



灵活适应不同业务的个性化推荐系统研究

陶彩霞¹, 袁海², 陈康¹, 马安华²

(1. 中国电信股份有限公司广东研究院 广州 510630;

2. 中国电信股份有限公司江苏分公司 南京 210037)

摘要: 互联网技术的迅速发展, 尤其是以个性化为主要特点的 Web 2.0 的不断成熟, 使得大量信息同时呈现在人们面前, 个性化推荐技术的价值日益凸显。提出了一种灵活适应不同业务的个性化推荐系统设计方案, 采集并分析用户的所有显式行为和隐式行为, 结合用户行为配置文件和熵值法分析用户对物品的兴趣度, 并引入时间遗忘函数, 解决用户兴趣漂移问题, 然后基于协同过滤技术得到用户的个性化推荐列表。通过实际数据进行测试, 给出了系统推荐效果评估分析。

关键词: 个性化推荐系统; 用户兴趣分析; 熵; 兴趣漂移; 协同过滤

doi: 10.3969/j.issn.1000-0801.2014.08.019

Research on Personalized Recommender System Adapting to Different Business

Tao Caixia¹, Yuan Hai², Chen Kang¹, Ma Anhua²

(1. Guangdong Research Institute of China Telecom Co., Ltd., Guangzhou 510630, China;

2. Jiangsu Branch of China Telecom Co., Ltd., Nanjing 210037, China)

Abstract: The rapid development of internet technology, especially the developed of Web 2.0 which has the main feature of personality, making a lot of information in front of people at the same time. The value of personalized recommendation technology has become increasingly prominent. A design of personalized recommender system adapting to different business was presented, including the collection and analysis of user's all explicit and implicit behavior, using user behavior profiles and entropy method to determine the weight of the behavior for analyzing user interest, the introduction of time forgetting function to solve the problem of user interest drift, and getting a list of the user's personalized recommendation based on collaborative filtering technology. Finally, the system test assessment analysis based on the actual data was provided.

Key words: personalized recommender system, user interest analysis, entropy, interest drift, collaborative filtering

1 引言

互联网技术的迅速发展, 尤其是以个性化为主要特点的 Web 2.0 的不断成熟, 使得大量信息同时呈现在人们面前。信息消费者和信息生产者都面临着很大的挑战: 信息消费者如何从海量信息中获得自己感兴趣的内容, 而信息

生产者如何让自己的信息被关注。传统的搜索引擎, 如谷歌、百度等, 只能呈现给所有用户相同的排序结果, 无法针对不同用户、不同时期的兴趣爱好提供相应的服务。个性化推荐技术就是针对这个问题而提出的, 其为不同用户提供不同的服务, 以满足不同的需求, 一方面帮助用户发现对自己有价值的信息; 另一方面使信息能够展现给对它感



兴趣的用户,从而实现信息消费者和信息生产者的双赢。

个性化推荐系统通过建立用户与物品/信息之间的二元关系,利用用户已有的操作行为或者相似性关系挖掘每个用户潜在感兴趣的对象,进而进行主动的个性化推荐。在日趋激烈的竞争环境下,个性化推荐系统已经不仅仅是一种商业营销手段,更重要的是能充分提高业务平台的服务质量和访问效率,增进用户的黏性,并吸引更多的用户。

用户兴趣分析是个性化推荐的基础和核心,只有系统很好地理解了用户的兴趣需求,才可能推荐出用户满意的信息。用户兴趣分析可以通过用户手工输入兴趣点或评分信息来建模,这种方式虽然简单易行,但完全依赖于用户,一般主动评分的用户不多,容易出现由于用户行为信息单一和数量较少带来的数据稀疏性的问题;并且用户的输入可能具有随意性或者输入错误,无法准确表达用户的兴趣。理想的方式是全面收集用户浏览、收藏、下载、分享的所有行为并进行自动兴趣建模。

不同的行为反映的兴趣程度不同,目前通常的做法是人工根据经验确定每个行为的权重系数,但通常很难人工确定什么行为更加重要,从而难以确定行为的权重系数,而且这种方式存在人为主观性和领域局限性,这也是目前很多推荐系统只针对特定领域甚至特定平台构建的一个原因。针对这些问题,本文采集并分析用户的所有显式行为和隐式行为,结合用户配置文件和熵值法,采用系统自动对用户兴趣进行建模的方式,基于协同过滤技术提出了一

种灵活适应不同业务的个性化推荐系统设计方案。

2 系统架构与模型算法设计

本文设计的推荐系统分为三大模块,包括进行数据预处理的用户行为记录预处理模块、作为连接模块的用户兴趣分析模块和个性化推荐算法模块,如图1所示。三大模块以前面模块为后面模块提供输入信息联结而成。

用户行为记录预处理模块获取用户行为信息,如用户对某款游戏的下载、分享、推荐等具体行为,对数据集进行去噪、归一化等预处理操作,并转化为统一的数据格式。

用户兴趣分析模块是不可或缺的连接模块,通过对上一模块行为记录数据进行统计分析,得到用户兴趣偏好,并将用户兴趣偏好输入个性化推荐算法模块,通过推荐算法模块得到推荐结果。用户兴趣偏好可以通过用户显式评分来确定,但这种方法的效果相当有限,因为用户通常都不愿意提供反馈,最终得到的信息也会由于它们主要来自于那些愿意提供反馈的用户而造成数据稀疏性问题。本文利用熵值法对收集到的用户的所有显式行为和隐式行为数据进行赋权,再取用户行为数据向量的加权之和,结合时间遗忘函数解决用户兴趣漂移问题,分析用户对物品的兴趣度,得到三元组(用户,物品,兴趣度),作为统一形式输入个性化推荐算法模块,这种方式在一定程度上缓解了数据稀疏性问题。

个性化推荐算法模块中,目前主要的推荐算法包括:

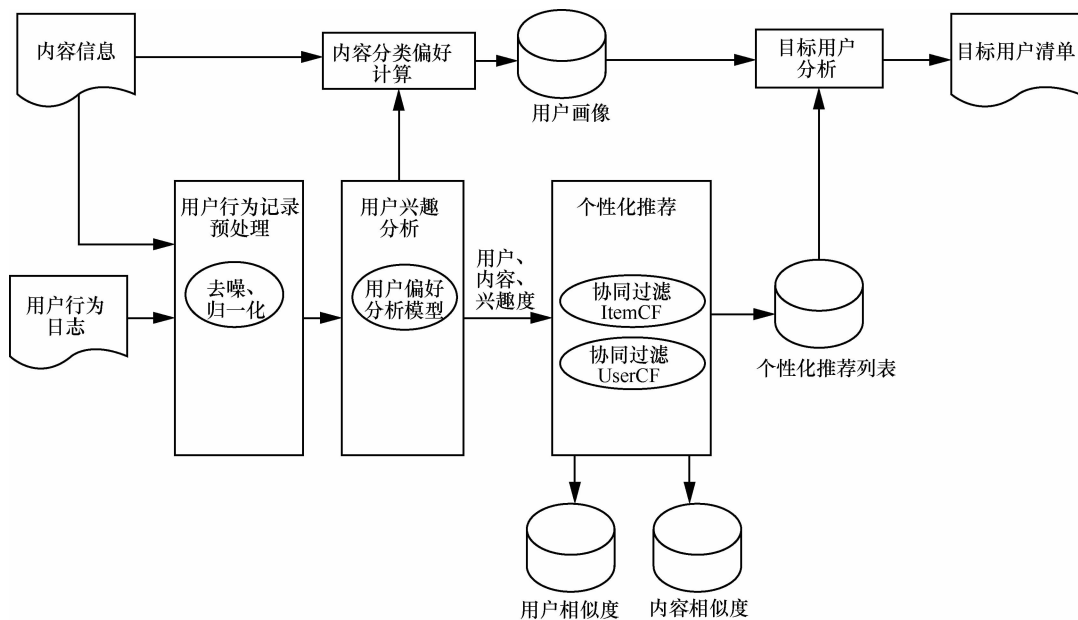


图1 系统架构

协同过滤推荐、基于内容的推荐、基于关联规则的推荐、基于效用的推荐、基于知识的推荐以及选取若干经典推荐算法组合使用的组合推荐。这里主要采用协同过滤推荐算法。

2.1 用户行为配置文件初始化

不同业务类型的用户行为也不同,如游戏、视频和电子商务平台(电商平台)的用户行为各不相同,见表 1。

表 1 各类业务的用户行为示例

业务类别	用户行为
游戏	下载、收藏、分享、推荐、评论、评分、使用频率、搜索、在线时长
视频	观看时长、收藏、搜索、评论、评分、分享、推荐
电商	浏览、加入购物车、购买、收藏、评论、评分、搜索、点击、分享、推荐

根据业务平台提供的用户行为数据,设置用户行为配置文件,主要包括业务类型、用户行为类型、行为编码、行为初始权重 weight0、行为动态权重 weight1 等,如图 2 所示。

```
<value> game </value> <!-- 游戏业务-->
<properties>
  <property>
    <name> collect </name> <!--收藏-->
    <value> 001 </value>
    <weight0> 0.07895 </weight0>
    <weight1> 0.07895 </weight1>
  </property>
  <property>
    <name> download </name> <!--下载-->
    <value> 007 </value>
    <weight0> 0.131579 </weight0>
    <weight1> 0.131579 </weight1>
  </property>
  <property>
    <name> recommend </name> <!--推荐-->
    <value> 003 </value>
    <weight0> 0.15789 </weight0>
    <weight1> 0.15789 </weight1>
  </property>
</properties>
```

图 2 用户行为配置文件示例

2.2 用户行为记录预处理模块

用户原始行为记录包括的信息主要有用户 ID、物品 ID、行为 ID、行为数值、行为发生时间,处理步骤主要包括规范处理和用户兴趣漂移(interest drift)处理两个步骤。

(1)规范化处理

用户行为记录预处理模块对原始数据记录进行去噪和规范化处理,转换为统一的数据格式。首先对各行为进行量化,即将原始用户行为记录处理为系统可识别的行为 ID 及行为数值,然后对用户行为记录进行规范化处理,不同的用户行为有不同的数据规范化处理方式,如视频用户的观看时长规范化公式为:

$$t' = \frac{t}{t_{-a}} \tag{1}$$

其中, t' 为规范化新数据, t 为观看时长, t_{-a} 为视频总时长。

(2)用户兴趣漂移处理

用户的兴趣会随着时间的推移而发生变化,一些用户原本感兴趣的主题会被渐渐遗忘,新的兴趣主题会逐渐产生,用户兴趣的这一渐变过程又被称为兴趣漂移。在个性化推荐应用中,信息的时效性至关重要,一般来说,用户近期访问过的物品对于推荐用户未来可能感兴趣的物品有比较重要的作用,而早期访问记录的影响则相对较小。引入了遗忘函数 $h(t)$,作用是增加最近访问记录的重要性,同时降低时间较长的访问记录的重要性。

首先计算用户对物品的第 i 个行为向量中,第 j 项行为的时间遗忘函数 $h(\Delta t_{ij})$:

$$h(\Delta t_{ij}) = (1 - \theta) + \theta \times \frac{\Delta t_{ij}}{L}, 0 \leq \theta \leq 1 \tag{2}$$

其中, L 为时间窗口, Δt_{ij} 为行为发生时间与时间窗口最开始时间(即当前时间减去时间窗口)之差, θ 初始值设为 0.5。

设有用户对物品的 m 个行为数据向量、 n 项行为(列),形成用户行为矩阵 $X = (x_{ij})_{m \times n}$,计算经时间遗忘函数修正过的用户行为数据 x'_j :

$$x'_{ij} = h(\Delta t_{ij}) \times x_{ij}, i = 1, 2, \dots, m, j = 1, 2, \dots, n \tag{3}$$

通过以上处理可解决用户兴趣漂移问题。

2.3 用户兴趣分析模块

该模块分析用户操作行为,结合用户行为配置文件和熵值法调整模型参数,灵活适应不同的业务类型,主要包括用户行为权重计算和用户对物品的兴趣度计算两个步骤。

(1)基于熵自动计算用户行为权重

用户对物品的兴趣度由用户对物品的各种操作行为确定,不同行为的重要程度不同,一般很难确定什么行为更加重要,常用方法是人工根据经验赋权,缺陷是人工赋权存在主观性和局限性,难以灵活适应多种业务。本文利用信息熵理论和用户行为数据,自动计算各项行为的权重系数。

在信息论中,信息熵 $H(x) = - \sum_{i=1}^n p(x_i) \ln p(x_i)$ 是系统无序程度的度量,而信息是系统有序程度的度量,二者的绝对值相等,符号相反。用户行为的差异性越大,对用户兴趣度分析的作用就越大,权重也越大。而信息熵是信息无序



程度的度量,用户行为的信息熵越小,行为取值的无序程度越低,即行为取值的差异性越大,则该行为对用户兴趣度分析的作用越大,权重就越大;反之则越小。由此可知,分析用户兴趣度的用户行为权重与用户行为熵值成反相关关系,可以根据各项行为值的差异程度,利用信息熵计算各行为的权重,并进一步确定用户偏好兴趣度。为了简化系统的计算量并降低系统复杂度,这里基于所有用户的行为熵值来确定行为的权重。

设有用户对物品的 m 个行为数据向量、 n 项行为(列),通过“用户行为记录预处理模块”进行规范化处理和零值处理后,形成用户行为矩阵 $X=(x'_{ij})_{m \times n}$,对各项行为分值进行同度量化处理,标记为:

$$p_{ij} = \frac{x'_{ij}}{\sum_{i=1}^m x'_{ij}} \quad (4)$$

然后计算第 j 项行为的熵值 e_j :

$$e_j = -k \sum_{i=1}^m p_{ij} \ln p_{ij} \quad (5)$$

其中, $k > 0, e_j \geq 0$ 。如果 x_{ij} 对于给定的 j 全部相等,那么:

$$p_{ij} = \frac{x'_{ij}}{\sum_{i=1}^m x'_{ij}} = \frac{1}{m} \quad (6)$$

此时 e_j 取极大值,即 $e_j = k \ln m$ 。若设 $k = 1/\ln m$,则有 $0 \leq e_j \leq 1$ 。

根据用户行为权重与用户行为熵值成反相关关系的原理,令:

$$g_j = 1 - e_j \quad (7)$$

再对向量 g 进行归一化处理,即可得到行为权重向量 $W=(w_1, w_2, \dots, w_n)$ 。

(2) 计算用户对物品 i 的兴趣度

用户对物品 i 的兴趣度为:

$$prefer_interest(i) = \sum_{j=1}^n w_j x'_{ij} \quad (8)$$

由此得到用户物品兴趣度记录表,包括用户 ID、内容 ID、兴趣度、计算时间。

2.4 个性化推荐模块与参数调优

这个模块采用业界最主流的推荐技术,即协同过滤推荐技术。协同过滤的主要特点是:利用海量用户的群体智慧采用关联关系分析来确定推荐列表,具有推荐新颖性,可以发现使用者潜在的但自己尚未发现的兴趣偏好。

目前主要有基于用户的协同过滤(UserCF)和基于物品的协同过滤(ItemCF)。ItemCF 通过用户对不同物品的评分来评测物品之间的相似性,基于物品之间的相似性做出推荐,适用于物品种类比较稳定的情况;UserCF 则是基于用户之间的相似性做出推荐,主要适用于物品种类变化较快、用户规模相对稳定的领域。一般情况下,物品的相似度要比用户的相似度稳定,同时 ItemCF 可以分为最近邻计算和产生推荐两个阶段,物品的最近邻计算可离线完成,这样在用户评分/兴趣度矩阵非常大的情况下也能做到实时计算推荐,因此,用户规模大的平台一般采用 ItemCF 推荐算法,如 Amazon 等。

在 ItemCF 的最近邻计算阶段,主要是确定相似度排名在前 K 名的物品,如图 3 所示, K 值是 ItemCF 非常重要的参数。

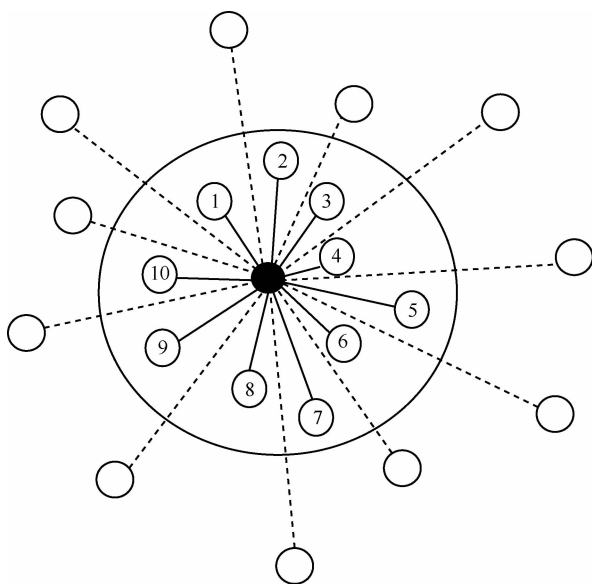


图3 最近邻计算

为了评测不同 K 值参数对推荐模型的效果,采用 TopN 预测准确率、召回率和覆盖率。准确率即系统推荐的物品中,用户喜欢的物品所占的比例;召回率即一个用户喜欢的物品被推荐的概率;为了评测系统发掘长尾物品的能力,将覆盖率也作为一个评测指标,覆盖率表示能够为用户推荐的商品占有所有商品的比率。

以中国电信某游戏平台 17 725 名游戏用户的历史数据为例进行参数调优试验,不同 K 值参数的评估指标如图4所示。

对于该平台数据, $K=10$ 时,预测准确率为 14.8%、召回率为 17.5%、覆盖率为 62.7%,在略微牺牲覆盖率的条件,整体推荐效果最优。

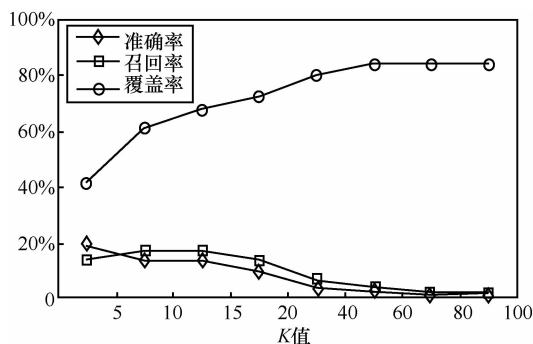


图4 不同K值参数的测试结果

3 推荐效果评估

本次评估测试实验中所采用的是中国电信某游戏平台的实际用户数据,从真实数据中选取游戏行为相对较丰富的用户,实验的用户总数为17 725个,用户行为日志的时间段是19天,游戏数量为5 116款。将每个用户对游戏的操作行为随机按7:3比例分为两份,记作A份和B份,将A份作为训练集,B份作为测试集。分别运用本系统推荐算法、热门推荐算法、随机推荐算法得出每个用户的TopN推荐列表,将推荐列表与测试集的结果进行对比,得到的结果如图5所示。

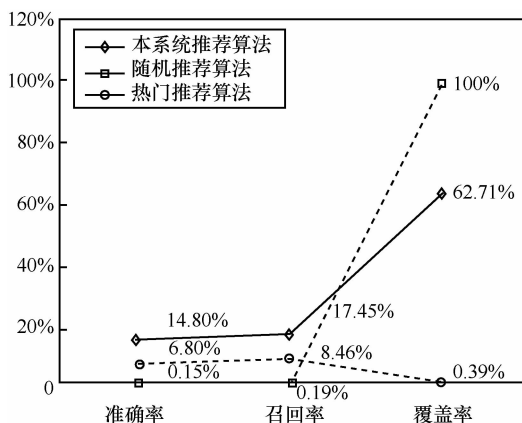


图5 推荐效果对比

由图5可以看出,随机推荐效果最差,虽然覆盖率为100%,但预测准确率只有0.15%,召回率为0.19%;热门推荐效果居中,预测准确率为6.80%,召回率为8.46%,但覆盖率很低,仅为0.39%;本系统推荐的预测准确率为14.8%,召回率为17.5%,覆盖率为62.7%。从对比效果可以看到,本系统推荐的预测准确率远远高于随机推荐,并且是热门推荐的2.18倍,召回率是热门推

荐的将近2倍,覆盖率更是远远高于热门推荐。实验结果表明,本系统的综合推荐效果良好,能很好地向用户推荐用户感兴趣的物品,并具有良好的发掘长尾物品的能力。

4 结束语

本文研究了用户兴趣建模和个性化推荐技术,针对目前很多个性化推荐系统的用户兴趣模型只针对特定领域甚至特定平台构建的情况,提出了一种灵活适应不同业务的个性化推荐系统设计方案,通过分析用户所有操作行为进行用户兴趣建模,结合用户配置文件和熵值法调整用户行为权重系数,以适应不同的业务领域,并采用协同过滤推荐技术,为各类业务提供用户行为分析和推荐能力。此外,提出的用户兴趣自动建模方法,以系统自动建模技术取代人工赋权的方式,提升了个性化推荐中用户兴趣模型的客观性和准确性,并引入时间遗忘函数,可以解决用户兴趣漂移问题,能够动态更新用户兴趣模型,使得推荐结果总能满足用户当前的兴趣需求。实验结果表明,本文提出的个性化推荐系统能够适应不同的业务领域,并具有良好的综合推荐效果。

参考文献

- 1 王国霞,刘贺平. 个性化推荐系统综述. 计算机工程与应用, 2012,48(7):66~76
- 2 Koychev I, Schwab I. Adaptation to drifting user's interests. Proceedings of ECML2000/MLnet Workshop on Machine Learning in the New Information Age, Barcelona, Catalonia, Spain, 2000
- 3 李宁,王子磊,吴刚等. 个性化影片推荐系统中用户模型研究. 计算机应用与软件, 2010,27(12):51~54
- 4 马宏伟,张光卫,李鹏. 协同过滤推荐算法综述. 小型微型计算机系统, 2009,30(7):1282~1288
- 5 丁振国,陈静. 基于关联规则的个性化推荐系统. 计算机集成制造系统, 2003,9(10):891~893
- 6 吴兵,叶春明. 基于效用的个性化推荐方法. 计算机工程, 2012,38(4):49~51
- 7 刘平峰,聂规划,陈冬林. 基于知识的电子商务智能推荐系统平台设计. 计算机工程与应用, 2007,43(19):199~201
- 8 Goldberg D, Nichols, Oki B M, et al. Using collaborative filtering to weave an information tapestry. Communications of the ACM, 1992,35(12):61~70
- 9 林霜梅,汪更生,陈弃疾. 个性化推荐系统中的用户建模及特征选择. 计算机工程, 2007(9)

(下转第148页)



到 IPv6 的过渡方案。

主机同时支持 3 个协议栈:TCP/IPv4、TCP/DN/IPv4 和 IPv6。先从 TCP/IPv4 平滑过渡到 TCP/DN/IPv4,再从 TCP/DN/IPv4 平滑过渡到 TCP/IPv6。从 IPv4 平滑过渡到 TCP/DN/IPv4 的方案已在前文详述,从 TCP/DN/IP 过渡到 IPv6 时,DNS 和 DNR 对于 TCP/DN/IPv4 报文按不同 IP 域处理,对 IPv6 报文按同 IP 域处理,即可完全平滑地实现过渡。

8 结束语

概而言之,双网络层方案有这样一些特点:地址空间无限、与 IPv4 网络高度兼容、升级过程非常平滑、投资极小,很好地满足了对下一代互联网技术方案的要求。

特别有趣的是,这个方案对升级的激励是正向的,所有各方都会有升级的主动性;对运营商而言,升级只需要极少的投资,就可以获得无限的地址空间;对 SP 而言,如果只是服务于 IP 域内用户那么就根本不需要升级,如果升级就可以服务于全球用户;对宽带用户而言,如果满足于目前的主动访问域内服务器的模式也不用升级,如果升级就可以双向访问全球网络,而且升级还很可能是完全自动的。

但一个新的协议从提出到成熟,毕竟需要较长的时间来充分发展。本文只是提出了一个简单的框架。笔者希望这个方案能得到大家的认同。更多的工作,还要期待业界共同推进和完成。

参考文献

- 1 RFC1550. IP: Next Generation (IPng) White Paper Solicitation, 1993
- 2 RFC1752. The Recommendation for the IP Next Generation Protocol, 1995
- 3 RFC2373. IP Version 6 Addressing Architecture, 1998
- 4 RFC1034. Domain Names-Concepts and Facilities, 1987
- 5 RFC1035. Domain Names-Implementation and Specification, 1987
- 6 RFC1123. Requirements for Internet Hosts—Application and Support, 1989
- 7 IETF. NAT444(draft-shirasaki-nat444-06), 2012
- 8 IETF. LAFT6: Lightweight Address Family Transition for IPv6 (draft-sun-v6ops-laft6-01), 2011
- 9 IPv6 发展为何如此纠结. 人民邮电报, 2013-10-10
- 10 姜明. IPv6 协议产生的背景过程和现状. http://tech.ccidnet.com/art/218/20030605/48972_1.html
- 11 Marsan D C. What happens if IPv6 fails. Network world, 2011

[作者简介]

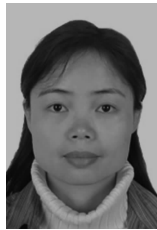


李实,男,中国电信股份有限公司上海研究院工程师、研究员,主要研究方向为基础网络。

(收稿日期:2014-07-02)

(上接第 135 页)

[作者简介]



陶彩霞,女,中国电信股份有限公司广东研究院高级工程师,主要研究方向为大数据、移动互联网、云计算。



陈康,男,中国电信股份有限公司广东研究院高级工程师,主要研究方向为大数据、云计算、电信支撑系统、移动互联网。



袁海,男,中国电信股份有限公司江苏分公司高级工程师,主要研究方向为互联网应用与平台开发运营。



马安华,男,中国电信股份有限公司江苏分公司工程师,主要研究方向为用户行为分析、数据挖掘。

(收稿日期:2014-03-12)