2022w2周进度总结 1.10~1.16

基于漏洞知识图谱的可视化系统的设计与实现

马嘉骥 2018211149

w2进展

一

阅读了以下文章，学习知识图谱基础概念。

1. 通过阅读综述了解规则匹配、传统机器学习、深度学习等构建知识图谱的方法。
2. 学习网络安全领域的本体构建，了解了利用Stanford NER使用线性CRF进行命名实体识别的过程即
   $P(y|x) = \frac{1}{Z(x)}\Pi_{t=1}^{T}\exp\{\sum_{k=1}^{K}\lambda_i f_i(y_{j-1}, y_j, x_j)\}$
   $Z(x) = \sum_y \Pi_{t=1}^{T}\exp\{\sum_{k=1}^{K}\lambda_i f_i(y_{j-1}, y_j, x_j)\}$
3. 图谱表示：若K表示知识图谱$K =< concept, instance, relation, properties, rule >$，其中
   $Concept = \{concept_i | i = 1, 2, \dots n\}$概念是各种抽象本体的集合，操作系统、软件、攻击都属于Concept。
   $Instance = \{instance_i | i = 1，\ \dots，\ m\}$，实例是具体例子的集合，如Windows 7、Adobe Acrobat PDF Reader、DDoS等。
   $Properties = \{< instance_i, property_{ij}, value_j >\}$，属性是实例属性值的集合。
   $Relation =< concept_i, R_{cc}, concept_j > | < concept_i, R_{ci}, instance_j > | < instance_i, R_{ii}, instance_j >$，表示实例之间的关系，如subClassOf、instanceOf、is a等。
   $Rule = \{rule | rule =< instance_i, newRule_{ij}, instance_j > | < concept_i, newRule_{ij}, instance_j > | < instance_i, property_{ij}, newValue_j >$，规则用来推演新的属性值和新的关系。
4. 知识推演：对于三个实例N_i, N_j, N_l属性推演的预测公式是

   $Value_{ik} = \sum_{j=1}^{m}\lambda_j \cdot f_{ij}(key_j, value_j) + \sum_{t=1}^{l}\sigma_t \cdot \sum_{j=1}^{m}\lambda_j \cdot f_{ij}(key_j + value_j)$。

   使用路径排序算法进行关系推演，就是使用连接两个实体的路径作为特征预测两个实体之间关系。对于三个实例N_i, N_j, N_l进行关系推演的预测公式是

   $Score(l, j) = \sum_{\pi \in Q} Path[e_l, e_j; length(\pi) \le n] \cdot \omega_\pi$，

   π是*l*到*j*所有的长度权值小于n的可达路径。在进行关系推演时对这些路径评分，如果$Score(l, j)$大于一个特定阈值，则认为这个关系（边）成立。如此可以得到实例之间新的关系。
5. 验证与评估：在信息检索和提取系统中主要使用精确率和召回率评估指标，可以使用$F - Measure$计算精确率和召回率的调和平均值来更全面地评估系统性能。对系统的预测结果有三种定义，分别是"真正TP"将正类预测为正类、"假正FP"将负类预测为正类、"假负FN"将正类预测为负类。

   精确率的公式定义是$Precision = \frac{TP}{TP+FP}$，

   召回率的公式定义是$Precision = \frac{TP}{TP+FN}$，

   $F - Measure$的$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$。

   根据以上公式，可以对系统的性能进行评估。评估方法可以采用交叉验证的方法。如采用十倍交叉验证评估模型质量，就是把数据均分为10块，9/10的数据训练模型，1/10的数据作为测试数据验证模型。

> Guo, Shu, Quan Wang, Lihong Wang, Bin Wang, and Li Guo. "Knowledge Graph Embedding with Iterative Guidance from Soft Rules." *ArXiv:1711.11231 [Cs]*, November 30, 2017. http://arxiv.org/abs/1711.11231.
>
> Jia, Yan, Yulu Qi, Huaijun Shang, Rong Jiang, and Aiping Li. "A Practical Approach to Constructing a Knowledge Graph for Cybersecurity." *Engineering*, Cybersecurity, 4, no. 1 (February 1, 2018): 53–60. https://doi.org/10.1016/j.eng.2018.01.004.
>
> Zhou Yuanchun, Qiao Ziyue, Du Yi, Wang Weijun, and Xiao Meng. "A survey on the construction methods and applications of sci-tech big data knowledge graph." *SCIENTIA SINICA Informationis* 50, no. 7 (July 1, 2020): 957–87. https://doi.org/10.1360/SSI-2019-0271.
>
> 张吉祥，张祥森，武长旭，赵增顺．知识图谱构建技术综述[J/OL]．计算机工程. https://doi.org/10.19678/j.issn.1000-3428.0061803
>
> 陶耀东，贾新桐，吴云坤． 一种基于知识图谱的工业互联网安全漏洞研究方法 [J] ．信息技术与网络安全， 2020，39(1) : 6-13，18．

二

了解CVE、CPE、CWE、CAPEC等网络安全漏洞领域术语。从CVE网站获取了csv格式漏洞数据。

三

学习使用Zotero管理文献。

w3预期工作内容

1. 学习具体模型的使用与训练方法，了解使用何种格式将爬虫原始数据清洗以获得可以进行训练的数据。
2. 对前端与后端技术栈进行调研与技术选型，并相应学习框架的使用。