оперативная память являются критичным ресурсами, а местом на физическом носителе можно пожертвовать. Степень сжатия, достигаемая в таком случае (около двух) вполне приемлема;

– если объём потребляемой алгоритмом во время работы оперативной памяти и процессорного времени не являются критичными, а важно максимальное сжатие данных, рекомендуется использовать размеры словаря более 16384 байт. Это позволит существенно увеличить степень сжатия данных.

Литература

1Ватолин Д., Ратушняк А., Смирнов М., Юкин В. Методы сжатия данных. Устройство архиваторов, сжатие изображений и видео. – М,: ДИАЛОГ-МИФИ, 2003. – 348с.

2Д.Сэломон. Сжатие данных, изображений и звука - Москва: Техносфера, 2004. - 368с.

И.В. Чеховских, Е.В. Симонова

РАЗРАБОТКА АВТОМАТИЗИРОВАННОЙ СИСТЕМЫ ПОИСКА СХОДСТВА ИССЛЕДОВАНИЙ В НАУЧНЫХ СТАТЬЯХ

(Самарский национальный исследовательский университет
имени академика С.П. Королёва)

Научная статья – это произведение, отражающее результаты исследовательской деятельности автора (авторов). Информация излагается четко, конкретно, детально [1].

В век информационных технологий исследователь может получить доступ к огромному количеству научных статей. Проблемой является поиск действительно полезной информации во множестве доступных статей, находящихся в интернете, для подготовки к различным конференциям или для написания собственных исследовательских работ.

Как правило, при поиске научных статей используют два основных подхода: по автору, если необходимо ознакомиться с полным списком исследований конкретного автора, или по ключевым словам, которые могут как содержаться, так и отсутствовать в названии статьи. Искать научные статьи двумя этими способами можно на большинстве сайтов. Но, к сожалению, данные методы поиска не дают полного представления об исследовании. Необходимо выполнять более продвинутый поиск научных статей на основании сходства текстового содержания научных исследований.

Для поиска похожих научных статей на основании контекста самой статьи необходимо первоначально выполнить канонизацию текста исследования – приведение оригинального текста к единой нормальной форме. Канонизация текста статьи выполняется с помощью Natural Language Processing (NLP).

Natural Language Processing (NLP) – общее направление искусственного интеллекта и математической лингвистики. Оно изучает проблемы компьютерного анализа и синтеза естественных языков [2].

Процесс канонизации текста состоит из двух этапов:

− очистка текста научной статьи от предлогов, союзов, знаков препинания и прочего шума в модели, который не должен участвовать в сравнении. В большинстве случаев также предлагается удалять из текста прилагательные, так как они не несут смысловой нагрузки;

− лемматизация.

Лемматизация – процесс приведения словоформы к лемме – её нормальной (словарной) форме [3].

После выполнения канонизации получается текст, очищенный от «мусора» и готовый для сравнения.

Для реализации метода поиска похожих научных исследований предлагается использовать модель Doc2Vec, принадлежащую набору моделей Word2Vec. Word2Vec – это набор моделей, принимающих на вход текст и получающих в результате работы представление слов в векторном пространстве на основе контекста. Принцип работы Word2Vec можно описать следующим образом: максимизация косинусной близости для векторного представления слов, которые появляются в похожих контекстах, и наоборот, её минимизация для слов, не встречающихся в похожих контекстах.

Векторное представление – общее название для различных подходов к моделированию языка и обучению представлений в обработке естественного языка, направленных на сопоставление словам (научным статьям) из некоторого словаря векторов из $\mathbb{R}^n$ для n, значительно меньшего количества слов в словаре [4]. Пример вектора размерностью 20 для научной статьи представлен на рисунке 1.

$$\overrightarrow{\text{вектор научной статьи}} = (-0.04096178, -0.18379952, -0.11860437, -0.20194705,$$
$$-0.50240123, 0.1839197\ , -0.15432355, 0.72807366, 0.5409689\ , -0.4779142, -0.10427655, 0.20272826,$$
$$0.00589776, 0.05693987, 0.16688912, -0.42739728, -0.01697599, -0.07323613, 0.26289058, 0.64651096)$$

Рисунок 1 – Вектор научной статьи

В отличие от Word2Vec, Doc2Vec принимает на вход вместе с текстом идентификатор текста. Существуют две модели Doc2Vec: Distributed Memory (DM) и Distributed-Bag-Of-Words (DBOW). DM сопоставляет каждому идентификатору текста вектор. Таким образом, при обучении векторов слов также обучается вектор идентификатора текста, в конце обучения он содержит числовое представление документа. DBOW, в свою очередь, предсказывает появление случайных слов в тексте только на основании вектора текста.

Модели Doc2Vec (DM и DBOW) представляют собой нейронную сеть (рисунок 2), задачей которой является реконструкция контекста слов [5].

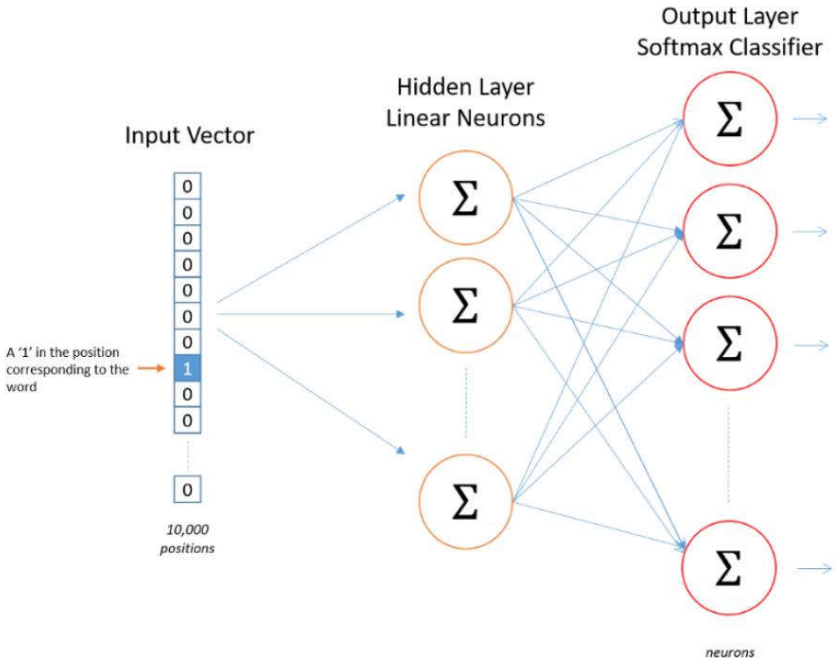

Рисунок 2 – Модель нейронной сети для обучения векторов текстов

Векторы слов и научных статей обучаются с использованием метода стохастического градиентного спуска и метода обратного распространения ошибки. Векторы научных статей являются уникальными, а векторы одинаковых слов в разных документах совпадают.

Конечная цель – обучить весовую матрицу скрытого слоя. Скрытый слой этой модели в действительности работает как справочная таблица, а именно выход скрытого слоя – это вектор научной статьи для входной научной статьи. Выходной слой отбрасывается после окончания обучения.

Благодаря возможности сравнения между собой полученных векторов можно определить сходство исследований в научных статьях. Главным преимуществом данного подхода является малая размерность векторов.

Литература

1. Обоснование и достоверность научных результатов [Электронный ресурс] // «Лекции.Орг». – 2015. – URL: https://lektsii.org/17-19946.html (дата обращения 16.05.2019).

2. Обработка естественного языка [Электронный ресурс] // Фонд Викимедиа. – 2001. – URL: https://ru.wikipedia.org/wiki/Обработка_ естественного_ языка (дата обращения 16.05.2019).

3. Лемматизация [Электронный ресурс] // Фонд Викимедиа. – 2001. – URL: https://ru.wikipedia.org/wiki/Лемматизация (дата обращения: 17.05.2019).

4. Векторное представление слов [Электронный ресурс] // Фонд Викимедиа. – 2001. – URL: https://ru.wikipedia.org/wiki/Векторное_представле

5. ние_слов (дата обращения: 17.05.2019).

6. Reynold, S. Xin. A Resilient Distributed Graph System on Spark [Текст] / S. Xin Reynold, E. Gonzalez Joseph, J. Franklin Michael, Stoica Graph X Ion. – Berkeley, 2013. – 6 с. (дата обращения 17.05.2019).

Н.С. Якошук, Ю.М. Заболотнов

АВТОМАТИЗИРОВАННАЯ СИСТЕМА МОДЕЛИРОВАНИЯ ПРОЦЕССА ФОРМИРОВАНИЯ ВРАЩАЮЩЕЙСЯ ТРОСОВОЙ СИСТЕМЫ С ПОМОЩЬЮ ПРОВОДЯЩЕГО ТОК ТРОСА

(Самарский национальный исследовательский университет
имени академика С.П. Королёва)

Введение

Космическая тросовая система (КТС) –это связка двух или более космических аппаратов, соединенных тросами длиной в десятки или сотни километров, причем еѐ развертывание и ориентация в пространстве обеспечивается в основном за счет действия гравитационных сил [1].

Космические тросовые системы представляют существенный практический интерес, так как позволяют решать широкий спектр задач, которые практически невозможно или неэффективно выполнять с помощью уже существующих технических средств. Например, они позволяют снизить затраты на поддержание и маневрирование на орбите спутников, сход с орбиты спускаемых аппаратов. Также КТС могут быть использованы для выполнения ремонтных работ космического аппарата (КА), для снабжения КА электроэнергией, для удаления КА, которые завершили срок службы, и другого космического мусора с орбиты, для обеспечения транспортных операций в космосе. Это всего лишь небольшая часть возможных практических применений КТС.

В связи с этим становится актуальной задача разработки автоматизированной системы, которая могла бы позволить моделировать процесс формирования вращающейся тросовой системы с помощью проводящего ток троса. Использование данной системы позволит рассчитать управление величиной тока для перевода системы во вращение для различных входных параметров. Разработанное программное обеспечение может быть использовано для исследования динамики электродинамической космической тросовой системы (ЭДКТС).

Постановка задачи

На рисунке 1 изображена схема ЭДКТС на орбите. Точка С– центр масс системы, точка O – центр масс Земли, m_1и m_2 – массы концевых тел, θ – угол отклонения троса от вертикали. $\bar{F}$– вектор силы Ампера, $\bar{B}$ – вектор магнитной индукции, I –величина тока в тросе.

Процесс формирования тросовой системы описывается системой дифференциальных уравнений (ДУ). Для исследования основных закономерностей динамики данного процесса целесообразно применять упрощенные модели, в