

SEWM2005 中文网页分类评测指南

2005 年 4 月 5 日北大网络实验室

1. 评测目的

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。本次的评测的目的在于评测大规模网页分类的核心技术，其中包括网页文本分类以及链接关系辅助。在给定分类体系的前提下，参赛队必须给出 CWT100g 中所有网页的分类结果，评测将考察分类结果的准确性以及全面性。

2. 数据集说明

2.1 类别设置

本次中文网页分类竞赛使用的分类目录如表 1 所示,共 11 个类别。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。

表 1 分类目录

类别编号	类别名称
01	人文与艺术
03	商业与经济
04	娱乐与休闲
05	计算机与因特网
07	教育
08	各国风情
10	自然科学
11	政府与政治
12	社会科学
13	医疗与健康
14	社会与文化

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 `examples.dat`，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 CWT100g 的说明

详见 CWT100g 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 CWT100g , 以及分类器训练集数据)。
- (2) 各参加评测单位建立分类系统, 给出 CWT100g 中所有网页的类别号。类别设置请见本文档第 2 章节“数据集”部分。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。同时注意主办方所提供的训练集类别为 3 层层次结构, 但是在本次评测中, 只统计到分类至第一层大类的结果。例如: 对于某一网页, 只需给出该网页属于 01 类别, 而不需给出该网页属于 01 类别下的哪一子类。
- (3) 各参加评测单位分类完成后, 务必于 2005 年 8 月 25 日零点之前将结果提交至龚笔宏 (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 http://www.cwirf.org/2005WebTrack/050404url.no.tar.gz)
cateno	该文档所属于的类别。只需给到第一层分类结果, (例如该文档属于 01 类别, 而不需要给出属于 01 下的“摄影”子类别)。
sim	文档与该类别的相似度计算值

- (4) 各参加评测单位分类结果收集完成后，我们将从 CWT100g 中随机抽取 1200 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。
- (5) 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。
- (6) 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，仅提供分类结果作为参考。

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 *presicion*, 召回率 *recall*, 宏观 F1 值, 微观 F1 值。

4.1 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：
$$P_i = \frac{I_i}{m_i}$$
 其中 m_i 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_i 中分类正确的文档个数。

4.2. 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：
$$R_i = \frac{I_i}{n_i}$$
 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

4.3. 第 i 类的 F1 值 (FI_i) 也称之为综合分类率:

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

4.4 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

4.5 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。

4.6 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。

5.其他

有更多疑问, 请联系龚笔宏(gbh@net.pku.edu.cn)