计算机研究与发展
Journal of Computer Research and Development
ISSN 1000-1239/CN 11-1777/TP
42(1): 79~84, 2005

一种改进的自适应文本信息过滤模型

马 亮 陈群秀 蔡莲红
(清华大学计算机科学与技术系智能技术与系统国家重点实验室 北京 100084)
(maliang00@mails.tsinghua.edu.cn)

An Improved Model for Adaptive Text Information Filtering

Ma Liang, Chen Qunxiu, and Cai Lianhong
(*State Key Laboratory of Intelligent Technology and Systems, Department of Computer Science & Technology, Tsinghua University, Beijing* 100084)

Abstract The information filtering technology is usually used to track favorite topics and eliminate garbage content from information stream. The adaptive information filtering, which requires little initial training resource and can actively improve itself in filtering process, provides a better performance and convenience than the old way. But there are still some difficulties in training and adaptive learning. In this paper, an improved filtering model for adaptive text filtering is proposed. In this model, two retrieval/feedback mechanisms are used respectively. One is based on vector space model and Rocchio feedback algorithm, and another mechanism is derived from a latest language model IR system. Based on them, an incremental learning method using multi-step pseudo feedback is introduced in profile training to keep a minimal bias to the original topic, and an adaptive profile adjusting mechanism in filtering process, which newly takes into account the document distribution and the decay rate of the topic feature, is also developed. The running system constructed using the new model got a high evaluation score in related international contest, indicating that the improvements in the filtering model are effective.

Key words information retrieval; Web; adaptive information filtering; language model; relevance feedback

摘 要 自适应信息过滤技术能够帮助用户从Web等信息海洋中获得感兴趣的内容或过滤无关垃圾信息.针对现有自适应过滤系统的不足,提出了一种改进的自适应文本信息过滤模型.模型中提供了两种相关性检索机制,在此基础上改进了反馈算法,并采用了增量训练的思想,对过滤中的自适应学习机制也提出了新的算法.基于本模型的系统在相关领域的国际评测中取得良好成绩.试验数据说明各项改进是有效的,新模型具有更高的性能.

关键词 信息检索; Web; 自适应信息过滤; Language Model; 相关性反馈

中图法分类号 TP391

1 引 言

互联网信息的迅速膨胀使Web用户难以从Internet上及时不断地获得自己感兴趣的信息.同时,大量Web垃圾信息也给Web使用和管理带来诸多麻烦.信息过滤[1]技术作为上述问题的有效解决方法,①可以向用户主动提供个人兴趣相关的Web信息;②过滤敏感性信息(如国家安全、暴力、色情和反动信息等).相比传统的批过滤技术,新的自适应过滤技术不需要大量初始训练文本,同时在过滤过程中可不断进行自主学习来提高过滤精度,因此更适合于Web环境的过滤要求.

在自适应过滤中,过滤的用户兴趣模型通常定

收稿日期:2003-07-03;**修回日期**:2004-05-18
基金项目:国家"八六三"高技术研究发展计划基金项目(2001AA14040)

义为 Profile. 一个典型的 Profile 包括定义主题的特征向量(包含若干主题特征和相应权重)和对应的阈值(用于进行相关性判定)两部分. 自适应过滤的处理过程包括两个步骤:

(1) 训练兴趣模型. 根据初始训练文本和附加训练集训练得到一个初始 Profile.

(2) 过滤测试及模型的自适应学习. 由 Profile 按顺序过滤文档流的每个文档. 特征向量计算文档的相关度,并由阈值确定是否与主题相关,是则作为正例结果输出,否则抛弃. 同时,Profile 不断通过已过滤文档的反馈信息进行自适应学习,以提高兴趣模型的准确度,相应提高后续过滤的精度.

Profile 定义的准确性对过滤结果有着直接的影响,有关 Profile 的训练和自适应学习一直是该领域研究的重点. 同时,相关性检索模型(计算特征权重和文档相关度)和反馈算法(自适应学习机制的基础)对于提高过滤结果的精度也很重要.

在本文中提出了一种改进的自适应文本过滤模型. 在后文面的各节中,首先分析了现有模型的不足;然后介绍了新模型中的相关改进. 最后通过相关实验,对数据进行了讨论并给出了结论.

2 现有的研究工作

在 Web 应用环境下,对原有的自适应信息过滤模型提出了一些新的要求:

(1) 由于用户兴趣的多样性和随机性,一个主题通常只能提供很少的初始正例用于训练;

(2) Profile 自适应学习中只用正例文档进行反馈(负例数量太多,且用户无法选择有效负例);

(3) 较早的正例文档往往不会被用户保留,因此只有近期过滤得到的正例文档可用于反馈.

现有模型中常使用向量空间模型[2]和概率模型[3]进行相关性检索,采用统计性信息(如 TF * IDF 或概率等)计算特征权重. 同时多采用标准 Rocchio 算法[4]进行反馈学习.

由于 Profile 训练中初始训练文档很少,只能提供有限的主题特征,因此会通过附加训练集扩展得到足够的特征. 常见的方法[5,6]是利用初始训练文档构造一个原始 Profile 特征向量,然后通过特征向量在训练集中选择较多数目(通常几十个)的伪正例文档,并利用它们做单次反馈扩展主题特征,并得到初始 Profile.

在过滤测试阶段,Profile 的自适应学习包括主题特征的学习和阈值的调整. 前者主要取决于选择的反馈算法(多使用 Rocchio 算法). 而阈值调整的思路是力求使已过滤结果达到某种评价标准(如 FBeta)的预期值. 如文献[7]采用基于过滤精度的最大期望值(expectation maximum)的调整机制.

在现有研究中,主要关注于 Profile 自适应学习机制[8,9],因为普遍认为其有助于提高最终的过滤精度. 实际上,Profile 训练对于过滤结果同样重要,自适应学习中使用的正反馈文档均基于初始 Profile 的判定. 如果初始 Profile 主题准确度较低,根本不可能得到好的过滤结果. 而当前简单的单次反馈机制很难保证 Profile 训练后不偏离原有主题.

同时,Profile 训练和自适应学习中使用的反馈算法[10]中,新特征的选择主要依靠候选特征的统计性权重信息(如 TF * IDF、概率等). 这些信息来自对反馈文档或整个训练集的统计,但实际由于反馈文档数目很少或训练集的均匀分布特性,这些统计性的数据并不能准确表示其与主题的相关度,导致许多噪音也被选择加入到 Profile.

针对现有研究存在的不足,我们分别提出了对应的改进方法,下面将具体介绍.

3 新的模型框架

3.1 相关性检索模型与反馈算法

系统采用两种独立的相关性检索模型:Vector Space Model(VSM)和 Language Model(LM). 前者已被广泛应用于信息检索领域,而 LM 作为新检索模型已在一些 Web 检索应用中有良好的表现. 这里引入该模型以测试其在自适应过滤中的性能,同时也使系统具有更大的适用性和灵活性.

VSM 中特征权重采用 TF * IDF(Okapi TF 公式[3])计算,并采用了改进的 Rocchio 算法(见第 3.2 节)用于相关性反馈. 对于 LM 则采用 SimpleKL 模型[11],并使用 Mixture[12]反馈算法.

3.2 改进的 Rocchio 反馈算法

由上述可知,在标准 Rocchio 反馈中,当反馈文档较少时,使用统计性权重选择特征容易引入噪音,因此,引入语言学属性来辅助提高特征选择的精度.

当反馈文档很少时,在文档中特征的语言属性比其统计性属性更为准确和稳定. 研究[13]表明,句法属性中的词性(part of speech)和句法功能最为有用. 通过对特征的词性和句法功能进行统计,我们发现:①大多数特征为名词、形容词和动词;②名词

(包括名词短语和专有名词)性特征比例最大;③主语和定语从句中的形容词、谓语动词也常作为特征.

利用上述规律,对每篇反馈文档进行句法分析,并从句法树中按特定句法属性抽取候选特征.这样每个文档就对应生成一个由候选特征组成的新文档.将所有反馈文档的对应新文档组成一个文档集 D_t,原文档集的大量噪音被过滤.同时,设 $W_{avg}(P)$ 是 Profile 特征向量中所有特征的权重平均值;设 $D_f(t)$ 是候选特征 t 在 D_t 中的文档频率(document frequency,DF),则 t 的权重 $W(t)$ 为

$$W(t) = W_{avg}(P) \times \sqrt{D_f(t)}. \quad (1)$$

权重计算公式中仅考虑了 DF 并不是常规的 TF * IDF 机制,是因为大量噪音已经被过滤,而且文档较少时 TF * IDF 并不具有更准确的定义能力.

得到 D_t 和新权重后,按标准 Rocchio 算法选择特征.设 N_p 为 D_t 中的文档数,则新特征数 K_m 为

$$K_m = 5 + 5 \times \lg(N_p + 1). \quad (2)$$

4　Profile 训练

如上所述,现有训练机制依靠大量伪正例进行单次反馈,由于原始特征向量中主题特征很有限,导致选择的伪正例精度波动较大,容易造成初始 Profile 与主题的偏差.对此我们提出了基于改进的 Rocchio 反馈算法(LM 中未使用)和多步反馈方式的增量训练机制.

设训练集为 S_t,定义一个正例集 U_t.增量训练的基本步骤如下:

(1) 由初始训练文档提取原始主题特征生成原始 Profile 特征向量,将初始训练文档加入 U_t;

(2) 用当前 Profile 计算 S_t 所有文档的相关性;

(3) 根据相关性评价结果选择新的伪正例文档;

(4) 利用新的伪正例文档对 Profile 特征向量进行反馈.然后将这些文档加入 U_t;

(5) 如达到终结条件则结束,否则返回(2).

步骤 3 中定义了两种伪正例选择机制.设文档 $d(d \in S_t - U_t)$ 的相关性评价为 $s(d)$,如 d 满足下列任一标准,则被认为是伪正例文档:

① 固定值 m:如 d 是 $S_t - U_t$ 中相关性评价结果最高的 m 个文档之一;

② 自适应阈值 T_h:如 $s(d) > T_h$, T_h 由 U_t 中所有伪正例文档的当前相关性评价结果确定.

当使用固定值 m 时,相应设置反馈次数作为训练终结条件.对于自适应阈值,当得不到新的伪正例时,训练终结(仍需要设置最大反馈次数).

在训练中,每次反馈时只有与当前 Profile 最相似的少量伪正例文档用于反馈,同时改进的反馈算法也保证只有最相关的特征被引入.利用这种增量学习机制,其目标不是尽可能获得更多的相关特征,而是在保证有效特征的前提下尽可能少地引入非相关特征,以避免与原主题的偏差.

在训练结束后,利用 Profile 特征向量重新计算 U_t 中所有伪正例文档的相关度,并选择最小评价值为初始 Profile 阈值.

5　Profile 自适应学习

在过滤测试中,Profile 在选择正例结果输出的同时,也不断利用它们进行自适应学习,以逐步提高自身的准确度.因此,自适应学习机制对提高 Profile 的过滤精度有重要的作用.

5.1　主题特征的学习

主题特征的学习主要用于从正例文档中获得新的主题特征.由于特征选择机制已由反馈算法确定,因此影响学习效果的主要因素是用于反馈的正例文档数和所选择的新特征数目.我们的研究表明,在一次反馈学习中,使用较少量(4～5 个)的正例文档比较有效.

至于选择的新特征数目,在现有的系统中通常为固定值(取决于反馈文档数).但实际上随着不断的学习,Profile 中有效主题特征越来越多,而必需的特征逐渐减少,此时应相应减少引入的特征以避免更多的引入噪音.所以引入特征衰减因子(decay rate)来逐渐减少新特征数目.

设 P 为特征学习中选择加入的特征数目.当第 n 次特征学习时,有 $P = P_0 \times d(n)$.其中,P_0 由式(2)确定;$d(n)$ 为特征衰减因子,定义为

$$d(n) = \alpha + (1.0 - \alpha) \times e^{-n\beta},$$
$$\alpha, \beta \in (0,1). \quad (3)$$

5.2　阈值的自适应调节

Profile 阈值用于从过滤文档中选择正例文档.当满足下面任一条件时,进行阈值调节:

条件 1.当 Profile 进行主题特征学习后;

条件 2.近期过滤得到的正例分布不符合预期.

条件 1 考虑了引入新特征的影响,基于原特征权重的阈值应根据新加入特征的权重进行相应调节;条件 2 则使用了基于分布的调节思想.如果近

期过滤的正例分布超出整体分布范围则升高阈值(以减少正例文档数)或下调阈值(以增加正例文档数). 这里不采用复杂的期望模型,是因为我们认为①训练和自适应学习中的特征学习机制是有效的;②由于主题文档在 Web 文档流中的分布随机性较大,基于已过滤文档的复杂期望模型未必准确,而基于整体分布统计应具有更好的稳定性,假设:

(1) t:过滤文档在文档流中的顺序编号;

(2) $n(t)$:截至文档 t 时过滤的文档总数;

(3) $n_R(t)$:截至文档 t 时过滤得到的正例文档数;

(4) $n_N(t)$:截至文档 t 时过滤得到的负例文档数;

(5) $T(t)$:截至文档 t 时的 Profile 阈值;

(6) $W_s(t)$:截至文档 t 时 Profile 特征权重总和;

(7) $S(t_k, t_{k+1})$:在(t_k, t_{k+1})过滤文档区间所有文档的相关性评价的平均值;

(8) $S_R(t_k, t_{k+1})$:在(t_k, t_{k+1})过滤文档区间所有正例文档的相关性评价的平均值;

(9) $D_R(t_k, t_{k+1})$:自上次阈值调节后得到的正例文档的分布密度.

可计算为

$$D_R(t_k, t_{k+1}) = \frac{n_R(t+1) - n_R(t)}{n(t+1) - n(t)},$$

则阈值调节公式如下:

条件 1:

$$T(t+1) = T(t) + T(t) \times \frac{Ws(t+1) - Ws(t)}{Ws(t)} \times e^{-\gamma}, \quad (4)$$

其中,$\gamma = S_R(t_k, t_{k+1})/S(t_k, t_{k+1})$.

条件 2:

$$T(t+1) = T(t) \times \theta(t+1), \quad (5)$$

其中,$\theta(t+1)$的调节算法如下:

If $D_R(t_k + t_{k+1}) < V_r \times D_{exp}$ Then

/* 降低阈值 */

$\theta(t+1) = A + (1.0 - A) \times S(t_k, t_{k+1})/T(t)$

Else

IF $D_R(t_k + t_{k+1}) > (1.0/V_r) \times D_{exp}$ Then

$\theta(t+1) = 1.0 + B \times S_R(t_k, t_{k+1})/T(t)$

/* 升高阈值 */

这里 D_{exp}为期望的正例分布密度;V_r 用于控制D_{exp}波动的许可范围. D_{exp}来源于对训练集的正例(包括训练得到的伪正例)分布统计(假设训练集和测试集具有近似的正例分布密度).

6 实 验

为了对改进的模型得到可信的评价结果,基于该模型的系统参加了 2002 年 TREC(Text Retrieval Conference)自适应过滤国际评测. 作为信息检索领域最重要的学术会议,TREC 致力于推动海量数据检索的研究工作. 2002 年的 TREC 自适应过滤评测[9]提供了统一的开放测试平台,以评价全世界各研究机构的相关系统(21 个报名机构). 实验数据来自路透社 1986～1987 年的英文新闻语料(reuters news corpus volume 1),所有新闻文档为 Xml 格式,并分为训练集(83650 个文档)和测试集(723141 个有序文档). 测试共使用 100 个不同的 Web 主题,每个主题仅提供 3 篇初始训练正例(来自训练集)和一个简短的任务描述. 由于过滤实际是一个非连续处理的过程,某个时间点只有少量文档被处理,因此可以不考虑时间代价.

6.1 改进的 Rocchio 反馈与 Profile 增量训练

由于反馈算法的效果需要多次反馈才能显现,所以对其评价可通过 Profile 训练的效果来衡量. 而由于 Profile 自适应学习效果的影响,Profile 训练的效果无法直接由过滤结果分析. 考虑到用于反馈的正例文档数量对 Profile 的准确度有直接的影响,所以采用训练结束后 U_t 伪正例文档中的正例文档(不包括初始训练正例)精度来评价训练效果. 设 P_n 为 U_t 中伪正例文档总数,P_r 为 U_t 中正例文档数,则训练精度定义为 $Precision = P_r/P_n$.

第 1 组实验用于测试增量训练机制的性能. 第一组训练(Run-1)使用了常规的单次反馈机制,在后两组结果(Run-2 和 Run-3)中,分别使用了基于固定值 m 和阈值 T_h(设置为 U_t 中所有文档的相关度评价结果的平均值)的增量训练方法. 每组结果均使用标准的 Rocchio 正反馈,相关参数和评价结果(100 个主题的平均值)见表 1.

Table 1 Training Using One-Step Feedback and Incremental Feedback

表 1 单次反馈训练和增量训练机制

Run	Schema	Parameters	P_n	P_r	*Precison*(%)
1	One-Step	$m=15$ Feedback times=1	15.0	4.5	32.73
2	Incremental	$m=3$ Feedback times=5	15.0	4.9	34.67
3	Incremental	Adaptive threshold T_h	12.1	4.3	39.26

第2组实验用于分析改进的Rocchio算法的效果. 根据特征的词性特点,我们只测试名词、形容词和动词(Run-4). 相关词性信息通过句法分析工具MINIPAR[14]得到. 为了避免MINIPAR的精度影响改进算法的效果,我们同时手工进行了句法分析(Run-5). 由于工作量较大,两组结果只随机选择了相同的20个主题进行测试,并采用了与Run-2相同的反馈参数. 同时为了比较,Run-2中相同主题下的评价结果(Run-2A)也一并列出. 相关数据见表2(均为20个所选主题的评价平均值).

Table 2　Improved Rocchio Feedback Algorithm

表2　改进的Rocchio反馈算法

Run	Parsing	Part of Speech	P_r	*Precision*(%)
4	MINIPAR	noun + adj + verb	4.6	33.25
5	By hand	noun + adj + verb	4.7	36.15
2A	none	none	4.6	34.15

6.2　Profile自适应学习机制与整体性能

由于过滤测试的输出即为系统运行结果,因此自适应学习效果结合系统运行效果一并评价.

TREC-2002自适应过滤提供两种评价标准(T11F和T11U)测试系统性能. 参赛系统可按照任一种标准优化性能和测试(我们选择T11F). 对某个主题的过滤输出正例集 FD_t,设 N 为文档总数,R_p 为主题相关文档数,R 为相关文档总数,则

$$\text{T11F} = \begin{cases} 0, & \text{if } R_p = 0, \\ (1.25 \times R_p) \times (N + 0.25 \times R), & \\ & \text{otherwise.} \end{cases}$$

我们按评测要求提交了过滤结果,一组结果(ThuT11af2)使用VSM(采用BM25 TF * IDF)和改进的Rocchio反馈. 另一组结果(ThuT11af3)使用LM(SimpleKL)和Mixture反馈. 两组结果均使用相同的增量训练[15]和自适应学习机制,并按照T11F优化. 表3列出两组结果的TREC官方评价(T11F标准)以及所有参赛结果性能平均值.

Table 3　TREC 2002 Official Evaluation(T11F)

表3　TREC 2002官方评价结果(T11F)

Runs	R101～R150	R151～R200
ThuT11af2	0.422	0.052
ThuT11af3	0.337	0.030
Average score of all participant runs	0.306	0.020

按TREC官方评价[9],在T11F标准下,系统(Tsinghua Group)的最好性能在所有评测系统中位于第4名(箱线图排名). 表4列出了前5名系统的最好结果(评价平均值,非箱线图成绩). 可以看到,前几名的最好结果非常接近.

Table 4　Top-5 Research Teams and Corresponding Score by T11F Evaluation

表4　T11F评价前5名的研究机构与最好测试成绩

Rank	R101～R150	R151～R200
1	0.428	0.062
2	0.426	0.056
3	0.401	0.054
4(Tsinghua University)	0.422	0.052
5	0.421	0.040

7　讨论与结论

从表1看出,增量训练机制比现有的单次反馈训练具有更好的性能,主要是因为减少了无关文档被引入的概率. 而在选择伪正例文档时,自适应阈值比固定值法具有更高的精度和灵活性.

在表2中,虽然Run-4的评价略低于Run-2A,主要归咎于MINIPAR的分析精度的影响,但基于准确的手工句法分析结果(Run-5),表明语义信息在反馈算法的特征选择中是有帮助的.

表3和表4的TREC官方评价表明,文中提出的自适应过滤模型比其他系统具有更优越的性能,同时也证实了模型的相关改进是有效的. 但新引入的Language Model在自适应过滤中未体现出较好的性能,一个重要原因是未对其对应的反馈算法加入辅助的语义信息进行噪音的过滤.

在本文中,针对现有模型的不足,我们提出了一个改进的自适应文本信息过滤模型. 实验表明,改进的机制使模型性能达到了较高水平. 后续将深入研究有效的句法属性和提高Language Model的应用性能.

参　考　文　献

1　Ricardo Baeza-Yates, Berthier Ribeiro-Neto. Modern Information Retrieval. Reading, MA: Addison-Wesley, 1999

2　Chris Buckley, *et al*.. The smart/empire TIPSTER IR system. In: Proc. of TIPSTER Phase 3 Workshop. San Francisco: Morgan Kaufmann Publishers, 1999. 107～121

3　Stephen Roberson, S. Walker. Okapi/keenbow at TREC-8. The 8th Text Retrieval Conf., Gathersburg, USA, 1999. http://trec.nist.gov/pubs/trec8/t8_proceedings.html

4 J. Rocchio. Relevance feedback in information retrieval. In: The SMART Retrieval System. Englewood Cliffs, NJ: Prentice-Hall, 1971. 313~323

5 Wu Lide, Huang Xuanjing, *et al*.. Filtering, QA, Web and video tasks. The 10th Text Retrieval Conf., Gathersburg, USA, 2001. http://trec.nist.gov/pubs/trec10/t10_proceedings.html

6 Zhai Chengxiang, Peter Jansen, Norbert Roma, *et al*.. Optimization in CLARIT TREC-8 Adaptive Filtering. The 8th Text Retrieval Conf., Gathersburg, USA, 1999. http://trec.nist.gov/pubs/trec8/t8_proceedings.html

7 Avi Arampatzis. Unbiased S-D threshold optimization, initial query degradation, incrementality, for adaptive filtering. The 10th Text Retrieval Conf., Gathersburg, USA, 2001. http://trec.nist.gov/pubs/trec10/t10_proceedings.html

8 Stephen Robertson, Ian Soboroff. The TREC 2001 filtering track report. The 10th Text Retrieval Conf., Gathersburg, USA, 2001. http://trec.nist.gov/pubs/trec10/t10_proceedings.html

9 Stephen Robertson, Ian Soboroff. The TREC 2002 filtering track report. The 11th Text Retrieval Conf., Gathersburg, USA, 2002. http://trec.nist.gov/pubs/trec11/t11_proceedings.html

10 Robert E. Schapire, Yoram Singer Amit Singhal. Boosting and Rocchio applied to text filtering. In: Proc. of 21st ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: ACM Press, 1998. 215~223

11 J. Lafferty, C. Zhai. Risk minimization and language modeling in information retrieval. In: Proc. of the 24th ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: ACM Press, 2001. 111~119

12 Zhai Chengxiang, John Lafferty. Model-based feedback in the language modeling approach to information retrieval. In: Proc. of the 10th Int'l Conf. on Information and Knowledge Management. New York: ACM Press, 2001. 403~410

13 Claire Cardie, *et al*.. Examining the role of statistical and linguistic knowledge sources in a general knowledge question-answering system. The 6th Applied Natural Language Processing Conf., 2000. http://www.cs.cornell.edu/home/cardie/papers/anlp-2000.ps

14 Lin Dekang. Dependency-based evaluation of MINIPAR. Workshop on the Evaluation of Parsing Systems, 1998. http://www.cs.ualberta.ca/~lindek/papers/granada.ps

15 Ma Liang, Chen Qunxiu, Ma Shaoping, *et al*.. Incremental learning for profile training in adaptive document filtering. The 11th Text Retrieval Conf., Gathersburg, USA, 2002. http://trec.nist.gov/pubs/trec11/t11_proceedings.html

Ma Liang, born in 1975. Currently a Ph. D. student, Research Interests: Web information retrieval and Chinese information processing.

马亮,1975 年生,博士研究生,主要研究方向为信息检索与中文信息处理.

Chen Qunxiu, born in 1947. Associate professor. Research Interests: Chinese information processing.

陈群秀,1947 年生,副教授,主要研究方向为中文信息处理.

Cai Lianhong, born in 1945. Professor and Ph. D. supervisor. Research Interests: Chinese speech processing.

蔡莲红,1945 年生,教授,博士生导师,主要研究方向为中文语音处理.

Research Background

In this paper, we introduce a new model for adaptive text filtering. Compared with the traditional batch filtering, adaptive information filtering needs little training resource and can improve its precision by adaptive learning in the filtering. These features make it the best choice for increasing instant filtering requirements now. Unfortunately, there are still some problems unsolved in this field, which set a limit to the application of this technology. We proposed some new ideas, including incremental learning and effective adaptive learning mechanism, in order to improve the existing mechanisms in training and adaptive learning. The performance of the new filtering model is demonstrated in related TREC evaluation. Now the filtering model can be used by Web users to track their favorite information in the Internet, and provide some personalized instant information filtered from Web. The paper is sponsored by 863 Hi-Tech Research and Development Program of China.