Algorithms for Building Highly Scalable Distributed Data Storages

Alexander Ponomarenko
2nd International Scientific Conference “SCIENCE OF THE FUTURE”
September 20-23 2016, Kazan, Russia.
Hierarchy vs Heterarchy
Hierarchy vs Heterarchy
Hierarchy vs Heterarchy
Hierarchy vs Heterarchy

Data

Key

Distributed Network

Peers

Data

- Fox
- The red fox runs across the ice
- The red fox walks across the ice

Key

- DFCD3454
- 52ED879E
- 46042841
DHT protocols and implementations

• Aeropike
• Apache Cassandra
• BATON Overlay
• Mainline DHT - Standard DHT used by BitTorrent (based on Kademlia)
• CAN (Content Addressable Network)
• Chord
• Koorde
• Kademlia
• Pastry
• P-Grid
• Riak
• Tapestry
• TomP2P
• Voldemort
Applications employing DHTs

- BTdigg: BitTorrent DHT search engine
- cjdns: routing engine for mesh-based networks
- CloudSNAP: a decentralized web application deployment platform
- Codeen: web caching
- Coral Content Distribution Network
- FAROO: peer-to-peer Web search engine
- Freenet: a censorship-resistant anonymous network
- GlusterFS: a distributed file system used for storage virtualization
- GNUnet: Freenet-like distribution network including a DHT implementation
- Hazelcast: Open-source in-memory data grid
- I2P: An open-source anonymous peer-to-peer network.
- I2P-Bote: serverless secure anonymous e-mail.
- JXTA: open-source P2P platform
- Oracle Coherence: an in-memory data grid built on top of a Java DHT implementation
- Retroshare: a Friend-to-friend network[17]
- YaCy: a distributed search engine
- Tox: an instant messaging system intended to function as a Skype replacement
- Twister: a microblogging peer-to-peer platform
- Perfect Dark: a peer-to-peer file-sharing application from Japan
Structured Peer-to-Peer Networks: Chord Protocol

Distance function: \( d(x,y) = (y - x) \mod 2^m \)

Each node, \( n \), maintains a routing table with (at most) \( m \) entries, called the finger table. The i-th entry in the table at node \( n \) contains the identity of the first node, \( s \), that succeeds \( n \) by at least \( 2^{i-1} \) on the identifier circle, i.e., \( s = \text{successor}(n+2^{i-1}) \), where \( 1 \leq i \leq m \).
Structured Peer-to-Peer Networks: Kademlia

Distance function: \( d(x,y) = x \text{xor} y \)

to Overcome DHT Disadvantages

• DHT uses very simple distance functions
• Hashing destroys semantic of the data
• It’s hard to perform complex queries

Use nearest neighbour search in high dimensional metric space instead of exact search
Nearest Neighbor Search

Let \( D \) – domain
\( d : D \times D \rightarrow \mathbb{R}_{[0;+\infty)} \) - distance function which satisfies properties:

– strict positiveness: \( d(x, y) > 0 \iff x \neq y \),
– symmetry: \( d(x, y) = d(y, x) \),
– reflexivity: \( d(x, x) = 0 \),
– triangle inequality: \( d(x, y) + d(y, z) \geq d(x, z) \).

Given a finite set \( X = \{p_1, \ldots, p_n\} \) of \( n \) points in some metric space \( (D, d) \), need to build a data structure on \( X \) so that for a given query point \( q \in D \) one can find a point \( p \in X \) which minimizes \( d(p, q) \) with as few distance computations as possible.
Examples of Distance Functions

- **$L_p$ Minkovski distance** (for vectors)
  - $L_1$ – city-block distance
  - $L_2$ – Euclidean distance
  - $L_\infty$ – infinity

- **Edit distance** (for strings)
  - minimal number of insertions, deletions and substitutions
  - $d(\text{‘application’}, \text{‘applet’}) = 6$

- **Jaccard’s coefficient** (for sets $A, B$)
  \[ d(A, B) = 1 - \frac{\left| A \cap B \right|}{\left| A \cup B \right|} \]
Max Common Subgraph Similarity

\[ sim(G_1, G_2) = \frac{(|V(G_1, G_2)| + |E(G_1, G_2)|)^2}{(|V(G_1)| + |E(G_1)|) \cdot (|V(G_2)| + |E(G_2)|)} \]

\[ d(G_1, G_2) = 1 - sim(G_1, G_2) \]
Kleinberg’s Navigable Small World

Local edges \( p \)
\[ A \rightarrow E := d(A, E) \leq p \]

Long-range edges \( q_i \)
\[ \Pr(A \rightarrow Z) \sim 1/[d(A, Z)]^\alpha \]
Inverse \( \alpha^{th} \)-power distribution

Lattice distance
\[ d(A, Z) = |t-u| + |w-v| \]

VoroNet, RayNet: A scalable object network based on Voronoi tessellations

Distance function: \[ d(x, y) = \sqrt{x^2 + y^2} \]

Metrized Small World Algorithm

\[ u = 1 \]
\[ u = 2 \]
\[ u = 3 \]

“Top level” – first (oldest) elements

Navigable small world

“Bottom” level – all elements

\[ u = \log(N) \]
\[ u = \log(N) - 1 \]

R₁
R₂

query element


Boolean non-linear programming formulation for optimal graph structure

Decision variables

\[ x_{ij} = \begin{cases} 1, & \text{if edge } (i, j) \text{ belongs to the solution} \\ 0, & \text{otherwise} \end{cases} \]  \hspace{1cm} (1)  

\[ y^k_{ij} = \begin{cases} 1, & \text{if vertex } k \text{ belongs to the greedy walk from } i \text{ to } j \\ 0, & \text{otherwise} \end{cases} \]  \hspace{1cm} (2)  

Objective function

\[ \min \sum_{i=1}^{n} \sum_{j=1}^{n} O(i, j) \]  \hspace{1cm} (2)  

\[ O(i, j) = \left| \left\{ l \in V : \exists k \text{ such that } x_{ik} = 1 \text{ and } y^k_{ij} = 1 \right\} \right| \]  \hspace{1cm} (4)  

Constraints

\[ x_{ii} = 0 \quad \forall i \in V \]  \hspace{1cm} (5)  

\[ y^i_{ij} = y^j_{ij} = 1 \quad \forall i, j \in V \]  \hspace{1cm} (6)  

\[ \sum_{k=1}^{n} x_{ik} y^k_{ij} \geq y^l_{ij} \quad \forall i, j, l \in V \]  \hspace{1cm} (7)  

\[ l^* = \arg \min_{l \in V : x_{ii} = 1} (d(l, j)) \Rightarrow y^l_{ij} \geq y^k_{ij} \quad \forall i, j, k \in V, j \neq i, k \neq j \]  \hspace{1cm} (8)
Search by greedy algorithm
Construction algorithm
Data sets

- CoPhIR (L2) is the collection of 208-dimensional vectors extracted from images in MPEG7 format.
- SIFT is a part of the TexMex dataset collection available [http://corpus-texmex.irisa.fr](http://corpus-texmex.irisa.fr) It has one million 128-dimensional vectors. Each vector corresponds to descriptor extracted from image data using Scale Invariant Feature Transformation (SIFT)
- Unfi64 is synthetic dataset of 64-dimensional vectors. The vectors were generated randomly, independently and uniformly in the unit hypercube.
Performance of a 10-NN search for $L_2$: plots in the same column correspond to the same data set

KL-divergence: \[ d(x, y) = \sum x_i \log \frac{x_i}{y_i} \]

Final16, Final64, and Final256: are sets of 0.5 million topic histograms generated using the Latent Dirichlet Allocation (LDA).
Wikipedia dataset

Vector Space Model

\[ d_j = (w_{1,j}, w_{2,j}, ..., w_{n,j}) \]
\[ q = (w_{1,q}, w_{2,q}, ..., w_{n,q}) \]

\[ \text{sim}(d_j, q) = \frac{d_j \cdot q}{\|d_j\| \cdot \|q\|} = \frac{\sum_{i=1}^{n} w_{i,j} w_{i,q}}{\sqrt{\sum_{i=1}^{n} w_{i,j}^2} \sqrt{\sum_{i=1}^{n} w_{i,q}^2}} \]

Wikipedia (cosine similarity): is a data set that contains 3.2 million vectors represented in a sparse format.
This set has an extremely high dimensionality (more than 100 thousand elements). Yet, the vectors are sparse: On average only about 600 elements are non-zero.
Wikipedia is dataset that contains 3.2 million vectors represented in a sparse format. Each vector corresponds to the frequency term vector of the Wikipedia page extracted using the gensim library. This set has an extremely high dimensionality (more than 100 thousand elements).
Scaling of MSW data structure
Summing up

• Algorithm is very simple
• Algorithm uses only distance values between the objects, making it suitable for arbitrary spaces.
• Proposed data structure has no root element.
• All operations (addition and search) use only local information and can be initiated from any element that was previously added to the structure.
• Accuracy of the approximate search can be tuned without rebuilding data structure
• Algorithm high scalable both in size and data dimensionality

Good base for building many real-world extreme dataset size high dimensionality similarity search applications
Source Code

https://github.com/searchivarius/NonMetricSpaceLib
https://github.com/aponom84/MetrizedSmallWorld
Questions?
Questions?

Why CERN doesn’t use DHT?