Московский Государственный Университет имени М. В. Ломоносова
Факультет Вычислительной Математики и Кибернетики
Кафедра Системного Программирования

Курсовая работа

Извлечение логической структуры из сканированных документов

Автор:
группа 328
Богатенкова Анастасия Олеговна

Научный руководитель:
Козлов Илья Сергеевич

Москва, 2020

Содержание

1 Введение

Большое количество текстовой информации представлено в виде pdf-документов, причем эти документы могут представлять собой сканированные копии других документов и у них может отсутствовать текстовый слой. При этом размер документов может быть очень большим. Зачастую требуется осуществлять поиск по содержимому таких документов и желательно осуществлять это более эффективным способом.

Как правило, документы имеют логическую структуру и имеют название, разбиение на главы, подглавы и т. д., содержат нумерованные и маркированные списки. Выделение такой логической структуры документа может помочь при решении задач автоматизированного анализа документов, а также при поиске по документам.

Применяется множество разнообразных подходов, которые позволяют выделять в тексте заголовки и распознавать логическую структуру документов. Для эффективного извлечения такой структуры может быть необходима метаинформация, такая как размер и тип шрифта, отступы, междустрочные интервалы и т. д. Поэтому извлечение логической структуры логично делать на этапе анализа сканированных документов.

2 Постановка задачи

Целью моей курсовой работы является разработка метода выделения логической структуры из документов в виде глав, подглав, элементов нумерованных и маркированных списков. Поставим задачу следующим образом: необходимо классифицировать каждую строчку документа как заголовок, элемент списка или текст и определить её уровень вложенности.

При решении задачи можно выделить следующие этапы ее выполнения:

1) Описание конкретной логической структуры, которую нужно выделить, т. е. разработка манифеста.

2) Разметка корпуса документов для обучения классификатора. Разметка проводится по правилам, указанным в манифесте.

3) Реализация метода и проведение экспериментальной проверки разработанного метода.

3 Обзор существующих решений

По извлечению структуры из документов существуют несколько подходов:

- на основе оглавления;
- на основе правил;
- на основе машинного обучения.

3.1 Извлечение структуры из документов на основе оглавления и правил

По анализу документов проводится очень много соревнований на ICDAR. В одном из таких соревнований [1] производилось извлечение структуры из книг, содержимое которых было получено с помощью оптического распознавания символов. Структура книг в виде разбиения на страницы, параграфы, главы извлекалась с использованием оглавления, которое присутствовало в большинстве книг.

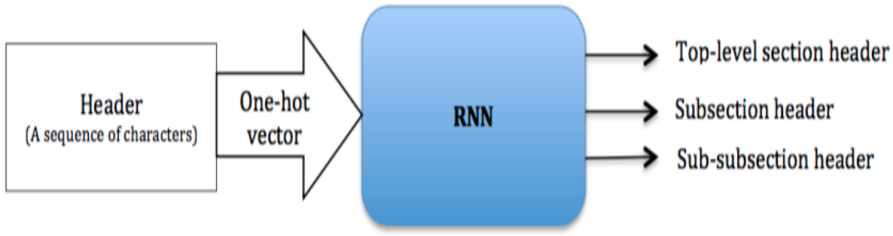
В 2019 году проводились соревнования FinTOC [2], где из финансовых документов извлекалась структура в виде иерархии уровней заголовков документов. Максимальная глубина уровней равна пяти. Одна из команд-участниц [3] извлекала необходимую структуру используя оглавление документов, а также систему правил, которые применялись для определения иерархии заголовков. Сначала идентифицировались страницы, содержащие текст оглавления, затем в документе находились страницы, соответствующие заголовкам, указанным в оглавлении. Последним шагом являлось выделение иерархии найденных заголовков, основанное на применении правил: анализировались такие признаки, как междустрочный интервал, отступ, шрифт, символы нумерации. Использованный подход позволил получить достаточно высокую точность, но низкую полноту, так как некоторые оглавления документов были неполными.

Извлечение структуры документов на основе оглавления имеет ряд недостатков. Во-первых, невозможно обрабатывать документы, в которых нет оглавления. Во-вторых, при использовании этого метода в структуру документа не будут включаться заголовки, которые не вошли в оглавление, например заголовки более низкого уровня. В-третьих, для нашей задачи необходимо извлекать элементы маркированных и нумерованных списков, которые не включаются в оглавление документа.

3.2 Извлечение структуры из документов на основе машинного обучения

В соревнованиях [2] кроме извлечения иерархической структуры документов решалась задача определения, является ли конкретный блок документа заголовком. Командам был дан набор pdf-документов, xml-файлов с выделенными блоками документов, а так же набор признаков для каждого блока: является ли шрифт блока жирным, курсивом, состоит ли текст из заглавных букв, начинается с заглавной буквы или с нумерации. Кроме данных признаков, каждая из команд использовала различные дополнительные морфологические, семантические, лингвистические признаки. На основе этих признаков обучались различные классификаторы: SVM, MNB, Extra Tree, Decision Tree, Gradient Boosting. Для оценки результатов использовалась F-мера, максимальный score в соревновании – 0,982.

В статье [4] 2017 года структура документа извлекалась с использованием методов машинного обучения, включая глубокое обучение. Цель данной работы – автоматически идентифицировать и классифицировать различные секции документов и понять их смысл в рамках документа (назначить семантическую метку). В рамках моей задачи интересна классификация секций документа.



Классификатор, который был использован при решении задачи, состоит из нескольких частей. Сначала строки документа подаются на вход классификатору (классификатор строк), который определяет, является ли строка заголовоком, затем строки-заголовки классифицируются точнее другими классификаторами (классификаторы секций). В этом решении структура документа имела вложенность 3, то есть предполагалось выделение секций, подсекций, подподсекций. Кроме того, в данной работе был размечен датасет, на котором происходило обучение модели. Метрика качества - F1-мера, при идентификации заголовков итоговый score – 0,96; при классификации секций средний F1-score – 0,81.

Существующие решения использовать сложно, так как документы, с которыми необходимо работать, могут содержать заголовки или элементы нумерованных списков глубокой вложенности, например 1.1.1.1 (уровней вложенности может быть и больше), а внутри этих уровней могут располагаться нумерованные/маркированные списки, также вложенные.

4 Описание практической части

5 Заключение

Список литературы

[1] AntoineDoucet, GabriellaKazai, SebastianColutto, Günter Mühlberger. Icdar 2013 competition on book structure extraction. In Document Analysis and Recognition (ICDAR), 2013 12th International Conference on, pages 1438–1443. IEEE, 2013.

[2] Rémi Juge, Najah-Imane Bentabet, Sira Ferradans. FinTOC-2019 Shared Task: Finding Title in Text Blocks.

[3] Gael Lejeune Emmanuel Giguet. Daniel fintoc-2019 shared task: Toc extraction and title detection. In The Second Workshop on Financial Narrative Processing of NoDalida 2019, 2019.

[4] Muhammad Mahbubur Rahman, Tim Finin. Deep nderstanding of a Documents Structure. In 4th IEEE/ACM International Conference on Big Data Computing, Applications and Technologies, 2017.