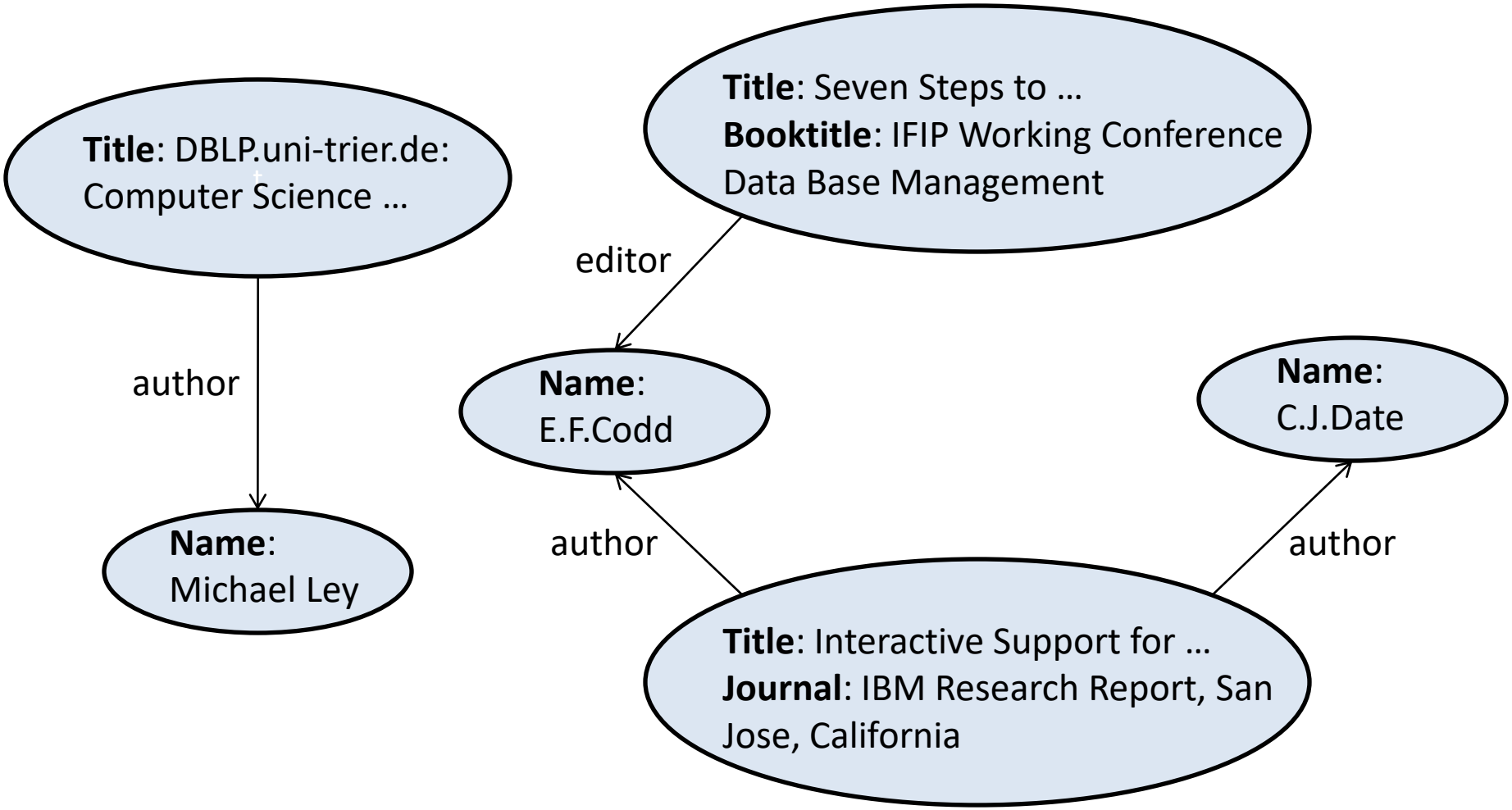
Выявление типов объектов в графовой базе данных на основе кластеризации

Соковикова Светлана, студентка 444 группы

Научный руководитель:
Д.ф.-м.н., проф. Новиков Б.А.

Рецензент:
Смирнов К.К.

Графовая база данных



Постановка задачи

Цель

- Создание инструментария для выделения структуры в слабоструктурированных данных

Задачи

- Разработка архитектуры инструментария
- Реализация модуля для выявления типов объектов в слабоструктурированных данных
- Реализация модуля конвертации слабоструктурированных данных в реляционную БД, в каждой таблице которой будут храниться данные об объектах одного типа
- Реализация модуля оценки качества кластеризации
- Апробация инструментария на базе данных научных публикаций DBLP

Актуальность задачи

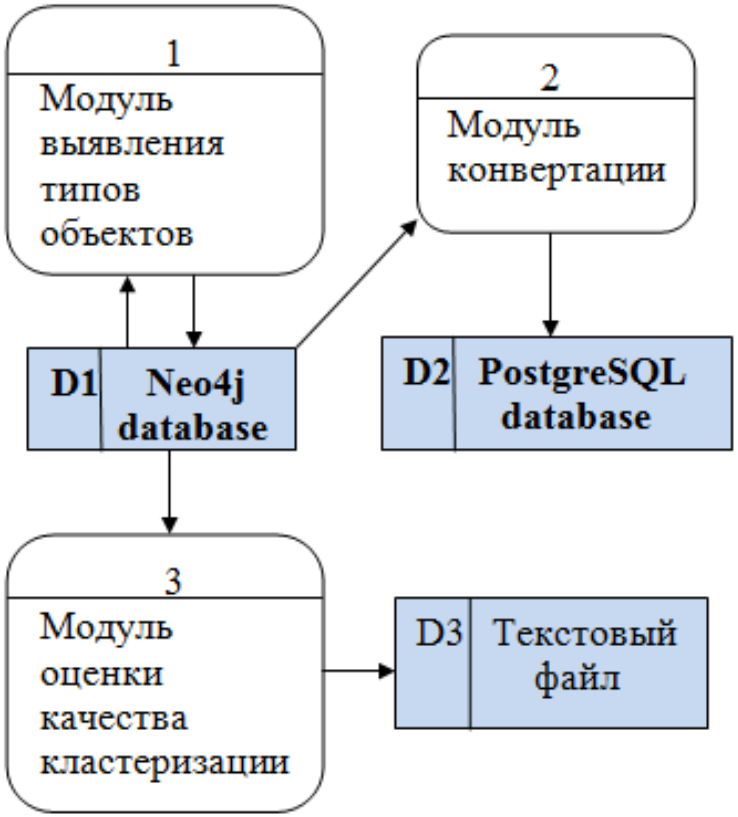
- **Семантическая паутина** (англ. Semantic web) — общедоступная глобальная семантическая сеть, формируемая на базе World Wide Web путём стандартизации представления информации в виде, пригодном для машинной обработки
- Linked data
- Находим структуру в слабоструктурированных данных
- Представляем эту структуру в реляционной базе данных

Подход к выявлению типов объектов

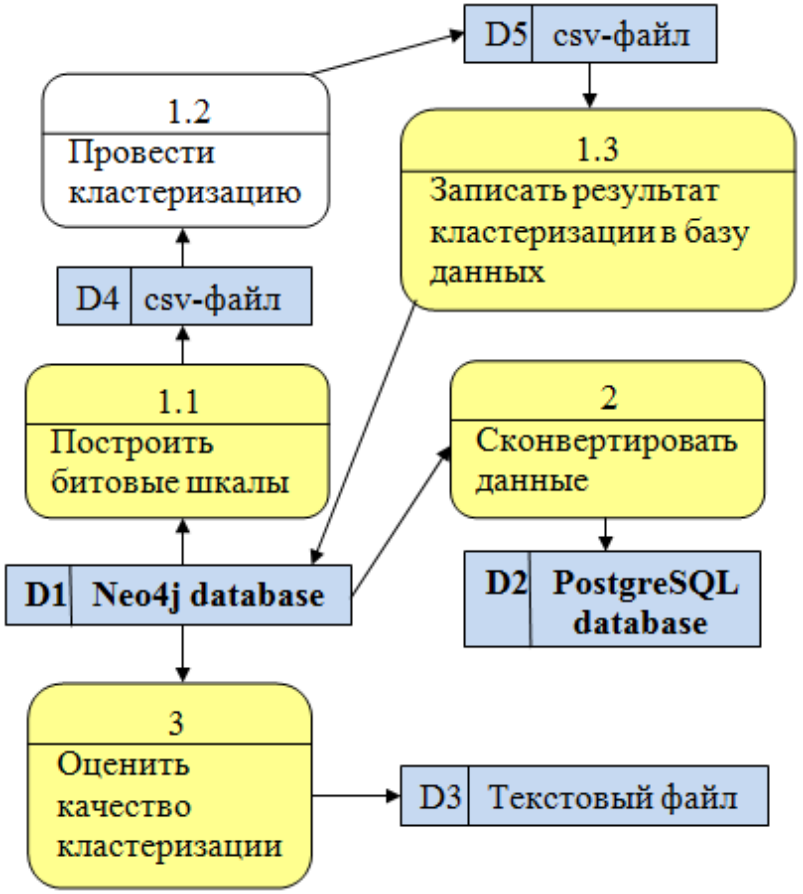
- Тип объектов считается одинаковым, если их наборы атрибутов «похожи»
- Упорядочиваются все атрибуты, какие есть в базе данных
- На каждый узел составляется битовая шкала: 1 – есть соответствующий атрибут, 0 – нет атрибута
- Кластеризуются полученные битовые шкалы

Диаграмма потоков данных

На уровне подсистем



На уровне процессов

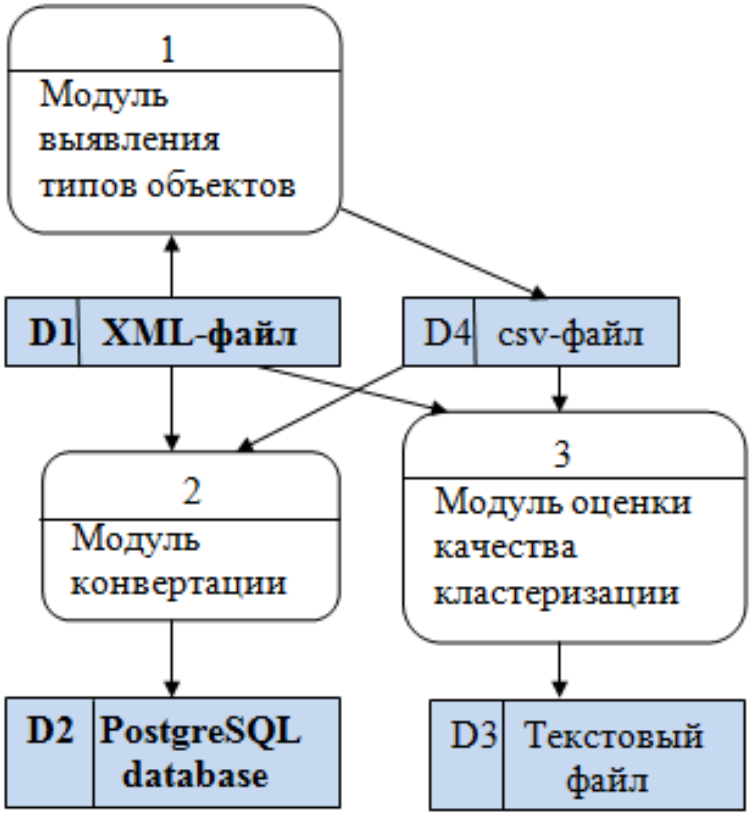


DBLP

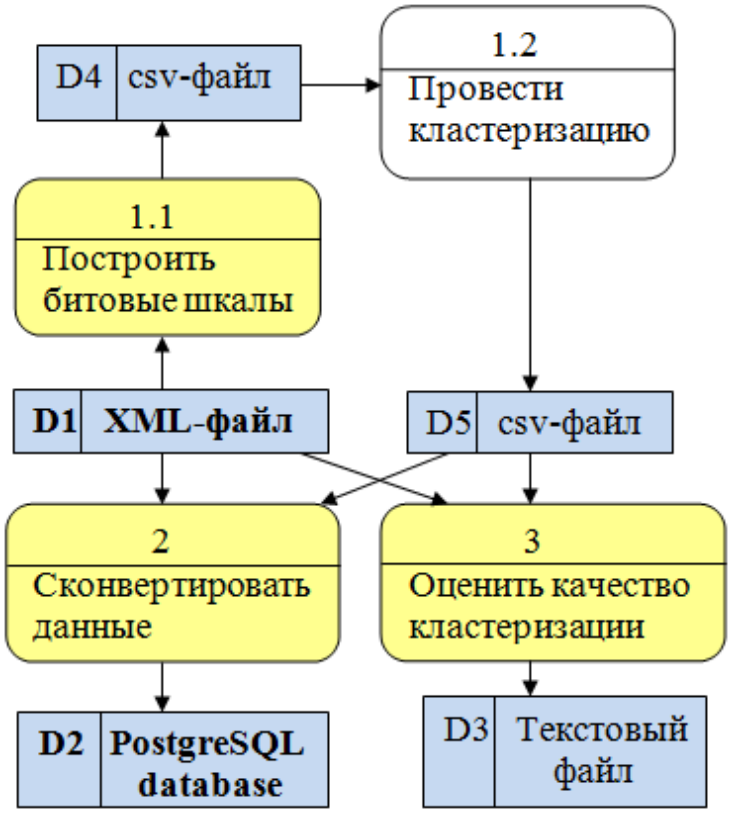
- Xml
- ~1.8 Gb
- ~5300000 публикаций
- Написана утилита для импортирования xml-файла в Neo4j базу данных
- Импортирование занимает неоправданно много времени (47 минут на 1% коллекции)
- Решение: расширить инструментарий
  - Получение необходимых данных из xml – файла без использования Neo4j

Диаграмма потоков данных из XML-файла

На уровне подсистем



На уровне процессов



Оценка качества кластеризации

- Рассматривается внешняя оценка качества

Идеальное разбиение

- Публикации - по типам, указанным в xml – файле (incollection, www, book, proceedings, inproceedings, phdthesis, article)

Кластеризация 1% публикаций

Номер кластера	incollection	www	book	proceedings	inproceedings	phdthesis	article
1	395	1	0	0	18833	0	0
2	0	18019	0	0	0	0	13
3	0	6	82	3	0	0	0
4	0	0	39	317	0	0	0
5	0	0	0	0	0	0	427
6	0	0	6	0	0	349	4
7	0	0	0	0	0	0	14889

Кластеризация 5% публикаций

Номер кластера	incollection	www	book	proceedings	inproceedings	phdthesis	article
1	1979	0	0	0	88488	0	0
2	0	90125	0	0	0	0	13
3	0	6	522	595	0	0	0
4	0	0	116	1006	0	0	0
5	0	1	0	2	5679	0	0
6	0	0	1	0	0	1748	4
7	0	0	0	0	0	0	76650

Кластеризация полной коллекции публикаций

Номер кластера	incollection	www	book	proceedings	inproceedings	phdthesis	article
1	20173	0	0	47	3502	0	0
2	0	1770051	77	10	0	0	521
3	0	1	10570	0	0	34962	6
4	3	5	2143	32007	3	0	3
5	19412	1	2	0	1879844	0	221
6	0	32589	1	0	0	0	13
7	0	0	4	0	0	0	1532590

Результаты

- Разработана архитектура инструментария
- Реализован модуль выявления типов объектов в слабоструктурированных данных
- Реализован модуль конвертации слабоструктурированных данных в реляционную БД PostgreSQL
- Реализован модуль оценки качества кластеризации
- Проведена апробация инструментария на базе данных научных публикаций DBLP