WebUIを対象とした広域スキャンにおけるIoT機器プロファイリングの自動化

林 嘉乃[1,a)]　合屋 琴江[3]　九鬼 琉[3]　藤井 翔太[2]　佐々木 貴之[4]　吉岡 克成[4,5]

概要：様々なIoT機器が，管理者の意図に反してインターネット上に公開され，外部からの攻撃に晒されている．これらの機器に対するリスク評価には，種類や属性を正確に把握することが不可欠であり，特に社会インフラで使用される重要施設設置機器では，攻撃による社会的影響が深刻なため優先的な対処が求められる．先行研究では，LLMによりWebインターフェースから機種名等を推定する手法が提案されているが，実環境での大規模実態調査への適用例はない．本研究では，HTMLとスクリーンショットを自動収集するWebUIスキャナと，LLMによるプロファイリング手法を統合し，機種名・メーカー名・シグネチャに加えて設置施設情報も自動推定するシステムを構築した．複数ASのIPアドレス空間（約20万IP）への適用により618機種を推定し，うち40機種が重要施設設置機器であった．生成したシグネチャにより新たに約1,300万件の機器を発見，重要施設設置機器486件の特定に成功し，提案手法の実用性を実証した．

キーワード：IoT機器，インターネット広域スキャン，大規模言語モデル

Automation of IoT Device Profiling in Wide-Area Scanning of WebUIs

Kano HAYASHI[1,a)]　Kotoe GOYA[3]　Ryu KUKI[3]　Shota Fujii[2]　Takayuki SASAKI[4]
Katsunari YOSHIOKA[4,5]

Abstract: Various IoT devices are unintentionally exposed to the Internet, leaving them vulnerable to external attacks. Accurate risk assessment of such devices requires precise identification of their types and attributes. In particular, devices installed in critical infrastructures demand prioritized protection, as attacks on them can have severe social consequences. Previous studies have proposed methods that leverage large language models (LLMs) to infer device information such as model names from web interfaces, but no prior work has demonstrated their applicability to large-scale, real-world measurements. In this study, we developed an integrated system that combines a web UI scanner, which automatically collects HTML data and screenshots, with an LLM-based profiling method. The system automatically infers device models, manufacturers, and signatures, as well as facility information. Applying this system to the IP address spaces of multiple autonomous systems (approximately 200,000 IP addresses), we identified 618 device models, of which 40 were deployed in critical facilities. Using the generated signatures, we further discovered approximately 13 million additional devices and successfully identified 486 IoT devices installed in critical facilities. These results demonstrate the practicality and effectiveness of the proposed approach.

Keywords: IoT devices, Internet-wide scanning, Large Language Model

[1] 横浜国立大学理工学部
College of Engineering Science, Yokohama National University
[2] 横浜国立大学
Yokohama National University
[3] 横浜国立大学大学院環境情報学府
Graduate School of Environment and Information Sciences, Yokohama National University
[4] 横浜国立大学先端科学高等研究院
Institute of Advanced Sciences, Yokohama National Univer-

1. はじめに

近年，センサーデータの収集や遠隔制御を目的とした Internet of Things（IoT）機器の導入が急速に進んでいる．特に，発電所や上下水道施設，工場といった社会インフラに導入されている IoT 機器は，施設の運用効率化や安全管理を担う重要な技術となっている．本研究では，このような重要施設に設置された IoT 機器に対し，セキュリティ上の課題に注目する．

多くの IoT 機器がインターネットに直接接続される中で，アクセス制御の不備，初期パスワードの放置，ファームウェアの未更新など，重大な脆弱性を抱えているケースが多数報告されている [1], [2]．こうした機器への攻撃は，施設の機能停止や情報漏洩といった深刻な被害につながる恐れがあるため，インターネット上の IoT 機器の実態把握とリスク評価を広範囲に行う仕組みが求められている．

既存の IoT 機器探索手法として，垂直展開スキャンと水平展開スキャンを組み合わせた 2 段階アプローチが提案されている [3]．一方，大規模言語モデル（LLM）の発展により，IoT 機器の Web インターフェースの UI から機器情報を自動推定する手法も提案されている [4], [5]．これらの研究では，HTML ソースやスクリーンショットを入力として，機種名やメーカー名の推定，シグネチャの自動作成，設置施設情報の自動抽出を実現している．

しかし，従来手法には以下の課題が存在する．垂直展開スキャン手法では，機種名・メーカー名等の推定や，シグネチャ（特定機種の検出用検索クエリ）の作成，設置施設情報を推定する機器のプロファイリングが手動で行われるため，1 件あたり数分から数十分を要し，大規模探索のボトルネックとなっている．また，LLM を用いた手法では個別の推定精度は向上したものの，探索プロセス全体の完全自動化は実現されていない．

そこで，本研究では，垂直・水平展開スキャンの 2 段階アプローチと LLM を用いた自動的な機器情報推定技術を統合し，IoT 機器の Web インターフェースホストの探索から属性推定，シグネチャ生成，水平展開スキャンに至る全プロセスを自動化するシステムを構築した．システムでは，収集した WebUI を HTML 構造の類似性に基づいてクラスタリングし，同一機種の IoT 機器をグループ化する．

4 つの AS（Autonomous System）を対象とした実証により，1,468 クラスタのうち，40 クラスタが重要施設に設置されている可能性が高いことを確認した．さらに，当該 40 クラスタに含まれる 1,167 ホストの機器を個別にプロファイリングした結果，486 ホストが重要施設に設置されている機器であることが判明した．また，1,468 クラスタから 933 個の有効なシグネチャを生成し，新たに約 1,300 万ホストの IoT 機器 Web インターフェースを発見した．これにより，従来の手動プロファイリングの 1/10 から 1/45 にあたる約 20 秒に短縮し，大規模スキャンにおけるボトルネックを解消した．

本研究の主な貢献は以下の通りである: (1) WebUI スキャンからシグネチャ生成まで，IoT 機器探索の全プロセスを自動化するシステムを構築，(2) 4 つの AS で 618 機種を推定し 933 個のシグネチャで約 1,300 万ホストの機器 Web インターフェースを発見，(3) 486 ホストの重要施設に設置された機器を発見し自動探索を実現，(4) プロファイリング作業を約 20 秒に短縮しボトルネックを解消．

2. 課題

従来の手法では，垂直展開スキャンと水平展開スキャンを組み合わせることで，インターネット上の IoT 機器を包括的に探索するアプローチが確立されつつある [3]．また，LLM を活用した属性推定に関する研究も進展している [4], [5]．しかし，これらの従来手法には以下の課題が存在する．まず，垂直展開スキャンの結果を踏まえてシグネチャ（検索クエリ）を作成するプロセスが人手に依存しており，大規模探索におけるボトルネックとなっている．次に，LLM を用いた既存の属性推定手法では，WebUI の HTML コードやスクリーンショット画像の収集・整形が手動で必要であり，対象機器の数が増えるほど作業負担が増加する．以上より，現在の LLM を用いた手法は，既知の機器や一部の特定環境での利用にとどまり，インターネット全体に分散する IoT 機器に対する自動化されたスキャンやリスク評価には対応できていない．本研究は，これらの課題を解決するため，探索からプロファイリング，水平展開スキャンに至る一連のプロセスを自動化し，大規模化にも対応可能な高精度の IoT 機器特定基盤の構築を目指す．

3. シグネチャとプロファイリング

3.1 シグネチャの定義

本論文におけるシグネチャの定義について説明する．「シグネチャ」とは，IoT サーチエンジンである Censys[6] において特定機種の検索に用いる検索クエリを指す．シグネチャは，HTML ソースに含まれる機種固有のキーワードに基づいて構成される．本用語は文献 [5] に基づくものであり，本研究でもこれを踏襲する．

3.2 プロファイリングの定義と出力

本研究において「プロファイリング」とは，LLM を用いて観測情報（HTML ソース，スクリーンショット等）から，対象 IoT 機器に関する多様な属性情報を推定・抽出するプロセスを指す．プロファイリングは処理単位に応じて

sity
[5] 横浜国立大学大学院環境情報研究院
Faculty of Environment and Information Sciences, Yokohama National University
[a)] hayashi-kano-jd@ynu.jp

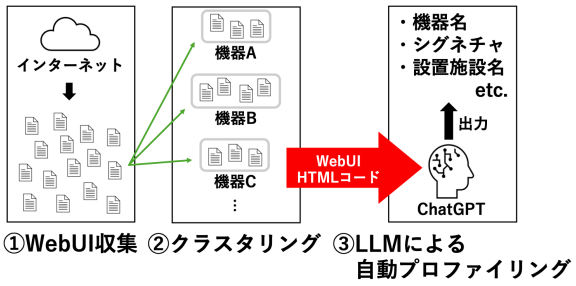

図 1 提案手法の概要図

以下の 2 段階で実行される．

機種単位（機種共通の情報）：機種名，メーカー名，機種カテゴリ，機種に関連するキーワード，検索クエリ（キーワードを組み合わせた Censys 用のクエリ），バージョン情報を抽出するための正規表現

個別 WebUI 単位（設置施設情報）：設置施設名，設置施設の重要性，設置施設の分類

4. 提案手法

4.1 提案手法概要

本研究では，WebUI スキャナによって収集した IoT 機器の情報（HTML コードおよびスクリーンショット）を LLM に入力し，機器のクラスタリング・プロファイリングを自動的に実行する統合手法を提案する．提案手法の全体構成を図 1 に示す．

提案手法は，WebUI 収集部，クラスタリング部，LLM による自動プロファイリング部の 3 つの構成要素から成る．WebUI 収集部では対象範囲から IoT 機器の WebUI を収集し，クラスタリング部では同一機種の WebUI をグループ化する．LLM プロファイリング部では各機器の機種情報と設置施設情報を自動推定する．この統合的なアプローチにより，手動処理時間の短縮を実現し，AS 全体やインターネット全域への適用が現実的となった．

以下，各構成要素の詳細について説明する．

4.2 WebUI 収集部

Censys の API を利用し，指定されたクエリにより，ホストを検索する．特定された IP アドレスに対して Nmap によるポートスキャンと Selenium WebDriver による WebUI のスクリーンショット・HTML ソースコード取得を順次実行し，収集データを JSON ドキュメントとして集約する．

4.3 クラスタリング部

同一の IoT 機種は同じ WebUI を持つため，クラスタリングにより同機種の IoT 機器をグループ化できる．一方で，通常の Web サイトは見た目が多様であるため，クラスタを形成しない．この特徴を利用し，IoT 機器の WebUI のみを抽出する．

表 1 機器カテゴリの分類

カテゴリ	主な機器例
カメラ	IP カメラ，ネットワークカメラ
映像保存	ビデオレコーダ，映像ストレージ機器
ネットワーク	ルータ，モデム，アクセスポイント
印刷機器	プリンタ，複合機，スキャナ
ストレージ	NAS，外部ストレージ機器
産業制御	産業制御システム，監視制御機器
クラウドサービス	クラウドベース API，SaaS
その他	上記以外の機器
不明	分類不能

文献 [7] の手法に基づき，HTML の構造的特徴を用いたクラスタリングを実行する．具体的には，HTML から script タグとテキストコンテンツを削除してタグ構造のみを抽出し，ssdeep アルゴリズムによるハッシュ値の算出，類似度計算，階層的クラスタリングを行う．

4.4 LLM による自動プロファイリング部

OpenAI GPT-4o（temperature=0.0）を用いて，3.2 節で定義した 2 段階のプロファイリングを実行する．出力形式の一貫性確保には LangChain の with_structured_output メソッドを使用する．プロンプトは文献 [4], [5] を参考に設計し，詳細は付録 A.1 に示す．

4.4.1 機種情報の推定

各クラスタの代表サンプルに対して以下の情報を推定・抽出する．

機種名，メーカー名の推定：スクリーンショットと HTML ソースから機種名，メーカー名を LLM で推定する．

機種カテゴリ分類：文献 [8] を参考とした 9 カテゴリ（カメラ，ネットワーク，印刷機器等）への分類を LLM を用いて実施する（表 1 参照）．

バージョン抽出正規表現：製品のバージョンを抽出するための正規表現を生成する．本手法はすべての機器に対して LLM を用いてバージョン推定をするのではなく，各クラスタの代表サンプルのみを LLM を用いてバージョンを抽出するための正規表現を生成している，これにより，LLM の呼び出し回数を最小限に抑え，コスト面において効率化をはかっている．

シグネチャ生成：HTML ソースから機器固有のキーワードを抽出し，Censys 検索用クエリを自動生成する．可変情報（IP アドレス，バージョン番号等）を避け，製品固有の識別子を LLM を用いて選定する．なお，本研究では Censys で 1 ホスト以上の検索結果を得られたシグネチャを**有効シグネチャ**と定義する．

4.4.2 設置施設情報の推定

各クラスタの代表サンプルに対して以下の情報を推定・抽出する．重要施設設置機器と推定されたクラスタについては，クラスタ内の全 WebUI に対して設置施設情報の推

定を行う．

設置施設情報の推定：スクリーンショットと HTML ソースから設置施設名，設置施設の重要性，設置施設の分類を LLM を用いて推定する．重要施設関連性評価では，以下の分類基準を適用する．

- **施設の分野**：情報通信，金融，航空，空港，鉄道，電力，ガス，政府・行政サービス（地方公共団体を含む），医療，水道，物流，化学，クレジット，石油，港湾，不明
- **施設の重要性**：重要，非重要，不明

重要施設の判定にあたっては，総務省の調査 [9] および ICT-ISAC の報告書 [10] を参照し，それらで示された基準を適用した．特に，稼働停止が社会全体や人命に影響を及ぼす可能性がある施設を重要施設として位置付けた．

4.5 実行手順

本節では，前述した手法を実現するシステムを用いた，具体的な IoT 機器の発見・プロファイリング手順について，文献 [3] の 2 段階スキャン手法に基づいて説明する．本手法では，垂直展開スキャンと水平展開スキャンの 2 段階で IoT 機器の発見とプロファイリングを行う．

垂直展開スキャン：特定の AS 等の限定的な範囲に対してシステムを実行する．WebUI 収集はクエリ「autonomous_system.asn:[AS 番号] and services.http.response.status_code=200」により，指定した AS 番号に属する IP アドレスで HTTP サービスが稼働しているホストを検索する．

本スキャンでは機種情報生成が主目的であるため，プロファイリングは各クラスタの代表ホストにのみ実施する．設置施設情報については，各クラスタの代表サンプルから重要施設設置機器と推定されたクラスタにおいて，クラスタ内の全 WebUI に対し，設置環境情報を推定・抽出する．

水平展開スキャン：垂直展開スキャンで生成されたシグネチャを用いて Censys 等で世界規模の機器検索を実行する．本スキャンにおけるクラスタリング処理は，対象機種以外の機器やノイズを別クラスタに分離し，高精度な機種特定を実現している．検出された機器に対しては，機種情報が既知であるため，個別 WebUI 単位のプロファイリング（設置施設情報の推定）のみを実施する．

5. 評価

5.1 評価手法概要

正確な精度評価には，機種ごと・個別機器ごとの正しい情報が必要であるが，網羅的な取得は困難である．そこで，本研究では 5.2 節の評価用データセットを用い，LLM による結果と人手での推定結果の一致率で評価した．本研究の目的は人手プロファイリングの自動化であり，人間の評価者との一致率による評価は妥当である．

表 2 各項目の一致率評価結果

評価項目	全数	一致数	一致率
機種名	100	92	92.0%
メーカー名	100	91	91.0%
機器カテゴリ	282	246	87.2%
施設名	100	86	86.0%
施設分類	100	93	93.0%
施設重要性	100	91	91.0%

5.2 評価用データセット

日本国内の 1 つの AS（約 4 万 IP）に対する垂直展開スキャンで収集・クラスタリングした 497 クラスタを評価に使用した．各クラスタには HTML，スクリーンショット，LLM 推定結果，および人手評価ラベルが含まれる．

先行研究 [4], [5] では各項目が明確に特定可能な機器のみを評価対象としていたが，本研究では実運用環境での実用性検証のため，プロファイリング情報の特定が困難な場合も「不明」として記録した．その結果，497 クラスタ中 211 個（42.5%）で機種名・メーカー名の少なくとも一方が不明，86 個（17.3%）は両方不明，14 個（2.8%）は全項目不明となった．これにより識別困難な機器も含む現実的な評価を実現した．

5.3 評価結果

5.3.1 機種名・メーカー名の一致率

100 件のクラスタをランダムに抽出し，人手による推定結果と比較したところ，機種名 92.0%（92 クラスタ），メーカー名 91.0%（91 クラスタ）の一致率が得られた（表 2）．これにより，提案手法の機種・メーカー特定能力が十分に高く，先行研究 [4] の機種名 89.1%，メーカー名 89.1%と同程度の性能を確認した．

5.3.2 機器カテゴリの一致率

497 クラスタ中 282 クラスタをランダムに抽出し評価を行った結果，表 2 に示すように，246 クラスタ（87.2%）で人手による推定と機器カテゴリが一致した．

不一致の多くは，情報が不十分であるにもかかわらずメーカー名からカテゴリを断定したケースであった．例えば，監視カメラで著名なメーカーの製品を一律にカメラと判定したが，同社はビデオレコーダー等も製造しているため誤判定が生じた．

一方で，LLM が人間の判断を上回る精度を示した事例もあった．具体的には，HTML ソースコード内の IP カメラ特有の言語設定や PTZ 制御機能を検出し，人間が見落としがちな複数の証拠を組み合わせて，カメラ機器であることを正確に特定した．この結果は，LLM が技術的特徴を根拠に多角的な推論を行えることを示している．

5.3.3 設置施設関連情報の一致率

497 クラスタ中，代表サンプルから施設名抽出に成功し

表 3 LLM 生成シグネチャと人手作成シグネチャの一致率分布

一致率	シグネチャ数	累積比率
100%	19	63.3%
80%以上 100%未満	7	86.6%
50%以上 80%未満	0	86.6%
0%より大きく 50%未満	1	90.0%
0%	3	100.0%

た 18 クラスタを対象として，各クラスタ内の全 WebUI に対して個別プロファイリングを実施した．18 クラスタに含まれる 633 ホストのうち，100 ホストを抽出して評価した．

表 2 に示すように，施設分類で最も高い 93.0%の一致率を示した一方，施設名は 86.0%，施設重要性は 91.0%であった．施設名の一致率が相対的に低い要因として，入力情報から施設名を特定できず，不明とすべきケースでメーカー名を施設名と誤判断する事例が見られた．

施設分類では，LLM が検索機能を持たないため，未知の施設名の判断が困難であることも確認された．例えば，「〇〇真空システム」という施設名において，Web 検索により浄水関連施設であることが判明したが，LLM は「不明」と出力したケースが確認された．

このように，実世界のスキャン結果には，施設名のみでは分野の特定に追加情報が必要な施設も含まれるため，LLM による判定が困難な場合がある．この課題に対しては，Web 検索機能付き LLM による外部情報の活用が有効な解決策と考えられる．詳細は 7.1 節で考察する．

5.3.4 シグネチャ生成の一致率

LLM によるシグネチャ自動生成の有効性を検証するため，生成されたシグネチャと人手作成シグネチャをそれぞれ Censys で検索し，検出されたホストとの一致率を算出した．

有効シグネチャ（1 ホスト以上検出成功）30 個を抽出した結果，19 個（63.3%）が一致率 100%，26 個（86.6%）が 80%以上であり，先行研究 [5] の 28.6%を大幅に上回った（表 3）．高一致率の要因として，評価対象を有効シグネチャに限定したことで，HTML から有効なキーワードを抽出しやすいケースのみを対象としたことが挙げられる．一方，一致率 0%の 3 個は，いずれかの生成シグネチャが過度に特異的であり，検出数に大きな差が生じていた．この課題に対する具体的な改善策については，7.2 節で詳しく考察する．

これらの結果により，LLM によるシグネチャ自動生成は実用的な品質を実現できることが示された．特に，有効シグネチャに対しては高い一致率を達成しており，手動作成の代替手段として十分な性能を有していることが確認された．なお，本評価は有効シグネチャに限定しており，全シグネチャでの性能評価が今後の課題である．

表 4 実ネットワークにおける実証結果（AS 別）

項目	日本	海外 1	海外 2	海外 3	統合 *
総 IP 数	4 万	15 万	3,221	598	**20 万**
総クラスタ数	497	884	83	4	**1,468**
メーカー名	131	223	30	2	**336**
機種名	228	362	42	1	**618**
有効シグネチャ	311	573	46	3	**933**
水平展開スキャン検出数	298 万	972 万	53 万	18 万	**1,341 万**
重要施設検出クラスタ数	18	11	11	0	**40**
施設名抽出ホスト数	385	77	86	0	**548**
重要施設検出ホスト数	337	70	79	0	**486**

* 統合にあたり，重複する推定結果を手動で取り除いた．

6. 実ネットワークにおける実証

6.1 垂直展開スキャンによる機器探索

構築したシステムの実用性を検証するため，国内 AS1 つと海外 AS3 つを対象とした実証を実施した．表 4 に各 AS での探索結果を示す．

日本 AS（約 4 万 IP）では，497 クラスタ[*1]から 131 メーカー，228 機種を推定し，311 個の有効シグネチャ（1 ホスト以上の検出に成功したシグネチャ）を生成した．水平展開スキャンにより全世界で 298 万ホストの機器を発見した．施設情報については，18 クラスタが重要施設であった．重要施設クラスタに含まれる 633 ホストを個別分析した結果，385 ホストで施設名抽出に成功し，最終的に 337 ホストが重要施設設置機器と判明した．

海外 3 つの AS（約 450IP〜15 万 IP）では 971 クラスタ[*2]から 267 メーカー，405 機種を推定した．生成した 622 個の有効シグネチャを用いて全世界を検索した結果，1,043 万ホストの機器を発見した．

4 つの AS の結果を統合して重複を取り除くと，**336 メーカー，618 機種，933 個の有効シグネチャを生成し，全世界で約 1,300 万ホストの機器を発見**した．生成した全 1,394 個のシグネチャのうち 933 個（66.9%）が有効シグネチャであった．検出に至らなかった 461 個のシグネチャについては，先行研究 [5] でも報告されているキーワード条件の厳しさが要因と考えられる．

設置施設情報の推定において，1,468 クラスタのうち各機種クラスタから選定した代表サンプル（1 ホスト）を用いた予備推定の結果，**40 クラスタ (機種) が重要施設に関連する**ことが確認された．さらに，これら 40 クラスタに属する全 1,167 ホストを詳細に分析した結果，**548 ホストにおいて施設名の抽出に成功し，そのうち 486 ホストが重要施設に設置された機器**であることが判明した．なお，本

[*1] 2 クラスタは FortiGate によるアクセスブロックを IoT 機器と誤検知したため除外．
[*2] 同様に，5 クラスタは誤検知であったため除外．

表 5 機器 X のプロファイリング結果（106 機器）

重要度		分野	
重要施設	101	水処理	97
非重要施設	0	電力	3
		化学	1
不明	5	不明	5

推定は代表サンプルに基づくため，実際にはさらに多くの重要施設設置機器が存在する可能性がある．

施設名抽出率は地域・AS 種別で大きく異なり，総 IP 数に対する抽出率は，海外 AS2（2.45%），日本 AS（0.76%），海外 AS1（0.05%），海外 AS3（0%）の順であった．日本 AS は海外 AS1 の約 15 倍の抽出率を示し，日本の IoT 機器で施設名明記の傾向が強いことが確認された．抽出率の高い海外 AS2 は教育カテゴリの AS で施設名を明記する大学が多く含まれるため，例外的に多く抽出された．この結果は，地域特性に加えネットワークの組織種別を考慮することの重要性を示唆している．

6.2 水平展開スキャン

水平展開スキャンで特定した機種の詳細分析として，日本の AS で発見した遠隔監視・制御機器（以下，機器 X）に着目し，個別 WebUI 単位のプロファイリングを実施した．垂直展開スキャンで生成したシグネチャにより Censys で世界規模の検索を実施し，106 台を特定した．これらの機器に対して LLM による重要施設判定と分野推定を行った結果を表 5 に示す．

結果として，(1) 機器 X が水処理分野に集中して導入されていること，(2) 95%以上が重要施設に分類され，社会インフラで重要な役割を担っていること，(3) シグネチャを用いた水平展開スキャンの有効性が確認できた．これらの知見は，特定機種の世界的な導入傾向の把握や，セキュリティリスクの優先度決定など，実運用における意思決定支援に直接活用できる価値を持つ．

6.3 バージョン情報の抽出

日本の AS から収集した 497 クラスタを対象に分析を行った結果，91 クラスタで正規表現の生成が可能であった．表 6 に示すように，80 クラスタ（87.9%）で少なくとも 1 ホストのバージョン情報抽出に成功し，72 クラスタ（79.1%）ではクラスタ内の全ホストからバージョン情報を抽出できた．

LLM は手動では数分を要する正規表現を約 5 秒で生成し，多様な表記形式に対応した．例えば `Version\s([\d\.]+)` でバージョン「1.11」を，`#version\s*=\s*([\d,]+)` でバージョン「7,1,0,161」を抽出できる．特筆すべきは，WebUI に混在する機器バージョンとライブラリバージョンを区別して，機器のバージョンのみを正しく抽出できた点も重要である．

表 6 バージョン情報抽出の成功率

項目	件数	成功率
正規表現生成可能クラスタ	91	—
部分抽出成功クラスタ*	80	87.9%
完全抽出成功クラスタ	72	79.1%
対象ホスト総数	3,943	—
抽出成功ホスト数	2,799	71.0%

* 少なくとも 1 ホストのバージョン情報抽出に成功したクラスタ

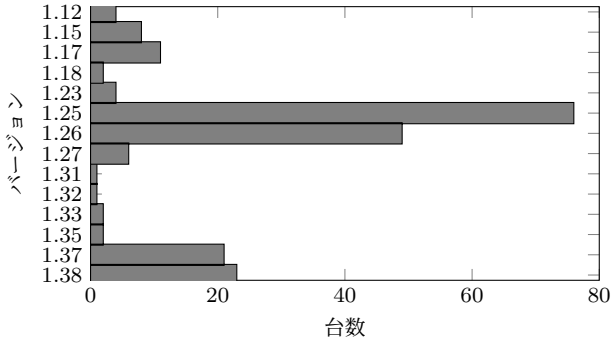

図 2 アクセスポイント機器のバージョン分布（210 台）

実用性検証として，特定のアクセスポイント機器 210 台に正規表現を用いてバージョン抽出を行ったところ，v1.25（76 台，36.2%），v1.26（49 台，23.3%）が多数を占めた一方で，v1.12〜v1.18 といった古いバージョンも 25 台（11.9%）存在した（図 2）．特に最古の v1.12（4 台）は既に脆弱性が報告され，メーカーから更新が推奨されていた．

以上より，実環境では新旧のバージョンが混在しており，正確なバージョン抽出が脆弱性評価に有効であることが示された．

7. 考察

7.1 実世界データ適用における課題と改善方針

本手法を実世界の大規模データに適用した結果，(1) 大規模適用時のコスト，(2) 情報不足による誤推定が課題として明らかとなった．

課題 (1) に対しては，個別 WebUI 単位（施設環境情報）の処理における 2 段階アプローチを提案する．第 1 段階で施設名抽出用の正規表現を生成後，第 2 段階では施設名が抽出されたクラスタのみ LLM による重要性判定と分類処理を実行する．全クラスタ処理から対象を限定した処理へ移行でき，コスト削減が可能となる．

課題 (2) に対しては，Web 検索機能付き LLM を活用した知識ベースを拡張するアプローチが効果的である．外部情報源からの正確な情報取得により，機器カテゴリ推定の高精度化（5.3.2 節），従来の LLM では対応困難であった施設情報における「不明」判定（5.3.3 節）が解決される．

上記の改善策を統合することで，より実用的な IoT 機器プロファイリングシステムの構築が可能となる．

7.2 シグネチャ品質向上への取り組み

本研究では，LLM を用いたシグネチャ自動生成の基盤技術を確立したが，実運用ではシグネチャが単一機種のみを正確に検出できることが求められる．具体的には，1 つのシグネチャが複数機種にマッチする事例や，当該機種を検索できない事例を最小化する必要がある．

対策として，元クラスタのホストを正解とし，生成シグネチャを Censys 検索で検証する．異種 WebUI の混入や検出漏れを確認し，誤検出要因に応じてキーワードの追加や条件調整を行う．この反復的改善により，誤検出抑制と高精度な機種特定を両立できる．

7.3 個別事例の分析

本研究で構築したシステムを実ネットワークに適用した際，重要施設と関連性が高い機器や施設名が抽出された機器について目視による追加調査を行った．その結果，適切なアクセス制御が施されておらず外部からログイン可能な機器や，認証なしで内部情報が閲覧可能な機器が多数確認された．特に，太陽光発電施設やダムの産業用制御機器など，攻撃を受けた場合に深刻な被害が想定される事例も含まれていた．

7.4 WebUI のコンテンツに基づいたリスク評価

プロファイリング結果と脆弱性データベース（CVE, NVD）の連携により，脆弱な機器の自動識別と，設置施設の重要度に基づくリスク優先度算出が可能となる．

このようなリスク評価に加えて，WebUI に表示される情報そのものを解析対象とすることで，「認証なしでアクセス可能な機能の危険度評価」「設定情報の露出レベル判定」「施設の社会的影響度（影響人口・経済規模）の推定」など，より詳細なリスク評価が可能となる．例えば，7.3 節で示した太陽光発電施設やダムの事例では，発電量や影響人口を踏まえたリスク分析が行えると期待できる．

7.5 研究倫理

インターネット上の IoT 機器 WebUI を対象としたスキャンにあたり，機器への影響を最小化するため，HTTP リクエストはトップページへのアクセスに限定した．また，スキャナサーバには研究目的・実施機関・連絡先を明記した Web ページを設置し，機器管理者が目的を理解できるよう配慮した．さらに，スキャン対象からの除外（オプトアウト）を希望する管理者が申請できる仕組みも提供した．本研究の成果は IoT セキュリティ向上に資する一方で，悪用の懸念もある．このため，論文中では機種名や施設名を匿名化し，攻撃に利用される恐れのあるシグネチャやキーワードの詳細は公開していない．加えて，研究過程で重要施設で利用されている可能性のある機器や 6.3 節で発見した脆弱な機器については，情報を精査した上で適切な機関へ提供し，成果がセキュリティ向上に資するよう努める．

8. 関連研究

体系的な IoT 機器の探索手法として，Censys を活用した垂直展開スキャンと水平展開スキャンの 2 段階構成によるアプローチが提案されている [3]．垂直展開スキャンで特定 AS や IP アドレス帯の IoT 機器を探索し，得られた機器情報からシグネチャを生成して，水平展開スキャンでインターネット全体の同種機器を探索する．この 2 つのスキャンにより，限られた範囲で得られた観測結果を起点として，より広域な探索へと発展させることが可能となる．

IoT 機器プロファイリング手法として，自然言語処理 [11] や，ニューラルネットワーク [12] による自動化が試みられているが，前者は複雑な推論を要するケース（例：機種特定に有用なキーワードの選定）への対応が難しく，後者は未学習機器への汎用性に乏しい．

これらの課題に対し，文献 [4], [5] では LLM を活用した Web インターフェース分析手法が提案されている．LLM の事前学習済み知識により未知機器にも対応でき，プロファイリング時間を約 20 秒に短縮し，キーワード抽出，検索クエリ生成，正規表現抽出を自動化している．

既存研究では個別の技術要素（探索，プロファイリング，シグネチャ生成）が提案されているが，これらを統合した自動化システムは実現されていない．本研究は，WebUI スキャナと LLM を統合し，探索からシグネチャ生成までの一連のプロセスを完全自動化している．

9. 結論

本研究では，WebUI スキャンと LLM プロファイリングの統合により，従来不可能であった大規模 IoT 機器探索の完全自動化を実現した．また，4 つの AS を対象とした実証により，933 個の有効シグネチャ生成と約 1,300 万ホストの機器発見を実現した．さらに，40 クラスタに属する重要施設設置の IoT 機器 40 機種を特定し，そのうち 486 ホストについては特に重要な IoT 機器であることを確認した．

謝辞 本研究の一部はＮＥＤＯ（国立研究開発法人新エネルギー・産業技術総合開発機構）の委託事業「経済安全保障重要技術育成プログラム／先進的サイバー防御機能・分析能力強化」（JPNP24003）によるものである．本研究は，国立研究開発法人情報通信研究機構 (NICT) の委託研究（JPJ012368C08101）により得られた成果を含む．

参考文献

[1] 笠間貴弘, 村上洸介. パスワード設定不備の IoT 機器における悪用リスクの実態解明. 情報通信研究機構研究報告, Vol. 70, No. 2, pp. 183–190, 2024.

[2] Rajasekar V R and Rajkumar Soundrapandiyan. A study on internet of things devices vulnerabilities using

shodan. *International Journal of Computing*, Vol. 22, , 2023.
[3] 平工瑞希, 佐々木貴之, 吉岡克成, 松本勉. 重要施設に設置された IoT 機器のインターネット全域探索. 電子情報通信学会技術研究報告; 信学技報, Vol. 121, No. 410, pp. 14–19, 2022.
[4] 合屋琴江, 佐々木貴之, 吉岡克成. LLM を用いた WebUI 解析による IoT 機器の機種と設置施設の推定. コンピュータセキュリティシンポジウム 2024 論文集, 2024.
[5] 合屋琴江, 佐々木貴之, 吉岡克成. IoT 機器の WebUI の LLM を用いたシグネチャ生成手法. 信学技報, pp. 120–125. 電子情報通信学会, 2025.
[6] Censys Inc. Censys - search engine for internet-connected devices. https://censys.io/, 2025.
[7] 内田佳介, 藤田彬, 吉岡克成, 松本勉. 管理 WebUI のカスタマイズに着目した遠隔監視制御用機器の探索手法. 暗号と情報セキュリティシンポジウム (SCIS), 2020.
[8] Xuan Feng, Qiang Li, Haining Wang, and Limin Sun. Acquisitional rule-based engine for discovering Internet-of-Things devices. In *27th USENIX security symposium (USENIX Security 18)*, 2018.
[9] 総務省. 重要 IoT 機器のセキュリティ対策に係る調査（令和 5 年度）. https://notice.go.jp/images/survey/survey_iot_security.pdf, 2024. 受託者: NTT コミュニケーションズ.
[10] 一般社団法人 ICT-ISAC. 脆弱な状態にある重要 IoT 機器の令和 5 年度調査及び注意喚起について（報告）. https://www.ict-isac.jp/news/news20240701.html, 2024.
[11] Ruimin Wang, Haitao Li, Jing Jing, Liehui Jiang, and Weiyu Dong. Iot device identification based on webui login pages. In *Sensors*, 2022.
[12] JinKe Song, Qiang Li, Haining Wang, and Limin Sun. Towards fine-grained fingerprinting of firmware in online embedded devices. Technical report, IEEE International Conference on Computer Communications, 2018.

付　　録

A.1 プロンプト

機種名・メーカー名抽出プロンプト

あなたは WebUI の画像と HTML から機種名とメーカー名を推定する専門家です。以下はある機器の Web 管理画面のスクリーンショットと、対応する HTML ソースコードです。この機器の機種名と機器のメーカー名を教えてください。推定できない場合は "Unknown" を返してください。
[HTML コード], [スクリーンショット]

機器カテゴリ推定プロンプト

あなたは IoT 機器の分類専門家です。Web 管理画面の画像、HTML、機器情報から適切なカテゴリを選択してください。
以下の情報から IoT 機器のカテゴリを分類してください：
メーカー：[メーカー名], 機種名:[機種名]
[HTML コード], [スクリーンショット]

シグネチャ生成プロンプト

あなたは HTML からキーワードを抽出する情報セキュリティの専門家です。指定された出力形式に必ず従ってください。
以下は、[メーカー名] というメーカーの [機種名] という機種の Web 管理画面の HTML ソースコードです。この機種を特定できるキーワードを HTML ソースコードから抽出してください。キーワードには，特定のバージョンの情報や IP アドレスなどを含めないようにしてください。できるだけ少ないキーワード数で、より多くの同機種の機器を検索できるように、かつ、別の機種の機器を検索しないようにしてください。また、Shodan 検索用のクエリとして、また，services.http.response.body: ‘キーワード 1’ and services.http.response.body: ‘キーワード 2’ and ... というフォーマットを作ってください．加えて、バージョンが含まれていれば、それを抽出する python の正規表現を作って下さい。バージョンがない場合は、Unknown を返してください。
[HTML コード]

設置施設関連情報を推定するプロンプト

あなたは WebUI の画像と HTML から施設名と施設の重要性を推定する専門家です。重要性と重要分野は、提供された選択肢の中から必ず選択してください。
以下は、あるネットワーク機器の Web 管理画面のスクリーンショットと、その HTML ソースコードです。この情報から、機器が設置されている施設名と、その施設の重要性および関連分野を推定してください。また、施設名が HTML ソースコードに含まれていた場合には、それを抽出する python の正規表現を作って下さい。施設名がない場合は、Unknown を返してください。判断に十分な情報がない場合は、適切な Unknown オプションを選択してください。

施設重要度判定基準：
■ 重要施設（Important）次のいずれかに該当する場合、Important と判断する：1. 人命・安全に直接関与している 2. 組織・地域の機能維持に関与する 3. 停止時の影響が重大であり、被害が複数人以上の生活・業務・安全に及ぶ可能性が高い
■ 重要でない施設（Not Important）次のすべてに該当する場合、Not Important と判断する：1. 停止しても人命や社会機能に即座の影響がない 2. 影響範囲が 1 世帯または個人レベルに限定される
■ 判断できない場合（Unknown）上記のいずれにも当てはまらない場合は、Unknown とする。
[HTML コード], [スクリーンショット]