CIS 895 – MSE Project

KDD-Research Entity Search Tool (KREST)
Presentation 1

Eric Davis
efd3467@ksu.edu
Outline

- Project Overview
- Project Requirements
- Project Schedule
- Cost Estimation
- Software Quality Assurance Plan
- Prototype Demonstration
- Questions / Comments
Project Overview

♦ Goal
  ■ To develop an application which allows web crawling, web searching and entity searching
  ■ To reproduce the results from the entity search work of Kevin Chang and Tao Cheng at UIUC

♦ Motivation
  ■ Improve upon current web searches by making it easier to find contact information
    ● Results up front
    ● No more sifting through results pages for contact info
  ■ By providing self-contained application, searching done on local machine rather than server
Project Overview (cont.)

- Primary Focus: Entity Search
  - Search for contact information
    - Works like a database query from user’s perspective
    - Breaks apart web pages to extract the entities
  - Sample Queries
    - Kansas State University database professor #email
    - Amazon Customer Service #phone
  - Provide contact information + back-links
Project Overview (cont.)

- KREST Environment
  - Update
  - Extract Data
  - Website DB
    - Allows Performance Of
    - Allows Performance Of
    - Allows Performance Of
      - Web Crawler
      - Web Searching
      - Entity Searching
Project Overview (cont.)
Project Overview (cont.)
Project Requirements

- Requirements Broken into 4 sections
  - Application Requirements
    - Loading / Saving Data
    - General System Level Requirements
  - Web Crawling Requirements
  - Web Searching Requirements
  - Entity Searching Requirements
- Each section has its own identifier + numbering system
- Each requirement has an associated build release noted in the Vision Plan
- Screenshot before each section
Application Screenshot

KDD-Research Entity Search Tool

File | Help
--- | ---
Load Dataset | Web Search | Entity Search
Save Results | Exit

Max Sites to Explore: 50
Log File to Use:
Min # of Backlinks: 0

Begin Crawl

Currently Crawling:
Crawled URLs:
Sites in the Queue:
Crawl Progress:
Application Requirements

- **ARI 100** - The program shall provide a GUI for user interaction
- ARI 101 - The application shall be executable in a single step (e.g. without having to perform any setup steps)
- ARI 102 - The application shall have a menu bar that contains at a minimum: a File menu and a Help menu
- **ARI 103** - The application shall allow the user to load a data set of web pages
- ARI 104 - The application shall allow the user to save entity search results
- ARI 105 - The application's Help menu shall contain at a minimum an About menu item
- ARI 106 - The application's menu bar shall contain shortcut keys
- **ARI 107** - The application shall be platform independent
- ARI 108 - The application shall be able to be minimized
- ARI 109 - The application shall be able to be closed without having to perform a Control-C from the command line

Bolded requirements represent Critical Project Requirements
Web Crawl Screenshot
Web Crawling Requirements

- WCRI 100 - The user shall have the ability to perform a web crawl based on a starting website
- WCRI 101 - The user shall be allowed to specify the starting website (if none is specified, http://www.cis.ksu.edu will be used)
- WCRI 102 - The user shall have the ability to specify the number of back-links required for a website to be maintained in the final list
- WCRI 103 - The user shall have the ability to specify a log file in which to save the results of the crawl
- WCRI 104 - The user shall be allowed to specify the maximum number of websites to crawl before stopping
- WCRI 105 - The user shall be allowed to stop the crawl at any time before it finishes
- WCRI 106 - The user shall be notified when the crawl is complete
- WCRI 107 - The user shall be kept apprised of the total number of pages left to crawl
- WCRI 108 - The user shall be kept apprised of the total number of pages crawled
- WCRI 109 - The crawler shall follow the robot exclusionary protocol
- WCRI 110 - The crawler shall use multiple threads to avoid putting too much stress on an individual web host
- WCRI 111 - The user shall have an option to search only within the specified domain

Bolded Requirements represent Critical Project Requirements
Web Search Screenshot
Web Search Requirements

- WSRI 100 - The user shall be allowed to search over previously crawled web pages
- WSRI 101 - The user shall have a box to enter search terms
- WSRI 102 - The user shall be allowed to specify the minimum number of back-links required for a page containing the search term to be considered a match
- WSRI 103 - The URLs that match the search terms shall be sorted in order of number of back-links
- WSRI 104 - The URLs that match the search terms shall be displayed in a scrollable text box

- Bolded requirements represent Critical Project Requirements
Entity Search Screenshot

The screen shows a search tool with the search string "machine learning #email" entered. The search results include:

- **William Hsu**
  - www.cis.ksu.edu/~bhsu
  - www.cis.ksu.edu/people/faculty
  - www.kddresearch.org/Groups/Machine-Learning/
  - bhsu@cis.ksu.edu

- **Doina Caragea**
  - people.cis.ksu.edu/~dcaragea/
  - www.cis.ksu.edu/people/faculty
  - dcaragea@cis.ksu.edu
Entity Search Requirements

- **ESRI 100** - The user shall have the ability to search for entities from previously crawled websites
- **ESRI 101** - The user shall have a box to enter search terms
- **ESRI 102** - There shall entities for at a minimum: email address, phone number, fax number, street address, and zip code
- **ESRI 103** - There shall be an overarching entity that gathers all contact info
- **ESRI 104** - The entity search results shall be ranked based on highest score
- **ESRI 105** - The user shall be allowed to specify search terms in addition to entity terms
- **ESRI 106** - The entities that match the search terms shall be displayed in a scrollable text box

- Bolded requirements represent Critical Project Requirements
Project Schedule

- Key Dates
  - Presentation 1: November 13
  - Presentation 2: February 15
  - Presentation 3: April 25
## Project Schedule (cont.)

<table>
<thead>
<tr>
<th>Task Name</th>
<th>Duration</th>
<th>Start Date</th>
<th>Finish Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Search</td>
<td>51 days</td>
<td>Mon 8/29/07</td>
<td>Tue 11/15/07</td>
</tr>
<tr>
<td>2. Design</td>
<td>6 days</td>
<td>Wed 8/29/07</td>
<td>Thu 9/4/07</td>
</tr>
<tr>
<td>3. Vision Document 1.0</td>
<td>10 days</td>
<td>Mon 9/2/07</td>
<td>Wed 9/19/07</td>
</tr>
<tr>
<td>4. Software Caddy Assurance Plan 1.0</td>
<td>4 days</td>
<td>Wed 9/26/07</td>
<td>Fri 9/28/07</td>
</tr>
<tr>
<td>5. Project Plan 2.0</td>
<td>15 days</td>
<td>Mon 10/1/07</td>
<td>Fri 10/26/07</td>
</tr>
<tr>
<td>6. Website</td>
<td>7 days</td>
<td>Tue 11/6/07</td>
<td>Fri 11/9/07</td>
</tr>
<tr>
<td>7. Data</td>
<td>2 days</td>
<td>Wed 11/7/07</td>
<td>Fri 11/9/07</td>
</tr>
<tr>
<td>8. Final Build</td>
<td>4 days</td>
<td>Mon 11/12/07</td>
<td>Thu 11/15/07</td>
</tr>
<tr>
<td>9. Final Build Cleanup</td>
<td>2 days</td>
<td>Tue 11/13/07</td>
<td>Thu 11/15/07</td>
</tr>
<tr>
<td>10. Edit Search Option</td>
<td>1 day</td>
<td>Wed 11/14/07</td>
<td>Thu 11/15/07</td>
</tr>
<tr>
<td>11. Web Site Search</td>
<td>1 day</td>
<td>Thu 11/15/07</td>
<td>Fri 11/16/07</td>
</tr>
<tr>
<td>12. Project Prototype 1.0</td>
<td>5 days</td>
<td>Mon 11/19/07</td>
<td>Fri 11/23/07</td>
</tr>
<tr>
<td>13. End User Training</td>
<td>1 day</td>
<td>Mon 11/26/07</td>
<td>Fri 11/30/07</td>
</tr>
<tr>
<td>14. Final Build</td>
<td>2 days</td>
<td>Wed 11/28/07</td>
<td>Fri 11/30/07</td>
</tr>
<tr>
<td>15. Update Schedules</td>
<td>7 days</td>
<td>Mon 12/3/07</td>
<td>Fri 12/7/07</td>
</tr>
<tr>
<td>16. Communication</td>
<td>5 days</td>
<td>Wed 12/5/07</td>
<td>Fri 12/7/07</td>
</tr>
<tr>
<td>17. Software Caddy Assurance Plan 2.0</td>
<td>3 days</td>
<td>Thu 12/6/07</td>
<td>Fri 12/7/07</td>
</tr>
<tr>
<td>18. Vision Document 2.0</td>
<td>3 days</td>
<td>Fri 12/7/07</td>
<td>Mon 12/10/07</td>
</tr>
<tr>
<td>19. Project Plan 2.0</td>
<td>3 days</td>
<td>Mon 12/10/07</td>
<td>Thu 12/13/07</td>
</tr>
<tr>
<td>20. Risk Management Plan</td>
<td>1 day</td>
<td>Fri 12/14/07</td>
<td>Sat 12/15/07</td>
</tr>
<tr>
<td>21. Assessment Evaluation</td>
<td>1 day</td>
<td>Sun 12/16/07</td>
<td>Mon 12/17/07</td>
</tr>
<tr>
<td>22. Test Plan 1.0</td>
<td>3 days</td>
<td>Wed 12/19/07</td>
<td>Fri 12/21/07</td>
</tr>
<tr>
<td>23. Assessment Evaluation 1.0</td>
<td>1 day</td>
<td>Fri 12/21/07</td>
<td>Sat 12/22/07</td>
</tr>
<tr>
<td>24. Technical Evaluation</td>
<td>1 day</td>
<td>Sun 12/23/07</td>
<td>Mon 12/24/07</td>
</tr>
<tr>
<td>25. Technical Evaluation 1.0</td>
<td>1 day</td>
<td>Sat 12/29/07</td>
<td>Sun 12/30/07</td>
</tr>
<tr>
<td>26. Final Build</td>
<td>2 days</td>
<td>Wed 1/2/08</td>
<td>Thu 1/3/08</td>
</tr>
<tr>
<td>27. Interface Manual - Shortcuts</td>
<td>2 days</td>
<td>Fri 1/4/08</td>
<td>Sat 1/5/08</td>
</tr>
<tr>
<td>28. Addendum - Backends for Webtop using</td>
<td>2 days</td>
<td>Sun 1/6/08</td>
<td>Mon 1/7/08</td>
</tr>
<tr>
<td>29. Addendum - Enhancing 1015Apr</td>
<td>2 days</td>
<td>Wed 1/9/08</td>
<td>Thu 1/10/08</td>
</tr>
<tr>
<td>30. Functional Testing</td>
<td>1 day</td>
<td>Fri 12/21/07</td>
<td>Sat 12/22/07</td>
</tr>
<tr>
<td>31. Project Strategy 2.0</td>
<td>1 day</td>
<td>Thu 12/27/07</td>
<td>Fri 12/28/07</td>
</tr>
<tr>
<td>32. Website 2</td>
<td>1 day</td>
<td>Fri 12/28/07</td>
<td>Sat 12/29/07</td>
</tr>
<tr>
<td>33. Phase Three - Production Phase</td>
<td>20 days</td>
<td>Mon 1/14/08</td>
<td>Fri 1/25/08</td>
</tr>
<tr>
<td>34. Audit/5</td>
<td>7 days</td>
<td>Mon 1/21/08</td>
<td>Fri 1/25/08</td>
</tr>
<tr>
<td>35. Communication</td>
<td>5 days</td>
<td>Mon 1/28/08</td>
<td>Fri 1/31/08</td>
</tr>
<tr>
<td>36. Software Caddy Assurance Plan 3.0</td>
<td>3 days</td>
<td>Mon 1/31/08</td>
<td>Wed 2/6/08</td>
</tr>
<tr>
<td>37. Vision Document 3.0</td>
<td>3 days</td>
<td>Wed 2/6/08</td>
<td>Mon 2/11/08</td>
</tr>
<tr>
<td>38. Project Plan 3.0</td>
<td>3 days</td>
<td>Mon 2/11/08</td>
<td>Thu 2/14/08</td>
</tr>
<tr>
<td>39. Component Design</td>
<td>1 day</td>
<td>Thu 2/14/08</td>
<td>Fri 2/15/08</td>
</tr>
<tr>
<td>40. Assessment Evaluation</td>
<td>1 day</td>
<td>Fri 2/15/08</td>
<td>Sat 2/16/08</td>
</tr>
<tr>
<td>41. User Manual</td>
<td>2 days</td>
<td>Sat 2/16/08</td>
<td>Mon 2/18/08</td>
</tr>
<tr>
<td>42. Internal Technical Report</td>
<td>3 days</td>
<td>Mon 2/18/08</td>
<td>Thu 2/21/08</td>
</tr>
<tr>
<td>43. Testing Guide</td>
<td>2 days</td>
<td>Thu 2/21/08</td>
<td>Fri 2/22/08</td>
</tr>
<tr>
<td>44. Project Evaluation</td>
<td>1 day</td>
<td>Fri 2/22/08</td>
<td>Sat 2/23/08</td>
</tr>
<tr>
<td>45. Final Build</td>
<td>1 day</td>
<td>Sat 2/23/08</td>
<td>Sun 2/24/08</td>
</tr>
<tr>
<td>46. Customer Audits</td>
<td>2 days</td>
<td>Sun 2/24/08</td>
<td>Mon 2/25/08</td>
</tr>
<tr>
<td>47. Website maintenance</td>
<td>4 days</td>
<td>Mon 2/25/08</td>
<td>Thu 2/28/08</td>
</tr>
<tr>
<td>48. Website Maintenance</td>
<td>4 days</td>
<td>Thu 2/28/08</td>
<td>Mon 3/3/08</td>
</tr>
<tr>
<td>49. Website Maintenance</td>
<td>3 days</td>
<td>Mon 3/3/08</td>
<td>Wed 3/5/08</td>
</tr>
<tr>
<td>50. Website Maintenance</td>
<td>3 days</td>
<td>Wed 3/5/08</td>
<td>Fri 3/7/08</td>
</tr>
<tr>
<td>51. Severe Contingency</td>
<td>4 days</td>
<td>Fri 3/7/08</td>
<td>Mon 3/10/08</td>
</tr>
<tr>
<td>52. Severe Contingency</td>
<td>4 days</td>
<td>Mon 3/10/08</td>
<td>Thu 3/13/08</td>
</tr>
<tr>
<td>53. Severe Contingency</td>
<td>4 days</td>
<td>Thu 3/13/08</td>
<td>Fri 3/14/08</td>
</tr>
<tr>
<td>54. Severe Contingency</td>
<td>4 days</td>
<td>Fri 3/14/08</td>
<td>Mon 3/17/08</td>
</tr>
<tr>
<td>55. Severe Contingency</td>
<td>4 days</td>
<td>Mon 3/17/08</td>
<td>Thu 3/20/08</td>
</tr>
<tr>
<td>56. Severe Contingency</td>
<td>4 days</td>
<td>Thu 3/20/08</td>
<td>Fri 3/21/08</td>
</tr>
</tbody>
</table>

### Notes
- The schedule is subject to change based on project progress and priorities.
- Key milestones are marked in bold.
- The project is expected to be completed by the end of the specified period.
Cost Estimation Formulas

- Intermediate COCOMO
  - Important Formulas:
    - Effort = 3.2 * EAF * (KLOC)^1.05
      - EAF represents Effort Adjustment Factor
      - KLOC represents Source Lines of Code (in thousands)
    - Time (in months) = 2.5 * (Effort)^0.38
## Effort Adjustment Factors

<table>
<thead>
<tr>
<th>Identifier</th>
<th>Classification</th>
<th>Value</th>
<th>Reasoning</th>
</tr>
</thead>
<tbody>
<tr>
<td>RELY</td>
<td>Low</td>
<td>0.88</td>
<td>Project is not safety critical, and does not have to be completely reliable</td>
</tr>
<tr>
<td>DATA</td>
<td>High</td>
<td>1.08</td>
<td>A large # of web pages are needed in order to perform a thorough search</td>
</tr>
<tr>
<td>CPLX</td>
<td>Nominal</td>
<td>1.00</td>
<td>Web crawling, Web Search, and Entity Search are not overly complicated</td>
</tr>
<tr>
<td>TIME</td>
<td>Nominal</td>
<td>1.00</td>
<td>Response time is important yet not overly critical</td>
</tr>
<tr>
<td>STOR</td>
<td>Very High</td>
<td>1.21</td>
<td>Crawling and searching will require a lot of memory usage</td>
</tr>
<tr>
<td>VIRT</td>
<td>Low</td>
<td>0.87</td>
<td>Low complexity of the hardware and software</td>
</tr>
<tr>
<td>TURN</td>
<td>Low</td>
<td>0.87</td>
<td>Since this is a single developer project, the turnaround time is low</td>
</tr>
<tr>
<td>ACAP</td>
<td>High</td>
<td>0.86</td>
<td>Developer has 4+ years experience in software engineering</td>
</tr>
<tr>
<td>AEXP</td>
<td>High</td>
<td>0.91</td>
<td>Developer has 3+ years experience in applications development</td>
</tr>
<tr>
<td>PCAP</td>
<td>High</td>
<td>0.86</td>
<td>Developer has applicable experience</td>
</tr>
<tr>
<td>VEXP</td>
<td>Nominal</td>
<td>1.00</td>
<td>Developer has 2+ years experience developing for Java virtual machine</td>
</tr>
<tr>
<td>LEXP</td>
<td>High</td>
<td>0.95</td>
<td>Developer has 2+ years experience developing using Java</td>
</tr>
<tr>
<td>TOOL</td>
<td>Nominal</td>
<td>1.00</td>
<td>Moderate experience with tools being used</td>
</tr>
<tr>
<td>MODP</td>
<td>Very High</td>
<td>0.83</td>
<td>Developer has 4+ years exp. using modern software engineering practices</td>
</tr>
<tr>
<td>SCED</td>
<td>Nominal</td>
<td>1.00</td>
<td>Project has a tight schedule, but some slippage is allowable</td>
</tr>
</tbody>
</table>
Cost Estimation Calculations

- KLOC estimated at: 2
  - Based on other available web crawlers + searchers
- EAF = 0.95
- Effort = 6.29 Staff-months
- Time = 5.04 Chronological months
  - Result: Project should be able to be accomplished within 2 semesters.
Software Quality Assurance Plan

- References
  - Vision Plan
  - Project Plan
- Supervisory Committee
  - Dr. Scott DeLoach
  - Dr. David Gustafson
  - Dr. William Hsu
- Major Professor
  - Dr. William Hsu
- Developer
  - Eric Davis
- Formal Technical Inspectors
  - TBD
Software Quality Assurance Plan (cont).

- Documentation
  - A listing of the required documentation is available at:
    - [http://mse.cis.ksu.edu/online/mse-portfolio.htm](http://mse.cis.ksu.edu/online/mse-portfolio.htm)

- Project Documentation will be available at:
  - [http://www.cis.ksu.edu/~efd3467](http://www.cis.ksu.edu/~efd3467)
Software Quality Assurance Plan (cont).

- **Standards, Practices, Conventions & Metrics**
  - Documentation – IEEE standards will be followed for all applicable documentation
  - Coding – Java naming conventions + Javadoc will be used
  - Metrics – COCOMO will be used to measure project effort

- **Reviews & Audits**
  - Supervisory committee will review all documentation at each milestone
  - TBD Formal Technical Inspectors will review the architecture before the second presentation
Software Quality Assurance Plan (cont).

- **Testing**
  - Defined in Software Test Plan
    - Will be available by Presentation 2

- **Problem Reporting**
  - Issues will be tracked in a spreadsheet
  - All issues will be reported to Major Professor
Software Quality Assurance Plan (cont).

- **Tools, Technologies, & Methodologies**
  - Eclipse IDE – for software development
  - Eclipse FatJar – for building executable JAR files
  - Eclipse Jigloo Plugin – for GUI development
  - Microsoft Word – for documentation development
  - Microsoft Excel – for risk and problem report tracking and time logs
  - Microsoft Powerpoint – for project presentation creation
  - Adobe Acrobat – for document conversion to PDF
  - Microsoft Project – for project planning
  - Microsoft Visio – for software design development
  - USE 2.3.1 – for developing formal specifications
Software Quality Assurance Plan (cont).

- Code & Media Control
  - CVS will be used for source code control
  - Repository is available at:
  - Change logs will be maintained for all documents
    - Versions will be maintained on the developer’s computer
Prototype Demonstration
Phase 2 Deliverables

- Vision Plan 2.0
- Project Plan 2.0
- Architectural Design Document
- Software Test Plan 1.0
- Technical Inspection List
- Presentation 2
- Prototype 2.0 Source Code
Current Obstacles / Questions

- Technical Inspectors
  - Two are still needed

- Masters Degree Final Examination?
Questions / Comments