Topic 3: Data-centric Applications

1 Guidelines

In this topic, you should present various data sharing applications that run on P2P networks, including file sharing systems, information retrieval applications, structured data management, and high dimensional data sharing. The common goal of these different applications is to enable data contributed by individual peers sharable in the whole P2P networks in a transparent (to the end user), easy, and flexible way. However, viewing from the aspects of data management, these applications deal with different data formats on different granularities, and they face different application requirements and design challenges. Your focus in this topic will be on the query processing issues in these applications.

In file sharing applications, various files (movies, songs, etc..) are shared as a whole in the system, and the most typical query is look up of files in the nodes of P2P networks based on matching their names. The processing of such query is essentially a search problem — searching in the online nodes in the P2P networks for the desired object. You will present various searching strategies and analyze their trade-offs.

In contrast to file sharing, information retrieval (IR) applications in P2P networks share data at finer granularities such that the content of files can be queried and results are expected to match the queries semantically. You should mainly discuss retrieval of textual documents in P2P networks as an example to illustrate how query processing is performed, where the queries are in the form of conjunction of keywords, or even a whole sentence. Traditional IR techniques should be adopted effectively to combine with efficient P2P-based index structures for processing such queries. You should introduce various methods devised in literature for processing keyword queries in IR applications in P2P networks.

Further, in addition to unstructured data, there is also need to share structured data, such as relational databases and XML documents, in P2P networks. Supporting such applications is more challenging than previous ones because there often exist plenty of structural and data heterogeneities among data sources residing at different autonomic peers. Although this problem does not newly arise in P2P networks, the decentralization and dynamism requirements of P2P paradigm make it especially tough, and consequently traditional approach cannot be directly applied. You should describe current state-of-the-art techniques for modeling and building schema mappings between the databases shared by different peers. Followed it, you should present various query processing methods for both keyword queries and structured queries by exploiting the built schema mappings.

Finally, you should introduce a broad range of methods for processing various queries, including multi-attribute query, multi-dimension query, KNN query,

skyline query, and partial match query, in P2P networks. These queries are frequently used in lots of applications such as multimedia retrieval, data mining, and spatial databases that deal with high-dimensional data. You should present various overlay structures for P2P networks and indexing techniques that support distributed processing queries over high-dimensional data.