计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.191100176

基于编辑距离的多实体可信确认算法

孙国梓　吕建伟　李华康
南京邮电大学计算机学院　南京 210003

摘　要　随着自媒体的蓬勃发展,任何人都可以在网上随意发布和转发信息,而这些信息可能是真实的,也可能是道听途说或被故意篡改的。互联网上数据的严重冗余和弱可信问题,导致现有数据的可用性很差。Bi-LSTM-CRF(Bi-Long Short Term Memory with Conditional Random Field Layer)网络虽然能够解决数据中命名实体识别的准确率问题,但不能满足识别出的实体是可信的这一要求。文中提出一种基于编辑距离的多实体可信确认算法,并通过人物命名实体识别实例对该算法进行验证。首先通过分布式爬虫抓取同一个邮箱地址在多个搜索引擎上的 Top N 网页记录,然后使用经过双语语料训练后的 Bi-LSTM-CRF 模型抽取每个页面内的人物命名实体,最后通过实体多参数融合确定邮箱所对应的人物命名实体。实验结果表明,多实体可信确认算法能够将邮箱地址与邮箱真实主人的匹配准确率 MRR(Mean Reciprocal Rank)提高到 91.32%,相比只使用词频的算法其 MRR 提升了 23.08%。实验数据充分说明,多实体可信确认算法能很好地从弱可信数据中获得强可信度的实体,降低海量数据中的低质特性,从而有效地增强实体数据源的可信度。

关键词:弱可信数据;双向长短时记忆循环-条件随机场网络;多实体可信确认算法;编辑距离
中图法分类号　TP311;TP391

MeTCa:Multi-entity Trusted Confirmation Algorithm Based on Edit Distance

SUN Guo-zi,LYU Jian-wei and LI Hua-kang
School of Computer Science and Technology,Nanjing University of Posts and Telecommunications,Nanjing 210003,China

Abstract　With the development of We-media,every individual can publish and forward information on the internet at will. The information may have real records,but it may also be hearsay or even contents being intentionally tampered with. The data on the Internet has serious redundancy and weak credibility problems,thus resulting in low availability of existing network media data. Although the Bi-LSTM-CRF network can solve the problem of the accuracy of named entity recognition in data,it cannot meet the requirement that the identified entity is credible. In this paper,a multi-parameter fusion credible confirmation algorithm based on multi-source weakly trusted data is proposed,which is verified by identifying instances of person named entities. This paper uses distributed spiders to crawl Top N pages with the same mailbox address on multiple search engines. Afterwards,Bi-LSTM-CRF algorithm trained by bilingual corpus is adopted to extract person named entities from each page. Finally,the person named entities corresponding to the mailbox are determined by multi-parameter entity fusion trusted confirmation algorithm. The experimental results show that the multi-parameter fusion credible confirmation algorithm can improve the accuracy of MRR (MRR) of the matching between the mailbox address and the real owner of the mailbox to 91.32%,which is 23.08% higher than the traditional algorithm using only the term frequency model. The experimental data reasonably demonstrates that the multi-parameter fusion credible confirmation algorithm can obtain strong credibility entities from weakly trusted data and reduce the low-quality characteristics of massive data,thus effectively enhancing the credibility of entity data sources.

Keywords　Weak trusted data,Bi-LSTM-CRF,Multi-parameter fusion trusted confirmation algorithm,Edit distance

收稿日期:2019-11-25　返修日期:2019-12-23　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金(61502247,11501302,61502243);中国博士后科学基金(2016M600434,2016M591840);江苏省博士后科研基金(1601128B);江西省经济犯罪侦查与防控技术协同创新中心开放基金资助课题(JXJZXTCX-015);数字工程与先进计算重点实验室开放课题(2017A10)
This work was supported by the National Natural Science Foundation of China(61502247,11501302,61502243),China Postdoctoral Science Foundation(2016M600434,2016M591840),Jiangsu Postdoctoral Research Foundation (1601128B),Economic Crime Investigation and Prevention and Control in Jiangxi Province Supported by the Open Fund of the Collaborative Innovation Center of Technology(JXJZXTCX-015) and Open Project of the Key Laboratory of Digital Engineering and Advanced Computing(2017A10).
通信作者:孙国梓(sun@njupt.edu.cn)

1 引言

在跨入工业 4.0 时代后，数据业务以指数级迅速增长，使得信息生产、传播到存储过程中的一系列方式方法产生了根本性变化。这为人们提供了快速、便利的信息渠道，但同时也产生了信息过载的问题[1]，进而大大增加了人们甄别“真”“假”消息的综合成本。

对于大部分未知问题，人们习惯通过搜索引擎来查询关键词，以获取问题的解决方案。然而，与关键词有关的重要信息经常被隐藏在不同 PageRank 算法排序后的页面中，增大了可信信息的辨识难度。因此，本文以分析用户关联关系为出发点，通过自动化的方法分析 Top N 的页面中有关用户姓名、地址、工作单位等的信息，以筛选出相关的有用信息，降低冗余信息的混乱程度。

目前，研究者针对弱可信数据源的研究甚少，在文本处理方面，现有的方法仍有很多不足：1)它们只是针对某一特定领域文本[2]，或单一语言的社交数据[3]，进行强可信信息的规则化、命名实体识别(Named Entity Recognition，NER)与抽取分析，强调对语料中实体的召回率，忽略了实体本身的可信性；2)它们对搜索页面中双语言文本等更加复杂的数据缺乏行之有效的分析方法。本文受 Lample 等[4]提出的双向长短时记忆循环-条件随机场网络(Bi-LSTM-CRF)命名实体识别方法的启发，以邮箱地址为切入点，分别从三大搜索引擎获取相关数据，采用了命名实体识别方法。首先使用 Bi-LSTM-CRF 模型，对复杂的多语言文本进行命名实体识别，然后根据邮箱地址的前缀命名规律，即邮箱的前缀与用户的姓名有直接的联系，如张三喜欢将邮箱命名为 zhangs@qq.com，Allen Smitth 喜欢将邮箱命名为 allenS@gmail.com 等，使用融合编辑距离与词频(Term Frequency，TF)的多参数融合可信确认算法来进一步提高邮箱地址和第一候选实体匹配的 MRR 准确率。通过一系列的实验证明，所提方法能为匿名邮箱推荐相关度较高的有效实体。

本文的贡献如下：

1)开发一套系统，在弱可信的数据分析中强化数据的可用性，并在特定业务的范围内有一定的伸缩性。

2)利用双语 Bi-LSTM-CRF 模型，提升了识别与抽取多语言、短文本中实体的效果，为进一步的融合排序提供了更坚实的基础。

3)提出的多参数融合可信确认算法大大提高了邮箱地址与相关候选实体的匹配度和可信度。

2 相关工作

本节将介绍可信数据下实体抽取中的相关方法及其不足，提出了冗余实体中可信度差的问题。

1)可信数据下的实体抽取方法目前主要分为两部分：基于机器学习的方法和基于深度学习的方法。Bikel 等[5]提出基于隐马尔可夫模型的命名实体识别方法。Lafferty 等[6]将条件随机场应用于命名实体识别任务。Zhou 等[7]提出将 4 种不同特征来提高隐马尔可夫模型在 NER 任务上的性能。McCallum 等[8]提出将更丰富、更高阶的马尔可夫模型的特征感应法和维特比法用于 NER 任务。Hochreiter 等[9]将神经网络 LSTM 引入 NER 任务中后，Collobert 等[10]提出了一种基于卷积神经网络的模型来同时完成多个自然语言处理任务，该模型在输入部分尽可能少地使用人工特征，在多个任务中都取得了较好的结果。Huang 等[11]提出了一系列模型并且首次将 Bi-LSTM-CRF 模型应用于一般领域的多个序列标记任务中。上述方法都是基于可信数据源进行实体识别的，再通过准确率或召回率来评价模型的优劣。可信数据源中，NER 方法对弱可信数据的命名实体识别是有参考意义的，然而弱可信数据本身就可能含有许多虚假内容，识别出的实体本身就有着巨大的冗余性和迷惑性，因此只进行命名实体识别不足以解决多实体中的实体可信问题，需要进一步确认冗余实体的可信度情况。

2)冗余实体相似度计算方法。Qiu 等[12]使用基于特征相似度的方法检测多数相似重复记录。减少冗余记录。Tan 等[13]融合使用了多种编辑距离的冗余字符串度量方法，来降低字符间的冗余。Ye[14]融合近似最长公共子串算法以进一步增强冗余信息的过滤能力。上述研究只是致力于如何降低信息的冗余度，而没有在实体抽取方面有所研究。另外，针对冗余实体只用一种特征相似度方法，不能解决实体的可信度问题，因此本文将融合字符串间的编辑距离和字符的词频两种方法进行可信实体的筛选与排序。只有去除冗余实体或降低弱可信实体的排名，才能使候选实体成为真正的可信实体。

3 MeTCa 模型

多实体可信确认算法模型(Multi-Entity Trusted Confirmation Algorithms，MeTCa)的整体框架如图 1 所示，主要有 3 个阶段：分布式弱可信数据获取阶段、命名实体识别阶段和多参数融合可信确认阶段。

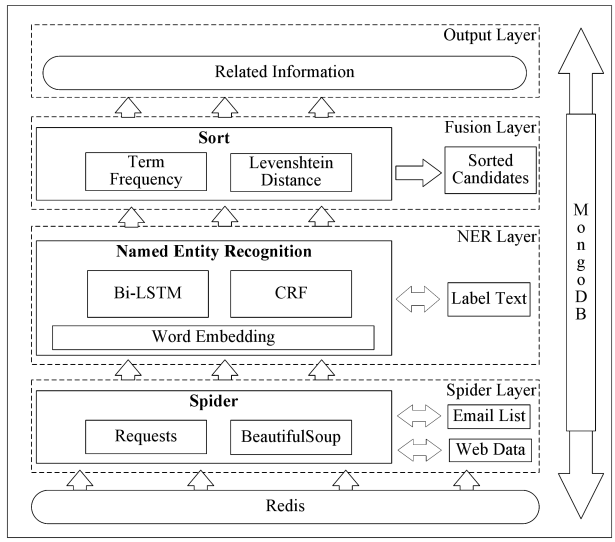

图 1 MeTCa 系统的整体架构

Fig. 1 Framework of MeTCa system

3.1 数据获取

测试的邮箱地址来自国内外各大高校公开的教师的邮箱地址，共 1000 个。通过自动化技术将邮箱地址分别输入到百度、必应和谷歌三大搜索引擎的搜索框中，爬取相关弱可信数

据。实验使用 3 台主机进行分布式爬取：一台主机作为控制节点，从文件中获取待爬取的邮箱地址，并将其传递给其余两台爬虫节点下的 URL 控制器，拼接成完整的搜索 URL 后分别进行数据的爬取与存储。爬取过程如图 2 所示。

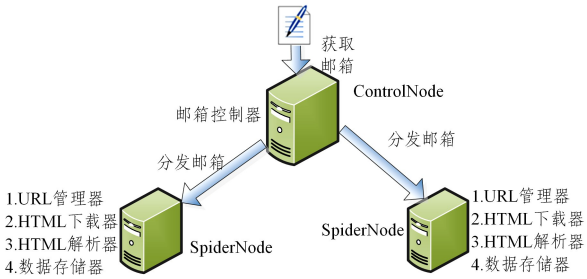

图 2 爬虫的架构图

Fig. 2 Framework of spiders crawling

3.2 命名实体识别

命名实体识别主要使用 Lample 等[4]提出的 Bi-LSTM-CRF 模型，但是由于针对单一语种训练的模型并不能在存在双语情况的网页信息中发挥效用，因此导致实体识别的准确率较低。

目标文本在进入模型识别前首先进行语种的识别，若网页内容中超过 50%的字符为英文字母，则判断该网页为英文网页，否则为中文网页[15]。然后，将它们分别投入针对中文或英文的训练的 Bi-LSTM-CRF 模型中进行实体识别。在区分语种后，实体的识别准确度得到了明显的提升。其中，Bi-LSTM-CRF 网络结构如图 3 所示，网络的输入为经过预处理的字词 one-hot 编码，将其传入网络输入层，经过嵌入层处理得到相应的字词嵌入向量[10]。将嵌入向量传入隐藏层，经过双向长短期记忆网络处理得到句子的特征序列，其中 L 层 LSTM 向右传播历史信息，R 层 LSTM 向左传播未来信息，C 层为 L 层和 R 层信息的结合。最后，将特征序列传入随机向量场层（Conditional Random Fields，CRF）并进行 Viterbe 解码，从而计算出最佳序列。

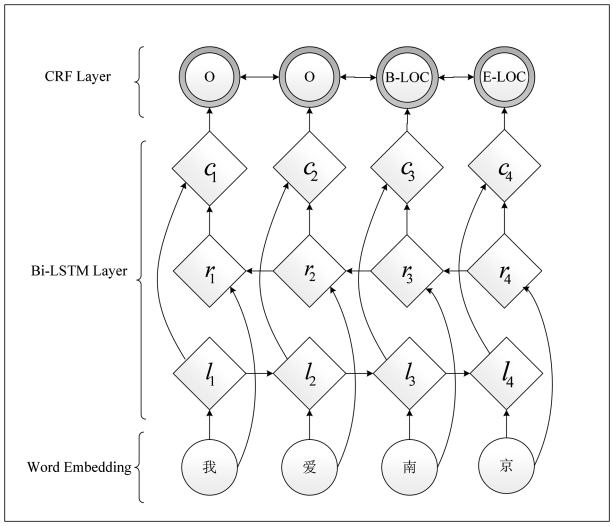

图 3 Bi-LSTM-CRF 网络的结构

Fig. 3 Structure of Bi-LSTM-CRF network

实验将预测结果进行修正后重新加入到训练集中，再次进行模型训练，使得模型在有原始语料的情况下进行修正。这样使得训练的结果更为准确，并且可以提高模型预测的准确率。

3.3 多参数融合可信确认算法

命名实体识别后的实体繁杂，因此从中挑选出与邮箱地址关联密切的实体，增强实体的可信度，是亟需解决的问题。

相关研究表明，在利用关键词进行查询时，越频繁出现的实体越能表明其与关键词有强相关性[16]。本文定义候选实体 i 为 e_i，N_{e_1} 表示第 i 候选实体的数量。实验将每个实体的频率 $freq_{e_1}$ 作为过滤弱可信实体的一个参考因子。

$$freq_{e_1}=\frac{N_{e_1}}{\sum_{i=1}^{n}N_{e_1}} \tag{1}$$

实验通过观察 500 个邮箱地址的命名方式发现，所有人的邮箱前缀命名都与自己的姓名相关，如姓名为张三，他的邮箱前缀命名为 zhangsan，zhangs，sanzh 等。基于这些规律，本文使用计算编辑距离[17]的方法，首先将中文的实体映射为对应的拼音，如“张三”对应为“zhangsan”，与邮箱的前缀进行相似度计算，获得编辑距离，具体公式如下：

$$sim(x_i,y_i)=\frac{\max(L_{x_1},L_{y_1})-\text{Levenshtein}}{\max(L_{x_1},L_{y_1})} \tag{2}$$

其中，L_{x_1} 表示中文实体转化后的拼音 x_i 或英文实体 x_i 的长度，L_{y_1} 表示对应邮箱地址 y_i 的前缀，$sim(x_i,y_i)$ 表示 x_i 与 y_i 的相似度。

利用加权的方法综合考虑这两个参考因子，统一为候选实体计算权重，为它们提供可信度的参考值，该值越大，可信度越高，将弱可信的实体转化为强可信的候选实体，具体公式如下：

$$confidence_{e_i}=\frac{\alpha * freq_{e_i}+(1-\alpha) * sim(x_i,y_i)}{\sum_{i=1}^{n}\alpha * freq_{e_i}+(1-\alpha) * sim_{e_i}} \tag{3}$$

其中，α 表示两者的相对权重，用于进一步优化实体排序情况；$confidence_{e_i}$ 表示实体 e_i 的可信度。

4 实验

4.1 不同搜索引擎所需抓取的平均页面数

实验随机选取了 20 个邮箱地址，并用这些邮箱地址在不同搜索平台上进行搜索，获取了 Top30 页面，Baidu 共有 593 条记录，Bing 有 597 条记录，Google 搜索反馈页面有 621 条记录，获得反馈网页记录中的重复页面数，用重复数与页面总数的比值来计算重复率，结果如表 1 所列。

表 1 三大搜索引擎的重复率

Table 1 Repeatrate of three search engines

重复率	Baidu	Bing	Google
Baidu	1	0.42	0.46
Bing	0.42	1	0.63
Google	0.46	0.63	1

由表 1 可以明显看出，不同搜索引擎关于同一关键词反馈的信息是不同的，且它们的重复率也较低，因此同时抓取 3 个搜索引擎中的反馈结果是非常有必要的，有利于获取更多的有效候选实体。

在邮箱候选集中，随机抽取 $m(m=20)$ 个邮箱，使用爬虫获取 Top N 的页面，并人工审核不同的 N 时含有实体的页面数(本实体是一般意义的实体，不一定是邮箱所对应的实体)。最终检测 K 次采样后前 N 个网页中实体数的平均值。

表 2 列出了随机抽取的 20 个邮箱地址采用不同搜索引擎时得到的 Top N 页面中含有的实体总数的平均值。由表 2 可以看出，随着页面数的增加，不同搜索引擎抓取的含有实体的页面数在不断减少，增长趋势减缓。

表 2 Top N 页面中的平均实体数

Table 2 Average number of entities in Top N page

(单位：个)

抓取页面数	Baidu	Bing	Google
Top 10	7.7	5.6	12.4
Top 20	14.4	9.6	18.3
Top 30	20.2	11.2	26.8
Top 40	25.3	17.3	32.6
Top 50	30.3	21.4	37.1

图 4 中，横坐标为排序前 N 的网页页面数，左纵坐标为 Top N 页面中的平均实体数(单位：个)，右纵坐标为平均有效实体的占比。在 Top 30 页时，平均有效占比达到了最大值，即 89%，后面的页面抓取到的实体多为与邮箱地址相关度低的无效实体。因此，在实验中采集数据时，选择获取三大引擎的 Top30 的页面，以降低实体的冗余程度，减少实验流程中的无效计算，提高系统的计算效率。

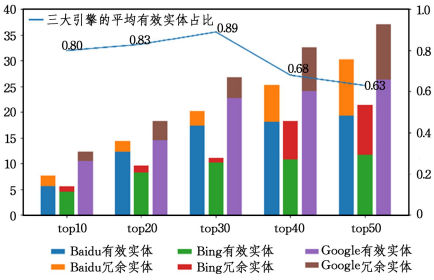

图 4 有效实体占总实体的堆积图

Fig. 4 Effective entities' stack in total entity

4.2 命名实体识别模型的选择与调优

本文利用中文模型训练的语料库是维基百科中文语料库，用英文模型训练的语料库是英语维基百科语料库。使用五折交叉验证法进行语料训练，分别使用 Ling 等[18]的 CRF 模型、Graves 等[19]的 Bi-LSTM 模型和 Lample 等[4]的 Bi-LSTM-CRF 模型进行训练。

本文选取 20 个邮箱地址，并对用这些邮箱地址搜索前 30 个页面得到的实体进行了人工标注，共有 522 个地名、771 个机构名、643 个人名，实验结果如表 3 所列。

表 3 不同模型对实体识别的实验结果对比

Table 3 Comparison of experimental results of different models for entity recognition

模型	Precision	Recall	F1-score
Bi-LSTM-CRF	**0.895**	**0.886**	**0.891**
Bi-LSTM	0.745	0.762	0.753
CRF	0.653	0.662	0.657

从表 3 可以看出，Bi-LSTM-CRF 模型在对人名、地名和机构名进行识别时精确率、召回率以及 F1 值都比其他模型更高，分别达到了 0.895，0.886，0.891。

选取 Bi-LSTM-CRF 模型后，实验特别针对模型网络参数进行精心调整，当 $dropout=0.5$，$batch\ size=512$，$learning\ rate=0.017$ 时，模型的效果最优。

在实体识别任务中，本文使用了 3 个评价指标：精准率、召回率和 F1。

精准率 Precision：正确预测为正例的样本占全部预测为正例的样本的比例。

$$Precision=\frac{TP}{TP+FP} \tag{4}$$

召回率 Recall：正确预测为正例的样本占全部正样本的比例。

$$Recall=\frac{TP}{TP+FN} \tag{5}$$

F1-score：precision 和 recall 调和均值的 2 倍。

$$F1=\frac{2*Precision*Recall}{Precision+Recall} \tag{6}$$

其中，TP(True Positive)指预测为正例且实际也为正例的样本数；FP(False Positive)指预测为正例而实际为负例的样本数；TN(True Negative)指预测为负例而实际为负例的样本数；FN(False Negative)指预测为正例而实际为负例的样本数。

4.3 实体融合算法对比

本实验随机选取了 100 个邮箱，并针对这 100 个邮箱地址人工整理出现实中对应的所有者的真实信息，其中包括邮箱所有者的真实姓名、单位和地址等。

系统将三大搜索引擎中获取的实体内容整合在一起，计算可信度权重，最后按可信度权重的大小进行排序输出。表 4 列出了经过TF-Levenshtein 融合算法计算的结果。

表 4 融合算法的实验结果

Table 4 Experimental results of fusion algorithm

(单位：%)

算法	MRR
词袋模型	68.24
编辑距离相似度模型	78.56
最长公共子序列[18]	72.34
融合模型	**91.32**

实验选取每个邮箱的前 10 个候选实体，若前 10 个实体中包含邮箱的真实主人，则再判断排在第一位的候选实体是不是邮箱相关的有效实体。其中，有效实体指与邮箱地址关联度接近的实体，包括邮箱主人或与邮箱主人合作过、有过联系的实体；而无效实体指那些识别出来与邮箱地址无关的实体。例如，邮箱 tmcrady@gmial.com 的实体有麦迪、姚明、Vince Carter 和 Donald Trump，假设邮箱是麦迪的，那么 Donald Trump 这个实体与这个邮箱的相关性是最弱的，可信度权重最小，因此排在候选实体的末端。实验以搜索算法中的有效性评价标准为评判准则(即 Mean Reciprocal Rank)：用第一候选实体是否为邮箱的主人作为算法有效性的准则，进行算法优劣性的考察与调优。

在 100 个邮箱中每个邮箱的前 10 个候选实体集中，包含

真实主人的有效候选实体集的比例为 78.43%，这些弱可信的有效实体集在经过融合算法计算排序后，邮箱的真实主人能够排列到第一候选人位置的比例为 91.32%，其中 MRR 的准确率与式(3)中的加权参数 α 的调整有关，为此，本文也做了一系列实验，验证 α 的最佳值，实验结果如图 5 所示。

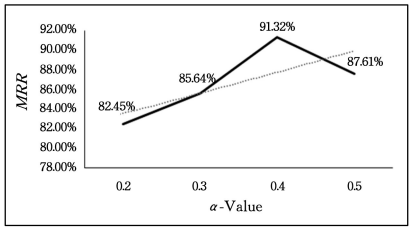

图 5 α 值与 MRR 准确率的关系

Fig. 5 Relationship between α-value and MRR-accuracy

从图 5 可以看出，第一候选实体为邮箱真实主人的 MRR 值随着 α 值的变化而波动。当 $\alpha=0.4$ 时，MRR 值达到了 0.9132。其中出现错误的部分是因为邮箱的命名并不是按照姓名的单词缩写，在后面的研究中，我们将进行更深入的探索，融入其他的算法，进一步提高所提方法的准确度。

结束语 针对海量、多源弱可信数据，本文提出了一种结合自然语言处理方法以及相似度计算的融合方法来分析弱可信数据，成功地从弱可信数据中整理出了强可信的命名实体，其中包括人名、地名和机构名等。

从实验中可以看出，搜索引擎中反馈的网页数据在 30 条过后基本上不再有参考价值。研究发现，多数人的邮箱地址命名方式与自己的姓名有关。中文由于没有天然间隔，在进行 Bi-LSTM-CRF 模型训练时需要先进行分词，因此在实体抽取前进行中英文分离是非常有必要的。中英文分开训练、抽取有利于提高中英文不同实体的识别。由于中文多是象形文字，无法与邮箱地址直接关联，在实体融合排序前有必要将相关实体进行拼音的转化，事实证明，经过融合频率和编辑距离算法排序后的实体更加合理可信。

由排序后的实体发现，针对没有用有关姓名命名邮箱的用户的分析准确率还是不够，这是未来需要改善的地方。另外，命名实体中除了邮箱主人外，其他的实体多数与他有关联，直接预示着他们在现实中可能是同事、朋友或者合作伙伴等，因此用户的社交网络关系也将是未来进一步研究的方向。

参 考 文 献

[1] GUO J, HUANG C S. Research progress of information overload in foreign network environment [J]. Information Science, 2018, 323(7): 172-178.

[2] GRIDACH M. Character-level neural network for biomedical named entity recognition[J]. Journal of biomedical informatics, 2018, 70(6): 85-91.

[3] CLICHE M. BB_twtr at SemEval-2017 task 4: twitter sentiment analysis with CNNs and LSTMs[J]. arXiv: 2017, 1704.06125.

[4] LAMPLE G, BALLESTEROS M, SUBRAMANIANS, et al. Neural architectures for named entity recognition[J]. arXiv: 2016, 1603.01360.

[5] BIKEL D M, SCHWARTZ R, WEISCHEDEL R M. An algorithm that learns what's in a name[J]. Machine Learning, 1999, 34(1/2/3): 211-231.

[6] LAFFERTY J, MCCALLUM A, PEREIRA F C N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[J]. Machine Learning, 2001, 7: 301-311.

[7] ZHOU G D, SU J. Named entity recognition using an HMM-based chunk tagger[C]// Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2002: 473-480.

[8] MCCALLUM A, LI W. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons[C]// Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4. Association for Computational Linguistics, 2003: 188-191.

[9] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[10] COLLOBERT R, WESTON J, BOTTOU L, et al. Natural language processing (almost) from scratch[J]. Journal of Machine Learning Research, 2011, 12(8): 2493-2537.

[11] HUANG Z, XU W, YU K. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv: 1508.01991, 2015.

[12] QIU Y F, TIAN Z P, JI W Y, et al. An Efficient Method for Detecting Similar Repetitive Records[J]. Chinese Journal of Computers, 2001(1): 69-77.

[13] TAN M C, CAO J J. A method for calculating string similarity with multiple editing distances [J]. Application Research of Computers, 2010, 27(12): 4523-4525.

[14] YE X. Approximate longest common substring matching and optimization algorithm for editing distance constraints [D]. Northeastern University, 2014.

[15] ZHANG C Z, MA S T, JIE C Y, et al. Study on Parallel Web Page Recognition Based on Beneficial URL Matching Mode Credibility[J]. Journal of Chinese Information Processing, 2018, 32(3): 91-100.

[16] YE X M, MAO X Q, XIA J C, et al. Improvement of text classification TF-IDF algorithm[J]. Computer Engineering and Applications, 2019, 55(2): 104-109.

[17] ZHENG K, OUYANG L Y, LIN Q, et al. Research on LCS Algorithm and Edit Distance Algorithm[J]. Information Communication, 2015(5): 22-23.

[18] LING W, LU T, MARUJO L, et al. Finding function in form: compositional character models for open vocabulary word representation[J]. Computer Science, 2015, 11: 1899-1907.

[19] GRAVES A, SCHMIDHUBER J. Framewise phoneme classification with bidirectional LSTM networks[C]// IEEE International Joint Conference on Neural Networks. IEEE, 2005, 40(7): 1482-1488.

SUN Guo-zi, born in 1972, Ph.D, professor, is a senior member of China Computer Federation. His main research interests include cyberspace security, digital forensics, and blockchain.