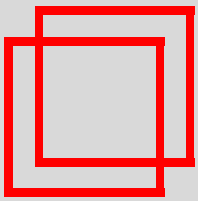

Math-Net.Ru
All Russian mathematical portal

A. A. Talalaev, I. P. Tishchenko, M. V. Khachumov, Allocation and clustering of text and graphic units on grayscale images,
Artificial Intelligence and Decision Making, 2008, Issue 3, 72–84

https://www.mathnet.ru/eng/iipr576

Use of the all-Russian mathematical portal Math-Net.Ru implies that you have read and agreed to these terms of use
https://www.mathnet.ru/eng/agreement

Download details:
IP: 18.97.14.84
May 21, 2025, 18:42:13

А.А. Талалаев, И.П. Тищенко, М.В. Хачумов

Выделение и кластеризация текстовых и графических элементов на полутоновых снимках

Аннотация Рассмотрена задача интеллектуального анализа документов, представленных в виде снимков, содержащих как текстовую (буквы, цифры), так и графическую части (рисунки, фотографии). Показано, что задачи выделения и кластеризации текстовой и графической информации в подобных документах могут решаться, несмотря на имеющиеся различия, с применением одних и тех же инструментальных средств, в том числе искусственных нейронных сетей (ИНС). Данный анализ рассматривается как первый шаг в решении общей задачи кластеризации составных документов на основе ИНС. Открытым остается вопрос о технологии обработки документов, представленных в различных форматах.

Ключевые слова: полутоновый снимок, графический объект, буквы, текст, шум, эталон, фильтрация, кластеризация, искусственная нейронная сеть, комитет.

Введение

Поиск заданных графических объектов по базам полутоновых снимков, как и поиск и кластеризация текстовых документов по ключевым словам в Интернет, являются актуальными самостоятельными задачами, которыми активно занимаются как в России, так и за рубежом. Большое внимание уделяется достижению приемлемого качества кластеризации, для чего могут быть привлечены самые разнообразные подходы [1-3]. Современные документы в электронных базах данных и ресурсах Интернета могут содержать, помимо фрагментов на естественном языке графические элементы, которые также могут быть использованы для кластеризации текстов. Это предопределяет постановку новых «смешанных» задач поиска и кластеризации текстовых документов совместно по текстовым и графическим данным, которые пока еще недостаточно изучены. При определенных допущениях указанные задачи, как раздельные по текстам и снимкам, так и «смешанные» могут быть унифицированы и сведены к формализованному анализу входных векторов признаков. С подобным анализом хорошо справляются нейронные сети, например ИНС Кохонена. Несмотря на то, что задачи и общие методы кластеризации достаточно хорошо исследованы, до сих остаются нерешенными проблемы, связанные с выбором числа кластеров и достижением требуемого качества кластеризации, как для текстов, так и для изображений [2,3]. Некоторые вопросы кластеризации с применением ИНС, рассмотрены в работах [4-7]. Как показывают эксперименты, решение задач поиска, классификации и кластеризации объектов на реальных снимках с помощью ИНС вызывает значительные сложности [8]. Использование различных ИНС и их комитетов, к сожалению, не гарантирует отсутствия ошибок. В особенности это касается случаев, когда наряду с выделенными объектами, представляющими интерес, на вход ИНС подается «шум». Трудности могут быть преодолены привлечением дополнительных методов обработки, способных отсекать шум на входе или выходе распознающей нейронной сети. Технология обработки снимков содержит, как правило, два последовательных этапа. Первый направлен на выделение и нормализацию потенциальных объектов. Второй – связан с кластеризацией и идентификацией объектов на основе ИНС. Аналогичная ситуация имеет место при кластеризации текстов на естественном языке, которая также требует выполнения этой

процедуры в два или три этапа [9]. В данной работе делается попытка проанализировать указанные направления кластеризации текстов и изображений. Ее авторы пока не готовы решить «смешанную» задачу для документов, представленных в различных форматах. Однако совершенно очевидна перспективность исследования такого «промежуточного» случая, когда текст и графика представлены одновременно на снимке. Строго говоря, в этом случае все классы объектов являются графическими, но каждый из них имеет определенные особенности. На стадии анализа подобных документов могут в равной степени использоваться одни и те же приемы выделения и распознавания объектов, а затем и кластеризации. Авторы надеются, что проведенные эксперименты, связанные с построением двухэтапной схемы кластеризации, позволят приблизиться к решению задачи обработки гипертекстовых документов.

1. Постановка задачи

Суть двухэтапного подхода к обработке текста и графики на полутоновых снимках заключается в следующем. ***На первом этапе*** производится выделение локальных объектов (букв, лиц, зданий, самолетов и т.д.) и их первичная фильтрация. Предполагается, что задача выделения и нормализации локальных объектов на снимке решена, например методами, описанными в работе [10]. В настоящей работе выделенные графические образы, подвергаются фильтрации с целью уменьшения числа объектов типа «шум». Отбор и предварительная кластеризация осуществляются на основе парного сравнения объектов с эталонами объектов и «шума» по серии признаков с использованием коэффициента корреляции. Основными инструментами для формирования и обработки векторов признаков применительно к изображениям являются ставшие классическими – аппарат ДПФ, гистограммы, контура объектов, инвариантные моменты. Первоначальная кластеризация на первом этапе позволяет проанализировать и существенно сократить количество признаков и выделенных объектов типа «шум». ***На втором этапе*** к выделенным объектам применяется кластеризация на основе ИНС. Традиционно используются сеть Хемминга и особым образом настроенная сеть Кохонена, реализующая алгоритм k-ближних соседей [11]. Выбор числа и первоначального размещения кластеров может осуществляться на основе принципа иерархической кластеризации. На данном этапе методы кластеризации текстов и графических образов в формализованной постановке имеют много общего, что может быть учтено при кластеризации гипертекстов.

2. Выбор информативных параметров и кластеризация графических образов

Кластеризация графических образов выполняется в два этапа. На первом производится сжатие признакового пространства и ограничение числа выделенных объектов класса «шум». На втором выполняется кластеризация нейронной сетью Хемминга или Кохонена.

2.1. Выбор информативных параметров и сокращение пространства признаков

В качестве исходных изображений использовались снимки со спутников. Пример такого снимка дан на Рис.1. Ставилась задача выделения летательных объектов (самолетов) и их классификации в соответствии с эталонами.

Для проведения вычислительного эксперимента использовались следующие компоненты:

1) множество выделенных объектов, включая объекты - «шумы» (фон, самолеты с нарушенными пропорциями, части самолетов) и полноценные изображения самолетов. Для выделения объектов использовалась методика, описанная в работе [10], основанная на принципе распространения волны. Примеры выделения и нормализации показаны на Рис.2;

2) набор эталонных изображений, представленных на Рис.3;

3) множество эталонов "шума", представленных на Рис.4.

Все изображения нормируются по технологии: авторазворот (приведение к стандартной ориентации), изменение до стандартного размера 128x128. Для множества выделенных на космическом снимке объектов, включая самолеты (с указанием типа), так и ошибочно принятые за самолет фоновые области и эталоны составляются таблицы парных сравнений. Таблицы содержат корреляционные коэффициенты Пирсона, полученные на основе сравнения:

1) матриц пикселей;
2) семи инвариантных моментов Ху;
3) коэффициентов ДПФ;
4) гистограмм;
5) геометрических разверток контуров.

Таблицы служат в качестве исходной информации для проведения более глубокого анализа. Результаты эксперимента и анализа применительно к сравнению объектов на основе ДПФ отражены в Табл.1.

Рис.1. Пример космического снимка

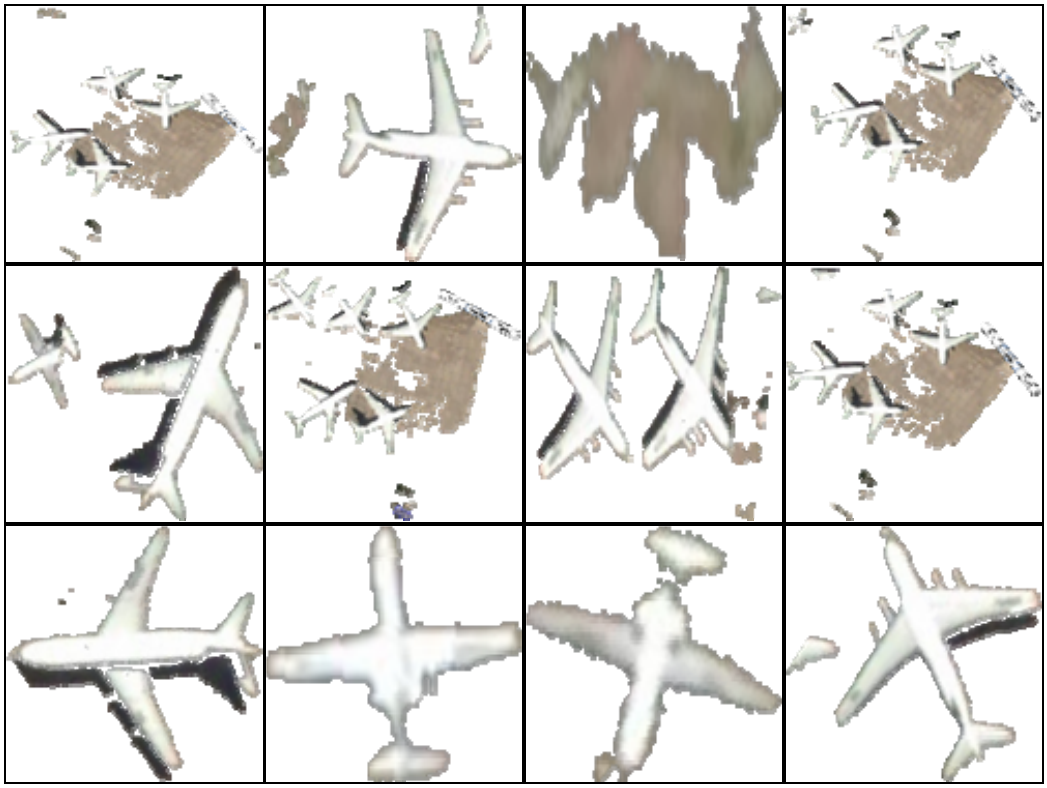


Рис.2. Примеры найденных и нормализованных объектов

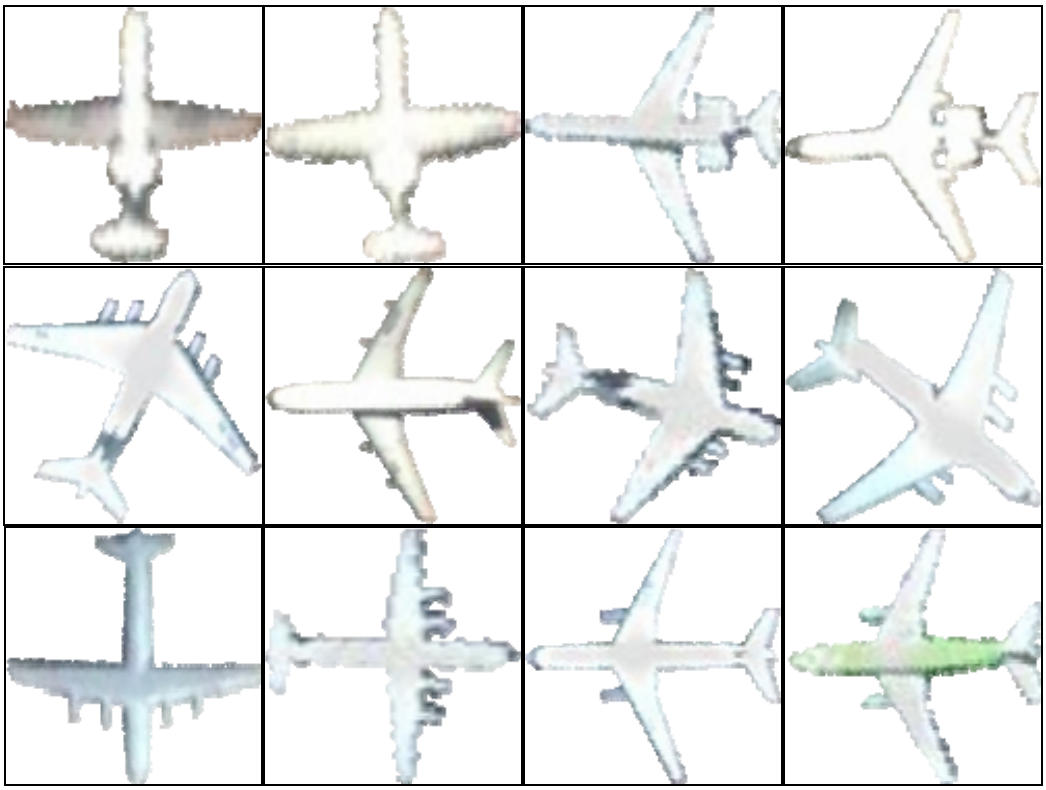


Рис.3. Набор эталонов самолетов

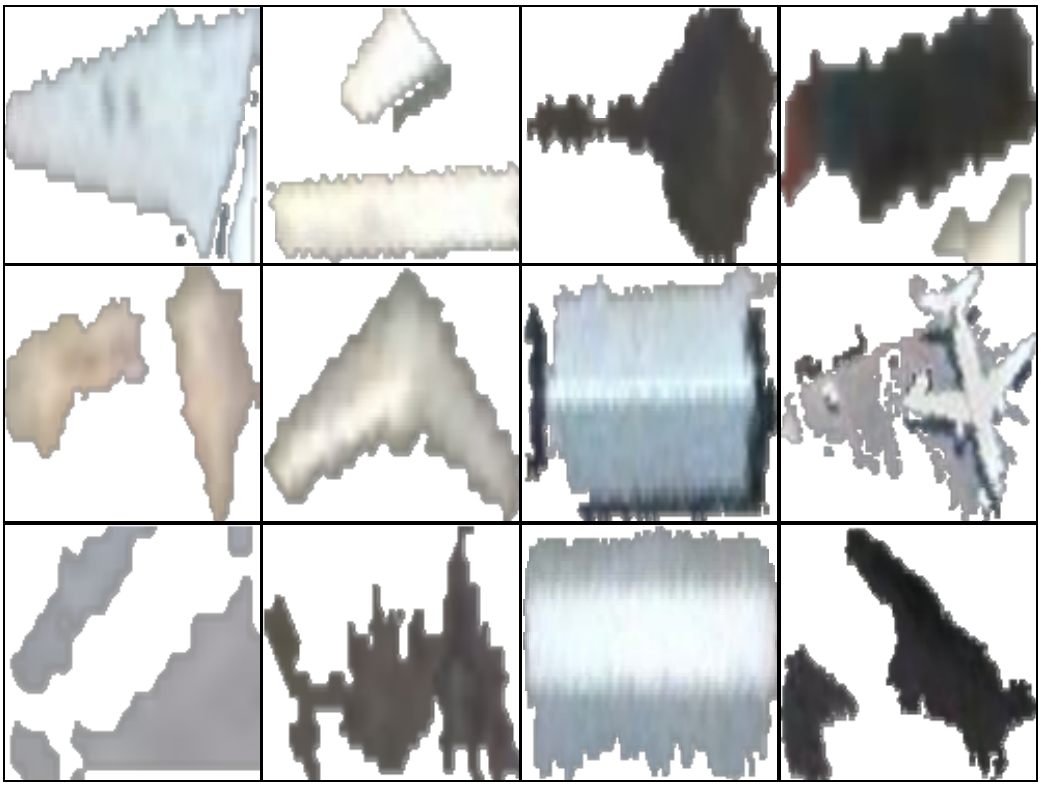


Рис 4. Эталоны «шума», выделенные на различных изображениях

Табл.. 1. Результаты сравнения выделенных объектов по коэффициентам ДПФ

Номер объекта	Тип выделенного объекта	Средняя корреляция с эталонами	Средняя корреляция с "шумом"	Максимальная корреляция с эталоном	Максимальная корреляция с "шумом"
1	Объект	0.739	0.765	0.782	0.826
2	Объект	0.746	0.777	0.785	0.826
3	Объект	0.760	0.771	0.796	0.804
4	Объект	0.754	0.782	0.793	0.839
5	Объект	0.722	0.713	0.768	0.771
6	Объект	0.720	0.747	0.751	0.790
7	Объект	0.695	0.692	0.737	0.739
8	Объект	0.697	0.713	0.734	0.767
9	Объект	0.752	0.781	0.798	0.849
10	Объект	0.722	0.729	0.761	0.764
11	Объект	0.687	0.712	0.745	0.760
12	Объект	0.672	0.693	0.740	0.760
13	Объект	0.691	0.708	0.720	0.746
14	Объект	0.676	0.696	0.719	0.745
15	Объект	0.721	0.699	0.761	0.781
16	Объект	0.731	0.735	0.768	0.784
17	Объект	0.727	0.757	0.770	0.799
18	Объект	0.786	0.792	0.836	0.848
19	Объект	0.690	0.675	0.718	0.755
20	Объект	0.695	0.697	0.732	0.766
21	Объект	0.705	0.688	0.730	0.756
22	Объект	0.692	0.685	0.738	0.760
23	Объект	0.783	0.777	0.815	0.833
24	Объект	0.753	0.785	0.792	0.826
25	Объект	0.702	0.693	0.734	0.762
26	Объект	0.700	0.713	0.737	0.769
27	Объект	0.732	0.743	0.763	0.774
28	Объект	0.722	0.735	0.758	0.782
29	Объект	0.733	0.728	0.788	0.772
30	Объект	0.762	0.754	0.799	0.810
31	Объект	0.706	0.692	0.735	0.756
32	Объект	0.694	0.663	0.733	0.746
33	Объект	0.721	0.723	0.754	0.787
34	Объект	0.700	0.678	0.735	0.758
35	Объект	0.769	0.770	0.808	0.819
36	Объект	0.777	0.773	0.835	0.820
37	Объект	0.721	0.706	0.749	0.768
38	Объект	0.705	0.679	0.736	0.754
39	Объект	0.758	0.752	0.804	0.818
40	Объект	0.714	0.692	0.743	0.765
41	Объект	0.711	0.698	0.735	0.767
42	Объект	0.702	0.685	0.734	0.757
43	Объект	0.705	0.695	0.732	0.771
44	Объект	0.741	0.748	0.770	0.838
45	Объект	0.799	0.798	0.830	0.845

Номер объекта	Тип выделенного объекта	Средняя корреляция с эталонами	Средняя корреляция с "шумом"	Максимальная корреляция с эталоном	Максимальная корреляция с "шумом"
46	Объект	0.812	0.802	0.838	0.832
47	Объект	0.692	0.728	0.722	0.775
48	Объект	0.800	0.786	0.831	0.825
49	Самолет	0.789	0.762	0.833	0.813
50	Самолет	0.803	0.785	0.838	0.824
51	Самолет	0.828	0.815	0.854	0.848
52	Самолет	0.816	0.808	0.848	0.850
53	Самолет	0.820	0.798	0.846	0.830
54	Самолет	0.777	0.756	0.815	0.806
55	Самолет	0.765	0.738	0.819	0.786
56	Самолет	0.772	0.746	0.828	0.796
57	Самолет	0.774	0.744	0.824	0.788
58	Самолет	0.770	0.741	0.828	0.786
59	Самолет	0.770	0.740	0.801	0.803

Табл. 2. Результаты сравнения информативности признаков

Способ формирования вектора признаков	Полезная фильтрация объектов	Ложное отбрасывание объектов	Отнесение «шума» к самолету
Пиксели изображения	2, 3, 6, 9, 29, 35, 39, 48		1, 7, 8, 10, 16, 25-28, 34, 42, 46
Пиксели развертки контура	1-4, 9, 10, 11, 13, 14, 17, 36, 39, 44,46	50, 54,59	6-8, 12, 15, 18, 20, 22, 24, 26, 28, 30, 31, 34, 37, 40,41
Инвариантные моменты Ху	9,18,19,20,23,29,36, 44	52, 54,59	
Гистограммы изображений	5-17, 26-30, 35,36, 39, 44, 47,48	50, 53, 55-58	1-4, 19-22 , 24-25, 31, 32, 34, 37, 38, 40-43, 45, 46
Коэффициенты ДПФ	1-4, 6, 8-14, 16-18, 20,24, 26-28, 33, 35, 44, 47		29, 36, 46, 48

Всего в таблице представлено 59 объектов, из которых 48 относятся к «шуму», 11 - к самолетам. Ошибки фильтрации в таблице выделены серым тоном. Их на первый взгляд достаточно много. Однако фильтрация не приводят к отбрасыванию целей, существенно сокращая число кандидатов для последующего анализа. Аналогичные исследования проводились и по другим признакам. В Табл.2. содержатся сравнительные результаты кластеризации по различным признакам. Цифрами в табл. 2 обозначены порядковые номера объектов из Табл.1.

Наилучшими показателями качества, как это следует из Табл. 2, обладает кластеризация на основе коэффициентов ДПФ. Она была выбрана в качестве основного метода предварительной фильтрации. Наличие данного механизма позволяет отбросить без потерь часть графических объектов до их подачи на ИНС.

2.2. Кластеризация на основе нейронной сети Хемминга

Следующим (заключительным) этапом обработки является кластеризация на основе ИНС. В эксперименте использовалась сеть Хемминга без фильтрации и с предварительной фильтрацией на основе ДПФ. Тестирование проводилось на шести космических снимках с использованием программной системы распознавания образов «ПС ИНС» [8], разработанной в ИПС РАН. Полученные результаты сведены в Табл. 3.

Эксперимент по системе «ДПФ+Хемминг» дает несколько лучший результат (Рис. 5), что определяет целесообразность включения данной схемы в технологическую цепочку обработки снимков.

Табл. 3

Тип обработки	Всего распознано объектов	Самолетов	«Шума»	Распознано корректно
Хемминг	60	37	23	13
ДПФ + Хемминг	45	34	11	12

Рис.5. Результаты кластеризации по Хеммингу

На Рис.5 видны ошибки кластеризации, что реально отражает в целом ограниченные возможности нейронных сетей при обработке сложных космических снимков.

3. Особенности выделения букв и кластеризации текстов

В соответствии с предлагаемой технологией на снимке выделяются буквы как соответствующие графические элементы, которые распознаются сетью Хемминга и сохраняются в виде обычного текста. Далее применяются специальные методы кластеризации текстов, в том числе и на основе ИНС.

3.1. Выделение и распознавание букв на полутоновых снимках

Для проведения экспериментов был выбран фрагмент научной статьи, представленной в виде полутонового снимка размером 3714x2166 пикселей (Рис.6).

Описываемый эксперимент решал задачу выделения и распознавания букв с использованием комплекса типовых модулей «ПС ИНС». Алгоритм обработки представлен в графическом виде на Рис.7. Он содержит следующие основные этапы анализа снимка.

1. Подготовка эталонных изображений и обучение ИНС:
   - Модуль ReadEtalons производит чтение исходных данных (эталонных изображений) и передает их модулю Resize1.
   - Модуль Resize1 изменяет размеры полученных эталонных изображений (подготавливая данные для ИНС) и передает их модулю Net в канал InputLearn.
   - Модуль Net получает все эталонные изображения из канала InputLearn и производит обучение ИНС.
2. Поиск и распознавание объектов:
   - Модуль ReadMap считывает изображения, на которых необходимо производить поиск объектов и передает их модулю MagicWand
   - Модуль MagicWand получает изображения, удаляет на них фон и передает данные модулю FindObjects
   - Модуль FindObjects получает изображения, выделяет объекты на изображении и передает их модулю Resize2.

АНАЛИЗ ЭФФЕКТИВНОСТИ ПРИМЕНЕНИЯ ИСКУССТВЕННЫХ НЕЙРОННЫХ СЕТЕЙ ДЛЯ РЕШЕНИЯ ЗАДАЧ РАСПОЗНАВАНИЯ, СЖАТИЯ И ПРОГНОЗИРОВАНИЯ

Талалаев А.А., Тищенко И.П., Фраленко В.П., Хачумов В.М.[1]

Аннотация. В статье обобщен опыт, накопленный в результате создания прототипа программной системы «ПС ИНС» для распознавания графических образов на основе искусственных нейронных сетей (ИНС). Показаны реальные возможности нейронных сетей на примерах решения задач распознавания, сжатия данных и экстраполяции. Приведены результаты экспериментальных исследований по ускорению решения задач на кластерном вычислителе.

Ключевые слова: *Искусственная нейронная сеть, алгоритм, распознавание, графический образ, эффективность, параллельные вычисления, кластерное вычислительное устройство.*

Рис. 6. Исходное изображение с текстом

- Модуль Resize2 изменяет размеры найденных изображений (подготавливая данные для ИНС) и передает их модулю Net в канал InputRecognize.
- Модуль Net получает изображения из канала InputRecognize и в случае, если к этому моменту обучение нейронной сети завершено, производит распознавание каждого из изображений, формируя структуры описывающие результаты распознавания. Эти структуры передаются модулю RecognitionFilter.
- Модуль RecognitionFilter производит фильтрацию результатов распознавания ИНС, удаляя результаты с малой вероятностью распознавания и возможные дубликаты. Обработанные данные передаются модулям Writer и TextWriter
- Модуль Writer получает результаты распознавания и сохраняет их в файле формата xml.
- Модуль TextWriter получает результаты распознавания и преобразует их в текстовый файл.

Для распознавания использовался набор из 262 эталонных изображений (букв русского алфавита и двух подклассов для распознавания символов ‘**.**’ и ‘**,**’).

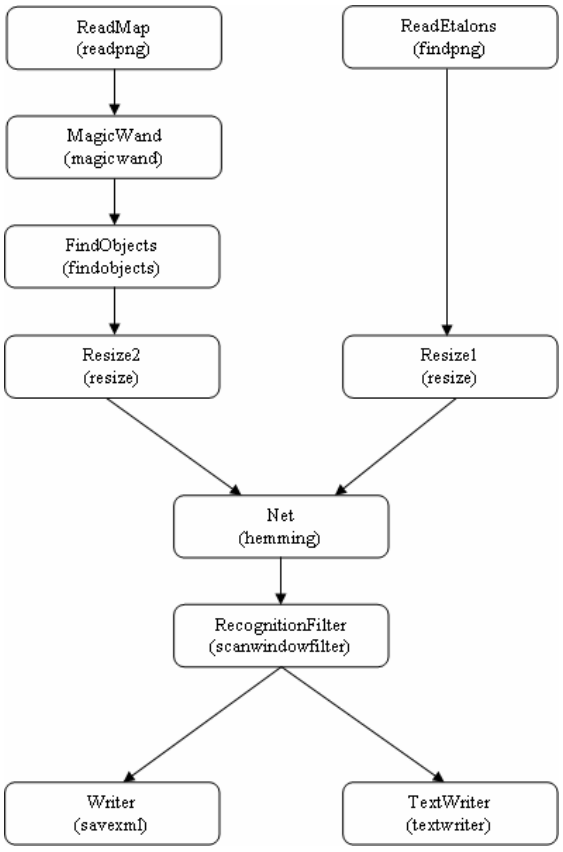
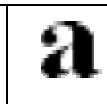
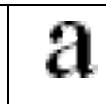
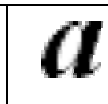
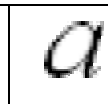
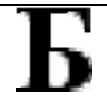
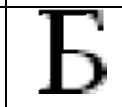
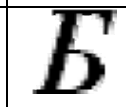
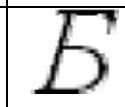
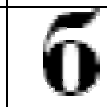
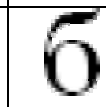
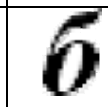
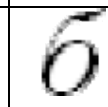
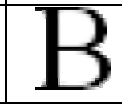
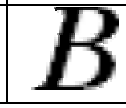
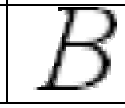
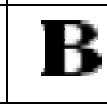
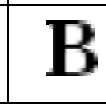


Рис.7. Графическая схема алгоритма работы с текстом

Табл. 4. Примеры эталонных изображений

Буква	Эталоны							
А								
Б								
В								

Результаты распознавания, полученные в виде текстового файла, представлены ниже.

АНАЛИЗ ЭФФЕКТИВНОСТИ ПРИМЕНЕНИЯ ИСКУССТВЕННЫХ НЕИРОННЫХ СЕТЕИ ДЛЯ РЕШЕНИЯ ЗАДАЧ РАСПОЗНАВАНИЯ, СЖАТИЯ И ПРОГНОЗИРОВАНИЯ
ТАЛАЛАЕВ А.А., ТИЩЕНКО И.П., ФРАЛЕНКО В.П., ХАЧУМОВ В.М.
АННОТАЦИЯ. В СТАТЬЕ ОБОБЩЕН ОПЫТ, НАКОПЛЕННЫИ В РЕЗУЛЬТАТЕ СОЗДАНИЯ ПРОТОТИПА ПРОГРАММНОИ СИСТЕМЫ ПС ИНС ДЛЯ РАСПОЗНАВАНИЯ ГРАФИЧЕСКИХ ОБРАЗОВ НА ОСНОВЕ ИСКУССТВЕННЫХ НЕИРОННЫХ СЕТЕИ ИНС ПОКАЗАНЫ РЕАЛЬНЫЕ ВОЗМОЖНОСТИ НЕИРОННЫХ СЕТЕИ НА ПРИМЕРАХ РЕШЕНИЯ ЗАДАЧ РАСПОЗНАВАНИЯ, СЖАТИЯ ДАННЫХ И ЭКСТРАПОЛЯЦИИ ПРИВЕДЕНЫ РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТАЛЬНЫХ ИССЛЕДОВАНИИ ПО УСКОРЕНИЮ РЕШЕНИЯ ЗАДАЧ НА КЛАСТЕРНОМ ВЫЧИСЛИТЕЛЕ
КЛЮЧЕВЫЕ СЛОВА. ИСКУССТВЕННАЯ НЕИРОННАЯ СЕТЬ, АЛГОРИТМ, РАСПОЗНАВАНИЕ, ГРАФИЧЕСКИИ ОБРАЗ, ЭФФЕКТИВНОСТЬ, ПАРАЛЛЕЛЬНЫЕ ВЫЧИСЛЕНИЯ, КЛАСТЕРНОЕ ВЫЧИСЛИТЕЛЬНОЕ УСТРОИСТВО.

Как видно из эксперимента все буквы достаточно уверенно выделены и распознаны правильно. Для сравнительной оценки качества распознавания тот же снимок был обработан с использованием программы FineReader 8 Professional Edition. В результате было получено:

АНАЛИЗ ЭФФЕКТИВНОСТИ ПРИМЕНЕНИЯ ИСКУССТВЕННЫХ
НЕЙРОННЫХ СЕТЕЙ ДЛЯ РЕШЕНИЯ ЗАДАЧ РАСПОЗНАВАНИЯ,
СЖАТИЯ II ПРОГНОЗИРОВАНИЯ
Талалаев А.А., Тищенко II.П., Фраленко В.П., Хачумов В.М.1
Аннотация. В статье обобщен опыт, накопленный в результате создания
прототипа программной системы «ПС ИНС» для распознавания графических образов на основе
искусственных нейронных сетей (ИНС). Показаны реальные
возможности нейронных сетей на примерах решения задач распознавания,
сжатия данных и экстраполяции. Приведены результаты экспериментальных исследований по ускорению решения задач на кластерном вычислителе.
Ключевые слова: Искусств в иная нейронная сеть, алгоритм, распознавание, гр а ф и ч е скы й о бр а з, э ф ф е к т и внос т ъ, п ар алл елъные вычисления, кла с т ерное в ычи ел \ I т елъное ус тр о й с т в о.

Видно, что программа FineReader в ряде случаев ошибается при распознавании некоторых букв, например заглавной буквы «И» и букв, набранных курсивом (например, *“ь”, “е”, “и”*). Однако нейросетевое распознавание при прочих равных условиях пока проигрывает по времени программе FineReader из-за использования трудоемких алгоритмов предобработки.

Не представляет принципиальной трудности выделение и распознавание не только букв, но и отдельных слов и предложений, но в данной статье эти операции не рассматриваются.

3.2. Особенности кластеризации текстовых документов

Выделенные на снимках буквы и слова образуют единый текстовый документ, который вместе с множеством других выделенных объектов такого рода проходит дальнейшую обработку в соответствии с выбранным типовым методом кластеризации. В Табл. 5 показаны основные методы кластеризации текстовых документов и их характеристики.

Открытыми для исследования остаются вопросы:

а) выбора начального числа кластеров и их размещения,

б) построения признакового пространства и формирования векторов признаков,

в) определения расстояний и нормирования векторов.

Основными признаками при векторизации документов являются веса слов и фраз, входящих в состав документа, при подсчете которых используется частота их появления в документе, обучающей выборке, а также данные глоссария. Размерность пространства признаков определяется числом возможных слов и фраз. Структура формируемого вектора признаков, в принципе, полностью зависит от разработчика системы кластеризации. Например, в модели терм – документ, принятой в работе [6], текст описывается лексическим вектором $\{\tau_i\}$, $i = 1,\ldots N_w$, где τ_i – важность (информативный вес) термина w_i в документе, N_w – полное количество терминов в словаре. Вес термина, отсутствующего в документе, принимается равным нулю. Для удобства веса нормируются, так что $\tau_i \in [0,1]$.

В нашем эксперименте вектор признаков содержал:

- массив ID-слов,
- массив весов слов,
- массив частот слов,
- массив числа документов кластера, содержащих эти слова,
- массив ID-фраз,
- массив весов фраз,
- массив частот фраз,
- массив числа документов кластера, содержащих эту фразу.

Важный вопрос – выбор меры близости векторов. В данном эксперименте для сравнения векторов одновременно использовались две меры близости: S_1- мера по ID-словам и S_2 - мера по ID-фразам, а также определялось их пересечение. С учетом сказанного двухэтапный алгоритм кластеризации текстов принимает следующий вид.

На первом этапе выполняется процедура формирования начальных кластеров на основе сочетания иерархического подхода и метода k-means [9]. Стандартная процедура иерархической кластеризации для ограниченной выборки документов хорошо известна. Она основана на методе парного сравнения текстов по сформированным векторам признаков и последовательного объединения наиболее близких пар до образования необходимого числа кластеров. Далее может быть подключен алгоритм k-means, который завершает процедуру формирования кластеров с использованием расширенной обучающей выборки. В основе алгоритма k-means лежит итеративный процесс стабилизации центроидов кластеров.

Табл.5. Сводка основных характеристик алгоритмов кластеризации текстов

Название метода	Наличие пересечения кластеров	Используемые числовые характеристики документов	Предварительное обучение	Оценка сложности работы (N-число документов, k-число кластеров)
LSI	-	Tfidf	-	N^2 k, (N=terms+docs, k-factors)
STC	+	-	-	$O(k^2 N)$
Single Link, Complete Link, Group Average	-	Similarity matrix	-	Single Link ~ $O(N^2)$ Complete Link ~ $O(N^3)$ Group Average ~ $O(N^2)$
Scatter/Gather	-	Similarity matrix	-	Buckshot ~ O(kN), Fractionation ~ O(mN),m=O(k)
K-means	-	Tfidf	-	O(N)
CI - необучаемый вариант	-	Similarity matrix	-	O(N*log k)
CI - обучаемый вариант		Similarity matrix или tfidf	+	-
SOM (сеть Кохонена)	+	Similarity matrix или tfidf	+	-

На втором этапе выполняется кластеризация всех текстовых документов по принципу отнесения каждого входного документа к тому или иному классу в зависимости от близости к сформированным кластерам. В Табл. 6 представлены результаты эксперимента по кластеризации текстовых документов на типовом ПК с формированием заданного числа кластеров на основе алгоритма k- means. В эксперименте использовалось до 1000 документов учебной выборки при числе кластеров порядка 20.

Временные характеристики работы алгоритма k-means в сравнении с алгоритмом STC отражены на Рис. 8 [9].

Из графиков видно, что алгоритм k-means опережает по времени работы алгоритм STC в достаточно широком диапазоне анализируемых текстовых документов.

Табл. 6.

Количество документов	Кластеризация с использованием алгоритма K-means		
	Время формирования кластеров (сек)	Время распределения всех документов по кластерам (сек)	Общее время (сек)
500	8	7	15
1000	14	10	24
2000	12	16	28
5000	13	41	54
7500	12	63	75
11000	10	163	173
15000	15	246	261
20000	14	352	366

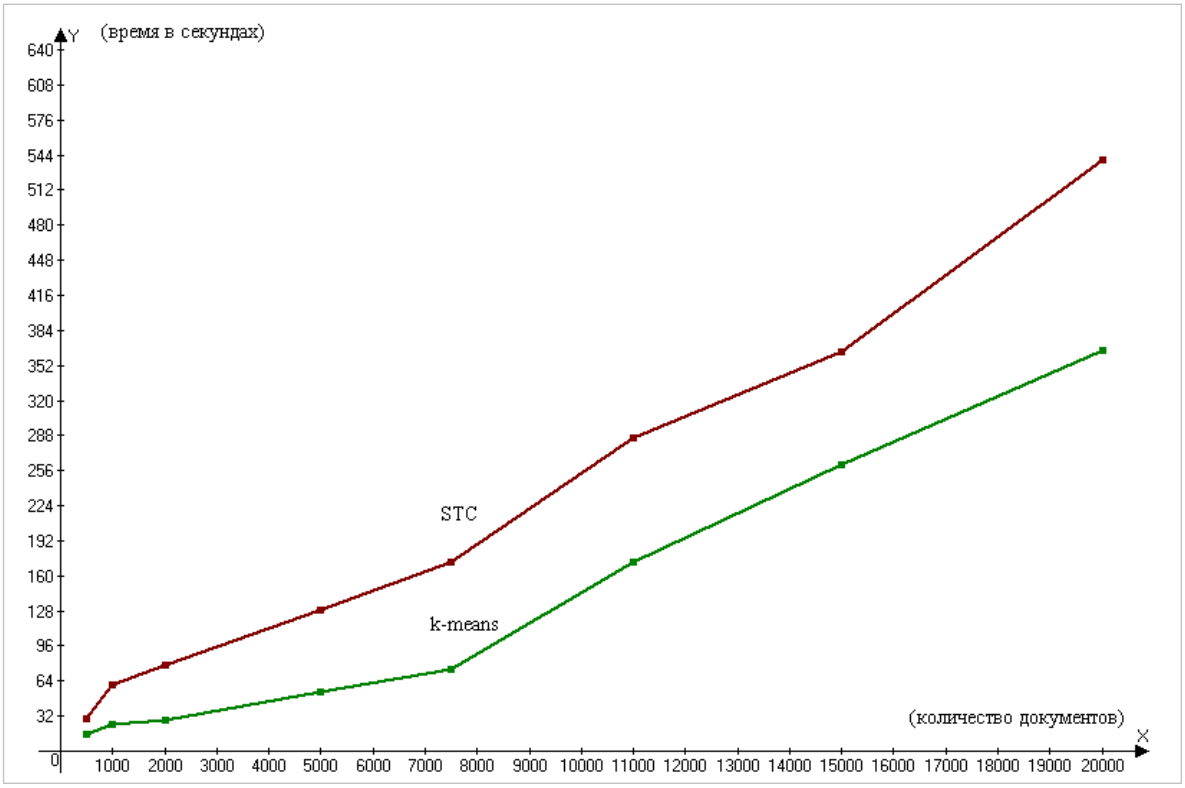


Рис. 8. Графики времени работы алгоритмов k-means и STC

4. Возможность применения ИНС для кластеризации текстов и изображений

В настоящее время крупные Интернет - компании внедряют алгоритмы индексирования и кластеризации на базе нейронных сетей, что улучшает релевантность отобранных документов к запросу. Примером служит система обработки естественного языка DISCERN, описание которой дается в работе [5]. С помощью этой системы можно подвергать текст реферативному сжатию, а также выполнять обратную процедуру. Задачи кластеризации документов можно решать, например, на основе единого аппарата сети Кохонена [12]. Свойства данной сети хорошо изучены как на задачах анализа текстов [4-6], так и на задачах анализа изображений [8,10, 12]. Объект подается на входы ИНС в виде вектора признаков $(x_1, x_{2,\ldots,} x_n)$ и далее анализируется нейронами $(c_1, c_{2,\ldots,} c_n)$, как это показано на Рис.9.

Начальная инициализация проводится, например, случайным распределением весовых векторов на гиперсфере единичного радиуса или в соответствии с распределением на основе предварительной иерархической кластеризации [9]. ИНС Кохонена целесообразно доверить реализацию алгоритма k-means. Настройка осуществляется путем коррекции весовых коэффициентов для всех нейронов, попадающих в зону соседства, в соответствии с алгоритмом Видроу-Хоффа или другим методом [2,3,11] до получения приемлемого качества кластеризации, удовлетворяющего пользователя, или до останова, предусмотренного алгоритмом настройки.

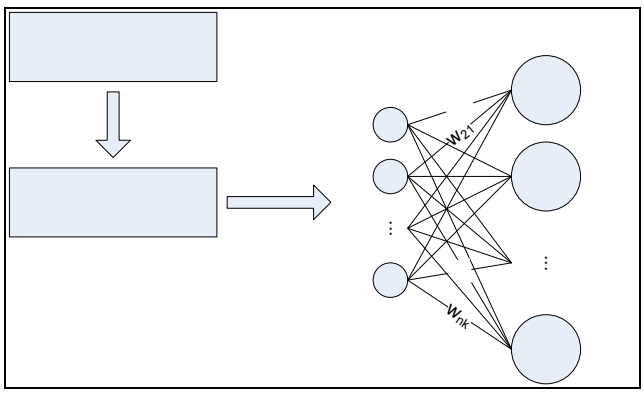


Рис. 9. Структура сети Кохонена для кластеризации изображений и текстов

Поскольку векторы, представляющие графические объекты, и векторы, формализующие тексты, все же значительно отличаются по структуре данных, то для кластеризации лучше использовать комитет нейронных сетей, часть из которых настраивается на анализ текстовой части, а другая на графические элементы. После обучения комитету ИНС можно доверить кластеризацию входных документов по комплексу выделенных признаков.

Заключение

В настоящей работе рассмотрены вопросы кластеризации документов, содержащих текстовые и графические фрагменты. Показано, что, несмотря на наличие особенностей в обработке текстов и изображений, возможно использование общих инструментов как на фазе выделения локальных объектов (букв и графических образов), так и на завершающей стадии кластеризации. На первом этапе предлагается решить задачи выбора наиболее информативных признаков и формирования первоначальных кластеров. Для распознавания объектов, в том числе букв, может быть использована нейронная сеть Хемминга. На втором этапе выполняется кластеризация документов с одновременным анализом их текстовых и графических частей, для чего целесообразно применять комитет ИНС Кохонена. Для достижения полноты исследования предлагаемый подход должен быть распространен на документы, представленные в различных типовых форматах. Очевидно, следует ожидать появления в ближайшее время поисковых систем, одновременно анализирующих все виды доступной информации, содержащейся в документах, что как можно надеяться, позволит повысить эффективность поиска и точность кластеризации.

Литература

1. Паклин Н. Алгоритмы кластеризации на службе Data Mining. http://www.basegroup.ru
2. Sloane N.J.A., Hardin R.H., Duff T.S., Conway J.H. Minimal-Energy Clusters. http://www.research.att.com/~njas/cluster/index.html
3. Hardin R.H., Sloane N.J.A., Smith W.D. Tables of Spherical Codes with Icosahedral Symmetry. http://www.research.att.com/~njas/icosahedral.codes/index.html

4.Борисов Е.С. Кластеризатор на основе нейронной сети Кохонена http://mechanoid.narod.ru/nns/kohonen/index.html
5. Яценко Д., Чернышев Ю.О. Перспективы применения нейронных сетей для построения поисковых систем в гипертекстовых распределенных системах http://itx.ru/articles/theory/neiro-search.html
6. Андреев А.М., Березкин Д.В., Морозов В.В., Симаков К.В. Автоматическая классификация текстовых документов с использованием нейросетевых алгоритмов и семантического анализа. http://www.inteltec.ru/publish/articles/textan/RCDL2003.shtml
7. Дударь З.В., Шуклин Д.Е. Семантическая нейронная сеть, как формальный язык описания и обработки смысла текстов на естественном языке. http://www.shuklin.com/ai/ht/ru/ai00001f.aspx
8. Талалаев А.А., Тищенко И.П., Фраленко В.П., Хачумов В.М. Анализ эффективности применения искусственных нейронных сетей для решения задач распознавания, сжатия и прогнозирования. - Искусственный интеллект и принятие решений, 2, 2008, с.24-33.
9. Хачумов М.В. Методы совершенствования алгоритмов кластеризации текстов. – Высокие технологии, фундаментальные и прикладные исследования, образование// Сборник трудов Четвертой международной научно-технической конференции «Исследование, разработка и применение высоких технологий в промышленности» (02-05.10.2007, Санкт-Петербург). – СПб.: Изд-во Политехнического университета, 2007, Т.11, с.135-136.
10. Виноградов А.Н., Калугин Ф.В., Недев М.Д., Погодин С.В., Талалаев А.А., Тищенко И.П., Фраленко В.П., Хачумов В.М. Выделение и распознавание локальных объектов на аэрокосмических снимках. – Авиакосмическое приборостроение, № 9, 2007, с.39-45.
11. Миркес Е.М. Нейроинформатика. Учебное пособие. – Красноярск: Издательство Красноярского государственного технического университета, 2003. http://www.softcraft.ru/neuro/ni/p00.shtml.
12. Круглов В.В., Борисов В.В. Искусственные нейронные сети. Теория и практика. – М.: Радио и связь. – 382 с.

Талалаев Александр Анатольевич. Институт программных систем РАН, м.н.с. В 2006 год окончил Университет г. Переславля им. А.К. Айламазяна. 8 печатных работ Область научных интересов: искусственный интеллект, машинная графика, распознавание образов, параллельные вычисления

Тищенко Игорь Петрович. Институт программных систем РАН, аспирант. В 2005 году окончил Университет г. Переславля им. А.К. Айламазяна. 4 печатные работы. Область научных интересов: искусственный интеллект, машинная графика, распознавание образов, нейронные сети, параллельные вычисления.

Хачумов Михаил Вячеславович. Российский университет дружбы народов, магистрант. 2 печатные работы. Область научных интересов: искусственный интеллект, машинная графика, кластеризация.