期望最大化算法在文本自动分类处理中的应用研究综述

杨文静，20131113，BUAA

**摘要：**人工对文本材料进行分类的方法是通读所有文章，然后再对它们进行归类保存，具有周期长、费用高、效率低的特点。

**关键字：**应用，期望最大化算法，文本自动分类

# 引言

现在，人们能通过因特网能很快地得到大量的资料，如何对所获得资料进行科学有效地管理是摆在人们面前一个的问题。对资料进行管理一个很常见的方法就是对它们系统地进行分类。 人工对文本材料进行分类的方法是通读所有文章，然后再对它们进行归类保存，具有周期长、费用高、效率低的特点，如何利用计算机实现文本自动分类是当前研究的热点问题。

# 文本分类技术

文本自动分类就是由计算机自动提取文本的特征项, 依据一定的算法, 将文本按内容或属性归到一个或多个类别的过程。文本自动分类的流程图图 1所示。



图 1 文本自动分类流程图[1]

其主要包括文本训练过程、分类过程二个部分[2]，其主要包括。它通过训练一定的文本集合得到类别与未知文本的映射规则，即计算出文本与类别的相关度, 再采取一定的阈值策略决定文本的类别归属[3]。

从数学角度来看，文本分类是一个映射的过程，它将未标明类别的文本映射到已有的类别中，该映射可以是一对一映射，也可以是一对多的映射，因为通常一篇文本可以同多个类别相关联[4]。

## 特征的抽取

现有的文本分类的方法可以分为基于外延的分类方法和基于语义的分类方法[5]：

基于外延的方法不关心文本的语义，根据文本的外在特征进行分类。最常见的方法是基于向量空间模型（Vector Space Module）的方法，该方法的思想是：把文本表征成由特征项构成的向量空间中的一个点，通过计算向量之间的距离，来判定文本之间的相似程度。采用该模型的文本分类方法一般步骤是：先通过对训练语料的学习对每个类建立特征向量作为该类的表征，然后依次计算该向量和各个类的特性向量的距离，选取距离大小符合域值的类别作为该文本所属的最终类别。这种方法有了很多的应用，但是其不足之处也是显而易见的：1. 正确率一般只能达到80%，且很难进一步向上发展。2. 对于不同体材的文本，则其归类正确率更是大打折扣。

基于语义的分类方法采用全部或部份理解文本的语义而进行归类。主要可以分为以下三类：1. 基于词的归类技术。文本的语义是基于概念之上的，而词是概念的基本构成单位，从文本抽取出能反映出该文本的关键词，通过对关键词归类而进行归类，显然这种方法产生的归类其实并不是真正通过语义来进行归类，属于较早采用的技术。2. 基于知识的归类技术。基于知识库的归类技术有一个明确的知识库，知识的表示方法主要有规则库、语义模型或格框架等。基于知识的分类技术的显著特点是需要手工建造知识库，且建造的知识库领域性极强，移植非常困难。最近的研究工作表明，在一定的领域内，基于知识库的系统能够进行快速准确的分类。3. 基于概念的归类技术。基于概念的归类技术是一种介于词的分类技术和基于知识的归类技术之间的技术，它只抽取那些对文本分类有用的概念，它抽取短语周围的文本和潜在的语义概念进行文本类别的确定。基于概念的归类技术并不需要理解全文的语义，这在当前对自然语言的理解水平尚处于初级阶段的现状来说无疑是一个较好的方法。

## 文本分类算法

在文本自动分类中，分类算法是其核心问题。现有的分类算法可分为：Rocchino法及其变异方法、贝叶斯分类，K紧邻分类，支持向量机分类，决策树的方法等。

### 贝叶斯算法

### K近邻算法( KNN)

### 支持向量机

### 决策树

### 其他

## 文本分类应用

# 期望最大化算法

EM算法是DemPster[6]等提出的一种求参数最大似然估计的迭代算法，在不完全数据处理中，有着广泛的应用。它的每一次迭代由两步组成:E步（期望步），M步(最大化步)"。E-步的作用是根据某一段时间内观测到的输入数据和输出数据将该段时间内其他隐藏的状态估计出来。M-步的作用是利用E-步的结果重新估计该系统的参数[7]。

期望最大化(EM)算法属于贝叶斯方法的一种，在解决包含非完整数据的统计估计和诸如混合估计[8]等类似的问题中是一个有效的工具[9]。

# 期望最大化算法在文本分类中的应用

对贝叶斯分类器采用EM算法的基本思路是首先利用由少量有类别标注的文档组成的原始训练集合,初始化贝叶斯分类器的参数,然后利用EM算法调整贝叶斯分类器,从而进一步优化贝叶斯分类器的参数,提高其分类性能。

为了提高文本自动分类的效率和准确性，很多研究者提出了基于各种模型EM算法。其改进分别针对E-step和M-step。

文献[10]提出了基于GMM((Gaussian Mixture Model)模型的EM算法，并提出了半监督式EM重构训练集算法。将文本向量空间模型化，并用EM进行分类，用与WEB信息的抽取和分类。传统的高斯分布描述了一种围绕某个聚集分布的随机变量，是一个单模态(只有一个最大值)的分布,不能对多模态的数据分布提供一个较好的近似。高斯混合模型(GMM)将数据看作从数个高斯分布中生成出来的,GMM可以变得任意复杂，它可以通过增加Model的个数，任意逼近任何连续的概率密度分布。半督导的方式既包括标注样本又包括未标注样本,它的思路是假设了同类别的已标注样本与未标注样本在文本特征空间上的“距离”的相近程度，从而执行EM过程分类。

张博锋等人[11]提出了基于自训练的改进的EM算法STEM，每步迭代的E-step中，将中间分类器最有把握对其类别进行预测的未标注样本转移至标注样本集，并应用到M．step中进行下一个中间分类器的训练，从而引入了利用中间结果的自训练机制。类实验表明STEM算法在大部分情况下的分类准确性都高于EM，并通过减少迭代提高了分类器学习的计算效率。

文献[12]将模拟退火算法和EM 算法结合，形成 Annealing-改进 EM 算法，再将遗传算法应用到改进 EM 算法中，同时估计模型阶数和模型参数，有效克服协方差矩阵出现奇异，并将混合算法应用到聚类中，采用 UCI 数据进行仿真，并比较查准率和查全率。

2007年，Das,A.S [13]等人提出了一种并行概率隐含语义分析的EM 算法。

文献[14]提出了基于EM算法的缺失数据插补算法，用户

文本分类技术与其它技术相结合有很广泛的应用前景。主要有[15]:

1.全文检索系统

传统的信息检索系统大多是基于自由关键词和布尔模型的,这种系统的检索结果并不十分理想,因此必须发展基于中文信息处理和现代汉语研究基础的具备汉语语言理解能力的全文检索系统。这种系统除了涵盖传统检索技术及其改进技术外,还得应用中文文本分类技术以及中文自动摘要技术等最新的自然语言理解研究的成果。

2.WEB信息检索系统

WEB网页资源是一个海量的、动态的信息资源。如何从中获取信息,是当今信息研究领域中极富挑战性的工作。基于相对静态的全文检索技术不能完全满足这一应用的需要。还必须针对WEB网页特征,采用特殊的技术,以适应海量信息的检索应用。

3.信息反馈系统

如何利用计算机根据用户查询的一些信息,主动地反馈相关信息。这与信息检索系统有相似之处。

4.网络信息过滤、萃取系统

每一天,都会得到大量的网上信息,必须根据用户己经关心过的信息过滤网上信息,然后主动形成用户需要的信息。这就是网络信息过滤系统。另外,根据某种需要,自动分析网页信息,萃取某一特定内容,进行分析,这就是网络信息萃取系统。它们都将综合运用分类技术和摘要技术。

5.文本库的建立与重建

机构或个人都会面临建立文本库或重新归类大量文本的任务。这就需要根据指定的一些文本和类别结构,自动地将所有的文本归于合适的类。若是将新的文本加入合适的文本类别中也必须采用此技术。

# 总结

**References:**

[1]. 张野与杨建林, 基于KNN和SVM的中文文本自动分类研究. 情报科学, 2011(09): 第1313-1317+1377页.

[2]. 王闰强与胡铁军, 中文文本自动分类研究进展. 医学情报工作, 2002(06): 第342-345+347页.

[3]. 张冬慧等, 文本自动分类关键技术研究. 微计算机信息, 2008(06): 第197-199页.

[4]. 苏金树, 张博锋与徐昕, 基于机器学习的文本分类技术研究进展. 软件学报, 2006(09): 第1848-1859页.

[5]. 苏伟峰, 李绍滋与李堂秋, 一个基于概念的中文文本分类模型. 计算机工程与应用, 2002. 38(6): 第193-195页.

[6]. Dempster, A.P., N.M. Laird and D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological), 1977: p. 1-38.

[7]. 张欣等, 基于期望最大化算法的车载导航定位. 信息技术, 2008(07): 第85-88页.

[8]. McLachlan, G. and T. Krishnan, The EM algorithm and extensions. Vol. 382. 2007: John Wiley & Sons.

[9]. 张和发与李立萍, 含噪独立分量分析的期望最大化算法. 电子科技大学学报, 2012(04): 第527-531页.

[10]. 乔爽爽, 基于EM算法和DOM树的WEB信息抽取的研究, 2013, 大连海事大学.

[11]. 张博锋, 白冰与苏金树, 基于自训练 EM 算法的半监督文本分类. 国防科技大学学报, 2007. 29(6): 第65-69页.

[12]. 曹红丽, 混合高斯模型的混合 EM 算法研究及聚类应用 [D], 2010, 新疆大学.

[13]. Das, A.S., et al. Google news personalization: scalable online collaborative filtering. in Proceedings of the 16th international conference on World Wide Web. 2007: ACM.

[14]. 庄朋与孟凡荣, 一种基于 EM 算法的缺失数据插补算法. 2009.

[15]. 孙丽华, 中文文本自动分类的研究, 2002, 哈尔滨工程大学.