УДК 004.89:004.657

ПОСПЕШНЫЙ А.С.

МЕТОД ИНФОРМАЦИОННОГО ДОПОЛНЕНИЯ НА ОСНОВЕ ОНТОЛОГИЙ ДЛЯ ПОВЫШЕНИЯ ЭФФЕКТИВНОСТИ ИНФОРМАЦИОННЫХ СИСТЕМ

В данной статье рассмотрено формальное описание информационной системы и предложено способ оценки количества информации, которое хранится и обрабатывается такой системой. Рассмотрено понятие дополняемой информационной системы, способной извлекать дополнительную информацию, основываясь на семантике данных. Доказано эффективность онтологий в качестве средства информационного дополнения. Рассмотрен способ существенного понижения сложности запросов к информационной системе и усовершенствования взаимодействия пользователя с системой. Предложено производить логическую валидацию входящих запросов, с целью дополнительного повышения эффективности информационных систем.

Ключевые слова: информационные системы, онтология, количество информации, информационный поиск.

The article presents a formal description of an information system and proposes a method for quantifying an amount of information that it stores and processes. A notion of complemented information systems is being considered, in which additional implicit information is extracted from the stored data based on its semantics. We discuss an application of ontologies to the task at hand, proving their effectiveness as an information complementing tool. We propose a new method to substantially reduce the complexity of quires to the information system and considerably simplify user's interaction with it. A suggestion is made to perform a preventive logical validation of incoming queries to the information system to further increase their effectiveness.

Keywords: information systems, ontology, amount of information, information retrieval.

Введение

В условиях современного динамического развития общества и усложнения технической и социальной инфраструктуры, информация становится таким же стратегическим ресурсом, как и традиционные материальные, энергетические и человеческие ресурсы.

Осознание мировым сообществом роли информации, как стратегического ресурса, стимулировало разработку новых информационных технологий для получения и обработки больших объемов информации, ее хранения и предоставления пользователям. Таким образом, любые попытки повысить эффективность работы таких систем будут иметь существенные положительные последствия для всех сфер нашей жизни.

В данной работе сделан особый упор на онтологии, как на средство, способное повысить эффективность поиска и хранения информации информационными системами. Предложенный в исследовании подход может играть значительную роль в расширении возможностей информационных систем и улучшении их эксплуатационных характеристик.

Также, ввиду огромного количества разнообразных информационных систем, которые уже созданы либо находятся на этапе разработки и внедрения, достаточно важно иметь средства оценки качества их работы в зависимости от реализуемых задач. В исследовании мы предлагаем метод оценки эффективности информационных систем основанный на методах теории информации и кодирования.

Цель работы

В данной работе преследуются цели, ориентированные на повышение эффективности работы информационных систем. Под информационной системой мы будем подразумевать программный комплекс, состоящий из хранилища данных, модуля обработки запросов и логического модуля. Основная функция данных систем заключается в обработке пользовательских запросов, обращении к внутреннему хранилищу данных и извлечении из него искомой информации в удобном для пользователя виде.

Первоочередной целью является поиск показателей и критериев эффективности работы информационных систем, а также количественная оценка сложности поисковых запросов к таким системам. Второй целью является обоснование целесообразности использования онтологий и метода информационного дополнения для повышения эффективности информационных систем. Третей целью этой работы являет-

ся усовершенствование формального процесса составления и обработки запроса в информационной системе, а также выполнение предварительного анализа его корректности.

Анализ существующих решений

Для оценки качества работы информационных систем в литературе наиболее часто применяются такие оценочные характеристики, как точность (precision) и полнота (recall). Точность определяется отношением количества релевантных документов, найденных информационной системой по произвольному запросу, к общему количеству найденных документов. Полнота определяет отношение числа найденных релевантных документов, к общему числу хранимых релевантных документов. Однако эти критерии целесообразно применять лишь в тех информационных системах, которые оперируют текстом, документами или другой слабо структурированной и неформализированной информацией. Информационные системы, рассматриваемые в этой работе, не относятся к подобным системам и по классу реализуемых задач более близки к реляционным базам данных, в которых данные подчиняются чётко определённой информационной схеме. В таких системах точность и полнота, как правило, составляет 100%, что делает бессмысленным использование этих характеристик для оценки качества их работы. В свою очередь, оценка эффективности подобных информационных систем производится чаще всего с учётом их эксплуатационных характеристик. В частности, оценочными критериями выступают время обработки запроса, количество используемой памяти, степень загруженности процессора(ов) и т.д. Несмотря на неоспоримую важность этих характеристик, они всё же не отображают действительной эффективности хранения и обработки информации информационной системой.

Применение семантических технологий постепенно становится обязательным моментом при проектировании и разработке информационных систем. Так, компания Oracle – лидер на рынке реляционных баз данных, заявила о поддержке онтологий в своей объектно-реляционной системе управления базами данных Oracle 11g [1]. На самом деле, эта RDBMS производит проекцию вводимых в виде RDF триплетов[2] данных на свою реляционную модель и применяет к ним множество порождающих правил в соответствии со спецификациями RDF(S) и OWL2-RL. При этом используется довольно слабая if-семантика, выражаемая в виде IF…THEN правил. Запрос к данным осуществляется через обёрнутый в SQL язык запросов SPARQL[3] или посредством специальных операторов ONT_RELATED и ONT_EXPAND. Используемые в Oracle 11g запросы чрезвычайно сложны и громоздки, а используемая ослабленная семантика и RDF-ориентированная модель данных не согласуется с представлением автора об онтологически-дополняемых системах.

Совершенно другой подход был рассмотрен в работе [2], где рассматривается возможность трансляции данных из реляционной базы данных в их онтологическое представление. Таким образом, появляется возможность интеграции RDBMS систем в канву Семантической паутины. Однако данный подход не учитывает анализ транслируемых данных с учётом их семантики и не предусматривает ни онтологического дополнения, ни последующего упрощения запросов к предоставляемой информации.

Похожий подход был предложен в работе [3], но в отличие от предыдущей работы, предлагалось использовать онтологии для доступа к «нижележащей» информационной системе. При этом рассматривается новая специальная дескрипционная логика DL-Lite$_A$ и механизмы генерации SQL-запросов на основе онтологий. К сожалению, предложенный авторами метод не универсален и не позволяет достичь поставленных в этой работе целей.

Понятие информационной системы

Вначале, рассмотрим понятие *информационной системы (ИС)*. В основе информационной системы лежит множество информационных *объектов* (X), природа которых определяет все её ключевые характеристики. Так, в качестве информационных объектов ИС могут быть пользователи, список товаров или услуг, данные полученные вследствие эксперимента, наблюдаемые процессы, события, список книг или запчастей и т.д. Объекты, в свою очередь, классифицируются посредством множества *атрибутов* (A). Например, «дата создания», «имя», «тип», «стоимость», «высота», «шири-

[2] http://www.w3.org/TR/rdf-primer/

[3] http://www.w3.org/TR/rdf-sparql-query/

на» и т.д. – всё это возможные атрибуты объекта информационной системы. Для каждого атрибута $a \in A$ определено множество допустимых значений V_a известное как *домен атрибута* a. Так, для атрибута «пол» $V_a = \{\text{мужчина}, \text{женщина}\}$, атрибут «возраст» может обладать доменом натуральных чисел от 0 до 150: $V_a = \{x \in \mathbb{N} \mid 0 \leq x \leq 150\}$, а доменом атрибута «масса» может быть множество рациональных чисел: $V_a = \mathbb{Q}$. Разумеется, атрибуты могут разделять общее пространство значений и, например, атрибуты «высота» и «длина» будут относиться к одному и тому же домену V_a.

Для того чтобы иметь возможность ассоциировать некую информацию об объекте информационной системы необходима *функция утверждения* ρ которая будет отображать множество $X \times A$ во множество $V = \bigcup_{a \in A} V_a$ таким образом, что $\rho(x, a) \in V_a$ для всех $x \in X$ и $a \in A$. Фактически посредством функции ρ мы задаём каждому объекту его описание в виде значений его атрибутов.

Если ρ – частично определённая функция, то определим информационную систему, как *неполную*, если же ρ – всюду определённая функция, то такая ИС определяется, как *полная* информационная система.

Таким образом, мы можем определить информационную систему S как

$$S = \langle X, A, V, \rho \rangle, \quad (1)$$

где X – множество объектов информационной системы, A – множество атрибутов, $V = \bigcup_{a \in A} V_a$ где V_a – множество значений атрибута $a \in A$ при условии что $|V_a| > 1$, и ρ – функция утверждения $X \times A \xrightarrow{\rho} V$.

Пример 1

Рассмотрим пример информационной системы:

$X = \{x_1, x_2, x_3, x_4, x_5\}$,

$A = \{\text{имя}, \text{пол}, \text{возраст}, \text{родители}\}$,

$V = \{V_{\text{имя}} \cup V_{\text{пол}} \cup V_{\text{возраст}} \cup V_{\text{родители}}\}$,

где $V_{\text{имя}} = \Sigma^*$, $V_{\text{пол}} = \{\text{мужчина}, \text{женщина}\}$, $V_{\text{возраст}} = \mathbb{N}$, $V_{\text{родители}} = X$.

Здесь и дальше под выражением Σ^* будет подразумеваться замыкание Клини – множество всех строк произвольной длины в соответствии с алфавитом Σ: $\Sigma^* = \bigcup_{n \in \mathbb{N}} \Sigma^n$

Функцию ρ в нашем примере выразим в виде таблицы:

Таблица 7. Функция ρ

X	имя	пол	возраст	родители
x_1	Ковальчук Тарас	мужчина	45	$\{x_5\}$
x_2	Ковальчук Ярослав	мужчина	12	$\{x_1, x_3\}$
x_3	Яремчук Инна	женщина	39	-
x_4	Ковальчук Юля	женщина	4	$\{x_1, x_3\}$
x_5	Ковальчук Андрей	мужчина	67	-

Оценка количества информации в ИС

Введём понятие ϱ_x^S как *совокупного описания объекта* x в информационной системе S (или ϱ_x, если информационная система S определена контекстом). Таким образом, ϱ_x^S для любого объекта $x \in X$ будет включать значения всех атрибутов этого объекта:

$$\varrho_x^S = \bigcup_{a \in A} \rho(x, a) \quad (2)$$

Так, для информационной системы S, представленной ранее в примере 1 совокупное описание объекта x_2 будет равно $\varrho_{x_2}^S = \{(\text{имя: Ковальчук Ярослав}), (\text{пол: мужчина}), (\text{возраст: 12}), (\text{родители: } \{x_1, x_2\})\}$.

Дополнительно введём понятие *совокупной хранимой информации* $\mathfrak{v}^S$ информационной системы S, которая будет определять множество всех утверждение в этой ИС:

$$\mathfrak{v}^S = \bigcup_{x \in X} \varrho_x^S \quad (3)$$

Теперь попробуем количественно оценить тот объём информации, которым располагает произвольная информационная система. Для этого обратимся к теории информации [4].

Основоположник теории информации Клод Шеннон определил *информацию* как снятую неопределенность. Задача, которая решается в ходе снятия неопределенности – уменьшение количества рассматриваемых вариантов, и в итоге выбор одного соответствующего ситуации варианта из числа возможных. В свою очередь *энтропия* – это величина, которая характеризует количество неопределенности и непредсказуемости.

Количество информации I и энтропия H характеризуют одну и ту же ситуацию, но с качественно противоположенных сторон. Можно сказать, что I – это количество информации,

которое требуется для снятия неопределенности H.

Для случайной величины X, имеющей конечное число значений так, что $P(x_i) = p_i$, $p_i \geq 0$, $i = 1,2,\dots,n$ и $\sum_{i=1}^{n} p_i = 1$ расчёт энтропии производится по формуле Шенона:

$$H = -\sum_{i=0}^{n} p_i \log_2 p_i = \sum_{i=0}^{n} p_i \log_2 \frac{1}{p_i}$$

В частном случае, когда все варианты равновероятны, остается зависимость только от количества рассматриваемых вариантов и количество информации I в этом случае определяется как двоичный логарифм числа состояний:

$$I = \log_2 n \qquad (4)$$

Опираясь на вышеупомянутые принципы, попробуем оценить количество информации, хранимое в информационной системе S:

$$I(S) = I(\mathfrak{s}^S),$$
$$I(\mathfrak{s}^S) = \sum_{x \in X} I(\varrho_x^S),$$
$$I(\varrho_x^S) = \sum_{a \in A} \log_2 |V_a|, \text{ для } \forall \rho(x,a) \neq \emptyset. \qquad (5)$$

Таким образом, можно записать что:

$$I(S) = \sum_{x \in X} \sum_{a \in A} \begin{cases} \log_2 |V_a|, & \rho(x,a) \neq \emptyset \\ 0, & \rho(x,a) = \emptyset \end{cases} \qquad (6)$$

Пример 2

Произведём оценку количества информации в информационной системе S, приведённой в *примере 1*:

$$I(S) = I(\mathfrak{s}^S) = \sum_{i=1}^{5} I(\varrho_x^S),$$

где

$$I\left(\varrho_{x_1}^S\right) = \log_2 |\Sigma^{15}| + \log_2 |V_{\text{пол}}| + \log_2 |\mathbb{N}| + \log_2 \left|V_{\text{родители}}\right| \approx$$
$$\approx \log_2 32^{15} + \log_2 2 + \log_2 2^{32} + \log_2 5 \approx \mathbf{110} \text{ бит},$$
$$I\left(\varrho_{x_2}^S\right) = \log_2 |\Sigma^{17}| + \log_2 |V_{\text{пол}}| + \log_2 |\mathbb{N}| + 2\log_2 \left|V_{\text{родители}}\right| \approx$$
$$\approx \log_2 32^{17} + \log_2 2 + \log_2 2^{32} + 2\log_2 5 \approx \mathbf{122} \text{ бит},$$
$$I\left(\varrho_{x_3}^S\right) = \log_2 |\Sigma^{12}| + \log_2 |V_{\text{пол}}| + \log_2 |\mathbb{N}| \approx$$
$$\approx \log_2 32^{12} + \log_2 2 + \log_2 2^{32} \approx \mathbf{93} \text{ бит},$$
$$I\left(\varrho_{x_4}^S\right) \approx \mathbf{103} \text{ бит}, \; I\left(\varrho_{x_5}^S\right) \approx \mathbf{113} \text{ бит}.$$

Таким образом, $I(S) \approx \mathbf{541}$ бит**.**

Хочется заметить, что в примере выражение $\log_2 |\mathbb{N}|$ было преобразовано в $\log_2 2^{32}$ так как, несмотря на бесконечное количество натуральных чисел, в цифровой технике значения численных доменов чаще всего выражены в виде величин с фиксированной разрядностью (short, int32, long64, и т.д.).

Информационное дополнение в ИС

Рассмотрим понятие *дополняемой информационной системы* S^+ как

$$S^+ = \langle X, A, V, \rho, \omega \rangle, \qquad (7)$$

где X – множество объектов информационной системы, A – множество атрибутов, $V = \bigcup_{a \in A} V_a$ где V_a – множество значений атрибута $a \in A$, ρ – функция определения $\rho : X \times A \to V$ и ω – функция дополнения $\omega : X \times A \to V$.

Назначение функции ω заключается в дополнительном определении атрибутов объекта x, исходя из множества уже определённых значений, заданных посредством функции ρ.

Таким образом, совокупное описание объекта $x \in X$ в дополняемой информационной системе S^+ будет равно:

$$\varrho_x^{S^+} = \bigcup_{a \in A} \rho(x,a) \cup \omega(x,a) \qquad (8)$$

Для реализации функции ω можно использовать множество методов, однако в этой статье для этих целей будет рассмотрен онтологический подход.

Онтологии – это формализация некоторой области знаний в виде множества понятий и отношений между ними. Онтологии нашли широкое применение в информатике как средство для представления знаний и обработки данных с учётом их семантики. Онтологии широко применяются для решения задач в области искусственного интеллекта, Семантического Веб, программной инженерии, биоинформатики, бизнес-процессов, библиотечного дела и множества других областей науки.

Язык онтологий определяет синтаксис и семантику онтологии таким образом, что бы она могла быть однозначно воспринята и интерпретирована компьютером. Самым развитым язы-

ком онтологий на сегодняшний день считается язык OWL, стандартизированный Всемирным Консорциумом паутины (W3C).[4] *Дескрипционная логика (DL)* [5], в свою очередь, обеспечивает подробно изученную логическую основу для онтологических языков. На сегодняшний день существует множество разнообразных DL языков, например, $\mathcal{EL}^{++}$, $\mathcal{FL}_0$, $\mathcal{ALC}$, $\mathcal{SHIF}$, $\mathcal{SHOIQ}$ и т.д., различающиеся по степени экспрессивности и вычислительной сложности.

Таким образом, рассмотрим *онтологически дополняемую информационную систему* $S_{\mathcal{O}}^{+}$ как расширение определения (8) вида $S_{\mathcal{O}}^{+} = \langle X, A, V, \rho, \omega, \mathcal{O} \rangle$, где $\mathcal{O}$ – онтология информационной системы. В такой ИС знания, представленные в виде онтологии, позволяют совершать логический анализ над информационными объектами $x \in X$ и производить доопределение их атрибутов, тем самым увеличивая количество информации в информационной системе. То есть для всех случаев будет верным утверждение $I\left(\mathfrak{s}^{S_{\mathcal{O}}^{+}}\right) \geq I(\mathfrak{s}^{S})$.

Пример 3

Рассмотрим пример онтологически дополняемой информационной системы $S_{\mathcal{O}}^{+}$ для мониторинга задач в распределённой вычислительной системе:

$$X = \{x_1, x_2, x_3, x_4, x_5\},$$

$$A = \begin{Bmatrix} \text{IP-адрес, задач в очереди,} \\ \text{запущенно задач, статус} \end{Bmatrix},$$

$$V = \begin{Bmatrix} \mathrm{V}_{\text{IP-адрес}} \cup \mathrm{V}_{\text{задач в очереди}} \cup \\ \mathrm{V}_{\text{запущенно задач}} \cup \mathrm{V}_{\text{статус}} \end{Bmatrix},$$

$$\mathrm{V}_{\text{IP-адрес}} = \{x \in \mathbb{N} \mid 0 \leq x \leq 2^{32} - 1\},$$

$$\mathrm{V}_{\text{задач в очереди}} = \mathbb{N}, \mathrm{V}_{\text{запущенно задач}} = \mathbb{N},$$

$$\mathrm{V}_{\text{статус}} = \{\text{простаивающий, занятый, недоступный}\}.$$

Функцию ρ выразим в виде таблицы:

Таблица 2. Функция ρ

X	*IP-адрес*	*задач в очереди*	*запущено задач*	*статус*
x_1	192.168.0.1	1		занятый
x_2	192.168.0.2	30	3	
x_3	192.168.0.3			проста – ивающий
x_4	73.24.101.11	1	1	
x_5	73.24.101.10	3	0	

[4] http://www.w3.org/TR/owl2-overview/

Допустим, информационная система была дополнена следующей онтологией:

$$\mathcal{O} = \begin{cases} \exists \text{статус. простаивающий} \equiv \exists \text{задач в очереди.}(=,0) \\ \quad \sqcap \exists \text{запущенно задач.}(=,0) \\ \exists \text{статус. занятый} \equiv \exists \text{запущенно задач.}(\geq,1) \\ \exists \text{статус. недоступный} \equiv \exists \text{задач в очереди.}(>,0) \\ \quad \sqcap \exists \text{запущенно задач.}(=,0) \end{cases}$$

В таком случае доопределённая информационная система будет иметь следующий вид:

Таблица 3. Функция ρ + ω

X	*IP-адрес*	*задач в очереди*	*запущено задач*	*статус*
x_1	192.168.0.1	1	≥ 1	занятый
x_2	192.168.0.2	30	3	занятый
x_3	192.168.0.3	0	0	проста – ивающий
x_4	73.24.101.11	1	1	занятый
x_5	73.24.101.10	3	0	недоступный

Воспользовавшись выражениями (5) и (8) не трудно рассчитать, что изначально информационная система содержала около **387 бит** информации. Поле онтологического дополнения количество информации в ИС составило **456 бит.** При этом для объекта x_1 была уменьшена неопределённость (энтропия) относительно параметра *«запущено задач»*.

Представленный пример демонстрирует лишь малую часть потенциала, заложенного в онтологическом дополнении информационных систем. Онтологии могут содержать тысячи утверждений и правил, отображая глубинные взаимосвязи между информационными объектами и их атрибутами. При этом накладные расходы, связанные с логическим выводом над онтологиями, зависят как от её размера, так и от языка дескрипционная логики, на которой она основана. Так для онтологий на базе языка $\mathcal{EL}^{++}$ характерна полиномиальная зависимость между размером онтологии и временем её обработки, что позволяет эффективно применять метод онтологического дополнения с учётом большого объёма данных [6].

Эффективность запроса в ИС

Основная функция рассматриваемого класса информационных систем – предоставление её пользователям запрашиваемой информации. Критерием эффективности подобного класса

систем априори можно считать степень удовлетворения информационных потребностей пользователя.

Ранее было показано, что применение онтологий для информационного дополнения ИС позволяет увеличить объём извлекаемой информации из изначально вводимых данных. Таким образом, только за счёт этого, пользователь дополняемой информационной системы может получать больше релевантной информации, чем пользователь традиционной ИС на основе того же набора исходных данных. Однако применение онтологий открывает возможность значительно усовершенствовать непосредственно сам процесс запроса к ИС, упростить его и сделать более удобным для пользователя.

Определим L_S как *язык запросов* к информационной системе S, который будет использоваться для извлечения информации, хранимой в этой ИС. Запрос к информационной системе может быть выражен в виде некого *термина* в рамках этой ИС, либо в виде *формулы*. В первом случае термины представляют собой подмножество всех информационных объектов ИС с указанными значениями атрибутов. В случае формул, в качестве запроса выступает логическое утверждение, результат которого может быть разрешён как «истинно» либо «ложно». В представленной работе внимание будет сфокусировано на запросах к ИС заданных в виде терминов, но все изложенные рассуждения будут в равной степени верны и для запросов заданных в виде формул.

Таким образом, будем считать, что запрос Q_S к информационной системе S, составленный с помощью языка L_S и множества его терминов T_s выполняется поисковой функцией $\varphi_s : T_s \to X$. При этом термином $t \in T_S$ может быть:

1. Универсальные термины $\bot$ и $\top$, обозначающие «ничего» и «всё» соответственно.
2. Атомарные *дескрипторы* информационной системы S, представляющие собой элемент множества $\{a\} \times V_a$, где $a \in A$.
3. Комплексный термин созданный с помощью булевых операторов $\neg$, $\wedge$, $\vee$.

Семантика поисковой функции φ_s при этом будет выглядеть следующим образом:

$$\begin{cases} \varphi(\bot) = \emptyset,\ \varphi(\top) = X \\ \varphi(\neg t) = X - \varphi(t) \\ \varphi(t \wedge s) = \varphi(t) \cap \varphi(s) \\ \varphi(t \vee s) = \varphi(t) \cup \varphi(s) \\ \varphi(a, v) = \{x \in X : \rho(x, a) = v\} \end{cases} \quad (9)$$

Для информационной системы, представленной в *Примере 3,* поисковым запросом может быть следующее выражение: (*задач в очереди*: > 20) $\wedge$ (*запущено задач*: ≤ 5). В этом случае, ответом будет $\{x_2\}$.

Воспользовавшись представленным ранее методом оценки количества информации в ИС, введём понятие *сложности запроса* к информационной системе S как количество информации, которое содержится в этом запросе:

$$I(Q_S) = \sum_{t \in Q_S} \log_2 |V_t|, \quad (10)$$

где V_t – множество значений термина $t \in T_S$. Вычисления в (10) согласуются и с интуитивной оценкой сложности запроса к ИС – чем больше атрибутов задействовано в запросе и чем больше область их возможных значений, тем сложнее его составить.

Теперь рассмотрим, каким образом онтологии могут существенно понизить сложность запроса к информационной системе.

Пример 4

Обратимся к онтологически дополняемой информационной системе из *Примера 3*. Составим поисковый запрос к ИС, который вернёт список всех простаивающих или вскоре освобождающихся вычислительных ресурсов в локальной сети:

$$Q_S = (\text{задач в очереди}: 0) \wedge (\text{запущено задач}: \leq 1) \wedge (\text{IP-адрес} : \geq 192.168.0.1) \wedge (\text{IP-адрес} : \leq 192.168.0.255).$$

Расширим онтологию информационной системы S_O^+ следующими выражениями:

$$\mathcal{O} \cup \begin{cases} \exists \text{IP-адрес}.(\geq 192.168.0.1, \leq 192.168.0.255) \sqsubseteq \text{Локальный ресурс} \\ \exists \text{статус}.\text{простаивающий} \sqcup (\exists \text{запущенно задач}.(=,1) \sqcap \exists \text{задач в очереди}.(=,0)) \sqsubseteq \text{Незагруженный} \end{cases}$$

При этом к множеству терминов T_S языка запросов L_S из онтологии будет добавлено два новых термина: *Локальный ресурс* и *Незагруженный*. С учётом новых терминов, перепишем поисковый запрос следующим образом:

$$Q_{S_0^+} = \text{Локальный ресурс} \wedge \text{Незагруженный}$$

Итого, нам удалось весомо упростить запрос к ИС, сохранив при этом показатели точности и полноты. Воспользовавшись формулой (10) можно посчитать, что количество информации в изначальном запросе составило **128 бит**. Во втором случае количество информации составило всего **2 бита**, так как определённые в онтологии термины не имеют области допустимых значений и только выражают определённое понятие. Естественно, запрос к ИС может и, скорее всего, будет содержать как термины, определённые в онтологии, так и обычные термины, основанные на дескрипторах.

Может сложиться ошибочное представление, что онтология повысила сложность информационной системы, так как на её создание уйдёт значительно больше времени, чем на составление комплексного запроса. Однако, онтологии составляются с целью описания целой области знаний, предоставляя широкий набор часто употребляемых терминов для многократного использования. Можно говорить о том, что термины, доступные через онтологию, будут основой для поисковых запросов к ИС, дополняемые уточняющими дескрипторами. Это во многом соответствует тому, как люди «запрашивают» информацию друг у друга. Например, онтологически дополняемая ИС может произвести поиск по запросу приблизительно следующего вида: «Дети старше 7 лет», «Многоядерные вычислительные узлы с поддержкой MPI», «Высокопроизводительные сервера фирмы Intel» и т.д. К тому же стоит сказать, что пользователь не всегда имеет достаточную квалификацию, мотивацию или время, необходимое для создания подобных запросов к ИС с помощью перечисления всех атрибутов искомого объекта со всеми их взаимосвязями. В то же время изучение и использование онтологии, зачастую дополнительно аннотированной пользовательской информацией и комментариями к предоставляемым терминам, значительно более быстрый и менее трудоёмкий процесс.

Логическая проверка запросов к ИС

Ещё одним преимуществом использования онтологий в информационных системах является возможность проведения предварительной проверки поискового запроса на предмет логических несоответствий. В том случае, если входящий запрос к ИС исключает всякую возможность того, что запрашиваемые данные могут существовать вообще, информационная система может тут же вернуть пустое множество в качестве ответа и сэкономить время на выполнении бессмысленного запроса.

Например, для информационной системы, приведённой в *Примере 4,* следующие запросы лишены всякого смысла:

$$Q_\perp = (\text{задач в очереди: -1})$$
$$Q_\perp = (\text{запущено задач: } \geq 1) \wedge (\text{статус: простаивающий})$$
$$Q_\perp = \text{Локальный ресурс} \wedge (\text{IP-адрес: 91.20.3.44})$$

Современные средства логического анализа онтологий могут помочь не только в выявлении логических коллизий, но и предоставляют подробное описание его причины, что будет немаловажным для пользователя, составившего такой ошибочный запрос.

Дополняющие онтологии

Онтологии – основа предложенной идеи информационного дополнения. Соответственно вопрос возникновения и развития онтологий является обязательным к освещению в данной работе.

По мнению автора, онтологии для предложенных в работе ИС могут быть определены тремя основными путями:

1. На этапе проектирования информационной системы её разработчиками. При этом производится анализ природы хранимой в ИС информации, аналитическая обработка семантики данных и формализация полученных результатов в виде онтологии. Именно полученная таким образом онтология будет играть основную роль в доопределении информации в ИС.
2. На этапе эксплуатация её пользователями. Пользовательские онтологии могут расширять онтологию ИС (см. п.1), дополняя и расширяя её. Главная задача созданных на этом этапе онтологий – расширить множество терминов языка запросов L_S.
3. Автоматизированными средствами на основе Data Mining анализа. На данный момент это наименее изученный подход для создания онтологий, который предусматривает сбор и анализ информации о хранимых в ИС данных и выполняемых запросах. На данный момент подобные инструменты не получили должного

развития и создаваемые таким образом онтологии, скорее всего, будут значительно уступать в качестве онтологиям, созданным человеком вручную. Однако с развитием подобных технологий этот вариант будет становиться более привлекательным для использования.

Все три подхода могут использоваться одновременно, не исключая, а дополняя друг друга.

Выводы

В данной работе представлено метод онтологического дополнения информационных систем, который заключается в применении онтологий для дополнения информации, хранимой в ИС. При этом знания о семантике данных представленные в онтологии помогают доопределить или уточнить значения атрибутов информационных объектов ИС, что позволяет увеличить объём хранимой в ИС информации при одинаковом количестве входных данных.

Доказано преимущество использования онтологий для выполнения запросов к информационной системе, так как их применение позволяет существенно понизить сложность используемых поисковых запросов к ИС и дополнительно позволяет совершать предварительный логический анализ их корректности.

Изложенные в статье предложения по повышению эффективности информационных систем были оценены с точки зрения теории информации и апробированы на практике в рамках проекта построения семантического информационного сервиса Грид [7].

Список литературы

1. Murray C. Oracle Database Semantic Technologies Developer's Guide, 11g Release 2 (11.2), E25609-03 // Oracle. – 2012. – Режим доступа: http://docs.oracle.com/cd/E11882_01/appdev.112/ e25609/toc.htm.
2. Sedighi S. Semantic Query in a Relational Database Using a Local Ontology Construction / S. Sedighi, R. Javidan // South African Journal of Science. – 2012. – 108(11/12). – pp. 1-10.
3. Poggi A. Linking Data to Ontologies / A. Poggi, D. Lembo, D. Calvanese, G. Giacomo, M. Lenzerini, R. Rosati // Journal on data semantics. – Springer, – 2012. – X. – pp. 133-173.
4. Shannon C.E. The Mathematical Theory of Communication. / C. Shannon, W. Weaver, R. Blahut. – Urbana: University of Illinois press, – 1949.
5. Baader F. The Description Logic Handbook: Theory, Implementation, and Applications. (2ed) / F. Baader, D. Calvanese, D. McGuinness, D. Nardi, P. Patel-Schneider. – Cambridge University Press, – 2007.
6. Поспешний А.С. Эффективный логический анализ больших онтологий за полиномиальное время // Вісник НТУУ «КПІ». Інформатика, управління та обчислювальна техніка: Зб. наук. пр. – К.: Век+, – 2012. – № 55. – С. 192-196.
7. Поспешный А.С. GRID-DL – семантический информационный сервис ГРИД / А.С. Поспешный, С.Г. Стиренко // Компьютинг. – 2011. – №3 (10). – С. 285-294.