МЕТОДЫ ПОИСКА БЛИЖАЙШИХ СОСЕДЕЙ В ЗАДАЧЕ АНАЛИЗА ГРАФИЧЕСКОГО ОБРАЗА СТРУКТУРИРОВАННОГО ДОКУМЕНТА

Пономаренко Б.А.
Научный руководитель – асс. Заворотная М.Г.
Харьковский национальный университет радиоэлектроники,
кафедра микропроцессорных технологий и систем,
e-mail: bohdan.ponomarenko@nure.ua

In some tasks of artificial intelligence, the problem arises of searching among a multitude of information objects most similar to this material. For example, the ability to recognize text, information retrieval, data compression, classification and clustering, building a database of images, video documents. In order to build an effective search method, it is necessary to take into account the peculiarities of this task. For example, strings must be processed in one way, and vectors in another. Therefore, the algorithms used will depend on how well it fits the description.

В некоторых задачах искусственного интеллекта возникает проблема поиска среди многих информационных объектов наиболее схожего на заданный материал. Например, способность распознавать текста, поиск информации, сжатие данных, построение баз изображений, видео документов. Для построения эффективного метода поиска нужно учитывать особенности данной задачи. К примеру, строки должны одним способом обрабатываться, а вектора иным. Следовательно, использованные алгоритмы будут зависеть от того насколько он подходит под описание.

Для распознавания образа графического документа нужно начинать с анализа самого документа. Во время анализа оцениваются многие причины. Например, оценивается перекос образа при сканировании, выделяются ли линии и текстовые фрагменты, изображение сегментируется. Для сканирования изображения понадобится, конфигурация пикселей, интервалов компонентов связности оцениваются на основании геометрических, цветовых и текстурных характеристик. Количество пикселей изучаемого объекта на изображении могут достигать от десятков до сотен тысяч, в этом случае требуется использовать специальные алгоритмы обработки для достижения наиболее приемлемого результата.

Наиболее общей задачей может считаться задача инкрементного поиска (неприрывный ближайший поиск соседей). Эта задача состоит из регулирования заданных объектов в порядке возрастания от объекта запроса. Поэтому оптимальнее будет воспользоваться похожим решением задачи, которое на каждом шагу находит следующего по отдаленности соседа. Перечислим методы для поиска ближайших соседей. В

оптимальное решение, которое будет опираться на бинарный поиск. В бинарном методе большой размерности используются диаграммы Вороного, случайные выборки (random sampling) и иные методы вычислительной геометрии. Если для исходной практической задачи требуется найти соседей в определенном участке, то можно будет использовать ассоциативные структуры. Иным способом решения задачи может являться использование вспомогательных структурных данных, которые описывают рекурсивное разбиение исходных множеств точек данных и самого пространства соответственно. Например, всевозможные варианты R-деревья, region quadtrees, kd-деревья и алгоритмы над этими структурами.

R-деревья применяются для организации доступа к пространственным данным, то есть для индексации многомерной информации. Построение, как правило, выполняется благодаря многократному вызову операции вставки элемента в дерево. Если добавление элемента приводит к переполнению то вершина разделяется.

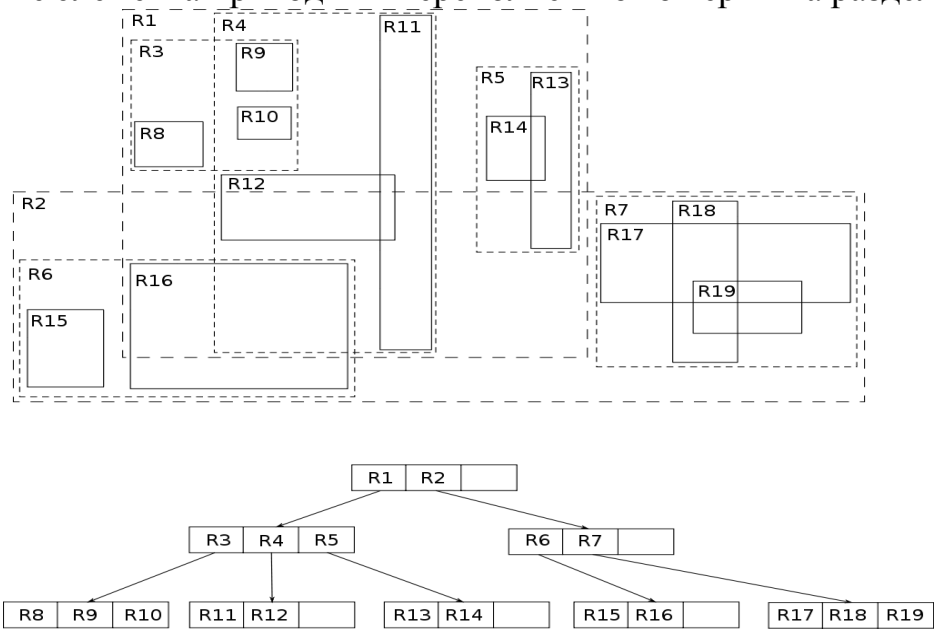

Рисунок 1

На рисунке 1 изображена схема построения R-дерева. Перечислим достоинства R-дерева: эффективно хранят локализованные в пространстве группы объектов; сбалансированы, то есть, быстрый поиск в худшем случае; вставка или удаления одной точки не изменит существенно дерево.

В данной работе были описаны методы поиска ближайших соседей с помощью R-дерева и других алгоритмах. О программной реализации этих методов с использованием алгоритмов быстрого поиска ближайших соседей, также о достоинствах и недостатков методов.

Список источников: 1 Методы поиска ближайших соседей [Электронный ресурс]. – Режим доступа: http://www.isa.ru/proceedings/images/documents/2007-29/302319.