DOI:10.3969/j.issn.1671-0673.2021.04.008

一种基于用户偏好分析和论坛相似度计算的改进 LFM 推荐算法

巨星海,周 刚

(信息工程大学,河南 郑州 450001)

摘要:针对用户的偏好推荐需求,提出一种改进的 LFM 算法 BBLFM 算法,通过引入隐含特征将稀疏的相关矩阵分解为两个相对稠密的矩阵,减少了空间复杂度,同时实现 LFM 的隐语义分析功能,深入挖掘了用户的潜在特征,提高了推荐的准确性。具体地,设计了一种基于 BM-25 的精确用户关注点查找与权重赋值方法,同时引入软概率情感分析方法的结果,合成出一种基于语义的标签体系。此外,还构建了一个基于 BERT 的用户偏好分析网络,根据用户曾经浏览或点击的历史论坛数据,来为用户画像,给出用户的主题偏好。在真实的百度贴吧数据集上进行的对比实验结果,表明算法在推荐准确性上优于比较的算法。

关键词:舆情分析;偏好分析;LFM 算法;BERT;情感分析

中图分类号:TP391.41 **文献标识码**:A **文章编号**:1671-0673(2021)04-0433-05

Improved LFM Recommendation Algorithm Based on User Preference Analysis and Forum Similarity Calculation

JU Xinghai, ZHOU Gang

(Information Engineering University, Zhengzhou 450001, China)

Abstract: In this paper, we propose an improved LFM algorithm for user preference recommendation, which decomposes the sparse correlation matrix into two relatively dense matrices by introducing hidden features. Our model greatly reduces the space complexity while realizing the hidden semantic analysis function of LFM, fully excavating the potential characteristics of users, and improving the accuracy of recommendation. Specifically, we design a precise user focus search and weight assignment method based on BM-25, and at the same time introduce the results of the soft probability sentiment analysis method to synthesize a semantic-based labeling system. In addition, we have also built a BERT-based user preference analysis network, which can profile the user and give the user's theme preference based on the historical forum data that the user has browsed or clicked on. Finally, we conduct comparative experiments on the real Baidu Tieba dataset. The experimental results show that our algorithm is superior to the compared algorithms in terms of recommendation accuracy.

Key words: public opinion analysis; preference analysis; LFM algorithm; BERT; sentiment analysis

收稿日期:2021-04-08;**修回日期**:2021-04-20

基金项目:国家自然科学基金资助项目(61702549)

作者简介:巨星海(1991-),男,博士生,主要研究方向为大数据分析、文本分析。

用户的关注点分析,作为舆情分析中的一个重要方向,近些年受到越来越多的关注。大量研究者对其进行了广泛研究。网络论坛作为社交媒体中的一个重要组成部分,是网民们获取信息、进行交流、了解时事热点的重要渠道。网络论坛中的用户在立场、价值观和社会需求千差万别,他们围绕的关注点与自身抒发的感受也往往存在较大的区别。针对同样的事件,不同用户间的关注点也可能存在明显差异。基于语义对网络论坛用户的关注点进行潜在的关联性分析,找出不同用户之间隐含的关注点异同,有助预测不同用户对时下热点事件实体的偏好,对网络舆情分析工作中的立场分析、用户画像等任务有着重要的意义和作用。

传统的隐含分析算法,包括 word2vec 词汇-向量算法[1]、LSA/LSI(Latent Semantic Analysis/Indexing)潜在语义分析[2-3],LDA(Linear Discriminant Analysis)降维线性分析算法[4]等。这些方法更多考虑将文本数据中的特征词向量化后进行分析,在大规模、高稀疏性的自由文本条件下表现不佳[5]。为了应用网络论坛中广泛存在的用户实体对应关系,进一步提高对用户在文本中隐含关注点的推荐准确性,本文基于 LFM 隐语义分析算法,构建了一个新的推荐算法框架。

LFM 隐语义算法是文献[6]最早提出的,它主要通过发掘用户与其所关注的实体间的隐含联系,从而判断出两者之间潜在的关注关系并做出推荐。针对 LFM 算法在其最初研究领域中依靠选取种子遍历图结构,因而导致效率不足的问题,研究者们已经做出了一些改进,如文献[7]通过选取“团体”集群代替种子,从而改善算法稳定性,并降低了算法的时间复杂度。文献[8]提出非负矩阵分解技术,对用户和项目评分矩阵进行规范,同时进行简化和降维,从而进一步提高算法效率。文献[9]针对 LFM 算法在寻优速率和准确性上的不足,通过提出带冲量的学习法和混合学习法,使算法效果得到一定的提升。LFM 算法作为一种主要应用于网络购物等平台的推荐算法,在文本分析领域的移植和研究工作还不深入,在面临真实文本数据特有的复杂性和稀疏性问题上,LFM 算法的性能也有进一步的提升空间。

本文在深入分析网络论坛文本数据的基础上,引入了特征词的 BM-25 分数作为用户对所关注实体的关注权重,并辅以情感分析算法,为特征关系增加了立场属性,提出了一种新的基于情感分析与特征权重的改进 LFM 推荐算法。另外,以所提出的算法为基础,本文选取百度贴吧中的部分公开文本数据作为实例,对百度贴吧用户的关注点相似度进行了分析,并对用户隐含的潜在关注点进行预测,实验结果表明相比于传统算法,本文的方法在推荐准确率上有较大的提升。

本文的主要贡献有:

(1)设计了一种基于 BM-25 的精确用户关注点查找与权重赋值方法,同时引入软概率情感分析方法的结果,合成出一种基于语义、能够更加直观表达文本情感强度的标签体系;

(2)构建了一个改进的基于 BERT 的用户偏好分析网络,能够根据用户曾经浏览或点击的历史论坛数据,来为用户画像,给出用户的主题偏好;

(3)提出了一个改进的 LFM 算法,通过引入一个隐含特征,从而将稀疏的相关矩阵分解为两个相对稠密的矩阵,极大减少了空间复杂度,同时实现 LFM 的隐语义分析功能,确保算法在文本数据前提下能够将网络信息充分利用,提高了算法的准确性。

1 改进的 LFM 推荐算法框架

1.1 总体框架和流程

本文的目的是为指定用户精准推荐其最感兴趣的帖子,为此构建了一种基于用户偏好分析和论坛主题相关性计算的改进 LFM 推荐算法,称为 BBLFM 算法,算法整体框架如图 1 所示。本文的模型主要分为 3 个基本模块:①用户偏好分析模块,根据用户曾经浏览或点击的历史论坛数据,来为用户画像,给出用户的主题偏好;②相关性计算模块,在综合分析论坛及其帖子数据后,利用改进的 BM-25 算法计算给出帖子与论坛之间的相关关系;③用户推荐模块,基于用户画像,利用相关性计算模块得出的帖子与论坛之间的相关性得分,来为用户精准推荐其感兴趣帖子。

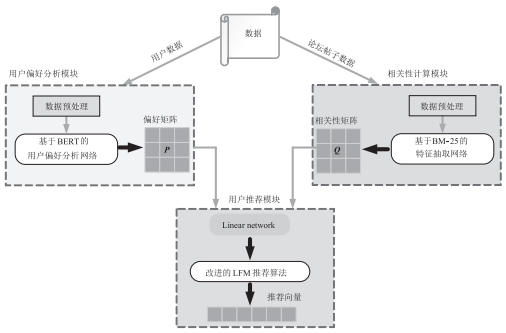

图 1 整体框架

1.2　LFM 推荐算法

隐语义模型(Latent Factor Model, LFM)推荐算法是协同过滤(Collaborative Filtering, CF)推荐算法的一种,可以基于用户的历史行为数据进行相关内容推荐。协同过滤算法中基于用户或基于推荐内容的推荐算法需要维护一张用户与推荐内容的相关矩阵,当用户或推荐内容数量很多时,矩阵维度很大,会占用大量存储空间。但实际上相关矩阵是稀疏的,LFM 算法引入一个隐含特征,从而可以将稀疏的相关矩阵分解为两个相对稠密的矩阵,极大减少了空间复杂度。LFM 计算思想如图 2 所示。

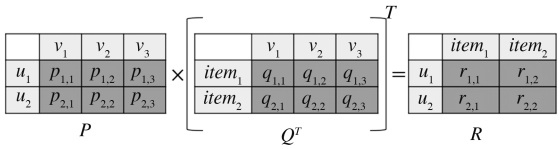

图 2　LFM 计算示意图

本文借助 LFM 算法的思想向用户推荐感兴趣的帖子,将用户记为 u,推荐内容帖子记为 *item*,LFM 算法通过以下模型得到用户 u 对帖子 *item* 的兴趣:

$$preference(u,i)=\boldsymbol{P}\times\boldsymbol{Q}^{T}=\sum_{v=1}p_{u,v}q_{v,i} \tag{1}$$

式中 v 为隐含特征,$p_{u,v}$ 为用户 u 对隐特征 v 的关注度,$q_{v,i}$ 为隐特征 v 与帖子之间的权重关系。然后根据 $preference(u,i)$ 值的大小,即用户 u 对帖子 *item* 的感兴趣程度,向其推荐最感兴趣且未浏览过的帖子。

传统 LFM 算法中需要根据优化目标函数来迭代计算获取 $\boldsymbol{P}$ 和 $\boldsymbol{Q}$ 的值,其中 $r_{u,i}$ 为用户 u 对帖子 *item* 的关注程度。计算过程的时间复杂度为:$O(N)=D\times F\times N$,其中 D 为用户数量,F 为隐特征的数量。在本文描述的场景下,用户和帖子的数量巨大,为了得到较为准确的 $\boldsymbol{P}$ 和 $\boldsymbol{Q}$ 的值,需要花费大量的时间进行训练。

为了节省训练时间,不选择矩阵分解的方式而是通过直接计算获取 $\boldsymbol{P}$ 和 $\boldsymbol{Q}$。将隐含特征定义为帖子主题,那么矩阵 $\boldsymbol{P}$ 中的元素 $p_{i,j}$ 实际上表示第 i 个用户对第 j 个主题的关注度;矩阵 $\boldsymbol{Q}$ 中的元素 $q_{i,j}$ 实际上表示第 i 个帖子与第 j 个主题的相关程度。

$$C=\sum_{(u,i)\in K}(r_{ui}-\hat{r}_{ui})^{2}=\sum_{(u,i)\in V}(r_{u,i}-\sum p_{u,v}q_{v,i})^{2}+\lambda\|p_u\|^{2}+\lambda\|q_i\|^{2} \tag{2}$$

下面在 1.3 节和 1.4 节中分别介绍 $\boldsymbol{P}$ 和 $\boldsymbol{Q}$ 矩阵的计算方式。获取 $\boldsymbol{P}$ 和 $\boldsymbol{Q}$ 矩阵后,计算 $preference(u,i)$ 即可得到用户对帖子的感兴趣程度,进而进行相关推荐。

1.3　用户偏好分析模块

定义 $\boldsymbol{U}=\{u_1,u_2,\cdots,u_{|U|}\}$ 是给定的用户集合,$\boldsymbol{V}=\{v_1,v_2,\cdots,v_{|V|}\}$ 是不同主题的论坛集合,$S_u=\{v_1^{(u)},v_2^{(u)},\cdots,v_t^{(u)},\cdots,v_n^{(u)}\}$ 是用户在某一段时间内浏览或点击的论坛的历史序列。偏好分析的目的就是根据用户的历史数据 S_u 来学习用户对不同主题的论坛的偏好,从而给出下一时刻用户感兴趣的主题分布。例如,有 4 个不同主题的论坛:电影、游戏、体育、动漫,已知某一用户在上周停留或关注最多的论坛依次为:游戏→动漫→游戏→电影→动漫→游戏,从用户的历史轨迹能够容易看出用户的属性为“偏宅”,并且在游戏之后,会更倾向于选择看电影或动漫进行放松,而不是户外活动。因此,在下一段时间,给用户优先推荐电影或动漫相关的内容更容易吸引他的注意。

但是,如何让机器去理解和学习用户的行为轨迹,从而做出判断依然是一个难点。过去,基于循环神经网络(RNN)[10]的模型在解决序列问题上取得了突破,有着很好的表现。在本文的问题背景下,不仅要考虑相邻序列之间的内在关系,还要从全局上去挖掘用户的潜在画像,基于 RNN 的模型不能完全胜任任务,而 BERT 预训练模型是一种双向模型,可以更完整地捕捉文本特征。因此,在用户偏好分析模块,基于预训练模型 BERT 搭建了一个深度神经网络来挖掘用户的深层的全局属性,进而为后续的推荐任务提供支撑。网络的整体框架如图 3 所示。

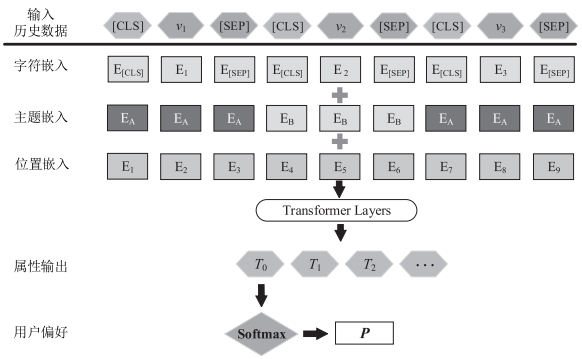

图 3　基于 BERT 的用户偏好分析模型

传统的 BERT 模型[11-12]将输入看成一个整体来进行嵌入学习,在本文的模型中为输入增加了一个主题嵌入,来区分不同主题的输入。为了去表示

独立的论坛 $v_1^{(u)}$,在序列 S_u 中每一项的前面插入“[CLS]”符号,在结尾插入“[SEP]”来区分不同的论坛输入。并且对于每一个 $v_i^{(u)}$ 都基于 i 的奇偶性为它分配一个主题嵌入(topicembedding) E_A 或者 E_B。例如,给定 $S_u=\{v_1^{(u)},v_2^{(u)},v_3^{(u)},v_4^{(u)},v_5^{(u)}\}$,分配的主题嵌入符号为[$E_A,E_B,E_A,E_B,E_A$]。随后,将合并后的嵌入向量输入到由 Transformer 块[13]构成的多层网络中,得到每一个 $v_i^{(u)}$ 的向量表示 $\boldsymbol{T}_i$,实际上就是符号“[CLS]”的嵌入,用最开始的向量 $\boldsymbol{T}_0$ 来表示用户的全局属性。最后,将 $\boldsymbol{T}_0$ 输入到一个 softmax 网络中,得到用户对于不同主题的偏好向量 $\boldsymbol{P}=[s_1,s_2,\cdots,s_{|V|}]$,其中 s_i 表示用户 u 对主题 i 的偏好程度。

1.4 相关性计算模块

定义 $\boldsymbol{I}=\{item_1,item_2,\cdots,item_n\}$ 是给定的帖子集合,$\boldsymbol{V}=\{v_1,v_2,\cdots,v_m\}$ 是不同主题的论坛集合,$K^{(m)}=\{k_1,k_2,\cdots,k_t\}$ 是描述主题 v_m 的关键词。根据每个主题中所包含的帖子计算词的 TF-IDF 值[14],选取部分 TF-IDF 值高的词作为对应主题的关键词。

选择 BM25 算法[15]来计算每个帖子与主题的相关程度,BM25 算法能够计算代表主题的每个关键词与帖子的相关性得分,并进行加权求和作为帖子与主题的相关性得分。BM25 算法的一般公式为

$$Score(Q,item)=\sum_i^t W_i * R(k_i,item) \tag{3}$$

式中 W_i 是关键词 k_i 的权重,$R(k_i,item)$ 为关键词 k_i 与帖子 $item$ 的相关性得分。

采用关键词 k_i 的 IDF 值作为权重,即

$$W_i=IDF(k_i)=log\frac{|I|-n(k_i)+0.5}{n(k_i)+0.5} \tag{4}$$

式中 $|I|$ 为所有帖子数量,$n(k_i)$ 为包含关键词 k_i 的帖子数量。从直观上很容易理解,对于给定的帖子集合,包含关键词 k_i 的帖子越多,k_i 的区分度就越低,权重也就越低。

关键词 k_i 与帖子 $item$ 的相关性得分可以根据公式计算,其中 b_1,b_2 为调节因子,$l(item)$ 为帖子 $item$ 的文本长度,$avgl(I)$ 为所有帖子的平均长度,f_i 为关键词 k_i 在主题中出现的频率,用关键词 k_i 归一化后的 TF-IDF 值进行表示。

$$R(k_i,item)=\frac{f_i\cdot(b_1+1)}{f_i+b_1\cdot(1-b_2+b_2\cdot\frac{l(item)}{avgl(I)})} \tag{5}$$

根据式(3)计算每个帖子 $item$ 与主题 v_j 的相关性得分,即可得到矩阵 $\boldsymbol{Q}$。

2 实验

2.1 实验环境与数据集

实验环境如下所示:硬件环境:Intel(R) Xeon(R) Gold 6136 CPU @ 3.00GHz128.0RAM 软件环境:相关代码使用 Python 进行编写。

为了更全面、更准确的评价算法的性能,从百度贴吧收集了 8 000 名用户的发帖、跟帖数据。共获取 114 056 条帖子,平均每个用户发帖 14.26 条。根据贴吧分类将各类帖子分为娱乐、综艺、追剧、电影、体育、小说、生活、闲趣、游戏、动漫、高校、地区、人文等 13 个主题,各主题帖子数量如表 1 所示。

表 1 帖子主题及对应数量

主题	数量
娱乐	9 660
综艺	7 318
追剧	9 897
电影	7 524
体育	9 787
小说	9 340
生活	9 682
闲趣	7 023
游戏	9 583
动漫	7 878
高校	9 736
地区	7 514
人文	9 114
合计	114 056

将与每个用户关联帖子的 80%作为训练集,20%作为测试集,分别测试了相关性计算的准确度以及推荐结果的准确度,并将算法效率与经典推荐算法 ItemCF 进行对比。

为了保护用户隐私,实验中出现的用户名均做匿名化处理,文本以词向量的形式展示。

2.2 推荐结果

首先根据各主题中所包含的帖子使用 TF-IDF 算法提取各个主题的关键词,表 2 中列举了每个主题中 TF-IDF 值排名前 5 的关键词。

然后根据公式计算每个帖子与主题的相关性得分,每个帖子与主题的相关性得分以一个 13 维向量表示,向量的每一维代表帖子与一个主题的相关性。比如:[18,4,6,5,4,2,1,0,4,3,2,1,2,6,21,0,7,…,0,0,0,0,1,5,3,2,0,0,0,0,0,0,0,0,0,1]这则帖子计算得到的相关性向量为[0.002,

0.001,0,0,0,0,0.12,0.03,0,0,0.757,0,0.09]。

表 2　主题与对应关键词

主题	关键词
娱乐	好声音、明星、超话、数据、偶遇
综艺	芒果、TVB、综艺、挑战、录制
追剧	电视剧、美剧、人气、背景、烂尾
电影	票房、复联、资源、票房、定档
体育	足球、篮球、NBA、健身、世界杯
小说	历史、穿越、科幻、诛仙、人气
生活	教程、旅游、二手、手工、背包客
闲趣	喵星人、吐槽、星座、搞笑、内涵
游戏	手游、春季赛、赛程、铂金、暴雪
动漫	漫画、声优、银魂、柯南、路飞
高校	高考、考研、大学、985、自习室
地区	中国、城市、东营、郑州、同城
人文	历史、朝代、惊悚、多肉、文玩

为了进一步验证相似性计算的有效性,将相关性最大的主题作为帖子的预测主题分类,并与其原主题分类做对比,结果如图 4 所示。图中对角线部分表示预测正确的数量,实际上根据主题相关性预测结果的准确率能够达到 87%以上,表明采用的相关性计算方法的有效性。

此外,本文对模型推荐的准确性(成功率)进行了评估,采用均值平均精度(Mean Average Precision, MAP)来评估模型的好坏。表 3 展示了网络不同网络深度下,模型在 MAP 上的得分。结果表明,本文的模型优于比较的模型,并且随着网络深度的增加本文模型的性能也会上升,当网络深度达到 6 层时,模型的性能达到峰值。再往后反而下降,可能的原因是随着网络越深训练的参数越复杂,在有限的数据集和时间内容易导致训练不充分,影响最终的预测。

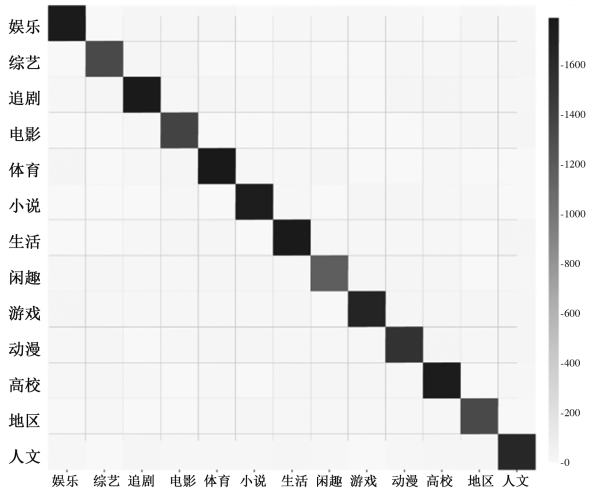

图 4　主题分类混淆矩阵

表 3　不同算法 MAP 得分比较

算法	MAP
ItemCF	0.45
UserCF	0.51
3-BBLFM	0.54
4-BBLFM	0.60
5-BBLFM	0.63
6-BBLFM	0.67
7-BBLFM	0.52

3　结束语

本文构建了一种基于用户偏好分析和论坛主题相关性计算的改进 LFM 推荐算法,称为 BBLFM 算法。构建了基于 BERT 的用户情感偏好模块和基于 BM25 的论坛相关性度量模块,用于分别计算 LFM 算法中的两个矩阵。实验证明,相较于传统的依赖矩阵分解的 LFM 算法,本文提出的 BBLFM 算法在效率上有明显提高,此外相较于其他推荐算法,BBLFM 算法有更高的推荐成功率。

参考文献:

[1] MIKOLOV T,CHEN K,CORRADO G,et al. Efficient estimation of word representations in vector space[EB/OL]. (2013-9-7)[2021-4-1]. https://arxiv.org/abs/1301.3781v3.

[2] PAPADIMITRIOU C H,RAGHAVAN P,TAMAKI H,et al. Latent semantic indexing:a probabilistic analysis[J]. Journal of Computer and System Sciences,2000,61(2):217-235.

[3] LANDAUER T K. Latent semantic analysis[M]. Berlin Heidelberg:Springer,2010.

[4] RIFFENBURGH R H,CLUNIES-ROSS C W. Linear discriminant analysis[J]. Chicago,2013,3(6):27-33.

[5] 杨焱,孙铁利,邱春艳. 个性化推荐技术的研究[J]. 信息工程大学学报,2005,6(2):84-87.

[6] LANCICHINETTIA,FORTUNATOS,KERTÉSZ J. Detecting the overlapping and hierarchical community structure in complex networks[J]. New Journal of Physics,2009,11(3):033015.

[7] 薛磊,唐旭清. 基于中心团的重叠社区检测算法[J]. 计算机科学,2020,47(8):157-163.

[8] 范敏敏. 非负矩阵分解与聚类方法在个性化推荐系统中的应用研究[D]. 南昌:华东交通大学,2012.

[9] 陈晔,刘志强. 基于 LFM 矩阵分解的推荐算法优化研究[J]. 计算机工程与应用,2019,55(2):116-120,167.

(下转第 449 页)

tion of knowledge (SoK): a systematic review of software-based web phishing detection[J]. IEEE Communications Surveys & Tutorials,2017,19(4):2797-2819.

[3] GUPTA B B, TEWARI A, JAIN A K, et al. Fighting against phishing attacks: state of the art and future challenges[J]. Neural Computing and Applications, 2017, 28(12):3629-3654.

[4] CANALI D, COVA M, KRUEGEL C, et al. A fast filter for the large-scale detection of malicious web pages[C]// Proceedings of the 20th international conference on World wide web, 2011:197-206.

[5] PRIYA M, SANDHYA L, THOMAS C. A static approach to detect drive-by-download attacks on webpages[C]// 2013 International Conference on Control Communication and Computing (ICCC), 2013:298-303.

[6] SHENG S, WARDMAN B, WARNER G, et al. An empirical analysis of phishing blaclists[C]//The 6th Conference on Email and Anti-Spam, 2009:59-78.

[7] ALEROUD A, ZHOU L N. Phishing environments, techniques and countermeasures: a survey[J]. Computers & Security, 2017, 68:160-196.

[8] 彭成维,云晓春,张永铮,等. 一种基于域名请求伴随关系的恶意域名检测方法[J]. 计算机研究与发展, 2019,56(6):1263-1274.

[9] SAHINGOZ O K, BUBER E, DEMIR O, et al. Machine learning based phishing detection from URLs[J]. Expert Systems With Applications, 2019, 117:345-357.

[10] ZHANG M, XU B Y, BAI S, et al. A deep learning method to detect web attacks using a specially designed cnn[C]//International Conference on Neural Information Processing, 2017:828-836.

[11] 崔艳鹏,刘咪,胡建伟. 基于 CNN 的恶意 Web 请求检测技术[J]. 计算机科学,2020,47(2):281-286..

[12] BAHNSEN A C, BOHORQUEZ E C, VILLEGAS S, et al. Classifying phishing URLs using recurrent neural networks[C]//2017 APWG Symposium on Electronic Crime Research (eCrime), 2017:1-8.

[13] LE H, PHAM Q, SAHOO D, et al. URLNet: learning a URL representation with deep learning for malicious URL detection[C]//Research Collection School Of Computing and Information Systems, 2018:1-13.

[14] GILMER J, SCHOENHOLZ S S, RILEY P F, et al. Neural message passing for quantum chemistry[C]//Proceedings of the 34th International Conference on Machine Learning, 2017:1263-1272.

(编辑:高明霞)

(上接第 437 页)

[10] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks[C]//Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 2, 2014:3104-3112.

[11] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017:6000-6010.

[12] LIU Y H, OTT M, GOYAL N, et al. RoBERTa: A robustly optimized BERT pretraining approach[EB/OL]. (2019-12-24)[2021-4-1]. https://openreview.net/forumid=SyxS0T4tvS.

[13] DEVLINJ, CHANGM W, LEE K, et al. BERT: Pre-Training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1, 2019:4171-4186.

[14] WU H C, LUK R W P, WONG K F, et al. Interpreting TF-IDF term weights as making relevance decisions[J]. ACM Transactions on Information Systems, 2008, 26(3):1-37.

[15] ROBERTSON S, ZARAGOZA H. The probabilistic relevance framework: BM25 and beyond[J]. Foundations and Trends in Information Retrieval, 2009, 3(4): 333-389.

(编辑:刘彦茹)