• Objectives:
  - Formulate “hypothesis mining” as contextualized comparative pattern mining. Develop algorithms for hypothesis mining and analysis. Build an easy-to-use system based on these algorithms.

• Novelty:
  - Hypothesis as a contextualized comparative pattern
  - P-value for controlling false-positive hypothesis
  - Comparison between hypotheses to identify actionable hypotheses, redundant hypotheses, Simpson’s paradoxes

• Scope & Deliverables:
  - Core algorithms for hypothesis generation and novel analyses described above, as well as OLAP operations for exploring hypotheses.
  - Graphical user interface (GUI) which supports basic functions for summarizing and visualizing data, visualization of discovered hypotheses, and visualization of OLAP operations

• Achievement #1
  - Novel formulation of hypothesis patterns as contextualized comparative patterns. E.g., (Race=Chinese), Drug=A, Response=positive
  - Novel data-driven paradigm of hypothesis generation and testing
  - Novel and efficient algorithms for generating significant hypotheses, isolating reasons behind significant hypotheses, and detecting confounding factors that form Simpson’s paradoxes with significant hypotheses

• Associated Technology
  - Implemented these algo’s into the EHTA system, the mining engine of iDIG in FR

• References
  - Liu, et al. Towards exploratory hypothesis testing and analysis. ICDE 2011, pp. 745-756
  - Liu, et al. Supporting exploratory hypothesis testing and analysis. ACM KDD, in press

• Achievement #2
  - Proving necessity of controlling false positives in class-association rule mining. Many spurious rules are produced if no correction is made
  - Proving that permutation-based approach is most effective in controlling false positives, and develop techniques to make it efficient

• Associated Technology
  - Implemented these techniques into ARminer

• References

• Achievement #4
  - Association rule visualization for exploratory data analysis
  - Relationship among rules reveal deep info of the data
  - Summarize this, with visualization, to help users understand the data and to suggest hypotheses to test

• Associated Technology
  - Implemented in AssocExplorer, the visualization engine of iDIG in FR

• References

• Achievement #3
  - Finding minimum representative rule set that can 1) represent all patterns with a minimum # of representative patterns and 2) restore the support of all patterns with error guarantee
  - Algorithms for doing the above efficiently: MinRPset (efficiently produces the smallest solution), FlexRPset (trades solution size for higher speed)

• Associated Technology
  - Implemented these into the EHTA system, the mining engine of iDIG in FR

• References

• Achievement #5
  - Management of large collections of frequent items for analysis and user exploration.
  - Refinement of CPFtree to index to provide efficient exact match, subset/superset search, etc. of frequent items

• Associated Technology
  - Implemented in the EHTA system, the mining engine of iDIG in FR

• References

Contact- Professor Wong Limsoon (wongls@comp.nus.edu.sg)
Funding- A*STAR SERC PSF grant # 102 101 0030