Агрегированное представление текстов для задач поиска в коллекциях текстовых документов

Фролов Дмитрий Сергеевич
Научн. руководитель: д.т.н. Миркин Б.Г.

Структура доклада

1. Актуальность, цель, новизна работы
2. Информационный поиск на основе аннотированных суффиксных деревьев
3. Методика интерпретации: подъем нечетких кластеров в таксономии предметной области
4. Алгоритм оптимального обобщения ПарГеНМ
5. Применение: анализ коллекций научных публикаций
6. Применение: рекламный таргетинг
7. Заключение

Структура доклада

1. **Актуальность, цель, новизна работы**
2. Информационный поиск на основе аннотированных суффиксных деревьев
3. Методика интерпретации: подъем нечетких кластеров в таксономии предметной области
4. Алгоритм оптимального обобщения ПарГеНМ
5. Применение: анализ коллекций научных публикаций
6. Применение: рекламный таргетинг
7. Заключение

1. Актуальность

- повышение скорости поиска документов;
- повышение качества поиска документов;
- автоматизация анализа текстовой информации, включая ее структурирование и интерпретацию.

Цель диссертации

Разработка эффективных методов для поиска и анализа текстовых данных на основе использования аннотированных суффиксных деревьев (АСД) для агрегированного представления коллекций документов

Научная новизна, 1/2: впервые

1. Разработана методика информационного поиска для коллекций документов, представленных в виде аннотированных суффиксных деревьев (АСД) и экспериментально доказана ее эффективность.

2. Разработан и применен метод интерпретации результатов поиска с помощью выявления и оптимального обобщения нечетких кластеров в таксономии предметной области (ПарГеНМ).

Научная новизна, 2/2: впервые

3. Разработан метод поиска релевантной аудитории для интернет-рекламы на основе оптимального обобщения сегментов пользователей в таксономии предметной области (ОПС). Метод позволяет расширить аудиторию среди пользователей интернета в 2-2.5 раза.

Структура доклада

1. Актуальность, цель, новизна работы
2. **Информационный поиск на основе аннотированных суффиксных деревьев**
3. Методика интерпретации: подъем нечетких кластеров в таксономии предметной области
4. Алгоритм оптимального обобщения ПарГеНМ
5. Применение: анализ коллекций научных публикаций
6. Применение: рекламный таргетинг
7. Заключение

2. Разработка метода информационного поиска на основе аннотированных суффиксных деревьев (АСД)

Дано множество (коллекция) документов и множество поисковых запросов. Требуется для каждого запроса предоставить множество наиболее релевантных ему документов из коллекции.

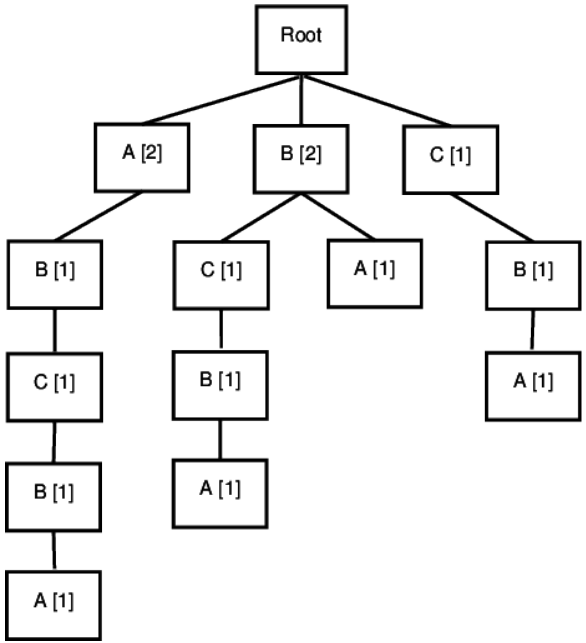
Предложен метод поиска АСДП:

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

Аннотированное суффиксное дерево (АСД)

Структура данных, представляющая текст через фрагменты текста (суффиксы) и их частоты (Миркин, Черняк-Артемова)

АСД для строки “*ABCBA*”:



Вычисление релевантности *S(x,T)* с помощью АСД (Миркин, Артемова, Чугунова, 2012)

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

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

Для строки $x=x_1 \ldots x_N$ релевантность *S(x, T)* АСД *T*:

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

$$S(x,T) = \frac{1}{N} \sum_{k=0}^{N} s(x_k, T)$$

k_{max} - длина наибольшего совпадения суффикса с АСД

Шаги метода АСДП

1. Разделить поисковый запрос на фрагменты и отобрать “документы-кандидаты” из обратного фрагментного индекса.

2. Рассчитать значения релевантности АСД, построенным для отобранных документов.

3. Отсортировать полученные значения по убыванию.

Обратный фрагментный индекс

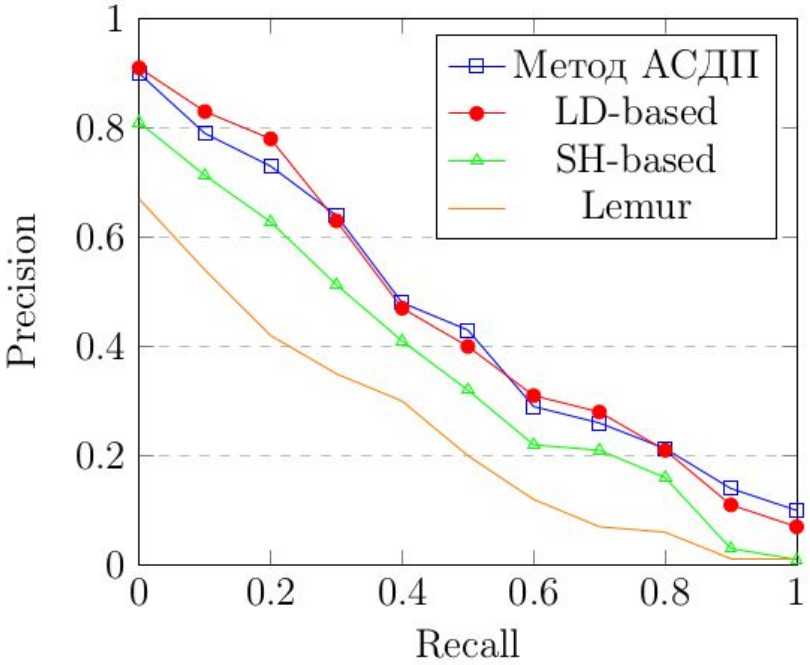
Ускорение: проверять степень вхождения строки-запроса не во все документы, а только в отобранные обратным индексом (например, вместо 20 000 документов - 200 отобранных)

Обратный индекс по фрагментам f_i:

$$\begin{cases} f_1 \leftarrow [(n_{11}, c_{11}), \ldots, (n_{1m_1}, c_{1m_1})] \\ f_2 \leftarrow [(n_{21}, c_{21}), \ldots, (n_{2m_2}, c_{2m_2})] \\ \ldots \\ f_K \leftarrow [(n_{K1}, c_{K1}), \ldots, (n_{Km_K}, c_{Km_K})] \end{cases}$$

Экспериментальное сравнение, 1/3

Популярные методы поиска: LD-based (Levenstein-based), SH-based (Signature hashing), Lemur (разработка MIT)



Коллекция Habr Extract

Метод	Место
АСДП	**1-2**
LD-based	**1-2**
SH-based	3
Lemur	4

Экспериментальное сравнение, 2/3

Производительность поиска, с

Коллекция	Метод АСДП	LD-based	SH-based	Lemur
#1	0.43	1.03	0.39	**0.08**
#2	0.15	0.34	0.16	**0.05**

Метод	колл. #1	колл. #2	Среднее
Lemur	1	1	1
SH-based	2	3	2-3
АСДП	3	2	2-3
LD-based	4	4	4

Экспериментальное сравнение, 3/3

АСДП **наиболее сбалансирован:**

Скорость + качество: сумма мест

Метод	Ранг	Место
АСДП	**4**	**1**
Lemur	5	2
SH-based	5,5	3-4
LD-based	5,5	3-4

Структура доклада

1. Актуальность, цель, новизна работы
2. Информационный поиск на основе аннотированных суффиксных деревьев
3. **Методика интерпретации: подъем нечетких кластеров в таксономии предметной области**
4. Алгоритм оптимального обобщения ПарГеНМ
5. Применение: анализ коллекций научных публикаций
6. Применение: рекламный таргетинг
7. Заключение

3. Метод формирования и обобщения нечетких кластеров в таксономии для разведочного поиска в коллекции

М1. Формирование таксономии предметной области.

М2. Построение таблицы Т оценок релевантности "тема таксономии - документ".

М3. Формирование нечетких кластеров "тем таксономии", соответствующих структуре коллекции документов.

М4. Оптимальный подъем тематических кластеров к верхним ярусам таксономии.

М5. Использование получаемых “головных” тем для содержательных выводов.

M1. Таксономия Науки о данных, 1/2

Темы из ACM Computing Classification System (2012)

Индекс темы	название темы
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

M1. Таксономия Науки о данных, 2/2

Фрагмент нижнего яруса (число листьев - 317)

5.2.3.8.		Rule learning	
5.2.3.8.1*			Neuro-fuzzy approach
5.2.3.9.		Instance-based learning	
5.2.3.10.		Markov decision processes	
5.2.3.11.		Partially-observable Markov decision processes	
5.2.3.12.		Stochastic games	
5.2.3.13.		Learning latent representations	
5.2.3.13.1.			Deep belief networks
5.2.3.14*		Multiresolution	
5.2.3.15*		Support vector machines	
5.2.4.	Machine learning algorithms		
5.2.4.1.		Dynamic programming for Markov decision processes	
5.2.4.1.1.			Value iteration
5.2.4.1.2.			Q-learning
5.2.4.1.3.			Policy iteration
5.2.4.1.4.			Temporal difference learning
5.2.4.1.5.			Approximate dynamic programming methods
5.2.4.2.		Ensemble methods	
5.2.4.2.1.			Boosting
5.2.4.2.2.			Bagging
5.2.4.2.3.**			Fusion of classifiers
5.2.4.3.		Spectral methods	
5.2.4.3.1*			Spectral clustering
5.2.4.4.		Feature selection	
5.2.4.5.		Regularization	
5.2.4.5.1*			Generalized eigenvalue
5.2.5.	Cross-validation		

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

M2.1. Формирование коллекции текстов

17685 статей с сайта SpringerOpen (springeropen.com) из 17 журналов по Науке о данных (1998 - 2017):

1 Pattern Analysis and Applications (Volume 1 / 1998 - Volume 20 / 2017)

2 World Wide Web (Volume 1 / 1998 - Volume 20 / 2017)

...

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

M2.2. Построение таблицы T оценок релевантности: “тема таксономии - документ”

Для получения оценок: метод аннотированного суффиксного дерева.

Матрица значений релевантности $T = (T_{ij})$ размера 317 x 17685: листовые темы таксономии x статьи

М3. Формирование нечетких кластеров

Матрица ко-релевантности $C=T N^{-1} T'$, где N - диагональная, n_{ii} - число текстов, релевантных теме i

Метод FADDIS после преобразования Лапласа минимизирует (Mirkin, Nascimento, 2012):

$$E = \sum_{t,t' \in T} (w_{tt'} - \xi u_t u_{t'})^2$$

Получено 6 нечетких кластеров; 3 - интерпретируемы.

Кластер L: Learning

$u(t)$	Code	Topic
0.300	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
0.192	5.2.1.4.1.	transfer learning
0.192	5.2.1.4.2.	lifelong machine learning
0.189	1.1.1.8.	online learning theory
0.166	5.2.2.2.	online learning settings
0.159	1.1.1.3.	unsupervised learning and clustering

Кластер C: Clustering

$u(t)$	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

Кластер R: Retrieval

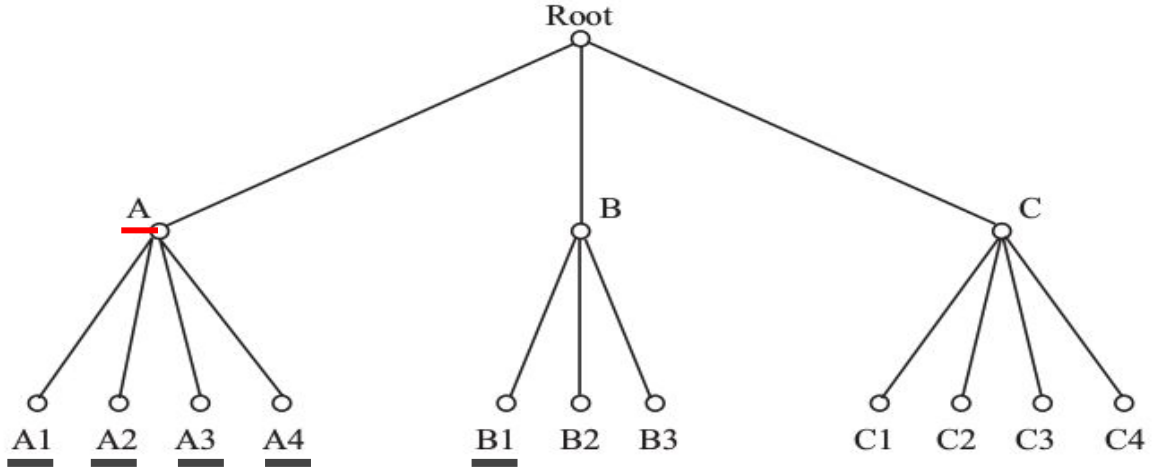
$u(t)$	Code	Topic
0.211	3.4.2.1.	query representation
0.207	5.1.3.2.1.	image representations
0.194	5.1.3.2.2.	shape representations
0.194	5.2.3.6.2.1	tensor representation
0.191	5.2.3.3.3.2	fuzzy representation
0.187	3.1.1.5.3.	data provenance
0.173	2.1.1.5.	equational models
0.173	3.4.6.5.	presentation of retrieval results
0.165	5.1.3.1.3.	video segmentation
0.155	5.1.3.1.2.	image segmentation
0.154	3.4.5.5.	sentiment analysis

Структура доклада

1. Актуальность, цель, новизна работы
2. Информационный поиск на основе аннотированных суффиксных деревьев
3. Методика интерпретации: подъем нечетких кластеров в таксономии предметной области
4. **Алгоритм оптимального обобщения ПарГеНМ**
5. Применение: анализ коллекций научных публикаций
6. Применение: рекламный таргетинг
7. Заключение

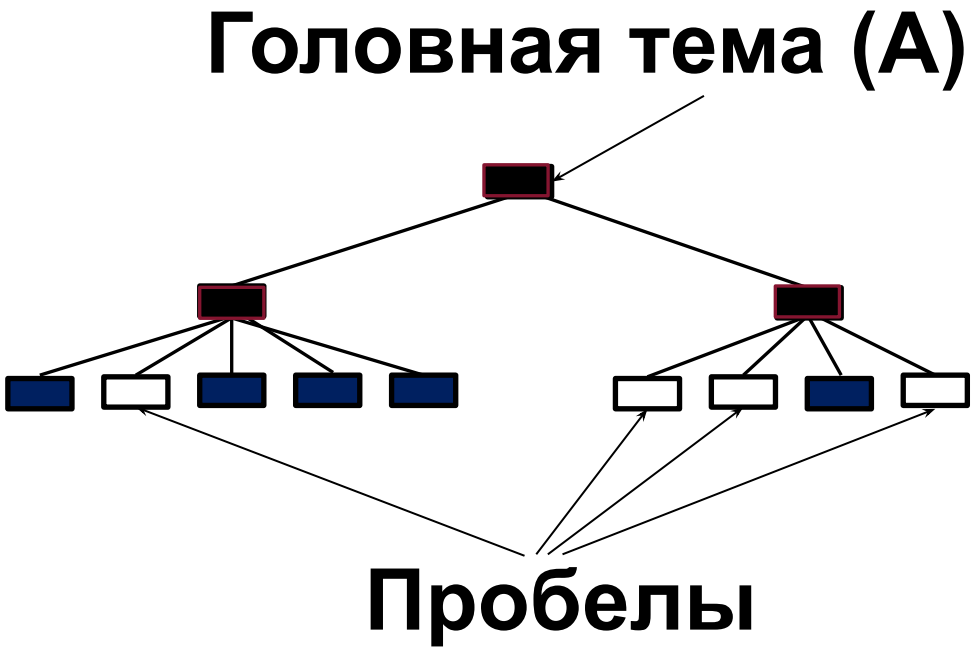
М4. Оптимальное обобщение нечетких кластеров в таксономии предметной области

Обобщение кластеров с помощью подъема в более общие вершины (Frolov, Mirkin, Fenner, Nascimento)

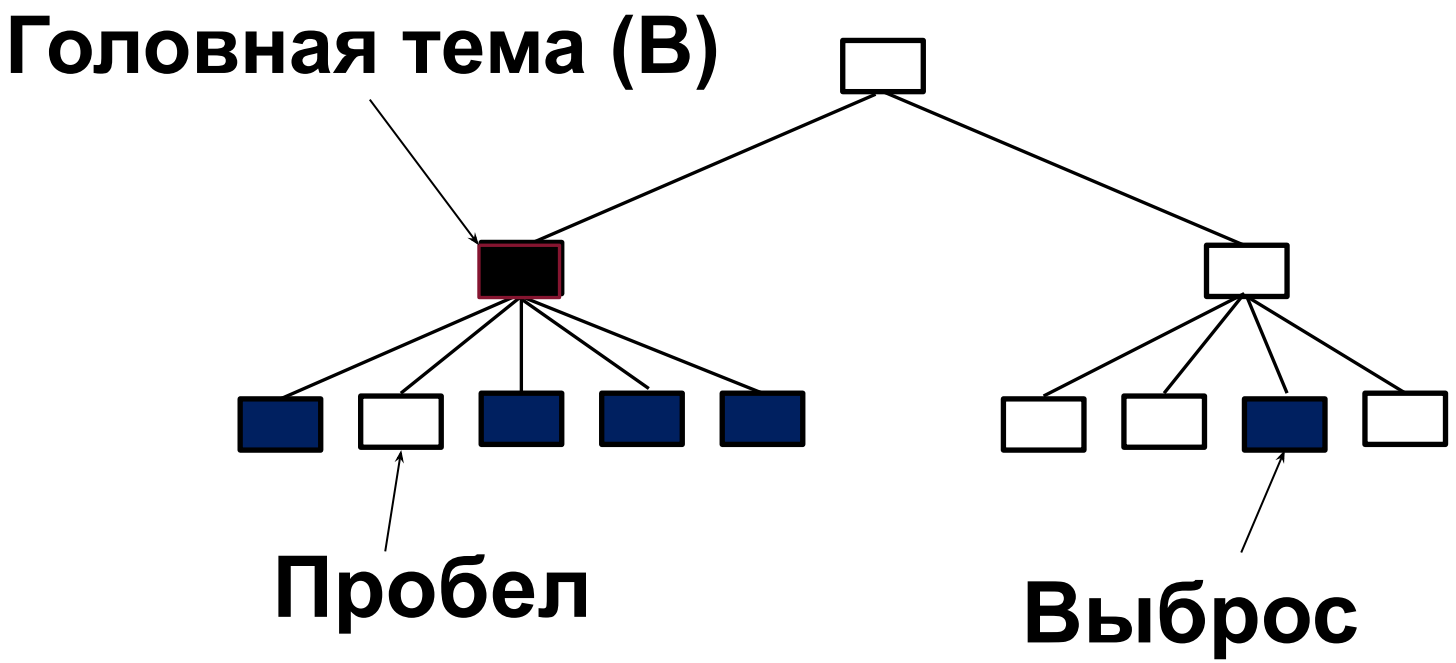


Если известна таксономия и получен кластер, в интерпретации удобно заменить листья на узлы более высокого уровня, если это возможно: **(A1, A2, A3, A4, B1) => (A)**, B1 - выброс.

Задача обобщения, 1/2



Задача обобщения, 2/2



Предложение: минимизировать суммарный штраф

|Головные темы| + λ|Пробелы| + γ|Выбросы|

Штраф за вариант A: $1+ 4\lambda$

Штраф за вариант B: $1+ \gamma + \lambda$

Алгоритм ПарГеНМ (Mirkin, Frolov, Fenner, Nascimento)

Находит множество головных вершин *H*, минимизирующее штраф

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

Веса: λ - за пробел, γ - за выброс, 1 за головную тему.

I - множество листьев таксономии, $\chi(h)$ - множество потомков h, $u(h)$ - функция принадлежности нечеткому множеству, $G(h)$ - множество пробелов вершины h, $v(h)$ - значимость пробела h, $V(h)$ - сумма значимостей всех пробелов вершины h

Алгоритм ПарГеНМ

Для каждого узла t таксономии алгоритм вычисляет два множества:

1. $H(t)$ – множество приобретенных узлов
2. $L(t)$ – множество потерянных узлов

и значение штрафа: $p(t)$. Алгоритм рекурсивно вычисляет эти множества и штрафы, начиная с листьев, пока не достигнет корня. В каждом узле t возможны два варианта: а) головная тема приобретается в t, б) головная тема не приобретается.

Алгоритм ПарГеНМ: Вариант а)

Головная тема возникает в *t*:

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

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

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

Алгоритм ПарГеНМ: Вариант б)

Головная тема возникает выше по дереву:

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

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

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

Из вариантов а) и б) выбирается тот, где значение штрафа *p(t)* меньше.

Структура доклада

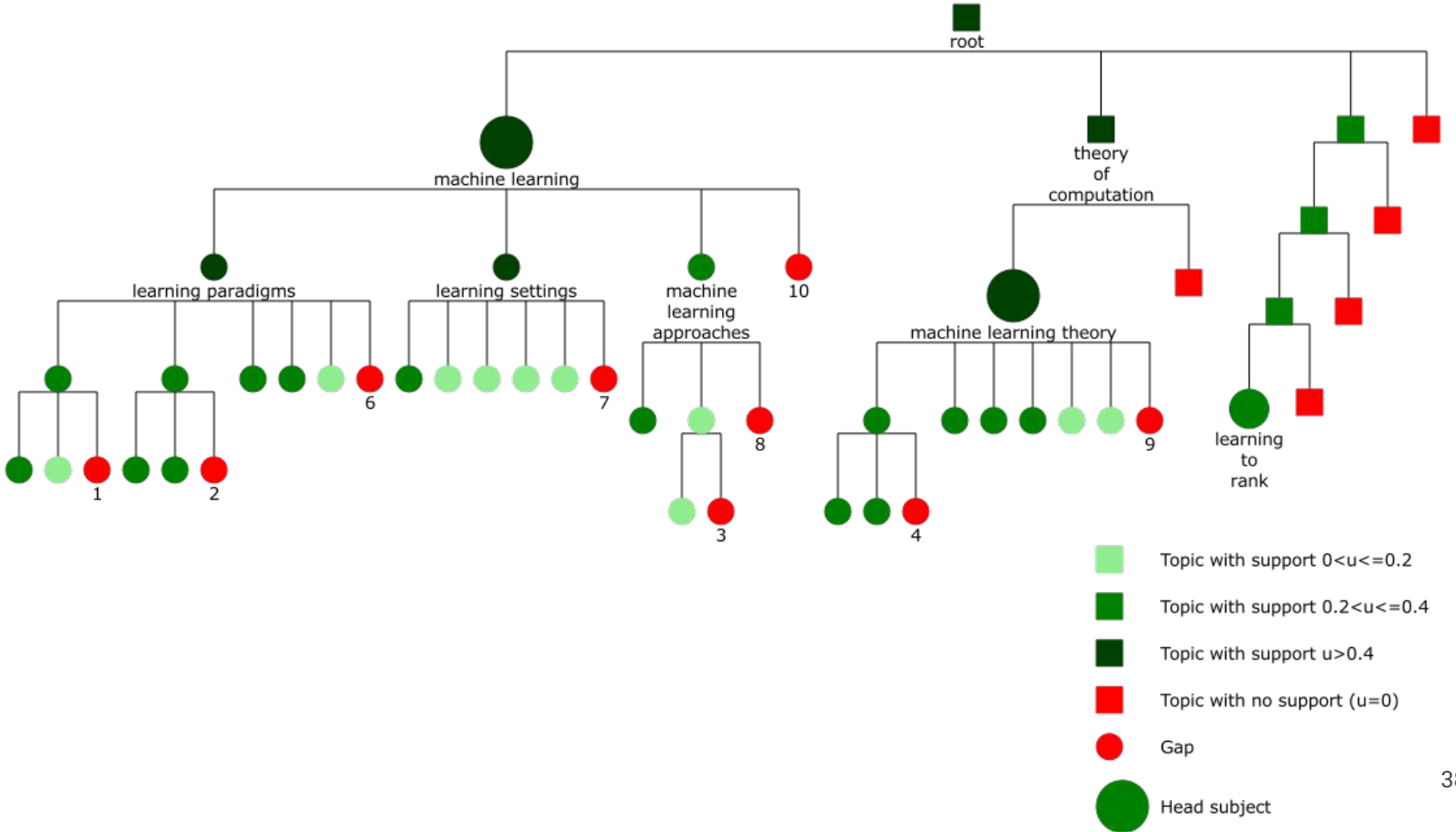
1. Актуальность, цель, новизна работы
2. Информационный поиск на основе аннотированных суффиксных деревьев
3. Методика интерпретации: подъем нечетких кластеров в таксономии предметной области
4. Алгоритм оптимального обобщения ПарГеНМ
5. **Применение: анализ коллекций научных публикаций**
6. Применение: рекламный таргетинг
7. Заключение

Результат обобщения кластера L “Learning”

Головные темы:

{Machine learning, Machine learning theory, Learning to rank}

Фрагмент дерева подъема кластера L “Learning”



Результат обобщения кластера R “Retrieval”:

Головные темы:

{Information Systems, Computer Vision}

М.5. Пример выводов: для кластера R

- Головные темы: (a) Information Systems, (b) Computer Vision
- Показывают тенденции
  - Работа с текстами
  - Сдвиг от текстов к изображениям и видео
  - Способы структурирования визуальной информации - предстоящий долгий путь к достижению семантической структуризации

Структура доклада

1. Актуальность, цель, новизна работы
2. Информационный поиск на основе аннотированных суффиксных деревьев
3. Методика интерпретации: подъем нечетких кластеров в таксономии предметной области
4. Алгоритм оптимального обобщения ПарГеНМ
5. Применение: анализ коллекций научных публикаций
6. **Применение: рекламный таргетинг**
7. Заключение

6. Применение ПарГеНМ в рекламном таргетинге (programmatic)

Таргетинг в интернет-рекламе по технологии programmatic осуществляется по сегментам таксономии:

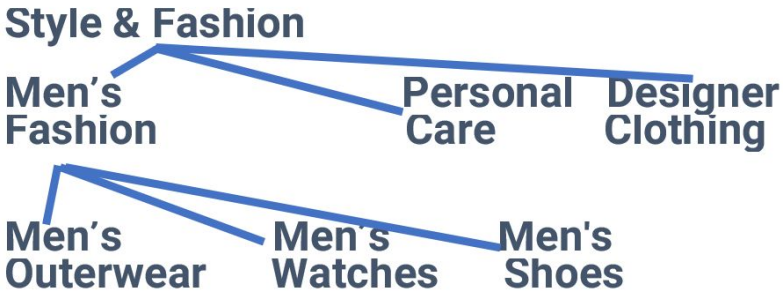
- Сегменты для пользователей (по их поведению)
- Сегменты для рекламных кампаний (по согласованию с рекламодателем)

Успешный таргетинг достигается при совпадении сегментов пользователя и кампании.

Таксономия Interactive Advertising Bureau (IAB)

Общепринятая таксономия сегментов: Interactive Advertising Bureau (IAB) Taxonomy

Фрагмент:



https://www.iab.com/guidelines/iab-quality-assurance-guidelines-qag-taxonomy/

Применение ПарГеНМ для эффективного расширения целевой аудитории, 1/2

Разработана методика обобщения сегментов на основе ПарГеНМ, названная ОПС, позволяющая расширять целевую аудиторию рекламы практически без потери качества: вместо ослабления порогов толерантности берется головная тема

Применение ПарГеНМ для эффективного расширения целевой аудитории, 2/2

Результаты рекламы программного продукта для родительского контроля. Методика позволила расширить аудиторию в 2,5 раза практически без потери качества по числу кликов.

Показатель	Классический таргетинг	**Таргетинг с использованием ОПС**	Таргетинг с использованием ППС
Impressions	378933	**942104 (+148.6% к классическому)**	1017598 (+168.5%)
Clicks	1061	**2544 (+139.8%)**	1526 (+43.8%)
CTR, %	0.28	**0.27 (-3.6%)**	0.15 (-46.4%)

Методика используется в коммерческой компании.

Структура доклада

1. Актуальность, цель, новизна работы
2. Информационный поиск на основе аннотированных суффиксных деревьев
3. Методика интерпретации: подъем нечетких кластеров в таксономии предметной области
4. Алгоритм оптимального обобщения ПарГеНМ
5. Применение: анализ коллекций научных публикаций
6. Применение: рекламный таргетинг
7. **Заключение**

7. Основные результаты

1. Разработан и экспериментально обоснован новый метод информационного поиска (АСДП).
2. Разработан и протестирован метод интерпретации результатов поиска с помощью оптимального обобщения нечетких кластеров в таксономии предметной области (ПарГеНМ).
3. Разработан и внедрен метод расширения аудитории интернет-рекламы на основе оптимального обобщения сегментов индивидуальных пользователей в таксономии предметной области (ОПС).

Публикации по теме диссертации, 1/3

Повышенный уровень:

1. D. Frolov, S. Nascimento, T. Fenner, B. Mirkin “Parsimonious Generalization of Fuzzy Thematic Sets in Taxonomies Applied to the Analysis of Tendencies of Research in Data Science”, Information Sciences. 2019. (WoS Q1) *To appear.*
2. D. Frolov, B. Mirkin, S. Nascimento, T. Fenner, “Method for Generalization of Fuzzy Sets”, Proceedings of the 18th International Conference on Artificial Intelligence and Soft Computing (ICAISC-2019), Springer Lecture Notes in Artificial Intelligence (LNAI). 2019. (Scopus Q2)
3. D. Frolov, B. Mirkin, S. Nascimento, T. Fenner, “Using Taxonomy Tree to Generalize a Fuzzy Thematic Clusters”, IEEE 2019 International Conference on Fuzzy Systems Proceedings. 2019. (CORE A)
4. D. Frolov, S. Nascimento, T. Fenner, Z. Taran, B. Mirkin “Computational Generalization in Taxonomies Applied to: (1) Analyze Tendencies of Research and (2) Extend User Audiences”. IDEAL 2019 Proceedings. (Scopus Q2) *To appear.*

Публикации по теме диссертации, 2/3

Стандартный уровень (Scopus, WoS, ВАК):

5. Frolov D.S., Annotated suffix tree as a way of text representation for information retrieval in text collections, Business Informatics. 2015.
6. D. Frolov, B. Mirkin, S. Nascimento, T. Fenner, Using Domain Taxonomy to Model Generalization of Thematic Fuzzy Clusters, IARIA Content 2019 Proceedings. 2019.
7. Dmitry Frolov, Zina Taran, Boris Mirkin, “A Method for Audience Extending in Programmatic Advertising by Using Parsimonious Generalization of User Segments”, Advances in Intelligent Systems and Computing, vol 1018. Springer, Cham.
8. D. Frolov, B. Mirkin, S. Nascimento, T. Fenner, “Globally Optimal Parsimoniously Lifting a Fuzzy Query Set Over a Taxonomy Tree”, Proceedings of the World Congress on Global Optimization, Springer, LNCS. 2019.

Публикации по теме диссертации, 3/3

9. Dmitry Frolov, Using Annotated Suffix Trees for Fuzzy Full Text Search, Communications in Computer and Information Science. Information Retrieval. RuSSIR 2016, Springer. 2016.

Препринт:

10. D. Frolov, B. Mirkin, S. Nascimento, T. Fenner, “Finding an appropriate generalization for a fuzzy thematic set in taxonomy”, Working paper WP7/2018/04, Moscow, Higher School of Economics Publ. House. 2018.

Доклады по теме диссертации, 1/3

1. ICAISC-2019, доклад на тему “Method for Generalization of Fuzzy Sets”, 16-20 июня 2019, г. Закопане, Польша.
2. World Congress on Global Optimization - 2019, доклад на тему “Globally Optimal Parsimoniously Lifting a Fuzzy Query Set Over a Taxonomy Tree”, 8-10 июля 2019, г. Мец, Франция.
3. IEEE 2019 International Conference on Fuzzy Systems, доклад на тему “Using Taxonomy Tree to Generalize a Fuzzy Thematic Clusters”, 25 июня 2019, г. Нью-Орлеан, США.
4. IARIA Content 2019, доклад на тему “Method for Generalization of Fuzzy Sets”, 5-9 мая 2019, г. Венеция, Италия.

Доклады по теме диссертации, 2/3

5. Общемосковский научный семинар “Математические методы анализа решений в экономике, бизнесе и политике”, доклад “Рубрикация коллекции документов с помощью формирования тематических нечетких кластеров и их оптимального подъема в таксономии предметной области”, 16 мая 2018, г. Москва.
6. 3-й Колмогоровский семинар по компьютерной лингвистике и наукам о языке, постерный доклад “Annotation of a Document Collection by Finding Thematic Fuzzy Clusters and Parsimoniously Lifting Them in a Domain Taxonomy”, 26 апреля 2018, г. Москва.
7. Общемосковский научный семинар “Математические методы анализа решений в экономике, бизнесе и политике”, доклад на тему “Обобщение в таксономиях: модель, метод, приложения”, 15 мая 2019, г. Москва.

Доклады по теме диссертации, 3/3

8. RuSSIR 2015, постерный доклад “Aggregate Text Representation for Information Retrieval in Collections of Text Documents”, 24-28 августа 2015, г. Санкт-Петербург.
9. Летняя школа Факультета компьютерных наук НИУ ВШЭ - 2016, постерный доклад “Using Annotated Suffix Trees for Fuzzy Full Text Search”, 27-29 мая 2016, п. Вороново. Московская обл.
10. RuSSIR-2016, постерный доклад: “Using Annotated Suffix Trees for Fuzzy Full Text Search”, 22-26 августа 2016, г. Саратов.
11. IHIET-2019, постерный доклад “A Method for Audience Extending in Programmatic Advertising by Using Parsimonious Generalization of User Segments”, 22-24 августа 2019, г. Ницца, Франция.

Спасибо за внимание!