***Газизуллина Яна Маратовна***

***Студентка 2 курса магистратуры, Институт информатики и кибернетики***
***Федеральное государственное автономное образовательное учреждение высшего образования «Самарский национальный исследовательский университет***
***имени академика С.П. Королева»***
***(Самарский университет)***

# РЕШЕНИЕ ЗАДАЧИ АВТОМАТИЧЕСКОГО ИЗВЛЕЧЕНИЯ СТРУКТУРИРОВАННЫХ ДАННЫХ ИЗ ТЕКСТОВЫХ ДОКУМЕНТОВ С ПОМОЩЬЮ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

**Аннотация:** В статье рассматривается решение задачи автоматического извлечения данных из текстовых документов с применением методов машинного языка. Будет предложено несколько методов распознавания текстовой информации с применением программного обеспечения вычислительных машин. Научная работа направлена на изучение и применение системы распознавания форм различными математическими и программными методами.

**Abstract:** The article considers the solution of the problem of automatic data extraction from text documents using machine language methods. Several methods of recognizing text information using computer software will be proposed. The scientific work is aimed at the study and application of the form recognition system by various mathematical and software methods.

**Ключевые слова:** извлечение данных из текстовых документов, машинное обучение, информационные технологии, модели текстовых документов, распознавание форм.

**Keywords:** data extraction from text documents, machine learning, information technology, models of text documents, form recognition.

## *Введение*

Распознавание и извлечение данных из текстовых документов осуществляется методом структурных образцов. Работа направлена на создание соответствующей модели файла. Для этого внедрена технология машинного обучения, которая строит модели документа, исходя из определенного количества изображений и схем, выполненных вручную.

Данная работа показывает решение задачи автоматического извлечения структурированных данных с использованием метода машинного обучения.

***Распознавание структурированных документов***

Структурированная информация хранится в жестких, полужестких, гибких формах и документах произвольной структуры. Формы отличаются структурой, основа которой состоит из статической части и сведений (полей). Примером такой формы могут быть документы о финансах, счета-фактур, платежные поручения, анкеты.

Извлечение структурированных данных из текстовых документов осуществляется через обработку изображения текста (например, отсканированных образцов) и локализацию информации.

У жестких форм расположение полей строго зафиксировано, за счет чего извлечение структурированных данных осуществляется без глубокого анализа. Для получения информации на изображении текста удаляют искажения и дефекты, появившиеся при сканировании (темные и светлые полосы, пятна). После исправления искажений изображение анализируется через специальный шаблон, в котором указаны координаты полей.

Полужесткие формы имеют более свободное расположение полей и разметку. К таким формам относятся документы, заполненные в электронном виде, и в дальнейшем распечатанные. В полужестких формах поля сдвигают для освобождения места под дополнительную информацию.

У гибких форм есть множество значительных вариаций раскладки. Каждое поле имеет локальный контекст (разделительные линии, заголовки). Расположение поля относительно контекста может быть изменено в незначительной степени.

Документы произвольной структуры являются самыми сложными для распознавания структурированных данных. В частных случаях требуется изучение естественного языка.

***Построение модели документа***

Распознавание и извлечение структурированных данных осуществляется с помощью системы ABBYY FlexiLayout. Технология основана на методах структурного распознавания образцов. Система ABBYY FlexiLayout использует специализированную модель документа (структурное описание). В модели описано множество структурных элементов, которые соответствуют реквизитам текстового документа. У каждого элемента есть фиксированное количеству допустимых значений его атрибутов. Это позволяет определить и локализировать структурный элемент на изображении текстового документа. Анализ осуществляется во время прочтения файла и его структурного описания.

Модель документа задает закономерность расположения структурных элементов в форме порядкового и метрического отношения, связи соседства. Можно задать точные или нечеткие отношения.

В структурном описании хранится интерпретируемый код инициализации параметров и отношений. Программа работает в процессе распознавания. Таким образом, вычисление параметров элементов проводится одновременно с построением модели текстового документа и его изображения.

Описание элементов логической структуры осуществляется одним из следующих способов:

- множество, включающее один графический объект;
- множество слов или словосочетаний (символьные строки);
- множество строк, включающих символы одного или нескольких заданных алфавитов;
- множество текстовых фрагментов с определенными параметрами (количество строк, ширина и высота фрагмента, выравнивание и др.);
- множество разделительных линий с определенными размерами.

Модель позволяет описать обобщенную структуру (обобщенный или шаблонный граф) или конкретное изображение (граф документа).

***Анализ метода машинного обучения для извлечения структурированных данных***

При использовании метода машинного обучения возникает типичная ситуация. Собирают шаблоны текстовых документов, по котором будет выстраиваться алгоритм. Шаблоны должны быть проанализированы и сравнены для их оценки качества и оптимизации метода.

Такие оценки качества не проводят через проверку шаблонов, которые были использованы при обучении. Конечный результат анализа будет слишком завышенным. Если алгоритм будет слишком простым, он выдает оценку шаблонам в 100% и при этом не может обработать новые документы, которые не были использованы при обучении.

Оценка качества осуществляется в 2 этапа: обучающее и тестовое множество. Алгоритм отрабатывается на обучающем множестве. Полученный код применяют на тестовом множестве, после чего определяют метрики качества (полнота, точность). Если тестовое множество будет слишком малым, оценка качества будет неточной. Чем больше это множество, тем эффективнее будет работать алгоритм.

При сравнении разных алгоритмов машинного обучения распознавание осуществляется случайным образом или по определенному параметру, который не влияет и не зависит от содержания текста (например, дата).

***Методы машинного обучения для извлечения структурированных данных из текстовых документов***

Задача извлечения структурированных данных может быть решена следующими способами:

- метод Байеса;
- метод k-ближайших соседей (k-nearest neighbours, k-NN);
- классификатор Роше (Rocchio classifier);
- нейронные сети;

- деревья решений (decision trees);
- построение булевых функций;
- метод опорных векторов (Support Vector Machines, SVM).

Метод Байеса основывается на сопоставлении совместных признаков документа и категорий. Данный способ имеет высокую скорость работы. Модель выстраивается быстра на основе математических вычислений. Метод Байеса чаще всего применяют, как базовый для сравнения алгоритмов машинного обучения.

Метод k-NN — единственная технология, для которой не требуется обучение. Для создания алгоритма большая часть времени тратится на вычисления, однако метод показывает высокую эффективность.

Классификатор Роше является наиболее простым методом извлечения структурированных данных. Все параметры быстро пересчитываются, если в алгоритм добавляют новые примеры. Данное свойство необходимо при решении задач адаптивной фильтрации: пользователь указывает в алгоритме, какие документы выбраны правильно, какие неверно. Система уточняет параметры с учетом добавленных объектов.

Решение задачи на основе искусственных нейронных сетей осуществляется на основе выполнения ряда тренировочных упражнений. Проводится подбор весов межнейронных связей, которые обеспечивают максимальную близость ответов системы к уже зафиксированным правильным решениям.

Метод на базе деревьев решений разбивает информацию на категории по значениям переменных. При определении классификации система отвечает на заранее установленные вопросы (они располагаются на вершинах схемы). Наиболее распространенный алгоритм, созданный на построении деревьев решений — CLS. Эффективная работа основана на процедурах усечения и преобразования схемы. Чем меньше количество узлов, тем качественнее машинное обучение.

Метод Support Vector Machines основан на принципе структурной минимизации рисков. При составлении классификации осуществляется контроль ошибок. Метод SVM наиболее часто используется при построении абстрактной векторной модели. С его помощью осуществляется точное распознавание образов и классификация форм.

***Заключение***

В настоящее время есть множество решений для извлечения структурных данных из текстовых документов на базе машинного языка. Некоторые методы находятся на стадии разработки, например, технология на базе нейронных сетей. Наиболее эффективными являются метод Байеса, k-nearest neighbours и Rocchio classifier. Метод булевых функций и метод деревьев решения сложны в применении из-за большого количества обрабатываемой информации и множества математических вычислений.

**Использованные источники:**

1. https://aws.amazon.com/ru/textract/
2. https://habr.com/ru/post/526984/
3. https://www.machinelearningmastery.ru/machine-learning-text-processing-1d5a2d638958/
4. http://www.ict.nsc.ru/jspui/bitstream/ICT/1455/1/2005_diss_ageev.pdf