УДК 004.832.32

# МЕТОДОЛОГИЯ ФОРМИРОВАНИЯ СИСТЕМООРГАНИЗУЮЩИХ ХАРАКТЕРИСТИК ТЕКСТОВЫХ ДАННЫХ

**Ломакин Д.В., Ломакина М.Д., Суркова А.С.**

*Нижегородский государственный технический университет им. Р.Е. Алексеева, Нижний Новгород, e-mail: ansurkova@yandex.ru, dlomakin@list.ru*

Разработана методология системного анализа многомерных объектов на основе принципа скрытых параметров, которая продемонстрирована на примере анализа текстовых данных. Предложена концепция формирования системоорганизующих характеристик текстовых данных как многомерных объектов. На основе диалектического метода введено определение понятия «система» с учетом иерархической организации многомерных объектов, что позволило сформировать системоорганизующие характеристики, в качестве которых выбраны скрытые параметры, вычисляемые как функции наблюдаемых параметров. Свойство скрытых параметров заключается в том, что каждому значению скрытого параметра соответствует некоторое подмножество наблюдаемых параметров. Количество скрытых параметров и выбор вида соответствующих функций определяются критерием, лежащим в основе решаемой задачи. Предложенная методология формирования системоорганизующих характеристик текстовых данных может найти практическое применение в задачах развития и совершенствования методов кластеризации, классификации и идентификации текстовых данных.

**Ключевые слова: система, многомерные объекты, текстовые данные, кластеризация, классификация, идентификация**

# FORMING METHODOLOGY OF SYSTEM-ORGANISING TEXT DATA CHARACTERISTICS

**Lomakin D.V., Lomakina M.D., Surkova A.S.**

*R.E. Alekseev Nizhny Novgorod State Technical University, Nizhny Novgorod, e-mail: ansurkova@yandex.ru, dlomakin@list.ru*

The methodology of the system analysis of multidimensional objects is based on the principle of hidden parameters, which is demonstrated in terms of text data analysis. The concept of forming of system-organising text data characteristics as multidimensional objects is proposed. The definition of the system as the hierarchical organization of multi-dimensional objects is introduced on the basis of the dialectical method. That allows to form system-organising characteristics as hidden parameters, which is calculated as a function of observable parameters. Property of hidden parameters is that a subset of observable parameters corresponds to each value of the hidden parameter. The number of hidden parameters and the choice of the corresponding functions are determined by the criterion that depends on the problem being solved. The proposed methodology of forming of system-organising text data characteristics may be used for practical applications in tasks of text data clustering, classification and identification.

**Keywords: system, multidimensional objects, text data, clustering, classification, identification**

В настоящее время при анализе и синтезе объектов различной физической природы активно используется системный подход. Положительные результаты его применения стимулировали возникновение и развитие системного анализа как самостоятельного научного направления. Накоплен большой эмпирический и теоретический материал, но, к сожалению, системный подход часто используется формально и поэтому полностью не раскрывает свои возможности при решении поставленной задачи. Это связано прежде всего с тем, что не раскрыты полностью объективное содержание, гносеологическая и методологическая значимость понятия системы, ее сущность и роль в познании и преображении действительности. Очевидно, системный подход только тогда эффективен, когда есть четкое представление о сущности системы и хорошо разработана методология применения системного метода при анализе и синтезе объектов. Объект проявляет себя в окружающей среде как совокупность свойств, которые представляют собой, в частности, наблюдаемые физические величины или параметры. Совокупность этих величин образует модель объекта как структурированный состав, т.е. систему, при этом количество наблюдаемых величин в общем случае не ограничено. Таким образом, объект и соответствующая модель могут быть многомерными, в частности текст рассматривается как многомерный объект. Проблема обработки многомерных данных в настоящее время решается за счет высокой производительности вычислительных средств, но тем не менее остается актуальной проблема их оптимизации [1]. Основные задачи, которые приходится решать при обработке текстовых данных, – это задачи классификации, кластеризации и идентификации текстов, которые в работе решаются на основе теории скрытых параметров.

**Модель объекта как система**

В настоящее время существует достаточно большое количество определений понятия система и все они в той или иной степени несут на себе след той сферы действительности, в которой они используются. Особого внимания заслуживает определение системы Л.А. Зеленовым как «структурированного состава», приведенное им в работе [3]. Понятия «состав» и «структура», считает он, являются первичными в определении системы, а все остальные характеристики системы – вторичными. Кроме этого, в работе достаточно полно сформулирована аксиоматика, лежащая в основе определения понятия «система». Это определение носит номологический характер, относится к номологическому мироосвоению с явно выраженным объектом и методологией исследования и может себя хорошо зарекомендовать в общей теории систем. Что касается философского определения системы, то целесообразно обратиться к работе А.Ф. Лосева [7] и рассмотреть понятие «система» в рамках философских категорий и диалектического метода познания действительности. А.Ф. Лосев начинает свои исследования с раскрытия содержания понятия «одно», которое он считает исходным пунктом диалектики и последовательно раскрывает его содержание в философских категориях, создавая тем самым категориальный эйдос. Категориальный эйдос (лик, образ) в конечном счете, по А.Ф. Лосеву, можно представить как «единичность подвижного покоя самотождественного различия». Далее А.Ф. Лосев раскрывает содержание пар категорий «движение – покой» и «самотождественное различие» и тем самым демонстрирует методологию их использования в процессе мироосвоения. Мы предлагаем использовать это определение в качестве определения понятия «система» на философском плане, которое является более общим по сравнению с известными и применимо к описанию любой действительности, доступной человеческому осмыслению. Необходимость обращения к такому определению продиктована наступившим кризисом в современной науке, который связан не с концепцией самой науки, а с возникшим противоречием между объектом научных исследований и его концептуальной моделью, методологией, математическим и логическим инструментарием. Человек растерялся перед рождающейся новой реальностью, испугался исчезновения материи. Разрешение указанного противоречия требует переосмысления понятия реальности, а вместе с ним и аксиоматического определения объекта как выражения концентрированного опыта, связанного с некоторым фрагментом реальности. Предлагаемый подход к определению понятия системы представляется более продуктивным по сравнению с известными и способствует преодолению кризиса, наступившего в науке.

Все сущее, доступное человеческому разуму, – едино, образует иерархически организованную целостность, которая является содержанием понятия «Универсум». Любой объект, который также является системой, принадлежит одному из уровней Универсума и входит как компонент в состав системы (объекта), принадлежащей вышележащему уровню.

Таким образом, Универсум можно представить как многоуровневую систему, каждый уровень которой организован не только законами вышележащего уровня, но и специфическими для данного уровня законами, причем нижележащий уровень является субстратом, на котором проявляется, организуется вышележащий уровень. Механизм взаимодействия между системой и надсистемой подробно исследован А.И. Субетто [8], который сформулировал законы взаимодействия системы с надсистемой и со всеми вышележащими системами. Вершиной этой последовательности вложенных друг в друга систем, по-видимому, будет приведенное выше философское определение системы. Одним из основных законов, открытых А.И. Субетто, является системный закон дуальности управления и организации, который раскрывает динамику взаимодействия системы с надсистемой. Сущность закона заключается в том, что «генетическое управление развитием» системы разделяется на управление от прошлого и будущего. Управление от прошлого обеспечивает устойчивость в развитии благодаря консервативности системы, а управление от будущего определяется той «системной нишей», которая задается надсистемой и в рамках которой подсистема может выбрать направление своего развития. Все законы, открытые А.И. Субетто, не противоречат категориям, которые входят в философское определение системы, а просто дополняют их содержание. Следует отметить еще одно важное свойство любой системы – это ее одновременную замкнутость и разомкнутость. Любая система замкнута относительно операций, задаваемых ее внутренними законами в том смысле, что указанные операции не выводят систему за пределы ее качественной определенности при неограниченном их повторении. Внутренние законы системы определяются ее структурой, в которой бытует сущность системы. Сущность системы определяет ее границу, за пределы которой система не может выйти, не изменив своей качественной определенности. В своем

внутреннем развитии система может приближаться к своей границе неограниченно долго, если у нее не будет изменена качественная определенность, а этого можно достигнуть только за счет управления системой от будущего, функцию которого выполняет надсистема и ее надсистемы. В этом смысле система становится разомкнутой (открытой), причем взаимодействие системы с надсистемой происходит на границе, процессы на которой управляются законами системы и ее надсистемы. Таким образом, система рождается в «системной нише» надсистемы, которая задает цель и «коридор» ее развития, управляет процессом развития. В процессе становления системы формируется ее структура, которая представляет собой ее информационное содержание. Структура системы, как ее сущность, обнаруживает себя вовне как совокупность измеряемых параметров (признаков). Эта совокупность параметров представляет собой модель системы, которая также является системой. Любая модель не может в полном объеме отобразить истинное строение структуры, ее информационное содержание, и поэтому возникает вопрос, какие параметры необходимо выбрать и какое должно быть их количество. Параметром, который доставляет максимально возможное количество информации о системе, является сама структура, но наблюдение ее и полный ее анализ возможен только с позиций надсистемы, что, как правило, не представляется возможным. Количество параметров, доступных наблюдению, может быть неограниченным, и поэтому возникают проблемы их эффективной обработки с целью выявления их структурных закономерностей, но тем не менее эта проблема в настоящее время стала разрешимой благодаря использованию мощных вычислительных средств. Хорошо зарекомендовал себя метод обработки многомерных объектов на основе выявления скрытых параметров (признаков), т.е. параметров не доступных непосредственному наблюдению и измерению. Скрытые параметры определяются как функция наблюдаемых параметров, каждому значению которой, в общем случае, соответствует некоторое подмножество наблюдаемых параметров. Если создать совокупность функций, то тем самым все пространство параметров (признаковое пространство) можно разбить на подмножества, количество которых и вид функций определяются спецификой конкретно решаемой задачи и соответствующим критерием. Таким образом, формируется новая параметрическая структура, которая до этого была скрытой, формируется новое параметрическое пространство. Вычисляемая таким образом совокупность скрытых параметров представляет собой системоорганизующие характеристики объекта.

**Модель текста как система**

Текст – это произведение речетворческого процесса, который первоначально формируется на ментальном плане с помощью языка, отражает структуру языка, образуя некоторую целостность. Существование структуры языка не требует специального доказательства. «Достаточно вспомнить о том, что знаки сосуществуют в рамках языка как целого, будучи связанными некой системой отношений» [2], которые определяются грамматикой языка. Грамматика организует знаки языка в некоторую систему, независимо от их конкретного содержания. Очевидной является иерархическая организация текста – это уровень букв, слогов, слов и предложений. Существуют авторские и универсальные инварианты текста, которые являются скрытыми системоорганизующими характеристиками текста, то есть не даны в непосредственном его наблюдении, причем универсальные инварианты в большей степени проявляются на низших уровнях иерархической организации текста (буквы, слоги), а авторские – на верхних (предложения). В частности, к универсальным относится априори известный закон распределения вероятностей появления букв в тексте. Следует отметить, что вид закона распределения определяется не только в результате экспериментальных исследований, как это принято в теории вероятностей, но и является следствием проявления универсального закона золотой пропорции, который проявляется в любом фрагменте действительности [4]. Учет указанных априорных сведений позволяет повысить эффективность решения поставленной задачи.

**Задачи кластеризации, классификации, идентификации**

Предложенный подход выявления системоорганизующих характеристик текстовых данных в виде скрытых параметров позволяет повысить эффективность решения широкого круга прикладных задач обработки текстов. Основные задачи можно разделить на три большие группы [6]: задачи **кластеризации –** разбиение корпуса текстов на отдельные кластеры; задачи **классификации** – отнесение неизвестного текста к одному из заданных классов и задачи **идентификации** – определение значимых признаков, структур и основных параметров текстовых данных.

**Задачи кластеризации.** К таким задачам можно отнести разбиение на группы научных текстов, статей в специализированных информационно-поисковых системах, например кластеризация патентной документации и заявочных материалов, извлекаемых из разных патентных баз. Для решения задач кластеризации широко используются

обучающие алгоритмы на основе нейросетевых технологий, например могут быть применены самоорганизующиеся карты Кохонена; также успешно могут быть применены методы сжатия и понятие Колмогоровской сложности при представлении текстов.

**Задачи классификации.** Примером задач классификации является задача категоризации (разбиения по тематическим категориям) текстов в информационно-поисковых системах, классификация по тематике сообщений в новостной ленте, определение эмоционального состояния автора текстовых сообщений в социальных сетях и рекомендательных системах. Для решения задачи классификации по предметным областям (задачи категоризации) успешно применяется системное представление текстов, в первую очередь векторное представление и использование *N*-грамм [5]. Существует большое число алгоритмов для решения задач классификации, которые могут быть модифицированы при обработке текстовых данных, например модифицированные алгоритмы *k* ближайших соседей (*k-NN*), метод опорных векторов и т.п.

**Задачи идентификации.** К задачам идентификации текстов относятся идентификация авторства и стиля художественных текстов, определение признаков автоматического перевода на основе выявления особенностей написания текстов на родном языке, проблемы переводного плагиата и заимствования в научных публикациях, определение ключевых слов, терминов и фраз с целью автоматического аннотирования и реферирования. При решении задач идентификации могут быть использованы разнообразные модели текстовых структур, такие как взаимная информация, Марковская модель текста, модели, основанные на N-граммах, энтропийные характеристики вычисления символьного разнообразия [9] и др.

В работе предложена концепция формирования системоорганизующих характеристик текстовых данных как многомерных объектов и разработана методология системного анализа многомерных объектов на основе принципа скрытых параметров, которая продемонстрирована на примере анализа текстовых данных. На основе диалектического метода введено определение понятия «система» с учетом иерархической организации многомерных объектов, что позволило обосновать выбор скрытых параметров в качестве системоорганизующих характеристик текстовых данных, которые вычисляются как функции наблюдаемых параметров. Количество скрытых параметров и выбор вида соответствующих функций определяются критерием, лежащим в основе решаемой задачи. Предложенная методология формирования системоорганизующих характеристик текстовых данных может найти практическое применение в задачах развития и совершенствования методов кластеризации, классификации и идентификации текстовых данных.

## Список литературы

1. Большаков А.А., Каримов Р.Н. Методы обработки многомерных данных и временных рядов: учебное пособие для вузов. – М.: Горячая линия – Телеком, 2007. – 522 с.

2. Жаров С.Н. Бытие и пространство: трансцендентальная перспектива: монография Пространство как трансцендентальная предпосылка познания реальности / Рос. акад. наук, Ин-т философии, ИФРАН, 2014. – 108 с.

3. Зеленов Л.А. Введение в общую методологию. – Н. Новгород, 2002. –155 с.

4. Ломакин Д.В., Панкратова А.З., Суркова А.С. Золотая пропорция как инвариант структуры текста. // Вестник Нижегородского университета им. Н.И. Лобачевского. – 2011. – № 4. – С. 196–199.

5. Ломакина Л.С., Мордвинов А.В., Суркова А.С. Построение и исследование модели текста для его классификации по предметным категориям // Системы управления и информационные технологии. – 2011. – № 1(43). – С. 16–20.

6. Ломакина Л.С., Суркова А.С. Прикладные аспекты концептуального анализа и моделирования текстовых структур // Фундаментальные исследования. – 2015. – № 7 (часть 3). – С. 540–544.

7. Лосев А.Ф. Бытие. Имя. Космос. – М.: Мысль, 1993. – 958 с.

8. Субетто А.И. Творчество, жизнь, здоровье и гармония. – М.: Изд. фирма «Логос», 1992. – 204 с.

9. Суркова А.С. Анализ и моделирование текстовых данных в задачах обеспечения кибербезопасности // Системы управления и информационные технологии. – 2015. – № 3.1(61). – С. 178–182.

## References

1. Bolshakov A.A., Karimov R.N. Metody obrabotki mnogomernyh dannyh i vremennyh rjadov: Uchebnoe posobie dlja vuzov. M., Gorjachaja linija, Telekom, 2007. 522 p.

2. Zharov S.N. Bytie i prostranstvo: transcendentalnaja perspektiva // Monografija Prostranstvo kak transcendentalnaja predposylka poznanija realnosti / Ros. akad. nauk, In-t filosofii, IFRAN, 2014, 108 p.

3. Zelenov L.A. Vvedenie v obshhuju metodologiju. N. Novgorod, 2002, 155 p.

4. Lomakin D.V., Pankratova A.Z., Surkova A.S. Zolotaja proporcija kak invariant struktury teksta. // Zhurnal «Vestnik Nizhegorodskogo universiteta im. N.I. Lobachevskogo», 2011, no. 4, pp. 196–199.

5. Lomakina L.S., Mordvinov A.V., Surkova A.S. Postroenie i issledovanie modeli teksta dlja ego klassifikacii po predmetnym kategorijam. //Sistemy upravlenija i informacionnye tehnologii, 2011, no. 1(43), pp. 16–20.

6. Lomakina L.S., Surkova A.S. Prikladnye aspekty konceptualnogo analiza i modelirovanija tekstovyh struktur // Fundamentalnye issledovanija, 2015, no. 7 (chast 3), pp. 540–544.

7. Losev A.F. Bytie. Imja. Kosmos. M., Mysl, 1993, 958 p.

8. Subetto A. I. Tvorchestvo, zhizn, zdorove i garmonija. M., Izd.firma «Logos», 1992, 204 p.

9. Surkova A.S. Analiz i modelirovanie tekstovyh dannyh v zadachah obespechenija kiberbezopasnosti // Sistemy upravlenija i informacionnye tehnologii, no. 3.1(61), 2015, pp. 178–182.

**Рецензенты:**

Баландин Д.В., д.ф.-м.н., профессор, заведующий кафедрой численного и функционального анализа, Нижегородский государственный университет им. Н.И. Лобачевского, г. Нижний Новгород;

Федосенко Ю.С., д.т.н., профессор, заведующий кафедрой «Информатика, системы управления и телекоммуникации», Волжский государственный университет водного транспорта, г. Нижний Новгород.