基于超链评价的搜集算法

倪良胜 邹若凡 丁伟

（东南大学 计算机系，210096 南京）

【摘要】搜集器是搜索引擎的重要组成部分，它的核心搜集算法的工作效率不仅对搜索引擎本身，也会对整个网络的性能产生重要的影响。本文介绍的基于超级链接评价搜集算法可以有效地减少搜集器产生的流量，同时提高搜集器的工作效率。该算法已被一个面向 WEB 信息资源的搜索引擎 GONIA_WEB 用于日常使用，并产生了良好的效果。

【关键词】 搜索引擎；搜集系统；超级链接；

中图分类号：TP393

A Gathering algorithm based on Hyperlink Rank[1]

Ni Liang Sheng ， Zou Ruo Fan ， Ding Wei

(SouthEast University Computer 210096 NanJing)

【Abstract】As a important role of search engine, the information gathering subsystem influences the performance of search engine not only, but also the traffic of backbone with its gathering algorithm. The paper introduces a new algorithm which works on a hyperlink rank and can improve gathering system and backbone traffic effectively. The algorithm has been used in the daily running of Gonia_web—a Internet web search engine for its terrific behavior.

【Key words】Search Engine; Gather System; Hyperlink

1:引言

随着 Internet 的快速发展，Internet 上的网络信息呈现爆炸的趋势，使得各种 WWW 信息检索工具得到愈来愈广泛的应用。目前搜索引擎一般基于的基本结构一般如下图所示：

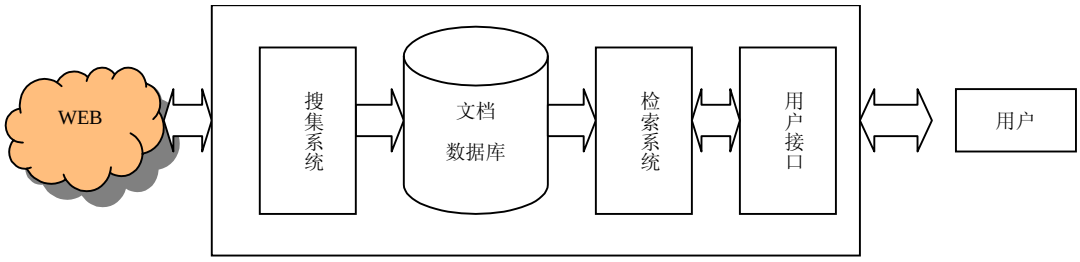

图 1：传统搜索引擎工作方式

由于 WWW 信息资源具有不断更新、海量数据的特征，因此给搜索引擎中的搜集系统造成很大的压力。由于 WWW 信息资源不断更新，搜集系统必须不断更新文档数据库；传统的搜索引擎实现中，搜集系统从一个 WEB 站点集合出发作为任务空间开始搜集文档。当遇到新站点，添加到这个 WEB 站点集合中去。当搜索结束，从集合开始点重新开始搜索，以防止信息的过时。但是对于海量数据，搜集系统往往效率有限。可以说搜集系统在很大程度上影响着搜索引擎的效率和查询效果。所以有必要研究新的搜集算法，以提高搜集系统的效率。本文所提出的基于链接评价的搜集算法，可用来降低搜集系统的强度，从而提高搜集系统的效率。

2：算法思想

---

[1]倪良胜：东南大学计算机系硕士研究生，主要研究方向为信息发现
邹若凡：东南大学计算机系硕士研究生，主要研究方向为信息发现
丁伟：东南大学计算机系教授

在大量浏览和分析WEB页面的同时，可以发现，不同的WEB站点之间，往往提供了对其他站点的链接，这在很大程度上方便了用户进行相关资料的搜索。而另外一方面，如果有一个站点被多个站点作为相关链接，显然该站点是一个重要站点；此外，如果一个站点被重要站点链接，显然比另外一个仅仅被普通站点链接的站点来得重要。

基于以上分析，可给每个站点赋予链接价值的特征，表示该站点在待搜集WEB页面空间中的重要情况。由站点间的链接情况，传递它们的链接价值。因而可给出一个链接评价算法，根据待收集WEB页面集合中的站点间的链接情况，不断修正WEB站点的链接价值，以此将待搜集WEB页面集合中的站点分成几个等级，并以此决定WEB站点的搜集频度。

给出主要数据结构：

No	站点标识	Linkval	rank	其他信息
1	WEB 1	linkval[1]	rank[1]	……
2	WEB 2	linkval[2]	rank[2]	……
……	……	……	……	……
M	WEB m	linkval[m]	rank[M]	……
……	……	……	……	……
Σ	WEB Σ	linkval[Σ]	rank[Σ]	……

相关变量定义：

RS：搜集系统用以搜集WEB页面的站点集合；

Σ：集合RS中站点数目；

rank[i]：第i个站点在搜集集合中的重要程度；

linkval[i]：第 i 个站点的链接价值。在搜集系统第一遍搜集之前，linkval[i] =0。在搜集过程中，如果当前第 k 个站点内含有对第i个站点的一个链接，则第k个站点的重要性就会传递给第 i 个站点，表现为：

$$linkval[i]=\begin{cases} linkval[i]+linkval[k] & linkval[k]\neq 0 \\ linkval[i]+1 & linkval[k]=0 \end{cases}, \text{其中} k\neq i \text{。}$$

3：该算法的主要步骤：

该算法的流程图：

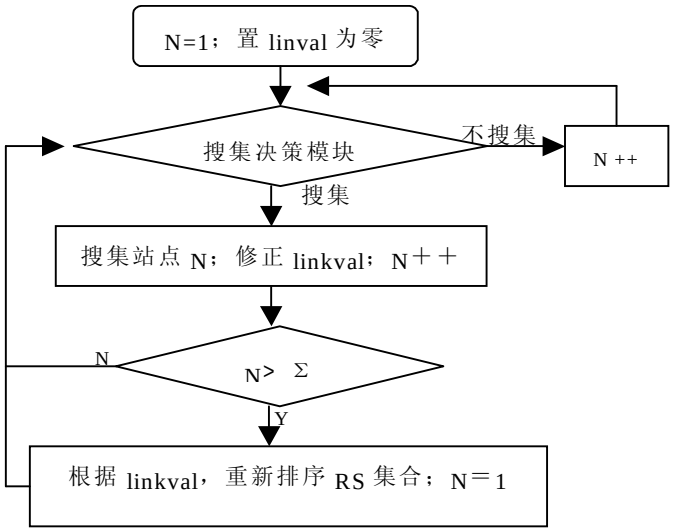

图 2：搜集系统流程图

（1）当搜集系统开始第一次搜集之前，初始化linkval表结构，为每个表项中的链接价值、重要性评价赋值为0。这是因为系统在第一次搜集之前没有学习到任何WEB站点之间链接的知识，只能把所有的待搜集站点视为普通站点，经过第一遍扫描才能给出每个站点的链接价值，从而区分站点类型，决定搜集策略。在一遍搜集完成之后，保存本遍

扫描所得的站点链接值，作为下一边搜集中区分站点类型的依据。

（2） 将 RS 集合中的站点根据其链接价值，分成若干个等级，等级越高，被收集的频度越大。该过程由搜集决策模块决定。下面给出一个简单的策略：

每次搜索 RS 集合前，根据站点的 rankval 由大到小排序后，再根据序号 No 从小到大，将 RS 集合中的站点等分成 4，2，1 三级，即：

$$rank[i]=\begin{cases}4 & No[i]\le\dfrac{\sum}{3}\\ 2 & \dfrac{\sum}{3}<No[i]\le\dfrac{2\sum}{3}\\ 1 & No[i]>\dfrac{2\sum}{3}\end{cases}$$

根据 rank[i]值，可以评价该站点的重要性，有以下结论：

$$\begin{cases}\text{重要站点} & rank[i]=4\\ \text{普通站点} & rank[i]=2\\ \text{次要站点} & rank[i]=1\end{cases}$$

再定义变量 searchtimes 为当前搜索遍数。让 searchtimes 从 1 到 4 不断循环。根据当前搜索遍数 searchtimes，决定搜索策略：

$$\begin{cases}\text{搜集} & rank\ [i]=4\\ \text{搜集} & rank\ [i]=2\text{且}(searchtime\quad s)\bmod 2=0\\ \text{不搜集} & rank\ [i]=2\text{且}(searchtime\quad s)\bmod 2=1\\ \text{搜集} & rank\ [i]=1\text{且}\ searchtime\quad s=0\\ \text{不搜集} & rank\ [i]=1\text{且}\ searchtime\quad s\neq 0\end{cases}$$

从而使得站点的搜集频度由它的重要性来决定，而不是一概而论。

（3）对于需要搜集的站点，根据深度遍历算法，来获取该站点所有网页的原始文档，并对网页进行文档过滤分析等操作，发掘出其中的有用超链，并计算相应的链接价值。搜集系统搜集 RS 集合中的站点时，搜集该站点页面中的超链情况，不断修正 RS 集合中的站点 i 的链接价值 linkval[i]。由于要搜索站点页面中的超链情况，会使得搜集系统的效率下降。但是因为搜集系统本来就需将页面进行切词分析，因而系统效率下降有限，而且由于非重要站点的搜集频度相对下降，所以整个系统的搜集频度是增加的。

（4） 搜集系统每完成一次轮回之后，根据本遍扫描所修正得出的站点的链接价值从大到小将 RS 集合重新排序。链接价值越大，对应站点的在数据结构中的序号 No 越小，其重要性越大，对应于上面给出算法，体现为 rank 值越小，该站点的被收集频度越大。

由于对于某些重要站点，有可能 linkval[i]增长极快。因此，有必要对 linkval[i]进行修正。因此，若计算出 linkval[i]>66535，则作如下修正：linkval[i]=sqrt(linkval[i])。

4：总结

以上给出了基于链接评价的搜集算法的基本实现。并应用到 GONIA-WEB 搜索引擎的实践中，由于对网站的重要性做出了区分，对于不同重要性的网站采取了不同的搜集策略，省去了对不重要的站点的反复搜集，加快了对重要站点的搜集频率。因此，基于链接评价的搜集算法，一方面节省了网络带宽，另外一方面，通过对重要站点的更新频率的加快，及时获得了用户所关心的站点的最新信息，满足了用户的需要。

5.参考文献（略）