А.В. Слива, *кандидат технических наук*
В.Н. Фокина, *кандидат социологических наук, доцент*
А.В. Абрамова
М.Е. Широкова, *кандидат социологических наук*

Методы совершенствования программных средств выявления плагиата

В статье рассмотрены проблемы использования программных средств выявления плагиата, среди которых: отсутствие учета легальных заимствований (цитат), списков литературы, наименований приведенных в тексте документов, самоцитирования автора. Необходима законодательная формализация и разработка единого стандарта по оформлению в электронном виде всех типов письменных работ, открытость экспертиз и их результатов, свободный доступ к полнотекстовым материалам научных работ, постоянный мониторинг новых ухищрений и разработка мер противодействия.

***Ключевые слова*:** *плагиат, программные средства выявления плагиата, оригинальность научных публикаций.*

В настоящее время проблема плагиата при написании рефератов, курсовых и дипломных работ, а также диссертаций во всем мире приобрела поистине огромные масштабы. По различным данным, в плагиате замешаны до 70% студентов, а скандалы с разоблачением ученых и политиков, «списавших» свои докторские диссертации, уже достаточно давно воспринимаются как нечто рутинное. На этом фоне вполне естественным является разработка мер по борьбе с плагиатом, в том числе массовое использование программных средств выявления плагиата.

За последние десятилетия во всем мире разработано значительное количество таких программных средств. Модули выявления плагиата встроены практически во все получившие широкое распространение за рубежом интеллектуальные системы оценивания эссе (IEA), такие, как, например, TurnItIn (http://www.turnitin.com), Criterion (ETS – Educational Testing Service) [1], Intelligent Essay Assessor (Pearson Education Technologies Inc.) [2, p. 319–330], IntelliMetric (компания Vantage Learning) [3, p. 49–62], Project Essay Grade (PEG) (Measurement Inc.) [4, p. 127–142] и т. д. Программные средства проверки на наличие плагиата разработаны и в России. Среди них наиболее популярными

являются Text.ru, Istio.com, Copyscape.com, Wsgu.ru/servis/copy.php, AntiPlagiat.ru, Advego Plagiatus (http://advego.ru/plagiatus), eTXT Антиплагиат (http://www.etxt.ru/antiplagiat), Miratools.ru. Ежегодно с помощью указанных программных систем проверяются миллионы письменных работ, и эти системы уже доказали свою эффективность: и в России, и за рубежом их используют ведущие вузы, научные учреждения, а также государственные структуры.

В России, особенно в последнее время, вопросы борьбы с плагиатом постоянно находились в фокусе внимания академической общественности. Серьезность проблемы вывела ее обсуждение даже на уровень руководства страны. Так, в сентябре 2012 г. в СМИ [5] появилось сообщение о том, что Председатель Правительства Российской Федерации Д.А. Медведев поручил чиновникам в кратчайшие сроки разработать порядок научных публикаций и ввести их тотальную проверку на наличие плагиата. Причем это должно касаться как научных статей и книг, так и диссертаций, дипломных работ и т. д., вплоть до курсовых и рефератов. Весьма важно, что диссертации и дипломные работы должны будут в обязательном порядке выкладываться в открытый доступ. Это, несомненно, важно с позиций расширения свободного доступа к научной информации. С точки зрения борьбы с плагиатом такая публичность – это, с одной стороны, возможность проверки любой работы на наличие плагиата, а с другой – постоянное расширение базы поиска незаконных заимствований в новых публикациях.

Казалось бы, что наличие такого внимания к оригинальности научных публикаций и учебных работ имеет одни только плюсы. Однако, анализ существующих в России программных систем антиплагиата и практики их использования выявил ряд принципиальных проблем, которые необходимо решить до перехода к практике тотальной проверки на наличие плагиата диссертационных, дипломных, курсовых работ, рефератов, научных публикаций и т. п. При этом одной из наиболее существенных является проблема отсутствия в этих системах учета легальных заимствований – правильно оформленных цитат. Системы, как правило, выдают долю заимствованного текста, включая в нее, в том числе, правильно оформленные цитаты.

На сегодня общими документами, регламентирующими оформление научных работ, являются ГОСТ 7.32-2001 «Система стандартов по информатизации, библиотечному и издательскому делу. Отчет о научно-исследовательской работе. Структура и правила оформления» и ГОСТ 7.0.11-2011 «Система стандартов по информатизации, библиотечному и издательскому делу. Диссертация и автореферат диссертации. Структура и правила оформления». Однако практически каждый научный журнал имеет свои собственные требования к оформлению статей, имеется также множество требований к оформлению письменных студенческих работ, которые устанавливаются каждым вузом самостоятельно. Кроме того, ни в одном ГОСТ нет строго узаконенных, формализованных единых требований и правил решения частного, но очень важного для систем антиплагиата вопроса оформления цитат. Самый распространенный способ оформления цитат – выде-

ление кавычками. Но применяются и другие способы, например, использование отчеркивающей линейки в отступе:

цитата должна быть уместной и точной, недопустимо использование чужого текста без ссылки на автора и произведение.

Если цитируется стихотворный текст, то кавычки обычно не ставятся; стихотворный текст размещается между строчками, например, в учебнике русского языка написано:

«В поэме А. Твардовского «Василий Теркин» лейтмотивом всего произведения являются следующие строки:

Бой идет святой и правый.
Смертный бой не ради славы,
Ради жизни на земле» [6].

Для того чтобы решить проблему, как не учитывать в тексте цитаты в процессе выявления плагиата, необходимо законодательно ввести единое правило оформления цитат. Тогда возможно программными средствами реализовать поиск плагиата по текстовым массивам, из которых исключены правильно оформленные цитаты.

Однако и это не решит окончательно проблему определения плагиата: остаются еще списки литературы и наименования приведенных в тексте документов. Их оформление также необходимо формализовать. С этой проблемой на практике при проверке студенческих работ столкнулась Современная гуманитарная академия (СГА) при внедрении интеллектуального робота контроля оригинальности и профессионализма (ИР КОП) [7, с. 16–25] (свидетельство о государственной регистрации от 31.03.2011 № 2011613972). Так, например, в одной из письменных студенческих работ, тематика которой была связана с анализом некоторой совокупности ГОСТ, ИР КОП отнес к плагиату все наименования этих ГОСТ. Анализ студенческих форумов показал, что аналогичная программа одного из вузов Южной Кореи признала заимствованными около 50% текста письменной работы, поскольку согласно требованиям вуза в ней был приведен текст каждого задания, которые программа восприняла как плагиат.

В целом, анализ практики вузов различных стран показал, что в большинстве из них работы выборочно передаются на проверку преподавателям после их автоматизированного контроля на плагиат (Франция, Испания и др.). Причина – нельзя исключать случаев, когда даже при наличии нелегальных заимствований, в работе могут быть ценные оригинальные мысли и решения, которые позволяют оценить такую работу положительно.

Вероятно, все же не следует безоговорочно полагаться на результаты работы компьютерных программ выявления плагиата. Прежде чем приступить к тотальному внедрению программных систем антиплагиата, необходима разработка единого ГОСТ по оформлению в электронном виде всех типов письменных работ (диссертации, статьи, дипломы, курсовые, рефераты, отчеты по НИР и т. д.). Уровень формализации такого ГОСТ должен являться достаточным

для исключения всех компонент, текст которых совпадает с текстом других документов, но не является плагиатом. Возможно, целесообразно сделать такой ГОСТ обязательным для всех издательств, что существенно упростит работу авторов при представлении работ.

Такой ГОСТ, по-видимому, должен включать следующие компоненты:

• общие правила форматирования текста (поля, размер шрифта, межстрочный интервал);

• однозначно трактуемые правила использования цитат;

• набор символов, выделяющий (открывающих и закрывающих) списки названий (документов, географических наименований, исторических персонажей и пр.);

• набор символов, отделяющий текст работы от списка литературы.

Еще одна проблема систем антиплагиата, которую также необходимо решить до начала массового их использования, – это проблема правомерного использования автором без выделения в виде цитат ранее написанных им лично оригинальных текстов. Такая ситуация вполне возможна, например, у целеустремленного студента, у которого дипломная работа вполне может явиться итогом его исследовательской деятельности, частично отраженной, например, в его курсовых работах, статьях и отчетах. И уже обязательно с этой проблемой столкнется соискатель ученой степени доктора наук, поскольку законодательно определено, что соискатель должен обязательно опубликовать не менее 50% содержания диссертации в рецензируемых научных журналах и изданиях.

По-видимому, при проверке на плагиат указанных работ необходимо решить вопрос о введении списка работ конкретного автора (диссертаций, статей, монографий, научных отчетов, курсовых работ и пр.), которые должны быть исключены программными средствами из базы сверяемых работ, особенно при проверке диссертаций. Наличие такого списка (пусть даже пустого, что будет означать анализ на плагиат по полной базе) следует отразить в стандарте на оформление работ, установив набор символов, выделяющих такой список.

При этом даже в случае выявления программой проверки плагиата, следует проявлять осторожность в оценках, так как нельзя исключить, что именно та работа, с которой могут быть выявлены текстовые совпадения, сама является «плагиатом» с более ранних работ автора. Поэтому следует аккуратно подходить к формированию базы системы выявления плагиата, тщательно проверяя на наличие плагиата каждую новую работу, включаемую в базу системы. А в случае выявления плагиата следует проверить на плагиат те элементы информационной базы системы, с которыми выявлено совпадение, взяв за базу работы автора, датируемые сроками, предшествующими датам публикации указанных элементов.

Кроме того, необходимо также при этом проверять уже совсем казусные ситуации, которые, тем не менее, имели место на практике. Например, возможное обвинение в плагиате соискателя степени доктора наук Петровой, за списывание с кандидатской диссертации Ивановой в случае, когда Иванова после защиты кандидатской диссертации поменяла фамилию на Петрову.

Особая аккуратность, к которой мы призываем при рассматриваемой вполне реальной перспективе перехода к тотальным проверкам работ на наличие плагиата, связана с тем, что, как было показано выше, результат проверки в значительной степени определяется особенностями программы проверки и базы априори оригинальных текстов, относительно которых анализируется наличие плагиата. Система может принимать за плагиат некоторые оригинальные и аутентичные тексты, например, самоцитирование автора. Необходимо сделать все, чтобы устранить даже потенциальную возможность недобросовестной конкуренции. Можно представить ситуацию, когда недоброжелатели, зная особенности конкретного программного комплекса поиска плагиата, организуют проверку оригинальной докторской диссертации конкурента именно через такую базу. В результате система обязательно покажет наличие плагиата, хотя на самом деле его не было. При планируемой открытости доступа к студенческим работам вузов, нетрудно представить себе аналогичную ситуацию с дипломными работами его выпускников, причем в массовом порядке, что даст возможность конкурентам подорвать деловую репутацию вуза. Отметим, что если даже в случае недобросовестной проверки в дальнейшем и будет восстановлена справедливость, то репутационный ущерб останется. Как еще в XIX в. писал А.Н. Апухтин: «он ли украл или у него украли... Главное то, что он был замешан в гадком деле» [8].

Здесь выход только один – обязательная открытость экспертиз, указание полного библиографического описания источников заимствований в случае обнаружения плагиата и публикации результатов экспертиз: на каких текстовых базах проводилась проверка, какие конкретно фрагменты списаны и откуда. Следует также обеспечить свободный доступ к полнотекстовым материалам, откуда проведено заимствование. Такой подход тем более важен на фоне требований к выкладыванию вузами ВКР и диссертаций в открытый доступ.

Необходимо обеспечить и возможность проведения повторной независимой экспертизы для оспаривания результатов первичной экспертизы в суде. Это должно сопровождаться жестким наказанием всех участников фальшивых экспертиз, публикацией опровержения их результатов, а также административным, а возможно, гражданским (возмещение морального вреда) и уголовным (закон о клевете) делопроизводством при обнаружении фальшивых экспертиз. Кроме того, при любых внешних проверках на плагиат как отдельных письменных работ, так и массовых проверках работ студентов какого-либо вуза, текстовая база системы антиплагиата должна быть сертифицирована, как и эксперты, проводящие проверку.

По нашему мнению, также необходимо узаконить предельную долю правильно оформленных цитирований в тексте письменных работ. Например, для кандидатских и докторских диссертаций установить ее на уровне 15–25%, иначе можно будет без боязни обвинения в плагиате «слепить» диссертацию целиком из правильно оформленных цитат.

Практика работы СГА по автоматизированному поиску плагиата в ИР КОП позволила определить еще несколько проблем такой проверки. В частности, было

выявлено использование отдельными недобросовестными студентами маскирующих плагиат слов-паразитов. Это слова типа «ежели», «промеж», «как бы», «само собой», «как многие выражаются», «вроде бы» и т. п., которые эти авторы в большом количестве вставляют с определенной периодичностью в избранный для списывания текст с целью выдать его за оригинальный. В процессе отработки специалистами СГА программного средства поиска плагиата, среди работ, определенных предыдущей версией системы антиплагиата как оригинальные тексты, были выявлены, например, работы на базе одного и того же исходного текста, который был «разбавлен» словами-паразитами и добавлением слов в искаженной форме типа «окромя», «опосля», «вопросец», «благоприязнь», «снутри» и т. д. Наблюдались и махинации с разделением слов на части, например вместо «верификация» было написано «вери фикация» – вроде бы это опечатка, но предыдущая версия программы антиплагиата списывание в этом случае уже не определяла. Аналогичную ситуацию отмечают исследователи систем антиплагиата, причем не только в студенческих, но и в других категориях работ [9, с. 86–97].

Итак, практика применения программных средств антиплагиата показала, что недобросовестные авторы, сталкиваясь с новым для себя препятствием, активно разрабатывают средства противодействия. В целом, рассмотренная ситуация является довольно естественной – на каждое усовершенствование средства защиты противоборствующая сторона находит в них новые лазейки. Любая система проверки не идеальна, имеет свои слабые стороны, всегда есть способ обмануть систему. Опыт использования ИР КОП в СГА показал, что студенты постоянно искали новые пути обхода, несмотря на постоянное усовершенствование системы. Если систематизировать деликты, то среди наиболее распространенных можно отметить следующие:

- общепринятые в употреблении слова заменены устаревшими, просторечными, уменьшительно-ласкательными;
- включено большое количество вводных слов;
- многократное разделение/слияние слов, слова не написаны полностью;
- намеренно заменен порядок слов;
- вставки лишних знаков (непечатных символов, знаков переноса, скрытого текста в Word);
- буквы в словах заменены цифрами, латинскими буквами;
- вставки бессмысленного текста;
- двойной перевод текста (с русского на английский и обратно);
- неоднократный повтор отдельных фрагментов текста.

Несмотря на доработку программных комплексов в части отдельных типов деликтов (замена букв цифрами, латинскими буквами, неоднократный повтор фрагментов текста, ссылки на устаревшие нормативные акты и др.) недобросовестные студенты используют для повышения оригинальности своей работы другие виды деликтов, например, вставки согласованного грамотного текста, но не относящегося к теме работы.

Выход из этой ситуации может быть только один, аналогичный тому, который ведется разработчиками антивирусных программ – непрерывный мониторинг экспертами новых «трюков» недобросовестных авторов путем выборочной проверки работ, как положительно, так и отрицательно оцененных системой антиплагиата, и самостоятельный поиск слабых мест в ее системе и их оперативное устранение – добавление «противоядий» к этим новым трюкам. Для этого, в частности, программные средства антиплагиата целесообразно пополнять различными словарями слов-паразитов, слов-искажений и др., а также частотным анализом употребления таких слов, что поможет фиксировать указанное мошенничество в выдаваемых за оригинальные тексты. Перед проверкой автоматически удалять слова-паразиты, слова-искажения и пр., тогда система поиска плагиата уже не будет обманута.

В совокупности с введением предложенного выше стандарта на оформление письменных работ и учетом при проверках особенности анализа текстов, включающих более ранние материалы авторов, и другие изложенные выше соображения, предложенные усовершенствования позволят значительно поднять эффективность программных средств поиска плагиата, хотя и не обеспечат их 100%-ю непогрешимость. Систематическая работа по совершенствованию программных средств может не только значительно сократить случаи плагиата, но и позволит формировать интеллектуальную аллергию к академической недобросовестности.

Литература

1. Burstein J., Chodorow M., Leacock C. Automated Essay Evaluation: The Criterion Online Writing Service // AI Magazine. 2004. Vol. 25. № 3.

2. Valenti S., Neri F., Cucchiarelli A. An Overview of Current Research on Automated Essay Grading // Journal of Information Technology Education. 2003. Vol. 2.

3. Dikli S. Automated Essay Scoring // Turkish Online Journal of Distance Education – TOJDE. 2006. January. Vol. 7. № 1. Art. 5.

4. Page E. B. New computer grading of student prose, using modern concepts and software // Journal of Experimental Education. 1994. № 62 (2).

5. Лемуткина М. Плагиату поставят мат // Московский комсомолец. 2012. 17 сентября.

6. Бархударов С.Г., Крючков С.Е. Русский язык. 8 класс: Учебник. М.: Просвещение, 2011.

7. Карпенко М.П., Фокина В.Н., Абрамова А.В. Интеллектуальные роботы для автоматизированного оценивания письменных творческих работ // Инновации в образовании. 2012. № 9.

8. Апухтин А.Н. Неоконченная повесть: Сочинения. Стихотворения и проза. М.: Художественная литература, 1985.

9. Фокина В.Н., Слива А.В., Семенова Т.Ю., Абрамова А.В. Проблема академической недобросовестности и некоторые пути ее решения в высшей школе // Инновации в образовании. 2012. № 11.

Sliva A.V., *Candidate of Technical Sciences*
Fokina V.N., *Candidate of Sociological Sciences, Associate Professor*
Abramova A.V.
Shirokova M.E., *Candidate of Sociological Sciences*

Methods to Improve Plagiarism Detection Software

The article covers the problem of the plagiarism detection software implementation, such as the lack of consideration of quotes (citations), reference lists, names of documents, self-citation of the author. It is necessary legislative formalization and development of a unified standard for electronic form of all types of written works, the openness of the examinations and their results, free access to full-text content of scientific works, constant monitoring of new tricks and countermeasures' development.

***Key words*:** plagiarism, plagiarism detection software systems, originality of scientific publications.