Задача кластеризации текстовых документов

М.В. Хачумов

Аннотация. В статье рассмотрены вопросы совершенствования технологии кластеризации текстовых документов на основе оптимизации числа кластеров и их первоначального размещения, а также выбора наиболее адекватных метрик. Полученные в ходе экспериментов результаты подтверждают эффективность предложенного подхода.

Ключевые слова: текст, кластеризация, класс, вектор, метрика, центр кластера, рубрика, эксперимент.

Введение

Одной из задач автоматического рубрицирования является кластеризация документов, заключающаяся в выявлении семантически связанных текстов в многомерном пространстве информационных признаков, а также определении центров кластеров, представляющих собой тематические рубрики (аннотации) [1]. Выбор метода определения первоначального числа кластеров весьма ограничен, плохо формализуется и осуществляется, как правило, экспертом.

Алгоритмы кластеризации текстовых документов предполагают сравнение объектов между собой на основе некоторой меры близости или расстояния [2], от которой существенным образом зависит конечный результат. Чаще всего применяют расстояние Евклида, которое позволяет формировать центры кластеров, но характеристики внутриклассового распределения при этом не берутся в расчет. Данное обстоятельство не позволяет учитывать корреляционные связи анализируемых объектов, что может сказаться на качестве кластеризации. Довольно большое распространение получила метрика Махаланобиса, лишенная указанного недостатка, но она не всегда позволяет вычислить искомое расстояние из-за проблемы получения обратной матрицы ковариаций. В настоящей работе предлагается использовать для кластеризации дополнительно обобщенную метрику Евклида–Махаланобиса [3], которая работает во всех случаях, но, как и метрика Махаланобиса, требует трудоемких вычислений.

Кластеризация является многоэтапной процедурой, на каждом шаге которой должна решаться отдельная задача выбора наиболее адекватного способа реализации, влияющего на последующие этапы. В работе дается сравнительный анализ алгоритмов, используемых на различных этапах кластеризации, эффективность которых проверяется экспериментальными исследованиями.

1. Постановка задачи кластеризации

Будем различать два крупных этапа кластеризации, характерных для обработки большого числа документов. На первом этапе происходит формирование кластеров на основе ограниченной выборки документов. На втором – окончательное распределение всего корпуса документов. Существенное значение при этом имеет выбор мер близости (расстояния).

1.1. Постановка задачи формирования центров кластеров

Пусть задано множество объектов-текстов $\{\omega_1,...,\omega_m\}$, каждый из которых формализуется n- мерным вектором признаков $(x_{j1}\ldots x_{jn})$, $j = 1,\ldots m$, или точкой в n- мерном пространстве признаков. Зададим меру расстояний $dist\left(x^i,\omega\right)$, где x^i – точка, характеризующая центр (ядро) i-го класса (кластера), а ω – точка из множества объектов. Тогда для заданного числа классов k необходимо подобрать k ядер таким образом, чтобы сумма внутриклассовых расстояний R была минимальной [4]:

$$R=\sum_{i=1}^{k}\sum_{x\in K_i} dist\left(x^i,\omega\right)\to \min, \qquad (1)$$

где K_i – множество объектов i-го класса.

Заметим, что решение задачи (1) предполагает знание числа кластеров, которое задается пользователем исходя из целесообразности с учетом особенностей предметной области. Если таких знаний нет, то значение k определяется экспериментально. В настоящей работе рассматривается метод, основанный на последовательном парном объединении близких объектов с усреднением их характеристик. Он позволяет автоматизировать процесс выбора числа кластеров, но окончательное решение принимает пользователь-эксперт.

1.2. Постановка задачи распределения документов по кластерам

Задача распределения документов может быть сформулирована как задача распознавания по Журавлеву Ю.И. [5]. Пусть дано множество М объектов $\{\omega_i\}$; на этом множестве имеется разбиение на конечное число классов $\Omega_l, l=1,...,k$, $\bigcup_{l=1}^{k}\Omega_l = M$. Объекты задаются значениями некоторых признаков $x_j, j=1,...,n$. Совокупность значений признаков x_j определяет описание объекта $I(\omega)=\{x_1,x_2,\ldots,x_n\}$. Информация о вхождении некоторого объекта ω в какой-либо класс представляется в виде вектора $\{I_1(\omega),I_2(\omega),\ldots,I_k(\omega)\}$, где $I_p(\omega)$ несет информацию о принадлежности объекта ω к классу Ω_p:

$$I_p(\omega)=\begin{cases}1, & \text{если } \omega\in\Omega_{\text{р}},\\ 0, & \text{если } \omega\notin\Omega_{\text{р}},\\ \Delta, & \text{если неопределенность.}\end{cases}$$

Решение о принадлежности объекта ω классу Ω_p принимается на основе сравнения расстояний между объектом и классами с некоторыми заранее установленными порогами-допусками.

1.3. Выбор метрик

Расстояние между образцом и классом может быть замерено в соответствии с формулой:

$$D^2=(X-W)^T A^{-1}(X-W), \qquad (2)$$

где X - вектор признаков исследуемого образца, W - вектор средних значений (или мат. ожиданий) значений признаков класса (кластера),

$$A=\begin{cases}E, \textit{ для метрики Евклида (E - единичная матрица),}\\ C, \textit{ для метрики Махаланобиса (C - матрица ковариаций),}\\ C+E, \textit{ для метрики Евклида - Махаланобиса}\end{cases}$$

Элементы матрицы C для p - го класса вычисляются по формуле:

$$c_{ij}=\frac{1}{K_p-1}\sum_{\beta=1}^{K_p}(x_i^{\beta}-\overline{x_i})(x_j^{\beta}-\overline{x_j}), \qquad (3)$$

где i,j $(i,j=1,...,n)$ – все возможные пары индексов измеряемых признаков; x_k^l - значение k - го признака l - го объекта; выражения в скобках есть отклонения от соответствующего среднего значения признака $\overline{x_k}$ для p - го класса. При $i=j$ вычисляются среднеквадратичные отклонения, которые соответствуют дисперсиям признаков, а при $i\neq j$ оценивается ковариация между двумя признаками.

Введение метрик Махаланобиса и Евклида-Махаланобиса [2-3] усложняет алгоритм кластеризации и способно существенно замедлить скорость его работы. В связи с этим необходимы дополнительные меры по ускорению счета, связанные, например, с использованием много-

процессорных систем и кластерных установок. Кроме того, следует установить конкретные места их целесообразного применения.

2. Особенности формирования кластеров

Выбор методов определения оптимального числа кластеров и их центров весьма ограничен и связан с трудоемкими процедурами. Рассмотрим некоторые подходы к решению этой задачи.

2.1. Методы формирования кластеров

Подходы, основанные на идеях алгоритмов DEL–ADD [6], используют при формировании кластеров принцип слияния-расщепления. На основе некоторого критерия качества принимается решение об удалении рассматриваемого класса из заданного начального множества, слиянии с другим классом или разбиении на два класса. Процесс останавливается, когда получают минимальную по размеру систему классов с приемлемым качеством. Критериями качества класса, отождествляемого с кластером могут быть:

1) Количественный критерий. Класс, количество точек в котором меньше заданного числа, определенного в пространстве признаков, считается пустым и полежит удалению. Порог выбирается экспертом в зависимости из смысла задачи и вида меры близости.

2) Критерий сферической разделимости. Два класса считаются сферически разделимыми, если сумма радиусов двух классов меньше расстояния между ядрами (центрами) этих классов. Если классы сферически неразделимы, то они сливаются в один.

3) *Средняя мера близости точек класса от ядра* (критерий равномерности). Средняя мера близости точек класса от ядра должна быть не менее половины или трети от максимума меры близости точек от ядра (радиуса класса). В противном случае класс разбивается на два (порождается еще одно ядро вблизи исходного).

4) *Часто воспроизводящийся класс.* Проводится достаточно большая серия классификаций с различным начальным выбором классов. Определяются частоты появления классов, которые служат основанием для получения «истинного» числа классов.

По способу формирования кластеров алгоритмы бывают двух типов: иерархические и неиерархические. Иерархические алгоритмы производят последовательное объединение исходных объектов и соответствующее уменьшение числа кластеров. Неиерархические алгоритмы основаны на оптимизации некоторой целевой функции, определяющей оптимальное в определенном смысле разбиение множества объектов на кластеры.

В настоящее время имеется множество методов и систем кластеризации, обзор которых дан в работе [1]. Наиболее популярные из них: Suffix Tree Clustering (STC) - кластеры образуются в узлах суффиксного дерева, которое строится из слов и фраз входных документов; K-means - кластеры представлены в виде центроидов, являющихся “центром массы” всех документов, входящих в кластер (данный алгоритм - один из основных алгоритмом кластеризации); Self-Organizing Maps (SOM) - производит классификацию документов с использованием самонастраивающейся нейронной сети Кохонена.

Выбор того или иного метода требует проведения экспериментальных исследований, анализа преимуществ и недостатков в конкретной предметной области. Перечисленные алгоритмы, к сожалению, достаточно трудоемки и сложны для сопоставительного анализа. Несмотря на внушительный перечень подходов и большое количество конкретных реализаций, проблема кластеризации документов на естественном языке остается в настоящее время в центре внимания исследователей в силу своей сложности и далека от окончательного решения.

2.2. Методы выбора числа кластеров и определения центров направленным объединением

Рассмотрим задачу объединения в группы (кластеры, классы) «близких» объектов. Пусть имеется m объектов $\omega_1, \omega_2, ..., \omega_m$; причем каждый i-ый объект характеризуется n признаками: $x_i = \{x_{i1}, ..., x_{in}\}$. Проведем парное сравнение объектов с определением близости между ними. Пусть r_{ij} - мера близости объектов i и j, тогда имеем матрицу размерности $m \times m$. Будем последовательно объединять па-

ры объектов. Для этого в матрице выбирается максимальное значение r_{ij}, $i \neq j$ и объекты ω_i, ω_j объединяются, создавая кластер $\{\omega_i, \omega_j\}$. Результат объединения можно рассматривать как новый объект ω_{i-j}, которому приписываются усредненные значения признаков. Из матрицы близости удаляем два объекта ω_i, ω_j и вводим новый объект ω_{i-j}. Рассчитываем близость нового объекта ко всем оставшимся и получаем матрицу размерности $(m-1)\times(m-1)$. Процесс повторяется, пока не будет получен один кластер. Задаваясь наперед степенью близости кластеров, можно остановить процедуру объединения, получив при этом вполне определенное число кластеров и расположение их центров.

Данная схема проста в реализации и может служить звеном иерархической кластеризации документов. Однако ее применение оправданно лишь при ограниченной выборке документов для первоначального формирования кластеров. При этом открытым остается вопрос принципа формирования самой выборки.

Представляет определенный интерес алгоритм SOM [7,8]. Он характерен тем, что количество кластеров и местоположение их центров устанавливается заранее. В процессе итерационной процедуры происходит подача объектов учебной выборки и уточнение местоположения центров. Настройка осуществляется путем коррекции весовых коэффициентов для всех нейронов, попадающих в зону соседства, в соответствии с алгоритмом Видроу-Хоффа до получения приемлемого качества кластеризации, удовлетворяющего пользователя, или до останова, предусмотренного алгоритмом настройки. Слабым местом данного подхода остается начальная инициализация нейронной сети, которая проводится, например, случайным распределением ядер на гиперсфере единичного радиуса или в соответствии с другими критериями [9-10]. Эффективность и реализуемость подобных теоретических методов начального распределения центров для задач реальной кластеризации текстовых документов пока остается под вопросом.

3. Особенности выбора метрик на различных этапах алгоритма кластеризации

Свойства метрик Евклида и Махаланобиса достаточно хорошо изучены. Менее известна обобщенная метрика [3]. Остановимся на некоторых ее особенностях. Расстояние между двумя классами Y_1 и Y_2, можно представить в виде квадратичной формы:

$$R_G^2(Y_1, Y_2) = (\bar{x}_1 - \bar{x}_2)^T A^{-1} (\bar{x}_1 - \bar{x}_2), \qquad (4)$$

где $\bar{x}_1$ и $\bar{x}_2$ – векторы средних выборочных классов Y_1 и Y_2.

Матрица A в выражении (4) должна быть симметрической и положительно определенной. Этим свойством, например, обладает матрица $A = C_1 + C_2 + E$, где C_1 и C_2 – корреляционные матрицы для классов Y_1 и Y_2 соответственно. Для любых двух классов Y_1 и Y_2, у которых совпадают центры $\bar{x}_1 = \bar{x}_2$, расстояние $R_G^2(Y_1, Y_2) = 0$. Если класс Y_1 представляет собой точку, то соответствующая ему корреляционная матрица состоит из нулей и получается расстояние, аналогичное расстоянию Махаланобиса, с той разницей, что $R_G^2(Y_1, Y_2)$ не стремится к бесконечности, если какая либо дисперсия обращается в ноль. Если оба класса представляют собой точки, то $A = E$, $R_G^2(Y_1, Y_2) = R_E^2(Y_1, Y_2)$, т.е. расстояние Махаланобиса совпадает с расстоянием Евклида $R_E(Y_1, Y_2)$.

С учетом многоэтапности алгоритма кластеризации целесообразно расширить его возможности путем использования нескольких метрик. Конкретные предложения заключаются в следующем.

1) На первом этапе следует выполнить кластеризацию ограниченной выборки. Для построения кластеров на этом этапе применяется метрика Евклида. Выборка должна быть представимой для статистической значимости распределений образцов внутри полученных кластеров. При достаточном количестве образцов полученные центры можно рассматривать как математические ожидания координат центров классов.

2) На втором этапе проводится вычисление коэффициентов ковариационной матрицы для каждого класса. Определяется целесообразность применения метрик и пороги близости (расстояний), разделяющие классы. Метрики Махаланобиса или Евклида-Махаланобиса следует применять, если число образцов каждого кластера достаточно для достоверной оценки его ковариационной матрицы. На этом этапе заканчивается процесс формирования кластеров.

3) Третий этап связан распознаванием или кластеризацией произвольных входных образцов на основе одной из метрик в соответствии с установленными порогами. Данный этап согласуется с предложениями работы [5].

Можно дать общие рекомендации по выбору метрики. В случае малого разброса параметров вероятностного распределения внутри классов и отсутствия пересечений кластеров целесообразно использовать метрику Евклида. В случае, когда имеет место значительный разброс параметров, целесообразно использовать метрику Махаланобиса. Обобщенная метрика удобна в случаях, когда характер классов заранее не известен. Она более надежна, т.к. в случае малых разбросов параметров практически совпадает с метрикой Евклида, а при больших разбросах – с метрикой Махаланобиса, позволяя всегда находить обратную матрицу ковариаций.

4. Построение алгоритма кластеризация текстовых документов и сравнительный анализ

В настоящем разделе проводится экспериментальное сравнение методов первоначального формирования кластеров и распределения документов на основе метрик Евклида и Евклида–Махаланобиса, выполненное в условиях экспертной системы «Айкумена» [11].

4.1. Образование учебной выборки и предварительный сбор статистики

По заданному текстовому фильтру (текстовому предикату) из базы данных извлекаются документы наиболее релевантные запросу. Эти документы в дальнейшем и будут подвергаться кластеризации на ее первом этапе. При этом для кластеров выделяются ключевые слова и фразы (кластер будет проаннотирован); формируется текстовый предикат, на основе которого будут фильтроваться остальные документы. Такое решение связано с временными ограничениями процесса кластеризации в соответствии с требованиями пользователей.

На вход блока кластеризации подаются проиндексированные документы (нормализованный и ненормализованный текстовый индекс документа), а также глоссарий. Нормализованный индекс участвует в процессе формирования кластеров, а ненормализованный используется для аннотирования (выделения ключевых слов и фраз) кластера.

4.2. Векторизация документов и сокращение признакового пространства

Для сравнения документов необходимо их векторизовать, т.е. выделить пространство признаков. Каждому документу, а также центрам кластеров соответствует объект (точка в пространстве признаков). Основными признаками являются веса слов и фраз, входящих в состав документа, при подсчете которых будем пользоваться частотой их появления в самом документе и обучающей выборке, а также данными глоссария. Для ID-слов и ID-фраз (от англ. слова identificator) подсчитывается статистика: частота, число документов кластера, в которых содержится данное слово или фраза. Размерность пространства признаков определяется числом возможных слов и фраз. В структуру (внутренний класс, реализующий объект пространства признаков) входят: массив ID-слов, массив весов слов, массив частот слов, массив номеров документов кластера, содержащих эти слова; массив ID-фраз, массив весов фраз, массив частот фраз (в документе и кластере), массив номеров документов кластера, содержащих эти фразы.

Специфика модификации кластеров в процессе их формирования связана с последовательным расширением признакового пространства и, соответственно, размерности векторов. Данная особенность характерна для центров кластеров, описания которых разрастаются при добавлении очередного документа за счет увеличения числа ненулевых частот и весов встре-

чаемости признаков. Такие «разрастающиеся» объекты трудно сравнивать, поэтому вводятся ограничения на размер векторов, которые реализуются за счет устранения («обнуления») малоинформативных признаков. На данном этапе используется накопленная статистическая информация. Выполняются следующие действия: убираются слова и фразы, встречающиеся менее двух раз, или же с частотой (вероятностью) встречаемости более 0.9 на число отобранных документов. Подготовленные таким образом векторы нормируются, после чего они готовы к дальнейшему анализу

4.3. Первоначальное формирование кластеров и размещение центров

Будем использовать две разные меры – первую на базе расстояния Евклида во время формирования кластеров и распределения обучающей выборки, а вторую на базе расстояния Евклида-Махаланобиса во время распределения всего корпуса документов. Эффективность такого подхода будем проверять экспериментально. Мера близости (на базе расстояния Евклида) включает две составляющие: S_1 - меру по ID-словам и S_2 - меру по ID-фразам. Начальные значения мер равны нулю, суммарная мера определяется как $S = (S_1 + S_2)$. Для сравнения векторов вычисляют пересечение множеств их элементов.

Рассмотрим два подхода к формированию кластеров для ограниченной выборки документов**.**

Базовый метод

Данный метод имеет следующие особенности: количество кластеров определяется исходя из опыта экспертов, центры кластеров выбирались случайно. Выполняется итерационная процедура**,** которая представляет собой модифицированный алгоритм k-means [1], относящийся к неиерархическим методам**.** Рассмотрим основные шаги этой процедуры.

- Распределяем выборку документов по кластерам. Для этого выбираем случайным образом очередной документ и сравниваем его векторизованный образ с векторами центров сформированных к этому моменту. Помещаем документ в ближайший кластер. При добавлении документа структура центра кластера должна быть заполнена модифицированной информацией по установленным правилам.
- Оптимизируем вектор центра кластера. Для этого оставляют слова и фразы с наибольшим весом. Критерий данной процедуры: ограниченное число слов и фраз.
- Нормируем вектор центра кластера.
- Оптимизируем центры с более жесткими требованиями по числу слов и фраз.
- Оптимизируем кластер по количеству входящих в него документов.

На выходе алгоритма имеем сформированные кластеры и их центры.

Метод направленного объединения

Первоначальное расположение центров кластеров будем определять направленным объединением векторов образцов учебной выборки. Данная процедура описана в п.2.2.

При объединении элементов используются следующие правила:

- для новых ID- слов и фраз добавляем их ID, копируем вес и частоту встречаемости;
- для совпавших ID- слов и фраз корректируем веса центров кластера;
- проводим нормирование объединенного вектора;
- ограничиваем число фраз и слов для недопущения разрастания кластера.

Критерием останова является пороговая величина расстояния между сравниваемыми векторами центров. Такой подход позволяет не задавать заранее число кластеров.

Как показали эксперименты иерархический подход разбивает рубрику на большее число полноправных кластеров, тем самым, показывая некоторое преимущество порогового определения числа кластеров.

4.4. Аннотирование кластеров

Для наглядности результатов и подтверждения, что полученные кластеры образуют определенные тематические группы, необходимо их проаннотировать. При этом для каждого кластера необходимо получить следующую информацию:

- название кластера – его образуют некоторые наиболее релевантные слова и фразы, максимальное число которых задается в качестве параметра;

- описание кластера – некоторое расширение названия кластера, которое наиболее полно описывает его тематику; максимальное число слов и фраз задается в качестве параметра;
- текстовый фильтр – на языке существующего поискового движка это фильтр, по которому попавшие в систему документы могут быть отнесены к данному кластеру;
- ключевые слова (аннотация) – те фразы и слова, которые будут выделяться (например, цветом) в документах кластера; максимальное число слов и фраз задается в качестве параметра.

Кроме того, для каждого документа, попавшего в кластер, необходимо определить релевантность отнесения к данному кластеру. Аннотирование требует выполнения следующих этапов:

Сбор статистики по фразам в рамках созданных кластеров

Просматриваем все фразы, которые встречаются в обучающей выборке, и для каждой из них подсчитываем число документов, в которых она встретилось.

Заполнение структур объектов

Просматриваем все сформированные кластеры и для каждого из них формируем описание. Процедура пошагово выглядит следующим образом:

1) Выбираем очередной кластер.

2) Отбираем те слова и фразы, из которых будем формировать описание кластера. Наиболее важным фактором при отборе, как слов, так и фраз является их распространенность по документам кластера, т.е. информация о том в скольких документах кластера встретилось данное слово или фраза.

3) Формируем имя кластера. На этом этапе используются ненормализованные фразы, т.е. фразы, в которые входят ненормализованные слова. Заметим, что это возможно благодаря тому, что хранятся как нормализованные, так и ненормализованные фразы, причем их ID совпадают. Выбираем некоторое количество (задается параметром) неповторяющихся слов и фраз.

4) Формируем описание кластера. Для этого используем уже существующее имя кластера и к нему дописываем новые отсортированные по убыванию весов слова и фразы, количество которых задается параметром. Здесь при расширении имени возможна ситуация, когда слова во фразах совпадают.

5) Формируем текстовый фильтр. На основе слов, отобранных для описания кластера, формируем текстовый запрос, который понимает поисковый движок системы.

6) Определяем порог релевантности. Для вычисления релевантности задаем поисковый запрос, текстовый фильтр которого был ранее сформирован.

7) Возврат к пункту 1)

4.5. Распределение корпуса документов по кластерам

После того, как сформированы кластеры и центры, определен минимальный порог релевантности и вычислены обратные матрицы, следует распределить весь корпус документов по кластерам. При этом будем также использовать подходы, основанные на метриках Евклида и Евклида-Махаланобиса. Если документ прошел порог релевантности самого близкого кластера, то приписываем документ к кластеру. Иначе в соответствии с работой [5] относим документ к кластеру “другие”, содержащему документы, не попавшие по релевантности ни в один кластер. Были проведены эксперименты по качеству отнесения документов к уже сформированным кластерам, которые показали, что кластеризация на основе Евклида-Махаланобиса выполняется точнее, что объясняется учетом статистических данных документов внутри кластера. Таким образом, в целом, обобщенная метрика точнее определяет кластер, к которому необходимо отнести документ.

Заключение

Задачи выбора числа кластеров и их первоначального расположения, а также вопросы распределения текстовых документов на естественном языке по кластерам продолжают оставаться актуальными в настоящее время. В ходе проведенных экспериментов показано, что для первоначального выбора числа кластеров и распределения их центров целесообразно использовать иерархический подход, основанный на последовательном объединении и усредне-

нии характеристик наиболее близких документов ограниченной выборки. Проведенные предварительные эксперименты показывают целесообразность использования набора метрик на разных этапах кластеризации. Причем метрику Евклида следует использовать непосредственно в процессе формирования кластеров, а метрику Евклида-Махаланобиса на этапе распределения всего корпуса документов. Применение обобщенной метрики на заключительном этапе дает некоторое улучшение качества кластеризации, оцениваемое в среднем, на 15-20%. Очевидно кластеры формируются точнее за счет использования данных вероятностного распределения документов внутри кластера. Выполненная экспериментальная проверка имеет ограниченный характер и, разумеется, не может гарантировать конечного качества кластеризации для всех возможных приложений. Тем не менее, ее результаты не противоречат выдвинутым предположениям и рекомендациям по построению алгоритмов кластеризации текстовых документов. Другие предложения автора по совершенствованию методов кластеризации содержатся в работах [12,13].

Литература

1.Кириченко К.М, Герасимов М.Б. Обзор методов кластеризации текстовой информации. http://www.dialog-21.ru/Archive/2001/volume2/2_26.htm

2. Паклин Н. Алгоритмы кластеризации на службе Data Mining. http://www.basegroup.ru

3. Амелькин С.А., Захаров А.В., Хачумов В.М. Обобщенное расстояние Евклида-Махаланобиса и его свойства. – Информационные технологии и вычислительные системы, № 4, 2006, с.40-44.

4.Миркес Е.М. Нейроинформатика. – Красноярск: Издательство Красноярского государственного технического университета, 2003. http://www.softcraft.ru/neuro/ni/p00.shtml

5.Журавлев Ю.И. Об алгебраическом подходе к решению задач распознавания или классификации. – Проблемы кибернетики, 1978, т.33, с. 5-68.

6.Загоруйко Н.Г. Прикладные методы анализа данных и знаний. – Новосибирск: Изд-во Института математики, 1999. –270 с.

7. 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

8. 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

9. 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

10.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

11. Экспертная система «Айкумена». http://www.iqmen.ru/

12. Атаманов В.В., Козачок М.А., Трушков В.В., Хачумов М.В. Выбор первоначального расположения кластеров в нейронной сети Кохонена. – Нейрокомпьютеры: разработка и применение, №1, 2009, с.73- 76.

13. Талалаев А.А., Тищенко И.П., Хачумов М.В. Выделение и кластеризация текстовых и графических элементов на полутоновых снимках. – Искусственный интеллект и принятие решений, № 3, 2008, с.72-84.

Хачумов Михаил Вячеславович. Аспирант РУДН. Окончил Российский университет дружбы народов в 2009 году. Имеет 5 печатных работ. Область научных интересов: искусственный интеллект, машинная графика, кластеризация. E-mail: khmike@inbox.ru, vmh48@mail.ru.