НАЙДИН Олег Павлович -
ведущий программист
отдела исследования
компьютерных систем Российской
государственной библиотеки (РГБ)
Адрес: 119019, г. Москва,
ул. Воздвиженка, 3/5
e-mail: oleg.naydin@gmail.com

КОНТРОЛЬ КАЧЕСТВА СКАНИРОВАНИЯ ТЕКСТОВ

Постановка задачи

В настоящее время большое количество документов массово оцифровывается на специализированных книжных сканерах, причем происходит это настолько быстро, что человек не в состоянии отследить качество получаемых изображений в режиме реального времени. После сканирования большинство электронных версий изданий не проверяется вообще или просматривается вручную. Так, например, в Российской государственной библиотеке (РГБ) создан отдел технического контроля качества сканирования (ОТК). Работники этого отдела вручную просматривают все оцифрованные документы до их включения в состав фонда электронной библиотеки РГБ. Автоматизация оценки качества изображений отсканированных документов является открытой проблемой. Для ее решения будет предложен ряд алгоритмов, которые отсеивают максимальное количество «хороших» страниц, а остальные оставляют оператору для ручной проверки.

В данной статье рассматривался довольно узкий класс документов - диссертации на соискание ученой степени кандидата или доктора наук. Большинство диссертаций датированы 2010-2011 годами, т.е. имеют неплохое исходное качество, что позволяет сосредоточиться лишь на ошибках оцифровки.

Типичные ошибки

При сканировании документов возникает большое количество ошибок, часть из которых вызвана сбоями в работе оборудования, часть - человеческим фактором, а некоторые - недостаточным качеством оригинала. Типичные ошибки, возникающие при сканировании, можно разбить на три основные категории:

1) искажения изображений, диаграмм;

2) пропуск, повтор и перестановка страниц;

3) искажение текста (поворот, растяжение, сторонний шум...).

Ошибки первой группы субъективны: оценить качество изображения или диаграммы можно, лишь зная, что там должно быть и как это должно выглядеть, поэтому их необходимо оставить оператору для отдельной проверки в любом случае. Таким образом, отнесем все страницы с изображениями и диаграммами к «подозрительным». Отметим, что наличие подобных объектов на странице нарушает ее строчную структуру. Это понятие является основным в данной работе, примеры строчных структур и их нарушений будут представлены ниже. Также для простоты отнесем в этот класс все таблицы, потому что они нарушают строчную структуру текста.

Для определения ошибок второй категории мы должны провести семантический анализ страницы (распознать ее номер), что неизбежно влечет несколько проблем: во-первых, отсутствуют четкие требования к оформлению диссертаций, поэтому авторы могут располагать номера страниц, минимум, в шести различных местах; во-вторых, зачастую нумеруются далеко не все страницы[1], поэтому отсутствие номера на странице не является признаком ее «подозрительности»; и, в-третьих, часто нумерация правится от руки уже после печати, что делает выделение номера практически невозможным. В рамках данного исследования эта группа ошибок не исследовалась, она представляет широкую область для дальнейших изысканий.

Ошибки третьей категории влекут нарушение строчной струк-

[1] *Часто не нумеруются, например, страницы с изображениями и таблицами с поворотом на 90°.*

туры текста: строки искривлены, непараллельны сторонам листа или теряются из-за шума. Страницы с такими искажениями тоже отнесем к «подозрительным». Таким образом, в данном исследовании основной задачей было отделение страниц с четкой строчной структурой от остальных. Отметим также, что, если мы хотим опираться на строчную структуру, мы вынуждены отнести неполные страницы к «подозрительным» для более устойчивой классификации. Это допущение несильно увеличивает нагрузку на проверяющего, причем мы сможем отследить еще один вид брака - пустые страницы. Также к «подозрительным» были отнесены и титульные листы, их структура сильно отличается как от обычного текста, так и друг от друга.

Итак, в данном исследовании была произведена попытка разделить все страницы на «хорошие», имеющие ярко выраженную строчную структуру, и «подозрительные» - все остальные: содержащие изображения, таблицы, наклоны (см. **рис. 1а**), посторонние объекты (см. **рис. 1б**), титульные листы и неполные страницы.

Основные обозначения

Для постановки задачи оптимизации и построения классификатора введем несколько основных показателей, на которые мы будем обращать внимание. Исходя из предположения, что все «хорошие» страницы оператор просматривать не будет, мы хотим минимизировать число брака в них, поэтому рассмотрим *FPR (false positive rate)* - долю ошибочно положительных классификаций, т.е. отношение числа пропущенных бракованных страниц к общему числу страниц с дефектами. Таким образом, первая задача оптимизации запишется в следующем виде:

$$FPR \to 0 \qquad (1).$$

Также будем рассматривать долю «хороших» страниц среди ответов классификатора - $P_{positive}$. Эта величина показывает, какую долю выборки библиотекарю не придется просматривать вручную, или долю сэкономленного времени. Значит, вторая задача оптимизации запишется как:

$$P_{positive} \to max \qquad (2).$$

Очевидно, задачи (1) и (2) противоречат друг другу и оптимизировать оба критерия параллельно невозможно, поэтому перейдем к задаче условной минимизации: потребуем, чтобы $FPR \le \delta$, *где* δ - некоторое число, близкое к 0, и будем максимизировать $P_{positive}$. Таким образом, итоговая задача запишется следующим образом:

$$\begin{cases} P_{positive} \to max \\ FPR \le \delta \end{cases} \qquad (3).$$

Также будем обращать внимание на качество классификации, т.е. долю объектов, для которых класс указан верно, но этот показатель не будет основным.

Обзор алгоритмов классификации

Будем считать, что исходный текст в формате PDF уже разбит на изображения страниц и бинаризован[2], тогда мы можем работать с каждой страницей отдельно. Прежде чем говорить о классификации, рассмотрим один предварительный этап, от которого зависит итоговый результат - выделение признакового описания страницы. Процесс выделения признаков одинаков для всех алгоритмов классификации, представленных ниже в этой главе в разделе 3, проходит в несколько независимых этапов и будет прослежен на примере двух тестовых страниц (см. **рис. 2**).

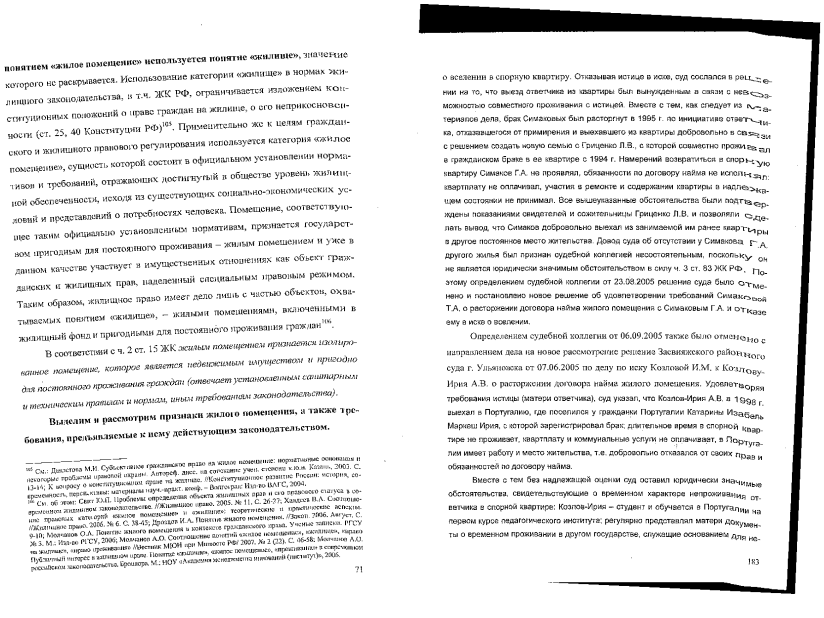
1) Предобработка изображений

Как указывалось выше, основным критерием оценивания страницы является наличие у нее ярко выраженной строчной структуры:

1) большинство строк начинается на одном расстоянии от левого края листа;

2) остальные строки имеют одинаковые отступы (красные строки, отступы списков);

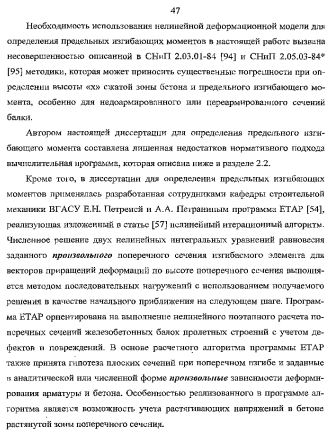
3) большинство строк заканчивается на одном расстоянии от правого края листа;



(а) Наклон (б) Посторонний объект

Рис. 1. *Примеры «подозрительных» страниц*

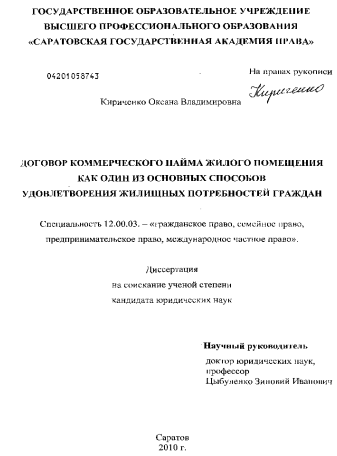
[2] *В данном исследовании для подготовки бинарных изображений использовался набор программного обеспечения Ghostscript [5].*


47

Необходимость использования нелинейной деформационной модели для определения предельных изгибающих моментов в настоящей работе вызвана несовершенностью описанной в СНиП 2.03.01-84 [94] и СНиП 2.05.03-84* [95] методики, которая может приносить существенные погрешности при определении высоты «x» сжатой зоны бетона и предельного изгибающего момента, особенно для недоармированного или переармированного сечений балок.

Автором настоящей диссертации для определения предельного изгибающего момента составлена лишенная недостатков нормативного подхода вычислительная программа, которая описана ниже в разделе 2.2.

Кроме того, в диссертации для определения предельных изгибающих моментов применялась разработанная сотрудниками кафедры строительной механики ВГАСУ В.Н. Петренёй и А.А. Петраншем программа ETAP [54], реализующая изложенный в статье [57] шаговый итерационный алгоритм. Численное решение двух нелинейных интегральных уравнений равновесия заданного *произвольного* поперечного сечения изгибаемого элемента для векторов приращений деформаций по высоте поперечного сечения выполняется методом последовательных нагружений с использованием получаемого решения в качестве начального приближения на следующем шаге. Программа ETAP ориентирована на выполнение нелинейного поэтапного расчета поперечных сечений железобетонных балок пролетных строений с учетом дефектов и повреждений. В основе расчетного алгоритма программы ETAP также принята гипотеза плоских сечений при поперечном изгибе и заданные в аналитической или числовой форме *произвольные* зависимости деформирования арматуры и бетона. Особенностью реализованного в программе алгоритма является возможность учета растягивающих напряжений в бетоне растянутой зоны поперечного сечения.


ГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
«САРАТОВСКАЯ ГОСУДАРСТВЕННАЯ АКАДЕМИЯ ПРАВА»

04201058743 — На правах рукописи

Кириченко Оксана Владимировна

ДОГОВОР КОММЕРЧЕСКОГО НАЙМА ЖИЛОГО ПОМЕЩЕНИЯ
КАК ОДИН ИЗ ОСНОВНЫХ СПОСОБОВ
УДОВЛЕТВОРЕНИЯ ЖИЛИЩНЫХ ПОТРЕБНОСТЕЙ ГРАЖДАН

Специальность 12.00.03. – гражданское право, семейное право, предпринимательское право, международное частное право.

Диссертация
на соискание ученой степени
кандидата юридических наук

Научный руководитель
доктор юридических наук,
профессор
Цыбуленко Зинаида Ивановна

Саратов
2010 г.

(а) Обычный текст (б) Титульный лист

Рис. 2. *Тестовые страницы*

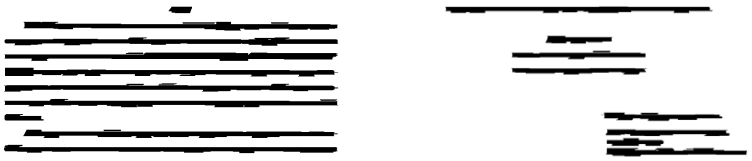
4) большинство строк имеет одинаковые ширину и высоту.

Эти и другие признаки необходимо выделить, чтобы обучить классификатор, поэтому сначала необходимо на изображении найти строки. Для этого использовалась горизонтальная дилатация[3], которая объединяет буквы в слова, а слова в строки (см. **рис. 3**[4]).

Затем производилось морфологическое открытие[5], которое убирает тонкие линии на изображении и делает строчную структуру еще более выраженной (см. **рис. 4**). Этот этап завершает предобработку изображений.

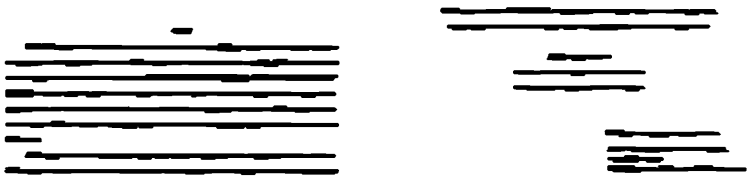
2) Вычисление признакового описания

Предыдущий этап выделил на изображении строки[6] и избавил их от шума. Теперь для каждой строки вычислим ряд признаков, на которых в дальнейшем будем обучать классификатор. В данном исследовании в качестве признаков вычислялись характеристики минимального прямоугольника, содержащего строку, стороны которого параллельны сторонам страницы.



(а) Обычный текст (б) Титульный лист

Рис. 3. *Тестовые страницы после горизонтальной дилатации*



(а) Обычный текст (б) Титульный лист

Рис. 4. *Тестовые страницы после морфологического открытия*

Имея граничные прямоугольники, для каждой строки, можно вычислить вектор из четырех характеристик: *[x y widht height]*, где *(x,y)* - координаты левого верхнего угла, *widht* - ширина, *height* - высота прямоугольника.

В связи с погрешностями дискретизации и вычислений даже на «хороших» страницах строки имеют немного разную высоту и отступ от левого края, поэтому построим гистограммы распределений каждого из четырех показателей на странице (см. **рис. 5**).

Есть несколько признаков, по которым можно отличить «хорошую» страницу от остальных:

1) обычный текст имеет два четких пика на гистограмме распределения начал строк;

2) все строки обычного текста имеют одинаковую высоту;

3) в обычном тексте строки отстоят на одинаковое расстояние друг от друга.

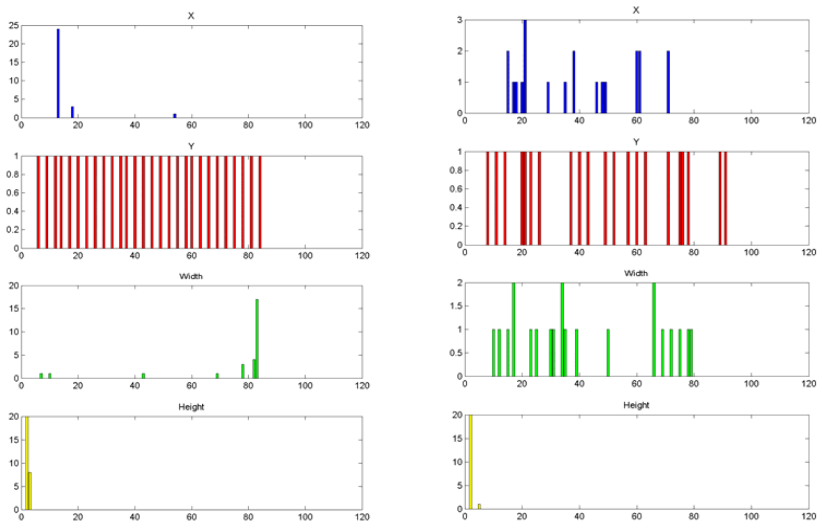
Таким образом, полученной с помощью этих гистограмм информации достаточно, чтобы произвести классификацию. Объединив четыре полученные гистограммы в один вектор, получим признаковое описание страницы, которое и будем использовать для классификации.

[3]Дилатация - морфологический прием обработки изображения, при котором каждый пиксель объекта «наращивается» согласно некоторому шаблону. В данном исследовании каждый пиксель «наращивался» в ширину.

[4]Здесь и далее используются фрагменты исходных изображений.

[5]Открытие - морфологический прием обработки изображения, который является последовательным применением эрозии (обратного процесса к дилатации) и дилатации. Эта операция удаляет небольшие объекты и сглаживает контуры остальных.

[6]В данном случае корректнее было бы говорить о компонентах связности на изображении, потому что помимо строк компонента связности может быть, например, изображением или шумовым объектом. Для краткости и простоты изложения здесь и далее будет использоваться термин «строка» вместо «компонента связности».



(а) Обычный текст (б) Титульный лист

Рис. 5. *Гистограммы распределения четырех признаков строк*

3) Обучение классификаторов

В этом разделе приведен краткий обзор двух принципиально различных алгоритмов обучения классификатора.

Комитет классификаторов

Основной идеей комитета классификаторов является следующее предположение: некоторая композиция базовых классификаторов может дать более сильный классификатор, чем каждый из базовых в отдельности. В данном исследовании в качестве базовых классификаторов были выбраны решающие деревья, а в качестве алгоритмов настройки комитета - два вида бустинга [1,2]: AdaBoost и Gentle Boost.

Данный метод имеет достаточно много параметров, часть из которых относится к базовым классификаторам (деревьям), остальные - к методу настройки композиции (бустингу). С одной стороны, хороший подбор параметров может значительно улучшить результаты классификации, но, с другой - одновременный подбор всех параметров занимает очень много времени и вычислительных ресурсов, поэтому практически невозможен.

Метрический классификатор

Метрические алгоритмы классификации основаны на более «прозрачной» идее: объект из тестовой выборки сравнивается с объектами из обучающей выборки по степени схожести и относится к тому классу, на объекты которого он больше всего похож. Для определения схожести элементов необходимо ввести некоторую мет-рику в пространстве признаковых описаний, в данном исследовании использовались Евклидова метрика[7] и EMD-метрика[8]. В качестве метрического классификатора использовался метод k-ближайших соседей [1].

Результаты исследования

В качестве платформы для тестирования был выбран компьютер со следующими характеристиками: процессор Intel® Core ™ i5-3570K 4.22GHz, оперативная память Corsair® Vengeance® 16GB DDR3 1600MHz. Тестирование производилось с использованием пакета прикладных программ MATLAB [4].

Для тестирования случайным образом были выбраны пять диссертаций, оценка качества классификатора оценивалась с помощью метода LOO[9]. Лучшие результаты из полученных для каждого из алгоритмов можно увидеть в **таблице 1**. Полужирным шрифтом в каждой колонке выделен лучший (или несколько лучших) результат по соответствующему критерию.

Таблица 1

Результаты классификации

Алгоритм обучения	FPR	$P_{positive}$	Качество классификации
AdaBoost + решающие деревья	**0.0572**	0.4546	0.8016
Gentle Boost + решающие деревья	0.0965	**0.6034**	**0.9125**
k-ближайших соседей + Евклидова метрика	0.0848	0.4948	0.8140
k-ближайших соседей + EMD-метрика	0.0758	0.4472	0.7819

Важным критерием работы алгоритма является время его работы. В **таблице 2** представлены как время обучения модели, так и время классификации с использованием полученной модели. Для

[7] *Евклидова метрика - геометрическое расстояние между двумя точками в многомерном пространстве, вычисляемое по теореме Пифагора.*

[8] *EMD (earth mover's distance) - метрика, введенная для сравнения гистограмм. Подробное описание можно найти в [3].*

[9] *LOO (leave-one-out) - контроль по отдельным объектам, т.е. обучение производится на всех документах, кроме одного, причем отложенный документ каждый раз изменяется. Полученные результаты усредняются.*

Таблица 2

Время работы алгоритмов

Алгоритм обучения	Время обучения (сек)	Время классификации (сек)
AdaBoost + решающие деревья	45.607627	4.581650
Gentle Boost + решающие деревья	35.847201	4.418844
k-ближайших соседей + Евклидова метрика	-	**0.133203**
k-ближайших соседей + EMD-метрика	-	**0.053356**

обучения использовались 1011 изображений страниц (именно столько страниц было в отобранных 5 диссертациях); время классификации указано для одной страницы.

Заключение

Данная статья описывает проблему автоматического контроля сканирования текстов, вводит некоторую формализацию этой задачи и демонстрирует первые результаты попыток ее решения.

На основании результатов из таблицы 1 для дальнейшей точной настройки были выбраны алгоритмы Gentle Boost с решающими деревьями и k-ближайших соседей с Евклидовой метрикой, причем бустинг представляется наиболее перспективным в силу большего числа варьируемых параметров.

Следующим шагом в решении поставленной задачи будет увеличение объемов обучающей и тестовой выборок, что позволит получить более точные статистические данные. По-прежнему остается открытым вопрос о перестановках/пропусках страниц, который требует отдельного алгоритма, рассматривающего не каждую отдельную страницу, а их последовательность.

Также отметим, что подсчитанный *FPR* сильно завышен относительно реального количества пропущенных ошибок, поэтому еще одним шагом к решению задачи должно стать тесное сотрудничество с ОТК РГБ и внедрение описанных алгоритмов для их проверки в реальных условиях.

Литература:

1. Bishop, C.M. Pattern Recognition and Machine Learning / C.M. Bishop. - New York: Springer, 2006. - XX, 738 p.: ill. - (Information science and statistics).

2. Schapire, R.E. The Strength of Weak Learnability / R.E. Schapire // Machine Learning. - 1990. - Vol. 5. - P. 197-227.

3. Rubner, Y. The Earth Mover's Distance as a Metric for Image Retrieval / Y. Rubner, C. Tomasi, L.J. Guibas // International Journal of Computer Vision. - 2000. - Vol. 40, № 2. - P. 99-121.

4. MATLAB [Электронный ресурс] / The MathWorks. - Режим доступа: http://www.mathworks.com/products/matlab/ (20.11.2013).

5. Ghostscript [Электронный ресурс] / Artifex Software. - Режим доступа: http://www.ghostscript.com (20.11.2013).

НАША ИНФОРМАЦИЯ

ФГБУ «Российское энергетическое агентство» Минэнерго России (РЭА) совместно с Технологической платформой «Комплексная безопасность промышленности и энергетики» провели секцию «Безопасность в ТЭК».

20 мая 2014 г. в Москве состоялся I Всероссийский съезд «Технологическая платформа "Комплексная безопасность промышленности и энергетики" - основа технологической модернизации России».

РЭА выступило одним из организаторов секции «Безопасность в ТЭК» (модератор - заместитель генерального директора А. Беднов).

В работе секции приняли участие ведущие ученые, руководители и специалисты в области обеспечения энергетической безопасности, а также представители экспертных организаций и органов власти.

В рамках работы секции были подняты следующие вопросы:

- Совершенствование нормативной правовой базы в области обеспечения комплексной безопасности объектов топливно-энергетического комплекса.
- Глобальная энергетическая безопасность. Проблемы и пути решения.
- Реализация ФЗ-256 «О безопасности объектов топливно-энергетического комплекса».
- Требования по обеспечению пожарной безопасности и антитеррористической защищенности объектов топливно-энергетического комплекса и атомной энергетики.
- Надежность систем энергетики, энергетическая безопасность и др.

По материалам сайта: *http://www.rosenergo.gov.ru*