文章编号:1001-5132（2008）03-0341-05

博客数据分析系统的设计与实现

黄丽丽，陈华辉*

（宁波大学 信息科学与工程学院，浙江 宁波 315211）

摘要：博客(Blog)网站作为近年来新型的网络媒体得到越来越多的个人和企业的关注，因此针对 Blog 数据设计了相应的数据分析系统. 并介绍了 Blog 数据分析与传统 Web 挖掘的区别和联系，阐明研究了 Blog 数据的必要性，其次详细分析了本系统的主要功能模块及其实现方法，最后采集中国博客网的数据对本系统进行验证，实验结果显示本系统是可行且有效的.

关键词：Web 结构挖掘；博客数据分析；链接挖掘

中图分类号：TP393 文献标识码：A

Blog 是近年来涌现出来的新的网络沟通工具. 据《2006 年中国博客调查报告》表明，我国 blogger 的规模已达到 1 750 万，用户规模较 2002 年增长了 30 多倍. 如何挖掘和利用 Blog 中有价值信息，是我们研究 Blog 的重要课题.

Blog数据分析则是一个新兴的研究领域，国内对Blog的研究工作主要还在传播学的基本框架下进行. 而国外对Blog相关研究工作早于国内，研究也更深入些. Kumar等人主要研究超链接关系Blog社区的兴起和演化[1] ;日本东京大学Ishida Kazunari等采用WP(Weakest Pair Algorithm)方法发现Blog中潜在的社区[2]；而NEC实验室Tatemura等人通过RSS Feeds抓取Blog数据，来获取Blog中讨论的主题[3].

国内外关于Web挖掘的研究已有相当一段时期了[4-6]. Web结构挖掘的研究为本系统开发提供了一定的基础，但是传统Web挖掘是针对一般的Web页面，而不是Blog数据. Blog数据分析与传统Web挖掘存在一定相似性，但又有较大的区别.

(1) Blog 通过 post 及对 post 的评论和链接形成某个主题的信息串.

(2) Blog 提供对 post 的附加评论和唯一的 URL 指定功能，使得 Blog 空间中相似度较高的信息可以通过 Blog 中的引用自主地连结在一起，从而形成局部小社区，因此称之为 Blog 社区(一般由几个到几十个 blogger 组成). 当出现共同兴趣话题时，这些社区逐渐兴起，而发展到一定阶段后又将慢慢地消失. 但传统 Web 中这种特征不明显.

(3) 传统Web研究均是通过网络蜘蛛获取Web上的静态数据，但关于抓取静态数据之前的数据则不能得到. 然而 Blog 数据分析可以通过将 blogger 的每条 post 都和某个特定时刻关联，确定每条 post 和链接被创建的精确时间.

本文融合现有数据挖掘、数据库及人工智能等技术，结合上述的 Blog 特性，设计并实现了相应的 Blog 数据分析系统.

收稿日期：2007-04-28. 宁波大学学报（理工版）网址：http://3xb.nbu.edu.cn
第一作者：黄丽丽（1983－），女，浙江丽水人，在读硕士研究生，主要研究方向：数据挖掘. E-mail: mokaly@163.com
*通讯作者：陈华辉（1964－），男，浙江宁波人，副教授，主要研究方向：数据挖掘. E-mail: chenhuahui@nbu.edu.cn

1 系统的基本模块

本系统对 Blog 站点进行抓取、提取其中的关键信息，并且对这些信息进行分析的同时，根据 post 的相互引用来研究 Blog 空间中占主导地位的 blogger.

本系统主要分为 4 个模块：数据抓取、数据预处理，数据分析和数据统计. 数据抓取模块抓取某个 Blog 站点数据；预处理模块将已抓取的 Blog 站点数据进行分析，并提取其中有价值信息存入数据库；数据分析模块对已分析的 Blog 站点数据进行 Blog 社区的兴起和演化识别，并分析社区的变化趋势、识别 Blog 社区中主要 blogger 等；数据统计模块根据已抓取的 Blog 站点数据，分析 Blog 空间中的热门 post、活跃 blogger 和 commenter 等等.

2 系统的具体实现

2.1 数据抓取

本模块利用网络蜘蛛技术抓取 Blog 站点数据，将获取的信息存入数据库；并以 1 组 URL 为起点，下载其网页源代码，从中提取新的 URL 地址；然后重复以上分析和提取 URL 地址过程，直到满足一定条件即终止数据抓取. 由于 Blog 站点信息不断在变化，因此需要将已抓取 Blog 站点数据进行更新. 本系统采用增量数据更新，即重新数据抓取更新数据时，只抓取该站点新的 post 信息.

2.2 数据预处理

本模块对已抓取的 Blog 站点数据进行分析，提取 blogger 所发表的 post、该 blogger 所推荐的友情链接以及对该 post 的评论等信息，将这些信息存入数据库，表结构设计见表 1～表 4.

表 1 name 表结构

列名	含义
id	序号
blogger	作者
link	主页地址
crawled	抓取标志

表 2 post 表结构

列名	含义
id	序号
blogger	作者
date	发表时间
title	标题
count	评论数

表 3 comment 表结构

列名	含义
id	序号
reader	评论者
date	评论时间

表 4 f_link 表结构

列名	含义
blogger	作者
f_blogger	友情链接的作者

2.3 数据分析

2.3.1 社区识别

由于 bloggers 的某种共同偏好，使得 bloggers 相互间产生了紧密的联系，当这种联系达到一定程度时，某个潜在的社区便产生，我们称之为 Blog 社区.

传统Web挖掘中对社区识别也有研究[7-10]，但这些研究均是针对Web社区的，不能直接应用于Blog社区. 本系统社区识别的主要思想如下：首先构造评论矩阵 A (其元素 a_{ij} 表示第 j 个blogger对第 i 条post的评论条数)；其次利用矩阵 A 构造同评矩阵 C (其元素 c_{ij} 表示第 i 个评论者和第 j 个评论者共同评论的post条数)；接着对矩阵 C 进行行列变换，把共同评论多的评论者聚集在一起；最后对行列变换后的矩阵进行二分分割. 第 1 次分割将全部blogger分成 2 个大的社区，递归地进行 m 次二分分割，将全部blogger划分成若干个分层次的社区，即实现了社区的识别. 社区的大小可通过限制每个社区中blogger的人数来控制，也可以通过计算社区的模块度来控制.

2.3.2 社区演化分析

Blog 的每个 post 提供了明确的时间标识和对该话题进行评论的 link，当 Blog 中出现大家都感兴趣的话题时，社区开始形成，也就是 2.3.1 节中讨论过的社区出现了. 随着话题讨论的进度，以前的参与者可能不再关心该话题，同时可能也会有新的成员参与该话题的讨论，因而社区中的成员会有离开和新加入. 当这个话题讨论到某一时刻，话题有可能会失去吸引力，从而社区就会慢慢的消失，这就是所谓的社区演化.

本功能模块主要针对某个特定话题的社区，跟踪它从产生到结束的过程中，社区成员以及社区大小的变化情况. 同样可以利用 2.3.1 节中分析得出的社区，观察它们之后社区成员的迁入与迁出情况，同时也可以观察社区中成员数目的变化情况.

2.3.3 重要博客识别

通过分析 Blog 空间中 comment 和 post 之间的评论与被评论关系，可以发现某个社区中存在处于重要地位的博客人物. 重要博客人物可分为重要 Blog 作者(称为重要 blogger)和重要 Blog 评论者(称为重要 commenter). 所谓重要 blogger 就是他所发表的 post 引起其他众多 blogger 的评论，且参与评论的 commenter 的重要性越高，该 blogger 的重要性也越高. 所谓重要 commenter 就是他评论了较多的 post，且他评论的 post 的作者的重要性越高，该 commenter 的重要性也越高.

参考HITS算法中对网页排序的思想[11]，对重要blogger和重要commenter可用下列参数来衡量：

(1) post 的 blogger 重要性 *IPB*(Importance of Post for Blogger)，其定义为：$IPB = \sum IC$.

(2) post 的 commenter 重要性 *IPC*(Importance of Post for Commenter)，其定义为：$IPC = \sum IB$.

(3) blogger 重要性 *IB*(Importance of Blogger)，其定义为：$IB = \sum IPB$.

(4) commenter 重要性 *IC*(Importance of Commenter)，其定义为：$IC = \sum IPC$.

其中 *IPB* 表示某个 post 的 blogger 重要性；*IC* 表示参与该 post 评论的全部 commenter 的重要性；*IPC* 表示某个 post 的 commenter 重要性；*IB* 表示发布该 post 的 blogger 重要性. 初始设全部 blogger 的 *IB* 和全部 commenter 的 *IC* 均为 1 个小的常数，通过迭代计算，得到各 *IB* 和 *IC* 的最终值，其值的大小分别确定各 blogger 和各 commenter 的重要性.

2.4 数据统计

本模块功能是分析 Blog 站点数据，并做相关的典型统计分析，如统计热门 post、活跃 blogger、活跃 commenter、热门的 blog 推荐以及分析活跃的 blogger 发布和评论 post 的情况.

所谓热门 post 就是 Blog 站点中大家都比较感兴趣的 post，本系统中规定 post 所拥有的评论数大于某个阈值的则为热门 post. 活跃博客包括活跃 blogger 和活跃 commenter.

所谓活跃 blogger 指发帖勤快的 Blog 作者，本系统中规定发布 post 的数量超过一定阈值的则为活跃 blogger. 同样活跃 commenter 就是评论 post 的数目超过一定阈值的. 同时还统计活跃 blogger 发布 post 的情况以及活跃 commenter 评论 post 的情况.

Blog 空间中提供友情链接的 blogger 都是该站点 Blog 作者所推荐的. 而热门 Blog 推荐是说被推荐次数大于某个阈值的就认为是热门 Blog 推荐.

3 实验结果和评价

整个 Blog 数据分析系统用 Java 实现，后台采用 SQL Server 2000 数据库. 本文选取中国博客网(http://www.blogcn.com)上的数据对系统进行测试.

测试环境：DELL PC 机，配置如下：3.00 GHz Pentium(R)4 CPU，1 G 内存，80 G 硬盘，操作系统 Windows XP.

选取博客目录页面为种子页面，设抓取深度为 3，由此所得数据量为 5.57 G. 经过数据预处理最后

将有效信息存入数据库，分别保存在 name 表、post 表、comment 表以及 f_link 表中，其中 name 表含记录条数 42 219 条，post 表 73 124 条，comment 表 154 352 条，f_link 表 20 512 条. 由于篇幅限制，在这里主要给出社区分析的实验结果.

在实验中，采用文献[12]提出的模块度思想来衡量社区识别的有效性. 针对社区识别的需要，按照模块度的思想，构造了 2 个指标.

第 1 个指标为总体模块度*TM*(Total Modularity)，它反映Blog站点中划分出的所有社区的有效性. *TM* 定义为：

$$TM = \sum_i e_{ii} - \sum_{ijk} e_{ij}e_{ki} = Tre - \left\| E^2 \right\|,$$

其中 e_{ij} 为矩阵 E 中的元素，表示第 i 个社区和第 j 个社区之间相互联系的边占总边数的比例，Tre 表示矩阵 E 的迹，$\left\| E^2 \right\|$ 表示矩阵 E 平方之后的矩阵的所有元素之和. 如果 *TM* 越大，说明社区内部的联系越紧密，而社区之间的联系越少. 因此 *TM* 越大说明社区划分越有效. *TM* 可用来评价社区总体划分的有效性.

第 2 个指标为社区模块度 *CM*(Community Modularity)，它反映了社区内部每个成员的紧密程度. 对某个社区，其 *CM* 的定义如下：

$$CM = \sum_{ij} c_{ij} / c_{sm},\ i \neq j,$$

其中 c_{ij} 表示该社区中的 blogger 所组成的同评矩阵的元素；c_{sm} 表示该社区包含的 blogger 人数，因而 *CM* 为该社区中共同评论数的平均值. 该值越大，说明社区中成员的共同兴趣越一致，该社区的划分越合理. *CM* 可用来衡量某一社区划分的有效性.

本系统中采用 *TM* 值来决定实验中某社区的最少人数，统计如图 1 所示. 由图 1 中可以看出 1 月份社区的最少人数为 10 人时，*TM* 值最大，也就是说每个社区的最少人数为 10 人时，产生的社区效果最好. 因此，根据社区的最少人数 10 人对 1 月份所产生的社区进行划分，同样可看出 2 月份社区最佳最少人数为 11 人，3 月份社区的最佳最少人数为 15 人，4 月份社区的最佳最少人数为 33 人.

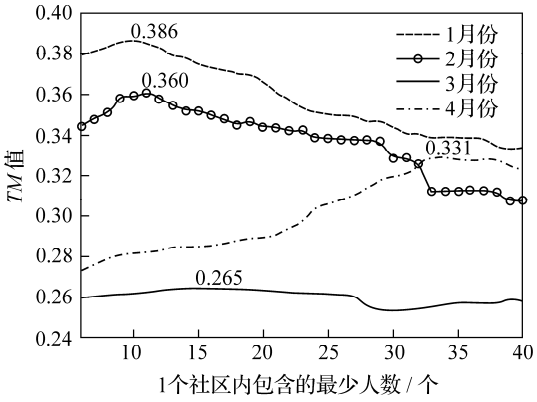

图 1　2006 年 1 至 4 月份的 *TM* 值

根据上面实验获得的数据对 2006 年 4 月份中国博客网所产生社区进行划分，共产生社区数为 54 个，其每个社区的 *CM* 值如图 2 所示. 由图 2 可知，每个社区的 *CM* 值均大于 1，说明社区中的评论者同评论的 post 平均大于 1 条；第 24 个社区的 *CM* 最大，其值为 20.05，说明该社区的成员关系比较紧密. 观察组成第 24 个社区的成员，这个社区大小为 70，这些成员主要对 xinxinaini、cathy124、weixiangmei8888、cuipidoufu1、qitongrenli、sunshine 780710 以及 nanaduo 这些 blogger 所发的帖子进行评论而聚集在一起，其讨论主题为情感和生活. 可见这 2 个主题是该社区中大多数 blogger 都比较感兴趣的主题.

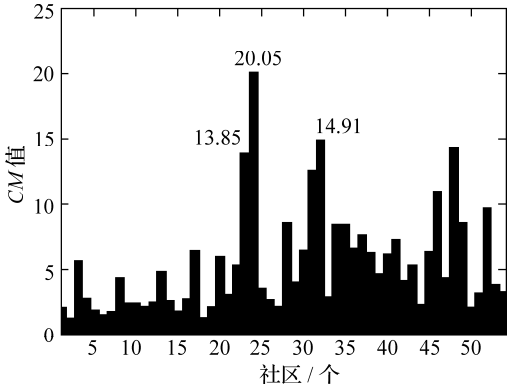

图 2　2006 年 4 月份每个社区 *CM* 值

4　总结

Blog 数据分析需要结合多个研究领域的方法，

目前该研究正处在起步阶段. 本文针对 Blog 数据，实现了 1 个分析系统，重点讨论了其实现，并将其应用于中国博客网的 Blog 数据分析中.

实验结果反映出一些尚须解决的问题，例如在对社区进行识别时，部分社区的 *TM* 值较低. 今后工作除了在本系统的基础上进行扩展，使得社区的识别更加完善外，另外将考虑如何将基于内容的挖掘融合到系统中.

参考文献：

[1] Kumar R, Novak J, Raghavan P, et al. On the bursty evolution of blogspace[C]//Proc of WWW, 2003:569-576.

[2] Ishida K. Extracting latent weblog communities- A partitioning algorithm for bipartite graphs[C]//Workshop on the Weblogging Ecosystem, 13th International World Wide Web Conference, 2004.

[3] Nakjima S, Tatemura J, Hino Y, et al. Discovering important bloggers based on analyzing blog threads[C]//WWW 2005 Workshop on the Weblogging Ecosystem, 2005.

[4] 宫秀军，史忠植. 基于 Bayes 潜在语义模型的半监督 Web 挖掘[J]. 软件学报, 2002, 13(8):1 508-1 514.

[5] 冯雁，王申康. Web 站点层次结构抽取算法的分析和实现[J]. 浙江大学学报: 理工版, 2005, 39(10):1 507-1 511.

[6] 张克君，李伯群，李欣. 基于 DWLMS 模型的分布式 Web 用户访问模式挖掘[J]. 清华大学学报: 理工版, 2005, 45(S1):1 762-1766.

[7] Brin S, Page L. The anatorny of a Large scale hypertextual web search engine [J]. Computer Networks and ISDN Systems, 1998, 30(1-7):107-117.

[8] Kleinberg J M. Authoritative source in a hyperlinked environment[C]//Proc of the 9th ACM-SIAM Symposium on Discrete Algorithm, 1998:668-677.

[9] Kumar R, Raghavan P, Rajagopalan S, et al. Trawling the web for emerging cyber-communities automatically[C]// Proc of the 8th ACM–WWW International Conference, 1999:1 481-1 493.

[10] Flakf G W, Lawrence S, Giles C L, et al. Self-organization of the Web and identification of communities[J]. IEEE Computer, 2002, 35(3):66-71.

[11] Kleinberg J M. Authoriative sources in a hyperlinked environment[J]. Journal of the ACM, 1999, 46 (5):604-632.

[12] Newman M E J, Girvan M. Finding and evaluating community structure in networks[J]. PhySical Review E, 2004, 69:026113.

Design and Implementation of a Blog Data Analysis System

HUANG Li-li, CHEN Hua-hui*

(Faculty of Information Science & Engineering, Ningbo University, Ningbo 315211, China)

Abstract: As a new-type network media, blog websites are becoming more and more popular and drawing more and more attention in recent years. With Blog data as the research object, a design of data analysis system is presented in this paper. The distinction and affiliation of Blog data analysis with conventional data mining are described. The implementing approaches are presented and the main function modules of the proposed system are introduced in detail. In the end, the developed system is validated using the data collected from blog website in China, and the results suggest that the system is both feasible and efficient.

Key words: web structure mining; blog data analysis; link mining

CLC number: TP393　　**Document code:** A

（责任编辑　章践立）