Chapter 5  
Dimension Modeling

5.1. INTRODUCTION

A concept which occurs frequently in the cluster documents provides enough information to represent a cluster. The frequent subgraphs are semantic patterns of a cluster. Monika (2008), Kiran (2010). They are indexed as cluster indices. Each cluster represents one dimension (with dominating concept). Therefore, a node can have n clusters of d dimensions. In each dimension, there can be m documents.

Say if the repeating pattern is \{cyclone, windstorm, weather\_condition\}, \{depression, cyclone\} in a cluster, then the topical semantics of the cluster it represents is \{cyclone, storm and weather\_condition\}. This constitutes cluster key (or summaries of keys pertaining to a domain) and it may contain m documents in this dimension.

5.2. SOLUTION OVERVIEW

A node P can contain d dimensions and n semantic vectors corresponding to each dimension. Semantic indices are assigned to cluster keys in each dimension. It is realized using hash structure with key value pairs. Since we have used Chord based overlay network, it would be more natural to adopt fully distributed index which is based on topical semantics. Clusters keys are used to map documents on a virtual dimensional space with distributed hashing in overlay network. Each cluster key is added to the system using hash values.
For a n bit identifier space, we can have $2^n$ virtual nodes on Chord ring. The dimension of Chord is determined based on semantic dimensions. Semantic dimensions are larger compared to the number of physical nodes. We have not created zones as in CAN preserving local storage autonomy of peer nodes. The notion is the peer nodes hold the ownership of the document they contain. Let $t_i$ be arbitrary number of semantic dimensions on a node $P_i$. Similarly every node will have its own dimensions. An artificially generated sample dimension on a peer is shown in figure 5.4. Therefore the dimensions on an individual peer can be modeled as three dimensional matrix $D$, where $t$ represents dimension vector, $d$ represents document vector, $c$ represents cluster vector as shown in figure 5.3. The cluster keys are mapped on to finger table.

$$D = [V_{t,d,c}]$$

When a query is initiated, the semantics of the query is extracted to match with cluster keys found locally from finger table, thereby it collects candidate peers usually neighbors with similar semantics. The cosine similarity is computed with query vector and semantic dimensions of peer nodes. The nodes with high similarity ratios are accessed to receive final result for the query initiated. The query processing is depicted in figure 5.1.

For $R$ keys and $N$ nodes, each peer will hold $B$ number of keys where $B = R / N$. By property of Chord, the keys will be evenly distributed with $B$ number keys on nodes. The index is rotated to get different dimensions of semantics. To understand the semantic indexing, let us see a small illustration.

Figure 5.2a map documents to topics or concept terms. With this matrix, it is possible to find documents mapped on to specific concepts. On rotating this index, it is possible to
find concepts mapped on to specific documents figure 5.2b. Figure 5.2c shows clusters after computing similarity using inverse of cosine similarity.

Figure 5.1. Query Processing and Peer Node Access
Figure 5.2a. Document to Topics (Concepts) index

Figure 5.2b. Concept to Documents index

Figure 5.2c. Cluster Formation after Similarity Computation
Figure 5.3. Semantic Dimensions against Documents and Clusters

Figure 5.4. Artificially Generated Dimensions Samples on Peer Node
5.3. CONCLUSION

Each peer is heterogeneous in nature. Each peer can contain n number semantic dimensions. The dimension modeling strongly depends on the underlying Distributed Hash Table (DHT) based Chord network (i.e.) the underlying semantic overlay network. For n bit identifier there will be $2^n$ identifiers (physical nodes). Each peer nodes implements local storage autonomy therefore the semantic domains are not forced. Naturally existing semantic dimensions are organized as clusters. Each cluster belongs to one dominating semantic domain.

The modeled dimensions in this chapter are implemented in P2P overlay search system explained in the next chapter 6.