Метод обобщения в таксономиях и его применение

Власов Александр Сергеевич

Научный руководитель: проф., д.т.н. Миркин Б.Г.

ФКН НИУ ВШЭ, Москва

29.05.2019

Постановка задачи

Дано:
a. Таксономия предметной области T.
b. Элемент предметной области E.

Каждый элемент предметной области принадлежит одному или нескольким листьям таксономии.

Обобщением назовем множество вершин (головных понятий), полностью покрывающее все листья, которым принадлежит этот элемент.
Проблема: найти оптимальное обобщение.

Актуальность

- В век больших данных задача структуризации и интерпретации текстовых коллекций крайне актуальна.
- Существующие методы включают в себя:
  - Кластеризацию
  - Тематическое моделирование (LDA и т.д.)
  - Суммаризацию текстов (выделение троек subject-verb-object, методы, основанные на глубинном обучении т.п.)
- **Проблема:** эти методы используют такой же уровень гранулярности, что и исходный текст, и, следовательно, не подразумевают обобщения!

Формализация задачи

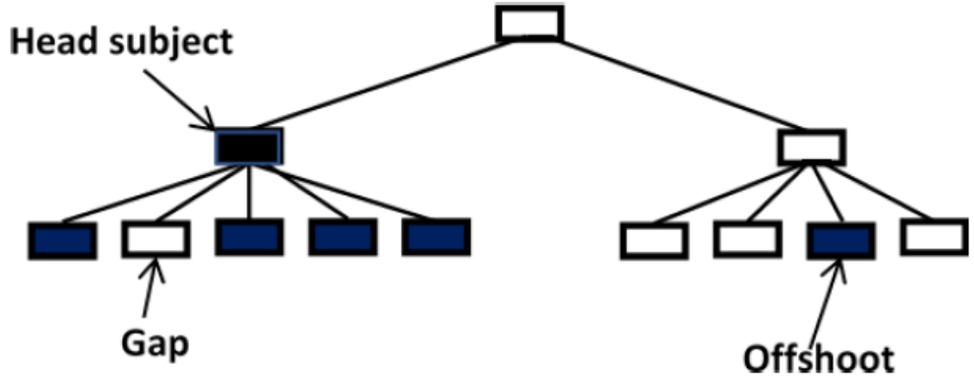
Дано: нечеткое множество S листьев таксономии.

Найти: множество вершин $t(S)$, которое как можно покрывает множество S так, чтобы:

- $t(S)$ имело как можно меньше элементов
- Минимизировать количество возникающих ошибок:
  - Непокрытых листьев из S (offshoot),
  - Ошибочно покрытых листьев, не принадлежащих S (gap)
- Иными словами: "объяснить" множество S наиболее *экономно*.

ParGenFS (Mirkin & Frolov, 2018)

Parsimonious Generalization of Fuzzy Sets



- Формализует и рекурсивно оптимизирует критерий максимальной экономичности (maximum parsimony):

$$p(H) = \sum_{h \in \text{heads}(H)} u(h) + \sum_{h \in \text{heads}(H)} \sum_{g \in G(h)} \lambda v(g) + \sum_{h \in \text{offshoots}(H)} \gamma u(h).$$

- Недостаток: необходимость выбора параметров γ и λ - штрафов за пропуски (gap) и выбросы (offshoot).

Задача дипломной работы

- Улучшить метод ParGenFS, изменив оптимизируемый критерий.
- Избавиться от необходимости ручного выбора штрафных коэффициентов.

Решение

Использовать метод максимального правдоподобия:

- ввести prior на вероятности приобретения и потери головных понятий
- новый критерий - максимизация вероятности сценария.

Предлагаемый алгоритм: MaLGenFS

Maximum Likelihood Generalization of Fuzzy Sets

1. Запуск ParGenFS для каждого элемента на используемом наборе данных.
2. Расчет вероятностей приобретения и потери головных понятий с помощью результатов последнего выполненного пункта.
3. Рекурсивное вычисление наиболее вероятного обобщения.

Повтор пунктов 2-3 до сходимости (вероятности в пункте 2 больше не меняются).

Таким образом, новый алгоритм уже *не зависит* от выбора гиперпараметров и не использует эвристический критерий максимальной экономности!

Определения

Возможные события в вершине:

- L - потеря головного понятия (loss),
- G - приобретение ГП (gain),
- P - отсутствие события (pass).

Для вершины таксономии t :

- **Сценарий** Sc_t - множество событий, произошедших в *поддереве* t.
  - Sc_t^I - сценарий при условии того, что вершина t *унаследовала* ГП от своего родителя
  - Sc_t^N - если вершина t *не унаследовала* ГП от своего родителя
- p_t^L, p_t^G - вероятности потери и приобретения головного понятия.

Алгоритм

Рассматриваем два возможных сценария:

I. Вершина унаследовала головное понятие от своего родителя, тогда:

$$p(Sc_t^I) = \max \begin{cases} p_t^L \prod_{w \in \text{children}(t)} p(Sc_w^N), & \text{(а)} \\ (1 - p_t^L) \prod_{w \in \text{children}(t)} p(Sc_w^I); & \text{(б)} \end{cases}$$

$$Sc_t^I = \begin{cases} \{L\} \cup \bigcup_{w \in \text{children}(t)} Sc_w^N, & \text{если (а)} \geq \text{(б)}, \\ \{P\} \cup \bigcup_{w \in \text{children}(t)} Sc_w^I, & \text{иначе;} \end{cases}$$

Событие в t должно максимизировать вероятность сценария в поддереве t.

В зависимости от вероятностей в сценарий добавляется событие L или P.

II. Вершина *не унаследовала* головное понятие от своего родителя, тогда:

$$p(Sc_t^N) = \max \begin{cases} p_t^G \prod_{w \in \text{children}(t)} p(Sc_w^I), & \text{(в)} \\ (1 - p_t^G) \prod_{w \in \text{children}(t)} p(Sc_w^N); & \text{(г)} \end{cases}$$

$$Sc_t^N = \begin{cases} \{G\} \cup \bigcup_{w \in \text{children}(t)} Sc_w^I, & \text{если (в)} \geq \text{(г)}, \\ \{P\} \cup \bigcup_{w \in \text{children}(t)} Sc_w^N, & \text{иначе.} \end{cases}$$

Для листьев дерева ($u_t \in \{0, 1\}$ - функция принадлежности в листе k):

$$p(Sc_t^I) = \max \begin{cases} 1 - p_t^L, & u_t = 1, \\ p_t^L, & u_t = 0; \end{cases} \qquad p(Sc_t^N) = \max \begin{cases} p_t^G, & u_t = 1, \\ 1 - p_t^G, & u_t = 0; \end{cases}$$

$$Sc_t^I = \begin{cases} \{P\}, & u_t = 1, \\ \{L\}, & u_t = 0; \end{cases} \qquad Sc_t^N = \begin{cases} \{G\}, & u_t = 1, \\ \{P\}, & u_t = 0. \end{cases}$$

Эксперимент

Основной набор данных для исследования:

- 26 799 аннотаций статей в области Data Science из 80 журналов издательств Springer и Elsevier за период 1971 - 2018 гг.

Название журнала	# Статей	# Томов	Период
Neurocomputing	3187	334	1992–2019
Expert Systems with Applications	2033	243	1998–2019
Procedia Computer Science	1933	139	2010–2019
Pattern Recognition	1360	301	1973–2019
Applied Soft Computing	1236	117	2003–2019
Information Sciences	1211	350	1998–2019
Pattern Recognition Letters	1001	292	1982–2019

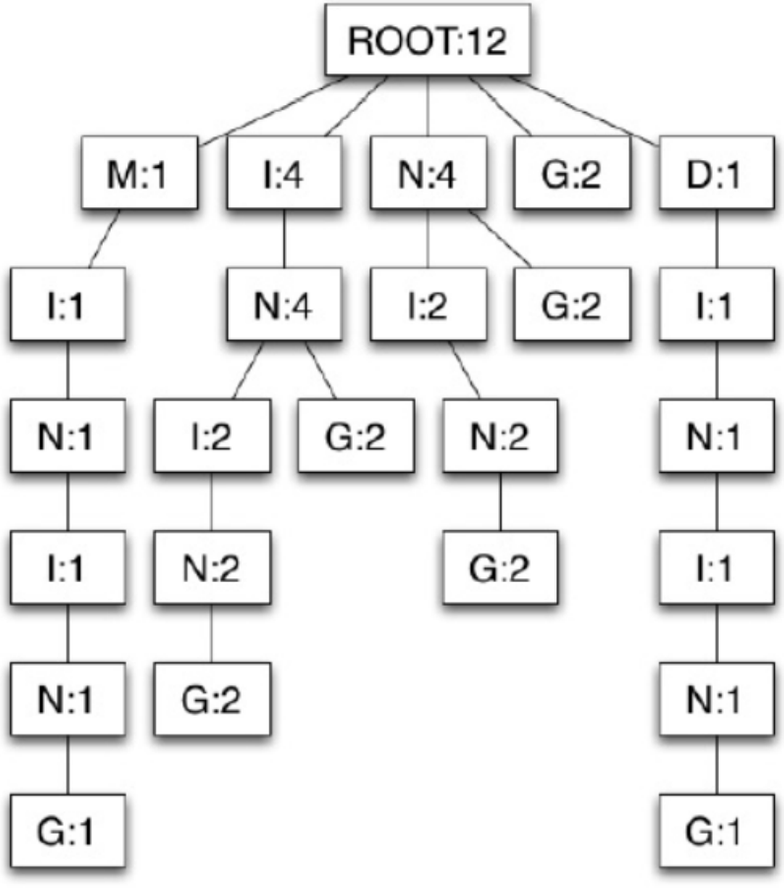
- Таксономия Data Science (Mirkin & Frolov), основанная на ACM Computing Classification System.

План эксперимента

1. Построить матрицу релевантности R "статья" $\rightarrow$ "листья таксономии" с помощью метода аннотированного суффиксного дерева (**AST**).
2. Построить матрицу корелевантности листьев таксономии $R^T R$.
3. Построить нечеткие кластеры на листьях таксономии с помощью метода **FADDIS**.
4. Подъем кластеров с помощью ParGenFS и вычисление статистик потерь/приобретений головных понятий $p^L,\ p^G$.
5. Инициализация MaLGenFS накопленными статистиками, повторный подъем кластеров.
6. Сравнение результатов, полученных с помощью ParGenFS и MaLGenFS.

Annotated Suffix Tree (Миркин и Черняк)

AST для строк "dining" и "mining":



Annotated Suffix Tree: оценка релевантности

1. Для каждой вершины u дерева T вычисляется условная вероятность:

$$p(u) = \begin{cases} \dfrac{f(u)}{f(\mathsf{parent}(u))}, & \mathsf{parent}(u) \neq R, \\ \dfrac{f(u)}{\sum_{v \in T:\ \mathsf{parent}(v)=R} f(v)}, & \mathsf{parent}(u) = R, \end{cases} \tag{13}$$

   где $f(u)$ — аннотация вершины u.

2. Для каждого k-суффикса строки x вычисляется коэффициент его релевантности тексту, хранимому в дереве T.

$$s(x^k, T) = \frac{1}{k_{max}} \sum_{i=1}^{k_{max}} p(x_i^k), \tag{14}$$

3. Релевантность строки x тексту, хранимому в T, вычисляется как среднее значение коэффициентов релевантности всех суффиксов строки:

$$S(x, T) = \frac{1}{N} \sum_{k=1}^{N} s(x^k, T). \tag{15}$$

LAPIN (Laplacian Pseudo-inverse Transform)

Предобработка матрицы схожести:

- $L_n = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$ - нормализованное преобразование Лапласа. D - диагональная матрица, $d_{tt} = \sum_{t' \in T} w_{tt'}$.
- $L_n^+ = Z \tilde{\Lambda}^{-1} Z^T$ - преобразование LAPIN.
  - Z - матрица собственных векторов, отвечающих ненулевым собственным значениям матрицы $L_n = Z \Lambda Z^T$,
  - $\tilde{\Lambda}$ - матрица, получаемая из матрицы Λ удалением нулевых значений на диагонали.

FADDIS (Mirkin & Nascimento)

Fuzzy Additive Spectral Clustering

$$r_{tt'} = \sum_{k=1}^{K} \mu_k^2 u_{kt} u_{kt'} + \varepsilon_{tt'},$$

Итеративный метод извлечения кластеров: минимизируем:

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

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

Эквивалентная задача - максимизировать отношение Релэя:

$$g(u) = \sqrt{G(u)} = \frac{u^T W u}{u^T u}.$$

FADDIS (Mirkin & Nascimento)

Решение в случае *безусловной* оптимизации: собственный вектор z, отвечающий максимальному собственному значению W.

Для задачи с ограничением $u : u_t \in [0, 1]$:

$$v_t = \begin{cases} 0, & z_t \le 0, \\ z_t, & 0 < z_t < 1 \\ 1, & z_t \ge 1. \end{cases}$$

Результаты: 7 наиболее репрезентативных кластеров (всего 35)

Интерпретация	Головные понятия и выбросы	Кол-во пропусков	Кол-во листьев в кластере
«Обучение» («Learning»)	1.1.1. – Machine learning theory 5.2. – Machine learning ⊙ 3.4.4.5. – Learning to rank	38	32
«Кластеризация» («Clustering»)	3.2.1.4. – Clustering ⊙ 1.1.1.3. – Unsupervised learning and clustering ⊙ 2.1.5.8. – Cluster analysis ⊙ 3.2.1.7.3 – Graph based conceptual clustering ⊙ 3.2.1.9.2. – Trajectory clustering ⊙ 3.4.5.8. – Clustering and classification ⊙ 5.2.1.2.1. – Cluster analysis ⊙ 5.2.3.2.5 – Kernel-based clustering ⊙ 5.2.4.3.1 – Spectral clustering	0	17

В сравнении с результатами, полученными на 18 000 статей (Mirkin & Frolov):

- Кластер **Learning** в точности совпадает.
- **Clustering** более плотный и содержит меньше ГП и выбросов.

Результаты

Интерпретация	Головные понятия и выбросы	Кол-во пропусков	Кол-во листьев в кластере
«Вероятностные представления» («Probabilistic representations»)	2.1.1. – Probabilistic representations 5.2.1.2. – Unsupervised learning 5.2.3.5. – Learning in probabilistic graphical models ⊙ 1.1.1.4.3. – Modelling ⊙ 1.1.1.6. – Bayesian analysis ⊙ 3.1.1.3.2. – Network data models ⊙ 3.3.1.4. – Web log analysis ⊙ 3.4.3.2. – Task models ⊙ 5.2.3.1.3 – Model trees ⊙ 5.2.3.13.1. – Deep belief networks ⊙ 5.2.3.7.2. – Factor analysis	11	31
«Извлечение» («Retrieval»)	3.1.4. – Query languages 3.4. – Information retrieval ⊙ 5.1.1.9. – Language resources	27	28

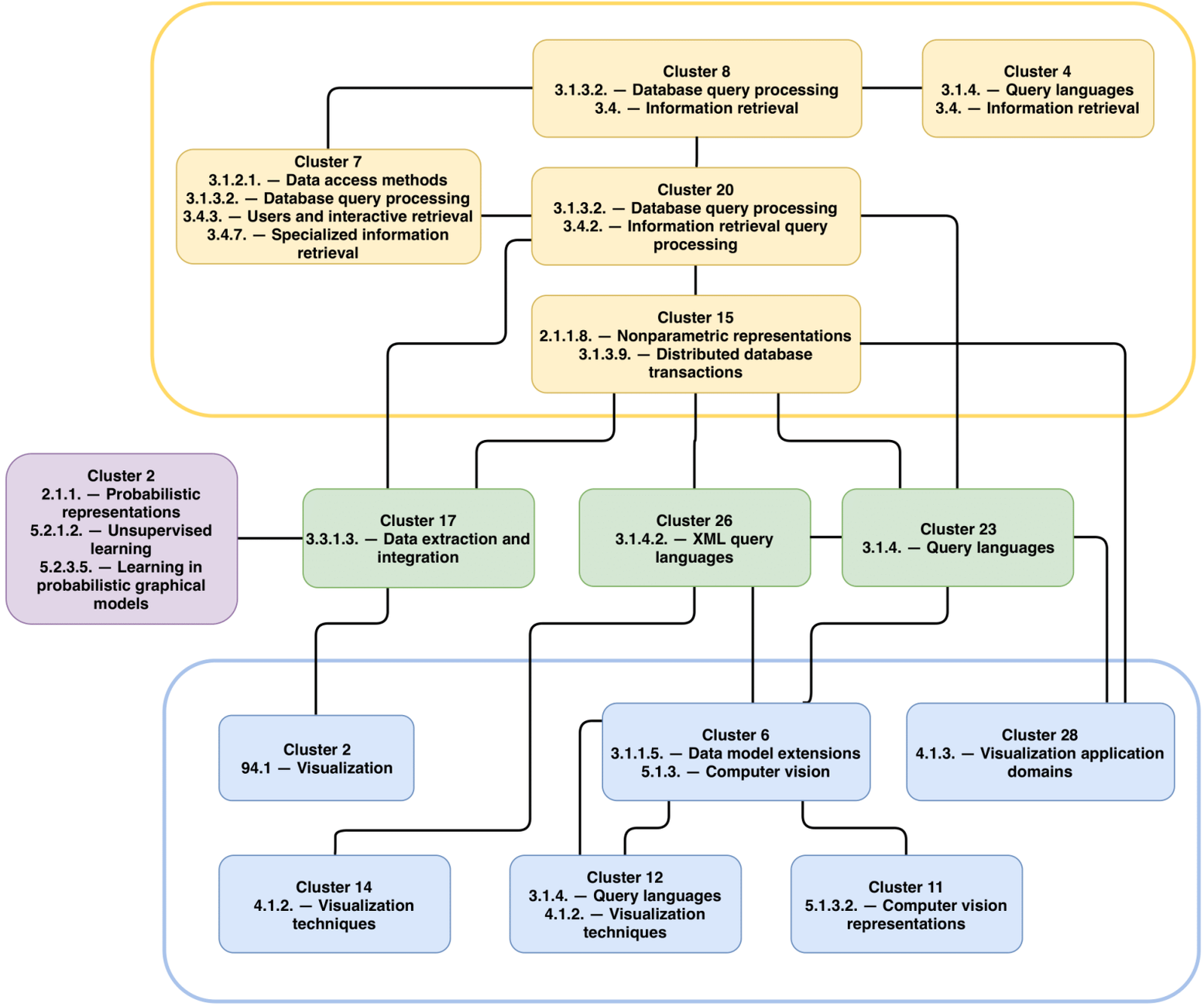
Результаты

«Структуризация» («Structuring»)	3.1.1.5. – Data model extensions 5.1.3. – Computer vision ⊙ 1.1.1.12. – Structured prediction ⊙ 1.1.2.10. – Logic and databases ⊙ 3.1.2.1.2. – Data scans ⊙ 3.1.3.3.3. – Database recovery ⊙ 3.1.3.7. – Database views ⊙ 3.1.4.1.1. – Structured Query Language ⊙ 3.1.5.9. – Federated databases ⊙ 3.2.1.4.5 – Feature weight clustering ⊙ 3.4.1.1. – Document structure ⊙ 3.4.2.1. – Query representation ⊙ 3.4.4.8. – Top-k retrieval in databases ⊙ 3.4.7.1.1. – Structured text search ⊙ 5.1.1.6. – Speech recognition ⊙ 5.2.1.1.5. – Structured outputs ⊙ 5.2.3.3.3.2 – Fuzzy representation ⊙ 5.2.3.6.2.1 – Tensor representation ⊙ 5.2.3.7.3.1 – 2D PCA	11	34

Результаты

Интерпретация	Головные понятия и выбросы	Кол-во пропусков	Кол-во листьев в кластере
«Представления в компьютерном зрении» («Computer vision representations»)	5.1.3.2. – Computer vision representations ⊙ 4.1.4.1. – Visualization toolkits ⊙ 5.2.3.3.3.2 – Fuzzy representation ⊙ 5.2.3.6.2.1 – Tensor representation ⊙ 5.2.3.7.3.1 – 2D PCA	0	13
«Запросы» («Querying»)	3.1.3.2. – Database query processing 3.4.2. – Information retrieval query processing ⊙ 2.1.5.1. – Queueing theory ⊙ 3.1.4.2.2. – XQuery ⊙ 4.1.2.5. – Dendrograms ⊙ 5.1.2.5. – Vagueness and fuzzy logic ⊙ 5.2.3.2.1.1 – Dynamic	3	15

Результаты: диаграмма пересечения кластеров



Основные источники

- The 2012 ACM Computing Classification System. Available: http://www.acm.org/about/class/2012
- Frolov, Mirkin, Nascimento, Fenner *Finding an appropriate generalization for a fuzzy thematic set in taxonomy*, 2018
- Nascimento, Felizardo, Mirkin *Laplacian Normalization for Deriving Thematic Fuzzy Clusters with an Additive Spectral Approach*, 2012
- Mirkin, Fenner, Koonin *Algorithms for computing parsimonious evolutionary scenarios for genome evolution...*, 2003
- Mirkin, Camargo, Fenner *Aggregating Homologous Protein Families in Evolutionary Reconstructions of Herpesviruses*, 2006
- Mirkin, Chernjak *Annotated Suffix Tree as a Way of String-To-Document Score Evaluating*, 2012

Метод обобщения в таксономиях и его применение

Власов Александр Сергеевич

Научный руководитель: проф., д.т.н. Миркин Б.Г.

ФКН НИУ ВШЭ, Москва

29.05.2019

Спасибо за внимание!

Полный текст работы доступен в моем репозитории на GitHub
(позже туда будут добавлены исходные коды всех программ)