УДК 159.9.072.533

ГРНТИ 15.01.77

Самойлова Татьяна Аркадьевна
ФГБОУ ВО «Смоленский государственный университет», Смоленск, Россия
Доцент кафедры «Информатики»Кандидат технических наук, доцент
E-mail: tatsamoilova24@gmail.com
РИНЦ: https://elibrary.ru/author_profile.asp?id=100995

Грибер Юлия Александровна
ФГБОУ ВО «Смоленский государственный университет», Смоленск, Россия
Профессор кафедры «Социологии и философии»
Директор «Лаборатории цвета»[1]Доктор культурологии
E-mail: y.griber@gmail.com
ORCID: https://orcid.org/0000-0002-2603-5928
РИНЦ: https://www.elibrary.ru/author_profile.asp?id=303167
Researcher ID: https://www.researcherid.com/rid/AAG-4410-2019
SCOPUS: https://www.scopus.com/authid/detail.url?authorId=56809444600

Интеллектуальный анализ цветовых предпочтений:поиск ассоциативных правил vs. кластерный анализ

Аннотация. Цель статьи заключается в том, чтобы представить опыт экспериментальной реализации на основе современных программных платформ и технологий двух различных методов интеллектуального анализа данных – (1) метода поиска ассоциативных правил в ответах испытуемых и (2) метода кластеризация ответов. Авторы анализирую возможности и ограничения использования этих методов в социально-психологическом исследовании цветовых предпочтений. Материалом для проводимого эксперимента стали данные социально-психологического исследования, в ходе которого испытуемым (N = 50) показывали цветовую палитру, содержащую 27 различных оттенков, и просили выбрать из нее те цвета, которые, по их мнению, лучше всего подходили для интерьера каждого из семи различных типов помещений: гостиной, прихожей, спальни, ванной, туалета, кухни и коридора. Средствами алгоритма Apriori получены ассоциативные правила, соответствующие взаимосвязям между цветовыми предпочтениями и типами помещений. Рассмотрены особенности применения метода иерархической кластеризации для получения выводов, к которым невозможно прийти, вычисляя лишь процентные показатели. Выполнен выбор правил объединения кластеров, дающий наиболее объективную и информативную оценку ответов. Предложено рассчитывать расстояние между цветовыми выборами респондентов по метрике цветового отличия CIEDE2000. На конкретных примерах показано, что оба метода интеллектуального анализа открывают широкие возможности визуализации выявленных психологических механизмов и закономерностей. Проведены эксперименты, в ходе которых установлено, что выбранные методы позволяют проводить эффективную оценку материала социально-психологических исследований. Установлено, что при значительном увеличении числа испытуемых и количества возможных вариантов ответов, задача поиска ассоциаций эффективно решается параллельными методами.

Ключевые слова: методы интеллектуального анализа данных; ассоциативные правила; алгоритм Apriori; кластерный анализ; иерархическая кластеризация; язык Python; цветовые предпочтения

=

Введение

Методы интеллектуального анализа данных получают все большую популярность в социально-психологических исследованиях (см., напр.: [1–3]), что связано с ростом объема обрабатываемой в этой сфере информации и с тем, что из всего арсенала методов интеллектуального анализа подавляющее большинство хорошо подходят для анализа связей между номинальными данными, присущими психической жизни и социальным явлениям. Интеллектуальный анализ позволяет не только выявлять статистические закономерности, но и строить модели, способные объяснить редко встречающиеся явления (см. напр.: [4–6]). В случае нечисловых данных, где привычные статистические пакеты могут применяться лишь ограниченно, свободно распространяемые программные средства с дружественным графическим интерфейсом и богатыми возможностями визуализации полученных результатов, такие как язык Python, показывают себя гораздо эффективнее и в совокупности с библиотеками Pandas, Numpy, Scikit-learn и многими другими предоставляют все необходимые инструменты для интеллектуального анализа.

Практически все предыдущие исследования показывают необходимость дополнительных вычислительных экспериментов для определения в каждом конкретном случае оптимального метода интеллектуального анализа, свободного от субъективных предпочтений исследователя (см. напр.: [7; 8]).

Цель статьи заключается в том, чтобы представить опыт экспериментальной реализации на основе современных программных платформ и технологий двух различных методов интеллектуального анализа данных – (1) метода поиска ассоциативных правил в ответах испытуемых и (2) метода кластеризация ответов – и сравнить возможности и ограничения использования этих методов в социально-психологическом исследовании цветовых предпочтений.

Методы

(1) Методы поиска ассоциативных правил

Методы поиска ассоциаций все чаще используются в современных социально-психологических исследованиях для обнаружения скрытых зависимостей между признаками: например, для выявления связи между страной проживания и устоявшимися стереотипами [9], закономерностями потребительского поведения [1], стилем активности личности [2].

Каждое ассоциативное правило характеризуется определенными параметрами. Параметр, называемый поддержкой (support), показывает частоту встречаемости данного правила в имеющемся множестве транзакций. Поддержка правила $X \Rightarrow Y$ вычисляется как процент транзакций, содержащий множество $X \cup Y$:$\mathrm{supp}(X \Rightarrow Y) = (\mathrm{N}(X \cup Y) / |\mathrm{D}|) \cdot 100\ \%$, где $\mathrm{N}(X \cup \mathrm{Y})$ – количество транзакций, содержащих множество $X \cup Y$.

Достоверность ассоциативного правила (confidence) показывает, с какой вероятностью из X следует Y. Вычисляется достоверность правила $X \Rightarrow Y$ как процент транзакций, содержащих как X, так и Y в базе транзакций, содержащих X: $\mathrm{conf}(X \Rightarrow Y) = \mathrm{supp}(X \Rightarrow Y) / \mathrm{supp}(X)$.

Лифт (lift) – показатель, позволяющий оценить значимость правила. Он связывает поддержку и достоверность и определяется по формуле: $\mathrm{lift}(\mathrm{X} \Rightarrow \mathrm{Y}) = \mathrm{conf}(X \Rightarrow Y) / \mathrm{supp}(Y)$. Лифт характеризует меру связи ассоциации: если лифт меньше 1 – связь отрицательная, равно 1 – отсутствует, больше 1 – связь положительная и чем больше, тем сильнее.

Поиск ассоциативных правил состоит в определении импликаций наборов ответов респондентов, поддержка которых не ниже, чем minsupport. Затем из найденных наборов выделяются правила с достоверностью не ниже minconfidence и лифтом не ниже minlift. Выбор алгоритма поиска ассоциативных правил определяется, прежде всего, объемом входных

=

данных и числом возможных вариантов ответов.

(2) Метод кластеризации

Объектами для применения метода кластеризации в современных социально-психологических исследованиях являются числовые или номинальные признаки, корреляции, результаты статистической обработки и т. д. (см. напр.: [3; 6; 7]). Этот метод обычно используется для выявления структуры данных и определения нетипичных объектов.

Задача кластеризации результатов исследования заключается в том, чтобы так разбить выборку, представляющую собой ответы респондентов в виде строковых констант или числовые результаты статистической обработки по категориям ответов, на непересекающиеся подмножества – кластеры, чтобы каждый кластер состоял из наиболее похожих объектов, которые, в то же время, будут значительно отличаться от объектов остальных групп.

Известны несколько подходов к решению задачи кластеризации: нейросетевой, статистический, теоретико-графовый, иерархический. При обработке данных, имеющих небольшой объем, наиболее эффективным считается метод иерархической агломеративной кластеризации [8]. Его характеризует содержательная ясность и относительная простота алгоритмов, допустимость контролируемого вмешательства в работу алгоритма, возможность визуализации данных в виде дерева (дендрограммы). В общем виде алгоритм иерархической агломеративной кластеризации включает создание первичного набор кластеров Y, каждый из которых содержит один элемент набора кластеризации X; расчет матрицы сходства между кластерами; определение наиболее близких по метрике сходства кластеров и их объединение по определенному правилу.

(3) Материал исследования

Материалом для проводимого эксперимента стали данные социально-психологического исследования, в ходе которого испытуемым (N = 50) показывали цветовую палитру, содержащую 27 различных оттенков, и просили выбрать из нее те цвета, которые, по их мнению, лучше всего подходили для интерьера каждого из семи различных типов помещений: гостиной, прихожей, спальни, ванной, туалета, кухни и коридора. Ответы непосредственно задавались цветовым кодом из множества A1÷A9, B1÷B9, C1÷C9 (см. подробнее: [10]).

Таким образом, имеющаяся база данных состояла из 50 транзакций, каждая из которых представляла собой набор ответов одного респондента. Ответы были представлены строковыми константами, позволяющими находить повторяющиеся наборы ответов в разных транзакциях (табл. 1).

Таблица 1

Фрагмент данных из нескольких транзакций

№	Гостиная	Прихожая	Спальня	Ванная	Туалет	Кухня	Коридор
1	A5	A5	A4	C3	C3	B5	B1
2	B1	B1	B1	A5	A5	B1	B1
3	A8	B2	A2	A1	A1	A6	B5
4	B6	B3	B9	A8	A1	A5	B1
…	...	...	...	...	...	...	...
49	B5	B9	B6	B8	A1	A9	B7
50	A3	B2	B9	B2	A1	A2	B2

Составлено авторами на основе полученных в ходе эксперимента ответов респондентов

Результаты и обсуждение

(1) Поиск ассоциативных правил в цветовых предпочтениях респондентов

Поскольку в проведенном опросе участвовало небольшое число испытуемых, несмотря

=

на существенное число вариантов (27 выбираемых цветов х 7 видов помещений = 189) ответов, наиболее эффективным для решения поставленных в исследовании задач оказался алгоритм Apriori, который предназначен для поиска всех частых множеств признаков и является поуровневым, поскольку использует стратегию поиска «в ширину» и осуществляет его снизу-вверх [11].

Обработка данных опроса алгоритмом Apriori включала три этапа.

1. *Объединение* – просмотр всей базы данных транзакций и определение частоты вхождения отдельных ответов респондентов по каждому из помещений (одноэлементные наборы).
2. *Сокращение* – перевод наборов ответов, которые удовлетворяют поддержке и достоверности (кандидаты), в следующий раунд (двухэлементные наборы).
3. *Повторение* предыдущих двух этапов для каждой величины набора до тех пор, пока не будет повторно получен ранее определенный размер.

Результат работы алгоритма – объединение всех множеств k-элементных наборов ответов *X* для k = 1, 2, ..., n, где n – максимальное число элементов в наборе ответов (рис. 1).

Применение алгоритма позволяет обнаружить значимые двух-, трех- и четырехэлементные ассоциации цветов семи видов помещений (рис. 2). Значения параметров support, confidence и lift определяется оператором вызова алгоритма Apriori.

```
        #Листинг 1
        association_rules  = list(apriori(transactions, min_support = 0.04,
min_confidence=0.5, min_lift = 2))
        print ("РЕЗУЛЬТАТЫ РАБОТЫ Apriori")
        print('Трехэлементные значимые ассоциации')
        print('-----------------------------------------')
        print ("         Rule       supp(%) conf(%)   lift")
        print('-----------------------------------------')
        for item in association_rules :
           x = list(item.items)
           if (len(x) == 3):   #Отбираем ассоциации из трех элементов
              items = [x1 for x1 in item[0]]
              print( items[0] + " => " + items[1]  + " , "+items[2],
'\t',int(100*item.support),'\t', int(100* item[2][0][2]),'\t', round(item[2][0][3],2))
```

Рисунок 1. *Фрагмент листинга Python-программы, реализующей алгоритм Apriori (пакет «Apriori 1.1.1» из центрального репозитария модулей языка Python – Python Package Index). Автор: Самойлова Т.А.*

К примеру, значение supp (Гостиная A8 ⇒ Прихожая B2, Коридор B5) = 0.1 означает, что 10 % из общего числа всех транзакций содержат одновременный выбор трех указанных цветов для гостиной, прихожей и коридора. Достоверность conf(Прихожая B9 ⇒ Ванная A1, Кухня C5) = 0.6 означает, что 60 % транзакций, содержащих выбранный прихожей цвет B9, содержат выбранный для ванной цвет A1 и выбранный для кухни цвет C5. Значение lift(Гостиная B5 ⇒Прихожая B9) = 5.8 говорит о достаточно сильной связи между указанным выбором. Полученное в результате ассоциативное правило

=

«Гостиная A8 ⇒ Прихожая B2» означает следующее: если для гостиной респондент выбрал цвет A8, то для прихожей он выберет цвет B2.

РЕЗУЛЬТАТЫ РАБОТЫ Apriori
Трехэлементные значимые ассоциации
['Li-Living Room', 'Ha-Hallway', 'Be-Bedroom', 'Ba-Bathroom', 'To-Toilet', 'Ki-Kitchen', 'Co-Corridor']

Rule	supp(%)	conf(%)	lift
To_A1 => Ba_A1 , Be_A7	6	100	2.17
Co_A1 => Ba_A1 , Ki_A9	6	75	5.36
Li_B1 => Co_B1 , Ba_A1	6	75	3.12
Ha_A2 => Ba_A1 , Ki_A2	6	75	4.17
Ha_A2 => Ba_A1 , Li_A8	6	75	6.25
Ha_A2 => Ba_A1 , To_A1	6	100	2.78
To_A1 => Ba_A1 , Ki_A2	6	75	2.08
To_A1 => Ba_A1 , Li_A8	10	100	2.17
Ba_A9 => Ha_B1 , Li_C2	6	75	12.5
Ba_A9 => Ha_B1 , To_A1	6	75	3.75
Ha_B1 => Be_A1 , To_A1	6	100	2.17
Co_A1 => To_A1 , Ki_A9	6	75	5.36
Li_B1 => Co_B1 , Ha_C2	6	100	4.17
Li_B1 => Co_B1 , Ki_B1	6	75	3.12
Li_B1 => Co_B1 , To_A1	6	50	2.08
Ha_A2 => To_A1 , Li_A8	6	100	2.17
Li_B1 => To_A1 , Ha_C2	6	100	4.17
Li_B1 => To_A1 , Ki_A3	6	100	2.17

Рисунок 2. *Фрагмент результатов выполнения программы поиска значимых цветовых ассоциаций для трех помещений и следующих значений параметров: minsupport = 0,4; minconfidence = 0,5; minlift = 2,0. Автор: Самойлова Т.А.*

(2) Кластеризация цветовых предпочтений респондентов

Проведенная экспериментальная реализация показала, что для обработки цветовых предпочтений испытуемых хорошо подходят разные варианты алгоритма кластеризации, отличающиеся расчетом метрики сходства и правилами объединения кластеров. В случае числовых значений кластеризуемых данных, меры сходства могут вычисляться стандартным способом по расстоянию Евклида. Для нечисловых данных, заданных цветовой палитрой, метрики сходства могут рассчитываться по методике, соответствующей международной стандартной цветовой модели CIELAB, в которую переводятся цвета, указанные в ответах респондентов. При этом хорошие результаты показывает расчет меры сходства по формуле цветового отличия CIEDE2000 [12], которая позволяет математически представить расстояние между двумя оттенками в цветовом пространстве.

В ходе программного эксперимента было исследовано влияние на результаты кластеризации правил объединения кластеров: 'average' – попарного среднего, 'single' – ближайшего соседа, 'complete' – дальнего соседа, 'ward' – Уорда. Экспериментальная реализация показала, что в случае числовых исходных данных количество кластеров практически не зависело от выбранного правила и поэтому вычисления проводились методом Уорда [8], который стремится создавать кластеры равных размеров. В частности, наилучший результат показало применение метода Уорда к результатам статистической обработки результатов исследования, представленных распределением оценок (%) выбора цвета респондентами по каждому из помещений (рис. 3)

=

```
# Листинг 2
from scipy.cluster.hierarchy import dendrogram, linkage
rows = ['Гостиная',' Прихожая',' Спальня','Ванная',' Туалет',' Кухня',' Коридор']
linked = linkage(statistic_data, 'ward', metric='euclidean')
plt.figure(figsize=(10, 4))
plt.title("Hierarchical Clustering Dendrogram (Помещения)")
plt.xlabel('distance (Ward)')
plt.ylabel('Помещения')
dendrogram(
        linked,
        orientation='right',
        labels=rows, leaf_font_size=16.,
        distance_sort='descending',
        show_leaf_counts=False
       )
plt.show()
```

Рисунок 3. *Фрагмент листинга Python-программы кластеризации методом Уорда. Автор: Самойлова Т.А.*

В методе Уорда для оценки степени сходства кластеров берется каждая их пара и рассчитывается, насколько увеличится дисперсия, если их объединить. Эта целевая функция известна как внутригрупповая сумма квадратов отклонений (СКО):

$$\text{СКО} = (\text{Xi})^2 - 1/(\text{n} * (\sum \text{Xi})^2,$$

где Xi – значение признака i-го объекта.

Если кластеры находятся очень далеко друг от друга, то при их объединении дисперсия также окажется очень большой. Избежать этого можно, лишь объединяя очень близко находящиеся друг к другу кластеры, для которых целевая функция СКО получает минимальное приращение.

В случае данных, задаваемых цветовой палитрой, число кластеров существенно зависело от выбора правила их объединения и варьировалось в диапазоне от 3 до 10. Наиболее оптимальным оказался результат, полученный для правила «дальнего соседа», когда степень сходства оценивается по степени сходства между наиболее отдаленными (несхожими) объектами кластеров (рис. 4). Согласно этому правилу, расстояние между кластерами определяется как максимальное расстояние между всеми возможными точками в двух кластерах.

```
# Листинг 3
import scipy.cluster.hierarchy as shc
import matplotlib.pyplot as plt
L = sch.linkage(distance_matrix, method= 'complete')
ind = sch.fcluster(L, 0.5*distance_matrix.max(), 'distance')
print(ind)
plt.figure(figsize=(6, 10))
sch.set_link_color_palette(['c', 'g', 'r', 'k'])
dn = sch.dendrogram(L, labels=rows, orientation='right')   #orientation='right')
'top'
plt.title(my_name[num])
plt.show()
```

Рисунок 4. *Фрагмент листинга Python – программы, разработанной для проведения кластеризации методом «дальнего соседа». Автор: Самойлова Т.А.*

При программной реализации средствами Python-пакета Hierarchical clustering из библиотеки SciPy в пакете имеется возможность вмешательства в работу алгоритма путем включения нестандартного блока расчета расстояний между ответами респондентов по формуле цветового отличия CIEDE2000 [12]. Он позволяет также выполнять кластеризацию средствами метода linkage, где есть возможность задания конкретного правила объединения кластеров.

(3) Визуализация результатов интеллектуального поиска

Оба метода интеллектуального анализа открывают широкие возможности визуализации

=

выявленных психологических механизмов и закономерностей.

Значения параметра lift-значимости трехэлементных ассоциаций, найденных алгоритмом Apriori могут быть представлены в форме диаграммы (рис. 5).

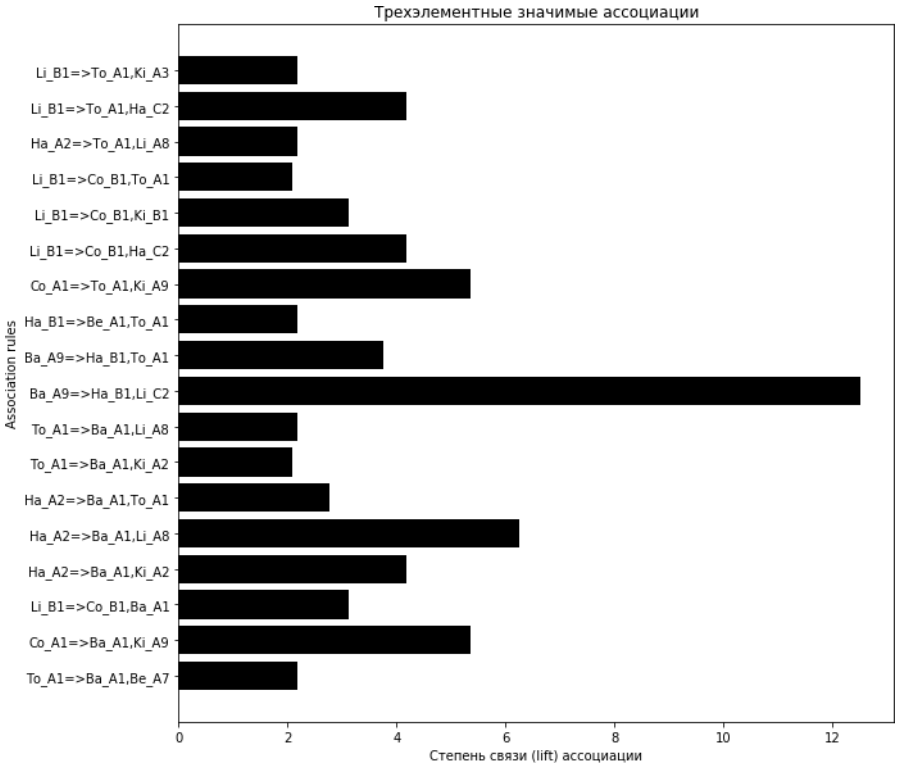

Рисунок 5. *Диаграмма значимости (lift) трехэлементных ассоциаций цветовых предпочтений респондентов. Автор: Самойлова Т.А.*

Применение метода кластеризации позволяет средствами языка Python 3.6 формировать дендрограммы с использованием базовой функции shc.dendrogram (рис. 6). Программа позволяет разрезать дерево дендрограммы на группы кластеров, которые изображаются в виде прямоугольников. Мера сходства ответов выполняется по стандарту CIEDE2000 посредством вычисления матрицы расстояний, непосредственно используемой алгоритмом кластеризации, или по метрике Евклида.

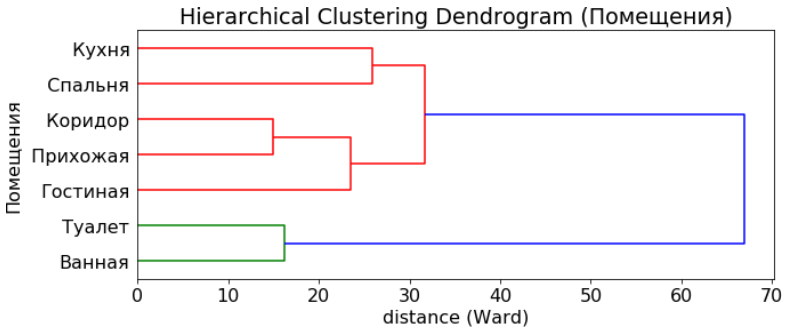

Рисунок 6. *Дендрограмма кластеризации цветовых предпочтений респондентов. Автор: Самойлова Т.А.*

=

Кроме того, визуализация результатов в обоих случаях может выполняться средствами специализированного Python-пакета «Tkinter» и в случае применения метода поиска ассоциаций может показывать наиболее заметные связи между цветами и типами помещений (рис. 7), а при применении различных алгоритмов кластеризации – пояснять данные дендрограмм (рис. 8).

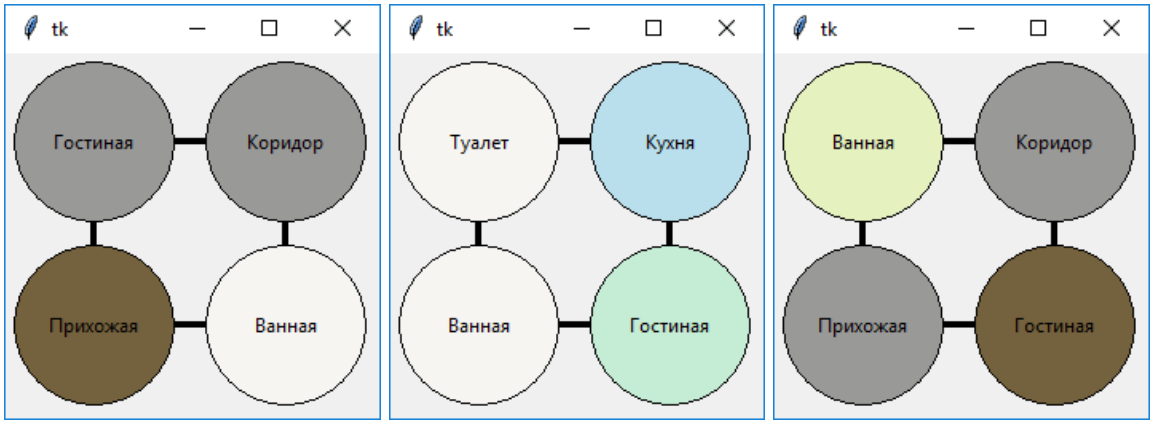

Рисунок 7. *Визуализация значимых ассоциаций цветовых предпочтений респондентов в цветовой палитре. Автор: Самойлова Т.А.*

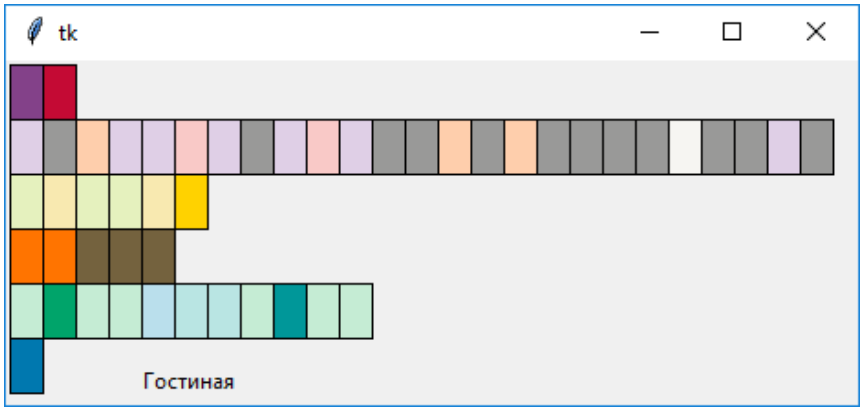

Рисунок 8. *Результаты кластеризации ответов в цветовой палитре. Автор: Самойлова Т.А.*

Выводы

В целом проведенные исследования показывают, что оба метода интеллектуального анализа данных (метод поиск ассоциативных правил и метод иерархической кластеризации), хорошо подходят для обработки данных социально-психологических исследований цветовых предпочтений.

Метод поиска ассоциаций позволяет, установив минимальные значения параметров таким образом, чтобы ограничить количество найденных правил, выявить новые психологические механизмы и закономерности в ответах респондентов, а также проверить зависимость выбираемых цветов от типов объектов и социальных характеристик участников (их пола, возраста, места проживания).

При значительном увеличении числа участников исследования и количества возможных вариантов ответов, задача поиска ассоциаций эффективно решается параллельными методами.

Предложенный подход к кластеризации, основанный на вычислении меры сходства ответов респондентов по формуле цветового отличия CIEDE2000 и допускающий выбор правила объединения кластеров, позволяет сделать выводы, к которым невозможно прийти, вычисляя лишь процентные показатели. При этом в зависимости от целей исследования, в

=

качестве правил объединения кластеров могут выбираться различные методы, отличающиеся решением вопроса о схожести объектов при их объединении в группу: метод дальнего соседа, одиночной связи, попарного среднего, Уорда.

=

ЛИТЕРАТУРА

1. Chong A.Y.L., Ch’ng E., Liu M.J., and Li B. Predicting consumer product demands via big data: the roles of online promotional marketing and online reviews // International Journal of Production Research. 2017. № 55 (17). P. 5142–5156. https://doi.org/10.1080/00207543.2015.1066519.
2. Fazzolari M., Petrocchi M. A study on online travel reviews through intelligent data analysis // Information Technology and Tourism. 2018. № 20. P. 37–58. https://doi.org/10.1007/s40558-018-0121-z.
3. Бродовская Е.В., Домбровская А.Ю., Иванов И.С. Изменение стратегий онлайн-поведения российской интернет-аудитории: результаты сравнительного кластерного анализа (2012–2014 гг.) // Мониторинг общественного мнения: Экономические и социальные перемены. 2016. № 3. С. 173–187. https://doi.org/10.14515/monitoring.2016.3.10.
4. Kyslova O. Big Data in the Context of Studying Problems of Modern Society // Visnyk V.N. Karazin Kharkiv National University. Series «Sociological Studies of Contemporary Society: Methodology, Theory, Methods». 2020. № 43. P. 26–33. https://doi.org/10.26565/2227-6521-2019-43-03.
5. Jonauskaite D., Abu-Akel A., Dael N. et al. Universal Patterns in Color-Emotion Associations are Further Shaped by Linguistic and Geographic Proximity // Psychological Science. 2020. № 31(10). P. 1245–1260. https://doi.org/10.1177/0956797620948810.
6. Uusküla M., Bimler D.L. From listing data to semantic maps: Cross-linguistic commonalities in cognitive representation of color // Folklore. 2016. № 64. P. 159–180. https://doi.org/10.7592/FEJF2016.64.colour.

=

7. Беликова М.Ю., Каранина С.Ю., Глебова А.В. Экспериментальное сравнение алгоритмов кластеризации в задаче группировки данных о грозовых разрядах // Кибернетика и программирование. 2018. № 1. С. 15–26. https://doi.org/10.25136/2306-4196.2018.1.25261.

8. Xu D., Tian Y. A Comprehensive Survey of Clustering Algorithms // Annals of Data Science. 2015. № 2. P. 165–193. https://doi.org/10.1007/s40745-015-0040-1.

9. Schnaudt Ch., Weinhardt M., Fitzgerald R., and Liebig St. The European Social Survey: Contents, Design, and Research Potential // Schmollers Jahrbuch. 2014. № 134 (4). P. 487–506. https://doi.org/10.3790/schm.134.4.487.

10. Грибер Ю.А., Самойлова Т.А., Двойнев В.В. Цветовые предпочтения пожилых людей в различных типах жилого интерьера // Урбанистика. 2018. № 4. С. 36–49. https://doi.org/10.7256/2310-8673.2018.4.28349.

11. Tank D. Improved Apriori Algorithm for Mining Association Rules // International Journal of Information Technology and Computer Science. 2014. № 6 (7). P. 15–23. https://doi.org/10.5815/ijitcs.2014.07.03.

12. Luo M.R., Cui G. and Rigg B. The Development of the CIE 2000 Colour Difference Formula: CIEDE2000 // Color Research & Application. 2001. № 26 (5). P. 340–350. https://doi.org/10.1002/col.1049.

=