ИССЛЕДОВАНИЕ ПЛАГИАТА В РАБОТАХ СТУДЕНТОВ

Шарапов Р. В. (info@vanta.ru),
Шарапова Е. В. (mivlgu@mail.ru)

Муромский институт (филиал) ФГБОУ ВПО «Владимирский государственный университет имени Александра Григорьевича и Николая Григорьевича Столетовых», Муром, Россия

В работе проводится исследование студенческих работ на наличие заимствований (плагиата). Дается обзор подходов к сокрытию фактов заимствований. Проанализированы базы контрольных и курсовых работ, а также база студенческих статей.
Анализ показал, что для сокрытия фактов плагиата наиболее часто употребляются: изменение заимствованного текста, сокращение текста, корректировка родов, чисел и времен слов, замена русских букв на сходные по написанию латинские. При усилении контроля за оригинальностью работ увеличиваются попытки скрыть факт заимствований за счет изменения текста и увеличения числа используемых источников.

Ключевые слова: плагиат, обнаружение плагиата, заимствование, сокрытие плагиата, студент, замена букв, изменение текста, курсовая работа, контрольная работа

RESEARCH OF PLAGIARISM IN THE STUDENT WORKS

Sharapov R. V. (info@vanta.ru),
Sharapova E. V. (mivlgu@mail.ru)

Murom Institute of Vladimir State University, Murom,
Russian Federation

In this paper we study plagiarism in student works. We consider as plagiarism a direct unattributed copy of a text. We describe how students try to hide plagiarism. We have analyzed a collection of tests, course papers, and student working papers. When writing their own work (a course paper or a student working paper), students usually copy material from just one source, while when writing a test, they usually use more sources. We found that the methods used to hide the plagiarism usually include changing of text; reduction of text; changing of certain pronouns, adjectives and verbs; replacement of Russian letters by similar Latin ones. Among the less frequently used methods are replacements of parts of text, changed punctuation, substitution of invisible characters for spaces, manual and automatic substitution of synonyms, copying texts from some sources and changing them. Methods of hiding plagiarism may change over time if they become less effective.

Key words: plagiarism, detection, of plagiarism, hiding plagiarism, students, replacement of letters, changing of text, course paper, student working paper.

Введение

Бурное развитие вычислительной техники привели к глубокому проникновению компьютеров в нашу жизнь. Компьютеры окружают нас везде — на работе, дома, в магазинах и общественных местах. Современное развитие информационных технологий и глобальной сети Интернет предоставило широким кругам пользователей доступ к огромным массивам информации. Появилось большое число online-библиотек, содержащих в электронном виде художественную и научно-техническую литературу. Стало возможным читать книги, новости и газеты непосредственно с экрана компьютера.

В сети Интернет стало доступно множество методических указаний, курсов лекций, учебников и т. д. Кроме того, появились огромные коллекции рефератов, готовых лабораторных работ, курсовых и дипломных проектов и даже диссертаций. Использование компьютерной техники сильно облегчило задачу поиска и копирования подобной информации. Если раньше для написания реферата или контрольной работы информацию было нужно, по крайней мере, найти в книгах и переписать (вручную, перепечатать или ввести в компьютер

с помощью сканера и программ распознавания текстов), то теперь достаточно ввести название темы в поисковую систему и скопировать найденные материалы. Стал распространяться метод написания работ, получивший название «Copy & Paste». Метод заключается в простом копировании информации из одного или нескольких источников с минимальным редактированием получающегося таким образом текста.

Аналогичная ситуация наблюдается с отчетными материалами внутри учебных заведений. В связи с тем, что большое число пояснительных записок по курсовым и дипломным проектам выполняется с использованием компьютеров, происходит их распространение и повторное использование среди учащихся.

В последнее время наблюдается бурный рост использования в учебном процессе подобной заимствованной информации. Ситуация усугубляется тем, что учащиеся иногда не знают (не читают) то, что написано и «их» работах.

Плагиат — умышленное присвоение авторства чужого произведения науки или искусства, чужих идей или изобретений [1].

Как можно убедиться из определения, подобные заимствованные работы можно отнести к разряду плагиата. Задача обнаружения недобросовестного использования заимствованных текстов в учебных кругах (фактов плагиата) приобретает высокую актуальность. Не менее важной задачей следует считать всестороннее исследование подобных заимствований, что может помочь в их дальнейшем автоматизированном выявлении.

Постановка задачи

Среди всех видов заимствований, встречающихся в современном обществе, мы будем рассматривать только наиболее распространенный в студенческой среде — использование в своих работах чужих материалов без какого-то ни было собственного вклада. Чаще всего это простое копирование доступных материалов в свою работу, без их переработки (как это могло бы быть, например, при пользовании учебниками, курсами лекций или методическими указаниями). Иногда такой вид заимствований называют «неприкрытым копированием».

Цель работы — провести всестороннее исследование заимствований, встречающихся в студенческих работах (как учебных, так и научных). Интерес представляет то, как часто делаются попытки скрыть факт заимствований, какие подходы используются для этого и т. д.

Подходы к сокрытию заимствований

Существует большое количество подходов к сокрытию фактов плагиата. В связи с тем, что чаще всего работы проверяются на плагиат различными

компьютерными системами (чаще всего системой «Антиплагиат» [2]), подходы призваны обмануть эти системы. По мере развития и совершенствования систем проверки подходы также эволюционируют.

Для того, чтобы скрыть факт заимствований, студенты могут применяться следующие подходы [3]:

1. Корректировка родов, чисел и времён входящих в текст слов. Например, замена слова «выполнил» на «выполнила» или «выполнили», использование местоимения «я» вместо «мы» в оригинальном тексте и т. д.
2. Незначительное изменение заимствованного текста. Например, изменение по одному слову в предложении.
3. Сокращение заимствованного текста путем удаления слов, предложений, абзацев, рисунков, формул и т. д.
4. Перестановка частей текста, абзацев и предложений местами.
5. Обход систем проверки на плагиат путем замены русских букв на аналогичные по написанию английские буквы и т. д.
6. Замена знаков препинания: «.» на «,» и обратно, « » на «.» и т. д.
7. Замена пробелов на невидимые буквы (написанные, например, белым цветом).
8. Осуществление ручной или автоматической синонимизации текста.

Исходные данные

В качестве объекта исследований мы использовали три базы работ.

Первая база («Conference») включает в себя материалы, подаваемые на научную конференцию студентами факультета (студенты были заинтересованы в подаче работ). Конференция проходит раз в год. В базу были включены материалы за три последних года. Количество работ ежегодно варьировалось в пределах от 200 до 400. Объем каждой работы — до 1000 слов.

Таблица 1. Состав работ базы «Conference»

Год	Число поданных работ	Оригинальных работ	Плагиат
2009	402	182	220
2010	369	201	168
2011	205	108	97

В 2009 году студенты не знали о факте проверки работ системой «Антиплагиат». В 2010 и 2011 году они были предупреждены заранее. Поэтому количество оригинальных работ увеличилось. С другой стороны, в работах стали применяться различные подходы к сокрытию фактов плагиата. Надо заметить, что под оригинальными работами мы понимаем работы, для которых не было найдено соответствия в сети Интернет и распространенных учебниках. Для оценки оригинальности кроме системы «Антиплагиат» использовались

и другие методы, в том числе авторская разработка «Автор.NET» [4]. Тем не менее, нельзя исключать тот факт, что некоторые работы, признанные оригинальными, все же могут содержать скрытые заимствования.

Вторая база работ («Course») включает в себя базу курсовых работ, выполненных студентами кафедры. Объёмы работ варьируются от 25 до 100 страниц.

Третья база («Control») состоит из контрольных работ, выполненных студентами заочной формы обучения по одной из гуманитарных дисциплин. Объёмы работ варьируются от 5 до 50 страниц.

Исследование подходов к сокрытию заимствований

Анализ базы «Conference» показал, что доля работ, применяющих различные подходы к сокрытию фактов заимствований, ежегодно растет (см. таблицу 2). Это объясняется, видимо, начавшейся борьбой с заимствованиями в студенческих работах.

Таблица 2. Процент работ в базе «Conference», скрывающих факты заимствований

Год	Работы, содержащие заимствования	Работы, скрывающие факт заимствований	Процент работ, скрывающих факт заимствований
2009	220	14	6,4%
2010	168	135	80,4%
2011	97	91	93,8%

Анализ показал, что подавляющее большинство работ копируются из одного источника (см. таблицу 2), чаще всего из учебников, статей из сети Интернет и публикаций региональной прессы. Тем не менее, число работ, компонуемых из двух и долее источников, ежегодно возрастает (см. таблицу 3). В 2011 году их доля достигла 15%. Это связано, вероятно, со знанием авторов о том, что их работы будут проверяться на плагиат.

Таблица 3. Количество источников для заимствований в работах базы «Conference»

Количество источников	Доля в 2009 г., %	Доля в 2010 г., %	Доля в 2011 г., %
Копирование текста из одного источника	98%	91%	85%
Копирование и компоновка текста из нескольких источников	2%	9%	15%

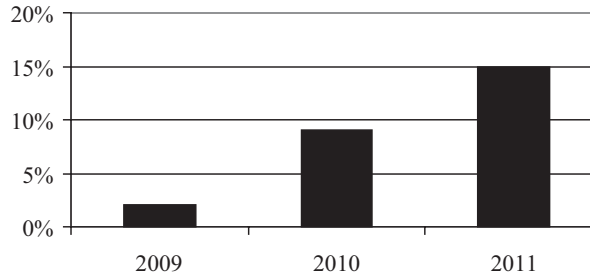


Рис. 1. Доля работ базы «Conference», скопированных и скомпонованных из нескольких источников

В связи с тем, что наиболее распространенной и доступной (как преподавателям, так и студентам) системой обнаружения заимствований является «Антиплагиат», большинство попыток скрыть факт заимствований нацелено именно на эту систему.

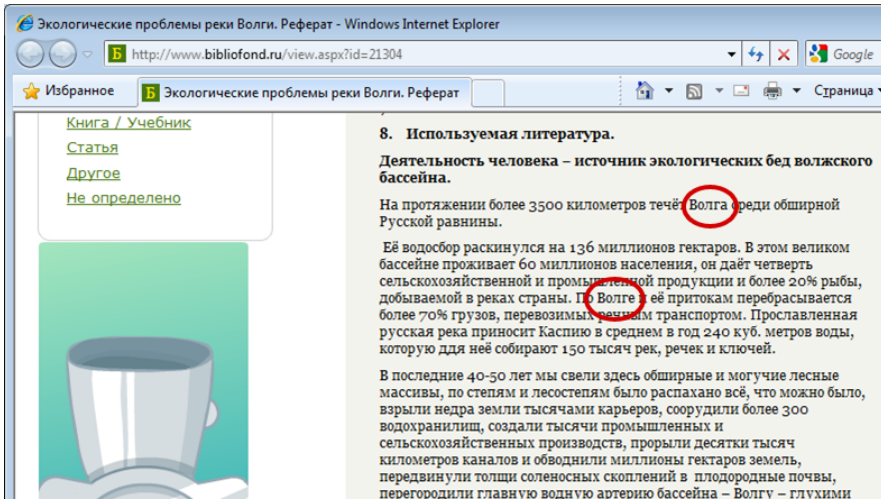
Анализ подходов, используемых студентами для сокрытия факта плагиата, показал, что в 33 % работ осуществлялась корректировка родов, чисел и времён слов (см. таблицу 4). Наиболее применяемый студентами подход к сокрытию факта плагиата — это незначительное изменение текста. Его доля достигла 89 % в работах 2011 года. Так, например, делались вставки слов и предложений в заимствованный текст, подвергались изменению названия населенных пунктов и рек (р. Волга в оригинале заменялась на р. Ока в статье). Надо заметить, что часть работ кроме заимствованных текстов содержала оригинальные блоки, чаще всего введение и заключение. Приведенная выше доля статей, подвергавшихся изменению, учитывает только заимствованные части таких текстов.

Экологические проблемы Оки

На протяжении более 3500 километров течёт Ока среди обширной Русской равнины.

Её водосбор раскинулся на 136 миллионов гектаров. В этом великом бассейне проживает 60 миллионов населения, он даёт четверть сельскохозяйственной и промышленной продукции и более 20% рыбы, добываемой в реках страны. По Оке и её притокам перебрасывается более 70% грузов, перевозимых речным транспортом. Прославленная русская

а)



б)

Рис. 2. Пример незначительного изменения текста
а) В проверяемой работе указана р. Ока
б) В оригинале указана р. Волга

Таблица 4. Частота использования подходов к сокрытию фактов плагиата в работах базы «Conference»

Подходы к сокрытию плагиата	Доля в 2009 г., %	Доля в 2010 г., %	Доля в 2011 г., %
Корректировка родов, чисел и времён слов	28,6%	33,3%	33%
Незначительное изменение текста	72%	70,3%	89%
Сокращение заимствованного текста	28,6%	43,7%	37,3%
Перестановка частей текста	0%	3%	2,2%
Замена букв	7%	25,9%	6,6%
Замена знаков препинания	0%	1,5%	1,1%
Замена пробелов на невидимые буквы	0%	0,7%	0%
Синонимизация текста	0%	1,5%	1,1%

Из работ, скопированных из одного источника, около 40% подвергались сокращению. В данном случае под сокращением подразумевалось исключение части предложений, графиков, рисунков из заимствованных текстов, а также исключение начальных или конечных блоков текста, по смыслу составляющих единое целое с заимствованным фрагментом. Копирование законченного фрагмента из текста (например, раздела или главы) сокращением не считалось.

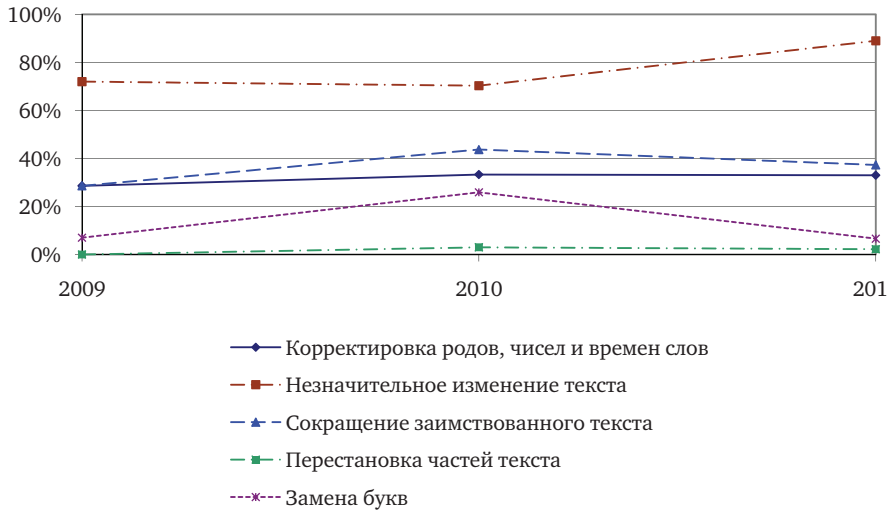


Рис. 3. Доля наиболее используемых подходов к сокрытию фактов плагиата в базе «Conference»

Замена букв активно использовалась в работах 2010 года и была обнаружена при проверке. Вследствие этого, в работах 2011 года доля замены букв снизилась до 6,6%. Чаще всего замене подверглись русские буквы «о», «е», «а», сходные по написанию с английскими буквами.

Сегодня воздействие человека на окружающую среду приняло угрожающие масштабы. Поэтому охрана природы – задача нашего века, проблема, ставшая социальной. Наше здоровье, благополучие и развитие человеческой цивилизации в целом зависят только от нас и наших действий Мы считаем, что достижение стабильной экологической обстановки станет возможным тогда, когда люди осознают, «что все взаимосвязано со всем», и природа всегда будет давать нам то, что мы сумели дать ей.

Рис. 4. Результаты проверки орфографии средствами Microsoft Word текста, содержащего замены русских букв «о», «е», «а» на сходные по написанию английские буквы

Замена знаков препинания («.» на «,» и обратно) осуществлялась в 1% работ. Замена пробелов на невидимые буквы написанные белым цветом обнаружена всего в одной работе. Ручная синонимизация проводилась только в 1% работ. Применения автоматической синонимизации в статьях замечено не было. В целом, последние три подхода применяются студентами достаточно редко.

Объем работ в базах «Course» и «Control» существенно больше, чем в базе «Conference». По этой причине обнаружение фактов заимствований в них — задача более сложная и трудоёмкая. Например, система «Антиплагиат» ограничивает объём проверяемого текста 3000 или 5000 символами (доступно после регистрации). В наше рассмотрение попали работы со значительной долей заимствований.

Следует отметить, что при сдаче указанных работ, проверка преподавателями оригинальности не проводилось. Этот факт оказал значительное влияние на содержание работ.

Анализ показал, что среди курсовых работ (база «Course) доля заимствований из сети Интернет достаточно мала — в основном это введение и теоретическая часть. С другой стороны, достаточно сильно работы пересекаются между собой. Другими словами заимствование происходит у своих же одногруппников или из работ старших курсов. Доля работ, скопированных из одного источника, составляет 78% (см. таблицу 5). Для того, чтобы скрыть факт заимствований, в основном применяется изменение визуального представления работы (оформления) — изменение шрифтов, интервалов, разбиение на абзацы и т. д. Иногда части работ переставляются местами. В большинстве работ (89%) производилась незначительная корректировка текстов (см. таблицу 6). Это объясняется в первую очередь «подгонкой» текстов под задание на курсовую работу.

Таблица 5. Количество источников для заимствований в работах баз «Course» и «Control»

Количество источников	**Курсовые работы, %**	**Контрольные работы, %**
Копирование текста из одного источника	78%	58%
Копирование и компоновка текста из нескольких источников	22%	42%

Таблица 6. Частота использования подходов к сокрытию фактов плагиата в базах работ «Course» и «Control»

Подходы к сокрытию плагиата	**Курсовые работы, %**	**Контрольные работы, %**
Корректировка родов, чисел и времен слов	14%	8%
Незначительное изменение текста	89%	76%
Сокращение заимствованного текста	17%	32%
Перестановка частей текста	3%	0,5%
Замена букв	0%	0%
Замена знаков препинания	0%	0%
Замена пробелов на невидимые буквы	0%	0%
Синонимизация текста	0%	0%

В базе контрольных работ «Control» ситуация несколько иная. 42% работ, замеченных в заимствованиях, копировались и компоновались из нескольких источников. Наиболее часто употребляется незначительное изменение текста (76% случаев), сокращение текстов (32%) и корректировка родов, чисел и времен слов (8%).

Выводы

Исследования показали, что для сокрытия фактов плагиата наиболее часто употребляются:

- изменение заимствованного текста,
- сокращение текста,
- корректировка родов, чисел и времен слов,
- замена русских букв на сходные по написанию латинские.

Такие подходы, как перестановка частей текста, замена знаков препинания, замена пробелов на невидимые буквы, ручная и автоматическая синонимизация текста применяются достаточно редко.

При усилении контроля за оригинальностью работ увеличиваются попытки скрыть факт заимствований за счет изменения текстов и увеличения числа используемых источников.

Подходы, используемые для маскировки плагиата, со временем могут меняться в зависимости от их эффективности против систем проверки на плагиат.

Литература

1. *Плагиат* — Википедия [Электронный ресурс]. — Режим доступа: http://ru.wikipedia.org/wiki/Плагиат
2. *Антиплагиат* [Электронный ресурс]. — Режим доступа: http://www.antiplagiat.ru/
3. *Шарапов Р. В.* Анализ подходов к обнаружению заимствованных текстов // Журнал «Современные наукоемкие технологии» — М: Российская академия естествознания, 2011 г. № 3, С. 47–49
4. *Шарапов Р. В., Шарапова Е. В.* Система проверки текстов на заимствования из других источников // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: XIII Всероссийская научная конференция «RCDL'2011». Воронеж, 19–22 октября 2011 г.: труды конференции — Воронеж: Издательско-полиграфический центр Воронежского государственного университета, 2011, с. 233–238

References

1. *Plagiarism* — Wikipedia, available at: http://ru.wikipedia.org/wiki/%D0%9F%D0%BB%D0%B0%D0%B3%D0%B8%D0%B0%D1%82
2. *Antiplagiat*, available at: http://www.antiplagiat.ru/
3. *Sharapov R. V.* (2011), Analysis of the approaches to the detection of plagiarism [Analiz podkhodov k obnaruzheniiu zaimstvovannikh tekstov], Sovremennye naukoemkie tekhnologii [Modern high technologies], no.3, pp. 47–49.
4. *Sharapov R. V., Sharapova E. V.* System of duplicate texts detection [Sistema proverki tekstov na zaimstvovaniya iz drugikh istochnikov]. // Elektronnye biblioteki: perspektivnie metody i tekhnologii, elektronnye kollektsii: XIII vserossiiskaia nauchnaia konferentsiia "RCDL'2011": Trudy Konferentsii [Digital Libraries: Advanced Methods and Technologies "RCDL'2011": Proceedings of the 13th Conference]. Voronezh, 2011, pp. 233–238.