Раздел IV. Анализ данных и управление знаниями

УДК 004.822 DOI 10.23683/2311-3103-2018-4-154-166

В.В. Бова, С.Н. Щеглов, Д.В. Лещанов

МОДИФИЦИРОВАННЫЙ АЛГОРИТМ ЕМ-КЛАСТЕРИЗАЦИИ ДЛЯ ЗАДАЧ ИНТЕГРИРОВАННОЙ ОБРАБОТКИ БОЛЬШИХ ДАННЫХ*

Приводится возможное решение проблем структуризации данных большого объема, а также их интегрированного хранения в структурах, обеспечивающих целостность, непротиворечивость их представления, высокую скорость и гибкость процессов обработки неструктурированной информации. Для решения указанных проблем предложен метод построения многоуровневой онтологической структуры, обеспечивающей решение взаимосвязанных задач выявления, структуризации и обработки больших массивов данных, преимущественно естественно-языковых форм представления. Разработанная на основе методов семантического анализа и онтологического моделирования многоуровневая модель, пригодна для интерпретации и эффективной интегрированной обработки неструктурированных данных, полученных из распределенных источников информации. Многоуровневое представление модели структуризации данных большого объема определяет способы и механизмы унифицированного метаописания элементов данных на логическом уровне, поиска закономерностей и классификации признакового пространства на семантическом уровне и лингвистический уровень реализации процедур выявления, консолидации и обогащения данных. В качестве возможного решения данной задачи предлагается метод и алгоритм кластерного анализа, который позволяет сократить размерность исходного набора данных и выявить семантические ареалы терминологического покрытия. Модификация данного метода заключается в применении масштабируемого и вычислительно эффективного генетического алгоритма поиска и генерации весовых коэффициентов, которые соответствуют разным мерам подобия множества наблюдаемых признаков, использующихся при формировании модели кластеризации данных. Полученные данные в серии вычислительных экспериментов подтвердили теоретическую значимость и перспективность применения метода кластеризации с ГА оценки семантической близости элементов данных, представленных в онтологии.

Семантическая близость; онтология; семантическая сеть; нестуктурированные данные; большие данные; семантический анализ; семантическая метамодель; генетический алгоритм; кластеризация.

V.V. Bova, S.N. Scheglov, D.V. Leshchanov

MODIFIED EM-CLUSTERING ALGORITHM FOR INTEGRATED BIG DATA PROCESSING TASKS

The paper presents a possible solution to the problems of structuring large-scale data, as well as their integrated storage in structures that ensure the integrity, consistency of their presentation, high speed and flexibility of processing of non-structured information. To solve these problems, we propose a method for constructing a multilevel ontological structure that provides a solution to the interrelated tasks of identifying, structuring, and processing large data sets, predominantly natural language forms of representation. Developed on the basis of methods of semantic

* Работа выполнена при финансовой поддержке РФФИ (проекты: № 18-07-00055, № 17-07-00446).

analysis and ontological modeling, a multilevel model is suitable for the interpretation and efficient integrated processing of unstructured data obtained from distributed sources of information. The multilevel representation of the large-scale data structuring model determines the methods and mechanisms of the unified meta-description of the data elements at the logical level, the search for patterns and the classification of the characteristic space at the semantic level, and the linguistic level of implementation of the procedures for identifying, consolidating and enriching data. As a possible solution to this problem, we propose a method and algorithm for cluster analysis that reduces the dimension of the initial data set and reveals the semantic areas of terminological coverage. Modification of this method consists in applying a scalable and computationally effective genetic algorithm for searching and generating weight coefficients that correspond to different measures of the similarity of the set of observed features used in the formation of the data clustering model. The data obtained in a series of computational experiments confirmed the theoretical significance and prospects of applying the clustering method with GA to assess the semantic proximity of data elements presented in ontology.

Semantic similarity; ontology; semantic network; unstructured data; Bigdata; semantic analysis; semantic meta-model; genetic algorithms; clustering

Введение. Современный этап развития направлений исследований в области обработки больших данных со сложной структурой и цифровых технологий определил переход к семантическим технологиям, что потребовало обязательного учёта проектирования формальных онтологий предметных областей современных многоцелевых информационных систем (МИС), представленных в виде семантических сетей. Актуальность проведения исследования вызвана необходимостью развития методов анализа сложной распределенной информации для задач разработки моделей структуризации данных больших объемов на основе интеграции методов онтологического моделирования и семантического анализа. В работе приводится возможное решение проблемы структуризации данных большого объема, а также их интегрированного хранения в структурах, обеспечивающих целостность, непротиворечивость их представления, высокую скорость и гибкость процессов обработки неструктурированной информации. Такие структуры характеризуются свойствами частичной наблюдаемости, динамики, непрерывности, стохастичности. Для решения указанных проблем предлагается метод формирования многоуровневой онтологической структуры с интерпретацией метаданных, для задач выявления и консолидации элементов данных из различных предметных областей (ПрО). Разработанные на его основе модель и алгоритм кластеризации позволит решить задачу обеспечения структурной и семантической интероперабельности больших данных, представленных в различных предметных областях МИС.В качестве возможного решения данной задачи предлагается метод и алгоритм кластерного анализа, который позволяет сократить размерность исходного набора данных и выявить семантические ареалы терминологического покрытия. Модификация данного метода заключается в применении масштабируемого и вычислительно эффективного генетического алгоритма поиска и генерации весовых коэффициентов, которые соответствуют разным мерам подобия множества наблюдаемых признаков, использующихся при формировании модели кластеризации данных.

Методологический анализ проблемы исследования. Задача кластеризации больших массивов данных встречается во многих прикладных областях, связанных с разработкой средств организации и структурирования гипертекстового пространства (концепция Semantic Web, разработка и администрирование веб-сайтов, веб-аналитика и др). На сегодняшний день разработано довольно много методов иалгоритмов кластеризации, однако далеко не все из предложенных способовмогут работать с потоками больших объемов [2–7], поступающей на обработку в online режиме [1]. Для решения этих задач может быть применен математический аппарат computational intelligence и, прежде всего, искусственные нейронные сети и soft computing [4–6].

Исследования и анализ современных подходов в области семантического анализа и обработки неструктурированных данных показал, что наибольшее распространение получили следующие методы кластеризации: алгоритмы на основе нейронных сетей, k-means, principal component analysis (PCA), EM-алгоритм [5–7].

Метод «k-средних» является количественным, т.е. непосредственно перед его использованием в той или иной системе необходимо все качественные характеристики, если таковые необходимо обработать, стоит привести к количественным, иначе обработка таким способом будет невозможна [2]. Одним из важных недостатков такого распространенного алгоритма является его неустойчивость к так называемым «выбросам». Здесь имеется ввидуситуация, когда в данных присутствует множество «шумов», что неизбежно приводит к ситуации, когда центры кластеров начинают сильно сдвигаться. Для решения этой проблемы используют медианы. Такая небольшая модификация алгоритма делает его менее чувствительным к «шумам» и «выбросам», поскольку медиана меньше подвержена их влиянию. В литературе такую модификацию алгоритма «k-средних» называют PartitioningAround Method (PAM) [24].

Другой подход алгоритмов кластеризации потоков данных основывается на основе анализа плотности обрабатываемых данных. Такие алгоритмы умеют различать кластеры исходя из формы, которая зависит от плотности обрабатываемых данных. Таким образом, если пара точек находится достаточно близко друг к другу, а область вокруг них плотная, то эти точки объединяются в кластер. Алгоритмы DBSCAN [20], OPTICS[21] и DENCLUE [22] являются примерами алгоритмов кластеризации на основе плотности. С недавнего времени алгоритмы такого типа стали применяться для анализа потоков данных. [21].

Так же примерами алгоритмов кластеризации такого типа могут служить алгоритмы GMDBSCAN [20-24] и ISDBSCAN [20–24]. Эти алгоритмы являются алгоритмами, которым требуется дважды проходить по обрабатываемым данным. За первый проход извлекается вся необходимая для анализа информация, а затем на ее основе происходит группировка данных в кластеры. Оба алгоритма не подходят для кластеризации потока данных, так как данные поступают непрерывно (потоком), а значит необходим алгоритм, который проводит обработку и формирование кластеров за одно сканирование поступающей информации.

DBSCAN-DLP это много-плотностный DBSCAN на основе разделения по уровням плотности [20]. Алгоритм определяет параметры каждого кластера с целью автоматического обнаружения кластеров с различающейся плотностью, используя разделение по уровням плотности. В этом способе сначала набор данных делится по разным уровням плотности на основе статистической информации об ее изменениях. Затем, для каждого уровня вычисляется параметр є. На последнем этапе работы алгоритма, для получения итоговых кластеров DBSCAN использует для кластеризации на каждом уровне плотности соответствующее значение параметра є. DBSCANDLP является алгоритмом кластеризации в два этапа обработки данных, который к тому же, имеет высокое вычислительное время, что делает его не подходящим для обработки потоков данных.

Каждый из методов имеет свои достоинства, но обладает и рядом ограничений: исходные данные должны иметь случайную природу и подчиняться нормальному закону распределения; возможно «застревание» процесса оптимизации в локальных экстремумах; вычислительная сложность; массив данных, подлежащих кластеризации, задан заранее и не изменяется в процессе обработки. Данные методы вычислительного интеллекта должны быть существенно модифицированы для обработки больших объемов информации. В качестве возможного решения данной задачи предлагается метод и алгоритм кластеризации потока данных, взвешенных по времени поступления. Алгоритм кластерного анализа позволяет сократить размерность исходного набора данных и выявить семантические ареалы терминологического покрытия на основе ГА оценки семантической близости.

1. Постановка задачи построения многоуровневой структуры онтологии. Для задачи структурирования данных в рамках одной системы понятийной концептуализации предлагается модель онтологии, ориентированная на обработку естественно-языковых описаний элементов данных, необходимых для создания и наполнения базы знаний онтологического типа.

В общем виде постановка задачи исследования может быть сформулирована следующим образом. Для заданной пятерки <O,F,D,Term,S>, где O – онтология ПрО, расширенная лингвистическим уровнем, F– множество моделей метаданных (схем фактов), D – текстовый фрагмент метаописания (далее документ), Term – терминологическое покрытие D, C – сегментное покрытие (кластеры) D, найти все семантические структуры, соответствующие онтологии O, покрывающие область D, которые можно получить в процессе применения правил из Fк Term с учетом C.

Выделение структуры предметной области (онтологии) – это основная задача по приведению неструктурированных данных к структурированному виду. Весь процесс построения онтологии разбит на несколько независимых этапов решения определенной задачи, результаты которой служат исходными данными для задачи следующего более сложного уровня. Выделим следующую последовательность действий: извлечение из документов на естественном языке терминов-кандидатов → разбиение терминов на группы (кластеризация) → присвоение обобщающего понятия-концепта каждой группе → определение отношений между концептами → формирование правил вывода (расширения концептов словаря метаописаний).

Проблемы семантического моделирования структуры данных обычно связаны именно с неструктурированными данными, представленными естественно-языковыми описаниями. На рис. 1 представлена логическая схема построения формальной онтологии на основе семантического анализа понятий и отношений во множестве естественно-языковых описаний данных.

Рис. 1. Обобщенная схема построения онтологической модели

Общим для всех формализаций является выделение множества объектов (концептов, понятий), алфавита отношений, правил установления отношений и аксиом, задающих правила вывода на множестве отношений.

Для построения модели онтологии как расширяемого тезауруса применяются: алгоритмы формирования и пополнения лингвистических шаблонов (словарей), механизмы распознавания значений атрибутов, эвристические правила, позволяющие выявлять зависимость между концептами в онтологии и отношения между объектами схемы фактов.

Тогда абстрактную модель онтологии O можно формально представить следующим образом:

$$O=\langle W\{X, E_x, N_o\}, V\{I, S, D_s\}, R, A\rangle,$$

где W – тезаурус ПрO, для которого представлено X – множество концептов онтологии – обобщающий класс понятий (терминов), обладающих одинаковыми свойствами и отношениями; Ex – множество экземпляров понятий и No – множество имен понятий, для которых задано отображение Ex: $X \rightarrow 2^{E_x}$;

V – словарь лингвистических образов (шаблонов, паттернов) метаописаний, включающийI – множество информационных входов (языковых выражений, значения которых представлены в W); Sij ={xi, ij} – множество отношений семантической связности между I и X;Ds– отображение множества схем фактов заданных документовна информационные входы и понятия тезауруса ПрOD_s: $(F, D)\rightarrow (I, X)$;

R – совокупность отношений $R = \{R_1, R_2, \dots, R_n\}$междупонятиями W и V, определяемых эвристическими алгоритмами поиска правил анализа связности;

A – аксиомы, основанные на свойствах транзитивности и наследования.

2. Модифицированный метод EM-кластеризации данных. Рассмотрим модифицированный метод EM-кластеризации, основанный на применении GA нахождения весовых коэффициентов для оценки семантической близости данных, последовательно поступающих на обработку в on-line режиме. Одним из широко известных и используемых алгоритмов кластеризации, который наиболее эффективно работает с большими объемами поступающих данных, является Expectation-Maximization (EM) алгоритм [3]. Область применения данного алгоритма достаточна широка, он используется не только для кластеризации данных, но и в дискриминантном анализе, а также для восстановления пропусков в данных. В основе алгоритма лежит методика интерактивного вычисления оценок максимального правдоподобия. В нем вместо центров кластеров предполагается наличие функции плотности вероятности распределения для каждого кластера. Основной алгоритм разделен на два шага.

На E-шаге вычислим $P(t|w,d)^{(r)}$. На этапе предварительной обработки и анализа будем рассматривать документ как «набор понятий» с численными характеристиками встречаемости терминов. Вероятность того, что термин w, принадлежащий формируемому тезаурусу W, встречается в метаописании d (множества D обрабатываемых документов) и принадлежит определенной ПрOt:

$$P(w|d) = \sum_{t\in T} P(w|t)\,(t|d), \qquad (1)$$

где t – элемент множества T предметных областей.

Параметры предварительного семантического анализа $P(w|t)$ и $P(t|d)$ определим следующим образом. Пусть r – число итераций.

$$P(t|w,d)^r = \frac{P(w|t)^{(r-1)}P(t|d)^{(r-1)}}{\sum_{t\prime\in T} P(w|t\prime)^{(r-1)}P(t\prime|d)^{(r-1)}}, \qquad (2)$$

На М-шаге оценим параметры:

$$P(w|t)^r = \frac{\sum_{d\in D} N(w,d)P(t|w,d)^r}{\sum_{w\prime\in W}\sum_{d\in D} N(w',d)P(t|w',d)^r}$$

и

$$P(t|d)^r = \frac{\sum_{w\in W} N(w,d)P(t|w,d)^r}{\sum_{t\prime\in T}\sum_{w\in W} N(w,d)P(t\prime|w,d)^r}, \quad (3)$$

где *N(w,d)* – число вхождения элемента тезауруса *w* в рассматриваемый документ *d*. Процесс обучения повторяется до сходимости параметров. Однако параметры часто попадают в область локального оптимума. Для повышения эффективности и управления скоростью обучения введем параметр $0<\beta\leq 1$.

Тогда на М-шаге выражение имеет вид:

$$P(t|w,d)^r = \frac{(P(w|t)^{(r-1)}P(t|d)^{(r-1)})^\beta}{\sum_{t\prime\in T}(P(w|t')^{(r-1)}P(t'|d)^{(r-1)})^\beta}. \quad (4)$$

Далее определим *W(w,t)* и *D(d,t)*суммарные вероятности как:

$$W(w,t)^r = \sum_{d\in D} N(w,d)P(t|w,d)^r$$

и

$$D(d,t)^r = \sum_{w\in W} N(w,d)P(t|w,d)^r. \quad (5)$$

Из (4) получим:

$$W(w,t)^r = \sum_{d\in D} \frac{N(w,d)(P(w|t)^{(r-1)}P(t|d)^{(r-1)})^\beta}{\sum_{t\prime\in T}(P(w|t')^{(r-1)}P(t'|d)^{(r-1)})^\beta}, \quad (6)$$

$$D(d,t)^r = \sum_{w\in W} \frac{N(w,d)(P(w|t)^{(r-1)}P(t|d)^{(r-1)})^\beta}{\sum_{t\prime\in T}(P(w|t')^{(r-1)}P(t'|d)^{(r-1)})^\beta}. \quad (7)$$

Для формирования отношений семантической сети тезауруса и вычисления семантической близости *sim (a, b)* между терминами *a* и *b* введем ϖ– весовую функцию, определенную над множеством семантических отношений *R(a; b)*, выражающую силу семантической связи между *a* и *b*. Отношение *R(a; b)* представляется набором лексических паттернов. Обозначим частоту встречаемости для пары *(a; b)* как *f(r; a; b)*.

$$\varpi\big(R(a,b)\big) = \sum_{r_i\in R(a,b)} w_i \times f(r_i,a,b), \quad (8)$$

где w_i – вес, связанный с r_i и определяемый с использованием обучающей выборки. Модифицированный алгоритм обрабатывает потоки данных, взвешенные по времени поступления. Весовой коэффициент введен для того, чтобы была возможность учитывать актуальность поступающих в кластер данных.

Для преодоления указанных ограничений и определения семантически связанных терминов разработан модифицированный алгоритм кластеризации (рис. 2), в котором приняты следующие обозначения: *P* – вектор частот пар *(a, b)*соответственно принадлежащих множествам *D*и *Term*, *f*(a_i; b_i; *p*), в лексических паттернах *p; Θ – порог подобия (задается экспертом); SORT – функция сортировки по общей встречаемости в парах (a, b). Косинусный коэффициент применяется для вычисления подобия между* p_i *и центроидом кластера* c_i.

Для повышения эффективности расчета меры семантической близости и определения семантически связанных терминов предлагается использовать генетический алгоритм (ГА) (рис. 3) [17]. Для нахождения оптимальных значений коэффициентов по методу максимизации оценки близости в ГА определена целевая функция:

$$F = max\ (\varpi\ (R\ (a, b)) \quad (9)$$

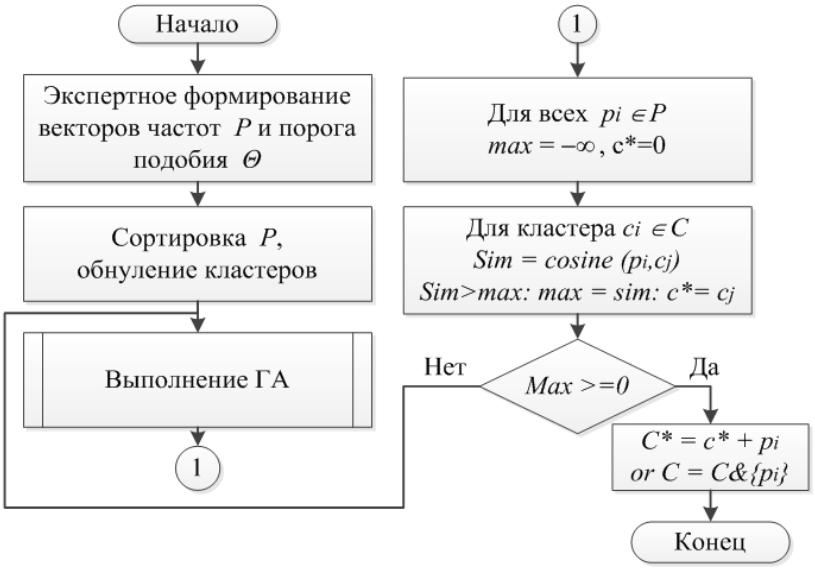


Рис. 2. Схема модифицированного алгоритма кластеризации

Рассмотрим работу модифицированного генетического алгоритма, который используется в представленном методе EM-кластеризации. Достоинства данного подхода заключаются в ограничении области поиска решений, удовлетворяющих запросам лица принимающего решения, сокращения времени обработки исходных данных, возможности интерактивной настройки параметров работы алгоритма от сложности поставленной задачи. Приведем основные этапы работы ГА расчета меры семантической близости и определения связанных терминов.

Создание начальной популяции – один из важнейших этапов эволюционного моделирования. В начальной популяции должны быть представлены все возможности генетического материала. В данном подходе предлагается использовать следующие принципы формирования начальной популяции: «одеяла», «дробовика», «фокусировки» и их комбинаций. Выбор того или иного принципа зависит от размерности поставленной задачи, предполагаемого времени получения приемлемого решения, оценки неструктурированности входных данных. В приведенных ниже экспериментальных исследованиях в основном использовался принцип дробовика для того, чтобы разнообразить генофод на начальных стадиях работы ГА.

Первый этап предполагает различные варианты формирования начальных параметров работы алгоритма. Возможна автоматическая генерация параметров элементов модели близости. Наиболее приемлемым является ввод значения весовых коэффициентов wi и вероятностей для операторов кроссинговера и мутации на основе экспертных данных, либо предпочтений лица принимающего решение.

Формирование начальной популяции происходит на основе имеющихся обучающих данных из множества C = {ci}. Для каждого элемента выявляются семантические отношения ri. Вектор, который описывает термины a и b в логическом представлении кластера ci, представляет элемент хромосомы (ген). Оценивание текущей популяции происходит на основе вычисления значения целевой функции по формуле (9). Исходя из полученных данных, происходит выполнение оператора селекции. В данном алгоритме применяется элитная селекция–выбираются лучшие решения.

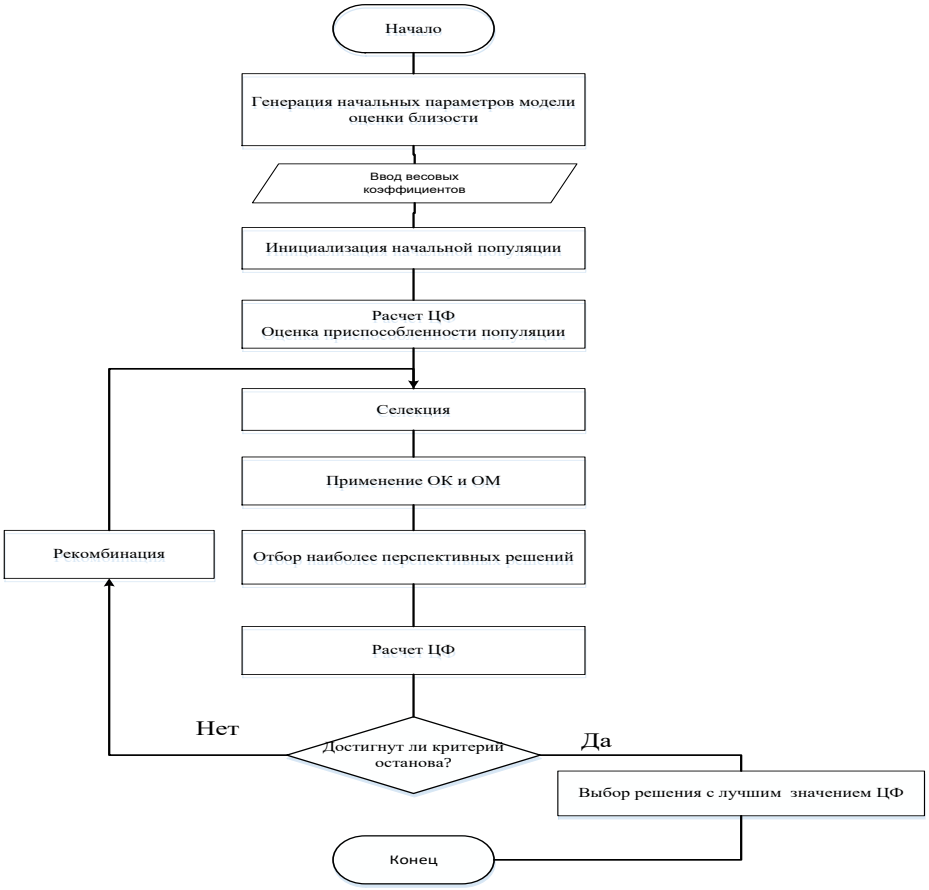


Рис. 3. ГА расчета меры семантической близости

Для каждой пары, отобранных в результате селекции родительских хромосом, применяются операторы кроссинговера и мутации с предварительно заданной вероятностью. Генерирование нового набора особей для каждой пары отобранных родительских хромосом производится с использованием операторов кроссинговера и мутации с предварительно заданной вероятностью. Скрещивание хромосом выполняется случайно с вероятностью Pc. Точка скрещивания определяется случайным образом в заданном интервале.

Исходя из результатов скрещивания, начинается процесс работы оператора мутации. Данный оператор производит изменения значения гена в хромосоме потомка. Изменение происходит на основе случайного выбора параметра из интервала [0,1] с вероятностью Pm. Отбор наиболее перспективных решений производится при помощи оператора рекомбинации.

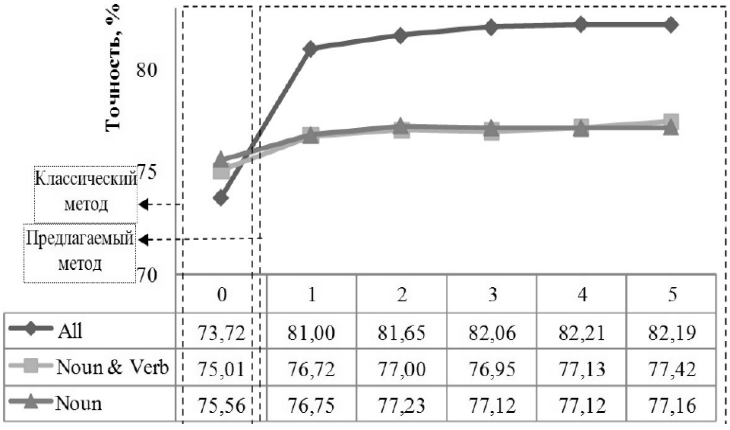
Завершение работы алгоритма осуществляется проверкой критерия останова. Если данный критерий не достигнут, то процесс переходит на следующую итерацию. Временная сложность разработанного алгоритма ориентировочно составляет O(n2).

3 Экспериментальные исследования. С целью анализа эффективности работы предложенных методов был проведен ряд вычислительных экспериментов.

В качестве тестовых данных используется набор данных (http://people.csail.mit.edu/jrennie/20Newsgroups), который предназначен для тестирования метода кластеризации.

При помощи этого набора были выполнены эксперименты для проверки точности предлагаемого метода кластеризации и размера межкластерной корреляционной матрицы совместной встречаемости. Зависимость точности классификатора от используемого алгоритма и значения расстояния z представлена на рис. 4. В качестве оценки близости векторов используется косинусная мера подобия. Горизонтальная ось представляет собой значения коэффициента *z*, а вертикальная ось – значения точности классификации двумя алгоритмами (EM и ГА). Считается, что термин t_i встречается вместе с термином t_j, если расстояние между ними в документе не превышает предела *z*: $|i–j|\leq z$.

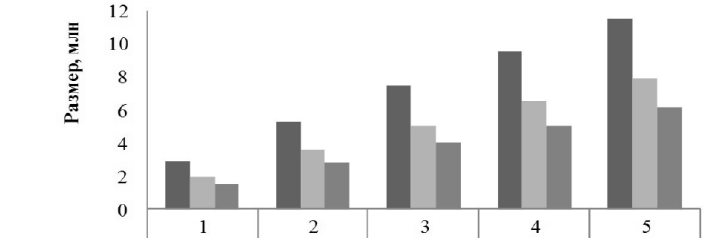
Три серии данных соответствуют значениям точности, полученным при использовании методов кластеризации разных терминов: *All* – используются все терминов; *Noun&Verb* – используются только существительные и глаголы; *Noun* – используются только существительные.



	0	1	2	3	4	5
All	73,72	81,00	81,65	82,06	82,21	82,19
Noun & Verb	75,01	76,72	77,00	76,95	77,13	77,42
Noun	75,56	76,75	77,23	77,12	77,12	77,16

Рис. 4. Графики оценки точности кластеризации EM-алгоритмом и ГА

Графики зависимости размера межкластерной корреляционной матрицы совместной встречаемости терминов и значения расстояния *z* представлены на рис. 5.



	1	2	3	4	5
All	2,89	5,31	7,53	9,59	11,54
Noun & Verb	1,96	3,59	5,11	6,54	7,89
Noun	1,56	2,83	4,01	5,12	6,17

Рис. 5. Графики зависимости размера корреляционной матрицы встречаемости терминов и значения расстояния

Исходя из полученных данных видно, что в случае *z=0* (классический EM-алгоритм), максимальная точность (75,56 %) получается при сохранении в коллекции документов только существительных. Однако в остальных случаях *z>0*

(используется оценка семантической близости ГА), наивысшая точность (≈82 %) получается при использовании всех терминов.

Исходя из полученных данных, видно, что в результате кластеризации на основе ГА оценки семантической близости терминов значительно уменьшается размер матрицы совместной встречаемости по сравнению со случаем использования всех терминов: приблизительно 30 % в случае использования только существительных и глаголов и приблизительно 50 % в случае использования только существительных.

В работе выполнено сравнение предлагаемого метода с известными методами кластеризации PCA и k-mean по критерию точности, полноты поиска и f-критерия (табл. 1).

Таблица 1

Результаты оценки и сравнения с методами кластеризации

Метод	Точность	Полнота поиска	f-критерий
Предложенный метод	0,81428	1,00	0,8976
K-means	0,7142	0,86	0,7803
PCA	0,3333	1,00	0,0073

Согласно результатам, представлены в табл. 1, эффективность предлагаемого метода составила 0,814 для критерия точности (высокая). По другим критериям предлагаемый метод также позволил получить более высокие результаты, что говорит о его эффективности.

Заключение. В работе предложен метод формирования многоуровневой онтологической структуры с интерпретацией метаданных, для задач выявления и консолидации элементов данных из различных ПрО. Разработанные модель и алгоритм кластеризации позволит решить задачу обеспечения структурной и семантической интероперабельности больших данных, представленных в различных предметных областях МИС. Полученные данные в серии вычислительных экспериментов подтвердили теоретическую значимость и перспективность применения метода кластеризации с ГА оценки семантической близости элементов данных, представленных в онтологии.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Bova V.V., Kureichik V.V., Leshchanov D.V.* The model of semantic similarity estimation for the problems of big data search and structuring // Application of Information and Communication Technologies - AICT 2017. – P. 27-32.
2. *Кравченко Ю.А., Марков В.В., Новиков А.А.* Семантический поиск в SemanticWeb // Известия ЮФУ. – 2016. – № 6 (179). – С. 65-75.
3. *Mumford C., Jain L.* Computational Intelligence. Collaboration, Fuzzyand Emergence. – Berlin: Springer-Verlag, 2009. – 726 p.
4. *Kureichik V., Zaporozhets D., Zaruba D.* Generation of bioinspired search procedures for optimization problems // Application of Information and Communication Technologies, AICT 2016 - Conference Proceedings 10. – 2016. – С. 7991822.
5. *Кулиев Э.В., Кравченко Ю.А., Логинов О.А., Запорожец Д.Ю.* Метод интеллектуального принятия эффективных решений на основе биоинспирированного подхода // Известия Кабардино-Балкарского научного центра РАН. – 2017. – № 62 (80). – С. 162-169.
6. *Бова В.В., Кулиев Э.В., Лещанов Д.В.* Концептуальные основы автоматизированной обработки неструктурированной информации в системах управления проблемно-ориентированными знаниями // В сб. "IS&IT'17". – 2017. – С. 341-350.
7. *Бова В.В., Лежебоков А.А., Лещанов Д.В.* Моделирование семантической сети представления знаний на основе онтологического подхода // Информатизация и связь. – 2018. – № 4. – С. 78-88.

8. *Kureychik V., Semenova A.* Combined method for integration of heterogeneous ontology models for big data processing and analysis // Advances in Intelligent Systems and Computing. – 2017. – Vol. 573. – P. 302-311.
9. *Kureichik V., Safronenkova I.* Integrated algorithm of the domain ontology development // Advances in Intelligent Systems and Computing. – 2017. – Vol. 573. – P. 146-155.
10. *Харченко А.М.* Адаптивный расчет функции для динамического ЕМ-алгоритма // Математика. – 2015. – С. 134.
11. *Хоанг В.К., Тузовский А.Ф.* Методы определения уровней безопасности элементов онтологии // Известия Томского политехнического университета. – 2013. – Т. 322, № 5. – С. 148-152.
12. *Loukachevitch N., Dobrov B.* Ontological resources for representing security domain in information-analytical system // Открытые семантические технологии проектирования интеллектуальных систем. – 2018. – Т. 2, № 8. – С. 185-191.
13. *Копайгородский А.Н., Семичева О.А* Семантическая информационная система для представления научной деятельности в сети интернет // Вестник Иркутского государственного технического университета. – 2014. – № 12. – С. 23-29.
14. *Кулиев Э.В., Лежебоков А.А., Лещанов Д.В., Шкаленко Б.И.* Механизмы роевого интеллекта и эволюционной адаптации на основе виртуального набора популяций для решения задач управления проблемно-ориентированными знаниями // Информатика, вычислительная техника и инженерное образование. – 2017. – № 1 (29). – С. 34-45.
15. *Gladkov L.A., Sheglov S.N., Gladkova N.V.* The application of bioinspired methods for solving vehicle routing problems // Procedia Computer Science, 120 (2017). 9th International Conference on Theory and Application of Soft Computing, Computing with Words and Perception, ICSCCW 2017. – P. 39-46.
16. *Кулиев Э.В., Шеглов С.Н., Пантелюк Е.А., Логинов О.А.* Адаптивный алгоритм стаи серых волков для решения задач проектирования // Известия ЮФУ. Технические науки. – 2017. – № 7 (192). – С. 28-38.
17. *Логинов О.А., Лежебоков А.А., Бова В.В., Щеглов С.Н.* Интеллектуальный анализ данных на основе биоинспирированного подхода // Информатизация и связь. – 2018. – № 4. – С. 66-71.
18. *Щеглов С.Н.* Использование онтологий в системах поддержки принятия решений // Конгресс по интеллектуальным системам и информационным технологиям IS-IT'17: Труды конгресса. – 2017. – Т. 1. – С. 242-252.
19. *Кравченко Ю.А., Коваленко М.С.* Разработка инструментальной среды обработки данных // Конгресс по интеллектуальным системам и информационным технологиям IS-IT'17: Труды конгресса. – 2017. – Т. 3. – С. 211-218.
20. *Esfandani Gholamreza, Abolhassani Hassan*. MSDBSCAN: multidensity scale-independent clustering algorithm based on DBSCAN // In: Advanceddata mining and applications. Chongqing, China: Springer, 2010. – P. 202-13.
21. *Carmelo Cassisi, Alfredo Ferro, Rosalba Giugno, Giuseppe Pigola, Alfredo Pulvirenti.* Enhancing density-based clustering: parameter reduction andoutlier detection // In: Syst. – 2013. – Vol. 38 (3). – P. 317-30.
22. *Макаров И.Е.* Автоматизация анализа проектных решений с применением методов интеллектуальной обработки // Интеллектуальные системы. – 2014. – № 10. – С. 26-27.
23. *Bernhard Pfahringer*. Data stream mining: a practical approach. – Режим доступа: http://voxel.dl.sourceforge.net/project/moadatastream/ StreamMining.pdf.
24. *Газиев Г.З., Курдюкова Г.Н., Курдюков В.В.* Кластеризация Big Data для их анализа и обработки // Сб. научных статей конференции «Направления и механизмы развития науки нового времени: от теории до внедрения результатов». – 2017. – С. 150-162.