自然言語処理 AI と知識グラフを活用した商材情報検索の効率化

Efficient Product Information Search using Natural Language Processing AI and Knowledge Graph

高　橋　諒　子

要　約　DNP では自然言語処理技術と知識グラフを活用し，曖昧な知識でも多岐にわたる商材を検索できるサービス（「スマートレファレンスサーチ」と呼称）を開発している．このシステムは商材 DB の他に商材を説明するキーワードタグである「商材関連用語」の DB を持つことで，入力テキストに関連する可能性の高い「商材関連用語」をサジェストして商材の検索をアシストする．

ユーザテストを実施した結果，「提示されるサジェストに納得感はあるか」「直感的に使用できるか」等の設問で高い満足度が得られた．今後は時系列を含めたより複雑なデータ構造を構築するとともに，商材検索以外のサービスにも本システムを発展させていく．

Abstract　DNP is developing a product search service named “Smart Reference Search”. Using natural language processing technology and knowledge graph database, this service provides a wide range of product information even if user has only vague knowledge of the products. This service has database of product and “product-related terms”, which are keywords that describe the products, and suggests these terms to help product searching.

The user’s test showed a high level of satisfaction with question such as “Are you satisfied with the suggested terms?”, “Is it intuitive to use?”, etc. In the future, we will develop this system with more complex data structures including time series data, and apply it to other than product search service.

1. は　じ　め　に

顧客のニーズに合致する商材の提案や商材の組み合わせによる新たな価値の創造には，幅広い商材の知識が必要とされる．企業の規模が大きくなるにつれ自社の扱う商材を広域にわたって把握することは困難となるため，全社的な商材検索システムの活用が求められる．しかし，単純な文字列一致による検索を行うシステムでは，商材の正式名称を知らず，専門知識を持たない場合に未知の商材に到達するのが困難となる．社内知識が少ない従業員こそ商材検索システムを必要とするため，曖昧な知識でも検索できたり，似た商材からサジェストが行われたりするような商材検索サービスが必要となってくる．

また，商材に関する情報は，社外に向けては紹介 web ページ・ニュースリリース・投稿サイト上の動画，社内に向けては提案資料・導入実績資料など多岐にわたる．各所に集積したこれらの情報に一度の検索で到達できるシステムにより，業務の効率化を見込める．

大日本印刷株式会社（以降，DNP）では業務効率化支援システムとして，DNP 社員向けの次世代商材検索システムである「スマートレファレンスサーチ」の開発を行っている．これは，自然言語処理技術および知識グラフを用いることで，正確な商材名や専門用語を知らないユー

ザでも簡単に商材を検索できることを目指したシステムである．

本稿では，2章で知識グラフの一般的な特徴，3章でスマートレファレンスサーチの機能概要，4章で自然言語処理 AI を用いたグラフ DB の構築手法，5章でユーザテストとその結果，6章で総括と将来像について述べる．

2. 知識グラフについて

スマートレファレンスサーチは，知識グラフを用いたデータベースに商材情報を格納している．商材の検索が発生すると，このデータベースに対してクエリ言語を用いた問い合わせを行い，検索結果を取得する．

本章では，知識グラフに関する基礎知識を説明する．

2.1 知識グラフの概要

知識グラフは多様な知識の関係性をグラフ構造で表す技術であり，検索・質問応答・推薦・意思決定支援など，様々なアプリケーションに応用されている．例えば Google 検索を行った際に表示される要約情報や関連情報は，膨大な情報を格納した知識グラフから提供される[1]．

知識グラフを記述するデータ方式として，RDF[2]が広く用いられる．RDF は事物や概念についてのメタデータ（情報についての情報）を記述するための汎用的な手法である．本システムでもデータ形式として RDF を採用している．

RDF では，データの構造を主語（Subject）・述語（Predicate）・目的語（Object）のセット（トリプル）の集合で表す．一つのモノを示すエンティティは固有の URI を持ち，主語・述語どちらにもなりうる．このため，あるトリプルにおいて目的語であるエンティティが別のトリプルでは主語となり，データ全体が繋がったグラフ構造をとる．

RDF データの一例を図1に示す．

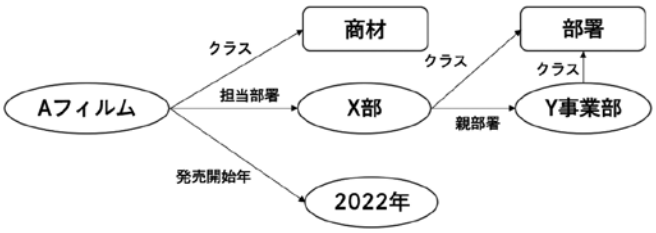

図1　グラフ構造の例

図1のグラフには，以下の内容が含まれる．一つの矢印が一つの内容に対応している．

1）A フィルムは商材クラスに属する．
2）A フィルムの担当部署は X 部である．
3）A フィルムの発売開始年は 2022 年である．
4）X 部は Y 事業部に属する．
5）X 部は部署クラスに属する．
6）Y 事業部は部署クラスに属する．

RDF に対しクエリ言語 SPARQL[3]を用いることで，指定したトリプルに一致する対象を検索することができる．上記の例を用いれば

・A フィルムの担当部署は何か

・Y 事業部に属する部が担当している商材は何か

といった内容を検索することができる．

さらに，RDF では推論を行うことができる．推論はルールに基づく論理的な演算であり，反射律や対称律，推移律などを扱える．

2.2　グラフ DB のメリット・デメリット

知識グラフを用いたデータベース（以降，グラフ DB）の利点として，大規模データに対し深い検索を行う際，RDB と比較して探索時間を削減できることが挙げられる．データ間の関係を何階層も辿っていく深い検索を行う場合，一般的な RDB では複数回のテーブル結合処理が必要になるため処理時間が増大する．一方グラフ DB では全ての情報が主語・述語・目的語のトリプルで管理され，関係を示す情報がデータごとに保持されているため，検索時間を小さくすることができる．

さらに，グラフ DB は 1 対 1 の関係の積み重ねであるために拡張が比較的容易であり，DB 制作時に想定したものを発展させたサービスが作成できる．

一方で欠点として，データ生成コストが高いこと，拡張を可能にするためにクラスをどう定義するかを事前によく検討する必要があることが挙げられる．

本システムでは商材情報をメインとして扱っているが，この情報は企業内部の部署や担当者等のデータと紐付くものであり，将来的にプロジェクト計画・人事データの管理に拡張させていくことができる．これらの情報は時間の経過とともに変動するものであるため，グラフ DB の利点が生かせる題材であるといえる．

3.　スマートレファレンスサーチについて

本章では，スマートレファレンスサーチの機能概要を述べる．

3.1　サービスの概要

スマートレファレンスサーチは DNP の商材情報検索を提供する DNP 社員向けサービスであり，既存の商材検索システムでの課題面を補完するために開発が進められている．商材知識の乏しいユーザこそ商材検索サービスを必要としている，という観点における既存システムの課題を以下に挙げる．

【既存システムの課題】

・名称・概要・紹介文章に対する完全一致のテキスト検索のみが可能である．このため，正確な名称を知らないと検索できず，意味的に関連が無くてもテキストが一致する場合に誤ヒットしてしまう．例えば「AI」で検索すると「detail」や「aichi」などの文字列をいずれかの情報に含む対象がヒットする．

・検索用キーワードの設定が可能だが，設定は商材担当者に任されるため商材によってキーワードの量が大きく異なる．

これらの課題を克服するため，スマートレファレンスサーチでは「商材関連用語」（商材に紐付けられたタグ）を用いた検索モード（3.2節を参照）を搭載した．

さらに，既に商材についてある程度知識を持つユーザにとって使いやすい「商材名による検索」「部署名による検索」の検索モード（3.3節を参照）を加えている．

3.2　商材用語による検索

「商材用語による検索」は，自由にテキストを入力して「商材関連用語」から商材を検索する，一種のタグ検索機能である．この検索モードの特徴として，入力ワードに加えて同義語や関連する用語を自動で追加し検索できる．検索の流れを図2に示す．

図2　検索の流れ

検索ボックスにテキストを入力すると，入力テキストに関連する用語のリストが表示される．

（ｉ）このリストから意図に近い用語を選択すると，選択用語とその用語に関連する用語を用いて商材が検索される．

（ii）用語を選択せず検索ボタン等で検索を実行すると，リスト内の用語をまとめて検索用語として商材が検索される．

図3に最終的な検索ワードの例を示す．

図3　「セキュリティ」を入力テキストとした場合の検索ワード

この例では，「セキュリティ」という文字列を含む商材関連用語である「情報セキュリティ」「セキュリティ対策」……「情報セキュリティ対策」に加え，「セキュリティ」と内部で結びついている「サイバー攻撃」「セキュア」「安全管理」を検索用語に加えている．

入力テキストから商材関連用語を検索する段階では，形態素解析による名詞等の抽出・表記ゆれ処理・ヒットした商材関連用語のスコアリングとランク付けを行っている．この処理により，入力テキストが単語ではなく文章の場合にも対応し，より入力意図に即したサジェストが

可能となっている．

検索結果画面では，以下の機能を提供している．

・外部サイトの商材紹介へのリンク機能
・商材に紐付く商材関連用語から別の商材を検索する機能
・商材用語によって検索結果を絞り込む機能
・検索に用いた用語の解説が閲覧できる機能
・検索には使われなかったが，関連すると予測できる用語の列挙

これらの機能により，検索結果から新たな商材関連用語を知ったり，その用語に関係する商材を調べたりすることができる．

3.3　商材名・部署名による検索

「商材名による検索」「部署名による検索」は，商材名やその商材を担当する部署名を検索対象にした一般的な検索機能である．現在は商材名・部署名の知識を持ったユーザを対象とし，「商材名による検索」では事前に登録された商材の正式名称による検索，「部署名による検索」ではプルダウンリストでの選択を用いた部署による検索を搭載している．

将来的には通称や旧名称による検索を行えるアップデートを検討しており，「うろ覚えの商材名・部署名」で商材を検索できることを目指している．

4.　グラフ DB の概要と構築

スマートレファレンスサーチでは，商材情報 DB に加え，商材関連用語のグラフ DB および社内知識のグラフ DB を構築している．このグラフ DB の構築には，社内に存在するグラフ形式ではない商材情報の DB を利用した．

本章では，これらのグラフ DB の概要および構築手法を述べる．

4.1　商材関連用語グラフ DB の概要

商材関連用語グラフ DB は商材に付与する汎用的なタグの DB であり，特定企業に限らない幅広い商材に使用できる．

3.2 節で述べた機能を実現するため，この DB に商材関連用語同士の相互関係を構築した．具体的には，商材関連用語は以下のプロパティを持つ．

・正式名称（文字列）
・別名称（文字列）
・上位語（他の商材関連用語を指す URI）
・下位語（他の商材関連用語を指す URI）
・関連語（他の商材関連用語を指す URI）

ここで，上位語は上位概念である用語，下位語は下位概念である用語を示す．例えば「球技」は上位語として「スポーツ」を持ち，下位語として「野球」を持つ．

これらのプロパティを利用することで，直接関連する可能性が高い「入力テキストが別名称として登録されている用語の上位語/下位語/関連語」や，間接的に関連する可能性を持つ「関

連語の関連語」や「同じ上位語を持つ他の用語」を検索することができる.

本システムでは，約 2 万語を収録した商材関連用語グラフ DB を構築した.

4. 2　商材関連用語グラフ DB の構築

商材関連用語グラフ DB の構築手順を以下に示す.

1）用語の抽出
2）抽出した用語の構造化
3）手作業による編集

1）では，既存の商材情報 DB から得られるテキスト群（商材名・キーワード・概要文・背景/仕様/特徴などの説明文）から利用できる単語を抽出した．テキスト群に対し mecab-ipadic-NEologd[4]を辞書として形態素解析を行い，固有名詞を抽出し用語候補リストを作成した．また，専門用語自動抽出システム「TermExtract」[5]を利用し，形態素解析辞書に含まれない用語を統計的に抽出して用語候補リストに追加した．作成した用語候補リストに含まれる各語に対して TF・IDF 法[6]等を用いてスコア付与を行った．さらに商材や技術に関する語の判定を別途手作業で行い，最終的な用語リストを作成した.

2）では，1）で抽出した用語間の関係（同義語関係・上位語 - 下位語関係・関連語関係）の RDF データを作成した．用語間の関係は NDL Authorities[7]・NDC9[8]の図書館分類系の分類データベース・Wikipedia[9]カテゴリを中心とした，外部の構造化済みデータを基に構築した．固有名詞については人手による構造化を行った.

3）では，2）の結果を人手で編集した．この工程では明らかな誤りを修正する他，以下の基準を用いた.

・複合語を避け，できるだけ最小単位の用語を採用する．例えば「AI 翻訳」ではなく「AI」「翻訳」のように分解する.
・専門用語より一般的に使われる用語を優先し，専門用語はできるだけ別名でカバーする.

4. 3　コンテンツへの単語の関連付け

商材関連用語を用いて商材検索を行うためには，適正な用語を商材に紐付けることが必要である．商材グラフ DB と商材関連用語グラフ DB の紐付け手順を以下に示す.

1）商材コンテンツ（商材タイトルおよび商材説明文）を商材関連用語の名称で単純一致テキスト検索し，合致するものを付与
2）付与された用語に対し，人手での編集を行う

人手での編集作業では，大きく以下の 2 点を行った.

ⅰ）付与すべき用語が漏れていた場合に追加
ⅱ）付与された用語が多すぎる場合の絞り込み

今回はプロトタイプ開発であるため，その商材に固有の特徴を表す用語を優先することを基準に人手での編集を行ったが，機械作業の比重をより大きくする手法を検討中である．

ⅰ）について，単語間の関連性や類似度を用いて付与すべき語のサジェストを行うツールを現在開発している．

ⅱ）について，用語の優先度を設定するアルゴリズムを検討中である．例えば，既に商材に多く付与されている語の優先度を高く設定することによって，検索したユーザに商材間の関連性を提示できると考えられる．

4. 4　社内知識グラフ DB

「商材名による検索」「部署名による検索」においては，企業内の情報が検索のキーとなる．こうした検索では，正式名称以外に通称や旧名称が検索キーに用いられることが考えられる．

今回開発した社内知識グラフ DB では，商材エンティティは正式名称に加え通称を持ち，両方を商材名称として利用できる．また，重層化された階層を表現するため，部署エンティティは自身の上位に位置する部署・下位に位置する部署をプロパティに持つことができる．このプロパティにより，様々な階層の部署名を使った部署名検索が可能となる．

将来的な発展性として，組織名の旧称や遷移，属する社員などの情報を社内知識グラフ DB に持つことで，商材を担当していた部署や社員の追跡を組織改編後でも実現することを検討している．

5.　ユーザテスト

本章では，スマートレファレンスサーチに対して行ったユーザテストについて述べる．

5. 1　ユーザテスト概要

本システムの有用性を検証するため，DNP グループ従業員を対象に利用感アンケートを行うユーザテストを実施した．テストでは特に検索する対象を指定せず，回答する従業員それぞれが業務上必要な内容の検索を行うこととした．このユーザテストを異なる領域の従業員を対象に実施した．

5. 2　ユーザテスト結果

DNP グループ従業員 162 名よりアンケート回答が得られた．図 4 に一部抜粋を示す．

自然言語処理 AI によるタグ付けの評価を問う設問である「入力文章とサジェストに挙がる用語の関連性に納得感があったか」に対しては，「納得感があった」「どちらかというとあった」の回答が計 64％と，半数以上のユーザから良い評価が得られた．

さらに，「直感的に使えると感じたか」に対する「直感的に使えた」「おおむね直感的に使えた」の回答が計 77％，「探し方が簡単だと感じたか」に対する「簡単だった」「どちらかというと簡単だった」の回答が計 78％，また社内サービスとして追加された場合の利用意向確認では，「とても利用したい」「どちらかというと利用したい」と回答したユーザが全体の 84％を占め，スマートレファレンスサーチの主な機能が若年層からベテラン層までの幅広い年代に支持されることが分かった．

一方で改善希望として，一部機能の使用方法が直感的に分かりづらいという評価が得られた．

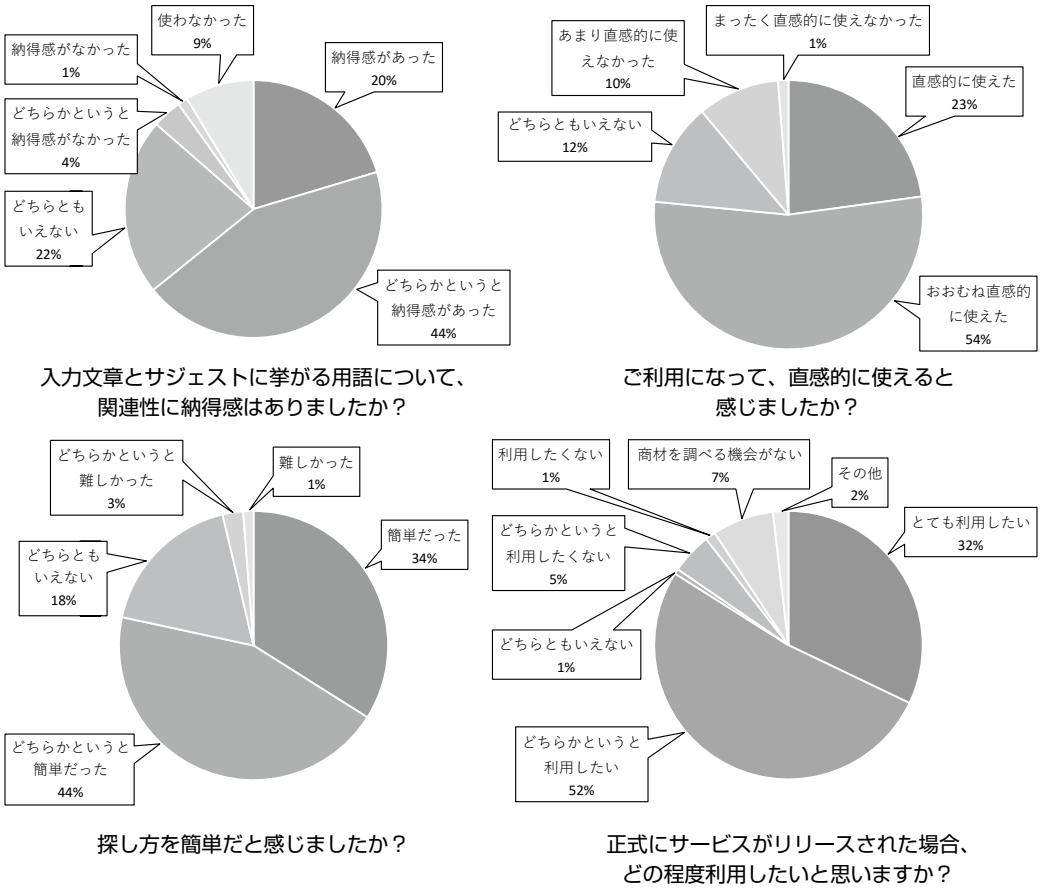

図 4　アンケート結果の一部抜粋

ユーザテストの結果，多くのユーザからスマートレファレンスサーチが直感的に使える商材検索システムとして好意的な評価を得られることが示せた．

6. お わ り に

本稿では，自然言語処理技術と知識グラフを用いた商材検索サービスであるスマートレファレンスサーチについて紹介した．最後に今後の展望を述べる．

現状のグラフ DB の構築においては，少なくない部分を人手の作業に依存している．人間による作業を完全に撤廃することは難しいが，機械学習による手法を加えることで不可欠な部分だけを人間による判断に委ねることができるようになるだろう．現在，DB 作成および検索の一部機能については採用技術の検討および選定を行っている．

より高い目標として，商材検索機能を超えて社内の様々な情報を横断するサジェスト・提案サービスに本サービスをアップデートしていくことを検討している．属人的になりやすい社内情報の検索性を上げ，必要な情報を必要な時に引き出せるようにすることで，さらなる業務効率化を図ることができるだろう．

参考文献
［1］ “Google のナレッジグラフの仕組み，”［オンライン］. Available: https://support.google.com/knowledgepanel/answer/9787176?hl=ja#
［2］ P. Hayes, “RDF Semantics, Technical report,” W3C Recommendation, 2004.
［3］ E. a. S. A. Prud’hommeaux, “SPARQL Query Language for RDF,” W3C Recommendation, 2008.
［4］ T. Sato, “Github:mecab-ipadic-neologd,”［オンライン］. Available: https://github.com/neologd/mecab-ipadic-neologd
［5］ “TermExtract，専門用語（キーワード）自動抽出システム”［オンライン］. Available: http://gensen.dl.itc.u-tokyo.ac.jp/
［6］ G. Salton, “Automatic Text Processing,” Addison-Wesley Publishing Company, 1989.
［7］ 国立国会図書館，“Web NDL Authorities,”［オンライン］. Available: https://id.ndl.go.jp/auth/ndla
［8］ 日本図書館協会，“日本図書館協会　NDC データ（NDC8 版および 9 版），”［オンライン］. Available: https://www.jla.or.jp/committees/bunrui//tabid/789/Default.aspx
［9］ “Wikipeda メインページ，”フリー百科事典 ウィキペディア日本語版，2022 年 12 月 21 日 17 時（日本時間）現在での最新版を取得. https://ja.wikipedia.org/wiki/%E3%83%A1%E3%82%A4%E3%83%B3%E3%83%9A%E3%83%BC%E3%82%B8

※　上記参考文献に含まれる URL のリンク先は，2022 年 12 月 21 日時点での存在を確認.

執筆者紹介　高　橋　諒　子（Ryoko Takahashi）
2019 年大日本印刷株式会社入社．AB センターにて自然言語処理技術，知識グラフを活用したサービスの開発に取り組む.