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

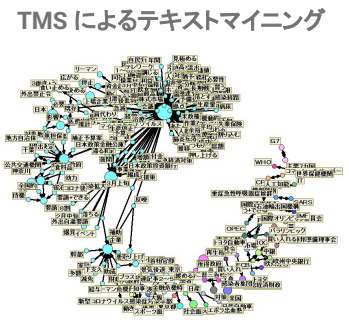
まとめ

- 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

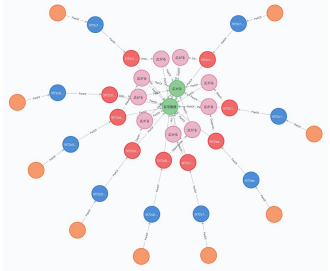
日経記事データ

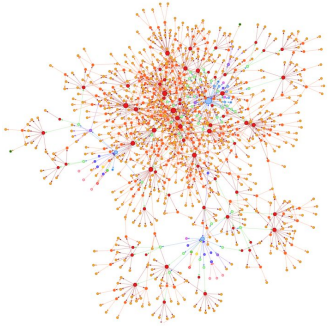
項目名	備考	サンプル
記事ID	"NRKDB20150715NKM0335" など	NRKDB20200628NKM0048
掲載日	"2016-09-25T23:33:26Z" など	2020-06-29T18:13:32+09:00
媒体略号	媒体のユニークコード	NKM
媒体名称	"日本経済新聞 朝刊"、"日経速報ニュースアーカイブ" など	日本経済新聞朝刊
紙・写真・表の有無	"○×"等	
記事本文段落数	段落数	6
記事本文文字数	文字数	721
記事見出し	見出し	ＰＣＲ需要、企業で拡大、大手の２割が実施検討、海外渡航に必要な例も。
記事本文	本文	新型コロナウイルス感染を判断する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５社に聞き取り調査したところ、三菱商事が国に検査を実施したお回答した。全体の約２割の７６社が「実施を検討中」と答えた。海外事業の出張が深い機械や自動車関連の企業で強弁渡航時に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） 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、最近は重症の疑いがない軽症者であっても受け付ける民間の診療所が増えている。 アンケートでは「希望する社員全員にＰＣＲ検査を受けさせる計画はあるか」と含む答に聞いた。三菱商事は海外赴任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。検査には当引診療所に加え外部の検査機関を活用している。 「検討中」はプリデストンやＤＭＧ森精機、日本航空やアルフレなどの１６社。東京エレクトロンは海外渡航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）指示証明書（１面参より）のこと
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,プリデストン,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(##Y~)、業界コード(##~)、話題コード等の分類コード(Y~、N~、PD~)、組織名等の記事分類キーワード($~)、コラム名(^~)	T60$58,PD43$17,E141,PD211,T924,PD$51,T835,PD239,T4543,PD313,T5108,PD131,T921,PD611,$~農,#K7,#B0320,#B0470,N000152,N0001154,N0038618,N0001738,N0013612,N0000697,N0001655,N0001941
株式コード	東証コード	8058,4543,5108,9201,6141,9024,8035

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

- 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

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

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

アプトポッド、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）が提供するサードパーティ製ソフトウェアを、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販路を開拓でき、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ドパーティ製ソフトウェアの検索、テスト、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 のコンソールパ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	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） 　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。 　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。 　「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
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,Ｎ０００００６９７,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	"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	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） 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。 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。 「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
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(S~)、コラム名(「~」)	T８０５８,P０４３１,T６１４１,P０２１１,T９０２４,P０５５１,T８０３５,P０２３９,T４５４３,P０３１３,T５１０８,P０１３１,T９２０１,P０６１１,$一面,＃Ｋ７,＃Ｂ０３２０,＃Ｂ０４７０,Ｎ０００１５９２,Ｎ０００１１５４,Ｎ００３８６１８,Ｎ０００１７３８,Ｎ００１３６１２,Ｎ０００００６９７,Ｎ０００１６５５,Ｎ０００１９４１
株式コード	東証コード	8058,4543,5108,9201,6141,9024,8035

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） 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。 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。 「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

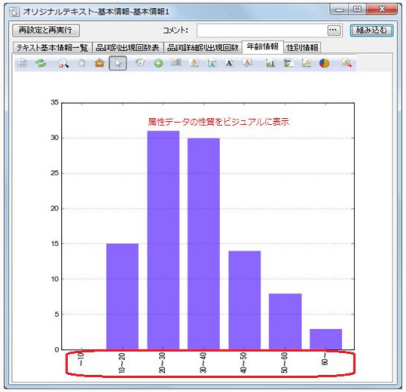
TMS 向け前処理カラム（イメージ）

NTTデータ数理システム Text Mining Studio

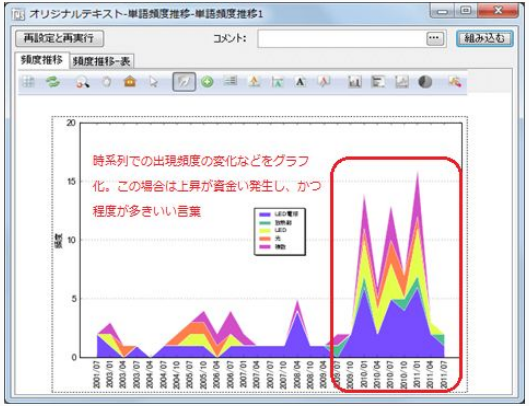
- NTTデータ数理システム Text Mining Studio
  - 手間のかかるテキストマイニングを簡単に実現
  - 日経テキストデータと相性が良いとの評価あり

- 分かち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

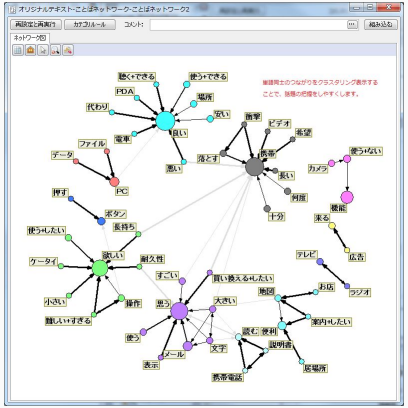
単語の頻度



単語の頻度の時系列変化



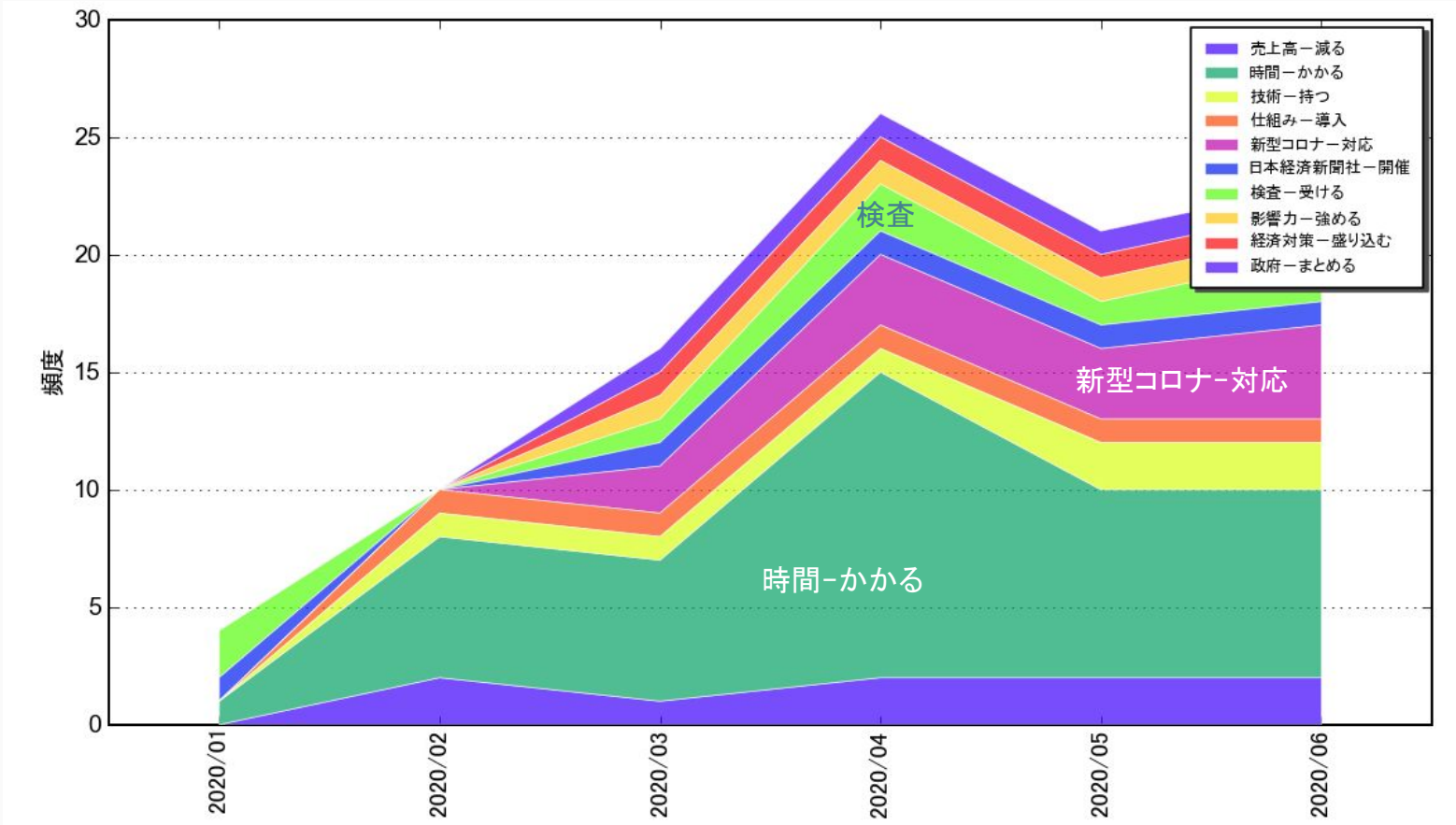
共起ネットワーク、係り受けネットワーク



係り受けの関係、時系列変化（上昇傾向）

Text Mining Studio

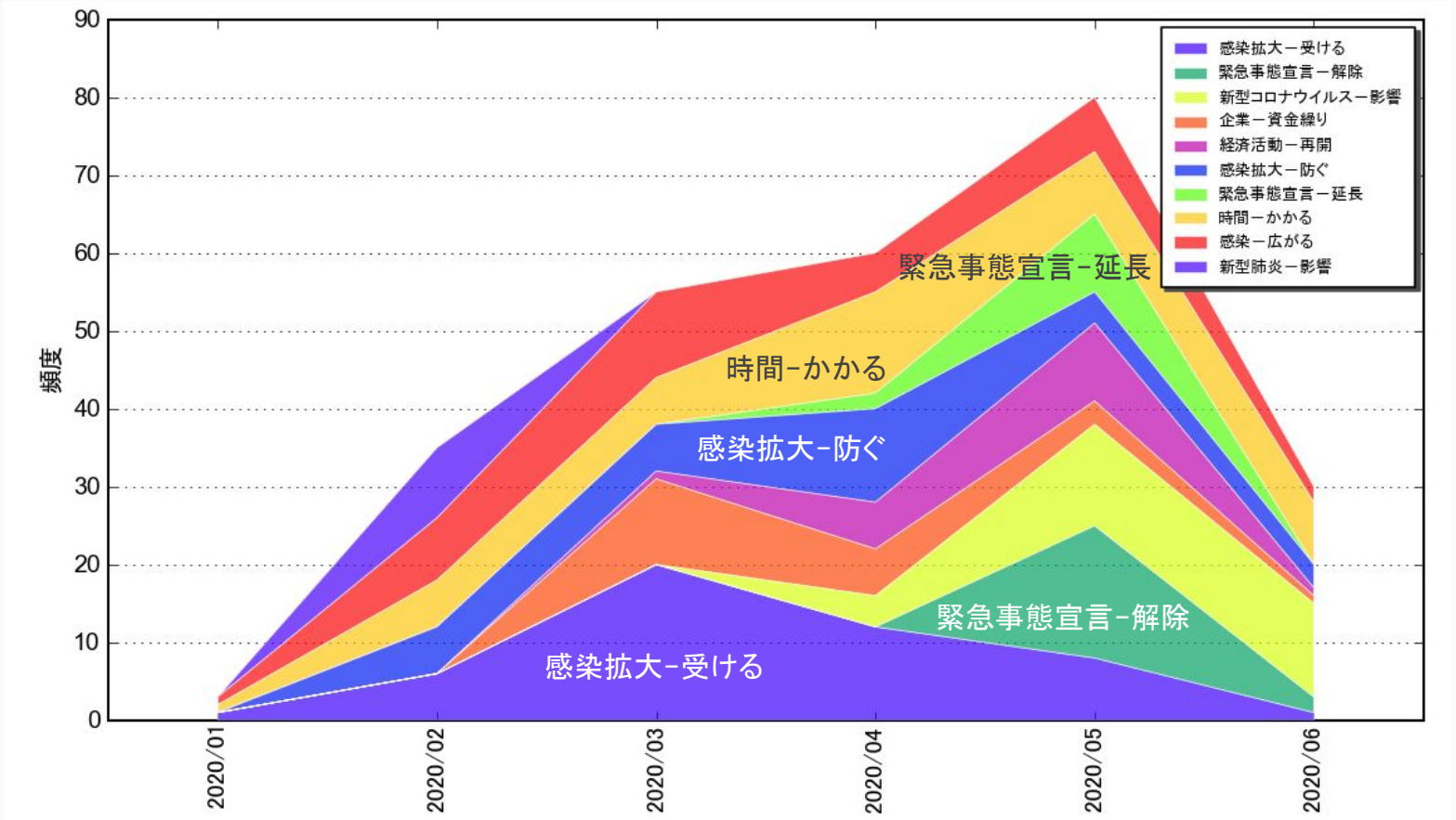
2020/01/01 ～ 06/30、日経新聞朝刊、一面に掲載された 9,819 記事（抜粋）



係り受けの関係、時系列変化（頻度の変動が大）

Text Mining Studio

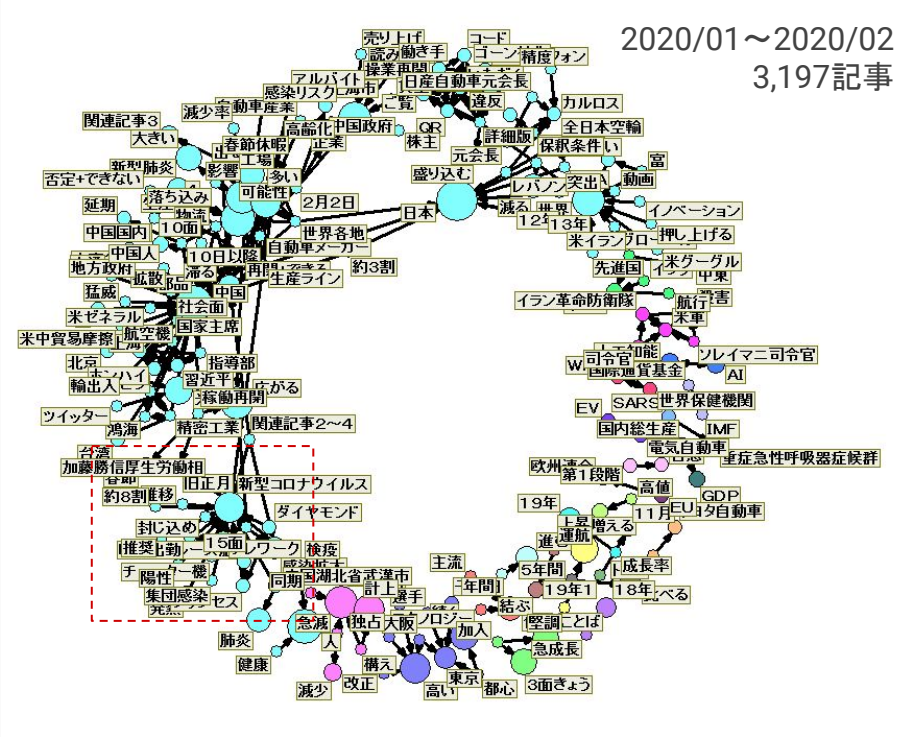
2020/01/01 ～ 06/30、日経新聞朝刊、一面に掲載された 9,819 記事（抜粋）



共起ネットワーク

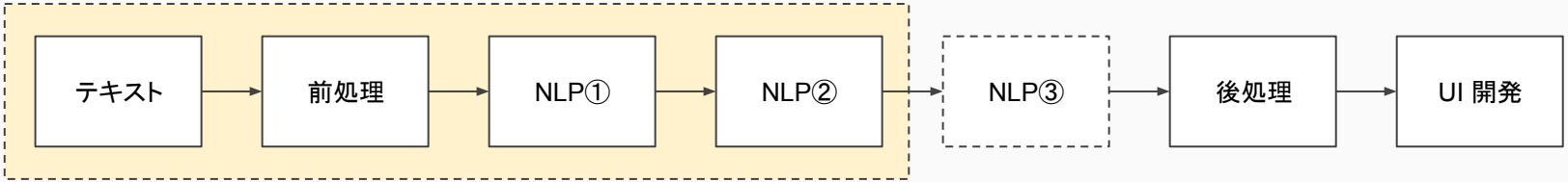
Text Mining Studio

- 「新型コロナウイルス」への集中を確認できる

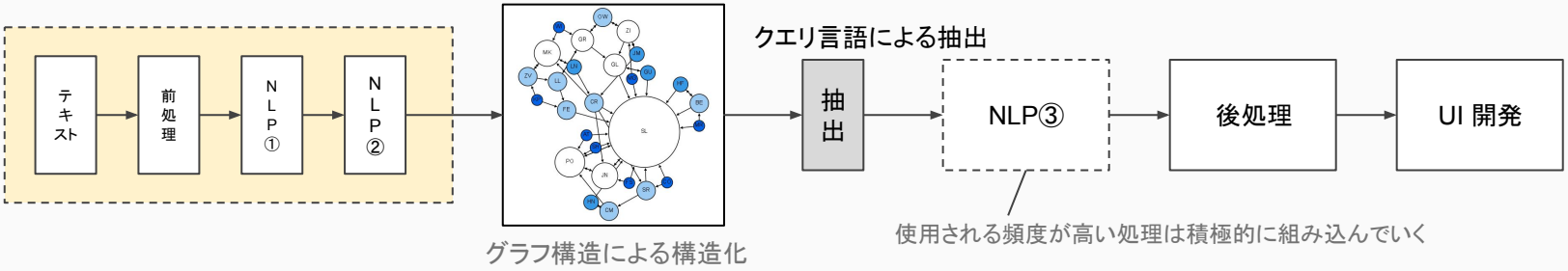


典型的な処理を構造化し、ソフトウェア資産を利活用

- 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



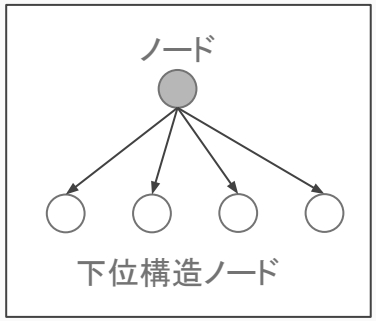
- 典型的な処理は柔軟な構造で蓄積しクエリ言語で抽出
  - 工程を典型的な処理と、個別の処理に分割する
  - 典型的な処理は柔軟なグラフ構造（グラフデータベース）で蓄積し、クエリ言語で抽出



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

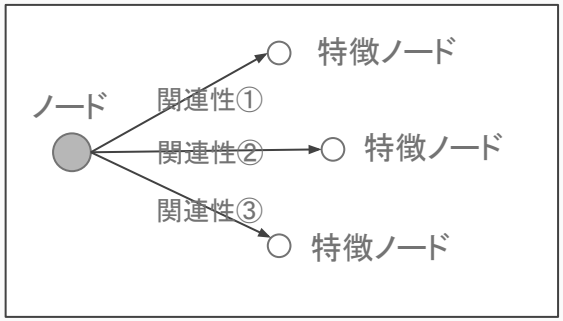
階層構造



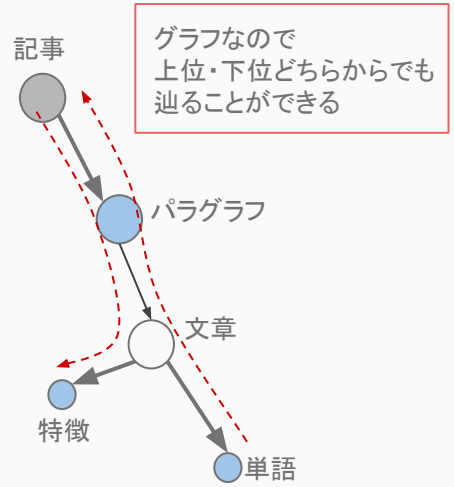
記事の階層構造 分類の階層構造など様々な下位構造を明示的に表現

+

情報抽出



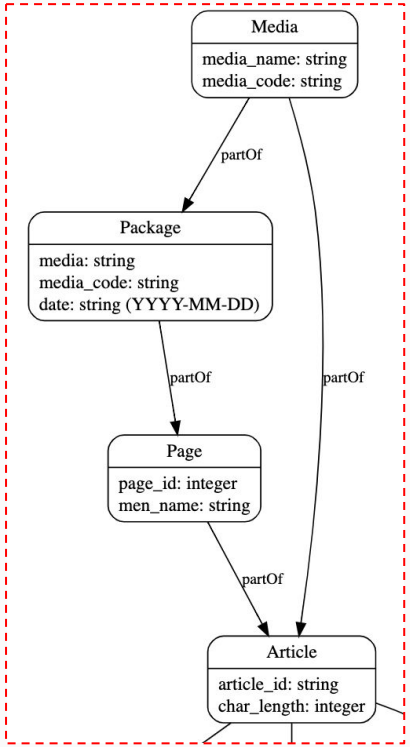
データの対象がテキストならば自然言語処理、動画像ならばオブジェクト認識など、様々な手法を用いる



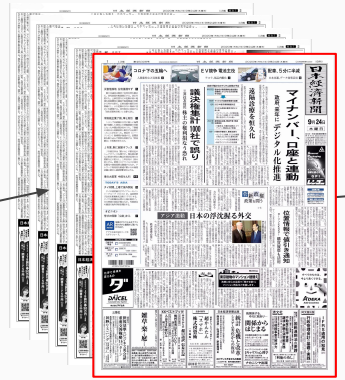
ある特徴(単語を含む、タグ)をもつ記事を抽出

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）



Media



Package

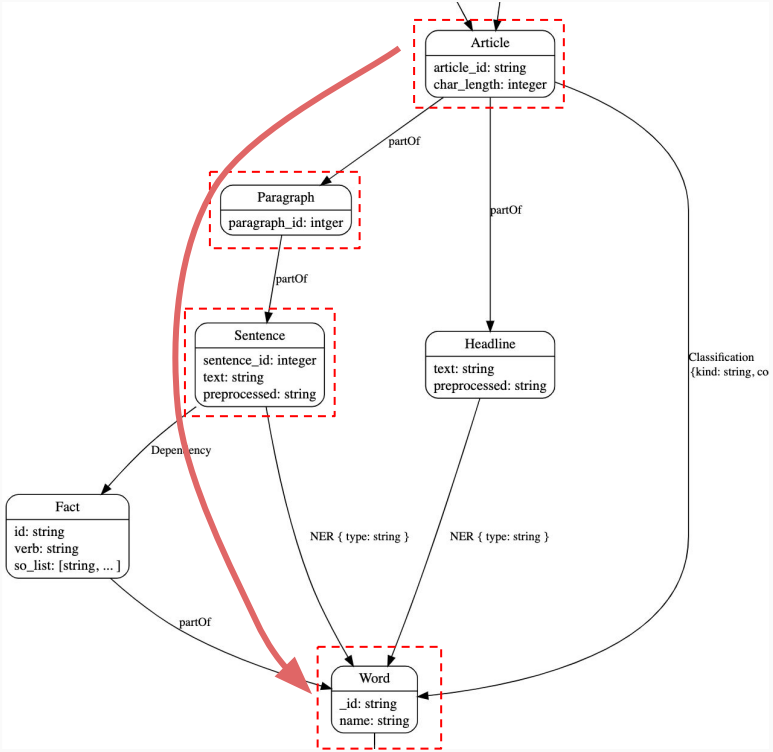
Page

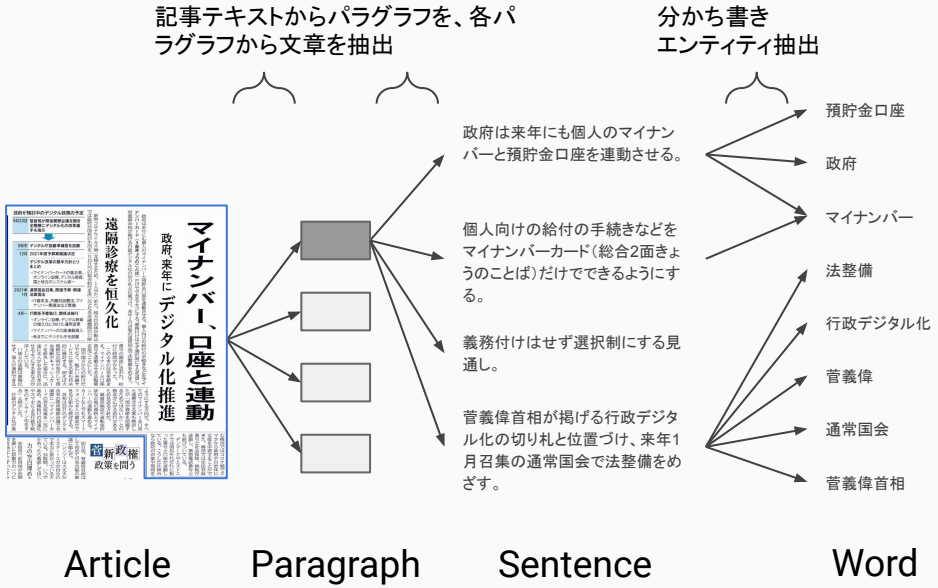
Article

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

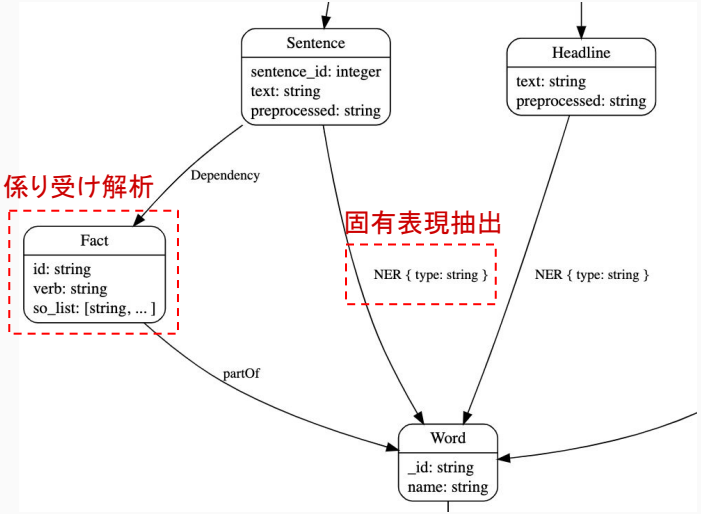




グラフ構造化による表現 ③

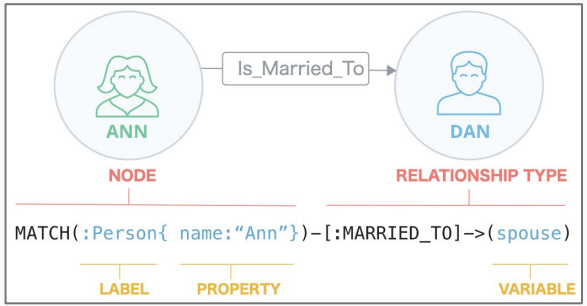
- 固有表現抽出 (NER)
  - 固有表現（企業・人名・地名）を利用
  - 抽出元の文章と語をつなぐ **エッジ** で表現
  - 文章によって役割が変わるような語句にも対応

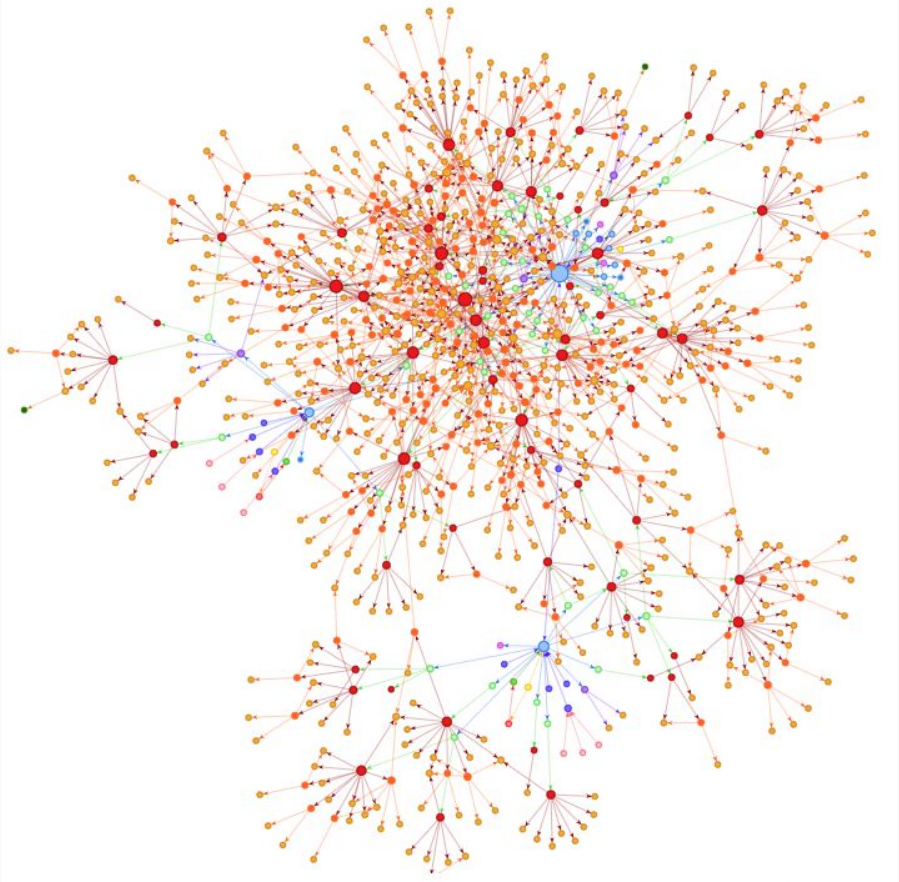
- 係り受け解析
  - 構文解析による SVO-構造を抽出
  - 抽出元の文章と語の間に **ノード** と **エッジ** で表現
  - 文章中に「特定の語を含む係り受け構造」があるかどうかを確認することができる



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

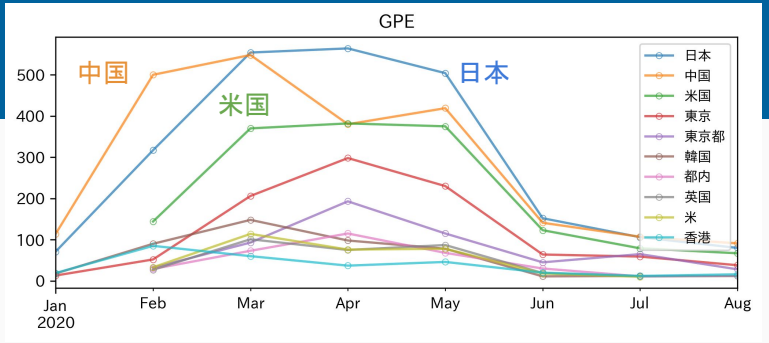


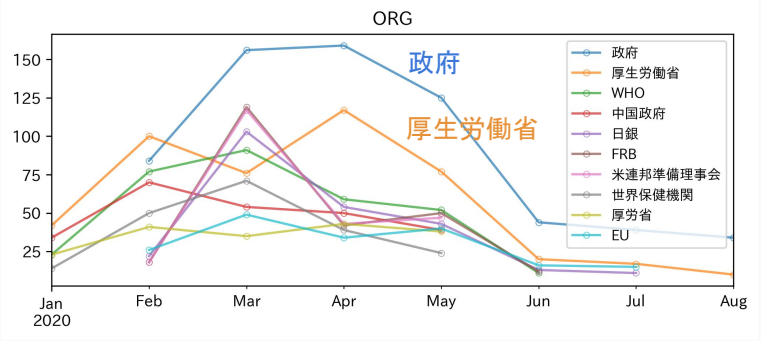


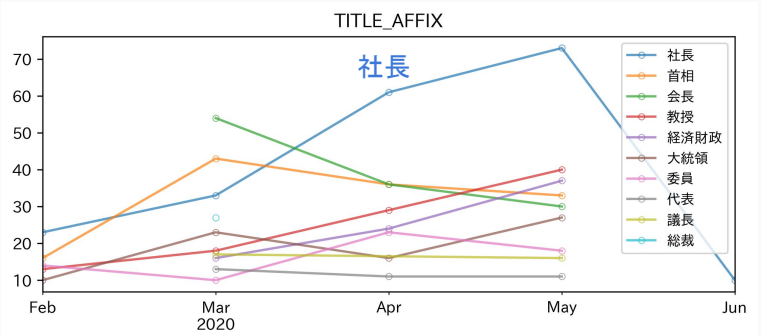
922ノード, 1351エッジのグラフ構造

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

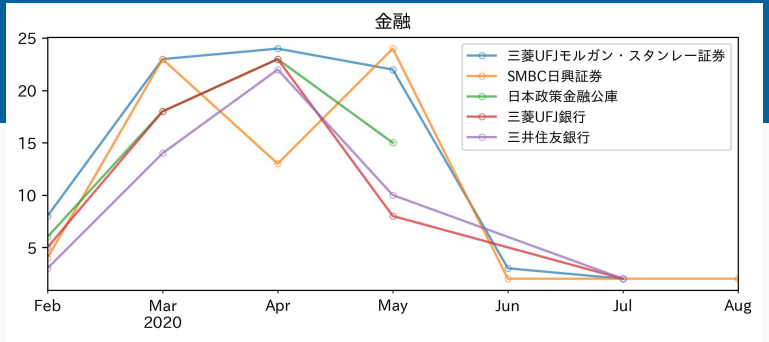


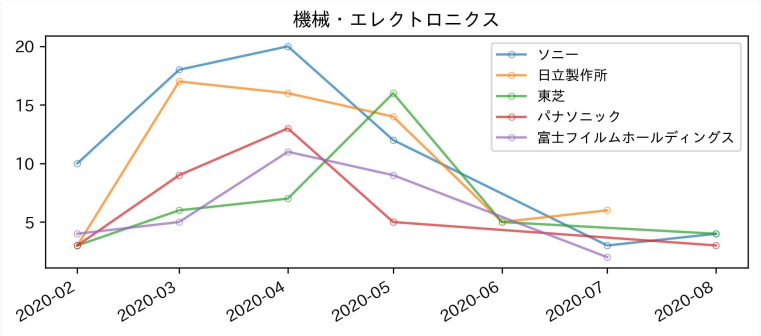


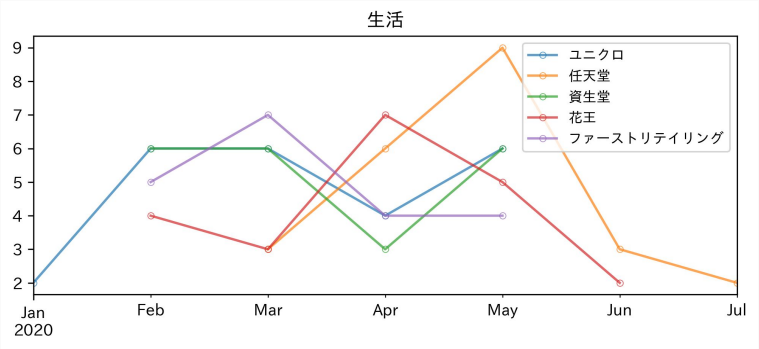


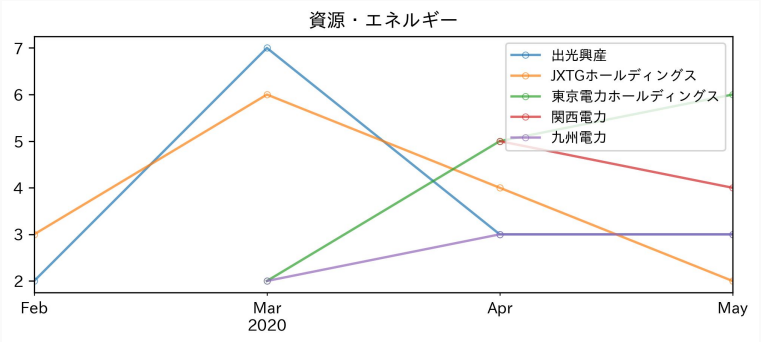
共起語の時系列変化

- 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



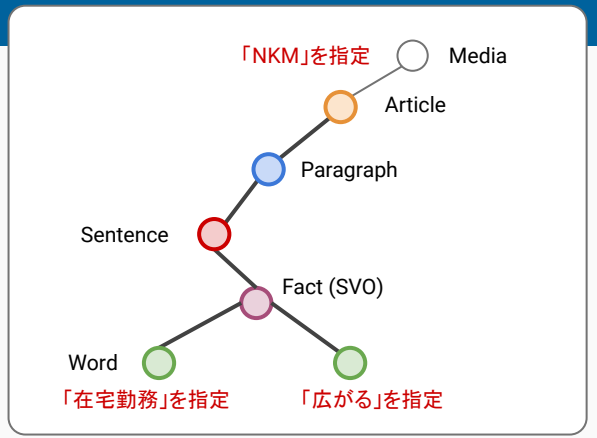


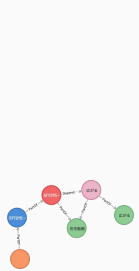




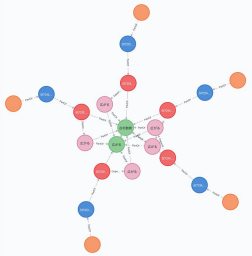
係り受けの時系列な変化

- 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

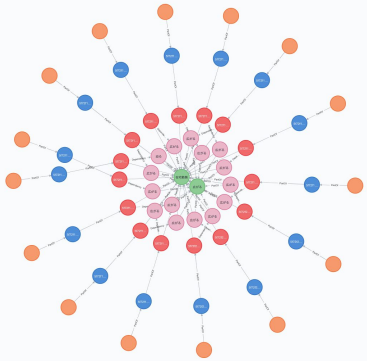




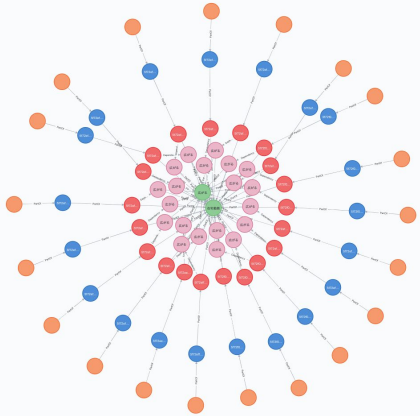
2020-02
1記事



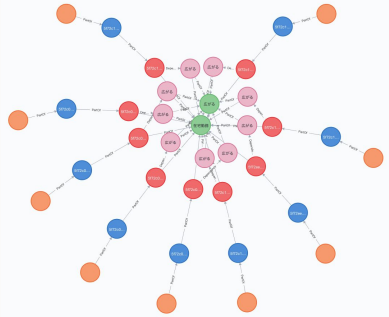
2020-03
5記事



2020-04
15記事



2020-05
20記事



2020-06
9記事

日経のメディア・情報サービスへの適用

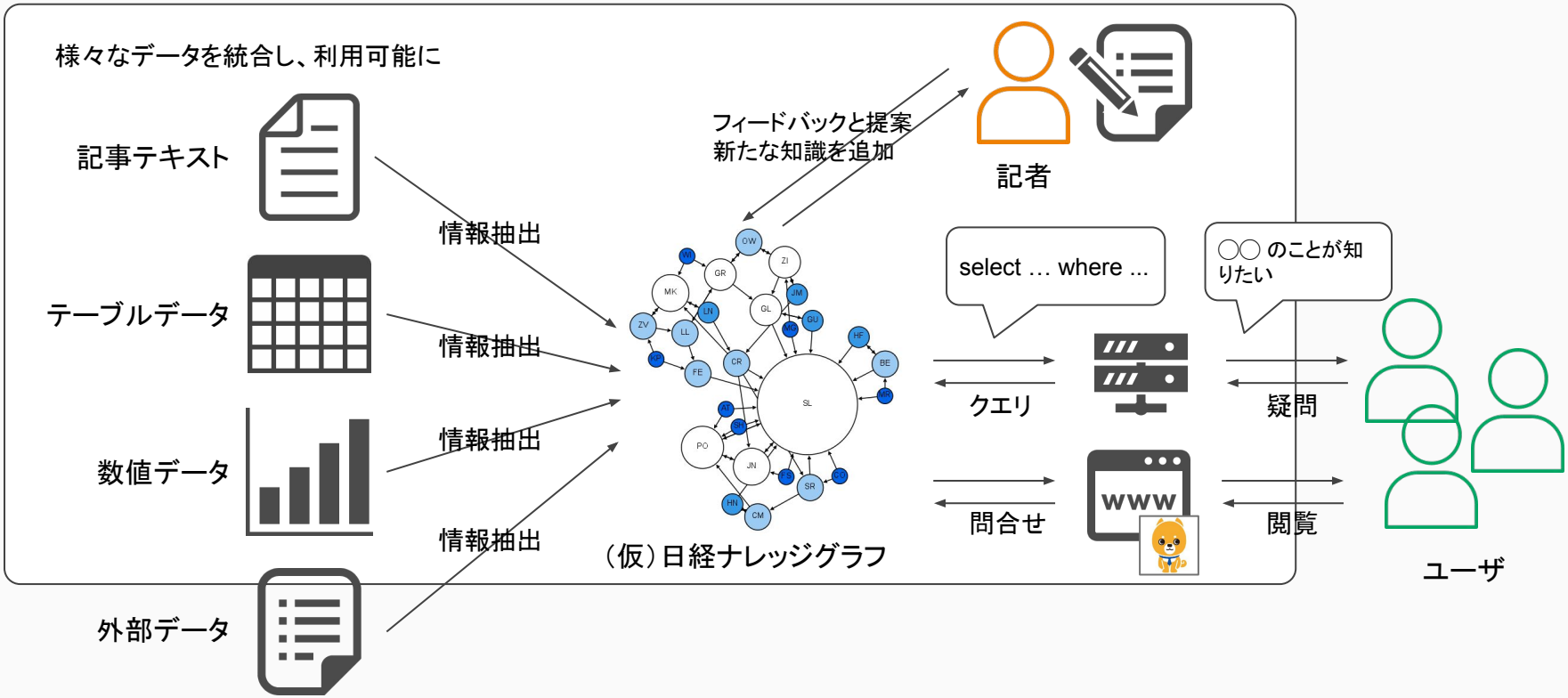
- 適用先のイメージ
  - メディア（電子版など）のバックエンド
  - 情報サービスへの適用
  - 記者へのフィードバック連携

- 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



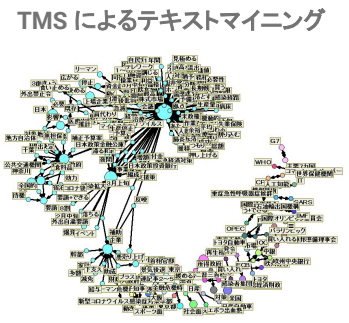
まとめ

- 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	備考	サンプル
記事ID	"NRKDB20150715NKM0335" など	NRKDB20200628NKM0048
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
紙・写真・表の有無	"○×"等	
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	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５社に聞き取り調査したところ、三菱商事が国に検査を実施したお回答した。全体の約２割の７６社が「実施を検討中」と答えた。海外事業の出張が深い機械や自動車関連の企業で強弁渡航時に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） 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、最近は重症の疑いがない軽症者であっても受け付ける民間の診療所が増えている。 アンケートでは「希望する社員全員にＰＣＲ検査を受けさせる計画はあるか」と含む答に聞いた。三菱商事は海外赴任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。検査には当引診療所に加え外部の検査機関を活用している。 「検討中」はプリデストンやＤＭＧ森精機、日本航空やアルフレなどの１６社。東京エレクトロンは海外渡航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）指示証明書（１面参より）のこと
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,プリデストン,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(##Y~)、業界コード(##~)、話題コード等の分類コード(Y~、N~、PD~)、組織名等の記事分類キーワード($~)、コラム名(^~)	T60$58,PD43$17,E141,PD211,T924,PD$51,T835,PD239,T4543,PD313,T5108,PD131,T921,PD611,$~農,#K7,#B0320,#B0470,N000152,N0001154,N0038618,N0001738,N0013612,N0000697,N0001655,N0001941
株式コード	東証コード	8058,4543,5108,9201,6141,9024,8035

TMS によるテキストマイニング



記事のグラフ構造化