УДК 004.9

# ПРОГРАММНЫЙ КОМПЛЕКС ДЛЯ ТЕМАТИЧЕСКОЙ КЛАССИФИКАЦИИ ТЕКСТОВЫХ ДАННЫХ

***Рождественская В.Б.***

*ФГАОУ ВПО Казанский (Приволжский) федеральный университет,*
*420008, г. Казань, ул. Кремлевская, д.18*

*e-mail: omni-equilibrium@mail.ru*

*поступила в редакцию 03 декабря 2014 года*

**Аннотация**

В статье приведены основные полученные результаты разработки программного комплекса для тематической классификации текстовых данных, автоматически собираемых из интернет-источников.

***Ключевые слова:*** *классификация текстовых данных, кластеризация, обработка текста, карта Кохонена, SOM, нейронные сети.*

**Введение.** С каждым годом количество и общий объем текстовых документов, используемых во многих областях человеческой деятельности, неуклонно растет.

Вследствие стремительного развития сети Интернет для поиска нужной информации, в частности, поисковыми системами, в больших и сверхбольших коллекциях данных стала необходима предварительная автоматическая систематизация текстового набора данных. Классификация текстового массива позволяет систематизировать коллекцию и сузить область рассматриваемых документов. Полученный массив можно классифицировать ещё несколько раз, разбивая полученные тематические множества на ещё более мелкие подмножества.

Разбиение текстовых массивов на систему схожих подмножеств становится важной задачей, без решения которой эффективная работа с текстами невозможна. По этой причине классификация любых данных, не только текстовых, на сегодняшний день является одной из самых актуальных задач.

Классификация документов разделяется на категоризацию и кластеризацию. В случае категоризации должна присутствовать обучающая выборка документов, для которых заранее известно, к какой группе, или классу, они принадлежат. В случае же кластеризации количество и параметры множеств, на которые можно разбить документы, неизвестны.

Кластеризация – один из фундаментальных методов современного анализа данных. Наиболее востребованной на сегодня и, вероятно, в ближайшем будущем является смысловая (тематическая) кластеризация текстовых документов. Этот вид кластеризации предполагает разделение текстовых коллекций на множества текстов (кластеры), такие, что тексты в пределах одного и того же кластера максимально схожи между собой по смыслу, тогда как тексты, относящиеся к разным кластерам, имеют различный смысл. При смысловой кластеризации текстов возможна «пометка» выделяемых кластеров их тематическими описателями.

Существенный толчок к развитию методов автоматической кластеризации текстов дали работы двух ученых из Корнельского Университета – G. Salton и A. Wang. В 1975 году группа ученых из Корнельского университета опубликовала статью, предлагавшую описывать тексты в виде векторов в многомерном пространстве и, соответственно, использовать в работе с такими текстами-векторами стандартные меры близости в векторных пространствах [1].

В настоящее время существуют десятки различных алгоритмов разбиения текстовых документов по темам, основанные на элементах математической статистики, нечёткой логики, нейронных сетей и тому подобном. Но даже при таком количестве возможных

решений развитие и усовершенствование этих способов ведётся и по сей день в рамках компьютерной лингвистики.

*Цели работы:* разработка и анализ программы для автоматической кластеризации текстовых данных, содержащихся в базе данных.

*Задачи:*

- выбор и сравнительный анализ подходящего алгоритма классификации текстов;
- разработка метода преобразования текстовых данных для выбранной задачи классификации;
- проектирование программы;
- программная реализация;
- исследование алгоритма и тестирование программы.

**Основная часть.** Для создания программы был проанализирован некоторый набор существующих алгоритмов классификации текстовых данных. В результате для классификации текстовых данных был выбран алгоритм самоорганизующихся карт Кохонена (SOM), впервые предложенный Тойво Кохоненом в 1982 [2]. Такой выбор следовал по нескольким причинам:

- этот алгоритм является алгоритмом классификации без учителя, то есть для его работы не требуется составлять обучающее множество документов – обучение происходит автоматически;
- для классификации количество классов (групп), на которые подразделяются все документы, определяется полностью автоматически;
- получаемая карта даёт наглядное представление о полученных классах;
- хотя алгоритм имеет квадратичную сложность [3], использование нейронных сетей даёт более высокую эффективность по сравнению с другими методами.

Программа для классификации текстовых данных состоит из 8 файлов скрипта, 4 из которых отвечают за соединение с базой данных с находящимся в ней набором текстов или создание новой, 2 файла отвечают за сбор и обработку текстовых данных из интернет-источников, и остальные 2 реализуют алгоритм классификации и построения карты Кохонена.

Перед началом непосредственно классификации обрабатываемый набор текстов преобразовывался в их образы, то есть векторы, состоящие из весов ключевых слов каждого текста. Затем далее весь алгоритм классификации работал только с этими векторами.

Алгоритм SOM работает согласно алгоритму обратного распространения ошибки и состоит из следующих основных шагов [3]:

1) По узлам карты размера *len* x *len* , где *len* – количество нейронов сети, распределяются случайные значения весов нейронов. Задаются пороговое значение допустимой ошибки обучения и радиус соседства;
2) Извлекается случайный документ (его образ) и вычисляется нейрон-победитель, т.е. нейрон, который является наиболее близким к предъявленному образу;
3) Происходит коррекция весов нейрона-победителя и его ближайших соседей;
4) Далее шаги 2-3 повторяются, пока не закончатся документы;
5) Вычисляется текущая ошибка обучения, и если она больше некоторой заданной малой величины, то повторяются шаги 2-5.
6) На выходе получаем веса множества нейронов сети.

Программа тестировалась на выборке из 6 заранее заданных текстов со своими определёнными наборами ключевых слов для возможности лёгкой последующей проверки и сравнения результата с верным (в скобках указана частота встречаемости в тексте):

0) [китайский (2), пекин (1)]
1) [китайский (2), шанхай (1)]
2) [китайский (1), макао (1)]
3) [токио (1), япония (1), китайский (1)]
4) [китайский (3), токио (1), япония (1)]

5) [токио (1), пекин (1)]

После запуска скрипта для проверки на экран выводятся получаемые векторы-образы текстовых документов, количество выполненных алгоритмом итераций и ошибка обучения на каждой итерации, массив получаемых при распределении классов с номерами документов, принадлежащих каждому классу, и, конечно же, графическое отображение карты Кохонена с отмеченными на ней идентификаторами документов.

Расположение документов на карте, а также вектор, показывающий привязанность того или иного документа какому-либо номеру получаемого класса, адекватно соответствуют реальному распределению тех же текстов на классы человеком. Так, например, в одну группу алгоритм внёс тексты с номерами 0 и 5, в другую – 3 и 4, в две отдельные группы входят документ 1 и документ 2.

На рисунке 1 приведён пример полученной карты Кохонена для приведённых выше документов, а на рисунке 2 – соответствующее распределение на классы:

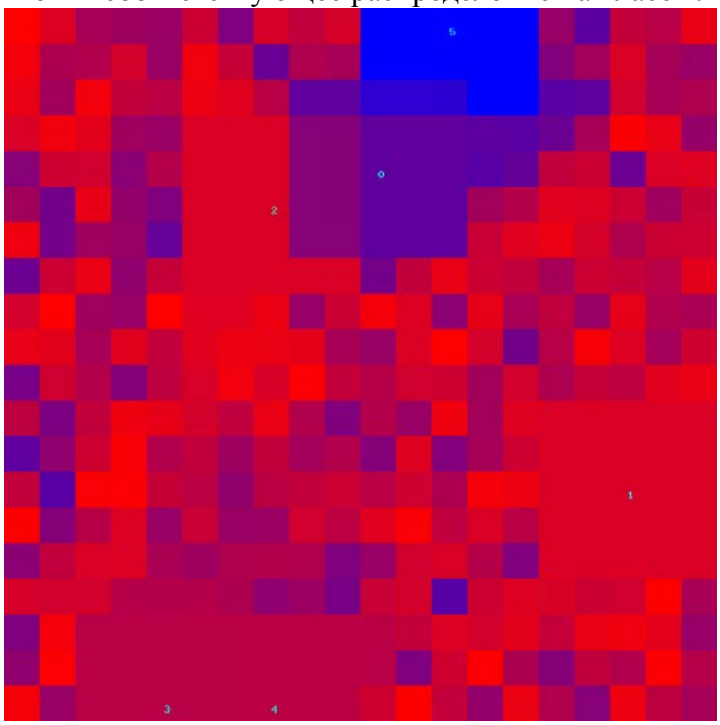


Рисунок 1. – Карта Кохонена.

Количество кластеров = 4
Кластеры:Array ( [0] => Array ( [0] => 0 [1] => 5 ) [1] => Array ( [0] => 1 ) [2] => Array ( [0] => 2 ) [3] => Array ( [0] => 3 [1] => 4 ) )

Рисунок 2. – Массив классов.

При разбиении на группы большого количества документов алгоритм требует достаточной вычислительной мощности компьютера, но при этом он также работает адекватно.

**Заключение.** Таким образом, была написана программа для кластеризации набора текстов, содержащихся в базе данных. В качестве алгоритма кластеризации был выбран нейросетевой алгоритм самоорганизующихся карт Кохонена (SOM). Результатом работы программы является построение карты Кохонена – наглядного графического отображения получаемого разделения на кластеры. Программу можно использовать на практике для кластеризации ограниченного набора текстовых данных.

**Благодарность.** Выражаю огромную благодарность своему научному руководителю Каримову В. С., а также Демьянову Д. Н. за помощь в написании и оформлении статьи.

## Список литературы

1) Шмулевич М.М. Метод автоматической кластеризации текстов, основанный на извлечении из текстов имен объектов и последующем построении графов совместной

встречаемости ключевых термов: дис. на соиск. учён. степ. канд. физ.-мат. наук. М.: МФТИ. 2009. 120 с.
2) Kohonen T. Self-Organizing Maps // Springer Series in Information Sciences. V.30. 3rd ed. Berlin; Heidelberg; New York; Barcelona; Hong Kong; London; Milan; Paris; Singapore; Tokyo: Springer, 2001. 501 p.
3) Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие. М.: МИЭМ, 2011. 272 с.