СОЗДАНИЕ БАЗЫ ЗНАНИЙ ИЗ НАУЧНЫХ СТАТЕЙ

Ю.А. Иванова, к.т.н., доц.,
М.А. Хайров[*]*, студент гр. 8ВМ03*
Томский политехнический университет
E-mail: mah9@tpu.ru

Введение

Исследование научной литературы является достаточно долгим и трудоёмким процессом. К тому же, всегда есть риск оставить без внимания тот или иной аспект исследуемого объекта. Исходя из сказанного, появляется спрос на автоматизацию процесса научного поиска и написание научной статьи.

Целью данной работы является создание базы знаний для системы умного научного поиска с возможностью генерации обзорной научной статьи с построением рисунков, таблиц, химических и математических формул по ключевым словам.

Описание алгоритма

Разработку системы умного поиска можно разделить на четыре крупных этапа:

1. разработка системы извлечения данных из научных статей;
2. создание базы знаний;
3. разработка алгоритмов генерации обзорных научных статей;
4. создание веб-сервиса.

В данной работе будут рассматриваться результаты первого и второго этапов. Здесь возникли следующие задачи:

1. извлечение текстовой информации научных статей;
2. извлечение нетекстовой информации.

Под «текстовой информацией» подразумеваются заголовки, абзацы, аффилиация, ссылки и контакты.

Под «нетекстовой информацией» подразумеваются рисунки, таблицы, математические и химические формулы. Также при выделении рисунков и таблиц, необходимо было получить их описания, для таблиц содержание их ячеек в структурированном виде (в формате json), для формул – номер формулы (при наличии).

Основы работы

Исходный набор данных представляет из себя коллекцию статей в формате PDF с наличием текстового слоя.

Структура данных в конечной системе будет в виде MAG/OAG (графовидная структура) с возможностью добавления новых данных. Финальный объём базы знаний оценивается в 100 Тб.

Было принято решение получать текстовую информацию, используя инструменты библиотеки машинного обучения GROBID (или Grobid, расшифровывается как GeneRation Of Bibliographic Data [1], что переводится как генерация библиографических данных), для извлечения данных (извлечение тестовой информации ссылок, колонтитулов, координат объектов документов и пр.) из документов PDF и дальнейшей её структуризации.

Всё же, для решения задачи получения «нетекстовой информации» система GROBID показала неудовлетворительные результаты для всех объектов кроме формул на отсканированных статьях. Инструменты GROBID систематически неверно идентифицировали объекты и описания к ним, а по полученной содержательной части таблицы сложно было восстановить её исходную структуру.

Исходя из сказанного выше, для повышения качества результата было принято решение написать свой обработчик таблиц и рисунков.

Извлечение рисунков

Алгоритм извлечения рисунков из статей имеет следующий общий вид:

Шаг 1. Детектирование объектов на изображении страницы;

Шаг 2. Сортировка объектов по расположению на странице;

Шаг 3. Для каждого объекта изображения выполняются шаги 4-6;

Шаг 4. В зависимости от класса и относительного расположения объектов для рисунка происходит поиск блока с описанием и объединяется с блоком рисунка (Рисунок 1);

Шаг 5. Производится распознавание текста описания по совместному изображению, полученному на шаге 5;

Шаг 6. Изображение и его описание сохраняются в базе данных по его идентификатору.

Для детектирования изображений и таблиц, а также распознавания содержимого последних используется модель mask_rcnn_X_101_32x8d_FPN_3x [2], обученная на датасете PubLayNet [3], входящая в состав библиотеки Layout Parser. Данная модель предназначена для детектирования таких классов, как текст, заголовок, список, таблица и рисунок.

Для распознавания текста на шаге 5 также используются инструменты агент TesseractOCR, встроенный в Layout Parser. В качестве идентификатора изображения использовался его номер. Например, Figure n преобразовалось бы в figure_n.

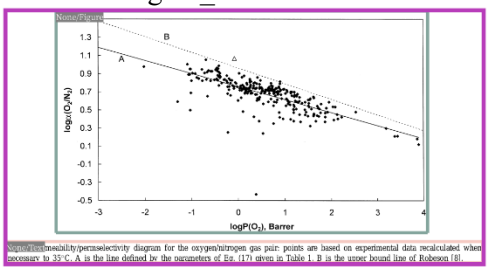

Рис. 1. Объединение блоков рисунка и описания для распознавания:
зелёный прямоугольник – рисунок; розовый – описание к рисунку; фиолетовый – объединение блоков

Извлечение таблиц

Извлечение таблиц и их описаний происходит по такому же алгоритму, главное отличие заключается в том, что у таблиц описание находится строго сверху.

После получения изображения таблицы происходит извлечение информации из самой таблицы.

Алгоритм для получения данных таблиц имеет следующий вид:

Шаг 1. таблица делится на две части – по горизонтальным линиям с нахождением границы раздела при помощи метода Хафа;

Шаг 2. распознавание оглавления таблицы при помощи агента TesseracOCR;

Шаг 3. агрегация распознанных блоков и организация мультииндексов;

Шаг 4. распознавание тела таблицы;

Шаг 6. организация элементов тела таблицы в соответствии с расстоянием и координатами столбцов;

Шаг 7. объединение таблицы.

Организация данных научных статей в базу знаний

С использованием разработанных алгоритмов, работающих с функционалом GROBID и Layout Parser, был произведён сбор данных научных статей.

Было обработано полностью 32 журнала, связанных с химией и химической технологией.

На рисунке 2 условно представлен вид графа знаний (точнее его ветвь).

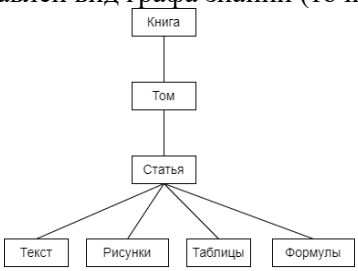

Рис. 2. Ветвь разработанного графа знаний

Заключение

В результате проведённой работы была подготовлена база знаний для создания прототипа системы интеллектуального поиска и генерации обзорных статей по ключевым словам. К созданию прототипа было переработано около 17 тыс. научных статей, объём базы знаний прототипа составляет 142 Гб. Далее планируется разработка самой системы когнитивного поиска.

Список использованных источников

1. GROBID Documentation [Электронный ресурс]: https://grobid.readthedocs.io [сайт]. Режим доступа: https://grobid.readthedocs.io/en/latest/., свободный (дата обращения 21.12.2021)
2. K. He, G. Gkioxari, P. Dollar, and R. Girshick. Mask ´ R-CNN. arXiv:1703.06870, 2017.
3. ibm-aur-nlp/PubLayNet [Электронный ресурс]: https://github.com [сайт]. Режим доступа: https://github.com/ibm-aur-nlp/PubLayNet., свободный (дата обращения 21.12.2021)