

Анализ методов машинного обучения для рекомендаций научных публикаций.

Научный руководитель – Федянин Денис Николаевич

Леонова Юлия Сергеевна

Студент (магистр)

Московский государственный технический университет имени Н.Э. Баумана,

Фундаментальные науки, Москва, Россия

E-mail: julia_leonova_123456@mail.ru

Современные большие языковые модели (LLM) активно применяются в различных задачах обработки текста, таких как классификация, перевод, генерация и анализ текста. Однако их использование часто ограничивается требованиями к конфиденциальности данных или необходимостью работы в рамках специализированного программного обеспечения. В таких условиях возникает потребность в разработке собственных методов машинного обучения, которые могли бы эффективно решать эти задачи.

В рамках исследования был проведен анализ сети "ученый-статья-журнал", где узлы — это аннотации статей, а связи — общие термины. Основное внимание работы уделено разработке и сравнению подходов к составлению рекомендательных подборок научных статей, в которых бы учитывались такие признаки, как актуальность, объем данных, методологическая значимость, влияние на область, цитируемость и историческая важность. Часть этих признаков может быть визуализирована с помощью диаграммы.

Рис. 1. Схематичная диаграмма, визуализирующая научные статьи в векторном пространстве.

На визуализации запрос пользователя (Q) обозначен зеленым кругом, а остальные статьи — синими. Радиус круга отражает количество цитирований статьи: чем больше круг, тем выше цитируемость. По оси OY показано время публикации, что позволяет определить статью, где впервые упоминался термин из интересующей области. По оси OX расстояние между кругами указывает на схожесть содержания статей: чем ближе статья к запросу, тем больше она соответствует его тематике.

Для составления рекомендательной подборки рассмотрены два метода: классификатор терминов и метод TF-IDF (Term Frequency-Inverse Document Frequency) [1]. Первый метод основан на экспертных знаниях, примером может служить классификатор терминов РНФ, который систематизирует научные проекты по направлениям и используется для подачи заявок на гранты. Используемый в работе классификатор терминов включает набор экспертно полученных терминов, разделенных по тематическим категориям. Его функция заключается в выделении ключевых слов и отнесении статьи к теме, сформулированной в запросе пользователя.

Рис. 2. Структура классификатора.

Тогда как метод TF-IDF автоматически оценивает важность терминов (TF) в контексте всего корпуса статей (IDF), что позволяет учитывать все слова в текстах, включая те, что не вошли в классификатор. Формулы для вычисления итогового веса термина:

где — количество раз, когда термин появляется в аннотации, — общее количество терминов в аннотации, — общее количество статей в коллекции.

Оба метода были протестированы на наборе данных из Информационной системы анализа научной деятельности (ИСАНД) [2].

Результаты исследования продемонстрировали, что классификатор терминов обладает высокой скоростью обработки данных, однако имеет существенное ограничение: его эф-

эффективность напрямую зависит от наличия в тексте статьи терминов, заранее определенных экспертами. Если статья содержит неизвестные классификатору слова или частицы "не", это может привести к ошибкам в классификации, что ограничивает его универсальность. Метод TF-IDF демонстрирует высокую эффективность при составлении рекомендательных подборок, предлагая не только статьи, наиболее релевантные пользовательскому запросу, но и включая работы, в которых ключевые понятия из запроса были впервые введены в научный оборот. Кроме того, метод выделяет статьи, содержащие обзорную или более общую информацию по исследуемой тематике, что позволяет пользователю получить полное представление о исследуемой научной области.

Проведенное исследование демонстрирует потенциал использования методов машинного обучения и анализа данных для решения задач управления в научной среде. Результаты работы имеют практическую значимость для оптимизации процессов поиска интересных научных публикаций и быстрого ознакомления с интересующей научной областью.

Источники и литература

- 1) Robertson S.E. Understanding Inverse Document Frequency: On Theoretical Arguments for IDF // 2004.
- 2) Информационная система а научной деятельности [Электронный ресурс] URL: <http://isand.ipu.ru/> (дата обращения: 20.11.24).

Иллюстрации

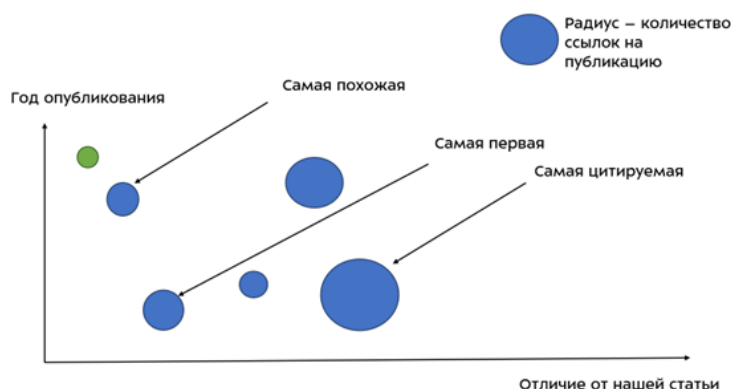


Рис. : Рис. 1. Схематичная диаграмма, визуализирующая научные статьи в векторном пространстве.

Факторы классификации	1-й уровень	2-й уровень	ТЕРМИНЫ
Общенаучные термины	Общенаучные термины	Общенаучные термины	абстрагирование, агент, агрегирование, адаптация, адекватный, аксиома, активность, актуальность, алгоритм, аспект, атрибут, величина, верификация, вид, внедрение, внешняя среда, вопрос, выбор, выборка, гипотеза, декомпозиция, деятельность, диффузия, дисперсия, доказательство, допустимое множество, единый, этап, идеализация, идентификация, иерархия, иммерсия, имитация, индигенд, индукция, институт, интерес, исследование, категория, качество, класс, классификация, компетентность, компетенция, компонент, норма
Математический аппарат	Алгебра и теория чисел		алгебра, теория чисел
		Теория чисел	алгебраические числа, взаимно простые числа, группа Гауэ, дзета-функция, конечные поля, простое
		Линейная алгебра, теория матриц	базис, векторное пространство, детерминант, евклидово пространство, линейная алгебра, норма, оператор
		Группы и алгебры Ли	алгебра Ли, алгебры Картана, группа Ли, мера Хаара, однородные пространства групп Ли, операторы
		Решетки и Булевы алгебры	булева алгебра, дистрибутивная решетка, квантовая логика, меры на дистрибутивных решетках

Рис. : Рис. 2. Структура классификатора.