ISSN 2307–3489 (Print), ISSN 2307–6666 (Online)

Наука та прогрес транспорту. Вісник Дніпропетровського національного університету залізничного транспорту, 2017, № 1 (67)

ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНІ ТЕХНОЛОГІЇ ТА МАТЕМАТИЧНЕ МОДЕЛЮВАННЯ

УДК 004.42:378.014.6

В. И. ШИНКАРЕНКО[1*], Е. С. КУРОПЯТНИК[2*]

[1*]Каф. «Компьютерные информационные технологии», Днепропетровский национальный университет железнодорожного транспорта имени академика В. Лазаряна, ул. Лазаряна, 2, Днипро, Украина, 49010, тел. +38 (056) 373 15 35, эл. почта shinkarenko_vi@ua.fm, ORCID 0000-0001-8738-7225
[2]Каф. «Компьютерные информационные технологии», Днепропетровский национальный университет железнодорожного транспорта имени академика В. Лазаряна, ул. Лазаряна, 2, Днипро, Украина, 49010, тел. +38 (056) 373 15 35, эл. почта elenadiit@rambler.ru, ORCID 0000-0003-2286-884X

ПРОБЛЕМЫ ВЫЯВЛЕНИЯ ПЛАГИАТА И АНАЛИЗ ИНСТРУМЕНТАЛЬНОГО ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ ДЛЯ ИХ РЕШЕНИЯ

Цель. Данное исследование направлено на: 1) определение понятия «плагиата» в текстах на формальных и естественных языках, построение таксономии плагиата; 2) выявление основных проблем обнаружения плагиата при использовании автоматизированных средств их решения; 3) анализ и систематизацию информации, полученной в ходе обзора, тестирования и анализа работы существующих систем обнаружения заимствований. **Методика.** Для выявления требований к программному обеспечению по обнаружению плагиата применяются методы анализа нормативной документации (законодательной базы) и конкурентного инструментария. Для проверки требований используются методы тестирования и обзора интерфейсов GUI. **Результаты.** В работе рассмотрено понятие «плагиата», вопросы его распространения и классификации. Выполнен обзор существующих систем выявления плагиата: настольных приложений и онлайн-ресурсов. Выделены их функциональных характеристики, определены форматы входных и выходных данных и ограничения на них, особенности настройки и доступа. Выполнена детализация требований к рассмотренным системам. **Научная новизна.** Авторами предложено дополнение к существующим иерархическим схемам таксономии плагиата. Выполнен анализ существующих систем с точки зрения функциональности и возможности использования для больших объемов данных. **Практическая значимость.** Практическая значимость определяется широтой проблемы плагиата в различных сферах. В Украине развивается законодательная база для борьбы с плагиатом, что требует активного решения задач разработки, совершенствования и внедрения соответствующего программного обеспечения (ПО). Данная работа способствует решению указанных задач. Обзор существующих программ-антиплагиатов, а также изучение и исследование опыта в этой области, уточнение понятия «плагиата», стратегии его выявления позволяет более полно сформулировать требования к функциональным характеристикам, входным и выходным данным разрабатываемого ПО, а также выявить особенности работы подобного ПО. В статье сделан акцент на особенности решения задачи выявления заимствований в академической среде.

Ключевые слова: плагиат; таксономия плагиата; заимствование фрагментов; системы обнаружения плагиата

doi 10.15802/stp2017/94034

ISSN 2307–3489 (Print), ISSN 2307–6666 (Online)

Наука та прогрес транспорту. Вісник Дніпропетровського національного університету залізничного транспорту, 2017, № 1 (67)

Введение

Интенсивность развития всех отраслей общественного производства, а также средств их информационной поддержки приводят к резкому росту объемов информации, в том числе представленной в текстовом виде. Одной из задач обработки текстов является их синтаксическое и семантическое сравнение с целью выявления заимствований.

Решению данной проблемы посвящено множество работ в правой и академической отраслях [4, 20, 24], а также в сфере информационных технологий [3, 15, 22].

Проблема выявления плагиата усложнена множеством вариантов определения понятия плагиата в разных контекстах. «Плагиат – акт взятия рукописей другого человека и выдачи их как свои собственные. Мошенничество тесно связанное с подделкой и пиратством на практике, как правило, в нарушение закона об авторских правах» [14]. Согласно Закону Украины «Про авторське право та суміжні права» редакции от 13.01.16 «плагіат – оприлюднення (опублікування), повністю або частково, чужого твору під іменем особи, яка не є автором цього твору…». На сегодня существует ряд объектов, охраняемых авторским правом: литературные произведения различного жанра, выступления, лекции, произведения искусства, производные произведения, а также «другие произведения» [11, 13], определены неохраняемые объекты [9]. Степень производности и специфика произведения требует особого внимания и подходов для решения задачи определении плагиата. В настоящее время получает все большую практику решение этой задачи с помощью IT-технологий, сфокусированных на выявление прежде всего текстовых заимствований, которые имеют широкое распространение в сфере науки, образовании, профессиональной деятельности, особенно в СМИ [10]. В таких случаях применяется термин «плагиат». В данной работе понятие «плагиат» используется в более широких семантических пределах, чем рамки нормативно-законодательной базы, что является традиционным для сферы разработки программного обеспечения. Под плагиатом будем понимать наличие в текстовых и иных документах фрагментов, заимствованных с различных источников без указания их автора и/или с нарушениями правил цитирования.

Существуют разные подходы для его классификации: по техническим средствам маскировки, по объему, степени маскировки [21, 23], сфере использования [10].

Одной из актуальных проблем является устранение последствий маскировки плагиата. В связи с этим выделяют такие типы плагиата [23]: дословный; скрытый плагиат с помощью перефразирования; скрытый плагиат с помощью технических трюков, использующих недостатки существующих систем антиплагиата, умышленное неточное использование ссылок; «жесткий плагиат» – тип плагиата, который особенно тяжело выявлять.

Для студенческих работ характерны такие виды плагиата [14]:

– текстуальные плагиаты: этот тип плагиата обычно делается студентами или исследователями в научных учреждениях, где документы являются идентичными или типичными для исходных документов, докладов, эссе научных работ и дизайнерского искусства;

– плагиат исходного кода компьютерных программ: также используется студентами в университетах, где студенты пытаются сдать копию полного или частей исходного кода, написанного кем-то другим, как свой собственный.

Таким образом, определение понятия плагиата неоднозначно, имеет много формулировок и включает множество различных аспектов.

Цель

Основной целью данной работы является построение таксономии видов плагиата, выявление основных проблем в задачах обнаружения плагиата и использовании автоматизированных средств для их решения; а также анализ и систематизирование информации, полученной в ходе обзора, тестирования и анализа работы существующих систем обнаружения заимствований.

Методика

Для выявления требований к программному обеспечению по выявлению плагиата применяются методы анализа нормативной документации (законодательной базы) и конкурентных

doi 10.15802/stp2017/94034

ISSN 2307–3489 (Print), ISSN 2307–6666 (Online)

Наука та прогрес транспорту. Вісник Дніпропетровського національного університету залізничного транспорту, 2017, № 1 (67)

продуктов, а также метод анализа и черного ящика. Для проверки требований используются методы тестирования и обзора GUI.

Результаты

В работе рассмотрено понятие плагиата, вопросы его распространения и классификации. Выполнен обзор существующих систем выявления плагиата: настольных приложений и онлайн-ресурсов. Выделены их функциональные характеристики, определены форматы входных и выходных данных и ограничения на них, особенности настройки. Выполнена детализация требований к рассмотренным системам.

Таксономия плагиата. Таксономия плагиата предполагают выделения различных его уровней по типу (виду материала, рис. 1), сложности и путям его реализации. Каждый тип работ (студенческая, исследовательская) может содержать заимствования такие, как цитаты первоисточника, ссылки на результаты экспериментов и апробаций. Правильное оформление подобных вставок является нормальной практикой в научно-образовательной сфере, пренебрежение ними – плагиатом.

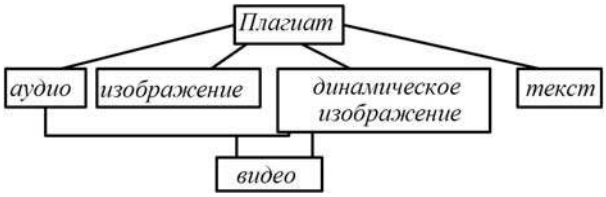

Рис. 1. Формы представления материалов, подвергаемых заимствованию

Fig. 1. The forms of presentation materials subjected to borrow

Умышленное неточное использование ссылок является одним из способов, применяемых при выполнении учебных заданий. Студенты могут использовать неправильные и неточные цитаты, проявляя неспособность определить цитируемый текст с необходимой точностью. Методы такого типа плагиата включают в себя [23]:

– обеспечение поддельной ссылки, то есть выдуманной ссылки, которой не существуют, и, следовательно, невозможно цитировать и текст ссылки точно;

– предоставление ложных ссылок: ссылка существует, но материал по ней не соответствует приведенному в работе;

– использование «забытых» или аннулированных ссылок на источники: добавление цитат или скобок, но непредставление информации о ссылке на источники.

Плагиат может быть полным и частичным в зависимости от процента заимствованных фрагментов. Классификация плагиата представлена в табл. 1.

Таблица 1

Характеристика плагиата

Table 1

Characteristics of plagiarism

Признак	Значение	
Объем	Полный	
	Частичный	
Количество источников	Один	Простой плагиат
	Много	Сложный плагиат
Структурный источник	Обзор	
	Постановка задачи	
	Основная часть	
	Примеры	
	Выводы (результаты)	
Непрерывность	Сплошной	
	Фрагментарный	
Степень важности	Насколько заимствованный фрагмент важен для данного документа (текста)	
Наличие изменений	Отсутствуют	
	Использованы «маскировочные трюки»	
	Перевод на другой язык	
	Перестановка фраз и/или других фрагментов	

Полный плагиат может быть классифицирован как простой и сложный. К первому типу можно отнести получение псевдо оригинального текста на основе одного документа: манипуляции с таким документом минимальны и не требуют сложного интеллектуального труда, и отчасти могут быть реализованы по средствам онлайн-сервисов или компьютерных про-

doi 10.15802/stp2017/94034

ISSN 2307–3489 (Print), ISSN 2307–6666 (Online)
Наука та прогрес транспорту. Вісник Дніпропетровського національного університету залізничного транспорту, 2017, № 1 (67)

грамм. Сложный плагиат, характеризующийся наличием нескольких источников, предполагает более сложную работу, связанную не только с поиском материалов, но и требует понимание предметной области исходных текстов.

Маскирование плагиата, или так званый латентный плагиат, является актуальным вопросом не только в сфере образования, но и информационных технологий, так как ряд программ, направленных на выявления заимствований, не имеют стопроцентной защиты от данной проблемы. Параллельно с программами-антиплагиатами разрабатываются антиплагиаты-киллеры, направленные на сокрытия заимствований в автоматическом режиме.

К приемам маскировки можно отнести:

– использование символов с разной кодировкой [23]: замена кириллических символов похожими по написанию латинскими символами;

– вставка в текст непечатных символов, в том числе добавление последовательностей из двух и более пробелов;

– допущения орфографических ошибок с определенной вероятностью;

– изменение регистра (изменение больших букв на малые и наоборот);

– добавление пустых абзацев и замена символа абзаца на символ разрыва строки;

– замена сокращений единиц измерения на их полные названия и наоборот;

– замена цифр их наименованием прописью.

При оценке текста на плагиат можно выделить такие задачи (табл. 2): определение типа документа по языку; определение уровня, на котором будет вестись поиск заимствований; определения лексических конструкций, на уровне которых будет вестись поиск заимствований; проверка на уникальность; анализ результатов.

Текст может быть написан на естественном языке или формализованном, а также содержать фрагменты обоих типов (табл. 3). Проверка текстов на естественном языке предполагает учет таких его особенностей:

– нестрогий порядок слов в предложении;

– наличие многозначных слов, синонимов, омонимов;

– изменения порядка слов может приводить к изменению смысла высказываний;

– эволюция языка.

Таблица 2

Этапы обнаружения плагиата

Table 2

Stages of plagiarism detection

Этап	Основные характеристики
Определение типа документа	ЕЯ, формальный язык, смесь (гибрид)
Определение уровня	Синтаксический, семантический, гибридный
Определение масштаба	Слова, словосочетания, предложения (фразы, абзацы)
Проверка на уникальность	
Анализ результатов	Объем, цитирование, пересечение фрагментов, структурный источник, важность фрагмента

Таблица 3

Типы документов

Table 3

Document Types

Документы	Примеры
Естественноязыковые	Издания СМИ Худ. лит-ра
Формальноязыковые	Мат. выкладки Программы UML-модели прочее
Смешанные документы	Техническая литература Пособия, монографии Диссертации Чертежи и конструкторская документация Учебные работы Документация к ПО

Общими проблемами являются:

– определение «границ» идиоматических единиц;

– изменения знаков препинания может приводить к изменению смысла фразы («Казнить

doi 10.15802/stp2017/94034

нельзя помиловать»).

Ряд языков, подобных uml [25], имеют графическую интерпретацию в виде геометрических фигур; распознания символов в формулах также является проблемой (частично она обусловлена различиями форматов редакторов формул).

Анализ результатов может быть выполнен по нескольким критериям (табл. 2), исходя из характеристик плагиата (табл. 1). По объему плагиат может быть единим, целым фрагментом, а может быть выборочными (отдельными, разрозненными) частями документа. По принадлежности текст может быть: свой (в том числе самоцитирование), общеизвестные вещи (например, народное творчество, правила языка), чужой.

Цитирование может быть сторонних источников и авторское (самоцитирование).

По принадлежности к структурному источнику фрагменты могут быть справочными или теоретическими сведениями (например, в лабораторных работах студентов), основным текстом, обзором аналогов, литературы, фрагментами, которые дополняют картину или вносят ясность в дальнейший текст. Последние могут быть допустимыми лишь в отдельных разделах документа (например, в докторской диссертации). По важности заимствованные фрагменты могут передавать основные мысли, а могут вспомогательные элементы работы (периферию): примеры решения задач, примеры начальных условий и т.п. По количеству источников фрагменты могут быть моно- и полизаимствованными.

Системы обнаружения плагиата. На сегодня существует ряд программ (настольных приложений и онлайн-сервисов), позволяющих выявлять заимствования текстов на естественном языке и на языках программирования. Среди них общего назначения: Etxt Анти плагіат, Advego Plagiatus, Double Content Finder (DCFinder), Praide Unique Content Analyser 2, Copyscape, istio.com и другие. А также специализированные – для использования в вузах: Anti-Plagiarism [6], пакет «Антиплагиат. ВУЗ», «Plagiarism» [12], strikeplagiarism.com, unplag.com. Описанию и сравнению различных систем антиплагиата посвящено ряд работ [5, 15, 19].

Рассмотрено 27 ресурсов по обнаружению плагиата: онлайн и настольные приложения с различными типами баз данных исходных текстов (рис. 2). Далее приведен перечень ресурсов:

1. eTXT Антиплагиат [1, 3, 5]
2. Advego Plagiatus [3, 5]
3. Double Content Finder [3, 5, 18]
4. Praide Unique Content Analyser II [5]
5. Viper [5]
6. Плагиата.НЕТ [5]
7. Duplichecker [5]
8. PaperRater [5]
9. Anti-Plagiarism [6]
10. strikeplagiarism.com [33]
11. Plagiarisma.Net [5, 26]
12. PlagiarismChecker [5, 27]
13. Plagium [5, 28]
14. PlagTracker [5, 31]
15. SeeSources [5]
16. PlagScan [5, 30]
17. Plagiarism Detector [5, 29]
18. Защита уникальности контента [5]
19. FindCopy [3, 5]
20. Docol©c [5, 17]
21. Grammarly [5]
22. Text.ru [5, 7]
23. Антиплагиат ру [3, 5, 8]
24. Copyscape [10, 16]
25. Miratools [3]
26. smallSeoTools.com/plagiarismChecker [32]
27. unplag.com [34].

В результате анализа [3, 5–8, 10, 16, 17, 26–34] были сформулированы требования к входным и выходным данным программ-антиплагиатов, а также их функциональным характеристикам. Непосредственная работа с ресурсом начинается с подачи документа на проверку, которая может быть осуществлена такими способами:

– url сайта, контент которого необходимо проверить (ресурсы 1–4, 7, 11, 12, 13, 16, 24, 25, 26);

– отдельный файл (1–4, 7, 28, 11, 13, 16, 17, 20, 27);

– пакет файлов (1, 25), несколько документов одним zip-архивом (16);

– проверяемый текст вводится в специальную экранную форму (1, 3, 4, 7, 8, 11–14, 17, 22, 26, 27).

doi 10.15802/stp2017/94034

ISSN 2307–3489 (Print), ISSN 2307–6666 (Online)
Наука та прогрес транспорту. Вісник Дніпропетровського національного університету залізничного транспорту, 2017, № 1 (67)

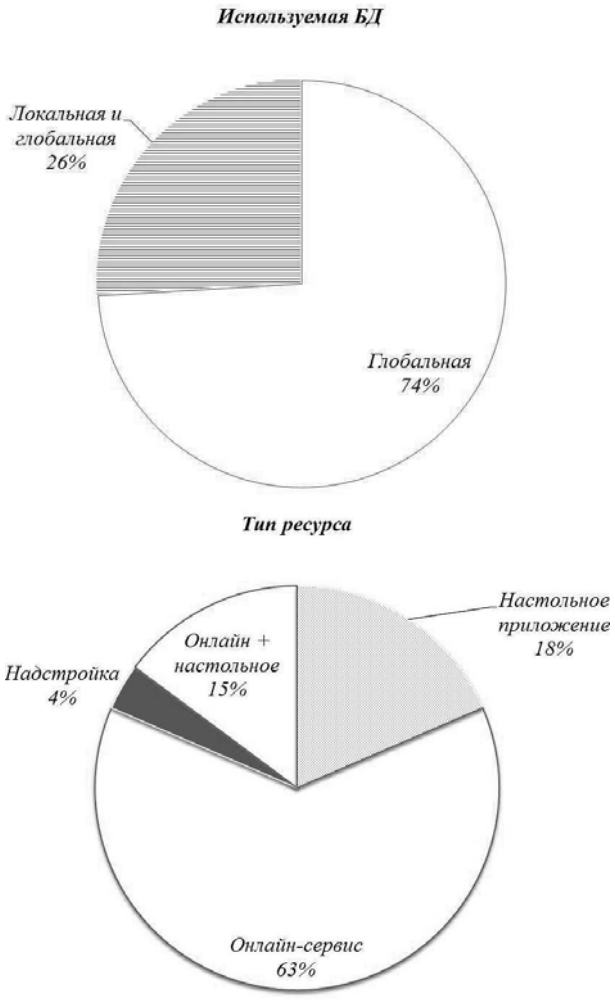

Рис. 2. Анализ ресурсов

Fig. 2. Resource analysis

Результаты работы по обнаружению плагиата представляются следующим образом:

– экранная форма с выделением заимствованных фрагментов (1, 2, 16, 17, 22–25);

– сохранение отчета в pdf-формате (16, 26, 27), в doc (1, 16), rtf (5);

– сохранение отчета о проверке в личном аккаунте (20, 27);

– отправка отчета о проверке по электронной почте (20).

По результатам работы система-антиплагиат формирует отчет, в котором могут быть представлены такие данные:

– список источников заимствования (1–5, 8, 11, 14, 16, 22, 23, 26);

– общий процент плагиата в работе (14, 16, 23, 24, 26);

– указания на фрагменты, где необходимо оформить текст как цитату и поставить ссылку на источник (14);

– процент уникальности (1, 3, 5, 8, 11, 22, 25, 26);

– проверяемый текст (26);

– сообщение о наличие плагиата (8);

– объем введенного текста (11, 22, 24);

– количество совпадений на источник (16).

Основными функциональными характеристиками являются:

– сравнение текстов с базой (выполняют все);

– сравнение текстов один к одному (тексты предоставляет пользователь) (13, 24, 27) или по url (7, 24);

– дословное сравнение и определение смысловых совпадений (1);

– формирование отчета о плагиате/уникальности (все с разной степенью детализации);

– замена латинских символов на кириллические (если есть такого вида маскировка) (6);

– предоставления сведений о плагиате с какого-либо сайта еженедельно или ежемесячно (7);

– проверка грамматики, правописания или стилистики встроенным литературным редактором (8);

– оповещение пользователя (по электронной почте) о плагиате его текстов (12);

– создание аккаунтов научных учреждений, институтов (13);

– настройка параметров проверки (1, 4, 25, 11);

– защита от проверки пустых документов (9);

– поиск по базам Google Scholar, Google Books (11).

Также предоставляется возможность выполнить настройку параметров проверки работы (сравнения):

– задание сайтов, с которыми не выполнять сравнение (25);

– задание параметра «размера выборки» (1, 27);

– наличие разных уровней проверки: обычная, экспресс, глубокая (1).

– установка «порога уникальности» (1);

– выбор поисковой машины (4, 11).

Рассмотрим возможности систем, имеющих внедрения [6, 8, 33].

doi 10.15802/stp2017/94034

ISSN 2307–3489 (Print), ISSN 2307–6666 (Online)

Наука та прогрес транспорту. Вісник Дніпропетровського національного університету залізничного транспорту, 2017, № 1 (67)

Система Anti-Plagiarism (Украина) [6] с 2008 разрабатывалась как открытая система, а с 2010 внедрена в Хмельницком национальном университете. Особенностями являются:

– проверка текстов на ЕЯ и ЯП;

– поддержка локальной и глобальной базы;

– проверка грамматических ошибок;

– защита от добавления символов, добавления или изменения пунктуационных знаков, изменения литерального состава;

– защита от проверки пустых документов.

«Антиплагиат. ВУЗ» является частью проекта компании Forecsys (Россия), предоставляющей онлайн сервис antiplagiat.ru. Особенностью этого пакета является:

– текстовая база продукта, включающая материалы, которые были собраны в рунете, а также дополнительные источники, включая методическую литературу и студенческие работы прошлых лет. Таким образом, ВУЗ имеет возможность вести собственную базу оригинальных работ, которая будет постоянно наращиваться;

– при использовании продукта преподавателем возможно оценивания работы, которая была проверена;

– работа в рамках организационно-штатной структуры учебного заведения, а также наличие различных типов пользователей с различными правами (преподаватели, менеджеры кафедр, администраторы);

– возможность учета результатов работы в статистических данных заведения [1];

– наличие локальной и Интернет-версии дает большие возможности для интеграции продукта в существующие в учебном заведении информационные системы.

К недостаткам можно отнести такие особенности продукта:

– студенты не являются пользователями данного пакета ([1]), информация о них используется только для персонализации работ;

– вся работа по загрузке работ (и пополнения баз), проверки их ложится на преподавателей, для пополнения базы используется специальный тип пользователей;

– не предусмотрено ведение статистики по отдельной дисциплине.

На сегодня пользователями данного продукта является ВАК России, Государственный университет – Высшая школа экономики (ГУ-ВШЭ), Московский институт экономики, менеджмента и права (МИЭМП), Московский городской психолого-педагогический университет (ММПП), Нижегородский государственный университет (ННГУ) и др.

Strikeplagiarism.com – система антиплагиата студенческих работ, разработанная в Польше. Имеет ряд таких особенностей [33]:

– интегрируется с Системами управления обучением (LMS) и подключается к Moodle;

– предоставляет подробный отчет подобия со списком источников заимствований;

– имеет функцию «Тревожного сигнала» для обозначения попыток помешать анализу на наличие плагиата;

– выполняет сравнение с ресурсами глобальной сети Интернет и базой данных учреждения или ВУЗа;

– выявляет сходства с помощью алгоритмов, основанных на анализе N-грамм, лингвистический анализ; постоянно совершенствуется поисковый алгоритм и предусмотрена адаптация алгоритмов обнаружения к специфике проверяемых документов;

– выполняет анализ на многих языках;

– предоставляет возможность добавлять документы в базу данных ВУЗа и присоединиться к Межвузовской программе обмена базами данных.

Рассмотренные ресурсы имеют ряд ограничений, которые являются препятствием для массового внедрения, например, в вузах, а именно:

– отсутствие поддержки или стадия тестирования, или проблемы с доступом (не найден/не отвечает);

– ограничения по объему проверяемых текстов и частоте запросов;

– длительная проверка документа;

– для полноценной работы необходим премиум-аккаунт;

– отсутствие подробного отчета;

– отсутствия поддержки русского и/или украинского языков.

Для технических ВУЗов актуальны также следующее проблемы:

– большинство ресурсов работают только с естественноязыковыми текстами;

– обработка только текстов (нет обработки изображений, формул и прочего).

doi 10.15802/stp2017/94034

Научная новизна и практическая значимость

Предложено дополнение существующих иерархических схем таксономии плагиата. Выполнен анализ существующих систем с точки зрения функциональности и возможности использования для больших объемов данных. Практическая значимость определяется широтой проблемы плагиата в различных сферах. В Украине развивается законодательная база для борьбы с плагиатом, что требует активного решения задач разработки, усовершенствования и внедрения соответствующего программного обеспечения (ПО). Данная работа способствует решению указанных задач. Обзор существующих программ-антиплагиатов, а также изучение и исследование опыта в этой области и уточнение понятия плагиата, стратегии его выявления позволяет более полно сформулировать требования к функциональным характеристикам, входным и выходным данным разрабатываемого ПО, а также выявить особенности работы подобного ПО. В статье сделан акцент на особенности решения задачи выявления заимствований в среде высших технических учебных заведений.

Выводы

Определение понятие плагиата является одним из основных этапов в формировании требований к системам обнаружения плагиата. Его определение, а также классификация видов плагиата является важной составляющей при формировании требований к функциональным характеристикам соответствующего программного обеспечения (ПО). Обзор и тестирование конкурентных продуктов позволили выделить основные требования к ПО, а также ряд дополнительных, которые являются полезными в зависимости от области применения ПО.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Антиплагиат. ВУЗ [Электронный ресурс] : рук-во пользователя. – Режим доступа: http://antiplagiat.nsu.ru/files/manual.pdf. – Загл. с экрана. – Проверено : 08.12.2016.
2. Антиплагиат eTxt.ru. Проверка уникальности текстов [Электронный ресурс] – Режим доступа: https://www.etxt.ru/antiplagiat/. – Загл. с экрана. – Проверено : 17.10.2016.
3. Болілий, В. О. Перевірка унікальності тексту при оцінюванні студентських робіт творчого або дослідницького характеру / В. О. Болілий, В. В. Копотій // Наук. записки НДУ ім. М. Гоголя. Серія: Психолого-педагогічні науки : зб. наук. пр. / Ніжин. держ. ун-т ім. М. Гоголя. – Ніжин, 2011. – № 7. – С. 134–145.
4. Голунов, С. В. Студенческий плагиат как вызов системе высшего образования в России и за рубежом / С. В. Голунов // Вопросы образования. – 2010. – № 3. – С. 243–257.
5. Лупаренко, Л. А. Інструментарій виявлення плагіату в наукових роботах: аналіз програмних рішень // Інформ. технології і засоби навчання. – 2014. – Т. 40, № 2. – С. 151–169.
6. Михайловский, Ю. Б. Система Anti-Plagiarism як інструмент запобігання плагіату в навчальній та науковій діяльності / Ю. Б. Михайловский, Н. А. Длугунович. – Вісн. Хмельн. нац. ун-ту. Технічні науки. – 2013. – № 3. – С. 162–168.
7. Онлайн сервис проверки текста на уникальность [Электронный ресурс] – Режим доступа: http://text.ru/. – Загл. с экрана. – Проверено : 17.10.2016.
8. Офіційна сторінка компанії Форексіс. Розділ «Продукти» [Электронный ресурс] – Режим доступа: http://www.forecsys.ru/ru/site/products/antiplagiat/. – Загл. с экрана. – Проверено : 15.11.2016.
9. Про авторське право та суміжні права [Электронный ресурс] : Закон України. – Режим доступа: http://zakon2.rada.gov.ua/laws/show/3792-12. – Загл. с экрана. – Проверено : 30.10.2016.
10. Романова, І. В. Явище плагіату: історія та сьогодення / І. В. Романова // Зовнішня торгівля: право, економіка, фінанси. – 2012. – № 3. – С. 267–272.
11. Цивільний кодекс України [Электронный ресурс] – Режим доступа: http://zakon0.rada.gov.ua/laws/show/435-15. – Загл. с экрана. – Проверено : 17.01.2017.
12. Шостак, І. В. Комп'ютеризація процесу виявлення плагіату у студентських роботах / І. В. Шостак, І. В. Груздо // Зб. наук. пр. військ. ін-ту Київ. нац. ун-ту ім. Тараса Шевченка. – Київ, 2013. – Вип. 41. – С. 99–109.
13. Штефан, О. Плагіат: поняття, ознаки, відповідальність / О. Штефан // Теорія і практика інтелектуальної власності. – 2011. – № 6. – С. 17–25.

doi 10.15802/stp2017/94034

14. Asim, M. El Tahir Ali. Overview and Comparison of Plagiarism Detection Tools / M. El Tahir Ali Asim, Hussam M. Dahwa Abdulla, V. Snasel. – Dateso, 2011. – P. 161–172.
15. Bin-Habtoor, A. S. A Survey on Plagiarism Detection Systems / A. S. Bin-Habtoor, M. A. Zaher // Intern. J. of Computer Theory and Engineering. – 2012. – Vol. 4, No. 2. – P. 185–188. doi: 10.7763/IJCTE.2012.V4.447.
16. Copyscape. Compare Articles or Web Pages [Электронный ресурс] – Режим доступа: http://www.copyscape.com/compare.php. – Загл. с экрана. – Проверено : 18.10.2016.
17. Docol©c [Электронный ресурс] – Режим доступа: http://docoloc.de/. – Загл. с экрана. – Проверено : 18.10.2016.
18. Double Content Finder [Электронный ресурс] – Режим доступа: https://textbroker.ru/main/dcfinder.html. – Загл. с экрана. – Проверено : 17.10.2016.
19. Culwin, F. A review of electronic services for plagiarism detection in student submissions / F. Culwin, Th. Lancaster // 8th Annual Conf. on the Teaching of Computing – Edinburgh, 2010. – P. 54–61.
20. Jarrah, A. Al. Plagiarism Detection based on studying correlation between Author, Title and Content / A. Al. Jarrah, I. Alsmadi, Z. Za'atreh // Intern. Conf. on Information Communication System (CICS) (22.05.2011). – Shanghai, 2011. – P. 22–24.
21. Kakkonen, T. AntiPlag – A Sampling-based Tool for Plagiarism Detection in Student Texts / T. Kakkonen, N. Myller // The Proc. of the 8th European Conference on E-learning (29.10–30.10.2009). – Bari, Italy, 2009. – P. 287–293.
22. Meuschke, N. State-of-the-art in detecting academic plagiarism / N. Meuschke, B. Gipp // Intern. J. for Educational Integrity. – 2013. – Vol. 9, No. 1. – P. 50–71.
23. Mozgovoy, M. Automatic Student Plagiarism Detection: Future Perspectives / M. Mozgovoy, T. Kakkonen, G. Cosma // J. of Educational Computing Research. – 2010. – Vol. 43. – Iss. 4. – P. 511–531. doi: 10.2190/ec.43.4.e.
24. Mozgovoy, M. Desktop Tools for Offline Plagiarism Detection in Computer Programs / M. Mozgovoy // Informatics in Education. – 2006. – Vol. 5, No. 1. – P. 97–112.
25. OMG Unified Modeling Language (OMG UML), Infrastructure : Version 2.4.1. – 2011. – 748 p.
26. Plagiarisma [Электронный ресурс] – Режим доступа: http://plagiarisma.net/. – Загл. с экрана. – Проверено : 17.10.2016.
27. Plagiarism-Checker.com [Электронный ресурс] – Режим доступа: http://www.plagiarism-checker.com/. – Загл. с экрана. – Проверено : 18.10.2016.
28. Plagiarism checker & plagiarism detection [Электронный ресурс] – Режим доступа: http://www.plagium.com/. – Загл. с экрана. – Проверено : 18.10.2016.
29. Plagiarism Detector [Электронный ресурс] – Режим доступа: http://plagiarismdetector.net/. – Загл. с экрана. – Проверено : 18.10.2016.
30. PlagScan [Электронный ресурс] – Режим доступа: http://www.plagscan.com/. – Загл. с экрана. – Проверено : 18.10.2016.
31. Plagtracker. The most accurate plagiarism checking service [Электронный ресурс] – Режим доступа: http://www.plagtracker.com/. – Загл. с экрана. – Проверено : 18.10.2016.
32. Small SEO Tools. Plagiarism Checker [Электронный ресурс] – Режим доступа: http://smallseotools.com/plagiarism-checker/. – Загл. с экрана. – Проверено : 17.10.2016.
33. StrikePlagiarism.com [Электронный ресурс] – Режим доступа: http://strikeplagiarism.com/ua/. – Загл. с экрана. – Проверено : 17.10.2016.
34. Unplag.com. Plagiarism Detection Engine [Электронный ресурс] – Режим доступа: https://ua.unplag.com/. – Загл. с экрана. – Проверено : 08.12.2016.

В. І. ШИНКАРЕНКО[1*], О. С. КУРОП'ЯТНИК[2*]

[1*]Каф. «Комп'ютерні інформаційні технології», Дніпропетровський національний університет залізничного транспорту імені академіка В. Лазаряна, вул. Лазаряна, 2, Дніпро, Україна, 49010, тел. +38 (056) 373 15 35, ел. пошта shinkarenko_vi@ua.fm, ORCID 0000-0001-8738-7225

[2*]Каф. «Комп'ютерні інформаційні технології», Дніпропетровський національний університет залізничного транспорту імені академіка В. Лазаряна, вул. Лазаряна, 2, Дніпро, Україна, 49010, тел. +38 (056) 373 15 35, ел. пошта elenadiit@rambler.ru, ORCID 0000-0003-2286-884X

doi 10.15802/stp2017/94034

ISSN 2307–3489 (Print), ISSN 2307–6666 (Online)

Наука та прогрес транспорту. Вісник Дніпропетровського національного університету залізничного транспорту, 2017, № 1 (67)

ПРОБЛЕМИ ВИЯВЛЕННЯ ПЛАГІАТУ ТА АНАЛІЗ ІНСТРУМЕНТАЛЬНОГО ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ ДЛЯ ЇХ ВИРІШЕННЯ

Мета. Дане дослідження спрямоване на: 1) визначення поняття «плагіату» в текстах на формальних і природних мовах, побудова таксономії плагіату; 2) встановлення основних проблем виявлення плагіату і використання автоматизованих засобів їх вирішення; 3) аналіз та систематизацію інформації, отриманої у ході огляду, тестування і аналізу роботи існуючих систем виявлення запозичень. **Методика.** Для формулювання вимог до програмного забезпечення з виявлення плагіату застосовуються методи аналізу нормативної документації (законодавчої бази) і конкурентного інструментарію. Для перевірки вимог використовуються методи тестування та огляду інтерфейсів GUI. **Результати.** У роботі розглянуто поняття «плагіату», питання його поширення та класифікації. Виконано огляд існуючих систем виявлення плагіату: настільних додатків та онлайн-ресурсів. Виділено їх функціональні характеристики, визначені формати вхідних та вихідних даних і обмеження на них, особливості налаштування і доступу. Виконана деталізація вимог до розглянутих систем. **Наукова новизна.** Авторами запропоновано доповнення до існуючих ієрархічних схем таксономії плагіату. Виконано аналіз існуючих систем із точки зору функціональності та можливості використання для великих обсягів даних. **Практична значимість.** Практична значимість визначається широтою проблеми плагіату в різних сферах. В Україні розвивається законодавча база для боротьби з плагіатом, що вимагає активного вирішення завдань розробки, вдосконалення та впровадження відповідного програмного забезпечення (ПЗ). Дана робота сприяє вирішенню зазначених завдань. Огляд існуючих програм-антиплагіатів, а також вивчення і дослідження досвіду в цій галузі, уточнення поняття «плагіату», стратегії його виявлення дозволяє більш повно сформулювати вимоги до функціональних характеристик, вхідних і вихідних даних розроблюваного ПЗ, а також виявити особливості роботи подібного ПЗ. У статті зроблено акцент на особливості вирішення завдання виявлення запозичень в академічному середовищі.

Ключові слова: плагіат; таксономія плагіату; запозичення фрагментів; системи виявлення плагіату

V. I. SHYNKARENKO[1*], O. S. KUROPIATNYK[2*]

[1*]Dep. «Computer and Information Technologies», Dnipropetrovsk National University of Railway Transport named after academician V. Lazaryan, Lazaryan St., 2, Dnipro, Ukraine, 49010, tel. +38 (056) 373 15 35, e-mail shinkarenko_vi@ua.fm, ORCID 0000-0001-8738-7225
[2*]Dep. «Computer and Information Technologies», Dnipropetrovsk National University of Railway Transport named after academician V. Lazaryan, Lazaryan St., 2, Dnipro, Ukraine, 49010, tel. +38 (056) 373 15 37, e-mail elenadiit@rambler.ru, ORCID 0000-0003-2286-884X

PLAGIARISM DETECTION PROBLEMS AND ANALYSIS SOFTWARE TOOLS FOR ITS SOLVE

Purpose. This study is aimed at: 1) the definition of plagiarism in texts on formal and natural languages, building a taxonomy of plagiarism; 2) identify major problems of plagiarism detection when using automated tools to solve them; 3) Analysis and systematization of information obtained during the review, testing and analysis of existing detection systems. **Methodology.** To identify the requirements of the software to detect plagiarism apply methods of analysis of normative documentation (legislative base) and competitive tools. To check the requirements of the testing methods used and GUI interfaces review. **Findings.** The paper considers the concept of plagiarism issues of proliferation and classification. A review of existing systems to identify plagiarism: desktop applications, and online resources. Highlighting their functional characteristics, determine the format of the input and output data and constraints on them, customization features and access. Drill down system requirements is made. **Originality.** The authors proposed schemes complement the existing hierarchical taxonomy of plagiarism. Analysis of existing systems is done in terms of functionality and possibilities for use of large amounts of data. **Practical value.** The practical significance is determined by the breadth of the problem of plagiarism in various fields. In Ukraine, develops the legal framework for the fight against plagiarism, which requires the active solution development tasks, improvement and delivery of relevant software (PO). This work contributes to the solution of these problems. Review of existing programs, Anti-plagiarism, as well as study and research experience in the field and update the concept of plagiarism, the strategy allows it to identify more fully articulate to the functional performance requirements, the input and

doi 10.15802/stp2017/94034

output of the developed software, as well as to identify the features of such software. The article focuses on the features of solving the problem of identification of borrowing in an academic environment.

Keywords: plagiarism; taxonomy of plagiarism; borrowing fragments; plagiarism detection system

REFERENCES

1. *ANTIPLAGIAT.VUZ. Rukovodstvo polzovatelya. Funktsionalnye vozmozhnosti rabochikh mest prepodavatelya, ministratora, menedzhera*. (2011). Moscow: NGU. Retrieved from http://antiplagiat.nsu.ru/files/manual.pdf
2. *eTxt Antiplagiat:Proverka unikalnosti tekstov*. (n.d.) Retrieved from https://www.etxt.ru/antiplagiat/
3. Bolilyi, V. O., & Kopotii, V. V. (2011). Perevirka unikalnosti tekstu pry otsiniuvanni studentskykh robit tvorchoho abo doslidnytskoho kharakteru. *Naukovi zapysky Naukovi zapysky Nizhynskoho derzhavnoho universytetu im. M. Hoholia. Seriia: Psykholoho-pedahohichni nauky*, *7*, 134-145.
4. Golunov, S. V. (2010). Plagiarism in students as challenge to higher education system in Russia and abroad. *Educational Studies, 3*, 243-257.
5. Luparenko, L. A. (2014). Plagiarism detection tools for research works: analysis of software solutions. *Informatsiini tekhnolohii i zasoby navchannia, 40*(2), 151-169.
6. Mikhaylovskiy, Y. B., & Dlugunovych, N. A. (2013). Anti-plagiarism system as a tool for plagiarism preventing in educational and research activities. *Visnyk Khmelnytskoho natsionalnoho universytetu: Tekhnichni nauky*, *3*, 162-168.
7. *TEXT.RU: Onlayn servis proverki teksta na unikalnost*. (n.d.) Retrieved from http://text.ru/
8. *Foreksis:Antiplagiat*. (n.d.) Retrieved from http://www.forecsys.ru/ru/site/products/antiplagiat/
9. Pro avtorske pravo ta sumizhni prava: Zakon Ukrainy 1994, No 1651-19 (2016). Retrieved from http://zakon2.rada.gov.ua/laws/show/3792-12
10. Romanova, I. V. (2012). Yavyshche plahiatu: istoriia ta sohodennia. *Foreign trade: economics, finance, law, 3*(62), 267-272.
11. Tsyvilnyi kodeks Ukrainy 2003, No 1666-19 (2016). Retrieved from http://zakon0.rada.gov.ua/laws/show/435-15
12. Shostak, I. V., & Hruzdo, I. V. (2013). Kompiuteryzatsiia protsesu vyiavlennia plahiatu u studentskykh robotakh. *Zbirnyk naukovykh prats Viiskovoho instytutu Kyivskoho natsionalnoho universytetu imeni Tarasa Shevchenka*, *41*, 99-109.
13. Shtefan, O. (2011). Plagiarism: the concept, attributes, responsibility. *Teoriia i praktyka intelektualnoi vlasnosti*, *6*, 17-25.
14. Ali, Asim M. El Tahir, Abdulla, Hussam M. Dahwa, & Snasel, V. (2011, April 20). Overview and Comparison of Plagiarism Detection Tools. *Proceedings of the Dateso 2011: Annual International Workshop on DAtabases, TExts, Specifications and Objects, Czech Republic,* 161-172. Retrieved from http://ceur-ws.org/Vol-706/poster22.pdf
15. Bin-Habtoor, A. S., & Zaher, M. A. (2012). A Survey on Plagiarism Detection Systems. *International Journal of Computer Theory and Engineering, 4*(2), 185-188. doi: 10.7763/IJCTE.2012.V4.447
16. *Copyscape. Compare Articles or Web Pages*. (n.d.) Retrieved from http://www.copyscape.com/compare.php
17. *Docol©c*. (n.d.) Retrieved from http://docoloc.de/
18. *Textbroker:Double Content Finder*. (n.d.) Retrieved from https://textbroker.ru/main/dcfinder.html
19. Culwin, F., & Lancaster, T. (2000). A review of electronic services for plagiarism detection in student submissions. *Proceedings of the First Annual Conference of the Learning and Teaching Support Network for Information and Computer Sciences*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.176.8211&rep=rep1&type=pdf
20. Jarrah, A. Al, Alsmad, I., & Za'atreh, Z. (2011). Plagiarism Detection based on studying correlation between Author, Title and Content. *International Conference on Information Communication System (CICS), May 22-24, Irbid, Jordan.*
21. Kakkonen, T., & Myller, N. (2009). AntiPlag – A Sampling-based Tool for Plagiarism Detection in Student Texts. *Proceedings of the 8th European Conference on E-learning, October 29-30, Bari, Italy*. 287-293.
22. Meuschke, N., & Gipp, B. (2013). State-of-the-art in detecting academic plagiarism. *International Journal for Educational Integrity*, *9*(1), 50-71. doi: 10.21913/IJEI.v9i1.847
23. Mozgovoy, M., Kakkonen, T., & Cosma, G. (2010). Automatic Student Plagiarism Detection: Future Perspectives. *Journal of Educational Computing Research, 43*(4), 511-531. doi: 10.2190/ec.43.4.e

doi 10.15802/stp2017/94034

24. Mozgovoy, M. (2006). Desktop Tools for Offline Plagiarism Detection in Computer Programs. *Informatics in Education, 5*(1), 97-112.
25. OMG Unified Modeling Language (OMG UML), Infrastructure: Version 2.4.1 (2011).
26. *Plagiarisma*. (n.d.) Retrieved from http://plagiarisma.net/
27. *Plagiarism-Checker.com*. (n.d.) Retrieved from http://www.plagiarism-checker.com/
28. *Plagium:Plagiarism checker & plagiarism detection*. (n.d.) Retrieved from http://www.plagium.com/
29. *Plagiarism Detector.* (n.d.) Retrieved from http://plagiarismdetector.net/
30. *PlagScan*. (n.d.) Retrieved from http://www.plagscan.com/
31. *Plagtracker. The most accurate plagiarism checking service*. (n.d.) Retrieved from http://www.plagtracker.com/
32. *Small SEO Tools:Plagiarism Checker*. (n.d.) Retrieved from http://smallseotools.com/plagiarism-checker/
33. *StrikePlagiarism.com*. (n.d.) Retrieved from http://strikeplagiarism.com/ua/
34. *Unplag.com. Plagiarism Detection Engine*. (n.d.) Retrieved from https://ua.unplag.com/

Статья рекомендована к публикации д.т.н., проф. В. В. Скалозубом (Украина); к.филос.н., доц. И. В. Агиенко (Украина)

Поступила в редколлегию: 03.11.2016
Принята к печати: 12.01.2017

doi 10.15802/stp2017/94034