



CASoRT 系统中基于聚集特性的在线流行度预测方法

陈皇卿^{1,2}, 钟晓峰^{1,2}, 孙 剑^{1,2}, 王 京^{1,2}

(1.清华大学信息科学与技术国家实验室 北京 100084; 2.清华大学电子工程系 北京 100084)

摘 要:少数在线热门内容会在短时间内吸引大量用户的访问,并占用大量的网络传输资源。如果能预知内容的热门程度(即流行度)并将热门内容广播给潜在用户,将极大地节省网络传输资源,这正是 CASoRT 系统的主要功能。通过对国内商业蜂窝通信系统中收集的相关数据进行分析,发现在用户行为、地理位置、数据内容等方面存在明显的聚集特性。根据上述特性给出了两个流行度预测算法,即对数线性和恒定比例模型,并使用最优观察门限改善两算法的性能。通过对两算法仿真结果的比较,对数线性模型表现更优,被选作系统的在线流行度预测方法。

关键词:流行度预测;聚集特性;无线蜂窝网;CASoRT

doi: 10.11959/j.issn.1000-0801.2015221

Online Popularity Prediction Scheme Based on Converging Property in Content Aware Soft Real Time Media Broadcast System

Chen Huangqing^{1,2}, Zhong Xiaofeng^{1,2}, Sun Jian^{1,2}, Wang Jing^{1,2}

(1. National Laboratory of Information Science and Technology of Tsinghua University, Beijing 100084, China;

2. Department of Electronic Engineering, Tsinghua University, Beijing 100084, China)

Abstract: A small number of online popular contents are often clicked by a great quantity of users in a short period, and take the most of the wireless cellular network traffic. With popularity prediction, the popular contents would be broadcasted to the potential users for saving a lot of transmitting resource, as illustrated in content aware soft real time media broadcast (CASoRT) system. With the data set collected from the Chinese commercial cellular network, the converging property of web contents, users and geographic positions in online news was shown. Then, two prediction schemes such as linear log and constant scaling model were proposed to estimate the popularity of online news, and improved by an optimal observation threshold. After comparison of simulation results, the linear log model performs better.

Key words: popularity prediction, converging property, wireless cellular network, CASoRT

收稿日期: 2015-06-23; 修回日期: 2015-09-06

基金项目: 国家重点基础研究发展计划(“973”计划)基金资助项目(No. 2012CB316000), 国家科技重大专项基金资助项目(No. 2015ZX03002010-002)

Foundation Items: The National Basic Research Program of China (973 Program) (No. 2012CB316000), National S&T Major Project (No. 2015ZX03002010-002)

论文引用格式: 陈皇卿, 钟晓峰, 孙剑等. CASoRT 系统中基于聚集特性的在线流行度预测方法. 电信科学, 2015221

Cheng H Q, Zhong X F, Sun J, *et al.* Online popularity prediction scheme based on converging properties in content aware soft real time media broadcast system. Telecommunications Science, 2015221



1 引言

随着智能手机和无线通信设备的广泛使用,多媒体数据传输占用着越来越多的无线蜂窝通信资源。多媒体数据分组包括新闻、文档、图片和视频等种类,其中新闻类数据无论在无线蜂窝网中还是互联网上都是访问最为频繁的网络数据。因此,主要针对通过无线蜂窝网访问的新闻数据进行分析 and 研究。事实上,网络新闻存在着明显的聚集特性,主要体现在少数热门新闻内容占据着网站总访问量的主体;用户存在明显的活跃用户和非活跃用户,对总访问量的贡献程度差别很大;用户访问新闻时的地理位置主要集中在少数几个区域。通过对数据进行分析,找出其中的热门内容,然后将热门内容在适当的时候(比如无线网络比较空闲时)进行广播,把热门内容提前存储到用户移动设备的本地存储器。用户在请求网络数据时首先对本地存储器进行搜索,若本地存储器有该数据则直接读取,未发现则向服务器申请传输该数据。这种方法可以大大地降低无线网络在通信繁忙时段的通信量,还可以节省无线蜂窝通信网络的传输资源,从而提高能效和谱效。CASoRT系统的无线蜂窝推送功能就是上述过程的具体实现^[1]。要想实现无线蜂窝推送,首先要找到预测热门内容的方法。

关于网络多媒体内容流行度预测的方法有很多,主要关注其使用的网络数据流行度指标和使用的预测方法的类型。大量文献使用了诸如评论、投票和点击量等流行度指标和各种不同类型的预测方法,比如对数线性(linear log)模型、恒定比例(constant scaling)模型、部分匹配(partial match)模型、对数正态分布(log-normal distribution)模型和Cox比例风险回归(cox proportional hazard regression)模型等。这些模型中的任意一个均无法对所有的网络数据进行准确的流行度预测,只能针对特定的数据集或者数据集中的一部分进行比较准确的流行度预测。

目前在流行度预测方法方面的研究主要有如下文献,和本文类似,它们都是通过对网络内容发布后某较短时间段(即观察时间)内的访问量大小来预测该网络内容的总访问量,也就是流行度。

- 参考文献[2]在对20Minutes网站中新闻评论数量研究的基础上,提出了一个简单的线性预测模型并分析了利用初期评论量来预测新闻流行度的局限性。
- 参考文献[3]通过对Telegraaf网站中新闻评论的研

究,给出了对数正态模型和负二项分布模型两种方法来预测新闻流行度,然后比较它们的性能。

- 参考文献[4]针对Slashdot网站中的新闻评论进行分析,提出用两个对数正态模型的组合进行流行度的预测和评论时间分布的估计。
- 参考文献[5]针对Digg网站中故事的评论和Youtube网站中视频的投票进行研究,然后利用对数线性模型和恒定比例模型来预测网络内容的流行度。

本文首先对从中国移动通信有限公司(以下简称中国移动)收集来的通过无线蜂窝网络传输的在线网络内容数据集进行研究,通过该数据集的特点来反映无线网络通信量的变化特征。经过数据处理,选取凤凰网(www.ifeng.com)上的新闻内容作为其有代表性的子数据集进行分析。该网站的在线新闻点击量作为流行度指标,即某新闻点击量越多,表示该新闻越热门。相比评论和投票,点击量能够更直接地反映出无线蜂窝网络中通信量的变化情况。研究发现,该数据集除了在数据内容、地理位置和用户行为方面的聚集特性以外,某些热门新闻在访问量(即点击量)快速上升之前,存在少量无规律访问,这些无规律访问会给流行度预测带来较大的负面影响。为了减轻这种影响,使用一个在固定时间段(如6 min)内访问次数的门限值来改变观察区间的起始时间点,使得该起始时间点能更加靠近访问量快速上升的阶段。在了解数据特性之后,提出了两种流行度预测方法,即对数线性模型和恒定比例模型,两种方法的预测准确度会因为门限值的使用而得到很大改善。基于对数据集的研究,本文给出了最佳观察门限值,然后将两个预测模型在使用最佳门限值时的预测准确度进行比较,仿真结果发现对数线性方法性能更好,因此被选定作为CASoRT系统的流行度预测方法。

2 数据特性

本文使用的数据集是中国移动通信集团有限公司浙江分公司(以下简称浙江移动)一周(2014年4月19-25日)的所有通过无线蜂窝网访问的互联网内容。该数据集由大量访问记录组成,记录中包含用户识别号、访问起始时间和终止时间、访问页面URL、区域位置号和小区位置号等属性。通过访问页面URL过滤,筛选出凤凰网上的在线网络新闻作为研究子数据集。由于凤凰网是我国第四大新闻门户网站,拥有丰富的新闻内容和各类不同的用户群体,而且该网站的数据量大小非常适合分析和研究,因此

被选定为研究对象。该子数据集的特性具有代表性,可以基本反映出所有网络内容的相关特性。图 1 展示了凤凰网所有网络新闻的访问量以 h 为单位随时间的分布情况,可以看出该网站每天的总访问量无明显差别。同时可以发现,在每一天中都有明显的访问量尖峰和访问量低谷,且尖峰和低谷出现的时间段相似。无线蜂窝推送策略需要在合适的时候对热门内容进行广播,如果在非访问量高峰时段特别是低谷时广播热门内容,在节省资源的同时还能达到平抑访问量尖峰的效果。

2.1 数据聚集特性

数据的聚集特性主要体现在用户行为、网络内容和地理位置 3 个方面^[6]。用户的聚集特性如图 2 所示。可以看到,少量的活跃用户占据着网站总访问量的主体,在前 20% 最活跃的用户中,其访问量之和占总访问量的比例达到 70%,因此,在研究无线蜂窝推送策略时应重点关注活跃用户。

另一个聚集特性体现在新闻内容上。用户在 7 天内共点击了凤凰网上 36 804 个不同的新闻标题,它们代表着不同的新闻内容。热门的网络内容和冷门的网络内容在其总访问量也就是流行度上的差别很大,最热门的标题有超过 3 万次访问,而很多非热门标题访问量仅为 1 次。如图 3

所示,少数热门标题占据着网站总访问量的主体。前 5% 最热门的内容的访问量占总访问量的 88%,而前 10% 最热门的标题的访问量约占总访问量的 92%。网络内容的聚集特性非常明显,因此,如果能从大量的网络数据中预测出热门内容,然后进行推送,必将节省大量的无线网络传输资源。

在地理位置方面,若从用户整体角度来看,聚集特性并不明显,但对于单个用户来讲,其访问网络内容时所在地理位置存在聚集特性。用户一般在少数几个固定的小区内访问网络内容,并且其访问量主要发生在若干个活跃小区内。统计发现,用户在其最活跃的小区内平均访问了 58% 的网络内容,而在其最活跃的前两个小区内平均访问了超过 80% 的内容。因此将热门内容推送给潜在用户时,会优先考虑向用户的最活跃小区进行推送。

2.2 数据随时间的变化特性

除了聚集特性之外,对于任一新闻标题而言,也存在访问量随时间变化的特性。每个标题访问量随时间变化的特性各不相同,但也存在着共性。不同点主要体现在每个标题的访问持续时间不同,有些标题的访问量主要集中在 24 h 内,而有些标题可能持续几天被大量点击。共性主要体现在每个标题访问量随时间的变化都符合图 1 所示的

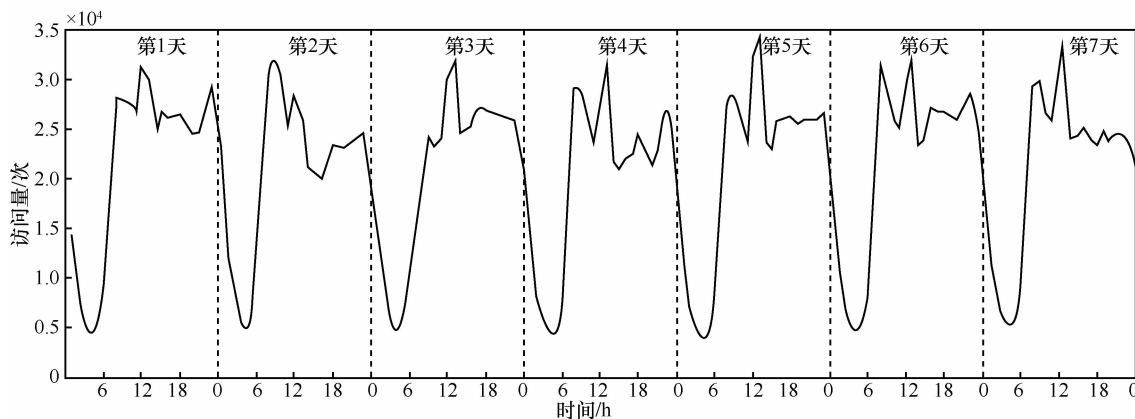


图 1 凤凰网所有网络新闻访问量随时间变化曲线

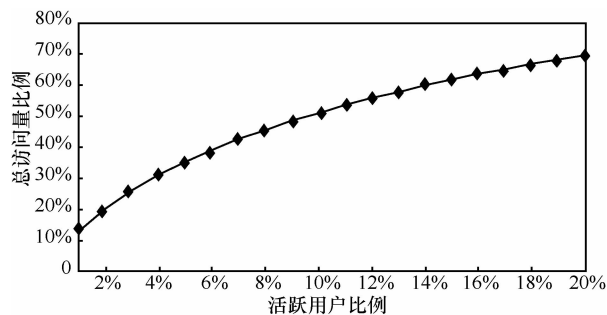


图 2 用户的聚集特性

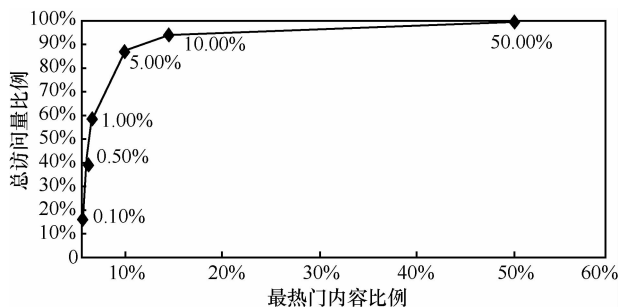


图 3 新闻内容的聚集特性



周期,即在一天内有明显的访问量高峰和低谷。在这些性质各不相同的标题中,那些访问量集中在某个较短时间比如 24 h 内的标题,在预测其流行度时可以达到较高的准确度;而那些访问持续时间比较长,访问量在时间上并不集中的标题,无法比较准确地预测其流行度,即预测流行度和实际流行度之间存在比较大的偏差。

另一个影响预测准确度的特点是部分标题在访问量迅速上升之前,存在少量的无规律点击,这可能是由于该新闻在发布时并未第一时间放到新闻网站的主页上,典型的该类型新闻标题如图 4 所示。这些点击会给预测带来严重的误差,甚至可能让预测算法将一个热门新闻内容误预测成一个非热门内容。使用访问量观察门限值来减轻该影响,即将标题观察时间的起始点定为某时刻,该时刻为第一个其之前 6 min 内的访问量超过门限值的时间点。

3 流行度预测算法

由数据特性可以推知,如果通过前若干个小时(如 2 h)观测到的访问量来预测该标题在整个访问持续时间里的总访问量,那么显然存在如下一般规律:观察时间里该标题的访问量越多,则该标题是热门内容的可能性越大。因此,考虑使用线性模型作为流行度预测方法,常用的线性预测方法主要有对数线性模型和恒定比例模型。这里需要注意的是,观测时段的起始时间点并不一定是新闻的发布时间,这主要取决于是否使用观察门限。

3.1 预测算法

线性预测算法是一类较为简单的流行度估计方法。在观察时间内的访问量和该标题的总访问量可能

存在线性关系,但事实上该线性关系并不显著。如果将观察时间内的访问量和总访问量置于自然对数尺度下,该线性关系会更为明显,特别是对于具有长尾效应的数据集。本文给出两种线性预测算法,并在随后对其性能进行比较。

(1) 对数线性模型

对数线性模型的主要目的是对变量进行对数变换,从而找出两个变量间的线性关系。该模型最早由参考文献[5]提出,该模型下的流行度预测公式如式(1)所示:

$$\ln N_i = k(t) \ln N_i(t) + b(t) \quad (1)$$

其中, N_i 为标题 i 的流行度预测值,而 $N_i(t)$ 表示标题 i 在观察时间内的访问量,该观察时间长度为 t 。 $k(t)$ 和 $b(t)$ 为反映线性关系的相关参数,它们随观察时间长度的变化而改变,可以在训练数据集中利用线性回归方法找到其最优值。

(2) 恒定比例模型

恒定比例模型最初也由参考文献[5]提出。该模型下的流行度预测公式如式(2)所示:

$$N_i' = \alpha(t) N_i(t) \quad (2)$$

其中, N_i' 为标题 i 的预测总访问量, $\alpha(t)$ 可通过式(3)计算得出:

$$\alpha(t) = \frac{\sum_i \frac{N_i(t)}{N_i'}}{\sum_i \left[\frac{N_i(t)}{N_i'} \right]^2} \quad (3)$$

3.2 最佳观察门限及算法性能比较

如前所述,门限值选取的目的是将观察时间的起点置于访问量快速上升之时,从而减轻偶然无规律访问对流行

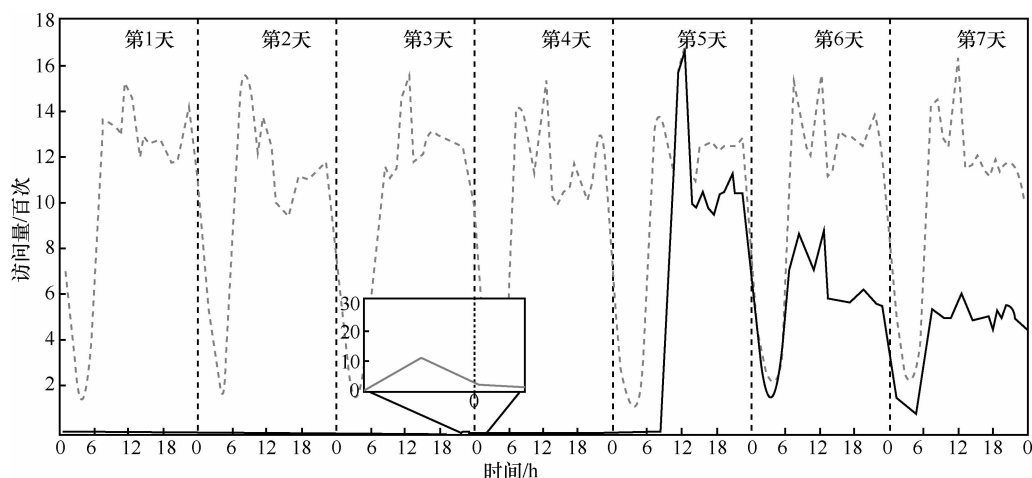


图4 典型网络标题访问量随时间变化特性

度预测准确性的影响。皮尔逊相关系数 (pearson correlation coefficient, PCC) 用来表征两个向量的线性相关度, 该值越大, 表示线性相关性越强。在此对观察时间内的所有标题访问量的自然对数和所有标题总访问量 (即流行度) 的自然对数之间的线性相关度进行分析。如图 5 所示, 当观察门限从 0 增加到 6 时, 任意长度观测时间的皮尔逊相关系数均逐渐增大, 而观察门限取 8 和 10 时, 皮尔逊相关系数相比门限值为 6 时呈下降趋势, 其中门限值取 4~10 时皮尔逊系数无明显差别。特别的, 在观察时间长度为 24 h, 门限值不小于 4 时, 皮尔逊相关系数接近 0.95, 也就是说此情况下 24 h 内的访问量和总访问量存在很强的线性关系, 这和数据集的特征是相符的。

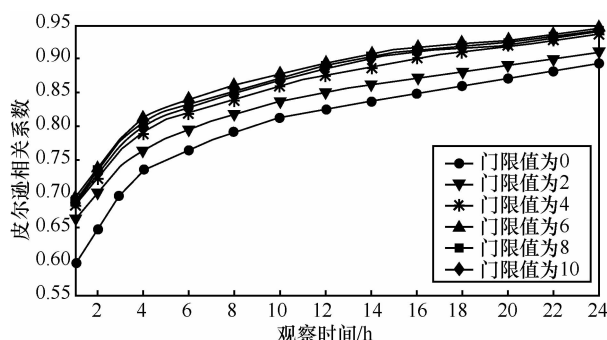


图 5 不同观察门限值下皮尔逊相关系数随观察时间长度变化规律

在选择最佳观察门限时, 需要考虑两个方面的因素: 一是希望尽可能多地保留数据集的信息量, 即被门限值排除在外的访问量所占比例应尽量小; 另一方面是取该门限值时的皮尔逊相关系数尽可能得大, 线性相关性越强, 预测性能也会越好。因此, 选择 4 作为最佳观察门限, 在取该门限时, 部分冷门新闻内容的访问和热门新闻内容的无规律点击可能不会计入观察时间段内的访问量, 从而造成部分信息的丢失。统计表明, 这部分被最佳观察门限值排除在外的访问量占所有标题总访问量的比例低于 8%, 这部分丢弃的信息基本不会对预测产生影响。

平均平方误差 (mean squared error, MSE) 被用作两种预测算法的性能评价指标。MSE 值越小, 说明该算法预测准确度越好。该指标的计算式见式 (4):

$$MSE = \sum_i^n (N_i' - N_i(t))^2 / n \text{ 或 } \sum_i^n (N_i - N_i(t))^2 / n \quad (4)$$

如图 6 所示, 在未使用观察门限时, 恒定比例模型在

不同观察时间长度的 MSE 值均小于对数线性模型。而使用最佳观察门限之后, 两个预测算法的性能都得到了很大提升, 更重要的是, 在观察时间长度为 1~8 h 时, 对数线性模型的 MSE 值小于恒定比例模型, 而观察时间长度取 9~24 h 时, 两个模型间的 MSE 值无较大差别。在能提供足够预测准确度的前提下, 如能越早预测出某新闻标题的流行度, 就可以在传输资源节省上获得更大收益。考虑到数据集的特性, 一般在 2~6 h 选择观察时间长度, 而在此区间内, 对数线性模型有着更好的性能, 因此被选作系统的流行度预测算法。显然, 若观察时间长度超过 24 h, 再进行流行度预测就没有意义了, 因为大部分标题的访问量都集中在 24 h 以内。

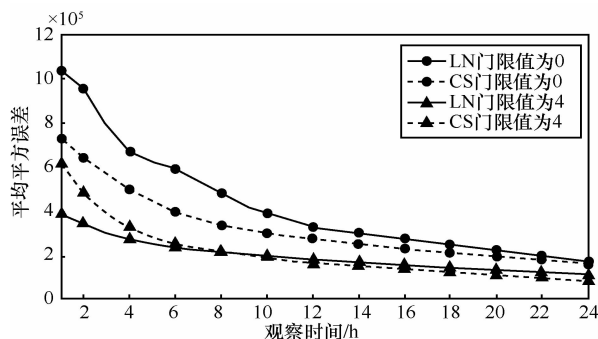


图 6 两线性算法在不同观察门限值下的 MSE 随观察时间长度变化规律

4 结束语

本文提出了一种基于最佳观察门限的对数线性预测方法, 该最佳观察门限大大提高了对数线性方法的预测准确度。观察门限的引入主要是基于对凤凰网在线新闻访问量的特点分析。了解到在线新闻不仅在内容、用户和地理位置方面存在着聚集特性, 部分内容还存在一些不规律的偶然访问, 滤除这些偶然访问可以大大改善预测性能。基于对本文数据集进行计算和分析, 确定了最佳观察门限为 4, 然后对两种预测算法使用最佳观察门限时的性能进行比较, 最后选定对数线性模型作为预测算法。

本文提出的预测算法能较好地估计网络内容的流行度, 同时在线新闻的相关特性也反映出无线蜂窝网通信量的变化情况和推送策略可能带来的能效收益。因此, 本预测算法能通过前若干个小时 (如 2 h) 观测到的访问量估计该网络内容总访问量从而预测该内容是否热门, 然后将热门内容通过无线蜂窝推送系统进行



广播,不仅能降低网络在高峰时段的拥塞度,还能节省大量的传输资源。

参考文献

- 1 Zhong X F, Zhao M, Zhou S D, *et al.* Content aware soft real time media broadcast (casort). Proceedings of the Third International Conference on Communications and Networking, Hangzhou, China, 2008
- 2 Tatar A, Leguay J, Antoniadis P, *et al.* Predicting the popularity of online articles based on user comments. Proceedings of the International Conference on Web Intelligence, Mining and Semantics, Sogndal, Norway, 2011
- 3 Tsagkias M, Weerkamp W, De R M. News comments: exploring, modeling and online prediction. Proceedings of the 32nd European Conference on Advances in Information Retrieval (ECIR2010), Berlin, Germany, 2010
- 4 Kaltenbrunner A, Gomez V, Lopez V. Description and prediction of slashdot activity. Proceedings of the 5th Latin American Web Congress, Santiago, Chile, 2007
- 5 Szabo G, Huberman B A. Predicting the popularity of online content. Communications of the ACM, 2010, 53(8):80~88
- 6 Sun J, Zhong X, Zhou X, *et al.* Recommendation scheme based on converging properties for contents broadcasting. Proceedings of the 11th International Conference on Natural Computation, Anaheim, California, USA, 2015

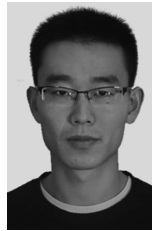
[作者简介]



陈皇卿,男,清华大学信息科学与技术国家实验室、清华大学电子工程系博士生,主要研究方向为未来无线网络用户行为与业务模型、内容感知的业务服务模式(CASoRT)。



钟晓峰,男,博士,清华大学信息科学与技术国家实验室、清华大学电子工程系副研究员,主要研究方向为无线网络技术,包括4G/5G移动通信系统、分布式无线通信系统、无线自组织与协作传输网络等。作为课题负责人承担“973”计划项目、国家科技重大专项等7项国家课题及相关部委课题等。在无线网络的传输、接入协议、业务特征等方面开展了深入的研究,提出了多个创新算法。已发表(录用)论文40余篇,申请专利11项。



孙剑,男,清华大学信息科学与技术国家实验室、清华大学电子工程系硕士生,主要研究方向为未来无线网络用户行为与业务模型、内容感知的业务服务模式(CASoRT)。



王京,男,清华大学博士生导师,清华大学信息科学与技术国家实验室、清华大学电子工程系副主任,清华大学信息技术研究院无线与移动通信技术研究中心主任,微波与数字通信国家重点实验室副主任兼无线传输与个人通信研究室主任;中国第三代移动通信研究开发项目(C3G)总体组成员,第三代移动通信技术实验(3G-TEG)专家组专家,国家“十五”“863”计划未来移动通信(FUTURE)项目总体组专家,中国电子学会高级会员,中国通信学会无线及移动通信委员会委员。主要研究方向为无线移动通信和宽带无线传输技术,包括B3G/4G移动通信、WCDMA、TD-SCDMA系统研究以及调制解调技术、纠错编译码技术、移动通信抗衰落技术、MIMO与时空联合数字信号处理技术、分布式无线通信系统等。