Блокирование Web-сайтов с неприемлемым содержанием на основании выявления их категорий

Зозуля Ю.В., Котенко И.В.
Лаборатория проблем компьютерной безопасности, СПИИРАН

РусКрипто'2010 , 1 - 4 апреля 2010 г.

Актуальность работы

В Интернете расположено множество абсолютно неприемлемых страниц для маленьких детей, которые могут появляться, даже если ребенок этого не желает.

В связи с этим огромную значимость имеют системы родительского контроля, предназначенные для обеспечения безопасности ребенка при работе с компьютером, в частности при его доступе в Интернет.

Общая характеристика работы

Цель:

- Разработка формальной модели, архитектуры и начального прототипа системы определения категории веб-сайтов для решении задачи родительского контроля.

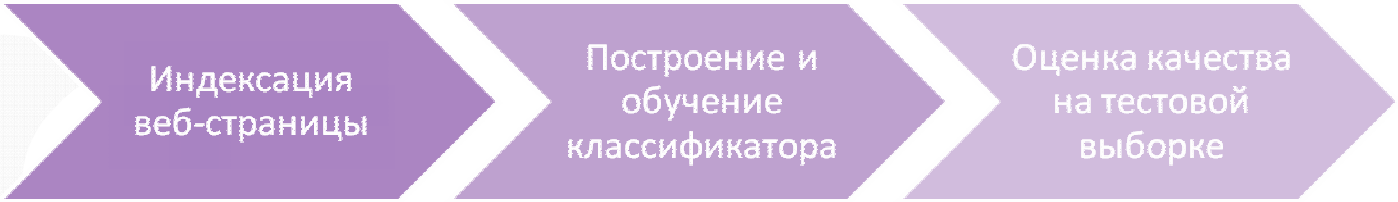
Задачи:

- Анализ существующих моделей и методов определения категории веб-сайтов.
- Разработка моделей:
  - Взаимодействия с пользователем.
  - Определения категории веб-сайтов.
  - Влияния категории веб-сайта на принятие решения о доступе к нему
- Разработка прототипа, реализующего выделенные модели.

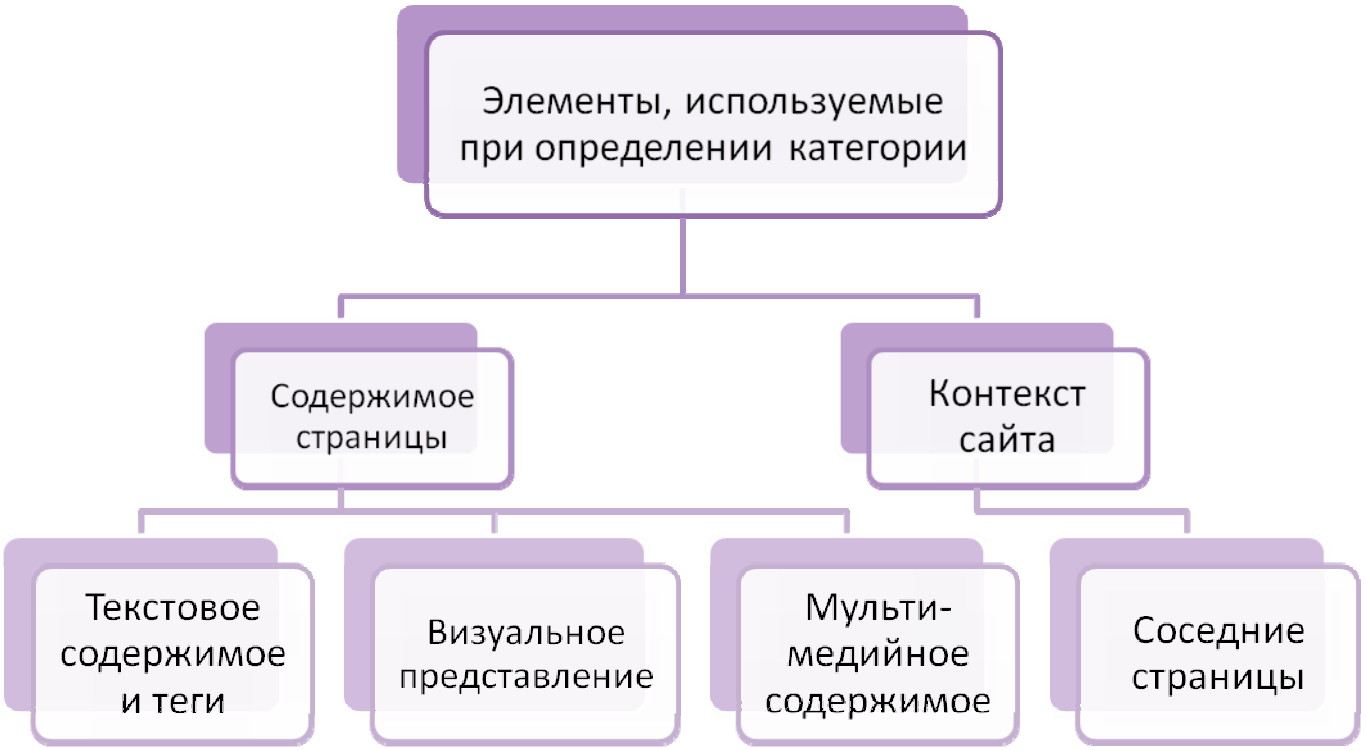
Анализ задачи классификации

- Классификация – отнесение веб-страницы к заранее определенным одной или нескольким категориям.

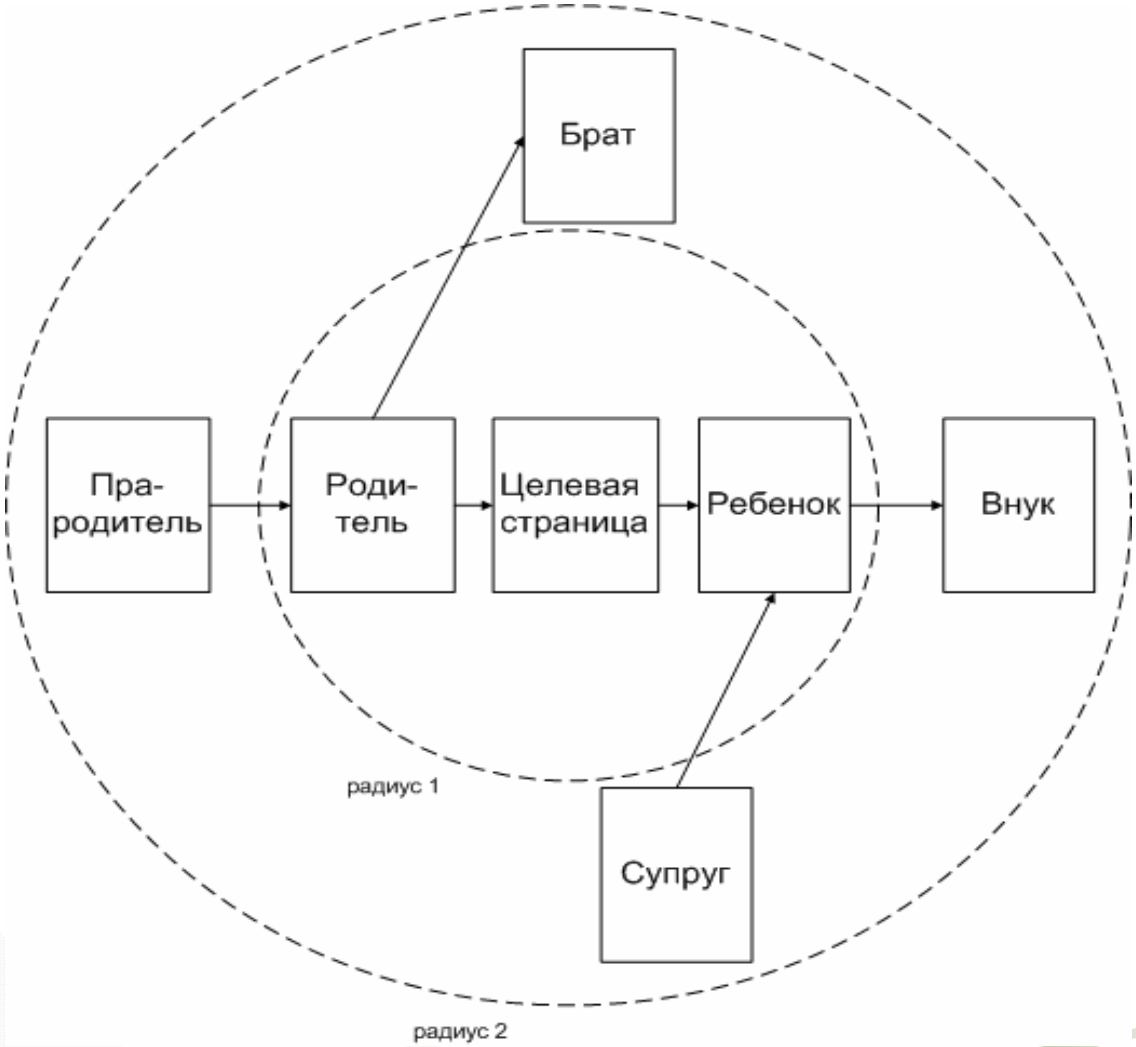
Автоматическая классификация состоит из 3 этапов:



Анализ существующих решений

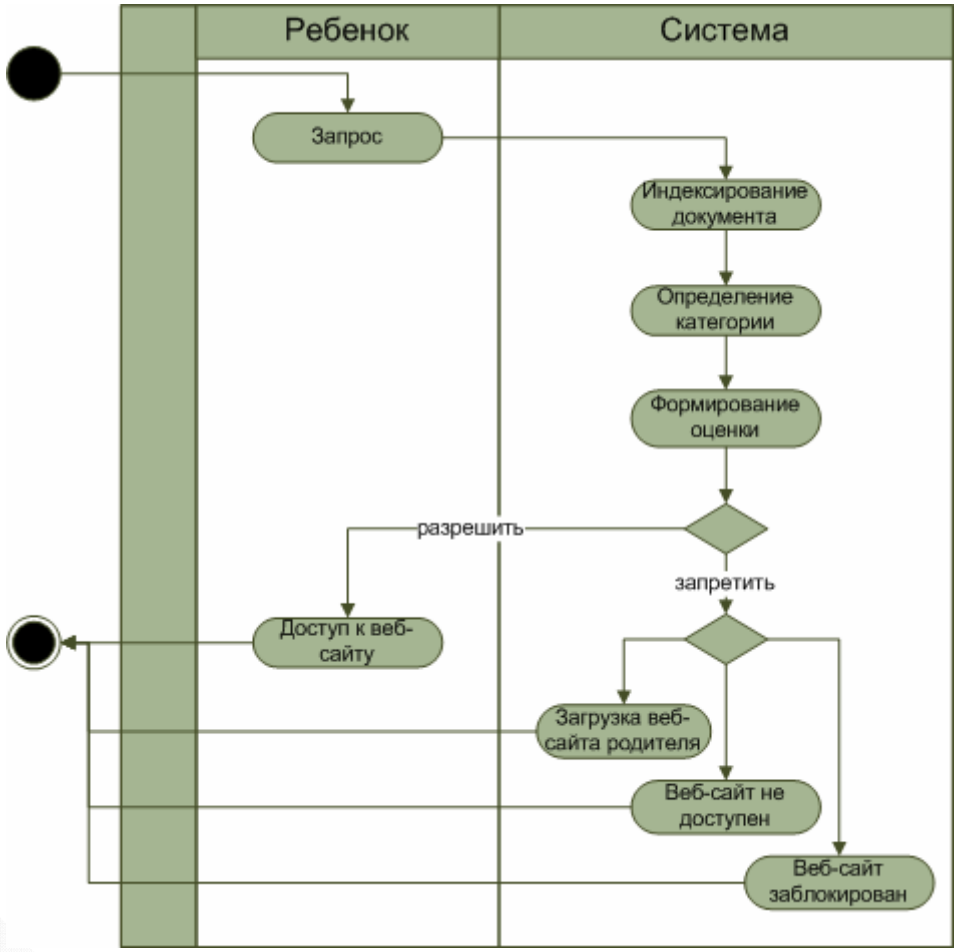


Виды соседних страниц



Требования к перечню настраиваемых параметров

Требования к перечню реализуемых функций



Качественные и количественные характеристики системы

Множество веб-страниц		Решение эксперта о доступе	
		ЗАПРЕЩЕН	РАЗРЕШЕН
Решение классификатора о доступе	ЗАПРЕЩЕН	TP	FP
	РАЗРЕШЕН	FN	TN

Точность и полнота принятого решения о запрете доступа:

$$\pi = \frac{TP}{TP + FP}, \qquad \rho = \frac{TP}{TP + FN}.$$

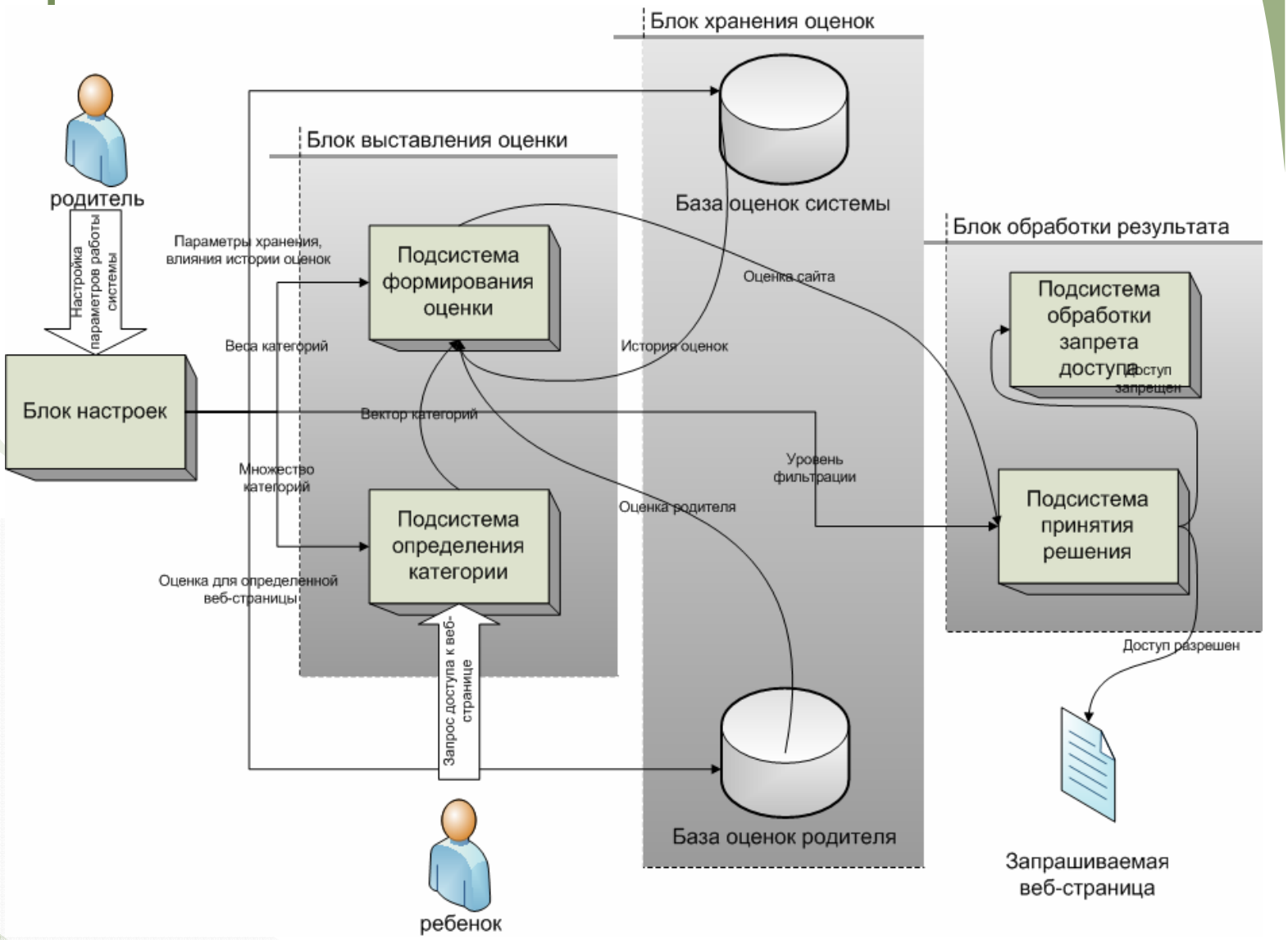
Интегрированная характеристика («точка безубыточности»):

$$\beta = \overline{\pi + \rho}.$$

Оперативность принятого решения:

$$T_d = t_f + \sum_i t_{ci}$$

Архитектура системы определения категории



Блок настраиваемых параметров

Учет оценки родителя:

$$G = \bar{G} = \omega_p G_p + \omega_s G_s,$$

$$\mathrm{A} = \omega_p A_p\langle p_i, c_i\rangle + \omega_s \Phi'(\langle p_i, c_i\rangle),$$

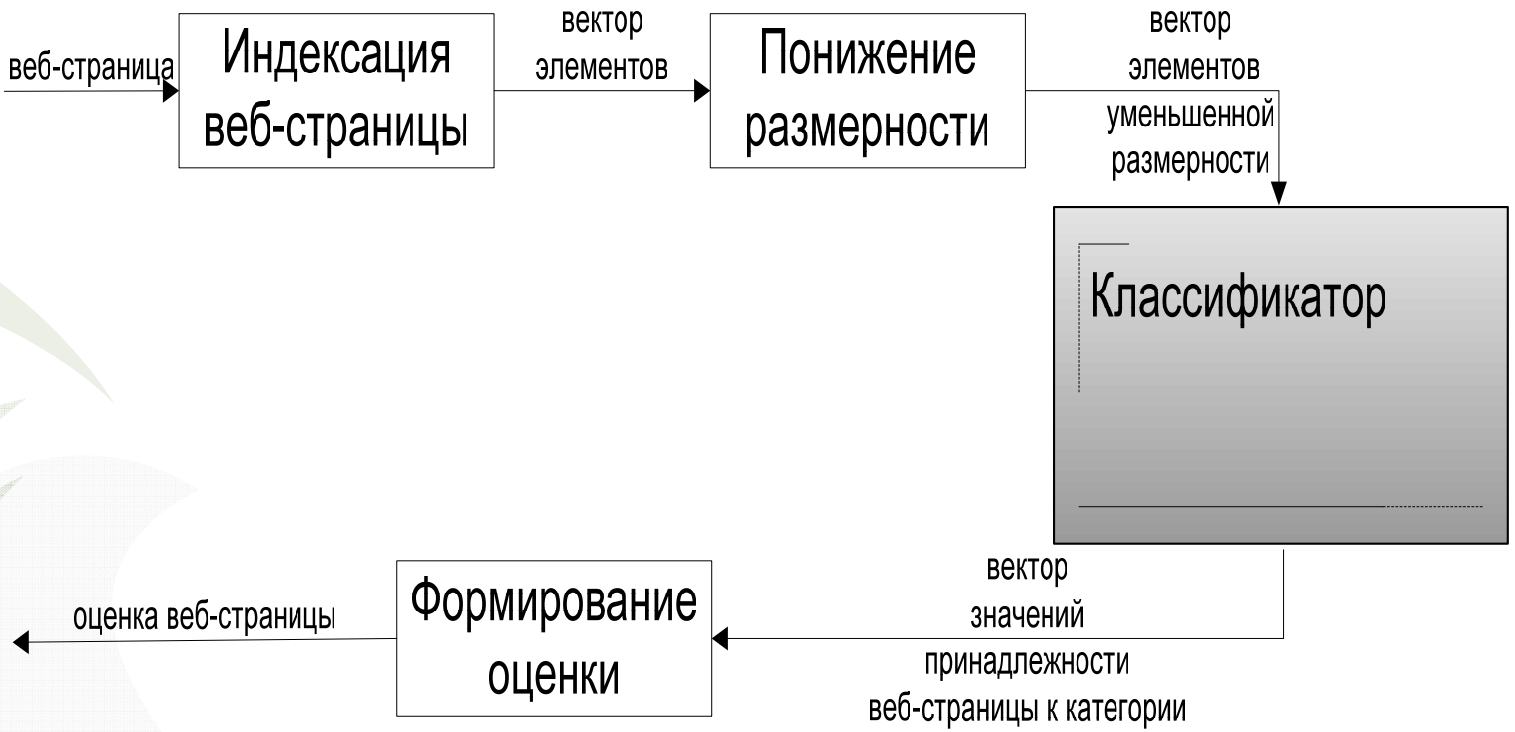
Влияние истории оценок:

$$G = \bar{G} = \omega_b G_b + \omega_s G_s,$$

Алгоритм изменения старшей оценки в базе данных:

1: $G_b \leftarrow G_N$
2: for grade of $n = N - 1$ to 1 **do**
3: $\quad G_b \leftarrow \omega_b G_b + \omega_s G_n$

Блок выставления системной оценки



Индексация страницы

Индексация страницы - представление страницы p_j в виде вектора весов элементов

$$p_j = \langle w_{1j}, \dots, w_{|T|j} \rangle$$

Вычисление веса w_{kj} элемента t_k :

$$tfidf(t_k, p_j) = \#(t_k, p_j) \cdot \log \frac{|Tr|}{\#_{Tr}(t_k)}$$

Уменьшение размерности векторов:

$$OR(t_k, c_i) = \frac{P(t_k | c_i) \cdot (1 - P(t_k, \overline{c_i}))}{(1 - P(t_k | c_i)) \cdot P(t_k | \overline{c_i})}$$

Классификаторы

- Наивный Байесовский классификатор.

$$\{P_{1t}, P_{1\bar{t}}, \ldots, P_{|\mathrm{T}|t}, P_{|\mathrm{T}|\bar{t}}\}$$

где P_{kt} краткая запись для $P(w_{kx} = 1|c_t)$

- *k*-NN классификатор.

$$\Phi'(p_j, c_i) = \sum_{p_z \in Tr_k(d_l)} F(p_j, p_z) \cdot [\![\Phi'(p_z, c_i)]\!],$$

- SVM-Классификатор. Попытка найти среди всех плоскостей $\sigma_1, \sigma_2, \ldots$ в $|\mathrm{T}|$ -мерном пространстве, отделяющих принадлежащие *P*-примеры от *N*-примеров, такую σ_t , которая разделяет их лучше остальных.

Формирование оценки

Обозначим a_i за степень принадлежности сайта к каждой из выбранных пользователем категорий.

Тогда системная оценка для целевого сайта:

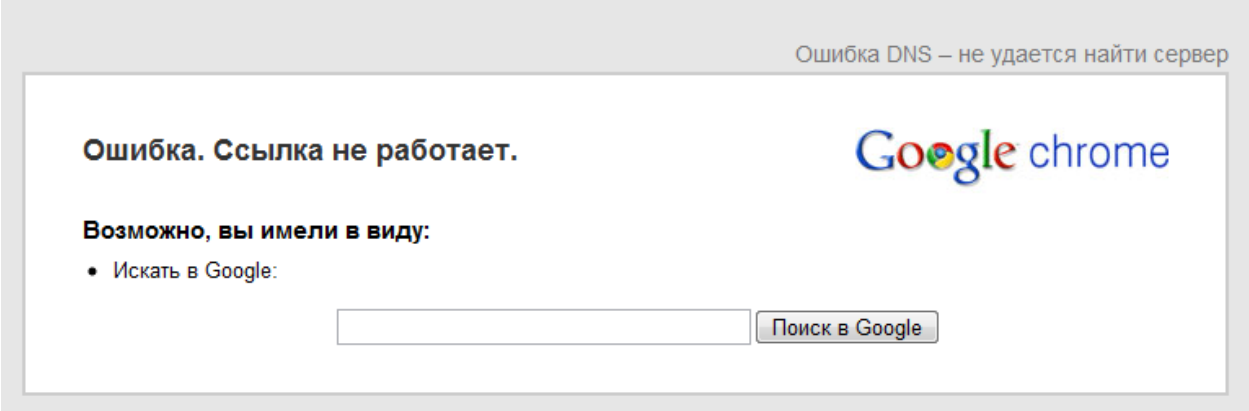
$$G_s = \max_{a_i \in A, \omega_i \in \Omega} a_i \cdot \omega_i$$

Тогда результирующая оценка с учетом настроенных параметров:

$$G = \bar{G} = \omega_p G_p + (1 - \omega_p)(\omega_b G_b + (1 - \omega_b) G_s).$$

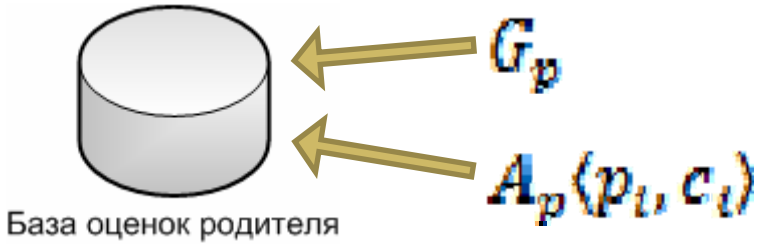
Блок обработки результата: Сценарии поведения системы

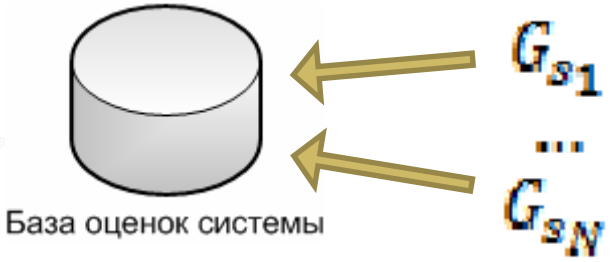
1. Невозможно отобразить страницу:



2. Переход на смоделированную веб-страницу с полезной информацией.
3. Переход на интернет-страницы, одобренные родителем.
4. Предупреждение о запрете доступа.

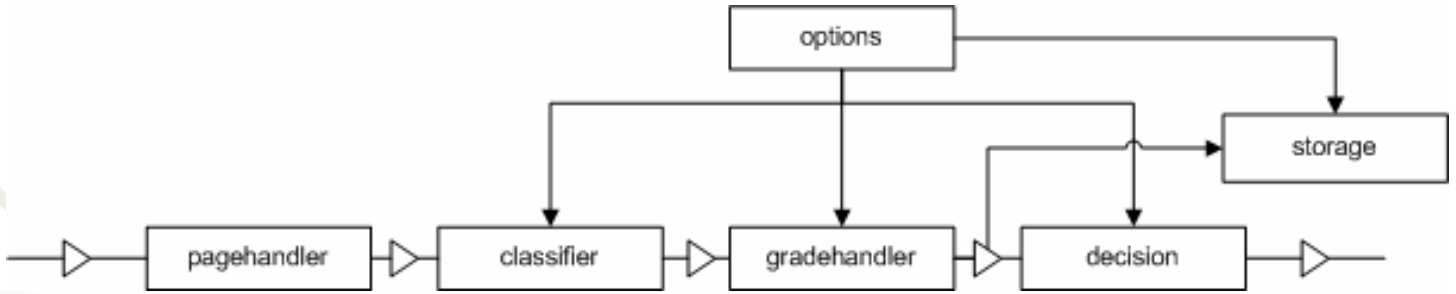
Блок хранения данных





Прототип системы

Прототип системы состоит из 5 основных модулей.



Спасибо за внимание