doi:10.3969/j.issn.1001-893x.2014.02.015
引用格式:褚衍杰,徐正国. 基于行为规律的搜索资源分配新算法[J]. 电讯技术,2014,54(2):195-200. [CHU Yan-jie,XU Zheng-guo. A New Algorithm for Allocation of Search Resources Based on Behavior Rule[J]. Telecommunication Engineering,2014,54(2):195-200.]

基于行为规律的搜索资源分配新算法*

褚衍杰**,徐正国

(盲信号处理重点实验室,成都 610041)

摘 要:针对具有行为规律的目标搜索问题,提出一种搜索资源分配算法。该方法以目标在各搜索区域的概率分布为基础,利用最优搜索理论分配区域搜索时长;以目标在各搜索区域的行为规律为基础,利用包络检测等方法决定区域开始搜索的时刻。针对网站关键词搜索的实验显示,根据目标行为规律在时间上相关程度的不同,本算法相对于最优搜索算法的性能提升在 15% ~50% 之间,在对大量信息源进行信息搜索时具有应用价值。

关键词:资源分配;最优搜索;行为规律;目标搜索
中图分类号:TP393 **文献标志码**:A **文章编号**:1001-893X(2014)02-0195-06

A New Algorithm for Allocation of Search Resources Based on Behavior Rule

CHU Yan-jie, XU Zheng-guo

(Key Laboratory of Science and Technology on Blind Signal Processing, Chengdu 610041, China)

Abstract: This paper proposes a new algorithm for search resources allocation to search targets with behavior rule. In the new algorithm, the optimal search theory is used to allocate the search periods for search zones based on the probability distribution of the targets, and a target-behavior based envelope detection algorithm is developed to decide the time instants when the search begins. The experiment results of keywords searching on network sites indicate that the proposed algorithm has much better performance than the optimal search method, and the performance gain is between 15% and 50% for different correlations of behavior rules. The algorithm will find application in searching mass information resources.

Key words: resources allocation; optimal search; behavior rule; target search

1 引 言

在第二次世界大战期间,由于战争中快速搜索对方运动目标的需要,George Kimball、Bernard Koopman 等人创立了最优搜索理论,并逐渐在犯罪学、矿藏勘探、市场调查、网络信息处理等领域得到了深入研究[1]。

最优搜索理论是关于如何以一种“最佳”的方式寻找某个事先已确定的对象(搜索目标)的理论。最优搜索问题的 3 个基本要素即目标位置概率分布函数、探测函数、代价函数从模型的角度来讲分别对应了目标初始概率密度、目标探测模型以及搜索资源模型。利用这些模型可以针对静止或运动目标的搜索问题展开研究,例如文献[2]研究了一维随机恒速运动目标的搜索问题,文献[3-4]将静止目标的搜索方法应用到网络信息处理领域,分别研究特

* 收稿日期:2013-11-19;修回日期:2014-01-17 Received date:2013-11-19;Revised date:2014-01-17
** 通讯作者:chuyanjie@mail.tsinghua.org.cn **Corresponding author**:chuyanjie@mail.tsinghua.org.cn

定信息搜索和入侵检测问题。

网络用户的浏览行为、通信行为、入侵行为等都受到作息时间、个人行为习惯、入侵方法等的限制，从而在网络数据上反应出一定的规律，例如文献[5]对用户信息查找行为的规律进行了分析，文献[6-7]将行为规律应用到异常检测和入侵检测领域，而以最优搜索为代表的众多优化搜索策略的方法，利用了目标的概率分布，但未对目标的行为规律进行进一步的探讨。

本文从网络数据中分析目标出现的规律，建立其行为规律模型，结合最优搜索理论提出基于行为规律的搜索资源分配算法，解决搜索多长时间和从何时开始搜索的问题，并通过网站关键词的实验证明了算法能够有效提高目标搜索的效率。

2 算法模型

搜索资源主要是指搜索的时间资源，搜索资源分配策略是指根据目标在各搜索区域出现的经验记录，制定出时间资源分配策略，具体包括在各搜索区域的搜索时长和开始搜索的时刻。

新算法的创新在于建立目标行为规律描述模型，并将其引入搜索资源分配策略中，其中目标行为规律是指目标在同一个区域不同时间出现规律的统计。算法结构如图 1 所示，首先统计目标概率分布，结合最优搜索理论制定各区域搜索时长的分配策略，其中目标概率分布是指目标在各搜索区域出现次数的概率统计；同时统计各区域内目标的行为规律，以获取更多信息为目标检测区域的最佳搜索时刻；最后结合区域搜索时长和最佳搜索时刻，制定目标搜索的时间资源分配策略。

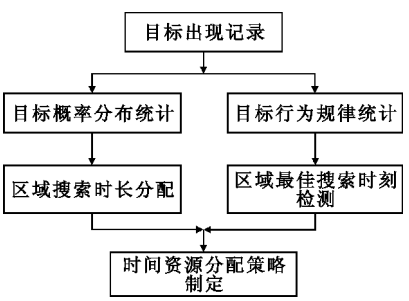


图 1 资源分配算法结构

Fig. 1 Structure of the algorithm

算法用到的符号如下：待搜索区域为 $S=\{S_1,S_2,\cdots,S_N\}$；目标在各区域的出现记录为 $G=\{g_1,g_2,\cdots,g_M\}$，其中 $g_i=\{S_i,T_i\}$，$1\leqslant i\leqslant M$，$S_i\in S$ 为第 i 条记录的出现区域，T_i 为第 i 条记录的出现时刻；算法的结果为 $R=\{r_1,r_2,\cdots,r_N\}$，其中 $r_i=\{S_i,TL_i,TS_i\}$，$1\leqslant i\leqslant N$，表示了从时间 TS_i 开始搜索区域 S_i，共搜索 TL_i 长的时间。

2.1 区域搜索时长分配

按照经典的最优搜索理论，区域搜索时长分配主要包含以下几个步骤。

(1)目标位置概率分布估计

由于目标的行为规律一般以天为单位，因此如果目标出现记录中包含多天的结果，可以进行加权融合。利用 $G=\{g_1,g_2,\cdots g_M\}$ 可以统计第 i 个区域第 j 天目标出现的总次数 C_{ij}，定义规律有效性衰减因子 $\beta=[\beta_1,\beta_2,\cdots,\beta_D]$，其中 $\beta_k(1\leqslant k\leqslant D)$ 表示第 k 天的规律的加权系数，D 为规律统计的总天数，则第 i 个区域的加权次数为 $C'_i=\sum_{j=1}^{D}C_{ij}\beta_j$，$1\leqslant i\leqslant N$，于是可以得到目标在各区域的概率分布估计 $P=\{p_1,p_2,\cdots,p_N\}$，其中目标在第 i 个区域的概率为

$$p_i=\frac{C'_i}{\sum_{j=1}^{N}C'_j},1\leqslant i\leqslant N \tag{1}$$

(2)目标探测函数

用探测函数 $b(i,t)$ 表示在区域 S_i 上，花费时间 t 能成功搜索到目标的概率。根据一般经验，若不考虑目标的概率分布，投入到区域 S_i 上的时间 t 越长，最后能成功找到目标的概率越大，因此可以将探测函数设为如下形式：

$$b(i,t)=1-e^{-\frac{t}{\tau}} \tag{2}$$

其中，$1\leqslant i\leqslant N$，$t>0$，τ 为经验常数。

(3)时长分配策略

设总的时间资源为 T，根据目标的概率分布及探测函数，可以得到时间 T 内发现目标的概率为

$$P[f]=\sum_{i=1}^{N}p_ib(i,f(i)) \tag{3}$$

其中，$f(i)$ 为分配给区域 S_i 的搜索时长，且 $\sum_{i=1}^{N}f(i)=T$。

代价函数 $c(i,t)$ 表示把时间 t 分配给区域 S_i 的代价，可以简单地令 $c(i,t)=t$，则策略 f 的总代价为

$$C[f]=\sum_{i=1}^{N}c(i,f(i))=\sum_{i=1}^{N}f(i)=T$$

于是问题转化为有约束的最优化问题：

$$\max P[f]$$

$$\text{s.t.}\ \sum_{i=1}^{N} f(i) \leqslant T$$
$$f(i) \geqslant 0, 1 \leqslant i \leqslant N \tag{4}$$

求解该问题有两个方法:第一种方法是首先忽略约束条件 $f(i)\geqslant 0, 1\leqslant i\leqslant N$,利用拉格朗日乘数法求得解析解:

$$\lambda = \mathrm{e}^{\frac{\left(\sum_{j=1}^{N}\tau\ln\left(\frac{p(j)}{\tau}\right)-T\right)}{N\tau}}$$
$$f_{\lambda}^{*}(i) = \tau\ln\left(\frac{p(i)}{\lambda\tau}\right), 1\leqslant i\leqslant N \tag{5}$$

然后在解析解的基础上进行结果调整,使得 $f(i)\geqslant 0, 1\leqslant i\leqslant N$;第二种方法是将该问题转化为有约束最小化问题:

$$\min\ -P[f]$$
$$\text{s.t.}\ \sum_{i=1}^{N} f(i) - T \leqslant 0$$
$$-f(i) \leqslant 0, 1 \leqslant i \leqslant N \tag{6}$$

则可以直接利用 Matlab 的 fmincon 函数求取数值解。

2.2　最佳搜索时刻检测

目标的活动一般具有规律性,例如以网络活动为例,朝九晚五的上班族一般在上午 10 点前后会集中处理邮件等日常工作,利用此类规律决定目标搜索的最佳时刻,可以提高搜索到目标的概率。

为了检测最佳搜索时刻,首先建立对目标行为规律的描述:对于区域 S_i,以时间分布向量 $[n'_{t1}, n'_{t2}, \cdots, n'_{tH}]$ 表示目标的行为规律,其中 $n'_{ti} = \sum_{j=1}^{D} n_{ti}^{j}\beta_j, 1\leqslant i\leqslant H$ 是多天统计结果按衰减系数的加权,n_{ti}^{j} 表示第 j 天目标在 $(t_{i-1}, t_i]$ 时间段内在该区域出现的次数。

按照上述方法建立每个区域目标行为的时间分布,该分布可能呈现一定的规律性,比如多峰,如图 2 所示的为双峰现象。

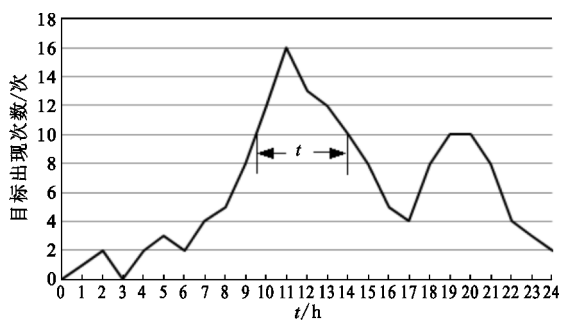


图 2　双峰现象及最佳搜索时刻检测示意图
Fig. 2 The sketch map for double-peak and detection of the optimal search moment

本文主要利用时间分布的多峰现象,将每个区域的搜索时间 $f(i)$ 分配到各个峰值位置。最佳搜索时刻的最优目标是图 2 中搜索时间段内曲线覆盖的面积最大,但是通过分析实际数据发现,对应每个区域的行为规律,图 2 中曲线的形状差异非常大,优化复杂,因此选取了一种相对较优的简化方法,便于在实际工程中使用。以双峰规律为例说明分配方法如下:

(1)利用包络检测的方法,检测最高峰 $[t_1, n_1]$ 和次高峰 $[t_2, n_2]$,其中 t_1、t_2 表示位置,n_1、n_2 表示峰值;

(2)检测 t_1 两侧值为 n_2 的位置,记录时间差 t;

(3)若 $f(i)\leqslant t$,则最佳接入位置为 t_1,时长为 $f(i)$,无次佳搜索时刻;否则最佳搜索时刻为 t_1,时长为 $\frac{f(i)+t}{2}$,次佳搜索时刻为 t_2,时长为 $\frac{f(i)-t}{2}$。

2.3　时间资源分配策略制定

对于双峰规律,得到每个区域的搜索时长和最佳/次佳搜索时刻后,问题转化为根据上述信息在时间轴上完成对搜索时长的分配。本算法采用根据各区域目标出现次数的排序,依次按照最佳/次佳搜索时刻分配区域搜索时刻,流程如图 3 所示。

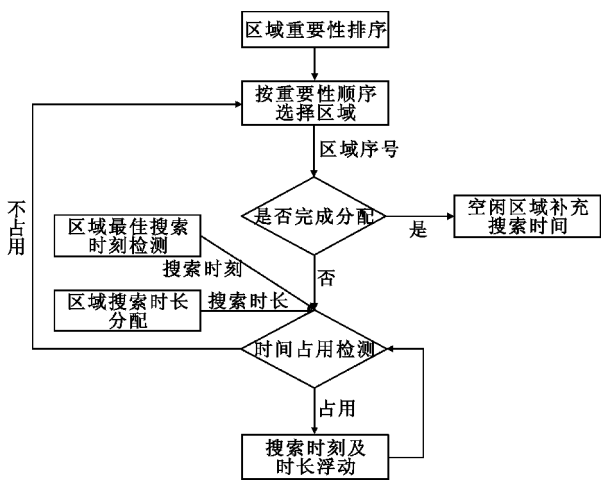


图 3　时间资源分配策略制定流程图
Fig. 3 The flow for time distributing strategy

对于流程的具体说明如下:

(1)统计目标的概率分布,并计算各区域的搜索时长;

(2)对于所有区域,统计每个区域 D 天的时间分布向量,并加权融合,得到每个区域的时间分布向量;

(3)利用包络检测方法检测时间规律是否具有

双峰现象;

(4)对于有双峰现象的区域,完成两段搜索时长的分配,并记录两个峰值为最佳/次佳搜索时刻;

(5)对于没有双峰现象的区域,检测单峰位置,记录为其最佳搜索时刻;

(6)对所有区域,按照目标出现次数进行重要性排序;

(7)按照排序结果,对所有区域,以其最佳/次佳搜索时刻为中心,按照预先分配的时间长度,检测该时间区间是否被占用,若未被占用,确认搜索时刻及时长并更新时间占用情况,否则在最佳/次佳搜索时刻一定范围内移动中心,检测时间是否被占用;若未被占用,确认搜索时刻及时长并更新时间占用情况,否则缩短搜索时长,以最佳/次佳搜索时刻为中心搜索,直至找到合适的搜索时刻;

(8)完成所有区域的初次分配后,检测时间占用情况,对于未被占用的时间段,若时长大于一定门限,根据在步骤 7 中缩短时长的区域的最佳搜索时刻,选取距离最近的区域填补空白,直至无法填补。

3 试验验证及结果分析

为了验证算法的性能,分两次采集了 254 个网站(数据集 1)以及1 219个网站(数据集 2) 9 天的数据,并按照 $g_i=\{S_i,T_i\}$ $(1\leqslant i\leqslant M)$ 的方式记录关键词的每次出现,以此作为试验验证的原始数据。验证过程模拟目标搜索过程,即假设目标为关键词,且关键词按照 $g_i=\{S_i,T_i\}$ $(1\leqslant i\leqslant M)$ 的方式在网站上出现,搜索目标时仅能搜索到网站实时出现的关键词,不能搜索到已经存在的关键词。验证方法是利用前 5 天的数据训练目标的概率分布和行为规律,然后利用后 4 天的数据分别按照制定的策略进行统计,得到每天可以搜索到的关键词次数,并对比不同算法搜索到关键词数量的性能;本算法主要分析一天内的行为规律,因此策略中总的搜索时间资源为 1 天。

3.1 行为规律分析

图 4 分别给出数据集 1 和数据集 2 中行为规律时间相关性最强的 3 个网站的相关性变化曲线,其中横坐标代表间隔时间(天),纵坐标代表相关系数,每个点表示当天数据的时间分布向量与第一天数据的时间分布向量之间的相关系数。

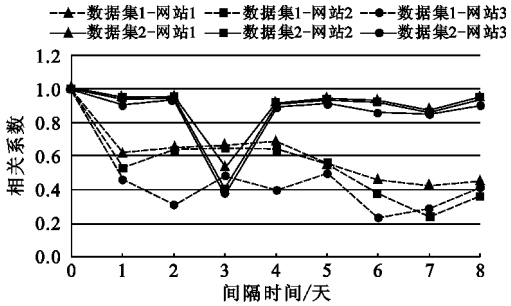


图 4 行为规律相关性变化曲线

Fig. 4 Relativity of behavior rule

由图 4 可以看出,数据集 1 中各网站的相关系数随着时间间隔的变大呈现变小的趋势,而数据集 2 中网站的相关系数保持稳定,没有明显的变小趋势,而且相关性比数据集 1 中更大。这种现象说明,不同的网站上关键词的行为规律随着时间延续在不断变化,行为规律有一定的时效性;也说明每个网站上不同日期的行为规律具有相关性,虽然相关程度不同,但可以利用前期的数据预测后期数据的行为规律,证明了本算法在原理上是可行的。

3.2 算法有效性对比

表 1 给出了利用数据集 1 和数据集 2 进行验证的结果,其中衰减指数均选择指数衰减。为了体现算法效果,将结果以平均分配方法(各区域分配相同的时长,顺序搜索)为基准进行了归一化,并给出了本算法以及最优搜索方法(只使用最优搜索不利用行为规律)的性能提升对比,性能提升表示相应算法获取的关键词数量与平均方法获取的关键词数量的比值;图 5 给出了不同算法和数据集的性能提升对比,其中横坐标代表不同日期,纵坐标代表性能提升倍数。

表 1 算法性能提升对比表

Table 1 Comparison of performance advance

数据集	日期/天	本文算法	最优搜索	平均方法
数据集 1	1	3.95	3.14	1.00
	2	5.22	3.89	1.00
	3	5.55	4.61	1.00
	4	4.37	2.87	1.00
数据集 2	1	29.09	24.82	1.00
	2	35.65	30.26	1.00
	3	28.31	22.75	1.00
	4	28.30	24.58	1.00

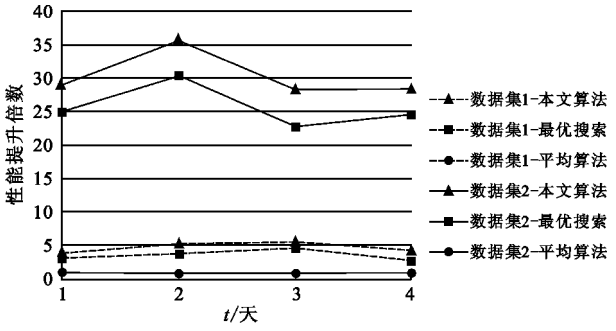


图 5　性能提升对比图

Fig. 5 Comparison of performance advance

从表 1 和图 5 可以看出,对于数据集 1,最优搜索方法的性能是平均算法的 3 ~5 倍,本文算法性能是平均算法的 4 ~6 倍,本文算法比最优搜索方法性能提升 20% ~50%;对于数据集 2,最优搜索方法的性能是平均算法的 20 ~30 倍,本文算法性能是平均算法的 28 ~35 倍,本文算法比最优搜索方法性能提升 15% ~25%;对照图 4 可以看出,由于数据集 2 的行为规律在时间上的相关性更强,因此相对平均算法的性能提升效果更明显。

3.3　训练数据对结果的影响

图 6 给出了对于数据集 1 和数据集 2,利用 5 天的数据进行概率分布统计和行为规律统计,且采用不同的衰减因子 β 时,算法相对于平均分配算法性能提升倍数的结果对比,其中横坐标代表不同日期,纵坐标代表性能提升倍数。无衰减的衰减因子为 β =[1,1,1,1,1],指数衰减因子为 β =[0.018 2, 0.049 8,0.135 3,0.367 9,1],线性衰减因子为 β =[0.2,0.4,0.6,0.8,1],一天的衰减因子为 β =[0,0,0,0,1]。

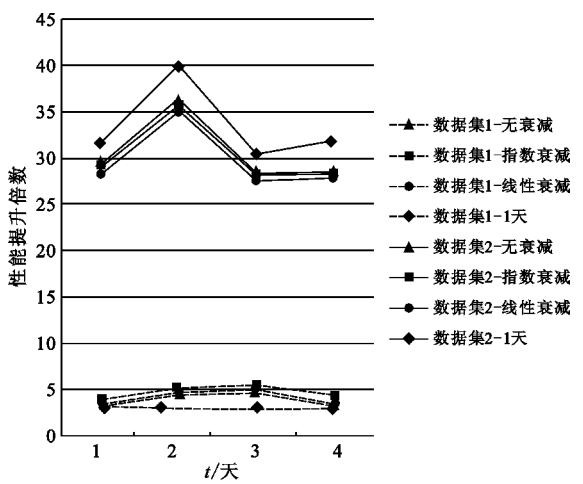


图 6　采用不同衰减因子 β 时性能提升对比图

Fig. 6 Comparison of performance advance with different β

由图 6 可以看出,对于数据集 1,采用指数衰减因子时算法性能提升最大,说明了不同日期行为规律相关性随时间降低时,应该采取快速衰减的 β,保证行为规律的及时性,而使用 1 天训练数据的性能提升最差,则是因为由 1 天数据得到的行为规律具有一定的不稳定性;对于数据集 2,使用 1 天训练数据时性能提升反而要好些,这是由于数据集 2 的 5 天训练数据中,行为规律非常稳定,只有第 3 天的数据偏差较大,因此使用 5 天训练数据相当于引入了一个噪声,效果反而变差。通过分析可以看出,在算法使用过程中,需要兼顾行为规律的稳定性和及时性,根据行为规律相关性的变化规律采用合理的衰减因子 β:时间规律衰减快则选择衰减快的 β,时间规律稳定则 β 的选择对性能影响较小,因此一般情况下,建议选择衰减较快的指数衰减因子。

4　结束语

本文通过分析目标的行为规律,结合最优搜索理论,提出了一种基于行为规律的搜索资源分配算法。利用网站上关键词出现记录进行的模拟搜索试验证明了目标行为规律的存在以及其在稳定性、时效性等方面的特点,实验结果显示,本文的方法相对于平均分配资源能够大幅度提高搜索效率,相对于单独使用最优搜索方法,搜索效率也有显著提高,在对大量信息源进行信息搜索时具有应用价值。另外,在实时处理条件下如何获取各搜索区域完整的历史信息问题未在文中涉及,将是下一步研究的重点。

参考文献:

[1] 朱清新. 离散和连续空间的最优搜索理论[M]. 北京:科学出版社,2005.
ZHU Qing-xin. The optimal search theory in discrete and continuous space [M]. Beijing: Science Press, 2005. (in Chinese)

[2] 陈建勇,王健. 对随机运动目标的一种最优搜索算法[J]. 海军航空工程学院学报,2012,27(4):456-458.
CHEN Jian-yong, WANG Jian. An optimal search algorithm for randomly moving target[J]. Journal of Naval Aeronautical Engineering Institute, 2012, 27(4): 456-458. (in Chinese)

[3] Chu Yanjie, Wei Qiang. A network specific information search system based on mobile agent[C]//Proceedings of 2012 Third Global Congress on Intelligent Systems. Wu-

han:IEEE, 2012:302-304.

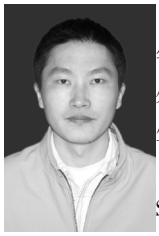
[4] 盛志伟,朱清新. 最优搜索理论在入侵检测系统中的应用研究[J]. 计算机应用与软件,2008,25(5):248-250.
SHENG Zhi-wei, ZHU Qing-xin. On appling optimal search theory in IDS[J]. Computer Applications and Software,2008,25(5):248-250. (in Chinese)

[5] 何惠芳. 网络环境下用户信息查找行为规律的实证分析[J]. 情报探索,2008(4):6-8.
HE Hui-fang. The analysis of behavior rule of information searching in network[J]. Information Research, 2008(4):6-8. (in Chinese)

[6] 苗强,周兴社. 基于行为规律的异常检测技术研究[J]. 计算机工程与应用,2010,46(15):211-214.
MIAO Qiang, ZHOU Xing-she. Research of outlier detection technique based on behavior rule[J]. Computer Engineering and Applications, 2010,46(15):211-214. (in Chinese)

[7] Mitchell R, ChenI R. Behavior rule based intrusion detection for supporting secure medical cyber physical systems[C]//Proceedings of 2012 International Conference on Computer Communications and Networks. Munich, Germany:IEEE,2012:1-7.

作者简介:



褚衍杰(1982—),男,山东枣庄人,2005 年于清华大学获学士学位,2008 年于盲信号处理重点实验室获硕士学位,现为博士研究生,主要研究方向为信息处理;

CHU Yan-jie was born in Zaozhuang, Shandong Province, in 1982. He received the B. S. degree from Tsinghua University and the M. S. degree from Key Laboratory of Science and Technology on Blind Signal Processing in 2005 and 2008, respectively. He is currently working toward the Ph. D. degree. His research concerns information processing.

Email:chuyanjie@ mail. tsinghua. org. cn

徐正国(1985—),男,湖北荆州人,2008 年于北京理工大学获学士学位,2011 年于盲信号处理重点实验室获硕士学位,现为博士研究生,主要研究方向为信息处理。

XU Zheng-guo was born in Jingzhou, Hubei Province, in 1985. He received the B. S. degree from Beijing Institute of Technology and the M. S. degree from Key Laboratory of Science and Technology on Blind Signal Processing in 2008 and 2011, respectively. He is currently working toward the Ph. D. degree. His research concerns information processing.