文章编号:1001 - 9081(2006)05 - 1099 - 03

基于增量学习和阈值优化的自适应信息过滤研究

王金宝[1,2]
(1. 大连理工大学 计算机科学与工程系,辽宁 大连 116024; 2. 西安陆军学院,陕西 西安 710108)
(wjbwxywxm@ hotmail. com)

摘 要:为了适应实时在线的网络信息过滤需求,提出了一种新的自适应过滤模型。在系统的初始化阶段,运用增量学习方法对附加的少量伪相关文档进行学习,采用改进的文档词频方法来抽取特征词,以此扩展需求模板,提高模板准确度。在系统测试阶段,以系统效能指标最优为目标,提出了将概率模型和文档正例分布统计方法相结合来实现阈值优化的新算法。

关键词:自适应信息过滤;伪相关反馈;增量学习;阈值优化
中图分类号: TP393.08 **文献标识码**:A

Research on adaptive information filtering based on incremental learning and threshold optimization

WANG Jin-bao[1,2]
(1. *Department of Computer Science and Engineering, Dalian University of Technology, Dalian Liaoning* 116024, *China;*
2. *Xi'an Military Academy, Xi'an Shaanxi* 710108, *China*)

Abstract: In order to meet web-based on-line information filtering requirement, a new realistic adaptive information filtering model was proposed in this paper. In the system initiations stage, the profile is improved by active topic term learning and term weight value updating based on incremental learning the few pseudo relevant feedback samples. An incremental feature selection method was presented based on document frequency. In the filtering test stage, aimed at the system's optimization utility, a new technique was proposed which combined the probabilistic distribution model and document stream statistical information to update and explore the dissemination threshold actively. Experimental results show that the new methods lead to a higher performance in the adaptive information filtering system.

Key words: adaptive information filtering; pseudo relevance feedback; incremental learning; threshold optimization

0 引言

传统的信息过滤系统需要大量的训练语料来获取能够准确表示用户信息需求[1](user profile)的模板,难以适应用户当前在线的信息过滤需求和对 Web 信息进行实时监测。自适应信息过滤[2,3]只需少量的用户需求信息来构建需求模板,并在过滤中能自主地学习用户的反馈信息来增强模板的准确性。

为了提高初始阶段过滤模板的准确性,通常附加少量的未标注训练语料,通过学习伪相关反馈信息来扩展模板。唐焕玲等[4]对于各种模板特征选择和特征权重调整方法进行了深入的研究。其他应用互信息[5]、KL 距离[6]、最大熵[7]等方法进行模板特征选择也都取得了较好的效果。何静等[8]利用向量中值法来更新需求模板,其他像 K 近邻,贝叶斯学习和加强算法等[9]都应用到模板的调整中来。在这些方法中,运用伪相关反馈方法学习附加的训练语料时,极易加入大量的噪音词,致使模板训练后的准确性很差。

在求取最优过滤阈值研究方面,夏迎炬等[10]运用启发式方法来求取最优过滤阈值。马亮等[5]提出的基于整体的正例文档分布密度的预测,Zhang 等[11]应用极大似然法对过滤阈值的估计方法取得了较好的效果。但是,复杂的期望模型没有应用文档流的分布变化信息来探测最优过滤阈值,而在正例统计分布的方法中,进行阈值调整的经验函数难于获取。针对这些不足,我们提出了一些改进的方法。

1 自适应过滤模型训练

在自适应过滤的初始阶段,一般只提供简短的用户信息需求描述和两三个正例,没有训练集用以需求模板的训练,所以,用这些信息形成的初始模板很难准确地表达用户的信息需求。而在附加的少量的训练语料上通过单次伪相关反馈的训练效果并不理想。因此,我们采用改进的文档词频方法选择特征,同时运用一种新的增量式迭代学习方法,以此实现过滤模板的训练。

1.1 改进的特征选择方法

在增量学习伪相关反馈信息的过程中,我们以相似度值排序较前的文档集为观察窗口,试图从中发现新出现的最能表示用户信息需求的特征词,因为这些文档距离用户信息需求的主题概念最近,从中选择的特征词必然最具信息量和鉴别力。在每次的增量反馈中,我们选择排序最前的文档作为伪正例,并从伪正例中发现新出现的特征词,统计新出现的关键词词频,并和其文档词频进行联合,作为特征词的评估函数,评估函数公式是:

$$w(t) = df(t) \times \log(tf_k(t,\vec{d}) + 1) \quad (1)$$

其中,$\vec{d}$ 指的是伪正例文档。最后选择适量的特征词加入到特征项空间中去,同时调整模板特征向量,公式为:

$$Q_{new} = Q_{old} + \beta \frac{\sum d_{rel}}{n_{rel}} \quad (2)$$

其中,Q_{new} 是新的模板特征向量,Q_{old} 是旧的模板特征向量,这

收稿日期:2005 - 11 - 29;**修订日期**:2006 - 02 - 28
作者简介:王金宝(1972 -),男,陕西扶风人,硕士,主要研究方向:中文信息处理、信息过滤.

里 β 是比例约束因子,实验中我们设为 1。

这个公式实际上是 Rocchio 公式的一种变形和改进,它是用选出的与模板非常接近的文档进行相关反馈学习,不断发现新的特征,并对特征的权值进行修正,通过增量式迭代反馈学习,使模板不断向真实的用户需求逼近。

1.2 增量式伪相关反馈学习

增量学习伪相关反馈信息的算法是:

设附加的训练集为 S_t,正例文档集为 R(包括初始阶段提供的两三个正例),伪正例文档集为 P。文档相似度排序前 m 个文档集为观察窗口 W。每次迭代学习都从观察窗口 W 和伪正例文档中发现新特征词。

(1) 用正例文档集 R 中的文档形成初始模板 Q_{old},并计算附加文档集 S_t 中的所有文档的相似度,按相似度大小排序。

(2) 迭代 N 次增量学习:

(a) 从 S_t 中选取相似度最大的一篇文档 D 作为伪正例文档进行反馈学习,并将其加入到伪正例文档集 P 中。

(b) 从选择的伪正例文档 D 和文档集观察窗口 W 中抽取适量的新特征词,新特征词数量用公式(3) 计算,最后计算新特征词的权重,用公式(2) 扩展模板向量,形成新模板 Q_{new}。

(c) 用新模板计算附加文档集 S_t 的相似度,从前 m 个文档中获取新出现的伪正例文档数量,记为 n。统计新出现的特征词,作为候选特征词。

(d) 如果 N 为零,则转到(e),否则转到(a);

(e) 选择第 m 个文档的相似度值作为初始阈值,结束训练。

在上述算法中,针对现有的特征选择方法在伪相关文档集中的不足,我们提出一种改进的文档词频特征选择方法,将新出现的特征词以其在正例和伪相关文档集中的文档词频 DF 及其在伪正例中的词频 TF 进行组合,并按大小排序,力图发现和用户主题需求最为接近的新特征。在特征数量的选择上采用式(3) 求取特征数量。其中 N 为增加特征数目,a 为初始特征词数目,n 为观察窗口 W 中每次增量反馈后新出现的文本数。

$$N = a + a * \lg(n+1) \tag{3}$$

实验中将参数 a 设为 5。

2 自适应过滤模型测试

2.1 过滤阈值的自适应优化调整

我们应用一种新的方法来探测和调整过滤阈值,主要是将概率模型和正例分布统计的方法相结合,实现系统在未标注的测试语料条件下自我监督学习,探测最优的过滤阈值。在新方法中,用初始的过滤阈值来过滤文档流中的每个文档,然后通过高斯指数模型[11,12] 来求解过滤后文档的相关性概率,依据 2.2 节所述的相关性判定规则,判定每个文档相关与否,最终得到文档的伪相关信息。在此基础上,我们根据近期过滤后正例分布信息,以系统效能指标 T11U 指标为优化目标,探索在使系统性能不断提高的条件下阈值的调整方向和幅度。

阈值调整算法表述如下:

假设:

(1) p_t 为累积文档集精度,即在上次用户相关信息反馈后,在累积的训练数据上得到的系统过滤精度。

(2) p_c 为当前时期段精度,即系统将相似度值大于当前阈值点的文档加入到累积文档中所形成的精度。

(3) s_l 为当前时间段的伪正例文档相似度平均值。

(4) s_g 为全局正例相似度平均值(累积文档集中正例文档相似度均值)。

(5) $v_{\max}$ 和 $v_{\min}$ 分别为当前时候段文档相似度最大值和最小值。

(6) T 为当前时候段的阈值。

阈值调整方法为:

```
If ( s_l > s_g )
  Threshold(v_min , T )
Else
  Threshold( T , v_max )
```

其中 $Threshold(x,y)$ 函数为在 $[x,y]$ 区间选择最优阈值的函数,选择最优阈值的条件是:

```
If ( p_c > p_t )
  If ( MaxUtility(x,y) )
    Return( Threshold)
    p_t = p_c
  ELSE
    Return( Pseudo Threshold)
ELSE
  Return( init threshold)
```

最后我们用这个设定的阈值作为下一时间段文档流中的每个文档相关与否的门限,并最终实现最优阈值的探测和反馈信息学习。

2.2 文档相关性判定策略

在测试集上我们采用较为准确的概率模型——高斯指数分布模型[11,12] 来推知当前文档的相关信息,运用用户多次反馈的累积文档来训练模型,该模型认为累积文档中的相关文档呈现高斯(正态)分布,而不相关文档呈现指数分布,并以累积文档中的相关信息作为先验知识,运用贝叶斯定理和全概率理论推知文档相关性的概率。公式为:

$$P(rel \mid score) = \frac{P(score \mid rel)P(rel)}{P(score \mid rel)P(rel) + P(score \mid nonrel)P(nonrel)} \tag{4}$$

根据高斯分布,求取当前给定相似度值的相关文档的先验概率用如下公式:

$$P(score \mid rel) = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{score} e^{-\frac{(t-\mu_t)^2}{2\sigma^2}}dt \tag{5}$$

根据指数分布,求取当前给定相似度值的不相关文档的先验概率用如下公式:

$$P(score \mid nonrel) = \int_{-\infty}^{score} \lambda e^{-\lambda t}dt \tag{6}$$

而 $P(rel)$ 和 $P(nonrel)$ 则为相关文档与不相关文档的先验概率,根据概率定义用如下公式:

$$P(rel) = r/n \tag{7}$$

$$P(nonrel) = 1 - P(rel) \tag{8}$$

运用公式(4) ~ (8),可以得出当前文档的相关性概率,为了准确地判别文档相关信息,我们必须依据 T11U 效用函数,求出最优概率阈值的门限,T11U 公式为:

$$T11U = 2a - b \tag{9}$$

对于采用 T11U 作为评估策略的系统,最优阈值的概率门限应该是大于 1/3[11,12]。因此,针对当前文档,我们可以准确地得到判定相关性的决策:

$$P(rel \mid score) >= \frac{1}{3} \tag{10}$$

式(10) 说明,文档相关性概率大于 1/3 时为相关文档。否则为不相关文档。通过上述规则的设定,系统可以快速得到当前时间段内文档相关性信息。

3 实验语料与实验方法

3.1 实验语料

由于中文信息过滤没有统一的语料供系统测试使用,所

以我们以国际 TREC 语料格式为标准,从中华网、新浪、搜狐和雅虎等权威网站获取特定主题网页,结合手工标注,建立了五个主题的语料库,共 500 多篇文档进行测试。每个主题 100 条文档,分为训练集和测试集两部分,提供系统训练和测试使用。

3.2　实验结果及分析

3.2.1　训练集上伪相关反馈学习实验

在自适应的训练阶段,特征选择是一个重要的环节,我们将互信息 MI 和改进的文档频次两种方法在以 AP(Average precision)指标基础上进行对比,结果见表 1 所示。

表 1　两种特征选择方法平均精度比较

互信息	Log *(tf)* * *df*
0.728	0.98

为了比较伪相关反馈学习中增量学习和单次反馈学习对过滤模板训练的效果,在实验中我们基于相同的训练语料规模(共 2 0 篇文档,相关文档,相近文档和不相关文档的比例为 6 : 7 : 7),实验中将两种学习方法结果进行比较,实验结果见图 1。

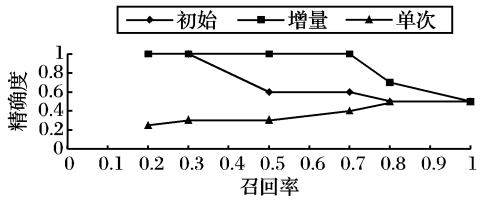

图 1　增量学习与单次反馈精度召回率比较

从图 1 中可以看到,增量反馈学习的效果远远高于单次反馈学习的效果,运用平均精度(Average Precision)指标比较实验数据,结果见表 2。

表 2　增量学习与单次反馈学习后平均精度比较

p@n	未训练	单次反馈	增量学习
6	0.716 1	0.37 -48%	0.869 1 +21.4%

3.2.2　测试集上阈值自适应调整的试验

在测试集上,将通过伪相关反馈学习后得到的主题模板运用于无标注的测试集中,进行自适应过滤阈值调整的测试,以此检验我们提出的探测最优阈值算法的性能。实验最终结果见图 2、图 3。

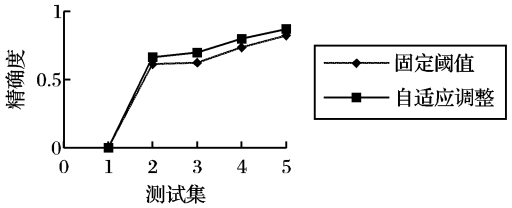

图 2　测试集上阈值调整精度比较

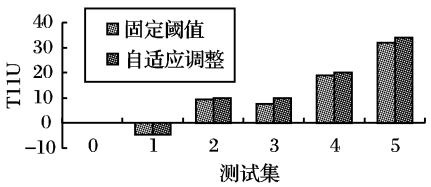

图 3　测试集上阈值调整 T11U 比较

实验中为了模拟较为客观真实的文本流数据,我们将测试集文档重新组织,共分为五组测试集,每组相关文档与不相关文档的数目比例配置为 0 : 1,1 : 1,1 : 2,2 : 1,1 : 0。每组测试集共 20 篇文档,其中不相关文档部分又细分为相近文档和不相关文档。这样做的目的在于检验系统进行过滤时面对动态的文档流的自适应能力和自主学习效果,以及系统对于与当前过滤主题概念有交叉的文档的区分能力。

从图 1 和图 2 中可以看出,在精度和 T11U 两个性能指标上,自适应阈值调整的方法明显要比固定阈值的方法好。

3.2.3　系统整体性能的比较

为了衡量和比较整个系统的过滤性能,我们首先再现了复旦大学的自适应过滤系统[2,10],该系统应用统计文档正例分布的方法探测最优的过滤阈值,与同类系统相比,过滤性能比较优越;然后基于相同的训练集和测试集,按照 3.2.2 节所述方法在复旦自适应过滤系统和本文所建系统上实验,实验结果数据如表 3 所示。

表 3　两种系统整体性能比较

系统	精确度	召回率	T11U	T11SU	T11F
本文所建系统	0.80	0.47	42.67	0.89	0.67
复旦过滤系统	0.73	0.33	27.67	0.86	0.59

从表中可以看出,本文提出的增量学习和阈值优化方法在自适应过滤中取得较为满意的结果。在测试集上系统在各个性能指标上明显要比复旦系统好,实验结果数据的对比充分说明了我们的方法能充分提高系统的过滤性能。

4　结语

我们应用增量学习方法学习伪相关反馈信息,改进了在小样本训练集上快速地进行模板特征选择的方法,提高了模板的准确性;同时在测试数据集上运用概率模型与正例分布统计的方法实现了自适应过滤阈值优化,并对运用用户反馈信息进行模板阈值调整进行了探索。实验结果表明新方法提高了系统整体性能,适应了当前在线信息过滤的迫切需求。

参考文献:

[1] BELKIN NJ, CROFT WB. Information filtering and information retrieval: two sides of the same coin[J]. Communications of the ACM, 1992, 35 (12): 29 - 38.

[2] 黄萱菁, 夏迎炬, 吴立德. 基于向量空间模型的文本过滤系统[J]. 软件学报, 2003, 14(3): 435 - 442.

[3] 马亮, 陈群秀, 蔡莲红. 一种改进的自适应文本信息过滤模型[J]. 计算机研究与发展, 2005, 42(1): 79 - 84.

[4] 唐焕玲, 孙建涛, 陆玉昌. 文本分类中结合评估函数的 TEF-WA 权值调整技术[J]. 计算机研究与发展, 2005, 42(1): 47 - 53.

[5] 赵林, 胡恬, 黄萱菁, 等. 基于知网的概念特征抽取方法[J]. 通信学报, 2004, 25(7): 46 - 54.

[6] 杨晔, 彭宏, 林嘉宜, 等. 一种有效特征词发现的贝叶斯文本分类方法[J]. 系统工程, 2004, 22(9): 107 - 110.

[7] 宋国杰, 唐世渭, 杨冬青, 等. 基于最大熵原理的空间特征选择方法[J]. 软件学报, 2003, 14(9): 1544 - 1550.

[8] 何静, 刘海燕, 宫云战. 内容过滤中过滤模板的改进技术研究[J]. 通信学报, 2004, 25(3): 112 - 117.

[9] 王斌, 潘文锋. 基于内容的垃圾邮件过滤技术综述[J]. 中文信息学报, 2005, 19(5): 1 - 10.

[10] 夏迎炬, 黄萱菁, 胡恬, 等. 自适应信息过滤中使用少量正例进行阈值优化[J]. 软件学报, 2003, 14(10).

[11] ZHANG Y, CALLAN J. Maximum likelihood estimation for filtering thresholds[A]. Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval[C], 2001, 24(9): 294 - 302.

[12] ARAMPATZIS A, HAMERAN A. The score-distributional threshold optimization for adaptive binary classification tasks[A]. Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval[C], 2001, 24 (9): 285 - 293.