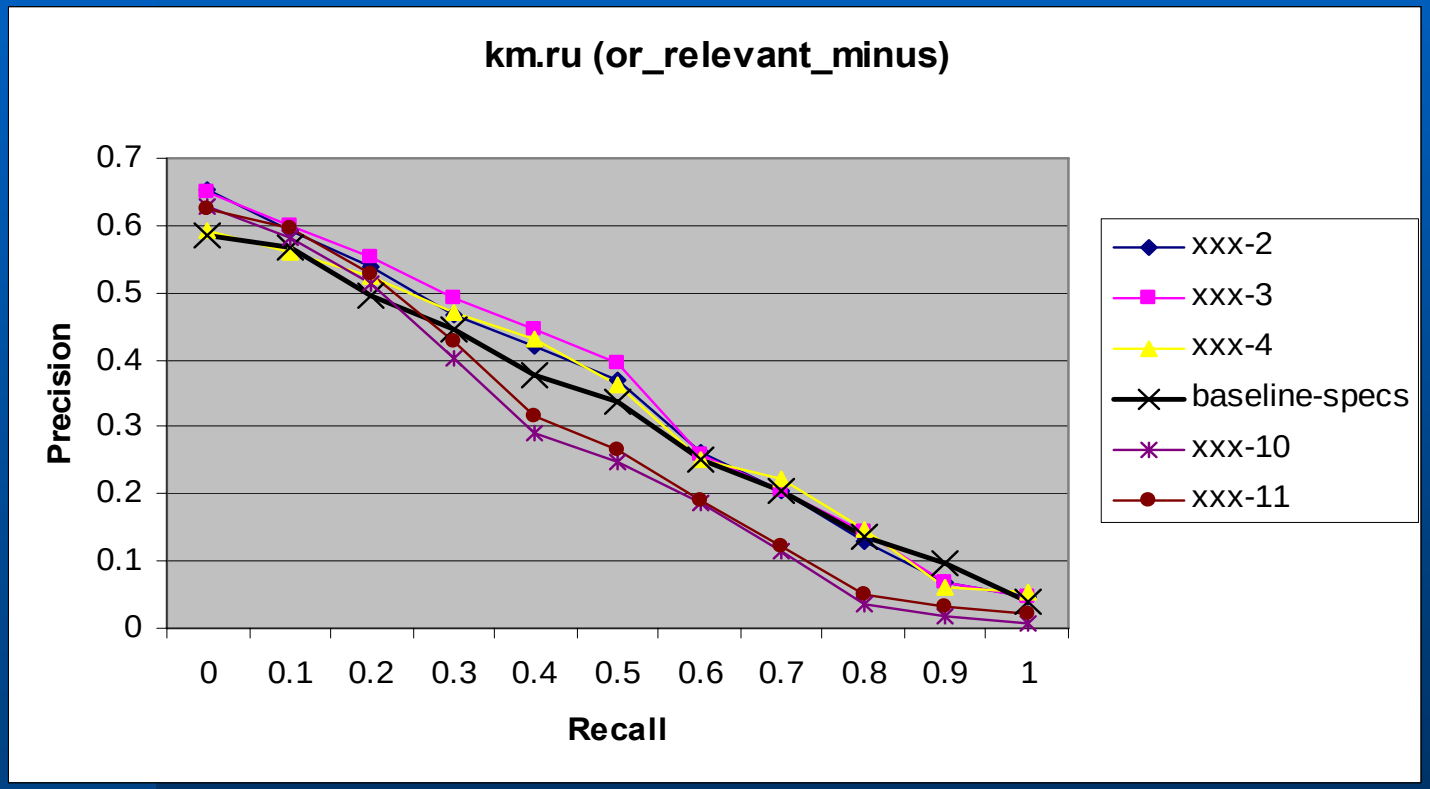
# Specs на РОМИП'2007

**Максаков Алексей**
**bruzz@yandex.ru**

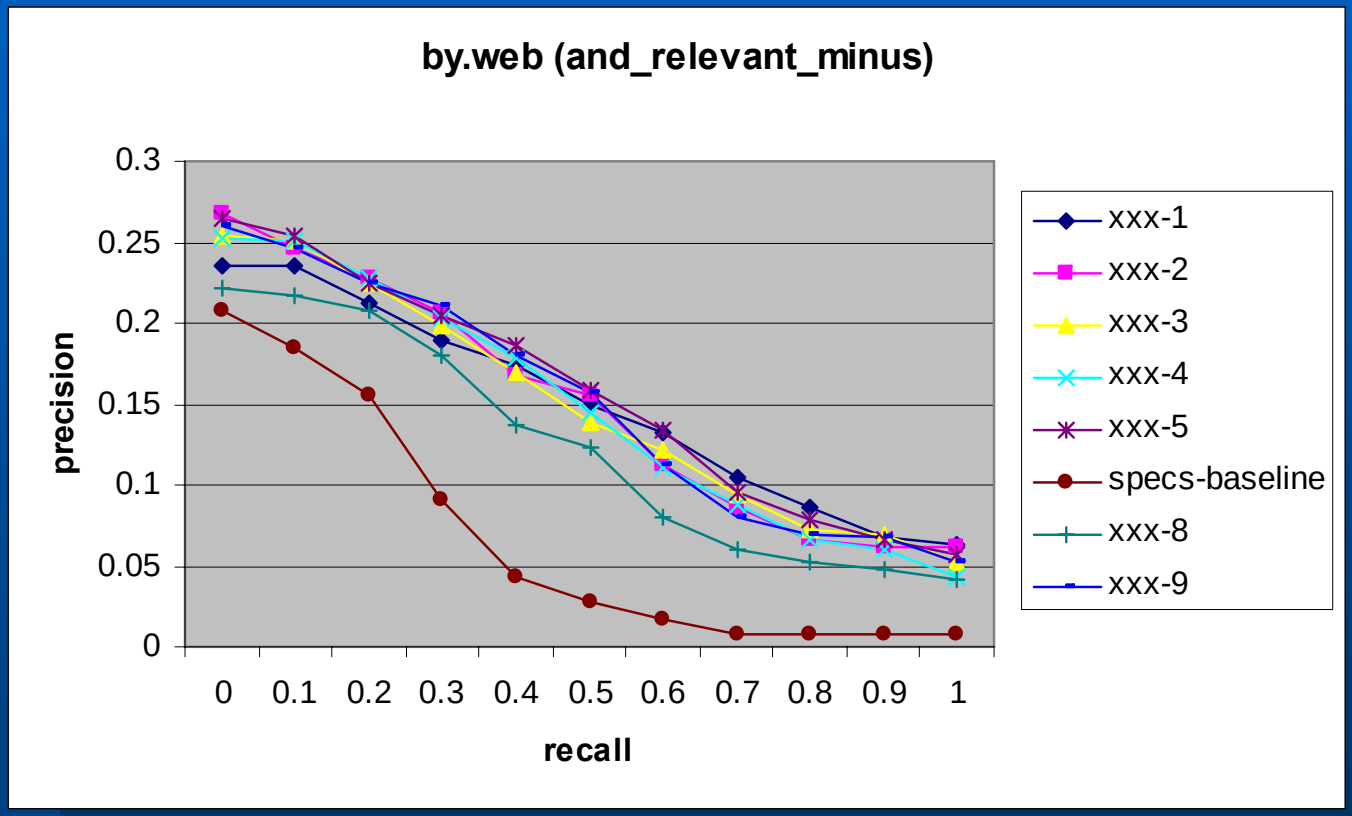
# Дорожка Web поиска

- **Традиционная оценка релевантности (Lucene)**
- **Словарный морфологический анализ**
- **Цель: сравнить “базовый” подход с представленными на РОМИП**

# Коллекция km.ru



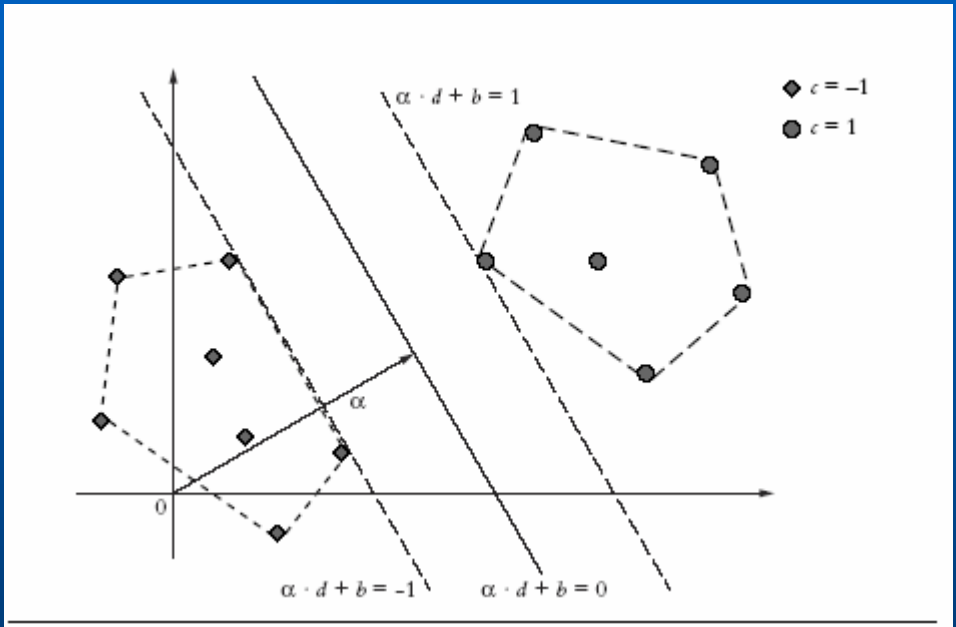
# Коллекция By.web



# Причины

- Традиционная оценка весов слов (tfidf)
- Нет оценки близости (в документе) слов из запроса
- Нет эвристического дополнения запрсов (прилагательные -> существительные, расшифровка понятий)

# Метод опорных векторов (SVM)




**минимизировать** $\frac{1}{2}\alpha^T \cdot \alpha + C_{light} \sum_i \xi_i$

**при** $c_i(\alpha \cdot x_i + b) \geq 1 - \xi_i, \forall i = 1...n$

$\xi_i \geq 0, \forall i = 1...n$ **(1)**

**Задача сводится к задаче квадратичной оптимизации:**

$$t_{об} = O(mn^{1,7})$$

**максимизировать** $\sum_i \lambda_i - \frac{1}{2}\sum_{i,j} \lambda_i \lambda_j c_i c_j (x_i \cdot x_j)$

**при** $\sum_i c_i \lambda_i = 0$

$0 \leq \lambda_i \leq C, \forall i = 1...n$

# Structural SVM

**минимизировать** $\frac{1}{2}\alpha^T \cdot \alpha + C\xi$ $\qquad C_{light} = \frac{C}{n}$

$$\forall c \in \{0,1\}^n : \frac{1}{n}\alpha^T \sum_{i=1}^{n} c_i d_i x_i \geq \frac{1}{n}\sum_{i=1}^{n} c_i - \xi \qquad \textbf{(2)}$$

$\frac{1}{n}\sum_{i=1}^{n} c_i$ **макс.доля ошибок обучения**

**Эквивалентно решению исх.задачи (1) при** $\xi^* = \frac{1}{n}\sum_{i=1}^{n} \xi_i^*$

# SVM-perf

## T. Joachims (KDD, август 2006): Training Linear SVMs in Linear Time

**Алгоритм Cutting Plane находит следующее решение задачи (2):**

$$(a, \xi + \varepsilon) \qquad C_{light} = \frac{1}{avg_i \|x_i\|} \qquad C_{perf} = \frac{n}{avg_i \|x_i\|}$$

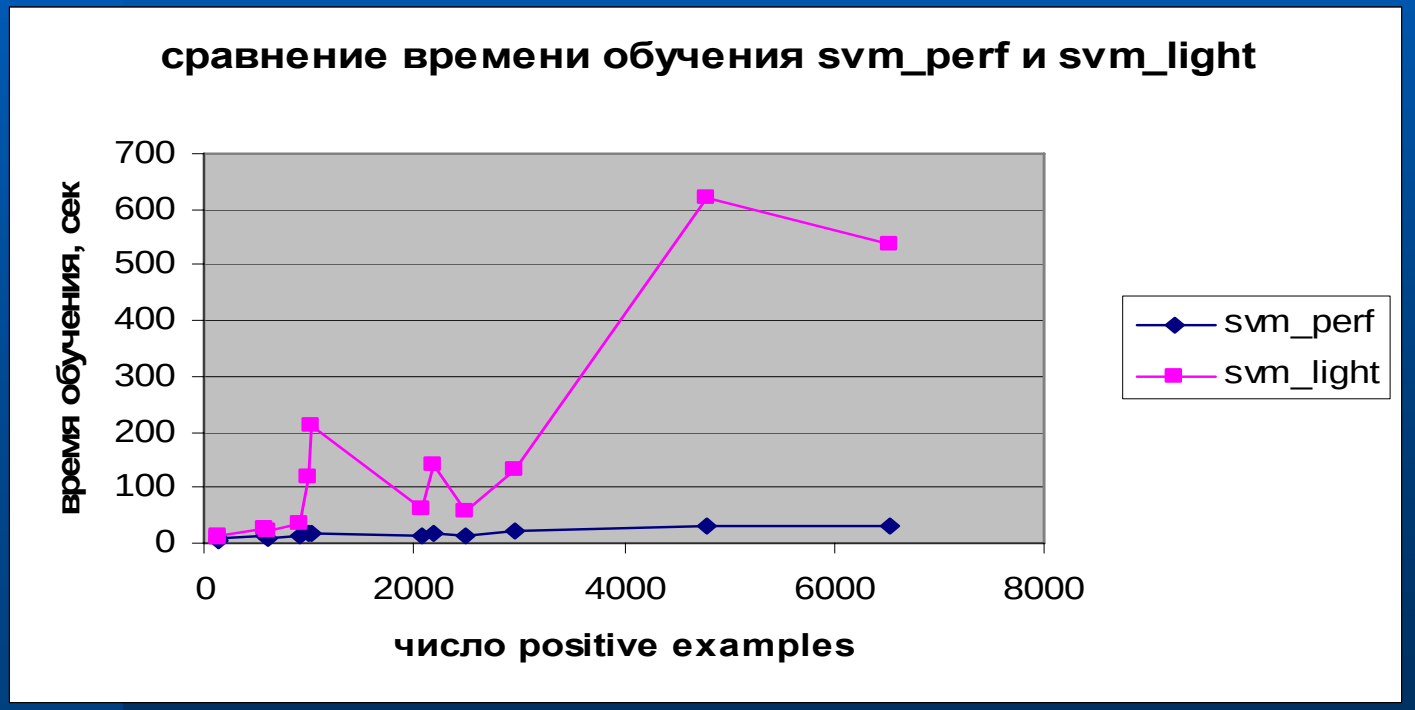
**Вычислительная сложность обучения: *O(sn)***
**1 шаг итерации: *O(sn)* операций, где s-|V|*коэффициент разреженности матрицы V×D , *n* -число примеров**

**Число итераций:** $\max\{\frac{2}{\varepsilon}, \frac{8C \max_{i=1..n} \| x_i \|}{\varepsilon^2}\}$ **(на практике 20-90)**

# Сравнение времени обучения

**|V|=80000, |D|=5200: $t_{об}$(SVM-perf)>2,5$t_{об}$(SVM-light)**

**|V|>250000, |D|=29700 (15-90 итераций):**



# Предстоит выяснить

- **Изменится ли качество классификации при применении svm_perf (всегда ли достаточно $\varepsilon = 0{,}001$ )?**
- **Имеют ли “право на жизнь” более простые алгоритмы, в том числе и при решении иерархической задачи классификации?**

# Спасибо за внимание