個人認証を見据えた検索クエリの類似性評価

宮野 祐輔[1]　山口 利恵[1]　坪内 孝太[2]　五味 秀仁[2]

概要：検索履歴にはユーザの属性情報が多く含まれており，この情報を用いて個人を認証できるのではないかと考えられている．特に検索クエリには多くの属性情報が含まれていることがプライバシー保護の観点から分析した先行研究によって明らかになっている．しかしそういった属性情報が，認証に適した永続性や唯一性などを有しているかは疑問が残る．そこで本研究では検索クエリの類似度をトピックモデルやベクトル空間モデルなどに基づいて算出し，本人や他者とのクエリ類似度を実データを用いて調査した．またその結果得られた類似度から，検索クエリがどの程度本人を同定する，あるいは他人と区別するための適性を保持しているかを考察した．

キーワード：ライフスタイル認証，多要素認証，行動認証，検索履歴，検索クエリ

Evaluation of Search Query Similarity for Personal Authentication

Yusuke Miyano[1]　Rie Shigetomi Yamaguchi[1]　Kota TSUBOUCHI[2]　Hidehito GOMI[2]

Abstract: It is known from previous research on privacy in search histories that search queries contain private information. This information thought to be useful for authentication, but it is not clear whether search queries have authentication aptitudes such as uniqueness and permanence. Therefore, this paper analyzed search query similarity based on various models such as topic models and vector space models. Furthermore, we evaluated how suitable each similarity index is for authentication.

Keywords: Lifestyle Authentication, Multi-factor Authentication, Active Authentication, Search History, Search Query

1. はじめに

検索履歴にはユーザの属性情報が多く含まれており，多くの先行研究によってプライバシー保護の必要が訴えられてきた．検索行動は電子メールや SNS への投稿などと異なり公開することを前提としないため，ユーザの行動特性や属性がより露骨な形で表れるであろうと想定される．そのため検索履歴の分析がより強いプライバシーの侵害に繋がる可能性は高く，過去には実際に公開された仮名処理済みの検索履歴から個人特定に至った例も報告されている．

一方で個人特定に結びつくユーザの属性情報が含まれていることを利用し，検索履歴を用いた認証ができるのではないかと考えられている．検索履歴にはユーザの行動特性を示す時間や検索内容などの情報が多く含まれていることから，ユーザの文脈情報から本人性を判断して認証する行動認証のひとつであるとみなせる．先行研究では検索クエリの入力パターンや検索回数・時間などから得られた特徴量が認証に利用できる可能性があることが明らかになっているが [1]，検索クエリの内容に基づいた特徴量の設計・評価については触れていない．その理由として，検索クエリの意味内容に対する自然言語処理的アプローチが意図推定や検索連動型広告などの観点からのものが多く，ユーザの本人性がどういった形で現れるのかといった側面からのアプローチは不十分である点が挙げられる．さらに検索クエリは一般的な文章と大きく異なる文法によって記述されることが多く，既存の手法がどの程度認証に利用するための

[1] 東京大学大学院 情報理工学系研究科
Graduate School of Information Science and Technology, The University of Tokyo
[2] ヤフー株式会社
Yahoo Japan Corporation

分析として有効かは明らかになっていない．

したがって本研究では，自然言語処理で広く用いられているトピックモデルやベクトル空間モデルを実際の検索履歴データに適用し，クエリの類似度を算出した．またその結果得られた類似度を比較することで，各モデルに基づく類似度がどのような特性を持ち，認証に利用する際にはどの程度本人のクエリであると同定できるか，あるいは他人のクエリとどの程度区別できるのかについて評価した．

その結果，クエリ類似度はユーザの検索対象をより抽象的にトピックレベルで表現することができるので，行動認証の弱点の一つであるユーザ行動における規則性・法則性のゆらぎに対しても頑強であることが分かった．

1.1 構成

本稿の構成について説明する．2 章では多要素認証について説明し，そのなかで検索履歴を用いた認証がどのような役割を果たしていくのかについて述べる．3 章では検索履歴を用いた認証に活用する文書の著者やトピックを推定する手法や，履歴情報を用いた認証技術などについて説明する．4 章では，今回行った実験の詳細について述べる．5 章では実験から得られた結果を元に検索クエリの特性および類似度の算出手法について考察し，認証技術に活用するための適性について検討する．6 章で全体をまとめて結論付ける．

2. 多要素認証と検索履歴

本章では認証方式の 1 つである多要素認証について説明する．多要素認証には多様な認証技術を組み合わせるのが安全性の観点から推奨されているが，検索履歴を利用した認証は行動認証の一形態として多要素認証の精度や利便性を高めるのに役立つと考えられる．しかし検索履歴は同一ユーザのものでも時々に応じて大きく変化し，既存の手法と比べてもそのゆらぎは大きいために認証の過程で緩和・吸収する必要がある．そのため検索履歴を用いた認証に利用する特徴量はそういったゆらぎに対応できるものでなければならない．以下では各認証方式やそれぞれの特性の詳細について述べる．

2.1 多要素認証

多要素認証とは，複数の認証方式を組み合わせることで，認証システムの安全性や利便性を向上させる認証方式のことである [2]．単一の認証方式のみを用いた認証システムの場合，そのシステムの安全性は当該方式の特性に大きく依存する．現在広く用いられているパスワードによる認証はパスワードの漏洩や安易なパスワードの設定などのリスクがあり，単体でのセキュリティ強度には課題が残る．

一方多要素認証では複数の認証技術を用いているため，一部の認証方式に脅威が生まれたり，適切に動作しなくなったりした場合でもシステム全体として認証の安全性を維持することができる．現状広く利用されている多要素認証の形態として，認証する際にあらかじめ登録されたスマートフォンなどの端末にワンタイムパスワードを送信し，その入力をもって認証を完了するという帯域外トークンを用いるものが挙げられる [3]．

しかし一方で多要素認証の弱点として，精度を上げるために要素を増やせば増やすほど個別の認証に要するコストが増大するという問題がある．セキュリティコストの上昇は結果としてユーザがセキュリティのレベルを下げ，全体の安全性が低下する要因になりうる．したがって多要素認証の要素技術として，ユーザへの負担が小さい行動認証などの要素を検討する必要がある．

2.2 行動認証

行動認証とはユーザの行動特性を用いた認証手法のことである．代表例としては物理的な行動に基づく歩容認証，ハードウェアを利用する際の特徴に基づくキーストローク認証，そしてソフトウェアの利用履歴にそして Web 閲覧履歴認証などが挙げられる．近年ではスマートフォンや活動量計などのモバイルデバイスが広く普及したため，それらの端末から得られた活動量 [4] や Wi-Fi アクセスポイント情報 [5] を用いた認証も提案されている．行動認証はユーザの行動特性を利用するため，パスワードを記憶したりデバイスを保持したりする必要がない．そのためユーザに対する負荷が小さく，多要素認証の要素技術として近年研究が盛んに行われている．

2.2.1 リスクベース認証

行動認証の一形態として，ユーザの行動特性や利用環境などからリスク評価を行い，一定以上のリスクが想定される場合に追加的な認証を行う**リスクベース認証**が普及している．リスクベース認証は通常通りにサービスを利用している場合には認証プロセスが発生しないため，認証を要求される回数が増えてしまうという多要素認証の弱点を補いうると期待されている．

実際に運用されているリスクベース認証の多くは端末情報や IP アドレス，位置情報やブラウザ情報などを基準にしているものが多い．すでに認証によって確かにユーザ本人からのアクセスであるということが確認されたアクセスのログをデータベースに管理しておき，そのログデータからユーザのアクセス環境が取り得る変動を分析して求めておく．そして新たにアクセスのリクエストがあった場合，そのリクエストに含まれる属性情報からユーザの現在のアクセス環境を特徴として抽出し，事前に求めたユーザのアクセス環境変動と比較する．ユーザのアクセス環境が登録されている，あるいは許容可能な変動幅以内であれば認証状態を維持し，一定の許容幅を超え著しく異なるアクセス環境からのリクエストに対しては新たに別要素での認証

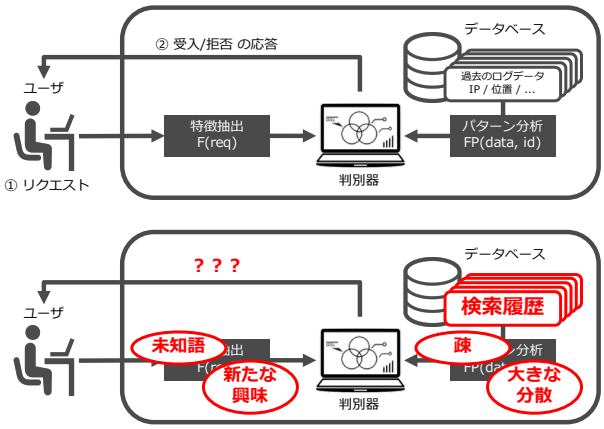

図 1 リスクベース認証の構成と検索履歴への適用

Fig. 1 Configuration of risk-based authentication and its application to search history.

（前述した帯域外トークンなどによる認証）を求めるというものである．別要素での認証が成功した場合にはそのログデータもデータベースに保管され，次回以降のリクエストでは正常にリクエストを受け付け，追加的な認証を求めることはしない．また近年ではキーストローク認証やマウス操作パターン認証を取り入れたリスクベース認証モデルも提案されている [6]．

2.2.2 ライフスタイル認証

検索履歴は本来認証に用いるために収集されたデータではなく，ユーザの日々の行動を記録したものである．したがって検索履歴を用いた認証を実用のサービスとして導入した場合，ユーザが認証目的で新たにデータを生成する必要はなく，通常通りの生活を送っていれば認証に必要なデータが揃うこととなる．このようにユーザが明示的に認証のための動作を行ったりデータを生成したりする必要なく，日々の生活パターンから認証を行う認証方式のことをライフスタイル認証という [7]．ライフスタイル認証は多要素認証における認証コストの増大という問題を緩和する効果が期待できる．

2.3 検索履歴を用いた認証

検索履歴を用いた認証は形態としては行動認証にあたるため，リスクベース認証への適用が考えられる．リスクベース認証として捉える場合，事前の検索履歴から学習したユーザモデルとの類似度を算出し，ユーザが行った検索行動のリスクとして評価する必要がある．しかし現在一般的なリスクベース認証と検索履歴を用いた認証とで大きく異なる点のひとつとして，時間経過に伴うゆらぎが大きい点が挙げられる．ログインする端末に基づくリスクベース認証であれば，今まで利用したことのない端末から操作する機会というのは毎日のように行う検索ほどは多くないため，新たな端末を利用する度に追加的に登録するという操作が大きな負荷とはならない．一方で検索行動というものは時間経過や外的刺激によって大きく変化するので，どの程度のゆらぎであれば許容するないしは拒絶するのかという判断が困難である．

本研究におけるゆらぎを吸収するためには，以下の 3 手法が考えられる．

（1） テンプレート更新によるゆらぎ吸収 [8]

（2） 0/1 でない，類似度によるゆらぎ吸収

（3） 多要素の組合せによるゆらぎ吸収

本稿では特に (2) に対応する類似度の算出手法について述べる．検索履歴における特徴語の一致など，特定の単語が含まれるか否かという 0/1 の評価基準では，対象とする語群があまりに疎で検索行動のゆらぎを十分に捉えられないと考えられる．たとえば「東北　旅行」と検索していたユーザが数日後に「秋田　温泉」と検索していた場合，人の目から見ればユーザが旅行の目的地を絞り込んだのだといったように判断できるものの，単語の一致は存在しないため機械的な判別では全くの別ユーザであるというように判断されかねない．

同様の問題は既存のリスクベース認証における位置情報の利用においても存在する．リスクベース認証に位置情報を用いる場合，GPS に基づく緯度経度などのように厳密な値ではなく，都道府県や市区町村単位でデータを取り扱うことが多い．これはユーザの位置というものが一定程度のゆらぎを持つ情報であり，そのゆらぎを吸収するために位置情報の粒度を荒くすることでゆらぎに対応する必要があるからである．

本研究ではこの位置情報の抽象化に倣い，検索クエリに現れる個別の単語ではなく検索クエリ自体が持つ検索意図・潜在的な興味関心に基づいて本人かどうか判別する．上述した例であれば「東北　旅行」と「秋田　温泉」がいずれも「(国内) 旅行」「レジャー」「観光」などといったより大きなジャンルでくくられることで，同一のユーザであるかどうかを判別しやすくするといったものである．

また検索行動自体はなりすましが容易なため，特徴語を一度だけ検索しただけでも本人と判定してしまうおそれがある．こういった状況でも検索全体の検索対象，ジャンル，傾向等を特徴量として捉えることができれば特徴語を推測してなりすます攻撃にも対応が可能である．

2.3.1 多要素認証における検索履歴の優位性

検索履歴を用いた行動認証は以下の様な優位点が存在する．

すでに身近に広く普及している

行動認証に利用するために新たなセンサやそれを備えたデバイスを開発しても，それが広くユーザに利用されなくては肝心のデータを収集することは難しい．し

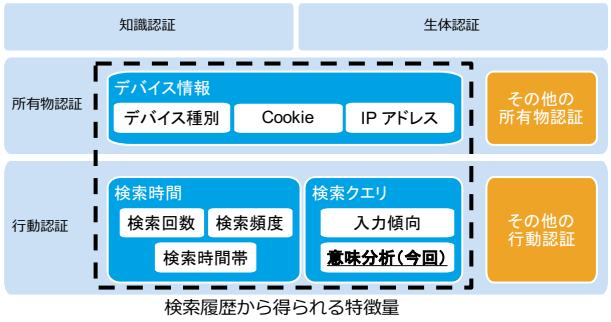

図 2 検索履歴の多要素認証における役割

Fig. 2 A role of authentication using search history in a multi-factor authentication system.

かし検索はすでに様々なサービスのなかで用いられており，ユーザが行動認証のためにわざわざ改めて導入する必要がなく負荷が小さい．

多様なデバイスで収集できる

検索はコンピュータやスマートフォン，タブレットなど，様々なデバイスで利用されている．これにより特定のデバイスを保持しなければならないという制約が緩和され，ユーザの行動をよりシームレスに追跡することができる．

分析のノウハウが多数存在する

今まで利用されてこなかったセンサ情報や履歴データを分析する場合，分析に用いる特徴量や手法を一から検討する必要がある．一方で検索クエリはクエリ推薦 [9] や検索の意図推定 [10] など，認証以外の分野で広く研究されており，分析のノウハウが蓄積されている．

ユーザの本性を強く反映した情報

検索行動は SNS での投稿などとは異なり，基本的に秘匿される内容である．そのためユーザ本来の興味が反映される可能性が高く，より強く本人性を表すデータであると考えられる．

多様な属性情報が含まれている

検索履歴にはユーザの使っているデバイスの種別，Cookie，IP アドレスなどのように所有物の属性を示す情報のほか，検索した時間，ならびに検索クエリなど多種の属性情報が含まれている．これらの情報は組み合わせることで全体の認証制度を高めたり，特定の情報源に対する攻撃への耐性を強めたりすることができる．

3. 関連研究

本章では文書から本人性を明らかにする技術である著者同定，および 4 章で行う実験で用いる自然言語処理モデルについて説明する．

3.1 著者同定

文書の読点の位置や語彙の偏りなどから，その文書の著者を特定することを**著者同定**という．著者同定は旧来より著者不明の文献などを対象として研究されてきたが，近年ではサイバー犯罪対策の観点から，電子メッセージのように短いテキストからその入力者を特定する研究が行われるようになった [11]．研究対象の電子メッセージは，メールや SNS への投稿など，以前からの対象であった文献などに比べて非常に短い．そのため絵文字など，電子メッセージが持つ固有の特徴量を活用して精度を高める手法が盛んに提唱されている．そういった電子メッセージのなかでも検索クエリは特別短く，かつ通常の文書とは形式が大きく異なるため，既存研究と比較しても検索クエリの著者同定はより困難であると考えられる．

3.2 ベクトル空間モデル [12]

ベクトル空間モデルとは文書を何らかの形でベクトル表現に置き換え，それらの類似度や距離を計算することで文書間の関係性を表現するモデルである．ベクトル空間モデルは情報検索の分野で幅広く用いられており，その手法や類似度の算出手法は多岐に渡る．以下では本稿の実験に用いた 2 種類のベクトル空間モデルについてそれぞれ述べる．

3.2.1 潜在的ディリクレ配分法 [13]

潜在的ディリクレ配分法 (LDA; Latent Dirichlet Allocation) とは，自然言語処理におけるトピックモデルの 1 つである．トピックモデルは文章が生成される過程でその背後に潜在的なトピック（話題）が存在し，そのトピックの分布に応じて単語や文書が生成されるという考えに基づいている．LDA は 1 つの文書に対して複数のトピックが存在すると想定した確率的モデルである．LDA を用いることで個別の単語のゆらぎではなくあるユーザが検索対象としているトピックの割合や傾向を分析の対象とすることができるので，検索対象の分散や遷移が激しい検索履歴に対しても有効に機能することが期待される．

3.2.2 word2vec [14–16]

Tomas Mikolov らによって提案された word2vec とは，単語を固定長の実数ベクトルで表現するためのアルゴリズムである．このように単語をベクトルで表現することは，**分散表現**ないしは**単語埋め込み**と呼ばれる．

word2vec によって生成された単語ベクトルは性別や首都，比較級の表現などを表しうることが先行研究によって示されており，「$king - man + woman = queen$」などのように加減算にも対応しているなど有用な性質を保持していることが知られている．

また，単語単位で分散表現を行う word2vec を文書単位で分散表現できるよう拡張した doc2vec [17] も発表されて

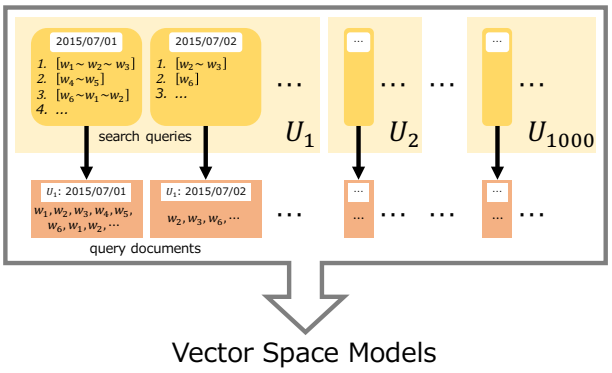

図 3 検索履歴を利用したベクトル空間モデルの構築
Fig. 3 Construction of vector space models using search history.

いる．

4. 実験

本章では，今回行った実験の詳細について述べる．

4.1 データセット

本節では実験に用いたデータセットの詳細について述べる．

本研究では， Yahoo! JAPAN[*1] で検索された検索履歴データを用いて実験を行った．対象として Yahoo! JAPAN ID（以降，YID と省略する）でログインしているユーザのうち，2015 年 7 月から 9 月までの 3 ヶ月間において毎日 1 件以上検索したユーザの検索履歴を収集した．実験に用いるにあたってユーザを 1000 人ランダムにサンプリングし，期間は 7 月 1 日から 7 月 26 日までの約 1 ヶ月間のデータを使用した．今回取得した検索履歴データには仮名化された YID，検索に用いたデバイスの Cookie および種別，検索日時，検索クエリが含まれている．

なお Cookie と YID は 1 対 1 に対応しておらず，複数の端末から同じ YID によって検索されたり，逆に同一の端末から複数の YID を用いて検索されたりすることがある．本研究では端末を複数所持するユーザや，目的に応じて YID を使い分けるユーザを考慮して，YID をユーザ単位として使用した．

4.2 実験に使用したデータおよび手法

実験に使用するデータとして，各ユーザの 1 日分の検索クエリから名詞を抽出しクエリ文書を作成した．このとき実験に使用する doc2vec が単語の位置に基づいて学習するモデルであるため，クエリ文書は Bag-of-Words 形式ではなく名詞の出現する順番を保持したリスト形式とした．

今回は 1000 人のユーザによる 26 日分の検索履歴をデータセットとして用いたため，クエリ文書の総数は 26000 と

[*1] http://www.yahoo.co.jp

表 1 最類似クエリ文書に基づくユーザ推定の精度．
Table 1 An accuracy of user estimation based on the most similar query document.

	第 1 候補	上位 3 候補	上位 5 候補
精度	44.5%	53.9%	58.2%

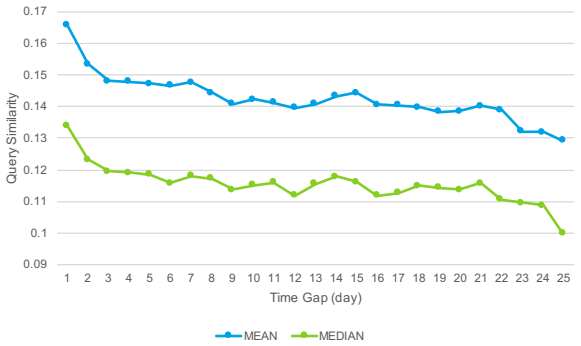

図 4 文書の日付間隔を変化させた場合のクエリ類似度の推移
Fig. 4 Secular change of query similarity.

なった．

本実験ではこれらの文書間で複数基準による類似度を算出し，ユーザ内・ユーザ間で条件を変化させた場合どのような値を取るのかを調査した．

形態素解析には Yahoo! JAPAN が提供している形態素解析ツールを用いた．本実験における処理では Python およびベクトル空間モデル用の Python ライブラリである gensim[*2] を用いた．

5. 結果および考察

本章では，実験で得られた結果を示す．

5.1 最類似クエリ文書とユーザ推定

gensim における doc2vec の機能として，ある文書が与えられた際に類似している単語ないしは文書を解析済み文書のなかから提示するものがある．本実験ではこの機能を用いて，各クエリ文書に最も類似している文書上位 n 件 $(n = 1, 3, 5)$ を提示させた．そして提示された候補文書のなかに元のクエリ文書と同一ユーザによるクエリ文書が存在するかどうかをカウントした．その結果を**表 1** に示す．

今回の実験では時系列を考慮していないため，厳密にはユーザの推定精度としてこの値が得られるわけではない．しかし比較対象となる 25999 件のクエリ文書のなかで最も類似しているクエリ文書が本人のものである確率が 44.5%という値を示したことから，多くのユーザが内容の類似している検索を行っていることが分かる．

5.2 時系列的変化

同一内容を定期的に検索するユーザもいるが，多くの

[*2] https://radimrehurek.com/gensim/

ユーザが検索する内容は時間の経過に伴って変化すると考えられる．したがって比較する文書の日付間隔が大きくなればなるほど，クエリ類似度は減衰していくことが予想される．**図 4** は文書の日付間隔を 1〜25 日で変化させた場合，クエリ類似度の平均値および中央値がどのように推移するかを示した図となっている．

図 4 から，日付間隔が空くと類似度の平均が僅かであるが減少傾向になることが分かる．また日付間隔が 7 の倍数のときに平均・中央値ともにやや類似度が上昇している傾向が見られるが，これは同じ曜日の検索クエリを比較していることが原因であると考えられる．この結果からはユーザの検索行動に弱い曜日周期性ないしは平日/休日の習慣性があることが分かる．

5.3 テンプレートとの比較

クエリ類似度を個人認証に転用するにあたって，バイオメトリクスで広く用いられているテンプレートを利用する形式が考えられる．すなわち事前に一定期間の履歴情報からユーザの特性を反映したテンプレートを作成した上で，そのテンプレートと認証リクエスト中のクエリ文書とのクエリ類似度を算出して認証の成否を判断するものである．認証の成否の判断基準として閾値を設定し，閾値を超える認証リクエストは受け入れ，下回る認証リクエストは拒否した．

本実験では 7 月 1 日〜7 日までの 1 週間分のクエリ文書をテンプレートとし，残り 19 日の各日のクエリ文書に対してクエリ類似度が閾値を上回るかどうかを確かめた．閾値は 0 から 0.5 までの間で 0.01 刻みで変化させた．このとき**図 5** は閾値を変化させたとき，テンプレートと異なるユーザを誤って受け入れてしまった他人受入率（FAR; False Acceptance Rate），およびテンプレートと同一ユーザであるにも関わらず誤って拒否してしまった本人拒否率（FRR; False Rejection Rate）をグラフとして表現したものである．

FAR と FRR は閾値の変化に対してトレードオフの関係にあるため，評価の指標として FAR と FRR のグラフ交点を調べ，そのエラー率を等価エラー率（EER; Equal Error Rate）と呼称して用いる．本実験における doc2vec と LDA の場合， EER はいずれも 24%前後であり双方に大きな差はない．

5.4 機械学習を用いたユーザ判別

先行研究では検索履歴から得られる特徴のなかに相関性を持ち，ユーザの個人特定に繋がるような特徴量がいくつかあることが明らかになっている [1]．それらの特徴量に加えて今回得られたクエリ類似度の特徴量を用いて認証した場合，本人と同定できる割合，ならびに各特徴量の寄与度をランダムフォレストを用いた 2 値判別を通して算出した．対象としたデータには，2 人 1 組のユーザ組を 1000 組作成した上で与えられたクエリ文書がどちらのユーザによるものかという値をフラグとして持たせた．今回使用した特徴量は [1] における**検索回数**，**単一クエリ率**，**文字数**，**単語数**，各時間（0〜23 時）に検索した回数を 24 次元のベクトルに変換した**時間ベクトル**，そして本稿で算出した**クエリ文書ベクトル**（doc2vec, LDA）の計 7 種類である．なお本実験ではテンプレートを用いた認証を前提としているため，いずれの特徴量も最初 7 日間の平均と比較した場合の割合・類似度を機械学習に利用している．

表 2 ランダムフォレストによるユーザ判別の正解率.
Table 2 Cross validation scores with random forest classifiers.

	傾向特徴のみ	+doc2vec	+LDA	+Both
正解率	75.5%	82.2%	82.9%	**85.5%**

ランダムフォレストの実装は scikit-learn[*3] に依った．判別器のパラメータはすべて規定値を使用した．5 分割交差検証を行った際のスコアを正解率とし，2 種類のクエリ類似度を特徴量として用いた場合と用いない場合の計 4 種類の結果を**表 2** に示す．

結果としてクエリ文書ベクトル 2 種類を用いたモデルが最も高い正解率を示した．

ランダムフォレストは特徴選択を学習の過程で行うため，それぞれの特徴量がどの程度判別に寄与したかを示す**寄与度**を算出することが出来る．その結果を**表 6** に示す．

2 種類のクエリ類似度の寄与度がほぼ同じ数値を示した．他の特徴量の寄与度に比べると大きく抜きん出ており，認証における検索内容の重要度が明らかになっている．

5.5 特徴量同士の比較

多要素認証では多種多様な特徴量を用いることが推奨されていることは既に述べたが，既存の研究で挙げられた特徴量ならびに今回新たに加えたクエリ類似度がそれぞれどの程度独立して機能しているのか，その相関性を実験により求めた．

5.4 節で用いた 7 種類の特徴量において，互いの相関係数（ピアソンの積率相関係数）を算出した．そのなかでも相関係数の絶対値が 0.5 を超え一定程度の相関がある特徴量の組を挙げると，(単語数，文字数) が 0.95，(単語数，単一クエリ率) が 0.58，(文字数，単一クエリ率) が 0.57 となった．これに次いで (doc2vec, LDA) が 0.47 という相関係数の値を取っている．

文字数，単語数，単一クエリ率はそれぞれ高い相関があり，ランダムフォレストにおける判別への寄与度も低いため，認証の要素としてこれら 3 種類を同時に用いる有用性は低いと考えられる．

また，2 つの手法に基づくクエリ類似度の相関は中程度

[*3] http://scikit-learn.org/

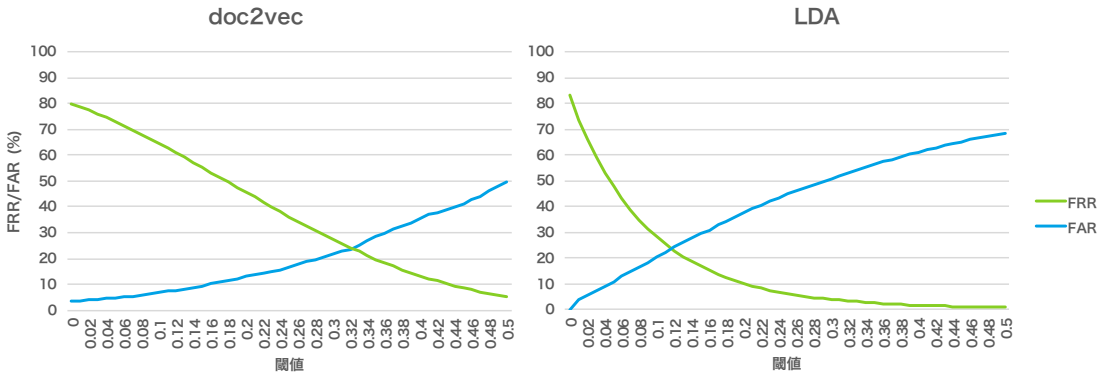

図 5 テンプレートとの類似度に対する閾値を変化させた場合の他人受入率/本人拒否率
Fig. 5 FAR/FRR of Authentication with query templates.

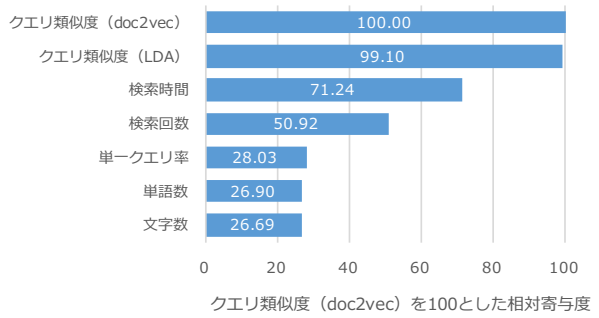

図 6 2 種類のクエリ類似度を用いたときののランダムフォレストにおける各特徴量の寄与度.
Fig. 6 Feature importances of random forest classifier with both features of query similarities.

であることが分かった．検索クエリの意味内容というほぼ同一の対象に関するベクトル表現であるにもかかわらず中程度の相関に留まったということは，両者が異なる観点から検索クエリを分析していることを表していると考えられる．

5.6 リスクベース認証との比較

本節では実験の結果からリスクベース認証と検索履歴を用いた認証を比較し，その共通点と差異についてまとめる．

本実験では検索履歴を用いた認証を，テンプレートとのクエリ類似度を利用して行った．クエリ類似度の指標として用いたコサイン類似度はコサイン距離という別の指標に変換することができ，その変換式は $CosineDistance = 1 - CosineSimilarity$ と表せる．一方リスクベース認証では事前に学習したユーザのテンプレートからどの程度逸脱しているかというリスク評価値を基準にして認証しているが，このリスク評価値はコサイン距離を使って表現できるため，検索履歴を使った認証はコサイン距離をリスク評価値に置き換えたリスクベース認証と見なすことができる．

一方で通常のリスクベース認証では具体例なアドレスやブラウザ情報を用いてリスク評価値を計算するものの，検索履歴を用いた認証ではクエリ類似度を算出するにあたって意味の抽象化というプロセスを経ている．これはユーザの検索行動が通常のリスクベース認証におけるユーザの行動に比べてより分散が大きく，意味を抽象化しないとユーザの典型的な行動パターンであるテンプレートを作成できないためである．

本実験では時間や行動回数など，行動認証やリスクベース認証で用いられることのある特徴量も使用したが，それらの寄与度より意味論から表現したクエリ類似度の寄与度のほうが大きくなった．このことからリスクベース認証として検索履歴を利用した認証に，抽象化した意味を表現する特徴量が有効であることが分かった．

6. 結論

本研究では検索履歴を用いた認証を多要素認証の一要素として活用することを目的として，検索クエリの類似度をどのように算出すべきか，またその類似度によって認証を行った場合どのような特性が表れるのかを実際の検索履歴を用いて分析した．

その結果，ユーザの検索内容に関しては，同一ユーザの検索クエリはよく似通っていることが分かった．またその内容は 1 ヶ月という期間であれば大きくゆらぐことはなく，時間間隔が空いてもその類似度は僅かずつしか減衰しないことが明らかになった．今後の課題として時間間隔を数ヶ月から数年に伸ばした上で検証を重ね，最適なテンプレートの更新間隔や検索行動の周期性について分析することが挙げられる．

また今回はクエリ類似度を用いた 2 種類の認証実験を行い，双方で一定の精度が得られることを確認した．クエリ類似度のみ用いて閾値以上か否かで判別する手法では doc2vec と LDA とで EER に大きな差は見られなかった．しかし互いの相関が中程度であったこと，ランダムフォレ

ストによる判別では 2 種類のクエリ類似度を用いたモデルが最も正解率が高かったことなどから，2 種類のクエリ類似度指標はそれぞれ違う観点から類似度を評価していることが分かった．そのため双方の手法はどちらかに優劣がつくものではなく，状況に応じて選択したり組合せたりすることで最大限の性能を発揮することが出来ると想定される．

以上の点から，クエリ類似度は検索履歴を用いた認証を多要素認証の一要素として組み込む際には有益な特徴量となりうることが判明した．今後の研究ではユーザの人数および実験期間を拡大することで，より一般的で確度の高い結果を導きたい．

謝辞 本研究は次世代個人認証技術講座（三菱 UFJ ニコス寄附講座）の助成を受けて実施された．

参考文献

[1] 宮野祐輔，山口利恵，坪内孝太，五味秀仁：個人認証を見据えた検索履歴からの行動分析，暗号と情報セキュリティシンポジウム 2016 (SCIS2016) (2016).

[2] 山口利恵，鈴木宏哉，小林良輔：認証精度の違う多要素・段階認証，コンピュータセキュリティシンポジウム 2015 (CSS2015) (2015).

[3] Google: Google 2 段階認証プロセス． `http://www.google.com/intl/ja/landing/2step/` 2016 年 8 月 9 日閲覧.

[4] Susuki, H. and Yamaguchi, R. S.: Cost-Effective Modeling for Authentication and Its Application to Activity Tracker, *Information Security Applications*, Springer, pp. 373–385 (2015).

[5] Kobayashi, R. and Yamaguchi, R.: A Behavior Authentication Method Using Wi-Fi BSSIDs around Smartphone Carried by a User, *2015 Third International Symposium on Computing and Networking (CANDAR)*, IEEE, pp. 463–469 (2015).

[6] Traore, I., Woungang, I., Obaidat, M. S., Nakkabi, Y. and Lai, I.: Combining mouse and keystroke dynamics biometrics for risk-based authentication in web environments, *Digital Home (ICDH), 2012 Fourth International Conference on*, IEEE, pp. 138–145 (2012).

[7] 小林良輔，疋田敏朗，鈴木宏哉，山口利恵：行動センシングログを元にしたライフスタイル認証の提案，コンピュータセキュリティシンポジウム 2016 (CSS2016) (2016).

[8] 小林良輔，疋田敏朗，鈴木宏哉，山口利恵：ライフスタイル認証におけるゆらぎ吸収を目的としたテンプレート更新手法の提案，コンピュータセキュリティシンポジウム 2016 (CSS2016) (2016).

[9] Baeza-Yates, R., Hurtado, C. and Mendoza, M.: Query Recommendation Using Query Logs in Search Engines, *Proceedings of the 2004 International Conference on Current Trends in Database Technology*, EDBT'04, pp. 588–596 (2004).

[10] Brenes, D. J., Gayo-Avello, D. and Pérez-González, K.: Survey and Evaluation of Query Intent Detection Methods, *Proceedings of the 2009 Workshop on Web Search Click Data*, WSCD '09, pp. 1–7 (2009).

[11] Layton, R., Watters, P. and Dazeley, R.: Authorship Attribution for Twitter in 140 Characters or Less, *Proceedings of the 2010 Second Cybercrime and Trustworthy Computing Workshop*, CTC '10, pp. 1–8 (2010).

[12] Salton, G. and McGill, M. J.: *Introduction to Modern Information Retrieval*, McGraw-Hill, Inc., New York, NY, USA (1986).

[13] Blei, D. M., Ng, A. Y. and Jordan, M. I.: Latent dirichlet allocation, *Journal of machine Learning research*, Vol. 3, No. Jan, pp. 993–1022 (2003).

[14] Mikolov, T., Yih, W. and Zweig, G.: Linguistic Regularities in Continuous Space Word Representations, *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, pp. 746–751 (2013).

[15] Mikolov, T., Chen, K., Corrado, G. and Dean, J.: Efficient Estimation of Word Representations in Vector Space, *CoRR*, Vol. abs/1301.3781 (2013).

[16] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. and Dean, J.: Distributed Representations of Words and Phrases and their Compositionality, *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pp. 3111–3119 (2013).

[17] Le, Q. V. and Mikolov, T.: Distributed Representations of Sentences and Documents, *CoRR*, Vol. abs/1405.4053 (2014).