Визуализация структуры текстовой коллекции

Александр Сигачёв
alexander.sigachov@gmail.com

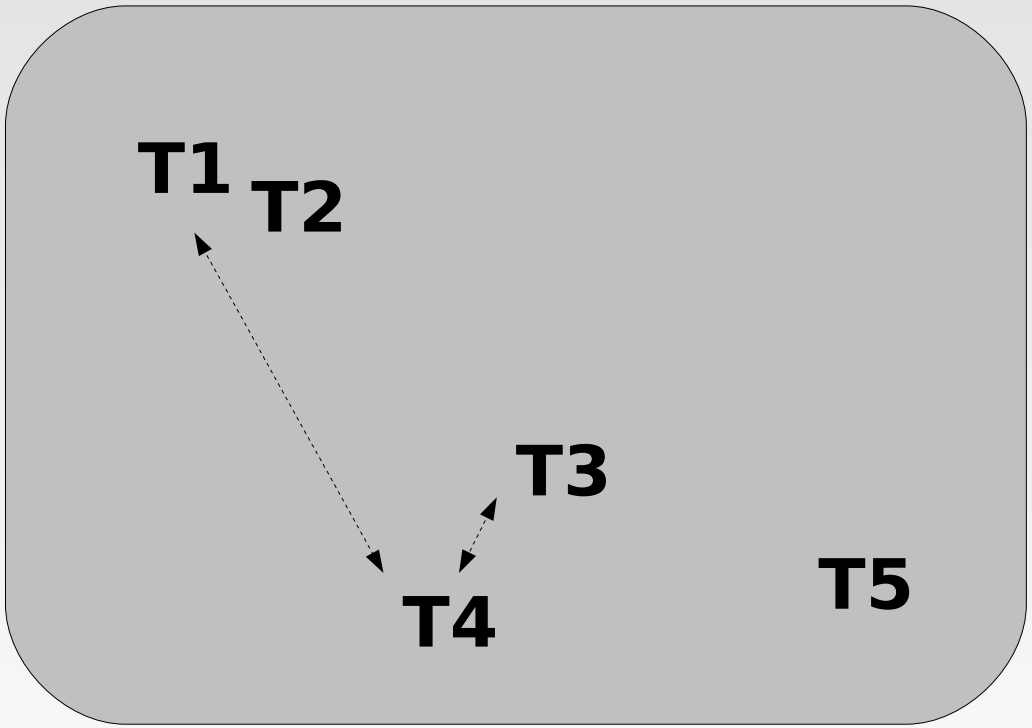
МГТУ им. Баумана
ИУ-5

Москва — 2007

Цель работы

Система анализа и визуализации тематических кластерных структур в коллекции текстов.

- Текст 1
- Текст 2
- Текст 3
- Текст 4
- Текст 5

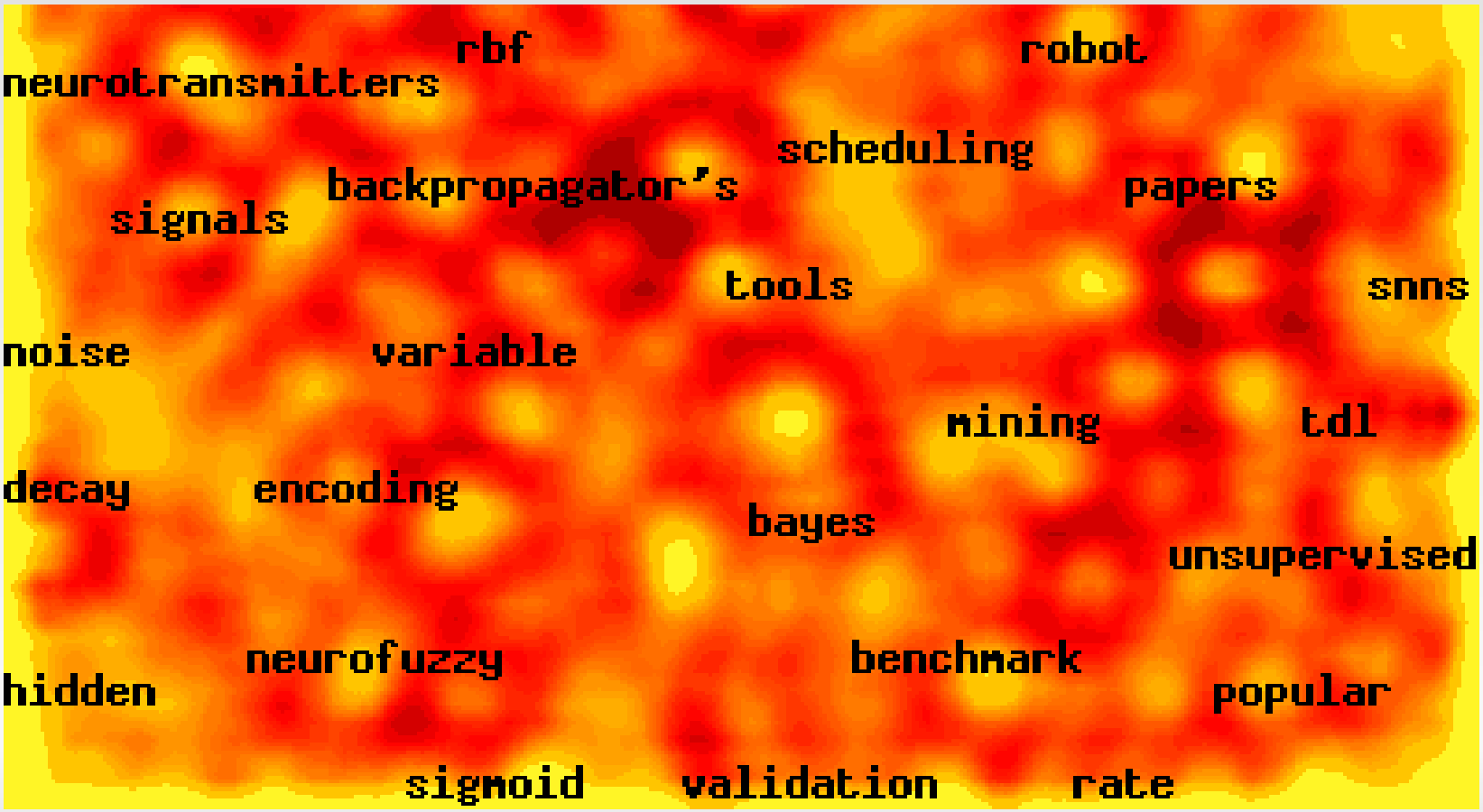


Применение

- визуальный тематический анализ коллекции документов в хранилище (Консультант+, Гарант);
- представление результатов поисковой системы (Google Scholar, Microsoft Academic Search);
- визуальная оценка нового документа (ручная категоризация);
- навигационная система сайта (карта сайта).

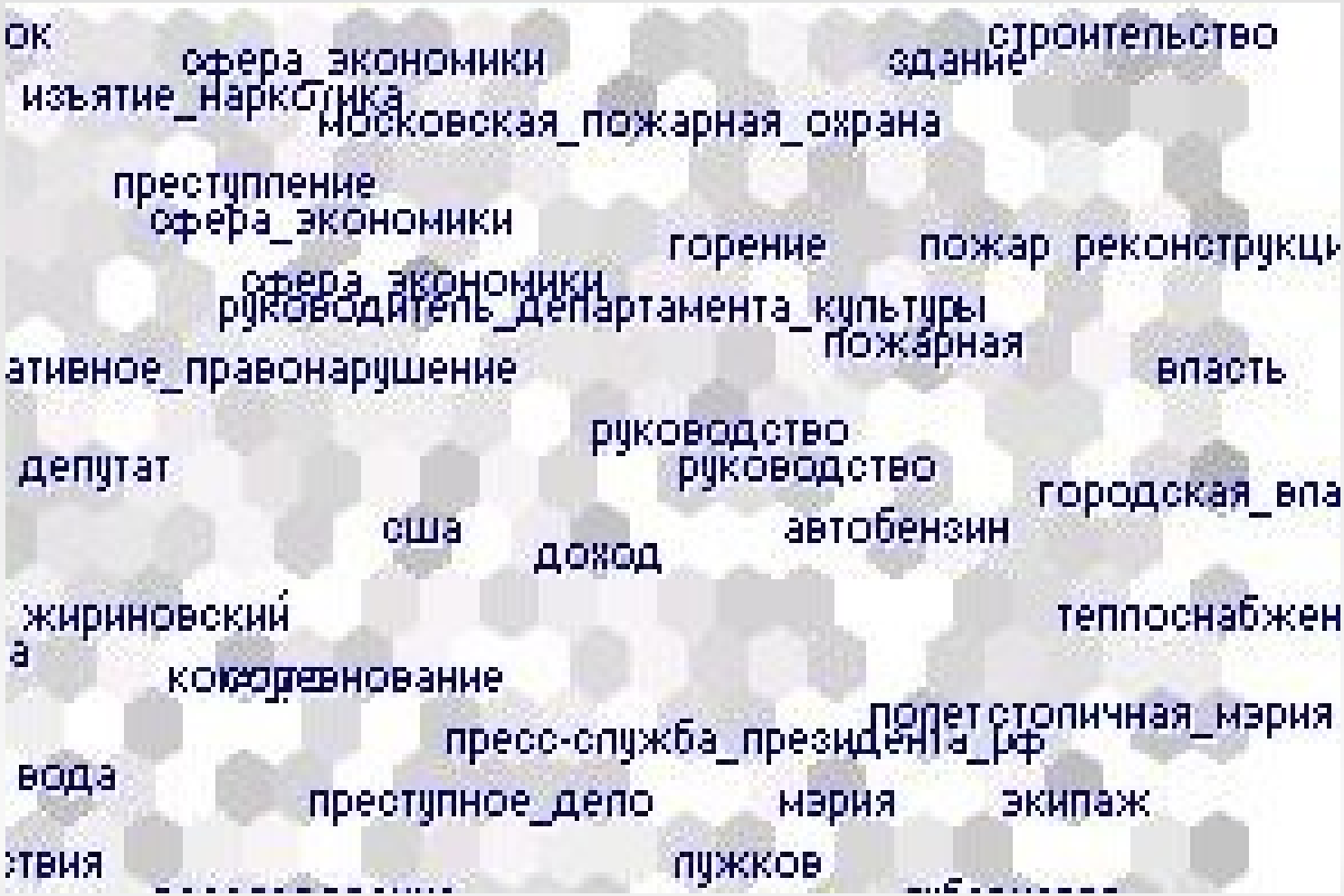
Прототипы: WEBSOM

Теуво Кохонен, визуализация UseNet,
более 1 мил. документов



Прототипы: TOPSOM

RCO, тысячи документов



Этапы обработки

- предобработка текста, лемматизация;
- расчёт частотных характеристик употребляемости слов в текстах;
- сокращение размерности пространства признаков;
- построение векторной модели текста;
- построение нейроной сети Кохонена;
- расчёт дополнительных параметров, визуализация сети;
- построение интерфейса.

Векторная модель текста

Текст как набор слов (модель терм-документ).

	Слово 1	Слово 2	Слово 3	Слово 4	...
Док. 1	0,54	0,02	--	0,11	
Док. 2	--	--	0,5	0,07	
Док. 3	0,67	--	--	0,22	

TFIDF

Функция определения веса слова в документе

TF-IDF.

$$TFIDF_{ij} = TF_{ij} \cdot \log\left(\frac{N}{DF_i}\right)$$

TF – частота слова в документе; DF – количество документов, где встречается термин; n – общее число документов.

- вес слова больше, если оно часто встречается в документе;
- меньше, если оно часто встречается в других документах.

Выбор фукции веса

Нормализация частоты

$$\frac{1+\log(TF)}{1+\log(MTF)}$$

Функция Яндекса для анализа близости

$$W_{single} = \log(p) * (TF_1 + 0.2 * TF_2)$$

$$TF_1 = \frac{TF}{TF + k_1 + k_2 * DocLength}, k_1 = 1, k_2 = 1/350$$

$$TF_2 = \frac{Hdr}{1 + Hdr}$$

$$p = 1 - \exp(-1.5 * \frac{CF}{D})$$

Сокращение размерности

- наиболее частотные слова;
- слова с наибольшим весом;
- слова со средней частотностью;
- слова, имеющие максимальную дисперсию веса;
- случайный выбор;
- предварительное разделение на две группы.

- решение о размерности вектора.

Мера сравнения векторов

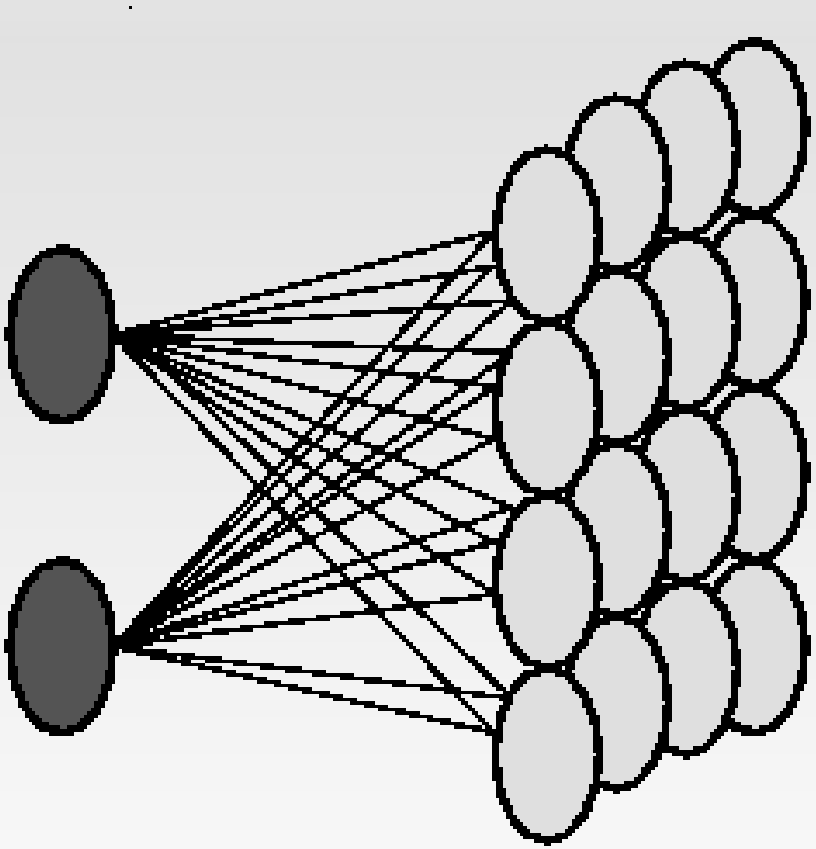
- евклидово расстояние;
- квадрат евклидова расстояния;
- косинусная мера;
- квадрат коэффициета корреляции;
- расстояние Чебышева;
- манхеттенское расстояние;
- процент несогласия;

Эксперимент

- Первый набор: 50 статей физической тематики из третьего издания БСЭ и 50 аналогичных статей из Википедии.
- Второй набор: 80 обзорных статей из издания Lenta.Ru по темам: политика, кино, экономика, оружие, Интернет, медицина.
- Строились и сравнивались вектора документов по различным методам.
- Целевая функция – отношение расстояния между аналогичными статьями к среднему расстоянию до других документов.

Победитель: выбор по дисперсии, 400 компонентов, логарифмическая функция

Карта Кохонена (SOM)



Элементы выходного слово конкурируют за соответствие входным данным.

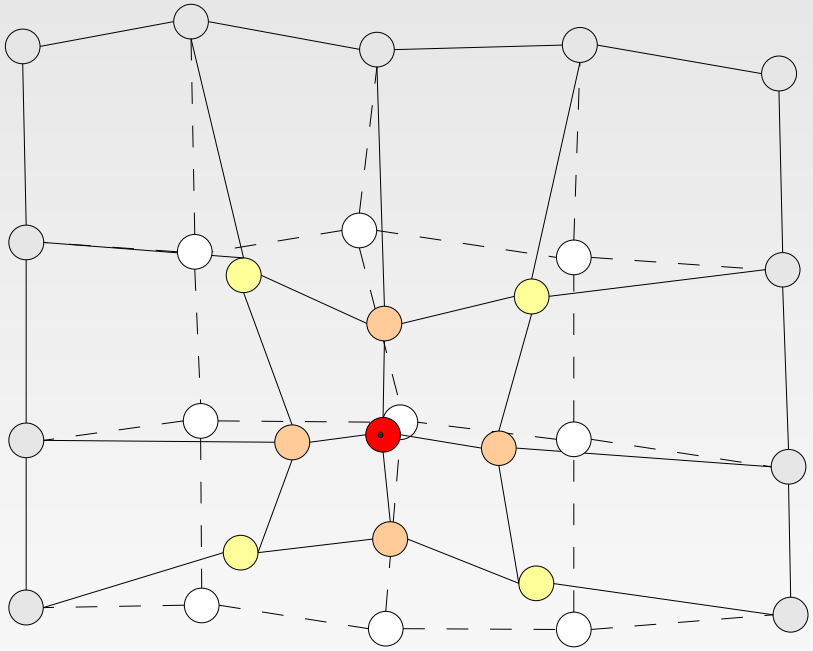
Определяется мера близости на пространстве выходного слоя.

Победитель получает больше (WTM).

Карта Кохонена: WTM

«Победитель получает больше» — WTM (winner takes most)

$$w_m := w_m - \eta(x_i - w_m)\Pi(\rho(x_i, w_m))$$



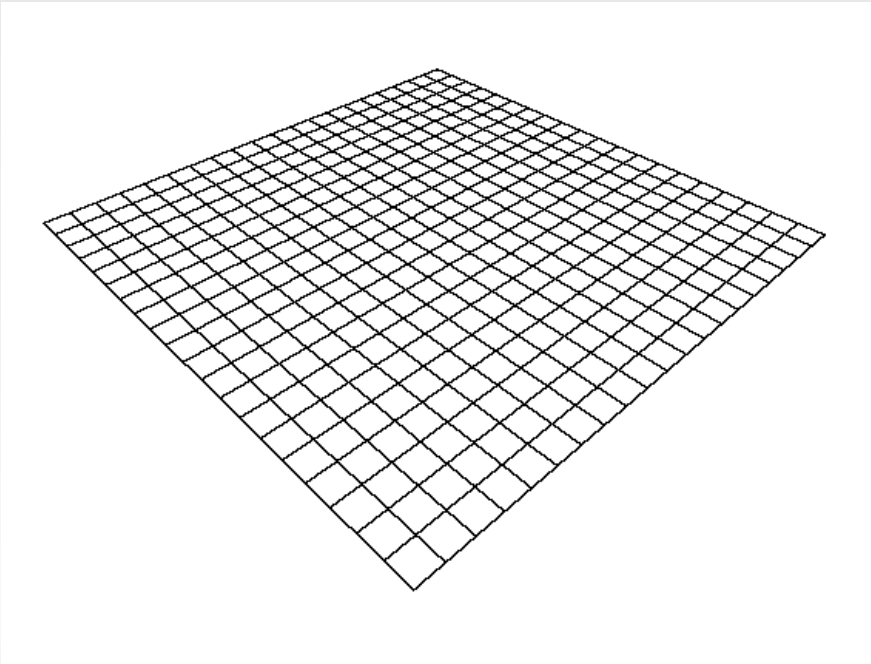
$$\Pi(\rho) = \exp(-\beta\rho^2), \beta > 0$$

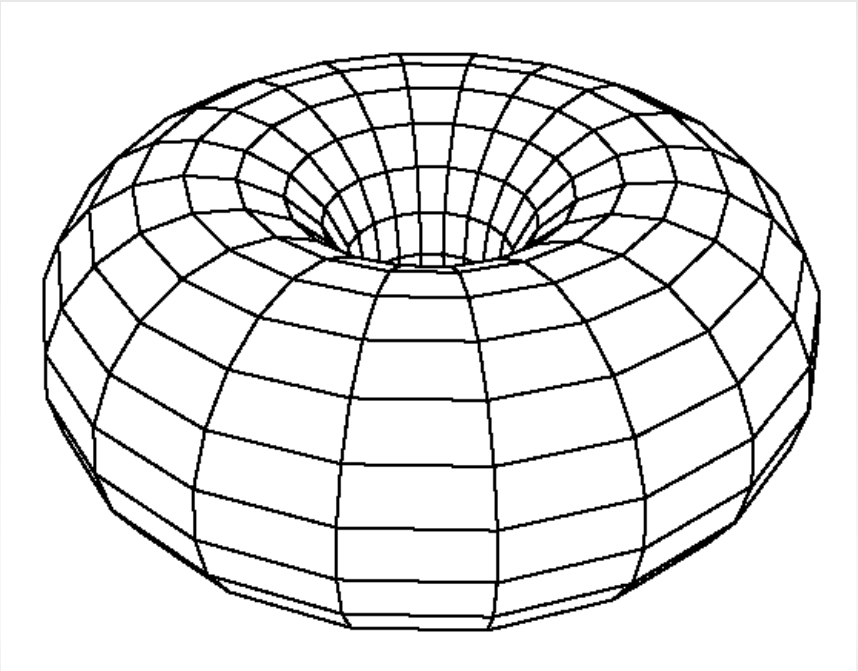
$$\rho(x, w) = \|x - w\|$$

$$a(x) = arg \min_{m \in Y} C_m \rho(x, w_m)$$

Топология SOM

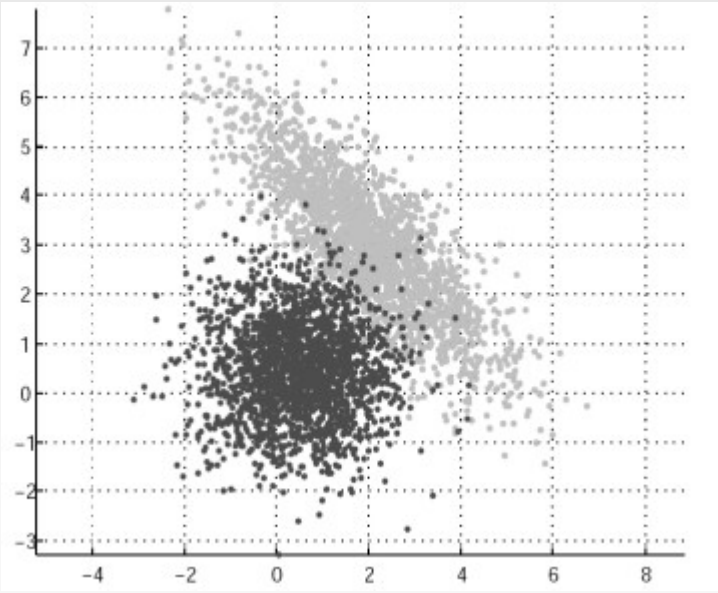
- большое количество нейронов выходного слоя (несколько тысяч);
- тороидальная топология выходного слоя.

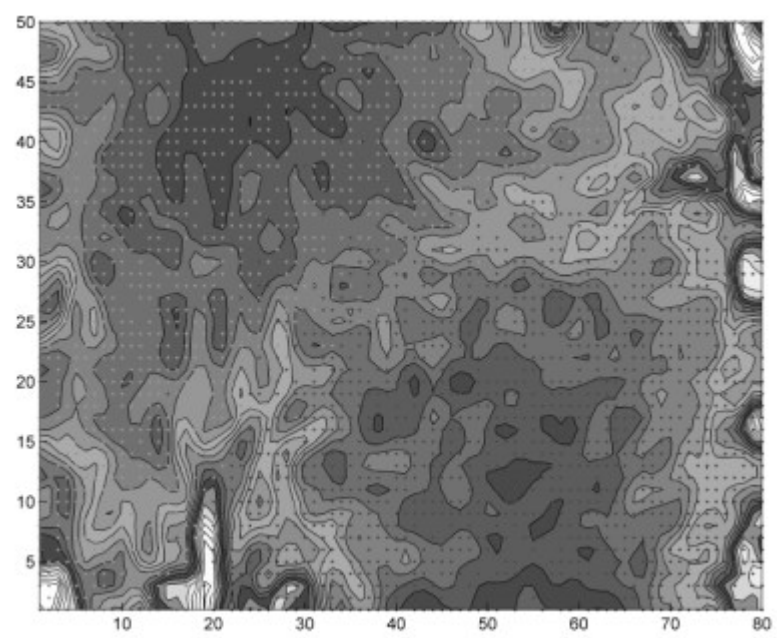




Визуализация

- Универсальная диаграмма расстояний.
  - Элементы матрицы определяют расстояние между весовыми коэффициентами нейрона и его ближайшими соседями. Большое значение — нейрон отличается от других.





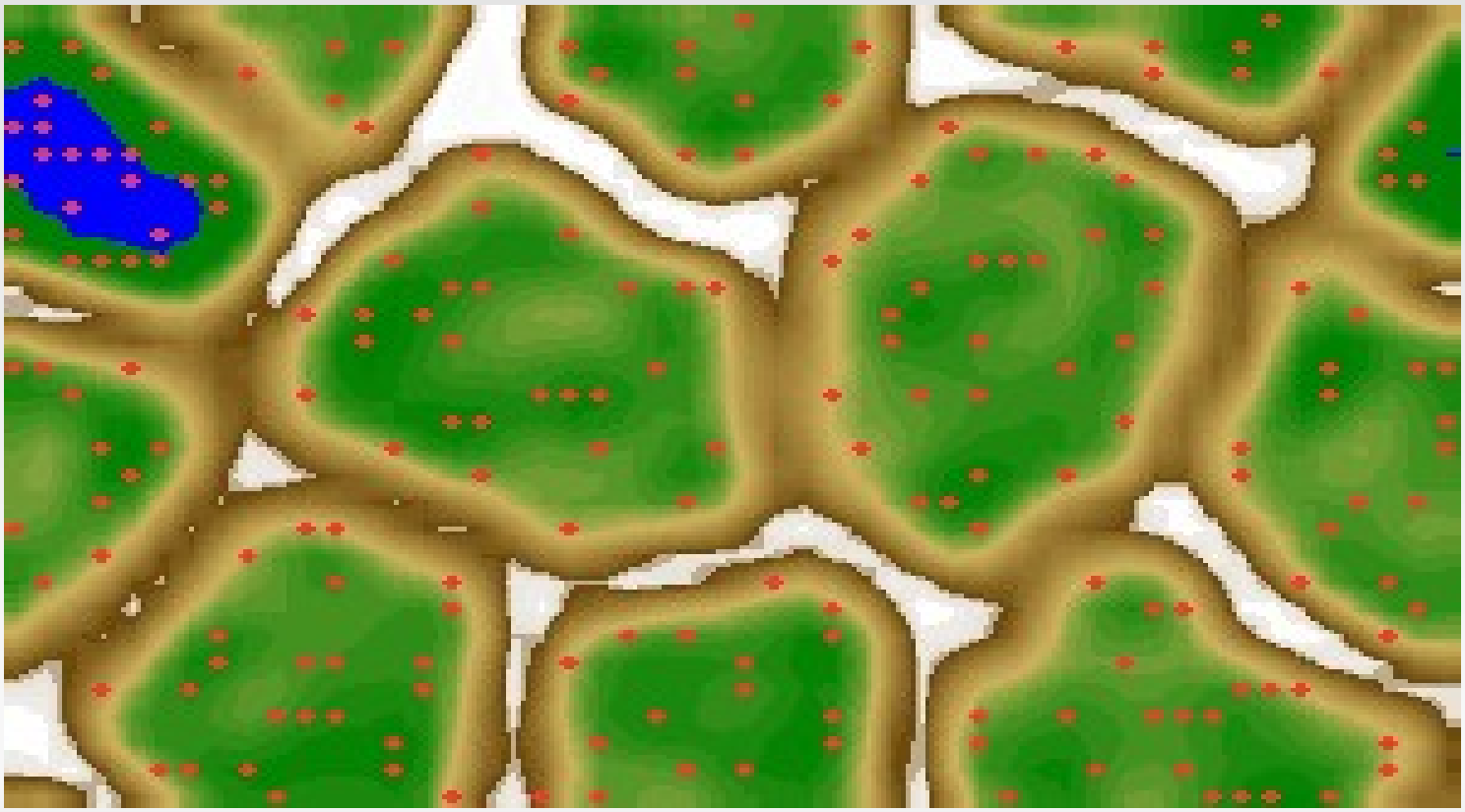
Коррекция визуализации

- Диаграмма плотности исходного пространства
  - Матрица плотности элементов входного набора данных около точки пространства входных данных отмеченной весами нейрона выходного слоя.
- Скорректированная унифицированная диаграмма расстояний
  - совмещает как методы кластеризации основанные на вычислении расстояния между элементами, так и методы основанные на плотности элементов в исходном пространстве

$$H_U*(n)=H_U(n)\cdot Q(n) \qquad Q(n)=\frac{H_P(n)-H_{cp}}{H_{cp}-H_{max}}+1$$

Метафора географической карты

Зелёный цвет — малые значения расстояний,
коричневый — большие.



Особенности разработки

- Количество элементов выходного слоя на порядок больше количества входных элементов.
- Размер исследуемых коллекций ограничен 30-200 документами.
- Тороидальная топология входного слоя.
- Использование скорректированной универсальной диаграммы расстояний.
- Учёт русской морфологии.

Направления дальнейших исследований

- Выбор модели построения текстовой коллекции в зависимости от параметров коллекции.
- Определение характерных параметров кластеров, именование кластеров.
- Автоматическое определение параметров обучения нейронной сети.
- Использование многословныхмоделей текста.
- Интерактивный режим работы.

Вопросы?

Ресурсы по теме:

- WEBSOM
  http://websom.hut.fi
- Databionic ESOM
  http://databionic-esom.sourceforge.net/
- РОМИП (Российский семинар по оценке методов информационного поиска)
  http://romip.ru/

alexander.sigachov@gmail.com