基于段落-句子互增强的自动文摘算法

谢 浩[1] 孙 伟[2]
(中山大学信息科学与技术学院 广州 510006)[1] (中山大学软件学院 广州 510006)[2]

摘 要 句子排序问题是文本自动摘要的核心问题,基于互增强关系(MRP)的基本思想,提出一种新的句子排序模型——段落-句子互增强模型。利用段落关系,通过段落句子的互增强,迭代计算出句子的显著度,抽取出文摘句。分析了模型中的内、外影响因子对算法效果的影响并对冗余处理进行了讨论。实验表明,将其运用在单文本自动摘要中,能取得高质量的文摘。

关键词 句子排序,互增强关系,自动文摘
中图法分类号 TP393 **文献标识码** A

Paragraph-Sentence Mutual Reinforcement Based Automatic Summarization Algorithm

XIE Hao[1] SUN Wei[2]
(School of Information Science and Technology, Sun Yat-sen University, Guangzhou 510006, China)[1]
(School of Software, Sun Yat-sen University, Guangzhou 510006, China)[2]

Abstract Sentence ranking is the key issue of text automatic summarization. Based on mutual reinforcement principles, we proposed a new sentence ranking model——paragraph-sentence mutual reinforcement model. With the relation between paragraphs and the mutual reinforcement between paragraphs and sentences, it iteratively computes the salience of the sentences and extract the summary sentences. We analyzed the effect of the internal and external reinforced factor and discussed the problem of redundancy remove. Experiments show that it can extract high quality summary when it applies to the single-document summarization.

Keywords Sentence ranking, Mutual reinforcement principle, Automatic summarization

1 引言

自动文摘系统,能让人们从海量的文档中迅速获取有用的信息,在这个大数据时代背景下其有很大的实用价值。它按实现方法主要分为两种:基于理解的自动文摘(Abstraction Based Summarization)和基于信息提取的自动文摘(Extraction Based Summarization)。前者要进行语法分析、语义分析,涉及到自然语言处理的技术,实现起来比较困难。后者是主流的自动文摘实现方法,它直接从文档中提取关键的句子,并用它们去表示文档摘要。按提取对象分主要有单文档摘要(Single-Document Summarization)和多文档摘要(Multi-Document Summarization)。前者是从一个文档里面提取摘要,后者是从一个文档集里面提取整个文档集的摘要。

基于信息提取的自动文摘的关键步骤就是提取文档(文档集)的关键句子,所以自动文摘的核心问题就转化为如何对句子进行排序。基于图的排序算法在网页搜索中有很成功的应用,其代表是 PageRank[1] 和 HITS[2]。基于图的句子排序算法将文档的句子表示成图中的节点,边的权值表示为句子间的相似度,迭代计算出每个句子的显著度(salience),抽取出关键句组成摘要。LexRank[3]、TextRank[4]、GRASSHOPPER[5]等算法根据 PageRank 的思想,认为一个句子与越多其他句子相似,该句子就越重要。MRP[6](Mutual Reinforcement Principle)借鉴 HITS,认为越重要的词会出现在许多越重要的句子中,而越重要的句子会包括越多越重要的词。

原始的 MRP 主要是针对单文本摘要,而在文献[7]中,提出了文档-句子互增强模型,在文献[8]中,提出了文档-句子-词语三层的互增强模型。这两种方法都充分利用了文档间关系信息,在多文档摘要中取得了不错的效果。

但在单文档摘要中,我们无法利用文档间关系信息,在文献[6,8]这两种互增强算法中,利用了词语间的关系,但是其通常要计算词语间的相似度,这又往往需要借助文档之外的信息,缺乏一个统一、高效的方法。例如,文献[9-11]利用搜索引擎为每个词建立上下文向量(context vector),它依赖于搜索引擎的有效性,文献[8]使用了 WordNet[14],即一种词语的语义网络,但是该语义网络目前还在完善当中,而且对中文还没有有效的支持。

我们在研究中发现,不利用词语间关系,我们还是可以设计出一个基于互增强的单文本摘要算法,因为我们可以利用段落间关系,建立一个段落-句子互增强模型。

本文第2节描述算法的思想和具体框架、相似度的计算、文摘句的选取和去冗余句问题;第3节为实验结果和对参数的选取进行讨论;最后做出总结和对未来工作的展望。

谢 浩(1988—),男,硕士生,主要研究方向为文本可视化分析,E-mail:xiehao1988@hotmail.com;孙 伟(1972—),男,博士,教授,主要研究方向为计算机图形学、多媒体安全。

2 算法框架

在一篇文章当中，不同段落有不同的作用，其重要性也不一样，开头和结尾段往往更加重要，它更可能包括文章的中心思想，它里面的句子很可能是文摘句。过渡段落、辅助说明段落在文章中的重要性会较低，因此它里面的句子就不太可能是文摘句。基于这样的观察，我们提出段落-句子互增强模型，其思想可以概括为：

1. 越重要的句子会出现在越多越重要的段落里，且与很多其他句子相似；

2. 越重要的段落会包括越多越重要的句子，且与很多其他段落相似。

2.1 段落-句子互增强框架

在 MPR 模型下，包括了两个带权无向图和由这两个图组成的带权二部图。本文提出的段落-句子互增强模型包括了一个段落图、句子图，还有这两个图组成的一个二部图，如图 1 所示。段落图或句子图的节点为表示段落或句子的向量，边的权值为节点之间的相似度。因此这两种图都可表示成相似度矩阵，关于相似度矩阵的计算将在 2.2 节讨论。

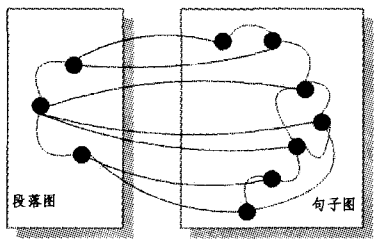

图 1 段落-句子互增强

我们称在同一层次图上的增强为内增强，层次之间的增强为外增强。两种增强的计算同时进行，并互相影响。最终可以同时迭代计算出句子的显著度和段落显著度。该迭代计算方程可以定义为：

$$\begin{cases} R_P^{(k+1)} = \alpha_1 \cdot P_P \cdot R_P^{(k)} + \beta_1 \cdot P_S \cdot R_S^{(k)} \\ R_S^{(k+1)} = \beta_2 \cdot S_P \cdot R_P^{(k)} + \alpha_2 \cdot S_S \cdot R_S^{(k)} \end{cases} \tag{1}$$

式中，R_P、R_S 表示段落集 P 的显著度和句子集 S 的显著度，P_P、S_S 表示段落相似度矩阵和句子相似度矩阵，P_S、S_P 表示段落-句子相似度矩阵，易知 $P_S=(S_P)^T$。$\Phi=\begin{bmatrix}\alpha_1 & \beta_1 \\ \beta_2 & \alpha_2\end{bmatrix}$是影响因子矩阵，用于控制内增强和外增强对显著度的影响程度。称 α_1、α_2 为内影响因子，β_1、β_2 为外影响因子，我们将在第 3 节论讨它们对算法的影响。

方程(1)可以映射为一个分块矩阵，

$$M=\begin{bmatrix}\alpha_1 P_P & \beta_1 P_S \\ \beta_2 S_P & \alpha_2 S_S\end{bmatrix} \tag{2}$$

令 $R=\begin{bmatrix}R_P \\ R_S\end{bmatrix}$，那么 R 即为 M 的特征向量：

$$M \cdot R = \lambda \cdot R \tag{3}$$

为了让 R 能最终收敛，根据文献[1]，要保证矩阵 M 是随机的和不可约的。因此，在迭代计算开始之前，我们要对矩阵 M 做变换。变换分为两步：

1. 随机化：首先，去除 P_P 和 S_S 中的全零行，即图中的孤立节点。然后，按列分别归一化 P_P、P_S、S_P 和 S_S 这 4 个矩阵。产生的一个新的随机化的矩阵，用 M' 表示。

2. 不可约化：设 X 表示 M' 中对角线上的矩阵。根据 PageRank 的思想，让 X 图中的节点，以某一概率与图中的其他节点连接起来，最简单的方法是设该概率向量为 $\vec{p}$，$\vec{p}_i=[1/n]$，n 为图 X 中的节点个数。这样，我们有：

$$X' = d \cdot X + (1-d) \cdot E \quad E = \vec{p} \times [1]_{1\times n} \tag{4}$$

d 是 PageRank 中的阻尼因子(damping factor)，本算法取 d 为 0.75。将 X' 代替 M' 中的 X，得到 M''。

经过上面的变换，我们可以证明：

引理 1 如果 Φ 是列随机的，则 M'' 是列随机。

证明：设 A，B 分别表示在 M'' 中的任何列的两个分块矩阵，α 和 β 为对应的系数，则有

$$\sum_i M_{ij}'' = \alpha \sum_i A_{ij} + \beta \sum_i B_{ij} = \alpha + \beta = 1$$

引理 2 M'' 是不可约的。

证明：因为 M'' 对角线上的两个矩阵对应的两个图是强连接的，即这两个矩阵是不可约的，且这两个图之间的连接是双向的，所以矩阵 M'' 所对应的那个图也是强连接的，即矩阵 M'' 是不可约的。

所以，M'' 是一个随机不可约的矩阵。我们可以找到它的一个惟一特征向量，亦即方程式(1)最终将收敛于向量 R。

以上是本文算法的整体框架，下面讨论两个关键问题：如何求相似度矩阵和如何选取文摘句。

2.2 相似度的计算

在上面的框架中，有一个关键的问题就是如何得到段落相似度矩阵、句子相似度矩阵和段落-句子相似度矩阵(P_P、S_S、P_S、S_P)。这里我们采用向量空间模型(Vector Space Model)。

由于段落和句子都属于文档中的文本单元，只是粒度不同，因此我们可以将它们统一表示成由词语元素构成的向量，其维度为文中袋口词(bag-of-word)个数。对于中文文本来说，袋口词要先通过对文本进行分词，然后去停词获得。对于单文本，袋口词的权值计算通常采用类似于 TF * IDF[15] 的 TF * ISF(term frequency inverse sentence frequency)。而我们认为，在本方案中组成段落向量的袋口词的权值使用段落层次的 IPF(inverse paragraph frequency)会更加合理：

$$ipf_i = \log_2(0.5 + \frac{|P|}{|\{p \in P : i \in P\}|}) \tag{5}$$

式中，$|P|$ 是段落总数，$|\{p \in P : i \in P\}|$ 是词 i 出现的段落的总数。

因为段落和句子的向量表示的形式是相同的，所以段落间、句子间和段落句子间的相似度，都可以统一计算为它们之间的余弦值。

设 $\overline{W}=\{W_1, W_2, \cdots, W_i, \cdots, W_n\}$ 为袋口词集合，$o_i=\{w_1, w_2, \cdots, w_i, \cdots, w_n\}$ 表示文本单元向量(段落或句子)。w_i 为袋口词 W_i 的权值。则有：

$$similarity(o_i, o_j) = \frac{\sum_{k=1}^{n} w_{ik} \cdot w_{jk}}{\sqrt{(\sum_{k=1}^{n} w_{ik}^2)(\sum_{k=1}^{n} w_{jk}^2)}} \tag{6}$$

在向量空间模型下计算的相似度，会得到一些过小的值，

导致图中有些边的权值过小，这会对迭代过程产生一定影响。通常我们设定一个相似度阈值，除去图中那些权值过小的边。本方案设相似度阈值为 0.01。这样，我们最终求得相似度矩阵。

2.3 文摘句的选取

在多文本文摘系统中，由于句子数多，文摘中会出现冗余信息，因此去冗余是文摘句选取中不可缺少的一部分。然而，我们在研究中发现，对于单文本文摘，文摘句的选取同样需要进行去冗余这一步，且去冗余阈值 *threshold* 应该随着文本的不同而动态变化。本方案中设 $threshold = avg * \eta$，其中 *avg* 为相似度矩阵中除对角线元素外的元素值的平均，而对于系数 η 的取值，我们会在下一章进行讨论。设 $S' = \{s_1, s_2 \cdots s_i \cdots s_n\}$ 为排序后的句子集，$\overline{S}$ 为文摘句集，初始为空，则文摘句选取的步骤如下：

1. 选取 s_1，令 $\overline{S} = \{s_1\}$，$S' = S' - \{s_1\}$；

2. 依次选取 S' 中的 $s_i (i \geqslant 2)$，若对于 $s_j \in \overline{S}$，$similarity(s_i, s_j) > threthold$，则抛弃 s_i，否则 $\overline{S} = \overline{S} + s_i$；

3. 重复步骤 2 直到选取一定个数的句子。

2.4 算法流程

根据以上分析，段落-句子互增强自动文摘算法步骤可以归纳如下：

输入：文本 text，影响因子矩阵 Φ(先归一化)，去冗余系数 η，文摘句子数 k

输出：文摘句子集 $\overline{S}$，$|\overline{S}| = k$

1. 由 text 得到句子集 S、段落集 P、词集 W；
2. 根据 2.2 节计算得到 P_P、S_S、P_S、S_P；
3. 根据 2.1 节由 Φ 和 P_P、S_S、P_S、S_P 得到分块矩阵 M 阵，对 M 随机化、不可约化得到 M″；
4. 初始化 R^0，使得 $|R^0| = 1$。迭代计算方程(1)直到收敛于 R；
5. 根据 R，对 S 进行排序，得到 S′；
6. 根据 2.3 节由 S_S 算出 avg 并由 η 得到 threthold 进行冗余处理，得到 $\overline{S}$。

3 实验结果与分析

段落-句子互增强算法涉及了几个比较重要的参数，我们分别做了不同的实验，测试这几个参数的影响。

3.1 数据集和评价方法

我们使用的数据集是哈尔滨工业大学提供的《哈工大信息检索研究中心单文档自动文摘语料库》[12]，包括了 6 种文章体裁共 211 篇语料，其中奥运 57 篇，记叙文 40 篇，说明文 40 篇，议论文 46 篇，应用文 18 篇和 03 年 863 评测语料 10 篇。每一篇文章都有 5 个人工摘要，分别按照原文 10% 和 20% 抽取文摘句子。

实验选择的评价方式是 DUC 2005 中使用的 ROUGE 方法[13]。它通过将系统摘要与人工摘要相比较来评价系统摘要的质量，方法是计算系统摘要与人工摘要的重叠词的数量。根据不同的重叠标准，可以分为 ROUGE-N，ROUGE-L，ROUGE-W，在这里我们使用的是 ROUGE-N 的方法，它也是 DUC 中最主要的评价方法。ROUGE-N 计算的是系统摘要和一组人工摘要的 n-gram 的 recall 值，它的公式如下：

$$ROUGE\text{-}N = \frac{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)} \tag{7}$$

式中，n 表示 n-gram 的长度，$Count_{match}(gram_n)$ 表示同时出现在人工摘要中和系统摘要中的 n-gram 的次数。

3.2 影响因子矩阵 Φ

影响因子矩阵 $\Phi = \begin{bmatrix} \alpha_1 & \beta_1 \\ \beta_2 & \alpha_2 \end{bmatrix}$，$\alpha_1$、$\alpha_2$ 为内影响因子，β_1、β_2 为外影响因子，为了分析简便，我们可以令 $\alpha_1 = \alpha_2$，$\beta_1 = \beta_2$，得到 $\Phi' = \begin{bmatrix} \alpha & \beta \\ \beta & \alpha \end{bmatrix}$。因为要保证方程式(1)收敛，必须保证 $\alpha + \beta = 1$，所以我们可以认为，α 和 β 分别表示内增强(同一层次)影响程度和外增强(不同层次)影响程度所占的比例。α 越大，表示内增强的影响越大，反之外增强的影响越大。

对于不同的 α、β 的取值，我们对每一种体裁的文章都做了实验，由于篇幅所限，我们只给出 6 种文章体裁的平均 ROUGE-N 值，如表 1 所列，对应折线图如图 2 所示，在这里去冗余系数 $\eta = 15$。

表 1 不同的 α、β 取值测试结果

β	评价方法/压缩率			
	ROUGE-1		ROUGE-2	
	10%	20%	10%	20%
0.1	0.4259	0.5287	0.3009	0.4023
0.2	0.4137	0.5228	0.2906	0.3949
0.3	0.4048	0.5212	0.2828	0.3919
0.4	0.4063	0.5248	0.2844	0.3969
0.5	0.3938	0.5028	0.2825	0.3851
0.6	0.4059	0.5306	0.2844	0.4027
0.7	0.4078	0.5323	0.2867	0.4041
0.8	0.4067	0.5357	0.2852	0.4092
0.9	0.4109	0.5389	0.2885	0.4114

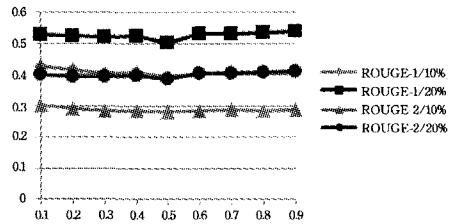

图 2 对应表 1 的折线图

从图 2 中可以看到，4 条曲线呈两边高、中间低的趋势。无论是哪一种评价方法和压缩率，都是当 $(\alpha, \beta) = (0.5, 0.5)$ 时取得最小值；压缩率为 10%，当 $(\alpha, \beta) = (0.9, 0.1)$ 时取得最大值；压缩率为 20%，当 $(\alpha, \beta) = (0.1, 0.9)$ 时取得最大值。

以上说明对于本方案，当内外影响程度相当时，效果较差，即要尽量使得内、外增强的影响程度有较大差别，才会有更好的效果。对于短文摘，应加大内增强影响，对于长文摘，应加大外增强影响。由于影响因子是算法的其中一个输入，因此我们可以根据不同的文摘要求，动态调整影响因子。

3.3 去冗余系数 η

由 2.3 节可知，$threshold = avg * \eta$，*threshold* 由 η 决定。本实验分析 η 的取值对最终文摘的影响，相似度阈值设为 0.01，所有 211 篇文章的平均相似度去冗余系数为 0.0104，分别固定 $(\alpha, \beta) = (0.9, 0.1)$ 和 $(\alpha, \beta) = (0.1, 0.9)$，变化 η，最

后令 η 取 1000，即去冗余这一步失效。这里也只给出 6 种文章体裁的平均 ROUGE-N 值，如表 2、表 3 所列。

表 2 β=0.1 时不同 η 取值测试结果

η	评价方法/压缩率			
	ROUGE-1		ROUGE-2	
	10%	20%	10%	20%
6	0.4140	0.5202	0.2710	0.3661
9	0.4222	0.5291	0.2918	0.3920
12	0.4253	0.5311	0.2989	0.4031
15	0.4259	0.5287	0.3009	0.4023
18	0.4179	0.5274	0.2965	0.4048
21	0.4105	0.5234	0.2905	0.4027
24	0.4077	0.5187	0.2900	0.3992
27	0.4044	0.5171	0.2882	0.3990
30	0.4033	0.5159	0.2875	0.3980
1000	0.3967	0.5059	0.2814	0.3873

表 3 β=0.9 时不同 η 取值测试结果

η	评价方法/压缩率			
	ROUGE-1		ROUGE-2	
	10%	20%	10%	20%
6	0.4038	0.5206	0.2653	0.3603
9	0.4132	0.5451	0.2803	0.4086
12	0.4134	0.5428	0.2891	0.4136
15	0.4109	0.5389	0.2885	0.4114
18	0.4103	0.5361	0.2911	0.4121
21	0.4065	0.5349	0.2894	0.4150
24	0.4058	0.5315	0.2912	0.4122
27	0.4052	0.5257	0.2912	0.4063
30	0.4048	0.5262	0.2913	0.4070
1000	0.3979	0.5171	0.2854	0.4002

可以看到，当没有去冗余这一步，或当 η 取值过小时，效果有明显的下滑。当 $\eta \in [9,18]$ 时效果比较好。因此对于单文本自动文摘系统，去冗这一步必不可少，但如何选取最佳的去冗余阈值，我们还要再作进一步的研究。

3.4 向量稀疏性影响分析

通过向量空间模型表示的句子和段落，当输入文本较长时，空间的维度较大，导致向量稀疏性较大。向量的稀疏性会影响相似度计算，从而影响图的构建及文摘提取的质量。

为了在一定程度上解决稀疏性对构建图及对计算结果准确性的影响，本算法在袋口词构建的过程中作了处理，在去停词的基础上，也过滤掉单个字的词，降低了空间维度。实验结果如表 4 所列，没有去单字的平均 ROUGE-N 值都要低于去单字的平均 ROUGE-N 值。

表 4 向量稀疏性对计算结果的影响

是否去单字	评价方法/压缩率			
	ROUGE-1		ROUGE-2	
	10%	20%	10%	20%
是	0.4253	0.5311	0.2989	0.4031
否	0.4034	0.5145	0.2927	0.4002

3.5 与 LexRank 比较

选择 $(\alpha,\beta)=(0.9,0.1)$、去冗余系数 $\eta=12$ 的一组数据与 LexRank 算法进行比较，如表 5、表 6 所列。

从图 3 可以看到两者的比较，压缩率为 10%时，ROUGE-1 提高了 6.14%，ROUGE-2 提高了 12.37%；压缩率为 20%时，ROUGE-1 提高了 2.53%，ROUGE-2 提高了5.72%。比起 LexRank，我们的算法有了一定的提高。

表 5 本算法测试结果

测试语料	评价方法/压缩率			
	ROUGE-1		ROUGE-2	
	10%	20%	10%	20%
03 年 863	0.5105	0.6173	0.3480	0.4605
奥运	0.4494	0.5308	0.3251	0.3978
记叙文	0.502	0.5534	0.3896	0.4292
议论文	0.4015	0.5232	0.2825	0.4083
应用文	0.2726	0.4141	0.1683	0.3058
说明文	0.4155	0.5477	0.2797	0.4168
平均	0.4253	0.5311	0.2989	0.4031

表 6 LexRank 测试结果

测试语料	评价方法/压缩率			
	ROUGE-1		ROUGE-2	
	10%	20%	10%	20%
03 年 863	0.4716	0.5648	0.3097	0.3932
奥运	0.4015	0.4957	0.2655	0.3616
记叙文	0.4620	0.5561	0.3492	0.4375
议论文	0.3534	0.4830	0.2370	0.3600
应用文	0.3303	0.4841	0.1949	0.3553
说明文	0.3853	0.5245	0.2398	0.3802
平均	0.4007	0.5180	0.2660	0.3813

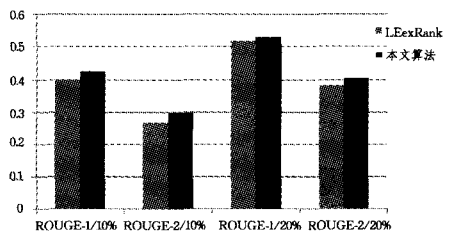

图 3 LexRank 与本算法平均 ROUGE-N 值的对比

结束语 本文提出了一种新的基于图的互增强模型——段落-句子互增强模型，它利用了段落间的关系，避免了计算词语间相似度，将其运用在单文本自动摘要中，获得了较高的文摘质量。我们还对内、外影响因子这两个重要参数作了分析，实验表明，应使内增强（同一层次）、外增强（层次之间）的影响程度有较大差别，文摘质量才更高，对不同长度的文摘，它们的影响程度也要有所区别。最后分析了文摘句提取中去冗余这一步的必要性。

未来的工作主要有两项：1. 我们在研究中发现，有些文章的段落较短，甚至只有一个句子，这样的短段落会对该模型的效果产生一定影响，未来可尝试先对段落进行聚类，再应用该模型；2. 本文并没有找到一个最佳的去冗余阈值，未来可以使用其他机器学习的方法对该参数进行估计，找到一个最佳的阈值。

参 考 文 献

[1] Langville A N, Meyer C D. Deeper inside PageRank[J]. Journal of Internet Mathematics, 2004, 1(3): 335-380

[2] Kleinberg J M. Authoritative Sources in a Hyperlinked Environment[J]. Journal of the ACM, 1999, 46(5): 604-632

[3] Erkan G, Radev D R. LexRank: graph-based centrality as salience in text summarization[J]. Journal of Aritficial Intelligence Research, 2004, 22: 457-479

[4] Mihalcea R, Tarau. TextRank: Bringing Order into Text[C]// Proceedings of EMNLP. Barcelona Spain, 2004: 404-411

[5] Zhu Xiao-jin, Goldberg A B, Van Gael J, et al. Improving Diversity in Ranking using Absorbing Random Walks[C]// Procee-

dings of NAACL HLT, 2007. Rochester NY, 2007: 97-104

[6] Zha Hong-yuan. Generic Summarization and Key Phrase Extraction using Mutual Reinforcement Principle and Sentence Clustering[C]// Proceedings of ACM SIGIR, 2002. Tampere Finland, 2002: 113-120

[7] Wei Fu-ru, Li Wen-jie, Lu Qin, et al. Applying Two-Level Reinforcement Ranking in Query-Oriented Multidocument Summarization[J]. Journal of the American Society for Information Science and Technology, 2009, 60(10): 2119-2131

[8] Wei Fu-ru, Li Wen-jie, Lu Qin, et al. Query-Sensitive Mutual Reinforcement Chain and Its Application in Query-Oriented Multi-Document Summarization[C]// SIGIR, 2008. Singapore, 2008

[9] Bollegala D, Matsuo Y, Ishizuka M. Measuring Semantic Similarity between Words using Web Search Engines[C]// Proceedings of WWW. 2007: 757-766

[10] Cilibrasi R L, Vitanyi P M B. The Google Similairity Distance [J]. IEEE Transactions on Knowledge and Data Engineering, 2007, 19(3): 370-383

[11] Sahami M, Heliman T D. A Web-based Kernel Function for Measuring the Similarity of Short Text Snippets[C]// Proceedings of WWW. 2006: 377-386

[12] Che Wan-xiang, Li Zheng-hua, Liu Ting. LTP: A Chinese Language Technology Platform[C]// Proceedings of the Coling 2010: Demonstrations. Beijing, China, 2010: 13-16

[13] Lin C-Y, Hovy E H. Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics [C]// Proceeding of 2003 Language Technology Conference (HLT-NAACL 2003). Canada, 2003

[14] Fellbaum C. WordNet [Z]. Theory and Application of Ontology: Computer Applications, 2010: 231-243

[15] Jones K S. A Statistical Interpretation of Term Specificity and Its Application in Retrieval[J]. Journal of documentation, 1972, 28(1): 11-21

(上接第 234 页)

及安全策略管理员制定策略视角的不同，都会导致安全策略合理性被质疑。因此在 TDSD-Droid 学习机制中，引入了人为的处理流程，特别强调有经验安全策略管理员的参与。在将来，希望能为该安全策略学习机制开发出一套仿人工智能测试系统，以便实现安全策略学习更全面的自动化。

结束语 这篇文章展现了 TDSD-Droid 基于强制访问控制机制的立体安全防御系统设计与实现，介绍了 TDSD-Droid 的架构，上、下一致的 MMAC 与 MAC 强制访问控制机制、FSP 灵活性安全策略机制，以及完整性验证机制等。本文也演示了 TDSD-Droid 面对 Android 攻击的防御功效，并给出了相关的性能测试结果。

参 考 文 献

[1] Llamas R, Restivo K, Shirer M. Android Marks Fourth Anniversary Since Launch with 75.0% Market Share in Third Quarter [EB/OL]. https://www.idc.com/getdoc.jsp?containerId=prUS23771812, IDC, 2012

[2] Kleidermacher D, Kleidermacher M. Embedded System Security Practical Methods for Safe And Secure Software and Systems Development [M]. Waltham, MA, USA: Elsevier Inc, 2012: 4-24

[3] Armando A, Merlo A, Verderame L, et al. An Empirical Evaluation of the Android Security Framework [C]// Proceedings of the 28th IFIP TC-11 International Information Security and Privacy Conference (SEC 2013). Auckland: Springer, 2013: 176-189

[4] Smalley S, Craig R. Security Enhanced (SE) Android: Bringing Flexible MAC to Android [C/OL]. http://selinuxproject.Org/~se-android/papers/NDSS2013-SEAndroid-Paper.pdf, NDSS, 2013

[5] Jhswx84. SELinux 详解[M/OL]. http://wenku.baidu.com/view/4d26594fc850ad02de804189.html, Baidu, 2012

[6] Enck W, Ongtang M, McDaniel P. Understanding Android security [J]. IEEE Security and Privacy Magazine, 2009 7(1): 50-57

[7] Sally. SELinux 学习笔记[M/OL]. http://wenku.it168.com/d_001220063.shtml. IT168, 2013

[8] Spencer R, Smalley S, Loscocco P, et al. The Flask security architecture: System support for diverse security policies [C]// Proceedings of The Eighth USENIX Security Symposium. Washington: USENIX, 1999: 123-139

[9] Carter J. Using gconf as an example of how to create an userspace object manager [C/OL]. http://www.nsa.gov/research/_files/selinux/papers/gconf07-paper.shtml, NSA, 2009

[10] NSA. SE For Android [EB/OL]. http://selinuxproject.org/page/SEforAndroid. NSA, 2013

[11] Ongtang M, McLaughlin M, Enck W, et al. Semantically rich application-centric security in Android[J]. Security and Communication Networks, 2012, 5(6): 658-673

[12] Bugiel S, Davi L, Dmitrienko A, et al. Practical and Lightweight Domain Isolation on Android[C]// Proceedings of the 1st ACM workshop on Security and privacy in smartphones and mobile devices (SPSM '11). New York: CCS, 2011: 51-62

[13] Enck I, Gilbert P, Chun B, et al. TaintDroid: An Information-Flow Tracking System for Realtime Privacy Monitoring on Smartphones[C]// proceeding of: 9th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2010. Vancouver, BC, Canada: USENIX, 2010: 1-6

[14] Bugiel S, Davi L, Dmitrienko A, et al. Towards Taming Privilege-Escalation Attacks on Android [C/OL]. http://www.trust.informatik.tu-darmstadt.de/fileadmin/user_upload/Group_TRUST/PubsPDF/NDSS_2012_Towards_Taming_Privilege-Escalation_Attacks_on_Android.pdf, NDSS, 2012

[15] Bea F. WhatsApp reads your phone contacts and is breaking privacy laws [CP/OL]. http://www.digitaltrends.com/mobile/whatsapp-breaks-privacy-laws/, DTDigital Trends, 2013

[16] Cai H, Shao Z, Vaynberg A. Certified Self-Modifying Code [C]// Proceedings of 2007 ACM SIGPLAN Conference on Programming Language Design and Implementation. San Diego: PLDI' 2007: 66-77

[17] AnTuTu Labs. AnTuTu Benchmark [CP/OL]. https://play.google.com/store/apps/details?id=com.antutu.ABenchMark, Google, 2013