

BIG DATA: Технологии добычи данных

Кравченко Ю.А.

- ❖ **повышение эффективности прикладных систем добычи данных на основе развития моделей, методов и алгоритмов семантического поиска, классификации, кластеризации, структурирования и интеграции данных.**

**Объект
исследования**
процессы
добычи данных
в условиях big
data

Предмет исследования
теоретические основы
разработки моделей, методов и
алгоритмов добычи данных

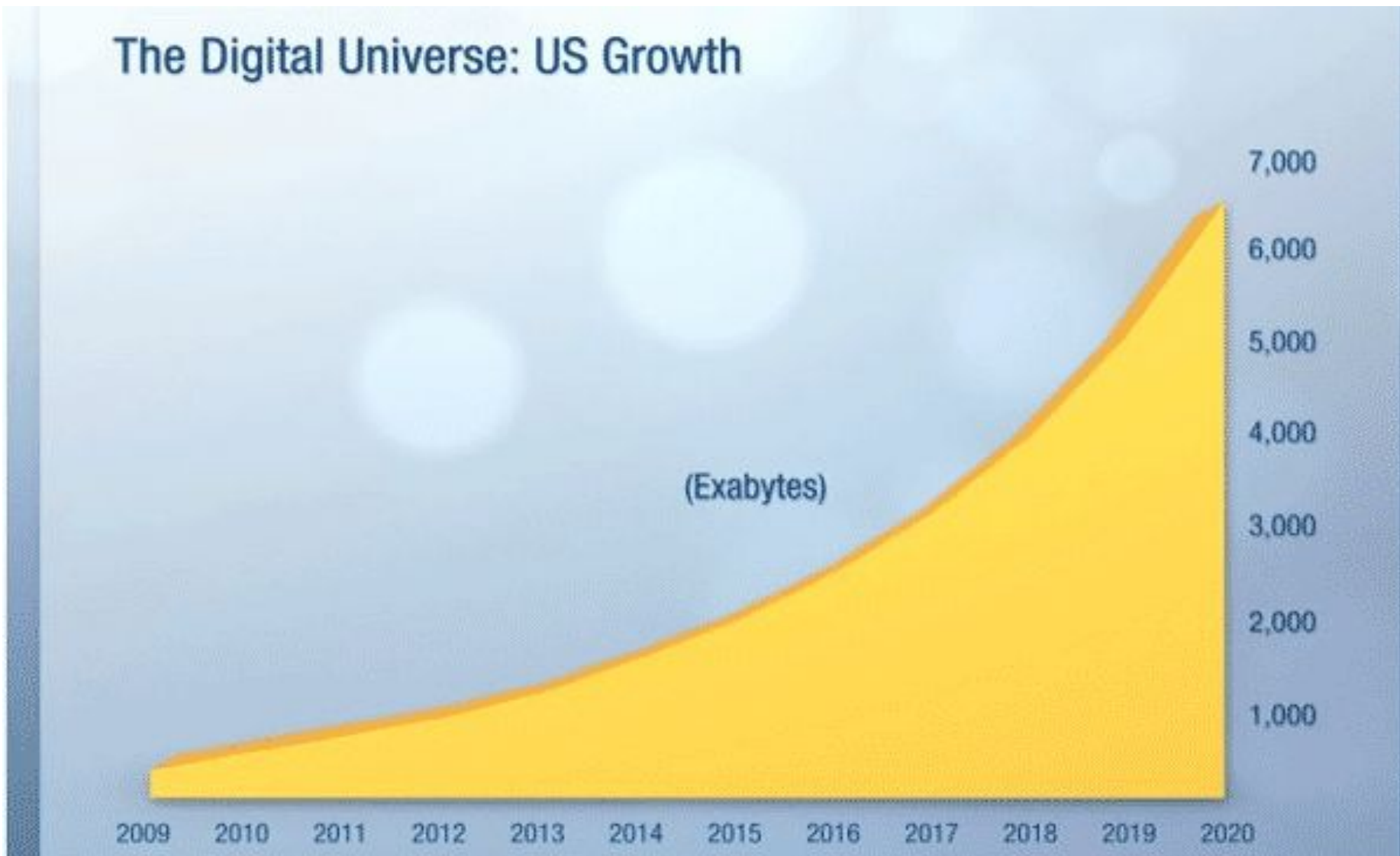
Инфографика схемы роста объемов информации

4



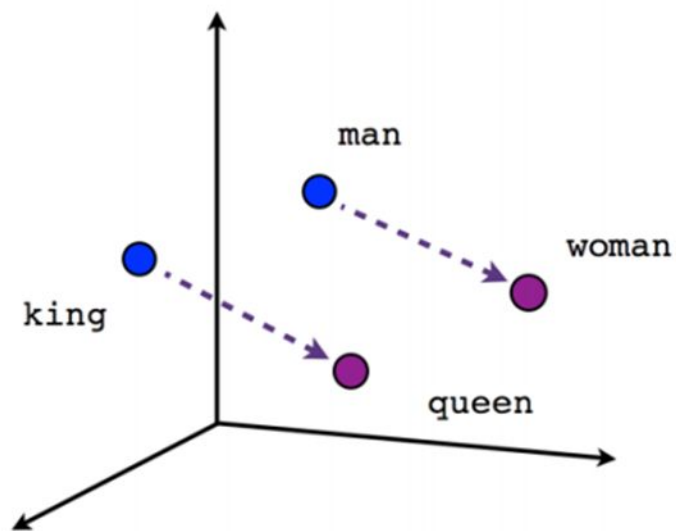
Рост объемов данных

5

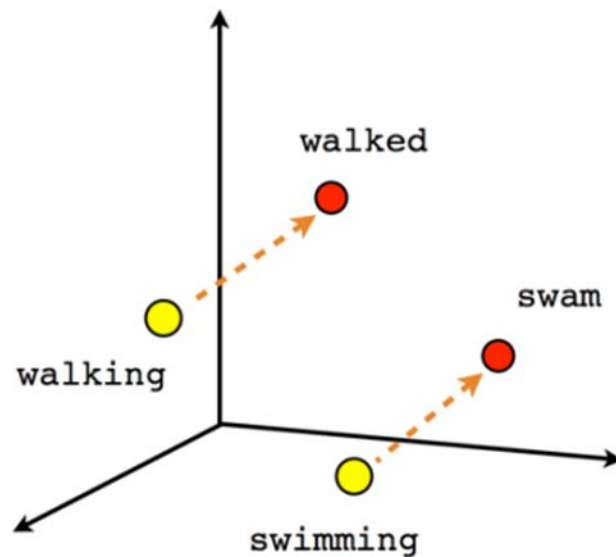


Векторная репрезентация слов

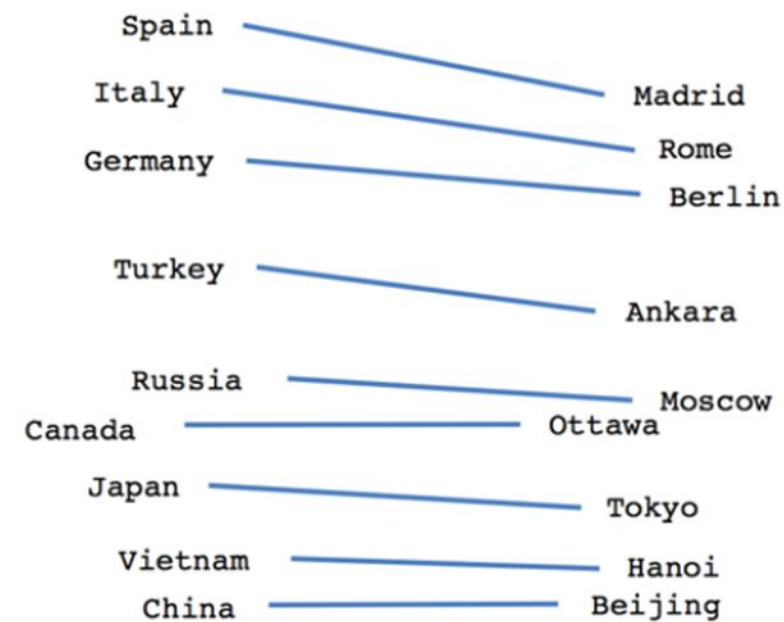
6



Male-Female



Verb tense



Country-Capital

Векторная репрезентация слов

7

Features:		<i>is a word</i>		<i>is a noun</i>		<i>is singular</i>		<i>related to food</i>		...
king		1		1		1		0		...
queens		1		1		0		0		...
eating		1		0		0		1		...

vector representation of words

Семантический вектор запроса и текста

8

семантический
вектор запроса

300 чисел



300 чисел

семантический
вектор страницы

- ❖ Идея семантического поиска заключается в описании поисковых запросов в виде набора триплетов. Пусть имеется запрос q , состоящий из набора триплетов $T(q)$. В таком случае результатом поиска в источнике знаний будет набор элементов знания $E = \{e_i \mid i \in [1, k]\}$, где k – количество элементов знания e_i , являющихся результатом поиска. Причем, семантические метаданные набора элементов знания $T(e)$ должны удовлетворять следующему условию семантической близости $sim(T(q), T(e))$ с описанием запроса $T(q): sim(e, q) = sim(T(q), T(e)) > \varepsilon$, где $sim(e, q)$ близость запроса q и элемента знания e , а ε – установленное пороговое значение релевантности. Результаты поиска ранжируются по значениям их семантической близости к запросу.

Постановка задачи классификации

10

❖ Пусть X – множество описаний элементов знаний, Y – множество наименований классов. Существует неизвестная целевая зависимость – отображение $y^*: X \rightarrow Y$, значения которой известны только на объектах конечной обучающей выборки $X^m = \{(x_1, y_1), \dots, (x_m, y_m)\}$. Требуется построить алгоритм $a: X \rightarrow Y$, способный классифицировать произвольный объект $x \in X$.

Представим вероятностную постановку задачи классификации знаний, которая считается более общей. Предполагается, что множество пар «элемент знания, класс» $X \times Y$ является вероятностным пространством с неизвестной вероятностной мерой P . Имеется конечная обучающая выборка наблюдений $X^m = \{(x_1, y_1), \dots, (x_m, y_m)\}$, сгенерированная согласно вероятностной мере P . Требуется построить алгоритм $a: X \rightarrow Y$, способный классифицировать произвольный объект $x \in X$.

Постановка задачи структуризации

11



Все системно значимые признаки элементов знания из определенной предметной области разобьем на m классов. Сформируем множество необходимых признаков системной значимости $F = \{F_1 \cup F_2 \cup \dots \cup F_m\}$.

$$F_1 = \{f_{11}, f_{12}, \dots, f_{1(i-1)}, f_{1i}\},$$

где $f_{11}, f_{12}, \dots, f_{1(i-1)}, f_{1i}$ – элементы множества F_1 , задающие 1-ый класс системно значимых признаков для элементов знания некоторой предметной области;

$$F_2 = \{f_{21}, f_{22}, \dots, f_{2(j-1)}, f_{2j}\},$$

где $f_{21}, f_{22}, \dots, f_{2(j-1)}, f_{2j}$ – элементы множества F_2 , задающие 2-ой класс системно значимых признаков для элементов знания некоторой предметной области;

$$F_m = \{f_{m1}, f_{m2}, \dots, f_{m(k-1)}, f_{mk}\},$$

где $f_{m1}, f_{m2}, \dots, f_{m(k-1)}, f_{mk}$ – элементы множества F_m , задающие m класс системно значимых признаков для элементов знания некоторой предметной области.

Постановка задачи структуризации

12



Зададим для каждого анализируемого элемента знания q_z ($z = 1 \dots n$) множество имеющих у него системно значимых признаков $Q_z = \{Q_{11} \cup Q_{12} \cup \dots \cup Q_{nm}\}$, где

$$Q_{11} \subset F_1, Q_{12} \subset F_2, Q_{nm} \subset F_m.$$

Тогда выражение определения соответствия элемента знания системно значимым требованиям предметной области представим в виде:

$$M_0 = Q_z \cap F.$$

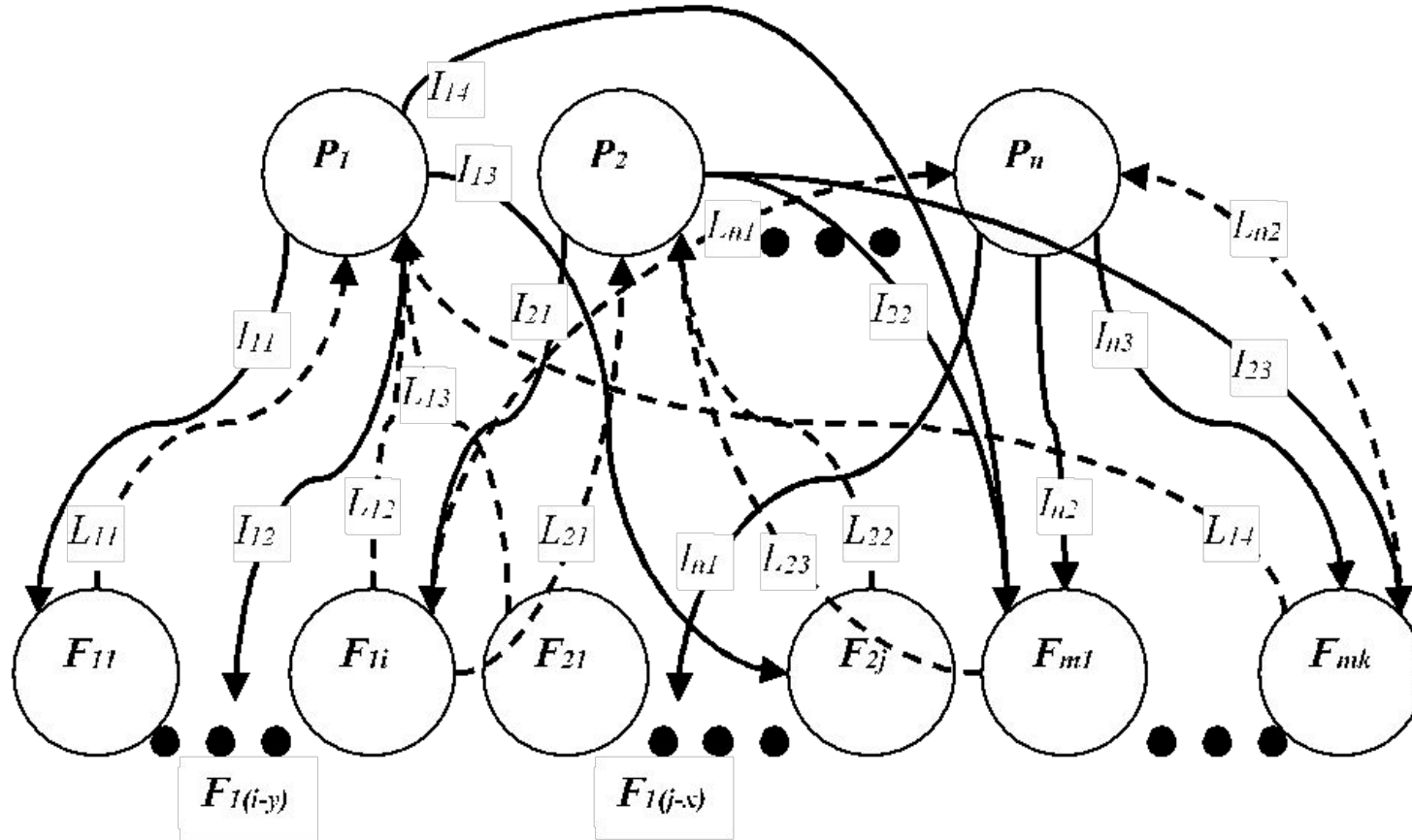
А целевая функция в таком случае примет вид:

$$M_0 = F.$$

Отсутствие заданного тождества указывает на неполное соответствие элемента требованиям, выдвигаемым к системно значимым признакам.

Абстрактный пример структуризации

13



Постановка задачи интеграции

14



Представим постановку задачи системной *интеграции знаний* множества онтологий в виде следующего выражения:

$$|O_i| = \sum_{j=1}^N |O_{ij}|, \quad |O_{ij}| = \overline{1, |O_{ij}|}$$

где O_i – онтограф (ОГ); i – номер предметной области; N – количество предметных областей.

При простой древовидной структуре выражение определения объема знаний в предметной области имеет следующий вид:

$$|O_i| = \sum_{p=1}^P \sum_{d=1}^{Z_{p,d}} |O_{i,p,d}|$$

где $Z_{p,d}$ – степень инцидентности вершины под номером d ; $p = \overline{1, P}$ – количество уровней ОГ; $d = \overline{1, D_p}$ – номер вершины на p -ом уровне ОГ.

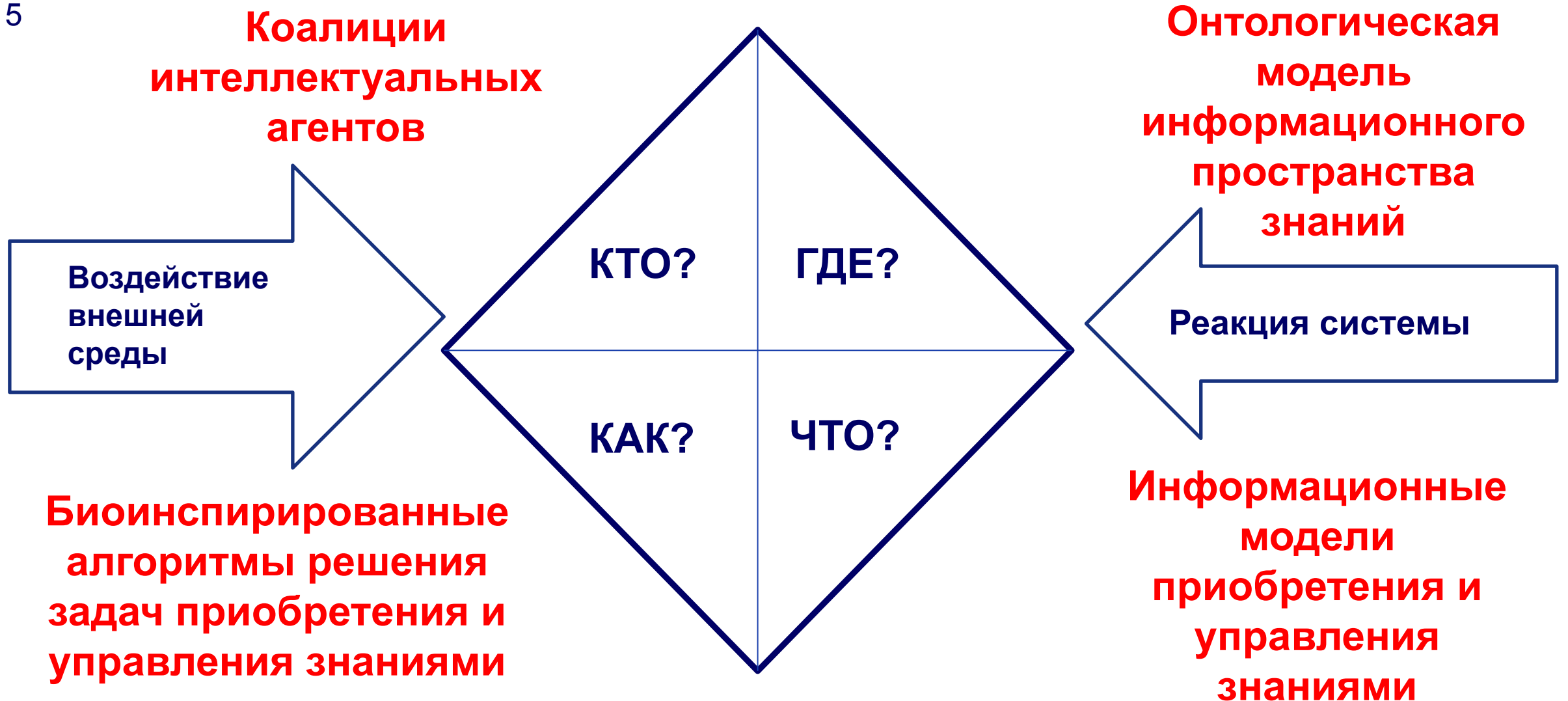
Учёт типов отношений и сложность функций интерпретации приводит к ОГ со взвешенными вершинами и ребрами. Выражение определения объема знаний в этом случае имеет вид:

$$|O_i| = \sum_{d=1}^{Z_{p,d}} \sigma_{d,j} |O_{i,p,d}| + \sum_{d=1}^{Z_{p,d}} \delta_{d,j} |O_{i,p,d}|$$

где γ_d и $\delta_{d,j}$ – значения весовых функций отношений и интерпретации.

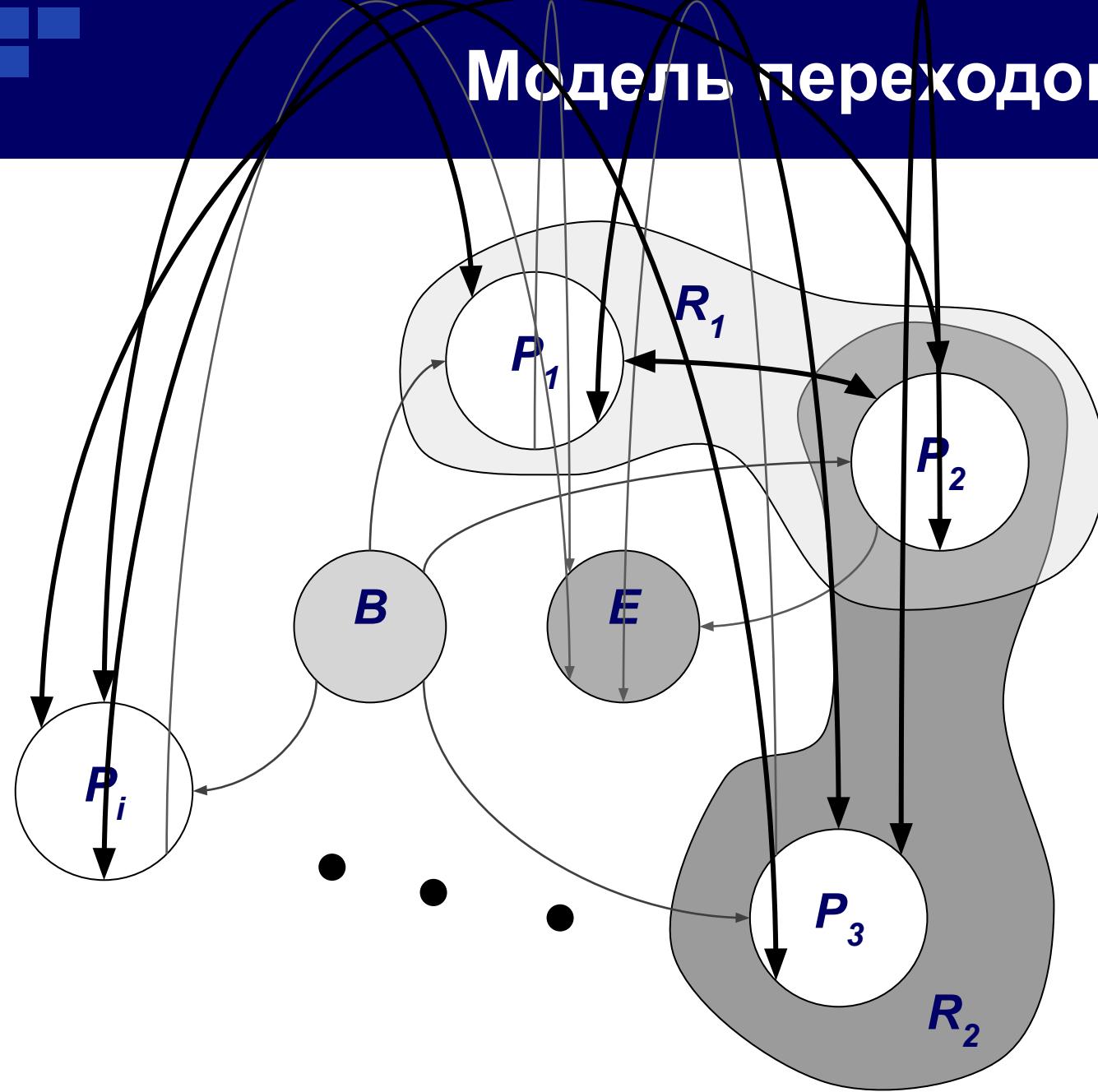
Путь исследования

15



Модель переходов агента

16

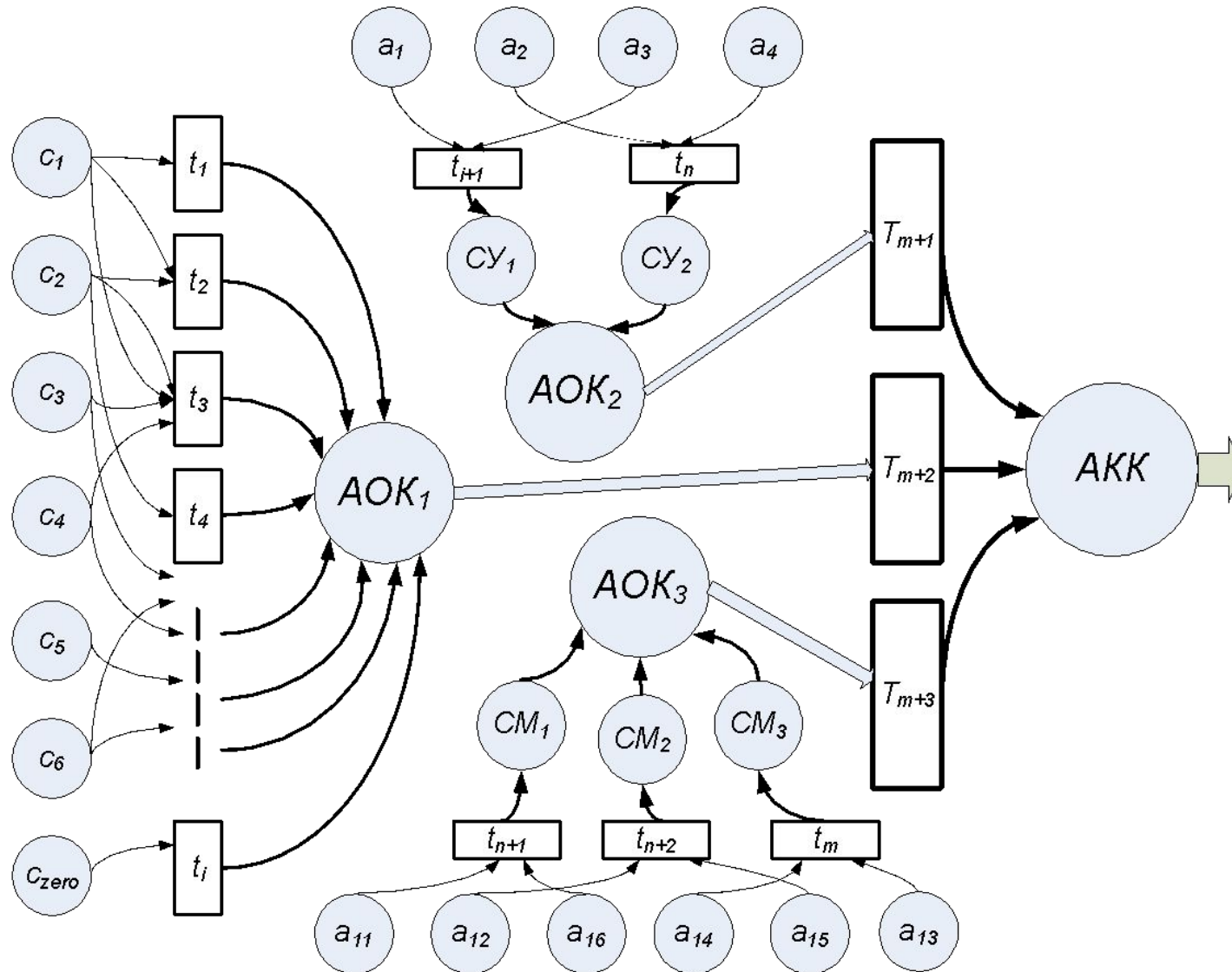


$$G = \langle B, E, P, U \rangle$$

B – вершина начала функционирования агента;
 E – вершина конца функционирования агента;
 P – множество вершин процессов функционирования агента;
 U – множество ребер переходов состояния агента.
Задание гиперграфа на множестве вершин процессов функционирования агента P отображает режимы функционирования агента R .

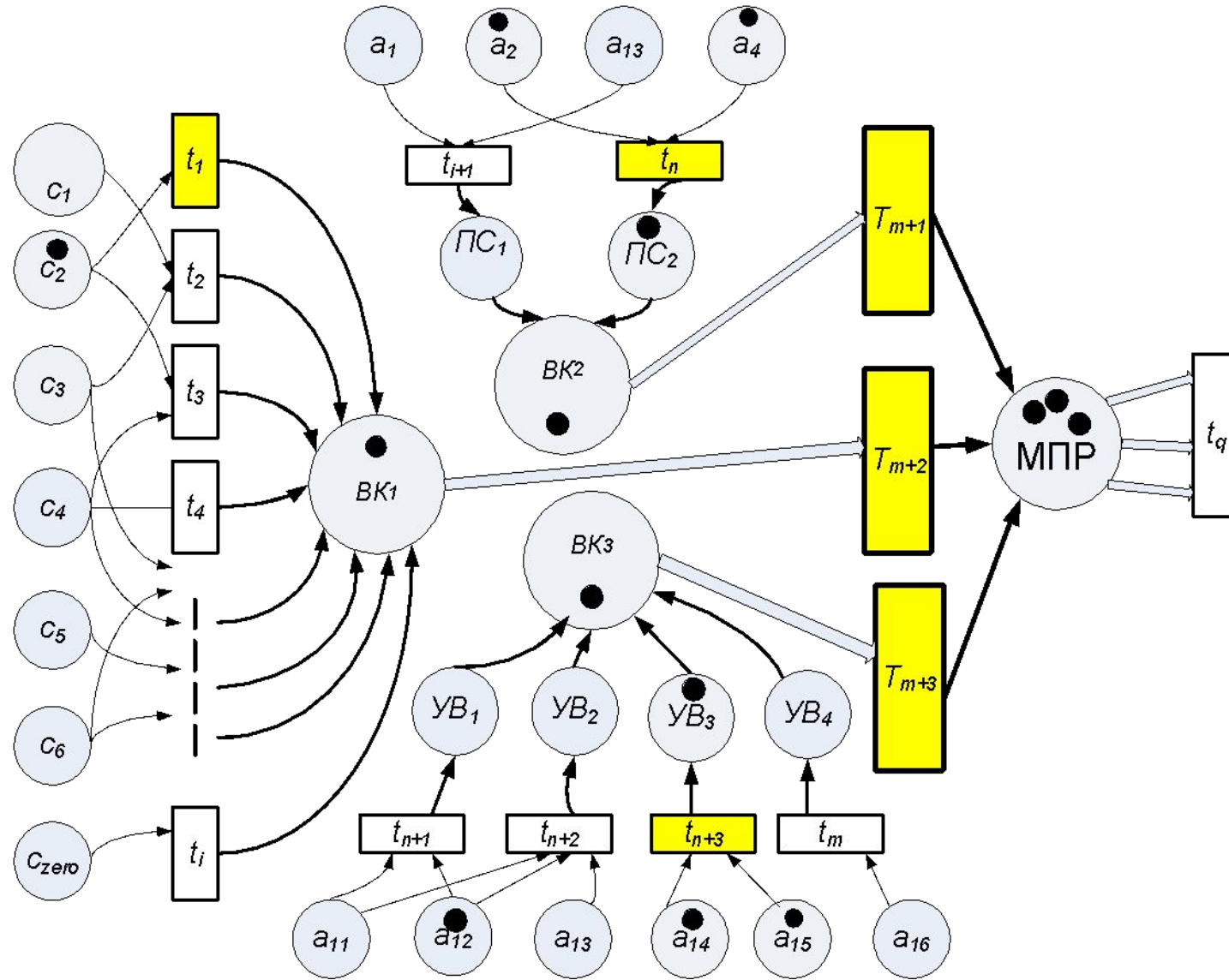
Имитационные модели прецедентов

17



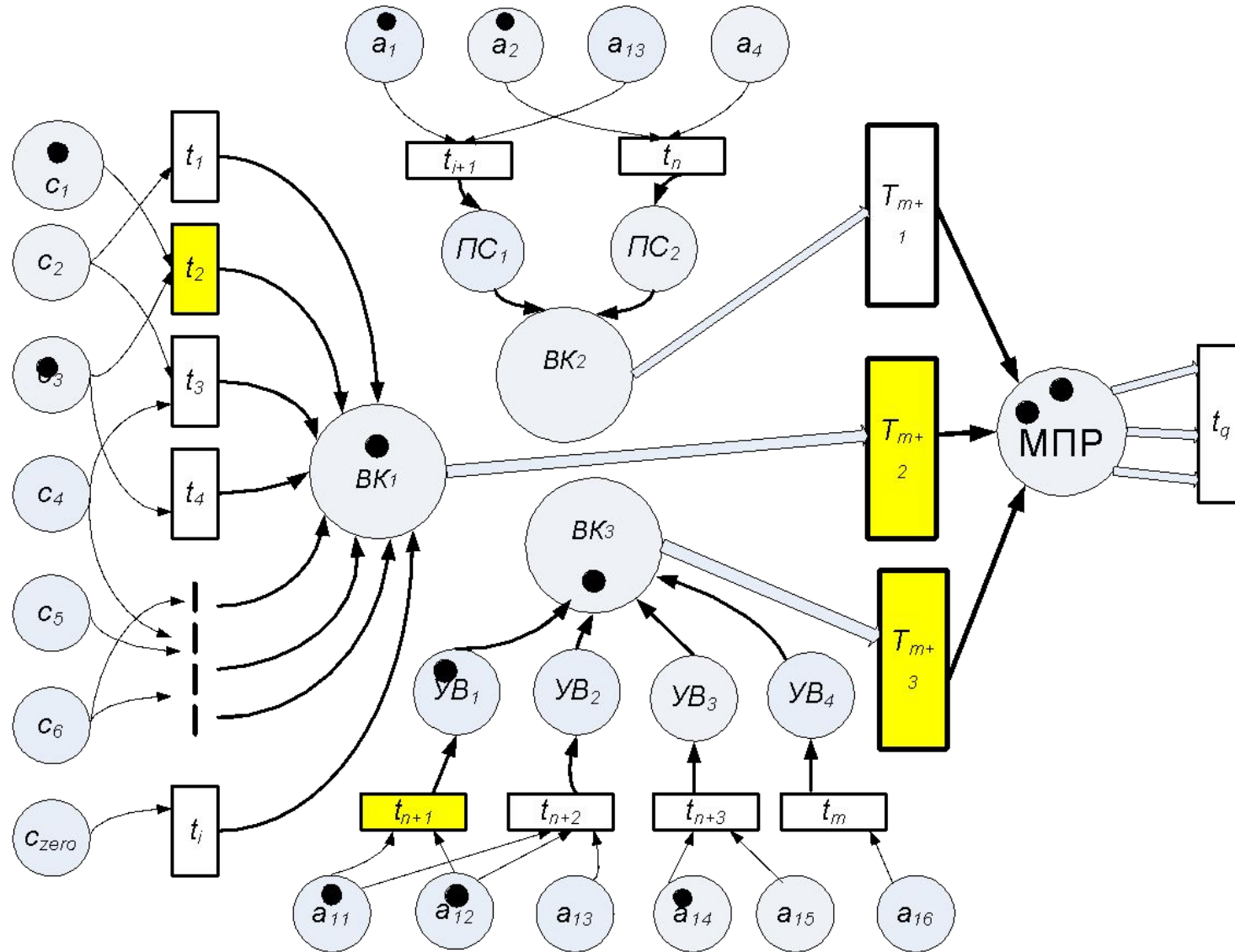
Выполнение условий достижимости

18



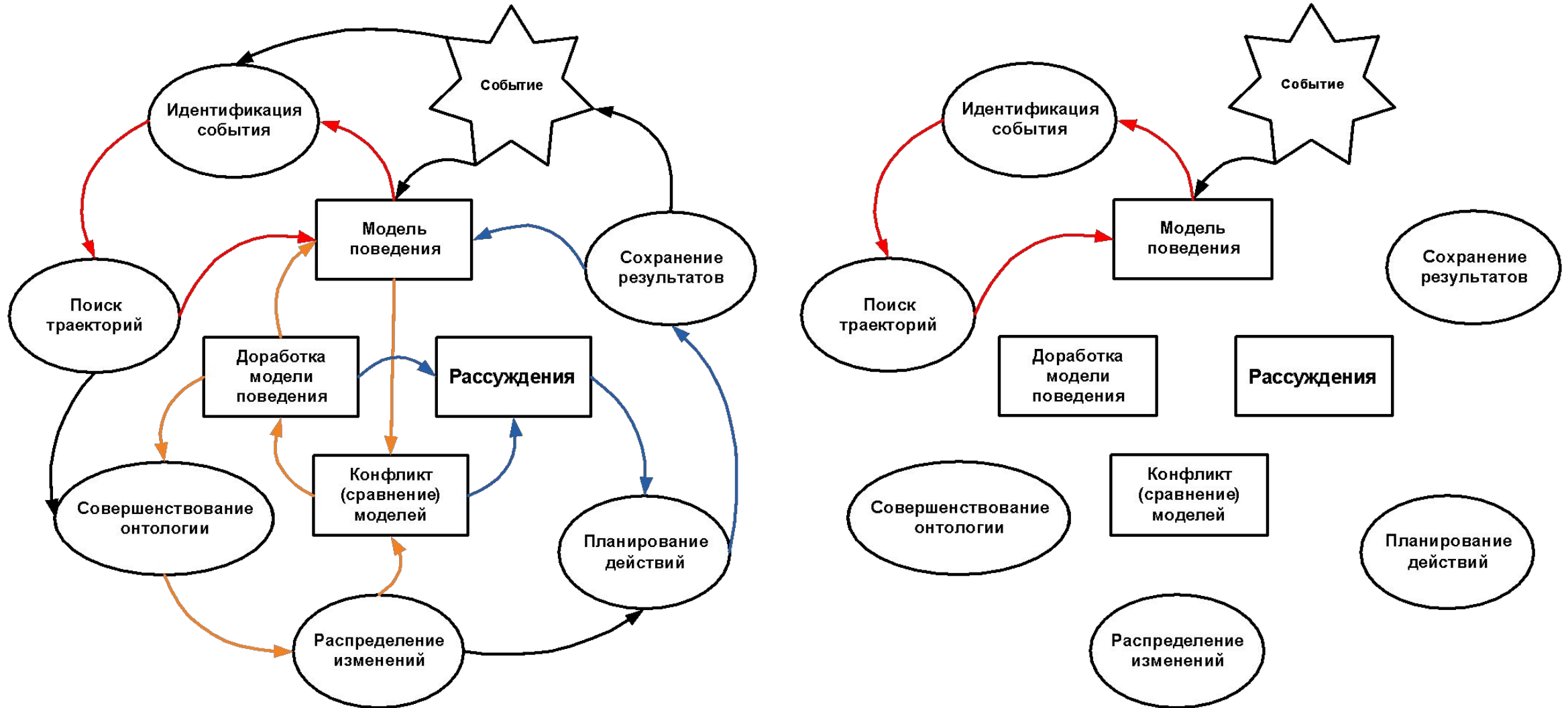
Нарушение условий достижимости

19



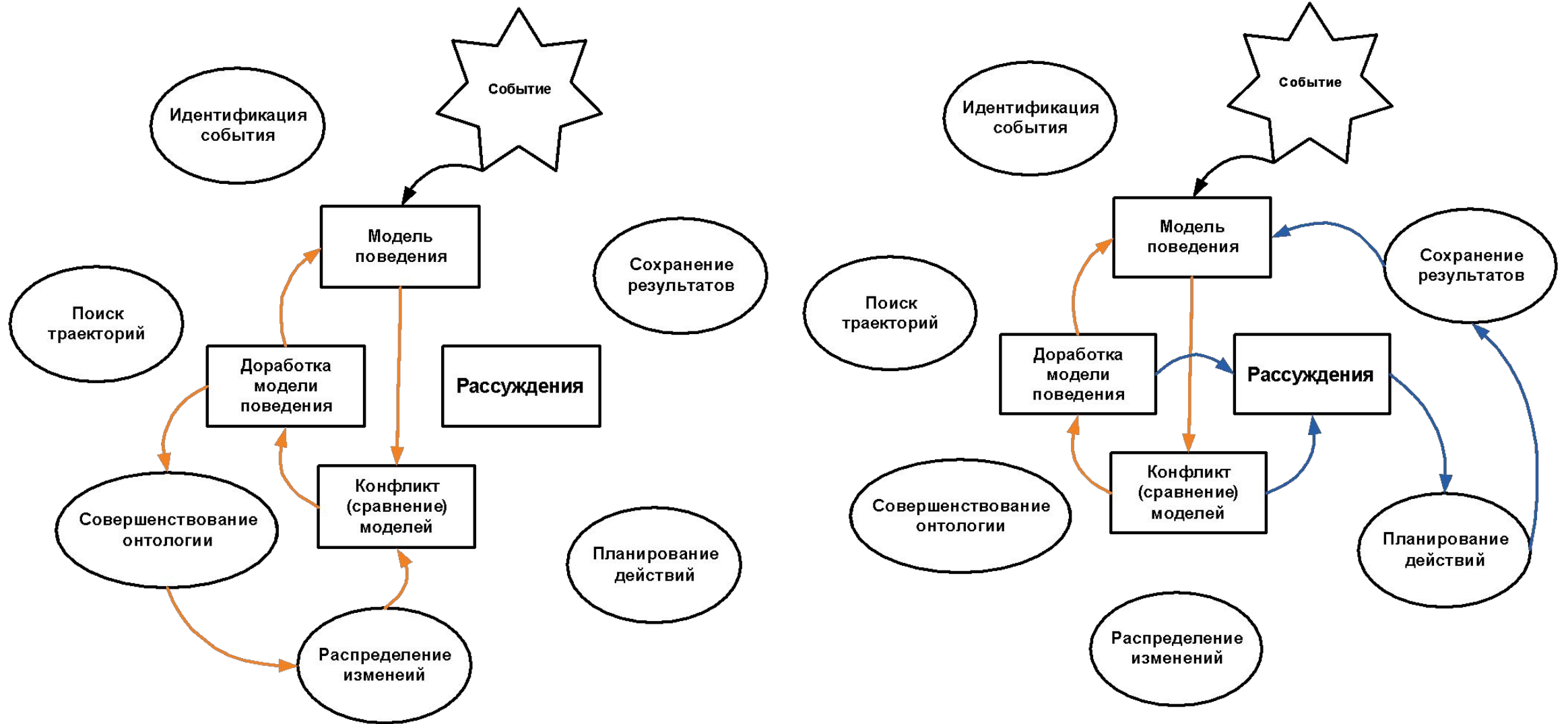
Модели поведения

20



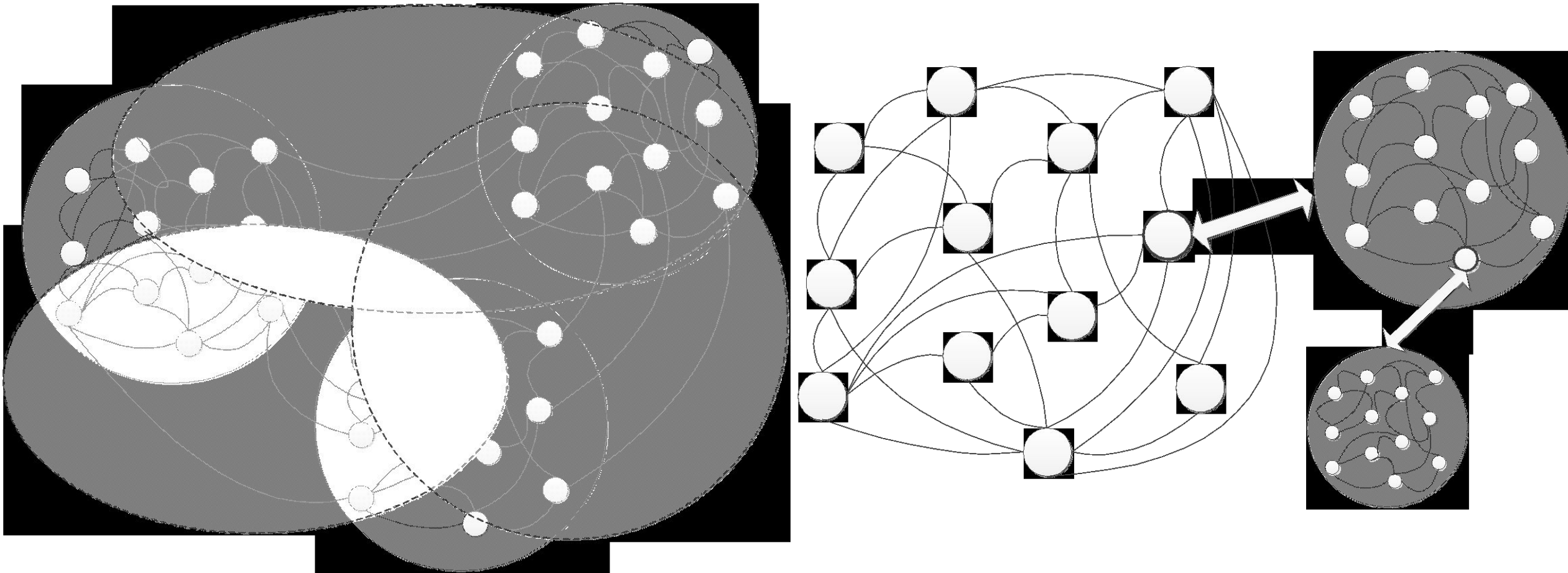
Модели поведения

21



Абстрактная модель среды поиска данных

22

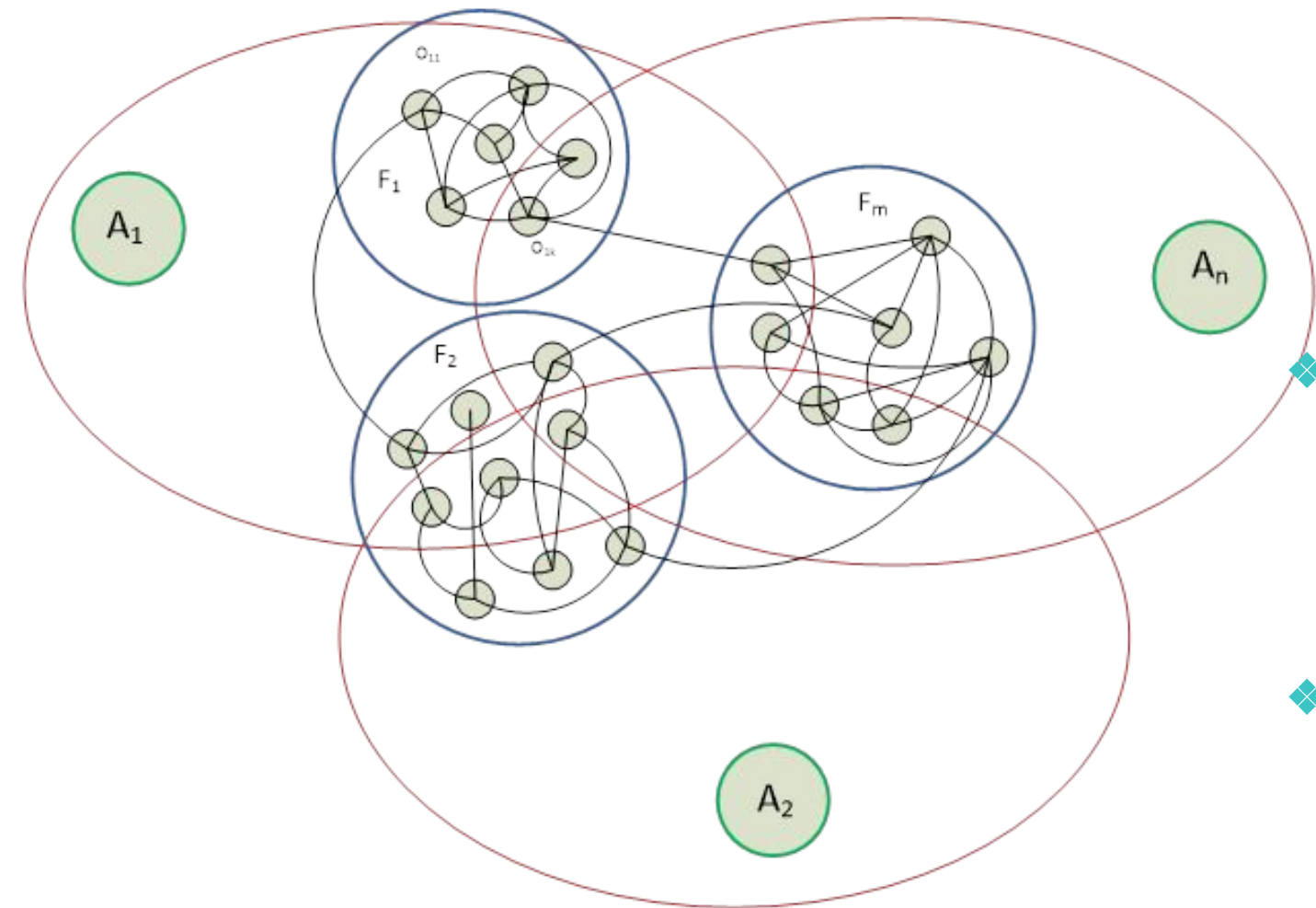


Модель онтологии

- 23 ❖ Онтология O представляет собой знаковую систему $O = \langle P, V, R, C \rangle$, где P – множество понятий (концептов); V – множество экземпляров понятий; R – множество предикатов – атрибутов понятий; C – множество отношений, которые задают следующие виды связи между сущностями:
- ❖ 1. Частичный порядок на множествах P и R , задающий отношения *is-a* – «подкласс-суперкласс».
 - ❖ 2. Отношение между понятиями, которое представляет собой триплет вида $\langle p_1 - r_1 - p_2 \rangle$, где $p_1, p_2 \in P$; $r_1 \in R$.
 - ❖ 3. Отношение между экземплярами, которое представляет собой триплет вида $\langle v_1 - r_1 - v_2 \rangle$, где $v_1, v_2 \in V$; $r_1 \in R$.
 - ❖ 4. Отношение между предикатами, которое представляет собой триплет вида $\langle r_1 - r_i - r_2 \rangle$, где $r_1, r_2, r_i \in R$.

Модель среды поиска данных

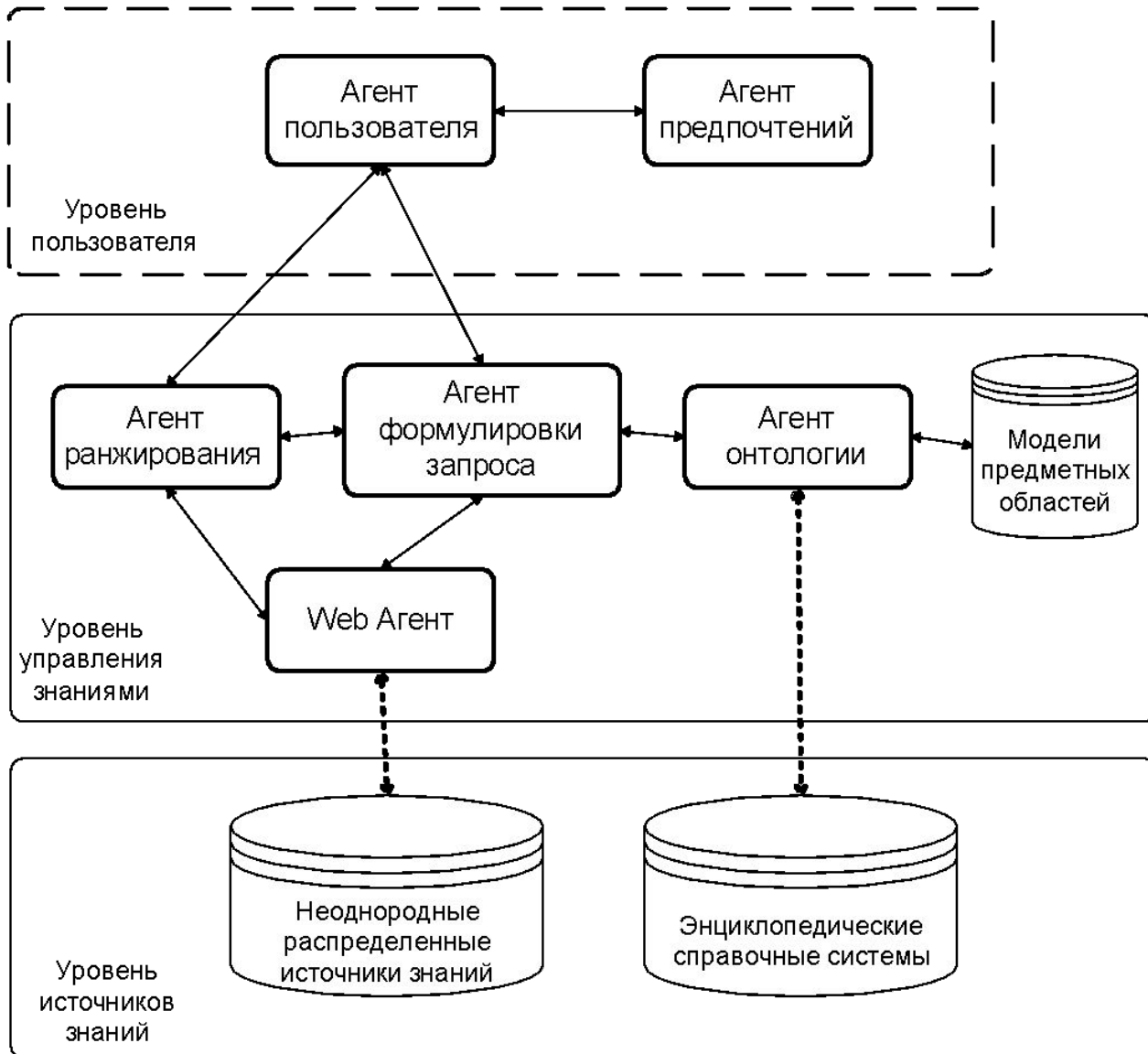
24



- ❖ 1) **акторы (A)** – фильтры знаний, представляющие активные элементы (агенты), создающие предметную область на основе поиска междисциплинарных отношений между разными функциональными областями, способные активизировать, выполнять или управлять информационным процессом;
- ❖ 2) **функциональная область (F)**, представляющая определенный информационный процесс, состоящий из последовательности операций и направленный на достижение поставленной цели;
- ❖ 3) **объекты (O)**, представляющие пассивные сущности, с которыми работает информационный процесс, обладающие собственными атрибутами (измеримыми свойствами) и отношениями между собой.

Модель фильтра данных семантического поиска

25

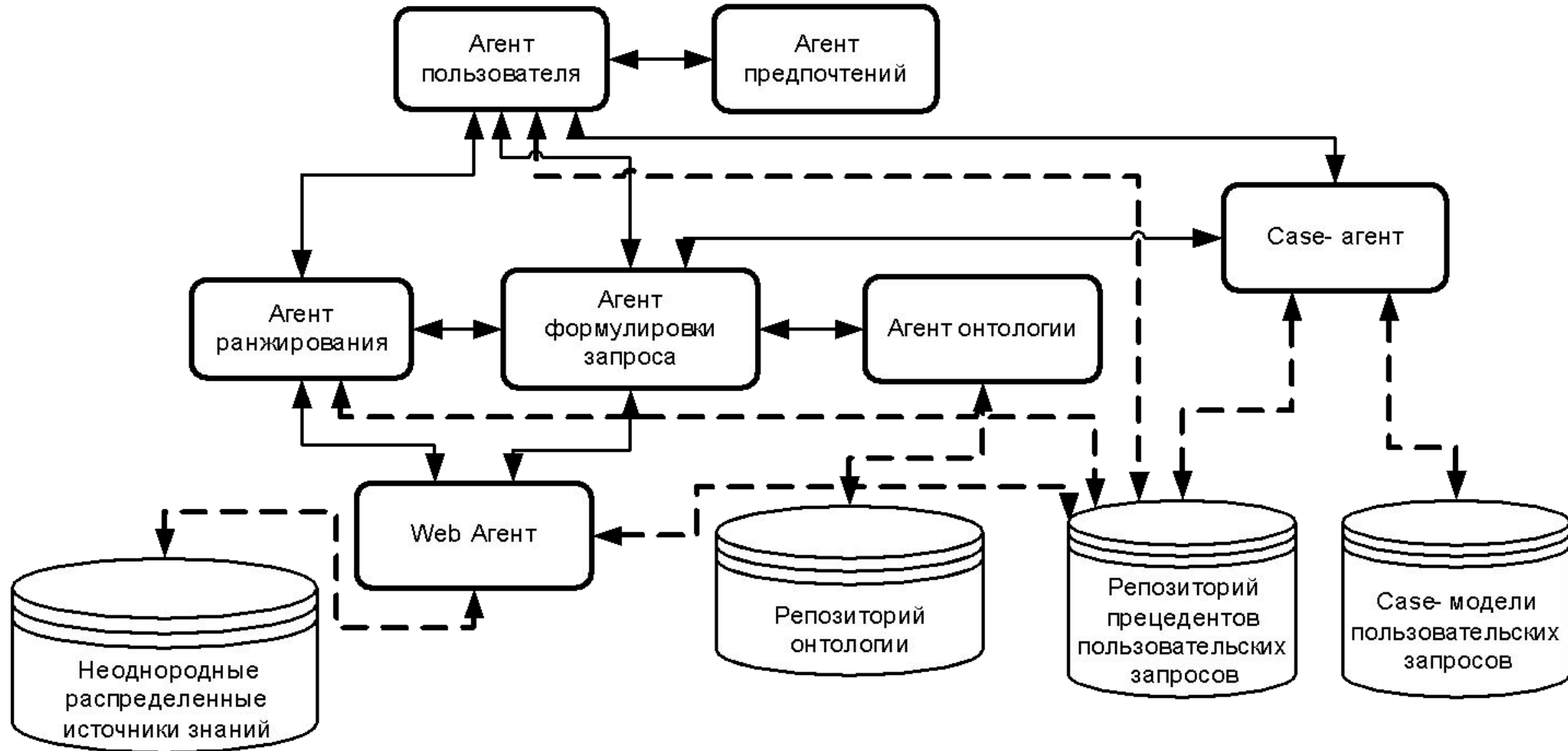


Основными достоинствами предлагаемой модели являются:

- 1) поддержка семантического поиска релевантных знаний на основе онтологических моделей;
- 2) использование информационных энциклопедических справочных систем различной функциональности для усовершенствования формы поискового запроса;
- 3) повышение эффективности запроса пользователя на основе использования репозитория прецедентов.

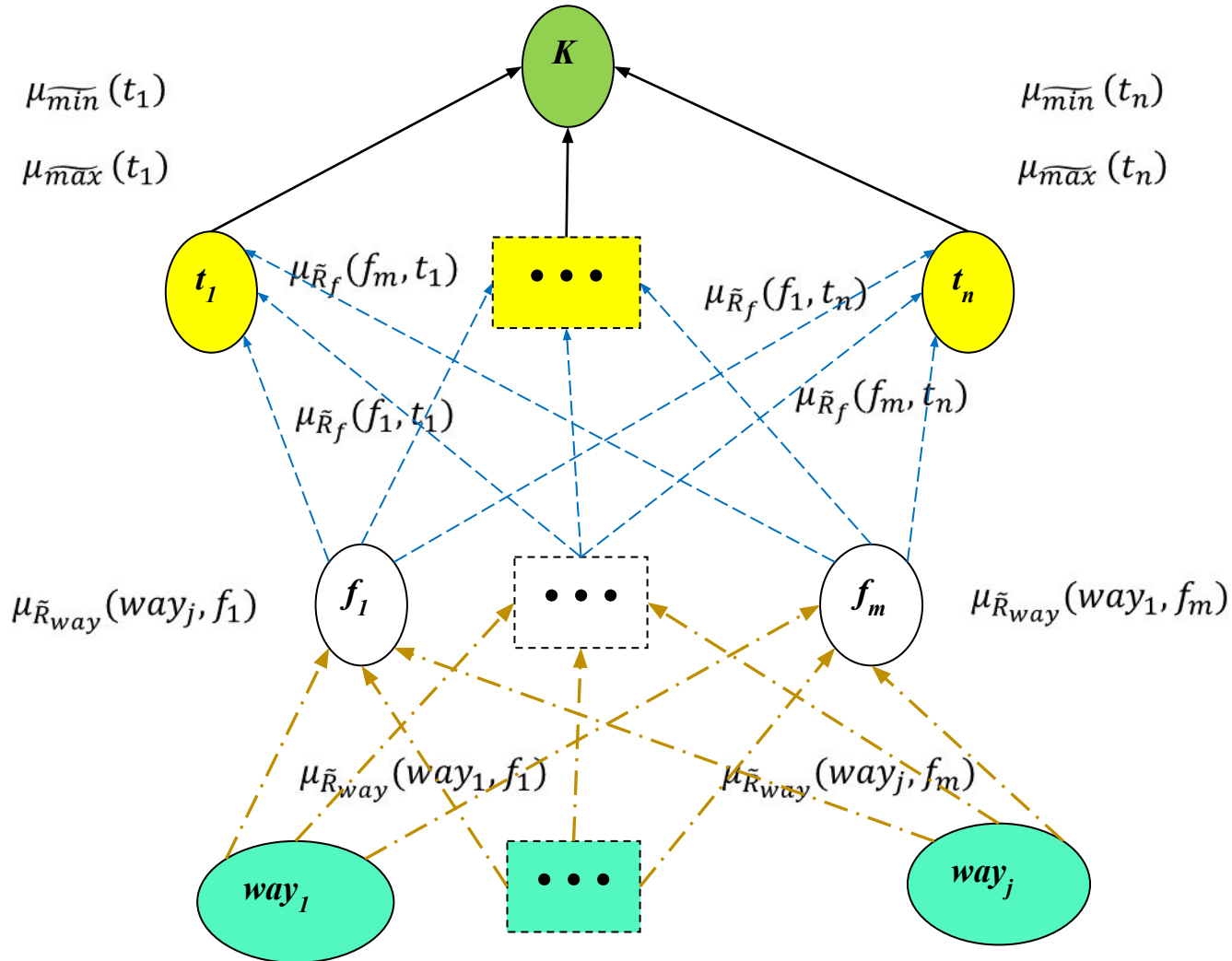
Case-модель фильтра данных

26



Нечеткая модель структуризации содержательной части данных

27



$$R \subset D \times D$$

$$R_t \subset T \times T, \text{ где } (t_i, t_j) \in R_t, i \in [1, n], j \in [1, n], i \neq j$$

$$\mu_{\min} : T \rightarrow P \text{ и } \mu_{\max} : T \rightarrow P$$

$$\overline{MIN} = \{t, \mu_{\min}(t)\}, \quad \overline{MAX} = \{t, \mu_{\max}(t)\}$$

$$\overline{MIN} \subset T \text{ и } \overline{MAX} \subset T$$

$$R_f \subset F \times T \quad (f, t) \in R_f,$$

$$\tilde{R}_f \subset R_f, \tilde{R}_f = \{(f, t), \mu_{\tilde{R}_f}(f, t)\}. \mu_{\tilde{R}_f} : R_f \rightarrow P$$

$$R_{way} \subset WAY \times F; (way, f) \in R_{way}; \mu_{\tilde{R}_{way}} : R_{way} \rightarrow P,$$

$$\tilde{R}_{way} \subset R_{way}, \tilde{R}_{way} = \{(way, f), \mu_{\tilde{R}_{way}}(way, f)\}$$

$$D_1 = T \cup F \cup WAY$$

$$R_1 \subset D_1 \times D_1$$

$$V_1 = \{R_t, R_f, R_{way}\}$$

$$\tilde{G}_1 = (D_1, R_1, \mu_{\tilde{G}_1}(d_1), \mu_{\tilde{G}_1}(r_1))$$

Модель классификации данных на основе обобщенного критерия

28

$Q_i = \left\{ \mu_{Q_i}(t_1)/t_1, \mu_{Q_i}(t_2)/t_2, \dots, \mu_{Q_i}(t_n)/t_n \right\}$ нечеткое множество степени соответствия альтернатив t_n критерию Q_i

$S = Q_1 \cap Q_2 \cap \dots \cap Q_k$ правило выбора наилучшей альтернативы

$$\mu_S(t_j) = \min \mu_{Q_i}(t_j), i = 1, \dots, k; j = 1, \dots, n$$

$$\mu_S(t^*) = \max \mu_S(t_j), j = 1, \dots, n$$
 лучшая альтернатива t^*

$$Q_1 = \sum_{i=1}^P \sum_{k=1}^S \varepsilon_{ij} \delta_{ik}$$

$$Q_2 = \sum_{i=1}^P \sum_{y=1}^S \sum_{k=1}^S \gamma_{iy} \delta_{ik}, y \neq k$$

$$Q_3 = \sum_{k=1}^S \sum_{i=1}^P \varepsilon_{ij} \delta_{ik} \varphi_{ij}$$

$$Q_{int} = \tau_1 Q_1 + \tau_2 Q_2 + \tau_3 Q_3 \rightarrow \max$$