

日経記事テキストに対する 自然言語処理を用いた情報抽出とグラフ構造化

安井雄一郎

日本経済新聞社 日経イノベーション・ラボ

数理システムユーザーコンファレンス 2020

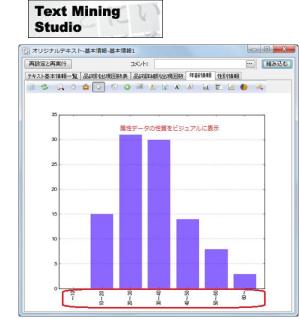
まとめ

- 日経記事データ
 - 記事データ販売、AWS Data Exchange での販売
 - 学生研究奨励賞への記事テキストデータ提供
- 日経記事データ x TMS
 - 日経記事テキストと相性が良い分析ツール
- 日経記事データの構造化
 - 記事がもつ階層構造と自然言語処理による情報抽出を用いて、グラフ構造を抽出
 - グラフ構造はグラフデータベースに蓄積して、クエリにより利活用

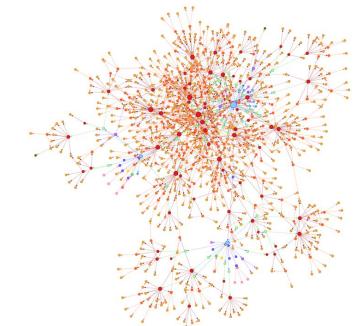
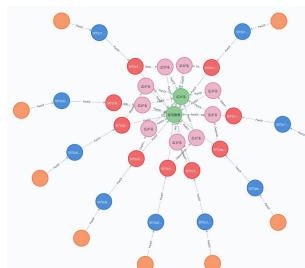
日経記事データ

項目名	値	サンプル
記事ID	"NPRI0302150719NO540351" など	NPRI0302200328NO50048
発表日	2020/06/20 16:18:32 +09:00	
用紙	NW	
取扱い	複数のニュースコード	
種別	日本経済新聞社、「日経経済ニューススクレイプ」以降	日本経済新聞社
登録元	日経社	
記事本文	本文	
登録者	"~君"	
記事本文拡張	6	
記事本文学年	大学教	721
記事本文表示	見出し	P.C.R測定、企業で記入、大学の2割が実験対応、海外移動に必要な規則。
記事本文	本文	新型コロナワクチン接種を実施するP.C.R検査の導入を大きな進展の例で広げている。日本政府は接種が完了する方へ「ワクチンパス」としてQRコードを発行。既存の電子機器にひびく検査結果のデータを組み込むことで、その有効性を高めよう。「デジタル接種証明」機能を備えた「ワクチンパス」が、多くの国で採用されている。三重県は8月26日から、ワクチン接種を受けた人の確認を容易にするため、国内で初めてQRコードを発行している。接種証明は原則としてP.C.R検査結果と組み合わせて発行されるが、接種済みの場合は「ワクチンパス」で示す。既存の電子機器との連携により、接種情報を簡単に確認できる。
キーワード	記者の文から自動的に取り出したワードまたはその近似名	三重県 P.C.R測定 対象 日本の接種率 ワクチンパス 国内接種率 デジタル接種証明 ワクチンパス QRコード ライブドア
分類	P.C.R測定 対象 日本の接種率 ワクチンパス 国内接種率 デジタル接種証明 ワクチンパス QRコード ライブドア	9.2, N_0.0115, A_N_0.3618, N_0.000173, N_0.0013, Z, N_0.000697, N_0.016
他ID	東京コード	8586-4843, 5108-9201, 6141, 9024, 8535

TMS によるテキストマイニング



記事のグラフ構造化



自己紹介

- 安井雄一郎(やすいゆういちろう)
- 学術研究(~2016/12)
 - コンピュータに適した(ハードウェアを考慮した)並列アルゴリズムの設計と実装
 - 主な対象はグラフ探索(幅優先探索)で、グラフ探索性能を競う Graph500 や Green Graph500 (7期連続世界一位)に挑戦
- 日経BP(2017/1~2019/3)
 - データサイエンティスト
 - トラフィックログ分析、マーケティング分野のデータ分析、分析基盤の設計と運用
- 日本経済新聞社(2019/4~)
 - 日経イノベーション・ラボ 上級研究員
 - 自然言語処理を用いた情報抽出、セマンティックウェブ技術やグラフデータベースを用いたデータ活用などに興味



安井雄一郎

日経の記事テキストデータ

- 新聞記事データは分析用に販売
 - 機械学習の学習データに
 - テキストマイニングなど分析に
- 豊富なメタ情報
 - 分類語の付与: テーマ, 業界, 地域, 記事種別
 - キーワード抽出: 人物名、企業名や企業コードなど
- AWS DataExchange で購入可能に
 - 期間・更新頻度: 1981年10月～で日次提供
 - 提供形式: CSV (UTF-8)
 - 記事データ以外に POS データも購入可能
 - ※ 現時点ではサンプルデータの提供まで

AWS DataExchange での記事データ販売



Nikkei Article Data (Trial)
Nikkei, inc

Nikkei News Article is a text dataset of articles from Japan's leading economic newspaper. Nikkei will provide the data of Japan's most influential financial, economic, and corporate news with precise tags and meta-information. This vast dataset is ideal data for natural language processing and machine learning.

Free
1 month subscription available.

プレスリリース (2020/09/30)

AWS Marketplace および AWS Data Exchange 上で、日本法人のソフトウェアベンダー、データプロバイダー、コンサルティングパートナーが自社のソフトウェアやサービスを提供開始

投稿日: Sep 30, 2020

アトボック、インサイトテクノロジー、テックビューロホールディングスなどの日本法人のソフトウェアベンダー (ISV) が提供するサーバーバイ製ソフトウェアを、世界中のお客様が容易に検索、導入、管理可能に

日経、QUICK、JMDC、など日本のデータプロバイダーは AWS Data Exchange を通じて、世界中の AWS ユーザーへデータセット販売が可能に

(東京、2020年9月30日発表) Amazon.com, Inc. の関連会社であるアマゾン ウェブ サービス ジャパン株式会社は本日、日本法人のソフトウェアベンダー (ISV)、データプロバイダー、AWS コンサルティングパートナーが、AWS Marketplace と AWS Data Exchange において自社のソフトウェアやサービスが販売できるようになったことを発表しました。今回、登録企業の範囲が日本に拡張したことにより、日本国内の ISV、コンサルティングパートナー、認定データプロバイダーは、世界中の AWS のお客様にリーチできる新たな戦略を開拓でき、お客様は利用できるソフトウェアやサービスの選択肢が広がります。

AWS Marketplace は、8,000 件以上の商品リストで構成される厳選されたデジタルカタログであり、50 カテゴリーあるサーバーバイ製ソフトウェアの検索、データ、導入、管理を容易に行えます。本サービスは世界中の 24 の AWS リージョンで展開されており、そこでお客様はソフトウェア製品を一元的に検索・比較でき、AWS のコンソールバネルへのログインからわずか数分で、事前構成済みのソフトウェアを迅速に起動できます。

記事データの構造、キーワードや分類タグを付与

項目名	備考	サンプル
記事ID	"NIRKDB20150715NKM0351" など	NIRKDB20200628NKM0048
掲載日	"2016-09-25T23:33:26Z" など	2020-06-29T18:13:32+09:00
媒体略号	媒体のユニークコード	NKM
媒体名称	"日本経済新聞 朝刊"、"日経速報ニュースアーカイブ" など	日本経済新聞朝刊
絵・写真・表の有無	"" or "有"	
記事本文段落数	段落数	6
記事本文文字数	文字数	721
記事見出し	見出し	PCR需要、企業で拡大、大手の2割が実施検討、海外渡航に必要な例も。
記事本文	本文	<p>新型コロナウイルス感染を判断するPCR検査の需要が大手企業の間で広がっている。日本経済新聞が主要85社に聞き取り調査したところ、三菱商事が既に検査を実施したと回答した。全体の約2割の16社が「実施を検討中」と答えた。海外事業の比重が高い機械や自動車関連の企業で海外渡航時にPCR検査を求められるケースが出てきている。（関連記事7面に）</p> <p>PCR検査はウイルス感染の判定などに使う遺伝子検査の一つ。感染拡大当初は発熱が一定期間続くなど実施には条件があったが、最近は感染の疑いがない無症状者であっても受け付ける民間の診療所が増えている。</p> <p>アンケートでは「希望する従業員にPCR検査を受けさせる計画はあるか」と各企業に聞いた。三菱商事は海外駐在を予定している社員など対象を限定。6月中旬から実施している。検査には社内診療所に加え外部の検査機関を活用している。</p> <p>「検討中」はブリヂストンやDMG森精機、日本航空やテルモなどの16社。東京エレクトロンは海外派遣する技術者に検査を受けさせることを検討している。「顧客から（PCR検査の）陰性証明書（3面きょうのこと）</p>
キーワード	記事の文中から主題語として切り出したワードまたはその正式名称	三菱商事,テルモ,ブリヂストン,大丸松坂屋百貨店,日本航空,DMG森精機,西武ホールディングス,東京エレクトロン,検査,PCR検査,PCR需要,感染,新型,日経調査,ウイルス,コロナウイルス,大手企業,実施検討
分類語	記事内容のテーマコード(#W～)・業界コード(#B～)、証券コード等の会社コード(T～、N～、PD～)、紙面名等の記事分類キーワード(\$～)、コラム名(「～」)	T8058,PD431,T6141,PD211,T9024,PD551,T8035,PD239,T4543,PD313,T5108,PD131,T9201,PD611,\$一面,#K7,#B0320,#B0470,N0001592,N0001154,N0038618,N0001738,N0013612,N0000697,N0001655,N0001941
株式コード	東証コード	8058,4543,5108,9201,6141,9024,8035

数理システム学生奨励賞への記事データ提供

- 数理システム学生研究奨励賞への記事データ提供
 - <https://www.msi.co.jp/userconf/student/index.html>
 - 日経の記事データを試せる良い機会です！
- 前処理済みデータを提供予定
 - 新聞固有(パラグラフの先頭の【】、末尾の丸括弧)の構造を除外
 - 日経記事データ固有の構造(表組みのテキスト書き起こし部分)を除外

項目名	備考	サンプル
記事ID	"NIRKOB20150715NM0351" など	NIRKOB20200628NM0048
掲載日	2016-09-25T23:33:26Z" など	2020-06-29T18:13:32+09:00
媒体番号	媒体のユニークコード"	NKM
媒体名	"日本経済新聞 報紙"、"日経伝媒ニュースアーカイブ" など	日本経済新聞社
絞り・写真・表の有無	" or "有"	
記事本文落款	段落数	6
記事本文文字数	文字数	721
記事提出出し	提出出し	P.C.R.需要、企業で忙しく、大手の2割が実施検査、海外直航に必要な要件。
記事本文	本文	<p>新型コロナウイルス感染症を判断するP.C.R.検査の需要が大手企業の間で広がっている。日本経済新聞が主催85社に聞き取り調査したこと、三菱商事が医療機器を実施したと回答した。全体会員の2割の16社が「実施を検討中」と答えた。海外事業の比率が高い機械や自動車関連の企業で導入検査にP.C.R.検査を実施されるケースが出てきている。(関連記事7選)</p> <p>PCR検査は、PCR検査の費用が高くなるうえ子供の検査一つ、通常大人相当は検査料一括請求額など豪華には条件が付いたが、最近は検査の安いがない、無症状者であっても受け付ける民間の診療所が増えている。</p> <p>アンケートでは「希望する検査実施にP.C.R.検査を受けさせる計画はあるか」と各企業に聞いた。三菱商事は海外駐在を決定している社員などを対象を設定。6月月中旬から実施している。検査には社内検査院に加え外部の検査機関も活用している。</p> <p>(出典)アサヒデジタルシステムDG新規機器、日本経営やテレワークなどの16社、東京エクスプレスは海外派遣する扶養者に検査を受けさせることを検討している。「顧客から(P.C.R検査の)陰性証明書(3選きょうのと)」</p> <p>三菱商事、ルモ・プリマス、大和松屋酒店貿易、日本航空、DMG森精機、西武ホールディングス、東京エクシトロン、P.C.R検査、P.C.R検査、P.C.R検査、感染症、新型、日経経営、ウイルス、コロナウイルス、大手企業、実施検査</p>
分類語	記事内のテーママーク(【】)～(NW～)、裏表マーク(【】～、N～、PD～)、登録名等の記事分類キー(ワード～)、コラム(「～」)	T8058.P0431.T6141.P0211.T9024.P0551.T8035.P0239.T4543.PD313.T5108.P0131.T9201.PD611.S-#.T#_B0320.B0470.N0001592.N0001154.N0038618.N0001738.N0013612.N0000697.N000155.N0001941
株式コード	東証コード	8058.4545.5106.8021.6141.9024.8035

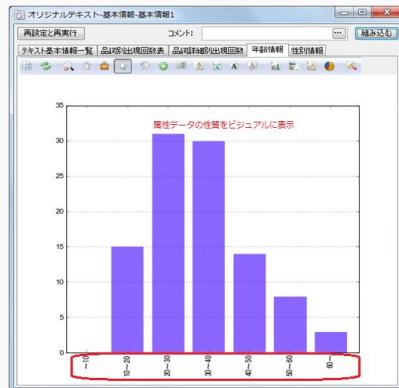
通常の記事データ

記事本文	本文	新型コロナウイルス感染を判断するP.C.R.検査の需要が大手企業の間で広がっている。日本経済新聞が主催85社に聞き取り調査したこと、三菱商事が医療機器を実施したと回答した。全体会員の2割の16社が「実施を検討中」と答えた。海外事業の比率が高い機械や自動車関連の企業で導入検査にP.C.R.検査を実施されるケースが出てきている。(関連記事7選)
------	----	---

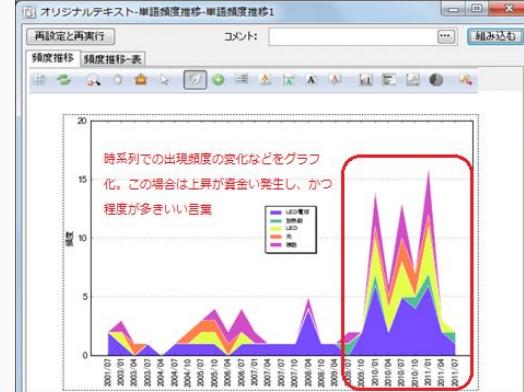
TMS 向け前処理カラム(イメージ)

- NTTデータ数理システム Text Mining Studio
 - 手間のかかるテキストマイニングを簡単に実現
 - 日経テキストデータと相性が良いとの評価あり
- 分かり書きからネットワーク分析まで様々な機能を GUI で
 - 典型的な分析を評価軸を変えて、簡単に試行錯誤できる

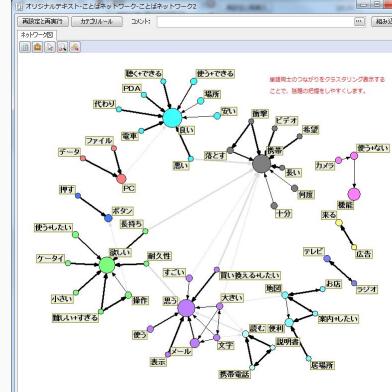
単語の頻度



単語の頻度の時系列変化



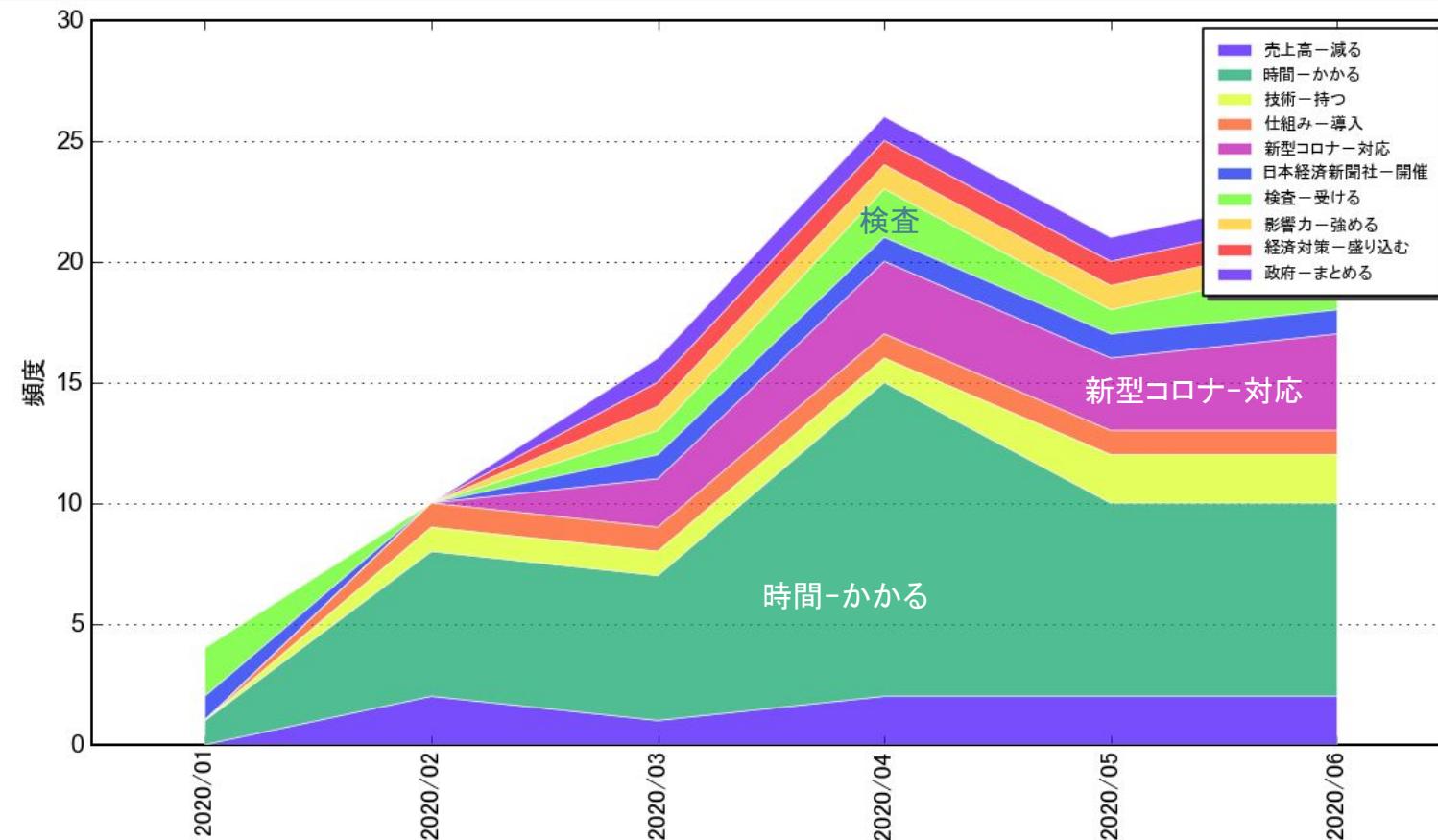
共起ネットワーク、係り受けネットワーク



係り受けの関係、時系列変化(上昇傾向)

Text Mining Studio

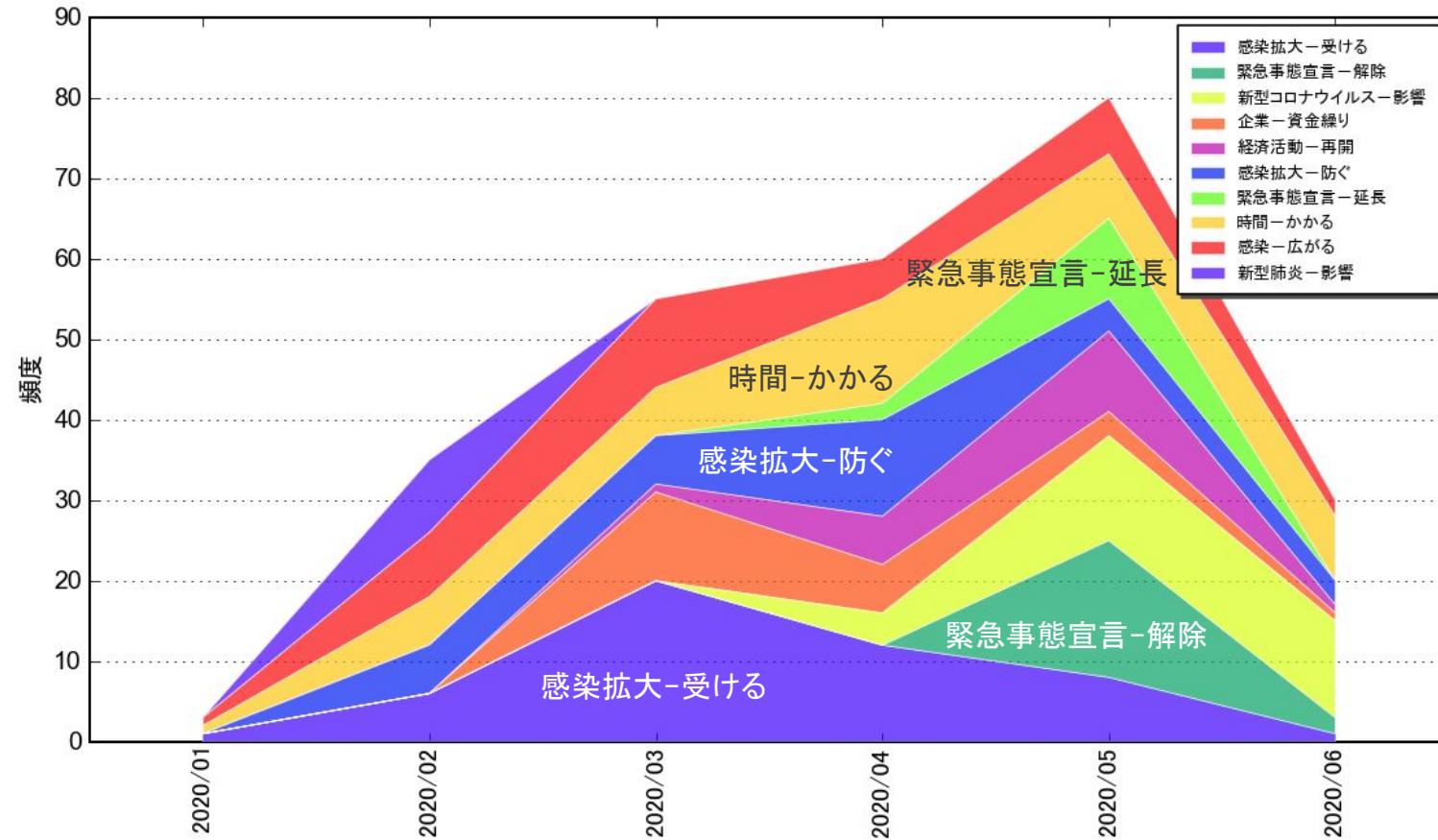
2020/01/01 ~ 06/30、日経新聞朝刊、一面に掲載された 9,819 記事(抜粋)



係り受けの関係、時系列変化(頻度の変動が大)

Text Mining Studio

2020/01/01 ~ 06/30、日経新聞朝刊、一面に掲載された 9,819 記事(抜粋)

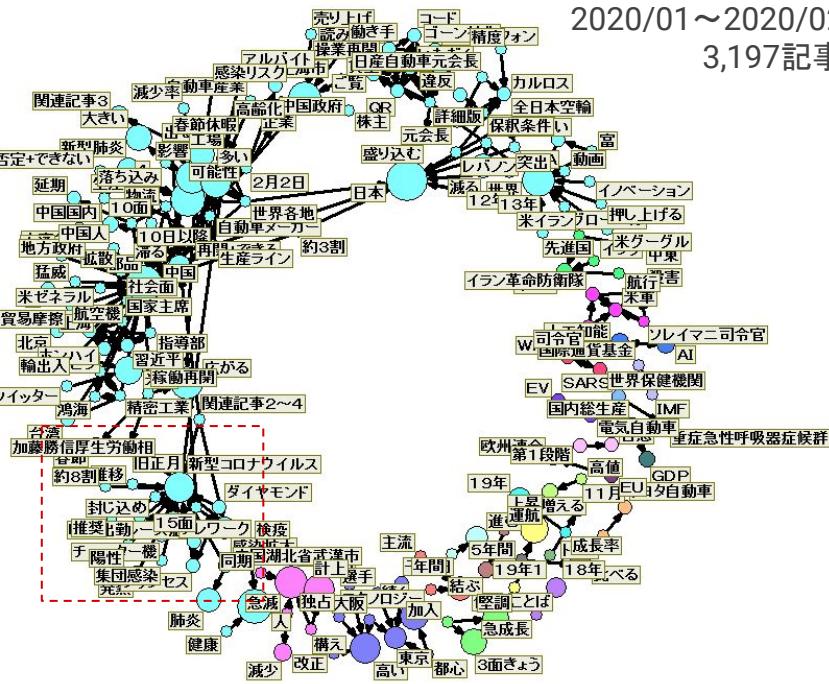


共起ネットワーク

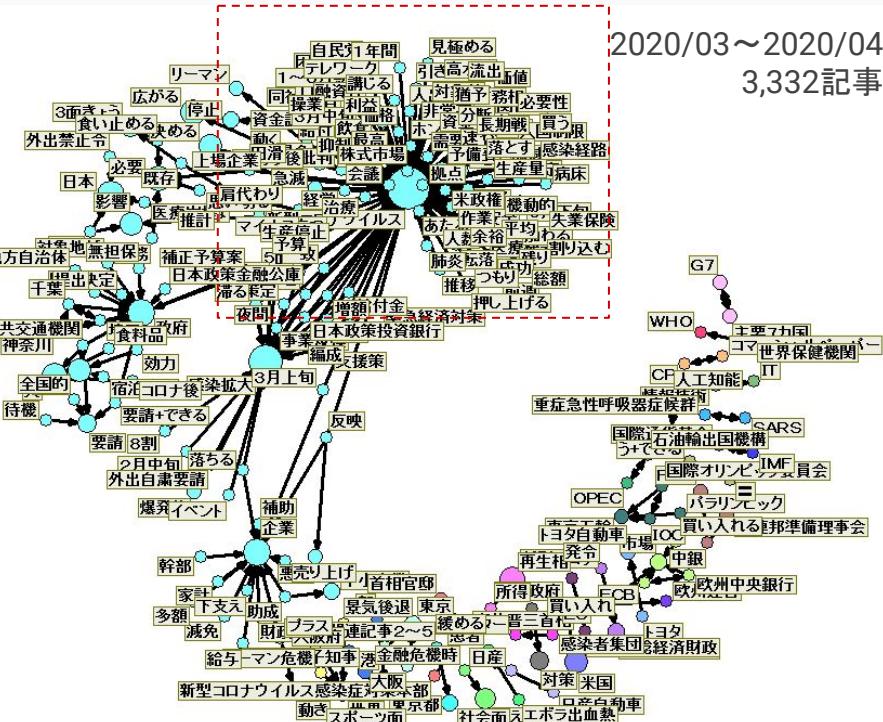
Text Mining
Studio

- 「新型コロナウイルス」への集中を確認できる

2020/01～2020/02
3,197記事

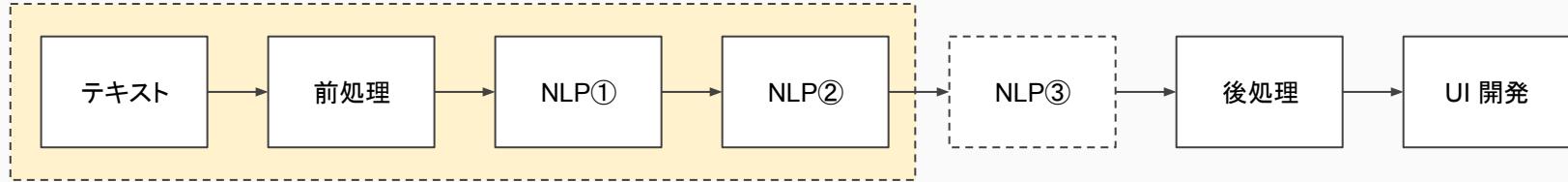


2020/03～2020/04
3,332記事

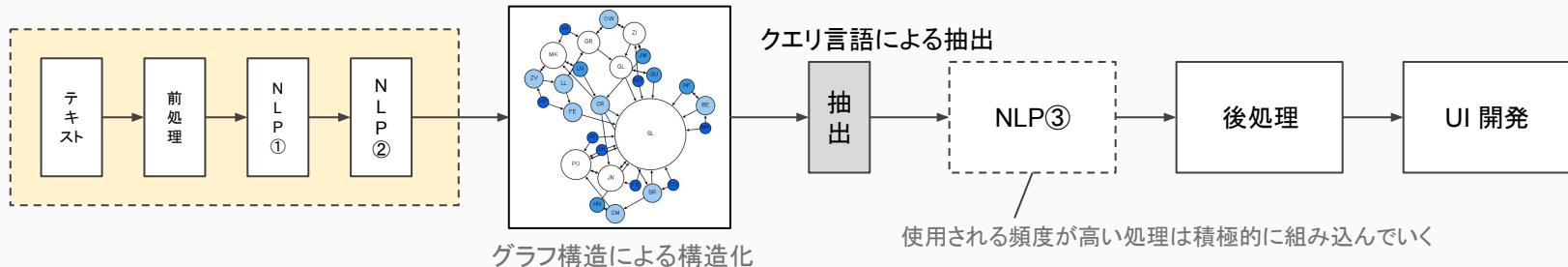


典型的な処理を構造化し、ソフトウェア資産を活用

- 自然言語処理を用いた技術検証は要素が多く、定型化しづらい
 - ソフトウェア資産を有効活用して、新しい技術検証の効率化を行いたい



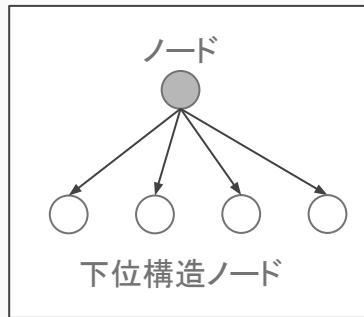
- 典型的な処理は柔軟な構造で蓄積しクエリ言語で抽出
 - 工程を典型的な処理と、個別の処理に分割する
 - 典型的な処理は柔軟なグラフ構造（グラフデータベース）で蓄積し、クエリ言語で抽出



記事がもつグラフ構造を抽出

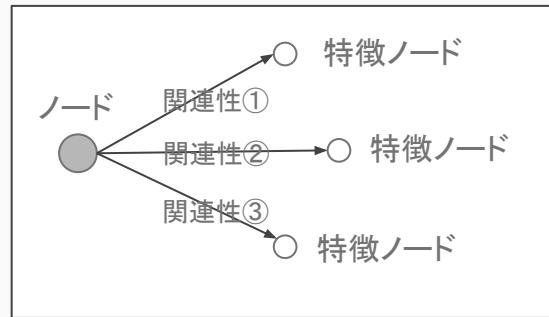
- 検索・抽出したい構造をグラフ(ノードとエッジ)で表現する
 - 階層構造: 記事の階層構造(媒体、面、記事、パラグラフ、文章)
 - 情報抽出: 固有表現、係り受け

階層構造



記事の階層構造、分類の階層構造など様々な下位構造を明示的に表現

情報抽出



データの対象がテキストならば自然言語処理、動画像ならばオブジェクト認識など、様々な手法を用いる

記事



グラフなので
上位・下位どちらからでも
辿ることができる

パラグラフ



文章



特徴



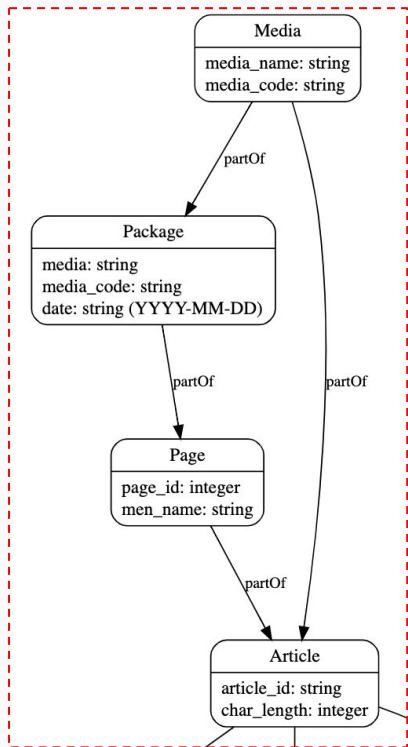
単語



ある特徴(単語を含む、タグ)をもつ記事を抽出

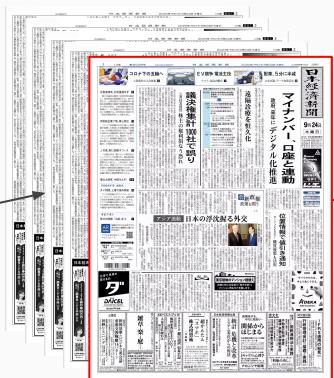
グラフ構造化による表現 ①

- 記事がもつ階層構造(メディア、パッケージ、面)



日本経済新聞

Media



Package



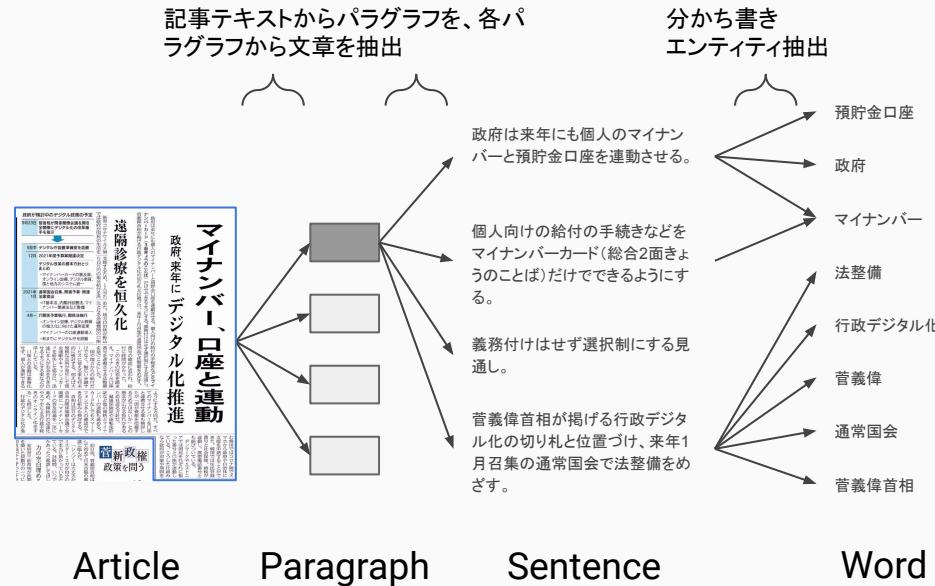
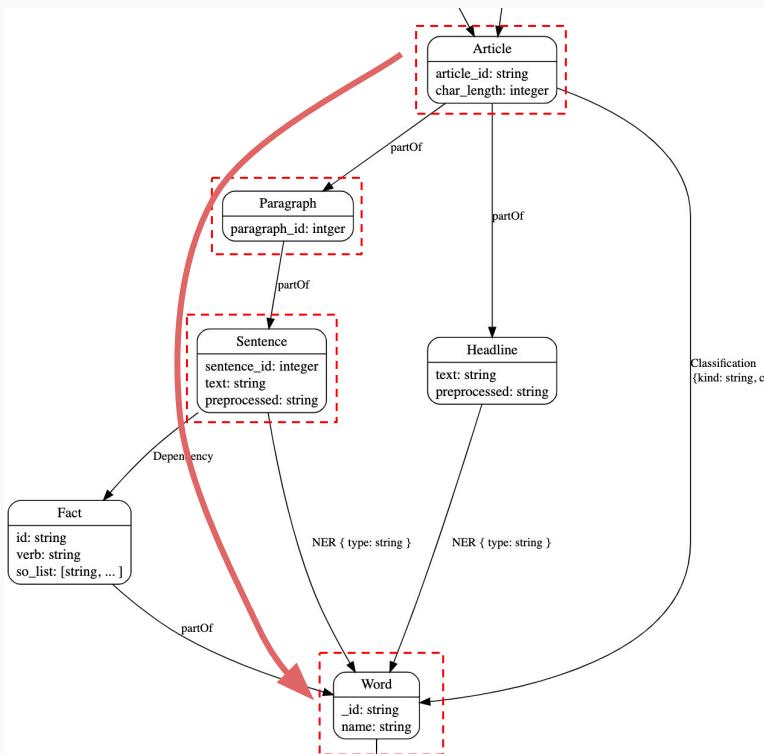
Page



Article

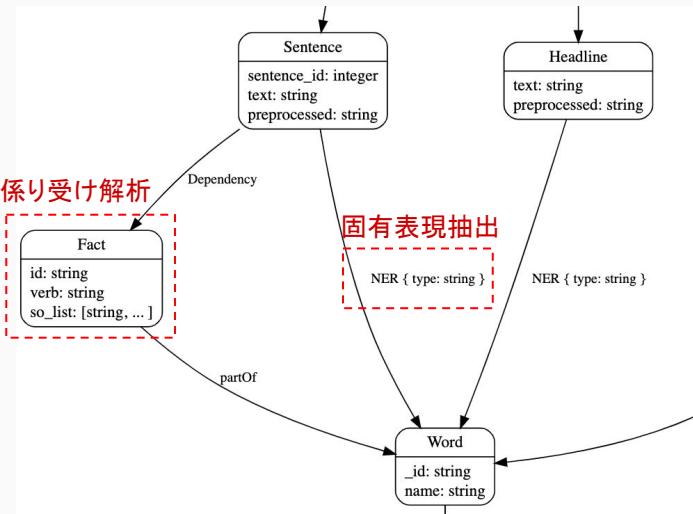
グラフ構造化による表現 ②

- ある単語を含む文章、パラグラフ、記事の抽出
 - 階層構造を



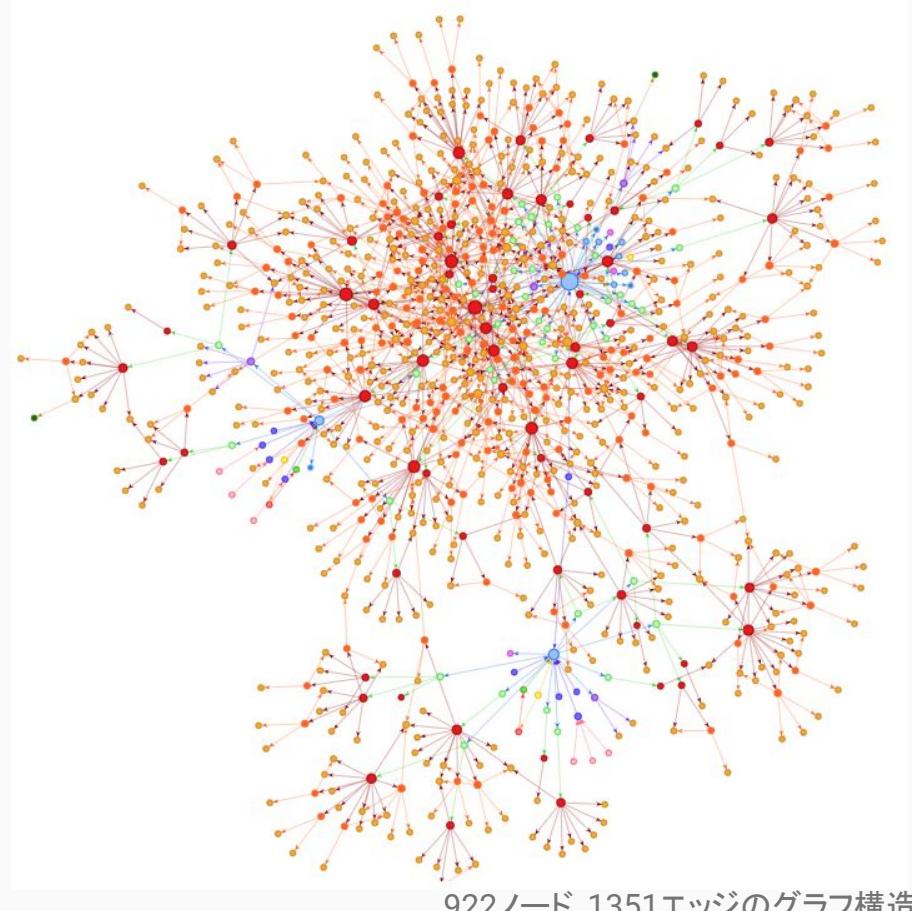
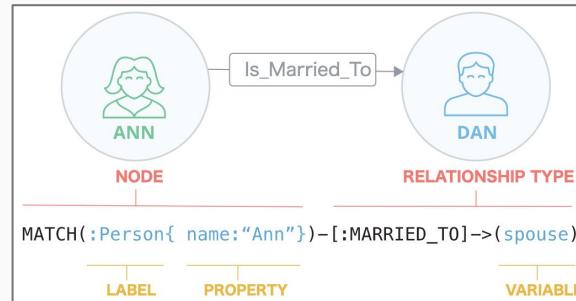
グラフ構造化による表現 ③

- 固有表現抽出 (NER)
 - 固有表現(企業・人名・地名)を利用
 - 抽出元の文章と語をつなぐエッジで表現
 - 文章によって役割が変わるような語句にも対応
- 係り受け解析
 - 構文解析による SVO-構造を抽出
 - 抽出元の文章と語の間にノードとエッジで表現
 - 文章中に「特定の語を含む係り受け構造」があるかどうかを確認することができる



抽出されたグラフ構造をグラフデータベースへ格納

- 記事からグラフ構造を抽出
 - 1記事が数百ノードの構造をもつ
 - 半年分の記事データで1000万ノード
- グラフデータベースに格納
 - グラフ構造をそのまま蓄積できるグラフデータベース Neo4j を用いる
 - クエリ言語 Cypher による直感的な抽出が可能



共起語の時系列変化

● データ

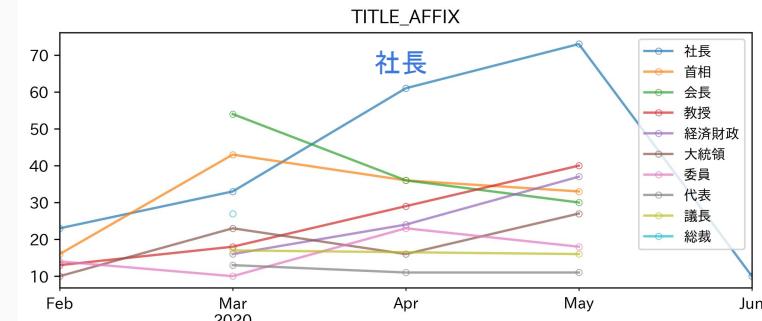
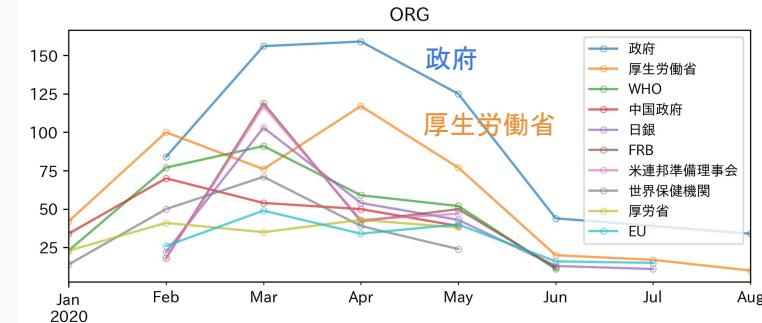
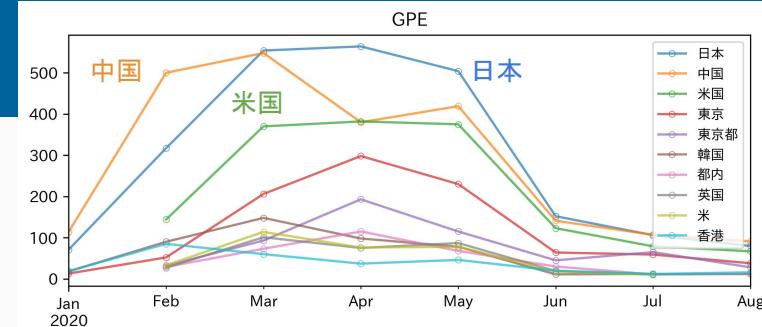
- 媒体: 日本経済新聞 朝刊
- 期間: 2020/01/01 ~ 2020/08/31
- 粒度: 月次集計

● 設定

- パラグラフでの共起
- 「新型コロナウイルス」との共起

● 固有表現ごとの単語の頻度

- 地域: 日本, 中国 (2月にピーク), 米国が多い
- 組織: 政府, 厚生労働省が多い
- 役職: 社長 (5月にピーク) が多い



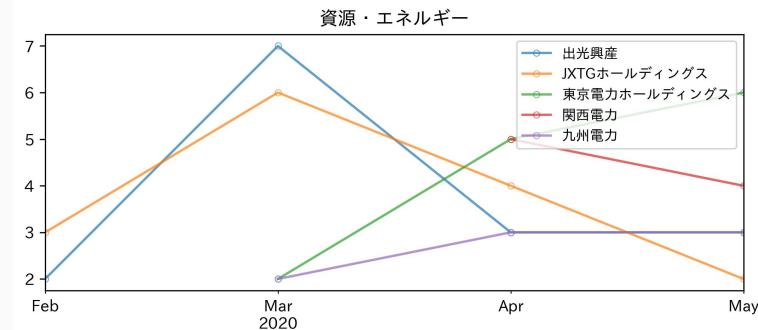
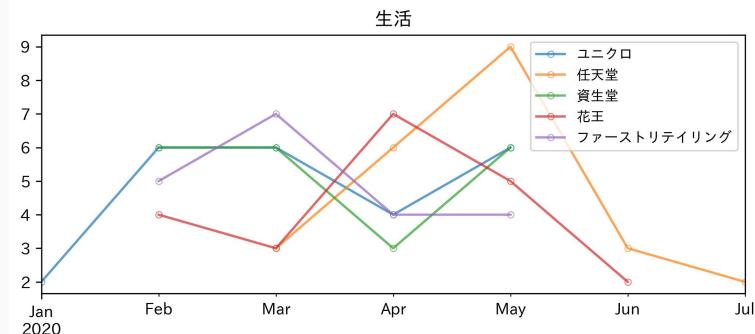
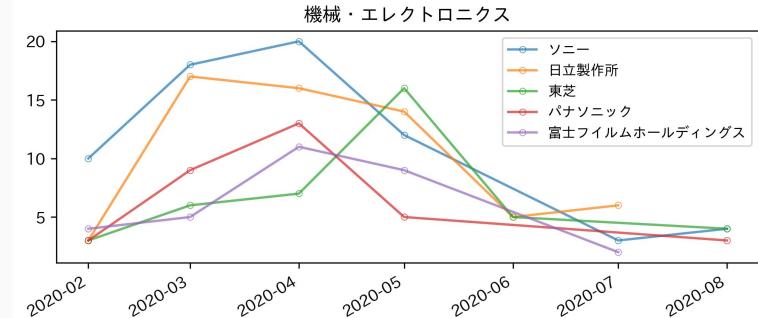
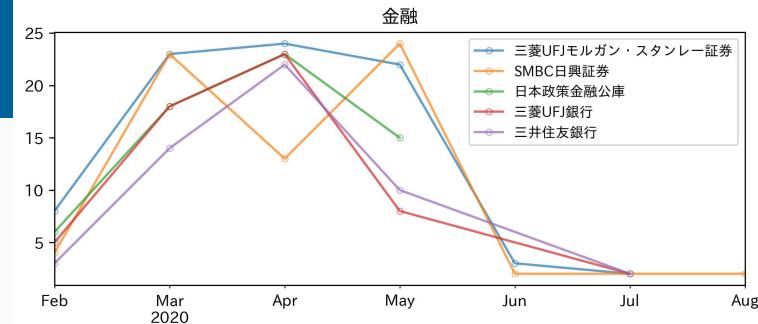
共起語の時系列変化

● データ

- 媒体: 日本経済新聞 朝刊
- 期間: 2020/01/01 ~ 2020/08/31
- 粒度: 月次集計

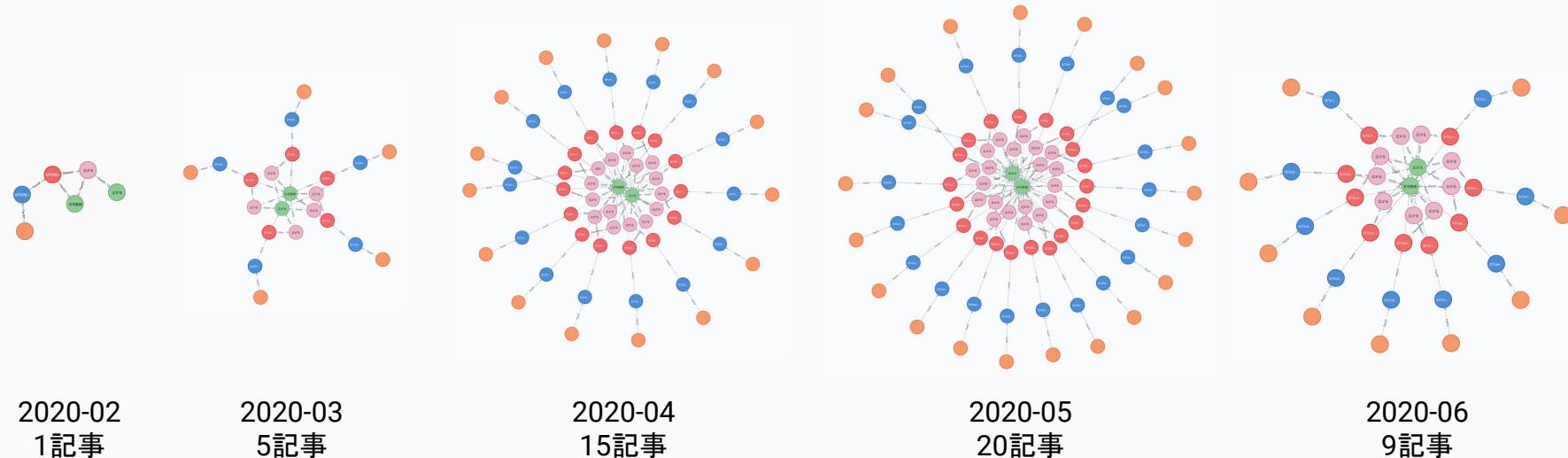
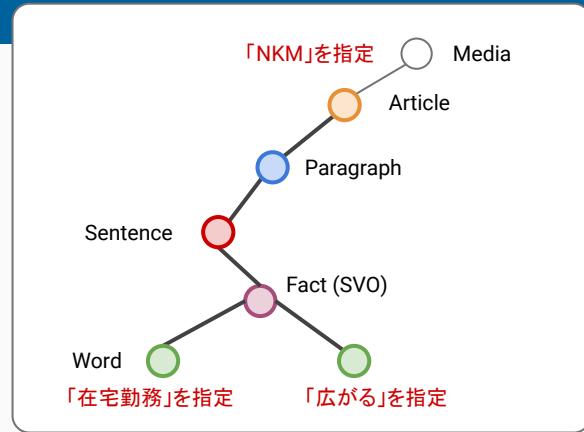
● 設定

- パラグラフでの共起
- 「新型コロナウイルス」との共起
- 固有表現が “ORG” かつ日経会社コードあり
- 業種ごとの単語の頻度



係り受けの時系列な変化

- データ
 - 日本経済新聞 朝刊, 2020/01/01 ~ 06/31, 月次集計
- 係り受けの構造を検索
 - 在宅勤務が普及していく様子を知りたい
 - 例えば「広がる」という表現に該当する記事を検索したい
 - 「在宅勤務」と「広がる」は係り受けの関係になる

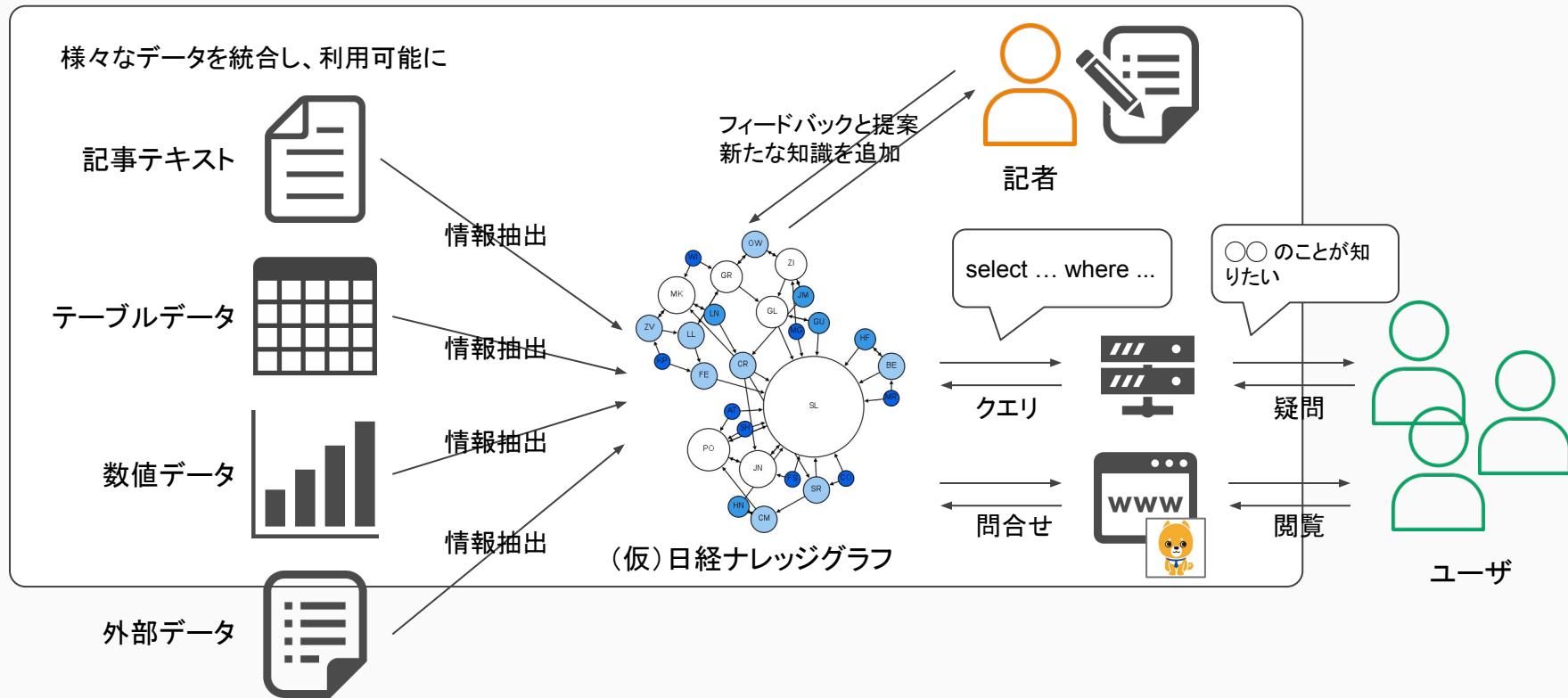


日経のメディア・情報サービスへの適用

- 適用先のイメージ
 - メディア(電子版など)のバックエンド
 - 情報サービスへの適用
 - 記者へのフィードバック連携
- 日経が展開する情報サービス
 - 「NEEDS」財務データ(数値データ)
 - 「日経バリューサーチ」企業情報データベースサービス
 - 「日経テレコン」記事検索サービス
- 様々な情報を構造化し、連携させることでより良い情報発信を実現
 - 様々な情報・データをつなげて、探索的な情報提供を実現したい

(仮)日経ナレッジグラフ構想

- 社内外のデータから知識体系を構築し、蓄積・成長させていく



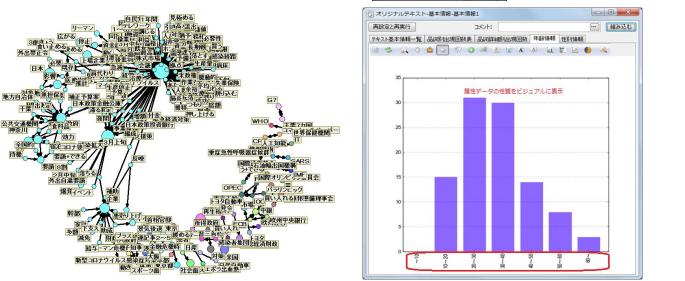
まとめ

- **日経記事データ**
 - 記事データ販売、AWS Data Exchange での販売
 - 学生研究奨励賞への記事テキストデータ提供
- **日経記事データ x TMS**
 - 日経記事テキストと相性が良い分析ツール
- **日経記事データの構造化**
 - 記事がもつ階層構造と自然言語処理による情報抽出を用いて、グラフ構造を抽出
 - グラフ構造はグラフデータベースに蓄積して、クエリにより利活用

日経記事データ

項目名	値	サンプル
記事ID	"NFRG0320150719NO540351" など	NFRG0320200218NO50046
発表日	2020-06-20T18:10:32Z (UTC)	
記事名	「政治のニーコード」	NAV
登録情報	「日経新聞新規」、「日経経済ニュースアーカイブ」など	日本経済新聞社
登録元	「政治」、政治家、政治家	
記事本文数	1	
記事本文字数	1文字	721
記事表示	見出し	P.C.R実験、企業で犯人、大学の2割が実験犯、海外移動に必要なもの。
本文		新型コロナワクチン接種を実施するP.C.R検査の需要が大企業の間で広がっている。日本経済新聞が2月8日付の報道によると、企業は従業員の健康状態を把握するための検査として、また、海外出張者に対する感染対策として、接種を実施する動きがある。政府は、接種券を交付する方針を示す一方で、接種の実施には個人の意思を尊重する方針だ。接種券を手に入れた従業員は、自ら接種を希望する場合、会社が費用を負担する形で接種できる。一方で、接種券を手に入れた従業員は、自ら接種を希望する場合、会社が費用を負担する形で接種できる。一方で、接種券を手に入れた従業員は、自ら接種を希望する場合、会社が費用を負担する形で接種できる。一方で、接種券を手に入れた従業員は、自ら接種を希望する場合、会社が費用を負担する形で接種できる。
キーワード	記者の文から自動的に切り取ったワードまたはその近似名	三井住友フルブルク・ラジオ・システムズ株式会社の社長の名前と、P.C.R検査の名前。
分類	政治	日本国内の「ニーコード」(約1万件)、東京コード(約1万件)、福岡コード(約1万件)、大阪コード(約1万件)、北海道コード(約1万件)、沖縄コード(約1万件)、その他(約1万件)。
特記事コード		8586-4843-5108-9201-0141-0024-8555

TMS によるテキストマイニング



記事のグラフ構造化

