基于用户-微博图分析的自适应 Twitter 用户排名

马凤龙，张宪超
（大连理工大学软件学院，大连 116621）

摘要：Twitter 是当今最流行的微博平台之一，本文的研究问题是 Twitter 上的权威用户排名。Twitter 具有大量的用户，并含有丰富的信息资源。权威用户经常发布有用的信息，因为有用的信息可以很快的传播。为了寻找权威用户，就要考虑到 Twitter 上的全部信息。然而，现有的算法仅仅考虑用户之间的关注关系，或者设定固定参数。本文提出了自适应 Twitter 用户排名，通过利用 PageRank 算法，对关注关系、发布关系以及转发关系之间的循环迭代来评估用户权威值。实验表明，该算法优于其他现有算法。

关键词：模式识别；用户排序；社交网络
中图分类号：TP181

Self-tuning Twitter User Ranking Based on User-Tweet Graph Analysis

Fenglong Ma, Xianchao Zhang
(Software School, Dalian University of Technology, Dalian 116621)

Abstract: In this paper, we focus on the problem of identifying authoritative users in Twitter , which is one of the most popular micro-blogging services. Owing to a large number of users, Twitter has been gaining a lot of information resource. In particular, authoritative users publish useful tweets usually, because useful information is disseminated quickly and widely. In order to find authoritative users, it is important to consider all the information in Twitter. However, existing algorithms only deal with following relationship among users, or set the parameters manually.In this paper, we propose the STURank (Self-tuning Twitter User Rank) algorithm, which employs PageRank algorithm and uses following relationship, publishing relationship and retweeting relationship. Experimental result shows STURank outperforms existing algorithms.

Key words: Pattern recognition; User Rank; Social Networks

0 引言

Twitter 是全球非常著名的微博服务平台，近年来注册用户急剧增长，发布的微博被称作 Tweet，用户可以随意关注（Follow）他们感兴趣的用户。用户可以转发（Retweet）感兴趣的微博，也可以在微博中提及（Mention）其他用户。

Twitter 中的用户种类繁多，例如普通的用户、公司、政府机构、新闻网站等。Java 等人[1]将用户划分为三类：（1）信息源：经常发布有用信息的用户；（2）朋友：这类用户是现实生活中的朋友、家人、同事等；（3）信息消费者：这些用户很少发布微博，但是关注大量用户，获取自己需求的信息。因此，在 Twitter 中寻找权威用户是一个巨大的挑战。

近些年来，有很多研究者已经展开了对 Twitter 中的权威用户排序的研究，Java[1]和 Weng[2]等人的研究主要是关注用户之间的 Follow 关系，然而在 Twitter 中存在两种实体：用

基金项目：教育部博士点基金 (20120041110046)；高等学校博士学科点专项科研基金（新教师类）项目（20100041120033）
作者简介：马凤龙（1986-），男，大连理工大学软件学院硕士研究生，主要研究方向：数据挖掘
通信联系人：张宪超（1971-），男，教授，主要研究方向：数据挖掘，信息检索，机器学习. E-mail: xczhang@dlut.edu.cn

户和微博，同时存在四种边：（1）用户与用户之间的关注边（Follow）；（2）用户发布微博（Publish）；（3）用户在微博中提及其他用户（Mention）；（4）用户转发其他用户的微博（Retweet）。因此，仅仅利用一种 Follow 关系是不能体现出 Twitter 的全部特性。Yamaguchi[3]的研究考虑到了用户之间的 Follow 关系，用户与微博之间的 Publish 关系以及微博之间的 Retweet 关系，但是他们的研究不能自动学习每一种边的权重，需要用户指定。所以，这种方法不能应用在实际生活中。

在本文中，我们提出了一种能够自动寻找 Twitter 中权威用户的方法，自适应 Twitter 用户排名（Self-tuning Twitter User Rank，简称 STURank）算法。STURank 算法不仅利用了 Twitter 中各种边信息，同时不需要用户输入指定的边权重，根据网络结构可以自动调节。

1 Twitter 结构分析

本文中所使用的数据为 2009 年 6 月到 9 月的 Twitter 数据，包括 5307392 个用户，178199070 条关注关系和 390797896 条 tweets。其中，含有转发（Retweet）的 tweets 49542429 条，含有@（Mention）的 tweets 118461782 条，包含 URL 的 tweets 137130347 条，以及含有#（Hashtag）的 tweets 40839189 条。

1.1 关注关系分布

图 1 和图 2 是关注者（Friend）和粉丝（Follower）的 log-log 尺度分布图。通过观察，我们可以发现关注者和粉丝的分布在某种程度上均服从幂律（power-law）分布，幂律分布在 Web 图中很常见[4]。

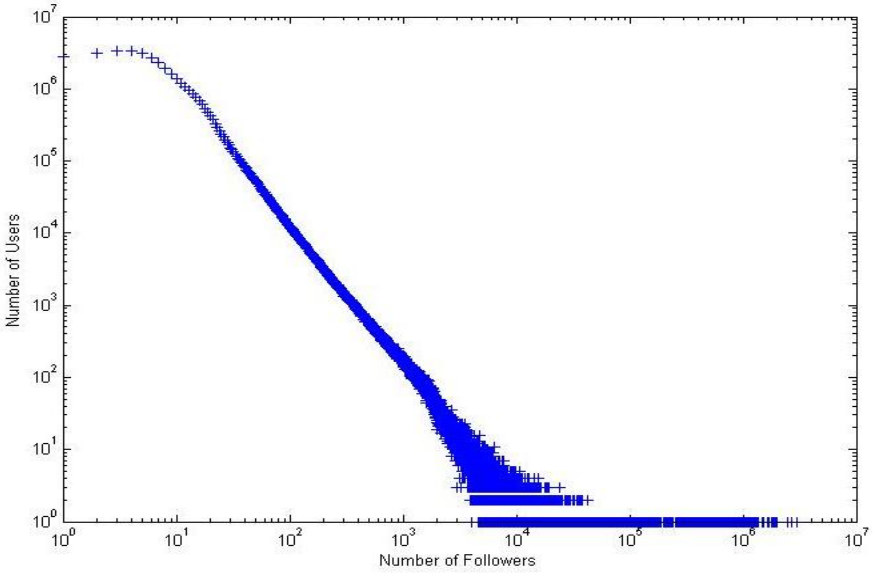

图 1　粉丝分布

Fig. 1　Follower Distribution

图 1 是粉丝的分布，即用户拥有的粉丝数，横轴代表粉丝的数量，纵轴代表拥有该粉丝数量的用户个数。通过观察图 1 可以发现，只有少部分用户的粉丝数量超过 100000，大部分的粉丝数低于 100。粉丝数高的用户都是社会中的名人、新闻机构等，如 aplusk，CNN 等。

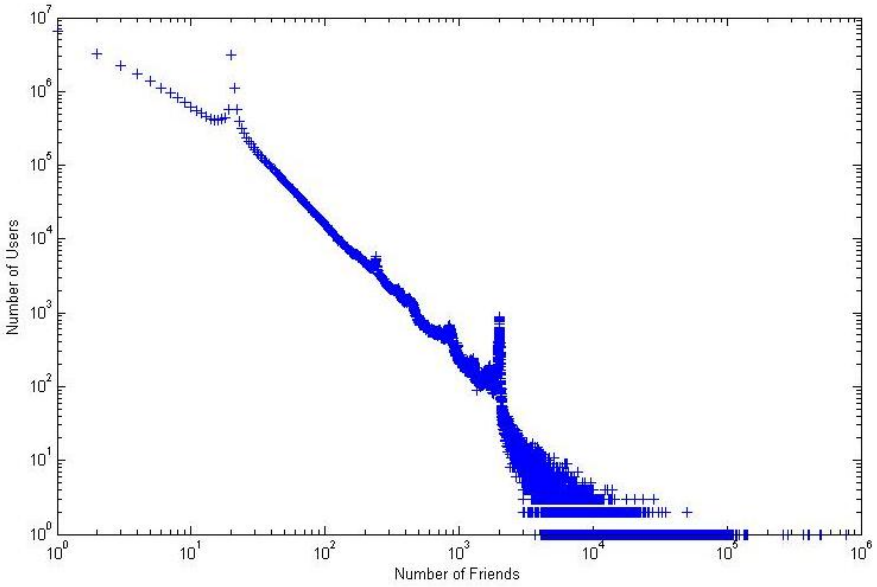

图 2　关注者分布

Fig. 2　Friend Distribution

图 2 是关注者的分布情况，粉丝关注不同的是，在 2000 处分布出现的小的变化，这是因为 Twitter 公司之前限制了用户的最大关注个数（2000），所以在图 2 中关注者为 2000 的人数出现了异常。通过分析图 2 可以发现，大部分用户的关注者数量在 20 左右，这是因为用户在注册 Twitter 账户的时候，系统会自动推荐 20 个用户给新用户。

1.2　转发关系分布

与图 1 和图 2 相似，图 3 和图 4 描述的转发关系的分布，其中图 3 是转发入度（Indegree）的分布，也服从幂律分布。与图 1 的尺度相比，转发入度的分布更接近于 Web 的真实分布。这是因为转发是用户根据自己的意愿主动发起的动作，与 Web 的超链接很相似：用户对 tweets 的内容感兴趣，并且对用户信任，才会产生转发行为，用户之间建立一条有向边。

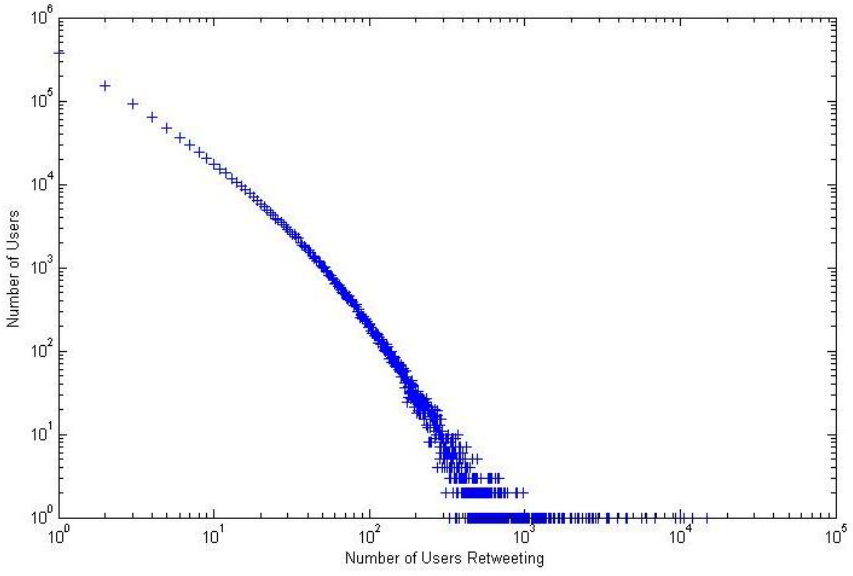

图 3　发起转发分布

Fig. 3　Retweeting Distribution

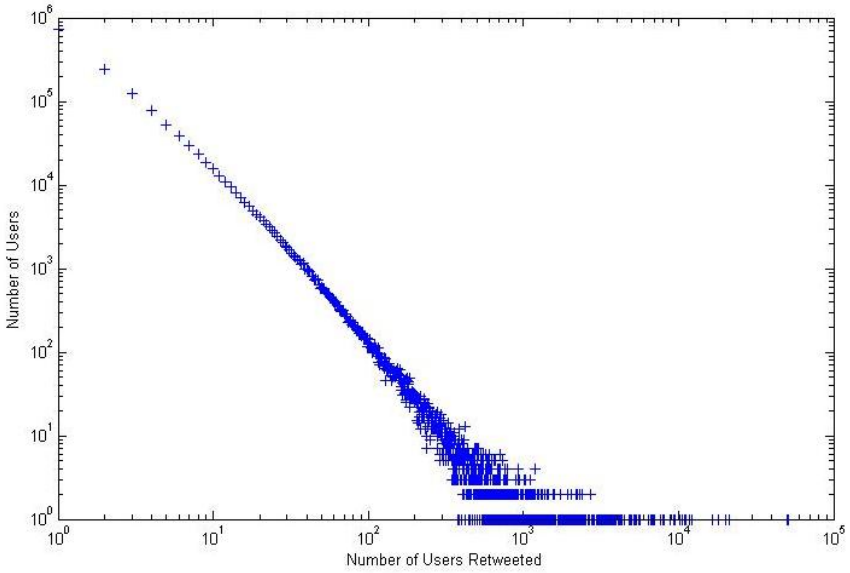

图 4　被转发分布

Fig. 4　Retweeted Distribution

1.3　Tweets 发布频率

Tweets 的发布频率可以反映出一个用户的活跃程度。如果用户经常发布一些重要的信息，那么他的粉丝可能急剧增长，与此相反，如果用户经常发布广告信息或者长时间不发布信息，那么他的影响力应该是下降的。

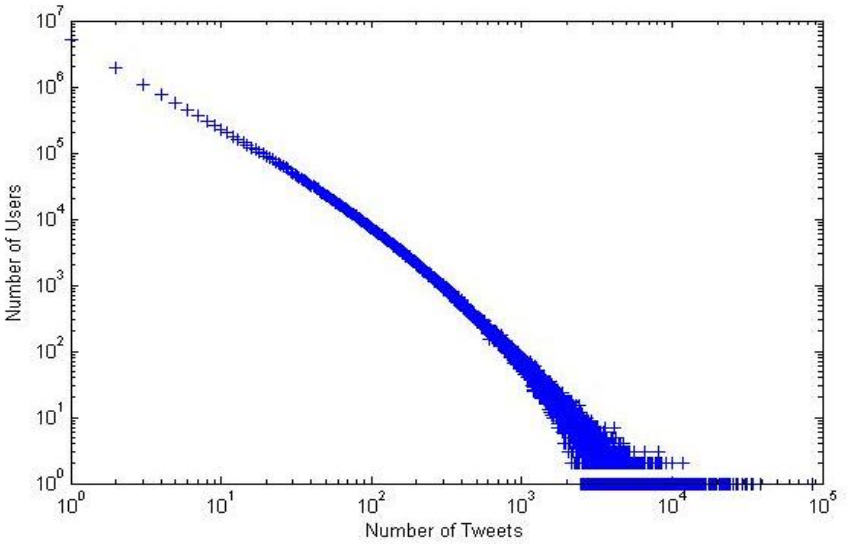

图 5　Tweets 发布频率

Fig. 5　Tweets Frequency

2　自适应 Twitter 用户排名算法

2.1　用户-微博图分析

在引言部分，我们已经介绍了 Twitter 中存在两种实体以及四种边关系，我们利用其中

的三种边（Follow，Publish 和 Retweet）来综合计算用户的权威性。我们的算法基于以下观察：

（1）如果用户的粉丝中，有很多权威用户，那么该用户很有可能是权威用户；

（2）如果用户的微博被很多权威用户转发，那么该微博的质量应该会很高；

（3）如果用户经常发布高质量的微博，那么该用户的权威性应该很好。

基于这些观察，我们建立了用户-微博图（Users-Tweets Graph）的简图，如图 6 所示，将用户和微博分为两层架构。在用户层，用户之间为 Follow 关系，在 Tweets 层，微博之间是 Retweet 关系，在两层之间是 Publish 关系，因此该用户-微博图模型可以表示 Twitter 的网络结构。

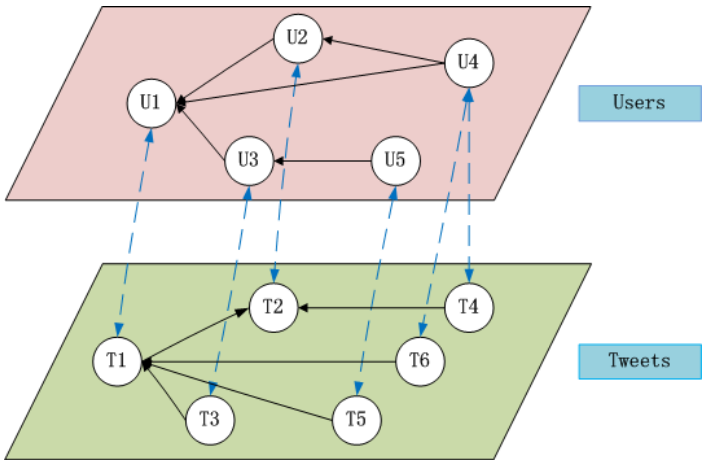

图 6　用户-微博图

Fig. 6　User-Tweets Graph

根据用户之间的 Follow 关系，我们可以利用 PageRank[5]算法计算用户的权威度，然后通过 Publish 边将用户的权威度平均分给用户所发布的微博，这样就可以将 Users 层的权威度转移到 Tweets 层。将 Users 层传下来的权威值，作为 Tweets 层的初始值，在 Tweets 层再利用 PageRank 算法进行迭代，用户的微博就会产生新的权威度，将该权威度相加在传递给 Users 层。经不断的循环迭代，就可以得到用户的权威度，同时也可以得到微博的权威度。下面我们将正式定义该算法。

2.2　自适应 Twitter 用户排名算法

令用户-微博图$UTG=(V,E)$，其中$V=\{V_U,V_T\}$，$E=\{E_F,E_P,E_R\}$，V_U 代表用户层的所有用户，V_T 代表微博层所有的微博，E_F 表示用户之间的 Follow 关系，E_P 表示用户层与微博层之间的 Publish 关系，E_R 代表微博层的 Retweet 关系。基于以上定义有：

（1）在用户层计算用户的权威值

根据 PageRank 算法的定义，有：

$$V_U=\alpha\times P_U\times V_U+(1-\alpha)\times S_U \qquad (1)$$

其中，V_U 代表用户层的用户权威值，P_U 代表用户层的概率转移矩阵，S_U 是随机跳转矩阵。下面给出P_U 和S_U 的定义：

假设$V_U^{(j)}\in Friend(V_U^{(i)})$，即用户$V_U^{(i)}$关注了$V_U^{(j)}$，则$V_U^{(i)}$到$V_U^{(j)}$的概率转移值为：

$$P_U^{(i\to j)} = \frac{1}{|Friend(V_U^{(i)})|} \tag{2}$$

其中，$|Friend(V_U^{(i)})|$是用户$V_U^{(i)}$所有关注者的个数。

$$S_U^{(i)} = \frac{1}{|V_U|} \tag{3}$$

其中，$|V_U|$代表用户层的用户总个数。

（2）将用户层的权威值传递给微博层

对于用户层的任一用户$V_U^{(i)}$，在微博层均具有一个集合$\{V_{TU_i}^{(1)}, V_{TU_i}^{(2)}, \ldots, V_{TU_i}^{(n)}\}$与之对应，即用户发的的微博集合。通过在用户层的计算，将权威值传递给微博层，作为初始值，对于任一微博结点$V_{TU_i}^{(j)} \in V_T$，有初始值：

$$V_{TU_i}^{(j)} = \frac{V_U^{(i)}}{|n|} \tag{4}$$

其中n是用户$V_U^{(i)}$发布微博的个数。

（3）在微博层计算微博的权威值

与公式（1）相似，根据 PageRank 定义，有：

$$V_T = \alpha \times P_T \times V_T + (1-\alpha) \times S_T \tag{5}$$

其中，V_T代表微博层的微博权威值，P_T代表微博层的概率转移矩阵，S_T是随机跳转矩阵。

对于P_T矩阵中的值每一行中只有一个是 1，即转发边，其余均为 0。对于随机跳转矩阵S_T，有：

$$S_T^{(i)} = \frac{1}{|V_T|} \tag{6}$$

其中，$|V_T|$代表微博层的微博总个数。

（4）将微博层的权威值传递给用户层

第 4 步是第二步的逆过程，将微博层的传递值作为用户层的初始值，有：

$$V_U^{(i)} = \sum_{j=1}^{n} V_{TU_i}^{(j)} \tag{7}$$

重复以上 4 个步骤，进行循环迭代，直至收敛，就可以得出用户的综合排名。具体实现过程见表 1。

3 实验结果及分析

3.1 评价标准

本文所用的对比实验为在用户层的 PageRank 用户排名，采用的标准为：

$$AVC(O_i) = \frac{1}{N}\sum_{j=1}^{N}(PageRank(V_U^{(j)}) - STURank(V_U^{(j)})) \tag{8}$$

其中，$AVG(O_i)$代表桶i中被转发用户排名提升平均桶数量；N 代表用户层的 PageRank

表 1　自适应 Twitter 用户排名算法

Tab. 1　Self-tuning Twitter User Rank Algorithm

自适应 Twitter 用户排名算法
输入：用户-微博图 $UTG=(V,E)$，$\alpha=0.85$
输出：用户权威值 V_U，微博权威值 V_T
迭代：
迭代：
执行公式（1）；
直至收敛
对于每一个用户 $V_U^{(i)}$：
执行公式（4）；
迭代：
执行公式（5）
直至收敛；
对于每一个用户 $V_U^{(i)}$：
执行公式（7）；
直至收敛；

用户排名桶 i 中被转发用户的总数；$PageRank(V_U^{(j)})$ 代表用户 $V_U^{(j)}$ 在用户层的 PageRank 排序中分到的桶编号；$STURank(V_U^{(j)})$ 用户 $V_U^{(j)}$ 在 STURank 排序中分到的桶编号。这种评价标准经常用于 Spam 用户检测[6]，本文的实验与 Spam 用户检测相似，所以采用此标准。

3.2　实验结果及分析

图 7 和图 8 给出了桶的数量分别是 50 和 3872 的用户排名提升情况。通过图可以分析，整体的排名趋势是在上升的，也就是我们的算法可以提高转发量比较大的用户，因为他们的微博质量好，所有权威度增加。但是可以观察到，排名靠前的用户的平局提升，也有低于 0 的，也就是说，整体趋势是上升的，但是排在前面的用户也会相应的降低排名。

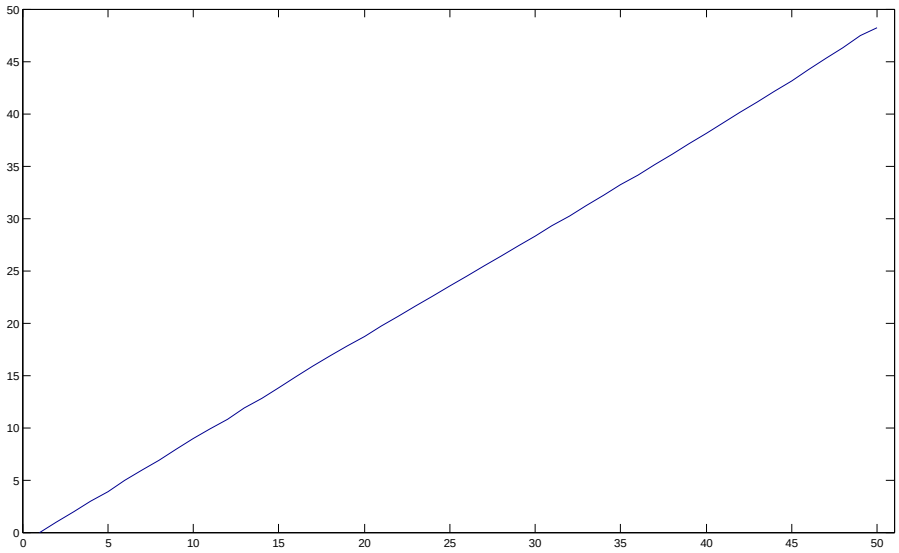

图 7　50 个桶

Fig. 7　50 Buckets

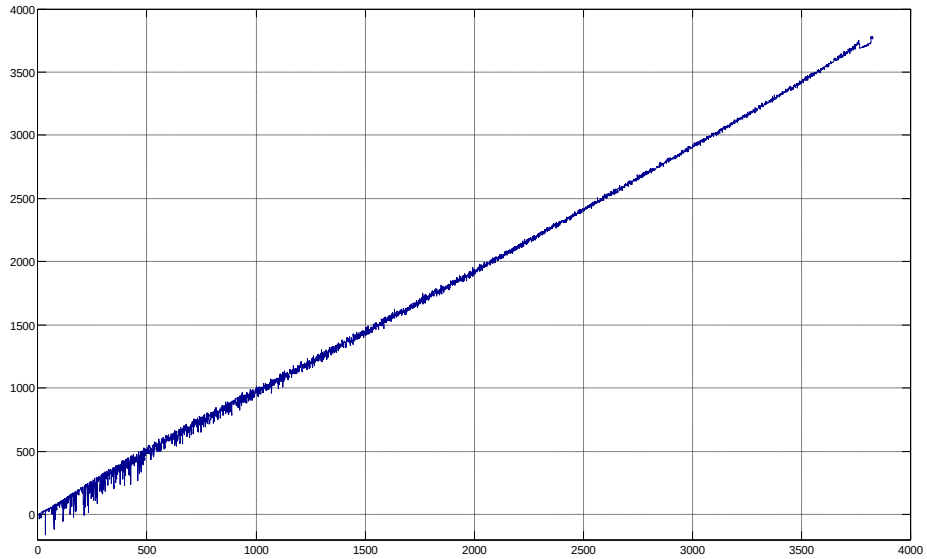

图 7　3872 个桶

Fig. 7　3872 Buckets

4　结论

本文提出了 Twitter 用户排名的新算法，自适应 Twitter 用户排名 STURank，该算法不仅考虑了用户之间的关注关系，同时考虑了用户发布微博，用户转发微博等行为，利用了 Twitter 的全部信息。实验结果表明，该方法可以有效的提升被转发用户的排名。

致谢

感谢张宪超教授多年来的指导与栽培，感谢实验室同学们关心与帮助。

[参考文献] (References)

[1] JAVA A, SONG X, FININ T, TSENG B. Why we twitter: understanding microblogging usage and communities[C]. Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis, 2007, 56-65.

[2] WENG J, LIM E, JIANG J, HE Q. Twitterrank: finding topic-sensitive influential twitterers[C]. Proceedings of the third ACM international conference on Web search and data mining, 2010, 261-270.

[3] YAMAGUCHI Y, TAKAHASHI T, AMAGASA T, KITAGAWA H. Turank: Twitter user ranking based on user-tweet graph analysis[C]. eb Information Systems Engineering--WISE 2010, 2010, 240-253.

[4] BRODER A, KUMAR R, MAGHOUL F, RAGHAVAN P, RAJAGOPALAN S, STATA R, TOMKINS A, WIENER, J. Graph structure in the web[J]. Computer networks, 2000, 33(1): 309-320.

[5] PAGE L, BRIN S, MOTWAIN R, WINOGARD T.The PageRank citation ranking: bringing order to the web [R].Stanford InfoLab, 1999.

[6] ZHANG X, WANG, Y, MOU N, LIANG W. Propagating Both Trust and Distrust with Target Differentiation for Combating Web Spam [C]. Twenty-Fifth AAAI Conference on Artificial Intelligence, 2011.