ТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ НА СЛУЖБЕ ОБЕСПЕЧЕНИЯ БЕЗОПАСНОГО ВЕБ-ПРОСТРАНСТВА

А.А. Новиков, К.А. Радкевич

ОАО «Гипросвязь», Минск, Беларусь

Анализ практического опыта применения многочисленных алгоритмов тематического моделирования, убеждает, что тематические модели обладают огромным «запасом решений», в том числе при разработке проектных решений в области обеспечения безопасного веб-пространства.

Благодаря глобальному распространению сети Интернет большинство пользователей получили возможность в невиданных ранее объемах получать и генерировать информацию. Это стало не только благом, но и породило ряд проблем, связанных с генерацией негативного контента, носящего деструктивный характер, как в отношении конкретных людей, так и сообществ, корпораций и государств в целом. Очевидно, что «красноармейца со штыком» к каждому защищаемому объекту (субъекту) не представишь, нужны новые подходы и технологии.

Все ранее используемые аналитические методы, столкнувшиеся с феноменом «больших данных» указывают на потребность в новых методах обработки и анализа, способных извлекать из этих, как правило, неструктурированных данных полезное «зерно», знание. Совокупность таких методов обозначается термином «интеллектуальный анализ данных» (data mining).

Из этой совокупности выделяется подмножество, специализирующееся на анализе текстовых данных – «интеллектуальный анализ текстовых данных» (text mining). Кроме того, выделяется группа методов, которые выполняют задачу тематического моделирования – построения статистических моделей, определяющих тематическую принадлежность каждого документа из корпуса [1].

Тематическое моделирование – это одна из современных технологий обработки естественного языка (англ. Natural language processing, NLP), активно развивающаяся с конца 90-х годов. Тематическая модель коллекции текстовых документов определяет, к каким темам относится каждый документ, и какие слова образуют каждую тему. Тематическое моделирование не претендует на полноценное понимание естественного языка (англ. natural language understanding, NLU), однако выявление тематики можно считать определенным шагом в этом направлении [2].

В сентябре 2022 года в Государственном комитете по науке и технологиям Республики Беларусь успешно прошел государственную экспертизу международный научно-технических проект «Система мониторинга и интеллектуального анализа веб-пространства “Безопасное веб-пространство”». Проект рассчитан на два года, стартует в мае 2023 года. В проекте участвуют представители научного коллектива Центра перспективных исследований в сфере цифрового развития ОАО «Гипросвязь» и ученые Центрального университета Раджастхана и Индийского института информационных технологий Коты, Республика Индия.

В рамках данного проекта предполагается на основе тематических моделей разработать алгоритмы автоматизированного мониторинга и интеллектуального

анализа контента веб-пространства в заданных пределах (областях, доменах) по разработанным правилам (частным политикам), для последующей передачи и использования результатов системой поддержки принятия решений. Также запланировано исследование возможности переноса некоторых разработанных алгоритмов на язык программирования Python.

Список литературы

1. Воронцов К.В. Вероятностное тематическое моделирование: теория, модели, алгоритмы и проект BigARTM, 2020.

2. Deerwester S., Dumais S.T., Furnas G.W. [et al.] Indexing by Latent Semantic Analysis // JASIS. 1990. Vol. 41. P. 391–407.