система учёта потребностей и возможностей работника и характеристик рабочего места. Для поиска различных вариантов решения вопроса организации трудовых отношений, рассматриваются две модели подбора пары работник – рабочее место:

– составление рейтинга работников среди общего количества соискателей для конкретного рабочего места. Отбор количества работников участвующих в рейтинге происходит по основным критериям рабочего места.

– составление рейтинга доступных рабочих мест для конкретного соискателя. Отбор количества рабочих мест происходит по основным критериям соискателя.

Данная информационная система рассматривает различные модели составления рейтинговых показателей в зависимости от общепринятых статистических данных приоритетов конкретных характеристик как работника, так и рабочего места, от выставляемых показателей поиска работником и работодателя.

Метод выделения областей математической нотации на растровых изображениях печатных документов

И.Г. Масков, А.Ю. Андреева
АлтГТУ, г. Барнаул

Одной из важных задач оптического распознавания документов является задача распознавания математических выражений. Проблема оцифровки технической и научной документации стоит перед издательствами, библиотеками. Поисковым системам для успешной организации поиска необходимо иметь средства выделения формул из растровых изображений.

Общепринятая структура механизма распознавания математической нотации, разделается на три этапа [1]:

1) выделения областей математических выражений на документе;

2) лексический анализ выделенных областей;

3) построение деревьев синтаксического разбора.

Первый этап работы реализуется с помощью набора эвристик, полученных статистическим анализом более 10000 документов в [2]. В результате анализа выявлено, что большинство областей с математическими выражениями располагаются в документе двумя способами: 1) непосредственно в строке текста, 2) на отдельной строке, с большим пространством пустоты сверху и снизу области.

В данной работе решается задача первичного выделения областей обоих типов.

Для работы метода определения областей математических выражений первого типа, представленного в [2], необходимо предварительно выделить области, содержащие только строки текста. В статье [3] предлагается подход, суть которого заключается в анализе гистограммы распределения чёрных пикселей по Y координате изображения документа. График гистограммы области со строками текста представляет собой «расчёску», т.к. среднее количество чёрных пикселей в строке текста по каждой Y координате примерно одинаково, а между строками оно приближается к нулю. Для определения периодичности в гистограмме используется дискретное преобразование Фурье. Если гистограмма имеет вид «расчёски», то на графике результата преобразования будет выделяться доминирующая частота, для оценки этого анализируется коэффициент эксцесса.

Однако, практическая реализация этого метода не дала эффективных результатов.

Авторами используется иной подход для выделения областей строковых блоков текста на основе подхода, предложенного в [4].

Метод использует штриховой фильтр (Stroke Filter) для определения областей-кандидатов на строковые блоки, после чего используется метод опорных векторов для уточнения границ блоков.

Лексемы в полученных строках выделяются с использованием метода из [5]. Здесь используются топографические признаки при анализе символов в градации серого и дальнейшая сегментация на графах.

После распознавания строк и лексем в них, строки документа анализируется на содержание 25 наиболее часто встречающихся символов (ключевые символы) в математических выражениях (это =,+,- и т.д.). Из выбранных строк выделяются выражения по следующему эвристическому алгоритму: 1) берётся первое слово слева, содержащее ключевой символ, ещё не включённое в математическую область 2) если слово содержит только бинарный оператор, то оба слова прилегающее слева и справа включаются в математическую область 3) слова смежные определяющему математическому символу включаются в математическую область если содержат следующее: другие математические символы, надстрочные и подстрочные символы, точки, числа. Шаги повторяются, пока все найденные ключевые символы не будут обработаны.

Области второго вида выделяются без распознавания лексем [2]. Здесь используются два важных свойства: 1) область окружена обширными (по сравнению с высотой символа текста) пустотами на до-

кументе 2) в строке математической формулы Y координата нижних левых граничных пикселей символов имеет в разы большую дисперсию по сравнению с той же величиной строки текста, ввиду того, что технических и научных публикациях используются «моновысотные» шрифты. Таким образом, сперва выделяются области, обрамлённые широкими пустотами, далее, анализируется дисперсия Y координаты самого левого нижнего пикселя каждого символа.

Полученные результаты позволяют сделать вывод о хорошем качестве выделения символов и возможности использования данного подхода для решения основной задачи – лексического анализа текста и выражений в математической нотации.

Библиографический список

1. Fateman R.J., Tokuyasu T. Progress in recognizing typeset mathematics // Proceedings of SPIE The International Society for Optical Engineering. – 1996. – Vol. 2660. – P. 37–50.

2. Garain, U., Chaudhuri, B.B., A Syntactic Approach for Processing Mathematical Expressions in Printed Documents // International Conference on Pattern Recognition'00. – Spain, 2000, vol. IV, P. 523-526.

3. Южиков В.С. Сегментация изображения страницы с текстом с текстом // Вестник ИжГТУ. – 2007. – №3. – С. 84-47.

4. Issam El-Naqa, Yongyi Yang, Miles N. Wernick. Support vector machine learning for detection of microcalcifications in mammograms. [Электронный ресурс]. – Режим доступа: http://www.ipl.iit.edu/IPL-Conferance_papers/isbl.pdf - Загл. с экрана.

5. Lee S-W., Lee D.-J., Park H.-S., A New Methodology for Gray-scale Character Segmentation and Recognition // IEEE Transactions on pattern analysis and machine intelligence. – IEEE, 1996, vol. 18, P. 1045-1050.

Анализ возможностей объектной системы CLOS[1]

О.Н. Половикова
АлтГУ, г. Барнаул

Функциональный язык LISP был на пике своей популярности в 70-80 годах прошлого столетия, когда активно использовался для написания программных систем в области искусственного интеллекта. На сегодняшний день растет интерес к этому языку, при этом исследова-

[1] Работа выполнена при поддержке аналитической ведомственной целевой программы «Развитие научного потенциала высшей школы (2009-2011 годы)» (код проекта №2.2.2.4/4278).