РАЗРАБОТКА АЛГОРИТМА ДЛЯ АВТОМАТИЗАЦИИ КОНТРОЛЯ ПОЛНОТЫ ДАННЫХ, ОТОБРАЖАЕМЫХ НА САЙТАХ КОМПАНИИ ДЛЯ СОБЛЮДЕНИЯ ФЗ И СОБЛЮДЕНИЯ РЕГУЛЯТОРНЫХ ОГРАНИЧЕНИЙ

Раджабова Н.Ш. Email: Radzhabova578@scientifictext.ru

Раджабова Наима Шамильевна – кандидат физико-математических наук, доцент,
кафедра дискретной математики и информатики,
Дагестанский государственный университет, г. Махачкала

Аннотация: *в статье рассматривается подход к разработке распределенной веб-системы, которая может с заданной регулярностью проверять страницы сайта и определять наличие необходимых файлов по требованиям ФЗ, их актуальность и общую возможность открытия/скачивания через браузер. Задача проверки ссылок весьма злободневна, так как «потеря ссылок» также может привести к материальным и другим издержкам.*

Разработка алгоритма на базе методов глубинного машинного обучения является новым способом отражения полноты информации и риска существенного расхождения с требованиями ФЗ и регулирующих органов.

Ключевые слова: *машинное обучение, распределённая система, мониторинг, проверка доступности, автотест, регуляторные ограничения.*

DEVELOPMENT OF AN ALGORITHM TO AUTOMATE THE CONTROL OF THE COMPLETENESS OF DATA DISPLAYED ON COMPANY WEBSITES TO COMPLY WITH THE FEDERAL LAW AND COMPLY WITH REGULATORY RESTRICTIONS

Radzhabova N.Sh.

Radzhabova Naima Shamilyevna – Candidate of Physical and Mathematical Science, Associate Professor,
DEPARTMENT OF THE DISCRETE MATHEMATICS, COMPUTER SCIENCE,
DAGESTAN STATE UNIVERSITY, MAKHACHKALA

Abstract: *the article discusses the approach to developing a distributed web system that can with a given regularity check the pages of the site and determine the availability of the necessary files according to the requirements of the Federal Law, their relevance and the general ability to open / download via the browser. The task of checking links is very topical, since the "loss of links" can also lead to material and other costs.*

The development of an algorithm based on methods of deep machine learning is a new way of reflecting the completeness of information and the risk of a significant discrepancy with the requirements of the Federal Law and regulatory bodies.

Keywords: *machine learning, distributed system, monitoring, accessibility checking, autotest, regulatory restrictions.*

УДК 004.891.3

Отсутствие на веб-сайтах компании и организаций необходимых документов и реестров на отчетную дату, отчетов для инвесторов на требуемую дату и т.д. сопряжено с риском регуляторных ограничений со стороны регулирующих органов, риском сбоев в предоставлении услуг сторонним организациям, подрядчикам. Повышается риск нарушения или остановки поддерживающих процессов (внешний инвестор не будет своевременно получать информацию о запланированной ставке купона, например, и т.д.).

Актуальным является разработка распределенной веб-системы, которая может с заданной регулярностью проверять страницы сайта и определять наличие необходимых файлов по требованиям ФЗ, их актуальность и общую возможность открытия/скачивания через браузер. Задача проверки ссылок весьма злободневна, так как «потеря ссылок» также может привести к материальным и другим издержкам.

Разработка алгоритма на базе методов глубинного машинного обучения является новым способом отражения полноты информации и риска существенного расхождения с требованиями ФЗ и регулирующих органов. Разрабатываемые методы способны обеспечить контроль рисков на невиданном ранее уровне и обеспечивают компании конкурентным преимуществом в текущих условиях.

Полных аналогов программных систем, которые могут с заданной регулярностью проверять страницы сайта и определять наличие необходимых файлов, их актуальность и общую возможность открытия/скачивания через браузер, с выдачей результата о полноте данных, отображаемых на сайтах компании для соблюдения ФЗ и соблюдения регуляторных ограничений данных, не обнаружено.

Наиболее близким аналогом является автоматизированная информационная система «Мониторинг

государственных объектов» [1] (далее – АИС «Мониторинг государственных объектов»), разработанная Министерством экономического развития Российской Федерации. АИС «Мониторинг государственных объектов» предназначена для оценки открытости информации о деятельности государственных органов и доступности государственных информационных ресурсов для граждан.

В рамках мониторинга проводится оценка соответствия требованиям к доступности информации о деятельности государственных органов и органов местного самоуправления, проверяются технические характеристики официальных сайтов и мнения пользователей о работе государственного органа. Его представительство в Интернете контролируется и принимается во внимание.

Методы мониторинга:

- анализ отзывов и запросов пользователей;
- анализ экспертных заключений;
- анализ автоматической обработки технических характеристик;
- анализ рейтинга посещаемости.

Недостатки данной системы:

- объектом мониторинга являются лишь официальные сайты государственных органов и местных органов власти, созданных и действующих на момент реализации мониторинга;
- методы мониторинга ресурсоемкие (анализ экспертных заключений).

Необходимо разработать алгоритм для автоматизации контроля полноты данных, отображаемых на официальных сайтах любых компаний и организаций, для соблюдения ФЗ, и соблюдения регуляторных ограничений, реализованный в виде веб-сервиса, с использованием методов машинного обучения.

В основу современной технологии Data Mining (discovery-driven data mining) [2, 3] положена концепция шаблонов (паттернов), представляющих собой закономерности, свойственные выборкам данных, которые могут быть компактно выражены в понятной человеку форме. Поиск шаблонов производится методами, не ограниченными рамками предположений о структуре выборки и виде распределений значений анализируемых показателей. Методика контроля полноты данных, отображаемых на сайтах компании, включает в себя построение таких шаблонов, с которыми и сравниваются реальные сайты.

Оценка технических параметров сайтов в автоматическом режиме включает необходимым образом определение следующих показателей:

- время отклика официального сайта компании;
- количество ошибок в разметке страниц официальных сайтов и проверка состояния ссылок, в том числе;
- количество ошибок в структуре каждого официального сайта по требованиям ФЗ;
- количество обновлений главной страницы в неделю.

Разработка описываемого алгоритма включает следующие этапы:

- разработка методики контроля полноты данных, отображаемых на сайтах компании;
- подбор платформы для машинного обучения;
- этап предобработки данных (feature engineering) – процесс формирования и отбора признаков правильно наполненного сайта;
- разработка алгоритма и реализация программной системы автоматизации контроля полноты данных и соблюдения регуляторных ограничений.

Список литературы / References

1. Мониторинг государственных сайтов. [Электронный ресурс]. Режим доступа: https://www.minfin.ru/ru/om/link/ratings/monitoring/ (дата обращения: 25.09.2019).
2. Data Mining – интеллектуальный анализ данных. [Электронный ресурс]. Режим доступа: https://blog.iteam.ru/data-mining-intellektualnyj-analiz-dannyh/ (дата обращения: 20.10.2019).
3. Witten I.H., Frank E. Data Mining. Practical Machine Learning Tools and Techniques. Morgan Kaufmann Publishers is an imprint of Elsevier, 2005. [Электронный ресурс]. Режим доступа: http://index-of.co.uk/Data-Mining/Data%20Mining%20Practical%20Machine%20Learning%20Tools%20and%20Techniques%203rd%20Edition-Mantesh.pdf/ (дата обращения: 20.12.2019).