

# 2006 SEWM 中文分类评测指南

2006年3月31日北大网络实验室

## 1. 评测目的

与面向英文的分类系统相比，中文分类系统的起步比较晚，而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，国内一直缺乏标准的大规模中文网页语料库。

通过本次中文分类评测项目，我们希望为这个领域的研究人员、用户、企业提供一个交流的机会，希望在国内外各个研究小组的共同参与下建立并完善以中文为主的分类训练集，以进一步促进中文 Web 分类技术的发展。

对于中文分类技术而言，训练数据集的质量对评测结果有很重要的影响。2005年的评测结束之后，参与人员就数据集问题提出了许多建设性建议。为了衡量数据集对评测结果的影响，以期能真实反应一个中文网页分类器的质量，本次评测给出了2套不同的分类体系以及不同的训练集（2套训练集采用了同样的数据格式以方便参与队使用）。参与评测的队伍必须给出所有待分类网页的2套分类结果，评测将据此来得到综合的评测指标。并希望能够根据不同参赛队在2套数据集上的不同表现，来研究数据集对评测结果的影响程度，并最终能改进下一次的评测流程。

## 2. 数据集说明

### 2.1 训练集

本次中文网页分类评测共有2套训练集：(content1.txt,example1.dat)以及(content2.txt,example2.dat)。其中数据集1 (content1.txt,example1.dat) 是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11个大类别中。数据集2(content2.txt,example2.dat)的分类体系是根据常见的新闻类别而设定，从新闻网站上抓取得到对应类别的新闻网页作为训练集页面。它包括960个训练网页和240个测试网页，分布在8个类别中。

此次分类评测规定对每个给定的测试网页最多返回一个类别结果。

### 2.2 分类训练集设置

参赛队需要事先完成分类器的训练工作。分类器使用的训练集为纯文本文件，文件名为example1.dat以及example2.dat。（两套训练集的文件格式相同），训练文档按结构顺序存放，一个结构的定义为：

网页编号： 5个字节；（注意删除空格，下同）

类别编号： 6个字节；

网页url： 255个字节；

标志： 1个字节（0为训练，1为测试，用户也可以自己决

定是否作为训练集)；

网页长度： 7 个字节（根据这个数据，随后为网页的内容）；

网页内容： 不定长， 其长度第 5 个字段给出了（网页内容完整的记录了网页的 HTML 源代码，没有做任何预处理）。

## 2.3 待分类文档说明

本次评测的待分类文档即 SEWM2006 Web track 中所使用的小数据集。数据集的大小大约为 20G，为 CWT200g 的一个子集。数据格式说明请详见 CWT200G 的说明。

## 3.评测流程

以下是整个评测的详细流程：

- 1) 参赛队申请获得所需数据（包括 CWT200G,以及 2 套训练集数据）。
- 2) 各参加评测单位根据训练集数据 1 (content1.txt, example1.dat) 建立分类系统，给出 CWT200G 中部分指定网页的类别号，保存为结果 1。根据训练集数据 2(content2.txt, example2.dat)给出 CWT200g 中部分指定网页的类别号，保存为结果 2。
- 3) 本评测不考虑一个网页属于多个类别的情况，因此给定一个网页，只需给出一个分类结果类别。
- 4) 各参加评测单位分类完成后，务必于 2006 年 6 月 10 日零点之前将结果提交至龚笔宏 ([gbh@net.pku.edu.cn](mailto:gbh@net.pku.edu.cn))，同时请提交分类系统的系统描述文档（系统描述文档至少应包括三个部分：系统模块结构说明；系统主要算法说明；以及系统运行环境说明。若分类过程中有人工参与，请务必说明人工参与的步骤）。逾期将视为自动放弃评测资格。结果格式如下：

（一个结果一行，各项之间用空格隔开）： docno cateno sim

docno	结果网页 url 对应的文档编号。(为文档数据中的 URL 字段内容对应的文档编号, url 与 docno 对照表参见 CWT200G 的对照表)
cateno	该文档所属于的类别。(例如该文档属于 01 类别, 类别号在 content.txt 文件中定义)
sim	文档与该类别的相似度计算值

- 5) 各参加评测单位分类结果收集完成后，我们将从待分类文档中随机抽取 1000 个网页作为目标网页进行评测，为了公平起见，网页抽取遵循三条原则：不是纯英文网页，不限编码格式，网页平均分布于各类别中。评测结束后，抽取的网页集将提供给各单位下载。
- 6) 对于抽取的网页集，通过对各参加评测单位的分类结果进行统计，给出每个网页的候选类别结果，随后人工标注正确类别。
- 7) 根据人工标注分类结果，对各参加评测单位的分类结果进行综合评测。为公平起见，北大天网将不参加评测。

## 4.评测指标

本次评测主要是考虑分类系统的准确性以及完整性，因此主要评测指标有精度

precision, 召回率 recall, 宏观 F1 值, 微观 F1 值。

1) 第  $i$  类的准确率 ( $P_i$ ):

此是指对于第  $i$  个类别, 所有待分类文本的分类结果正确比率。数学公式为:

$$P_i = \frac{I_i}{m_j}$$
 其中  $m_j$  是经分类系统输出分类结果为第  $i$  类的文档个数,  $I_i$  是在  $m_j$  中分类正确的文档个数。

2) 第  $i$  类的召回率 ( $R_i$ ):

指对于第  $i$  个类别, 分类结果的完整性。数学公式为:  $P_i = \frac{I_i}{n_i}$ , 其中  $n_i$  为所有测试文档中, 属于第  $i$  类的文档个数;  $I_i$  是经分类系统输出分类结果为第  $i$  类且结果正确的文档个数

3) 第  $i$  类的 F1 值 ( $F1_i$ ) 也称之为综合分类率:

其公式为  $F1_i = \frac{2 P_i R_i}{P_i + R_i}$  其中  $P_i$  为第  $i$  类的准确率,  $R_i$  为第  $i$  类的召回率。

4) 宏平均精度:

其公式为  $MacroP = \frac{1}{n} \sum_{j=1}^n P_j$ 。其中  $P_j$  为第  $j$  类的准确率,  $n$  为所有类别的总数。

5) 宏平均召回率:

其公式为  $MacroR = \frac{1}{n} \sum_{j=1}^n R_j$ 。其中  $R_j$  为第  $j$  类的召回率,  $n$  为所有类别的总数。

6) 宏平均 F1 值:

其公式为  $MacroF1 = \frac{MacroP \times MacroR \times 2}{MacroP + MacroR}$  其中  $MacroP$  为宏平均精度,  $MacroR$  为宏平均召回率。

## 4.其他

如有疑问, 请联系龚笔宏([gbh@net.pku.edu.cn](mailto:gbh@net.pku.edu.cn))。