Методы нейросетевого распределенного представления и поиска сходных символьных последовательностей в задачах классификации на основе рассуждений по примерам

05.13.23 – системы и средства искусственного интеллекта
Диссертация на соискание степени кандидата технических наук

Соколов Артем Михайлович
Отдел нейросетевых технологий обработки информации МНУЦИТиС

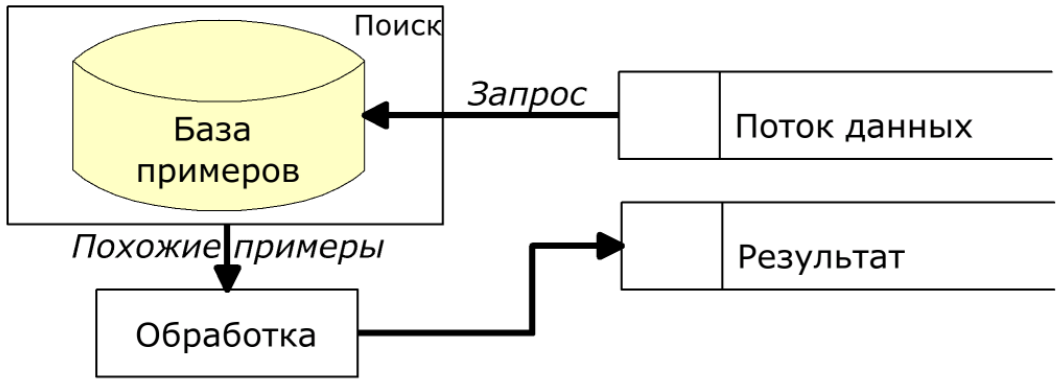
Научный руководитель: д.т.н. Рачковский Дмитрий Андреевич

Обработка последовательностей и рассуждения на основе примеров

Задачи

- поиск дубликатов
- фильтрация спама
- поиск генов
- обнаружение вторжений

Рассуждения на основе примеров



Этап поиска похожих примеров – центральный в данном подходе

Для $x, y \in \Sigma^n$

Расстояние редактирования $ed(x, y)$

Минимальное число операций **замены, удаления и вставки** символов, необходимое для преобразования x в y.

Вычисление $ed(x, y)$

- Динамическое программирование – $O(n^2)$
- Лучший результат – $\frac{O(n^2)}{\log n}$
- Слишком ресурсоемко для больших n

Распределенные представления и вложения пространств

Распределенное представление (РП) [Амосов, Куссуль, Рачковский]

Форма векторного представления, где любой объект представлен совокупностью элементов псевдослучайного (бинарного) вектора

$$x \to v(x) = (0, \boxed{1}, 0, 0, 0, \boxed{1}, 0, 0, 0, 0, \boxed{1}, \boxed{1}, 0, \boxed{1}, 0, 0, 0, \boxed{1}, \ldots, 0, \boxed{1})$$

Подход к вычислению «сложных» метрик

Идея – вложить в «простое» пространство: $(X, \rho_1) \xrightarrow{v} (Y, \rho_2)$

- векторное (желательно малой размерности)
- с простой метрикой ρ_2 (ℓ_1, ℓ_2, ℓ_∞,$Hamming$)

Качество вложения

Указать такие $k_1 \le k_2, d_1 \le d_2$, что

$$\text{если } \rho_1(x,y) \le k_1, \text{ то } \rho_2(v(x), v(y)) \le d_1,$$
$$\text{если } \rho_1(x,y) > k_2, \text{ то } \rho_2(v(x), v(y)) > d_2.$$

Чем меньше $k_2 - k_1$, тем точнее аппроксимация

Научная задача, цель, задачи работы

Научная задача

Разработка методов распределенного представления последовательностей, их поиска и классификации, для эффективной оценки сходства последовательностей в системах ИИ, применяющих модели рассуждений на основе примеров.

Цель работы

Повышение эффективности оценки сходства информации с последовательной структурой для решения прикладных задач классификации и поиска.

Задачи работы

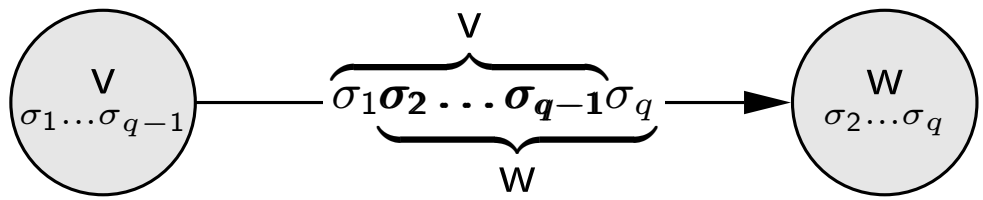
- Разработать методы векторного представления последовательностей, учитывающие порядок элементов
- Разработать методы рандомизированного представления последовательностей, которые могут быть представлены как РП и теоретически их проанализировать.
- Разработать и исследовать методы поиска сходных последовательностей с помощью РП.
- Развить методы классификации последовательностей в прикладных задачах на основе моделирования рассуждений по примерам за счет повышения эффективности путем использования РП.
- Разработать программные средства, реализующие предложенные методы представления, поиска и классификации приближенно ближайших последовательностей.
- Экспериментально исследовать эффективность и качество разработанных методов в задачах поиска и классификации информации с последовательной структурой на основе рассуждений по примерам.

Метод вложения расстояния редактирования в векторное пространство

- q_1, q_2 — мин. и макс. длины q-грам
- w — ширина скользящего окна
- $\Sigma^n \to (\mathbb{N} \cup 0)^{|\Sigma^n|(n-w+1)(q_2-q_1+1)}$ — вложение
  $x \mapsto (v_{q_1}(x[1,w]), v_{q_1+1}(x[1,w]), \ldots, v_{q_2}(x[1,w]), v_{q_1}(x[2,w+1]), \ldots,$
  $v_{q_2}(x[n-w+1,n]))$
- $D(x,y) = \frac{\sum_{i=1}^{n-w+1} \sum_{q_1}^{q_2} d_q(x[i,i+w-1],y[i,i+w-1])}{(n-w+1)(\Delta q+1)}$ — метрика

Графы де Брейна

- $B(\Sigma; q) = G(V, E)$
  $V = \Sigma^{q-1}$
  $E = \Sigma^q$

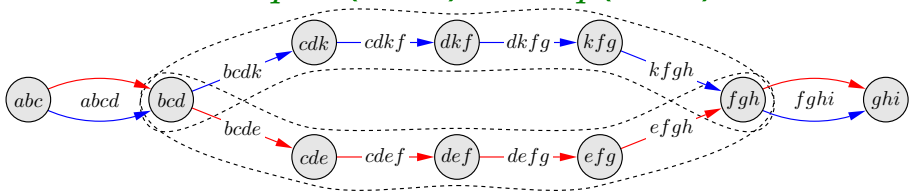


- $(v_q(x))_j = \sum_{i=1}^{n-q+1} [[x[i, i+q-1] = \sigma_j]], \sigma_j \in E$ — q-грамный вектор
- $d_q(x,y) = \sum_j |(v_q(x))_j - (v_q(y))_j|$ — q-грамное расстояние

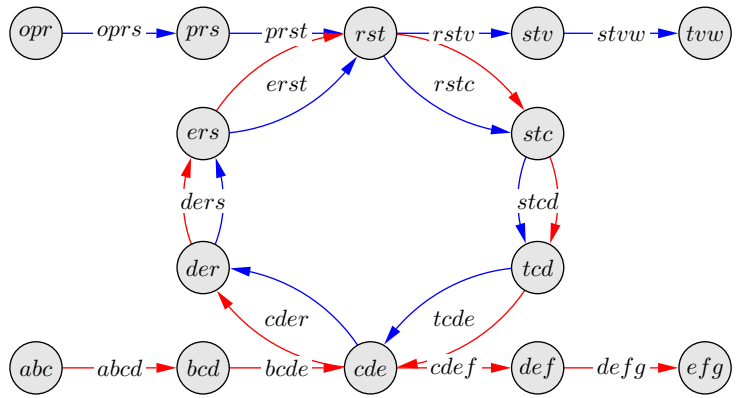
Теорема о точности вложения расстояния редактирования в векторное пространство

При увеличении q количество различных дуг ведет себя по-разному в зависимости от конфигурации

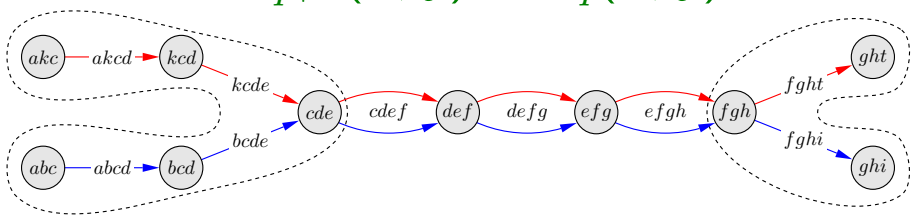
Петля – $d_{q+1}(x,y) = d_q(x,y) + 2$



Цикл



Вилка – $d_{q+1}(x,y) = d_q(x,y)$



Теорема: При $w \geq 6$, $k_1 \geq 1$, $q_1 = 2w/3$, $n > w(k_1+1)+1$,
$\Delta q = \frac{1}{2}(-7+\sqrt{57+16(w-q_1)})$, $Q = (\Delta q+1)(\Delta q+2)$, $t = w - \Delta q + 1$

$$\text{если } ed(x,y) \leq k_1, \text{ то } D(x,y) \leq \frac{2k_1[w^2+(n+1)]}{n-w+1}$$

$$\text{если } ed(x,y) > k_2, \text{ то } D(x,y) \geq \frac{Qt(\frac{k_2}{2(\Delta q+1)}-2)}{(n-w+1)(\Delta q+1)}.$$

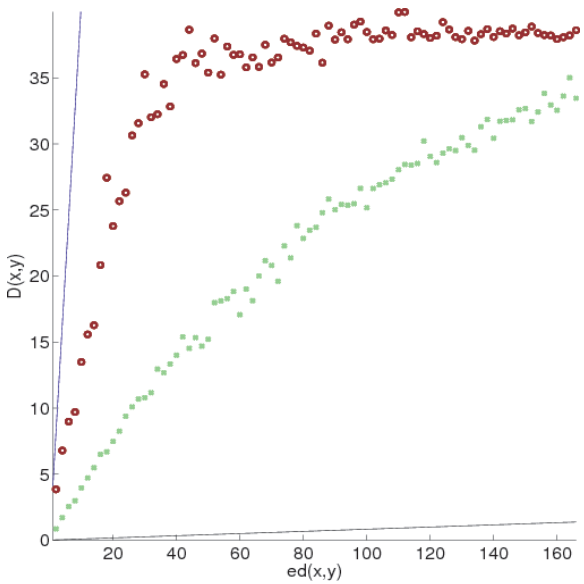
При $w = n^{1/2}$ время построения – $O(n^{3/2})$, кол-во ненулевых элементов – $O(n^{5/4})$

Теоретическое и экспериментальное сравнение

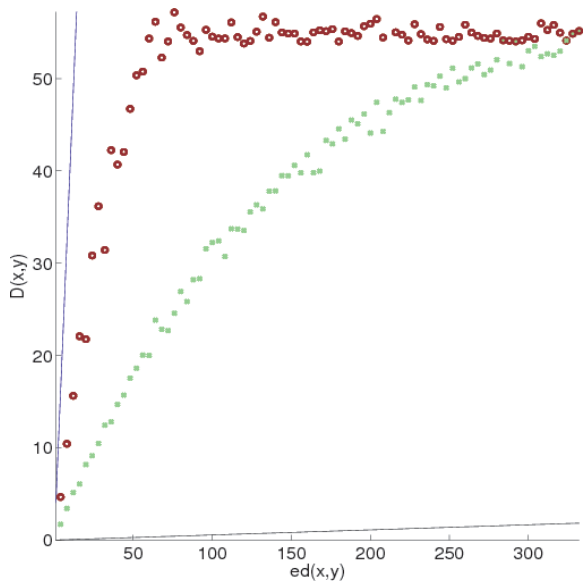
n – длина последовательности k – параметр точности

ссылка	пространства	k_1	k_2	размер
[Andoni et al, 03]	$ed \to \ell_1$	искажение $> \frac{3}{2}$		не констр.
[Batu et al, 03]	$ed \to ed$	$O(n^\alpha)$	$\Omega(n)$	$\tilde{O}(n^{\max(\frac{\alpha}{2}, 2\alpha-1)})$
[Bar-Yossef et al, 04]	$ed \to$ Hamm.	k	$(kn)^{2/3}$	$O(1)$
[Ostrovsky et al, 05]	$ed \to \ell_1$	k	$k2^{O(\sqrt{\log n \log\log n})}$	$O(n^2)$
дисс. работа	$ed \to \ell_1$	k	$k\sqrt{n}$	$O(n^{5/4})$

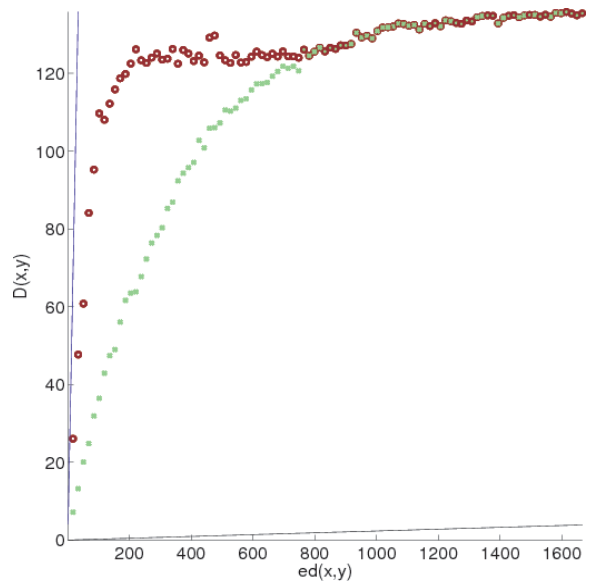
Численный эксперимент



$n = 5000$



$n = 10000$



$n = 50000$

Метод рандомизированного вложения и поиск близких строк

Приближенные ближайшие соседи

- приближенные соседи часто достаточны для приложений
- часто данные известны с некоторой точностью
- «проклятие размерности» для поиска точных соседей

Задача поиска (k_1, k_2)-ближайшего соседа (k_1, k_2)-БС

Дано:

- $P \subset \Sigma^n$ – база строк
- $k_1 < k_2$ – параметры
- $z \in \Sigma^n$ – запрос к базе

Задача:
Если $\exists x \in P$, такое что $ed(x, z) \leq k_1$, то возвратить любую $y \in P$, такую что $ed(y, z) \leq k_2$

Разработанные рандомизированные представления позволили

- уменьшить ресурсоемкость
- получить эффективный метод поиска ближайших соседей
- получить нейросетевую распределенную схему представления последовательностей

Локально-чувствительная функция для ed

Определение [Indyk, Motwani, 98]

Семейство $H = \{h : (X, \rho) \to Y\}$ – локально-чувствительное для метрики ρ, если для $\forall x, y \in X$ и любой i.i.d. $h \in H$ выполняется:

если $\rho(x, y) \le k_1$, то $Prob[h(x) = h(y)] > p_1$,

если $\rho(x, y) > k_2$, то $Prob[h(x) = h(y)] < p_2$,

$$k_1 < k_2, p_1 > p_2$$

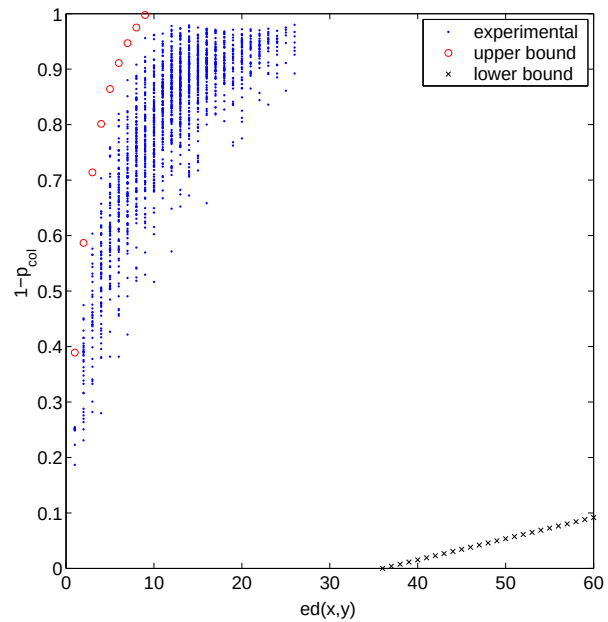
Построение локально-чувствительной функции для ed:

- i – i.i.d. равновероятная случайная величина из $[1, \ldots, n - w + 1]$
- v_{q_1,q_2} – конкатенация q-грамных векторов окна $x[i, i + w - 1]$ для $q = q_1, \ldots, q_2$
- ϕ – случайный вектор, распределенный по Коши $p(x) = \frac{1}{\pi(1+x^2)}$
- $b \in \mathbb{R}$ – случайная величина из $[0, r]$.

LSH-семейство функций

$$h(x) = \left\lfloor \frac{(v_{q_1,q_2}(x[i, i + w - 1]), \phi) + b}{r} \right\rfloor$$

Вероятность коллизии



$n = 1000$

Генерация распределенного представления

Формирование РП для строки $x = \sigma_1\sigma_2\ldots\sigma_n \in \Sigma^n$

- $\Sigma \ni \sigma \mapsto r(\sigma) \in \mathbb{R}^K$ «элемент σ», столбцы $\mathbf{R}$
- $1,\ldots,n \ni i \mapsto c(i) \in \{0,1\}^K$ «позиция i», строки $\mathbf{C}$
- $t(\sigma_i) := r(\sigma) \times c(i)$ поэлементно, «элемент σ в позиции i»
- $v(x) := \sum_{i=1}^n t(\sigma_i) = \sum_{i=1}^n r(\sigma) \times c(i)$ «последовательность x»

$$K \times K \left\{ \begin{pmatrix} \mathbf{1} & 0 & 3 & \ldots & 2 \\ 0 & \mathbf{-2} & 4 & \ldots & 0 \\ 2 & -1 & \mathbf{0} & \ldots & 1 \\ 1 & 0 & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 5 & 1 & 0 & \ldots & \mathbf{3} \end{pmatrix} \right. =$$

$$\underbrace{\begin{pmatrix} \mathbf{0.38} & -1.3 & 0.74 & 0.47 \\ \mathbf{-1.1} & -1.1 & -0.2 & 1.73 \\ \mathbf{-0.9} & 5.43 & 1.99 & -1.4 \\ \ldots & \ldots & \ldots & \ldots \\ \mathbf{-1.0} & 0.34 & 10.2 & 0.69 \end{pmatrix}}_{\substack{\mathbf{R} \\ K \times |\Sigma| \\ \text{столбцы} \sim \text{по Коши}}} \cdot \underbrace{\begin{pmatrix} 0 & 1 & 0 & 0 & \ldots & 0 \\ \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{1} & \ldots & \mathbf{0} \\ 0 & 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 1 & 0 & \ldots & 1 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & 0 & \ldots & 0 \end{pmatrix}}_{\substack{\mathbf{L} \\ |\bigcup_{q=q_1}^{q_2} |\Sigma|^q| \times n \\ \text{индикаторная матрица}}} \cdot \underbrace{\begin{pmatrix} 1 & 0 & \ldots & 0 \\ \mathbf{1} & \mathbf{0} & \ldots & \mathbf{0} \\ 1 & 1 & \ldots & 0 \\ 1 & 1 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 1 \end{pmatrix}}_{\substack{\mathbf{C} \\ n \times K \\ \text{строки} - \text{случ. окна шириной } w}}$$

Поиск (k_1, k_2)-ближайших соседей

Доказано, используя

- результат для детерминированного вложения
- свойства распределения Коши,

что $h(x)$ – локально-чувствительная для расстояния редактирования

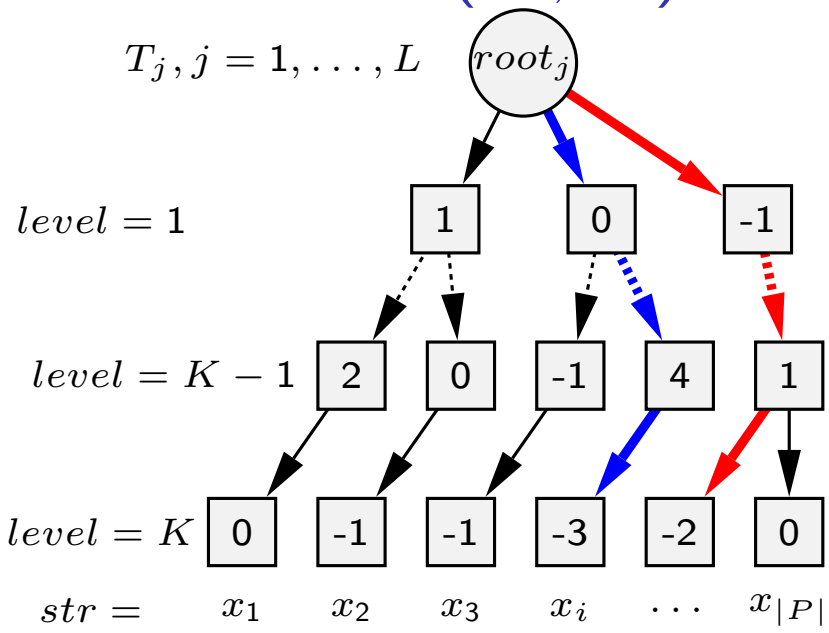
LSH-алгоритм

1. Для $\forall x \in P$ создать L векторов $h^j(x) = (h_{1j}(x), h_{2j}(x), .., h_{Kj}(x))$
2. В ячейках с «адресами» $h^j(x)$ запомнить x
3. Для запроса z отобрать до $2L$ строк из ячеек $h^j(z), j = 1, \ldots, L$
4. Если для найденной строки x_i, $ed(x_i, z) < k_2 \Rightarrow$ это сосед

Теорема

Если $K = \log_{1/p_2} |P|, L = |P|^{\frac{\ln p_1}{\ln p_2}}$, то LSH-алгоритм с функцией $h(x)$ с вероятностью $> 1/2$ находит $(k_1, O(zk_1 n^{1/3} \ln n))$-ближайшего соседа по расстоянию редактирования за время $O(|P|^{\frac{1}{1+z}}), z > 1$

Поиск (k_1, k_2)-БС с помощью LSH-леса

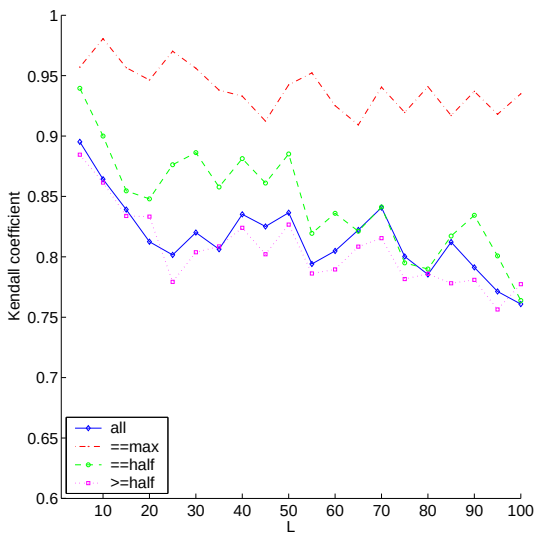


1. $x \in P$, $h^j(x)$, $j = 1, \ldots, L$
2. z, $h^j(z)$, $j = 1, \ldots, L$
3. $m = \max_j level(T_j, h^j(z))$
4. $S \leftarrow str(T_{j'}, h^{j'}(z))$, где $level(T_{j'}, h^{j'}(z)) = m$
5. while $|S| < 2L$: $m = m - 1$, goto 4

На выходе

S – упорядоченное по уровню множество кандидатов на приближенных ближайших соседей

Упорядоченность множества S



$|S| = 100$

Точность ($\frac{TP}{|S|}$) от L и $|S|$

L	$\|S\| = L$	σ_p	$\|S\| = 2L$	σ_p	$\|S\| = 3L$	σ_p	$\|S\| = 4L$	σ_p
1	0.950	0.048	0.945	0.029	0.927	0.025	0.893	0.031
2	0.895	0.056	0.853	0.054	0.825	0.032	0.810	0.028
3	0.885	0.050	0.816	0.035	0.784	0.030	0.770	0.025
4	0.842	0.041	0.810	0.031	0.777	0.023	0.757	0.021
5	0.824	0.039	0.786	0.021	0.759	0.014	0.735	0.014
6	0.797	0.040	0.782	0.025	0.760	0.019	0.730	0.014
8	0.791	0.024	0.755	0.016	0.729	0.013	0.724	0.010
10	0.787	0.024	0.749	0.015	0.722	0.009	0.689	0.008
20	0.762	0.014	0.708	0.006	0.682	0.005	0.658	0.004

Методы работы с базами строк разной длины

Выравнивание длины

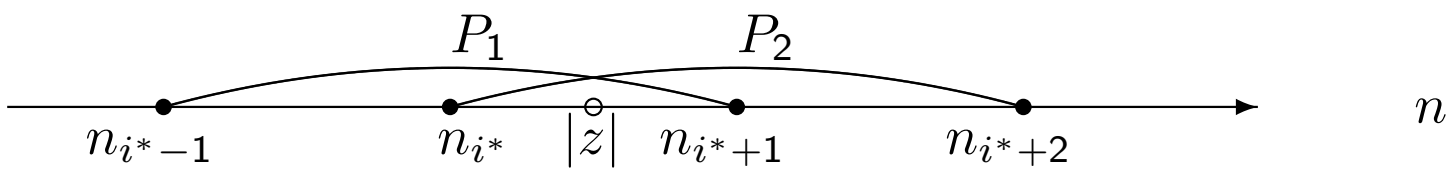
Дополнение спецсимволом \$ до длины n

Кластеризация базы по длине

- $0 < \varepsilon < 1$ — ширина кластера
- $n_0 = \min_{t \in P} |t|$ — нач. знач. центра
- $n_{i+1} = \lceil (1+\varepsilon) n_i \rceil$ — след. значения

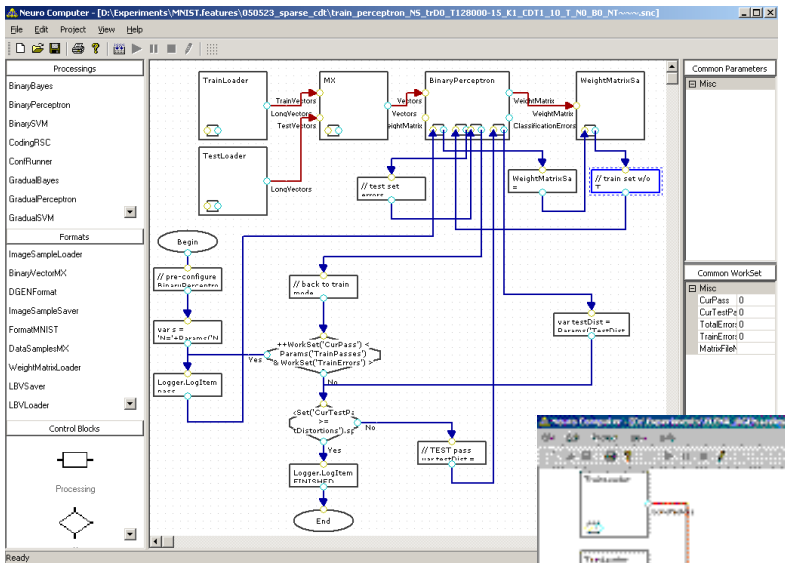
Выполнение запроса

- z — запрос
- $i^* = arg\max_i \{n_i \mid n_i \le |z|\}$ — ближайший снизу центр
- $P_1 = \{t \in P \mid n_{i^*-1} \le |t| \le n_{i^*+1}]\}, P_2 = \{t \in P \mid n_{i^*} \le |t| \le n_{i^*+2}]\}$

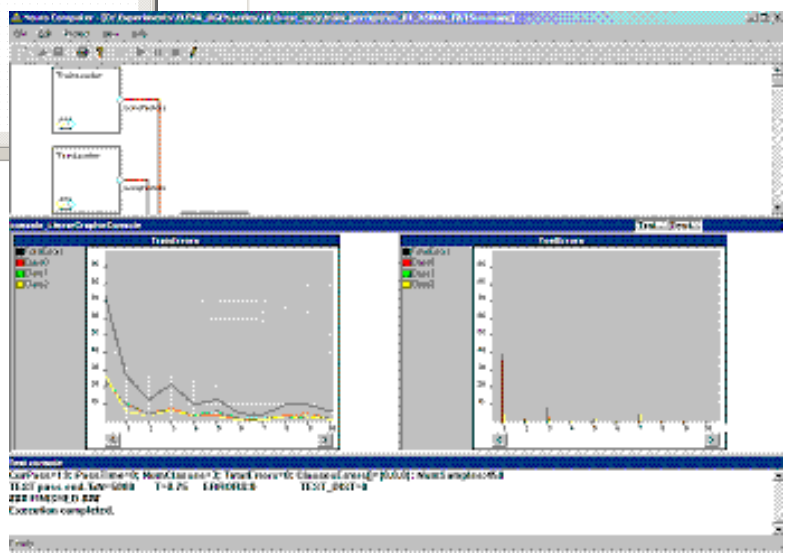


Программные реализации методов представления, поиска и классификации последовательностей

- LSHLibrary
- DuplClassifier
- EmailClassifier
- NuclClassifier
- SessionClassifier
- TextInputTools



Режим САПР



Режим выполнения

Поиск дубликатов среди веб-документов

База РОМИП/Яндекс

- 800 тысяч веб-страниц $\sim$ 0.3% Рунет
- 10 млн. пар дубликатов – эталон
- полнота $r = \dfrac{\text{кол-во найденных дубликатов}}{\text{кол-во действительных дубликатов}}$
- точность $p = \dfrac{\text{кол-во найденных дубликатов}}{\text{кол-во найденных документов}}$

Результат

ссылка	сходство документов	$F = \frac{2rp}{r+p}$
[Кузнецов, 05]		0.14-0.49
[Косинов, 07]	0.85	0.53
	0.90	0.58
	0.95	0.64
	1.00	0.63
дисс. работа $\varepsilon = 0.005$	0.85	0.66
	0.90	0.75
	0.95	0.82
	1.00	0.89

Оценка количества спама

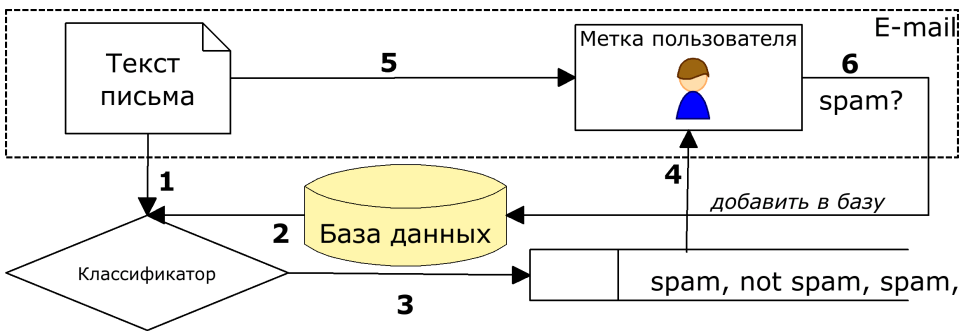
Спам

- 80-85% всей электронной почты
- массовость одинакового/схожего спама
- искажение слов для обхода систем обнаружения спама:
  `'mortage' 'buy viagra'`
  `'m0rt@ge' '6uy v1agraa'`

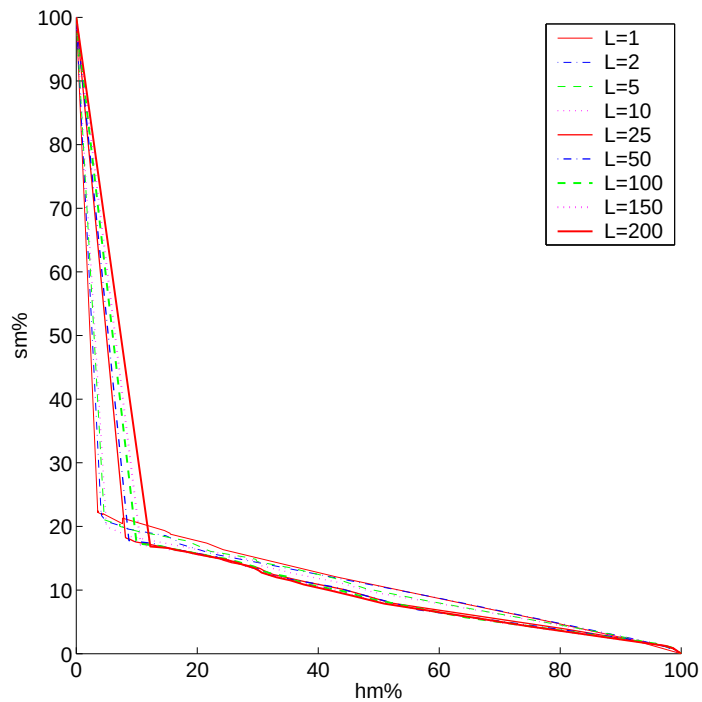
Коллекция Spam Track 2006

- $\sim$ 38000 писем (189Mb)
- из них 66% – спам
- sm% – false negative
- hm% – false positive

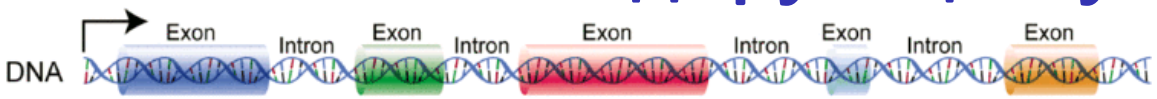
Классификация



Зависимость sm% от hm%



Поиск кодирующих участков ДНК



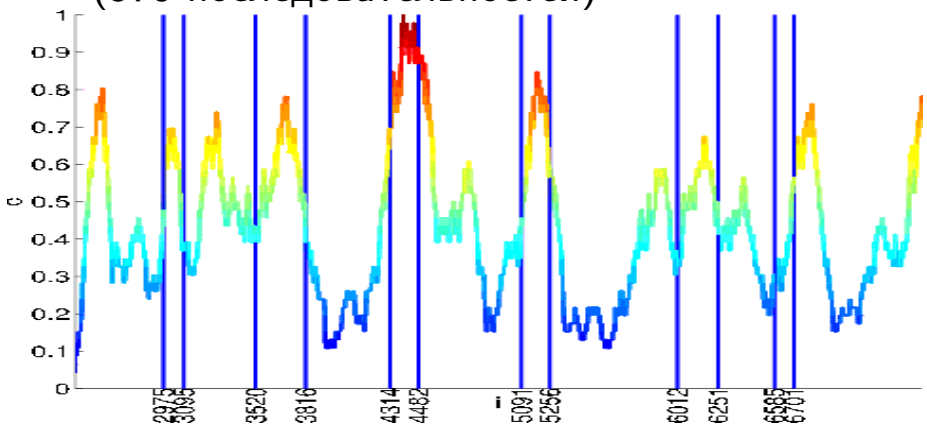
$\Sigma = \{A, T, G, C\}$ – основания

Объемы генетических данных

- 10^{10} Mb/год
- GenBank удваивается ежегодно
- длины до $3.2 \cdot 10^9$ символов

Наборы ДНК

- тренировочный база HMR195 (948 экзонов)
- тестовый набор BursetGuigo (570 последовательностей)



Методика поиска участков ДНК

- z запрос, экзон из трен. выборки
- t тестовая последовательность
- $c_j,\ j = 1, \ldots, |t|$ счетчики для $t[j]$
- $P = \{t[i, i + |z| - 1]\}$ база
- S множество кандидатов на ближайших соседей
- если $ed(t[i', i' + |z| - 1], z) = \min_{x \in S} ed(x, z)$, то $c_i = c_i + 1$, $i = i', \ldots, i' + |z| - 1$
- T – порог на значение c_i

если $c_i \geq T$, то $t[i]$ принадлежит экзону

Мера качества

$$AC = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TP}{TP + FP} + \frac{TN}{TN + FP} + \frac{TN}{TN + FN}\right) - 1$$

Результат

ссылка	AC	время на 1 PC
[Costello, 03]	0.49	$\sim$ 6 лет (класс. алг.)
дисс. работа	0.47	70 часов

Классификация сессий пользователей

Обнаружение вторжений

- $\Sigma = \{\text{'ls'}, \text{'mail'}, \text{'rm'}, \dots\}$, $|\Sigma| \sim 10^3$
- $\sim 10^3$ процессов/час
- аномалия – необычное поведение
- replay-атаки

База аудит-сессии ОС FreeBSD

- период – $\sim$ 3 года
- >500 пользователей
- $\sim$ 20 млн. комманд

Методика

- U множество пользователей
- u^* действительный пользователь
- t его тестовая сессия
- $c_u,\ u \in U$ счетчики

Результат

n	K	L		
		1	5	10
10	5	0.470	0.984	0.997
	7	0.431	0.945	0.987
20	5	0.440	0.942	0.983
	7	0.403	0.741	0.942
40	5	0.354	0.848	0.967
	7	0.230	0.390	0.836

- P_u базы для $\forall u \in U$ (все скользящие окна их сессий)
- $z = t[i, i+n-1]$ запрос, содержимое окна текущей сессии
- S_u множество ближайших окон, найденных в P_u
- если $S_u \neq \emptyset$, то $c_u = c_u + 1$, если $c_{u^*} = \max_u c_u$, то аномалии нет

Выводы

1. Разработанные методы распределенного представления обеспечивают сохранение сходства данных с последовательной структурой по расстоянию редактирования.
2. Анализ с помощью теории метрических вложений показал, что детерминированный вариант предложенного векторного представления последовательностей обеспечивают более высокую точность аппроксимации расстояния редактирования по сравнению с известными результатами.
3. Разработанные на основе локально-чувствительного хеширования рандомизированные методы поиска приближенных ближайших последовательностей имеют малую ресурсоемкость и сублинейное по размеру базы время поиска приближенно ближайших последовательностей.
4. Распределенное представление последовательностей, использующее рандомизацию векторных представлений и связывание элементов последовательности с их позициями, обеспечивает унификацию формата представления и возможность использования мер сходства векторных представлений для оценки сходства последовательностей.
5. Решение прикладных задач поиска текстовых дубликатов, спама, участков ДНК и обнаружения вторжений подтверждает перспективность разработанных методов.