мирования отчета об использовании ранее полученных бланков по требованию инспекции Министерства по налогам и сборам Республики Беларусь.

Разработанная подсистема учета бланков строгой отчетности предполагает практическое применение в проектах на средних и крупных предприятиях Республики Беларусь, занимающихся производством, оптовой торговлей, оказанием услуг, строительством и т.д. при работе с бланками строгой отчетности.

Результатом внедрения подсистемы является полная автоматизация учета бланков строгой отчетности с соблюдением всех норм действующего законодательства.

УДК [004.78:33](075.8)

Исследование ключевых проблем при реализации агрегатора новостей

Ляховец П.В., Кулаков А.Т.
Белорусский национальный технический университет

Под агрегатором новостей в рамках рассматриваемой задачи понимается Интернет-портал, собирающий новости из различных источников, для представления пользователям в удобном виде.

В данной работе рассматриваться лишь небольшой раздел задачи, связанных с наполнением базы данных – это создание структуры базы данных, набор кодов на PHP для наполнения, система отсеивания нечетких дубликатов текстов.

База данных должна быстро обрабатывать большое количество запросов на чтение. Система наполнения базы данных должна своевременно наполнять базу данных свежими новостями, предоставлять не только текст новости, но и сопутствующую информацию необходимую для качественного отображения собранных данных конечным пользователям. Кроме этого, система должна исключать добавление в базу данных нечетких дубликатов текстов, задействовав при этом минимально возможное количество системных ресурсов.

Сложность и новизна разработки агрегатора новостей заключается в создании системы поиска нечетких дубликатов текстов.

Происхождение копий документов в Интернете может быть различным. Один и тот же документ на одном и том же сервере может отличаться по техническим причинам: быть представлен в разных кодировках и форматах; может содержать переменные вставки – рекламу или текущую дату.

Широкий класс документов в глобальной сети активно копируется и редактируется – ленты новостных агентств, документация и юридические документы, прейскуранты магазинов, ответы на часто задаваемые вопросы и т.д. Популярные типы изменений: корректура, реорганизация, ревизия, реферирование, раскрытие темы и т.д. Наконец, публикации могут быть скопированы с нарушением авторских прав и изменены злонамеренно с целью затруднить их обнаружение.

Кроме того, индексация поисковыми машинами страниц, генерируемых из баз данных, порождает еще один распространенных класс внешне мало отличающихся документов: анкеты, форумы, страницы товаров в электронных магазинах.

Очевидно, что с полными повторами особых проблем нет. Достаточно сохранять в индексе контрольную сумму текста и игнорировать все остальные тексты с такой же контрольной суммой. Однако этот метод не работает для выявления хотя бы незначительно измененных документов.

Проблема обнаружения нечетких дубликатов является одной из наиболее важных и трудных задач анализа веб-данных и поиска информации в Интернете.

Основным препятствием для успешного решения рассматриваемой задачи является гигантский объем данных, хранимых в базах современных поисковых машин. Такой объем делает практически невозможным (в разумное время) ее «прямое» решение путем попарного сравнения текстов документов. Поэтому в последнее время большое внимание уделяется разработке методов снижения вычислительной сложности создаваемых алгоритмов за счет выбора различных эвристик (например, хеширования определенного фиксированного набора «значимых» слов или предложений документа, сэмплирования набора подстрок текста, использование дактилограмм и др.).

При применении приближенных подходов наблюдается уменьшение (иногда весьма значительное) показателя полноты обнаружения дублей.

Важным фактором, влияющим на точность и полноту определения дубликатов в задачах веб-поиска, является выделение содержательной части веб-страниц с помощью надежного распознавания элементов оформления документов и их последующего удаления.

И, наконец, еще одним ключевым требованием, предъявляемым к качеству алгоритмов детектирования нечетких дубликатов, является их устойчивость к «небольшим» изменениям исходных документов и возможность уверенно обрабатывать короткие документы.

УДК [004.78:33](075.8)

Интеграция системы контроля доступа «PERCO» и системы учета и планирования рабочего времени «Босс-Кадровик»

Савченко М.И., Кулаков А.Т.
Белорусский национальный технический университет

С недавнего времени стали широко использоваться системы контроля доступа (СКД). СКД – совокупность программно-технических средств и организационно-методических мероприятий, с помощью которых решается задача контроля и управления посещением как отдельных помещений, так и задача оперативного контроля за персоналом и временем его нахождения на территории объекта. СКД прошли длительный эволюционный путь от простейших кодовых устройств, управляющих дверным замком, до сложных компьютерных систем, охватывающих целые комплексы зданий. Современные системы контроля доступа имеют множество функциональных возможностей и применений. Традиционными потребителями СКД являются небольшие офисы, предприятия розничной торговли. Однако с развитием технологий область применения СКД значительно расширилась. В качестве примера можно привести такие предприятия как «МАЗ», «Коммунарка», «Минский завод строительных материалов» и многие другие, которые используют СКД более высокого уровня. В состав данных систем включают, как уже давно сложилось турникеты, калитки, электронные замки, считыватели бесконтактных карт, системы пожарной охраны и оповещения, системы видеонаблюдения, а так же комплексы про-