ВЫСШАЯ ШКОЛА ЭКОНОМИКИ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

В.Ф. Хорошевский

ОБ ОДНОМ МЕТОДЕ СЕМАНТИЧЕСКОЙ ИНТЕРПРЕТАЦИИ ПАТТЕРНОВ ДАННЫХ НА ОСНОВЕ СТРУКТУРНОГО ПОДХОДА

Препринт WP7/2012/08
Серия WP7
Математические методы
анализа решений в экономике,
бизнесе и политике

Москва
2012

УДК 519.2:332.1
ББК 65в6
Х64

Редакторы серии WP7
«Математические методы анализа решений в экономике,
бизнесе и политике»
Ф.Т. Алескеров, В.В. Подиновский, Б.Г. Миркин

Х64 **Хорошевский, В. Ф.** Об одном методе семантической интерпретации паттернов данных на основе структурного подхода : препринт WP7/2012/08 [Текст] / В. Ф. Хорошевский ; Нац. исслед. ун-т «Высшая школа экономики». – М.: Изд. дом Высшей школы экономики, 2012. – 28 с.

В работе обсуждается структурный подход к обработке изображений паттернов данных, а также вопросы семантической интерпретации паттернов данных. Приведены краткие сведения о состоянии исследований и разработок в области структурного подхода к распознаванию образов и анализу сцен. Описание предлагаемого структурного подхода ведется в контексте обсуждения вопросов анализа изображений паттернов данных, а также семантической интерпретации паттернов данных, полученных с помощью классических методов статистической обработки индикаторов науки, образования и инновационной деятельности в регионах РФ с использованием данных, предоставленных ИСИЭЗ НИУ ВШЭ.

УДК 519.2:332.1
ББК 65в6

Хорошевский В.Ф. – Центр информационно-аналитических систем ИСИЭЗ НИУ ВШЭ, Вычислительный центр им. А.А. Дородницына РАН; vkhoroshevsky@hse.ru, khor@ccas.ru.

Работа выполнена при финансовой поддержке Министерства образования и науки Российской Федерации по государственному контракту от 14.06.2012 г. № 07.514.11.4144 в рамках ФЦП «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2007–2013 гг.

Препринты Национального исследовательского университета
«Высшая школа экономики» размещаются по адресу: http://www.hse.ru/org/hse/wp

Содержание

1. Введение

В настоящее время одной из проблем, решение которой определяет достойное положение России в рамках мирового постиндустриального общества, является инновационное развитие всех ее регионов. В связи с этим особую важность приобретает мониторинг состояния науки, образования и инноваций. Уже сегодня в этом направлении ведутся активные исследования и разработки, но сложность задачи настолько велика, что требуются новые подходы и методы ее решения.

Одним из таких подходов является анализ паттернов, выделенных в результате применения классических методов обработки статистических данных, который поддерживает поиск взаимосвязей исследуемых объектов, их классификацию и исследование процессов развития объектов во времени. При этом особую роль играют методы, обеспечивающие структурные спецификации паттернов, поскольку именно эти характеристики являются наиболее важными для идентификации трендов и прогнозирования.

В настоящей работе обсуждается структурный подход к обработке изображений паттернов данных, а также вопросы семантической интерпретации паттернов данных, базой которого являются результаты статистической обработки индикаторов науки, образования и инновационной деятельности в регионах РФ с использованием данных, предоставленных ИСИЭЗ НИУ ВШЭ [1–4], а также методы и алгоритмы анализа паттернов данных [5–7] и результаты формирования паттернов данных науки, образования и инновационной деятельности, полученные в работе [8].

Организовано изложение следующим образом. В разделе 2 приведены краткие сведения о состоянии исследований и разработок в области структурного подхода к распознаванию образов и анализу сцен. Показано, что данный подход дает хорошо интерпретируемые результаты в случае обработки сложных изображений различной природы. В заключительной части раздела дается краткое описание языка PDL (Picture Description Language), который используется для описания изображений на базе специальных грамматик.

Описание предлагаемого структурного подхода представлено в разделе 3 и ведется в контексте обсуждения вопросов анализа изо-

бражений паттернов данных, а также семантической интерпретации паттернов данных, полученных с помощью методов статистической обработки индикаторов науки, образования и инновационной деятельности в регионах РФ на базе материалов ИСИЭЗ НИУ ВШЭ [1–4].

В Заключении приведены выводы и рекомендации по использованию структурного подхода для анализа паттернов данных в аналитических задачах оценки состояния регионов России в части индикаторов науки, образования и инновационной деятельности, а также возможностей прогнозирования их развития.

2. Структурные методы описания изображений: состояние исследований и разработок

Структурный подход к описанию изображений разных классов развивается уже около полувека. И в рамках этого подхода, который возник в связи с необходимостью совершенствования методов и средств разработки распознающих систем, уже получены серьезные научные и практические результаты [9–12]. Становлению и развитию структурного подхода способствовали многие дисциплины (статистика, лингвистика, вычислительная математика, теория управления, исследование операций и др.). Как представляется с сегодняшних позиций, основы структурного подхода к анализу изображений были заложены в работах по теории решений в рамках дискриминантного подхода [13]. А затем, в первую очередь в силу необходимости решения проблем распознавания изображений и анализа сцен, появился структурный (синтаксический) подход. При этом синтаксический подход базируется на аналогии между структурой изображений и синтаксисом языка, что позволяет использовать в данной области аппарат математической лингвистики, хотя в общем случае в рамках синтаксического подхода используются и нелингвистические методы, что позволяет рассматривать его как гибридный. Дискриминантный и синтаксический подходы отличаются друг от друга тем, что в первом для описания или моделирования образов пользуются вероятностными распределениями реализаций образа, а во втором – синтаксическими правилами, или грамматиками. При этом эффективность каждого из под-

ходов зависит от конкретной задачи и часто возникает необходимость в одновременном их применении. Акцент синтаксического подхода на таких задачах, где важна информация, описывающая структуру каждого анализируемого объекта, а от процедуры распознавания требуется, чтобы она давала возможность не только отнести объект к определенному классу, но и специфицировать те свойства объекта, которые исключают его отнесение к другому классу.

Типичным примером таких задач служит распознавание изображений и/или анализ сцен, где объекты сложны, а число требуемых для описания признаков велико. В таких условиях становится эффективной идея описания сложного объекта в виде иерархической структуры образов более простых объектов и генерации общего описания с помощью правил соответствующей грамматики. Для примера, на рис. 1 представлено описание объекта «прямоугольник» (а) при помощи операции композиции из базовых (непроизводных) элементов (б), а на рис. 2 – изображение цифры 9 и ее структурное описание.

Рис. 1. Описание объекта «Прямоугольник»(а), базовые элементы описания (б)

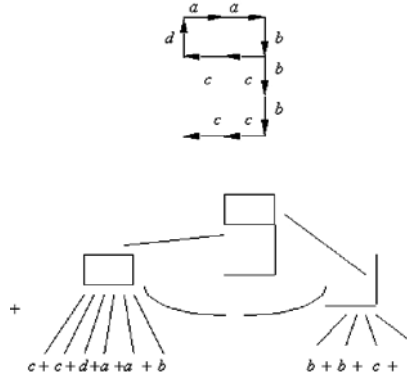


Рис. 2. Изображение цифры 9 и соответствующее ей структурное описание

Соответствующие грамматики языков описания объектов формируются на этапе обучения с использованием обучающей выборки и/или с помощью экспертов, а теоретической базой данного подхода является теория формальных языков и лежащих в их основе порождающих грамматик.

Как известно, пионером структурного подхода является Р. Нарасимхан (R. Narasimhan), опубликовавший в 1962 г. работу «Лингвистический подход к распознаванию образов» [14]. По сути дела, именно в этой работе и были заложены теоретические основы структурного подхода, а практическим его применением была разработанная автором грамматика распознавания рукописных изображений латинских букв и цифр. Чуть позже Нарасимхан, Ледли и другие исследователи использовали структурный метод для анализа изображений треков частиц в пузырьковой камере и изображений хромосом [9, 15, 16]. Примерно в это же время Аланом Шоу (A. Shaw) из Стэнфорда был разработан специальный язык PDL (Picture Description Language), ориентированный на анализ изображений [17].

В 70–80-х годах прошлого века активные исследования и разработки в области применения грамматик для распознавания изображений и анализа сцен велись в университете Пердью (Purdue University) под руководством К.-С. Фу (King-Sun Fu). В это же время была опубликована первая фундаментальная монография по структурным методам распознавания образов [13].

В дальнейшем развитие работ в данной области пошло по пути создания стохастических языков [18, 19], поскольку наличие шума в изображениях часто приводило к неоднозначности в лингвистических представлениях объектов. В стохастических грамматиках для распознавания/порождения цепочек, принадлежащих таким языкам, правила вывода снабжаются вероятностными оценками полезности их применения, что в случае неоднозначности обеспечивает выбор в качестве синтаксического описания таких цепочек наиболее вероятного вывода.

В настоящее время синтаксический подход развивается в направлении интеграции используемых здесь методов и средств с классическими дискриминантными методами, что позволяет говорить о гибридном подходе к структурному описанию и обработке изображений [20].

Как отмечалось выше, язык PDL является одним из самых известных и активно используемых формализмов описания изображений. С учетом этого в рамках настоящей работы предлагается использовать этот формализм для семантической интерпретации паттернов данных. Ниже кратко обсуждаются основные конструкции языка PDL и его свойства.

Основная идея языка PDL состоит в том, чтобы строго специфицировать алгебру описания произвольных графических изображений на основе конечного множества графических примитивов и грамматики, порождающей (распознающей) все нужные изображения и только их.

Терминальные символы грамматики описания изображений или, как их определяет А. Шоу, базовые (непроизводные) элементы выбираются в зависимости от предметной области. При этом любой терминал определяется как объект с двумя выделенными точками – точкой начала (tail) и точкой конца (head). На вид самих терминалов никаких ограничений не накладывается, но их объединение в более сложные объекты может происходить только «через» начальные и концевые точки. При таком подходе любое изображение может быть представлено графом его примитивов, каждый из которых задается своим списком атрибутов:

PRIMITIVE CLASS = (<NAME>, <спецификация tail>, <спецификация head>, <атрибут-1>, <атрибут-2>, ...<атрибут-N>).

Кроме того, в языке допускается использование «пустых» (невидимых – invisible) примитивов, которые могут использоваться для связи отдельных фрагментов изображения или спецификации геометрических отношений между ними. Один специальный примитив (null point) λ играет в языке PDL особую роль. Для этого примитива хвост (tail) и голова (head) совпадают. Таким образом, данный примитив представляет в графе помеченный узел (node).

С учетом вышесказанного синтаксис языка PDL определяется следующими правилами:

$S \rightarrow p \mid (S\ \theta\ S) \mid (\sim S) \mid (\neg S) \mid T(\omega)\ S \mid S^{\iota}$

$\theta \rightarrow + \mid \times \mid - \mid * \mid \sim$

p – примитивы

{+, ×, –, *} – бинарные операторы конкатенации

{~, ¬, T(ω)} – унарные операторы

$^{\iota}$ – пометки в графе.

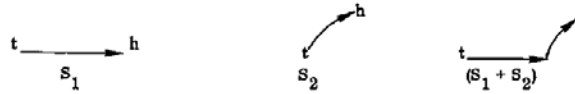
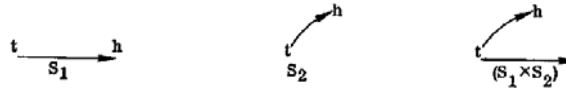
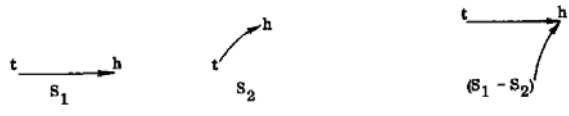
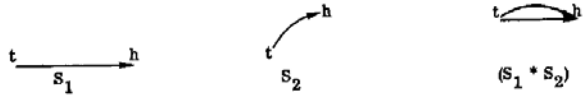
Все бинарные операторы конкатенации определяются следующим образом:

Tail $((S_1 \theta S_2)) =$ **Tail** (S_1)

Head $((S_1 \theta S_2)) =$ **Head** (S_2), $\theta \in \{+, \times, *, \sim\}$.

При этом семантика основных операторов языка PDL определяется правилами, представленными в табл. 1.

Таблица 1. Основные операторы языка PDL

№	Оператор	Семантика оператора
1	$(S_1 + S_2)$	
2	$(S_1 \times S_2)$	
3	$(S_1 - S_2)$	
4	$(S_1 * S_2)$	
5	$(S_1 \sim S_2)$	$\equiv (S_1 + (\sim S_2))$ для бинарного оператора «~»
6	$(\sim S_2)$	**Tail** $((\sim S)) =$ **Head** (S) **Head** $((\sim S)) =$ **Tail** (S)
7	$(\neg S)$	**Head** $(\neg S) =$ **Head** (S) **Tail** $(\neg S) =$ **Tail** (S)

В дополнение к операторам, представленным выше, в языке PDL определен унарный оператор $T(\omega)$, который используется для аффинных преобразований примитивов и/или классов примитивов,

и унарный оператор «$^{\iota}$» – для присваивания меток объектам в графе изображения.

Эквивалентность структур, представленных в языке PDL, определяется следующим образом:

S_1 слабо эквивалентен S_2 ($S_1 \equiv_w S_g$), если существует изоморфизм между графами S_1 и S_2 такой, что их соответствующие дуги имеют одинаковые имена.

S_1 эквивалентен S_2 ($S_1 \equiv S_2$), если

$S_1 \equiv_w S_2$ и

Tail(S_1) = Tail (S_2)

Head(S_1) = Head (S_2).

Для выполнения преобразований над структурами, представленными в языке PDL, используются следующие алгебраические свойства введенных выше операторов:

Ассоциативность бинарных операторов PDL

$((S_1 + S_2) + S_3) \equiv (S_1 + (S_2 + S_3))$

$((S_1 \times S_2) \times S_3) \equiv (S_1 \times (S_2 \times S_3))$

$((S_1 - S_2) - S_3) \equiv (S_1 - (S_2 - S_3))$

$((S_1 * S_2) * S_3) \equiv (S_1 * (S_2 * S_3))$

Коммутативность оператора *

$(S_1 * S_2) \equiv (S_2 * S_1)$

Слабая коммутативность операторов × и –

$(S_1 \times S_2) \equiv_w (S_2 \times S_1)$

$(S_1 - S_2) \equiv_w (S_2 - S_1)$

Стандартные правила булевой алгебры для оператора ~

$(\sim(S_1 + S_2)) \equiv ((\sim S_2) + (\sim S_1))$

$(\sim (S_1 * S_2)) \equiv ((- S_2) * (\sim S_1))$

Правило де Моргана для оператора ~ по отношению к операторам × и –

$(\sim(S_1 \times S_2)) \equiv ((\sim S_2) - (\sim S_1))$

$(\sim(S_1 - S_2)) \equiv ((- S_2) \times (\sim S_1))$

Правила преобразования «пустых» примитивов

$(S\ \theta\lambda) \equiv (\lambda\theta\ S)$ $\quad \theta \in \{+, \times, -, *\}$

$(S\ \varphi\lambda) \equiv S$ $\quad \varphi \in \{+, \times, -\}$

$(\sim \lambda) \equiv \lambda$

$(\lambda\theta\lambda) = \lambda$

и некоторые другие алгебраические правила.

Для иллюстрации возможностей языка PDL ниже приведен пример описания класса «Дом» с различными типами крыш. Для этого определены шесть терминальных символов (примитивов), представленных на рис. 3.

Рис. 3. Примитивы описания изображений класса «Дом»

С учетом приведенных на рис. 3 примитивов и правил языка PDL изображение объекта «Дом с круглой крышей» можно представить следующим образом:

$$\mathbf{H_2 \rightarrow ((p_6 * p_3) * ((p_4 + p_3) + p_5))}$$

Соответствующий граф изображения показан на рис. 4.

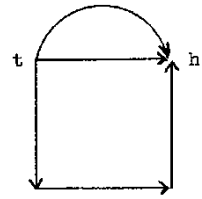


Рис. 4. Граф изображения «Дом с круглой крышей»

При этом класс всех различных домов H специфицируется грамматикой с правилами вывода вида:

$H \rightarrow (R * B)$
$R \rightarrow R1 \mid R2$
$R1 \rightarrow ((p_1 + p_2) * p_3)$
$R2 \rightarrow (p_6 * p_3)$
$B \rightarrow ((p_4 + p_3) + p_5)$

Нетрудно заметить, что язык описания изображений PDL является, по существу, метаязыком, а конкретное множество предложений, выводимых в этом языке, определяется подъязыком, порождаемым соответствующей грамматикой.

С практической точки зрения язык PDL интересен и тем, что для него существует эффективный алгоритм распознавания изображений, общая схема которого представлена на рис. 5.

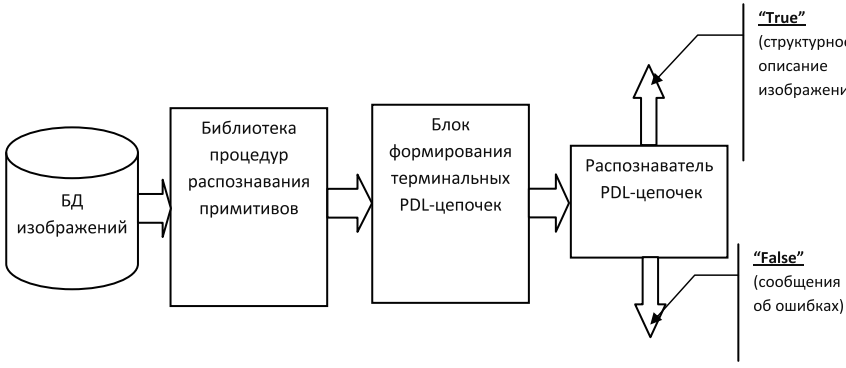


Рис. 5. Общая схема распознавания изображений, специфицированных в языке PDL

Библиотека процедур распознавания примитивов обеспечивает переход от изображений, представленных в БД, к синтаксическим примитивам. Следующий блок – блок формирования терминальных PDL-цепочек – обеспечивает символьное представление изображений, поступающих на вход PDL-распознавателя. В зависимости от типа грамматики распознавания изображений этот блок реализуется конечным автоматом, МП-автоматом или распознавателем расширенных сетей переходов Вудса [21, 22].

В случае успешного распознавания на выходе PDL-распознавателя формируется дерево грамматического разбора входной PDL-строки, которое дает структурное описание изображения. В случае ошибок во входной PDL-строке на выходе формируется список сообщений о причинах невозможности распознавания.

Таким образом, структурный подход к распознаванию изображений обеспечивает получение формальных описаний достаточно широкого класса изображений, их эффективное распознавание, а так-

же, благодаря алгебраическим свойствам операторов языка PDL, возможности эквивалентных преобразований изображений.

3. Структурный подход к семантической интерпретации паттернов данных

Учитывая все вышесказанное, ниже для семантической интерпретации паттернов данных, полученных в результате работы алгоритмов кластеризации и классификации временных рядов, предлагается использовать структурный подход.

Детальное описание конкретных алгоритмов семантической интерпретации паттернов данных, основанное на использовании специализированной библиотеки процедур распознавания примитивов и разработанной распознающей грамматики, приводится в следующих подразделах.

3.1. Онтологическая модель исходных данных

В [8] зафиксировано, что паттерны данных представляются как кусочно-линейные аппроксимации соответствующих временных рядов. Для иллюстрации, на рис. 6 приведены результаты расчетов паттернов данных науки, образования и инновационной деятельности.

Анализ этих результатов показывает, что для построения онтологических моделей исходных данных можно опираться на модели отрезков, имеющих достаточно четкую ориентацию.

По аналогии с названиями трендов, которые используются при анализе временных рядов при торгах на биржах, введем в рассмотрение три класса отрезков: «бычьи» (Ox), «медвежьи» (Bear) и «боковые» или «побочные» (Flat).

В соответствии с концепцией структурного подхода и основными конструкциями языка PDL рассмотренными выше, элементы всех классов будем представлять отрезками с определенной ориентацией и спецификацией типа отрезка, а также фиксацией координат его начала (tail) и конца (head). С учетом вышесказанного описание любого элемента может быть специфицировано следующим образом:

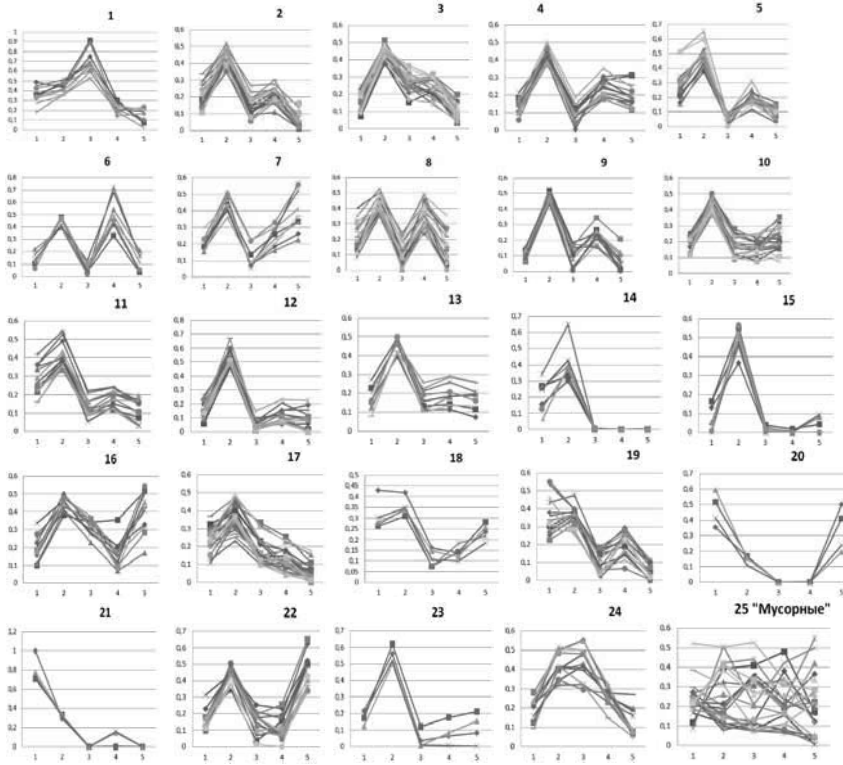


Рис. 6. Результаты расчетов паттернов данных науки, образования и инновационной деятельности

ELEM ::= TYPE(Tail Point Head Point)

TYPE ::= O | B | F

Tail Point ::= (tail_Xtail_Y)

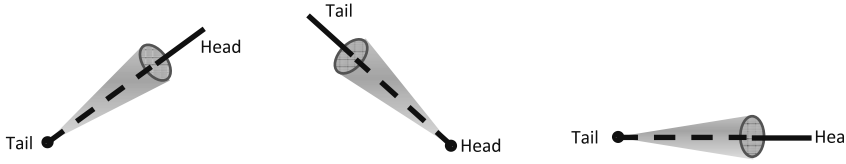
Head Point::= (head_Xhead_Y),

где **tail_X**, **tail_Y**, **head_X** и **head_Y** задают координаты начала и конца элемента.

При необходимости различения элементов одного типа будем снабжать их индексами. Таким образом, элементы O_1 и O_2 – различны, а все элементы (O_x, B_y, F_z) с одинаковыми индексами являются синонимами.

Анализ положения отдельных элементов паттернов данных, представленных на рис. 6, показывает, что для спецификации типа эле-

мента целесообразно ввести в рассмотрение еще один параметр – угол наклона отрезка, соответствующего данному типу элемента. При этом понятно, что элементы одного типа будут расположены внутри некоторой области, которую удобно представлять конусами с вершиной в точке начала множества элементов одного типа. С учетом этого основные классы отрезков для спецификации паттернов данных науки, образования и инновационной деятельности представлены на рис. 7.



«Бычий» (O) элемент «Медвежий» (B) элемент «Боковой» (F) элемент

Рис. 7. Основные классы отрезков для спецификации паттернов данных науки, образования и инновационной деятельности

Для задания области экземпляров конкретных типов элементов в онтологической модели необходимо задать аксиомы принадлежности экземпляра типу. Для этого целесообразно использовать формулу для вычисления угла наклона отрезков следующего вида:

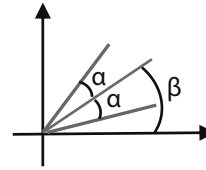
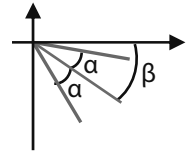
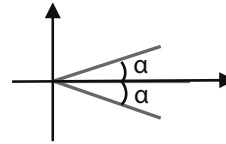
$$\varphi = \cos^{-1}\left(\frac{head\ X - tail\ X}{\sqrt{(head\ X - tail\ X)^2 + (head\ Y - tail\ Y)^2}}\right);$$

и формулы вычисления ограничений на экземпляры отрезков разных типов, представленные в табл. 2.

Теперь у нас имеется вся информация, необходимая для построения онтологической модели исходных данных, которые будут использоваться в рамках семантической интерпретации паттернов данных.

Согласно результатам, полученным в процессе выполнения первого этапа НИР, в рамках которой проводилось настоящее исследование, для построения онтологической модели исходных данных семантической интерпретации паттернов данных использовался инструментарий онтологического инжиниринга Protégé [23]. Экранные формы результатов моделирования представлены на рис. 8.

Таблица 2. Ограничения на экземпляры отрезков для спецификации паттернов данных

Тип элемента	Графическое представление	Ограничения
«Бычий» (O) элемент		$\beta + \alpha \geq \varphi \geq \beta - \alpha$
«Медвежий» (B) элемент		$\beta + \alpha \geq \varphi \geq \beta - \alpha$
«Боковой» (F) элемент		$\alpha \geq \varphi \geq -\alpha$

В соответствии с общей схемой обработки данных первичная информация, полученная в результате построения паттернов данных, хранится в таблицах MS Excel. Поэтому библиотека формирования примитивов (рис. 5) представляется совокупностью процедур конвертации данных из этих таблиц в экземпляры онтологической модели исходных данных семантической интерпретации паттернов. При этом валидация исходных данных осуществляется средствами инструментария Protégé. В данном случае для этого могут использоваться валидаторы онтологий HermiT 1.3.3 или FaCT++ [24, 25].

При указанных выше условиях блок формирования терминальных PDL-цепочек (рис. 5) реализуется как конвертор экземпляров онтологической модели исходных данных семантической интерпретации паттернов в текстовую строку, которая поступает на вход распознавателя PDL-цепочек.

Соответствующая грамматика языка паттернов данных обсуждается в следующем подразделе.

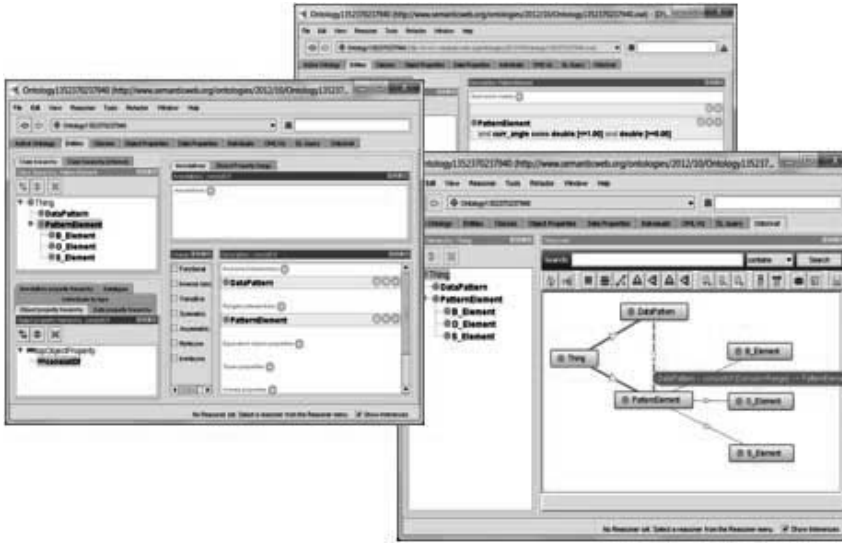


Рис. 8. Экранные формы результатов онтологического моделирования исходных данных для семантической интерпретации паттернов данных

3.2. Грамматика семантической интерпретации паттернов данных

Следуя общим принципам построения распознающих грамматик [21], определим грамматику языка паттернов G_{DP} данных упорядоченной четверкой вида

$G_{DP} = \langle N, T, P, S \rangle$,

где T – множество терминальных символов грамматики (в нашем случае это экземпляры отрезков, которые специфицируют примитивы паттернов данных), N – множество нетерминалов, фиксирующих изображения отдельных фрагментов паттернов данных и сами паттерны данных, S – начальный символ грамматики G_{DP}, причем $S \in N$, а P – множество правил вывода грамматики G_{DP}.

В нашем случае в терминальный словарь входят объекты трех типов – ***O***-, ***B***- и ***F***-элементы, из которых, собственно, и формируются терминальные цепочки PDL-описания разных типов паттернов. Таким образом, будем полагать, что $T = \{O_i, B_j, F_k\}$, где O_i – элементы класса «бычьих» отрезков, B_j – элементы класса «медвежьих» отрезков и F_k – элементы класса «боковых» отрезков соответственно.

В качестве начального символа грамматики выберем понятие паттерна данных (DataPattern) и перейдем к построению множества правил вывода $\boldsymbol{P}$ грамматики $\boldsymbol{G_{DP}}$, поскольку оно, в конечном счете, фиксирует и множество нетерминальных символов $\boldsymbol{N}$ этой грамматики.

В соответствии с целями и задачами настоящей работы, основу для семантической интерпретации паттернов данных науки, образования и инновационной деятельности могут дать типовые изображения кластеров, полученных в результате классификации множества паттернов, сформированных в результате применения строгих математических методов, рассмотренных в соответствующих разделах и подразделах [8].

Как показывает анализ результатов кластеризации (рис. 6), для дальнейшего распознавания и интерпретации целесообразно провести агрегацию паттернов данных внутри каждого кластера и выбрать для дальнейшего анализа центроиды кластеров, представленные на рис. 9.

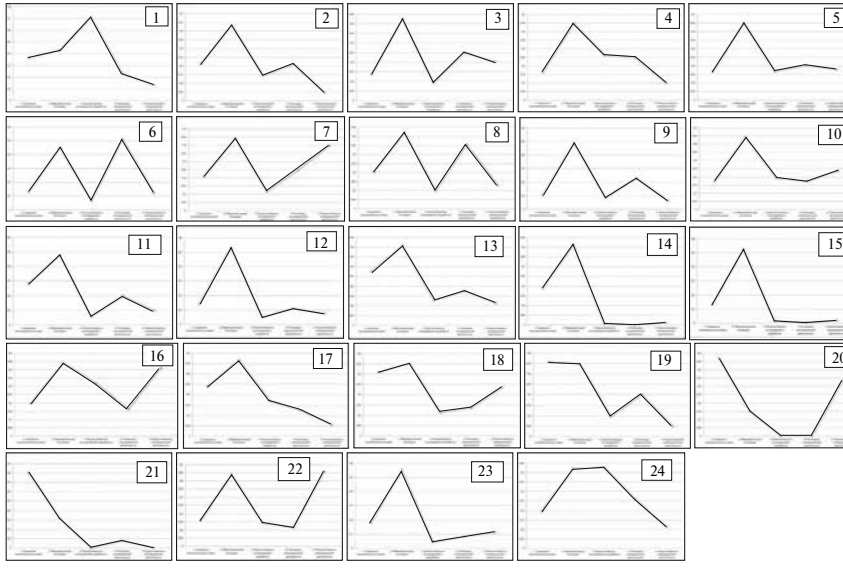


Рис. 9. Агрегация кластеров паттернов данных

В качестве первого приближения к дальнейшему обобщению полученных агрегатов примем допущение, что одинаковыми являются паттерны данных, имеющие одинаковую структуру изображений. В таком случае можно выделить из представленных выше центроидов кластеров следующие 14 групп типовых изображений:

$\mathbf{G_1}$ ={2, 5, 9, 11, 12, 13}; $\mathbf{G_2}$ ={10, 14, 15};$\mathbf{G_3}$ ={6, 8};$\mathbf{G_4}$ ={7, 23}; $\mathbf{G_5}$ ={1};$\mathbf{G_6}$ ={4}; $\mathbf{G_7}$ ={16};$\mathbf{G_8}$ ={17};$\mathbf{G_9}$ ={18};$\mathbf{G_{10}}$ ={19}; $\mathbf{G_{11}}$ ={20};$\mathbf{G_{12}}$ ={21};$\mathbf{G_{13}}$ ={22}; $\mathbf{G_{14}}$ ={24},

которые после преобразования в PDL-цепочки дают следующее множество «текстов» языка, порождаемого грамматикой G_{DP}:

L_{DP} = {$O_1 + B_2 + O_3 + B_4$; $O_1 + B_2 + B_3 + O_4$; $O_1 + B_2 + O_3 + B_4$; $O_1 + O_2 + B_3 + B_4$; $O_1 + B_2 + B_3 + B_4$; $O_1 + B_2 + O_3 + O_4$; $O_1 + B_2 + B_3 + O_4$; $O_1 + B_2 + B_3 + B_4$; $O_1 + B_2 + O_3 + O_4$; $B_1 + B_2 + O_3 + B_4$; $B_1 + B_2 + F_3 + O_4$; $B_1 + B_2 + O_3 + B_4$; $O_1 + B_2 + B_3 + O_4$; $O_1 + F_2 + B_3 + B_4$}

После удаления из множества L_{DP} одинаковых цепочек нетрудно получить PDL-цепочки, соответствующие каждому из 8 базисных паттернов (табл. 3), а после выделения общих подстрок – сформировать множество правил порождающей грамматики G_{DP}, приведенное ниже.

DataPattern$\rightarrow O_1 X \mid B_1 Y$

$X \rightarrow B_2 X' \mid O_2 X' \mid F_2 X'$; $X' \rightarrow O_3 X'' \mid B_3 X''$; $X'' \rightarrow O_4 \mid B_4$

$Y \rightarrow B_2 Y'$; $Y' \rightarrow O_3 Y'' \mid F_3 Y''$; $Y'' \rightarrow B_4$

Анализ правил порождения PDL-представлений паттернов данных показывает, что для их распознавания можно использовать следующие правила:

O_4 (tail(x_1, y_1), head(x_2, y_2)) $\rightarrow X''$ (tail(x_1, y_1), head(x_2, y_2))

B_4 (tail(x_1, y_1), head(x_2, y_2)) $\rightarrow X''$ (tail(x_1, y_1), head(x_2, y_2))

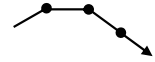
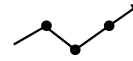
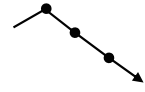
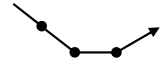
O_3 (tail(x_1, y_1), head(x_2, y_2)) X'' (tail(x_2, y_2), head(x_3, y_3)) $\rightarrow X'$(tail(x_1, y_1), head(x_3, y_3))

B_3 (tail(x_1, y_1), head(x_2, y_2)) X'' (tail(x_2, y_2), head(x_3, y_3)) $\rightarrow X'$(tail(x_1, y_1), head(x_3, y_3))

F_2 (tail(x_1, y_1), head(x_2, y_2)) X'(tail(x_2, y_2), head(x_3, y_3)) $\rightarrow X$ (tail(x_1, y_1), head(x_3, y_3))

O_2 (tail(x_1, y_1), head(x_2, y_2)) X'(tail(x_2, y_2), head(x_3, y_3)) $\rightarrow X$ (tail(x_1, y_1), head(x_3, y_3))

Таблица 3. PDL-цепочки базисных паттернов данных науки, образования и инновационной деятельности

№	Название паттерна	Графическое представление	PDL-представление паттерна
1	**Pattern-1**		$O_1 + O_2 + B_3 + B_4$
2	**Pattern-2**		$O_1 + F_2 + B_3 + B_4$
3	**Pattern-3**		$O_1 + B_2 + O_3 + O_4$
4	**Pattern-4**		$O_1 + B_2 + O_3 + B_4$
5	**Pattern-5**		$O_1 + B_2 + B_3 + O_4$
6	**Pattern-6**		$O_1 + B_2 + B_3 + B_4$
7	**Pattern-7**		$B_1 + B_2 + O_3 + B_4$
8	**Pattern-8**		$B_1 + B_2 + F_3 + O_4$

B_2 (tail(x_1, y_1), head(x_2, y_2)) X'(tail(x_2, y_2), head(x_3, y_3)) → X (tail(x_1, y_1), head(x_3, y_3))

B_4 (tail(x_1, y_1), head(x_2, y_2)) → Y'' (tail(x_1, y_1), head(x_2, y_2))

F_3 (tail(x_1, y_1), head(x_2, y_2)) Y'' (tail(x_2, y_2), head(x_3, y_3)) → Y'(tail(x_1, y_1), head(x_3, y_3))

O_3 (tail(x_1, y_1), head(x_2, y_2)) Y'' (tail(x_2, y_2), head(x_3, y_3)) → Y'(tail(x_1, y_1), head(x_3, y_3))

B_2 (tail(x_1, y_1), head(x_2, y_2)) Y'(tail(x_2, y_2), head(x_3, y_3)) → Y (tail(x_1, y_1), head(x_3, y_3))

B_1 (tail(x_1, y_1), head(x_2, y_2)) Y (tail(x_2, y_2), head(x_3, y_3)) → ***DataPattern***(tail(x_1, y_1), head(x_3, y_3))

O_1 (tail(x_1, y_1), head(x_2, y_2)) X (tail(x_2, y_2), head(x_3, y_3)) → ***DataPattern***(tail(x_1, y_1), head(x_3, y_3))

Нетрудно показать, что с помощью приведенных выше правил распознаются все паттерны данных и только они.

3.3. Алгоритм интерпретации

Правила распознавания языка L_{DP} обладают важным свойством – все они имеют структуру вида $ab \rightarrow A$ или $aA \rightarrow B$, где a и b – терминалы, а A и B – нетерминалы.

Как известно, языки, порождаемые с помощью правил такого вида, относятся к классу языков типа «3» по Хомскому, а это, в свою очередь, обеспечивает реализацию распознавателей таких языков с помощью конечных автоматов [26]. При этом важно, что емкостная сложность распознавания цепочек равна 1, а временная сложность – линейно зависит от длины распознаваемой цепочки.

Учитывая вышесказанное, в заключение настоящего подраздела рассмотрим алгоритм реализации распознавателя паттернов данных науки, образования и инновационной деятельности.

Будем считать, что каждому правилу вида $ab \rightarrow A$ ставится в соответствие фрагмент конечного автомата, представленный на рис. 10(а), а правилу вида $aA \rightarrow B$ – фрагмент конечного автомата, представленный на рис. 10(б).

Рис. 10. Базовые фрагменты конечного автомата для распознавания паттернов данных науки, образования и инновационной деятельности

Тогда общую структуру конечного автомата, распознающего все PDL-цепочки изображения паттернов данных науки, образования и инновационной деятельности, можно описать графом, представленным на рис. 11.

Программная реализация такого конечного автомата является достаточно очевидной и потому здесь не обсуждается.

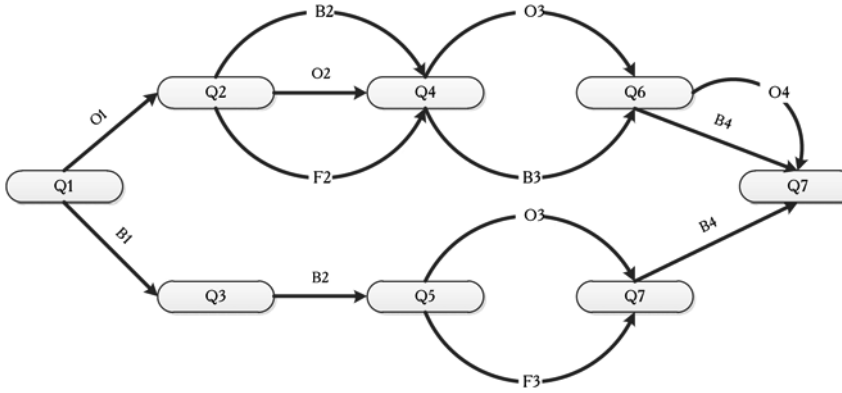


Рис. 11. Конечный автомат для распознавания паттернов данных науки, образования и инновационной деятельности

Заключение

В настоящей работе представлены результаты разработки структурного подхода к обработке изображений паттернов данных науки, образования и инновационной деятельности в регионах РФ.

При этом основное внимание уделено изложению методов и алгоритмов построения грамматик семантической интерпретации паттернов данных, предложена грамматика распознавания паттернов данных науки, образования и инновационной деятельности, а также эффективный конечный автомат для распознавания PDL-цепочек представления изображений паттернов данных.

Как показывает анализ последних работ [27, 28], данное научно-техническое направление переживает в настоящее время новый всплеск исследований и разработок, обусловленный тем, что семантическая интерпретация результатов анализа паттернов данных предполагает явную спецификацию получающихся здесь структур.

Предложенный структурный подход может быть использован не только для распознавания и классификации паттернов данных, но и для их постобработки в процессе вывода на знаниях.

Литература

1. Индикаторы науки: 2008. Статистический сборник. М.: ГУ ВШЭ, 2008 (http://issek.hse.ru/news/49369919.html).

2. Индикаторы образования: 2008. Статистический сборник. М.: ГУ ВШЭ, 2008 (http://issek.hse.ru/news/49370471.html).

3. Индикаторы инновационной деятельности: 2008. Статистический сборник. М.: ГУ ВШЭ, 2008 (http://issek.hse.ru/news/49369377.html).

4. Рейтинг инновационного развития субъектов Российской Федерации: аналитический доклад / под ред. Л.М. Гохберга. М.: Национальный исследовательский университет «Высшая школа экономики», 2012.

5. Aleskerov F., Alper C.E. A clustering approach to some monetary facts: a long-run analysis of cross-country data // The Japanese Economic Review. 2000. Vol. 51. No. 4. P. 555–567.

6. Few S. Multivariate Analysis Using Parallel Coordinates, 2006 (http://www.perceptualedge.com/articles/b-eye/parallel_coordinates.pdf).

7. Aleskerov F., Nurmi H. A Method for Finding Patterns of Party Support and Electoral Change: An Analysis of British General and Finnish Municipal Elections // Mathematical and Computer Modelling. 2008. P. 1225–1253.

8. Алескеров Ф.Т., Гохберг Л.М., Егорова Л.Г., Мячин А., Сагиева Г.С. Анализ данных науки, образования и инновационной деятельности с использованием методов анализа паттернов: препринт WP7/2012/07. М.: Изд. дом Высшей школы экономики, 2012.

9. Narasimhan R.N. Syntax-directed interpretation of classes of pictures, Comm. ACM, 9, 166—173 (1966). (Рус. перевод: Нарасимхан Р. Синтаксическая интерпретация классов изображений // Автоматический анализ сложных изображений, 1969).

10. Narasimhan R. On the description, generation, and recognition of classes of pictures // Automatic Interpretation and Classification of Images / A. Grasselli (ed.). N. Y.: Academic Press, 1969.

11. Преображенский А.Б., Хорошевский В.Ф. Структурная модель восприятия окружающей среды // Вопросы радиоэлектроники. Серия «Общетехническая». Вып. 13, 1971.

12. Stallings W.W. Recognition of printed Chinese characters by automatic pattern analysis, Comput. Graphics and Image Process, 1, 47–65 (1972).

13. Фу К. Структурные методы в распознавании образов: пер. с англ. (ред. М.А. Айзерман). М.: Мир, 1977.

14. Narasimhan R. A Linguistic Approach to Pattern Recognition, Rep 121. Digital Comput Lab, UnivofIllinois, Urbana, 1962 (Русский перевод: Нарасимхан Р. Лингвистический подход к распознаванию образов // Автоматический анализ сложных изображений. М.: Мир, 1969).

15. Ledley R.S. High-Speed Automatic Analysis of Biomedical Pictures. // Science. October 1964. No. 146. P. 216–223.

16. Kirsch R.A. Computer Interpretation of English Text and Picture Patterns // IEEE Trans. Elec. Comp. EC-13. August 1964. P. 363–376.

17. Shaw A.C. A formal picture description scheme as a basis for picture processing system // Information and Control. 1969. No. 14. P. 9–52.

18. Swain P.H., Fu K.S. Stochastic programmed grammars for syntactic pattern recognition, Pattern Recognition. 1972. No. 4. P. 83–100.

19. Dimitrov V.D. Multi-layered stochastic languages and stochastic grammars for syntactic pattern recognition, Int. Symp. Theor. Problems and Syst. for Pattern and Situation Recognition, Varna, Bulgaria, October 8–12, 1972.

20. Goldfarb L. Pattern representation and the future of pattern recognition: A program for action // Proc. of ICPR 2004 satellite workshop. St Catharine's College, Cambridge, UK, Aug 22, 2004.

21. Гладкий А.Б. Формальные грамматики и языки. М.: Наука, 1973.

22. Вудс В.А. Сетевые грамматики для анализа естественных языков // Кибернетический сборник. Новая серия. Вып. 13. М.: Мир, 1976.

23. Protege Homepage, http://protege.stanford.edu/.

24. Glimm B., Horrocks I., Motik B., Stoilos G. Optimising Ontology Classification. In Proc. of the 9th Int. Semantic Web Conf. (ISWC 2010). Vol. 6496 of LNCS. P. 225–240, Shanghai, China, November 7–11 2010.

25. Tsarkov D., Horrocks I. Efficient reasoning with range and domain constraints.In Proc. of the 2004 Description Logic Workshop (DL 2004), Vol. 104. P. 41–50. CEUR (http://ceur-ws.org/), 2004.

26. Mylopoulos J. On the application of formal language and automata theory to pattern recognition // Pattern Recognition. 1972. No. 4. P. 37–52.

27. Hancock E.R., Wilson R.C., Windeatt T., Ulusoy I., Escolano F. (eds.). Proceedings Joint IAPR International Workshop "Structural, Syntactic, and Statistical Pattern Recognition", SSPR&SPR 2010, Cesme, Izmir, Turkey, August 18–20, 2010.

28. Gimel'farb G.L., Hancock E.R., Imiya A., Kuijper A., Kudo M., Omachi S., Windeatt T., Yamada K. (eds.): Proceedings Joint IAPR International Workshop "Structural, Syntactic, and Statistical Pattern Recognition", SSPR&SPR 2012, Hiroshima, Japan, November 7–9, 2012.

Препринт WP7/2012/08
Серия WP7
Математические методы анализа решений
в экономике, бизнесе и политике

Хорошевский Владимир Федорович

Об одном методе семантической интерпретации паттернов данных на основе структурного подхода

Зав. редакцией оперативного выпуска *А.В. Заиченко*
Технический редактор *Ю.Н. Петрина*

Отпечатано в типографии
Национального исследовательского университета
«Высшая школа экономики» с представленного оригинал-макета

Формат 60×84 $^{1}/_{16}$. Тираж 20 экз. Уч.-изд. л. 1,6

Усл. печ. л. 1,6. Заказ № . Изд. № 1526
Национальный исследовательский университет
«Высшая школа экономики»
125319, Москва, Кочновский проезд, 3
Типография Национального исследовательского университета
«Высшая школа экономики»