サイバー攻撃分析のためのセキュリティレポート検索システム

添田 綾香[1] 長澤 龍成[1] 長田 侑樹[1] 白石 善明[1]
冨田 裕涼[2] 箕浦 翔悟[2] 毛利 公美[2] 森井 昌克[1]

概要： インターネットのインフラ化に伴い，企業や組織を標的としたサイバー攻撃が高度化・多様化してきている．組織の被害を最小限にするために，セキュリティ管理者は現在起こっているインシデントに関する情報を収集し，迅速に対応しなければならない．分析対象の事象に関わる具体的なキーワードがあらかじめわかっている場合には一般の検索エンジンで情報を得ることは可能である．しかしながら，適当なキーワードをセキュリティ管理者が想起できない場合は一般の検索エンジンでは良い結果が期待できない．そこで本論文ではシステムに残った Indicator of Compromise（IoC）情報をもとに，汎用的な検索エンジンを利用するより速く望ましい脅威情報を収集できるセキュリティレポート検索システムを提案している．本システムの検索には，トピックモデルの一種である LDA（Latent Dirichlet Allocation）を用いて付与した検索用ラベルと特徴ベクトルを用いている．提案システムは汎用検索エンジンよりも単純なキーワード検索では到達できない文書に容易に到達できることを確認している．

Security Information Retrieval System for Cyber Attack Analysis

AYAKA SOEDA[1], RYUSEI NAGASAWA[1], YUKI OSADA[1]
YOSHIAKI SHIRAISHI[1], YUSUKE TOMITA[2], SHOGO MINOURA[2],
MASAMI MOHRI[2], MASAKATU MORII[1]

1. まえがき

インターネットのインフラ化に伴ってサイバー攻撃の被害の影響度は高く，攻撃手段は高度化・多様化してきている．サイバー攻撃の事後対応が遅れると関係する組織に多大な損害が生じることもあり，組織のセキュリティ管理者には迅速な事後対応と事前の予防的対策が求められる．このとき，限られた人的資源で効果的なセキュリティオペレーションを行うためにオペレーションの省力化が望まれている．

事後対応や予防的対策のためのセキュリティ管理者による情報収集ではインターネットが活用される．検索対象として，過去にあったインシデント事例について，セキュリティベンダーや個人の研究者が脆弱性や攻撃手法，攻撃対象といった内容をまとめたものを含むセキュリティレポートは有益である．

セキュリティレポートは Google や Bing といった汎用的な検索エンジンを使って一般に検索される．セキュリティ管理者が当該インシデントに関する調査を行うとき，検索したいマルウェアの名前や特徴的な検索キーワードを想起できており，かつ，インシデントの様子がいくらか想像できている場合は，これらの検索エンジンを使用して適当なセキュリティレポートが得られる．しかしながら，攻撃者が使用するマルウェアのファイル名や IP アドレスなどの Indicator of Compromise（IoC）情報を用いて検索を行うとき，汎用的な検索エンジンの検索結果には様々な情報が混在し，セキュリティレポート以外も出力される．例えば IP アドレスをクエリに検索すると，当該あるいは類似の IP アドレスに対して行われたマルウェアの攻撃傾向に関するセキュリティレポートではなく，その IP アドレスの所有者情報を記載したページが出力される．そのためセキュリティ管理者が検索結果一つ一つを確認してセキュリティレポートを抽出することとなる．

セキュリティ管理者はシステムに不審な点があったとき，まず調査対象を絞るためにマルウェアの操作したレジストリや攻撃元 IP アドレスなどの初期に手に入る IoC 情報から ExploitDB[1]や VirusTotal[2]などに入力した結果のハッシュ値を二次的な IoC 情報として入手する．そのハッシュ値を Google のような汎用的な検索エンジンに入力してもハッシュ値が羅列されたページが多く出力され，適切な情報を得るには検索を何度も繰り返さなければならない．そこでこのハッシュ値を入力として攻撃者の使用したマルウェアや脆弱性に関連する情報をより速く的確に得ることを本研究の目的とする．

提案する検索システムでは事前に収集したセキュリティレポートをデータベースに蓄積し，まず，マルウェアのハッシュ値（MD5，SHA256，etc.）やその通信先の IP アドレスといった IoC 情報をクエリとして検索を行う．次に，トピックモデルの一種である LDA により作成した検索用のラベルと特徴ベクトルを用いて関連文書を出力する．

本論文では提案システムについて述べたあと，ユーザ実験によりセキュリティレポートを探し出す時間が短縮すること，応答性能に関する実験によりデータ件数が増加しても検索にかかる時間が変化しないことを確認している．

1 神戸大学大学院工学研究科
Dept. of Electrical and Electronic Engineering, Kobe University
2 岐阜大学大学院自然科学技術研究科
Dept. of Electrical, Electronic and Computer Engineering, Gifu University

2. 関連研究

2.1 関連論文の推薦・検索

セキュリティレポートは NVD(National Vulnerability Database)[3] や CVE(Common Vulnerabilities and Exposures)[4]のような構造化された文書ではなく自然言語で書かれた専門性の高い非構造化文書である．自然言語で書かれている専門性の高い文書である学術論文の検索に関する研究での知見はセキュリティ文書の検索にも応用できると考えられる．文書本文に付与された検索タグを用いる検索方法と，文書本文を何らかのベクトルで表現した検索方法の二つに文書検索は大きく分けられる．

前者のタグによる検索の長所はそのタグを用いて論文をカテゴリー分けできることである．論文検索ではまず文レベルで検索を行うために各文に背景，手法，結論といった注釈を付ける AZ(Argumentative Zoning)の研究[5，6]がある．AZ はアノテーションスキーム(付与を行う方法)とそれに基づいた各文へのアノテーション(注釈の付与)からなる．学術論文のような専門的な文書に文単位でのアノテーションを行うには，労力を要する．そこで文より大きな単位の章レベルで検索を行うために，ルールベースの章へのタグの付け方を利用した Section tagger を用いた手法が提案されており，機械学習を用いることが今後の課題となっている[7]．テキストにタグをつける方法でデータセットを用いて検索モデルを訓練するとき，学術論文は専門性の高い文書であり，大規模なデータセットを用意するのは．容易ではない．しかし，カテゴリーを表すタグは利用者の理解を助けるという点で有用である．

後者の文書本文をベクトル化して検索に用いる手法の長所は個別の文書間の距離を求めることができることである．ベクトル化の方法は大きくテキストベースとグラフベースに分けられる．テキストベースでは検索システムのユーザーの発表論文を TF-IDF を用いてベクトル化し，それにより研究の関心をプロファイルして学術論文を推薦するシステム[8]がある．一方グラフベースでは LINE という埋めこみグラフを利用して，ノードである論文の引用・被引用の関係から隣接行列をもとにベクトル化し，学術論文の分類に利用するもの[9]がある．ここでは，学術論文を単語の出現頻度によってベクトル化した BoW(Bag of Words)や，埋めこみグラフの情報だけを利用しているため，単語を固定長のベクトルで表現する Word2Vec[10], 引用機能(citation function)[11]といった，より高次な情報を利用することで望ましい検索結果が得られると期待できる．例えば，文書の潜在的な意味を解析するトピックモデルの一つ，LSA(Latent Semantic Analysis)を用いて神経科学学会のポスターをベクトル化し，気に入ったポスターから類似ポスターを検索するシステム[12]がある．このように単純な BoW 形式のベクトルをそのまま用いて検索を行うのではなく，文書の潜在意味解析を行ったモデルによりベクトル化することで関連文書検索の品質向上が期待できる．

本論文の提案システムは，前者の検索タグとなるラベルと後者の特徴ベクトルの両者による検索を行い，セキュリティレポートのそれぞれに対するそのラベルの付与と特徴ベクトルへの変換を，セキュリティレポートの集合に対して潜在意味解析を行う教師無し機械学習のモデルを用いて行う．

2.2 トピックモデルによる検索タグと特徴ベクトルの作成

2.2.1 トピックモデル

トピックモデルは文書中に出現する単語の種類と出現頻度に基づいて，その文書の潜在的意味を解析する手法の一つである．トピックモデルは，BoW 形式の文書ベクトルを基に単語の共起関係を学習し，文書に潜在的に存在するトピック（話題）を解析する．この学習したモデルにより各文書の話題を考慮したベクトル化ができる．

2.2.2 LDA(Latent Dirichlet Allocation)

2007 年に Blei らによって提案された LDA(Latent Dirichlet Allocation)[13]は BoW 形式の文書ベクトルをモデル化する教師なし学習のトピックモデルの一つである．各文書には観測できない複数のトピックが潜在的に存在すると仮定し，観測される文書中の単語の共起性を潜在トピックという確率変数によって定式化する．図 1 に LDA のグラフィカルモデルを示す．

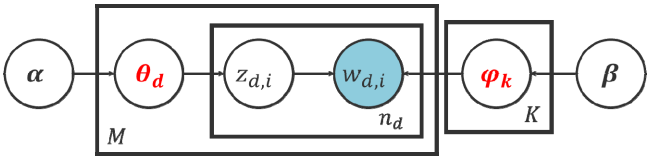


図 1　LDA のグラフィカルモデル．
Figure 1　Graphical model of LDA.

LDA における文書ベクトルの生成過程は以下のようになる．文書数をM，トピック数をK，ある文書dに現れる総単語数をn_dとする．LDA では，ある文書dは複数のトピックから構成されていると仮定し，トピック分布θ_dという確率分布でベクトル化される．文書dでトピックkが出現する確率を$\theta_{d,k}$とすると，$\theta_d = \{\theta_{d,1}, \theta_{d,2}, \ldots, \theta_{d,K}\}$である．ここで，$\theta_d$はディリクレ分布$Dir(\alpha)$から生成され，$\alpha = (\alpha_1, \ldots, \alpha_K)$は$K$次元ベクトルである．また，各トピック$k = \in \{1,2, \ldots, K\}$は単語分布$\phi_k$という確率分布でベクトル化される．トピック$k$における単語$v$の出現確率を$\phi_{k,v}$とすると，$\phi_k = \{\phi_{k,1}, \phi_{k,2}, \ldots, \phi_{k,v}\}$である．ここで，$\phi_k$はディリクレ分布$Dir(\beta)$から生成され，$\beta$は単語の種類数$V$を用いて$\beta = (\beta_1, \ldots, \beta_V)$と表される．文書$d$中の$t$番目の単語$w_{d,t}$が潜在的に持つトピックを$z_{d,t} \in \{1,2, \ldots, K\}$とすると，単語$w_{d,t}$は単語分布$\phi_{z_{d,t}}$から生成される．潜在トピック$z_{d,t}$は多項分布$Multi(\theta_d)$から生成され，単語$w_{d,t}$は多項分布$Multi(\phi_{z_{d,t}})$から生成される．なお，トピック数$K$とディリ

クレ分布のパラメータαおよびβは，学習時に予め指定するパラメータである．

2.2.3 ラベルの付与と特徴ベクトルへの変換

脅威情報へのラベル付け手法は文献[14-17]をはじめとしていくつか存在するが，これらの手法は脅威情報に単語単位でラベル付けをすることでセキュリティ管理者の理解を助けるものである．本論文の提案システムでは文献[18]の LDA を用いたセキュリティレポートに対するラベルの付与手法を用いる．この手法は学習済みモデルからセキュリティレポートのそれぞれに文書ベクトルθと単語分布ϕを求め，それらをもとにθとϕによって複数個のラベルを生成することができるマルチラベル付与手法である．例えば，攻撃手法と攻撃対象の話題が含まれているセキュリティレポートがあったときに，単一のラベルよりも複数のラベルを付与する方がより適当な検索結果が得られると期待できる．

文書ベクトルθを特徴ベクトルとして各セキュリティレポートに付与する．なお，セキュリティレポートは学術論文と異なり，セキュリティベンダーが発行する月次レポートのような複数の話題を取り扱う文書（“**サマリー文書**”と呼ぶ）が存在し，サマリー文書とそれ以外の文書を区別することなく同じアルゴリズムでベクトル化するとラベルの品質が下がる．そこで文献[18]の手法では，サマリー文書の集合を特定し，特徴ベクトルの再計算とラベルの再付与を行っている．

3. セキュリティレポート検索システム

3.1 システムの要件

本論文でのセキュリティレポートを検索するフローを図 2 に示す．ユーザは初めに IoC 情報などのキーワードをクエリとした全文検索を行う．検索サーバは最初の全文検索の結果に付いているラベルを用いてバックエンドで検索を行い，全文検索とラベルによる関連文書検索の結果を出力する．フロントエンドに出力された検索結果から，ユーザが注目するレポートを選択し，そのレポートの特徴ベクトルを用いて関連文書検索を行い，セキュリティ情報を収集する．このような検索を可能にするシステムの要件は以下のように定められる．

要件 1：全文検索が可能

IoC 情報などのキーワードをクエリとし，キーワードが含まれる文書を提示する．提示された文書に付いているラベルが要件 2 のクエリとなる．

要件 2：ラベルをクエリとした関連文書検索が可能

文書に付いているラベルに基づいて，関連文書を提示する．提示された文書が要件 3 のクエリとなる．

要件 3：文書をクエリとした関連文書検索が可能

文書の関連性に基づいて，文書を提示する．

要件 4：文書を検索可能な形式に加工・保存が可能

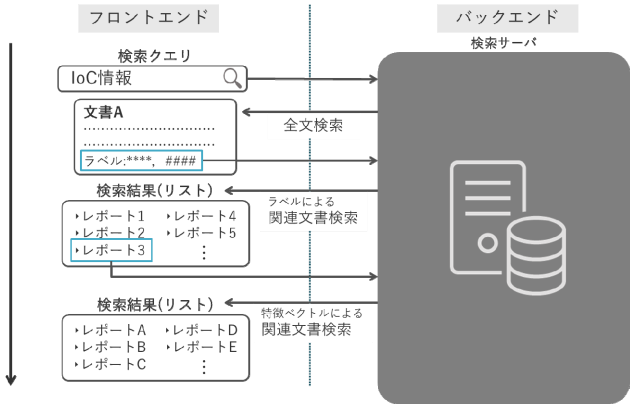


図 2　提案する検索システムのフロー．
Figure 2　Proposed search system flow.

関連文書検索に必要なラベルや特徴ベクトルを作成し，全文検索に必要な情報と紐づけてデータベースに保存する．

要件 5：オリジナル文書を閲覧可能

検索結果で提示された文書の詳細が必要な際に，オリジナル文書の閲覧を可能にする．

3.2 提案システム

セキュリティレポート検索システムの構成と機能を図 3 に示す．システムの主な機能としては“全文検索機能”と“関連文書検索機能”がある．この機能を運用するために“ラベル取得機能”，“特徴ベクトル取得機能”，“検索用ラベル作成・保存機能”，“特徴ベクトル作成・保存機能”，“文書の加工・登録機能”，“オリジナル文書閲覧機能”があり，個々の機能の動作は次のとおりになる．

全文検索機能：検索フォームに入力されたキーワードから全文検索を行う．【要件 1 に対応】

関連文書検索機能：クエリとなる文書に対応するラベルや特徴ベクトルを用いて関連文書検索を行う．【要件 2・3 に対応】

ラベル取得機能：全文検索でヒットしたレポートのラベル

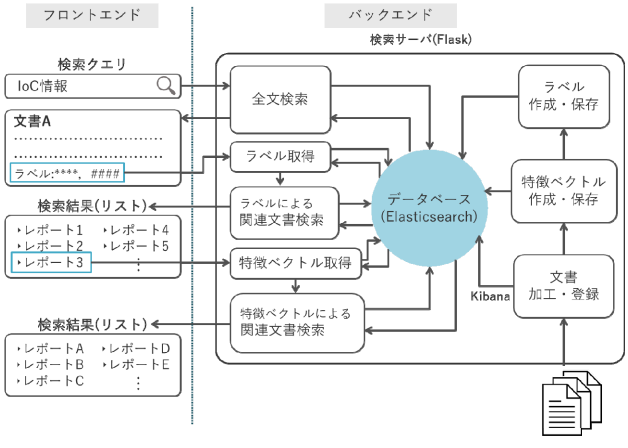


図 3 提案するセキュリティレポート検索システムの構成と機能．
Figure 3　Proposed security report search system configuration and functions.

を取得する．【要件 2 に対応】

特徴ベクトル取得機能：クエリとして入力された文書に付与された特徴ベクトルを取得する．【要件 3 に対応】

検索用ラベル作成・保存機能：文書からラベルを作成する．関連文書検索で利用する．【要件 4 に対応】

特徴ベクトル作成・保存機能：文書から特徴ベクトルを作成する．関連文書検索で利用する．【要件 4 に対応】

文書の加工・登録：検索対象の文書にラベルと特徴ベクトルを付与し，データベースに登録する．【要件 4 に対応】

オリジナル文書閲覧：検索結果で提示される文書のオリジナル文書を表示する．【要件 5 に対応】

要件 2 と要件 3 に対応するための検索用ラベルと特徴ベクトルの生成は 2.2.3 で述べた方法により行う．各レポートには 50 個ずつのラベルを付与している．

4. Web システムの実装

4.1 実装環境

データベースに Elasticsearch[19]を利用し，バックエンドは Python の Flask フレームワーク，検索対象の文書やラベルの登録には Python と Kibana[20]，特徴ベクトルの距離計算にはユークリッド距離を用いた．

4.2 Web システムの検索手順

図 4 から図 6 にフロントエンドでの検索画面，及び検索結果の表示例を示す．

Step1．全文検索（フロントエンド）

Web システムのフォーム（図 4）にマルウェアのハッシュ値（MD5，SHA256，etc.）などの IoC 情報を入力し，submit ボタンをクリックすると Elasticsearch 内で全文検索が行われる．

Step2．ラベルによる関連文書検索（バックエンド）

全文検索の結果に付いているラベルを用いて関連文書検索を行う．具体的には，レポートへの関連度が高い上位 15 個のラベルを用いて検索を行う．50 個のラベルはレポートごとに異なるため，どのセキュリティレポートのラベルを用いて検索しても 10～100 件のセキュリティレポートが出力されるようにレポートへの関連度が高い上位 15 個のラベルを用いることとした．

図 4　Web システムのキーワード入力フォーム画面．
Figure 4　Keyword input form screen of the Web system.

Step3．注目するレポートの選択（フロントエンド）

Step2 の実行した後に図 5 のような画面が Web ブラウザに出力される．最上段にラベルによる関連文書検索に用いたラベル，その下に全文検索でヒットして Step2 にラベルを提供したセキュリティレポート，次にラベルによる関連文書検索結果のレポート一つ一つがリスト形式で表示される．検索結果表示画面の「続きをよむ」（黒色）ボタンを押すと本文，発行元を確認することができる．「このレポートに関連したレポートを探す」（青色）ボタンをクリックすると当該レポートの特徴ベクトルをクエリとした関連文書検索が行われる．

Step4．特徴ベクトルによる関連文書検索（バックエンド）

Step3 でユーザが選んだレポートの特徴ベクトルに対して Elasticsearch に登録されている全てのセキュリティレポートの特徴ベクトルとのユークリッド距離を計算し，距離が短い上位 20 件のレポートをリスト形式で関連レポートとして出力する（図 6）．検索結果表示画面の「続きをよむ」ボタンを押すと本文，発行元，注目するレポートとして選んだレポートとそのレポートとのユークリッド距離を確認することができる．

Step5．複数回検索

関連文書検索で出力されたレポートからさらに注目す

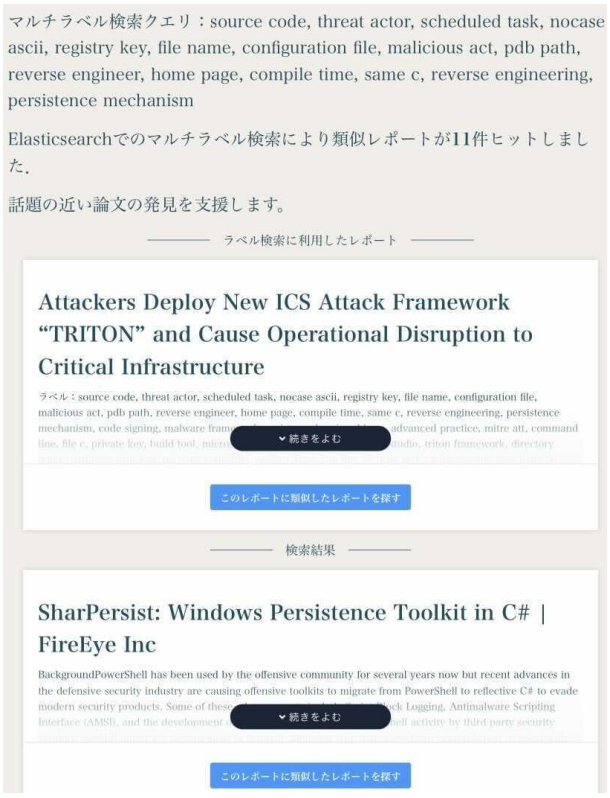


図 5　Web システムのキーワード検索・ラベルによる関連文書検索結果出力画面
Figure 5　Keyword search and label-based related document search result output screen of the Web system.

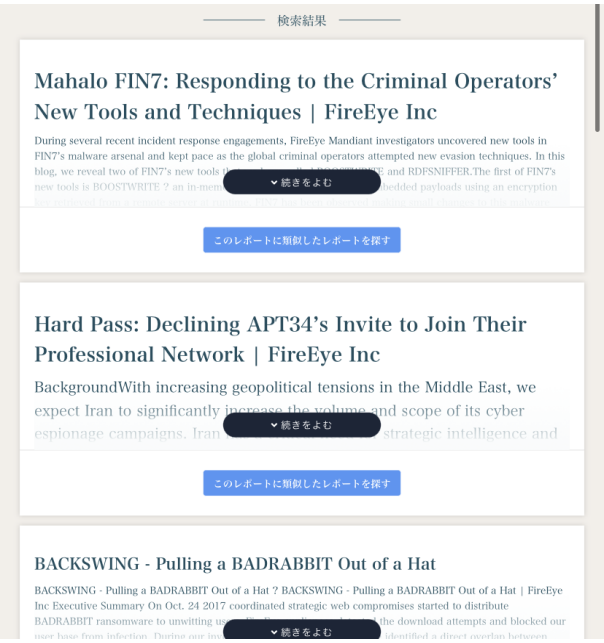


図 6　Web システムの特徴ベクトルによる関連文書検索結果出力画面.
Figure 6　Feature vector based related document search result output screen of the Web system.

るレポートを選択して「このレポートに関連したレポートを探す」ボタンをクリックし，Step3 と Step4 を繰り返すことができる.

5.　検索システムのユーザ実験

提案するセキュリティレポート検索システムを用いることで検索時間が短縮されることを確認するための，既存の汎用的な検索エンジンである Google を用いた検索と比較するユーザ実験について述べる.

5.1　データセット

セキュリティベンダー8 社(Trend Micro[21], Cisco[22], Druva[23], Broadcom [24], Barracuda[25], FireEye[26], NETSCOUT [27], Paroalto[28])が 2017 年 1 月 1 日～2019 年 12 月 31 日に発行した 2385 件のセキュリティレポートを検索対象の文書とし，タイトル，本文，発行日，発行元のデータにラベルと特徴ベクトルを付与して CSV ファイルにまとめ，データベースに登録した.

5.2　実験方法

被験者は神戸大学でセキュリティに関する研究を行っている学生 8 名である．被験者をグループ A，グループ B に 4 名ずつランダムに分けた．被験者にはセキュリティインシデントが発生した状況にて IoC 情報が得られたとし，当該インシデントに関する情報を検索エンジンからできるだけ速く正確に収集することが目的であると伝えている．被験者にはセキュリティインシデントに関する穴埋め問題を提示し，IoC 情報と問題文中に含まれるキーワードをもとに被験者が任意の検索クエリを設定して情報を収集し，解答を記入する．実験中は操作画面を動画で記録する．実施日は 2021 年 1 月 22 日である.

問題は練習 1，練習 2，問 1（1-1, 1-2），問 2（2-1, 2-2）の計 6 問を用意し，どちらのグループにも同じ問題を提示する．被験者の能力差が結果に影響を与える可能性があるため，問 1 と問 2 の大問 2 種類を用意してどちらのグループにも提案システムと Google 検索の両方を使用してもらう．また，最初に解答した大問で解答の要領を掴んだことでその後の問題のスコアが良くなるということが考えられるため，大問で用いる検索エンジンをグループ毎に入れ替える．それぞれのグループごとの実験の手順は以下のとおりである.

【グループ A】
(1) 実験方法の説明
(2) Google 検索で練習問題を 1 問解答
(3) Google 検索で本番問題を 2 問（大問 1 問）解答
(4) 休憩
(5) 提案手法の検索システムで練習問題を 1 問解答
(6) 提案手法の検索システムで本番問題を 2 問（大問 1 問）解答
(7) アンケートの回答

【グループ B】
(1) 実験方法の説明
(2) 提案手法の検索システムで練習問題を 1 問解答
(3) 提案手法の検索システムで本番問題を 2 問（大問 1 問）解答
(4) 休憩
(5) Google 検索で練習問題を 1 問解答
(6) Google 検索で本番問題を 2 問（大問 1 問）解答
(7) アンケートの回答

問題の内容は 5.1 で述べた 2385 件のデータセット内のセキュリティレポートから答えられる範囲で作成し，問題の量や難易度が同等になるように作成している．練習 1 と練習 2 は制限時間を 15 分，5 点満点とし，問 1（1-1，1-2）と問 2（2-1，2-2）は制限時間を 20 分，11 点満点とした．図 7 は練習 1 の問題である．md5 や sha256 のハッシュ値を基に各検索エンジンでセキュリティ情報を収集し，穴埋め問題に解答する.

（練習 1）
　中国の標的型攻撃グループ APT10 の活動.
　関連問題:アジア圏を標的としたマルウェア活動.
（問 1）
（問 1－1）バックドアマルウェア LOWKEY.
　関連問題:その他のバックドアマルウェア.
（問 1－2）APT33 の活動.
　関連問題:過去の APT33 のその他の攻撃活動.
（練習 2）

マルウェア Triton.

関連問題:Triton の詳細な解析.

（問 2）

（問 2－1）マルウェア Nanocore.

関連問題:アンチウイルスソフトを回避する攻撃.

（問 2－2）DDoS ボットネット Flusihoc.

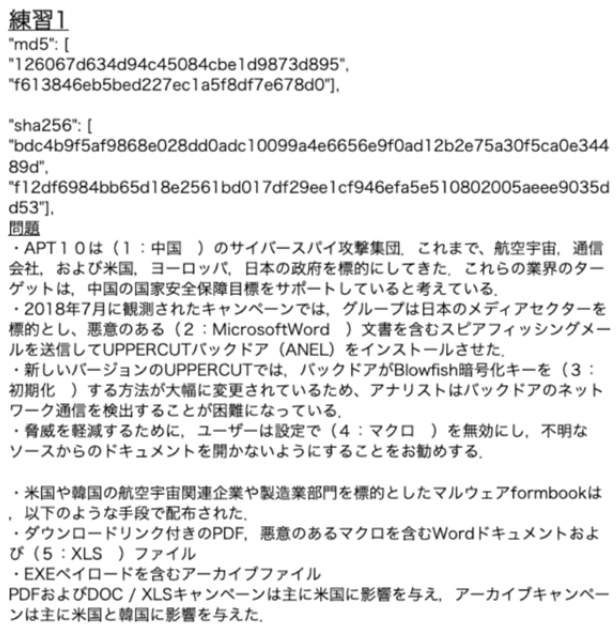
関連問題:IoT.

5.3 実験結果

ユーザビリティは ISO9241-11 において「特定の利用状況下で，特定のユーザによって，ある製品が指定された目標を達成するために用いられる際の，有効さ，効率，ユーザの満足度の度合い」と定義されている．この定義に基づき提案システムの有効性, 効率性, 満足度の評価を行った．有効性及び効率性の評価のため，穴埋め問題をどの程度正解できたか，解答完了までにどの程度時間がかかったかを計測した．また満足度の評価のため，提案システムに対するアンケート（5 段階と自由記述）を行った．

表 1 と表 2 に，問 1 および問 2 におけるグループごとの解答完了までにかかった平均時間と平均正解数を示す．問 1 では提案システムを用いたグループ B の方が平均で 4 分 59 秒短い時間で解答を完了でき，正解数も多かった．問 2 においても同様に，提案システムを用いたグループ A の方が平均で 4 分 49 秒短い時間で解答を完了でき, 正解数も多かった．このことから，検索エンジンを使う順番によらず Google 検索を使う場合に比べて提案システムを使う場合が良い結果になることがわかる．また，グループ A の平均時間が，問 1 から問 2 で 4 分 50 秒短縮された．グループ B の平均時間は 4 分 58 秒延長された．このことから，個人の能力差によらず提案システムが有用であると言える．

表 1 より Google 検索を用いたグループ A の正解数の標準偏差が大きかった．11 問全て正解した被験者がいる一方, 3 問しか正解しなかった被験者も存在した．そして，問 1, 問 2 ともに Google 検索を用いたグループにおける解答時間の標準偏差は小さかった．19 分 34 秒ないし 36 秒の平均時間であることから, 多くの被験者が 20 分の制限時間内に解答完了できなかったことを表している．これらの原因として，Google 検索では関連情報を収集する際にユーザが検索クエリを自身で選定する必要があり，被験者の能力に応じて結果に差が出たと考えられる．実際，短時間で正解数の多い被験者は上手に検索クエリを選定して数回の検索で目的の情報にたどり着いていた．悪い結果の被験者は様々なクエリで繰り返し検索を行っており，時間内に目的の情報にたどり着けなかった．一方の提案システムでは，トピックモデルによるラベルと特徴ベクトルにより関連情報がシステムにより提示されるため，被験者が検索クエリを選定する必要がなく，解答完了時間にばらつきがなく，総じて短くなったと考えられる．


練習1
"md5": [
"126067d634d94c45084cbe1d9873d895",
"f613846eb5bed227ec1a5f8df7e678d0"],

"sha256": [
"bdc4b9f5af9868e028dd0adc10099a4e6656e9f0ad12b2e75a30f5ca0e34489d",
"f12df6984bb65d18e2561bd017df29ee1cf946efa5e510802005aeee9035dd53"],
問題
・APT１０は（１：中国　）のサイバースパイ攻撃集団．これまで，航空宇宙，通信会社，および米国，ヨーロッパ，日本の政府を標的にしてきた．これらの業界のターゲットは，中国の国家安全保障目標をサポートしていると考えている．
・2018年7月に観測されたキャンペーンでは，グループは日本のメディアセクターを標的とし、悪意のある（２：MicrosoftWord　）文書を含むスピアフィッシングメールを送信してUPPERCUTバックドア（ANEL）をインストールさせた．
・新しいバージョンのUPPERCUTでは，バックドアがBlowfish暗号化キーを（３：初期化　）する方法が大幅に変更されているため、アナリストはバックドアのネットワーク通信を検出することが困難になっている．
・脅威を軽減するために，ユーザーは設定で（４：マクロ　）を無効にし，不明なソースからのドキュメントを開かないようにすることをお勧めする．

・米国や韓国の航空宇宙関連企業や製造業部門を標的としたマルウェアformbookは，以下のような手段で配布された．
・ダウンロードリンク付きのPDF，悪意のあるマクロを含むWordドキュメントおよび（５：XLS　）ファイル
・EXEペイロードを含むアーカイブファイル
PDFおよびDOC / XLSキャンペーンは主に米国に影響を与え，アーカイブキャンペーンは主に米国と韓国に影響を与えた．

図 7　練習問題 1 の内容.
Figure 7　Contents of Exercise 1.

表 1　問 1 におけるグループごとの解答完了までの平均時間と平均正解数.
Table 1　Average time and average number of correct answers for each group in Question 1.

問 1（1-1,1-2）	時間	
	平均	標準偏差
A（Google 検索）	19 分 35 秒	0 分 44 秒
B（Web システム）	14 分 36 秒	3 分 41 秒

問 1（1-1,1-2）	正解数	
	平均	標準偏差
A（Google 検索）	5.8	3.1
B（Web システム）	10.5	0.5

表 2　問 2 におけるグループごとの解答完了までの平均時間と平均正解数.
Table 2　Average time and average number of correct answers for each group in Question 2.

問 2（2-1,2-2）	時間	
	平均	標準偏差
A（提案システム）	14 分 45 秒	4 分 01 秒
B（Google 検索）	19 分 34 秒	0 分 45 秒

問 2（2-1,2-2）	正解数	
	平均	標準偏差
A（提案システム）	9.0	1.4
B（Google 検索）	6.5	1.7

5.4 アンケート内容

ユーザ実験後，被験者に提案システムに関するアンケートを実施した．6 つの質問項目については 5 段階で評価してもらい，2 つの項目は自由記述とした．質問項目は以下のとおりである．

[5 段階評価]

- 出力文書について

(1) 出力されたレポートにまとまりや関連性はありましたか？(1:関連していなかった，2:どちらかというと関連していなかった，3:どちらともいえない，4:どちらかというと関連していた，5:関連していた)

(2) 検索キーワードから有益な関連情報を含むレポート

を得られたと思いますか？(1:得られなかった，2:どちらかというと得られなかった，3:どちらともいえない，4:どちらかというと得られた，5:得られた)

- システムの使用感について

(3) システムの使い方はわかりやすかったですか？(1:わかりにくい，2:ややわかりにくい，3:どちらともいえない，4:ややわかりやすい，5:わかりやすい)

(4) システムを使いこなせたと思いますか？(1:使いこなせなかった，2:やや使いこなせなかった，3:どちらともいえない，4:やや使いこなせた，5:使いこなせた)

(5) 慣れたら使いやすそうだと感じましたか？(1:慣れても使いにくそう，2:やや使いにくそう，3:どちらともいえない，4:やや使いやすそう，5:慣れたら使いやすそう)

(6) Google 検索よりも有用だと感じましたか？(1:有用でない，2:やや有用でない，3:どちらともいえない，4:やや有用，5:有用)

[自由記述項目]

A) Google 検索で問題を解くうえで難しかった点を教えてください.

B) 評価用システムで問題を解くうえで良かった点，難しかった点や改善点を教えてください.

5.5 アンケート結果

5 段階評価の結果については表 3 のようになった．質問項目(1)と(2)については，どちらのグループも平均評価値が 4.0 以上となったことから，提案システムは検索クエリに関連した情報を的確に出力できていると考えられる．質問項目(3)については全員が 4.0 以上の評価になっており，短い時間でも使い方を理解できるシステムであるといえる．また，質問項目(4)については，どちらのグループも平均評価値が 3.0 という結果になった．このことから，使い方は理解できたが，システムを駆使して幅広い関連情報を収集するには実験時間が短かったと考えられる．しかし，質問項目(5)と(6)の平均評価値が 4.3 を上回っており，システムに慣れることで Google 検索より有効に使えると感じた被験者が多かった.

自由記述項目(A)について，Google 検索でセキュリティ情報を検索する際の難点として「マルウェアのハッシュ値のデータベースなどセキュリティレポート以外のサイトが表示され，情報を収集し辛い」,「検索キーワードを考えるのが難しい」などが挙げられた．本実験の被験者はセキュリティに関する専門知識を有する学生であったが，多くの被験者が検索クエリを選定する作業を負担に感じていた．自由記述項目(B)について，提案システムの良かった点として「ハッシュ値や IoC さえ入力すれば，解答に必要な情報に関連の深い文献が上位に表示されていた点で優秀であると感じた」という点が挙げられた．トピックモデルによる話題分析の自動化によって情報収集の省力化を実現していると考えられる.

一方，提案システムの難点として「ハッシュ値の前後に半角の空白や[.]を含むクエリで検索するとレポートが出力されない」など入力方法に関する意見があった．入力方法についての検討は今後の課題である.

表 3 アンケート結果
Table 3 Questionnaire results.

質問項目	グループ	評価分布					平均	標準偏差
		1	2	3	4	5		
(1)	A	0	0	0	4	0	4	0
	B	0	0	1	1	2	4.3	0.8
(2)	A	0	0	0	2	2	4.5	0.5
	B	0	0	0	0	4	5	0
(3)	A	0	0	0	0	4	5	0
	B	0	0	0	2	2	4.5	0.5
(4)	A	0	1	2	1	0	3	0.7
	B	0	1	2	1	0	3	0.7
(5)	A	0	0	0	2	2	4.5	0.5
	B	0	0	1	1	2	4.3	0.8
(6)	A	0	0	1	1	2	4.3	0.8
	B	0	0	0	1	3	4.8	0.4

6. 応答性能

純増するセキュリティレポートをシステムに追加することを想定し，データ数が増えたときの Web システムの応答に関して計測した.

6.1 データセット

セキュリティベンダー17 社(Trend Micro[21], Cisco[22], Broadcom[23], Barracuda[24], Druva[25], FireEye[26], NETSCOUT [27], Paroalto[28], CrowdStrike[29], ESET[30], Darktrace[31], SecPod[32], Check Point[33], Fortinet[34], Netlab Blog[35], Secureworks[36], JPCERT[37])が 2016 年 1 月 1 日～2021 年 4 月 9 日に発行した 13161 件のセキュリティレポートを検索対象の文書とし，タイトル，本文，発行日，発行元，記事のタイプ，言語，URL のデータにラベルと特徴ベクトルを付与してデータベースに登録した.

6.2 方法と結果

まず 13161 件のデータセットから 2500, 5000, 7500, 10000, 13161 件のデータセットを作成した．本システムにおいてラベルによる関連文書検索(Step2)と特徴ベクトルによる関連文書検索(Step4)は異なるクエリを用いて検索を行なっているため，それぞれの検索方法においてデータ件数が増えたときの応答性能を確認した. Step1 と Step3 はユーザが入力，選択する部分でありデータ件数には依存しないため割愛する．また，特徴ベクトルによる関連文書検索は連続で複数回行うことができるが本実験ではどちらも一回の検索を基準として 10 回ずつ行い，検索時間を計測した.

計測結果は表 4 のようになった．この結果から，データ件数を増加させても検索にかかる平均時間は大きく変化することなく，また検索時間のばらつきも小さかった. なお，ラベルによる関連文書検索が特徴ベクトルによる関連文書

表 4　応答性能の結果
Table 4　Response performance.
(a) ラベルによる関連文書検索

データ件数[件]	2500	5000	7500	10000	13161
平均[msec]	49.61	46.49	45.48	48.50	57.64
標準偏差[msec]	8.58	6.64	6.15	7.44	15.88

(b) 特徴ベクトルによる関連文書検索

データ件数[件]	2500	5000	7500	10000	13161
平均[msec]	37.43	47.98	42.76	35.86	50.53
標準偏差[msec]	5.26	13.41	12.29	6.03	6.80

検索よりも時間がかかっているのは検索結果の出力数がラベル検索は 50 件，ベクトル検索は 20 件であることに起因している．

この結果からデータ件数を増やしても検索にかかる時間が大幅に増加することはないと予想される．

7. むすび

本論文では IoC 情報などをクエリとした全文検索と，トピックモデルによる潜在意味解析や外れ値検出アルゴリズムを行なって付与したラベルや特徴ベクトルを用いて関連文書検索を行うセキュリティレポート検索システムを提案した．既存の汎用的な検索エンジンでは入力を IoC 情報だけにした場合，所望の情報を人が選択的に抽出することとなり情報取得まで時間がかかる．提案システムを用いることで，情報取得までの時間が短縮されることを確認するためにユーザ実験を行った. 結果として Google 検索よりも提案システムを用いた方が検索時間の短縮や正解数の増加が見られ，単純なキーワード検索では到達できない，あるいはキーワードを選びながら検索を何度も繰り返して到達できるような文書に容易に到達できることを確認した．

本システムでは使用する各種アルゴリズムは教師なし学習のものを適用しているため，訓練データの準備は不要という意味でシステム構築の負担は大きくない．また，データセットの入力からラベルと特徴ベクトルの生成まで教師無し学習で自動実行可能としており，純増するセキュリティレポートに対して適当なタイミングでラベルと特徴ベクトルの更新を行うことが可能なシステムである．

提案システムを拡張し，ダークネットやハニーポットの観測網で捕捉された IoC 情報から，例えば使用された脆弱性情報を自動的に出力するエンジンとして構築し，サイバー攻撃ハイブリッド分析プラットフォーム[38]に結合する予定である．本論文の提案システムはセキュリティ管理者の操作を想定して構築しているので，人が介在せずに望ましいセキュリティ情報を出力する方法を検討することが今後の課題である．

謝辞　本研究は総務省の「電波資源拡大のための研究開発(JPJ000254)」における委託研究「電波の有効利用のための IoT マルウェア無害化／無機能化技術等に関する研究開発」によって実施した成果を含みます．

参考文献

[1] Exploit Database: Exploit Database, Exploit Database(online), available from<https://www.exploit-db.com/.> (accessed 2021-04-08)
[2] VirusTotal: VirusTotal, VirusTotal(online), available from<https://www.virustotal.com/gui/.> (accessed 2021-04-08)
[3] NIST: National Vulnerability Database, National Vulnerability Database(online), available from<https://nvd.nist.gov/.>
[4] MITRE: Common Vulnerabilities and Exposures, Common Vulnerabilities and Exposures(online), available from<https://cve.mitre.org/.> (accessed 2021-04-08)
[5] Guo, Y., Korhonen, A., Liakata, M., et al.: Identifying the Information Structure of Scientific Abstracts: An Investigation of Three Different Schemes, Proc. *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, pp.99-107, (2010).
[6] Contractor, D., Guo, Y. and Korhonen, A.: Using Argumentative Zones for Extractive Summarization of Scientific Articles, Proc. *Proceedings of COLING 2012*, pp.663-678, (2012).
[7] Kafkas, Ş., Pi, X., Marinos, N., et al.: Section level search functionality in Europe PMC, *Journal of biomedical semantics*, Vol. 6, No.1, p.7, (2015).
[8] Sugiyama, K. and M.Y. Kan: A comprehensive evaluation of scholarly paper recommendation using potential citation papers, *International Journal on Digital Libraries*, Vol. 16, No. 2, pp. 91-109, (2015).
[9] Tang, J., Qu, M., Wang, M., et al.; LINE: Large-scale Information Network Embedding, Proc. *Proceedings of the 24th International Conference on World Wide Web*(WWW '15), pp.1067-1077, (2015).
[10] Mikolov, T., Sutskever, I., Chen, K., et al.: Distributed representations of words and phrases and their compositionality, Proc. *Proceedings of the 26th International Conference on Neural Information Processing Systems*(NIPS'13), vol.2, pp.3111-3119, (2013).
[11] Teufel, S., Siddharthan, A. and Tidhar, D.: Automatic classification of citation function, Proc. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing* (EMNLP '06), pp.103-110, (2006).
[12] Achakulvisut, T., Acuna, D.E., Ruangrong, T., et al.: Science Concierge: A fast content-based recommendation system for scientific publications," PLoS One 11(online), DOI: 10.1371/journal.pone.0158423 (2016).
[13] Blei, D.M., Ng, A.Y. and Jordan, M.I.: Latent Dirichlet Allocation, *Journal of Machine Learning Research*, pp.993-1022, (2003).
[14] Bridges, R.A., Jones, C.L., Iannacone, M.D., et al.: Automatic Labeling for Entity Extraction in Cyber Security, arXiv preprint, available from<https://arxiv.org/abs/1308.4941.> (accessed 2021-04-08) (2013).
[15] Mulwad, V., Li, W., Joshi, A., et al.: Extracting Information about Security Vulnerabilities from Web Text, Proc. *Proceedings of the 2011 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, vol.3, pp. 257-260, (2011).
[16] McNeil, N., Bridges, R.A., Iannacone, M.D., et al.: PACE: Pattern accurate computationally efficient bootstrapping for timely discovery of cyber-security concepts, *2013 12th International Conference on Machine Learning and Applications*(ICMLA), vol.2, pp.60-65, (2013).

[17] Joshi, A., Lal, R., Finin, T., et al.: Extracting cybersecurity related linked data from text, *2013 IEEE 7th International Conference on Semantic Computing*, pp.252–259, (2013).
[18] 長田 侑樹，瀧田 愼，古本 啓祐，白石 善明，高橋 健志，毛利 公美，高野 泰洋，森井 昌克：セキュリティ情報検索のためのトピックモデルによるマルチラベリング，暗号と情報セキュリティシンポジウム(SCIS2021)，(2021).
[19] elastic: Elasticsearch, Elasticsearch(online), available from<https://www.elastic.co/jp/elasticsearch/.> (accessed 2021-04-08)
[20] elastic: Kibana, Kibana(online), available from<https://www.elastic.co/jp/kibana.> (accessed 2021-05-08)
[21] Trend Micro: Simply Security News, Views and Opinions from Trend Micro, Trend Micro(online), available from<https://blog.trendmicro.com/.> (accessed 2021-05-08)
[22] Cisco: Cisco Talos Intelligence Group - Comprehensive Threat Intelligence, Cisco(online), available from<https://talosintelligence.com/.> (accessed 2021-05-08)
[23] Broadcom: Symantec Blogs, Broadcom(online), available from <https://www.symantec.com/blogs/.> (accessed 2021-05-08)
[24] Barracuda: Journey Notes, Barracuda(online), available from<https://blog.barracuda.com/.> (accessed 2021-05-08)
[25] Druva: Tech/Engineering Archives, Druva(online), available from<https://www.druva.com/category/tech-engineering/.> (accessed 2021-05-08)
[26] FireEye: Threat Research, FireEye(online), available from<https://www.fireeye.com/blog/threat-research.html.> (accessed 2021-05-08)
[27] NETSCOUT: NETSCOUT's ASERT Blog, NETSCOUT(online), available from<https://www.netscout.com/.> (accessed 2021-05-08)
[28] Palo Alto Networks: Unit 42 - Latest Cyber Security Research, Palo Alto Networks(online), available from<https://unit42.paloaltonetworks.jp/.> (accessed 2021-05-08)
[29] CrowdStrike: Endpoint Protection Blog from Crowdstrike, CrowdStrike(Online), available from<https://www.crowdstrike.com/blog/category/endpoint-protection/.> (accessed 2021-05-08)
[30] ESET: WeLiveSecurity, ESET(Online), available from<https://www.welivesecurity.com/.> (accessed 2021-05-08)
[31] Darktrace: Darktrace Blog, Darktrace(Online), available from<https://www.darktrace.com/en/blog/.> (accessed 2021-05-08)
[32] SecPod: SecPod Research Blog, SecPod(Online), available from<https://www.secpod.com/blog/.> (accessed 2021-05-08)
[33] Check Point: Check Point Software - Blog, Check Point(Online), available from<https://blog.checkpoint.com/.> (accessed 2021-05-08)
[34] Fortinet: Threat Research, Fortinet(Online), available from<https://www.fortinet.com/blog/threat-research.> (accessed 2021-05-08)
[35] Netlab Blog: Network Security Research Lab at 360, Netlab Blog(Online), available from<https://blog.netlab.360.com/.> (accessed 2021-05-08)
[36] Secureworks: Cybersecurity and Compliance Resources, Secureworks(Online), available from<https://www.secureworks.com/services/security-consulting.> (accessed 2021-05-08)
[37] JPCERT: JPCERT/CC Eyes, JPCERT(Online), available from<https://blogs.jpcert.or.jp/en/.> (accessed 2021-05-08)
[38] Takahashi, T., Umemura, Y., Han, C., et al.: Designing Comprehensive Cyber Threat Analysis Platform: Can We Orchestrate Analysis Engines?, Proc. of *IEEE International Conference on Pervasive Computing and Communications* (PerCom), (2021).