Поиск в электронных коллекциях на основе тексто-графической информации

А. Н. Талбонен, А. А. Рогов

Петрозаводский государственный университет
antal@sampo.ru, rogov@psu.karelia.ru

Аннотация

Данная статья посвящена вопросам организации поиска в коллекциях цифровых изображений, представляющих историческую ценность. В статье описываются возможности программного комплекса «ПФЦИК», специально разработанного для решения задач организации поиска по таким коллекциям на основе содержания как текстовой, так и графической информации.

1. Введение

Для активно создаваемых в настоящее время электронных коллекций оцифрованных фотографических фондов музейных и архивных организаций требуется организация современных методов поиска по ним. Их создание вызывает определенные затруднения, связанные с поиском объектов на изображениях. Кроме того, качество фотографий и их оцифрованных изображений является недостаточным для применения существующих алгоритмов поиска объектов.

На данный момент не существует комплексных средств, позволяющих на основе заданной подобной коллекции и соответствующей предметной области организовать полнотекстовый и атрибутивный поиск, необходимый для представления коллекции широким массам. В связи с этим был разработан программный комплекс для построения поисковой системы для цифровых исторических коллекций «Поисковый фреймворк для цифровых исторических коллекций» («ПФЦИК»).

2. Описание программного комплекса

2.1. Размер страницы

Программный комплекс позволяет решать комплексную задачу поиска информации различного рода, содержащейся на оцифрованных изображениях. К объектам поиска можно отнести текстовую информацию, представленную графически на цифровых изображениях, различные объекты или их части (к ним можно отнести здания, предметы, транспортные средства, людей и тд), контуры данных объектов и различные поверхности, представленные текстурами (вода, деревянная или кирпичная кладка, деревья).

Основная идея комплексного построения поисковой системы заключается в использовании существующих поисковых движков, создании базы данных ключевых слов и соответствующего полнотекстового индекса. Таким образом, разработанный программный комплекс представляет собой некий поисковый робот, основная задача которого, извлечение информации для последующего индексирования. Общая схема работы данного робота представлена на рисунке 1.

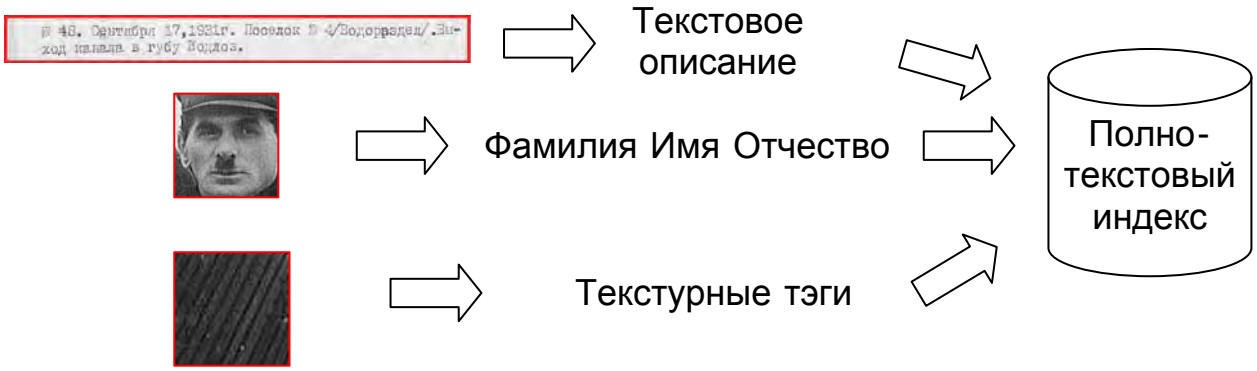

Рис. 1. Общая схема работы поискового робота, реализованного в данном программном комплексе

Интернет и современное общество: сборник научных статей XVI Всероссийской объединенной конференции IMS-2013, Санкт-Петербург, 9 - 11 октября 2013 г.

Комплекс сочетает в себе несколько программных и информационных систем, использующих как существующие программные и алгоритмические средства (см. [4, 5]):

1. современные системы распознавания текста («FineReader» [3]);
2. морфологический анализатор «Mystem» [2];
3. словарь словоформ русского языка;
4. СУБД «MS SQL Server» [1] и др;

так и собственные методы и алгоритмы (подробно описанные в [4, 5]):

1. метод организации полнотекстового поиска на основе текстовой информации, содержащей ошибки, и текстовых меток, полученных с помощью аннотирования изображений;
2. метод извлечения текстовой информации, содержащей ошибки, из коллекции изображений «низкого» качества;
3. метод поиска лиц людей на изображениях, основанный на повышении точности и полноты результатов работы алгоритма Виолы-Джонса [8] с помощью локальных бинарных шаблонов [6];
4. метод аннотирования изображений с помощью текстурного классификатора, построенного на основе метода моментов [7];
5. метод повышения точности и полноты текстового запроса с помощью лексических правил.

Программный комплекс предоставляет пользователю следующие функции:

1. Выполнять обработку коллекции изображений и извлекать текстовую информацию;
2. Формировать и индексировать текстовую коллекцию на основе извлеченной из коллекции изображений текстовой информации;
3. Редактировать текстовую коллекцию путем добавления новых слов, автоматического исправления ошибок и поиска различных словосочетаний по наборам лексических правил;
4. Задавать различные лексические правила поиска и соответствующие команды редактирования для поиска словосочетаний в текстовой коллекции;
5. Формировать полнотекстовую таблицу, индексируемую для выполнения соответствующего поиска;
6. Выполнять поиск по коллекции изображений с помощью текстовых запросов и дополнительных атрибутов;
7. Выполнять поиск лиц людей в коллекции изображений различными методами;
8. Организовывать коллекции лиц из результатов обнаружения лиц;
9. Выполнять экспертное оценивание некоторой выборки из коллекции изображений, включая специально разработанное быстрое экспертное оценивание на основе коллекции обнаруженных лиц из этой выборки;
10. Выполнять оценку точности и полноты коллекции лиц с помощью экспертного оценивания;
11. Создавать и редактировать текстурные классификаторы и выполнять по ним поиск;
12. Выполнять автоматический подбор параметров отдельных текстур классификатора;
13. Выполнять экспертное оценивание некоторой выборки из коллекции изображений на предмет наличия текстур из тестовой выборки;
14. Выполнять оценку точности и полноты результатов поиска текстур;
15. Выполнять обработку изображений различными методами, получать и сравнивать гистограммы;
16. Добавлять и тестировать различные модификации алгоритма локальных бинарных шаблонов;
17. Тестировать алгоритм текстурной сегментации методом моментов с различными параметрами, выполнять автоматический подбор оптимальных параметров для данного алгоритма на основе обучающей выборки.

Все функции различных подсистем комплекса можно разделить на две группы: общие функции, которые не зависят от предметной области (например, функции обработки изображений, поиска лиц людей и д.р.), и предметно-зависимые функции, которые определяются предметной областью и особенностями заданной коллекции, и которые для другой предметной области будут другими. К подобным функциям можно отнести различные правила и закономерности разбора текстовой информации, представленной на изображениях коллекции, функции настройки модулей программных систем.

Поскольку каждая отдельная коллекция обладает своими характерными особенностями, требуется отдельная работа по разработке предметно-зависимых функций, настройке модулей системы и подборе соответствующих параметров. Поэтому для подобных проектов целесообразно построение архитектуры с выделением ядра и поддержкой подключаемых модулей (так называемых плагинов).

Текущее состояние разработанного программного комплекса не предполагает разделение функций, поскольку он разрабатывался для решения задачи организации поиска по одной конкретной коллекции. В дальнейшем планируется переработать архитектуру комплекса с целью поддержки подключаемых модулей, реализующих предметно-зависимые функции.

Разработанный комплекс был апробирован на коллекции фотографий со строительства ББК из фондов Национального музея республики Карелия. Данная коллекция содержит 6,5 тыс. изображений. Результатом применения комплекса стала разработанная информационно-поисковая система «ИПС ББК» (см. рис. 2), использующая построенный с помощью комплекса полнотекстовый индекс.

Рис. 2. Внешний вид информационно-поисковой системы «ИПБ ББК» – результата применения программного комплекса к коллекции фотографий со строительства ББК

Литература

[1] Компонент Full-Text Search (SQL Server). URL: http://msdn.microsoft.com/ru-ru/library/ms142571.aspx. (дата обращения: 01.07.2013)

[2] О программе mystem — Компания Яндекс. URL: http://company.yandex.ru/technology/mystem/. (дата обращения: 01.07.2013)

[3] Программа для распознавания текста ABBYY FineReader. http://www.abbyy.ru/finereader/. (дата обращения: 01.07.2013)

[4] Талбонен А. Н., Рогов А. А. Анализ машинописных подписей к фотографиям в цифровом историческом альбоме. Ученые записки Петрозаводского государственного университета. Серия «Естественные и технические науки», 2012, № 2 (123), С. 109-113.

[5] Талбонен А. Н., Рогов А. А. Модели и методы поиска людей на фотографиях из исторического альбома. Ученые записки Петрозаводского государственного университета. Серия «Естественные и технические науки», 2012, № 6 (127), С. 113-117.

[6] Ahonen T., Hadid A., Pietikäinen M. Face Recognition with Local Binary Patterns. URL: http://masters.donntu.edu.ua/2011/frt/dyrul/library/article8.pdf. (дата обращения: 01.07.2013)

[7] Tuceryan M. Moment Based Texture Segmentation. URL: http://cs.iupui.edu/~tuceryan/pdf-repository/Tuceryan1992.pdf. (дата обращения: 01.07.2013)

[8] Viola P., Jones M. Robust Real-time Object Detection. URL: http://research.microsoft.com/en-us/um/people/viola/Pubs/Detect/violaJones_IJCV.pdf. (дата обращения: 01.07.2013)

Search in digital collections based on textual and graphical information

A. N. Talbonen, A. A. Rogov

This article is devoted to the organization in search of a collection of digital images representing historical value. The paper describes the capabilities of the software specially designed to meet the challenges for providing search these collections on the basis of the content of both textual and graphical information.