Journal of Computer Applications
计算机应用,2012,32(2):322 -325
ISSN 1001-9081
CODEN JYIIDU
2012-02-01
http://www.joca.cn

文章编号:1001 -9081(2012)02 -0322 -04
doi:10.3724/SP.J.1087.2012.00322

位置服务社交网络用户行为相似性分析

袁书寒*,陈维斌,傅顺开
(华侨大学 计算机科学与技术学院,福建 厦门 361000)
(*通信作者电子邮箱 bookcold@hqu.edu.cn)

摘　要:基于位置的社交网络(LBSN)能够支持用户分享地理位置信息,网站中保存用户访问真实世界地理位置的记录构成用户的行为轨迹,但LBSN用户相似性的分析并没有从用户的地理位置轨迹上加以考虑。为此,提出基于划分层次,在不同的邻域半径下密度聚类的方法,探索基于位置的服务(LBS)平台上用户地理位置上相似性的度量。该方法在不同空间位置比例尺下观察用户访问各个聚类区域的次数,进而利用向量空间模型(VSM)计算用户在各个层级的相似性,最终以不同权重叠加各层级的用户相似性值,得出用户在地理空间行为上的相似性。基于国内某大型位置社交网站真实用户数据的实验结果表明,该方法能有效识别出访问地理位置相似的用户。

关键词:用户相似性;轨迹相似性;基于位置的服务;空间数据挖掘;聚类
中图分类号:TP311.13;TP393.094　　**文献标志码**:A

User behavior similarity analysis of location based social network

YUAN Shu-han*, CHEN Wei-bin, FU Shun-kai
(*College of Computer Science and Technology, Huaqiao University, Xiamen Fujian* 361000, *China*)

Abstract: Location-based social network allows users to share location information. The complete geographical record about users kept by social network plays as the basis for analyzing the behaviors of the users in geographical track. For Location-Based Service (LBS) platform did not take the users' geographical location on the track into consideration, this paper proposed a new hierarchical density based clustering approach. It determined the similarity among users in different scales by classical Vector Space Model (VSM), with vectors composed of users' visiting frequencies about different cluster area. Overlapping the different scale user similarity value with different weighted obtained the geospatial similarity of the user behaviors. The experiments based on user data from a large LBS social network site demonstrate that the proposed approach can effectively identify similar users.

Key words: user similarity; trajectory similarity; Location-Based Service (LBS); spatial data mining; clustering

《计算机应用》Journal of Computer Applications

0　引言

在线社交网络服务已经成为互联网上发展最快的应用。通过社交网站,用户之间可以相互联系、分享照片或视频等信息。根据最近的研究[1]显示,社交网站的访问量已经占到互联网总访问量的25%;全球互联网用户中有近2/3的人使用社交网站。基于位置的社交网络(Location Based Social Network, LBSN)通过整合移动互联网和互联网的新型社交网络服务,支持用户随时随地自由记录并分享地理位置等信息。用户可以通过个人电脑或手机客户端软件签到所处位置信息,并且告知朋友。当签到信息发生变更时,用户能够通过社交网站同步更新,方便快捷地与好友分享更新内容。在使用此类社交网络时,用户通常希望系统能够推荐一些与他们行为/兴趣相似的好友,或者更多自己可能感兴趣的地点,从而帮助用户去更好地发现自己身边的世界。然而,目前国内外主流的基于位置的社交网络均未提供此功能,而要实现此推荐功能,用户相似性的判断扮演着关键的作用。

因为位置服务社交网络的流行,对位置服务社交网络用户行为的分析[2-5]是研究重点之一。文献[2]得出用户的签到次数和访问的位置服从正态对数分布,同时分析出用户的签到次数并不会随着用户好友数的增加而增加。文献[3]主要分析位置服务社交网络用户的使用习惯,例如用户的使用时间、用户的主要访问地区等。进而也有文献利用大量的地理位置数据研究用户在行为轨迹上的相似性。文献[6-7]都是利用全球定位系统(Global Positioning System, GPS)日志分析用户的行为轨迹,由于GPS日志可以详细记录用户的轨迹,适合分析用户的行为模式。文献[6]通过用户的访问轨迹识别用户社团;文献[7]利用GPS日志计算用户活动轨迹的相似性,首先识别出用户访问的地理位置点,并对位置点聚类,匹配用户在聚类后的位置上的访问序列计算用户相似性。文献[8]利用位置服务社交网站Foursquare数据,考察用户在位置上的语义相似性。由于Foursquare对地理位置数据均有语义上的分类,文章利用位置的分类信息计算用户访问行为的语义相似性。由于GPS日志可以持续跟踪用户的行为轨迹,而在位置服务的社交网络中,用户仅在到达某位置后签到,没有对用户的行为轨迹进行持续的跟踪,且用户签到具有一定的随意性,因此序列模式的方法并不适用于社交网络。而在推荐系统中,为了给用户推荐合适的好友和感兴趣的地点仅通过语义相似性是不合适的,例如对于两位在不同城市的用户,即使他们访问的地点在语义上有相似性,但是推荐不同城市的好友或者地点对于用户来说都没有实际意义。因此,在用户活动轨迹记录并不连续的情况下,利用用户访问的

收稿日期:2011-07-18;**修回日期**:2011-09-15。　**基金项目**:福建省重大产学研项目(2010N5008);泉州市科技计划项目(2009G5)。
作者简介:袁书寒(1987 -),男,湖南常宁人,硕士研究生,主要研究方向:数据仓库、数据挖掘;　陈维斌(1954 -),男,福建泉州人,教授,主要研究方向:数据库、数据仓库、决策支持;　傅顺开(1978 -),男,福建仙游人,讲师,博士,主要研究方向:数据挖掘、信息检索、基于位置的应用。

实际地理位置坐标计算用户相似性,可以为用户提供适合的好友和地点推荐。

本文从用户历史签到的地理位置上研究用户相似性,提出基于划分层次密度聚类的方法。该方法首先对用户签到的地理位置进行聚类操作,得到用户访问的位置区域;并通过改变聚类的邻域半径,在不同空间位置比例尺下观察用户访问各个位置区域的情况,进而利用向量空间模型计算用户在不同空间比例下的相似性,并最终得到用户行为轨迹上的相似性。给用户推荐在访问轨迹上具有相似性的地点和好友,能更好地为用户服务。

1　用户相似性分析

在位置服务社交网络中记录用户完整的行为轨迹,这些轨迹由一系列的时空点组成。通过分析用户轨迹,可以得出用户在真实世界地理位置上的行为方式。在用户访问轨迹中,对于两位普通用户,当他们签到的位置都是在北京市时,则认为两位用户有一定的相似性;假如两位用户同时都在中关村签到,可以认为这两位用户的在行为轨迹上的相似性更高。同时,若这两位用户经常在中关村签到,则要比只是偶尔在中关村签到的用户具有更高的相似性。因此本文在判断用户在地理位置上的行为相似性基于如下经验:

1)用户所访问的地理位置越接近,用户行为轨迹越相似;

2)用户访问相近的地理位置次数越多越相似。

1.1　POI 划分层次聚类

在位置服务社交网络中,用户访问的每一个地理位置称作一个地理兴趣点(Point Of Interest, POI)。典型的 POI 库中记录每个 POI 的唯一内部 ID、名称和经纬度,即 POI = (VenueID, Name, Location)。用户在现实世界中到达某地理位置后,通过手机在社交网站上签到(Checkin)记录访问过的 POI。一次签到记录在内部通常包括用户标识信息、用户访问的 POI、访问时间以及用户对访问 POI 的评价,其中用户评价是可选的:Checkin = (UsrID, POI, Time, Tips)。

由于在某一地理位置附近,可能有很多不同的 POI,因此用户访问相近的地理位置并不一定在同一个 POI 上签到,仅通过判断用户是否访问同一个 POI 计算得到的用户间相似性是趋近于零的。因此,本文提出利用划分层次密度聚类的方法将相近的 POI 在不同的距离比例尺下聚类,根据用户访问的各个聚类区域,度量用户相似性。利用密度聚类的方法,可以聚集用户经常访问的 POI,而用户偶尔访问的区域或只有少数用户才访问的区域作为噪声而过滤掉。本文利用 DBSCAN[9] 对用户访问的兴趣点进行聚类操作。

定义 1　地理兴趣区域(Region Of Interest, ROI)[10]。经 DBSCAN 聚类后的 POI 集合,称为 ROI。

DBSCAN 需要给定邻域 ε 和邻域内包含的最小 POI 数目 *MinPts* 两个参数。根据不同的邻域值,可以调整聚类的地理位置区域的大小。由于用户所访问的地理位置越接近,用户行为轨迹越相似。因此,本文利用聚类的邻域 ε 对聚类区域划分层次,这样能够在不同的地理空间比例下计算用户相似性。邻域 ε 从大到小,用户在更小的邻域上访问相同的聚类区域就有更高的相似性。如图 1,从上至下,随着划分层次的增高,空间位置比例尺也逐渐变大,邻域参数则逐步减小,对应的聚类结果逐步细化,从大 ROI 到较小的 ROI 直至最底层的 POI。基于用户在更高层次上访问相同 ROI 的相似性计算结果将比用户在低层次上访问相同 ROI 的相似性要高。在 LBSN 中采用基于这种层次特征的 ROI 计算的优势是显而易见的,包括:1)地理比例尺和邻域参数存在着自然且合理的对应关系;2)计算结果将比基于单个 POI 的计算更稳定,效率也将更高,这对于在线应用很重要;3)允许在不同层级下获得不同的相似群体,满足不同粒度要求的 LBS 应用需求(例如社区、街道、区等);4)可以根据计算资源的富裕程度来决定层级数目的选取。

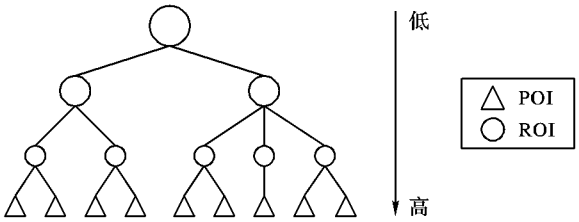


图 1　对应不同空间地理比例尺的划分层次聚类示意图

1.2　用户相似性计算

本文利用向量空间模型(Vector Space Model, VSM)计算用户相似性。将每位用户所签到的 ROI 记为向量 $\boldsymbol{R} = [r_1, r_2, \cdots, r_n]$,为了体现用户访问相同 ROI 次数越多,用户行为越相似,设定 a_i 为用户访问第 i 个 ROI 的次数。则所有用户访问的 ROI 构成一个用户访问位置矩阵 $\boldsymbol{V}_{m\times n}$,可用余弦相似性方法计算用户间的相似性。

定义 2　用户位置矩阵。

$$\boldsymbol{V}_{l(m\times n)} = \begin{bmatrix} v_{1,1} & v_{1,2} & \cdots & v_{1,n-1} & v_{1,n} \\ v_{1,2} & v_{2,2} & \cdots & v_{2,n-1} & v_{2,n} \\ \vdots & \vdots & & \vdots & \vdots \\ v_{m,1} & v_{m,2} & \cdots & v_{m,n-1} & v_{m,n} \end{bmatrix}$$

其中:m 为用户数,n 为在特定 DBSCAN 邻域 ε 下的 ROI 数,v_{ij} 为第个 i 用户对 第 j 个 ROI 的访问次数,l 为划分聚类层次后的第 l 层。

把用户位置看作 n 维位置空间上的向量,用向量间的余弦夹角度量用户间的相似性。设用户 A 和用户 B 在 n 维位置空间上分别表示为向量 $\boldsymbol{U}_A$ 和 $\boldsymbol{U}_B$,则用户 A 和用户 B 之间的相似性为

$$sim(A,B) = \cos(\boldsymbol{U}_A, \boldsymbol{U}_B) = \frac{\boldsymbol{U}_A \cdot \boldsymbol{U}_B}{\| \boldsymbol{U}_A \| \| \boldsymbol{U}_B \|}$$

由于对 POI 在不同邻域 ε 进行聚类,因此在计算了用户在不用的邻域半径下的相似性后需要计算用户总体的相似性。则跨聚类层次的用户相似性为:

$$sim_{\text{overall}} = \sum_{i=1}^{H} \mu sim_i;\ \mu = \beta_i \Big/ \sum_{i=1}^{H} \beta_i$$

其中:H 为划分的总层次数目;sim_i 为第 i 层上的用户相似性;β_i 为第 i 层的相似性权重,用户在越高层次的 ROI 上相似,权重越高。

2　实例分析

2.1　实验数据集

本文以真实的位置服务社交网络作为案例计算用户相似性。嘀咕网是中国最大的基于手机客户端的地理位置签到服务商,可以利用爬虫获取嘀咕网的用户信息。由于社交网站隐私政策的限制,实验中只能获取到对所有用户公开自己签到历史记录的用户相似记录,对于选择只能好友见到详细签

Journal of 《计算机应用》 Computer Applications

到记录的用户,本实验无法获得他们的信息。用户以图 2 的方式使用社交网络。

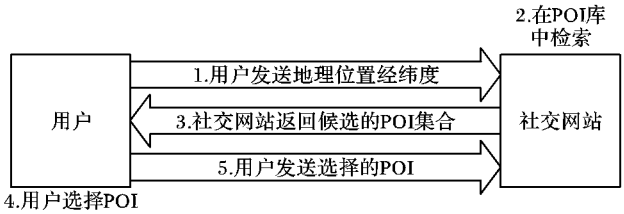
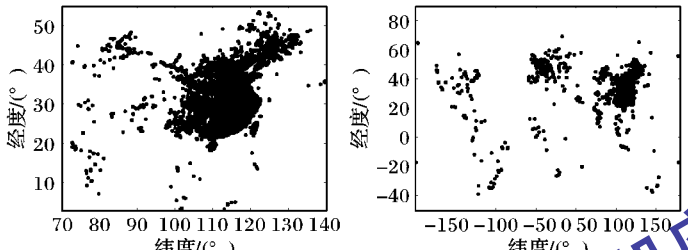


图 2 社交网站使用方式示意图

在实验中,本文共收集 13443 位用户,总共 1346003 次签到,这些用户总共访问 377702 个 POI。将收集到的所有签到数据以散点图的方式展现为图 3 所示,构成了这些用户访问位置的分布地图。从图 3(b)可以看出,国内的签到最为密集,其次是欧洲地区;说明该社交网站用户主要的出国目的地是欧洲地区。当主要观察图 3(a)国内用户的签到分布时,可以看出签到地区主要集中在中东部地区,越往东部用户签到越密集。



(a) 中国签到分布示意图 (b) 全球签到分布示意图

图 3 用户签到位置分布示意图

2.2 查找相似用户

本文利用 DBSCAN 算法对 POI 进行聚类,由于 DBSCAN 算法是基于密度的聚类算法,需要人工设置聚类的地理位置距离 ε 和该地理空间内聚类最少 POI 数目 *MinPts*。实验中,以广州、深圳、北京 3 个用户最多的城市为样本,计算 3 个城市主城区的平均 POI 密度,即平均每平方公里内拥有 POI 的数目,最终确定表 1 的划分为 5 个层次聚类参数。由于聚类的地理面积是以 n^2 增长,因此地理空间内聚类的最少 POI 数目也以相应比例增长,各层次的权重 $\beta_i = 2^i$。后续实验能够很好地找出相似的用户,说明表 1 的 DBSCN 各层聚类参数是合理的。

表 1 DBSCAN 各层参数值

层级	邻域 ε/km	*MinPts*
5	2000	10
4	4000	40
3	8000	160
2	16000	640
1	32000	2560

由于社交网站中多数用户并非活跃用户,因此在计算用户相似性时,本文选择社交网络中签到次数最多的 500 位用户,这些用户总共访问 133782 个地理位置。图 4 为标志在 Google Earth 上,利用本文跨层次聚类计算用户相似性的方法找到的两位最相似用户访问的地理位置。

2.3 方法有效性分析

本文通过查准率和查全率评价所提出的方法。实验中计算 500 位用户间的两两相似性,从中找出各个用户的最相似用户,进而比较找出的每一对用户访问的 POI。若用户与查找出的最相似用户访问相同的 ROI 越多,则说明算法准确度越高。在计算用户相似性时,假定用户访问相近的地理位置次数越多越相似,因此在计算查准率和查全率时,本文比较用户访问次数最多的 Top-*K* 个地理位置的情况。

令 $f(u)$ 为用户访问次数最多的 Top-*K* 个地址所属于的第 i 层的 ROI 集合,则在第 i 层上的查准率的计算公式为:

$$Precision = \frac{|f(u) \cap f(u_r)|}{|f(u_r)|}$$

则两用户在所有聚类邻域 ε 下的准确率为:

$$Precision_{overall} = \sum_{i=1}^{H} \mu Precision_i;\ \mu = \beta_i / \left(\sum_{i=1}^{H} \beta_i \right)$$

其中:u_r 为选择出的和用户 u 相似性最高的用户,H 为总的层次数目,β_i 为第 i 层的相似性权重。

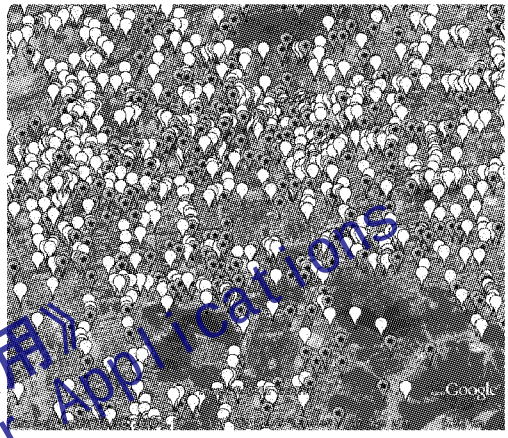


图 4 最相似两用户的访问地理位置

类似地,计算查全率的公式为:

$$Recall = \frac{|f(u) \cap f(u_r)|}{|f(u)|}$$

$$Recall_{overall} = \sum_{i=1}^{H} \mu Recall_i;\ \mu = \beta_i / \left(\sum_{i=1}^{H} \beta_i \right)$$

本文对所有相似用户计算出的查全率和查准率求平均值,并将方法和不划分层次仅在最高层级一次聚类以及 Jaccard 系数[11]计算相似性的方法进行比较,得到图 5 的结果。从图 5 中可以看出,本文的方法要比仅在最高层级聚类判断用户相似性的方法有更高的查准率,可见同时衡量不同空间比例尺的用户相似性要比仅在最细粒度上比较更准确。同时比 Jaccard 系数计算相似性的方法有更高的查准率和查全率,并且相似的用户最经常访问的位置相似度最高。因为,Jaccard 系数只能考虑用户所访问的地理位置上的相似性,而无法计算用户访问地理位置的次数。

由于利用 Top-*K* POI 验证本文方法,为了体现用户访问 Top-*K* POI 的次数已经覆盖用户大部分的访问次数,定义 *TopKCoverage*(*K*)[5] 为用户所访问的 Top-*K* 个 POI 的次数占用户访问全部 POI 次数的比例。

$$TopKCoverage(K) = \frac{\sum_{u \in U} \frac{Topk(u)}{Total(u)}}{|U|}$$

其中:U 为所选用户集合,$Topk(u)$ 为第 u 为用户访问 Top-*K* POI 的次数,$Total(u)$ 为第 u 位用户访问的总次数。

从图 6 可以看出,用户最经常访问的前 50 个 POI 的次数占到用户使用总访问次数的 50.9%,而访问 Top-150 的 POI 已经占到总访问次数的 77.2%。因此利用相对小的 Top-*K* 个 POI 集合可以覆盖用户最经常访问的地理位置。

Journal of Computer Applications 《计算机应用》

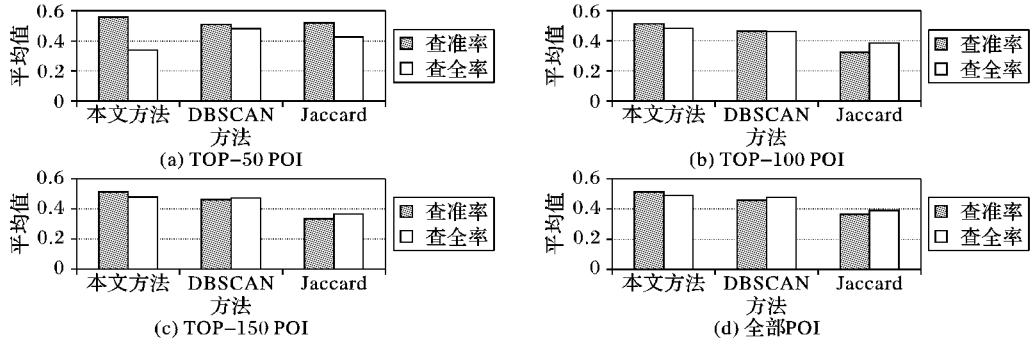


图 5　Top-*K* POI 的方法查准率和查全率

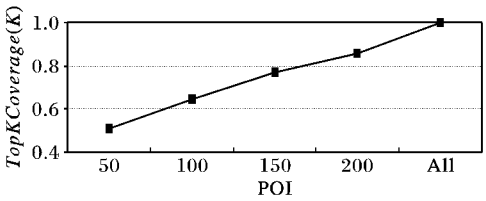


图 6　Top-*K* POI 签到次数覆盖率

3　结语

本文基于位置服务社交网络真实的用户访问数据,提出划分空间层次地理位置聚类的方法,该方法在各种邻域范围下对用户访问的地理位置聚类,考察用户访问各聚类区域的次数,并利用向量空间模型计算余弦相似性的方法度量用户行为轨迹上的相似性。通过统计分析查找出的相似用户及其访问的地理位置,运用计算查准率和查全率表明了本方法的有效性。

参考文献:

[1] BENEVENUTO F, RODRIGUES T, CHA M, *et al*. Characterizing user behavior in online social networks [C]// Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement. New York: ACM, 2009: 49 – 62.

[2] SCELLATO S, MASCOLO C. Measuring user activity on an online location-based social network [C]// Proceedings of Third International Workshop on Network Science for Communication Networks. Piscataway: IEEE, 2011: 918 – 923.

[3] CHENG Z, CAVERLEE J, LEE K, *et al*. Exploring millions of footprints in location sharing services [C]// ICWSM 11: Fifth International AAAI Conference on Weblogs and Social Media. Barcelona, Spain: AAAI, 2011: 282.

[4] NOULAS A, SCELLATO S, MASCOLO C, *et al*. An empirical study of geographic user activity patterns in foursquare [C]// ICWSM 11: Fifth International AAAI Conference on Weblogs and Social Media. Barcelona, Spain: AAAI, 2011: 570 – 573.

[5] CHO E, MYERS S A, LESKOVEC J. Friendship and mobility: User movement in location-based social networks [C]// Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2011: 1082 – 1090.

[6] HUNG C-C, CHANG C-W, PENG W-C. Mining trajectory profiles for discovering user communities [C]// Proceedings of the 2009 International Workshop on Location Based Social Networks. New York: ACM, 2009: 1 – 8.

[7] LI Q, ZHENG Y, XIE X, *et al*. Mining user similarity based on location history [C]// Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. New York: ACM, 2008: 1 – 10.

[8] LEE M-J, CHUNG C-W. A user similarity calculation based on the location for social network services [C]// DASFAA'11: Proceedings of the 16th International Conference on Database Systems for Advanced Applications, LNCS 6587. Berlin: Springer-Verlag, 2011, 1: 38 – 52.

[9] ESTER M, KRIEGEL H P, SANDER J, *et al*. A density based algorithm for discovering clusters in large spatial databases with noise [C]// Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining. Portland: AAAI, 1996: 226 – 231.

[10] LEE R, SUMIYA K. Measuring geographical regularities of crowd behaviors for Twitter-based geo-social event detection [C]// Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks. New York, ACM, 2010: 1 – 10.

[11] TAN P N, STEINBACH M, KUMAR V. Introduction to data mining [M]. Boston: Pearson Addison Wesley, 2006.

《计算机应用》Journal of Computer Applications

第一届中国互联网学术会议征文通知

Internet Conference of China (CCF ICoC 2012)

由中国计算机学会主办,互联网专委会协办,清华大学承办的"第一届中国互联网学术会议(CCF ICoC 2012)"将于 2012 年 5 月 31 至 6 月 1 日在北京清华大学召开。会议将就互联网领域相关理论与技术的最新研究进展和发展趋势开展广泛深入的学术交流,并特邀著名专家学者作大会报告。

征文涉及的领域包括但不限于:未来互联网体系结构、互联网路由、网络安全、网络管理、数据中心网络、绿色网络、无线网络、P2P 网络、移动互联网、物联网,以及其他互联网研究领域。会议录用论文将分别推荐到《计算机学报》(EI)、《中国科技论文》(核心期刊正刊)和《小型微型计算机系统》(核心期刊正刊)等刊物上发表。会议还将评选优秀论文奖。

来稿内容应属于作者的科研成果,数据真实、可靠,未公开发表过,引用他人成果已注明出处,署名无争议,字数一般不超过 1 万字(不作严格限制)。来稿请按《计算机学报》的格式排版。此次将采取网上投稿方式,投稿地址请见会议网站 http://www.ccf-internet.edu.cn/ccf2012/。

重要时间

论文提交截止日期:2012 年 03 月 15 日

论文录用通知日期:2012 年 04 月 30 日

正式论文提交日期:2012 年 05 月 15 日

联系人:刘畅

(62603001,ccf-internet@cernet.edu.cn)