УДК 004.91

Д.В. Тимошенко

ПРИМЕНЕНИЕ МЕТОДА ГЛАВНЫХ КОМПОНЕНТ ДЛЯ КЛАСТЕРИЗАЦИИ СИМОВЛОВ

Аннотация. В статье рассматривает алгоритм финальной кластеризации с помощью метода главных компонент. Оптимальное квантование с автоматическим выбором количества интервалов, применяемое к вектору первой главной компоненты символов одного кластера, уточняет кластеризацию по скалярным характеристикам.

Актуальность темы. С ростом цифровых ресурсов возрастает актуальность проблемы хранения, сжатия и передачи электронных растровых документов. В свое время для ее решения были разработаны несколько форматов хранения электронных растровых документов: DJVU[1], JPEG2000/Part 6[2] и LuraDocument. Другой проблемой, порожденной развитием Интернета, является управление доступом к информации. Перечисленные выше форматы, являясь по своей сути локальными, позволяют пользователю распространять документы самостоятельно без контроля со стороны авторов или правообладателей (издательств, библиотек и пр.). Для решения этой задачи был создан формат электронных документов ALD [7]. В данной работе рассмотрен один из этапов работы кодера ALD – финальная часть обработки символьного слоя электронного растрового документа, связанной с кластеризацией символов.

Анализ последних публикаций. Обычно задача анализа символа как элемента текстовой информации сводится к задаче классификации, а точнее – к задаче OCR (Optical Character Recognition) [3,4,5]. В этом случае, в отличие, от традиционной задачи кластеризации, заданы центры кластеров. В случае классификации множества символов такими центрами выступают элементы системных шрифтов. Однако в такой постановке, во-первых, изначально нужно знать язык электронного растрового документа (кроме того, часто документ многоязычный), во-вторых, в системе может не оказаться шрифтов,

которые соответствуют написанию символов обрабатываемого документа.

Одним из ключевых вопросов кластеризации символов является вопрос выбора критерия качества [4]. Наиболее часто в качестве характеристик символов для их кластеризации используют различные геометрические характеристики, такие как периметр, площадь, соотношение сторон объекта и пр. Проблема выбора геометрических характеристик перекликается с проблемой кластеризации в отсутствии информации о системных шрифтах, поскольку на вход к алгоритму может попасть электронный растровый документ, содержащий, например, иероглифы или другие символы, для которых не были предусмотрены геометрические факторы, хорошо отделяющие символы один от другого.

Постановка задачи. Целью работы является построение алгоритма финальной кластеризации символов, минимизирующего ошибки кластеризации по выбранным характеристикам без информации о языке документа и шрифте текста.

Основные результаты. Для решения проблем хранения и контроля над растровыми документами группой авторов – Лигуном А.А, Шумейко А.А и Тимошенко Д.В., была разработана технология ALD (ALLDocument), ориентированная на использование в сети [7]. Данный подход основан на разделении информации, содержащейся в электронных документах на слои.

Авторами был предложен подход, который позволил получить формат, конкурентоспособный в своей нише.

Данная статья является логическим продолжением работ, посвященных локализации слоя символов на изображении [8] и последующей кластеризации символов с помощью оптимального квантования [9].

Для того чтобы перейти к формальной постановке задачи, введем необходимые определения и понятия.

Будем считать множество 8-ми связным, если для любой его точки (x_1, y_1) существует такая точка (x_2, y_2), что выполняется условие:

$$|x_1 - x_2| \le 1, |y_1 - y_2| \le 1.$$

 ISSN 1562-9945

Таким образом, символом S_k мы будем считать 8-ми связное множество точек, полученное с помощью алгоритма локализации. Соответственно, $\{S_k\}_1^m$ – все символы, извлеченные из электронного растрового документа.

Кластером K_v будем называть множество «похожих» друг на друга символов S_k. На первом «грубом» этапе кластеризации, критерий «похожести» сводится к отличию неких скалярных величин – таких как высота, ширина, количество точек символов. Одни и те же скалярные величины у похожих символов группируются около одного и того же среднего значения соответствующей характеристики. Количество таких средних значений и сами значения можно найти с помощью оптимального (в смысле минимизации среднеквадратического уклонения) квантования [10]. Таким образом, благодаря этому подходу, задача кластеризации значительно упрощается и мы можем использовать оптимальное квантование для разбиения на кластеры скалярных величин.

В работе [9] был предложен метод иерархического разбиения на кластеры.

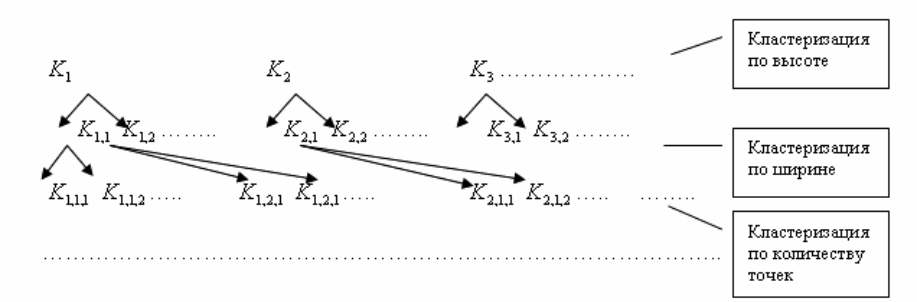

Рисунок 1 – Древовидная структура последовательной кластеризации по скалярным величинам

Очевидно, что такое разбиение нельзя считать окончательным. Рассмотрим применение такой кластеризации к последовательности символов (рисунок 2).

Interest in image compression dates back more than 35 years. The initial focus of research efforts in this field was on the development of analog methods for reducing video transmission bandwidth

Рисунок 2 – Фрагмент текста для экспериментов

Результатом работы кластеризации по 3-м характеристикам (ширина, высота, количество точек в символе) явилось разбиение указанного множества на кластеры. Есть кластеры, которые объединили в себе действительно одинаковые символы. Пример такого кластера представлен на рисунке 3.

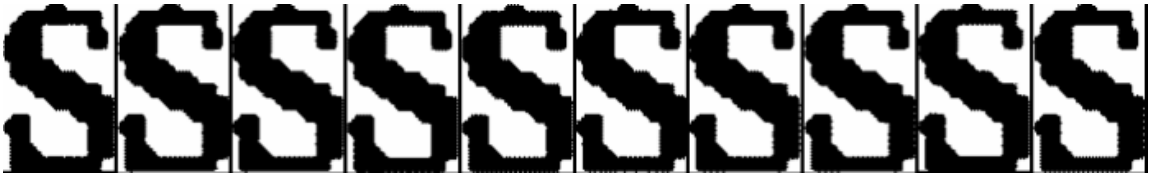

Рисунок 3 – Удачно выделенный кластер

В то же время есть кластеры, которые содержат в себе разные символы. К таким символам, для которых не будет достаточно приведенных выше критериев, относятся, например, “b”,“d”,“p”. Но не только они, в целом, ситуация достаточно типична.

Рисунок 4 – Фрагмент кластера, состоящий из разных символов

Таким образом, актуальна задача финального этапа кластеризации, целью которой является сокращение числа ошибочных кластеров.

В работе [9] дан критерий, какой кластер считать окончательным. Данный критерий основан на методе главных компонент (МГК) [11].

Поставим в соответствие каждому элементу S_k кластера K_v вектор $M_k = \{m^k{}_{i,j}\}$ с размерами $H_v \times W_v$

$$m^k{}_{i,j} = \begin{cases} 1, (i - x_k + W_v/2, j - y_k + H_v/2) \in S_k; \\ 0, (i - x_k + W_v/2, j - y_k + H_v/2) \notin S_k, \end{cases} \quad i = \overline{1, W_v}, j = \overline{1, H_v}, k = 1 \ldots N_v,$$

где

$$H_v = \max_k (H_k \mid S_k \in c_v),$$

$$W_v = \max_k (W_k \mid S_k \in c_v),$$

(x_k, y_k) - центр тяжести символа S_k.

 ISSN 1562-9945

Каждому кластеру K_v будет соответствовать шаблон T_v - вектор, который характеризует все элементы кластера K_v.

Для заданных векторов M_k необходимо найти векторы $\widetilde{T}$ и $\widetilde{\alpha}$, которые реализуют решение экстремальной задачи

$$\sum_{k=1}^{N_v}(M_k - \alpha_k T)^2 \to \min, \ \sum_{k=1}^{N_v}(\alpha_k)^2 = 1. \qquad (1)$$

При этом вектор $\widetilde{T}$, дающий решение задачи (1), называется первой главной компонентой. Существуют разные способы нахождения вектора $\widetilde{T}$. Традиционно решение сводится к нахождению максимального собственного числа (и соответствующего собственного вектора) корреляционной матрицы векторов M_k. В нашей интерпретации первая главная компонента – это шаблон кластера. Вектор $\widetilde{\alpha}$, соответствующий $\widetilde{T}$, это величина, характеризующая влияние k-го символа на формирование шаблона кластера.

Финальным кластером K_v называется множество символов S_k, для которых шаблон $\widetilde{T}$, построенный исходя из условия (1), для фиксированного τ обеспечивает неравенство

$$\sqrt{\frac{1}{N_v}\sum_{k=1}^{N_v}(M_k - \widetilde{\alpha}_k \widetilde{T})^2} < \tau \qquad (2)$$

Невыполнение неравенства (2) говорит о наличии разных символов в кластере.

Для того, чтобы разделить отличающиеся символы, которые попали в один кластер, используем итерационный алгоритм:

1. Создать кластер из любого символа первичного кластера, который нужно разделить.

2. Присоединить в новый кластер следующий символ.

3. Пересчитать с помощью МГК вектора $\widetilde{T}$ и $\widetilde{\alpha}$. Если вектора удовлетворяют неравенству (2), оставляем добавленный символ в кластер. Если же вектора не удовлетворяют неравенству (2), добавленный символ изымается. В этом случае он может образовать новый кластер или же присоединиться к другому кластеру.

Повторяем шаги 2 и 3, пока не будет проверены все символы.

Нетрудно заметить, что это достаточно трудоемкий алгоритм, так как он заключается в полном пересчете шаблона $\widetilde{T}$ и вектора $\widetilde{\alpha}$,

как минимум, для количества шагов, равного количеству символов кластера без одного.

Альтернативой данному подходу является идея об использовании для кластеризации информации, которая содержится только в векторе $\tilde{\alpha}$. Как уже говорилось выше, каждый коэффициент $\tilde{\alpha}$ отражает, насколько символ S_k похож на шаблон $\tilde{T}$. Следовательно, у похожих символов должны быть похожи и их $\tilde{\alpha}_k$. Рассмотрим коэффициенты, которые соответствуют приведенному фрагменту кластера на рисунке 4.

Таблица 1

Коэффициенты кластера с неоднородными символами

Коэффициент $\tilde{\alpha}$	Маска символа S_k
0.2146526	
0.2764183	
0.2124491	
0.2767134	
0.2055073	
0.2782143	

Нетрудно заметить, что символы “a” и “e” тяготеют к разным значениям: 0.28 и 0.21 соответственно, если округлить до 3-го знака после запятой. Таким образом, задача о кластеризации с использованием метода главных компонент сводится к задаче о кластеризации вектора коэффициентов $\tilde{\alpha}$.

Таким образом, для заданного вектора $\tilde{\alpha}$ нужно найти такой вектор b с координатами $b_{k+1/2}$ и такой вектор $\tilde{b}$ с координатами $\tilde{b}_k$, что

$$\sqrt{\frac{1}{N}\sum_{k=1}^{N}\sum_{i:b_{k/2}<\tilde{a}_i<b_{k+1/2}}(\tilde{\alpha}_i-\tilde{b}_k)^2}\to\min \qquad (3)$$

где N – количество интервалов,

$$\min_i \tilde{\alpha}_i = b_{1/2} < b_{3/2} < \ldots < b_{n+1/2} = \max_i \tilde{\alpha}_i,\ b_{k/2} < \tilde{b}_k < b_{k+1/2}.$$

$\widetilde{b}_k$ – оптимальные квантовочные числа. В нашем случае это те центры кластеров, к которым тяготеют значения вектора $\widetilde{\alpha}$.

Заметим, что задача поиска кластеров усложняется тем, что заведомо не известно какое количество кластеров может быть получено. В работе [9] , был предложен итеративный алгоритм поиска количества кластеров.

Обозначим через δ_j ошибку приближения вектора $\widetilde{\alpha}$ после разбиения на j интервалов оптимальным образом.

$$\delta_j = \sqrt{\frac{1}{j}\sum_{k=1}^{j}\sum_{i:b_{k/2}<\widetilde{a}_i<b_{k+1/2}}(\widetilde{\alpha}_i - \widetilde{b}_k)^2} ,$$

Последовательно применяя оптимальное квантование для j=1,2,.., пока не будет выполнено условие стабилизации

$$\left|\delta_j - \delta_{j+1}\right| < \varepsilon . \qquad (4)$$

Например, для случая с “e” и “a” графическое представление процесса стабилизации количества интервалов имеет вид, приведенный на рисунке 5.

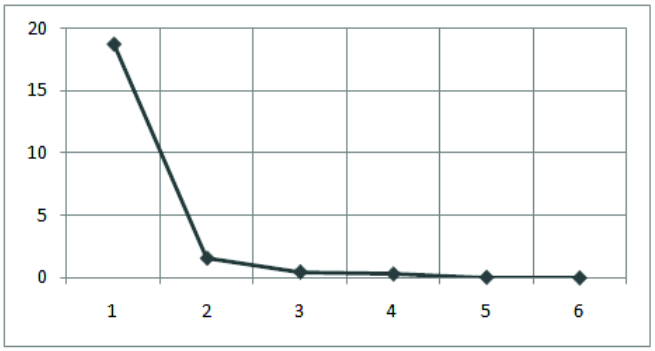

Рисунок 5 – График уменьшения ошибки восстановления с ростом количества интервалов

Нетрудно заметить, что стабилизация кластеров происходит на двух интервалах. То есть первичный кластер разбивается на 2 подкластера: “e” и “a”.

Таким образом, символы S_k, у которых характеристики $\widetilde{a}_k$ попадают в один интервал, образуют кластеры K_v.

Заметим, что для того, чтобы полученные кластеры были финальными, должно выполняться условие (2). В большинстве случаев новые кластеры будут удовлетворять этому условию, но, тем не менее, может оказаться часть новых кластеров ему не соответствует. Это

может случиться в случае, если первоначальный кластер состоял более чем из 3-х разнородных видов символов и построенный шаблон оказался слишком общий. В этом случае для кластеров, которые не удовлетворили условию (2) процедуру кластеризации с помощью метода главных компонент следует повторить. В целом, можно предложить альтернативный способ остановки, например, продолжать разбивать новые кластеры с помощью метода главных компонент до тех пор, пока на очередном шаге ни один кластер не будет разбит на подкластеры.

Выводы

Описанный кластеризации позволяет корректировать ошибки, которые сделаны на этапе первичной кластеризации и делает процесс разделения символов менее чувствительным к первоначальному выбору характеристик, тем самым делая его более универсальным и подходящим для разных шрифтов и языков.

Вместе с тем, предложенный алгоритм рекомендуется использовать именно на финальном этапе в качестве уточнения, поскольку применение его на оригинальном, первичном множестве символов приведет к долгому процессу стабилизации и, как следствие, к большому времени расчета кластеров.

Выражаю благодарность своему научному руководителю Шумейко Александру Алексеевичу за полезное обсуждение и советы.

ЛИТЕРАТУРА

1. Specification of DJVu Image Compression Format / AT&T .– 1999 .– 39 p.
2. Information technology Jpeg2000 Image Coding System. Final Committee Draft .– 2006 .– 205 p.
3. J. Mantas.An Overview of Character Recognition Methodologies.Pattern Recognition, Vol. 19, No 6, p. 425-430, 1986.
4. S. Kahan, T. Pavlidis & H. S. Baird.On the Recognition of Printed Characters of Any Font and Size.IEEE T-PAMI, Vol. 9, No.2, p. 274-288, March 1987.
5. Шамис А.Л. Принципы интеллектуализации автоматического распознавания изображений и их реализация в системах оптического распознавания символов. / А.Л. Шамис // Новости искусственного интеллекта .– 2002. – №1 .– С. 27-30.
6. Выбор признаков для распознавания печатных кириллических символов / И.А. Багрова, А.А. Грицай, С.В. Сорокин и др. // Вестник Тверского Государственного Университета .– 2010 .– 28 .– С. 59-73.
7. Лигун А.О. ALLDocument – технологія нового покоління для збереження, передачі та відображення електронних документів. / А.О. Лигун, О.О.

Шумейко, Д.В. Тимошенко // Вісник Східноукраїнського національного університету імені Володимира Даля .– №9(103), Частина 1 .– 2006 .– С. 83-85.
8. Лигун А.А. О локализации и формировании символов в электронных растрових документах со сложным фоном. / А.А. Лигун, А.А. Шумейко, Д.В. Тимошенко // Системные технологии. Региональный межвузовский сборник научных трудов .– Днепропетровск, 2008 .– № 1(54) .– С. 13-24.
9. Лигун А.А. Кластеризация символов в электронных растрових документах / А.А. Лигун, А.А. Шумейко, Д.В. Тимошенко // Вісник Східноукраїнського національного університету імені Володимира Даля .– 2008 .– №8 (126), Частина 1 .– С. 111-117.
10.Gray R. Quantization. / R. Gray, D. Neuhoff // IEE Transactions on Information Theory .– 1998 .– 44(6) .– P. 1-63.
11.Зиновьев А.Ю. Визуализация многомерных данных. / А.Ю. Зиновьев .– Красноярск : Изд. КГТУ, 2000. – 168 с.