Eine Kurzfassung zu „ASdb: A System for Classifying Owners of Autonomous Systems“

Dominik Stahmer*
Technische Hochschule Ingolstadt
dominik.stahmer@posteo.de

Maya Ziv†
Stanford University
mziv@cs.stanford.eu

Liz Izhikevich†
Stanford University
lizhikev@stanford.edu

Kimberly Ruth†
Stanford University
kcruth@cs.stanford.edu

Katherine Izhikevich†
UC San Diego
kizhikev@ucsd.edu

Zakir Durumeric†
Stanford University
zakir@cs.stanford.edu

ABSTRACT

Das Internet als „Netz aus Netzen“ setzt sich aus den durch das Protokoll BGP definierten Autonomen Systemen zusammen. Autonome Systeme werden von verschiedensten Organisationen mit unterschiedlichen Interessen betrieben. Regierungen, Internet Service Provider, Krankenhäuser, etc. können alle bei Bedarf an BGP Teilnehmen. Die Einordnung Autonomer Systeme in Kategorien ist dabei ein vielfach erforschtes, aber keineswegs abschließend gelöstes Problem. Der Vorliegende Artikel „ASdb: A System for Classifying Owners of Autonomous Systems“ [15] leistet einen wichtigen Beitrag in der Kategorisierung Autonomer Systeme und soll in diesem Dokument zusammengefasst, auf dem Niveau eines Master-Studenten an der THI erklärt und bewertet werden.

1 VORWORT

Dieses Dokument ist eine Zusammenfassung und damit auch eine Vereinfachung. Dabei müssen unweigerlich Details ausgelassen werden. Die Lektüre des eigentlichen Paper ist daher empfehlenswert. Der Natur einer Zusammenfassung nach sind beinahe alle Ideen in diesem Dokument aus [15] übernommen. Es wird Aufgrund der Länge dieses Dokuments zu Gunsten der Lesbarkeit auf korrekte Zitierung jeder Idee aus [15] verzichtet.

Jegliche hier geübte Kritik an dem Artikel ist als konstruktiv zu verstehen. An dieser Stelle ist es wohl angebracht hervorzuheben, dass das Paper im Allgemeinen meiner Meinung nach inhaltlich und formal sehr hochwertig ist und einen wesentlichen Beitrag zur Forschung über Autonome Systeme leistet.

*Autor dieser Kurzfassung
†Autor von „ASdb: A System for Classifying Owners of Autonomous Systems“

SEMI.1 SS2022, 12.04.2022–05.07.2022, Ingolstadt, Deutschland
2022.

2 HINTERGRUND

Zum Verständnis des vorliegenden Paper sind einige grundlegende Kenntnisse über BGP, Autonome System und das WHOIS-Protokoll notwendig. Da die Zielgruppe dieser Zusammenfassung Master-Studenten an der Technische Hochschule Ingolstadt sind und dort meines Wissens diese Themen nur sehr kurz und oberflächlich behandelt wurden, wird in diesem Abschnitt ein kurzer Überblick über die genannten Technologien gegeben.

2.1 BGP

Das Border Gateway Protokoll (BGP) ist eines der grundlegendsten Bausteine des modernen Internet. Es ermöglicht Routing über topologisch beliebig angeordnete Autonome Systeme (AS). Ein AS ist ein Zusammenschluss von Netzwerken, das üblicherweise von *einer* Entität betrieben wird, etwa von einer Regierung, einer Hochschule oder einem Krankenhaus. Innerhalb eines AS wird Routing üblicherweise über dynamische Routing-Protokolle wie etwa Open Shortest Path First (OSPF) ermöglicht. Die Rolle von BGP als Routing Protokoll *zwischen* Autonomen Systemen wird in Abbildung 1 beschrieben.

Die Topologie des Internets ist damit also die Gesamtheit aller BGP Nachbarschaftsbeziehungen. Es wird dabei üblicherweise zwischen zwei verschiedenen Arten von BGP Nachbarschaftsbeziehungen unterschieden:

- Man sprint von *„Peering“*, wenn zwei BGP Nachbarn ähnlicher Größe Routing gegenseitig unentgeltlich über ihre Netzwerke ermöglichen.
- Im Gegensatz dazu spricht man von einer *„Transit“* Verbindung, wenn ein BGP Teilnehmer einen anderen bezahlt, um Zugriff auf dessen Routen zu bekommen.

Grundsätzlich ist es jeder Organisation beliebiger Größe mit dem notwendigen technischen Know-How möglich, an BGP teilzunehmen. Viele BGP Teilnehmer sind große oder für das Internet zentrale Unternehmen wie etwa Internet Service Provider (ISPs) oder Web Hosting Anbieter. Aber auch

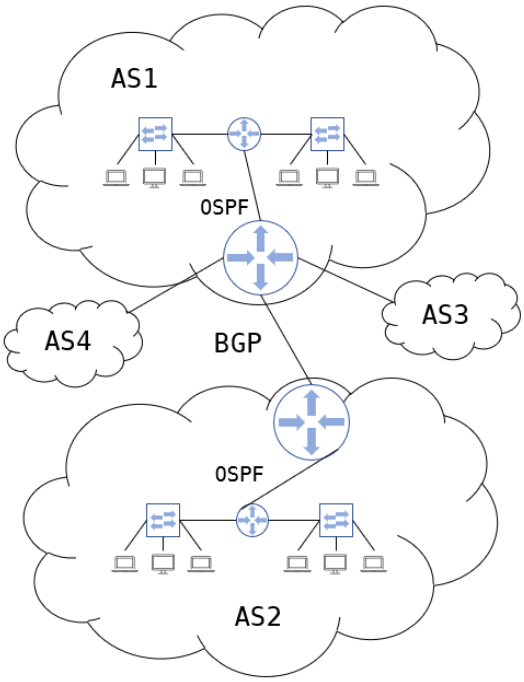

Abbildung 1: Schaubild zur Rolle von BGP als Routing Protokoll zwischen Autonomen Systemen

manche Unternehmen mittlerer und kleiner Größe nehmen an BGP Teil, um zum Beispiel sogenanntes „Multihoming“ zu betreiben, also die Ausfallsicherheit ihrer Internetverbindung zu erhöhen, indem sie für Transit Verbindungen zu mehreren ISPs bezahlen.

Wichtig zum Verständnis des „ASdb“ Papers ist hier vor allem, dass Unternehmen aller Größen und Branchen, sowie Regierungen aller Welt aus verschiedensten Gründen an BGP teilnehmen und die dadurch entstehende Topologie keinesfalls trivial ist.

Zur Teilnahme an BGP ist abgesehen von entsprechender Hardware, Software und Know-How eine Registrierung als Autonomes System notwendig. Diese Registrierung übernehmen die von der Internet Assigned Numbers Authority (IANA) beauftragten Regional Internet Registries (RIRs). Die RIRs vergeben dabei an jedes AS eine weltweit eindeutige AS Nummer (ASN).

2.2 WHOIS

WHOIS ist ein Protokoll, mit dem Informationen zu Internet-Domains, IP Adressen und Autonomen Systemen abgefragt werden können [8]. Dabei ist das Protokoll denkbar einfach: Nach Aufbau einer TCP Verbindung schickt der Client die IP Adresse, Domain oder AS Nummer als Text an den Server, gefolgt von einem Newline Character. Der Server antwortet mit den ihm verfügbaren Informationen als Freitext. In dem Protokoll ist nicht definiert, auf welchem Server welche Informationen verfügbar sein sollten. Diese Aufgabe ist damit den Clients überlassen. So heißt es beispielsweise in der Anleitung zu dem in Debian GNU/Linux eingesetzten WHOIS Client: „This version of the WHOIS client tries to guess the right server to ask for the specified object.“ [10].

Üblicherweise finden sich Informationen zu einer AS Nummer auf dem WHOIS Server der für das AS zuständigen RIR. Bei einer Anfrage nach Informationen zum Klinikum Ingolstadt (AS51378) über den in Debian eingesetzten WHOIS Client antwortet beispielsweise die europäische RIR „Réseaux IP Européens Network Coordination Centre“ (RIPE).

Meist werden in dem Format der Antwort gewisse Konventionen eingehalten, um eine maschinelle Weiterverarbeitung zu ermöglichen. Leider gibt es verschiedene Konventionen, in denen viele relevante Felder nur optional sind [1, 2, 5].

3 EINFÜHRUNG

In diesem Abschnitt werden Leser mit den grundlegenden Ideen und Zielen des Papers zu „ASdb“ vertraut gemacht.

3.1 Zielsetzung

Wie der Titel „ASdb: A System for Classifying Owners of Autonomous Systems“ bereits verrät, ist es Ziel des Papers, Autonome Systeme in verschiedene Klassen einzuordnen. Damit sollen scheinbar einfache Fragen beantwortet werden können, etwa „In welchen Branchen (abgesehen von IT) wird am häufigsten BGP eingesetzt?“. Im Allgemeinen soll damit eine Grundlage für weitere Forschung geschaffen werden.

3.2 AS Klassifizierung: Ausgangssituation

Die Zielsetzung impliziert bereits die Schlussfolgerung dieses Abschnitts: Die existierenden Systeme zur Klassifizierung Autonomer Systeme sind durch ihre Mängel ungeeignet, genaue und zuverlässige Aussagen über die Kategorisierung aller Autonomen Systeme zu treffen. Zur Begründung dieser Schlussfolgerung werden im Paper dazu die folgenden Texte und Webseiten behandelt:

- Die Organisation CAIDA (Center for Applied Internet Data Analysis) verwaltet einige verschiedene Datensätze rund um das Internet. Einer dieser Datensätze war bis vor einigen Monaten eine Tabelle zur Klassifizierung Autonomer Systeme [7]. Da die Genauigkeit der Daten über die Jahre stark abnahm, entschied sich CAIDA im Januar 2021, die Daten vorübergehend nicht mehr zur Verfügung zu stellen.
- Baumann und Fabian [6] haben 2014 auf Grundlage einer Text-Klassifizierung von WHOIS Datensätzen eine Einteilung ca 57% aller ASe in 18 Kategorien vorgenommen. Zusätzlich haben Sie die Daten ca. 500

Autonomer Systeme über Daten der U.S. Securities and Exchange Commission (SEC) angereichert.
- Dhamdhere und Dovrolis [9] haben 2011 aufgrund ihrer Topologie ca. 80% aller ASe in vier grobe Kategorien eingeteilt.
- Die Webseite PeeringDB [4] teilt ca. 15% aller ASe in 11 grobe Kategorien ein. Einzelne Einträge werden von Nutzern gepflegt.
- Die Webseite IPInfo [3] teilt über eine nicht spezifizierte Methodik ca. 30% aller ASe in vier grobe Kategorien ein.

Aus dieser Liste wird ersichtlich, dass alle vorhandenen Klassifizierungssysteme in mindestens einer der Dimensionen Abdeckung, Genauigkeit und Granularität der Kategorien (starke) Mängel aufweist. Hieraus begründet sich die Relevanz von ASdb.

3.3 NAICSlite

Eine offensichtlich zu klärende Frage bei jeder Kategorisierung ist die Entwicklung der zu verwendenden Kategorien. Für ASdb wurde hierfür zunächst das North American Industry Classification System (NAICS)[1] in Betracht gezogen. Dabei handelt es sich um ein durch die US-Regierung standardisiertes und weit verbreitetes Klassifizierungssystem für Unternehmen aller Branchen. NAICS ist *sehr* ausführlich. Es gibt in etwa 2000 Kategorien[2] und das Handbuch umfasst mehr als 900 Seiten[3]. Andererseits ist NAICS im Bereich von Technologie-Unternehmen oft nicht granular genug: So umfasst die Kategorie 518210 „Computing Infrastructure Providers, Data Processing, Web Hosting and Related Services“ sowohl ISPs als auch Web Hosting Anbieter. Eine Differenzierung zwischen diesen Kategorien ist also mit NAICS nicht möglich.

Aufgrund der hohen Komplexität von NAICS und der trotzdem im Bereich von Technologie-Unternehmen unzureichenden Granularität haben sich die Autoren von ASdb dazu entschieden ein eigenes, an NAICS angelehntes Klassifizierungssystem zu erstellen: *NAICSlite*.

NAICSlite ist einfach strukturiert: Es werden 17 Oberkategorien definiert (im Weiteren als „Layer 1“ Kategorien bezeichnet), die jeweils bis zu 9 Unterkategorien (im Weiteren als „Layer 2“ Kategorien bezeichnet) enthalten. Ein ISP wie etwa die deutsche Telekom würde beispielsweise in NAICSlite auf Layer 1 als „Computer and Information Technology“ und auf Layer 2 als „Internet Service Provider (ISP)“ bezeichnet werden. NAICSlite ist über den Internet-Auftritt von ASdb als `.csv`-Datei verfügbar [15]. NAICSlite ist durch seine insgesamt 100 Kategorien deutlich granularer als die in Abschnitt 3.2 beschriebenen existierenden Lösungen.

[1] https://www.census.gov/naics
[2] https://www.census.gov/naics/2022NAICS/6-digit_2022_Codes.xlsx
[3] https://www.census.gov/naics/reference_files_tools/2022_NAICS_Manual.pdf

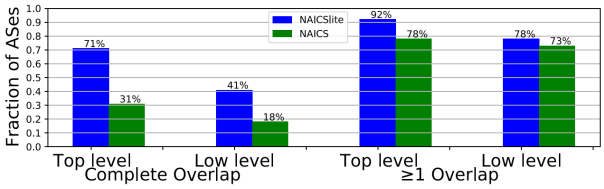

Abbildung 2: Konsens in der Klassifizierung der ASe im Gold-Standard per NAICS und NAICSlite – „Complete Overlap“ bedeutet, dass beide Wissenschaftler, die das betrachtete AS klassifiziert haben, die exakt gleichen Kategorien wählen. „≥ 1 Overlap“ hingegen bedeutet, dass beide Wissenschaftler mindestens eine der gleichen Kategorien wählen.

3.4 Der Gold-Standard

Zur Bewertung verschiedener (Kombinationen) externer Datenquellen (siehe Abschnitt 4) wird für ASdb zunächst ein im Weiteren als „Gold-Standard“ bezeichneter Datensatz entwickelt. Für diesen Datensatz werden sechs Wissenschaftler aus dem Fachgebiet der Netzwerktechnik beauftragt, 150 zufällig gewählte ASe nach NAICS (nicht NAICSlite) zu kategorisieren. Dabei werden die ASe so auf die Wissenschaftler aufgeteilt, dass jedes AS von zwei der Wissenschaftler unabhängig eingeordnet wird.

Der Gold-Standard wurde vor der Entstehung von NAICSlite entwickelt. Tatsächlich war der geringe Konsens der beauftragten Wissenschaftler bei der Kategorisierung der ASe für den Gold-Standard ein Auslöser für die Entwicklung von NAICSlite. Es wurde daher eine leider nicht näher beschriebene automatische Übersetzung der für den Gold-Standard gefundenen NAICS Kategorien (zuzüglich nicht näher beschriebenen „zusätzlichen Codes“) zu NAICSlite durchgeführt.

Aus dieser Übersetzung lässt sich bestätigen und quantifizieren, dass NAICSlite den Konsens stark erhöht. In Abbildung 2 ist zu erkennen, dass beispielsweise bei der Zuweisung von „Top level“ (Layer 1) Kategorien sich beide beauftragte Wissenschaftler bei NAICSlite zu 71% vollständig einig waren, während es bei NAICS nur 31% waren.

Es sei erwähnt, dass eine Vereinfachung von NAICS keineswegs eine neue Idee ist. Baumann und Fabian [6] haben bereits 2014 ähnliche Überlegungen angestellt. Allerdings ist NAICSlite mit seinen etwa 100 Kategorien deutlich granularer und der Vorteil in der Konsens-Findung wurde hier (wenn auch durch eine nicht vollständig transparente Methodik) quantifiziert.

Quelle	Durchsuchbar	Klassifizierung	Kostenlos
D&B	N,W,P	NAICS	
Crunchbase	N,W	Proprietär	✓
ZoomInfo	N,W,P	NAICS	
Clearbit	N,W	NAICS*	
PeeringDB	N,A	Proprietär	✓
IPInfo	N,A	Proprietär	
Zvelo	W	Proprietär	

Tabelle 1: Potentielle externe Datenquellen – Aufgeführt sind die zur Suche verfügbaren Attribute (N = Name, W = Webseite, P = Physische Adresse, A = ASN), das verwendete Klassifizierungssystem und ob automatisierter Zugriff auf die Quelle kostenlos verfügbar ist. *Clearbit unterstützt nur zwei der sechs Ziffern von NAICS und stellt granularere Informationen in einem proprietären Klassifizierungssystem bereit.

4 BEWERTUNG EXTERNER DATENQUELLEN

Wie in Abschnitt 3.2 geschildert, weisen existierende Quellen zur Klassifizierung Autonomer Systeme (starke) Mängel auf. Für ASdb sollen deshalb zusätzlich Datenquellen mit einbezogen werden, die sich nicht direkt nach AS Nummer durchsuchen lassen. Gemeint sind vor allem an Unternehmen und Marketing Abteilungen gerichtete (meist kommerzielle) Angebote wie etwa Dun & Bradstreet (D&B) oder Zvelo. Diese lassen sich alle nach der Webseite des Unternehmens durchsuchen, zum Teil erlauben sie auch die Suche nach Namen oder physischen Adressen. Tabelle 1 bietet einen Überblick.

In diesem Abschnitt soll eine Entscheidung getroffen und rechtfertigt werden, welche dieser Datenquellen in ASdb eingesetzt werden sollen. Dazu werden zunächst die Qualitätsmerkmale Abdeckung, Trefferquote und Genauigkeit anhand des Gold-Standards ermittelt. Weiterhin wird die automatisierte Ermittlung der Suchparameter diskutiert (für Datenquellen ohne Möglichkeit zur Suche direkt nach AS Nummer). Zuletzt wird noch ein den Gold-Standard ergänzender Datensatz entwickelt, der zur Bewertung der Abdeckung, Trefferquote und Genauigkeit in den weniger vertretenen NAICSlite Kategorien eingesetzt wird.

4.1 Abdeckung

Die „Abdeckung“ einer Datenquelle ist hier definiert als der prozentuale Anteil der in der Datenquelle (händisch) auffindbaren ASe des Gold-Standards. Die Abdeckung der verschiedenen Datenquellen ist in Abbildung 3 dargestellt. Es fällt auf, dass Dun & Bradstreet zusammen mit Zvelo die deutlich höchste Abdeckung erreichen.

Source	Coverage	Tech	Non-Tech
D&B	122/148 (82%)	73/96 (76%)	49/52 (94%)
Crunchbase	55/148 (37%)	28/96 (29%)	27/52 (52%)
ZoomInfo	101/148 (68%)	55/96 (57%)	46/52 (88%)
Clearbit	91/148 (61%)	77/96 (80%)	57/52 (90%)
Zvelo	138/148 (93%)	86/96 (90%)	52/52 (100%)
PeeringDB	22/148 (15%)	21/96 (22%)	1/52 (2%)
IPinfo	45/148 (30%)	37/96 (39%)	8/52 (15%)
All - ZI, CL	148/148 (100%)	96/96 (100%)	52/52 (100%)

Abbildung 3: Abdeckung des Gold-Standards durch externe Datenquellen – insgesamt sowie für separat für Technologie-Unternehmen und Unternehmen außerhalb der Technologie-Branche. Die Bedeutung der Abkürzungen „ZI“ und „CL“ werden im Paper leider nicht definiert.

ISPs im Gold-Standard	ISPs in Datenquelle *D*
Vodafone	Vodafone
Telekom	Telekom
O2*	Klinikum Ingolstadt**
	Google**

Tabelle 2: Beispiel-Daten zur Definition der Metriken Korrektklassifikationsrate, Trefferquote und Genauigkeit im Kontext von ASdb
*: Fehlt in Datenquelle *D* (falsch negativ)
**: Fehlerhaft in *D* als ISP eingeordnet (falsch positiv)

4.2 Korrektklassifikationsrate, Trefferquote und Genauigkeit

Bei den Begriffen Korrektklassifikationsrate (englisch: *accuracy*), Trefferquote (englisch: *recall*) und Genauigkeit (englisch: *precision*) handelt es sich um wohldefinierte statistische Metriken. Am einfachsten sind sie hier an einem Beispiel erklärt. Angenommen, es wird die Anfrage „Liste alle ISPs“ eine Datenquelle *D* gestellt. Die Antwort von *D* könnte dann so aussehen, wie in Tabelle 2.

Die Ergebnisse „Vodafone“ und „Telekom“ sind richtig positiv. Bei den Ergebnissen „Klinikum Ingolstadt“ und „Google“ handelt es sich um falsch positive Ergebnisse. Das fehlende Ergebnis „O2“ ist falsch negativ.

Die Definitionen der genannten Metriken können wie folgt formuliert werden:

$$\text{Korrektklassifikationsrate} = \frac{\text{richtig positiv} + \text{richtig negativ}}{\text{Alle Ergebnisse}}$$

$$\text{Trefferquote} = \frac{\text{richtig positiv}}{\text{richtig positiv} + \text{falsch negativ}}$$

$$\text{Genauigkeit} = \frac{\text{richtig positiv}}{\text{richtig positiv} + \text{falsch positiv}}.$$

Damit gilt in dem geschilderten Beispiel

$$\text{Korrektklassifikationsrate}_{\text{ISP}} = \frac{2+0}{5} = 0,4$$

$$\text{Trefferquote}_{\text{ISP}} = \frac{2}{2+1} \approx 0,66$$

$$\text{Genauigkeit}_{\text{ISP}} = \frac{2}{2+2} = 0,5.$$

Die Begriffe werden hier so genau definiert, da sie im behandelten Paper leider oft recht unscharf verwendet werden. Beispielsweise behandelt Tabelle 4 in [15] laut ihrer Beschriftung die Korrektklassifikationsrate („accuracy“ im Englischen, hier als „correctness“ bezeichnet und im Fließtext definiert). Im Fließtext davor hingegen (Abschnitt 3.3, Unterabschnitt „Recall and Precision“, S.706) heißt es „The data sources with the highest overall layer 1 *recall* (96%) are D&B and IPinfo (Table 4) [...]“; Hier wird also von der Trefferquote („recall“ im Englischen) gesprochen. Die Mehrheit des Fließtextes scheint sich auf diese Tabelle als Quelle der Trefferquote zu beziehen, vermutlich handelt es sich hier also um einen Fehler in der Beschriftung.

Allgemein ist die Behandlung solcher Messergebnisse der inhaltlich schwächste Aspekt des Artikels. Zusätzlich zur Unschärfe in der Verwendung der Begriffe werden viele Daten nicht in Tabellen aufgeführt und müssen deshalb mühsam aus dem Fließtext extrahiert werden. Selbst nachdem diese Daten aus dem Fließtext gelesen werden, sind sie noch immer unvollständig. Zur schlüssigen Argumentation werden diese fehlenden Daten zwar nicht benötigt, für zukünftige weitere Forschung wären sie aber sicherlich interessant gewesen. Die Tabelle 3 (in diesem Dokument) gibt einen Überblick über die verfügbaren Daten zur Trefferquote und Genauigkeit.

Aus Tabelle 3, Abbildung 3 und Tabelle 4 in [15] kann entnommen werden, dass die verfügbaren Datenquellen im Allgemeinen vielversprechend hohe Werte für Abdeckung, Genauigkeit und Trefferquote bei der Kategorisierung von Unternehmen außerhalb der Technologie-Branche erreichen. Unternehmen in dieser Branche hingegen – speziell die zwei häufigsten Kategorien (ISPs und Web Hosting Anbieter) – können nicht zuverlässig unterschieden werden.

Datenquelle	Trefferquote (aus Tabelle 4)	Genauigkeit (aus Fließtext)
D&B	96%	
Crunchbase	80%	
ZoomInfo	70%	66%
Clearbit	34%	55%
Zvelo	86%	
PeeringDB	95%	95%
IPinfo	96%	96%

A: Genauigkeit und Trefferquote für die Gesamtheit aller Layer 1 NAICSlite Kategorien

Datenquelle	Genauigkeit	Trefferquote
Alle	96%-100%	86%-100%

B: Genauigkeit und Trefferquote für die Layer 1 NAICSlite Kategorien „Bildung“ und „Finanzen“ zusammen (alle Daten aus dem Fließtext übernommen)

Datenquelle	Trf. WH	Gen. WH	Trf. ISP	Gen. ISP
D&B Crunchbase ZoomInfo Clearbit Zvelo	45% 25%	78% 86%	70% 	89%
PeeringDB IPinfo			100%	100%

C: Genauigkeit (Gen.) und Trefferquote (Trf.) für die Layer 2 NAICSlite Kategorien „Web-Hosting“ (WH) und „ISP“ (alle Daten aus dem Fließtext übernommen)

Tabelle 3: Genauigkeit und Trefferquote externer Datenquellen bezüglich des „Gold-Standards“

4.3 Der Uniforme Gold-Standard

Die meisten Einträge im Gold-Standard sind aus den Layer 1 NAICSlite Kategorien „Computer and Information Technology“, sowie „Finance and Insurance“ und „Education and Research“. Da außerhalb dieser drei Kategorien so wenige Einträge im Gold-Standard vorhanden sind, wurde zur besseren Bewertung der Abdeckung, Trefferquote und Genauigkeit der Datenquellen in anderen Layer 1 NAICSlite Kategorien der „uniforme Gold-Standard“ entwickelt. Dabei handelt es sich um einen neuen Datensatz 320 kategorisierter Autonomer Systeme, die näherungsweise gleichmäßig über alle Layer 1 NAICSlite Kategorien verteilt sind. Zu kritisieren ist, dass an keiner Stelle erwähnt wird, *wie* diese neuen 320 ASe kategorisiert wurden. (Wurde wieder eine Gruppe Wissenschaftler beauftragt?)

Tabelle 11 in [15], die hier aus Platzgründen nicht übernommen wird, zeigt die Abdeckung und Genauigkeit im uniformen Gold-Standard nach Layer 1 NAICSlite Kategorie. Es fällt auf, dass in den meisten Kategorien mindestens eine Datenquelle annähernd 100% Genauigkeit erreicht. Wenn sich mehrere Datenquellen einig sind, steigt die Genauigkeit weiter.

4.4 Ermittlung der Suchparameter

Die Daten aus den letzten beiden Abschnitten wurden durch händische Identifizierung der ASe in den Datenquellen ermittelt. Dieser Ansatz skaliert natürlich nicht auf die Gesamtheit aller ASe. Daher bedarf es eines Algorithmus, der automatisch eine AS Nummer einem Eintrag in einer Datenquelle zuordnen kann. Für manche Datenquellen ist dies trivial, da sie sich direkt nach ASN durchsuchen lassen können (IPInfo und PeeringDB), für alle anderen müssen mindestens einer der Suchparameter Name, physische Adresse oder Webseite aus der ASN ermittelt werden (siehe Tabelle 1). Da Clearbit und ZoomInfo beide keinen vollen Datenzugriff für akademische Zwecke anbieten, werden sie nicht weiter betrachtet. Damit bleiben noch D&B, Zvelo und Crunchbase, für die die genannten Suchparameter ermittelt werden müssen.

Ermittlung der Webseite. Die Webseite eines AS ist meist nicht direkt in dem WHOIS Eintrag des RIRs eingetragen. Allerdings werden oft E-Mail Adressen (zur Meldung von Fehlern oder Missbrauch) angegeben, die häufig die Webseite im Domain-Teil der Adresse enthalten. Auf dieser Grundlage entwickeln die Autoren von ASdb zwei Strategien:

- „*least common domain*“: Wähle die im WHOIS Eintrag am seltensten vorkommende Webseite. Die Begründung für diese Strategie ist, dass viele der angegeben E-Mail Adressen bei Drittanbietern (GMail, Outlook, etc.) registriert sind. Diese Einträge sollen durch diese Strategie ignoriert werden.
- „*most similar domain*“: Wähle die Webseite aus dem WHOIS Eintrag, deren Titel „am ähnlichsten“ zu dem Namen des AS ist. Leider wird hier nicht klargestellt, wie genau „Ähnlichkeit“ definiert wird (Menge gleiche Zeichen, Soundex, oder ein anderes Verfahren?). Die Methodik zur Ermittlung des Namens wird in Appendix A in [15] definiert. Der Titel der Webseite wird nicht klar definiert, auch wenn hier wohl von dem `<title>` Element im HTML `<head>` ausgegangen werden kann. Wenn die Webseite momentan nicht erreichbar ist, wird der Name der Webseite direkt anstatt dem Titel verwendet.

Mit der Strategie „most similar domain“ wird die Korrekte Webseite zu 91% gefunden, mit der Strategie „least common domain“ mit einer nur geringfügig kleineren Wahrscheinlichkeit von 90%. Daher wird bei der Implementierung von ASdb in Abschnitt 6 „most similar domain“ verwendet werden. Es ist davon auszugehen, dass diese Werte auf Grundlage des Gold-Standards ermittelt wurden, auch wenn das nicht konkret erwähnt wird.

Zvelo. Da Zvelo ein System zur automatischen Klassifizierung von Webseiten ist, bedeutet eine korrekt identifizierte Webseite direkt, dass ein Eintrag in Zvelo zugeordnet werden kann. Daher ist unter der Verwendung von „most similar domain“ bei Zvelo mit der korrekten Identifizierung eines gegebenen AS von 91% zu rechnen.

Crunchbase. Suchanfragen an Crunchbase lassen als Parameter zusätzlich zur Webseite auch den Namen des Unternehmens zu. Wenn Crunchbase eine korrekte Webseite als Parameter erhält, wird eine Abdeckung des Gold-Standards von 12% erreicht und innerhalb der Abdeckung zu 100% der Suchanfragen der korrekte Eintrag gefunden. Wenn die Webseite nicht verfügbar ist, wird im Kontext von ASdb stattdessen eine Suche nach dem in Token übersetzten Namen des AS durchgeführt. Dabei werden zu 15% der ASe im Gold-Standard Ergebnisse gefunden, die zu 95% die Organisation korrekt identifizieren.

Es bleibt unklar ob, wenn eine Webseite verfügbar ist, Crunchbase im Kontext von ASdb den Namen als zusätzlichen Parameter erhält. Das Verfahren zur Ermittlung des Namens eines AS wird im Artikel in Appendix A beschrieben.

Dun & Bradstreet. Dun & Bradstreet lässt sich nach Name, Webseite und physischer Adresse durchsuchen. Dabei wird in der Antwort die durch den Algorithmus von D&B für die Ergebnisse ermittelte Konfidenz mit angegeben. Solange dieser über einem Wert von 6 liegt, wird in über 80% der Fälle das gesuchte AS in der Datenbank gefunden. Sobald dieser Konfidenz-Wert unter 6 liegt, sinkt diese Treffergenauigkeit auf etwa 50% ab. Leider wird nicht angegeben, wie oft welche Konfidenz-Werte vorkommen, so dass die tatsächliche Genauigkeit/Abdeckung nicht aus dem Text hervorgeht.

Es ist wohl davon auszugehen, dass D&B nach dem gleichen Schema wie Crunchbase durchsucht wird (nach der Webseite wenn verfügbar, sonst über eine in Token übersetzte Form des Namens). Allerdings wird das nicht angegeben. Interessant wäre außerdem auch, ob eine Suche in D&B durch die Angabe einer über einen Kartendienst automatisch ermittelten Adresse die Ergebnisse bessere Ergebnisse liefern würde.

Ergebnisse. Mit der Kombination aller fünf Datenquellen (PeeringDB, IPInfo, Zvelo, Crunchbase und D&B) kann für 99% aller global registrierten ASe eine Organisation in mindestens einer der Datenquellen gefunden werden. Allerdings sind sich bei 14% der ASe im Gold-Standard mindestens

zwei Datenquellen uneinig, um welche Organisation es sich handelt.

4.5 Konsens und Dissens in der Kategorisierung

Wenn mindestens zwei Datenquellen sich in der Kategorisierung einig sind, ist die Genauigkeit der Kategorisierung sehr hoch, oft sogar bei 100% (siehe hierzu Tabelle 11 in [15]). Allerdings sind sich die Datenquellen bei 13% der ASe im Gold-Standard und 40% der ASe im uniformen Gold-Standard vollständig uneinig. Dieser Dissens entsteht vor allem aus drei Gründen:

- Alle Datenquellen ermitteln an sich zutreffende, aber unterschiedliche Kategorien. (Betrifft etwa 6% der ASe im Gold-Standard.)
- Eine der Datenquellen liegt schlicht falsch. (Betrifft etwa 7% der ASe im Gold-Standard.)
- Die gefundenen Einträge in den Datenquellen beschreiben unterschiedliche Organisationen. Dieses Problem wurde bereits ausführlich in Abschnitt 4.4 behandelt. (Betrifft etwa 14% der ASe im Gold-Standard.)

4.6 Zusammenfassung

Bezüglich der meisten NAICSlite Layer 1 Kategorien erreichen die untersuchten externen Datenquellen bei Übereinstimmung mindestens zweier Datenquellen eine sehr hohe, meist nahe zu perfekte Genauigkeit. Leider sind sich die Datenquellen oft vollständig uneinig (siehe Abschnitt 4.5).

Außerdem scheitern existierende Datenquellen oft an der granularen Kategorisierung von Technologie-Unternehmen (etwa bei der Unterscheidung von ISPs und Web Hosting Anbietern) – obwohl Technologie-Unternehmen ca. 64% der Gesamtheit aller ASe ausmachen.

5 MACHINE LEARNING IN ASDB

Zur besseren Unterscheidung von ISPs und Web Hosting Anbietern wird für ASdb Machine Learning (ML) in Form zweier binärer Klassifikatoren eingesetzt. Machine Learning wird nur auf diese beiden Kategorien angewendet, da es sich hierbei erstens um die zwei häufigsten Kategorien handelt (und somit ausreichend Trainings-Daten vorhanden sind) und zweitens Zvelo bereits selbst ein ML basiertes System zur weiteren Klassifizierung zur Verfügung stellt, welches vermutlich schwierig zu übertreffen wäre.

5.1 ML Pipeline

Der grundlegende Ablauf von Einsatz und Training der beiden Klassifikatoren ist in Abbildung 4 dargestellt.

Zunächst werden zum Erzeugen der Trainingsdaten alle verfügbaren Webseiten der ISPs und Web Hosting Anbieter im Gold-Standard heruntergeladen. Diese werden dann wenn

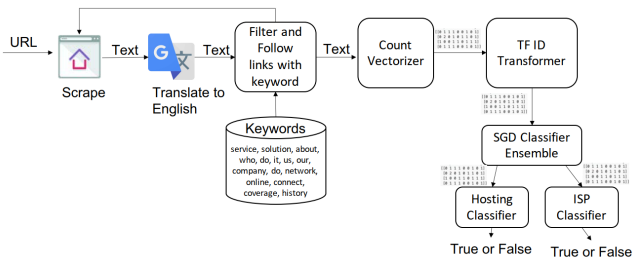

Abbildung 4: Schaubild zum Training und Einsatz von Machine Learning in ASdb

nötig durch den Google Übersetzer auf englisch übersetzt. Da viele Webseiten Autonomer Systeme ihre Dienstleistungen nicht direkt auf ihrer Landing-Page beschreiben, werden zusätzlich bis zu fünf verlinkte Seiten herunterladen, wenn der Link-Titel eines der abgebildeten Schlüsselwörter enthält. (Wieso genau diese Schlüsselwörter gewählt wurden, bleibt leider unklar.)

Danach kommt das sogenannte „TF-IDF“ Verfahren zum Einsatz, um die Relevanz bestimmter Schlüsselwörter zu bestimmen. Das Verfahren hat den Vorteil, dass zwar häufig vorkommende, aber vermutlich uninteressante Wörter wie Artikel oder Pronomen ignoriert werden. Denkbare so ermittelte Schlüsselwörter zur Erkennung eines ISPs währen beispielsweise „Infrastructure“, „Network“ oder „connect“.

Die Häufigkeit dieser Schlüsselwörter zu einer Webseite inklusive der Information, ob es sich dabei um einen ISP beziehungsweise Web Hosting Anbieter handelt, werden dann als Trainings-Daten für zwei „Stochastic Gradient Descent“ (SGD) Klassifikatoren verwendet. Diese können nach einer kurzen Trainingsphase dann zu einem neuen Datensatz mit den Häufigkeiten der Schlüsselwörter vorhersagen, ob es sich um einen ISP handelt oder nicht beziehungsweise ob es sich um einen Web Hosting Anbieter handelt oder nicht.

Der Einsatz eines SGD Klassifikators in der Praxis ist durch Verwendung entsprechender Bibliotheken (etwa `sklearn` für Python) recht einfach. Ein für diese Zusammenfassung verfasstes minimales Beispiel kann bei Bedarf im Anhang in Listing 1 eingesehen werden.

Leider werden die Parameter oder die genaue verwendete Implementierung des SGD Klassifikators nicht genannt. Diese können das Trainings-Verhalten und somit die Güte des Klassifikators stark beeinflussen und wären daher zur Reproduktion des Artikels interessant.

5.2 Bewertung

Laut den Autoren erreichen die beiden Klassifikatoren für ISPs und Web Hosting Anbieter jeweils eine Korrektklassifikationsrate von 94% und 90% und irren sich dabei nur in jeweils 1% und 3% falsch positiven Ergebnissen. Falsch

negative Ergebnisse sind etwas wahrscheinlicher mit jeweils 5% und 7%, aber damit immer noch vertretbar gering.

Die Autoren geben außerdem einen „Area Under Curve“ (AUC) Score an, der bei jeweils 94% und 80% liegt. Der AUC Score ist eigentlich eine Metrik, mit der sich die Güte eines Klassifikators über verschiedene Werte eines Parameters quantifizieren lässt. Bei anderen Machine Learning verfahren wie der logistischen Regression ist dieser Parameter meist ein Grenzwert, der die Unterscheidung zwischen dem positiven und negativen Fall bestimmt. Üblicherweise vergleicht man dann den Einsatz verschiedener Machine Learning Verfahren über den AUC Score. Nach längerer Recherche ist mir nicht klar, welcher Parameter zur Berechnung eines AUC Scores eines SGD Klassifikators eingesetzt wird. Die Werte werden auch im Paper nicht weiter behandelt oder bewertet.

Unabhängig des AUC Scores zeigt sich die Qualität der trainierten Klassifikatoren durch die hohe Korrektklassifikationsrate und dem niedrigen Anteil falsch positiver und falsch negativer Antworten.

6 ASDB

In diesem Abschnitt wird das eigentliche „Ergebnis“ der Arbeit vorgestellt: Das AS Klassifizierungssystem ASdb. Es wird die Architektur geschildert und die Leistung auf Basis verschiedener Metriken bewertet. Die aktuelle Version der Daten (sowie NAICSlite) ist als eine .csv-Datei auf der Webseite von ASdb verfügbar[4]. (Ein Archiv zum Download älterer Versionen fehlt.)

6.1 Architektur

Die Architektur von ASdb ist in Abbildung 5 illustriert. Zunächst werden zur Klassifizierung eines AS die zugehörigen WHOIS Daten (bereitgestellt von der zuständigen RIR) betrachtet. Wenn das AS noch nicht klassifiziert wurde, dann werden zunächst PeeringDB und IPInfo nach der gegebenen AS Nummer durchsucht. Wenn dabei ein Ergebnis mit „hoher Konfidenz“ gefunden wird, wird die Klassifizierung ohne weitere Umwege nach einer automatischen Übersetzung zu NAICSlite übernommen. In der konkreten Implementierung von ASdb wird lediglich eine Einordnung als ISP durch PeeringDB als ein Ergebnis mit „hoher Konfidenz“ betrachtet.

In jedem anderen Fall geht ASdb in die nächste Phase über. Dafür wird für das gegebene AS zunächst eine Webseite ermittelt, die dann als Suchparameter für weitere Datenquellen verwendet wird. Dafür werden schlicht alle in den Ergebnissen von PeeringDB und IPInfo, sowie den WHOIS Daten vorkommenden Webseiten gesammelt, einige häufige Webseiten (etwa von E-Mail Anbietern wie GMail) gefiltert und schließlich die in Abschnitt 4.4 beschriebene Strategie „most similar domain“ angewandt.

[4]https://asdb.stanford.edu/

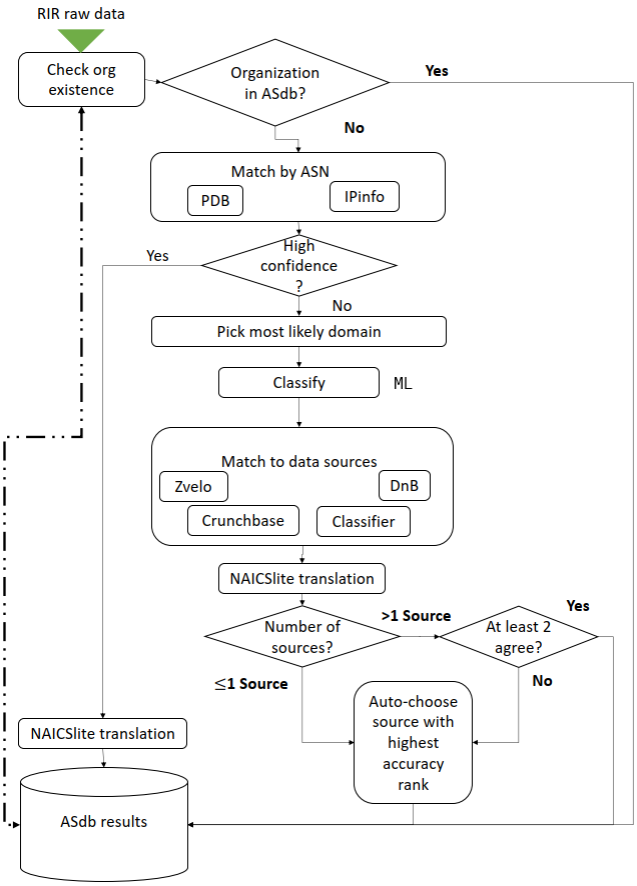

Abbildung 5: Übersicht zur Architektur von ASdb

Die so gefundene Webseite wird dann durch die in Abschnitt 5 beschriebene Machine Learning Pipeline bearbeitet, um so festzustellen, ob es sich um einen ISP oder einen Web Hosting Anbieter handelt. Wenn das der Fall ist, wird das Ergebnis übernommen.

Wenn nicht, dann werden in einem dritten Schritt auf Grundlage der Webseite und den WHOIS Daten die restlichen Datenquellen (D&B, Zvelo und Crunchbase) nach Name, Webseite und physischer Adresse durchsucht. (Die Ermittlung dieser Parameter wird im Artikel in Appendix A beschrieben.) Wenn zwei der Datenquellen sich bei der Kategorie einig sind, wird diese Kategorie übernommen. Wenn nicht, dann wird die Kategorie der Datenquelle mit der höchsten Trefferquote nach Tabelle 3 A gewählt. (Zu kritisieren ist, dass im Artikel hier (Abschnitt 5.1) von der Korrektklassifikationsrate gesprochen wird, obwohl der referenzierte Abschnitt 3.3 diese mit keinem Wort erwähnt. Stattdessen geht es in dem referenzierten Abschnitt um die Abdeckung, Genauigkeit und Trefferquote. Außerdem passen die angegebenen Werte zu Zvelo und Crunchbase nicht exakt zu den

in Tabelle 3 in [15] angegebenen, was einem Schreibfehler geschuldet sein könnte.)

6.2 Bewertung

Zur Bewertung von ASdb wird zusätzlich zu dem Gold-Standard und dem Uniformen Gold-Standard ein neuer Datensatz nach der gleichen Methode wie der Gold-Standard erstellt.

ASdb erreicht in diesem neuen Datensatz eine Abdeckung von 96% und eine Korrektklassifikationsrate der Layer 1 NAICSlite Kategorien von 93%.

Im direkten Vergleich muss sich ASdb nur mit PeeringDB und IPInfo messen, denn diese beiden Datenquellen sind außer ASdb die einzigen, die nach AS Nummer durchsucht werden können. Im Vergleich zu diesen beiden Systemen verfügt ASdb über etwa 90 zusätzliche Kategorien, ist also deutlich granularer. Zum weiteren Vergleich wird das F_1-Maß herangezogen (ein gewichtetes harmonisches Mittel aus Trefferquote und Genauigkeit). Das für ASdb berechnete F_1-Maß ist in der Kategorie der Web Hosting Anbieter um den Faktor 2,5-6 größer, für ISPs um den Faktor 1,3-2,5, 1,1-5 für die Kategorie „Education and Research“ und 1,3-12 für alle restlichen Kategorien. Damit ist ASdb nicht nur granularer sondern in allen Kategorien auch etwas bis deutlich genauer als existierende Lösungen. Trotzdem erreicht ASdb nur ein F_1-Maß von 65% für Web Hosting Anbieter. Detailliertere Messwerte können in [15] unter Abschnitt 5.2 nachgelesen werden.

ASdb ist also ein wesentlicher Fortschritt zur Kategorisierung Autonomer Systeme gegenüber PeeringDB und IPInfo, allerdings werden keinesfalls perfekte Werte erzielt. Hier könnte weitere Forschung ansetzen.

7 CROWDWORK

Um die restlichen Fehler in ASdb zu reduzieren, wurde der Einsatz von Crowdwork über „Amazon Mechanical Turk“ (AMT) in Betracht gezogen, aber schließlich verworfen. Speziell wurde an zwei Stellen angesetzt:

(1) Die in Abschnitt 5 beschriebenen ML Klassifikatoren zur Erkennung von ISPs und Web Hosting Anbietern produzieren einige falsch negative Ergebnisse. Durch den Einsatz von AMT Arbeitern wäre nach einem durchgeführten Experiment möglich, nahezu alle dieser Fehler zu korrigieren, indem die Arbeiter angewiesen werden, alle diejenigen ASe zu klassifizieren, die in mindestens einem der externen Datenquellen als ISP beziehungsweise Web Hosting Anbieter gekennzeichnet wurden, nicht aber von der ML Pipeline. Dabei würden allerdings vermutlich in etwa 31,000$ an Kosten anfallen, was als deutlich zu teuer eingestuft wurde.

(2) In Fällen, in denen kein Konsens zwischen den externen Datenquellen über die Kategorisierung besteht, könnte Crowdwork zur Konsensfindung eingesetzt werden. Es konnte allerdings experimentell gezeigt werden, dass dadurch kein relevanter Vorteil über der einfachen tatsächlich eingesetzten Heuristik „wähle die Datenquelle mit der höchsten Trefferquote“ entstehen würde.

Die Autoren sind in ihrem Artikel noch deutlich weiter ins Detail gegangen und haben unter anderem auch über faire und effiziente Entlohnung, alternative Crowdwork Plattformen zu AMT und ethische Fragestellungen geschrieben. Da diese Überlegungen schlussendlich aber für den Einsatz in ASdb verworfen wurden, wird hier nicht näher darauf eingegangen.

8 BEWERTUNG

In diesem Abschnitt sollen kurz einige formale Aspekte und die meistzitierten Quellen betrachtet werden. Außerdem wird ein Überblick über die Arbeit der Autoren gegeben. Inhaltliche Kritik wurde bereits an entsprechenden Stellen angebracht. Insgesamt sei zu der inhaltlichen Kritik nur noch einmal erwähnt, dass es sich meiner Meinung nach im Allgemeinen um ein hochwertiges Paper handelt, das trotz der kritisierten Details einen wertvollen Beitrag durch die gewonnenen Erkenntnisse leistet. Da das Ergebnis (die ASdb Datenbank) einfach einsetzbar und leistungsfähig ist (siehe Abschnitt 6.2), würde ich von einer hohen Relevanz des Papers für zukünftige Artikel ausgehen.

8.1 Formale Aspekte

Die Struktur des Artikels ist durchdacht und konsequent umgesetzt. Dass der „große Überblick“ erst im letzten Abschnitt gegeben wird, ist aufgrund der Komplexität der einzelnen Teile von ASdb gerechtfertigt. (Daher wurde es auch hier so übernommen.) Die Sprache ist wie für eine wissenschaftliche Arbeit zu wünschen klar und präzise, ohne viele Füllwörter oder Ausschmückungen.

Lediglich die Lesbarkeit in Abschnitt 3 (hier Abschnitt 4) ist zu kritisieren. Dort müssen Daten oft mühsam aus dem Fließtext gelesen werden und durch die an den entsprechenden Stellen erwähnten inhaltlichen Schwachpunkte ist der Fortschritt bei genauerem Lesen recht träge.

8.2 Quellen

In diesem Abschnitt sollen die meistzitierten Quellen, auf die sich der Artikel bezieht, kurz in wenigen Sätzen beschrieben werden. „Meistzitiert“ ist dabei ein dehnbarer Begriff; Es gibt viele verschiedene Dienste (Google Scholar[5], Scite[6], Semantic

[5] https://scholar.google.com
[6] https://scite.ai/

Scholar[7], Research Rabbit[8], etc.), welche unterschiedliche Werte zur Anzahl der Zitationen liefern. Hier werden nur die Werte von Google Scholar betrachtet.

„Text categorization with Support Vector Machines: Learning with many relevant features". [12]

- Anzahl Zitationen laut Google Scholar: 12053
- Inhalt: Das Paper behandelt die Anwendung von „Support Vector Maschinen" (SVMs), eine Methode des Machine Learning, zur Text Klassifikation. Die Eignung solcher SVMs wird theoretisch und empirisch gezeigt.
- In ASdb wird die Quelle auf Seite 709 verwendet, um die Entscheidung zur Verwendung eines SGD Klassifikators zu rechtfertigen. Laut dem Zitat werden SGD Klassifikatoren oft zur Klassifikation von Text aufgrund ihrer Skalierbarkeit eingesetzt. Ohne Hintergrund im Machine Learning war ich nicht in der Lage, die Richtigkeit dieses Zitats zu prüfen; Es *scheint* ein Zusammenhang zwischen SVMs und SGD Klassifikatoren zu existieren, so dass bei der Wahl gewisser Parameter die Funktionsweise ähnlich oder gleich ist. Die Verwendung der Quelle verliert in diesem Kontext aufgrund des Veröffentlichungsjahrs 1998 etwas an Aussagekraft.

„Using tf-idf to determine word relevance in document queries". [14]

- Anzahl Zitationen laut Google Scholar: 2274
- Inhalt: Es wird die Eignung des TF-IDF Verfahrens zur schlüsselwort-basierten Suche über eine Menge an Dokumenten geprüft und schließlich gezeigt. Es wird außerdem auf die größte Limitation von TF-IDF hingewiesen: Es werden keine Synonyme erkannt.
- In ASdb wird das Paper schlicht nach dem ersten Vorkommen des Wortes TF-IDF auf Seite 709 zitiert. Die Stelle des Zitats könnte den Leser zu der Annahme verleiten, dass es sich bei der Quelle um den Ursprung des TF-IDF Verfahrens handelt. Das ist nicht der Fall. Allerdings zeigt die Quelle die Effizienz des Verfahrens.

„Reputation as a sufficient condition for data quality on Amazon Mechanical Turk". [13]

- Anzahl Zitationen laut Google Scholar: 1443
- Inhalt: Das Paper zeigt, dass die Verwendung von „Attention Check Questions" (ACQs) zur Sicherung der Qualität der auf Amazon Mechanical Turk durch Crowdwork gewonnenen Daten nicht notwendig ist, wenn die Aufgaben von Arbeitern mit hoher Reputation bearbeitet werden.
- In ASdb wird der Artikel auf Seite 714 zitiert, um die Entscheidung zu rechtfertigen, ausschließlich „Master Worker" für die unternommenen Crowdwork Versuche zu beschäftigen.

8.3 Autoren

In diesem Abschnitt wird ein kurzer Überblick über die Autoren des Artikels gegeben.

Maya Ziv.

- h-index und i10-index: Nicht Verfügbar (kein Google Scholar Eintrag)
- Webseite: https://mayaziv.com/
- Maya Ziv ist eine Master Absolventin der Stanford University. Abgesehen von ihrem Beitrag zu ASdb hat sie laut ihrer Webseite an einem Paper mit zwei weiteren Studenten im Kontext eines belegten Kurses gearbeitet.

Liz Izhikevich.

- h-index: 4, i10-index: 2
- Webseite: https://lizizhikevich.github.io/
- Liz Izhikevich ist eine Doktorandin an der Stanford University.
- Meistzitierte Veröffentlichungen[9]:
  - „Sprocket: A serverless video processing framework"
  - „On the origin of scanning: The impact of location on Internet-wide scans"
  - „LZR: Identifying Unexpected Internet Services"

Kimberly Ruth.

- h-index: 4, i10-index: 4
- Webseite: https://kcruth.com/
- Kimberly Ruth ist eine Doktorandin an der Stanford University.
- Meistzitierte Veröffentlichungen[10]:
  - „Towards security and privacy for multi-user augmented reality: Foundations with end users"
  - „Securing augmented reality output"
  - „Secure Multi-User Content Sharing for Augmented Reality Applications"

Katherine Izhikevich.

- h-index: 2, i10-index: 0
- Webseite: https://kizhikevich.github.io/
- Kimberly Ruth ist eine Studentin an der Stanford University.
- Meistzitierte Veröffentlichungen[11]:
  - „ASdb: a system for classifying owners of autonomous systems"

[7] https://www.semanticscholar.org/
[8] https://researchrabbitapp.com/
[9] https://scholar.google.com/citations?user=jO0eK0AAAAAJ&hl=en
[10] https://scholar.google.com/citations?user=VImQVZ4AAAAJ&hl=en
[11] https://scholar.google.com/citations?hl=en&user=SpNMl4kAAAAJ

– „SARS-CoV-2 emergence very likely resulted from at least two zoonotic events“

Zakir Durumeric.

- h-index: 25, i10-index: i10-index: 29
- Webseite: https://zakird.com/
- Zakir Durumeric ist ein Assistenzprofessor an der Stanford University.
- Meistzitierte Veröffentlichungen[12]:
  – „Understanding the Mirai Botnet“
  – „The Matter of Heartbleed“
  – „ZMap: Fast Internet-Wide Scanning and its Security Applications“

9 SCHLUSSBETRACHTUNG

Mit ASdb haben die Autoren durch den vorliegenden Artikel ein neues Werkzeug zur Klassifizierung Autonomer Systeme geschaffen, dass existierende Werkzeuge in Granularität, Trefferquote, Genauigkeit und Korrektklassifikationsrate teilweise leicht und oft stark übertrifft. Dadurch können in Zukunft weitere Fragen besser erforscht werden. Als Beispiel für die Möglichkeiten, die ASdb bietet, haben die Autoren ASdb mit einem existierenden Telnet Scan aus [11] verbunden und so feststellen können, dass ASe in Kategorien von Organisationen, die zum Teil für kritische Infrastruktur verantwortlich sind (wie etwa Elektrizitätsversorgungsunternehmen oder Regierungen), mit einer höheren Wahrscheinlichkeit ungesicherte Telnet Verbindungen erlauben.

Es bleibt zu hoffen, dass ASdb wie geplant weiterhin betrieben und aktualisiert wird oder die Ergebnisse von anderen Autoren reproduziert werden können.

LITERATUR

[1] 2022. 4.2.1 Description of the AUT-NUM Object — RIPE Network Coordination Centre. https://www.ripe.net/manage-ips-and-asns/db/support/documentation/ripe-database-documentation/rpsl-object-types/4-2-descriptions-of-primary-objects/4-2-1-description-of-the-aut-num-object (Zugriff am 28.04.2022).

[2] 2022. Aut-Num – APNIC. https://www.apnic.net/manage-ip/using-whois/guide/aut-num/ (Zugriff am 28.04.2022).

[3] 2022. Comprehensive IP Address Data, IP Geolocation API and Database - IPinfo.Io. https://ipinfo.io/ (Zugriff am 01.05.2022).

[4] 2022. PeeringDB. https://www.peeringdb.com/ (Zugriff am 01.05.2022).

[5] 2022. Searching Whois Using a CLI - American Registry for Internet Numbers. https://www.arin.net/resources/registry/whois/rws/cli/#interpreting-whois-results (Zugriff am 28.04.2022).

[6] Annika Baumann and Benjamin Fabian. 2014. Who Runs the Internet? - Classifying Autonomous Systems into Industries:. In *Proceedings of the 10th International Conference on Web Information Systems and Technologies*. SCITEPRESS - Science and and Technology Publications, Barcelona, Spain, 361–368. https://doi.org/10.5220/0004936803610368

[7] Center for Applied Internet Data Analysis (CAIDA). 2015. AS Classification. https://www.caida.org/catalog/datasets/as-classification/ (Zugriff am 01.05.2022).

[8] Leslie Daigle. 2004. *WHOIS Protocol Specification*. Request for Comments RFC 3912. Internet Engineering Task Force. https://doi.org/10.17487/RFC3912

[9] A. Dhamdhere and C. Dovrolis. 2011. Twelve Years in the Evolution of the Internet Ecosystem. *IEEE/ACM Transactions on Networking* 19, 5 (Oct. 2011), 1420–1433. https://doi.org/10.1109/TNET.2011.2119327

[10] Marco d'Itri. 2022. Rfc1036/Whois Manual. https://github.com/rfc1036/whois/blob/1b89adf90b4ee9ff374f11128bf5cbb3b93285f5/whois.1 (Zugriff am 01.05.2022).

[11] Liz Izhikevich, Renata Teixeira, and Zakir Durumeric. 2021. LZR: Identifying Unexpected Internet Services. In *30th USENIX Security Symposium (USENIX Security 21)*. USENIX Association, Virtuelle Veranstaltung, 3111–3128.

[12] Thorsten Joachims. 1998. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In *Machine Learning: ECML-98*, Claire Nédellec and Céline Rouveirol (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 137–142.

[13] Eyal Peer, Joachim Vosgerau, and Alessandro Acquisti. 2014. Reputation as a Sufficient Condition for Data Quality on Amazon Mechanical Turk. *Behavior Research Methods* 46, 4 (Dec. 2014), 1023–1031. https://doi.org/10.3758/s13428-013-0434-y

[14] Juan Ramos. 2003. Using TF-IDF to Determine Word Relevance in Document Queries. In *Proceedings of the First Instructional Conference on Machine Learning (1, Vol. 242)*. Citeseer, 29–48.

[15] Maya Ziv, Liz Izhikevich, Kimberly Ruth, Katherine Izhikevich, and Zakir Durumeric. 2021. ASdb: A System for Classifying Owners of Autonomous Systems. In *Proceedings of the 21st ACM Internet Measurement Conference*. ACM, Virtuelle Veranstaltung, 703–719. https://doi.org/10.1145/3487552.3487853

ANHANG

[12] https://scholar.google.com/citations?user=TxPSRHIAAAAJ&hl=en

```python
from sklearn.linear_model import SGDClassifier
import numpy as np

# (Natürlich sind das hier viel zu wenige Daten, um einen echten, sinnvoll
# einsetzbaren Classifier zu trainieren.)
training_data = [
    {
        'keywords_count': {
            'cloud'  : 5,  'performance': 4,  'scalable': 5,
            'connect': 0,  'coverage'   : 0,  'service' : 1,
        },
        'is_isp': True
    },
    {
        'keywords_count': {
            'cloud'  : 0,  'performance': 1,  'scalable': 0,
            'connect': 5,  'coverage'   : 5,  'service' : 4,
        },
        'is_isp': False
    }
]

# Die Daten müssen zur Verwendung mit sklearn noch etwas transformiert werden.
#
# X ist hierbei eine Liste von Listen mit der Anzahl der Vorkommen der einzelnen
# Schlüsselwörter:
# [
#     [5, 4, 5, 0, 0, 1],
#     [0, 1, 0, 5, 5, 4],
#     ...
# ]
#
# y ist eine Liste der is_isp Werte: [True, False, ...]

X = list(map(lambda el: list(el['keywords_count'].values()), training_data))
y = list(map(lambda el: el['is_isp'], training_data))


sgd_classifier = SGDClassifier() # Classifier Objekt initialisieren
sgd_classifier.fit(X, y)         # --- Classifier trainieren ---

to_predict_1 = {
    'cloud'  : 5,  'performance': 3,  'scalable': 4,
    'connect': 1,  'coverage'   : 1,  'service' : 0,
}
to_predict_2 = {
    'cloud'  : 5,  'performance': 3,  'scalable': 4,
    'connect': 20,  'coverage'  : 20, 'service' : 20,
}

to_predict_1 = np.array(list(to_predict_1.values())).reshape(1, -1)
to_predict_2 = np.array(list(to_predict_2.values())).reshape(1, -1)

print(sgd_classifier.predict(to_predict_1)) # -> True
print(sgd_classifier.predict(to_predict_2)) # -> False
```

Listing 1: Minimales Beispiel eines Stochastic Gradient Descent Classifiers in Python mit der Bibliothek `sklearn` – Dient der Veranschaulichung der praktischen Anwendung eines SGD Klassifikators