结合信息检索技术的半监督文本分类方法

贾志洋[1],高 炜[2,3],王勇刚[1]

(1. 云南大学旅游文化学院,云南 丽江 674100; 2. 苏州大学数学科学学院,江苏 苏州 215006;
3. 云南师范大学信息学院,云南 昆明 650092)

摘 要: 搜索引擎的查询结果和查询关键词与某一个文本类别应该具有一定关联. 基于这样的假设,针对文本分类问题,根据小样本集提取特征词构建查询并从查询结果中下载网页样本,将下载的网页样本进行去重、去噪、提取正文等处理后,判断其类别并扩充到初始样本集,最终使用扩充后的实验样本集学习训练朴素贝叶斯文本分类器,并对分类器的分类效果进行了测试. 实验结果表明,结合信息检索技术的半监督分类器的分类准确率相对于使用小样本构建的分类器具有较大的提高.

关键词: 文本分类; 半监督学习; 信息检索; 搜索引擎

中图分类号: TP393.04 **文献标识码:** A **文章编号:** 1000 - 2073(2011)04 - 0034 - 06

Semi-supervised text classification with information retrieval techniques

Jia Zhiyang[1], Gao Wei[2,3], Wang Yonggang[1]

(1. Tourism and Culture College, Yunnan University, Lijiang 674100, China; 2. School of Mathematical Sciences, Soochow University, Suzhou 215006, China; 3. Department of Information, Yunnan Normal University, Kunming 650041, China)

Abstract: It supposes that the search results bear some relation both to the key words of the query and a certain text category. As such, queries are constructed according to the feature words extracted from the initial sample set, then queries are send to the search engine and web pages are downloaded from the search results which response from the search engine. Downloaded web pages are processed by eliminating of duplicated content, noise reduction and extraction of text content. These samples are expanded into the sample set after the category of the samples is predicted. Finally a Naive Bayes text classifier is retrained by the enlarged sample set. The classification effect of the classifier is also experimented. Experimental results show that the precision of semi-supervised text classification method with information retrieval techniques is significantly better than the classifier constructed by small sample set.

Key words: text classification; semi-supervised learning; information retrieval; search engine

0 引 言

随着计算机技术的飞速发展和互联网的普及,用户可以获得越来越多的数字化信息,但同时也需要投入大量时间对信息进行组织和整理. 为了减轻这种负担,学者们研究使用计算机对文本进行自动分类. 文本分类就是在给定的分类体系下,由计算机根据文本的内容确定与它相关联的类别. 文本分类属于人工智能技术和

收稿日期: 2011 - 07 - 01
基金项目: 国家自然科学基金项目(60903131); 云南省教育厅科学研究基金项目(2010Y108)
作者简介: 贾志洋(1980 -),男,吉林吉林市人,讲师,硕士,主要研究方向为机器学习、知识管理.

信息检索(Information Retrieval) 技术相结合的研究领域. 国内外学者在文本分类以及相关的信息检索、信息抽取等领域进行了较为深入的研究. 在研究初期,自动文本分类以知识工程的方法为主,根据领域专家对给定本领域的分类经验,人工提取出各种逻辑规则,作为计算机自动文本分类的依据. 随着机器学习的不断发展,学者开始研究实现各种基于机器学习的分类算法,主要涉及监督学习和无监督学习. 监督学习需要大量的有标号数据; 无监督学习在学习时只用无标号数据,有标号数据只是作为检测其学习性能的测试数据[1]. 半监督学习在学习过程中可以同时利用这两类数据,从而提高分类的准确性.

半监督学习是近年来机器学习领域的一个研究热点,已经出现了很多半监督学习算法[2],其中较具代表性的算法有半监督 EM 算法[3]、协同训练(co-training) 算法[4]、直推式支持向量机[5]等. 在很多实际应用中,随着数据采集技术和存储技术的发展,获取大量的无标号样本已变得非常容易,而获取有标号样本通常需要付出很大代价. 因而,相对于大量的无标号样本,有标号的样本数量有限. 传统的无监督学习只利用无标号样本学习,监督学习则只利用少量的有标号样本学习,而半监督学习的优越性体现在能够同时利用大量的无标号样本.

文本分类相关研究的核心问题之一就是实验样本集的构建与选择. 英文文本分类方面,国外已经有了 TREC 等语料库,中文的网页分类语料库有北京大学网络实验室的中文网页分类训练集. 语料库的构建既耗费人力、物力,又在某些领域中难以应用,比如现有的语料库就很难满足从大量自然灾害的文本对地震、海啸、台风、洪水等各种自然灾害的文本进行分类预测的任务. 同时在处理不均衡样本集时,分类器预测倾向于多数类,少数类分类误差大[6].

互联网的发展,特别是基于关键字的全文搜索引擎相关的信息检索技术的发展,为解决类似的问题提供了一个构建实验数据集的良好条件. 全文搜索引擎通过网络爬虫遍历 Web 空间,并沿着网页上的链接从一个网页到另一个网页采集网页的最新信息,然后对下载的网页内容进行分析,根据信息检索的相关度算法进行计算并建立关键字倒排索引,当用户输入关键词进行查询时,搜索引擎会从关键词倒排索引中查找符合该关键词的所有相关网页,并按一定的排名规则呈现给用户[7]. 本文设计的半监督文本分类方法就利用了搜索引擎的这一特性构建查询并从查询搜索引擎的查询结果中自动下载大量有类别倾向的无标号网页样本以扩充初始样本集. 由于无标号样本是从搜索引擎中下载的有类别倾向的样本,故分类算法在自学习训练的过程中不仅根据基分类器对无标号样本分类,而且将样本的类别倾向作为分类因素.

1　半监督分类器

本文设计的分类器由两部分组成: 首先根据有标号的初始样本集构建查询,并根据查询从搜索引擎的查询结果中获取有类别倾向的无标号网页样本; 然后由基准弱分类器从无标号样本中挑选合适的网页样本并扩充到实验样本集; 最后根据已增加的实验样本集迭代训练分类器,直至迭代完成构建最终的分类器,其工作流程如图 1 所示.

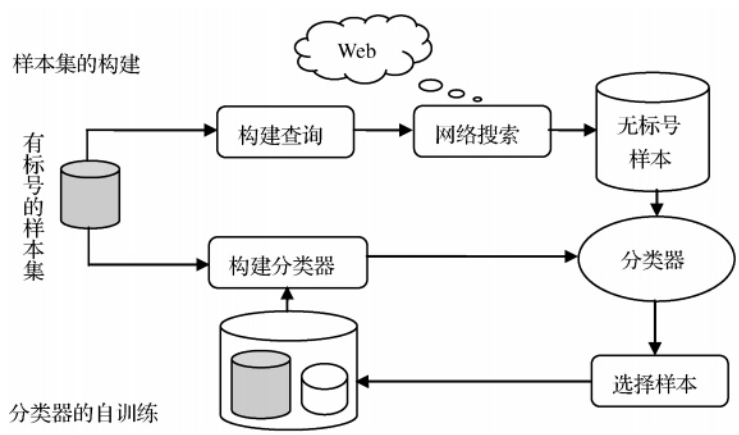

图 1　分类器工作流程图

1.1 实验样本集的构建

本文的分类方法需要一个初始样本集,初始样本集需要拥有少量的有标号样本(每个类别的样本数量不超过10个),在分类器自学习的过程中通过搜索引擎从Web自动下载无标号样本网页以扩充样本集.运行此过程需要构建一系列供搜索引擎使用的查询,这些查询由各类别的特征词组成.

为了构建搜索引擎使用的查询,需要为每一个类别构建查询序列,这些查询由每个类别的特征词组合而成.本文根据互信息从初始样本集中选取特征词,互信息度量了类 c 和特征词 w 之间的关联信息,用于表征两个变量的相关性[8].特征词 w 与类别 c 的互信息采用下式计算:

$$MI(w,c) = \log\frac{p(w,c)}{p(w)\cdot p(c)} = \log\frac{A\cdot N}{(A+B)\cdot(A+C)}, \tag{1}$$

式中,$p(w,c)$ 表示特征词 w 和 c 类文档同时发生的概率;$p(w)$ 为特征词 w 发生的概率;$p(c)$ 为 c 类的概率,就是特征词 w 和文档 c 发生时的互信息,即特征词 w 与类 c 之间的相关程度;A 为 w 和 c 同时出现的次数;B 为 w 出现而 c 没有出现的次数;C 为 c 出现而 w 没有出现的次数;N 为所有训练文档总数量.当特征词的出现只依赖于某一类时,特征词与该类的互信息量很大;当特征词与该类相互独立时,互信息为0;当特征词很少在该类型文中出现时,它们之间的互信息为负,即负相关.

为每一个类别 c_i 挑选与此类别相关特征词 w_i 的要求如下:

(1) 特征词 w_i 在类 c_i 出现的次数要高于所有特征词在类 c_i 出现的次数的平均值.

(2) 特征词 w_i 与类 c_i 的互信息为正,即满足如下公式:

$$MI(w,c) > 0. \tag{2}$$

每个类别的特征词提取完成后,就可根据这些特征词为每个类别构造对应的查询序列,每个查询应为若干个特征词的组合.根据Xu的研究成果[9],本文分别选择两个和三个特征词的组合构建查询,即每个类别的特征词共 m 个$\{w_1,\cdots,w_m\}$,可以构建 n 个查询$\{q_1,\cdots,q_n\}$,其中 n 满足公式(3),查询 $q=\{w_1,w_2,w_3\}$ 与类 c 的相关度的计算采用公式(4).

$$n = C_m^3 = m!/3!(m-3)!, \tag{3}$$

$$I_c(q) = \sum_{i=1}^{3} f_{w_i}^c \cdot MI(w_i,c). \tag{4}$$

如果一个类需要根据查询下载 N 个新样本,那么从查询 q_i 中抽取的网页样本数量为

$$\mathrm{Amount}_c(q_i) = \frac{N}{\sum_{k=1}^{M} I_c(q_k)} \cdot I_c(q_i). \tag{5}$$

由于每一个下载的新样本都有一个类别倾向,故这个类别倾向可以在半监督学习的过程中充分利用.

1.2 半监督分类器工作流程

半监督分类器的学习目标为:通过迭代从Web下载新的网页样本,以增加样本集的大小,从而提高分类器的准确率.根据这一目标,假设初始样本集为 T,本文设计的半监督分类器学习算法流程如下:

(1) 使用训练样本集 T,基于朴素贝叶斯分类算法训练一个弱分类器 C_1.

(2) 使用分类器 C_1 对从Web下载的无标号样本集 E 中的网页进行分类预测.

(3) 根据分类结果从样本集 E 中为每一个类别挑选 m 新样本组成样本集 E_m,其中 E_m 满足公式(6),选择的依据为:分类器 C_1 的预测结果与此网页的类别倾向一致,即此网页是根据此类别的特征词构造的查询获取的网页;分类器 C_1 的分类预测结果的置信度最高的 m 网页样本.

$$E_m \subseteq E \tag{6}$$

(4) 将挑选出来的样本集 E_m 扩充到样本集 T,T 满足公式(7),构建出新的训练样本集.将挑选出的样本网页从无标号样本集中剔除,E 满足公式(8).

$$T \leftarrow T \cup E_m, \tag{7}$$

$$E \leftarrow E - E_m. \tag{8}$$

(5) 迭代步骤(1)-(4)若干次,其迭代次数为 δ,δ 为阈值.

(6) 使用训练样本集 T 训练最终的分类器.

1.3 查询与网页样本处理

本文的方法在查询序列构造完成后,根据查询序列查询搜索引擎并分析搜索引擎的查询结果,然后从查询结果中下载网页样本,本文为每一个类别下载了 1 000 个网页样本以扩充训练网页集. 在根据搜索引擎的查询结果下载网页样本时,需要对搜索引擎返回的结果进行处理,目的是防止样本集中包含重复文本样本,因为不同的查询可能获取同样的结果网页,同时某一查询的查询结果中的网页内容可能重复. 为了解决这一问题,需要维护一个网页 URL 列表,其中包含了已下载网页样本的 URL 的 Hash 值[10],在每一个网页在下载前先判断此 URL 是否已经下载,采用了基于 Bloom Filter 算法[11] 以防止下载内容重复的网页,网页样本的处理流程如图 2 所示.

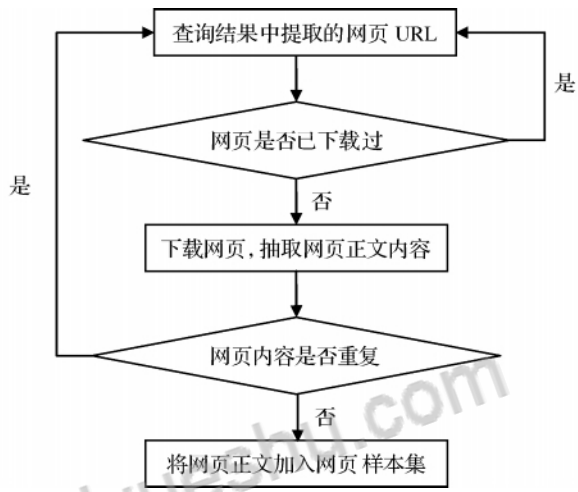

图 2　网页样本的处理流程

由于下载下来的网页样本包含了大量的噪音,需要进行去噪处理,并将网页的正文抽取出来. 本文采用基于分块的网页正文提取 TVPS 算法[12] 提取网页正文: 首先将已下载网页的 HTML 源代码进行标准化处理,使用 HTML tidy 工具将不合符标准的网页标签进行清理,然后构造网页标签树,将网页中标签按照嵌套关系整理成一个树状结构并将树中无文本节点裁剪掉,再根据 TVPS 算法对网页内容分块,最后提取正文块并提取正文块中文本内容.

2 实验与实验分析

2.1 查询与网页样本处理

为测试本文的半监督文本方法的分类效果,本文设计实现了 STCIRS (Semi-supervised Text Classification with Information Retrieve System) 系统. STCIRS 系统是基于 Windows 环境开发的,基于. NET Framework 3.5 框架,采用 Visual C# 2008 作为开发工具,在 Intel 双核 2.0GHz、2G 内存的计算机开发实现的. STCIRS 系统由两个模块组成,分别是基于朴素贝叶斯的文本分类模块和网络爬虫模块. 根据本文提出的文本分类方法,首先由文本分类模块根据训练文本集训练基准分类器并构造查询,然后由网络爬虫模块根据查询从 Web 下载网页并将处理后的网页文本入库,再由分类模块训练新分类器,完成这个过程后根据迭代次数迭代此流程.

2.2 查询与网页样本处理

为了测试本文的半监督分类方法的分类性能,本文针对地震、旱灾、海啸、洪水、泥石流、台风等六类文本进行了测试. 数据集并没有选用已有的数据集,而是手动收集了训练基准弱分类器使用的实验样本和测试分类器效果的测试样本. 训练样本集由通过维基百科查询对应类别前 10 条相关条目的解释组成,如地震对应的前 10 条相关条目分别为地震、地震波、地震列表、地震学、震级、地震带、地震预测、震源、震中、地震云; 测试样本集由 1 023 个自然灾害类文本组成,其中大部分文本通过手工查询百度新闻 2003 年 1 月 1 日至 2006 年 1

月1日的新闻组成,查询关键词为对应类别的名称,其中地震文本172个,旱灾文本170个,海啸文本160个,洪水文本样本182个,泥石流文本样本165个,台风文本样本174个.

表1 分类器的样本集自扩充过程

迭代次数	分类准确率/%	构造的查询	扩充到数据集的新闻网页
基准分类器	36.7	(震动,整个,坍塌)	玉树地震72小时逃生录: 幸存者称到处是伤者
1	42.2	(震源,震动,整个)	5.8级! 凌晨,强震袭击睡眠中的意大利
2	46.5	(震源, 震级,地震)	钢结构建筑抗震性能最好
3	52.0	(震动,整个,国家地震局)	争议地震局
4	48.5	(坍塌,典型,波长)	30岁还是五千万岁? "最年轻黑洞"透露天文玄机
5	47.1		

将初始数据集大小设置为1(即为每个类别准备1条训练文本),其中每条训练文本都是根据维基百科中对应类别名称的条目的解释,构建查询后查询百度新闻,从百度新闻下载2006年1月1日至2011年1月1日之间的新闻网页作为扩充样本集的样本,将 m 值设置为1,迭代次数 δ 设置为5,运行STCCIRS系统并测试其分类效果. 表1为STCCIRS系统的文本分类过程中每次迭代后分类准确率以及为地震类别构建的查询.

2.3 查询与网页样本处理

表2 分类器的准确率

初始数据集大小	基准弱分类器准确率/%	m 值	分类器的准确率/%				
			1次迭代	2次迭代	3次迭代	4次迭代	5次迭代
1	36.7	1	42.2	46.5	52.0	48.5	47.1
2	44.2	1	48.2	52.8	56.3	57.4	61.1
5	52.5	1	57.2	59.5	63.6	65.4	64.7
10	70.1	1	75.0	78.5	82.8	81.8	83.2
1	36.7	5	49.8	53.3	52.8	55.6	50.9
2	44.2	5	58.6	63.6	60.1	54.3	58.9
5	52.5	5	64.7	76.2	83.3	85.6	86.1
10	70.1	5	78.3	84.7	88.7	89.0	91.3
1	36.7	10	48.6	51.5	52.8	49.2	47.8
2	44.2	10	58.9	64.7	65.4	65.1	64.9
5	52.5	10	72.4	78.3	84.8	88.4	89.8
10	70.1	10	82.1	86.5	89.4	89.9	91.2

如表2所示,表中的初始样本集大小表示每个类别的样本数量,基准弱分类器为使用初始样本集训练的分类器,m 值为每次迭代从已下载的无标号网页样本扩充到初始样本集的网页样本数量,迭代次数 δ 的取值分别实验了1-5. 通过实验数据可以发现,本文设计的半监督文本分类方法的分类效果与初始的小数据集的分类效果相比有很大的提高,但由于初始数据集样本的选择制约了分类器的分类效果,并且初始样本集的数量也制约了分了器的分类效果,m 的取值也是影响本文设计的方法准确率的重要参数.

3 结　语

本文提出了一种新的半监督文本分类方法,此方法与其他半监督学习方法相比,优点为无需事先准备好的无标签样本集,可以自动从 Web 下载无标签样本; 其次,该方法在从无标签样本中挑选样本并扩充训练集时充分利用了其分类倾向(构造查询的特征词所属的类别) .

根据本文的实验结果可以得出如下结论:

(1) 根据类别的特征词构造查询是可行的,根据构造的查询并从查询结果中下载的网页样本与构建查询的类别具有较高的相关性,将这些相关网页样本下载并扩充到训练样本集可以显著提高分类器的分类准确率.

(2) 根据在搜索引擎中的查询结果下载网页样本的工作方式使得本文的分类方法无需过多的训练样本即可获得比较好的分类器准确率.

(3) 在处理不均衡样本集时本文的方法更具优势.

(4) 在训练样本集较难获取时,比如本文的自然灾害类文本分类问题,该方法可以根据小样本集从 Web 下载样本并自动扩充样本集.

虽然本文取得了一些研究成果,但是很多方面没有进行深入的研究和实验,还有许多工作需要进一步完善: 通过实验可以发现,在样本集每次迭代时为每个类别扩充的样本数量应一致,但参数 m 与 δ 的取值及其对分类效果的影响也应在后续的工作中完善.

参考文献:

[1] 苏金树,张博锋,徐昕. 基于机器学习的文本分类技术研究进展 [J]. 软件学报,2006,17(9) : 1848 - 1859.
[2] 彭岩,张道强. 半监督典型相关分析算法 [J]. 软件学报,2008,19(11) : 2822 - 2832.
[3] 张博锋,白冰,苏金树. 基于自训练 EM 算法的半监督文本分类 [J]. 国防科技大学学报,2007,29(6) : 65 - 69.
[4] Blum A, Mitchell T. Combining Labeled and Unlabeled Data with Co-Training [C] // Proceedings of the 11th Annual Conference on Computational Learning Theory. New York: ACM Press, 1998: 92 - 100.
[5] 肖建鹏,张来顺,任星. 基于增量学习的直推式支持向量机算法 [J]. 计算机应用,2008,28(7) : 1642 - 1648.
[6] 林智勇,郝志峰,杨晓伟. 不平衡数据分类的研究现状 [J]. 计算机应用研究,2008,25(2) : 332 - 336.
[7] Boldi P, Santini M, Vigna S. PageRank: Functional dependencies [J]. ACM Transactions on Information Systems,2009,27 (4) : 192 - 195.
[8] 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 [J]. 中文信息学报,2004,18(1) : 26 - 32.
[9] Xu Zenglin. Learning with Unlabeled Data [D]. Hong Kong: The Chinese University of Hong Kong, 2009.
[10] 苏国荣,杨岳湘,邓劲生. 一种去除重复 URL 的算法 [J]. 广西师范大学学报: 自然科学版,2010,28(1) : 122 - 126.
[11] 张刚,刘挺,郑实福,等. 大规模网页快速去重算法 [C] // 中国中文信息学会二十周年学术会议论文集. 北京: 清华大学出版社,2001.
[12] 于满泉,陈铁睿,许洪波. 基于分块的网页信息解析器的研究与设计 [J]. 计算机应用,2005,25(4) : 974 - 976.

(责任编辑:子　实)

阅读此文的还阅读了：

1. 文本信息检索技术
2. 文物信息的分类与检索
3. 文本分类技术研究进展
4. 结合文本信息量和聚类的文本裁剪算法
5. 结合句子级别检索的信息检索模型
6. 文本分类技术在信息检索中的应用
7. 文本信息检索技术
8. 结合信息检索技术的半监督文本分类方法
9. 个性化信息检索中的文本分类方法
10. 面向信息检索的文本自动分类技术研究
11. 分层分类与监督分类相结合的遥感图像信息提取方法研究
12. 网络信息资源检索方法
13. 融合分类特征的信息检索技术研究
14. 文本信息检索实验方法研究
15. 用于信息检索的文本聚类技术
16. 超文本技术与信息检索
17. 信息检索技术应用
18. 结合边缘信息的图像检索技术
19. 网上信息检索技术的改进方法
20. 结合TFIDF方法与Skip-gram模型的文本分类方法研究
21. 文本挖掘技术研究及其在信息检索中的应用
22. 文本信息检索装置以及文本信息检索方法
23. 基于半监督学习算法在文本分类中的应用研究
24. 基于半监督学习算法在文本分类中的应用研究
25. 文本与轨迹交叉检索技术

26. 结合半监督学习和LDA模型的文本分类方法
27. 基于CPC分类号在信息检索的有效检索策略
28. 半监督文本分类综述
29. 结合聚类的半监督分类方法
30. 文本分类技术进展
31. 基于降维技术的文本分类技术
32. 信息检索理论与技术结合与梳理的经典之作——评《信息检索理论与技术》
33. 现代技术与传统方法交融下的信息检索技术
34. 网页文本分类技术研究
35. 网络信息检索技术探析
36. 信息检索技术
37. WEB中文本信息检索的关键技术研究
38. 网上科技信息的分类与检索
39. 基于本体的文本信息检索研究
40. 一种简单实用的文本信息检索方法
41. 文本信息检索中的概率模型
42. 基于半监督的SVM迁移学习文本分类算法
43. 基于本体的文本信息检索
44. 信息检索与文本挖掘
45. 文本分类技术探究
46. 基于领域本体的文本信息检索探讨
47. CPC分类号在信息检索领域的检索应用
48. 信息分类检索的技术演进及模式
49. 探讨文本挖掘技术研究在信息检索中的应用
50. 浅述专题信息的检索工具和检索方法