УДК 004.624, 004.912

Конвертер PDF-документов для использования в системе семантического анализа текстов на естественном языке

Ткешелашвили Н.М., Анисимова М.А., Ткешелашвили А.М., Университет ИТМО, Санкт-Петербург,
научный руководитель: Клименков С.В., Университет ИТМО, Санкт-Петербург

Извлечение информации из неструктурированных документов является одной из приоритетных областей развития прикладных систем датамайнинга. Современная цифровая экономика хранит данные в множестве офисных форматов, с различной структурой документа. Коллективом авторов была проведена работа по созданию автоматического конвертера, извлекающего информацию из широко распространенных форматов, таких как DOC, ODT, TXT, XLS и других, и преобразующих их в компактный текстовый формат, уточняя и сохраняя семантически значимую информацию и связи в документе, такие, как последовательность токенов в документе, выделение заголовков и относящейся к ним информации, последовательность и принадлежность перечислений, выделение главной информации и др. Конвертер документов построен на базе API LibreOffice и состоит модулей, обрабатывающих соответствующие типы документов: текстовые документы, электронные таблицы и иллюстрационные документы, к которым относится формат PDF. Ввиду особенностей формата хранения, наибольшую сложность представляет обработка PDF документов. Текст внутри PDF хранится блоками, при этом исходный абзац или строка текста оказываются разбиты на несколько блоков, порядок хранения которых не всегда соответствует порядку их следования в документе. В случае наличия в документе нескольких колонок, разрывающих текст изображений или таблиц, восстановление исходной структуры абзацев становится нетривиальной задачей.

Целью работы является разработка алгоритма извлечения структуры из PDF документа и преобразование его к внутреннему формату разрабатываемой системы, с сохранением семантически значимой информации, а также создание программного модуля, реализующего предложенный алгоритм.

В ходе работы был проведён анализ существующих конвертеров из PDF, выявлены их слабые места, сделан обзор современных подходов к обработке PDF документов. В результате авторами предложен алгоритм извлечения текста из PDF документа с восстановлением исходной структуры документа путем объединения снизу вверх текстовых блоков в строки, а строк в абзацы и т.д.. Алгоритм опирается на взаимное расположение элементов, порядок их создания и набор эмпирических правил. Алгоритм реализован в виде программного модуля на языке Java и интегрирован в систему семантического анализа.

Автор _____ / Ткешелашвили Н.М.
Научный руководитель _____ / Клименков С.В.
Руководитель образовательной программы _____ / Алиев Т.И.

Контакты автора: ninomt@cs.ifmo.ru, 8 (812) 325-25-53