

標準間の関連情報作成のための 辞書構築手法の提案

高橋 雄志^{†1} 太田 悟^{†2} 間形 文彦^{†3}
西垣 正勝^{†4} 佐々木 良一^{†1} 篠宮 紀彦^{†2} 勅使河原 可海^{†1}

近年では、様々なコンシューマデバイスの登場、発展により生活の様々な場面で ICT システムが活用されている。各組織においてそれらを安心して安全に活用していくためには、共通のルールによって運営されていることが望ましい。それは、セキュリティ標準や安全基準といった規格群を定め、運用することによって実現されつつある。しかし、新しい分野、システムが普及速度と規格の策定速度とに隔たりがあり、追いついていないのが現状である。我々はこれまで、セキュリティ標準に着目し、標準間の関連情報や、それぞれの標準の固有項目を自動抽出する研究を行ってきた。これらを抽出することによって、策定された規格がその目的に準拠していることを明確化できることを確認している。また、その過程で有効な情報を得るためには、解析に用いる複合名詞の辞書も重要な要素であるという知見を得ることができた。しかし、その辞書は、一般的に専門的な知識を用いるか、システム運用者の主観によって設計されている。本稿では、専門的な知識を用いずに、標準のドキュメントを文書解析することで取得できる情報に基づき辞書を作成する方法を提案し、実験によってその有効性を確認した。

A Proposal of a Dictionary Building Method for Pertinent Information Creation between Standards

YUJI TAKAHASHI^{†1} SATORU OTA^{†2} FUMIHIKO MAGATA^{†3}
MASAKATSU NISHIGAKI^{†4} RYOICHI SASAKI^{†1}
NORIIHIKO SHINOMIYA^{†2} YOSHIMI TESHIGAWARA^{†1}

Recent years, an ICT system is utilized by various appearance and development of a consumer device in various living situations. It is desirable to be managed by a common rule to utilize these ICT systems safely and securely in each organization. That is being achieved to set and use the related standards such as security and safety. However, a new field and system are spreading faster than developing the standard. Therefore, the current status shows the difficulty in utilization of standards. We have been studying, in focusing on security standards, to automatically extract pertinent information between standards and unique requirements of the standards. By extracting such information we confirmed to clarify that the developed standard is complied with their requirements. In addition, we obtained the knowledge that a dictionary of the compound nouns for the analysis is also an important element to get effective information in this process. The dictionary is generally designed with specialized knowledge or by the subjective view of the system operators. In this paper, we proposed a dictionary building method based on the information which can be acquired by analyzing standard documents without using technical knowledge. Moreover, we confirmed the validity of the proposed method by experimental data.

1. はじめに

近年、組織はサイバー攻撃から情報資産を守る自己防衛だけでなく、踏み台として利用されるような二次的な加害者になるリスクも防ぐ必要もあり、情報セキュリティの観点から対処すべき問題の範囲が拡大している[1]。そのため、組織はそのようなリスクを適切に管理する必要がある。それらのリスク管理について、政府機関や業界団体は IT システムが満たすべきセキュリティ水準を定めたガイドライン（以下、セキュリティ標準という）を策定している[2]。こ

れに伴い、組織の情報資産に対するリスクへの対策状況を、外的機関からの認証を取得することによって、利害関係者からの信頼を獲得することが重要視されている[3]。

具体的な認証評価として ISMS 適合性評価制度に基づく情報セキュリティマネジメントシステム（以下、ISMS : Information Security Management System という）認証の取得がある。この ISMS 認証は認証制度ができて以来取得件数が増加し続けており、2014年11月27日現在で4,548件と多くの企業や組織が取得している[4]。

また、近年における、スマートフォンやタブレット端末など（以下、まとめてスマートデバイスという）の普及に伴い、私物端末を業務に利用する BYOD(Bring Your Own Device) という考え方が、提唱されている。BYOD は、私物端末を業務に利用するため、個人データと業務データが混在している可能性が高い。しかし、スマートデバイス使用者の情報セキュリティに対する意識が低く、機密情報が漏洩してしまう危険性がある[5]。また、私物端末上に業務

†1 東京電機大学総合研究所サイバーセキュリティ研究所
Cyber Security Laboratory, The Research Institute of

†2 創価大学大学院工学研究科
Graduate School of Engineering, Soka University

†3 NTT セキュアプラットフォーム研究所
NTT Secure Platform Laboratories

†4 静岡大学大学院情報学研究科
Graduate School of Informatics, Shizuoka University

データが存在していなくても、組織の情報資産へ私物端末からアクセスすることは、新しい重大なリスクを伴う。このため、単純なセキュリティ対策だけでは、業務データのセキュリティを十分に確保することは困難である[6]。

そこで、BYOD に使用される私物端末の管理に関して、慎重な計画による十分なセキュリティプロセスおよびセキュリティコントロールを確実に実現し、私物端末上に存在する機密情報と機密性の高いアプリケーションを保護する必要がある。そのため強力なユーザ認証、アイデンティティライフサイクル管理、Web アクセス管理、情報の保護、および暗号化などの領域を含めて、BYOD に使用される端末に関するセキュリティ要件の重要性が高まっており、様々な標準化活動も積極的に行われている[7]。

現在、BYOD の運用手順に関する対策は、組織によって異なる[6]。そのため、BYOD の運用に関する標準が策定され、その認証を取得する際に、組織が既に対応済みである対策と、これから新たに対応しなければならない対策を漏れなく判断する必要がある。

ISMS などのセキュリティ認証の多くは、記載されている要求事項を満たすことにより、組織などの対象が標準に適合していることを評価する。こういったセキュリティ標準での要求事項には、保護すべき情報資産の管理体制やセキュリティ対策の基本的な考え方が示されている。また、セキュリティ対策の基本的な考え方は、いかなる組織や IT システムにも対応できるように、汎用的な記述になっている。そのため、適切な対応策の実現には、IT システムとセキュリティの専門知識の両方が要求される。しかし、IT システムの専門知識を持つ人材は不足しており、その中でも特にセキュリティ人材は不足していると言われている[8]。

そのため、セキュリティ標準の認証評価時に、取得するセキュリティ標準の要求事項すべてに対策がとられているか分からないという網羅性の問題があげられる。

そこで、認証取得に必要な要求事項の達成度を確認するためのセキュリティ評価システムが活用されている[9]。同様に、セキュリティ標準を知識ベースとして用いた IT システムの設計評価システムの提案もなされている[10]。しかし、標準は時代の変化に合わせて頻りに内容が変更される。中でもセキュリティ標準については、変更が行われる回数が他の標準に比べて多く、新たなセキュリティ標準も多数策定されている[11]。さらに、組織の規模や製品によって、認証の取得範囲やセキュリティ標準の満たすべき要求事項が異なる。そのような違いにより、認証取得のために新たな体制を作り、それぞれの認証取得にあわせた個別の評価ツールや人員を用いてセキュリティ対策実施状況の評価をやり直さなければならないといった状況を作り出す原因となっている[12]。このような問題を解決するためには、個別のセキュリティ評価ツールではなく、標準の内容に依存しない統合的なセキュリティ評価ツールを実現する必要性

があると考えられる。

これらの問題に対して我々は、統合的なセキュリティ対策実施状況の評価が可能なセキュリティ評価プラットフォームを提案してきた[13]。このプラットフォームでは、異なる標準間における対応策のデータ移行機能を提供している。この機能を実現するために、異なる標準間における同一の対応策を求める要求事項を示す情報（以下、関連情報という）が定義されている必要がある。しかし、その異なる標準同士の関連情報が必ず定義されているとは限らないという問題がある。そこで、情報検索の分野で使われている、自然言語処理を用いたテキスト間の近似度算出手法[14] を応用し、各標準の項目同士の近似度から、項目の相関を取ることで、関連情報を作成する実験を行い、その有効性を確認した[15]。

本稿では、標準間の関連情報作成の際に行う相関分析で重要な要素の 1 つである形態素解析に使う辞書（以下、辞書という）を作成する手法を提案する。本手法では、最初に関連情報を作成したい標準の形態素解析を行い、各形態素の出現数に基づいたパレートを引き、出現回数が上位の形態素を組み合わせることで辞書の作成を行っている。この手法を用いることで、関連情報を作成する際に、同じような単語が多く出現しても、それぞれの単語に違う意味を与えることで、正確な関連情報を取得できるようになった。

2. 関連する規格および技術

2.1 マネジメントシステム規格(MSS : Management System Standard)

MSS とは、組織が特定の目的を達成するために方針、プロセス及び手順を策定し、それらを体系的に管理するための要求事項又は指針を提供する規格である[16]。MSS では、PDCA サイクルに基づいた経営を行うことにより、組織の目標を達成するための力を継続的に改善していくことを求めている。代表的な MSS の標準として、製品やサービスの品質向上のための規格である ISO 9001 や、環境への悪影響を防ぐための規格である ISO 14001、また後述するセキュリティに関する規格である ISO/IEC 27001 などが存在する。これら MSS の標準については、MSS 同士の整合性をはかるために、国際標準化機構によって、MSS の上位構造と共通テキスト(以下、MSS 共通テキストという)、共通用語の定義の指針が開発された[16]。そのため、MSS に基づく標準を新たに策定、もしくは改訂を行う場合においては、常に文献[16] に記載されている定義に従って作成し、妥当性の評価を行わなければならない。今後、MSS の標準については、MSS 共通テキストに従って策定・改訂が行われるため、我々の提案手法で作成された辞書の精度が向上すると考えられる。

ISMS に関する MSS として、ISO/IEC 27001 がある。これは、国際標準化機構と国際電気標準会議の共同によって

策定された規格である[17]. この規格は, ISMS に必要な要求事項を規定し, ISMS の開発, 実施, 改善を支援するための指針から構成されている. そのため, いかなる規模や形態の組織にも適用可能な規格となっている. この規格の認証を取得するために, まず, 組織は情報セキュリティに関するリスクを分析, 評価し, 必要に応じて適切な情報セキュリティ制御を実装する必要がある. また, 情報セキュリティの運用は, 状況に応じてリスクや対策が変化していくため, 他の MSS 同様, PDCA サイクルにより継続的な見直しと改善が要求される. ISO/IEC 27001 は, 2008 年からの定期見直しにより文献[16] に基づき, MSS 共通テキストの内容に沿って改訂が行われ, 2013 年 10 月 1 日に ISO/IEC 27001:2013 要求事項が発行された.

また, ISO/IEC 27000 ファミリーとして, ISMS に必要な要求事項である ISO/IEC 27001 を含む様々な規格が検討され, 発行されている. ISO/IEC 27000 ファミリーは, 多くの分野においての基準となる標準群となり, ISMS に基づく PDCA サイクルによる運営の重要性を示している.

ISO/IEC 27000 ファミリーは, 2013 年 12 月の時点で 14 種類の標準が策定済みであり, 他にも多くの標準が策定中となっている[18].

2.2 情報セキュリティ対策マップ

NPO 日本ネットワークセキュリティ協会の配下の標準化部会の情報セキュリティ対策マップ WG によって, 組織全体の情報セキュリティ対策の状況を確認するための情報セキュリティ対策マップの作成が行われている[12]. この WG では, 多くのセキュリティ標準からセキュリティ対策を収集し, 対策の構造についてのモデル化, ツリー化を行っている. この WG での対策収集の過程で各セキュリティ標準内での記述や表現, 対策の粒度 (抽象度), 対策を行う時期や対象が異なっていることが指摘されている.

その上で, あらゆる対策を同じ表現で記述する正規化や, 対策同士を比較できるように, 極限まで対策を具体化させる原子化を行っている. 今後, 対策の正規化, 原子化が行われたセキュリティ標準が新しく策定されるようになれば, 我々の提案手法によって作成された辞書の精度が向上すると考えられる.

なお, 本稿では, 後述の相関分析の中で標準の形態素解析を行った際の各形態素の出現頻度がパレートの法則[19]に当てはまると予測し, 解析を行った. また, パレートの法則とデシル分析[20]に基づき後述の実験の対象となる形態素の特定を行った.

パレートの法則とは, x を所得金額, R_x を x 以上の所得のある人数, A を正の常数, α をパレート係数としたとき, 経済や社会体制が異なる国々または異なる時期においても, $R_x = Ax - \alpha$ が成り立つと指摘されているものであり, 80 対 20 の法則とも呼ばれている. デシル分析とは, 顧客の購入金額に応じて売上構成比を分析する手法のことである.

2.3 関連情報の作成

先行研究では, 評価の対象となる標準を整理したデータの入れ替えのみで, 他の標準と同様にセキュリティ評価を行なうことができる統合的なセキュリティ評価プラットフォームを提案してきた[13]. このセキュリティ評価プラットフォームの主な機能の 1 つである異なる標準間のデータ移行機能では, 関連情報が必要となるが定義されているとは限らないといった問題がある. そのため我々は, 1 章で述べたように関連情報を作成する手法を提案している[13]. 提案手法は異なる標準の項目間の相関から, 関連情報の作成を行う手法である.

2.3.1 相関分析手法

我々は, 異なる標準の項目間の相関を取る方法として, 文書の分類や情報検索に関する研究分野において使われている自然言語処理によって文書間の近似度を算出する手法を用いている[14].

はじめに, 関連情報算出の対象となるセキュリティ標準を決定し, テキスト情報を取得する. 次に, 取得したテキスト情報を標準の項目ごとに「茶釜システム」[21]などを用いて形態素解析を行い, 形態素に分割する. 形態素とは, 文書の形態素解析によって得られた言語における意味を持つ最小単位のことである. そして, 得られた形態素から文書の内容を表す単語を索引語と定義し抽出する. 形態素のうち, 文書の内容を特徴付ける上で, 役に立たない語を不要語として定義し削除する. 不要語を削除した項目ごとの索引語が, その文書の内容にどれだけ密接に関係しているのかを, 索引語の重要度として付与するために, 重み付けを行う. 重み付けの手法として, 文書中に出現する索引語の頻度を用いた TF (Term Frequency) や他の文書中の索引語の分布を考慮した IDF (Inverse Document Frequency), それらを組み合わせた TFIDF がよく用いられる[14]. その後, 各項目の重みをベクトルや行列で表現する. 重み付けによって作成した各標準の項目のベクトルや行列の全組み合わせに対して余弦[14]を計算し, 項目間の近似度を算出する. 近似度の計算には余弦の他に, Dice 係数や Jaccard 係数などもある[14]. 最後に, 各セキュリティ標準の項目間の近似度が最大となる項目の組のうち, どちらの標準から見ても一致しているものを相関がある項目の組と定義する.

3. 辞書構築手法

先行研究では, 辞書を用いずに相関を求める場合と, なんらかの辞書を用いて相関を求める場合では, 辞書を用いて相関を求めた方がより高い精度で関連情報を作成できるという知見を得ることができている[22].

標準に特化した辞書を作成するには, IT に関する知識や, 関連情報を作成したい標準に関する知識が求められる. しかし, そういった専門知識を持った人材の確保や, 新しい分野の標準の関連情報を作成する際に, 適切な人材を集めることは困難であると言われている[8].

我々は、何度も出現している索引語は、標準内において、同一の意味ではなく、異なる意味として使われているのではないかと考えた。そこで2章で述べた様々な技術に着目し、以下の手法を提案する。

3.1 辞書構築手順

① 標準の文書に対して形態素解析を行い、パレート図を作成

最初に、辞書を用いずに形態素解析を行う。その際に各形態素の出現回数をカウントする。カウントした出現回数で降順ソートを行い、上位の形態素からそれぞれの出現比、累積出現比を求める。また、形態素の総数から構成比も求める。

② デシル分析により、各形態素のD値を算出

①で作成した累積出現比を元に10%充足ごと、かつ出現回数が増えたタイミングでD値を増加させ、各形態素にD01からD10の値を与える。なお、出現回数が0となる形態素にはDZZという値を与える。

③ 情報検索の範囲を確定

パレートの法則を参考に出現累積比80%を満たす範囲内で、辞書登録用の複合名詞を作成する範囲を決定する。

本稿で行う実験では、ここの検索範囲を初期の段階で固定して最適な範囲を探るものとする。

④ 検索範囲の形態素から複合名詞を作成し辞書登録

標準の各項目の中で、③の検索範囲内の形態素同士が隣り合わせになっていた場合は複合名詞として辞書登録を行う。

⑤ 辞書を使って①から④を繰り返す

①の工程で、繰り返す時には作成された辞書を用いて形態素解析を行う。その後は、辞書への新規登録件数が0となるまでこの工程を繰り返す。

4. 関連情報作成のための辞書構築実験

本提案手法を用いて、辞書構築を行う際に対象とするD値の範囲の最適値を探るために、複数の検索範囲で辞書構築を行った。その際に、構築された辞書を用いて関連情報の作成を行って再現度を比較して辞書の評価を行った。また、辞書の構築を人の手で行ったと想定した場合との差異を示し、本手法で自動的に辞書を作成することが有効であることの確認も行う。

4.1 実験概要

本提案手法の有効性を検証するために、実際に項目間の関連情報が存在するISO/IEC 27001:2005 附属書A（以下、文書Aという）とISMS 認証基準 Ver.2.0 附属書「詳細管理策」（以下、文書Bという）の2つの標準を用いて辞書構築を行った。そして、実際にその辞書を用いて関連情報を作成して評価値を算出し、辞書の有効性比較を行った。

評価値については、それぞれの辞書を用いて作成された関連情報が元の関連情報をどの程度再現できているのかを再現率、元の関連情報に含まれる割合を適合率、再現率と

適合率の調和平均をF値として、3軸比較をするためバブルチャートを用いて有効な辞書がどれであるかの検討を行った。また、作成された複合名詞を確認して、本手法における問題点、改良すべき点の考察を行った。

4.2 実験手順

① 基準値の作成

最初に、辞書を用いずに関連情報を作成して、関連情報の再現率、適合率、F値を求める。ここで求めた数値が基準値となる。

② 形態素解析およびパレート図の作成

形態素解析を行いその結果より各形態素の出現回数取得する。取得データから出現比、累積出現比を求めてパレート図を作成する。

③ D値の計算

②と同様に各形態素の累積出現比を10%単位の充足値を基準としてD値を与える。閾値を超えた形態素が次の形態素の出現回数が一緒であった場合は上位のD値を与える。

例えば、9.99%、10.01%、10.02%、10.03%と並んだ時の各形態素の出現回数が31、30、29、28となった場合は通常通り10.02%の形態素のD値からD02となるので、D値はD01、D01、D02、D02となる。しかし、出現回数が31、30、30、29となった場合は10.01%、10.02%の2つの形態素に優劣をつけることができないので2つともD01となるので、D値はD01、D01、D01、D02となる。

各形態素に与えられたD値とパレート図から情報検索の範囲を決める。

④ 複合名詞の作成および辞書登録

③で決めた範囲に含まれる形態素同士が標準の項目内で隣接していた場合は、その組み合わせを複合名詞として定義する。例えば、「情報」という形態素と「セキュリティ」という形態素のD値が共にD01（検索範囲内）であり、標準に「情報セキュリティ」という記述があれば、「情報セキュリティ」という複合名詞として定義する。このように定義した複合名詞を辞書に登録していく。

⑤ 関連情報の作成

④で作成した辞書を用いて関連情報の作成を行う。作成された関連情報について、再現率、適合率、F値を求める。

⑥ 辞書のブラッシュアップ

④で新たな複合名詞が定義されなくなるまで、②～⑤までの工程を繰り返し辞書のブラッシュアップを行う。ただし、本実験では検索範囲は固定として③の工程で決めた検索範囲で辞書作成を繰り返すものとする。例えば最初にD03vsD03の範囲に決めた場合は、その後のパレート図の変化に関わらずD03vsD03で②～⑤の工程を繰り返す。

⑦ バブルチャートの作成

辞書構築が完了したら、完了までのそれぞれの時点での関連情報の評価値を、再現率を横軸、適合率を縦軸、グラフの交点をそれぞれの平均値、バブルのサイズをF値とし

てバブルチャート作成する。そして、バブルの移動や位置によって辞書の有効性を比較する。

⑧ 複合名詞に関する考察

最終的に構築された辞書の内容がどういったものであるのか、また人が考える複合名詞とどのような関係にあるのかを考察する。

⑨ 人によって作成された辞書との比較

これまでの工程から得られた情報から人による辞書構築と異なる要素を洗い出し比較検討する。

4.3 実験結果

① 基準値の作成

辞書を用いずに関連情報を作成した結果を表1に示す。

文書Aと文書Bはもともと近い標準であったため初期段階でも高い評価値を示していることがわかった。

表1 辞書未使用時の評価値

	再現率	適合率	F値
辞書未使用	83.54%	94.96%	88.89%

② 形態素解析およびパレート図の作成

辞書を使用せずに形態素解析を行って、各形態素の出現回数をカウントしてパレート図を作成し結果を図1に示す。

図から視覚的にパレートの法則に沿ったデータであることが確認できた。

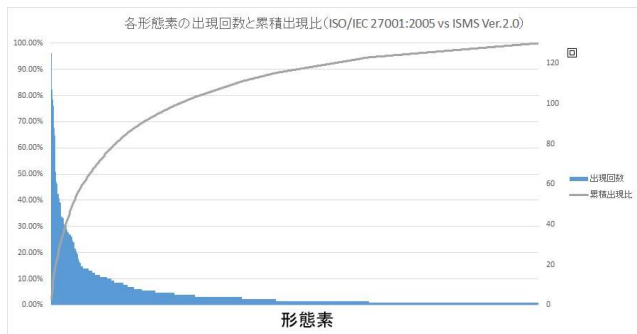


図1 各形態素の出現回数と累積出現比

③ D値の計算

それぞれの形態素のD値を求めた結果の件数、出現回数の最小、最大、構成比(累積)を表2に示す。また、各形態素の平均値、標準偏差を表3に示す。2つの表より複合名詞検索の母集団は、平均を大きく超える値として+3σの値を超える出現数のみで構成されているD01からD03とすることにした。母集団と組み合わせる範囲は、同じD01からD03をはじめとしパレートの法則に則りD01からD08の範囲について、計6回の実験を行った。以下、D03vsD0xとして検索範囲を示す。

表2 各D値におけるサマリ

D値	件数	最小	最大	構成比(累積)
D01	3	142	207	0.40%
D02	6	61	107	1.20%
D03	9	46	60	2.40%
D04	13	34	45	4.14%
D05	26	18	33	7.61%
D06	30	14	17	11.62%
D07	54	8	13	18.83%
D08	81	5	7	29.64%
D09	268	2	4	65.42%
D10	259	1	1	100.00%
DZZ	0	0	0	0.00%

表3 出現回数の平均と標準偏差

平均	標準偏差	+σ	+2σ	+3σ	+4σ	+5σ	+6σ	+7σ
6.4	12.6	19.0	31.6	44.2	56.8	69.4	82.0	94.6

④ 複合名詞の作成および辞書登録

D03vsD03 から D03vsD08 の辞書への複合名詞登録件数の推移を表4に示す。

D03vsD07 以外の組み合わせではいずれも6回目のトライアルで新規に作成される複合名詞がなくなり辞書登録が完了となった。D03vsD07の時だけ、6回目のトライアルではなく、7回目のトライアルで辞書登録が完了となった。

表4 複合名詞登録件数の推移

		1回目	2回目	3回目	4回目	5回目	6回目	7回目
D3vsD3	合計	23	32	38	53	55	55	
	新規	23	9	6	15	2	0	
D3vsD4	合計	45	67	96	101	106	106	
	新規	45	22	29	5	5	0	
D3vsD5	合計	77	116	139	151	152	152	
	新規	77	39	23	12	1	0	
D3vsD6	合計	96	150	162	172	182	182	
	新規	96	54	12	10	10	0	
D3vsD7	合計	120	190	210	218	246	248	248
	新規	120	70	20	8	28	2	0
D3vsD8	合計	150	227	262	273	278	278	
	新規	150	77	35	11	5	0	

⑤ 関連情報の作成

D03vsD03 から D03vsD08 の各辞書での関連情報作成結果の評価値を表5に示す。表中の青文字は、再現率、適合率、F値の最大値となっているデータを示している。

表 5 辞書使用時の関連情報の評価値

		1回目	2回目	3回目	4回目	5回目	6回目
		件数	構成比	件数	構成比	件数	構成比
D03vsD03	再現率	84.18%	82.91%	82.91%	82.91%	82.91%	
	適合率	95.68%	95.62%	95.62%	95.62%	95.62%	
	F値	89.56%	88.81%	88.81%	88.81%	88.81%	
D03vsD04	再現率	82.91%	82.91%	83.54%	84.81%	84.81%	
	適合率	95.62%	95.62%	95.65%	96.40%	96.40%	
	F値	88.81%	88.81%	89.19%	90.24%	90.24%	
D03vsD05	再現率	83.54%	85.44%	85.44%	86.71%	86.71%	
	適合率	95.65%	97.83%	97.83%	97.86%	97.86%	
	F値	89.19%	91.22%	91.22%	91.95%	91.95%	
D03vsD06	再現率	83.54%	86.71%	86.71%	86.71%	86.71%	
	適合率	96.35%	97.86%	97.86%	97.86%	97.86%	
	F値	89.49%	91.95%	91.95%	91.95%	91.95%	
D03vsD07	再現率	84.18%	86.08%	86.08%	86.08%	86.08%	
	適合率	96.38%	97.86%	97.84%	97.84%	97.84%	
	F値	89.86%	91.95%	91.58%	91.58%	91.58%	
D03vsD08	再現率	83.54%	86.08%	84.81%	84.81%	85.44%	
	適合率	96.35%	97.84%	97.81%	97.81%	97.83%	
	F値	89.49%	91.58%	90.85%	90.85%	91.22%	

⑥ 辞書のブラッシュアップ

D03vsD03 から D03vsD08 における各 D 値の件数推移を表 6 から表 11 に示す。

表 6 D03vsD03 の各 D 値の件数推移

D03vsD03	D値	1回目の辞書		2回目の辞書		3回目の辞書		4回目の辞書	
		件数	構成比	件数	構成比	件数	構成比	件数	構成比
D01	5	0.65%	6	0.77%	6	0.77%	6	0.76%	
D02	8	1.68%	9	1.92%	9	1.92%	9	1.77%	
D03	11	3.11%	10	3.21%	11	3.32%	13	3.42%	
D04	15	5.05%	18	5.51%	16	5.36%	17	5.58%	
D05	26	8.42%	27	8.97%	29	9.07%	34	9.89%	
D06	35	12.95%	31	12.95%	45	14.81%	38	14.70%	
D07	65	21.37%	65	21.28%	52	21.46%	54	21.55%	
D08	137	39.12%	137	38.85%	136	38.83%	139	39.16%	
D09	201	65.16%	203	64.87%	203	64.75%	202	64.77%	
D10	269	100.00%	274	100.00%	276	100.00%	278	100.00%	
DZZ	0	-	1	-	3	-	12	-	

表 7 D03vsD04 の各 D 値の件数推移

D03vsD04	D値	1回目の辞書		2回目の辞書		3回目の辞書		4回目の辞書	
		件数	構成比	件数	構成比	件数	構成比	件数	構成比
D01	6	0.76%	6	0.74%	6	0.72%	6	0.72%	
D02	8	1.78%	8	1.74%	9	1.81%	9	1.81%	
D03	12	3.30%	16	3.72%	13	3.38%	13	3.37%	
D04	17	5.46%	22	6.45%	23	6.16%	31	7.10%	
D05	35	9.91%	24	9.43%	34	10.27%	25	10.11%	
D06	37	14.61%	38	14.14%	37	14.73%	48	15.88%	
D07	54	21.47%	89	25.19%	80	24.40%	71	24.43%	
D08	140	39.26%	110	38.83%	115	38.29%	115	38.27%	
D09	203	65.06%	209	64.76%	219	64.73%	219	64.62%	
D10	275	100.00%	284	100.00%	292	100.00%	294	100.00%	
DZZ	7	-	11	-	16	-	17	-	

表 8 D03vsD05 の各 D 値の件数推移

D03vsD05	D値	1回目の辞書		2回目の辞書		3回目の辞書		4回目の辞書	
		件数	構成比	件数	構成比	件数	構成比	件数	構成比
D01	6	0.74%	6	0.71%	6	0.70%	6	0.69%	
D02	9	1.84%	8	1.66%	9	1.74%	9	1.73%	
D03	12	3.31%	16	3.56%	15	3.48%	17	3.69%	
D04	21	5.89%	27	6.77%	24	6.27%	31	7.27%	
D05	39	10.67%	32	10.57%	49	11.96%	35	11.30%	
D06	28	14.11%	40	15.32%	38	16.38%	41	16.03%	
D07	88	24.91%	75	24.23%	63	23.69%	104	28.03%	
D08	111	38.53%	118	38.24%	121	37.75%	83	37.60%	
D09	216	65.03%	219	64.25%	222	63.53%	224	63.44%	
D10	285	100.00%	301	100.00%	314	100.00%	317	100.00%	
DZZ	10	-	21	-	25	-	31	-	

表 9 D03vsD06 の各 D 値の件数推移

D03vsD06	D値	1回目の辞書		2回目の辞書		3回目の辞書		4回目の辞書	
		件数	構成比	件数	構成比	件数	構成比	件数	構成比
D01	6	0.72%	6	0.69%	6	0.69%	6	0.68%	
D02	9	1.81%	9	1.73%	9	1.72%	9	1.70%	
D03	13	3.37%	15	3.46%	16	3.55%	19	3.85%	
D04	20	5.78%	24	6.22%	32	7.21%	28	7.03%	
D05	36	10.12%	45	11.41%	35	11.21%	36	11.11%	
D06	47	15.78%	39	15.90%	66	18.76%	61	18.03%	
D07	74	24.70%	104	27.88%	77	27.57%	79	26.98%	
D08	114	38.43%	83	37.44%	85	37.30%	87	36.85%	
D09	218	64.70%	227	63.59%	228	63.39%	236	63.61%	
D10	293	100.00%	316	100.00%	320	100.00%	321	100.00%	
DZZ	13	-	29	-	35	-	37	-	

表 10 D03vsD07 の各 D 値の件数推移

D03vsD07	D値	1回目の辞書		2回目の辞書		3回目の辞書		4回目の辞書		5回目の辞書	
		件数	構成比	件数	構成比	件数	構成比	件数	構成比	件数	構成比
D01	6	0.71%	6	0.67%	6	0.66%	6	0.66%	6	0.65%	
D02	9	1.77%	9	1.67%	9	1.65%	9	1.64%	12	1.95%	
D03	15	3.54%	16	3.46%	17	3.53%	21	3.94%	14	3.46%	
D04	24	6.38%	29	6.70%	28	6.62%	28	7.01%	29	6.99%	
D05	28	9.68%	38	10.94%	38	10.80%	44	11.83%	43	11.24%	
D06	57	16.41%	60	17.63%	59	17.31%	50	17.31%	51	16.76%	
D07	65	24.09%	73	25.78%	63	24.26%	161	34.94%	154	33.41%	
D08	117	37.90%	156	43.19%	169	42.89%	72	42.83%	77	41.73%	
D09	226	64.58%	179	63.17%	182	62.95%	184	62.98%	192	62.49%	
D10	300	100.00%	330	100.00%	336	100.00%	338	100.00%	347	100.00%	
DZZ	20	-	41	-	50	-	52	-	68	-	

表 11 D03vsD08 の各 D 値の件数推移

D03vsD08	D値	1回目の辞書		2回目の辞書		3回目の辞書		4回目の辞書	
		件数	構成比	件数	構成比	件数	構成比	件数	構成比
D01	6	0.69%	6	0.65%	6	0.64%	6	0.64%	
D02	9	1.73%	9	1.64%	10	1.71%	10	1.70%	
D03	14	3.34%	17	3.49%	19	3.74%	23	4.14%	
D04	25	6.21%	27	6.43%	28	6.73%	22	6.48%	
D05	28	9.44%	37	10.47%	43	11.32%	42	10.95%	
D06	56	15.88%	62	17.23%	52	16.88%	53	16.58%	
D07	89	26.12%	61	23.88%	147	32.59%	149	32.41%	
D08	88	36.25%	164	41.77%	70	40.06%	68	39.64%	
D09	241	63.98%	186	62.05%	203	61.75%	205	61.42%	
D10	313	100.00%	348	100.00%	358	100.00%	363	100.00%	
DZZ	28	-	57	-	73	-	79	-	

⑦ バブルチャートの作成

各組み合わせでの評価値をバブルチャートで表現したものを図 2 から図 7 に示す。

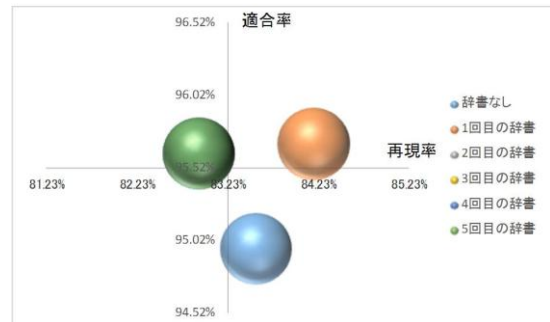


図 2 D03vsD03 の評価値バブルチャート

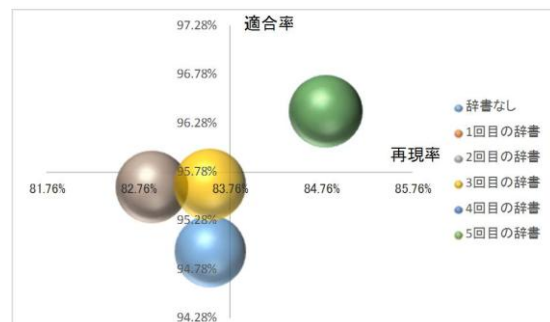


図 3 D03vsD04 の評価値バブルチャート

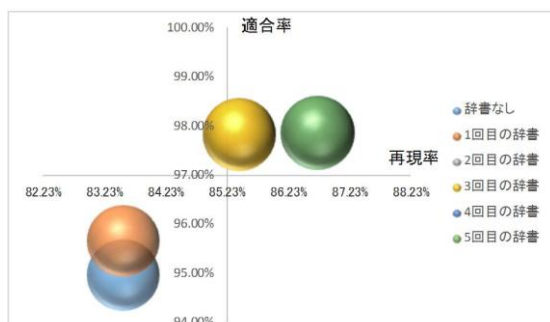


図 4 D03vsD05 の評価値バブルチャート

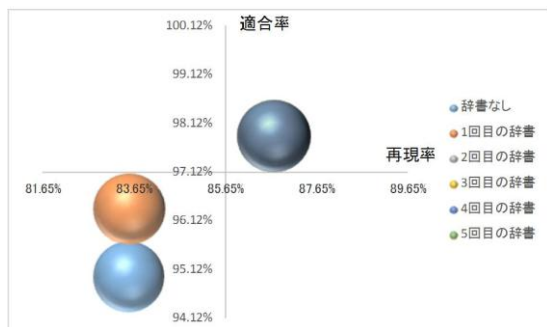


図5 D03vsD06の評価値バブルチャート

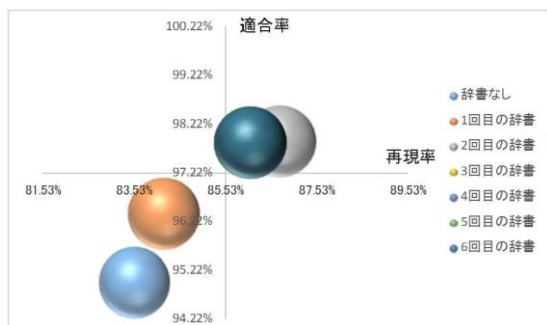


図6 D03vsD07の評価値バブルチャート

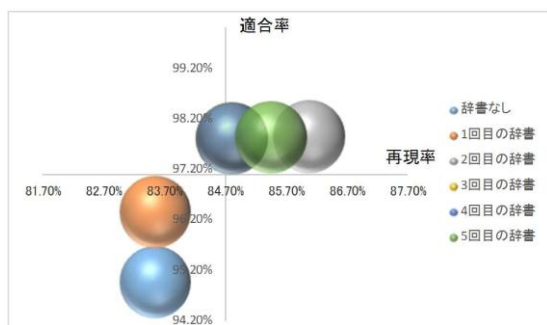


図7 D03vsD08の評価値バブルチャート

⑧ 複合名詞に関する考察

表5の結果より D03vsD05, D03vsD06, D03vsD07 で作成された複合名詞を中心に確認し、その他の実験結果と合わせることで、以下の傾向があることがわかった。

すべての組み合わせにおいて、辞書なしと1回目の辞書構築後の結果を比較して改善と見て取ることができた。しかし、辞書が構築された後は組み合わせによっては改善が見られず、かえって結果が悪くなる組み合わせもあった。以上のことより、辞書構築1回目とそれ以降では意味合いが異なると推察した。

改善が見られた組の2回目以降の複合名詞を確認すると、複合名詞がさらに他の形態素とつながって他の複合名詞として登録されていることや、ひとつの形態素が複数の種類の形態素とそれぞれ組み合わせることで異なる意味を持つ複合名詞が登録されていた。このように、構築を繰り返すことで単名詞では伝わらない技術的な要素を読み取ることができるような複合名詞が作成される。これらは標準の要求事項ともいえる項目ともなり得る項目であると推察した。

⑨ 人によって作成された辞書との比較

本手法を用いて辞書作成にかかる時間を計測したところ形態素解析から複合名詞を作成するまでにかかる時間は、辞書のサイズなどに依存してある程度の誤差が生じるものの秒オーダーで結果が出るということがわかった。また、作成された辞書を次の形態素解析に使えるように変換するために必要な時間も分オーダーで実行できることが分かった。

実際に、専門家に依頼して辞書を作ることを想定した場合には、文章を読み込み時間、そこから複合名詞として定義すべき語を選ぶ時間がかかることになる。しかし、それらは少なく見積もっても時間オーダーの作業となることは想像に難くない。多くの場合で日をまたいで作業することになると考える。

正確性については、人が作成する場合は当然作成者の専門知識のレベルによってできあがる辞書の精度が決まってくる。登録される件数も同様に知識レベルによって決まる。一方、提案手法では専門知識を一切使用しないので、誰が作成に携わっても同じ時間、同じ精度、同じ件数の辞書を作成することができる。

以上のことを表にまとめると表12に示すようになる。

表12 人による作成と本手法の結果比較

	人による作成	提案手法による作成
正確さ	作成者の知識量に依存する。登録件数も同一になるとは限らない。	誰が実行しようと一定のレベルの辞書が作成できる。登録件数は検索範囲に依存するが一定の数となる。
作成に要する時間	実際に、標準の文書を読みこんでから、必要な語を抽出することになるので、早くも時間オーダーの時間がかかり、複雑な文書であれば日オーダーの時間がかかる。	データが電子化されている条件下であれば、形態素解析は秒オーダー、辞書を次の形態素解析に使用できるようにコンバートするのに分オーダー、その繰り返し分の時間となるため10分から1時間の時間が必要となる。

4.4 考察

本手法によって構築された辞書を用いて関連情報を作成した場合に、辞書を用いずに関連情報を作成した場合に比べすべてのケースで評価値が改善されることがわかった。このことより本手法で構築された辞書は関連情報を作成する際に有効な辞書であることがわかった。

また、実験結果からただ単に多くの形態素を組み合わせるよりも、範囲を絞って必要な形態素のみを使用して辞書を構築するほうが有効であることがわかった。しかし、本手法ではそれぞれの形態素の種別を区別することなく複合名詞を定義しているため、接尾語が先頭に来てしまう複合名詞が作られてしまった。このような複合名詞は時にはブラッシュアップの段階で正しい複合名詞となるケースもあった。例えば、「的セキュリティ」という複合名詞が「人的セキュリティ」という複合名詞に次のトライアルで置き換わるといったものがあつた。

本手法では、辞書を機械的に作成できるため使用者の知識レベルによらず、高い精度の辞書が短時間で構築される

ことがわかった。

5. 今後の課題

本実験では、カタカナ語や英語が初期段階で適切に形態素にわかれておらず、かつ辞書構築でも D 値が低かったために最後まで不自然な単語として残っていた。こうした問題に対して最初に形態素解析をする段階で取り込めるような仕組みが必要であるとする。

また、4.4 節でも述べたように接尾辞のように語の接続関係が決まっている形態素を正しくつなぐことで、より正確な複合名詞を作り出すことができると考える。この課題については、複合名詞を作る工程で形態素の種別を加味して複合名詞を作る仕組みを検討したい。

本稿での実験では、形態素の出現回数を文書 A、文書 B の合計値としてカウントしてパレート図および D 値の設定を行っている。各標準で別々に形態素も出現回数を出してそれぞれの標準用の辞書を構築してそれらをマージした辞書を用いて関連情報を作成する場合との比較を行い、どちらの方がより正しい関連情報を導き出すことができるのかの検証が必要と考える。

6. まとめ

本稿では、標準間の関連情報作成時に用いる辞書を構築する手法を提案した。そのために、経営の分野などで用いられるパレートの法則やデシル分析などを応用し辞書構築の対象範囲を決めて構築実験を行い、その評価を行った。

実験では、提案手法を用いて作られた辞書を使って関連情報を作成して、辞書を用いずに作成した関連情報よりも制度の高い関連情報を作成することができた。これにより本手法を用いることで、専門知識がなくても関連情報を作成する際に使用する辞書が構築できることがわかった。

そして、この手法を応用することで辞書から各項目の要求事項の抽出ができれば、要求事項に合わせて対策リストのようなものと組み合わせることで、それぞれの標準に特化したセキュリティ対策案の提案ができるようになると思われる。

今後は 5 章で述べた課題に取り組み、手法の改良と追加の実験を行うことで、有効性を高めていく。

参考文献

- 1) 総務省：情報通信白書平成 24 年度版(online), 入手先 <<http://www.soumu.go.jp/johotsusintokei/whitepaper/h24.html>> (2012).
- 2) 内閣官房情報セキュリティセンター：政府機関の情報セキュリティ対策のための統一技術基準(平成 24 年度版)(online), 入手先 <<http://www.nisc.go.jp/active/general/pdf/k305-111.pdf>>(2012).
- 3) 日本情報処理開発協会：情報セキュリティマネジメントシステム (ISMS) の国際動向と取り組みの実際<2004 年版>, (2004.5).
- 4) 情報マネジメントシステム推進センター：認証取得組織数推移, 認証機関別・県別認証取得組織(online), 入手先 <<http://www.isms.jp/dec/jp/1st/ind/sui.html>>

- 5) 情報処理推進機構：2013 年度情報セキュリティの脅威に対する意識調査-調査報告書- (online), 入手先 <<https://www.ipa.go.jp/les/000035983.pdf>>(2013.12).
- 6) 総務省：スマートフォン・クラウドセキュリティ研究会-最終報告-(online), 入手先<<http://www.soumu.go.jp/main/content/000166095.pdf>>(2013.6).
- 7) TechTarget ジャパンホワイトペーパー：コンシューマデバイスのセキュリティ戦略計画のために考慮すべきポイント, (online), 入手先<<http://wp.techtarget.itmedia.co.jp/contents/?cid=11501>>.
- 8) 情報処理推進機構：「情報セキュリティ人材の育成に関する基礎調査」報告書について(online), 入手先 <<http://www.ipa.go.jp/security/fy23/reports/jinzai/>>.
- 9) 情報処理推進機構：セキュリティ設計評価支援ツール V03 (online), 入手先 <<http://www.ipa.go.jp/security/fy13/evalu/ccsystem/CCtoolV03/secevtoolV03.htm>>.
- 10) 芦野佑樹, 高橋雄志, 森田陽一郎, 島成佳, 岡村利彦, 勅使河原可海, 佐々木良一：セキュリティ標準に基づいた IT システム評価支援ツールの開発, 情報処理学会コンピュータセキュリティシンポジウム(2013.10).
- 11) 日本情報経済社会推進協会：情報セキュリティマネジメントシステム適合性評価制度の概要(online), 入手先 <<http://www.isms.jp/dec.or.jp/doc/ismspanf.pdf>>.
- 12) 日本ネットワークセキュリティ協会：情報セキュリティ対策マップ WG 情報セキュリティ対策マップ検討 WG 活動報告 (online), 入手先<http://www.jnsa.org/seminar/2013/0607/video_t1.html>.
- 13) 高橋雄志, 篠宮紀彦, 勅使河原可海：国際標準に基づいたセキュリティ評価プラットフォームの提案, 日本セキュリティ・マネジメント学会学会誌 Vol.27, No.2, pp.16-29(2013.9).
- 14) 徳永健伸：情報検索と言語処理, 東京大学出版会(1999).
- 15) 高橋雄志, 篠宮紀彦, 勅使河原可海：セキュリティ標準間の関連情報作成手法の検討とその適応, 情報処理学会論文誌コンシューマデバイス&システム第 3 巻, pp.22-32.(2013.12).
- 16) 日本規格協会：ISO/IEC 専門業務用指針, 第 1 部, 統合版 ISO 補足指針-ISO 専用手順 2013, 入手先 <<http://www.jsa.or.jp/itn/pdf/shiryu/isohosoku taiyaku1304.pdf>>
- 17) ISO/IEC 27001 Information technology - Security techniques - Information security management systems - Requirements(2013)
- 18) 情報マネジメントシステム推進センター：国際動向「ISO/IEC 27000 ファミリーについて」, 入手先 <http://www.isms.jp/dec.or.jp/27000family_20131212.pdf>
- 19) 情報システム用語事典：パレートの法則 (ばれーとのほうそく) - ITmedia エンタープライズ, (online), 入手先 <<http://www.itmedia.co.jp/im/articles/0408/01/news004.html>>
- 20) デシル分析とは ~ exBuzzwords 用語解説, (online), 入手先 <http://www.exbuzzwords.com/static/keyword_3510.html>
- 21) 松本祐治, 北内啓, 山下達雄, 平野善隆, 松田寛, 高岡一馬, 浅原正幸：形態素解析システム『茶釜』version 2.0 使用説明書 第二版, NAIST Technical Report, NAIST-IS-TR99012, 奈良先端科学技術大学院大学(1999).
- 22) 太田悟, 高橋雄志, 勅使河原可海, 篠宮紀彦：セキュリティ評価プラットフォームにおける国際標準間の関連情報作成手法の提案と実装, 情報処理学会第 76 回全国大会(2014.3)