УДК 004.622

РАЗРАБОТКА МЕТОДОВ ОЦЕНКИ КАЧЕСТВА ИСТОЧНИКОВ ИНФОРМАЦИИ ДЛЯ МАШИННОГО ОБУЧЕНИЯ

Примак Л.А. (ИТМО)
Научный руководитель – доктор технических наук, Бессмертный И.А.
(ИТМО)

Введение. Количество областей, где внедряется машинное обучение, с каждым годом постепенно растёт. Неизбежно увеличивается и число ошибок в результатах работы алгоритмов. Основа машинного обучения – это данные, поэтому одним из вариантов повышения качества выходной информации является более усиленный контроль за входными значениями. Это можно реализовать с помощью фильтрации источников исходной информации.

Основная часть. Разработка метода оценивания качества источников идёт от частного к общему. Показательным примером может служить оценка качества научных статей, т.к. количество публикаций постоянно увеличивается.

Сам процесс состоит из трёх циклов. На каждом этапе выделяются несколько критериев. Затем определяются их значения для определённого объекта (научная статья) и сверяются с установленными метриками контроля. По результатам сравнения определяется, переходит ли объект на следующий цикл проверки или нет.

Первый блок – оценка методанных объекта. Второй – анализ названия, ключевых слов и аннотации. Последний этап – полный анализ текста статьи. Разберём более подробно первый цикл.

Методанные, в данном конкретном примере, это общая информация о статье. К ним относятся: сведенья о журнале и авторе, библиометрические показатели (входит в РИНЦ, входит в ядро РИНЦ, наличие рецензии, количество цитирований), альтметрики. Оценка идёт по следующим метрикам:

1) Контроль качества журнала – частота публикаций (>1 раза в месяц), количество статей (>1000 в год), отсутствие рецензирования и индексов, нет информации о участниках редколлегии и правил для авторов, ложные данные о географическом положении.
2) Оценка уровня автора – количество и частота публикаций (> 10 в год), индекс Хирша, различные области исследования.
3) Проверка общих данных по статье – отсутствие рецензии, маленькое количество цитирований, отзывов и просмотров, некорректность ссылок на источники.

В результате статьи классифицируются на две группы: проходящие на второй блок проверки или нет.

Выводы. Определены метрики контроля качества источников данных на первом этапе фильтрации на примере выбора научных статей. Проведено тестирование и получены объекты для следующих этапов анализа.

Список использованных источников:

1. Beall’s list – of potential predatory journals and publishers. – URL: https://beallslist.net/ (date of treatment: 05.02.2024).
2. Высшая аттестационная комиссия: офиц. сайт. – URL: https://vak.minobrnauki.gov.ru (дата обращения: 06.02.2024).
3. Иностранные хищные журналы в Scopus и WoS: переводной плагиат и российские недобросовестные авторы // Российская академия наук: офиц. сайт. – URL: https://kpfran.ru/wp-content/uploads/plagiarism-by-translation-2.pdf (дата обращения 06.02.2024).