

浅析搜索引擎技术

—Google 的 PageRank 技术剖析

李维君 (兰州商学院陇桥学院 甘肃兰州 730101)

1. PageRank 的基本原理

PageRank 的发明者对网络超链接结构和文献引文机制的相似性进行了研究,把引文分析思想借鉴到网络文档重要性的计算中来,利用网络自身的超链接结构给所有的网页确定一个重要性的等级数,当从网页 A 链接到网页 B 时,就认为网页 A 投了网页 B 一票,增加了网页 B 的重要性。最后根据网页的得票数评定其重要性,以此来帮助实现排序算法的优化,而这个重要性的量化指标就是 PageRank 值。

但是网页和学术上的出版文献的差别是很大的。首先学术论文的出版发表非常的严格,而网页的出版非常自由、成本很低并缺乏控制,用一个简单的程序就可以产生大量的网页和很多链接。另外学术出版物的引文一般和文章的领域有关系,而网页的链接范围领域却很广。可见简单的链接数量计算并不能客观真实地反映网页的重要性。所以 PageRank 除了考虑网页得票数(即链接)的纯数量之外,还要分析为其投票的网页的重要性,重要的网页所投之票有助于增强其他网页的重要性。简单地说,PageRank 就是要从链接结构中获取网页的重要性,而网页的重要性决定着同时也依赖于其他网页的重要性。

2. PageRank 的定义

根据上面的基本原理, L. Page 等给出 PageRank 的简单定义:令 u 为一个网页, $N(v)$ 表示从网页 v 向外的链接数目, $B(u)$ 表示链接到网页 u 的网页集合, $R(u)$ 表示网页 u 的 PageRank 值, C 为规范化因子,作用是保证所有网页的 PageRank 总和为常量。例如为保证总的 PageRank 值为 1,可以通过网页 PageRank 总和的倒数求得。

$$R_{(u)} = C * \sum_{v \in B(u)} R_{(v)} / N_{(v)} \quad (\text{定义 } 1)$$

必须注意的是定义 (1) 有一个假设前提,即所有的网页形成一个牢固的链接图(即每个网页能从其他网页通过超链接达到)。从定义 (1) 可以看出,网页的 PageRank 是一个由网络的超链接结构所产生的一个网页重要性等级值,所有的网页的 PageRank 值都可以根据其他网页的 PageRank 值和链接的数量来计算得到,即所有链接到它的网页的 PageRank 值除以各自向外的链接数的商进行求和。

为了更加容易地理解 PageRank 的定义,可以用非常直观的模拟冲浪模型来进行解释。假设一个网络冲浪者通过随机的点击超链接在网上冲浪,在前面的假设前提下,每个网页都是可能达到的,只不过是达到的可能性的大小不同。很显然,链接到哪个网页的超链接多,那么到达哪个网页的可能性大。这个网页就相对重要,PageRank 值也就高。而重要的网页链接到的网页,冲浪者到达的可能性当然也就大,其 PageRank 值也就相对高。同时可见重要性权值 (PageRank) 是整个网页的一个重要性概率分布结果。所以所有网页 PageRank 的总和应该是 1。

3. PageRank 的计算

前面给出的定义 1 本身是一个 PageRank 的计算公式,利用这个公式,可以计算网页集合中所有网页的 PageRank 值。假设 S 为整个网页的总和。因为所有的网页的 PageRank 值开始是未知的。所以我们进行平均的分配,给每个网页的 PageRank 都赋以 $1/S$ 。再根据公式 1 进行计算。然后对得到的值再次利用公式 1 计算。这样反复地计算。直到计算得到的 PageRank 值收敛于一个相对固定的数 (ϵ)。也就是说,根据超链接结构计算出的所有网页的重要性等级趋于稳定,这时停止计算。

$$\forall .u \in s : R_{(u)0} = 1 / |s|$$

$$\text{while} (|R_{(u)i} - R_{(u)i-1}| > \epsilon)$$

$$\{$$

for each $u \in s :$

$$R'_{(u)i} = R'_{(u)i-1} + \sum_{v \in B(u)} \frac{R_{(v)i-1}}{N_{(v)}}$$

$$C = 1 / \sum_{u \in s} R'_{(u)i}$$

for each $u \in s :$

$$R_{(u)i} = C * R'_{(u)i}$$

}

算法 1

为了方便论述,我们将网络结构简单化,假设仅仅有 5 个网页。

次数	P (1)	P (2)	P (3)	P (4)	P (5)
1	0.25	0.2	0.15	0.2	0.2
2	0.25833	0.21667	0.13333	0.19167	0.2
3	0.2625	0.22708	0.12708	0.18542	0.19792
4	0.26521	0.23417	0.12433	0.18083	0.19542
5	0.26717	0.23954	0.12316	0.17734	0.19299
6	0.26868	0.24331	0.12266	0.17459	0.19075
7	0.26989	0.24649	0.12254	0.17236	0.18873
8	0.27089	0.24909	0.12261	0.17050	0.18691
10	0.27245	0.25313	0.12306	0.16757	0.18378
12	0.27363	0.25615	0.12367	0.16535	0.18120
14	0.27456	0.25851	0.12429	0.16360	0.17903
16	0.27532	0.26043	0.12491	0.16218	0.17717
20	0.27649	0.26335	0.12603	0.15998	0.17414
24	0.27736	0.26551	0.12700	0.15836	0.17178
28	0.27803	0.26718	0.12784	0.15709	0.16986
36	0.27903	0.26962	0.12921	0.15523	0.16691
37	0.27913	0.26989	0.12936	0.15504	0.1666
38	0.27923	0.27011	0.12950	0.15486	0.16630

从上面可以看出,PageRank 的总和为 1 (因为采用舍入计算,所以没有完全精确)。其 PageRank 的分布也是完全合理的,这里网页 1 的 PageRank 值高是因为有 3 个网页链接到它,很显然在冲浪模型中冲浪者到达网页 1 的可能性大。而网页 2 有相同的 PageRank 值是因为网页 1 链接到网页 2,而且只有一个链接。在冲浪模型中,冲浪者访问了网页 1 之后肯定会访问网页 2。所有到达网页 2 的可能性和到达网页 1 的可能性是一样的。

4. PageRank 存在的问题

PageRank 对于 google 的成功,所起的作用是非常大的。而随着 google 的成功,和 PageRank 类似的超链接分析排序技术也纷纷被搜索引擎技术公司所采用。那么在这种排序算法已经成为主流的环境下,商业网站为了在搜索引擎中获得好的排名,必然会费尽心机去影响本来天然客观的网络超链接结构。被 google 所标引的网站中每个网页都有一个 PageRank,并非一个网站的 PageRank 就是网站的首页,并且站内的链接也被用来计算 PageRank。网站制作者通过站点链接和网站地图等来提高站内的 PageRank 的反馈值。那么 google 会不会对这样的站点内部的链接投票值打折呢?而这样做会不会影响这种基于链接的投票计算方法因为这本身偏偏正是 PageRank 工作的最基本的机理,更何况庞大网络链接的计算量非常的巨大。有了这些人为的影响,那么 google 所说的客观公正会不会只是 google 的一厢情愿:另外,网络开始显著地改变,现在的一个链接很多是因为应用许可的需要或像交换链接这样的利益交换,实际上并无推荐之意。所以有人认为,PageRank 在 Google 排序算法中的作用已经下降。但是有一点必须指出,对巨大的网络超链接结构人为的影响,其难度和 webSpamming 相比,是不可同日而语的。

从目前情况看,多媒体在课堂教学中的应用还处于初级阶段,实践和理论上的一些问题尚需探讨和解决。

3.1 应当认清多媒体在课堂教学中的地位,注意使用方式,处理好多媒体呈现与整个教学流程的关系。多媒体尽管具有先进性、科技性的特点,但它毕竟是作为一种教学手段进入课堂的,它的使用必须有助于教学目标的实现。如果通过多媒体补充的课外知识过多,就可能喧宾夺主,导致大学语文课的异化和本学科知识的弱化,冲淡甚至淹没了教学主题。而多媒体也不是万能的,它不能代替教师的讲解和分析,只能帮助教师更好地表达思想和观点。因此,多媒体的使用应当与教学内容相契合并注意使用的方式,而且要把握好多媒体的使用量和呈现时机。大学语文课堂教学特别强调教学内容的动态性和生成性,作为课堂教学的组织者和引导者的教师,要善于分析具体的教学情景,选择最有利于学生掌握学习内容或使训练效果达到最优化的时机呈现多媒体,使多媒体的呈现成为整个教学流程中的一个有机组成部分。那种追求时尚,为形式而形式,形式偏离内容,不设置问题,不分析问题,不引发讨论,满堂放音影制品的做法,显然是有悖于多媒体使用原则的。

3.2 教师在课堂教学中的主导地位不能也不应让位于多媒体。课堂教学中应当处理好多媒体使用与教师活动、学生活动的关系。一堂课中,教师和学生之间的交流、学生之间的交流应该是课堂构成的主要部分,教师是主导,学生是主体,教师在教学活动中应时时处于主导地位,但其主要任务是激发学生的学习兴趣,帮助学生形成学习动机,创设符合教学内容需要的多种情境和提示新旧知识之间的联系,帮助学生进行知识的迁移,组织“协作学习”,学生的“学习——质疑探究——讨论释疑——迁移拓展”应是课堂教学的中心环节,而多媒体的使用只是起一个辅助的作用,只能视教学流程的需要择而用之,不能成为教学的唯一手段,更不能成为课堂的主宰。教师不能成为电脑操作员和投影解说员,把教学的主动权让位于多媒体。

(上接第 200 页)

牵引变压器高压侧 IA 电流模值最大, Ia 由装置的平衡关系式从低压侧向高压侧平衡后的 ICDA、ICDB、ICDC 都近似等于 0,且 IZD 远大于 ICD,这说明此变电所的差动保护接线正确,且整个系统(包括主变压器微机保护装置)一切运行正常。

3. 结论

差动保护是牵引变压器的主要保护之一,其目的就是要保护高压电气设备,在牵引变压器内部发生短路故障时,差动保护必须无延时可靠

(上接第 215 页)

现。连接关系单元含有多个入口和一个出口,通过初始化过程完成逻辑关系的设定。例如图 1 中控制单元 B 和 D 的连接关系初始化为:

B. SetParent (OP_RELATION_AND, &A, TRUE);

D. SetParent (OP_RELATION_OR, &B, &C);

其中,OP_RELATION_OR 代表连接关系为或关系,OP_RELATION_AND 代表连接关系为与关系,&A、&B 和 &C 分别是上级控制单元的指针。每个控制单元可以通过指针实现对上级控制单元的访问,判断其当前状态,以决定自身操作或上级控制单元状态改变时自身状态。在大型仿真系统中,&B 和 &C 等也可以用标识号代替,通过全局注册表实现访问。

3. 成绩评估

成绩评估的核心是判断操作目标是否实现,操作目标一般是控制单元状态的组合,因此在系统实现上可以采用虚拟控制单元表示操作目标,并按照操作目标之间的顺序关系形成操作目标表,评估成绩时只要按照顺序经过此表,即可完成评估。

成绩评估算法为: 1) 确定当前操作目标; 2) 获取当前控制单元状态; 3) 判断这些控制单元的状态组合; 4) 根据组合结果判断当前操作目标是否完成; 5) 给出错误提示; 6) 记录成绩; 7) 确定下一操作目标; 8) 重复 a-g, 直到操作结束。

3.3 教师应当提高自身的业务素质,提高课件的制作水平。多媒体进入大学课堂,对教师的业务素质提出了更高的要求。教师不仅应当熟练掌握本专业专业知识,还应了解与本专业相关的边缘性知识,更要努力学习计算机应用方面的知识。多媒体在教学中的应用,目前的关键是要制作出适用于本学科的多媒体课件,它的制作需要在教学理论的指导下,搜集大量素材,做好教学设计,将教材、学生等因素综合考虑进去,并在教学实践中反复修改,这样才能使制作出的课件符合教学规律,取得良好的使用效果。

21 世纪的人类社会是信息化社会。随着多媒体技术的日益成熟,它在教育中的应用也越来越普遍。多媒体计算机辅助教学是当前国内外教育技术发展的大趋势。我们应当把握时代的脉搏,不断拓展知识领域,以适应现代化社会对教育工作者的多元化需求。

参考文献:

[1] 傅苏黎,杨万利主编.《基础课程现代化教学改革研究》,北京文化艺术出版社 2002 年 2 月第 1 版。
 [2] Edward Sallis 著.《全面质量管理》,何瑞译,华东师范大学出版社 2005 年 6 月第 1 版。
 [3] 查有梁著.《新教学模式之建构》,广西教育出版社 2003 年 5 月第 1 版。
 [4] 周军著.《教学策略》,北京教育科学出版社 2003 年 12 月第 1 版。

王霁娟:(1971—)女,辽宁大连人,辽宁对外经贸学院教师,副教授
 研究方向:中文学科教育

动作,切除故障点。通过武广线荣家湾牵引变电所等三牵引所的现场运行实践证明,对南京电力自动化总厂的 WBZ-61A 主变压器微机保护装置的差动保护外围接线原则为:平衡牵引变压器高、低压侧差动臂始终保持相同的相位,且数值上成一定比例关系。

参考文献:

《电气化铁道施工手册牵引变电所》中国铁道出版社。

4. 加权与否定

在上述成绩评估算法中,对于错误操作的评估没有严格的级别区分。而实际设备工作时,情况却不是这样。一些无关操作可能没有太大的影响,而另外一些错误操作不仅可能使机组状态失去控制,严重时还可能导致事故的发生。这种区别对于评价操作人员的操作熟练程度非常重要,因此在评判系统中必须针对不同情况进行处理。对不同的错误操作进行等级划分,在统计错误时进行加权处理,对致命错误实行否定制。

5. 结论

某型雷达操作仿真训练系统中的应用,证明上述成绩评估算法具有逻辑结构清晰、容易实现、评估灵活和可靠性高等特点,且对机械和电器控制系统可以同等对待,适应性较强。同时,由于软件采用 C++Builder 面向对象的开发环境,系统的软件复用率和可靠性都很高,可维护性也较强,很适合在其他操作仿真系统中使用。

参考文献:

[1] 谢明与编著. BorlandC++编程实例剖析. 北京:科学出版社,1993.
 [2] 姚毅军等编著. 防空兵器与操作教程. 北京:解放军出版社,1999.
 [3] 余明兴等编著. BorlandC++Builder6 程序设计经典. 北京:科学出版社,2004.