硕士学位论文

直推式迁移学习及其应用研究

RESEARCH ON TRANSDUCTIVE TRANSFER LEARNING AND THE APPLICATIONS

秦彦霞

哈尔滨工业大学

2012 年 7 月

国内图书分类号：TP391.2　　学校代码：10213
国际图书分类号：681.37　　密级：公开

工学硕士学位论文

直推式迁移学习及其应用研究

硕士研究生：秦彦霞
导　　师：郑德权 副教授
申请学位：工学硕士
学　　科：计算机科学与技术
所在单位：计算机科学与技术学院
答辩日期：2012 年 7 月
授予学位单位：哈尔滨工业大学

Classified Index: TP391.2
U.D.C: 681.37

Dissertation for the Master Degree in Engineering

RESEARCH ON TRANSDUCTIVE TRANSFER LEARNING AND THE APPLICATIONS

Candidate: Qin Yanxia
Supervisor: Associate Prof. Zheng Dequan
Academic Degree Applied for: Master of Engineering
Speciality: Computer Science and Technology
Affiliation: School of Computer Science and Technology
Date of Defence: July, 2012
Degree-Conferring-Institution: Harbin Institute of Technology

摘 要

传统机器学习方法从训练数据中学习得到的数据模型能够在测试数据中取得良好效果的前提是：有充足的训练数据且训练数据与测试数据同分布。然而，这种强约束性的前提往往难以得到满足。迁移学习方法有效地削弱了这种约束前提的限制，使得不同领域的知识可以被用于辅助学习目标领域模型。

直推式迁移学习方法是迁移学习方法的一种，用于解决源领域与目标领域不同时的迁移问题。针对目标领域缺乏已标注语料的问题，本文探索了一种基于 EM 的直推式迁移学习方法。该方法旨在从已标注的源领域数据中获取到迁移知识，然后借助 EM 算法将迁移知识与未标注的目标领域数据结合，以协助目标领域任务的完成。本文的主要研究工作与创新如下：

（1）探索了一种基于 EM 的直推式迁移学习模型构建方法（EMTL），该模型利用 EM 算法对隐含变量的极大似然估计能力，从已标注源领域数据中获取辅助知识，用于解决目标领域数据未标注情况下的学习任务。同时，给出了基于 EM 的直推式迁移学习方法中的迁移知识形式及其获取和应用方法。

（2）针对传统文本分类所面临的问题，研究实现了基于 EM 的文本分类迁移学习方法（EMTC），该方法是基于 EM 的直推式迁移学习方法在文本层次上的应用。该方法从源领域已标注数据中学习得到朴素贝叶斯分类器，以其模型参数作为迁移知识，辅助完成目标领域文本分类任务。基于传统中文文本分类语料构建了中文文本分类迁移学习语料库，并通过相关实验验证了 EMTC 方法的有效性，在不同领域间文本分类任务取得了较好的分类结果。

（3）探索了将迁移学习及分类的思想融入到中文术语抽取任务中的方法。基于分类的术语抽取方法主要采用分类器估计候选术语的术语度，以此为根据抽取文档中术语。同时，本文研究了基于 EM 的术语抽取迁移学习方法（EMTE），该方法是基于短语层次的直推式迁移学习方法的实现。实验证明 EMTE 方法能够解决目标领域缺乏标注语料问题，获得较理想的效果。

关键词：直推式迁移学习；文本分类；术语抽取；朴素贝叶斯；EM 算法；迁移知识

Abstract

Only on a premise that sufficient training data are available and test data have the same data distribution with training data, traditional machine learning methods can yield satisfied results. While the strongly constrainted premise is difficult to be fulfiled. Transfer learning methods can impair the constrained limitation effectively by applying the transfer knowledge obtained from a source domain to assist learning the target domain model.

As one kind of transfer learning, transductive transfer learning method is used to achieve transfer tasks when target domain is different from source domain. This thesis proposed an EM-based transductive transfer learning method to deal with problems of lacking labeled target corpus. The EM-based transductive transfer learning method aimed at achieving target task by obtaining transfer knowledge from labeled source data and making a combination with unlabeled target data with the assistance of EM algorithm. The main research work and innovations are presented as follows:

(1) explored a method to build an EM-based transductive transfer learning model, which effectively explored the estimating ability of EM algorithm on log likelihood of hidden variables and learnt auxiliary knowledge from labeled source data to achieve target tasks on the circumstances that the target data are unlabeled. Meanwhile, transfer knowledge content and the methods to obtain transfer knowledge and application methods are introduced.

(2) studied and implemented an EM-based text classification transfer learning method (EMTC) to handle the problem that cannot be solved by traditional text classification algorithms, which is an document level based version of EM-based transductive transfer learning method. The EMTC method learnt a naive bayesian classifier from source data, taking the classifier parameter as transfer knowledge to assist the text classification task on target domain. The paper built a Chinese text classification transfer corpus on the basis of traditional Chinese text classification corpus. Related experiment results show that the proposed EMTC method was effective to solve text classification task between different domains.

(3) made an research on how to apply transfer learning and classification to deal with Chinese terminology extraction task. Classification based terminology extraction method utilizes classifier estimating the termhood of extracted candidate terminologies to yield terminologies in documents. This paper proposed an EM-based terminology extraction transfer method (EMTE) which was a phrase level based transductive transfer learning method. The experiment result proved our

EMTE method can yield an satisfied terminology list even when we lack of labeled target data.

Keywords: Transductive transfer learning, Text classification, Terminology extraction, Naive bayesian, EM algorithm, Transfer knowledge

目 录

第 1 章 绪　论

1.1 研究目的和意义

计算机的问世，大大地方便了人们的生活。数百年来，研究者们也试图进一步探索计算机的智能性，以求更好的服务于社会。其中，研究者们探索的一个问题就是如何让计算机进行自我学习。20 世纪 80 年代，机器学习研究热潮在 CMU 召开的机器研讨会上拉开了序幕。机器学习研究计算机程序根据已有数据获取经验知识完成相应学习任务，并不断提高性能[1]。根据可用数据的不同，目前机器学习方法的主流方法分为三种：有指导学习（Supervised Machine Learning）、无指导学习[2,3]（Unsupervised Machine Learning）和半监督学习[4-6]（Semi-supervised Machine Learning）。

传统机器学习的流程是针对一个学习任务，在给定的充足训练数据中学习经验知识，获得数据模型，然后利用该数据模型对测试数据进行学习任务的分析。在此过程中，我们假设训练数据和测试数据具有相同数据分布，故而传统的机器学习方法才能够将训练数据中获得的经验知识成功应用于分析测试数据上的学习任务。然而，实际应用中的数据并不总是能够满足这种理想的假设。训练数据过期、大量网络新词的出现等情况都会导致假设不成立。此外，很多新领域中的任务很难得到理想规模的训练数据，而大量训练数据的标注会耗费过多的人力物力资源。这些问题都限制了传统机器学习方法的使用。迁移学习作为一种新的学习框架被提出用于解决上述难题。迁移学习[7]目标是将从旧领域中学习经验知识，帮助解决新领域中的学习任务，探索人们已经学到的知识在新任务解决中起到的作用。

日常生活中处处可见迁移学习思想的应用。人们在学会骑自行车后很容易学会骑摩托车，掌握 Perl 编程语言的人很容易学会 Python 编程语言。这些都是人们自动将已经学到的知识用于解决新任务的例子，是人类的迁移学习能力的体现。研究者们改变传统学习模式，将迁移学习引入科学研究中，尝试借助原有领域知识解决新领域任务。

基于此，本文重点研究了一种直推式的迁移学习方法，该方法首先从已标注的源领域数据中获取迁移知识，然后采用 EM 算法将迁移知识与未标注的目标领域数据结合，协助目标领域任务的完成。同时，为了验证本文所提方法的有效性，选择文本分析中的分类和术语抽取两个经典问题进行迁移应用。

1969 年，在美国国防部高级研究计划管理局（ARPA）使用 4 台计算机建

立名为 ARPAnet 的小范围网络之后，互联网萌生并迅速发展起来。互联网的产生带给人们更多的知识，并使每个人都参与到知识的建设中去。如何在这形式多样的网络知识库中快速获取人们感兴趣的信息吸引了越来越多的关注。信息抽取任务将直接从非结构化/半结构化的网络自然语言文本中抽取关键信息，并以一种结构化的形式呈现给人们，方便人们快速理解和阅读[8]。作为信息抽取技术的重要组成部分，术语自动抽取技术研究如何从专业领域语料中识别并提取该领域内的专业术语[9]。利用术语抽取技术获取的专业术语可以辅助构建、管理术语数据库；辅助机器翻译的短语对齐，提高译文质量；辅助自动文摘识别文章中的重要语句，提高文摘可读性和专业性。

当前的术语抽取方法分为两种：基于统计的方法和基于语言学的方法。基于语言学的术语抽取方法使用术语词典和规则模板，依赖于领域和语言学知识。基于统计的术语抽取方法分为两种：基于统计度量的术语抽取方法和基于机器学习的术语抽取方法。基于统计度量的术语抽取方法不局限于领域、语法信息和资源，但是不能很好的解决数据稀疏问题。基于机器学习的术语抽取方法无需专家的领域知识和语言知识，但是约束在于训练语料的构建成本和训练语料及测试语料的同构性要求。为了解决训练语料稀疏或者数据标注需要大量人力劳动的问题，本文尝试借助源领域的术语抽取知识帮助完成目标领域的术语抽取，并且期望达到至少与原来方法至少相同的效果。

文本分类技术(Text Classification)利用计算机自动将未知类别的文档划分到类别集合中，帮助人们方便有效地组织管理大量文档数据。研究如何利用计算机进行自动文本分类，也成为自然语言处理和人工智能领域中的重要课题。当前主流文本分类技术主要借助机器学习方法，从大量的训练数据获取知识，得到学习模型，并对测试数据进行分类。然而，新领域中的文本分类往往缺少充足的训练数据或出现训练数据和测试数据分布不同等问题。对此，迁移学习可以提供一种有效的解决方法。

1.2 国内外研究现状

1.2.1 迁移学习方法的国内外研究现状

自20世纪90年代被引入机器学习领域后，迁移学习方法以其高效利用已经学习到的知识解决新任务的优势，引起了研究者们的广泛关注。迁移学习方法[10-12]以各种各样的名字“Transfer Learning”、“Knowledge Transfer”、“Inductive Transfer”、“Learning to Learn”、“Life-long Learning”，出现在机器学习、数据挖掘、及其应用领域的顶级期刊会议

（ICML、ECML、AAAI、IJCAI、ACM KDD，SIGIR、ACL 等）中。迁移学习的目标是在源领域中学习源任务的知识，辅助完成目标领域中目标任务。

迁移学习方法的研究存在着三大问题：迁移什么、如何迁移和何时迁移[7]。1）迁移什么：迁移学习方法中迁移的是知识。而领域中知识有些是领域特有的，有些是领域间共享的。只有领域间共享知识才能够帮助提高目标领域中任务性能的提高，才能作为迁移知识。2）如何迁移：如何迁移主要研究采用何种学习算法将获取到的迁移知识进行领域间的迁移。3）何时迁移：在哪种情况下适合利用迁移学习完成学习任务。当源领域和目标领域完全不相关时进行迁移学习会对任务性能的提高有害无益。

根据源领域和目标领域以及源任务和目标任务是否相同，可将迁移学习分为如下三种：

（1）归纳迁移学习方法，源任务和目标任务不一致但相关。当源领域有大量标注语料时，归纳学习类似于多任务学习。当源领域没有标注语料时，归纳学习方法与自学习方法情况相似。

（2）直推式迁移学习方法，源领域和目标领域不一致但相关，源任务及目标任务相同。直推式迁移学习方法适用情况为：源领域中有大量标注语料而目标领域缺少标注语料。当源领域和目标领域的特征空间相同而概率分布不同时，直推式学习类似于领域自适应方法。

（3）无指导学习方法，源领域和目标领域不一致但相关，源任务和目标任务不一致但相关。且此时源领域和目标领域都缺乏标注语料。

迁移学习方法可按照所用迁移知识形式，分为四种：

（1）基于实例的迁移学习，假设源领域中的部分数据在更改权重之后可以被用于解决目标领域的学习问题。实例加权和重要性抽样是基于实例的迁移学习常用的从源领域中选择迁移知识的方法。

（2）基于特征表示的迁移学习，期望通过将迁移知识用一种好的特征表示方法表示来提高目标任务的性能。

（3）基于参数的迁移学习，当源任务和目标任务共享参数模型中的参数或者先验分布时，将迁移知识表示成共享的参数，从而完成目标任务。

（4）基于关系知识的迁移学习，在源领域和目标领域的数据相关的情况下，解决相关领域的迁移学习问题。

在以往的研究中，归纳学习任务一般采用基于实例的、基于特征表示的、基于参数的、基于关系知识的迁移学习；直推式学习任务通常采用基于实例的和基于特征表示的迁移学习方法；无指导迁移学习方法常用的方法为基于特征表示的迁移方法。

戴等人基于 AdaBoost 算法提出了 TrAdaBoost 算法以解决源领域和目标领域数据分布不同时的分类问题[13]。TrAdaBoost 利用两种数据进行分类：少量与目标领域数据同分布的已标注新数据；大量与目标领域数据分布不同的已标注旧数据。其中，新数据的作用是判定旧数据是否能够辅助分类。和 AdaBoost 思想相似，能够被分类器分到正确类别的旧数据的权重被增大，分错的旧数据权重被减小。该方法中，被转移的知识是旧数据本身，判断旧数据是否对分类器有辅助作用的过程就是知识转移的过程。

Jiang 等针对源领域和目标领域的特征空间相同而概率分布不同时的直推式迁移学习问题，对源领域和目标领域数据进行加权，共同学习数据模型[12]。该方法去除源领域中的有错误导向的训练实例，赋予已标注的目标领域实例高权重，已标注的源领域实例低权重，且借助预测标签的目标领域实例扩充训练实例集合大小。迁移知识是源领域中没有错误导向的数据实例。

Structural Correspondence Learning(SCL)算法[14]是基于特征表示的迁移学习方法。SCL 算法采用枢特征表示不同领域中特征间的对应关系。枢特征在不同领域的识别学习中作用相同。若不同领域中非枢特征都与同一枢特征相关，则假设这些非枢特征也是相关的，并在学习模型时被赋予相同的权重。该方法中，迁移知识是从中提取出的枢特征。

Xue 等基于 Probabilistic Latent Semantic Analysis (PLSA) 提出了一种跨领域的文本分类算法 Topic-bridged PLSA (TPLSA)[15]。TPLSA 以源领域和目标领域间的公共话题为桥梁进行知识迁移。PLSA 中的隐变量被作为话题，对应于分类问题中的类别标签。领域间的公共知识被抽取为先验知识，并以约束的形式融入到 TPLSA 模型中。其中，迁移知识为从训练数据中提取出的类别结构。

1.2.2 文本分类的国内外研究现状

文本分类技术是对未标记类别的文本进行类别标注的技术。文本分类技术在国外起源于 50 年代末，Luhn 认为词频对于类别标注具有贡献作用，首次提出基于词频的文本分类方法。70 年代，向量空间模型(Vector Space Model)被提出用于表示文本，继而成为文本表示的经典模型。90 年代以后基于机器学习的文本分类方法成为了主导的分类方法，并取得了良好效果。当前文本分类方法分为三类：基于统计的分类方法，如 Naïve Bayesian、K 近邻分类、回归模型、支持向量机等方法；基于规则的分类方法，如决策树等；基于连接的分类方法，以人工神经网络为代表。基于统计的文本方法是目前研究与应用的主要方法，尤其是 SVM 方法被广泛应用。

近年来，研究者们对于文本分类的研究主要包括特征选择技术[16]、多类文本分类[17]，网络文本分类[18]等。Feng 等人针对网络文本中的类别偏斜及特征稀疏问题，提出一个生成式概率模型，引入排除-包含隐含向量描述类别分布，并在已标注文本集合学习过程中进行特征选择[16]。文献[17]中，作者采用一个一体化异构多类别分类器解决样本失衡及类别标记交叉问题。该方法重点在于研究将什么样的多类文本分类技术整合到一体而不是探索整合技术。Chen 针对网络文本中的多种不规范短语及特征，采用文本聚类及特征聚类技术抽取出网络文本中的有意义短语，以提高网络文本分类性能[18]。

研究者们将迁移学习思想用于提高不同领域间的文本分类准确率。Pan 等[19]提出谱特征对齐算法将不同领域间的领域相关词映射到一个通用的簇集合中，以减小不同领域的评论词差异性对情感分析造成的不良影响。该方法通过对领域相关词和领域独立词进行协同聚类，更好地表示了跨领域数据。文献[20]中，Zhu 等人提出一种迁移增量式支持向量机方法以解决不断增加的语料对于文本分类的影响。该方法采用与现有增量学习方法不同的更新规则计算逆矩阵，并借助基于实例的权重调整策略保证辅助数据集合和目标数据集合间的迁移满足迁移学习目标一致性。Dai 等人提出基于 EM 的朴素贝叶斯分类方法，借助已标注语料估计初始分类器模型参数，然后将模型适用到不同分布下的未标注测试语料的分类任务中，并取得一定的实验结果[21]。

1.2.3 术语抽取的国内外研究现状

作为信息抽取的重要组成部分，术语抽取技术受到研究者们的广泛关注。术语自动抽取技术大致分为两类：基于统计的方法和基于规则的方法。

基于统计的术语自动抽取方法主要使用相关统计量衡量一个字符串是否是领域术语。常用的统计量有频率、互信息、假设检验（T 检验、卡方检验）、C-value 等。其中，频率、互信息和假设检验等是通过计算词语作为完整的一部分出现概率衡量字符串或词串结合紧凑程度，以判定候选短语是否是术语[22]。C-value 是通过边界自由度，即以一个符号串的边界上出现多种字符串的可能性大小衡量字符串的独立性[23]。

基于规则的术语抽取方法大多是针对特定领域利用语言学知识构造术语抽取规则进行术语抽取。这种方法不但领域依赖性强，而且需要大量的人工参与，且移植性差。术语抽取所用的语言学信息有术语的前缀信息、上下文统计信息以及上下文语义信息等。国外的 Justeson 等只提取前缀为名词的字符串作为术语[24]。融合了统计量和术语边界信息的 NC-value 被 Frantzi 等人用于抽取术语[23]。针对医学领域，Maynard 等人综合考虑了统计信息、上下文

的语义信息，开发了术语自动抽取系统 TRUCKS[25]。Ji 等人使用计算机领域的现有术语库衡量候选术语的术语度从而判断当前短语是否是计算机领域术语[26]。该文献分别探索了基于语义的术语抽取方法和基于语法规则的术语抽取方法，并进行了方法的融合。相关实验证明融合语义和语法的术语抽取方法能够取得更好的效果。

研究者们对于跨领域术语抽取之间的帮助研究较少，多采用基于统计的术语抽取技术，效果不太理想。Yang 提出采用基于指示词的方法一定程度上解决了不同领域间的术语抽取任务[27]。基于指示词的领域自适应术语候选抽取，利用由虚词和领域无关的实词构成的指示词标记术语边界，抽取出术语候选，从而实现了在不同领域间进行术语抽取任务的通用方法。

1.3 本文的主要研究内容

传统术语抽取方法大多使用基于统计或规则的术语抽取方法，很少采用基于机器学习的术语抽取方法。并且目前研究者们提出的迁移学习方法大多用于解决文本分类迁移问题，很少有人针对术语抽取这一任务提出有效的迁移学习策略。而本文研究的内容则是构建一个新的迁移学习方法框架，并在文本分类和术语抽取任务中进行应用，以充分验证方法的可行性。

本文针对源领域和目标领域分布相似但不同，源领域中的语料有标记，而目标领域语料无标记的迁移学习问题，提出直推式迁移学习方法解决该问题。本文所做工作的前提假设为，本文选取的不同领域之间满足进行迁移学习的条件，即“能够迁移”。对于“迁移什么”和“如何迁移”的问题，本文探索了一种新的基于 EM 的直推式迁移学习方法，首先经过机器学习方法获取源领域数据集合中的迁移知识，然后采用 EM 算法在目标领域进行隐含参数的估计，在此过程中将迁移知识进行领域间迁移，辅助目标任务的解决。不同的应用任务中迁移知识的具体形式及迁移方法均有不同。本文的研究重点在于新的迁移学习方法的构建过程，以及迁移知识的形式和迁移学习方法在不同学习任务中的应用。

1.4 论文的组织结构

本文的各章内容组织如下：

第 1 章，首先介绍本文的研究目的和意义，然后分别对迁移学习方法、文本分类任务、术语抽取任务的研究现状进行了分析，针对当前研究存在的一些问题，提出了本文的研究内容。最后，说明了本文的内容组织结构。

第 2 章，阐述了直推式迁移学习方法的相关概念，探索了几种现有的直

推式迁移学习方法。随后介绍了本文所用的基于 EM 的迁移学习方法的模型。然后，回顾了迁移学习方法中迁移知识概念及常见的几种形式，并解释了基于 EM 的迁移学习方法在文本分类任务和术语抽取任务中的迁移知识形式。最后，对本章内容进行小结。

第 3 章，首先介绍了传统文本分类技术，包括预处理、文本表示、特征选择及文本分类算法。然后着重介绍基于 EM 的直推式迁移学习方法在文本分类中的具体应用算法。并通过相关实验证明基于 EM 的文本分类迁移方法对于解决不同领域的文本分类性能效果显著。最后总结了本章内容。

第 4 章，研究了传统术语抽取方法：候选术语抽取、术语抽取算法，并提出基于 EM 的术语抽取迁移学习方法作为一种新颖的术语抽取算法抽取目标领域的术语。实验效果说明基于 EM 的直推式迁移学习方法能够在术语抽取任务中取得良好效果。最后简要总结本章内容。

第 2 章 基于 EM 算法的直推式迁移学习模型

2.1 引言

本章将首先介绍直推式迁移学习的相关概念及现有的几种直推式迁移学习方法，并将从模型建立的角度阐述基于 EM 算法的直推式迁移学习方法。此外，讲述了迁移知识的概念和不同任务中获取的迁移知识形式及知识获取方法。

根据经典的迁移学习方法分类体系，根据源任务与目标任务是否相同，可以将迁移学习方法分为三种不同形式的迁移学习，如下表 2-1 所示：

表 2- 1 迁移学习的分类

	源领域 VS 目标领域	源任务 VS 目标任务
归纳式迁移学习	相同	不同
直推式迁移学习	不同	相同
无指导迁移学习	不同	不同

其中，直推式迁移学习方法研究的是源任务和目标任务相同，源领域与目标领域不同但相关情况下的迁移学习方法。直推式迁移学习方法中，目标领域没有标记数据可用来学习目标数据模型，而源领域中有大量已标记数据可起到辅助作用。本文研究的就是源领域辅助目标领域完成同一任务的直推式迁移学习方法。首先介绍本文中提到的领域和任务的概念。

定义 2.1：领域 领域 $\mathcal{D}$ 包括特征空间 $\mathcal{X}$ 及其边缘概率分布 $P(X)$，其中 $X=\{x_1,\cdots,x_n\}\in\mathcal{X}$。特征空间不同或者相同特征空间的边缘概率分布不同，都会导致两个领域不同。

定义 2.2：任务 给定一个领域 $\mathcal{D}=\{\mathcal{X},P(X)\}$，任务 $\mathcal{T}$ 包含两部分，标记空间 $\mathcal{Y}$ 和映射函数 $f(\cdot)$，其中 $y_i=f(x_i)$ 表示对应于实例 x_i 的标记，$x_i\in\mathcal{X}$ 且 $y_i\in\mathcal{Y}$。映射函数即是任务模型，通常从训练数据中学习得到。

2.2 直推式迁移学习方法

由上文可知，直推式迁移学习方法解决的问题是目标任务和源任务相同时，目标领域数据无标记，而源领域中有大量可用已标注数据。在这种情况下，我们做出如下假设：不同领域中，相同实例的标记相同，即相同实例的标

记信息不依赖于领域。常用的迁移学习方法分为四种：基于实例的、基于特征的、基于参数的和基于相关知识的迁移学习方法，本文针对现有的直推式迁移学习中的几种迁移学习方法进行描述。

2.2.1 基于实例的迁移学习方法

基于实例的迁移学习方法主要通过更改源领域实例的权重来辅助目标领域目标任务的完成。学习目标领域模型的过程中，一般采用极小化期望风险的方法。

$$\theta^* = \arg\min_{\theta} \sum_{(x,y)\in D_T} P_{D_T}(x,y) l(x,y,\theta) \tag{2-1}$$

其中，θ为目标领域模型的参数，$l(x,y,\theta)$为θ的损失函数。由于上式中，无法获得目标领域D_T中的标记数据y，故将公式（2-1）转换为：

$$\begin{aligned}\theta^* &= \arg\min_{\theta} \sum_{(x,y)\in D_S} \frac{P_{D_T}(x,y)}{P_{D_S}(x,y)} P_{D_S}(x,y) l(x,y,\theta) \\ &= \arg\min_{\theta} \sum_{(x_i^s,y_i^s)\in D_S} \frac{P_{D_T}(x_i^s,y_i^s)}{P_{D_S}(x_i^s,y_i^s)} P_{D_S}(x_i^s,y_i^s) l(x_i^s,y_i^s,\theta)\end{aligned} \tag{2-2}$$

由上式可知，若对源领域数据重新设置权重，则能够用于学习训练目标领域模型。由概率公式可知，$P(x_i^s,y_i^s) = P(y_i^s \mid x_i^s)P(x_i^s)$，则权重可以表示为：

$$w(x_i^s) = \frac{P_{D_T}(x_i^s)}{P_{D_S}(x_i^s)} \tag{2-3}$$

文献[28]通过匹配再生核希尔伯特空间中源领域数据和目标领域数据的平均值表示权重$w(x_i^s)$。该方法能够有效的避免小数据集上的常见的源领域实例在源领域或目标领域中的密集型评估问题。

文献[29]中，Suyiyama 提出基于 Kullback-Leibler 的重要性评估程序（KLIEP）采用 K-L 参数评估实例的重要性，以表示权重$w(x_i^s)$。该方法能够在选择模型时进行交叉验证。Bickel 等在前人基础上，将 KLIEP 中的评估权重和训练模型步骤融合在一起，提出一个内核逻辑回归分类器[30]。

2.2.2 基于特征的迁移学习方法

基于特征的迁移学习方法不对源领域数据的$w(x_i^s)$权重进行直接估计，而是获取领域间共同特征或类别结构等信息进行迁移。

Blitzer 提出的结构化对应学习方法[14]（SCL）即是一种典型的基于特征的迁移学习方法。枢特征是指在不同领域的识别学习中作用相同的特征。SCL 算

法采用枢特征来表示不同领域特征的相关性。枢特征可以解释成不同领域间的共享知识的存在方法，源领域及目标领域均将知识映射到枢特征上。不同领域中和相同枢特征相关的非枢特征被假设是相关的，在学习数据模型时被赋予相同的权重。虽然研究者们验证了枢特征对于目标任务模型的学习具有较好的辅助作用，但是却很难获取到。研究者们也尝试了很多不同的方法试图获取较好的枢特征[31]。

跨领域的文本分类算法 TPLSA[15]以概率隐含语义分析 PLSA 方法为基础，使用公共话题建立源领域和目标领域之间知识迁移的桥梁。以 PLSA 中的隐变量，即分类问题中的类别标签为话题。TPLSA 方法借鉴半监督聚类方法中提出的“必相关”和“必无关”约束，分别表示两个实例必定具有相同的类别标签以及两个实例的类别标签必定不相同。并从领域中获取这种约束以融入 TPLSA 模型。

2.3 基于 EM 的直推式迁移学习模型

本文研究的迁移问题是目标领域语料未标注的情况下的领域间迁移。因此，无法获取目标领域数据的标记信息进行数据模型的学习。EM 算法以其独特的对隐含参数寻求似然估计值的能力被众多研究者们推崇。本文也将充分利用 EM 算法的优势构建一个基于 EM 的迁移学习方法框架，框架图如图 2-1 所示。

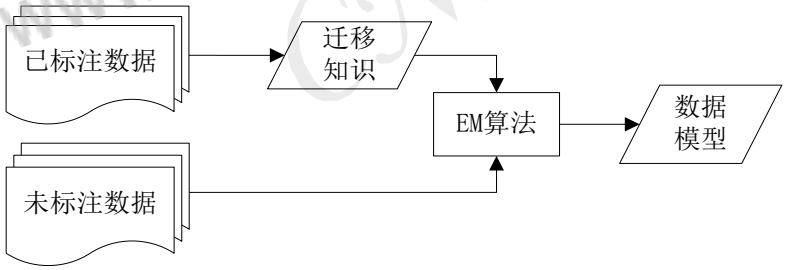

图 2- 1 基于 EM 的直推式迁移学习方法框架图

2.3.1 EM 算法

Dempster 等于 1977 年首次提出 EM（Expect-Maximum）算法，解决了数据缺失情况下的参数极大似然估计问题[32]。此后，EM 算法作为一种简单高效的迭代算法，被频繁用于对隐含参数进行极大似然估计。现对 EM 算法原理进行详细描述。

首先，我们用 D_W 表示可观察数据集合，D_U 表示隐含数据集合，则 $D = D_W + D_U$ 代表全部可用数据。记 h 为针对隐含数据 D_U 做出的假设，在 EM 算法开始时对其进行随机赋值。h'表示 EM 算法每次迭代后产生的新假设。

$P(D|h')$表示在已知假设 h'时计算得到的数据集合 D 的似然统计值。EM 算法寻找使得 $E[\log P(D|h')]$达到最大值的 h'作为 h 的最大后验估计值。EM 算法每次迭代分为两步，Expect 步和 Maximum 步。

（1）Expect 步：根据当前假设 h 和可观察数据 D_W 来估计全部可用数据 D 上概率分布以计算 $Q(h'|h)$。

$$Q(h'|h) = E[\log P(D|h')|h, D_W] \tag{2-4}$$

（2）Maximum 步：将假设 h 替换为 h'。

$$h = \arg\max_{h'} Q(h'|h) \tag{2-5}$$

重复以上迭代过程直到函数 Q 收敛到 $P(D|h')$的一个固定点。

本文拟采用 EM 算法估算迁移学习问题中目标领域未标注数据的所属标记信息。

2.3.2 迁移学习模型的构建

本文将要解决的迁移学习问题可描述为：已知一个已标注数据集合 D_l 和一个未标注数据集合 D_u，已标注数据集合与未标注数据集合数据分布不同但相关，分别服从数据分布 Ω_l 和 Ω_u。在这种情况下，迁移学习的任务是从已标注数据集合中获得迁移知识，用于学习得到未标注数据集合上的任务假设。本文的迁移学习问题可通过建模表示为：

$$h_{MAP} = \arg\max_{h} P_{\Omega_u}(h) \cdot P_{\Omega_u}(D_l, D_u|h) \tag{2-6}$$

其中，h_{MAP} 为最大后验假设，$P_{\Omega_u}(h)$ 为在数据分布 Ω_u 下，后验假设 h 的概率估计，$P_{\Omega_u}(D_l, D_u|h)$ 为给定后验假设 h 后，标注数据集合 D_l 与未标注数据集合 D_u 在数据分布 Ω_u 下的共现概率估计。最优后验假设是在同时考虑 D_l 和 D_u 的情况下计算得到的。值得提出的是，本文的迁移学习问题重点在于找到数据分布 Ω_u 下的最优后验假设，因此忽略对数据分布 Ω_l 的探索，以下所有概率计算均是以 Ω_u 为数据分布。对公式（2-6）中的 $P_{\Omega_u}(h) \cdot P_{\Omega_u}(D_l, D_u|h)$，为避免直接求最大值，我们求其对数似然最大值。于是公式转化为寻求 $l(h|D_l, D_u)$ 的最优估计。

$$\begin{aligned} l(h|D_l, D_u) &= \log P_{\Omega_u}(h) \cdot P_{\Omega_u}(D_l, D_u|h) \\ &= \log P_{\Omega_u}(h) + \log P_{\Omega_u}(D_l, D_u|h) \\ &= \log P_{\Omega_u}(h) + \sum_{d \in D_l} \log P_{\Omega_u}(d|h) + \sum_{d \in D_u} \log P_{\Omega_u}(d|h) \end{aligned} \tag{2-7}$$

EM 算法能够解决具有隐含参数情况下的参数极大似然估计问题，对应于

本文的迁移学习问题，隐含参数则为未标注数据集合 D_u 中实例的标注信息。故基于 EM 的直推式迁移学习方法旨在借助 EM 算法的参数估计能力，对 D_u 的标注信息进行极大似然估计，即寻求最优后验假设。

基于 EM 的直推式迁移学习方法的算法描述为：

输入：类别集合 $CSET=\{c_1,c_2,\cdots,c_k\}$，训练文本集合 $D_l=\{(d_1,c_1^{'}),(d_2,c_2^{'}),\cdots,(d_m,c_m^{'})\},c_i^{'}\in CSET$，与 D 不同分布但相关的未标注测试文本集合 $D_u=\{d_1,d_2,\cdots,d_m\}$，最大迭代次数 $MAXI$。

输出：测试文本集合文本的所属类别

步骤：

1、初始化参数阶段：

从已标注数据集合 D_l 中学习，得到相关初始假设概率。

2、EM 迭代阶段：

将当前假设概率应用于数据分布 Ω_u 中，迭代获得 Ω_u 中的最优假设。迭代次数小于 $MAXI$，单次迭代过程为：

（1）Expect：通过当前假设概率计算隐含变量值，即为数据分布 Ω_u 下的未标注数据集合 D_u 的标注信息；

（2）Maximum：采用计算得到的 D_u 标注信息计算新的假设，替换原假设。

图 2-2 基于 EM 的直推式迁移学习模型（EMTL）算法流程图

2.4 迁移知识获取技术

2.4.1 迁移知识

迁移知识是迁移学习方法三大问题中的“迁移什么”的答案。在各种各样的迁移学习方法中，迁移知识也各有不同。迁移知识既可以是源数据本身，也可以是从源数据中提取出来的特征或者类别结构信息，还可以是根据源数据训练得到的学习模型。不同的迁移知识对应着不同的迁移学习方法。基于实例的迁移学习方法中迁移的是更改权重后的源数据，基于特征的迁移学习方法迁移的是源数据中提取到的相关特征及类别结构信息，而基于参数的迁移学习方法中迁移知识是从源数据中学习到的任务模型。

基于 EM 的直推式迁移学习方法首先从源领域中获取迁移知识，即为源领域模型中的参数，然后采用 EM 算法对迁移知识进行迭代更新，应用于目标领域模型的建立。下面，本文将介绍基于 EM 的直推式迁移学习方法在不同的学习任务中迁移知识的形式及获取方法。

2.4.2 文本分类中的迁移知识获取

针对文本分类问题，基于 EM 的文本分类迁移学习方法采用朴素贝叶斯分类器计算初始模型参数值，然后采用初始模型参数值进行逐步迭代计算目标领域模型隐含参数，并更新模型参数值。迁移知识为模型参数值。朴素贝叶斯分类器的模型参数为数据集合中类别概率 $P(c)$及某一类别中特征词的概率值 $P(w|c)$。隐含参数为目标领域文本属于某一类别的概率 $P(c|d)$。

迁移知识通过朴素贝叶斯分类器计算得到。朴素贝叶斯分类器根据源领域数据学习模型，获得的参数值即为迁移知识。具体模型参数为类别概率 $P(c)$和特征词概率 $P(w|c)$，计算公式见本文的 3.3.1 节。

2.4.3 术语抽取中的迁移知识获取

基于 EM 的术语抽取迁移学习方法采用分类思想完成术语抽取。与文本分类相似，该方法采用的迁移知识亦为从已标注的源领域数据中学习得到的模型参数，隐含变量为文本中的候选术语的术语度。但与文本分类不同的是，文本分类迁移任务以文本为处理对象计算其概率值，术语抽取迁移任务中的迁移知识则以文本中出现的候选术语即短语，为处理对象计算其属于术语类别的概率。迁移知识具体形式表现为以下模型参数：术语出现在数据集合中的概率 $P(yes)$、术语集合中候选术语 t 的概率 $P(t|yes)$、非术语出现在数据集合中的概率 $P(no)$ 和 $P(t|no)$。隐含变量为：候选术语属于术语类别的概率 $P(yes|t)$、候选术语 t 不属于术语类别的概率 $P(no|t)$ 及候选术语的术语度 $P(t)$。

基于 EM 的术语抽取迁移学习方法中的迁移知识也是通过朴素贝叶斯分类器计算得到的。模型参数即为上述概率值 $P(yes)$、$P(t|yes)$，$P(no)$和 $P(t|no)$，具体计算方法见本文的 4.4.1 节。迁移知识应用于解决术语抽取的方法为采用 EM 算法迭代更新模型参数值。

2.5 本章小结

本章首先对直推式迁移学习问题进行了描述，并介绍了几种常用的直推式迁移学习方法。然后针对本文研究的迁移学习问题，构建了新的迁移学习模型：基于 EM 的直推式迁移学习模型。随后，本文讲述了迁移知识的几种常见形式，以及文本分类和术语抽取两种应用任务中的迁移知识形式及知识获取方法。

第 3 章 迁移学习方法在文本分类中的应用

3.1 引言

文本分类技术是根据文本内容对未知类别的文档标记类别的方法。利用计算机进行文本自动分类，帮助人们方便有效地组织管理大量文档数据，成为工程应用领域中的重要研究内容。目前的文本分类技术主要借助机器学习方法，从大量的训练数据中获取知识，得到数据模型，并对测试数据进行分类。然而，新领域的文本分类往往缺少充足的训练数据或训练数据和测试数据分布不同。本文将基于 EM 的直推式迁移学习方法应用于文本分类中，研究了一种文本层次的基于 EM 的文本分类迁移学习方法解决文本分类迁移学习中相关约束限制。

3.2 传统文本分类方法

传统文本分类方法的流程包括两个阶段：训练阶段和测试阶段。训练阶段和测试阶段均包括 4 个模块。其中，共用模块为数据预处理、文本表示和特征选择。训练阶段还包含构建分类器模块；测试阶段包括文本分类模块。流程图如下所示：

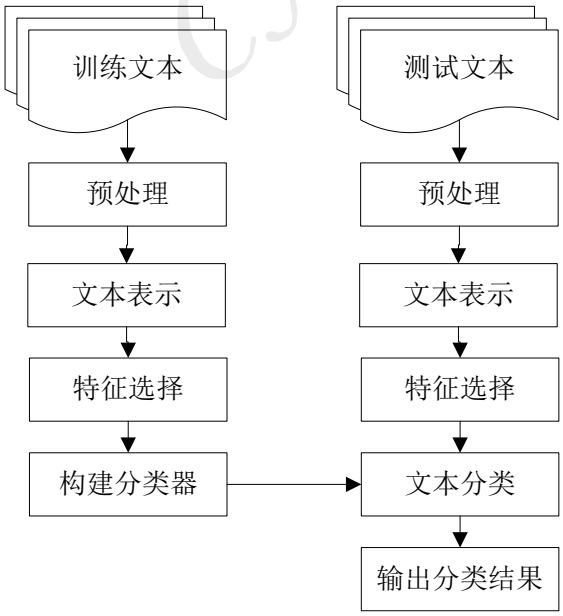

图 3- 1 文本分类流程图

下面本文将从预处理、文本表示、特征选择、文本分类算法四个方面对传

统的文本分类方法进行系统的描述。

3.2.1 预处理

一般来说，语料的获取途径为互联网、期刊或报纸，获得的文本形式多样化。文本分为三种：自由文本、半结构化文本和结构化文本。其中，文本分类的处理对象大多为自由文本和半结构化文本，这两类文本包含若干对文本分类具有负面混淆影响的噪音文字及符号。因此需要首先采用预处理技术对原始文本进行相关处理：分词、停用词过滤。

（1）分词

文本由段落、句子和字组成。英文中字称为单词，单词之间由空格分隔，每个单词能够作为独立的语义单元。中文作为独特的汉藏语系，每个中文句子是一个连续的字序列，字一般不能表达具体的语义，而由多个字组成的词语能够表征独立的含义。因此，首先需要对中文文本进行分词，由字序列转换为词序列。国内外分词效果较好、应用较为广泛的为斯坦福大学开发的中文分词器[1]和中科院开发的分词器[2]等。

（2）停用词过滤

文本中包含大量的助词、介词等虚词以及标点符号，并与实词一起组成有意义的句子。虚词和标点符号不代表具体含义，对于类别信息没有区分作用，因此对于文本分类无辅助作用。而且，虚词在文本中所占的比例较大，过滤掉这些虚词和标点符号后，能够减小文档大小，使得计算机更快更准确的进行类别标注。本文采集常用的助词、介词等虚词以及标点符号等形成停用词表。分词之后过滤掉停用词表中出现的停用词。

特殊地，对于从互联网上获取的半结构化网络文本，还需要去除<html>，<body>，<body>等格式符，只保留文本内容。在对原始文本进行上述预处理之后，可以将每篇文本表示为一系列较有意义的词语的集合。这些词语称为特征词。采用文本表示模型能够将这些特征词组织形成特征向量形式，以方便计算机进行下一步处理。

3.2.2 文本表示

文本自动分类的处理对象是预处理后的训练文本集合和测试文本集合。将预处理之后语料文本表示为计算机可识别形式是对文本建立分类模型的基础。常用的文本表示模型有布尔模型[33]（Boolean Model）、向量空间模型[34]（Vector

[1] 斯坦福分词器, http://nlp.stanford.edu/software/segmenter.shtml
[2] 中科院分词器, http://ictclas.org/

Space Model）和概率模型[35]（Probabilistic Model）。下面将介绍前两种文本表示模型。

（1）布尔模型

布尔模型中，每个文本的特征向量的基本元素是“特征对”，特征对的形式为（特征词：特征词权重）。特征词可采用原始语言形式，也可以采用编号表示。特征词的权重为 0 或 1，0 表示该特征词未出现在当前文本中，1 表示该特征词出现在当前文本中。例如，（机器：1）表示该文本中出现了“机器”一词，（神经元：0）表示“神经元”一词未出现在文本中。布尔模型简单，适合快速分类。但该模型对特征的表示能力差，只能表示特征词是否出现，遗漏了特征词的出现频率等信息。

（2）向量空间模型

20 世纪末 Salton 提出了著名的向量空间模型表示文本，其思想是：统计文本集合中出现的全部词语数，记为 n。以每个词语作为一维向量，构建 n 维向量空间。n 维向量空间中的一个点代表了原来的一篇文本。模型表示如下所示：

$$CorpusMatrix=\begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{m1} & w_{m2} & \cdots & w_{mn} \end{bmatrix} \quad (3\text{-}1)$$

其中，*CorpusMatrix*为文本矩阵，每一行代表文本集合中的一篇文本，每一列代表文本集合中出现的一个特征词。w_{ij} 为第 j 个特征词在第 i 篇文本中所占的权重，n 为全部特征词数，m 为文本集合中文本数量。若 w_{ij} 为 0，说明特征词 j 未出现在文本 i 中。由于 n 远远大于每个文本中出现的特征词数，故该文本矩阵是一个包含大量 0 的稀疏矩阵。为了节省存储空间减少运算量，对稀疏文本矩阵进行压缩。文本表示过程中只记录出现在当前文本中的特征词的权重，并加入特征词 ID 号进行区分。

向量空间模型在布尔模型的基础上选用更复杂的特征权重计算方法来衡量特征词对于所在文本的贡献度。常用的权重计算方法为：TF、TFIDF、熵权重等。本文使用 TF-IDF 方法计算特征的权重值。

（a）TF 权重

一般来说，一个词在文本中出现的次数和它对文本的重要程度成正比关系。词频 TF（Term Frequency）将特征词的权重表示为该特征词在当前文本中的出现频率。文本 doc_j 中特征词 $term_i$ 的 TF 值计算如公式（3-2）所示：

$$TF(term_i, doc_j) = \frac{freq(term_i)}{|doc_j|} \quad (3\text{-}2)$$

$freq(term_i)$ 表示文本 doc_j 中特征词 $term_i$ 的出现次数。相比布尔模型的权重表示方法，TF 权重增加了特征词的出现频率信息以更准确地衡量特征词对文本的贡献度。而且 TF 权重计算方法简单有效，是文本表示时常用的权重表示方法。

（b）TF-IDF

TF 指特征词的词频，计算方法如公式（3-2）所述。IDF（Inverse Document Frequency）为特征词的逆文档频率，特征词 $term_i$ 的 IDF 值计算方法如下：

$$IDF(term_i) = \log_2 \frac{M}{numOfDoc(term_i)} \quad (3\text{-}3)$$

$numOfDoc(term_i)$ 指包含特征词 $term_i$ 的文本个数。M 为文本集合中文本个数。

研究者认为，一个词对文本类别分析的贡献度与出现次数成正比，但是若一个特征词出现在所有文本中，则该特征词对文本的类别信息没有任何贡献，即我们无法通过该特征词判定文本的所属类别。特征词的文档频率和特征词对文本类别分析的贡献度成反比。将特征词的 TF 和 IDF 值组合形成 TF-IDF 值能够同时考虑特征词的频率和普适度信息，计算公式如下：

$$TF-IDF(term_i, doc_j) = TF(term_i, doc_j) * IDF(term_i) \quad (3\text{-}4)$$

3.2.3 特征选择

特征选择的目标是将经过上述处理后的庞大的特征集合进行进一步的缩减处理。尽管去除停用词之后减少了特征词的数量，但是特征集合中依然存在大量的特征词，其中很多特征词对于文本的类别信息表征能力非常小甚至具有混淆作用。因此，利用特征选择技术从特征集合中选择出少部分具有较好表征力的特征能够使得文本分类更准确更高效。特征选择技术不仅用于文本分类，还对其他应用任务具有较好的辅助作用。常用的特征选择方法有：文档频率[36]、互信息、信息增益[37]、卡方统计[38]、交叉熵等。

（1）文档频率（DF）

文档频率 DF 表示包含特征词的文本个数，特征词 $term$ 的 DF 值计算如下：

$$DF(term) = \frac{numOfDoc(term)}{M} \quad (3\text{-}5)$$

其中，M 是文本集合的规模大小。DF 太小，限制了特征词的频率大小，对文本类别信息无较好表征作用；DF 太大，特征词频繁出现在不同类别的文本

中，对于文本类别信息无辨别作用。用 DF 值进行特征选择时，首先需要设定一个阈值，特征词的 DF 值在阈值范围内则保留，否则过滤掉 DF 值太大或者太小的特征词，以达到特征降维的目的。DF 特征选择方法时间复杂度小，适合大规模文档集合的特征选择。

（2）互信息（Mutual Information）

互信息（MI）是衡量两个变量相关度的统计量。其一般形式为，

$$MI(a,b)=\log\frac{p(a,b)}{p(a)p(b)} \tag{3-6}$$

a 和 b 代表两个变量，$p(\cdot)$ 表示变量的概率。若变量 a 和 b 相关，则 a 和 b 共同出现的概率较大，互信息值就大；若不相关，则互信息值小。

特殊地，对于文本分类，特征词和类别的互信息值能够表征特征词对于类别的贡献度。特征词和类别的互信息计算如下：

$$MI(t,c)=\log\frac{p(t,c)}{p(t)p(c)} \tag{3-7}$$

特征词的平均互信息值在一定程度上衡量了对类别的区分能力。计算公式如下：

$$MI(t)=\sum_{c\in CSET}MI(t,c) \tag{3-8}$$

上式中，$CSET$ 是要归类到的类别集合。

（3）信息增益（Information Gain）

信息增益[37]（IG）定义为特征词出现前后类别的信息熵变化量。信息论中，将信息量定义为不确定性因素的多少。获得信息意味着不确定性减少，信息量也相应减少。信息熵是衡量随机事件的不确定性的统计值，在随机事件发生前后，信息熵值不同。

针对文本分类问题，特征词是否出现在类别 c 中是一个随机事件，信息增益可表示为：

$$IG(t)=H(CSET)-H(CSET\,|\,t) \tag{3-9}$$

$H(CSET)$ 是类别集合 $CSET$ 的信息熵，$H(CSET\,|\,t)$ 是特征词 t 出现在类别集合 $CSET$ 后的信息熵。计算方法分别如公式（3-10）、（3-11）所示：

$$H(CSET)=-\sum_{c\in CSET}p(c)\log p(c) \tag{3-10}$$

$$\begin{aligned}H(CSET\,|\,t)=&-p(t)\sum_{c\in CSET}p(c\,|\,t)\log p(c\,|\,t)\\&-p(\bar{t})\sum_{c\in CSET}p(c\,|\,\bar{t})\log p(c\,|\,\bar{t})\end{aligned} \tag{3-11}$$

将公式（3-10）、（3-11）带入公式（3-9）得 IG 的计算公式为：

$$IG(t)=\sum_{c\in CSET}(p(c,t)\log\frac{p(c,t)}{p(c)p(t)}+p(c,\bar{t})\log\frac{p(c,\bar{t})}{p(c)p(\bar{t})}) \quad (3\text{-}12)$$

$p(\cdot)$表示事件发生的概率，$p(c,\bar{t})$表示特征词 t 未出现在类别 c 的联合概率，$p(c,t)$为特征词 t 出现在类别 c 的联合概率。

（4）卡方统计量（CHI Statistics）

卡方统计量（CHI）衡量的是一个特征与一个类别的相关程度，统计值越高，相关性越大。特征词 t 的卡方统计值计算如下：

$$\chi^2(t)=\sum_{c\in CSET}P(c)\chi^2(t,c) \quad (3\text{-}13)$$

其中，$\chi^2(t,c)$表示特征词 t 与类别 c 的相关度统计值，计算如公式（3-14）所示：

$$\chi^2(t,c)=\frac{NUM\times(N_1\times N_4-N_2\times N_3)}{(N_1+N_2)\times(N_1+N_3)\times(N_2+N_4)\times(N_3+N_4)} \quad (3\text{-}14)$$

上式中，NUM 是特征集合中特征总数，N_1 代表包含特征 t 而且类别为 c 的文档数，N_2 为包含特征 t 但不属于类别 c 的文档数量，N_3 是不包含特征 t 类别为 c 的文档集合大小，N_4 代表不包含特征 t 而且不属于类别 c 的文档数。

计算得到特征集合中特征的统计值后，按照卡方统计量进行排序，筛选出特征总数的一定比例作为文本分类任务中所用的特征集合。本文采用卡方统计量对原始特征集合进行特征选择。选择阈值为 0.5。

3.2.4 文本分类方法

经典的分类体系将文本分类方法划分为三类：1）以朴素贝叶斯分类[39]、K 近邻分类[40]、支持向量机分类[41]、最大熵模型分类等为代表的基于统计的文本分类方法；2）以决策树分类方法[42]、关联规则分类方法为代表的基于规则的分类方法；3）以人工神经网络[43]为代表的基于连接的分类方法。下面对一些将常用分类算法进行简要介绍。

（1）K 近邻

K 近邻（K-Nearest Neighbor）分类是基于统计的分类方法，其基本思路是，在 N 维特征向量空间中寻找与待分类的文本距离最近的 K 个训练文本，待分类文本所属类别由这 K 个最近邻文本所在的类别决定。

一般来说，文本距离由欧式距离计算得到。文本 $d_1=(t_{11},t_{12},\cdots,t_{1n})$ 和文本 $d_2=(t_{21},t_{22},\cdots,t_{2n})$ 的欧氏距离计算如下：

$$\mathrm{dis}(d_1,d_2)=\sqrt{\sum_{i=1}^{n}(t_{1i}-t_{2i})^2} \quad (3\text{-}15)$$

记与待分类文本距离最近的训练文本集合为 $DOC=(d_1,d_2,\cdots,d_k)$，所属类别集合为 $CLASS=(c_1,c_2,\cdots,c_k)$，$c_i$ 是预定义类别集合中的类别。待分类文本的类别为 K 个近邻文本的类别加权，计算公式为：

$$Category(d)=\arg\max_{c}\sum_{d_i\in DOC,\ c_i=c} dis(d,d_i) \qquad (3\text{-}16)$$

K 近邻文本分类方法的重点在于近邻文本数量 K 的选择。K 的不同直接影响分类结果，使得分类结果具有较高的偶然性。K 的选择通常采用开发集进行确定。K 近邻分类方法并不需要训练复杂的分类模型，简单方便。但当训练文本集合和测试文本集合较大时，分类效率会有大幅度下降。

（2）支持向量机

支持向量机（Support Vector Machine，SVM）统计分类方法以其独特的优势，自产生以来一直被广泛应用于众多领域。支持向量机分类方法将数据表示成向量形式，建立分类面将线性可分的正反例样本数据分隔开来，并寻求使得分隔间距最大的最优分类面。若样本数据线性不可分，则采用核函数将数据映射到更高维向量空间，使其间接地成为线性可分数据。线性可分数据的 SVM 几何解释图如图 3-2 所示：

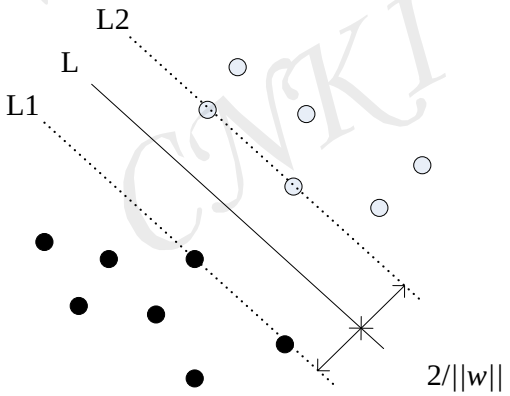

图 3- 2 SVM 模型几何解释图

图 3-2 中，实心、空心圆点分别表示 N 维空间中的正反例数据。L1 和 L2 为一对相互平行的边界线，L 为平行于 L1、L2 的最优分类线，且使得 L1 和 L2 之间的分类距离最大。L1 和 L2 上训练数据称为支持向量。支持向量机的根本目标就在于寻求 L。

支持向量机的优势表现如下：具有强大的数学理论支持，能够表示为简单的数学形式，几何解释直观易懂；对大样本数据的良好支持；较好地解决非线性分类问题；能够获得全局最优解。支持向量机的劣势在于核函数参数寻优过程过于耗时。

台湾大学的林智仁等开发设计操作简单的通用 SVM 软件包-LibSVM 工具

包[3]，提供了多种核函数，能够简单方便的完成分类、回归或估计分布，方便了研究者们对 SVM 的深入研究。

（3）决策树

决策树（Decision Tree）是应用最广泛的基于规则的分类方法之一。决策树分类算法将若干个由实例中获得的 IF-Then 规则组织形成树状结构，能够方便地将决策树表示为析取表达式的形式。决策树最常用的样例为，根据周六的天气情况判断“周六上午是否适合打网球”[44]。“是否适合打网球”样例如下图所示：

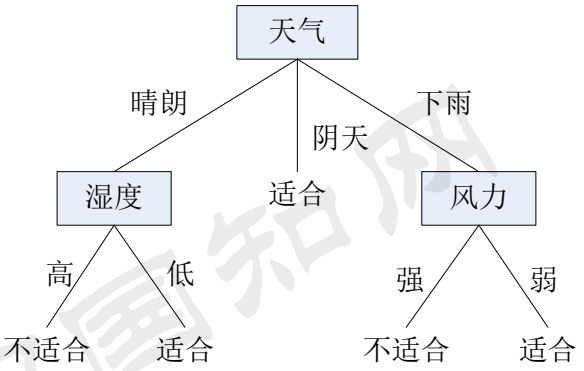

图 3-3 “是否适合打网球”的决策树

决策树分类算法的实例数据表示为属性对形式，即“（属性：属性值）”，其中属性值取值为少量离散值。决策树的重点问题在于如何选择最佳的分类属性。最常用的方法是使用信息增益作为衡量某个属性对于类别分析的贡献。该算法的优点在于对于错误训练数据具有较好的健壮性：能够允许部分训练实例的类别标注错误；允许训练数据部分属性值缺失或者属性值错误。

（4）人工神经网络

人工神经网络，Artificial Neural Network（ANN），是常用的基于连接的分类方法。人工神经网络通过对生物神经系统运行机制进行简单模拟实现分类。人工神经网络采用结点模拟处理神经信号的神经元，结点处理过程为接收信号、运算处理、输出信号，其中接收信号和输出信号都由结点之间的连接通道完成。结点和连接通道共同形成一个网状结构，并行处理信号。人工神经网络分为输入层、隐含层、输出层，其中输入层的结点接受输入数据，通过连接通道映射到隐含层的结点，多层隐含层之间亦进行相同步骤的运算，最后一层隐含层通过连接通道映射到输出层，得到输出结果。人工神经网络模型如图 3-4 所示：

[3] LibSVM 工具包, http://www.csie.ntu.edu.tw/~cjlin/

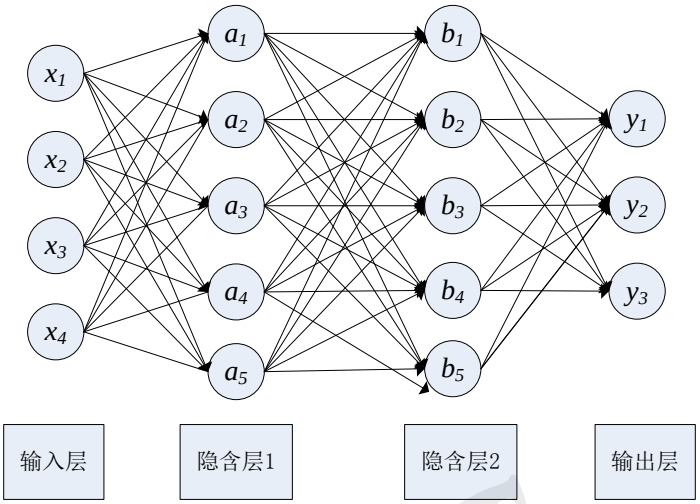

图 3- 4 4 层人工神经网络模型结构图

上图描绘出的人工神经网络具有两层隐含层，隐含层结点数均为 5。隐含层结点数需要在构建人工神经网络模型时由人工指定。输入层结点数为 4，表示输入数据中的特征数，输出层结点数为 3，说明为三类文本分类问题。相邻的每两层结点之间的连接通道具有一定的权重参数。

人工神经网络处理的文本分类问题是非线性文本分类问题。分类效果较好，但缺点是计算量大存在过拟合问题以及只能获得局部最优解。

3.3 文本分类中的迁移学习方法

目前在文本分类任务中，迁移学习的研究内容包括 Web 文本分类，垃圾邮件过滤及文本情感分析等。

（1）Web 文本分类中的迁移学习方法

众所周知，互联网的快速发展使得 Web 信息日新月异，种类繁多，形式多样化，更新速度非常快。随着时间的推移，即使对于形式较为单一的 Web 文本，其数据分布也会不断变化。而且，不同领域的 Web 文本的数据分布差异性更大。然而，在对新领域 Web 文本进行分类时，丢弃大量过时已标注数据或者其他领域已标注数据是非常可惜的。Web 文本分类中的迁移学习方法能够使得这些过时及旧领域辅助新的文本分类任务。文献[45]认为两个分布不同但相关的领域必定有共享的概念空间存储领域相关性知识，而好的特征表示方法能够解读共享概念空间。该文献采用领域自适应方法最小化领域间分布差距以及新领域的经验损失以获得领域间共享概念空间的参数。

（2）垃圾邮件过滤中的迁移学习方法

电子邮件已经成为网络用户常用的沟通交流工具，但是随之而来的是日益

严重的广告邮件等垃圾邮件。垃圾邮件过滤通常采用分类的思想解决：将邮件分为到合法邮件类别和垃圾邮件类别中。然而，很难得到一个适用于所有人的垃圾邮件分类模型，因为每个人具有不同的各种属性，比如爱好、兴趣、年龄、专业等。因此不同人的垃圾邮件过滤模型需要借助迁移学习辅助完成。文献[46]采用基于互信息的方法衡量垃圾邮件分类前后样本和特征信息量变化，实现基于信息论的垃圾邮件过滤方法。

（3）文本情感分析中的文本分类方法

自然语言处理领域的情感分析任务，引起了研究者们的重视。情感分析任务的目标是对带有主观情感的文本进行类别分析，判定是正面情感还是负面情感。迁移学习方法能够有效地解决不同领域的情感分析，避免了新领域的情感分析语料的标注工作。基于特征表示的迁移学习方法-SCL[14]被用于解决文本情感分析任务，并取得良好的效果。SCL 通过寻找不同领域间的枢特征，建立领域间的相关性。

本文采用基于 EM 的文本分类迁移学习方法完成不同领域之间的文本分类任务。首先采用朴素贝叶斯分类方法对源领域标注语料进行学习，获得分类模型，然后用 EM 算法将该初始模型适用到目标领域未标注语料中。

3.3.1 朴素贝叶斯分类方法

朴素贝叶斯（Na ïve Bayes）分类方法是一种被广泛应用的贝叶斯学习方法。贝叶斯学习方法以贝叶斯公式为学习基础，计算事件的假设概率。

贝叶斯公式如下所示：

$$P(A|B)=\frac{P(A,B)}{P(B)}=\frac{P(B|A)P(A)}{P(B)} \qquad (3\text{-}17)$$

其中，A 和 B 是随机事件，$P(\bullet)$ 表示随机事件的发生概率。事件 B 发生后事件 A 发生的后验概率 $P(A|B)$ 较难计算时，可以间接通过计算事件 A、B 共同发生的概率 $P(A,B)$ 与事件 B 的先验概率得到。贝叶斯公式提供了通过事件的先验概率计算后验概率的数学方法。

朴素贝叶斯分类方法基于贝叶斯公式计算文档属于类别的后验概率，将具有最大后验概率值的类别定义为该文档的所属类别。朴素贝叶斯分类的处理对象为采用向量空间模型表示的文本数据，设训练文本集合为 $D=\{d_1,d_2,\cdots,d_m\}$，m 为训练文本集合中文本数量。给定类别集合为 $CSET=\{c_1,c_2,\cdots,c_k\}$，$k$ 为类别集合中类别个数。文档 d 的类别标记为 c 的后验概率 $P(c|d)$ 由公式（3-18）计算。

$$P(c\mid d)=\frac{P(d\mid c)P(c)}{P(d)} \tag{3-18}$$

该式中，$P(d)$ 指文档的先验概率，对于类别集合 CSET 中所有类别 c，$P(d)$ 值不变，故可将该概率值省略。$P(c)$ 是类别 c 的先验概率，计算方法如公式（3-19）所示。$P(d\mid c)$ 是类别 c 中包含文档 d 的概率，计算方法如公式（3-20）所示。

$$P(c)=\frac{num(c)}{\sum_{c'\in CSET} num(c')} \tag{3-19}$$

$$P(d\mid c)=P(t_1,t_2,\cdots,t_{|d|}\mid c)=\prod_{i=1}^{|d|}P(t_i\mid c) \tag{3-20}$$

公式（3-19）中，$num(c)$ 表示训练文本集合中属于类别 c 的文本数量。公式（3-20）中，文档 d 表示为 $d=\{t_1,t_2,\cdots,t_{|d|}\}$。在此，我们做出假设：在给定所属类别 c 之后，文档 d 中的各个特征项之间相互独立，即贝叶斯假设。$P(t_i\mid c)$ 为特征词 t_i 出现在类别 c 的文本中的概率，计算如公式（3-21）所示。

$$P(t_i\mid c)=\frac{appnum(c,t_i)+1}{appnum(c)+WordNUM} \tag{3-21}$$

其中，$appnum(c)$ 为所属类别为 c 的所有文本中特征词位置数，$appnum(c,t_i)$ 为 $appnum(c)$ 中特征词 t_i 的出现次数，$WordNUM$ 为训练文本集合中的特征总数。

综上，文档 d 的所属类别为：

$$\begin{aligned} c_{NB}&=\arg\max_c P(c\mid d)=\arg\max_c P(d\mid c)P(c)\\ &=\arg\max_c P(c)\prod_{i=1}^{|d|}P(t_i\mid c)\end{aligned} \tag{3-22}$$

下图描述了朴素贝叶斯分类方法的具体实现。

输入：类别集合 $CSET=\{c_1,c_2,\cdots,c_k\}$，训练文本集合 $D=\{(d_1,c_1^{'}),(d_2,c_2^{'}),\cdots,(d_m,c_m^{'})\},c_i^{'}\in CSET$，测试文本 *doc*。

输出：测试文本集合文本的所属类别

步骤：

1、训练阶段：

1）统计训练文本集合中的出现的所有特征词，形成特征词集合 *TERM*

2）对于类别集合中的每个类别 *c*

a） 计算 $P(c)$

b） 对于所属类别为 *c* 的所有训练文本整合到同一个文件中，统计该文件中所有特征词出现位置。计算 *TERM* 中所有特征词在此文件中出现的次数，并计算特征词在该类别中的特征词权重 $P(t_i|c)$。

2、测试阶段：

统计出现在测试文本 *doc* 及 *TERM* 中的所有特征词，通过公式（3-22），获得测试文本所属类别 c_{NB}。

图 3- 5 朴素贝叶斯分类方法

3.3.2 基于 EM 的文本分类迁移方法

基于 EM 的文本分类迁移方法是基于 EM 的直推式迁移学习方法在文本分类问题上的具体实现算法。基于 EM 的直推式迁移学习方法是利用 EM 算法计算得到隐含信息，即未标注数据集合 D_u 中的标注信息。在文本分类中，隐含信息为未标注文本集合中文档的所属类别信息。基于 EM 的直推式迁移学习方法将寻求最优后验假设 h_{MAP} 的过程转化为最大化对数似然值 $l(h|D_l,D_u)$。

$$l(h|D_l,D_u)=\log P_{\Omega_u}(h)+\sum_{d\in D_l}\log P_{\Omega_u}(d|h)+\sum_{d\in D_u}\log P_{\Omega_u}(d|h) \quad (3\text{-}23)$$

针对文本分类问题，EM 算法通过 E-M 迭代寻求 $l(h|D_l,D_u)$ 的局部最大值，即 $P(c|d)$ 的最大似然估计值。以下所有概率均是服从数据分布 Ω_u。

（1）Expect 步：通过当前的概率值 $P(c)$ 和 $P(t|c)$ 求解 $P(c|d)$，

$$P(c|d)\propto P(c)\prod_{t\in d}P(t|c) \quad (3\text{-}24)$$

（2）Maximum 步：根据 $P(c|d)$ 更新概率值 $P(c)$ 和 $P(t|c)$。

$$P(c)=\sum_{i\in\{l,u\}}P(D_i)P(c|D_i) \quad (3\text{-}25)$$

$$P(t|c)=\sum_{i\in\{l,u\}}P(D_i)P(c|D_i)P(t|c,D_i) \quad (3\text{-}26)$$

其中，$P(D_i)$ 是 D_l 和 D_u 的权衡参数，由于 D_u 直接从数据分布 Ω_u 中提取出，

故 $P(D_u) > P(D_l)$。$P(c|D_i)$ 计算公式如下：

$$P(c|D_i) = \sum_{d \in D_i} P(c|d)P(d|D_i) \quad (3\text{-}27)$$

$P(t|c,D_i)$ 计算如下所示：

$$P(t|c,D_i) = \frac{1+n(t,c,D_i)}{|TERM|+n(c,D_i)} \quad (3\text{-}28)$$

其中，$n(c,D_i)$ 是数据集合 D_i 中类别 c 的出现概率，$n(t,c,D_i)$ 是数据集合 D_i 中特征词 t 在类别 c 中的出现概率，分别计算如下：

$$n(c,D_i) = \sum_{d \in D_i} |d| P(c|d) \quad (3\text{-}29)$$

$$n(t,c,D_i) = \sum_{d \in D_i} |d| P(t|d)P(c|d) \quad (3\text{-}30)$$

基于 EM 的文本分类迁移方法，简记为 EMTC，算法流程图如下：

输入：类别集合 $CSET = \{c_1, c_2, \cdots, c_k\}$，训练文本集合 $D_l = \{(d_1, c_1'), (d_2, c_2'), \cdots, (d_m, c_m')\}, c_i' \in CSET$，与 D 不同分布但相关的未标注测试文本集合 $D_u = \{d_1, d_2, \cdots, d_m\}$，最大迭代次数 *MAXI*。

输出：测试文本集合文本的所属类别

步骤：

1、初始化参数阶段：

采用朴素贝叶斯分类器，对训练文本集合文本得到测试文本集合所属数据分布 Ω_u 下的初始参数值，$P(t|c)$，*P(c)*，*P(d)*。

2、EM 迭代阶段：

使用当前参数值逐步迭代，迭代次数小于 *MAXI*：

Expect：对于每个类别以及 D_u 中的文本，根据当前概率值 *P(c)*和 $P(t|c)$ 使用公式（3-24）计算 $P(c|d)$；

Maximum：对于类别集合中的类别 *c*，根据公式（3-25）和 $P(c|d)$ 计算 *P(c)*，对于每个特征 *t*，由公式（3-26）及 $P(c|d)$ 计算 $P(t|c)$。

图 3- 6 基于 EM 的文本分类迁移方法（EMTC）算法流程图

3.4 实验与结果分析

3.4.1 实验语料与评价方法

由于迁移学习的特殊研究内容，所用的实验语料具有一定的特性：相关但分布不同。完全不相关的语料没有共同知识，无法进行迁移学习。目前多数分类语料均是基于传统机器学习方法的，并不满足该条件，需要为迁移学习建立

语料。一般从层次分类语料中构建迁移学习语料。设层次语料的 A、B 两个大类下分别有两个子类，记为 A_1、A_2，B_1、B_2。子类之间具有一定的相关性，但却不完全相同。因此，可将 A_1、B_1 类别下的文本作为训练集合，将 A_2，B_2 类别下的文本作为测试集合。构建示例图如下所示。

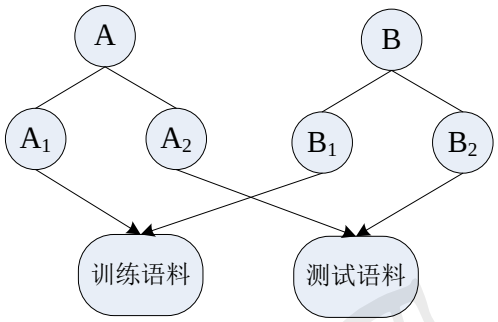

图 3- 7 迁移学习语料构建过程

文本分类任务中，国内外学者常用的迁移学习文本语料多为英文语料，包括 20Newsgroup[4]、SRAA[5]和 Reuters-21578[6]。这三个语料已被研究者们处理成为适合迁移学习研究的形式。然而，基于中文文本语料的迁移学习却非常稀少。文献[47]中使用中文文本分类语料谭松波语料作为实验数据集。

本文从谭松波层次分类语料[48]中建设出实验语料 TLCorpus1。原始谭松波层次分类语料共包括文本 14150 篇，分为 12 个一级分类，60 个二级分类，原始语料分类结构统计见表 3-1。本文将取其中五类一级分类及下面部分子类构建迁移学习文本分类语料 TLCorpus1，构建方法如上所述，TLCorpus1 语料结构统计如表 3-2 所示。

作为本文的迁移学习语料，训练语料是有标注的语料，测试语料是无类别标注语料。另外，本文从复旦语料中抽取出部分类别文档作为迁移学习文本分类语料 TLCorpus2。复旦语料中与 TLCorpus1 类别结构相同的部分语料结构统计如表 3-3 所示。

[4] 20Newsgroup, http://people.csail.mit.edu/jrennie/20Newsgroups
[5] SRAA, http://www.cs.umass.edu/~mccallum/data/sraa.tar.gz
[6] Reuters-21578, http://www.daviddlewis.com/resources/testcollections/

表 3- 1 谭松波语料分类结构统计表

一级分类	二级分类	文本数	一级分类	二级分类	文本数
体育（2806）	足球	1317	电脑（2943）	电子商务	693
	篮球	962		电脑病毒	631
	网球	131		电脑科技	574
	乒乓球	112		电脑网络	517
	水上	94		电脑软件	426
	田径	84		电脑游戏	102
	羽毛球	55			
	棋牌	50			
财经（819）	金融	267	教育（808）	校园	226
	证券	214		考试	173
	企业	164		就业	146
	消费	91		招生	127
	人物	64		留学	67
	财富	19		出版	48
				培训	21
人才（608）	人才管理	412	艺术（546）	舞台艺术	185
	人才薪金	40		文学艺术	153
	人才创业	39		美学艺术	84
	人才猎取	39		音乐艺术	73
	人才履历	39		古董艺术	51
	人才应试	39			
汽车（590）	汽车快讯	258	科技（1040）	生命科学	459
	汽车行驶	176		自然科学	229
	汽车百科	118		考古科学	183
	汽车政策	38		天文科学	169
卫生（1406）	保健	625	房产（935）	私宅	433
	医药	383		组屋	254
	两性	335		装修	172
	心理	63		城建	76
地域（150）	地域城市	71	娱乐（1500）	综艺娱乐	500
	地域风俗	47		音乐娱乐	500
	地域美食	32		电影娱乐	500

表 3- 2 TLCorpus1 语料结构统计表

训练语料			测试语料		
一级分类	二级分类	文本数	一级分类	二级分类	文本数
电脑	电脑科技	500	电脑	电脑病毒	600
	电脑网络	500		电脑软件	400
体育	篮球	900	体育	网球	100
	乒乓球	100		足球	900
科技	自然科学	220	科技	生命科学	300
	考古科学	180		天文科学	100
房产	私宅	350	房产	组屋	250
	城建	50		装修	150
娱乐	电影娱乐	500	娱乐	音乐娱乐	500

表 3- 3 与 TLCorpus2 类别结构相同的复旦语料结构统计表

类别映射		文档数	类别映射		文档数
TLCorpus1	TLCorpus2		TLCorpus1	TLCorpus2	
电脑	Computer	1000/1358	体育	Sports	1000/1254
财经	Economy	819/1601	卫生	Medical	1406/53
教育	Education	808/61	艺术	Art Literature	546/776
汽车	Transport	590/59			

由于复旦语料中部分类别规模较小，我们只从电脑、体育类别中选择 1000 篇文档作为 TLCorpus2。表 3-3 中的第一行即为 TLCorpus2 的类别结构。

文本分类的评价指标主要有准确率、召回率、F_1 值、宏观平均值、微观平均值。对于类别集合中的每个类别 *i*，分类器对语料集合的分类结果分为如下四种情况，如表 3-4 所示。

表 3- 4 分类器的分类结果

	属于 *i* 类的文档数	不属于 *i* 类的文档数
分类器分到 *i* 类的文档数	*NUM_YY*	*NUM_YN*
分类器未分到 *i* 类的文档数	*NUM_NY*	*NUM_NN*

由上表的结果，可以直观表示各评价指标的定义如下：

（1）准确率、召回率、F_1 值，是衡量单个类别的分类效果的指标。

1) 准确率(Precision)，分类器划分到该类别的文档中属于该类别的文档所占的比例，定义为：

$$P_i = \frac{NUM\_YY}{(NUM\_YY + NUM\_YN)} \tag{3-31}$$

2) 召回率(Recall)，属于该类别的文档中被分类器划分到该类中的文档所占的比例，定义为：

$$R_i = \frac{NUM\_YY}{(NUM\_YY + NUM\_NY)} \tag{3-32}$$

3) F_1 值，由准确率和召回率计算的综合评测指标，定义为：

$$F_{1i} = \frac{2 \cdot P_i \cdot R_i}{(P_i + R_i)} \tag{3-33}$$

（2）宏观平均值、微观平均值，是衡量全局分类效果的指标。

1) 宏平均，首先计算类别结构中所有类别的准确率、召回率及 F_1 值，通过下列公式计算宏平均准确率、宏平均召回率、宏平均 F_1 值：

$$Macro\_P = \frac{\sum_{i=1}^{|CSET|} P_i}{|CSET|} \tag{3-34}$$

$$Macro\_R = \frac{\sum_{i=1}^{|CSET|} R_i}{|CSET|} \tag{3-35}$$

$$Macro\_F_1 = \frac{\sum_{i=1}^{|CSET|} F_{1i}}{|CSET|} \tag{3-36}$$

其中，*CSET* 为类别集合中的类别数量。

2) 微平均，先计算整个文档集的分类结果，再计算整体的准确率、召回率、F_1 值，公式如下所示：

$$Micro\_P = \frac{\sum_{i=1}^{|CSET|} NUM\_YY}{\sum_{i=1}^{|CSET|} NUM\_YY + \sum_{i=1}^{|CSET|} NUM\_YN} \tag{3-37}$$

$$Micro\_R = \frac{\sum_{i=1}^{|CSET|} NUM\_YY}{\sum_{i=1}^{|CSET|} NUM\_YY + \sum_{i=1}^{|CSET|} NUM\_NY} \tag{3-38}$$

$$Micro\_F_1 = \frac{2 \cdot Micro\_P \cdot Micro\_R}{Micro\_P + Micro\_P} \quad (3\text{-}39)$$

本文以宏平均准确率、宏平均召回率、宏平均 F1 值、微平均准确率、微平均召回率、微平均 F1 值作为实验指标衡量实验效果。

3.4.2 实验设置

本文以朴素贝叶斯（NB）分类方法，支持向量机（SVM）分类方法作为基准实验方法测试基于 EM 的文本分类迁移学习方法（EMTC）的实验效果。具体实验设置详见下表 3-5。

表 3- 5 文本分类迁移学习方法实验设置

实验属性			实验方法		
二元分类	实验组 1	Comp vs Sports	NB1	SVM1	EMTC1
	实验组 2	电脑 vs 体育	NB2	SVM2	EMTC2
	实验组 3	电脑 vs 科技	NB3	SVM3	EMTC3
	实验组 4	科技 vs 房产	NB4	SVM4	EMTC4
多元分类（五类）	实验组 5	电脑、体育、娱乐、科技、房产	NB5	SVM5	EMTC5

我们从迁移学习分类语料 TLCorpus1 中抽取出不同类别，进行实验组 2、3、4、5，每组实验均采用 NB、SVM 和 EMTC 方法进行实验。实验组 1 采用 TLCorpus1 中的电脑和体育类别文本进行训练，在 TLCorpus2 上的 Computer 和 Sports 类别文本集合中进行 NB、SVM 和 EMTC 方法的测试。实验方法采用进行卡方统计特征选择方法，阈值设为 0.5。实验组 1、2、3、4 为二元分类实验，实验组 5 为多元分类，具体为五类文本分类。二元分类实验组使用不同类别的语料进行实验，以评估实验语料不同时，具体表现为不同类别分布间差异对实验方法效果的影响。

3.4.3 实验结果与分析

文本分类迁移学习方法的实验结果如表 3-6 所示：

表 3- 6 文本分类迁移学习方法实验结果

		Macro_P	Macro_R	Macro_F_1	Micro_P	Micro_R	Micro_F_1
实验组 1	NB1	77.47%	60.5%	53.28%	60.5%	60.5%	60.5%
	SVM1	76.92%	76.92%	76.92%	59.65%	59.65%	59.65%
	EMTC1	84.14%	83.95%	83.93%	83.95%	83.95%	83.95%
实验组 2	NB2	89.4%	89.4%	89.39%	89.4%	89.4%	89.4%
	SVM2	81.39%	81.39%	81.39%	79.6%	79.6%	79.6%
	EMTC2	91.53%	91.45%	91.45%	91.45%	91.45%	91.45%
实验组 3	NB3	86.14%	86.75%	86.44%	87.07%	87.07%	87.07%
	SVM3	76.18%	76.18%	76.18%	62.71%	62.71%	62.71%
	EMTC3	89.18%	90.38%	89.75%	91.5%	91.5%	91.5%
实验组 4	NB4	84.19%	83.5%	83.47%	83.5%	83.5%	83.5%
	SVM4	65.56%	65.56%	65.56%	42.18%	42.18%	42.18%
	EMTC4	87.31%	86.87%	86.84%	86.88%	86.88%	86.88%
实验组 5	NB5	80.67%	77.48%	78.38%	80.78%	80.78%	80.78%
	SVM5	51.29%	51.29%	51.29%	40.12%	40.12%	40.12%
	EMTC5	90.65%	60.08%	71.57%	90.17%	61.45%	73.09%

对表 3-6 中的结果进行分析，总结得到以下结论：

（1）实验组 2、3、4、5 中 NB、SVM 及 EMTC 方法的评价指标值均是依次递减的，说明机器学习方法的性能与实验语料的规模的相关性。训练语料规模越大，越有可能学习到更好的模型，获得好的分类效果。由此也证明了缺少大量训练语料时，文本分类性能会大大降低。

（2）实验组 2、3、4、5 中 NB 与 SVM 分类方法的性能相比说明，在迁移学习问题中 NB 效果优于 SVM。宏平均指标大表示各类别分类结果差距小，微平均指标值大说明分类器的总体性能较好。各组实验中 NB 分类方法微平均指标值略高于宏平均指标值，说明 NB 方法的分类性能稳定，且对训练语料中各类别文本数量比例不敏感。同时说明 SVM 不稳定，容易出现某一类文本分类结果全部正确，另外一个类别分类结果大量错误的情况。

（3）实验组 2、3、4、5 中 EMTC 的评价指标值均大于 NB 与 SVM 方法的评价指标值，且评价指标差值随 NB 评价指标值的降低而增大。这说明在 NB 方法不适用的迁移学习方法中 EMTC 方法能够起到更明显的效果。由此证明基于 EM 的直推式迁移学习方法的有效性。

（4）和实验组 2 结果相比，实验组 1 中 NB1、SVM1、EMTC1 的评价指标值略低，原因是该实验组中所用的 TLCorpus1 与 TLCorpus2 虽然类别结构相

同，但是特征集合与分布均不同，训练语料与测试语料的较大差异性导致 NB1 与 SVM1 的效果较差。但是 EMTC1 方法能够将也得到 84%的准确率，说明本文的基于 EM 的文本分类迁移方法效果显著。

3.5 本章小结

本章中，首先介绍了传统文本分类方法的概念及流程，包括预处理、文本表示、特征选择及文本分类方法。然后阐述了基于 EM 的直推式迁移学习方法在文本分类中的应用。最后，讲述了实验语料的构建方法及文本分类的评价方法。以朴素贝叶斯分类方法及支持向量机分类方法为 Baseline，设置了 5 组实验对文本分类迁移学习方法进行效果验证。实验结果证明，本文的方法对异构空间下的文本集合的分类准确率等指标均有促进作用。

第 4 章 迁移学习方法在术语抽取中的应用

4.1 引言

作为承载领域知识的最基本单元，术语对于术语库的编撰和管理以及众多工程应用具有重要意义。从文档中抽取出术语则是术语抽取任务(Terminology extraction)的目标。术语抽取技术包括两部分：根据单元度抽取出文档中的候选术语；通过术语度判定候选术语是否是术语。单元度（Unithood）标示了一个字串作为有意义短语出现的可能性，术语度（Termhood）标示了一个短语在所属文本中是术语的可能性。

术语的领域性使得当前效果较好的术语计算机自动抽取技术均不具有领域普适性，而是领域相关的。然而对于一个新的领域术语抽取任务，我们缺乏大量的术语抽取训练语料，而术语抽取语料标注工作所耗费的人力物力无可估量。因此进行领域间的术语抽取工作意义重大。本文所提出的术语抽取迁移方法采用基于机器学习的算法实现计算机领域术语知识完成对环境领域术语的抽取。

4.2 术语分析

4.2.1 术语定义

“术语”至今还没有一个普遍认可的定义。部分研究者们提出的定义列举如下：

（1）术语是通过语音或文字来表达或限定专业概念的约定性符号。这一解释由我国较早从事术语研究的冯志伟在《现代术语学引论》中提出的。

（2）术语是各门学科中的专门用语，用来标记生产技术、科学艺术、社会生活等各专门领域中的事物、现象、特性、关系和过程。这是《中国大百科全书》中对术语的描述。

综上，本文将术语定义为：表示某一特定领域的特定概念的特定符号。

4.2.2 术语类别体系

根据不同的角度，可以将术语划分到不同的类别体系。

（1）根据术语的不同构成方式，可将术语划分为单词型术语和词组型术语。

定义 4.1：单词型术语，由一个单词组成，又称为“简单术语”。对于英文来说就是一个独立单词（Word），例如“Memory”，中文含义为“内存”。对于

中文来说，一个单词是一个及一个以上数量的字组成的字序列。例如“词频”包含两个字。计算机领域中用“词频”表示一个词语在一篇文本中出现的次数。

定义 4.2：词组型术语，由两个及以上数量的单词组成的固定搭配，又称为“复杂术语”。组合成词组型术语的单词可能是普通单词，也可能是单词型术语。例如词组型术语“机器翻译”由普通单词“机器”和普通单词“翻译”组合而成；术语“神经网络”由普通单词“神经”和单词型术语“网络”组成。

（2）从术语的专业性来看，把术语分为两类。

定义 4.3：纯术语，指某个领域特有的专业术语，是专业性较强的术语。纯术语使用范围较小，仅供领域专业人士研究。例如计算机领域的“支持向量机”、“IBM 模型”等。

定义 4.4：一般术语，指各领域中的基本术语，专业性弱于纯术语。例如化学领域术语“还原剂”，生物专业术语“神经元”。一般术语使用范围广，能够被其他领域引用。

除上述分类标准之外，根据所属领域还可将术语划分为计算机领域术语、物理领域术语、化学领域术语、人文领域术语等。

4.2.3 术语语言特性分析

作为领域信息的载体，术语具有其特殊的结构和一定的语言特性，能够辅助识别术语。本文将对各领域术语的通用语言特性进行简单分析。

（1）术语长度 中文术语长度约为 2~8 字，单字术语和 8 字以上术语所占比例极小，可忽略之。

（2）词性特点 术语多为名词性词语或名词性短语。

（3）边界信息 部分术语在文本中经常作为定义出现，具有明显的前后界标志。例如“定义为”，标点符号冒号等。

（4）组合模式 词组型术语由两个及以上词语构成，组合模式具有一些普遍规律，如“名词+名词”、“形容词+名词”等。

4.3 传统术语抽取方法

传统术语抽取方法包括候选术语抽取和术语抽取两部分内容。其中候选术语抽取工作同样适用于本文进行的术语抽取任务的前期工作。

4.3.1 候选术语抽取

候选术语抽取的任务是从原始语料中抽取出候选术语，包括去噪、分词、重新切分三个步骤。

（1）去噪

本文所选术语抽取语料大部分均是来自于期刊中发表的论文。文本中包含图表等中文术语抽取的噪声信息。在进行候选术语抽取之前，我们将这些信息进行删除，只保留原始中文句子。

（2）分词

中文术语抽取的任务是从中文文本中抽取出其中所包含的术语。文本中的句子是一个包含标点符号的连续字序列。在英文句子中，字和字之间由空格分隔开来，每个字能够表征一定的含义。和英文不同的是，中文字与字之间不包括空格，且中文的一个字无法表达具体的含义，词语才能作为承载语义信息的单元。中文分词就是把中文句子由连续的字序列变成由空格分隔开来的词序列的过程。本文使用中科院分词器对初始文本进行分词。需要额外提出的是，考虑到系统的时间，我们首先将语料中若干文件整合到一个大的语料文件中，然后进行分词。该语料文件中的一行代表原来的一个文本。

（3）重新切分

经验统计，中文领域术语的长度约为 2~8 字。普通分词器无法将领域术语全面准确的切分出来，我们有必要对切分后的词进行重新组合，形成 2~8 字长度的候选术语集合。这个过程也就是根据一定的切分标志，对词串进行重新切分。切分标志可选定为对中文术语抽取无意义的标点符号、中西文数字以及助词、介词等虚词。将文本中的切分标志替换成候选术语分隔符以得到更具有代表意义的候选术语。

表 4- 1 重新切分结果样例

	候选术语部分列表
重切分	智能 Agent ###协同 工作 模型 ###计算机 支持 ###协同 工作 ###CSCW ###灵活 ###开放 ###扩充 ###模型 结构 ###本文 ###分布式 人工 智能 研究 ###智能 Agent ###系统 基本 单元 ###智能 Agent ###协同 工作 模型 ###具体 实现 ###关键 计算机 支持 ###协同 工作 ###智能 Agent ###分布式 人工 智能 ###计算机 支持 ###协同 工作 ###CSCW ###Computer Supported Collaborative Work

4.3.2 术语抽取算法

经过上述预处理后得到的候选术语集合，术语抽取算法将根据候选术语的术语度从中抽取出领域术语。

术语抽取技术可被分为两类：基于统计的方法和基于规则的方法。基于规则的方法根据术语的语法特性，构建若干术语抽取规则，从候选术语集合中抽

取出领域术语。术语抽取规则大多针对特定领域由人工总结得到，得到的术语形式较为规范，但却需要大量人工参与，且抽取召回率指标随着规则的变化而不同，移植性也较差。

基于统计的方法采用一定的统计量衡量候选术语的术语度。常用的统计量有频率、互信息、假设检验和 C-value[22]等。互信息统计量通过计算候选术语作为一个整体在语料库中出现概率，用来衡量字串内部结合紧密度。C-value 是通过边界自由度，即一个字串的边界上出现多种字符串的可能性大小衡量字串的独立性。

（1）频率

频率表示词语出现在语料集合中的次数。术语的频率往往很高，但频率最高的不是术语，而是助词“的”等虚词。频率常作为辅助统计量进行术语抽取。

（2）互信息

基于互信息的术语抽取方法，从字串的角度出发，计算短语的互信息。记候选术语为字串 $\mathrm{s}=s_1,\cdots,s_n$，n 为字串 s 的长度。特殊地，术语抽取任务中要求 $n\geq 2$。字串 s 的两个子串记为 $a=s_1,\cdots,s_{n-1}$ 和 $\mathrm{b}=s_2,\cdots,s_n$。$prob(s)$ 表示 s 中子串的共现概率。字串 s 的互信息 $MI(s)$ 计算如下：

$$MI(s)=\log_2\left(\frac{prob(s)}{prob(a)prob(b)}\right) \qquad (4\text{-}1)$$

其中，$prob(s)=\dfrac{frequency(s)}{Num}$，$frequency(s)$ 表示字串 s 的共现频率，Num 表示语料库中候选术语的数量。根据极大似然估计，在大规模语料库中，$prob(s)$ 可近似为 $frequency(s)$。

字串 s 的互信息值较大时，子串 a 和 b 的共现概率与 s 的共现概率差值较小，进一步说明字串 s 内部结合紧密。互信息值较小说明字串 s 内部结合的不够紧密。对于语料库中提取的所有候选短语计算互信息后，根据互信息阈值选择互信息值较大的候选短语作为候选术语。

（3）C-value/NC-value

C-value/NC-value 不仅包括统计量 C-value，还融入了语言学信息 NC-value[22]。C-value 用来衡量候选术语的术语度，由下列统计值组合运算得出：

1）候选术语的出现频率 frequency。一般来说，高频短语的术语度高于低频短语，故术语度正比于频率 frequency。

2）父候选术语频率 frequencyOfFatherString。父候选术语是指其中一个子串是当前候选术语的候选术语。父候选术语比当前候选术语更有可能是一个具有完整意义的术语。研究学者认为术语度与父候选术语频率

frequencyOfFatherString 成反比关系。

3）父候选术语的数量。候选短语的父候选术语越多，该候选术语可结合性更强，更容易和其他词结合形成新的候选术语，术语度越小。

4）候选术语的长度（术语包含的字数）。众所周知，在语料中，长短语比短短语更难获得高频率，故在相同频率下，长候选术语的术语度更高。

则候选术语 c 的 C-value 值计算公式如下：

$$C\_value(c)=\begin{cases}\log_2|c|*freq(c) & \text{c无父候选短语}\\ \log_2|c|*(freq(c)-\dfrac{1}{num(FS_C)}\sum_{s\in FS_C}freq(s)) & \text{c有父候选术语}\end{cases} \quad (4\text{-}2)$$

其中，$freq(\bullet)$ 指语料库中候选术语的出现频率，FS_C 是候选术语 c 的父候选术语集合，$num(FS_C)$ 是父候选术语集合大小。

术语的边界信息被作为语言学信息融入 C-value 统计量形成 NC-value。术语的边界信息即为术语前后出现的边界标志词，如“称为”等。从语料中抽取出边界词表时，通过边界权重阈值进行选择，计算如下：

$$boundaryWeight(word)=\frac{num(T_{Word})}{num\_\text{Term}} \quad (4\text{-}3)$$

其中，$word$ 为要计算边界权重阈值的边界候选词，T_{Word} 为边界候选词 $word$ 前后出现的术语集合，$num(T_{Word})$ 为 T_{Word} 中术语个数，num_Term 为语料库中所有术语个数。

候选术语 c 的 NC-value 值计算如下：

$$NC\_value(c)=\alpha*C\_value(c)+(1-\alpha)*\sum_{w\in B_C}freq(w)*boundaryWeight(w) \quad (4\text{-}4)$$

α 为加权系数，此处选为 0.8，B_C 为候选术语 c 前后出现的边界词集合。

4.4 术语抽取中的迁移学习策略

研究者们对于不同领域的术语抽取技术进行了初步探索。本文拟采用基于机器学习的术语抽取方法，即基于分类的方法完成不同领域中的术语抽取任务。而后，将迁移学习的思想融入基于分类的术语抽取方法中，实现跨领域的术语抽取。下面进行详细介绍。

4.4.1 基于朴素贝叶斯分类的术语抽取方法

本文采取分类的思路进行文档中的术语抽取：文档中的短语要么是术语，要么不是术语，正确地将文档中所有短语划分到术语或非术语类别，就完成了文档术语抽取。本文将采用朴素贝叶斯分类器对文档中的候选术语进行类别分

析，识别出文档中术语。本文采用的基于朴素贝叶斯分类的术语抽取流程为：

（1） 候选术语抽取，具体方法详见本文的 4.3.1 节。

（2） 基于朴素贝叶斯分类的术语抽取，采用朴素贝叶斯分类器对候选术语进行二元分类，获得各候选术语属于术语类别的概率，按概率值大小排序。

（3） 后处理过程，对上述列表进行后处理，获取 Top N（N>=5）个术语作为文档术语。

本文中将使用候选术语的两种统计特征训练分类器，分别是 TF-IDF 值和首次出现位置，这两种统计值均能在一定程度上表征候选术语的重要性。

（1）TF-IDF：候选术语的 TF-IDF 值探索的是候选术语在某篇文档中的出现频率在整体语料中的出现频率的比重。候选术语 P 的 TF 值表示在文档中出现的频率，在一定意义上表征其重要性。一般说来，文档中高频出现的候选术语比低频候选术语更加重要。候选术语 P 的 IDF 值是文档频率 DF 的倒数。文档频率 DF 值表达的是候选术语在语料中的分布信息，即语料中包含该短语的文档数。由于助动词等在某个文档中出现的频率可能高于某些有意义的候选术语。故将候选术语的分布信息 DF 值考虑在内，能够制衡助动词等无意义词的 TF 值。故而，TF-IDF 能够更充分的表征一个短语的意义。候选术语的 TF-IDF 值计算方法见本文的 3.2.2 节。

（2）首次出现位置 Pos：首次出现位置指文档中候选术语 P 第一次出现之前的单词个数。这种位置特征统计值表现了术语的独特性质。该值最后得到 0 到 1 之间的值。

值得说明的是，上述两种特征值均是连续值，为了使其适用于分类，我们对特征值进行了离散化，采用离散值 $v \in \{1,2,3,4,5\}$ 表示特征值。

对于已经抽取出的候选术语列表中的每个术语，首先要计算该候选术语 t 属于术语类别 yes 的概率。根据贝叶斯法则可转化为先验概率，具体计算如公式（4-5）所示。

$$\begin{aligned} P(yes \mid t) &= \frac{P(yes)P(t \mid yes)}{P(t)} \propto P(yes)P(t \mid yes) \\ &\propto \frac{NUM\_Y}{NUM\_Y + NUM\_N} \cdot P_{tf-idf}(v \mid yes) \cdot P_{pos}(p \mid yes) \end{aligned} \tag{4-5}$$

其中，NUM_Y 是训练语料中的人工标注的术语个数，NUM_N 是训练语料中非术语个数。$P_{tf-idf}(v \mid yes)$ 是候选术语 t 的 TF-IDF 离散值概率，$P_{pos}(p \mid yes)$ 是 pos 的离散值概率，计算如下：

$$P_{tf-idf}(v \mid yes) = \frac{num_{tf-idf}(v, yes) + 1}{num_{tf-idf}(yes) + Y}, v \in \{1,2,3,4,5\} \tag{4-6}$$

$$P_{pos}(p \mid yes) = \frac{num_{pos}(p, yes) + 1}{num_{pos}(yes) + NUM\_Y}, p \in \{1,2,3,4,5\} \quad (4\text{-}7)$$

$num_{tf-idf}(v, yes)$为文本集合中所有 TF-IDF 值离散值为 v 的术语出现次数之和，$num_{tf-idf}(yes)$为文本集合中所有术语出现次数之和。

与公式（4-5）类似，可计算候选术语属于非术语类别 *no* 的概率为

$$P(no \mid t) = \frac{NUM\_Y}{NUM\_Y + NUM\_N} \cdot P_{tf-idf}(v \mid no) \cdot P_{pos}(p \mid no) \quad (4\text{-}8)$$

则候选术语 *t* 是术语的概率计算如下：

$$P(t) = \frac{P(yes \mid t)}{P(yes \mid t) + P(no \mid t)} \quad (4\text{-}9)$$

根据此概率值将所有候选短语进行排序，得到候选术语的有序列表。

后处理过程包括：1）将具有相同概率值的候选术语按照原始 TF-IDF 排序。2）若列表中的候选术语具有更高概率值的父候选术语，将该候选术语删除。经过后处理后的候选术语列表，取 Top N 为文档术语结果。

基于朴素贝叶斯分类的术语抽取方法描述见图 4-1。

输入：术语集合$T = \{t_1, t_2, \cdots, t_k\}$，训练文本集合$D = \{d_1, d_2, \cdots, d_m\}$，测试文本 *doc*。

输出：测试文本所包含的术语。

步骤：

1、训练阶段：

1）统计训练文本集合中的出现的所有候选术语，将所有候选术语进行是否是术语进行类别标注，形成候选术语集合$T = \{(t_1, c_1), (t_2, c_2), \cdots, (t_n, c_n)\}, c_i \in \{yes, no\}$，计算术语数量，记为 Y，非术语数量记为 N。

2）对于训练文本集合中的每个文本 *d*，

计算文本 *d* 中包含的候选术语的 TF-IDF 值及 Pos，同时统计 TF-IDF 和 Pos 特征的最大、最小值。

3）将所有候选术语 TF-IDF 及 Pos 值离散到区间[1-5]中的离散值上。

4）对于候选术语集合中的每个类别标注为 *yes* 及 *no* 的候选术语 *t*

利用公式（4-6）和（4-7）计算$P_{tf-idf}(v \mid yes)$、$P_{pos}(p \mid yes)$、$P_{tf-idf}(v \mid no)$、$P_{pos}(p \mid no)$。

2、测试阶段：

统计测试文本 *doc* 中的所有候选术语，通过公式（4-9），计算每个候选术语的术语概率，形成有序候选术语列表，取 Top N 作为 *doc* 中出现的术语。

图 4- 1 基于朴素贝叶斯分类的术语抽取

4.4.2 基于 EM 的术语抽取迁移方法

基于 EM 的术语抽取迁移方法是基于 EM 的直推式迁移学习方法在术语抽取问题上的具体实现。文本分类问题中的隐含信息为未标注文本集合中文档的所属类别信息，而在术语抽取中，隐含信息为未标注术语的文本集合中候选术语是否属于术语的信息。本方法采用 EM 算法得到后验假设对数似然值 $l(h|D_l,D_u)$ 的最大值。EM 算法过程为：

（1）Expect 步：由当前的概率值 $P(yes)$、$P(t|yes)$、$P(no)$ 和 $P(t|no)$ 根据公式（4-5）、（4-8）、（4-9）求解 $P(yes|t)$、$P(no|t)$、$P(t)$。

（2）Maximum 步：根据 $P(yes|t)$、$P(no|t)$、$P(t)$ 更新概率值 $P(yes)$、$P(t|yes)$、$P(yes)$ 和 $P(t|yes)$。

$$P(yes)=\sum_{i\in\{l,u\}}P(D_i)P(yes|D_i) \tag{4-10}$$

$$P(t|yes)=\sum_{i\in\{l,u\}}P(D_i)P(yes|D_i)P(t|yes,D_i) \tag{4-11}$$

其中，$P(D_i)$ 是 D_l 和 D_u 的权衡参数，且 $P(D_u)>P(D_l)$。$P(yes|D_i)$ 计算公式如下：

$$P(yes|D_i)=\sum_{d\in D_i}\sum_{t\in d}P(yes|t)P(t|d)P(d|D_i) \tag{4-12}$$

$P(t|yes,D_i)$ 计算如下所示：

$$P(t|yes,D_i)=P_{tf-idf}(v|yes,D_i)\cdot P_{pos}(p|yes,D_i) \tag{4-13}$$

其中，$P_{tf-idf}(v|yes,D_i)$、$P_{pos}(p|yes,D_i)$ 分别计算如下：

$$P_{tf-idf}(v|yes,D_i)=\frac{num_{tf-idf}(v,yes,D_i)+1}{num_{tf-idf}(yes,D_i)+NUM\_Y} \tag{4-14}$$

$$P_{pos}(v|yes,D_i)=\frac{num_{pos}(p,yes,D_i)+1}{num_{pos}(yes,D_i)+NUM\_Y} \tag{4-15}$$

基于 EM 的术语抽取迁移方法，简记为 EMTE，算法流程图见图 4-2。

输入：训练文本中术语集合 $T=\{t_1,t_2,\cdots,t_k\}$，训练文本集合 $D=\{d_1,d_2,\cdots,d_m\}$，与 D 不同分布但相关的未标注术语测试文本集合 $D_u=\{d_1,d_2,\cdots,d_m\}$，最大迭代次数 *MAXI*。

输出：测试文本集合中文本的包含的术语。

步骤：

1、初始化参数阶段：

采用朴素贝叶斯分类器，对训练文本集合文本得到测试文本集合所属数据分布 Ω_u 下的初始参数值，$P(yes|t)$、$P(no|t)$、$P(t)$、$P(yes)$、$P(t|yes)$、$P(no)$ 和 $P(t|no)$。

2、EM 迭代阶段：

使用当前参数值逐步迭代，迭代次数小于 *MAXI*：

Expect：对于每个类别以及 D_u 中的文本，根据当前概率值 $P(yes)$、$P(t|yes)$、$P(no)$ 和 $P(t|no)$ 使用公式（4-5）、（4-8）、（4-9）计算 $P(yes|t)$、$P(no|t)$、$P(t)$。

Maximum：根据公式 4-10 计算 $P(yes)$、$P(no)$

对于 TF-IDF 及 Pos 离散后的每个值，计算 $P(t|yes)$、$P(t|no)$。

图 4- 2 基于 EM 的术语抽取迁移方法（EMTE）算法流程图

4.5 实验与结果分析

4.5.1 数据集与评价指标

1、数据集

本文从复旦语料中选择计算机和环境两个类别中的部分文章作为术语抽取的语料库。首先，从计算机和环境领域中过滤掉文档小于 6KB 的文本，小于 6KB 的文本一般内容为新闻，术语个数小于 3。其次，从经过初步筛选的 1102 篇计算机领域文本和 924 篇环境领域文本分别选出 150 篇作为实验语料。

2、评价指标

对于术语抽取的评价，一般采用准确率指标。准确率 Precision 是抽取出的术语中真实领域术语所占的比率。评价方法为：从已抽取出的术语列表中选取 Top N，判断哪些术语属于正例。继而计算准确率。

本文精确评价和覆盖评价的思想融入准确率指标中。精确评价指只有当术语和人工标注的术语完全匹配时才被算作是正确提取出的术语；覆盖评价是指当抽取出的术语和人工标注出的术语大致相同时即可被算作是正确提取出的术语。因此，我们引入精确准确率（Exact Precision，EP）和覆盖准确率（Cover

Precision，CP）及二者的平均值（Average Precision，AvgP）作为评价指标。各评价指标计算公式如下：

$$EP = \frac{|TermSet_S \cap TermSet_M|}{|TermSet_M|} \quad (4\text{-}16)$$

$$CP = \frac{|TermSet_S \tilde{\cap} TermSet_M|}{|TermSet_M|} \quad (4\text{-}16)$$

$$AvgP = \frac{EP+CP}{2} \quad (4\text{-}16)$$

上式中，$TermSet_S$ 表示由本文的术语抽取方法抽取出的术语集合，$TermSet_M$ 表示人工标注的术语集合。$\cap$ 符号表示集合交操作，即获取两集合中完全相同的共同元素，$\tilde{\cap}$ 符号表示模糊的集合交操作，即获取两集合中大致相同的共同元素。

4.5.2 实验设置

本文将采用构建的术语抽取迁移学习语料对基于 EM 的术语抽取迁移方法进行效果验证。设置如下实验：

1）以基于互信息的术语抽取方法（MITE）为 Baseline1；

2）基于朴素贝叶斯分类的术语抽取方法（NBTE）为 Baseline2；

3）实验方法为基于 EM 的术语抽取迁移方法（EMTE）。

三种术语抽取方法均以相同的候选术语抽取方法抽取出候选术语，候选术语抽取方法详见本文的 4.3.1 节。MITE 是一种基于统计的术语抽取方法，可独立地对计算机领域和环境领域语料进行术语抽取，不需要训练语料和测试语料。NBTE 方法为基于机器学习的术语抽取方法，需要训练语料和测试语料，且一般假设训练语料和测试语料是属于同一领域的同分布语料。EMTE 方法为本文提出的术语抽取迁移学习方法，是一种实现了跨领域术语抽取的学习方法。该方法需要训练语料和测试语料，但训练语料和测试语料不必为同一领域的同分布语料。

本文设置两组不同的实验，实验组 1 采用计算机领域标注语料进行实验。MITE 方法直接对 50 篇文本进行术语抽取，NBTE 与 EMTE 以 100 篇计算机文本进行训练，50 篇测试。

实验组 2 采用语料为计算机领域标注语料和环境领域未标注语料。MITE 方法直接对环境领域术语进行抽取。NBTE 方法和 EMTE 方法以计算机领域标注文本集合为训练数据集合，对环境领域术语进行抽取。

表 4- 2 术语抽取迁移学习方法实验设置

实验方法		训练语料		测试语料	
		计算机	环境	计算机	环境
实验组 1	MITE1	-	-	50	-
	NBTE1	100	-	50	-
	EMTE1	100	-	50	-
实验组 2	MITE2	-	-	-	150
	NBTE2	150	-	-	150
	EMTE2	150	-	-	150

实验组 1 旨在探索三种术语抽取方法在同一领域语料中的实验效果，验证基于机器学习方法的术语抽取方法的有效性，即 NBTE 方法效果是不弱于传统基于统计的术语抽取方法。实验组 2 旨在验证本文提出的 EMTE 术语抽取方法在不同领域间术语抽取任务中的效果。

4.5.3 结果与分析

术语抽取任务结果样例如表 4-3 所示。表 4-3 中用斜体表示由本文术语抽取方法获取的和人工标注的术语完全相同的术语。术语抽取迁移学习方法实验结果如表 4-3 所示。

表 4- 3 术语抽取任务结果样例表

文档及实验方法		术语抽取结果（Top 5）
文档 1（计算机）	MITE1	*遗传算法*；*遗传 算子*；简单 比例 因子；自适应 模糊 控制器；仿真 结果 差异
	NBTE1	*遗传算法*；神经 模糊 控制；*模糊 控制器*；*遗传 算子*；变异 概率
	EMTE1	*遗传算法*；*模糊 控制器*；*遗传 算子*；神经 模糊 控制；变异 概率
文档 2（环境）	MITE2	*有机物*；*三氯乙烯*；催化 还原 脱氯 特性；*催化剂*；*零价*；催化剂 制备 试剂
	NBTE2	*碱性 溶液*；*三氯乙烯*；*零价*；*加氢 催化剂*；恒温 摇床
	EMTE2	*脱氯 反应*；*三氯乙烯*；*零价*；*催化剂*；表面 杂质

表 4-4 术语抽取迁移学习实验结果

		Top 5			Top 10		
		EP	CP	AvgP	EP	CP	AvgP
实验组 1	MITE1	56.87%	62.37%	59.62%	61.28%	65.48%	63.38%
	NBTE1	65.42%	69.29%	67.36%	61.33%	72.29%	69.31%
	EMTE1	66.01%	72.26%	69.14%	67.57%	79.32%	73.45%
实验组 2	MITE2	50.19%	59.32%	54.76%	55.93%	60.79%	58.36%
	NBTE2	58.58%	69.01%	63.80%	59.01%	69.98%	64.50%
	EMTE2	62.93%	74.42%	68.68%	63.13%	77.78%	70.46%

上表展示了取 Top N（N = 5、10）作为术语抽取结果时，两组实验中三种术语抽取方法的评价指标值，趋势图如图 4-3、图 4-4 所示。

图 4-3、图 4-4 中，每条趋势线的前三个节点和后三个节点分别表示 N 取值为 5 或 10 时的评价指标值。

首先，由实验结果可知，术语抽取方法准确率普遍较低。且一般来说，覆盖准确率 CP 的值均略高于精确准确率 EP 的值。覆盖准确率 CP 最高也只能达到 80%左右，而精确准确率 EP 最高值也只有 65%。这说明术语抽取方法仍有很大进步空间，尤其是不同领域间的术语抽取迁移方法还有待进一步探索。

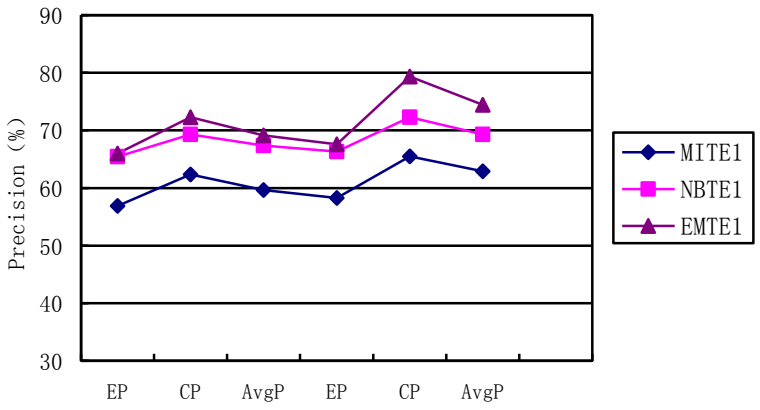

图 4-3 实验组 1 的实验结果图

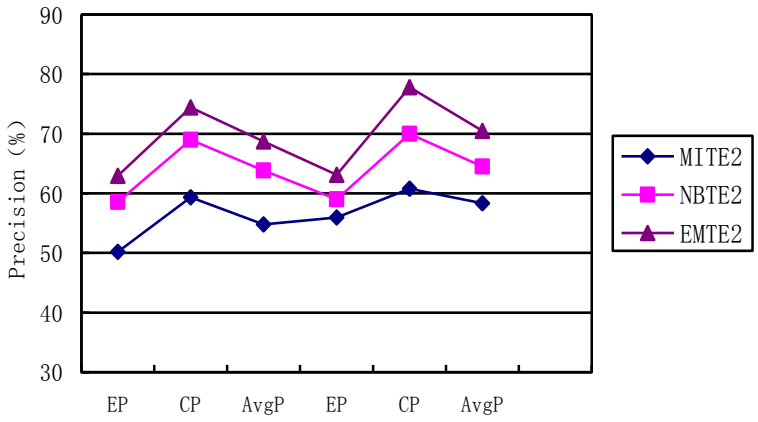

图 4-4 实验组 2 的实验结果图

其次，由上述实验结果图中的三种方法实验结果对比可知，无论是同领域或跨领域术语抽取任务，EMTE 术语抽取方法效果优于 NBTE 术语抽取方法，优于 MITE 术语抽取方法。

具体来说，实验组 1 中，NBTE 方法比 MITE 方法考虑了更多的候选术语信息，故效果优于 MITE。相比于 NBTE 方法，EMTE 术语抽取方法除了训练数据模型外，还将 100 篇计算机领域文本中的相关知识应用于 50 篇测试文本的术语抽取中，使得术语抽取效果稍高于 NBTE。实验组 2 的实验结果中 EMTE 算法平均准确率高于 NBTE 及 MITE 说明借助了计算机领域知识后的迁移学习方法在抽取环境领域术语时起到了良好的辅助作用。

实验组 2 中 EMTE 方法与 NBTE、MITE 方法的效果差异大于实验组 1 中的效果差异。这说明，在术语抽取迁移问题上，EMTE 方法展现了其独特优势。

可以看出，实验组 2 中各方法的准确率均在一定程度上低于实验组 1 中准确率。说明跨领域的术语抽取方法的实验效果略低于同一领域中的术语抽取效果。

实验组 1 和 2 的实验结果对比，证明了在传统机器学习问题和迁移学习问题上，本文提出的基于 EM 的术语抽取迁移学习方法都更有利于提高领域术语的抽取效果。

4.6 本章小结

本章在介绍了术语的概念、类别体系、分析了其相关语言特性后，介绍了传统术语抽取方法，其中本文采用的候选术语抽取方法和传统术语抽取方法中的候选术语抽取方法相同。随后，阐述了一种新颖的基于朴素贝叶斯分类的中文术语抽取方法，并将基于 EM 的直推式迁移学习方法应用于术语抽取迁移任

务中，提出基于 EM 的术语抽取迁移方法，最后相关实验结果，验证了该方法的有效性。

结　论

针对缺少目标领域标注语料的迁移学习问题，本文构建了一种基于 EM 的直推式迁移学习模型。该方法从已标注的源领域数据中学习得到一个数据模型，以数据模型参数为迁移知识，通过 EM 算法将迁移知识用于辅助目标领域数据学习得到目标领域数据模型完成应用任务。本文完成的主要工作有：

（1） 构建了基于 EM 的直推式迁移学习模型；

为解决本文中的迁移学习问题，以目标领域未标注数据的标记信息为隐含变量，并利用 EM 算法估计隐含变量值的优势，从源领域中获取迁移知识，用 EM 算法将迁移知识适用于目标领域的任务学习。同时分析说明了基于 EM 的直推式迁移学习方法中的获取的迁移知识形式及知识获取方法。

（2） 针对文本分类问题，研究了基于 EM 的文本分类迁移学习方法；

基于 EM 的文本分类迁移学习方法是基于 EM 的直推式迁移学习方法在文本层次上的具体应用。该方法从源领域中学习得到一个朴素贝叶斯分类器，以其模型参数作为迁移知识，辅助完成目标领域文本分类任务。本文基于中文文本分类语料构建了中文文本分类迁移语料，并通过不同领域间的迁移实验验证了该方法的有效性。

（3） 研究了基于 EM 的迁移学习方法在术语抽取中的应用；

探索了一种基于机器学习方法的中文术语抽取方法。采用朴素贝叶斯分类器获取候选术语的术语度，进而抽取出文档中出现的领域术语。同时，在基于 EM 的直推式迁移学习方法的基础上，研究实现了基于短语层次的术语抽取迁移学习方法。在从复旦语料中构建的术语抽取语料上的实验结果证明基于 EM 的术语抽取迁移学习方法能够解决目标领域缺乏标注语料问题。

除文本已完成工作之外，还有一些可改进之处：

（1）本文分别构建了实验所用的中文文本分类和术语抽取迁移语料，原始语料并非专用的迁移学习语料，故语料的内容普适性较低，构建方法还有待进一步科学化和规范化，研究界也有待共同规范化针对中文应用任务研究所用的迁移语料。

（2）在术语抽取方法中，采用分类器获取候选术语列表后，还可考虑融入语言学知识，对获得的候选术语列表进行重排序，以提高术语抽取的召回率。一个思路是，根据上下文信息建立一个指示词词表，根据候选术语附近产生的指示词信息计算候选术语的术语度，并与分类后获得的候选术语的术语度概率值进行加权平均，从而得到最后的术语度统计值。由于各领域的指示词表大同小异，这种基于指示词的术语重排序技术具有一定的迁移性。

（3）基于 EM 的直推式迁移学习方法属于基于参数的迁移学习方法，迁移知识是分类模型的参数。术语抽取是基于短语层次的文本分析任务。可以尝试使用基于特征的迁移学习方法实现术语抽取迁移学习任务。

参考文献

[1] 曾华军, 张银奎等译, Tom M. Mitchell. Machine Learning[M]. 北京: 机械工业出版社, 2003:3.

[2] Rothenhausler K, Schutze H. Unsupervised Classification with Dependency Based Word Spaces[C]. Athen:Association for Computational Linguistics. Proceedings of the EACL 2009 Workshop on GEMS: GEometical Models of Natual Language Semantics. 2009:17-24.

[3] Slonim N, Friedman N., Tishby N. Unsupervised Document Classification using Sequential Information Maximization[C]. Tampere: Association for Computing Machinery. SIGIR. 2002:11-15.

[4] Halder A, Ghosh S, Ashish G. Ant Based Semi-supervised Classification[C]. Brussels:Springer Verlag. Proceedings of the 7th International Conference on Swarm Intelligence. 2010:376-383.

[5] Yang G, Xu X, Yang G, et al. Semi-supervised Classification by Local Coordination[C]. Sydney:Springer Verlag. Proceedings of the 17th International Conference on Neural Information Processing. 2010:517-524.

[6] De L. Arcanjo F, Pappa Gisele L, Bicalho Paulo V., et al. Semi-supervised genetic programming for classification[C]. Dublin: Association for Computing Machinery. Proceedings of the 13th Annual Conference on Genetic and Evolutionary Computation. 2011:1259-1266.

[7] Pan Sinno J.L, Yang Q. A Survey on Transfer Learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(10):1345-1359.

[8] 李保利, 陈玉忠, 俞士闻. 信息抽取研究综述[J]. 计算机工程与应用, 2003, 39(10):1-5.

[9] 周浪. 中文术语抽取若干问题研究[D]. 南京:南京理工大学, 2009: 6-6.

[10] Pan J.L. Feature-Based Transfer Learning with Real-World Applications[D]. Hong Kong: The Hong Kong University of Science and Technology, 2010:1-95.

[11] Tan S.B, Cheng X.Q. Improving SCL Model for Sentiment-Transfer Learning[C]. Boulder:Association for Computational Linguistics. Proceedings of NAACL HLT. 2009:181-184.

[12] Jiang J, Zhai C.X. Instance Weighting for Domain Adaptation in NLP[C]. Prague:Association for Computational Linguistics. Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. 2007:264-271.

[13] Dai W.Y, Yang Q, Xue G.R, Yu Y. Boosting for Transfer Learning[C].

Corvalis:Association for Computational Linguistics. Proceedings of the 24th International Conference on Machine Learning, 2007:193-200.

[14] Blitzer J, McDonald R, Pereira F. Domain Adaptation with Structural Correspondence Learning[C]. Sydney:Association for Computational Linguistics. Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, 2006:120-128.

[15] Xue G.R, Dai W.Y, Yang Q, et al. Topic-bridged PLSA for Cross-Domain Text Classification[C]. Singapore:Association for Computational Linguistics. SIGIR. 2008:627-634.

[16] Feng G.Z, Guo J.H, Jing B.Y, et al. A Bayesian Feature Selection Paradigm for Text Classification[J]. Information Processing & Management, 2012, 48(2):283-302.

[17] Tahir M.A, Kittler J, Bouridane A. Multilabel Classification Using Heterogeneous Ensemble of Multi-label Classifiers[J]. PATTERN RECOGNITION LETTERS, 2012, 33(5):513-523.

[18] Chen J, Tan J.L, Liao H. Meaningful String Extraction Based on Clustering for Improving Webpage Classification[J]. CHINA COMMUNICATIONS, 2012, 9(3):68-77.

[19] Pan Sinno J.L, Ni X.C, Sun J.T, et al. Cross-Domain Sentiment Classification via Spectral Feature Alignment[C]. Raleigh:Association for Computing Machinery. Proceedings of the 19th International Conference on World Wide Web, 2010:751-760.

[20] Zhu Z.F, Zhu X.Q, Guo Y.F, et al. Transfer incremental learning for pattern classification[C]. Toronto:Association for Computing Machinery. Proceedings of the 19th International Conference on Information and Knowledge Management and Co-located Workshops. 2010:1709-1712.

[21] Dai W.Y, Xue G.R, Yang Q, et al. Transfer Naïve Bayes Classifiers for Text Classification[C]. Vancouver:American Association for Artificial Intelligence. Proceedings of the 22nd AAAI Conference on Artificial Intelligence. 2007:540-545.

[22] 张锋, 许云, 侯艳, 樊孝忠. 基于互信息的中文术语抽取系统[J]. 计算机应用研究, 2005, (5):72-73.

[23] Frantzi K, Ananiadous S, Mima H. Automatic Recognitation of Multi-Word Terms: The C-value/NC-value Method[J]. International Journal on Digital Libraries, 2000, 3(2):115-130.

[24] Justeson J.S, Slava M. K. Technical terminology: some linguistic properties and an algorithm for identification in text[J]. Natural Language Engineering, 1995:9-27.

[25] Maynard D, Ananiadou S. Identifying Contextual Information for multi-Word

Term Extraction[C]. Sandrini. Terminology and Knowledge Engineering, 1999:212-221.

[26] Ji L.N, Sum M.T, Lu Q, et al. Chinese Terminology Extraction Using Window-Based Contextual Information[C]. Mexico:Springer Verlag. 8th Annual Conference on Intelligent Text Processing and Computational Linguistics. 2007:62-74.

[27] 杨宇航. 领域自适应的弱指导信息抽取关键技术研究[D]. 哈尔滨: 哈尔滨工业大学, 2009:22-49.

[28] Huang J.Y, Alexander J.S, Archur G, et al. Correcting sample selection bias by unlabeled data[C]. Vancouver:MIT Press. In Proceedings of the 19th Annual Conference on Neural Information Processing Systems, 2007.

[29] Sugiyama M, Nakajima S, Kashima H, et a1. Direct importance estimation with model selection and its application to covariate shift adaptation[C]. Vancouver:Curran Associates Inc. In Proceedings of the 20th Annual Conference on Neural Information Processing Systems, 2008.

[30] Bickel S, Bruckner M, Scheffer T. Discriminative learning for differing training and test distributions[C]. Corvalis:Association for Computing Machinery. In Proceedings of the 24th international conference on Machine learning. 2007:81-88.

[31] Blitzer J, Dredze M., Pereira F. Biographies, bollywood, boomboxes and blenders: Domain adaptation for sentiment classification[C]. Prague:Association for Computational Linguistics. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. 2007:432-439.

[32] Dempster A.P, Laird N.M, and Rubin D.B. Maximum likelihood from incomplete data via the EM algorithm[J]. Journal of the Royal statistical Society, 1977, 39:1-38.

[33] Lashkari A.H, Mahdavi F, Ghomi V. A Boolean Model in Information Retrieval for search Engines[C]. Kuala Lumpur:IEEE Computer Society. Proceedings of the 2009 International Conference on Information Management and Engineering, 2009:385-389.

[34] Salton G., Wong A, Yang C.S. A Vector Space Model for Automatic Indexing[J]. Communications of the ACM, 1975, 18(1):613-620.

[35] Kuncheva L.I. Fitness Functions in Editing KNN Reference Set by Genetic Algorithms[J]. Pattern Recognition. 1997, 30(6):1041-1049.

[36] Wang K, Zhou S.Q. Hierarchical Classification of Real Life Documents[C]. Valencia:INSTICC. In Proceedings of the First Siam International Conference on Data Mining. 2001.

[37] Lang K. Newsweeder: Learning to Filter Netnews[C]. Tahoe:Association for Computing Machinery. In International Conference on Machine Learning. 1995:331-339.

[38] Salton G. Term Weighting Approaches in Automatic Text Retrieval[J]. Information Proeessing and Management.1998, 24(5):513-523.

[39] Ronald R. Yager. An Extension of the Na ïve Bayesian Classifier[J]. Information Sciences. 2006, 176(5):577-588.

[40] Wang L, Khan L, Thuraisingham B. An Effective Evidence Theory Based K-Nearest Neighbor (KNN) Classification[C]. Sydney:Inst. of Elec. and Elec. Eng. Computer Society. Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology. 2008:797-801.

[41] Manikandan J, Venkataramani B. Design of a modified one-against-all SVM classifer[C]. San Antonio:Institute of Electrical and Electronics Engineers Inc.. Proceedings of the 2009 IEEE International Conference on Systems, Man and Cybernetics. 2009:1869-1874.

[42] Breslow Leonard A, Aha David W. Simplifying Decision Trees: A survey[J]. The Knowledge Engineering Review. 1997, 12(1):1-40.

[43] Kavzoglu T. Increasing the Accuracy of Neural Network Classification using Refined Training Data[J]. Environmental Modelling & Software, 2009, 24(7):850-858.

[44] Quinlan J. R. Introduction of Decision Trees[J]. Machine Learning. 1986, 1(1):81-106.

[45] Chen B, Lam W, Tsang I, et al. Extracting Discriminative Concepts for Domain Adaptation in text Mining[C]. Paris:Association for Computing Machinery. In Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2009:179-187.

[46] Zhang X, Dai W.Y, Xue G.R, et al. Adaptive Email Spam Filtering Based on Information Theory[C]. Nancy:Springer Verlag. In Proceedings of the 8th International Conference on Web Information Systems Engineering. 2007:159-170.

[47] 杜俊卫. 基于聚类的文本迁移学习算法研究及应用[D]. 太原:山西财经大学, 2011: 19-19.

[48] 谭松波，王月粉．中文文本分类语料库-TanCorpV1.0[OL]. http://www.searchforum.org.cn/tansongbo/text_corpus.jsp.

攻读硕士学位期间发表的论文及其它成果

[1] Yanxia Qin, Dequan Zheng, Bing Xu. Search Results Optimization Method Combined with Multi-features[C]. The 8th International Conference on Fuzzy Systems and Knowledge Discovery, 2011, Vol.2: 1167-1171.（EI 收录号：20114014401118）

[2] Yanxia Qin, Dequan Zheng, Tiejun Zhao. Research on Search Results Optimization Technology with Category Features Integration[J]. International Journal of Machine Learning and Cybernetics, 2012, 3(1), 71-76.（EI 收录号：20121014833479）

学位论文原创性声明

本人郑重声明：此处所提交的学位论文《直推式迁移学习及其应用研究》，是本人在导师指导下，在哈尔滨工业大学攻读学位期间独立进行研究工作所取得的成果。据本人所知，论文中除已注明部分外不包含他人已发表或撰写过的研究成果。对本文的研究工作做出重要贡献的个人和集体，均已在文中以明确方式注明。本声明的法律结果将完全由本人承担。

作者签名：秦彦霞　　日期：2012 年 7 月 1 日

致　谢

两年的硕士学习生涯即将结束，预示着下一个更有挑战性的博士生涯的到来。从三年前第一次来哈尔滨，并特地跑来瞻仰哈尔滨工业大学的那一刻开始，也许就注定了自己将在这里度过人生中最重要的六年。两年间我取得的进步都要归功于在哈尔滨工业大学遇到的所有人。

感谢我的导师郑德权副教授。慈父般的郑老师为人和善，做人处事态度积极，热爱运动，关心学生，总是给人以希望。这一切都让我非常庆幸能够成为郑老师的学生，也坚定了我若以后成为老师，必定以郑老师为榜样。

感谢实验室的其他所有老师们。高瞻远瞩的李生教授教育我们要规划硕士生涯，他的言传身教让实验室所有人受益终生。赵铁军教授作为语言技术中心负责人和学院副院长，勤恳认真的工作科研态度为我们树立了良好榜样，也希望博士期间能在实验室良好氛围中成长更快。杨沐昀副教授的严格认真让我深刻了解了“严师出高徒”这一道理。感谢朱聪慧老师带领我们将代码中心的项目工作引入正轨，关心我们的学习生活，还传授给我们做人做研究的道理。感谢徐冰老师向我展现了自信的女研究学者风范。感谢曹海龙老师为实验室的学术发展做出的努力。感谢邹馨蕊秘书对实验室的无私付出，保证了实验室的后勤事务正常运行。

感谢实验室的各位师兄师姐师弟师妹。与陈宇、李风环等实验室同学们交流让我学到了丰富的科研方法及经验，对你们给予的无私帮助表示感谢，与你们的友谊让我收益匪浅。祝你们研究学术的博士早毕业，工程的工作顺利。感谢实验室同届的其他 10 位同学，韩威、孙振龙、何小春、田乐逍、肖冬青、吴建伟、胡鹏龙、王澍、吴松、高景斌，有了你们，咱们这届才成为一个欢乐的小团体。

感谢我的室友李清及其他朋友，与你们成为朋友是我一生的财富。

感谢我的家人和男朋友，你们陪我度过一次一次的困难，在我疲惫时你们的爱给予了我希望和勇气。

感谢其他所有帮助关心过我的老师和同学们。谢谢你们！

阅读此文的还阅读了：

26. 学习迁移与样例研究的发展

27. 化学信息题的学习迁移原理及其应用

28. 学习迁移能力研究综述

29. 概念迁移研究及其现实意义

30. 竖直上抛运动及其迁移应用

31. 解析英语学习迁移的实际应用

32. 迁移理论及其应用

33. 学习迁移研究的新视角:角色导向迁移观

34. 直推式迁移学习及其应用研究

35. 迁移学习研究综述

36. 应用学习迁移,加快教学进程

37. 合作学习的理论研究及其应用

38. 学习者与学习迁移的矛盾

39. 课程整合背景下的学习迁移理论研究

40. 学习的迁移规律及其应用

41. ThinkQuest项目学习模式及其应用研究

42. 学习迁移理论在农民工培训中的应用研究

43. 基于压缩编码的迁移学习算法研究

44. 迁移学习相关理论研究

45. 迁移学习和领域自适应在表示学习中的应用

46. 解析英语学习迁移的实际应用

47. 学习迁移在计算机教学中的应用研究

48. 基于案例推理的学习迁移研究

49. 选择性直推式迁移学习

50. 基于直推式学习和迁移学习方法改进的支持向量机分类方法及应用研究