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]。

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]：

## **基于外延方法的分类方法**

这种类型的分类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**

这种类型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：

#### 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。

#### 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。

#### 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。

文献[6]提出了一个中文文本自动分类模型，它是基于概念的归类技术，建立在《知网》基础之上的，从文章中抽取出关键词，利用词产生的概念对全文进行分析，然后根据概念的类别综合进行分类。

# 期望最大化算法

EM算法是DemPster[7]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[8]。

# 期望最大化算法在文本分类中的应用

2007年，Das,A.S [9]等人提出了一种并行概率隐含语义分析的EM 算法。

# 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]. 苏伟峰, 基于概念的文本自动分类研究, 2002, 厦门大学. 第 56页.

[7]. 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.

[8]. 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.

[9]. 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.

# 标题1：二号 宋体 加粗 居中

## 标题2：三号 黑体

### 标题3：四号 宋体 加粗

#### 标题4：小四 黑体

##### 标题5：小四 宋体 加粗

点分类

* + 点分类符号1
    - 点分类符号2
      * 点分类符号3
      * 点分类符号3
    - 点分类符号2
  + 点分类符号1

文档正文：小四宋体+Times New Roman西文体，缩进2字符+1.5倍行距离，段落前后间距0行；

<文档正文：插入定义

定义 1‑1

文档正文：小四宋体+Times New Roman西文体，缩进2字符+1倍行距离。

<文档正文：插入文档正文公式

公式 1‑1 *y*=*f*(*x*)；

*u*=*g*(*v*) 式 1‑1

<文档正文：插入专业的MathType公式，之前需要安装MathType公式编辑器。例如，

在行内插入公式；但以下是独立居中成为一行，分别是无编号、左编号和右标号的公式。







文档正文：小四宋体+Times New Roman西文体，缩进2字符+1.5倍行距离，段落前后间距0行；

<文档正文：插入图的分章题注，如图表 1‑1所示。

首先建立(插入)文本框，然后在此插入图表

图表 1‑1 <此处居中插入图表的标题，如xxx系统结构>

文档正文：小四宋体+Times New Roman西文体，缩进2字符+1.5倍行距离，段落前后间距0行；

<文档正文：插入图的无章题注，如图 1所示。

首先建立(插入)文本框，然后在此插入图表

图 2 <此处居中插入图表的标题，如xxx系统结构>

文档正文：小四宋体+Times New Roman西文体，缩进2字符+1.5倍行距离，段落前后间距0行；

<文档正文：插入表格的分章题注，如表格 1‑1所示。

表格 1‑1

首先建立(插入)文本框，然后在此插入图表

文档正文：小四宋体+Times New Roman西文体，缩进2字符+1.5倍行距离，段落前后间距0行；

<文档正文：插入表格的无章题注，如表 1所示。

表 1

首先建立(插入)文本框，然后在此插入图表

文档正文：小四宋体+Times New Roman西文体，缩进2字符+1.5倍行距离，段落前后间距0行