Progress Report for Reproducing the Paper "Generating Semantic Annotations for Frequent Patterns with Context "

Raj Datta, Weidi Ouyang, Ye Xu
Team TR Squirrels

1) Which tasks have been completed?

- Thorough Study of Paper and understanding of options
- Presentation draft prepared covering overview of paper
- Tightly defined scope, in discussion with Bhavya, lead TA
    - To choose only one dataset (chosen to be DBLP)
    - To choose and use only one clustering algorithm
    - To target completing the context modeling and frequent pattern mining steps through part 4.1 of the paper
    - Consider 4.2 and 4.3 to be out of scope unless time allows after completing 4.1
- Obtained clarification on various open questions and direction from Bhavya, lead TA
- Obtained the raw Dataset (DBPL.xml)
- Cleaned & Parsed the Dataset into csv format with "author" and "title" columns
- Decided the proper algorithm and libraries for "Closed Frequent Pattern" mining and Redundancy reduction (test pre-processings and pattern clustering).
- Using UIUC paper authors, created a toy dataset to test the concepts, algorithms and libraries.
- Initial trial on the DBLP dataset to generate author itemset frequent patterns

2) Which tasks are pending?

- Finalize selection of toolset/library for closed frequent patterns mining on author list and title list
- Generate formal itemsets and sequential frequent patterns for clean data set
- Using hierarchical clustering algorithm, remove Redundancy from the initial closed frequent patterns
- Finalize the context modeling by implementing the weighting function on context indicator and pattern pairs
- Analysis of Results
- Finalize presentation and report based on implementation and results

3) Are you facing any challenges?

- Understanding the concepts and the algorithms in the paper
    - Paper was written for many general scenarios (e.g. graphs/subgraphs) which aren't necessarily applicable to our implementation; trying to make it more generic made it more difficult to understand the applicability and our relevant extracts
    - Some of the concepts weren't covered in depth in class (e.g. closed frequent patterns, maximal frequent patterns.)
    - Some ambiguity (e.g. stop word removal, specifics of laplace smoothing)

    - Since the paper is dated, we needed to find/explore some of our own more recent libraries and tools, hoping there would be minimal impact on the end result
- Implementing weighting functions (Mutual Information) to build Context indicator vectors
- Timeline for completing all remaining optional parts of the paper.