DOI 10.12851/EESJ201501C06ART03

Nadezhda V. Margazova*;*
Director,
Transbaikalian Technology Transfer Center;

Yury V. Andreev*,*
ScD, associate professor;

Sergey Y. Andreev*,*
ScD, associate professor,
Institute of Atmospheric Optics,
Siberian Branch of the Russian Academy of Sciences

Работа поддержана грантом РФФИ (№ 14-07-98005)

Fuzzy Modelling Methods for Semantic Information Analysis Using

Key words: *semantic, interdisciplinary researches, knowledge fuzzy hypergraph, stability factor, significance factor*

Annotation: *In article are described the basic approaches to modeling integrating and generation of interdisciplinary knowledges processes. Semantic data structuring with fuzzy hypergraph methods are described. Introduced the concept of stability and significance factor of semantic relations. The effect on knowledge using character and information space stratification of these factors is described.*

Стало почти аксиомой утверждение, что наиболее значимые и принципиально новые результаты дают исследования, в основе которых лежит принцип интеграции знаний, полученных в отдельных научных дисциплинах. В этой связи в научной среде активно обсуждаются стратегии исследований, базирующихся на принципах полидисциплинарности, междисциплинарности и трансдисциплинарности (5), (2).

В стадии зарождения знания на стыке наук принципиальную значимость приобретает сам факт возникновения связи между информацией, полученной в различных предметных областях. Основное назначение этой связи – открыть возможности применения семантик и методик, используемых в различных дисциплинах, в процессе генерации новых знаний.

Для моделирования процесса согласования информационных структур, принадлежащих разным предметным областям, необходимо выяснить, то каким образом понимание человеком информации, созданной в информационной среде, относящейся к предметной области, изначально не входившей в сферу его интересов, становится отправной точкой для генерации междисциплинарного и трансдисциплинарного знания. Определить почему это новое знание – совсем не то же самое, которое было в изучаемой предметной области, почему и каким образом оно вплетается в знакомые информационные структуры, становясь «точкой кристаллизации» нового междисциплинарного направления исследований?

И.В. Лысак, рассматривая междисциплинарный и трансдисциплинарный подходы к исследованиям в гуманитарной сфере, обращает внимание на проблему, несовпадения специализированных языков, присущих не только отдельным научным дисциплинам, но даже различным сферам культуры. Она отмечает: «Если гуманитарии еще могут «договориться о терминах», то адекватный диалог между представителями естественных и гуманитарных наук «языковая несовместимость» делает практически невозможным», но при этом автор приходит к выводу, что «интеграция естественнонаучного и гуманитарного знания на основе прочного методологического фундамента может открыть новые перспективы ...».

Различия в сложившейся терминологии, условных обозначениях, используемых методах обработки и хранения знания, исторически сложившиеся контакты формируют своего рода «диалекты» научных школ. И чем более узкой является их специализация, тем более специфичными становятся их знаковые системы. Нередки даже случаи, когда одни и те же обозначения или термины в разных дисциплинах могут иметь совершенно различное смысловое наполнение.

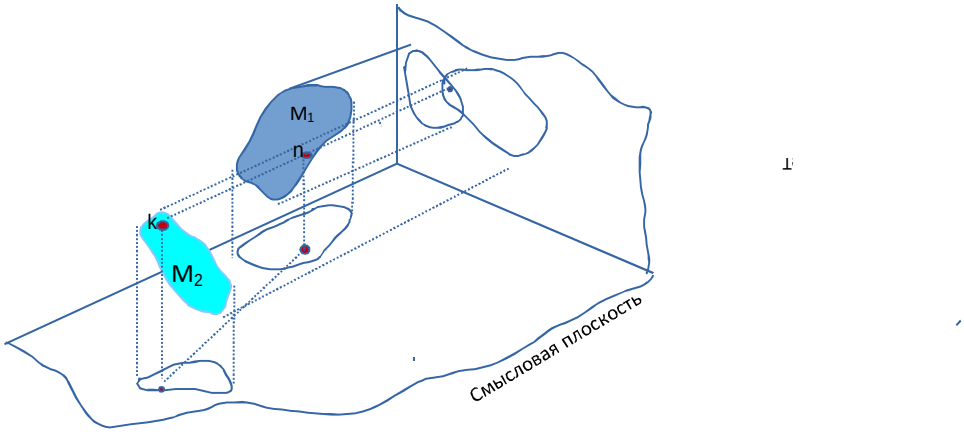

Рис. 1 Визуализация отображения понятий

Пусть мы имеем две неких научных дисциплины, каждая из которых характеризуется своим множеством понятий M1 и M2. Расположим эти множества в n-мерном пространстве, в котором для простоты визуализации выделим две ортогональные плоскости, одну из них назовем «терминологической», а другую - «смысловой» (рис. 1).

Предположим, что множество M1 включает в себя понятия, относящиеся к нейробиологии, а множество M2 — к кибернетике. И в нейробиологии, и в кибернетике активно используется понятие «нейронная сеть». Таким образом это понятие будет принадлежать и множеству M1 и множеству M2. При этом в нейробиологии нейронная сеть это совокупность нейронов - структурно-функциональных единиц нервной системы, выполняющих специфические физиологические функции. В кибернетике же под нейронной сетью подразумевается математическая модель, а также её программная или аппаратная реализации, построенная по принципу организации и функционирования сетей нервных клеток живого организма.

Обозначим индексом n точку, относящуюся к множеству M1, соответствующую понятию «нейронная сеть», а индексом k аналогичную точку

множества M2. Проекции этих точек на терминологическую плоскость будут совпадать, проекции на смысловую плоскость будут различаться.

С точки зрения генерации междисциплинарного знания этот пример не показателен, однако важно акцентировать внимание на том, что восприятие словосочетания «нейронная сеть» изначально вызовет у биолога и программиста ассоциации, связанные с разными объектами.

Как известно, память человека ассоциативна. Мы не помним во всех подробностях многих увиденных когда-либо объектов и событий. В памяти остаются лишь некие смысловые указатели на сущности и явления реальности, которые ассоциируются у наблюдателя с тем или иным событием, и по мере необходимости позволяют памяти каждый раз заново воссоздать образ объекта или картину события. При этом задействуется только та часть связей, которая необходима в данный момент, хотя дополнительные семантические структуры могут быть в любой момент вызваны из памяти по соответствующим им цепочкам ассоциативных связей. Здесь под семантическими структурами мы будем понимать смысловое наполнение, как отдельных слов, так и словосочетаний.

В работе (6) отмечалось, «...что, хотя способность сознания человека напрямую оперировать смыслами является врожденной и проявляется у него до освоения языков и письменности, однако для обеспечения устойчивых коммуникативных процессов в группах применяются различные системы знаков. Когнитивная семантика определяет, что существуют проекции смыслов в доступные для участников коммуникативного процесса форматы. Примерами таких проекций являются язык, письменность, условные обозначения, жесты, формулы, термины и т.д.» Экстралингвистическая природа смысла (9) позволяет, опираясь на семантику, абстрагироваться от частных знаковых представлений.

Процесс преобразования семантических конструкций в сообщения, зафиксированные в формате, присущем той или иной знаковой системе, может быть представлен как проецирование абстрактной многомерной семантической структуры на символьное множество. Процесс проецирования можно также назвать форматированием, т.к. для каждой знаковой системы характерен свой уникальный формат. Одновременно могут существовать проекции смыслов в различные символьные множества, при этом в каждом формате существуют только те элементы множеств, семантические прообразы которых принадлежат областям интересов сообщества, создавшего данную знаковую систему, а, следовательно, можно утверждать, что ни одна знаковая система не содержит проекций всех возможных смыслов.

Очевидно, что каждая семантическая единица в свою очередь может быть раскрыта через ассоциативные, смысловые или структурные связи с другими смыслами. В работах Н.Н. Заличева (14), (15), А.Н. Печникова (7), А.П. Леонтьева (4) для выполнения количественного семантического анализа информации используется понятие элементарной семантической единицы (ЭСЕ).

Н.Н. Заличев в качестве элементарной семантической единицы когнитивной структуры информации принимает законченную мысль в виде утверждения. А.П. Леонтьев под ЭСЕ информации понимает неопределяемые и первичные понятия, а

также понятия, привнесенные из других областей знания, факты и события, представленные в лингвистико-знаковой форме без семантической и прагматической избыточности.

Оба эти определения не противоречат друг другу в том смысле, что ЭСЕ является количественной семантической мерой обрабатываемой информации. В своей диссертации, посвященной разработке энтропийной теории семантического анализа научной информации, Н.Н. Заличев отмечает, что информация, получаемая и анализируемая в процессе формирования новых научных знаний, характеризуется определенной нечеткостью, расплывчатостью, приблизительностью оценок. Он пишет: «Характерно, что формируемая в процессе научных исследований когнитивная структура информации (нечеткое знание), все более адекватно отражающая изучаемый фрагмент действительности, является информационным фантомом (ИФ) этого фрагмента действительности, не связанным напрямую со значимостью информации». Под информационным фантомом автор понимает «информационное отражение фрагмента анализируемой реальности, описываемый обрабатываемой информацией и состоящий из «мозаики» элементарных семантических единиц».

Все вышесказанное дает нам основание использовать для семантического анализа научной информации методы теории нечетких множеств. Такой подход, как мы считаем, имеет право на существование, исходя хотя бы из того, что ряд таких понятий, как элементарная семантическая единица, информационный фантом, практически соответствуют понятиям лингвистическая переменная, лингвистическое множество, введенным Л.А. Заде — основателем теории нечётких множеств и нечёткой логики (10, 11, 12, 13), а также другими исследователями (8)(1).

Каждое слово, обозначение или понятие мы можем объяснить, интерпретировать, дать определение, т. е. выразить его значение с помощью других слов или знаков, каждому из которых соответствует некая смысловая единица. Это доказывает, что ни одна семантическая единица не существует сама по себе, а объединяет целый комплекс взаимосвязанных смыслов, с помощью которых может быть раскрыта ее сущность. Следовательно, семантическим прообразом слова или обозначения в конечном счете является не смысловая «точка», а некоторое множество взаимосвязанных смыслов. А сам процесс раскрытия одного смысла через некое множество других взаимосвязанных смыслов является рекурсивным.

Применительно к семантическому анализу, базирующемуся на методах теории нечетких множеств, целесообразно оперировать понятием семантическое множество Xs, представляющее по своей сути информационный фантом Н.Н. Заличева, элементами которого являются семантические единицы xsi, которым в символьном множестве языка соответствуют лингвистические переменные Л.А. Заде.

Степень принадлежности каждого семантического значения к семантическому множеству характеризуется функцией принадлежности $\mu_A(X)$, которая ставит в соответствие каждой ассоциативной связи некоторое число из интервала [0, 1].

$$\mu_A(Y): \rightarrow [0,1], \forall_Y \in X \qquad (1)$$

Использование устоявшихся знаковых систем является основой долговременного сохранения ассоциативных связей, используемых в неизменном виде в контексте решения различных задач, что служит важнейшей предпосылкой

формирования научно-технического и культурного базиса сообщества. Поэтому в структуре любого семантического множества, возникающего при мысленном моделировании реальности, формируются устойчивые ассоциативные связи, которые воспринимаются как достоверные, не подлежащие сомнению или уточнению. Одновременно с этим между семантическими единицами могут существовать и менее устойчивые связи, требующие уточнения и подкрепления путем получения дополнительной информации, т.е. более полного раскрытия через другие смыслы.

Информационная структура предметной области в общем случае может быть представлена в виде нечеткого гиперграфа G.

$$G=(X,E) \quad (2)$$

где X={xi}, $i \in I$ – множество вершин графа;

$E=\{\mu_E(x_j,x_i) \mid (x_i,x_j) \in X\}$ – нечеткое множество связей между семантическими единицами $x_i,x_j \in X$

μE — функция принадлежности для ребра (x_i,x_j)

Объединение n нечетких гиперграфов семантических множеств, имеющих проекции в знаковое пространство той или иной предметной области, может в целом характеризовать структуру знаний об этой предметной области.

$$G_1 \cup \ldots \cup G_n = (X_1 \cup \ldots \cup X_n, E_1 \cup \ldots \cup E_n) \quad (3)$$

Для описания возможности существования в нечетком гиперграфе связей с различной степенью устойчивости, от таких, которые пользователь воспринимает как достоверные, до связей, воспринимаемых как гипотезы, введем понятие коэффициента устойчивости связи между семантическими единицами $S(x_i,x_j)$, который может быть определен на основе статистического анализа выбранных пользователем траекторий обхода графа, включающих его ребро (x_i,x_j) .

Тот факт, что у любого предмета исследования потенциально существует значительно больше характеристик, чем необходимо и достаточно для решения той или иной задачи, означает, что в случае, когда модель объекта учитывает n связей $E_{1\ldots n}$, в открытой системе существует характеристика E_{n+1} . При этом очевидно, что набор учитываемых характеристик зависит от понимания исследователем того, какие из них являются наиболее значимыми в контексте исследуемой проблемы, и, как следствие, от используемых методов и подходов.

Определим ассоциативные связи, значимые в контексте некой конкретной задачи, как имеющие вес равный 1, а остальные, незначимые, но потенциально существующие связи, как имеющие вес равный 0.

$$\mu_E(X) = \begin{Bmatrix} 0, \lim K(x_i,x_j) \to 0 \\ 1, \lim K(x_i,x_j) \to 1 \end{Bmatrix} \quad (4)$$

где К - критерий значимости.

В этом случае структура информации может быть приведена к виду, описываемому четким графом, а следовательно можно говорить, что веса ребер графа стремятся либо к 0, либо к 1 в зависимости от значимости соответствующих связей в

контексте решаемой задачи. Соответственно, $\mu_E(X) = 1, \forall_y \in E$, где E – область связности графа информационной структуры.

Поскольку нечеткий граф является обобщением обычного графа (3), можно предположить, что любая иерархически структурированная информация также рассматривается как подграф нечеткого гиперграфа знаний, определяемый в соответствии с семантическим пространством контекста конкретной задачи.

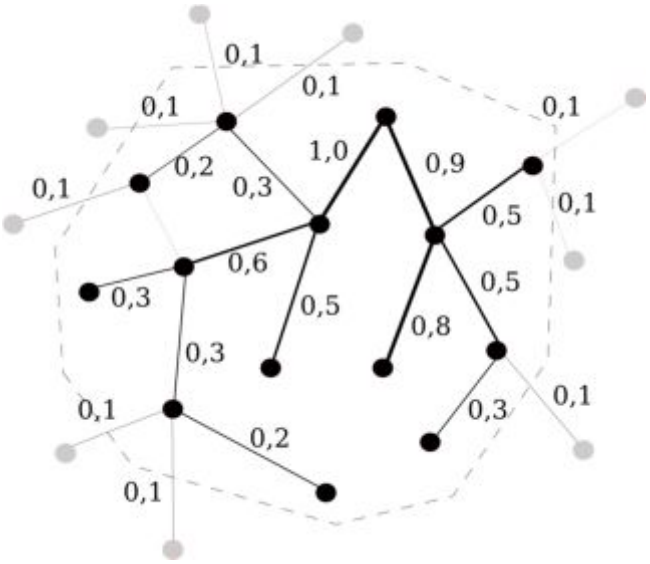

Рис. 2. Критерий значимости

В том случае, когда различные ассоциативные связи внутри семантического множества имеют различную значимость, можно утверждать, что коэффициент значимости $K(x_i, x_j)$ принимает любое значение в диапазоне от 0 до 1. Если $\lim K(x_i, x_j) \to 0$, то соответствующая связь (x_i, x_j) не является значимой. Соответственно, элементы семантического множества образуют область связности нечеткого гиперграфа. Для реальных задач имеет смысл задать нечеткий критерий $K(x_i, x_j)_{\min}$, который определяет условие значимости связи между семантическими единицами.

Рис. 2 показывает пример локализации множества семантических единиц при формировании информационной структуры путем выявления нескольких наиболее значимых признаков. Установка порогового критерия значимости, например, на уровень $K(x_i, x_j)_{\min} = 0{,}15$, позволяет «отсечь» все «лишние» связи, значимость которых в данном случае невелика, осуществляя таким образом приведение нечеткой граф-схемы к каноническому четкому виду. Варьирование этого коэффициента, в свою очередь, позволяет либо расширить диапазон исследуемых свойств, либо сузить его, выделяя наиболее важные.

Возможность получения различных нечетких информационных структур, базирующихся на едином семантическом пространстве, в общем случае подразумевает ненулевые значения весов ребер на объединении нечетких гиперграфов частных семантических множеств.

Очевидно, что коэффициент устойчивости связей между семантическими единицами не зависит от их значимости, и наоборот. Устойчивые связи могут не иметь

никакого значения в контексте решаемой задачи, но при этом наиболее значимые связи могут оказаться неустойчивыми. Таким образом, ребра нечеткого графа характеризуются вектором двух независимых компонент – фактора значимости и фактора устойчивости.

Значимость — это контекстозависимая характеристика. Как следствие, использование исследователем тех или иных методов и подходов к решению задачи, а также глубина знаний предметной области, вносят субъективную составляющую в процесс семантического моделирования информационных структур. Поэтому можно говорить о формировании семантической системы, в которую помимо объективно существующих данных включается исследователь, как распознающая и интерпретирующая кибернетическая составляющая этой системы, что позволяет персонифицировать особенности структурирования информационного пространства в зависимости от уровня знаний и областей интересов исследователя.

Рассмотрим распределение величин значимости и устойчивости связей внутри семантической системы (рис. 3). Объем накопленных знаний в той или иной области, способствует формированию большого количества связей, обладающих высокой устойчивостью. Поэтому семантические единицы, объединенные устойчивыми связями, образуют «слой» знаний, который можно назвать базовым. Поскольку наиболее сильные связи наименее изменчивы, данная область характеризуется наименьшей мобильностью. Повышение значимости связей из этой области в процессе решения тех или иных задач приводит к формированию наиболее стабильных и предсказуемых семантических структур, которые можно использовать для получения экспертных оценок. Количество связей с высокой устойчивостью непрерывно увеличивается в результате обработки поступающих сведений. Область устойчивых связей является базисом для формирования поисковой активности и определяет восприимчивость к входящим потокам информации.

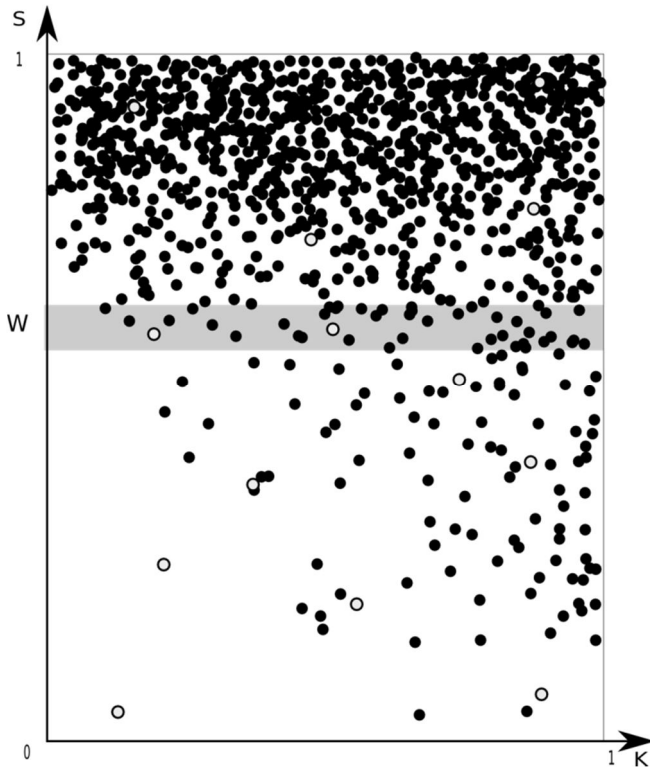

Рис. 3. Распределение значимости и устойчивости связей между семантическими единицами.

Активное развитие той или иной тематики, существование массы гипотез, нерешенных вопросов способствуют активному формированию связей с низкой устойчивостью и высокой значимостью. Подмножество семантических единиц, объединенных такими связями, можно охарактеризовать как область поисковых интересов. Семантические структуры этого слоя наименее стабильны, значения функции принадлежности наиболее сильно варьируются, происходят процессы генерации новых связей, разворачивания и сворачивания цепей ребер нечеткого гиперграфа. Эволюция семантических систем приводит к постепенному повышению устойчивости ассоциативных связей и, как следствие, к миграции сформировавшихся структур в базовый слой. Таким образом, эти процессы за счет изменения устойчивости связей влияют на структуру базового уровня знаний. В большинстве случаев нечеткий гиперграф области поисковых интересов имеет несколько семантических ядер, вокруг каждого из которых происходят наиболее активные изменения.

Поскольку информационная система в общем случае является открытой, значительное влияние на структурирование информации оказывают внешние факторы. Часть информации, с которой взаимодействует семантическая система, поступает извне и не может контролироваться пользователем. На рис. 3 связи, формирующиеся вследствие интерпретации входящих сигналов, обозначены как светлые точки, распределенные случайным образом. Прогнозировать, семантическое наполнение таких сигналов невозможно, однако характер их воздействия на структуру и динамические характеристики семантических систем напрямую зависят от того, какую значимость и устойчивость получают связи, сформированные в результате взаимодействия входящих сигналов с существующим информационным пространством.

«Попадание» такого сигнала в область высокой устойчивости не вызывает интереса, поскольку полученная информация воспринимается пользователем как давно известная. Если значимость и устойчивость одновременно не высоки, то полученные сведения не вызывают ни доверия, ни интереса, а следовательно интерпретируются как информационный шум. Наибольшее влияние на семантическую систему оказывает информация, относящаяся к области высокой значимости, но небольшой устойчивости, особенно попадающая в область поисковых интересов. Новости, факты, публикации, имеющие значимое для пользователя семантическое наполнение, в условиях неустойчивости связей могут как существенно повлиять на выбор методов и подходов к решению задач, так и способствовать подтверждению или опровержению гипотез, привлечь особое внимание к тем или иным аспектам исследования и даже вовлечь в область интересов ранее не рассматривавшиеся смыслы.

Таким образом, используя методы нечеткого моделирования для изучения информационных структур, мы можем определить их качественные и количественные характеристики, а также стратифицировать семантическое пространство, выделив три слоя.

1. Базовый слой знаний, влияющий на все аспекты интерпретации и использования информации, описываемый нечетким гиперграфом устойчивых семантических связей.

2. Слой поисковых интересов, характеризующийся высокой изменчивостью величин значимости и динамикой, связанной с непрерывным формированием новых связей и цепочек ребер гиперграфа, изменением их значимости и устойчивости.

3. Слой семантических структур, формирующихся при взаимодействии с входящей извне информацией, который можно охарактеризовать как новостной контекст.

Предположив существование двух независимых компонент — фактора значимости и фактора устойчивости связей между семантическими единицами, мы можем исследовать комплексную динамику развития информационных структур, как в долгосрочной перспективе, так и в контексте решаемых исследователем задач.

References:

1. *Bertram T, Svaricek F. and other Fuzze Control. Zusammenstellung und Bschreibung Wichtiger Begriffe. Autotisierungstechnik. 1994, vol.42, №7; 322-326.*
2. *Knyaseva EN. Transdistiplinary research strategies: Messenger of Tomsk state pedagogical university. 2011, №10; 193-201.*
3. *Kofman A. Introduction into the Theory of Fuzzy Sets. Moscow, 1982; 432.*
4. *Leontiev AP, Gochman OG. Problem management education process (mathematical models). Riga, 1984.*
5. *Lysak IV. Interdistiplinarity and transdistiplinarity as approach to human research: Historical, philosophical, political and juridical sciences, cultural studies and art history. Theoretical and practical queries. 2014, № 6, Part II; 134–137.*
6. *Margazova NV. Overview of modern concepts and approaches to communicative process in heterogeneous information environment modeling: Messenger of Chita State University. 2012, № 6 (85); 76-82.*
7. *Pechnikov AN. Theoretical basis of automated education systems psycho-pedagogical engineering. Petrodvorets, 1995; 322.*
8. *Piegat A. Fuzzy modeling and control. -2-nd ed. Moscow. Binom. Knowledge laboratory, 2013; 798. (Adaptive and intelligent systems)*
9. *Sljusareva HA. Semantics as a extra-linguistic phenomenon: How to prepare an interesting lesson in a foreign language. Moscow, 1963; 185-199.*
10. *Zadeh LA. The Linguistic variable and its application to approximate reasoning. Moscow, 1976; 165.*
11. *Zadeh LA. From Computing with Numbers to Computing with Words – From Manipulation of Measurements to Manipulation of Perceptions: IEEE Transactions on Circuits and Systems. 1999, Vol.45; 105-119.*
12. *Zadeh LA. Toward a Theory of Fuzzy Information Granulation and Its Centrality in Human Reasoning and Fuzzy Logic. Fuzzy Sets and Systems. 1997, Vol. 90; 111-127.*
13. *Zadeh LA. Toward a Generalized Theory of Uncertainty: Information Sciences – Informatics and Computer Science. 2005, Vol. 172; 1-40.*
14. *Zalichev NN. Entropy of information and the essence of life, Moscow, 1995; 192.*
15. *Zalichev NN. Development and practical application of the semantic analysis methodology in automated science information processing systems. Thesis for the degree of Doctor of Technical Sciences.*