中文博客标签的若干统计性质*

刘知远[1]，司宪策[2]，郑亚斌[3]，孙茂松[4]
(中国北京市清华大学计算机科学与技术系
清华信息科学与技术国家实验室(筹)
智能技术与系统国家重点实验室 100084)
[1]liuliudong@gmail.com，[2]adam.si@gmail.com, [3]yabin.zheng@gmail.com, [4]sms@tsinghua.edu.cn

摘　要：随着 Web2.0 理念日益深入人心，博客作为一种网络日志的形式，成为网络上的主要应用之一。而主要出现在博客、网络相册等系统上的，依靠大量用户使用自由选择的词汇作为标签(Tag)来对事物进行标记的人工分类的“大众分类法”也逐渐成为研究热点。本文将焦点集中在中文博客标签上，着重考查其统计性质、齐夫定律和复杂网络性质，从多方面初步了解中文博客标签的性质和特点。

关键字：博客，标签，齐夫定律，复杂网络

Statistical Properties of Tagging System on Chinese Blogs

Liu Zhiyuan[1], Si Xiance[2], Zheng Yabin[3], Sun Maosong[4]
(State Key Laboratory on Intelligent Technology and Systems
Tsinghua National Laboratory for Information Science and Technology
Department of Computer Science and Technology, Tsinghua University, Beijing 10084, China)
[1]liuliudong@gmail.com，[2]adam.si@gmail.com, [3]yabin.zheng@gmail.com, [4]sms@tsinghua.edu.cn

Abstract: As the thoughts of Web2.0 are accepted by people all over the world, weblog, as an online log, is one of the most widely used web applications. Tags, appearing in weblogs, online albums, etc., with which users assign to mark different affairs and things, named as “folksonomy”, are focused more recently by many researcheres. In this article, we focus on the tags of Chinese blogs. We observe and analyse some interesting statistical properties, Zipf’s Law and properties of complex network.

Keywords: blog, tag, Zipf’s law, complex network

1. 介绍

博客是一种日志性质的网站，主要由按新旧顺序排列的带有日期的文章及对应的评论组成。不同的博客之间通过链接、评论和反向链接互相联系，带有明显的社区性质[1, 2] 。博客反映了作者群体的观点和生活，并且对网络和现实的世界都有一定的影响，虽然说博客的作者群体存在偏向性，但是其中还是蕴含了重要的信息[3, 4] 。经过对大量博客的索引和挖掘，可以得到例如博客群体对商品和公司的意见、广告投放的效果、博客群体对某些事件的关注程度和看法等等对商业和社会有价值的信息[5, 6] 。这里值得特别注意的一个要点是 “大众分类法”(Folksonomy，意谓 Folk 的 Taxonomy)在博客中的应用。这种主要出现在博客、网络相册和网络书签系统上的、依靠大量用户使用自由选择的词汇作为标签(Tag)来对事物进行标记的人工分类方法，被称为“大众分类法”。这种分类方法显然和传统的基于少数专家的分类体系不同，每个分类者的自由度很大，通常也没有层次关系[7] 。现在比较流行的英文自由分类法有 del.icio.us 网络书签和 Flickr 网络相册中使用的体系，中文自由分类法的代表有豆瓣网(www.douban.com)等。

与大众分类法在网络上的大行其道相比，对其结构性质、动态演化过程和机理的研究仍然处在刚刚起步阶段。本文将对中文博客网站的标签体系的统计性质和演化模型进行研究，以期对中文博客的标签标注机理有更为深入的了解。

2. 前人工作

随着博客、网络相册和网络书签系统等 Web2.0 应用的大量涌现，“大众分类法”也开始受到越来越多研究者的关注。对大众分类法的研究中，最近提出了一个话题一直争论不休，从而成为研究热点，那就是如何解释大众分类法的演化模式并为之建模。实际上，大众分类法按照标签标注方式可以分为两种：协同标注(Collaborative Tagging)和非协同标注。前者的特点是，所有用户可以对同一个资源进行标注，如 del.icio.us 的用户可以对同一 URL 标注自己认为有用的标签；而后者则只能是资源发布者才能够对资源添加标注，如博客系统中只有发布文章的用户能够为文章添加标签。目前相关的工作都是在协同大众分类体系上进行的：

Cattuto [8] 对来自 del.icio.us 和 connotea 的大众分类体系进行了分析，发现与某一标签同

*本研究承国家自然科学基金(项目号 60573187, 60621062 和 60520130299)资助。

现的其他标签的 Frequency-Rank 大致遵守齐夫定律，但在低频处的斜率较小。作者基于“用户的标注行为对其它用户公开”和“标签的时间性质”建立了描述标签标注行为模型。Schmitz [9] 分析了 del.icio.us 和 BibSonomy 上(tag, user, resource)三元组构成三部图(tri-partite graph)的复杂网络性质，发现其具有小世界性质。Cattuto [10] 分析 del.icio.us 标签增长特性发现，无论整体还是针对某一特定资源的标签，时间与标签数总呈幂律分布，斜率小于 1。作者还发现，不同资源的标签随时间的增长性质具有很强的一致性。Halpin [11] 发现 del.icio.us 用户对流行网站 URL 添加的标签呈幂律分布。作者通过优先添加模型较好地拟合了真实数据。

而对博客等非协同大众分类体系，研究工作相对较少，尤其是对其演化模式的研究，更是凤毛麟角。本文将在中文博客的大众分类真实数据上展开研究，对其结构性质、动态演化模式进行初步的探讨。

3．统计性质

3.1 数据集

本文使用的博客数据集是通过我们设计的一个聚焦抓取系统(focused crawling)来实现的，该抓取系统可收集并定期查看已经发现的博客所发布的 RSS 列表，下载其中出现的新文章，并且收集博客文章中附带的标签信息。整个系统在无人值守的情况下连续运行，并且只关注中文博客。我们已利用此系统发现了 48,647 个博客站点，包括 362,889 篇文章，其中标签出现的总次数为 890,935，互不相同的标签 129,001 个。

3.2 统计信息

图 1 显示了数据集中标签按照所含字数的分布情况。可以看到，标签多数为 2 字词；从频度最高的十个标签全为 2 字词也可窥豹一斑，这与现代汉语以多字词为主的现象相符。在该数据集中，频度排名前 10 位的标签为：爱情，日记，生活，情感，女人，心情，男人，希望，妈妈，公司。可以发现，这些经常被用来作为标签的词，其语义具有一般意义，这种词汇具有更强的概括能力，因此会被更频繁地使用。图 2 显示了数据集中每次标注按照标签数目的分布情况。每次标注的标签数目，以 1 个最多，但标注 2~5 个的情况也较多，这使研究标签同现网络成为可能。表 1 显示了“爱情”，“日记”，“生活”，“情感”等四个指定标签的同现统计信息。

表 1：指定标签的同现统计信息

标签	使用该标签的文章数	与该标签同现的标签次数	与该标签同现的标签种类
爱情	9,066	11,481	3,220
日记	7,839	6,816	1,147
生活	7,769	12,726	2,494
情感	7,569	5,635	1,249

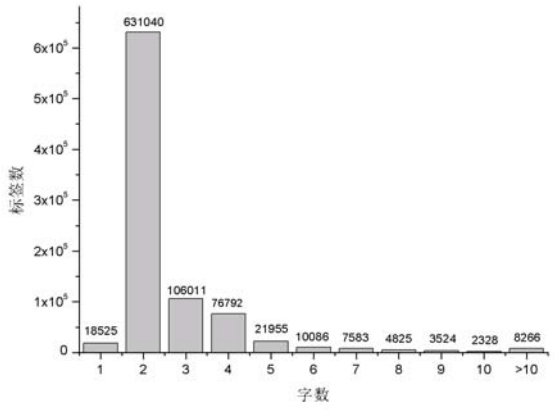

图 1：数据集中标签按照标签字数的分布图

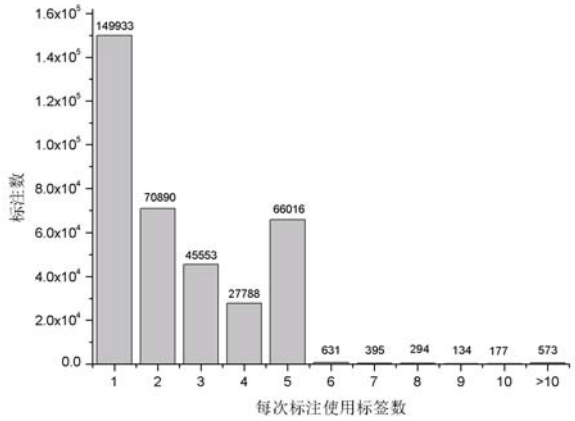

图 2：数据集中标注按照标注中的标签数目的分布图

对标签按照频度由高到低排序，前 N 个标签及其频度覆盖率如表 2 所示。如前所述，互不相同的标签数为 129,001 个，而标签频度高于等于 6 的 15,845 个标签已经覆盖了标签使用总频度的 80%以上。如图 3 所示，是频度高于等于 6 的 15,845 个标签按照所含字数的分布情况，可以发现，与图 1 相比，在覆盖率较高的这些标签中，虽然 2 字词仍然占主要地位，但是 3 字词和 4 字词的比例明显增大。

表 2：按频度由高到低排序的标签数目及其频度覆盖率对应关系

标签频度	标签数目(N)	覆盖率(%)
≥ 6	15,845	81.93
≥ 5	18,482	83.41
≥ 4	22,274	85.11
≥ 3	28,689	87.27
≥ 2	41,797	90.21

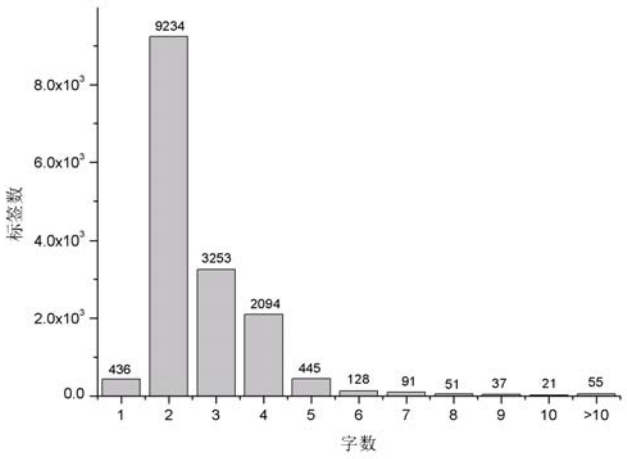

图 3：频度高于等于 6 的标签按照标签字数的分布图

3.3 齐夫定律

如图 4 所示，是所有标签的频度统计的排名分布图；图 5 则显示了指定标签的同现标签的频度统计排名分布图。可以明显看出，分布基本符合齐夫定律[12] 。然而，也可以发现在排名较高的部分相对平坦，这主要是有两个原因[8] ：(1)语义相近或重叠的常用词语会在使用上存

在竞争关系，如“情感”和“心情”之间就存在这种关系。(2)标签在语义上存在潜在的层次结构，对于更为通用的标签如“爱情”、“生活”相对于“杜鹃”等会更多地与其他标签搭配出现。

每次标注中的任意两个标签之间存在同现关系。图 6 是对同现的两个标签(可以称为 BiTag)按照频度统计得到的排名分布图。可以看到也明显基本符合齐夫定律。图中还标注了频度最高的几个同现标签。

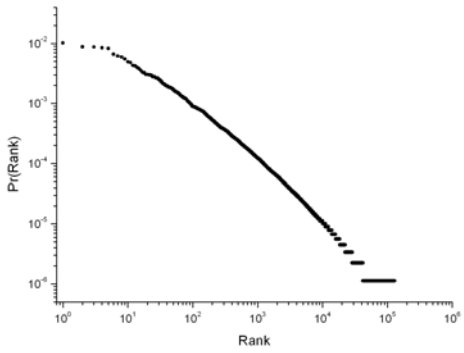

图 4：所有标签的频度排名分布图

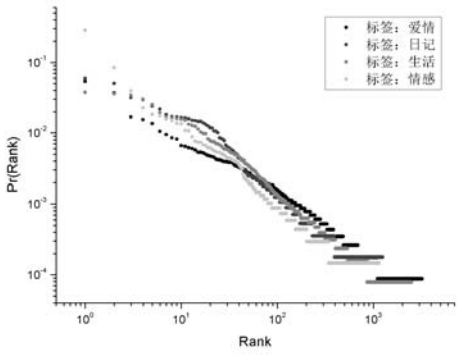

图 5：选定标签的同现标签频度排名分布图

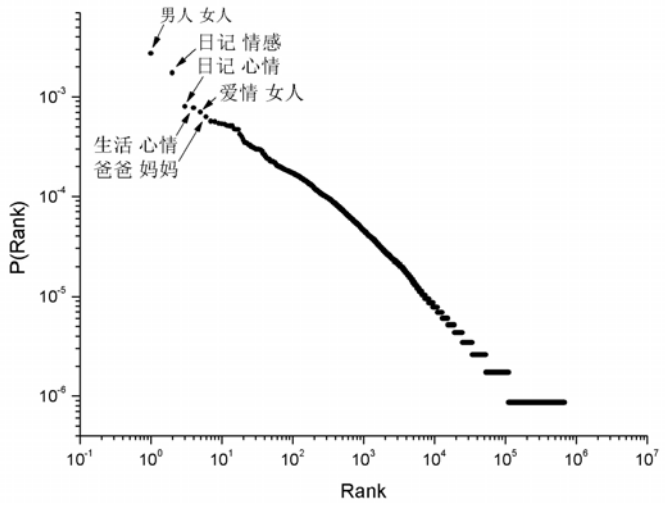

图 6：同现两个标签的频度排名分布图

3.4 复杂网络性质

最近复杂网络的研究取得了引人注目的成果。研究发现各领域中有着不同拓扑结构的复杂系统，如 Internet、食物网以及社会网络等，都表现出相似的统计规律。首先是复杂网络的小世界效应，即网络的聚集程度相对随机网络明显较高，同时平均最短路径比较小[13] 。此外，复杂网络表现出无标度特性，即节点连接度呈幂律分布[14] 。

大众分类体系由于各种关系(如同现关系等)也构成规模巨大的复杂网络。然而目前对于大众分类体系复杂网络性质的研究还刚起步[8, 9] ，其中对非协同的大众分类体系的研究，就作者了解，到目前还未相关工作发表。本文将在中文博客上的标签同现网络进行这方面的尝试。

设网络节点个数为 N ，边数为 E 。以下是复杂网络的三个重要参数。

平均最短路径长度。网络中两节点之间的平均距离。具有小世界性质的网络的平均最短路径会很短，远小于网络规模[13] （这也是“小世界”命名的原因）。设平均最短路径为 d ，网络节点平均度为 $\overline{k}$ ，对“小世界”网络，则有 $d \approx \ln(N)/\ln(\overline{k})$ 。

聚合系数。一个节点的聚合系数反映了其相邻节点所构成集合的聚集程度。整个网络的聚合系

数 C 是每个节点 i 的聚合系数 C_i 的平均值($0 \le C \le 1$)。对一个包含 N 个节点的ER随机图网络，当 N 很大时，有 $C \approx \bar{k}/N$，即其聚合系数远小于1。而大规模的实际复杂网络表现出显著的聚合效应[15-17] 。

节点连接度分布。大量研究表明，实际复杂网络的度分布明显不同于 Poisson 分布，而更接近于幂律分布（无标度分布），即 $\Pr(k) \propto k^{-\gamma}$，其中 $\Pr(k)$ 是度为 k 的节点出现在网络中的概率，γ 为常数。

设标签同现网络节点数为 N，网络边的条数为 E，网络节点的平均连接度为 $\bar{k}$，平均最短路径长度为 d，同等规模随机网络的平均最短路径长度为 d_{random}，聚合系数为 C，同等规模随机网络的平均最短路径长度为 C_{random}，这些统计参数值列在表 3 中。可以发现标签同现网络的 d 很小，而聚合系数 C 相对于 C_{random} 较大，表现出复杂网络的小世界性质。如图 8 所示，标签同现网络的最短路径长度集中分布在 3、4 数值上，也就是说，从一个标签到任意另外一个标签，平均只需要 3 到 4 跳，这对研究用户进行标签标注的行为模式具有一定的启发意义。

表 3：标签同现网络各参数值

参数	标签同现网络
N	129,001
E	680,367
$\bar{k}$	10.5
d	3.5
d_{random}	5.0
C	0.4
C_{random}	8.1×10^{-5}

如图 7 所示，标签同现网络的累积度分布大致呈幂律分布，即 $\Pr(k) \propto k^{-\gamma}$，对曲线拟合得到 $\gamma \approx 2.28$。表现出复杂网络的无标度性质。

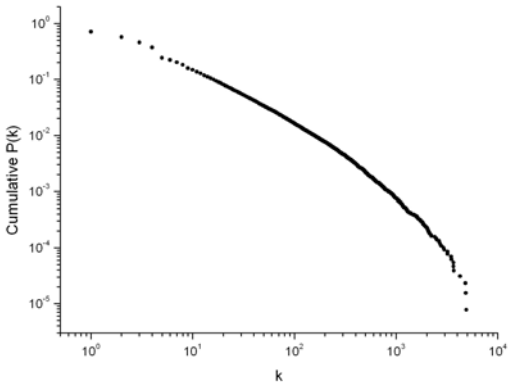

图 7：标签同现网的累积度分布曲线

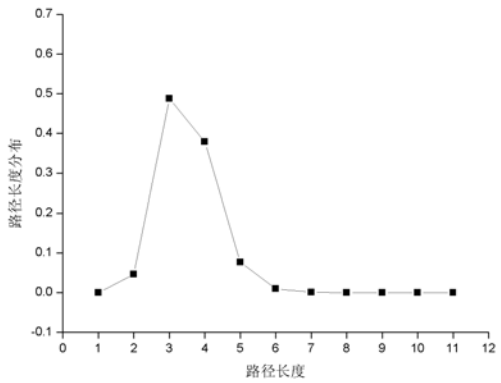

图 8：标签同现网的最短路径长度分布曲线

4．结论与展望

本文从齐夫定律、复杂网络等几个方面对中文博客标签数据集进行考查，初步揭示了非协同标注的“大众分类法”的一些有意义的性质。这些性质为接下来更深入的工作提供实证基础。在此基础上，我们可以开展以下工作：(1)分析用户标注的行为模式，为“大众分类法”的演化建模，用于预测未来趋势；(2)通过标签在标签网络中表现的性质，对其进行语义分类；(3) 分析 “大众分类法”与“专家分类法”的异同，并尝试两者的融合,等等。

参考文献

[1] Lin, Y., et al. Discovery of Blog Communities Based on Mutual Awareness. in WWW-2006 Workshop on the Webloging Ecosystem. 2006.

[2] Zhou, Y. and J. Davis. Discovering Web Communities in the Blogspace. in 40th Annual Hawaii International Conference on System Sciences (HICSS'07). 2007. Los Alamitos, CA, USA.

[3] Fukuhara, T., T. Murayama, and T. Nishida. Analyzing concerns of people using Weblog articles and read world

temporal data. in WWW-2005 Workshop on the Webloging Ecosystem. 2005.
[4] Gill, K. How can we measure the influence of the blogosphere? in WWW-2004 Workshop on the Webloging Ecosystem. 2004.
[5] Thelwall, M. Blogs During the London Attacks: Top Information Sources and Topics. in WWW-2006 Workshop on the Webloging Ecosystem. 2006.
[6] Avesani, P. Learning Contextualised Weblog Topics. in WWW-2005 Workshop on the Webloging Ecosystem. 2005.
[7] Mathes, A., Folksonomies-Cooperative Classification and Communication Through Shared Metadata, . 2004.
[8] Cattuto, C., V. Loreto, and L. Pietronero, Semiotic dynamics and collaborative tagging. Proceedings of the National Academy of Sciences (PNAS), 2007. **104**(5): p. 1461-1464.
[9] Schmitz, C., et al. Network Properties of Folksonomies. in Workshop "Tagging and Metadata for Social Information Organization" in the 16th International World Wide Web Conference (WWW2007). 2007. Banff, Alberta, Canada.
[10] Cattuto, C., et al., Vocabulary growth in collaborative tagging systems. oai:arXiv.org:0704.3316, 2007.
[11] Halpin, H., V. Robu, and H. Shepherd. The complex dynamics of collaborative tagging. in Proceedings of the 16th international conference on World Wide Web. 2007. Banff, Alberta, Canada: ACM Press.
[12] Zipf, G.K., Human Behavior and the Principle of Least Effort. 1949, Cambridge, MA: Addison-Wesley.
[13] Watts, D.J. and S.H. Strogatz, Collective dynamics of 'small-world' networks. Nature, 1998. **393**: p. 440-442.
[14] Barab, A.L. and AlbertRéka, Emergence of Scaling in Random Networks. Science, 1999. **286**: p. 509-512.
[15] Strogatz, S.H., Exploring complex networks. Nature, 2001. **410**(6825): p. 268-276.
[16] Jeong, H., et al., The large-scale organization of metabolic networks. Nature, 2000. **407**(6804): p. 651-654.
[17] Montoya, J.M. and R.V. Sole, Small world patterns in food webs. Journal of Theoretical Biology, 2002. **214**(3): p. 405-412.