# SPECS на РОМИП'2007

© Максаков А.В

Московский Государственный Университет им. М.В. Ломоносова
bruzz@yandex.ru

**Аннотация**

В статье рассмотрены эксперименты по применению свободно распространяемой системы индексирования и поиска Lucene при решении задачи поиска по ad-hoc запросам. Также приведены сравнительные оценки производительности пакетов SVM-perf и SVM-light при решении задачи классификации текстов.

## 1. Введение

При проведении экспериментов по исследованию различных алгоритмов классификации текстов на предыдущих семинарах библиотека Lucene [1] использовалась для обеспечения оперативной индексации больших объемов данных, а также получения необходимой статистической информации. Было бы досадным упущением не провести оценку алгоритмов ранжирования и поиска по ключевым словам, относящихся к основной функциональности этой библиотеки. Целью участия на РОМИП в 2007 году было сравнение алгоритмов, реализованных в Lucene с другими алгоритмами, представленными на семинаре.

Также, в 2006 году был опубликован алгоритм SVM-perf [2] с линейной зависимостью времени обучения от числа примеров, что теоретически позволяет применять его при обучении на больших коллекциях документов. Но поскольку этот алгоритм является итерационным и теоретическая верхняя граница числа итераций весьма высока, практический интерес представляет сравнение этого алгоритма по времени обучения с традиционным SVM [3] на реальных данных.

## 2. Дорожка поиска по ad-hoc запросу

При проведении экспериментов на дорожке поиска по запросу использовалась библиотека Lucene с традиционной TFIDF оценкой релевантности, без внесения каких либо изменений в алгоритм ранжирования, реализованный в библиотеке.

Однако на этапе предварительной обработки документов был привнесен ряд изменений.

Во-первых, поставляемый лемматизатор, основанный на отсечении окончаний слов, был заменен на собственный словарный морфологический анализатор, основанный на словарях проекта aot.ru.

Также при определении начальной формы слов для разрешения морфологической омонимии использовался синтаксический анализатор, созданный на основе алгоритма, описанного в [4]. При этом для достижения приемлемой производительности был упрощен фрагментационный анализ.

### 2.1 Результаты экспериментов

Асессорам в РОМИП было представлено два прогона: по одному для коллекций km.ru и by.web. Прогоны проводились по отдельным индексам для каждой коллекции, так что для поиска по

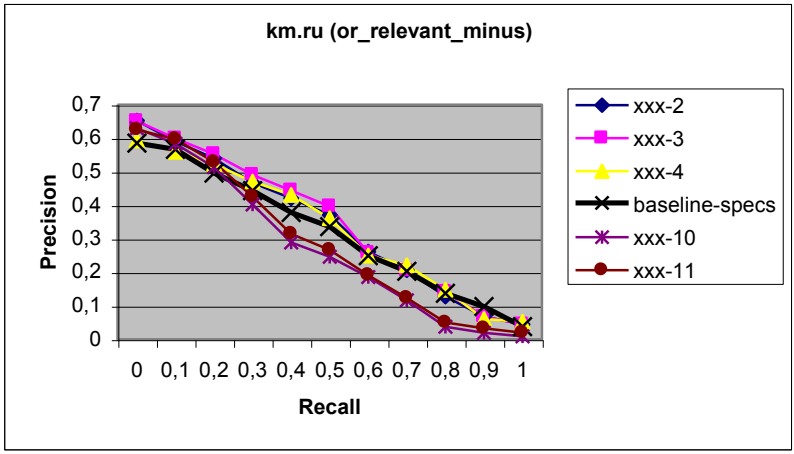


рис.1. 11-точечный графики TREC участников дорожки поиска по коллекции km.ru с оценкой or_relevant_minus

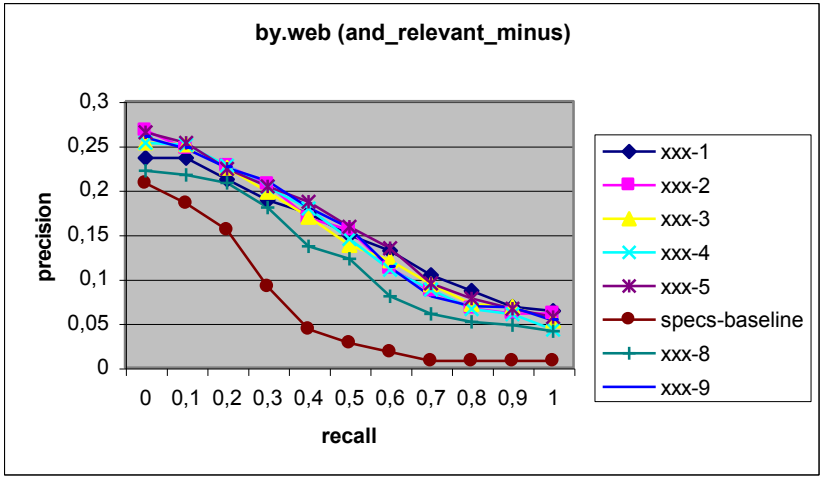


рис.2. 11-точечный графики TREC участников дорожки поиска по коллекции by.web с оценкой or_relevant_minus

смешанной коллекции результаты по каждому из прогонов ожидаемо низкие. 11-точечные графики TREC по каждому из прогонов приведены на следующих рисунках.

Несмотря на простоту использованных решений, и способов оценки близости документов запросу, на коллекции km.ru были получены вполне достойные результаты.

Следует отметить, что результаты прогона по коллекции by.web могли быть существенно ухудшены в результате технической накладки. Дополнительно проверить качество поиска по этой коллекции планируется в следующем году.

### 2.2 Анализ полученных результатов

Несмотря на то, что для коллекции km.ru по метрике or_relevant_minus были получены хорошие результаты, относительное качество поиска по метрике and_relevant_minus оказалось существенно хуже.

Проведенный анализ результатов поиска по оцениваемым запросам показал, что сравнительное ухудшение качества поиска наблюдалось в тех случаях, когда слова в запросе сильно связаны и

формируют устойчивые словосочетания или термин, использующийся в запросе, имеет очевидные синонимы.

В связи с этим, для улучшения качества поиска необходимо учитывать расстояние между терминами при оценке релевантности, использовать методы расширения запросов.

## 3. Классификация нормативно-правовых документов

На коллекции нормативно-правовых документов производилось сравнительное исследование времени обучения алгоритма SVM с линейным ядром и итерационного алгоритма SVM-perf, обеспечивающего линейную зависимость времени обучения от числа примеров в обучающей выборке.

Традиционный алгоритм SVM [3] решает следующую задачу оптимизации (1): необходимо минимизировать

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

при

$$c_i(\alpha \cdot x_i + b) \geq 1 - \xi_i, \forall i = 1...n$$
$$\xi_i \geq 0, \forall i = 1...n$$

В структурной SVM решается задача минимизации (2):

$$\frac{1}{2}\alpha^T \cdot \alpha + C\xi$$
$$\forall c \in \{0,1\}^n : \frac{1}{n}\alpha^T \sum_{i=1}^{n} c_i d_i x_i \geq \frac{1}{n}\sum_{i=1}^{n} c_i - \xi$$

при этом

$$C_{light} = \frac{C}{n}$$

что эквивалентно решению задачи (1) при

$$\xi = \frac{1}{n}\sum_{i=1}^{n} \xi_i$$

Алгоритм Cutting Plane находит следующее приближенное решение задачи (2):при этом число итераций алгоритма в рамках

$$(a, \xi + \varepsilon)$$

$$\max\{\frac{2}{\varepsilon}, \frac{8C \max_{i=1..n} \| x_i \|}{\varepsilon^2}\}$$

SVM-perf ограничено сверху

В статье [2] показано, что на ряде коллекций

$$\varepsilon = 0{,}001$$

достаточно для получения качества классификации близкого к качеству SVM при max $\|x_i\|$ = 1. Однако по причине наличия большого числа итераций, время обучения SVM-perf может быть больше времени обучения обычного SVM с линейным ядром. В частности на подмножестве коллекции нормативно-правовых документов образца 2004-2006 годов (число документов в обучающей выборке – около 5400, размерность пространства признаков – 80000) время обучения SVM-perf было в среднем более чем в 2,5 раза больше времени обучения метода опорных векторов в пакете SVM-light. На коллекции нормативно-правовых документов 2007 года (число документов – 29700, размерность пространства признаков – более 250000) SVM-perf оказался минимум на порядок быстрее в обучении, чем SVM-light. Было обнаружено, что выигрыш от применения SVM-perf оказывался большим в случае обучения на более сбалансированной выборке (см. Рис 3).

Приведенные наблюдения позволяют предположить, что наряду с общим числом примеров в обучающей выборке, критерием при выборе между алгоритмами SVM-light и SVM-perf в задачах, имеющих ограничения на время обучения классификаторов, может быть и сбалансированность этой выборки.

## 4. Заключение

В этом году участие в семинаре, к сожалению, оказалось далеко не таким продуктивным, каким бы оно могло быть. Тем не менее, достаточно положительные результаты применения средств поиска и индексирования с открытым кодом оказались приятной неожиданностью.

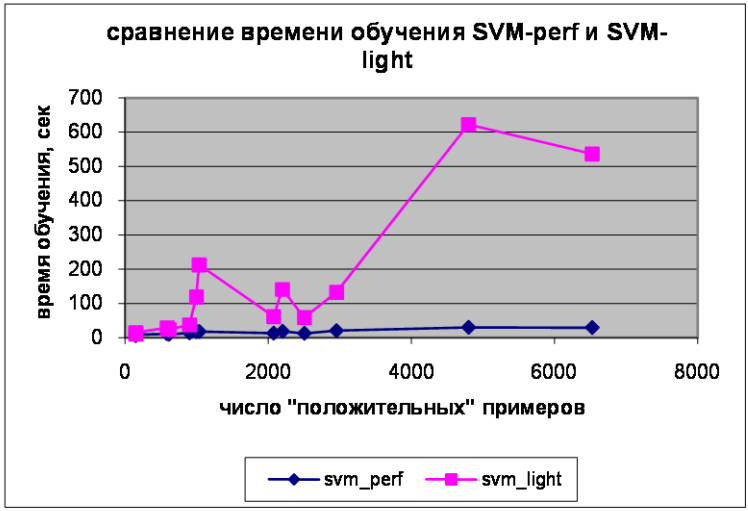


**Рис.3. Зависимость времени обучения алгоритмов SVM-perf и SVM-light в зависимости от числа “положительных” примеров.**

## Литература


[1] Lucene web site. http://lucene.apache.org

[2] T. Joachims. Training Linear SVMs in linear time. *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2006

[3] T. Joachims. Making Large-Scale SVM Learning Practical. *Advances in Kernel Methods. Support Vector Learning*. MIT-Press, 1999.

[4] Сокирко А. Семантические словари в автоматической обработке текста (по материалам системы ДИАЛИНГ).// Дис. канд. техн. наук: 05.13.17. Москва, 2001.


## SPECS at RIRES’2007


A. Maksakov



In this paper results of experiments with open-source indexing and search library Lucene in ad-hoc retrieval task are discussed. Also comparative training performance analysis between SVM-perf and traditional SVM algorithm is made.