METHOD FOR TAXONOMIC ANALYSIS OF TEXT COLLECTIONS

Boris Mirkin

1) International Center for Decision Choice and Analysis, NRU HSE Moscow
2) Faculty of Computer Sciences, NRU HSE Moscow RF
3) Birkbeck University of London UK
4) Supported by Academic Fund of NRU HSE Moscow (RSG 2019-2020, 2 mln ruble)

5)Joint work with T. Fenner (Birkbeck ULondon), S. Nascimento (New University Lisbon), D. Frolov, Zh. Airapetian, D. Babin, A. Vlasov, A. Guzharina, A. Denisenko, A. Sitnikov, A. Ushakova (HRU HSE Moscow RF)

CONTENTS

1. EXAMPLE OF TEXT COLLECTION: 17685 ABSTRACTS FROM 17 SPRINGER JOURNALS IN DATA SCIENCE (1998-2017)

1. Pattern Analysis and Applications (V. 1 /1998 – V. 20 /2017)

2. Journal of Classification (V. 15 /1998 – V. 34 /2017)

3. Annals of Mathematics & Artificial Intelligence (23/1998 - 80/2017)

4. Social Network Analysis and Mining (V.1/2011—V. 7/2017)

….

17. Machine Learning (V. 30/1998—V.106/2017)

CHALLENGE

- Given: 17685 abstracts from 17 Springer journals in Data Science (1998-2017)
- Wanted: Provide a brief description of main contents

- I know of 5 ways for doing that using computer:
  - 1. Content-analysis
  - 2. Summarization
  - 3. Co-citation and mutual citation graphs
  - 4. Topic modeling
  - 5. Taxonomic content-analysis (here)

METHOD

- Given: 17685 abstracts from 17 Springer journals in Data Science (1998-2017)
- Wanted: Provide a brief description of main contents

- **Taxonomic content-analysis** – proposed here
  - 1. Find (build) a taxonomy of the domain (Data science)
  - 2. Take the 317 taxonomy leaf concepts as units of the analysis
  - 3. Compute 17685×317 relevance matrix "text-to-leaf_concept"
  - 4. Find fuzzy clusters of leaf concepts
  - 5. Generalize a fuzzy cluster by optimally lifting it in the taxonomy tree to a head subject
  - 6. Conceptualize the result

TAXONOMY 1: DATA SCIENCE ITEMS IN ACM CCS

Subject index	Subject name
1.	Theory of computation
1.1.	Theory and algorithms for application domains
2.	Mathematics of computing
2.1.	Probability and statistics
3.	Information systems
3.1.	Data management systems
3.2.	Information systems applications
3.3.	World Wide Web
3.4.	Information retrieval
4.	Human-centered computing
4.1.	Visualization
5.	Computing methodologies
5.1.	Artificial intelligence
5.2.	Machine learning

DS TAXONOMY LEAF SUBJECTS

DS IN ACM COMPUTING CLASSIFICATION SYSTEM, LOWER RANKS

3.2.1.	Data mining
3.2.1.1.	Data cleaning
3.2.1.2.	Collaborative filtering
3.2.1.2.1**	Item-based
3.2.1.2.2**	Scalable
3.2.1.3.*	Association rules
3.2.1.3.1**	Types of association rules
3.2.1.3.2**	Interestingness
3.2.1.3.3**	Parallel computation
3.2.1.4.	Clustering
3.2.1.4.1**	Massive data clustering
3.2.1.4.2**	Consensus clustering
3.2.1.4.3**	Fuzzy clustering
3.2.1.4.4**	Additive clustering
3.2.1.4.5**	Feature weight clustering
3.2.1.4.6**	Conceptual clustering
3.2.1.4.7**	Biclustering
3.2.1.5.	Nearest-neighbor search

DATA SCIENCE TAXONOMY (FROLOV ET AL. 2018)

- Based on Classification of Computing Systems by ACM (456 items; 317 are lowest layer subjects (leaves)

https://www.hse.ru/mirror/pubs/share/213924179

5.1.2.			Knowledge representation and reasoning	
5.1.2.1.				Description logics
5.1.2.2.				Semantic networks
5.1.2.3.				Nonmonotonic, default reasoning and belief revision
5.1.2.4.				Probabilistic reasoning
5.1.2.5.				Vagueness and fuzzy logic
5.1.2.6.				Causal reasoning and diagnostics
5.1.2.7.				Temporal reasoning
5.1.2.8.				Cognitive robotics
5.1.2.9.				Ontology engineering
5.1.2.10.				Logic programming and answer set programming
5.1.2.11.				Spatial and physical reasoning
5.1.2.12.				Reasoning about belief and knowledge
5.1.3.			Computer vision	
5.1.3.1.				Computer vision problems

3. Leaf topic – to – text Relevance Matrix:

Annotated Suffix Tree (AST)

AST method (Pampapathi et al. 2006, Mirkin, Chernyak, 2014)

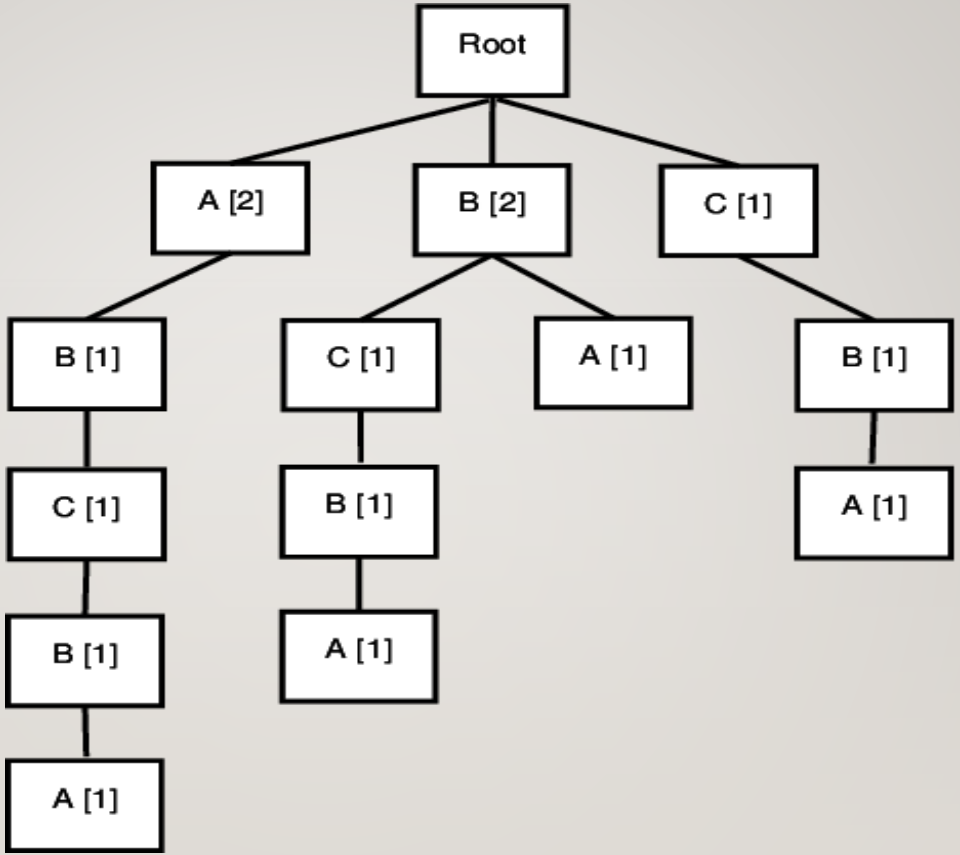
$\Rightarrow$ matrix R = (R_{vt}) 317x17685

Relevance index matrix: taxonomy_leaf_topic x text

AST: a way to keep text fragments (suffixes) and their frequencies

AST INSTANCE

String ABCBA



RELEVANCE INDEX: SUMMARY CONDITIONAL PROBABILITY OF NEXT SYMBOL IN MAXIMAL ALIGNED FRAGMENTS

Advantages of Annotated Suffix Trees

- **No pre-processing (lemmatization, stemming) needed**
- **Admits random errors in texts**

4. FINDING THEMATIC FUZZY CLUSTERS

13

- Convert rectangle topic-to-text relevance matrix R into square topic-to-topic co-relevance matrix C
- Apply Laplacian normalization B=L(C) to sharpen the cluster structure in C
- Fuzzy clustering in the space of eigen-vectors of B=L(C) [EigenMap (Belkin, Niyogi, 2003), FADDIS (Mirkin, Nascimento, 2012)]

CO-RELEVANCE MATRIX 317×317

- Given $T \times V$ $\boldsymbol{R=(r_{vt})}$, define $V \times V$ $\boldsymbol{C=(c_{vw})}$, $c_{vw} = \sum_{t=1}^{T} r_{vt}\, r_{wt}/n_t$

- n_t = *number of leaf subjects v such that* $r_{tv} > 0.2$
  (topics relevant to text v)

# Texts	n_t = # Relevant subjects
1237	0 [attention to be given]
2353	1
7114	2 – 4
6124	5 – 11
857	12 or more

FADDIS METHOD (2012, IN HOUSE): ONE CLUSTER AT A TIME

15

- $\mathbf{Min}_{\mathbf{u},\xi} \Sigma_{t,t' \in T} (b_{tt'} - \xi u_t u_{t'})^2$

- Equivalent to maximum of Rayleigh quotient (**max** eigenvalue)

  Max $\mathbf{uBu^T/(u^Tu)}$

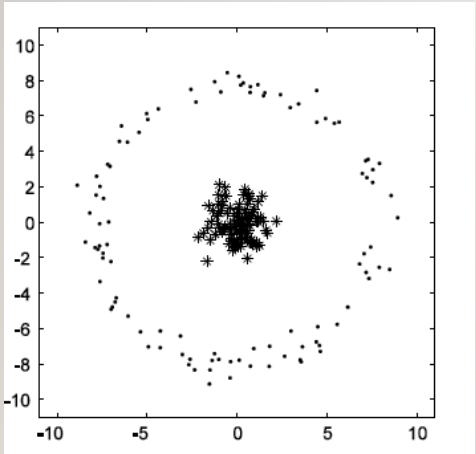
- **Spectral approach (Shi, Malik, 2000):** find **min** eigenvalue and its vector, adjust the latter to fuzzy membership
- To make consistent [max], apply **pseudo-inverse** transformation to B
- Found 6 fuzzy clusters of which 3 are more or less homogeneous: **L- Machine Learning, C- Clustering, and I – Information Retrieval**

LAPLACIAN PSEUDO-INVERSE (LAPIN):

- **Given B, convert into L^+**

$$D=\mathrm{diag}(B*I_N)$$

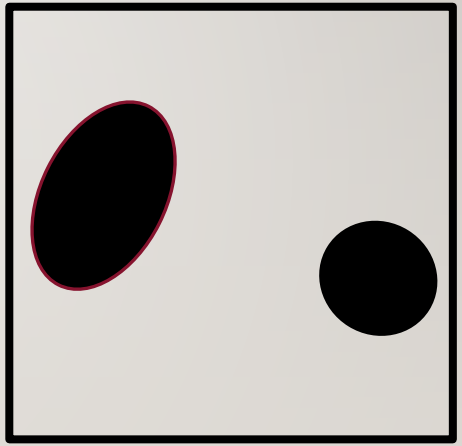
B



$$L = I - D^{-1/2}BD^{-1/2}$$

$$L^+ = \mathrm{pinv}(L)$$

L^+



Cluster L: Learning

Membership	Code	Topic
0.299	5.2.3.8.	rule learning
0.282	5.2.2.1.	batch learning
0.276	5.2.1.1.2.	learning to rank
0.217	1.1.1.11.	query learning
0.216	5.2.1.3.3.	apprenticeship learning
0.213	1.1.1.10.	models of learning
0.203	5.2.1.3.5.	adversarial learning
0.202	1.1.1.14.	active learning
0.191	5.2.1.4.1.	transfer learning
0.191	5.2.1.4.2.	lifelong machine learning
0.188	1.1.1.8.	online learning theory
0.165	5.2.2.2.	online learning settings
0.158	1.1.1.3.	unsupervised learning and clustering
0.141	5.2.2.6.	active learning settings
0.136	5.2.1.1.4.	supervised learning by regression
0.128	5.2.2.5.	learning from implicit feedback

Cluster C: Clustering

Membership	Code	Topic
0.327	3.2.1.4.7	biclustering
0.286	3.2.1.4.3	fuzzy clustering
0.248	3.2.1.4.2	consensus clustering
0.220	3.2.1.4.6	conceptual clustering
0.192	5.2.4.3.1	spectral clustering
0.187	3.2.1.4.1	massive data clustering
0.159	3.2.1.7.3	graph based conceptual clustering
0.151	3.2.1.9.2.	trajectory clustering
0.148	3.1.3.7.	database views
0.143	5.1.1.9.	language resources
0.141	3.4.4.3.	language models
0.138	3.2.1.4.4	additive clustering
0.136	3.2.1.4.5	feature weight clustering
0.136	3.4.5.8.	clustering and classification
0.135	3.1.3.12.	stream management
0.131	3.4.7.2.4.	music retrieval

CLUSTER R "RETRIEVAL": $U_I \geq 0.15$

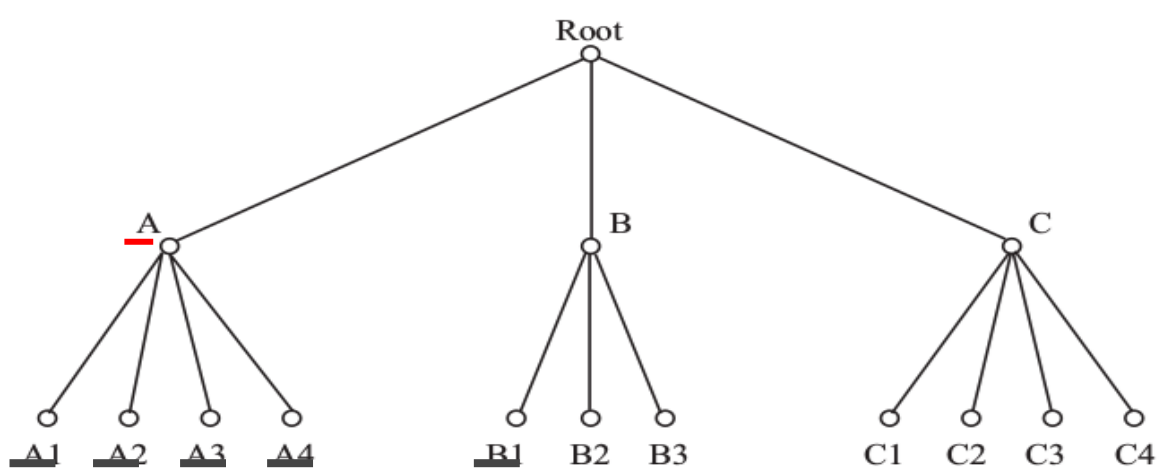
18

0.211 & 3.4.2.1. & query representation
0.207 & 5.1.3.2.1. & image representations
0.194 & 5.1.3.2.2. & shape representations
0.194 & 5.2.3.6.2.1 & tensor representation
0.191 & 5.2.3.3.3.2 & fuzzy representation
0.187 & 3.1.1.5.3. & data provenance
0.173 & 2.1.1.5. & equational models
0.173 & 3.4.6.5. & presentation of retrieval results
0.165 & 5.1.3.1.3. & video segmentation
0.155 & 5.1.3.1.2. & image segmentation

“TO **GENERALIZE**” ACCORDING TO MERRIAM–WEBSTER (USA)

- A meaning:
  - ``to give a general form to''
  - ``to derive or induce (a general conception or principle) from particulars''

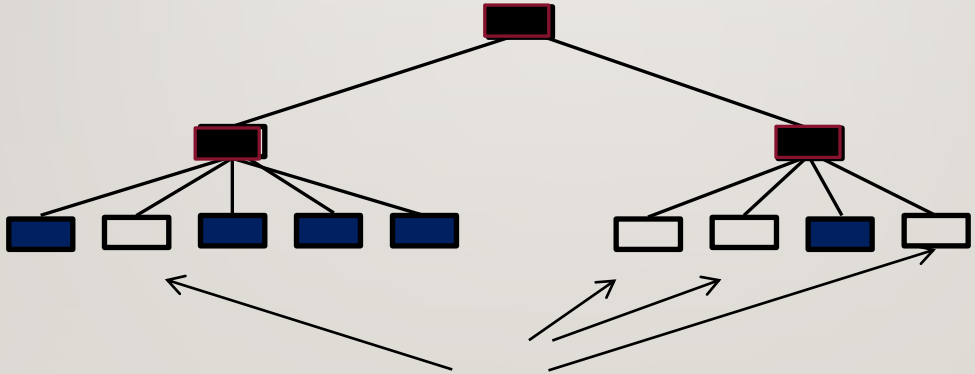
GENERALIZATION: APPROPRIATELY LIFTING CLUSTERS TO COMMON ROOT CONCEPTS



Given a taxonomy and a crisp leaf cluster, lift the leaves to a higher rank node: **(A1, A2, A3, A4, B1) => (A)**, B1 disregarded as an **offshoot**.

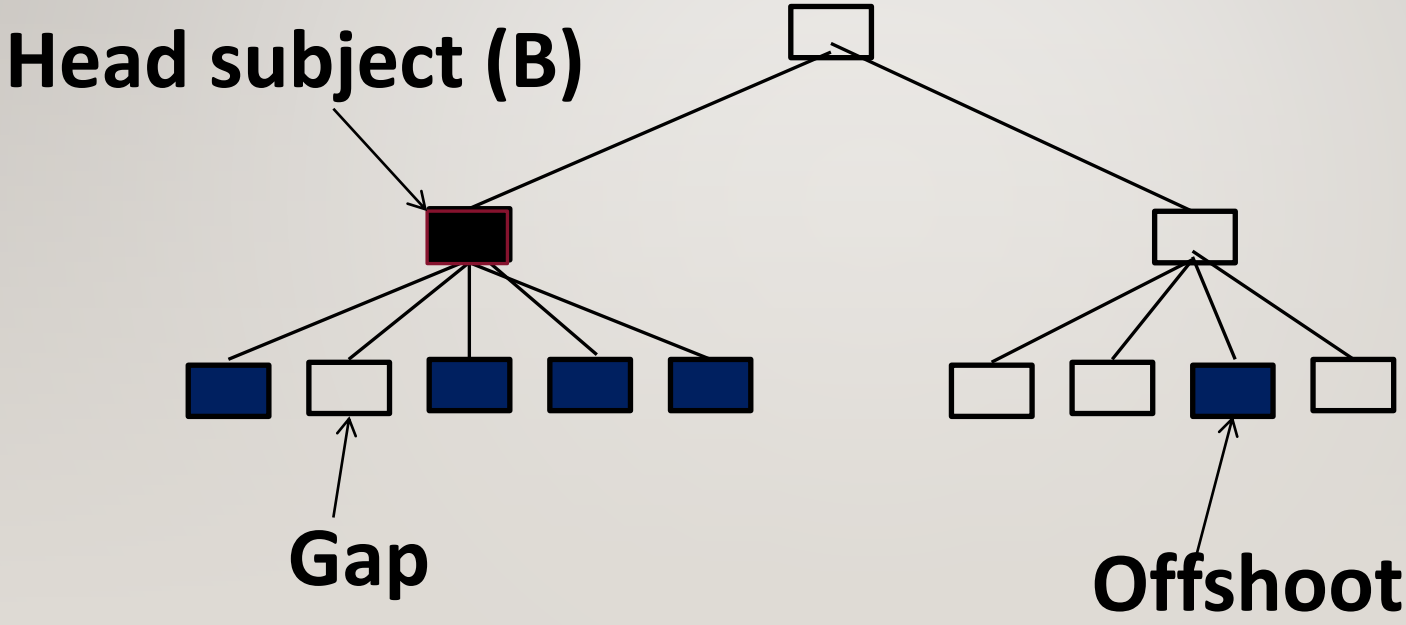
GENERALIZE: GIVEN 5 LEAVES IN A CLUSTER, WHERE TO LIFT THAT? OPTION A

Head subject (A)



Gap

GENERALIZE: GIVEN 5 LEAVES IN A CLUSTER, WHERE TO LIFT THAT? OPTION B



MINIMIZE THE PENALTY!

Penalty:

#Head_Subject + λ#Gap + γ#Offshoot

Penalty at option A: $1+4\lambda$

Penalty at option B: $1+\gamma+\lambda$

ALGORITHM **PARGENFS:**

Parsimonious generalization

Output: set of head subjects H, minimizing

$$p(H) = \sum_{h \in H - I} u(h) + \sum_{h \in H - I} \sum_{g \in G(h)} \lambda v(g) + \sum_{h \in H \cap I} \gamma u(h)$$

Penalties: λ – for a gap, γ for an offshoot, 1 for a head subject

I **– leaf set of the taxonomy rooted tree,**

u(h) **– query fuzzy set membership function**

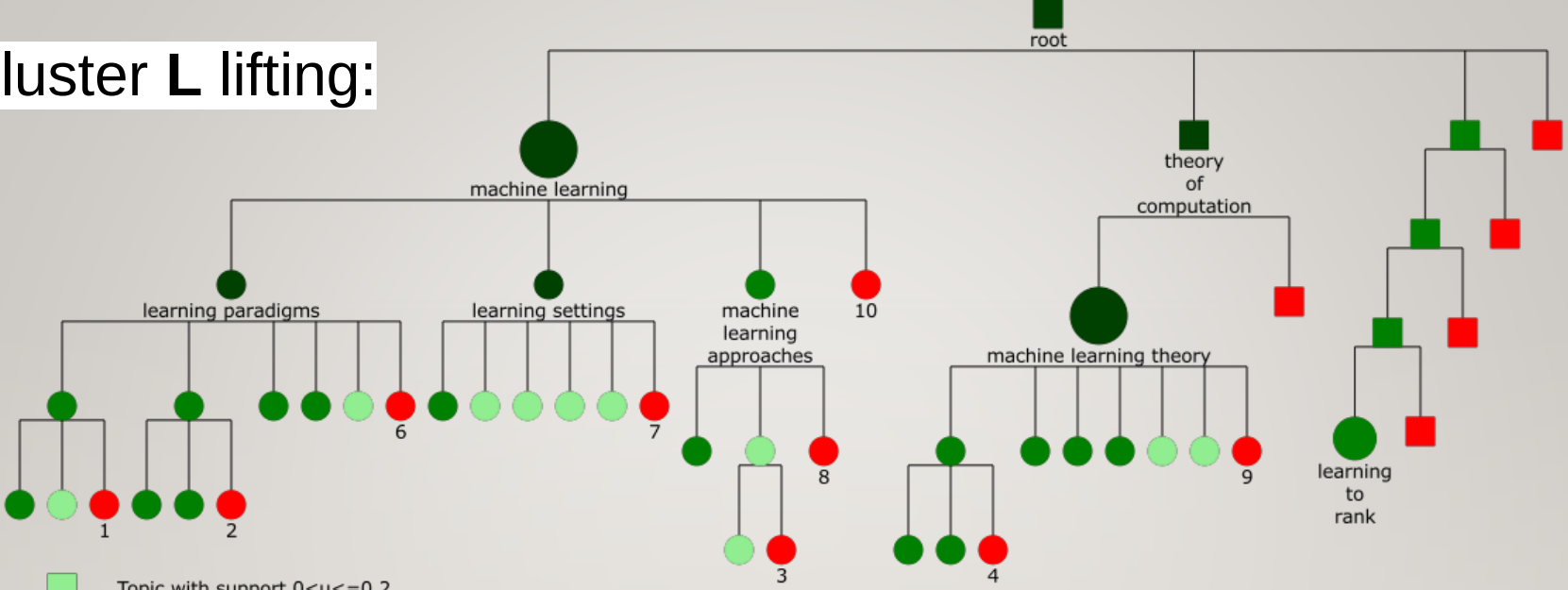
TCAN SOFTWARE

- GOT package (under renovation)
- Includes
  - Relevance and co-relevance matrices
  - FADDIS clustering (including LAPIN)
  - Parsimonious lifting
  - Visualization of taxonomy and lifting results
- Site URL: https://github.com/dmitsf/GOT
- Technical documentation: https://got-docs.readthedocs.io/

APPLYING GOT software TO the abstracts sample

- Lifting parameters (according to structure of DST)
- gap penalty: **λ=0.1**,
- offshoot penalty: **γ=0.9**
- 3 out of 6 clusters are interpretable (learning L, retrieval R, clustering C)
- Each of L, R, and C clusters is lifted with ParGenFS

Cluster **L** lifting:



Head subjects: {Machine learning,

Machine learning theory, Learning to rank}

TENDENCIES OF DATA SCIENCE RESEARCH (PARTLY)

- A page long description

  according to TCAN

TENDENCIES OF DATA SCIENCE RESEARCH (PARTLY)

- Three clusters out of six:
  - **Learning**
  - **Information retrieval**
  - **Clustering**

CLUSTER LIFTING RESULTS SUGGEST, 1:

- "Learning" lifted: **Conceptualization**
  - main work still on theory and method rather than applications.
  - Expanding from learning subsets and partitions towards learning of ranks and rankings.
  - Many subareas are not covered by publications.

CLUSTER LIFTING RESULTS SUGGEST 2:

- "Information retrieval" lifted: **Conceptualization**
- Head subjects:

  (a) Information Systems, (b) Computer Vision
  - Text management
  - Moving from text to embrace images and video.
  - Ways for structuring visual information probably leading to a **future "wordnet" for images**

CLUSTER LIFTING RESULTS SUGGEST 3:

- "Clustering" cluster C lifted: **Conceptualization**
  - 16 (!) head subjects, to be raised to higher ranks in Taxonomy of Data Science
  - Should be lifted in the taxonomy from auxiliary roles to a main concept, and instrument, in knowledge engineering.

COLLECTION 2: ABSTRACTS FROM SPRINGER И ELSEVIER, GOOGLE OUTPUTS TO QUERIES}

- **Queries:** clustering, machine learning, neural networks, algorithm, classification, information retrieval, natural language processing, software, computing, pattern recognition, deep learning, probabilistic, artificial intelligence, support vector, Bayesian, regression, search engine
- **Collection 2: 26 799** abstracts from 80 journals 1971-2019

OTHER APPROACHES TO THE CHALLENGE

- Given: 17685 abstracts from 17 Springer journals in Data Science (1998-2017)
- Wanted: Provide a brief description of main contents

- Other approaches:
  - 1. **Conventional content-analysis**
  - 2. **Summarization**
  - 3. **Co-citation and mutual citation graphs**
  - 4. **Topic modeling**

APPROACH 1: 1. CONVENTIONAL CONTENT-ANALYSIS

- Given: 17685 abstracts from 17 Springer journals in Data Science (1998-2017)
- Wanted: Provide a brief description of main contents

- (http://www.audiencedialogue.net/kya16a.html)

Content analysis is a method for summarizing any form of content by counting various aspects of the content, like user-specified words or concepts.

What for? For comparisons:

“27% of programs on Radio XXX in April 2017 mentioned at least one aspect of peacebuilding, compared with only 3% of the programs in 2010 [or with only 3% of the programs on Radio YYY]."

APPROACH 2: SUMMARIZATION

- Given: 17685 abstracts from 17 Springer journals in Data Science (1998-2017)
- Wanted: Provide a brief description of main contents

- **2. Summarization**
  - **Extractive summarization**: Automatic selection of "key" sentences from text
  - **Abstractive summarization**: Deep learning using Recurrent NN and Convolutional NN for text embedding in vector spaces – seems a very promising direction for the future.

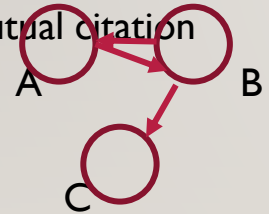
APPROACH 3: CITATION AND CO-CITATION GRAPHS

- Given: 17685 abstracts from 17 Springer journals in Data Science (1998-2017)
- Wanted: Provide a brief description of main contents

- 2. Graph of co-citation or mutual citation between papers or authors:
- papers **A, B, C**
- **A** List of references:

1.B, 2. X, 3.Y, 4. Z, 5. D

Mutual citation



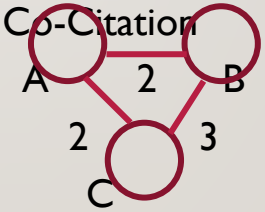
B list of references

1.A, 2. E, 3.Y, 4. Z, 5. F

C list of references

1.B, 2.Y, 3.F, 4. E

Co-Citation



APPROACH 3: EXAMPLE

- Co-citation or citation graph between papers or authors: cluster analysis

- Example: Cluster **“Information retrieval”** Chen, Ibekwe-SanJuan, Hou (2010):

  - prominent members of a cluster as the intellectual base (books by G. Salton and C. Van Rijsbergen, and a paper by S. Robertson)

  - themes identified in the citers of the cluster as research fronts (``information retrieval", ``probabilistic model", ``query expansion", ``using heterogeneous thesauri“)

ПОДХОД 4: ТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ

- Given: 17685 abstracts from 17 Springer journals in Data Science (1998-2017)
- Wanted: Provide a brief description of main contents

- Data: Probability(word/text)
- Model: Matrix Factorization

$$\Pr(\text{word/text}) = \sum_{topic} \Pr(\text{word/topic}) * \Pr(\text{topic/text})$$

4. EXAMPLE: TOPIC MODELING (MUCH POPULAR), 1

- 1 Information retrieval?

0.018*"software", 0.018*"inform", 0.013*"query", 0.012*"retrieve",
0.012*"study", 0.011*"develop", 0.011*"product", 0.010*"document",
0.010*"user", 0.009*"engine", 0.009*"research", 0.008*"model",
0.008*"approach", 0.008*"search", 0.008*"busy", 0.008*"knowledge",
0.007*"manage", 0.007*"service", 0.006*"semantic", 0.006*"provide"

- 2 Text and images?

0.020*"image", 0.012*"language", 0.010*"model", 0.010*"retrieve",
0.010*"feature", 0.009*"propose", 0.009*"method", 0.009*"approach",
0.008*"inform", 0.008*"recognition", 0.008*"paper", 0.007*"process",
0.007*"network", 0.007*"base", 0.006*"present", 0.006*"result",
0.006*"differ", 0.006*"system"

4. TOPIC MODELING (MUCH POPULAR),2

3 Classifiers?

0.018*"classify", 0. 016*"feature", 0.013*"method", 0.011*"classification", 0.011*"result", 0.010*"data", 0.009*"perform", 0.009*"accuracy", 0.009*"propose", 0.009*"model", 0.009*"recognition", 0.008*"base", 0.007*"image", 0.007*"study", 0.007*"differ", 0.006*"extract", 0.006*"predict", 0.006*"pattern", 0.006*"inform"

- 4 Clusters in networks?

0.013*"algorithm", 0.012*"propose", 0.012*"cluster", 0.010*"graph", 0.009*"method", 0.009*"base", 0.008*"paper", 0.008*"result", 0.008*"inform", 0.007*"data", 0.006*"function", 0.006*"network", 0.006*"model"

SAME APPROACH TO OTHER COLLECTIONS

- «Visit to a restaurant» (18036 user reviews of restaurants and cafee in Moscow, in Russian, TripAdvisor, 2019, 267 leaf subjects)
- «Car» (35785 user reviews of cars)
- «Research journal contents» (abstracts of all 461 papers "Journal of Classification" 1984-2019, leaf subjects 106)

- **Results found, but nothing sensational; probably, our in-house taxonomies lack substance**

CONCLUSION

- TCAN explicitly involves the contents and structure of a taxonomy of the domain
- There is an original component: Parsimonious lifting as a model of generalization
- **TCAN's use much depends on the usage of taxonomies in the development of specific domains**
- Future work: use of the maximum likelihood criterion in the problem of optimal lifting
- Future work: Use of the optimal lifting in other applications (reconstruction of the history of individual genes in phylogenetic trees, optimization of the targeted advertising over Internet).

PUBLICATIONS

44

- D. Frolov, B. Mirkin, S. Nascimento, T. Fenner (2018) Finding an appropriate generalization for a fuzzy thematic set in taxonomy, preprint HSE, https://wp.hse.ru/data/2019/01/13/1146987922/WP7_2018_04_____.pdf

- Frolov, D., Nascimento, S., Fenner, T., & Mirkin, B. (2020). Parsimonious generalization of fuzzy thematic sets in taxonomies applied to the analysis of tendencies of research in data science. *Information Sciences*, *512*, 595-615.

SOME PUBLICATIONS

- Ekaterina Chernyak and Boris Mirkin (2014) An AST method for scoring string-to-text similarity in semantic text analysis, In "Clusters, Orders, and Trees: Methods and Applications" pp 331-340 (Springer, SOIA, v. 92, 2014)
- Mirkin, B., & Nascimento, S. (2012). Additive spectral method for fuzzy cluster analysis of similarity data including community structure and affinity matrices. *Information Sciences*, *183* (1), 16-34.
- Mirkin, B. G., Nascimento, S., Fenner, T. I., & Pereira, L. M. (2010). Building Fuzzy Thematic Clusters and Mapping Them to Higher Ranks in a Taxonomy. *Int. J. Software and Informatics*, *4*(3), 257-275.