PROJECT PROPOSAL

Name : Vaibhav Saharan
Term : Spring 2008
Advisors : Prof. Gail Kaiser, Albert Boulanger
Course Section : E6901 Sec. 14
Date : September 26, 2008

OVERVIEW

This document outlines a proposal along with a tentative schedule for a project with the Center for Computational Learning Systems, Columbia University under the supervision of Prof. Gail Kaiser. This project will be undertaken as a 3-credit course under section 14 of COMS E6901 “Projects in Computer Science”.

The CCLS has an ongoing project for ConEdison, which involves using machine learning algorithms to rank various feeders, transformers, and cables according to their rate of susceptibility to failure. These tools are used by Con Edison employees to better understand the components of a feeder (cables, joints, terminators, & transformers) that are susceptible to failure. This helps them better track down the root causes of failures and to use susceptibility to failure to guide how the system operates during periods of stress and guide replacement of weak components. My work this semester will concentrate on researching available cable data, to find replaced structures in the feeder hierarchies and to build a model to collate joint information after mining available cable data.

PROJECT DESCRIPTION

My project this semester involves two main categories of work:

The first category of work will require writing a program to read in cable failure data, data from other sources like Splice tickets, monthly cable section data dump from Vision and the daily less-detailed cable section data and use that to come up with records of cables and joints that have been replaced both presently and historically. This will mainly involve studying data from the CAJAC database (please refer to Fig. 1), available daily on VanSpliceSQL at ConEdison and copied at CCLS. This will provide regular failure reports and these reports will be used to get an estimate of which cables and joints have been replaced.

The second component will involve coming up with a model to generate a complete set of joints by using various sources of data and collating them into a single joint inference resource. A major source among these will be a table containing the entire set of cable
sections which will be available on a daily basis and the existing joint inference logic as well as past inferred-joint tables, available from Jeopardy and requests to Con Edison’s IR (their IT group) for joint tables. Another resource will be the Splice Ticket data that includes information regarding recent joint splices, which are also available daily on VanSpliceSQL at Con Edison and copied at CCLS. A third resource will be FOWR data which is a work repository for trouble tracking (please refer to Fig. 1 for more information on all of these data sources). These resources will be utilized to collate and mine data from to create a possibly exhaustive set of joint data.

For both the components, many of the mentioned data resources are not yet available on a regular basis at CCLS and there will be coordination work with ConEdison personnel, especially Bill McGarrigle, to understand what data is available where.

![Cable/Splice with Connectors](image)

**Fig. 1 Data sources at ConEdison**

Most of the development work will be carried out in Python. The database used will be Microsoft SQL Server.
**DEVELOPMENT PLATFORM**

I shall use my own machine for most of the development and research work. I will use CS lab machines for any platform specific work. I might also need to use CCLS servers at LDEO including:

- tecac-x2100j.ldeo.columbia.edu (development server)  
  OS: Windows Server 2003

- hertz.ldeo.columbia.edu (database server)  
  OS: Windows Server 2003

- cengdev1 at ConEdison Homenet  
  OS: Windows Server 2003

**TIMELINE**

Sept 26, 2008 to Oct 05, 2008: Analyze the current structure of data sources and get more information about update frequency and access rights. Set up the basic platform for the two components.

Oct 05, 2008 to Nov 25, 2008: Coordinate with ConEdison personnel to set up usage of the data sources that we may not have already. Begin work on the first component and start generating data on changes based on failure reports. Simultaneously, start work on the second component, by creating an algorithm to consolidate data from various sources and mining it, to generate joint information. It should be mentioned that the second project does not depend very much on data sources that we don’t already have so will be something to work on while investigating the best set of data and course for the first program – there is an investigational aspect to the first program that the second does not share.

Nov 25, 2008 to End of Semester: Conclude work on the first component and consolidating the change model. Complete work on second component and consolidate it into a seamless joint inference engine.