Гусеница Я. Н., Шерстобитов С. А.
Y. N. Gusenitsa, S. A. Sherstobitov

НАУЧНО-МЕТОДИЧЕСКИЙ ПОДХОД К ФОРМАЛИЗАЦИИ АДЕКВАТНОСТИ ИНФОРМАЦИИ

RESEARCH AND SYSTEMATIC APPROACH TO THE ADEQUACY OF INFORMATION FORMALIZATION

Гусеница Ярослав Николаевич – кандидат технических наук, начальник лаборатории испытательной (информатики и вычислительной техники) Военного инновационного технополиса «ЭРА» (Россия, Анапа). E-mail: yaromir226@mail.ru.
Mr. Yaroslav N. Gusenitsa – PhD, Head of Testing Laboratory (computer science and computing) Military Innovation Technopolis «ERA» (Russia, Anapa). E-mail: yaromir226@mail.ru.

Шерстобитов Сергей Александрович – кандидат технических наук, старший научный сотрудник Главного научного метрологического центра Минобороны России (Россия, Мытищи). E-mail: radosti_yad@mail.ru.
Mr. Sergey A. Sherstobitov – PhD, Major Researcher of Main Scientific Metrology Center (Russia, Mytischi). E-mail: radosti_yad@mail.ru.

Аннотация. В работе представлен научно-методический подход, который позволяет формализовать адекватность информации с позиции её синтаксических, семантических и прагматических свойств. В основе предложенного подхода лежит идея представления тезауруса с использованием семантических сетей и фреймов, а также определения его информационной энтропии. Установлены базовые закономерности прагматической адекватности информации. На конкретных примерах показаны зависимости, влияющие на её изменение. Даны рекомендации формирования тезауруса, имеющего наилучшую адекватность информации. Указана практическая значимость разработанного научно-методического подхода. Определены дальнейшие направления его развития, связанные с разработкой методов объединения тезаурусов нескольких систем, использованием теории нечётких множеств при определении семантической и прагматической адекватности, а также исследованием вопросов оценивания адекватности неоднородных моделей систем, которые построены на основе различных методов моделирования.

Summary. This paper presents a scientific and methodical approach that allows to formalize the adequacy of information from the perspective of its syntactic, semantic and pragmatic properties. At the core of this approach is the idea of submission of the thesaurus with semantic networks and frames, as well as determining its information entropy. Established the basic laws of pragmatic adequacy of the information. Specific examples are shown depending on, to roll on its change. The recommendations form a thesaurus having the best adequacy of the information. It contains the practical importance of the developed scientific and methodological approach. The further directions of its development related to the development of methods for combining multiple thesauri systems using fuzzy set theory in determining the semantic and pragmatic adequacy and research issues of evaluation of the adequacy of inhomogeneous models of systems that are based on various modeling techniques.

Ключевые слова: тезаурус, синтаксическая адекватность информации, семантическая адекватность информации, прагматическая адекватность информации, исчисление предикатов, семантическая сеть, фрейм, предикат, ориентированный граф, зона, матрица смежности, информационная энтропия.

Key words: thesaurus, syntax adequacy of information, semantic information the adequacy, adequacy of information pragmatic, predicate calculus, semantic network, frame, predicate, directed graph, area, adjacency matrix, information entropy.

УДК 519.172: 519.722

Материал поступил 29.05.2019

Введение

Современный этап развития общества характеризуется активным использованием робототехники и интеллектуальных процессов, созданием и интеграцией информационных и телекоммуникационных систем различного назначения в единое информационное пространство. И уже сейчас не существует такой области науки и техники, такой сферы практической деятельности людей, таких систем, где одним из решающих факторов прогресса не были бы информационные технологии. Они составляют неотъемлемую часть любой человеческой деятельности, в том числе и в сфере обороны и безопасности.

Основу информационных технологий составляет информация, циркулирующая в системах. С позиции материалистической философии информация представляет собой отражение объективной реальности. Информация не материальна, но она является свойством материи. Поэтому, как и любой материальный объект, информация обладает определённым качеством, причём от этого качества зависит эффективность функционирования систем в целом.

В настоящее время по вопросам определения качества информации отечественными и зарубежными теоретиками написано значительное количество трудов.

Фундаментальной работой в формализованном описании свойств информации считается статья К. Шеннона [13]. В ней приводится формула информационной энтропии для определения количества информации. Эта работа является основой теории информации, теории передачи информации, теории алгоритмов и т. д.

Другое направление по формализованному описанию свойств информации связано с определением точности и достоверности. Данное направление основано на теории вероятностей и математической статистике и широко используется на практике при обработке измерительной информации и других экспериментальных данных.

На сегодняшний день существуют работы, посвящённые описанию таких свойств информации, как репрезентативность, содержательность, полнота, доступность, актуальность, своевременность и др. [1; 7]. Однако, как это было отмечено в статье [3], указанные свойства характеризуют не столько информацию, сколько её потребителя. Поэтому для них сложно подобрать количественные показатели, которые могли бы быть использованы для формулирования объективных требований к качеству информации.

Особое внимание заслуживают работы, посвящённые количественному описанию адекватности информации. С точки зрения адекватности у информации выделяют три свойства: синтаксис, семантику, прагматику. Синтаксические свойства адекватности отражают структурный аспект информации. Семантические свойства адекватности выражают смысловой аспект информации. Прагматические свойства адекватности отражают потребительский аспект информации. Таким образом, перечисленные свойства информации соответствуют трём ступеням познания истины: от живого созерцания к абстрактному мышлению и от него – к практике, – таков диалектический путь познания истины, объективной реальности. В настоящее время по данному направлению активно ведутся исследования, основной акцент в которых сделан на формализации семантических [8] и прагматических [4; 9; 10] особенностей информации. Меньше работ посвящено комплексному определению синтаксических, семантических и прагматических особенностей информации [3]. В то же время в существующих работах недостаточное внимание уделено аналитической взаимоувязке перечисленных свойств информации.

Целью данной статьи является формализация адекватности информации с позиции её синтаксических, семантических и прагматических свойств на основе теории графов и теории информации.

Содержание научно-методического подхода

Пусть имеется система A, лингвистическое обеспечение которой построено на языке, содержащем априорную информацию о некоторой исследуемой системе B.

Элементами языка, как это описано в работе [11], являются конечное множество V существительных («индивидуалов») и конечное множество E прилагательных («предикатов»), на основе которых можно построить простые предложения. Так, если $v \in V$ есть существительное,

а $e \in E$ – прилагательное, то предложение ev читается как «v имеет свойство e» или, проще сказать, «v есть e», при этом каждое простое предложение описывает определённое состояние системы B.

Для построения более сложных предложений могут использоваться логические связки:

$\neg$ – отрицание (логическое *не*);

$\vee$ – дизъюнкция (логическое сложение);

$\vee$ – конъюнкция (логическое умножение);

$\Rightarrow$ – импликация (логическое следствие).

С позиции ассоционистской теории, представленной в работе [5], восприятие системой A системы B происходит через понятия, которые определены существительными. Понятия являются частью информации о системе B. Они же связаны с другими понятиями. При этом связи представляют собой свойства системы B. Совокупность понятий и связей между ними образуют семантическую сеть, представляющую собой ориентированный граф, вершины которого соответствуют понятиям, а дуги – ассоциациям между этими понятиями. Следовательно, совокупность всех простых предложений, описывающих состояния системы B, графически может быть представлена в виде ориентированного графа $G(V, E)$, который состоит из множества вершин V и множества дуг E. Данный ориентированный граф является тезаурусом системы A, то есть словарём, содержащим понятия и определения о системе B.

Пример простейшего тезауруса, состоящего из четырёх существительных и трёх предикатов, представлен на рис. 1.

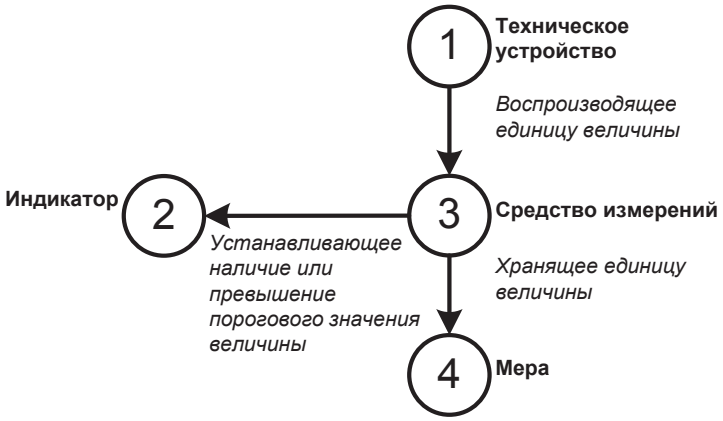

Рис.1. Графический пример простейшего тезауруса

Одной из важных особенностей представления тезауруса с использованием ориентированного графа G является то, что направления дуг между двумя смежными вершинами зависят от языка, на котором построено лингвистическое обеспечение.

Другой важной особенностью представления тезауруса с использованием ориентированного графа G является возможность графического изображения принципа иерархического наследования свойств понятий [12]. Например, в приведённом на рис. 1 тезаурусе понятия «мера» и «индикатор» обладают всеми свойствами понятия «средство измерений».

Для графического представления сложного предложения, состоящего из двух простых, необходимо осуществить отождествление смежных вершин v и u ориентированного графа G, которые для некоторой дуги e являются начальной и конечной вершинами соответственно. Дуга e, инцидентная обеим вершинам, должна соответствовать предикату, который составляет первую часть сложного предложения. Эта же дуга удаляется из ориентированного графа G. Вторую часть сложного предложения составляет дуга e', для которой инцидентная вершина u является начальной. Аналогично осуществляется графическое представление сложных предложений, состоящих

из большего количества простых предложений. Результатом применения описанной операции является новый ориентированный граф G' с меньшим количеством вершин и дуг. Для приведённого выше примера тезаурус, содержащий сложные предложения, представлен на рис. 2.

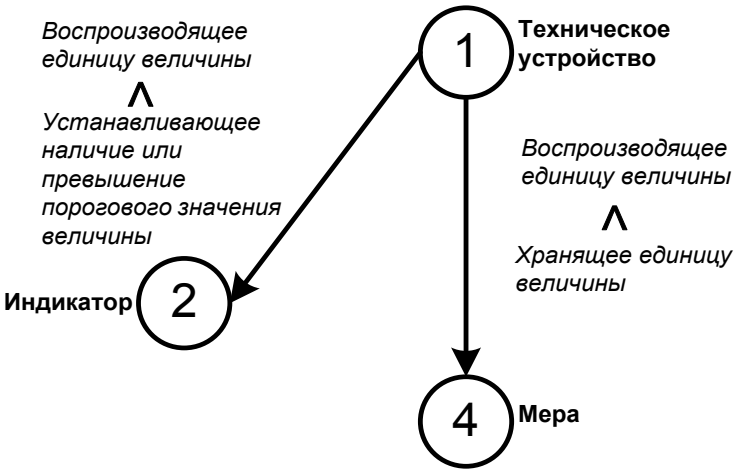

Рис. 2. Графический пример простейшего тезауруса со сложными предложениями

Если система A получает сведения о системе B, то её тезаурус может расширяться, и в ориентированном графе G могут появляться новые вершины v и новые дуги e. Пример тезауруса, расширенного одним существительным и одним предикатом, показан на рис. 3.

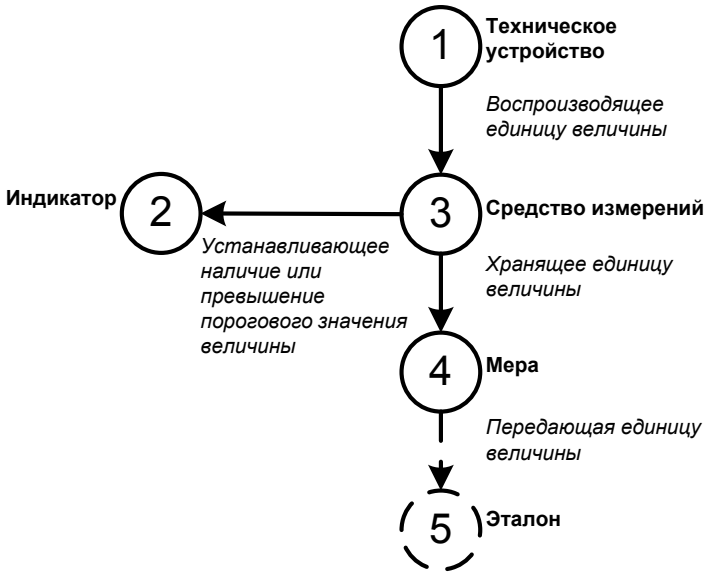

Рис. 3. Графический пример простейшего тезауруса, дополненного одним существительным и одним предикатом

Для ориентированного графа G показателем структуры является матрица смежности R, под которой понимается квадратная матрица с элементами a_{ij}, принимающими значение 1, если вершины v_i и v_j смежные, и 0 – в противном случае. При этом порядок матрицы R определяется количеством вершин. Следовательно, если тезаурус системы A содержит n существительных и

m не перекрывающихся по смыслу предикатов, то в качестве показателя синтаксической адекватности информации о системе B может быть использована матрица смежности R порядка n.

Матрица смежности ориентированного графа G, представленного на рис. 1, имеет следующие значения:

$$R(G)=\begin{vmatrix}0 & 0 & 1 & 0\\ 0 & 0 & 0 & 0\\ 0 & 1 & 0 & 1\\ 0 & 0 & 0 & 0\end{vmatrix}.$$

Описание семантической адекватности информации предполагает использование дополнительных исходных данных.

Прежде всего необходимо знать для каждого i-го существительного вероятность r_i его использования системой A. С позиции математической статистики r_i представляет собой частоту использования i-го существительного из n возможных. Поэтому $\sum_{i=1}^{n} r_i = 1$.

Кроме того, для всех предикатов, по аналогии с существительными, необходимо определить вероятности p_{ij}, которые характеризуют достоверность наличия свойств у системы B. При этом если из вершины v_i выходит несколько дуг, то $\sum_{\forall j: e_{ij}=(v_i,u_j)} p_{ij} = 1$. Все вероятности p_{ij} являются элементами матрицы вероятностей переходов $M(G)$. Данная матрица получается из матрицы $R(G)$ путём замены элементов a_{ij} на вероятности p_{ij}. При получении системой A новых сведений о системе B вероятности могут изменяться.

Чтобы определить вероятности состояний системы B на основе информации, которая имеется у системы A, необходимо преобразовать $R(G)$ в матрицу $S(G)$ следующим образом:

$$S(G)=\begin{vmatrix}S_{11} & S_{12} & \dots & S_{1n}\\ S_{21} & S_{22} & \dots & S_{2n}\\ \dots & \dots & \dots & \dots\\ S_{n1} & S_{2n} & \dots & S_{nn}\end{vmatrix},$$

где $s_{ij} = r_i \cdot p_{ij}$.

В связи с тем, что каждый элемент матрицы $S(G)$ является вероятностью определённого состояния системы B, то в качестве показателя семантической адекватности информации может использоваться информационная энтропия

$$H = -\sum_{i=1}^{n}\sum_{j=1}^{n} s_{ij}\log_2 s_{ij}. \qquad (1)$$

Значение показателя семантической адекватности информации зависит от значения показателя синтаксической адекватности информации. Если изменить структуру информации о системе B, то изменится её восприятие системой A. В частности, если в тезаурусе системы A будет недоставать понятий и определений, требуемых для восприятия сведений о системе B, то информационная энтропия будет достаточно высока. Это объясняется тем, что в ориентированном графе G будут отсутствовать определённые вершины и дуги, а следовательно, будет иметь другое значение матрица $S(G)$.

Описание прагматической адекватности информации основано на изменении информационной энтропии H. Если информационная энтропия H уменьшается, то для системы A полученные сведения о системе B являются полезными. Если информационная энтропия H не изменяется, то для системы A полученные сведения о системе B являются бесполезными. Наконец, если информационная энтропия H увеличивается, то для системы A полученные сведения о системе B являются дезинформацией.

С учётом этого факта для количественного определения показателя прагматической адекватности информации необходимо найти разность значений информационной энтропии системы A после и до получения сведений о системе B:

$$I = H_{\tau} - H_{\tau-1}, \tag{2}$$

где $H_{\tau-1}$ – информационная энтропия системы A до получения сведений о системе B; H_{τ} – информационная энтропия системы A после получения сведений о системе B.

Как видно из формулы (2), показатель прагматической адекватности информации зависит от показателя семантической адекватности информации. Если система A не воспринимает сведения о системе B, то такая информация ей не нужна.

Кроме того, показатель прагматической адекватности информации обладает следующими свойствами:

1. Если $I = 0$, то полученные сведения о системе B являются бесполезными.
2. Если $I > 0$, то полученные сведения о системе B являются полезными.
3. Если $I < 0$, то полученные сведения о системе B являются дезинформацией.

Следует отметить, что перечисленные особенности показателя прагматической адекватности информации схожи со свойствами ценности информации, предложенными в работе [9].

Другие способы представления информации

В многочисленных работах в области искусственного интеллекта помимо семантических сетей используют другие способы представления информации. Основными из них являются: исчисление предикатов, продукция, фреймы.

В работе [5] отмечается, что семантические сети мало чем отличаются от исчисления предикатов и продукции. Использование ориентированных графов в семантических сетях позволяет лишь наглядно продемонстрировать отношения между понятиями. Значительным преимуществом семантических сетей, по сравнению с исчислениями предикатов и продукцией, является возможность первых графически изображать принцип иерархического наследования свойств понятий. Эта же особенность определяет их сходство с фреймами, которые впервые описаны в работе [6].

Вместе с тем считается, что фреймы обладают преимуществом перед семантическими сетями, так как не только имеют возможность изображения иерархического принципа наследования свойств понятий, но и позволяют описывать табличное представление информации. Однако, на наш взгляд, семантические сети являются более общим способом представления информации и пригодными для табличного представления информации.

Продемонстрируем наше утверждение на следующем примере (см. табл. 1). В табл. 1 имеется информация о количестве сотрудников метрологического воинского подразделения, включённых в распределение календарного фонда рабочего времени.

Таблица 1

Количество сотрудников

Категория	На год, чел.	Фактически выполняло работы (участвовало в мероприятиях) в течение года, чел.
Военнослужащие	2	1
Гражданский персонал Вооружённых сил РФ	6	5

Для представления таблицы в виде ориентированного графа необходимо осуществить следующие действия. В качестве начальной вершины ориентированного графа выбрать имя таблицы. Из начальной вершины построить дугу в вершину, соответствующую заголовку первого столбца таблицы. Из неё построить дугу к вершине, которая соответствует заголовку второго столбца таблицы. Аналогично необходимо построить дуги для вершин, соответствующих заголовкам остальных столбцов таблицы. Каждая запись в таблице представляет собой путь с вершинами, соответствующими значениям полей. Эти же вершины являются конечными для дуг, исходящих из вершин, соответствующих заголовкам столбцов таблицы.

В результате описанных действий будет получен ориентированный граф, представленный на рис. 4.

Если таблица имеет достаточно много строк, то семантическая сеть получается довольно сложной. Поэтому отдельные фрагменты семантической сети целесообразно заменять таблицами.

Таким образом, графическое изображение тезауруса системы A о системе B является комбинацией семантических сетей и фреймов, объединяющей достоинства обоих способов представления информации.

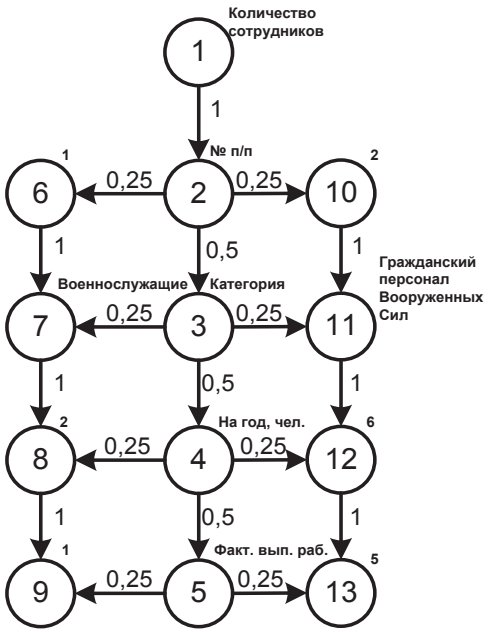

Рис. 4. Ориентированный граф, соответствующий табл. 1

Расчётные примеры

Пример 1. Рассмотрим расчётный пример для тезауруса системы A, представленного в виде ориентированного графа $G_{\tau-1}$, содержащего априорную информацию (см. рис. 5, *а*), и ориентированного графа G_{τ}, содержащего апостериорную информацию (см. рис. 5, *б*) о системе B, при этом ориентированный граф $G_{\tau-1}$ отличается дополнительным существительным со связующими предикатами.

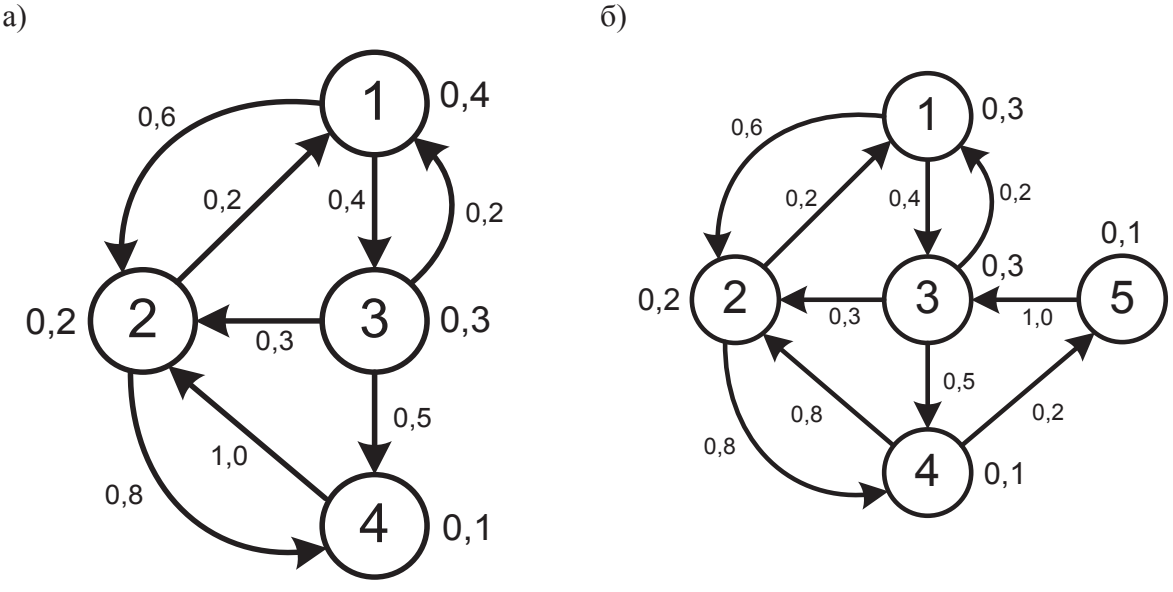

Рис. 5. Графический пример простейшего тезауруса системы A:
а – с априорной информацией о системе B; б – с апостериорной информацией о системе B

Показатели синтаксической адекватности информации с априорной и апостериорной информацией равны соответственно:

$$R_{\tau-1}(G)=\begin{vmatrix}0&1&1&0\\1&0&0&1\\1&1&0&1\\0&1&0&0\end{vmatrix} \text{ и } R_{\tau}(G)=\begin{vmatrix}0&1&1&0&0\\1&0&0&1&0\\1&1&0&1&0\\0&1&0&0&1\\0&0&1&0&0\end{vmatrix}.$$

На основе $R_{\tau-1}(G)$ и $R_{\tau}(G)$ получим матрицу $S_{\tau-1}(G)$ тезауруса системы A с априорной информацией о системе B и матрицу $S_{\tau}(G)$ тезауруса системы A с апостериорной информацией о системе B:

$$S_{\tau-1}(G)=\begin{vmatrix}0&0,24&0,16&0\\0,04&0&0&0,16\\0,06&0,09&0&0,15\\0&0,1&0&0\end{vmatrix} \text{ и } S_{\tau}(G)=\begin{vmatrix}0&0,18&0,12&0&0\\0,04&0&0&0,16&0\\0,06&0,09&0&0,15&0\\0&0,08&0&0&0,02\\0&0&0,1&0&0\end{vmatrix}.$$

С использованием формулы (1) рассчитаем показатели семантической адекватности информации для тезауруса системы A с априорной и апостериорной информацией о системе B:

$$H_{\tau-1}=0,372\,[\text{бит}] \text{ и } H_{\tau}=0,332\,[\text{бит}].$$

Теперь определим показатель прагматической адекватности информации. На основе формулы (2) получим $I=0,372-0,332=0,04\,[\text{бит}]$.

Таким образом, в рассмотренном примере добавление в тезаурус системы A сведений о системе B снижает общую энтропию, а сведения, полученные о системе B, являются полезными.

Пример 2. Воспользуемся исходными данными для ориентированного графа G_{τ} из примера 1 (см. рис. 3, *б*). Добавим предикат, связывающий существительное *5* с существительным *1*. Перераспределив вероятности, получим новый тезаурус системы A с информацией о системе B, представленный на рис. 6 в виде ориентированного графа $G_{\tau+1}$.

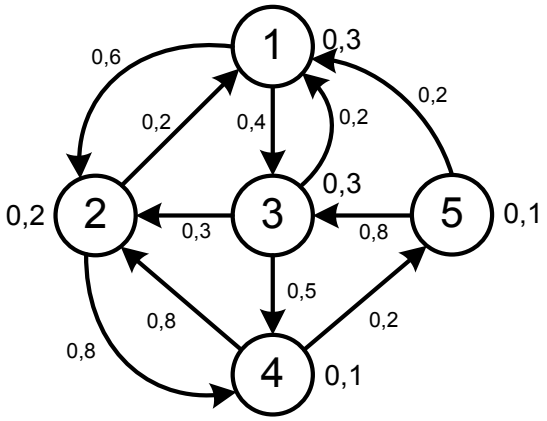

Рис. 6. Графический пример простейшего тезауруса системы A с дополнительным предикатом

Аналогичным образом рассчитаем для него информационную энтропию, получим $H_{\tau+1} = 0,328$ [бит]. Определим показатель прагматической адекватности информации $I = H_{\tau+1} - H = 0,004$ [бит]. Следовательно, добавление в тезаурус системы A сведений о системе B также снижает общую энтропию, а сведения, полученные о системе B, являются полезными.

Пример 3. Рассмотрим несколько ориентированных графов, представляющих изменение тезауруса системы A в зависимости от поступления информации о системе B (см. рис. 7).

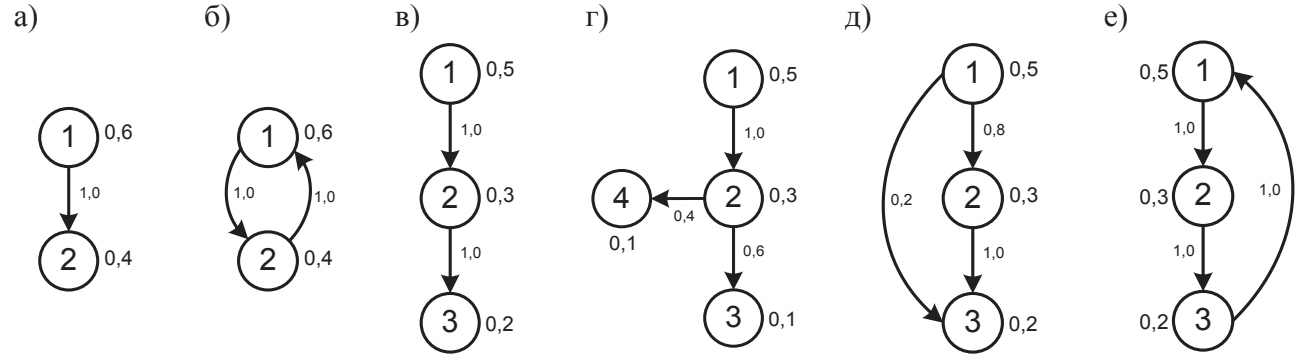

Рис. 7. Графический пример изменения тезауруса системы A в зависимости от получения сведений о системе B

С помощью формулы (1) рассчитаем показатели семантической адекватности информации. Результаты расчётов приведены в табл. 2.

Таблица 2

Результаты расчётов (*пример 3*)

Обозначение тезауруса в соответствии с рис. 6	а)	б)	в)	г)	д)	е)
Значение показателя семантической адекватности	0,814	1,117	0,673	0,612	0,505	0,759

Из полученных расчётов видно, как изменяется значение показателя семантической адекватности информации при добавлении новых вершин и новых дуг. Так, например, добавление новых вершин и новых дуг приводит к снижению энтропии. Это объясняет используемые на практике методы проектирования информационных систем, в том числе нормализацию баз данных, когда одна таблица разбивается на несколько связанных между собой таблиц. Кроме того, значение показателя семантической адекватности информации возрастает при появлении замкнутых структур, которые нежелательны при формировании тезауруса. В теории графов подобные структуры называются зонами. Каждая зона представляет собой подграф $C_k = (X_k, U_k)$ ориентированного графа G, в котором для любой пары вершин найдётся путь из одной в другую. Для отыскания зон в управляющем графе может быть использован алгоритм, представленный в работе [2].

Заключение

Таким образом, представленный научно-методический подход позволяет формализовать адекватность информации с позиции её синтаксических, семантических и прагматических свойств. В основе предложенного подхода лежит идея представления тезауруса с использованием семантических сетей и фреймов, а также определения его информационной энтропии. Установлены базовые закономерности прагматической адекватности информации. На конкретных примерах показаны зависимости, влияющие на её изменение. Даны рекомендации формирования тезауруса, имеющего наилучшую адекватность информации.

Представленный научно-методический подход может быть использован при проектировании баз данных, хранилищ данных и баз знаний, а также для исследования систем различного назначения.

Дальнейшее направление развития предлагаемого научно-методического подхода авторы связывают с разработкой методов объединения тезаурусов нескольких систем, использованием теории нечётких множеств при определении семантической и прагматической адекватности, а также исследованием вопросов оценивания адекватности неоднородных моделей систем, которые построены на основе различных методов моделирования.

ЛИТЕРАТУРА

1. Горяев, Ю. А. Информатика: учеб. пособие / Ю. А. Горяев. – М.: МИЭМП, 2005. – 116 с.
2. Гусеница, Я. Н. Алгоритм поиска зон в управляющих графах / Я. Н. Гусеница // Информатика и системы управления. – 2017. – № 3(53). – С. 119–124.
3. Князев, В. В. Качество информации в прикладных информационных системах сферы сервиса / В. В. Князев // Вестник Ассоциации вузов туризма и сервиса. – 2008. – № 1. – С. 11–19.
4. Корогодин, В. И. Информация как основа жизни / В. И. Корогодин, В. Л. Корогодина. – Дубна: Изд. центр «Феникс», 2000. – 205 с.
5. Люгер, Дж. Ф. Искусственный интеллект: стратегии и методы решения сложных проблем / Дж. Ф. Люгер. – 4-е изд. – М.: Вильямс, 2003. – 864 с.
6. Минский, М. Фреймы для представления знаний / М. Минский. – М.: Мир, 1979. – 152 с.
7. Основы современных компьютерных технологий: учебник / Г. А. Брякалов, С. В. Войцеховский, Е. Г. Воробьев [и др.]; под ред. А. Д. Хомоненко. – СПб.: КОРОНА-принт, 2005. – 672 с.
8. Парамонов, И. Ю. Мера информационной мощности тезауруса и её применение / И. Ю. Парамонов, В. А. Смагин // Интеллектуальные технологии на транспорте. – 2016. – Вып. 8. – С. 5–9.
9. Харкевич, А. А. О ценности информации / А. А. Харкевич // Проблемы кибернетики. – 1960. – Вып. 4. – С. 53–72.
10. Шанкин, Г. П. Ценность информации. Вопросы теории и приложений / Г. П. Шанкин. – М.: Филоматис, 2004. – 128 с.
11. Carnap, R. An Outline of a Theory of Semantic Information / R. Carnap, Y. Bar-Hillel // The Journal of Symbolic Logic. – 1954. – Vol. 19(3). – P. 230–232.
12. Collins, A. M. Retrieval time from semantic memory / A. M. Collins, M. R. Quillian // Journal of verbal learning and verbal behavior. – 1969. – Vol. 8. – № 2. – P. 240–247.
13. Shannon, C. E. A Mathematical Theory of Communication / C. E. Shannon // Bell System Technical Journal. – 1948. – Vol. 27. – P. 379–423.