2016年度情報検索委員会　1-3WG
「新検索技術の到来と特許調査の今後」

2017/11/21　関東部会
2017/11/27　関西部会

［メンバー］

藤田　尚吾	凸版印刷株式会社	（関東部会発表者）
片山　博子	住友化学株式会社	（関西部会発表者）
鈴木　憲	日油株式会社	
六坂　明彦	ルネサス エレクトロニクス株式会社	
加地　英之	住友重機械工業株式会社	
大久保　武利	キヤノン技術情報サービス株式会社	
山崎　勇二	三菱化学株式会社	

はじめに：テーマ決定の経緯

昨今の「人工知能による既存業務の代替」に関する多数の記事
（○○年後になくなる職業、人工知能に置き換わる可能性△△％　etc..）
弁理士や知財業務も対象に挙がることが多い。

⇩ しかし・・・

・そもそも人工知能とは何か？何ができるのか？
・弁理士や知財業務の具体的にどの作業が代替されうるのか？
を追究した記事は少ない。

⇩ そこで・・・

特許調査業務に着目し、以下の調査・考察を行うことにした。
・特許検索に用いられる技術(人工知能に限らず)の概要・動向
・その技術が調査業務に与える影響、今後の可能性

目次

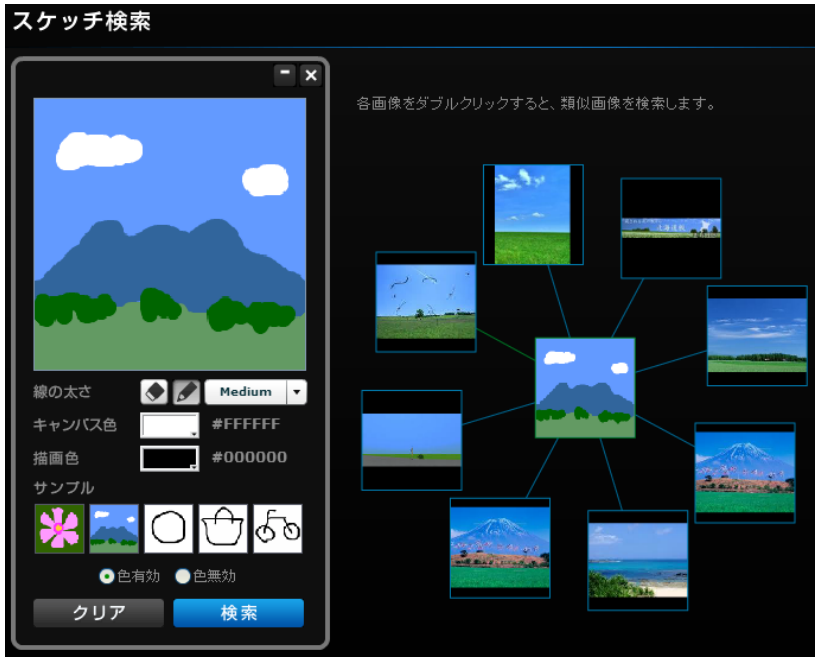
1）背景、活動目的

背景

セマンティック検索とは：
「サーチャーの意図やキーワード（文章）の意味を理解した上で行われる情報検索」

- 情報検索技術の進化
  - ◇セマンティック検索や人工知能（機械学習）
  - ◇画像データそのものを検索項目とした検索（類似画像検索）

（http://www.apple.com/jp/ios/siri/）

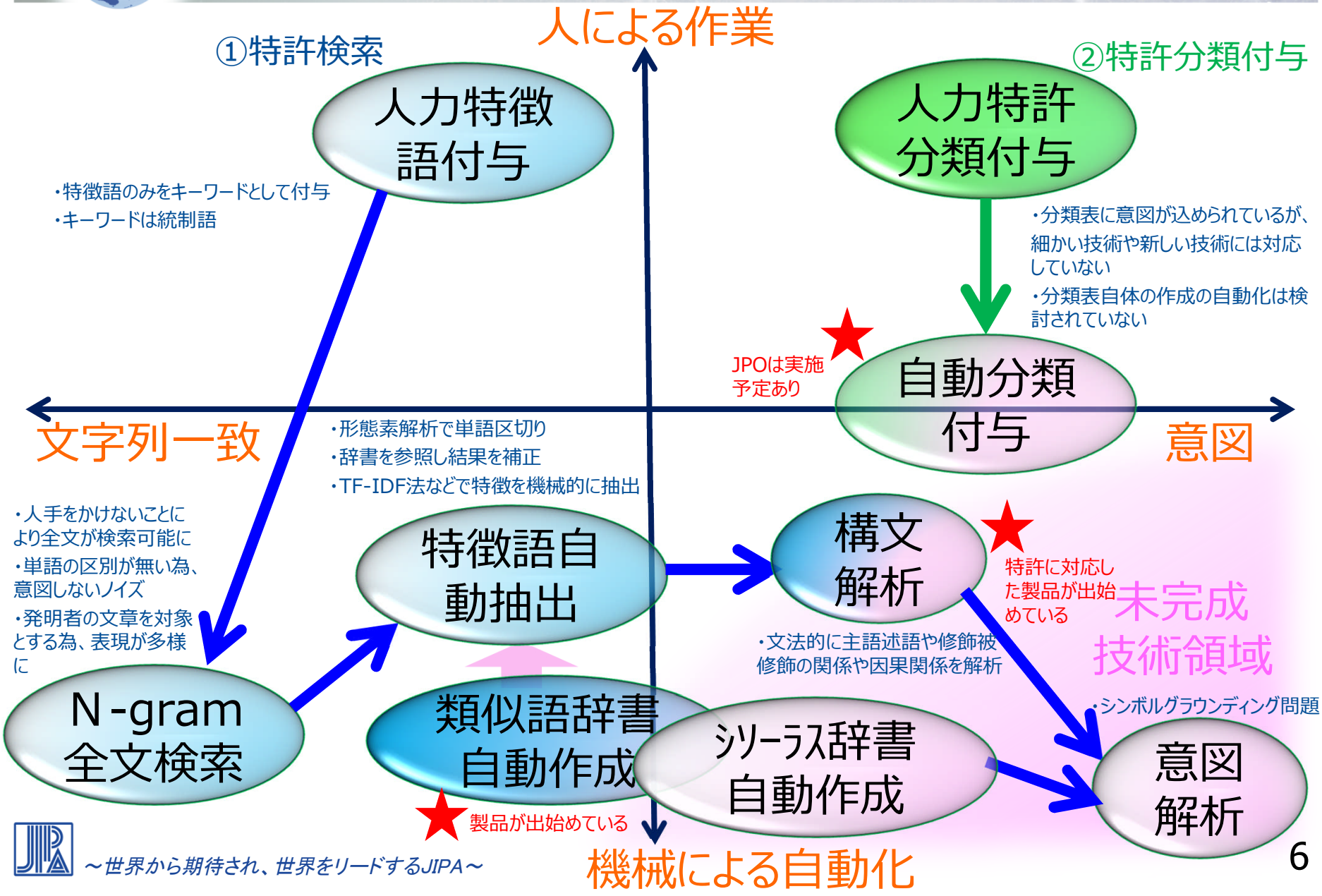


（http:///visseeker2.yahoo-labs.jp/vs）

知財の世界にもこれらの技術が浸透し始めている！！

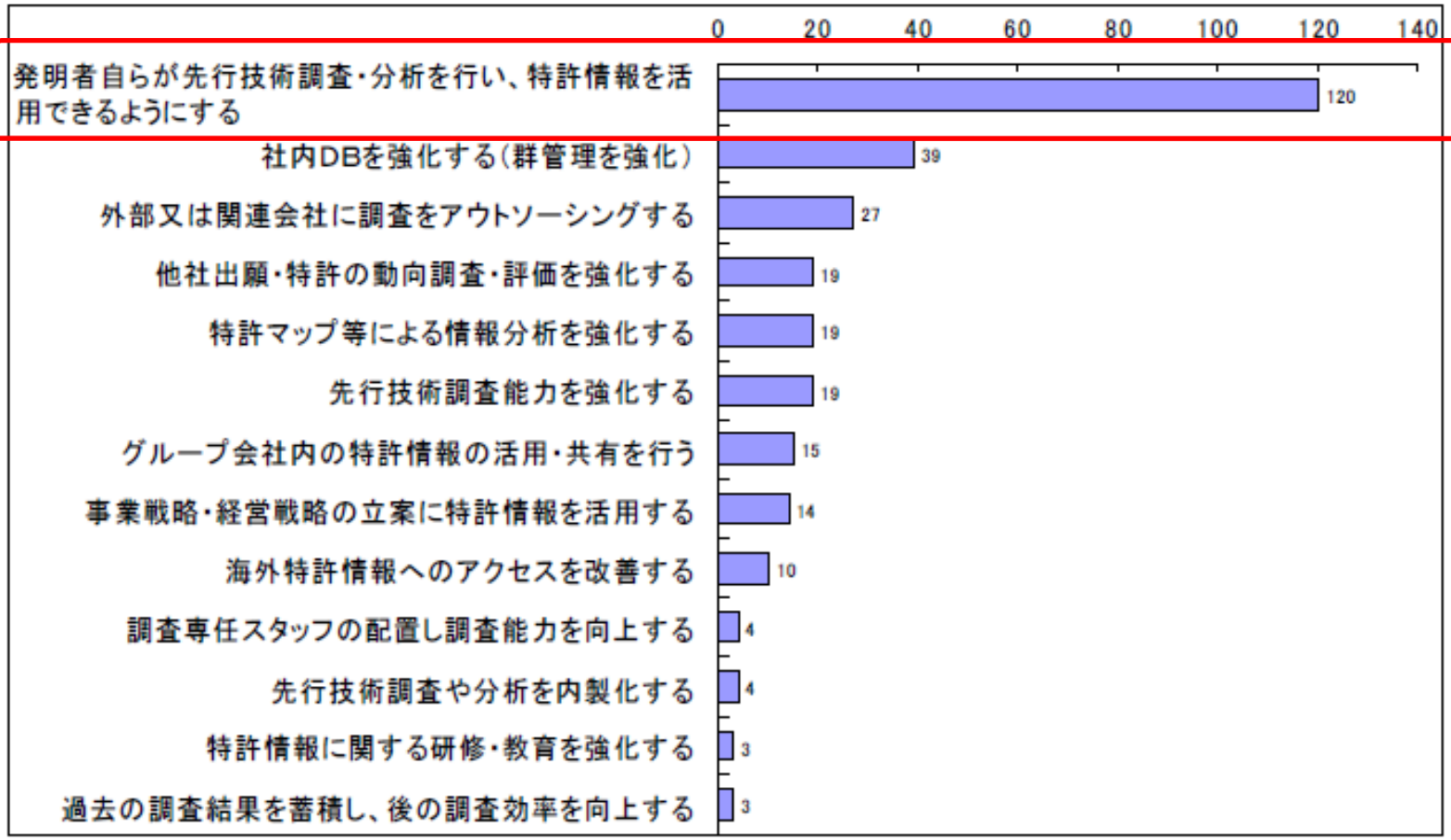
特許検索システム技術の変遷

★：現在研究中の技術



企業におけるニーズ

図1　特許情報の活用に関する将来像（理想像）（自由記載）



国内外企業150社に対するヒアリング調査の自由記載部分。
自由記載にも関わらず8割もの企業が**発明者自らによる調査・分析を期待**している。

引用元：戦略的な知的財産管理に向けて-技術経営力を高めるために-　2007年　特許庁
http://www.jpo.go.jp/torikumi/hiroba/pdf/chiteki_keieiryoku/01.pdf

活動目的

- 背景
  - 検索技術の高機能化
    ⇒ 検索式不要の調査手法の普及（概念検索、意味検索、機械学習・・・）
  - 企業ニーズ
    ⇒ 技術者自身による特許調査・特許情報活用の要望

「誰でも手軽に＆高精度な特許検索」が時代の潮流！！

今後、サーチャーとしての地位を確立していくには、
①検索技術に精通した圧倒的な調査スキル（スペシャリスト化）
②調査以外の分野（分析、戦略etc..）の研鑽（ゼネラリスト化）
のいずれか（両方）が必要！？

But

手軽＆高精度＝機能の複雑化。検索機能の“中身”が見えにくい

- 目的

高度な検索機能を使いこなすには、その技術を知ることが重要。

  - 近年の検索技術をサーチャー目線で解釈し、会員企業に紹介
  - 今後の調査業務のあり方を考察し、会員企業に提言

２）特許検索ツールに用いられている技術

２－１）既存技術

①概念検索

新しい特許検索技術は、概念検索を系譜とするものが多いです。
ですので、最新技術の紹介に入る前に、概念検索の技術のおさらいを・・・

概念検索の要素技術

input **自然文**

概念検索

output **類似特許**

要素技術

- **形態素解析**
  単語の切り出し
- **TF-IDF法**
  特徴語の選別
  単語の重み付け
- **ベクトル空間モデル**
  類似性の評価

形態素解析

文書中の単語の抽出

- 形態素（≒単語）を抽出する処理

日本語：分かち書きされていない → 形態素に分けるのが難しい

品詞の判断(辞書が必要）、接続のルールで判断

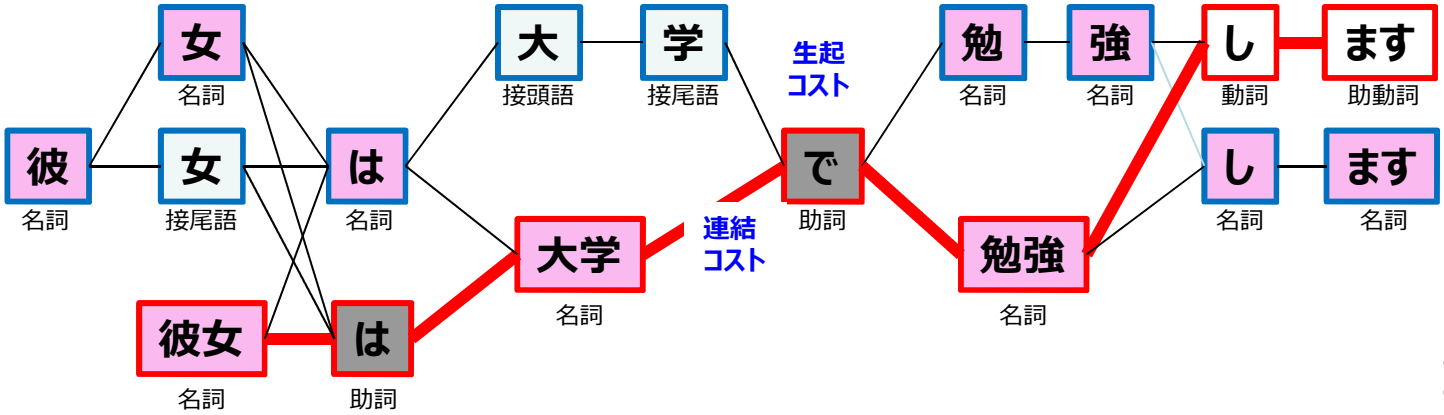
さらに　最長一致法
　　一番長い形態素を順番に割り当てる。
文節最小化法
　　できる文節の数が最少になる候補を選択する
接続コスト最小化法
　　連接コスト：二つの単語のつながりやすさ
　　生起コスト：一つの単語の出現しやすさ
　　コストの合計が最少となるつながりを探す

N-Gram
単語ではなく文字単位（N+1文字）で分解

東京都庁
東京+京都+都庁

京都で検索しても東京都庁もヒット

例：「彼女は大学で勉強します」⇒「彼女／は／大学／で／勉強／し／ます」



TF-IDF

文書中の単語の重み付け

- **TF**（Term Frequency，索引語頻度）
  - 文書内その単語が生じでる頻度
    - 何度も繰り返して言及される概念は重要な概念

$$TF_{ij} = \frac{\text{文書}_i\text{中で単語}_j\text{の出現回数}}{\text{文書}_i\text{中で最も多く出現する単語の出現回数}}$$

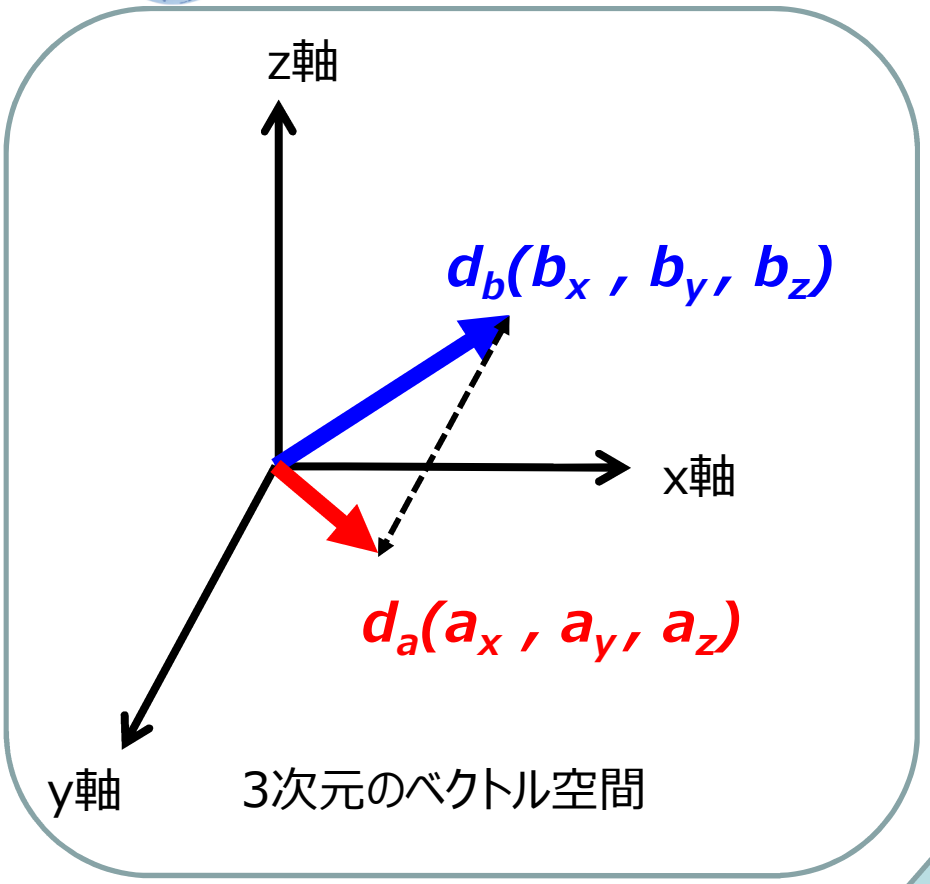
- **IDF**(Inverse Document Frequency;逆文書頻度)
  - 文書集合中である単語が現れる頻度の逆数
    - 多くの文書に現れる単語は文書の検索には役に立たない

$$IDF_j = log\frac{\text{文書集合中の文書の総数}}{\text{文書集合中で単語}_j\text{を含む文書の数}}$$

- **TF-IDF**
  - 文書$_i$の単語$_j$のTF-IDF値 → 文書ベクトルdiのj番目の要素の値(重み付け)

$$W_{ij} = TF_{ij} \times IDF_j$$

ベクトル空間モデル



文書の類似性の評価

これをn次元で考える

文書$_i$のベクトル

$$d_i(W_{i1}, W_{i2}, W_{i3}, W_{i4}, .., W_{in})$$

⋮

文書$_k$のベクトル

$$d_k(W_{k1}, W_{k2}, W_{k3}, W_{k4}, .., W_{kn})$$

↓

N次元空間でのベクトルの近さを算出

↓

文書を近い順に並べる→概念検索

N次元空間を2次元に最も特徴が出るよう(主成分分析)に投影 →ヒートマップ

２）特許検索ツールに用いられている技術

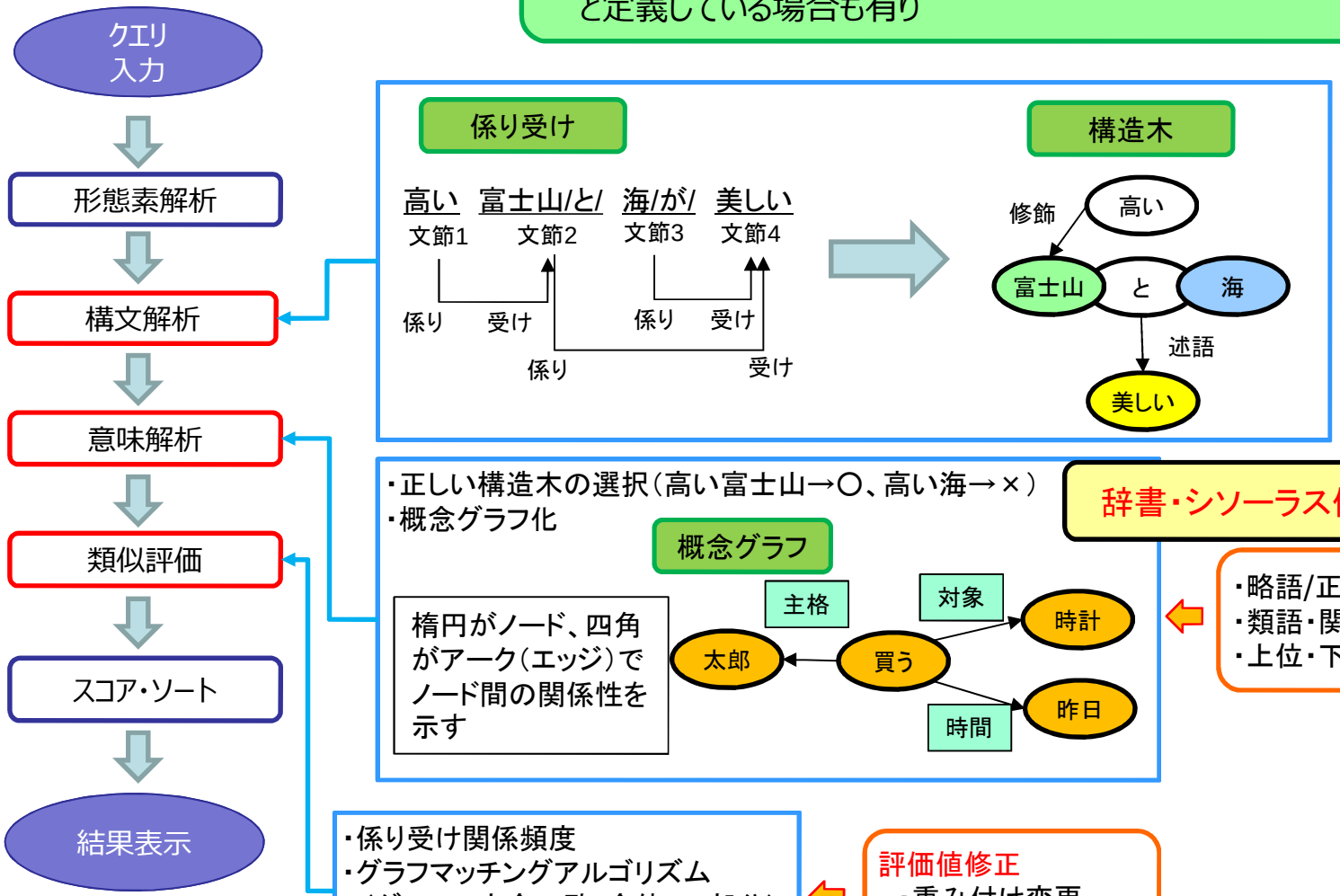
２－１）既存技術

②意味検索

意味検索

意味(セマンティック)検索とは

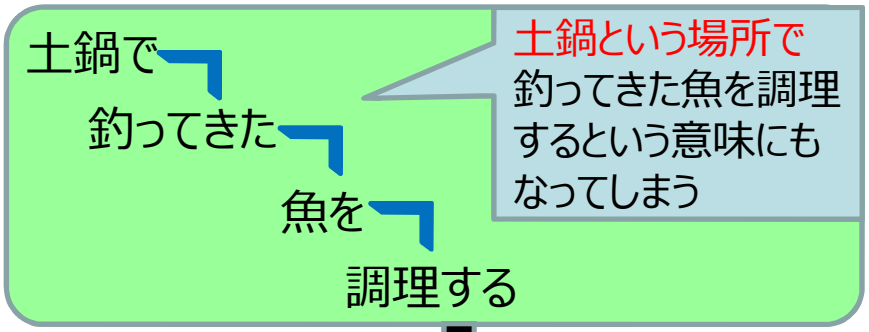
自然文をクエリとして品詞間の係り受けを解析して検索する方式
※単語をクエリとして類語・関連語等を追加して検索する方式も意味検索と定義している場合も有り

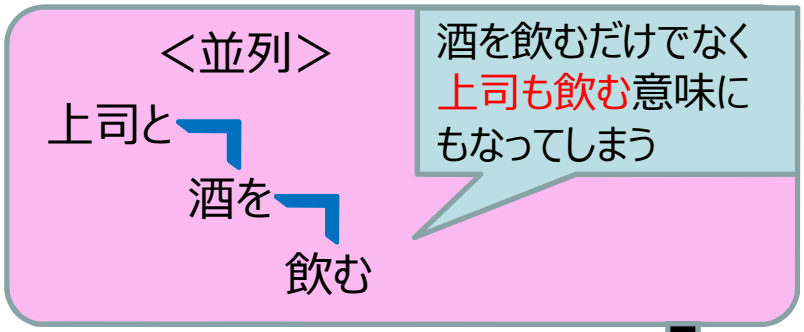


係り受け（構文解析、意味解析）

構文解析とは

自然文を形態素に分割した後、修飾-被修飾等の関係を定義された文法に従い、文構造を明確化する。

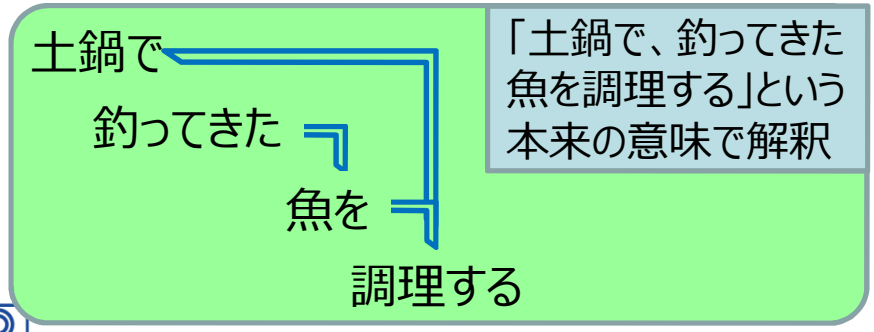


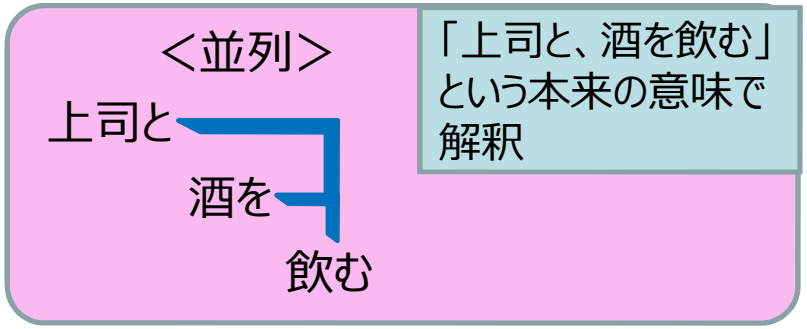


辞書（シソーラス）を使って意味を考慮

意味解析とは

複数の文構造が文法上成り立つ場合、どの文構造が確率的により多くあり得るかを計算する。



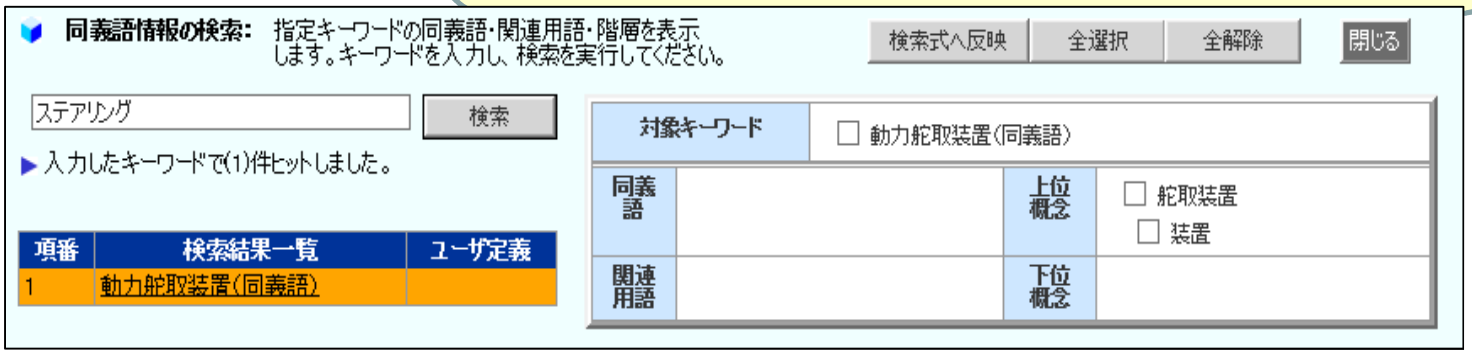


辞書／シソーラスとは

辞書とは？

●特定の言語単位、例えば「単語」や「形態素」に関する様々な情報（品詞情報等）が記述されている知識データベース。

代表的な辞書：
日本語WordNet、 日本語動詞の結合価
分類語彙表、 IPAL辞書、 EDR電子化辞書



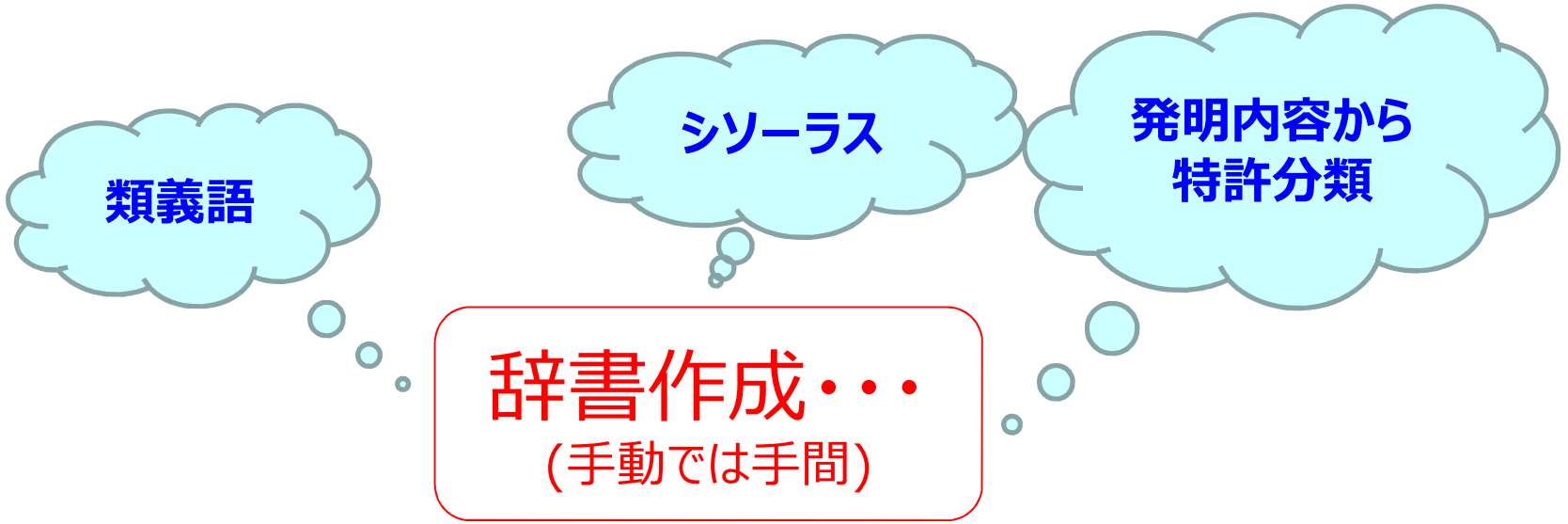
シソーラスとは？

●階層関係を持つ用語集。類語辞書の一種。
語彙の持つ意味から、大⇒細へと階層をたどることができる。
シソーラスを利用している類似検索システムもあるが、辞書により類似や上下関係のまとめ方が異なる。

２）特許検索ツールに用いられている技術

２－２）最新技術による検索の変化

辞書作成・検索の自動化のニーズ



・辞書の自動作成できたらな～

・従来の概念検索に不満　(もっと精度を上げたい !!)

☆　人工知能（機械学習）の登場　☆

現在の技術レベルと期待レベル

	人	機械（コンピュータ）	期待
技術の把握、特徴の抽出	・研究者にヒアリングし、技術を理解した上で、調査条件、技術のバックグラウンド、特徴別に構造化した概念を理解(文章化)する	・実現していない	・機械がインタラクティブ・インターフェースで研究者から情報を引き出す ・あらゆる社内文章からデータマイニングで特徴を抽出する
類似語・特許分類の追加	・類似語、シソーラス、翻訳、特許分類辞書の作成 ・分野ごとに特化した辞書のカスタマイズ ・辞書を利用しながら技術内容に応じて類似語や上位下位語、妥当な特許分類を判断する	・単語の出現パターンから新語・類似語候補の選択 ・接頭語や語尾変化などの文法から類似語の作成 ・出現頻度から特許分類の選択	・機械が概念を理解し、辞書に加えるべき単語か否かを判断する (辞書に新語を自動追加することを特に期待)
検索式作成や類似文章の抽出	・技術の特徴部分を比較し、検索式に反映し、類似性の判断を行う ・技術の構成要素別に情報を整理し、過不足を抽出する。	・特許独自の文章(クレームなど)の構文(パターン)解析 ・モデル文章との単語の出現頻度、構文の文法的類似性の比較 ・人が予めプログラムした特徴や典型文章との類似性の比較	・技術の構成要素別に情報を整理し、人が理解できる構造化を行う ・機械が技術の概念を理解し、OUTPUTの妥当性の判断を行う

現在の技術レベルでは機械は概念を理解していないが、
近い将来、人工知能によって解決できると期待されている。

機械学習と人工知能の関係

人工知能

人間の”知能”を機械で人工的に再現したもの

なので、定義は幅広くあいまい

機械学習は、人工知能の一要素

機械学習　(旧世代・機械学習)

集められたデータから未知のものを予測するのに必要な特徴や情報を抽出し、ルールを自動で取得する手法

データ

↓

特徴抽出
人手(経験必要)

↓

規則性やルールを発見

ディープラーニング
(新世代・機械学習)

脳科学の研究成果を基礎にデータの分類や認識の基準を人間が教えなくても、データを解析することで、自ら見つけ出すことができる

機械学習の手法。ニューラルネットを多層にしたもの。

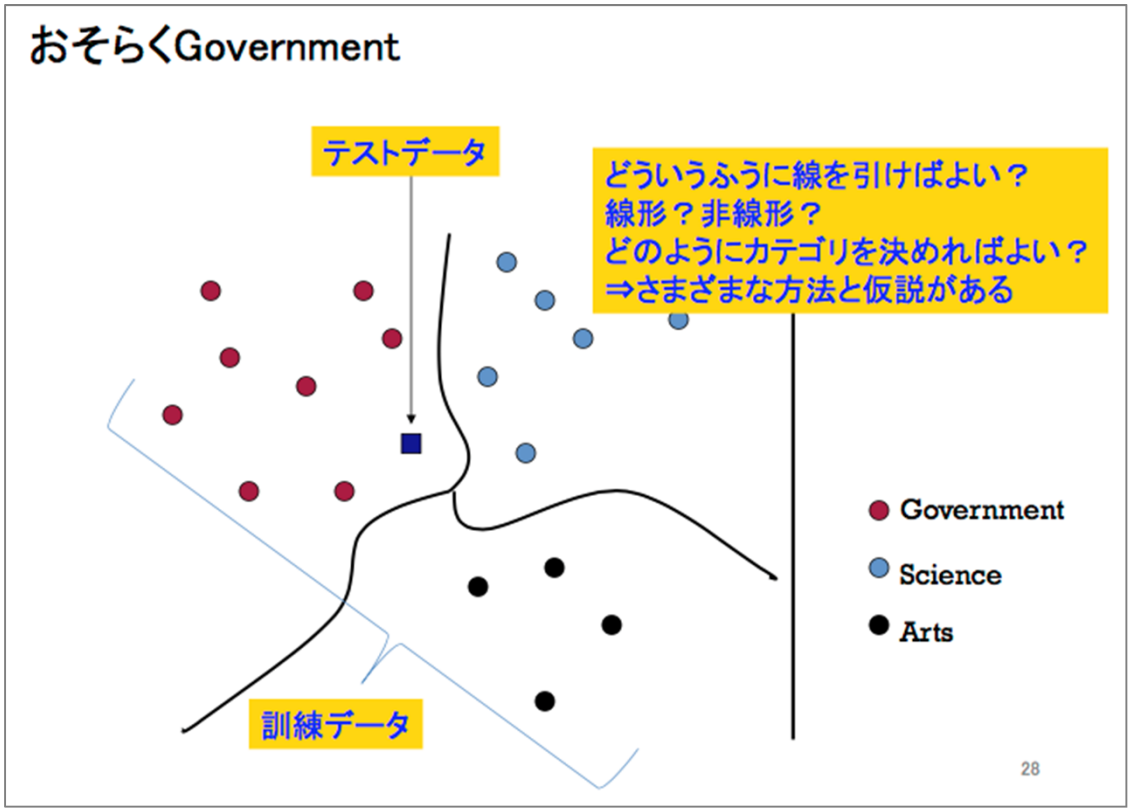
特徴を抽出しながら学習

1950～1970年代　1980～2000年代　2010年代～

http://www.itmedia.co.jp/enterprise/articles/1608/31/news086.html

機械学習とは

機械がデータの分け方（線の引き方）を学習する仕組み



www.gdep.jp/seminar/20150526/DLF2015-01-MATSUO.pdf

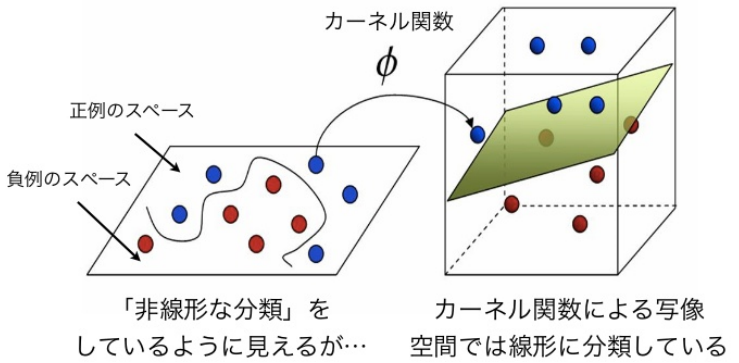
一旦学習すれば、未知のデータがどのカテゴリに属するか予測できる

機械学習の種類

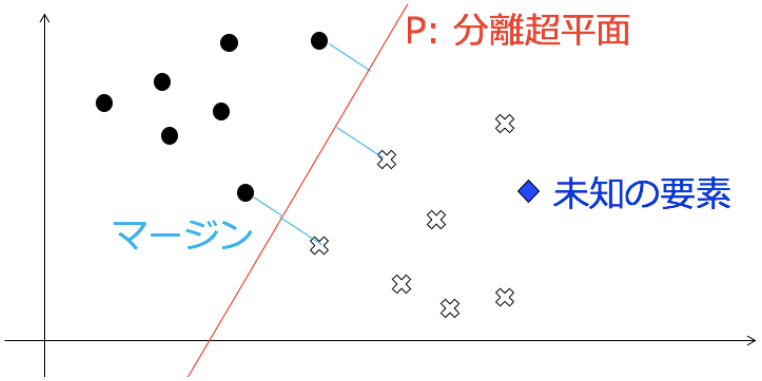
①教師あり学習 (教師データあり)

事前に学習させたデータの分類基準に基づいて、入力データを分類する。

例.サポートベクターマシン



学習モデル構築

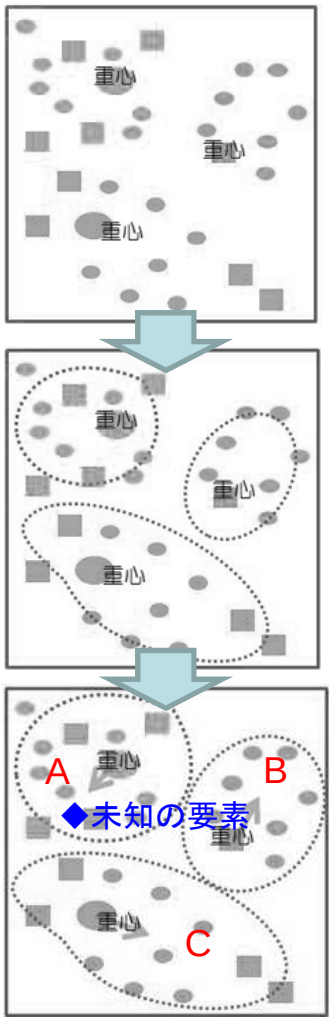


http://www.tsjshg.info/udemy/Lec82-83.html

②教師なし学習(教師データなし)

入力データのみから何らかの基準を設けて分類する。

例.クラスタリング（K-means法等）



A.いくつかランダムに重心を決める

B.近い点同士をグループ化

C.クラスターの中心にくるように重心を移動し、再びB.へ(収束するまで繰返し)

http://news.infoseek.co.jp/article/enterprisezine_4727/

機械学習の問題点とディープラーニング

■これまでの機械学習の問題点■

①素性(そせい、特徴量)設計

データのどこに着目すべきか、人間が決めなければならない

②シンボルグラウンディング問題

機械は言葉の意味を理解できない（青りんご ＝ りんご ＋ 青い　が理解できない）

③フレーム問題

限られた枠組み（フレーム）の中でしか有用でない（例外に対応できない）

解決の可能性

■ディープラーニング(深層学習)の出現

データから機械が自動的に特徴を抽出できる

すなわち、どこに着目すればよいか（特徴量）を教える必要がなくなった

＜期待される分野＞

1.音声認識
2.画像認識
3.自然言語処理
4.その他・・・ロボットなどの制御、異常検知、マーケット分析、不正検知など

より高精度化できる

研究発展中

ディープラーニングの問題点と可能性

既にディープラーニングを利用した特許検索システムも見られるが、
他の機械学習と比較し、まだ画期的な精度とはなっていない。

課題

・大量の学習データと
　正解データが必要
・大量の学習に大量の
　リソースが必要
・現状の学習モデルは機能
　特化型(汎用性が・・・)

・あらゆるデータが電子化され、利用で
　きるデータは加速度的に増加している
・GPUが手軽に使える計算資源提供業
　(クラウドサーバ貸業)が起っている
・機械は疲れないので24時間学習可能
　且つ、学習した結果を共有可能
・新アルゴリズムや学習済みモデルを無料
　で一般公開する傾向が業界にある

一旦、特許に適合した学習済みモデルが出来上がれば
加速度的に改良され、普及する土壌がある。

期待!!

３）調査業務の今後と提言

10年後の特許調査は?

・文献Aに近い文献をリストアップする
・クレームを構成要件に分解し、
　構成ごとに類似度の高い文献
　(文献中の文章)を示す
・新語、類似語、シソーラス辞書の作成
・精度の良い翻訳
・特許分類の付与

手を動かす

10年後には検索システムによる全自動化が実現しているかも!!

機械に任せてしまっても良い分野

・研究者が意識する前の発明を発掘する
・データから将来を予測する
・技術動向や企業動向を解釈し、自社の特許戦略や開発戦略に結びつけ、妥当な対策を生み出す
・裁判や交渉に勝てるロジックを構築し、技術の素人に理解させる

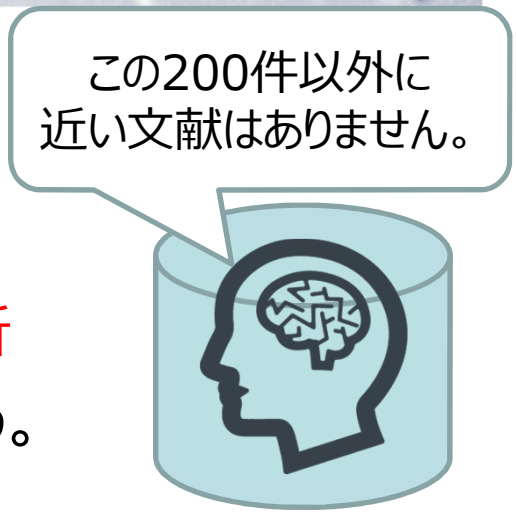
分析、判断、予測

10年では自動化は難しいだろう

人間がやらなくてはならない分野

仮想事例：10年後の無効資料調査

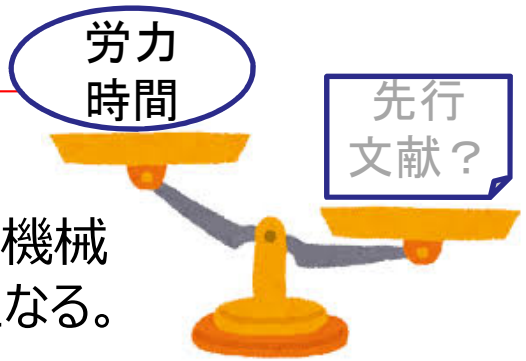
- 障害となる他社特許を発見。
- 機械にその特許を登録し、先行文献を
  自動調査させたところ、類似文献は200件と提示。
  ※公報自体を登録すれば機械が記載内容を自動解析
- その200件を人手で精査するも、無効化は厳しそう。



あなたはどのような判断を下しますか？

①人手で追加調査を行う（機械の判断を信じない）

②無効化を諦め、ライセンス・購入・事業撤退を検討する
（機械の判断を信じる）



上記調査であれば他部門でも実施可能。

ビジネス面からの人手による追加調査の要否判断と、機械が到達しえないスキルを駆使した調査が知財の仕事となる。

提言

・あなたはどういうキャリアを歩むか？
・知財部をどういう組織に変えていくべきか？

今後、サーチャーとしての地位を確立していくには、
　①ツールの機能に精通した圧倒的な調査スキル（スペシャリスト化）
　②調査以外の分野（分析、戦略etc..）の研鑽（ゼネラリスト化）
のいずれか（両方）が必要！？

これまでの延長で検索スキルを磨いても、
ニーズが無くなる時代が到来するかもしれません。

ご清聴有難うございました

論説は11月号掲載です。ぜひご一読ください！