

WEB 超链分析及应用

□向桂林

摘要 文章分析了传统的基于文本处理的信息检索算法在处理 WEB 页面时遇到的问题,指出在处理 WEB 页面时,应分析页面中超链的意义,给出了超链分析在网络爬行和检索结果排序两个方面的应用及相应的算法。

关键词 WEB 页面 超链分析 信息检索

1 传统信息检索方法遇到的问题

信息检索是计算机科学的一个子领域,其目的是在给定的文献集合中找出与用户需求相关的所有文献。从这个意义上讲,信息检索也叫文献检索。在 WEB 未出现之前,信息检索系统安装在图书馆或信息研究机构中,供咨询馆员使用。这些信息检索系统的算法常常只是对文档中的词语或字进行分析。WEB 页面的出现,改变了传统文本文档的性质:WEB 页面中有一些超链信息和标记。如果我们不顾处理对象性质上的变化,仍旧沿用传统的基于文本处理的信息检索方法,会有什么弊端?让我们举一个传统的基于文本处理的信息检索算法:向量空间模型^[1],看看它的工作原理,以及在处理 WEB 页面时会出现的弊端。

向量空间模型算法认为一篇文献是按词构成的一个高维向量空间。每一篇文献和查询被表示为该向量空间中的一个词语向量。在文献中出现了的词语所对应的相应向量入口为正值,未在文献中出现的词语所对应的向量入口为零。而且,词语所对应的向量入口值常常是一个函数,其值随该词语在每一文献中出现的频次增高而增高,随该词在不同文献中出现的频次增高而减少。可表述为:

$$\text{Weight}_{ik} = \text{Freq}_{ik} / \text{DocFreq}_k$$

其中 Weight_{ik} 表示词 k 在文献 i 中的权重; Freq_{ik} 为词 k 在文献 i 中的出现频次; DocFreq_k 为含有词 k 的文件数量^[2]。换句话说,词 k 在越多的文献中出现,那么,词 k 对文献 i 的特征描述越少;词 k 在文件 i 中出现的频次越高,那么,词 k 也越能反映文献 i 的特征。词语向量可以被标准化(比如维数、取值范围等),以此来适应不同长度的文献。换言之,不管文献长度如何,只提取固定数量的

词语来表达文献的内容。文献与查询之间的相似性通常是由它们的词语向量间的点积来计算。对于一个给定的查询,点积运算会赋给每一篇文献一个非零值,在响应查询时,被赋有正值的文献以分值递减的次序返回给用户(即,点积的值越小,说明该文献与查询越匹配)。

基于上述原理,一个 WEB 页面的作者,如果他的页面与商业利益有关的话,会想尽办法使他的页面在检索结果中被排在前面。这里有如下一些办法可以钻这些传统的基于文本处理的信息检索算法的空子,比如:某页面是做汽车广告的,就可以在页面的<meta>标记中重复写“汽车”这个词,以增大汽车这个词在页面中的频次,或者用<Form>表格中的<hidden>标记,大量书写“汽车”;更有甚者,干脆用不可见字体在页面中书写“汽车”。有一些网络广告公司就专门在研究如何钻搜索引擎的空子,使得他们的客户的网页在搜索引擎返回结果的排名中更靠前。如何来消除这种问题呢?

2 超链分析的用处

超链,即指向网页 B 的链接位于网页 A 之中,从其完成的功能上讲,是简单的,对信息检索也没有什么直接的用处。但是,网页作者使用超链的行为,有可能是指出更有价值的内容。作者常常创建一些对访问者有用的超链:一些超链起着导航的作用,譬如让访问者退回到主页面,另一些超链则提供访问比当前页更多内容的途径。后一种超链有可能指出与本页同主题的但质量更高的网页。WEB 信息检索系统要能够利用这种超链信息来优化对相关文献的查询。显然,超链分析能够极大地提高检索结果的相关性,以至于几乎所有的 WEB 搜索引

擎都宣称他们使用了超链分析技术。但是,很少有WEB搜索引擎暴露他们的算法,这主要是为了避免被网络广告公司这类利益追逐者钻算法上的空子。

超链分析的应用很广泛^[3]。主要用于网络爬行和检索结果排序。除此之外,网页的按例查询(QBE)——比如,给出一个网页,现在需要检索出与该网页内容相似的网页——就属于这种问题;寻找镜像站点;WEB页面分类;计算网页的地理空间或兴趣空间:比如,一个天气预告的网页,仅仅对其所覆盖的地域范围有价值,而中国政府的税收政策网页,可能对全世界欲来华投资的商人都有价值。

下面给出超链分析在网络爬行和检索结果排序中的应用。

3 超链分析在网络爬行中的应用

网络爬行就是收集WEB页面的过程。WEB信息检索不同于传统信息检索,在于这种收集不是“给”WEB搜索引擎,而是WEB搜索引擎自己去“找”。怎样才是一个好的“找”法?在保证信息收集完整的情况下,尽可能少找一些网页,即寻找精品。这既缩短了爬行时间,也减轻了服务器的存储压力。按照J.Kleinberg^[4]的说法,就是直接抓回好的权威页(authority page)。如何来找到权威页?这需要有一个原始积累和评价过程。先给出一批源网页(起始网页),让网络爬行者从这些源网页开始爬行,收集回WEB页,此时还没有权威页的概念,爬行的时候,既可按深度优先的算法,也可以按广度优先的算法^[5],当把网页爬回来后,采用如下的超链分析技术来计算出权威页。

首先,一个网页的质量是由指向它的页面的数量来决定的。简而言之,一个页面的质量由别的网页来决定,这是超链分析的核心思想。但这种思想并没有区别如下情况:由大量低质量网页所指的网页的质量与同数量高质量网页所指的网页的质量是不同的。这就有可能使一些网络广告公司为提升其客户网页的排名,而人为地制造一些网页来指向客户的网页。

其次,我们把爬行回来的所有网页看成一个有向图(Directed Graph),每个页面被看成一个顶点。如果页面A中有一超链指向页面B,就称顶点A与顶点B之间存在一条有向边(A,B)。顶点的入度(Indegree)——指向该顶点的有向边的条数——即

指向该网页的超链数量;顶点的出度(Outdegree)——离开该顶点的边的条数——即该页面中含有的超链数量。在给定的有向图G中,设R(A)为网页A的质量,可递归定义R(A):

$R(A) = \varepsilon/n + (1-\varepsilon) * \sum R(B)/\text{Outdegree}(B)$ 且 $(B,A) \in G$

其中,

ε ——常数,取0.1;

n——有向图中的顶点数量,即参与计算的网页的数量;

Outdegree(B)——离开顶点B的边的条数,即网页B中含有的超链数量;

(B,A)——存在一条由顶点B指向顶点A的有向边,即网页B中有超链指向网页A。

假设有一个有向图(图1),可计算各页面的质量如下:

$R(A) = 0.1/4 + (1-0.1) * R(D)/1$ 指向顶点A的边只有一条(D,A),且D的出度为1;

$R(B) = 0.1/4 + (1-0.1) * [R(A)/3 + R(C)/2]$;

$R(C) = 0.1/4 + (1-0.1) * R(A)/1$;

$R(D) = 0.1/4 + (1-0.1) * [R(A)/3 + R(C)/2]$;

化简:

$R(A) = 0.025 + 0.9 * R(D)$;

$R(B) = R(D)$;

$R(C) = 0.0325 + 0.27 * R(D)$;

$R(D) = R(B)$;

这是一个不定方程。我们取 $R(B) = R(D) = 1$,其含义是:网页质量最高为1,质量高低,取决于数值的大小。(也可以取 $R(B) = R(D) = 0$,其解释方法:网页质量最高为0,质量高低取决于数值的小大,即,0为最大,比0还大者,质量在下降)。可得:

$R(A) = 0.925$

$R(B) = 1$

$R(C) = 0.302$

$R(D) = 1$

所以,权威页为网页B和网页D。此递归定义算法能有效消除如下的情况:某些WEB广告商为提升其客户网页被搜索引擎收录的机会,人为编造一些网页来指向其客户的网页。有兴趣的读者可以试一下。有了权威页,网络爬行者在下次爬行时,就有针对性,能有效提高爬行效率和爬行质量。

4 超链分析在检索结果排序中的应用

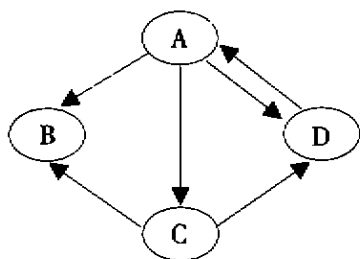


图 1 有向图

现在 WEB 搜索引擎的后台数据库容量大, 用户输入一个检索词, 如果是当成任意词检索的话, 有可能会有 1 万、10 万甚至更多的匹配结果, 假设每屏最多显示 20 条记录, 在如此庞大的检索结果中(假设为 1 万条), 先把哪 20 条记录送给用户呢? 用户理想化的认识是这样: 先从 1 万个匹配的结果中, 找出最相关的 20 条记录, 然后再对选出的 20 条记录作进一步的排序, 最后显示出来。事实上, WEB 搜索引擎为了提高响应时间, 不会这么做的^[6]。搜索引擎不会做第一步: 从 1 万条容量的匹配结果中, 选出最相关的 20 条, 是因为搜索引擎负担不起这么大的开销。它能做的就是: 按记录号的顺序提出结果集中的前 20 条记录, 再按毗连图(Neighbor Graph)算法对这 20 条记录排序, 最后显示给用户。下面是该算法:

首先, 该算法使用了超链分析技术, 算法中涉及三个概念:

一是初始集。初始集就是检索结果集(比如有 1 万条记录)中的一部分, 比如, 前 20 条记录。二是毗连集。毗连集是这样一个页面集合: 集合中的页面要么指向初始集中的页面, 要么被初始集中的页面所指。三是毗连图。毗连图由初始集和毗连集中所有页面构成, 可以把页面看成毗连图的顶点, 当页面 A 拥有一超链指向页面 B, 就称顶点 A 与顶点 B 之间存在一条有向边(Directed Edge)。毗连图的示意图请见图 2。

其次, 我们称被很多(相对而言)页面所指的页面为权威页(Authority Page); 指向很多(相对而言)页面的页面为中心页(Hub Page), 即中心页内含有很多(相对而言)超链。该算法可同时计算毗连图中各顶点的权威值(表示该页面成为权威页面的相对大小值)和中心值(表示该页面成为中心页面的相对大小值)。算法如下:

- ① 设 N 是毗连图 G 中的顶点数量;
- ② 对毗连图 G 中的任一顶点 A , 设 $Hub(A)$ 是中心值, $Aut(A)$ 为权威值;
- ③ 初始化 $Hub(A)$ 的值为 1, A 为毗连图 G 中的任意一个顶点;
- ④ 对毗连图 G 中任一顶点 $A: Au(A) = \sum Hub(B)$ 且 $(B, A) \in G$;
- ⑤ 对毗连图 G 中任一顶点 $A: Hub(A) = \sum Aut(B)$ 且 $(A, B) \in G$;
- ⑥ 标准化 Hub 和 Aut 。

设有一个毗连图(见图 2), 各页面的权威值和中心值计算如下:

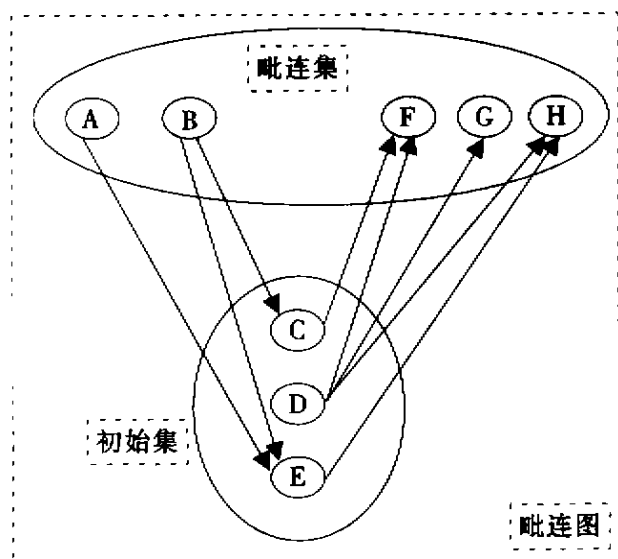


图 2 毗连图

第一步: 初始化

$$Hub(A) = Hub(B) = Hub(C) = Hub(D) = Hub(E) = Hub(F) = Hub(G) = Hub(H) = 1;$$

第二步: 计算权威值

$$Aut(A) = 0; \text{ 没有边指向 } A;$$

$$Aut(B) = 0$$

$$Aut(C) = Hub(B) = 1; \text{ 有一条边 } (B, C) \text{ 指向 } C, \text{ 且 } Hub(B) \text{ 为 } 1;$$

$$Aut(D) = 0$$

$$Aut(E) = Hub(B) + Hub(A) = 2$$

$$Aut(F) = Hub(C) + Hub(D) = 2$$

$$Aut(G) = Hub(D) = 1$$

$$Aut(H) = Hub(D) + Hub(E) = 2$$

第三步: 计算中心值

$$Hub(A) = Aut(E) = 2; \text{ 有一条边 } (A, E) \text{ 离开顶点 } A, \text{ 且 } Aut(E) = 2;$$

$$Hub(B) = Aut(E) + Aut(C) = 3$$

$$Hub(C) = Aut(F) = 2$$

$$\text{Hub}(D) = \text{Aut}(F) + \text{Aut}(G) + \text{Aut}(H) = 5$$

$$\text{Hub}(E) = \text{Aut}(H) = 2$$

$$\text{Hub}(F) = 0$$

$$\text{Hub}(G) = 0$$

$$\text{Hub}(H) = 0$$

第四步: 标准化 Hub 和 Aut

$$\text{Aut}(A,B,C,D,E,F,G,H) = (0,0,0.125,0,0.25,0.25,0.125,0.25)$$

$$\text{Hub}(A,B,C,D,E,F,G,H) = (0.14,0.21,0.14,0.36,0.14,0,0,0)$$

最后,把初始集中的页面以 Hub 和 Aut 的递减次序显示给用户。显示分成两部分:一是中心页(所谓的目录页)排序输出;二是权威页排序输出,好比是 Yahoo 网站的返回结果。所以:

页面 E 被排在权威页的第一位。

页面 D 被排在中心页的第一位。

综观上述毗连图算法,一是它减轻了搜索引擎的压力(把对 1 万个结果的操作,缩减到对毗连集中的页面进行操作。事实上,一个初始集的毗连集中,最多几百个页面。);二是它利用了超链分析技术,对显示结果进行了排序。Yahoo 网站的搜索引擎

就使用了类似与毗连图这样的算法。

参考文献

- 1 G. Salton et al. The SMART System—Experiments in Automatic Document Processing. Prentice-Hall, Engle-wood Cliffs, N.J., 1971
- 2 赖茂生. 计算机情报检索. 北京:北京大学出版社,1996
- 3 Monika R. Henzinger. Hyperlink Analysis for the Web. IEEE Internet Computing, Jan./Feb. 2001
- 4 J. Kleinberg. Authoritative Sources in a Hyperlinked Environment. Proc. Ninth ACM-SIAM Symp. Discrete Algorithms, ACM Press, New York, 1998
- 5 韩家伟. WEB 挖掘研究. 计算机研究与发展,2001(4)
- 6 S. Brin, L. Page. The Anatomy of a Large-scale Hyper-textual Web Search Engine. Proc. Seventh Int'l World Wide Web Conf., Elsevier Science, New York, 1998

作者单位:中国科学院文献情报中心,北京,100080

收稿日期:2001年11月13日

(上接第 43 页)因此,必须建立三维目标管理和三重评估激励机制。

三维目标管理要求建立采访部目标系统、采访项目组目标系统和个人目标系统。首先考察采访部整体经费执行情况和整体馆藏建设的质量,其次考察各个采访业务项目的完成质量,再次重点考察每个学科采访馆员工作任务完成的质量和数量、工作创新能力、信息导航的合作完成情况和协作态度。

三重评估激励机制是指对个体、项目组 and 整个部门三个层次进行奖赏和激励的机制。肯定每个学科采访馆员在学科专业化建设中的贡献并进行奖励,可以激发个体的积极性,使每个人为集体做出尽可能多的贡献;根据业务项目组的整体表现,对参与者进行奖励,可以激励团体积极接受新任务;对整个部门进行奖励,可以加强团体成员的自豪感和归属感。

(5)重视继续教育,树立学科采访馆员终生学习的理念。

在网络环境下,学科采访馆员面对的是庞大的传统文献资源和数字化资源的选择以及网上信息的调查、评价和获取,学科信息的无限性和专业知

识、工作技能的有限性的矛盾构成学科采访馆员挑战自我的永恒课题,只有通过继续教育和终生学习才能跟上时代发展的步伐。

学科采访馆员继续教育和终生学习重点关注五个方面:①学习信息管理学知识和加强采访综合业务技能,优化学科知识结构。②现代化信息技术及应用技能学习,转换角色到真正的网络信息选取者和知识导航者,成为采访工作技能和信息技术相结合的复合型人才。③外语训练。学科采访工作必须使用中文、英文或其他语种。④掌握现代管理理论。了解图书馆组织和管理理论的变革及其发展方向,才能提高自我管理和自我控制的能力,发挥协作精神,在提高服务质量中达到自我实现。⑤加强职业道德教育,牢牢树立敬业精神。

参考文献

- 1 王春生.网络环境下高校图书馆采访工作的几个问题.大学图书馆学报,2000(1)
 - 2 周群.文献采访工作科学管理思路.图书馆建设,2000(3)
 - 3 杨薇桐.参考馆员的建立与实施.图书馆论坛,2000(3)
- 作者单位:厦门大学图书馆,厦门,361005
收稿日期:2001年10月18日