Vision Plan

For KDD-Research Entity Search Tool (KREST)

Version 2.1

Submitted in partial fulfillment of the Masters of Software Engineering degree.

Eric Davis
CIS 895 – MSE Project
Department of Computing and Information Sciences
Kansas State University
## Change Log

<table>
<thead>
<tr>
<th>Version #</th>
<th>Changed By</th>
<th>Release Date</th>
<th>Change Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Version 1.0</td>
<td>Eric Davis</td>
<td>11/05/07</td>
<td>Initial Release</td>
</tr>
<tr>
<td>Version 1.1</td>
<td>Eric Davis</td>
<td>11/08/07</td>
<td>Updating Section 1.2 and changing references to APA format</td>
</tr>
<tr>
<td>Version 2.0</td>
<td>Eric Davis</td>
<td>12/12/07</td>
<td>Updating with comments from Presentation 1</td>
</tr>
<tr>
<td>Version 2.1</td>
<td>Eric Davis</td>
<td>03/08/08</td>
<td>Removing domain specific web crawling requirement</td>
</tr>
</tbody>
</table>
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1</td>
<td>Project Overview</td>
<td>2</td>
</tr>
<tr>
<td>Figure 2</td>
<td>Project Block Diagram</td>
<td>3</td>
</tr>
<tr>
<td>Figure 3</td>
<td>KREST Data Flow Diagram</td>
<td>4</td>
</tr>
<tr>
<td>Figure 4</td>
<td>System Use Case</td>
<td>5</td>
</tr>
</tbody>
</table>
# Table of Contents

Change Log ........................................................................................................................ ii
List of Figures ......................................................................................................................... iii

1. Introduction ................................................................................................................. 1
   1.1 Motivation ......................................................................................................... 1
   1.2 KDD-Research Entity Search Tool (KREST) .................................................. 1
   1.3 Terms & Definitions ......................................................................................... 1
   1.4 References ......................................................................................................... 2

2. Project Overview .......................................................................................................... 2
   2.1 Introduction ....................................................................................................... 3
   2.2 Project Goal ...................................................................................................... 4
   2.3 Project Purpose ................................................................................................. 4

3. Project Requirements .................................................................................................. 5
   3.1 Application Requirements ................................................................................ 5
      3.1.1 ARI 100 [Critical Requirement] ................................................................. 6
      3.1.2 ARI 101 ....................................................................................................... 6
      3.1.3 ARI 102 ....................................................................................................... 6
      3.1.4 ARI 103 [Critical Requirement] ................................................................. 6
      3.1.5 ARI 104 ....................................................................................................... 6
      3.1.6 ARI 105 ....................................................................................................... 6
      3.1.7 ARI 106 ....................................................................................................... 6
      3.1.8 ARI 107 [Critical Requirement] ................................................................. 6
      3.1.9 ARI 108 ....................................................................................................... 6
      3.1.10 ARI 109 ..................................................................................................... 7
   3.2 Web Crawler Requirements .............................................................................. 7
      3.2.1 WCRI 100 [Critical Requirement] .............................................................. 7
      3.2.2 WCRI 101 [Critical Requirement] .............................................................. 7
      3.2.3 WCRI 102 ................................................................................................... 7
      3.2.4 WCRI 103 ................................................................................................... 7
      3.2.5 WCRI 104 [Critical Requirement] .............................................................. 7
      3.2.6 WCRI 105 ................................................................................................... 8
      3.2.7 WCRI 106 ................................................................................................... 8
      3.2.8 WCRI 107 ................................................................................................... 8
      3.2.9 WCRI 108 ................................................................................................... 8
      3.2.10 WCRI 109 [Critical Requirement] ............................................................. 8
      3.2.11 WCRI 110 [Critical Requirement] ............................................................. 8
   3.3 Web Search Requirements ................................................................................ 8
      3.3.1 WSRI 100 [Critical Requirement] .............................................................. 8
      3.3.2 WSRI 101 [Critical Requirement] .............................................................. 9
      3.3.3 WSRI 102 ................................................................................................... 9
      3.3.4 WSRI 103 ................................................................................................... 9
      3.3.5 WSRI 104 [Critical Requirement] ............................................................. 9
   3.4 Entity Search Requirements .............................................................................. 9
      3.4.1 ESRI 100 [Critical Requirement] .............................................................. 9
      3.4.2 ESRI 101 [Critical Requirement] .............................................................. 9
3.4.3  ESRI 102 [Critical Requirement]................................................................. 9
3.4.4  ESRI 103 .................................................................................................... 10
3.4.5  ESRI 104 ................................................................................................... 10
3.4.6  ESRI 105 [Critical Requirement]............................................................... 10
3.4.7  ESRI 106 [Critical Requirement]............................................................... 10
4.  Assumptions...................................................................................................... 10
5.  Constraints ...................................................................................................... 10
6.  Environment................................................................................................. 10
1. Introduction

1.1 Motivation
The motivation for this project is to improve upon the current state of web searching. Searching for contact information in the web is a tedious task in the current state: it involves trying to determine a proper search string, followed by wading through matching pages looking for the contact information desired. The goal of this project is to improve upon this process by providing the contact information after one search, without requiring the user to wade through pages that match the search string.

1.2 KDD-Research Entity Search Tool (KREST)
The KDD-Research Entity Search Tool (KREST) is the answer to the search related problems mentioned above. It breaks apart web pages into entities, such as email addresses, phone numbers, and fax numbers. Specific entity results are returned to the user, rather than the old way of returning page matches. KREST also allows the user to perform a web crawl from a given starting webpage, and to perform a traditional web search in addition to performing an entity search.

Entity search works by allowing the user to specify what specific type of information they are looking for. To the end user, using entity search will seem somewhat like using a database to query for information -- just the information requested will be returned, without any additional filler. Rather than being forced to search for a general term like “Amazon Customer Service”, entity search will allow the user to specify that they are looking for the Amazon Customer Service phone number by entering a search term like “Amazon Customer Service #phone”. Alternatively, if the user was looking for the email address of the professors at Kansas State University that teach database courses, they could search for that specifically by a search term such as “Kansas State University professor database #email”. Upon receiving a search term, the entity search tool will look for the pages that match the search text. Those pages that match the search text will then be broken to extract the requested entities (if they exist on those pages). The entities that match will then be returned to the user with links to the pages that contained the information in case the user wishes to verify the information.

1.3 Terms & Definitions
Actor – For UML purposes, the actor is the end user of the system.
Entity – A specific piece of information, such as an email address or a phone number.
Knowledge Discovery in Databases (KDD) – A group headed by Dr. William Hsu whose primary focus is data-mining.
Sequence Diagram – A graphical design used to display the order in which objects interact during a certain period.
Unified Modeling Language (UML) – A standard notation used to describe real-world objects.

Use Case Diagram – A behavioral diagram defined by UML. It provides a graphical depiction of system functionality in terms of actors.

1.4 References


2. Project Overview

The Project Overview section provides information about the structure and goals of the KREST project.

![Figure 1: Project Overview](Image)
2.1 Introduction

Figure 1 provides a high level overview of what the KREST project is working to achieve. It will allow the user to perform web crawls, web searches, and entity searches all within the same tool. It will be a self-contained application that works separately from the user’s normal Internet browser. The KREST environment will update and extract data from a database that stores previously crawled web pages.

Figure 2: Project Block Diagram

Figure 2 provides a block diagram of how the KREST project will operate. The user will interact with the KREST tool within the KREST environment. The KREST environment makes calls to the Application Level where the Web Crawler Service, the Web Search Service or the Entity Search Service perform the work. Each of these services makes use of the Website Database in the Storage
Level. All of the work being performed is done on the JAVA Virtual Machine, which in turn runs on the user’s actual system hardware.

![KREST Data Flow Diagram](image)

**Figure 3: KREST Data Flow Diagram**

Figure 3 provides a view of how data will be used throughout the program, especially for entity searching. The database will contain web pages, which are linked to for specific entity instantiations.

### 2.2 Project Goal

The goal of the KREST project is to create an application that provides the ability to perform entity searches on either previously loaded data or crawled web pages. The project should be able to reproduce the findings by [2], which is searching for contact information based on a publicly available dataset of web pages.

The results will have to be cut in size in comparison to [2] because the original dataset is over 1 TB in size. It will also be different from [2] in that the entity searches will be limited to contact information, and the results will be displayed as they are found, and not ranked according to a complex ranking formula. In addition, KREST will be GUI based, and will be capable of running on a single PC, both things which are not available in [2].

### 2.3 Project Purpose

The purpose of the KREST project is to provide a tool that allows enhanced web searching by way of entity search. It is also to provide a standalone application that will speed up searches on the client end. The developed application will act as a platform for future KDD students to perform entity search testing, and provide a good base for future entity search enhancements.
3. Project Requirements
The Project Requirements section will detail all of the requirements for the KREST project. Each requirement will be discussed in detail, as well as the associated requirement number, and the planned release that will fulfill the requirement (i.e. Demo 1, Demo 2, or Final Release). All of the project’s critical requirements will be noted.

Figure 4: System Use Case
The requirements are broken out into four distinct sections based on the Use Case diagram found in Figure 4: Application Requirements, Web Crawler Requirements, Web Search Requirements, and Entity Search Requirements. This makes it easier to track the requirements between different parts of the application, and also makes it easier to refine and add requirements as the project progresses.

3.1 Application Requirements
This section details all of the requirements related to the main application that are not specific to the web crawler, the web search, or the entity search pieces. The requirements are numbered ARI 1XX, where ARI stands for Application Requirement Item.
3.1.1 ARI 100 [Critical Requirement]
The program shall provide a GUI for user interaction. This is a critical requirement because the usefulness of the system would be extremely limited if done in a command line format.
- **Build Release Applicability:** Demo 1, Demo 2, Final Release

3.1.2 ARI 101
The application shall be executable in a single step (e.g. without having to perform any setup steps).
- **Build Release Applicability:** Demo 1, Demo 2, Final Release

3.1.3 ARI 102
The application shall have a menu bar that contains at a minimum: a File menu and a Help menu.
- **Build Release Applicability:** Demo 1, Demo 2, Final Release

3.1.4 ARI 103 [Critical Requirement]
The application shall allow the user to load a data set of web pages. This is a critical requirement because in order to reproduce the findings of [2], the same data set needs to be used.
- **Build Release Applicability:** Final Release

3.1.5 ARI 104
The application shall allow the user to save entity search results.
- **Build Release Applicability:** Final Release

3.1.6 ARI 105
The application's Help menu shall contain at a minimum an About menu item
- **Build Release Applicability:** Demo 2, Final Release

3.1.7 ARI 106
The application's menu bar shall contain shortcut keys.
- **Build Release Applicability:** Demo 2, Final Release

3.1.8 ARI 107 [Critical Requirement]
The application shall be platform independent. This is a critical requirement because while the application is being developed using Windows, the goal is to also allow it to be used on both Linux and Unix as well.
- **Build Release Applicability:** Demo 1, Demo 2, Final Release

3.1.9 ARI 108
The application shall be able to be minimized.
- **Build Release Applicability:** Demo 1, Demo 2, Final Release
3.1.10 **ARI 109**
The application shall be able to be closed without having to perform a Control-C from the command line.

- **Build Release Applicability:** Demo 1, Demo 2, Final Release

3.2 **Web Crawler Requirements**
This section details all of the requirements related to the web crawling portion of the project. The requirements are numbered WCRI 1XX, where WCRI stands for Web Crawling Requirement Item.

### 3.2.1 WCRI 100 [Critical Requirement]
The user shall have the ability to perform a web crawl based on a starting website. This is a critical requirement because without the web crawling portion of the project, the usefulness of the project is extremely limited (it would be limited to only using user loaded data sets). By allowing user specified web crawls to be performed, the user can tailor the search to their needs.

- **Build Release Applicability:** Demo 1, Demo 2, Final Release

### 3.2.2 WCRI 101 [Critical Requirement]
The user shall be allowed to specify the starting website (if none is specified, http://www.cis.ksu.edu will be used). This is a critical requirement because allowing the user to specify the start point to crawl from allows a good web crawl to take place. Without allowing the user to specify the start point, there would not be any usefulness to the web crawler.

- **Build Release Applicability:** Demo 1, Demo 2, Final Release

### 3.2.3 WCRI 102
The user shall have the ability to specify the maximum depth of the web crawl.

- **Build Release Applicability:** Demo 2, Final Release

### 3.2.4 WCRI 103
The user shall have the ability to specify a log file in which to save the results of the crawl.

- **Build Release Applicability:** Demo 2, Final Release

### 3.2.5 WCRI 104 [Critical Requirement]
The user shall be allowed to specify the maximum number of websites to crawl before stopping. This is a critical requirement because without allowing the user to specify how many websites to search, it would have to be bounded by the application, which is not a good solution. By allowing the user to specify the maximum number of websites, it allows much better control over the web crawl.

- **Build Release Applicability:** Demo 1, Demo 2, Final Release
3.2.6 WCRI 105
The user shall be allowed to stop the crawl at any time before it finishes.
- **Build Release Applicability**: Demo 2, Final Release

3.2.7 WCRI 106
The user shall be notified when the crawl is complete.
- **Build Release Applicability**: Demo 2, Final Release

3.2.8 WCRI 107
The user shall be kept apprised of the total number of pages left to crawl.
- **Build Release Applicability**: Demo 1, Demo 2, Final Release

3.2.9 WCRI 108
The user shall be apprised of the total number of pages crawled.
- **Build Release Applicability**: Demo 1, Demo 2, Final Release

3.2.10 WCRI 109 [Critical Requirement]
The crawler shall follow the robot exclusionary protocol. This is a critical requirement because it keeps the web crawler from crawling pages that are not intended to be crawled. If a robot protocol is not specified for a domain, all pages will considered to be able to be crawled.
- **Build Release Applicability**: Demo 2, Final Release

3.2.11 WCRI 110 [Critical Requirement]
The crawler shall use multiple threads to avoid putting too much stress on an individual web host. This is a critical requirement because it will help prevent overloading a web host with numerous requests one right after another.
- **Build Release Applicability**: Demo 2, Final Release

3.3 Web Search Requirements
This section details all of the requirements related to the web search portion of the project. The requirements are numbered WSRI 1XX, where WSRI stands for Web Search Requirement Item.

3.3.1 WSRI 100 [Critical Requirement]
The user shall be allowed to search over previously crawled web pages. This is a critical requirement because it is important to provide a web search functionality similar to what is available on the web for comparison to the entity search portion of the project.
- **Build Release Applicability**: Demo 1, Demo 2, Final Release
3.3.2 WSRI 101 [Critical Requirement]
The user shall have a box to enter search terms. This is a critical requirement because without a box for user to enter search terms, it would not be possible to provide an entity search capability.

- **Build Release Applicability**: Demo 1, Demo 2, Final Release

3.3.3 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.

- **Build Release Applicability**: Demo 2, Final Release

3.3.4 WSRI 103
The URLs that match the search terms shall be sorted in order of number of back-links.

- **Build Release Applicability**: Demo 2, Final Release

3.3.5 WSRI 104 [Critical Requirement]
The URLs that match the search terms shall be displayed in a scrollable text box. This is a critical requirement because all of the results need to be shown to the user in a useful fashion.

- **Build Release Applicability**: Demo 1, Demo 2, Final Release

3.4 Entity Search Requirements
This section details all of the requirements related to the entity search portion of the project. The requirements are numbered ESRI 1XX, where ESRI stands for Entity Search Requirement Item.

3.4.1 ESRI 100 [Critical Requirement]
The user shall have the ability to search for entities from previously crawled websites. This is a critical requirement because providing an entity search capability is the primary thrust of the project.

- **Build Release Applicability**: Demo 2, Final Release

3.4.2 ESRI 101 [Critical Requirement]
The user shall have a box to enter search terms. This is a critical requirement because without a box for user to enter search terms, it would not be possible to provide an entity search capability.

- **Build Release Applicability**: Demo 2, Final Release

3.4.3 ESRI 102 [Critical Requirement]
There shall entities for at a minimum: email address, phone number, fax number, street address, and zip code. This is a critical project requirement, because this is the minimum amount of information required to reproduce the findings of [2].
• Build Release Applicability: Demo 2, Final Release

3.4.4 ESRI 103
There shall be an overarching entity that gathers all contact info.
• Build Release Applicability: Demo 2, Final Release

3.4.5 ESRI 104
The entity search results shall be ranked based on highest score.
• Build Release Applicability: Final Release

3.4.6 ESRI 105 [Critical Requirement]
The user shall be allowed to specify search terms in addition to entity terms. This is a critical requirement because without being allowed to specify additional search terms, it would not be possible to return any interesting results.
• Build Release Applicability: Demo 2, Final Release

3.4.7 ESRI 106 [Critical Requirement]
The entities that match the search terms shall be displayed in a scrollable text box. This is a critical requirement because all of the results need to be shown to the user in a useful fashion.
• Build Release Applicability: Demo 2, Final Release

4. Assumptions
• Java Runtime Environment 1.3.1 or later will be installed on the computer running the application.
• In order to run a search, the user will have an active Internet connection.
• In order to perform a Web Crawl in a reasonable amount of time, the user will have a high-speed Internet connection (DSL or better).
• The user will need a minimum of 512 MB of memory.
• The user will have a computer with a minimum speed of 1.6 GHz.

5. Constraints
• Java will be used for the web crawling. While it will not be as efficient as using other languages, there is much web functionality defined in the JDK, making it easier to write the web crawling.
• Entity Search is being limited to searching for contact info entities. An excellent future enhancement would be to add other entity types.

6. Environment
• Eclipse 3.3.0 will be used as the IDE.
• Java version JDK 1.5 will be used.
• The Jigloo plugin for Eclipse will be used for GUI development.