·先进计算与数据处理· 文章编号: 1000-3428(2015)08-0013-05 文献标识码: A 中图分类号: TP391

基于 Ceph 的分布式存储节能技术研究

沈良好,吴庆波,杨沙洲

(国防科学技术大学计算机学院,长沙 410073)

摘 要: 分布式存储作为目前流行的数据中心存储系统,在具有高性能、高扩展性的同时,面临着系统能耗增加的问题。为此,基于 Ceph 分布式存储,分析其数据布局在节能方面的不足,提出划分功耗组的节能优化算法,以提升系统节能比例。建立 Ceph 的多级功耗模型并给出管理策略,设计并实现 Ceph 系统的多级功耗管理框架,以进行 Ceph 系统功耗的动态管理。实验结果证明,该框架能够有效降低 Ceph 分布式存储的能耗,并保证系统的服务质量和数据可用性。

关键词: 分布式存储; 节能计算; Ceph 存储; 功耗管理; 数据布局

中文引用格式: 沈良好,吴庆波,杨沙洲. 基于 Ceph 的分布式存储节能技术研究[J]. 计算机工程,2015,41(8):13-17.

英文引用格式: Shen Lianghao,Wu Qingbo,Yang Shaozhou. Research on Distributed Storage Energy Saving Tech-nologies Based on Ceph[J]. Computer Engineering,2015,41(8):13-17.

Research on Distributed Storage Energy Saving Technologies Based on Ceph

SHEN Lianghao,WU Qingbo,YANG Shazhou

(School of Computer,National University of Defence Technology,Changsha 410073,China)

【Abstract】 Distributed storage systems are widely used in data centers,because of high performance and scalability,yet most of them are not energy-efficient. This paper,based on Ceph,analyses the data layout's disadvantage in energy saving and proposes a power group partition algorithm to increase the energy-saving proportion,builds a multi-level power mode of Ceph and proposes a multi-level power management strategy,besides,it designs and implements a power management framework based on the two former points,to manage Ceph's power dynamicly. Experimental results show that this framework can reduce the power consumption of Ceph effectively,meanwhile,the quality of service,data availablity are preserved.

【Key words】 distributed storage; energy saving computing; Ceph storage; power consumption management; data layout

DOI: 10.3969/j.issn.1000-3428.2015.08.003

1 概述

能耗管理是数据中心面临的重要挑战。一方面,随着基础架构规模的扩大,数据中心需要为日益剧增的能源消耗买单,在数据中心总体总拥有成本(Total Cost of Ownership,TCO)中,用于能源的费用已经成为重要的组成部分。另一方面,随着数量的增加与规模的扩大,数据中心所消耗的能源,在整个社会能源消耗中所占的比重也越来越大,据统计,全球数据中心在 2010 年对电能的消耗超过了2×10^{11}千瓦时,约占全球总用电量的 1.3%,且呈逐年上升的趋势在数据中心的组件中,存储系统是能源消耗主要来源之一,仅次于计算资源,约占 30%。因此,降低存储系统的能耗是达到数据中心节能目的的重要手段。近年来,存储系统节能技术受到了广泛关注,从单个磁盘到磁盘阵列,再到分布式存储,均出现了大量的研究工作[2-5]。

分布式存储是目前大部分数据中心所采用的存储形式。作为近年热门的分布式存储,Ceph[6] 因具备了高扩展性、高性能、高可靠性的特点,而备受关注。因此,本文基于开源项目 Ceph,进行分布式存储系统节能技术的研究。本文的工作主要有: 以节能为目的的 Ceph 数据布局优化; 多级功耗管理策略; 功耗管理框架的设计与实现。该框架结合了上述数据布局优化方法、多级功耗管理策略以及硬件的节能功能,实现了 Ceph 分布式存储的动态功耗

基金项目: 国家"863"计划基金资助项目"智能云服务与管理平台核心软件及系统"(2013AA01A212)。

作者简介: 沈良好(1986-),男,硕士,主研方向: 分布式存储技术; 吴庆波,研究员; 杨沙洲,副研究员。

收稿日期: 2014-09-09 修回日期: 2014-10-07 E-mail: shenlianghao@kylinos.com.cn

管理。

2 相关工作

分布式存储系统的节能技术近年来受到了广泛的关注,是存储领域热门的研究方向之一。

微软剑桥研究院的 Thereska E 等人设计并开发了 Sierra[7] 分布式存储系统,该系统使用了能耗感知的数据布局和基于负载预测的节点状态管理,在系统低负载时,关闭部分节点,从而增加系统的能源利用率。为保证系统的容错能力以及数据的一致性,Sierra 使用了分布式虚拟日志(DVL)技术。经过测试,该系统在作为 Hotmail 和 Windows Messenger 服务的后端存储时能够节省 23% 以上的能耗,且性能的损失相当微小。

UIUC 的 Kaushik R 等人基于标准的 HDFS 提出了其节能的衍生版 GreenHDFS[8]。在 GreenHDFS 中,数据节点最初被划分为热区和冷区,处于热区的数据(约 70%)有着更高的访问频率,所以为保证性能,热区的节点是一直处于活跃状态的;而冷区的数据(约 30%)使用率非常低,所以冷区的节点将会进入省电模式。一个自适应的划分策略用来动态地指定节点所属的区,并使得冷区节点能够达到一定的数量,进而增加整个集群节能的程度:在 Yahoo 的一个真实负载环境中,GreenHDFS 在 3 个月的测试时间内,节省了 26% 的能耗。

此外,针对 Ceph 分布式存储的节能技术也受到了关注,文献[9]提出了一种 Ceph OSD(对象存储设备)的自适应的磁盘降速算法。该算法针对单个 OSD,在其低负载时降低所对应的磁盘速度,进入节能状态。他们的工作针对的只是部分 OSD 上磁盘的节能,所以对整个系统能耗的影响十分有限。

3 数据布局优化

分布式存储中的数据放置算法及其产生的数据布局是影响系统可靠性、扩展性的重要因素之一。Ceph 存储系统基于 CRUSH 算法的数据布局是其具备可靠性、高性能、高扩展性的基础,但同时也限制了系统节能的能力。

3.1 Ceph 数据布局

作为 Ceph 存储的关键技术之一,CRUSH 算法由 Weil S A 于 2006 年提出[10]。CRUSH 基于伪随机的哈希算法产生确定的均匀的数据分布。CRUSH 有多个输入,包含了对象 id,Crushmap 和放置规则。其中,对象 id 一般用于区别需要存放的数据对象,实际运行时对象会被映射到不同的放置组(PG),所以作为输入 id 的实际是 PG 的 id。Crushmap 描述了数据节点的层级关系,用树形结构表示,包含 bucket 和 device 类型的节点,其中 bucket 节点可以包含其他类型的 bucket 和 device 节点,通常用于描述故障域(failure domain),如主机、机架、机柜等,把副本分布于不同故障域,是 Ceph 保证数据可靠性的重要措施;而 device 节点只能作为叶子节点,表示对象存储设备(OSD)。放置规则用于指定副本放置的策略,包含了 take,select,emti 等语句,其中 select 可以指定副本个数和副本放置的故障域。在一个包含 4 台主机的集群中,副本分布策略为:select(2,host),即数据的 2 个副本放置于不同的主机上,则通过 CRUSH 算法产生的数据布局如图 1 所示。

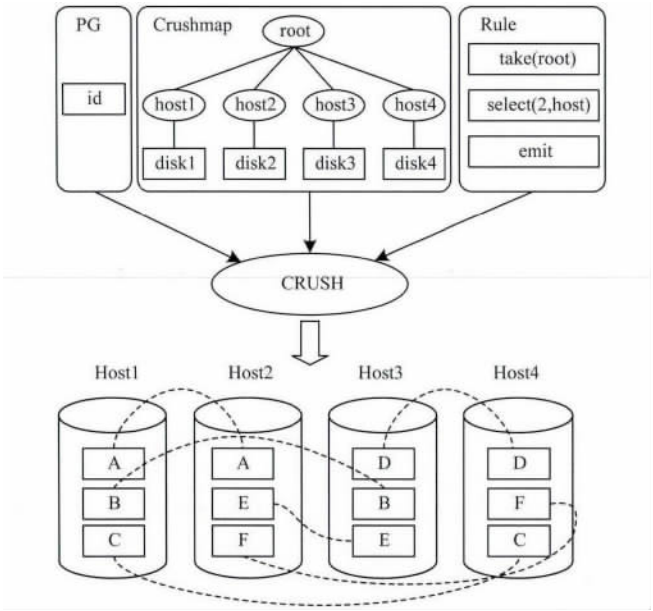

图 1 Crush 算法数据布局

3.2 存储系统存在的不足

很多应用场景中都存在低负载时期,此时,为了节省系统的能耗,可以关闭部分数据节点。出于以下考虑,关闭节点时应保证数据全集是可用的(即任意的数据对象至少有一个以上副本是未被关闭的):一方面,如果发生不可用数据的访问,则会产生非常大的访问延迟;另一方面,频繁开启相应的节点会产生不可忽视的额外能耗。在 Ceph 存储系统中,数据节点数为 N,被划分为 n 个故障域 fd,副本分布策略为 select(r,fd),则在系统低负载系统时期最多可以关闭的故障域的个数为 $n'<r$,因为任何大于或者等于 r 个故障域的组合中,必然包含了某些数据的所有副本。如图 1 中,可以关闭的主机个数为 1 台,当关闭 2 台以上主机时,A 个 ~ F 个数据块必然有一个无法访问。随着集群规模的增加,最多只能关闭 $r-1$ 个故障域的节点,可以达到最大节能比例为 $(r-1)/n$,当集群规模较大时,节能的效果微乎其微。

3.3　节能优化

结合 CRUSH 的特性,在 Crushmap 中引入了功耗组(PowerGroup) 的 bucket,用于对故障域集合进行再次划分,即数据副本在放置于不同故障域前,将首先被分布于不同的功耗组。同一个功耗组的节点处于相同的能耗状态,功耗组的个数等于副本个数 r。优化算法的描述如下:

```
{
  detect target failure domain; /* 确定分布的故障域* /
  declare PowerGroup bucket;
  PowerGroup [] pgs;
  for fd in all failure domains do
    PGid = fd. id mod r;        /* 故障域功耗组划分* /
    pgs [PGid] . items. append( id) ;
  rule = 'select( r,PowerGroup) ';
  insert_rule( rule) ;          /* 插入新 select 规则* /
}
```

以图 1 中场景为例,则经过优化的数据布局如图 2 所示。

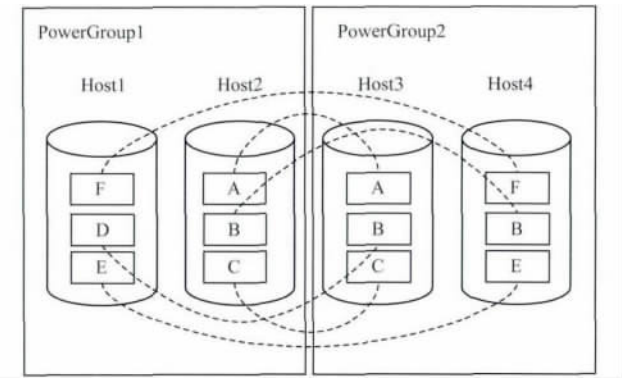

图 2　优化后的数据布局

此时,可以关闭的节点为 2 个,且所有数据仍然可用。经过优化的数据布局中,数据的副本分别位于 r 个不同的功耗组中,在保证数据集可用的情况下,则最多可以关闭 $r-1$ 个功耗组的节点,系统能够达到的节能比例为 $r-1/r$,当集群规模较大时,节省的能耗是非常可观的。同时,由于很好地实现了副本的分布,可以关闭的功耗组个数可以是 $1 \sim r-1$ 的任何一个,从而为系统的多级功耗管理提供了基础。

4　多级功耗管理

基于对数据副本分布的优化,使得系统中活跃的功耗组个数可以根据系统 I/O 负载的情况按需调整,让系统处于不同的功耗级别,从而实现系统的多级功耗管理,减少能耗。

4.1　多级功耗模型

在 Ceph 中,系统的主要能耗来自于 OSD 节点。在一个包含 n 个 OSD 节点的 Ceph 集群中,副本个数设置为 r,经过数据布局优化后,节点被划分至 r 个功耗组,若单个 OSD 节点的功耗为 p,则单个功耗组的功耗为:

$$P_g = p \times n/r$$

系统中活跃(未关闭或休眠) 的功耗组个数为 r_{active},则系统功耗为:

$$P_{total} = P_g \times r_{active} = (n/r) \times p \times r_{active} = n \times p \times \left(\frac{r_{active}}{r}\right)$$

其中,r_{active} 的取值可以为 $1 \sim r$,即系统可以处于 P_1,P_2,…,P_r 不同的功耗级别。在一段时间 T 内,系统所消耗的能耗为:

$$E = \sum_{i=1}^{r} P_i t_i + E_t \tag{1}$$

其中,t_i 是系统处于相应功耗级别的时间; E_t 是系统用于级别切换所消耗的能耗之和。

4.2　功耗级别管理

功耗级别管理的主要任务是根据 I/O 负载状态动态调整功耗级别,在保证服务质量的同时,尽量减少功耗。I/O 负载状态可以通过统计分析或者预测的方式确定,本文采用的是前者,即收集并统计系统在一定时间内的 I/O 数据如 I/O 次数、I/O 数据量等,并以此确定系统的 I/O 负载状态。为描述不同场景下的 I/O 负载状态,需要对随机 I/O 和顺序I/O 都进行收集与统计,因此,在可配置的时间窗口 W 内,对系统的 I/O 状态数据可统计为:

$$IO_{ran} = IO_{rtotal}/W$$

$$IO_{seq} = IO_{stotal}/W$$

其中,IO_{rtotal} 为 W 时间内发生的随机 I/O 的次数; IO_{stotal} 为 W 时间内顺序 I/O 请求的数据量。两者分别与预先测得的系统峰值 I/O 能力对比,得出系统 I/O 状态的量化描述,即 I/O 负载率 L:

$$L = \max\left\{\frac{IO_{ran}}{IO_{rpeak}}, \frac{IO_{seq}}{IO_{speak}}\right\} \tag{2}$$

其中,L 被用于与当前功耗率 $P_l = r_{active}/r$ 比较,确定系统下一阶段所处的功耗级别: 当系统的 I/O 负载率高于能耗率时,活跃的功耗组已经不能满足 I/O 负载的要求,需要更多的功耗组提供服务; 当系统的 I/O 负载率低于能耗率时,则系统中有部分功耗组可以被关闭,以达到节能的目的。功耗级别确定的伪代码为:

```
while ( overtime( W) ) {
      if( L < p_l) {                  /* 切换至低级别* /
while ( ( r_active > 1) && ( L < = p_l) )
  r_active -;
      }
      if ( L > p_l{                   /* 切换至高级别* /
        while( ( r_active < r) && ( L > = p_l) )
```

r_{active} + +;
}
r_{active}' = r_{active};
}

输出的 r_{active} 将作为下一个 W 时间段内的系统所处的功耗级别，此时需要关闭/开启的功耗组的个数为 $|r_{active}' - r_{active}|$。为保证系统的高可用性，可以设置允许的最少活跃的功耗组个数为 2，即允许系统数据副本至少有 2 个是可用的。

5 系统实现与实验评估

基于第 3 节、第 4 节所述的优化方法，结合 Ceph 自身技术特点，本文设计并实现了一个能耗管理框架，并进行了实验评估。

5.1 系统架构

如图 3 所示，该功耗管理框架由 4 个模块组成，其中 Layout Optimizer 模块实现针对 CRUSH 的数据布局的优化算法，并生成新的 Crushmap 和放置规则；I/O Tracer 模块用于跟踪与统计系统的 I/O 数据，即以一定的频率采集分析 Ceph Log 中的 I/O 相关信息记录；Lever Shifter 则是管理系统功耗级别的模块，根据跟踪统计到的 I/O 数据，分析当前系统的 I/O 负载状态，基于所选择的的策略确定是否需要切换功耗级别；Status Manager 模块负责功耗级别切换的执行，首先利用 Ceph 的 OSD 状态管理工具设置 OSD 在 Ceph 集群中的状态为 noout，保证不会因主动停止 OSD 而发生数据迁移，接着通过远程休眠/网络唤醒(WOL[11])等技术控制 OSD 所在的服务器电源管理状态，基于休眠/唤醒的方式比传统的关闭/开启服务器的方式更为节能，式(1)中用于级别切换的能耗对系统的总能耗的影响几乎可以忽略不计，且响应时间更短。

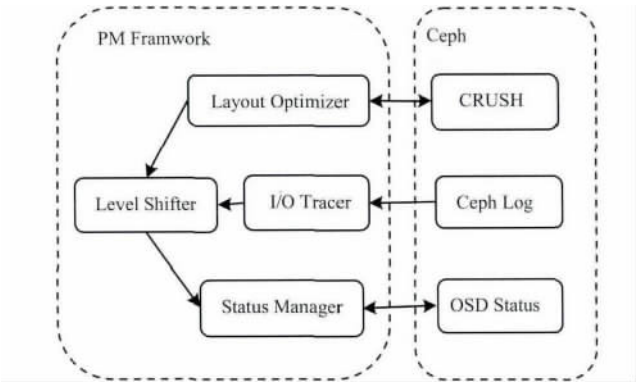

图 3 系统架构

5.2 实验评估

在一个有 6 个 OSD 节点的 Ceph 系统中进行了该功耗管理框架的实验评估。Ceph 版本为 0.80.5，操作系统为 Kylin3.2，每个节点上配置一个 OSD，副本策略为 select(3,host)。为模拟不同的负载场景，用 fio[12] 测试工具进行测试。

如图 4 所示内容为连续 8 h 测试(级别切换的超时时间设置为 10 min，通过 thinktime 参数使得系统约 4 h 处于低负载时期)中，系统分别处于各能耗级别的次数。

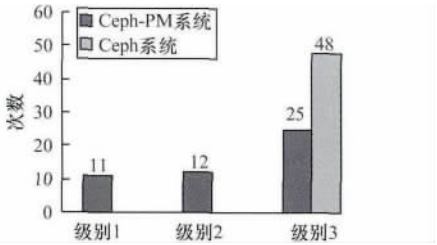

图 4 系统所处功耗级别的次数

在未使用节能框架时，系统运行 8 h 的能耗约为 48×6×10 min×400 W，在开启了节能框架(Ceph-PM)后，系统运行 8 h 的能耗约为(11×2+12×4+25×6)×10 min×400 W，系统达到的节能比例约为 25%，如果系统规模增大，节省的能源开支将会非常可观。但需要注意的是，在真实环境中，系统负载的变化可能更为频繁、剧烈，所以需要更精确、复杂的负载级别切换策略，这也是本文未来的工作内容之一。

图 5、图 6 描述的是系统在低功耗状态下对随机 I/O 和顺序 I/O 的响应时间的影响，总体上，读操作平均响应时间有略微变化，但不会太大地影响服务质量。

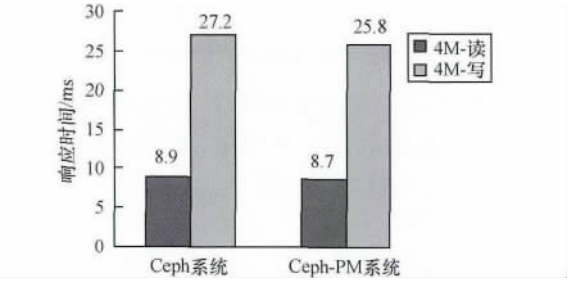

图 5 随机 I/O 的平均响应时间

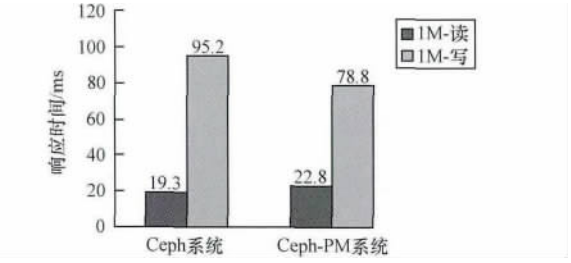

图 6 顺序 I/O 的平均响应时间

而写操作时，由于低功耗状态下需要写副本的次数变少，对于客户端来说，响应时间反而更小一些，且对于顺序写影响更为明显。

6 结束语

本文基于 Ceph 系统,研究分布式存储技术,分析基于 CRUSH 算法的数据布局存在的不足,提出以节能为目的的优化算法和系统多级功耗管理策略,并实现了 Ceph 的多级功耗管理框架。实验结果表明,该能耗管理框架能够根据系统负载变化动态地调整系统功耗级别,有效地降低系统能耗。

参考文献

[1] Koomey J. Growth in Data Center Electricity Use 2005 to 2010 [EB/OL]. (2011-10-11). http://www.analytic spress.com/datacenters.html.

[2] Gurumurthi S, Sivasubramaniam A, Kandemir M, et al. DRPM: Dynamic Speed Control for Power Management in Server Class Disks [C]//Proceedings of the 30th Annual International Symposium on Computer Architecture. San Diego, USA: IEEE Press, 2003: 211-219.

[3] 李海东. 磁盘阵列节能技术研究与实现 [D]. 武汉: 华中科技大学, 2009.

[4] Verma A, Koller R, Useche L, et al. SRCMap: Energy Proportional Storage Using Dynamic Consolida-tion [C]// Proceedings of FAST'10. San Jose, USA: USENIX Association, 2010: 148-155.

[5] 廖 彬, 于 炯, 孙 华, 等. 基于存储结构重配置的分布式存储系统节能算法 [J]. 计算机研究与发展, 2013, 50(1): 3-18.

[6] Weil S A, Brandt S A, Miller E L. Ceph: A Scalable, High-performance Distributed File System [C]// Proceedings of OSDI'06. Seattle, USA: USENIX Association, 2006: 269-277.

[7] Thereska E, Donnelly A, Narayanan D. Sierra: Practical Power-proportionality for Data Center Storage [C]// Proceedings of EuroSys'11. Salzburg, Austria: ACM Press, 2012: 153-161.

[8] Kaushik R T, Bhandarkar M. Greenhdfs: Towards an Energy-conserving, Storage-efficient, Hybrid Hadoop Compute Cluster [C]//Proceedings of USENIX Annual Technical Conference. Boston, USA: USENIX Association, 2010: 159-167.

[9] Bisson T, Wu J, Brandt S A. A Distributed Spin-down Algorithm for an Object-based Storage Device with Write Redirection [C]//Proceedings of the 7th Work-shop on Distributed Data and Structures. Santa Clara, USA: ACM Press, 2006: 459-468.

[10] Weil S A, Brandt S A, Miller E L, et al. CRUSH: Controlled, Scalable, Decentralized Placement of Replicated Data [C]//Proceedings of 2006 ACM/IEEE Conference on Supercomputing. Tampa, USA: ACM Press, 2006: 367-378.

[11] Wake-on-Lan [EB/OL]. (2013-10-10). http://en.wikipedia.org/wiki/Wake_on_lan.

[12] Fio [EB/OL]. (2013-10-10). http://freecode.com/projects/fio.

编辑 索书志

(上接第 12 页)

虽然本文能够预测电视剧的点播排名,但预测准确度还有待提高,并且不能准确预测电视剧的具体点播次数,下一步研究工作主要分为 2 个部分: (1) 考虑更多相关因素,由于视频点播系统中电视剧受很多因素的影响,一些因素本文未使用,如电视台对电视剧的影响因素; (2) 进行更精确的文本数据处理,由于使用电视剧名作为关键字进行匹配的方式会忽略很多信息,因此今后将通过关键词扩展进一步提高预测性能。

参考文献

[1] 刘 珊. 购剧有道: 电视剧交易中的买方角色 [EB/OL]. (2013-11-14). http://www.meijiezazhi.com/zt/nr/2013-11-14/13144.html.

[2] 洪皓轶. 电视剧收视率预估的市场化操作模式构建探析 [J]. 电视研究, 2013, (2): 71-73.

[3] Abrahamsson H, Nordmark M. Program Popularity and Viewer Behavior in a Large TV-on-demand System [C]// Proceedings of 2012 ACM Conference on Internet Measurement. New York, USA: ACM Press, 2012: 199-210.

[4] Qiu T, Ge Z, Lee S, et al. Modeling Channel Popularity Dynamics in a Large IPTV System [C]//Proceedings of the 11th International Joint Conference on Measurement and Modeling of Computer Systems. New York, USA: ACM Press, 2009: 275-286.

[5] Morris M R, Teevan J, Panovich K. What Do People Ask Their Social Networks, and Why? A Survey Study of Status Message Q & A Behavior [C]//Proceedings of SIGCHI Conference on Human Factors in Computing Systems. New York, USA: ACM Press, 2010: 1739-1748.

[6] Panaligan R. Quantifying Movie Magic with Google Search [EB/OL]. (2013-05-18). http://www.tuicool.com/articles/mei2Qf.

[7] Asur S, Huberman B A. Predicting the Future with Social Media, HPL-2010-53 [R]. Hewlett Packard Company, 2010.

[8] Sadikov E, Parameswaran A G, Venetis P. Blogs as Predictors of Movie Success [C]//Proceedings of the 3rd International Conference on Weblogs and Social Media. New York, USA: ACM Press, 2009: 304-308.

[9] Mishne G, Glance N. Predicting Movie Sales From Blogger Sentiment [C]//Proceedings of 2006 Spring Symposium on Computational Approaches to Analyzing Weblogs. Menlo Park, USA: AAAI Press, 2006: 301-304.

[10] Jansen B J, Zhang Mimi, Sobel K. Twitter Power: Tweets as Electronic Word of Mouth [J]. Journal of the American Society for Information Science and Technology, 2009, 60(11): 2169-2188.

[11] Nielsen. Nielsen Twitter TV Ratings [EB/OL]. (2013-10-07). http://www.nielsen.com/us/en/press-room/2013/nielsen-launches-nielsen-twitter-tv-ratings.html.

编辑 陆燕菲