研究報告

依頼対応記録表からの効果的な情報検索とその応用について

学習院コンピュータシステム支援組織　大久保　秀
京都大学大学院情報学研究科　小林靖明
学習院コンピュータシステム支援組織　安部健太
学習院コンピュータシステム支援組織　神谷匠

1. はじめに

学習院コンピュータシステム支援組織（以下，支援組織）は教職員の ICT に関するスキルアップや授業での利用を促進させる取り組みのひとつとして，学校法人学習院が運営する各教育機関（学習院大学，学習院女子大学，学習院高等科，学習院中等科，学習院女子中・高等科，学習院初等科，学習院幼稚園）を対象に，授業で使用する教室機器に関することや，研究や業務において必要なソフトウェアの操作等に関して，様々な質問やトラブルの解消を目的とした依頼を受け付けている[1]．そのような依頼は年間 3,000～4,000 件程寄せられており，これらの依頼が来るたびに，支援組織では独自に開発した Web アプリケーションである「依頼対応記録表」に，いつ，誰が，どのような依頼をして，どのようにして解決に至ったかを詳細に記録している[2, 3, 4]．2017 年 8 月現在，依頼対応記録表には 54,000 件以上の依頼データが記録されている．支援組織のスタッフはこれらの依頼データを作成および閲覧することで，すぐに解決できなかった問題を翌日以降へ引き継ぐための共有資料として利用したり，類似した依頼があったときの解決手段のヒントとして活用している．

支援組織では寄せられた依頼に対して，所属するスタッフがその依頼の解決にあたるが，大半は非常勤のスタッフ（学生を含む）で構成されおり，必ずしもその依頼で生じた問題に対しての知識や経験が十分にあるとは限らない．このような依頼に対して，これまでの支援組織では比較的に経験が豊富なスタッフを中心として，依頼対応記録表から適切に情報を検索および選択しつつ依頼の解決に取り組まなければならなかった．特に，起きた依頼が学習院固有の環境に依存している場合については，一般的な情報とは異なることもあるため，過去の依頼データから必要な情報を探しだすことが求められる．従来の依頼対応記録表には単純なキーワード検索のみ実装されているため，この検索でマッチした膨大な過去の依頼データのリストから必要な情報や解決につながる記述を発見する必要がある．そのため，効果的に情報を引き出すためには，過去に解決すべき依頼と類似する依頼の対応を行った人物やそれを見聞きした人物の記憶に基づいた検索の絞り込みやキーワード選択を行い，アクセスしたい情報に到達しているのが現状である．このような問題のため，

得たい情報に素早くアクセスするのは困難であり，常にスムーズな依頼解決を行える状態には至っていないという課題がある．

本研究は，この依頼対応記録表に高度な情報検索機能を実装し，それらを用いて問題解決のために必要な記述を高速に発見するためのシステム構築が目的である．この目的のために，自然言語処理と機械学習の手法を用いて依頼データ間の類似度を計算し，対象の依頼と類似した依頼データに高速にアクセスするためのインターフェースを実現する．依頼内容に多少の差異があっても内容の近い記述には解決方法が記載されている可能性が高いと考えられるため，類似した依頼データを発見することは速やかな問題解決に対して有効である．将来的には依頼データ間の類似度に基づいて異なる依頼データ間の関係性を分析し，頻度の高い依頼データを自動的に抽出することで，支援組織のスタッフの新人教育用のマニュアルを生成するために活用したり，問題が起きたときにその解決策を推薦できるようなシステムを視野に入れて研究を進める．

2. 学習院コンピュータシステム支援組織依頼対応記録表

2.1. 従来の依頼対応記録表の問題点

支援組織では，受け付けた依頼に対して所属するスタッフが依頼の解決にあたっており，その依頼の詳細や解決方法などを依頼対応記録表に記録している．そうすることで，今後に同様の依頼があった場合やその場で解決できなかった場合に，解決手段のヒントを得るために依頼対応記録表を利用している．しかし，従来の依頼対応記録表は単純なキーワード検索のみが実装された状態であり，キーワードにマッチした依頼データが新規登録された順番でリスト表示される．検索するキーワードの数が少ないほど，マッチする依頼データの数は多くなり，得たい情報を速やかに得ることが困難になる．例えば，メールに関する依頼が寄せられたときに，「メール」というキーワードで検索すると，9,887 件の「メール」というワードが含まれたさまざまな依頼内容の依頼データがリスト表示される．これらの中には，メールに関するトラブルの依頼データだけでなく，単純に依頼者との連絡としてメールが用いられた依頼データについても発見されてしまう．そのため，これらの中から当該の依頼内容についての解決手段が記述された依頼データを探し出すのは時間がかかる．これにより，依頼解決の遅延を招いたり，寄せられた依頼が過去に解決したことのある類似依頼であることを見落とすという課題があった．

また，このような問題を解決するために既存のシステムに機能追加を行う上でも，システムの拡張性に問題点が存在する．既存の依頼対応記録表は 10 年以上前の支援組織設立当初から運用されており，既に使用していない機能の実装が残っていたり，機能追加のたびにそのときに在籍していた支援組織のスタッフがコードをつぎ足してきた状態である．さらにはネイティブな PHP やネイティブな JavaScript で記載されてきたため，各人の裁量による記述がなされていき，可読性・拡張性ともに低い状態であり，本研究で追加する機

能を運用しているコードに追加で実装するのは困難な状態であった．

2. 2. 依頼対応記録表の改良

2.1 節で述べたシステムの拡張性の問題点の解決と，より利便性の高い依頼対応記録表の構築を目指し，以下の改良を行った．なお，依頼対応記録表から素早く必要な情報にアクセスするために，依頼データ間の類似度を自然言語処理と機械学習の手法を用いて計算し，それらを依頼データの検索システムに適用する試みを行った．この機能については 3 節で説明する．

基本システムの再構築

従来の依頼対応記録表の使用状況をもとに依頼データとして記載する項目の再検討を行った．また，新機能の追加の妨げとなる可読性・拡張性の低さを改善させるために，支援組織設立当初からつぎ足されてきたソースコードを一から構築し直した．その際，一般的に確立された効率の良い開発手法を取り入れるために，PHP の開発フレームワークである CodeIgniter [5]と，JavaScript のライブラリである jQuery [6]を利用した．これらの技術を用いることで，ロジックとデザインの分離が可能となり，さらには今後の機能追加に関しても一定のガイドラインが強制されるようなシステムを構築することができた．

データベースの再構築

従来の依頼対応記録表の使用状況をもとに，データベース内のデータについて要・不要の分類を行った．また，必要と判断された各データについても，テーブルを再構築する上で適切にテーブルの変更を加えた．さらにはデータベース内の自然言語で記述されたデータについては，本研究の目的を達成するために適切に修正を行った．

ユーザーインタフェースの改良

ユーザーインタフェースについては，高度な検索機能を実現するために設計し直した．また，依頼対応記録表を実際に利用している支援組織スタッフの要望なども参考にし，改良を行った（図 1，図 2）．

図 1. 旧依頼対応記録表

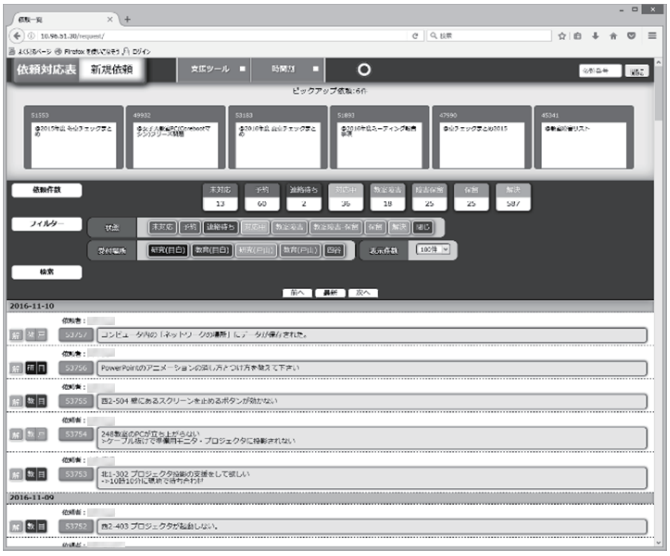

図 2. 新依頼対応記録表

3. 効果的な情報検索手法のための準備

本節では，依頼対応記録表から効果的に情報検索するための依頼データ間の類似度計算方法について説明する．

3. 1. doc2vec

一般的に，テキストデータに対して様々なアルゴリズムを適用する際にそれらのデータを数値ベクトルとして表現する方法はよく用いられており，自然言語処理や機械学習の分野では重要なタスクとして考えられている．特に，ふたつのテキスト間の類似度を計算したり，k-means や SVM などの既存の数値ベクトルに対する機械学習アルゴリズムをテキストデータに対して適用するためには必要不可欠な処理であるといえる．

このタスクに対して，最も単純な方法として知られる方法は bag-of-words（BOW）表現である．この手法は各文書を単語の多重集合として考えることで，次元がコーパスに出現する語彙と一致し，単語の出現回数を要素とするようなベクトルとして表現する方法である．この手法はそれ自体の単純さから様々な研究で用いられているが，出現する単語の順序が失われているという問題点や，単語のセマンティクスを無視した表現であるといった弱点が存在する．

これらの問題に対し，Le と Mikilov [7]は，今日では doc2vec と呼ばれている文書の分散表現手法を提案した．この手法は，Mikilov ら[8]が提案した単語の分散表現（つまり数値ベクトル）を得る手法（word2vec）を文書に対して拡張したモデルであり，単語のコンテキストの情報をうまく取り込みながらニューラルネットに学習させることで上述の問題を解消している．doc2vec は，文書のベクトル化を行う手法としては様々な場面で利用されている（例えば[9]など）．本研究では Python のライブラリである gensim [10]にある doc2vec の実装を利用した．

3. 2. doc2vec の依頼対応記録表への適用

依頼対応記録表を分析するために，データベースから問題の内容や解決方法などを自然言語で記載した項目を抽出し，それらを形態素解析器に掛け，動詞と名詞のみを抽出した．本研究では形態素解析器として Mecab [11]を利用した．ICT に関わる最新の用語や学習院内で独自に使われている用語などを学習するために，Wikipedia の記事のタイトルから得られる辞書やカスタム辞書を作成して利用した．このように作成したデータに対して，doc2vec を用いて数値ベクトル化した．

3. 3. 効果検証

本研究では以下のような実験を行った．前節で示した方法で数値ベクトル化した各依頼データについて，依頼データ間のコサイン類似度を算出し，指定した依頼データとのコサイン類似度が高い依頼データ上位 10 件を抽出した．コサイン類似度の値が 1 に近いほど，依頼データ間の類似度も高いことを表している．実験結果を表 1，表 2 に示す．また，「メール」というキーワードを含む依頼データをいくつか抽出し，抽出した各依頼データ間のコサイン類似度を求めた結果を表 3 に示す．

【実験例 1】依頼 No. 47000 とのコサイン類似度が高い依頼データ上位 10 件を抽出

依頼 No.47000 とのコサイン類似度が高い依頼データについて，依頼内容，解決方法，依頼 No.47000 とのコサイン類似度の一覧を表 1 に示す．依頼 No.47000 は，「教室に備え付けられている PC の画面をプロジェクタで投影できない」という依頼内容であり，「マルチメディア操作卓主電源とプロジェクタの電源をオンにする」ことで解決している．

依頼 No.47000 と同様に「プロジェクタで投影できない」という依頼内容の依頼データは，10 件中 8 件であった．特に上位 3 件の依頼内容は，プロジェクタで投影させたい機材が備付 PC である点も同じであった．また，「プロジェクタの電源をオンにした」ことで解決された依頼データは，10 件中 6 件であった．

上位 2 件の依頼データは，依頼内容と解決方法ともに，依頼 No.47000 と，依頼が発生した教室名は異なるがほぼ同じ内容であり，類似度の高さを確認できた．3 位以下の依頼データも，依頼内容，解決方法ともに似ている内容のものが多かったが，7 位の依頼 No.48923 や，10 位の依頼 No.49742 のように，関連性は高いが内容に若干ずれがあるものも確認できた．

表 1. 依頼 No. 47000 とのコサイン類似度が高い依頼データ上位 10 件

依頼 No.	依頼内容の要約	解決方法の要約	コサイン類似度
47000	教室備付 PC の画面をプロジェクタで投影できない	マルチメディア操作卓主電源とプロジェクタの電源をオンにした	_
52054	教室備付 PC の画面をプロジェクタで投影できない	プロジェクタの電源をオンにした	0.9884686470031738
46130	教室備付 PC の画面をプロジェクタで投影できない	プロジェクタの電源をオンにした	0.9868957996368408
37067	教室備付 PC の画面をプロジェクタで投影できない	支援組織のスタッフが教室につくと自己解決されていた	0.9863526225090027
48784	プロジェクタがつかない	マルチメディア操作卓主電源とプロジェクタの電源をオンにした	0.9852731227874756
37349	貸出 PC の画面をプロジェクタで投影したい	マルチメディア操作卓主電源とプロジェクタの電源をオンにした	0.9851259589195251
48923	プロジェクタの電源が切れない	リモコンでプロジェクタの電源を切った	0.983698844909668
35931	プロジェクタが投影されない	マルチメディア操作卓を再起動した	0.9836828708648682
29565	書画カメラがプロジェクタに投影できない	書画カメラを再起動した	0.9836283922195435
44118	プロジェクタがつかない	プロジェクタの電源を入れた	0.9835773706436157
49742	教室備付 PC の画面が手元のモニタに映らない	マルチメディア操作卓を再起動した	0.983302891254425

【実験例 2】依頼 No. 50979 とのコサイン類似度が高い依頼データ上位 10 件を抽出

依頼 No.50979 とのコサイン類似度が高い依頼データについて，依頼内容，解決方法，依頼 No.50979 とのコサイン類似度の一覧を表 2 に示す．依頼 No.50979 は，「Thunderbird 上で同じメールを何度も受信する」という依頼内容であり，「npop でメールサーバの当該メールを削除した」ことで解決している．なお，Thunderbird はオープンソースのメールクライアントソフトであり，npop はメールサーバ上のメールを表示するソフトのことである．

依頼 No.50979 と同様に「Thunderbird 上で同じメールを何度も受信する」という依頼内容の依頼データは，10 件中 5 件であった．また，「npop でメールサーバの当該メールを削除した」ことで解決された依頼データは，10 件中 7 件であった．

依頼 No.50979 と比較し，依頼内容より解決方法の方がほぼ同じ内容である依頼データが多く見受けられたが，解決方法が同じで依頼内容が異なる 6 位の依頼 No.48740 や 10 位の依頼 No.48772 も，実際発生した問題は依頼 No.50979 と同じである．

また実験例 2 でも，2 位の依頼 No.44916 や，5 位の依頼 No.32530，9 位の 53030 のように，内容にずれがあるものも確認できた．

表 2. 依頼 No. 50979 とのコサイン類似度が高い依頼データ上位 10 件

依頼 No.	依頼内容の要約	解決方法の要約	コサイン類似度
50979	Thunderbird 上で同じメールを何度も受信する	npop でメールサーバの当該メールを削除した	_
51981	Thunderbird 上で同じメールを何度も受信する	npop でメールサーバの当該メールを削除した	0.9859085083007812
44916	Thunderbird 上で受信中のまま先へ進まない	受信完了まで待った	0.9825268387794495
52020	Thunderbird 上で同じメールを何度も受信する	npop でメールサーバの当該メールを削除した	0.9804210662841797
48411	Thunderbird 上で同じメールを何度も受信する	npop でメールサーバの当該メールを削除した	0.9766550660133362
32530	Thunderbird 上でメールが見れない	Thunderbird を再起動した	0.9757219552993774
48740	Thunderbird 上でメールが受信できない	npop でメールサーバの当該メールを削除した	0.9743168354034424
51670	Thunderbird 上で同じメールを何度も受信する	npop でメールサーバの当該メールを削除した	0.9734225273132324
50951	Thunderbird 上で同じメールを何度も受信する	npop でメールサーバの当該メールを削除した	0.973110020160675
53030	Thunderbird 上で迷惑メールを多数受信する	迷惑メール多発時期であることと，迷惑メールの削除を案内した	0.9728596210479736
48772	同じメールを何度も受信してほかのメールが受信できない	npop でメールサーバの当該メールを削除した	0.9721630811691284

【実験例 3】「メール」というキーワードを含む依頼データ間のコサイン類似度

ここではまず,「メール」というキーワードを含む依頼データとして依頼 No.52641 をもととし，その他の「メール」を含む依頼データを任意で取り上げ，依頼 No.52641 とのコサイン類似度を算出した．依頼 No.52641 は,「学習院ドメインのメールアドレスに届くメールの転送方法を教えてほしい」という依頼内容であり,「転送方法の具体的な手順を教えた」ことで解決している．学習院ドメインのメールアドレスとは，学習院で勤務する教職員が学習院内のシステム部門に申請することによって付与される，個々のメールアドレスを示す．

依頼 No.52641 と同じ依頼内容，解決方法の依頼データである依頼 No.16359，40060 のコサイン類似度を測ったところ，高い値を得られた．また，依頼 No.51815，50979，51411 は,「メール」というキーワードを含み，依頼内容もメールに関連する依頼データだが，依頼 No.52641 とは異なる内容の依頼であり，これらとのコサイン類似度は比較的低い値となった．さらに，依頼 No.51276 は,「メール」というキーワードを含むが，依頼内容はメールとは関連性のないものであり，解決方法の伝達手段としてメールを利用した依頼である．この依頼と依頼 No.52641 とのコサイン類似度は非常に低い値となった．

表 3. 依頼 No.52641 と「メール」を含む依頼データとのコサイン類似度

依頼 No.	依頼内容の要約	解決方法の要約	コサイン類似度
52641	学習院ドメインのメールアドレスに届くメールの転送方法を教えてほしい	転送方法の具体的な手順を教えた	−
16359	学習院ドメインのメールアドレスに届くメールの転送方法を教えてほしい	転送方法の具体的な手順を教えた	0.975090742111206
40060	学習院ドメインのメールアドレスに届くメールの転送方法を教えてほしい	転送方法の具体的な手順を教えた	0.956198053098
51851	迷惑メールを頻繁に受信するのが心配	迷惑メールは添付ファイルを絶対に開かず削除するようお伝えした	0.943197821713
50979	Thunderbird 上で同じメールを何度も受信する	npop でメールサーバの当該メールを削除した	0.906692519741
51411	Thunderbird 上で受信メールの文章をコピーしたい	コピーしたい文章を選択し、右クリック→コピーを伝えた	0.844182030266
51276	液晶モニターを３つ目のモニターとして接続する方法を知りたい	メールでモニター接続のための機器が必要になることを案内した	0.58200212602

【考察】

実験例 1，2 ともに，対象の依頼データと依頼内容，解決方法ともに一致した依頼データは上位 10 件中 50%以上の割合で抽出することができ，依頼内容，解決方法のどちらかが一致した依頼データは上位 10 件中 70%以上の割合で抽出することができた. 特に実験例 2 で示したように，依頼内容の記述に差異があっても，根本の問題が同じならば同じ方法で解決することが可能であり，これは類似した依頼を高速に発見したいという本研究の目的に合致する.

一方で，実験例 1，2 ともに，依頼内容，解決方法とも対象の依頼データと内容が異なる依頼データも抽出された．これは，doc2vec によって変換された各依頼データの数値ベクトルが近かったためである．doc2vec を適用するデータは依頼データを形態素解析器に掛けた結果の動詞と名詞を羅列したデータである．そのため，依頼内容が異なっていても，多くの動詞と名詞が順序を保存しながら出現すれば，似た数値ベクトルが生成される可能性が高い. 例えば実験例 2 において，2 位の依頼 No.44916 は，内容が異なるにもかかわらず，比較対象の依頼データとの類似度が高かった．しかし，実際の依頼データを確認すると，両者は「Thunderbird」「メール」「受信」などの単語が共通していることや，「メールの受信がうまくいかない」という観点では同じと言えることから，類似度が高いことも納得できる.

実験例 3 で取り扱った依頼データは，すべてが「メール」というキーワードを含むが，その中でも依頼内容と解決手段が似ている依頼データ間の類似度は高く，異なる依頼データ間の類似度は比較的低いという結果になった．従来のキーワードベースの検索では，依頼データに優先度をつけられず，検索時に依頼データをリスト表示させる順番は新規依頼の順番であった．実験例 3 の結果により，類似度を測ることで依頼データに優先度をつければ，例えば類似度の高い順に依頼データをリスト表示させることで，得たい情報を発見しやすくなることを示せた.

4. おわりに

本研究では，依頼対応記録表に記載された依頼データ間の類似度を計算し，それを用いて高速に目的の依頼データに到達するためのシステムの構築に取り組んだ．これを達成するための準備として，長年の間使用されてきた既存のシステムを大幅に見直すことで，プログラムの可読性やシステム拡張性を改善した．これにより，今後の依頼対応記録表の機能拡張を目的とした開発を効率的に進められると考えられる．依頼データ間の類似度計算については，様々なアプリケーションにおいて用いられている代表的な手法である doc2vec を利用し，計算機実験においてその有効性を確認することができた.

なお，本研究では 2 節で示した依頼対応記録表の改良に膨大な時間を費やしたために，研究期間内に高度な情報検索機能の実装を終えることができなかった．そこで今後の課題

として，得られたデータ分析結果を用いた効果的な情報検索機能の構築を行う必要がある．例えば，依頼が来たときにその問題点を記述することで，自動的に過去の依頼データとの類似度を計算し，それを元に解決方法を推薦するシステムなどが考えられる．また，今回使用した doc2vec は単語間の類似度を計算する機能が含まれているため，これを使用することで，表記の揺れなどを考慮したキーワード検索の構築も可能である．さらなる展開としては，現在では手動で依頼データをカテゴリ分けしているが，機械学習の手法を取り入れることでその半自動化に取り組んだり，頻出する依頼をうまく抽出することで，その対応方法を文書化したマニュアルの作成にも取り組みたいと考えている．

また，本研究では取り扱わなかったが，各依頼が問題解決の上でどれだけ重要であるかを指し示すような重要度計算は非常に有効であると考えている．これは，ある依頼を別の依頼データを参考に解決した場合や類似した依頼があった場合には，依頼データ間にリンクを張るような記述が依頼対応記録表上で多くなされているため，これらから参照関係のネットワークを構築し，PageRank [12]を計算することで，依頼データの重要度が計算できるであろう．この重要度も今回実装したシステムに組み込むことで，より高度な検索機能が実現できるのではないかと考えられる．

謝辞

本研究は，平成 28 年度計算機センター特別研究プロジェクトの助成を受けたものである．

参考文献

[1] 入澤寿美，市川収，小倉統，松本喜以子: 学習院コンピュータシステム支援組織を学内に設置した効果について．学習院大学計算機センター年報, Vol.21, pp.55-79, 2000.

[2] 因幡哲男，黒崎茂樹，水上悦雄，横山悦郎，入澤寿美: 依頼対応の効率化と実績評価について． 学習院大学計算機センター年報, Vol.24, pp.64-78, 2003.

[3] 浦上大輔，海老澤賢史，佐藤友彦，勝野弘康，久保山哲二，横山悦郎，入澤寿美，坂本孝治郎: 学習管理システムによる自己学習型マルチメディア教育支援体制．学習院大学計算機センター年報, Vol.31, pp.2-11, 2010.

[4] 勝野弘康，鏑木崇史，山口健二，久保山哲二，横山悦郎，入澤寿美，坂本孝治郎: IT 活用教育支援業務組織における効率的なデータベースの開発と運用方法の検討． 学習院大学計算機センター年報, Vol.32, pp.2-13, 2011.

[5] CodeIgniter: CodeIgniter Web Framework https://codeigniter.com/ (2017.8.22)

[6] jQuery: jQuery https://jquery.com/ (2017.8.22)

[7] Q. Le, T. Mikolov: Distributed Representations of Sentences and Docoments. In Proc. of ICML 2014,　pp. 1188 - 1196, 2014.
[8] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean: Distributed Representationsof Words and Phrases and their Compositionality. In Proc. of NIPS 2013, pp. 3111 – 3119,2013.
[9] 塩野剛志: 文書の分散表現と深層学習を用いた日銀政策変更の予想．第 16 回人工知能学会金融情報学研究会，2017.
[10] gensim: Topic modeling for humans. https://radimrehurek.com/gensim/ (2017.8.21)
[11] MeCab: Yet Another Part-of-Speech and Morphological Analyzer. http://taku910.github.io/mecab/ (2017.8.22)
[12] S. Brin, L. Page: The Anatomy of a Large-Scale Hypertextual Web Search Engine. In Proc. of WWW 1998, pp. 107 - 117, 1998.