自动文本分类中的智能处理技术

孙晋文　肖建国
（北京大学计算机科学技术研究所　北京100871）

The Intelligent Treatment Technology in Automatic Text Classification

SUN Jin-Wen　XIAO Jian-Guo
(The Institute of Computer Science and Technology, Peking University, Beijing 100085)

Abstract　Text automatic classification has become an important technology along with development of Internet and the increment of information, because of the complexity of text, it is very difficult to achieve better effect only depending on the different classification methods, it need to use multi- ways to resolve. Based on the retrospection of text classification, this paper gives a comprehensive ways to enhance the performance of text classification, which will provide good instruction to the application of text classification.

Keywords　Text, Classification, Automatic, Intelligent, Feedback

1　引言

文本自动分类技术是自然语言处理的一个重要的应用领域，是替代传统的繁杂人工分类方法的有效手段和必然趋势，特别是随着互联网技术的发展，网络成为人们进行信息交互和处理的最有效的平台，各种数字化的信息每天以极高的速度增长，面对如此巨大的信息，人工分类选择已经无能为力，计算机自动分类已成为网络时代的必然选择。通过利用先进的计算机技术、人工智能技术，不仅可以实现方便快捷的分类效果，节省大量的人力物力，并且可以进一步进行更深层次的信息挖掘处理，提高信息的利用效率。

文本分类处理的研究是计算机、信息处理领域的重要内容，特别是随着网络技术的快速发展，这种应用也变得更加迫切。

自动文本分类技术的研究最早可追溯到20世纪60年代的Maron[1]的研究工作，从那时起，该技术便逐渐应用到信息检索、文档组织、文档过滤等方面；1970年，Salon 等人提出了VSM 模型，由于该模型在良好的统计学方法基础上简明地实现了对文本特性的抽象描述，从而成为文本分类处理的一种经典模型；到80年代末，在文本分类领域，基于知识工程的方法一直占主导地位，其中最著名的是CONSTRUE[2]系统，虽然该方法取得了较好的分类效果，然而该方法具有分类规则制定困难、推广性差的缺点，很难大规模推广应用；进入90年代以来，随着互联网技术的快速发展，文档自动分类的研究也进入了一个新的阶段，各种分类方法相继得到了发展，包括机器学习技术为主的信息分类技术逐渐取代了基于知识工程的方法，成为文本自动分类研究的主要形式，如 Naive Bayes、Decision Tree、Linear Classifiers、神经网络等等[3~6]，1998年Dortmund 大学的 T. Joachims[7]探讨了支持向量机（SVM）方法进行文本分类，取得了很好的效果。此外，一些学者还采用 Boosting[8]方法来探讨提高分类处理的方法。国内，许多研究院所也对中文信息分类技术进行了大量的研究[9~11]，在具体分类算法上与国外是相同的，只是由于中文的词与词之间没有明显的分割，因此需要首先进行切词处理。

根据目前对于文本分类技术的研究，大多数研究者的精力主要放在各种不同分类的方法探索与改进上。然而，根据目前的结果表明，虽然不同的分类方法在进行分类处理时性能上确实存在一些差异，但并非是唯一因素，而且，单纯从算法上进一步提高文本分类的效果已经相当困难。事实表明，分类系统作为一个复杂系统，其它因素对分类性能的影响也是非常大的，包括文档集的选择、特征词的处理等等。对于具体文本分类技术的应用，需要从文本分类处理的多个环节着手，用综合的方法来改善和提高分类的性能。

2　文本分类的特性

文本分类的基本原理是将待处理文本集 $D=\{d_1, d_2, \cdots, d_n\}$ 按照一定的规则划分到预定义的类别 $C=\{c_1, c_2, \cdots, c_k\}$ 中的过程，其基本处理流程如图1。从具体处理上分为训练与分类两个阶段，因此，文本分类是一种有监督的学习过程，在训练阶段，需要人工提供大量的进行了类别标记的事例文档进行学习，在此之间，需要首先进行文档的向量化，即将文档用其特征组成的向量来表示。

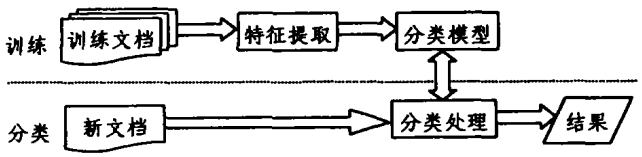

图1　文档自动分类基本流程

总体来讲，文档分类处理具有以下特点：

·文本分类需要先训练再使用，因此训练样本的质量对分类有较大影响；

·分类模型是根据训练样本而得到，因此不可避免地具有局限性。面对实际使用中样本的多样性，若系统不具有相关的

孙晋文　博士后，研究方向：人工智能、数据挖掘、网络系统集成、智能控制技术。肖建国　教授，博导，研究方向：信息处理、人工智能、网络系统集成。

自我反馈学习能力，则性能将会逐渐下降；

·文本本身具有复杂性、特征的广泛性、稀疏性等特点，使得仅仅依靠单一的分类处理模型，很难使分类处理进一步提高，必须采用多种策略加以解决；

·在分类处理上，分类准则的模糊性是其一个重要特征，因此，在分类模型中引入模糊分类处理技术将有助于分类性能的提高。

针对以上特点，结合我们在分类过程的研究，本文具体介绍进行分类处理中的一些智能解决策略，从系统整个处理的各个环节来优化分类处理，从而达到完善解决文档自动分类系统的目的。

3 智能文本分类处理策略

由于文本本身的复杂性、不规律性的特征，文本自动分类系统是一个涉及多方面综合的系统，想获得良好的文本分类效果，不仅仅是单纯的分类处理算法的问题，必须运用多种手段加以解决，特别是文档分类系统作为一个有指导的学习系统，与其它控制系统具有类似的特性，可以借鉴其它的智能控制技术加以解决。为此，根据文本自动分类处理的特点，我们给出一种文本分类系统的多策略智能解决方案(图2)，从影响分类处理的几个主要环节入手，来优化处理分类系统的流程，从而从效果上大大提高分类处理效果，为文本分类处理提供综合的解决方法。

处理上主要从以下几方面对分类系统进行了改善：

·*训练文档的优化* 从整个系统的入口环节入手，对系统进行学习的样本进行控制，提高学习样本的质量，从而为分类模型的建立提供较好的保证。

·*分类模型的运用策略* 从具体分类模式的运用上，进一步增强系统的分类效果。

·*分类系统的反馈学习* 实现系统在使用过程中不断的自我学习、自我完善，从而达到其分类性能不断提高的目的。

·*模糊分类处理* 提高分类处理的智能化，使分类处理结果更能反应文本类别的真实特征，从而达到减小误分类、提高分类精度的目的。

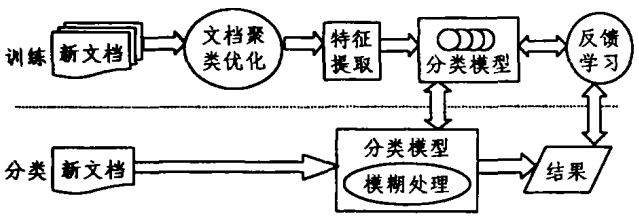

图2 智能文档分类处理

下面，对相关环节进行具体介绍：

1. 训练文档优化技术

分类系统需要先进行训练学习然后再使用，其训练过程需要大量的训练文本，因此，训练文本集越大、代表性越强，其训练效果越好，分类性能才能越高。可以想象对于存在大量不准确的分类文档的学习是很难获得良好分类效果的。1999年Yang[12]对多种分类方法在Reuters的不同版本的语料库进行的比较实验充分说明了这一点，由于Reuters2中一些未标记文本的存在，使得分类效果远远地低于其它的版本。

然而，在实际使用过程中，大量高质量的训练样本集合的获得是非常困难的，通过人工的方法对每篇文档进行筛选也不现实，一般的情况下，人们比较容易得到一些大体比较相关的文档，比如从专业资源库、互联网、新闻组等等，然而，这些资源中的文本的质量是参差不齐的，有些甚至是错误的，若直接在这样的文本上进行学习，将使分类模型的效果大大降低，因此，要实现良好的分类效果，必须对现有的训练样本进行必要的预处理，从初始的训练文档集中，选取高质量的训练文档。为实现这一目的，可以运用聚类技术对训练文档进行优化处理，实现去粗取精的目的，具体方法如下：

对于给定的初始训练文档集合 Q，$Q_i(i=1,2,\cdots,k)$分别为属于各个类别的文档，c_i 分别为各个类别文档聚类中心，分别计算各类中文档与各自聚类中心的距离，通过给定距离参数 d 来具体控制文档优化处理，将各个文档集内与聚类中心过大的文档删除，然后重新计算各个类别文档的聚类中心，循环进行；若剩余文档数量少于希望得到的文档数量，可通过调节参数 d 或重新补充新的文档来进行优化选取(图3)。

通过训练文档聚类优化，提高训练文档的类别相关性，减小噪声数据的干扰，从而为分类处理提供有效的保证。

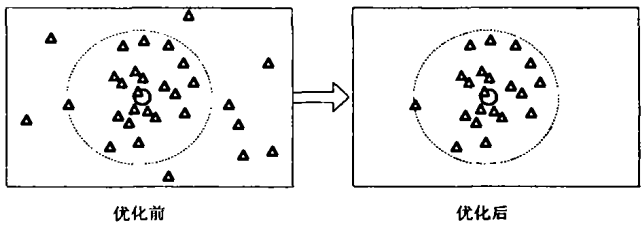

图3 文档聚类优化

2. 多模型处理技术

对于信息分类技术的研究，长期以来形成了各种各样的方法，如 Rule-based、Naive Bayesian、kNN、Decision Tree、SVM、Boosting 等，不同的方法都有各自不同的特点，是从不同的方面实现了对分类问题的描述，并且，一些简单的分类方法往往也可以达到一个较好的效果。就目前的研究来看，SVM 方法作为性能较好的分类处理方法，比其它的分类方法具有一定的优越性。但从实验结果表明[7]，其分类性能比传统的简单的分类方法，如 kNN 也并没有一个太大的提高，这是由我们所提到的文本本身的复杂性所决定的。同时，采用 Boosting 方法的试验结果表明其也可取得较好的分类效果，Boosting 的主要思想是用一些弱的分类规则实现较高分类效果的目的。因此，针对这种情况，在具体处理时，我们可以将几种不同的方法结合起来进行处理，如将支持向量机方法与基于规则的方法相结合等，使各种分类方法取长补短，互相补充，即几个不同分类器的结合，其整体分类性能将高于任何一个，从而提高分类的精度与效率。

3. 分类反馈学习

反馈是控制领域常用的技术方法，同时也是信息处理中的重要手段，通过对系统处理后的信息通过一定方式作用于原有系统，从而对原有模型进行修正，达到优化系统的性能的目的。针对文本分类处理的基本原理可以看出，在该流程中的分类处理是一个开环处理过程，无法获得处理后的结果反馈信息，特别是无法在实际使用过程中根据实际的情况进行分类模型的调整，从而使分类系统往往随着使用时间的延长、范围的扩大而性能逐渐降低，这也是制约分类系统实用化的重要因素。

而反馈技术则为弥补这一不足提供了有效的手段，通过

引入反馈学习技术，实现系统的闭环控制，从而可以使分类系统可以应对实际使用中的复杂情况。因此，信息反馈是信息处理的重要技术，反馈技术在信息处理中的应用最早是在信息检索中，通过对检索结果的反馈，从而实现更加精确的查询处理。其基本原理是，Rocchio[13]提出的相关反馈公式：

$$Q_1=Q_0+\frac{1}{n_1}\sum_{i=1}^{n_1}R_i-\frac{1}{n_2}\sum_{i=1}^{n_2}S_i$$

其中，Q_0、Q_1分别是原始查询及修正后的特征向量；R_i、S_i分别是相关文档和不相关文档的特征向量，n_1、n_2为相关、不相关文档数；后来，Ide 提出了 dec-hi 方法，实现证明这个方法对于信息处理是较好的相关反馈技术。

根据具体反馈处理的不同，在分类处理中起作用的是负反馈，即对错误分类文本的学习，而正确分类文本对反馈学习基本没有多少作用。在系统中通过对分类过程中的错误分类文档与不可分类文档进行反馈学习，不断对原有分类模型进行修正，从而实现分类性能的逐步改善提高(图4)。

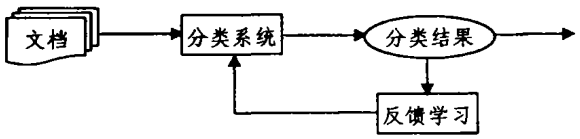

图4　文本分类系统反馈学习

4. 模糊分类处理技术

模糊性是客观事物的本质特性。在分类处理中，对于待分类的文本，都是在某种程度上属于某一个类别，而并非是绝对二值逻辑。在我们根据具体的分类模型进行分类处理时，我们得到的往往也是待分类文档属于各个类别的数值信息，在一般情况下往往是根据最大的结果数值来判定系统的分类结果，这将会丢失大量的信息，造成大量的误分类情况。而模糊处理技术正是根据事物本身模糊性的特征，在处理过程中根据模糊规则进行处理，从而更能真实地反映事物的本来面目。因此，为提高分类的智能性、准确性，在进行分类处理时，可以运用模糊处理技术，对分类结果进行模糊规则处理，即先对分类模型的分类结果进行模糊化处理，将具体的数值量转换成模糊变量，然后根据具体情况制定相应的分类处理规则，实现模糊推理(图5)。

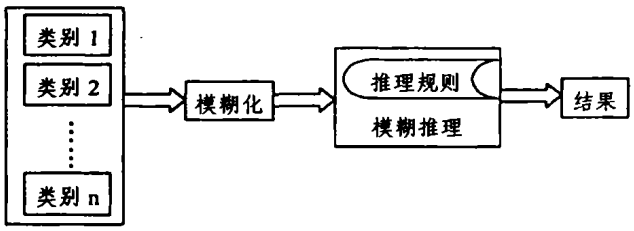

图5　模糊分类处理流程

运用模糊分类处理，也可以很好地处理文本分类中的兼类、拒类等情形。当只有属于某一个类别的可信度为高时，则该类别为其所对应分类；当同时对应于两个或多个类别的可信度都高时，则该文档可同时被分为多个类，即是兼类；而当文档所对应的所有分类的可信度皆为低时，则为拒类；从而使分类处理具有了专家分类处理时的智能性，当然也就更能准确地反应文本本身所具有的实际类别特征。

结束语　信息技术的发展，使得文本自动分类技术的应用日渐迫切，而文本本身的复杂性，使得文本分类处理不是一个简单的过程，需要通过综合的策略加以解决。本文从分类系统的特点出发，针对分类处理的多个环节，给出了综合的智能解决方案，通过该方法的实施，为进一步提高分类系统的效果，提高分类系统的实际应用性能提供了有效的途径。

参 考 文 献

1 Maron M. Automatic indexing: an experimental inquiry. Journal of the Association for Computing Machinery, 1961(3)

2 Hayes P J, Weinstein S P. CONSTRUE/TIS: a system for content-based indexing of a database of news stories. In: Proc. of IAAI-90, 2nd Conf. on Innovative Application of Artifical Intelligence, 1990

3 Lewis D D, Ringuette M. A comparison of two learning algorithms for text categorization. In: Proc. of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, US, 1994

4 Cohen W W, Singer Y. Context-sensitive learning methods for text categorization, SIGIR-96

5 Dumains S, Platt J, Heckerman D. Inductive Learning Algorithms and Representations for Text Categoriztion, in CIKM-98, 1998

6 Ruiz M E, Srinivasan P. Automatic text categorization using neural networks. In: E. Efthimiadis, ed. Proc. of the 8th ASIS/SIGCR Workshop on Classification Research, Washington, US, 1997

7 Joachims T. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In: Proc. of the European Conference on Machine Learning, Berlin, Springer, 1998. 137～142

8 Schapire R. BoosTexter: A Boosting-based System for Text Categoriztion, Machine Learning 2000, 39

9 黄萱菁，吴立德. 基于向量空间模型的文档分类系统. 模式识别与人工智能，1998(2)

10 邹涛，王继成，黄源，张福炎，等. 中文文档自动分类系统的设计与实现. 中文信息学报，1999(3)

11 范炎，郑诚，等. 用 Naive Bayes 方法协调分类 Web 网页. 软件学报，2001(9)

12 Yang Yiming. An Evaluation of Statistical Approaches to Text Categorization. Journal of Information Retrieval, 1999(1/2)

13 Rocchio J J. Relevance feedback in information retrieval. In the SMART Retrieval System-Experiments in Automatic Document Processing, Prentice Hall Inc. 1971