Statement of Research Interests
Prasenjit Mitra

I intend to pursue research on the principles of database systems. I have worked on the problems in information integration and interoperation especially from autonomous, heterogenous data sources over the Internet. In the future, I plan to extend my research in the areas of information processing, databases and the Internet where problems of interoperation occur. I expect that my research will apply to areas requiring the handling of huge amounts of information like bio-informatics, supply-chain management, etc.

Prior Research:

Rewriting queries using views:
Often, databases provide views of the underlying data to their users. A typical end-user does not have access to the entire database – say on the Internet site amazon.com – but only gets a view of the database. A view can be defined to provide data that one typically gets by executing a query on a database. In our work, we considered queries and views that are conjunctive queries with arithmetic comparisons. End-users pose queries based on a set of predicates called base predicates. In order to optimize the execution of a query, we need to generate a query plan by rewriting the query using views. We have conducted research on finding query rewritings that are a) *maximally-contained* in the query and b) *equivalent* to the query. The problem of rewriting queries using views is closely tied to the problem of *query containment* since the rewriting needs to be contained in the original query. In our work, we extended the existing results on query containment available in the literature. Based on our new results, we designed efficient algorithms for answering queries using views, especially in the presence of arithmetic comparisons.

Semi-automatic mapping of schemas and ontologies:
The Internet has enabled people to publish information without constraints, resulting in millions of information sources and vast amounts of data of mixed quality. To access this information, end-users use a search engine, like Google, and then sift through a large number of web-pages manually to extract the information they need. Often, the information end-users need is not available from a single source but needs to be composed manually from several sources.

Several distributed information-processing systems have been built to help the end-user automate data integration on the Web. However, most of these efforts involve significant manual effort. Clearly, manually constructing data integration engines becomes impossible when the number of sources being integrated is large. Also, automating the construction of a data-integration engine has not worked well, because the semantics of the terms used in the sources differ and are not explicitly defined.

To assist automated processing of information, scientists who need to publish information are designing *ontologies* expressed as graphs that contain the metadata that defines terms used in information sources and the relationships among different terms.

Before we can compose information from multiple sources, we need to resolve their semantic heterogeneity. I have designed and implemented several heuristic algorithms that generate articulation rules – Horn rules expressing the semantic mapping between objects in different sources – among the ontologies semi-automatically. Experiments I conducted show that an articulation generator employing multiple heuristics – like looking up thesauri and dictionaries, generating word similarity based on a corpus of texts, matching objects across ontologies based on the structure of the ontology graphs and on the instances and attributes of objects, significantly reduces the manual effort required to articulate information sources.

Ontology Composition:
Often, we need to compose ontologies. For example, when two corporations merge, they need to integrate their ontologies into one. Articulation rules establish the semantic mapping between ontologies and form the basis for such composition.

To formalize the composition of schemas or ontologies, I have proposed an Ontology-Composition Algebra. I have designed ONION – a system for ONtology compositION – based on the algebra. In relational algebra, queries are optimized exploiting the properties of the operators. Similary, optimization

of tasks involving composition of ontologies can be performed based on the properties, such as commutativity and associativity, of the operators. I have shown that the properties of the operators depend upon the articulation generation function that establishes the semantic mapping. Prior research has not considered the effects an articulation-generation function has on the eventual scalability of information composition tasks. My work shows that while designing articulation generation functions one needs to ensure that the algebraic operators that use them have desirable properties so that a system can use these properties to optimize the composition of ontologies.

Future Work:

My previous research raises a number of questions that need intensive investigation:

Query Rewriting and Optimization in Peer-to-Peer Systems: Peer-to-peer systems have several similarities to the distributed, autonomous and heterogeneous systems that I have worked on. My work on query rewriting and optimization has the potential to be used in peer-to-peer systems. I intend to investigate the similarities and differences and accordingly modify our algorithm to design an efficient query rewriting and optimization algorithm for peer-to-peer systems.

Information Extraction and Data Mining: In order to obtain maximum automation, especially in rapidly changing sources where manual extraction of information is inefficient, I would like to study the problems of extracting structured and unstructured data from the Web and mining the data to extract accurate information.

Maintenance and its effect on Information Integration: Though we have designed the system to be easily maintainable, I have not extensively studied the problems caused by frequent changes of websites. As information and software inventories grow, one needs more and more resources to keep them up-to-date. What modifications or enhancements does ONION need to allow rapid re-deployment of the toolkit in the face of frequent updates? Are the operators that we have sufficient or do we need other operators to formalize different real-world applications?

Applications and Scalability: Apart from the Semantic Web for which ONION was developed, the problem of data integration and interoperation is acute in several fields. I intend to work closely with scientists at the university who use large sources of distributed information like in domains such as bio-informatics, medical informatics, and operations research to find settings to deploy the tool I have developed. We have not evaluated ONION for large systems and I intend to research the issues of scalability in real-world applications.

While researching information processing systems, I intend to design and use tools and algorithms from a wide range of fields from database system design, query optimization, data mining and information extraction to natural language understanding, graph visualization and human computer interaction, and computational logic. I believe my research experience as a graduate student makes me a good candidate to pursue a career in research in databases and information processing systems.