Двухступенчатые модели и проблема переобучения в латентном семантическом анализе

В. А. Лексин

Московский Физико-Технический Институт

Научный руководитель К. В. Воронцов

Определения и обозначения

U — множество субъектов (клиентов, пользователей: users);
R — множество объектов (ресурсов, товаров, предметов: items);
Y — пространство описаний транзакций;

Сырые исходные данные:

$D = (u_i, r_i, y_i)_{i=1}^m \in U \times R \times Y$ — протокол транзакций;

Агрегированные данные:

$F = \|f_{ur}\|$ — матрица кросс-табуляции размера $|U| \times |R|$,
где $f_{ur} = \text{aggr}\{(u_i, r_i, y_i) \in D \mid u_i = u,\ r_i = r\}$

Задачи:

- прогнозирование незаполненных ячеек f_{ur};
- оценивание сходства: $K(u, u')$, $K(r, r')$, $K(u, r)$;
- выявление скрытых интересов $p(t|u)$, $q(t|r)$ относительно заданного либо неизвестного набора тем $t = 1, \ldots, T$.

Два основных подхода

1. **Анамнестические алгоритмы**
   (Memory-Based Collaborative Filtering)
   - хранение всей исходной матрицы данных F;
   - сходство клиентов — это корреляция строк матрицы F;
   - сходство объектов — это корреляция столбцов матрицы F.

2. **Модельные алгоритмы**
   (Model-Based Collaborative Filtering)
   - оценивание профилей клиентов и объектов (*профиль — это вектор скрытых характеристик*);
   - хранение профилей вместо хранения F;
   - сходство клиентов и объектов — это сходство их профилей.

Модельные алгоритмы

По данным D оцениваются векторы:
$(p_{tu})_{t \in T}$ — профили клиентов $u \in U$;
$(q_{tr})_{t \in T}$ — профили объектов $r \in R$.

Типы модельных алгоритмов:

1. Вероятностный латентный семантический анализ.
2. Латентный семантический анализ (матричные разложения).
3. Двухступенчатая (симметризованная) вероятностная латентная модель.

Латентный семантический анализ (матричные разложения)

T — множество тем (интересов): $|T| \ll |U|$, $|T| \ll |R|$;

p_{tu} — неизвестный *профиль клиента* u; $\quad P = \left(p_{tu}\right)_{|T|\times|U|}$;

q_{tr} — неизвестный *профиль объекта* r; $\quad Q = \left(q_{tr}\right)_{|T|\times|R|}$;

Задача: найти разложение $f_{ur} = \sum\limits_{t\in T} \lambda_t p_{tu} q_{tr}$; $\quad F = P^{\mathrm{T}}\Lambda Q$;

Методы решения:

SVD — сингулярное разложение (плохо интерпретируется!);

NNMF — неотрицательное разложение: $p_{tu} \geqslant 0$, $q_{tr} \geqslant 0$;

Вероятностная интерпретация:

$$\underbrace{p(u,r)}_{f_{ur}?} = \sum_{t\in T} \underbrace{p(t)}_{\lambda_t} \cdot \underbrace{p(u|t)}_{p_{tu}} \cdot \underbrace{q(r|t)}_{q_{tr}};$$

$$q(t|r) = \frac{q_{tr}p(t)}{\sum\limits_{\tau\in T} q_{\tau r}p(\tau)}; \quad p(t|u) = \frac{p_{tu}p(t)}{\sum\limits_{\tau\in T} p_{\tau u}p(\tau)}$$

Вероятностный латентный семантический анализ

Вероятностная модель:

$$p(u, r) = \sum_{t \in T} p(t) p(u|t) q(r|t),$$

$p(t)$ — априорная вероятность темы t;
$p(u|t)$ — апостериорное распределение клиентов по теме t;
$q(r|t)$ — апостериорное распределение ресурсов по теме t.

Задача максимизации правдоподобия:

$$L = \ln \prod_{u \in U} \prod_{r \in R} p(u, r)^{f_{ur}} = \sum_{u \in U} \sum_{r \in R} f_{ur} \ln p(u, r) \to \max_{p(t),\, p(u|t),\, q(r|t)}.$$

Вероятностный латентный семантический анализ. EM-алгоритм

E-шаг:

$$p(t|u,r) = \frac{p(t)p(u|t)q(r|t)}{\sum\limits_{t' \in T} p(t')p(u|t')q(r|t')}, \quad u \in U,\ r \in R,\ t \in T.$$

M-шаг:

$$p(t) = \frac{\sum_{u \in U} \sum_{r \in R} f_{ur} p(t|u,r)}{\sum_{u \in U} \sum_{r \in R} f_{ur}},$$

$$q(r|t) = \frac{\sum_{u \in U} f_{ur} p(t|u,r)}{\sum_{u \in U} \sum_{r' \in R} f_{ur'} p(t|u,r')},$$

$$p(u|t) = \frac{\sum\limits_{r \in R} f_{ur} p(t|u,r)}{\sum\limits_{u' \in U} \sum\limits_{r \in R} f_{u'r} p(t|u',r)}.$$

Двухступенчатая (симметризованная) вероятностная латентная модель

T — множество тем (интересов);
$p_{tu} = p(t|u)$ — неизвестный *профиль клиента* u;
$q_{tr} = q(t|r)$ — неизвестный *профиль объекта* r;
$p_u = p(u)$ — априорная вероятность клиента u;
$q_r = q(r)$ — априорная вероятность объекта r;

Вероятность посещения (u, r) записывается двумя способами:

$$p(u,r) = \begin{cases} \sum\limits_{t\in T} p_u p_{tu} q(r|t,u); \quad q(r|t) = \dfrac{q_{tr} q_r}{\sum_{r'\in R} q_{tr'} q_{r'}}; \\ \sum\limits_{t\in T} q_r q_{tr} p(u|t,r); \quad p(u|t) = \dfrac{p_{tu} p_u}{\sum_{u'\in U} p_{tu'} p_{u'}}; \end{cases}$$

Задача: оценить профили p_{tu}, q_{tr}.

Принцип максимума правдоподобия: $\sum\limits_{i=1}^{m} \ln p(u_i, r_i) \to \max\limits_{p_{tu}, q_{tr}}$.

Общая идея: алгоритм согласования профилей

Повторять итерации, пока профили не сойдутся:

1. Настройка профилей клиентов p_{tu} при фиксированных q_{tr}:

$$\begin{cases} \sum\limits_{i=1}^{m} \ln\left(\sum\limits_{t\in T} p_u p_{tu} q(r|t)\right) \to \max\limits_{p_{tu}}; \\ \sum\limits_{t\in T} p_{tu} = 1, \quad \forall u \in U; \end{cases}$$

2. Настройка профилей объектов q_{tr} при фиксированных p_{tu}:

$$\begin{cases} \sum\limits_{i=1}^{m} \ln\left(\sum\limits_{t\in T} q_r q_{tr} p(u|t)\right) \to \max\limits_{q_{tr}}; \\ \sum\limits_{t\in T} q_{tr} = 1, \quad \forall r \in R; \end{cases}$$

EM-алгоритм (настройка профилей клиентов)

Скрытые переменные $H_{tr}(u) \equiv p(t|r,u)$ — апостериорная вероятность темы t при посещении объекта r клиентом u.

EM-алгоритм:
повторять, пока профили p_{tu} не сойдутся

- **E-шаг** (вычисление скрытых переменных):
  для всех объектов $r \in R$, клиентов $u \in U$, тем $t \in T$

$$H_{tr}(u) := \frac{p_{tu} q(r|t)}{\sum_{t' \in T} p_{t'u} q(r|t')};$$

- **M-шаг** (максимизация правдоподобия):
  для всех клиентов $u \in U$, тем $t \in T$

$$p_{tu} := \frac{1}{|D_u|} \sum_{r \in D_u} H_{tr}(u), \quad \text{где} \quad D_u = \{r : (u,r) \in D\};$$

Симметризованный EM-алгоритм

Инициализировать профили q_{tr} и p_{tu};
Повторять итерации, пока все профили не сойдутся:

1. Фиксировать q_{tr};
   Вычислить $q(r|t)$ по формуле Байеса;
   Повторять, пока профили клиентов не сойдутся:
   - E-шаг: вычислить скрытые переменные $H_{tr}(u)$;
   - M-шаг: вычислить профили клиентов p_{tu};
2. Фиксировать p_{tu};
   Вычислить $p(u|t)$ по формуле Байеса;
   Повторять, пока профили объектов не сойдутся:
   - E-шаг: вычислить скрытые переменные $H_{tu}(r)$;
   - M-шаг: вычислить профили объектов q_{tr};

Эксперименты

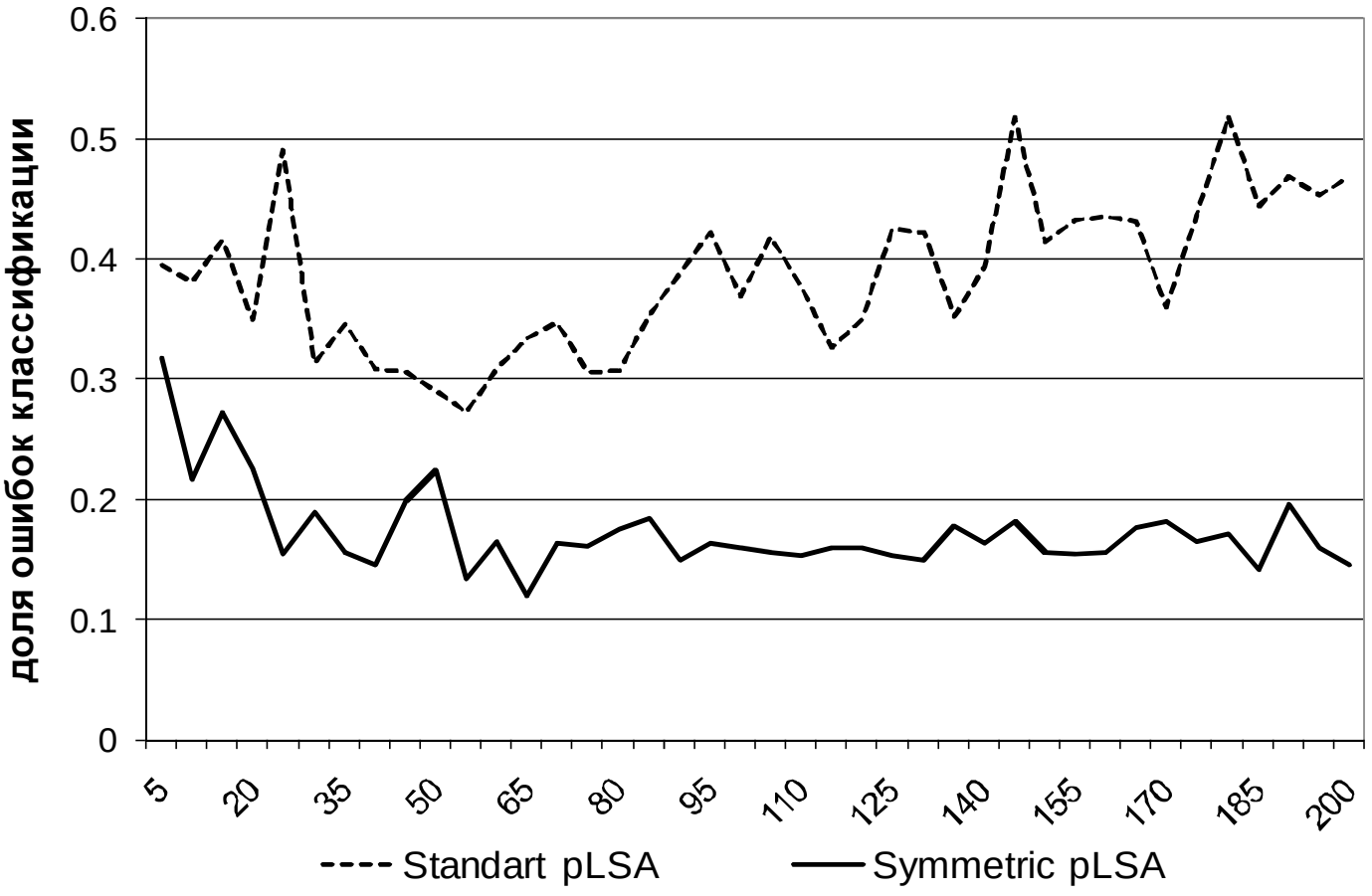
Данные поисковой машины Яндекс

7 дней работы поисковой машины Яндекс; объём лога 3.7 Гб;
14 606 пользователей;
207 696 запросов;
1 972 636 документов было выдано;
129 600 документов были выбраны пользователями.

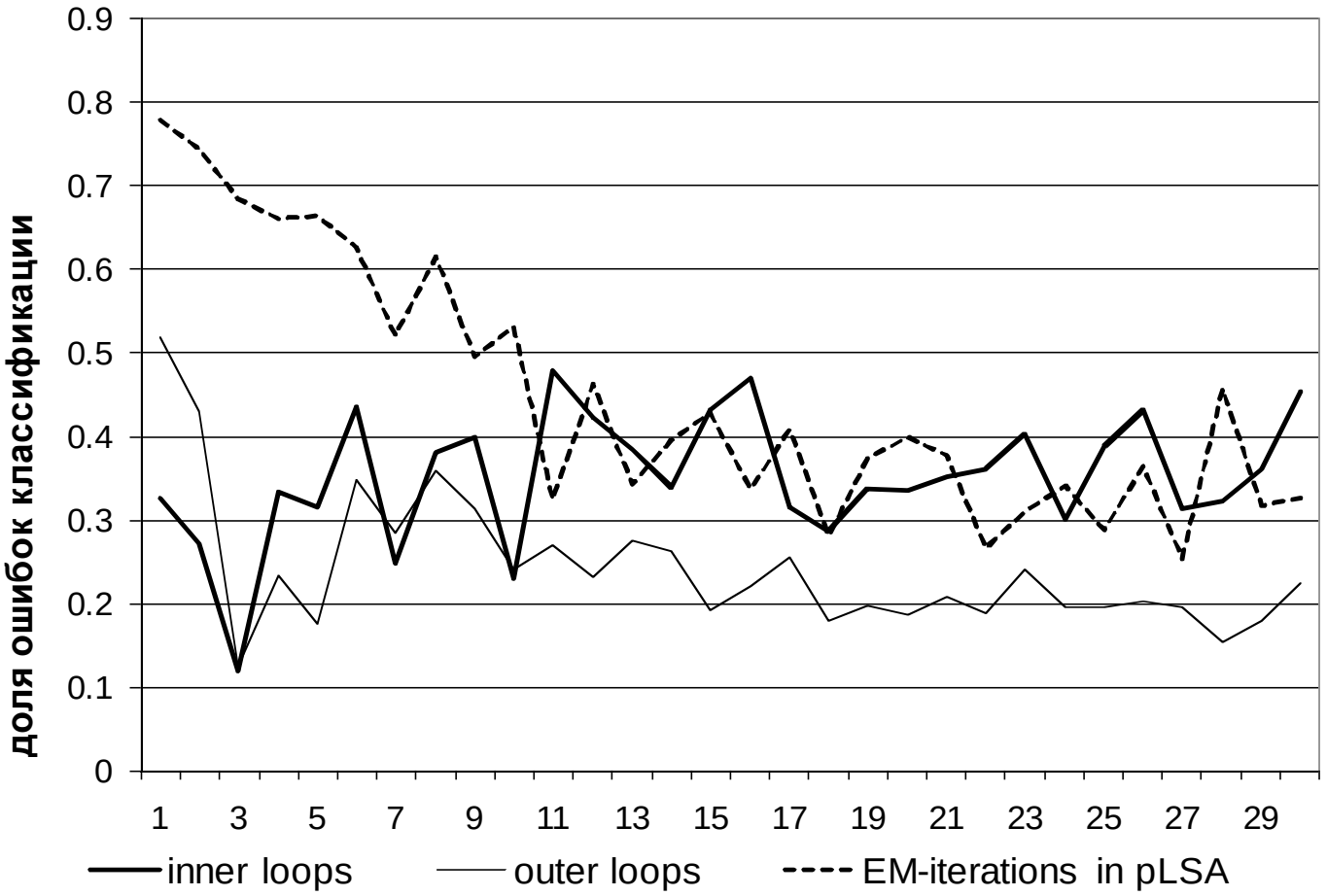
Схема эксперимента:

- Отбор наиболее посещаемых сайтов, $|R| = 1024$.
- Отбор наиболее активных пользователей, $|U| = 1902$.
- Введение критериев качества профилей:
  - 400 сайтов заранее классифицированы на $|T| = 12$ тематических классов;
- Оптимизация параметров по критерию качества.
- Построение профилей и оценок сходства сайтов.
- Визуализация: карты сходства.

Результаты: оптимизация количества тем



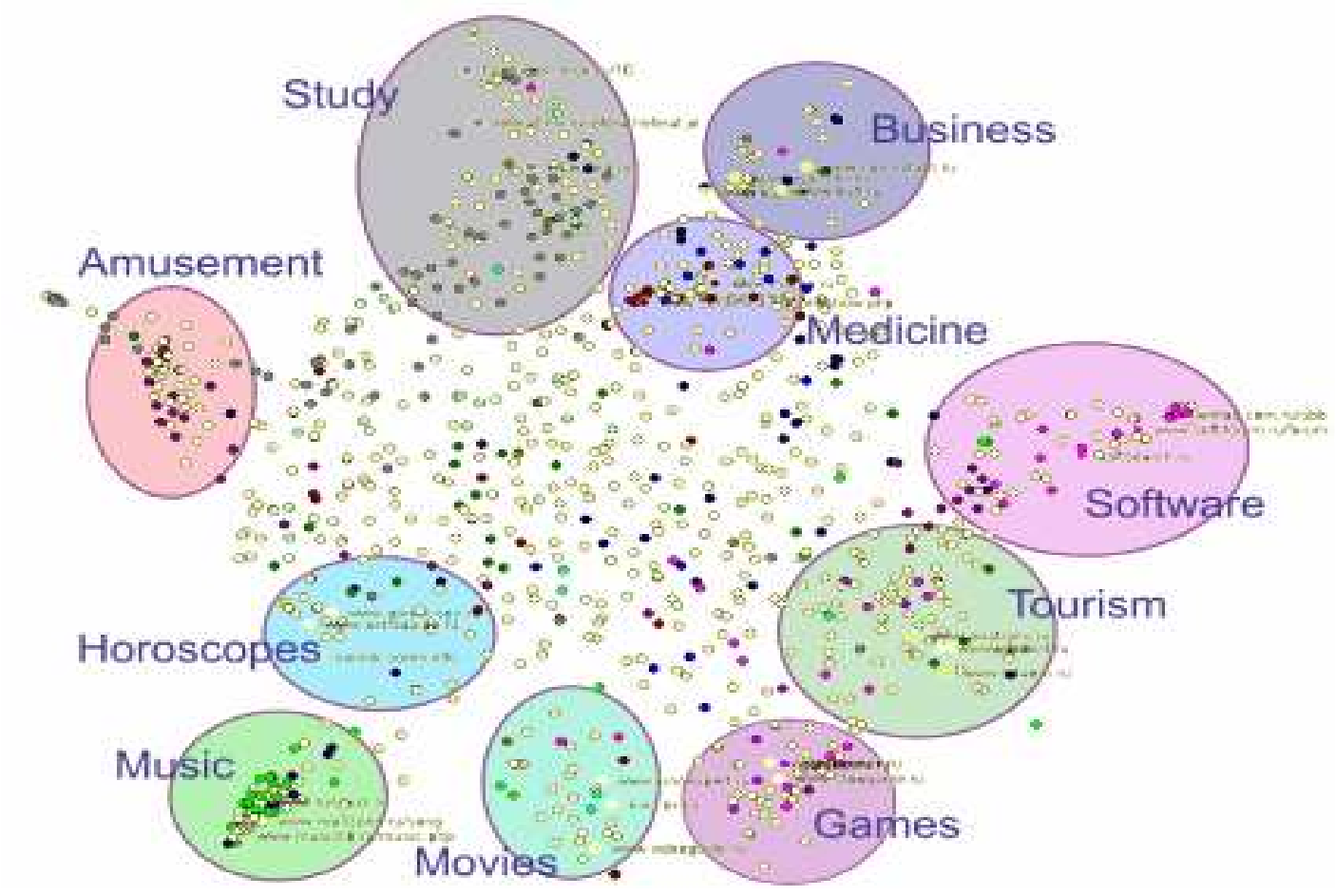
Оптимизация количества итераций



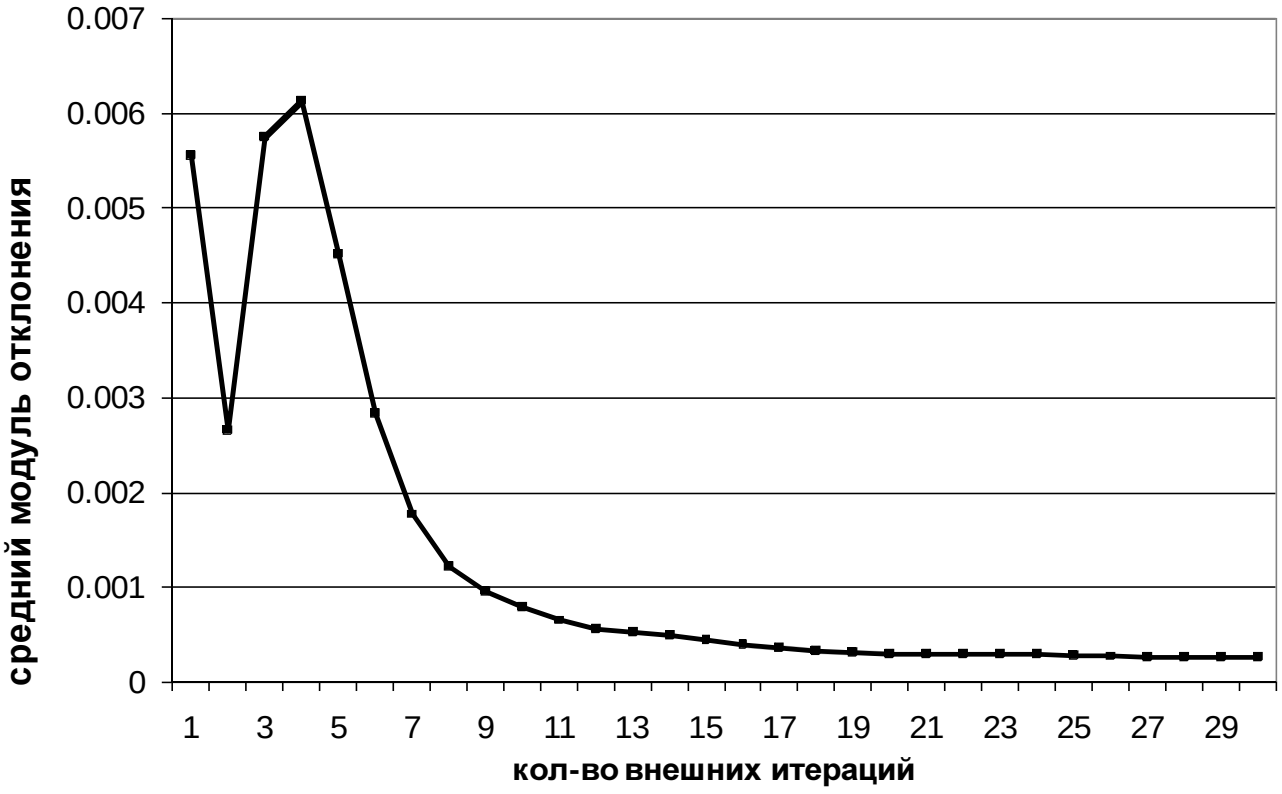
Примеры восстановления профилей сайтов

	Компоненты профиля											
Сайт	1	2	3	4	5	6	7	8	9	10	11	12
Музыка												
www.mp3real.ru	0	0.01	**0.86**	0	0.02	0.04	0.01	0	0.03	0	0.01	0.01
mp3.musicfind.ru	0	0	**0.96**	0	0	0	0	0	0	0.02	0	0.01
akkordi.ru	0	0.01	**0.85**	0.02	0.03	0.02	0.01	0	0.01	0.02	0.01	0.03
www.muzzone.com	0.01	0	**0.94**	0	0	0	0.02	0	0	0.01	0	0.02
mp3forum.ru	0.01	0.01	**0.85**	0.02	0	0.01	0.04	0.01	0.01	0.03	0	0.01
Сотовая связь												
mindmix.ru/mobile	0.01	**0.83**	0.02	0	0.01	0.01	0.04	0	0.01	0.05	0	0
www.sotoman.ru	0.01	**0.78**	0.01	0.02	0.04	0.01	0.04	0.02	0.01	0.03	0.01	0.02
www.mobyline.ru	0.02	**0.74**	0.02	0.01	0.02	0.01	0.03	0.03	0.07	0.02	0.02	0.01
www.eurotel.ru	0.01	**0.87**	0.04	0	0.01	0.01	0.01	0	0	0.01	0.02	0.03
www.sota1.ru	0.01	**0.91**	0.01	0.01	0.01	0	0.02	0	0	0.01	0.01	0
Рефераты, учебные ресурсы												
www.zachetka.ru	0	0	0	0.01	**0.16**	**0.56**	0	0	0.02	0.01	0.21	0
edu.mton.ru	0	0	0	0.01	**0.45**	**0.41**	0	0	0.01	0	0.1	0
forstudent.msk.ru	0	0	0.01	0.01	**0.39**	**0.44**	0.01	0.01	0.02	0	0.1	0
www.5ka.ru	0.01	0.01	0	0.02	**0.11**	**0.65**	0.01	0.01	0.02	0.01	0.14	0.01
school.edu.ru	0.01	0.06	0.01	0.05	**0.53**	**0.17**	0.01	0.02	0.03	0.01	0.1	0.01
Игры												
gameguru.ru	0.01	0.01	0	0.01	0.02	0.03	**0.77**	0.01	0.02	0.09	0.01	0.02
www.gameland.ru	0.08	0.01	0.02	0.02	0	0	**0.73**	0.05	0.02	0.05	0.01	0
www.ag.ru	0	0.02	0.04	0.01	0.01	0.02	**0.84**	0.01	0	0.01	0.01	0.04
www.neogame.ru	0.02	0.01	0	0	0.04	0.01	**0.81**	0.04	0.01	0.04	0.01	0.02

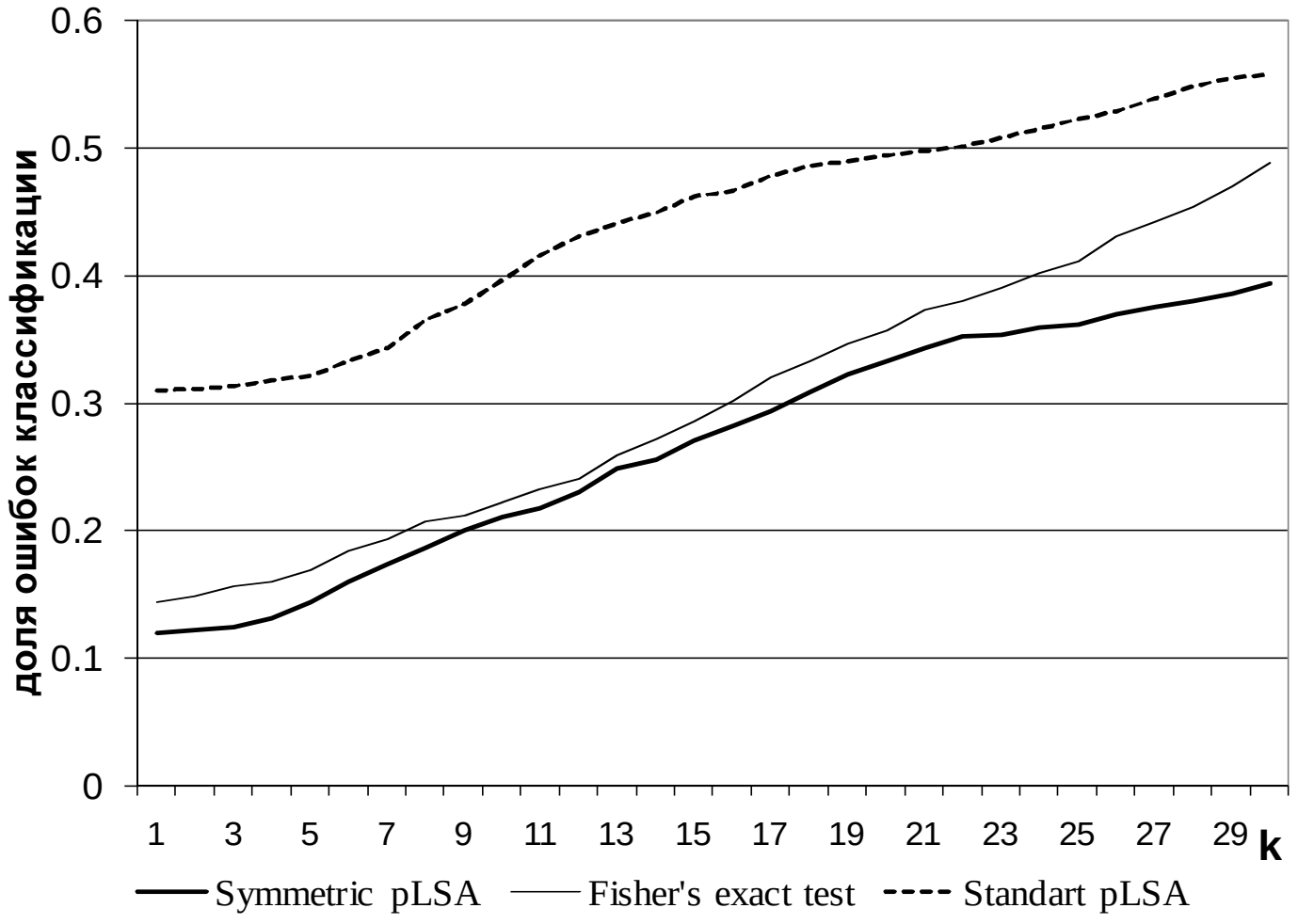
Карта сходства



Средний модуль отклонения вероятностей $H_{tr}(u)$ и $H_{tu}(r)$



Сравнение различных метрик по kNN



Данные мебельной компании

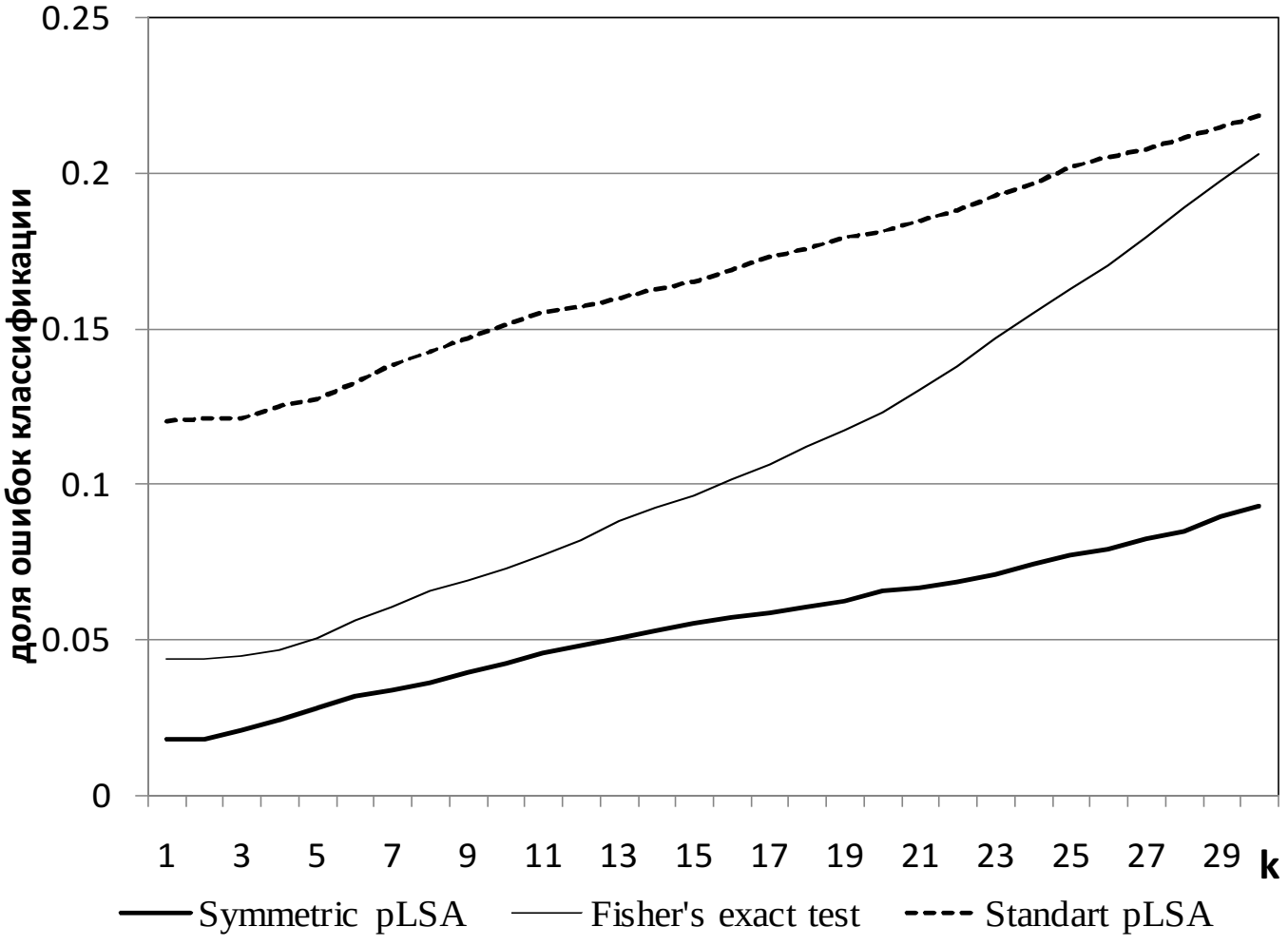
Исходные данные:

- история продаж за 3 года работы компании;
- 1 920 товаров;
- 1 328 постоянных клиентов;
- выборка из 112 256 фактов покупки товаров;
- для оценки качества по товарам использовалось разбиение 403 товаров на 12 категорий;
- для оценки доли правильно классифицированных товаров использовался метод k ближайших соседей при $k = 5$;

Результаты:

- оптимальное значение функционала — 3% ошибок классификации
- оптимальные параметры: $|T|$=30, 4 внутренних и 4 внешних итераций

Сравнение различных метрик по kNN



Анализ сообщений форума

Исходные данные:

- 564 сообщения;
- 3 841 ключевых слов;
- анализировалась встречаемость ключевых слов в сообщениях.

Результаты:

- построены профили длины $|T|=10$ для каждого сообщения;
- построена функция сходства сообщений на основе профилей.

Направления дальнейших исследований

- Если $f_{ur} \in Z = \{1, 2, \ldots, z_{\max}\}$ — рейтинги,
  то вместо $p(u, r) = \mathsf{P}(f_{ur} \neq \varnothing)$ надо оценивать
  $(z_{\max} - 1)$ вероятностей $p_z(u, r) = \mathsf{P}(f_{ur} \leqslant z)$, $z \in Z$;
- Динамическое обновление профилей при пополнении D;
- Иерархические профили;
- Учёт априорной информации через
  начальное приближение профилей:
  - тематический каталог объектов;
  - соц-дем (анкеты) клиентов;
- Связь с матричной факторизацией;
- Унифицированный профиль объектов и клиентов;
- Долгосрочный и краткосрочный профили;
- Оценивание сходства по частям профиля.