第16卷 第5期
2006年5月

计算机技术与发展
COMPUTER TECHNOLOGY AND DEVELOPMENT

Vol. 16 No. 5
May 2006

# 基于概念学习的过滤模板获取方法

朱祥玉,侯德文
(山东师范大学 信息管理学院,山东 济南 250014)

**摘 要**:基于内容的文本过滤关键在于建立有效的过滤模板。一种高效的过滤模板可以降低整个文本过滤系统对机器学习机制的要求,提高系统的过滤效率。提出了一种基于概念学习的过滤模板获取方法。该方法结合处理文本特征项的需要改进了概念学习方法中的寻找极大特殊假设算法,并应用新的算法从给定的少量训练文本中提取用户过滤模板。实验结果表明,与直接使用主题描述作为过滤模板的方法相比,较大地提高了过滤精度,可以达到比较令人满意的过滤效果。

**关键词**:文本过滤;过滤模板;概念学习;TFFind-S算法
**中图分类号**:TP301 **文献标识码**:A **文章编号**:1673-629X(2006)05-0053-03

# Method of Filtering Profile Extraction Based on Concept Learning

ZHU Xiang-yu, HOU De-wen
(Dept. of Infor. Management, Shandong Normal Univ., Jinan 250014, China)

**Abstract**: The key to content-based text filtering consists in constructing an effective filtering profile. An effective filtering profile can debase the request from the whole text filtering system to the machine learning mechanism and improve filtering efficiency of system. This paper brings forward a method for constructing filtering profile. The method improves the find-maximum-special-supposion algorithm in the methods of concept learning by combining the need for dealing with the text feature items and constructs filtering profile from a few training texts by using the new algorithm. The result of experiments shows that, compared with the method which uses the subject-description as filtering profile straight, this method improves filtering precision markedly, and it can obtain the satisfying filtering purpose.

**Key words**: text filtering; filtering profile; concept learning; TFFind-S algorithm

## 0 引 言

在自适应文本过滤中,系统需要根据非常少的训练文档建立初始过滤模板;根据建立的过滤模板,系统对输入的文档流进行过滤,并将系统认为相关的文档提交给用户,用户对提交的文档给以反馈;然后系统再根据用户的反馈调整过滤模板,以此提高系统的过滤性能。而如果能够找到一种方法使得初始化得到的过滤模板不经过反馈学习就能够取得较为令人满意的过滤效果,则可降低系统对机器学习机制的要求,为系统实现的后继工作顺利进行奠定基础。

## 1 基于概念学习的过滤模板获取方法

### 1.1 概念学习

概念学习(concept learning)指给定某一类别的若干正例和反例,从中获得该类别的一般定义的过程。它也可以看作是一个搜索问题的过程,在预定义的假设空间中搜索假设,使其与训练样例有最佳的拟合。多数情形下,为了高效地搜索,可以利用假设空间中一种自然形成的结构即一般到特殊偏序结构,利用这种结构可以在无限的假设空间中进行彻底地搜索,而不需要明确地列举所有的假设。

Find-S(寻找极大特殊假设)算法[1]是一种常用的利用一般到特殊偏序结构搜索与训练样例相一致假设的算法。Find-S算法从假设空间中最特殊假设开始,在假设覆盖正例失败时将其一般化(当一个假设能正确地划分一个正例时称该假设覆盖该正例),Find-S算法精确描述如下:

(1)将某假设初始化为假设空间中最特殊假设;
(2)对每一个正例
对该假设的每一个属性约束
如果正例满足约束
那么不做任何处理
否则将假设中的该属性替换为该正例满足的另一个更一般约束;
(3)输出该假设。

### 1.2 基于概念学习的过滤模板获取方法

该方法的基本思想是:对给定的少量训练文本,利用改进的Find-S算法进行概念学习,从而获得相关主题的一般概念,即用户过滤模板。由此法获得过滤模板的过程

**收稿日期**:2005-08-29
**基金项目**:山东省中青年科学家奖励基金(03BS009)
**作者简介**:朱祥玉(1980-),男,山东潍坊人,硕士研究生,研究方向为文本过滤、数据挖掘;侯德文,副教授,硕士生导师,研究方向为计算机网络、Web数据挖掘、图像处理。

如图 1 所示。

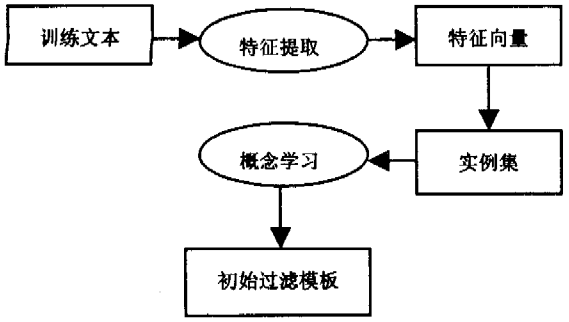


图 1 初始过滤模板获取过程

首先对给定的训练文本进行特征抽取,分别得到对应的特征向量;然后将所有训练文本对应的特征向量经整理后转变为实例,并由所有训练文本对应的实例构成实例集;对实例集进行概念学习,从而获得初始过滤模板。

图中概念学习部分涉及的具体子概念定义描述如下:

(1)实例集 $X$。

$X$ 中每个实例对应一个训练文本的特征向量,特征向量经整理转变为由多个特征项词性作为属性的描述形式,即每个属性对应着一个同一词性的特征项的集合,所有这些属性集合共同构成了原来的特征向量,且每个属性的取值范围为任意本词性所允许的特征项(即字、词或短语等);

(2) 假设集 $H$。

$H$ 中每个假设 $h$ 描述为多个词性的所有取值的合取;在描述形式上等同于任何一个实例。

(3) 目标概念 $c$。

$c$ 是一个布尔函数,其变量为实例集 $X$ 中的实例;得出的结果为实例是否属于相关主题的结论。

(4) 训练样例集 $D$。

$D$ 由目标函数(或目标概念) 的正例和反例组成,其中每一个样例由一个实例以及该样例是否属于目标概念 $c$ 的标注构成。

用此法获得过滤模板的概念学习部分由 Find - S 算法改进而来,命名为 TFFind - S算法,具体算法描述如下:

① 将 $h$ 初始化为 $H$ 中最特殊假设;

② 对于实例集 $X$ 中的每个正例 $x$。

对 $h$ 的每个属性特征项集合 $s_i$

如果 $s_i$ 不包含 $x$ 的相应属性特征项集合

那么将 $x$ 中对应属性特征项集合中的元素添加进 $h$ 中该属性特征项集合:

对于要添加的每一元素

如果 $h$ 的对应属性特征项集合中原来没有

那么直接添加进去

否则只将 $h$ 中对应元素的个数加 1;

③ 输出目标假设 $h$。

然后对目标假设 $h$ 进行人工评测:首先去掉 $h$ 的各属性特征项集合中词义相同或相近的特征项;然后根据一定的阈值筛选出各属性特征项集合中的最优特征项,阈值的设定以保持筛选后 $h$ 中的特征项总和与各实例中特征项数目相同为原则。

对训练文本实例集利用 TFFind - S 算法进行概念学习即可获得相关主题的用户过滤模板。

### 1.3 实验测试

考虑到人工评测工作量的庞大,只选取了关于"法轮功"和"台独"两个主题来进行性能评测,所使用的语料选自中国新闻网、人民网、中华网、解放军报网络版、新浪、雅虎等网站上共 150 个页面的新闻语料,其中人工评测出有关反对、揭批"法轮功"的相关新闻语料共 76 篇、反对"台独"的相关新闻语料共 57 篇;另外从各个主题中选出正例文本各 10 篇作为训练文本。

经人工整理,给出"法轮功"和"台独"两个主题的描述分别为:

法轮功:揭批,法轮功,邪教,组织

台　独:反对,台独,和平,统一

所有文本在分词后经预处理继而表示为特征向量形式,特征向量取平均维数 146,按照一般标准选择 2% ~ 5%的最佳特征来表示文本[2],采用评估函数做特征提取[3]后各文本特征向量维数取为 4。使用给出的 20 篇训练文本经概念学习后获得目标假设(即"法轮功"和"台独"的一般概念)为:

filtering profile1 = {<法轮功,83.088738 >,<邪教,42.840315>,<是,25.437070>,<组织,6.083116 >}

filtering profile2 = {<台独,176.000393 >,<反对,143.237009 >,<中国,136.210078 >,<统一,22.368091 >}

这就是最终得到的有关"法轮功"和"台独"两个主题的过滤模板。

系统采用信息检索中常用的准确率(Precision)和查全率(Recall)两个指标及 TREC[4]提出的评测指标 Utility 来进行评价,它们的描述如下:

(1)准确率是信息检索的性能指标,定义为被检出的相关文档数除以所有检出的文档数的值。其数学公式表示如下:

$$\text{准确率(precision)} = \frac{\text{检出的相关文档数}}{\text{检出的文档数}}$$

(2)查全率是信息检索的另一个性能指标,定义为检出的相关文档数除以集合中全部相关文档数的值。其数学公式表示如下:

$$\text{查全率(recall)} = \frac{\text{检出的相关文档数}}{\text{全部相关文档数}}$$

准确率表明系统的精确性;查全率反映了系统的覆盖性。这两个量不是独立的,其中一个指标的提高往往以另一个指标的降低为代价。

(3)给定主题和文本,文本可能相关,也可能不相关;过滤系统可能检出该文本,也可能未检出,由此可建立如下表格[5]:

| | 相关 | 不相关 |
|---|---|---|
| 检出 | R1/A | N1/B |
| 未检出 | R0/C | N0/D |

检出相关文本和未检出不相关文本都是过滤正确的情况。而未检出相关文本意味着遗漏,检出不相关文本意味着错检。线性 utility 函数对这4种情况赋相应的权重:

Utility = A * R1 + B * N1 + C * R0 + D * N0

这里的 R1/R0/N1/N0 指的是每个主题4种文本的数量,A,B,C,D 决定了每种情况的代价。Utility 值越大,系统的过滤性能就越好。TREC10 中选择 A=2,B=-1,这样得到的指标称为 T10U[6]。

实验测试结果如表1所示。

表1　概念学习获取过滤模板测试结果

| | 相关文本数(法轮功/台独) | 不相关文本数(法轮功/台独) |
|---|---|---|
| 检出 | 54/37 | 25/18 |
| 未检出 | 12/10 | 39/65 |

在表1结果上的三项评测指标结果如表2所示。

表2　概念学习获取过滤模板评测指标结果

| Precision(法轮功/台独) | Recall(法轮功/台独) | T10U(法轮功/台独) |
|---|---|---|
| 0.6835/0.6727 | 0.8181/0.7872 | 110/111 |

将给出的主题描述作为用户模板,并将训练文本中对应特征项的平均评估值作为其各个主题词的权值,以此方法进行实验得测试结果如表3所示。

表3　主题描述作为过滤模板测试结果

| | 相关文本数(法轮功/台独) | 不相关文本数(法轮功/台独) |
|---|---|---|
| 检出 | 46/33 | 29/19 |
| 未检出 | 20/14 | 35/64 |

在表3结果上的三项评测指标结果如表4所示。

表4　主题描述作为过滤模板评测指标结果

| Precision(法轮功/台独) | Recall(法轮功/台独) | T10U(法轮功/台独) |
|---|---|---|
| 0.6133/0.6346 | 0.6969/0.7021 | 78/97 |

以上两种方法的评测指标结果对比如表5所示。

表5　两种方法的评测指标结果对比

| | 平均准确率 | 平均查全率 | 平均T10U |
|---|---|---|---|
| 概念学习用户模板 | 0.6781 | 0.8027 | 111 |
| 主题描述用户模板 | 0.6239 | 0.6995 | 88 |

实验结果分析:在没有模板学习即系统不具有反馈功能的条件下进行试验并获得了上述结果,从结果来看,由概念学习一次性获取的过滤模板用于文本过滤系统,与直接使用主题描述作为用户模板的方法相比,过滤效果已有明显提高;另外,其相对于不同主题的平均准确率已接近70%,平均查全率已超过80%,过滤效果已比较令人满意,但距理想结果仍差距很大,如果能够在自适应过滤系统上进行实验,就有可能更大幅度地提高过滤性能。

## 2　结束语

基于内容的文本过滤包括3个基本环节[7]:确定用户的信息需求,即建立过滤模板;确定文本与过滤模板的匹配机制以及利用用户评注作为相关反馈动态改进过滤模板。其中如何建立过滤模板是一个关键问题。文中提出了一种基于概念学习的过滤模板获取方法,即通过对给定的少量训练文本进行概念学习来获取用户过滤模板。实验部分验证了此法具有一定的可行性。

**参考文献:**


[1] Mitchell T M. 机器学习[M]. 曾华军,张银奎译. 北京:机械工业出版社,2003.

[2] 李　凡,鲁明羽,陆玉昌. 关于文本特征抽取新方法的研究[J]. 清华大学学报(自然科学版),2001(7):98-101.

[3] 张鹏飞,李　贤,刘建毅,等. 基于相对词频的文本特征抽取方法[J]. 计算机应用研究,2005(4):23-26.

[4] 王　斌. TREC之文本过滤技术[R]. 北京:中科院计算所软件室,2001.

[5] 夏迎炬. 文本过滤关键技术研究[D]. 上海:复旦大学,2003.

[6] 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(7):46-54.

[7] 黄钢石,张亚非,陆建江,等. 基于NMF的用户模板构造方法[J]. 情报学报,2004(8):394-398.


(上接第52页)


ICDE 1995[C]. Taipei:[s.n.],1995.3-14.

[2] Nielsen J. 可用性工程[M]. 刘正捷译. 北京:机械工业出版社,2004.

[3] Srikant R, Agrawal R. Mining Sequential Patterns: Generalizations and Performance Improvements[A]. In: Proc. 5th Int. Conf. Extending Database Technology[C]. Avignon, France:[s.n.],1996.3-17.

[4] Han J, Pei J, Mortazavi-Asl B, et al. FreeSpan: Frequent pattern projected sequential pattern mining[A]. In: Proc. 2000 Int. Conf. Knowledge Discovery and Data Mining[C]. Boston, MA:[s.n.],2000.355-359.

[5] Pei J, Han J, Mortazavi-Asl B, et al. PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth[A]. In: ICDE 2001[C]. Heidelberg, Germany:[s.n.],2001.215-224.