Human-Powered Data Cleaning for Probabilistic Reachability Queries on Uncertain Graphs

Abstract:

Graph data arises in a large variety of real-world applications, such as social networks, knowledge graphs, and protein-protein interaction networks. However, the data in these applications is often uncertain for various reasons that include unreliable data sources, ambiguity in the content of data, or approximate data models. For example, duplicate author names may confuse the coauthor networks that are automatically extracted from academic websites (e.g., DBLP and Google Scholar); in a machine-generated knowledge graph, the relationships between entities may be uncertain since they are usually extracted from web documents by using natural language processing techniques; in a protein-protein interaction network, the interactions between proteins are typically derived by a statistical model.

Existing System:

Uncertain graph models are widely used in real-world applications such as knowledge graphs and social networks. To capture the uncertainty, each edge in an uncertain graph is associated with an existential probability that signifies the likelihood of the existence of the edge. One notable issue of querying uncertain graphs is that the results are sometimes uninformative because of the edge uncertainty. In this paper, we consider probabilistic reachability queries, which are one of the fundamental classes of graph queries. To make the results more informative, we adopt a crowd sourcing-based approach to clean the uncertain edges. However, considering the time and monetary cost of crowd sourcing, it is a problem to efficiently select a limited set of edges for cleaning that maximizes the quality improvement.

Disadvantages:

- Takes more time in edge selection.
- The problem of data cleaning for probabilistic reachability queries on uncertain graphs.

Proposed System:

We proposed a correlated factor P↤(e) to facilitate the edge selection in crowdscouring-based cleaning. We developed a number of optimization techniques and pruning heuristics for reducing the computation time in edge selection for both single queries and multiple queries. Extensive experimental results demonstrate the effectiveness and efficiency of our proposed algorithms under various system settings.

Advantages:

- Takes less time in edge selection.
- The problem of data cleaning for probabilistic reachability queries on uncertain graphs is solved.

Modules:

- Single-edge selection algorithm for single query.
- Single-edge selection for multiple queries.
- Multiple-edge selection algorithm for multiple queries.

\ **SYSTEM REQUIREMENTS**

H/W System Configuration:-

Processor	-	Pentium –III
RAM	-	256 MB (min)
Hard Disk	-	20 GB
Key Board	-	Standard Windows Keyboard
Mouse	-	Two or Three Button Mouse
Monitor	-	SVGA

S/W System Configuration:-

Operating System	:	Windows95/98/2000/XP
Application Server	:	Tomcat5.0/6.X
Front End	:	HTML, Jsp
Scripts	:	JavaScript.
Server side Script	:	Java Server Pages.
Database	:	MySQL 5.0
Database Connectivity	:	JDBC