Presentation 1 - Action Items

Eric Davis

CIS 895 – MSE Project

- Create a substantial formal specification for a portion of the project
  - **Solution:**
    - Decided to formally specify the Web Crawling portion of the project. This was chosen instead of the Entity Search portion of the project because it will allow for a more substantial formal specification. It will include methods for adding to, deleting from, and searching in the database, as well as contain uniqueness conditions. It will be formally specified in OCL using USE 2.3.1.

- Email the committee with the part of the system that will be specified and the tools used for the specification
  - **Solution:**
    - Committee emailed on 12/07/07 with updated status

- Update the 'Tools Used' section of the SQA Plan when decided
  - **Solution:**
    - Update is unnecessary because the USE tool was already included in the SQA Plan as the anticipated tool for formally specifying a portion of the project.

- Investigate depth limited crawlers as the project is currently breadth first (e.g. Wget, Teleport Crawl)
  - **Solution:**
    - These crawlers provide COTS crawling capability, along with following the Robot Exclusionary Protocol, and the ability to limit crawl depth
    - The decision was made to create a new crawler for this project
      - It allows the developer to learn about web crawling
      - The majority of the code for the crawler has already been developed
      - Adding a depth-limiting crawl capability can be easily implemented
        - A new requirement will be added for this capability
      - Using a COTS crawler may be a nice enhancement for future work for someone who is already experienced with web crawling tasks

- Move Back-links field from Web Crawler tab to Web Search tab
  - **Solution:**
    - The back-links field was moved to the Web Search tab

- Re-look at the effort adjustment factors for the complexity and data fields
  - Is the storage needed for the data exponential or logarithmic? Report back on during the second presentation
  - **Solution:**
    - The storage requirements for the data are linear. Although the amount of data available on the internet grows exponentially, the number of websites being stored is restricted by user input -- therefore the storage requirements are linear based on the number of websites that the user wishes to crawl.
    - Since the required storage can be very high depending on the number of websites to be crawled, the Complexity and the Data Effort Adjustment Factors (EAF) will remain unchanged, as they were both at the high end of the EAF sliding scale.
• Delineate the scope compared to Tao's work for feasibility
  • **Solution:**
    • Differences:
      • Project will be limited to contact information entities
      • Project will use smaller data sets
      • Project will have the ability to run on a single PC rather than the need to run on a large cluster
      • Project will be GUI based
      • Project will not have a complex ranking formula for entity search results, nor will the results be compared to numerous other methods