Фильтрация и тематическое моделирование коллекции научных документов

Сергей Воронов

Научный руководитель: д.ф.-м.н. К.В.Воронцов

Московский физико-технический институт (государственный университет)
Факультет управления и прикладной математики
Кафедра «Интеллектуальные системы»

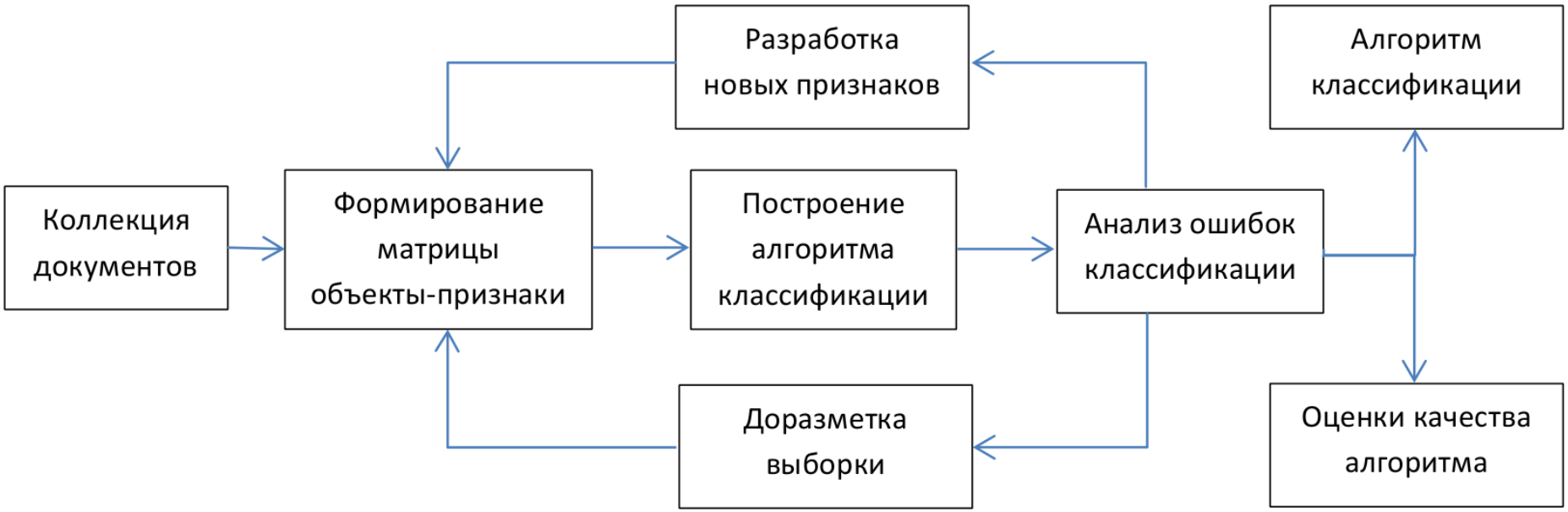
Москва, 2014 г.

Постановка задачи

- **Конечная цель**: создать систему тематического поиска научного контента в Интернете
- **Цель данного исследования**: разработать обучаемый алгоритм распознавания научных документов
- **Сложность задачи**:
  - Большой размер коллекции
  - Несбалансированная выборка (научных документов $\leq 2\%$)
  - Изначально нет системы признаков
  - Изначально нет представительной обучающей выборки

Процесс разработки алгоритма классификации

Автоматизация процесса постепенного наращивания обучающей выборки, разработки признаков и улучшения классификатора.



Линейный классификатор с ненастраиваемыми весами

Линейный алгоритм классификации:

$$a(x, w) = \operatorname{sign}\left(\sum_{j=1}^{n} w_j f_j(x) - w_0\right),$$

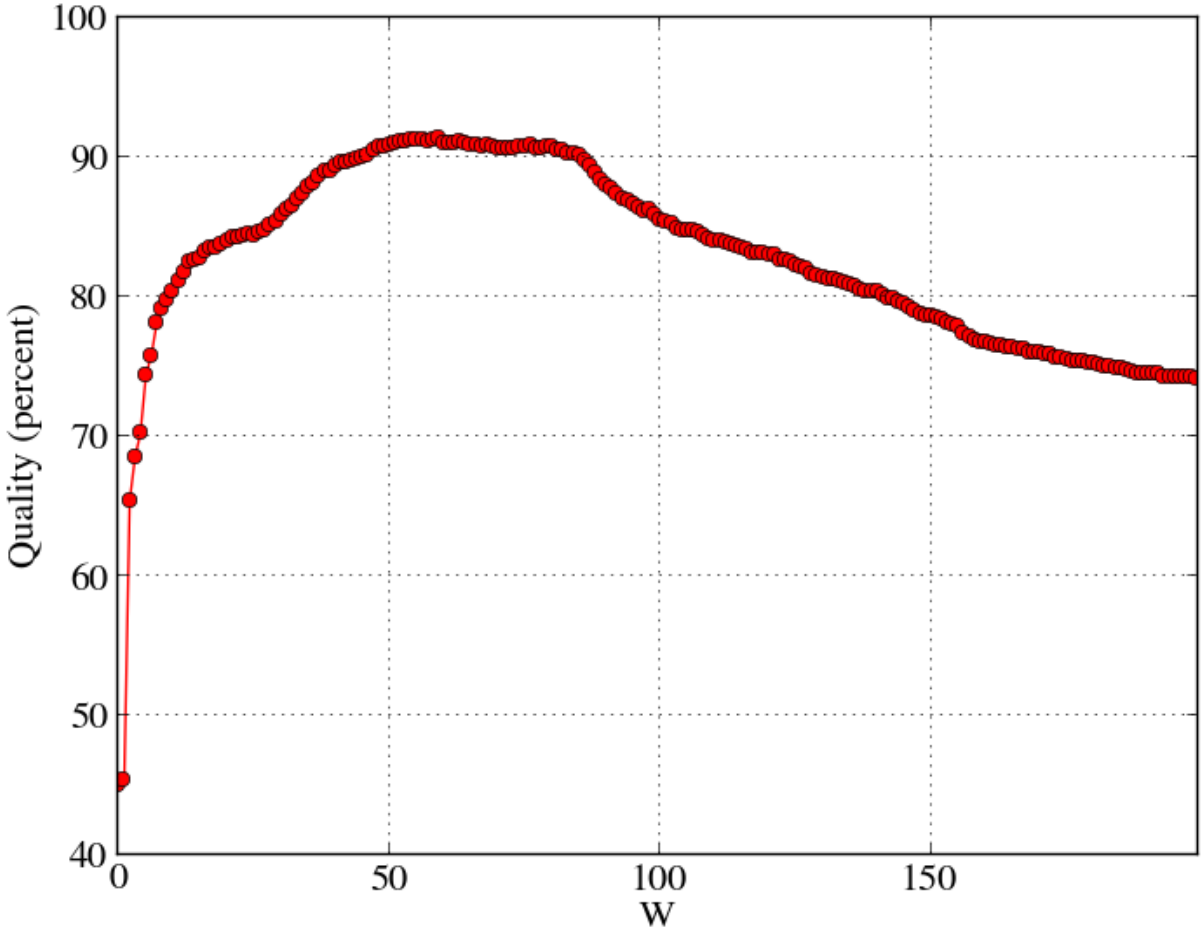
где w_j — вес j-го признака, w_0 — порог принятия решения,
$w = (w_1, \ldots, w_n)$ — вектор весов признаков.

В базовой необучаемой версии w_i задаются экспертом.

Использовался для разметки первоначальной выборки; после этого использовались SVM и регуляризованная логистическая регрессия.

Зависимость качества базового классификатора от w_0

Качество — это доля верных классификаций.



Вывод: слишком высокий процент ошибок.

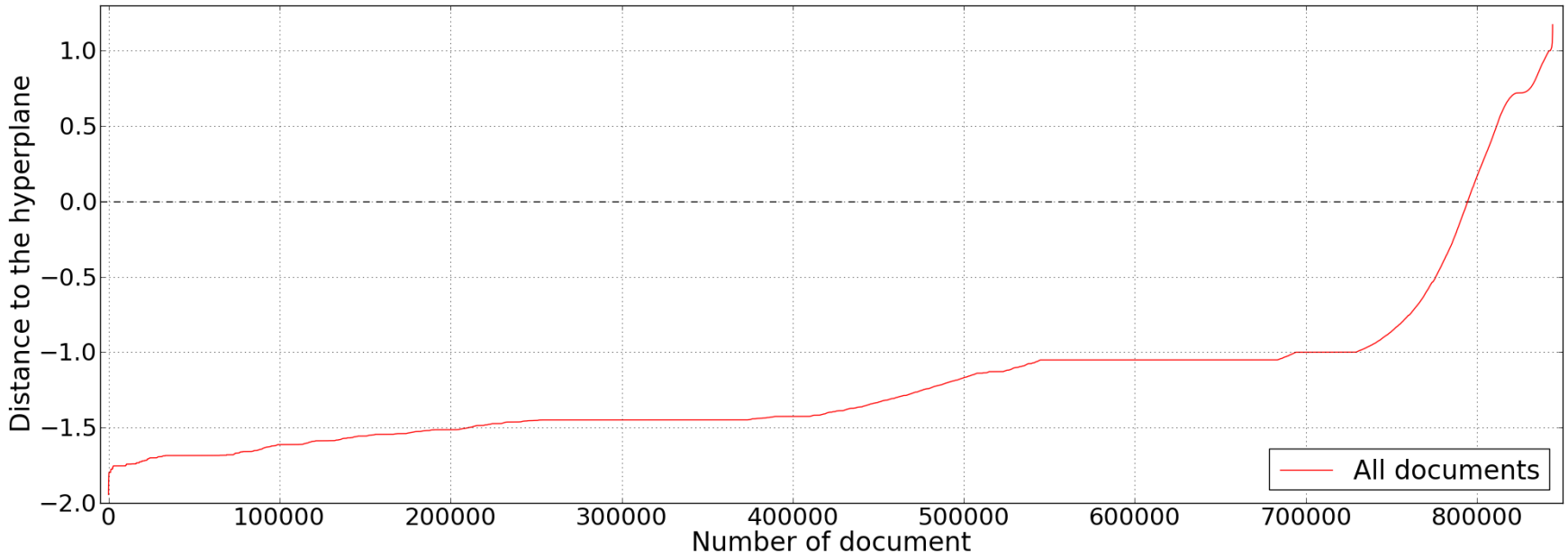
Признаковое описание документов

1. Число греческих букв в тексте
2. Логарифм длины текста в символах
3. Число математических символов в тексте
4. Научные термины
5. Типовые фразы, относящиеся к грантам («работа выполнена при поддержке РФФИ» и т.д.)
6. Имена, часто встречающиеся в научных работах (именные теоремы, критерии, ...)
7. Слова, относящиеся к оформлению (редактор, оппонент и т.д.)
8. Минус-слова, присутствие которых понижает “научность” (количество часов, профком, экзамен)
9. Длина текста (длина текста больше заданного порога)
10. Число цифр в документе

Проблема несбалансированности выборки

Проблема: доля научных документов $< 2\%$

В обучающую выборку лучше брать объекты из граничной зоны



Решение: оценивать ошибку классификации по формуле

Хансена-Гурвица: $Avr(f) = \frac{1}{n}\sum\limits_{d\sim p} f(d) \quad \to \quad Avr(f) = \frac{1}{n}\sum\limits_{d\sim q} f(d)\frac{p(d)}{q(d)},$

$p(d)$ – изначальное распределение документов, $q(d)$ – измененное.

SVM и RLR

Линейный алгоритм классификации:

$$a(x, w) = \operatorname{sign}\left(\sum_{j=1}^{n} w_j f_j(x) - w_0\right),$$

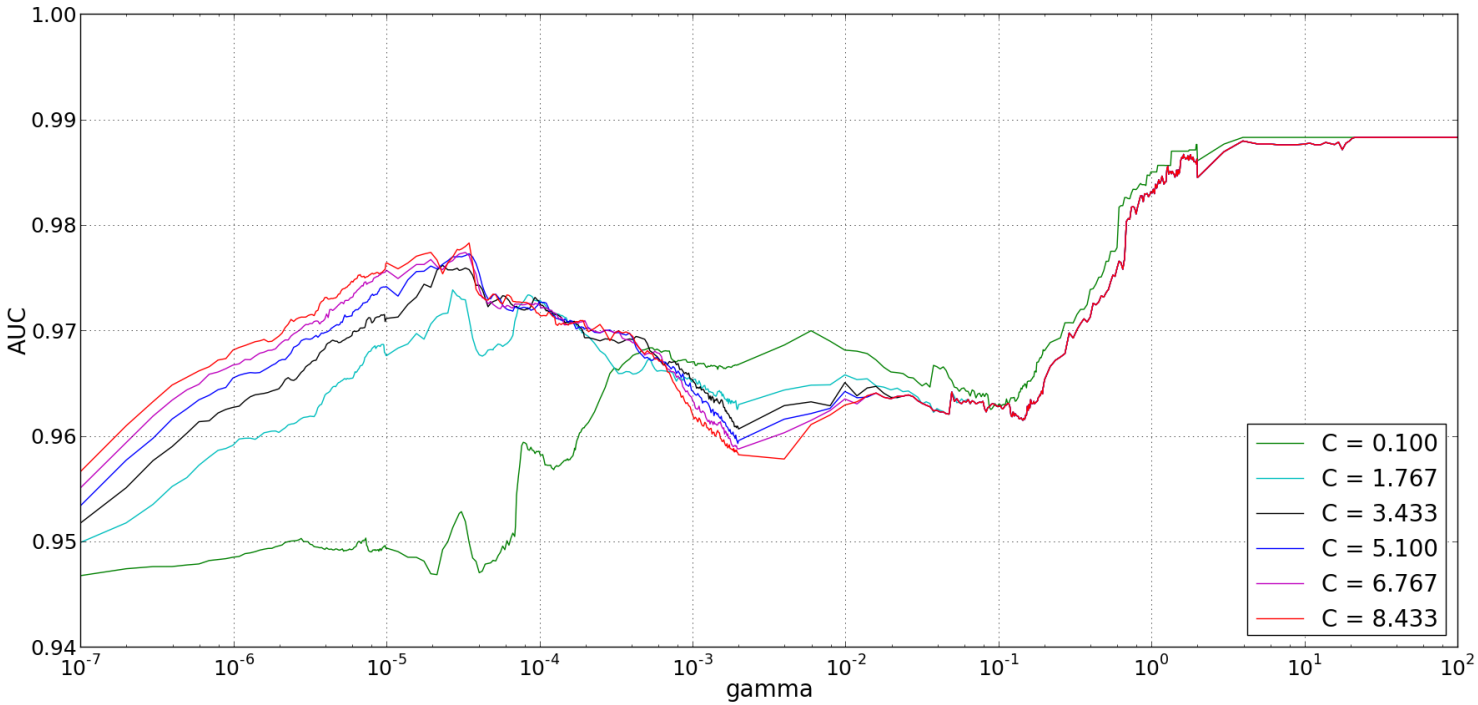
w_i определяются из условия:

SVM: $$Q(w, w_0) = \sum_{i=1}^{m} (1 - y_i(\langle x_i, w\rangle - w_0))_+ + \frac{1}{2C}|w|^2 \to \min_{w, w_0}.$$

RLR: $$Q(w) = \sum_{i=1}^{m} \ln\left(1 + \exp(-y_i\langle x_i, w\rangle)\right) + \frac{\lambda}{2}|w|^2 \to \min_{w}$$

Подбор параметров SVM

- ядро rbf (радиальные базисные функции, $e^{-\gamma(x-x_0)^2}$)
- γ и C (штраф за неверную классификацию) — по скользящему контролю



Оптимальные параметры: $C > 4$ и $\gamma > 10$.

Различные алгоритмы настройки весов RLR

- Градиентный спуск

$$w = w - \alpha \left(\frac{\partial Q(w)}{\partial w} \right)$$

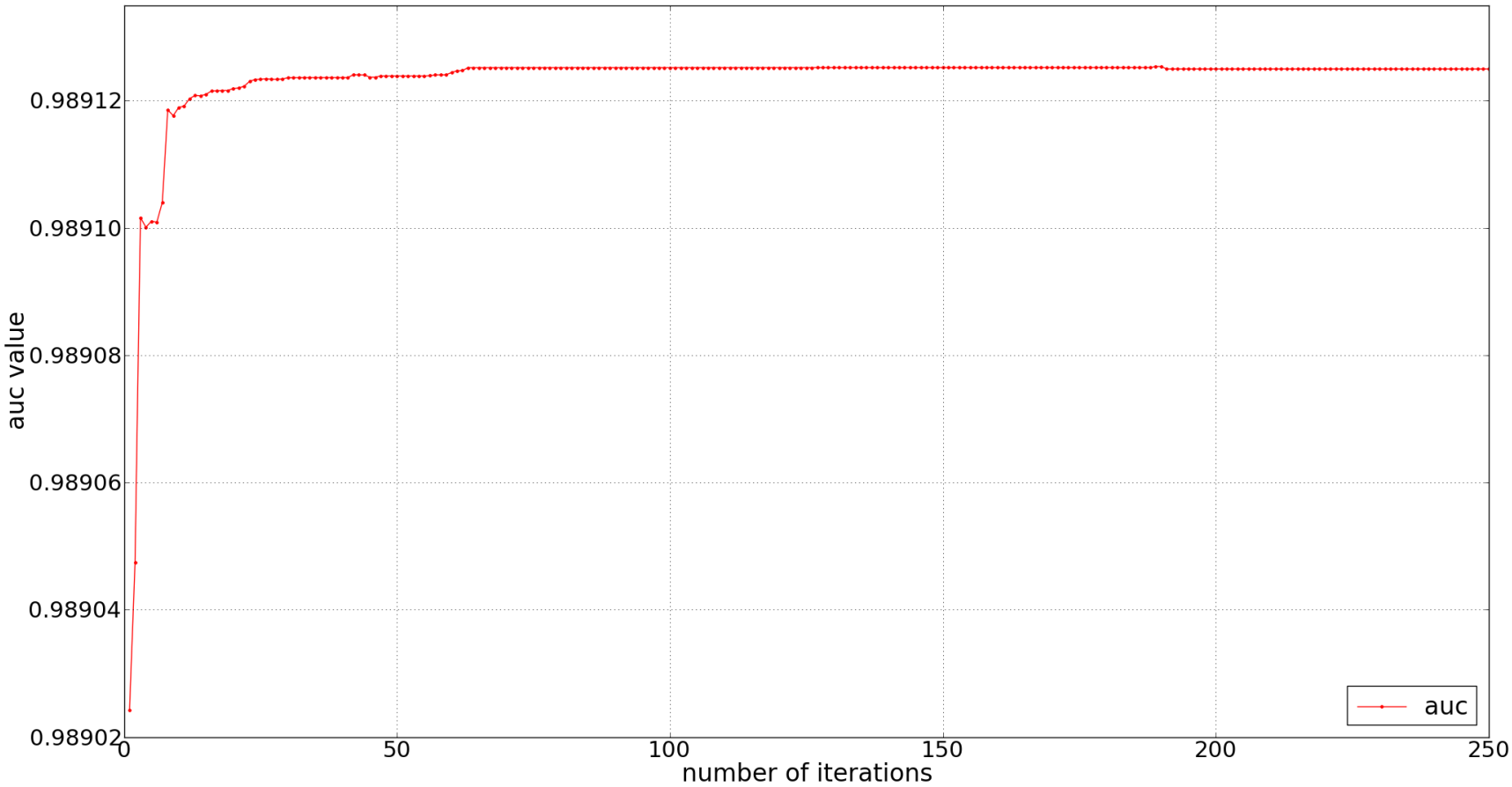
- Метод Ньютона

$$w = w - \alpha \left(\frac{\partial^2 Q(w)}{\partial w \partial w^T} \right)^{-1} \frac{\partial Q(w)}{\partial w}$$

- Метод Левенберга-Марквардта

$$w_i = w_i - \alpha \left(\mu + \frac{\partial^2 Q(w)}{\partial w_i \partial w_i} \right)^{-1} \frac{\partial Q(w)}{\partial w_i}$$

Настройка параметров LR (количество итераций)



Вывод: после 50 итераций не происходит существенного изменения AUC

Применение тематического моделирования

Один из информативных признаков — доля научных научных слов.

Проблема: сформировать словари научных слов вручную затруднительно.

Тривиальный подход: считать каждое слово за признак; затем сделать отбор признаков.

Проблема: большая трудоемкость (>70000 слов, встречающихся в нескольких документах).

Предлагаемый подход: автоматическая генерация признаков по обучающей коллекции методом тематического моделирования.

Тематическая модель классификации

Тематическая модель появления слов в документах:

$$p(w|d) = \sum_{t \in T} p(w|t)p(t|d) = \sum_{t \in T} \varphi_{wt}\theta_{td}$$

Тематическая модель классификации документов:

$$p(c|d) = \sum_{t \in T} p(c|t)p(t|d) = \sum_{t \in T} \psi_{ct}\theta_{td}$$

где c – класс, w – слово, t – тема, d – документ коллекции.

Задача максимизации регуляризованного правдоподобия

$$\sum_{d,w} n_{dw} \ln \sum_{t} \varphi_{wt}\theta_{td} + \tau \sum_{d,c} m_{dc} \ln \sum_{t} \psi_{ct}\theta_{td} + \sum_{i} \tau_i R_i(\Phi, \Psi, \Theta) \to \max$$

Регуляризаторы

Аддитивная регуляризация тематической модели (ARTM):

- Разреживание:

$$R_1(\Phi) = -\beta \sum_{t \in T} \sum_{w \in W} \ln \varphi_{wt}$$

$$R_2(\Psi) = -\gamma \sum_{t \in T} \sum_{c \in C} \ln \psi_{ct}$$

$$R_3(\Theta) = -\alpha \sum_{d \in D} \sum_{t \in T} \ln \theta_{td}$$

- Ковариационный регуляризатор (повышение различности тем)

$$R(\Phi, \Psi, \Theta) = -\frac{\tau}{2} \sum_{t \in T} \sum_{s \in T \setminus t} cov(\varphi_t, \varphi_s)$$

EM-алгоритм для ARTM в матричной форме

Алгоритм 1 EM-алгоритм для ARTM, одна итерация, матричный вид

1: $\Phi_{new} = \Phi \otimes \left[N^T \oslash (\Phi\Theta)\right]\Theta^T$

2: $\Psi_{new} = \Psi \otimes \left[M^T \oslash (\Psi\Theta)\right]\Theta^T$

3: $\Theta_{new} = \Theta \otimes \Phi^T\left[N^T \oslash (\Phi\Theta)\right] + \tau\Theta \otimes \Psi^T\left[M^T \oslash (\Psi\Theta)\right]$

4: Нормировка столбцов $\Phi_{new}, \Psi_{new}, \Theta_{new}$.

5: $\Omega = \sum\limits_{t\in T} \varphi_{wt}$

6: Регуляризация Φ : $\Phi_{new} = (\Phi_{new} - \beta 1_{W\times T} - \eta\Phi \otimes [\Omega 1_{1\times T} - \Phi])_+$

7: Регуляризация Ψ : $\Psi_{new} = (\Psi_{new} - \gamma 1_{C\times T})_+$

8: Регуляризация Θ : $\Theta_{new} = (\Theta_{new} - \alpha 1_{T\times D})_+$

9: Нормировка столбцов $\Phi_{new}, \Psi_{new}, \Theta_{new}$.

Примеры научных тем

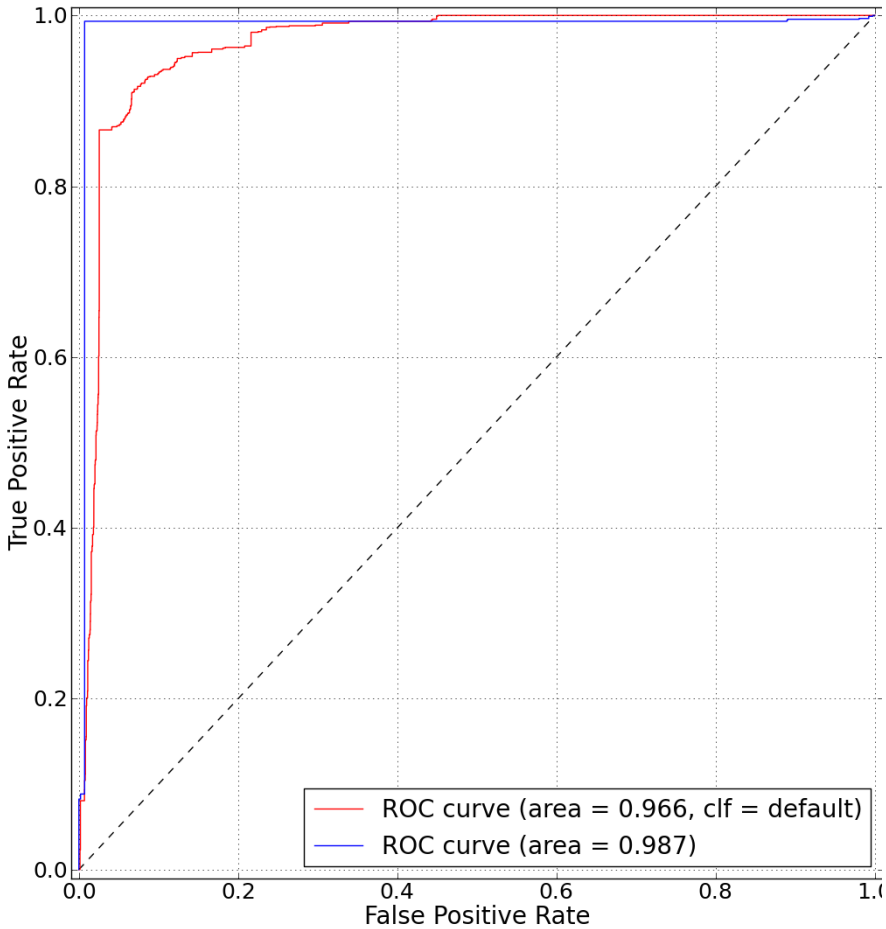
$c = -1$	0.495	0.544	0.983	0.932	1.000	0.117	1.000	0.605	...
$c = +1$	0.505	0.456	0.017	0.068	0.000	0.883	0.000	0.395	...

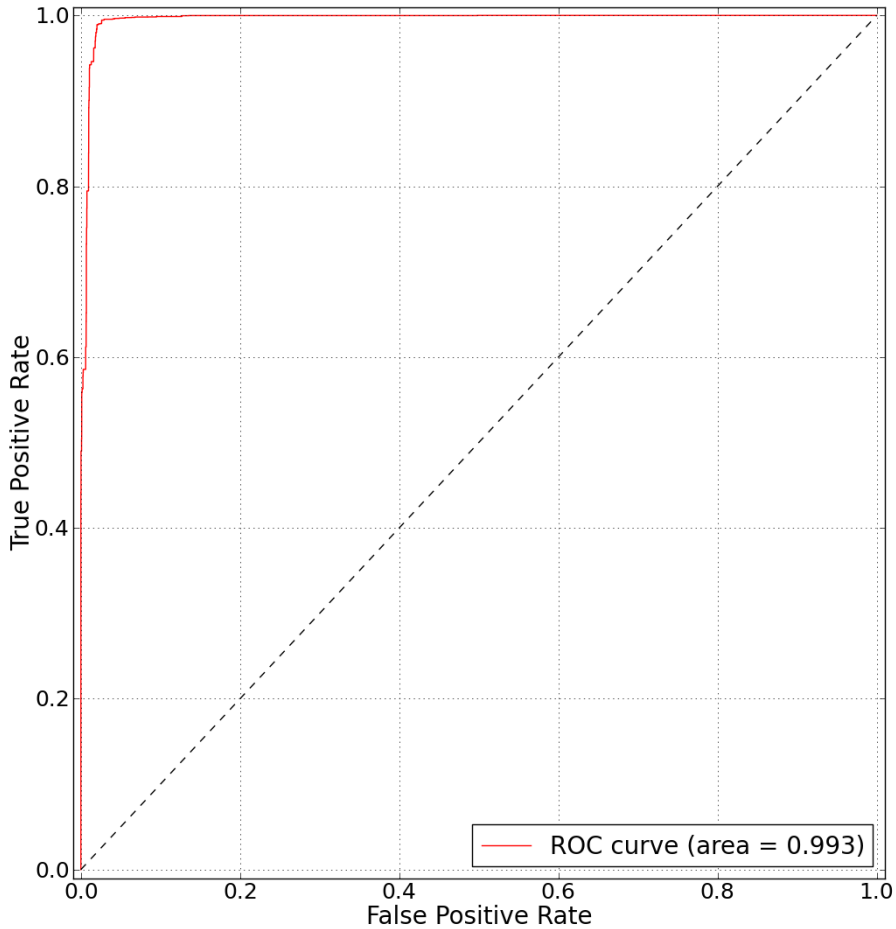
30-40 наиболее часто встречающихся в научной теме слов — признак.
Примеры топа слов двух из отобранных научных тем:

p_w	w
0.0575	система
0.0450	рис
0.0305	модель
0.0299	функция
0.0286	значение
0.0284	параметр
0.0235	характеристика
0.0232	уравнение
0.0230	процесс
⋮	⋮

p_w	w
0.0398	результат
0.0342	анализ
0.0296	вид
0.0296	исследование
0.0286	решение
0.0279	метод
0.0245	задача
0.0217	использование
0.0201	фактор
⋮	⋮

ROC-кривая для SVM и RLR (после добавления признаков)





Результаты

	Стадия работы	Ошибка	AUC
	Базовый классификатор	10,2%	-
SVM	Базовая версия	9,6%	0,91
	Настройка параметров	8,0%	0.93
	Признаки научных слов	7,4%	0.95
	Улучшение признаков	4,0%	0,983
	+ признаки из TM	3,7%	0.985
RLR	Градиентный спуск	5,0% (3,2–6,3)	0.981–0.991
	Метод Левенберга-Марквардта	4.6% (2,7–6.6)	0.983–0.992
	+ признаки из TM (гр. сп.)	5,2% (3,3–6,4)	0.976–0.994
	+ признаки из TM (Л-М)	4,0% (2,2–5,0)	0.985–0.996

Результаты, выносимые на защиту

- Разработана система признаков для линейной модели распознавания научных документов
- Предложен метод формирования словарных признаков на основе регуляризованной тематической модели
- Выполнена программная реализация и проведены численные эксперименты показавшие, что использование данных признаков улучшает качество классификации