化学データ利活用サービスに向けたナレッジグラフ構築
―化合物と製造方法の理解―

Knowledge Graph Construction for Chemical Data Utilization Service

株式会社富士通研究所
田中　一成

人工知能研究所　ナレッジテクノロジー PJ、テキストマイニング技術の研究、特許読解支援システムの開発、ナレッジグラフの研究に従事。

tanaka.kazunari@jp.fujitsu.com　TEL 044-754-2652

株式会社富士通研究所
池田　紀子

R&D 戦略本部　企画部、技術士（応用理学／総合技術監理部門）、電子・光学デバイス材料の設計および分析、並列処理および分子モデリングの研究、特許読解支援システムの開発に従事。

1　はじめに

一般社団法人 日本化学工業協会のアニュアルレポートによると、プラスチック製品とゴム製品を含めた日本の化学工業の 2015 年の出荷額は 44 兆円、その付加価値は 16 兆円である。日本の産業別状況では、いずれも輸送用機械器具製造業に次ぐ第 2 位で、日本の経済に大きく貢献している。また、2016 年の従業者数は 87 万人にのぼり、雇用面でも国民の生活を支えている。しかし、製造しているものがあまりにも多岐にわたるため、その姿が見えにくい[1]。そのため、化学・バイオ分野の特許や論文等から、原料や助剤等の化合物名を抽出し、最終生成物の製造方法を理解するには、高度なスキルに多大な時間と労力が必要である。化学・バイオ分野の特許には、化学物質や製造方法の特定が容易でない発明がある。さらに、用途や物性発明のように、化学・バイオ分野に特徴的なものが多くあり、他の分野とは異なる特質がある。化学分野の特許（以下、化学特許）には、論文以上に詳細で現実的な製造の実施例が埋もれている。これまで取り組んできた化合物名の抽出方法をさらに進化させると共に、製造方法情報を特許文書から抽出する方法を考案した。化合物や製造方法についての知識をナレッジグラフとして連携させることで、知識の利活用拡大と調査の負担削減の可能性が高まる。

第 2 節では、ナレッジグラフの基本的な考え方について説明する。第 3 節では、化学情報のアプリケーションの 1 つとして、検索・読解支援システムについて提案する。第 4 節では、化合物のナレッジグラフについて示す。第 5 節では、化合物の製造方法情報の抽出と製造方法のナレッジグラフについて説明する。

2　ナレッジグラフ

ナレッジグラフは、グラフ形式で表現された知識ベースであり、実世界における知識を表現、統合、解析することが可能である。

2.1　ナレッジグラフ技術

ナレッジグラフ技術は、各分野で扱われる多種多様な知識データを統一的なグラフ形式のデータに変換し、統合的な一つの知識ベース（ナレッジグラフ）として扱えるようにする。従来、個別に扱われてきた知識を統合することで、今まで記述が困難であった複雑な関係性など

を記述できるようになった[2]。

2.2　ノードとエッジ

ナレッジグラフの基本的な要素は、「ノード」と「エッジ」である。「ノード」（円）が概念、「エッジ」（矢印）が関係を表す。例えば、図 1 の上段は化合物「2,2-ビス（4-ヒドロキシフェニル）プロパン」の名前の 1 つが「ビスフェノール A」であるという関係を表している。図 1 の中段は「2,2-ビス（4-ヒドロキシフェニル）プロパン」の化学構造が右の構造式で表せることを示している。また、図 1 の下段は、「ビスフェノール A」が「特願 2000-xxxxxx」で示される特許文書中に現れるという関係を示している。

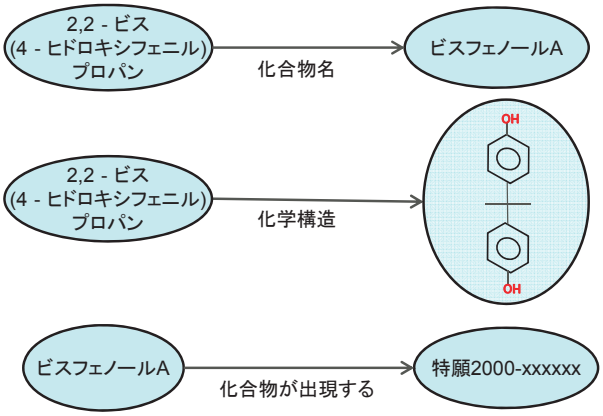

図 1　関係の例

「ノード」や「エッジ」に、名前または場所などを識別する識別子、URI（Uniform Resource Identifier）を使用する。ナレッジグラフでは、URI によって示された項目間に意味を表す関係を定義することでデータを蓄積する。例えば、図 1 の中段の場合、「2,2-ビス（4-ヒドロキシフェニル）プロパン」の構造式を示す場合には、ナレッジグラフの中に画像データを登録しなくても、構造式を公開しているアドレスを URI として持っておくことにより、必要に応じで画像データを参照できるようになる。

ナレッジグラフを利用した検索は、ノードやエッジを指定したクエリによって行う。例えば、図 1 の下段のグラフを利用して、「ビスフェノール A」というノードの URI と「化合物が出現する」という関係を表す URI を指定したクエリにより、「特願 2000-xxxxxx」という特許を得ることができる。さらに、図 2 に示すように、「2,2-ビス（4-ヒドロキシフェニル）プロパン」の名前が「ビスフェノール A」であるという関係と組み合わされることで、「2,2-ビス（4-ヒドロキシフェニル）プロパン」を入力して検索を行う場合でも、「2,2-ビス（4-ヒドロキシフェニル）プロパン」の名前の 1 つが「ビスフェノール A」であり、「ビスフェノール A」が「特願 2000-xxxxxx」の中に現れるという関係をたどることができるため、「ビスフェノール A」で書かれている特許「特願 2000-xxxxxx」も漏れなく探すことができる。

図 2　ナレッジグラフによる検索

化学情報には、膨大な化合物に多数の化合物名があり、多種多様なデータが存在する。このため、ナレッジグラフにデータを蓄積しておき、必要に応じてエッジをたどってアクセスできるようにしておくことで、知識の利活用拡大に有効ではないかと考えられる。

3　化学情報検索・読解支援

化学情報の利活用支援のために、化学情報検索・読解支援システムを試作した。図 3 に、本システムの構成を示す。Japio YEAR BOOK 2016 では、従来の特許検索システムに読解支援を追加した構成について示した[3]。これに対し、今回は、検索機能として、全文検索エンジンに、ナレッジグラフを組み合わせる構成にした。

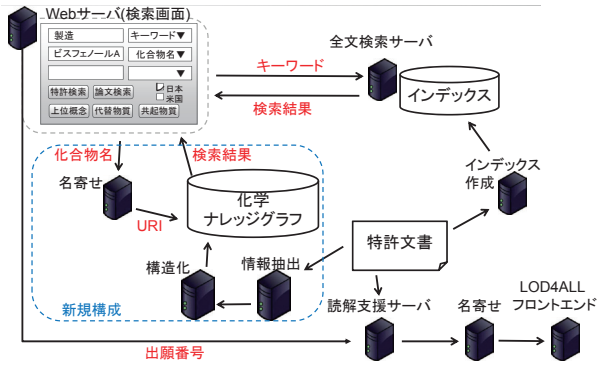

図 3　化学情報検索・読解支援システムの構成

3.1　全文検索エンジン

全文検索は文字列一致で検索するもので、特許検索などで広く使われている検索エンジンである。一般的なキーワードで検索するようなシステムでは、全

文検索エンジンが有効である。実績の高さで有名なElasticsearch を採用した。

3.2　ナレッジグラフ

全文検索を用いて、化合物名で検索を行う場合には、長い化合物名や別称、抽象的な名称の問題が調査の妨げになる。化合物は多くの別称を持つものが少なくなく、特に慣用名は文字列の類似性では判別できないものが多いため、全文検索のように文字列の部分一致で検索をすると、目的の化合物について書かれている化学特許が漏れてしまう原因になる。そこで、検索の対象となる化学特許文書から化合物名を抽出して名寄せを行った。そして、図 2 に示したように、URI に変換してから特許のURI と関連付けてナレッジグラフとして蓄積することにした。

さらに、化学情報には、例えば、「アルカン」のように 1 つの化合物を特定しない抽象的な名称も存在し、そうした抽象的な名称で検索したい場合もある。そこで、化合物名を抽象的な名称と関係付けをしておく。化学特許をもれなく検索するためには、検索条件として与えられた化合物名や抽象的な名称を URI に変換し、URI の接続関係をたどって検索する。

4　化合物のナレッジグラフ

4.1　化合物情報抽出

図 2 で示したような化合物と特許との関係を化学特許から構築するためには、化学特許中の化合物名を抽出する必要がある。Japio YEAR BOOK 2017 では、特徴パターンと機械学習を組み合わせることにより、化合物名を抽出する手法を提案した[4]。この手法では、抽出する化合物名候補の文字列だけを見て抽出するため、学習データを作りやすいという長所がある。しかし、新たに抽出しようとする化合物名候補の文字列が学習データの中に含まれない場合には判別できないという短所があった。

化合物名抽出の精度と漏れの少なさの両立を図る方法として、固有表現抽出を取り入れることにした[5]。固有表現抽出では、抽出する文字列前後の文脈を利用し、文字列前後の数単語の表記や品詞、化合物名表記のパターンを利用した抽出ルールを学習する。事前に人手で化合物名にタグを付けた学習データを用意しておき、機械学習によって、化合物名を抽出するためのルールを学習する。このように、辞書に登録されていない化合物名の候補を抽出することが可能となる。

この学習には distant supervision の考え方を応用することができる[6]。固有表現抽出では、多くの学習データが必要となるが、学習データを作成するには多くのコストがかかる。distant supervision では、予め化合物名の辞書を用意しておいて、辞書にマッチする文字列にタグを付けることで学習データを増やすことにより、より効率的に学習を行う考え方である。化合物名の場合、日化辞を辞書として活用できる[7]。抽出したい化合物名のバリエーションは膨大にあるが、頻繁に出てくる化合物名については、ある程度は化合物 DB に登録されていると考えられる。そのため、distant supervision の考え方は有効ではないかと考えられる[8]。

以前の方式と固有表現抽出には一長一短があるため、これらをうまく組み合わせて用いることにより精度向上が期待できる。

4.2　関係付け

抽出された化合物名は名寄せをして化合物を表す URI を付与する。化合物を表す URI と化学構造の情報を表す URI を関係付けたり、特許の URI とその特許に出現する化合物の URI を関係づけたりということをしていくことで、ナレッジグラフが組み上がっていく。化合物名の名寄せとは、同じ化学構造を持つ化合物名を同じと扱えるようにする処理である。例えば、「2,2-ビス（4-ヒドロキシフェニル）プロパン」と「ビスフェノール A」は、化合物名が異なる表記であるが、同一の化学構造を持つ化合物として関係づける。化合物のナレッジグラフでは、図 2 に示したように関係付けておいて、検索の際には「化合物名」という関係をたどるようにしておくことで、同一の化学構造を持つ化合物を扱うことができる。別称を持つ化合物が多く、名寄せの処理が重要になる。名寄せの方法としては、Japio YEAR BOOK 2017 で提案した言い換えルール[4] や機械翻訳技術の応用が可能と考えられる。特許や論文から抽出された化合物名は、名寄せされて URI に変換され、文献と化合物とのエッジが付けられる。

さらに、機能・用途、代替物質の候補といった情報も

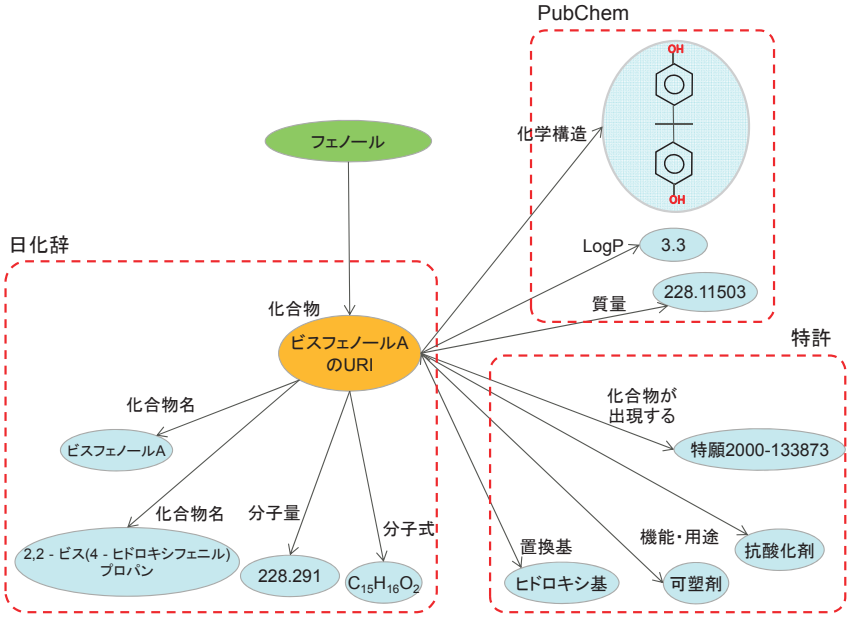

図４　化合物ナレッジグラフ

化学特許から抽出し、化合物の URI と関係付けて蓄積することもできる。機能・用途とは、化合物が何に使われるかという情報で、例えば、可塑剤や界面活性剤といったものがある。代替物質の候補としては、例えば、可塑剤という用途において、「ジブチルフタレート」以外に「ジオクチルフタレート」も使われるといった情報である。

これらの情報もナレッジグラフとして蓄積することによって、利用が容易になる。図４は「ビスフェノール A」に関係する情報のナレッジグラフを模式的に描いたものである。ナレッジグラフを用いることにより、化合物に関する様々な情報を蓄積することができる。ナレッジグラフを応用することによって、化合物の関連情報として、分子式や分子量のような基本的な情報、日化辞で公開している別称情報、情報抽出によって得られた特許と化合物の関係、および、用途情報、アメリカの環境保護庁で公開している製品情報など、様々な情報を統合して表示することができる。

4.3　化学構造の包含関係

化学構造が一意に決まる化合物名以外にも、曖昧性がある化合物名が特許文書や論文の中に出現する場合がある。例えば、「アルカン」のような総称や「ジヒドロキシベンゼン」のような化学構造が曖昧な名称がある。２価フェノールの「ジヒドロキシベンゼン」は、２つのヒドロキシ基の結合位置が特定されていない化学構造である。２つのヒドロキシ基の結合位置が特定された化合物として、「1,2-ジヒドロキシベンゼン」や「1,3-ジヒドロキシベンゼン」、「1,4-ジヒドロキシベンゼン」がある。これらの関係は、図５のように表すことができる。

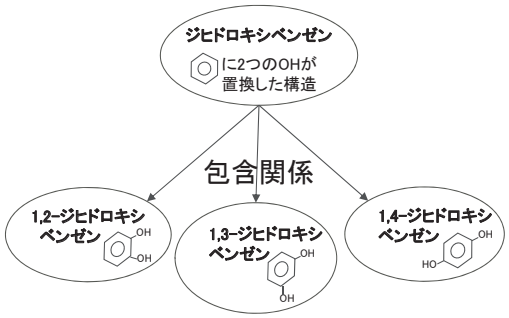

図５　化合の化学物構造の包含関係

化学特許の「ジヒドロキシベンゼン」は、置換基位置が特定されていない抽象的な意味で書かれている場合や、置換基位置が特定された化合物の混合物を表す場合などが考えられる。

ナレッジグラフでは、これらの関係性をエッジによって表現し、必要に応じて検索に利用することができる。このようなナレッジグラフで情報を蓄積しておけば、置換基位置が特定されていない化学構造について書かれている化学特許でも必要に応じてノード間の関係をたどって情報検索や情報表示をすることができる。場合によっては、継承できる情報もあるため、推論や学習にも活用できるものと思われる。

4.4　ナレッジグラフの可視化

Japio YEAR BOOK 2016 で紹介した LOD4ALL フロントエンドという技術を用いて可視化することができる[2][9]。LOD4ALL フロントエンドでは、必要に応じてページをカスタマイズできるため、多様な化合物情報の中からユーザに必要な情報を提供することができる。また、Linked Data の特性を活かして、ユーザの操作によりエッジをたどって掘り下げてみていくこともできる。

5　化合物の製造方法の抽出

化合物の製造では、複数の原料、さまざまな設備に加え、高度な技術と経験の伝承が必要であると言われている。化合物の製造に関する電子化日本語特許は約 40 万件で、多くの関連情報が開示されている。この情報を利用して、化合物の製造に関する情報を抽出・整理して、製造方法と製造プロセスの理解を支援することができる。

製造方法を記述する表現は、ある程度パターン化できると考え、記述パターンを利用した抽出ルールを想定し、ルールベースでの情報抽出を検討した。さらに、製造方法に関する情報をナレッジグラフで表現した。図 6 は化学特許中に書かれた、ビスフェノール A からポリカーボネートを製造する実施例である。

5.1　原料や助剤、最終生成物情報

有機化合物名は階層構造（元素の組合せを階層で構成して命名、炭素で骨格を形成）を持っているので、化合物名を解析して、化合物の部分構造 (置換基) をリスト化することができる[10]。

化合物の製造は複数の化学反応の組み合わせである。多くの場合、助剤を活用して、反応を促進させる。原料の例として、図 6 に、明示的に「モノマーとして」「を用い」「反応器に仕込み」「重合を行った。」等と示されている。助剤の例として、図 6 に、「触媒として」等と示されている。生成された化合物の例として、図 6 に、「白色粉末状」「を得た」等と示されている。有機化合物と原料や助剤、最終生成物の対応関係を用いて、原料や助剤、最終生成物をリスト化することが可能である。さらに、化学量を加えて、以下に示すように、表 1 に原料、表 2 に助剤、表 3 に最終生成物の各リストを生成することができる。

モノマーとして、2,2-ビス(4-ヒドロキシフェニル)プロパンを用い、この100g(0.255モル)と8重量%NaOH水溶液550mℓと塩化メチレン400mℓと分子量調節剤としてのp-t-ブチルフェノール1gと触媒としてのトリエチルアミンの10重量%水溶液3mℓとを、じゃま板付反応器に仕込み、激しくかきまぜながら、これにホスゲンガスを10分間10℃で吹き込んで縮重合を行った。
次いで、反応混合物を塩化メチレン1ℓで希釈したのち、水1ℓ、0.01規定NaOH水溶液500mℓ、水500mℓ、0.01規定HCℓ水溶液500mℓ、水500mℓの順で洗浄した。得られた重合体の塩化メチレン溶液をメタノール3ℓ中に注ぎ、再沈精製して白色粉末状の高分子量ポリカーボネートを得た。
このポリカーボネートの還元粘度は0.48dℓ/g、ガラス転移温度Tgは131℃であった。

図 6　特許明細書からの抜粋例：製造例

表 1　原料リスト

原料	化合物（下線は官能基を示す）	化学量
モノマー	①2,2-ビス(4-ヒドロキシフェニル)プロパン	100g(0.255モル)
	②NaOH	8重量%水溶液550mℓ
	③塩化メチレン	400mℓ
	④ホスゲン	ガスを10分間10℃

表 2　助剤リスト

助剤	化合物	化学量
分子量調節剤	⑤p-t-ブチルフェノール	1g
触媒	⑥トリエチルアミン	10重量%水溶液3mℓ

表 3　最終生成物リスト

生成物	化合物	化学量
重合体	⑦ポリカーボネート	

5.2　処理情報

反応処理の例として、図 6 に、「を用い」「反応器に仕込み」「かきまぜながら」「吹き込んで」「重合を行った。」「希釈した」「の順で」「洗浄した。」「得られた」「中に注ぎ」「再沈精製」等と示されている。反応条件の例として、図 6 に、「g」「モル」「重量 %」「水溶液」「ガス」「分」「℃」「0.01 規定」「溶液」等と示されている。原料や助剤と反応処理や反応条件の対応関係を用いて、反応や精製処理をリスト化することが可能である。それらは、製造経路理解につながる。

以下に示すように、表 4 に反応処理、表 5 に精製処理の各リストを生成することができる。最終生成物が得られても、原料の残存や生成した副産物が混合した状態であり、精製処理も重要である。精製方法によって、純度が大きく異なり、最終価格に影響するからである。

表 4　反応処理

反応順	化合物	処理
1	①②③⑤⑥	じゃま板付反応器に仕込み、
2		激しくかきまぜながら、
3	④	吹き込んで

表５　精製処理

精製順	化合物	処理
1	③	1ℓで希釈したのち、
2		水1ℓ、
3	②	0.01規定　水溶液500mℓ
4		水500mℓ、
5	⑧HCℓ	0.01規定　水溶液500mℓ
6	⑨メタノール	3ℓ中に注ぎ、
7		再沈精製して

5.3　物性情報

物性の例として、図６に、「還元粘度」「ガラス転移温度」「分子量」等と示されている。有機化合物と物性の対応関係を用いて、最終生成物の物性リストを表示することが可能である。以下に示すように、表６に最終生成物の物性リストを生成することができる。高分子を特徴づける基本的な要素は構造と分子量である。一般に、重合で得られる最終生成物は、構造と分子量が異なった分子の集合体である。そこで、「還元粘度」から高分子濃度を、「ガラス転移温度」から耐熱性を把握できる。

表６　最終生成物の物性

物性	値
還元粘度	0.48dℓ/g
ガラス転移温度Tg	131℃

5.4　製造方法のナレッジグラフ

図７に示すように、ポリカーボネートの製造プロセスをナレッジグラフで表すことができる。ここで、灰色のノードは、この製造方法に固有のURIを持ち、それ以外の化合物名や処理などは、他の化学情報とエッジで結ばれたノードとして蓄積する。この製造方法を３つのステップからなると捉えた。ステップ１では原料と分子量調節剤、触媒をじゃま板付反応器に仕込むまでとして、その出力を空ノードで表した。実際にはこの空ノードにも何らかのURIを付ける。ステップ２では、ステップ１での出力を入力として、激しくかきまぜるまでと捉えた。ステップ３では、激しくかきまぜるという処理とホスゲンを吹き込むという処理を同時に処理するという形に表した。ナレッジグラフにおいて、矢印は時間順序を表すとは限らないため、このナレッジグラフでは、時間順序の情報を処理順として、各ステップに持たせることにより、時間順序を表した。

ナレッジグラフにすることで、製造方法を化合物間の関係を用いて検索したり、製造プロセスを可視化したりすることが可能になる。

6　まとめ

化学物質は使い方や使用量により、有益か有害か異なってくるので、反応性と安全性を確保して有効利用するための知識の集約がますます望まれている。そのために、物質の化学的属性（物性）や化学反応性に注目した原子団や化合物の全体構造を把握することが求められている。さらに、静的、動的、潜在的な分子間相互作用を把握できれば、新規化合物や新しい製造方法にたどり着

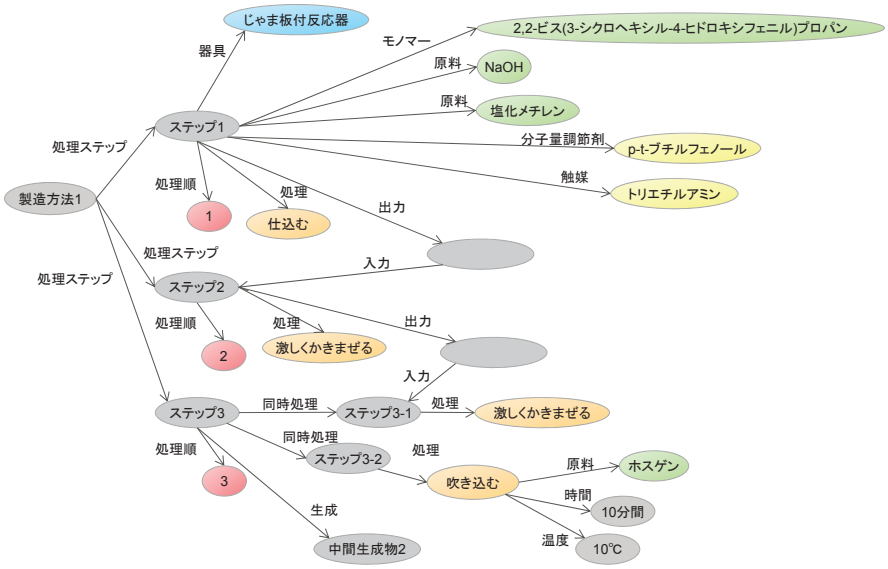

図７　製造方法のナレッジグラフ

く支援ができると考える。また、道具として使ってみたいツールにするためにわかりやすさを兼ね備えることも重要である。ナレッジグラフ等を用いて効率的な集合知の利活用を確立して探索と推薦を可能にしたい。そこから、臨機応変さに答えるリアルタイムな製造計画の実行や、乱雑な環境に耐えるロバスト性の問題対応につなげていきたい。製造の最適化は効率だけではなく、コスト、時間、安全性などが含まれ、与えられた制約条件下で何らかの評価指標を最良にすることである。多くの場合、評価指標は複数に及び、しかも、評価基準はトレードオフの関係にある、多目的最適化問題である。この最適化問題にも挑戦したいと考える。

参考文献

[1] 暮らしと産業を支える日本の化学工業 https://www.nikkakyo.org/upload_files/chemical_industry/overview.pdf

[2] 富士秀，森田一，後藤啓介，丸橋弘治，穴井宏和，井形伸之：Deep Tensor とナレッジグラフを融合した説明可能な AI，雑誌 FUJITSU2018-7 月号，p. 90-96（Vol. 69, No. 4）

[3] 池田紀子，田中一成：特許文書から抽出した化学物質情報の知識化，Japio YEAR BOOK 2016，p. 204-209（2016）

[4] 田中一成，池田紀子：オープンデータを用いた化学特許情報活用へのアプローチ，Japio YEAR BOOK 2017，p. 206-211（2017）

[5] Tomoya Iwakura. A Named Entity Recognition Method using Rules Acquired from Unlabeled Data. Proc. of RANLP'11. Pp. 170-177.

[6] Mintz, Mike and Bills, Steven and Snow, Rion and Jurafsky, Dan. Distant Supervision for Relation Extraction Without Labeled Data. Proc. Of ACL'09. pp. 1003-1011. 2009.

[7] 日本化学物質辞書（日化辞）〈http://dbarchive.biosciencedbc.jp/jp/nikkaji/desc.html〉

[8] Kazunari Tanaka, Tomoya Iwakura, Yusuke Koyanagi, Noriko Ikeda, Hiroyuki Shindo, Yuji Matsumoto. Chemical Compounds Knowledge Visualization with Natural Language Processing and Linked Data. Proc. of LREC. pp. 2250-2253. 2018

[9] LOD4ALL フロントエンド〈http://lod4all.net/frontend/〉

[10] 池田紀子，田中一成：特許文書からの化学物質情報の抽出，Japio YEAR BOOK 2015，p. 274-281（2015）