УДК 004.04

АВТОМАТИЗАЦИЯ ОПРЕДЕЛЕНИЯ ПЛАГИАТА В УЧЕБНОМ ПРОЦЕССЕ

Попова Ю.Б., Голобурда А.А.
Белорусский национальный технический университет
Минск, Республика Беларусь

В последнее время в учебном процессе наблюдается бурный рост использования заимствованной информации, которую можно отнести к разряду плагиата. Согласно определения, приведенного в [1], плагиат — это умышленное присвоение авторства чужого произведения науки или искусства, чужих идей или изобретений. Поэтому задача обнаружения недобросовестного использования чужих работ (фактов плагиата) в учебных заведениях приобретает высокую актуальность.

Для проверки документов на плагиат в сети Интернет существует несколько программных продуктов, рассмотренных в [2]. Однако не все обучающиеся размещают свои работы во всемирной сети и часто передают их младшим курсам. Поэтому в учебных заведениях существует проблема плагиата работ с прежних лет. Одним из решений указанной проблемы является использование автоматизированных систем управления обучением (англ., Learning Management System, LMS) [3]. Обучающиеся закачивают свои лабораторные, контрольные и курсовые работы в такие системы, образуя архивы прошлых лет, и преподавателю не составляет труда проверить работы на плагиат. Ключевым моментом здесь является возможность LMS делать такие проверки. Поэтому авторами предлагается веб-сервис, который может быть интегрирован в любую LMS, для поиска заимствованных работ и составления отчетов с указанием процентов схожести с аналогами.

Разработанный веб-сервис реализует модифицированный алгоритм векторной модели текстового документа. Предлагаемая модификация заключается в формировании отдельного *N*-списка для каждого анализируемого документа. Вследствие чего будет происходить попарное сравнение документов и формирование образа одного документа относительно *N*-списка другого. Таким образом, в *i*-й строку матрицы схожести будут записываться коэффициенты схожести всех рассматриваемых документов относительно *i*-го документа. Предлагаемая модификация позволяет ускорять процесс вычислений, поскольку нет необходимости искать общие термы для всех документов.

Следует отметить, что в ходе учебного процесса преподаватель сталкивается с большим количеством работ студентов, которые необходимо проверять одновременно, например, контрольные или курсовые работы целого потока обучающихся. Поэтому проверять на плагиат каждую работу отдельно – это очень затратный процесс как по времени выполнения, так и по вычислительным ресурсам. Для решения

данной проблемы авторы предлагают использовать кластерный анализ текстовой информации, основная цель которого — разбить множество объектов на группы таким образом, чтобы объекты внутри одной группы были максимально похожи друг на друга, но в то же время максимально отличались от объектов другой группы. Алгоритм кластерного анализа текстовой информации реализован на языке программирования Java, интегрирован в систему управления учебным процессом, разработанную и используемую на кафедре программного обеспечения вычислительной техники и автоматизированных систем БНТУ, и находится на стадии апробации.

Для проведения вычислительных экспериментов было подготовлено два тестовых набора документов объемом 1500-2000 слов. Первый набор состоял из 50 документов, а второй – из 100. Результаты вычислений показали, что время проверки одного документа на плагиат из набора пятидесяти документов составляет 88 секунд. Такое же количество времени требуется для проверки на плагиат всех пятидесяти документов, применяя кластерный анализ. С увеличением количества проверяемых документов в два раза время вычислений увеличивается приблизительно в такое же количество раз. Таким образом, проведенные эксперименты показали преимущество использования кластерного анализа при определении плагиата: данный подход позволяет за один раз найти сразу все варианты схожих работ, что существенно экономит время поиска.

В ходе экспериментов было также замечено, что при повторных проверках документов время вычислений может быть существенно сокращено за счет первоначального сохранения результатов морфологического разбора. Поскольку сравнение каждого нового документа происходит с ранее проверенными документами, то для них нет необходимости делать морфологический разбор каждый раз, когда формируется матрица схожести, а можно брать сохраненные варианты предыдущих разборов.

1. Бобкова, О. В. Плагиат как гражданское правонарушение / Бобкова О. В., Давыдов С. А., Ковалева И. А. // Патенты и лицензии. – 2016. – № 7. – С. 31–41.
2. Голобурда, А. В. Проверка плагиата в веб-приложениях / А. В. Голобурда, Ю. Б. Попова // Информационные технологии в образовании, науке и производстве: IV Международная научно-техническая интернет-конференция, 18-19 ноября 2016 г. Секция Информационные технологии в производстве и научных исследованиях [Электронный ресурс]. – Режим доступа: http://rep.bntu.by/handle/data/27126 – Дата доступа: 25.11.2017.
3. Попова, Ю.Б. Классификация автоматизированных систем управления обучением / Попова Ю.Б. // Системный анализ и прикладная информатика. – 2016. – №2. – С. 51–58.