ПОИСК ВЗАИМОСВЯЗЕЙ МЕЖДУ КОНЦЕПТАМИ С ИСПОЛЬЗОВАНИЕМ ГРАФА ЗНАНИЙ

Пашук А. В., Гуринович А. Б., Волорова Н. А., Смирнов В. Л.
Кафедра информатики, кафедра вычислительных методов и программирования, кафедра информатики, кафедра вычислительных методов и программирования, Белорусский государственный университет информатики и радиоэлектроники
Минск, Республика Беларусь
E-mail: {pashuk, gurinovich, volorova, pom_rektora}@bsuir.by

С каждым годом увеличивается количество публикуемых научных статей в области биомедицины. Это приводит к тому, что исследователям становится все сложнее быть в курсе текущего положения дел даже в своей специализации. Это приводит к замедлению исследований и отсутствию новых открытий.

Введение

Автоматический поиск взаимосвязей в больших объемах информации сегодня является одной из актуальных задач в области анализа данных.

I. Получение взаимосвязей из научных статей

Для получения научных концептов и взаимосвязи между ними для построения графа знаний используется два источника, разработанных в ходе исследования:

- Приложение для автоматического поиска концептов и взаимосвязей между ними на основе семантического анализа текстов опубликованных статей. Это источник возможных фактов, т.к. они генерируются автоматически и не проверяются людьми.
- Приложение для преобразования новых статей в машиночитаемые форматы с проверкой авторами. Это источник достоверных фактов, т.к. они проверены людьми.

Для обработки статей была разработана система автоматической семантизации текстов [1], которая на первом этапе осуществляет поиск и выделение концептов в тексте статьи, затем выделяет взаимосвязи между найденными концептами. Взаимосвязи представляют собой факты, каждый из которых представляет собой так называемый трипл (triple), включающий в себя субъект (subject), объект (object) и предикат (predicate) [2]. Трипл - это выражение, связывающее две сущности (субъект и объект) через отношение (предикат). Также, каждый трипл хранит информацию о концептах, содержащихся в нем. Ниже приведен пример трипла:

`Tryptophan hydroxylase` (субъект) `adds` (отношение) `a hydroxyl group to the 5 position of tryptophan to form 5-hydroxytryptophan (5-HTP)` (объект).

Приложение для преобразования новых статей представляет собой веб-приложение, позволяющее загрузить текст научной статьи в текстовом формате, добавить дополнительную информацию о терминах и их взаимосвязях в тексте статьи и затем экспортировать их в разрабатываемый граф. Интерфейс приложения [3] приведен на рисунке 2.

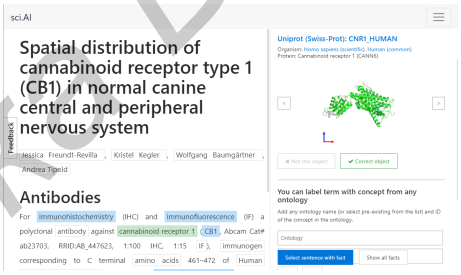

Рис. 1 – Интерфейс приложения для семантизации статей

Разработанное приложение позволяет добавить дополнительную информацию о ключевых терминах научной статьи и их взаимосвязях. После загрузки текста статьи в приложение происходит анализ текста и поиск в нем терминов из нескольких наиболее популярных биомедицинских онтологий (Uniprot, MeSH, ChEBI, ICD-10 и др.). Все найденные термины помечаются и требует валидации автором (рисунок 3).

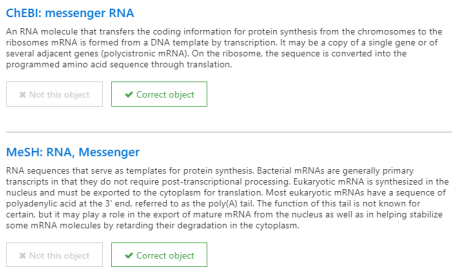

Рис. 2 – Пример термина, найденного в статье

Аналогично происходит анализ и выделение ключевых фактов в тексте статьи (рисунок 4).

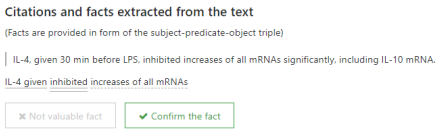

Рис. 3 – Пример термина, найденного в статье

Автоматическая обработка статей происходит аналогичным образом, исключая процесс валидации экспертом.

II. Построение графа знаний

Граф знаний представляет собой совокупность всех фактов извлеченных из статей. Факты объединяются с помощью уникальных идентификаторов концептов, содержащихся в субъекте и объекте факта (трипла). На рисунке 5 приведен пример цепочки, образуемой фактами для публикаций, связанных с болезнью Альцгеймера. Правая часть рисунка представляет текстовое представление фактов в статьях, в то время как левая - концепты, найденные в этих фактах (в данном случае, все концепты ссылаются на онтологию UniProt [4], специализирующуюся на протеинах).

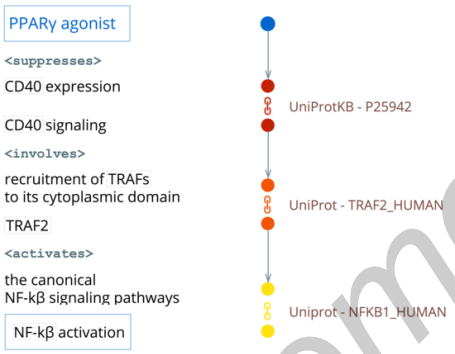

Рис. 4 – Пример взаимосвязей в графе

Для хранения графовой базы данных используется OrientDB [5]. Одним из главных преимуществ данной СУБД является возможность сочетать возможности документо-ориентированных и графо-ориентированных баз данных, что позволяет хранить в базе дополнительную информацию об элементах графа. Также стоит отметить нетребовательность к ресурсам, открытый код и отсутствие сторонних зависимостей.

Для визуализации графов используется библиотека cytoscape.js [6], разработанная для визуализации и анализа сетей и графов. Данная библиотека обладает широким набором функций, необходимых для работы с графами. Так, например, в ней реализованы алгоритмы поиска кратчайших путей: алгоритм Дейкстры, поиск A* (A звезда) и др. С другой стороны, cytoscape.js обладает необходимой гибкостью для внедрения специфических алгоритмов расположения узлов графа, фильтрации и других задач.

На рисунке 4 приведен пример пути между двумя заданными пользователем концептами.

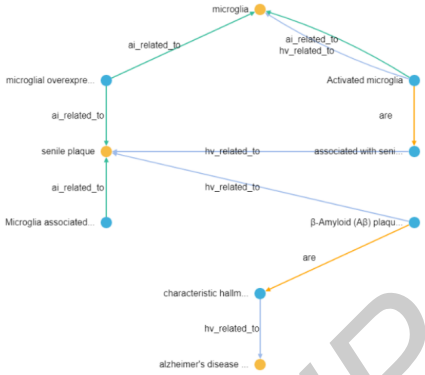

Рис. 5 – Путь между концептами микроглия [7] и болезнь Альцгеймера [8]

Заключение

В ходе исследования был построен прототип графа на основе нескольких сотен публикаций, связанных с болезнью Альцгеймера. Несмотря на небольшое количество научных статей, полученный граф можно использовать для поиска связей между концептами, связанными с данной болезнью.

Стоит отметить, что при каждой дополнительной загрузке фактов из статей в граф, появляется множество дополнительных связей между концептами, что затрудняет восприятие информации. Следующей задачей исследование является разработка способа поиска и фильтрации несущественных связей.

III. Список литературы

1. Пашук, А. В. Проблема распознавания именнованных сущностей в биомедицинских публикациях / А. В. Пашук, А. Б. Гуринович, Н. А. Волорова, А. П. Кузнецов // Big Data and Advanced Analytics : материалы 3 науч.-практич. конф., Минск, 3–4 мая 2017 г. / Белорус. гос. ун-т информатики и радиоэлектроники ; редкол.: М. П. Батура [и др.]. – Минск : БГУИР, 2017. – С. 350.
2. Resource Description Framework (RDF): Concepts and Abstract Syntax [Electronic resource]. - Mode of access: https://www.w3.org/TR/2004/REC-rdf-concepts-20040210/. - Date of access: 15.08.2017.
3. sci.AI Papers [Электронный ресурс] / Режим доступа: http://app.sci.ai/. – Дата доступа: 26.08.2017.
4. UniProt [Electronic resource]. - Mode of access: http://www.uniprot.org/. - Date of access: 16.08.2017.
5. OrientDB - Distributed Graph/Document Multi-Model Database [Electronic resource]. - Mode of access: http://orientdb.com/. - Date of access: 16.08.2017.
6. Cytoscape.js [Electronic resource]. - Mode of access: https://js.cytoscape.js. - Date of access: 16.08.2017.
7. MeSH RDF Explorer: Microglia [Electronic resource]. - Mode of access: https://id.nlm.nih.gov/mesh/D017628.html. - Date of access: 16.08.2017.
8. International Classification of Diseases, Version 10 - Alzheimer's disease [Electronic resource]. - Mode of access: http://purl.bioontology.org/ontology/ICD10/G30.0. - Date of access: 16.08.2017.