7X-03

事前学習言語モデルによる関連文書検索
～コンピュータセキュリティシンポジウム論文ナビゲーションシステム～

冨田 裕涼† 毛利 公美† 白石 善明‡

岐阜大学工学部電気電子・情報工学科† 神戸大学大学院工学研究科‡

1.はじめに

文書集合から取り出したい文書に含まれている用語が既知の場合の検索は容易であるが，適当な検索ワードを想起できないときには検索は困難となる．検索結果として期待する文書が明らかでない場合に望ましい結果が出たと利用者が思う検索の実現を目的とし，本稿ではルールベースの検索を初期操作とする制限を設け，その検索結果に対する関連文書を利用者に提示する検索システムを提案する．

2.文書検索

本稿での文書を検索するフローを図1に示す．利用者は初めにキーワードをクエリとした全文検索を行い，検索結果を得る．そこから入手したい文書に近しい文書をもとに関連文書検索を行い，その検索結果より入手したい文書を取り出す．このような検索を提供するシステムの要件は以下のよう定める．

要件1：全文検索ができる

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．ここで提示される文書が要件2のクエリとなる．

要件2：文書をクエリとした関連文書検索ができる

文書の関連性に基づいて，文書を提示する．

要件3：文書を検索可能な形式に加工し保存することができる

全文検索に必要な索引作成，関連文書検索に必要な特徴ベクトル計算ができることが求められる．

要件4：オリジナル文書の閲覧ができる

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が可能なことが求められる．

3.関連文書検索システム

提案する文書検索システムの構成を図2に示す．システムの主な機能としては“全文検索機能”，“特徴ベクトル取得機能”，“関連文書検索機能”，“文書の加工・登録機能”，“特徴ベクトル計算機能”，“索引作成・保存機能”，“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：検索窓に入力されたキーワードをもとに全文検索を行う．事前に作成された索引を活用する．【要件1に対応】

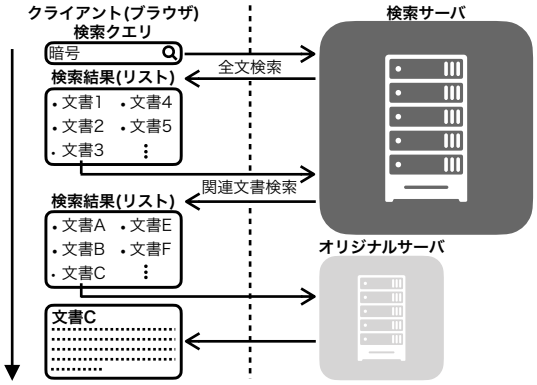

図1 検索のフロー

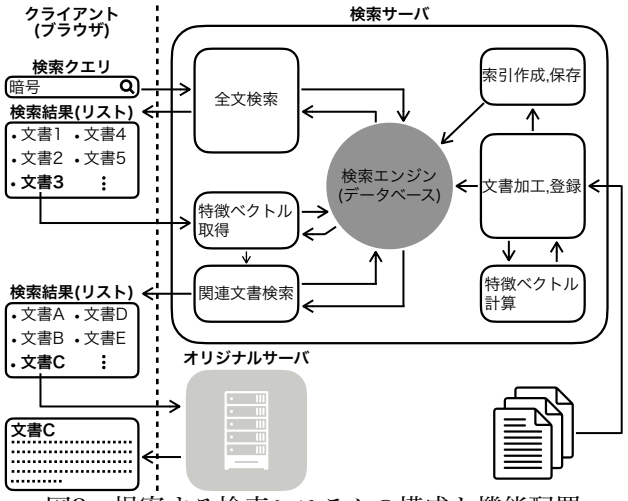

図2 提案する検索システムの構成と機能配置

特徴ベクトル取得：クエリとして入力された文書に対応する特徴ベクトルを取得する．【要件2に対応】

関連文書検索：クエリとなる文書に対応する特徴ベクトルをもとに関連文書検索を行う．【要件2に対応】

文書の加工・登録：検索対象の文書に特徴ベクトル，文書の所有元情報(URL)を付与し検索エンジンに登録する．【要件3に対応】

特徴ベクトル計算：“文書の加工・登録”で必要となる特徴ベクトルを計算する．【要件3に対応】

索引作成・保存：文書に含まれる用語に対して索引を作成する．全文検索で利用する．【要件3に対応】

オリジナル文書閲覧：検索結果で提示される文書リストに含まれるリンクからオリジナル文書を保有するサイトにアクセスする．【要件4に対応】

Related Paper Search by Pretrained Language Model -Paper Navigation System for Computer Security Symposium-

† Yusuke TOMITA, Masami MOHRI・Gifu University

‡ Yoshiaki SHIRAISHI ・ Kobe University

4.実装

クライアントサイド（ブラウザ）はHTMLで記述する．図3は検索エンジン（データベース），索引作成にElasticsearchを利用し，サーバーサイドはPythonのFlaskフレームワーク，検索対象の文書の登録にはPython，特徴ベクトル計算にはbert-as-service[1]，　事前学習モデルはSciBERT[2]を利用した際の実装例である．

2009年から2019年のコンピュータセキュリティシンポジウムの論文（2010年を除く）を検索対象の文書とし，情報処理学会電子図書館から入手した論文題目及び概要を用いて特徴ベクトルを計算した．

クライアントサイド（ブラウザ）での検索画面及び，検索結果の表示例を図4に示す．図4の検索画面上部の検索窓にキーワードを入力し全文検索を行う．検索結果は検索窓の下に表示される．検索結果の文書一つひとつがリスト形式で表示され題目をクリックするとオリジナル文書にアクセスできる，“この論文に関連した論文を探す”ボタンをクリックすると文書をクエリとした関連文書検索を行う．

5.評価

キーワードをクエリとした単純な全文検索は単語の出現頻度を表す“Bag of Words”を特徴ベクトルとして用いる関連文書検索であると見なすことができる．そこで，“Bag of Wordsを特徴ベクトルとした関連文書検索(≒全文検索)”と”SciBERTを特徴ベクトルとした関連文書検索”について文書の関連性についてクラスタリングを用いて比較した．図5はSciBERTを特徴ベクトルとした際のクラスタリングの結果を可視化したものである．クラスタリングにはコサイン類似度を距離指標としたK-meansを用いた．K=26のクラスタのうちの10個のクラスタに特定の分野の文書が集中していることを確認した．それに対しBag of Wordsを特徴ベクトルとした際のクラスタリングでは特定の分野の文書が集まったクラスタを見出すことはできなかった．本稿の検索システムが入手したい文書に含まれている用語が未知の場合でも，特定の分野においては特徴ベクトルの近さに基づき望ましい検索結果を提示することが可能であることを示している．

6.まとめ

本稿では検索結果として期待する文書が明らかでない場合に望ましい結果が出たと利用者が思う検索の実現を目的とした検索システムを提案した．ルールベースの検索を初期操作とする制限を設けているが，その検索結果に対する関連文書を利用者に提示することで，検索クエリを想起できない場合にも望ましい検索結果を提示することができた．可視化により定性的に関連文書検索の有用性を述べたが，異なる特徴ベクトルの作成方法の検討と定量的な評価を今後の課題とする．

なお，本検索システムはコンピュータセキュリティシンポジウム2020のポータルサイトにて，2020年の論文のみを対象とした論文ナビゲーションシステムとして運用した．

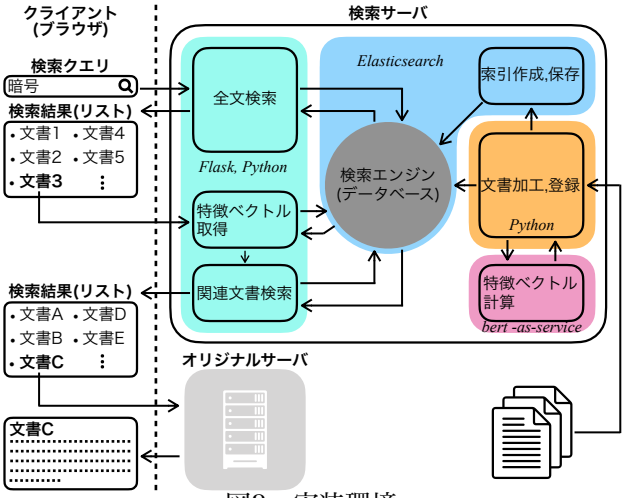

図3　実装環境

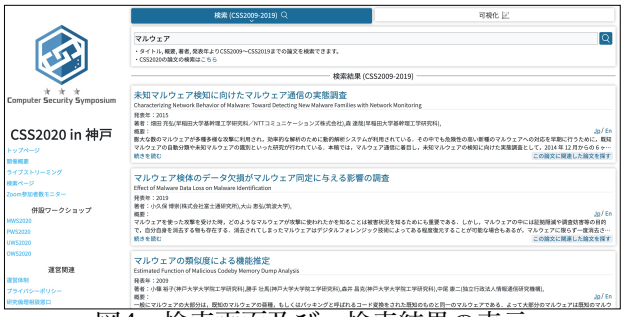

図4　検索画面及び，検索結果の表示

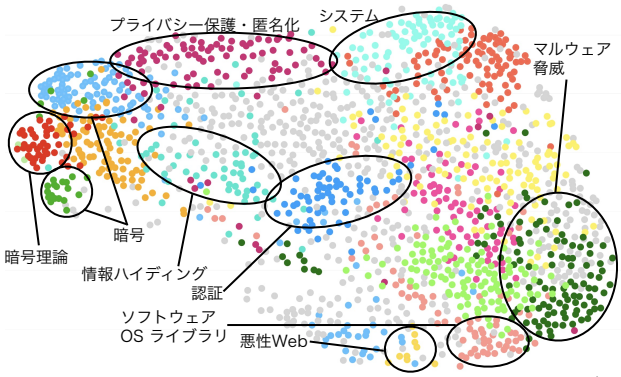

図5 SciBERTによる特徴ベクトルのクラスタリング

謝辞　本研究の一部は国立研究開発法人情報通信研究機構の委託研究「機械学習に基づくサイバー攻撃情報分析基盤技術の研究開発」およびJSPS　科研費JP19K11963，JP18K04133　の支援のもとに行われた．

参考文献

[1] https://github.com/hanxiao/bert-as-service
[2] Iz Beltagy, Kyle Lo, Arman Cohan, SCIBERT: A Pretrained Language Model for Scientific Text, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp.3615–3620, 2019.