中文图书分类号：TP18
密　　级：公开
UDC：004
学 校 代 码：10005

硕 士 学 位 论 文

MASTERAL DISSERTATION

论 文 题 目： 基于半监督学习的文本分类算法研究

论 文 作 者： 杜芳华

学　　科： 计算机科学与技术

指 导 教 师： 冀俊忠 教授

论文提交日期： 2014 年 6 月

UDC：004
中文图书分类号：TP18
学校代码：10005
学　　号：S201107089
密　　级：公开

北京工业大学工学硕士学位论文

题　　目： 基于半监督学习的文本分类算法研究

英文题目： RESEARCH ON TEXT CLASSIFICATION ALGORITHMS BASED ON SEMI-SUPERVISED LEARNING

论 文 作 者： 杜芳华
学 科 专 业： 计算机科学与技术
研 究 方 向： 计算机应用技术
申 请 学 位： 工学硕士
指 导 教 师： 冀俊忠教授
所 在 单 位： 计算机学院
答 辩 日 期： 2014 年 6 月
授予学位单位： 北京工业大学

独 创 性 声 明

本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得北京工业大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

签　名：杜芳华

日　　期：2014 年 6 月 12 日

关于论文使用授权的说明

本人完全了解北京工业大学有关保留、使用学位论文的规定，即：学校有权保留送交论文的复印件，允许论文被查阅和借阅；学校可以公布论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存论文。

（保密的论文在解密后应遵守此规定）

签　　名：杜芳华　　日　　期：2014 年 6 月 12 日

导师签名：冀俊忠　　日　　期：2014 年 6 月 12 日

摘 要

文本分类技术是目前文本挖掘领域中的研究热点之一。由于现实世界中存在着大量的无标记的文本数据，而获取有标记的文本数据，需要耗费巨大的人力、物力和财力，因此能够充分使用无标记数据的半监督文本分类技术具有非常重要的科研价值。而且伴随大数据时代的来临，半监督文本分类在组织、管理和处理海量无序的互联网文本数据上更能显现出技术的优势。于是，半监督文本分类越来越多地引起了国内外众多学者的广泛关注。本文基于蚁群算法和迁移学习相关技术，针对半监督文本分类中数据分布不一致可能导致分类器性能下降的问题，进行了比较深入的研究，主要工作包括：

(1) 提出了一种基于蚁群聚集信息素的半监督文本分类算法。算法将聚集信息素与传统的文本相似度计算相融合，首先，基于聚集信息素的作用，利用 Top-k 策略选取出未标记蚂蚁可能归属的种群；然后，依判断规则，判定未标记蚂蚁的置信度；最后，采用随机选择的策略，把置信度高的未标记蚂蚁加入到对其最有吸引力的一个训练种群中。与朴素贝叶斯算法和 EM 算法在标准数据集上对比实验表明：该算法在精确率、召回率以及 F_1 度量方面，都取得了更好的效果。

(2) 提出了一种基于特征映射的半监督文本分类算法。首先通过不同的特征选择方法，分别在训练集的已标记数据、未标记数据以及测试集数据选取各自的特征集，并初始化特征的权值；在此基础之上，建立已标记数据与未标记数据、已标记数据与测试集数据、未标记数据与测试集数据之间的映射函数，利用这三个特征映射函数重新计算特征的权重；最后利用 EM 算法进行半监督文本分类。在标准数据集上的实验表明，本文提出的基于特征映射的半监督文本分类算法是有效的，能够解决数据分布不一致可能导致半监督文本分类器性能下降的问题。

关键词：蚁群算法；聚集信息素；特征映射；半监督学习；文本分类

Abstract

Text classification is one of the hotspots in the field of text mining. There are a large number of unlabeled text data in the real world, and it takes huge manpower and financial resources to get labeled text data, so it is very significant to study semi-supervised text classification that make use of unlabeled text data. With the entry of big data era, the semi-supervised text classification is more efficient than other methods in the organization and management of massive disorder Internet text data. It also attracts more and more attention and research of scholars both at home and abroad. This paper research on the problem in semi-supervised text classification based on ant colony algorithm and transfer learning, which inconsistent data distribution may lead to performance degradation, the main work includes:

(1) This paper presents a semi-supervised text classification algorithm based on aggregation pheromone, which was used for species aggregation in real ants and other insects. The proposed method, which has no assumption regarding the data distribution, can be applied to any kind of data distribution. Firstly, in light of aggregation pheromone, colonies that unlabeled ants may belong to are selected with a top-k strategy. Then the confidence of unlabeled ants is determined by a judgment rule. Finally, unlabeled ants with higher confidence are added into the most attractive training colony by a random selection strategy. Compared with naïve Bayes and EM algorithm, the experiments on benchmark dataset show that this algorithm performs better on precision, recall and Macro F1.

(2) We have proposed a semi-supervised text classification algorithm based on feature mapping. First, we select three respective sets of features from labeled data, unlabeled data and test data using different feature selection methods, and initialize their value; Second, three feature mapping functions are learned , then the weight of each feature is recalculated by them; Finally, the EM algorithm classifies text data. Experiments on standard data sets show that the proposed algorithm is effective.

Keywords： ant colony algorithm; aggregation pheromone; feature mapping; semi-supervised learning; text classification

目 录

第1章 绪论

1.1 研究背景与意义

近年来随着信息技术的高速发展以及大数据时代的来临，互联网文本数据呈现出海量性和无序性的特性，这种特性使人们无法轻易的获得文本数据中潜在的价值信息。因此，能够有效处理文本数据的文本分类技术，逐渐引起了大家的关注和研究[1]。

基于机器学习的分类算法可以分成有监督的分类[2-6]和无监督（或非监督）的聚类[7-8]。有监督的分类由训练学习和类别判定两部分组成。在训练学习过程中，通过利用大量的有标签数据训练学习得到一个分类模型；在类别判定过程中，利用已获得的分类模型对无标签的测试数据进行类别判定。无监督的聚类则是指只利用无标签的数据，将它们划分成不同的簇，使得同一簇中的数据尽可能地相关，不同簇中的数据尽可能地无关。它们作为分类领域中两种最重要的技术，都有自己的优点和缺点。有监督分类技术的优点是，通过大量已标注数据训练学习到的分类模型比较准确，其整体的分类效果相对较好；缺点是现实世界中并不存在大量的已标注数据，而获取大量已标注数据由于所需代价太高往往不太现实。无监督聚类的优点是，整个聚类过程是自动完成的，不用借助外在影响；缺点是由于没有已标记数据的帮助，无法达到令人满意的性能。面临这一情况，半监督学习技术的产生就水到渠成了 [9-12]。

不同于传统的分类方法，半监督学习需要的是一个包含较少的已标注数据和较多的无标注数据的训练集，从而改善了传统的分类技术的劣势。按照对无标签数据和有标签数据使用方法的差异，能够将半监督学习技术分成半监督分类和半监督聚类两种技术。基于半监督学习的分类通过使用较多的没有类别信息的数据来帮助有监督学习，从而提高分类器的性能。而基于半监督学习的聚类算法则使用有标签数据的类别信息，来指导对无标签数据进行簇的划分。

在文本挖掘领域中，半监督学习技术具有非常明显的优势。由于文本数据的特殊性，因此对文本数据进行大量标记比较困难。原因是：首先，大数据时代下的互联网文本数据数量巨大且类别繁多，若标记大量文本将花费非常多的人力、物力和财力；其次，随着时间的推移，新获取的文档中可能会包含新的主题种类，而且原有的主题类也有可能消失，出现类别减少的情况，导致原来的分类器将不再适用于新增文档；最后，不同的用户有着自己独有的文本分类需求，因此单一

的标记规则已经无法赢得用户的认可。因此，半监督学习技术特别适用于文本分类领域。

1.2 国内外研究现状

1.2.1 文本分类研究现状

文本自动分类是一个交叉学科的产物，其涉及的领域包括数理统计、机器学习、文本挖掘和自然语言处理等，其理论研究能够追溯到上世纪 60 年代早期[13]。它的发展过程可以分为三个阶段：

第一阶段是二十世纪六十年代至八十年代。20 世纪 60 年代初，Maron[14]首先提出对自动文本分类的研究。那时文本自动分类主要应用于信息检索系统的科学文章自动标引。同时，信息检索和模式识别逐渐发展成为两门具有重要影响力的学科。Rosenblatt[15]设计了感知器（Perceptron），利用阈值神经元来进行完成二元分类任务；Gerald Salton[16]提出了对于文本挖掘领域产生深远影响的向量空间模型（Vector Space Model, VSM）用来表示文本数据。在这个阶段，学者们主要研究文本分类理论。

第二阶段是 20 世纪 80 年代。这一阶段以知识工程（Knowledge Engineering）技术为主，首先利用专家提供的专业领域知识抽象出规则，然后依靠手工建立分类器。由于规则只适用于某些特定的领域，因此很难将它们移植到新的领域。这一时期的代表是由 Hayes[17]等人设计实现的 Construe 系统。

第三阶段是 20 世纪 90 年代以后。这一阶段，文本分类技术借鉴了机器学习和统计分析的思想，利用训练集来学习到效果更优的分类模型。文本分类模型可以自动创建，不再依赖于相关领域的专家。而且分类的效率和精确率都比之前有着大幅度的提高[18]。

当前所说的文本分类一般是指以机器学习和统计理论为基础的文本分类。因此，从现在的技术角度来看文本分类的研究历史只有短短的二十年时间。本文的研究内容也是以机器学习和统计理论为基础的文本自动分类技术。

国内对文本自动分类技术的研究时间相对较短，到上世纪八十年代，国内的学者才开始从事文本自动分类技术的研究，但是经过这些年的发展，已经获得了长足的进步。由于汉语与英语具有不同的语言特点，因此直接照搬人家的研究是不可取的。国内文本分类的研究主要是通过学习国外文本分类的技术思想，再融合我们中文语言的特点，最后形成了我们自己的文本分类研究体系。中文文本分类主要有两种方法：一种是基于语义理解的分类方法，另一种是基于机器学习的

分类方法。基于语义理解的方法其主要思想是让机器从人类思维的角度来尝试理解文本的语言含义来进行分类，但是这种方法的缺点是容易受到其他相关技术的限制；基于机器学习的分类方法是从统计的角度出发，利用组成文本数据的基本单位来完成分类。到目前为止，我国在文本自动分类技术的应用研究中已经取得了很多令人瞩目的成果，例如百度、搜狗和 360 等搜索引擎都得到了广泛的应用，并且已经融入到了我们的日常生活之中。而且文本自动分类作为信息管理等领域的核心技术，受到了国内很多学者的关注和研究。

1.2.2 半监督学习研究现状

对于半监督学习的研究，一般认为是由 Shahshahani 和 Landgrebe[19] 的工作开始的。但在二十世纪八十年代后期，就已经有学者逐渐认识到如何利用和挖掘未标注数据的潜在价值了。近年来，一是机器学习技术开始进入了高速发展的时期，二是如何利用未标注数据的潜在价值变的越来越急切。在这两大外在因素的推动下，半监督学习迅速引起了很多学者的关注，并开始成为数据挖掘和机器学习领域的一个研究热点。

目前，对半监督学习的研究工作主要依据两种基本假设[20]：一是聚类假设（cluster assumption）；二是流形假设（manifold assumption）。

聚类假设认为处在同一个簇中的样本有较大的可能具有相同的标签。根据这一假设，半监督学习的决策函数边界应当通过样本密度比较小的地方，从而避免处于样本密度较大的数据却被分到不同的类中这一情况的发生。在这种假设下，大量未标注样本的作用其实就是为了找出数据分布密度不同的样本空间区域，从而不断优化学习到的决策函数，最终达到让它以最大概率通过样本分布密度较小的区域的目的。

流形假设认为，如果样本处在一个小的局部邻域内，那么他们具有类似的属性，所以他们的标签也应该是类似的。从而表明决策边界具有局部平滑的特性。由上可知，聚类假设主要考虑整体，而流形假设更加注重部分特征。依据这一假设，大量的未标注样本的作用其实就是为了增大样本空间的密度，进而更加准确的描述局部区域，最终使决策模型可以较好地拟合数据。

目前，半监督学习算法可以分为以下几种：

(1) 生成式模型算法

这种算法直接使用了第一种聚类假设，将生成式模型（Generative Models）训练学习成为分类器，而生成式模型的参数其实就是未标注数据属于所有类别的概率值，最后使用 EM（Expectation Maximization）算法来判定样本的标注，并

计算出模型参数。EM 算法是半监督学习中使用的第一种方法，其主要思想是将半监督学习问题转化为求解不完全数据问题。生成式模型算法实际上可以被视为是在较少的已标注样本附近进行聚类。Nigam[21]等人将其用于文本分类；Baluja[22]等人将其用于人脸识别；Fujino[23]等人通过使用“偏差修正”（bias correction）和最大熵原理区分训练扩展了生成式模型。

(2) 基于图的半监督学习算法

这种学习算法使用了流形假设。它的主要思想是，首先需要构建一张包含已标注和未标注数据的图，图中的节点表示训练集数据，边则表示节点之间的相似度（或距离），然后确定优化目标函数，最后用平滑的决策模型来计算出最优的模型参数。这种算法其实是将半监督学习问题转化为图的求解问题。图中边的权值决定了其连接的两个节点是否属于同一个类。权值大的边比权值小的边拥有更高的优先级来扩散类别信息，因此权值大的边上的节点更容易将类别信息扩散到其相邻的节点。

这种算法是无参数识别方法，其核心是求解一个图的函数，该函数还必须满足以下两个条件：一是在已标注节点它计算出的类别应该与给定的类别比较接近；二是它必须在整个图上是光滑的。

Blum 和 Chawla[24]把半监督学习问题变换成图的最小割问题，首先以数据间的相似度为基础生成一个图，然后通过最小化带有不同类别标签相似对数据个数来进行分类。Pang 和 Lee[25]基于彼此相似的句子应该属于相同类别的假设，改善了句子分类问题。Brian[26]等学者实现并证明了半监督学习在不同表示类型数据上的等价性。

(3) 协同训练（Co-Training）算法

这种算法使用了前面所述的两种假设。其采用两个或更多的学习器，在学习时，它会挑选多个比较可信的没有类别信息的数据进行相互判定，以便更新模型。目前，有很多学者对其进行了深入的研究，取得了很多优异的成绩，使协同训练算法成为半监督学习中最重要的一种典范（paradigm）[20]，而不仅仅只是一种方法。

1998 年 Blum 和 Mitchell[27]第一次提出了协同训练算法。他们假定数据集有两个完全冗余的视图，也就是符合设定条件的属性集合：第一，如果训练集数据非常多，那么通过在属性集上的训练学习，就能够获得一个比较强的学习器；第二，在给定标签时，它们之间都是条件相互独立的。他们的算法在不同的视图上使用有标签的数据分别学习到一个模型，然后，每个模型对无标签数据中比较可信的数据进行标注，然后将这些有标签的数据作为另外一个不同的学习器的训练集数据，以此来更新学习模型。协同训练是一个重复循环训练的过程，直到其满

足给定的终止条件。但是在现实世界中，他们假定的条件往往很难实现。因此，很多学者开始研究不需要充分冗余条件的算法。

Goldman 和 Zhou[28]提出了一种改进的协同训练算法。他们使用不同的决策树算法，在预测阶段，首先通过对比不同模型对无标签数据进行标注的置信度，然后选择置信度高的模型来进行判定。但这种算法对学习模型的类别有特殊要求，而且算法的计算效率比较低。为了使协同训练算法更具普适性，Zhou 和 Li[29]提出了一种 tri-training 算法。这种算法使用了三个学习模型，可以轻松应对标签置信度计算问题和无标签数据的判定问题，还可以通过使用集成学习（ensemble learning）来扩展算法的应用范围。首先通过选取不同的有标签数据来建立三个不同的训练集，然后利用这三个不同的训练集分别训练学习得到一个分类模型。在协同训练时，每个模型所添加的新标注数据都是由另外两个不同的模型协作提供的。在对未标注数据进行类别判定时，这种算法与之前的算法使用了不同的方法，他们使用集成学习中的投票策略来对未标注数据进行类别判定。此后，Li 和 Zhou[30]扩展了 tri-training 算法，提出了更加高效的 Co-Forest 算法。

(4) 其他算法

自训练（Self-Training）算法是半监督学习中比较常见的一种算法。它的基本思想是，首先只使用较少的有标签的数据训练学习得到一个分类模型，然后给无标签数据进行类别判定，并把置信度高的无标签数据加入到训练集当中重新训练分类模型，重复循环训练直到满足给定的停止条件。自训练算法已经被用于自然语言处理领域。Yarowsky[31]通过自训练方法来进行语义消歧，例如：在给定的语境下预测单词“plant”是“植物”还是“工厂”。Riloff 等人[32]用其识别主题名词。Maeireizo 等人[33]将其用于解决情感语句识别问题。

直推式支持向量机（Transductive Support Vector Machines，TSVM）传统的支持向量机方法进行了扩展，它是一种直接使用有标签数据来对特定的无标签数据进行判别的技术，这种方法不仅在利用了有标签数据，而且考虑了学习模型会受无标签数据影响的情况，还结合支持向量机算法，最终实现了一种高效的半监督学习算法。在进行直推式学习时，需要利用较少的有标签数据和较多的无标签数据来训练学习。其核心在于，通过对混合数据的训练，分类模型学习到了测试集样本的数据分布信息。由于无标签数据占整个样本数据的很大比例，因此可以更加准确地描绘整个样本空间分布，从而能够学习获得更好泛化能力的的分类模型。

1.2.3 基于半监督学习的文本分类研究现状

目前，已有许多学者对半监督文本分类进行了研究。Nigam 等人[21]将 EM 算法与朴素贝叶斯分类器（Naïve Bayes Classifier）相结合，实现了一种半监督的文本分类算法。首先，利用有标签的数据来学习到一个初始的分类模型，然后用这个分类模型对无标签数据进行不确定性判定，从而使每个无标签数据都部分地属于某一类别。再然后使用所有的数据训练学习到一个新的分类器模型，重复循环上述过程一直到分类模型状态稳定时停止。张博峰等人[34]提出了一种基于自训练的改进 EM 算法。他们在使用 EM 算法训练中间分类模型时，加入了自训练机制来利用中间结果，也就是说将置信度高的无标签数据添加到有标签数据集中，不断缩小无标签数据集规模，从而加快迭代速度，最终提高分类模型的性能。但是由于训练出的中间分类器也可能存在比较大的误分类，所以该方法并不适用于已标记类别比较相似的情况。郑海清等人[35]提出了一种基于紧密度的半监督文本分类方法。这种方法首先在负例集中提取出一些置信度较高的数据，然后再根据无标签数据集合中的数据与正例以及选取出的负例之间的紧密度来对可信的负例集进行相应的扩展，从而得到用于分类训练所用的负例集。最后结合训练集中原来给出的正例集，进行文本分类。但是由于此方法只扩展了训练集中的负例集合，因此并不适用于正例集合数据本身就比较少的情况。上述这些半监督文本分类算法，本质上都是基于数据统计理论上的分布假设，即将已标记样本与未标记样本视为具有独立同分布的文本数据。然而，这个假设很难在现实世界中成立，原因是大量的无标签数据可能来自于一个不同的数据分布或者毫不相关的环境，并且这些无标签数据可能混有脏数据。当无标签数据的分布与有标签数据不同时，那么使用较多无标签数据反而可能会导致分类器的性能下降[20]。因此，如何在有标签数据的分布与无标签数据的分布不同的情况下，有效提高半监督文本分类模型的性能，已经成为半监督学习领域的一个研究热点。

1.3 本文的主要研究内容

基于半监督学习的文本分类研究，本课题主要进行了如下两个方面的研究：

第一，提出了一种基于蚁群聚集信息素的半监督文本分类算法。算法将聚集信息素与传统的文本相似度计算相融合，首先，基于聚集信息素的作用，利用 Top-k 策略选取出未标记蚂蚁可能归属的种群；然后，依判断规则，判定未标记蚂蚁的置信度；最后，采用随机选择的策略，把置信度高的未标记蚂蚁加入到对其最有吸引力的一个训练种群中。在标准数据集上地实验表明：本文提出的算法在精确率、召回率和 F1 度量三个方面表现更好。

第二，提出了一种基于特征映射的半监督文本分类算法。首先利用不同的特

征选择方法，分别在训练集的已标记数据、未标记数据以及测试集数据选取各自的特征集合，并给这些特征赋予初值；然后，建立已标记数据和未标记数据之间的映射函数，已标记数据和测试集数据之间的映射函数，未标记数据与测试集数据之间的映射函数，利用这三个特征映射函数重新计算特征的权重；最后利用 EM 算法进行半监督文本分类。大量的实验表明，我们提出的算法是有效的。

1.4 本文的组织结构

第 1 章：绪论。首先介绍了课题的研究背景和意义，然后详细阐述了国内外学者在文本分类、半监督学习以及半监督文本分类等领域的研究现状，最后阐述了本文的主要研究内容。

第 2 章：半监督文本分类的关键技术。首先对文本分类进行了概述，然后对常用的文本表示模型、文本预处理技术、文本分类算法以及分类性能评估进行了详细的介绍。

第 3 章：基于蚁群聚集信息素的半监督文本分类算法。首先介绍了基于蚁群聚集信息素的半监督分类算法；然后分析了其无法处理文本分类的原因，并提出了一种基于蚁群聚集信息素的半监督文本分类算法；最后，详细分析了实验结果。

第 4 章：基于特征映射的半监督文本分类算法。首先介绍了相关背景；然后提出了基于特征映射的半监督文本分类算法，给出了其主要思想和算法描述；最后，详细分析了实验结果。

最后，在结论部分对全文进行了总结，并对未来工作进行了展望。

第2章 半监督文本分类的关键技术

2.1 文本分类定义和过程

文本分类（Text Classification, TC）是指，在已经训练学习完成的分类系统中，通过基于文档数据的内容来计算出其所属的类别信息。文本分类其实是一个映射的过程，其把无标签的文档数据映射到有标签的文本集合当中。因为一个文档往往包含多个主题，所以这种映射可能是一对一映射，也可能是一对多映射。数学定义如下：

$$f:T\to C \tag{2-1}$$

其中，T 是无标签的文本集合，C 是有标签的文本集合。

文本分类可以分成学习和预测两个部分。如图 2-1 所示，首先利用训练集数据即已标记的样本集合（$\mathbf{t}_i$，$\mathbf{c}_i$），其中 $\mathbf{t}_i$ 表示已标记的文本，$\mathbf{c}_i$ 表示文本所归属的类别，学习得到一个分类模型。对于测试文本 t_{n+1}，利用训练学习得到的分类模型来预测它的类别信息。

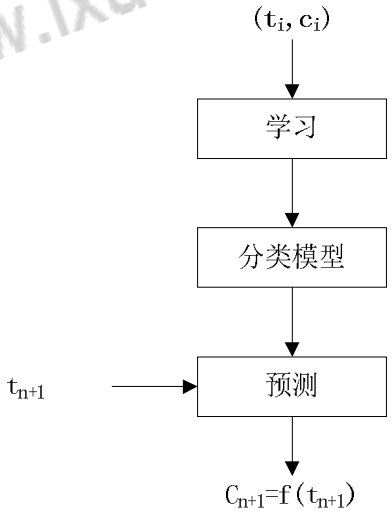

图 2-1 文本分类

Fig. 2-1 Text classification

2.2 文本表示

由于文本数据是一种计算机无法直接处理的半结构化数据，为了易于处理，需要尽可能保留语义信息的前提下将文本数据表示成计算机可以识别处理的形式。

目前，常用的文本表示模型主要有以下几种：

(1) 向量空间模型

向量空间模型（Vector Space Model，VSM）是一种最流行的文本表示方法。它被上世纪著名的 SMART[36]检索系统中第一次使用。VSM 的主要思想是：将每一个文本数据分别表示成一个特征向量，然后通过计算特征向量之间的相似度来确定文档内容是否相关。在 VSM 中，每篇文档可以表示为 D（T_1，W_1；T_2，W_2；…，T_n，W_n），简记为 D（W_1，W_2，…，W_n）。其中，D 表示文档，T_i 表示特征项，W_i 表示特征项 T_i 的权值。特征项是指文档数据中的基本语法单位，目前最有效的方法是以词作为特征项。文档 D_1 和 D_2 的相似度计算如图 2-2 所示，计算公式为：

$$Sim(D_1,D_2)=\cos(q)=\frac{\sum_{i=1}^{n} w_{1i}\times w_{2i}}{\sqrt{\left(\left(\sum_{i=1}^{n} w_{1i}^{\ 2}\right)\left(\sum_{i=1}^{n} w_{2i}^{\ 2}\right)\right)}} \tag{2-2}$$

由公式 2-2 可以看出，相似度计算实际上就是计算两个特征向量的余弦值。

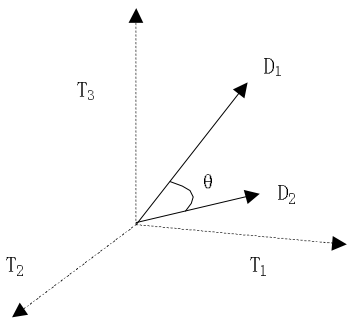

图 2-2 文档 VSM 及相似度

Fig. 2-2 VSM and document similarity

(2) 布尔模型

布尔模型（Boolean Model）是一种比较简单的文本表示模型，它主要基于集合论和布尔代数等数学理论。这种模型实际上是 VSM 的一种特殊情况。早期的搜索引擎都将布尔模型作为主要的检索模型。这种文本表示模型最大的特点是特征项的取值只有两种，True 或者 False。布尔模型适合用于特征项数比较少的情况，在常见的分类器中，决策树是基于布尔模型的。在大部分情况下，由于布尔模型不能够涵盖所有的特征项，并且布尔函数的匹配标准过于严格，这些都限制了该模型的使用。

(3) 概率模型

概率模型（Probabilistic Model）与上述的两种模型都不相同，其原理是利用数理统计中的条件概率来作为判定无标签文档类别信息依据。该模型以概率论为基础，采用相关反馈策略，在理论上能够比较好的解决文本数据分类问题。但是该模型所需的空间和时间复杂度都比较大，并且它必须借助理论假设来简化参数估计的难题。如果现实情况与理论假设不一致时，那么该模型的分类效果就会比较差。

(4) N 元模型

N 元模型（N-Gram Model）是一种统计语言模型[37]，其最大的优点是与具体的语言类型无关。该模型不考虑文档语言组成的基本单位，它将整篇文档看作是由不同语言符号组成的字符串，因此该模型可以方便地表示各种语言的文档数据。该模型基于 N-1 阶马尔可夫假设，计算公式为：

$$P(t_i \mid t_1, t_2, ..., t_{i-1}) = P(t_i \mid t_{i-N+1}, t_{i-N+2}, ..., t_{i-1}) \tag{2-3}$$

该模型假设文档中的每个特征项 t_i 和前面的 N-1 个特征项有关，而与更前面的特征项无关，因此当前特征项 t_i 的概率只取决于前面 N-1 个特征项。由于该模型需要的空间复杂度很高，并且它不能够覆盖所有的语言现象，因此在实际应用中，N 的取值都比较小。

2.3 文本预处理

因为文本数据具有非结构化的特点，所以在对文本数据建模之前，需要对每一篇文档数据进行文本预处理操作，去掉无用信息，提取出有用信息，统一文本格式，为下一步文本处理工作奠定基础。

(1) 文本格式化

文本格式化首先是过滤掉文档中的非法字符、特殊字符、标点符号、数字（或

者按照需求保留）等无用的信息，然后将文档中的有用字符统一换成小写或者大写，最后将文档变成只有合法字符和空格符组成的字符串。

(2) 分词

对于汉语等语言来说，由于文字之间没有明显的分割符，因此我们需要通过分词技术来获得每个词。目前，常用的分词算法主要有以下几种：

l 基于词库的分词算法

基于词库的分词算法是一种基于字符串匹配的分词算法，首先需要一个比较完备的词库，然后将需要分词的句子按照给定的规则与词库中的词条进行完全匹配，如果匹配成功，则提取这个词语；如果匹配失败，那么就进行其他相关处理。目前常用的匹配方式有：根据扫描方向分为正向、逆向和双向匹配；根据匹配长度分为最大和最小匹配。这种算法的缺点是不具有自适应性，不能够切分出词典中未出现的新词。

l 基于统计的分词算法

基于统计的分词算法以数理统计理论为基础，其核心思想是通过计算字与字相邻出现的概率可以比较准确的估计出一个词的可信度，也就是说越经常相邻出现的字，它们组成词的可能性就越大。因此可以统计出语料库中经常相邻出现的字的组合，通过计算它们的相邻概率，以确定它们之间的相关性。定义两个字的相关度为：

$$P = \frac{P(C_1, C_2)}{P(C_1)P(C_2)} \tag{2-4}$$

其中 $P(C_1,C_2)$是字 C_1,C_2 相邻出现的概率，$P(C_1)$、$P(C_2)$分别表示 C_1、C_2 在语料库中单独出现的概率。如果相关度 P 大于或者等于给定的阈值时，则认为相邻出现的字串是一个词。该方法首先提取出所有与词典中的词完全匹配的词，也就是找出全部可能的词条，然后使用统计语言模型和决策模型来得到最好的分词结果。但是这种方法的缺点是，通过统计筛选出来的一些相邻出现的字，它们组成的词从语言学的角度看是没有意义的，例如“子的”、“向的”等都会在文档中大量的相邻共现，但是实际上它们应该属于不同的词语；并且现在没有一种非常全面的歧义解决方法，因此它额外需要大量的已标注语料库，同时该方法的运行效率会很容易受到搜索空间的限制。目前，这种算法主要应用的模型有：互信息、N 元模型、隐马尔可夫模型以及最大熵模型等。

l 基于理解的分词算法

基于理解的分词算法旨在改进上述两种算法的不足，它由词典、知识库和推

理机构成。其中词典里存放着常用的词条；知识库中存储着已经标准化的各种语法和句法知识和语言学家总结出来的经验知识；推理机则通过使用大量的存储信息，来模拟人的推理判断过程，从而实现自动分词。该分词算法需要借助非常多的自然语言数据知识。由于中文本身的语言知识非常复杂，而且机器直接读取自然语言数据还比较困难，因此这种分词系统还不能实际应用。

(3) 词干提取

词干提取就是把文档中的同义词和近义词使用同一个词来表示的步骤。英文中的名词有单复数形式，动词则有时态、人称的变化，这些情况都会导致一个单词会出现多种不同的形式，但对于文本分类来讲，它们都属于同一个语义单位，例如名词 student 的复数形式为 students，动词 study 有 studies，studying，studied 等多种形式；而中文里同样有非常多的同义词和近义词，例如：“帮助”、“辅助”、“协助”等都具有相似的含义。通过提取词干，可以减少特征向量空间的维度，并改善文本分类器的效率。目前提取词干的方法主要有两种：基于词典的方法和基于规则的方法。根据中文及相似语言的特点，只能通过构建专业的同义词或者近义词词典，来进行词干的抽取。而对于英文等语言来说，两种方法可以单独使用，也可以混合使用。著名的 Porter Stemming[38]算法就是基于规则的方法。

(4) 去除停用词

停用词往往都是一些没有实际语义的助词，由于它们在语法规则中扮演着重要的角色，所以这种词在文本数据中占有很大比例。但是，第一，它们无法体现出文本内容所要表达的主题；第二，它们对文本分类任务几乎没有任何的贡献；第三，如果存储它们，则会浪费很大的存储空间。所以必须将它们从文本中过滤掉。通常做法是首先按照需求构建一个停用词库，然后把文本中出现在停用词库中的词条删除。这样的做法不仅简单，而且能够去除掉大部分对文本分类没有作用的词条。同时，我们也可以将一些在文本数据集中出现频率很低的稀有词条作为停用词进行处理，因为稀有词条大部分情况下对文本分类的贡献有限。而且去除停用词也是一种降低特征空间维度的方法。

(5) 特征选择与降维

经过预处理后的文本数据的特征集是一个具有高维的特征空间，这严重影响了分类器的效率，并且某些特征的存在会降低分类器的分类效果，因此我们需要对高维的特征空间进行降维，选取出对文本分类作用最大的特征，从而提高分类器的性能，并能够有效防止出现过拟合现象。现有的做法是利用特征评估函数对特征集中的所有特征项进行评估，并将评估值最高的 N 个特征项构成新的特征

集。下面介绍几种比较常用的算法：

- 文档频率

文档频率（Document Frequency，DF）表示的是文本数据集中包含某个特征项的文档数。对于一个特征项，若 DF 值小于下限阈值时，通常认为其没有代表性；若 DF 值大于上限阈值时，则认为它无法区分不同类别的文档。对于上述不在阈值范围内的特征项，都可以删掉，这样既可以将特征空间降维，又可以在一定程度上改善分类的效果。该方法是一种非常简单的特征选择方法，计算量小，选择特征速度快，但缺点是容易误删那些出现频率较低却对分类贡献较大的特征项。

- 信息增益

信息增益（Information Gain，IG）是一种非常有效的特征选择方法。它以信息论中的熵理论为基础，利用每个特征项出现前后的信息熵的变化来进行特征选择。IG 计算公式为：

$$IG(t)=P(t)\sum_{i=1}^{n}P(C_i\mid t)\log\frac{P(C_i\mid t)}{P(C_i)}+P(\bar{t})\sum_{i=1}^{n}P(C_i\mid \bar{t})\log\frac{P(C_i\mid \bar{t})}{P(C_i)} \tag{2-5}$$

其中 $P(t)$表示特征项 t 在文档集中出现的概率，$P(C_i)$表示 C_i 类文档在文档集中出现的概率，$P(C_i|t)$表示包含特征项 t 的文档属于 C_i 类文档的概率，$P(\bar{t})$ 表示特征项 t 在文档集中未出现的概率，$P(C_i\mid\bar{t})$ 表示不包含特征项 t 的文档属于 C_i 类文档的概率。

该方法的优点是综合考虑了特征项出现和不出现两种情况。但是 IG 方法只能够针对全局进行特征选择，也就是说文档集中所有类别的特征都是相同的，这就导致了 IG 无法选择出某个类别独有的一些特征。

- 互信息

互信息（Mutual Information，MI）也是一种基于信息论的方法，它可以衡量特征项与类别信息之间的相互关系。MI 计算公式为：

$$MI(t,C_i)=\log\frac{P(t,C_i)}{P(t)P(C_i)} \tag{2-6}$$

其中 $P(t,C_i)$表示既包含特征项 t 又属于类别 C_i 的文档出现的概率。由式(2-6)可以看出，若特征项 t 依赖于类别 C_i 时，互信息值则比较大；若特征项 t 与类别 C_i 相互独立时，互信息值则为零。由于低频特征词通常具有较高的 MI 值，因此该方法更倾向于选择出现次数较少的特征词。

- 开方检验

开方检验（Chi-square，CHI）是一种基于数理统计理论且效果比较好的特征选择方法。CHI 计算公式为：

$$CHI(t,C_i)=\frac{P(t,C_i)P(\overline{t},\overline{C}_i)-P(\overline{t},C_i)P(t,\overline{C}_i)}{P(t)P(C_i)P(\overline{t})P(\overline{C}_i)} \tag{2-7}$$

由式(2-7)可以看出，CHI 方法同时考虑了特征项出现和未出现的情况，CHI 值越大，说明特征项 t 与类别 C_i 越相关，则该特征项对分类贡献也就越大。

上述的几种方法基于不同的理论，通过大量的实验发现，并不存在一种可以在任何数据集上表现都很好的通用算法。因此，特征选择需要综合考虑文本数据集的特点以及特征选择方法的优缺点，然后做出最优的选择。

(6) 权重计算

在用向量空间模型表示文本数据时，在特征选择过程完成以后，还需要计算选取的每个特征词的权值。由于不同的特征词对文本分类影响各不相同，特征词的权值应该准确反映出其对分类的贡献度。对于那些贡献较大的特征词，应该赋予较大的权值；而对于那些对分类贡献很少的特征词的，应该赋予较小的权值。目前，最常用的权重计算方法是 TFIDF 算法。

TFIDF 算法综合考虑了特征项的频率以及逆文档频率，该算法计算得出的权重值，既可以充分地表示文本内容，又可以很好地区分不同的类别。计算公式为：

$$W_{ik}=tf_{ik}\times idf_k \tag{2-8}$$

其中，tf_{ik} 表示特征项 t_k 在文本 D_i 中出现的次数，idf_k 表示特征项 t_k 的逆文档频率。

2.4 常用的文本分类算法

预处理阶段主要是将非结构化的文本数据，转换成结构化的数据。首先，通过选择出的特征集合来表示文本数据的内容，并把所有需要处理的文本数据都用已选取出的特征集表示。然后，利用训练集中的文本数据，训练学习得到一个分类器。最后，对测试集中的未标记的文本数据进行分类。目前，常用的文本分类算法主要有以下几种：

(1) 朴素贝叶斯分类算法

朴素贝叶斯分类器（Naïve Bayes Classifier）是一种非常简单的概率关系模型，它基于数理统计中的贝叶斯定理，并假设文本数据中的所有特征项都是相互独立的。分类器的训练过程实际上是通过训练集中的已标注的文本数据来估计概

率关系模型中参数的过程。

对于测试集中新的文档 d，Naïve Bayes 分类器需要分别计算文档 d 属于每个类别的后验概率，最后将文档 d 分到具有最大后验概率值的类别中。该模型的计算公式为：

$$P(C_i \mid d) = \frac{P(C_i) \times P(d \mid C_i)}{P(d)} = \frac{\prod_{k=1}^{n} P(t_k \mid C_i) \times P(C_i)}{P(d)} \tag{2-9}$$

其中，$P(d|C_i)$表示在给定的文本类别 C_i 条件下，文档 d 出现的概率。$P(t_k|C_i)$ 表示在给定的文本类别 C_i 条件下，特征项 t_k 属于类别 C_i 的概率。

(2) 支持向量机

支持向量机（Support Vector Machine，SVM）是一种非常有效的二元分类器，它具有非常坚实的统计学理论基础。其主要思想是，通过使用结构风险最小化（Structural Risk Minimization，SRM）原理，寻找一个最大边缘超平面，使得训练集中不同类别的数据恰好分布在决策边界的两侧。对于线性 SVM 来说，它的决策边界函数为：

$$w \cdot x + b = 0 \tag{2-10}$$

其中，w 和 b 是决策模型的参数。SVM 通过训练集样本数据的训练学习，通过估计得到决策模型中的参数值，使其能够找到一个最优的决策边界模型。

(3) K 近邻算法

K 近邻（K Nearest Neighbor，KNN）分类器是一种消极的（或懒惰的）学习算法，它不需要事先通过训练集数据训练学习得到一个分类模型，而是直接利用训练集文本数据本身对测试集文档进行分类。该算法的主要思想是，对于每一个测试集文档，通过计算其与训练集中所有文档的距离或者相似度，选取出与其距离最近或者最相似的 K 个文档，通过这 K 个训练集文档的类别标记信息，来确定测试集文档的类别。具体的算法如表 2-1 所示。

(4) 决策树算法

决策树分类器是一种基于规则的分类算法，它的模型非常简单。其核心思想是根据训练集数据，构建一棵决策树，学习得到一些规则，然后根据学习到的规则，对测试集数据进行分类。决策树算法的关键内容是如何选择具有最佳分类效果的数据特征。与上述的几种基于统计理论的分类算法相比，基于规则的决策树分类器在某些领域的分类效果更加准确。从应用的角度看，决策树算法是目前应用最为广泛的一种分类器。

表 2-1 K 近邻算法

Table 2-1 K nearest neighbor algorithm

K 近邻算法
1： 令 K 为最近邻的数值，D 为训练集文档集合
2： For 每个测试文档 $z=(t',c')$ do
3： 计算 z 和每个训练集样本 $(t,c)\in D$ 的相似度 $Sim(t',t)$（或者距离）
4： 选择与 z 最相似的 K 个训练集样本的集合 $D_z\subseteq D$
5： $c'=\arg\max_v \sum_{(t_i,c_i)\in D_z} I(v=c_i)$ //v 是类别标记，c_i 是最近邻的类别标记，I()为判别函数，若参数为 Ture，返回 1，否则，返回 0
6： End for

2.5 分类性能评估

由于文本分类在机器学习和数据挖掘等领域中，发挥着越来越重要的作用，对其进行研究的学者也日益增多。因此，大家不得不面对如何评价各种分类器分类效果的问题。影响分类器性能的原因有很多种，例如与所使用的分类算法、文档数据的类型、文档表示模型、特征选取方法等有关。目前，大家一致的做法是通过实际的实验，利用适当的评价方法和标准，来比较分类器的性能高低。

2.5.1 评价方法

我们通常将实验数据分为训练集数据和测试集数据，首先利用训练集数据学习得到分类模型，然后利用测试集数据对分类器进行性能评测。由于训练集数据和测试集数据有可能包含离群点、噪声等脏数据，这些都有可能影响分类器的性能。因此，将实验数据分割一次是远远不够的。目前，常用的数据分割技术有：

(1) K 折交叉验证（K-fold Cross-validation）

该方法将实验数据集 S 随机的划分成 K 个大小相同的数据子集 S_i。每次实验时挑选其中一个子集作为测试集数据，其余的 K-1 个子集作为训练集数据，这样将数据切分成训练集和测试集。重复上述做法 K 次，就可以获得 K 个不同的训练集和测试集数据。通过 K 次实验，然后取实验结果的平均值来评估分类算法的性能。

(2) 留一交叉验证（Leave-one-out）

该方法是一种比较常见的交叉验证方法，其主要思想是：假设数据集的大小

为 N，每次实验选取其中一个数据作为测试数据，其余的 N-1 个数据作为训练集数据。重复实验 N 次，最后将实验结果的均值作为分类器的评价标准。

(3) 5×2 折交叉验证

该方法的主要思想是，将实验数据集 S 随机的划分为两个大小相同的子集，S_1 和 S_2。首先将 S_1 作为训练集数据，S_2 则作为测试集数据，进行一次实验。然后，将 S_2 作为训练集数据，S_1 作为测试集数据，再进行一次实验。这样的做法称为一次对折。重复上述的做法，进行 5 次对折实验。

2.5.2 评价标准

目前常用的评价标准主要有以下几种：

(1) 精确率和召回率

精确率（precision，p）和召回率（recall，r）是目前文本分类中最常使用的评价标准。精确率（或准确率、查准率）是指被正确划分到类别 C_i 的文档数与被划分到 C_i 类的所有的文档数的比值。而召回率（或查全率）是指被正确判定归属类别 C_i 的文档数目与实际上归属 C_i 类的文档数之比。具体的计算公式为：

$$p = \frac{x}{x+y} \tag{2-11}$$

$$r = \frac{x}{x+z} \tag{2-12}$$

其中，x，y，z 如表 2-2 所示。x 是被分类器正确判断归属 C_i 类的文档数；y 是被分类器错误判断归属 C_i 类的文档数；z 是被分类器错误判断属于其他类的 C_i 类文档数。

表 2-2 类别 C_i 列联表

Table 2-2 Contingency table of category C_i

	实际属于 C_i 的文档数	实际不属于 C_i 的文档数
分类器判定属于 C_i 的文档数	x	y
分类器判定不属于 C_i 的文档数	z	w

(2) F 度量

F 度量（F-measure）是一种综合考虑精确率和召回率的评价标准。因为精确率和召回率是相互影响的，如果仅仅追求提高其中一项性能则可能会降低另外一项性能。具体的计算公式为：

$$F=\frac{(1+a^2)pr}{a^2p+r} \tag{2-13}$$

其中，a 是一个可变的参数，用来改变精确率和召回率之间的相对重要性。由式(2-13)可以得知，当a 取值为零时，F 度量就变成了精确率；当a 趋向于无穷大时，F 度量就变成了召回率。目前，a 通常取值为 1，即认为精确率和召回率重要性是相同的，也就是比较常见的 F_1 度量。计算公式为：

$$F_1=\frac{2pr}{p+r} \tag{2-14}$$

(3) 宏平均和微平均

宏平均（macro-average）和微平均（micro-average）主要对分类器在数据集中所有类别上的性能进行评测。宏平均是直接利用各个类别的评价标准计算其平均值，微平均却是将各个类别的列联表求和如表 2-3 所示，然后再计算其平均值。具体的计算公式为：

表 2-3 所有类的列联表

Table 2-3 Contingency table for all classes

	实际属于类 C	实际不属于类 C
分类器判定属于类 C	$x=\sum_{i=1}^{\|C\|}x_i$	$y=\sum_{i=1}^{\|C\|}y_i$
分类判定不属于类 C	$z=\sum_{i=1}^{\|C\|}z_i$	$w=\sum_{i=1}^{\|C\|}w_i$

$$macro\_p=\frac{\sum_{i=1}^{|C|}p_i}{|C|} \tag{2-15}$$

$$macro\_r=\frac{\sum_{i=1}^{|C|}r_i}{|C|} \tag{2-16}$$

$$macro\_F_1=\frac{2\times macro\_p\times macro\_r}{macro\_p+macro\_r} \tag{2-17}$$

$$micro\_p=\frac{\sum_{i=1}^{|C|}x_i}{\sum_{i=1}^{|C|}(x_i+y_i)} \tag{2-18}$$

$$micro\_r = \frac{\sum_{i=1}^{|C|} x_i}{\sum_{i=1}^{|C|} (x_i + z_i)} \tag{2-19}$$

$$micro\_F_1 = \frac{2 \times micro\_p \times micro\_r}{micro\_p + micro\_r} \tag{2-20}$$

2.6 本章小结

本章主要是对半监督文本分类所需的关键技术进行了介绍，首先对文本分类进行了概述，然后对常用的文本表示模型、文本预处理技术、文本分类算法以及分类性能评估进行了详细的介绍。

第3章 基于蚁群聚集信息素的半监督文本分类算法

本文借鉴了基于蚁群聚集信息素浓度的半监督分类算法(aggregation pheromone density based semi-supervised classification, APSSC)[39]的基本思想，提出了一种基于蚁群聚集信息素浓度的半监督文本分类算法。针对文本数据，算法将文本相似度融合到聚集信息素的计算之中，从而扩展了聚集信息素浓度的计算模型。在半监督学习过程中，首先，基于聚集信息素的作用，利用 Top-k 策略选取出未标记蚂蚁可能归属的种群；然后，依判断规则，判定未标记蚂蚁的置信度；最后，采用随机选择的策略，把置信度高的未标记蚂蚁加入到对其最有吸引力的一个训练种群中，来对训练种群进行扩展。大量的实验结果表明：与常用的朴素贝叶斯分类和 EM 分类相比，本文提出的算法，在精确率、召回率以及 F_1 度量方面，都有一定的提高。

3.1 基于蚁群聚集信息素的半监督分类算法

2010 年，Halder 等人[39]将蚁群聚集信息素引入到半监督分类中，提出了 APSSC 算法，它是一种自训练的算法，并在非文本数据集上取得了比较好的效果。聚集信息素(Aggregation Pheromone)[40]是蚂蚁等昆虫分泌的一种化学物，可用于招引同种个体一起栖息，共同取食，攻击异种对象，并最终形成群集智能的整体行为。

APSSC 算法把每一个样本都看作一只蚂蚁，每个类别当作蚂蚁可能归属的种群，其中已标记蚂蚁 $a_i^{l_k}$ 表示类别 k 中的已标记样本 $x_i^{l_k} \in C_k^0$，未标记蚂蚁 a_i^u 表示未标记样本 $x_i^u \in U$ 。初始时，训练种群中只包含已标记蚂蚁，各训练种群对未标记蚂蚁释放的聚集信息素浓度均为零。在每一代的学习过程中，需要对每一只未标记蚂蚁进行如下的操作：

(1) 计算训练种群 k 向未标记蚂蚁 a_j^u 释放的聚集信息素浓度：

$$\Delta \bar{t}_{jk}^{t} = \frac{1}{\left|C_k^t\right|} \sum_{x_i^{l_k} \in C_k^t} \Delta t^{t}\left(x_i^{l_k}, x_j^u\right);\ \forall j, \forall k \tag{3-1}$$

其中 $|C_k^t|$ 为训练种群 k 包含的蚂蚁数量，$\Delta t^t(x_i^{l_k}, x_j^u)$ 为已标记蚂蚁 $a_i^{l_k}$ 向未标记蚂蚁 a_j^u 释放的信息素浓度，计算模型为：

$$\Delta t^{t}\left(x_i^{l_k}, x_j^{u}\right)=\frac{1}{(2p)^{d/2}\left(\det\left(\sum_k^t\right)\right)^{\frac{1}{2}}}\exp\left(-\frac{1}{2}\left(x_j^{u}-x_i^{l_k}\right)^{T}\left(\sum_k^t\right)^{-1}\left(x_j^{u}-x_i^{l_k}\right)\right) \tag{3-2}$$

其中，$\sum_k^t$ 是训练种群 k 的协方差矩阵，$\det(\sum_k^t)$ 是协方差矩阵的行列式，d 是数据集的维度。

(2) 更新训练种群 k 向 a_j^u 释放的聚集信息素浓度：

$$t_{jk}^{t}=(1-r)t_{jk}^{t-1}+r\Delta\bar{t}_{jk}^{t} \tag{3-3}$$

其中，$r\in[0,1]$ 是蒸发常量。

(3) 当所有的训练种群都已经向未标记蚂蚁 a_j^u 释放完聚集信息素时，需要对更新后的 t_{jk}^{t} 进行归一化处理，得到归一化的聚集信息素浓度：

$$m_{jk}^{t}=\frac{t_{jk}^{t}}{\sum_{k=1}^{K}t_{jk}^{t}};\forall j,\forall k \tag{3-4}$$

其中 K 为训练种群总数。

(4) 根据算法 1 选取置信度高的未标记蚂蚁加入到对其最有吸引力的一个训练种群 h 中，它将在 t+1 代的自训练过程中作为种群 h 的已标记蚂蚁。

(5) 当训练种群全部稳定时，自训练过程结束，此时训练种群已经得到了扩展；否则，进入 t+1 代学习。

从上可见，APSSC 算法的自训练过程是一个利用大量置信度高的未标记蚂蚁对训练种群进行扩展的过程。因此，如何选取置信度高的未标记蚂蚁即算法 1，是 APSSC 算法的核心。自训练过程完成以后，进入测试过程。

在测试过程中，根据种群 k 对测试蚂蚁 a_n 释放的聚集信息素浓度 $\Delta\bar{t}_{nk}$（见式(3-5)），把测试蚂蚁 a_n 分到对它释放了最多聚集信息素的种群 h 中，即测试蚂蚁 a_n 归属于种群 h。

$$\Delta\bar{t}_{nk}=\frac{1}{|C_k^t|}\sum_{x_i\in C_k^t}\frac{1}{(2p)^{d/2}(\det(\sum_k^t))^{\frac{1}{2}}}\exp(-\frac{1}{2}(x_n-x_i)^{T}(\sum_k^t)^{-1}(x_n-x_i)) \tag{3-5}$$

表 3-1 算法 1

Table 3-1 Algorithm 1

算法 1 选取置信度搞的未标记蚂蚁

输入：未标记蚂蚁 a_j^u

输出：扩展后的训练种群 h

begin

for m_{jk}^t (k≠h)

if ($\frac{m_{jk}^t}{m_{jh}^t = \max(m_{jk}^t)} \le \frac{1}{K}$) then

flag_variable = 1

else

flag_variable = 0

break

end if

end for

if (flag_variable == 1) then

把未标记蚂蚁 a_j^u 加入到训练种群 h 中

else

未标记蚂蚁 a_j^u 不加入到任何一个种群中

end if

end

3.2 基于蚁群聚集信息素的半监督文本分类算法

3.2.1 存在的问题和缺陷

APSSC 算法在 Ionosphere，Balance Scale，Sonar 等来自 UCI 的数据集上取得了比较好的效果，证明了蚁群聚集信息素在半监督学习中的有效性，但由于该算法存在着如下不足，使其无法完成文本分类。

(1) 由公式(3-2)可以看出，其聚集信息素计算公式中数据集的维度 d 作为 $2p$ 的指数出现在分母中，由于 $2p > e$，当 d 取值比较大时，公式(3-2)趋向于零，因此该算法不能处理像文本这样的高维度数据。

(2) 公式(3-2)中 $\det(\sum_k^t)$ 作为一个因子出现在分母当中，这就要求数据集的协方差矩阵 $\sum_k^t$ 必须是可逆的，因此该算法很难适用于具有稀疏性的文本数据。

(3) 在算法 1 中，选取置信度高的蚂蚁的判定条件是必须满足条件：

$$\frac{m_{jk}^t}{m_{jh}^t = \max(m_{jk}^t)} \le \frac{1}{K} \tag{3-6}$$

也就是说，当遇到类别比较多且相似性较大的文本数据集时，该算法可能会学习不到置信度高的蚂蚁，从而无法达到扩展训练种群的目的。

3.2.2 算法思想

本文对 APSSC 算法进行了扩展，提出了基于蚁群聚集信息素浓度的半监督文本分类算法（a semi-supervised text classification based on ant colony aggregation pheromone, SSTCACAP），使其能有效处理文本数据。

首先，根据文本数据高维稀疏的特点，结合文本相似度与聚集信息素之间的关系，将文本相似度作为蚁群聚集信息素浓度计算公式的一个影响因子，扩展了蚁群聚集信息素浓度的计算模型：

$$\Delta t^{t}(x_i^{l_k}, x_j^{u}) = \exp^{-\frac{[1-Similarity(x_i^{l_k}, x_j^{u})]^2}{2d^2}} \tag{3-7}$$

其中，$Similarity(x_i^{l_k}, x_j^{u})$ 是余弦相似性公式，d 是高斯函数的扩散速度系数。

相应地扩展了测试过程中种群 k 对测试蚂蚁释放的聚集信息素浓度计算模型：

$$\Delta \bar{t}_{nk} = \frac{1}{|C_k^t|}\sum_{x_i \in C_k^t} \exp^{-\frac{[1-Similarity(x_i, x_u)]^2}{2d^2}} \tag{3-8}$$

其次，根据文本数据集类别多且相似性高的特点，利用 Top-k 策略和随机选择策略扩展了 APSSC 算法中的选取置信度高的未标记蚂蚁算法即算法 1，得到扩展后的算法 2，具体的扩展过程如下：

在半监督学习过程中，因为文本数据本身可能携带多个类别特征，并且初始的训练种群包含的种群特征信息非常有限，所以各个训练种群对其释放的聚集信息素的差别很小，时甚至可能出现几个种群对其释放的聚集信息素浓度相同的情况。因此，在判定未标记蚂蚁置信度的过程中，我们采用在 Web 搜索引擎中被广泛使用的 *Top-k* 查询策略，选取对其最有吸引力的 k 个种群，组成一个候选种群集合 T_{top-k}。根据集合 T_{top-k} 中包含的种群数量来判定未标记蚂蚁的置信度，判定规则为：

$$\frac{|T_{top-k}|}{K} \le b \tag{3-9}$$

其中 $|T_{top-k}|$ 是 T_{top-k} 包含的种群数，K 是训练种群总数，b 是设定的阈值。

当未标记蚂蚁满足判定公式(3-9)时，我们就判定它的置信度高。这种 *Top-k* 策略，可以有效处理因训练种群对未标记蚂蚁释放的聚集信息素差别较小而可能学习不到置信度高的未标记蚂蚁的情况。

一旦未标记蚂蚁 a_j^u 的置信度被判定为高时，就需要将其加入到 T_{top-k} 中的一个种群中，来达到扩展训练种群的目的。由于 T_{top-k} 中的训练种群对未标记蚂蚁的吸引力相差不大，也就是说置信度高的未标记蚂蚁 a_j^u 归属于 T_{top-k} 中任何一个训练种群的概率是相等的，因此在 T_{top-k} 中随机选择一个种群 k，把 a_j^u 加入到种群 k 中。那么 a_j^u 就可以在下一代的学习过程中，作为种群 k 的已标记蚂蚁，从而扩展了训练种群 k。

3.2.3 算法描述

在 SSTCACAP 算法的每一代学习中，需要判定每一只未标记蚂蚁的置信度，并利用置信度高的未标记蚂蚁扩展训练种群。算法描述如下：

表 3-2 算法 2

Table 3-2 Algorithm 2

算法 2 选取置信度高的未标记蚂蚁

输入：未标记蚂蚁 a_j^u

输出：扩展后的训练种群 h

```
begin
    T_top-k = {h}
    for  m_jk^t (k≠h)
        if ( m_jk^t / m_jh^t ≥ α )  then
            将种群 k 加入到候选种群集合 T_top-k 中
        end if
    end for
    if ( |T_top-k| / K ≤ β )  then
        在候选种群集合 T 中随机选取一个种群 k   /*当 T 中只有一个训练种群时，T=k*/
    else
        未标记蚂蚁不加入到任何一个种群中
    end if
end
```

每一代学习完成以后，便得到了扩展后的训练种群。

3.3 实验结果与分析

本章所有实验都是在配置为，处理器：英特尔 酷睿 2 四核 Q9450 2.66GHz；内存：4 GB DDR2 667MHz ；操作系统：Windows7 32 位的 PC 机上进行的。

3.3.1 实验数据集与评价标准

实验中我们采用了一个英文数据集和一个中文数据集来进行测试。我们采用的英文数据集是著名的语料库 20 Newsgroups（20news-18828），为了减少实验等待时间，从中挑选了 5 个 comp.*类别数据来进行实验，其中 comp.graphics 有 973 篇文档，comp.os.ms-windows.misc 有 985 篇文档，comp.sys.ibm.pc.hardware 有 982 篇文档，comp.sys.mac.hardware 有 961 篇文档，comp.windows.x 有 980 篇文档。我们采用的中文数据集是复旦大学提供的中文语料库，同样为了尽快得到实验结果，从中挑选了 C19-Computer、C31-Environment、C32-Agriculture、C38-Politics 和 C39-Sports 五个类来进行实验，其中有 1076 篇 C19-Computer 类文档，1087 篇 C31-Environment 类文档，1021 篇 C32-Agriculture 类文档，1024 篇 C38-Politics 类文档，1077 篇 C39-Sports 类文档。实验时，我们固定的在每一类文档中抽取 20%作为测试集数据，抽取 50%作为未标记数据集，其余的作为已标记数据集，标记比例为 5%-25%。

我们实验中对算法的评价指标采用了精确率（precision）、召回率（recall）和宏平均 F_1 度量（Macro F_1）三个比较通用的评价标准，具体的计算方法详见 2.5.2 节。

3.3.2 预处理及参数设置

实验时，本文利用向量空间模型（VSM）来表示文档数据，本文对实验数据的预处理操作具体包括：文本格式化来清理文档数据中的无用信息和噪声等；对英文数据集使用 Porter Stemming 算法进行了词干提取、然后去除停用词；对中文数据集利用“结巴”分词进行了分词处理；使用了信息增益算法对已标记数据进行特征选择；利用 TFIDF 算法进行权重计算。

参数设置为：实验中 SSTCACAP 算法使用的参数配置 δ=1.0，α=0.9，β=0.6。EM 算法采用文献[21]的终止条件。

3.3.3 实验结果比较与分析

本文在 20news 英文数据集和复旦大学中文数据集上进行了多组实验，并比较了 SSTCACAP、EM 以及 naïve Bayes 算法的分类效果。具体地，我们比较了各种算法在精确率、召回率和宏平均 F_1 值三个方面的表现，并通过不断改变已标记文本集的大小，得出算法的精确率、召回率以及宏平均 F_1 度量与已标记样本集大小之间的关系图。

(1) 20news 英文数据集

首先，分析三种算法在 20news 英文数据集上的分类效果。

由图 3-1 可以看出，在 20news 数据集上，两种半监督学习方法 SSTCACAP 算法和 EM 算法在引入未标记样本后，都在一定程度上提高了分类的精确性，尤其在已标记样本较少的情况下，半监督学习方法的分类性能表现更好，且本文提出的 SSTCACAP 算法的精确率比传统的有监督的方法 naïve Bayes 算法提高了 5%。SSTCACAP 算法在大部分情况下，分类精确率都高于 EM 算法和 naïve Bayes 算法，只有在 20%已标记样本时，精确率低于 EM 算法。例外产生的一个可能原因是由于某个训练种群的吸引力远大于其他的训练种群，使得这个训练种群学习到了大部分的置信度高的未标记蚂蚁，从而出现了相对较多的误分情况。随着已标记样本数量的增加，出现了 naïve Bayes 算法的分类精确率高于 EM 算法的情况。其中一个可能的原因是此时已标记样本的数据分布与未标记样本的数据分布存在较大的差异，导致 EM 算法学习到的分类器的性能下降。而本文提出的 SSTCACAP 算法没有出现分类性能低于 naïve Bayes 的情况，表明 SSTCACAP 算法的稳定性要优于 EM 算法。

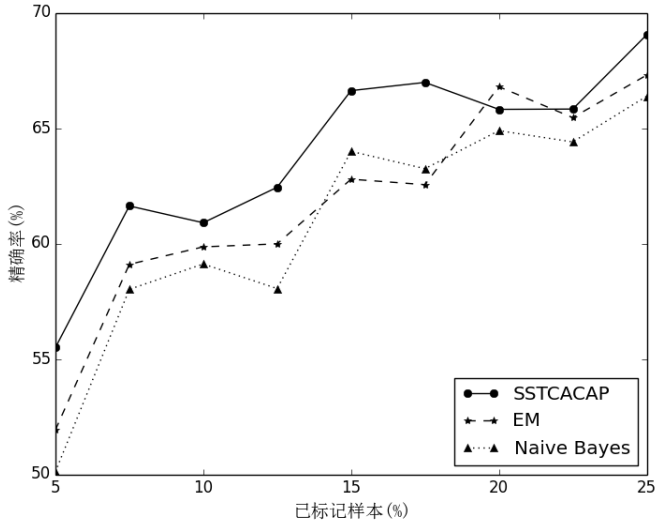

图 3-1 在 20news 数据集上的精确率(precision)比较

Fig. 3-1 Comparison of precision on 20news

表 3-3　不同算法在 20news 数据集上分类结果评价值对比

Table 3-3　The comparison of different algorithms' results on 20news

评价标准	标注比例	算法		
		SSTCACAP	EM	naïve Bayes
精确率	5%	**0.5549**	0.5192	0.5011
	7.5%	**0.6163**	0.591	0.5801
	10%	**0.609**	0.5985	0.5911
	12.5%	**0.6244**	0.5998	0.5805
	15%	**0.6662**	0.6279	0.6399
	17.5%	**0.6698**	0.6255	0.6324
	20%	0.6581	**0.6679**	0.6489
	22.5%	**0.6583**	0.6546	0.644
	25%	**0.6904**	0.673	0.6638
召回率	5%	**0.5545**	0.4796	0.5039
	7.5%	**0.6143**	0.5827	0.575
	10%	**0.6023**	0.5815	0.5857
	12.5%	**0.6171**	0.5909	0.5788
	15%	**0.6618**	0.6135	0.6231
	17.5%	**0.6686**	0.6172	0.6212
	20%	**0.6548**	0.6511	0.6395
	22.5%	**0.6582**	0.6446	0.6365
	25%	**0.6888**	0.6587	0.6574
宏平均 F1 值	5%	**0.5545**	0.4481	0.4739
	7.5%	**0.612**	0.5749	0.5622
	10%	**0.6015**	0.5779	0.576
	12.5%	**0.6179**	0.5821	0.5674
	15%	**0.6579**	0.6143	0.6198
	17.5%	**0.6653**	0.6179	0.6175
	20%	0.6523	**0.6529**	0.6367
	22.5%	**0.6566**	0.6455	0.634
	25%	**0.6872**	0.6579	0.6543

由图 3-2 可以看出，在 20news 数据集上，本文提出的 SSTCACAP 算法在召回率方面表现是最好的，大部分情况下召回率高于 60%，要优于 EM 算法和 naïve Bayes 算法。且在 5%已标记样本时，本文提出的 SSTCACAP 算法分类性能比 EM 算法和 naïve Bayes 算法，分别提高了 7.5%和 5.1%。而另一种半监督学习方法 EM 算法和传统的有监督的学习方法 naïve Bayes 算法在已标记样本低于 20%时，分类性能出现了交替上升的情况。我们认为出现这种情况的原因是基于样本数据分布理论假设的半监督学习方法，在无法保证已标记样本与未标记样本的分

布一致的情况下，导致半监督学习方法学到的分类器的性能下降。

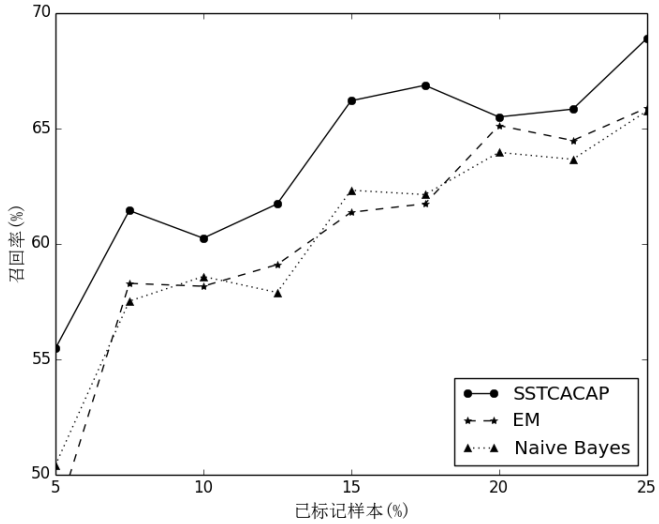

图 3-2　在 20news 数据集上的召回率比较

Fig. 3-2　Comparison of recall on 20news

由图 3-3 可以看出，在 20news 数据集上，本文提出的 SSTCACAP 算法在宏平均 F_1 度量（Macro F_1）方面，分类性能也是最好的。只有在 5%已标记样本的情况下，SSTCACAP 算法的宏平均 F1 度量低于 60%，但是 SSTCACAP 算法比 EM 算法和 naïve Bayes 算法，在性能上分别提高了 10.7%和 8.1%。且只有在 20%已标记数据时，另一种半监督学习方法 EM 算法的分类性能达到了本文提出的 SSTCACAP 算法的水平。而 EM 算法在已标记样本低于 20%时，分类性能与 naïve Bayes 算法相当，并没有明显的提升。

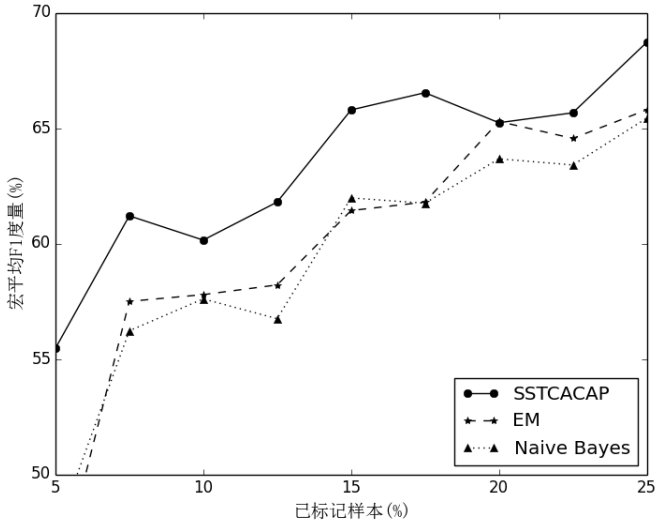

图 3-3　在 20news 数据集上的宏平均 F_1 度量(Macro F_1)比较

Fig. 3-3　Comparison of Macro F_1 on 20news

表 3-4　不同算法在复旦大学中文数据集上的分类结果评价值对比

Table 3-4　The comparison of different algorithms' results on Fudan dataset

评价标准	标注比例	算法		
		SSTCACAP	EM	naïve Bayes
精确率	5%	**0.5939**	0.5093	0.3434
	7.5%	**0.586**	0.5741	0.3593
	10%	**0.7778**	0.6378	0.4024
	12.5%	**0.7749**	0.5946	0.384
	15%	**0.7481**	0.5668	0.3726
	17.5%	**0.7657**	0.6122	0.3826
	20%	**0.7648**	0.589	0.4045
	22.5%	**0.7576**	0.6464	0.4
	25%	**0.7501**	0.6172	0.4069
召回率	5%	**0.5002**	0.4849	0.3319
	7.5%	0.5364	**0.5492**	0.3595
	10%	**0.7417**	0.6286	0.3963
	12.5%	**0.7067**	0.5812	0.3787
	15%	**0.6929**	0.5643	0.3705
	17.5%	**0.7191**	0.6071	0.3746
	20%	**0.7296**	0.5866	0.4005
	22.5%	**0.7288**	0.6245	0.397
	25%	**0.7201**	0.604	0.4041
宏平均 F1 值	5%	0.4597	**0.484**	0.3354
	7.5%	0.4794	**0.5526**	0.3574
	10%	**0.7129**	0.6302	0.3979
	12.5%	**0.6893**	0.5848	0.3786
	15%	**0.6736**	0.5645	0.3683
	17.5%	**0.6896**	0.6082	0.3749
	20%	**0.69**	0.5874	0.4002
	22.5%	**0.6902**	0.6269	0.3956
	25%	**0.6785**	0.605	0.404

(2)　复旦大学中文数据集

然后，分析三种算法在复旦大学中文数据集上的分类性能。

由图 3-4 可以看出，在复旦大学中文数据集上，本文提出的 SSTCACAP 算法在精确率方面表现是最好的，优于另外一种半监督文本分类算法 EM 算法以及朴素贝叶斯算法。SSTCACAP 算法比 EM 算法在精确率上高出 1%到 18%，并且在标记比例为 10%时，SSTCACAP 算法的精确率达到了最高值 77.8%，而 EM

算法的精确率是 63.8%，朴素贝叶斯算法精确率是 40%。随着已标记文档数据的增多，三种算法在精确率上都有一定程度的提高。但是从图中可以得知，SSTCACAP 算法在已标记文档低于 10%时，与另外一种半监督文本分类算法的对比不太明显，而当已标记文档高于 10%时，SSTCACAP 算法对另外两种算法的优势更加明显。出现这一现象的原因是，在标记比例比较少时，各训练种群对未标记蚂蚁的吸引力相差不大，我们采用的随机选择策略导致了相对较多的误分情况。这说明本文提出的 SSTCACAP 算法会受到标记比例因素的影响。同时，我们可以从图中看出，朴素贝叶斯算法在精确率方面的表现与其他两种半监督文本分类算法相差比较大，这充分说明了半监督文本分类算法是可行而且有效的。

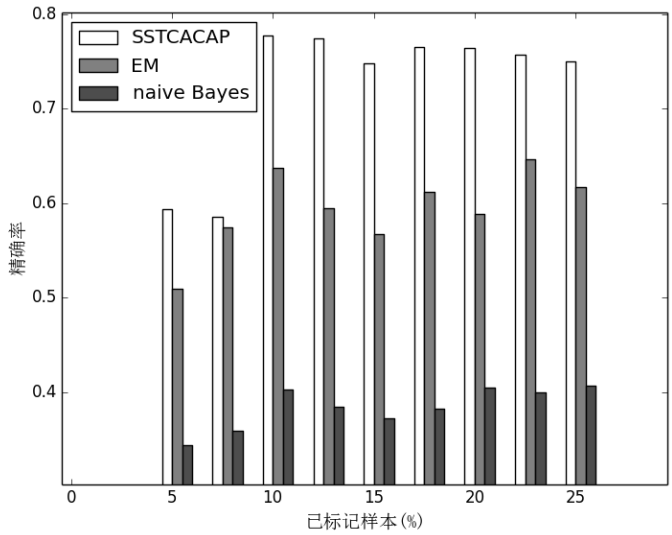

图 3-4　算法在复旦大学中文数据集上的精确率比较

Fig. 3-4　Comparison of precision on Fudan dataset

从图 3-5 中可以看出，两种半监督文本分类算法 SSTCACAP 算法和 EM 算法在召回率上的分类性能要好于另外一种普通的朴素贝叶斯分类算法，这也再一次表明半监督文本分类算法是非常有效的。而且，SSTCACAP 算法在已标记比例小于 10%时，召回率超过了 50%，当标记比例大于 10%时，召回率大部分都超过了 70%，与另外两种算法相比，有着比较大的优势 。同时，我们可以看到 SSTCACAP 算法只有在已标记比例为 7.5%时，召回率略低于另外一种半监督文本分类算法 EM 算法，相差只有 1.2%，其余情况下的召回率都要高于 EM 算法。我们认为出现这种例外的原因是由于会发生某一个训练种群对未标记蚂蚁的吸引力大于其他训练种群的情况，从而使得该训练种群吸引了较多的未标记蚂蚁来扩展该训练种群，导致发生了较多的误分情形。

并且图 3-6 也再一次表明半监督文本分类算法的分类性能要优于普通的文本

分类算法。同时，不难发现本文提出的 SSTCACAP 算法在宏平均 F1 值方面大部分情况下是表现最好的，并在标记比例 10%时达到了最高的 71.3%，比 EM 算法高 8.3%，比朴素贝叶斯算法高 31.5%。但是，我们也可以发现，在已标记比例低于 10%时，EM 算法的性能比 SSTCACAP 算法更好一些。而朴素贝叶斯算法在大部分情况下，宏平均 F1 值低于 40%。

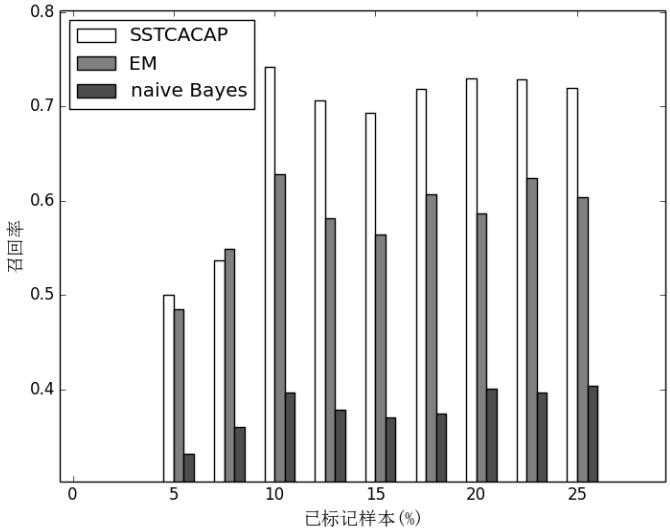

图 3-5　算法在复旦大学中文数据集上的召回率对比

Fig. 3-5　Comparison of recall on Fudan dataset

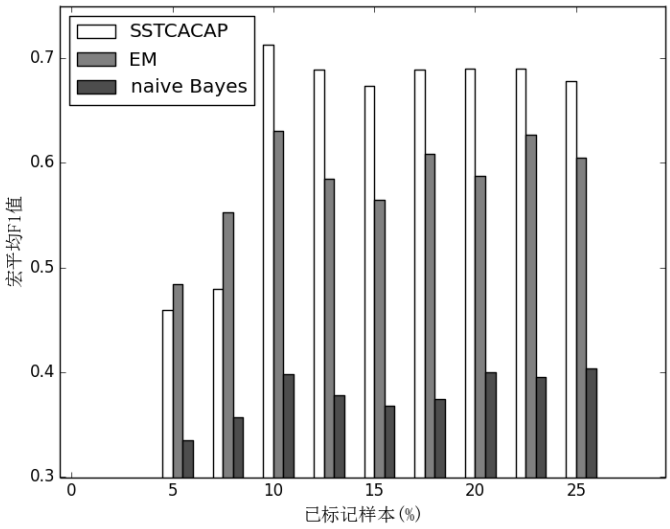

图 3-6　算法在复旦大学中文数据集上的宏平均 F_1 值对比

Fig. 3-6　Comparison of Macro F_1 on Fudan dataset

综合图 3-1 到图 3-6，我们可以看出，本文提出的 SSTCACAP 算法在两个数据集上的分类性能是最好的，优于另外两种算法 EM 算法和朴素贝叶斯算法。因

此，可以表明 SSTCACAP 算法是一种可行且有效的半监督文本分类算法。

3.4 本章小结

本文提出了一种基于蚁群聚集信息素浓度的半监督文本分类算法。算法在计算蚁群聚集信息素浓度时，把文本相似度作为其中的一个重要的因素，扩展了聚集信息素浓度计算模型。在半监督学习过程中，首先，基于聚集信息素的作用，利用 Top-k 策略选取出未标记蚂蚁可能归属的种群；然后，依判断规则，判定未标记蚂蚁的置信度；最后，采用随机选择的策略，把置信度高的未标记蚂蚁加入到对其最有吸引力的一个训练种群中，达到对训练种群进行扩展的目的。大量的实验表明：该算法在精确率、召回率以及 F_1 度量方面，都有着良好的表现。

在本文提出的基于蚁群聚集信息素浓度的半监督文本分类算法中，有一个重要的步骤就是需要对算法中的参数进行人工设置，并且参数选择的好坏也会影响分类的效果，因此下一步的工作，将是对算法中的参数优化进行研究，以期进一步提高该算法的性能。

第4章 基于特征映射的半监督文本分类算法

4.1 相关背景

4.1.1 背景介绍

现有的基于半监督学习的文本分类算法，一般是基于聚类假设和流形假设等统计理论中的样本分布一致性假设，即认为已标注数据的样本分布与未标注数据的样本分布是一致的。但是，这种假设往往在现实世界中是很难成立的，一旦出现未标注数据与已标注数据分布不同的情况，通过半监督学习有可能会导致学习器的性能下降。同样，当训练集数据的样本分布与测试集数据的样本分布不同时，也有可能导致通过训练集学习到的分类器性能下降[20]。

迁移学习是一种解决不同数据分布的方法[41-48]，20 世纪 90 年代学者们开始逐渐关注使用知识间的相互联系，通过利用已获得的知识来学习新的知识。迁移学习是一个非常容易理解的方法，例如：对于一个精通 Java 编程语言的程序员来说，他很容易就学会 C#、Python 等其他编程语言。因此，人类本身就具有这种迁移学习的能力。同样，对于机器学习来讲也可以具有这种迁移学习的能力。目前，基于特征的方法是迁移学习其中的一种主要方法。这种方法的主要思想是，认为相似领域的知识可能会有一些公共的特征，因此可以通过特征来进行迁移学习，来解决样本分布不一致的问题。Dai[49]等人提出了一种基于特征翻译的迁移学习算法，其主要思想是利用跨领域的特征来解决训练集数据与测试集数据分属于不同特征空间的问题，该方法能够借助不相关的数据来辅助目标领域的分类和聚类学习，实验结果证明该方法是有效的。Meng[50]等人提出了一种基于特征映射的迁移学习算法，该方法首先通过构造一个源领域和目标领域的新的特征子空间，在此基础之上，利用训练集数据和测试集数据相似的特征，建立一个特征映射函数，然后利用该映射函数，重新计算训练集和测试集数据的权重，实验结果也表明该方法的有效性。

本文借鉴了迁移学习中特征映射的思想，提出了一种基于特征映射的半监督文本分类算法（Semi-supervised Text Classification Algorithm based on Feature Mapping，SSTCFM），首先通过不同的特征选择算法，分别在训练集的已标记数据、未标记数据以及测试集数据中选择各自相应的特征集，并初始化特征的权值；在此基础之上，建立已标记数据和未标记数据之间的映射函数，已标记数据和测

试集数据之间的映射函数，未标记数据与测试集数据之间的映射函数，利用这三个特征映射函数重新给选取的特征赋予权重；最后通过 EM 算法进行半监督文本分类。

4.1.2 问题描述

为了便于下面的讨论，本文约定一些符号。$D=\{d_1,d_2,\ldots,d_n\}$表示数据集，D_L表示训练集中的已标记文档，D_U表示训练集中的未标记文档，D_T表示测试集文档。$F=\{f_1,f_2,\ldots,f_m\}$表示特征集合，F_L 表示在已标记文档中选取的特征，F_U 表示从未标记文档中选取的特征；F_T 表示从测试集选取的特征。半监督文本分类就是利用少量的 D_L 数据和大量的 D_U 数据学习一个分类模型，然后对 D_T 数据进行分类。

4.2 基于特征映射的半监督文本分类算法

4.2.1 算法主要思想

本文提出的算法由三部分组成：第一，通过不同的特征选择算法分别对训练集中的已标记文档、未标记文档以及测试集文档选取各自的特征集，并对已选取的特征赋予初始的权值；第二，建立三个特征映射函数，即已标记文档与未标记文档之间的特征映射，已标记文档与测试集文档之间的特征映射，未标记文档与测试集文档之间的映射函数，利用这三个特征映射函数来重新计算已选取特征的权值；第三，利用 EM 算法进行半监督文本分类。

首先，对于训练集中的已标记文档，由于其具有比较准确的类别标记信息，因此需要充分利用这一信息，我们采用 IG 算法（具体算法详见 2.3 节）进行特征选择，选取已标记文档的特征集 F_L；对于训练集中的未标记文档和测试集文档数据来说，由于它们都没有包含类别信息，因此我们利用 DF 算法（具体算法详见 2.3 节）进行特征选择，选取训练集中未标记文档的特征集 F_U，选取测试集文档的特征集 F_T。利用 F_L、F_U、和 F_T 构建一个初始的特征表示空间 $F=\{f_i, f_i \in F_L \cup F_U \cup F_T\}$，并对 F 中的每个特征 f_i 赋予初始的权值 w_i。

然后，建立已标记文档与未标记文档之间的特征映射函数 Φ_1，已标记文档与测试集文档之间的特征映射函数 Φ_2，未标记文档与测试集文档之间的特征映射函数 Φ_3。在映射函数 Φ_1 下，更新特征的权重，公式为：

$$w_i' = \begin{cases} a\,w_i & \text{if } f_i \in F_L \cap F_U \\ w_i & \text{else} \end{cases} \tag{4-1}$$

在特征映射函数 Φ_2 下，更新特征权重的公式为：

$$w_i'' = \begin{cases} b\,w_i' & \text{if } f_i \in F_L \cap F_T \\ w_i' & \text{else} \end{cases} \tag{4-2}$$

在特征映射函数 Φ_3 下，更新特征权重的公式为：

$$w_i''' = \begin{cases} l\,w_i'' & \text{if } f_i \in F_U \cap F_T \\ w_i'' & \text{else} \end{cases} \tag{4-3}$$

其中 α、β 和 λ 为权重调节参数，其取值都大于 1。

在特征权值更新完成后，对特征按照权值大小排序，选取其中 m 个特征作为新的特征表示空间。

最后，利用 EM 算法在训练集中的已标记文档和未标记文档学习到半监督文本分类模型，对测试集文档进行分类。

4.2.2 算法描述

具体的算法描述如表 4-1。

表 4-1 算法描述

Table 4-1 Algorithm Description

输入：训练集中的已标记文档 D_L，训练集中的未标记文档 D_U，测试集文档 D_T
输出：测试集文档 D_T 的类别信息
1： 对训练集中的已标记文档、未标记文档和测试集文档进行预处理，分别对 D_L 采用 IG 算法，D_U 和 D_T 采用 DF 算法进行特征选择，获得各自的特征集 F_L、F_U、和 F_T；构建初始的特征表示空间 $F=\{f_i, f_i \in F_L \cup F_U \cup F_T\}$，每个特征对应的权值为 w_i。
2： 建立三个特征映射函数 Φ_1、Φ_2 和 Φ_3，依次利用式子 4-1、4-2 和 4-3 更新权值 w_i；根据权值的大小，选取 m 个特征作为新的特征表示空间。
3： 利用 EM 算法进行半监督文本分类。

4.3 实验结果与分析

本章所有实验都是在配置为，处理器：英特尔 酷睿 2 四核 Q9450 2.66GHz；内存：4 GB DDR2 667MHz ；操作系统：Windows7 32 位的 PC 机上进行的。

4.3.1 实验数据集与评价标准

实验时我们使用了一个英文数据集和一个中文数据集。我们采用的英文数据集是著名的语料库 20 Newsgroups（20news-18828），挑选的类别与 3.3.1 节相同。我们采用的中文数据集是复旦大学提供的中文语料库，具体使用的类别详见 3.3.1 节。实验时，我们固定的在每一类文档中抽取 20%作为测试集数据，抽取 50%作为未标记数据集，其余的作为已标记数据集，标记比例从 5%-25%。

我们实验中对算法的评价指标采用了精确率（precision）、召回率（recall）和宏平均 F1 度量（Macro F1）三个比较通用的评价标准，具体的计算方法详见 2.5.2 节。

4.3.2 预处理及参数设置

实验时，本文对实验数据的预处理操作具体包括：文本格式化来清理文档数据中的无用信息和噪声等；对英文数据集使用 Porter Stemming 算法进行了词干提取、然后去除停用词；对中文数据集利用“结巴”分词进行了分词处理。

SSTCFM 算法的参数设置为 α=2，β=$\sqrt{3}$，λ=$\sqrt{2}$，w_i=1，m=500。

4.3.3 实验结果比较与分析

本文在 20news 英文数据集和复旦大学中文数据集上进行了多组实验，并比较了 SSTCFM 算法、EM 算法和朴素贝叶斯算法的分类效果。具体地，我们比较了三种算法在精确率、召回率和宏平均 F1 值三个方面的表现，并通过不断改变已标记文本集的大小，得出算法的精确率、召回率以及宏平均 F1 度量与已标记样本集大小之间的关系图。

(1) 20news 英文数据集

首先，对三种算法在 20news 数据集上的分类性能进行分析。

由图 4-1 可以看出，本文提出的 SSTCFM 算法在精确率方面表现是最稳定的，大部分情况下都要优于另外一种半监督文本分类算法 EM 算法，并且没有出现精确率低于 naïve bayes 算法的情况。而且，在已标记文档只有 5%时，本文提出的 SSTCFM 算法精确率达到了 55%，分别高出 EM 算法、naïve bayes 算法 3%和 5%。同时，我们还可以发现 EM 算法在已标记文档为 15%和 17.5%时，出现了精确率低于朴素贝叶斯算法的情况。我们认为这种情况的发生是由于 EM 算法基于样本分布一致性假设原理，当已标记文档与未标记文档的样本分布不一致时，通过半监督学习，可能会降低分类器的性能。通过对比可以发现，本文提出

的 SSTCFM 算法可以在一定程度上解决由于样本分布不一致而导致的半监督文本分类器性能下降的问题。

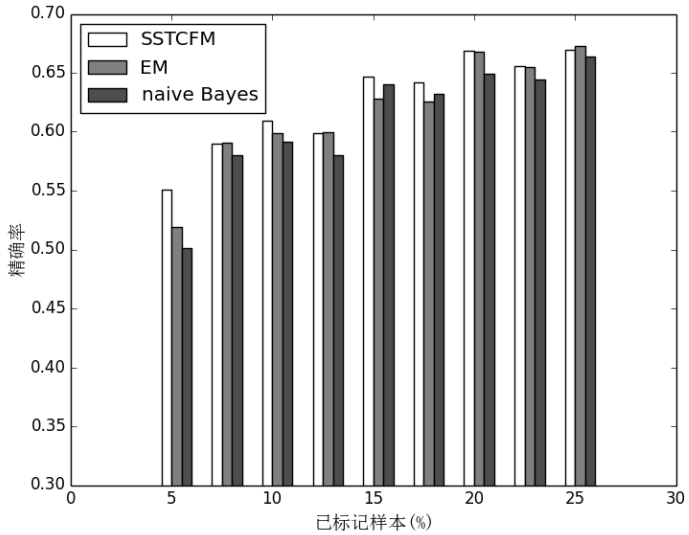

图 4-1　在 20news 上的精确率比较

Fig. 4-1　Comparison of precision on 20news

通过分析图4-2我们可以得知，SSTCFM算法在召回率上的表现是比较好的，分别在已标记文档为 5%，7.5%，10%，15%，17.5%，22.5%时取得了最好的分类效果。而另外一种半监督文本分类算法 EM 算法，在已标记文档为 12.5%，20%，25%时表现是最好的。但在已标记文档为 5%，10%，15%，17.5%时，出现了 EM 算法的召回率低于 naïve bayes 的情况。而本文提出的 SSTCFM 算法没有出现召回率表现不如朴素贝叶斯算法的情况，这也说明了 SSTCFM 算法是有效的。

图 4-3 是三种算法在宏平均 F_1 度量方面的表现对比图，通过观察不难发现：本文提出的 SSTCFM 算法的分类效果是比较不错的。在 9 组不同比例的已标记文档的实验中，SSTCFM 算法在其中 6 组的实验效果都是最优的，而 EM 算法在其余的 3 组实验中取得了最好的分类效果。但是，EM 算法在其中有 4 组的实验效果不如 naïve bayes，与图 4-1、4-2 中的情况类似，我们认为这是由于已标记文档与未标记文档，以及训练集文档与测试集文档的样本分布不一致，导致了半监督文本分类算法 EM 的分类性能下降。而本文提出的 SSTCFM 算法却没有发生这种特殊的情况，也再一次说明了 SSTCFM 算法的稳定有效性。

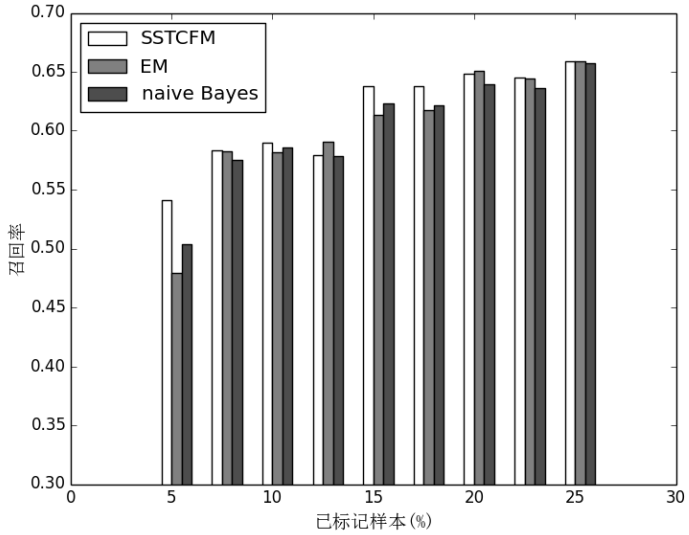

图 4-2　在 20news 上的召回率比较

Fig. 4-2　Comparison of recall on 20news

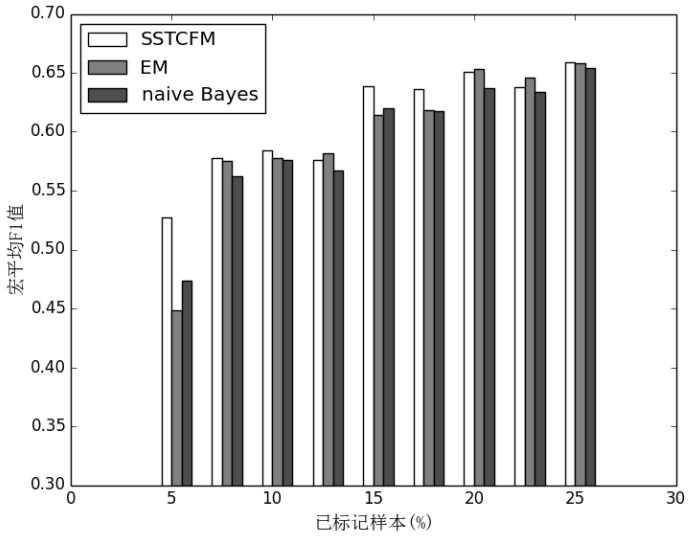

图 4-3　在 20news 上的宏平均 F_1 值比较

Fig. 4-3　Comparison of macro-F_1 comparison on 20news

综合对比图 4-1、4-2 和 4-3 的数据，我们可以看出，随着已标记文档的比例的提高，三种算法的分类性能都有一定程度的提高。在已标记文档比例低于 10%时，本文提出的 SSTCFM 算法的优势更加明显。并且 SSTCFM 算法与另外一种半监督文本分类算法不同的是，没有出现分类效果低于朴素贝叶斯算法的情况，这也说明了通过引入特征映射的思想，可以在一定程度上解决样本分布不一致而可能导致分类器性能下降的问题。

(2) 复旦大学中文数据集

然后，对三种算法在复旦大学中文数据集上的分类性能进行分析。

由图 4-4 可以看出，本文提出的 SSTCFM 算法在精确率上表现是最好的，与另外两种算法相比具有明显的优势。且 SSTCFM 算法的精确率在绝大部分情况下高于 70%，只有在已标记比例为 5%时，精确率为 65.6%，但仍然比 EM 算法高 14.7%，比朴素贝叶斯算法高 31.3%。另一种半监督文本分类算法 EM 算法的精确率大部分情况下在 50%和 60%之间，其在已标记比例为 10%时，达到了最高值 63.8%，但仍然比 SSTCFM 算法低 10.3%。而朴素贝叶斯算法的精确率大部分情况下低于 40%。由此可见，半监督文本分类算法的有效性。

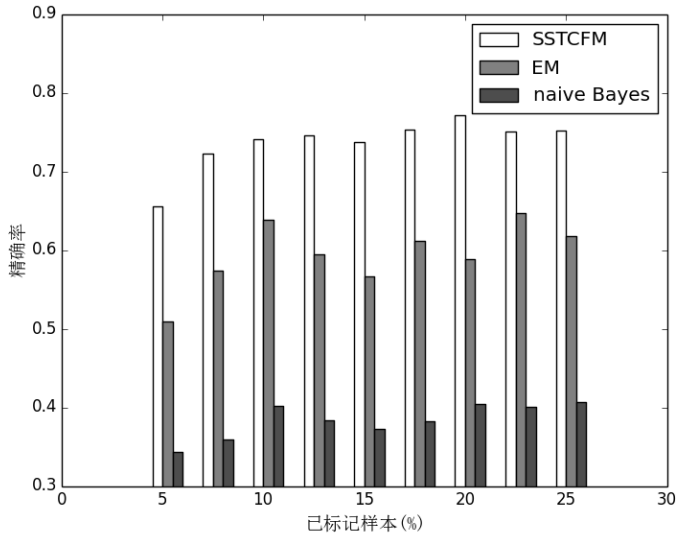

图 4-4　在复旦大学数据集上的精确率比较

Fig. 4-4　Comparison of precision on Fudan dataset

由图 4-5 可以发现，本文提出的 SSTCFM 算法在召回率上的表现仍然是最好的，没有出现召回率低于 EM 算法和朴素贝叶斯算法的情况。SSTCFM 算法在标记比例为 20%时达到了最高值为 76.3%，标记比例是 5%时出现了最低值 64%。EM 算法的召回率在标记比例为 10%时达到了最高值 62.8%，在标记比例为 5%时达到了最低值 48.5%。朴素贝叶斯算法的召回率在已标记样本比例为 25%时达到了最高值 40.4%，在标记比例为 5%时达到了最小值 33.2%。同时，我们可以看出两种半监督文本分类算法 SSTCFM 算法和 EM 算法在召回率方面的效果全部优于朴素贝叶斯算法。且 SSTCFM 算法的召回率比 EM 算法平均高 10%以上，比朴素贝叶斯算法平均高 30%以上。

由图 4-6 可以得知，我们提出的 SSTCFM 算法在宏平均 F1 值方面表现是最好的，全部高于 EM 算法和朴素贝叶斯算法。而且 EM 算法在宏平均 F1 值上全部高于朴素贝叶斯算法。同时，可以发现 SSTCFM 算法在标记比例大于 7.5%时，

macro-F1 值稳定在 70%以上，没有出现 macro-F1 值下降较多的情况。而 EM 算法的 macro-F1 值在标记比例为 10%时达到了最高值 63%，但是在标记比例为 12.5%、15%分别是 58.5%、56.4%，出现了比较明显的降低。这也表明，我们提出的 SSTCFM 算法要比 EM 算法稳定。

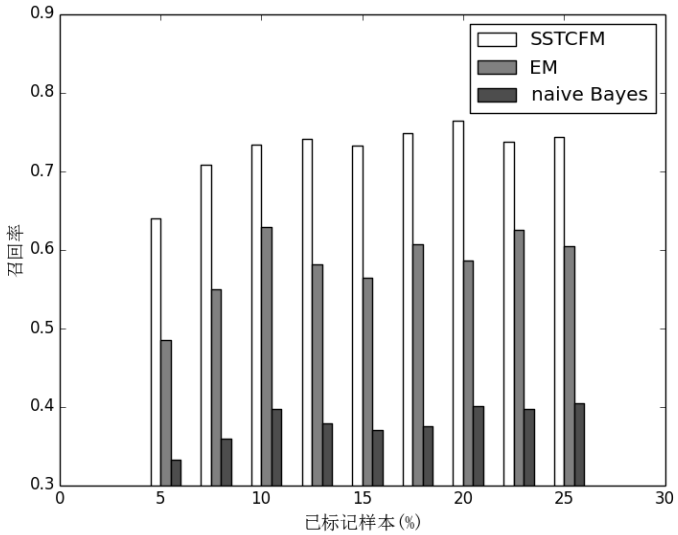

图 4-5　在复旦大学数据集上的召回率比较

Fig. 4-5　Comparison of recall on Fudan dataset

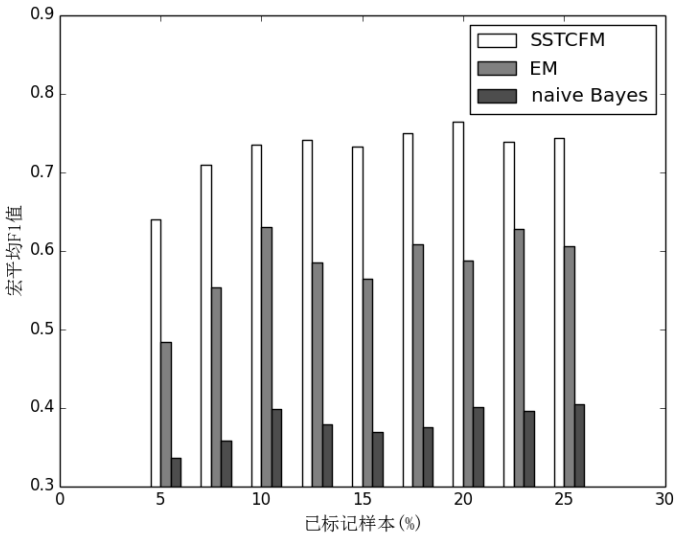

图 4-6　在复旦大学数据集上的宏平均 F_1 值比较

Fig. 4-6　Comparison of macro-F_1 on Fudan dataset

综合观察图 4-4，4-5，4-6 的数据，不难发现本文提出的 SSTCFM 算法，不仅在精确率、召回率和宏平均 F_1 值三个方面性能是最优的，而且随着已标记数据比例的不断变化，SSTCFM 算法的分类性能是最稳定的。这也再一次说明了

SSTCFM 算法的有效性。

4.4 本章小结

本文通过借鉴迁移学习中的特征映射思想，提出了一种基于特征映射的半监督文本分类算法，首先通过不同的特征选择算法，分别在训练集的已标记数据、未标记数据以及测试集数据选取三个不同的特征集，并给每个选取的特征赋予初始的权重；在此基础之上，建立已标记数据和未标记数据之间的映射函数，已标记数据和测试集数据之间的映射函数，未标记数据与测试集数据之间的映射函数，通过这三个特征映射函数来重新计算特征的权重；最后利用 EM 算法进行半监督文本分类。在标准数据集上的实验表明，本文提出的算法在精确率、召回率和宏平均 F_1 值三个方面都取得了比较好的效果。

结 论

随着互联网技术的飞速发展，人们能够很容易地获取大量的无标记的文本数据。但是传统的有监督学习技术却无法利用这些无标记的文本数据，显然这就造成了对资源的极大浪费。而且迈入大数据时代以后，如何充分利用海量的无标记文本数据的潜在价值信息，已经成为文本挖掘领域的前沿课题之一。因此能够利用无标记文本数据进行学习的半监督文本分类技术逐渐引起了国内外学者的广泛关注。本文对基于半监督学习的文本分类算法进行了比较深入的研究。

(1) 针对基于半监督学习的文本分类中已标记数据与未标记数据分布不一致可能导致分类器性能较低的不足，提出了一种基于蚁群聚集信息素的半监督文本分类算法。算法将聚集信息素与传统的文本相似度计算相融合，首先，基于聚集信息素的作用，利用 Top-k 策略选取出未标记蚂蚁可能归属的种群；然后，依据判断规则，判定未标记蚂蚁的置信度；最后，采用随机选择的策略，把置信度高的未标记蚂蚁加入到对其最有吸引力的一个训练种群中。大量的实验表明：该算法在三个不同的评价标准上，都有着良好的表现。

(2) 为了解决传统的半监督文本分类算法严重依赖于数据分布的问题，提出了一种基于特征映射的半监督文本分类算法。首先通过不同的特征选择方法，分别在训练集的已标记数据、未标记数据以及测试集数据选取各自的特征集，并初始化特征的权值；在此基础之上，建立已标记数据和未标记数据之间的映射函数，已标记数据和测试集数据之间的映射函数，未标记数据与测试集数据之间的映射函数，利用这三个特征映射函数重新计算特征的权重；最后利用 EM 算法进行半监督文本分类。在标准数据集上的实验表明，本文提出的算法是非常有效的。

本文提出的两种算法，都在一定程度上解决了数据分布不一致可能导致半监督文本分类器性能下降的缺点。但是，它们具有各自不同的优势，基于蚁群聚集信息素的半监督文本分类算法的分类性能更好，但是基于特征映射的半监督文本分类算法的分类效率更高。

未来的研究工作包括：第一，使用无参的方法改进基于蚁群聚集信息素的半监督文本分类算法，使其不需要人工调整参数，能有更好的自适应性；第二，将基于特征映射的半监督文本分类学习框架应用到更多的半监督文本分类中。

参考文献

[1] Aggarwal C C, Philip S Y. On clustering massive text and categorical data streams[J]. Knowledge and information systems, 2010, 24(2): 171-196.

[2] Wang Q, Li P. Randomized Bayesian Network Classifiers[M]//Multiple Classifier Systems. Springer Berlin Heidelberg, 2013: 319-330.

[3] James G, Witten D, Hastie T, et al. Support Vector Machines[M]//An Introduction to Statistical Learning. Springer New York, 2013: 337-372.

[4] Qi Z, Tian Y, Shi Y. Robust twin support vector machine for pattern classification[J]. Pattern Recognition, 2013, 46(1): 305-316.

[5] Hsu C W, Lin C J. A comparison of methods for multiclass support vector machines[J]. Neural Networks, IEEE Transactions on, 2002, 13(2): 415-425.

[6] Weinberger K Q, Saul L K. Distance metric learning for large margin nearest neighbor classification[J]. The Journal of Machine Learning Research, 2009, 10: 207-244.

[7] Frey B J, Dueck D. Clustering by passing messages between data points[J]. science, 2007, 315(5814): 972-976.

[8] Cai D, He X, Han J. Locally consistent concept factorization for document clustering[J]. Knowledge and Data Engineering, IEEE Transactions on, 2011, 23(6): 902-913.

[9] 朱岩. 面向文本数据的半监督学习研究[D]. 北京：北京交通大学, 2012.

[10] Zhu X, Goldberg A B. Introduction to semi-supervised learning[J]. Synthesis lectures on artificial intelligence and machine learning, 2009, 3(1): 1-130.

[11] Zhou Z H, Li M. Semi-supervised learning by disagreement[J]. Knowledge and Information Systems, 2010, 24(3): 415-439.

[12] Wang Y, Chen S, Zhou Z H. New semi-supervised classification method based on modified cluster assumption[J]. Neural Networks and Learning Systems, IEEE Transactions on, 2012, 23(5): 689-702.

[13] Sebastiani F. Machine learning in automated text categorization[J]. ACM computing surveys (CSUR), 2002, 34(1): 1-47.

[14] Maron M E. Automatic indexing: an experimental inquiry[J]. Journal of the ACM (JACM), 1961, 8(3): 404-417.

[15] Rosenblatt F. Principles of neurodynamics. perceptrons and the theory of brain mechanisms[R]. CORNELL AERONAUTICAL LAB INC BUFFALO NY, 1961.

[16] Salton G, Wong A, Yang C S. A vector space model for automatic indexing[J]. Communications of the ACM, 1975, 18(11): 613-620.

[17] Hayes P J, Weinstein S P. CONSTRUE/TIS: A System for Content-Based Indexing of a Database of News Stories[C]//IAAI. 1990, 90: 49-64.

[18] 苏金树，张博锋，徐昕．基于机器学习的文本分类技术研究进展 [J]．软件学报，2006, 17(9): 1848-1859.

[19] Shahshahani B M, Landgrebe D A. The effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes phenomenon[J]. Geoscience and Remote Sensing, IEEE Transactions on, 1994, 32(5): 1087-1095.

[20] 周志华，王珏．机器学习及其应用[M]．北京：清华大学出版社, 2007: 259-275.

[21] Nigam K, McCallum A K, Thrun S, et al. Text classification from labeled and unlabeled documents using EM[J]. Machine learning, 2000, 39(2-3): 103-134.

[22] Rowley H A, Baluja S, Kanade T. Neural network-based face detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(1): 23-38.

[23] Fujino A, Ueda N, Saito K. A hybrid generative/discriminative approach to semi-supervised classifier design[C]//Proceedings of the National Conference on Artificial Intelligence. Menlo Park, CA, AAAI Press, 2005, 20(2): 764-769.

[24] Blum A, Chawla S. Learning from Labeled and Unlabeled Data using Graph Mincuts[C]//Proceedings of the Eighteenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc., 2001: 19-26.

[25] Pang B, Lee L. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts[C]//Proceedings of the 42nd annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2004: 271-278.

[26] Kulis B, Basu S, Dhillon I, et al. Semi-supervised graph clustering: a kernel approach[J]. Machine Learning, 2009, 74(1): 1-22.

[27] Blum A, Mitchell T. Combining labeled and unlabeled data with co-training[C]//Proceedings of the eleventh annual conference on Computational learning theory. ACM, 1998: 92-100.

[28] Goldman S A, Zhou Y. Enhancing Supervised Learning with Unlabeled Data[C]//Proceedings of the Seventeenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc., 2000: 327-334.

[29] Zhou Z H, Li M. Tri-training: Exploiting unlabeled data using three classifiers[J]. Knowledge and Data Engineering, IEEE Transactions on, 2005, 17(11): 1529-1541.

[30] Li M, Zhou Z H. Improve computer-aided diagnosis with machine learning techniques using undiagnosed samples[J]. Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on, 2007, 37(6): 1088-1098.

[31] Yarowsky D. Unsupervised word sense disambiguation rivaling supervised methods[C]//Proceedings of the 33rd annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, 1995: 189-196.

[32] Riloff E, Wiebe J, Wilson T. Learning subjective nouns using extraction pattern bootstrapping[C]//Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4. Association for Computational Linguistics, 2003: 25-32.

[33] Maeireizo B, Litman D, Hwa R. Co-training for predicting emotions with spoken dialogue

data[C]//Proceedings of the ACL 2004 on Interactive poster and demonstration sessions. Association for Computational Linguistics, 2004: 28.

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

[35] 郑海清, 林琛, 牛军钰. 一种基于紧密度的半监督文本分类方法[J]. 中文信息学报, 2007, 21(3): 54-60.

[36] Salton G. The Smart document retrieval project[C]//Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1991: 356-358.

[37] 吴军. 数学之美[M]. 北京：人民邮电出版社, 2012: 27-33.

[38] Porter M F. An algorithm for suffix stripping[J]. Program: electronic library and information systems, 1980, 14(3): 130-137.

[39] Halder A, Ghosh S, Ghosh A. Aggregation pheromone metaphor for semi-supervised classification[J]. Pattern Recognition, 2013, 46(8): 2239-2248.

[40] Tsutsui S. Ant colony optimisation for continuous domains with aggregation pheromones metaphor[C]//Proceedings of the The 5th International Conference on Recent Advances in Soft Computing (RASC-04). 2004: 207-212.

[41] Pan S J, Yang Q. A survey on transfer learning[J]. Knowledge and Data Engineering, IEEE Transactions on, 2010, 22(10): 1345-1359.

[42] Raina R, Battle A, Lee H, et al. Self-taught learning: transfer learning from unlabeled data[C]//Proceedings of the 24th international conference on Machine learning. ACM, 2007: 759-766.

[43] Dai W, Yang Q, Xue G R, et al. Boosting for transfer learning[C]//Proceedings of the 24th international conference on Machine learning. ACM, 2007: 193-200.

[44] Pan S J, Kwok J T, Yang Q. Transfer Learning via Dimensionality Reduction[C]//AAAI. 2008, 8: 677-682.

[45] Perlich C, Dalessandro B, Raeder T, et al. Machine learning for targeted display advertising: Transfer learning in action[J]. Machine Learning, 2013: 1-25.

[46] Shao H, Tong B, Suzuki E. Extended MDL principle for feature-based inductive transfer learning[J]. Knowledge and information systems, 2013, 35(2): 365-389..

[47] Yang L, Hanneke S, Carbonell J. A theory of transfer learning with applications to active learning[J]. Machine learning, 2013, 90(2): 161-189.

[48] Yang Q. Big data, lifelong machine learning and transfer learning[C]//Proceedings of the sixth ACM international conference on Web search and data mining. ACM, 2013: 505-506.

[49] Dai W, Chen Y, Xue G R, et al. Translated Learning: Transfer Learning across Different Feature Spaces[C]//NIPS. 2008: 353-360.

[50] 孟佳娜. 迁移学习在文本分类中的应用研究[D]. 大连: 大连理工大学, 2011.

攻读硕士学位期间所发表的学术论文

1 杜芳华，冀俊忠，吴晨生，吴金源. 基于蚁群聚集信息素的半监督文本分类算法. 计算机工程.(中文核心期刊，已录用)

2 杜芳华，冀俊忠，吴晨生. 软件著作《科技词条关系分析软件》[简称：词条分析软件] V1.0,2013 登记号：2013SR102082

3 吴金源，冀俊忠，吴晨生，杜芳华. 基于类别加权和方差统计的特征方法研究. 北京工业大学学报.(中文核心期刊，已录用)

4 杜芳华，冀俊忠，吴晨生，吴金源. 基于特征映射的半监督文本分类算法.(投稿中)

致 谢

首先要感谢我的导师冀俊忠教授！在三年的研究生学习生涯中，冀老师为我提供了干净整洁的实验室，使我获得了良好的科研学习环境。读研期间，冀老师认真负责地教会我如何搞科研、做学问，使我能够接触到文本挖掘等科研工作的前沿领域，让我受益颇丰。并且，冀老师在做学问方面治学严谨的态度、踏实肯干的作风都对我产生了深远的影响，让我走在了一条正确的科研道路上，并顺利完成硕士研究生的学业。再一次，对我的导师冀俊忠教授表示衷心的感谢和崇高的敬意！

感谢吴晨生老师对我工作的细心指导，吴老师那发散的思维方式、看待问题的独特角度等都对我产生了深刻的启迪！

感谢实验室已经毕业的宋向京、刘红欣、房小娟、刘志军等师兄师姐，他们在学习、生活、工作等各方面都给予了我极大的帮助；感谢吴金源、焦朗、宋辰、杨翠翠、张玲玲、吕嘉伟、潘飞、贝飞、柴鹰、韩跃等同学，他们在研究思路、方法和编程技术等方面都给了我具体的帮助和建议！

感谢我的室友武辰之、姚治成、管洋洋，他们在计算机技能方面给予了我非常多的帮助，使我取得了很大的进步！

感谢我的家人，他们对我的理解和支持，是我能够顺利完成学业的动力源泉！

最后，特别感谢能够在百忙之中参与本文评阅和答辩的各位专家、学者和老师！

阅读此文的还阅读了：

26. 基于VSM的文本分类挖掘算法综述

27. 基于PU学习算法的文本分类研究与实现

28. 基于改进 TFIDF 算法的文本分类研究

29. 基于SVM的半监督两分类算法的EICMM研究

30. 基于半监督SVM主动学习的文本分类算法研究

31. 基于机器学习的专利文本分类算法研究综述

32. 基于半监督学习算法在文本分类中的应用研究

33. 基于ESA的文本分类算法研究

34. 一种基于半监督学习的非平衡分类算法

35. 基于半监督LDA的文本分类应用研究

36. 基于SVD和部分聚集分类的文本挖掘算法

37. 基于半监督的SVM迁移学习文本分类算法

38. 基于改进潜在语义分析算法的文本情感分类研究

39. 基于Matlab的极限学习机分类算法

40. 一种新的基于SVM的文本分类增量学习算法

41. 基于半监督的SVM迁移学习文本分类算法

42. 基于机器学习聚类算法的学习者自动分类研究

43. 基于近邻分类的增量学习分类算法研究

44. 基于机器学习的文本分类技术研究进展

45. 中文本体的自动学习算法研究

46. 基于机器学习的Web文本自动分类

47. 诱导半监督聚类学习算法

48. 基于深度学习的文本分类研究进展

49. 一种基于聚类密度的文本分类算法研究

50. 基于交叉覆盖算法的文本分类研究