С.П.Орлов//Вестник Волжского университета им. Татищева. – 2015. – № 2(24). – С.63-71.

4. Орлов, С.П. Техническая диагностика радиоэлектронных блоков по тепловым полям элементов/ С.П.Орлов, Е.А. Ахполова // Перспективные информационные технологии (ПИТ 2016): труды Международной научно-технической конференции/под ред.С.А.Прохорова - Самара: Изд. Самарского научного центра РАН, 2016. - С.139-142.

5. Орлов С.П. Метод термографии при контроле электронной аппаратуры авиационной техники/С.П. Орлов, О.Ю. Уютова// Наука и образование транспорту: труды IX Международной научной конф. (Самара, 19-21 октября 2016). – Самара, 2016. – Т. 2. – С. 70-71.

6. Haykin S. Neural networks. A Comprehensive Foundation. Second Edition. Prentice Hall, 1999.

7. LeCun Y., Bottou L., Bengio Y., Haffner P., Gradient-based learning applied to document recognition. (pp. 306-351). IEEE Press, 1998.

8. Гирин Р.В. Двухстадийная нормализация выходных сигналов искусственных нейронных сетей /Р.В.Гирин, С.П. Орлов//Вестник Самарского гос. тех. ун-та. Серия «Технические науки». – 2017. – № 4(56). – С.7-16.

В.И. Жирнов*, Н.М. Виштак**, И.А. Штырова**

МОДУЛЬ ИНДЕКСАЦИИ СЛАБОСТРУКТУРИРОВАННЫХ ДАННЫХ В СИСТЕМЕ ЭЛЕКТРОННОГО ДОКУМЕНТООБОРОТА

(*Федеральное государственное автономное образовательное учреждение высшего профессионального образования «Национальный исследовательский ядерный университет (МИФИ)», г. Москва.
**Балаковский инженерно-технологический институт – филиал федерального государственного автономного образовательного учреждения высшего образования «Национальный исследовательский ядерный университет (МИФИ)», г. Балаково)

В настоящее время автоматизированные информационные системы на предприятиях обрабатывают колоссальные объемы данных. В том числе, системы электронного документооборота (СЭД), используемые на предприятиях, предназначены для централизованной обработки больших массивов слабоструктурированных или неструктурированных данных [1,2,3 и др.].

Слабоструктурированными являются данные, формат хранения которых предполагает, что структура документов не может быть задана заранее и может изменяться во время эксплуатации информационной системы [4,5 и др.].

. В отличие от полностью неструктурированных данных, слабоструктурированным данным характерно иметь некоторые форматы и правила в общем виде, что позволяет с небольшими затратами привести их к структурированному виду.

Большая часть данных, хранимых в СЭД имеют неструктурированный или слабоструктурированный вид. Эффективность поиска по таким данным очень низкая, что ведет за собой снижение эффективности процесса документооборота на предприятии. Соответственно, актуальным вопросом является разработка программного обеспечения, осуществляющего интеллектуальный поиск по слабоструктурированным и неструктурированным данным.

Разрабатываемый программный модуль к СЭД должен эффективно преобразовывать неструктурированные или слабоструктурированные данные в структурированные. Основным требованием является приемлемое время поиска необходимой информации. Рассмотрим основные типы неструктурированных и слабоструктурированных документов, обрабатываемых в СЭД:

1) Документы Microsoft Office, Open Office и аналогичные xml документы.
2) Portable Document Format.
3) Графические документы PNG и JPEG.

Время поиска информации, хранящейся в выше перечисленных документах, растет прямо пропорционально их количеству, вследствие чего получаем неприемлемое время ожидания результата. Решением данной проблемы является выявление полезной информации из документов и сохранение текстового отпечатка. Поиск по структурированной информации имеет малое время ожидания, что является желаемым результатом в работе СЭД.

Для получения текстового отпечатка, разработан модуль сканирования и обработки документов. На рисунке 1 представлен алгоритм работы разработанного модуля.

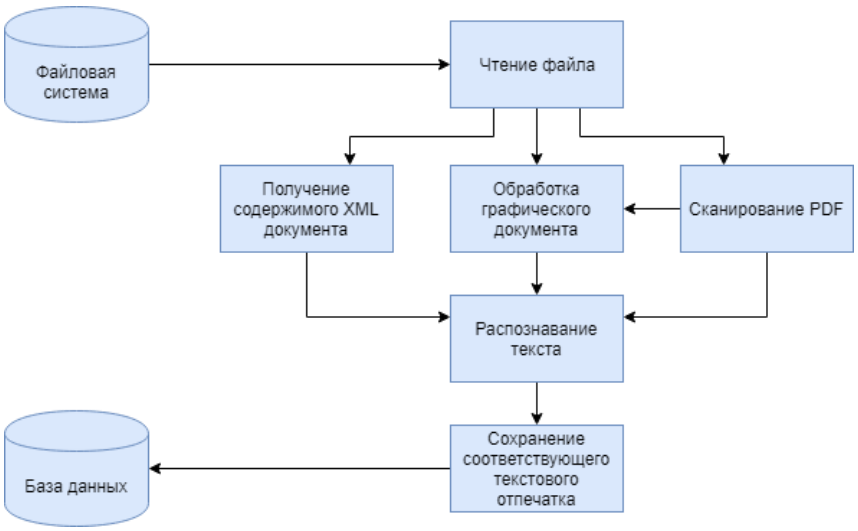

Рисунок 1 – Алгоритм работы модуля сканирования документов

Модуль сканирования документов работает параллельно работе СЭД и позволяет за малое время преобразовать неструктурированную информацию в полностью структурированную. Несмотря на такое очевидное преимущество, как быстрый поиск информации в СЭД, данный модуль имеет недостаток в виде дополнительного потребления дисковой памяти СЭД. В таблице 1 представ-

лена информация о потреблении дискового пространства текстовых отпечатков каждого типа файлов, анализируемых разработанным модулем, изначальный объем информации и время на обработку. В таблице представлены усредненные значения.

Таблица 1 – информация о потреблении дискового пространства

	Реальный объем	Объем отпечатка	Время на сканирование и распознание
XML документы	1 МБайт	100 Кбайт	2 секунды
Графические документы	5 Мбайт	150 Кбайт	5 секунд
PDF	3 Мбайт	150 Кбайт	2 секунды

Исходя из данных в таблице 1, получаем увеличение занимаемого дискового пространства в среднем на 10% и уменьшение время поиска с $3*N_1$ секунд (где N_1 – количество документов в системе) до N_2 секунд (N_2 – время выборки подходящей информации из базы данных).

Внедрение данного модуля в СЭД позволит эффективно выполнять поиск по неструктурированным или слабоструктурированным данным.

Литература

1. Электронный документооборот: что такое электронный документооборот, основные понятия, виды, преимущества, задачи, критерии выбора, классификация систем, требования [Электронный ресурс] – Режим доступа: http://www.docflow.ru/edu/glossary/detail.php?ID=27946, свободный.

2. Гладких Н. А. Применение интеллектуального анализа данных в системах электронного документооборота // Ученые записки. Электронный научный журнал Курского государственного университета. 2010. №2 (14). Режим доступа: https://cyberleninka.ru/article/n/primenenie-intellektualnogo-analiza-dannyh-v-sistemah-elektronnogo-dokumentooborota.

3. Виштак О.В., Жирнов В.И., Ремаренко С.А. Минимизация информационных рисков при использовании систем электронного документооборота в организации. // Сборник трудов. III Международной научно-практической конференции «Проблемы развития предприятий энергетической отрасли в условиях модернизации российской экономики и общества». Балаково: НИЯУ МИФИ, БИТИ НИЯУ МИФИ, 2017. С.85-89

4. Слабоструктурированные данные [Электронный ресурс]: Национальная библиотека им. Н. Э. Баумана / 2017. – Режим доступа: https://ru.bmstu.wiki/Слабоструктурированные_данные, свободный.

5. Semi-structured_data [Электронный ресурс]: Материал из Википедии — свободной энциклопедии: — Режим доступа: https://en.wikipedia.org/wiki/Semi-structured_data.