研究者と大学のベクトル化とその応用

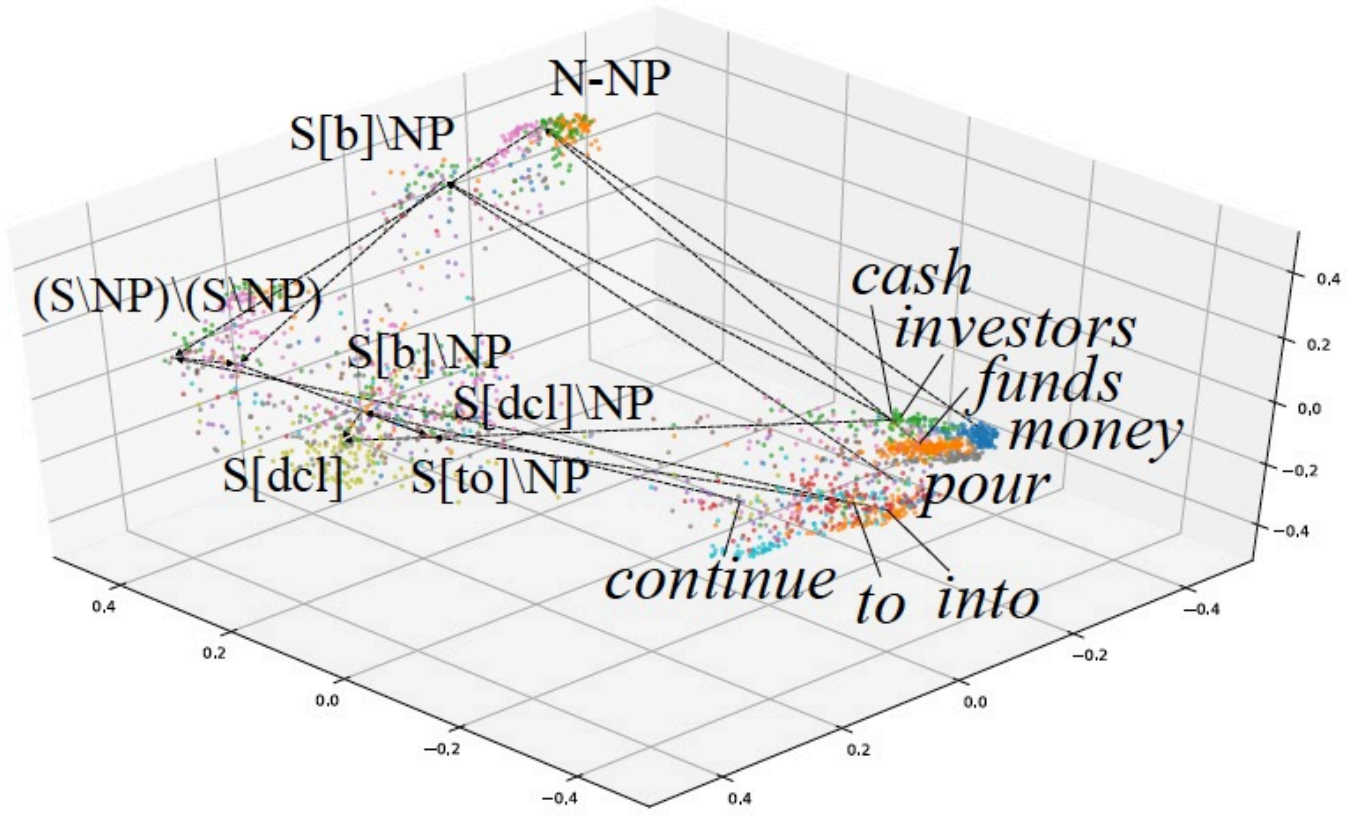
持橋大地
統計数理研究所 統計基盤数理研究系
日本学術振興会 学術情報分析センター
daichi@ism.ac.jp

第1回 Science of Science研究会
2024-3-17 (日)

自己紹介

- 持橋大地
  情報・システム研究機構 統計数理研究所 (立川市)
- 略歴：
  東大教養・基礎二→NAIST(博士)
  →ATR→NTT CS研→統数研
- 専門：自然言語処理、機械学習
  - 自然言語処理: 英語や日本語のような言語をコンピュータで扱う分野、計算言語学ともいう
  - ロボティクス、計量政治学、音楽情報科学、心理統計学など、多数の分野と共同研究しています

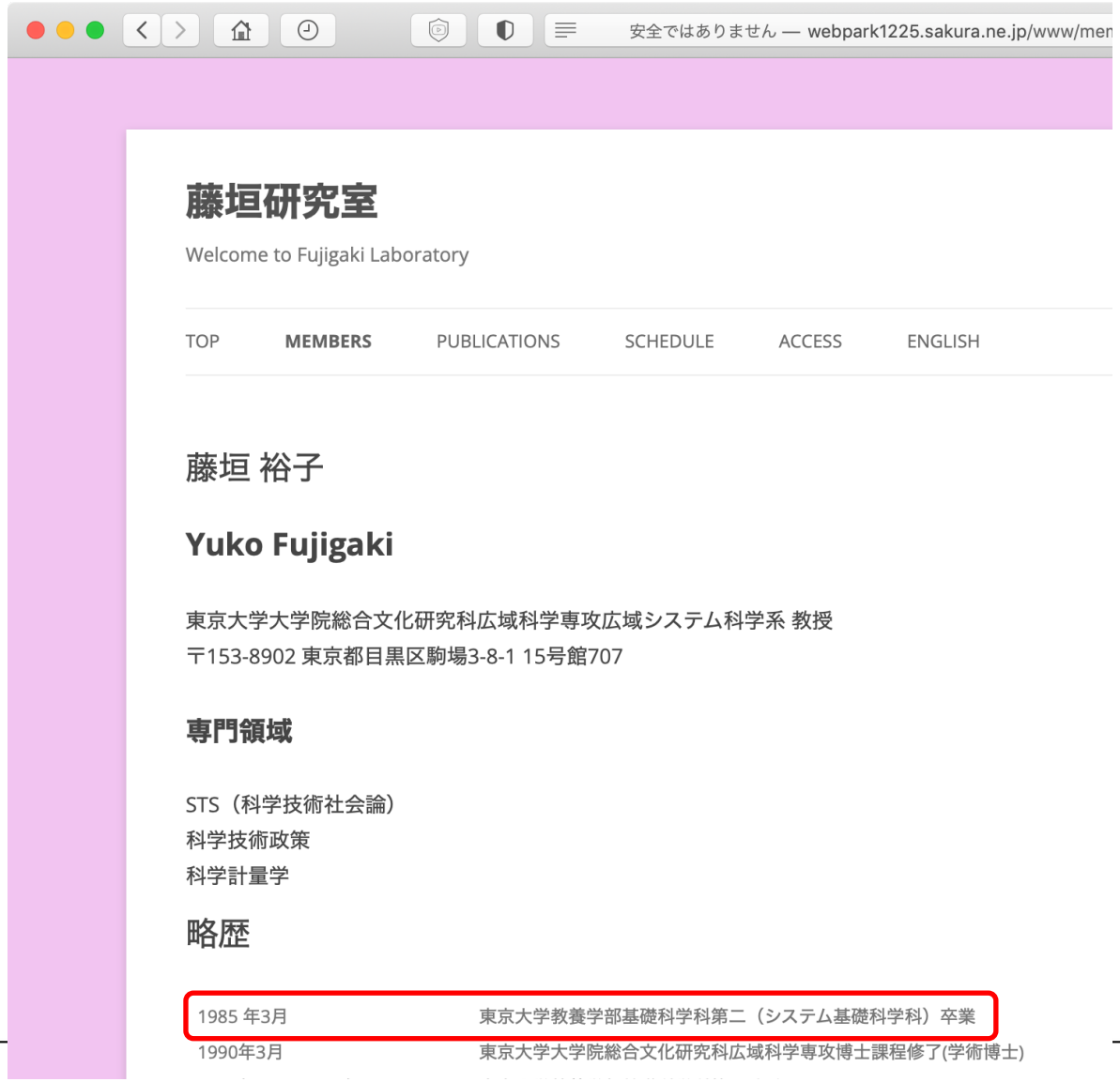
Holographic CCG Parsing (ACL 2023)



- 768次元の埋め込み空間上で、組み合わせ範疇文法(CCG)による構文解析をベクトルの再帰的合成として実現

"Holographic CCG Parsing", Ryosuke Yamaki, Tadahiro Taniguchi, Daichi Mochihashi. ACL 2023 (long), pp.262-276, 2023.

東大教養学部・基礎科学科第二



- 別名: システム基礎科学科
- 現在の広域システム科学系
- 藤垣先生が基礎科学科第二の1期生
- 基礎二の学部教育で、科学技術計画論の存在を叩き込まれた

学振・学術情報分析センター

学振・学術情報分析センター

□ 構成

役職	氏名	所属等
所長	安西 祐一郎	(独)日本学術振興会 顧問
分析研究員（副所長）	沼尾 正行	大阪大学 産業科学研究所 教授
分析研究員	調 麻佐志	東京工業大学 リベラルアーツ研究教育院 教授
分析研究員	持橋 大地	統計数理研究所 数理・推論研究系 准教授
分析研究員	遠藤 悟	

・事務局

事務長	樋口 和憲
分析調査員	田島 直人、矢代 寿寛、和田 匡路、岩瀬 文達

- 学振内部の一種のシンクタンク→科研費・学振特別研究員・海外事業などの改善に繋がる分析を行う

背景

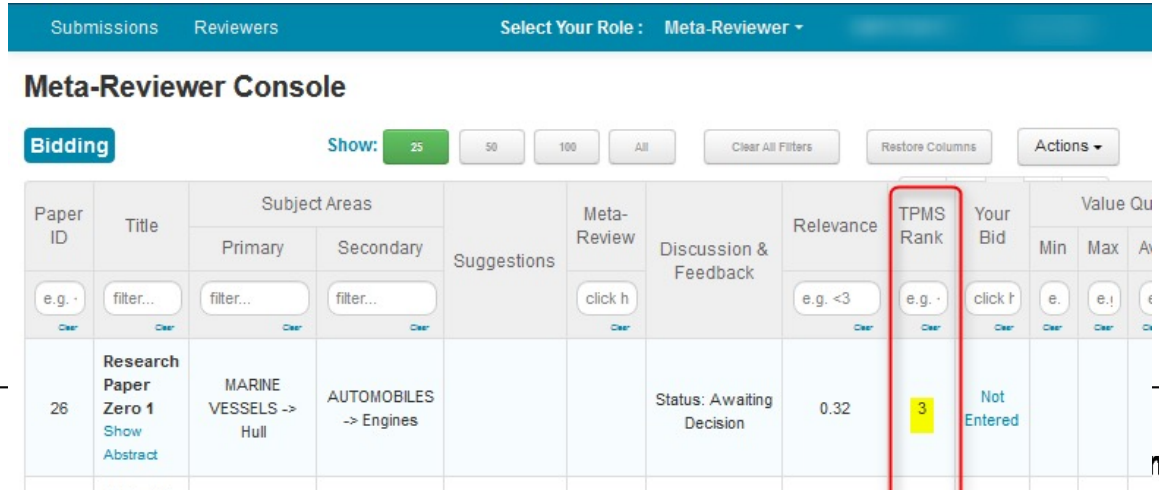
- 研究費申請が爆発的に増え、厖大な数にのぼっている
  - 米国: NSF 43,614件 (2021)
  - 中国: NSFC 10,084件 (2019)
  - 日本: 科研費 95,208件 (2021)
- 各申請は、内容を理解できる専門家によって評価される必要がある

→ 誰がやるのか？

- 分野が非常に広範になっており、経験豊かな研究者でさえ、自分の分野のすべてを網羅するのは実質不可能

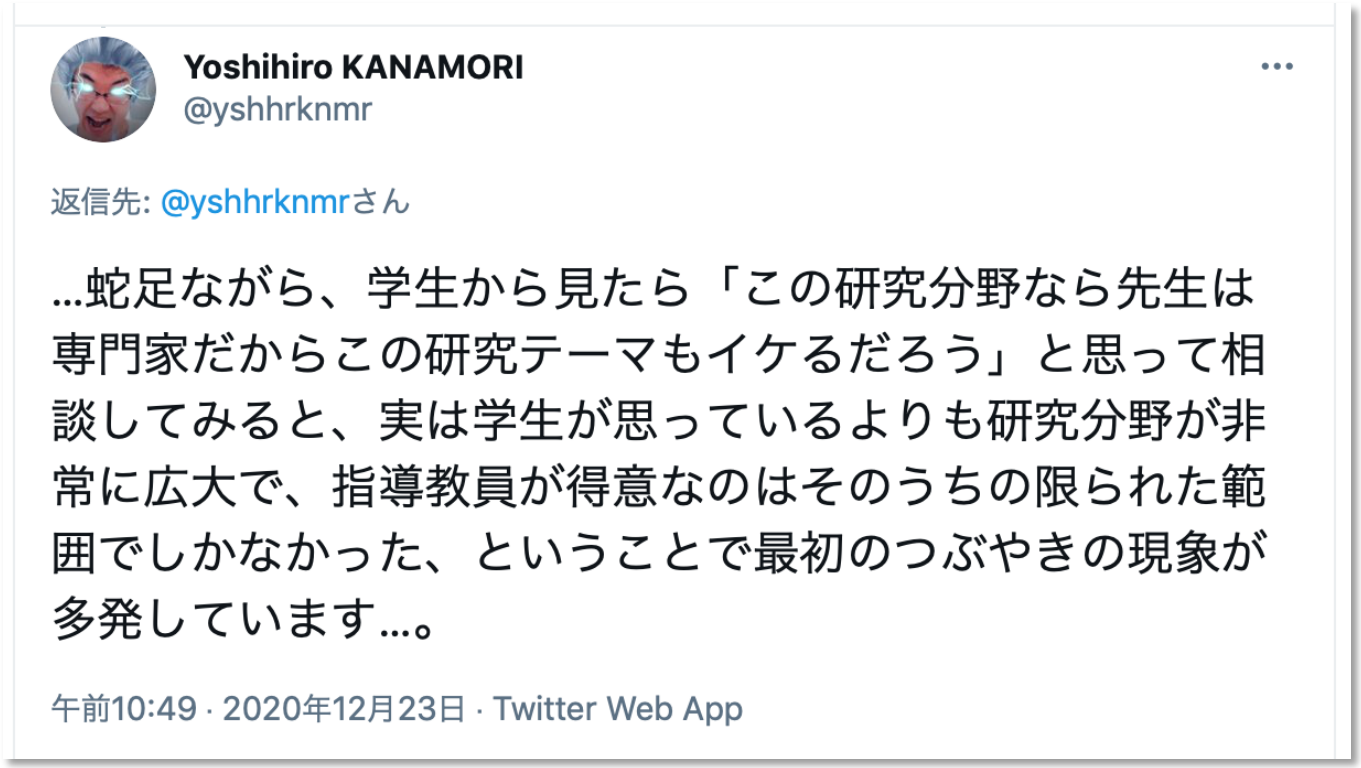
背景 (2)

- コンピュータサイエンス分野では特に、論文数が激増
  - NeurIPS 2021…9121本、CVPR 2021…7500本の論文投稿
  - 人手による査読割り当ては、もはや限界
  - 上の採択数の4〜10倍×3〜5程度の査読数が必要 (CVPR 2021では20000個以上の査読が必要)
  - TPMSが使われているが、研究者別情報は非公開
  - 未だに研究者の専門性を人手で調べる必要がある



背景 (3)

- 研究の興味が多様化しており、ある分野についてなら誰に聞いても同じ、ではもはやない



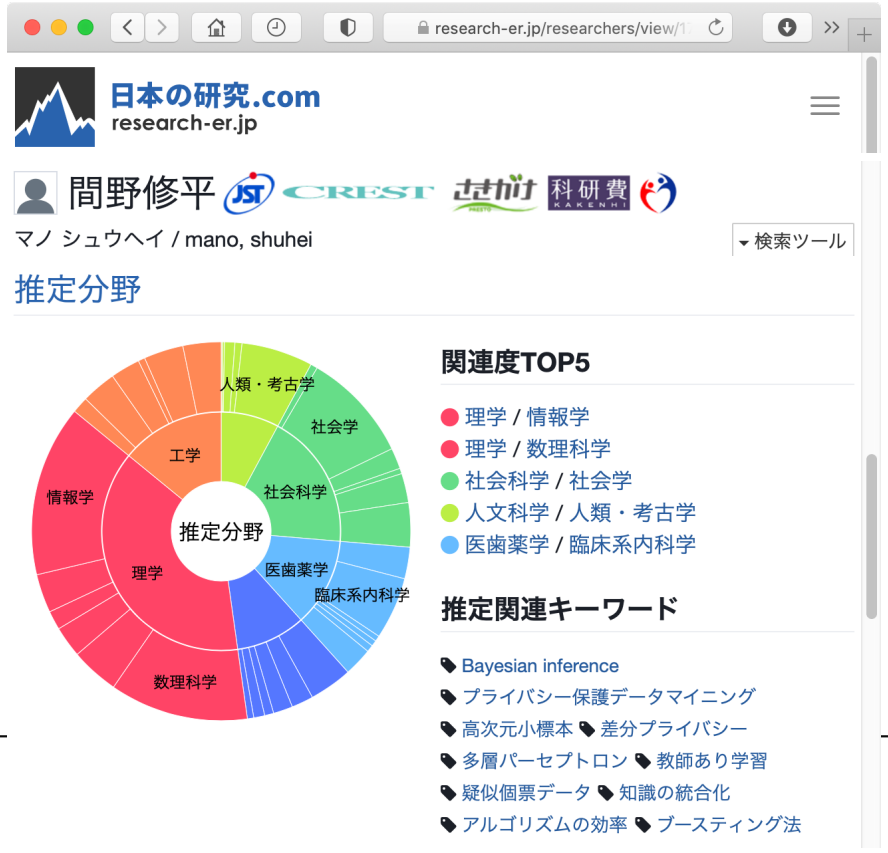
問題意識

- 自分に近い研究をしている論文・人はどれか？
  → 近い論文・研究者を自動的に発見
- この分野の研究をよく知っている先生は誰か？
  → 適切な指導者の発見 (学生、企業とも)
- 会議やジャーナルへの適切な研究者のリクルート
  → コネがなくてもコミュニティに加われるようにしたい

- Institutional Research (IR) の視点：
  各組織の研究動向や強みを、内容に応じて直接可視化したい

既存システム

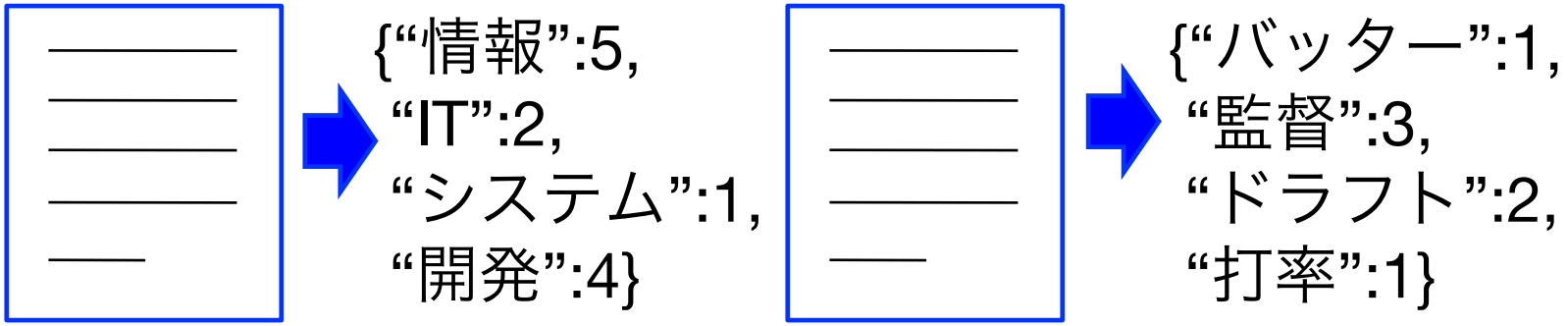
- Springer Nature Reviewer Finder、「日本の研究.com」，JDream Expert Finder, JSTサイエンスマップ、論文の内容ではなく、引用などメタ情報がベース
  - 本当の詳しい専門性は分からない、共著関係に依存



← 全然間違っている例

単語集合データ(bag of words)

- 論文は文書なので、文書の統計的な扱いを考える
- 文書の中に同時に現れる単語群には相関がある
  → 文書の中に現れた“単語”とその頻度だけを考える



- これは、文書ー単語行列で表すことができる

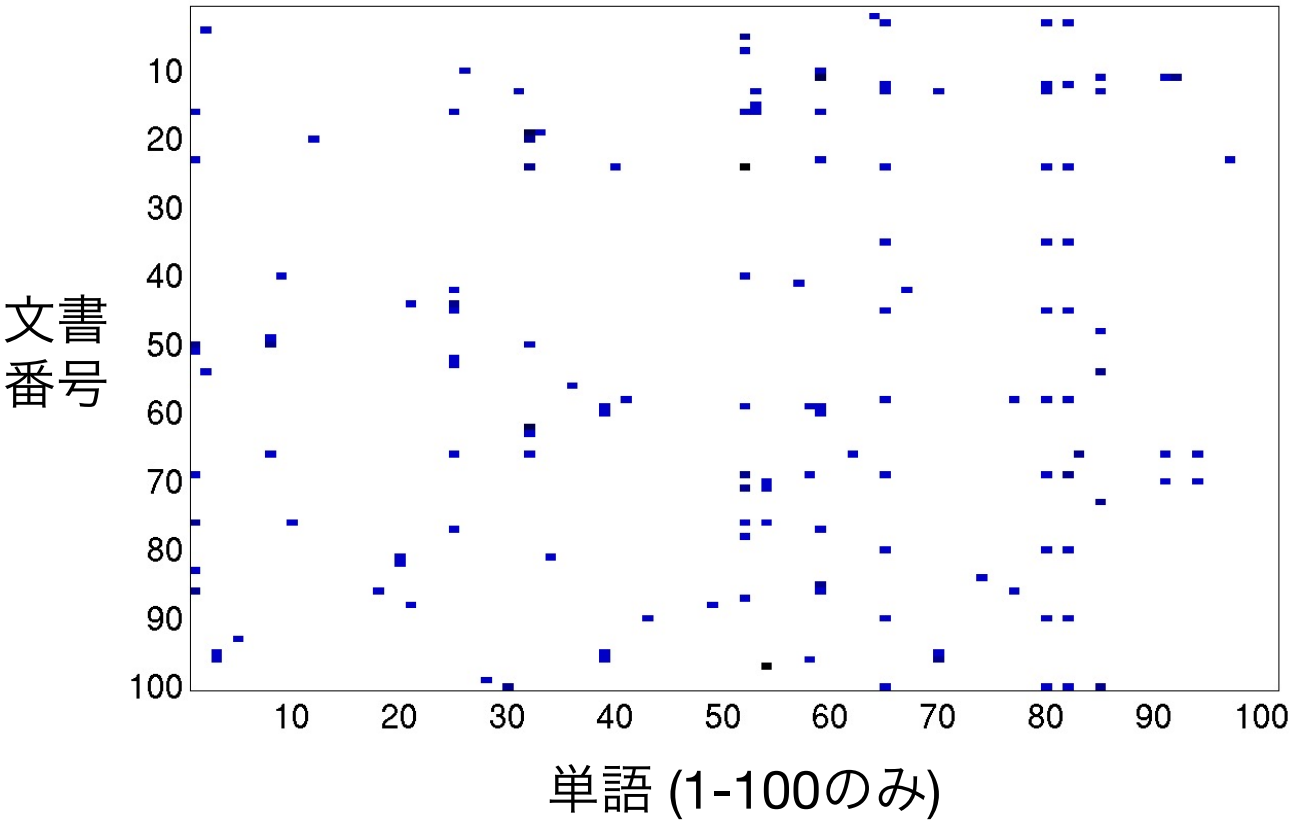
文書ー単語行列

$$D = \begin{array}{c} \\ d_1 \\ d_2 \\ d_3 \end{array} \begin{array}{c} \begin{array}{ccccccc} w_1 & w_2 & w_3 & w_4 & w_5 & w_6 & w_7 \end{array} \\ \begin{pmatrix} 1 & 2 & 1 & & 1 & & \\ & 2 & & & 1 & 1 & 1 \\ 1 & & 1 & 1 & & 2 & \end{pmatrix} \end{array}$$

- 単語集合表現は、文書ー単語行列として表せる
  - 縦が文書、横が含まれる単語、数字は出現回数
- 上では文書が3個、語彙が7個の非常に簡単な場合
  - 文書1では、w_1が1回、w_2が2回、w_3とw_5が1回出現

実際のデータの一部

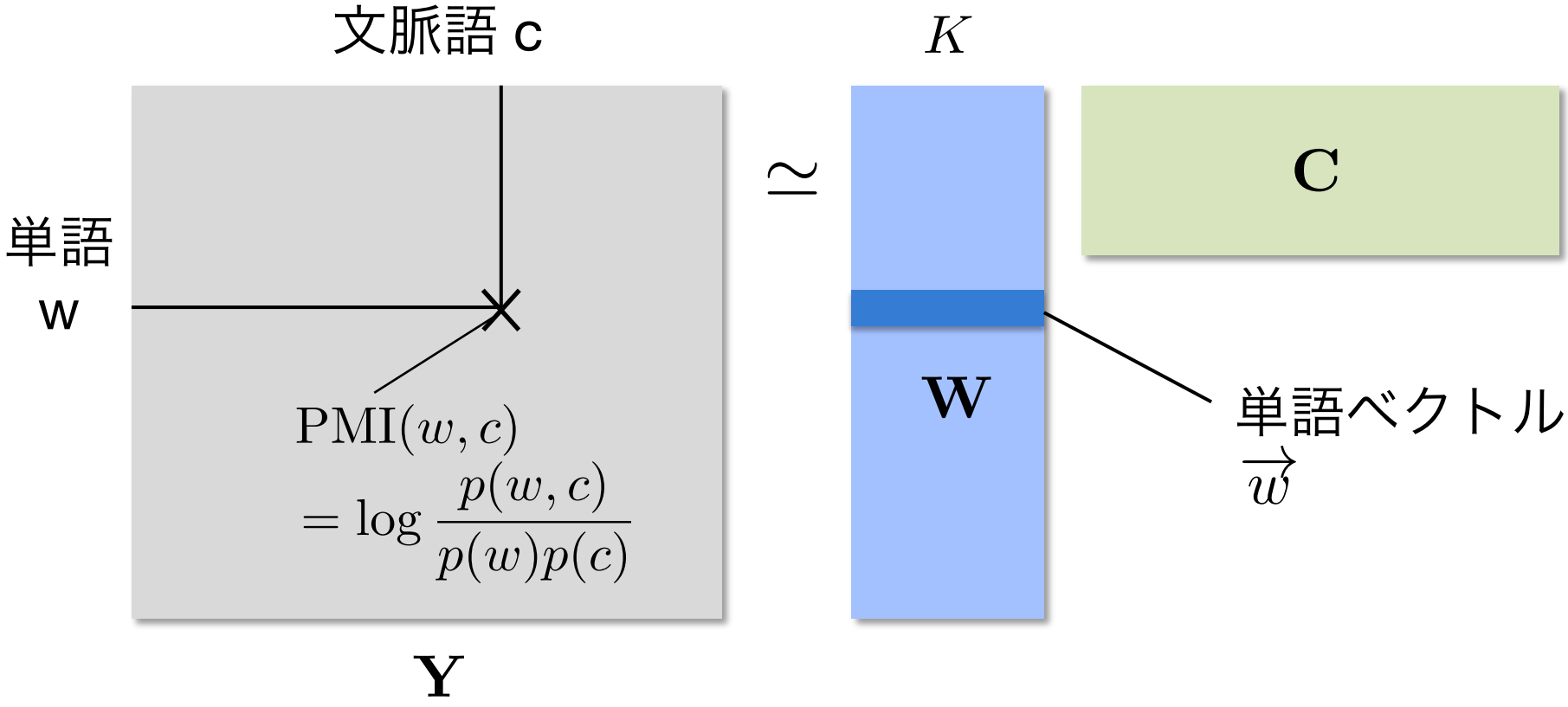
- DailyKOS (新聞)のよく使われるデータセット (一部)
  - 1行が1つの文書



- ほとんどの値は0
- 非負の値も1か、非常に小さい値

Word2vecの数理

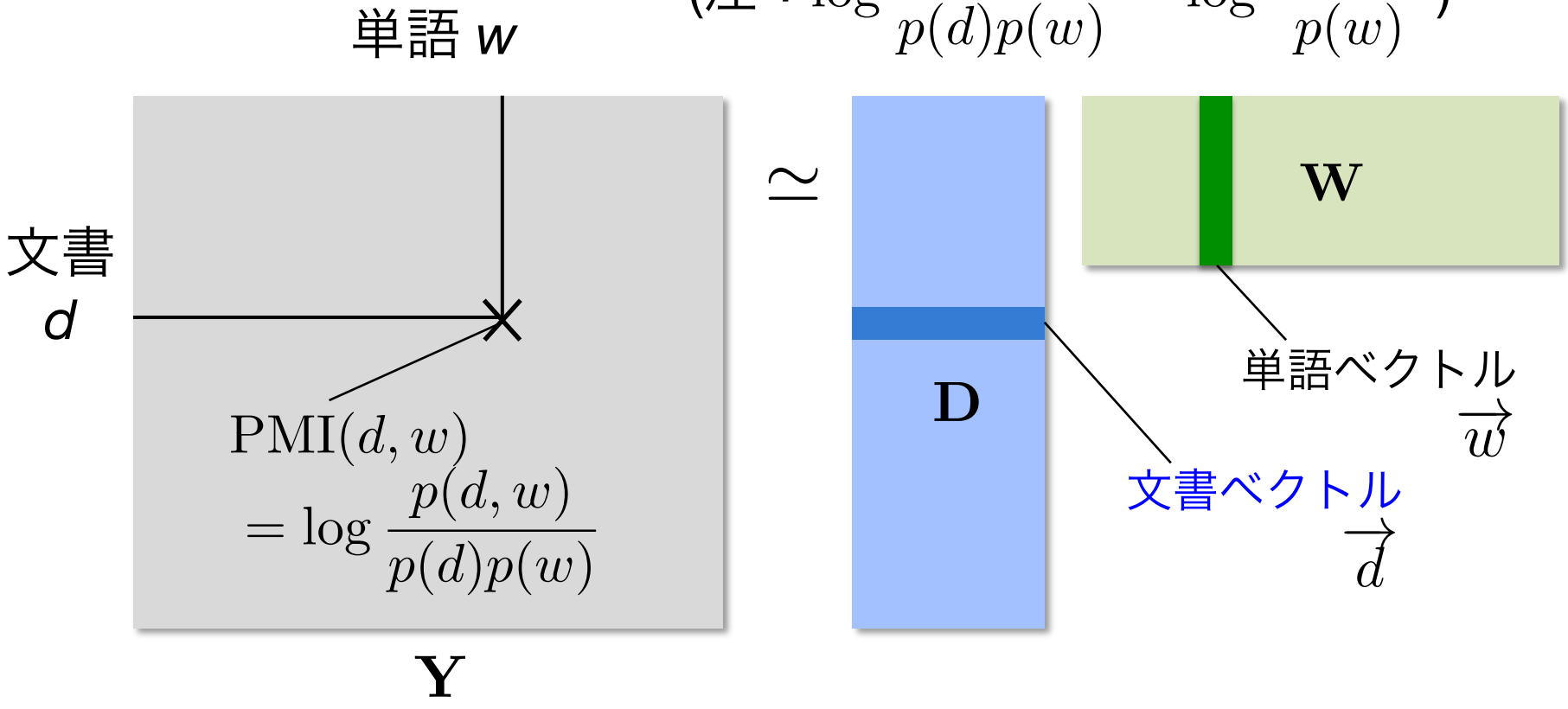
- 単語をベクトル化する、有名なWord2vec (Mikolov+ 2013)は、以下の自己相互情報量(PMI)の行列分解と等価であることが示されている (Levy and Goldberg 2014)



Word2vecから文書ベクトルへ (docvec)

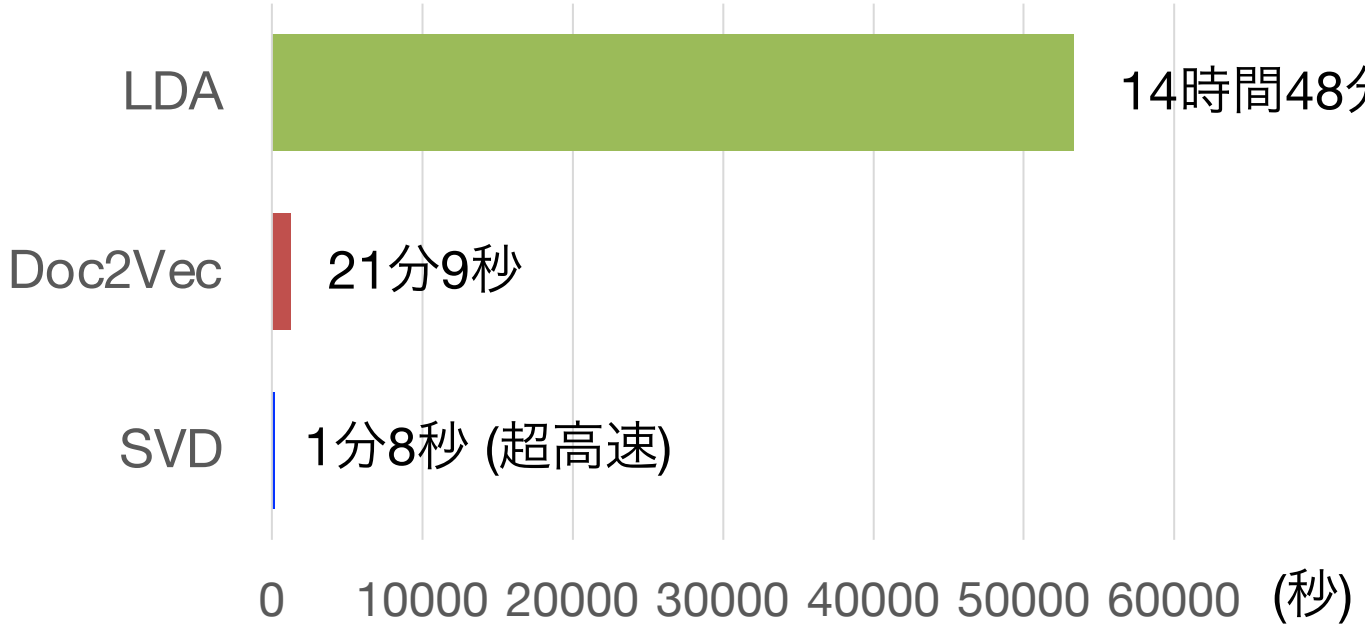
- 単語→文書、文脈語→含まれる単語 に置き換えれば、SVDで簡単に「文書ベクトル」と「単語ベクトル」を計算できる

(注：$\log \frac{p(d,w)}{p(d)p(w)} = \log \frac{p(w|d)}{p(w)}$)



計算時間の比較

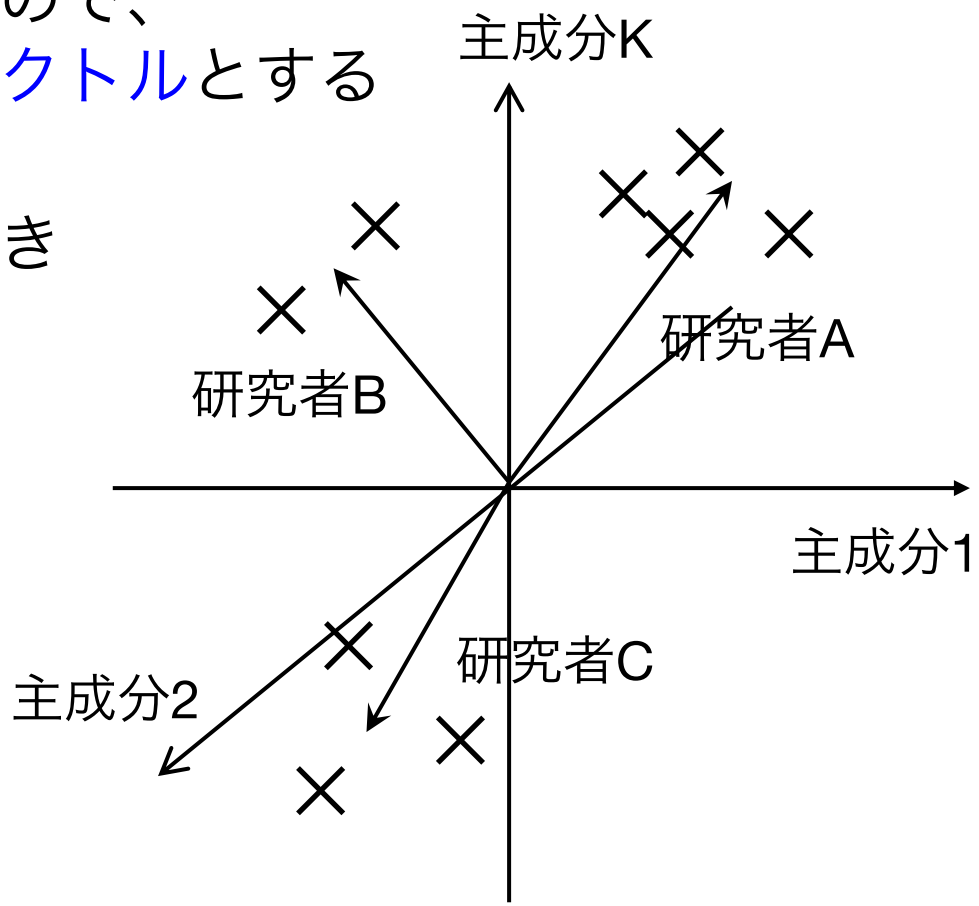
- K=1,000次元(トピック)で同じコーパスから学習した場合、



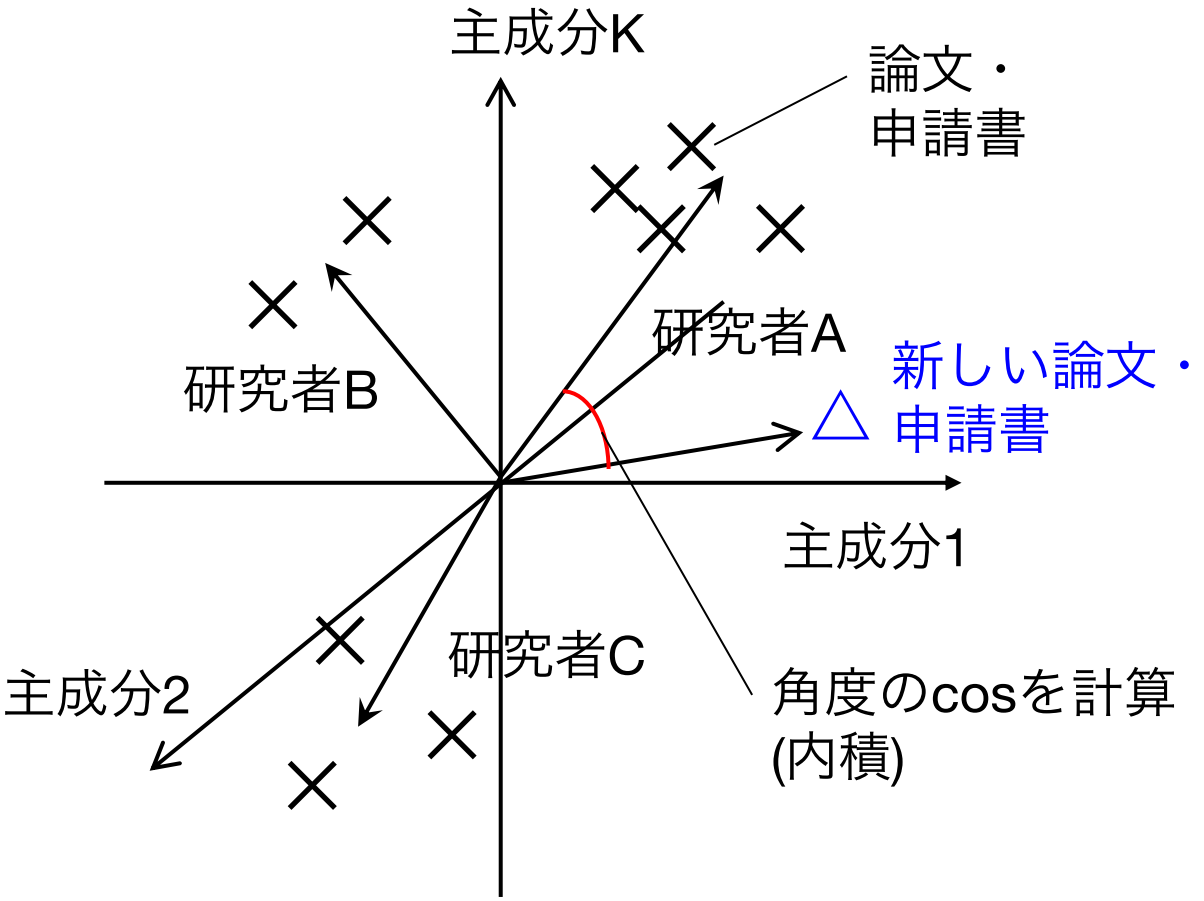
  - LDAはGibbs 1,000 iteration, Doc2Vecは100 epochs
  - データが巨大な場合、線形代数の様々な高速化が可能

文書ベクトルから研究者ベクトルへ

- 各研究者について、書いた論文/科研費申請書の文書ベクトルが得られるので、それらの平均を研究者ベクトルとする(最尤推定)
- 本来は分散も推定するべき
- 次元Kは、最大値はmin(文書数, 語彙数)

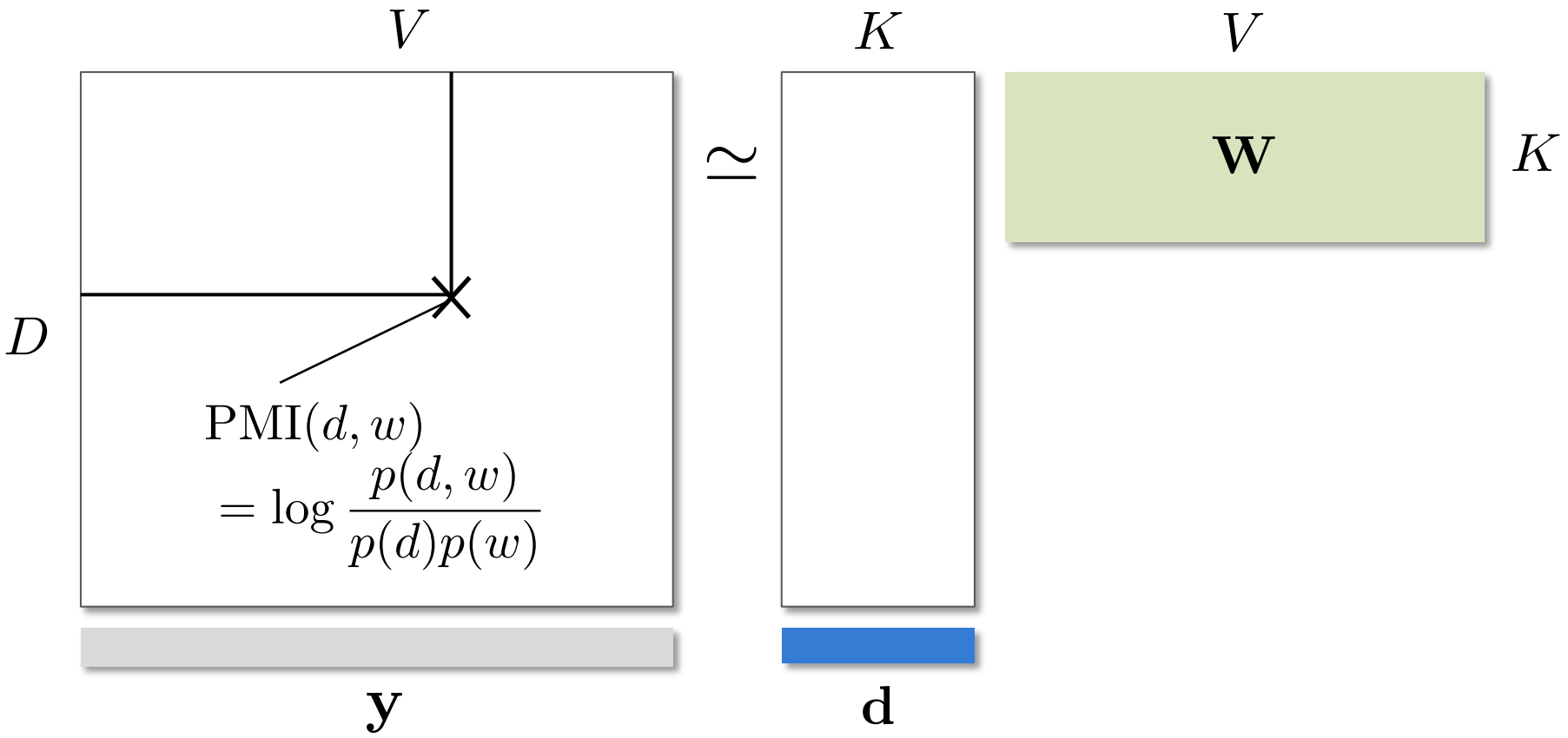


研究者ベクトルの計算と検索



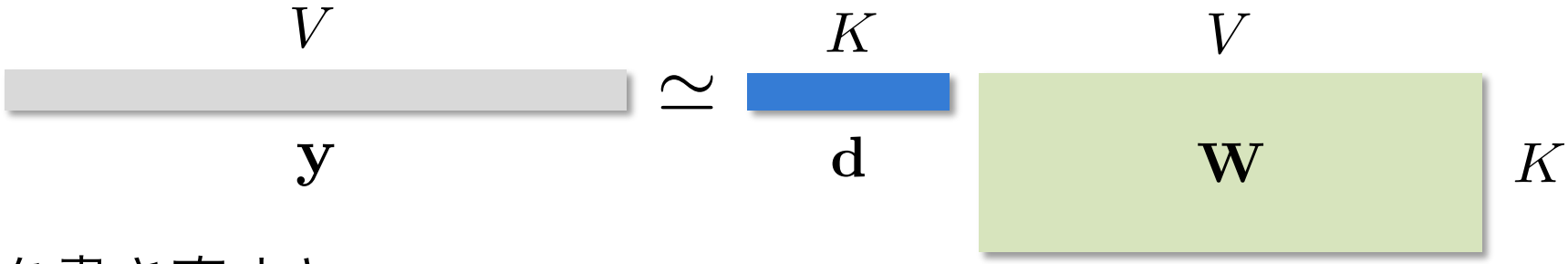
- 新しい申請書の文書ベクトルが得られれば、それを各研究者の研究者ベクトルと比べればよい

キーワード検索の方法

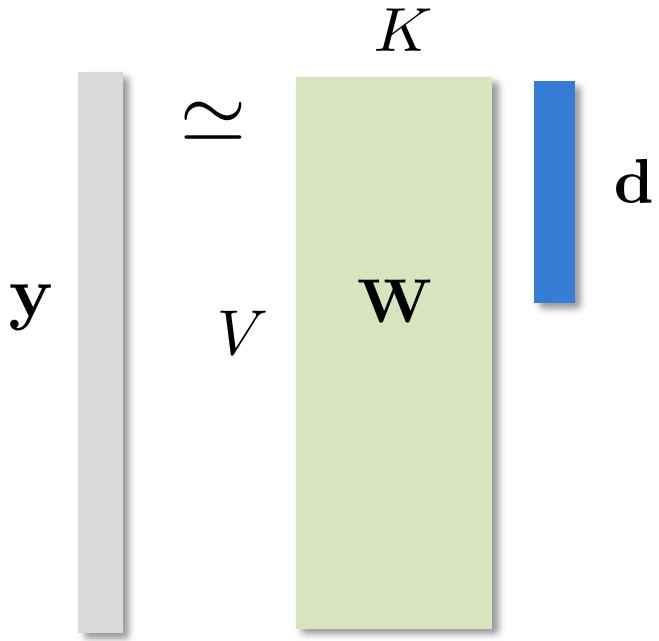


- クエリを仮想的な「文書」yと思うと、$\mathbf{y} \simeq \mathbf{dW}$ が成立
- dを研究者ベクトルと比べればよい

キーワード検索の方法 (2)



を書き直すと、



- これは線形回帰モデル！

$$\mathbf{y} \simeq \mathbf{W}\mathbf{d}$$

- よって、dの最適解は通常のOLSで、

$$\mathbf{d} = (\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\mathbf{y}$$

キーワード検索の方法 (3)

- 線形回帰の基本ですが、二乗誤差を最小化したいので

$$\begin{aligned} E &= |\mathbf{y} - \mathbf{W}\mathbf{d}|^2 \\ &= (\mathbf{y} - \mathbf{W}\mathbf{d})^T(\mathbf{y} - \mathbf{W}\mathbf{d}) \\ &= \mathbf{y}^T\mathbf{y} - 2\mathbf{d}^T\mathbf{W}^T\mathbf{y} + \mathbf{d}^T\mathbf{W}^T\mathbf{W}\mathbf{d} \end{aligned}$$

- よって

$$\frac{\partial E}{\partial \mathbf{d}} = -2\mathbf{W}^T\mathbf{y} + 2\mathbf{W}^T\mathbf{W}\mathbf{d} = 0$$

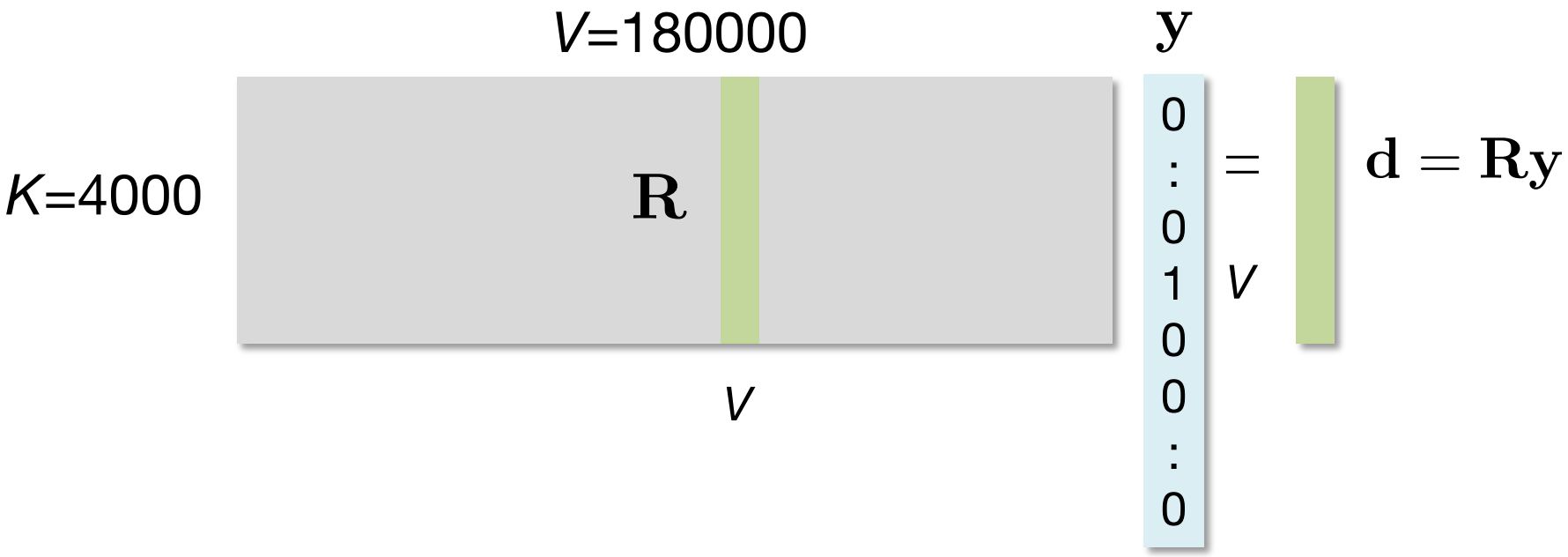
$$\therefore \ \mathbf{d} = (\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\mathbf{y}$$

- 事前に $\mathbf{R} = (\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T$ を計算しておけば、

$$\mathbf{d} = \mathbf{R}\mathbf{y}$$

で一瞬で求まる

キーワード検索の方法 (4)



- yの要素はほとんど0なので、Ryの掛け算は結局、Rの対応する列を取り出してくればよい
- Rはディスクにmmap()しておけば、メモリ使用も最小(対応済み)

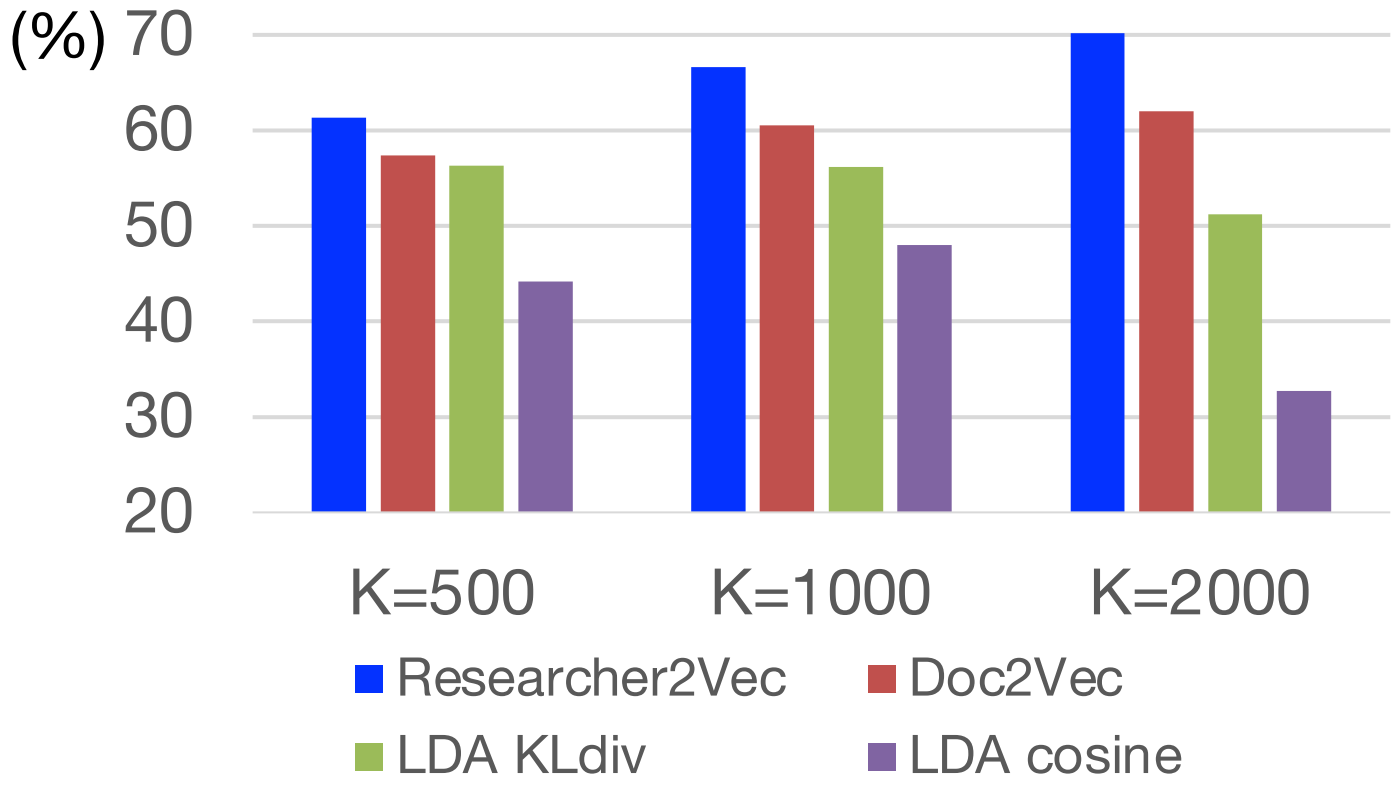
実験とデータ

- (a) 言語処理学会年次大会の20年分の論文データと
  (b) arXivのunarXiveデータセットを使って実験
- 対象となる著者は一定以上の論文がある(a) 499人および
  (b) 13989人
- テストデータの文書の著者を推定し、スコア順に並べた際の平均適合率 (Mean Average Precision, MAP)を計算
  - MAP=1：スコア最上位がすべて真の著者
  - Doc2Vec, LDAと比較

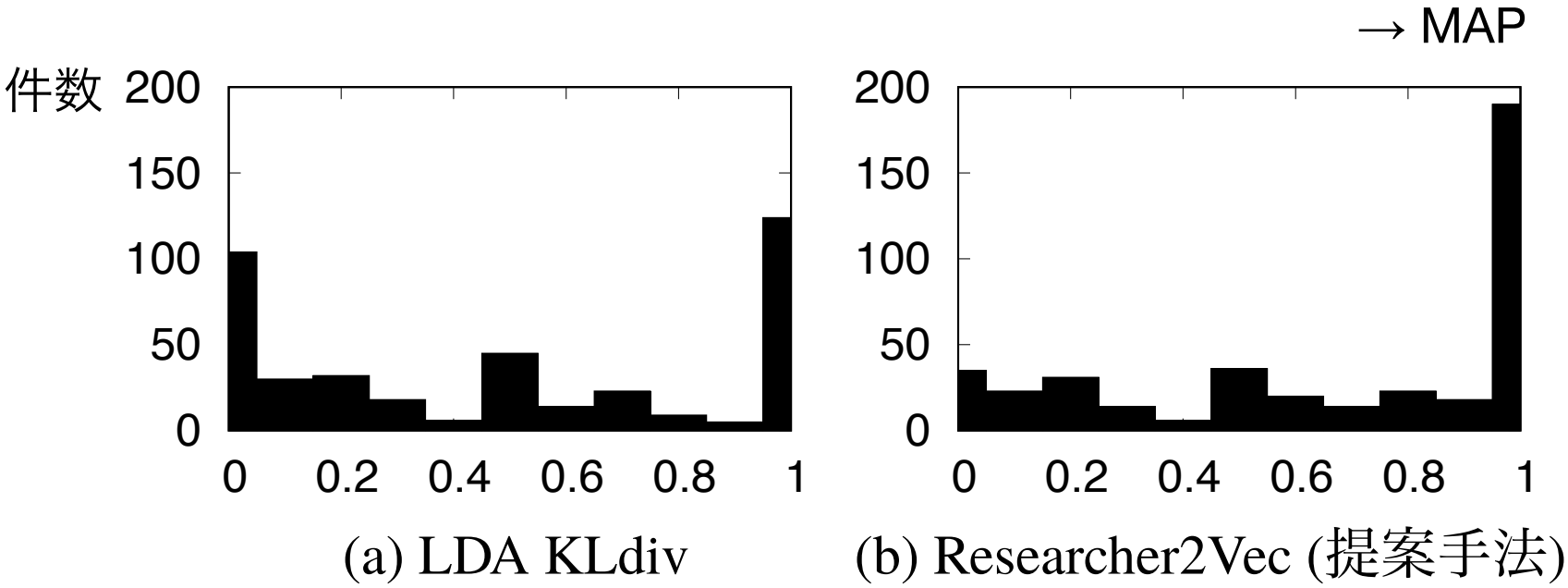
"Researcher2Vec: Neural Linear Model of Scholar Recommendation for Funding Agency". Daichi Mochihashi. International Society for Scientometrics and Informatics (ISSI 2023), Vol. 2, pp.329-335, 2023.

論文著者の推薦精度 (平均適合率)

- Doc2Vecを超えて、提案法が常に最高精度
  - LDAは、桂井ら(2016)の確率分布のコサイン類似度よりKLダイバージェンスで測った方がよい



平均適合率(MAP)の分布

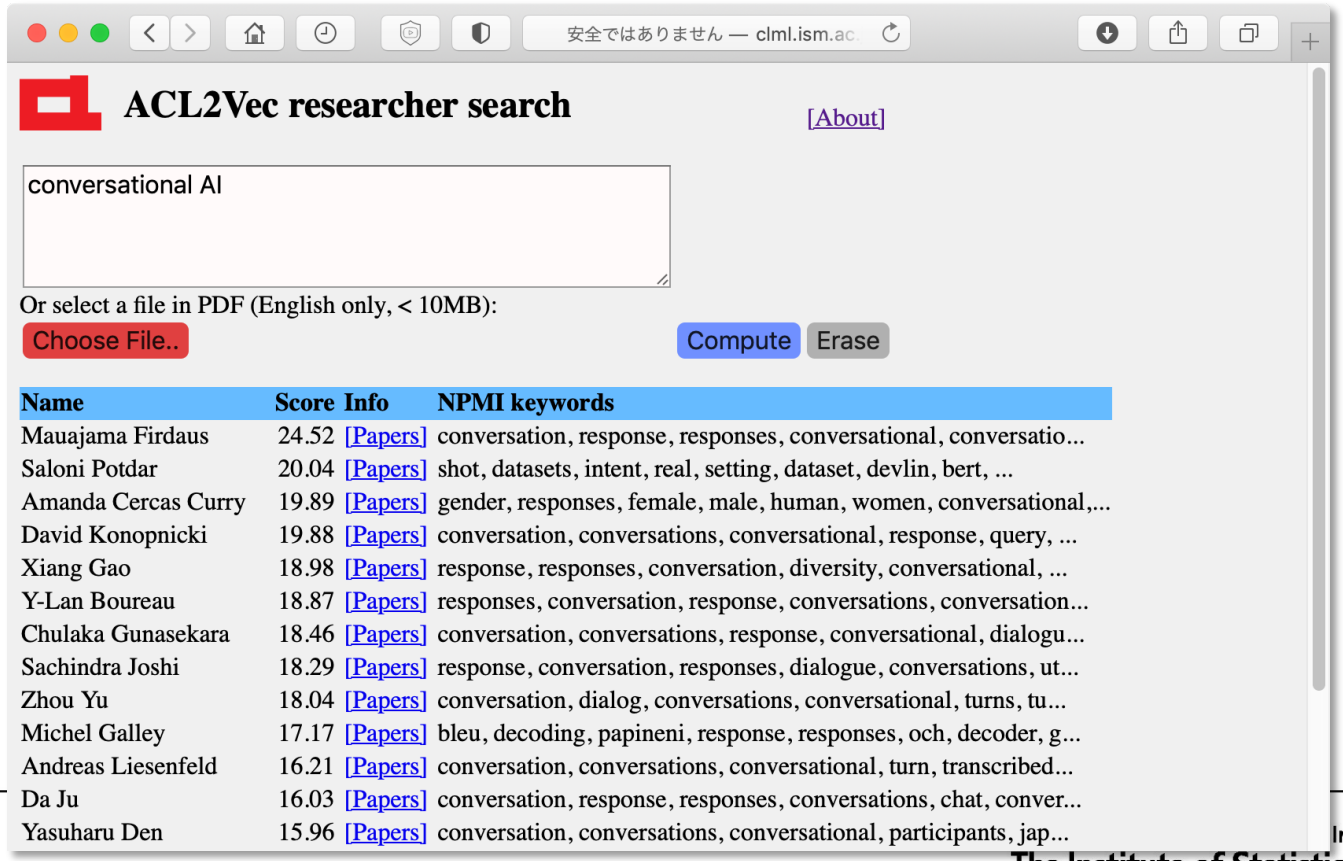


- 提案手法では、ほとんどの場合に平均適合率＝1
  → 著者をスコア順に並び替えたとき、真の著者が最上位を占める
- それ以外は、先生と興味が違う学生の論文や、英語論文

デモ

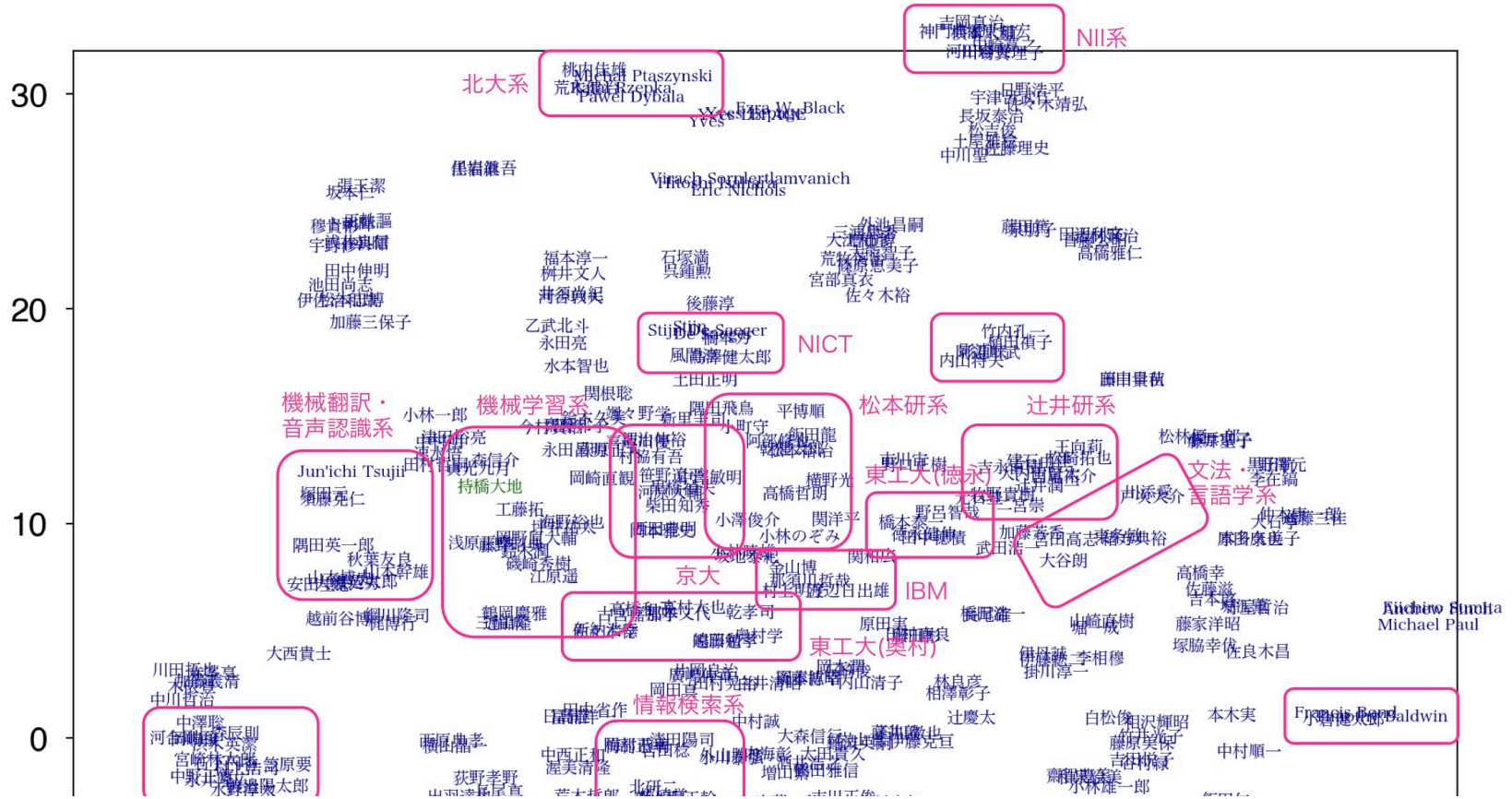
http://clml.ism.ac.jp/ACL2Vec-authors/

- 公開サーバですので、誰でも検索を試すことができます
  - 自然言語処理の国際会議論文のビューアにもなります



Name	Score	Info	NPMI keywords
Mauajama Firdaus	24.52	[Papers]	conversation, response, responses, conversational, conversatio...
Saloni Potdar	20.04	[Papers]	shot, datasets, intent, real, setting, dataset, devlin, bert, ...
Amanda Cercas Curry	19.89	[Papers]	gender, responses, female, male, human, women, conversational,...
David Konopnicki	19.88	[Papers]	conversation, conversations, conversational, response, query, ...
Xiang Gao	18.98	[Papers]	response, responses, conversation, diversity, conversational, ...
Y-Lan Boureau	18.87	[Papers]	responses, conversation, response, conversations, conversation...
Chulaka Gunasekara	18.46	[Papers]	conversation, conversations, response, conversational, dialogu...
Sachindra Joshi	18.29	[Papers]	response, conversation, responses, dialogue, conversations, ut...
Zhou Yu	18.04	[Papers]	conversation, dialog, conversations, conversational, turns, tu...
Michel Galley	17.17	[Papers]	bleu, decoding, papineni, response, responses, och, decoder, g...
Andreas Liesenfeld	16.21	[Papers]	conversation, conversations, conversational, turn, transcribed...
Da Ju	16.03	[Papers]	conversation, response, responses, conversations, chat, conver...
Yasuharu Den	15.96	[Papers]	conversation, conversations, conversational, participants, jap...

研究者の可視化



- t-SNEによる可視化 (可視化は1通りではありません)

日本学術振興会での取り組み

- 科研費の大型種目(特別推進、基盤S、Aなど)について、事前審査を行う審査員の推薦に本システム (の前駆のLDA版) を提供中
  - 主任研究員からは、非常に助かるとのご意見をいただいています
- 海外種目については、10万人程度の海外審査員候補をまず統計的に選定し、その中で本システムで審査員候補をスコア順に100名提示
  - 実際に採択された審査員の上位は、100名中で上位10%程度の順位!

前半のまとめ

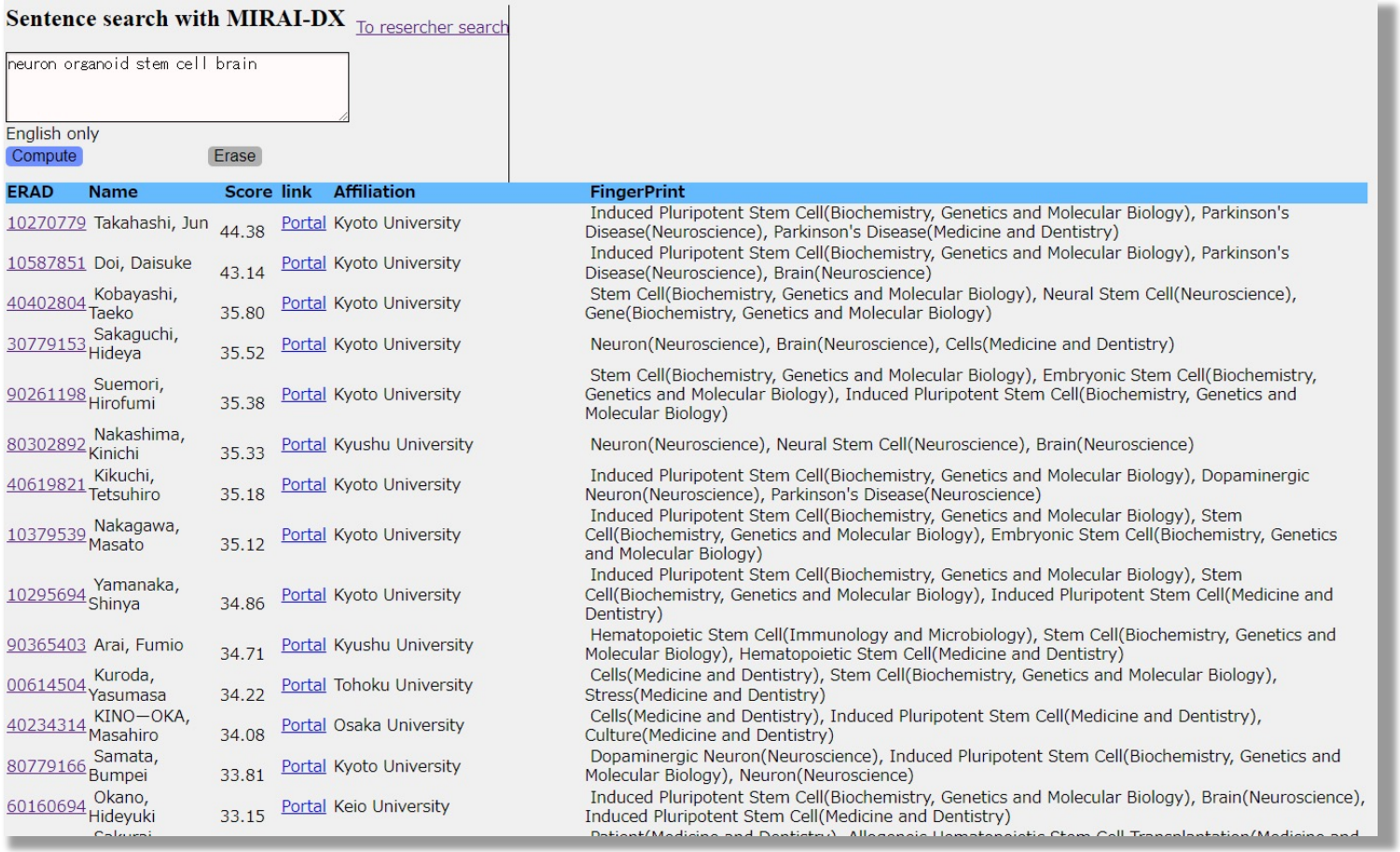
- 論文からWord2vecと同等のニューラル文書ベクトルをSVDで高速に求め、それを書いた研究者の “研究者ベクトル”を計算する手法を提案
  - Doc2Vecの20〜40倍高速、解析的な検索解、省メモリ
  - Doc2VecおよびLDAを超えて最高精度
- 実際にACL Anthologyの全文コーパスから、論文に含まれる単語で研究者を検索できるシステムを公開(ACL2Vec)
- 日本学術振興会で同様のシステムが審査の補助として運用中

MIRAI-DXプロジェクトでの取り組み

- MIRAI-DX: 各大学のURAの方々を中心とした研究支援DX
  - 自然科学研究機構の小泉先生がリーダー

MIRAIDX2Vec

https://miraidx2vec.nins.jp/



Sentence search with MIRAI-DX To resercher search

neuron organoid stem cell brain

English only

Compute Erase

ERAD	Name	Score	link	Affiliation	FingerPrint
10270779	Takahashi, Jun	44.38	Portal	Kyoto University	Induced Pluripotent Stem Cell(Biochemistry, Genetics and Molecular Biology), Parkinson's Disease(Neuroscience), Parkinson's Disease(Medicine and Dentistry)
10587851	Doi, Daisuke	43.14	Portal	Kyoto University	Induced Pluripotent Stem Cell(Biochemistry, Genetics and Molecular Biology), Parkinson's Disease(Neuroscience), Brain(Neuroscience)
40402804	Kobayashi, Taeko	35.80	Portal	Kyoto University	Stem Cell(Biochemistry, Genetics and Molecular Biology), Neural Stem Cell(Neuroscience), Gene(Biochemistry, Genetics and Molecular Biology)
30779153	Sakaguchi, Hideya	35.52	Portal	Kyoto University	Neuron(Neuroscience), Brain(Neuroscience), Cells(Medicine and Dentistry)
90261198	Suemori, Hirofumi	35.38	Portal	Kyoto University	Stem Cell(Biochemistry, Genetics and Molecular Biology), Embryonic Stem Cell(Biochemistry, Genetics and Molecular Biology), Induced Pluripotent Stem Cell(Biochemistry, Genetics and Molecular Biology)
80302892	Nakashima, Kinichi	35.33	Portal	Kyushu University	Neuron(Neuroscience), Neural Stem Cell(Neuroscience), Brain(Neuroscience)
40619821	Kikuchi, Tetsuhiro	35.18	Portal	Kyoto University	Induced Pluripotent Stem Cell(Biochemistry, Genetics and Molecular Biology), Dopaminergic Neuron(Neuroscience), Parkinson's Disease(Neuroscience)
10379539	Nakagawa, Masato	35.12	Portal	Kyoto University	Induced Pluripotent Stem Cell(Biochemistry, Genetics and Molecular Biology), Stem Cell(Biochemistry, Genetics and Molecular Biology), Embryonic Stem Cell(Biochemistry, Genetics and Molecular Biology)
10295694	Yamanaka, Shinya	34.86	Portal	Kyoto University	Induced Pluripotent Stem Cell(Biochemistry, Genetics and Molecular Biology), Stem Cell(Biochemistry, Genetics and Molecular Biology), Induced Pluripotent Stem Cell(Medicine and Dentistry)
90365403	Arai, Fumio	34.71	Portal	Kyushu University	Hematopoietic Stem Cell(Immunology and Microbiology), Stem Cell(Biochemistry, Genetics and Molecular Biology), Hematopoietic Stem Cell(Medicine and Dentistry)
00614504	Kuroda, Yasumasa	34.22	Portal	Tohoku University	Cells(Medicine and Dentistry), Stem Cell(Biochemistry, Genetics and Molecular Biology), Stress(Medicine and Dentistry)
40234314	KINO—OKA, Masahiro	34.08	Portal	Osaka University	Cells(Medicine and Dentistry), Induced Pluripotent Stem Cell(Medicine and Dentistry), Culture(Medicine and Dentistry)
80779166	Samata, Bumpei	33.81	Portal	Kyoto University	Dopaminergic Neuron(Neuroscience), Induced Pluripotent Stem Cell(Biochemistry, Genetics and Molecular Biology), Neuron(Neuroscience)
60160694	Okano, Hideyuki	33.15	Portal	Keio University	Induced Pluripotent Stem Cell(Biochemistry, Genetics and Molecular Biology), Brain(Neuroscience), Induced Pluripotent Stem Cell(Medicine and Dentistry)

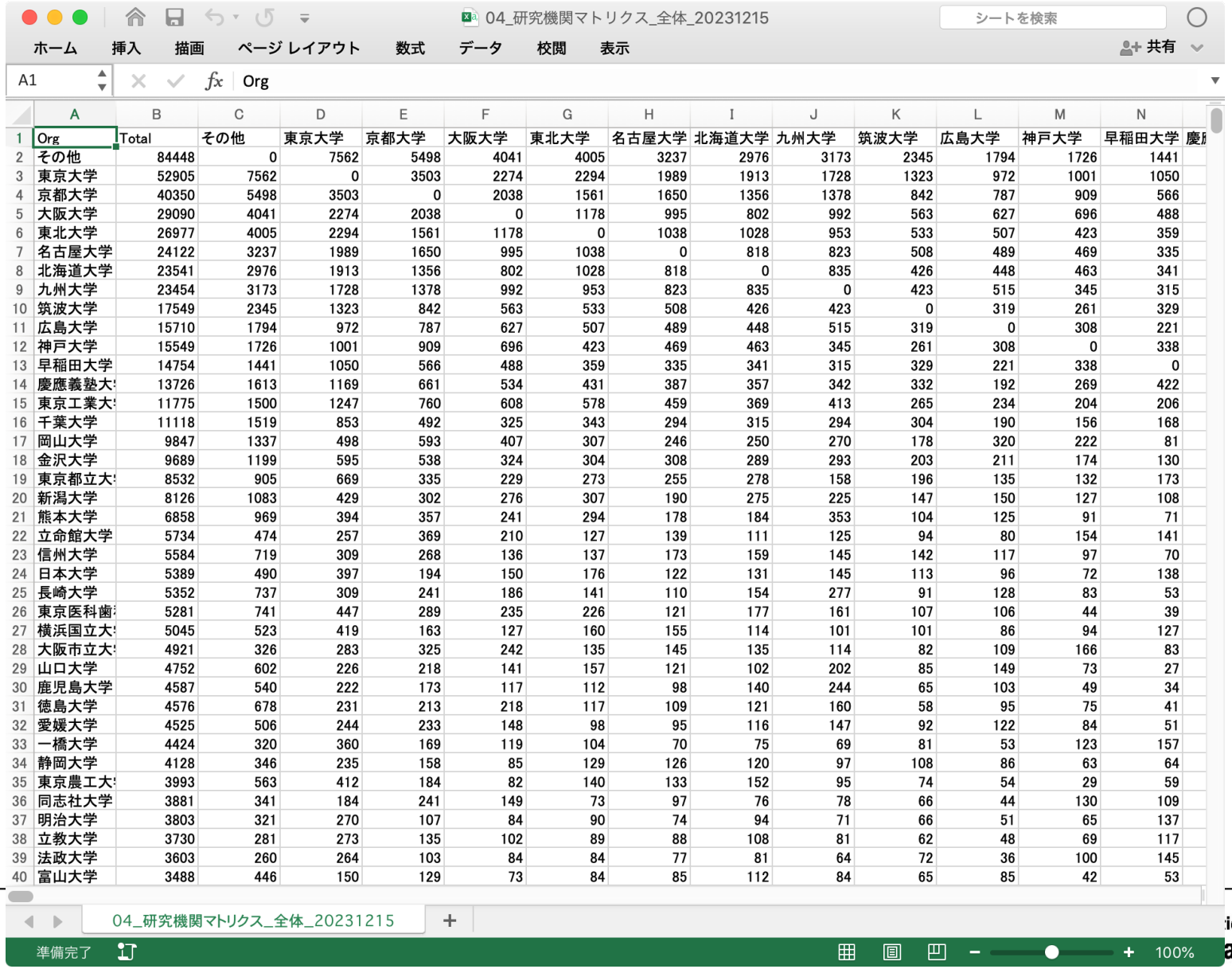
- E-radの研究費ファンディング情報をベクトル化して同様に検索可能に

大学・機関間の共同研究情報

- MIRAI-DXで持っている研究の申請データから、
  同じ研究で代表者・分担者として名前を連ねている場合に
  頻度を+1したデータ
  - 例：研究プロジェクト (Aさん(東北大)・Bさん(東工大)・
    Cさん(医科歯科大)
    → (東北大, 東工大)++, (東工大, 医科歯科大)++,
    (東北大, 医科歯科大)++

- この行列から、各研究機関の特徴について何がいえるか？
  (信州大URA 久保さんとの共同研究)

大学・機関間の共同研究マトリクス



04_研究機関マトリクス_全体_20231215

シートを検索

ホーム　挿入　描画　ページ レイアウト　数式　データ　校閲　表示　共有

A1　Org

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	
1	Org	Total	その他	東京大学	京都大学	大阪大学	東北大学	名古屋大学	北海道大学	九州大学	筑波大学	広島大学	神戸大学	早稲田大学	慶
2	その他	84448	0	7562	5498	4041	4005	3237	2976	3173	2345	1794	1726	1441	
3	東京大学	52905	7562	0	3503	2274	2294	1989	1913	1728	1323	972	1001	1050	
4	京都大学	40350	5498	3503	0	2038	1561	1650	1356	1378	842	787	909	566	
5	大阪大学	29090	4041	2274	2038	0	1178	995	802	992	563	627	696	488	
6	東北大学	26977	4005	2294	1561	1178	0	1038	1028	953	533	507	423	359	
7	名古屋大学	24122	3237	1989	1650	995	1038	0	818	823	508	489	469	335	
8	北海道大学	23541	2976	1913	1356	802	1028	818	0	835	426	448	463	341	
9	九州大学	23454	3173	1728	1378	992	953	823	835	0	423	515	345	315	
10	筑波大学	17549	2345	1323	842	563	533	508	426	423	0	319	261	329	
11	広島大学	15710	1794	972	787	627	507	489	448	515	319	0	308	221	
12	神戸大学	15549	1726	1001	909	696	423	469	463	345	261	308	0	338	
13	早稲田大学	14754	1441	1050	566	488	359	335	341	315	329	221	338	0	
14	慶應義塾大	13726	1613	1169	661	534	431	387	357	342	332	192	269	422	
15	東京工業大	11775	1500	1247	760	608	578	459	369	413	265	234	204	206	
16	千葉大学	11118	1519	853	492	325	343	294	315	294	304	190	156	168	
17	岡山大学	9847	1337	498	593	407	307	246	250	270	178	320	222	81	
18	金沢大学	9689	1199	595	538	324	304	308	289	293	203	211	174	130	
19	東京都立大	8532	905	669	335	229	273	255	278	158	196	135	132	173	
20	新潟大学	8126	1083	429	302	276	307	190	275	225	147	150	127	108	
21	熊本大学	6858	969	394	357	241	294	178	184	353	104	125	91	71	
22	立命館大学	5734	474	257	369	210	127	139	111	125	94	80	154	141	
23	信州大学	5584	719	309	268	136	137	173	159	145	142	117	97	70	
24	日本大学	5389	490	397	194	150	176	122	131	145	113	96	72	138	
25	長崎大学	5352	737	309	241	186	141	110	154	277	91	128	83	53	
26	東京医科歯	5281	741	447	289	235	226	121	177	161	107	106	44	39	
27	横浜国立大	5045	523	419	163	127	160	155	114	101	101	86	94	127	
28	大阪市立大	4921	326	283	325	242	135	145	135	114	82	109	166	83	
29	山口大学	4752	602	226	218	141	157	121	102	202	85	149	73	27	
30	鹿児島大学	4587	540	222	173	117	112	98	140	244	65	103	49	34	
31	徳島大学	4576	678	231	213	218	117	109	121	160	58	95	75	41	
32	愛媛大学	4525	506	244	233	148	98	95	116	147	92	122	84	51	
33	一橋大学	4424	320	360	169	119	104	70	75	69	81	53	123	157	
34	静岡大学	4128	346	235	158	85	129	126	120	97	108	86	63	64	
35	東京農工大	3993	563	412	184	82	140	133	152	95	74	54	29	59	
36	同志社大学	3881	341	184	241	149	73	97	76	78	66	44	130	109	
37	明治大学	3803	321	270	107	84	90	74	94	71	66	51	65	137	
38	立教大学	3730	281	273	135	102	89	88	108	81	62	48	69	117	
39	法政大学	3603	260	264	103	84	84	77	81	64	72	36	100	145	
40	富山大学	3488	446	150	129	73	84	85	112	84	65	85	42	53	

04_研究機関マトリクス_全体_20231215

準備完了　100%

機関間関係行列

Org	Total	その他	東京大学	京都大学	大阪大学	東北大学	名古屋大学	北海道大学	九州大学	筑波大学	広島大学	神戸大学	早稲田大学
その他	84448	0	7562	5498	4041	4005	3237	2976	3173	2345	1794	1726	1441
東京大学	52905	7562	0	3503	2274	2294	1989	1913	1728	1323	972	1001	1050
京都大学	40350	5498	3503	0	2038	1561	1650	1356	1378	842	787	909	566
大阪大学	29090	4041	2274	2038	0	1178	995	802	992	563	627	696	488
東北大学	26977	4005	2294	1561	1178	0	1038	1028	953	533	507	423	359
名古屋大学	24122	3237	1989	1650	995	1038	0	818	823	508	489	469	335
北海道大学	23541	2976	1913	1356	802	1028	818	0	835	426	448	463	341
九州大学	23454	3173	1728	1378	992	953	823	835	0	423	515	345	315
筑波大学	17549	2345	1323	842	563	533	508	426	423	0	319	261	329
広島大学	15710	1794	972	787	627	507	489	448	515	319	0	308	221
神戸大学	15549	1726	1001	909	696	423	469	463	345	261	308	0	338
早稲田大学	14754	1441	1050	566	488	359	335	341	315	329	221	338	0
慶應義塾大	13726	1613	1169	661	534	431	387	357	342	332	192	269	422
東京工業大	11775	1500	1247	760	608	578	459	369	413	265	234	204	206
千葉大学	11118	1519	853	492	325	343	294	315	294	304	190	156	168

- 生の頻度では、有力大学が非常に大きな数
  → この影響を抑える統計的な分析が必要
- 自己相互情報量 (Pointwise Mutual Information) を用いる
  - x, y をそれぞれ研究機関として、

$$\mathrm{PMI}(x, y) = \log \frac{p(x, y)}{p(x)p(y)}$$

自己相互情報量の意味

$$\mathrm{PMI}(x, y) = \log \frac{p(x, y)}{p(x)p(y)}$$

- ベイズの定理から、

$$\log \frac{p(x, y)}{p(x)p(y)} = \log \frac{p(y|x)}{p(y)}$$

- よってPMIは、「xがyと共同研究する確率」と「yと共同研究する一般的な確率」との比
  - この比が1＝一般的な確率と同じなら、PMIはlog1 = 0
  - PMI>0なら標準より多い、<0なら標準より少ない
  - [一般に、東大や京大はp(y)が大きい]

機関間関係行列

Org	Total	その他	東京大学	京都大学	大阪大学	東北大学	名古屋大学	北海道大学	九州大学	筑波大学	広島大学	神戸大学	早稲田大学
その他	84448	0	7562	5498	4041	4005	3237	2976	3173	2345	1794	1726	1441
東京大学	52905	7562	0	3503	2274	2294	1989	1913	1728	1323	972	1001	1050
京都大学	40350	5498	3503	0	2038	1561	1650	1356	1378	842	787	909	566
大阪大学	29090	4041	2274	2038	0	1178	995	802	992	563	627	696	488
東北大学	26977	4005	2294	1561	1178	0	1038	1028	953	533	507	423	359
名古屋大学	24122	3237	1989	1650	995	1038	0	818	823	508	489	469	335
北海道大学	23541	2976	1913	1356	802	1028	818	0	835	426	448	463	341
九州大学	23454	3173	1728	1378	992	953	823	835	0	423	515	345	315
筑波大学	17549	2345	1323	842	563	533	508	426	423	0	319	261	329
広島大学	15710	1794	972	787	627	507	489	448	515	319	0	308	221
神戸大学	15549	1726	1001	909	696	423	469	463	345	261	308	0	338
早稲田大学	14754	1441	1050	566	488	359	335	341	315	329	221	338	0
慶應義塾大	13726	1613	1169	661	534	431	387	357	342	332	192	269	422
東京工業大	11775	1500	1247	760	608	578	459	369	413	265	234	204	206
千葉大学	11118	1519	853	492	325	343	294	315	294	304	190	156	168

- 機関xと機関yの共同研究頻度を $n(x,y)$ とすると、

$$p(x,y) = n(x,y)/N$$

- よって、

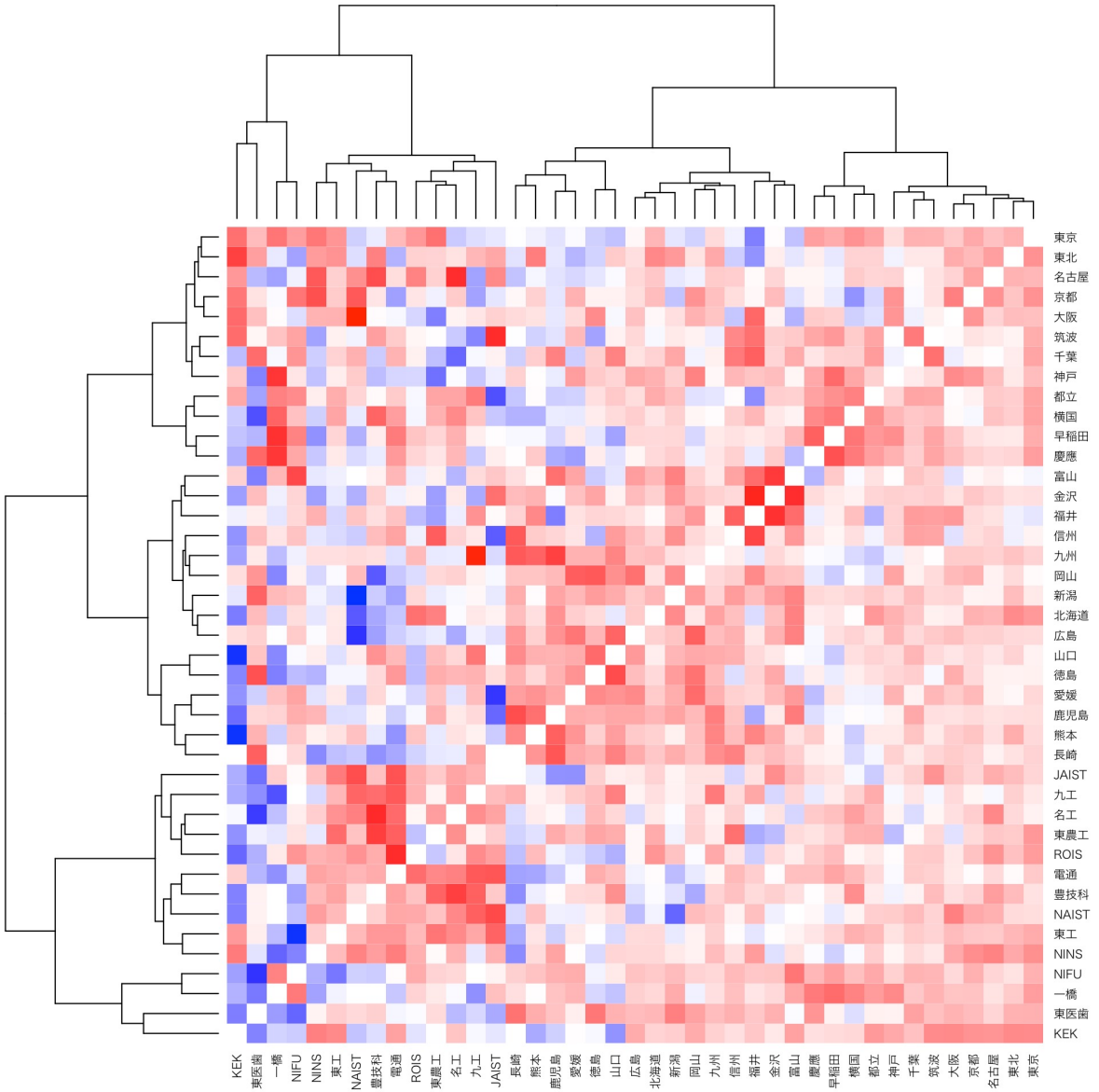
$$\mathrm{PMI}(x,y) = \log \frac{p(x,y)}{p(x)p(y)} = \log \frac{n(x,y)/N}{n(x)/N \cdot n(y)/N}$$

$$= \log \frac{n(x,y) \cdot N}{n(x)\,n(y)}$$

機関間のPMI行列 (一部)

Org	東京大学	京都大学	大阪大学	東北大学	名古屋大学	九州大学	北海道大学	筑波大学	広島大学	神戸大学	東京工業大学	慶應義塾大学	早稲田大学
東京大学	#NUM!	0.22	0.12	0.20	0.14	0.06	0.18	0.19	-0.04	0.05	0.30	0.26	0.22
京都大学	0.22	#NUM!	0.27	0.08	0.22	0.10	0.11	0.01	0.01	0.22	0.07	-0.04	-0.13
大阪大学	0.12	0.27	#NUM!	0.12	0.04	0.09	-0.10	-0.07	0.11	0.27	0.17	0.07	0.04
東北大学	0.20	0.08	0.12	#NUM!	0.16	0.13	0.23	-0.05	-0.03	-0.15	0.19	-0.07	-0.18
名古屋大学	0.14	0.22	0.04	0.16	#NUM!	0.07	0.09	-0.01	0.02	0.04	0.05	-0.09	-0.17
九州大学	0.06	0.10	0.09	0.13	0.07	#NUM!	0.16	-0.14	0.13	-0.21	0.00	-0.16	-0.18
北海道大学	0.18	0.11	-0.10	0.23	0.09	0.16	#NUM!	-0.11	0.02	0.11	-0.09	-0.09	-0.07
筑波大学	0.19	0.01	-0.07	-0.05	-0.01	-0.14	-0.11	#NUM!	0.05	-0.09	-0.03	0.21	0.27
広島大学	-0.04	0.01	0.11	-0.03	0.02	0.13	0.02	0.05	#NUM!	0.15	-0.10	-0.26	-0.06
神戸大学	0.05	0.22	0.27	-0.15	0.04	-0.21	0.11	-0.09	0.15	#NUM!	-0.17	0.14	0.43
東京工業大学	0.30	0.07	0.17	0.19	0.05	0.00	-0.09	-0.03	-0.10	-0.17	#NUM!	-0.19	-0.04
慶應義塾大学	0.26	-0.04	0.07	-0.07	-0.09	-0.16	-0.09	0.21	-0.26	0.14	-0.19	#NUM!	0.71
早稲田大学	0.22	-0.13	0.04	-0.18	-0.17	-0.18	-0.07	0.27	-0.06	0.43	-0.04	0.71	#NUM!
千葉大学	0.20	-0.09	-0.18	-0.05	-0.12	-0.06	0.03	0.37	-0.02	-0.16	0.03	-0.03	0.04
金沢大学	-0.05	0.11	-0.08	-0.06	0.03	0.04	0.05	0.07	0.19	0.05	-0.39	-0.02	-0.11
岡山大学	-0.22	0.22	0.17	-0.04	-0.17	-0.03	-0.08	-0.04	0.62	0.31	-0.29	-0.39	-0.57
東京都立大学	0.24	-0.18	-0.24	0.01	0.03	-0.40	0.19	0.22	-0.08	-0.03	0.18	0.25	0.36
新潟大学	-0.15	-0.24	-0.01	0.18	-0.22	0.00	0.23	-0.02	0.08	-0.03	-0.59	-0.02	-0.07
自然科学研究	0.40	0.41	0.16	0.17	0.50	0.00	-0.25	-0.40	-0.25	-0.52	0.01	-0.64	-1.03
熊本大学	-0.12	0.05	-0.02	0.25	-0.16	0.57	-0.05	-0.24	0.01	-0.24	-0.19	0.00	-0.37
情報・システ	0.29	-0.01	-0.27	0.01	0.34	0.18	0.33	-0.05	-0.39	-0.42	0.18	-0.07	0.06
東京医科歯科	0.17	0.00	0.11	0.16	-0.39	-0.04	0.07	-0.05	0.01	-0.81	-0.26	0.58	-0.80
信州大学	-0.11	0.01	-0.34	-0.26	0.06	-0.06	0.06	0.32	0.20	0.07	-0.38	-0.05	-0.13
長崎大学	-0.07	-0.06	0.01	-0.19	-0.35	0.62	0.06	-0.09	0.33	-0.05	-1.11	-0.31	-0.37
横浜国立大学	0.31	-0.37	-0.30	0.01	0.07	-0.31	-0.16	0.09	0.00	0.15	0.27	0.47	0.58
山口大学	-0.27	-0.04	-0.15	0.03	-0.14	0.42	-0.24	-0.04	0.59	-0.06	-0.71	-0.49	-0.93
徳島大学	-0.22	-0.04	0.31	-0.24	-0.22	0.21	-0.04	-0.40	0.17	-0.01	-0.41	-0.21	-0.49
人間文化研究	0.26	0.30	-0.24	-0.37	-0.18	-0.22	-0.05	0.23	0.05	-0.14	-2.45	0.20	0.50
東京農工大学	0.42	-0.11	-0.60	0.01	0.05	-0.24	0.26	-0.08	-0.33	-0.89	0.77	-0.17	-0.05
愛媛大学	-0.08	0.14	0.01	-0.33	-0.27	0.22	0.01	0.15	0.51	0.19	-0.22	-0.88	-0.18
鹿児島大学	-0.16	-0.14	-0.21	-0.18	-0.23	0.74	0.21	-0.18	0.35	-0.33	-0.80	-0.80	-0.57
一橋大学	0.40	-0.09	-0.11	-0.17	-0.48	-0.45	-0.34	0.12	-0.23	0.67	-0.98	0.87	1.04
電気通信大学	0.17	-0.34	-0.15	-0.33	0.00	-0.29	-0.35	0.09	-0.49	-0.43	0.48	0.28	0.54
富山大学	-0.27	-0.16	-0.40	-0.19	-0.09	-0.05	0.27	0.10	0.44	-0.21	-0.31	-0.71	0.15
高エネルギー	0.42	0.30	0.44	0.43	0.27	-0.57	-0.61	0.42	-0.05	-0.01	0.45	-0.78	-0.71
福井大学	-0.50	-0.05	0.45	-0.42	-0.11	-0.25	-0.28	0.41	-0.10	0.01	-0.61	-0.44	-0.19

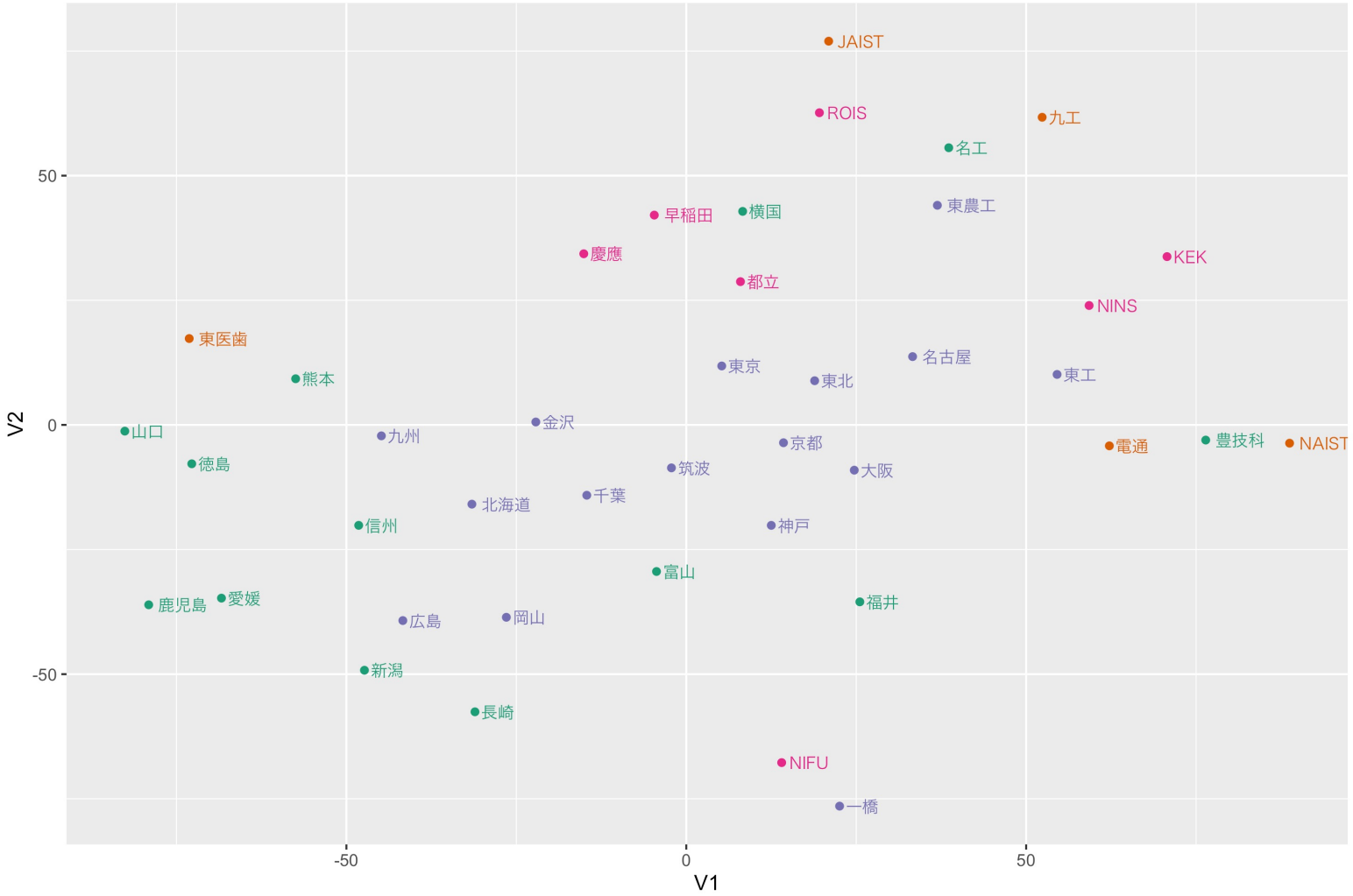
機関間のPMI行列のヒートマップ



- 大きくは
  1) 旧帝大
  2) 首都圏大学
  3) 地方国立
  4) 工学系研究大
  5) 分野特化型大
  にわかれる
  (人間の主観を使わずに)
- この時系列変化などが、今後の研究課題

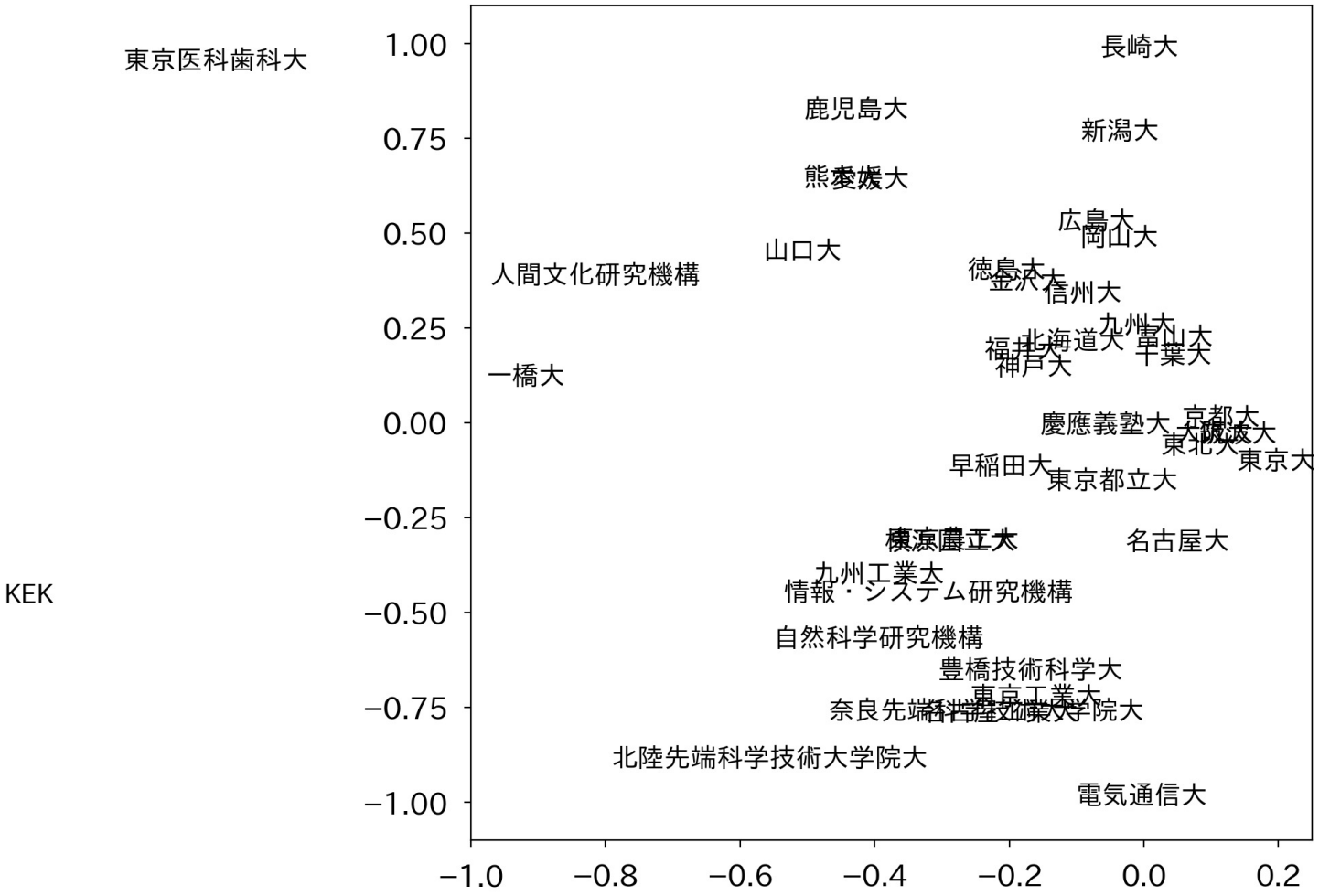
各機関のPMIベクトルの可視化

- t-SNEで2次元に可視化
  (共同研究関係が似た機関が似た位置)



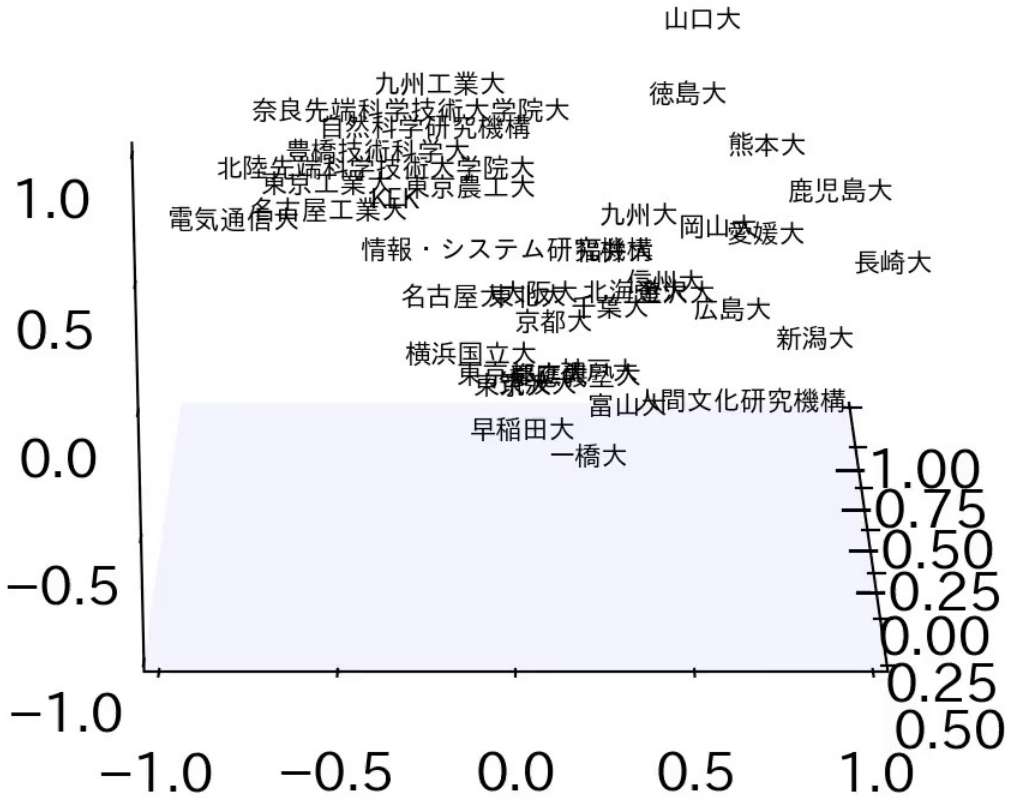
PMI行列のベクトル化

- 行列の各行はPMIなので、SVDで次元圧縮すれば Word2Vecと同じ「大学ベクトル」が得られる



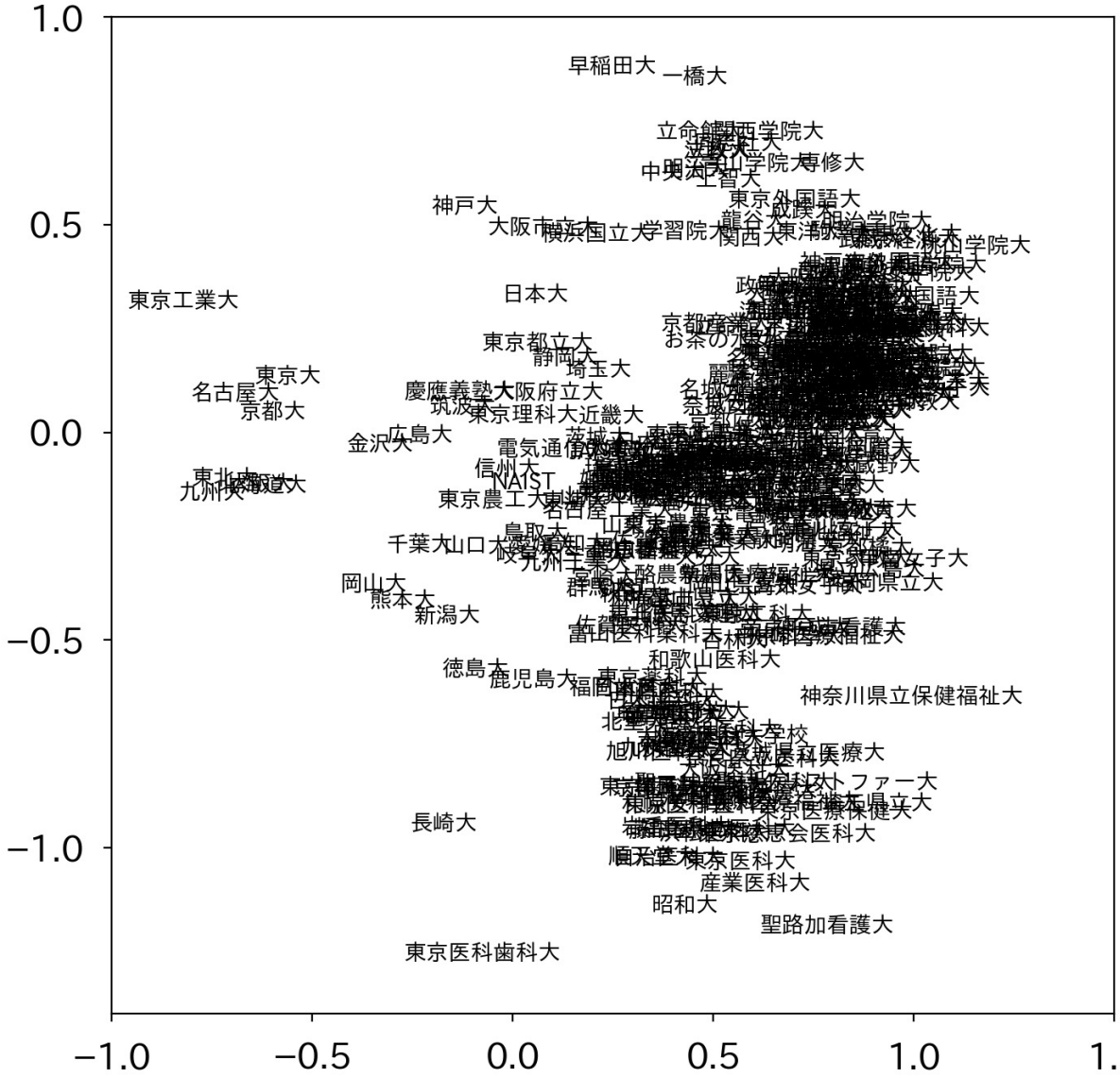
PMI行列のベクトル化

- 行列の各行はPMIなので、SVDで次元圧縮すれば Word2Vecと同じ「大学ベクトル」が得られる



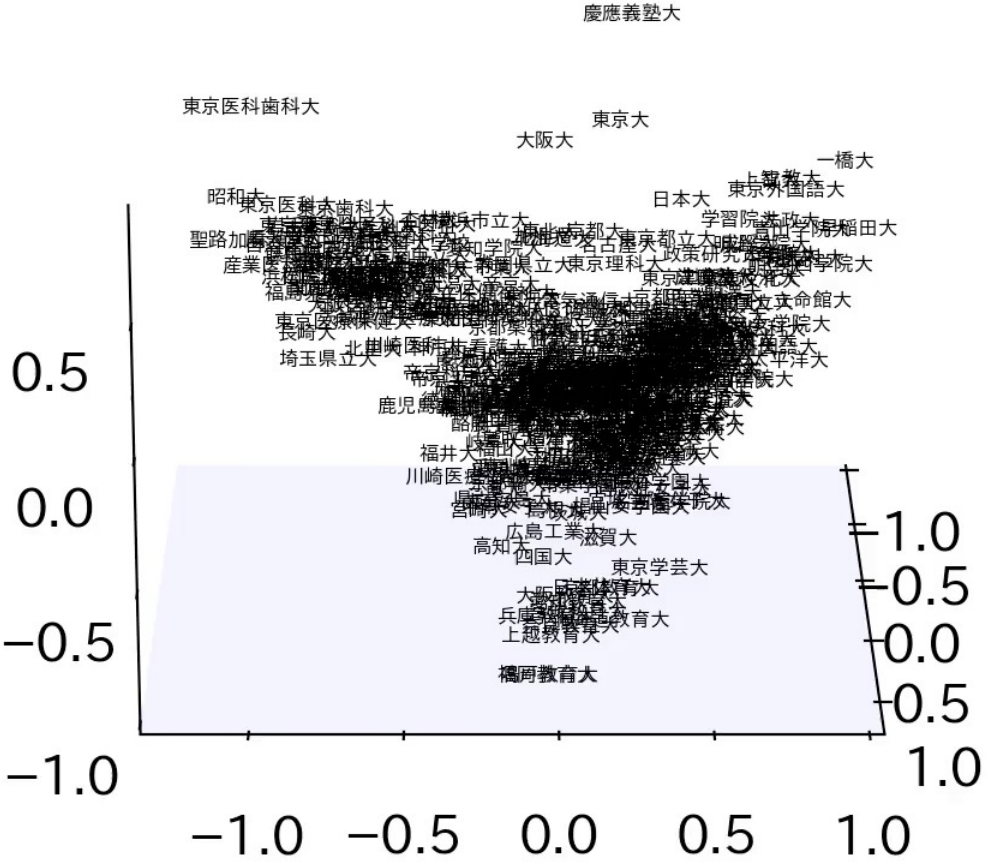
対象の大学を増やした場合

- 共同研究数の多い上位350大学ほど

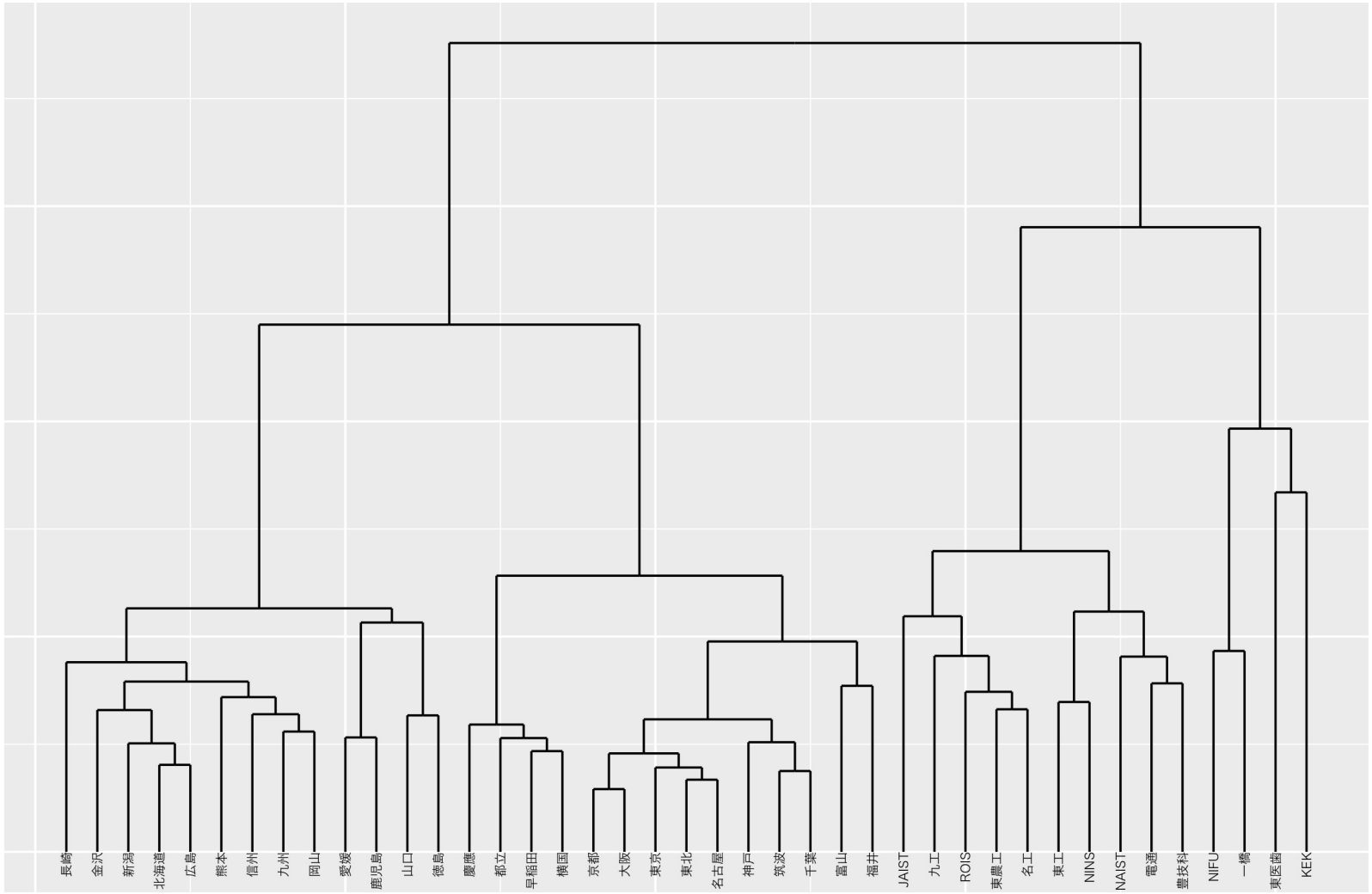


対象の大学を増やした場合

- 医学系、人文系、教育系の大きな3方向がみられる
  (“方向”があることが重要)



各機関のベクトルの階層クラスタリング



MIRAI-DX: 今後の課題

- 多数あるが、
  - 年ごとの、「大学の位置」の変化 (大学IR)
  - 医学、工学など各分野の中での位置付けの導出
  - 共同研究ネットワークにおけるハブの検出
  - 研究者検索の高度化
  - 産業界との連携：特定のテーマに専門性の高い研究者と共同研究を行うためのツール

結論

- 論文データや共同研究データなど、Science of Scienceの対象となるデータについて、適切な統計的取り扱いが重要
  - 頻度を適当に扱うと、結果が信頼できず、ノイズ等もうまく除くことができない
  - 本講演では、主に自己相互情報量(PMI)を使用して様々な分析を行った
- 今後の課題：
  積極的な統計モデル化
  - 本講演の範囲では記述統計学に近いが、複雑な関係をあぶり出すためには、統計モデルとして学習を行うことが必要 (機械学習の必要性)