БЫСТРОЕ ПОСТРОЕНИЕ ГРАФОВОЙ БД WEB-САЙТА. ГРАФ ЗНАНИЙ КАК СРЕДСТВО ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ИНТЕРНЕТ ИСТОЧНИКА

Зорко П. А., Кулевич А. О.

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Пилецкий И. И. – канд. физ.-мат. наук, доцент

Приводится описание архитектурных решений быстрого построения компонента тематического прототипа графической БД, графа знаний из открытых интернет-источников с целью глубокого анализа данных сайта, выявления скрытых зависимостей в некоторой научной области. Описываются принятые решения, демонстрируются, результаты работы компонента получение данных с web-сайта.

Одна из больших технических задач — это получение данных из конкретного сайта [1]. Но, еще более сложная проблема — это анализ данных размещенных на данном сайте. Что приводит к появлению ряда новых возможностей моделирования структуры сайта, одна из которых - возможность предложить простой способ представления свойств отношений в виде графовой модели. Что позволяет анализировать свойства сайта с помощью графовых моделей. Многие сайты формально поддерживает тройки RDF (Resource Description Framework), в которых тройка имеет вид «субъект — предикат — объект» и называется триплетом (тройкой). Множество RDF-утверждений образует ориентированный граф, в котором вершинами являются субъекты и объекты, а рёбра отображают отношения. С помощью графовых технологий можно преобразовать представление сайта в графовую БД со свойствами.

На сайте Open Food Facts [2] имеется общедоступный набор данных с 6,2 миллионами троек по пищевым продуктам, их ингредиентам, аллергенам, фактам питания и многому другому. Импортируя тройки RDF получается тематическая графовая БД.

Общее представление полученной БД продемонстрировано на рисунке 1.

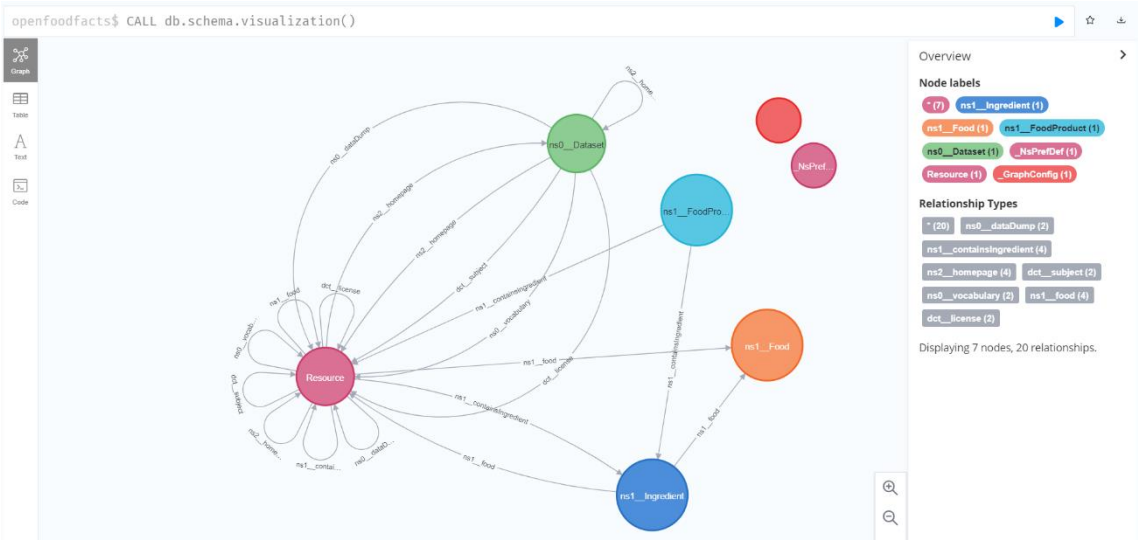

Рисунок 1 – Общее представление схемы графовой базы данных

В данной БД основными типами узлов являются:

- FoodProduct – продукты;
- Food – ингредиенты, которые могут содержаться в продуктах;
- Ingredient – промежуточные узлы, соединяющие узлы FoodProduct и Food, в которых указывается, сколько ингредиента содержится в продукте.

Основными типами связей являются:

- containsIngredient – соединяет продукт FoodProduct и промежуточный узел Ingredient;
- food – соединяет промежуточный узел Ingredient и ингредиент Food.

В полученной базе данных можно искать различную информацию о продуктах и составе их ингредиент. К примеру, с помощью запроса, представленного ниже, можно узнать, какой набор общих ингредиентов у двух продуктов. Результат запроса показан на рисунке 2.

```
MATCH (prod1:Resource { uri: 'http://world-fr.openfoodfacts.org/produit/5053827196642/coco-pops-kellogg-s'})
MATCH (prod2:Resource { uri: 'http://world-fr.openfoodfacts.org/produit/5010358227290/2-snack-pork-pies-spar'})
MATCH (prod1)-[:ns1__containsIngredient]->(x1)-[:ns1__food]->(sharedIngredient)<-[:ns1__food]-(x2)<-[:ns1__containsIngredient]-(prod2)
RETURN prod1, prod2, x1, x2, sharedIngredient
```

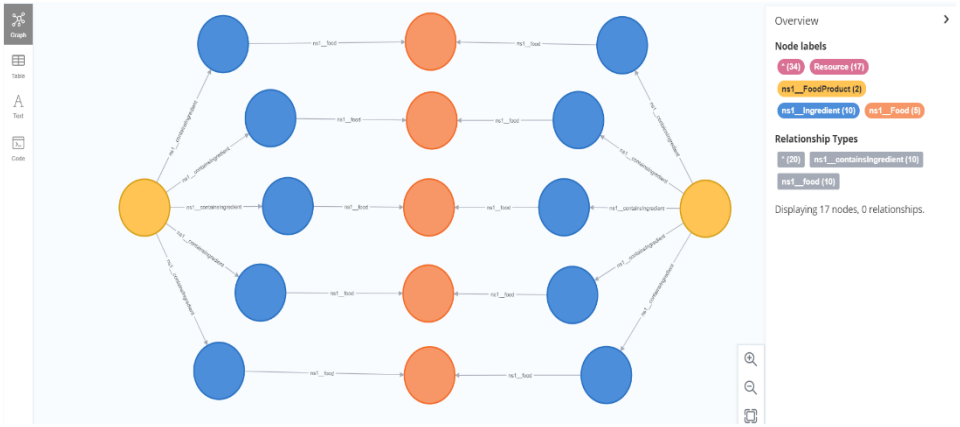

Рисунок 2 – Вывод общих ингредиентов у двух продуктов

Используя данный граф знаний, мы легко можем продемонстрировать 15 самых популярных ингредиентов, которые используются в продуктах. Для этого нам понадобится рассмотреть только узлы FoodProduct, Ingredient, Food и связи между ними [3].

Следующий запрос используется для подсчета использования каждого ингредиента и выводит топ 15 из них:

```
MATCH (:ns2__FoodProduct)-[:ns2__containsIngredient]->()-[:ns2__food]-(i:ns2__Food)
RETURN i.ns2__name as name, count(*) as Popularity
ORDER BY Popularity DESC
LIMIT 15
```

Сохраним полученные данные в файл формата csv, чтобы в дальнейшем их использовать.

Используя возможности языка Python, построим гистограмму популярности ингредиентов. Результат представлен на рисунке 3.

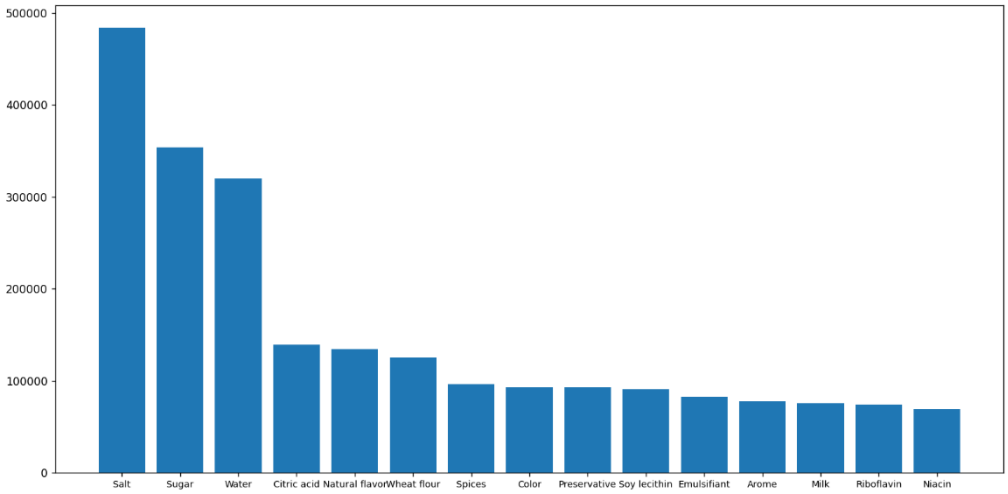

Рисунок 3 – Гистограмма популярности ингредиентов

Результатом научной работы является граф знаний продуктов питания Open Food Facts, к которому в дальнейшем можно применять различные алгоритмы машинного обучения. К примеру, разбить ингредиенты на «плохие» и «хорошие», а затем классифицировать продукты на полезные и нет.

Список использованных источников:

1. Ян Робинсон, Джим Вебер, Эмиль Эифрем. Графовые базы данных: новые возможности для работы со связанными данными / пер. с англ. Р.Н. Рагимова; науч. ред. А.Н, Киселев. – 2-е изд. – М.: ДМЛ Пресс, 2016. – 18 с.

2. Open Food Facts [Электронный ресурс] / Режим доступа: https://world.openfoodfacts.org Дата доступа: 25.03.22.
3. Neo4j [Электронный ресурс] / Режим доступа: https://neo4j.com/developer/graph-data-science/graph-algorithms/ Дата доступа: 10.02.22