**УДК 004.021:004.912**

**Алгоритм анализа структур документа с шаблоном**

**И.Н. Магомедов**

*Балтийский государственный технический университет «ВОЕНМЕХ» им. Д.Ф. Устинова*

Для приложения структурного анализа [1] документа был разработан алгоритм осуществляющий поиск, сравнение, выделение структур и предложений. В данной работе описан алгоритм деления структуры и иерархии документа на отдельные блоки по определённому алгоритму, для упрощения и ускорения поиска, анализа и сравнения структур документов. Алгоритм, представленный в данной работе предназначен для решения проблемы заполнения шаблонов на основе имеющихся данных.

Сравнение блоков происходит по следующему алгоритму:

1. Первые выделенные блоки шаблона сравниваются с первыми выделенными блоками текста. Блоки выделяются на основе структуры документа [2].
2. При совпадении блоков текста с блоками шаблона, следуют следующие этапы:
   1. Морфологический анализ
   2. Синтаксический анализ
   3. Семантический анализ
   4. Проблемный анализ
3. В случае несовпадении блоков запускается процесс сравнения предложений блоков шаблона и текса. Номера блоков и предложений для которых не удалось найти совпадений записываются в отдельный файл для последующего анализа. Найденные части проходят второй этап.
4. Блоки и предложения, для которых не было найдено совпадений проходят дополнительный этап сравнения. В данном случае в предложениях и блоках выделяются схожие слова (например синонимы), части предложений. Найденные блоки и предложения проходят этап под номер 2. Для оставшейся части предложений или блоков применяется сравнение по смысловому подтексту, метаданным и разметке.
5. Следующий этап выявление ключевых слов. Данный этап позволяет ввести пользователю ключевые слова, которые позволят заполнить шаблон. Имеется возможность запустить алгоритм поиска ключевых слов.
6. После создается новый документ, в который копируется шаблон и запускается процесс заполнения шаблона на основе результатов предыдущих этапов. Найденные части целиком подсвечиваются зелёным, части частично совпадающие подсвечиваются желтым и не найденные структуры, и предложения подсвечиваются красным [1].

Алгоритм возможно применять как для не сортированных или структурированных данных, а также частично заполненных шаблонов, а также для отношения один шаблон к множеству текстов (документов).

Спроектированное приложение и разработанный алгоритм имеют широкий круг применения, от учебных заведений до корпоративного сектора, например, может применяться при оформлении договоров, технических заданий, рабочих и учебных планов.

Алгоритм позволит ускорить процесс создания документов за счёт быстрой проверки документа на ошибки и наличие нужных блоков структуры. Даёт возможности ручной настройки процесса заполнения и выделения требуемых частей и ключевых слов для поиска, анализа и сравнения. В следствии настройки алгоритма, позволяет создавать новые шаблоны на основании имеющегося шаблона и документа, что предоставляет гибкость в процессе автоматизированного оформления документа.

**Библиографический список**

1. Магомедов И.Н. Проектирование приложения для структурного анализа документа с шаблоном // Молодёжь. Техника. Космос: материалы X Общероссийской молодёжной науч.-техн. конф. Том 3/Балт. гос. техн. ун-т. – СПБ БГТУ «Военмех», Изд-во «Инфо-Да», 2018. - С. 31.
2. Магомедов И.Н. Сопоставление структур форматов документов odt и docx // Старт-2017: Тезисы докладов Ⅲ Общероссийской молодёжной науч.-техн. конф. – СПб: Балт. гос. техн. ун-т., 2017. – С. 48.