

УДК 004.622; 004.912; 004.932.75'1

**ИССЛЕДОВАНИЕ ФУНКЦИОНАЛЬНЫХ ВОЗМОЖНОСТЕЙ ПРОГРАММЫ
ДЛЯ АВТОМАТИЗАЦИИ ПРОЦЕССА ОБРАБОТКИ ИНФОРМАЦИИ
БИБЛИОГРАФИЧЕСКИХ КАРТОЧЕК**

Коробковский В.А. (Университет ИТМО)

Научный руководитель – к.т.н., доцент Горлушкина Н.Н.

(Университет ИТМО)

Введение. В XXI веке прослеживается четкая тенденция на проведение цифровизации. Исключением не стали и библиотеки, которые активно занимаются созданием электронных каталогов из информации, хранящихся в их фондах. Наиболее популярной системой автоматизации, предназначенной для создания и ведения электронной библиотеки, является ИРБИС64+ [1]. Библиографические данные добавляются в систему на основе коммуникативного формата RUSMARC, являющегося национальной адаптацией формата UNIMARC. Однако на текущий момент этот процесс выполняется вручную, что занимает большое количество времени и сил, поэтому конечной целью стала разработка программы для его автоматизации с сопутствующим изучением возникающих проблем и предложением вариантов для их преодоления.

Основная часть. В рамках данного исследования было продолжено улучшение программы, которая была реализована в процессе подготовки к защите бакалаврской дипломной работы [2, 3]. Имеющийся программный код позволял проводить предобработку изображений, локализацию необходимых областей и считывание информации из них, а также её разбиение на поля формата RUSMARC. В зависимости от количества информации он позволял обрабатывать от 120 до 200 карточек в час.

Текущей задачей стало проведение исследования для определения возможных функциональных и оптимизационных улучшений. Программа была запущена на 1200 карточках, в процессе её работы проводился замер времени выполнения каждой функции. На основании полученной информации было принято решение о необходимости ускорения процесса поворота изображения и переработки функции для построчного распознавания из-за ненадёжности выбранного метода и возникающих ошибок.

Исправление функции построчного распознавания было проведено с помощью расчёта профиля горизонтальной проекции [4, 5] путём вычисления суммы значений пикселей по горизонтали и его нормализации для последующего поиска пиков. Вместо локальных максимумов производится поиск минимумов, так как на вход подаётся инвертированное изображение. Пики, значения которых превышают 5% от максимальной суммы, удаляются. После добавления индекса первой и последней строк в список с пиками его можно использовать для разделения.

Что касается функциональных изменений, то все они связаны с форматом RUSMARC [6, 7]. Была переработана функция для учёта авторов: теперь разделение проводится по каждой второй запятой, исключая первую, так как она всегда отделяют фамилию автора от его инициалов. Также был добавлен учёт годов жизни: если отделённая часть начиналась с цифры, то она записывалась к предыдущему автору, а алгоритм продолжал свою работу со следующей запятой при её наличии.

Помимо этого, были внесены более мелкие доработки в виде учёта периодики в области шифра и хранения и ссылок в области заглавия посредством поиска последнего двойного переноса строк. Также были удалены лишние скобки в подполях серии и ответственных лиц.

Выводы. Результатом оптимизации программы стало ускорение процесса выравнивания изображения более чем в 4 раза благодаря распараллеливанию

вычислений сумм гистограммы. Всё это положительно сказалось и на скорости обработки информации из каждой области, поскольку перед ней изображение выравнивается по горизонтали. Построчное распознавание стало выполняться чуть медленнее, однако результат оказался более точным, так как ошибок больше не возникало, а увеличение времени работы этой функции незначительно по сравнению с тем, которое удалось сократить.

Таким образом, итогом упомянутых изменений оказалось не только добавление нового функционала, но и снижение времени выполнения программы с 6 до 4 часов для 1200 изображений.

Список использованных источников:

1) Система ИРБИС64+ [Электронный ресурс]. — URL: https://elnit.org/index.php?option=com_content&view=article&id=255:irbis64&catid=18:kharakteristiki-produktov (дата обращения: 04.02.2025).

2) Коробковский В.А. (науч. рук. Горлушкина Н.Н.) Разработка программы для автоматизации процесса обработки информации библиографических карточек // Сборник тезисов докладов конгресса молодых ученых. Электронное издание. – СПб: Университет ИТМО, [2024]. URL: <https://kmu.itmo.ru/digests/article/12882>

3) Коробковский В.А., Горлушкина Н.Н., Белинская М.А. Разработка алгоритма автоматизации ретроконверсии для создания электронного каталога. *Научные и технические библиотеки*. 2025;(2)

4) Ashwin R. Dobariya, Prof. V. R. Rathod. Horizontal and Vertical Projection techniques for Line & word Segmentation Process in Offline Handwritten Gujarati Text [Электронный ресурс]. — URL: <http://www.publishingindia.com/GetBrochure.aspx?query=UERGQnJvY2h1cmVzfC8zMzI1LnBkZnwvMzMzMyNS5wZGY=> (дата обращения: 03.01.2025)

5) Silva, Chamari. Segmenting Sinhala Handwritten Characters. *International Journal of Conceptions on Computing and Information Technology*. 2. [Электронный ресурс]. — URL: https://www.researchgate.net/figure/Use-of-horizontal-projection-profile-for-line-segmentation_fig4_270280869 (дата обращения: 05.01.2025)

6) Российский коммуникативный формат представления классификационных данных (российская версия UNIMARC Classification Format) [Электронный ресурс]. — URL: <http://www.rusmarc.ru/commun/index.html> (дата обращения: 04.02.2025)

7) Официальный сайт ФГБУ “Президентская библиотека имени Б.Н. Ельцина”. Сводная таблица использования полей RUSMARC для различных уровней архивного описания [Электронный ресурс]. — URL: https://www.prilib.ru/sites/default/files/u1/descrm_table.pdf (дата обращения: 04.02.2025)