ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.19, No.11, November 2008, pp.2887−2898
DOI: 10.3724/SP.J.1001.2008.02887

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

信息过滤中基于二元近似关系分布的噪声屏蔽算法*

洪 宇+, 张 宇, 郑 伟, 刘 挺, 李 生

(哈尔滨工业大学 计算机科学与技术学院 信息检索研究室,黑龙江 哈尔滨 150001)

Algorithm of Shielding Noises in Information Filtering Based on Distribution of Two-Dimension Similarity Relation

HONG Yu+, ZHANG Yu, ZHENG Wei, LIU Ting, LI Sheng

(Information Retrieval Lab., School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

+ Corresponding author: E-mail: hy@ir.hit.edu.cn

Hong Y, Zhang Y, Zheng W, Liu T, Li S. Algorithm of shielding noises in information filtering based on distribution of two-dimension similarity relation. *Journal of Software*, 2008,19(11):2887−2898. http://www.jos.org.cn/1000-9825/19/2887.htm

Abstract: This paper investigates the reasons for generation of noises in feedbacks of filtering system by analyzing the knowledge expressions and text structures, and builds a two-dimension similarity (DTS) model of information based on the opposed relation between the noises and user profiles. At the same time, by using the algorithm of AdaBoost based on the LMS (least mean square) classifier, this paper builds a classification curve between the noises and related information according to their distribution in two-dimension similarity space, which helps information filtering system detect and filter the noises in feedback. Experiments validated this algorithm substantially improved the capability of filtering system to rule out the noises.

Key words: information filtering; noise; profile; two-dimension similar relation

摘 要: 针对信息过滤反馈中充斥噪声的缺陷,提出一种基于二元近似关系分布(distribution of two-dimension similarity,简称 DTS)的过滤策略.DTS 根据噪声和用户模型的相悖关系,为信息流建立二元近似关系模型.同时,根据信息在二维近似关系空间中的分布,采用基于 LMS(least mean square)分类器的 AdaBoost 算法建立噪声和相关信息的分类曲线,从而辅助信息过滤系统识别和屏蔽反馈中的噪声.通过实验验证,该算法显著提高了过滤系统屏蔽噪声的能力.

关键词: 信息过滤;噪声;用户模型;二元近似关系

中图法分类号: TP301 **文献标识码**: A

信息过滤是一项从动态信息流中自动获取相关信息的技术[1,2].它在知识获取领域具有重要的应用意义,如垃圾邮件过滤和个性化信息推荐系统等.信息过滤往往需要为用户预先建立相对固定的需求模型(profile),并针对信息流中的每项信息进行实时的相关性判断,在此基础上向用户反馈相关信息和屏蔽不相关信息.与此不同,

* Supported by the National Natural Science Foundation of China under Grant Nos.60435020, 60503072, 60736044 (国家自然科学基金); the National High-Tech Research and Development Plan of China under Grant No.2006AA01Z145 (国家高技术研究发展计划(863))
Received 2007-10-03; Accepted 2008-01-10

信息检索针对用户动态变化的信息需求(query),从静态知识库中查询所有相关信息,并基于相关度排序的方式全部反馈于用户.因此,信息过滤与信息检索如同“硬币的两面”[3],采用不同的信息服务方式实现相同的服务目的,两者获取知识的方法往往不能通用.信息过滤也因此在知识获取领域成为相对独立并富有活力的研究方向.

文本检索会议(Text Retrieval Conference,简称 TREC)将信息过滤方向的研究划分为 3 个子任务:分流(routing)[4]、批过滤(batch filtering)[5]和自适应过滤(adaptive filtering)[6].综合上述任务的特点,过滤系统可划分为如下 4 个部分:用户模型、过滤模型、相关性阈值和自适应学习.用户模型用于描述用户的信息需求;过滤模型用于估计信息与用户模型的相关度;阈值用于设置相关性的判定标准;自适应学习则兼顾用户模型的更新和过滤模型的优化.

针对文本信息的过滤系统主要采用向量空间模型(VSM)对用户需求进行建模,即由词或短语组成高维特征向量,其中特征的权重通过 TFIDF[7],BM25[8]等方法进行估计.在此基础上,早期的过滤模型通过向量的余弦夹角计算相关度;近期的相关研究则将信息过滤解释为二元分类问题[7],采用贝叶斯模型[8]和支持向量机[9]等机器学习方法进行相关性的判别.与此对应,阈值的设置方法也由经验性的估计[1](训练语料中最优过滤性能对应的阈值)逐步深化为相关度分布的概率模型,如基于相关和不相关信息分布的联合概率模型[10]、MLR[11]算法等.自适应学习在用户模型更新中,基于相关或伪相关反馈重估和扩展特征向量,如 Rocchio[12]算法;而在过滤模型优化中,则基于反馈不断迭代地修正过滤模型的参数,如 Okapi[8]系统.

目前,影响信息过滤性能的主要因素是反馈中充斥着大量噪声.噪声是被过滤系统误判为相关的不相关信息.噪声对分流和批过滤任务的直观影响是降低精确率;而在自适应过滤任务中,自学习机制需要根据反馈更新用户模型和优化过滤模型,反馈中的噪声将误导自学习过程并使后续过滤产生偏差.噪声的成因如下:

(1) 用户模型与不相关信息存在共有的特征,这些特征在不同上下文中描述不同的语义;

(2) 用户模型与不相关信息基于共有特征计算得到的相关度高于特定阈值.

现有方法屏蔽噪声的效果并不理想,如BM25 将高频出现于初始用户模型、相关信息和待测信息,同时,低频出现于不相关信息的特征赋予更高权重[8],从而定量地保证了用户模型中的重要特征,但由于自然语言的歧义性**,无法定性保证上述特征正确描述用户需求的语义[7].这一现象在欠缺训练的自适应过滤中尤为明显,原因是用户模型中容易造成歧义的特征无法通过训练语料中的不相关信息进行削弱或屏蔽.不同的是,支持向量机在用户模型中融合多种特征,如文本特征、用户点击[13]以及浏览时间[14]等,通过将用户模型映射到高维空间,并基于训练语料估计分类超平面实施过滤.其优点是用户点击等特征更直观地描述了用户的需求偏好,不仅有利于相关信息与不相关信息的区分,同时也有利于削弱歧义性在相关性估价中的负面作用.但由于支持向量机需要大量的训练,同时对大规模信息的运算复杂度过高,使其无法有效应用于自适应过滤任务.

噪声在阈值估计和自适应学习中往往产生链式的负面效应.如 MLR[11]分别估计相关信息和不相关信息的分布边界,并基于两者边界内的带状区域建立阈值的概率模型,同时随系统反馈不断调整边界和阈值.由于 MLR 忽视反馈中的噪声屏蔽,造成两者的边界趋向于一致,从而逐渐削弱阈值的判别能力,由此产生的更多噪声将进一步影响后续的阈值估计,最终相关信息和不相关信息的分布边界不可分,并使系统持续采用低辨别能力的阈值进行后续过滤.同理,无指导的自适应过滤方法 Rocchio[12]也由于没有屏蔽伪相关反馈中的噪声,造成用户模型的更新不断被误导.与此相比,有指导的自适应过滤系统(如 Okapi[8]系统)可以获得较好的性能,并始终保持自学习机制的可信性,但同时增加了实际应用中用户的负担.

针对上述缺陷,本文提出一种基于二元近似关系分布(distribution of two-dimension similarity,简写为 DTS)屏蔽噪声的过滤算法.DTS 将信息的相关度映射至二维空间,并借助相关信息与噪声固有的相悖关系推动两者

** 歧义性:自然语言处理领域中的“歧义”往往指单一文法的多种语义理解,如“吃了战士的狗”既可理解为“狗吃了战士”,也可理解为“战士的狗被吃了”.本文所指的“歧义”略有区别,指的是基于独立假设的特征向量包含的多种语义组合,如向量{金大中,陈水扁,获得,诺贝尔,奖}既可以组合为“金大中获得诺贝尔奖”,也可组合为“陈水扁获得诺贝奖”,这种语义组合本身并不存在,但往往出现于向量间的相关性匹配,并误导向量间的真实关系.具体分析参见第 1.1 节.

的概率分布趋向可分,从而辅助过滤系统屏蔽反馈中的噪声信息.本文第 1 节分析噪声成因,并探讨 DTS 屏蔽噪声的可行性.第 2 节论述基于 DTS 的噪声识别及屏蔽算法.第 3 节介绍 DTS 在批过滤和自适应过滤中的应用.第 4 节介绍语料、评测及实验流程.第 5 节分析实验结果.第 6 节对全文进行总结.

1 二元近似关系及分布

1.1 噪声成因分析

噪声是被系统误判为相关的不相关信息.Nallapati 将噪声的成因归结为“自然语言固有的歧义性和复杂性,以及不相关信息总会包含相关信息的特征”[7].图 1 演示了噪声的产生过程,其中特征基于 BM25 计算权重,用户模型包含初始用户需求的特征,以及采用 RSV[15]算法从前 30 篇最相关信息中扩展得到的特征.

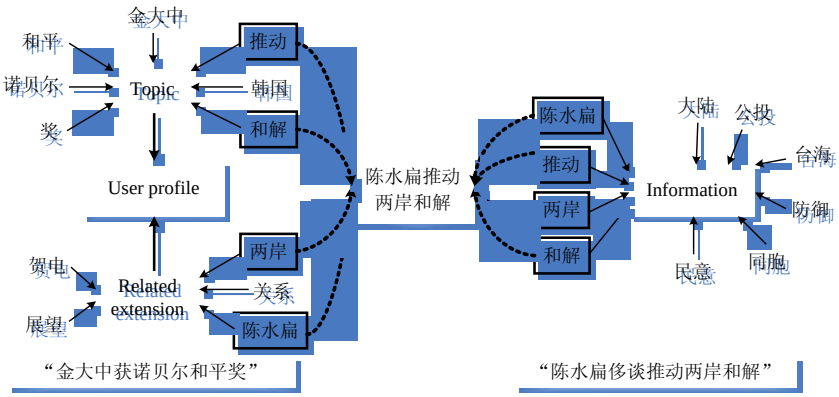


Fig.1 An example about generation of noises

图 1 噪声的生成样例

图 1 中的用户模型可以划分为两个结构,一个是主题“金大中因推动南北韩和解而荣获诺贝尔和平奖”;另一个是围绕主题进行论述的外延“陈水扁向金大中致贺电并借此展望两岸关系”.现有基于词间独立假设的“词包”无法体现这一结构化关系,从而导致用户模型承载多种歧义与信息流进行匹配.如图 1 中主题的特征“和解”和“推动”,与外延中的特征“陈水扁”和“两岸”形成歧义“陈水扁推动两岸和解”.该歧义与题为“陈水扁侈谈推动两岸和解”的不相关信息具备强相关性,从而增大了过滤系统误判的概率.

1.2 二元近似关系

虽然用户模型中的某些特征可以形成歧义并误导后续过滤,但武断地削弱甚至屏蔽这些特征将影响系统的召回率,因为它们也是描述用户需求的重要元素.此外,噪声成因的分析过程也显示,不相关信息的特征没有全部参与匹配,而被遗漏的特征对削弱歧义和区别于用户模型都有重要意义,如图 1 中“公投”和“民意”.

事实上,仅仅观察信息相关于用户模型的绝对指标不足以评价其相关性[8].假设将用户模型类比为“正”极,同时存在相悖于用户模型的“负”极,那么信息与两极的相对指标更为真实地反映信息与用户模型的关系,即越是趋近于“正”极并背离“负”极则越相关,否则越不相关.这一过程无须人为地削弱或屏蔽造成歧义的特征,而是根据信息与两极的相关性趋势自发地削弱歧义,同时该过程可以利用上述遗漏的特征降低匹配过程中的偏见性.基于这一假设,本文提出一种二元近似关系,表示形式如下:

$$R_p(d)=\{r(d,p),r(d,\overline{p})\} \tag{1}$$

其中,p 表示用户模型;d 表示某一文本信息;$R_p(d)$表示 d 与 p 的二元近似关系;r 表示相关度计算的函数;$\overline{p}$ 作为相悖于 p 的概率模型,概括了所有与当前用户模型不相关的信息.二元近似关系将用户模型和信息之间的相关度映射至二维空间,如图 2 所示,其中实心圆代表相关信息,空心圆代表不相关信息.

一维相关度中,大部分不相关信息与用户模型的相关度很低,可通过优化阈值进行屏蔽,而噪声往往与用户

模型的相关度很高,甚至超过某些相关信息,如图 2 中 k 与用户模型 p 的相关度高于 j,该现象使系统难于恰当地设置阈值:若为屏蔽噪声而提高阈值,则使部分相关信息遗失,如图 2 中阈值 θ 遗失相关信息 j;反之,若为获取更多相关信息而降低阈值,则反馈中的噪声也随之增加,如图 2 中阈值 θ' 引入噪声 k 和 l.

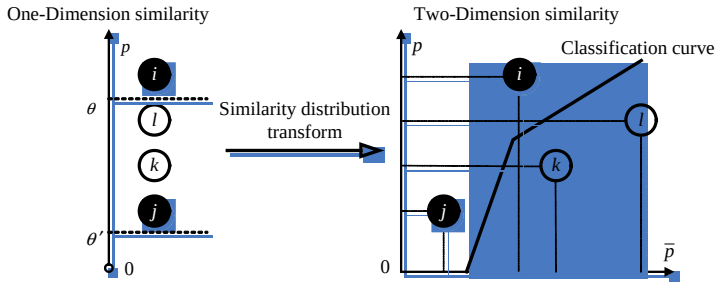


Fig.2 Examples of both one-dimension and two-dimension similarity space

图 2 相关度一维空间和二维空间样例

二维相关度可以描述信息与正、负两极(p 和 $\overline{p}$)的相关性分布趋势,从而为估计噪声和相关信息的分类曲线提供了条件,如图 2 所示.此时,信息与 p 具有高相关度并不等价于两者一定相关,如果该信息与 $\overline{p}$ 具有更高的相关度,则被划分为噪声的概率将增大,如图 2 中的信息 l.相关度的计算往往受篇幅影响,篇幅很短的相关信息往往与用户模型共有的重要特征偏少,导致相关度 r 随之偏低,甚至低于某些噪声的相关度,如图 2 中信息 j,但如果该信息在二维空间中与 $\overline{p}$ 的相关度更低,则仍然可以利用预先训练的分类曲线正确划分.

1.3 二元近似关系分布(DTS)

利用二元近似关系屏蔽噪声的重要步骤是合理选择 $\overline{p}$. $\overline{p}$ 涵盖相悖于用户模型的所有信息,但海量信息流中与用户模型不相关的信息包罗万象, $\overline{p}$ 难以利用有限的特征对其进行统一描述.而如前文所述,大量不相关信息可以通过优化阈值进行屏蔽,影响过滤性能的主要因素是少量相关度很高的噪声.因此,构造 $\overline{p}$ 的信息源可选择特定阈值上的噪声.相应地,二元近似关系则描述信息在用户模型和噪声 $\overline{p}$ 之间的相关度分布.

本节介绍 6 种 $\overline{p}$ 的构造方法,并针对每种 $\overline{p}$ 分析信息在二维相关度空间的分布情况.图 3 中实心圆为相关信息,空心圆为噪声;用户模型 p 描述 2005 年国家高技术研究发展计划(863)评测中的主题“姚明在 NBA 的表现”,横、纵坐标分别是信息与 $\overline{p}$ 和 p 的相关度. $\overline{p}$ 和 p 采用 VSM 进行构造,特征基于 BM25 计算权重,相关度采用向量余弦进行计算, $\overline{p}$ 和 p 的特征数量分别为 100 和 30.此外,图 3 中的所有信息与用户模型 p 的相关度均高于阈值 θ=0.15.

图 3(a)在所有语料中随机选择某一信息构造 $\overline{p}$;图 3(b)则随机选择相关度高于阈值 θ 的某一信息构造 $\overline{p}$. 信息在两图中的分布情况都难以辅助过滤系统建立分类曲线,尤其是图 3(a)中的信息几乎不受 $\overline{p}$ 的影响,这说明虽然语料中大部分信息满足 $\overline{p}$ 的构造条件,即相悖于用户模型,但它们与相关信息和噪声的相关性基本等价且近似为 0,无法描述两者的二元近似关系;图 3(b)虽然形成了信息针对 $\overline{p}$ 的分布趋势,但由于 $\overline{p}$ 采用单一信息进行构造,无法涵盖所有噪声的特征,从而使二元近似关系的可区分性不强.

图 3(c)选择相关度高于阈值 θ 的后 10 项信息构造 $\overline{p}$, $\overline{p}$ 为上述信息的特征向量质心.如图所示,部分噪声挣脱了相关信息的分布区域,更加趋近于 $\overline{p}$.但由于上述构造 $\overline{p}$ 的信息中存在相关信息(高于阈值 θ 的后 10 项信息既有噪声也有相关信息),使得 $\overline{p}$ 与用户模型 p 存在大量共有特征,导致相关信息也趋向于 $\overline{p}$,并使两种分布仍然不可分.图 3(d)延用图 3(c)构造 $\overline{p}$ 的方法,并在此基础上滤除其中出现于 p 的特征,从而使相关信息与噪声的分布区域趋向可分.但图 3(d)中部分噪声存在与 $\overline{p}$ 相关度极低,甚至为 0 的情况,增加了估计分类曲线的复杂度.造成这一现象的原因是 $\overline{p}$ 并没有涵盖所有噪声的特征.图 3(c)中所有噪声与 $\overline{p}$ 的相关度都大于 0,是由于 $\overline{p}$ 包含了部分用户模型 p 的特征,这些特征也是噪声与 p 的相关度高于阈值 θ 的成因,当图 3(d)滤除上述特征后,部分噪声因与 $\overline{p}$ 不存在共有特征而使相关度为 0.

针对图 3(d)的不足,图 3(e)采用所有噪声的质心构造 $\overline{p}$,使相关信息和噪声分布于邻接的两个带状区域内.

在此基础上,图 3(f)滤除 $\overline{p}$ 内出现于 p 的特征,使两种分布近似可分.但是,图 3(e)和图 3(f)要求过滤系统必须预先知道哪些信息是噪声,不适用于无指导的自适应过滤任务.因此,基于所有噪声作为先验知识训练分类曲线的方法只适用于信息路由、批过滤和有指导的自适应过滤任务.

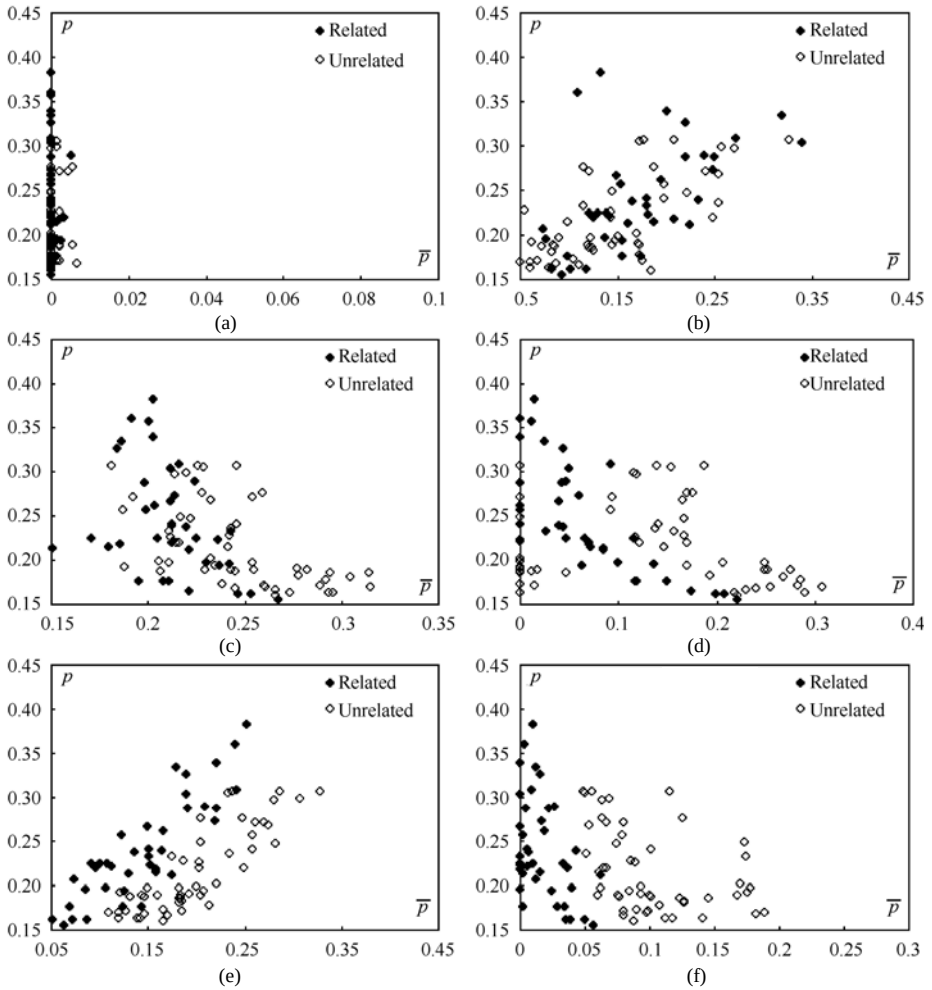


Fig.3 Examples of distributions of two-dimension similarity relations

图 3 二元近似关系分布样例

2 基于二元近似关系的噪声识别及屏蔽算法

基于二元近似关系及其分布(DTS),过滤系统需要选择划分相关信息与噪声的方法,即如何估计两者在二维空间的分类曲线.本文采用线性分类器实现这一划分.线性分类器的核心思想是为不同类别寻找线性判别函数[16],其几何解释为介于不同类别之间的判定面,定义如公式(2).其中,$W=[w_1,...,w_n]$ 为权向量; $X=[x_1,...,x_n]$ 为类别属性;w_0 为偏置.对于 DTS,线性判别函数退化为二维空间的判定线,定义如公式 3.

$$G(x)=W^tX+w_0 \tag{2}$$

$$g(x)=\begin{bmatrix} w_0 \\ w_1 \\ w_2 \end{bmatrix}\cdot[1,x_1,x_2] \tag{3}$$

估计线性判别函数需要定义准则函数 $J(w)$.$J(w)$实际上是当前权向量 W 与最优权向量 W'的距离,当 W达到

最优解时,$J(w)$为最小.因此,线性判别函数的估计转化为求解 $J(w)$极小值的问题,通常可以采用梯度下降法对其进行求解.假设∇J_s 是 $J(w)$收敛的梯度,则梯度下降法通过在训练样本集上的反复迭代,使权向量总是沿着梯度∇J_s最陡的方向移动.因此,相关信息和噪声分类曲线的建立流程如下:(1) 建立由相关信息和噪声组成的训练样本,样本属性为二元近似关系;(2) 选择线性分类器的收敛梯度∇J_s;(3) 基于训练样本反复迭代,求解 $J(w)$的极小化解,将极小化解对应的 W 作为分类曲线的权向量.在此基础上,后续过滤过程将信息的二元近似关系作为分类器的输入,基于分类器输出的判定值来裁决信息是否为噪声并进行屏蔽.

2.1 基于LMS(least mean square)分类器的分类曲线求解

本文选择 LMS[17]分类器来描述分类曲线,LMS 分类器是基于最小均方算法(简写为 LMS)构造的,同时也是梯度下降法的一种改进.其准则函数的下降梯度∇J_s 定义如下:

$$\nabla J_s = \eta X^t(Xa-b) \tag{4}$$

其中, $a=[w^1,...,w^m]^t$; $X=[x^1,...,x^m]$; m 为迭代的次数; $w^i=[w_0^i,w_1^i,...,w_n^i]^t$ 为第 i 次迭代时判定面的权向量.针对 DTS, w^i 描述当前相关信息与噪声分类线的权向量,可表示为 $w^i=[w_0^i,w_1^i,w_2^i]^t$;而 $x^i=[1,x_1^i,...,x_n^i]$ 代表参与第 i 次迭代的样本分布属性,基于二元近似关系的定义,如果训练样本相关于用户模型,则该样本的分布属性 $x^i=[1, r(d,p),r(d,\bar{p})]$,否则 $x^i=[-1,-r(d,p),-r(d,\bar{p})]$,其中 $r(d,p)$为样本 d 与 p 的相关度, $r(d,\bar{p})$ 为 d 与 $\bar{p}$ 的相关度,见第 1.2 节.此外, $\eta=[\eta^1,...,\eta^m]$, $\eta^i=\eta^1/i$, η^1 和 b 为任意正常数.

2.2 DTS的非线性可分问题求解

DTS 往往非线性可分,如图 3 所示,此外,LMS 分类器并不总能获得最优解.因此,本文采用 LMS 线性分类器作为分量分类器,并结合 AdaBoost 算法促进分类曲线对信息分布的拟合.Boosting[18]的目标是提高分类器的准确率,其核心思想是依次设计一组分量分类器 $g(x)$,每个分量分类器的样本集都选择前期各分类器没有正确分类的样本进行构造,并基于该样本集对自身进行训练,最终的分类结果根据所有分量分类器共同决定.

自适应 Boosting 算法[18](简称 AdaBoost)是基于 Boosting 方法的一种变形.该算法不断训练并嵌入新的分量分类器,直到使分类效果达到某个预设的误差率为止.在 AdaBoost 方法中,每个样本都被赋予一个权重 V,表示该样本被后续分量分类器选入训练样本集的概率,其公式如下:

$$V_{k+1}(i)=\frac{V_k(i)}{Z_k}\times\begin{cases}e^{-\alpha(k)}, & g_k(x^i)=y_i\\ e^{\alpha(k)}, & g_k(x^i)\neq y_i\end{cases} \tag{5}$$

其中,$g(x)=y$ 代表样本 x 被正确分类,$g(x)\neq y$ 代表样本 x 被错分;y 为判定系数,二元分类器将 $g(x)>0$ 的情况统一判定为 $g(x)=y=1$,否则 $g(x)=0$;Z_k 为归一化系数; $\alpha(k)$ 为误差系数,其公式如下:

$$\alpha(k)=\frac{1}{2}\ln[(1-E_k)/E_k] \tag{6}$$

其中,E_k 是按照权重 $V_k(i)$采样的前一分量分类器的误差.如果某个样本点被准确分类,则在构造下一分量分类器时,该样本被选入训练集的概率将会减小;反之则会增大.因此,AdaBoost 算法能够始终关注那些很难分类的样本点,并采用它们不断训练后续分量分类器.因此,只要迭代的次数足够大,总体分类器的训练误差就可以达到任意小.最终总体分类器的判断可以使用各个分量分类器的加权平均来获得,其公式如下:

$$h(x)=\left[\sum_{k=1}^{k\max}\alpha_k g_k(x)\right] \tag{7}$$

本文采用 LMS 算法作为分量分类器,同时基于二元近似关系建立样本属性,即 $x=[1,r(d,p),r(d,\bar{p})]$,并将该属性作为 LMS 分类器的输入,在此基础上利用 AdaBoost 算法构造总体分类器.

3 嵌入DTS的过滤系统设计

DTS 属于过滤系统后处理.在过滤过程中,DTS 将系统输出的判定结果作为输入,并在此基础上通过 $\bar{p}$ 建

模、DTS 估计以及训练分类曲线等步骤实现噪声的检测与屏蔽,具体系统框架如图 4 所示.其中,“原型系统”表示尚未嵌入 DTS 屏蔽噪声的现有过滤系统;“非自适应”和“自适应”标明两种不同的 DTS 训练流程.“非自适应”流程基于大量训练语料统计相关信息和噪声的分布趋势,并采用 LMS 及 AdaBoost 算法估计分类曲线,其测试阶段利用该分类曲线检测“原型系统”输出的结果,识别并屏蔽其中的噪声.“自适应”流程则不经过训练语料估计分类曲线,而是基于“原型系统”的输出,实时地构建 $\bar{p}$ 、二元近似关系及分类曲线.

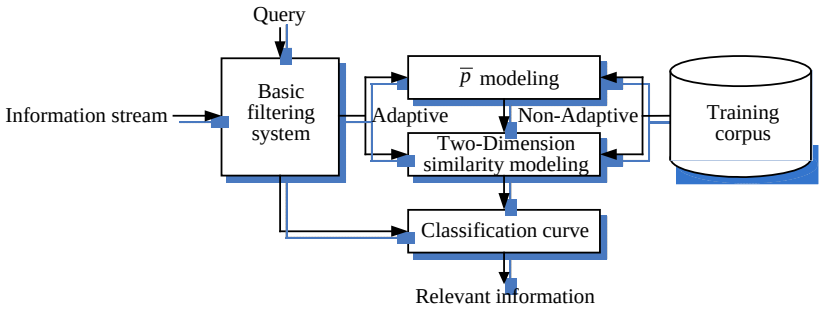


Fig.4 Structure of filtering system shielded from noises by embedding with DTS

图 4 嵌入 DTS 屏蔽噪声的过滤系统框架

3.1 基于二元近似关系的批过滤系统

批过滤系统是一种“非自适应”性过滤系统,它往往基于相对较多的相关信息建立初始用户模型,并可以利用全部训练语料及其人工评价优化用户模型,但在测试阶段,用户模型必须保持固定不变[1].批过滤的 DTS 训练过程如下:首先,计算所有训练样本与用户模型的相关度,并估计“原型系统”的阈值;其次,选择高于阈值的所有不相关样本(噪声),构建相悖于用户模型的 $\bar{p}$,并滤除 $\bar{p}$ 中与用户模型共有的特征,如图 3(f)所示;最后,建立所有训练样本的二元近似关系,并结合先验的相关性评价(相关或噪声)建立 DTS,其中,相关样本的分布属性表示为 $x=[1,r(d,p),r(d,\bar{p})]$,噪声的分布属性表示为 $x=[-1,-r(d,p),-r(d,\bar{p})]$.批过滤系统基于 DTS,采用 LMS 及 AdaBoost 算法估计相关信息与噪声的分类曲线.在测试阶段,系统首先判断信息与用户模型 p 的相关度是否高于阈值,如果低于阈值,则直接屏蔽,否则建立该信息的二元近似关系并代入分类曲线的判别式,判断它是否为噪声,若不是,则作为相关信息输出.

3.2 基于二元近似关系的自适应过滤系统

自适应过滤区别于批过滤,初始用户模型的训练数据较少(2~4 篇),对用户模型的更新完全通过后期的自学习来完成[1].因此,自适应过滤系统嵌入 DTS 的方式遵循图 4 中的“自适应”流程.自适应过滤包括有指导和无指导两种自学习方式,前者假设用户实时地对过滤结果进行评价,并根据用户的相关反馈更新用户模型和优化过滤模型;后者则假设用户较长时期内甚至从不评判过滤结果,仅依靠系统自身的伪相关反馈进行自学习.

有指导自适应过滤的DTS训练过程与批过滤基本类似,区别在于,DTS不是基于大规模先验语料一次训练完成,而是根据用户周期性[***]的反馈逐步训练.针对某一周期内过滤系统的输出,自学习机制检测其中是否存在用户认为不相关的信息,若存在,则触发DTS训练,其使用的语料仅包括当前周期内过滤系统输出的噪声,以及所有已知的相关信息.DTS屏蔽噪声的方法作为一种后处理,不直接影响“原型系统”内部的操作.因此,嵌入自适应过滤系统的DTS必须在周期性的训练中兼顾“原型系统”自身进行的自学习.假设当前“原型系统”已借助自学习更新了用户模型、阈值及过滤模型参数,则DTS训练中的相关度$r(d,p)$(如公式(1)所示),将采用更新后的用户模型和过滤模型进行计算; $\bar{p}$ 也将利用当前训练语料中高于新阈值的噪声进行建模.

无指导自适应过滤系统自学习使用的训练语料是未经人工评价的伪相关反馈,其中哪些信息与用户模型相关,哪些信息为噪声,对自学习机制并不透明.DTS 既无法直接抽取训练语料中的噪声构造 $\bar{p}$,也无法利用已

*** 某些过滤系统不基于固定周期进行自学习,而是每识别出一个相关信息就立刻进行自学习,如参加 TREC-7 的 Okapi 系统.

知的样本类别估计分类曲线.因此,DTS 选择训练语料中与 Profile 相关度最低的 10 个样本构造 $\overline{p}$,并滤除 $\overline{p}$ 中与用户模型共有的特征,如图 3(d)所示.样本类别则基于如下方法进行估计:首先对训练语料进行层次聚类;然后分别计算每个聚类与用户模型 p 和 $\overline{p}$ 的相关度,如果前者高于后者,则该聚类包含的所有样本作为相关信息参与 DTS 训练,否则作为噪声.其他 DTS 训练流程与有指导自适应过滤相同.

4 语料、评测及实验设计

4.1 语 料

实验选择 100G 的天网语料作为训练和测试语料,同时使用 2005 年参加国家高技术研究发展计划(863)评测的检索系统(以下简称 IR863-System)针对 30 个主题分别进行检索.检索结果由 10 名学生进行评测,其中每两名学生为一组,同时对 6 个主题进行人工评测.此后,实验投入 6 名学生对评测结果进行校验,得到关于 29 个主题的 3 937 篇相关文本,其中一个主题没有发现相关文本.

由于本文更关心如何屏蔽最有可能成为噪声的不相关信息,因此,在构造语料的过程中忽略了与用户模型相关度很低的文本.基于这种需要,实验构造了一个规模略小的语料.首先使用 IR863-System 从原始语料中对每个主题进行检索,其反馈的结果按照相关度进行排序;然后选择每个主题相关度靠前的 1 000 篇文档,与人工评测结果取并集构成规模为 29 897 篇文档的语料库;最后选择 50%作为训练,50%作为测试语料,每个主题对应的相关文本均分到训练和测试语料集中.

4.2 评测体系

实验主要采用的评测方法是 T11SU,该方法在 TREC 评测以及 TDT 的研究与实验中被广泛采纳.信息过滤的结果可分为 4 种情况,见表 1.其中 R^+代表系统从信息流中筛选出相关文本的数量;N^+代表系统筛选出不相关文本的数量,即误检指标;R^-代表被系统过滤掉但相关的文本数量,即漏检指标;N^-代表被系统过滤掉且不相关的文本数量;A,B,C,D 分别控制 4 种指标对性能的影响强度.

Table 1 Sorts of results in information filtering

表 1 信息过滤结果类别

	Relevant (by human)	Irrelevant (by human)
Relevant (by system)	R^+/A	N^+/B
Irrelevant (by system)	R^-/C	N^-/D

TREC 为过滤任务定义的评测规范,如公式(8).TREC-10 和 TREC-11 使用的评测方法如公式(9).其中,参数 β 控制评测系统对误检指标影响过滤性能的重视程度,参数 η 用以平滑评测指标.TREC-10 和 TREC-11 将 β 和 η 分别设置为 0.5 和–0.5.此外,该实验还采用精确率和召回率辅助 T11SU 对过滤效果进行评测.

$$Utility = A\cdot R^+ + B\cdot N^+ + C\cdot R^- + D\cdot N^- \tag{8}$$

$$T11SU_{\beta,\eta} = \frac{\max\left(\frac{R^+ - \beta\times N^+}{R^+ + R^-}, \eta\right) - \eta}{1-\eta} \tag{9}$$

4.3 实验流程

实验旨在检验 DTS 屏蔽噪声作为后处理是否能够改进现有过滤系统的性能.在对语料进行分词和去停用词等预处理的基础上,实验建立如下 6 个过滤系统,并进行 3 组对比测试.

System 1:基于 VSM 建立主题的用户模型和信息流的概率模型,并采用 BM25 计算特征权重,用户模型与信息的相关度采用余弦夹角进行计算;初始用户模型利用主题提供的描述信息进行构造,然后从训练语料中选择最相关的前 30 篇相关信息扩展用户模型,其中特征选择采用 RSV[15]算法,用户模型的特征数量为 30;基于 T11SU 训练阈值θ,如果信息的相关度高于 θ,则判定为相关.该系统为参加 TREC-3 的 Okapi 系统[15].

System 2:以 System 1 为“原型系统”,基于 DTS 的“非自适应”训练流程估计相关信息和噪声的分类曲线,如

图 4 所示,训练中 $\bar{p}$ 包含的特征数为 100.在测试阶段,信息首先经“原型系统”判别相关性,如果作为相关信息输出,则再经分类曲线判别是否为噪声,如果判定为真(即噪声)则屏蔽,否则输出为相关信息.

System 3:训练阶段与 System 1 类似,不同点包括相关度计算采用罗杰斯特回归(logistic regression,简称 LR)模型修正刻度;用户模型扩展使用的相关信息只有 2 篇;经验性地降低初始阈值θ.测试阶段采用有指导的自学习机制,系统每识别出一个相关信息就触发自学习,并将已检测过的所有信息作为训练语料.自学习利用牛顿下降法修正 LR 模型的参数,并基于一定梯度提高θ.该系统为参加 TREC-7 的 Okapi 系统[8].

System 4:以 System 3 为“原型系统”,基于 DTS 有指导的“自适应”训练流程实时估计分类曲线,如图 4 所示,$\bar{p}$ 包含的特征数为 100.测试过程与 System 2 基本相同,不同点是“原型系统”每识别出一个相关信息,“原型系统”和 DTS 就立刻先后进行自学习,见第 3.2 节.

System 5:训练阶段与 System 3 相同,测试阶段则采用无指导的自学习机制,每识别出一个相关信息就利用 Rocchio[12]算法对用户模型进行更新.

System 6:以 System 5 为“原型系统”,基于 DTS 无指导的“自适应”训练流程实时估计分类曲线,如图 4 所示,$\bar{p}$ 包含的特征数为 100.测试过程与 System 4 相同.

实验中的 3 组对比测试分别是:(1) System 1 vs. System 2;(2) System 3 vs. System 4;(3) System 5 vs. System 6.它们检验嵌入 DTS 屏蔽噪声能否改进“原型系统”的性能.实验还在(1)中检验阈值对 DTS 屏蔽噪声的影响.

5 实验结果与分析

5.1 批过滤系统:System 1 vs. System 2

该测试中 BM25[15]的参数设置为 k_1=2,k_2=0,k_3=1000,b=0.75;训练中 System 1 在 T11SU 最优时对应的阈值 θ 为 0.14.采用这一阈值,System 1 和 System 2 基于测试语料获得的评测结果见表 2.该结果显示 System 2 的性能显著优于 System 1,其平均精确率提高约 13 个百分点.如第 4.3 节所述,System 2 以 System 1 为“原型系统”,利用预先训练的分类曲线识别并屏蔽 System 1 输出中的噪声.该测试验证了 DTS 屏蔽噪声的有效性.

Table 2 Testing results of System 1 and System 2

表 2 System 1 和 System 2 测试结果

	T11SU	Precision	Recall
System 1	0.4597	0.4351	0.5404
System 2	0.5486	0.5675	0.5216

如表 2 所示,System 2 的召回率略低于 System 1,原因如下:(1) System 2 在 System 1 的基础上嵌入基于 DTS 屏蔽噪声的过程,而这一过程的处理对象是“原型系统”(System 1)的输出,因此 System 2 的召回率必然小于或等于 System 1;(2) 阈值θ 过高,DTS 的训练过程遗漏了部分相关信息,使得分类曲线失真.此外,System 2 虽然改进了 System 1 的精确率,但明显仍有大量噪声未被检测并屏蔽,该现象的成因之一也是训练阶段阈值θ 过高,造成用于构造 $\bar{p}$,DTS 及分类曲线的资源不够充分.因此,“原型系统”训练获得的最优阈值θ 并不一定适用于嵌入 DTS 的系统.为此,实验设置阈值θ 在[0,0.14]范围内以 0.01 为粒度变化,System 2 基于不同阈值重新训练 DTS 及其分类曲线,则 System 1 和 System 2 在测试语料上的精确率变化趋势如图 5 所示,召回率如图 6 所示.

如图 5 和图 6 所示,适当降低训练中 System 2 的阈值(如θ =0.13),有助于进一步改进过滤性能,此时 System 2 的精确率提高至 0.5816,召回率提高至 0.5885.虽然召回率仍然略低于 System 1,但这一损失从θ 为 0.14 时的 1.9%(见表 2)降低到θ 为 0.13 时的 0.96%(如图 6 所示).其原因在于,降低 System 2 训练阶段的阈值θ,将增加 DTS 及其分类曲线可用的训练数据.其优点是:一方面使 $\bar{p}$ 能够涵盖更多噪声的特征,从而促进相关信息和噪声分布的可区分性;另一方面,新增的训练数据可进一步完善 DTS 的分布细节,使分类曲线更为精确.

但是过分降低阈值θ 将起到负面作用,甚至使系统性能不可控.其原因是,阈值太低将产生过饱和的训练数据,大量不影响“原型系统”性能的信息融入 DTS,其分布将误导分类曲线的估计;此外,这些信息将作为噪声被训练过程用于构造 $\bar{p}$,从而使 $\bar{p}$ 偏离噪声的质心,同时, $\bar{p}$ 有限的特征空间因覆盖过多信息而被泛化,致使真正需

要区分的相关信息和噪声无法利用 DTS 分布于二维空间,如图 3(a)所示,由此训练获得的分类曲线往往等同于随机曲线,造成 System 2 的性能极为不稳定,如图 5 和图 6 中阈值θ 低于 0.9 时 System 2 的性能指标.

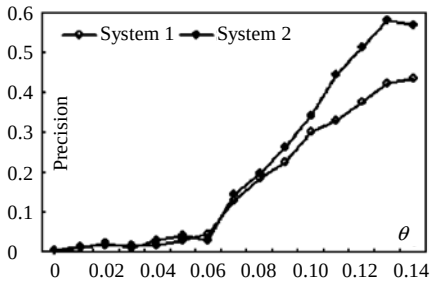


Fig.5 The trend of precision changing with θ in System 1 and System 2

图 5 System 1 和 System 2 精确率随θ的变化趋势

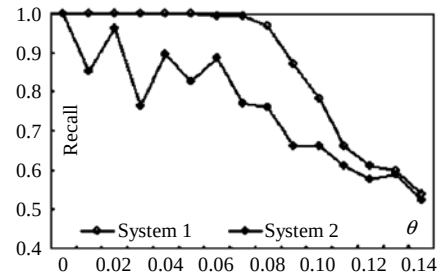


Fig.6 The trend of recall changing with θ System 1 and System 2

图 6 System 1 和 System 2 召回率随θ的变化趋势

5.2 有指导的自适应过滤系统:System 3 vs. System 4

尽管 System 2 依靠优化阈值θ进一步改进批过滤的性能,但平均精确率最优时为 0.581 6,说明部分噪声仍无法屏蔽,其原因包含:(1) 基于 LMS 的 AdaBoost 算法没有训练出分类曲线的最优解;(2) 分类曲线训练阶段最优,但测试阶段存在误判.通过人工观察,在参与测试的 29 个主题中,因分类曲线不是最优解而影响 System 2 性能的情况有 4 个(1 个奇异解);其他均属于误判,原因是训练中构造的 $\bar{p}$ 不能涵盖所有测试语料中的噪声,导致这些噪声基于 $\bar{p}$ 建立的二元近似关系位于相关信息的分布区域,从而难以有效检测并屏蔽.

针对批过滤系统的局限性,实验建立了有指导的自适应过滤系统 System 3 及 System 4,其中 System 4 以 System 3 为“原型系统”并嵌入 DTS,见第 4.3 节.作为有指导的自适应过滤系统,System 4 可以根据“原型系统”的输出动态更新 DTS 及分类曲线,借以改进恒定分类曲线无法有效屏蔽新噪声的缺陷.

如表 3 所示,System 4 的性能优于 System 3,前者的精确率高于后者约 7.7 个百分点,说明 DTS 在有指导的自适应过滤系统中也可以发挥屏蔽噪声的作用.但是,结果显示 System 3 与 System 4 的性能都低于批过滤系统,见第 5.1 节.原因在于初始用户模型只基于 2 篇相关文本进行训练,同时阈值θ 经验性地设置为较低指标,见第 4.3 节,从而导致过滤初始阶段的精确率很低.为此,实验将测试语料按时序划分为 10 个时段,并检验 System 3 与 System 4 在不同时段的平均精确率,如图 7 所示.

Table 3 Testing results of System 3 and System 4

表 3 System 3 和 System 4 测试结果

	T11SU	Precision	Recall
System 3	0.407 3	0.390 4	0.506 3
System 4	0.475 0	0.466 9	0.495 1

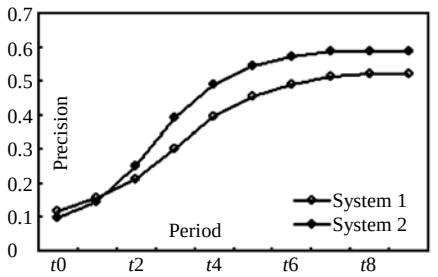


Fig.7 Alteration trend of precision with time in System 3 and System 4

图 7 System 3 和 System 4 精确率随时间的变化趋势

如图 7 所示,System 3 与 System 4 最初的平均精确率都很低,尤其是 System 4 更低于 System 3.原因是初始阈值θ偏低,System 4 的“原型系统”(即 System 3)输出过多噪声,导致每次自学习构造的 $\bar{p}$ 存在质心偏移和泛化现象.由此训练出的分类曲线往往不是最优解,增强了当前时段中误判的概率.但随着“原型系统”性能的提高,输出的相关信息也逐渐增多,同时,前期过滤中已积累了一定数量的相关信息.因此,在后续时段的 DTS 训练中,相关信息与噪声的比例逐渐趋向均匀,既削弱了 $\bar{p}$ 的质心偏移和泛化现象,也逐渐完善相关信息的二元近似关系分布,从而使 System 4 的平均精确率取得显著改进,并大幅优于 System 3.System 3 与 System 4 后期的精确率都高于批过滤系统,最优时 System 3 与 System 4 的平

均精确率分别为 0.521 8 和 0.586 9.

5.3 无指导的自适应过滤系统:System 5 vs. System 6

实验为检验 DTS 在无指导情况下对自适应过滤系统的影响,建立系统 System 5 和 System 6,见第 4.3 节,其测试结果见表 4.System 5 和 System 6 的性能明显低于批过滤系统,原因是用户模型的初始训练语料较少(2 篇相关信息),以及初始阈值设置偏低,导致过滤初期误判率较高,见第 5.2 节.此外,System 5 和 System 6 的性能也低于有指导的自适应过滤系统.其中,System 5 的精确率低于 System 3 约 3.4 个百分点,原因是 System 5 使用伪相关反馈更新用户模型,反馈中因无人指导而包含的大量噪声使用户模型逐渐产生偏差.

System 6 改进了“原型系统”(即 System 5)的性能,准确率提高约 1.4 个百分点.但是,DTS 中相关信息与噪声都基于聚类和用户模型的相关性进行估计,换言之,它们是非人工判断的伪相关信息和伪噪声,见第 4.3 节,因此 DTS 内的分布关系先天地存在误差.此外,System 6 无法借助人工判断,选择所有噪声构造 $\bar{p}$,而是经验性地选择相关性最低的 10 篇伪相关反馈构造 $\bar{p}$,造成 DTS 中某些噪声与 $\bar{p}$ 的相关度极低,甚至为 0,如图 3(d)所示,从而增大训练分类曲线的复杂度.基于 LMS 的 AdaBoost 算法在有限迭代次数内无法有效训练分类曲线的最优解.因此,System 6 对“原型系统”的改进幅度远远低于有指导的自适应过滤系统 System 4.

Table 4 Testing results of System 5 and System 6

表 4 System 5 和 System 6 测试结果

	T11SU	Precision	Recall
System 5	0.362 4	0.356 2	0.452 9
System 6	0.376 9	0.370 3	0.436 5

通过上述 3 组对比测试来看,基于 DTS 屏蔽噪声的方法对“原型系统”有很高的依赖性.当“原型系统”过滤性能较低时,DTS 屏蔽噪声的效果十分有限.与此对照,DTS 在“原型系统”性能较优时可以有效屏蔽噪声,如 System 2 最优时的精确率涨幅约为 16 个百分点.此外,虽然 DTS 显著提高了有指导自适应过滤系统的性能,但在无指导自适应过滤中取得的改进并不明显,而在实际应用中,用户提供的相关反馈往往是有限的,且可信性低.因此,如何进一步屏蔽无指导自适应过滤中的噪声,将是该领域未来研究的一项重点.

6 结 论

本文通过建立信息的二元近似关系,使一维相关度空间中不可分的相关信息和噪声映射为二维空间中近似可分的不同分布,并采用基于 LMS 分类器的 AdaBoost 算法估计相关信息和噪声的分类曲线,从而实现过滤系统对噪声的识别和屏蔽.实验将该算法嵌入批过滤、有指导和无指导的自适应过滤系统,并分别对比系统嵌入该算法前后的性能,实验结果表明,该方法有效提高了过滤系统屏蔽噪声的能力.尽管如此,实验中同时发现该算法对过滤系统本身的性能存在依赖性.针对这一现象,今后的工作将尝试采用错误驱动的方式触发噪声屏蔽算法,使其总是在过滤性能出现衰减趋势时进行噪声的检测和屏蔽,并辅助系统持续筛选相关信息进行自学习,从而避免后续过滤的偏差.

References:

[1] Hang XJ, Xia YJ, Wu LD. A text filtering system based on vector space model. Journal of Software, 2003,14(3):435−442 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/14/435.htm

[2] Hanani U, Shapira B, Shoval P. Information filtering: Overview of issues, research and systems. User Modeling and User-Adapted Interaction, 2001,11(3):203−259.

[3] Belkin NJ, Croft WB. Information filtering and information retrieval: Two sides of the same coin? Communications of the ACM, 1994,35(12):29−38.

[4] Lewis DD. Applying support vector machines to the TREC-2001 batch filtering and routing tasks. In: Voorhees E, Harman D, eds. Proc. of the TREC 2001. Gaithersburg: NIST Special Publication, 2001. 286−292.

[5] Robertson S, Hull DA. The TREC-9 filtering track final report. In: Voorhees EM, Harman DK, eds. Proc. of the 9th Text Retrieval Conf. (TREC-9). Gaithersburg: NIST Special Publication, 2001. 25−40.

[6] Ault T, Yang YM. *k*NN, Rocchio and metrics for information filtering at TREC-10. In: Voorhees EM, Harman DK, eds. Proc. of the 10th Text Retrieval Conf. (TREC 10). Gaithersburg: Department of Commerce, National Institute of Standards and Technology,

2001. 84−92.

[7] Nallapati R. Discriminative models for information retrieval. In: Sanderson M, Jarveln K, Allan J, Bruza P, eds. Proc. of the 27th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Sheffield: ACM Press, 2004. 64−71.

[8] Robertson SE, Walker S, Beaulieu M. Okapi at Trec-7: Automatic ad hoc, filtering, VLC and interactive track. In: Voorhees EM, Harman DK, eds. Proc. of the 7th Text Retrieval Conf. (TREC-7). Gaithersburg: NIST Special Publication, 1998. 253−264.

[9] Burges CJC. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, 1998,2(2): 121−167.

[10] Zhang Y, Callan J. Maximum likelihood estimation for filtering thresholds. In: Croft BW, Harper DJ, Kraft DH, Zobel J, eds. Proc. of the 24th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR 2002). New York: ACM Press, 2002. 294−302.

[11] Yang YM, Kisiel B. Margin-Based local regression for adaptive filtering. In: Donald K, Ophir F, Joachim H, Sajada Q, Len S, eds. Proc. of the 12th Int'l Conf. on Information and Knowledge Management. New Orleans: ACM Press, 2003. 88−95.

[12] Allan J. Incremental relevance feedback for information filtering. In: Hans PF, Donna H, Peter S, Ross W, eds. Proc. of the 19th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: ACM Press, 1996. 270−277.

[13] Joachims T. Optimizing search engines using clickthrough data. In: Randy G, David H, Daniel K, Raymond N, eds. Proc. of the 8th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (KDD2002). Edmonton: ACM Press, 2002. 102−110.

[14] Zeng C, Xing CX, Zhou LZ. A survey of personalization technology. Journal of Software, 2002,13(10):1952−1961 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/13/1952.pdf

[15] Robertson SE, Walker S, Jones S. Okapi at TREC-3. In: Harman DK, eds. Overview of the 3rd Text Retrieval Conf. (TREC-3) NIST. Gaithersburg: NIST Special Publication, 1995. 109−126.

[16] Grove AJ, Littlestone N, Schuurmans D. General convergence results for linear discriminant updates. Machine Learning, 2001,43(3): 173−210.

[17] Xie CF, Li X. A sequence-based automatic text classification algorithm. Journal of Software, 2002,13(4):783−789 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/13/783.pdf

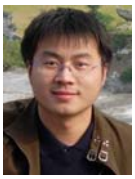
[18] Collins M, Schapire RE, Singer Y. Logistic regression, AdaBoost and Bregman distance. Machine Learning, 2002,48:253−285.

附中文参考文献:

[1] 黄萱菁,夏迎炬,吴立德.基于向量空间模型的文本过滤系统.软件学报,2003,14(3):435−442. http://www.jos.org.cn/1000-9825/14/435.htm

[14] 曾春,邢春晓,周立柱.个性化服务技术综述.软件学报,2002,13(10):1952−1961. http://www.jos.org.cn/1000-9825/13/1952.pdf

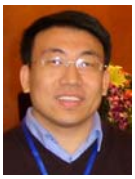
[17] 解冲锋,李星.基于序列的文本自动分类算法.软件学报,2002,13(4):783−789. http://www.jos.org.cn/1000-9825/13/783.pdf



洪宇(1978－),男,黑龙江哈尔滨人,博士生,CCF 学生会员,主要研究领域为话题检测与跟踪,信息过滤,个性化信息检索.

张宇(1972－),男,博士,副教授,CCF 高级会员,主要研究领域为信息过滤,自动问答,自然语言处理.

郑伟(1984－),男,硕士生,主要研究领域为话题跟踪,社会关系网络及应用.



刘挺(1972－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为自然语言处理,信息检索,认知心理学.

李生(1943－),男,博士,教授,博士生导师,CCF 会员,主要研究领域为自然语言处理,信息检索,机器翻译.