ОБ ИСПОЛЬЗОВАНИИ СИСТЕМЫ АВТОМАТИЧЕСКОГО РАСПОЗНАВАНИЯ ВОСПРОИЗВЕДЕННЫХ ФРАГМЕНТОВ ТЕКСТОВЫХ ДОКУМЕНТОВ В УЧЕБНОМ ПРОЦЕССЕ

Крапивин Ю. Б.
Кафедра интеллектуальных информационных технологий, Брестский государственный технический университет
Брест, Республика Беларусь
E-mail: ybox@list.ru

Приводятся результаты работы системы автоматического распознавания воспроизведенных фрагментов текстовых документов, полученные при обработке дипломных работ студентов, с целью повышения качества контроля учебного процесса на этапе подготовки дипломных работ студентов ВУЗов.

Введение

Последние десятилетия характеризуются стремительным развитием информационных технологий во всем мире, что находит свое отражение в популяризации электронной формы хранения, накопления и обработки информации во всех сферах человеческой деятельности. С каждым годом все больше информации размещается в сети Интернет, которая выступает в качестве средства не только ее распространения, но и хранения. При этом быстрый рост числа Интернет-ресурсов с разнообразным контентом, предоставляющих доступ к полнотекстовым базам данных и обладающих простым и понятным интерфейсом, существенно облегчает работу даже неподготовленного пользователя, позволяя удовлетворить практически любую информационную потребность.

I. Постановка задачи

Постоянно увеличивающийся объем информации, которая главным образом представлена в виде текста – документов различных форматов, доступной в сети Интернет, кроме очевидных преимуществ создает множество проблем. Одной из них является автоматическое распознавание плагиата, под которым обычно понимают умышленное присвоение авторства на чужое произведение литературы, науки, искусства, изобретение или рационализаторское предложение (полностью или частично). Случаи плагиата могут быть и непреднамеренными, например, вследствие сильного внешнего информационного влияния, которое может проявляться в использовании идей или характерного способа их выражения, а также несоблюдения общепринятых правил цитирования, в случае информации, представленной в текстовой форме [1]. Решение указанной проблемы целесообразно осуществлять в два этапа:

- распознавание эквивалентных (точное совпадение или совпадение с точностью до лексической и грамматической синонимии) фрагментов у заданного текстового документа и текстовых документов из заданной базы данных или доступных Интернет-источников;
- анализ, как правило с привлечением экспертов, эквивалентных фрагментов на предмет их заимствования, т.е. на предмет наличия плагиата.

Таким образом, речь идет о распознавании воспроизведенных фрагментов текстовых документов, т.е. тех фрагментов данного (входного) документа, которые заимствованы из других документов, представленных, в конечном счете, в некоторой заданной многоязычной полнотекстовой базе данных, в нашем случае – белорусско-русской [1]. В [2] нами была определена базовая функциональность, а также структурно-функциональная схема системы автоматического распознавания воспроизведенных фрагментов текстового документа.

II. Тестирование системы

Указанная система тестировалась на пояснительных записках к дипломным работам студентов как гуманитарного, так и технического профилей УО БрГТУ, всего 94 текстовых документа. Тестирование осуществлялось по следующей схеме: на вход системы поступал текстовый документ, затем из него автоматически выделялись ключевые слова, на основании которых строился поисковый запрос к информационно-поисковой системе Google [3], которая, в свою очередь, решала задачу отбора документов, релевантных входному, для последующего анализа на предмет наличия в них заимствований из полученного множества релевантных документов. При этом только первые 20 найденных документов, рассматривались как источники потенциальных заимствований для анализируемого документа. Что касается формирования поискового запроса, то для этих целей использовался интерактивный режим работы программы, при котором пользователю предоставлялась возможность выбора из автоматически выделенных ключевых слов, тех, которые, по его мнению,

наиболее полно отражают информационное содержание анализируемого документа. Такой подход позволил избавить пользователя от необходимости детального анализа входного документа, что обычно связано со значительными временными затратами. А поскольку пользователь являлся, как правило, экспертом (например, руководителем дипломного проектирования) в области знаний, к которой принадлежал анализируемый документ, то к тому же он имел возможность точного и оперативного, насколько это возможно, выражения информационной потребности в терминах ключевых слов, предложенных системой.

В результате удалось установить, что при написании работ авторы придерживались определенной стратегии, которая заключалась в использовании групп тематических Интернет-ресурсов для получения текстовых фрагментов и их последующего применения в определенных разделах пояснительной записки. При этом различия в тематике дипломных работ и структуре пояснительных записок также имели место. В качестве тематических Интернет-ресурсов выступали как сайты, предоставляющие доступ к бесплатным коллекциям рефератов, дипломных и диссертационных работ, а также оказывающие подобные услуги за оплату, вплоть до написания работы «под заказ», с гарантией отсутствия плагиата, которые широко представлены в доменной зоне «.ru» (в том числе http://bibliofond.ru, http://knowledge.allbest.ru, http://bestreferat.ru), так и сайты новостных агентств, Интернет-площадки бизнес-организаций и пользовательских сообществ (например, http://ru.wikipedia.org, http://habrahabr.ru). Как правило, авторы использовали коллекции рефератов, дипломных и диссертационных работ для представления теоретических выкладок и описания методик решения задач, а также документы, отражающие состояние дел предметной области или объекта автоматизации (например, технологические или экономические показатели) и размещенные на новостных и Интернет-площадках бизнес-организаций. Т.е. в анализируемых работах отмечаются факты заимствований как на уровне идей, так и способов их выражения, причем первое, скорее всего, связано с использованием типовых подходов к решению задач, выносимых на дипломное проектирование.

Как уже отмечалось, системой были проанализированы 94 документа, размер каждого в среднем составлял 91974 символа. При этом количество ссылок на использованные источники составили всего 1328, в расчете на один документ – 14,12, в том числе печатные издания (книги и пособия) – 7,0, Интернет-источники – 3,18. Суммарный размер эквивалентных фрагментов, обнаруженных системой, составил 532003 символа, в расчете на один документ 5660 символов (6,15% от общего размера проанализированных документов).

Количество найденных источников, содержащих фрагменты, составило 205, в расчете на один документ – 2,18. Наиболее часто (в 44,68% работ) эквивалентные фрагменты встречались в коллекциях рефератов, из них в коллекции http://bibliofond.ru в 26% случаев, в коллекции http://knowledge.allbest.ru в 11,7%, в коллекции http://bestreferat.ru в 6,38%. В то же время, в самих работах на коллекции рефератов ссылались в 6,38% работ, и наиболее часто заявляемыми в списке использованных источников (23,4% работ) были http://ru.wikipedia.org, http://habrahabr.ru.

При анализе в 84% (79 из 94) работ обнаружен хотя бы один эквивалентный фрагмент. В 36,17% случаев (34 работы) размер эквивалентных фрагментов превышал среднее его значение (5660 символов) в анализируемой группе, при этом в них было заявлено 104 ссылки, что составляет 34,78% от общего числа ссылок на Интернет-источники.

Среди этих 34 анализируемых работ число работ со ссылками на Интернет-источники составило 35,29% (12 работ), из них наиболее часто заявляемыми в списке использованных источников были http://ru.wikipedia.org – 17,64%, http://knowledge.allbest.ru и http://habrahabr.ru – 8,82%, соответственно. При этом наиболее часто (в 88,23% работ) эквивалентные фрагменты встречались в коллекциях рефератов, из них в коллекции http://bibliofond.ru в 47,05% случаев, в коллекции http://knowledge.allbest.ru в 29,41%, в коллекции http://bestreferat.ru в 11,76%.

Заключение

В целом полученные результаты позволяют сделать вывод о возможности применения разработанной системы для повышения качества контроля учебного процесса на этапе подготовки дипломных работ студентов ВУЗов.

III. Список литературы

1. Крапивин, Ю. Б. Автоматический поиск заимствованных из Интернет-источников фрагментов / Ю. Б. Крапивин // Искусственный интеллект. – 2012. – № 4. – С. 183–189.
2. Крапивин, Ю. Б. К задаче автоматического распознавания воспроизведенных фрагментов текстовых документов / Ю. Б. Крапивин // Вестник БрГТУ: Физика, математика, информатика. – 2009. – № 5 (59). – С. 120–123.
3. Google [Электронный ресурс] / Google. – 2013. – Режим доступа: http://www.google.com. – Дата доступа: 10.09.2013.