中 文 信 息 学 报
第 17 卷 第 5 期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol. 17 No. 5

文章编号:1003 - 0077(2003)05 - 00034 - 07

关于“中文网页自动分类竞赛”结果的分析①

冯是聪,王继民

(北京大学 计算机科学与技术系,北京 100871)

摘要:在最近召开的“全国搜索引擎与网上信息挖掘学术研讨会”上,举办了一场“中文网页自动分类竞赛”,共有来自全国各地的 10 个队参加。本文在介绍本次竞赛活动规则和过程的基础上,详细分析了竞赛的结果,从而使我们对于目前中文网页自动分类技术的现状有了一种具体的认识:目前已有分类器的性能没有呈现出明显的差距,中文网页的分类比普通文本的分类要困难的多。同时,本文还尝试推出一个标准的中文网页分类的实例样本集,希望通过不断完善,最终作为中文网页分类技术研究的基本语料。

关键词:计算机应用;中文信息处理;机器学习;中文网页自动分类;TREC 评测

中图分类号:TP391 **文献标识码**:A

The Analysis of a Contest Result on Chinese Web Page Automatic Categorization

FENG Shi-cong, Wang Ji-min

(Department of Computer Science and Technology, Peking University, Beijing 100871, China)

Abstract: A Chinese Web page automatic categorization contest was hold in national symposium on Search Engine and Web Mining and ten teams took part in this contest. After describing the contest rules, this paper analyses the contest results in details and we can have an explicit view on the present technologies of Chinese Web page automatic categorization: no explicit difference is shown among those classifiers had been developed and Chinese Web page categorization is more difficult than plain text categorization. This paper also attempt to provide a standard Chinese Web page categorization instance examples and develops them to be a base corpus of Chinese Web page categorization by continuous modification.

Key words: computer application; Chinese information processing; machine learning; Chinese Web page automatic categorization; TREC evaluation

一、引言

随着 Internet 的飞速发展,Web 上的网页数据量呈指数增长[1,2]。海量的网页在为人们提供包罗万象、无比丰富的信息资源的同时,也向人们提出了如何快速从中获取其所需信息的挑战。为了能够有效地组织和处理海量的 Web 信息,人们希望能够按照其内容实现对网页的自动分类。在 Web 出现之前,人们研究过许多普通文本分类的方法[3],例如支持向量机(Support Vector Machines,SVM)、k-Nearest Neighbor(kNN)、简单贝叶斯(Naïve Bayes, NB)、Linear Least Squares Fits (LLSF)、Neural Network (NNet)等等,形成了各种文本自动分类(Auto-

① **收稿日期**:2003 - 03 - 27
基金项目:国家 973 重大基础研究项目资助(G1999032706)
作者简介:冯是聪(1974 —),男,博士研究生,主要研究方向为个性化搜索引擎.

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

matic Text Classification, ATC)技术。所谓文本自动分类就是指定文本和预先定义类之间的类属关系,分类的工作由计算机自动完成[4]。随着 Web 技术日益得到广泛的应用,海量网页信息的涌现,ATC 技术的处理对象从普通文本扩展到网页信息,自然地,ATC 技术成了实现网页自动分类的基础。目前,网页自动分类技术在数字图书馆、主题搜索、个性化信息检索、搜索引擎的目录导航、信息过滤、主动信息推送服务等领域得到了广泛地应用。网页自动分类技术已经成为 Web 领域的研究热点。

为了评价各种文本自动分类算法的优劣,推进信息检索领域的发展,由美国国家标准和技术研究院(NIST)、信息技术实验室(ITL)检索小组、美国国防部高级研究计划署(DARPA)信息技术处、高级研究开发机构(ARDA)等单位共同主持召开了信息检索会议 TREC[5],至今已召开了11 次(从 1992 年起,每年一次,今年将召开第 12 次会议)。TREC 会议实际上是信息检索系统的擂台赛,可以说,在 TREC 上展示的文本分类系统代表了文本分类领域的最新研究成果。一些大学,如 CMU、BERKLEY、CORNELL 等和一些公司带着自己开发的文本分类系统参加会议,由大会使用相同的训练集和测试集对这些系统进行评测。中国科学院计算所,清华大学等单位近几年也有派队参加,并取得了不错的成绩。我们注意到,由于 Web 技术的发展,TREC 也逐步开始提供标准的英文网页语料[6]来评测 Web 信息检索系统。

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,到目前为止,标准的中文网页语料库还没有出现,因此也没有出现针对中文网页分类系统的评测。近年来,国内外的研究人员针对中文网页已经做了大量有意义的研究工作[7,8]。但是,由于没有标准的中文网页语料,无法比较不同分类系统性能的优劣。因此,为了促进中文网页自动分类技术研究的不断发展,加强研究人员之间的交流,2003 年 3 月 14 - 15 日在北京大学举办的"全国搜索引擎与网上信息挖掘学术研讨会"组织了一次"中文网页自动分类竞赛"活动。希望通过本次活动来评测已有中文网页分类器的质量和性能,并且尝试推出一个标准的中文网页分类的数据集,通过不断完善,最后作为中文网页分类技术研究的基本语料,从而对中文网页自动分类的研究起到推进作用。

这次"中文网页自动分类竞赛"活动受到了国内外研究人员的热烈响应。先后有 16 个队报名参赛,最后共有包括北大、清华、上海交大等 10 个代表队参加了现场测试。

本文的随后部分是这样安排的:第 2 节将详细介绍本次"中文网页自动分类竞赛"活动的竞赛规则;第 3 节将给出本次竞赛的结果并简要地进行了分析;第 4 节对全文进行了总结,并指明了本次活动的不足和将来的改进方向。

二、竞赛规则

为了保证本次"中文网页自动分类竞赛"的公开、公平和公正,以及测试结果的客观真实性,我们在 2003 年 2 月 5 日公布了本次活动的竞赛规则。下面介绍有关要点。

2.1 训练集和测试集

中文网页训练样本集是实现中文网页自动分类的前提和基础。但是到目前为止还没有出现标准的中文网页语料库。为此,我们通过动员不同专业的几十个学生,人工选取形成了一个中文网页训练样本集①。该样本集共包括 13973 个网页实例,其中训练网页为 11059 个,测试

① 可以免费提供给有兴趣做类似研究工作的同行,燕穹产品号:YQ - WPBENCH - V1.0

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

表 1　训练集和测试集的分布情况

类别编号	类别名称	训练集	测试集
01	人文与艺术	398	101
02	新闻与媒体	124	19
03	商业与经济	818	211
04	娱乐与休闲	1482	369
05	政府与政治	285	83
06	社会与文化	1070	292
07	教育	284	82
08	自然科学	1807	473
09	社会科学	1708	463
10	计算机与因特网	831	217
11	医疗与健康	2252	604
合计		11059	2914

网页为 2914 个,大小约 200MB。这些网页被划分为 11 种类型,如表 1 所示。其中,"训练网页"和"测试网页"的区分主要是为了方便参赛队预先测试各自的分类器,随机挑选的。同时我们也声明了参赛队也可以用自己的训练、测试集。后来从各参赛队谈的情况看,都用的是我们提供的这个样本集。

竞赛中使用的测试网页集则由所有参赛队和其他参会人员在现场共同收集整理的。按照竞赛规定,可以(但不必须)为每个测试网页确定两个类别(一个主类,一个兼类)。

2.2　结果评价标准

在信息检索领域,评测分类器质量的指标通常为查准率和查全率,或者是综合考察查准率和查全率的 F1 值[3]。这些指标通常用来评价一个文本有一个结果类别的情况。但是事实上,许多网页的类别边界比较模糊,通常可以属于多个不同的类别,一般来讲,为这些网页取多个类别比仅取一个类别要合理的多,应用上述的评价指标来处理一个网页有多个结果类别的情况是不合理的。因此,我们应用公式(1)来评价中文网页分类器的质量。为简单起见,本次竞赛规定每个给定的测试网页最多取两个结果类别。考虑到不同的分类系统使用不同的系统平台,无法得到客观公正的测试时间,因此,分类器的效率问题不是本次竞赛考察的重点,仅规定平均 1 个网页的分类时间不得超过 3 秒。

$$\text{分类器得分 } S = \sum_{i=1}^{n} \text{对每个网页的分类得分 } Si \qquad (1)$$

其中:n 为测试网页数;S_i 为第 i 个网页的分类得分,其取值根据由表 2 得到。

表 2　不同情况下 S_i 的得分

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,位置同原来的不同	0.5
个数和次序都不对	0

表 2 显示,决定 Si 取值的因素有两个:(1)分类器给出的结果类别的个数:1 个或 2 个,但是不能为 0 个,也就是说分类器必须对一个网页给出结果,不能不做任何判断。(2)结果类别的次序。

例如:某个网页 p 的类别为 01,08 类,不同分类结果下的得分分别为:

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
(6)02	得分为 0
(7)02,07	得分为 0

例如:某个网页 p 的类别为 01 类,不同分类结果下的得分分别为:

分类结果　　得分

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

(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

2.3　现场测试

为了保证测试结果的客观性和公平性,整个测试过程由一位所有参赛队以外的专家主持,最终的竞赛结果和中间结果都保存在他的机器上。并且规定,从现场收集网页开始,所有参赛人员不能再操作机器,测试集的导入和分类结果的导出均由主持人完成。整个竞赛过程都是"透明的":所有参会人员都能够通过投影仪看到分类过程的执行情况。测试集是在主持人的主持下,由每个参赛队出一个代表,现场搜集整理,按照多数原则,共同决定测试网页的类别。经过现场 3 个小时的紧张整理,我们最后共得到 44 个测试网页,每个类别有 4 个网页。44 个网页中只有一个被确定为有兼类。

三、竞赛结果及其分析

最终参加本次"中文网页自动分类竞赛"的共有 10 支代表队,表 3 是按表 2 给出的评分规则,参赛队得分的情况(为便于叙述,这里参赛队按得分高低顺序编号)

表 3　本次"中文网页自动分类竞赛"的结果

参赛队	1	2	3	4	5	6	7	8	9	10
得分	30.9	30.7	30.0	29.5	29.0	28.5	27.1	24.7	22.0	21.0

从表 3 可以看出前 6 名的得分非常接近,分类器的分类查准率相差不大,没有呈现出明显的差距,相继两名之间的差距相差不到 1 分。也就是说,按照表 2 的评分规则,前一名完全正确分类的网页个数比后一名完全正确分类的网页个数不超过 1 个。

为了考察每个参赛分类器的分类质量,我们分别统计分析了所有参赛队的分类结果。根据文献[3]中查准率和查全率的定义,我们分别把被分到每个类的网页数以及其中包含的分类正确的网页个数统计出来,并计算相应的查准率和查全率①,最后得到结果如表 4 所示。其中横向从 1 到 10 的编号代表上述参赛队的编号,纵向从 01 到 11 的编号表示表 1 中 11 个类别的编号,中间的每个单元格表示不同参赛队的分类器在不同类别下的查准率和查全率,符号"/"前后的两个数值表示其查准率和查全率。这里需要特别说明的是,第 3 队有 1 个测试网页,第 7 队有 11 个测试网页没有给出结果类别。根据查准率和查全率的定义:

每个类的查全率 = 正确分到该类的网页个数 / 该类测试网页的总数　　(2)

每个类的查准率 = 正确分到该类的网页个数 / 分到该类的网页总数　　(3)

其中该类测试网页的总数为 4。这里的"正确分到该类的网页个数"没有异议,但是"分到该类的网页总数"就无法统计了,因为他们没有给出这 12 个测试网页没有给出结果类别。为了在表 4 中能够统一表示,在这里我们把"查准率"的定义做了修正:

① 当某个网页有两个结果类别时,根据查准率和查全率的定义,这种情况下的查准率和查全率计算比较困难,因此,为简单起见,我们仅比较第一个结果类别是否正确,只要第一个结果正确,就算该网页分类正确。当第二个类别正确,而第一个类别不正确时,尽管根据表 2,此时分类器可以得分,但是在这里不算该网页分类正确。因此,表 4 计算得到的 10 个队的分类器的查准率同 10 个队最后的得分不是正比关系:这里的查准率最大并不意味着该队的得分最多,因为每个队的得分是根据表 2 计算得到的。

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

查准率 = 正确分到该类的网页个数 / (分到该类的网页总数 + 没有给出结果的网页数/总类别数) (4)

因此,对于第 3 队,查准率 = 正确分到该类的网页个数 / (分到该类的网页总数 + 1/11);对于第 7 队,查准率 = 正确分到该类的网页个数 / (分到该类的网页总数 + 11/11)。

每个参赛队分类器的平均查准率和查全率是分别根据公式(5)和公式(6)计算得到的,也就是把 11 个类的查准率和查全率求和然后再取平均。根据文献[3],这里求得的查准率和查全率是分类器的 macro-p 值和 macro-r 值。

分类器的平均查准率 = $\sum_{i=1}^{11} P_i$ (5)

其中 P_i 为第 i 个类别的查准率。

分类器的平均查全率 = $\sum_{i=1}^{11} R_i$ 正确分类的网页个数 / 被分类的网页总数 (6)

其中 R_i 为第 i 个类别的查全率。

表 4 不同参赛队的分类器对 11 个不同类别的查准率和查全率

队编号 / 类别编号	1	2	3	4	5	6	7	8	9	10
01	.75/.75	.67/.50	.97/.75	1/.75	1/.50	1/.75	.50/.50	1/.75	1/.50	1/.50
02	.44/1	.50/.50	0/0	1/.25	0/0	0/0	.33/.25	.24/1	0/0	0/0
03	1/.75	0/0	.73/.75	0/0	0/0	0/0	.92/.25	0/0	0/0	0/0
04	0/0	.50/1	.79/1	.57/1	.33/1	.36/1	.50/.50	0/0	.50/.75	0/0
05	1/.25	1/.25	.48/.25	.50/.25	.33/.25	1/.25	.50/.25	1/.25	1/.25	0/0
06	.67/1	.75/.75	.58/.75	1/.50	.75/.75	.80/1	.80/1	1/.50	.67/.50	.67/.50
07	1/.75	1/1	.92/.25	1/.50	1/.75	1/.50	.67/.50	1/1	0/0	1/.25
08	1/1	.80/1	.79/1	.80/1	1/1	.80/1	.80/1	.57/1	1/.75	1/1
09	.60/.75	.50/.75	.42/.75	.29/1	.67/1	.33/.75	.57/1	0/0	.21/1	.50/1
10	1.0/1.0	1/1	.79/1	1/1	1/1	1/1	.75/.75	1/.75	1/.75	.33/1
11	1.0/1.0	.80/1	.66/1	1/1	1/1	1/1	.80/1	1/1	.80/1	.80/1
平均	.77/.75	.68/.70	.65/.68	.74/.66	.64/.66	.66/.66	.65/.64	.62/.57	.56/.50	.48/.48

从表 4 可以看出,这 10 个分类器的分类查准率和查全率都不高。最好的查准率为 77 %,最好的查全率也才为 75 %;而最差的查准率仅为 48 %,最差的查全率为 48 %。这与通常的普通中文文本分类结果相比有相当差距。这也从一定的程度上说明了中文网页的分类比普通文本的分类要困难的多。这主要是因为,同普通文本相比,网页的设计比较随意,通常包含大量的广告、设计人员的注释以及版权声明等无关信息,这些“噪音”降低了分类的查准率。

从表 4 可以看出参赛各队的分类器质量。为了考察所有参赛分类器对不同类别的分类质量,我们统计分析了所有参赛队的分类结果。根据查准率的定义,我们分别把被分到每个类的网页数以及其中包含的分类正确的网页个数(同表 4 不同,这里只有两个结果都正确,才算该网页分类正确)统计出来,并计算相应的查准率和查全率,结果如表 5 所示,其中“被分到某类的网页数”的总和为 428,小于 440 个(440 是这样计算得到的:11 个类别,每个类别有 4 个网页,共 10 个参赛队,所以为 11 ×4 ×10 = 440),这是因为第 3 队有 1 个网页,第 7 队有 11 个网页没有给出结果类别,所以没有统计在内。

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

表 5　分类器针对不同类别的分类查准率

编号	类别名称	被分到该类的网页数	其中正确的网页数	查准率	查全率
01	人文与艺术	28	22	.79	.50
02	新闻与媒体	35	9	.26	.20
03	商业与经济	12	5	.42	.11
04	娱乐与休闲	61	23	.38	.52
05	政府与政治	13	7	.54	.16
06	社会与文化	38	27	.71	.61
07	教育	22	21	.95	.48
08	自然科学	46	39	.85	.89
09	社会科学	82	29	.35	.66
10	计算机与因特网	46	35	.76	.80
11	医疗与健康	45	38	.84	.86

图 1 为表 5 中每个类别查准率的柱状图表示(其中横轴从 1 到 11 的数字表示从 01 到 11 的类别编号,纵轴表示查准率的取值)。

从图 1 可以清楚地看出,“新闻与媒体”、“社会科学”、“商业与经济”等类别的查准率较低,而“自然科学”、“医疗与健康”等类别的查准率较高。分析其原因,这主要是因为所有参赛分类器都使用的是基于统计学习的“向量空间模型”,无法正确描述特征项之间的语义信息。对于“自然科学”、“医疗与健康”等类别,其包含的特征项比较特别,其它类别中很少出现,因此,分类的查准率比较高。而对于“新闻与媒体”、“社会科学”、“商业与经济”等类别,由于其包含的特征项在其它类别中也会同时出现,因此,分类的查准率比较低。

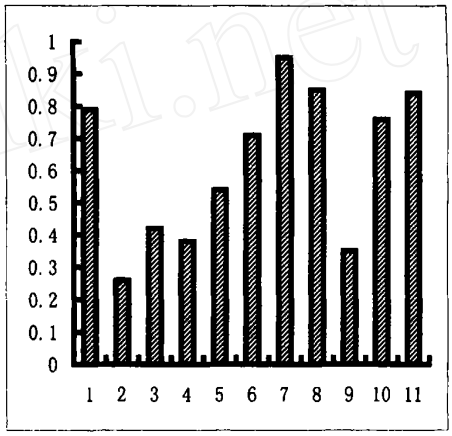


图 1　分类器在不同类别上的分类查准率

下面比较本文图 1 和文献[9]中图 1 的结果(尽管文献[9]使用 Macro-F_1 值作为衡量分类质量的指标,但是在一定程度上可以同查准率进行比较)。虽然在本次测试中仅使用了有限的 44 个测试网页,远远小于文献[9]中使用的 2800 个,但是两者的结论却基本一致:分类器在“自然科学”、“医疗与健康”等类别上有较高的分类质量;在“新闻与媒体”、“社会科学”等类别上的分类效果不佳。因此,我们可以得到一个很有意义的结论:使用少量有限的测试网页就能够说明分类器的质量。这也说明了本次竞赛的测试结果是可靠的,随机性不大。

四、总结及进一步的工作方向

为了促进中文网页自动分类技术研究的不断发展,加强研究人员之间的学术交流,在北京大学举办的“全国搜索引擎与网上信息挖掘学术研讨会”期间组织了一次“中文网页自动分类竞赛”活动。希望通过本次活动来评测已有中文网页分类器的质量和性能,并且尝试推出一个标准的中文网页分类的实例样本集,通过不断完善,最后作为中文网页分类技术研究的基本语料,从而对中文网页自动分类的研究起到推进作用。本文还详细介绍的本次竞赛活动的规则以及最终的竞赛结果。经过分析可知,尽管本次竞赛活动使用了有限的测试网页样本,但是得

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

到的结果同大规模测试的结果基本一致,因此可以说明本次的测试结果是可靠的,随机性不大。下一步的主要工作包括:(1)对本次竞赛使用的中文网页语料库进行全面系统的修订,使其更加科学和规范;(2)对现有的中文网页分类器进行全面系统地评测,找出最适合中文网页的分类算法和策略;(3)在考察分类器的分类质量的同时,还要考察分类器的分类效率。

参 考 文 献:

[1] 李晓明.对中国曾有过静态网页数的一种估计[J].北京大学学报(自然科学版). 2003,39(3).

[2] 阎宏飞,李晓明.关于中国 Web 的大小、形状和结构[J].计算机研究与发展.(Grid 特刊)2002.8.

[3] Yiming Yang, Xin Liu. A re-examination of text categorization methods [C]. Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99, pp 42 - 49), 1999

[4] Sebastinai Fabrizio. A tutorial on Automated text categorization. http://faure.iei.pi.cnr.it/~fabrizio/ATCtutorialprogram.html [Z].

[5] http://trec.nist.gov/ [Z].

[6] http://www.ted.cmis.csiro.au/TRECWeb/access-to-data.html [Z].

[7] LI Xiaoli and SHI Zhongzhi. Innovating Web Page Classification Through Reducing Noise [J]. In Journal of Computer Science & Technology,2002,17(1).

[8] 李蓉,叶世伟,史忠植. SVM-KNN 分类器 —一种提高 SVM 分类精度的新方法[J].电子学报. 2002, 30(5).

[9] 龚笔宏,冯是聪. kNN 与 NB 算法在中文网页分类上的比较研究[C].搜索引擎与 Web 挖掘进展, 2003:73 - 79.

(上接第 18 页)

参 考 文 献:

[1] Ronald Rosenfeld, Two decades of statistical language modeling: Where do we go from here? Proceedings of the IEEE [C], 88(8), 2000.

[2] DeMori, R., and M. Federico, Language Model Adaptation, [A]. In Computational Models of Speech Pattern Processing, Keith Pointing (ed.), NATO ASI Series, Springer Verlag, 1999.

[3] R. Kuhn and R. D. Mori, A cache-based natural language model for speech reproduction [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAM2 - 12(6):570 - 583, 1990.

[4] Daniel Gildea and Thomas Hofmann, Topic-based language models using EM. In Proceedings of the 6th European Conference on Speech Communication and Technology(EUROPEANSPEECH)[C], 1999.

[5] G. Salton, Automatic text processing: The transformation, Analysis, and Retrieval of Information by Computer [M], Addison-Wesley 1989.

[6] P. Clarkson and A. Robinson, Language model adaption using mixture and an exponentially decaying cache[C]. In Boc. ICASSP - 97, 1997.

[7] A. P. Dempster, N. M. Laivd, and D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm [J]. Journal of the Royal Statistical Society B,1977,39:1 - 38.

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net