Название моего доклада

Иван Иванов

Данные

Уже похоже на нужные
нам данные

160 GB
Логов активности

97M
Пользователей

3M
Сообщест

13B
Сессий

Problem Formulation

Exact k-NN (Survey[1])

Definition 1 (Exact k-NN): *Given $(\mathcal{X}, \rho)$ - metric space, $X \subseteq \mathcal{X}$ - set of n points and $q \in \mathcal{X}$ - query point, task of Exact k-NN is to find a set $kNN(q) \subseteq X$ such that $|kNN(q)| = k$ and*

$$\forall x \in kNN(q)\ \forall x' \in X \setminus kNN(q)\ \ \big(\rho(q, x) \leq \rho(q, x')\big)$$

[1] Nitin Bhatia et al. "Survey of nearest neighbor techniques". In: *arXiv preprint arXiv:1007.0085* (2010).

Problem Formulation

Exact k-NN (Survey[1])

Definition 2 (Exact k-NN): *Given $(\mathcal{X}, \rho)$ - metric space, $X \subseteq \mathcal{X}$ - set of n points and $q \in \mathcal{X}$ - query point, task of Exact k-NN is to find a set $kNN(q) \subseteq X$ such that $|kNN(q)| = k$ and*

$$\forall x \in kNN(q)\ \forall x' \in X \setminus kNN(q)\ \ \big(\rho(q, x) \leq \rho(q, x')\big)$$

Examples of spaces

- $(\mathbb{R}^d, \|\cdot\|_2)$ — Euclidian space
- $\left(\mathbb{R}^d, \arccos \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\| \cdot \|\mathbf{y}\|}\right)$ — $\mathbb{R}^d$ with cosine distance
- ...

[1] Bhatia et al., "Survey of nearest neighbor techniques".

Part I

Tools and Systems for Big Data Storage and Processing

1. Hadoop and MapReduce
2. Apache Spark
3. Spark SQL

Table of Contents

Ключевые особенности

- Большое количество данных и признаков ($> 10^6$)
- Сильно разряженные данные
- Категориальные признаки большой размерности

Table of Contents

Конференции

KDD

Knowledge Discovery and Data Mining

RECSYS

Recommender System Conference

WWW

World Wide Web Conference

Вопросы?

Иван Иванов