**Распределённый масштабируемый алгоритм для приближенного поиска ближайшего соседа в метрическом пространстве**[1]

А.А. Пономаренко

*Национальный исследовательский университет «Высшая школа экономики», Лаборатория алгоритмов и технологий анализа сетевых структур, ул. Родионова, 136, Нижний Новгород 603093, Россия*
*e-mail: aponomaremko@hse.ru*

Задача поиска ближайшего соседа является важной составляющей для таких областей как, распознавание образов [1], машинное обучение [2], семантический поиск документов [3].

Задача построения структуры данных для эффективного поиска ближайшего соседа формулируется следующим образом. Для функции метрики $\sigma : U \times U \to R$ определенной на пространстве всевозможных объектов $U$, требуется так организовать объекты из конечного множества $X$ в структуру данных $S$, так чтобы операция поиска ближайшего объекта к запросу $q$ из $X$ требовала как можно меньше вычислений функции метрики $\sigma$ .

Ранее было предложено множество алгоритмов, как для формулировки поиска точного ближайшего соседа, так и для приближенного поиска [5]. Вычислительная сложность всех точных алгоритмов экспоненциально зависит от размерности пространства. Причина этого лежит в «проклятье» размерности [4]. Алгоритмы для приближенной версии задачи, в меньшей степени зависят от размерности пространства.

В докладе речь пойдет об алгоритме построения структуры $S$ для приближенного поиска ближайшего соседа, в виде графа $G(V, E)$ , где $V = X$ . Предлагается строить граф $G$ обладающий навигационными свойствами тесного мира и содержащий в себе аппроксимированный граф Делоне. В алгоритме поиска основанного на «жадном алгоритме» заложена возможность варьирования точности поиска без необходимости изменений в структуре.

Вычислительные эксперименты на общедоступных тестовых наборах данных, демонстрируют производительность, значительно превосходящую все ранее известные методы.

Алгоритм не использует координатные представления и не требует привлечения свойств линейных пространств, а основан только на вычислении метрики между объектами и потому применим для данных из произвольных метрических пространств.

## ЛИТЕРАТУРА

1. T. M. Cover and P. E. Hart *Nearest neighbor pattern classification* Information Theory, IEEE Transactions on, vol. 13, no. 1, pp. 21-27, Jan. 1967
2. Salzberg, S. Cost, and Steven *Weighted Nearest Neighbor Algorithm for Learning with Symbolic Features*, Machine Learning, vol. 10, no. 1, pp. 57-78, 1993
3. S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman *Indexing by Latent Semantic Analysis*, J. Amer. Soc. Inform. Sci., vol. 41, pp. 391-407, 1990
4. E. Chavez, G. Navarro, R. Baeza-Yates, and J. L. Marroquin, *Searching in metric space*, Journal ACM Computing Surveys (CSUR), vol. 33, no. 3, pp. 273-321, Sep. 2001

[1]Работа выполнена при поддержке Лаборатории алгоритмов и технологий анализа сетевых структур НИУ ВШЭ, грант правительства РФ дог. 11.G34.31.0057