PIRE 2009 Project Proposal

Student Name: Lester Melendez

Student's School: FIU

Student Email: lmele002@cs.fiu.edu

Student Home Page: http://www.cs.fiu.edu/~lmele002

Student Rank: PhD

Student Expected Graduation Date: Fall 2010

Supervisor's Name and Title at FIU/FAU: Dr. Naphtali Rishe, Professor, HPDRC Director

Name of the PIRE International Partner's Institution: Universidad Politécnica de Cataluña/Barcelona Supercomputing Center

Supervisor's Name and Title at the PIRE International Partner's Institution: Rosa Badia, Jordi Torres, David Carrera

Project Title: Computing Object Similarity with MapReduce

Problem Statement: *There is a great deal of inefficiency in computing the similarity between a given object and all objects in a given dataset when conducted in a sequential manner. I aim to explore the possibility of utilizing a distributed computing paradigm, MapReduce, to bring this task to efficiency levels that allow for the dynamic addition of functionality to various applications.*

Motivation and Impact: *The motivation for this work began while working at IBM Almaden with the Midas group. There was a great deal of data out there but, turning it into useful information took considerable effort. I felt it necessary to find a way that allows one to turn that raw data into useful information dynamically and in as close to real-time as possible. The impact lay in the notion of providing dynamic functionality to applications. I am focusing on computing similarities between and object and objects in a given dataset. My work will provide a MapReduce template and infrastructure for this problem.*

Current Status: *Currently I have completed the first version of the MapReduce object similarity computation template application. It is running on various implementations of Hadoop. The main goal right now is to tweak a few parameter passing and file reading issues so as to improve efficiency. I will be completing a second application template in the next couple of weeks. The second template will allow a developer to treat an entire file as an object and add more data mining functions other than similarity computation.*

Research Roadmap:

- **Week 1**
  - Transition from working with a Hadoop 0.18.3 implementation through cygwin and work with Hadoop 0.20 running on Ubuntu 9.4
  - Find all available Hadoop and MapReduce related official documentation
  - Build repository of MapReduce examples and tutorials
  - Explore possible datasets to use in application development
    - Spatial Data

        - Linked Open Data
        - Sports Statistics
        - Social Network Profiles

- **Week 2**
    - Confer with various human and internet sources to get feedback on what has already been done related to this matter
    - Evaluate current sequential similarity computation paradigm to see where it needs the most improvement
    - Evaluate datasets explored last week for feasibility of exploitation

- **Week 3**
    - Go through Google's MapReduce on-line course
    - Begin running initial MapReduce helloWorld style applications in pseudo-distributed mode
    - Begin processing datasets and preparing them for processing
    - Set up IBM systemT, JAQL, and other Midas development tools
    - Implement the Explore, Extract, and Scrub phases based on IBM's Midas

- **Week 4**
    - Develop non-MapReduce classes in Java for use in the similarity comparison application
    - Classes such as a web crawler, data extraction tools, spatial functions, search tools, etc.
    - Ensure that the development is enough to quickly implement sequential versions of the final applications

- **Week 5**
    - Tackle MapReduce related issues concerned with processing heterogeneous datasets
    - Create classes for processing files as records as well as multiple records per file
    - Test developed MapReduce classes
    - Begin integration of these classes into previous MapReduce examples

- **Week 6**
    - Full scale integration begins
    - Run first full application tests using the datasets that were chosen
    - Using gathered results begin tweaking the application in order to make a solid errorless template out of it

- **Week 7**
    - Begin documenting results in slides and producing template
    - Continue development of MapReduce application adding functionality and user customization ability

Relation to PIRE Core Research Projects: *This project fits well into "Data Mining Software Tools". It provides a framework for quick and efficient data mining using MapReduce; spefically, Hadoop. A researcher will be able to simply provide a "similarity calculation" function, a "query object", and a dataset then my application template will do the rest. Without any knowledge of MapReduce or the inner workings of Hadoop the researcher will be taking advantage of the MapReduce paradigm.*