中文网页自动分类竞赛规则

1. 分类目录

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。

表 1 分类目录

类别编号	类别名称
01	人文与艺术
02	新闻与媒体
03	商业与经济
04	娱乐与休闲
05	政府与政治
06	社会与文化
07	教育
08	自然科学
09	社会科学
10	计算机与因特网
11	医疗与健康

2. 训练集和测试集

a.训练集: 在现场比赛之前，参赛人员需要事先完成分类器的训练工作。分类器使用的训练集，参赛人员可以有两种选择：

1） 参赛人员自己准备。按照表 1 的分类目录，参赛人员自己准备网页训练集。

2） 北大网络实验室提供。

请参赛人员直接与我们联系，签署数据使用协议后，即可获得该数据集(约 1.4 万网页，约 200MB)。我们将免费提供(跟我们 mail 联系后,可从网上直接下载)。

训练集为纯文本文件，文件名为 examples.txt, 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）；
网页内容的长度： 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）。

说明: 1) 获得该数据集的步骤:

a: 到 http://net.cs.pku.edu.cn/~webg/infomall/reg_infomall.html 上注册.

b. 下载许可协议,打印,签字后请寄:

北京大学计算机系网络实验室,黄蕊收 100871

c: 同我们 mail 联系后,直接下载.

2) 训练集中没有一个网页属于两个类别的情况,需要参赛人员自己准备(我们自己也没有)

3) 其中的测试集格式同训练集一样,只是为方便测试,随机选的.

b.测试集：现场使用的测试集，将由程序委员会组织专家人工收集整理，比赛前不公布。最后在线测试的测试集格式同上面的训练集格式相同.文件名为 tests.txt.

3. 结果评价标准

本次竞赛仅考察分类器的分类查准率（没有考察查全率和分类效率）。为简单起见，每个给定测试网页最多取两个结果类别。中文网页分类器的质量按公式 1 计算：

分类器最后得分 $S = \sum_{i=1}^{n}$ 每个网页分类的得分 S_i (1)

其中：

决定每个网页的分类得分 S_i 的因素有两个：
1 ）结果类别的个数；一个或两个。
2 ）结果类别的次序。

S_i 由表 2 计算

分类结果	得分
个数和次序都对	1.0
个数对，次序不对	0.8
本来只有一个类别，却得到两个结果，正确的结果被放在第一位	0.7
本来只有一个类别，却得到两个结果，正确的结果被放在第二位	0.5
本来有两个类别，却只有一个结果，这个结果为第一位的结果	0.7
本来有两个类别，却只有一个结果，这个结果为第二位的结果	0.5
本来有两个类别，有两个结果，正确的结果被放在第一位	0.7
本来有两个类别，有两个结果，正确的结果被放在第二位	0.5
本来有两个类别，有两个结果，只有一个正确的结果,位置同原来的的不同	0.5
个数和次序都不对	0

例如：某个网页 p 的类别为 01,08 类，分类结果为：
分类结果 得分
1) 01,08 得分为 1
2) 08,01 得分为 0.8
3) 01 或 01,07 得分为 0.7
4) 08 或 02,08 得分为 0.5
5) 08,09 或 02,01 得分为 0.5
5) 02 得分为 0
6) 02,07 得分为 0

例如：某个网页 p 的类别为 01 类，分类结果为：
分类结果 得分
1) 01 得分为 1
2) 01,08 得分为 0.7
3) 08,01 得分为 0.5
4) 02 得分为 0
5) 02,07 得分为 0

4. 现场测试

1) 分类结果的输出格式

分类结果请输入到文件名为 results.txt 的纯文本文件，每一行的格式为：
网页编号 类别编号
或： 网页编号 类别编号 1 类别编号 2

2) 测试时时必须对每篇测试网页给出类别，不能为空.

3) 需现场编译.测试使用的系统平台为 Solaris,内核版本为:Sun OS 5.8;
gcc 版本为: gcc version 2.95.3 20010315 (release)
(经过商量,现在也允许使用 windows 系统.)

4) 测试集的大小(这得专家商量后决定了，估计不会很多)