Dijkstra 算法在 Web 结构挖掘的应用

林耀进[1]，王晨曦[2]
（1、漳州师范学院计算机科学与工程系 福建 漳州 353000 2、漳州职业技术学院计算机工程系 福建 漳州 353000）

【摘 要】：该文从 Web 结构挖掘角度出发，利用概率论分析了 Web 结构挖掘的 PageRank 算法，得出挖掘结果，最后介绍 Dijkstra 算法在其挖掘结果的应用。

【关键词】：Web 结构挖掘，PageRank 算法，Dijkstra 算法，权重

搜索引擎 Google 的成功，取决于它采用了有效的 Web 信息挖掘技术。Web 挖掘指在 WWW 上挖掘潜在的、有用的模式及隐藏的信息过程[1]。Web 挖掘分为 Web 内容挖掘、Web 结构挖掘和 Web 使用挖掘。其中结构挖掘则是从人为地链接结构中获取有用知识的过程。在设计搜索引擎等服务时，对 Web 页面的连接结构进行挖掘以得出有用的知识是提高检索的重要手段。

本文从 Web 结构挖掘入手，对 Web 结构挖掘的 PageRank 算法结合概率论进行研究应用。然后介绍 Dijkstra 算法在挖掘结果中的应用。

1. 算法的介绍

1.1 PageRank 算法

在 PageRank 方法中的 PageRank 被定义为[2]：设 u 是一个 Web 页，Fu 为所有 u 指向的页面的集合，Bu 为所有指向 u 的页面的集合。设 Nu=|Fu|为从 u 发出的链接的个数，那么 u 页面的 PageRank 可以定义为：

$$R(u) = c\sum_{v\in Bu} R(v)/Nv \qquad (1)$$

其中 c(<1)为归一化因子(因为所有页面的 RankPage 之和为一个常数)。PageRank 算法的实现过程：将网页的 URL 对应成唯一的整数，把每一个超链接用其整数 ID 存放到索引数据库中，经过预处理后，设每个网页的初始 PR 值为 1，通过以上的递归算法计算每一个网页的 PageRank 值，反复进行迭代，直至结构收敛。显然，PageRank 值越大，该页面权威性越高。

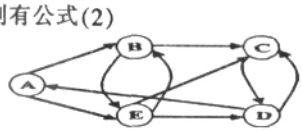
1.2 Dijkstra 算法

Dijkstra 算法是由荷兰计算机科学家艾兹格·迪科斯彻发现的。Dijkstra 算法是图论学中求解最短路径问题的经典算法，Dijkstra 算法建立在抽象的网络模型上，把路抽象为网络中的边，以边的权值来表示路相关的参数，算法确定了赋权网络中从某点到所有其它结点的具有最小权的路。权的含义是广泛的，可以表示距离、费用、数量等等[3]。

Dijkstra 算法的输入包含了一个有权重的有向图 G，以及 G 中的一个来源顶点 S。我们以 V 表示 G 中所有顶点的集合。每一个图中的边，都是两个顶点所形成的有序元素对。(u,v)表示从顶点 u 到 v 有路径相连。我们以 E 所有边的集合，而边的权重则由权重函数 w: E → [0, ∞]定义。因此，w(u,v)就是从顶点 u 到顶点 v 的非负花费值 (cost)。边的花费可以想像成两个顶点之间的距离。任两点间路径的花费值，就是该路径上所有边的花费值总和。已知有 V 中有顶点 s 及 t，Dijkstra 算法可以找到 s 到 t 的最低花费路径 (i.e. 最短路径)。这个算法也可以在一个图中，找到从一个顶点 s 到任何其他顶点的最短路径。

2. 仿真实验

2.1 根据图(1)的 Web 页面结构和公式(1)的 PageRank 定义，很容易列出 Web 的 PageRank 之间的线性关系，不妨引入记号 u=R(u)则有公式(2)

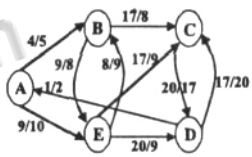


图(1) Web 页面结构

对公式(2)求解得公式(3)

$$\begin{cases} A=\frac{D}{2} \\ B=\frac{A}{2}+\frac{E}{3} \\ C=\frac{B}{2}+\frac{D}{2}+\frac{E}{3} \\ D=C+\frac{E}{3} \\ E=\frac{A}{2}+\frac{B}{2} \end{cases} \qquad (2) \qquad \begin{cases} 2A=D \\ 9A=10E \\ 4A=5B \\ 17B=8C \\ 9B=8E \\ 20C=17D \\ 9C=17E \\ 20E=9D \end{cases} \qquad (3)$$

对于页面 B, E, 其中页面 B、E 存在直接链接关系：把 W(B, E)声明为代表 B->E 的回顾因子权重。W(,)表示两个页面之间的紧密链接关系，W(,)值越大说明两个页面之间关系越紧密。如 W(B, E)=9/8, W(E, B)=9/8。根据图(1) Web 页面结构和表达式(3)，很容易得出图(2) Web 带权重的页面结构。



图(2)Web 带权重的页面结构

2.2 在图(2)中，我们发现对于每一对页面 Pi, Pj∈P(P 代表页面的集合)，Pi≠Pj，若从 Pi 到 Pj 存在路径数≥2，W(Pi, Pj)都是一样的。如对图(2)，页面 A 到 D 存在多条路径，其中：R1: A->B->C->D; R2: A->E->D; R3: A->B->E->D; R4:A->E->B->C->D。W(A, D)都为 2。依据概率论的积事件，说明若从页面 A 搜索页面 D，无论走哪条路经，概率是等同，即 P(A∩B∩C∩D)=P(A∩E∩D)=P(A∩B∩E∩D)=P(A∩E∩B∩C∩D)。

同样，在图(2)中，我们发现各条路径的回顾因子权重和不一定相等。如路径 R1: W_{R1}=4/5+17/8+20/17=2789/680；路径 R2: W_{R2}=9/10+20/9=281/90；路径 R3: WR3=4/5+9/8+20/9=1259/360；路径 R4: WR4=9/10+8/9+17/8+20/17=62306/12240。即 $W_{Ri} \neq W_{R2} \neq W_{R3} \neq W_{R4}$ 其中 W_{Ri} 代表路径 Ri 的权重和。依据概率论的和事件，WRi 越大说明，路径 Ri 的页面出现的概率越大，如在图(2)中，当从页面 A 搜索页面 D 时，路径 R4 中的页面出现概率较大。

2.3 根据前面推出的结果，利用 Dijkstra 算法对带回顾因子权重的 Web 页面结构图求两点之间的最短路径，其伪代码如下(其中 u,v 为顶点，d[v]中储存的便是从 s 到 v 的最短路径，w(u,v)代表顶点 u 与顶点 v 之间的权重)：

```
1  function Dijkstra(G, w, s)
2      for each vertex v in V[G]                    // 初始化
3          d[v] := infinity
4          previous[v] := undefined
5      d[s] := 0
6      S := empty set
7      Q := set of all vertices
8      while Q is not an empty set                  // Dijstra 算法主体
9          u := Extract_Min(Q)
10         S := S union {u}
11         for each edge (u,v) outgoing from u
12             if d[v] > d[u] + w(u,v)              // 拓展边(u,v)
13                 d[v] := d[u] + w(u,v)
```

（下转第 103 页）

模型。

(4) 索引数据库：就是存放索引器所检索到的信息。

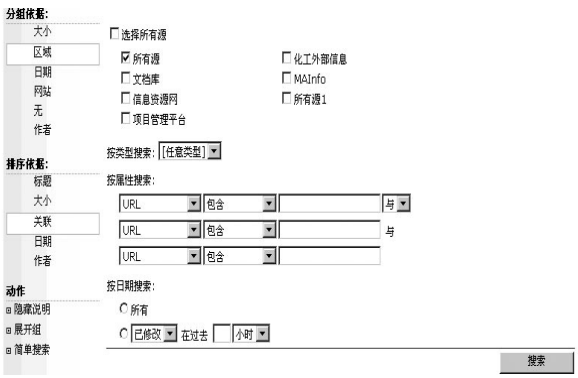
(5) 检索器：从索引中找出与用户查询请求相关的文档。首先采用与分析、索引文档相似的方法来处理用户查询请求。

(6) 用户接口：为用户提供可视化的查询输入和结果输出界面。在查询输入界面中，用户按照搜索引擎的查询语法指定待检索词条及各种简单，高级检索条件。

5. 门户内搜索引擎的用户界面和功能展示

本搜索引擎的用户接口界面：

高级搜索界面：



当输入"大庆"时的搜索情况：

6. 结语

企业信息门户(EIP)不能简单视为一个企业网站，它提供一些企业生产、销售、产品、服务、财务管理信息，它还能帮助企业实现多业务系统(例如 OA、ERP、MIS、CRM、HR)的集成、能对客户的各种要求做出快速响应、并且能对上述异构系统的数据进行集成、展现、应用。员工可以通过企业信息门户管理系统(EIP)访问企业的生产数据、管理信息、客户信息、销售信息、库存信息、财务信息，以最低的成本共享和利用企业的所有信息、文档、知识。因此，高性能的搜索引擎更能发挥门户的优势，给中石油勘探与生产分公司的员工提供了更进一步的方便。

但此搜索引擎仍需改进的地方有：

提高它的个性化：搜索引擎门户信息检索系统通常作为一种服务器程序运行，同时响应多个用户的请求。事实上，不同领域背景、知识结构的用户对检索结果的要求是不一样的。故更应提高它的智能性，更适合用户需求，满足用户的兴趣变化而动态调整。

主动查询问题：搜索引擎的检索工作是用户驱动的，即由用户显式地提出检索请求，系统给出响应。这是一种较为被动的信息获取方法。还有待建立一种主动协助用户获取信息的机制，将门户中用户需要的信息主动通知给用户。

参考文献：

1.蒋凯，武港山 基于 Web 的信息检索技术综述[J]，计算机工程，2005(12)

2.雷景生，林冬雪，符浅浅 基于改进向量空间模型的 Web 信息检索技术研究[J]，计算机工程，2005(1)

3. 吴栋，滕育平 中文信息检索引擎中的分词与检索技术[J]，计算机应用，2004(7)

4.王枝军，强俊，程效军，基于 Web 的信息检索系统的设计与实现[J]，计算机工程与设计，2006(3)

5.彭波 大规模搜索引擎检索系统框架与实现要点[j],计算机工程与科学，2006(3)

(上接第 74 页)

14　　previous[v] := u

在上面的算法中，u:=Extract_Min(Q)在在顶点集 Q 中搜索有最小的 d[u]值的顶点 u。

现在我们可以通过迭代来回溯出 s 到 t 的最短路径。

```
1 S := empty sequence
2 u := t
3 while defined u
4 insert u to the beginning of S
5 u := previous[u]
```

得出序列 S 就是从 s 到 t 的最短路径的顶点集。从顶点集发现：路径越短的页面出现的概率越低。此结论可以在搜索引擎中得到应用。

3.总结

Web 结构挖掘技术是 Web 数据挖掘的重要内容之一，它主要对 Web 页之间的链接结构进行的数据挖掘。本文把网络中的页面链接图转换为带权重有向图，以该图为模型，对 Web 结构挖掘的 PageRank 算法结合概率论进行研究。然后介绍 Dijkstra 算法在挖掘结果中的应用。

参考文献：

1.韩家炜，孟小峰，王静等 Web 结构研究[J] 计算机研究与发展 2001，35(4)

2.Brin S, Page L. The anatomy of a large- scale hypertextual Web search engine[J/OL] http:/www.db.standford.edu/backrub/google.html 2003

3. 鲍培明 距离寻优中 Dijkstra 算法的优化 [J] 计算机研究与发展 2001，38(3)

阅读此文的还阅读了：

1. HITS算法在Web挖掘中的应用与改进
2. 一种基于云计算的Web结构挖掘算法
3. Web结构挖掘算法探讨
4. 基于Dijkstra算法的Web服务合成选择策略研究
5. 专业网站零输入导航引擎设计与实现
6. Dijkstra算法在智能公交查询系统中的应用
7. Dijkstra改进算法在地震救援中的应用
8. Dijkstra改进算法在机器人避障问题的应用
9. Web数据挖掘算法
10. Dijkstra算法在Web结构挖掘的应用
11. 基于 Dijkstra算法的
12. 结构挖掘中web有向图模型的改进算法
13. Trawling算法在Web结构挖掘中的应用
14. Web结构挖掘及其算法分析
15. Web结构挖掘算法研究
16. Dijkstra算法在最优投资策略问题中的应用

17. 改进的PrefixSpan算法在Web挖掘中的应用

18. Web结构挖掘及HITS算法分析

19. Web结构挖掘与其基于超链接结构的算法

20. 基于Web结构挖掘算法的网站构建

21. Dijkstra算法的优化

22. Web结构挖掘

23. Dijkstra算法在企业成本控制的应用

24. 计算机行业挖掘结构性机会

25. Dijkstra算法在物流网络设计中的应用

26. 结构挖掘中web有向图模型的改进算法

27. Web结构挖掘研究

28. 基于Web页面链接结构的挖掘算法

29. XML的DOM树结构在WEB挖掘中的应用

30. 应用Web结构挖掘的PageRank算法的改进研究

31. 二进制挖掘算法在Web使用挖掘中的应用

32. Web结构挖掘中HITS算法的改进

33. 模式恢复算法在Web使用挖掘中的应用

34. 云计算在Web结构挖掘算法中的运用研究

35. Dijkstra算法在配电网抢修中的应用

36. Dijkstra算法在人群疏散上的应用

37. Dijkstra算法在路由选择中的应用

38. Dijkstra算法在最优投资策略问题中的应用

39. 基于Web结构挖掘算法的网站构建

40. 基于Web结构挖掘的HITS算法分析及改进

41. Dijkstra矩阵算法

42. Web结构挖掘研究

43. 云计算在Web结构挖掘算法中的运用研究

44. 数据聚类算法在web数据挖掘中的应用

45. Dijkstra算法在最短旅游路径中的应用

46. Dijkstra算法在部队快速行进中的应用

47. 基于PageRank和HITS的Web结构挖掘算法研究

48. 基于GIS的Dijkstra算法在运输系统的应用

49. 改进Dijkstra算法在PGIS中的应用

50. 基于Web结构挖掘的HITS算法研究