# 对AS属性进行推断的相关研究总结-未完成

## IMC-21-“Identifying ASes of State-Owned Internet Operators”论文解读

这篇论文的主要目的是**对全球范围内的国有网络operator以及他们的ASN/研究网络Operator的ASN以及其国有属性**。

这个研究的作用主要在于为网络审查、国有网络识别、国际关系以及信息战等研究提供基础

同时，在这篇文章的研究之前，并没有一个公开的数据库或者数据集提供了这方面的信息；少部分的商业数据库提供了全球公司的归属权的数据，但是这部分数据库（1）并没有提供与AS相关的数据；（2）这些数据库的数据不够全面和准确，因此这项研究的工作具有开创性的意义

这项研究所面临的困难：

* **缺乏关于国有公司（公司-to-是否国有）公开数据库**，不存在一个全球级别的国有企业登记数据库（少部分国家如北欧三国会公开，但是数量远远不够）；提供电信公司所属权的两个数据库Orbis和GlobalComms两个数据本身又存在非常大的错误
* **缺乏公司-AS之间的映射数据**，虽然WHOIS数据可能能做到这点，但是WHOIS数据更新不及时，并且WHOIS数据的公司名称和法定名称可能不一样
* **公司的所属权本身也是个很复杂的问题**，部分公司的所属权可能存在共同拥有或者间接拥有的情况，导致判断所属权的问题变得十分复杂，并且对于电信公司，这个问题会更加复杂：
  + 国家可以通过各种渠道传达某个公司被私有化/国有化
  + 电信公司经常以各种方式被国有化
  + 由于国际市场变更太多，规模很大，所以持续监控很难

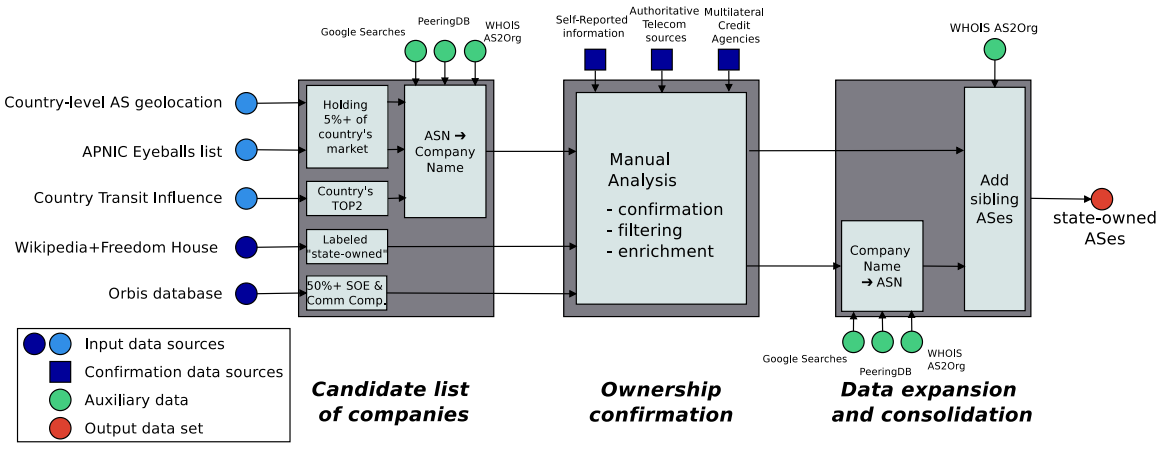
这篇文章的主要方法：

首先，明确国有公司的定义，多个比较权威的机构（IMF、DECO和欧洲委员会）对国有公司的定义标准不同，但是有几个主要的共识：

* 公司有自己独立的法人实体
* **该实体由政府单位部分控制**
* 实体从事商业或经济活动

作者团队主要采用第二个共识来定义本文中的国有企业：**国家控股占比超过50%的公司是国有公司**。

在本文的上下文中，作者将国有 AS 定义为国有互联网运营商拥有的 AS（即IRR中由国有互联网运营商控制）



接下来的数据挖掘和分类主要分为三个阶段：

* **识别 ASN 和公司名称**以获得要手动验证的候选公司名单
* 在手动验证期间，对这些数据进行**过滤和丰富**。 过滤的一个例子是排除少数国家参与的运营商，而丰富包括增加子公司
* 通过获取选定公司运营的 ASN 的最终列表并生成最终数据集（第 3 阶段）

总结：第二阶段的判断过程依旧包括很多手工操作，导致工作量上升，无法做到全自动；并且作者自己评价也称自己所收集的数据具备一定的时效限制性，即会因为公司所有权转让等原因，导致数据的时效性降低

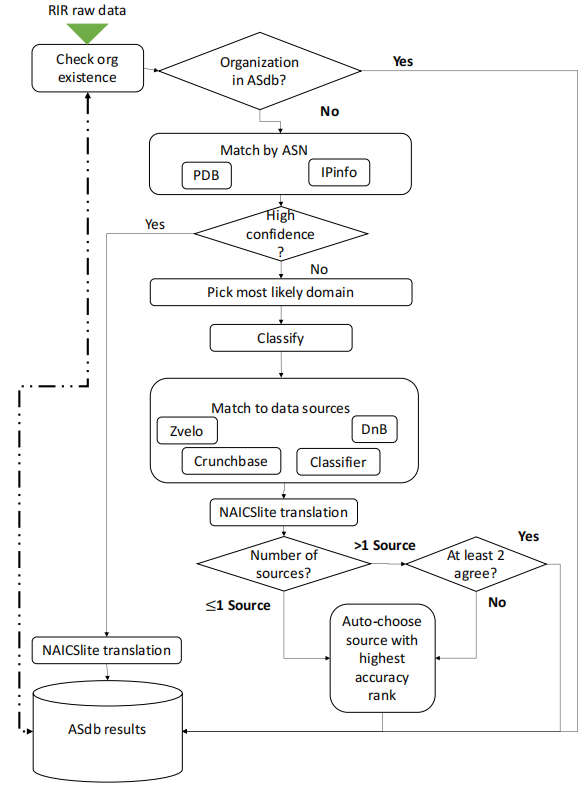
## “ASdb: A System for Classifying Owners of Autonomous Systems”论文解读

作者团队使用一些商务数据库加上机器学习的方法来准确分类AS，并构建了一个数据库ASdb

之前大部分的AS分类的相关研究，只将AS分类到大致的行业以及在互联网拓扑中的作用上，ASdb将组织分为 17 个行业类别和 95 个子类别

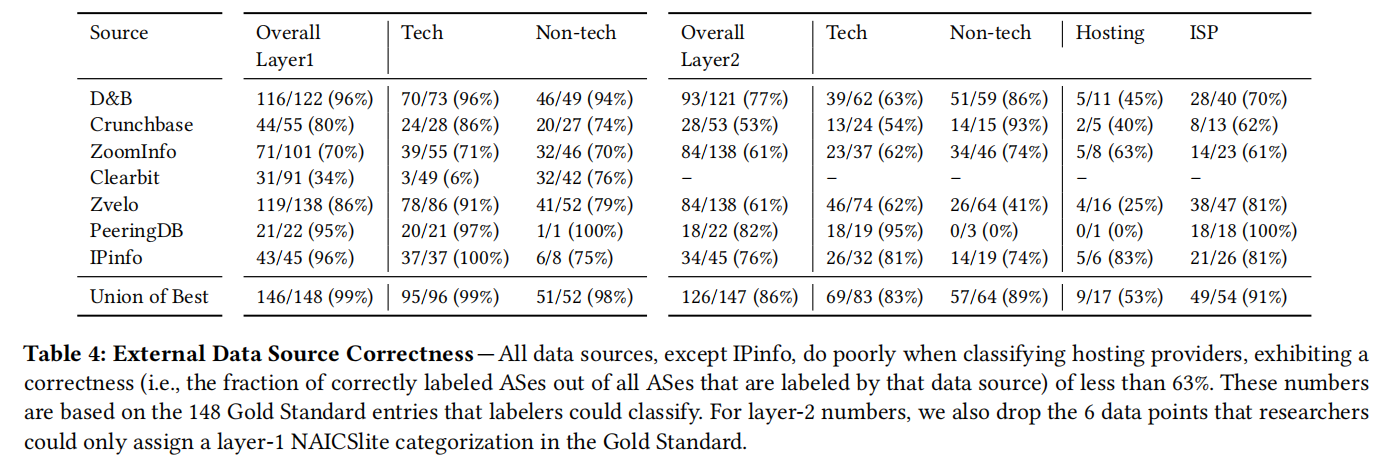
* 首先，虽然没有数据源可以提供有关 AS 的足够数据，但几乎所有 AS 都属于可识别的组织，并且存在维护和提供对业务记录的访问的成熟行业
* 并且所有AS都有可供访问的网站，并且网站上一般都有说明的文档

基于以上两点，作者首先使用**一些商业数据库，网站分类器和现有的AS分类数据集**来对AS进行分类；作者的实验结果表明，上面几种方式的组合可以很好地对大部分非技术领域的公司进行分类，但是无法处理两类AS：ISP和托管服务提供商

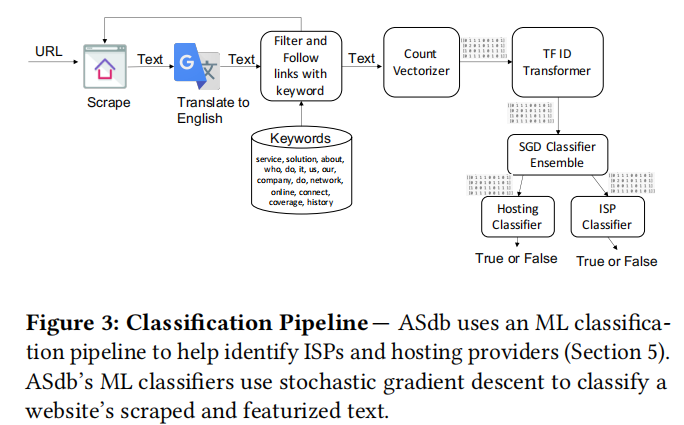


在前面的实验的基础上，作者提出了结合机器学习和使用crowdwork的方法来分类AS。机器学习可以分别以 94% 和 90% 的准确率正确分类 ISP 和云/托管服务提供商。 进一步的，crowdworkers 可以分别以 98.7% 和 94% 的准确率捕获 ML 故障并解决数据源分歧。 然而，激励众包准确性所需的货币成本引入了一个障碍，最终使众包对ASdb系统不切实际。

在基于前面所有方法的基础上，作者开发了ASdb。ASdb 使用可配置的内部匹配算法来统一所有组件，对数据源不一致和丢失信息进行了处理。 作者针对 620 个手动标记的 AS 评估 ASdb。 ASdb 为 96% 的 AS 提供了多层分类，顶级类别的准确率达到 93%，子类别的准确率达到 75%。



作者首先研究了当前AS分类数据集的分类能力，结果如上图所示；结果表明，当前的商业或者开源AS分类数据集在对非技术类的AS的分类能力比较强，但是对技术公司的区分能力很弱。超过一半的 AS 属于技术公司，业务数据集难以正确识别这些公司（例如，Zvelo 和 D&B 在识别托管服务提供商时实现了 25% 和 45% 的召回率）。 其次，对于 21% 的 AS，多个数据源不一致。



ASdb通过以上流程对一个AS公司的类型进行加以判断。主要是对ISP和服务托管商进行分类。在爬取下AS的网站后，通过关键词和SGD对AS进行分类

接着作者团队使用众包平台，通过人力解决两个问题：

* 判断前面的ML模型在哪些case出现错误
* 解决外部数据源不一致或覆盖不完整的情况

## “Revealing the Autonomous System Taxonomy: The Machine Learning Approach”论文解读

这篇文章介绍了一种基于机器学习技术的全新方法，将互联网中的所有 AS 映射到自然的 AS 分类中，并且介绍了分类的类别和分类所依据的一系列变量。

对AS进行分类有以下好处：

* 在拓扑建模中，AS 类型的知识对于使用现实的 AS 内和 AS 间路由器级拓扑来增强综合构建或测量的 AS 拓扑是必不可少的
* 使用AS类型在 研究AS级别的拓扑图的Internet演化的先决条件，因为不同类型的 AS 表现出不同的增长模式
* AS分类有利于对IP地址的用户类型映射

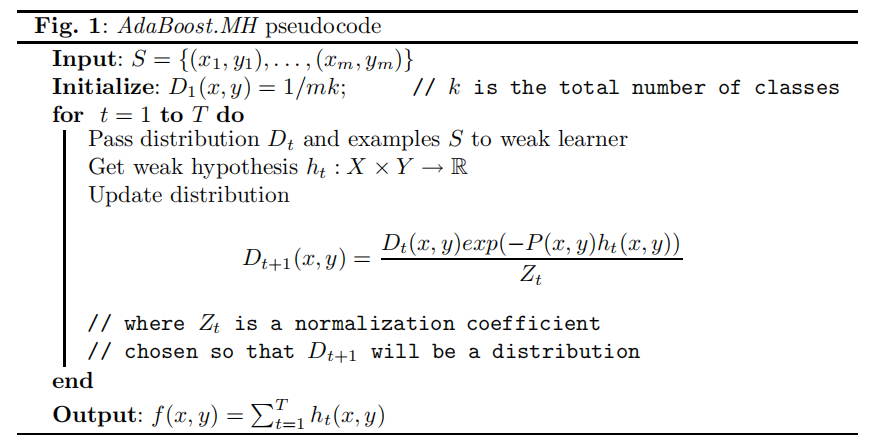
作者首先使用IRR和RouteViews的数据来对不同类型的AS进行区分，找出不同AS类型之间的区别；利用这些数据进行进一步推断，推断以下的属性作为判断依据的变量：

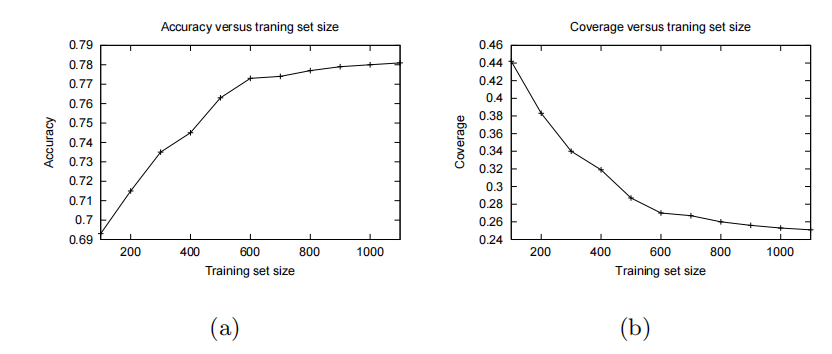
* 组织描述记录（描述属性）
* 推断的客户（customer）数量（客户属性）
* 推断的提供者（provider）的数量（提供者属性）
* 推断的对等点（peer）的数量（对等点属性）
* 通告的IP前缀（IP prefix）数（前缀属性）
* 等量的/24 前缀覆盖所有通告的IP 空间（空间属性）。

然后基于机器学习方法建造一个分类算法，使用这些差别将AS分成6个大类（每类拥有不同的网络属性和网络设施）：

1. 大型 ISP：大型骨干供应商，一级 ISP，拥有洲际网络。
2. 小型 ISP：具有小型城域或较大区域网络的区域和接入提供商。
3. 客户 AS：运行自己的网络但与前两类成员不同的公司或组织不提供 Internet 连接服务。 我们在这个类中找到了广泛的 AS，如网络托管公司、技术公司、咨询公司、医院、银行、军事网络、政府网络等。
4. 大学：大学或学院网络。 我们将这些网络与客户 AS 类别的成员区分开来，因为它们通常具有服务于数千个终端主机的更大的网络。
5. 互联网交换点（IXP）：作为前两类成员的互连点的小型网络。
6. 网络信息中心 (NIC)：承载重要网络基础设施的网络，例如根或 TLD 服务器。

然后作者使用AdaBoost算法，构建分类算法，构建分类模型





## “Towards an AS-to-Organization Map”论文解读

这篇文章提出了一种从WHOIS数据中推断AS类型的方法。

当前的AS拓扑相关研究存在这样的问题：

* 将网络抽象为一堆点与电和点之间的线会使得很多AS级别网络的语义信息丧失
* 当前的测量中获取准确的AS级别拓扑图很难

在这篇文章的研究中，从AS拓扑图升级为AS级别生态系统（ecosystem，我不知道这么翻译是否完全准确），AS生态系统主要是为了找到组织，以及组织的AS，还有IXP。

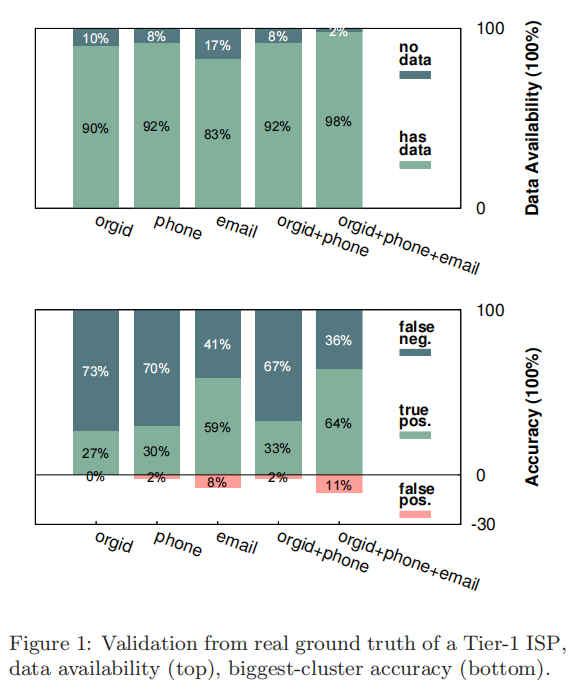
首先，作者采用一系列自动组合的从WHOIS数据库中提取AS-组织映射

作者主要关系三类属性：AS，组织和联络点（point-of-contact）

可以使用AS登记数据来寻找AS-to-organization关系，但是AS登记数据既不完善，格式也不对；不同的登记机构的格式也不一样，数据缺失很严重

在一些AS的记录中，存在OrgID和组织记录数据；OrgID用来指向organization，一些AS记录有OrgID；但是由于RIR允许机构使用多个OrgID，因此导致OrgID无法用于AS-to-organization映射。虽然OrgID不好用，但是AS记录中带有很多其他信息，包括联系人电话，邮件地址，鉴于这些电话拥有者多半是org管理者，可以很好地利用这些信息进行推断，进一步的，作者根据这些属性，提出了一种聚类算法：

* 从源数据中提取原生参数（raw attribute）-AS的对
* 转化为简单参数（simple attribute）-AS的对
* 放弃许多组织共有的通用属性
* 使用相关的简单、非通用属性对所有 AS 进行聚类：
  + 使用一些未聚类的 AS ai 发起一个新聚类（cluster） ci。 然后重复寻找AS直到没有AS再能合并到 ci 中
  + 通过ci中所有 AS 的属性的联合来识别 ci
  + 通过一些匹配的属性将其他一些 AS aj 与ci合并



主要使用了以下几个属性：

* OrgID
* 电话号码
* 电子邮件域名
* 混合参数（混合前三种）

然后，使用验证数据进行验证，根据validation结果，对WHOIS数据的质量和时效性进行评估

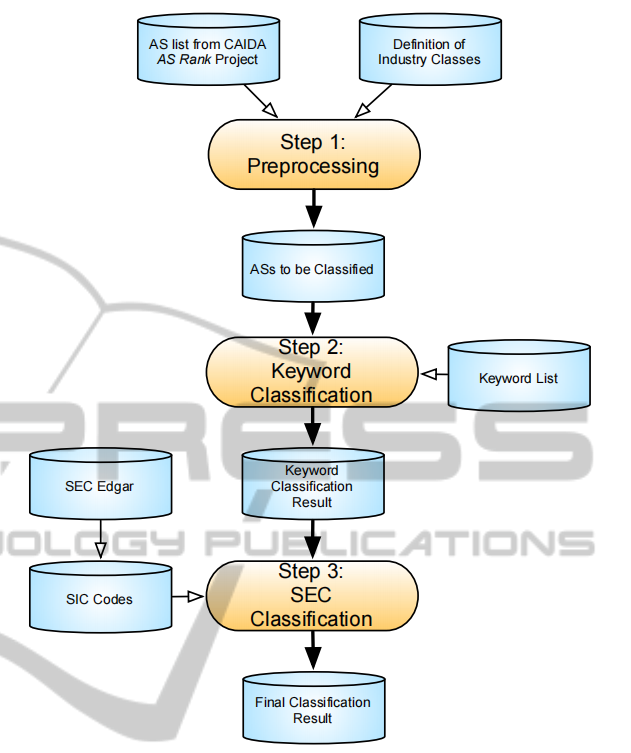
# "Who Runs the Internet? Classifying Autonomous Systems into Industries"论文解读

这篇文章的作者提出一种新颖的分类方法，将AS特定的数据与美国证券交易局的业务数据相结合。并且根据NAICS，提出了更多的AS分类。

作者的方法主要是分析公开ASN的登记信息，并且根据交易数据对AS进行更加细致的分类，分成更细致的行业类型，对网络的组织架构和经济模式进行更好的理解。

数据来源：

* CAIDA AS Rank
* The U.S. Securities and Exchange Commission (SEC)：对AS的Organization进行细致分类
* RIR



首先对NAICS的类别进行处理，为了适应AS分类这一研究，作者在NAICS的分类基础上进行了一系列合并，删除和修改操作

预处理：首先用AS Rank的数据生成AS List，然后通过IRR的数据洗掉一部分已经不再活动的AS;然后将没有声明organization的AS记录删除

关键词识别：使用bi-grams和tri-grams来统计关键词的词频，（tri-grams）三个字的出现最少5词，（bi-grams）2字的出现至少10次，1字的至少出现20次才被标记为某个类别。并且通过循环迭代减少错误分类的次数；一次分类完成后进行手动查询，识别错误，更换或者删除关键词；同时再对每次循环没有被成功分类的AS的List进行检查，根据没被分类的情况选择添加或删除关键词

SEC分类：通过SEC的信息进行补充，查找，并且去掉查找结果多个的结果

## 总结

| 发表时间 | 发表会议 | 论文标题 | 研究内容 | 研究方法 | 分类属性 |
| --- | --- | --- | --- | --- | --- |
| 2006 | PAM | Revealing the Autonomous System Taxonomy: The Machine Learning Approach | 介绍了一种基于机器学习技术的全新方法，将互联网中的所有 AS 映射到自然的 AS 分类中，并且介绍了分类的类别和分类所依据的一系列变量 | 使用**IRR和RouteViews**的数据来对不同类型的AS进行区分，然后利用这些数据进行进一步推断，利用推断出的属性进行**机器学习AdaBoost** | 网络属性 |
| 2010 | IMC | Towards an AS-to-Organization Map | 从**WHOIS信息**中提取AS类别 | 根据AS的OrgID以及各**种联络方式信息**进行判断，并进行聚类 | 行业属性 |
| 2016 | ICWST | Who Runs the Internet? Classifying Autonomous Systems into Industries | 将AS特定的数据与美国证券交易局的业务数据相结合。并且根据NAICS，提出了更多的AS分类 | 核心方法是使用**关键词**进行识别；用到了多次迭代的方法来提升准确度 | 行业属性 |
| 2021 | IMC | Identifying ASes of State-Owned Internet Operators | 识别全球范围内的国有网络operator以及他们的ASN/研究网络Operator的ASN以及其国有属性 | 通过**多数据源进行验证**的方法来识别国有ASN | 国有属性 |
| 2021 | IMC | ASdb: A System for Classifying Owners of Autonomous Systems | 使用一些商务数据库加上机器学习的方法来准确分类AS，并构建了一个数据库ASdb | 在商业数据库，网站分类器和现有AS分类数据集的基础上，加上了结合**机器学习和众包**的方法，提高识别技术类AS的准确率 | 行业属性 |