# ELPS:一种高效的微博信息传播轨迹提取算法

王 悦[1] 黄威靖[2]

(中央财经大学信息学院计算机系 北京 100081)[1] (北京大学信息科学技术学院 北京 100871)[2]

**摘 要** 近年来,随着社会性网络服务应用(SNS)的流行与发展,SNS已成为人与人之间重要的交流渠道。SNS中大量用户产生的数据内容包含了社会网络中信息传播的客观知识,由此SNS可用于研究社会网络中公众舆论的变化趋势及信息传播的相关规律。由于SNS服务中节点规模大、其用户间的信息传播通常出现离散而稀疏的情况,需要高效的信息传播观察手段。为解决该问题,提出信息传播轨迹用于研究社会网络中信息传播的基本规律,具体的方法为:(1)提出信息传播轨迹(info-trajectory)模型以记录社会网络中信息传播的具体路径;(2)针对微博社会网络,提出几个高效的信息传播轨迹抽取算法;(3)根据已获取的信息传播轨迹研究用户间转发信息行为的时序规律;(4)提出算法K-advocators-discover用于发现社会网络中促进信息传播的top-k名用户;(5)提供充分的实验测试来将所提方法用于抽取新浪微博上热点话题信息的传播轨迹,并采用K-advocators-discover算法分析新浪微博中促进信息传播的用户。实验结果验证,所提方法能高效地提取微博中信息传播轨迹,挖掘其中促进信息传播的用户。

**关键词** 社会网络,图挖掘,信息传播轨迹

**中图法分类号** TP392 **文献标识码** A

## ELPS:An Efficient Information Trajectory Extracting Algorithm in Microblog

WANG Yue[1] HUANG Wei-jing[2]

(Department of Computer Science,School of Information,Central University of Finance and Economics,Beijing 100081,China)[1]

(School of Electronics Engineering and Computer Science,Peking University,Beijing 100871,China)[2]

**Abstract** With the development of the Social Networking Services (SNS),SNS has become an important tool for people to communicate with each other. The rich user generated contents (UCG) of SNS have contained useful knowledge about information propagation rules. Thus SNS can be used to study public opinion and information propagation rules in the social networks. As information propagation happens in the online social networks discretely and sparsely,it is hard to directly observe and study the propagation process of information in an online social network with over 10 million nodes. To meet the challenges,the paper (1) provided a model "info-trajectory" to capture the information propagation pathways in the online social network,(2) proposed several algorithms to extract info-trajectory from some practical microblog social networks efficiently by employing repost timeline (a kind of public available repost notification data of microblog),(3) studied the temporal relations of the repost actions for the users on the obtained info-trajectories, (4) proposed algorithm K-advocators to discover the advocators from the information propagation trajectories and information propagation patterns in the microblog,and (5) in the experiment section,provided the sufficient experiments to study info-trajectories for several topics on Sinamicroblog(a prevalent microblog application in China),and several popular SNSs. The results show that the proposed methods are efficient to extract the info-trajectories,and useful to discover advocators for specific topics in the microblogs.

**Keywords** Social network,Graph mining,Information trajectory

## 1 介绍

随着在线社会网络应用(Social Network Services,SNS)的发展,其用户规模以惊人的速度增加。截至2012年末,Twitter、新浪微博等微博类SNS应用的用户数均已超过5亿。每天通过微博了解周边新鲜事,并与朋友交流已成为人们生活的时尚。SNS中大量的用户交互、传播行为记录为社会网络性科学的研究提供了绝佳机会。SNS所包含的真实可见的信息交互、传播数据可应用于社会网络信息传播动力学、复杂网络等科学研究及电子商务的在线广告投放、新闻影响力分析、在线公众舆情分析等实际应用中。SNS用户规模惊人,且其用户间的信息传播通常出现离散而稀疏的情况,这

到稿日期:2013-05-09 返修日期:2013-06-20 本文受国家自然科学基金(60970143,61272398),北京市教委共建项目,中央财经大学研究生教育改革项目资助。

王 悦(1981—),男,博士,讲师,主要研究方向为数据库与知识工程,E-mail:wangyuecs@cufe.edu.cn;黄威靖(1986—),男,博士生,主要研究方向为文本挖掘、Web挖掘。

成为相关研究的巨大挑战。

本文的主要任务是研究高效抽取微博社会网络中信息传播的轨迹，并应用抽取结果发现真实社会网络中促进信息传播的用户。信息传播轨迹是信息在社会网络中传播路径的集合：它包含社会网络的拓扑结构信息（比如：社会网络中的朋友关系）以及社会网络中的信息传播的时序路径。若信息传播轨迹得到合理应用，可客观地回答如下相关问题：(1)当突发新闻“日本地震，福岛核电站泄漏”在微博上发布后，信息是怎样在用户间传播的？(2)针对微博中的热门新闻事件博文，是哪些用户群体在推动该事件的发展？

微博的特性使其中各用户均可被看作是社会网络中的独立的信息源与中转站。微博中用户间的关系是公开、单向的：用户可以单向地成为(follow)其他用户的粉丝(follower)，同时该用户也可以被别人 follow 成为他人的 followee。在成为他人粉丝后，用户可在私人 public timeline 看到其 followee 的公共行为。

理想情况下，微博上观察信息传播轨迹可采取如下步骤实现：(1)获取社会网络中讨论过给定话题的用户列表；(2)从微博社会网络中抽取这些用户的拓扑关系；(3)使用图模型表达这些用户间的朋友关系；(4)根据信息传播顺序抽取出信息传播轨迹。图 1 展示了这些步骤。

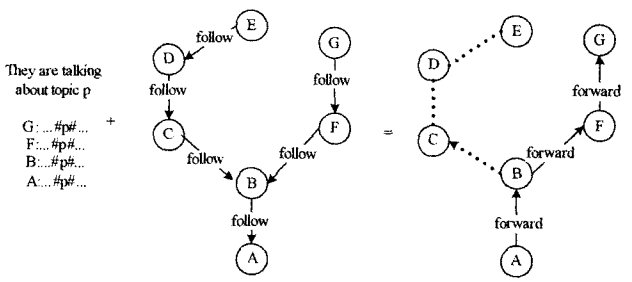



图 1　获取信息传播轨迹的基本步骤

分析微博获取的实际数据发现：在转发微博时，用户的转发内容不一定会加入引用符号(“#”或“@”)，这使得信息传播的轨迹很难被直接观察到。同时，由于在实际 SNS 应用中海量的用户数及 hub 用户的存在使信息传播非常稀疏，这造成在实际应用中观察信息传播轨迹更加困难。为应对这些挑战，本文的目标即为提供利用从微博所获的公共数据观察，研究信息传播规律的技术手段。

基本思路：微博类应用通常提供公开的转发信息 API（如：repost timeline=[tweetid, userid, time]），其记录了用户转发微博文的信息。我们尝试使用从 repost timeline 中所获取的数据来推测信息的真实传播规律。

针对一些热点话题，我们从新浪微博中获取了 2011 年 2 月 26 日到 2011 年 3 月 26 日间部分博文的 repost timeline。发现大部分博文都是由转发者直接从原作者转发，甚至包含大量与原作者没有直接 follow 关系的转发者；这说明用户间微博转发的记录(@或#等标记)丢失在了转发过程中。我们分析了从 2011 年 2 月 28 日到 2011 年 3 月 7 日间转发数过万的博文，直接被原作者粉丝转发的博文仅占所有转发数的 37.1%。这说明超过一半的用户并不是从原作者处看到并转发该博文，而是从自己的 followee 处看到并转发。根据前面描述 repost timeline 的数据模式可知，其包含数据并没有直接记录用户间的信息传播路径，而只记录了转发选定博文的用户列表及时间点。由此，若希望推测用户间信息的真实传播轨迹，需结合 SNS 的图结构及转发数据进行综合分析。

本文尝试结合微博的 repost timeline 数据及用户间 follow 关系来推测信息传播的真实传播的轨迹，主要贡献为：(1)提出信息传播轨迹模型(info-trajectory)，给出从微博抽取信息传播轨迹模型问题的形式化定义及相关分析；(2)针对微博社会网络模型，提出基于种子节点的局部结构信息传播轨迹抽取算法，该算法根据 repost timeline 上的种子用户迭代实现抽取；(3)提出“K-Advocator-Discover(KAD)”算法用于从 info-trajectory 中分析促进信息传播的用户；(4)在新浪微博、Twitter、Flickr 等 SNS 应用的数据上测试所提的信息传播轨迹抽取算法，并尝试用“KAD”算法找出其中促进信息传播的关键用户，同时将所提方法与相关算法进行比较，测试其有效性及性能。

本文第 2 节讨论相关研究工作进展情况；第 3 节给出信息传播轨迹模型的基本概念，并提出信息传播轨迹抽取及信息倡导者挖掘问题；第 4 节提出信息传播轨迹抽取算法及倡导者发现算法，并分析所提算法的正确性及误差率；第 5 节提供了充分的实验分析所提算法的性能和效率，并在新浪微博等 SNS 应用的数据上比较“KAD”算法与“PageRank”算法的实际分析效果；最后总结全文的工作并给出未来研究方向。

## 2　相关工作

通常研究使用人与人之间的通讯或信息交互的记录[1]来创建信息传播的网络结构并使用图论的相关理论[2]来研究信息传播的基本规律。根据这一思路，G. Kossinets 等[3]研究了由一所大学邮件系统的交互数据构建的真实社会网络中信息传播的潜在快捷路径(potential short-pathway)。他们将信息传输的主要路径定义为“信息脊梁”(backbone)，并提供算法尝试从真实社会网络中抽取出该结构。G Kossinets 在自己的研究[4]中提到了社会网络中信息传播子结构的边界确定问题(boundary specification problem)[5]及其界方法。考虑从完整社会网络中直接分析信息传播结构的困难性[4]，部分研究者采用事件驱动以及局部搜索的方法：Jure Leskovec 等在文献[6]中提供了一种利用用户行为的方法研究电子商务市场中使用电子邮件进行病毒营销的策略。D. Liben-Nowell 则研究[7]使用链式内容 email 数据(chain-letter，一种要求接受者转发的 email 邮件)跟踪信息传播模式。此外，近年也有一些针对社会网络结构[8]及用户行为对其中信息传播的影响[9]的研究。

研究信息传播轨迹涉及对整体社会网络结构的采样，旨在根据部分信息推测信息真实的传播轨迹。L. Qin 等人提出的 DC 算法[10]从种子节点出发分析未知图结构，这一思路启发了我们对信息传播轨迹的研究。由此，我们尝试将 repost timeline 中的记录作为种子节点，跟踪抽取信息传播结构，提高信息轨迹抽取效率并提出了 ELPS 算法。为了与传统的结构采样算法比较，我们改进 snowball[11]采样算法为 K-DFS，使其按照深度优先遍历方式执行，并控制其访问深度。实验分析了两种算法性能的优劣，验证了 ELPS 是优秀的信息传播结构抽取方法。

## 3　问题定义

### 3.1　信息传播轨迹及信息传播轨迹抽取算法

当前分析社会网络结构的方法都基于图假设，即假设整

个社会网络是一张巨大的图 $G=\langle V,E\rangle$，其中节点集合 $V$ 代表所有包含在社会网络 $G$ 中的用户，而边集合 $E$ 为 $G$ 中的拓扑结构关系（$\forall e=\langle v_1,v_2\rangle\in E,\forall v_1,v_2\in V,v_1\neq v_2$）。在微博或在线相片分享类 SNS 应用中，边集 $E$ 代表用户间相互 follow 的关系。在此基础上，我们认为信息传播轨迹是 $G$ 中参与信息传播节点及相互关系构成的 $G$ 的子图。在给出信息传播迹的正式定义前，首先有如下假设：

假设 1　设 $A$ 与 $B$ 是微博社会网络 $G$ 中两名用户，信息从 $A$ 传播到 $B$ 需要满足如下规则：(1) $B$ 是 $A$ 的 follower；(2) $A$ 与 $B$ 都转发或发表了相同的微博文；(3) $A$ 在 $B$ 之前转发或发表该博文。

基于假设 1，信息传播轨迹可正式定义如下。

**定义 1**（信息传播轨迹，Info-trajectory）　设 $G=\langle V,E\rangle$ 是微博社会网络，$p$ 为指定话题博文，$V_p$ 是在 $G$ 中发表或转发博文 $p$ 的用户集合。则信息传播轨迹 $\Omega_p=\langle V_p,E_p\rangle$ 是 $G$ 的子图，且 $\Omega_p$ 满足：对 $\forall e\in E_p,e=\langle v_1,v_2\rangle,v_1,v_2\in V_p,v_2$ 是 $v_1$ 的 follower，$t_p(v_1)<t_p(v_2)$。其中 $t_p(v_i)$ 是用户 $v_i(i=1,2,\cdots,|V_p|)$ 发表或转发博文 $p$ 的时间。

定义 1 记录了信息在社会网络中传播的时序路径，则信息传播轨迹抽取问题可按如下方式定义。

**问题 1**（信息传播轨迹抽取）　设 $G$ 为微博社会网络，$p$ 为一则指定博文，$V_p$ 是发表或转发过博文 $p$ 的用户集合。$L_p$ 是博文 $p$ 的 repost timeline 列表，对 $\forall l_p\in L_p,l_p=\langle v_i,t_p(v_i)\rangle,v_i\in V_p$。则本问题为：从 $G$ 中抽取子图 $\Omega_p=\langle V_p,E_p\rangle$，使其满足定义 1 的设定。

问题 1 分析：要从 $G$ 中抽取信息传播轨迹，需(1)从社会网络中根据用户的朋友关系及对博文 $p$ 的发表转发情况获得所有参与信息传播的用户子图；(2)根据 repost timeline 按照定义 1 中的规则过滤子图中的节点及边集合。

按该步骤实现抽取的时间复杂度为 $O(|V|*|V_p|)$，因此若从节点规模庞大的实际微博社会网络中直接获得给定信息的信息传播轨迹非常困难。于是我们尝试使用局部抽取轨迹的办法研究信息传播轨迹。

基于种子的局部方法：由于 $L_p$ 包含了可观察到的所有与博文 $p$ 相关的用户，我们尝试以 $L_p$ 中部分节点为种子，从整个社会网络中获取参与信息传播的节点并获取信息传播轨迹。为确定从种子节点向外获取网络子结构的范围，下面给出 K-Route 网络的概念。

**定义 2**（K-Route 网络）　设 $G=\langle V,E\rangle$ 为微博社会网络，$p$ 为需要观察的博文，$V_p$ 是发表或转发博文 $p$ 的所有用户集合。$path_{i,j}=\{v_i,v_{i(0)},v_{i(1)},v_{i(2)},\cdots,v_j\}$ 是从 $v_i$ 至 $v_j$ 的路径记录，其中对 $path_{i,j}$ 中的每对相邻节点 $v_{i(n)},v_{i(n+1)}$，存在边 $e=\langle v_{i(n)},v_{i(n+1)}\rangle\in E$；设 $PATH_{i,j}(G)$ 代表在 $G$ 中 $v_i$ 与 $v_j$ 间所有路径的集合，则 $PATH^K_{i,j}(G)$ 代表在 $G$ 中 $v_i$ 与 $v_j$ 间所有长度小于 $K$ 路径的集合。由此 K-Route 网络 $G_k=\langle V_K,E_K\rangle$ 是 $G$ 的子集，且满足：$V_p\subseteq V_K$，对 $\forall v_i,v_j\in V_K(i\neq j)$，$PATH^K_{i,j}(G_K)=PATH^K_{i,j}(G)$。

因为 K-Route 网络包含了两点间所有长度在 $K$ 以内信息传播的必要通路，使用它替代原始社会网络 $G$ 近似抽取信息传播轨迹可以减少所涉及的节点及边的数量，可以提高信息传播轨迹抽取的效率，由此根据 *K-Route* 网络的概念，问题 1 可以改写为如下近似问题的求解：

**问题 2**（K-path 信息传播轨迹抽取）　在问题 1 的设置下，(1)首先从 $G$ 中获取指定博文 $p$ 的包含最小节点数的 K-Route 网络；随后(2)根据定义 1 的规则使用 $L_p$ 过滤 $G_K$ 以获取信息传播轨迹。

本文的剩余部分，为书写简单，在上下文一致的情况下，我们将包含最小节点数的 K-Route 网络简称为 K-Route 网络。

### 3.2　信息传播倡导者发现问题

信息传播轨迹记录了社会网络中用户间信息传播的历史，可用于很多有意义的研究。我们应用信息传播轨迹可以发现在微博社会网络中对信息传播起到促进、倡导作用的用户。假设该类用户可加速信息在微博中的传播速度，则该假设的形式化描述如下：

假设 2　设在社会网络 $G$ 中，若将任意对信息传播起到倡导作用的用户及其相关的边从 $G$ 中去掉，则 $G$ 中剩余的所有用户对信息的平均响应时间会增长。

假设中的平均响应时间是指一则博文从单个用户发出到社会网络中其他用户转发的平均时间。我们分析了 2011 年 2 月至 2011 年 4 月间新浪微博中热门话题的用户平均响应时间。如图 2 所示，我们发现用户的平均响应时间与其 follower 数呈反比：用户的 follower 数越多，则其响应时间越短。

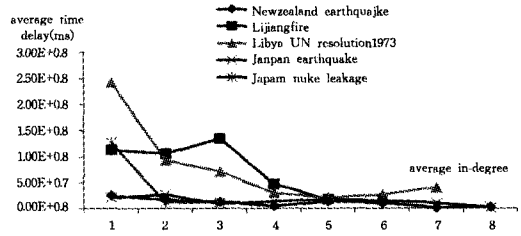



图 2　平均响应时间与 follower 数量的关系

根据假设 2，平均响应时间（Average Response Delay，ARD）越短，信息传播轨迹中部分用户对信息传播起到的倡导作用越强。我们扩展该规则，提出 ADV-Rank（Advocating ability rank）衡量信息传播轨迹中用户促进信息传播的能力。记给定用户 $v$ 的 ADV-Rank 为 $I(v)$，则 $I(v)$ 可由如下表达式计算：

$$I(v)=(1-\alpha)\sum_{i\in IN(v)}I(i)+\alpha\frac{|OUT(v)|}{\sum_{v'\in OUT(v)}|t(v')-t(v)|}\qquad(1)$$

其中，$OUT(v)$ 表示信息传播轨迹中 $v$ 的 follower 集合，$IN(v)$ 表示信息传播轨迹中 $v$ 的 followee 集合；按式(1)的设置，参数 $\alpha$ 是用户调整响应时间对 $I(v)$ 值影响大小的权值：$\alpha$ 越大，则具有短响应时间用户的 $I(v)$ 值越大；$\alpha$ 越小，则对网络拓扑结构越重要节点的 $I(v)$ 值越大。根据前述概念，ADV-Rank 可用于衡量用户对信息传播起到的倡导、促进作用，由此发现信息传播轨迹中倡导者任务的形式化可描述如下：

**问题 3**（Top-K 信息传播倡导者发现）　给定社会网络 $G$，设 $\Omega_p$ 是给定博文 $p$ 的信息传播轨迹，则从 $\Omega_p$ 中找出使 ADV-Rank 值最大的前 $K$ 个用户节点。

## 4　信息传播轨迹抽取算法及应用

### 4.1　信息传播轨迹抽取算法

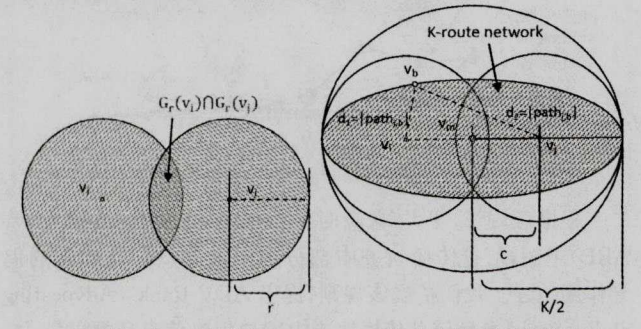
(1)K-DFS 信息传播轨迹抽取算法。snowball[11] 是从网络结构中遍历抽取子集的经典算法，我们通过将其每次递归深度控制在 K-hop 设计实现了 K-DFS 算法，用于抽取给定博文在网络中的传播轨迹结构。

K-DFS算法的主要思路是从 repost timeline 中用户为种子节点开始，以 DFS 方式遍历网络，获取所有与种子节点距离 $K$ 步内的节点及其网络结构。主要步骤可描述为：(a) 初始化栈 paths 存储种子节点向外遍历时的路径节点；(b)将所有种子节点加入 paths；(c)从栈 paths 的顶节点开始遍历网络，迭代选择 $G$ 中未被访问且度数最大的节点及边加入子网络 $G_k$ 中直到 paths 为空；(d)完成结构抽取后，根据定义 1 中规则过滤 $G_k$ 中的节点及边，获取信息传播轨迹。

K-DFS 获取包含信息传播轨迹有效路径的效率不高。在实验中，K-DFS 算法通常不能获得足够的子网络结构。其优点是它实现简单且能分别计算多个节点的周围子网络，该特性使其可分布式实现，以提高效率。

(2)ELPS 圆形扫描范围的信息传播轨迹抽取算法。受到算法 DC[10] 启发，我们提出了 ELPS 算法，ELPS 可高效地从未知全局的网络结构中抽取局部的近似信息传播轨迹，它可在满足信息传播轨迹抽取需求时，尽量减少所获取的网络结构。

图 3 展示了 ELPS 算法的基本思想，其主要步骤为：(1)从 $V_p$ 中所有节点开始($\forall v \in V_p$)，获取 r-hop 范围内的子图 $G_r(v)$，并求出其交集 $G_r(v_1) \cap G_r(v_2) \cap G_r(v_3) \cdots$(见图 3(a))；(2)从所有交集中的节点出发，获取 $K/2$ 距离范围内的子网络结构(见图 3(b))；(3)按定义 1 过滤第 2 步的结果获取信息传播轨迹，ELPS 算法的伪码列在算法 1 中。




(a) 获取 r-hop 中间节点　　(b) 获取所有路径的节点

图 3　两个种子节点的 ELPS 算法主要计算思路

**算法 1**　ELPS 算法伪码

```
Data: V_p, G, K, r, L_p
Result: information trajectory Ω
begin
    G_K = ⟨V_K, E_K⟩ ← ∅;
    foreach ∀v_i, v_j ∈ V_p(v_i ≠ v_k) do
        G_r(v_i) ← get r-hop subgraph centered at v_i;
        G_r(v_j) ← get r-hop subgraph centered at v_j;
        V_m ← V_r(v_i) ∩ V_r(v_j);
        foreach ∀v_m ∈ V_m do
            G_{K/2}(v_m) = ⟨V_r(v_m), E_r(v_m)⟩ ← get K/2-hop subgraph
            centered at v_i;
            foreach ∀v_b ∈ V_{K/2}(v_m) do
                calculate shortest path_{i,b}, path_{j,b} in G_{K/2}(v_m);
                if |path_{i,b}|+|path_{j,b}| ≤ K then
                    add related nodes, edges of path_{i,b}, path_{j,b} into
                    G_K;
                end
            end
        end
    end
    G_K = ⟨V_K, E_K⟩;
    foreach e=⟨v_1, v_2⟩ ∈ E_K do
        if v_1 repost p after v_2 then
            delete edge e from Ω;
        end
    end
    return Ω;
end
```

在 ELPS 算法中，从交集节点中向外搜索 $K/2$ 距离的子网结构能保证获取到全部长度不长于 $K$ 的路径(图 3(b)中 $d_1+d_2 \leqslant K$)。我们采用欧几里德空间的几何性质证明 ELPS 算法的正确性。

**定理 1**　设 $G=\langle V,E\rangle$ 为社会网络，对 $\forall v_i, v_j \in V(v_i \neq v_j)$，函数 $dist(v_i,v_j)$ 代表 $v_i, v_j$ 之间的距离；$G_r(v)=\langle V_r(v), E_r(v)\rangle$ 为 $G$ 的子图且满足：$\forall v' \in V_r(v), dist(v',v) \leqslant r$；$V_m = V_r(v_i) \cap V_r(v_j)$；则对 $\forall v_m \in V_m$，从 $G$ 中抽取出 $G_{\frac{K}{2}}(v_m)$ 将获得足够基于 $v_i, v_j$ 节点的 K-Route 网络结构。

证明：根据定义 2，K-Route 网络是 G 中包含所有 $PATH_{i,j}^{K}(G)$ 中路径的子网络，如图 3(b)所示，对任意 $v_i, v_j \in G.V$。若边界集 $V_B$ 对 $\forall v_b \in V_B$ 满足下式：

$$dist(v_b,v_i)+dist(v_b,v_j) \leqslant K \tag{2}$$

则边界集 $V_B$ 中所有节点将以 $v_i, v_j$ 为焦点、长轴 $2a=K$、焦距 $2c=dist(v_i,v_j)$ 的椭圆边界分布。假设 $V_m$ 非空，由此 $c \leqslant r$。因此该椭圆的短轴为 $b=\sqrt{\frac{K^2}{4}-r^2}$。因为 $r \geqslant 0$，且 $\max(b)=K/2$，所以包含椭圆边界的路径范围 K-Route 网络一定包含在以交集 $V_m$ 中节点为圆心、$K/2$ 为半径的原型范围中，即对任意 $v_i, v_j \in V$，$G_{\frac{K}{2}}(v_m)$ 一定包含所有 $PATH_{i,j}^{K}(G)$ 中的路径。证毕。

### 4.2　Top-K 倡导者发现算法

根据第 2 节定义，计算 $I(v)$ 则需要计算 $v$ 所有 follower 的 ADV-Rank 值，由此所有 $\forall v_i \in \Omega.V_L$ 的 ADV-Rank 值构成向量 $IR=\{I(v_0),I(v_1),I(v_2),\cdots\}^{\mathrm{T}}$。根据式(1)，IR 满足如下表达式：

$$IR^{n+1}=(1-\alpha)M^{\mathrm{T}}IR^{n}+\alpha IS \tag{3}$$

其中，$M$ 是信息传播轨迹 $\Omega$ 的邻接矩阵，$IS=\{s_1,s_2,s_3,\cdots\}^{\mathrm{T}}$，其中 $s_n$ 对 $\forall v_n \in \Omega$ 满足如下表达式：

$$s_n=\frac{|OUT(v_n)|}{\sum\limits_{\forall v' \in OUT(v_n)} |t(v')-t(v_n)|} \tag{4}$$

由于信息传播轨迹 $G$ 的子图中可能存在有零出度节点或圈，这可能会导致式(3)无解。我们根据文献[12]中提供的随机跳转形式，通过引入随机跳转概率 $p_r$，式(3)可转换为如下形式：

$$IR^{n+1}=(1-\alpha)p_rM^{\mathrm{T}}IR^{n}+(1-\alpha)(1-p_r)A+\alpha IS \tag{5}$$

其中，$A=\{a_1,a_2,a_3,\cdots\}$ 是个矢量，$a_i=1$ 代表节点 $v_i$ 的出度为零或位于圈路径上。我们采用随机游走的方式对式(5)迭代求解，则可以获取信息传播轨迹上所有节点的 ADV-Rank 值。

Top-K 信息传播倡导者的发现：经过观察发现，节点之间会相互影响彼此的信息传播行为及模式。由此我们按“去掉-重计算(Removed and recalculate)”策略设计了算法 K-Advocator-Discover(算法 2)。

直观上，算法 K-Advocator-Discover 采用如下几个步骤实现 Top-K 信息传播倡导者的挖掘：(1)计算并找到信息传播轨迹中 ADV-Value 值最大的节点 $v$；(2)从信息传播轨迹中输出并删除 $v$ 生成新的信息传播轨迹；(3)重新从信息传播轨迹中计算所有节点的 ADV-Rank 值，反复运行第(1)与第(2)步骤直到结果集中包含了 $K$ 个节点，此时输出最终 Top-K 信息传播倡导者结果。

**算法 2** K-Advocator-Discover(KAD)

```
Data: Ω, R (iterate rounds)
Result: top-k list L of users with ADV-Rank in Ω
begin
    L ← ∅;
    while Ω is not ∅ and |L| < k do
        foreach ∀v_n ∈ Ω do
            Calculate s_n ← |OUT(v_n)| / Σ_{v'∈OUT(v)} [r(v')−r(v_n)]
        end
        IS←{s_1, s_2, s_3,...};
        A← calculate a_i by judging out-degree for each node;
        Iteratively R times to solve
        "IR^{n+1} = (1 − α)p_r · M · IR^n + (1 − α)(1 − p_r)A + αIS";
        Select v with Maximum ADV-Rank into L;
        Remove v and its related edges from Ω;
        re-generate Ω by checking reposted relations;
    end
    return L;
end
```

## 5 实验及相关讨论

### 5.1 数据集与实验设置

我们选择了目前流行的 SNS 应用——新浪微博用于实验。实验小组实现了一个可根据公开 API(repost timeline)的新浪微博数据抓取系统,获取了从 2011 年 2 月 26 日至 2011 年 4 月 7 日间一些与流行事件相关的微博文的发布及转发情况。总共获取了 61 个关于"新西兰地震"、"利比亚内战"、"日本地震及核泄漏"等事件的相关微博文的数据。所有跟踪到的博文在指定时间段内均有超过 833 个用户的转发数。

表 1 用到数据汇总

| 来源 | 节点数 | 边数 | 平均度数 | Repost timeline |
|---|---|---|---|---|
| Sina | 410298 | 882010 | 2.15 | 50546 |
| Twitter | 11316811 | 85331846 | 15.08 | N/A |
| Delicious | 103144 | 1419519 | 27.52 | N/A |
| Flickr | 80513 | 5899882 | 146.56 | N/A |
| Douban | 154907 | 654188 | 8.45 | N/A |

在收集到的数据集上,我们将新浪微博相关博文的 repost timeline 中的部分用户作为种子节点,分别实现了 K-DFS 及 ELPS 两种方法用于信息传播轨迹发现。为了比较不同数据集上信息传播轨迹抽取算法的性能及效果,我们还使用了从 AS social computing data repository[13] 获取的 4 个 SNS 拓扑结构数据,包含 Twitter、Deli. ci. ous、Flickr 以及 Douban。表 1 展示了本文用到的所有数据集的统计信息,新浪数据使用真实 repost timeline,其余数据均是在真实拓扑结构上仿真了信息传播情况。

本文的实验包含两部分:(1)信息传播轨迹抽取算法测试相关实验;(2)信息传播倡导者发现算法测试实验。实验(1)在一台具有 2.393GHZ 频率的四核 CPU,16G 内存的服务器上实现;实验(2)在一台具有 2.93GHz 频率的单核 CPU、2GB 内存的 PC 机上实现。原型系统采用 JAVA 实现,所有的数据均存储在安装了 MySQL 的数据库服务器中,实验中所有网络图形化结果均使用开源软件 JUNG[14] 绘制。

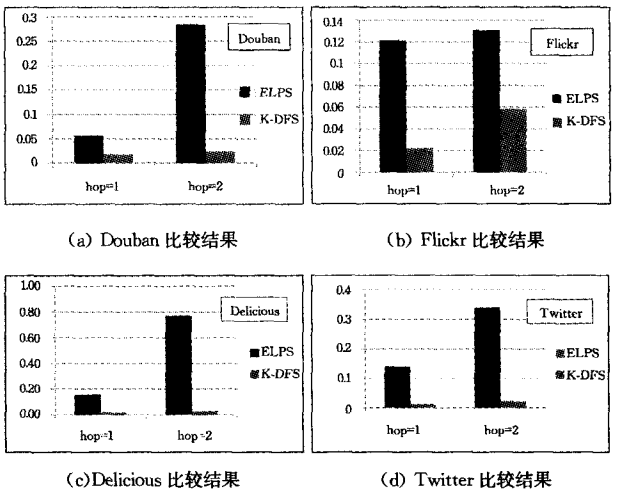
### 5.2 信息传播轨迹抽取实验

我们实现并测试了算法 K-DFS 及 ELPS 的信息传播轨迹抽取性能及效率。所采用的评价指标为 $K$ 路径覆盖率(记为 $r_K$)。该指标用于测量信息传播轨迹抽取结果中可以覆盖到的路径比率。比率越高说明算法越可能找到尽量多的信息传播通路。对任意网络 $G'=\langle V',E'\rangle$,$r_K$ 计算方式如下:

$$r_K(G')=\frac{|\{v_i|v_i\in path,\forall\, path\in PATH^K(G')\}|}{|V_r|} \quad (6)$$

其中,$Path^K(G')$是由 $G'$包含的种子节点间长度不超过 $K$ 的路径构成的集合。式(6)中分子部分代表在已获取结构中覆盖到的目标路径中的节点数量,而分母代表获取到的结构中所有包含的节点数。

我们将两种算法分别应用于从 Twitter、Flickr、Douban 以及 Delicious 数据集上抽取仿真信息传播轨迹的任务中。本部分仿真信息传播的方式为:从种子节点出发,每次随机从邻居节点选择 5 个节点,迭代 20 次。图 4 展示了在 4 个数据集上的 K-DFS 与 ELPS 算法的 $K$ 路径覆盖率比较结果。



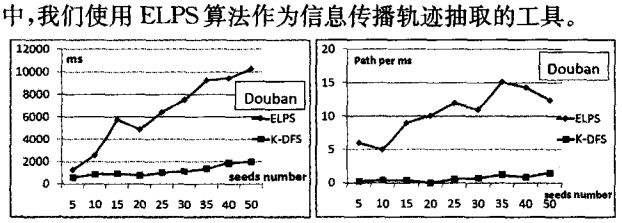

(a) Douban 比较结果 (b) Flickr 比较结果

(c)Delicious 比较结果 (d) Twitter 比较结果

图 4 K-DFS 与 ELPS 算法的路径覆盖率比较

如图 4 所示,ELPS 算法的 $K$ 路径覆盖率比 K-DFS 高,说明 ELPS 算法可以更好地获取到社会网络中可能是信息传播通路的中间结构。

为测试 K-DFS 与 ELPS 两种算法的效率,我们将每个算法以参数 hop=1、$K$=10 运行 20 遍,则算法的时间效率及信息传播路径抽取效率的结果如图 5 所示。如图 5(a)中结果,随着信息传播种子节点数增加,K-DFS 与 ELPS 算法的时间开销基本都成线性增长,ELPS 算法的时间开销基本是 K-DFS 算法的 5 倍;图 5(b)列出了 ELPS 与 K-DFS 算法每毫秒找到可覆盖到信息传播轨迹的路径数量,如图中所示,ELPS 算法能找到可覆盖到信息传播轨迹的路径为 K-DFS 算法的 15.6 倍。由此为了获得更大的性价比,在实验剩余的部分中,我们使用 ELPS 算法作为信息传播轨迹抽取的工具。




(a) 时间效率比较 (b) 每毫秒有效路径数搜索

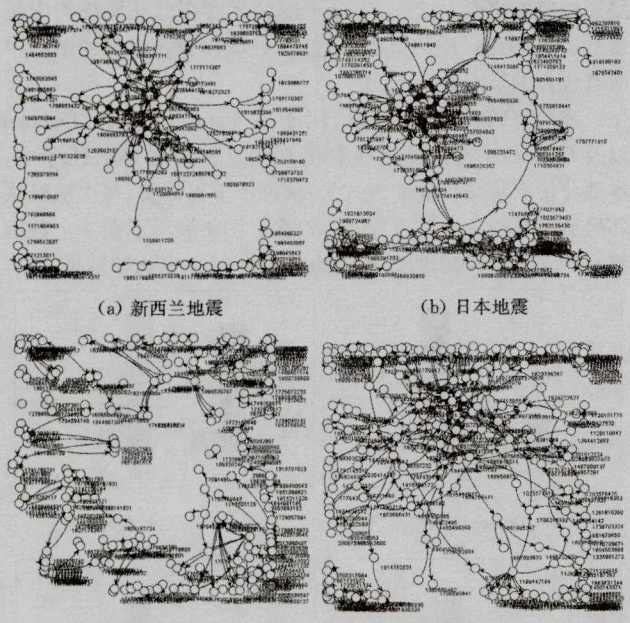
图 5 ELPS 与 K-DFS 运行效率详细比较

### 5.3 信息传播倡导者发现应用

我们使用 ELPS 算法抽取了从 2011 年 2 月 28 日到 2011 年 3 月 7 日的新浪微博中一些热点话题的信息传播轨迹,结果展示在图 6 中。

在图 6 结果的基础上,分析 KAD 算法的实际分析效果。根据假设 2,信息传播倡导者可以加速信息传播的过程及速度,假如一个较强的信息传播倡导者从信息传播轨迹中移除,

则该信息传播轨迹中所有节点的平均延迟时间将增大。由于PageRank也经常被用于网络中有影响力节点的发现任务，我们实现了PageRank(PR)算法作为KAD的比较算法。

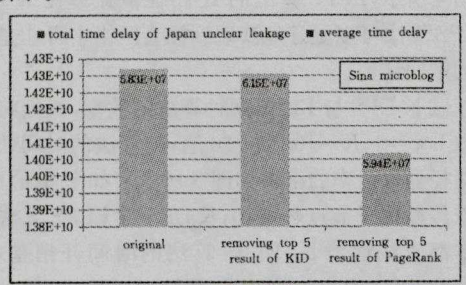


(a) 新西兰地震　(b) 日本地震

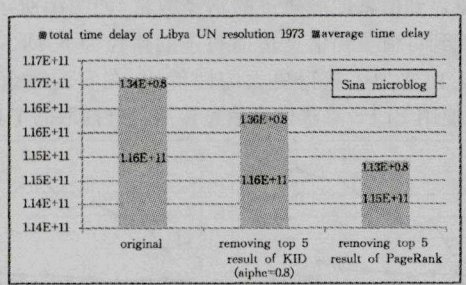
(c) 日本核泄漏　(d) UN利比亚1973号决议

图6　新浪微博中流行话题的信息传播轨迹

KAD vs. PR：在本部分实验中，将KAD与PR算法的随机跳转率都设为0.85，并设置信息传播倡导参数$\alpha=0.8$，则KAD与PR两个算法发现的信息传播倡导用户的效果比较如图7所示。




(a) 关于日本核泄漏话题信息传播效果影响




(b) UN关于利比亚1973号决议话题信息传播效果影响

图7　KAD与PR搜索效果的比较

图7分别展示了将KAD与PR算法找到的Top5结果从原始信息传播轨迹中去掉后，“日本核泄漏话题”及“UN关于利比亚1973号决议话题”的信息传播轨迹中节点的平均响应时间ARD的变化情况。如图所示，将KAD算法找到的结果去掉后，“日本核泄漏话题”与“UN关于利比亚1973号决议话题”的ARD分别为$6.15\times10^7$毫秒与$1.36\times10^8$毫秒，长于去掉PR算法找到结果后的ARD值(分别为$5.94\times10^7$毫秒及$1.13\times10^8$毫秒)。根据假设2，该结果说明本文提出的K-Advocator-Discover算法比PageRank在信息传播倡导者搜索任务中的效率更高。

信息转发分布影响：除比较平均响应时间，我们还比较了分别去掉KAD及PR找到的Top5信息传播倡导者对信息转发分布的影响。分别计算了原始信息传播轨迹($\Omega_{ori}$)、去掉KAD结果的信息传播轨迹($\Omega_{KAD}$)及去掉PR结果的信息传播轨迹($\Omega_{PR}$)的离差平方和(Sum of Square，SS，用于分析组间差距)，详细结果及比较见表2。

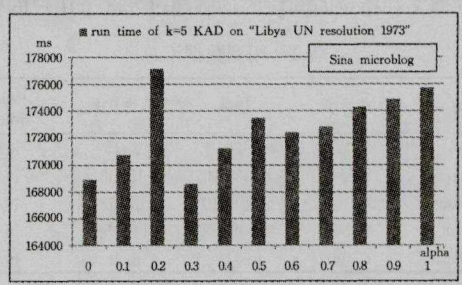
表2　离差平方和比较结果

| | | | |
|---|---|---|---|
| 日本核泄漏 | SS | $\Omega_{KAD}$ | $\Omega_{PR}$ |
| | $\Omega_{ori}$ | 60 | 5 |
| UN利比亚1973决议 | SS | $\Omega_{KAD}$ | $\Omega_{KAD}$ |
| | $\Omega_{ori}$ | 44 | 29 |

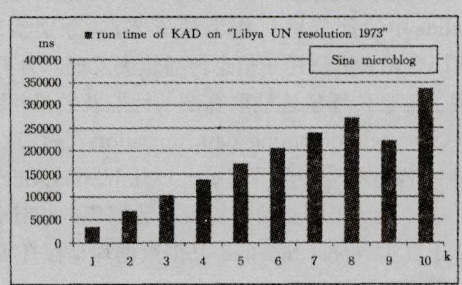
从表2中结果可看到从原始信息传播轨迹中去掉KAD算法找出的信息传播倡导者后，对信息传播分布的影响远大于去掉PR算法找出的结果的影响。

一些观察和发现：我们比较了KAD找到的“联合国关于利比亚1973号决议”与“日本核泄漏”相关的信息传播倡导者的用户profile发现：此类新闻80%的信息传播倡导者都是40岁左右的男性用户，这符合中年男性比较关心世界政治话题的直观判断。

算法性能测试：我们测试了KAD算法在不同参数取值下的性能，具体实验结果展示在图8中。




(a) KAD算法在参数$\alpha$不同取值下的性能表现




(b) KAD算法在参数$K$不同取值下的性能表现

图8　KAD算法在不同参数取值下的性能表现

从图8中可以看到，在“联合国关于利比亚的1973号决议”相关话题的信息传播轨迹结构上，KAD的总运行时间在168610ms到173500ms之间。当$\alpha$取0时，KAD的运行效率和PageRank相同，此时的运行时间为168922ms。另，从图8看出，KAD算法运行的时间开销随$K$的增大而线性增加。

**结束语**　我们提出了信息传播轨迹用于建模、分析信息在微博社会网络中的传播过程。主要贡献为：(1)提出了微博社会网络中的信息传播轨迹相关概念及如何获取的问题；(2)在信息传播轨迹结构的基础上，提出了信息传播倡导者的相关概念及如何发现信息传播倡导者的问题；(3)提出了基于种

(下转第255页)

模块给系统性能带来提升的效能要小于模块占用系统资源的性能，所以造成了小幅度的系统响应时间增加，但随着暴发式任务请求强度的增加，其所占系统资源的比重不断减小，其给系统性能提升带来的表现逐渐显现。因此得出结论，通过引入 BP 神经网络模块可使云计算系统在应对暴发式任务请求时的性能得到有效提高，本文提出的模型是一种可行的方法。

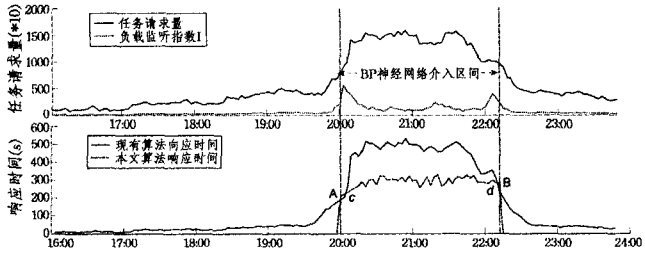



图 3　添加 BP 神经网络模块前后性能对比

**结束语**　本文针对暴发式任务请求对云计算系统性能带来的冲击，设计了基于 BP 神经网络算法改进的资源部署模型，它能够在判断暴发式任务到来的同时，改变资源部署策略，并通过 BP 神经网络模块，实时优化调整参数，达到优化系统全局的目的。本文首先分析了现有的应对暴发式任务请求资源部署模型，并归纳总结了现有的优势与不足，随后引入网络模块来解决动态跟随底层资源与请求量变化的问题。从部署模型的应用原理与流程出发，对其网络结构与学习算法进行设计，最后通过仿真实验对模型进行验证，结果表明本文方法可行、有效，对提高云计算系统应对暴发式任务请求能力起到积极作用。

## 参 考 文 献


[1] 文雨，孟丹，詹剑锋. 面向应用服务级目标的虚拟化资源管理[J]. 软件学报，2013，24(2)：358-377

[2] Caniff A, Lu Lei, Mi Ning-fang, et al. Fastrack for Taming Burstiness and Saving Power in Multi-Tiered Systems[C]//22nd International Teletraffic Congress (ITC 22). Amsterdam, the Netherlands, September 2010

[3] Tai Jiang-zhe, Meleis W, Zhang Jue-min, et al. ARA: Adaptive Resource Allocation for Cloud Computing Environments under Bursty Workloads, 978-1-4673 [R]. Northeastern University, Boston, USA, 2011

[4] 高刃，唐龙，伍爵博. 基于神经网络的无线传感器网络数据预测应用研究[J]. 计算机科学，2012，30(5)：44-47

[5] 马锐. 人工神经网络原理[M]. 北京：机械工业出版社，2010

[6] Tirado J M, Higuero D, Isaila F, et al. Predictive Data Grouping and Placement for Cloud-based Elastic Server Infrastructures [C]//2011 11th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing. IEEE DOI/CCGrid, 2011：285-294

[7] C12G Labs S. L. Private cloud computing with OpenNebula 1. 4 [EB/OL]. http://opennebula. org/_ media/software: ecosystem:private_cloud_computing_with_opennebula_1. 4. pdf, 2010

[8] 刘进军，赵生慧. 面向云计算的多虚拟机管理模型的设计[J]. 计算机应用，2011，31(5)：1417-1419

[9] Arlitt M, Jin T. Workload characterization of the 1998 World Cup Web site[R]. HPL-1999- 35R1. HP Laboratories, 1999

[10] 李强，郝沁汾，肖利民，等. 云计算中虚拟机放置的自适应管理与多目标优化[J]. 计算机学报，2011，34(12)：2253-2264

[11] Rodrigo N C, Ranjan R, Beloglazov A, et al. CloudSim: A Toolkit for Modeling and Simulation of Cloud Computing Environments and Evaluation of Resource Provisioning Algorithms[R]. Cloud Computing and Distributed Systems Laboratory, Australia, 2010


(上接第 238 页)

子节点的局部信息传播轨迹获取算法(K-DFS, ELPS)，以从微博社会网络中获取信息传播轨迹；(4)提出了信息传播倡导者发现算法(KAD)；(5)提供了充分的实验在真实的微博社会网络(新浪微博、Twitter、Flickr、Douban)中讨论分析本文所提算法的效果及效率。在实验部分，我们验证了信息传播轨迹是一个可用于研究社会网络中信息传播规律不错的数据结构。除此外，本文所提方法可不局限于使用在微博领域的信息传播规律分析领域中，还可用于分析其他 SNS 应用、在线广告投放或电子商务领域的其他应用中。我们未来的工作方向为：提高所提算法性能，根据观察 SNS 应用中用户对不同新闻事件话题信息的传播方式，尝试建立新闻事件与微博信息传播轨迹相关的联系。

## 参 考 文 献


[1] Granovetter M. The strength of weak ties[J]. American Journal of Sociology, 1973, 78(6)：1360-1380

[2] Huberman B A, Adamic L A. Information Dynamics in the Networked World [J]. Lect. Notes Phys. , 2004, 650：371-398

[3] Kossinets G, Kleinberg J M, Watts D J. The structure of information pathways in a social communication network[C]// Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2008：435-443

[4] Kossinets G. Effects of missing data in social networks[J]. Social Networks, 2006, 28：247-268

[5] Laumann E, Marsden P, Prensky D. The boundary specification problem in network analysis[J]. Applied Network Analysis, 1983(10)：18-34

[6] Leskovec J, Adamic L A, Huberman B A. The dynamics of viral marketing[J]. ACM Transactions on the Web (TWEB), 2007, 1 (1)

[7] Liben-Nowell D, Kleinberg J. Tracing information flow on a global scale using Internet chain-letter data[J]. Proc. Natl. Acad. Sci. USA, 2008, 105(12)：4633-4638

[8] Bakshy E, Rosenn I, Marlow C, et al. The role of social networks in information diffusion[C]//WWW. 2012：519-528

[9] 樊鹏翼，王晖，姜志宏，等. 微博网络测量研究[J]. 计算机研究与发展，2012(4)：691-699

[10] Qin L, Yu J X, Chang L. Keyword search in databases: the power of RDBMS[C]//SIGMOD Conference. 2009：681-694

[11] Illenberger J, Kowald M, Axhausen K W, et al. Insights into a spatially embedded social network from a large-scale snowball sample[C]//The European Physical Journal B-Condensed Matter and Complex Systems. 2011：1-13

[12] Song Xiao-dan, Chi Yun, Hino K, et al. Identifying opinion leaders in the blogosphere[C]//CIKM 200. 72007：971-974

[13] Zafarani R, Liu H. Social Computing Data Repository at ASU [OL]. http://socialcomputing. asu. edu. Tempe, AZ: Arizona State University, School of Computing, Informatics and Decision Systems Engineering, 2009

[14] http://jung. sourceforge. net/