

ビッグデータ分析技術ワークショップ
～高速グラフマイニングが切り拓く世界～



Graponトレンド分析の紹介

NTTソフトウェアイノベーションセンタ
2018年3月9日



高速グラフマイニング技術“Grapon”を、トレンド抽出，論文発見に活用した事例を紹介し、グラフが有効だった点、課題、今後の展望を述べる。

<アジェンダ>

- はじめに
- 事例 1：ニュース記事からのトレンド抽出
- 事例 2：関連論文の発見
- まとめと今後の展望



はじめに

- グラフ技術の注目度
- グラフマイニングの利点・分析例

[illegible]

グラフ関係の発表は全セッションのうち4分の1を占める

- 24 session のうち6セッションがグラフ関係
- 香港, 中国, シンガポール, 韓国からの発表が多い

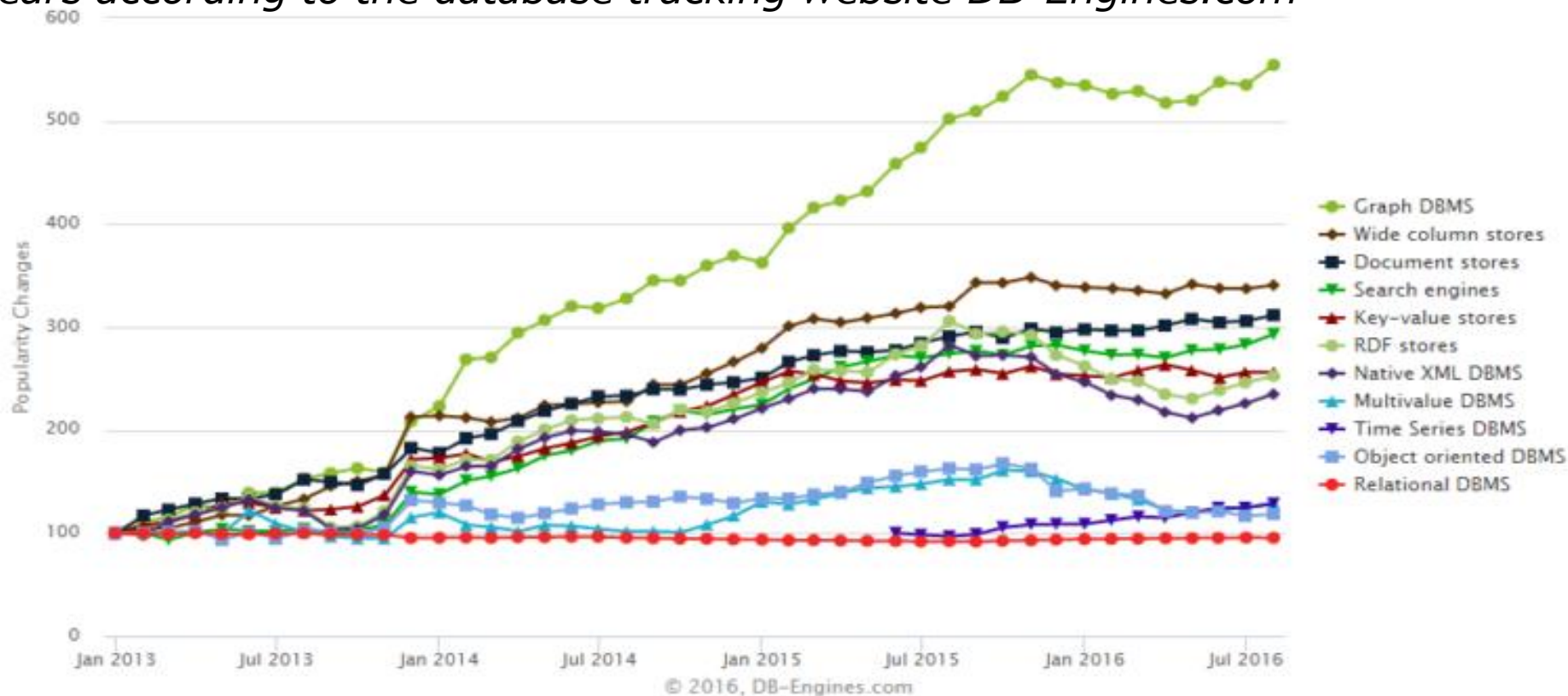
出典：Gartnerレポート（2015年時点）



グラフ技術の動向

実際，グラフDBの人気はこの3年で台頭。

Graph databases have seen the greatest rise in popularity over the past three years according to the database tracking website DB-Engines.com



グラフ技術の動向



昨今では各社がグラフDBをサポート。ITPro記事によれば2018年の注目ITインフラのひとつ。

企業	紹介時期	製品・サービス名	備考
Oracle	2014年10月	Oracle Spatial Graph	
SAP	2015年6月	HANA Graph	
GE	2016年9月	Predix	Titanがベース
Linux(Google, IBMの支援によるもの)	2017年1月	JanusGraph	Titanがベース
Microsoft	2017年10月	SQL Server 2017	グラフDBを搭載
Amazon	2017年11月	Neptune	SPAQL, Gremlin

2018年のITインフラ大胆予測

超広域分散DBやローコードなど14分野、注目株のITインフラ技術

島津 忠承 = 日経SYSTEMS

2018/02/02

SYSTEMS

グラフDBは
2018年の注目
分野のひとつ

グラフデータベース

NoSQLデータベース(DB)の一種。人や物などのノード同士のつながり、関係をグラフ構造で表現する。主な製品にはオープンソースソフトウェア版と商用版の両方がある「Neo4j」、米オラクルの「Oracle Spatial and Graph」などがある。

RDBであれば大量の結合が必要で処理が遅くなる複雑なデータを扱う用途に向く。例えば、ネット通販のレコメンド機能、地図の経路探索などである。さらに、富士通が独自のAI技術体系「Jinrai」において中核技術の一つとしてグラフDBを取り込むなど、用途が広がりつつある。

出典：IT Pro 2018年2月2日

実務のグラフ活用はアカデミアでも関心事項。2017年のVLDB論文には「グラフは実務でどのように使われているか？」の調査結果が掲載された。

VLDB2017 The Ubiquity of Large Graphs and Surprising Challenges of Graph Processing

Table 4: Real-world entities represented by the participants' graphs and studied in publications. Legend for non-human entities: *Products* (NH-P), *Business and Financial Data* (NH-B), *Web Data* (NH-W), *Geographic Maps* (NH-G), *Digital Data* (NH-D), *Infrastructure Networks* (NH-I), *Knowledge and Textual Data* (NH-K).

Category	Human	RDF	Scientific	Non-Human	NH-P	NH-B	NH-W	NH-G	NH-D	NH-I	NH-K
Total	45	23	15	60	13	11	4	7	5	9	11
R	18	11	9	22	1	6	2	4	1	7	6
P	27	12	6	38	12	5	2	3	4	2	5
A	54	8	11	63	2	8	30	11	0	2	3

驚くべきことは、RDBMSが完全にフィットするタスクもグラフで扱うことがあるということ

- NH-P：取引グラフ（製品・注文・購買）、製品と部品の調達グラフ ⇒ 似た製品をクラスタリングに関心
- NH-B：企業資産の保有、資金、ビットコイン、マルウェア検体の関係
- NH-G：道路、自転車シェアリングステーション、景観スポット
- NH-D：ファイルとフォルダ、ビデオとキャプション
- NH-I：油井とパイプライン、無線センサーネットワーク
- NH-K：キーワード、用語定義

グラフを活用した分析例

NTTソフトウェアイノベーションセンタではグラフ活用を進めている。

グラフ構造が自明な場合（交友関係，道路の接続関係など）での分析例

災害下でのTwitterインフルエンサー発見



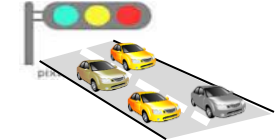
奈良先端科学技術大学院大学様との共同研究

交通制御に向けた広域道路網の最適メッシュ分割



NTTデータ様での活用

予測・信号制御・渋滞回避



ベクトル量からグラフを構築（センサーデータ，画像，文書など）

状態の類似性で満空を分類



周りの混雑状況をもとに空いている駐車場を推定

慶應義塾大学様との共同研究

画像の類似性から，路面状態を分類

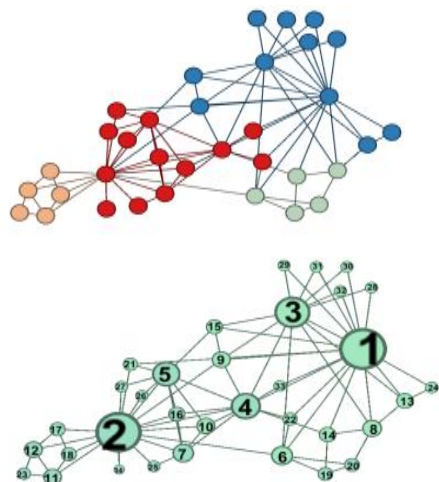


路面状態の類似性から正常な路面と劣化した路面とを判別

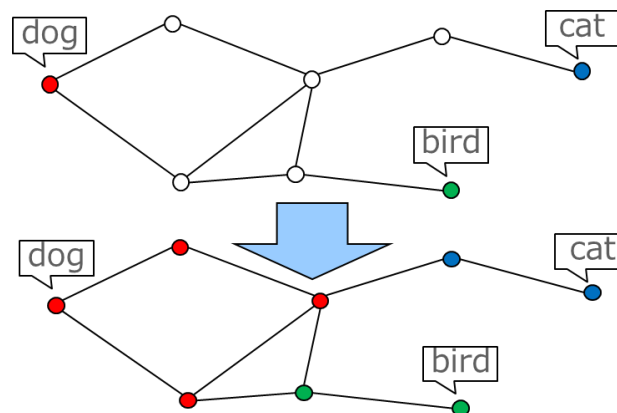
慶應義塾大学様との共同研究

- グラフ構造データの検索、マイニングを**効率的に処理可能**
- グラフ構造があるために**教師データが少なくても高精度な分析が可能**
- **可視化**との親和性が高い

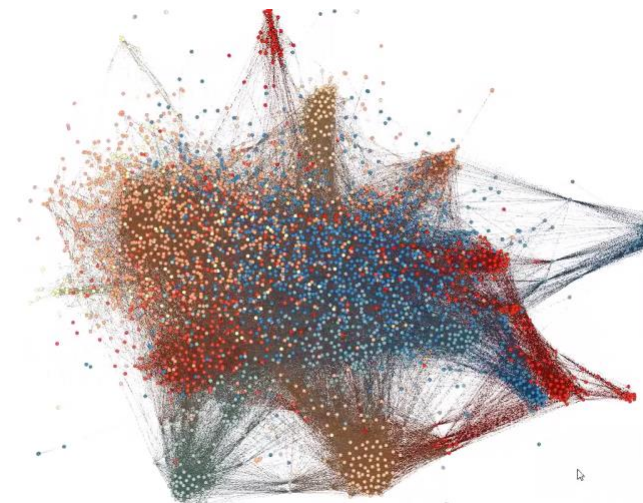
つながりを効率的に処理



少ない教師データから高精度に分析
(ラベル伝播)



可視化による情報把握の容易化
特徴的な部分をすばやく発見



事例 1 : ニュース記事からのトレンド抽出

- 小規模なデータセットでトレンド抽出試行を開始。検討状況を紹介

ニュース記事からトレンド抽出を試行

2017年のIT Proのうち、技術的なトピックスを題材に選定

データセット1

[AI／機械学習]のみの20記事

IT Pro 2017年アクセスランキング 詳細

◆ 期間：2017年1月1日～12月10日

◆ 対象分野：8分野のニュースを検証用に抽出（1分野あたり約20記事）

2017年 アクセスランキング



データセット2

8分野のTOP20記事

[xTECH]
[ネットワーク]
[クラウド]
[IoT]
[システム]
[総合]
[セキュリティ]
[AI／機械学習]

1. [電子行政] マイナポータルと情報連携に振り回された1年
2. [FinTech] いったいどれだけの人が財布を無くすのか
3. [スキルアップ] エンジニアなら誰もが気になる、昇進に役立つIT資格
4. [スタートアップ] 勝者しか住めないシリコンバレー、浮沈の激しさにも翻弄
5. **[xTECH] トヨタもコマツもソフトバンクも、クロスこそが変革の要**
6. [ソフト開発] なくならない受発注者間の争い、なぜシステム裁判は起こるのか
7. **[ネットワーク] 使いものにならない光回線、グーグル発の大規模障害が話題に**
8. **[クラウド] Office 365が上位独占、クラウド女子にも熱視線**
9. **[IoT] 日立Lumadaの売上高は9000億円、トイレ難民をIoTが救う**
10. [モバイル] iPhone Xが話題独占、曲がり角迎えた格安スマホ
11. [IT経営] AIやスパコン、華やかな新技術に潜む闇
12. **[システム] いよいよ移行本番か、Windows 10のアップデートに注目が集まる**
13. [PC] 圧倒的に読まれた「Excel方眼紙」、超軽量PCも人気
14. **[総合] ユーザーの怒りや不満、ベンダーとの衝突が上位独占**
15. **[セキュリティ] WannaCryやリスト型攻撃の猛威、覆される常識**
16. **[AI／機械学習] ディープラーニングの仕組み、AI脅威論への反論が注目集める**

キーワード抽出

- KHCoderを活用し, 記事から名詞・組織名などを抽出
- 表記揺れを対処 (Aamazon/amazon/アマゾン等)

名詞	名詞	システム	情報	企業	データ	技術	ユーザー	一つ	ネットワーク	製品	内容	画面	事業	同社	理由	環境	記事	基本	...
	名詞B	きっかけ	いくつか	そのもの	めど	あと	それなり	うち	した	つめ	もの	あたり	おかげ	おすすめ	つながり	どこか	あれこれ	がん	...
	名詞C	年	月	人	例	次	他	別	基	手	円	図	気	形	目	割	先	上	...
	サ変名詞	利用	開発	サービス	機能	提供	関連	管理	説明	活用	対応	発表	実現	出所	確認	導入	実施	表示	...
	形容動詞	必要	可能	重要	新た	様々	簡単	同様	主	明らか	安全	非常	大量	便利	有効	容易	確か	高度	...
	固有名詞	オブ	ソ	リ	クック	トー	マック	創	日本国	さいたま	アンド	エン	コト	サイエンス	ソト	ターミネーター	タメ	ドラえもん	...
	組織名	自社	日経	富士通	日立製作所	ソフトバンク	トヨタ	テック	ドコモ	ブレ	日立	トヨタ自動車	野村総合研究所	コム	三菱電機	インテル	ゼネラル	デル	...
	人名	ハード	オン	リー	伊藤	伸弘	太郎	辻	楠	クレイグ	浩	松村	鈴木	アグリ	ケイ	サー	シニア	テン	...
	地名	日本	米	米国	東京	ジャパン	韓国	日	中国	サンフランシスコ	英	ポー	欧州	台湾	アマゾン	シリコンバレー	ドイツ	英国	...
動詞	動詞	使う	見る	持つ	考える	増える	思う	変わる	異なる	分かる	呼ぶ	言う	作る	受ける	感じる	進める	知る	得る	...
	動詞B	する	ある	できる	なる	いう	いる	よる	しれる	いく	つながる	かかる	みる	かける	いえる	でる	にる	まとめる	...
形容詞	形容詞	多い	高い	大きい	少ない	難しい	新しい	強い	良い	低い	近い	小さい	悪い	正しい	無い	広い	安い	長い	...
	形容詞B	ない	よい	うまい	いい	すごい	うれしい	おかしい	にくい	やすい	ややこしい	ありがたい	かつこいい	からい	くらい	たまらない	つばい	もったいない	...
	ナイ形容	問題	違い	間違い	とんでも	仕方	容赦	致し方	頼り										...
	形容詞(非自立)	やすい	よい	にくい	ほしい	いい	づらい	良い	がたい	難い	欲しい								...
助詞等	副詞B	さらに	どう	もう	まず	そう	まだ	なぜ	こう	あまり	ほとんど	すぐ	ほぼ	かなり	やはり	より	もちろん	もっと	...
	副詞	実際	既に	特に	最も	同時に	初めて	全く	比較的	少し	ある程度	改めて	当然	必ず	本当に	次々	続々	同じく	...
	副詞可能	場合	結果	ほか	多く	今回	時間	一部	現在	前	今	今後	全体	それぞれ	最近	従来	全て	当初	...
	感動詞	と	ありがとう	あれよあれよ	え	お	そうですね	なるほど	ねえ	はい	まあ	キャッ	フェ						...
	否定助動詞	ない	ぬ	ん	まい														...
その他	未知語	IT	Web	アプリ	AI	ベンダー	PC	Windows	NTT	OS	Google	グーグル	Amazon	AWS	iPhone	Android	ITpro	for	...
	タグ	クラウド	サイト	サーバー	IoT	スマートフォ	マイクロソフト	スマホ	ロボット	サイズ	インチ	サイバー	マルウェア	アイコン	サムスン	マルチ	ランサム	利便性	...

グラフ構築

2つの手法を試行的に適用した。

- **スパースグラフ**： 相関の高いワード間にエッジを張る
- **共起グラフ**： 記事内に共起したワードにエッジを張る

スパースグラフ

	news1	news2	news3	news4	news5	news6	news7	news8	
データ	23	5	3	10	0	0	1	3	...
技術	2	18	0	8	1	1	3	0	...
音声	0	0	3	3	1	1	0	0	...
システム	14	11	0	6	0	0	0	8	...
機械	16	0	0	15	0	0	0	1	...
人間	0	4	4	8	0	0	0	5	...
...	...	...	...	...	...	...	...	...	...

ドキュメント内での出現回数でベクトル化

ベクトル間の相関
の高さを計算

グラフ構築

共起グラフ

同時に出現したドキュメント数をカウント

	news1	news2	news3	news4	news5	news6	news7	news8	
データ	23	5	3	10	0	0	1	3	...
技術	2	18	0	8	1	1	3	0	...
音声	0	0	3	3	1	1	0	0	...
システム	14	11	0	6	0	0	0	8	...
機械	16	0	0	15	0	0	0	1	...
人間	0	4	4	8	0	0	0	5	...
...	...	...	...	...	...	...	...	...	...

エッジ

エッジ

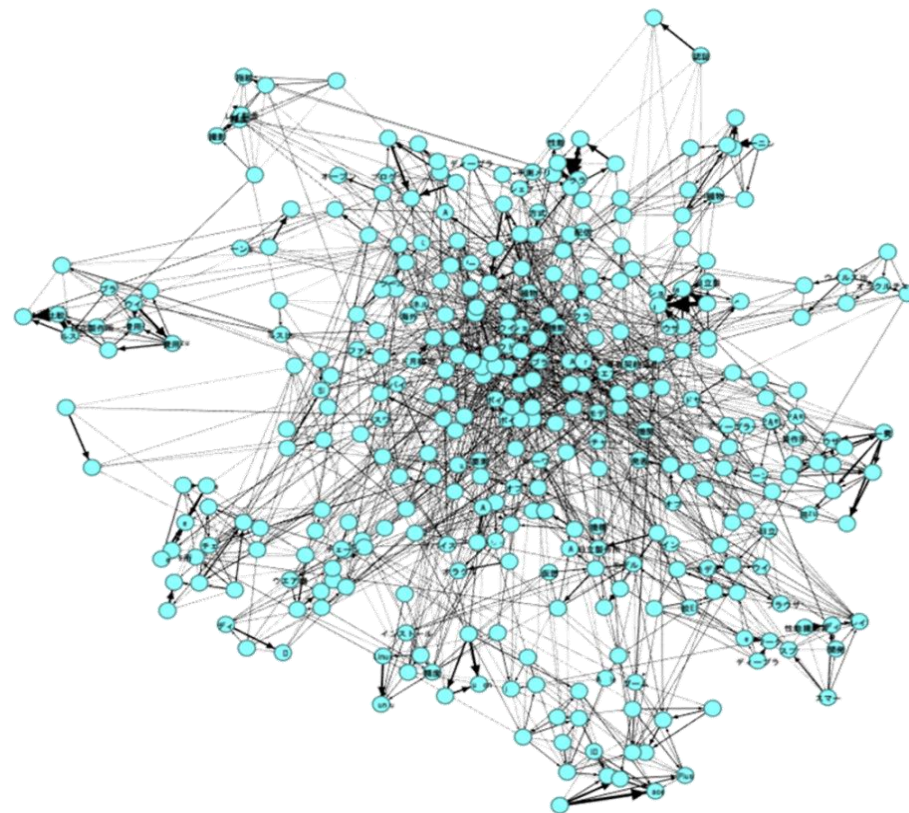
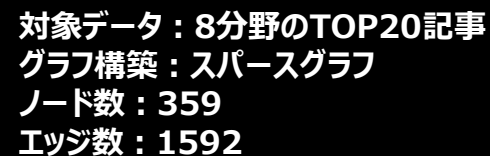
エッジ

ドキュメント内での出現回数

共起の回数
を計算

グラフ構築

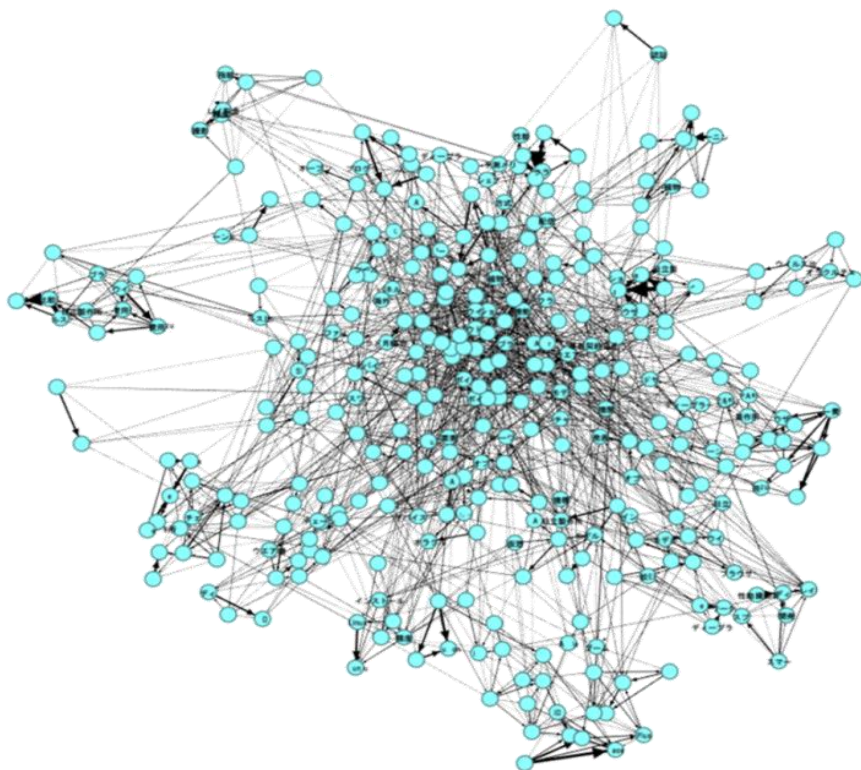
対象データ：[AI／機械学習]のみの20記事
 グラフ構築：スパースグラフ
 ノード数：164
 エッジ数：284



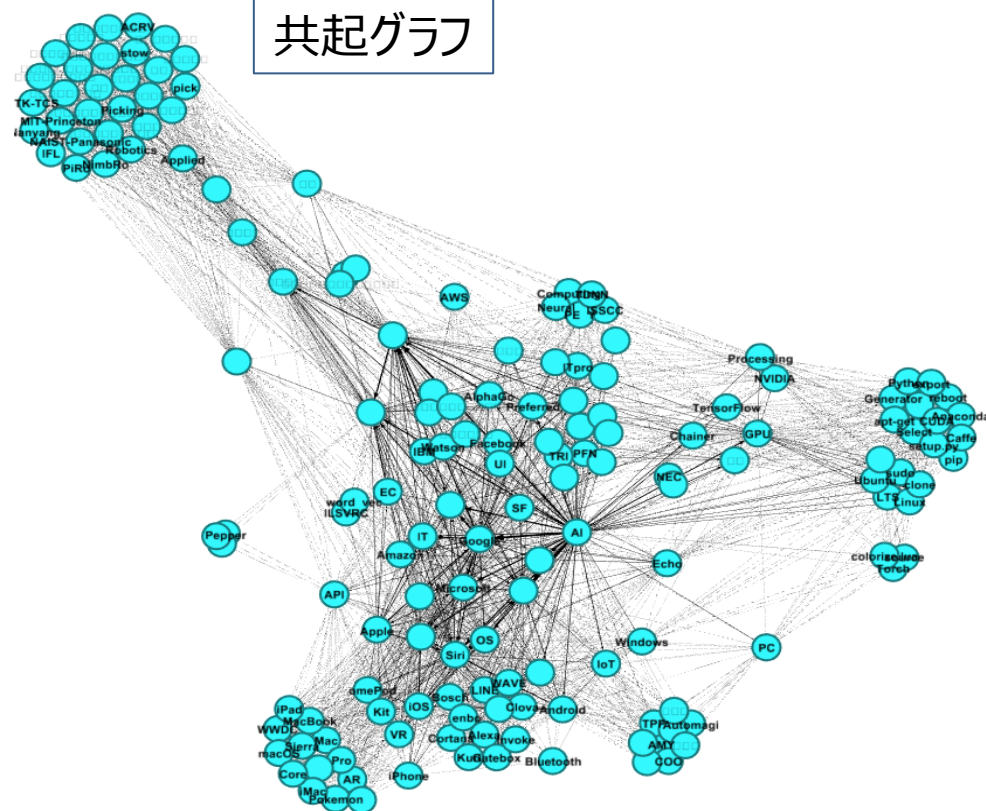
可視化例 続き

スパースグラフ（相関の高いワード間にエッジ）に比して、共起グラフ（記事内に共起したワードにエッジ）のほうが自然な構造を示したため、主に共起グラフで検証した。

スパースグラフ



共起グラフ



注目ワード抽出

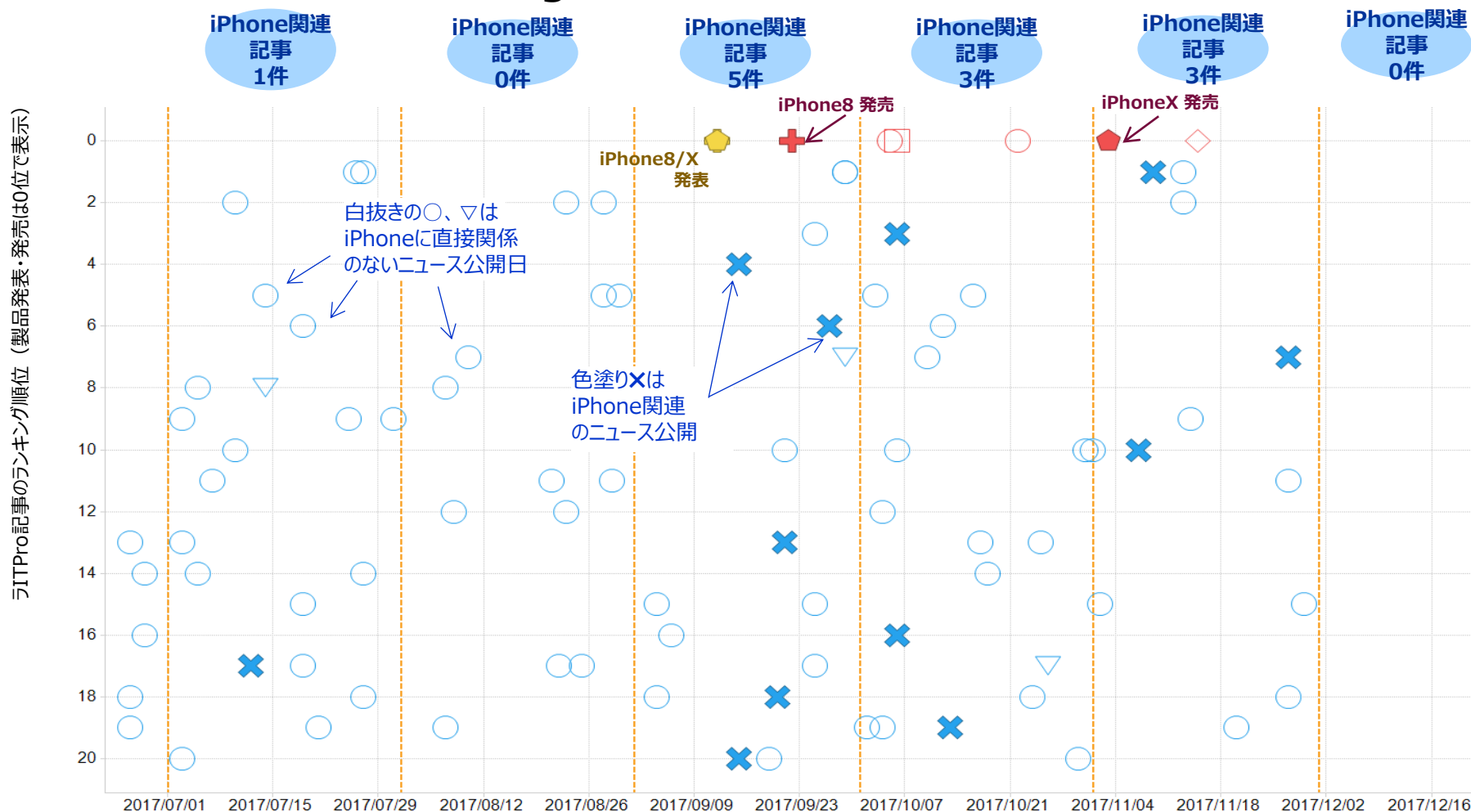
1年分の全体でPageRankを計算した結果、およそ妥当な結果が得られている。

<データセット1, 共起グラフ>

Id	Label	PageRank	Id	Label	PageRank	Id	Label	PageRank
1	AI	0.019884	49	Kit	0.006724	149	ユーシー	0.005835
152	ロボット	0.014375	77	Processing	0.006707	20	Clova	0.005832
135	ディープラーニング	0.012421	64	NVIDIA	0.006706	98	WAVE	0.005832
33	Google	0.012408	139	ニューラルネット	0.006694	125	スマートスピーカー	0.005831
85	Siri	0.010298	3	AlphaGo	0.006691	51	LINE	0.005831
126	スマートフォン	0.010054	90	TensorFlow	0.006569	0	ACRV	0.005826
115	クラウド	0.009736	53	LTS	0.006454	110	カーlsruエ	0.005822
65	OS	0.009343	94	Ubuntu	0.006454	116	コム	0.005822
57	Microsoft	0.009194	19	clone	0.006454	80	Robotics	0.005822
140	ニューラルネット ワーク	0.009137	52	Linux	0.006454	37	IFL	0.005822
158	東京大学	0.00854	88	sudo	0.006454	87	stow	0.005822
34	GPU	0.008389	136	ディストリビューション	0.006452	117	コンサルタンシ	0.005822
10	Apple	0.008114	141	パナソニック	0.006433	127	スマートロボティクス	0.005822
119	サムスン	0.008027	9	API	0.006356	112	カーンブル	0.005822
138	トヨタ	0.00788	43	IoT	0.006346	114	クイーンズランド	0.005822
75	Preferred	0.007865	99	Windows	0.006165	71	Picking	0.005822
28	Echo	0.007677	83	SF	0.006081	123	ジョーイ	0.005822
47	IT	0.007665	106	エヌビディア	0.006069	132	ダラム	0.005822
42	iOS	0.007446	48	ITpro	0.006069	70	pick	0.005822
153	ロボティクス	0.007349	66	PC	0.006069	122	シドニー大学	0.005822
8	Android	0.007288	163	富士通	0.006044	131	タタ	0.005822
118	サーバー	0.00727	97	Watson	0.00598	73	PiRo	0.005822
161	日経	0.007209	36	IBM	0.00598	63	NimbRo	0.005822
18	Chainer	0.007192	45	iPhone	0.00591	133	チーフテクノロジー	0.005822
159	東芝	0.007125	160	奈良先端科学技 術大学院大学	0.005835	38	IITK-TCS	0.005822
4	Amazon	0.007015	143	プリンストン大学	0.005835	58	MIT-Princeton	0.005822
156	三菱電機	0.006984	147	マサチューセッツ工科大 学	0.005835	59	NAIST-Panasonic	0.005822
96	VR	0.006726	150	ライトナー	0.005835	60	Nanyang	0.005822
11	Applied	0.006725	144	ヘルムートシュミット	0.005835	137	デンソー	0.0058
						103	アーキテクチャー	0.005497

注目ワード抽出

iPhone8に関する記事に着目して、PageRankの影響をしてみる。



対象期間① 2017/5~8 67件

対象期間② 2017/6~9 60件

対象期間③ 2017/7~10 64件

対象期間④ 2017/8~11 54件

注目ワード抽出

iPhoneに関する記事の増加とともに、関連ワードのPageRankも増大する。

【対象期間①】

Rank	Personalized PageRank	Label
1	0.02153	IT
2	0.01802	クラウド
3	0.01646	MS ▼
4	0.01521	AI
5	0.01417	Google
6	0.01355	マイクロソフト ▼
7	0.01315	ベンダー
8	0.01204	スマートフォン
9	0.01096	Windows
10	0.01084	インターネット・オブ・シングス ▼
11	0.01084	IoT ▼
12	0.01023	サーバー
13	0.01009	日経
14	0.01008	AWS
15	0.00992	ロボット
16	0.00992	NTT
17	0.00895	Amazon.com
18	0.00812	API
19	0.00809	NRI
20	0.00792	Azure
21	0.00771	トヨタ
22	0.00769	NTT コミュニケーションズ
23	0.00714	IPA
24	0.00708	iOS
25	0.00699	IP
26	0.00689	Apple
27	0.00681	KDDI
28	0.00681	Linux
29	0.00677	ORACLE
30	0.00672	ドコモ
31	0.00654	SNS
32	0.00645	IBM
33	0.00639	マルウェア
34	0.00638	ブログ
35	0.00634	サムスン
36	0.00626	BP
37	0.00626	ソフトバンク
38	0.00622	Android
39	0.00604	iPhone
40	0.00602	FinTech

【対象期間②】 iPhone8/X発表
・iPhone8発売

Rank	Personalized PageRank	Label
1	0.02323	IT
2	0.01882	クラウド
3	0.01466	AI
4	0.01463	MS
5	0.01412	Google
6	0.01381	スマートフォン
7	0.01376	ベンダー
8	0.01126	Apple
9	0.01096	AWS
10	0.01091	ロボット
11	0.01051	iPhone
12	0.01041	Amazon.com
13	0.00987	サーバー
14	0.00985	Windows
15	0.00972	インターネット・オブ・シングス
16	0.00951	Android
17	0.00947	NRI
18	0.00920	日経
19	0.00920	API
20	0.00901	iOS
21	0.00898	NTT
22	0.00863	Touch
23	0.00771	Siri
24	0.00770	サムスン
25	0.00768	ORACLE
26	0.00766	マルウェア
27	0.00763	トヨタ
28	0.00759	GCP
29	0.00756	NTT コミュニケーションズ
30	0.00754	Linux
31	0.00746	富士通
32	0.00737	IBM
33	0.00727	野村総合研究所
34	0.00720	マルチ
35	0.00715	IPA
36	0.00710	スワイプ
37	0.00710	BP
38	0.00710	Galaxy
39	0.00705	三菱
40	0.00705	UFJ

【対象期間③】

Rank	Personalized PageRank	Label
1	0.01938	IT
2	0.01450	スマートフォン
3	0.01450	クラウド
4	0.01247	Google
5	0.01204	AI
6	0.01160	MS
7	0.00934	ベンダー
8	0.00880	Windows
9	0.00825	iPhone
10	0.00792	NRI
11	0.00775	Amazon.com
12	0.00753	AWS
13	0.00752	インターネット・オブ・シングス
14	0.00752	ロボット
15	0.00748	アンドロイド
16	0.00748	Android
17	0.00748	API
18	0.00737	Apple
19	0.00694	サーバー
20	0.00692	NTT
21	0.00659	トヨタ
22	0.00633	Touch
23	0.00630	日経
24	0.00629	iOS
25	0.00622	IBM
26	0.00621	スワイプ
27	0.00610	Linux
28	0.00603	ORACLE
29	0.00602	ブログ
30	0.00598	SNS
31	0.00597	IPA
32	0.00588	Creators
33	0.00588	Fall
34	0.00582	アイコン
35	0.00575	マルウェア
36	0.00570	GUI
37	0.00562	セキュアテクノロジーズ
38	0.00548	ISP
39	0.00547	EL
40	0.00543	サムスン

【対象期間④】 iPhoneX発売

Rank	Personalized PageRank	Label
1	0.01818	スマートフォン
2	0.01746	IT
3	0.01571	クラウド
4	0.01351	Google
5	0.01243	iPhone
6	0.01202	AI
7	0.01011	Windows
8	0.01007	MS
9	0.00982	EL
10	0.00969	ベンダー
11	0.00906	AWS
12	0.00900	NTT
13	0.00899	Face
14	0.00898	Touch
15	0.00896	Apple
16	0.00877	スワイプ
17	0.00865	サーバー
18	0.00838	インターネット・オブ・シングス
19	0.00833	SNS
20	0.00801	Twitter
21	0.00799	IPA
22	0.00798	IBM
23	0.00790	Amazon.com
24	0.00787	Android
25	0.00787	アンドロイド
26	0.00775	NRI
27	0.00763	API
28	0.00755	日経
29	0.00754	Linux
30	0.00753	iOS
31	0.00739	アイコン
32	0.00738	ISP
33	0.00691	Edge
34	0.00690	ORACLE
35	0.00679	Creators
36	0.00679	Fall
37	0.00678	NTT コミュニケーションズ
38	0.00675	ブログ
39	0.00669	Galaxy
40	0.00669	利便性

注目ワード抽出

全体的にPageRankの高いワードに引きずられやすい。

【トヨタ+Google】

Rank	PPR	Label
1	0.311192	Google
2	0.304952	トヨタ
3	0.017643	Microsoft
4	0.017148	IT
5	0.015748	AI
6	0.012596	マイクロソフト
7	0.009403	日経
8	0.009221	Amazon
9	0.009215	Windows
10	0.007941	AWS
11	0.007845	NTT
12	0.007549	NRI
13	0.007306	ロボット
14	0.007159	クラウド
15	0.00653	TRI
16	0.00653	インスティテュート
17	0.006324	Azure
18	0.006238	IoT
19	0.005453	IBM
20	0.005295	ベンダー
21	0.005171	Apple
22	0.005034	日産
23	0.004569	iOS
24	0.004469	ドコモ
25	0.004382	GCP
26	0.004239	API
27	0.004165	ブログ
28	0.003996	ORACLE
29	0.003949	MUSIC
30	0.0038	サーバー
31	0.003795	Android
32	0.003575	IaaS
33	0.003574	HomePod
34	0.003574	Siri
35	0.003527	Macintosh
36	0.003527	Lambda
37	0.003248	Linux
38	0.003203	ニューラルネットワーク
39	0.003179	レイヤー
40	0.002995	富士通
41	0.002875	セキュアテクノロジーズ
42	0.002875	IETF
43	0.002831	NEC
44	0.002825	MIT
45	0.002758	ディープラーニング
46	0.002715	LINE
47	0.002672	CTC
48	0.002592	ロボティクス
49	0.002558	BP
50	0.002549	Twitter

【Google+AWS】

Rank	PPR	Label
1	0.310062	Google
2	0.307635	AWS
3	0.018231	Microsoft
4	0.018124	IT
5	0.013075	マイクロソフト
6	0.012058	Amazon
7	0.010797	クラウド
8	0.010609	Windows
9	0.010124	AI
10	0.009684	NTT
11	0.00912	Azure
12	0.008764	NRI
13	0.007845	ベンダー
14	0.006611	ロボット
15	0.006508	API
16	0.006277	サーバー
17	0.005537	GCP
18	0.005196	ORACLE
19	0.004885	三菱
20	0.004885	UFJ
21	0.00481	日経
22	0.004711	ドコモ
23	0.004643	トヨタ
24	0.004607	IaaS
25	0.004573	野村総合研究所
26	0.004352	Apple
27	0.004101	IBM
28	0.003848	三井
29	0.003846	Macintosh
30	0.003846	Lambda
31	0.00375	ソリューションアーキテクト
32	0.003733	JAWS-UG
33	0.003694	ビジナリー
34	0.003533	ブログ
35	0.003404	Linux
36	0.003357	Group
37	0.003357	Japan
38	0.003141	Java
39	0.003075	Android
40	0.002956	iOS
41	0.002924	Suite
42	0.002879	ASIMO
43	0.002811	日産
44	0.002802	XML
45	0.002709	IoT
46	0.002684	BP
47	0.002443	SNS
48	0.002403	レイヤー
49	0.002382	MUSIC
50	0.002381	VMware

【トヨタ+SAP】

※赤字はPPRにおけるシード

Rank	PPR	Label
1	0.301583350	トヨタ
2	0.301033258	SAP
3	0.025306644	IT
4	0.017300745	NTT
5	0.013390614	クラウド
6	0.012003399	IoT
7	0.010467812	IBM
8	0.009759893	アプリ
9	0.009499968	AI
10	0.008750618	オラクル
11	0.008111504	アクセシビリティ
12	0.008047689	ドコモ
13	0.007683809	API
14	0.00766169	コム
15	0.007607354	ICT
16	0.00741826	ERP
17	0.007177922	旭川医科大学
18	0.007177916	スルガ銀行
19	0.006961199	キーパー
20	0.006655585	ファナック
21	0.006305102	Web
22	0.006068079	Microsoft
23	0.00517172	日経
24	0.005160125	Google
25	0.004924572	サーバー
26	0.004767707	Amazon
27	0.003998442	PC
28	0.003917587	富士通
29	0.003422199	NEC
30	0.003366988	Windows
31	0.002996751	スマートフォン
32	0.002969217	AWS
33	0.002830484	ITpro
34	0.002798993	NRI
35	0.002748261	TRI
36	0.002630678	OS
37	0.002444703	OSS
38	0.002253071	日立製作所
39	0.002218992	デック
40	0.002215136	Apple
41	0.002077638	Android
42	0.002056028	Cloud
43	0.002028187	日産
44	0.001941768	Echo
45	0.001934968	スマートスピーカー
46	0.001854962	iOS
47	0.001834036	Alexa
48	0.001797046	ディープラーニング
49	0.001781985	日立
50	0.001775314	Azure



技術トレンド抽出

4つのクラスタ(トレンド)を抽出。今後、各クラスタの特徴をうまく表現する方法の検討が必要。

スマートフォン・音声クラスタ

Id	Label	PageRank
85	Siri	0.0103
126	スマートフォン	0.0101
115	クラウド	0.0097
65	OS	0.0093
57	Microsoft	0.0092
10	Apple	0.0081
119	サムスン	0.0080
42	iOS	0.0074
153	ロボティクス	0.0073
8	Android	0.0073
118	サーバー	0.0073
96	VR	0.0067
35	HomePod	0.0067
49	Kit	0.0067
9	API	0.0064
43	IoT	0.0063
99	Windows	0.0062
66	PC	0.0061
45	iPhone	0.0059
20	Clova	0.0058
98	WAVE	0.0058
125	スマートスピーカー	0.0058
51	LINE	0.0058
2	Alexa	0
6	AMY	0
13	AR	0
14	Automagi	0
15	Bluetooth	0
16	Bosch	0
21	colorize.lua	0
23	COO	0
24	Core	0

ディープラーニング・企業クラスタ

Id	Label	PageRank
152	ロボット	0.0144
135	ディープラーニング	0.0124
33	Google	0.0124
140	ニューラルネットワーク	0.0091
138	トヨタ	0.0079
75	Preferred	0.0079
47	IT	0.0077
161	日経	0.0072
4	Amazon	0.0070
139	ニューラルネット	0.0067
3	AlphaGo	0.0067
83	SF	0.0061
106	エヌビディア	0.0061
48	ITpro	0.0061
163	富士通	0.0060
97	Watson	0.0060
36	IBM	0.0060
137	デンソー	0.0058
103	アーキテクチャー	0.0055
95	UI	0.0055
5	AWS	0
22	Computing	0
27	EC	0
30	Facebook	0
46	ISSCC	0
61	NEC	0
62	Neural	0
68	PEZY	0
69	PFN	0
89	TDNN	0
93	TRI	0
104	アティフ	0
105	インスティテュート	0
109	ガートナー	0
113	ギル	0
124	ステマ	0
128	セーフティ	0
134	チャットボット	0
145	ベンダー	0
151	レイヤー	0
154	ワークスアプリケーションズ	0
162	富国生命保険	0
39	ILSVRC	0
100	word2vec	0

AI基盤クラスタ

Id	Label	PageRank
1	AI	0.0199
34	GPU	0.0084
28	Echo	0.0077
18	Chainer	0.0072
77	Processing	0.0067
64	NVIDIA	0.0067
90	TensorFlow	0.0066
53	LTS	0.0065
94	Ubuntu	0.0065
19	clone	0.0065
52	Linux	0.0065
88	sudo	0.0065
136	ディストリビューション	0.0065
72	pip	0.0054
79	reboot	0.0054
82	setup.py	0.0054
7	Anaconda	0.0054
12	apt-get	0.0054
17	Caffe	0.0054
29	export	0.0054
26	CUDA	0.0054
81	Select	0.0054
32	Generator	0.0054
78	Python	0
142	ブラウザー	0

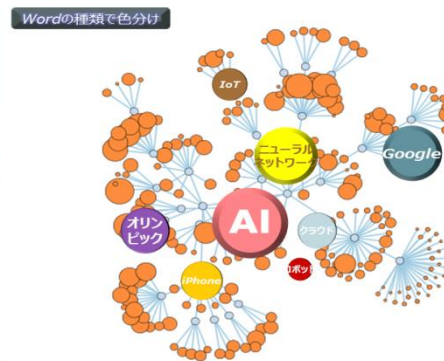
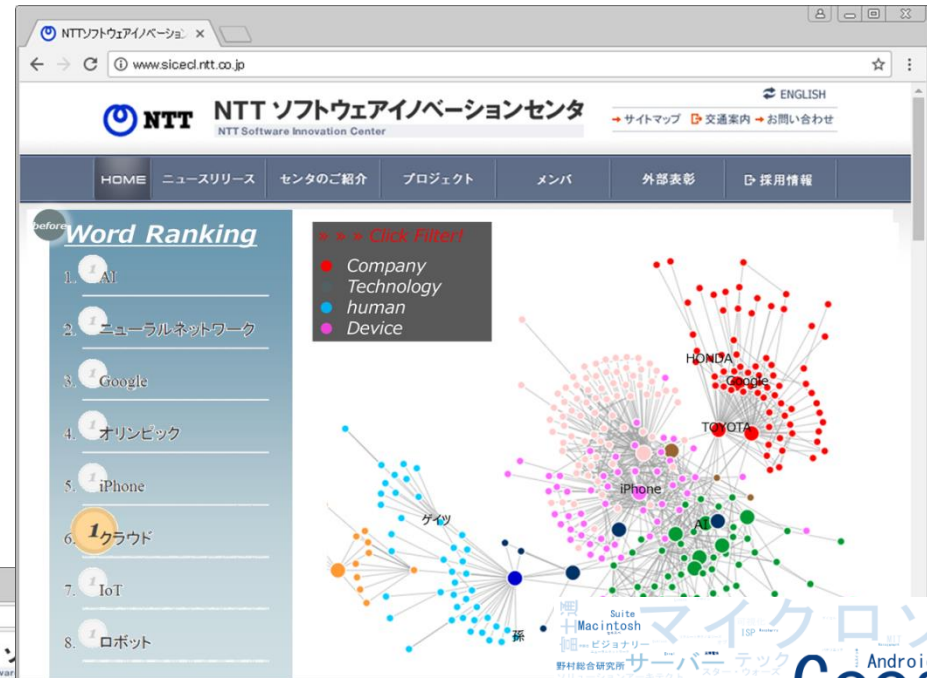
大学・企業クラスタ

Id	Label	PageRank
158	東京大学	0.0085
159	東芝	0.0071
156	三菱電機	0.0070
11	Applied	0.0067
141	パナソニック	0.0064
160	奈良先端科学技術大学院大学	0.0058
143	プリンストン大学	0.0058
144	ヘルムートシュミット	0.0058
147	マサチューセッツ工科大学	0.0058
150	ライトナー	0.0058
149	ユニーシー	0.0058
0	ACRV	0.0058
110	カールスルーエ	0.0058
116	コム	0.0058
80	Robotics	0.0058
117	コンサルタンシ	0.0058
127	スマートロボティクス	0.0058
37	IFL	0.0058
87	stow	0.0058
112	カーンブル	0.0058
114	クイーンズランド	0.0058
123	ジョーイ	0.0058
71	Picking	0.0058
132	ダラム	0.0058
131	タタ	0.0058
122	シドニー大学	0.0058
70	pick	0.0058
63	NimbRo	0.0058
73	PIRo	0.0058
59	NAIST-Panasonic	0.0058
133	チーフテクノロジスト	0.0058
38	IITK-TCS	0.0058
58	MIT-Princeton	0.0058
60	Nanyang	0.0058
107	オントロジ	0
108	オントロジスト	0
164	北陸先端科学技術大学院大学	0

可視化イメージ

トレンドを分かり易く表現するグラフ可視化方法が課題。

※いずれもイメージです



事例 2：関連論文の発見

- 電子情報通信学会の論文データ（I-Scover）から論文発見する方法を提案

2017年のI-Scover利活用コンテストへの応募作品

I-Scoverシステムで、キーワード検索で出てきた論文のみならず、それらの 関連論文まで同時に拾えないか？

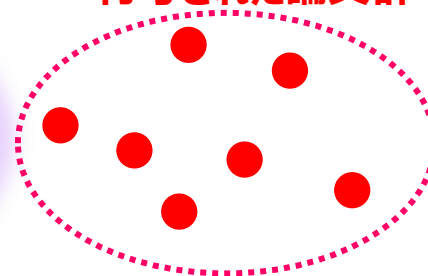
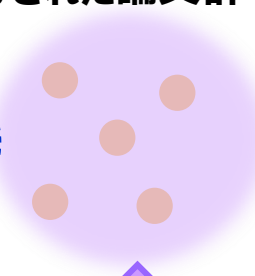
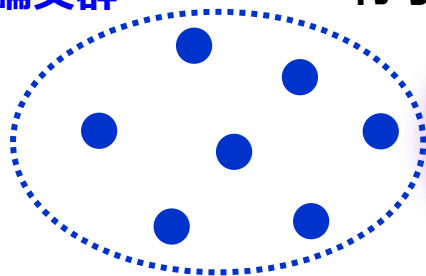
技術名が変遷してしまうと、キーワード検索では変遷の狭間にある論文を拾えない

「ユビキタス網」と「IoT」の2つでキーワード検索しても、現状は「M2M」の論文までは拾えない

ユビキタス網に関するキーワード
が付与された論文群

M2Mのキーワードが
付与された論文群

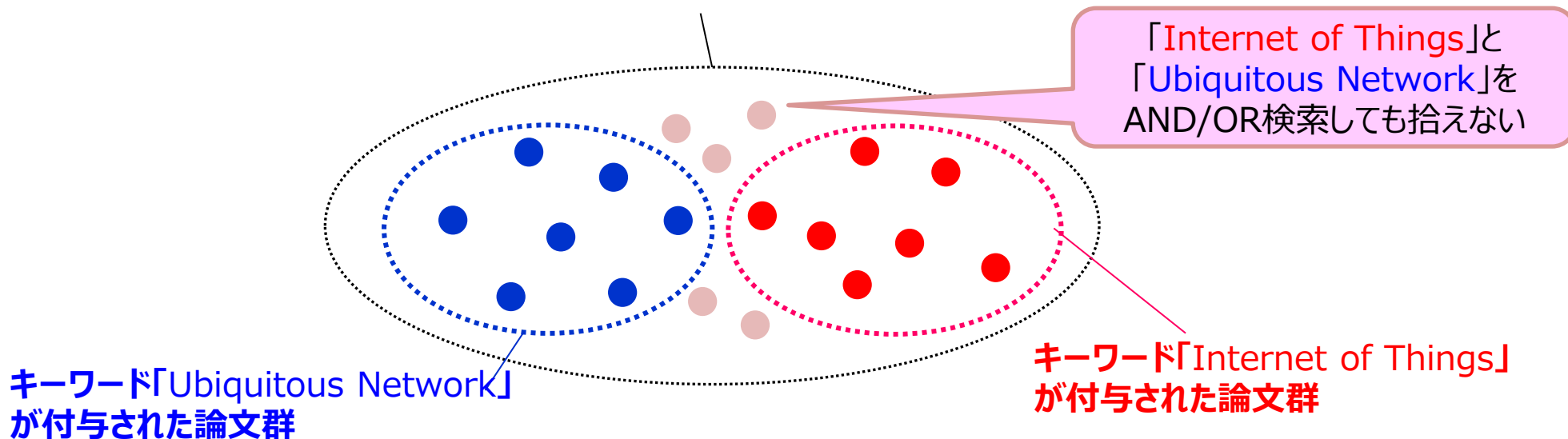
IoTに関するキーワードが
付与された論文群



ユビキタス網とIoTの2つで検索すれば
ここにある**関連論文**まで拾えないか？

キーワードを複数指定すれば，それらの**技術領域の狭間にある関連論文**まで取得できる手法を提案する

取得したい論文（ユビキタス網とIoTの**関連論文**）



凡例

- 「Internet of Things」がキーワードに付与された論文
- 「Ubiquitous Network」がキーワードに付与された論文
- 「M2M」，「センシング網」など，その他のIoT関連のキーワードが付与された論文

関連論文を発見する手順

1. 論文の類似性グラフ構築
2. 論文分類のための技術カテゴリ構築
3. 検索キーワードを技術カテゴリにマッピング
4. ラベル伝播で関連論文を抽出

技術カテゴリ
IoT
ユビキタス網
センサ網
センサ
ユビキタス計算
RFID

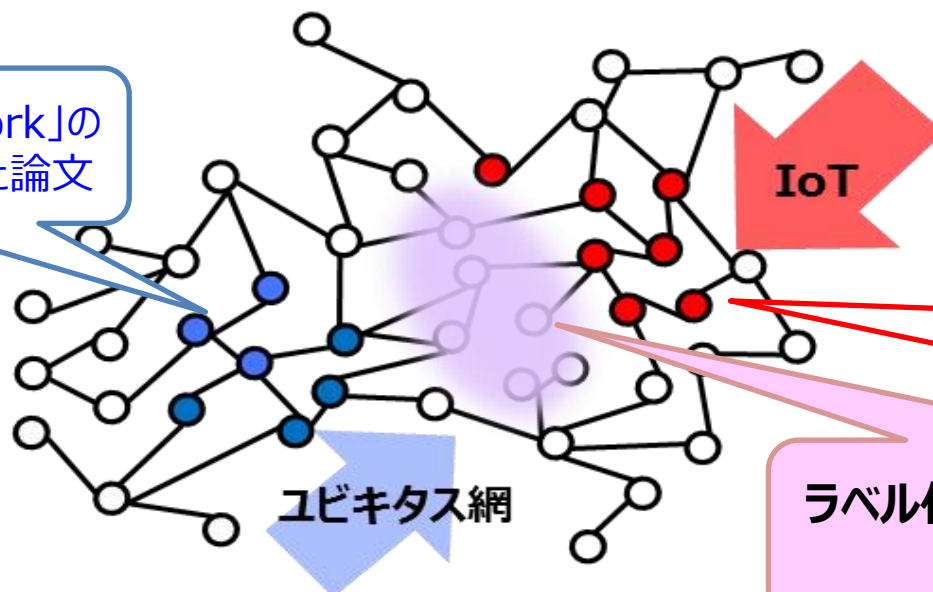
検索キーワード

Internet of Things

Ubiquitous Network

論文の類似性グラフ

「Ubiquitous Network」の
キーワードが付与された論文



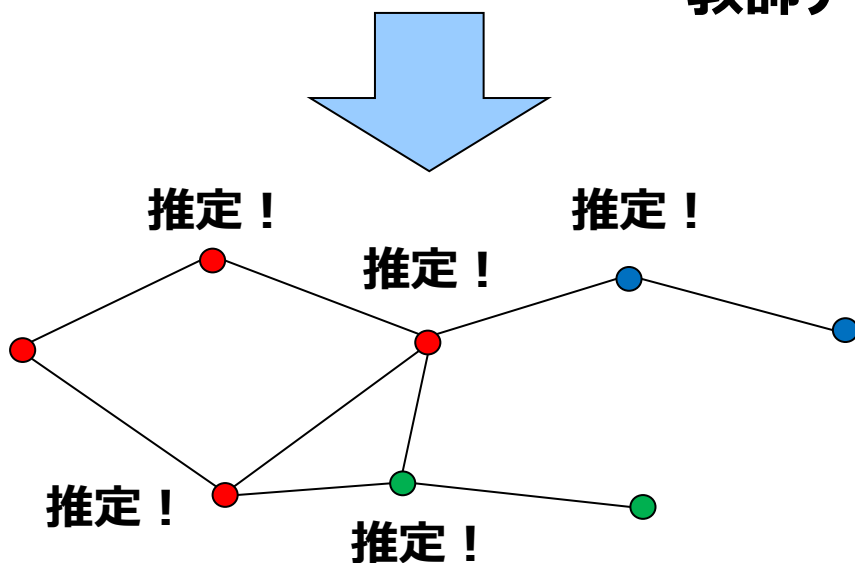
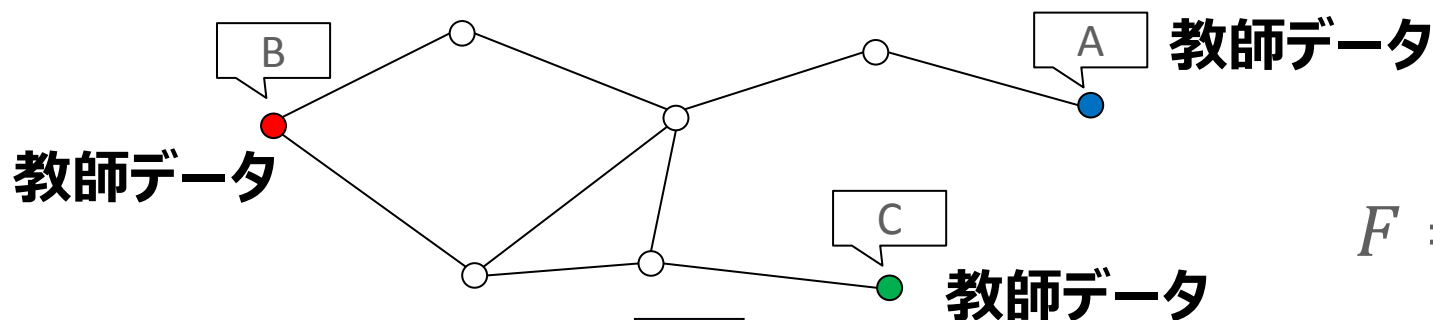
「Internet of Things」の
キーワードが付与された論文

ラベル伝播でIoTとユビキタス網の狭間
にある関連論文を抽出！

前提知識：ラベル伝播とは？

グラフ(ノード, エッジ, ラベル)から, **教師データのラベル**をもとに未知のノードのラベルを推定する半教師ありグラフベース機械学習

⇒ ノードを複数の**カテゴリに分類**する技術



$$F = (I - \alpha S)^{-1} Y$$

F : ラベルのスコア行列 ($n \times c$)

I : 単位行列

α : 定数

S : グラフ構造を表した行列

Y : 教師データのラベル行列

n : ノード数

c : カテゴリ数

1. 論文の類似性グラフ構築

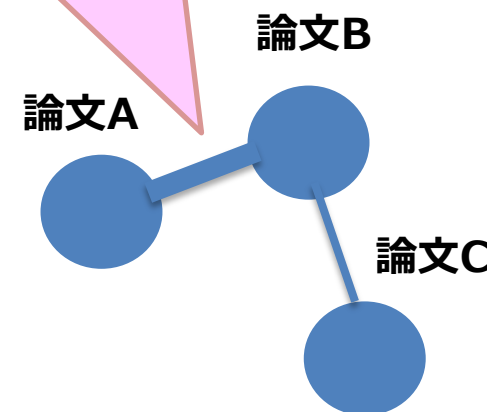
- 要約にある頻出用語（単語共起）で論文の特徴ベクトルを作る
- 特徴ベクトル間の類似性の高さ（近さ）に応じて論文間にエッジを張る

約23万の文献から和文あり・要約ありの16万文献に着目.

さらにIoTとその周辺領域の計2667件の論文をSPARQLで抽出し分析対象とした.

要約にある頻出用語（88語）

類似性が高いほど
太いエッジを張る

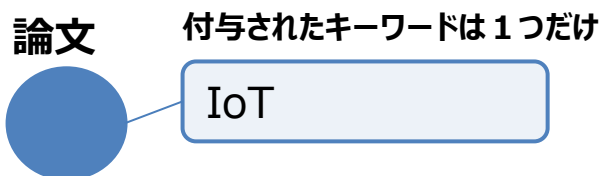


抽出語	出現文書数	抽出語	出現文書数	抽出語	出現文書数	抽出語	出現文書数	抽出語	出現文書数
電力	895	リンク	384	インターネット	265	加速度	153	駆動	129
分散	794	中継	376	帯域	264	広域	149	非常	129
マルチ	755	スループット	375	光	258	要素	148	クラスタ	128
移動	660	クラウド	346	チャンネル	246	コンテンツ	143	ソフトウェア	128
経路	633	空間	338	伝搬	230	パラメータ	141	エネルギー	127
パケット	619	高速	329	インフラ	227	障害	141	周期	127
ホップ	610	自律	321	トラヒック	219	前提	141	コンピュータ	126
アドホック	593	干渉	319	リアルタイム	218	レート	140	シンク	121
消費	527	サーバ	318	波	185	予測	139	素子	120
信号	498	アンテナ	314	タグ	184	セキュリティ	138	回路	116
プロトコル	489	アルゴリズム	313	アーキテクチャ	177	ポイント	132	誤り	116
伝送	486	転送	313	小型	176	レベル	132	変換	116
アクセス	468	災害	296	フレーム	175	一定	132		
遅延	462	電波	296	センサー	169	生活	132		
精度	440	モバイル	295	リソース	163	発見	132		
モデル	433	メッシュ	287	衝突	161	対策	131		
アプリケーション	431	品質	284	自動	157	年	131		
負荷	401	ルーティング	276	監視	156	指標	130		
周波数	400	デバイス	272	分割	155	柔軟	130		

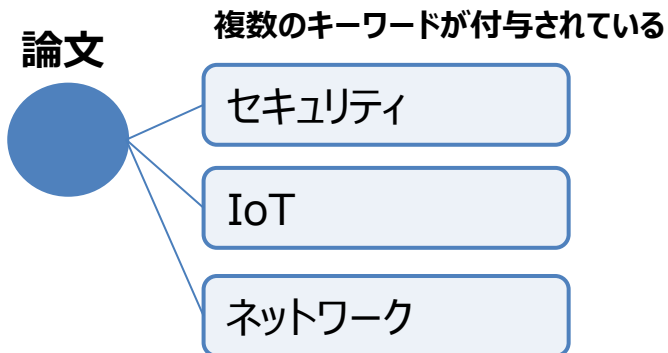
2. 論文分類のための技術カテゴリ構築

キーワードがただ1つだけ設定された論文は，今回の分析における教師データとすべき論文．また，教師データと1：1で対応する技術カテゴリを設定．

教師データに適した例



教師データに適さない例



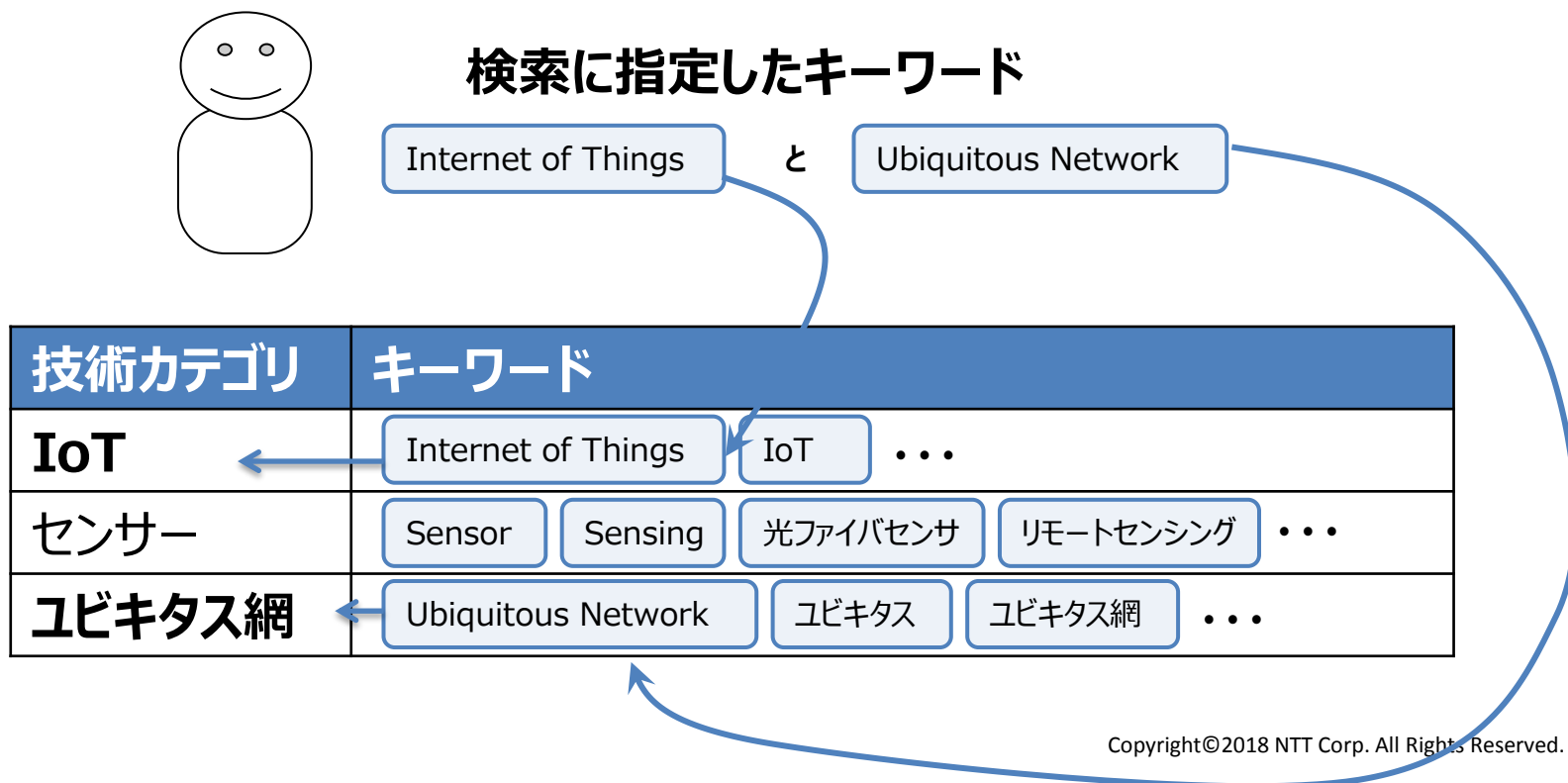
技術カテゴリ	教師データの論文数
IoT	100論文
ユビキタス網	162論文
センサ網	10論文
センサ	160論文
ユビキタス計算	50論文
RFID	236論文

3. 検索キーワードを技術カテゴリにマッピング

指定された検索キーワードの属する技術カテゴリを、以降で詳述する共通化するラベルにする。これで分析の準備は完了。

下図の例では、2つの技術カテゴリを、この後の分析で1つの仮想的な“ラベル”に共通化される

論文検索を行うI-Discoverユーザ

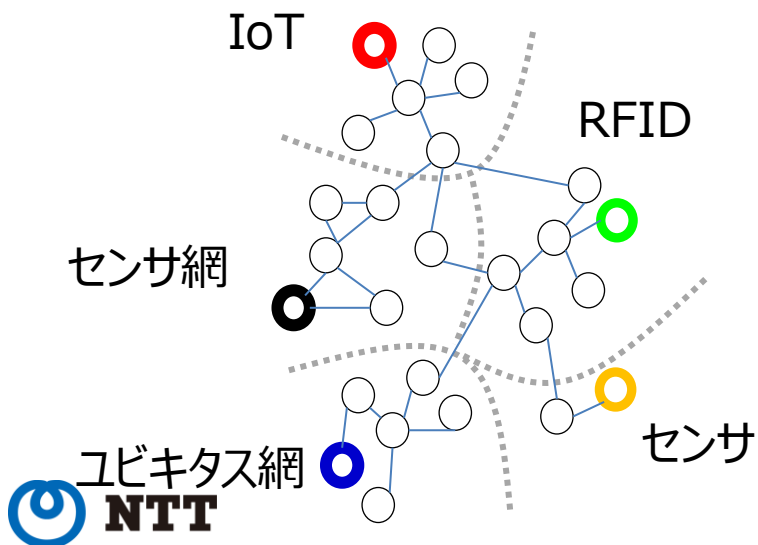


4. ラベル伝播で関連論文を抽出

- IoTとユビキタス網の2つのラベルを, 1つの新たなラベル「IoT×ユビキタス網」に共通化
- この共通化ラベルが伝播した範囲のうち, IoTとユビキタス網の狭間に染み出してきた論文に着目

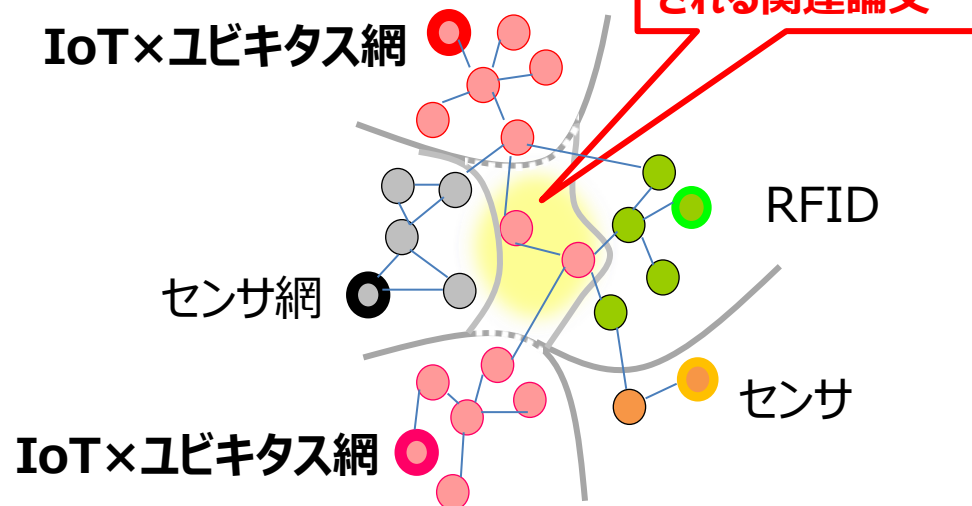
Before

類似性グラフと教師データが存在
(ラベルを共通化する前)



After

ラベルを共通化した後, ラベル伝播を実行



分析結果



本手法により,「IoT」と「ユビキタス網」に関するキーワードを指定した際,同じ技術領域と考えられる「M2M」の論文を発見することができた

No.	title	(参考) “M2M”を含む箇所	ラベル伝播
1	"Study on the diffraction of WPAN in building"@en	Abstract	I×U
2	"仮想グループ化を利用したスリープ制御によるM2Mネットワークの省電力方式"@ja	Keyword, Abstract, Title	I×U
3	"M2M の番号動向"@ja	Keyword, Abstract, Title	I×U
4	"A Study of Accommodation Design for Sensor Devices in User-centric Wireless Networks"@en	Abstract	I×U
5	"センサネットワークとインフラ監視システム"@ja	Abstract	I×U
6	"既存システムと新たなシステムの共存のための信号検出方式の解析的な特性検討 (その2) "@ja	Keyword, Abstract	I×U
7	"短メッセージの大量トラフィック処理を実現するインメモリ分散KVSを適用した高速メッセージングシステム"@ja	Keyword, Abstract	I×U
8	"M2M通信のためのランダムアクセスにおける中継器適用効果に関する一検討"@ja	Keyword, Abstract, Title	I×U
9	"M2M通信のためのランダムアクセスにおける中継器および干渉除去適用効果に関する一検討"@ja	Keyword, Abstract, Title	I×U
10	"NetInfを適用したM2Mネットワークにおける分散経路構築法"@ja	Keyword, Abstract, Title	I×U
11	"M2Mネットワークのための脆弱性診断システムの試作"@ja	Keyword, Abstract, Title	I×U
12	"分散協調M2Mアーキテクチャに基づく洪水防災システムの提案"@ja	Keyword, Abstract, Title	I×U
13	"IEEE1888ネットワークへのZigBeeデバイスのプラグイン・アーキテクチャ"@ja	Keyword	I×U
14	"無線LAN・Bluetooth混在型マルチホップネットワークシステム"@ja	Keyword	I×U
15	"NetInfを適用した複数M2Mサービスを実現するセンサネットワークの実装"@ja	Title	I×U
16	"Innovation of Agriculture by Information Science"@en	Abstract	I×U
17	"900MHz帯センサー無線用低消費電力ダイバシティ受信機の開発"@ja	Abstract	I×U
18	"仮想グループ化を利用したスリープ制御によるIEEE 802.11ahシステムの省電力化"@ja	Keyword, Abstract	I×U
19	"情報指向型ネットワーク方式に基づくセンサネットワークにおけるクラスタリングを用いた電力消費削減法の検討"@ja	Keyword, Abstract	I×U
20	"920MHz帯無線を用いたM2Mゲートウェイの同時接続数削減手法の提案"@ja	Keyword, Abstract, Title	I×U
21	"M2M無線アクセスの間欠動作時における基地局との同期精度向上技術 (I) "@ja	Keyword, Abstract, Title	I×U
22	"NetInfを適用した複数M2Mサービスを実現するセンサネットワークの評価"@ja	Keyword, Abstract, Title	I×U
23	"M2M データの活用法と処理基盤"@ja	Keyword, Abstract, Title	I×U
24	"社会インフラ無線通信の拡大とその取り組み"@ja	Keyword	I×U
25	"センサネットワーク向けインタフェース変換端末"@ja	Keyword	I×U
26	"CSMA/CA方式における干渉量の定量的評価"@ja	Keyword, Abstract	I×U
27	"M2Mエリアネットワークの仮想化に関する検討"@ja	Keyword, Abstract, Title	I×U
28	"ランダムAIFS制御を用いたIEEE 802.11ah準拠M2Mネットワークのスループット特性改善"@ja	Keyword, Abstract, Title	I×U
29	"NetInfによる複数のM2Mサービスのためのセンサネットワーク実現方式"@ja	Keyword, Abstract, Title	I×U
30	"CA-EXI : 運用コンテキストとEXI符号化を用いたIEEE1888メッセージ圧縮手法"@ja	Keyword	I×U
31	"高信頼制御通信研究への招待"@ja	Keyword	I×U
32	"M2M通信に向けた無線LANのポーリング方式の一検討"@ja	Keyword, Title	I×U
33	"数万端末競合環境を実現するM2M 無線アクセスネットワーク"@ja	Keyword, Title	RFID

全33件のM2M論文のうち, 32件を発見!

これら33件のM2M論文は, キーワードに一切,「IoT」,「ユビキタス網」を含んでいないことがポイント

まとめと今後の展望

グラフを用いたことで互いに密につながったまとまりをトレンドとして捉えたり、つながりに隠れた発見を視認性よく行えることを示した。

今後、より分かりやすくグラフをレイアウトする方法や粗視化、増大するグラフデータの効率的な管理方法などが課題。

SQL Server 2017に

- ・ **グラフマイニングはデータ分析ツールとしてだけでなく、検索・コミュニケーション支援、推薦等のシステムに組み込んだ活用も可能。**
- ・ **今後、多くの方と議論しながら、良いユースケースを獲得していく。**

文献分析
センサーデータ分析

医療（診断データ）

状態推定

特許動向

研究トレンド

34



ご清聴ありがとうございました

Advanced Processing and Analytics for Large-scale Graphs
NTT Technical Review, Feb 2016, Vol. 14, No. 2

