УДК 004.01

«УМНЫЙ» ДОКУМЕНТООБОРОТ: СЭД И МАШИННОЕ ОБУЧЕНИЕ

К.К. Малыхин
Магистрант БГУИР

Silk Data AI, инженер-программист (software developer)

Аннотация. Целью работы является провести анализ возможности внедрения машинного обучения в систему электронного документооборота. Как показал анализ, уже сейчас могут быть применены алгоритмы классификации, алгоритмы семантического анализа текста для ускорения обработки документа, однако финальное слово все еще остается за человеком.

Ключевые слова: документооборот, машинное обучение, классификация, тегирование.

Одной из характерных тенденций наших дней является переход на электронный документооборот и электронное хранение документов. Многие организации сталкиваются с проблемой выбора и внедрения систем электронного документооборота (СЭД). Кроме того, развитие машинного обучения (МО) предоставило возможности внедрения препроцессинга данных, тем самым ускоряя обработку данных.

В первую очередь необходимо определить сферы применения МО. За основу данного анализа был взят документооборот в государственных организациях. Это было обусловлено несколькими причинами.

Во-первых, госучреждения обрабатывают колоссальный объем документов с регламентированным временем их отработки, и от качества и эффективности документационного взаимодействия во многом зависит оперативность и эффективность работы органов государственной власти. Функции СЭД в госорганах не ограничиваются внутренним управлением. Большой объем занимает внешний документооборот – коммуникация с гражданами и организациями по предоставлению государственных сервисов. В связи с развитием электронного правительства количество обрабатываемых запросов может достигать нескольких тысяч в день.

Во-вторых, в государственных структурах типизированы как процессы, так и документы, поэтому применение интеллектуальных алгоритмов будет более эффективным, нежели в структуре, где сложная и уникальная организационная структура.

Машинное обучение может ускорить процесс обработки документов, подготовить все необходимые для принятия решения человеком данные, а еще предотвратить человеческие ошибки и без помощи супер мощного компьютера.

«Умные» технологии призваны, в первую очередь, помочь человеку избавиться от рутинных операций, не требующих принятия каких-либо решений. Благодаря применению алгоритмов МО документ может пройти весь путь от регистрации до формирования с минимальным вмешательством человека в процесс. Однако, следует учитывать, что машине необходимо будет научиться получать различные представления из массива данных (Big Data) в данном случае – это результаты обработки конкретных документов человеком.

Машинное обучение представлено в наши дни множеством алгоритмов, некоторые из которых являются довольно универсальными и могут использоваться для решения разных задач. Чтобы понять, какое место занимают алгоритмы машинного обучения в процессе документооборота, разберем блок задач по обработке текстов.

Одна из первоочередных задач анализа данных, полученных из системы электронного документооборота – это построение кластерной модели данных. Кластерный анализ представляет собой разбиение базы данных на кластеры – группу похожих элементов – и имеет широкий круг применимости. Учитывая объемы документов, которые подвергаются обработке, умение системы разбивать их на кластеры будет полезно перед применением любых алгоритмов машинного обучения в ней. Кластеризация будет полезна для упрощения решения таких задач, как поиск дубликатов, поиск близких/похожих документов и др., а также позволит построить алгоритм для более точного предсказания атрибутов документов. Наиболее очевидные варианты практического применения результатов кластеризации – это автоматическая классификация (или тегирование) новых документов.

Ниже (рисунок 1) представлены результаты кластеризации базы данных реальных документов. Система, обнаруживая сходства в тексте, определяет документ к одному из кластеров, таким образом упорядочивает объекты в сравнительно однородные группы.

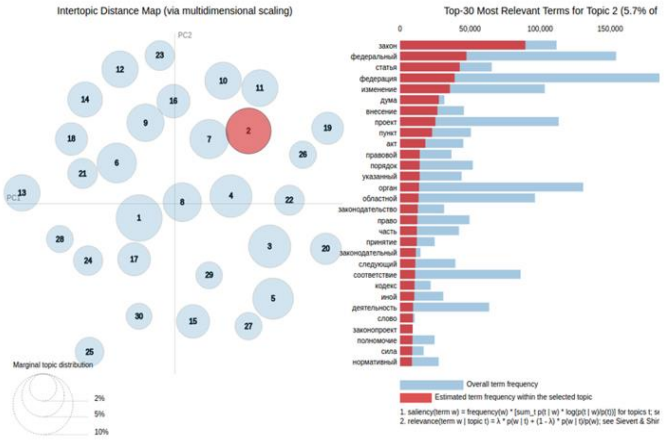

Рисунок 1.

Любой электронный документ сопровождается набором атрибутов (автор, подразделение, вид документа, исполнитель, и др.), которые необходимо заполнить для его дальнейшей обработки, а также последующего поиска документов и формирования отчетов. Это значит, что процесс обработки документа полностью зависит от набора его атрибутов: например, документы, поступившие от определенного адресата и по конкретной теме, должны обрабатываться конкретным подразделением и по вполне конкретным правилам. Сейчас эта процедура обработки каждого документа выполняется на 100% вручную. Но, учитывая структурированность этой информации, тем же правилам легко обучить и алгоритм МО. Обработав базу данных, в которой документы структурированы в соответствии с правилами организации, алгоритмы МО будут готовы самостоятельно прогнозировать новые атрибуты и маршруты обработки для новых документов, а также прогнозировать количество дней, требующихся для выполнения задания, и определять исполнителя. Для того, чтобы алгоритмы научились это делать с высокой точностью, необходима база структурированных и не очень структурированных данных огромных объёмов.

Так, в ходе испытаний алгоритмов машинного обучения специалистам «Диджитал дизайн» удалось достигнуть 95% точности определения подразделения, ответственного за обработку документа, по его содержанию.

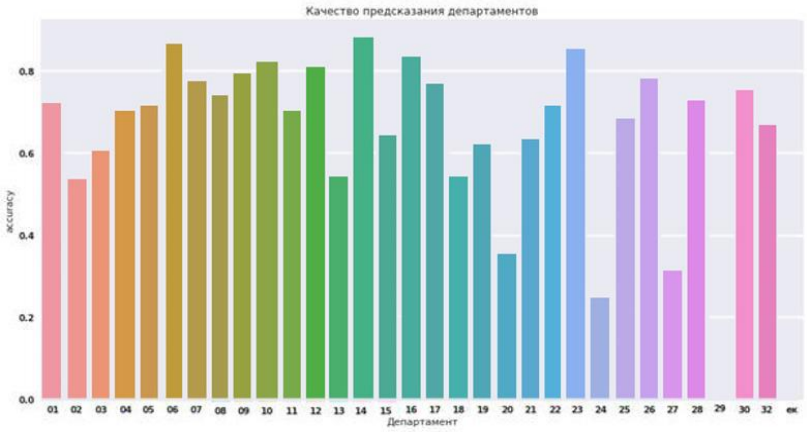

Рисунок 2.

За последние годы data scientists смогли достичь больших успехов в разработке алгоритмов, технологиях семантического анализа текста, что позволило решать задачи с достаточно высокой точностью. Затем появился электронный документооборот, и эксперты стали стремиться к повсеместному внедрению СЭД. Как следствие, компании накопили достаточное количество электронных документов для того, чтобы можно было найти общие закономерности данных, интересные зависимости и, наконец, применить алгоритмы МО на практике.

На сегодняшний день созданы практически все условия, чтобы исключить человека из процесса обработки документов, и полностью самостоятельный организм СЭД в тандеме с алгоритмами машинного обучения заработает. Однако принятие решений по-прежнему остается за человеком. Сколько бы алгоритмов ни подключили к обработке документа, они не смогут, например, принять важное управленческое решение или решение о сокращении бюджета. Кроме того, алгоритмы должны применяться к структурированным базам данных, тогда они смогут выполнять задачи с высокой точностью. Поэтому уже сейчас стоит приступать к проектированию такой модели данных путем применения алгоритмов машинного обучения и анализа данных. Без участия человека на первом этапе обучения алгоритмов не обойтись.

Литература

[1] http://www.tadviser.ru/index.php/Статья:Применение_машинного_обучения_в_СЭД:_от_теории_к_практике#.D0.9F.D1.80.D0.B8.D0.BA.D0.BB.D0.B0.D0.B4.D0.BD.D1.8B.D0.B5_.D0.B7.D0.B0.D0.B4.D0.B0.D1.87.D0.B8_.D0.BC.D0.B0.D1.88.D0.B8.D0.BD.D0.BD.D0.BE.D0.B3.D0.BE_.D0.BE.D0.B1.D1.83.D1.87.D0.B5.D0.BD.D0.B8.D1.8F

«SMART» DOCUMENT MANAGEMENT SYSTEM: ECM AND MACHINE LEARNING

K.K. MALYHIN
Master degree student

Abstract. The main goal was to analyze the possibility of integrating machine learning into the electronic document management system. As the analysis has shown, classification algorithms and semantic text analysis algorithms can now be applied to speed up the processing of a document, but the final word still remains with the human.

Keywords: document flow, machine learning, classification, tagging.