Март, № 2 **Физико-математические науки** **2012**

УДК 004.9

АНДРЕЙ НИКОЛАЕВИЧ ТАЛБОНЕН
преподаватель кафедры теории вероятностей и анализа данных математического факультета, Петрозаводский государственный университет
antal@sampo.ru

АЛЕКСАНДР АЛЕКСАНДРОВИЧ РОГОВ
доктор технических наук, профессор, заведующий кафедрой теории вероятностей и анализа данных математического факультета, Петрозаводский государственный университет
rogov@psu.karelia.ru

АНАЛИЗ МАШИНОПИСНЫХ ПОДПИСЕЙ К ФОТОГРАФИЯМ В ЦИФРОВОМ ИСТОРИЧЕСКОМ АЛЬБОМЕ

В данной статье описывается метод построения поисковой системы для узкоспециализированной текстовой коллекции, допускающей наличие текстовых ошибок, на примере коллекции фотографий со строительства Беломорско-Балтийского канала. Предложенные методы можно использовать для других проектов, связанных с работой над коллекциями исторических документов.

Ключевые слова: фильтрация изображений, распознавание текста, морфологический анализ, лексический анализ, поисковый индекс

ОСОБЕННОСТИ ЗАДАЧИ

Необходимость организации поиска в большом массиве цифровых исторических фотографий, относящихся к одной более крупной тематике, потребовала решения рассмотренных в данной статье задач. В качестве исходного материала используется электронная коллекция черно-белых изображений строительства Беломорско-Балтийского канала (ББК) в формате JPEG, сделанных около 80 лет назад. Данная коллекция состоит из 8 альбомов в среднем по 800 снимков в каждом, что в общей сложности составляет почти 6,5 тыс. изображений. Коллекция находится в Карельском государственном краеведческом музее. Пример изображения представлен на рис. 1. Каждое изображение коллекции – сфотографированный лист бумаги с фотографией и напечатанным текстом, который содержит информацию о времени, месте объекта снимка и краткое описание самого объекта и сюжет.

Извлечение информации из текстовых подписей к фотографиям и последующая логическая организация этой информации позволяют построить информационную систему с возможностью поиска по данной коллекции на основе текстового запроса.

Следующие характеристики изображений коллекции существенно затруднили их анализ: цифровые изображения были получены методом фотографирования при достаточно низком разрешении (75 dpi); на текстовом фоне и на участках литер присутствуют размытые участки и эффекты «соль и перец» [1; 194], которые обусловлены изношенностью бумаги и ошибками при нанесении текста; отсутствует резкость линий букв текста; уровни серого текста и фона на некоторых фотографиях отличаются незначительно.

Рис. 1. Пример изображения коллекции, посвященной строительству ББК

ФОРМИРОВАНИЕ ТЕКСТОВОЙ КОЛЛЕКЦИИ

Общая схема обработки коллекции

Процесс извлечения текстовой информации из изображения основан на распознавании символов. Были протестированы 3 наиболее популярные системы оптического распознавания символов (Optical Character Recognition – OCR), поддерживающие кириллицу: FineReader, Google Tesseract, CuneiForm. Наилучшие результаты показал FineReader, однако даже он смог получить информацию только с 38 % правильно распознанных символов. На качество распознавания основное влияние оказало низкое качество самих изображений, прежде всего размытость и зашумленность. Кроме того, были отмечены многочисленные случаи выделения текста на участках изображения, где его не было. Для

повышения качества итоговой коллекции было решено использовать дополнительные методы. В результате анализа и серии экспериментов была установлена общая последовательность операций преобразования изображений и текста, содержащая следующие шаги.

1. Выделение областей, содержащих подписи.
2. Предварительная обработка выделенных областей различными методами повышения резкости изображений, получение альтернативных изображений подписей.
3. Распознавание с помощью OCR.
4. Анализ полученных текстовых файлов с целью отбора среди них наиболее качественных.

Выполнение данных шагов позволило увеличить общую долю правильно распознанных символов до 65 %.

Предварительная обработка изображений

Выделение областей, содержащих подпись. Для выделения областей, содержащих подпись, был использован эвристический метод, описанный в [4]. Данный метод выполняет поиск прямоугольной области изображения, содержащей текст, контрастирующей с цветом фона, с точностью около 92 %. Ошибки в работе данного алгоритма возникли в результате наличия изображений, на которых присутствовали детали, помешавшие распознаванию границ областей подписей. Например, к таким деталям относятся выступающие нижние края листа бумаги, не закрытые подписью целиком. Кроме того, подписи не удавалось распознать, если они были наклеены под большим углом к краям листа. Пример результата работы эвристического метода представлен на рис. 2.

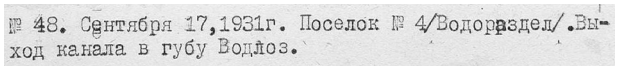
№ 48. Сентября 17,1931г. Поселок № 4/Водоразд/.Выход канала в губу Водлоз.

Рис. 2. Выделенная подпись к фотографии, представленной на рис. 1

Методы обработки изображений. С целью повышения качества распознавания было опробовано 12 методов.

1. Эвристический метод порогового отсечения, который использует в качестве порогового значения среднее значение яркости над всей обрабатываемой областью (Cut) [4].
2. Методы пространственной фильтрации, основанные на выборе ядра (11 методов).
   a. Методы, применяющие Лапласиан [1; 200], [4]: аддитивное изображение с использованием простого Лапласиана (ALaplace), выровненное аддитивное изображение с использованием простого Лапласиана (EALaplace), аддитивное изображение с использованием сложного Лапласиана (AELaplace), выровненное аддитивное изображение с использованием сложного Лапласиана (EAELaplace).
   b. Методы на основе выделения границ [4]: оператор Собеля (ASobel) [1; 212], [6], оператор Робертса (ARoberts) [1; 212], оператор Прюитт (APrewitt) [5], оператор Щарра (AScharr) [8].
   c. Методы сглаживания изображения: простое сглаживание (Smooth) [1; 190], [2], [7], размытие по Гауссу (Gauss) [2], [8].
   d. Метод медианного фильтра (Median) [1; 194].

Результаты, полученные после обработки изображения подписи различными методами, будем называть альтернативными изображениями. Набор изображений подписей, обработанных одним методом, назовем альбомом.

Оценка качества распознавания

Общие идеи оценивания качества. Результат распознавания альбома будем называть текстовым альбомом. Элемент текстового альбома назовем текстовым файлом. В данной работе применительно к рассматриваемой коллекции были выделены следующие элементы подписи: номер, дата и основной текст. Разделение текстового файла на элементы осуществляется эвристическим образом на основе закономерностей, выявленных у текстов подписей [4]. Для оценки качества работы методов были рассчитаны веса слов и механизм поиска ошибок.

Алгоритм определения веса слова опирается на список словоформ слов русского языка, а также дополнительные списки словоформ, не встречающихся в основных словарях. Поиск ошибок основан на использовании расстояния Левенштейна [9]:

1. Каждое слово ищется среди множества всех словоформ.
2. Для каждого ненайденного слова определяются слова-кандидаты с минимальным расстоянием Левенштейна. Если минимальное расстояние превышает заданный порог или длину слова, то слово считается нераспознанным. Данный порог может быть как абсолютным, так и относительным.
3. Для каждого слова рассчитывается вес w по формуле $w = \frac{n - L}{n}$, где n – длина слова, L – расстояние Левенштейна до слова-кандидата. В случае, когда $L > n$, вес считается равным 0. Вес слова, найденного в словаре, будет равен 1.
4. Общий вес отдельного элемента текстового файла (как и всего файла), состоящего из нескольких слов, определяется как сумма весов составляющих его слов.
5. Общий вес всего текстового альбома определяется как сумма весов составляющих его текстовых файлов.

Сравнение результатов распознавания. Сравнение методов основано на сравнении общих оценок текстовых альбомов. В качестве оценки всего текстового альбома используется среднее

арифметическое оценок его текстовых файлов. Когда средние оценки двух текстовых альбомов примерно равны, производится сравнение максимальных весов текстовых файлов. В табл. 1 приведены результаты сравнения общих оценок альтернативных текстовых альбомов. Серым цветом выделены наиболее качественные методы.

Таблица 1

Пример сравнения различных методов

Метод	Средняя оценка метода	Максимальная оценка метода
Cut	0,67	0,88
EALaplace	0,6	0,82
AELaplace	0,66	0,82
EAELaplace	0,66	0,93
AScharr	0,53	0,8
Smooth	0,55	0,86
Median	0,56	0,82
Original	0,52	0,77

Вместо сравнения оценок отдельных текстовых файлов было решено сравнивать оценки соответствующих элементов текстовых файлов. Обозначим через F_i текстовый файл (в данном контексте будем называть его просто файл), соответствующий одному и тому же исходному изображению и полученный с помощью обработки этого изображения методом i ($i = \overline{1,n}$, где n – число альтернативных методов). Тогда можно представить файл как множество составляющих его элементов следующим образом: $F_i = \{t_j \mid j = \overline{1,m}\}$, где m – количество элементов файла. Предполагается, что у альтернативных файлов число элементов совпадает, а элементы с одинаковым номером соответствуют одной и той же области текста подписи исходного изображения. Кроме того, необходимо учитывать наличие случайных символов и слов.

Пусть W_{ij} – общий вес, а N_{ij} – количество слов элемента t_{ij}. Тогда оценка элемента t_{ij} будет рассчитываться следующим образом:

$$R_{ij} = S_{ij} \cdot D_{ij},$$

где $S_{ij} = \begin{cases} \frac{W_{ij}}{N_{ij}}, N_{ij} > 0 \\ 0, N_{ij} = 0 \end{cases}$, $D_{ij} = \begin{cases} \sqrt{1 - \left(\frac{N_{ij} - \overline{N_j}}{\overline{N_j}}\right)^2}, N_{ij} \le 2 \cdot \overline{N_j} \\ 0, N_{ij} > 2 \cdot \overline{N_j} \end{cases}$ и $\overline{N_j} = \frac{\sum_{i=1}^{n} N_{ij}}{n}$.

В приведенной выше формуле S_{ij} играет роль относительной оценки элемента, определяющей долю правильно распознанных в нем символов. Фактор D_{ij} в данном случае определяет отклонение количества слов элемента t_{ij} от среднего, то есть от того количества слов, которое следовало бы ожидать после распознавания.

В результате сравнения формируется новый текстовый файл:

$$F^* = \{t_j^* \mid t_j^* = t_{i^* j}, i^* = \arg\max_i (R_{ij}), i = \overline{1,n}, j = \overline{1,m}\}.$$

Результаты сравнения альтернативных текстовых файлов, соответствующих одному исходному изображению, приведены в табл. 2. Серым цветом выделены элементы с наибольшей оценкой.

Таблица 2

Пример сравнения альтернативных текстовых файлов

Метод	Оценка текста	Оценки атрибутов	
		Дата	Номер
Cut	0,8	1	0,6
EALaplace	0,78	0,85	0,8
AELaplace	0,85	0,75	0,75
EAELaplace	0,82	0,9	0,9

АНАЛИЗ ТЕКСТОВОЙ КОЛЛЕКЦИИ

Общий алгоритм анализа текстовой коллекции

Общая схема анализа представлена на рис. 3.

Данная схема представляет собой условно-бесконечный цикл обработки коллекции, в котором на каждой итерации происходит постепенное улучшение качества коллекции. При этом схема предполагает наличие условия выхода. Момент, когда можно завершить цикл, либо определяется человеком, либо задается двумя пороговыми значениями: долей неизвестных слов и долей слов с ошибками.

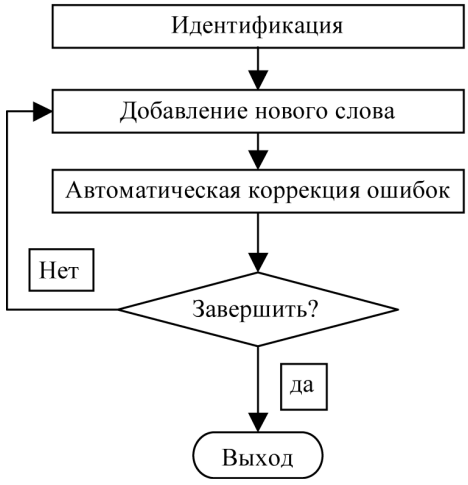

Рис. 3. Общая схема анализа текстовой коллекции

Идентификация слов

На данном этапе производится расчет весов слов для всех текстовых файлов. Для каждого распознанного слова извлекается тематическая информация на основе имеющихся тематических словарей, а также с помощью морфологического анализа, в результате чего определяется тематика слова. В данной работе используются следующие тематические словари: имен, географических названий, строительной техники,

сооружений, общий словарь (приведен здесь, поскольку является словарем по умолчанию). В качестве морфологического анализатора используется Mystem (Яндекс) [3], выявляющий флаг соответствия имени, фамилии, отчества или географического названия.

Добавление нового слова

Данный процесс требует участия пользователя. При добавлении нового слова кроме самого слова указывается также дополнительная информация: часть речи, нормальная форма слова, словоформы (наиболее часто встречающиеся), тематика.

Автоматическая коррекция ошибок

Данный процесс может выполняться автоматически, в случае определения необходимых пороговых значений. В результате добавления нового слова часть слов, которые до этого были признаны нераспознанными или ошибочными, могут изменить свой статус, поэтому на следующем шаге требуется пересчет весов всех слов, имеющих вес меньше 1. В случае, когда очередное рассматриваемое слово будет достаточно близко к новому слову, можно выполнить коррекцию слова без участия пользователя. Критерий близости может определяться превышением как абсолютным, так и относительным (разделенным на длину слова) значением расстояния Левенштейна заданного порогового значения.

Кроме решений о коррекции система должна также принимать решение о близости к тому или иному слову в случае, когда есть несколько слов-кандидатов с одинаковым расстоянием Левенштейна. Основным решением этой проблемы является определение тематики каждого кандидата и отнесение рассматриваемого слова к одной из них, однако в данной работе указанный метод не реализован.

Построение индекса

Построение полнотекстового индекса выполняется с помощью приведения всех слов к нормальной форме. Для этого используется как морфологический анализ (Mystem), так и дополнительные словари, заполненные пользователем в процессе добавления новых слов. Формирование индекса можно осуществить с помощью любой СУБД, поддерживающей полнотекстовый поиск. Суть данной операции заключается в заполнении специально подготовленной таблицы, включающей в себя информацию о файлах коллекции, атрибуты, найденные в текстах подписей, а также поле для хранения нормальных форм слов основного текста. Последнее поле индексируется для полнотекстового поиска средствами СУБД. В данной работе использовалась СУБД MS SQL Server.

Для повышения точности индекса были разработаны специальные правила лексического анализа, позволяющие выполнять поиск и подстановку отдельных словосочетаний внутри индекса с заменой на уникальные ключевые слова. Каждое правило можно представить в виде ориентированного набора элементов определенного типа. В данной работе были выделены следующие типы элементов правил: строковая константа, лексема, например слово, состоящее из букв алфавита, лексическая группа (часть речи), тематическая группа (слово относится к одному из тематических словарей), онтологическая группа (слово принадлежит определенному таксономическому узлу).

Данные правила позволяют выполнять поиск ключевых словосочетаний не только в текстовой коллекции, но и в поисковом запросе, что, в свою очередь, позволяет уточнять также и результат поиска. Имеется возможность осуществлять поиск различных сокращений, которые в большинстве случаев никак не распознаются.

УТОЧНЕНИЕ И РАСШИРЕНИЕ ПОИСКОВОГО ЗАПРОСА

Уточнение поискового запроса реализовано с помощью описанного выше правила лексического анализа, при котором отдельные словосочетания, содержащиеся в запросе, заменяются ключевыми словами.

С помощью онтологии / тезауруса можно расширить запрос за счет включения понятий онтологии / тезауруса, связанных с ключевыми словами запроса различными отношениями:

1. С помощью синонимов;
2. В глубину за счет отношений «род – вид» и «часть – целое»;
3. В любую сторону на 1 шаг, рассматривая и другие отношения, например ассоциативные;
4. С помощью комбинации различных отношений.

ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ

Для выполнения операций по обработке изображений и анализа текстовой коллекции был разработан программный комплекс «BBKManager». Для выполнения поиска по рассматриваемой коллекции была разработана специальная программа «BBKClient», осуществляющая полнотекстовый поиск по сформированному индексу. Поиск осуществляется средствами СУБД и основан на алгоритме ранжирования BM25 [10]. На рис. 4 представлен пример работы программы. Данная программа позволяет выполнять следующие операции:

1. Обычный поиск по текстовому запросу.
2. Расширенный поиск по текстовому запросу и атрибутам (дата, номер).
3. Поиск по текстовому запросу в обоих случаях (пункты 1 и 2) с дополнением операций уточнения и расширения запроса.
4. Настройка клиента с возможностью указать источники изображений, а также включаемые связи по умолчанию для метода расширения запроса.

Рис. 4. Пример работы программы-клиента

Операция уточнения запроса реализована с помощью специального индекса, в котором для каждого правила определены константы, являющиеся ключами в данном индексе. Для операции расширения запроса используется тезаурус, поддерживающий отношения «род – вид», «часть – целое» и синонимы. Данный тезаурус представляет собой XML-файл, который полностью загружается в память программы для ускорения поиска по узлам тезауруса.

ЗАКЛЮЧЕНИЕ

В результате работы была обработана часть коллекции изображений строительства ББК, предоставленная Карельским государственным краеведческим музеем, объемом 2000 снимков. Для построения индекса была разработана отдельная программная система «BBKManager». Кроме того, для поиска была разработана специальная программа, использующая алгоритм ранжирования BM25 на базе СУБД MS SQL Server (рис. 4) и поддерживающая операции уточнения и расширение запроса.

В дальнейшем планируется провести анализ методов обработки изображений с целью обнаружения лиц, контуров и текстур. Будут исследованы возможности библиотеки OpenCV, в частности функции обнаружения объектов на основе вейвлета Хаара. Предполагается применить методы распознавания для поиска объектов с характерной текстурой, например воды, стенок карьеров или бревенчатых домов.

СПИСОК ЛИТЕРАТУРЫ

1. Гонсалес Р., Вудс Р. Цифровая обработка изображений: Пер. с англ. М.: Техносфера, 2005. 1072 с.
2. Каньковски П. Как работают фильтры размытия [Электронный ресурс]. Режим доступа: http://www.computerra.ru/gid/rtfm/graphic/35934/
3. О программе mystem [Электронный ресурс]. Режим доступа: http://company.yandex.ru/technology/mystem/
4. Талбонен А. Н., Рогов А. А. Анализ машинописных подписей к фотографиям в цифровом альбоме // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды XII Всероссийской научной конференции RCDL'2010. Казань: Казан. ун-т, 2010. С. 422–429.
5. Edge Detection [Электронный ресурс]. Режим доступа: http://www.cse.unr.edu/~bebis/CS791E/Notes/EdgeDetection.pdf
6. Engel K. Real-Time Volume Graphics. AK Peter, 2006. P. 112–113.
7. Image Processing for Dummies with C# and GDI+ Part 2 – Convolution Filters [Электронный ресурс]. Режим доступа: http://www.codeproject.com/KB/GDI-plus/csharpfilters.aspx
8. Scharr H. Optimal Operators in Digital Image Processing [Электронный ресурс]. Режим доступа: http://archiv.ub.uniheidelberg.de/volltextserver/volltexte/2000/962/
9. Wagner R. A., Fischer M. J. The String-to-string correction problem [Электронный ресурс]. Режим доступа: http://www.daimi.au.dk/~cstorm/courses/AiBS_e08/papers/WagnerFisher_EditDist.pdf
10. Zaragoza H., Craswell N., Taylor M., Saria S., Robertson S. Microsoft Cambridge at TREC-13: Web and HARD tracks [Электронный ресурс]. Режим доступа: http://trec.nist.gov/pubs/trec13/papers/microsoft-cambridge.web.hard.pdf