АВТОМАТИЗИРОВАННАЯ СИСТЕМА ПОИСКА ПЛАГИАТА ДЛЯ КАФЕДР ВУЗОВ

Е.Н. Унучек, А.О. Позняк, М.М. Радакович

Белорусский государственный университет информатики и радиоэлектроники, Минск, Беларусь, e.unuchek@gmail.com, Hanna_Pazniak@epam.com, m.radakovic@vildisi.net

Abstract. The main purpose of this project is using of special tools to identify plagiarism as a verification tool and enhance the effectiveness of teachers. Such service allows a direct comparison with the earlier loaded documents and samples of works. After the comparison user can view the summary statistics of the matches found represented in the form of graphics, or download a detailed report.

Использование дистанционного обучения дает такие очевидные преимущества как: персонализация процесса обучения (возможность обучения различных категорий людей, в том числе с ограниченными способностями); отсутствие географических и временных барьеров; повышение интенсивности обучения; отимизация и автоматизация процесса передачи знаний; экономия (расходы на обучение одного обучаемого при использовании дистанционной формы намного меньше, чем при очном обучении).

Основные недостатки, связанные с дистанционным обучением можно разделить на психологические, связанные с высокими требованиями к самоорганизации и технические, которые обусловлены несовершенством контента, технологий и телекоммуникационной инфраструктуры. Но если развитие информационных систем и технологий в перспективе позволит минимизировать технические недостатки, то проблемы психологического характера, связанные с необходимостью высокой мотивации и отсутствием живого общения, скорее всего, будут решаться по мере развития общества.

Традиционно, дистанционное обучение подразумевает под собой наличие специально разработанной, опубликованной через сеть Интернет учебно-методической информации, педагогическое общение в реальном и отложенном времени между участниками образовательного процесса, некоторые организационно-административные функции и самое важное – систему итоговой проверки полученных знаний. Существует огромное количество апробированных методов проверки результатов самостоятельной работы обучаемых, например выполнение типовых расчетов, контрольных работ, тестирование с использованием специализированного программного обеспечения, курсовое и дипломное проектирование. Однако в условиях современного массового обучения возникает вопрос: как повысить эффективность этих методов контроля и снизить нагрузку на преподавателя?

Использование специализированных оболочек для тестирования является хорошим решением, но они чаще всего недостаточно гибки и не всегда реализуют поддержку психолого-педагогических особенностей обучающих и обучаемых. А если необходимо проводить контроль по дисциплинам гуманитарной специальности?

Как и раньше, сегодня по прежнему главными проблемами обучения являются нехватка самоорганизации, мотивации, недостаточная ответственность обучаемого.

Использование общедоступной информации, в первую очередь из сети Интернет, может как положительно влиять на процесс обучения, так и отрицательно. Процесс индексации практически всей информации в сети Интернет значительно упрощает поиск, удаленный доступ и копирование необходимых ресурсов. Нерадивый студент может найти требуемую информацию, в том числе и готовые варианты работ,

прикладывая к этому минимум усилий. Данное благо современности привело к массовому распространению плагиата. Значительная часть докладов, рефератов, курсовых или дипломных проектов и работ частично или полностью списаны из Интернет, умышленно присвоены себе как автору данного произведения. Сегодня аналитики считают плагиат одной из основных причин кризиса в образовании[1].

Минимизировать случаи некорректного использования общедоступной информации, позаимствованной из Всемирной паутины, в материалах итогового контроля образовательного процесса поможет использование в учебном процессе автоматизированной системы поиска плагиата.

В настоящее время существует значительное количество автоматизированных систем, позволяющих осуществлять поиск плагиата. Наиболее известными системами поиска нарушения авторских прав ближнего зарубежья ориентированных на поиск плагиата в Интернете при помощи ресурсов различных поисковых систем (Google, Yandex, Yahoo) являются программы «Детектор плагиата», «PlagiatInform», «АнтиПлагиат». В качестве ведущих на мировом рынке зарубежных онлайновых средств проверки документов на наличие плагиата можно назвать «Turnitin tool», «Plagiarism-Finder», «CopyChecker» и многие другие [2].

Для решения задачи определения дубликатов документов, содержащих текст, популярным подходом является метод Андрея Бродера (Andrei Broder) разработанный в 1997[3]. Он создал алгоритм, использующий для анализа схожести двух документов пересекающиеся куски текста или *шинглы* (от слова англ. shingles, «черепички, чешуйки»).

Пусть D – совокупность слов некоторого документа. Перекрывающие друг друга последовательности, содержащиеся в D, будем называть их *шинглами*. Определим $S(D,w)$, как совокупность всех уникальных шинглов размера w содержащихся в документе D.

Например, строка, разбитая с помощью пятисловных шинглов (w=4) «Клара у Карла украла кораллы, Карл у Клары украл кларнет» будет выглядеть так:

{(Клара у Карла украла), (у Карла украла кораллы), (украла кораллы Карл у), (кораллы Карл у Клары), (Карл у Клары украл кларнет)}.

Для шинглов одинакового размера сходство двух документов A и B определяется как:

$$\mathrm{r(A,B)} = \frac{|S(A) \bigcap S(B)|}{|S(A) \bigcup S(B)|},$$

отношение числа одинаковых для обеих страниц шинглов к общему количеству различных шинглов.

Подобно этому вложенность A в B определяется:

$$\mathrm{c(A,B)} = \frac{|S(A) \bigcap S(B)|}{|S(A)|},$$

отношением числа одинаковых для обеих страниц шинглов к числу шинглов страницы A.

Зачастую коллекция документов, для которой происходит поиск дубликатов, имеет большой размер, что накладывает серьезные ограничения на быстродействие системы. Сравнение всех шинглов одного документа со всеми шинглами другого документа займет слишком много времени. Решением проблемы в классическом алгоритме Бродера является использование шинглов кратных какому-нибудь небольшому числу (10-30). Критерий выбора, в данном случае, получается не привязанным к особенностям текста, так как значения контрольных сумм для разных

документов распределены равномерно. Преимущества такого подхода для оптимизации метода шинглов очевидны, так как с его помощью существенно сокращается количество сравниваемых величин без значительного ухудшения качества работы алгоритма.

Выбор значения кратности шинглов, используемых для анализа сходства, осуществляется исходя из желательности проведения расчетов сравнения исключительно с помощью оперативной памяти компьютера. Кроме того, необходимо принять во внимание, что для коротких документов алгоритм отбора шинглов может не выбрать ни одного подходящего шингла или выбрать слишком мало для качественного сравнения.

Для решения вышеперечисленных проблем, предлагается реализация автоматизированной системы поиска плагиата, в основе которой лежит следующий механизм работы: загруженный документ разбивается на фрагменты, которые сравниваются с содержимым базы данных при помощи алгоритма Андрея Бродера. Поиск плагиата происходит на базе сравнения исходного текста с содержимым внутренних баз данных. База данных, главным образом, должна пополняться за счет загрузки других работ, выполнявшихся ранее.

В предлагаемой автоматизированной системе пользователю предоставляется возможность пополнения и редактирования базы, хранящей уже проверенные работы студентов. Кроме того, база может пополняться произведениями классиков, учебными и научными работами студентов, преподавателей.

Система предоставляет возможность непосредственного сравнения документа с ранее загруженными образцами работ на основе полного текстового совпадения или же сравнение по абзацам и предложениям. После проведения сравнения пользователю предоставляется возможность просмотреть итоговую статистику о найденных совпадениях, представленную в виде графика, либо скачать подробный отчет, открываемый при помощи любого текстового редактора.

Тестирование приложения показало стабильность работы предоставляемого функционала и логичную структуру интерфейса программы. Вместе с тем дальнейшее использование автоматизированной системы требует совершенствования алгоритма поиска плагиата, например посредством добавления предварительной лингвистической обработки текстовой информации, и расширения функциональности, за счет реализации возможности сравнения исходного документа с документами, опубликованными в сети Интернет.

Уникальность предлагаемой автоматизированной системы состоит в её реализации на базе сервис-ориентированной архитектуры, что предоставляет возможность её использования как подсистемы в рамках взаимодействия с другими системами, в частности с комплексной системой проверки результатов самостоятельной работы обучаемых.

Внедрение автоматизированной системы поиска плагиата в деятельность кафедры ВУЗа позволит перевести качество проверки знаний студентов на новый уровень, сократит временные затраты преподавателя на проверку индивидуальных заданий, повысит мотивацию обучающихся и качество из работы с первоисточниками.

Литература

1. A. Madray. Developing Students' Awareness of Plagiarism: Crisis and Opportunities, 2009

2. Поисковые системы, каталоги и интернет-бизнес. Статьи и новости об электронной коммерции. URL: http://www.iskati.com/

3. Andrei Z. Broder, Steven C. Glassman, Mark S. Syntactic Clustering of the Web. Manasse WWW, 1997