D]: [硕士学位论文]. 兰州: 兰州大学, 2019.

[10] 钱海军. 基于遗传算法的开放教育排课系统研究[D]: [硕士学位论文]. 广州: 广东技术师范学院, 2018.