针对中文文本自动分类算法的评估体系*)

徐 威[1] 董 渊[1] 白若鹞[1] 张素琴[1]
(清华大学计算机科学与技术系 北京 100084)[1]

摘 要 中文文本自动分类能够帮助人们更有效地利用不断膨胀的海量中文信息。现有中文文本自动分类算法基于不同原理,性能各异,适用于不同情况。对于分类算法的比较评估能够确定某个分类算法的适用环境和性能特征。目前缺乏针对中文文本自动分类算法的系统评估体系。本文将引入一个评估体系,并基于该体系实现一个开放的研究平台,得出若干已有中文文本自动分类算法的比较结果。

关键词 文本分类,评估体系,中文分词,特征选择,语料库

An Evaluation System for Algorithms of Automated Text Categorization on Chinese

XU Wei[1] DONG Yuan[1] BAI Rou-Yao[1] ZHANG Su-Qing[1]
(Department of Computer Science and Technology, Tsinghua Univ., Beijing 100084)

Abstract Automated text categorization on Chinese helps people make more effective use of the growing Chinese information on the Internet. Current algorithms have different performance according to the environment. Evaluation to the algorithms tells the feature of certain algorithm. At present, there is no evaluation system form algorithms of automated text categorization on Chinese. This paper introduces such a evaluating system. An open research platform based on it is also introduced, with results of some algorithms.

Keywords Text categorization, Evaluation system, Chinese word division, Feature selection, Corpus

1 引言

随着计算机和互联网技术的不断发展,对海量信息,特别是非结构信息的检索、过滤、管理成为一个突出的问题。自动分类的出现很好地解决了对信息的分类,使得人们能对信息各取所需。其中最主要的中文文本自动分类已经有了40多年的历史,基于英文的分类算法、评估标准以及语料库已经有了一套成熟的体系,而基于中文的相关方面初具规模。

但是中文文本自动分类算法受分词、特征选取、权值调整等因素的影响比较大,而具体的准确率等指标又和应用领域、训练文本和测试文本甚至实验环境有关,因此需要一个统一的平台和统一的语料库来比较各种算法的优劣。

英文方面 Yiming Yang 等人在同一个平台实现了各种算法,并对各算法的表现作了比较[1]。但是在中文文本自动分类方面,目前还没有人利用统一的平台以及一个合适的语料库对各分类算法进行比较。此外,中文文本和英文文本的分类有非常大的差别[2]。

本文的目的就在于实现一个开放的中文文本自动分类平台,为对中文文本自动分类进行的研究提供一个实验和评估的平台,使人们能在此平台上实现、研究和比较各类算法。

2 中文文本自动分类

中文文本自动分类是自然语言处理的一个重要应用领域。由自然语言构成的文本并不能被计算机直接处理,需要对其进行数学抽象,建立模型后进行自动分类的处理。

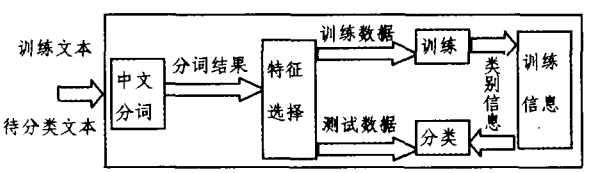

图1 中文文本自动分类过程

图1是中文文本自动分类的一般过程,文本分类器由训练文本训练产生后,才能对测试文本进行有效的分类。训练文本或者测试文本在训练或者测试过程中一般都要经过分词、特征选择、训练分类三个步骤。

2.1 分词算法

中文文本与英文文本不同,中文字符流中各词汇之间没有固有的分隔符,进行中文文档的词频统计前,需要对中文文本进行分词。根据是否利用机器可读词典和统计信息,可将当前主要的分词方法分为两大类:基于词典的方法、基于统计的方法[3]。

目前中文分词算法主要存在的困难在于切分歧义消解、未登录词语处理和语言资源建设[4]。本文考虑的分词问题是直接为中文文本自动分类服务的,因此对分词算法的效果评估将集中在对分类结果的评价。

2.2 特征提取

将文本集合里所有词都作为特征词汇,不仅会有一些噪声词汇产生影响,而且文本空间维数过高会导致系统开销过大,因此需要特征选择来抽取一些最能代表类别特征的词语构成文本的特征向量。文[5]认为利用类别信息的特征选取

*)基金项目:国家自然科学基金(No.60573017)。**徐 威** 硕士生,主要研究方向:软件工程,文本自动分类;**董 渊** 讲师,博士,研究方向:系统软件,软件工程。**白若鹞** 硕士生,主要研究方向:文本检索;**张素琴** 教授,研究方向:程序设计语言设计与实现,编译优化。

在不加以修正的情况下并不适合中文文本分类，并提出了可能的矫正措施，包括增大训练语料的规模和采用组合的特征抽取方法。

特征选择的方法很多[6]，如 TFIDF 方法、x^2 统计量（x^2 Statistic）、文本证据权（Weight of Evidence for Text）等等[7]。利用多种特征函数综合选取特征词可以达到较好分类效果。

2.3 文本自动分类算法

目前的分类算法有许多种类，如基于概率论的分类算法（naïve bayes 算法、bayes 神经网络算法等）、基于决策树的分类算法、基于决策规则的分类算法、基于回归模型的分类算法（LLSF 算法：Linear Least Squares Fit）、基于在线线性分析的分类算法（基于原形的算法、批处理归纳方法 batch induction、在线归纳方法 On-line induction）、Rocchio 分类算法、基于神经网络的分类算法、基于样本的分类算法（k-NN 算法）、联合分类算法（boosting 算法）[8]。和特征选择一样，采用多个分类算法联合进行分类同样可以达到较好的分类效果[9]。

3 分类研究评估体系

目前中文文本自动分类并没有一个标准的平台和评估体系，从上一节分类的三个方面可以看出中文文本自动分类与英文文本分类最大的不同在于分词算法的影响，但是对其他两方面的考察也不能忽略。因此，我们希望利用已知的分类知识去构建一个分类评估体系，并给出针对中文文本自动分类算法的评估基准。

3.1 分类效果评估体系

文本分类从根本上说是一个映射过程，所以评估文本分类系统的标志是映射的准确程度和映射的速度。映射的速度取决于映射规则的复杂程度，而评估映射准确程度的参照物是通过专家思考判断后对文本的分类结果（这里假设人工分类完全正确并且排除个人思维差异的因素），与人工分类结果越相近，分类的准确程度就越高，这里隐含了评估文本分类系统的两个指标：查准率和查全率。

查准率 precision 是所有判断的文本中与人工分类结果吻合的文本所占的比率：

$$\text{precision}=\frac{n_{correct}}{n_{classified}}$$

其中 $n_{correct}$ 表示分类正确的文本数，$n_{classified}$ 表示实际分类的文本数。

查全率 recall 是人工分类结果应有的文本中分类系统吻合的文本所占的比率：

$$\text{recall}=\frac{n_{correct}}{n_{belongs}}$$

其中 $n_{correct}$ 表示分类正确的文本数，$n_{classified}$ 表示实际分类的文本数。

查准率和查全率反映了分类质量的两个不同方面，两者必须综合考虑即 F1 测试值：

$$F_1=\frac{\text{precision}\times\text{recall}\times 2}{\text{precision}+\text{recall}}$$

对于每一个类上述参数的计算叫做微平均，而对全部类的参数计算叫做宏平均。

在实际的算法评测中需要将一些算法作为基准，以建立一个合理的评估体系，方便评估研究结构。每个类别 C_i 的微平均 $F1$ 测试值 $MicroF1_i$ 考查了每个类别的查准率 precision_i 与查全率 recall_i，是对分类结果比较精确的判断。但分类各部分算法的比较不可能对每个类别 $MacroF1$ 进行一一比较。采用宏平均 F1 测试值 $MacroF1$ 则考查了整个测试结果的查准率 precision 与查全率 recall，但是并没有考查每个类别对于整个测试结果的贡献，因此在对微平均 $F1$ 测试值的基础上对每个类别进行加权，总和是一个比较均衡的考查方式：

$$\text{score}=\sum_i w_i * MicroF1_i$$

其中 w_i 为类别 i 文档数在全类别中占的比重。

利用上述公式计算的 score 值可以明显地考查整个测试文本集上的分类效果，此外还可以建立基准算法对各种算法进行评估，目前的基准如下：

1. 分词：以指定字典的正向最大分词算法作为分词算法标准；
2. 特征选择：不进行特征选择；
3. 分类：以 Naïve Bayes 算法[10]作为分类算法标准；
4. 语料库：复旦大学提供的测试语料库。

实验结果 score_{base} 为 82.11%，以此作为基准准确率，若其他算法结果 score_{test}，则得分如下：

$$point_{test}=10^{2\times\frac{score_{test}}{score_{base}}}$$

3.2 评估体系实现

评估体系的实现利用了模块化的设计思想：分类过程中的三个步骤被设计成独立模块，对某一方面的研究直接实现该模块的接口，加入平台进行实验，避免了重复实现其他模块。而每个模块与输入输出模块整合，可以很方便地投入到应用中去。

以下是该系统的模块设计图：

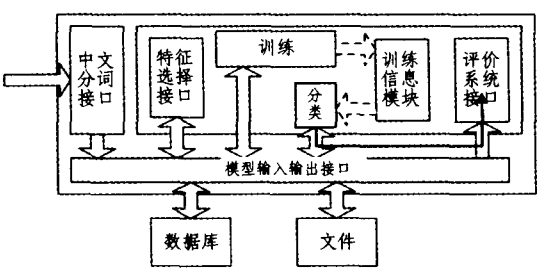

图 2 分类平台模块设计图

由图 2 可见，这个平台是基于图 1 的流程设计的。平台为所有模块建立接口，并实现一部分基准算法，能完成一个完整的中文文本自动分类过程。

分词模块提供了分词接口以便于利用已有字典与分词算法将文本输入转化为词频向量，建立文本模型，进行训练和分类，从而考查不同分词算法对中文文本自动分类效果的影响。分词模块的输出将会被存储在数据库或是文件系统中。

利用分词接口建立了文本空间模型后，就可以通过（或者不通过）实现的特征选择接口进行特征选择产生新的文本空间模型，进行文本分类，从而确定特征选择算法参数的影响。

分类算法包括了训练模块和分类模块，同时训练信息同样可以输出到文件或数据库中，为下次分类准备训练内容。分类模块的结果一方面可以直接输出到评估系统接口，另一方面可以存入文件或数据库以备事后分析。

该平台的任一接口都能与输入输出接口结合而单独应用于商业应用，因此，研究不同模块算法只需实现不同模块的接口，其他部分则可利用已实现的部分。

4 实例测试

以下将以一个简单的例子来说明如何利用该平台。

本实例是用来比较文本证据权 WET 和 χ^2 统计量两种特征选择算法在特定环境下的效果。在分词接口实现正向匹配算法，在特征选择接口实现两种方案，在分类模块中选择 Naïve Bayes 算法，利用宏平均精度来展示分类精度，预料库利用复旦大学提供的测试语料库，采用 10-fold cross validation 的方法。

表 1 分类测试方法比较

分词算法	特征选择算法	文本分类算法
正向最大匹配法	文本证据权 WET	Naïve Bayes
	χ^2 统计量	

下表是比较结果。

表 2 分类测试结果比较

特征选择算法	score	point
χ^2 统计量 选取 40%特征	79.87%	88.19
WET 选取 40%特征	82.27%	100.90
WET 选取 10%特征	75.15%	67.68

结果显示，在选择相同数目特征的情况下，文本证据权 WET 算法的特征选择算法相对于 χ^2 统计量算法表现了较好的分类精度。同时，文本证据权 WET 算法在选取特征比较少的情况下，分类结果下降较明显。

总结与展望 本文介绍了一个中文文本自动分类评估体系，用于对中文文本自动分类过程中使用到的分词算法、特征选择算法以及分类算法进行综合评价，并实现了一个中文文本自动分类研究的开放平台，供人们实现和比较各种算法。利用该平台，得出了若干已有中文文本自动分类算法的实验结果，评价了不同特征选择算法和不同参数对分类的影响。

目前中文方面并没有一个公开的、相对标准的语料库，所以本评估体系的下一步方向就是利用自动下载工具对权威网站进行定期下载，建立一个标准的语料库，并实现语料库的动态更新。

参 考 文 献

1 Yang Yiming, Liu Xin. A Re-Examination of Text Categorization Methods. In: 22nd Annual International SIGIR. 1999. 42～49
2 刘延章，余义芳. 近五年来网络信息分类组织研究的现状及其展望. 情报学报，2004，23(2)
3 张春霞，郝天勇. 汉语自动分词的研究现状与困难. 系统方针学报，2005(1)
4 孙茂松，邹嘉彦. 汉语自动分词研究评述. 当代语言学，2001(1)：22～32
5 代六玲，黄河燕，陈肇雄. 中文文本分类中特征抽取方法的比较研究. 中文信息学报，2004，18(1)：26～32
6 Wang Yi, Wang Xiao-Jing. A new approach to feature selection in text classification, Machine Learning and Cybernetics, 2005. In: Proceedings of 2005 International Conference on, Aug. 2005, 6：3814～3819
7 李凡，鲁明羽，陆玉昌. 关于文本特征抽取新方法的研究. 清华大学学报(自然科学版)，2001，41(7)：98～101
8 Sebastiani F. Machine Learning in Automated Text Categorization. In: 18th International Conference on Computational Linguistics (COLING'00), Nancy, France, July 2000
9 Jain G, Ginwala A, Aslandogan Y A. An approach to text classification using dimensionality reduction and combination of classifiers, Information Reuse and Integration, 2004. IRI 2004. In: Proceedings of the 2004 IEEE International Conference on, Nov. 2004. 564～569
10 Wang Baoyi, Zhang Shaomin. A Novel Text Classification Algorithm Based on Naïve Bayes and KL-Divergence, Parallel and Distributed Computing, Applications and Technologies, 2005. PDCAT 2005. In: Sixth International Conference on, Dec. 2005. 913～915

(上接第 129 页)

任务的异构性比较低时，CBU-Min-min 算法的性能比 CBU-Sufferage 算法的性能要好(这一点与传统方法认为 Sufferage 算法总是比 Min-min 算法好的结论不一样)，但差距不是很大；从图 2 和图 4 可以看出，当任务的异构性较高时，CBU-Min-min 算法的性能没有 CBU-Sufferage 算法的性能好，且二者的性能差异较大。

总结 由于网格的广域、异构和动态特性，网格任务调度存在随机、模糊、不确知、中介和突发等多种不确定因素，任何单纯的不确定性方法都难于真实描述和表达这种综合不确定性问题。本文利用集对分析联系数来研究和处理网格调度的这种综合不确定性问题，并借鉴传统网格任务静态调度算法提出了相应的不确定网格静态调度新算法。研究结果表明，这些算法能较好地描述网格任务预期执行时间的动态性和不确定性，不仅在动态和不确定网格环境中有良好的理论和实际应用价值，还将成为网格任务调度理论建模的一个新的研究方法。

参 考 文 献

1 Foster I, Kesselman C 著. 金海，袁平鹏，石柯译. 网格计算[M](第 2 版). 北京：电子工业出版社，2004
2 Foster I, Kesselman C. The Anatomy of the Grid: Enabling Scalable Virtual Organizations[J]. International Journal of Supercomputer Applications, 2001, 15(3): 200～222
3 罗红，蔡德俊，邓智群，王晓东. 网格计算中任务调度研究综述[J]. 计算机应用研究，2005(5)：16～19
4 Brauny T D, Siegely H J, Becky N, et al. A Comparison Study of Static Mapping Heuristics for a Class of Meta-tasks on Heterogeneous Computing Systems[C]. In: Proceedings of the 8th Heterogeneous Computing Workshop (HCW'99), Apr. 1999. 15～29
5 陈志刚，刘安丰，熊策，张连明. 一种有效负载均衡的网格 Web 服务体系结构模型[J]. 计算机学报，2005，28(4)：458～466
6 刘安丰，陈志刚，陆静波，张连明. 网格环境中一种有效的 Web 服务资源组织机制[J]. 计算机研究与发展，2004，41(12)：2141～2147
7 张伟哲，刘欣然，云晓春，等. 信任驱动网格作业调度算法[J]. 通信学报，2006，27(2)：73～79
8 Schopf J M, Nitzberg B. Grids: Top Ten Questions. Scientific Programming, special issue on Grid Computing, 2002, 10(2): 103～111
9 李季，钟将，吴中福. 具有模糊处理时间的网格任务调度免疫算法[J]. 计算机科学，2006，33(2)：35～38
10 赵克勤. 集对分析及其初步应用[M]. 杭州：浙江科学技术版社，2000
11 黄德才，赵克勤. 用联系数描述和处理网络计划中的不确定性[J]. 系统工程学报，1999，14(2)：112～117
12 薛根元，王国强. 不确定性理论集对分析在预报模型建立中的应用研究[J]. 气象学报，2003，61(5)：592～599
13 黄兵，周献中. 不完备信息系统中基于联系度的粗集模型拓展[J]. 系统工程理论与实践，2004，24(1)：88～72
14 李志辉，夏少云，查建中. 基于案例推理的同异反产品设计及其应用[J]. 计算机辅助设计与图形学学报，2003，15(11)：1397～1403
15 蒋云良，徐从富. 集对分析理论及其应用研究进展[J]. 计算机科学，2006，33(1)：205～209
16 Ali S, Siegel H J, Maheswaran M, Hensgen D, Ali S. Representing Task and Machine Heterogeneities for Heterogeneous Computing Systems[J]. Tamkang Journal of Science and Engr, 2000, 3(3): 195～207