面向社会目标的复杂网络态势感知与取证分析

孙国梓　　南京邮电大学

ISC 互联网安全大会 中国 · 北京
Internet Security Conference 2018 Beijing · China

关于

关于讲者

- 博士，南京邮电大学教授、计算机技术研究所副所长
- CCF YOCSEF南京主席 (2016-2017)
- CCF 南京分部副主席 (2018)
- 中国计算机学会高级会员
- 中国电子学会高级会员
- 中国网络空间安全协会理事
- CCF区块链专业委员会委员，东南大学区块链技术实验室
- 中国电子学会计算机取证专家委员会委员
- 电子数据鉴定司法鉴定人
- IEEE CS SSA会员，香港ISFS会员

关于南邮

- 1942年诞生于山东抗日根据地的八路军战邮干训班
- 南邮精神：信达天下自强不息
- 南邮校训：厚德、弘毅、求是、笃行
- 南邮校风：勤奋、求实、进取、创新
- 国家“双一流”建设高校和江苏高水平大学建设高校
- 2002年信息安全专业开始招生
- 信息安全国家特色专业建设点（全国15所）
- 中国网络空间安全协会常务理事单位
- 江苏省无线传感网高技术研究重点实验室
- 江苏省大数据安全与智能处理重点实验室

关于团队

王志伟
博士 / 副教授

李华康
博士 / 讲师

曹晓梅
博士 / 副教授

Privacy Protection

李群
博士 / 讲师

BIG DATA

CLOUD

汪莹
讲师

Digital Forensics

沙乐天
博士 / 讲师

付雄
博士 / 教授

孙国梓
博士 / 教授

ZERO TRUST SECURITY

ISC 互联网安全大会 中国 · 北京
Internet Security Conference 2018 Beijing · China

目录

背景概述

关键技术

态势感知分析

电子数据取证视角

背景概述

背景概述

- ✓ 智能终端、移动应用、社交网络、物联网等为社会目标分析提供了丰富的素材
  ——掀起了新的社会目标分析热潮

- ✓ 社会中的每个目标不是单独存在的
  ——目标与目标之间、事件与事件之间、目标与事件之间存在复杂动态的网络关系

- ✓ 多维数据分析
  ——需要一个分析方法，能够对数据进行维度化分析后的度量聚集统计

- ✓ 维度
  ——即观察事物的角度，综合考虑多个维度的因素，能够更清晰的认识事物的本质

关键技术

关键技术

系统简要概述

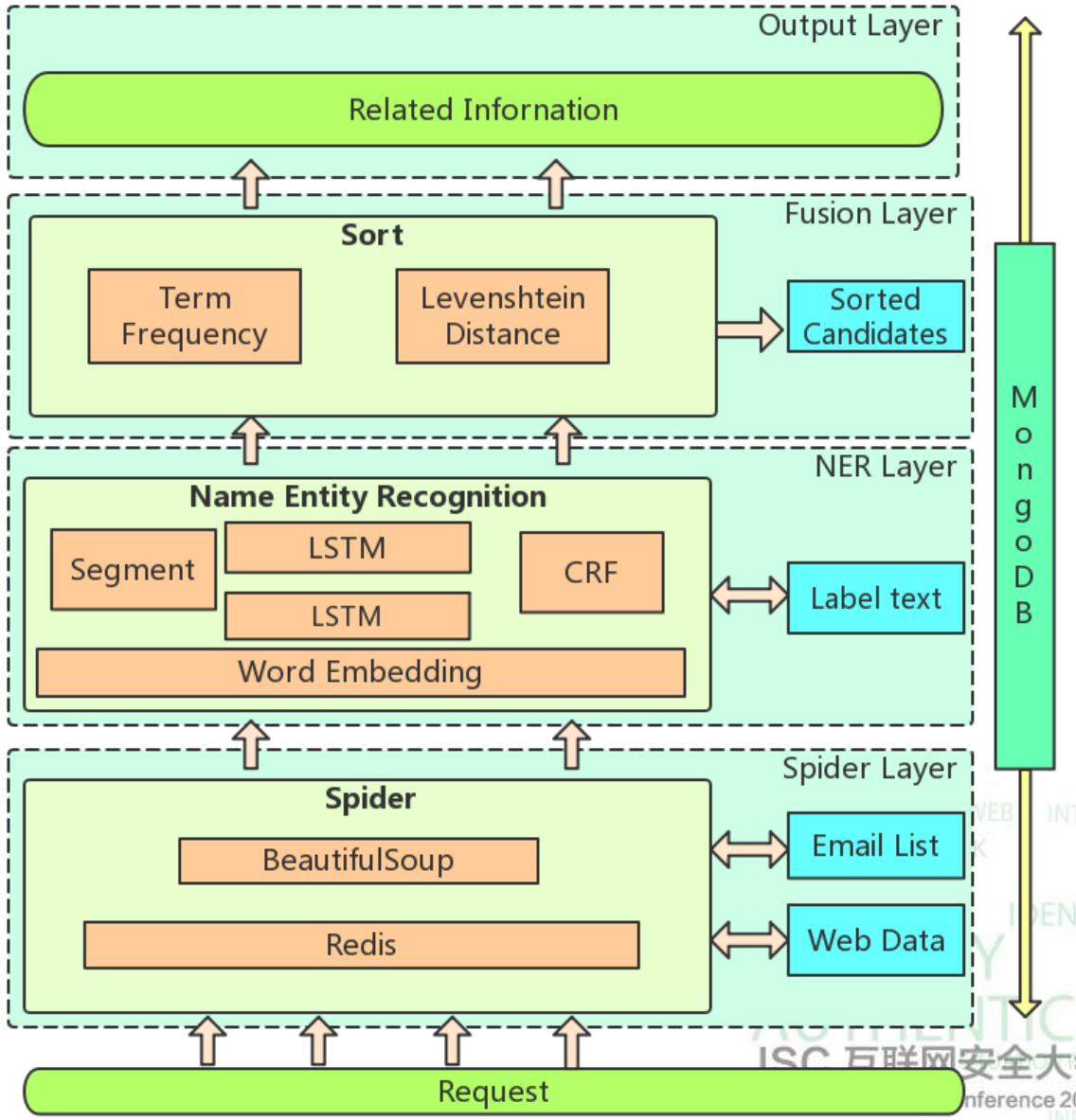
系统框架

系统重点关注三部分：

“分布式数据采集”

“命名实体识别”

“融合排序”

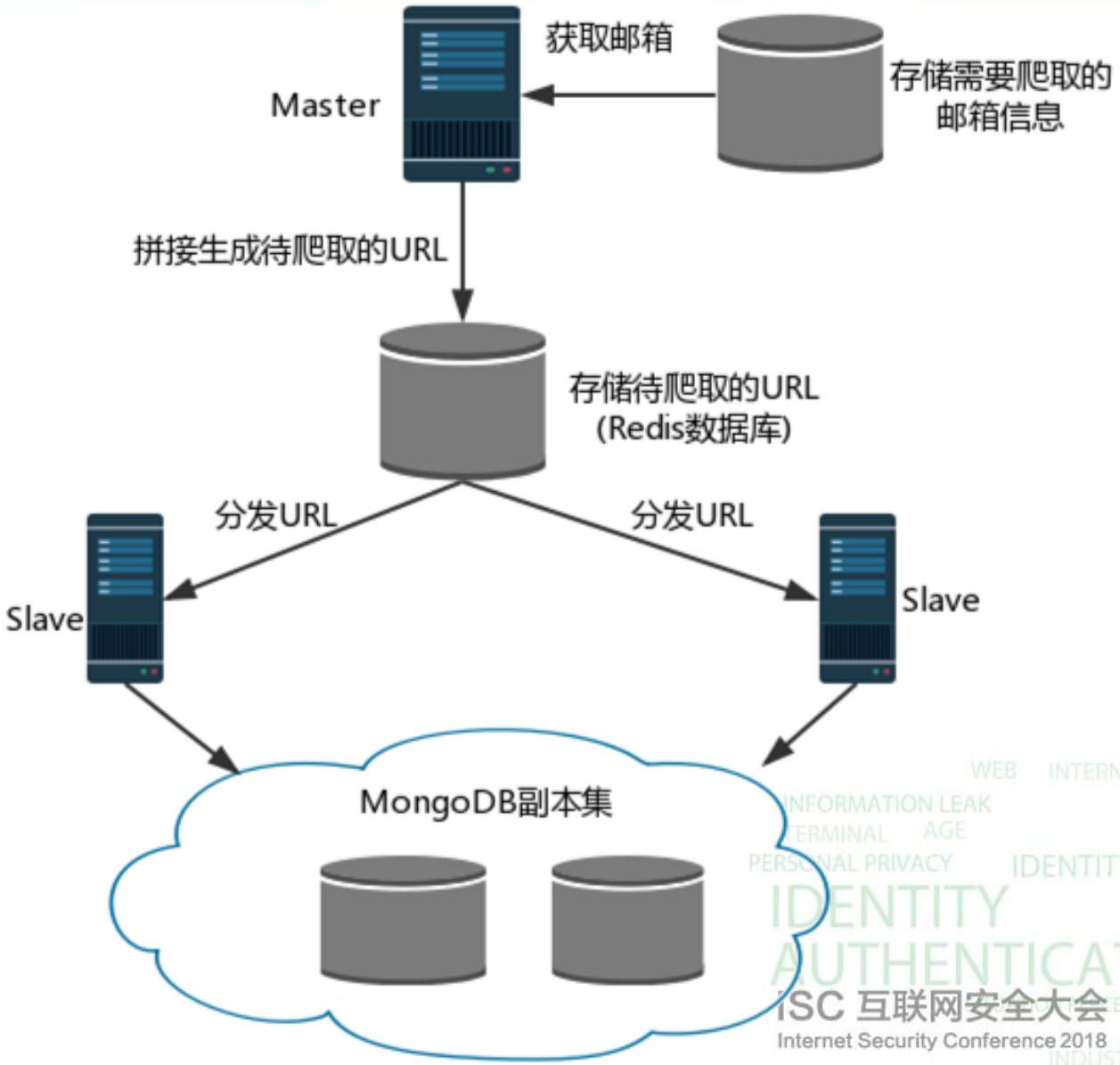


关键技术

关键点描述一

分布式数据采集

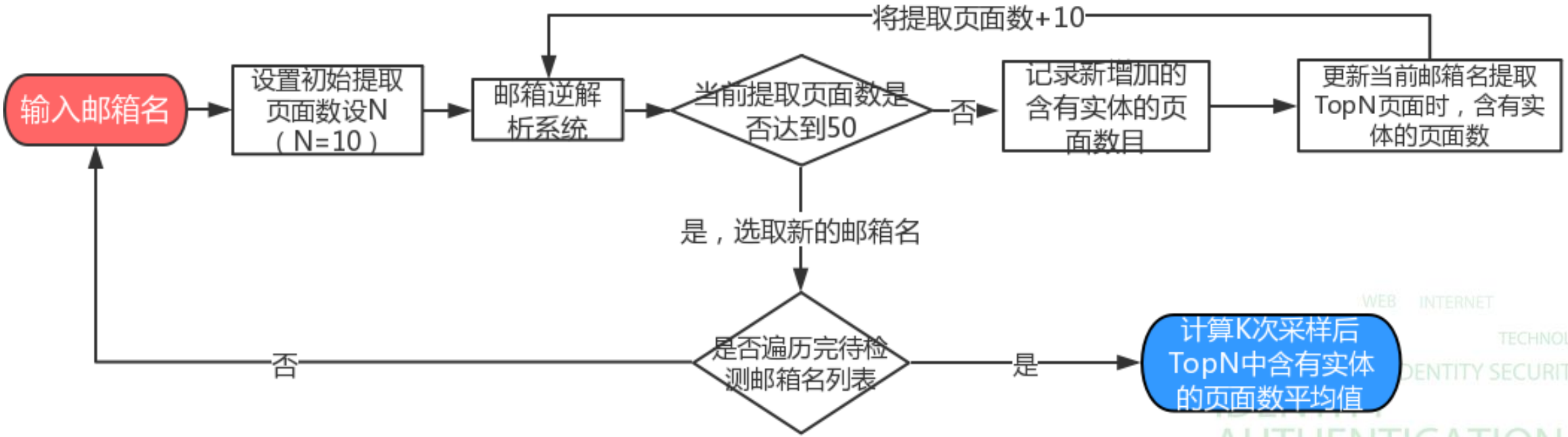
1. 分布式大批量数据采集
2. 分布式存储



关键点描述一

分布式数据采集

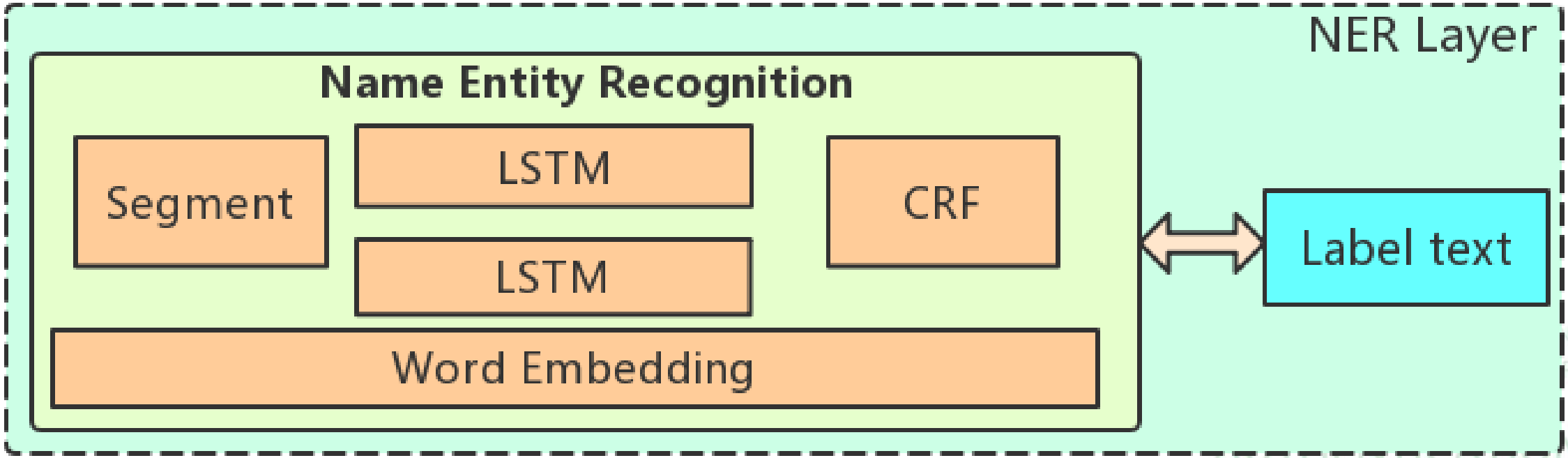
3. 确认爬取页面中的条数：TopN



关键点描述二

命名实体识别

1. 命名实体识别算法：LSTM+CRF



关键点描述二

命名实体识别

2. 命名实体识别方面

1）**词向量进行预处理**：维基百科中英文语料

2）**训练集**：现有语料基础上个性定制训练集

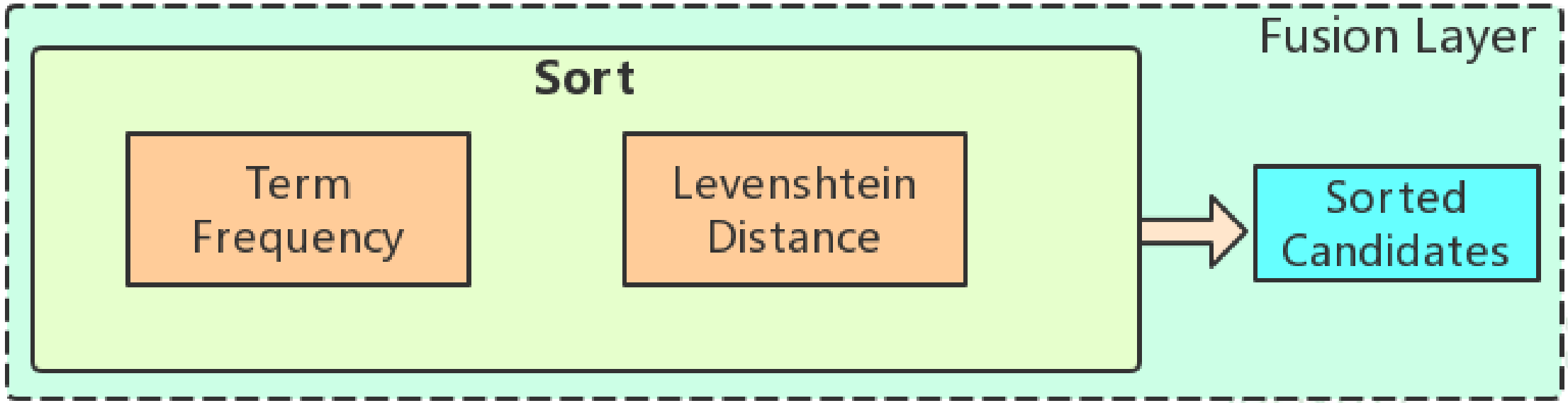
3）**自定义实体识别种类**：可按需定制

4）**模型的改进**：中英文双向识别

关键点描述三

融合排序

1. 多算法融合排序：TF-Levenshtein Distance

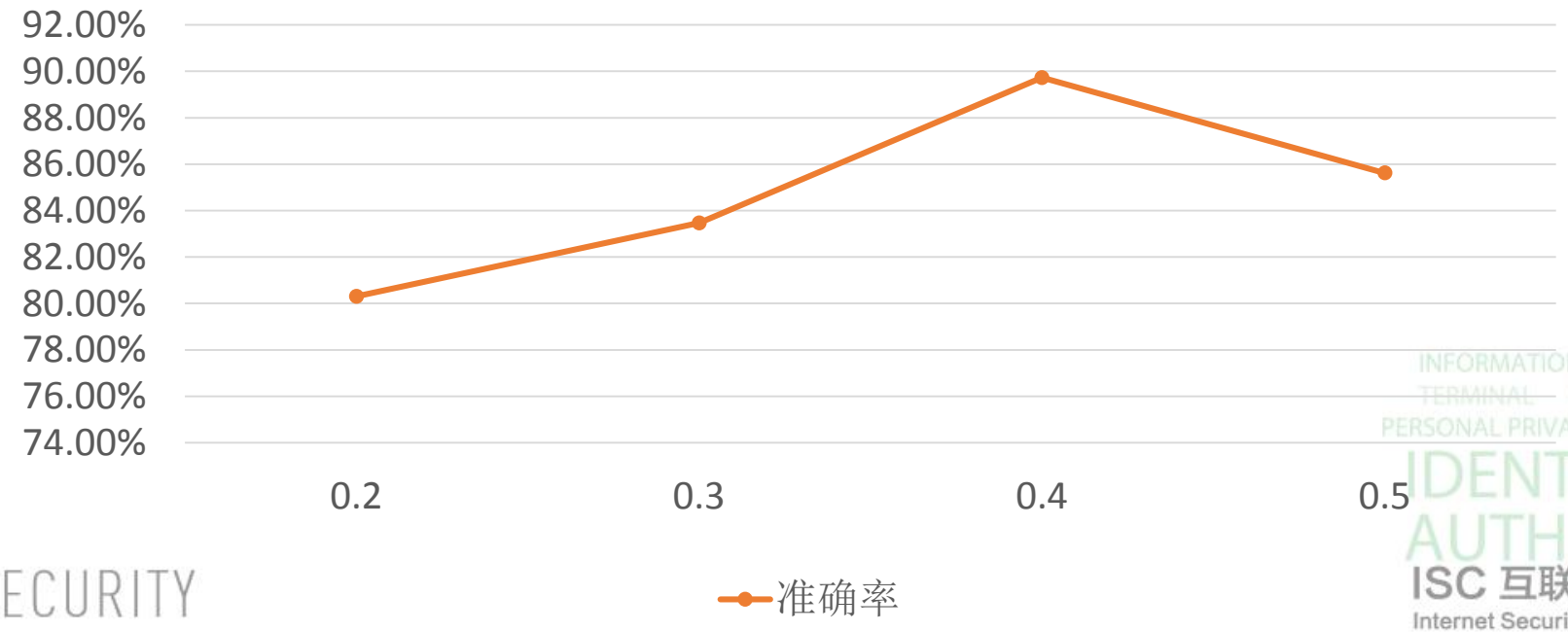


关键点描述三

融合排序

2. 多算法融合排序：TF-Levenshtein Distance

$$weights = \alpha * TF + (1 - \alpha) * Levenshtein$$



测试结果

测试一 不同搜索引擎抓取TopN页面中含有实体的平均页面数

抓取页面数	百度	Bing
top10	7.65	5.6
top20	14.4	9.6
top30	20.2	11.2
top40	25.3	12.3
top50	30.3	13.4

关键技术

测试结果

测试二 LSTM+CRF、CRF、HMM模型对于不同搜索引擎得到的结果的实体识别率（F1值）

1. 对于地点名 Location 的识别率

F1	LSTM+CRF	CRF	HMM
baidu	**89.28%**	77.18%	71.44%
bing	**90.18%**	78.98%	71.67%

2. 对于机构名 Organization的识别率

F1	LSTM+CRF	CRF	HMM
baidu	**84.87%**	50.27%	37.51%
bing	**86.42%**	62.18%	42.95%

3. 对于人名 Person的识别率

F1	LSTM+CRF	CRF	HMM
baidu	**92.12%**	74.45%	64.26%
bing	**90.35%**	65.33%	66.00%

测试结果

测试三 测试多个平台检测试题结果融合算法

1. 排序结果前2的人名实体中，包含邮箱所有者真实姓名的邮箱数所占比例

百度	Bing	融合
67%	72%	94%

2. 排序结果前5的人名实体中，包含与邮箱所有者相关的实体平均个数

百度	Bing	融合
3.4	3.8	4.2

3. 非重复的地名、机构名且与邮箱所有者相关的实体平均个数

百度	Bing	融合
2.8	2.6	3.1

结论

- 对于中文实体来说，搜索条数在30条以上时，新实体的个数增加的很少
- 实体识别，采用LSTM+CRF模型比采用HMM或只用CRF结果更好
- 采用多平台融合的处理结果单平台准确率更高

关键技术

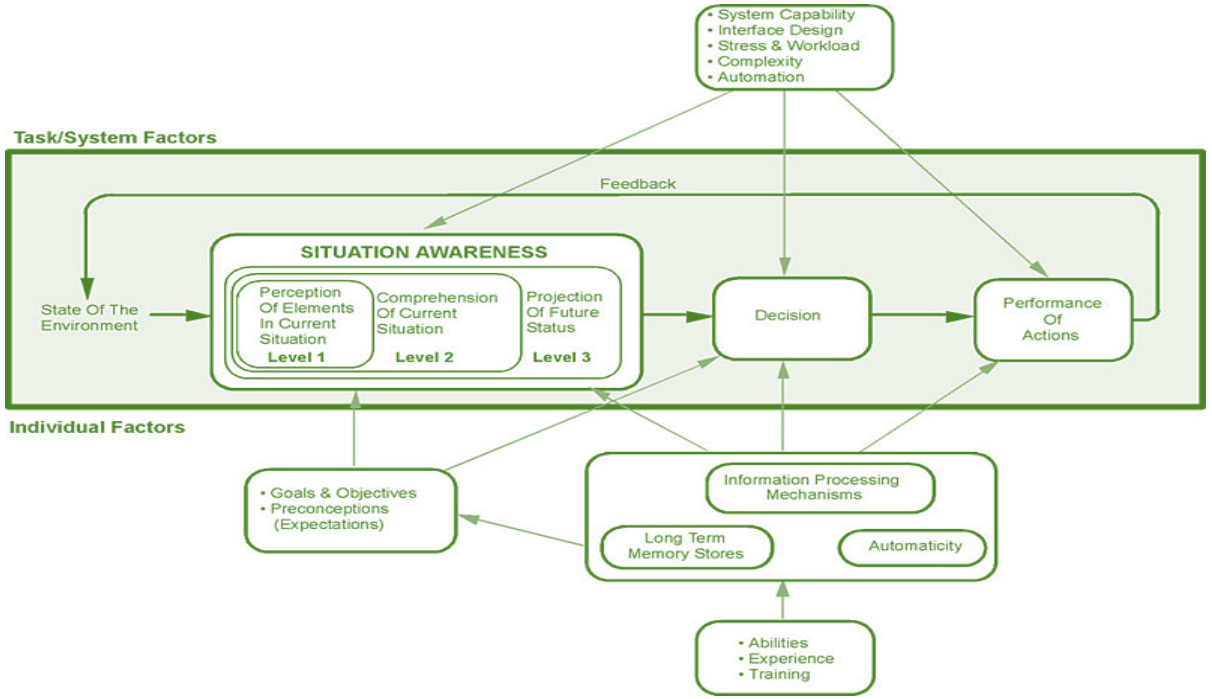
建议

- 针对搜索结果的特性，可以考虑在实体识别中加入一些人工规则，过滤掉可能的无用信息，如：“百度文库”、“豆丁网”等
- 搜索条数可以控制在30条以内，应采用多搜索平台信息融合的方式进行信息获取
- 地名和机构名实体在实际应用中可看作同一类别

态势感知分析

网络态势感知

- 态势感知： “在一定的时空条件下，对环境因素的获取、理解以及对未来状态的预测。”(Endsley，1988) "
- 网络态势感知： “在大规模网络环境中，对能够引起网络态势发生变化的安全要素进行获取、理解、显示以及预测最近的发展趋势。”(Tim Bass, 1999)



➢ 态势觉察

➢ 态势理解

➢ 态势预测

Endsley, "Design and evaluation for situation awareness enhancement" 1988.

网络态势感知——能力构建

- 感知能力（觉察）
- 发现能力（理解）
- 预测能力（预期）

风险和威胁说

风险

某一特定**环境**下，在某一特定的**时间段**内，某种**损失**发生的**可能性**

通常是**泛指**， 强调**使命保障**，强调其中的未来结果的不确定性或损失，**关注可能性**

威胁

用武力、权势胁迫；使……面临危险

通常是**特指**，围绕**任务保障**，关注危及当前任务的攻击来源和手段，**关注现实性**

行为

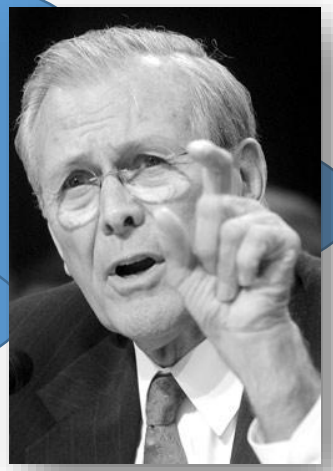
“基于个人意志而具体表现於外的举止动作。”

在网络空间，人类的行为是通过网络乃至软件（或代理）的行为来实现的，虚拟世界的网络行为是人类行为的延续和拓展

我们把**软件运行时作为主体，依靠其自身的功能对实体的施用、操作或动作称为行为**

风险的分类

- 『据我们所知，有「已知的已知」，有些事，我们知道我们知道；我们也知道，有「已知的未知」，也就是说，有些事，我们现在知道我们不知道。但是，同样存在「未知的未知」——有些事，我们不知道我们不知道。』
  - 美国前国防部长拉姆斯菲尔德

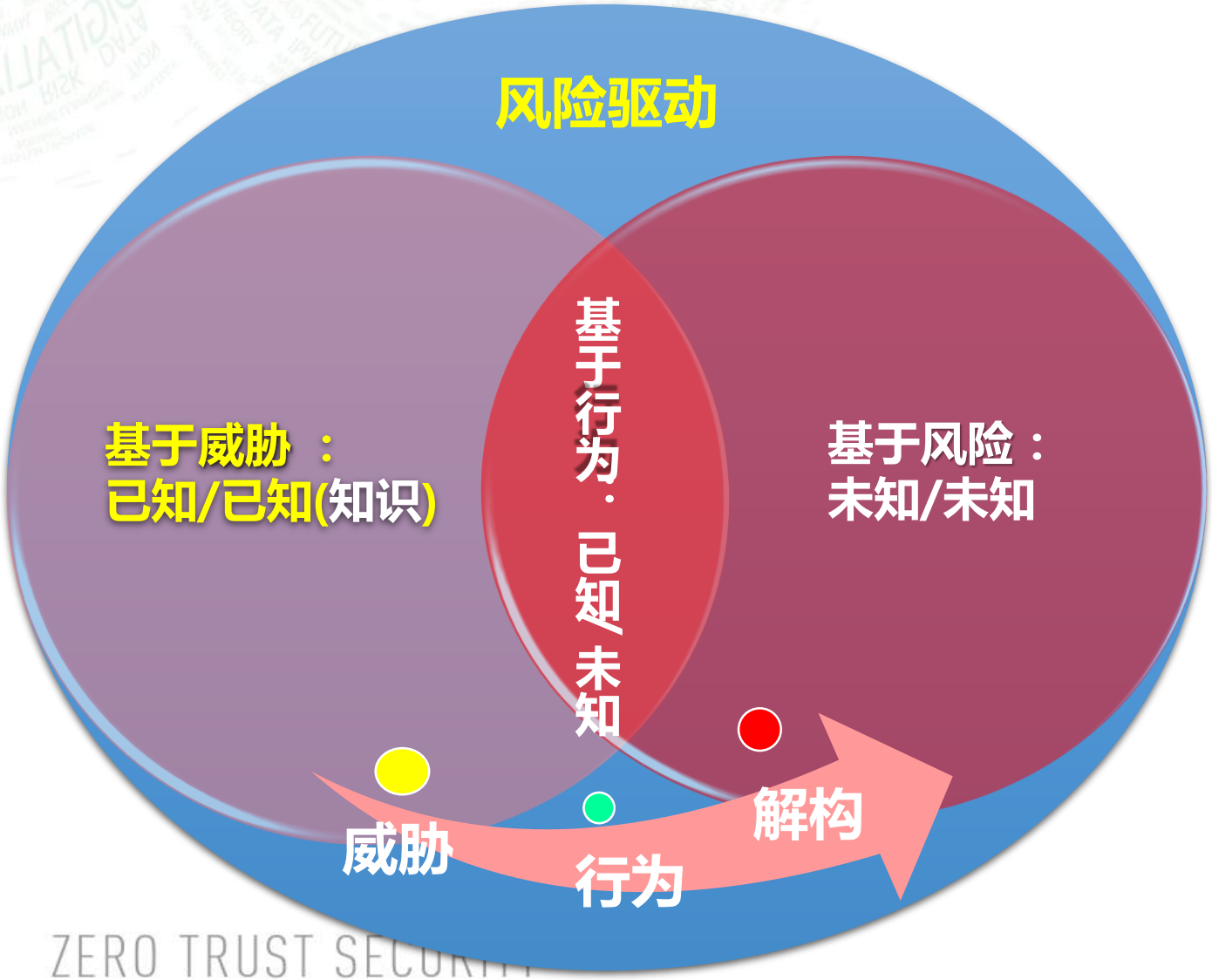


从管理学的角度，风险被分为三种类型：

- **『已知/已知』**——知道可能发生什么风险，且对风险发生的可能性和影响有准确了解。（如：人的死亡）
- **『已知/未知』**——知道可能发生什么风险，但对风险发生的可能性和影响并不了解。（如：机器故障）
- **『未知/未知』**——不知道可能发生什么风险，也不知道风险发生的可能性和影响的严重性。（如：美国总统小布什在演讲的时候被扔鞋。）

当前网络安全的最大危机：不知道我们的网络中存在着什么东西，发生了什么事情！（盲，瞎）

威胁和风险



- 从网络安全的角度，网络安全态势感知应该考虑：

  - I 已知/已知： 基于威胁的实时恶意行为预处理能力(实时分析)
  - II 已知/未知：基于行为的准实时威胁发现能力（准实时分析）
  - III 未知/未知：基于风险的多元多维威胁解构能力（实时/准实时/中长期分析）

态势感知需求概述

数据采集	数据要丰富（网络结构、服务、恶意代码、漏洞、入侵等）
安全要素	安全分析全面（保密性、完整性、可用性）
感知流程	流程规范，算法简单，易操作模型，并能实时准确评估
评估预测	多层次、角度评估威胁、脆弱、安全等状况，支持多种结果预测
结果显示	支持多形式可视化，与用户交互，生成评测报表和提供加固方案

电子数据取证视角

电子数据取证视角

人物属性分析

目标人物关系分析

人物行为序列分析

事件序列分析

证据链的形成


ISC
2018
ISC互联网安全大会
360互联网安全中心
谢 谢！
ISC 互联网安全大会 中国 · 北京
Internet Security Conference 2018 Beijing · China