УДК: 004.021
DOI: 10.30987/2658-6436-2020-2-24-32

П.Ю. Шалимов

МОДЕЛИ КОЛИЧЕСТВЕННОЙ ШКАЛЫ ОЦЕНКИ СЕМАНТИЧЕСКОЙ ИНФОРМАЦИИ

Проанализирована проблема математического моделирования семантической информации. Разработана модель информационной среды как носителя семантической информации. Представлена процедура оценки семантической информации в имитации функционирования в режиме накопления и оценки. Разработан программный комплекс формирования информационной среды и проведения имитационных исследований. Рассмотрена постановка вычислительных экспериментов при сравнительном анализе количества семантической информации текстового ресурса. Показаны наиболее актуальные приложения модели.

Ключевые слова: *семантическая информация, количественная шкала, математическая модель, имитационное моделирование, семантическая емкость.*

P.Yu. Shalimov

MODELS OF A QUANTITATIVE SCALE FOR EVALUATING SEMANTIC INFORMATION

The problem of mathematical modeling of semantic information is analyzed. A model of the information environment as a carrier of semantic information is developed. The procedure for evaluating semantic information in the simulation of functioning in the accumulation and evaluation mode is presented. A software package for creating an information environment and conducting simulation studies has been developed. We consider the setting of computational experiments in the comparative analysis of the amount of semantic information of a text resource. The most current applications of the model are shown.

Keywords: *semantic information, quantitat ivescale, mathematical model, simulation modeling, semantic capacity.*

Введение

Современные подходы к исследованию понятия семантическая информация формируют несколько позиций, в определенном смысле противоречивых, по отношению к принципиальному вопросу: что считать основным элементарным носителем семантической информации. Первая позиция полагает основным элементарным носителем смысла при человеческом общении предложение [1], при низкой значимости словосочетаний и слов (понятий). Ряд источников [2, 3] фиксирует отрицание такого положения.

Понятие информация описывается с лингвистической и философской точки зрения, что не позволяет перейти к эмпирическому исследованию семантической информации. При том, что существуют и достаточно широко используются модели, использующие качественные оценки семантической информации [4, 5]. Синергетический и физико-информационные подходы к понятию информатика позиционируются в [6 - 8].

Постановка задачи

Представляемая модель предполагает возможность вычисления количественной оценки семантической информации, что будет актуальным для атрибутирования текстовых материалов количественной оценкой смыслового содержания. Появляется возможность поставить в соответствие тексту количественную оценку смыслового содержания и провести семантическое оценивание энциклопедий, словарей, тезаурусов, учебников.

Прототип модели количественной оценки семантической информации – любой

субъект, существующий во множестве экземпляров. Наряду с субъектом, в роли которого выступает человек, к носителям семантической информации относятся текстовые материалы. Для обобщения носителей семантической информации используется понятие информационная среда [9].

В основе разработанной модели лежит постулат субъективизма информации: семантическая информация генерируется человеком и материально проявляется при синтаксическом формулировании в виде последовательности предложений. Постулат включает положение генерации семантической информации и положение передачи информации.

Семантическая информационная среда (СИН) – субъект, обладающий множеством элементов имитации памяти с целью сопоставления с входным несемантическим предложением, выраженным в синтаксической форме, и вычисляющий оценку количества семантической информации. Любое синтаксически выраженное сообщение имеет количественную оценку значения семантического отображения для конкретного экземпляра СИН. Сообщение в синтаксической форме проходит через схему, обрабатывается в массиве ЭИП (элемент имитации памяти), формирует значение количества семантической информации. Ценность - количественный атрибут ЭИП.

ЭИП в процессе активации выполняет два основных действия: формирование нового элементарного сообщения, изменение значения важности. ЭИП характеризуется координатами в топологическом пространстве схемы, значением ценности, значением такта в локальной процедуре. Новое (выходное значение ЭИП) сообщение формируется при линейном умножении ценности v на значение входного сообщения. Ценность меняет значение в зависимости от номера такта локальной активации t.

К основным свойствам семантической информации, которые следует отразить в модели, относятся субъектность и синтаксическая непредставимость, материальность. Свойство субъектности означает, что семантическая информация формируется, хранится и формулируется у конкретного субъекта информационного обмена. С позиций семантической информации единственным субъектом является человек (прототип для модели). Семантическая информация не представима на физическом уровне.

Словарь несемантической (асемантической) информационной среды включает:

- Конечное множество лексем $L=\{l_i|i\in I\}$, где l_i– лексема, I – множество индексов.
- Размер асемантической среды $N=|L|$.

Приведенная асемантическая среда множество лексем $X=\{X_i|X_i=i\}$. Несемантическое предложение n-кортеж выборка лексем (l) множества L размера $\{k|k<m\}$

$$B[i]=\{1|t\in|a;0|t\notin a\}.$$

Описание модели

Обработка исходного текста выполняется в простом линейном порядке. Модели основаны на использовании элементов имитации памяти (ЭИП), из которых составлены схемы семантической информационной среды. Расчетные схемы интерфейсов соединения ЭИП представлены на рис. 1.

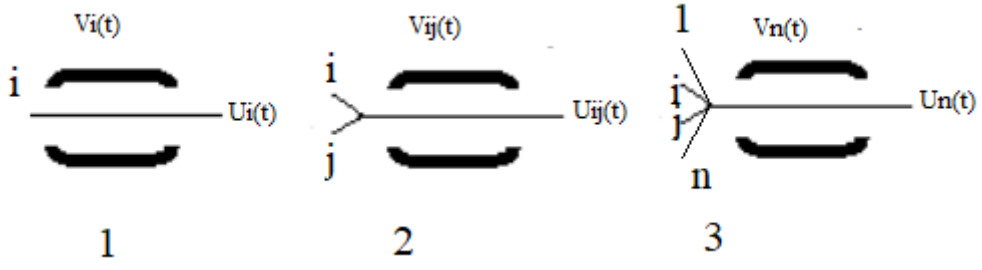


Рис. 1. Расчетные схемы интерфейсов ЭИП

К параметрам ЭИП относятся:

- арность - Ar (определяется числом одновременно обрабатываемых лексем предложения исходного текста);
- ценность - V (количественный атрибут памяти);
- такт локального функционирования - t (порядковый номер активации).

На рис.1. показаны ЭИП с одним входом со значениями арности Ar=1, «бипольный ЭИП» (Ar = 2), «n-мерный» ЭИП ($Ar = n$).

Сообщение в синтаксической форме проходит через схему, обрабатывается в массиве ЭИП (вычисляется и фиксируется новое значение ценности ЭИП), формирует значение количества семантической информации, фиксируется значение порядкового типа локального такта. Ценность V- количественный атрибут ЭИП. $V \in [0..1]$.

ЭИП в процессе активации выполняет два основных действия:

- формирование нового элементарного сообщения:
- изменения значения ценности.

ЭИП характеризуется координатами в схеме, значением ценности, значением такта в локальной процедуре. Такт в локальной процедуре – элементарная операция функционирования модели, при которой вычисляются и регистрируются изменения конкретного ЭИП.

Ценность $V_i(t)$ меняет значение в зависимости от номера такта локальной активации t

$$V_i(t)=F(V_i(t\text{-}1),t).$$

Без учета значения ценности на предшествующем такте используется зависимость

$$V_i(t+1)=F_s(t),$$

где F, F_s функции преобразования элементов входного сообщения в количественную характеристику выходного значения ЭИП; t – номер локального такта функционирования ЭИП; i – координата в схеме.

В качестве функции преобразования $F(t)$ используется сигмоидальная двухпараметрическая зависимость

$$F(t) = \frac{1}{1 + \exp(-a(t - d)}\,,$$

где a – параметр «крутизны» графика функции; d – параметр начального смещения; t – номер локального такта функционирования ЭИП. Параметры a и d - индивидуальные атрибуты конкретного экземпляра информационной среды. Применительно к субъекту информационного обмена параметр a именуется как атрибут интереса, параметр d - атрибут начальной информации. Выходное значение U формируется при линейном умножении ценности V на значение входного сообщения B.

Модель оценки семантической информации реализуется в схеме семантической информационной среды (рис. 2). Входное предложение 1 показанное на рис.2, как вектор B получается в результате предобработки исходного естественно-языкового предложения оцениваемого текста. Предобработка предполагает приведение исходного предложения к размерности схемы информационной среды. Размерность схемы информационной среды N соответствует максимальному количеству лексем в текстах, обрабатываемых конкретным экземпляром информационной среды.

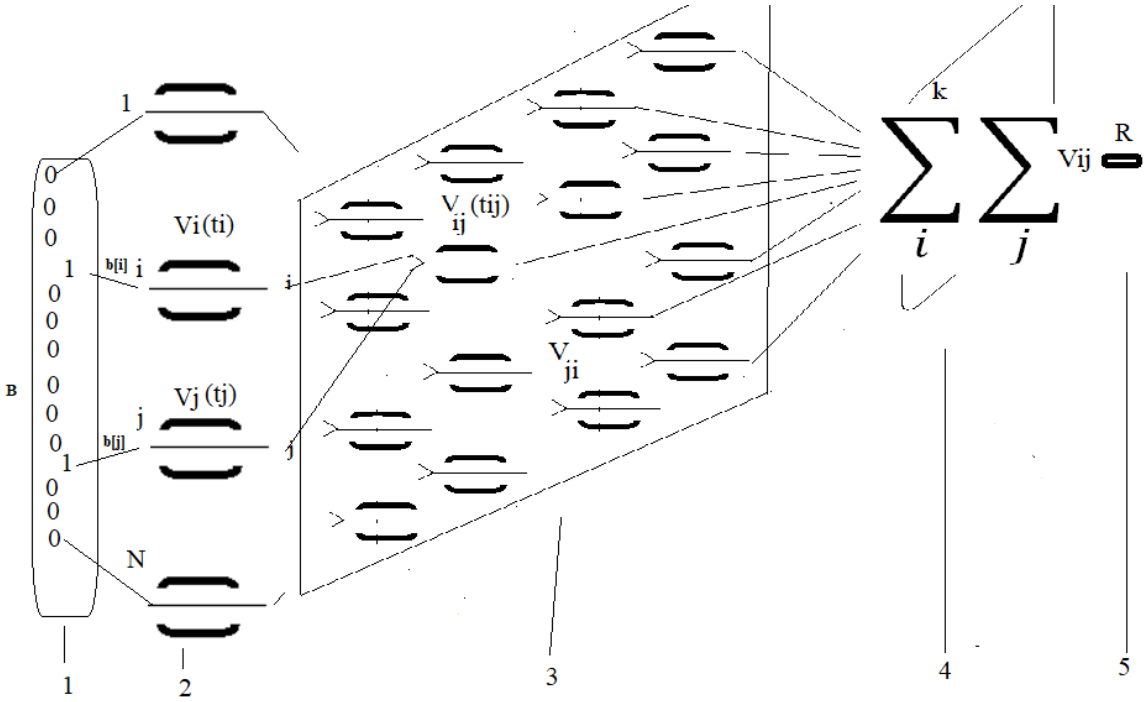
Предложение, обрабатываемое в схеме, приводится к вектору размерностью N содержащим числа 0 и 1.

Входное приведенное предложение B обрабатывается в слое ЭИП одного входа с вычислением выходных сообщений $U_i(t_i)$

$$U_i(t_i)=b[i]V_i(t_i),$$

где t_i – значение локального такта координаты i, $b[i]$ – приведенное значение входного

вектора по координате *i*, *Vi* – значение важности ЭИП одного входа. Обработка предложения происходит в такте глобальной активации при моделировании среды. Такт глобальной активации содержит локальные активации ЭИП.



Входное предложение | Слой ЭИП одного входа | Слой бипольных ЭИП | Сумматор | Выходное значение количества семантической информации от входного предложения

Рис. 2. Схема семантической информационной среды

Вычисляются новые значений ценности $V_i(t_i+1)$

$$V_i(t_i+1)=F(t_i),$$

где $F(t_i)$ - функция преобразования.

Выполняются инкрементные операции с значениями локальных тактов координаты *i*

$$t_i=t_i+1.$$

Выходные сообщения $U_i(t_i)$ становятся входными сообщениями для слоя бипольных ЭИП. При обработке слоя бипольных ЭИП выполняется процедура комплектования пар входных компонент в простом линейном порядке. Общее количество комплектов входных компонент Q, определяется как число сочетаний из n по 2 без повторений

$$Q = C_n^2,$$

где n – количество ненулевых компонент выходного предложения слоя одного входа. Компоненты выходного сообщения U_{ij} бипольного ЭИП рассчитываются по формуле

$$U_{ij}(t)=U_iU_jV_{i,j}(t),$$

где *i*-номер первой лексемы пары; *j*- номер второй лексемы пары; U_i, U_j – компоненты выходных сообщений ЭИП первого слоя; $V_{i,j}$ - значение ценности бипольного элемента.

Значение ценности, получаемое при выполнении такта обработки сообщения, устанавливается отношением

$$V_{ij}(t+1)=F(t),$$

где $F(t)$ – функция преобразования сообщения в слое бипольных элементов.

Вычисляется оценка по результатам этапа функционирования схемы

$$R = \sum_{i=1}^{N} \sum_{j=1}^{N} U_{i,j} .$$

Схема вычислительной информационной среды используется для реализации процедур, выполняющих режимы функционирования:

- только оценка – определяется семантическая информация входного сообщения относительно конкретного экземпляра среды;
- только накопление – определяются операции изменения значений ценности для ЭИП конкретного экземпляра среды;
- комбинированный режим с оценкой и накоплением информации.

Укрупненная блок-схема алгоритма системы накопления и оценивания семантической информации описывается в последовательности этапов:

- ввод асемантических предложений как массива лексем;
- анализ массива лексем и определение размерности схемы;
- формирование схемы;
- (с параметрами *N*, одномерного массива слоя ЭИП одного входа размерностью *N*, массива ЭИП бипольных элементов размерностью *N*x*N*, начальные условия моделирования и параметры ЭИП);
- приведение текущего асемантического предложения к виду схемы информационной среды. Получение вектора *B* по формуле, приведенной выше;
- обработка вектора *B* в слое ЭИП одного входа. Получение выходного вектора *U*;
- определяется количество и координаты комплектов бипольных элементов, получающих активацию на следующем такте.

Применение модели

Постановка вычислительных экспериментов. В имитационном моделировании с вычислительной информационной средой реализуются апостериорные и априорные численные исследования. Практическое применений априорных и апостериорных исследований состоит в реализации сравнительных вычислительных экспериментов. Сравнительные эксперименты используют одинаковые значения параметров модели.

Апостериорный численный эксперимент предполагает использование модели семантической среды с накопленной информацией. Пример практического применения апостериорных исследований – точное оценивание результатов закрытого тестирования.

Точное оценивание результатов тестирования возвращает количественную оценку успешно завершенных тестов. До начала процесса оценивания выполняется процедура накопления семантической информации в вычислительной информационной среде. Информация для накопления представляется в наборе асемантических предложений, полученных на основе литературы для изучения.

Априорные исследования основаны на использовании комбинированного режима функционировании с оценкой и накоплением информации. Практическое применение – сравнительная оценка текстовых ресурсов по критерию количества семантической информации. В исследовании определяется значение семантической информации для каждого ресурса, с формированием отдельного экземпляра информационной среды для каждого текста. Модели для всех экземпляров сформированы при одном значении атрибутов интереса и начальной информации (*a*,*b*).

Семантическая информация и ее количественные атрибуты рассматриваются безотносительно субъекта, принимающего участие в информационных процессах.

Сравниваемые значения количества информации относятся к атрибутам текстовых ресурсов.

Инструментарий моделирования

Проведение вычислительных экспериментов с моделью выполняется посредством специально разработанного программного комплекса ScuS. Функционал моделирования с использованием программного комплекса разделен между программами: ScuS_t – сканирование исходных данных, формирование схемы информационной среды, подготовку электронного образа данных; ScuS_m – формирование экземпляра информационной среды на основе образа данных.

Электронный образ данных формируется на основе набора асемантических предложений, находящихся в порядке моделируемого ресурса. Порядок определяется правилом чтения лексем и предложений при сканировании ресурса. Используется правило «простой линейный порядок»: чтение лексем «слева направо» и предложений «вниз по странице».

Электронный образ данных - набор предложений, приведенных к схеме информационной среды и находящихся в порядке исходных данных. Исходные данные – текстовый ресурс, записанный как набор предложений в асемантической форме, подлежащий количественной семантической оценке. В качестве примеров исходных данных для семантического оценивания: словари, энциклопедии, учебники, тексты интернет ресурсов.

Формирование экземпляра информационной среды происходит при обработке электронного образа данных программой ScuS_m. Работа программы эквивалентна имитационному моделированию процесса чтения субъектом текстового источника и получения при этом семантической информации.

Формирования экземпляра информационной среды состоит в расчете значений ЭИП слоя бипольных элементов и значений ЭИП элементов одного входа. Процесс формирования экземпляра информационной среды соответствует чтению текста человеком.

Модель, представленная в виде схемы информационной среды, относится к простейшим (примитивным) и соответственно не учитывает множество явлений, процессов которые известны для прототипа: явление забывания; не мгновенное накопление информации; появление новых «интересов» в процессе накопления информации.

Постановка эксперимента и результаты числительного моделирования

Численный эксперимент для демонстрации возможностей модели проводился в режиме сравнительных исследований. Основные группы исследований: демонстрация семантики; семантическая емкость синтаксической информации.

Цель группы экспериментов демонстрация семантики заключается в иллюстрации возможностей модели обеспечивать основные качества семантической информации. Основная особенность семантической информации, которую желательно учесть в модели – зависимость количества новой информации, от объема имеющейся. Эта особенность отмечена в работах [10-13].

Подготовка экспериментов демонстрации семантики включает формирование наборов данных с уникальными предложениями и набора данных с повторяющимися предложениями. Основной набор данных с уникальными предложениями на основе словаря основных биологических терминов и понятий. Набор данных с повторяющимися предложениями формировался на основе представительной выборки основного набора данных. Схема формировалась с параметрами основного набора данных.

Имитационное моделирование проводилось в комбинированном режиме с накоплением и оценкой количества информации на каждом глобальном такте. Полученное значение оценки семантической информации отражено на графике рис. 3. Общие для наборов данных исходные параметры моделирования: размер схемы 1411 ед.; количество предложений

181; атрибут интереса a=0,95; атрибут начальной информации d=2.

Точка R на графике показывает изменение в наборах данных. В наборе данных, график показан линией 2, используются периодически повторяющиеся данные и увеличение количества информации практически не происходит. Набор данных с уникальными предложениями (линия 1) приносит постоянное увеличение количества информации.

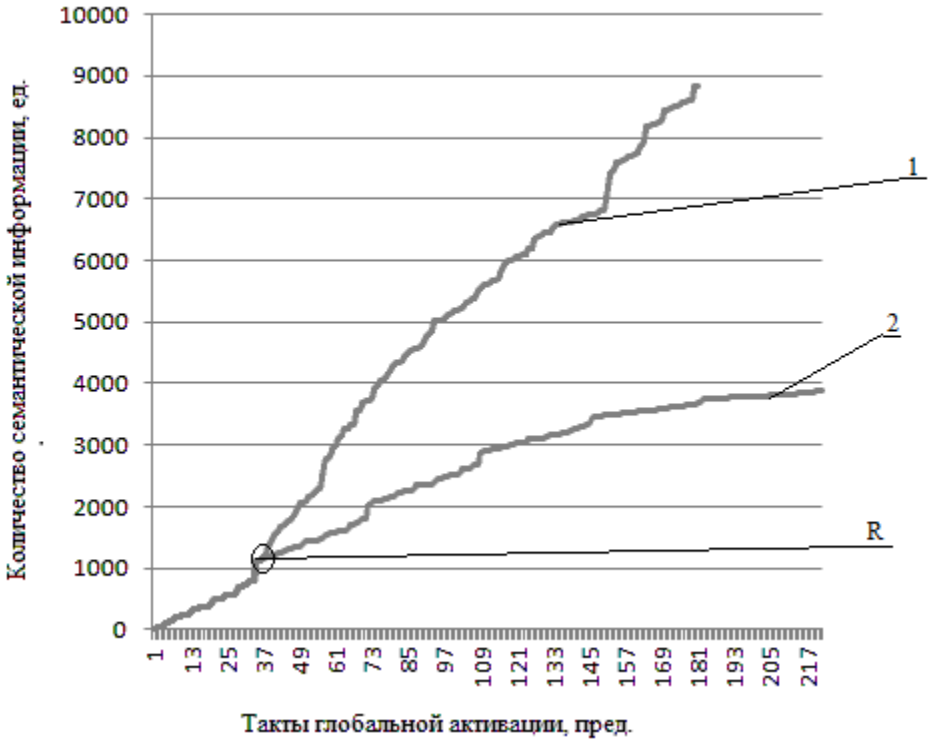


Рис. 3. Накопление семантической информации:
1 - Набор уникальных предложений; 2 - Набор повторяющихся предложений;
R – начало повтора предложений в наборе данных

Точка R иллюстрирует ситуацию, когда добавление синтаксической информации не приводит к появлению семантической.

Цель исследований семантическая емкость синтаксической информации – продемонстрировать возможность модели выполнять оценочную функцию массивов текстовых данных. Оцениваемая при моделировании характеристика может использоваться для лингвистической, семантической и прагматической оценки ресурса.

Пример сравнения текстовых материалов по критерию количества семантической информации предполагает анализ существенно различающихся по смысловому содержанию ресурсов. Исследование проводится при одинаковых синтаксических характеристиках наборов данных, но существенно различающихся смысловых параметрах текстов:

1. Художественный текст (https://ilibrary.ru/text/11/p.91/index.html).
2. Словарь основных биологических терминов (https://infotables.ru/biologiya/811-osnovnye-biologicheskie-terminy).
3. Глава учебника по биологии 6 класс (http://tepka.ru/biologiya_6/index.html).
4. Текст для детей до 5 лет (https://mishka-knizhka.ru/skazki-suteeva).

Наборы данных имели синтаксический размер 4850-5000 Бт, атрибут интереса a=0,95; атрибут начальной информации d=2. Результаты численных экспериментов показаны на рис. 4. Соотношение значений на графиках позволяет ввести целый ряд параметров, относящихся к качеству текстового ресурса.

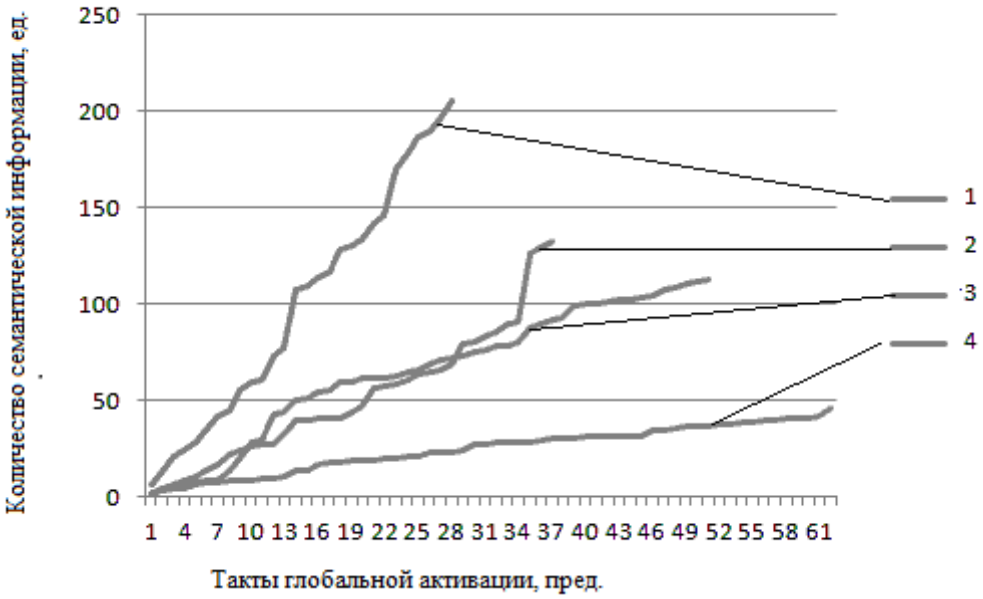


Рис. 4. Семантическая емкость синтаксической информации:
1 - художественный текст; 2 - биологический словарь; 3 – учебник; 4 - текст для детей 5 лет

Рассмотренная модель учитывает в качестве информационного элемента лексемы, словосочетания и предложения. На каждом такте вычисляются значения атрибутов текста: количество семантической информации экземпляра среды (ед. семантической информации); скорость приобретения информации (ед./пред.); семантическая энергия (ед.×пред.); семантическая емкость данных (ед/байт). Текст 1 показывает скорость накопления информации на порядок большую, чем текст 4, что может говорить об интеллектуальном уровне потребителя информации (ребенок, учащийся, студент, специалист). Количественные оценки параметров позволяют с позиций эмпирических знаний подойти к описанию параметров эффективности текстовых данных.

Выводы

Конечному пользователю нужен смысл, заключенный в синтаксических данных. Любые качественные оценки текстовых ресурсов будут носить элемент субъективизма. Использование модели оценки семантической информации предоставляет конечному пользователю возможность выбора материала (учебника, энциклопедии, справочника, тезауруса) по количественным атрибутам в зависимости от персональных запросов. Знание количественных атрибутов семантической емкости, может лежать в основе технологии финансовой оценки ресурсов.

Список литературы:

1. Звегинцев, В.А. Предложение и его отношение к языку и речи / В.А. Звегинцев. – М.: Изд-во Моск. ун-та, 1976. - 307 с.
2. Белоногов, Г.Г. Ещё раз о гносеологическом статусе понятия «информация» / Г.Г. Белоногов, Р.С. Гиляревсий // Науч.-техн. информация. - Сер. 2. Информ. процессы и системы, 2009. - №2. - С. 1-6.
3. Белоногов, Г.Г. О природе информации / Г.Г. Белоногов, Р.С. Гиляревсий, А.А. Хорошилов // Науч.-техн. информация. - Сер. 2. Информ. процессы и системы, 2009. -№1. - С. 1-6.

References:

1. Zvegincev, V.A. Predlozheniei ego otnoshenie k yazykuirechi / V.A. Zvegincev. – M.: Izd-vo. Mosk. unt.-ta - 1976. - 307 s.;
2. Belonogov, G.G. Eshchyoraz o gnoseologicheskom statuse ponyatiya «informaciya» / G.G. Belonogov, R.S. Gilyarevsij // Nauch.-tekhn. informaciya. Ser. 2. Inform. process i sistemy. - 2009. - №2. - S. 1-6.
3. Belonogov, G.G. O prirodeinformacii / G.G. Belonogov, R.S. Gilyarevsij, A.A. Horoshilov // Nauch.-tekhn. informaciya. Ser.2. Inform. process i sistemy. - 2009. -№1. - S. 1-6.

4. Осгуд, Ч. Приложение методики семантического дифференциала к исследованиям по эстетике и смежным проблемам / Ч. Осгуд, Дж. Суси, П. Танненбаум // Семиотика и искусствометрия - М., 1972.
5. Серкин, В. П. Методы психологии субъективной семантики и психосемантики : учеб. пособие для вузов. - М.: Изд-во ПЧЕЛА, 2008. - С. 253-254.
6. Волченков, Е.Я. О природе информации: физико-семантический подход / В.Я. Волченков // Науч.-техн. Информация, 2010. - №3. - С. 1-7.
7. Вяткин, В.Б. Синергетический подход к определению количества информации / В.Б. Вяткин // Информ. Технологии, 2009. - №12. - С. 68-73.
8. Лысак, И.В. Информация как общенаучное и философское понятие: основные подходы к определению [Text] / И.В. Лысак // Философские проблемы информационных технологий и киберпространства, 2015. № 2, vol. 10. – С. 9–26.
9. Шалимов, П.Ю. Математическая модель информационной среды / П.Ю. Шалимов // Вестник БГТУ, 2008. - №1. - С. 54-60.
10. Громов, Ю. Ю. Материалы к разработке теории информации. Меры количества и качества информации / Ю.Ю. Громов, В.М. Тютюнник // Фундаментальные исследования. - 2011. - № 8(2). - С. 347–355.
11. Гуревич, И.М. Информация как универсальная неоднородность / И.М. Гуревич // Информ. Технологии, 2010. - №4. - С. 66-74.
12. Киселев, Ю. А. Современное состояние электронных тезаурусов русского языка: качество, полнота и доступность / Ю.А. Киселев, С.М. Поршнев, М.Ю. Мухин // Программная инженерия, 2015. - № 6. - С. 34–40.
13. Шрейдер, Ю.А. Семантика и категоризация / Ю.А. Шрейдер // М.: Наука, 1991. - 168 с.

4. Osgud, CH. Prilozhenie metodiki semanticheskogo differenciala k issledovaniyam po estetike i smezhnym problemam / CH. Osgud, Dzh. Susi, P. Tannenbaum // Semiotika i iskusstvometriya. - M., 1972.
5. Serkin, V. P. Metody psihologii sub"ektivnoj semantiki i psihosemantiki : ucheb. posobie dlya vuzov. - M.: Izd-vo PCHELA, 2008. - Pp. 253-254.
6. Volchenkov, E.YA. O prirode informacii: fiziko-semanticheskij podhod / V.YA. Volchenkov // Nauch.-tekhn. Informaciya, 2010. - №3. - Pp. 1-7.
7. Vyatkin, V.B. Sinergeticheskij podhod k opredeleniyu kolichestva informacii / V.B. Vyatkin // Inform. Tekhnologii, 2009. - №12. - Pp. 68-73.
8. Lysak, I.V. Informaciya kak obshchenauchnoe i filosofskoe ponyatie: osnovnye podhody k opredeleniyu [Text] / I.V. Lysak // Filosofskie problemy informacionnyh tekhnologij i kiberprostranstva, 2015. № 2, vol. 10. – Pp. 9–26.
9. SHalimov, P.YU. Matematicheskaya model' informacionnoj sredy / P.YU. SHalimov // Vestnik BGTU, 2008. - №1. - Pp. 54-60.
10. Gromov, YU. YU. Materialy k razrabotke teorii informacii. Mery kolichestva i kachestva informacii / YU.YU. Gromov, V.M. Tyutyunnik // Fundamental'nye issledovaniya. - 2011. - № 8(2). - Pp. 347–355.
11. Gurevich, I.M. Informaciya kak universal'naya neodnorodnost' / I.M. Gurevich // Inform. Tekhnologii, 2010. - №4. - Pp. 66-74.
12. Kiselev, YU. A. Sovremennoe sostoyanie elektronnyh tezaurusov russkogo yazyka: kachestvo, polnota i dostupnost' / YU.A. Kiselev, S.M. Porshnev, M.YU. Muhin // Programmnaya inzheneriya, 2015. - № 6. - Pp. 34–40.
13. SHrejder, YU.A. Semantika i kategorizaciya / YU.A. SHrejder // M.: Nauka, 1991. - 168 p.

Статья поступила в редколлегию 30.04.2020.
Рецензент: д-р. техн. наук, доц., Брянский государственный технический университет Захарова А.А.
Статья принята к публикации 07.05.2020.

Сведения об авторах:

Шалимов Пётр Юрьевич
к.т.н., доцент кафедры «ИиПО» Брянского государственного технического университета
E-mail: shalimov.petr@gmail.com

Information about authors:

Pyotr Yuryevich Shalimov
Ph. D., associate Professor of the Iipo Department of Bryansk state technical University
E-mail: shalimov.petr@gmail.com