

邮件过滤中特征选择算法的研究及改进

卢扬竹, 张新有, 祁 玉

(西南交通大学 信息科学与技术学院, 成都 610031)

(luyangzhu117@163.com)

摘 要:对基于内容的垃圾邮件过滤技术尤其是特征选择算法进行了研究。在此基础上,对其中的互信息算法进行了分析,并将其与邮件过滤的特点结合起来进行,在频度、集中度及分散度三个指标上进行改进,在原互信息算法已考虑分散度的基础上,引入词频来表征频度,以类别贡献比来衡量特征对分类的贡献,即表征集中度,并给出了改进后的互信息计算公式及算法。最后使用真实邮件训练集进行了邮件分类的实验,实验结果证明对互信息算法的改进能有效提高邮件分类性能。

关键词:垃圾邮件;文本分类;特征选择;互信息

中图分类号: TP393 **文献标志码:** A

Improvement of feature selection method in spam filtering

LU Yang-zhu, ZHANG Xin-you, QI Yu

(School of Information Science and Technology, Southwest Jiaotong University, Chengdu Sichuan 610031, China)

Abstract: Spam filtering techniques based on content, especially feature selection algorithm was studied. Based on that, Mutual Information (MI) algorithm, combined with the feature of spam filtering, was analyzed and improved according to frequency, divergence, and concentration. Comparing with conventional mutual information algorithm, word frequency was introduced, and ratio of mutual information was used to evaluate the contribution to classifying provided by features. The improved formula and algorithm were given. At last, simulation test with real E-mail set, was conducted, which shows that the improved mutual information algorithm provides a better result for spam classification.

Key words: spam; text classification; feature selection; mutual information

0 引言

目前垃圾邮件过滤技术研究集中在基于内容解析和基于行为解析两个方面。基于内容的解析,其理论与应用的发展较为成熟,其对邮件的解析过程大致为:提取邮件文本,邮件文本预处理,分词,特征选择,权重计算及文本分类,最终达到垃圾邮件过滤的目的。在以上流程中,一封邮件经过预处理后即被表示为一个特征空间,但此特征空间的维数必然相当大,如果直接在此基础上进行分类,将对分类算法造成很大的负担,并且原始特征空间中必定含有相当一部分与类别无关的、冗余的特征,将对分类结果造成影响,因此必须在分类之前对特征空间进行维数约减,保留下对分类贡献最大的那些特征,这个过程便是特征选择,其算法优劣直接影响到分类效果。

特征选择的常用方法有文档频数(Document Frequency, DF)、信息增益(Information Gain, IG)、CHI统计和互信息(Mutual Information, MI)等。本文将对以上特征选择算法进行研究并对其中的互信息算法进行改进,以选出对类别最具代表性的特征集合,为垃圾邮件的后续分类过程提供更好的基础。

1 特征选择方法及在邮件过滤中的应用

1.1 常用特征选择方法

1.1.1 文档频数

文档频数是最简单的特征选择方法,它是指文档集中包

含某特征的文档数,这种方法用某类文档集中包含某特征的文档数与此类别总文档数之比来衡量特征对类别的贡献^[1]。包含此特征的文档数越多,这个比值越大,则说明此特征越能代表此类别。而比值小于预先设定的阈值的特征,在特征选择的结果中将不被保留。但是在特征提取(Information Extraction, IE)的研究中,通常认为这些文档频率低的特征,也可能对类别具有较多的贡献,不应该全部排出特征子集。

1.1.2 信息增益

信息增益采用包含某特征的文档频率和不包含此特征的文档频率来衡量特征对分类的贡献^[1],如式(1):

$$IG(t) = - \sum_{i=1}^m P(C_i) \log P(C_i) + P(t) \sum_{i=1}^m P(C_i | t) \log P(C_i | t) + P(\bar{t}) \sum_{i=1}^m P(C_i | \bar{t}) \log P(C_i | \bar{t}) \quad (1)$$

其中: m 为类别总数, $P(C_i)$ 为文档集中 C_i 类出现的概率, $P(t)$ 为包含特征 t 的文档频率, $P(C_i | t)$ 为文档包含特征 t 时其属于 C_i 类的概率, $P(\bar{t})$ 为不包含特征 t 的文档频率, $P(C_i | \bar{t})$ 为文档不包含特征 t 时其属于 C_i 类的概率。

这种方法的特点是考虑了不包含某特征的文档数对分类的贡献,但是实验表明^[2],考虑特征不出现的情况对于分类的干扰比贡献大。

1.1.3 CHI统计

CHI统计方法使用特征 t 出现与不出现的文档频数来衡

收稿日期:2009-04-22。

作者简介:卢扬竹(1985-),女,四川蓬溪人,硕士研究生,主要研究方向:计算机网络;张新有(1971-),男,河南人,副教授,博士研究生,主要研究方向:计算机网络、网络计算;祁玉(1975-),男,安徽颍上人,硕士,主要研究方向:计算机网络。

量特征 t 与类别 C 之间的相关程度^[1], 设 a, b, c, d 的含义如表 1。

用 CHI 统计计算特征 t 与类别 C 的关联程度即 χ^2 值的公式如式(2):

$$\chi^2(t, C) = \frac{N \times (ad - bc)^2}{(a + c)(b + d)(a + b)(c + d)} \quad (2)$$

其中 N 为文档集中的文档总数。

特征 t 对于整个文档集的 CHI 值由式(3) 计算:

$$\chi_{\max}^2(t) = \max_{i=1}^m \chi^2(t, C_i) \quad (3)$$

同样, 预先设定一个阈值, 将 CHI 值超过此阈值的特征保留下来, 得到特征子集。

表 1 文档频数

文档数	特征 t	特征 \bar{t}
属于 C 类的文档数	a	c
不属于 C 类的文档数	b	d

1.1.4 互信息

信息论中用互信息方法来衡量两个信息之间的相关程度, 在特征选择中, 用互信息量表示特征 t 和类别 C 之间关联的程度^[3]。特征 t 与类别 C 之间的互信息计算公式如式(4):

$$I(t, C) = \log \frac{P(t, C)}{P(t) \times P(C)} \quad (4)$$

其中: $P(t, C)$ 表示特征 t 与类别 C 共现的概率, $P(t)$ 表示特征 t 在整个训练集中出现的文档频率, $P(C)$ 表示类别 C 在训练集中出现的概率。

1.2 邮件过滤中的特征选择算法的特点

在垃圾邮件过滤中, 邮件文本的分类实际上是一个文本二类分类问题: 设一个邮件文本为 d , 经过预处理之后, 文本 d 被表示为 n 个特征词(设为 t_1, t_2, \dots, t_n) 及其词频。邮件过滤即是判定 d 属于某类文本 $C_k (k = 1, 2)$, 其中 C_1 代表垃圾邮件类别, C_2 代表正常邮件类别。

在文本分类中, 特征词的频度、集中度以及分散度是对分类效果影响最大的。其中, 频度是指特征词在邮件训练集的某一类的所有文档中出现的次数, 出现次数越多, 其与此类别越相关, 越能代表此类别。集中度是指含有此特征词的类别数, 含有此特征词的类别越少, 越说明此特征词集中在这几类或一类中出现, 越能代表这些或这一类别。而分散度是指在某类中含有此特征词的文档数, 文档数越多, 说明该特征词在此类中越均匀、越分散出现, 那么此特征词越能代表此类。

本文在邮件过滤中引入这三大指标对特征算法的改进。其中, 频度用词频数(Term Frequency, TF) 来表征, 集中度用 2.2 节中的类别贡献比来表征, 而分散度用文档频数 DF 来表征。

2 互信息算法研究及改进

2.1 互信息算法用于邮件过滤的分析

如上一节所述, 在文本分类领域中, 互信息用来度量特征词和类别之间相互关联的程度。由条件概率公式, 互信息公式可以得出如式(5) 的推导:

$$I(t, C) = \log \frac{P(t, C)}{P(t) \times P(C)} = \log \frac{P(t|C) \times P(C)}{P(t) \times P(C)} = \log \frac{P(t|C)}{P(t)} \quad (5)$$

在垃圾邮件分类中, 设 N 为邮件训练集中的文本总数, C_s 代表垃圾邮件类 spam, C_l 代表合法邮件类 legit。 $DF(t, C_s)$ 、 $DF(t, C_l)$ 、 $DF(\bar{t}, C_s)$ 以及 $DF(\bar{t}, C_l)$ 代表的意义如表 2 所示。

表 2 邮件过滤中的 DF

文档数	特征词 t	特征词 \bar{t}
属于垃圾邮件类 C_s 的文档数	$DF(t, C_s)$	$DF(\bar{t}, C_s)$
属于正常邮件类 C_l 的文档数	$DF(t, C_l)$	$DF(\bar{t}, C_l)$

那么, 特征 t 与垃圾邮件类 C_s 之间的互信息可表示为:

$$I(t, C_s) = \log \frac{P(t|C_s)}{P(t)} = \log \frac{\frac{DF(t, C_s)}{N_s}}{\frac{N_t}{N_{\text{total}}}} = \log \frac{DF(t, C_s) \times N_{\text{total}}}{N_s \times N_t} \quad (6)$$

其中: N_s 为垃圾邮件总数, N_t 为包含特征 t 的邮件总数, N_{total} 为训练集中所有邮件总数。用文档频数来表示, 设 $A = DF(t, C_s)$, $B = DF(t, C_l)$, $C = DF(\bar{t}, C_s)$, $D = DF(\bar{t}, C_l)$, 即:

$$I(t, C_{\text{spam}}) = \log \frac{A \times N_{\text{total}}}{(A + C)(A + B)} \quad (7)$$

t 与正常邮件类 C_l 之间的互信息 $I(t, C_l)$ 可以同理求出。

在分类中使用互信息算法进行特征选择, 即计算特征 t 的互信息量, 一般有两种计算方式, 即最大互信息:

$$MI_{\max}(t) = \max_{i=1}^m I(t, C_i); \quad (8)$$

$i = 1, 2, \dots, m$ (在垃圾邮件分类中, m 为 2)

以及平均互信息:

$$MI_{\text{avg}}(t) = \sum_{i=1}^m P(C_i) \times I(t, C_i); \quad (9)$$

$i = 1, 2, \dots, m$ (在垃圾邮件分类中, m 为 2)

通过以上研究, 可以分析得出互信息算法的几个缺陷。

1) 在特征词对分类效果影响的指标中, 互信息算法仅考虑了分散度即文档频数 DF, 并没有考虑到词条频数 TF 的影响。此算法只考虑词条出现和不出现, 而不考虑词条在文档中出现的次数。即对于文档频数相同的特征, 其在一个文档中出现一次和出现很多次, 对计算出来的互信息量没有影响。而经验认为, 出现次数越多(即词频大) 的特征词与类别的相关程度更大, 更能代表此类别。

2) 最大互信息以及平均互信息这两种计算方法都无法很好体现出特征词对各类别的贡献差异。如特征 t 与 C_1 和 C_2 的互信息量, 即 $I(t, C_1)$ 和 $I(t, C_2)$, 都较大但是相差无几, 那么 t 会被选入特征集, 但因为它与两个类别的关联程度相差不大, 即集中度不大, 所以实际上这个特征词对邮件分类的作用很有限。

3) 从式(5) 可以看出, 当 $P(t|C_1)$ 比 $P(t)$ 小即 $\frac{P(t|C_1)}{P(t)} < 1$ 时, 互信息量 $I(t, C_1)$ 将是一个负数。如果这个负数很大, 表明特征 t 与垃圾邮件类 C_1 的关联很小, 但是这也说明 t 与正常邮件类 C_2 的关联很大, 这种情况称作“负相关”^[3]。所以在实际特征选择的计算过程中, 互信息为负数的特征, 有可能会比互信息量为正数的特征具有更大的分类贡献, 应该在互信息公式中针对负相关情况进行处理。

2.2 对互信息算法的改进

根据以上分析, 针对垃圾邮件过滤领域中的特征选择, 兼顾

频度、分散度以及集中度,本文提出一种改进的互信息算法。

首先,在互信息算法中考虑频度的指标,引入特征词 t_i 在类别 C_j 中出现的词频 $TF(t_i, C_j)$,这里用 P_{df} 表示:

$$P_{df}(t_i | C_j) = \frac{1 + TF(t_i, C_j)}{\sum_{i=1}^n TF(t_i, C_j)}; \quad (10)$$

$i = 1, 2 \dots, n; j = 1, 2 \dots, m$

其中 $TF(t_i, C_j)$ 为特征词 t_i 在类别 C_j 的所有文档中出现的次数总和。在邮件分类中, n 为特征个数, m 的取值为 2。考虑词频 $TF(t_i, C_j)$ 为 0 的情况,在分子上作加 1 的处理。

则特征词 t_i 与类别 C_j 的互信息的计算公式成为:

$$MI(t_i, C_j) = P_{df}(t_i | C_j) \times \left| \log \frac{P_{df}(t_i | C_j)}{P_{df}(t_i)} \right|; \quad (11)$$

$i = 1, 2 \dots, n; j = 1, 2$

上一节提到的负相关问题在此公式中用绝对值进行处理^[3]。式中 P_{df} 指词频概率, P_{df} 指文档概率,计算方法如式(12):

$$P_{df}(t_i | C_j) = \frac{DF(t_i, C_j)}{N_{C_j}}, P_{df}(t_i) = \frac{N_t}{N_{total}}; \quad (12)$$

$i = 1, 2 \dots, n; j = 1, 2$

其次,对于集中度问题,本文提出采用“类别贡献比” $r(t_i)$,来衡量特征词对于各个类别的贡献差异。综合以上研究与分析,改进的互信息特征选择算法计算如式(13):

$$r(t_i) = \begin{cases} \frac{MI(t_i, C_s)}{MI(t_i, C_l)}, & MI(t_i, C_s) > MI(t_i, C_l) \text{ 且皆不为 } 0 \\ \frac{MI(t_i, C_l)}{MI(t_i, C_s)}, & MI(t_i, C_s) < MI(t_i, C_l) \text{ 且皆不为 } 0 \\ \lambda, & MI(t_i, C_s) = 0 \text{ 或 } MI(t_i, C_l) = 0 \end{cases} \quad (13)$$

其中当 $MI(t_i, C_s) > MI(t_i, C_l)$ 时, $r(t_i)$ 的计算如式(14):

$$r(t_i) = \frac{MI(t_i, C_s)}{MI(t_i, C_l)} = \frac{P_{df}(t_i | C_s) \times \left| \log \frac{P_{df}(t_i | C_s)}{P_{df}(t_i)} \right|}{P_{df}(t_i | C_l) \times \left| \log \frac{P_{df}(t_i | C_l)}{P_{df}(t_i)} \right|}; \quad (14)$$

$i = 1, 2 \dots, n$

其中: $MI(t_i, C_j)$ 表示特征 t_i 对类别 C_j 的贡献, $r(t_i)$ 为特征 t_i 对垃圾邮件类 C_s 和合法邮件类 C_l 的类别贡献比。当 $MI(t_i, C_s) < MI(t_i, C_l)$ 时,计算公式同理可得。特别地,当 $MI(t_i, C_l)$ 为 0 时,说明特征 t_i 与类别 C_l 相互独立,而只与类别 C_s 有关,则将 $r(t_i)$ 设为一个较大的常数 λ , 这里 λ 取 1 000。当 $MI(t_i, C_s)$ 为 0 时,同理, $r(t_i)$ 也设为 1 000。计算出所有特征的类别贡献比并排序之后,在其中从大到小筛选出一定数目的特征,或筛选出大于预先设定的阈值的所有特征,即得到特征子集。

本文的互信息计算公式对原互信息算法进行了如下两方面的改进。

1) 引入了频度的指标。使得互信息计算公式不再是只考虑特征的文档频数,而是将特征在文档中的出现次数对特征选择的影响考虑了进来,更合理地选择与类别最有关联的特征。

2) 加入了集中度的影响因素。这种特征选择方法考虑到了邮件过滤领域中的特殊情况:只进行二类分类。针对上一节提到的,“特征词与两个分类的关联程度都很大但是相差无几”,即互信息量较大但分类作用有限的情况,选择出

“类别贡献比”大的特征词,而不仅仅是与某一类的互信息量大,也就是在“最能代表类别”的基础上,选择出“最能区分类别”的特征。

2.3 算法描述

算法 1

输入:原始特征集合 F , 类别贡献比阈值 δ (或特征子集维数 d)

输出:特征子集 G

统计 DF_s, DF_l ;

计算 DF_{total} ;

for $i = 1$ to n

统计 $DF(t_i, C_s), DF(t_i, C_l)$;

计算 $DF(t_i)$;

计算 $P_{df}(t_i | C_s), P_{df}(t_i | C_l)$;

统计 $TF(t_i, C_s), TF(t_i, C_l)$;

计算 $MI(t_i, C_s), MI(t_i, C_l)$;

根据公式计算 $r(t_i)$;

$i++$;

选出 $r(t_i)$ 大于阈值 δ 的所有特征 (或按 $r(t_i)$ 从大到小选出 d 个特征) 作为特征子集;

3 实验结果

1) 实验内容:为了验证改进的互信息算法的分类效果,分别采用原互信息算法和改进互信息算法进行特征选择,再进行分类实验,最后将所得的结果进行对比。

2) 实验数据:采用 CCERT(中国教育和科研计算机网紧急响应组)提供的电子邮件数据集,该数据集包括一个垃圾邮件集和一个合法邮件集,是 CCERT 于 2005 年 8 月发布的,其内容均来源于真实邮件。此数据集可以在 CCERT 的网站: <http://www.ccert.edu.cn/spam/sa/datasets.htm>^[4] 下载得到。

3) 实验工具:对邮件文本的分词及预处理采用中科院的汉语词法分析系统 ICTCLAS(Institute of Computing Technology Chinese Lexical Analysis System)^[5],特征选择实验采用新西兰怀卡托大学开发的数据挖掘软件 WEKA(怀卡托智能分析环境, Waikato Environment for Knowledge Analysis)^[6],分类算法采用 WEKA 中的 Naive Bayes 分类算法。

4) 实验方法:采用十次交叉验证,对所得结果取平均值。

5) 评价指标:采用 F1 值,其计算方法如式(15):

$$F1 = \frac{2PR}{P + R} \quad (15)$$

其中: P 为准确率(Precision)即垃圾邮件检出率: $P = TP / (TP + FP)$; R 为查全率(Recall)即垃圾邮件检出率: $R = TP / (TP + FN) = TP / N_s$ 。其中 N_s 为垃圾邮件总数, TP 、 FP 、 FN 及 TN 代表的意义如表 3 所示。

表 3 参数意义表

参数意义	实际为垃圾邮件	实际为合法邮件
系统判定为垃圾邮件(Positive)	TP	FP
系统判定为合法邮件(Negative)	FN	TN

在数据集中随机选取 600 封垃圾邮件和 600 封合法邮件,进行分词、去停用词、去除单字、统计词频并删除低频词等预处理之后,将统计结果转换成 weka 需要的 aff 文件,其中的一列即一个属性,代表一个特征词;其中的一行即一个实例,代表一个邮件文本。将此结果用 weka 进行训练和测试,其分类方案参数为:weka.classifiers.bayes.NaiveBayes,特征选

择的维数分别为 500 维、1 000 维、1 500 维、2 000 维和 2 500 维。实验结果的精确率 precision、查全率 recall 如表 4 所示, $F1$ 值如表 5 所示。

表 4 分类实验 Precision 值及 Recall 值

特征维数	MI		本文改进的 MI	
	准确率	查全率	准确率	查全率
500	0.38	0.25	0.55	0.4
1 000	0.48	0.53	0.67	0.73
1 500	0.61	0.49	0.71	0.83
2 000	0.62	0.77	0.75	1.00
2 500	0.76	0.65	0.84	0.89

表 5 $F1$ 值

特征维数	MI/%	本文改进的 MI/%
500	30.20	46.30
1 000	50.40	69.90
1 500	54.30	76.50
2 000	68.70	85.70
2 500	70.10	86.40

由表 5 中的 $F1$ 值可得到如图 1 的直观表示。

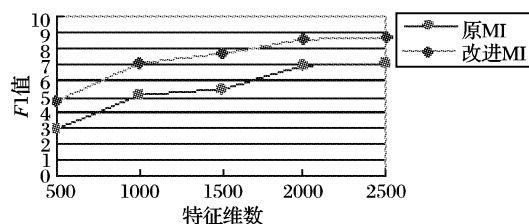


图 1 原 MI 与改进的 MI 算法的 $F1$ 值比较

(上接第 2798 页)

更新速度较快的论坛主题,在 20 次的更新计算中,有 3 次两种方法的计算结果相同,其余 17 次的计算中,主题更新模型均能够使首页受欢迎程度得到提高。

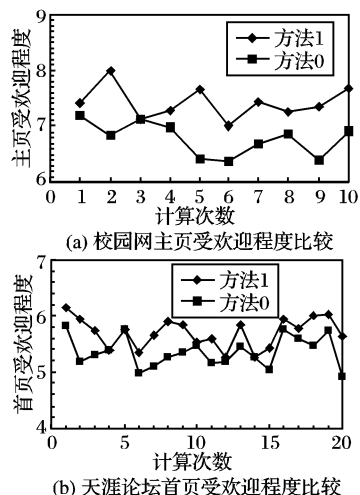


图 2 两种更新方法的受欢迎程度比较

4 结语

不同于个性化推荐,本文给出面向整个网站用户群基于 Web 挖掘技术的多个主题更新策略,通过这种更为合理的更新,力求达到主页中主题有更大浏览点击次数,提高主页的受欢迎程度。

1) 通过对提取实际新闻主题点击次数、用户停留时间或回帖等数据的分析,提出基于主题有效点击次数来建立主题

通过图 1 中的对比可以看出,本文提出的特征选择方法有效提高了分类能力,并且较高的维数比低维数的分类效果更佳。由于本文是在中文邮件数据集上进行实验,鉴于预处理过程中的词典选择等因素,本文实验的绝对结果受到一定影响,但这并不影响验证相对效果。

4 结语

本文对邮件过滤中的特征选择算法进行了研究,对其中的互信息方法结合邮件过滤的特点进行了具体分析,得出了其在频度、集中度以及分散度上对分类效果的影响,并进一步针对这三方面进行了改进,提出了使用类别贡献比 $r(t_i)$ 来衡量特征对分类的贡献,给出了改进后算法的具体计算公式,最后在实际环境中的真实邮件集上进行了实验验证。由于中文分词以及预处理效果比英文分词效果相对较差,本文实验的绝对结果不够理想,中文分词及预处理技术的改进将是以后的研究方向。

参考文献:

- [1] 代六玲,黄河燕,陈肇雄. 中文文本分类中特征抽取方法的比较研究[J]. 中文信息学报, 2004, 18(1): 26 - 32.
- [2] 陈平,刘晓霞,李亚军. 文本分类中改进型互信息特征选择的研究[J]. 微电子学与计算机, 2008, 25(6): 194 - 196.
- [3] 卢新国,林亚平,陈治平. 一种改进的互信息特征选取预处理算法[J]. 湖南大学学报: 自然科学版, 2005, 32(1): 104 - 107.
- [4] CCERT Data Sets of Chinese Emails(CDSCE) [EB/OL]. [2009 - 05 - 01]. <http://www.ccert.edu.cn/spam/sa/datasets.htm>.
- [5] ICTCLAS [EB/OL]. [2009 - 05 - 01]. <http://ictclas.org/>.
- [6] WEKA [EB/OL]. [2009 - 05 - 01]. <http://www.cs.waikato.ac.nz/ml/weka/>.

更新模型。

2) 提出基于主题点击率的主题更新模型,并结合模型指出如何进行单个主题和多主题更新,从而提高主页的受欢迎程度。

本方法也可通过参数的调节推广应用到对时新性要求更高的门户网站的主页更新中。本主题更新模型未来的改进方向是如何合理地将主题重要性以及主题关联两个因素融入模型中。

参考文献:

- [1] 涂承胜,鲁明羽,陆玉昌. Web 挖掘研究综述[J]. 计算机工程与应用, 2003, 39(10): 90 - 93.
- [2] HAN J, KAMBER M. Data mining: Concepts and techniques[M]. San Mateo, CA: Morgan Kaufmann, 2000.
- [3] 韩家伟,孟小峰,王静,等. Web 挖掘研究[J]. 计算机研究与发展, 2001, 38(4): 405 - 414.
- [4] SRIVASTAVA J, COOLEY R, DESHPANDE M, et al. Web usage mining: Discovery and application of usage patterns from Web data[J]. SIGKDD Explorations, 2000, 1(2): 12 - 23.
- [5] 邵华,高风荣,邢春晓,等. 基于 VSM 的分层网页推荐算法[J]. 计算机科学, 2006, 33(11): 86 - 88.
- [6] 谭艳霞,徐珂. 基于大型网站的搜索引擎网页更新方法研究[J]. 微计算机信息, 2005, 21(11-3): 125 - 127.
- [7] 吕韩飞,王申康. 一种重要性与时新性结合的网页更新策略[J]. 计算机应用研究, 2005, 22(11): 212 - 214.
- [8] DOU ZHI-CHENG, SONG RUI-HUA, WEN JI-RONG, et al. Evaluating the effectiveness of personalized Web search[J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(8): 1178 - 1190.