反馈方法在文本分类系统中的应用

庞剑锋 程学旗
中国科学院计算技术研究所 100080
E-mail:pangjf@ncic.ac.cn

摘要：在基于向量空间模型的文本分类系统中，训练文档的数量和质量是决定系统性能的很重要的因素。而收集、筛选和整理训练文档通常是一件费时费力的事情。本文通过在文本分类系统中应用反馈方法，大大地减少了系统在训练过程中对训练文档数量的要求。实验表明随着分类结果的不断反馈，系统的性能也能随之逐步提高，从而逐步达到较满意的结果。

关键词： 文本分类 反馈方法 向量空间模型

Employing Feedback methods in Text Categorization System

Pang Jianfeng Cheng Xueqi
Institute of Computing Technology, CAS 100080
Email:Pangjf@ncic.ac.cn

Abstract: In Text Categorization systems based on VSM, the quality and quantity of the training documents is one of the most important factors which affect performance. But gathering, filtering and classifying the training documents is very difficult. This paper employs Feedback methods for Text Categorization systems and reduces the need for labeled training documents. The experimental results show that the system's performance can be improved by the Feedback procedure.

Key words: Text Categorization Feedback methods Vector Space Model

1. 引言

文本分类是指在给定的分类体系下，根据文本的内容自动地确定与文本关联的类别。当前通常采用机器学习的方法构造文本分类系统，系统一般由训练过程和分类过程两部分组成。其中，训练过程的训练结果决定了文本分

类系统所具备的分类能力，这个分类能力在分类过程中是固定不变的。

因此，很多研究文本分类系统的学者都竭尽全力地改善训练结果，以此作为提高文本分类性能的唯一途径。而改善训练结果的方法一般不外乎两种，一是改进算法，二是增加训练文档的数量，或提高训练文档的质量。

但是不幸的是，收集整理训练文本，为训练文本标注正确的类别是一件很费时费力的事。细细阅读上万篇的训练文本，筛选文本并精确地标注类别即使对专业人员也不轻松。而且，如果更改了分类体系，那么原先的训练文本整理工作将全部推倒重来，这将造成人力物力的巨大浪费。

那么我们是否可以在训练过程中投入较少的精力，而力争在分类过程中通过再学习的反馈过程提高分类性能呢？答案是肯定的，笔者将在本文的以下部分讨论反馈方法在文本分类算法中的应用。

2. 反馈方法的基本思想和适用情况

反馈方法的实质是将分类系统的“训练——分类”过程扩展为“训练——分类——反馈”过程，结合了反馈过程的文本分类系统和传统的文本分类系统相比，在训练过程和分类过程上没有任何不同，唯一的区别是增加了反馈过程。反馈过程的主要任务是系统判断或由人工判断分类结果是否满意，对结果满意的情况，利用新文本的分类结果调整先前的训练结果。

反馈过程主要包括两个问题：什么样的分类结果是满意的结果；如何调整训练结果。关于第一个问题，有两种解决方法，第一，可以人为参与反馈判断过程，决定分类结果是否可以反馈。另一种解决方法是机器自反馈，如果新文本与类别的相关性大于该类别的阈值，那么机器将对训练过程进行反馈，从而改进和加强训练结果。

第二个问题，如何调整训练结果相对容易解决，最简单的反馈方法是将新文本及其类别信息添加到训练文本中，对整个分类系统再重新进行一次训练，显然这是不可取的。训练过程一般都十分耗时，这样也失去了反馈的意义。应在已经产生的训练结果上进行调整。

反馈方法主要适用于训练不充分，整理训练文本比较困难的文本分类系统，另外，也适用分类体系可能会经常变更的情况。在这些情况下，结合反馈方法能取得事半功倍的效果。

当前存在很多种基于向量空间模型的文本分类算法，例如 Bayes 分类算法、支持向量机分类算法、类中心向量最近距离判别分类算法和 K-近邻分类算法等等。几乎所有这些文本分类算法都可以结合反馈方法，一般而言，结合反馈方法后，这些算法过程扩展为：

训练——>分类——>反馈判断——>反馈

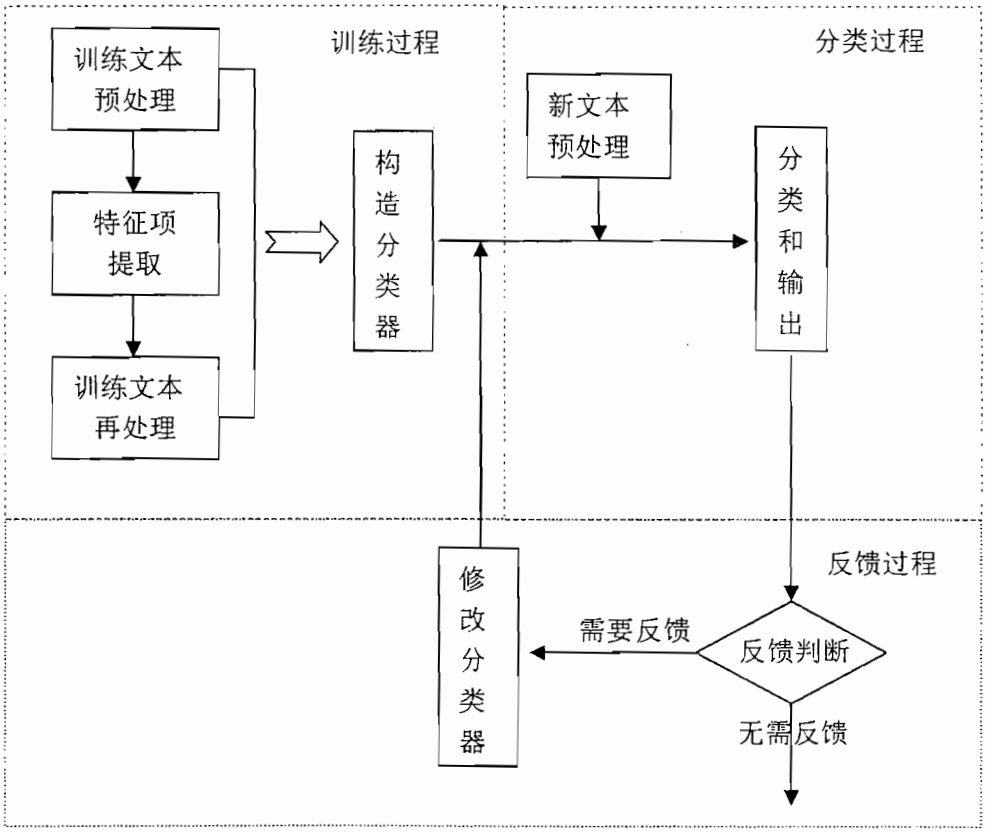
训练过程和分类过程与传统的分类系统相同，只是需要确定类别的相关

性阈值。反馈判断过程将分类过程得到的文本与类别的最大相关性同训练过程得到的该类别的阈值比较，如果最大相关性大于阈值，则进行反馈，否则，放弃反馈。反馈过程则根据新文本向量及其对应的类别信息相应地调整训练结果，对于不同的分类算法，反馈调整步骤也不相同，但是必须要使反馈过程快速高效。

3. 应用反馈方法的文本分类系统的结构和关键算法

3.1 应用反馈方法的文本分类系统的结构

本系统在传统的文本分类系统上结合了本文以上介绍的反馈方法，结构如下图所示：



如图所示，系统主要由三个部分组成，它们分别是训练过程、分类过程和反馈过程。训练过程的关键任务是构造分类器，其他模块都是为该任务服务的，分类过程是在构造完成的分类器上对新文本分类的过程，而反馈过程

主要实现了第二节的反馈方法，对分类器进行调整，该过程为可选的部分，对于训练文档充分的系统，可以略去该过程。

3.2 关键算法描述

3.2.1 结合反馈方法的 Bayes 分类算法

在传统的 Bayes 算法上的反馈过程如下：

输入：新文本向量 $D(w_1, w_2, \ldots\ldots w_n)$，对应的类别 C

过程：

STEP 1：读出类别 C 的几率向量 $C(p_1, p_2, \ldots\ldots p_n)$ 和该类包含的所有词数 m 及特征项数 n（训练过程中已经得到）

STEP 2：求出类别 C 的总词数向量 $C'(q_1, q_2, \ldots\ldots q_n)$，其中：

$$q_i = (m+n) \cdot p_i - 1$$

STEP 3：按照如下公式：

$p_i' = \dfrac{1 + q_i + w_i}{n + \sum_{j=1}^{n} (q_i + w_i)}$ 计算类别 C 的调整后的几率向量

$(p_1', p_2', \ldots\ldots p_n')$

3.2.2 结合反馈方法的类中心向量最近距离判别分类算法

在类中心向量最近距离判别算法上的反馈过程描述如下：

输入：新文本向量 $D(w_1, w_2, \ldots\ldots w_n)$，对应的类别 C

过程：

STEP 1：读出类别 C 的中心向量 $C(p_1, p_2, \ldots\ldots p_n)$ 和该类包含的

所有词数 m 及特征项数 n（类中心向量在训练过程中由每类所有训练文本向量算术平均而得到）

STEP 2：求出类别 C 的几何平均前的中心向量$C'(q_1,q_2,.....q_n)$，其中：

$$q_i = n \cdot p_i$$

STEP 3：按照如下公式：

$p_i^{'} = \frac{q_i + w_i}{n+1}$ 计算类别 C 的调整后的中心向量

$(p_1^{'}, p_2^{'},.....p_n^{'})$

3.2.3 结合反馈方法的 K-近邻算法

K-近邻算法增加的反馈过程十分简单，只是将新文本向量$D(w_1,w_2,.....w_n)$和对应的类别 C 添加到原来记忆的训练文本集中，以供下一次分类过程使用。

4. 实验结果与分析

4.1 分类体系和语料

我们在两个语料库上测试我们的系统，这两个语料库分别是：新闻语料库 2830 篇[1]，北图法律专业语料库 1780 篇。

其中，新闻语料库 2830 篇中的文本都是新闻电讯稿，绝大部分采自新华社，还有 200 余篇采自中国新闻社和人民日报。所有的新闻稿都由领域专家事先进行分类，按照中图分类法分成经济类、政治类、等 38 个大类。

北图法律专业语料库 1780 篇的文本全部由北京图书馆提供，文本大多来自法律类报刊杂志和法律书籍，这些文本预先经北图信息处理员分类，包括案例类、财政法类、自然资源等共 23 个类。

4.2 测试内容和测试结果

4.2.1 测试方法

测试时，我们将这些分好类的语料平均分成十份，任意选择其中一份作为最初构造分类器的训练样本，再任意选择一份作为开放测试集，剩余的八份作为反馈分类文本集供系统分类反馈使用。最初构造的分类器逐渐对这八份反馈分类文本集进行分类反馈处理，即逐步地学习积累，改进分类器。在分类反馈完一份、二份、……八份文本集后，都对开放测试集进行准确率测试，记录实验数据。

同时，我们也进行没有反馈过程的系统准确率的对比实验，即，将上面提到的最初构造分类器的一份训练样本和八份反馈分类文本集都作为本次实验的训练文本，以此为基础构造分类器，对剩余的一份样本进行开放测试。

4.2.2 测试结果

1. Bayes 算法

反馈份数	新闻语料库准确率	法律语料库准确率
0（未反馈）	59.2%	44.1%
1	64.1%	50.2%
2	74.6%	54.8%
3	78.6%	58.8%
4	80.1%	60.0%
5	80.5%	58.9%
6	81.3%	60.3%
7	82.4%	61.2%
8	82.7%	62.9%
对比实验	85.8%	64.1%

2. 类中心向量最近距离分类算法

反馈份数	新闻语料库分类效率	法律语料库分类效率
0（未反馈）	72.6%	48.2%
1	75.6%	50.8%
2	78.2%	57.6%
3	78.7%	60.0%
4	79.4%	60.0%
5	80.1%	60.6%
6	79.8%	61.2%
7	81.6%	62.3%
8	82.3%	61.2%
对比实验	83.1%	62.3%

3. K-近邻分类算法

反馈份数	新闻语料库分类效率	法律语料库分类效率
0（未反馈）	58.4%	40.6%
1	63.3%	50.6%
2	74.9%	55.3%
3	79.4%	58.2%
4	79.0%	59.4%
5	83.5%	60.6%
6	84.0%	61.2%
7	84.6%	61.2%
8	87.3%	61.8%
对比实验	88.4%	62.9%

4.2.3 结果分析

从上面的数据中，我们可以很容易地发现：反馈方法对文本分类系统性能提高有很大作用，尤其是在应用分类反馈的初期，系统性能的提高是很明显的，因为最初构造的分类器是训练不完全的，随着反馈的应用，分类器不断进行调整，分类性能也不断提高。随着分类反馈的逐渐增加，分类器也逐渐从训练不充分的阶段走向训练充分的阶段，因此，在分类反馈的后期，系统性能的提高不明显，甚至个别情况还出现了微弱的性能下降。

同对比实验比较，对比实验中分类系统使用 9 份训练样本构造分类系统，应用反馈方法的分类系统仅使用 1 份训练样本构造分类系统，其实验结果的准确率显示两个系统的准确率并没有很大的差距，但是收集整理一份训练文本比收集整理九份训练文本的工作量大大降低。从而说明反馈方法对于训练不充分的文本分类系统的分类性能有很大的改进作用。

5. 进一步的工作

本文实现的结合反馈方法的文本分类系统对文本的分类结果仅考虑两种情况，反馈的分类结果和不反馈的分类结果，并不区分反馈的力度。同时在修改训练结果时也不区分原始的训练文本和分类过程中反馈给训练过程的新文本，而事实上，如果增加权重区分这些不同的文本可能会对系统的性能有所帮助，将来可以进行进一步的研究。

参考文献

[1] 黄萱菁，复旦大学，大规模中文文本的检索、分类与摘要研究，1998，博士论文。

[2] 黄萱菁、吴立德：独立于语种的文本分类方法，2000 International Conference on Multilingual Information Processing , pp 37-43 , 2000

[3] 鲁松、白硕等：文本中词语权重计算方法的改进，2000 International Conference on Multilingual Information Processing , pp 31-36 , 2000

[4] 卜东波：聚类/分类理论研究及其在大规模文本挖掘中的应用，博士论文 2000 11

[5] McCallum and Nigam, Employing EM and pool-based active learning for text classification. American Association for Artificial Intelligence Press 1998.