DATA WAREHOUSING RESEARCH TRENDS

- Research trends in analytics relevant to data warehousing and OLAP include [Cuzzocrea et al]:
  - Data source heterogeneity and incongruence
  - Filtering out uncorrelated data
  - Strongly unstructured nature of data sources
  - High scalability
  - Combining the benefits of RDBMS and NoSQL database systems
  - Query optimization issues in HiveQL
  - Integrating multidimensional data models and techniques with Hadoop
  - Visualization of big data analytics
DATA MINING RESEARCH TRENDS

• Data mining research trends include exploration of the following [Han & Kamber]:

  • Application exploration
  • Scalable and interactive data mining methods
  • Integration of data mining with search engines, database, data warehouse and cloud database systems
  • Mining social and information networks
  • Mining spatiotemporal, moving-objects and cyber-physical systems
  • Mining multimedia, text, and web data
  • Mining biological and biomedical data
• Data mining with software engineering and system engineering
• Visual and audio data mining
• Distributed data mining and real-time data stream mining
• Privacy protection and information security in data mining

• The above areas for both data warehousing and data mining largely have a computer technology focus.

• One of our original definitions of a data warehouse was “… a collection of technologies aimed at enabling the knowledge worker to make better and faster decisions”.

• Hence, business goals also need to be addressed in considering the use of these technologies.

• Also, it is increasingly recognised that the complexity of the data and multiple technologies used by a “knowledge worker” introduces its own problems.
PRACTICAL LESSONS FROM LARGE-SCALE DATA MINING

• The importance of domain knowledge. “Look and know your data.”

• Many data mining projects fail because of the lack of adequate data: suitable data pre-processing, sampling… are crucial.

• Understand the business goals when developing a mining model.

• Important metrics in training and deployment of mining models: offline training time, online prediction time, model size or memory footprint.

• Understand the impact of architecture and scalability issues on algorithm design and implementation.

• Importance of people in considering the above.

• Importance of two under-researched areas:
  • visualization
  • real-time interaction with large datasets.
VISUAL ANALYTICS

• Visualization of complex data and the results of analysis of that data requires techniques which include the following:

  • Software architectures which support visualization functionality as an integral service.

  • Support for a workflow of operations which enables data to be visualised while being filtered, summarised and analysed.

  • Support for dynamic data so that visualization of changes without a need for a global recomputation from scratch is possible.
Reading (Optional)


