一种降低微博僵尸粉影响的方法*

原福永[1]　冯　静[1]　符茜茜[1,2]　曹旭峰[3]
[1](燕山大学信息科学与工程学院　秦皇岛 066004)
[2](秦皇岛职业技术学院经济系　秦皇岛 066100)
[3](潞安集团高河能源有限公司　长治 046000)

【摘要】以新浪微博平台为研究对象,针对微博平台存在的虚假粉丝——僵尸粉问题进行分析,从僵尸粉的定义、发展和目前采取的措施进行研究。根据微博用户存在的形式和用户间关系的特征,从链接分析的角度提出用户被关注度的概念及计算方法。实验通过对用户被关注度、用户人气值和用户影响力进行比较和分析,证明用户被关注度可以有效地降低僵尸粉带来的虚假粉丝问题。

【关键词】新浪微博　僵尸粉　虚假粉丝　用户被关注度
【分类号】TP302

A Method to Reduce the Impact of Zombie Fans in Micro – blog

Yuan Fuyong[1]　Feng Jing[1]　Fu Qianqian[1,2]　Cao Xufeng[3]
[1](College of Information Science and Engineering, Yanshan University, Qinhuangdao 066004, China)
[2](Economics Department, Qinhuangdao Institute of Technology, Qinhuangdao 066100, China)
[3](Gaohe Energy Co., LTD, Luan Group, Changzhi 046000, China)

【Abstract】The paper takes Sina micro – blog platform as the research object, and analyzes the fake fans——zombie fans problems, and studies the definition and development of zombie fans and the current solutions. According to the user's existing form and the characteristics of relationship between them, it proposes the concept of user attention and its calculation method from the point of view of the link analysis. The experiment compares and analyzes the user attention, user popularity value and user influence, and the result is that the user attention can effectively reduce the fake fans problem caused by zombie fans.

【Keywords】Sina micro – blog　Zombie fans　Fake fans　User attention

1　引　言

从 Twitter 到新浪微博的兴起,微博得到了空前的发展和广泛的使用。2011 年 5 月,首届中国微博大会公布的数据中,中国活跃微博用户数量已达到 2.2 – 2.4 亿[1],其中新浪微博用户超过了 2 亿[2]。微博在蓬勃发展的同时也遇到了许多亟待解决的问题,僵尸粉问题是其中之一,即微博的虚假粉丝问题。微博从主体到粉丝,从粉丝到内容,都存在虚假问题。虚假导致的微博信任危机严重影响了微博的发展[3]。针对僵尸粉问题,新浪采取了一些措施,但还没有找到行之有效的解决方法。目前对微博的研究更多地集中在微博的基础性研究、微博环境下的沟通研究和微博沟通价值的相关进展[4],关于僵尸粉问题解决途径的研究成果还很少。本文针对微博僵尸粉

收稿日期: 2012 – 03 – 27
收修改稿日期: 2012 – 05 – 06
* 本文系河北省自然科学基金项目"面向协同过滤的可信推荐算法研究"(项目编号: F2011203219) 的研究成果之一。

导致的虚假粉丝问题，从链接分析的角度提出了微博用户被关注度的概念，考察用户粉丝数量和质量两个方面。实验将得到的用户被关注度与用户人气值和用户影响力进行量化对比，表明用户被关注度可以有效地降低僵尸粉带来的虚假粉丝问题，较为合理地展现用户真实的受关注程度。

2　僵尸粉

(1) 僵尸粉的定义

“僵尸粉”一词，经国家语言资源监测与研究中心等机构专家审定入选 2010 年年度新词语，并收录到《中国语言生活状况报告》中。释义：微博上长期没有动态的粉丝，这些用户大都无图片、无评价、无关注，也称“死粉丝”、“僵尸粉丝”。互动百科称其为：“空头微博”，指微博上的虚假粉丝[5]；百度百科定义为：微博上的虚假粉丝，指花钱就可以买到“关注”，有名无实的微博粉丝，通常是由系统自动产生的恶意注册用户[6]。关于僵尸粉的定义目前还没有一个统一说法，但僵尸粉的发展已使得既存的僵尸粉定义与其真实的状况存在差别[7]。最初僵尸粉丝，都是无头像、无粉丝，并且不评论、不转发，由此得以“僵尸”闻名。但目前“僵尸粉”已经“复活”，发展为了“活粉”[8]，“活粉”是由人工添加的虚假粉丝，有头像、有资料、有评论以及有转发，与真粉丝非常相似。

(2) 僵尸粉的发展

僵尸粉制造者和经营者从开始在各大论坛、微博等平台出没，发出“专业团队为你快速增加高质量粉丝，有意者欢迎加 QQ…”等类似的邀请，发展到后来直接在淘宝等网店做起生意。2010 年，淘宝网通过搜索关键词关闭了一批“僵尸粉”网店，当众人认为“僵尸粉”现象有望得到遏制时，僵尸粉却在不断进化。目前卖家不再叫卖僵尸粉，而开始出售高质量的“活粉”。文献[9]把僵尸粉的发展归纳为三个阶段，称之为僵尸粉三代：第一代僵尸用户是无头像、无微博以及无粉丝的，关注的对象一半是名人，一半是僵尸用户；第二代则有头像、有微博、有粉丝，但微博数量一般低于 5 条，粉丝数量也一般在 100 以下，关注的有真实用户也有僵尸用户；而发展到第三代，僵尸粉已是有头像、有微博、有粉丝，每天定时更新微博，粉丝数量 100 到 300 之间，关注多为真实用户。

3　解决僵尸粉问题所采取的措施

3.1　粉丝管理软件

针对微博的僵尸粉问题，新浪微博应用中出现了很多粉丝管理软件，如粉丝工具箱[10]、关系趋势分析[11]等，还有清除僵尸粉的软件，如僵尸粉清理[12]、僵尸粉克星[13]等，但软件的设置都过于简单，而且目前“僵尸复活”，僵尸用户变得有头像、有粉丝、主动转发微博，所以如果把没有头像、发微博数少于 10 条、粉丝少于一定数目的不活跃用户都划为僵尸粉，很多真实用户将被误认为是僵尸粉。新注册用户在短时间内没有头像，发布微博数很少，甚至没有粉丝都很常见，所以这些软件目前并不可靠，因此新浪也叫停了平台上一些清除僵尸粉的应用。

3.2　用户影响力

僵尸粉的存在使得“人气排名”不真实[14]，新浪微博推出了用户影响力排名。影响力是由用户活跃度、传播力及覆盖度三方面组成，其中用户活跃度由用户原创微博数、原创被转发、评论次数以及与粉丝间私信的数目共同得到；用户传播力由用户原创微博被转发、评论的次数决定；而覆盖度则是由用户的当日新粉丝和粉丝的互动决定的[15]。用户影响力较人气排名更具合理性，能客观反映一段时间内用户的活动状态，但对解决僵尸粉问题并没有起到作用。本文按照规则的方法分别对选取的 8 位微博用户进行考察[16,17]，其中考察条件主要包括：注册时间、粉丝数、关注数/粉丝数比值、是否相互关注、个人资料等，对不同的条件赋予不同的权重，得到初步的用户僵尸粉数量[17]。用户粉丝中僵尸粉的统计结果如表 1 所示：

表 1　用户僵尸粉比率

用户	1	2	3	4	5	6	7	8
僵尸粉	5.32%	6.16%	13.67%	16.32%	10.12%	15.71%	17.1%	16.4%

通过对新浪微博开放 API 爬取的用户人气值，即粉丝数和用户影响力统计[15]，得出用户人气值与影响力的比较，如图 1 所示。

为了便于数据比较将用户影响力值进行统一缩放。图 1 中用户 3、5、7、8 人气值高于其影响力，而用户 1、2、6 人气值却低于其影响力，表明用户人气值与用户影响力之间不存在直接的线性关系；用户 1 较用户 2 人气值低，同时僵尸粉比用户 2 低 0.84% 的情况

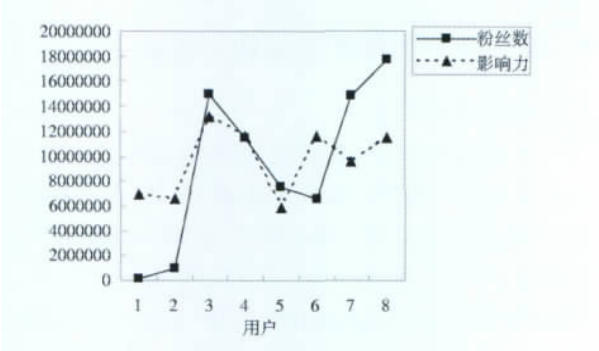

图 1　用户人气值与影响力比较

下影响力高出了用户 2;用户 6 比用户 5 人气值低,但僵尸粉高 5.59% 的情况下,影响力同样高出了用户 5;而用户 4 在僵尸粉占到了用户粉丝 16.32% 的情况下还能保持影响力不变,显然用户影响力与僵尸粉之间也没有直接的关系,且达不到降低僵尸粉影响的效果。

3.3　实名认证机制

2011 年 12 月 1 日,新浪微博推出了实名认证机制,意在通过用户对真实身份存在的责任来实现用户的自我约束,从而实现诚信微博平台的构建。这样不仅可以解决微博虚假信息泛滥的状况,同时也让僵尸粉无处藏身,这一举动受到一些用户的欢迎,但同时也还有很多不同声音的存在。65.91% 的人表示在微博中不使用真实姓名,并有超过 62% 的被调查者认为微博中没有必要实名制。新浪微博也指出,目前更多的用户不愿意以真实身份出现在网络平台,表现出较强的自我保护与隐匿意识[14,18,19],所以通过实名认证实现诚信微博平台的构建,解决微博的僵尸粉问题还难以实现。

Twitter 在 2008 年也遇到了僵尸粉问题[18],相比 Twitter 对营销僵尸的账户采取关闭的严厉措施,新浪和腾讯只将“僵尸”消除,而没有采取进一步的措施。新浪微博公关部负责人告知,会用技术手段来甄别、删除一些长期没有动态、同一 IP 地址申请多个微博的微博账号。但“博远科技”开发的系列软件可以使得每个账户都显示是不同 IP 地址[9]。总之,僵尸粉问题目前还没有找到较好的解决途径。

4　微博用户的用户被关注度

4.1　用户被关注度的提出

新浪微博给出的用户描述中,有三个显现的用户数据:关注的用户数、粉丝数和发布微博数(包括转发的微博)。其中用户粉丝数体现了用户的受关注程度,是用户之间选择关注与否的一个重要参考值。对用户数据分析发现,用户粉丝不仅存在数量差异也存在质量区别,用户粉丝可能是受关注度很高的认证用户、普通的微博用户、也可能是僵尸用户。用户是以个人主页形式存在的,每个粉丝都是用户主页的一条入链,表示对用户的支持,而用户自身的关注则是用户主页的出链,文献[20]指出链接分析更多地体现了一种用户关系,所以本文使用引文分析的思想,把链接分析法的 PangRank(PR) 算法引入微博平台,计算出用户主页的 PR 值,并把这个权值定义为微博用户的被关注度。实验证明用户被关注度可以削弱僵尸粉的影响,较为合理地体现用户受关注程度。

4.2　用户被关注度的算法描述

PR 算法存在着如主题漂移、平均权值和忽视用户兴趣等问题[21,22],其中主题漂移和用户兴趣问题对用户被关注度没有影响,不予考虑,但是平均权值问题在微博用户被关注度计算过程中是不能接受的,高人气用户和僵尸粉的被关注度是截然不同的。本文针对平均权值问题,将改进的不平均分配权值的 PR 算法应用到微博平台[22],实验结果表明该算法可以达到降低僵尸用户被关注度的效果。

改进后的 PR 算法基本思想是:重要页是被多个网页引用,或者是能被其他重要页引用的网页,同时被引用的次数大大超过其本身引用其他网页的次数。通过计算页面的入链数和出链数的比值,得到其获得 PR 值的能力;再根据其在作为入链时在所有入链中所占的比重,得到其带给所关注不同用户时的不同 PR 值;算法较好地区分了不同页面在作为入链页面时的重要程度。

算法描述为:

$$PR(u) = (1-d) + d\sum_{v_i \in P(u)} m_{v_i} PR(v_i)$$

其中,$d = 0.85$ 为经验值,$P(u)$ 为指向页面 u 的页面 v_i 的集合,$m_{v_i} = \frac{IO_i}{\sum_{i=0}^{n} IO_i}$ 为页面 u 从其链入页面 v_i 得到的 PR 值的比重。$IO_i = \frac{InL_i}{OutL_i}$ 为该网页 v_i 从其入链页面获取 PR 值的能力,IO 值越大,则越容易获得高 PR 值。考虑到用户出链数为 0 时,IO 值将为无穷大的情况,将出链数为 0 的页面的 IO 值设为较小的常数 m_0,如新

浪微博加“V”用户郎咸平，其主页只有数以百万记的入链数和 0 条出链。InL_i 和 $OutL_i$ 分别表示页面 v_i 的入链和出链数，相当于输入和输出 PR 值。在指向页面 u 的所有页面 $v_i \in P(u)$ 中，网页 v_i 的全部 n 个出链 $O_1, O_2, \cdots, O_n$ 的入链总数和出链总数的比值 IO_i，分别为 $IO_1, IO_2, \cdots, IO_n$ 对应页面从其链入页面中获得 PR 值的能力。

PR 算法对用户主页进行加权计算后，不同用户有不同的 PR 值，同时为所关注的用户带去不同的 PR 权值。僵尸粉在作为带入 PR 值的入链时，由于几乎没有正常的用户去关注毫无意义的僵尸用户，所以其被关注度较低，即带入 PR 值的能力较小，4 位僵尸用户的被关注度如表 2 所示：

表 2　僵尸用户的被关注度

用户	1	2	3	4
被关注度	0.1721	1.0712	0.6233	0.7941

在用户被关注度的加权计算和矩阵的循环迭代中会扩大高 PR 值的比重，而降低低 PR 值的比重，进而达到降低僵尸粉影响的目的。

5　实验结果分析

僵尸粉目前更多出现在认证名人用户中，实验数据取认证的高人气用户为研究对象。但微博用户的粉丝数量庞大且数据动态变化，而实验条件和设备存在局限性，为了实验可行、方便，选取 8 位用户某一时刻的数据状态进行实验，并对结果进行对比分析。2012 年 3 月 5 日爬取的用户数据如表 3 所示：

表 3　爬取的部分用户数据

用户	关注数	粉丝数	微博数
1	83	136 597	2 285
2	174	959 600	2 732
3	58	1 870 964	213
4	301	11 474 460	4 743
5	0	7 573 563	193
6	86	6 577 967	29 972
7	355	14 794 387	4 093
8	548	17 690 015	5 363

可以看出用户粉丝数量的庞大，实验数据中还用到了表 3 中粉丝用户的入链和出链的数目计算得到 IO_i、m_{v_i}，进而进行迭代计算。

计算过程使用 PR 的幂法[23]，把问题转换为求解 $\lim_{n\to\infty} AX$ 的值，在 Matlab 环境实现页面 PR 值的收敛计算，其中向量 X 为页面初始 PR 值，设为 1，矩阵 $A = \alpha \times P_1^T + (1-\alpha) \times ee^T / m_{v_i}$，$\alpha$ 为阻尼系数，取经验值 0.85，P_1^T 为概率转移矩阵，e^T 为 n 维的全 1 行，m_v 为每个页面贡献出的 PR 值。由于微博用户之间的关系、特征可以发现初始矩阵 P 是一个对角线均为 0，下三角为 1，而上三角稀疏的特殊矩阵，将每一行除以该行非零数字之和得到矩阵 P_1，矩阵 P_1 记录了每个网页跳转到其他网页的概率。再将矩阵 P_1 转置得到 P_1^T 进行计算。为了简化计算和运行的方便，将矩阵分解，转换为求解 $\lim_{n\to\infty}(L+U)X$，其中 L 为上三角和对角线都为 0 且下三角稀疏的矩阵，U 为下三角和对角线都为 0 而上三角元素非 0 的矩阵，由于 L 矩阵的特殊在迭代过程中 X 变为了 0，所以变为求解 $\lim_{n\to\infty} UX$ 的过程。经过迭代计算，向量 X 稳定后用户 1－8 的用户被关注度、用户影响力和用户粉丝中的僵尸粉比率如表 4 所示：

表 4　用户被关注度与僵尸粉对比

用户	1	2	3	4	5	6	7	8
影响力	696	661	1 312	1 157	592	1 156	593	1 145
僵尸粉	5.32%	6.16%	13.67%	16.32%	10.12%	15.71%	17.10%	16.40%
被关注度	2.874	3.114	5.433	5.971	3.547	5.872	3.762	6.213

用户人气值、影响力、被关注度的对比（为了方便对比，对用户人气值和用户影响力进行了等量缩放）如图 2 所示：

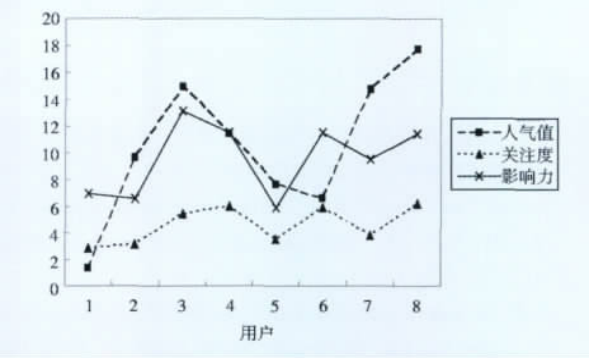

图 2　用户人气值、影响力、被关注度对比

用户 1 选取的不是认证名人而是新浪的一个认证应用，其人气值、影响力和被关注度都明显低于其他人气用户，事实上又明显高于普通的微博用户。调查结果中用户 1 的僵尸粉比率较少，占其粉丝数 5.32%，其中高人气粉丝也较少，绝大多数粉丝为普通的真实粉丝用户，相比其他用户无论其粉丝的数量和质量都明显较低，所以虽然其被关注度相对高于其自身人气值

却明显低于其他用户;用户 7 人气值在所有用户中较高,而影响力在等量级人气值用户中出现了跌落,用户被关注度更是出现了一个低谷,通过对其分析发现,用户 7 的入链中高被关注度的粉丝用户数量与用户 3 相当,但是其僵尸粉比用户 3 多出 3.43%,即在其入链中包含大量 PR 很低的用户——僵尸粉,所以出现了被关注度低于用户 3 的情况;用户 4、6、8 的情况看似出现了异常,在僵尸粉分别高达 16.32%、15.71%、16.40% 的情况下,用户被关注度仍然很高。对其迭代过程做进一步分析可以发现,在其迭代过程中,向量 X 中出现了明显多于其他用户迭代过程中的高 PR 值,即高 PR 的入链用户,所以图 2 中的情况并非异常,相反是符合用户真实状况的结果。

6 结 语

本文分析了微博平台的僵尸粉问题,根据微博用户存在的形式和用户间关系的特征,从链接分析的角度提出了用户被关注度的概念,将用户的关注与粉丝作为用户页面的出链和入链,由于不同用户的入链存在明显的数量和质量的差异,所以使用改进的不平均分配权值的 PR 算法计算得到用户页面的被关注度。实验证明用户被关注度可以削弱僵尸粉带来的虚假粉丝问题,缓解当前虚假人气的状况。

参考文献:

[1] 陈渊,林磊,孙承杰,等. 一种面向微博用户的标签推荐方法[J]. *智能计算机与应用*,2011,10(3):21 - 22. (Chen Yuan, Lin Lei, Sun Chengjie, et al. A Tag Recommendation Method for Microblog Users [J]. *Intelligent Computer and Applications*, 2011, 10(3):21 - 22.)

[2] 何毅. 新浪微博注册用户已超 2 亿[EB/OL]. [2011 - 08 - 27]. http://article.pchome.net/content - 1379385.html. (He Yi. The Registered Users of Sina Micro - blog has Exceeded 200 Million[EB/OL]. [2011 - 08 - 27]. http://article.pchome.net/content - 1379385.html.)

[3] 李畅. 中国微博发展隐忧[J]. *西南民族大学学报*,2011,8:170 - 171. (Li Chang. The Development Malaises of Micro - blog in China[J]. *Journal of Southwest University for Nationalities*, 2011, 8:170 - 171.)

[4] 闫幸,常亚平. 微博研究综述[J]. *情报杂志*,2011,30(9):61 - 65. (Yan Xing, Chang Yaping. A Review on the Research of Microblog [J]. *Journal of Intelligence*, 2011, 30(9):61 - 65.)

[5] 互动百科: 僵尸粉[EB/OL]. [2012 - 02 - 10]. http://www.hudong.com/wiki/% E5% 83% B5% E5% B0% B8% E7% B2% 89. (Hudong: Zombie Fans [EB/OL]. [2012 - 02 - 10]. http://www.hudong.com/wiki/% E5% 83% B5% E5% B0% B8% E7% B2% 89.)

[6] 百度百科: 僵尸粉[EB/OL]. [2012 - 02 - 10]. http://baike.baidu.com/view/4047998.html. (Baidu Baike: Zombie Fans [EB/OL]. [2012 - 02 - 10]. http://baike.baidu.com/view/4047998.html.)

[7] 孙大平. 社会媒介场域话语符号权力的探索与反思—以新浪微博为例[D]. 合肥: 中国科学技术大学,2011: 94 - 96. (Sun Daping. Exploration and Reflection on Discourse Symbolic Power in Social Media Field: Case Study for Sina Microblog [D]. Hefei: University of Science and Technology of China, 2011: 94 - 96.)

[8] 马凌霜. 微博有僵尸粉出没[N]. 广东科技报,2011 - 09 - 17(7). (Ma Lingshuang. The Micro - blog has Emerged the Zombie Fans [N]. GUANGDONG KE JI BAO. 2011 - 09 - 17(7).)

[9] 秦旺,陈震霖. 活捉僵尸粉[J/OL]. *南都周刊*,2011(37). [2011 - 02 - 10]. http://www.nbweekly.com/news/special/201109/27547.aspx. (Qin Wang, Chen Zhenlin. Capture Alive the Zombie Fans [J/OL]. *Southern Metropolis Weekly*, 2011(37). [2011 - 02 - 10]. http://www.nbweekly.com/news/special/201109/27547.aspx.)

[10] 广州赢新网络科技有限公司. 粉丝工具箱[CP/OL]. [2011 - 09 - 27]. http://tejiatai.sinaapp.com/. (Guangzhou Yingxin Network Technology Company Limited. Fans Tool Case [CP/OL]. [2011 - 09 - 27]. http://tejiatai.sinaapp.com/.)

[11] 安卓游戏精选. 关系趋势分析[CP/OL]. [2011 - 09 - 27]. http://weibo.bibifa.com/. (Android Games Handpick. Relationship Trend Analysis [CP/OL]. [2011 - 09 - 27]. http://weibo.bibifa.com/.)

[12] 王洋. 僵尸粉清理[CP/OL]. [2011 - 09 - 27]. http://apps.weibo.com/jiangshifen. (Wang Yang. Clean up the Zombie Fans [CP/OL]. [2011 - 09 - 27]. http://apps.weibo.com/jiangshifen.)

[13] 网络科技. 僵尸粉克星[CP/OL]. [2011 - 09 - 27]. http://apps.weibo.com/kill_zombies. (Network Technology. The Nemesis of Zombie Fans [CP/OL]. [2011 - 09 - 27]. http://apps.weibo.com/kill_zombies.)

[14] 段婷婷. 微博影响力产生机制的研究[D]. 济南: 山东大学,2011: 36 - 54. (Duan Tingting. The Production Mechanism of Microblog Impact [D]. Jinan: Shandong University, 2011: 36 - 54.)

[15] 新浪微博: 用户影响力[EB/OL]. [2012 - 02 - 10]. http://data.weibo.com/index. (Sina Micro - blog: User Influence [EB/OL]. [2012 - 02 - 10]. http://data.weibo.com/index.)

[16] 百度百科: 僵尸粉统计器[EB/OL]. [2012 - 02 - 10]. http://

baike. baidu. com/view/5777226. htm. (Baidu Baike: Zombie Fans Statistical System [EB/OL]. [2012 - 02 - 10]. http: //baike. baidu. com/view/5777226. htm.)

[17] 僵尸粉识别器[EB/OL]. [2011 - 02 - 10]. http: //zombies. sinaapp. com. (The Zombie Fans Recognizer[EB/OL]. [2011 - 02 - 10] http: //zombies. sinaapp. com.)

[18] Mendoza M, Poblete B, Castillo C. Twitter Under Crisis: Can We Trust What We RT? [C]. In: *Proceedings of the 1st Workshop on Social Media Analytics (SOMA '10)*. 2010.

[19] 张红军,王瑞. 关于微博实名制的思考[J]. *新闻爱好者*,2011(20): 10 - 11. (Zhang Hongjun, Wang Rui. The Real Name Thinking of Micro - blog [J]. *Journalism Lover*,2011(20): 10 - 11.)

[20] Cha M, Haddadi H, Benevenuto F, et al. Measuring User Influence in Twitter: The Million Follower Fallacy [C]. In: *Proceedings of International AAAI Conference on Weblogs and Social*. 2010.

[21] 王冬,雷景生. 一种基于 PageRank 的页面排序改进算法[J]. *微电子学与计算机*,2009,26(4): 210 - 214. (Wang Dong, Lei Jingsheng. An Improved Ranking Algorithms Based on PageRank [J]. *Microelectronics & Computer*,2009,26(4): 210 - 214.)

[22] 王德广,周志刚,梁旭. PageRank 算法的分析及其改进[J]. *计算机工程*,2010,36(22): 291 - 293. (Wang Deguang, Zhou Zhigang, Liang Xu. Analysis of PageRank Algorithm and Its Improvement [J]. *Computer Engineering*,2010,36(22): 291 - 293.)

[23] 金迪,马衍民. PageRank 算法的分析及实现[J]. *经济技术协作信息*,2009(18): 118. (Jin Di, Ma Yanmin. The Analysis and Realism of the PageRank Algorithms [J]. *Economic and Technological Cooperation Information*,2009(18): 118.)

(作者 E - mail: kefeng510@163. com)

数据管理需要更多关注

据全球最大的市场情报提供商国际数据公司(IDC) 指出,2011 年创造和复制的数据大约是 1.8 泽字节(万亿 GB) 。这相当于是 2000 亿部 2 小时时长的高清晰度电影,不停地播放这些电影需要大概 4 700 万年。

“大数据”的指数增长对任何大规模的组织来说都正在成为一个紧迫的问题,信息和数据的采集、存储、检索、分享和分析变得越来越难以控制,影响了气象、基因学、网络搜索、金融和商业信息等诸多领域。大数据是信息技术术语,指数据和内容的增长如此之大和复杂,几乎不可能使用现有的数据库管理工具来处理。

数据管理问题,即如何将这些大量的原始数据转化为可操作的信息,从而使得公司能够从中获得商业价值,已经成为近期各大会议的主题。Smartlogic 创始人兼 CEO,Semaphore 内容智能管理平台的开发者 Bentley 解释: “数据管理的重点是从非结构化的大数据中抽取事实和观点,为组织提供透明化管理企业内部内容的方法,不论内容处于何位置、以何格式存储,从而帮助他们找出最相关的信息。Smartlogic 的 Semaphore 能够理解内容,并使用词汇和元数据来充实内容,因此内容便更为可见,此外,Semaphore 还能通过扩充查询的上下文从而放大查询,以得到精确和相关的查询结果。”

Bentley 最近在 Lucerne 创新会议(2012 年 5 月 7 日 - 10 日在波士顿召开) 上做了题为“大数据遇到了智能内容元数据”的报告,强调了元数据管理的重要性,他指出: 随着信息变得更加普及,对理解和挖掘内容来说,元数据管理变得越来越重要。

Bentley 认为: “元数据是解锁信息资产的钥匙。元数据管理得好,信息资产就更加有价值; 管理不善,信息资产就没有价值,因为元数据为这些资产带来了更大的代价和风险。”

元数据是检索、发现和分类任何内容形式包括大数据的一个关键因素。基于叙词表或是本体或是通过情感分析和事实抽取来为这些内容添加辅助元数据,能够改进内容的质量和深度,从而获得更高的相关度、查准率和查全率。

(编译自: http: //taxodiary. com/2012/06/data - management - demanding - attention/ ~ ~ V)

(本刊讯)