

Mail.ru на РОМИП'2005

Андрей Федоровский, Михаил Костин, Андрей Проскурин

Mail.ru

Структура системы

- Инверсные и прямые индексы
- Комплексная функция релевантности
 - Частота термов
 - Совместная встречаемость термов
 - Зоны документа
 - Неполное вхождение термов запроса
 - Особая обработка длинных документов

Функция релевантности

$$W = k_f W_f + k_p W_p + k_{ps} W_{ps}$$

W_f Частотность термов запроса по TF*IDF

W_p Встречаемость пар соседних слов запроса

W_{ps} Вес наилучшего пассажа в документе

TF*IDF

$$TF * IDF_{term} = \frac{f_{term} \times IDF_{term}}{f_{term} + k_1(b + L(1 - b))}$$

$$L = \frac{L_w + k_4}{k_3 + k_4}$$

$$W_f = \sum_{term \in query} TF * IDF_{term}$$

Пассажи

реформа законодательства России

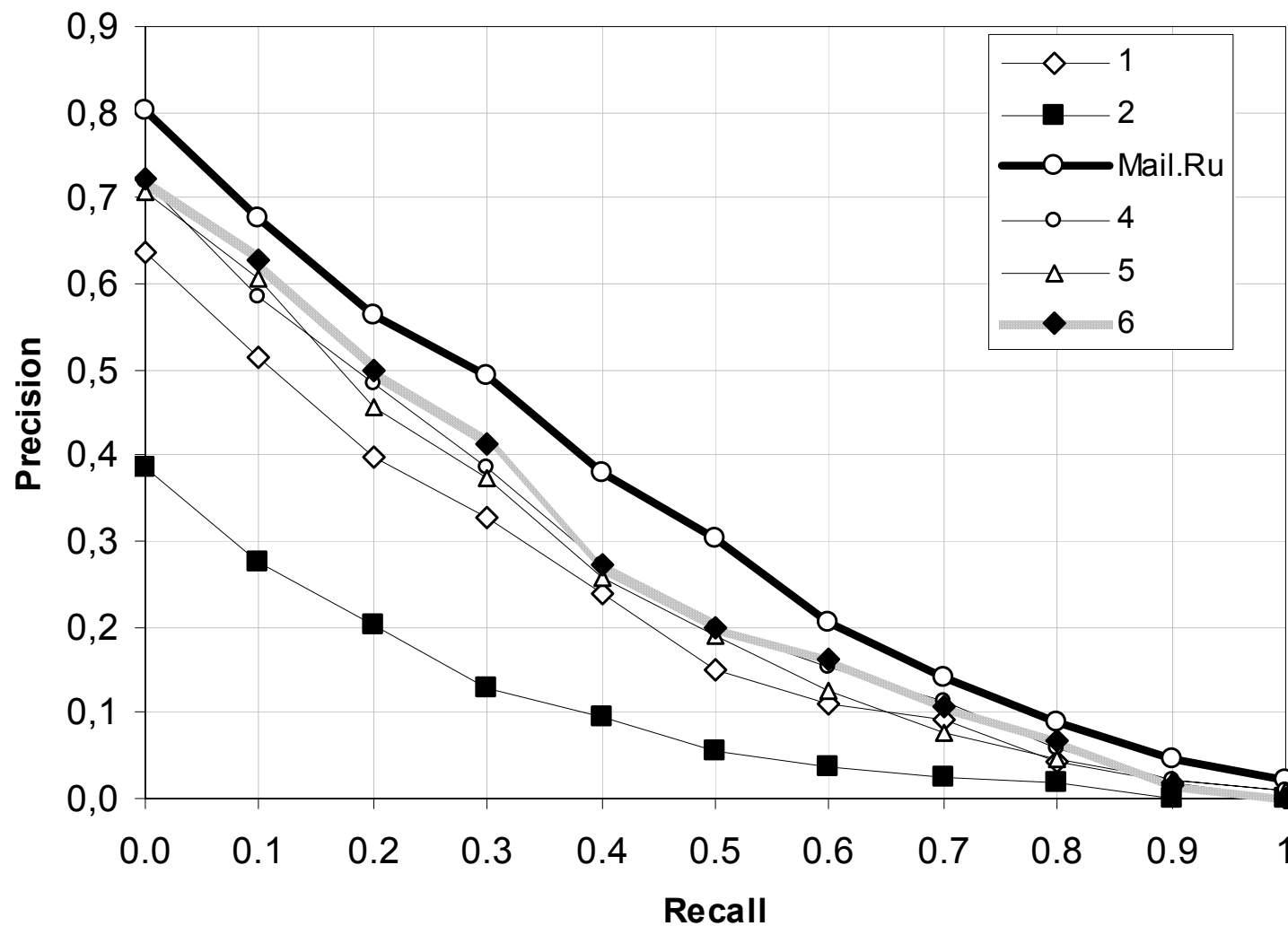
“Введение в политику права” (“Киевские Университетские Известия” за 1896 — 97 гг.). предисловие к ст. “Предстоящая **реформа** акц. **законодательства** в **России**”, (в “Русском экономическом обозрении”, 1896, май — декабрь).

«Картофельные бунты» государственных крестьян в **России**. Отмена в **России** Литовского статута (действовал с 1588). Распространение на западные губернии общероссийского **законодательства**.

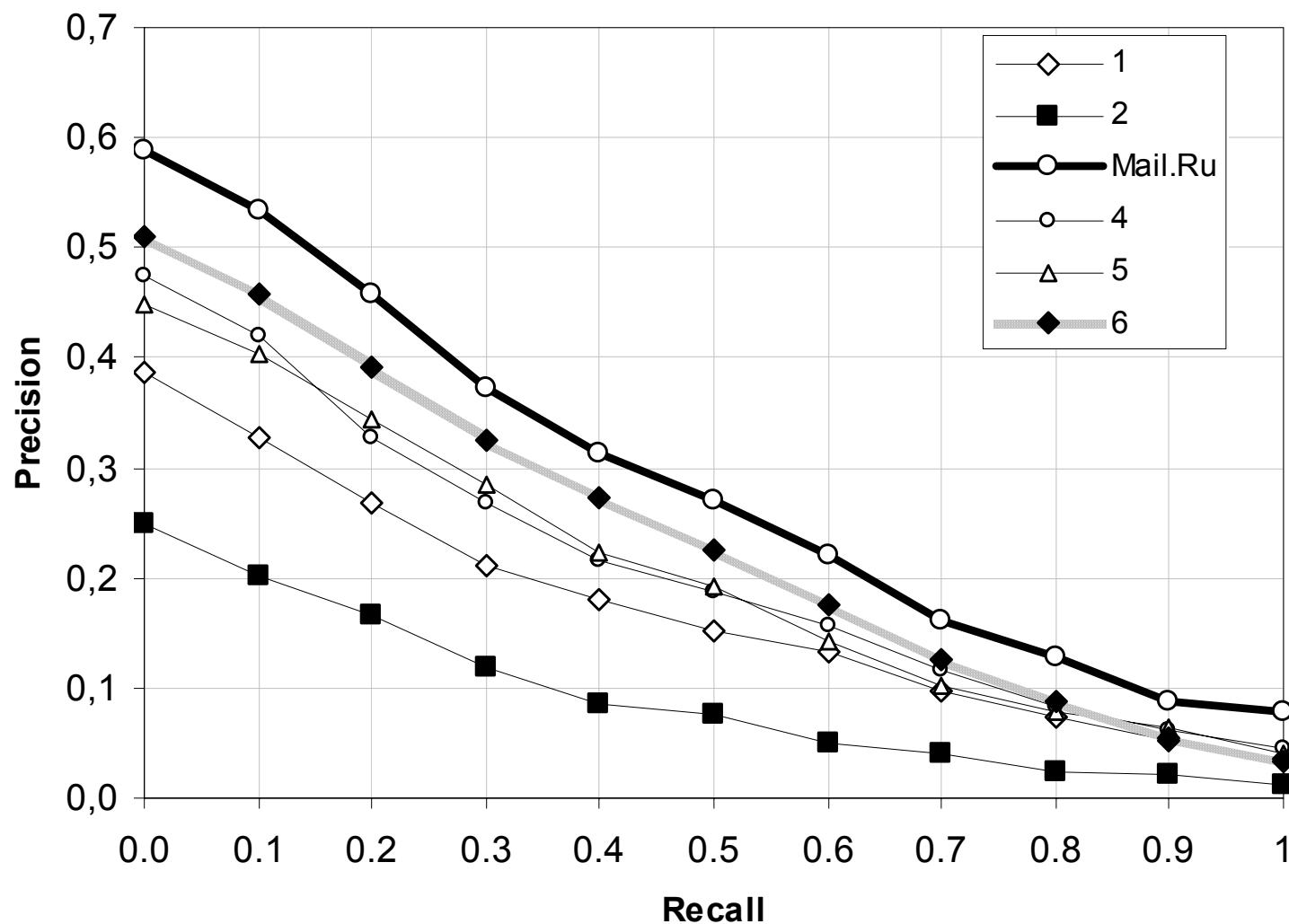
Основные характеристики пассажей

- Полнота
- Длина
- Порядок слов
- Зона документа
- Близость к началу

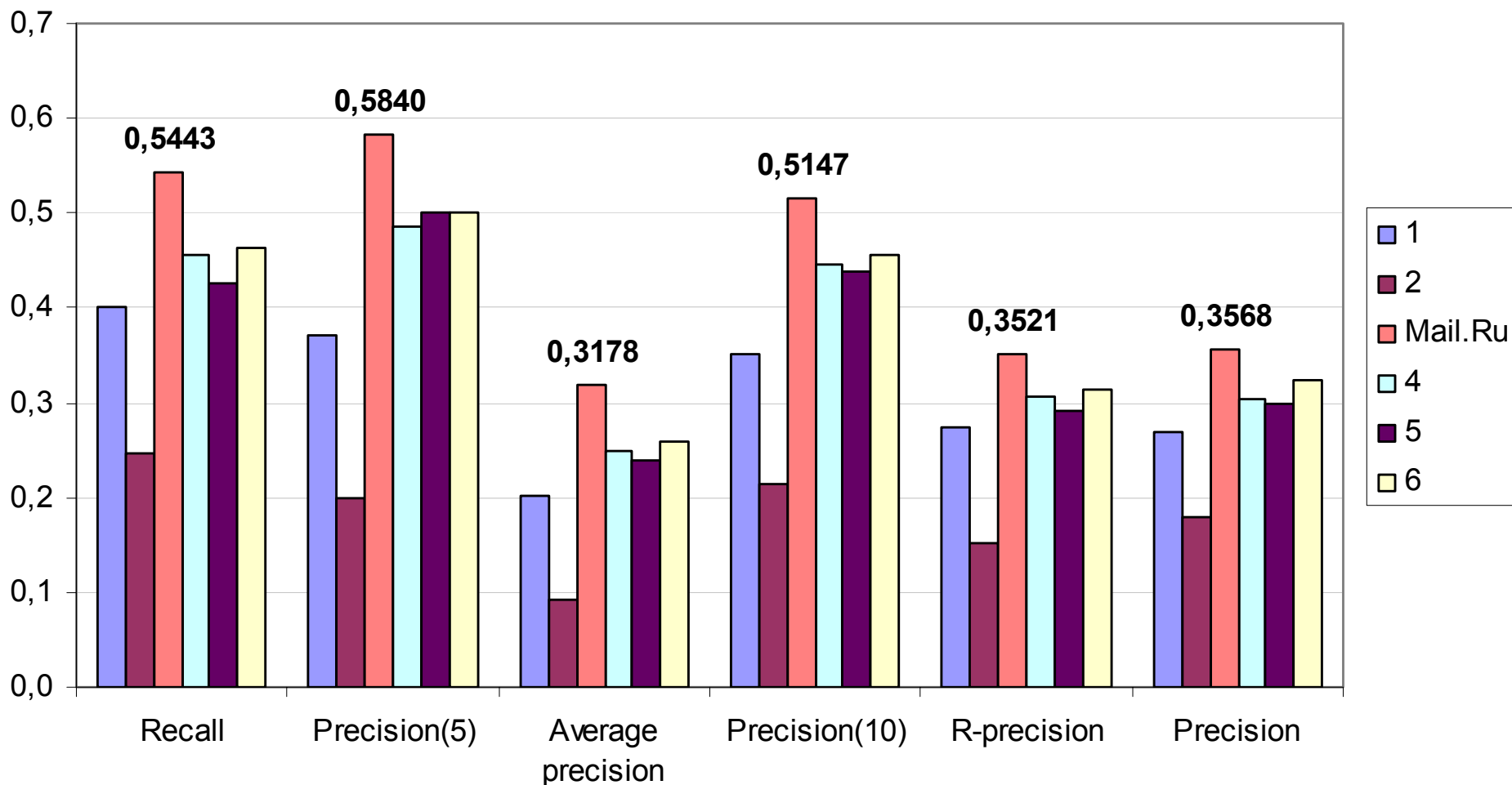
Web adhoc all(2004+2005) pd50 OR



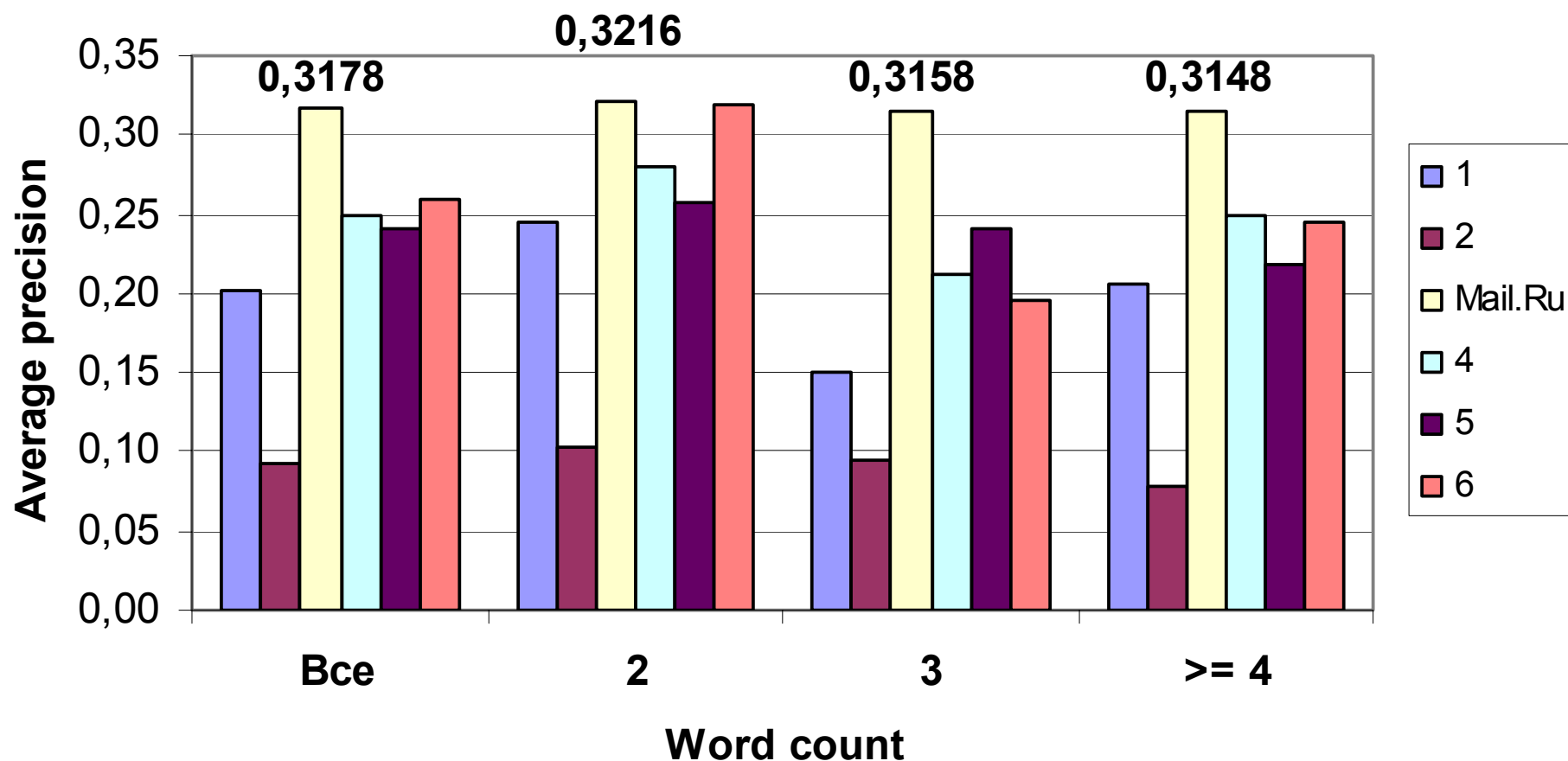
Web adhoc all(2004+2005) pd50 AND



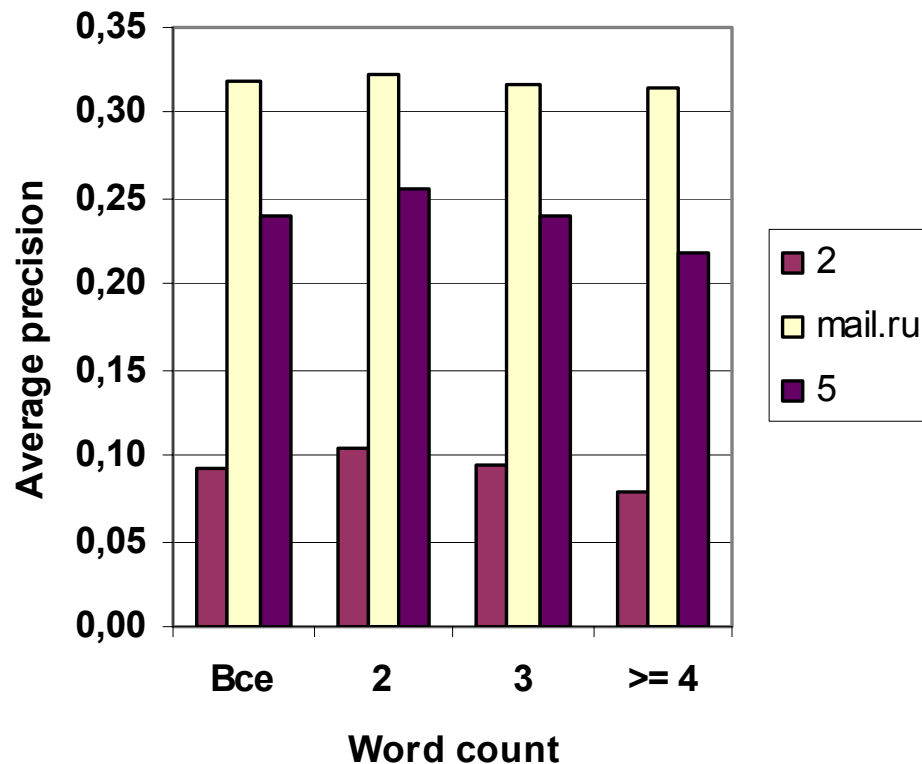
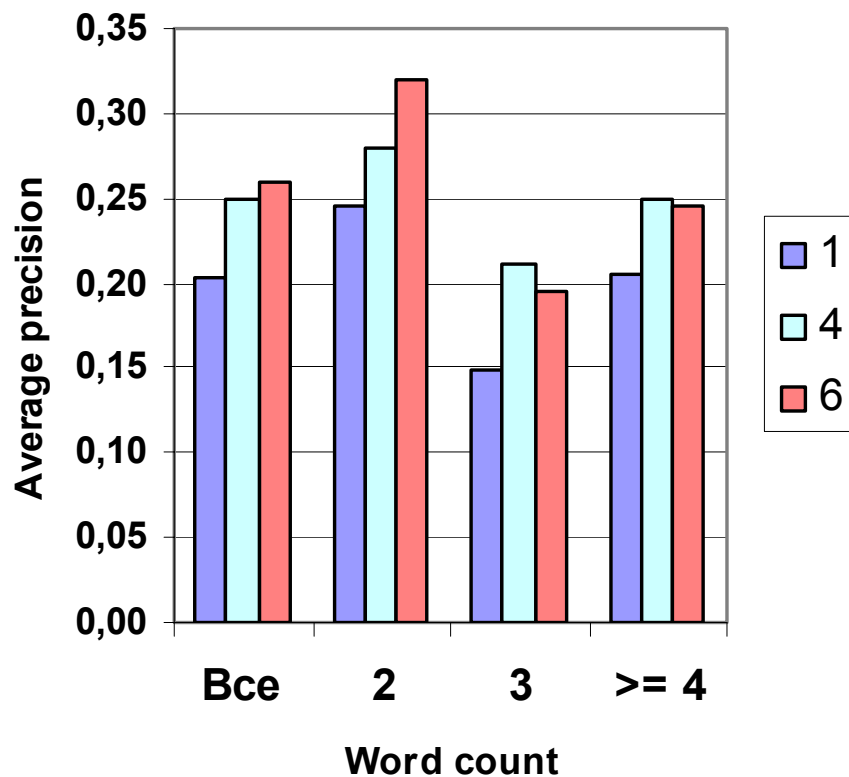
Web adhoc all(2004+2005) pd50 OR



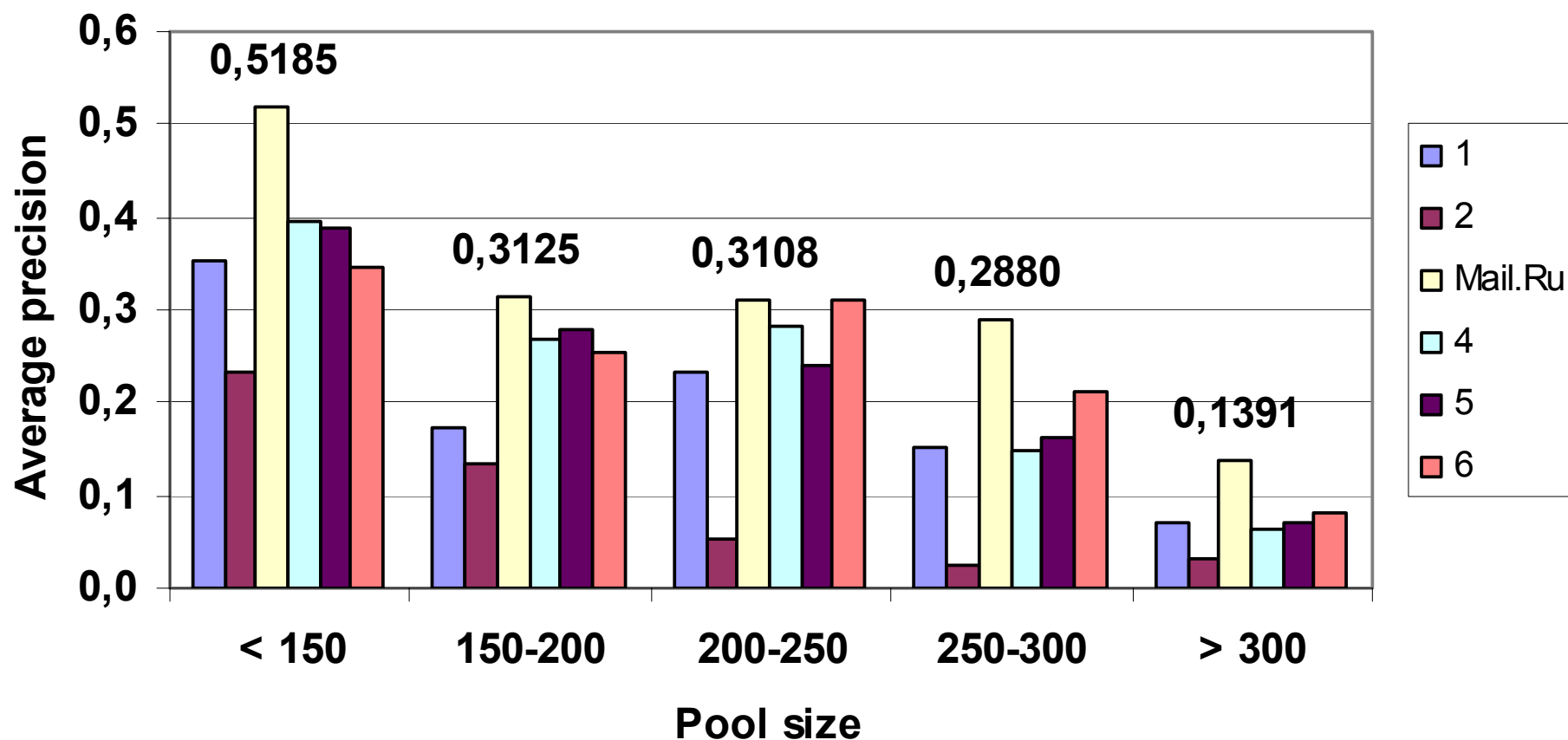
Разбиение по кол-ву слов



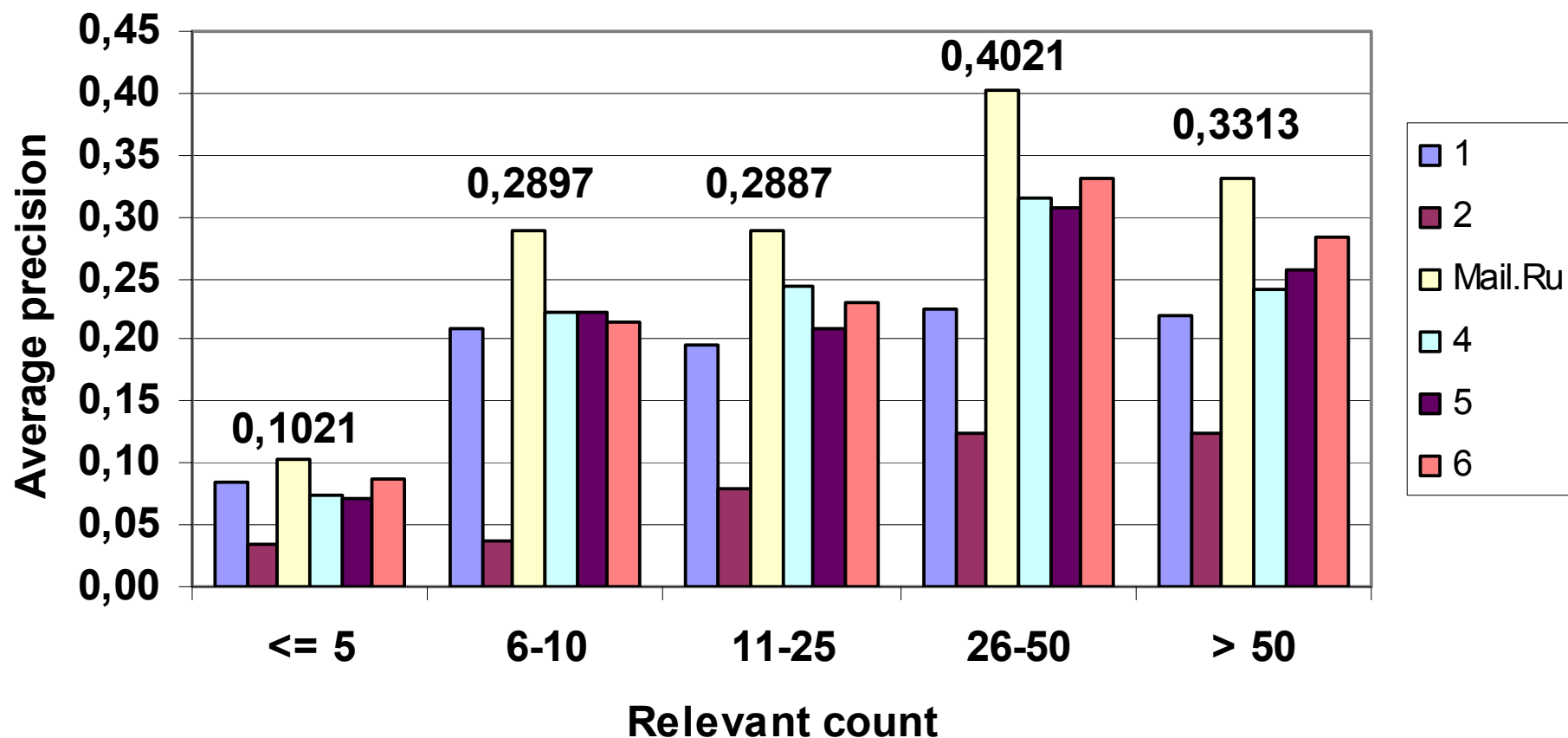
Разбиение по кол-ву слов



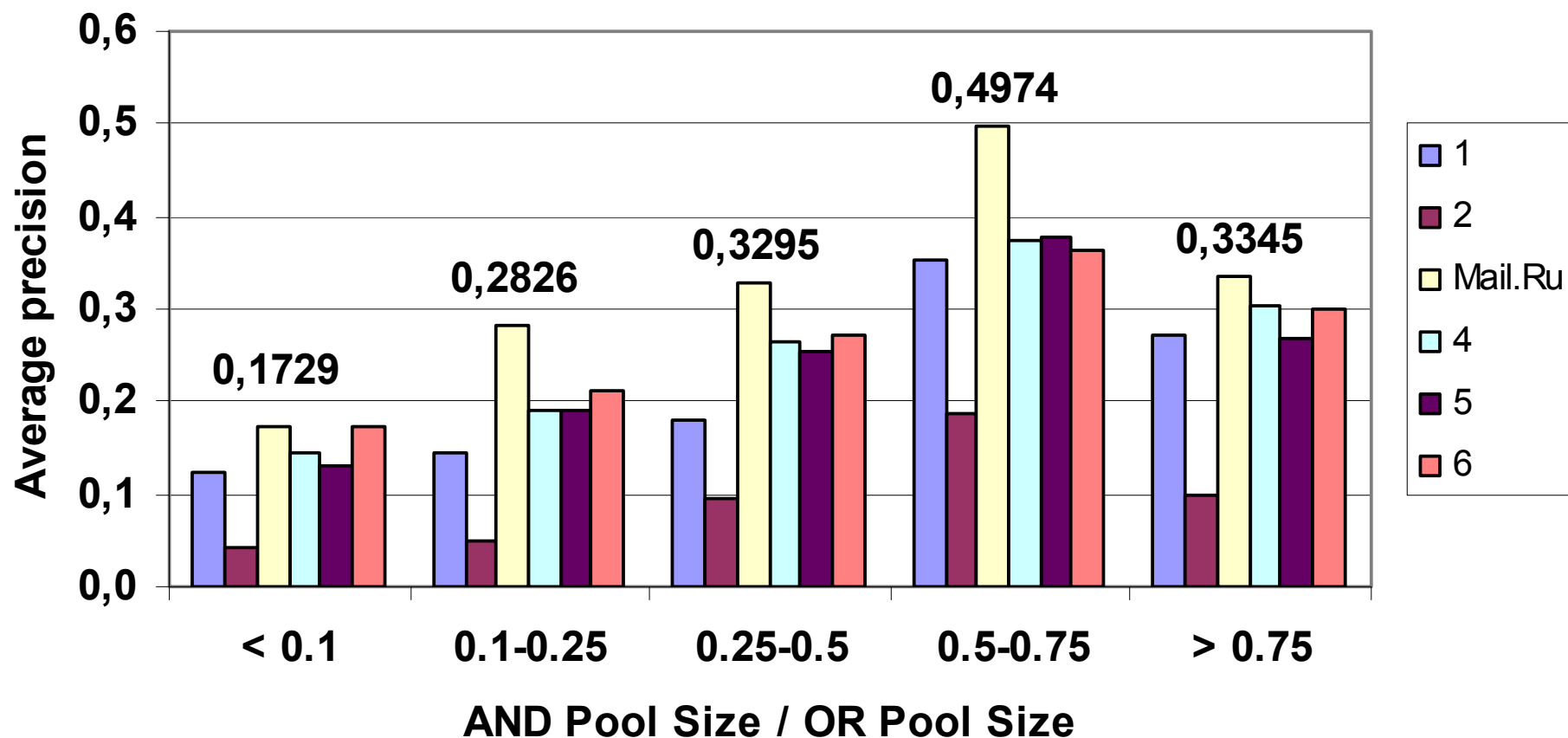
Разбиение по числу найденных



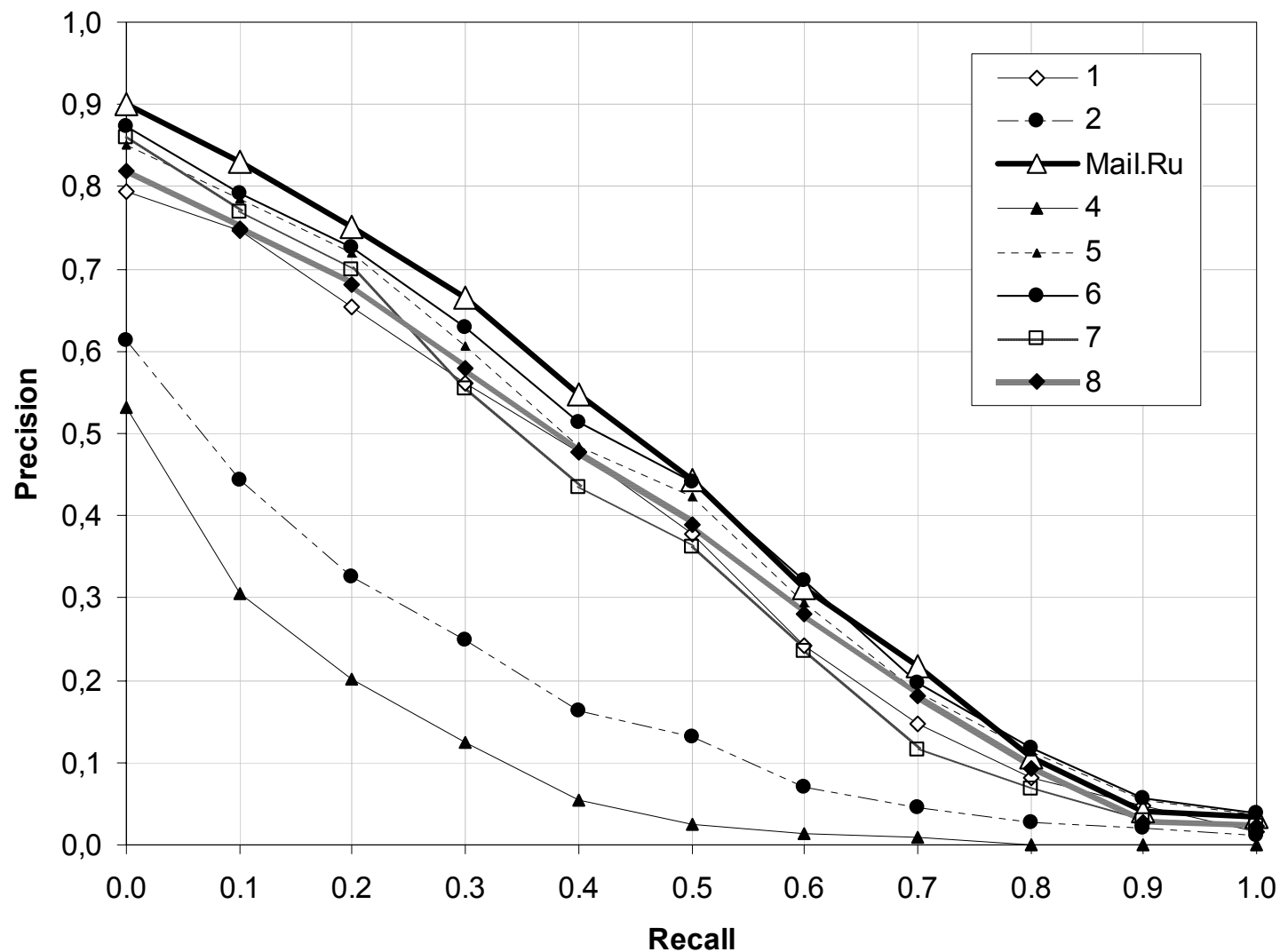
Разбиение по числу релевантных



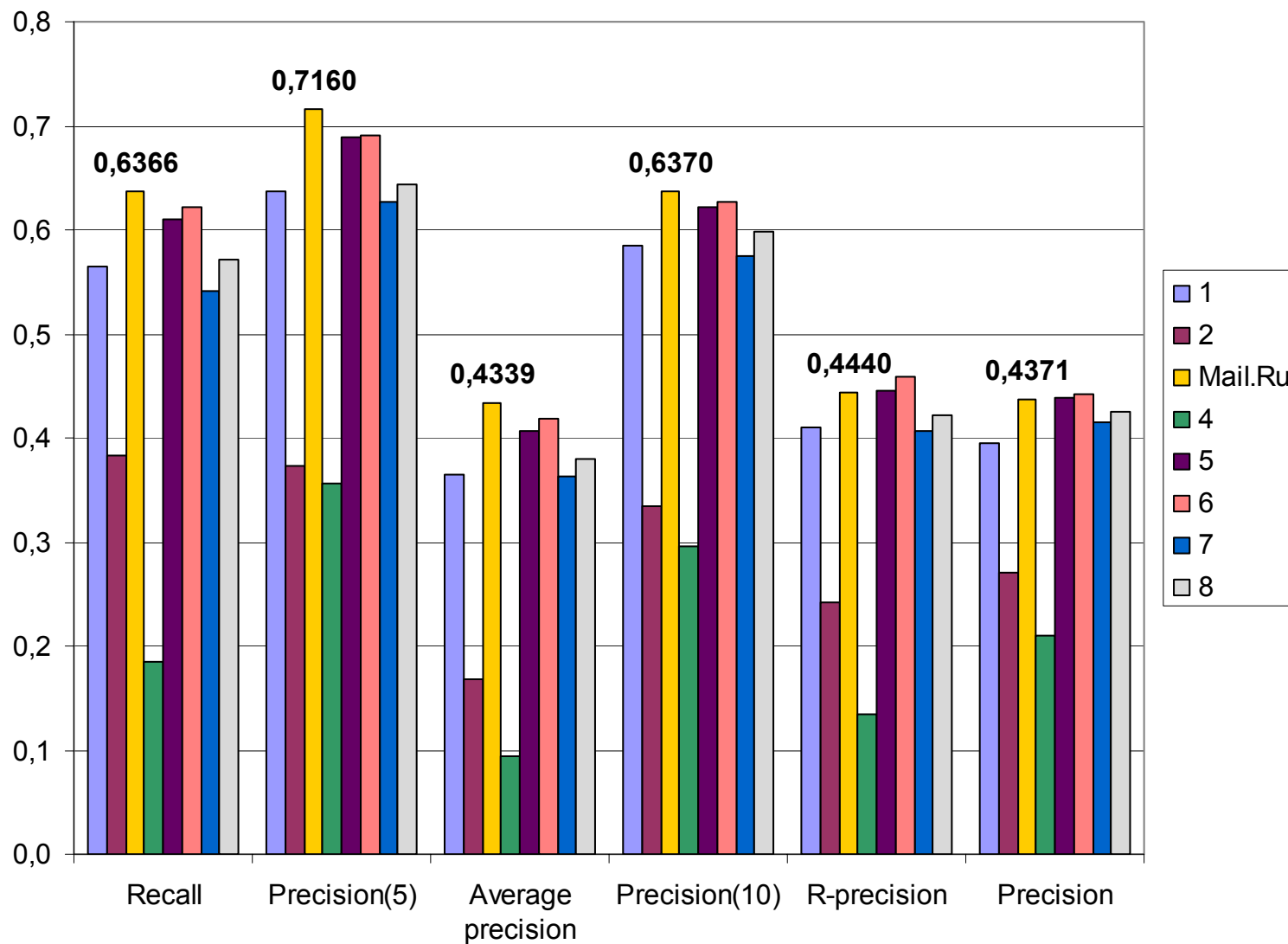
Разбиение по отношению AND/OR



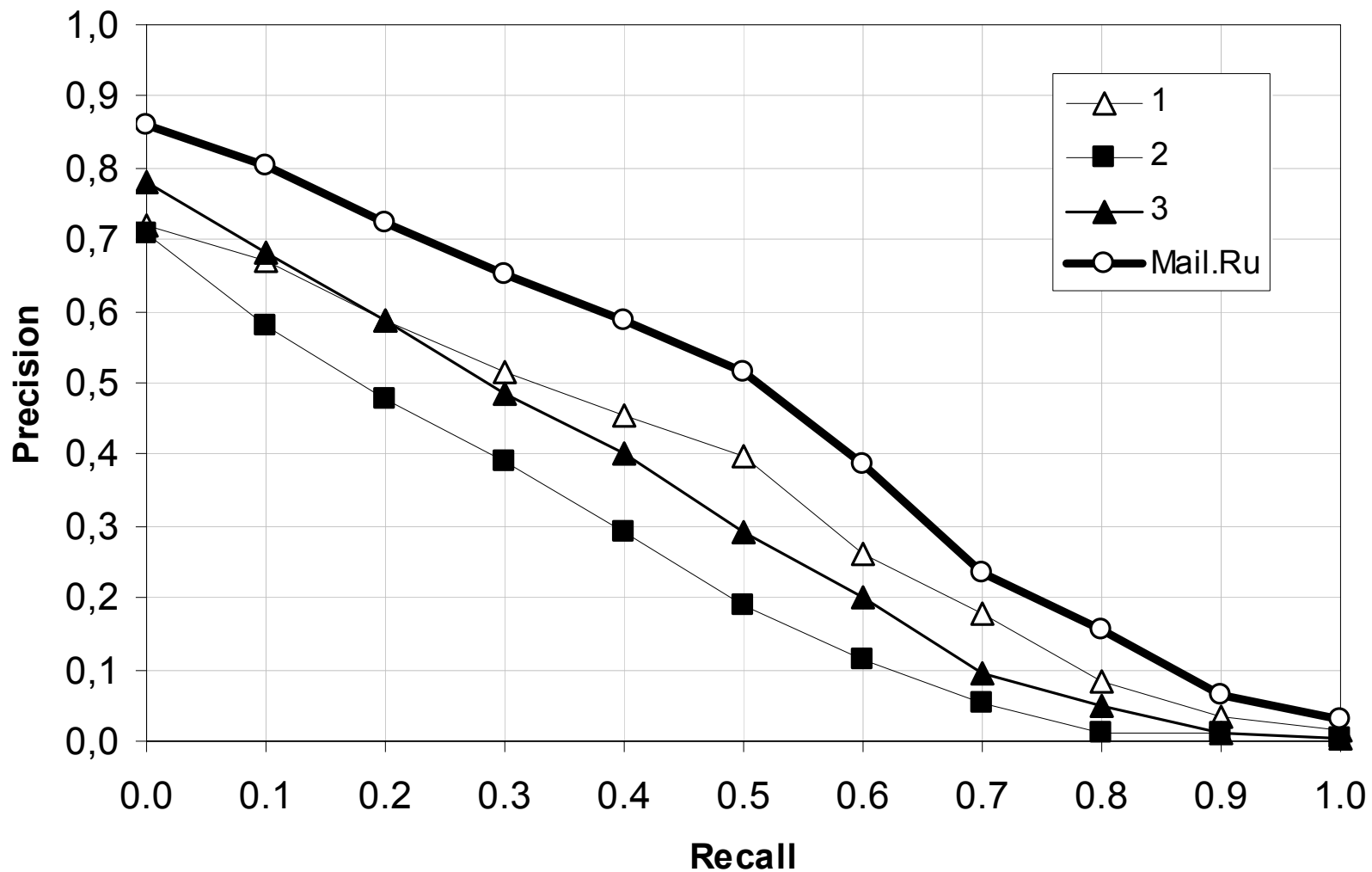
Legal adhoc all(2004+2005) pd50



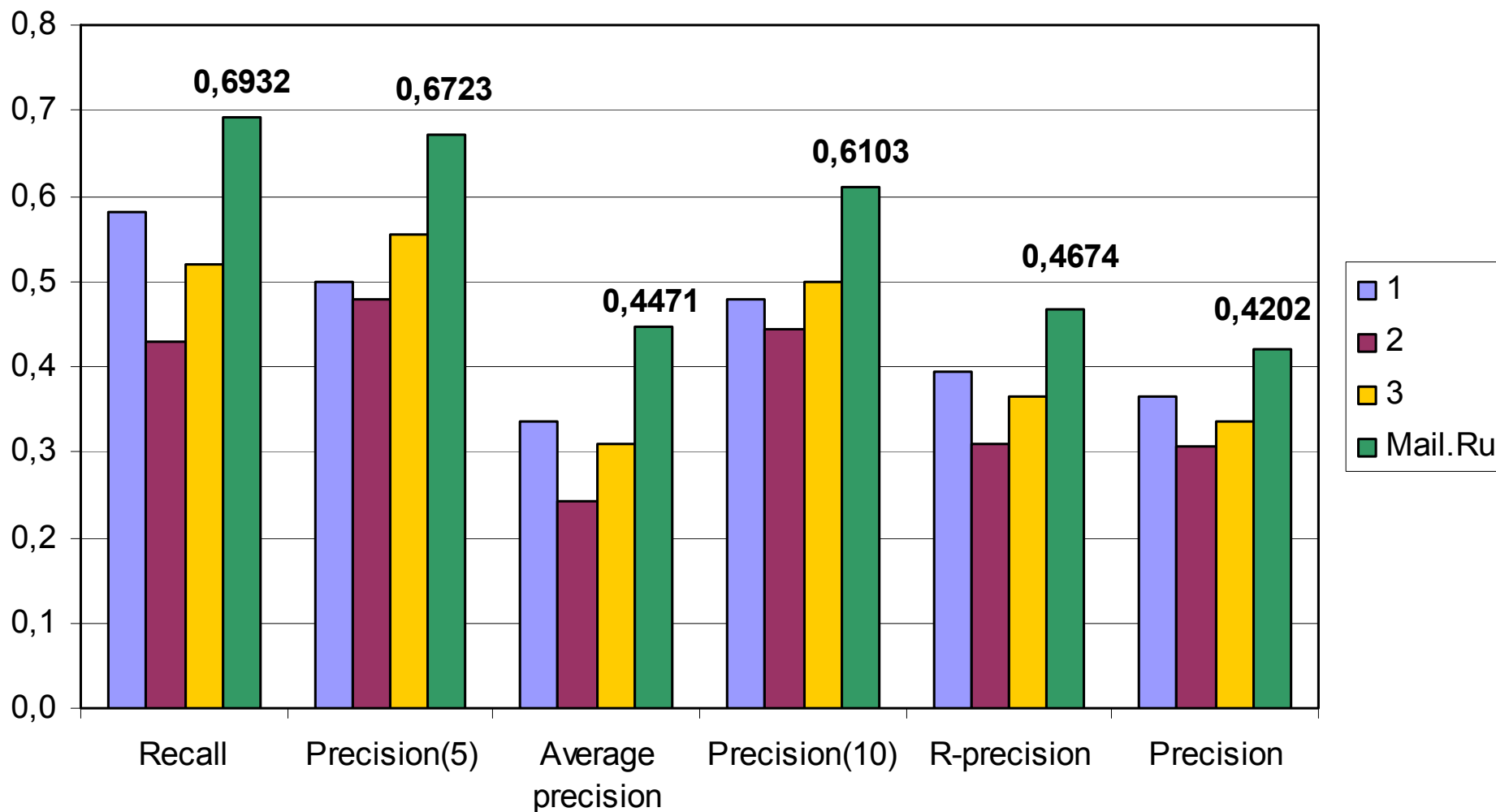
Legal adhoc all(2004+2005) pd50



Mixed adhoc pd50 OR



Mixed adhoc pd50 OR



Корреляция мер

	Recall	P(5)	AvP	P(10)	R-prec	Precision	TREC
Recall	1	0,5020	0,8617	0,3902	0,7291	-0,0469	0,7684
P(5)	0,5020	1	0,8601	0,9824	0,9348	0,8046	0,9270
AvP	0,8617	0,8601	1	0,7808	0,9645	0,4293	0,9808
P(10)	0,3902	0,9824	0,7808	1	0,8927	0,8690	0,8782
R-prec	0,7291	0,9348	0,9645	0,8927	1	0,6180	0,9907
Precision	-0,0469	0,8046	0,4293	0,8690	0,6180	1	0,5712
TREC	0,7684	0,9270	0,9808	0,8782	0,9907	0,5712	1

P(5) и P(10)

AvP, R-prec, TREC avg

SVM. Линейно разделяемый случай

$$\{x_i, y_i\}, \quad i=1\dots l, \quad y_i \in \{1, -1\}, \quad x_i \in R^d$$

$$w \cdot x + b = 0 \quad - \text{уравнение разделяющей плоскости}$$

$$x_i \cdot w + b \geq 1 \quad \forall i: y_i = 1$$

$$x_i \cdot w + b \leq -1 \quad \forall i: y_i = -1$$

$$y_i(x_i \cdot w + b) - 1 \geq 0 \quad \forall i \quad - \text{окончательный вид ограничений}$$

SVM. Неразделяемый случай. Ядра

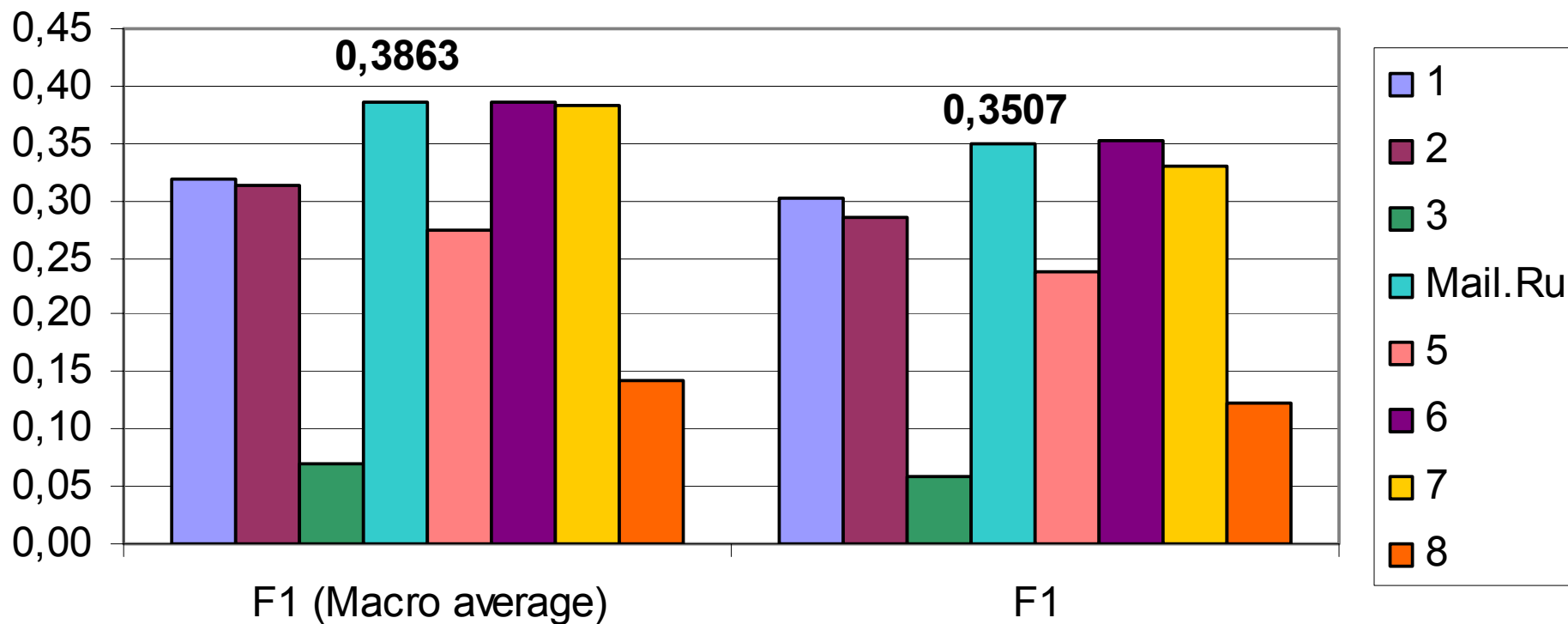
$$\phi: R^d \rightarrow H$$

$$K(x_1, x_2) = \phi(x_1) \cdot \phi(x_2)$$

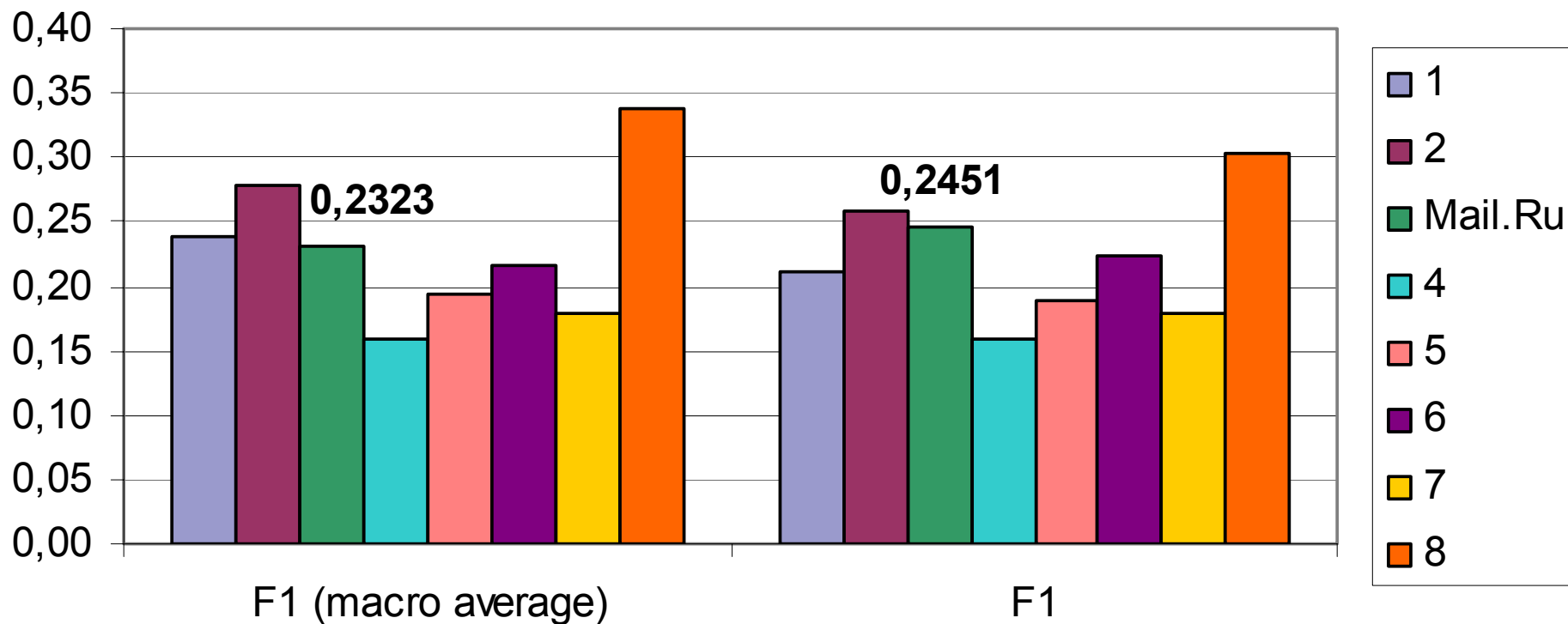
$$K(x, y) = (x \cdot y + 1)^p \quad - \text{гомогенное полиномиальное ядро}$$

$$K(x, y) = e^{-\|x-y\|^2 / 2\sigma^2} \quad - \text{гауссово ядро (RBF)}$$

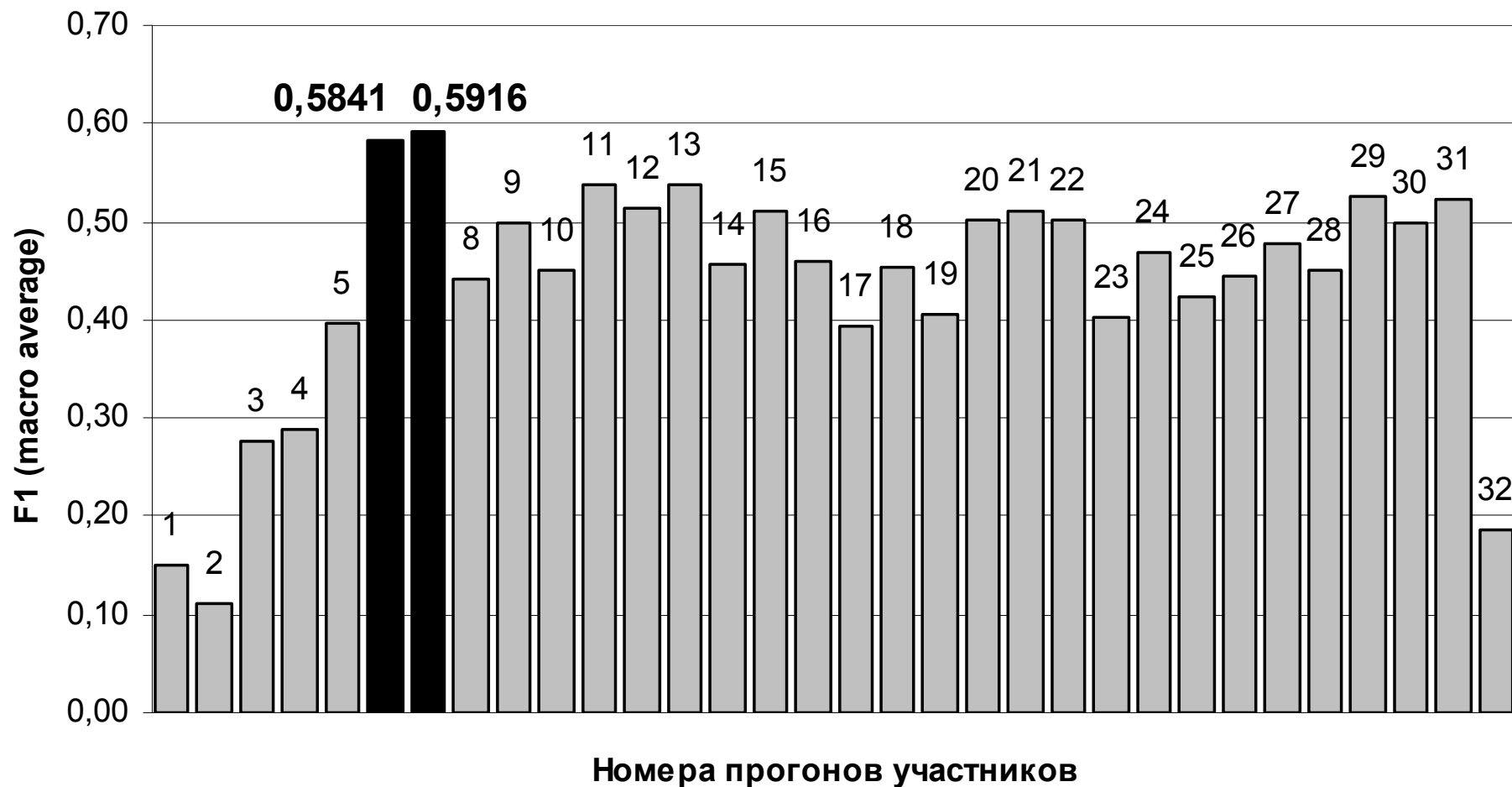
Классификация страниц



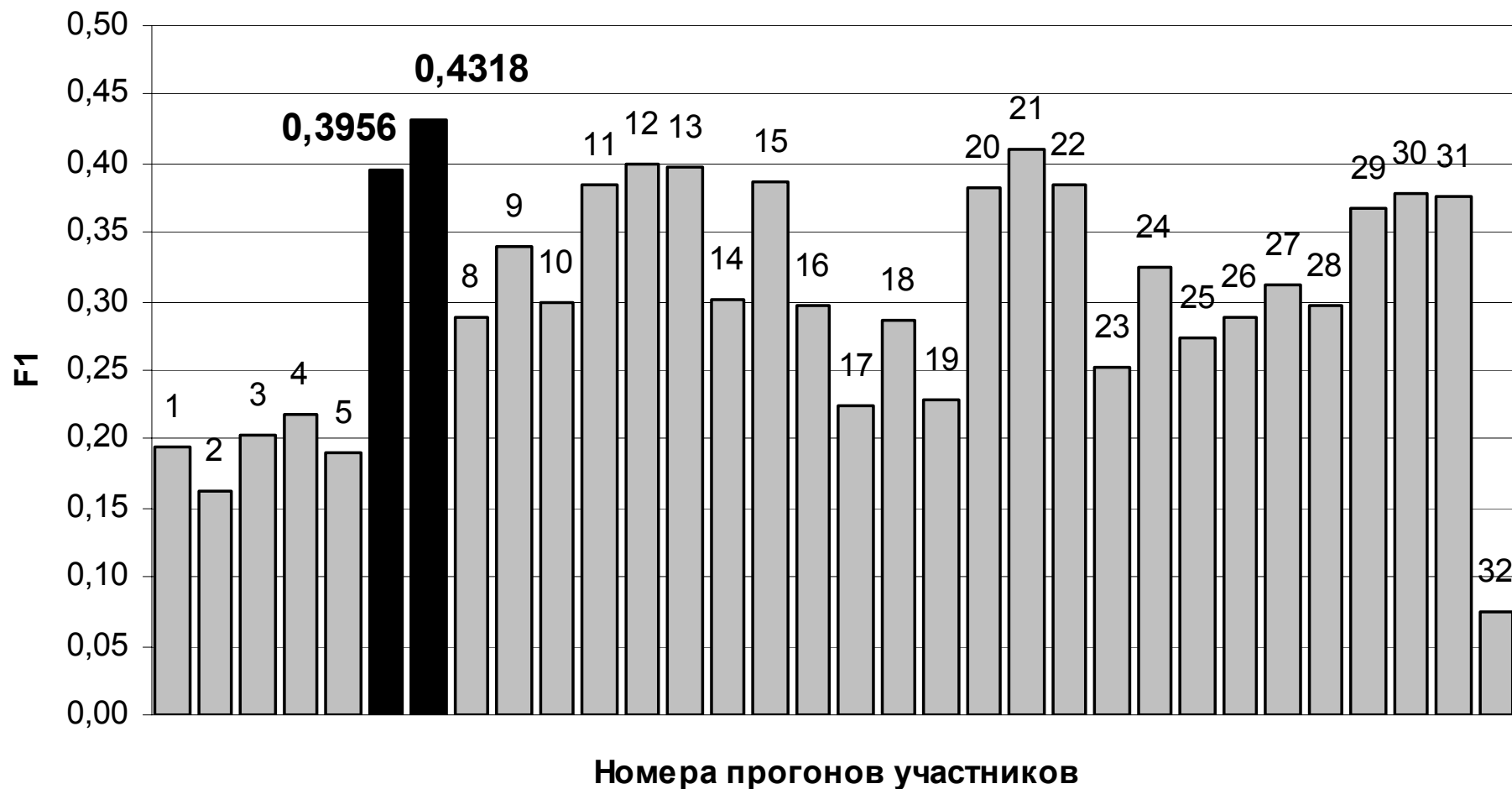
Классификация сайтов



Классификация нормативных документов



Классификация нормативных документов



Есть вопросы? Обращайтесь!

Андрей Федоровский

Руководитель отдела разработки

Mail.ru

fedorovsky@corp.mail.ru

Тел. (095) 725-6357