УДК 004.912+004.021+004.023

ИЗВЛЕЧЕНИЕ ЗАГОЛОВКОВ ИЗ PDF-ДОКУМЕНТОВ НАУЧНОЙ ТЕМАТИКИ

Д. С. Филиппов

Высшая школа информационных технологий и интеллектуальных систем Казанского (Приволжского) федерального университета
dmitriyfil1995@gmail.com

Аннотация

Актуальность представленного исследования обусловлена бедностью существующих подходов к извлечению заголовков из PDF-документов, предложенных в более ранних исследованиях, которые используют либо машинное обучение, либо простые эвристики. Цель настоящего исследования – предоставить более проработанные подходы к общей задаче извлечения заголовка документа и предложить лучший алгоритм выделения его из документов научной тематики. Основная методика, использованная нами при выборе решения, – рассмотреть, как можно большее количество различных ситуаций относительно форматирования заголовка, возникающих в разных документах, и предложить решение для каждой из них, а затем обобщить их в полноценный подход. Результаты выбранного подхода показали его эффективность по сравнению с методами других исследователей, если в нашем распоряжении находятся документы с различными вариациями оформления, структурной организации и форматирования. Данное исследование показало, что глубокое исследование задачи – перспективный путь для разработки лучших решений и инструментов. Статья будет полезна исследователям и разработчикам, которые часто встречаются с проблемой извлечения заголовков как одной из подзадач анализа документов.

Ключевые слова*: Pdf processing, title extraction, header extraction, strategy based approach, title heuristic, structural analysis, style information, text analysis, document analysis, information extraction, анализ текстов, автоматическая обработка документов, структурирование информации.*

ВВЕДЕНИЕ

В последние несколько лет активно развивается тенденция к структурированию информации, основную часть которой составляют текстовые документы. Так как подавляющая часть этих документов находится в формате PDF (Portable Document Format), то задача их обработки и анализа является как никогда актуальной. В число распространённых подзадач обработки входит проблема автоматического формирования библиотечных записей (метаданных), например, применительно к электронным библиотекам. Библиотечные записи вмещают в себя такую информацию, как авторы публикации, год, место издания, тип издания и, в частности, заголовок. Структура и содержание библиотечной записи, её взаимосвязь с информацией на титульных страницах публикации являются весьма интересной темой и заслуживают отдельного исследования.

Извлечение заголовка является также необходимым компонентом выделения и анализа структуры документа. В частности, комплексный структурный анализ документа может включать в себя: выделение заголовка, структуризацию оглавления, суммаризацию и структурирование контента с целью извлечения сущностей определённых типов. Например, моделирование и формализация публикаций в области математики, описанные в [7–9], могут применяться в сочетании с выделением заголовка для составления картотеки обработанных публикаций, формирования метаданных, при составлении тренировочных и тестовых корпусов и для иных целей. Активно применяется и обобщённый анализ научных публикаций различной тематики с опорой на корпуса статей известных журналов и конференций, таких, как ACM, IEEE и другие [10–15].

Однако даже задача выделения заголовка является не такой простой, как может показаться на первый взгляд. Имеется множество проблем, которые не только затрудняют анализ PDF-документов в целом, но и усложняют задачу анализа и выделения заголовков. Среди них, например, отсутствие текстового слоя, пользовательская кодировка, разнообразное и иногда весьма специфическое оформление титульных страниц, расположение и форматирование заголовка и другие. Поэтому данная работа посвящена извлечению заголовков из PDF-документов, а также возникающим при этом проблемам и возможными подходами к их разрешению. Для некоторых из проблем, освещённых в настоящей

работе, варианта решения предложено не будет, причиной чему являются технологические ограничения, для преодоления которых требуются отдельные исследования и разработки. В данной статье рассмотрены различные случаи, касающиеся специфического оформления и структурной организации заголовка.

Для извлечения заголовков из документов исследователи предлагают ряд методов. Большинство из них основано на одних и тех же принципах [1–2], [4] или близких техниках [3, 4], [10], [15, 16]. Многие предлагают готовые инструменты в виде приложений или веб-сервисов для множества задач, включая извлечение заголовка [2–6], [10], [14–16].

Большинство предложенных методов опирается на одну единственную эвристику: заголовок расположен на первой странице в блоке текста с самым большим шрифтом. Если таких блоков несколько, выбирается наидлиннейший. Другие же предлагают использовать алгоритмы машинного обучения. Серьёзный недостаток таких подходов состоит в том, что даже небольшие вариации в формате или структурной организации документа с высокой вероятностью приведут к неправильным результатам, извлечению только части заголовка, или же вовсе алгоритмы завершат свою работу неудачно. Нами предложен способ извлечения, преимущественно опирающийся на стилевую информацию текста документа, ищущий заголовки не только на первой странице и, кроме того, более устойчивый к флуктуациям форматирования заголовка. Необходимо отметить, что данный метод применим только к документам научной тематики, множество которых будет рассмотрено ниже в секции «Описание целевых документов», так как другие типы документов могут иметь серьёзные отличия в стилевом оформлении и структуре титульных страниц. Для остальных документов предложен механизм выбора стратегий извлечения заголовков на основе их жанров и типовых особенностей. Также даны рекомендации для некоторых из стратегий. Кроме того, обсуждены особенности обработки документов в других форматах (офисных, HTML и других).

Таким образом, основная цель настоящего исследования – предложить набор более качественных эвристик для извлечения заголовка из PDF-документов научной тематики, таких, как методички, учебники, научно-популярная литература и др., потому что методов, предложенных более ранни-

ми исследованиями, недостаточно для успешной обработки из Всемирной паутины научных книг в формате PDF.

Дальнейшее изложение организовано следующим образом: вначале идут описание и обоснование подхода, основанного на стратегиях, далее – более подробное описание свойств целевых документов, к которым будет применяться основной алгоритм извлечения. Третья секция перечисляет основные технические проблемы, возникающие при обработке и, в частности, извлечении заголовка из PDF-документов. Четвёртая секция предлагает возможные решения для этого, а также подробно описывает метод выделения заголовков из документов научной тематики. Последние секции показывают результаты описываемого алгоритма на тестовом множестве PDF-документов и сравнивают их с результатами, полученными с использованием методик предыдущих исследований. Также в последней секции сделаны замечания о других возможных стратегиях извлечения заголовка из документов в других форматах.

ОБЩИЙ ПОДХОД К ИЗВЛЕЧЕНИЮ

В различных по стилю и тематике документах титульные страницы и заголовки, в частности, оформляются по-разному. Отличия порой настолько велики, что говорить о методе выделения заголовков, общем для всех форматов и типов документов, не приходится. Даже в пределах одной тематики различия в оформлении могут быть критическими. Для решения рассматриваемой задачи в общем виде можно было бы использовать машинное обучение, как это сделано в [6]. Но тогда на выходе получится «тяжёлая» модель, которую будет трудно изменять и сопровождать. При этом каждое изменение повлечёт за собой переобучение модели с возможными регрессиями. Поэтому оптимальным для этой задачи (как и для многих других в области анализа неструктурированных документов) представляется подход, основанный на стратегиях.

В данном контексте будем понимать стратегию как отдельный алгоритм, рассчитанный на успешное выполнение задачи (выдачу правильного результата) для некоторого набора частных случаев. В контексте настоящей статьи частный случай определяется набором следующих параметров: стилистика документа (научная, деловая, художественная, публицистическая), его тип (статья, исследование, учебник, книга, собрание сочинений и другие), свойства оформления,

формат документа (DOCX, PDF, rich text formats, plain text, HTML). Тогда при обнаружении нового частного случая все старые стратегии остаются без изменений, и всё, что нужно, – добавить новую стратегию, учитывающую этот частный случай.

Таким образом, общая задача выделения заголовка (вне зависимости от формата) сводится к следующему: определяется и реализуется набор стратегий для известных частных случаев. Документ обрабатывается каждой из этих стратегий, пока не будет получен ненулевой результат (нулевой результат – значит, заголовок не найден). Возможна также вариация: вначале документ проходит предобработку с целью выделить лишь некоторое подмножество (в идеале одну) стратегий обработки, чтобы минимизировать риск получения лишних, неправильных результатов из других стратегий.

Вышеописанный подход и применен ниже, а именно, определена стратегия для извлечения заголовков из документов, удовлетворяющих условиям, описанным в следующем разделе.

ОПИСАНИЕ ЦЕЛЕВЫХ ДОКУМЕНТОВ

Как уже было сказано ранее, для анализа будут браться только документы научной тематики в формате PDF. Помимо этого, такие документы должны обладать следующими свойствами: иметь одну или несколько выделенных титульных страниц, где размещается заголовок, который, в свою очередь, будет иметь отличительные признаки, по которым его можно однозначно идентифицировать как заголовок. Допустимы документы на английском и русском языках.

ОПИСАНИЕ ПРОБЛЕМНЫХ СИТУАЦИЙ

Теперь, когда множество целевых документов описано, можно подробно рассмотреть стратегию выделения заголовка. Во многих исследованиях, например, [2–4], авторы опираются на единственную эвристику – заголовком является текстовый блок с самым крупным шрифтом на первой странице или, если таких несколько, самый крупный из них. Однако есть ряд существенных и довольно распространённых отклонений от этого простого правила, которые будут рассмотрены в настоящем разделе.

Первая категория проблем связана с форматом. Формат PDF изначально не был предназначен для автоматической обработки. Хотя его последние версии

поддерживают включение метаданных, облегчающих автоматическую обработку, подавляющая часть документов не содержит этих данных, и приходится прилагать существенные усилия, чтобы точно выделить необходимую информацию. Кроме того, есть следующие факторы, препятствующие успешной обработке таких документов:

- Текст, особенно на титульных страницах, может быть представлен в виде иллюстрации, и тогда единственный способ считать его – это OCR (Optical Character Recognition). Документы, для которых данная особенность актуальна, не будут в полной мере учитываться в этой статье. Однако при наличии OCR-технологии, позволяющей извлекать не только чистый текст, но и стилевую информацию, с ограничениями, допустимыми для успешного выполнения задачи, можно также обрабатывать подобные документы, используя алгоритм, описанный ниже.

Следует отметить, что некоторые из документов с титульной страницей, оформленной в виде обложки, можно обработать и без OCR-инструментов. В данном случае необходимо использовать информацию, считываемую с других титульных страниц, о чём будет сказано ниже.

- PDF-документ имеет внутреннюю кодировку, и считывание текста в первоначальном виде невозможно. Такой документ не может быть обработан ввиду отсутствия информации об этой внутренней кодировке в самом документе.
- На титульных страницах есть рисунки с текстовыми подписями, вносящие искажения и лишний текст.
- Искажения при считывании текста, связанные с неточной шириной и расстановкой пробелов, из-за чего одно слово может быть разделено на несколько слов или, наоборот, несколько слов могут быть склеены в одно. Возможны и смешанные случаи. Иногда также встречается ситуация, когда строка разделена по буквам, что характерно в большей степени для многострочных заголовков, состоящих из коротких слов.

Вторая проблема состоит в том, что заголовок не обязательно будет на первой странице. Фактически документы можно разделить на четыре категории: с одной, двумя, тремя титульными страницами и когда у документа есть не-

сколько страниц с вступительным текстом, а уже затем идёт заголовок. Часто первая страница оформлена, как обложка, и текст из неё выделить трудно или практически невозможно ввиду проблем из первой категории, описанной выше. Кроме того, возможны документы, у которых на разных страницах указан различный заголовок (например, в сокращённом варианте на первой странице и в полном варианте – на следующей).

Следует помнить, что указанные категории весьма условны и будут содержать самые разные вариации от документа к документу. К примеру, на иллюстрации 1 приведён документ с двумя заголовками на разных языках, расположенных на одной странице.

Иллюстрация 1. Пример документа с двумя заголовками на одной странице (одна из многочисленных вариаций титульных страниц


БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИНФОРМАТИКИ И РАДИОЭЛЕКТРОНИКИ

АДМИНИСТРАЦИЯ ПАРКА ВЫСОКИХ ТЕХНОЛОГИЙ

Открытые семантические технологии проектирования интеллектуальных систем

OSTIS-2012

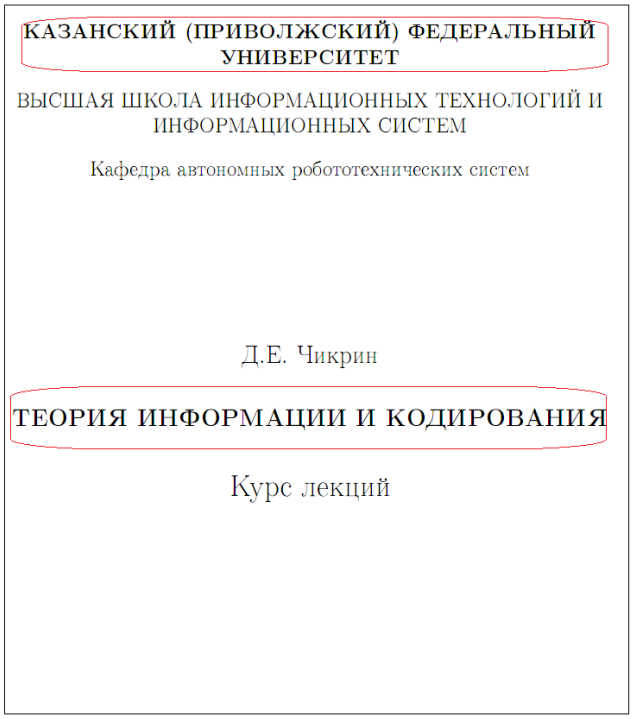
Open Semantic Technologies for Intelligent Systems

МАТЕРИАЛЫ
II МЕЖДУНАРОДНОЙ
НАУЧНО-ТЕХНИЧЕСКОЙ КОНФЕРЕНЦИИ

Следующая трудность в выделении заголовков – наличие шапок, в основном, с названиями учреждений, к которым относится данный труд. Проблема с ними – в том, что они могут иметь такой же по величине шрифт, что и заголовок,

однако почти во всех случаях текстовый блок шапки будет больше блока заголовка по площади. Значит, алгоритм отдаст приоритет шапке (см. иллюстрацию 2). В примере из приложения у заголовка и шапки одинаковый шрифт, как по гарнитуре, так и по стилю и размеру. При этом алгоритм на простой эвристике отдаст предпочтение шапке, как большей по площади.

Иллюстрация 2. Демонстрация коллизий шапок и основного заголовка документа



Следующая проблема возникает в случае, когда заголовок состоит из нескольких частей с разными шрифтами, как в примерах титульных страниц документов, изображённых на иллюстрациях 3 и 4. Пренебрегать второй частью с более мелким шрифтом нельзя, так как выделенный заголовок будет неполным.

Иллюстрация 3. Пример заголовка из двух частей с разным форматированием №1

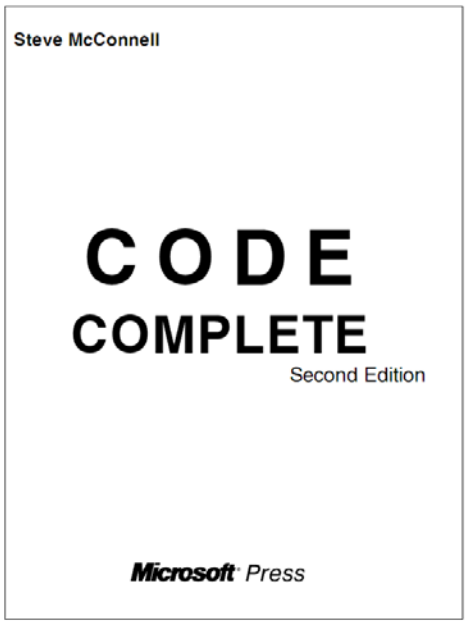
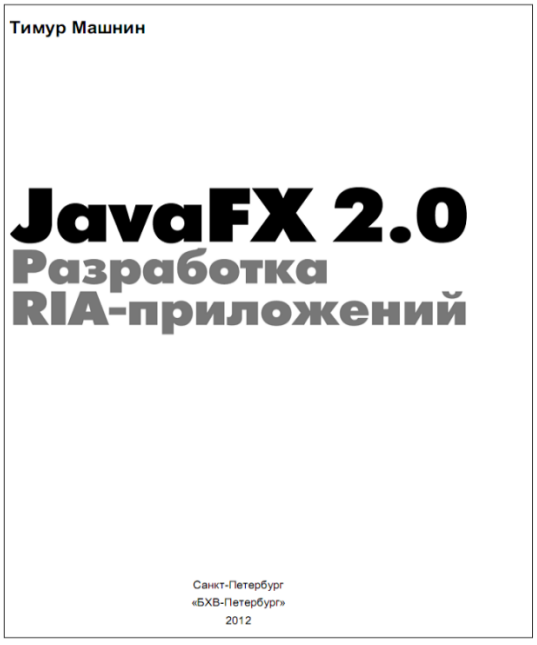


Иллюстрация 4. Пример заголовка из двух частей с разным форматированием №2

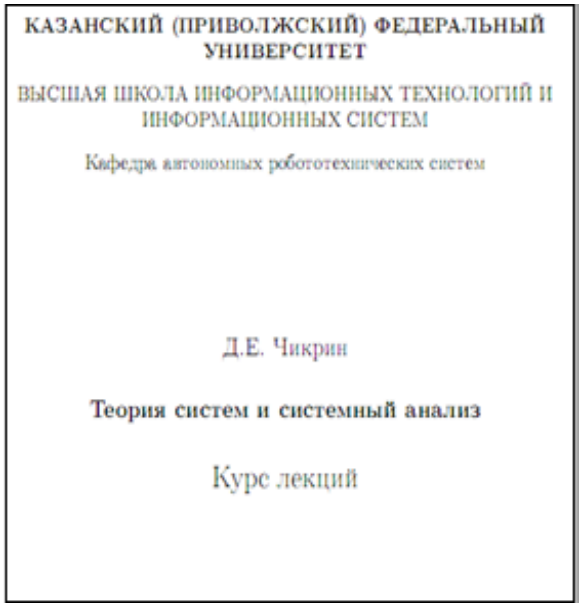


Наконец, ещё одна, самая коварная трудность при выделении заголовка – это то, что заголовки могут не иметь самый большой шрифт.

Рассмотрим пример на иллюстрации 5. Очевидно, заголовок здесь «Теория систем и системный анализ», однако самые большие шрифты здесь – у ша-

пок с названием учреждения и строки «Курс лекций». При этом алгоритм с учётом всех предыдущих проблем отсеет шапки, но разметит «Курс лекций» как заголовок.

Иллюстрация 5. Пример документа с фальш-заголовком


КАЗАНСКИЙ (ПРИВОЛЖСКИЙ) ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ

ВЫСШАЯ ШКОЛА ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ И ИНФОРМАЦИОННЫХ СИСТЕМ

Кафедра автономных робототехнических систем

Д.Е. Чикрин

Теория систем и системный анализ

Курс лекций

Однако можно заметить некоторые особенности, в частности, то, что настоящий заголовок выделен жирным шрифтом. Возникает вопрос об учёте дополнительной подкатегории заголовков, учитывающих эти особенности.

МЕТОДЫ

В предыдущем разделе были рассмотрены некоторые распространённые группы проблем, возникающих при выделении заголовков даже из документов весьма узкой категории. Ниже предлагаются решения некоторых из этих проблем.

Выделение текстов из иллюстраций и исследование внутренних кодировок PDF-документов находятся за рамками этой статьи, однако другие проблемы с форматом можно частично обойти или решить. Ниже предложены возможные варианты:

- Для учёта рисунков с текстовыми подписями рекомендуется предусмотреть наборы условий и параметров, согласно которым текстовый блок

можно классифицировать как подпись (или иного рода запись, не несущую смысла) и исключить из рассмотрения.

- Для исправления искажений при считывании текста, связанных с неточной шириной и расстановкой пробелов, можно предусмотреть словарь «несуществующих» сочетаний; если последовательность символов встречается в этом словаре, значит, это часть слова, которую надо объединить со следующей либо предыдущей частью. В случае склеивания слов нужно смотреть на расстояния между символами в исходном документе. Часто фактические отступы между словами в документе больше отступов между символами, и, выбрав или иным образом определив граничное значение отступа, можно классифицировать последние как символьные или словарные.

Для решения второй проблемы – наличия нескольких титульных страниц и неизвестного местонахождения верного заголовка – необходимо детектировать по возможности все титульные страницы и пытаться выделить заголовок из каждой. В своём программном решении мы использовали отношение суммарной площади символов к площади страницы, отбирая титульные страницы по экспериментально подобранному пороговому значению. При этом при наличии минимум двух титульных страниц, значит, и двух заголовков встаёт проблема выбора одного из них. В идеале алгоритм должен выделять со всех страниц одинаковый заголовок, но гарантировать такой результат нельзя. Так как чаще всего встречаются документы первых трёх категорий (то есть с одно, двумя или тремя титульными страницами), а первая страница обычно оформлена как обложка, из которой трудно выделить текст, то авторы в своём программном решении за некоторыми исключениями отдавали предпочтение заголовку с последней титульной страницы.

Чтобы отсеять шапки из документов, можно определить словарь типичных слов, встречающихся в них: типы учебных учреждений, научных министерств и т. д., и проверять их наличие в текстовых блоках на этапе выделения заголовка либо применить машинное обучение.

Для решения проблемы с заголовком, состоящим из нескольких частей с разным форматированием, потребовалось ввести новое понятие: ***добавочный заголовок (small title)*** – часть заголовка, меньшая его основной части по размеру шрифта (также может отличаться по цвету) и находящаяся рядом с ним. Для

того чтобы понять, как выделять добавочный заголовок, нужно знать отличительные признаки этой сущности. Нами выделены следующие признаки добавочного заголовка:

- шрифт является наибольшим на странице за исключением шрифта основного заголовка;
- шрифт добавочного заголовка совпадает со шрифтом заголовка по стилю, гарнитуре и регистру (uppercase, lowercase or mixed-case);
- находится рядом с заголовком (т. е. расстояние до блока заголовка – наименьшее среди всех расстояний до других блоков);
- комплекс из заголовка и добавочного заголовка значительно отдалён от других текстовых блоков страницы;
- может находиться как сверху, так и снизу основного заголовка;
- расстояние от заголовка до добавочного заголовка не должно превышать 1,5 * max (vertSpacing); если у обоих 0 (оба однострочные), то может быть любым;
- шрифт добавочного заголовка предполагается не более чем в 1,5 раза меньшим шрифта основного заголовка.

Учитывая вышеперечисленные признаки, схема учёта добавочного заголовка при выделении заголовка следующая: так как добавочного заголовок обязан находиться рядом с основным заголовком, то достаточно знать три расстояния – расстояние между заголовком и предполагаемым добавочным заголовком, расстояние от заголовка до ближайшего блока с другой стороны и расстояние от добавочного заголовка до ближайшего блока с противоположной стороны. Если первое расстояние значительно меньше двух последних (хотя бы в 1,5 раза), совпадают гарнитура и стиль шрифта с аналогичными у заголовка, и добавочный заголовок имеет второй наибольший шрифт среди блоков, то этого кандидата можно считать добавочным заголовком и включить его в основную часть.

Наконец, для решения последней из проблем, описанных в предыдущей секции, в данном исследовании введено новое понятие: ***фальшивый заголовок*** – текстовый блок, не подходящий под заголовок по признаку размера шрифта, но имеющий отличительные особенности, которые обычно присущи заголовкам.

Нельзя отрицать, что фальшивый заголовок может состоять более чем из одного блока.

Как и для решения предыдущей подзадачи, необходимо определить свойства фальшивого заголовка:

- у него шрифт не более, чем в 1.5 раза, меньше, чем у основных кандидатов;
- фальш-заголовок оформлен жирным стилем, а основные кандидаты нет и/или фальш-заголовок оформлен uppercase-регистром, а основные кандидаты нет.

Значит, для обнаружения фальшивого заголовка требуется проверять все строки, не попадающие под понятие чистого (primary) заголовка по размеру шрифта, но обладающего вышеперечисленными свойствами. При этом ни один из кандидатов на заголовок не должен быть выделен жирным шрифтом или заглавными буквами, иначе нарушатся эти условия-свойства. Необходимо отметить, что ни один текстовый блок не может являться добавочным заголовком и фальшивым заголовком одновременно. Это легко доказать, сопоставляя свойства-признаки каждого из типов нестандартных заголовков, приведённых выше.

С учётом описанных ранее проблем общая стратегия выделения заголовка из целевых документов научной тематики выглядит следующим образом:

- выделить титульные страницы; вначале проверить первые три страницы: если ни одна из них не титульная, то просматривать весь документ постранично до тех пор, пока не будет найдена первая титульная страница;
- выбрать самый большой по длине из кандидатов на заголовок, отобранных с учётом шапок, добавочных заголовков и фальшивых заголовков.

Если выделено больше одного заголовка, то выбрать один из них.

РЕЗУЛЬТАТЫ

Для тестирования описанного выше алгоритма были взяты 60 PDF-документов научной тематики, удовлетворяющих условиям, приведённым в главе «Описание целевых документов». Для сравнения были также рассмотрены следующие алгоритмы и инструменты: простая эвристики на основании блока с самым крупным шрифтом на первой странице, Docear’s PDF Inspector из [2] и метод на основе машинного обучения из [6]. Первичный запуск всех перечисленных инструментов и методов дал следующие результаты:

Таблица 1. Первичные результаты тестов

Наименование инструмента/ алгоритма	**Общее число документов**	**Число успешно обработанных документов (правильных заголовков)**	**Точность (%)**
Простая эвристика	32	12	37,5
Docear’s PDF Inspector	32	12	37,5
Подход, основанный на машинном обучении	32	20	62,5
Метод, изложенный в данной статье	32	28	87,5

Как видно, у техник, основанных на простых эвристиках (первые два подхода), резко снижается точность, что вызвано наличием исключительных случаев, разобранных в главе «Методы». Однако точность может быть искажена ещё одной группой факторов: искажениями, связанными с пробелами, подписями к рисункам и другими проблемами обработки формата. При этом заголовок может быть распознан правильно, но из-за подобных искажений может не совпасть с ожидаемым результатом. Чтобы исключить влияние неверной расстановки пробелов в заголовках, будем убирать все пробельные символы из них (например, вместо “Genetic algorithm: theory and practice” получится “geneticalgorithm:theoryandpractice”). Таким образом, результаты после удаления пробелов следующие:

Таблица 2. Результаты тестов с применением нормализации

Наименование инструмента/ алгоритма	**Общее число документов**	**Число успешно обработанных документов (правильных заголовков)**	**Точность (%)**
Простая эвристика	32	18	56.25
Docear’s PDF Inspector	32	17	53.13
Подход, основанный на машинном обучении	32	24	75.0
Метод, изложенный в данной статье	32	31	96.88

ОБСУЖДЕНИЕ

Как видно из таблицы 2, влияние неверно распознанных пробелов существенно сказывается на результатах обработки документов. Однако метод, изложенный в статье, показывает существенно лучший результат. Это связано с тем, что простая эвристика и Docear's PDF Inspector не учитывают исключительных случаев, встречающихся в документах. Кроме того, простая эвристика не учитывает того факта, что заголовок не обязательно может быть выделен на первой странице. Подход же, основанный на машинном обучении, показывает довольно высокие результаты на широком круге документов. Однако он не всегда верно ведёт себя в случаях с фальшивыми или составными заголовками.

Следует также отметить, что выделение заголовков человеком отчасти является творческой задачей, и результаты, полученные разными людьми, могут не сойтись, значит, могут быть оспорены и результаты, показанные выше. Это связано с разнообразием информации, представленной на титульных страницах. Помимо заголовков и метаданных на них размещены различные пояснения, дополнения и подзаголовки, которые одними могут включаться в состав заголовка всего документа, другими – нет. Вопрос о том, что же включать в общем случае в состав заголовка, даже в рамках отдельно взятой категории документов, требует серьёзного исследования, выходящего за рамки данной статьи.

ЗАКЛЮЧЕНИЕ

Рассмотрены два основных принципа извлечения заголовков из PDF-документов: подход, основанный на стратегиях для выбора конкретного метода извлечения, и метод с использованием сложных эвристик для выделения заголовков из документов научной тематики. Оба метода были разработаны для улучшения качества выделения заголовка в общем случае, так как предыдущие исследования были ориентированы на простые методы либо машинное обучение. Однако алгоритмы машинного обучения имеют естественные пределы точности и полноты, которые не могут быть преодолены ни посредством выбора большего и лучшего тренировочного корпуса, ни какой-либо комбинацией готовых моделей допустимой точности. Поэтому можно заключить, что предлагаемые подходы имеют широкие перспективы в области автоматической структуризации документов и/или выделения отдельных структурных частей. А именно,

подход, основанный на стратегиях, позволит обрабатывать документы различных типов и жанров, а извлечение заголовков из научных документов описанным выше алгоритмом предоставляет лучшую точность относительно методов предыдущих исследователей.

Следует заметить, что подход, основанный на стратегиях, может быть также успешно использован и при формировании библиотечных записей, в частности, в журнале «Электронные библиотеки», позволяя индексировать не только книжные издания, но и статьи, отчёты и другую документацию. Также анализ стилевой информации и стратегии позволит углубиться в анализ титульных страниц в целом, например, для определения авторов, учреждения, которому принадлежит документ, издательства, типа труда и прочей информации.

Результаты, показанные выше, полностью оправдали ожидания авторов. Предложенный метод будет использован в дальнейших исследованиях (например, при извлечении заголовков секций или обнаружении оглавления документа) и уже может быть применён в различных задачах автоматизированной обработки документов (к примеру, в задачах суммаризации, классификации по темам и других), особенно научных книг и трудов.

СПИСОК ЛИТЕРАТУРЫ

1. *Lipinski M., Yao K., Breitinger C., Beel J., Gipp B.* Evaluation of Header Metadata Extraction Approaches and Tools for Scientific PDF Documents // 13th ACM/IEEE-CS Joint Conf. on Digital Libraries, Indianapolis, USA, 2013. ACM: 2013. P. 385–386.

2. *Beel J., Langer S., Genzmehr M., Müller M.* Docear's PDF Inspector: Title Extraction from PDF Files // 13th ACM/IEEE-CS Joint Conf. on Digital Libraries, Indianapolis, USA, 2013. ACM: 2013. P. 443–444.

3. *Marinai S.* Metadata Extraction from PDF Papers for Digital Library Ingest // 10th Int. Conf. on Document Analysis and Recognition (ICDAR). 2009. P. 251–255.

4. *Васильев А., Самусев С., Шамина О., Козлов Д.* Создание электронной библиотеки русскоязычных научных статей // Сб. работ участников конкурса науч. проектов по информ. поиску под ред. П. И. Браславского, Екатеринбург, Россия, 2007. Изд-во Урал. ун-та, 2007. С. 37–45.

5. *Beel J., Gipp B., Shaker A., Friedrich N.* SciPlore Xtract: Extracting Titles from Scientific PDF Documents by Analyzing Style Information (Font Size) // Research and Advanced Technology for Digital Libraries. 2010. P. 413–416.

6. *Hu Y., Li H., Cao Y., Teng L., Meyerzon D., Zheng Q.* Automatic extraction of titles from general documents using machine learning // 5th ACM/IEEE-CS Joint Conf. on Digital Libraries, New York, USA, 2005. ACM: 2005. P. 145–154.

7. *Elizarov A. M., Kirillovich A. V., Lipachev E. K., Nevzorova O. A., Solovyev V. D., Zhiltsov N. G.* Mathematical knowledge representation: semantic models and formalisms // Lobachevskii Journal of Mathematics. 2014. No 4. P. 348–354.

8. *Elizarov A. M., Lipachev E. K., Nevzorova O. A., Solovyev V. D.* Methods and means for semantic structuring of electronic mathematical documents // Doklady Mathematics. 2014. № 1. P. 521–524.

9. *Solovyev V. D., Zhiltsov N. G.* Logical Structure Analysis of Scientific Publications in Mathematics // Int. Conf. on Web Intelligence, Mining and Semantics, Sogndal, Norway, 2011. ACM: 2011, P. 21:1–21:9.

10. *Han H., Giles C.L., Manavoglu E., Zha H., Zhang Z., Fox E.A.* Automatic document metadata extraction using support vector machines // 3rd ACM/IEEE-CS Joint Conf. on Digital Libraries, Houston, USA, 2003. ACM: 2003. P. 37–48.

11. *Peng F., McCallum A.* Information Extraction from Research Papers Using Conditional Random Fields // Inf. Process. Manage. 2006. No 4. P. 963–979.

12. *Nakagawa K., Nomura A., Suzuki M.* Extraction of logical structure from articles in mathematics // Int. Conf. on Mathematical Knowledge Management, 2004. Springer: 2004. P. 276–289.

13. *Beel J., Gipp B., Langer S., Genzmehr M., Wilde E., Nürnberger A., Pitman J.* Introducing Mr. DLib, a Machine-readable Digital Library // 11th Annual Int. ACM/IEEE Joint Conf. on Digital Libraries, Ottawa, Ontario, Canada, 2011. ACM: 2011, P. 463–464.

14. *Granitzer M., Hristakeva M., Knight R. and Jack K.* A Comparison of Metadata Extraction Techniques for Crowdsourced Bibliographic Metadata Management // 27th Annual ACM Symposium on Applied Computing, Trento, Italy, 2012. ACM: 2012, P. 962–964.

15. *Yilmazel O., Finneran C. M., Liddy E. D.* MetaExtract: an NLP system to automatically assign metadata // 4th ACM/IEEE-CS Joint Conf. on Digital Libraries, Tuscon, USA, 2004. ACM: 2004. P. 241–242.

16. *Mayank S., Barnopriyo B., Priyank P., Manvi G., Sidhartha S.* OCR++: A Robust Framework For Information Extraction from Scholarly Articles // arXiv preprint arXiv:1609.06423. 2016. P. 1–9.

TITLE EXTRACTION FROM ENGLISH SCIENTIFIC BOOKS IN PDF FORMAT

D. S. Filippov

Higher School of Information Technologies and Intelligent Systems at Kazan (Volga region) Federal University

dmitriyfil1995@gmail.com

Abstract

Relevance of the issue under study is due to tenuity of methods proposed by other researchers that use simple heuristics or machine learning algorithms. The purpose of the article is to provide better way to extract titles from scientific PDF documents and offer better and more reasonable approach to title selection generally. The leading approach to the study is regard as many cases and problems appeared during extraction as possible and find an approach to solve all of them. The results showed the efficiency of chosen approach in case of having a document set with all of considered problems. The research highlights that deep analysis of current task problem is a perspective to make the best solutions and tools. The article may be useful for all researchers and developers who often encounter the problem of document structural analysis or title detection as secondary task of a main program workflow.

Keywords*: Pdf processing, title extraction, header extraction, strategy based approach, title heuristic, structural analysis, style information, text analysis, document analysis, information extraction*

REFERENCES

1. *Lipinski M., Yao K., Breitinger C., Beel J., Gipp B.* Evaluation of Header Metadata Extraction Approaches and Tools for Scientific PDF Documents // 13th ACM/IEEE-CS Joint Conf. on Digital Libraries, Indianapolis, USA, 2013. ACM: 2013. P. 385–386.

2. *Beel J., Langer S., Genzmehr M., Müller M.* Docear's PDF Inspector: Title Extraction from PDF Files // 13th ACM/IEEE-CS Joint Conf. on Digital Libraries, Indianapolis, USA, 2013. ACM: 2013. P. 443–444.

3. *Marinai S.* Metadata Extraction from PDF Papers for Digital Library Ingest // 10th Int. Conf. on Document Analysis and Recognition (ICDAR). 2009. P. 251–255.

4. *Vasilyev A., Samusev S., Shamina O., Kozlov D.* Digital library of Russian-language scientific articles creation // coll. of academic papers of information search competetion participants, edited by P. I. Braslavsky, Ekaterinburg, Russia, 2007. Publishing Office of Ural University, 2007. P. 37–45.

5. *Beel J., Gipp B., Shaker A., Friedrich N.* SciPlore Xtract: Extracting Titles from Scientific PDF Documents by Analyzing Style Information (Font Size) // Research and Advanced Technology for Digital Libraries. 2010. P. 413–416.

6. *Hu Y., Li H., Cao Y., Teng L., Meyerzon D., Zheng Q*. Automatic extraction of titles from general documents using machine learning // 5th ACM/IEEE-CS Joint Conf. on Digital Libraries, New York, USA, 2005. ACM: 2005. P. 145–154.

7. *Elizarov A. M., Kirillovich A. V., Lipachev E. K., Nevzorova O. A., Solovyev V. D., Zhiltsov N. G.* Mathematical knowledge representation: semantic models and formalisms // Lobachevskii Journal of Mathematics. 2014. No 4. P. 348–354.

8. *Elizarov A. M., Lipachev E. K., Nevzorova O. A., Solovyev V. D.* Methods and means for semantic structuring of electronic mathematical documents // Doklady Mathematics. 2014. No 1. P. 521–524.

9. *Solovyev V. D., Zhiltsov N. G.* Logical Structure Analysis of Scientific Publications in Mathematics // Int. Conf. on Web Intelligence, Mining and Semantics, Sogndal, Norway, 2011. ACM: 2011. P. 21:1–21:9.

10. *Han H., Giles C.L., Manavoglu E., Zha H., Zhang Z., Fox E. A.* Automatic document metadata extraction using support vector machines // 3rd ACM/IEEE-CS Joint Conf. on Digital Libraries, Houston, USA, 2003. ACM: 2003. P. 37–48.

11. *Peng F., McCallum A.* Information Extraction from Research Papers Using Conditional Random Fields // Inf. Process. Manag. 2006. No 4. P. 963–979.

12. *Nakagawa K., Nomura A., Suzuki M.* Extraction of logical structure from articles in mathematics // Int. Conf. on Mathematical Knowledge Management, 2004. Springer: 2004. P. 276–289.

13. *Beel J., Gipp B., Langer S., Genzmehr M., Wilde E., Nürnberger A., Pitman J.* Introducing Mr. DLib, a Machine-readable Digital Library // 11th Annual International ACM/IEEE Joint Conf. on Digital Libraries, Ottawa, Ontario, Canada, 2011. ACM: 2011. P. 463–464.

14. *Granitzer M., Hristakeva M., Knight R., Jack K.* A Comparison of Metadata Extraction Techniques for Crowdsourced Bibliographic Metadata Management // 27th Annual ACM Symposium on Applied Computing, Trento, Italy, 2012. ACM: 2012. P. 962–964.

15. *Yilmazel O., Finneran C. M., Liddy E. D.* MetaExtract: an NLP system to automatically assign metadata // 4th ACM/IEEE-CS Joint Conf. on Digital Libraries, Tuscon, USA, 2004. ACM: 2004. P. 241–242.

16. *Mayank S., Barnopriyo B., Priyank P., Manvi G., Sidhartha S.* OCR++: A Robust Framework For Information Extraction from Scholarly Articles // arXiv preprint arXiv:1609.06423. 2016. P. 1–9.

СВЕДЕНИЯ ОБ АВТОРЕ

ФИЛИППОВ Дмитрий Сергеевич – бакалавр Высшей школы информационных технологий и интеллектуальных систем Казанского (Приволжского) федерального университета, студент 1 курса магистратуры.

Dmitriy Sergeevich FILIPPOV – has bachelor's degree of the Higher School of Information Technologies and Intelligent Systems at Kazan (Volga region) Federal University, 1st year graduate student.

e-mail: dmitriyfil1995@gmail.com

Материал поступил в редакцию 4 июня 2018 года