МЕТОД ТАКСОНОМИЧЕСКОГО КОНТЕНТ-АНАЛИЗА ТЕКСТОВЫХ КОЛЛЕКЦИЙ

Б.Г. Миркин, Д.С. Фролов

[1] Международный Центр Анализа и Выбора Решений, НИУ ВШЭ

[2] Факультет компьютерных наук НИУ ВШЭ (участники: T. Fenner (Лондон), S. Nascimento (Лиссабон) и студенты ФКН НИУ ВШЭ Ж. Айрапетян, Д. Бабин, А. Власов, А. Гужарина, А. Денисенко, А. Ситников, А. Ушакова)

[3] Поддержано Научным Фондом НИУ ВШЭ 2019-2020 (НУГ 2 млн)

СОДЕРЖАНИЕ

1. Подходы к обобщенному представлению коллекций текстов
2. Этапы таксономического контент-анализа на примере коллекции статей по науке данных:

   2.1 Таксономия науки данных и ее листья

   2.2 Формирование матрицы релевантности лист-текст

   2.3 Отыскание нечетких кластеров листовых тем

   2.4 Обобщение нечетких кластеров как оптимальный подъем в таксономии

3. Анализ двух коллекций аннотаций, 17000 и 26000
4. Заключение: что сделано и что надо бы

ПРИМЕР ТЕКСТОВОЙ КОЛЛЕКЦИИ: 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)

ПРОБЛЕМА

- **Дано: 17685** abstracts from 17 Springer journals in Data Science (1998-2017)
- Надо: Кратко описать, на страницу, о чем они, что в них главное

- КАК? Я знаю о 5 направлениях:
  - 1. **Content-analysis** (Контент-анализ)
  - 2. **Summarization** (Реферирование)
  - 3. **Co-citation and mutual citation graphs** (Анализ графа ко-цитирования/взаимного цитирования)
  - 4. **Topic modeling** (Тематическое моделирование)
  - 5. **Taxonomic content-analysis** (Таксономический контент-анализ) **-новое**

ПРИМЕР АННОТАЦИИ:

KOVALEVA, MIRKIN, JOURNAL OF CLASSIFICATION, 2015, 32(3), 414-442.

5

- **Abstract:** The paper presents a least squares framework for divisive clustering. Two popular divisive clustering methods, Bisecting K-Means and Principal Direction Division, appear to be versions of the same least squares approach. The PDD recently has been enhanced with a stopping criterion taking into account the minima of the corresponding one-dimensional density function (dePDDP method). We extend this approach to Bisecting K-Means by projecting the data onto random directions and compare thus modified methods. It appears the dePDDP method is superior at datasets with relatively small numbers of clusters, whatever cluster intermix, whereas our version of Bisecting K-Means is superior at greater cluster numbers with noise entities added to the cluster structure.

ПРОБЛЕМА 1

- **Дано: 17685** abstracts from 17 Springer journals in Data Science (1998-2017)
- Надо: Кратко описать, на страницу, о чем они, что в них главное

- 1. Контент-Анализ (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]."

ПРОБЛЕМА 2

- **Дано: 17685** abstracts from 17 Springer journals in Data Science (1998-2017)
- Надо: Кратко описать, на страницу, о чем они, что в них главное

- **2. Summarization (Реферирование):**
  - **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 – по-моему, очень перспективное направление

ПРОБЛЕМА 3

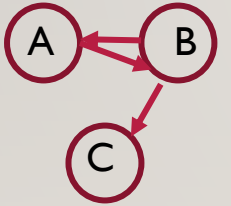
- **Дано: 17685 abstracts from 17 Springer journals in Data Science (1998-2017)**
- **Надо: Кратко описать, на страницу, о чем они, что в них главное**

- 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

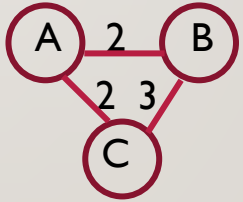
  B list of references
  1.A, 2. E, 3.Y, 4. Z, 5. F

  C list of references
  1.B, 2.Y, 3.F, 4. E

Mutual citation



Co-Citation



ПРОБЛЕМА 3

- **Дано: 17685 abstracts from 17 Springer journals in Data Science (1998-2017). Надо: Кратко описать, о чем они, что в них главное**

- 3. 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 basis (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”)

ПРОБЛЕМА 4

- **Дано: 17685 abstracts from 17 Springer journals in Data Science (1998-2017). Надо: Кратко описать, о чем они, что в них главное**

- 4. Topic modeling (much popular)

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

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

4. TOPIC MODELING (MUCH POPULAR), I

- Дано: 17685 abstracts from 17 Springer journals in Data Science (1998-2017). Надо: Кратко описать, о чем они, что в них главное
- 1 Информационный поиск?

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 Текст и изображения?

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

- **Дано: 17685 abstracts from 17 Springer journals in Data Science (1998-2017). Надо: Кратко описать, о чем они, что в них главное**
- 3 Методы классификации?

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 Кластеры на сетях?

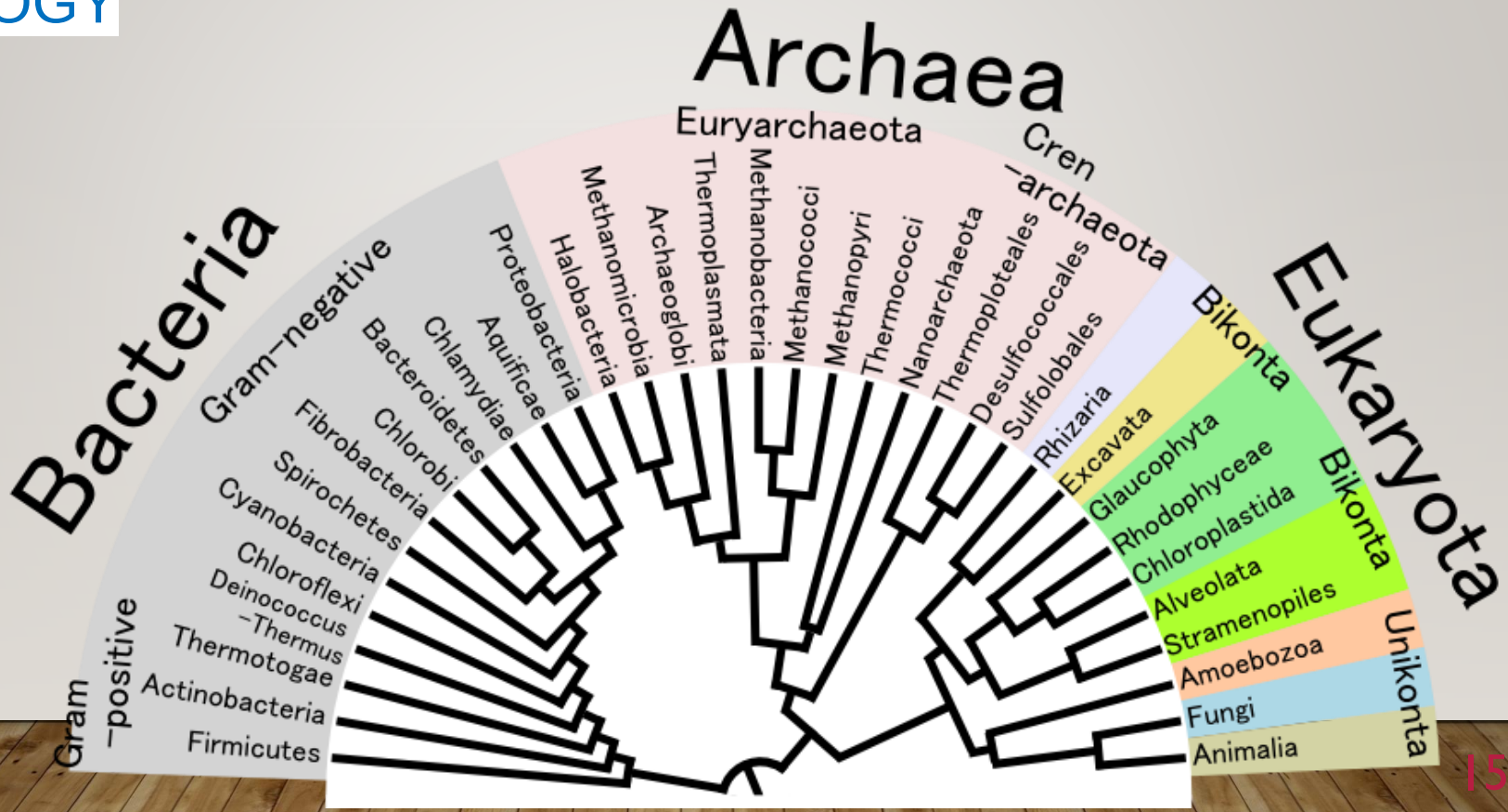
0.013*"algorithm", 0.012*"propose", 0.012*"cluster", 0.010*"graph", 0.009*"method", 0.009*"base", 0.008*"paper", 0.008*"result", 0.008*"inform", 0.007*"data", 0.006*"function", 0.006*"network", 0.006*"model"

ПРОБЛЕМА 5

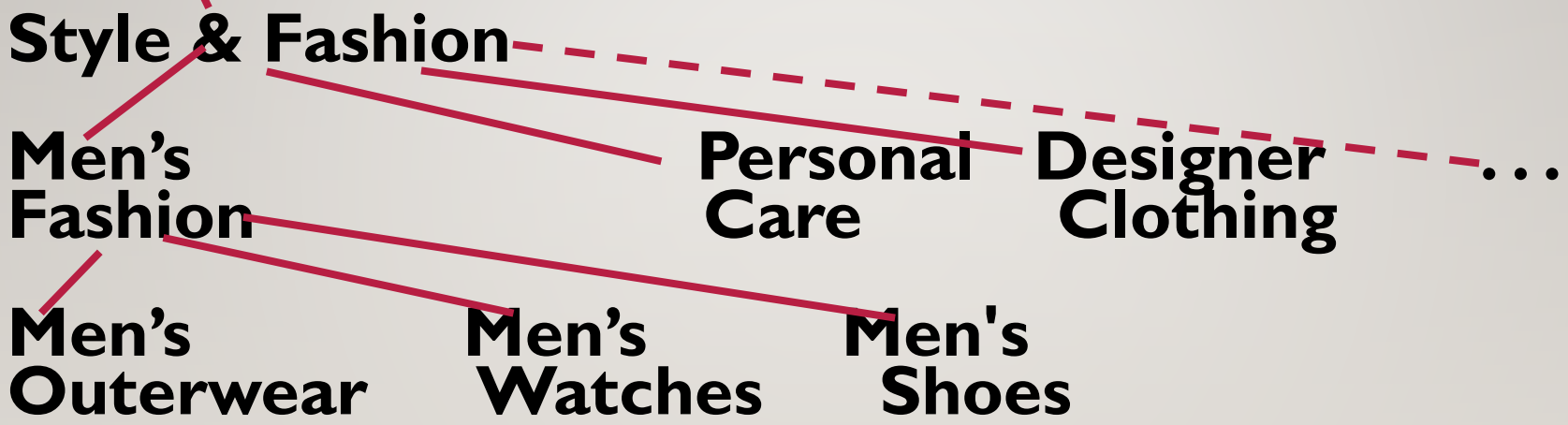
- **Дано: 17685** abstracts from 17 Springer journals in Data Science (1998-2017)
- Надо: Кратко описать, о чем они
- 5. **Taxonomic content-analysis** (Таксономический контент-анализ) – предлагаемый метод
  - 1. Find (build) a taxonomy of the domain
  - 2. Take taxonomy leaf concepts as units of the analysis
  - 3. Compute 17685×317 relevance matrix "text-to-leafconcept"
  - 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. Interpret the result

1. TAXONOMY FINDING/MAKING

TAXONOMY 1: A TREE-LIKE ARRANGEMENT OF OBJECTS OVER AN EXTENT OF SIMILARITY. EXAMPLE 1: BIOLOGY



TAXONOMY 2: IAB (THE INTERACTIVE ADVERTISING BUREAU) CONTENT TAXONOMY FRAGMENT HTTPS://WWW.IAB.COM/



TAXONOMY 3: DATA SCIENCE ITEMS IN ACM CCS 2012

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

2. TAXONOMY LEAF SUBJECTS

DATA SCIENCE IN ACM CCS 2012, 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 2012 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. Матрица релевантности между текстами и листовыми словосочетаниями:

Аннотированное суффиксное дерево (АСД)

AST method (Mirkin, Artemova-Chernyak, 2014)

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

Relevance matrix: taxonomy_leaf_topic x text

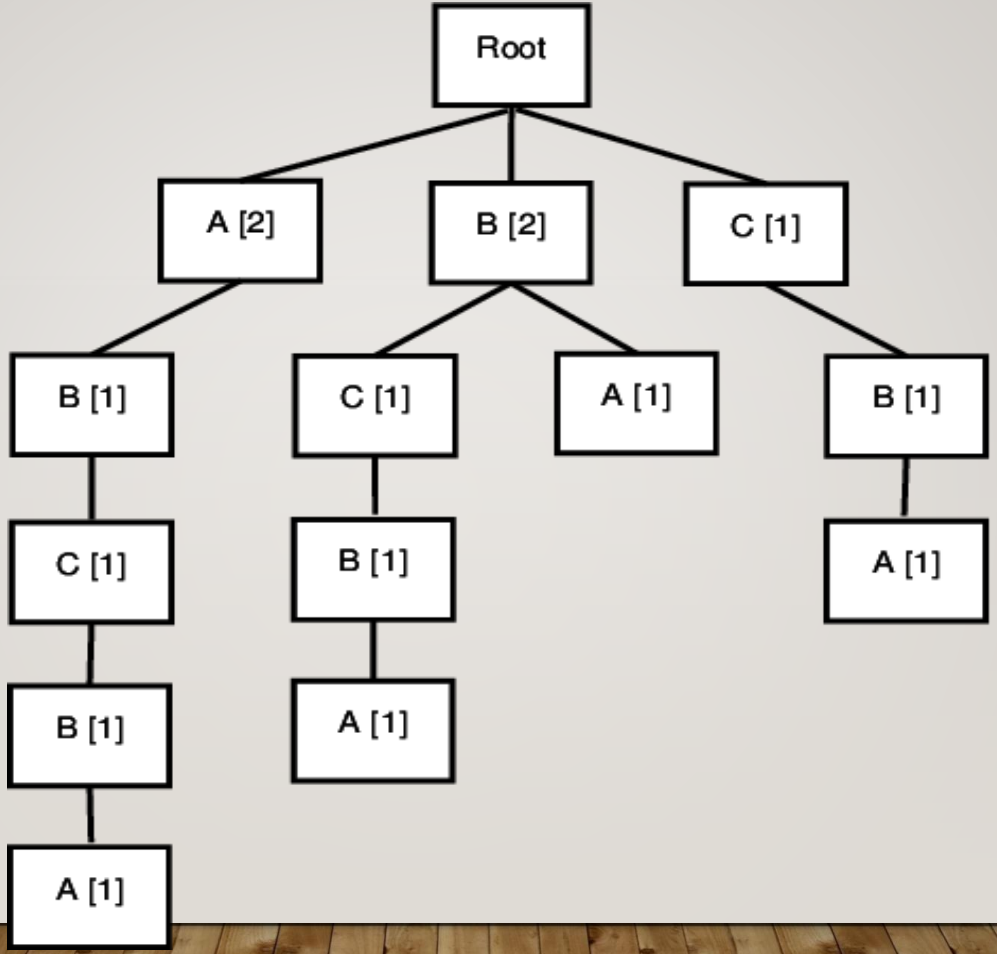
АСД способ хранить фрагменты текста (суффиксы) вместе с их частотами

4. ВЫЧИСЛЕНИЕ РЕЛЕВАНТНОСТИ ТЕМА-ДОКУМЕНТ

- **Метод определения степени вхождения строки в текст на основе аннотированного суффиксного дерева (АСД) (Б. Миркин, Е. Черняк, 2012)**
- **АСД позволяет хранить фрагменты текста (суффиксы) вместе с их частотами**

ПРИМЕР АСД

Строка ABCBA



ВЫЧИСЛЕНИЕ РЕЛЕВАНТНОСТИ СТРОКИ ТЕКСТУ ПО СОВПАДЕНИЮ U

Условная вероятность **p(u)** узла **u** (**f(u)** - его частота):

$$p(u) = \frac{f(u)}{\sum_{v \in T: ancestor(v) = ancestor(u)} f(v)}$$

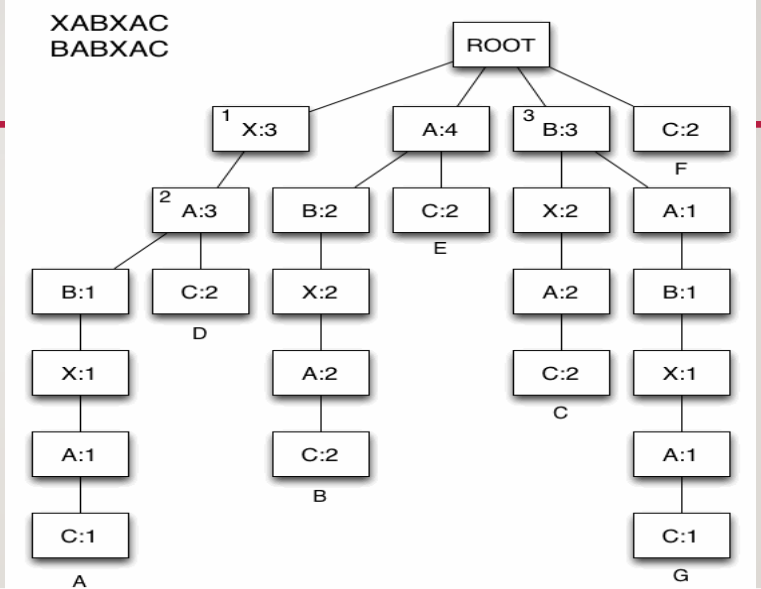
Для строки $\mathbf{x = x_1 \ldots x_m}$ степень вхождения **S(x, T)** в АСД **T**:

$$s(x_k, T) = \frac{1}{k_{max}} \sum_{i=0}^{k_{max}} p(x_i^k)$$

k_{max} - длина наибольшего совпадения суффикса с АСД

TOOL: A **UNIFORM PHRASE-TEXT SIMILARITY MEASURE** ACCORDING TO MATCHES IN SUFFIX TREE ANNOTATED BY FREQUENCIES

Suffix tree for strings
XABXAC and
BABXAC
annotated with substring frequencies, and
the **similarity score** for string **VXACA**



Suffix	Match	Score
'VXACA'	None	0
'XACA'	'X'->'A'->'C'	3/12 + 3/3 + 2/3=1 11/12
'ACA'	'A'->'C'	4/12 + 2/4=5/6
'CA'	'C'	2/12
'A'	A'	4/12

ВЫЧИСЛЕНИЕ РЕЛЕВАНТНОСТИ СТРОКИ ТЕКСТУ

Преимущества метода АСД:

- **Позволяет избежать предобработки текстов (лемматизация, стеммирование)**
- **Допускает тексты с искаженными признаками (ошибки, опечатки в словах)**

4. НАЙТИ НЕЧЕТКИЕ КЛАСТЕРЫ ЛИСТОВЫХ ТЕМ

- Преобразовать прямоугольную матрицу релевантности «тема-текст» в квадратную матрицу корелевантности «тема-тема»
- Провести предобработку матрицы корелевантности, чтобы сделать ее кластерную структуру четче
- Применить метод аддитивного спектрального нечеткого кластер-анализа ФАДДИС (Mirkin, Nascimento, Inf.Sci., 2012)

МАТРИЦА КОРЕЛЕВАНТНОСТИ 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* $r_{tv} > 0.2$
  (темы, релевантные тексту)

# Texts	n_t = # Relevant subjects
1237	0
2353	1
7114	2 – 4
6124	5 – 11
857	12 or more

ADDITIVE FUZZY CLUSTERING **FADDIS** (MIRKIN, NASCIMENTO, 2012) , I

29 Observed:

- **Similarity $B=(b_{ij})$, $i,j \in I$: co-relevance C after pseudo-inverse Laplacian transformation (LAPIN)**

Laplacian: $L = I - D^{-1/2}CD^{-1/2}$ where D is diagonal, $d_{ii} = \sum_j c_{ij}$

PIN: $B = Z\Lambda^{-1}Z^T$

To be found:

- **Cluster membership $u_k=(u_{ik})$**
- **Intensity $\mu_k > 0 \Rightarrow$**

Fuzzy cluster similarity $A_k = \mu_k^2 u_k u_k^T$

ADDITIVE FUZZY CLUSTERING FADDIS (MIRKIN, NASCIMENTO, 2012), 2

30 Observed:

- **Similarity $B=(b_{ij})$, $i,j \in I$**

To be found:

- **Cluster membership $u=(u_i)$**
- **Intensity $\mu > 0 \Rightarrow$**

Fuzzy cluster similarity $A= \mu^2 uu^T$

K clusters:

$B = A_g + A_1 + A_2 + \ldots + A_K + E$ (g- universal background)

$\langle B'- A_k, B'- A_k \rangle \Rightarrow \min_{\mu,u}$ one-by-one

ADDITIVE FUZZY CLUSTERING STEP

31

Given:

- **Residual similarity $B'=(b'_{ij})$, $B' = B - A_g - A_1 - \ldots - A_{k-1}$**

To be found:

- **Cluster membership $u=(u_i)$,**
- **Intensity $\mu > 0 \Rightarrow$**

Fuzzy cluster similarity $A= \mu^2 uu^T$, $\|u\| = 1$, $u \geq 0$

$\langle B'- \mu^2 uu^T, B'- \mu^2 uu^T \rangle \Rightarrow \min_{\mu,u}$ with no $u \geq 0$ constraints:

μ - maximum eigenvalue, u – corresponding eigenvector

Spectral clustering: adjust u to $u \geq 0$

FADDIS METHOD: ONE CLUSTER AT A TIME

- $\mathbf{Min}_{\mathbf{u},\xi} \Sigma_{t,t' \in T} (b_{tt'} - \xi u_t u_{t'})2$
- Equivalent to maximum of Rayleigh quotient

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

- **Spectral approach (She, Malik, 2000):** find max eigenvalue and its vector, adjust the latter to fuzzy membership
- Found 6 fuzzy clusters of which 3 are more or less homogeneous: **L- Machine Learning, C- Clustering, and I – Information Retrieval**

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": $U_I \geq 0.15$

34

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. & <u>equational</u> 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

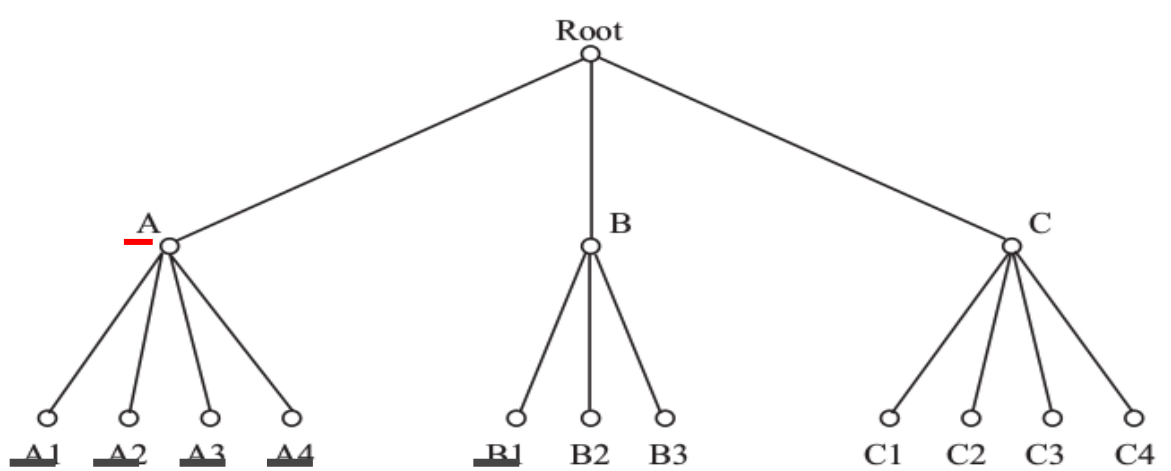
ОБОБЩЕНИЕ КАК ПОДЪЕМ В ТАКСОНОМИИ

- Толковый словарь Современного русского языка Ефремовой:
- **«Обобщать» - Объединять, соединять вместе**.

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

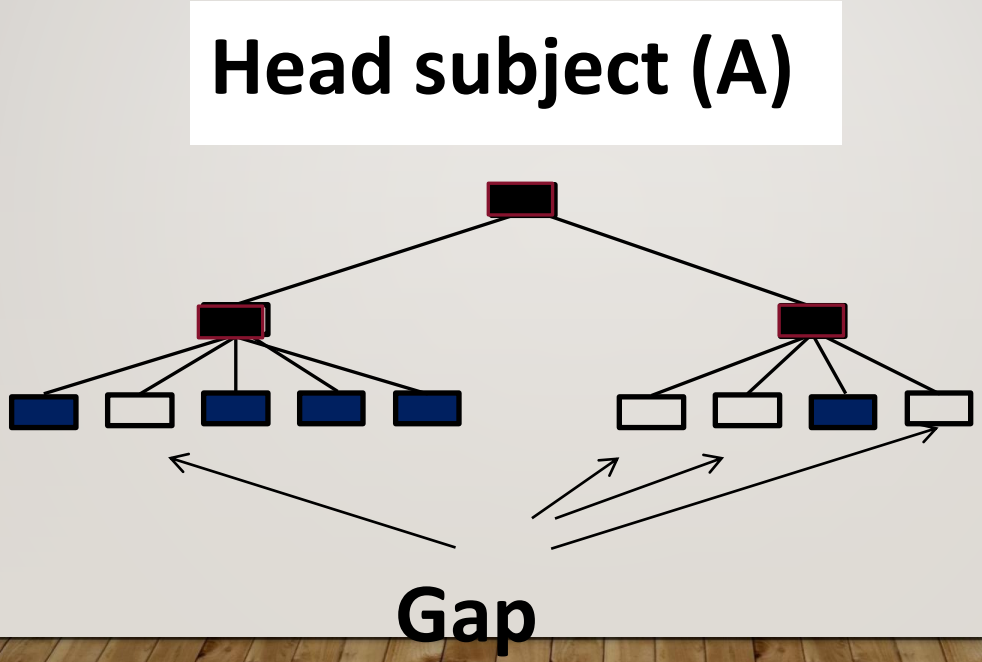
- ``to give a general form to"
- ``to derive or induce (a general conception or principle) from particulars"

GENERALIZATION: APPROPRIATELY LIFTING CLUSTERS TO COMMON ROOT CONCEPTS

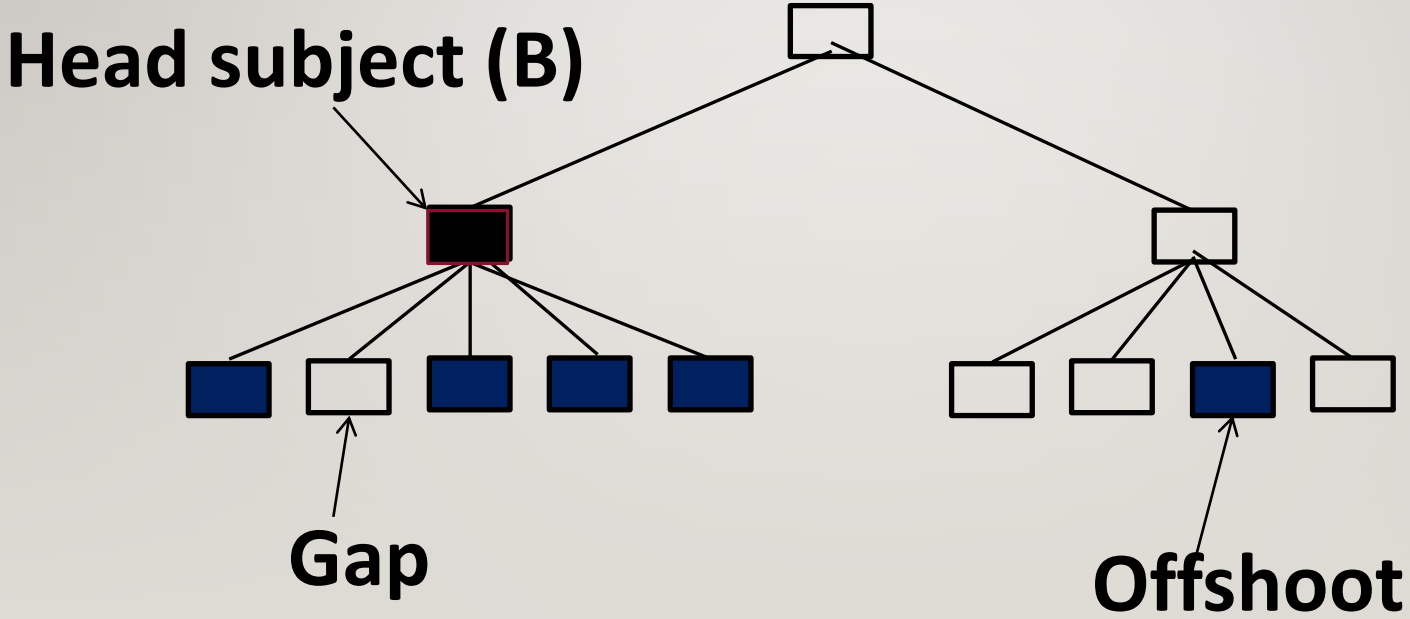


Given a taxonomy and a leaf cluster, lift the leaves to a higher rank node: **(A1, A2, A3, A4, B1) => (A)**, B1 disregarded as an **offshoot**.

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



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 A: $1+4\lambda$

Penalty at 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**

Algorithm computes two sets at each node **t** :

1. **H(t)** – set of newly emerged subjects
2. **L(t)** – set of lost subjects

and penalty value **p(t)**.

Starting from the leaves, algorithm recursively computes these till root is reached.

At each node **t**, two scenarios: (a) head subject emerges at **t**, or, (b) it does not.

Scenario (a) – Head subject emerges at t. Then:

$$H(t) = \{t\}$$

$$L(t) = G(t)$$

$$p(t) = u(t) + \lambda V(t)$$

Scenario (b) – head subject does not emerge. Then sets H и L are derived as unions of the corresponding sets in all the children:

$$H(t) = \bigcup_{w \in \chi(t)} H(w)$$

$$L(t) = \bigcup_{w \in \chi(t)} L(w)$$

$$p(t) = \sum_{w \in \chi(t)} p(w)$$

Scenario (a) or (b) is chosen depending on the value ***p(t)*** – the smaller wins.

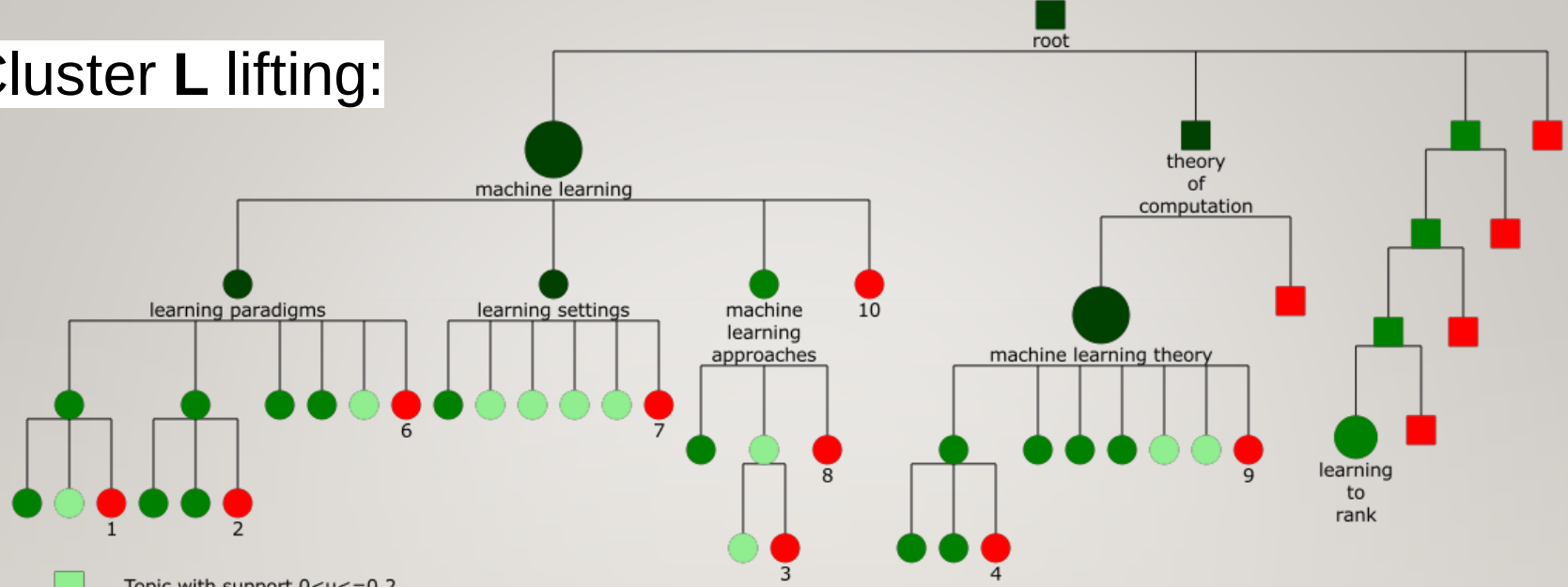
TCAN SOFTWARE

- GOT package
- Includes
  - Relevance and co-relevance matrices
  - FADDIS clustering
  - Optimal 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:



Topic with support 0<u<=0.2

Topic with support 0.2<u<=0.4

Topic with support u>0.4

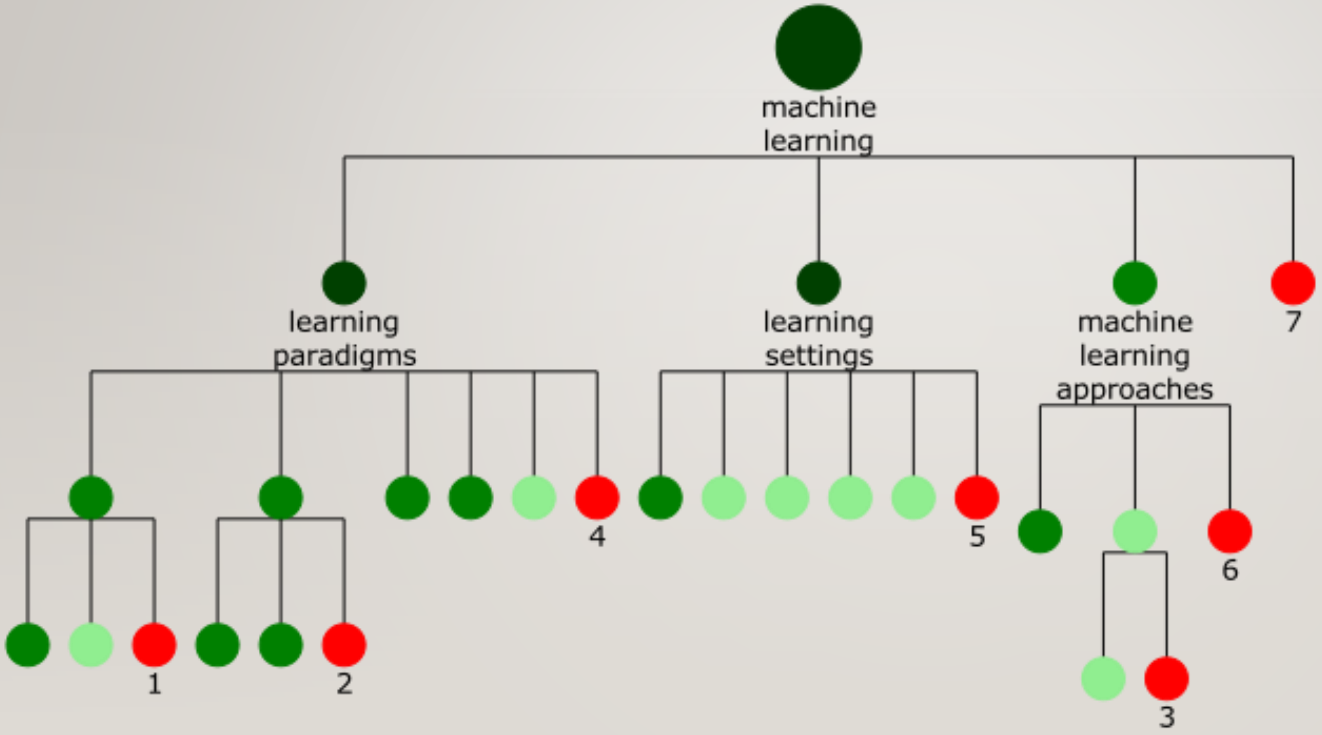
Topic with no support (u=0)

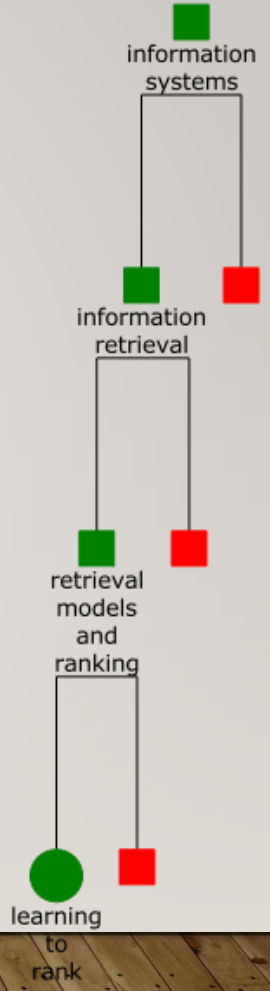
Gap

Head subject

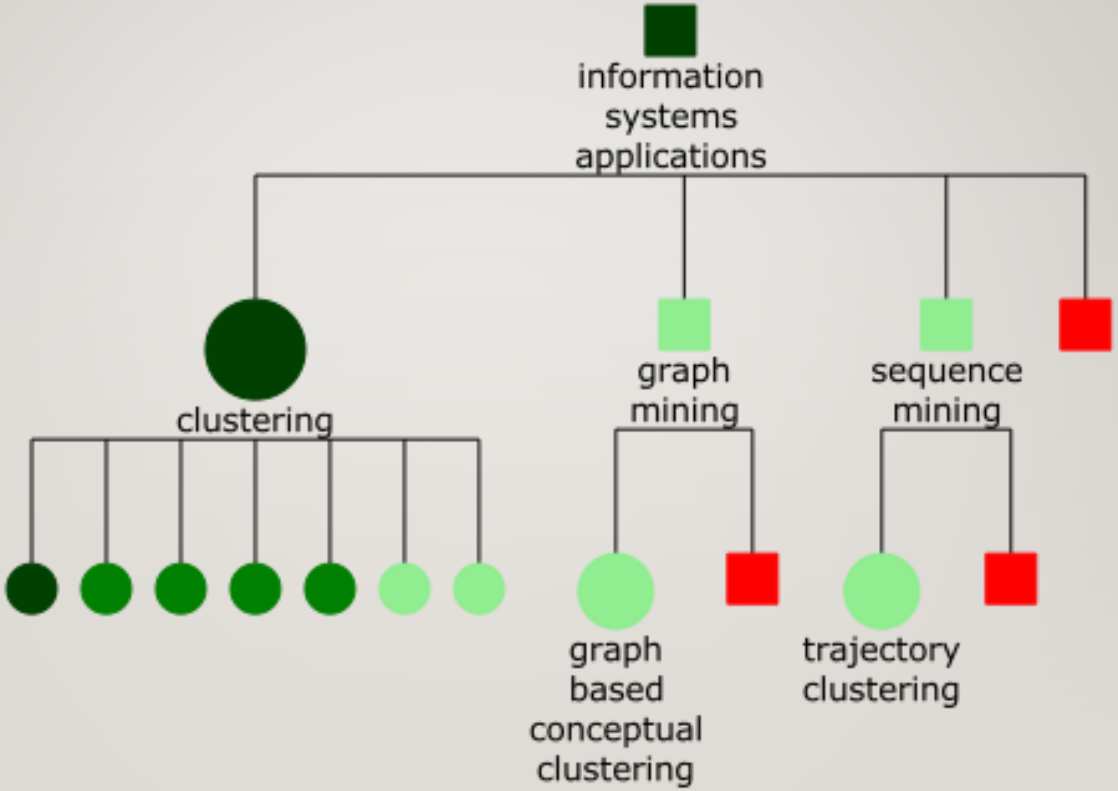
Head subjects: {Machine learning, Machine learning theory, Learning to rank}

Cluster **L** lifting (fragment):





Cluster **C** lifting: a fragment



CLUSTER LIFTING RESULTS SUGGEST, I:

- "Learning" lifted:
  - 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" cluster R lifted:
- 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
- Compare this with results of citation graph analyses

CLUSTER LIFTING RESULTS SUGGEST 3:

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

КОЛЛЕКЦИЯ 2: АННОТАЦИИ СТАТЕЙ SPRINGER И ELSEVIER, ПО ЗАПРОСАМ ИЗ НАУКИ ДАННЫХ}

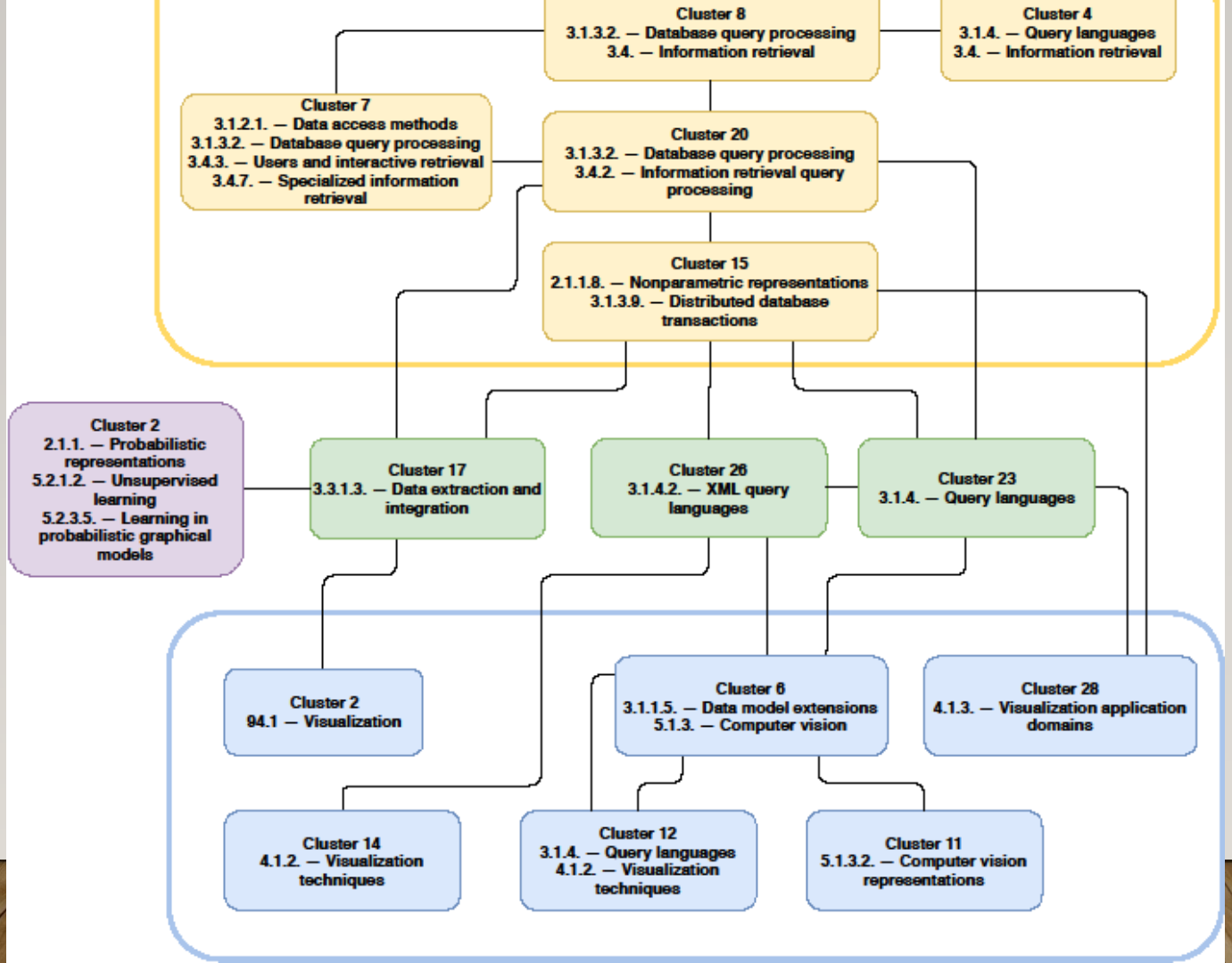
- **Запросы:** clustering, machine learning, neural networks, algorithm, classification, information retrieval, natural language processing, software, computing, pattern recognition, deep learning, probabilistic, artificial intelligence, support vector, Bayesian, regression, search engine
- **Коллекция2:** аннотации **26 799** статей из 80 журналов 1971-2019

МАТРИЦА КОРЕЛЕВАНТНОСТИ 317×317

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

- n_t = *number of leaf subjects v such that* $r_{tv} > 0.2$

# Col.1	Col.2	n_t = # Relevant subjects
1237	2401	0
2353	3867	1
7114	10868	2 – 4
6124	8921	5 – 11
857	742	12 or more


Cluster 8
3.1.3.2. — Database query processing
3.4. — Information retrieval
Cluster 4
3.1.4. — Query languages
3.4. — Information retrieval
Cluster 7
3.1.2.1. — Data access methods
3.1.3.2. — Database query processing
3.4.3. — Users and interactive retrieval
3.4.7. — Specialized information retrieval
Cluster 20
3.1.3.2. — Database query processing
3.4.2. — Information retrieval query processing
Cluster 15
2.1.1.8. — Nonparametric representations
3.1.3.9. — Distributed database transactions
Cluster 2
2.1.1. — Probabilistic representations
5.2.1.2. — Unsupervised learning
5.2.3.5. — Learning in probabilistic graphical models
Cluster 17
3.3.1.3. — Data extraction and integration
Cluster 26
3.1.4.2. — XML query languages
Cluster 23
3.1.4. — Query languages
Cluster 2
94.1 — Visualization
Cluster 6
3.1.1.5. — Data model extensions
5.1.3. — Computer vision
Cluster 28
4.1.3. — Visualization application domains
Cluster 14
4.1.2. — Visualization techniques
Cluster 12
3.1.4. — Query languages
4.1.2. — Visualization techniques
Cluster 11
5.1.3.2. — Computer vision representations

ЗАМЫСЛОВАТАЯ СХЕМА КЛАСТЕРОВ НА ПРЕДЫДУЩЕМ СЛАЙДЕ

- **Две группы кластеров** :
  - Извлечение и хранение данных (выделены желтым).
  - Визуализация данных (выделены синим)
- **Связка**
  - Кластеры языков для извлечения информации

ФАДДИС НА КОЛ.2, I: 35 КЛАСТЕРОВ, ИЗ НИХ ИНТЕРПРЕТИРУЕМЫ 7, C1,C2,C3,C4,C5,C6,C7

57

- Cl. L ~ C1
- Cl. C ~ C2
- Cl. R: Кластеры C3, C4, C5, C6, C7 (след. слайд)

ФАДДИС НА КОЛ.2, 2:
5 КЛАСТЕРОВ, СООТВЕТСТВУЮЩИХ КЛАСТЕРУ R ПО КОЛ. I

- C3 2.1.1. -- Probabilistic representations
- C4 Retrieval: 3.1.4.—Query languages \\ 3.4– Information Retrieval
- C5 Structuring: 3.1.1.5. -- Data model extensions\\ 5.1.3. -- Computer vision
- C6 5.1.3.2. -- Computer vision representations
- C7 Querying: 3.1.3.2. -- Database query processing\\ 3.4.2. -- Information retrieval query processing

ПЕРЕНОС ПОДХОДА НА ДРУГИЕ КОЛЛЕКЦИИ

- «Посещение ресторана» (18036 отзывов о ресторанах и кафе г. Москвы, TripAdvisor, 2019, leaf subjects 267)
- «Осмотр автомобиля» (35785 отзывов пользователей об автомобилях различных марок)
- «Содержание академического журнала» (abstracts of all 461 papers "Journal of Classification" 1984-2019, leaf subjects 106)
- **Результаты не очень интересны; возможно, таксономии носят слишком академический характер**

ИСПОЛЬЗОВАНИЕ БОЛЕЕ ГРУБОЙ ТАКСОНОМИИ

- Анализ 2000 объявлений hh.ru о вакансиях в сфере ВТ
- Таксономия
  - Т: Требования к подготовке
  - О: Рабочие обязанности
  - У: Предоставляемые условия
- С использованием АСД получены оценки корелевантности на декартовом произведении Т×О×У
- Найдены трикластеры – тесно связанные подмножества требований, обязанностей, условий
- На этом пути перспективно: получить более адекватную типологию профессий в ВТ

ЗАКЛЮЧЕНИЕ

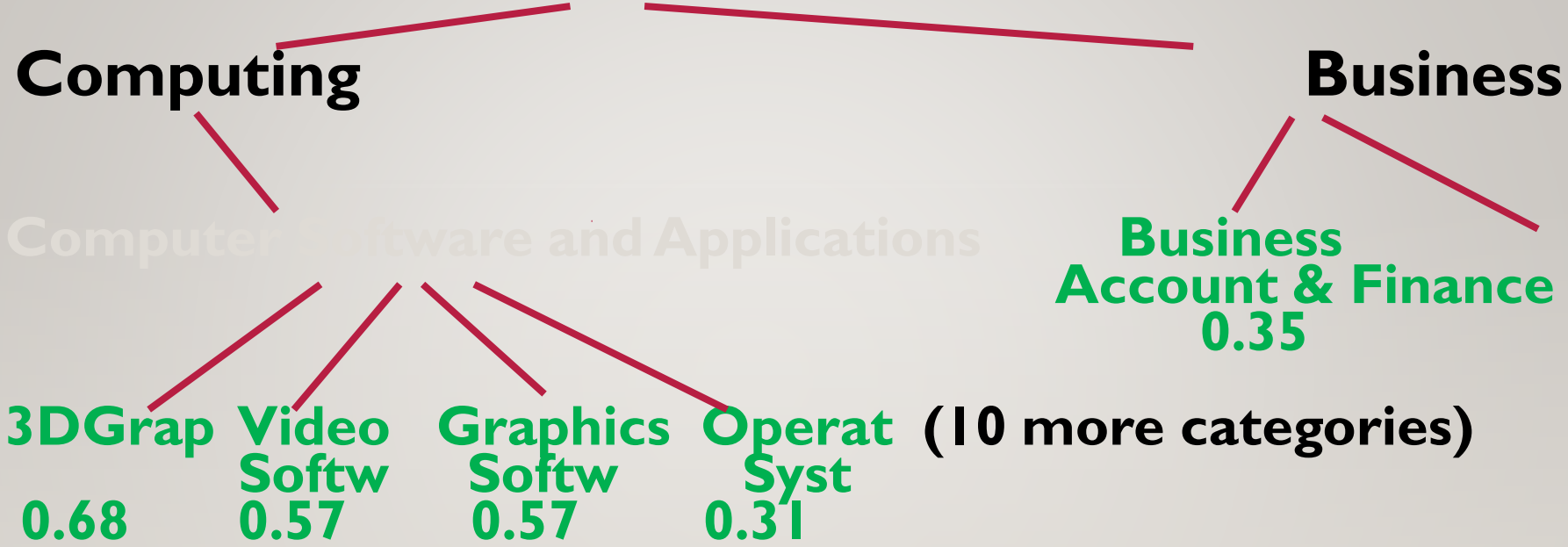
61

- Предложен метод ТКАН контент-анализа текстовых коллекций, в котором явно используется состав и структура таксономии предметной области
- Метод ТКАН включает оригинальную компоненту: модель обобщения в таксономии как задачу о наиболее экономном подъеме заданного нечеткого множества листьев
- Метод ТКАН перспективен постольку, поскольку перспективна идея разработки и поддержания таксономических классификаций предметных областей
- Первоочередная теоретическая задача – замена критерия максимальной экономии на критерий максимального правдоподобия
- Перенос оптимального подъема в другие приложения: оптимизация таргетированного адвертайзинга (Фролов)

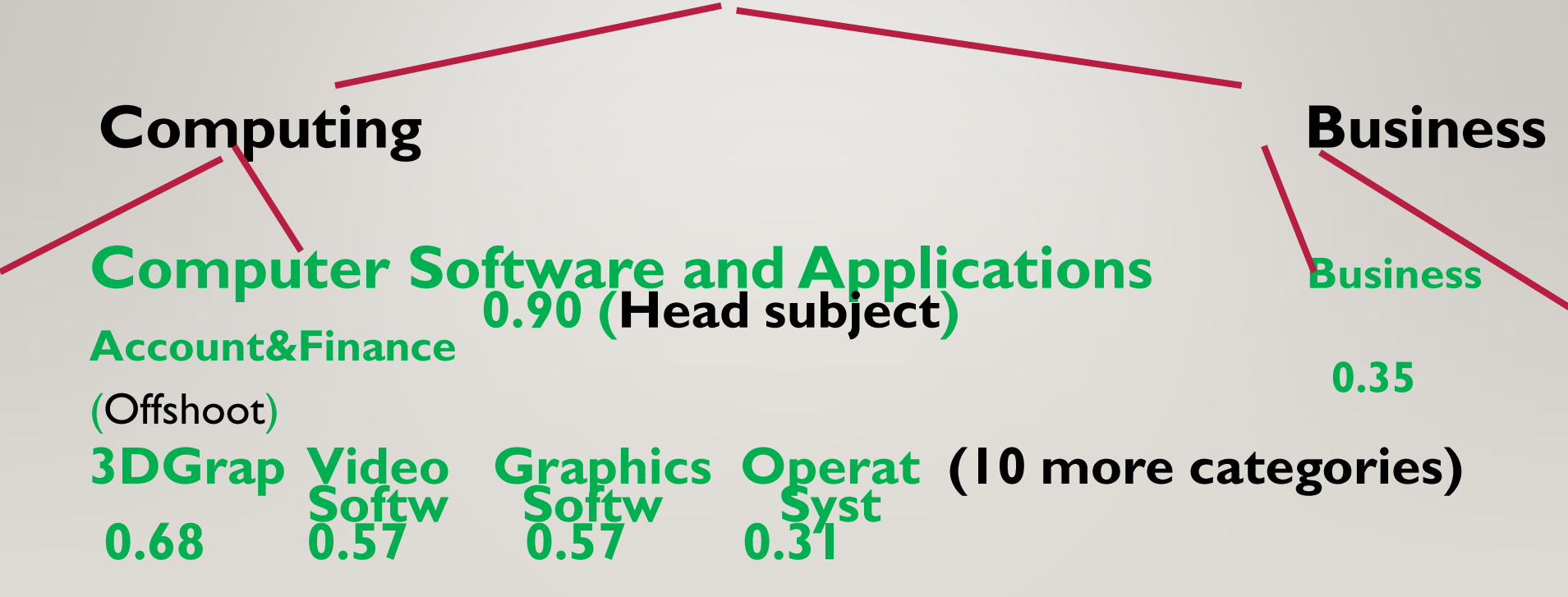
Применение к таргетированному адвертайзингу

D. FROLOV («НАТИМАТИКА»): ARCHIVE OF ~ 20 MLN VISITORS OF POPULAR SITES, EACH TAGGED BY LABELS FROM IAB TAXONOMY ACCORDING TO THEIR INTERNET BEHAVIOR

A USER PROFILE IN IAB IN NATIMATICA (GREEN)



LIFTING USER PROFILE TO A HEAD SUBJECT



APPLYING TO (PROGRAMMATIC) TARGETED ADVERTS, 1: EXAMPLE

USER'S TARGET for an advert on

***** AS: Antivirus Software *****

Solution: Show that to all users whose tag for AS is >0.3

Issue: Not too many visitors

APPLYING TO TARGETED ADVERTS, 2: EXAMPLE

USER'S TARGET for an advert on

*****AS: Antivirus Software*****

Solution 1: Show that to all users whose tag for AS is >0.3

Issue: Not too many visitors

Solution 2: Show that to all users whose tag for AS is >0.1

Issue: Falling efficiency

APPLYING TO TARGETED ADVERTS, 3: EXAMPLE

USER'S TARGET for an advert on

***** AS: Antivirus Software *****

Solution 1: Show that to all users whose tag for AS is >0.3

Issue: Not too many visitors

Solution 3: LIFT. Show that to all users whose lifted node contains AS

Feature: Surprising efficiency

APPLYING TO TARGETED ADVERTS, 4: EXAMPLE

USER'S TARGET: an advert on

Daycare and Pre-School; Internet Safety; Parenting Children Aged 4-11; Parenting Teens; Antivirus Software

Impressions	378933	942104	1017598
Clicks	1061	**2544**	1526
	Standard	**LIFT**	**Solution 2**

CONCLUSION: DONE

- A novel computational model for generalization, and an effective algorithm
- Applied to the analysis of research publications leading to observations on tendencies of the research, which cannot be produced with a popular approach of the analysis of co-citation graphs

CONCLUSION: TO DO:

- **Deeper into generalization: inserting a node to reduce the penalty**
- **Shifting to the criterion of maximum likelihood**
- **More applications**

PUBLICATION SO FAR

- 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

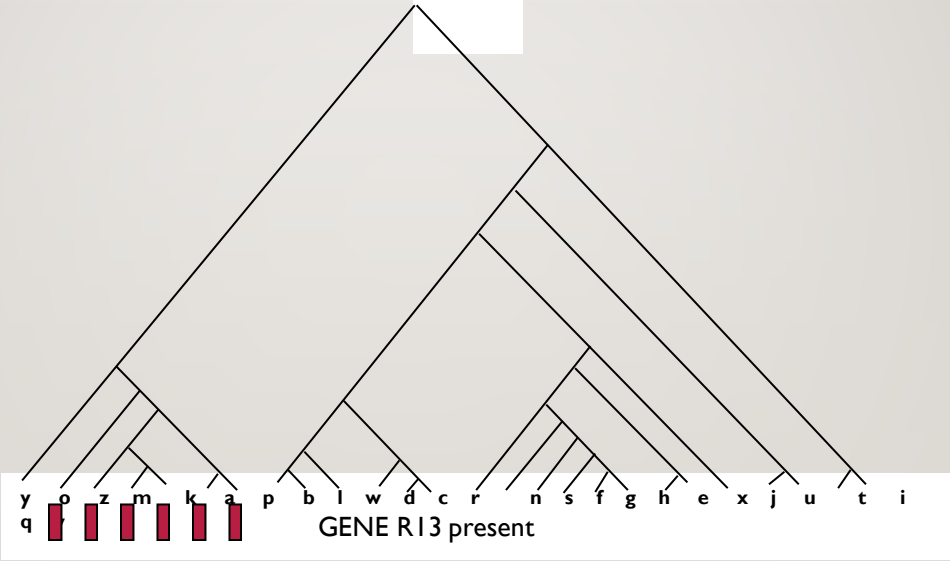
RELATED STUFF

SOME PUBLICATIONS

- Ekaterina Chernyak and Boris Mirkin (2014) An AST method for scoring string-to-text similiarity 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.

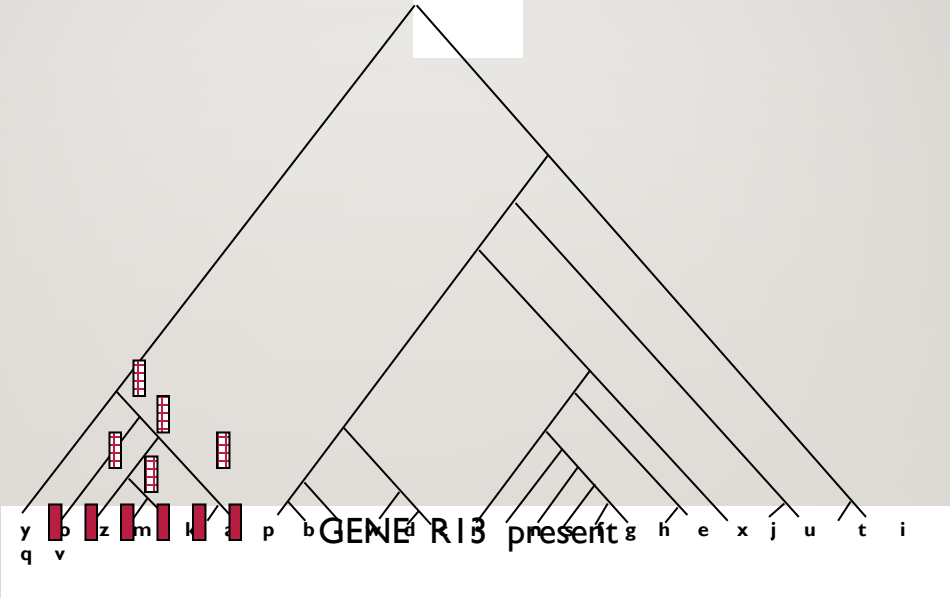
ORIGIN: ROUND 1 FROM EUGENE KOONIN (NCBI NIH USA)
RECONSTRUCTION OF GENE HISTORY

EVOLUTIONARY TREE OVER 26 EXTANT BACTERIAL GENOMES: BOX (PRESENCE OF R13 GENE). QUESTION: AT WHAT PLACE R13 EMERGED (WAS GANED)



EVOLUTIONARY TREE

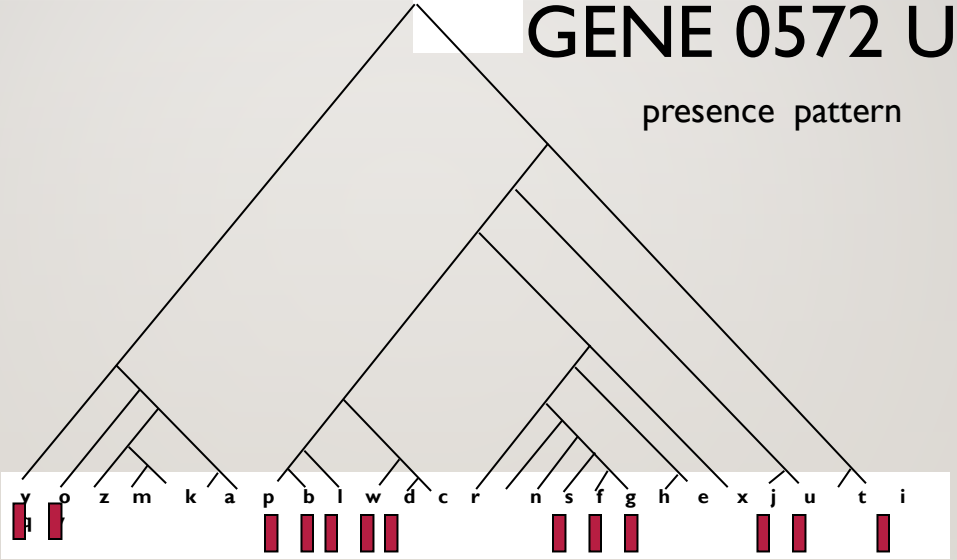
Parsimoniously Reconstructing R13 history



EVOLUTIONARY TREE: MORE REALISTIC CASE

GENE 0572 Uridine Kinase:

presence pattern



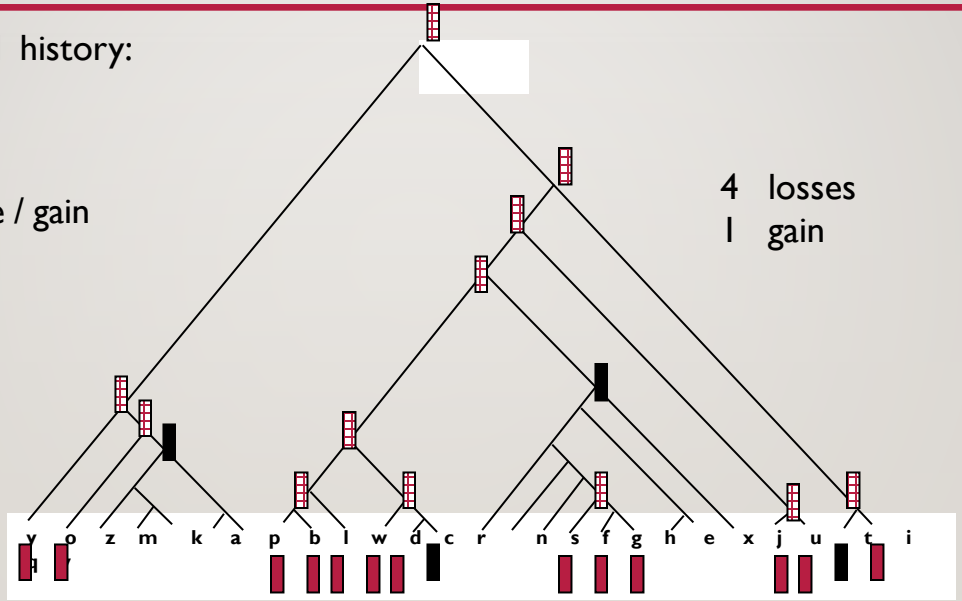
GENE 0572 Uridine Kinase Evolutionary history Reconstruction at equal weights

Reconstructed history:

presence

inheritance / gain

loss



4 losses
1 gain

FINAL RESULTS (MIRKIN,…,KOONIN, 2003):

20 RECONSTRUCTIONS MADE OVER 3066 GENES EACH

FROM GAIN WEIGHT 10, LOSS WEIGHT 1
TO GAIN WEIGHT 1, LOSS WEIGHT 10

LUCA CONTENTS DIFFER EACH TIME; THE BEST MATCHING THE THEORETIC EXPECTATION (MODEL "GENOME SIMPLE") GW=LW=1

LUCA 567 GENES

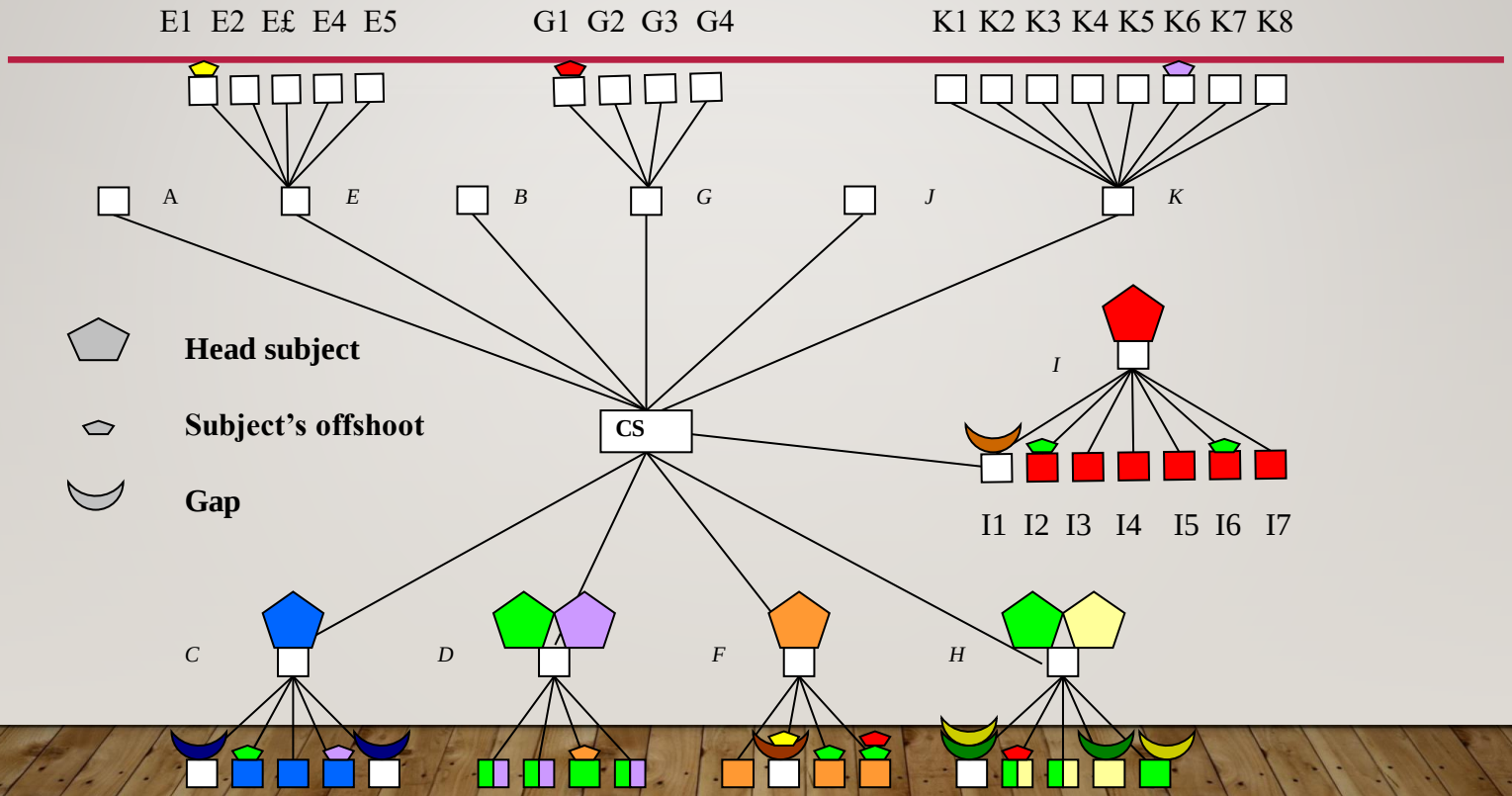
ORIGIN: ROUND 2 FROM PROF LUIS MONIZ PEREIRA (UNIVERSIDADE NOVA LISBOA PORTUGAL)

POSITIONING OF CENTRIA OVER THE CS TAXONOMY

C. COMPUTER SYSTEMS ORGANIZATION D. SOFTWARE AND H. INFORMATION SYSTEMS

F. THEORY OF COMPUTATION D. SOFTWARE H. INFORMATION SYSTEMS

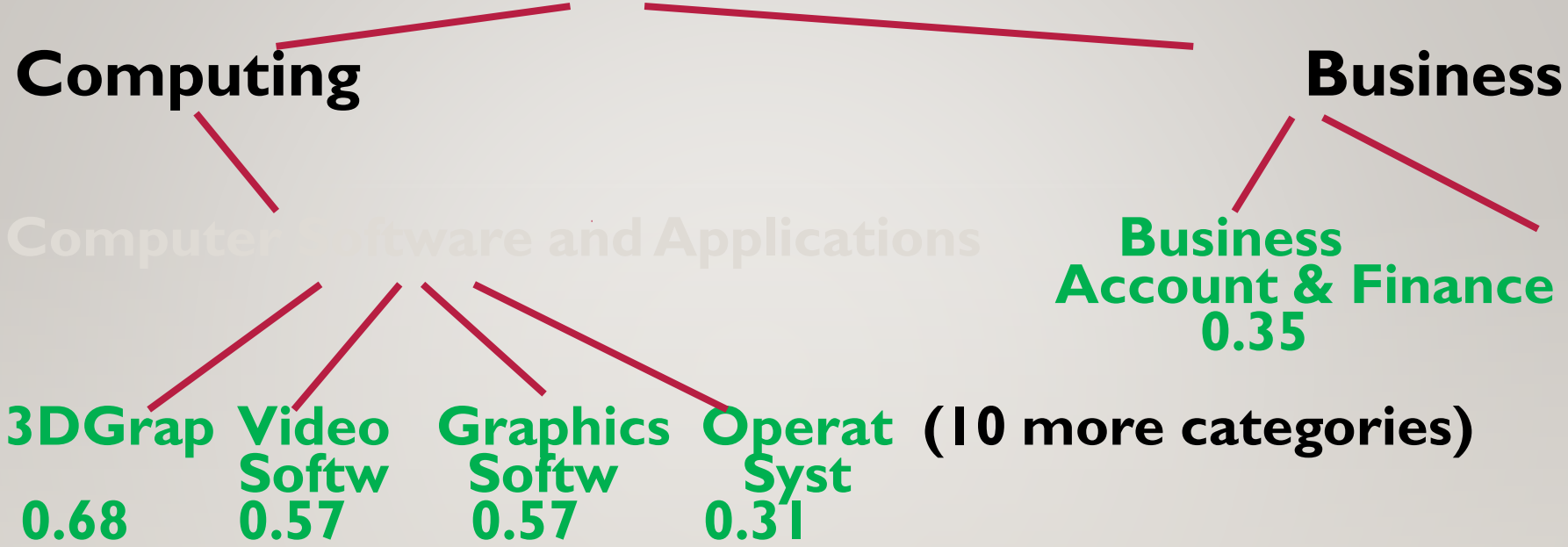
I. COMPUTING METHODOLOGIES



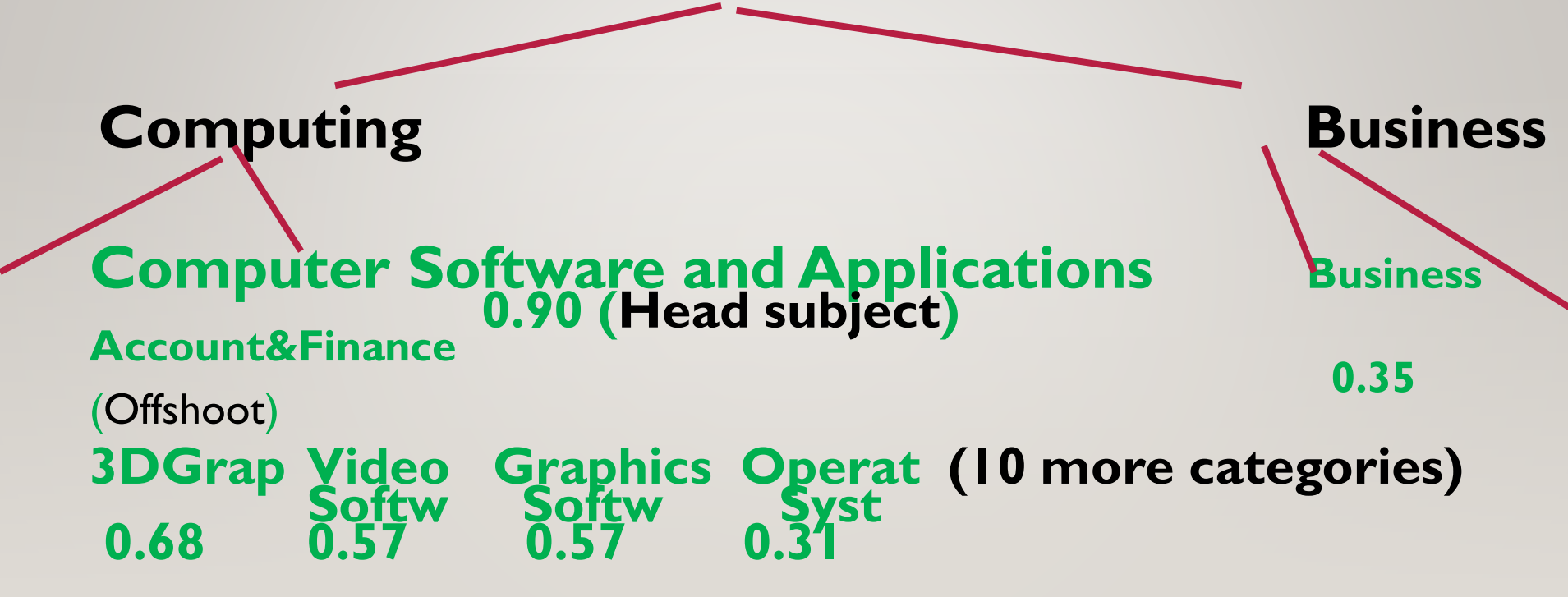
D. FROLOV («НАТИМАТИКА»):

ARCHIVE OF
~ 5 MLN VISITORS OF POPULAR SITES, EACH TAGGED BY LABELS FROM IAB TAXONOMY ACCORDING TO THEIR INTERNET BEHAVIOR

A USER PROFILE IN IAB IN NATIMATICA (GREEN)



LIFTING USER PROFILE TO A HEAD SUBJECT



APPLYING TO (PROGRAMMATIC) TARGETED ADVERTS, 1: EXAMPLE

USER'S TARGET for an advert on

***** AS: Antivirus Software *****

Solution: Show that to all users whose tag for AS is >0.3

Issue: Not too many visitors

APPLYING TO TARGETED ADVERTS, 2: EXAMPLE

USER'S TARGET for an advert on

AS: Antivirus Software

Solution 1: Show that to all users whose tag for AS is >0.3

Issue: Not too many visitors

Solution 2: Show that to all users whose tag for AS is >0.1

Issue: Falling efficiency

APPLYING TO TARGETED ADVERTS, 3: EXAMPLE

USER'S TARGET for an advert on

***** AS: Antivirus Software *****

Solution 1: Show that to all users whose tag for AS is >0.3

Issue: Not too many visitors

Solution 3: LIFT. Show that to all users whose lifted node contains AS

Feature: Surprising efficiency

APPLYING TO TARGETED ADVERTS, 4: EXAMPLE

USER'S TARGET: an advert on

Daycare and Pre-School; Internet Safety; Parenting Children Aged 4-11; Parenting Teens; Antivirus Software

Impressions	378933	942104	1017598
Clicks	1061	**2544**	1526
	Standard	**LIFT**	**Solution 2**