ABSTRACT

Recent advances in Peer-to-Peer technology have led Industries to transform from Client-Server computing to Peer-to-Peer computing due to advantages like fault tolerance, low cost and scalability. Substantial information in the form digital documents is distributed over these Peer-to-Peer systems. Hence, it becomes difficult to extract relevant information in response to query. Inherently, two issues need to be addressed to retrieve relevant documents: 1. Locate the cluster and node containing relevant documents 2. Retrieve the relevant document despite of missing semantics i.e. no meta data stating class of document or no specific domain ontology is available.

This thesis addresses the above stated issues. First, fairly a novel P2P coordination model is proposed to extract cluster and node information from large P2P system. Second, concept based clustering method is incorporated to derive conceptual semantics of documents in order to cluster them. Further, cluster keys which would represent clusters more aptly are extracted. Third, semantic dimensions are modeled and indexed over peer nodes to facilitate search process. Fourth, the cluster and neighboring nodes containing relevant content are extracted to bootstrap the search process in order to visit fraction number of nodes.

We introduce P2P catalog system to extract cluster and node information in order to initiate the search process and retrieve documents efficiently. The documents distributed over P2P network are conceptually clustered and characterized by cluster keys to represent the dimension of the cluster. Every peer will carry $n$ number of dimensions. Where, these dimensions cannot be mapped directly on to peer nodes due to curse of dimensionality. Therefore, the clusters are mapped on to logical nodes in DHT based CHORD ring. On initiating a query from an arbitrary peer, the peer node consults its own finger table to discover similar clusters available on neighboring peers. The cluster indices are shipped to get the reduct from neighbors. The peers containing relevant content are the candidate peers. The indices corresponding to the query are shipped to candidate peers and semi join is performed to retrieve the content in response to initiated query.
Thereby our solution provides means to extract cluster and node information to locate the relevant content. It also offers local storage autonomy without being forced to store contents like other Content Addressable Networks. Indirectly it minimizes the document movement since only indices of clusters are shipped for search. Documents are clustered according to semantics. Therefore, even a document of unknown class can also be clustered under the proper class. The search for content is not initiated at random point. Instead the search starts with the node where the relevant content is available based on semantic distribution of clusters. The peer nodes containing relevant content are gathered first. Then the search query is communicated only to those peers on the virtual ring.

The algorithms proposed in this thesis offer flexibility and local storage autonomy to achieve results similar to Content Addressable Semantic Networks.