Система проверки текстов на заимствования из других источников

© Р.В. Шарапов, Е.В. Шарапова

Муромский институт (филиал) ГОУ ВПО "Владимирский государственный университет имени Александра Григорьевича и Николая Григорьевича Столетовых"

info@vanta.ru

Аннотация

В статье обсуждается проблема обнаружения заимствования текстов. Рассматриваются основные подходы к обнаружению заимствований, проводится обзор существующих на сегодняшний день программ. Дается описание разработанной системы «Автор.NET», способной проводить проверку заимствований по внутренним источникам и сети Интернет.

1. Введение

Современное развитие информационных технологий и глобальной сети Интернет предоставило широким кругам пользователей доступ к огромным массивам информации. Появилось большое число online-библиотек, содержащих 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 [11].

Как можно убедиться из определения, подобные заимствованные работы можно отнести к разряду плагиата. Задача обнаружения недобросовестного использования заимствованных текстов в учебных и ученых кругах (фактов плагиата) приобретает высокую актуальность.

2. Формы заимствований текстов

Рассмотрим формы заимствований, встречающиеся в практике учебных заведений и подлежащие выявлению.

1. Полное или частичное копирование текста из одного источника.

2. Копирование и компоновка текста из нескольких источников.

3. Копирование текста из другого источника и изменение порядка следования частей текста.

Для того, чтобы скрыть факт заимствований, могут применяться следующие подходы:

1. Корректировка родов, чисел и времен входящих в текст слов. Например, изменение слова «выполнил» на «выполнила» или «выполнили», использование местоимения «я» вместо «мы» в оригинальном тексте и т.д.

2. Незначительное изменение заимствованного текста.

3. Сокращение заимствованного текста путем удаления предложений, абзацев, рисунков, формул и т.д.

Труды 13[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2011, Воронеж, Россия, 2011.

4. Обход систем проверки на плагиат путем замены русских букв на аналогичные по написанию английские и т.д.

5. Осуществление ручной или автоматической синонимизации текста.

Все вышеописанное должно учитываться при создании и использовании систем проверки на заимствования. О правомочности того или иного заимствования решение выносит сам проверяющий.

3. Подходы к обнаружению заимствований

Существует несколько подходов к обнаружению заимствований (или, как их еще называют, нечетких дублей текстов). Достаточно подробный обзор приведен в [9].

Наибольшую известность получил метод «шинглов» [2]. Метод основан на представлении текстов в виде множества последовательностей фиксированной длины, состоящих из соседних слов. При значительном пересечении таких множеств документы будут похожи друг на друга. Одна из модификаций метода, получившая название «супершинглов», используется для быстрого обнаружения подобных документов [9].

Существует ряд методов, использующих сигнатурную лексическую информацию документов. В [4] для этих целей используется I-Match сигнатура, вычисляемая для слов со средним значением IDF (инверсной частоты слов в документах). Другим сигнатурным подходом, основанным на лексических принципах, является метод «опорных» слов [3]. В данном случае для документов составляются по определенным правилам наборы опорных слов, для которых строятся сигнатуры документов. Совпадение сигнатур говорит о подобии самих документов. Эта группа методов, несмотря на большую сложность реализации, показывает более хорошие результаты в обнаружении похожих документов [9].

Для обнаружения заимствований иногда используются алгоритмы, построенные на классических принципах информационного поиска, таких как TF, TF*IDF и т.д. [13]. В [10] предлагается использовать функцию схожести Джаккарда, применение которой позволяет добиться неплохих результатов даже в текстах с использованием синонимов и наличием орфографических ошибок.

4. Обзор существующих систем

Рассмотрим практическое использование описанных подходов в задачах обнаружения плагиата. В настоящее время существует достаточно большое количество сервисов и программ, позволяющих, так или иначе, выявить заимствованный контент. Большую известность получила система «Антиплагиат», разработанная компанией «Форексис» [8]. Система осуществляет поиск по большому количеству коллекций рефератов, контрольных работ и учебников, хранящихся в собственной базе системы. Тем не менее, система имеет ряд недостатков. Во-первых, система не осуществляет поиск по всем документам, доступным в сети Интернет. Особенно это касается тематических сайтов и новостных порталов: большое число заимствований осуществляется именно с таких источников. Соответственно, даже при полном дублировании подобной информации, система «Антиплагиат» соответствий не обнаружит. Во-вторых, присутствует ограничение размера проверяемого текста 3000 или 5000 символами (доступно после регистрации). В-третьих, ограничен просмотр документов, частично соответствующих проверяемому тексту. Кроме того, система ограничивает возможность проверки по базе имеющихся работ.

Программа Advego Plagiatus осуществляет проверку с использованием поисковых систем [1]. Использует разные поисковые системы и проверяет их доступность. В отличие от аналогичных систем, Advego Plagiatus не использует Яндекс.XML. Качество обнаружения плагиата – достаточно высокое. Программа выдает процент совпадения текста и выводит найденные источники. Недостатком является отсутствие преобразования букв, отсутствие поддержки поиска по собственной базе. Из-за особенностей работы программы возникают ситуации, когда результаты проверки отличаются от раза к разу.

Сервис www.miratools.ru позволяет осуществлять On-line проверку текста на плагиат [6]. Система использует результаты выдачи поисковых систем. К достоинствам можно отнести возможность замены английских букв на русские. Имеются возможности изменять длину и шаг шинглов (используемых для проверки). По результатам проверки выдается процент совпадений и найденные источники. Система не работает с собственной базой. Присутствует ограничение на длину текста в 3000 символов и на число проверок в течение суток.

Сервис www.istio.com осуществляет проверку текста на наличие заимствованного контента с использованием поисковых систем [7]. Для этих целей используют Яндекс.XML и Yahoo.com. Возможности сервиса несколько слабее по сравнению с Miratools. По результатам проверки выдается сообщение о том, является ли текст уникальным или нет, и выдается список подобных сайтов. Преобразование букв и поддержка поиска по собственной базе отсутствует. Сервис предоставляет дополнительные средства для анализа текстов, например, проверку орфографии, анализ наиболее частотных слов и т.д.

Программа Praide Unique Content Analyser II [12] имеет широкие возможности по проверке текстов с использованием поисковых систем. Имеется возможность выбора используемых поисковых систем, содержит средства добавления новых

поисковых систем. Проверка осуществляется пассажами и шинглами, длину которых можно изменять. Можно задавать количества слов перекрытия шинглов. Выводится подробный отчет по проверке в каждой поисковой системе. К недостаткам можно отнести отсутствие замены букв и обработки стоп-слов. Нет поддержки работы с собственной базой.

Система Plagiatinform, по заверениям авторов, имеет наиболее широкий функционал [5, 14]. Она умеет проверять документы на наличие заимствований, как в локальной базе, так и в сети Интернет. Система умеет обрабатывать документы, скомпонованные из перемешенных кусков текста нескольких источников. Проверка может осуществляться с использованием быстрого или углубленного поиска. Результаты проверки выдаются в виде наглядного отчета. Авторы не предоставляют возможности свободного использования или тестирования системы, и оценить качества ее работы невозможно.

Результаты сравнения функциональности рассмотренных сервисов проверки на плагиат приведены в таблице 1. Несмотря на большое количество существующих решений, ни одно из них не может служить универсальным средством проверки на плагиат. Основной недостаток большинства существующих систем – это направленность поиска либо на сеть Интернет, либо на собственную базу. Очевидно, что более точная и универсальная проверка будет в случае использования обоих видов источников. Кроме того, большинство систем не способны обрабатывать замену букв, чем часто пользуются недобросовестные авторы (чаще всего студенты).

Таблица 1 - Сравнение функциональности сервисов проверки текстов на плагиат

Система	Поиск в Интернет	Поиск в локальной базе	Обработка замены букв	Под-робный отчет
Advego Plagiatus	+	-	-	+
Антиплагиат	-	+	-	+/-
Istio	+	-	-	-
Miratools	+	-	+	+
Plagiat-inform	+	+	-	+
Praide Unique Content Analyser II	+	-	-	+

Большинство рассмотренных систем используют в своей работе метод «шинглов». По исследованиям [9] этот метод демонстрирует высокую точность обнаружения дублированных текстов. Тем не менее, из-за особенностей реализации результаты проверки в каждой системе сильно отличаются от других. Минусом метода является отсутствие возможности обработки синонимов [10]. Это является значительным недостатком существующих систем.

5. Практическая реализация

На базе Владимирского государственного университета авторами была разработана система проверки текстов на наличие заимствований из других источников (проверки на плагиат) «Автор.NET». Система осуществляет проверку как по источникам, доступным в сети Интернет, так и по собственным источникам (базам статей, курсовых и контрольных работ, дипломных проектов и т.д.). По результатам проверки формируется отчет с подсветкой найденных заимствований и возможностью просмотра найденных источников.

Рассмотрим структуру системы (см. рис. 1).

Проверяемый исходный текст подвергается предварительной обработке, в которую входят:

1. Исключение из текста знаков препинания и спецсимволов.
2. Преобразование регистра.
3. Обработка замены символов (преобразование латинских букв в русских словах на аналогичные буквы русского алфавита для текстов на русском языке).
4. Удаление стоп-слов и знаков препинания (предлоги, наречия и т.д.).
5. Фильтрация текста (удаление не информативных слов).
6. Стемминг (обработка окончаний слов).

Фильтрация текста заключается в удалении наиболее частотных слов, редко встречающихся слов, не информативных слов и т.д. Кроме того, фильтрации подвергаются слова, содержащие спецсимволы, слова большой длины и т.д. Эта процедура позволяет существенно сократить объемы вычислений (длину проверяемого текста).

Стемминг заключается в обработке окончания слов. В нашем случае они просто отбрасываются. Это позволяет исключить влияние таких модификаций текста, как изменение единственного и множественного числа, мужского и женского рода, настоящего и прошедшего времени и т.д.

Система проверки на плагиат «Автор.NET» включает в себя два модуля, каждый их которых функционирует независимо друг от друга.

Первый модуль осуществляет проверку по внутренней базе источников. База источников включает в себя базу статей, курсовых и контрольных работ, дипломных проектов, а также учебников и курсов лекций. Источники хранятся как в виде полных текстов, необходимых для оценки значимости заимствований (по результатам проверки), и в виде специально организованного поискового индекса. Последний необходим для быстрой проверки на совпадения текста и базы источников. Нет необходимости при каждой проверке просматривать все имеющиеся тексты и производить их достаточно трудоемкую обработку. Вся необходимая для поиска информация уже включена в структурированный поисковый индекс, с которым и работает модуль. Поисковый индекс

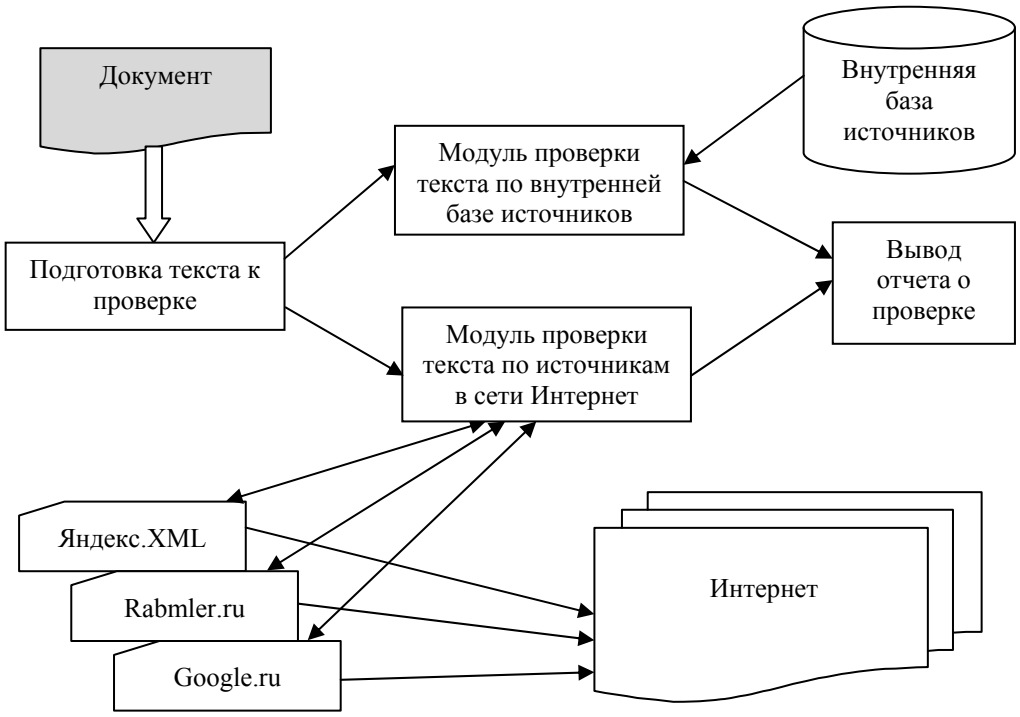

Рис. 1 Структура системы проверки текстов на заимствования

формируется из текстов, прошедших предварительную обработку, описанную выше.

Второй модуль осуществляет проверку по источникам сети Интернет. Для этих целей текст проверяемого документа разбивается на информативные куски (разбиение проводится по полному тексту документа без проведения фильтрации и стемминга). Число таких кусков зависит от размера документа. Далее с использованием поисковых систем проводится поиск источников, содержащих указанные информативные куски. Для осуществления поиска модуль использует Яндекс.XML, а также доступ к on-line поиску систем Google.ru, Rambler.ru, Aport.ru, Поиск.Mail.ru, Nigma.ru и т.д. Полученные таким образом источники проверяются затем на соответствие исходному документу. Для этого определяется формат источника (html-документ, txt-файл, doc- или rtf-документ, pdf-файл). В случае html-документа из источника удаляются теги разметки. Файлы *.doc, *.rtf и *.pdf преобразуются, если это возможно, в обычный текстовый формат без разметки. Далее источники проходят предварительную обработку и затем проводится оценка их сходства с исходным документом.

Для оценки сходства исходного документа и источников используется некая модификация алгоритма «шинглов». Модификация алгоритма заключается в том, что рассматриваются не оригинальный документ, а его обработанная и отфильтрованная копия с исключением неинформативных объектов. Основное требование к системе – полнота и точность оценки совпадений. Мы не ставили задачей сокращение времени проверки, проведение экспресс оценки на полные дубли и т.д.

В настоящее время локальная база системы содержит дипломные проекты, выполненные за последние 6 лет и курсовые проекты, выполненные за последние 3 года студентами одной из специальностей. Также в базе содержится ряд контрольных работ, выполненных студентами заочной формы обучения.

6. Результаты исследования

Для оценки частоты использования тех или иных форм плагиата мы провели следующий эксперимент. Студентам двух групп гуманитарных специальностей было предложено написать статьи на тему экологической ситуации региона (Владимирской области). Студенты были предупреждены о том, что статьи будут проверяться на наличие плагиата. Из полученного набора статей по результатам проверки системой «Автор.NET» были исключены оригинальные статьи. Анализ статей, содержащих заимствованный контент, показал, что большинство из них скомпонованы из нескольких (реже одного) источников, чаще всего из учебников, статей из сети Интернет и публикаций региональной прессы (см. таблицу 2). Тот факт, что доля статей, полностью или частично скопированных только из одного источника, составила всего 36% (в реальных условиях она часто бывает больше), вероятно связан со знанием авторов о том, что работы будут проверяться. Доля работ, составленных путем копирования текста из

другого источника и изменения порядка следования частей текста, оказалась незначительной (2%).

Таблица 2 – Формы плагиата

Форма плагиата	Доля, %
Полное или частичное копирование текста из одного источника	36%
Копирование и компоновка текста из нескольких источников	62%
Копирование текста из другого источника и изменение порядка следования частей текста	2%

Анализ подходов, используемых студентами для сокрытия факта плагиата, показал, что в 32% работ осуществлялась корректировка родов, чисел и времен слов (см. таблицу 3). В 38% работ (составленных как из одного, так и из нескольких источников) осуществлялось незначительное изменение заимствованного текста. 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. Из работ, скопированных из одного источника, 44% 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. Замена букв осуществлялась в 4% работ. В одной из работ замене подверглись практически все русские буквы, сходные по написанию с английскими буквами. В остальных работах заменялись одна-две гласные буквы. Ручная синонимизация проводилась только в 2% работ. Применения автоматической синонимизации в статьях замечено не было. Надо заметить, что около 40% рассматриваемых работ вообще не подвергались каким либо изменениям, призванным скрыть факты плагиата.

Таблица 3 – Частота использования подходов к сокрытию фактов плагиата

Подходы к скрытию плагиата	Доля, %
Корректировка родов, чисел и времен входящих в текст слов	32%
Незначительное изменение текста	38%
Сокращение заимствованного текста	44%
Замена букв	4%
Синонимизация текста	2%

Для проверки работоспособности системы «Автор.NET» нами были составлены тесты трех видов:

1. Заимствования с изменением в тексте времен и родов слов (Т1).
2. Заимствования из одного источника с измененным порядком следования предложений и добавлением оригинального текста между предложениями (Т2).
3. Заимствования, взятые из нескольких источников, с измененным порядком следований предложений (Т3).

Все тесты имели приблизительно одинаковый размер в 2000 символов и содержали в среднем по 400 слов. В качестве источника текстов для составления тестов использовалась коллекция рефератов, широко доступная в сети Интернет. Было составлено по 10 тестов каждого вида.

Для оценки качества обнаружения заимствований мы сравнили результаты работы системы с результатами систем Антиплагиат, Advego Plagiatus и Miratools. В связи с тем, что каждая система имеет свои принципы подсчета оригинальности документа, в качестве метрики оригинальности мы использовали процентное отношение оригинальных слов в документе к общему количеству слов.

Для оценки качества обнаружения заимствований использовался показатель полноты (Recall), показывающий, какой процент заимствований был обнаружен (см. таблицу 4). Точность обнаружения (Precision) во всех системах была на высоком уровне и стремилась к 1.

Таблица 4 – Результаты тестирования

Система	Т1	Т2	Т3
Антиплагиат	0	1	0.97
Miratools	0	0.9	0.83
Advego Plagiatus	0.14	1	0.62
Автор.NET	0.99	1	0.98

Как можно заметить, ни одна из трех рассматриваемых систем не справилась с тестом на замену окончаний (Т1). Показатель Advego Plagiatus объясняется наличием в измененном тексте цепочек из 5 слов, для которых окончания не менялись. Применение стемминга в системе «Автор.NET» позволило ей стравиться с указанной задачей и обнаружить заимствования.

С задачей обнаружения изменения порядка следования предложений, взятых из одного источника (Т2), справились все системы. Чуть худший результат Miratools (полнота 0.9) объясняется, видимо, особенностями реализации алгоритма сравнения в этой системе.

С задачей обнаружения предложений, взятых из разных источников с изменением порядка их следования (Т3), рассматриваемые системы справились немного хуже. Система Антиплагиат показала хорошее значение полноты (0.97). Результаты система Miratools оказались более скромными (полнота 0.83). В системе Advego Plagiatus полнота иногда опускалась до 0.45 при

среднем значении в 0.62. Система «Автор.NET» хорошо справилась с указанной задачей, продемонстрировав полноту в 0.98.

Как можно заметить, система «Автор.NET» хорошо справилась со всеми видами тестов и показала результаты, не уступающие, а иногда и превосходящие результаты работы существующих систем.

7. Выводы

Таким образом, разработанная система «Автор.NET» проверки текстов на плагиат показала достаточно хорошие результаты. Использование фильтрации текста, стемминга и преобразования символов, позволило системе находить заимствованные тексты даже при их незначительной модификации.

Особенностью системы является возможность проведения проверки, как по внутренней базе источников, так и по источникам сети Интернет. Это делает систему достаточно универсальным средством проверки текстов и выгодно отличает ее от существующих систем. Выдаваемые системой отчеты позволяют оценивать правомерность найденных заимствований текстов.

Система может использоваться для проверки уникальности студенческих работ (курсовых и дипломных проектов, рефератов и контрольных работ). Еще одной областью применения может служить использование системы для проверки докладов, представляемых на студенческие и молодежные научные конференции.

Литература

[1] Advego Plagiatus - проверка уникальности текста [Электронный ресурс]. — Режим доступа: http://advego.ru/plagiatus/ (дата обращения: 11.04.2011)

[2] Broder A. On the resemblance and containment of documents // Compression and Complexity of Sequences (SEQUENCES'97). IEEE Computer Society, 1998. P. 21-29.

[3] Ilyinsky S., Kuzmin M., Melkov A., Segalovich I. An efficient method to detect duplicates of Web documents with the use of inverted index // WWW Conference 2002.

[4] Kolcz A., Chowdhury A., Alspector J. Improved Robustness of Signature-Based Near-Replica Detection via Lexicon Randomization // KDD 2004, 22-25 August, 2004, Seattle, Washington, USA.

[5] SearchInform Плагиат-Информ - система для определения плагиата в документах [Электронный ресурс]. — Режим доступа: http://www.searchinform.ru/main/full-text-search-plagiarism-search-plagiatinform.html (дата обращения: 11.04.2011)

[6] www.miratools.ru - Сервис проверки уникальности контента [Электронный ресурс]. — Режим доступа: http:// www.miratools.ru/ (дата обращения: 11.04.2011)

[7] Анализировать текст, поиск плагиата | istio.com [Электронный ресурс]. — Режим доступа: http://istio.com/rus/text/analyz/ (дата обращения: 11.04.2011)

[8] Антиплагиат [Электронный ресурс]. — Режим доступа: http://www.antiplagiat.ru/ (дата обращения: 11.04.2011)

[9] Зеленков Ю.Г., Сегалович И.В. Сравнительный анализ методов определения нечетких дубликатов для WEB-документов // Труды 9-ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» RCDL'2007: сб. работ участников конкурса – Переславль-Залесский, 2007. – Т. 1. – С. 166-174.

[10] Неелова Н.В., Сычугов А.А. Сравнение результатов детектирования дублей методом шинглов и методом Джаккарда // Вестник РГРТУ. № 4 (выпуск 34). Рязань, 2010 – с. 72-78.

[11] Плагиат - Википедия [Электронный ресурс]. — Режим доступа: http://ru.wikipedia.org/wiki/Плагиат (дата обращения: 11.04.2011)

[12] Проверка уникальности текста в Интернете - очень полезная программа для качественной раскрутки сайтов [Электронный ресурс]. — Режим доступа: http://www.nado.su/downloads.html (дата обращения: 23.03.2011)

[13] Шарапов Р.В., Шарапова Е.В. Пути расширения булевой модели поиска // Информационные системы и технологии. Известия Орел ГТУ – Орел: ОрелГТУ, 2009 №6(56) – С. 74-78

[14] Ширяев М.А., Мустакимов В. Plagiatinform избавит от плагиата в научных работах // Educational Technology & Society 11(1) 2008, с. 367-374

System of Duplicate Texts Detection

© R.V. Sharapov, E.V. Sharapova

In the article we discuss the problem of duplicate texts detecting. The basic approaches to detection of text duplicates are considered. We review the existing programs of duplicate texts detecting. We create a system «Autor.NET» which checks text duplications at the internal sources and at the Internet.