УДК 004.912

Е.В. Костерин*

РАЗРАБОТКА СПОСОБОВ, АЛГОРИТМИЧЕСКОГО И ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ АНАЛИЗА УНИКАЛЬНОСТИ ДОКУМЕНТОВ В ИНФОРМАЦИОННЫХ СИСТЕМАХ

В настоящее время существует свободный доступ к информации в Интернете. Учащимся вузов стало легче делать доклады, курсовые работы, писать статьи и дипломные работы. Они копируют большинство информации из Интернета, не вникая в суть проблемы, что негативно сказывается на знаниях. Как следствие, наши вузы выпускают специалистов низкого уровня. Также много работ выполняется в графическом виде, а средств – для поиска и анализа уникальности графического материала очень мало.

---

* Работа представлена в отборочном туре программы У.М.Н.И.К. 2012 г. в рамках Седьмой научной студенческой конференции «Проблемы техногенной безопасности и устойчивого развития» ассоциации «Объединенный университет им. В.И. Вернадского» и выполнена под руководством д-ра техн. наук, профессора ФГБОУ ВПО «ТГТУ» В.Е. Дидриха.

Система анализа уникальности документов (САУД) широко используется для проверки на предмет заимствований материала из открытых источников.

Целью работы является разработка САУД для учебно-исследовательской и проектной деятельности, посредством принципиально нового подхода к процессу анализа информации в сети Интернет.

Достижение поставленной цели обеспечит повышение уровня работ учащихся вузов. Программный комплекс подразумевает повышенный контроль анализируемой информации, уникальный алгоритм анализа информации и принятия решения о ее качестве и уникально спроектированную базу данных для хранения и обработки запросов.

Построенные математические модели, алгоритмы, технологические решения, предложенные для достижения поставленной цели, лягут в основу программного комплекса для анализа уникальности документов с текстовым и графическим наполнением, с возможностью определения уровня плагиата, зависящего от положения в документе и от величины неуникальной информации.

В целом, программное обеспечение позволит повысить уровень ответственности обучаемых, улучшить качество получаемых знаний и автоматизировать процесс анализа уникальности информации.

Научная новизна САУД заключается в способе анализа информации (принятия решения об уникальности документов). Программный комплекс отличается от известных существующих программ для анализа уникальности текста, тем, что все программные продукты анализируют только текстовую составляющую, в то время как САУД анализирует документ в целом (с изображениями). Когда анализ будет закончен, программный комплекс сообщит также и о качестве работы. Например, в хороших работах, в начале текста неуникальный контент встречается чаще, чем в основной части.

На сегодняшний день представлено множество программ анализа текста, но они в большинстве своем работают через поисковые системы.

Рассмотрим имеющиеся на рынке аналоги:

1. «Антиплагиат» – это продукт, предназначенный для поиска неуникального текста и позиционирующий себя как средство для борьбы с плагиатом в учебных заведениях, и созданный на основе Интернет/Интранет технологий. Для анализа текста в нем используются уникальные алгоритмы.

2. Система «Плагиат-Информ», разработанная компанией Софт-Информ. Программа на первом этапе сравнивает сдаваемую работу с уже имеющимися в базе рефератами и курсовыми целиком. Если плагиат не отслежен, то программа повторно проверяет ее, предварительно разбив на абзацы. Программный продукт может использоваться в

рамках одного вуза и в сети вузов, что позволяет сравнивать сдаваемые работы в разных вузах между собой.

3. Advego Plagiatus – программа поиска в Интернете частичных или полных копий текстового документа.

4. eTXT Антиплагиат – программа проверки уникальности текста. Осуществляет поиск совпадений текста в Интернете.

Разрабатываемая САУД обладает схожим функционалом с программным продуктом «Антиплагиат», но в отличие от него анализирует не только текстовый материал, но и графический. Таким образом, она выдает более точные результаты уникальности.

Остальные программы направлены на анализ текста для продвигаемых ресурсов в сети Интернет. Для проверки же уникальности изображения существует, например, сервис tineye.com, однако добавлять по одной картинке для поиска не очень удобно и задачи этот сервис выполняет другие.

В создаваемую САУД должны входить следующие основные компоненты:

1) поисковая машина;
2) база данных ресурсов в сети Интернет;
3) сервис анализа уникальности текста;
4) сервис анализа уникальности изображений;
5) веб-интерфейс администратора;
6) веб-интерфейс пользователя;
7) набор интерфейсов прикладного программирования и протоколов взаимодействия компонентов.

Кроме того, в САУД планируется построить оптимизированную базу данных, основанную на разработанных моделях, для ускорения работы с данными.

Для анализа уникальности текста планируется реализовать метод выявления дубликатов, заключающийся в признании документов дубликатами, если у них совпадает более 6 из 12 отобранных по статистическим критериям ключевых слов [1].

Продукт должен обеспечивать:

1) обмен файлами, сообщениями электронной почты и другими видами информации;
2) независимость от пользовательской платформы;
3) прозрачность на уровне доступа;
4) масштабируемость;
5) целостность и конфиденциальность данных.

Программное обеспечение должно функционировать под управлением следующих операционных систем:

1) компоненты операторской части – ОС Linux;
2) компоненты клиентской части – ОС Windows, ОС Linux.

Рассмотрим контингент покупателей и предполагаемый объем платежеспособного рынка.

Согласно статистике Росстата, за 2011 год в России насчитывалось 56 тысяч средних и высших образовательных учреждений. Учитывая область применения разрабатываемой САУД, мы можем сказать, что потенциальными потребителями нашего продукта являются все образовательные учреждения РФ. Если принять за ориентир, что 5% учреждений согласны внедрить в процесс обучения разрабатываемую САУД, мы получаем 2800 образовательных учреждений. В данном случае непосредственный контингент покупателей системы составляют администрации соответствующих учреждений или муниципальные образовательные контролирующие органы.

Также были проанализированы поисковые запросы по ключевым словам, относящимся к САУД, и к учебно-исследовательской деятельности в целом. За последние 12 месяцев пользователи поисковой системы Яндекс обращались к системе со следующими запросами (в среднем за месяц):

– анализ уникальности – 197 запросов;

– уникальность текста – 15 543 запроса;

– уникальность изображений – 141 запрос;

– проверить на уникальность онлайн – 1780 запросов.

Полученные результаты дают основания полагать, что потенциальный рынок содержит не только образовательные учреждения, но и индивидуальных пользователей проектируемой САУД.

Таким образом, согласно российскому классификатору деятельности ОКВЭД разрабатываемый продукт может применяться в следующих областях:

– 80.10.3 Дополнительное образование детей;

– 80.21 Основное общее и среднее (полное) общее образование;

– 80.22.2 Среднее профессиональное образование;

– 80.30.1 Обучение в образовательных учреждениях высшего профессионального образования (университетах, академиях, институтах и в др.);

– 80.30.2 Послевузовское профессиональное образование.

Рассмотрим ориентировочную цену и себестоимость (в расчете на единицу продукции), планируемую прибыль на единицу продукта.

Учитывая ценовую политику конкурентов и существующие способы распространения продуктов, ориентировочная стоимость годовой подписки для образовательного учреждения не будет превышать 15 000 рублей.

Для индивидуальных пользователей предполагается предоставление сервиса на срок от 1 месяца, при этом месячная подписка не будет превышать 500 рублей с ограничением количества анализа документов в день (не более 15 в день).

Рассмотрим ценовую политику конкурентного продукта.

Продукт «Антиплагиат» полный пакет услуг стоит 22 тыс. р. в год. Для преподавателей бесплатно, но с ограничением по отчетам и дополнительным коллекциям.

Таким образом, в статье была рассмотрена проблема плагиата в вузах и информационные системы, позволяющие бороться с ним. А также предложены принципы построения САУД, которая позволит выполнять комплексный анализ уникальности документов.

СПИСОК ЛИТЕРАТУРЫ

1. Ландэ, Д.В. Интернетика: Навигация в сложных сетях: модели и алгоритмы / Д.В. Ландэ, А.А. Снарский, И.В. Безсуднов. – М. : Книжный дом «ЛИБРОКОМ», 2009. – 264 с.

Кафедра «Информационные системы и защита информации»
ФГБОУ ВПО «ТГТУ»