Parallelizing the Itemset Tree Data Structure

Project proposal for CSCE 561 – Fall 2015
Jennifer Lavergne
Project Goals

- **Phase 1:** Modify to itemset tree algorithm specified in general mining paper to either:
  - Build separate trees
  - Break apart a completed tree

- **Phase 2:** Modify the itemset tree search algorithm to run parallel on multiple trees
  - Run on multiple trees
  - Combine results

- **Phase 3:** (a) Test different ways to split the tree and recombine the results. (b) Test different support calculations and minimum support thresholds.
BACKGROUND
Association Rules

- **Association rule** - an implication \( \{X \Rightarrow Y, \text{support, confidence}\} \). Where \( X \) and \( Y \) are subsets of the itemset \( I \) and \( X \cap Y = \emptyset \)
  - Example: \( \{\{\text{bread, milk}\} \Rightarrow \{\text{cheese}\}, 30\%, 75\%\} \)

- **Support** = \#occurrences of \( I \) in database/\#rows in database
  - **Minsup** – The minimum support threshold for an itemset \( I \) to be considered frequent

- **Confidence** = \( \text{Support}(X \cup Y)/\text{Support}(X) \) for itemset \( I = X \cup Y \).
  - **Minconf** – a user specified threshold that indicates the interestingness of a candidate rule \( I \): \( \text{conf}(I) \geq \text{minconf} \)
Itemset Trees

- A data structure which aids in users querying for a specific itemset and its support: Targeted Association Mining

- Item mapped to numeric values: \{bread\} = \{1\}, \{cheese\} = \{2\}
  - Numbers must be in ascending order within the itemset
  - Ex: I = \{1, 2, 56, 120\}

- **Note:** Can be used to find all or specific rules within a dataset.
PROJECT DESCRIPTION
Phase 1: Itemset Tree Code

• Modify existing code or write your own.
  – Build separate trees
  – Break apart a completed tree

• Read papers:
  – Itemset trees
  – Ordered Min-Max Itemset Tree
Phase 2: Parallelize search

• Modify existing code or write your own:
  – Run parallel search algorithm on multiple trees
  – Combine results of the parallel

• Possibly use a modified support calculation and min-sup threshold based upon the number of subtrees and the overall support of each subtree.
  – support = count (itemset in subtree)/total in main tree
  – support = count (itemset in subtree)/total in subtree
  – Try others?
Phase 3: Tests

Test different ways to split the tree and recombine the results.

- **Test 1**: Generate and split an unordered tree v.s. generating multiple unordered trees.

- **Test 2**: Generate and split an ordered tree v.s. generating multiple ordered trees.

- **Test 3**: Test other support calculations and different minimum support thresholds.
Questions?

Jennifer Lavergne
jjslavergne@louisiana.edu
References

• M. Kubat, A. Hafez, V. V. Raghavan, J. Lekkala, and W. K. Chen, “Itemset trees for targeted association mining”, *IEEE Trans. on Knowledge and Data Engineering*, 2002
