РАЗРАБОТКА РЕКОМЕНДАТЕЛЬНОЙ СИСТЕМЫ С ЕСТЕСТВЕННО-ЯЗЫКОВЫМ ИНТЕРФЕЙСОМ НА ОСНОВЕ МАТЕМАТИЧЕСКИХ МОДЕЛЕЙ СЕМАНТИЧЕСКИХ ОБЪЕКТОВ

А.А. Правиков,
аспирант кафедры «Информационные технологии» Российского Государственного технологического университета им. К.Э. Циолковского, e-mail: alex_pravikov@mail.ru.
В.А. Фомичев,
доктор технических наук, профессор кафедры «Инновации и бизнес в сфере информационных технологий» факультета бизнес-информатики Государственного университета – Высшей школы экономики, e-mail: vfomichov@hse.ru.
Адрес: г. Москва, ул. Кирпичная, д. 33/5.

Статья описывает новый метод проектирования информационного и алгоритмического обеспечения естественно-языковых интерфейсов рекомендательных систем на основе разработки математических моделей семантических объектов. Приводятся сведения о программной реализации предложенного метода в среде PHP + SQL и результатах тестирования программы.

Ключевые слова: рекомендательная система, естественно-языковой интерфейс, теория К-представлений, СК-языки, концептуальный базис, семантическое представление текста.

1. Введение

С начала 2000-х годов в области электронной коммерции развивается направление, цель которого заключается в разработке рекомендательных систем (РекС) с естественно-языковым интерфейсом (ЕЯ-интерфейсом). Такие системы предназначены для того, чтобы помочь Интернет-пользователю принять решение о выборе товаров и/или услуг [1 – 4]. Исследование, проведенное в США, показало, что 79 процентов пользователей предпочитают взаимодействовать не с традиционной системой предлагаемых меню, а с ЕЯ-интерфейсом (в эксперименте применялся англоязычный интерфейс) [1 – 2].

Под естественным языком (ЕЯ) в теоретической и компьютерной лингвистике понимается совокупность языков, на которых говорят люди, пишутся книги и т.д., т.е. совокупность русского, англий-

ского и других языков. Важным преимуществом РекС с ЕЯ-интерфейсом является то, что пользователь не должен быть знаком со структуризацией предметной области, отраженной в базе данных о товарах РекС. Представление первоначального запроса пользователя РекС на ЕЯ позволяет быстро выделить из большого перечня товаров подмножество тех товаров, которые могут быть интересны для пользователя, и приступить к уточнению желательных характеристик товаров из этого подмножества. Исследование, проведенное в США, показало, что в случае использования РекС с ЕЯ-интерфейсом время навигации до момента выбора пользователем товара уменьшается на 33 % [2].

Область разработки ЕЯ-интерфейсов прикладных компьютерных систем относится к наукоемким направлениям техники. В проектировании технических объектов многих видов существенно используются формальные методы. Однако в настоящее время методы формализации проектирования ЕЯ-интерфейсов развиты еще недостаточно, причем это касается не только русского, но и английского языка. В частности, недостаточно исследована задача математического описания смысловой структуры ЕЯ-запросов пользователя РекС.

Между тем, проведенный анализ показал, что важную роль в разработке информационного и алгоритмического обеспечения ЕЯ-интерфейса РекС должно играть построение математической модели многообразия смысловых структур, соответствующих входным запросам пользователя РекС. Значение такой модели заключается в описании общих свойств смысловой структуры разнообразных входных запросов РекС. Подобная модель должна лежать в основе технического задания на разработку семантико-синтаксического анализатора входных запросов РекС.

В данной статье описываются теоретические основы нового метода проектирования информационного и алгоритмического обеспечения естественно-языковых интерфейсов рекомендательных систем, в основе метода лежит высказанная выше идея. Модельной предметной областью является организация взаимодействия на естественном (русском) языке с интеллектуальной базой данных автомобильного салона.

Статья посвящена, во-первых, разработке математических моделей таких объектов семантического уровня, которые существенно влияют на проектирование информационного и алгоритмического обеспечения естественно-языковых интерфейсов рекомендательных систем. Основными решенными задачами являются: разработка математической модели системы первичных единиц концептуального уровня, используемой рекомендательной системой; построение математической модели многообразия смысловых структур, соответствующих запросам пользователей на нормализованном естественном языке. Во-вторых, приводятся сведения о программной реализации предложенного метода в среде PHP + SQL и результатах тестирования программы.

2. Неформальное описание структуры и принципов обработки первоначального запроса пользователя на естественном языке

Рассмотрим особенности естественно-языковых запросов пользователей рекомендательной системы (РекС). Гибкость русского языка значительно затрудняет интерпретацию запросов. Чаще всего запросы образуются из существительных, прилагательных и предлогов. Например, такими запросами могут быть выражения «недорогие немецкие автомобили», «автомобили из Германии», «мобильные телефоны фирмы Сименс», «дорогой мобильный телефон».

В запросах часто присутствуют ограничения на числовые значения некоторых параметров, например, на цену и вес. Такие ограничения формулируются с помощью выражений «не дороже четырёхсот тысяч рублей», «не старше пяти лет», «цена от 300 до 400 т.р.» и т.п. Выражения, задающие ограничения, могут входить как в состав простых фраз, так и в состав связных текстов, или дискурсов. Для формулировки числовых ограничений могут использоваться краткие прилагательные и логическая связка «отрицание». Например, на вход РекС может поступить запрос «немецкие машины не дороже 20 т. евро».

Пользователей часто интересует не просто какой-либо предмет, а значения ряда критериев, по которым можно подобрать товар. В качестве параметров отбора могут выступать цена, цвет, вес, страна-производитель, марка и так далее. Например, запрос «недорогой ноутбук фирмы Asus с процессором core 2 duo и диагональю экрана не меньше 12 дюймов» показывает, что один товар может быть охарактеризован целым рядом параметров. Проанализировав данный запрос, система сформирует

критерии отбора с соответствующими значениями и сделает выборку среди ноутбуков. Мы могли указать вместо ноутбука любой другой товар, дав таким образом понять РекС, в какой категории производить поиск.

Следует добавить, что пользователи часто формируют запросы, похожие на запрос «телевизор с диагональю от 12 до 15, с ценой от 3000 до 6000 т.р.». Отличительная особенность таких запросов - неявное указание размерности. В подобных случаях система должна понять, что диагональ чаще всего указывается в дюймах.

Часто бывает, что запросы, звучащие по разному, имеют один и тот же смысл. Например, запросы «автомобили, сделанные в Германии» и «немецкие авто», несмотря на абсолютно разное написание, одинаковы по смыслу. Многие запросы могут быть сформулированы в виде вопроса, в отличие от ранее рассмотренных примеров. В частности, предыдущий запрос может быть сформулирован так: «Какие есть немецкие авто?» или «Какие автомобили сделаны в Германии?».

Для обеспечения единообразия структуры естественно-языковых запросов и упрощения их обработки целесообразно предложить пользователю РекС формулировать запросы в виде описания объектов интереса и их свойств, используя существительные, прилагательные, предлоги, краткие прилагательные и логическую связку «отрицание» (передаваемую приставкой «не» или словом «не») и не используя причастные обороты, придаточные предложения и вопросительные предложения. Такой подход к нормализации структуры естественно-языковых запросов пользователя РекС лежит в основе данного исследования.

Первый этап обработки запроса пользователя РекС заключается в построении по запросу его семантического представления (СП), являющегося выражением некоторого СК-языка (стандартного концептуального языка). Определение класса СК-языков можно найти в [5 – 7]. Так строится К-представление (КП) запроса. Затем КП запроса преобразуется в SQL-выражение, которое соотносится с содержанием базы данных о товарах. После этого начинается диалог с пользователем. После ввода и интерпретации запроса часто выводится не один товар, а целый список. После этого РекС вступает в диалог с пользователем, предлагая ему сократить список с товаром при помощи дополнительных уточнений.

3. Разработка дополнительных предположений о структуре рассматриваемого концептуального базиса

На первом этапе формализации смысловых структур первоначальных запросов пользователя РекС введём дополнительные предположения об используемой системе первичных единиц концептуального уровня. Математическая модель этой системы единиц задаётся в теории К-представлений [5 – 7] определением понятия концептуального базиса.

Каждый концептуальный базис *B* является упорядоченный тройкой вида *(S, Ct, Ql)*, где *S, Ct, Ql* – упорядоченные наборы формальных объектов, называемые соответственно *сортовой системой, концептуально-объектной системой и системой кванторов и логических связок*.

Первым компонентом произвольной сортовой системы *S* является конечное множество символов *St*, эти символы называются сортами и интерпретируются как обозначения наиболее общих понятий из рассматриваемой группы предметных областей. Например, *St* может включать элементы *простр.об* (сорт «пространственный объект»), *физ.об* (сорт «физический объект»), *дин.физ.об* (сорт «динамический физический объект»), *орг* (сорт «организация»), *сит* (сорт «ситуация»), *соб* (сорт «событие», т.е. динамическая ситуация).

Произвольная концептуально-объектная система *Ct* является упорядоченной четвёркой вида

$$(X, V, tp, F),$$

где X – счётное множество символов, называемое *первичным информационным универсумом*, V – счётное множество символов, называемых *переменными*, F – непустое конечное подмножество множества X, состоящее из обозначений функций (другими словами, из функциональных символов). Элементы множества X интерпретируются как первичные (т.е. неструктурированные) единицы концептуального (или семантического) уровня. Отображение *tp* связывает с каждым элементом *d* из объединения множеств X и V некоторую цепочку *tp(d),* являющуюся формальной характеристикой элемента *d* и называемую *типом* элемента *d*.

Например, *типом* имени функции *Вес* может быть цепочка *{(физ. об, вещ. число)}*. Эта цепочка отображает информацию о том, что аргументом функции *Вес* может быть только физический объ-

ект, а значением функции является какое-то вещественное число.

В формулируемых ниже предположениях будем использовать обозначение *B* для рассматриваемого концептуального базиса.

Предположение 1. Первичный информационный универсум *X(B)* включает подмножества *Nt* и *Re*, где *Nt* – множество всех цепочек вида $d_1 \ldots d_n$, где $n \geq 1$, и для $k = 1,\ldots, n$ d_k – цифра из множества {'0','1',...,'9'}; *Re* – множество всех цепочек вида *b, c*, где $b,c \in Nt$.

Пример.
Множество *Nt* включает цепочки 123 и 4125; множество *Re* включает, в частности, цепочки 12,78 и 0,315.

Предположение 2. Множество сортов *St(B)* рассматриваемого концептуального базиса *B* включает различные элементы *нат* и *вещ*, причем для всякой цепочки $d \in Nt$ *tp(d)=нат,* и для всякой цепочки $h \in Re$ *tp(h)=вещ.*

Пример.
tp(123)= tp(4125)=нат; tp(12,78)= tp(0,315)= вещ.

Предположение 3. Первичный информационный универсум *X(B)* включает выделенное конечное подмножество *Units*, множество сортов *St(B)* включает сорт parameter-unit, и для каждого сорта *u* из подмножества *Units* выполняются соотношения

tp(u) $\in$ *St(B),(parameter-unit, tp(u))* $\in$ *Gen(B),*

т.е. сорт *tp(u)* является конкретизацией сорта *parameter-unit* для отношения общности *Gen(B)*.

Пример. Для рассматриваемого концептуального базиса *B* множество *X(B)* может включать элементы *рубль, евро,* множество *St(B)* может включать элементы *единица-стоимости, parameter-unit,* и выполняются соотношения

tp(рубль)=tp(евро)=единица-стоимости,
(parameter-unit, единица-стоимости) $\in$ *Gen(B),*

т.е. *parameter-unit* $\rightarrow$ *единица-стоимости* (поскольку для произвольных сортов *s, w* обозначение *(s, w)* $\in$ *Gen* равносильно обозначению $s \rightarrow w$).

Предположение 4. Множество сортов *St(B)* включает выделенный сорт *digit-param-value,* первичный информационный универсум *X(B)* включает подмножество *Param-values*, состоящие из всех цепочек вида *g/h*, где *g* – цепочка из объединения множеств *Nt* и *Re, h* $\in$ *Units*, и для каждого элемента *d* из подмножества *Param-values*

tp(d) $\in$ *St(B), digit-param-value* $\rightarrow$ *tp(d).*

Пример.
Можно определить концептуальный базис *B* таким образом, что *X(B)* $\supset$ *Param-values* $\supset$ *{210000/руб, 4,80/м, 14000/euro}.*

Определение. Пусть *B* – произвольный концептуальный базис, тогда упорядоченный набор *Dig-par-system* вида

(нат, вещ, Units, digit-param-value, Param-values) (1)

называется *разметкой числовых параметров для концептуального базиса B* $\Leftrightarrow$ когда для базиса *B* и компонентов набора (1) выполнены предположения 1 – 4.

Предположение 5. Множество сортов *St(B*) включает выделенный сорт *ling-value*, первичный информационный универсум *X(B)* включает различные элементы *small, middle, big,* причем *tp(small)=tp(middle)=tp(big)=ling-value;* кроме того, X(B) включает такой бинарный реляционный символ Лингв–оценка, что

tp(Лингв-оценка)={(digit-param-value, ling-value)}.

Определение. Пусть *B* – произвольный концептуальный базис, тогда упорядоченный набор *Ling-par-system* вида

(ling-value, small, middle, big, Лингв-оценка) (2)

называется *разметкой лингвистических параметров для концептуального базиса B* $\Leftrightarrow$ когда для базиса *B* и компонентов набора (2) выполнено предположение 5.

Определение. *Концептуальной сигнатурой* называется упорядоченный набор *Consign* вида

(B, Dig-par-system, Ling-par-system) (3)

где *B* – произвольный концептуальный базис, *Dig-par-system – разметка числовых параметров для базиса B* вида (1), *Ling-par-system – разметка лингвистических параметров для базиса B* вида (2), и выполняются предположения 1 – 5.

Данное определение будем интерпретировать как результат первого этапа разработки *математической модели системы первичных единиц концептуального уровня,* используемой рекомендательной системой.

Предположения 1 – 5 вводят специальные единицы концептуального уровня (=семантические единицы), которые будет удобно использовать

для представления содержания (смысла) запросов пользователя РекС.

Пример.

Семантику прилагательного «недорогой» из запроса «недорогой немецкий легковой автомобиль не старше пяти лет» раскрывает К-цепочка

Лингв-оценка(Цена(y_1),(small ∨ middle)),

где y_1 – обозначение произвольного объекта интереса пользователя РекС.

Обозначение. *Пусть Consign* – концептуальная сигнатура вида (3). Тогда будем обозначать концептуальный базис *B* через *B(Consign)*.

4. Разработка математической модели многообразия смысловых структур первоначального запроса пользователя рекомендательной системы

Используем введённое выше определение класса концептуальных сигнатур в качестве отправной точки построения математической модели, описывающей многообразие смысловых структур первоначальных естественно-языковых запросов пользователя РекС.

Определение 1. *Пусть Consign* – произвольная концептуальная сигнатура, *B = B(Consign)*, var – произвольная переменная из множества *V(B)*. Тогда *Lrel1(B, var)* – это множество всех цепочек СК-языка *Ls(B)*, представимых в виде *r(var, d)* или в виде ¬ *r(var, d)* или в виде *r(h(var),d)* или в виде ¬*r(h(var),d),* где *r* – бинарный реляционный символ из *X(B), d* ∈ *X(B), h* – некоторый одноместный (унарный) функциональный символ из *F(B)*.

Пример.

Можно построить такую концептуальную сигнатуру *Consign*, что *B = B(Consign),* y_1 ∈ *V(B)* и язык *Lrel1(B,* y_1*)* включает цепочки ¬*Старше(*y_1*, 5/год) и Больше(Цена(*y_1*, 250000/руб)*.

Определение 2. Пусть *Consign* – произвольная концептуальная сигнатура, *B = B(Consign), var* – произвольная переменная из множества *V(B)*. Тогда *Lrel2(B, var)* – это множество всех цепочек СК-языка *Ls(B)*, представимых в каком-либо из видов

$$r(var, (d_1 \vee ... \vee d_n)),$$
$$\neg r(var,(d_1 \vee ... \vee d_n)),$$
$$r(h(var),(d_1 \vee ... \vee d_n)),$$
$$\neg r(h(var),(d_1 \vee ... \vee d_n)),$$

где *r* – бинарный реляционный символ из *X(B), n>1,* $d_1, ..., d_n \in X(B)$, *h* – одноместный функциональный символ из *F(B)*.

Пример.

Нетрудно задать такую концептуальную сигнатуру *Consign*, чтобы $y_1 \in V(B(Consign))$ и *Lrel2(B(Consign),* y_1*)* включал цепочки

Фирма (y_1*, (BMW ∨ Volkswagen)),*
Цвет(y_1*, (тёмно-зелёный ∨ тёмно-синий))*.

Определение 3. *Пусть Consign* – произвольная концептуальная сигнатура, *B = B(Consign), var* – произвольная переменная из множества *V(B*). Тогда *Lfunc1(B, var*) – это множество всех цепочек СК-языка *Ls(B)*, представимых в виде (*f(var) ≡ d)* или в виде ¬*(f(var) ≡ d),* где *f* – одномерный (или унарный) функциональный символ из *F(B), d* ∈ *X(B)*.

Пример.

Можно построить такую концептуальную сигнатуру *Consign,* что $y_1 \in V(B(Consign))$ и язык *Lfunc1(B(Consign),* y_1*)* включает цепочку *(Цена(*y_1*) ≡ 354000/руб)*.

Определение 4. Пусть *Consign* – произвольная концептуальная сигнатура, *B = B(Consign), var* – произвольная переменная из множества *V(B)*. Тогда *Lfunc2(B, var)* – это множество всех цепочек СК-языка *Ls(B)*, представимых в виде *(f(var)* ≡ *(*$d_1 \vee ... \vee d_n$*))* или в виде ¬*(f(var)* ≡ *(*$d_1 \vee ... \vee d_n$*))*, где *f* – одноместный (или унарный) функциональный символ из *F(B), n>1,* $d_1, ..., d_n \in X(B)$.

Пример.

Легко определить такую концептуальную сигнатуру *Consign*, что $y_1 \in V(B(Consign))$ и язык *Lfunc2(B(Consign),* y_1*)* включает цепочку
(Страна-производитель (y_1*) ≡ (Германия ∨ Бельгия))*.

Используем введённые определения для формализации смысловой структуры тех фрагментов первоначального запроса пользователя РекС, которые описывают дополнительную информацию об объекте интереса. В частности, таким фрагментом является выражение «*Цвет – тёмно зелёный или тёмно-синий, не старше 5 лет*».

Определение 5. Пусть *Consign* – произвольная концептуальная сигнатура, *B = B(Consign)*, *var* – произвольная переменная из множества *V(B*). Тогда обозначим через *Lmany(B, var)* множество всех цепочек вида $(z_1 \wedge ... \wedge z_n)$, где $n>1, z_1, ..., z_n \in$ *Lrel1(B, var)* ∪ *Lrel2(B, var)* ∪ *Lfunc1(B, var)* ∪ *Lfunc2(B, var)*.

Пример.

Можно задать концептуальную сигнатуру *Consign* так, что $y_1 \in V(B(Consign))$ и язык $Lmany(B(Consign), y_1)$ включает цепочку

(Меньше(Цена(y_1),350000/руб) ∧ Цвет(y_1,(тёмно-зелёный ∨ тёмно-синий)) ∧ Страна-производитель (y_1,(Германия ∨ Бельгия))).

Определение 6. Пусть *B* – произвольный концептуальный базис, *var* – произвольная переменная из множества *V(B)*. Тогда пусть *Laddinf(B,var)=Lrel1(B,var) ∪ Lrel2(B,var) ∪ Lfunc1(B,var) ∪ Lfunc2(B,var) ∪ Lmany(B,var)*.

Введём понятие первичного семантического образа первоначального запроса пользователя РекС.

Важным фактором, который необходимо учитывать, является то, что, вероятно, большинство запросов пользователя будут состоять из двух частей. Часть 1 кратко обозначает объект интереса пользователя (например, «Немецкий легковой автомобиль»). Часть 2 перечисляет дополнительные требования, которым должен удовлетворять объект интереса (например, «не старше 5 лет, цвет тёмно-зелёный или тёмно-синий»).

СК-языки, определяемые теорией К-представлений [5 – 7], позволяют строить обозначения упорядоченных наборов формальных объектов как цепочки вида $<w_1,w_2,\ldots,w_n>$, *где* $k>1$, $w_1,\ldots,w_k$ – выражения СК-языка *Ls(B)*, и *B* – некоторый концептуальный базис.

Поэтому в данной работе предлагается строить первичный семантический образ запроса пользователя РекС в виде

<Semrepr1, Semrepr2>,

где *Semrepr1* – семантическое представление (СП) краткого описания объекта интереса пользователя (например, СП выражения *«Немецкий легковой автомобиль»*), а *Semrepr2* – СП фрагмента, перечисляющего дополнительные требования, которым должен удовлетворять объект интереса (например, СП выражения *«не старше 5 лет, цвет тёмно-зелёный или тёмно-синий»)*.

Во вводимом ниже определении параметр *n* интерпретируется как порядковый номер запроса пользователя, поэтому $n \geq 1$.

Определение 7. Пусть *Consign* – произвольная концептуальная сигнатура, $B = B(Consign)$, $n \geq 1$. Тогда *Lsemimage(B,n)* – это множество всех цепочек вида

<Semrepr1, Semrepr2>,

где *Semrepr1* – цепочка СК-языка *Ls(B)* вида

все concept $*(r_1,d_1)\ldots(r_k,d_k)$,

где *concept* ∈ *X(B)*, тип элемента *concept* – цепочка *tp(concept)* – начинается с символа ↑, $k \geq 1$, $r_1,\ldots, r_k$ – бинарные реляционные символы из *X(B)*, $d_1, \ldots, d_k \in X(B)$, $Semrepr2 \in Laddinf(B,y_n)$.

Множество цепочек *Lsemimage(B,n)* будем называть языком первичных семантических образов с параметрами *B* и *n*.

Пример.

Пусть запрос1= «*Немецкие легковые автомобили не старше 5 лет. Цвет – тёмно-зелёный или тёмно-синий, не дороже 350000 рублей*», тогда можно задать концептуальную сигнатуру *Consign* так, что язык *Lsemimage(B(Consign), 1)* будет включать выражение

<Semrepr1, Semrepr2>,

где *Semrepr1=все авто * (Страна-производитель, Германия)(Вид-авто, легковой)* ,

Semrep2=(¬Больше(Возраст(y_1),5/год) ∧ Цвет(y_1,(тёмно-зелёный ∨ темно-синий)) ∧ ¬Больше(Цена(y_1), 350000/руб)).

Выражение *<Semrepr1, Semrepr2>* будем интерпретировать как первичный семантический образ запроса 1.

В построенном в предыдущем примере выражении вида *<Semrepr1, Semrepr2>* значения параметров *Страна-производитель* и *Вид-авто* задаются в подцепочке *Semrepr1* не в той же форме, что и значения параметров *Возраст, Цвет* и Цена в подцепочке *Semrepr2.*

Анализ выразительных механизмов СК-языков позволил предложить такую форму первичного семантического представления (ПСП) исходного естественно-языкового запроса пользователя РекС, которая дает возможность единообразно отображать значения всех параметров объекта интереса пользователя, что удобно для последующей обработки запроса. Для задания такой формы ПСП запроса потребуется ввести дополнительные предположения о рассматриваемом концептуальном базисе.

Предположение 6. Множество сортов *St(B)* включает выделенные элементы *физ.об* и *инф.об,* первичный информационный универсум *X(B)* включает различные элементы *Объекты-интереса,Элемент, Описание1,* причем

tp(Объекты-интереса) ={(инф.об, {физ.об}, P})} ,
tp(Элемент)={(физ.об, { физ.об})} ,
tp(Описание1)=)={(инф.об, P})} ,

где $P = P(B)$ – выделенный сорт «смысл сообщения» концептуального базиса B; подмножество интенсиональных кванторов 1-го вида $Int_1(B)$ множества $X(B)$ включает элемент *произв;* подмножество интенсиональных кванторов 2-го вида *Int2(B*) множества $X(B)$ включает элемент *все.*

Выделенные элементы *физ.об* и *инф.об* называются соответственно сортом «физический объект» и сортом «информационный объект», элемент произв называется интенсиональным квантором «произвольный».

Определение 8. Пусть B – произвольный концептуальный базис, тогда упорядоченный набор *Output-mark-system* вида

(физ.об, инф.об, произв, все, Объекты-интереса, Элемент, Описание1) (4)

называется *выходной разметкой для концептуального базиса* B ⇔ когда для базиса B и компонентов набора (4) выполнено предположение 6.

Предположение 7. Множество переменных $V(B)$ включает выделенные подмножества *Vrequest, Vclass и Vobject,* где *Vreques*t состоит из элеменов вида $запрос_n$, *Vclass* состоит из элеменов вида S_n, *Vobject* состоит из элементов вида y_n, где $n \geq 1$, причем $tp(запрос_n) = инф.об$, для каждой переменной z из *Vclass* $tp(z) = \{физ.об\}$, для каждой переменной z из *Vobject* $tp(z) = физ.об$.

Определение 9. *Проблемно-ориентированным концептуальным базисом* называется упорядоченный набор *Probs* вида

(B, Dig-par-system, Ling-par-system, Output-mark-system) (5)

где упорядоченная тройка *(B, Dig-par-system, Ling-par-system)* является произвольной концептуальной сигнатурой, *Output-mark-system* – выходная разметка вида (4) для концептуального базиса B, и выполняются предположения 6 и 7.

Данное определение будем интерпретировать как *математическую модель системы первичных единиц концептуального уровня,* используемой рекомендательной системой.

Предположения 1 – 7 вводят специальные единицы концептуального уровня (другими словами, семантические единицы), которые будет удобно использовать для представления содержания (смысла) запросов пользователя РекС.

Обозначение. Пусть *Probs* – проблемно-ориентированный концептуальный базис вида (5). Тогда будем обозначать концептуальный базис B через *B(Probs).*

Определение 10. Пусть *Probs* – произвольный проблемно-ориентированный концептуальный базис вида (5), $n \geq 1$, z – цепочка из языка *Lsemimage(B,n)* вида <*Semrepr1, Semrepr2*>, где для цепочек *Semrepr1* и *Semrepr2* выполнены предположения определения 7. Тогда для $m = 1, ..., k$ выражение $Form(r_m,d_n,y_n)$ в случае $r_m \in F(B)$ обозначает цепочку вида $(r_m,(y_n) \equiv d_m)$, а в случае $r_m \in X(B) \setminus F(B)$ обозначает цепочку вида $r_m(y_n,d_m)$.

Определение 11. Пусть *Probs* – произвольный проблемно-ориентированный концептуальный базис вида (5), $n \geq 1$, *semrequest* – цепочка языка первичных семантических образов *Lsemimage(B, n)*, имеющая вид <*Semrepr1, Semrepr2*>, где для цепочки *Semrepr1* выполнены предположения определения 7. Тогда отображение *Secondary* задаётся следующим образом: *Secondary(semrequest)* – цепочка вида

Объекты-интереса(запрос$_n$, все concept(Элемент, S_n),*
Описание1(произв concept(Элемент, S_n):y_n,*
(Form(r_1,d_1,y_n) ∧....∧ Form(r_k,d_k,y_n) ∧ Semrepr2))).

Выражение *Secondary(semrequest)* будем называть *первичным семантическим представлением* входного запроса, соответствующим первичному семантическому образу *semrequest.*

Пример. Пусть запрос 1 = *«Немецкие легковые автомобили не старше 5 лет. Цвет – тёмно зелёный или тёмно-синий, не дороже 350000/рублей»,* и *semrequest* – это построенный выше первичный семантический образ запроса 1 вида <*Semrepr1, Semrepr2*>. Тогда *Secondary(semrequest)* – первичное семантическое представление рассмотренного запроса – является цепочкой вида

Объекты-интереса (запрос$_1$, все авто(Элемент, S_1),*
Описание1(произв авто(Элемент, S_1):y_1,*
(Страна-производитель(y_1, Германия) ∧
Вид-авто(y_1,легковой) ∧ ¬Больше(Возраст(y_1), 5/год) ∧
Цвет(y_1, (тёмно-зелёный ∨ тёмно-синий)) ∧
¬Больше (Цена(y_1), 350000/руб)).

Определение 12. Пусть *Probs* – произвольный проблемно-ориентированный концептуальный базис, $B = B(Probs)$, $n \geq 1$. Тогда *Requests(B,n)* = {z | найдется такая цепочка *semrequest* из языка первичных семантических образов *Lrequest1(B,n), что z = Secondary(semrequest)}.*

Определение формального языка *Requests(B,n)* для параметров *B* и *n* будем интерпретировать как математическую модель многообразия смысловых структур, соответствующих первоначальным запросам пользователей рекомендательной системы, сформированным на нормализованном естественном (русском) языке.

5. Преобразование запроса в К-представление и затем в SQL-выражение

Построенная модель послужила отправной точкой для разработки технического задания на проектирование лингвистической базы данных (ЛБД) и алгоритма семантико-синтаксического анализа запросов, преобразующего запрос в семантическое представление, являющееся выражением некоторого СК-языка, т.е. К-представлением запроса. Основу ЛБД составляют лексико-семантический словарь и словарь предложных семантико-синтаксических фреймов, структура которых формально определена в [5, 7].

В простейшем случае К-представление запроса можно представить в виде строки, главную роль в которой играет объект, далее идут его свойства, описанные как названия свойств и их значения, например: *все авто*(Страна, Германия)*. Мы видим, что в качестве объекта указан автомобиль, у него представлено только одно свойство – страна и его значение *Германия*. Таким образом, мы можем понять, что пользователь запрашивает все автомобили, выпущенные в Германии. Подобного рода выражение может быть преобразовано в формальный запрос вида

*Объекты-интереса (запрос$_1$, все авто * (Элемент, S_1), Описание1 (произв авто * (Элемент, S_1) : y_1, Страна-производитель (y_1, Германия))).*

Слово «авто» в данном случае указывает предметную область. Слово «страна» указывает на поле в таблице с характеристиками автомобиля. Название поля в базе данных (БД), по которому нужно производить выборку, мы получаем из дополнительной таблицы, содержащей названия полей и их русскоязычные наименования. В результате преобразований запрос к SQL серверу будет выглядеть следующим образом: select * from auto where auto. country=*'Германия'*.

Разработанный подход не предполагает жёсткой привязки к структуре БД, так как в подобном случае его внедрение будет нерентабельным. Для связи интерпретатора и таблиц с товаром и его характеристиками используются промежуточные таблицы, в которых русскоязычные значения и их семантические представления связаны с соответствующими полями товара. Связь может осуществляться напрямую (например, поле «Цена» соответствует полю «price»), а может быть выполнена в виде подзапроса.

Поскольку большинство готовых БД имеют хотя бы первую степень нормализации, рассмотренный выше пример запроса будет некорректен. Как правило, в таблице с товаром будет храниться только идентификатор страны, а само значение – в другой таблице. Поскольку предлагаемая структура имеет гибкую форму, запрос может выглядеть следующим образом:

select from country, country_kategories,_tovar, kategories where country.country_name = 'Германия' and country. country_id = country_kategories.country_id and tovar.id _kategorii = kategories.id_kategories and kategories.id _pred_kategorii = country_kategories.id_kategories).*

Запрос может содержать более одного параметра, например, может являться цепочкой

Объекты-интереса (запрос$_2$, все авто(Элемент, S_2), Описание1 (произв авто *(Элемент, S_2) :y_2, ((Страна-производитель (y_2, Франция) $\wedge$ Тип-кузова (y_2, седан) $\wedge\neg$ Больше1(Цена(y_2), 14000/USD))))).*

Такому запросу соответствует SQL-выражение *select * from auto where auto.country=' Франция' and cars.body.type='седан' and cars.price<=14000.*

Преобразование К-выражения в SQL-запрос, то есть запрос, который пригоден для обращения к БД, происходит, главным образом, путём подстановки полученного выражения в заранее известный шаблон. После анализа шаблона мы можем сопоставить ряд значений с реальной базой данных.

6. Применение бизнес-правил для преобразования первичного К-представления в глубинное К-представление

В формировании рекомендации пользователю могут использоваться бизнес-правила, предназначенные, во-первых, для продвижения того или иного товара или групп товаров. Подобная необходимость может быть связана с ярко выраженной сезонной принадлежностью или проведением рекламной компании того или иного бренда. Другим аспектом применения бизнес-правил является уточнение нечётких характеристик [8 – 10], упоминаемых в запросе. Например, уточнение разме-

ра *(большой, средний, крупный)* или цены *(дорогой – дешевый)*. Подобные характеристики не являются чёткими, и по ним нельзя непосредственно построить SQL-запрос. Однако их можно сравнить с заранее заготовленными значениями, связав тип товара и набор констант, обозначающих его характеристики.

Так, запрос «Дорогой мобильный телефон», адресованный РекС в области бытовой электроники, может быть интерпретирован как «телефон, цена которого выше средней цены всех мобильных телефонов или его цена выше 10 000 руб.».

7. Программная реализация и экспериментальные результаты

Реализацией изложенного метода стала первая версия рекомендательной системы в программной среде PHP + MySQL [8 – 10]. Тестирование показало, что средний запрос, включающий две характеристики, выполняется за 0.09 сек. и обращается к SQL серверу 12 раз. Добавление неоднозначного параметра приводит к незначительному увлечению времени работы скрипта, а именно, до 0.15сек., и количество SQL запросов в таком случае возрастает до 16.

Для проверки работоспособности разработанных скриптов они были внедрены на сайт с реально существующей базой данных автомобилей. В ходе эксперимента, на основании рейтинга LiveInternet, был отмечен рост посещаемости от 20 до 28 процентов по сравнению с аналогичным периодом неделей ранее. При этом увеличилось среднее время посещения сайта и количество просмотров в целом в среднем на 37 процентов.

8. Заключение

Основной результат проведенного исследования заключается в разработке нового метода ЕЯ-интерфейсов рекомендательных систем (РекС). Метод базируется на построении математической модели многообразия смысловых структур, соответствующих первоначальным запросам пользователя РекС. Модель использует выразительные механизмы СК-языков, определяемых теорией К-представлений. Предложенный метод программно реализован в среде PHP + SQL. Экспериментальные данные показывают, что внедрение разработанного метода является эффективным способом повышения популярности ресурса. ■

Литература

1. Chai J., Horvath V., Nicolov N., Stys-Budzikowska M., Kambhatla N., and Zadrozny W. Natural Language Sales Assistant – A Web-based Dialog System for Online Sales // Proceedings of the Thirteenth Innovative Applications of Artificial Intelligence Conference. The AAAI Press, 2001, pp. 19-26.
2. Chai J., Horvath V., Nicolov N., Stys M., Kambhatla N., Zadrozny W., Melville P. Natural Language Assistant – A Dialog System for Online Product Recommendation //AI Magazine, 2002, V. 23, No. 2, p. 63-76.
3. Жигалов В.А. Естественное общение с приложением // Открытые системы, № 12, 2001, с. 22-27.
4. Жигалов В.А. Естественно-языковой интерфейс в электронной коммерции // Труды Междунар. семинара Диалог'2002 по компьютерной лингвистике и ее приложениям. Том 2. Прикладные проблемы, URL http://www.dialog-21.ru/archive.asp?y=20012 & vol=6078 &parent_menu_id=711.
5. Фомичёв В.А. Формализация проектирования лингвистических процессоров. Москва, Макс Пресс, 2005.-368 с.
6. Фомичёв В.А. Математические основы представления содержания посланий компьютерных интеллектуальных агентов. М., ГУ-ВШЭ, изд-во «ТЕИС», 2007.-176 с.
7. Fomichov V.A. Semantics-Oriented Natural Language Processing: Mathematical Models and Algorithms. Springer: New York, Dordrecht, Heidelberg, London, 2010.-354 p.
8. Правиков А.А. Некоторые принципы и средства организации диалога с пользователем рекомендательной системы // Научные труды. МАТИ, 2009. Вып. 15(87), с. 192-193.
9. Правиков А.А. Разработка и программная реализация методов математического моделирования содержания диалога с ЭВМ пользователя рекомендательной системы // Научные труды Международной молодежной научной конференции XXXV Гагаринские чтения, Москва, 7-10 апреля 2009.
10. Правиков А.А. Элементы формализации диалога с пользователем русскоязычного интерфейса рекомендательной системы // Научные труды Международной молодежной научной конференции XXXVI Гагаринские чтения, Москва, 8-9 апреля 2010.