

基于非平衡学习与交互式标注的引文情感动机标注系统

孙亦昕¹, 许露², 郑翼斐¹, 朱妍¹, 唐媛¹,
董猛¹, 刘宇¹, 胡凯¹



(1.江南大学物联网工程学院, 江苏 无锡 214122;

2.江苏省行政学院决策咨询研究中心, 江苏 南京 210009)

✉1160116214@stu.jiangnan.edu.cn; x109_05@126.com; zhengcharon@163.com; 1034170102@stu.jiangnan.edu.cn;
1034170408@stu.jiangnan.edu.cn; 1030416629@stu.jiangnan.edu.cn; 1033170635@stu.jiangnan.edu.cn; hukai_wlw@jiangnan.edu.cn

摘要: 随着现代文献的数据量不断增长, 仅仅依靠论文引用次数的传统方式无法很好的描述论文的真实影响力, 基于引文内容的情感动机的研究成为新的研究重点。虽然当前的研究中存在大量自然语言处理工具可以处理文献文本, 但并没有专门针对文献引文情感动机的处理工具。为解决引文动机样本分布极度不均衡和引文动机标注数据生成效率低下的问题, 本文提出构建一个基于非平衡学习和交互式标注的引文系统。经过实验证明, 本文提出的系统可以较好地提高引文情感动机标注效率。

关键词: 非平衡学习; 引文情感动机; 交互式; 标注系统

中图分类号: TP311.5 **文献标识码:** A

Citation Emotive Tagging System based on Unbalanced Learning and Interactive Tagging

SUN Yixin¹, XU Lu², ZHENG Yifei¹, ZHU Yan¹, TANG Yuan¹, DONG Meng¹, LIU Yu¹, HU Kai¹

(1.School of Internet of Things Engineering, Jiangnan University, Wuxi 214122, China;

2.Decision Consultation and Research Center of Jiangsu Administrative College, Nanjing 210009, China)

✉1160116214@stu.jiangnan.edu.cn; x109_05@126.com; zhengcharon@163.com; 1034170102@stu.jiangnan.edu.cn;
1034170408@stu.jiangnan.edu.cn; 1030416629@stu.jiangnan.edu.cn; 1033170635@stu.jiangnan.edu.cn; hukai_wlw@jiangnan.edu.cn

Abstract: With the increasing data volume of modern literature, the traditional method of paper citation cannot fully represent the real influence of the paper, so emotional motivation (Emotive) study based on citation content has become a new research focus. Although a large number of natural language processing tools exist in the current research to process literature texts, special tools to process literature citation emotional motivation have not been found yet. In order to solve the problems of extremely unbalanced sample distribution of citation motivation and low efficiency of citation motivation annotation data generation, this paper proposes building a citation system based on unbalanced learning and interactive annotation. Experiment results show that the system proposed in this paper can improve the tagging efficiency of citation emotive.

Keywords: unbalanced learning; citation motivation; interaction; tagging system

1 引言(Introduction)

引文情感动机分析作为重要的评价科学成果的方法越来越成为当前科学文献研究的重要问题^[1], 相关的研究表明, 引用行为不仅是对相关工作的认可, 而且受很多非科学因素的影响, 呈现出高度复杂性的特征^[2]。现有的关于引文情感动机的研究, 倾向于自动化方法效果的提升等领域, 却忽视了当

前这类研究中缺乏大量的标注数据的问题, 这种矛盾在文献数据爆炸性增长的当下显得尤其突出, 如何解决这样的矛盾是本文的核心。若通过引入自动化分类的方法, 用分类器代替人工从而实现高效分类(标注), 则存在两方面问题: 首先, 由于科学论文的文本具有较强的领域特点, 正面或者无感情偏好的样本规模往往大于负向的情感偏好, 因而导致了样本

分布的不均衡^[3-5]，使得分类器的分类效果很差；其次，对于引文上下文的标注往往需要领域专家的大量背景知识，并不容易获取^[6]，因此大规模的标注数据的生产与获取仍存在困难。

针对以上问题，本文提出构建交互式标注系统，区别于传统的标注方式，将标注的过程融合交互过程。用户在标注的过程中，系统不断的对引文的标注内容进行非平衡学习的算法模型训练，用生成的模型将产生的数据进行分类标注，只将不容易确定的标注类型推送至专家，通过这样的机制提升标注的效率，以达到高效而准确标注的目标。

2 研究背景(Research background)

引文情感动机识别属于引文类别识别，后者的相关研究很早就成为引文分析的重要研究课题。早在1999年Teufel的博士论文就对引文类型识别研究的过程分成以下几个步骤：引文类型定义、引文内容抽取和引文类型标引^[7]。引文类型定义是随着引用动机的理论和实践研究产生的，比如Small将引用动机分为驳斥、注释、评论、应用和支持五类^[8]，Brooks通过调查将引用动机分为三大类：说服读者、积极引用、为了交流、社会认可；消极引用；提醒读者、操作性信息^[9]，并将引文的动机分成延伸、批评、比较、提高等^[10]。虽然不同研究者对引文类型有不同的理解，但基本都遵循了三个共同的原则：被引次数不能够从本质上揭示引文之间的关联，需结合引文的内容分析；被引文献对施引文献的作用和重要程度并不相同，不能同等看待被引文献；引文的上下文文本是引文类型定义的基础和关键^[11]。引文类型识别从引用动机和引文内容分析出发，能够在一定程度上揭示出施引文献和被引文献之间的语义关联^[2]，可以进一步分析被引文献对施引文献的贡献程度，是对传统引文分析方法的语义增强和重要补充，具有重要的理论和实践意义。

然而，目前引文分析仍然主要以被引次数为基础，将所有的引文同等看待，施引文献和被引文献之间的关联性也通常不加区分^[12]，无法很好的描述论文的真实影响力。这样的现状主要是由于，过去文献的全文，也就是引文的上下文并不容易获得，虽然在定义上学者们给出了较好的研究思路，但是基于引文情感动机的论文评价方法未被广泛采纳。

这一情况在新的环境下得到了扭转，随着文献的全文数据大量可得，文献数量急剧增加等也让引文情感动机的研究具有较好可行性背景^[13]。自然语言处理与分析的工具为解决这类问题提供了基本思路，例如通过统计语言学模型^[14]和主题模型^[15]，基本的文本特征可以被提取和表示。结合自然语言处理与机器学习的方法，部分情报学领域学者也结合这种思路对

引文动机的研究进行了一系列的探索，如多标签的引文文本分类方法^[16]，对方法的扩展如与影响力分析结合^[17]，与学科主题分析的结合^[18]，针对具体类型的引文内容分析^[19,20]，基于现有引用内容类型的标引框架^[16]和标准数据集合的构建^[21]等。然而这些研究都需要对引文进行标注，并在标注数据的基础上进行研究，标注数据的缺乏仍然是一个严重问题，因此，我们提出了一套交互式标注的系统，并通过非平衡学习的方式来提升标注的效率。

3 系统整体设计(Overall system designing)

系统采用B/S构架，前端提供用户交互界面，用户可以通过前端页面交互的方式，对引文的内容进行标注，同时在标注的过程中，用户产生的标注数据不断跟后台的系统进行交互，后台的非平衡学习模型通过对用户生成的数据和现有的预训练模型，不断进行训练和更新模型，同时将敏感的数据推送至交互前端，并自动对不敏感数据进行分类标注(1为正向或中性，0为负向)。

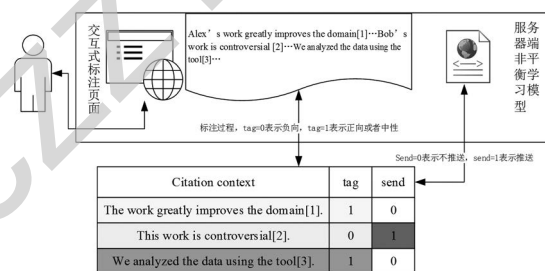


图1 系统整体设计图

Fig.1 Diagram of the overall system designing

如图1所示，其中tag列由用户标注生成，标识情感极性；send列由服务器模型生成，标识是否发送给用户端进行标注，引文内容即包含引用文本的内容，通过用户的标注，情感极性首先通过用户的认识和理解，将其内容划分为正向或者负向。以图中的文本为例“Alex's work greatly improves the domain”这句话表达了作者对于该工作的极大肯定，因此，在此标注过程中，这句话被标注成“1”也就是正向情感。而第二句话“this work is controversial”表达了该研究比较有争议，可能代表着负面评价，因此被标注为“0”负向的情感。第三句话中，“We analyzed the data with the tool”，是一个事实陈述类型的引文，情感极性为中性，被标注成“0”。在send列中，0代表不推送，1代表推送，这主要是由于0通常代表着正向的情感和中性的情感，算法对其不敏感，需要推送给专家进行人工分类。

4 数据和方法(Data and methods)

4.1 原始数据与中间数据

在文本中，原始数据的获取主要通过自行编写爬虫对

Nature Communication网站的开源文本数据爬取获得,网站通常具有一定的格式,为爬虫的爬取工作奠定了基础。如图2所示,网页中的开源论文全文数据结构被展示出来。

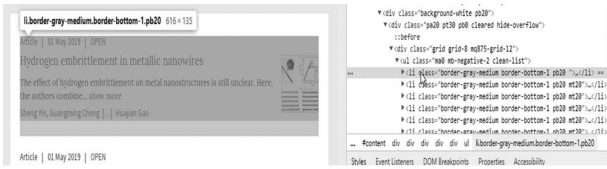


图2 开源网站中开源论文页面内容结构

Fig.2 The structure of web page for open-source articles

通过不同的爬虫爬取方法,论文就可以被逐行提取下来并存在本地的数据库中,本文主要采用了MongoDB数据库,按照不同的章节,论文的全文被分开存储在不同的字段中。例如摘要(abstract)、引言(introduction)、讨论(discussion)等,这些字段按照论文的结构对其进行了基础的组织,如图3所示。

abstract	authors	discussion	introduction	methods	references	results	title	url		
The dimeric lindenaneBao Du Zhongson in summary, weThe dimeric lindenaneAll reactions w(1)Cao, C. -M., We initiated sya unified stra/articles/41467-019-09850-8	organic compounds perClaudio U. Bail, We have presenThe interaction oThe nanocrystal(1)Goban, A. et We use a gold-cNanoprinting o/articles/41467-019-09877-5	Self-powered implantationDuyang, Zhous In summary, wMillions of patieThe fabricated r(1)Kim, J., GwThe SPW consistSymbiotic card/articles/41467-019-09851-1	Transmembrane proteoTriss L. Zhiguan In this study,Cell surface expoH293T cells (1)Bovens, E. WIn order to state inner actio/articles/41467-019-09779-7	Brucodominant-containLiaofeng Wang,So Taken togetherSMBR2(1) also knoG601, H1299, HCT(1)Huang, X., HuWe performed a BRD9 defines a/articles/41467-019-09899-7	Interfaces betw tV. Crispel, N. ClaWhile our anaThe Integer QuantWe come back to t(1)Alalila, C. D. Chemical biologa chemical too/articles/41467-019-09672-2	Brucodominant (BRD9) (Qin Wu, David HeiRecent effort Genetic and epigenethman TBCs were(1)Alalila, C. D. Chemical biologa chemical too/articles/41467-019-09672-2	Stream CO(2) emissioA. Campanu, F. BiThis study perThe flensing of tThe study was co(1)Quist, M. G.Radiocarbon anaCurrent forest/articles/41467-019-09922-3	Actic feedbacks accoDmitry Tsvetkov, We have inventThe Arctic regionWe defined the r(1)Overland, J. We base the PFCClimate policy/articles/41467-019-09869-2	Innovative materials Greer, J., HorvitzSABP is aThe worldwide onDue to the moit(1)IntermittentThe crystal str(1)H1299S2(1) articles/41467-019-09642-9	African society is peElizabeth J. KenThe results belatit now, futureThe models used (1)Christensen, The wettest 3=Enhanced futur/articles/41467-019-09776-9

图3 MongoDB中全文数据存储的视图

Fig.3 View of the full-text data store in MongoDB

已被存储在MongoDB的数据库中的内容只是论文全文,由于全文并非每一句话都包含有引用,因此还需对全文进行引用内容的抽取。引用内容在施引文献中会伴随一定的引用标签出现,常见的引用标签包括方括号+序号、圆括号+作者日期等,这些引用标签是识别引文内容的关键。识别到引用标签后更重要的是确定引文内容抽取的范围,目前引文内容抽取的范围主要有两种:一种是狭义范围,仅指包含引文的句子,即引文句;另一种是广义范围,除包含引文句子外,还包括跟引文句子相关的其他上下文句子,即引文上下文。狭义范围的抽取较为容易,准确率和效率较高;广义范围的抽取相对较难,但是能够获得更多引文相关的内容,可为揭示语义关联提供更多的信息。在本文中,我们仅采用了较为简单的基于规则的提取法,例如方括号+序号的方式,对文献文本进行分割提取,并且存储在数据库中。

4.2 非平衡分类方法

在引文标注过程中,如果全部采用人工标注的方式,不仅标注的过程劳神费力,更严重的问题是无法保证标注的准确性,尤其是在某些特殊的研究领域中,大量出现的术语通常属于术语低频词汇,领域外的人员无法领悟其中的含义,容易产生误解,因此,完全依赖人工必然存在巨大的问题。

但要在标注过程完全通过机器自动进行,同样存在着巨

大挑战。有研究指出,负向的情感引用往往存在于特定的文章段落;例如评价性的引用通常在引文或者讨论部分出现,其他地方如数据和方法阶段的引文则通常是基本的中性的引用。这样就导致了以下问题,一方面引文具有极性特殊的聚集性,另一方面它们本身的分布业绩不均衡,往往正面的引用占有较多篇幅,而负面的引用只有很少内容。综上所述,简单的采用机器分类的方法无法满足高效推送引文标注的需求。因此,本文引入了非平衡学习的分类方法,如图4所示。

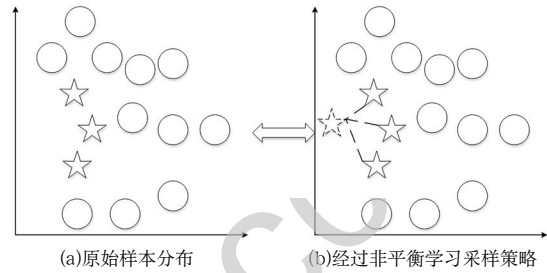


图4 非平衡学习示意图

Fig.4 Diagram of unbalanced learning

在图4(a)原始样本分布中,两个类型的分布并不均衡,非平衡学习则可以通过插值的方式,生成新的样本,增加小类的样本数量,图4(b)显示了经过非平衡学习采样策略小类的样本生成新的样本的过程。这样的策略同样可以用来增强引文标注的过程,增加小类样本数量,比如负向情感引文的样本数量。通过这种方式,使最终分类的效果取得提升,更好的识别出负向情感的引文文本,进而实现标注过程效率的提升。

5 系统实现(System implementation)

5.1 系统的实现界面

系统的前端界面主要采用了Vue框架,通过模块化的方式,分别设计标注系统的多个可视化页面,最后通过JavaScript函数和后端的Web Service接口调用的方式,实现前端数据向后端的传输。如图5所示,图中展示了页面交互的详细可视化元素,用户主要通过勾选方式进行句子的情感极性标注,页面中句子的内容由后端的数据服务结构提供。

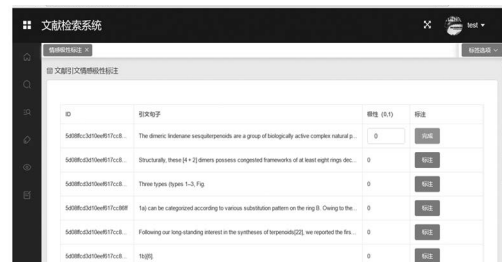


图5 标注交互页面

Fig.5 Interactive annotation pages

5.2 后端核心代码

服务器端主要使用Python的机器学习框架Scikit-learn和

Python的Web服务框架Flask构成数据标签分类服务。其主要核心工作可以有两部分组成，第一部分主要是通过预训练的语义模型获得文本的向量，第二部分则是通过非平衡学习的分类方法，采用上采样的方式增加小众样本的比例，从而提升自动化分类的效率，同时在推送的过程中，采用阈值设定的方法，将分类准确度较明确的对象自动分类打上标签，而将不确定的靠近分类临界面的对象send设置为1,并推送给用户进行交互性标注。

第一部分通过预训练语义模型获得文本向量：

```
sent_tokenizer=nlk.data.load(r'tokenizers/punkt/english.pickle')

def all_embedding():
    embeddings_index={}
    with open(r'numberbatch-en.txt',
encoding='utf-8') as f:
        for line in f:
            values=line.split(' ')
            word=values[0]
            embedding=np.asarray(values[1:],
dtype='float32')
            embeddings_index[word]=embedding
    return embeddings_index
```

第二部分通过非平衡学习的分类方法，使用了sklearn.utils.resample方法对语义样本进行重新采样。

```
X_train,X_test,y_train,y_test=train_test_split(train,
drop(['0_y'],axis=1),y_train,te st_size=.2,random_
state=1)
```

```
X=pd.concat([X_train,y_train],axis=1)
nonneg=X[X['0_y']==1]//非负样本
neg=X[X['0_y']==0]//负样本
neg_upsample=resample(neg, replace=True, #
sample with replacement
n_samples=len(nonneg), random_state=21) #
reproducible results
```

5.3 标注性能评估

本文性能评估主要从效率和精度的两个方面展开。效率主要是指，在单个用户操作的情况下，单位时间内用户标注数量加上机器标注数量的总和。采用本系统，标注过程只需要标注三分之一的工作量就可以，就可以获得具有较高精度的正常数据量的标注结果，具体数量见表格表1所示。

表1 传统手工标注与本系统标注工作量对比

Tab.1 Comparison of traditional annotation and our system

指标	传统标注	本系统标注
需要数量	1500次	300次
获得数量	1500个	1500个

在获得标注数据之后，通过对数据集的精度考查，即主要是通过设置训练集合和验证集合，分别对比未采用非平衡学习和采用了非平衡学习的方式的分类Precision-Recall曲线图，如图6所示，左图为未经非平衡学习的分类效果，右图为经过非平衡学习后的分类效果。

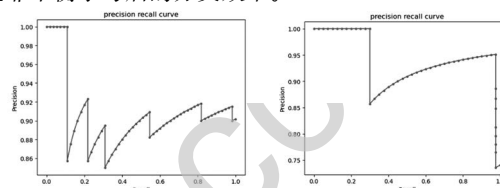


图6 分类性能测试

Fig.6 Classification performance test

6 结论(Conclusion)

传统引文内容分析方式通常从归纳类型动机角度出发展开研究。本文则主要从可行性与数据源头的可得性角度出发，发掘当前的引文动机分析存在标注数据不足的问题，由于纯粹使用人工或者机器的方式都无法较好满足标注的需要，因此本文提出了一种结合机器学习与人工交互的标注系统，从原理和设计的角度对系统的构成进行了完整的描述，并通过实验验证了本文提出系统的有效性。虽然本文提出的系统对标注效果具有一定的提升，但是仍然存在一些可优化之处，例如提高现有模型的分类精度。

参考文献(References)

- [1] 赵洁.知识场论与知识流动——兼论科学引文机理分析[J].情报探索,2020(03):11-15.
- [2] 尹莉,郭璐,李旭芬.基于引用功能和引用极性的一个引用分类模型研究[J].情报杂志,2018,37(07):139-145.
- [3] 王杰,李德玉,王素格.面向非平衡文本情感分类的TSF特征选择方法[J].计算机科学,2016,43(10):206-210;224.
- [4] 万志超,胡峰,邓维斌.面向不平衡文本情感分类的三支决策特征选择方法[J].计算机应用,2019,39(11):3127-3133.
- [5] 田锋,王媛媛,吴凡,等.超平面距离的非平衡交互文本情感实例迁移方法[J].西安交通大学学报,2018,52(10):1-7.
- [6] 田锋,兰田,Kuo-Ming C,等.领域实例迁移的交互文本非平衡情感分类方法[J].西安交通大学学报,2015,49(04):67-72.
- [7] Teufel S, Siddharthan A, Tidhar D. An annotation scheme for

(下转第52页)