DOI: 10.11992/tis.201809024
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.TP.20181225.1603.004.html

基于 Hadoop 的大规模网络安全实体识别方法

秦娅[1,2], 申国伟[1,2], 余红星[1,2]
（1. 贵州大学 计算机科学与技术学院, 贵州 贵阳 550025; 2. 贵州大学 贵州省公共大数据重点实验室, 贵州 贵阳 550025）

摘　要: 随着大数据时代的到来, 如何从多源异构数据中准确地识别网络安全实体是构建网络安全知识图谱的基础问题。因此本文针对网络安全相关文本数据, 研究支持海量网络数据的安全实体识别算法, 为构建网络安全知识图谱奠定基础。针对海量的文本类网络数据中安全实体的高效精准抽取问题, 本文基于 Hadoop 分布式计算框架提出改进的条件随机场 (conditional random fields, CRF) 算法, 对数据集进行有效分割, 实现安全实体的高效准确识别。在大规模真实网络数据集上的实验证明, 本文提出的算法达到了较高的网络安全实体识别准确率, 同时提高了识别的效率。

关键词: 大数据; 异构数据; 网络安全; 知识图谱; 安全实体; 实体识别; 网络数据; Hadoop; CRF 算法
中图分类号: TP391.0　　**文献标志码:** A　　**文章编号:** 1673−4785(2019)05−1017−09

中文引用格式: 秦娅, 申国伟, 余红星. 基于 Hadoop 的大规模网络安全实体识别方法 [J]. 智能系统学报, 2019, 14(5): 1017–1025.
英文引用格式: QIN Ya, SHEN Guowei, YU Hongxing. Large-scale network security entity recognition method based on Hadoop[J]. CAAI transactions on intelligent systems, 2019, 14(5): 1017–1025.

Large-scale network security entity recognition method based on Hadoop

QIN Ya[1,2], SHEN Guowei[1,2], YU Hongxing[1,2]
(1. Department of Computer Science and Technology, Guizhou University, Guiyang 550025, China; 2. Guizhou Provincial Key Laboratory of Public Big Data, Guizhou University, Guiyang 550025, China)

Abstract: In this era of big data, a fundamental problem for constructing network security knowledge graphs is how to efficiently and accurately identify the network security entities present in multi-source heterogeneous data. This study focuses on text data related to network safety and investigate the use of a security entity recognition algorithm that supports massive-network text data, thereby laying a foundation for building the network security knowledge graph. To efficiently and accurately extract the security entities in massive-network text data, we propose an improved conditional random fields (CRF) algorithm based on the Hadoop distributed computing framework to segment data sets effectively, which realize efficient and accurate recognition of security entities. The experimental results reveal that the proposed security entity recognition algorithm achieved a high precision rate on a large-scale real network data set and improved the efficiency of network security entity recognition..

Keywords: big data; heterogeneous data; network security; knowledge graph; security entity; entity recognition; network data; Hadoop; CRF algorithm

近年来, 随着信息技术的快速发展, 逐步进入了大数据[1]时代, 网络空间安全面临全新的挑战, 因此网络威胁情报这一新的安全技术应运而生。威胁情报[2](threat intelligence), 主要是通过大数据、分布式系统或其他特定收集方式收集的用于评估和应用的数据集, 针对一个现存的或新兴的

收稿日期: 2018−09−13.　**网络出版日期:** 2018−12−28.
基金项目: 国家自然科学基金项目 (61802081); 贵州省公共大数据重点实验室开放课题 (2017BDKFJJ024); 贵州省自然科学基金项目 (20161052).
通信作者: 申国伟. E-mail: gwshen@gzu.edu.cn.

威胁，可用于做出相应决定的知识。从 2014 年开始，威胁情报逐渐成为网络安全领域的热点，成为一种新的网络安全技术[3-4]。

当今社会正处于大数据时代，同时具有信息碎片化的特征，从而赋予了网络安全信息海量化与碎片化特点，导致网络威胁情报分析人员很难对信息进行获取和整合。因此，针对网络安全信息的碎片化和海量化的特点，将其进行过滤、分类以及关联，从而形成一个网络安全知识体系，衍生成为网络安全知识图谱。网络安全知识图谱构建的前提就是对信息进行抽取，信息抽取是网络安全知识图谱构建的最为关键的一步，其中最为关键就是网络安全实体识别。

网络安全实体识别是命名实体识别[5]中一种特定领域的实体识别，其目的是对网络安全领域专业的词汇进行分类；而通用领域的命名实体识别，主要识别文本中具有特定意义的实体，主要包括人名、组织名和地名等。目前，常见的是英文网络安全实体识别，针对中文的网络安全实体的识别研究工作很少。Jones 等[6]在 Bootstrapping 算法指导下，实现了网络文本中的安全实体和关系自动识别；Joshi 等[7]实现了一种网络文本数据的信息识别方法，利用 CRF 算法来识别网络安全相关实体及关系；Lal[8]提出了一种基于 SVM 算法的信息识别方法，实现了从网络文本数据中识别网络安全相关概念和术语；Mulwad 等[9]设计了基于 SVM 算法的信息识别系统，检测和识别网络文本中的漏洞与攻击信息。

总的来说，网络安全实体的识别方法主要分为基于规则和基于统计的实体识别方法[10-12]。基于规则的实体识别方法对于较小规模的数据具有效果好和速度快的特点，但是规则的编写十分困难，且移植性较差。基于统计的识别方法利用人工标注语料进行训练，对具体语言特性依赖相对较少，移植性强，主要识别方法有隐马尔科夫模型[13]（hidden Markov mode, HMM）、最大熵模型[14]（maximum entropy markov model, MEMM）和条件随机场模型[15-16]（conditional random fields, CRF）等。

目前，网络安全实体的识别主要存在以下难点：

1) 网络安全实体数量众多且类型多种多样，难以满足自然语言处理领域中的命名实体定义，且不断地会有未登录词作为新的安全实体出现。

2) 网络文本数据中的实体具有不同的结构，比如网络安全实体出现大量的嵌套、别名、缩略词等问题，没有严格的构词规律可以遵循。

3) 在大规模数据条件下，基于机器学习模型的算法运行效率较低，单机上的安全实体识别算法难以满足安全实体识别需求。

针对上述问题，本文提出了基于 Hadoop 的 Map/Reduce 分布式计算框架，提出了与规则相结合的改进 CRF 算法实现对安全实体的高效、准确识别。本文的主要工作包括：

1) 针对网络安全实体识别，对安全实体识别进行问题抽象及形式化描述，给出了基于 Hadoop 的网络安全实体识别框架。

2) 分析网络安全数据中的实体结构特征，给出了网络安全实体识别规则，并进一步提出了改进的 CRF 算法，对算法进行分析。

3) 在真实的数据集上，针对提出的网络安全实体识别方法，结合评测标准进行对比实验，结果表明本文提出的方法在准确率和效率上都有所提高。

综上所述，针对网络安全实体识别问题，本文基于 Hadoop 分布式计算框架提出改进的 CRF 算法，对数据集进行有效分割，解决网络安全实体识别的问题，实现准确识别网络安全实体的意义。

1　问题定义

网络安全威胁情报分析可为复杂网络环境下的网络攻防提供情报支撑。在网络威胁情报分析中，网络数据主要识别黑客组织、单位、漏洞、恶意程序等类型网络安全实体，如图 1 所示。

近日，国家信息安全漏洞共享平台收录了Joomla! com_fields组件存在的SQL注入漏洞（CNVD-2017-06861、对应CVE-2017-8917）。远程攻击者无需任何身份认证，可获取数据库敏感信息，包括管理员登录信息并控制网站后台。

图 1　Web 文本数据中的安全实体识别

Fig. 1　Security entity recognition in web text data

本文重点分析 17 类网络安全实体，图 2 给出了网络安全实体的本体模型[17-18]，通过人工编写的方式构建了网络安全领域的本体模型，通过 JSON 语言实现。该模型是一个基于多维标签的网络安全本体模型，其中多维标签包括来源信息、属性信息、元信息等标签信息。

针对海量的文本数据 $D=\{D_1,D_2,\cdots,D_N\}$，经过预处理和分词后，任意一个文本数据 D_i 形成 M_i 个词，构成词序列 $D_i=\{x_1,x_2,\cdots,x_{M_i}\}$。本文的目标是经过算法处理后，从 D 中抽取 K 类网络安全实体 $E=\{E_1,E_2,\cdots,E_K\}$，且每一类安全实体标记为 $E_i=\{e_i^1,e_i^2,\cdots,e_i^{M_i}\}$。

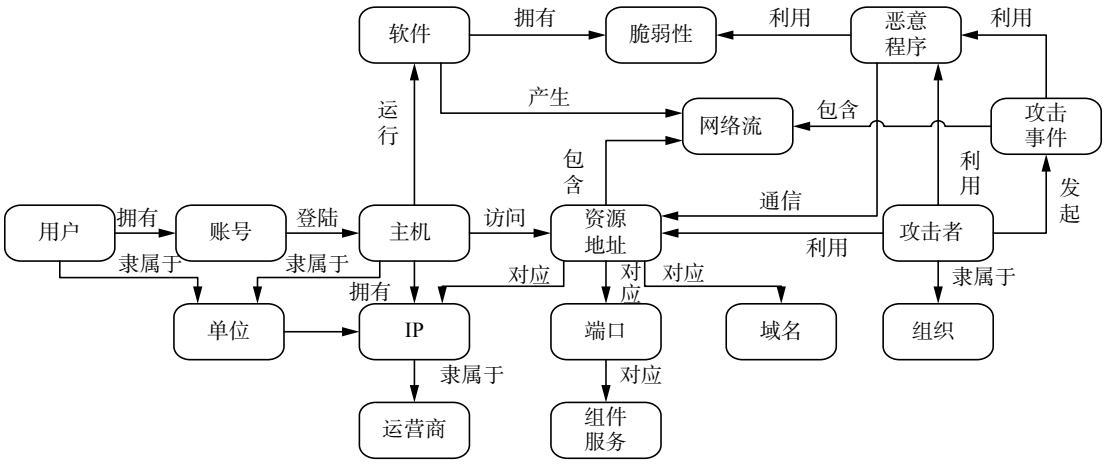

图 2　网络安全实体的本体模型

Fig. 2　Ontological model of network security entity

2　基于 Hadoop 的安全实体识别框架

针对海量的网络安全数据, 本文提出基于 Hadoop 平台的网络安全实体识别框架, 利用 Map/Reduce[19] 分布式计算模型实现高效的数据处理。本文针对大规模数据的网络安全实体识别的工作, 主要运用了 Hadoop 中的 HDFS 和 MapReduce 这两个组件, 对数据进行并行化处理。具体的抽取过程为: 首先, 将预处理的数据存储在 HDFS 中, HDFS 会将这些数据切分成许多独立的小数据块, 存储到若干个节点上, 这些小数据块就会被多个 Map 任务并行处理; 其次, 在 Hadoop 上提交任务进行网络安全实体识别, MapReduce 会为每个任务输入一个数据子集, 同时调用 CRF 算法进行网络安全实体识别, Map 任务生成的结果会继续作为 Reduce 任务的输入; 最后, 由 Reduce 任务输出最后结果, 并写入 HDFS。本文除了将识别出的网络安全实体存入 HDFS, 也将网络安全实体存入图数据库 Neo4j, 为将来构建网络安全知识图谱奠定基础。图 3 为网络安全实体识别的框架图。

1) 数据预处理

本文主要对网页文本数据进行实体识别, 因此在抽取之前要对数据预处理, 处理过程如下:

①使用正则表达式对网页文本进行预处理, 去除网页中的关于 HTML 的标签。

②通过使用 Stanford CoreNLP 提供的分词工具, 将去除标签后的文本数据进行分词。

③构建语料库, 由于网络安全领域没有统一的语料库, 因此在对安全实体识别前, 需要对其构建语料库。对已经分词的文本数据进行实体标注, 特征实体时, 可以通过程序先将所有实体标注为 O, O 表示未识别实体; 然后进行网络安全实体标注, 由人工判断手动标注为 E_n, E_n 表示安全实体。

④最后, 训练网络安全实体模型。在训练过程中, 根据训练工具的格式要求将前面的所有标注后的数据转化成特定的数据格式, 然后利用 CRF 算法进行模型训练。

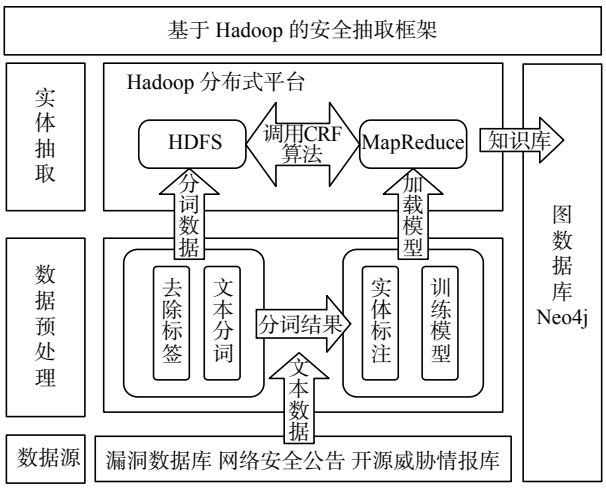

图 3　网络安全实体识别框架

Fig. 3　Network security entity recognition framework

2) 中文网络安全实体识别

本文主要是针对中文网络文本数据的安全实体识别, 数据的输入为中文分词文本数据, 在此之前, 需要利用 CRF 算法进行模型训练, 训练数据主要来自于部分网络安全文本数据。对于中文网络安全实体数据, 进行人工手动标注, 标注完成后, 将其放入训练工具中进行训练, 实现中文网络安全实体模型的建立, 最后通过 CRF 算法实现网络安全实体的识别。

在对网络文本数据进行分词的过程中, 对于网络攻击事件, 一般都是由“动词+名字”组合, 才能完整而清楚描述一次攻击, 如: XSS 跨站脚本

攻击、木马攻击、蠕虫蔓延等。所以在攻击事件名的分词上，本文采用基于规则进行识别，不进行分词，因为分词会导致对攻击事件的整体叙述在语义上描述不清楚，无法理解到底发生了什么样的攻击事件。

3 基于 Hadoop 的 CRF 改进算法

3.1 Hadoop 算法描述

本文采用基于 Map/Reduce 的 CRF 算法并行化处理以缩短识别时间，实现大量数据的网络安全实体识别。MapReduce 模型两个核心函数为 Map 函数和 Reduce 函数，它们的输入都为<key, value>键值对，按一定的映射规则转换为另一个或一批<key,value>。Map 和 Reduce 任务函数有下列通用格式：

$$\text{Map}:< K_1, V_1 >\rightarrow< \text{list}(K_2, V_2) > \tag{1}$$

$$\text{Reduce}:< K_2, \text{list}(V_2) >\rightarrow< K_3, V_3 > \tag{2}$$

式中：Map 函数将输入的数据元素转换成<K_1, V_1>形式的键值对，K_1 和 V_1 的类型是任意的。每一个输入的<K_1, V_1>都会输出一批<K_2, V_2>，<K_2, V_2>是 Map 计算的中间结果，然后输入到 Reduce 函数进行处理，输入形式为<K_2, list(V_1)>，输出为<K_3, V_3>。

在网络安全实体识别的过程中，对于每一个要进行安全实体识别的文本数据，首先将训练好的模型加载进来，然后在 Map 阶段调用 CFR 算法识别网络安全实体，最后在 Reduce 阶段将数据存储到 HDFS 和图数据库 Neo4j。具体的基于 Hadoop 的网络安全实体识别算法如算法 1 所示。

算法 1　基于 Hadoop 的网络安全实体识别核心算法：

输入　网络文本数据 $D_i=\{x_1, x_2, \cdots, x_{M_i}\}$；

输出　网络安全实体 $E=\{E_1, E_2, \cdots, E_K\}$。

1) 调用 Map 函数；

2) 对数据 D 中的文档进行分词并标注，形成观察序列 $D_i=\{x_1, x_2, \cdots, x_{M_i}\}$；

3)CRFClassifier(Di) //调用 CRF 算法识别网络安全实体；

4) 经过模型预测形成标注序列 $Y=\{y_1, y_2, \cdots y_i, y_{M_i}\}$；

5) 调用 Reduce 函数；

6) 根据标注序列进行分类，得到 K 类实体的 $E=\{E_1, E_2, \cdots, E_K\}$；

7) 根据平均值获得最终的 $E_i=\{e_i^1, e_i^2, \cdots, e_i^{M_i}\}$；

8)EntityStore.CreateNeo4j(key)//存储到 Neo4j

实际上，在对网络安全实体识别进行评测时，不需要一个合并的输出，因为合并输出后会影响最后的评测结果，因此可以在对网络安全实体进行评测时省去 Reduce 阶段，那么 Map 函数的输出将不会有中间输出，数据将直接存储至 HDFS。

3.2 CRF 算法描述

在算法 1 中，CRF 是网络安全实体识别的核心，分别对应算法 1 中的 3)~6)步。CRF 又称为马尔可夫随机域，最早由 Lafferty 等[20] 于 2001 年提出，是一种对有序数据进行标注和切分的条件概率模型，拥有 HMM 和 MEMM 的特点。从形式上来讲，可以将 CRF 看作一种概率无向图模型，定义一个无向图 $G=(V, E)$，节点和边用 v 和 e 表示，在图 G 中，$v\in V$ 表示 G 中的节点，V 表示节点集合，$e\in E$ 表示 G 中的任意一条 E 为边集合；$\boldsymbol{X}$、$\boldsymbol{Y}$ 是两个随机变量，$P(\boldsymbol{Y}|\boldsymbol{X})$ 是定义在 $\boldsymbol{X}$ 的条件下的条件概率分布。如果在图 G 上，每个基于 $\boldsymbol{X}$ 的随机变量 $\boldsymbol{Y}$ 都服从马尔可夫特性，即

$$p(\boldsymbol{Y}_v|\boldsymbol{X}, \boldsymbol{Y}_w, w\neq v)=p(\boldsymbol{Y}_v|\boldsymbol{X}, \boldsymbol{Y}_w, w\sim v) \tag{3}$$

式中对任意节点 v 成立，则称条件概率分布 $P(\boldsymbol{Y}|\boldsymbol{X})$ 为条件随机场，式 (3) 中 $w\sim v$ 表示两个节点 w 和 v 之间存在连接边，表示两个节点 $G=(V, E)$ 在中位置相邻。$\boldsymbol{Y}_v$, $\boldsymbol{Y}_w$ 为节点 v 和 w 所对应的随机变量。

最常用和最简单的 CRF 图结构是线性链结构，可用于序列标注等问题，图 4 为线性链 CRF。由图 4 可知，线性链 CRF 在各个输出序列节点之间做了一阶马尔可夫独立性假设，在给定一个输入序列 X 的标注序列的情况下，令 $\boldsymbol{X}=\{x_1, x_2, \cdots, x_n\}$ 表示被观察的输入序列，$\boldsymbol{Y}=\{y_1, y_2, \cdots, y_n\}$ 表示有限状态的集合。根据线性链 CRF，线性链的 $\boldsymbol{Y}$ 的条件概率分布的形式为

$$p(y|x, \lambda)\propto \exp\left(\sum_{i,j}\lambda_j t_j(y_{i-1}, y_i, x, i)+\sum_{i,k}u_k s_k(y_i, x, i)\right) \tag{4}$$

式中：t_j 为转移特征函数；s_k 是状态特征函数；λ_j 和 u_k 是对应的权重。可以将两个特征函数统一为 f_j。

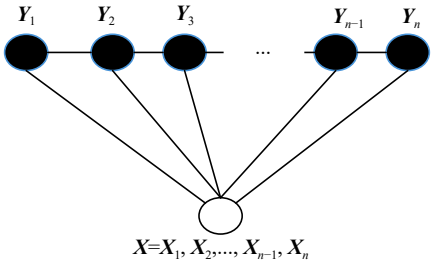

图 4　链式条件随机场

Fig. 4　Chain conditional random field

因此，线性链 CRF 可表示为

$$p(y|x,\lambda)=\frac{1}{Z(x)}\exp(\sum_{j=1}^{j}\sum_{i=1}^{n}\lambda_j f_j(y_{i-1},y_i,x,i)) \tag{5}$$

$$Z(x)=\sum_{y}\exp(\sum_{j=1}^{j}\sum_{i=1}^{n}\lambda_j f_j(y_{i-1},y_i,x,i)) \tag{6}$$

式中：λ_j 是特征函数对应的权值,是待训练的参数。

在 CRF 算法中主要有 3 个关键的问题，分别为特征函数的选择、参数估计和模型推断。CRF 模型中特征函数的形式定义为 $f_j(y_{i-1},y_i,x,i)$，它是状态特征函数和转移特征函数的统一形式表示。

$$f_j(y_{i-1},y_i,x,i)=\begin{cases} b(x,i), & y_{i-1}=<\text{titile}>,y_i=<\text{author}> \\ 0, & \text{其他} \end{cases} \tag{7}$$

特征函数通常是二值函数，取值为 1 或 0。在定义特征函数 $f_j(y_{i-1},y_i,x,i)$ 的时候，首先构建 i 时刻的观察值 x 的真实特征 $b(x,i)=\{0,1\}$ 的集合，结合其对应的标注结果，就可以获得模型的特征函数集。本文根据网络安全实体的特点，选取部分特征，主要包括词、词性、词边界和网络安全实体列表，每一个特征都对应一个特征函数。

参数估计是条件随机场最为关键的问题，主要是从已经标注好的训练数据集学习条件随机场模型的参数，即各特征函数的权重向量 $\boldsymbol{\lambda}$，通常可以通过最大似然估计来实现。目前对于 CRF 模型参数进行估计的方法有 3 种，其中基于 IIS 和 GIS 两种算法是属于迭代的方法。目前广泛使用的条件随机场参数估计算法是 L-BFGS 算法，它是一种近似的二阶方法。与传统的迭代梯度方法相比，此方法的收敛速度更快。下面是 L-BFGS 算法的计算公式：

$$\frac{\partial L(\lambda)}{\partial \lambda_j}=E_{\beta(x,y)}f_j(x,y)-\sum E_{p(y|x^k,\lambda)}f_j(x^{(k)},y)-\frac{\lambda_k}{\sigma^2} \tag{8}$$

模型推断是在给定条件随机场模型参数 λ 下，预测出最可能的状态序列。

4　实验及分析

4.1　实验环境及数据集

本实验是在 Windows 环境下的 Eclipse 下进行开发的，使用 Java 编程语言。由于本实验是基于 Hadoop 的网络安全实体识别，Hadoop 集群环境部署在实验室所提供的 5 台服务器上，Hadoop 平台的拓扑图如图 5 所示，其中服务器使用的是 Linux 操作系统——CentOS 6.8，表 1 为 5 台服务器的硬件配置。

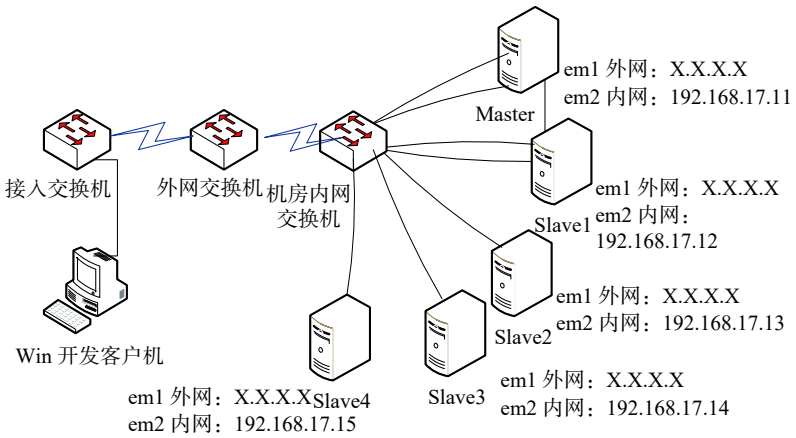

图 5　Hadoop 平台的拓扑结构

Fig. 5　Topological diagram of the Hadoop platform

表 1　服务器的硬件配置

Table 1　Server hardware configuration

服务器	CPU	内存/GB	硬盘/T
Master	24核Intel(R) Xeon(R)	192	4
Slave1	24核Intel(R) Xeon(R)	96	7
Slave2	24核Intel(R) Xeon(R)	96	6
Slave3	24核Intel(R) Xeon(R)	64	10
Slave4	24核Intel(R) Xeon(R)	32	2

本实验采用的数据集主要来自于乌云漏洞数据库，数据主要包括 2010~2016 年公开的漏洞数据，共有 40 292 条漏洞数据。这些数据主要包括漏洞标题、漏洞缺陷编号、漏洞类型、漏洞作者、攻击事件名以及漏洞公开时间。本实验先对乌云漏洞数据集进行去标签，再进行分词，然后进行实体标注，形成了语料库。

为了对算法进行有效的测试，本文对网络安全实体进行人工标注。在实验中用语料库中的 70% 进行训练，30% 进行测试，采用 CRF 算法，以词为单位进行网络安全实体识别。通过 Hadoop 平台，本实验对 30% 的语料库数据进行测试，对漏

洞数据中的 8 种网络安全实体类型进行识别，图 6 为 8 种网络安全实体类型在语料库中的统计信息。

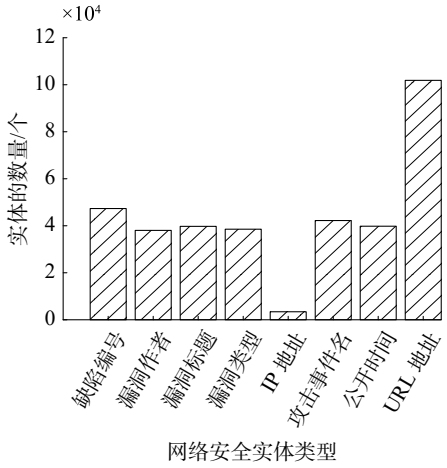

图 6　语料库统计信息

Fig. 6　Network security entity types

4.2　小规模识别率对比实验

本文以准确率 P、召回率 R 和 F 值作为评价指标，具体的定义如下：

$$P=\frac{N_2}{N_1} \tag{9}$$

式中：N_2 表示识别正确的网络安全实体的总个数；N_1 表示识别出来的网络安全实体的总个数。

$$R=\frac{N_2}{N} \tag{10}$$

式中：N_2 表示识别正确的网络安全实体的总个数；N 表示测试语料的网络安全实体的总个数。

$$F=\frac{2\times P\times R}{P+R} \tag{11}$$

本文利用 CRF 算法识别网络安全实体，将识别出来的网络安全实体作为候选网络安全实体，然后利用基于规则的方法，对候选网络安全实体进行修正，将修正过的结果和未修正的结果进行对比。本文利用基于规则的方法对基于 CRF 的网络安全实体的识别进行修正，实验过程中首先建立简单的规则，然后将规则加入到网络安全实体的识别中进行比较。本文制定了以下几条规则：

规则一：如果词的前缀是“腾讯”“优酷”“微软”等厂商名，且该词带有“漏洞”结束符，那么该词应标记为漏洞名称，例如“腾讯某分站地址跳转漏洞”。

规则二：如果词的前缀是“WooYun”，将此类词标记为漏洞缺陷编号。

规则三：如果词的前缀出现“SQL”“XSS”等词，且该词带有“注入”“攻击”“传播”“泄露”等结束符，那么该词应标记为漏洞类型，例如：“XSS 跨站脚本攻击”。

经过以上规则对结果进行纠正，网络安全实体的识别效率都有所提高。图 7 是对修正和未修正结果的准确率的对比，图 8 是召回率的对比，图 9 是 F 值的对比。

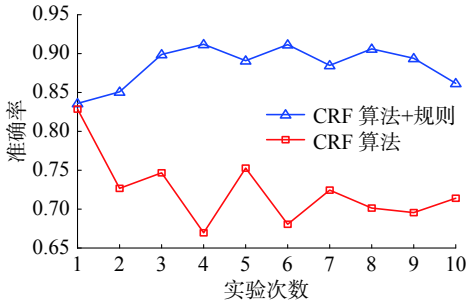

图 7　准确率对比结果

Fig. 7　Comparisons of precision of results

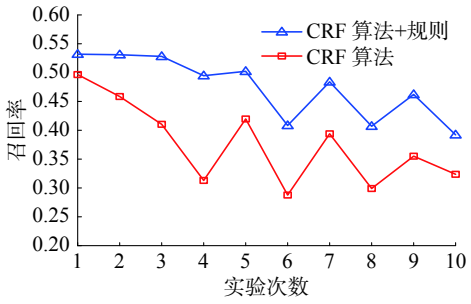

图 8　召回率对比结果

Fig. 8　Comparisons of recall results

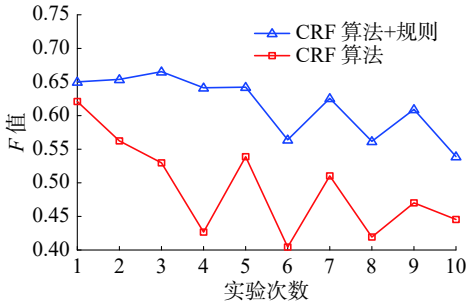

图 9　F 值对比结果

Fig. 9　Comparisons of F-value results

图 7~9 列出了网络安全实体 10 次实验的识别结果，从实验结果可以看出，在使用规则对于基于 CRF 算法的网络安全实体识别的结果进行修正，识别效果有了一定的提高。就准确率而言，基于 CRF 算法与规则相结合的准确率能达到 85% 以上，10 次实验中准确率最高达到了 91%。但是就召回率而言，从实验结果来看，识别效果比较低，主要是因为 CFF 模型泛化能力不够和训练的语料库非常小。

4.3　大规模对比实验

本实验采用 Hadoop 框架，主要利用 MapReduce 对大规模数据进行分割，对网络安全实体的识别并行化处理。本文将 Hadoop 安装在 5 个节点的集群中，文本数据块的大小为 128 MB。为了

更好地说明分布式计算效率，本实验在不同的数据规模下，基于不同的节点个数测试网络安全实体识别时间。实验中分为 3 个节点、4 个节点以及 5 个节点，同时加上伪分布式集群。在 Hadoop 集群上，运用 4 组数据进行实验，4 组数据大小分别为 1.3 GB、6 GB、13 GB、28 GB。实验结果如图 10 所示。

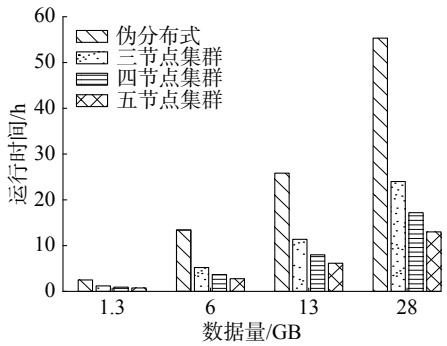

图 10　不同节点数下的运行时间对比结果

Fig. 10　Comparison of running times for different node numbers

从图 10 可以看出，随着计算节点个数的增加，网络安全实体的识别时间也随之加快。在数据量为 1.3 GB 的时候，随着节点数的增加，网络安全实体识别时间变化不大，识别效率提高不明显。随着数据量的增大，在伪分布式的情况下，28 GB 数据耗时近 55 h，5 个节点耗时近 13 h，识别效率明显提高。

4.4　算法的可扩展性分析

本文提出的基于 Hadoop 的 CRF 算法的网络安全实体识别算法具有很好的扩展性。图 11 展示了 28 GB 数据的运行时间，从图中可以看出随着计算节点数的增加数据运行时间逐渐下降。实验证明，增加节点数可以有效增加网络安全实体识别效率，因此本文基于 Hadoop 的网络安全实体识别算法具有良好的可扩展性，适用于大规模数据的集群计算。

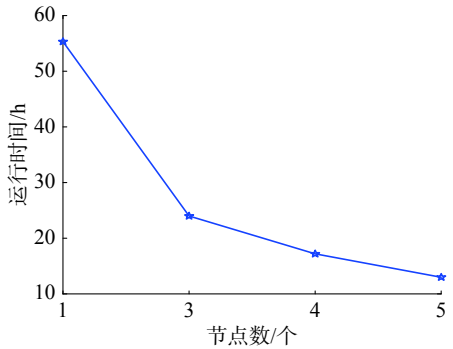

图 11　28 GB 数据运行时间对比

Fig. 11　Comparison of 28 GB data running times

4.5　网络安全实体识别实例分析

为了进一步直观展示本文算法在网络安全实体识别方面的结果，安全实体词云图如图 12 所示。“DDOS 攻击”“SQL 注射漏洞”等网络安全实体，具有典型的中英文混合结构，传统的命名识别方法较少关注中英文混合结构的命名实体识别。通过词云图可以直观地看出，本文提出的基于规则的 CRF 算法能够有效处理中英文混合的网络安全实体，进一步提升了安全实体识别的准确率，为基于网络安全知识图谱的威胁情报分析奠定了基础。

图 12　网络安全实体词云图

Fig. 12　Word cloud map of network security entity

5　结束语

本文对网络安全实体识别的常用算法进行了总结，详细分析了基于 CRF 算法的网络安全实体识别方法，并针对大规模数据在 Hadoop 框架下对网络安全实体识别进行并行化处理。实验表明，本文采用基于 Hadoop 的 CRF 算法的网络安全实体识别，取得了良好的效果，并大大地缩短了识别时间。在后续的工作中，会考虑融合更多网络安全领域的知识使得安全实体识别具有更好的泛化能力，从而提高实体的识别率，并扩展至多机分布式平台，进一步提高性能。

参考文献：

[1] 廖建新. 大数据技术的应用现状与展望 [J]. 电信科学, 2015, 31(7): 1–12.
LIAO Jianxin. Big data technology: current applications and prospects[J]. Telecommunications science, 2015, 31(7): 1–12.

[2] 单琳. 网络威胁情报发展现状综述 [J]. 保密科学技术, 2016(8): 28–33.
SHAN Lin. Overview of the development status of cyber threat intelligence[J]. Security science and technology,

2016(8): 28–33.

[3] 南湘浩, 陈钟. 网络安全技术概论 [J]. 计算机安全, 2003(30): 76.
NAN Xianghao, CHEN Zhong. Introduction to network security technology[J]. Computer security, 2003(30): 76.

[4] 陈兴蜀, 曾雪梅, 王文贤, 等. 基于大数据的网络安全与情报分析 [J]. 工程科学与技术, 2017, 49(3): 1–12.
CHEN Xingshu, ZENG Xuemei, WANG Wenxian, et al. Big data analytics for network security and intelligence[J]. Advanced engineering sciences, 2017, 49(3): 1–12.

[5] 张晓艳, 王挺, 陈火旺. 命名实体识别研究 [J]. 计算机科学, 2005, 32(4): 44–48.
ZHANG Xiaoyan, WANG Ting, CHRN Huowang. Research on named entity recognition[J]. Computer science, 2005, 32(4): 44–48.

[6] JONES C L, BRIDGES R A, HUFFE K M T, et al. Towards a relation extraction framework for cyber-security concepts[C]//Proceedings of the 10th Annual Cyber and Information Security Research Conference. Oak Ridge, USA, 2015: 11.

[7] JOSHI A, LAL R, FININ T, et al. Extracting cybersecurity related linked data from text[C]//Proceedings of 2013 IEEE Seventh International Conference on Semantic Computing. Irvine, USA, 2013: 252–259.

[8] LAL R. Information extraction of security related entities and concepts from unstructured text[D]. Baltimore County: University of Maryland, 2013.

[9] MULWAD V, LI Wenjia, JOSHI A, et al. Extracting information about security vulnerabilities from web text[C]//Proceedings of 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology. Lyon, France, 2011: 257–260.

[10] 翟菊叶, 陈春燕, 张钰, 等. 基于 CRF 与规则相结合的中文电子病历命名实体识别研究 [J]. 包头医学院学报, 2017, 33(11): 124–125, 130.
ZHAI Juye, CHEN Chunyan, ZHANG Yu, et al. A study on the named entity recognition of Chinese electronic medical record based on combination of CRF and rules[J]. Journal of Baotou Medical College, 2017, 33(11): 124–125, 130.

[11] 张晓艳, 王挺, 陈火旺. 基于混合统计模型的汉语命名实体识别方法 [J]. 计算机工程与科学, 2006, 28(6): 135–139.
ZHANG Xiaoyan, WANG Ting, CHEN Huowang. A mixed statistical model-based method for chinese named entity recognition[J]. Computer engineering and science, 2006, 28(6): 135–139.

[12] 徐梓豪. 基于统计模型的中文命名实体识别方法研究及应用 [D]. 北京: 北京化工大学, 2017.
XU Zihao. Statistical model based Chinese named entity recognition methods and its application to medical records[D]. Beijing: Beijing University of Chemical Technology, 2017.

[13] RABINER L R. A tutorial on hidden Markov models and selected applications in speech recognition[M]//WAIBEL A, LEE K F. Readings in Speech Recognition. San Francisco: Morgan Kaufmann, 1990: 267–296.

[14] KOELING R. Chunking with maximum entropy models[C]//Proceedings of the 2nd Workshop on Learning Language in Logic and the 4th Conference on Computational Natural Language Learning. Lisbon, Portugal, 2000: 139–141.

[15] 郑秋生, 刘守喜. 基于 CRF 的互联网文本命名实体识别研究 [J]. 中原工学院学报, 2016, 27(1): 70–73, 95.
ZHRNG Qiusheng, LIU Shouxi. Research of web text named entity recognition based on CRF[J]. Journal of Zhongyuan University of Technology, 2016, 27(1): 70–73, 95.

[16] 朱颢东, 杨立志, 丁温雪, 等. 基于主题标签和 CRF 的中文微博命名实体识别 [J]. 华中师范大学学报(自然科学版), 2018, 52(3): 316–321.
ZHU Haodong, YANG Lizhi, DING Wenxue, et al. Named entity recognition of Chinese microblog based on theme tag and CRF[J]. Journal of Central China Normal University (Natural Sciences), 2018, 52(3): 316–321.

[17] TELNOV Y, SAVICHEV I. Ontology-based competency management: infrastructures for the knowledge intensive learning organization[C]//Proceedings of the 1st International Early Research Career Enhancement School. Cham, Switzerland, 2016: 249–256.

[18] IANNACONE M, BOHN S, NAKAMURA G, et al. Developing an ontology for cyber security knowledge graphs[C]//Proceedings of the 10th Annual Cyber and Information Security Research Conference. Oak Ridge, USA, 2015: 12.

[19] DEAN J, GHEMAWAT S. MapReduce: simplified data processing on large clusters[C]//Proceedings of the 6th Conference on Symposium on Opearting Systems Design and Implementation. San Francisco, USA, 2004: 10.

[20] LAFFERTY J D, MCCALLUM A, PEREIRA F C N. Conditional random fields: probabilistic models for segmenting and labeling sequence data[C]//Proceedings of the 18th International Conference on Machine Learning. Williamstown, USA, 2001: 282–289.

作者简介:

秦娅, 女, 1992 年生, 硕士研究生, 主要研究方向为网络安全知识图谱。

申国伟，男，1986 年出生，副教授，主要研究方向为大数据、网络与信息安全、数据挖掘。

余红星，男，1993 年生，硕士研究生，主要研究方向为大数据技术。

第三届人工智能与大数据国际会议（ICAIBD 2020）
2020 3rd International Conference on Artificial Intelligence and Big Data (ICAIBD 2020)

2020 3rd International Conference on Artificial Intelligence and Big Data (ICAIBD 2020) will take place on May 28-31, 2020 inChengdu, China. It is sponsored by Sichuan Province Computer Federation and technically assisted by many local and international universities. This conference provides you opportunity to meet with academicians as well as practitioners in the fields of Artificial Intelligence and Big Data from all over the world, and get the latest insights from every area of Artificial Intelligence and Big Data theory and practice.

ICAIBD features invited keynote speech, peer-reviewed paper presentations, and academic visit. The conference is completely open (one needs to register first…), you will not have to be an author or a discussant to attend. Submissions will be peer-reviewed and evaluated based on originality, relevance to conference, contributions, and presentation. We invite the submission of original research contributions. Accepted papers will be collected in the Conference Proceedings, which is published by IEEE and reviewed by the IEEE Xplore, and then sent for indexing by Ei Compendex, and Scopus.

会议官址：www.icaibd.org