利用链接分析技术提高搜索引擎查找质量的研究

Research on Utilization of Link Analysis Technology in Searchengine to Improve the Query Quality

中国科学院计算技术研究所 刘悦 杨志峰 程学旗 王斌 (北京 100080)

摘 要：文章对链接分析技术在搜索引擎的页面排名中的应用进行了深入细致的研究，对 google 的利用链接分析的页面排名算法 PageRank 算法进行了改进，并将其应用到我们自己的搜索引擎中，给出了它在提高查找质量的页面排名中的具体应用策略。

关键词：链接分析，搜索引擎，PageRank 算法，Web 图

1 引言

WWW 的产生为信息检索（IR）领域提出了挑战，当我们需要在 Web 上查找某一个方面的信息的时候，通常会使用 Web 搜索引擎，这时搜索引擎往往会返回成千上万甚至几百万个搜索结果。检索结果以每页十几个到几十个结果的方式以相关度降序排列提交给用户。在查看这些返回结果时，我们发现返回的页面或多或少都与查询相关联，但是检索的页面质量不够高，主要体现在以下几个方面：

(1) 不同主题的页面混杂在一起

一词多义是自然语言里面一种十分普遍的现象，同一个词在不同领域里面所表达的含义是不一样的，查询用户希望了解的一般都是某一个领域的的信息内容，而搜索引擎将不同领域的内容混杂在一起提供给用户，显然这样的页面质量就不是很高。

(2) 低质量的页面大量返回

这里提到的低质量的页面是指：页面虽然包含查询信息，但页面本身没有什么内容或者包含的内容价值不是很大，例如一些类似与广告，导航之类的页面。对查询用户来讲，它们的实用价值不大。

(3) 重复页面

由于镜像，常用文档的广泛传播等原因，Web 上存在着许多合法的重复页面，把众多重复页面提交给用户，显然没有达到高质量的检索的目的。

(4) Web 的规模和异构性

Web 的规模迅速增长和数据的异构性也是造成页面质量低下的原因。

为了能够提高查找的质量，产生了各种流派的搜索引擎，用户在查看搜索引擎返回结果时，一般只会看头几页的结果，很少有人会看排序在 100 或 200 之后的结果页面，几乎没有人会穷尽数量庞大的检索结果。基于链接分析的搜索引擎就是一个很好的解决方案。

收稿日期：2001－09－10

基金项目：国家 973 课题资助项目(G1998030413)

2 链接分析技术的背景

链接分析也称结构分析，它是把 WWW 上无数互相链接的页面看成一个巨大的链接有向图，也许是受到论文引用排名的影响，被引用的多的，或者是被权威期刊引用的论文的档次和质量就比较高。而在 WWW 上超链接上隐含了非常有用的这种引用信息。

这种方法是基于这样一个假设：即把超链接看作是对它所指向的页面的赞许。在这种假设下，某个页面 1 通过超链接指向页面 2，它隐含了这样的意义：页面 2 与页面 1 是主题相关的，页面 2 对于页面 1 来讲是值得关注的页面。为了能更好地说明，先对链接分析中用到的一些基本概念作定义。

如果将页面看作顶点，链接看作有向边，整个 Web 就可以看作一个有向图。我们可以利用复杂网络理论对其进行研究分析。其形式化的定义如下：

定义 1 有向图 G(V; E) 称为 Web 图，若 V 表示有限页面顶点集合，E 是由 V 中不同元素组成的有序对的集合，它表示页面之间的超链接。

$\forall v, w \in V$, 且 $v \neq w$, $\langle v, w \rangle \in E$

表示从页面 v 指向页面 w 的一个超链接。

目前利用链接分析对 Web 页面性质的研究做得比较好的有两家，一家是 Google 采用的方式；一家是 IBM 的算法。

在 Google 的算法中，他们假设 Web 上有一个随机的浏览者，这个随机的浏览者从一个任意给定的页面出发，按照页面上的链接前进，在每一个页面，浏览者都有可能不再对本页的面的链接感兴趣，从而随机选择一个新的页面开始新的浏览，Pagerank 是他访问到页面 A 的概率，Google 全局地为每个页面计算一个 pagerank 值，作为页面的质量评分。

在 IBM 的 CLEVER 系统中的 HITS 算法中，认为 Web 页面都有被指向，作为权威(Authority)和指

向其它页面作为资源中心（Hub）的两个方面的属性，其取值分别用 A(p) 和 H(p) 表示，A(p) 为所有指向 p 的页面 q 的中心权重 H(q) 之和，同样页面 p 的中心权重 H(p) 值是所有 p 所指向的页面 q 的权重 A(q) 之和，如下式：

$$A(p) = \sum H(q_i)$$

其中：q_i 是所有链接到 p 的页面。

$$H(p) = \sum A(q_i)$$

其中：q_i 是所有页面 p 所链接到的页面。

从这两个概念的名称可以推想到，authority 是重要的信息资源，hub 是指向信息资源的中心点。相应的 Web 图是一个二分图。Hub 和 authority 之间是相互加强的关系，一个好的 hub 必然指向许多好的 authority，同样一个好的 authority 必然被许多好的 hub 链接。

无论是 CLEVER 系统还是 Google 都是利用链接分析技术对搜索引擎的结果排序进行改进，但通过比较发现 Google 是利用 Web 图以及浏览行为的随机属性全局地为每个页面计算一个 pagerank 值。而 IBM 的 CLEVER 系统采用的策略，则是在收到查询请求时，临时生成一个相关页面集合，利用 HITS 算法从该集合中蒸馏（distill）出权重最高的中心页面和权威页面。从以上的比较可以看出，CLEVER 系统的 HITS 算法在对 Web 的认识和计算上更加深入细致，但由于计算页面权重使其效率受到很大限制，难以形成供大量用户使用的商用搜索引擎，另外 CLEVER 系统的结果限制在由传统搜索引擎的最前面的搜索结果为种子，进行扩展得到的集合范围中，对链接分析在超大规模范围挖掘全局信息的特长能力没有利用到。因为链接分析是挖掘深层隐藏信息，需要迭代计算，计算量比较大，我们认为链接分析是适合预处理阶段使用的技术，而 Google 的所有计算都是在预处理阶段完成的，所以它的页面评分比较客观，我们的系统中链接分析部分就采用了 pagerank 算法，并对其进行了改进。

3　pagerank 算法简介

PageRank 算法是 Standford 大学的研究人员开发的 Google 搜索引擎的页面质量评价算法。Google 的研究人员观察到在 Web 图这一有向图中，不同的结点被访问到的概率是不同的，他们为用户在 Web 上的浏览行为建立了一个模型：即以概率 d 顺着朝链接点击访问，或者以概率 1 – d 从一个新的页面开始访问，在该模型下，某一个页面 t 被访问到的概率只与指向它的页面的概率有关。为了能更清楚地表达，我们对算法中涉及的有关概念加以说明。

定义 2　设 G = (V; E) 是一个 Web 图，且 $\forall t \in V$，称如下集合为页面 t 的前集：

$$\{y \mid \langle y, t\rangle \in E, \forall y \in V\}$$

记为 $\cdot t$。

定义 3　设 G = (T; E) 是一个 Web 图，且 $\forall t \in V$，称如下集合为页面 t 的后集：

$$\{y \mid \langle t, y\rangle \in E, \forall y \in V\}$$

记为 $t^{\cdot}$。

定义 4　$\forall t \in T$，记 $|\cdot t|$ 为 t 的前集中的元素的个数。$|t^{\cdot}|$ 为 t 的后集中的元素的个数。

在PageRank 算法中页面 t 被访问到的概率 Pr(t)（通常称为 rank 值）是按照下式给出的：

$$Pr(t) = (1-d)/MAX + d(\Sigma(Pr(t_i))/|t_i^{\cdot}|) \quad (1)$$

其中 $t_i \in \cdot t$，MAX 为页面的总数，d 称为影响因子。它一般根据经验进行指定。

概率 Pr(t) 反映了 t 的重要程度，在 PageRank 算法中将之用作页面质量的评价参数，它采用的方法就是根据公式（1）进行迭代运算，直到计算出的值收敛为止。为每一个页面计算一个 rank 值。

4　对算法的改进

直观上讲从公式（1），觉得该算法的计算很简单，但在具体实现的过程中，发现算法的许多方面是可以改进的，改进后的算法能够更迅速地获得更能反映实际情况（好页面的 rank 值就排在前面）的 rank 值。具体的改进如下：

（1）在进行迭代之前，对页面进行预处理，过滤掉噪音链接，标记出不参与迭代的页面。

通过分析数据发现并不是所有的页面都存在前集和后集，对于前集为空集的页面，PageRank 算法给出的公式（1）对它们是没有迭代效果的，所以这样的页面可以不参加迭代，而是直接指定一个固定的 rank 值。剩下的页面，为了加快查找的效率，按页面号进行排序。

（2）对于后集中结点数量很大的结点的预处理

对于公式（1），如果 $|t_i^{*}|$ 很大的话，$Pr(t_i)/|t_i^{*}|$ 的值就会很小，直观上讲，也就是说在 Web 图上页面 t_i 对页面 t 的 rank 值的贡献很小。所以对于这样一类结点，没有必要简单的迭代，在改进的算法中，预先给定一个阈值，对于后集中结点数量大于阈值

的结点，指定其 rank 值为 1/MAX。这样进一步缩减了迭代的规模。

(3) 规范化和收敛性的判定规则

根据概率本身的含义，每一次迭代计算出来的结果规范化之后的∑和最好为 1，但是这样得到的每一个页面的 rank 值就非常小(小数点之后 6 位才有数据)，不利于收敛性的判定。在改进后的算法中，为了让计算所得的结果具有可比性，又不至于太小，我们给出了如下的规范化规则：

在迭代开始时，已经给每一个页面指定了一个初始的 rank 值 (由于它对最终的结果没有影响，所以是任意指定的)。将所有页面的初始 rank 值相加，得到一个常量 M。

① 对第 i 次迭代所得的每一个页面的 rank 值求和，得到 M'；

② 令 K = M/ M'；

③ 对第 i 次迭代所得的每一个页面的 rank 值做如下规范化：

$rank^{(i)}[j] = k * rank^{(i)}[j], j = 1, 2, \cdots, MAX$；

④ END 规范化。

对于收敛性的判定，采用的判定性规则是：

预先指定一个充分小的 δ，整个过程都在用下面的公式 (2) 对相邻两次的迭代结果进行比较，如果满足公式(2)，就停止迭代。

$$MAX(rank^{(i)} - rank^{(i+1)}) - MIN(rank^{(i)} - rank^{(i+1)}) < \delta \quad (2)$$

从公式 (2) 可以看出，对于收敛性的判定不仅仅局限于所有页面的 rank 值都有减小收敛的趋势，而且要求所得的结果波动性要小 (MAX 与 MIN 的差足够小，保证了这一点)。

算法收敛之后，对每一个页面都得到一个 rank 值，这个值将作为对搜索出来的页面进行排名依据之一。

5 链接分析在页面排名中的应用

在实验中，把基于内容和基于链接分析的结果综合考虑，利用如下公式得到最后的页面排名。

$$W_d = W_{dc} + W_{dl}$$

$$W_{dl} = \frac{W_{dc}}{\log \frac{PR_{max} * k}{PR_d}}$$

其中，W_d 为文档 d 的权重，W_{dc} 为用基于内容的算法得到的文档 d 的权重，W_{dl} 为用链接分析得到的文档 d 的权重，PR_d 为文档 d 的 PageRank 值，PR_{max} 为最大的 PageRank 值，k 为大于 1 的常数。

如果分别给 W_{dc} 和 W_{dl} 赋一个权值，那么需要不断调整这两个权值以达到最好效果，引入的人工干预过多，降低了自动化程度，因此，根据 W_{dc} 和 PR_{max}/PR_d 的值来确定 W_{dl} 的值。因为基于内容的算法已经非常成熟，这种方法更强调 W_{dc} 的作用，提高了结果的可信度。

6 实验结果

利用 TREC WEB TRACK WT10G 的数据进行了实验，用改进的算法在 TREC Web Track 中的 WT10g 的数据上进行了实验，实验的结果表明的改进方案是有效的，首先在预处理阶段，经过预处理之后，in_link 和 out_link 中只保留了 1/5 左右的有用超链接。经过 20 次左右的迭代，就得到了波动比较小，已经收敛的运行结果。

下面是基于文本内容的实验结果，由 treceval 输出。这个结果作为其它测试的基准。

Queryid (Num): 46
Total number of documents over all queries
Retrieved: 25410
Relevant: 2484
Rel_ret: 1071
Interpolated Recall - Precision Averages:
at 0.00 0.3170
at 0.10 0.2152
at 0.20 0.1826
at 0.30 0.1501
at 0.40 0.1143
at 0.50 0.0973
at 0.60 0.0656
at 0.70 0.0527
at 0.80 0.0254
at 0.90 0.0214
at 1.00 0.0115
Average precision(non - interpolated)over all rel docs
0.1025
Precision:
At 5 docs: 0.1565
At 10 docs: 0.1522
At 15 docs: 0.1304
At 20 docs: 0.1283
At 30 docs: 0.1130
At 100 docs: 0.0793
At 200 docs: 0.0645
At 500 docs: 0.0381
At 1000 docs: 0.0233

R – Precision(precision after R(= num_rel for a query) docs retrieved):

Exact: 0.1244

下面是使用内容加链接算法所得的结果

Queryid (Num): 46

Total number of documents over all queries

Retrieved: 25410

Relevant: 2484

Rel_ret: 1071

Interpolated Recall – Precision Averages:

at 0.00 0.3143

at 0.10 0.2142

at 0.20 0.1838

at 0.30 0.1504

at 0.40 0.1151

at 0.50 0.0981

at 0.60 0.0655

at 0.70 0.0526

at 0.80 0.0254

at 0.90 0.0214

at 1.00 0.0115

Average precision (non – interpolated) over all rel docs

0.1024

Precision:

At 5 docs: 0.1565

At 10 docs: 0.1435

At 15 docs: 0.1290

At 20 docs: 0.1239

At 30 docs: 0.1109

At 100 docs: 0.0800

At 200 docs: 0.0648

At 500 docs: 0.0381

At 1000 docs: 0.0233

R – Precision(precision after R(= num_rel for a query) docs retrieved):

Exact: 0.1238

根据结果，虽然使用链接分析算法之后与基准测试基本持平。但它为页面排名增加了一个十分有利的依据，通过分析数据，发现在给定的封闭的页面集合中有用的超链接数量比较少，也就是说真正参与迭代，并在 pagerank 的计算中起作用的链接数量不足。如果链接数量是充足的话，应该可以得到更好的结果，因为从单纯的 pagerank 的排名上，发现它已经客观地对页面质量的好坏给出了一个排名。所以从这一点来讲，链接分析在页面质量的评价上是客观的和有效的。

7 结束语

从实验的结果和在算法的改进过程中我们发现很多值得进一步探讨和研究的问题：首先在超链接的预处理方面，还存在着许多可以进一步拓展的方向，其次可以考虑将静态的封闭页面集合变成可扩张的开放集合；另外还可以考虑扩大实验的数据集合，增加有效链接的数量，来提高 pagerank 算法的效率，加大它在搜索引擎页面排名中所占的分量，提高页面质量的检索效果。

参考文献

[1] S Brin and L Page. The Anatomy of Alarge Scale Hypertextual Web Search Engine Proc 7th www, 1998.

[2] OGAWA Yasushi, MANO Hiroko, NARITA Masumi, HONMA Sakiko: Structuring and Expanding Queries in the Probabilistic Model. TREC – 9 **论文集**.

[3] Lawrence Page, Sergey Brin, Rajeev Motwani, Terry Windograd, The Pagerank Citation Ranking: Bring Order to the Web. Jan, 29, 1998.

[4] CLEVER **项目组成员**. **网络的超搜索**. **科学**, 1999, 10.

[5] Sergey Brin, Larry Page. Google search engine. http: // google. standford. edu.

[6] J Kleinberg. Authoritative sources in a hyperlinked environment. Proc 9th ACM – SIAM SODA, 1998.

[7] **冯国臻**. **基于结构分析的大规模 WWW 文本信息检索技术的研究**: [**博士学位论文**]. **中科院计算所**.

LIU Yue, YANG Zhi-feng, CHENG Xue-qi, WANG Bin

(Institute of Computing Technology Chinese Academy of Sciences, Beijing 100080)

Abstract: In this paper, we research the link analysis technology in detail. Under the base of the analysis for the PageRank algorism, we proposed an improving method for the exised algorism and applied the results in our own search engine, giving a strategy that used link analysis technology to improve the query quality.

Key words: Link analysis, Search engine, PageRank algorism, Web graph.

刘　悦　博士研究生。研究方向为海量信息处理、知识检索与数据挖掘、算法设计与分析、petri 网理论与应用等。

杨志峰　博士研究生。研究方向为海量信息处理、搜索引擎、智能信息处理、知识检索等。

程学旗　博士研究生，副研究员。研究方向为 Internet 高性能软件、智能信息处理、知识检索与算法分析、信息安全等。

王　斌　博士，副研究员。主要从事 Internet 信息处理、自然语言处理等方向的研究。