

# 由排序支持向量机抽取博客文章的摘要

何海江<sup>1</sup>, 陈 姝<sup>2</sup>

(1. 长沙学院计算机系 长沙 410003; 2. 中南大学信息科学与工程学院 长沙 410083)

**【摘要】**提出了一种用平滑型排序支持向量机(Rank-sSVM)抽取博客文章摘要的方法。使用该排序算法抽取的摘要,反映了评论者的意见和博客文集的特性。自动摘要过程中,首先经人工从文章选择重要句子标记为摘要,作为训练对象;再由机器生成表示文章语句的特征集,共14个特征,包含标签、评论等博客文章独有的信息;最后用Rank-sSVM学习人工摘要后,将文章所有句子排序,选取最靠前的若干语句构成摘要。该方法在一个中文博客数据集上取得良好效果。

**关键词** 博客; 评论; 信息检索; 排序学习; 支持向量机; 摘要

**中图分类号** TN391

**文献标识码** A doi:10.3969/j.issn.1001-0548.2010.04.026

## Extraction of Blog Post Summarization by Using Ranking SVM

HE Hai-jiang<sup>1</sup> and CHEN Shu<sup>2</sup>

(1. Department of Computer Science and Technology, Changsha University Changsha 410003;

2. School of Information Science and Engineering, Central South University Changsha 410083)

**Abstract** A new approach is presented for blog post summarization based on ranking smooth support vector machine (Rank-sSVM). The use of ranking algorithm for this task allows one to adapt summaries to the commenter needs and to the blog corpus characteristics. To use Rank-sSVM, firstly, key sentences are extracted manually from blog posts as training samples. Feature set representing post sentences, which consist of 14 features including tag, comment and other unique blog information, is generated by machine. After all the sentences are ranked by the ranking model, the most important ones in front are selected to summarize the post. The experimental results show that the proposed method has good performance on Chinese blog datasets.

**Key words** blog; comment; information retrieval; learning to rank; support vector machine; summarization

博客(blog)属Web2.0的重要应用,是一种传播信息和思想的新媒体。博客作者发表文章(post),用以记录日常生活,向外发布信息,讨论新闻事件及人物。随着博客社区的繁荣,出现了许多新的应用形式,如专门的博客搜索引擎(百度BlogSearch、Technorati等)和可阅读博客的手机等小型移动设备。

自动产生博客文章摘要使读者可以先浏览博客的摘要部分,再评估文章是否值得通篇阅读;或者只阅读摘要,克服小尺寸屏幕的限制。

可将博客文章视为一系列句子的组合。每条语句不仅包括段落号和句子号等结构特征、关键词密度等语义特征;还包括标签(tag)、评论(comment)等博客特征。与普通文章不同,作者在博客文章上常常添加标签,用以改善Web检索质量<sup>[1]</sup>。另外,读者被鼓励在阅读博客文章后发表评论,与作者或其他读者交流。博客网站会将评论多的文章列为热门,

搜索引擎也将评论作为博客评级的重要指标。实验结果表明,阅读评论后人们会更改文章的摘要<sup>[2]</sup>。

本文提出了一种平滑型排序支持向量机(Rank-sSVM)作为自动摘要的排序学习算法。依据结构风险最小化原则,Rank-sSVM将排序问题转换成两类别分类问题;分类优化目标变成无约束的数学规划问题,并具有唯一最优解;使目标函数二次可微,用平滑多项式逼近 $(1-x)_+$ ;最后由Newton-YUAN<sup>[3-4]</sup>算法直接求解。依据排序学习算法,可将语句按照重要程度从大到小排列,选取最靠前的若干句子形成博客文章的摘要。

## 1 相关研究工作

总的来说,摘要可分为概括型摘要和抽取型摘要。无论是单文档,还是多文档,抽取型摘要受到大多数研究者的关注,本文亦属于抽取型摘要研究。

收稿日期: 2008-11-28; 修回日期: 2009-07-23

基金项目: 湖南省教育厅科学研究项目(09c123)

作者简介: 何海江(1970-), 副教授, 主要从事Web挖掘、机器学习方面的研究。

从技术路线划分,产生抽取型摘要的方法有3类:(1)基于特征的方法,如关键词位置信息、词频信息(SumBasic<sup>[5]</sup>、SumFocus<sup>[6]</sup>)等;(2)基于图的方法<sup>[7-8]</sup>;(3)机器学习方法。分类是一种重要的机器学习算法,将语句划分为“重要”或“不重要”,再由贝叶斯<sup>[9]</sup>、Logistic回归<sup>[10]</sup>、支持向量机(support vector machine, SVM)<sup>[11]</sup>等算法将句子分类。文章的语句重要与否是相对的,与分类相比,排序学习更适用。Logistic回归排序<sup>[12]</sup>、神经网络排序<sup>[13-14]</sup>等排序学习算法在自动摘要领域的应用成为研究热点。排序支持向量机用于自动摘要则还未见报道。

Rank-sSVM与以前的排序支持向量机有两点不同<sup>[15]</sup>:误差项是二次函数;最优化目标为无约束问题,直接用Newton法求解,训练速度更快。

互联网的发展要求自动摘要技术适应Web文档,包括Web页面和E-mail等。然而,博客文章摘要的研究却很少,文献[2]和[8]虽然能概括博客文章,但完全面向评论,忽略了作者的意图及其他博客特征,属于偏重型摘要。本文的研究则综合了作者和评论者的表达思想,属于通用型摘要。

## 2 平滑型排序支持向量机Rank-sSVM

令 $s_a >_* s_b$ 为优先关系,若博客文章 $d$ 的句子 $s_a$ 和 $s_b$ 存在 $s_a >_* s_b$ ,则在挑选 $d$ 的摘要语句时, $s_a$ 比 $s_b$ 有优先权。文档 $d$ 有 $n$ 条句子,其中 $v(v < n)$ 条句子被人工标注为摘要,令SST为摘要语句集,SNS为非摘要语句集,可构造 $v*(n-v)$ 个优先关系:

$$\forall (s_a \in \text{SST}, s_b \in \text{SNS}) \quad s_a >_* s_b \quad (1)$$

### 2.1 排序SVM

博客集有 $m$ 篇文章,全部包含人工标注的摘要。对所有优先关系,排序SVM的优化目标为问题1<sup>[15]</sup>:

$$\min_{w, \xi} \frac{1}{2} w^* w + C \sum \xi_{k,i,j}^r \quad (2)$$

使得:

$$\forall (s_{1,i} >_* s_{1,j}): w^* \phi(s_{1,i}) \geq w^* \phi(s_{1,j}) + 1 - \xi_{1,i,j},$$

$$p(x, \eta) = \begin{cases} 1-x & x \leq 1-\eta \\ -\frac{1}{16\eta^3}(x+3\eta-1)(x-1-\eta)^3 & 1-\eta < x < 1+\eta \\ 0 & x \geq 1+\eta \end{cases} \quad (7)$$

式中  $0 < \eta < 1$ ;  $p(x, \eta)$ 具有二阶光滑性,  $\eta$ 越小,在区间 $(1-\eta, 1+\eta)$ 越逼近 $(1-x)_+$ 。

定理 2  $p(x, \eta)$ 二阶光滑。

定理 3 (1)  $p(x, \eta) \geq (1-x)_+$ ;

(2)  $p(x, \eta)^2 - (1-x)_+^2 \leq 0.0515\eta^2$ 。

$$\begin{aligned} \forall (s_{2,i} >_* s_{2,j}): w^* \phi(s_{2,i}) &\geq w^* \phi(s_{2,j}) + 1 - \xi_{2,i,j}, \dots, \\ \forall (s_{m,i} >_* s_{m,j}): w^* \phi(s_{m,i}) &\geq w^* \phi(s_{m,j}) + 1 - \xi_{m,i,j} \\ \forall k, i, j: \xi_{k,i,j} &\geq 0 \end{aligned} \quad (3)$$

式中  $s_{k,j}$ 是文章 $d_k$ 的第 $j$ 条语句;  $\Phi(\cdot)$ 是一个特征映射函数,将原始空间的向量映射到高维空间,以解决线性不可分问题;  $*$ 代表内积运算;  $\xi$ 是使约束条件成立的误差标量;  $C$ 是误差和模型复杂度间的平衡因子。

将样本对 $s_a$ 和 $s_b$ 组合成一个新样本 $x_{a,b} = s_a - s_b$ ,令 $y_{a,b} = +1(s_a >_* s_b)$ 或 $-1(s_b >_* s_a)$ ,则排序问题1可转化成分类问题2:

$$\min_{w, \xi} \frac{1}{2} w^* w + C \sum_i \xi_i^r \quad (4)$$

使得:

$$\begin{cases} \forall i: y_i(w^* \sum_j K(x_j, x_i)) \geq 1 - \xi_i \\ \forall i: \xi_i \geq 0 \end{cases} \quad (5)$$

核函数 $K(x_i, x_j) = \Phi(x_i)^T \Phi(x_j)$ 。设 $w^*$ 是问题2的解,则分类超平面为 $f_w = w^* \sum_j K(x_j, s_a - s_b) = 0$ 。除缺少超平面的偏移量外,问题2与常规SVM相同,固然可由常规SVM方法求解<sup>[15]</sup>。

### 2.2 分段光滑多项式

目前已有的排序SVM集中于一次误差函数的情况,即问题1、2的 $r=1$ ,而提出的Rank-sSVM使用二次误差函数( $r=2$ )。引入函数 $(1-x)_+ = \max(0, 1-x)$ 后,问题2( $r=2$ 时)与无约束问题3等价:

$$\min_w f(w) = \frac{1}{2} w^* w + C \sum_i (1 - y_i(w^* \sum_j K(x_j, x_i)))_+^2 \quad (6)$$

定理 1 式(6)是严格的凸函数,问题3有唯一最优解。

$(1-x)_+$ 不可微,为了直接用Newton法求解问题3,定义分段多项式函数 $p(x, \eta)$ 近似 $(1-x)_+$ ,有:

### 2.3 Rank-sSVM的模型及收敛特性

将 $p(x, \eta)$ 代入式(6),Rank-sSVM的优化目标为问题4:

$$\min_w \psi_\eta(w) = \frac{1}{2} w^* w + C \sum_i p(y_i(w^* \sum_j K(x_j, x_i)), \eta)^2$$

(8)

式(8)是严格的凸函数, 问题4有唯一最优解。随着 $\eta$ 趋向于零, 式(8)的最优解收敛于式(6)的最优解。

**定理 4** 若 $w^*$ 是问题3的最优解,  $w^\#$ 是问题4的最优解, 数据集共 $h$ 个样本, 则有:

$$(w^\# - w^*)^2 \leq hC \times 0.051 5\eta^2$$

由于 $\psi_\eta(w)$ 二次可微, 可结合Newton法和YUAN<sup>[3]</sup>的一维精确搜索算法求解。将文献[4]中Newton-YUAN算法的 $(w, b)$ 变为 $w$ ,  $\eta$ 修改为 $(\frac{\epsilon_1}{0.051 5hC})^{0.5}$ 即可求解Rank-sSVM。 $\epsilon_1$ 是式(8)最优解与式(6)最优解的最大误差, 梯度精度 $\epsilon_2$ 为Newton法迭代终止条件, 以近似最优解<sup>[17]</sup>替代实际解。

### 3 表示博客文章语句的特征

若文章Post有 $n$ 个段落, 它的第 $k$ 段 $P_k$ 的句子总数为 $m$ , 记段落 $P_k$ 的第 $t$ 个句子为 $S_{kt}$ 。

向量空间模型是常用的文档表示模型,  $TF_{w,content}$ 是词 $w$ 在content(语句、段落、文章、文章标题、评论)内出现的次数,  $IDF_w$ 是词 $w$ 的反文档频率, 则 $w$ 在content的信息熵 $E_{w,content} = \ln(1 + TF_{w,content}) \times IDF_w$ 。

评论往往有多条, 将它们当作一个整体, 记为Comments。有些评论者为强调自己的观点, 将同一句话或几个词在评论段反复拷贝, 因此, 设计了两种模式Bernoulli、Multinomial考察它们对摘要学习的影响。若词 $w$ 在一条评论中出现 $x$ 次, 依照Bernoulli模式,  $TF_{w,Comments}=1$ ; 而依照Multinomial模式,  $TF_{w,Comments}=x$ 。

简要描述句子的3种类型的共14个特征, 依此可生成表示文章语句的特征集, 如表1所示。

表1 博客文章句子的特征

类型	特征名称	简要描述
结构特征	LP	段落位置信息
	LS	句子位置信息
	RL	句子的相对长度
	ST	与标题的相似度
	KWD	关键词密度
语义特征	ICu <sup>[13]</sup>	是否包含线索词
	RE	相对信息熵
	SF	与同段落首句的相似度
	Sum <sub>Avr</sub> <sup>[6]</sup>	平均词频
	ICe	是否交谈语句
	RT	标签词比例
	博客特征	RCE
RCN		语句评论名词比例
CB <sub>Avr</sub>		语句平均评论词频

LP: 若句子位于首段或尾段, 其值为1, 否则 $LP(S_{kt})=(n-k)\div n$ 。

LS: 若句子是Post的首句, 其值为1; 尾句或每段的第一句, 值为0.8; 首段和尾段的段落系数 $coep=0.6$ , 其他段 $coep=0.4$ , 句子 $S_{kt}$ 的 $LS(S_{kt})=(t-1)\div m \times coep$ 。

RL: 语句包含字符数称为长度,  $RL(S_{kt})= S_{kt}$ 的长度 $\div$ Post语句最大长度。

ST: 语句 $S_{kt}$ 与文章标题Title的相似度,

$$ST(S_{kt}) = \frac{\sum_{w \in S_{kt} \cap Title} E_{w,S_{kt}} \times E_{w,Title}}{\sqrt{\sum_{w \in S_{kt}} (E_{w,S_{kt}})^2} * \sqrt{\sum_{w \in Title} (E_{w,Title})^2}}$$

KWD: 先计算Post所有词在文章的信息熵, 前1/3最高信息熵值的词称为关键词,

$$KWD(S_{kt}) = \frac{S_{kt} \text{的关键词总数}}{S_{kt} \text{的词总数}}$$

ICu<sup>[13]</sup>: 布尔量, 描述是否包含诸如“总的说来”、“综上所述”等线索词。

RE: 语句的信息熵 $\zeta(S_{kt}) = \sum_{w \in S_{kt}} E_{w,S_{kt}}$ , 归一化的

$$RE(S_{kt}) = \frac{\zeta(S_{kt})}{\max_{S_{ki} \in Post} (\zeta(S_{ki}))}$$

SF: 语句 $S_{kt}$ 与所在段落 $P_k$ 首句 $S_{k1}$ 的关联度。类似于ST。

Sum<sub>Avr</sub><sup>[6]</sup>:  $w$ 的文章词频 $p(w) = \frac{TF_{w,post}}{\text{post的词总数}}$ ,

$$\text{而 Sum}_{avr}(S_{kt}) = \frac{\sum_{w \in S_{kt}} p(w)}{S_{kt} \text{的词总数}}$$

Ice: 布尔量, 描述语句是否表现为对话形式。

RT:  $RT(S_{kt})=S_{kt}$ 包含标签词种类数 $\div$ 标签词数目。

RCE: 语句 $S_{kt}$ 的评论信息熵 $\tau(S_{kt}) =$

$$\sum_{w \in S_{kt} \cap Comments} TF_{w,S_{kt}} \times E_{w,Comments}, \text{ 经归一化 } RCE(S_{kt}) = \frac{\tau(S_{kt})}{\max_{S_{ki} \in Post} \tau(S_{ki})}$$

。计算RCE时有两种模式。

RCN: 评论提到的名词往往是文章的焦点,

$$RCN(S_{kt}) = \frac{S_{kt} \text{出现在Comments中的名词数}}{S_{kt} \text{的词总数}}$$

CB<sub>Avr</sub>: 类似Sum<sub>Avr</sub>, 只不过词频计算资源不再是文章, 而是评论集, 词 $w$ 的评论词频

$$p(w) = \frac{TF_{w,Comments}}{\text{Comments的词总数}}, \text{ 相应地 } CB_{Avr}(S_{kt}) = \frac{\sum_{w \in S_{kt} \cap \text{Comments}} p(w)}{S_{kt} \text{的词总数}}.$$

计算  $p(w)$  时有 Bernoulli 和 Multinomial 两种模式。

## 4 实验数据集及实验结果分析

考虑到自动摘要系统的实用要求, 将每篇文章的摘要固定为5个句子。实际上, 更改句子数目, 并不影响本文的摘要方法, 甚至都不需要重新学习。使用中科院计算所汉语分词软件ICTCLAS分词, 除RL外, 其他特征在计算之前先消除停用词。

### 4.1 中文博客数据集

编写一个博客抓取器, 从新浪博客随机下载120篇文章, 涵盖科技、财经、娱乐、美食等多个方面。这些文章经解析后构成实验数据集, 表2是数据集的基本情况, 其中标签数是指标签经分词后的词数目。

表2 数据集基本信息

参数	值或范围
句子数目范围	10~79
每篇文章平均句子数	28.2
标签数目范围	1~20
每篇文章平均标签数	6.3
评论数目范围	4~239
每篇文章平均评论数	51.7

由3个人完成摘要编写, 在浏览文章、标签及所有评论后, 3个人分别从文章中抽取5条语句作为摘要。

### 4.2 对比方法和性能评估测度

为了比较自动摘要学习的有效性, 设计了一种称为Baseline的基准方法: 先抽取文章的首句和末句, 在保证不重复的前提下, 再随机抽取每段的第一句, 若仍不足, 则从文章的第二句开始逐一选择, 直到满5句为止。另外, 为考察标签和评论对自动摘要方法性能的影响, 将Rank-sSVM和特征组合成3种学习方法: 博客特征不参与训练及测试的称为PostOnly, 采用Bernoulli模型计算评论相关特征的称为PCmBerl, 采用Multinomial模型的则称为PCmMnom。

NDCG和ROUGE是评估自动摘要系统性能的常用测度, ROUGE系列测度选用ROUGE-1, 以中文的词为基本单元。

精度  $Precision = \frac{\text{自动摘要与人工摘要重合句子数}}{5}$ ,

也被选作评估测度。为适用NDCG的计算, 按照如下方法对文章句子分级: 若一语句被R个人标注为摘要, 级别为R, 无一人标注, 则级别为0。

### 4.3 实验结果及分析

Rank-sSVM选用高斯核, Newton-YUAN算法参数  $\epsilon_1=10^{-3}$ ,  $\epsilon_2=10^{-4}$ 。表3、表4和表5分别比较Baseline、PostOnly、PCmBerl和PcmMnom等4种方法的Precision、NDCG和ROUGE-1。所有实验结果皆为5折交叉的平均值。

$H_x(x=1,2,3)$ 表示以第x人的摘要为理想摘要, 也就是基准,  $H_a$ 则是将3个人的摘要合并, 视作理想摘要, 显然此时摘要句子数在5~15之间, Precision也比 $H_x$ 高得多。

表3 Precision对比表(%)

方法	$H_1$	$H_2$	$H_3$	$H_{1-3}$ 平均	$H_a$
Baseline	31.2	33.2	31.8	32.1	52.7
PostOnly	43.8	44.3	39.8	42.6	67.7
PCmBerl	47.8	46.5	40.7	45.0	71.3
PcmMnom	47.3	46.5	40.7	44.8	71.2

表4 NDCG对比表

方法	NDCG值
Baseline	0.54
PostOnly	0.642
PCmBerl	0.667
PcmMnom	0.666

表5 ROUGE-1对比表

方法	$H_1$	$H_2$	$H_3$	$H_a$
Baseline	0.403	0.412	0.41	0.405
PostOnly	0.603	0.616	0.584	0.605
PCmBerl	0.633	0.624	0.583	0.617
PcmMnom	0.632	0.626	0.585	0.618

比较表3、表4和表5的结果, 自动摘要学习方法非常有效。即使PostOnly与Baseline相比,  $H_x$ 的平均精度提高32.7%,  $H_a$ 精度提高28.5%, NDCG提高18.9%, ROUGE-1( $H_a$ )提高49.3%。Baseline得方法分别以3人的摘要为基准, 其精度都达到31%以上。事实上, 通过观察也可发现, 博客作者往往用第一条语句直入主题, 最后一条语句发表总结性言论, 每段的第一句表现段落主旨。

加入博客特征, 学习效果尽管没有显著提高, 但也有明显改善, PcmBerl与PostOnly相比,  $H_x$ 平均高2.4%, NDCG高0.025, ROUGE-1高0.012。性能差别没有预想的大, 估计与下列因素有关:

(1) 评论集有噪声, 存在不少与文章和博客无关的垃圾评论。

(2) 许多评论者并非讨论文章的主题, 只是与作者或其他评论者打招呼或聊天。

当然, 人的因素也不可忽视, “理想”摘要受人的情绪、态度干扰, 并非总是理想。从表3和表5可看出,  $H_3$ 与 $H_1$ 、 $H_2$ 相比, 明显对评论和标签重视不够。总的说来, 在基于机器学习的博客文章自动摘要系统中, 选择博客特征能明显改善性能。

Bernoulli模型和Multinomial模型则几乎没有差别, PCmBerl和PcmMnom相比, 无论是精度和NDCG, 还是ROUGE-1, 都近乎相等。明确提示, 在计算博客特征时, 不必在评论模型选择上费工夫。

## 5 结束语

本文提出了一种新的排序支持向量机, 其误差项为二次函数, 可用Newton法直接求解。并首次将排序学习用于博客文章的自动摘要抽取。Rank-sSVM结合博客文章的结构特征、语义特征、博客特征, 自动抽取摘要, 性能良好。虽然实验在中文环境完成, 但本文方法能够适用于其他语言的博客。

下一步, 将分析14个语句的特征的权, 降低特征提取时间复杂度, 缩短摘要时间。

### 参 考 文 献

- [1] BAO Sheng-hua, WU Xiao-yuan, FEI Ben, et al. Optimizing web search using social annotations[C]//Proceedings of the 16th World Wide Web Conference. New York, USA: ACM, 2007: 501-510.
- [2] HU Mei-shan, SUN Ai-xin, LIM Ee-peng. Comments-oriented blog summarization by sentence extraction[C]//Proceedings of 16th ACM International Conference on Information and Knowledge Management. New York, USA: ACM, 2007: 901-904.
- [3] YUAN Ya-xiang. Step-sizes for the gradient method[C]//Proceedings of the Third International Congress of Chinese Mathematicians. Ann Arbor, USA: American Mathematical Society, 2004: 785-796.
- [4] 何海江. 代价与样本相关的简约核支持向量机[J]. 计算机应用, 2008, 28(11): 2863-2866.  
HE Hai-jiang. Costs sensitive to examples learning in reduced support vector machine[J]. Journal of Computer Applications, 2008, 28(11): 2863-2866.
- [5] NENKOVA A, VANDERWENDE L, MCKEOWN K. A compositional context sensitive multi-document summarizer: exploring the factors that influence summarization[C]//Proceedings of the 29th ACM Special Interest Group on Information Retrieval. New York, USA: ACM, 2006: 573-580.
- [6] VANDERWENDE L, SUZUKI H, BROCKETT C, et al. Beyond sumbasic: task-focused summarization with sentence simplification and lexical expansion[J]. Information Processing & Management, 2007, 43(6): 1606-1618.
- [7] WAN Xiao-jun. Using only cross-document relationships for both generic and topic-focused multi-document summarizations[J]. Information Retrieval, 2008, 11(1): 25-49.
- [8] HU Mei-shan, SUN Ai-xin, LIM Ee-peng. Comments-oriented document summarization: understanding documents with readers' feedback[C]//Proceedings of 31st ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM, 2008: 291-298.
- [9] DAUME H, MARCU D. Bayesian query-focused summarization[C]//Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of ACL. Stroudsburg, USA: Association for Computational Linguistics, 2005: 305-312.
- [10] AMINI M, GALLINARI P. The use of unlabeled data to improve supervised learning for text summarization[C]//Proceedings of the 25th ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM, 2002: 105-112.
- [11] LE N M, SHIMAZU A, PHAN X H, et al. Sentence extraction with support vector machine ensemble[C]//Proceedings of the First World Congress of the International Federation for Systems Research. Japan: JAIST Press, 2005.
- [12] AMINI M R, USUNIER N, GALLINARI P. Automatic text summarization based on word-clusters and ranking algorithms[C]//Proceedings of the 27th European Conference on Information Retrieval. Berlin: Springer, 2005: 142-156.
- [13] SVORE K, VANDERWENDE L, BURGESS C. Enhancing single-document summarization by combining RankNet and third-party sources[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg, USA: Association for Computational Linguistics, 2007: 448-457.
- [14] SVORE K, VANDERWENDE L, BURGESS C. Using signals of human interest to enhance single-document summarization[C]//Proceedings of the 23rd AAAI Conference on Artificial Intelligence. Menlo Park, USA: AAAI Press, 2008: 1577-1580.
- [15] XU Jun, CAO Yun-bo, LI Hang, et al. Cost-sensitive learning of SVM for ranking[C]//Proceedings of the 17th European Conference on Machine Learning. Berlin: Springer, 2006: 833-840.
- [16] 袁玉波, 严杰, 徐成贤. 多项式光滑的支撑向量机[J]. 计算机学报, 2005, 28(1): 9-17.  
YUAN Yu-bo, YAN Jie, XU Cheng-xian. Polynomial smooth support vector machine(PSSVM)[J]. Chinese Journal of Computers, 2005, 28(1): 9-17.
- [17] ZHANG Xiao-ling, DU Li, ZHANG Guang-wei, et al. An improved genetic algorithm with quasi-gradient crossover[J]. Journal of Electronic Science and Technology of China, 2008, 6(1): 47-51.

编辑 蒋晓