Объединенный институт проблем информатики
Национальной академии наук Беларуси

XXI Международная
научно-техническая конференция

РАЗВИТИЕ ИНФОРМАТИЗАЦИИ И ГОСУДАРСТВЕННОЙ СИСТЕМЫ НАУЧНО-ТЕХНИЧЕСКОЙ ИНФОРМАЦИИ

РИНТИ-2022

17 ноября 2022 г., Минск

Доклады

Минск
ОИПИ НАН Беларуси
2022

УДК 004.04

ВОПРОСЫ ПРОВЕРКИ МУЛЬТИМЕДИЙНОГО КОНТЕНТА В СИСТЕМАХ АНТИПЛАГИАТА

М. Т. Саидова, Р. Ш. Гасанова, Ф. Ш. Аскеров
Институт информационных технологий НАН Азербайджана, Баку

Представлена информация о плагиате и способах борьбы с ним. Приведены примеры плагиата известных политиков. Исследованы проблемы, связанные с проверкой мультимедийного контента в системах антиплагиата. Предложено использование алгоритма Copyright Match Tool для выявления плагиата видеоконтента в системах антиплагиата.

Введение

Плагиат – это представление мыслей и идей другого автора от своего имени. Он включает такие виды киберпреступности, как компьютерные вирусы, спам, фишинг и т. п. [1]. Плагиат происходит путем копирования текста из первоисточников (книг, журналов, материалов из сети Интернет и др.) полностью или с небольшими изменениями.

В настоящее время чаще используются два вида плагиата – текстовый и исходного кода. При текстовом плагиате информация копируется в неизменном, перефразированном виде (или с использованием синонимов) из оригинальной статьи, диссертации, научной работы и других источников без ссылок. При плагиате исходного кода весь текст или его часть копируется путем изменения кода без указания использованной литературы, поэтому этот вид плагиата обычно бывает трудно определить.

Стремительное развитие Интернета, доступность любой информации в виртуальном пространстве приводят к распространению случаев плагиата. В наше время крадут и текстовую информацию, и мультимедийные файлы, поэтому так актуальна их проверка в системах антиплагиата.

1. Методы борьбы с плагиатом

Плагиат можно определить двумя способами: физическим и автоматическим [1]. При физическом определении эксперт проверяет материал и решает, является он плагиатом или нет. Однако этот метод не эффективен для проверки большого количества документов, так как требует от специалиста больших затрат времени и сил. Автоматическое обнаружение осуществляется с помощью различных систем антиплагиата, таких как Turnitin, AntiPlagiarism.NET, PlagiarismDetection, PlagAware, iThenticate, Ouriginal. Кроме применения таких систем для борьбы с плагиатом используются также правовые и просветительские методы, в том числе общественное осуждение [2].

Юридический метод связан с формированием правовой базы. На сегодняшний день во многих странах плагиат считается уголовным преступлением, таким как мошенничество либо кража, и правонарушитель обязан возместить ущерб.

Просветительский метод – еще один метод борьбы с плагиатом. Этот процесс включает в себя подготовку учебников и специалистов в области борьбы с плагиатом, а также преподавание соответствующих дисциплин в высших учебных заведениях.

Международный опыт показывает, что общественное осуждение является одним из работающих механизмов наказания плагиаторов. Виновные в плагиате лишаются своих должностей и ученых званий без привлечения к юридической ответственности.

2. Случаи плагиата в научной среде и политике

С развитием Интернета стремительный рост плагиата отразился и на науке в Азербайджане. Так, согласно отчету Высшей аттестационной комиссии (ВАК) при Президенте Азербайджана количество соискателей ученых степеней и ученых званий после 2015 г. уменьшилось, а количество отказов в связи с выявлением плагиата в представленных работах увеличилось. Согласно статистическим данным ВАК в 2021 г. было рассмотрено 400 заявлений соискателей: 42 (10,5 %) были отклонены по причине плагиата, из них 30 соискателям (в том числе пяти из-за рубежа) было отказано в присвоении ученой степени доктора философии, 10 соискателям – звания доцента, двум – звания профессора (URL: http://www.aak.gov.az/news/102).

Если рассмотреть некоторые интересные случаи плагиата в мировой практике, то можно увидеть, что его элементы были выявлены в работе известного европейского политика (URL: https://www.politico.eu/article/politicians-plagiarism/). Вики-блог VroniPlag, который проверяет, не являются ли академические публикации плагиатом в Интернете, обнаружил, что 12 % его работ были плагиатом. Элементы плагиата были обнаружены также на 200 из 215 страниц докторской диссертации президента одной европейской страны. В 2012 г. в еженедельнике Heti Vilaggazdasag сообщалось, что 16 страниц его диссертации скопированы дословно.

С такими случаями можно столкнуться не только в научной среде, но и в политике (URL: https://blog.scanmyessay.com/2019/03/26/the-top-five-famous-cases-of-plagiarism/). Например, во время инаугурации президента США в 2016 г. в речи первой леди был обнаружен плагиат. Ее выступление по содержанию и структуре было идентично выступлению другой первой леди в 2008 г., что вызвало большой интерес в социальных сетях. Еще один интересный факт о плагиате в политической среде связан с одним из лидеров афроамериканского правозащитного движения. Так, выражения, которые он использовал во время выступления на митинге численностью 250 тыс. чел. в Вашингтоне в 1963 г., были взяты из его же докторской диссертации, написанной в 1955 г. Из этого нашумевшего инцидента можно сделать вывод, что использование автором своих прошлых работ без ссылок также считается плагиатом.

3. Ограничения в системах защиты от плагиата

Поскольку алгоритмы, используемые в системах антиплагиата, могут проверять только текстовые файлы, они не могут анализировать аудио- и видеофайлы и их субтитры. Хотя системы защиты от плагиата анализируют информацию текстового типа, существуют также и связанные с ними ограничения. Например, могут обрабатываться только текстовые типы файлов, ранее введенные в системы антиплагиата, такие ограничения есть в системе антиплагиата Turnitin, используемой во многих международных академических базах данных, издательствах и топовых университетах без языковых ограничений. Система может проверять сходство только файлов типа .html, .doc/.docx, .hwp, .odt, .rtf, .txt, .wpd, .ps (URL: https://help.turnitin.com/feedback-studio/canvas/plagiarism-framework/student/the-similarity-report/accepted-file-types-and-sizes.htm). Тип файлов .pdf не может быть проверен, поскольку отсканированные данные обычно хранятся в виде изображения, а не текста. Вместе с тем система легко проверяет pdf-файлы с выбираемым текстом, проверяются также файлы PowerPoint. Так система очищает отправленный файл от видео и анимации (теней, 3D-эффектов и т. д.), конвертирует его в формат pdf и проверяет исходный текст (URL: https://supportcenter.turnitin.com/s/article/Troubleshooting-PowerPoint-files). Файлы Excel проверяются аналогичным образом.

Многие системы антиплагиата при проверке могут сопоставлять представленный контент с текстами в разделе описания видео на YouTube, а также с расшифровками видеоконтента (при наличии в Интернете), а некоторые системы могут даже обнаружить плагиат в формате изображения. Тем не менее проблема прямой обработки аудио- и видеоконтента в системах антиплагиата остается нерешенной.

В настоящее время проверить, звучал ли ранее аудиоконтент, т. е. не является ли он плагиатом, можно с помощью программ Shazam, SoundHound, Musixmatch, Musipedia и др. (URL: https://www.makeuseof.com/original-melody-avoid-plagiarism-tips/). Каждый музыкант должен удостовериться, что его произведение ранее не звучало ни в одной песне. Музыкант не ворует чью-то песню осознанно, а считает ее оригиналом, когда воспроизводит мелодию, которую слышал много лет назад, подсознательно помнил и давно забыл. Примером плагиата является популярная в 1970-е гг. песня «My Sweet Lord» Джорджа Харрисона из группы «Битлз» (URL: https://consequence.net/2018/01/10-famous-instances-of-alleged-music-plagia-rism/). В суде было доказано, что она идентична песне He is So Fine группы The Chiffons (URL: https://www.radiox.co.uk/features/x-lists/most-famous-accusations-of-musical-plagiarism/).

В соцсети YouTube есть инструмент для автоматической идентификации видеоконтента в социальных сетях (URL: https://support.google.com/youtube/answer/7648743/). Если будет обнаружено, что видеоконтент ранее был размещен на другом канале, то YouTube предоставляет пользователю три варианта защиты авторских прав: включение видео в свой архив, при этом трансляция видео в соцсети прекращается, его может видеть только пользователь; удаление видео со своего канала; возможность переписки с пользовательским каналом, который ранее разместил видео.

Заключение

Системы антиплагиата не могут обрабатывать графические, аудио- и видеофайлы, поскольку предназначены для анализа только текстового контента. Для обеспечения возможности обнаружения плагиата в аудиофайлах в системах антиплагиата предлагается использовать алгоритм, применяемый в программном обеспечении обнаружения плагиата в музыкальных файлах. Целесообразно также в системе антиплагиата применить алгоритм, основанный на принципе работы Copyright Match Tool (в соцсети YouTube), для выявления плагиата видеоконтента.

Наряду с вышеуказанным одним из ключевых моментов является сравнительный анализ разных видов контента. Сравнительный анализ содержания аудио- или видеофайла путем преобразования его в файл текстового типа позволит получить более полное представление о том, является ли содержание плагиатом или нет.

Опыт преподавания предмета «Борьба с плагиатом» в зарубежных высших учебных заведениях показывает, что подготовка учебников и исследовательские работы студентов по данной теме могут быть одними из лучших способов борьбы с плагиатом.

Список литературы

1. Hasanova, R. Sh. Conception of creating the national anti-plagiarism service in Azerbaijan / R. Sh. Hasanova, F. Sh. Asgarov // Problems of Information Society. – 2021. – № 1. – P. 88–93.

2. Əliquliyev, R. M. Azərbaycanda plagiatliqla mübarizə problemləri və həll yolları / R. M. Əliquliyev, R. M. Alıquliyev, R. Ş. Mahmudov // İnformasiya cəmiyyəti problemləri. – 2019. – № 1. – S. 34–43.