Метод обобщения в таксономиях и его применение

Method for Appropriate Generalization in a Taxonomy

Выполнил:
Власов Александр Сергеевич
студент группы мНоД17-ИССА

Руководитель:
Миркин Борис Григорьевич
д.т.н. профессор

Национальный исследовательский университет
«Высшая школа экономики»

Факультет компьютерных наук

13 июня 2019

Содержание

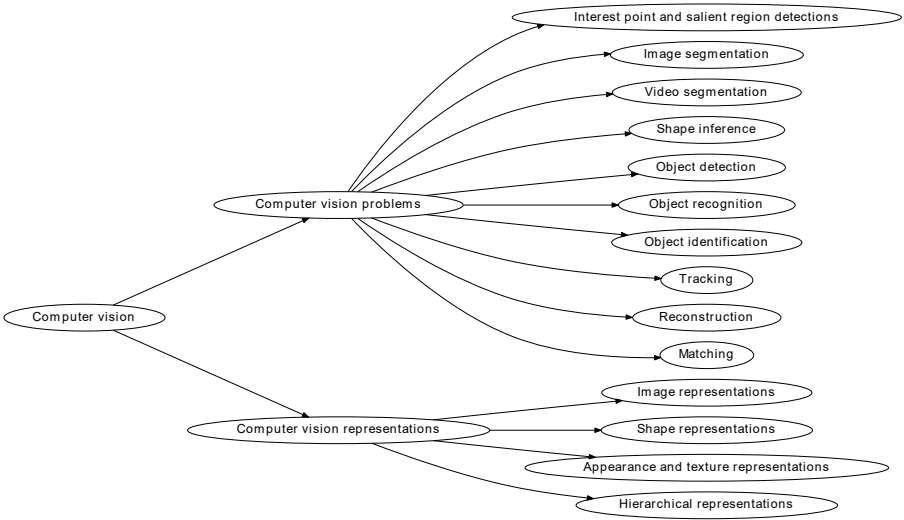
Цели работы

- Применить метод оптимального обобщения, разработанный Фроловым Д.С. и Миркиным Б.Г. [1], к анализу тенденций развития науки о данных.
- Предложить модификацию метода для критерия максимального правдоподобия.
- Провести экспериментальное исследование на расширенной текстовой коллекции.

[1] Finding an appropriate generalization for a fuzzy thematic set in taxonomy / Dmitry Frolov [и др.] // Series WP7 "Математические методы анализа решений в экономике, бизнесе и политике". — 2018. — Т.4.

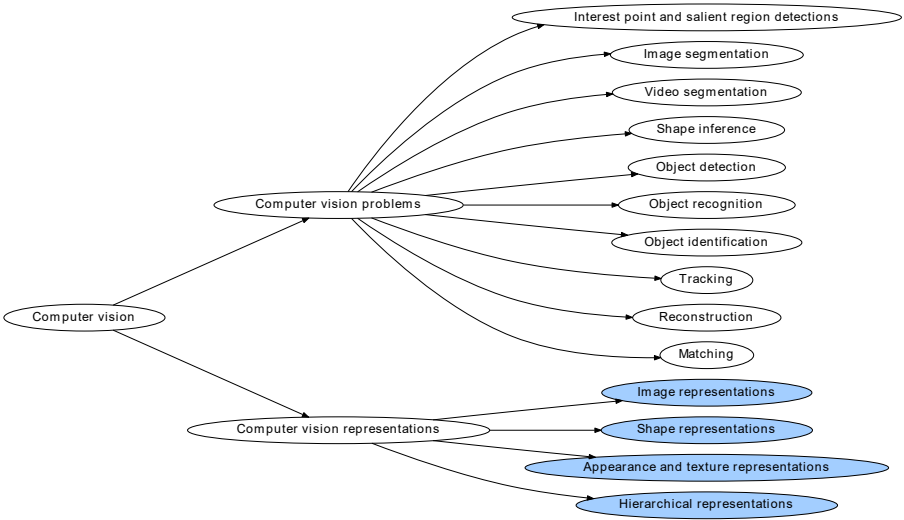
Таксономия и обобщение в таксономии

Фрагмент таксономии компьютерных наук ACM Computing Classification System 2012



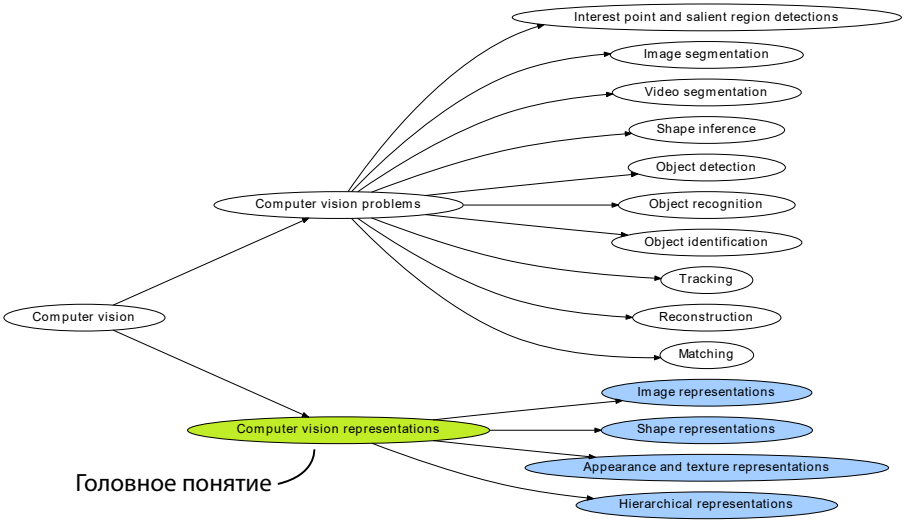
Таксономия и обобщение в таксономии

Тривиальный пример



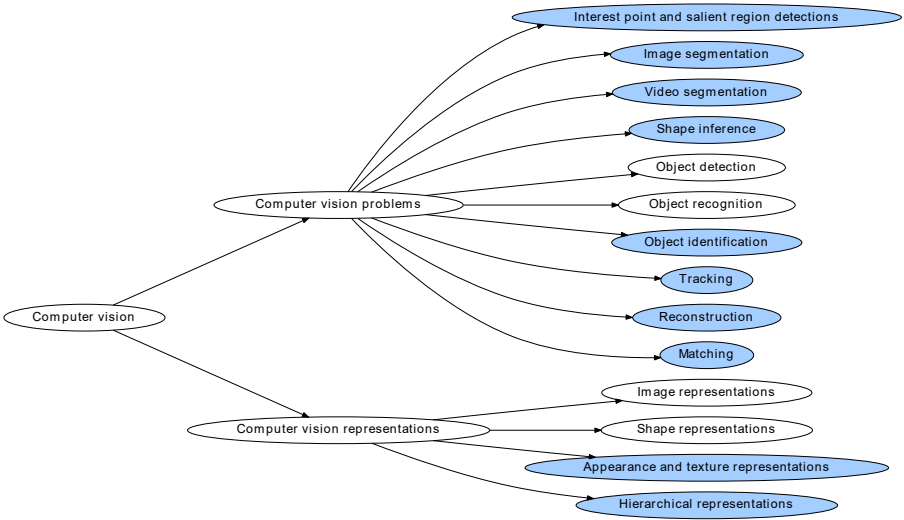
Таксономия и обобщение в таксономии

Тривиальный пример: обобщение



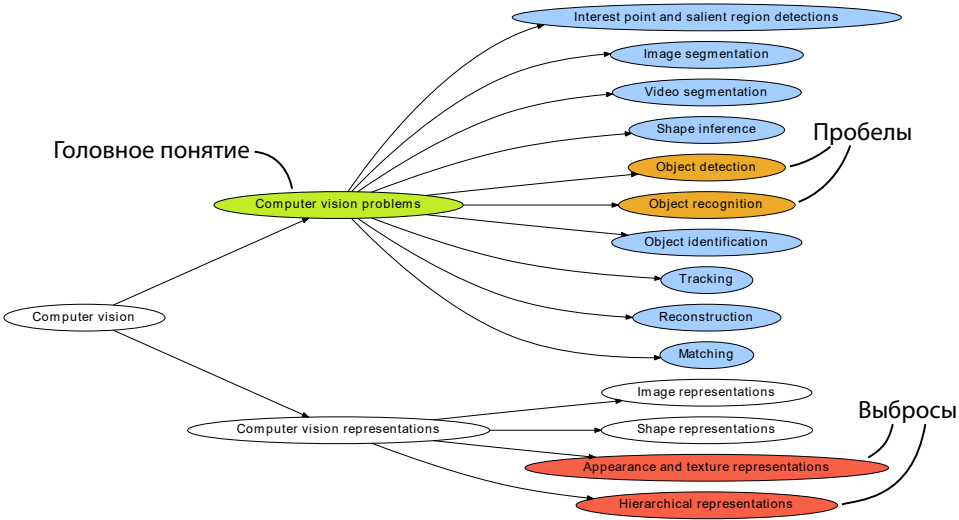
Таксономия и обобщение в таксономии

Нетривиальный пример



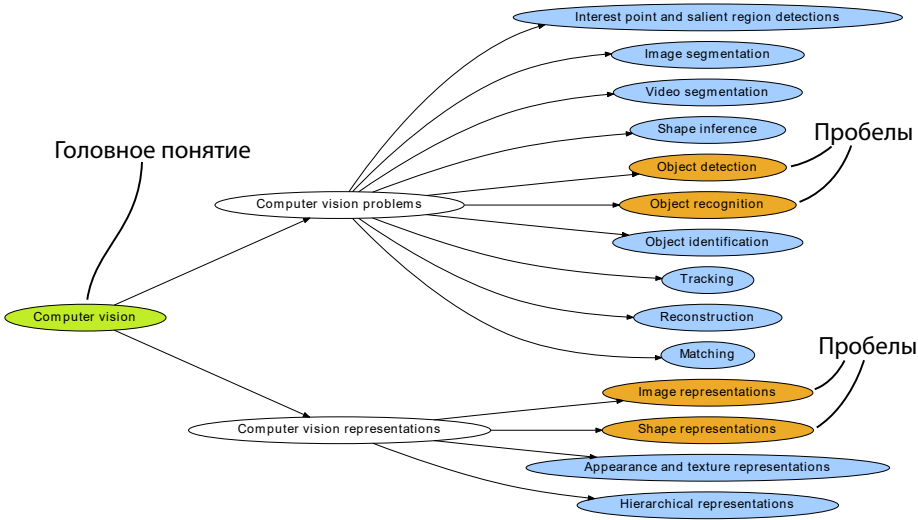
Таксономия и обобщение в таксономии

Нетривиальный пример: обобщение 1



Таксономия и обобщение в таксономии

Нетривиальный пример: обобщение 2



Выбор оптимального обобщения

Метод наибольшей экономии

Идея: оптимальное обобщение имеет *наименьшее* количество элементов (головных понятий, пробелов, выбросов).
Штрафная функция:

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

- $u(\cdot)$ — функция принадлежности, определенная на листьях таксономии,
- 1 — штраф за головное понятие
- γ — штраф за пробел,
- λ — штраф за выброс.

Выбор оптимального обобщения

Метод максимального правдоподобия (собственная разработка)

Идея: максимизировать вероятность **сценария**.

- Каждой вершине в соответствие ставится событие: приобретение, потеря или передача головного понятия.
- Сценарий: множество событий в каждой из вершин.

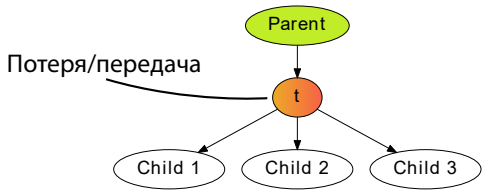
Особенности:

- Не требует явного задания штрафных коэффициентов.
- Использует априорные вероятности приобретений и потерь головных понятий в вершинах.

Выбор оптимального обобщения

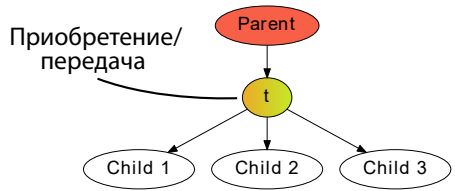
Наследование

I. Головное понятие *унаследовано* от родителя:



$$p(Sc_t^I) = \max \begin{cases} p_t^{\text{loss}} \prod_{w \in C(t)} p(Sc_w^N), \\ (1 - p_t^{\text{loss}}) \prod_{w \in C(t)} p(Sc_w^I); \end{cases}$$

II. Головное понятие *не унаследовано* от родителя:



$$p(Sc_t^N) = \max \begin{cases} p_t^{\text{gain}} \prod_{w \in C(t)} p(Sc_w^I), \\ (1 - p_t^{\text{gain}}) \prod_{w \in C(t)} p(Sc_w^N); \end{cases}$$

Выбор оптимального обобщения

Пример сценария

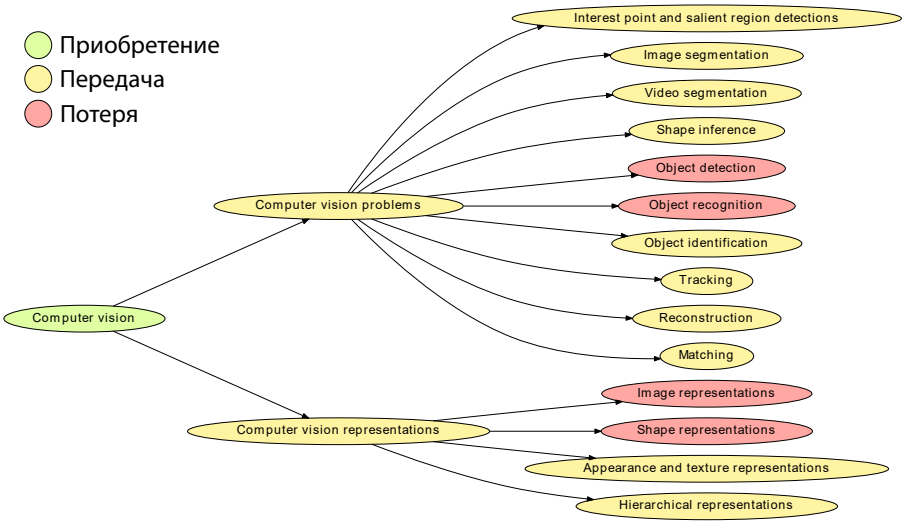


Схема применения метода обобщения

Коллекция текстов и таксономия науки о данных

- 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, основанная на ACM Computing Classification System:
  - максимальная глубина равна 7,
  - 456 вершин,
  - 353 листа.

Схема применения метода обобщения, 1
Тематические кластеры на листьях таксономии

Этапы расчетов:

1. Построение матрицы релевантности $\mathcal{R}$ текстов к листьям таксономии с помощью метода **аннотированного суффиксного дерева** (26799×353).
2. Расчет матрицы $\mathcal{C}$ корелевантности листьев таксономии как взвешенного матричного произведения $\mathcal{R}^T$ и $\mathcal{R}$ (353×353) .
3. Применение **псевдо-обратного преобразования Лапласа** (LAPIN).
4. Извлечение тематических кластеров на листьях таксономии с помощью метода **нечеткой аддитивной спектральной кластеризации** (FADDIS).

Схема применения метода обобщения, 2
Оптимальное обобщение

Полученные тематические кластеры обобщаются двумя методами:

1. Методом наибольшей экономии (алгоритм **ParGenFS**),
2. Методом максимального правдоподобия (алгоритм **MalGenFS**).

Далее:

1. Обобщенные кластеры интерпретируются,
2. Проводится сравнение с ранними результатами,
3. Методы обобщения (исходный и модифицированный) сравниваются между собой.

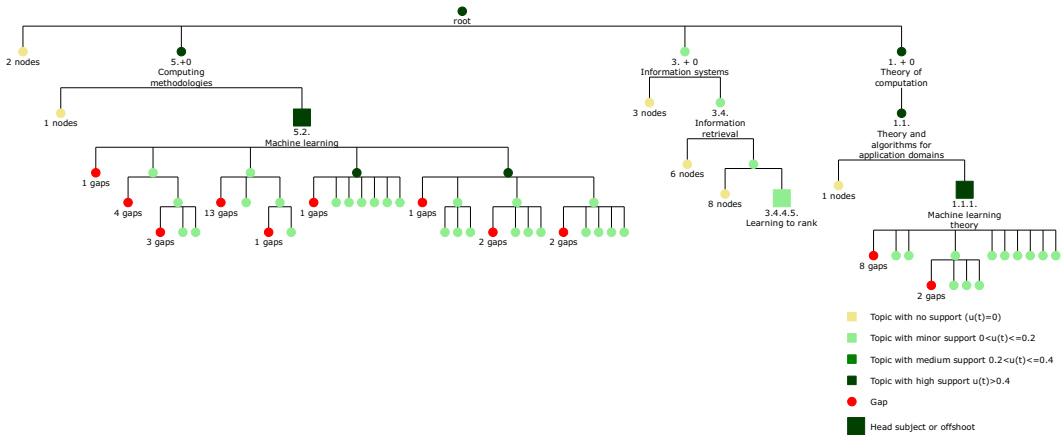
Результаты экспериментов

Тематические кластеры

- На 26 799 аннотациях статей получено 35 тематических кластеров, к ним применены методы обобщения.
- 7 обобщенных кластеров оказались хорошо интерпретируемыми:
  - Learning
  - Clustering
  - Probabilistic representations
  - Structuring
  - Computer vision representations
  - Retrieval и Querying
- Похоже на более ранние результаты Фролова и Миркина на 17000 статей:
  - **Learning**: в точности совпадает с полученным в данной работе.
  - **Clustering**: более плотный и содержит меньше головных понятий и выбросов.
  - **Retrieval**: разделен на несколько более маленьких кластеров.

Результаты экспериментов

Визуализация кластера Learning на таксономии



Результаты экспериментов

Сравнение методов оптимального обобщения

1. Исходный алгоритм: метод наибольшей экономии,
2. Модифицированный алгоритм: метод максимального правдоподобия.

Результат работы модифицированного алгоритма **практически полностью совпал** с исходным алгоритмом.

При этом модифицированный алгоритм:

- Работает только с *жесткими* кластерами.
- Не требует явного задания параметров штрафа за выбросы и пробелы.

Заключение

Основные результаты работы:

- Предложена модификация метода оптимального обобщения.
- Разработан комплекс программ для анализа текстовых коллекций: предобработки, кластеризации, отображения на таксономии, обобщения и визуализации.
- Обработано 26 799 аннотаций статей в области наук о данных.
- 7 кластеров проинтерпретированы в контексте тенденций развития наук о данных.
- Показано, что предложенный метод наиболее правдоподобного обобщения согласуется с методом максимальной экономии.
- Сделанные ранее выводы (Фролов и Миркин) подтверждаются и значительно детализируются.

Заключение

Дальнейшее развитие:

- Разработка метода автоматического расширения таксономии в рамках алгоритма обобщения.
- Адаптация обобщения методом максимального правдоподобия на случай нечетких кластеров.

Спасибо за внимание!

Метод обобщения в таксономиях и его применение

Выполнил:
Власов Александр Сергеевич
студент группы мНоД17-ИССА

Руководитель:
Миркин Борис Григорьевич
д.т.н. профессор

Национальный исследовательский университет
«Высшая школа экономики»

Факультет компьютерных наук

13 июня 2019

Диаграмма пересечения кластеров