И.А. Бессмертный

Контекстный подход к оценке количества информации в базах знаний

Аннотация. В статье рассматривается задача количественной оценки информативности формализованных источников информации – баз знаний и данных. Основным отличием от известных методов является использование контекста – словаря понятий, используемых в предметной области. В рамках контекста вычисляется также количество информации, содержащееся в элементарных фактах вида субъект-предикат-объект. Приводятся примеры для небольших предметных областей.

Ключевые слова: база знаний, контекст, информативность.

Введение

Один из основоположников искусственного интеллекта Алан Тьюринг, автор знаменитого теста, высказал предположение, что при емкости памяти 1 млн. бит компьютер сможет успешно пройти его тест [1]. К настоящему времени этот уровень превышен, по меньшей мере, на 4 порядка даже для мобильных устройств, но это не приблизило нас к решению задачи. Более того, нет надежных метрик для оценки интеллекта, которые могут применяться как к естественному, так и искусственному интеллекту. Современные исследования концентрируются вокруг оценки информационной емкости мозга [2] в байтах или косвенных методов оценки интеллекта через время, затраченное на обучение. В статье [3] знания оцениваются числом понятий и связей между ними (фактов), но количество информации, содержащееся в этих понятиях и фактах, остается за рамками исследований.

В данной работе описывается подход к измерению информативности понятий и фактов, который далее может использоваться как для оценки количества информаций в базах знаний, так и для количественных измерений в естественном интеллекте.

1. Оценка информативности понятий

К. Шеннон в 1948 г. предложил единицу информации бит [4] как величину энтропии для одного из двух равновероятных событий. Общая формула энтропии для n состояний случайной величины x

$$H(x) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i). \qquad (1)$$

Для двоичной пары (0,1) энтропия равна единице при условии равной вероятности нуля и единицы. При условии, что для всех разрядов двоичного числа вероятности нулей и единиц равны, число разрядов в точности соответствует количеству информации I в числе, определяемому как двоичный логарифм числа состояний:

$$I = \log_2 n.$$

Количество информации в одной букве русского алфавита (в предположении, что вероятности всех букв равны) приблизительно равно пяти битам, что соответствует пяти двоичным разрядам, которыми русский алфавит как раз и может быть закодирован.

Механическое перенесение энтропии алфавита на составляемые с его помощью слова лишено смысла, иначе слова из редко встречающихся букв будут более информативными,

а английский текст менее информативным, чем русский, поскольку не только латинский алфавит содержит меньше знаков, но и английские слова обычно короче русских или немецких. Количество информации в одном и том же слове «скорость» на английском языке (*speed*) и немецком (*Geschwindigkeit*) будет различаться в три раза.

Слово – это код для обозначения понятия, причем заведомо избыточный. Пусть средняя длина слова равна шести знакам, каждый из которых кодируется пятью битами, т.е. имеет 32 значения, если игнорировать букву «ё». В таком случае максимальное число слов равно $32^6 \approx 1$ млрд слов. Это означает, что избыточность составляет не менее двух порядков. Если рассматривать слово как минимальную лексическую единицу (атом), то к его информативности можно применить подход Шеннона. Применяя формулу (1) для словаря D, получим значение энтропии

$$H(D) = -\sum_{i=1}^{Card(D)} p(w_i)\log_2 p(w_i), \qquad (2)$$

где $card(D)$ – мощность словаря, w_i – i-е слово словаря, $p(w_i)$ – вероятность появления слова w_i, которая может быть вычислена следующим образом:

$$p(w_i) = \frac{n_i}{\sum n_i},$$

где n_i – число экземпляров слова w_i во множестве документов.

Таким образом, энтропия словаря зависит от набора документов, использующих данный словарь. Если пренебречь разной частотностью слов, составляющих словарь, то количество информации $I(w,D)$ в слове w, определенном на множестве D (словаре), можно определять

$$I(w,D) = log_2(card(D)), \qquad (3)$$

где $card(D)$ – мощность словаря, $w \in D$.

Здесь мы сталкиваемся с проблемой определения $card(D)$. Автор текста владеет одним словарным запасом, а его читатели – другим. Следовательно, информативность текста становится субъективной характеристикой, что, впрочем, соответствует действительности. Например, если читатель встречает незнакомое слово, лежащее вне его словаря, то его информативность для него равна нулю. Кроме того, богатый словарный запас из одной области знаний бесполезен для написания или восприятия текста из другой области. В работе [5] предлагается оценивать информативность понятий в рамках контекста, что делает вычислимыми мощности словарей. В статье [6] определяется роль контекста в задаче анализа технических текстов, а в статье [7] – описывается процесс автоматического создания прикладных онтологий на базе тезаурусов, составляющих контекст предметной области. Проект, представленный в работах [6, 7], убедительно демонстрирует возможность анализа в рамках контекста даже неформализованных документов.

Пусть в студенческой группе 25 человек, тогда информативность выбора конкретного студента s, $s \in S$, в контексте группы S: $I(s,S) = \log_2 25 = 4{,}64$. Если речь идет о потоке, состоящем из четырех групп, то информативность потока y $I(g) = \log_2 4 = 2$, а информативность выбора студента в контексте потока $I(s,g) = I(s,S) + I(g) = \log_2 25 + \log_2 4 = 6{,}64$. Правда, простым сложением можно вычислять информативность одного и того же понятия на разных уровнях контекста только в простых случаях, когда каждое подмножество контекста имеет одинаковую мощность. В противном случае $I(s,g) = \log_2(card(S,g))$, где $card(S,g)$ - мощность множества S в контексте g, что в рассматриваемом примере соответствует числу студентов на потоке. Таким образом, количество информации, содержащееся в любом понятии достаточно просто определить в рамках определенного контекста.

2. Контекст и количественная оценка информативности фактов

В работе В. Петровского [8] дается определение информативности научного факта, как меры неопределенности, устраняемой в ходе эмпирической проверки гипотезы. При этом вполне допускается, что результаты прогнозов, касающихся исхода эксперимента (эксп), могут соответствовать эмпирическим (эмп) по критерию совпадения выборочных средних значений ($M_{эксп} = M_{эмп}$). В то же время разброс экспертных оценок может существенно превосходить разброс реальных эмпирических данных. Эмпирической мерой информативности научного факта Петровский предлагает считать статистически значимое различие между выборочными дисперсиями ($D_{эксп} - D_{эмп}$) или, соответственно, между выборочными стандартными отклонениями прогнозируемого и реального

положения дел ($d_{эксп} - d_{эмп}$). В том случае, если разброс экспертных оценок меньше разброса эмпирических (при равенстве средних), то статистически достоверная разница выборочных дисперсий (стандартных отклонений) свидетельствует о неоднозначности научного факта, или, что, скорее всего, о том, что научный факт описывает частный случай. В данном методе декларируется энтропийный подход, что вполне оправданно, но при этом количественная мера каждого факта становится привязанной к метрикам событий, которые описывает факт. Например, информативность факта о том, что вода кипит при +100°C, будет измеряться не в битах, а в градусах Цельсия.

В книге Н. Валгиной [9] делается анализ информативности текста как степени его смыслосодержательной новизны для читателя, которая заключена в теме и авторской концепции, системе авторских оценок предмета мысли. С этой точки зрения читатели составляют три группы:

а) соответствующие авторской ориентации;

б) не достигшие уровня тезауруса автора;

в) тезаурус читателей превышает тезаурус автора.

Для читателей третьей группы информативность текста приближается к нулю, а избыточность – к 100%. Читателям второй группы не хватит базовых знаний, что отрицательно скажется на полезности информации. И только для читателей первой группы текст будет максимально полезен (Рис.1).

Таким образом, мы получаем еще одно подтверждение необходимости определения информативности фактов в определенном контексте. Предположим, что количество информации $I(w,D)$ в слове w, определенном на словаре (контексте) D, известно. Минимальной же смысловой единицей текста является не слово, а фраза или предложение подобно тому, как единицей количественных данных является число, а не цифра. При этом разрядность числа далеко не всегда отражает его информативность. Нельзя утверждать, что число 5 менее информативно, чем число 1000000 потому, что информативность числа определяется исключительно числом его возможных состояний. Например, температура тела человека в градусах измеряется трехзначным числом с одним знаком после запятой. Исходя из разрядности числа температуры, можно сделать ошибочный вывод о том, что количество информации в

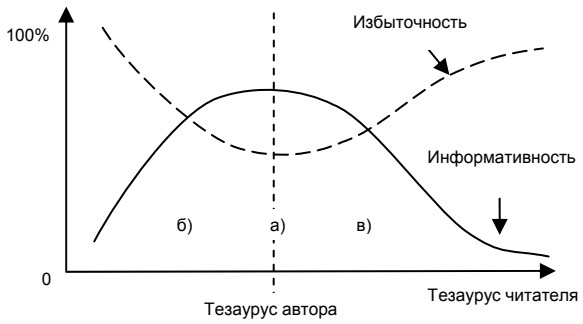

Рис. 1. Информативность и избыточность текста для читателей

значении температуры тела человека определяется исходя из количества состояний трехзначного числа (0,0 – 99,9°C):

$$I(t°) = \log_2 10^3 \approx 10 \text{ бит}.$$

Между тем, диапазон температуры тела человека лежит в пределах 35-42°C, следовательно, число возможных значений измеряемой температуры составляет не 1000, а только 70, и $I(t°) = \log_2 70 \approx 6{,}13$ бит. Аналогичным образом, информативность $I(s)$ фразы s определяется не числом слов, а множеством их возможных размещений (не сочетаний, поскольку порядок слов несет смысл):

$$I(s) = \log_2 A_n^k,$$

где k – длина фразы, $n = card(D)$ – мощность словаря.

Однако данная формула не может быть применена, поскольку:

- не отражает подмножество возможных размещений, а включает в себя все размещения;

- невозможно установить значение k, поскольку все фразы имеют разное число слов, или здесь должна использоваться максимально возможная длина фразы;

- мощность словаря – величина субъективная, как показано выше.

На основании вышесказанного мы приходим к невозможности точного вычисления количества информации в текстах, написанных на естественных языках.

Невозможность вычисления информативности фразы не означает, что число слов в тексте не может служить приблизительной мерой его информативности. Как показано в работе [5], два выражения «британец» и «подданный ее Величества Королевы Елизаветы II» эквива-

лентны, но различаются по количеству слов, а значит, информации, в шесть раз. Однако при внимательном рассмотрении тождественность этих выражений оказывается не абсолютной. На самом деле второе выражение более информативно, поскольку британец может быть подданным короля Ричарда Львиное Сердце или любого британского короля. Понятие «современный британец» уже более приближено ко второму выражению, но все же не идентично, поскольку привязано к времени, уточнение которого потребует дополнительных слов. Любые попытки привести первое выражение по смыслу ко второму неизбежно увеличат число слов.

Не претендуя на измерение информации в произвольных текстах, тем не менее, можно попытаться вычислить количество информации, содержащееся в формализованных знаниях, а именно в элементарных фактах. Если лексической единицей является слово (атом), то семантической единицей (аналогом молекулы) является факт. Пусть имеется факт вида (*s,p,o*), где s – субъект, $s \in S$, p – предикат, $p \in P$, o – объект, $o \in O$. Предположим, что, руководствуясь методом, рассмотренным выше, мы имеем оценки информативности субъекта $I(s)$, предиката $I(p)$ и объекта $I(o)$, вычисленные по отдельности для некоторого контекста. Очевидный способ вычисления информативности факта

$$I(s,p,o) = I(s) + I(p) + I(o),$$

что эквивалентно выбору факта из декартова произведения $F = S \times P \times O$ – множества всех возможных сочетаний субъектов, предикатов и объектов. Размерность F является чрезмерной, поскольку включает в себя заведомо несуществующие сочетания субъектов, предикатов и объектов.

Рассмотрим простой мир (контекст) $T = (E, P, F)$, где E – множество сущностей, P – множество предикатов, для которого имеется набор фактов $F = \{f\}$:

Волга впадает в Каспийское море.
Нева впадает в Балтийское море.
Самара стоит на Волге.
Рига стоит на Даугаве.

Таким образом, данный контекст насчитывает семь сущностей $E = \{e\}$ = {Волга, Даугава, Нева, Каспийское море, Балтийское море, Самара, Рига}, которые могут выступать в качестве субъекта или объекта, и два предиката P = {впадает, стоит на}. Исходя из этих данных, количество информации, содержащееся в любом из фактов $f = (s,p,o)$ контекста T, будет равно:

$$I(f) = I(e) + I(p) + I(e) = \log_2 7 + \log_2 2 + \log_2 7 = 2{,}8+1+2{,}8 = 6{,}6 \text{ бит},$$

что соответствует выборке из декартова произведения $E \times P \times E$, мощность которого $card(E \times P \times E) = 98$. Объем информации, содержащейся во всем контексте T

$I(T) = I(\{f\}) = I(f) \cdot card(F) = 6{,}6*4 = 26{,}46$ бит.

Табл. 1. Фрагмент декартова произведения $E \times P \times E$

Субъект	**Предикат**	**Объект**
Волга	впадает	Волга
Волга	впадает	Нева
Волга	впадает	Самара
Волга	впадает	Рига
Волга	**впадает**	**Каспийское море**
Волга	**впадает**	**Балтийское море**
Волга	впадает	Даугава
Нева	впадает	Волга
Нева	впадает	Нева
…	…	…

В составе декартова произведения присутствуют все возможные комбинации «сущность-предикат-сущность» (*e,p,e*), включая бессмысленные, как «Волга впадает в Самару» и т.п. Фрагмент декартова произведения $E \times P \times E$ приведен в Табл. 1.

Декартово произведение дает максимально возможное сочетание сущностей с предикатами, следовательно, информативность фактов, вычисленная на нем, является верхней оценкой.

Найдем пересечение множеств понятий, включающих субъект, предикат и объект, как показано на Рис. 2. Пересечение $S \bigcap P$ включает в себя подмножество кортежей $\{(s,p)\}$, встречающихся в фактах F контекста T, а $O \bigcap P$ – аналогичное подмножество кортежей $\{(p,o)\}$. Пересечение $S \bigcap P \bigcap O$ даст подмножество допустимых всех кортежей $\{(s,p,o)\}$, среди которых есть неверные, как “Нева, впадает, Каспийское море”, но нет семантически невозможных.

В Табл. 1 кортежи, входящие в $S \bigcap P \bigcap O$, выделены полужирным шрифтом. Пересечение $S \bigcap P \bigcap O$ приведено в Табл. 2.

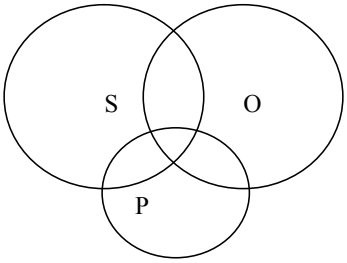

Рис. 2. Пересечение понятий на множестве фактов

Табл.2. Пересечение $S \bigcap P \bigcap O$ контекста T

Субъект	**Предикат**	**Объект**
Волга	впадает	Каспийское море
Волга	впадает	Балтийское море
Нева	впадает	Каспийское море
Нева	впадает	Балтийское море
Самара	стоит на	Волга
Самара	стоит на	Даугава
Рига	стоит на	Волга
Рига	стоит на	Даугава

Количество информации, содержащееся в триплете $f = (s,p,o)$ на подмножестве $S \bigcap P \bigcap O$ для контекста T

$$I(f, S \bigcap P \bigcap O) = \log_2(card(S \bigcap P \bigcap O)) =$$
$$= \log_2 8 = 3\text{бит},$$

где $s \in S$, $p \in P$, $o \in O$.

Общее количество информации в контексте T составит:

$$I(F, S \bigcap P \bigcap O) = I(f, S \bigcap P \bigcap O) \cdot card(F) =$$
$$= 3 \cdot 4 = 12\text{бит},$$

Однако поступив таким образом, мы часть знаний выносим в контекст верхнего уровня, а именно, обобщенные знания о том, что реки впадают в моря, в озера или в другие реки, что горы имеют высоту, а города – население и т.д.

Введем в контекст множество классов $C = \{c\}$ = {река, море, город} и определим принадлежность сущностей e к классам c, т.е. введем множество предикатов для отношений между экземпляром и классом P_{ec} = (*is a* или *ISA*) и установим множество фактов $F_{ec} = \{f_{ec}\}$:

Волга *ISA* река.
Даугава *ISA* река.
Нева *ISA* река.
Рига *ISA* город.
Самара *ISA* город.
Каспийское море *ISA* море.
Балтийское море *ISA* море.

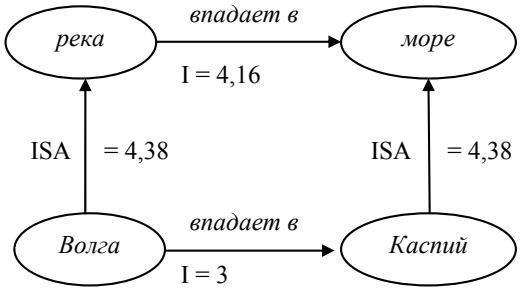

Рис.3. Информативность фактов в таксономическом контексте

Количество информации, заключенное в любом из этих фактов

$$I(f_{ec}) = I(e,ISA,c) = \log_2 7 + \log_2 1 + \log_2 3 =$$
$$= 2{,}8+0+1{,}58 = 4{,}38 \text{ бит},$$

а объем информации, содержащейся во всех этих фактах

$$I(F_{ec}) = I(e,ISA,c) \cdot card(\{e,ISA,c\}) = 4{,}38 \cdot 7 = 30{,}66 \text{ бит}.$$

Введем факты $F_{cc} = \{f_{cc}\}$, определяющие отношения между классами:

Река впадает в море.
Город стоит на реке.

Количество информации, содержащееся в любом из этих фактов,

$$I(f_{cc}) = I(c,p,c) = \log_2 3 + \log_2 2 + \log_2 3 =$$
$$= 1{,}58 + 1 + 1{,}58 = 4{,}16 \text{ бит},$$

а суммарная информация, содержащаяся в этих фактах

$$I(F_{cc}) = I(f_{cc}) \cdot card(F_{cc}) = 4{,}16 \cdot 2 = 8{,}32 \text{ бит}.$$

Таким образом, суммарное количество информации, заключенное в контексте, подвергнутом классификации, или в таксономическом контексте $T' = (E, E,C,F,P,P_{ec},F_{cc},F_{ec})$,

$$I(T') = I(F, S \bigcap P \bigcap O) + I(F_{ec}) + I(F_{cc}) =$$
$$= 12+30{,}66 + 8{,}32 = 50{,}98 \text{ бит}.$$

Эта величина почти в два раза превосходит информативность контекста Т, но и число фактов увеличилось почти в 3 раза, с 4 до 13. Рис. 3 иллюстрирует информативность фактов в контексте T'. Факт «Волга впадает в Каспийское море» в данном контексте имеет информативность всего 3 бита, но это значение вытекает из фактов, что Волга – это река, Каспий это море, а реки (в данном контексте) впадают в моря, а не в озера или в другие реки.

Итак, классификация в пределах контекста снижает количество информации, несущей каждым фактом, а значит, и величину энтропии,

что является ожидаемым результатом. Общая величина энтропии, однако, возрастает, поскольку вводятся новые сущности и, следовательно, система не является замкнутой.

3. Пример оценки информативности контекста

Рассмотрим в качестве примера контекст, представленный в онтологии в формате OWL с помощью редактора Protégé (http://protege.stanford.edu) и посвященный рецептам пиццы (www.co-de.org/ontologies/pizza/2005/10/18/pizza.owl). Фрагмент классификации сущностей данной предметной области представлен на Рис. 4.

Сущности, содержащиеся в данной онтологии, приведены в Табл. 3. Понятие *Pizza Topping* имеет иерархическое деление вида *Pizza – PizzaTopping – CheeseTopping – CheeseVegetableTopping*.

Как показано на Рис. 4, экземпляры пиццы располагаются на разных уровнях иерархии: часть экземпляров является подмножеством класса *Pizza*, в то время как другая часть – подмножеством *Named Pizza*.

Помимо отношений иерархии классов (*AKO*) данная онтология содержит отношения *hasCountryOfOrigin, hasIngredient* (*hasBase* и *hasTopping*), *hasSpiciness*.

Рис. 5 показывает информативность фактов в контексте пиццы. В скобках указано число экземпляров класса. В частности, можно видеть, что факт “Pizza hasTopping PizzaTopping” имеет нулевую информативность, поскольку пиццы без начинки не бывает. Это означает, что факт наличия начинки у пиццы лежит в контексте уровнем выше, где пицца должна описываться как подмножество выпечных изделий. Самую большую информативность имеют факты, определяющие разновидности начинки для каждого типа. Общая информативность онтологии «Пицца» равна сумме информативностей фактов, составляющих данную онтологию:

$$I(Pizza) = 2{,}32+1{,}59+3{,}32+1+3{,}17+4{,}52+5{,}17 = 21{,}09 \text{ бит.}$$

Это означает, что каждый экземпляр пиццы, скажем, «Нью-Йоркер», несет в себе 21,07 бит информации. Поскольку в базе знаний имеется 33 экземпляра пиццы, то общее количество информации в данной онтологии

$$I(OntoPizza) = 21{,}09 \cdot 33 = 695{,}97 \text{ бит.}$$

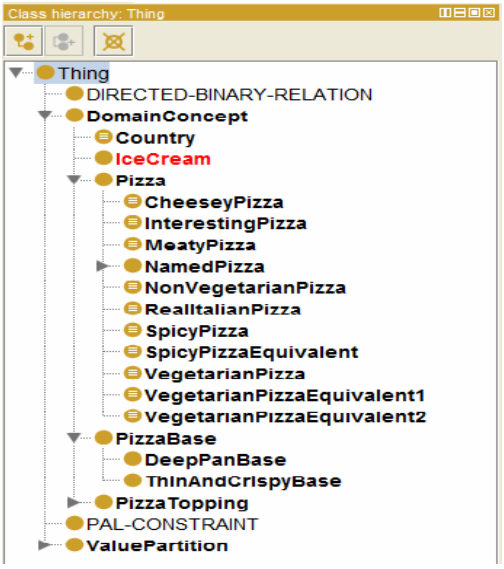

Рис.4. Онтология «Пицца»

Табл.3. Сущности в онтологии «Пицца»

Сущность	Разновидность	Число
Pizza	-	10
Named Pizza	-	23
Country	-	5
Pizza Base	-	2
Spiciness	-	3
Pizza Topping		36
-	*Cheese Topping*	6
-	*Fish Topping*	3
-	*Fruit Topping*	1
-	*Meat Topping*	4
-	*Nut Topping*	1
-	*Sauce Topping*	1
-	*Spicy Topping*	1
-	*Vegetable Topping*	18
-	*Vegetarian Topping*	1

Таким образом, чтобы оценить информативность конкретной предметной области, необходимо определить объем словаря, а также его подмножеств, встречающихся в фактах, составляющих базу знаний. Очевидно, что вычислить информативность можно только для полностью формализованной предметной области, поэтому оценить в битах информативность текстовых документов не представляется возможным, хотя по формуле (3) можно определить информативность всех понятий, составляющих текст. Однако, очевидно, что суммарное количество информации в словах не равно информативности составленного из них текста.

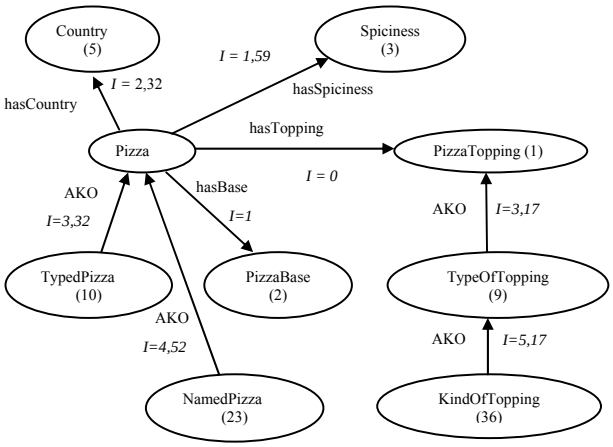

Рис. 5. Информативность фактов в онтологии «Пицца»

4. Оценка информативности баз данных

Базы данных представляют собой пример формализованных знаний, к которым также может быть применен предлагаемый метод. Обычно объем баз данных измеряется размером занимаемой памяти, но такая оценка далеко не всегда отражает количество содержащейся в них информации. Между тем, располагать такой оценкой полезно для оценки качества проектирования баз данных, сравнения разных СУБД и других задач управления знаниями.

Рассмотрим процесс измерения количества информации на очень простой базе данных, содержащей сведения об автомобилях в контексте автосалона. Для каждого автомобиля имеются данные о производителе, типе кузова (седан, купе, универсал), двигателе (бензин или дизель) и трансмиссии (механическая или автоматическая). Базы данных имеют изначально ограниченный контекст, и поэтому вычисление мощности словарей не представляет проблемы. Инфологическая модель базы «Car» приведена на Рис. 6. Задача проектирования базы данных для такой простой предметной области заключается в том, что универсальное отношение (плоская таблица) разбивается на более мелкие для устранения избыточности, потенциальной противоречивости, аномалий включения и удаления [10].

Такое разбиение автоматически дополняет текстовые описания сущностей их численными идентификаторами, что позволяет устранить дублирование текстов в случае повторения в разных строках таблиц. Для наших целей эти идентификаторы позволяют легко вычислить информативность понятий, а информативность сложного понятия определяется как сумма информативностей понятий, входящих в данное понятие. Например, понятие Country имеет 5 значений, и количество информации, содержащееся в таблице *I(Country)*= 3,17 бит. Таблица Brand имеет 4 строки, но включает в себя код страны (CountryID), имеющий информативность 3,17 бит. Таким образом, каждая строка данной таблицы имеет информативность 4·3,17=12,68 бит, а для идентификации одной

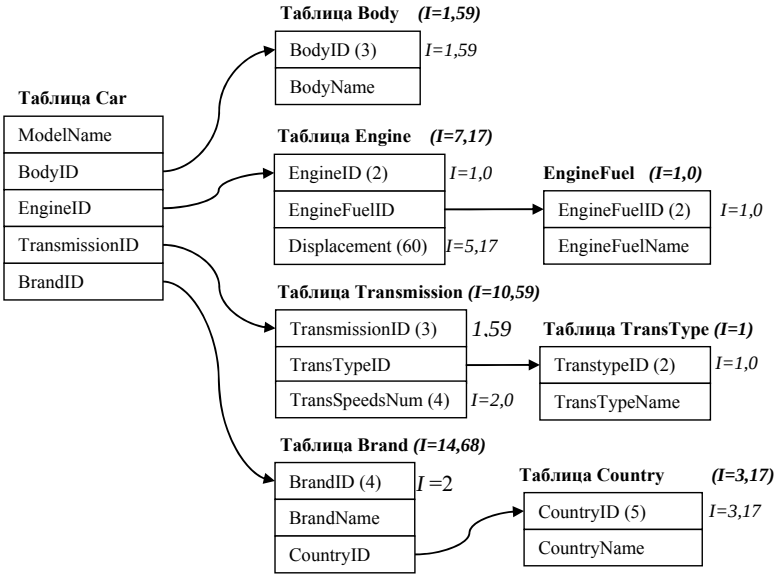

Рис. 6. Инфологическая модель базы данных «Car»

из четырех строк требуется 2 бита. Следовательно, информативность таблицы *I(Brand)*= 12,68+2=14,68 бит. Аналогично, таблица типов трансмиссий TransType, состоящая из двух строк, имеет информативность 1 бит, а каждая строка таблицы Transmission, включающая в себя ссылку на тип трансмиссии TransTypeID и число ступеней трансмиссии TransSpeedsNum=4 (*I=2*), имеет информативность 3 бита, а каждая из трех строк таблицы идентифицируется $\log_2(3)$=1,59 битами. Таким образом, информативность таблицы *I(Transmission)*=3·3+1,59=10,59 бит.

Информативность каждой строки таблицы Car складывается из информативностей ссылок на таблицы *Body, Engine, Transmission* и *Brand*. Каждая из этих информативностей определяется только числом строк в соответствующей таблице:

$$I(CarLine) = I(Body) + I(Engine) + I(Transmission) + I(Brand) = 1{,}59+1+1{,}59+2 = 6{,}18 \text{ бит.}$$

Информативность таблицы *Car,* состоящей из *N* строк, определяется как сумма информативностей всех строк плюс информативность выбора строки: $I(Car) = N \cdot I(CarLine) + \log_2 N$. Если в базе данных имеется 8 автомобилей, $I(Car) = 6{,}18 \cdot 8 + \log_2(8) = 52{,}44$ бит. Общая информативность базы данных составит: 52,44+1,59+7,17+10,59+14,68+1+1+3,17=91,64 бит.

Обобщая данные рассуждения, можно определить следующую формулу вычисления количества информации в базе данных *db*:

$$I(db) = \sum_i I(T_i), \quad I(T_i) = n_i \sum_j I(F_j) + \log_2 n_i,$$

где n_i – число строк *i*-й в таблице базы данных T_i, $I(F_j)$ – количество информации в *j*-м поле таблицы T_i.

Заключение

Предложенный подход базируется на очень простых соображениях, позволяющих подвергнуть декомпозиции и сделать решаемой задачу оценки количества информации, содержащейся в формализованных источниках – базах данных и знаний. Полезность такой оценки обусловлена ее энтропийным характером и может служить критерием качества проектирования баз данных и знаний, а при условии создания тезаурусов, как например, в работе [7], - для анализа информативности текстовой документации. Растущий интерес к формализации Интернет-ресурсов в стандарте Семантической паутины и идея создания на этой платформе глобального искусственного интеллекта [11] также могут потребовать оценок количества информации в создаваемых базах знаний.

Литература

1. Turing Alan. Computing machinery and intelligence. // Mind. -October, 1950, -pp. 433-460.
2. Шумилов В. Н. Информативная емкость мозга. //http://www.scorcher.ru/theory_publisher/show_art.php?id=26&readonly =1.
3. Богданов И. В. Учебная информация и единицы ее измерения. // Труды СГУ. Гуманитарные науки, -2002, -с. 44.
4. Shannon Claude. A Mathematical Theory of Communication. // Bell System Technical Journal. -July, October 1948, -Vol. 27, -pp. 379-432, 623-656.
5. Бессмертный И. А. Оценка количества информации в базах знаний. // Научно-технический вестник СПбГУ ИТМО, -№2, -2011г. –с.146-149.
6. Невзорова О.А., Федунов Б.Е. Система анализа технических текстов "ЛоТА": основные концепции и проектные решения. // Изв. РАН. ТиСУ. – 2001. – № 3.– С. 138-149.
7. Добров Б.В, Лукашевич Н.В., Невзорова О.А., Федунов Б.Е.Автоматизированное проектирование прикладной онтологии: методы, средства и задачи применения. Изв.РАН. ТиСУ. №.2, 2004.
8. Петровский В.А**.** Общая персонология: Наука личности. // http://petrowskiy.ru/publish/personalogia.html.
9. Валгина Н. С. Теория текста: Учебное пособие. // - М. : Мир книги, 1998. – 210с. http://www.hi-edu.ru/e-books/xbook029/01/about.htm.
10. Кириллов В.В., Громов Г.Ю. Введение в реляционные базы данных. // – СПб.,:БХВ-Петебрург, 2009. – 464 с., ил. +CD ROM.
11. Бессмертный И.А. Семантическая паутина и искусственный интеллект // Научно-технический вестник СПбГУИТМО. - Санкт-Петербург: СПбГУИТМО, 2009. – Т. 64, вып. 6. – С. 77-83. – 122 с.

Бессмертный Игорь Александрович. Доцент Санкт-Петербургского национального исследовательского университета информационных технологий, механики и оптики. Окончил Рижский институт инженеров гражданской авиации в 1976 году. Кандидат технических наук. Автор 40 печатных работ. Область научных интересов: продукционные системы, системы интеллектуального управления, язык Prolog. E-mail: igor_bessmertny@hotmail.com