

СТРУКТУРА КНИГИ «ГАРРИ ПОТТЕР & ФИЛОСОФСКИЙ КАМЕНЬ» В ТЕОРИИ ГРАФОВ

Д. А. Дриня

drinc2002@gmail.com

Научный руководитель – И. В. Пинчук, кандидат социологических наук, доцент

Романы часто предлагают сложное повествование, которое может быть трудным для читателя на протяжении всей книги. Поэтому предложение инструментов, помогающих лучше понять повествование, – это интересный путь, который пока еще мало изучен. Действительно, несмотря на недавнее развитие алгоритмов НЛП, таких как GPT-3, которые демонстрируют выдающиеся результаты в широком спектре задач (перевод, ответы на вопросы, генерация текста...), получение истинного понимания книги или даже хорошего резюме все еще остается недоступным. Тем не менее, существуют и другие способы получить полезную информацию из романа. В этой заметке мы обратимся к концепции, которая находится на пересечении обработки естественного языка (NLP) и сетевой науки (NS) – новой и развивающейся ветви прикладной теории графов, которая объединяет традиции многих дисциплин, включая математику, социологию, экономику и информатику.

Ключевые слова: социальные сети; динамическая гетерогенная социальная сеть персонажей; динамическая социальная сеть

Современная теория графов позволяет проводить многоуровневый анализ практически любого социального объекта, используя математические и статистические методы, при этом не обладая зависимости от статичности времени. В рамках данной работы мы создадим соответствующую динамическую гетерогенную социальную сеть персонажей из текстового содержания книги, а затем используем информацию из этого графа для улучшения понимания романа. В частности, мы фокусируемся на пяти различных приложениях: важности персонажей, обнаружении сообществ, обобщении и сравнении книг. Их изучение дает нам ценные подсказки о сюжете книги, стиле автора или важности, отношениях и ролях персонажей.

Мы применим наш анализ к знаменитой книге Дж. К. Роулинг: «Гарри Поттер и философский камень» (HP1) – хотя анализ может быть применен к абсолютно любой книге.

Возможно, в данный момент вы все еще задаетесь вопросом: «Но что такое динамическая гетерогенная социальная сеть персонажей?». Чтобы ответить на ваш вопрос, сеть персонажей – это не что иное, как граф, то есть набор узлов V и ребер E , где узлы представляют персонажей книги, а ребра – взаимодействия между ними. В то время как подавляющее большинство подходов в литературе сосредоточено на статических методах,

которые по определению фиксированы во времени – один граф для всей книги, мы придерживаемся динамического подхода – граф постоянно развивается во времени и позволяет сохранить временную информацию повествования. Чтобы создать этот динамический аспект, мы дополнительно делаем граф гетерогенным, то есть с несколькими типами ребер и узлов. Другими словами, не все узлы будут представлять персонажей, а все ребра не будут представлять взаимодействие между ними. Но к этому мы вернемся позже.

Для того чтобы создать своего рода динамический граф символов, нам сначала нужно провести некоторую НЛП над исходным текстом. Цель модуля предварительной обработки – перехватить все случаи появления всех персонажей в книге в различных формах и сохранить следующую информацию: имя персонажа (например, 'Гарри'), количество лексем с начала (например: 30490), индекс главы (например: 2), соответствующая сущность (например: 'HARRY POTTER').

Мы получаем книгу из Интернета в формате .txt, используя библиотеку Project Gutenberg, и разбиваем ее на главы. Почему? Потому что глава – это единица, представляющая интерес для развития сюжета.

Мы запускаем BERT NER – предварительно обученный google BERT, тонко настроенный для задач распознавания сущностей – на всей книге, глава за главой, чтобы обнаружить все вхождения символов. Мы сохраняем только сущности «PER» (лица), которые мы храним вместе с их положением в тексте и соответствующим индексом главы.

Мы сопоставляем каждое упоминание персонажа с уникальной сущностью (= главный герой). Социальная сеть связывает сущности, и поэтому очень важно сгруппировать все со-референты вместе. Например, что касается Гарри Поттера, мы хотим, чтобы Гарри, Поттер, мистер Поттер, Х. Поттер были связаны с одной сущностью. В данном случае ГАРРИ ПОТТЕР. Мы следуем стратегии, разработанной M.Ardanuy и C.Sporleder в [1] для этой задачи разрешения символов. Это было сделано в 4 этапа:

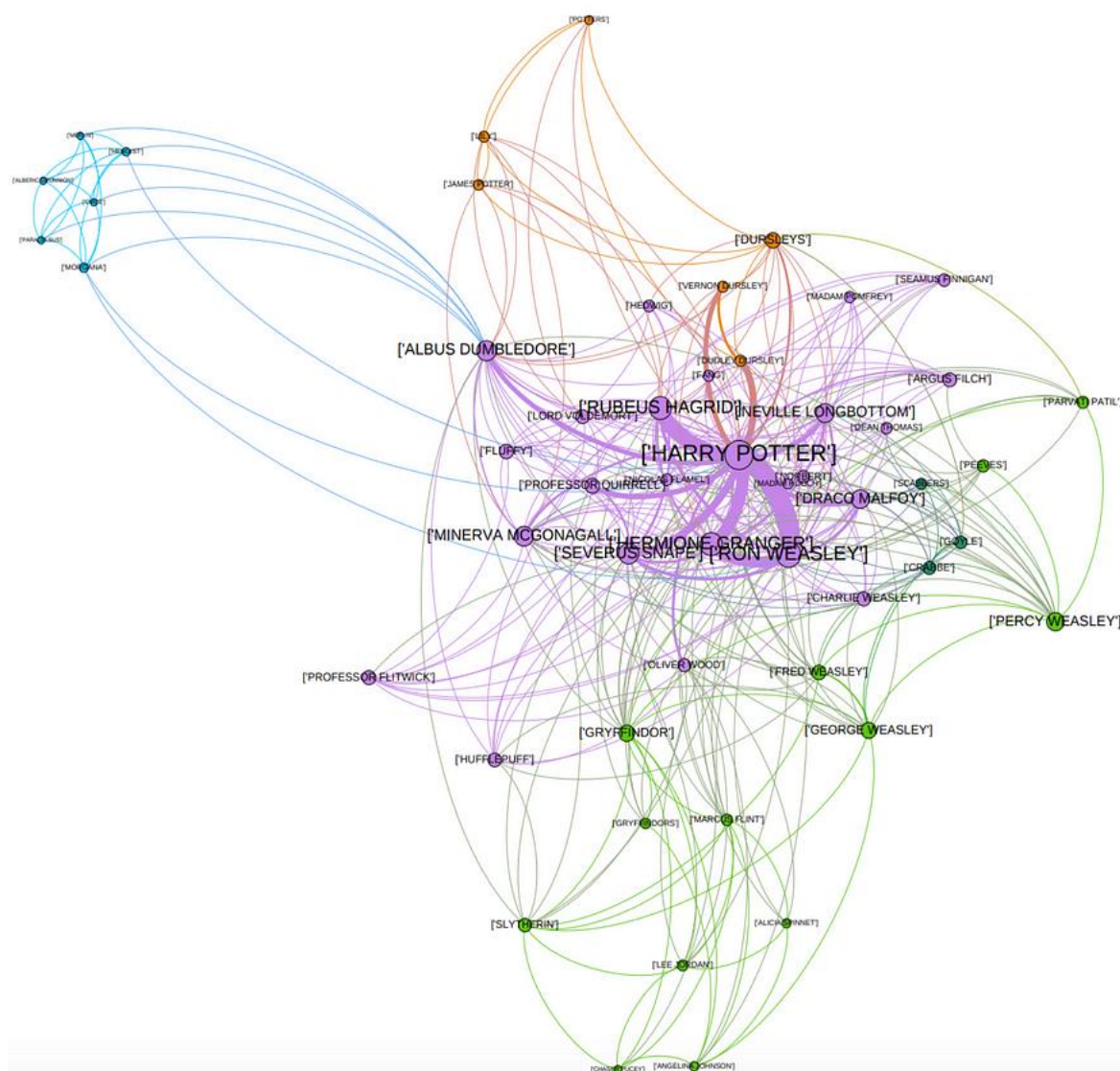
Разбор имени человека: использование фреймворка NameParser для разбора имени персонажа в общую структуру. Например, имя Mr. Harry Potter будет разобрано как: {title: Mr, first: Harry, last: Поттер}

Назначение пола. Определите, если возможно, жанр персонажа. Здесь мы просто оцениваем его по названию и по имени, используя базы данных. Например, если название – «Мистер», то жанр – «Мужской».

Теперь мы переходим к важнейшему моменту нашего подхода: как создать граф, хранящий временную информацию о динамике повествования?

Мы решили построить граф, используя информацию, сохраненную выше, и представляя следующую структуру.

Алгоритм сопоставления: сопоставляет каждое вхождение с сущностью, группируя сопоставимые референты вместе. Не углубляясь в детали, скажем, что сначала мы рассматриваем все вхождения с названием, именем и фамилией – вхождения с наиболее полной формой. Если ни одна из существующих сущностей не имеет имени, фамилии и жанра, мы создаем новую. Так мы поступаем со всеми упоминаниями, имеющими одинаковую структуру, а затем фокусируемся на оставшихся, которые отображают только имя и фамилию.



Динамический граф по книге «Гарри Поттер и философский камень»

Граф сущностей, в частности, полезен для некоторого глобального анализа книги. Это статическая свернутая версия полного графа, которая содержит просто один тип узлов: entity-node и один тип ребер: interact-with, которые взвешиваются по количеству взаимодействий между двумя

сущностями. Часто бывает интересно разрезать его по главам, чтобы встроить временную информацию. Таким образом, мы сворачиваем динамический граф в последовательность графов сущностей, по одному на каждую главу.

Мы можем расширить это представление еще дальше и создать тонкий динамический граф взаимодействия сущностей, где временная шкала – это просто количество лексем в тексте, т.е. «время» идет от 0 до количества лексем в романе.

Наш сетевой анализ подтвердил некоторые ожидания и позволил по-новому взглянуть на богатую фантазиями книгу «Гарри Поттер и философский камень». Обратите внимание, что мы рассмотрели причудливое применение сетевой науки, чтобы дать представление о ее возможностях. Другие связанные задачи включают кластеризацию романов, предсказание жанра, автора или даже новых взаимодействий персонажей. В целом, существует множество более серьезных приложений, и наука о сетях обещает быть бесценной для понимания нашей современной сетевой жизни.

Библиографические ссылки

1. Кластеризация романов на основе структуры, Mariona Coll Ardanuy и Caroline Sporleder. 2014.
2. Рассказывание историй о динамических сетях с помощью графических комиксов. Бенджамин Бах и др. 2016 г.
3. Сеть престолов. Эндрю Беверидж и Цзе Шань. 2016.
4. Извлечение социальных сетей из литературной беллетристики. Дэвид Элсон и др. 2010.
5. Extraction and Analysis of Fictional Character Networks: A Survey. Винсент Лабатут и Ксавье Бост. 2019.
6. How to Tell Stories with Networks: Исследование повествовательных возможностей графиков на примере «Илиады». Томмазо Вентурин и др. 2017.