A Dimension Reducing Indexing Scheme for Guaranteed Keyword Searches in Peer-to-Peer Storage Systems

Cristina Schmidt  Manish Parashar
TASSL, Department of Electrical and Computer Engineering, Rutgers University

Goal
Develop a P2P storage system that enables complex keyword searches, guaranteeing that all data elements matching a query will be found with reasonable costs in terms of number of messages and queried nodes.

Motivation
• Recent proliferation of P2P storage systems
• The need to retrieve data stored in P2P systems, using partial keywords

System architecture – components
Locality preserving mapping
• Maps data elements to indices based on the document’s keywords
• Use Space-Filling Curves (SFC)

Properties:
• Continuous mapping from d-dimensional space to 1-dimensional space

Space-Filling Curves
• The need to retrieve data stored in P2P systems, using partial keywords
• Recent proliferation of P2P storage systems

Recent proliferation of P2P storage systems makes the use of space-filling curves (SFC) needed for indexing and querying data in a P2P system. The SFC-based index space preserves keyword locality, which means that clusters in the index space will be close together in the 1-dimensional space.

Keyword 1: A document is a point in a multidimensional keyword space.

Keyword 2: Digital causality: points on the segment contained in a cell have the first digit identical to the digit of the previous approximation of the line in that cell.

Note: Cluster = a group of grid points that are consecutively connected by a mapping (or a curve)

Figure: Recursive refinement of the query (011, *) viewed as a tree. Each node is a cluster, and the bold characters are the cluster’s prefixes.

Figure: Embedding the leftmost tree path (solid arrows) and the rightmost path (dashed arrows) onto the overlay network topology.

Experimental Evaluation
Evaluating the Query Engine
Experiment: 2D keyword space, the system size increases from 1000 nodes to 5400 nodes, and the number of keys stored increases from $2 \times 10^4$ to $10^6$.

Figure: Results for queries with one keyword or partial keyword, e.g. (computer, *).

Figure: Results for queries with two keywords or partial keywords (at least one partial keyword), e.g. (comp*, net*).

Figure: Results for queries with one keyword or partial keyword, e.g. (computer, *).

Evaluating the Load Balancing Mechanisms
The distribution of the keys at nodes when using only the load balancing at node join technique.

The distribution of the keys at nodes when using both the load balancing at node join technique and the local load balancing.

Future work
• Improving the system’s availability and response time (caching and redundancy)
• Making the system fault tolerant
• Developing new overlay topologies
• Address issues like hot-spots, ranking the documents, security