

文章编号:1001-9081(2015)10-2733-04

doi:10.11772/j.issn.1001-9081.2015.10.2733

基于稀疏矩阵面向论文索引排名的启发式算法

万晓松, 王志海^{*}, 原继东

(北京交通大学 计算机与信息技术学院, 北京 100044)

(*通信作者电子邮箱 zhhwang@bjtu.edu.cn)

摘要:为了提高学术论文检索的精准性,进而为学术研究提供便利,提出了针对学术论文检索问题的排名策略。首先,介绍了基于网页排名算法面向论文索引排名的启发式方法,其中利用 Hash 索引技术有效地减少了稀疏矩阵计算对内存的消耗;其次,定义了论文间引用关系图的密集度均衡值,并通过大量实验阐明了不同排名算法的迭代次数与图密集度均衡值之间的关系;最后,将所提出的基于论文索引排名的启发式算法应用于科学引文索引(SCI)数据库中,并与原被引频次降序的排序结果进行比较与分析。实验结果表明:在三种基于网页排名技术的算法中,基于链接结构分析的随机过程算法比较适合于按关键词搜索得到的相关领域学术论文的排名。

关键词:网页排名算法;稀疏矩阵;Hash 索引;论文索引排名;SCI 数据库

中图分类号: TP391.4 **文献标志码:**A

Heuristic algorithms for paper index ranking based on sparse matrix

WAN Xiaosong, WANG Zhihai^{*}, YUAN Jidong

(School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China)

Abstract: In order to enhance the accuracy of retrieved academic papers, so as to facilitate academic research extensively, a series of ranking strategies for academic paper retrieval problem were proposed. Firstly, the heuristic methods based on page ranking algorithm for paper index ranking were described, taking advantage of a Hash indexing technique to effectively reduce memory consumption of the sparse matrix computation. Secondly, the definition of intensive equilibrium value of reference relationship among papers was presented, at the same time, the correlation between iterations of different ranking algorithms and intensive equilibrium value was clarified by a large number of experiments. Finally, the proposed heuristic algorithms for paper index ranking were tested on the SCI index database, and compared with the classical citation descending sort results. The experimental results show that, in the proposed three kind of algorithms based on page ranking techniques, the stochastic process approach with link-structure analysis is much more suitable for the ranking of papers, which obtained by the searching results according to keywords in a certain field.

Key words: page ranking algorithm; sparse matrix; Hash index; paper index ranking; SCI index database

0 引言

搜索引擎的核心技术是网页排名算法^[1-2]。最简单直接的网页排名方式是按照人工设定的目录结构进行搜索,典型代表是 Yahoo 文本搜索算法^[3];以 PageRank^[4]和中枢-权威(Hyperlink-Induced Topic Search, HITS)^[5]为代表的网页排名算法是基于链接分析的搜索排名算法,典型的代表是 Google 搜索算法。网页排名算法可应用在许多不同场景^[6-7],针对推荐系统的排名算法同样也可有很多应用场景^[8]。

目前科学引文索引(Science Citation Index, SCI)数据库论文检索的排序方式共有 13 种,如图 1 所示。对于某个研究领域的新手来说,“出版日期”为近期的论文,其价值难以准确估计,也许不够权威;“被引频次”较高的论文所提出的方法可能已经有了实质性的扩展或改进;而目前的“相关性”排序方式仅考虑了关键词的语义关联。并且,所有的这些排序方式都是将检索到的论文看作是一个孤立的检索结果,而没有在检索过程中体现出这些结果间的相互关系,这正是本文

研究的主要动因。



图 1 SCI 数据库论文排序方式

本文提出了面向论文索引排名的启发式算法,为 SCI 数据库论文检索提供有效排序方式。

1 理论基础与研究背景

网页排名算法的改进包括:基于主题敏感的研究^[9-10]、基于主题漂移的研究^[11]、基于自适应方法的研究^[12]、基于网页缩放更新规则的研究^[13-14]、基于无超链接情况下的研

收稿日期:2015-06-01;修回日期:2015-07-01。 基金项目:国家自然科学基金资助项目(61370130)。

作者简介:万晓松(1991-),女,吉林通化人,硕士研究生,主要研究方向:数据挖掘、机器学习; 王志海(1963-),男,河南安阳人,教授,博士,CCF 会员,主要研究方向:数据挖掘、机器学习; 原继东(1989-),男,河南焦作人,博士研究生,主要研究方向:数据流挖掘、模式识别。

究^[15]等。本文所关注的排名算法均是假定以网页之间的链接关系为基础的,经典的算法有 PageRank 算法^[2]、HITS 算法^[5]以及 SALSA (Stochastic Approach for Link-Structure Analysis) 算法^[16]等。这些算法都是基于网络链接的结构关系,而与网页内的文本内容无关^[17]。

1.1 基本 PageRank 链接分析算法

PageRank 算法用于衡量特定网页在搜索引擎中相对于其他网页而言的重要程度^[2, 4]。算法中定义邻接矩阵为 N ,其中 N_{ij} 为节点 v_i 在一次更新中传递给节点 v_j 的网页排名份额,如果 v_i 没有到 v_j 的链接,则 $N_{ij} = 0, N_{ii} = 1$;否则, N_{ij} 为 v_i 出度数的倒数。以向量 r 表示所有节点的网页排名值,基本 PageRank 算法网页排名的更新规则表示为:

$$r \leftarrow N^T r \quad (1)$$

注意:反复运行更新规则后,可能排名泄露。在大范围的网络中,选择一个缩放因子 s ,限定在 $0 \sim 1$ 。因此,缩放网页排名算法 (PageRankScale) 的更新规则定义 $\tilde{N}_{ij} = sN_{ij} + (1 - s)/n$ 。当反复执行网页排名更新规则 n 次时,

$$r^{<k>} = (\tilde{N}^T)^k r^{<0>} \quad (2)$$

所有节点的网页排名值收敛于相应的极限值,与初始值无关,只与网络结构有关,即 $r^{<*>} = \tilde{N}^T r^{<*>}$,其中 $r^{<*>}$ 是 \tilde{N}^T 的特征向量,且对应的特征值是 1。用 Perron 定理^[2]可证明,当 \tilde{N}^T 是正数矩阵时,反复运用更新规则后, $r^{<*>}$ 收敛,这个向量就是要寻找的网页排名的收敛极限值。

1.2 中枢-权威 (HITS) 链接分析算法

HITS 是由 Kleinberg 在 20 世纪 90 年代末提出的基于链接分析的网页排名算法^[5]。该算法描述了两种类型的网页,“权威型 (Authority) 网页”与“中枢目录型 (Hub) 网页”。设有向图 $D = \langle V, E \rangle$, $V = \{v_1, v_2, \dots, v_n\}$, M_{ij} 为节点 v_i 邻接到节点 v_j 的边数^[19],当 $M_{ij} = 1$ 时代表有链接指向, $M_{ij} = 0$ 时代表无链接指向,这样就形成了一个链接关系的邻接矩阵 $M(D)$ 。这个图一般情况下是非强连通图^[18]。每个网页节点都有中枢值和权威值两个属性。以向量 h 表示所有节点的中枢值,向量 a 表示所有节点的权威值,中枢更新规则可以表示为:

$$h \leftarrow Ma \quad (3)$$

权威更新规则可以表示为:

$$a \leftarrow M^T h \quad (4)$$

注意,权威值代表了网页内容的真正价值,所以在查询关键词后,是用其权威值进行排名。在邻接矩阵 $M(D)$ 中,反复执行更新规则 n 次后:

$$a^{<k>} = (M^T M)^{k-1} M^T h^{<0>} \quad (5)$$

$$h^{<k>} = (MM^T)^k h^{<0>} \quad (6)$$

中枢和权威值更新后会增大,只有对其归一化处理,才能收敛。当 k 趋近于无穷大时,存在常数 c ,向量序列 $h^{<k>}/c^k$ 收敛于极限向量为 MM^T 的特征向量。

1.3 SALSA 链接分析算法

SALSA 算法的初衷希望能够结合 PageRank 算法和 HITS 算法两者的主要特点,既可以利用 HITS 算法与查询相关的特点,也可以采纳 PageRank 的“随机游走模型”^[10]。PageRank 算法与 HITS 算法均利用了特征向量作为收敛性依据,在文献

[19] 中详细叙述了它们的不同点。实际应用中,用户大多数情况下是向前浏览网页,但是也会回退浏览网页。基于上述直觉知识,Lempel 和 Moran 提出了 SALSA 算法^[16],具体计算迭代细节在文献[10, 16]中有明确说明。

本文研究的问题是论文排名。由于论文间的引用关系与网页间的链接关系相似,所以利用网页排名算法的技术手段,来解决所关注的问题。注意:由于论文发表存在先后关系,因此,论文引用关系形成的图是拓扑图,这点与网页链接关系图不同,这就导致了排名值传递有极大的方向性。

2 基于稀疏矩阵与 Hash 索引技术的排名算法

本章将利用 Hash 索引技术减少稀疏矩阵对内存的消耗,同时引入一种启发式的策略来定义图密集程度均衡值。

2.1 链接关系的稀疏矩阵表示

计算机算法永远是在时间与空间中权衡与妥协。第 1 章介绍了邻接矩阵表示链接结构的方法。但在实际中,一篇论文的引用论文数量是有限的,搜索得到的论文集所生成的邻接矩阵极大且稀疏。如果不利用 Hash 索引技术,在 2 GB 的内存主机中,本文介绍的排名算法是处理不了表 1 后 7 行实验数据的。本节介绍的 Hash 索引技术表示稀疏矩阵的方法,就是用来缓解内存空间有限性所带来的问题。

本文实验所用到的表示论文之间引用关系的数据格式如图 2 所示。加载进 HashMap 的数据结构形式如图 3 所示,不记录论文之间没有引用的 0 项。在搜索 HashMap 数据结构的内容时,可以用迭代器来进行搜索。所以在每次更新某一篇论文的排名值时,只需要计算引用(映射)到这篇论文的那些论文的 HashMap 结构的内容就可以了。

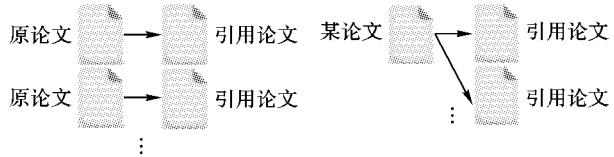


图 2 论文间引用关系数据格式

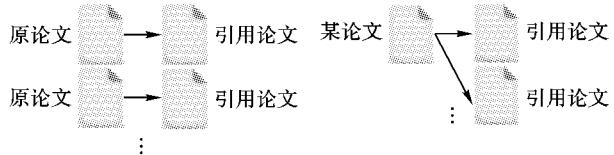


图 3 HashMap 数据结构

注意,1.1 节引入缩放因子所改进的缩放网页排名算法 (PageRankScale) 不能利用 HashMap 思想,因为它将所有 $N_{ij} = 0$ 的项都变为很小的值,以防止排名泄漏的情况。

2.2 启发式策略

论文引用关系形成的图为有向图,在有向图中,设计启发式策略的主要定义如下。

定义 1 图的密集程度定义为 $D(\text{density}) : D = L/N$ 。其中: L 为论文间的引用链接数, N 为总的论文节点数。 D 越大,则图越密集;反之越稀疏。

定义 2 图中有出度的节点的占比定义为 $ODR(\text{Out Degree Ratio}) : ODR = ODN/N$ 。其中: ODN (Out Degree Node) 为有出度的论文节点数, N 为总的论文节点数。

定义 3 图密集程度均衡值 $IEV(\text{Intensive Equilibrium Value}) : IEV = D \times ODR$ 。其中: D 为图的密集程度, ODR 为图中有出度的节点占比。在论文引用关系图中,有些节点只有出度或只有入度,是图当中较为“劣质”的节点,没有紧密地连接到图中其他的节点。没有出度的节点会可能造成排名泄露。

的情况,排名算法的迭代次数将会有所变化。基于以上原因,在计算 IEV 时,考虑到有出度节点的占比,可以较全面地把握图的整体属性,并保证 IEV 作为一个表示图性质的比值,单位为 1。本文通过实验分析不同算法的迭代次数与所定义的 IEV 之间的关系。

2.3 排名算法描述

算法 1 是论文排名算法的入口,可以在四种排名算法中选择,得出相应算法的迭代次数以及对应算法每篇论文重新排名的次序以及排名值。在 SALSA 排名算法中,本文只进行一次前向与后退更新,所以迭代次数始终为 2。算法 2 是 PageRank 算法排名过程。注意并不排序所有的论文,只是排序所关注的被引频次降序排序方式的检索得到的前 1 000 篇论文。算法 3 是 HITS 算法实现一次更新的伪代码,主要进行每篇论文中枢排名值(authority)和权威排名值(hub)的一次更新,在执行 HITS 算法达到收敛后,综述类文章的中枢排名值将会排名较高,因为综述类文章相当于枢纽会被较多论文引用。排名算法判断迭代收敛的依据是所有每篇论文的排名值一次更新前后相差均少于 $10E - 6$,即精度为 $10E - 6$,如 HITSOnceUpdate()方法。注意:在 HITS 算法中以每篇论文的权威排名值收敛作为迭代结束标志。本节只介绍算法中比较有代表性的方法。另外本文介绍的 SALSA 没有极限值的概念,即不存在算法迭代次数的讨论,SALSA 算法的实现要注意 mapData 集的双向索引,具体实现方法请参照文献[20]中的内容。

算法 1 PapersRank()。

输入 论文引用关系的数据 HashMap mapData; 论文总数 int numPapers。

int iterations = 0; //记录排名算法的迭代次数

switch(判断选择的算法)

case "PageRank";

iterations = PageRank(mapData, numPapers);

展示论文重新排名结果; //选择一种排名算法

end;

算法 2 PageRank()。

输入 HashMap mapData, numPapers。

输出 iterations。

iterations = 0;

for(每一篇论文)

原始排名值 $\leftarrow 1/numPapers$;

更新排名值 $\leftarrow 1/numPapers$;

end;

while(true)

iterations += 1;

if(排名值不收敛, 即所有论文排名值更新前后相差超过 $10E - 6$)进行式(1)的一次排名值更新;

end 对收敛的排名值进行更新;

return iterations;

算法 3 HITSONCEUpdate()。

输入 HashMap mapData, Rerank originalHITS [], Rerank updateHITS [], numPapers。

输出 更新后排名值 updateHITS []。

for(每一篇论文)

updateHITS[i].hub = updateHITS[i].Authority = 0;

```

if ( mapData. getByIndex( i ) == null )
    updateHITS[ i ] = updateHITS[ i ] + originalHITS[ i ];
    while( 迭代 mapData )
        j = 找到 mapData. next( ). leftIndex 左键;
        if( 对应的键值为 1, 即 mapData. getBy Index( j ). ( i ) == 1 )
            进行式(3)与式(4)的更新
        end
    end
    updateHITS[ ]. normalize();

```

3 实验及其结果分析

实验代码可从文献[20]中获得。

3.1 迭代次数与 IEV 的关系

表 1 数据集及每种算法在其上方的迭代次数

数据集	链接数	节点数	IEV	A1	A2	A3
Dataset1	13	8	1.625	85	8	32
Dataset2	7	6	0.973	35	27	21
Dataset3	6	4	1.500	30	50	19
SogouT_Link	9 884	5 604	0.060	2	6	空
Link_First	3 713	1 726	0.121	3	5	空
Link_Second	6 171	3 889	0.039	2	3	空
Link_Odd	4 942	3 328	0.084	3	5	空
Link_Over	4 942	3 308	0.086	2	3	空
recommended_system	46 763	43 770	1.046	11	5	空
data_mining	17 524	16 262	1.056	11	5	空

注:A1 为 PageRank, A2 为 HITS, A3 为 PageRankScale。
PageRankScale 算法选择的缩放因子 $s=0.8$ 。

Dataset1 ~ Dataset3 是为了描述本文实验所构造的人工数据集。SogouT_Link 为搜狗实验室提供的数据集,Link_First 和 Link_Second 为 SogouT_Link 进行的前后部分随机拆分形成的数据集,Link_Odd-Over 为进行奇偶行拆分形成的数据集,这 4 个数据集增加了数据随机性。recommended_system 为在 SCI 数据库输入“recommended system”后获得的论文引用关系数据集;同理获得“data_mining”。

从图 4 可看出,在小数据集上(Dataset1 ~ Dataset3)呈现的规律是:随着图密集程度均衡值的增加,3 种排名算法的迭代次数都相对增加。在大数据集上 PageRank 与 HITS 也基本呈现这样的规律。另外,在引用链接数与节点数较多的图中,PageRank 与 HITS 的迭代次数都会很小,即在大规模数据的论文的排名中,迭代次数并不高。因为排名算法的迭代次数与有向图的结构有关,所以如果其他有向图与本文分析的链接图同构,则有相同的迭代次数等性质。

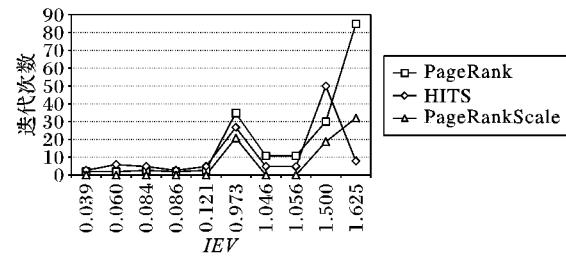


图 4 迭代次数与图密集程度均衡值之间的关系

3.2 基于 SCI 数据库的数据集

以图 2 介绍的论文引用形式作为实验数据的格式,采用

唯一表示一篇文献的方法是:用该论文的题目,去掉非英文字母以外的所有字符,并且统一转换为小写的形式。利用正则表达式来写网络爬虫可以非常容易得到想要的网页信息^[21],获得的文件格式为:

论文名 论文被引论文信息的一条链接 URL 被引用频次

本实验先暂定第 1 层搜索的论文数为 1000;进入某篇论文的被引论文集后,相当于第 2 层,每篇论文爬取的被引用论文数为 50。第 1 层的 1000 篇论文中有些被引频次不足 50,所以本应得到的连接数为 $1000 \times 50 = 50000$ 个链接,实验只得到了 46763 个链接。

3.3 参考文献排名的实验结果与分析

图 5~6 是应用了 PageRank 排名算法,对“recommended system”集文献按照原排序方式(被引频次降序)的前 1000 文章重新排名的结果。图 5、图 7、图 8 的横坐标为某篇论文在原排序方式(被引频次降序)中的排名次序(OriginalRank),纵坐标为运行不同排名方法(分别为 PageRank、HITS、SALSA)后论文的排名值(RankValue)。以图 5 为例分析,这 1000 篇论文的排名值整体呈缓慢的下降趋势,这说明在原排序方式中排名靠后的论文,PageRank 排名值也较低。同时,这 1000 篇论文的排名值都在 0.001 附近,说明论文引用关系图中的排名值基本集中在这 1000 篇论文中,流动到其他 42770 篇论文的排名值较少,这是因为在论文引用链接关系图中,这 1000 篇论文多以入度节点的形式存在,而不一定有出度,所以排名值会慢慢流向这 1000 篇论文。在论文引用链接关系图中,很多论文处于图中的相同的地位,即出度结构与入度结构相同,所以这些论文的排名值是相同的。图 6 的横坐标仍为排名次序,纵坐标为不同排名方式重新排名次序(Rank),即在原排序方式中排名第一的论文,运用了 PageRank 方式的论文排名算法后,在这 1000 篇论文中排名第 300 名。图 5 中,在 PageRank 方式中排名值最高的论文名转化处理后的形式是“theinternationalassociationforthestudyoflungcancerinternationalstagingprojectonlungcancer”,排名值为 0.00684258624628742。由图 6 观察得到论文重新排名的趋势与原排序方式拟合,呈上升趋势,这与图 5 同样证明了这个理论。但图 6 中 PageRank 方式走势高低差距较大,说明了 PageRank 方式的论文排名结果与原排名结果还是有一定差距的。注意:在 HITS 方式(见图 7)排名算法分析中,一篇论文有权威排名值也有中枢排名值,本文均以这篇论文的权威(authority)排名值(RankValue)进行重排名,这是因为论文的权威排名值才真实反映其在关键领域具体技术方面的价值,所以图 7 纵坐标为权威排名值。

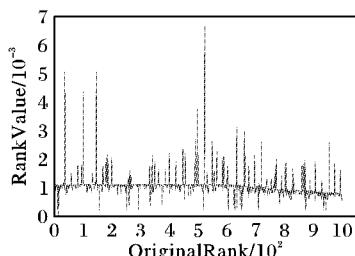


图 5 PageRank 算法论文排名值

在这三种论文排名算法的比较中,就论文排名值的比较而言,HITS 方式(图 7)论文之间的排名值差距较大,PageRank 方式(图 5)次之,SALSA(图 8)最小,SALSA 方式排名值平稳过渡,与原被引频次降序排列比较拟合。就论文重排名次序比较而言,PageRank 方式论文的排名次序与原排序方式拟合程度较差,HITS 方式较好,SALSA 方式最好。由于 SALSA 方式的论文排名算法,不更新到极限值,运行时间也较短,所以 SALSA 方式的论文排名算法比较适合文后参考文献的排名。

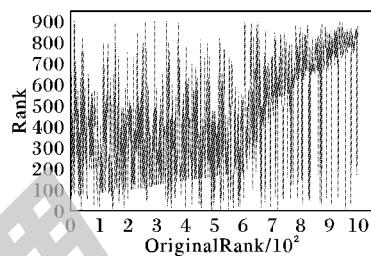


图 6 PageRank 算法论文排名次序

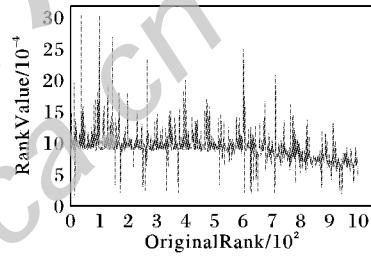


图 7 HITS 算法论文排名值

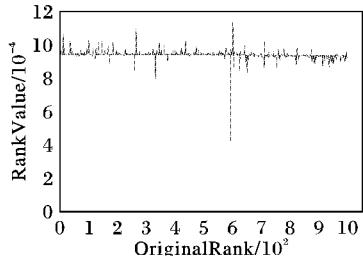


图 8 SALSA 算法论文排名值

4 结语

本文基于经典的网页排名算法,研究了 SCI 数据库论文检索的有效方式,主要考虑到检索到的论文之间的相关性。介绍了利用 Hash 索引技术减少排名算法对内存的消耗,并研究了排名算法的迭代次数与 IEV 的关系。在实验部分,分析不同的论文索引排名的启发式算法对实验数据的排名结果的影响,并给出评价。最后关于如何唯一标识一个作者及确定这个作者写过某篇文章需要更多的研究。

参考文献:

- PAGE L, BRIN S, MOTWANI R, et al. The PageRank citation ranking: bringing order to the Web[EB/OL]. [2014-10-10]. <http://ilpubs.stanford.edu/422/1/1999-66.pdf>.
- LANGVILLE A N, MEYER C D. Google's PageRank and beyond: the science of search engine rankings[M]. Princeton: Princeton University Press, 2011:1~4.

(下转第 2741 页)

研究的可能方向有:一是量化分析高层架构的结构特征;二是分析高层架构中高层结构元素,如社群和中心节点的演化情况;三是利用高层架构解决网络分析应用问题,如在线销售客户群管理和定向广告发布等问题。

参考文献:

- [1] NEWMAN M E J. Networks: an introduction [M]. Oxford: Oxford University Press, 2010.
- [2] BONCHI F, CASTILLO C, GIONIS A, et al. Social network analysis and mining for business applications [J]. ACM Transactions on Intelligent Systems and Technology, 2011, 2(3) 22: 1–22: 37.
- [3] ZHU Z. Discovering the influential users oriented to viral marketing based on online social networks [J]. Physica A: Statistical Mechanics and its Applications, 2013, 392(16): 3459–3469.
- [4] LI Y-M, SHIU Y-L. A diffusion mechanism for social advertising over microblogs [J]. Decision Support Systems, 2012, 54(1): 9–22.
- [5] RAO W, CHEN L, BARTOLINI I. Ranked content advertising in online social networks [J]. World Wide Web-internet & Web Information Systems, 2014, 18(3): 1–19.
- [6] HASSAN N R. Using social network analysis to measure IT-enabled business process performance [J]. Information Systems Management, 2009, 26(1): 61–76.
- [7] WU Y, LIU S, YAN K, et al. OpinionFlow: visual analysis of opinion diffusion on social media [J]. IEEE Transactions on Visualization and Computer Graphics, 2014, 20(12): 1763–1772.
- [8] ALBERT R A L, BARABASI A L. Statistical mechanics of complex networks [J]. Reviews of Modern Physics, 2002, 74(1): 47–97.
- [9] BARABASI A L, BONABEAU E. Scale-free networks [J]. Scientific American, 2003, 288(5): 60–69.
- [10] FORTUNATO S, LATORA V, MARCHIORI M. Method to find community structures based on information centrality [J]. Physical Review E: Statistical, Nonlinear, and Soft Matter Physics, 2004, 70(5): 056104.
- [11] TANG L, WANG X, LIU H. Community detection via heterogeneous interaction analysis [J]. Knowledge Discovery and Data Mining, 2012, 25(1): 1–33.
- [12] FORTUNATO S. Community detection in graphs [J]. Physics Reports, 2010, 486(3/4/5): 75–174.
- [13] ROSVALL M, BERGSTROM C T. Maps of random walks on complex networks reveal community structure [J]. Proceedings of the National Academy of Sciences of the United States of America, 2008, 105(4): 1118–1123.
- [14] CHEN W, WANG C, WANG Y. Scalable influence maximization for prevalent viral marketing in large scale social networks [C]// Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2010: 1029–1038.
- [15] LIN S, CHALUPSKY H. Discovering and explaining abnormal nodes in semantic graphs [J]. IEEE Transactions on Knowledge & Data Engineering, 2008, 20(8): 1039–1052.

(上接第 2736 页)

- [3] QIU Z, FU T, WANG X. Develop its own search engine [M]. 2nd ed. Beijing: People's Posts and Telecommunications Press, 2010: 4–6. (邱哲, 符滔滔, 王学松. 开发自己的搜索引擎 [M]. 2 版. 北京: 人民邮电出版社, 2010: 4–6.)
- [4] BRIN S, PAGE L. The anatomy of a large-scale hypertextual Web search engine [J]. Computer Networks and ISDN Systems, 1998, 30(1): 107–117.
- [5] KLEINBERG J M. Authoritative sources in a hyperlinked environment [J]. Journal of the ACM, 1999, 46(5): 604–632.
- [6] MIHALCEA R, TARAU P, FIGA E. PageRank on semantic networks, with application to word sense disambiguation [C]// COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics. New York: ACM Press, 2004: 1126.
- [7] ESULI A, SEBASTIANI F. Pageranking WordNet synsets: an application to opinion mining [EB/OL]. [2014-10-10]. <http://www.cl.uni-heidelberg.de/courses/ws10/graphs/elefterios.pdf>.
- [8] RICCI F, ROKACH L, SHAPIRA B, et al. Recommender systems handbook [M]. Berlin: Springer-Verlag, 2011: 1–10.
- [9] HAVELIWALA T H. Topic-sensitive PageRank [C]// Proceedings of the 11th International Conference on World Wide Web. New York: ACM Press, 2002: 517–526.
- [10] ZHANG J. It is search engine: detailed core technology [M]. Beijing: Publishing House of Electronics Industry, 2012: 146–162. (张俊林. 这就是搜索引擎: 核心技术详解 [M]. 北京: 电子工业出版社, 2012: 146–162.)
- [11] HUANG D, QI H. Pagerank algorithm research [J]. Computer Engineering, 2006, 32(4): 145–146.
- [12] KAMVAR S, HAVELIWALA T, GOLUB G. Adaptive methods for the computation of PageRank [EB/OL]. [2014-10-10]. <http://ilpubs.stanford.edu:8090/774/1/2003-26.pdf>.
- [13] FRANCESCHET M. PageRank: standing on the shoulders of giants [J]. Communications of the ACM, 2011, 54(6): 92–101.
- [14] EASLEY D, KLEINBERG J. Networks crowds and markets: reasoning about a highly connected world [M]. Cambridge: Cambridge University Press, 2010: 397–417.
- [15] KURLAND O, LEE L. PageRank without hyperlinks: Structural reranking using links induced by language models [J]. ACM Transactions on Information Systems, 2010, 26(4): 18.
- [16] LEMPEL R, MORAN S. The Stochastic Approach for Link-Structure Analysis (SALSA) and the TKC effect [J]. Computer Networks, 2000, 31(1): 387–401.
- [17] BERKHIN P. A survey on PageRank computing [J]. Internet Mathematics, 2005, 2(1): 73–120.
- [18] GENG S, QU W, ZHANG L. Discrete mathematics [M]. Beijing: Tsinghua University Press, 2004: 162. (耿素云, 屈婉玲, 张立昂. 离散数学 [M]. 北京: 清华大学出版社, 2004: 162.)
- [19] HE X, WU Q, WU Z. Comparative analysis of HITS algorithm PageRank algorithm [J]. Journal of Information, 2004, 23(2): 85–86. (何晓阳, 吴强, 吴治蓉. HITS 算法与 PageRank 算法比较分析 [J]. 情报杂志, 2004, 23(2): 85–86.)
- [20] WAN X. PageRank algorithm and its application [EB/OL]. [2015-02-06]. <http://download.csdn.net/detail/waxdhgj/8428995>. (万晓松. 网页排名算法及其应用 [EB/OL]. [2015-02-06]. <http://download.csdn.net/detail/waxdhgj/8428995>.)
- [21] LUO G. Technical combat secret search engine: Lucene & Java essentials [M]. Beijing: Publishing House of Electronics Industry, 2011: 33–58. (罗刚. 揭秘搜索引擎的技术实战: Lucene&Java 精华版 [M]. 北京: 电子工业出版社, 2011: 33–58.)