基于遗传算法的定题信息搜索策略*

许欢庆[1] 王永成[1] 孙强[1]

[1]（上海交通大学计算机系 上海 200030）

E-mail: xuhuanqing@sjtu.edu.cn

摘要: 定题检索将信息检索限定在特定主题领域，提供主题领域内信息的检索服务。它是新一代搜索引擎的发展方向之一。定题检索的关键技术是主题相关信息的搜索。本文提出了基于遗传算法的定题信息搜索策略，提高链接于内容相似度不高的网页之后的页面被搜索的机会，扩大了相关网页的搜索范围。同时，借助超链 Metadata 的提示信息预测链接页面的主题相关度，加快了搜索速度。对比搜索试验证明了算法具有较好的性能。

关键词: 定题检索; 定题信息搜索; 遗传算法; Hub; Authority

引言

网络信息的高速增长，带给用户丰富的可供选择信息资源的同时，也带来了寻找有用信息的困扰。如何正确、快速、便捷地获取到相关信息成为研究热点。搜索引擎作为网络信息检索工具被用户广泛接受，但其依然存在诸多不足之处。近年来，研究学者提出了新一代搜索引擎的发展方向，定题检索是其中尤为突出的一种。所谓定题搜索引擎，就是将信息检索限定在特定主题领域，就主题相关的信息提供检索服务。不同于通用搜索引擎，定题搜索引擎的检索范围相对小，查准率和查全率易于保证。显然，主题相关信息的搜集是定题搜索引擎的核心。本文中，我们讨论定题检索系统中主题相关信息的搜索算法，提出了一种基于进化算法的定题信息搜索方案。在信息搜索过程中，引入遗传算法，借助遗传算子选择下一步搜索的网页。算法提高了链接于内容相似度不高的网页之后的页面被搜索的机会，扩大了相关网页的搜索范围，提高相关信息的查全率。同时，直接分析超链 Metadata 提示信息，以此预测链接页面的主题相关度，减少了计算时间，加快了搜索速度。

1 相关研究进展

近年来，一些研究学者开始定题检索技术的研究，工作重点主要围绕主题相关信息的搜集算法展开。

P. DeBra 等人首次提出称为“Fish-Search”的定题 Crawler 搜索算法[1]。定题 Crawler 动态维护一个按搜索优先权值排序的未搜集 Url 列表，并根据它选择下一步搜集目标。在信息搜索过程中，相关网页包含的超链被赋予比不相关网页包含的超链更高的优先权值，插入到未搜索 Url 列表中。

M. Hersovici 等人在“Fish-Search”算法基础进行了完善，提出了“Shark－Search”算法[2]。其算法在 Url 的优先权计算时考虑了超链描述文字的提示作用，同时，采用向量空间模型计算网页的相似度，细化了搜索优先权值的计算。

J. Cho 等人提出基于网页重要性优先的定题搜索算法[3]，列举出几种网页重要性评价指标：网页内容

* 本项目受到国家自然科学基金资助(60082003)。

相似性、网页的入度、网页的出度、网页的 PageRank 值和 Url 提示信息等。

F.Menczer 等人设计的 InfoSpider 引入神经网络[4]，通过抽取网页超链提示信息作为神经网络的输入，输出结果作为选择进一步搜索超链的依据。被搜索网页的相关度作为反馈训练神经网络。

S. Mukherjea 设计的 WTMS[5]（Web Topic Management System）中采用的 Crawler 在信息搜索开始之前，分析给定的起始页面集，生成一个特征表示向量。搜索过程中，分析被搜集的网页，计算其与特征向量之间的相似度。相似度超过预定值的页面包含的超链加入待搜集 Url 列表中。同时，引入启发规则提高搜索效率：只搜集离相关页面近的页面（父节点、兄弟节点和子节点）；优先搜集包含关键词的 Url 对应的页面；设立阈值，对一些相关度低于阈值的目录放弃搜索。

上述算法，搜索路径的选择基于如下假设：相关或重要的网页相互链接。事实上，存在许多相关网页链接在不相关的网页之后，譬如，许多研究学者设计个人主页时，将撰写的文章以及研究领域相关的文章目录列举出来。由于个人主页相似度不高，从而致使链接的页面搜索优先权偏低，甚至失去被搜索的机会。同时，基于相关或重要网页优先的搜索算法搜索区域局限在与起始页面相连的范围内，实现局部寻优，相关信息的搜索范围狭窄。

Chen H.等人提出设计客户端智能搜索引擎[6]。其 Spider 信息搜索方法与我们采用的方法类似，搜索算法中引入了遗传算法。搜索开始之前，首先将用户提问表达式提交给通用搜索引擎，获得搜索起始网页集。同时，从 Yahoo 中获取与主题相关的检索结果组成相关网页集。搜索过程，通过遗传算法的变异操作，将相关页面集中的页面作为新个体加入群体。此算法一定程度上扩大信息搜索范围，但搜索过程依然是基于网页相似度优先。

M. Diligenti 等人将网页之间的链接关系表示成层次关系[7]。根据给定起始页面集，建立称为“Context Graph”的层次图。图中每一层建立一个 Naive Bayes 分类器，对应一个队列。搜索过程中，将被搜索网页分类到相应的类别中，离中心越近的队列，网页搜集优先权越大。此算法引入分类器，分类器的性能将直接影响搜索的效果。

2 定题信息搜索算法

定题信息搜索的任务是在尽可能短的时间内，搜集尽可能多的主相关信息，而尽可能少的无关信息。搜索进行过程中，路径选择是最为关键问题，直接影响搜索的质量和速度。

分析相关网页构成的网络拓扑图，可以看到称为 Authority 和 Hub 的两种页面节点：被多个相关页面链结的页面称为 Authority 节点，hub 节点链接多个相关页面。通常，Authority 节点的内容具有主题权威性，有较强的主题相关度。Hub 节点对主题相关的信息进行汇总，提供主题以及相关主题的信息目录，其本身页面内容的相关度不高。基于内容相关度或权威度优先的搜索算法中，Authority 节点获得搜集机会较大，而 Hub 节点通常搜索优先权较低。事实上，Hub 页面的目录特性对主题相关信息的搜索有很大的提示作用，同时，Hub 页面的更新通常能及时反映网络中主题相关信息的变化。一定程度地提高 Hub 节点搜索机会可以实现扩大相关检索范围，实现全局寻优搜索。基于上述分析，我们在信息搜索过程中引入遗传算法，利用它全局寻优概率搜索的特点，引导搜索过程。待搜索的网页 Url 作为遗传个体，交叉和变异操作中，通过预定的概率引入 Hub 作为新个体加入群体中，进入下一代遗传进化。

2.1 网页Authority和Hub

互联网上，网页间链接关系类似科技文献间相互引用关系。文献引用关系可以抽象成有向图 G＝(N, E)，图中弧带有权值用于表示节点间的认同度。Kleinberg 在文献引用关系的基础上提出网络节点的 Hub 和 Authority 概念[8]。通过迭代计算网页的 Hub 和 Authority 度。算公式如下：

$$Authority(p) = \sum_{p' \to p} Hub(p') \qquad (1)$$

$$Hub(p) = \sum_{p \to p'} Authority(p') \qquad (2)$$

其中 $p \to p'$ 表示网页 p 有链指向页面 p'。

2.2 超链Metadata

超链是整个互联网信息库的重要组成部分。网页设计者设计页面时，通常借助超链完成进行整体信息框架的组织。为了便于用户选择继续浏览的超链，页面作者用较为简洁、正确的文字概括超链指向网页的内容。我们将之称为超链 Metadata。超链 Metadata 不仅对用户选择浏览有很好的提示作用，对定题信息搜索同样具有很强的提示意义。搜索模块可以根据超链 Metadata 信息进行搜索路径选择。我们选择 AnchorText 和 HREF 信息作为超链 Metadata，其中，对 HREF 的信息进行相应处理，剔除无用信息。下述是超链 Metadata 一个实例：

表 1 超链 Metadata 实例

* a smart web query method for semantic retrieval of web data* 超链 "Http://naxun.sjtu.edu.cn/retrieval/semantic"的 *Metadata*： *A smart web query method for semantic retrieval of web data + retrieval+ semantic+ naxun +sjtu*

2.3 GA算法实现

遗传算法是一种模拟生物在自然环境中遗传和进化过程而形成的自适应全局优化概率搜索算法。60 年代，美国密执安大学 Holland 教授最先提出。80 年代，Goldberg 对其进行了归纳总结，形成算法的基本步骤。

我们在定题信息搜索过程中引入遗传算法，借助选择、交叉、变异三个主要遗传算子进行搜索路径选择。算法的基本步骤如下：

- **Step1**. 初始化。定题信息搜集的初始条件是给定主题对应的检索提问式。将检索提问式提交给通用搜索引擎，检索结果构成搜索初始 URL 集。对搜索引擎返回的结果集进行处理，选择一定数目的 URL 组成初始群体。同时，计算出 Hub 页面集。
- **Step2**. 交叉操作。搜索模块下载当前群体 P(0)中个对 URL 对应的网页。抽取网页包含的超链，从未被搜集过的超链中挑选出被多个 P(0)个体页面指向的超链组成集合 C。
- **Step3**. 变异操作。按照预定的变异概率从 Hub 页面集中提取出一定的数量的未被搜索的 URL，同时根据交叉概率从集合 C 中提取相应数目的 URL，共同组成新的集合 Q。
- **Step4**. 选择操作。抽取集合 Q 中 URL 对应页面包含的所有超链，以及超链对应 Metadata，计算超链 URL 的 Fit 值，经过筛选组成群体 P(t+1)。
- **Step5**. 终止条件判断。根据已下载的页面数是否超过用户设定的阈值，或者进化代数 t 是否超出最大进化代数 T，决定是否中止进化进程。是则中止进化操作；否则跳转到交叉步骤，继续进化过程

步骤细节描述如下

2.3.1 初始化

我们选用了 Altavista 搜索引擎。给定主题对应检索提问式返回的结果集中选择前 n 个结果页面对应的 URL 组成集合 S'，$S' = (s_1, s_2, ..., s_n)$。对集合 S'，进行如下扩展，获得集合 S：

根据集合 S' 中网页的入链进行扩展。包含指向集合 S' 中页面的超链的页面组成集合 S''；

- 集合 S'' 中 URL 对应页面包含的超链 URL 组成集合 S'''；
- URL 集合 $S = S'+S''+S'''$，对集合 S 进行 URL 去重。

将集合 S 构造成有向图 $G'=(N,E)$，集合 S 包含的 URL 组成节点集 N，节点间的超链组成弧集 E（图 1）。根据公式（1）和（2），计算图中各节点的 Authority 和 Hub 值。按照节点 Authority 值从大到小的顺序，选取前 α 个节点组成初始群体 P(0)。同样，按照节点 Hub 值从大到小的顺序，组成集合 H，H = (h_1, h_2, …)。

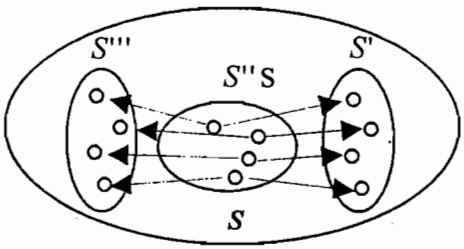

图 1　集合 S 的建立

2.3.2　交叉

交叉操作中，信息搜索器下载当前群体 P(t)对应的网页。网页经过分析，抽取出所有包含的超链。所有超链进行过滤，剔除指回上一页的超链，根据一定规则剔除广告链接、书签等。剩余未被搜集过的超链将相对 URL 转化成绝对 URL，构成集合 C。

计算指向集合 C 中的每个 URL 的 P(t)个体个数。根据计算值，降序排列集合中 URL。

2.3.3　变异

为了增加链接于不相关网页的页面被搜集的机会，扩大相关网页搜索范围，我们在变异操作中，按照预定的变异概率引入 Hub 节点。

根据变异概率 p_m，从集合 H 中提取出前 m 个未被搜索过的 URL，$m \le p_m \cdot \alpha$；根据交叉概率 p_c，从集合 C 中提取出 n 个未被搜索的 Url，$n \ge p_c \cdot \alpha$。抽取出来的 Url 组成集合 Q。其中，交叉概率 p_c 与变异概率 p_m 之和等于 1。

2.3.4　选择

选择操作将群体中适应度较高的个体按某种规则遗传到下一代群体中。

我们抽取集合 Q 中 URL 对应页面包含的所有超链，以及超链 Metadata，进行如下处理：

剔除已被搜集过的超链；

- 去除重复超链，合并超链 Metadata；
- 根据超链 Metadata，计算超链对应 Url 的适应度；
- 根据 Fit 值的大小，从所有超链中选取前 α 个组成群体 P(t+1)，剩余超链插入到集合 P 末尾。

个体适应度计算公式如下：

$$Fit(link_i) = S(q, MetaData(link_i)) \qquad (3)$$

其中，Metadata($link_i$)表示超链 $link_i$ 的 Metadata 信息，q 为定题领域对应的检索提问式。$S(d_1,d_2)$ 表示文本 d_1 和 d_2 间的相似度。为了缩短计算时间，我们对相似度计算公式进行了简化。假设，经过分词和剔除禁用词，检索提问式表示为 $q=\{t_1^q, t_2^q, ..., t_m^q\}$，$t_i^q$ 表示特征词。同样，超链 Metadata 表示为 $d=\{t_1^d, t_2^d, ... t_n^d\}$，$t_j^d$ 表示特征词，n 为特征词个数。相似度计算公式如下：

$$s(q,d)=\frac{Cnt(q\ \mathrm{I}\ d)}{Cnt(q\ \mathrm{Y}\ d)} \tag{4}$$

其中，$Cnt(x)$ 表示集合 x 的元素个数。

3 性能评价

3.1 试验设计

为了检验提出算法的性能，我们选定信息检索主题进行搜索测试。主题对应的检索提问为“Information Retrieval”。以下二种定题搜索算法被选择作为对比试验方案：

- **BestFirst 算法** BestFirst 算法的搜索过程中，从待搜集 URL 列表中挑选相似度值最大的 URL，抽取 URL 对应网页包含的超链，下载链接页面。计算下载页面与主题的相似度，并根据相似度大小，插入到待搜集 URL 列表中。这里，我们选用向量空间模型表示网页，网页的相似度采用向量间夹角余弦公式计算。
- **HITS 算法** HITS 算法在信息搜索过程中，计算待搜集 URL 列表中网页的 Authority 和 Hub 值，优先搜集 Authority 值大的网页。

采用不同搜索算法的 Robot 依次对给定主题进行搜索，下载的网页根据向量空间计算与主题的相似度，相似度超过给定阈值 γ 的网页被认为是相关网页，反之为不相关网页。统计搜索过程中下载页面数与其中相关页面数，分析之间比例关系。同时，分析不同交叉概率和变异概率下，我们提出算法的 Robot 下载页面数与其中相关页面数的比例关系。

3.2 测试结果与分析

图 2 显示了采用三种不同搜索的 Robot 在搜索进程中，下载页面数与相关网页比例变化情况（提出算法的变异概率为 20%，交叉概率为 80%）。曲线表明：搜索过程初期，提出的算法(GA)下载的相关网页占整个下载页面数比例为 54.3%，低于 BF 算法的 57.3%。随着搜索进一步深入，GA 算法搜索性能优势进一步明显，下载相关网页数所占比例为 61.8%，高于 BF 的 41.7%和 HITS 算法的 31%。

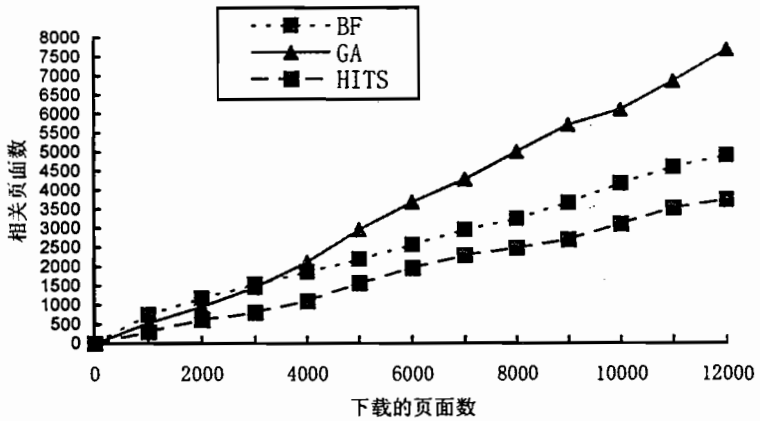

图2. 三种算法搜索性能比较

分析三种搜索算法下载的相关网页集，计算集合之间的重合度，结果如表 2。GA 算法搜索到的相关网页

集与其他两种算法的重合度远小于两种对比搜索算法的重合度。一定程度上表明 GA 算法的搜索范围较其他二者更为广泛。

表 2　三种算法下载的相关页面集合间重合度

	BF 与 HITS	GA 与 HITS	GA 与 BF
重合度	76.3%	42.7%	56.7%

为了检验变异和交叉概率的选择对算法搜索性能的影响。我们选用了四组概率组合，分别进行搜索测试。测试结果如图 3 显示。交叉概率设定为 10%时，其搜索特性比较类似 BestFirst 算法。交叉概率超出 20%时，其搜索性能有明显下降。测试表明：交叉概率选择对整个搜索性能有较大影响。选择过大的交叉概率将会因为引入过多的 Hub 节点，而致使整个搜索性能下降。较为合理的交叉概率选择范围为(15%～25%)。

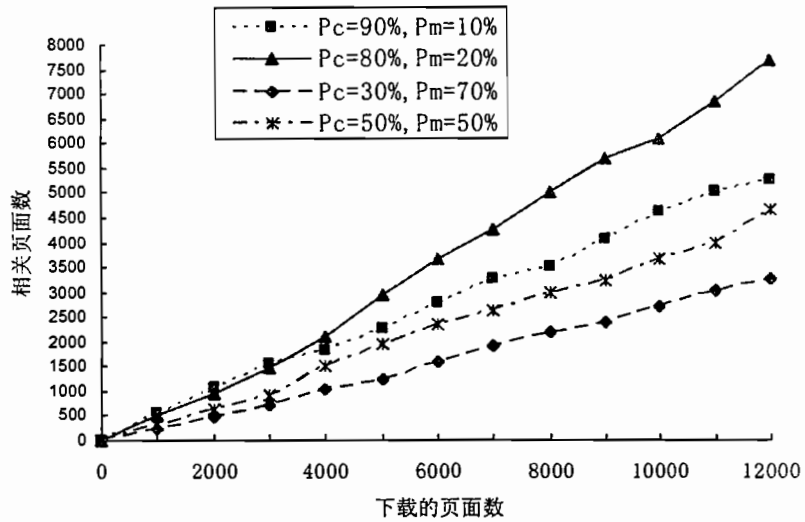

图3. 采用不同概率组合的GA算法搜索性能

4 结束语

定题信息搜索算法是定题检索的关键技术。本文中，我们提出了一种基于进化算法的定题信息搜索方案，扩大相关网页的搜索范围，一定程度上提高了 Robot 的查全率。同时，我们借助超链 Metadata 的提示信息预测链接指向页面的主题相关度，加快了搜索速度。下一步，我们将重点研究如何根据超链 Metadata 更为准确预测连接网页相关度，拟采用的方案借助概念间联想关系进行分析。

参考文献:

[1] P. DeBra, G. Houben, Y. Kornatzky and R. Post. Information Retrieval in Distributed Hypertexts, In Proceedings of the 4th RIAO Conference, 481 - 491, New York, 1994.

[2] M. Hersovici, M. Jacovi, Y. Maarek, D. Pelleg, M. Shtalhaim and S. Ur. The Shark-Search Algorithm - An Application: Tailored Web Site Mapping, In Proceedings of the Seventh International World Wide Web Conference, Brisbane, Australia. April 1998

[3] J. Cho, H. Garcia-Molina, L. Page. Efficient Crawling Through URL Ordering, L. Page. In Proceedings of the 7th International WWW Conference, Brisbane, Australia, April 1998.

[4] F.Menczer and R.Belew. Adaptive retrieval agents: Internalizing local context and scaling up to the web. Machine Learning, 39(2/3):203-242, 2000

[5] S. Mukherjea. WTMS: A System for Collecting and Analysing Topic-Specific Web Information, In Proceedings of the 9th International World Wide Web Conference, Amsterdam, Netherlands, May 15-19, 2000.

[6] Chen, H., Chung, Y.M., Ramsey, M. & Yang, C.C.: "An intelligent personal spider (agent) for dynamic internet/intranet searching",

Decision Support Systems v23 n1, 1998, pp. 41—58

[7] M. Diligenti, F. Coetzee, S. Lawrence, C. Giles and M. Gori. Focused Crawling Using Context Graphs, In Proceedings of the 26th International Conference on Very Large Databases (VLDB 2000), Cairo, Egypt, September 2000.

[8] J. Kleinberg, Authoritative Sources in a Hyperlinked Environment, Proc. Ninth Ann. ACM-SIAM Symp. Discrete Algorithms, ACM Press, New York, 1998, pp. 668-677.

作者简介： 许欢庆，1973 年生，博士研究生，主要研究领域为智能信息检索，Web 挖掘，信息过滤与推荐。 王永成，1939 年生，博士生导师，主要研究领域为信息检索，自动摘要，自然语言理解。孙强，1974 年生，博士研究生，主要研究领域为自然语言理解，信息家电。

Focused Crawling based on Genetic Algorithm

XU Huan-qing[1], WANG Yong-cheng[1], SUN Qiang[1]

[1](Department of Computer Science, Shanghai Jiao Tong University, Shanghai 200030, China)

E-mail: xuhuanqing@sjtu.edu.cn

Abstract: The exponential growth of information available on the WWW makes it increasingly difficult to crawl and index the entire internet for general-purpose crawlers. Rather than collecting and indexing all accessible web documents to answer all possible ad-hoc queries, focused crawler analyzes its crawl boundary to find the links that are likely to be most relevant for the crawl, and avoids irrelevant regions of the Web. In this paper, a new focused crawling approach based on Generic Algorithm is proposed. The method electively seeks out pages that are relevant to a pre-defined set of topics using Generic Algorithm, increases the crawling chance of the web page following the web page with the low content-relevance, and broadens the relevant-searching scope of crawlers. Meanwhile, the hyperlink metadata is used to predict the topic-relevance of the web page pointed and quickens the information crawling. Experimental results indicate that our approach has better performance.

Key words: Topic-Specific retrieval; Focused crawling; GA; Hub; Authority