

中文训练集 version 1.1 说明

2005 年 4 月 5 日北大网络实验室

与面向英文的分类系统相比，中文分类系统的起步比较晚。从第五次 TREC 会议开始，增加了对中文分类系统的评测。实际上参加 TREC-5 的中文分类系统处理的重点还停留在中文的分词问题上，而且处理的对象还是新华社的新闻稿这类普通的中文文本。基于案例的有指导的机器学习方法是实现中文网页自动分类的理论基础。因此，中文网页训练集是实现中文网页自动分类的前提条件。但是，到目前为止，还没有出现标准的中文网页语料库，因此也没有出现针对中文网页分类系统的评测。

为了解决这一问题，2002 年秋天 北京大学网络与分布式实验室天网小组通过动员不同专业的几十个学生，人工选取形成了一个全新的基于层次模型的大规模中文网页样本集。它包括 11678 个训练网页实例和 3630 个测试网页实例，分布在 726 个类别中，每个类别平均有 20 个训练实例和 5 个测试实例。样本集中类别及实例数量的分布情况如表 1 所示。

该版本于 2003 年 version1.0 版本的区别在于修改了部分类别设置（主要是对于原有的“新闻与媒体”类别，因为类别界限模糊，与其他类别内涵重叠较大，因此在本版本中，去掉了该类别。）

表 1 样本集中类别及实例数量的分布情况表

类别编号	类别名称	类别数	训练样本数	测试样本数
1	人文与艺术	24	407	120
3	商业与经济	48	795	240
4	娱乐与休闲	88	1421	440
5	计算机与因特网	58	871	290
7	教育	18	279	90
8	区域	53	846	265
10	自然科学	113	1788	565
11	政府与政治	18	281	90
12	社会科学	104	1711	520
13	医疗与健康	136	2214	680
14	社会与文化	66	1065	330
共计		726	11678	3630

下面简要地介绍一下上述中文网页样本集的分类体系。我们主要借鉴了国外的一些分类标准，比较分析它们的特色和不足，并针对中文网页这一特殊对象，提出了我们的分类体系。

经过分析整理，本文最终采用的分类体系如图 1 所示。它包含三个层次，11 个大类，共 726 个类别。从总体上可以分为学术性和非学术性两大类。其中学术性类别按国家标准 GB/T 13745-92《学科分类与代码》分类。选用该分类体系的主要原因是它分类层次关系简单明了，中国用户比较熟悉。

