User Manual

For KDD-Research Entity Search Tool (KREST)

Version 1.0

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

Version #	Changed By	Release Date	Change Description
Version 1.0	Eric Davis	03/05/08	Initial Release

Table of Contents

1. Introduction

This document describes how to setup and run the KDD-Research Entity Search Tool (KREST). It will explain how to run web crawls, web searches, and entity searches, as well as detailing how to load in available data.

2. Application Set-up

This section details what things are necessary in order to run KREST.

2.1 Required Software

- Java Runtime Environment 1.3.1 or later

2.2 Recommended Hardware

- Minimum recommended processor speed: 1.6 GHz
- Minimum recommended RAM: 512 MB
- Minimum recommended internet connection: DSL or better

2.3 Required Files

- KREST.jar – This jar file contains everything necessary to run KREST. If you desire to see or make modifications to the source code, it is available in KREST-Source-final.zip. Simply download the source, make any modifications deemed necessary, and rebuild the project. The FatJar plugin was used with eclipse to package everything necessary into the executable jar file.

2.4 Recommended Files

- WebBase Datasets – These can be created from WebBase at: http://dbpubs.stanford.edu:8091/~testbed/doc2/WebBase/. They represent previously crawled pages. If you want to load in a large section of crawled pages for web or entity searching, you should consider downloading datasets from there. Instructions for how to download datasets are available on the WebBase website.

3. KREST

3.1 Running KREST

- Double click on the KREST.jar executable Jar file to start up the application. You should see a screen like the one below.

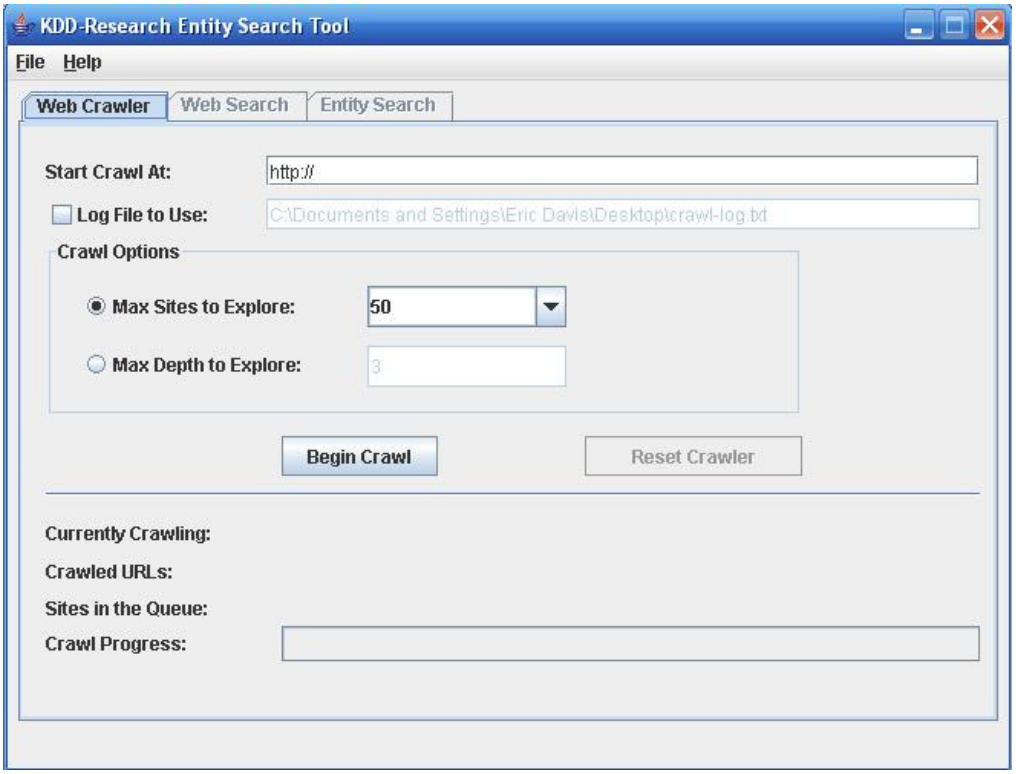

Figure 1: Opening KREST Screen

3.2 Performing a Web Crawl

So you want to perform a web crawl. Before you can do that though, there are several decisions that you need to make:

- Where do you want to start the web crawl at
- Do you want to perform a breadth-first crawl? If so, how many pages do you want to explore?
- Or would you rather perform a depth-limited crawl? If so, how many levels deep would you like to explore?

3.2.1 Breadth-First Crawling

This is the type of crawling where you limit the scope of the web crawl by the number of websites that you want to explore. First, enter the website that you would like to begin exploring at. After that, make sure that the 'Max Sites to Explore' circle is selected, and enter the maximum number of websites that you want to have explored. There is a drop down box containing different amounts, or you can enter a specific number.

It is important to note that if the crawler runs out of web pages to explore before it reaches your maximum number of sites to explore, it will stop crawling. (However, it is extremely rare for this to happen.).

Next, once you are satisfied with the start page and the maximum number of sites to explore, press the 'Begin Crawl' button. You should see the fields at the bottom of the KREST form start updating with the progress bar moving to tell you how much progress has been made in your web search. When the web crawl is complete a box will pop up telling you that the crawl has completed.

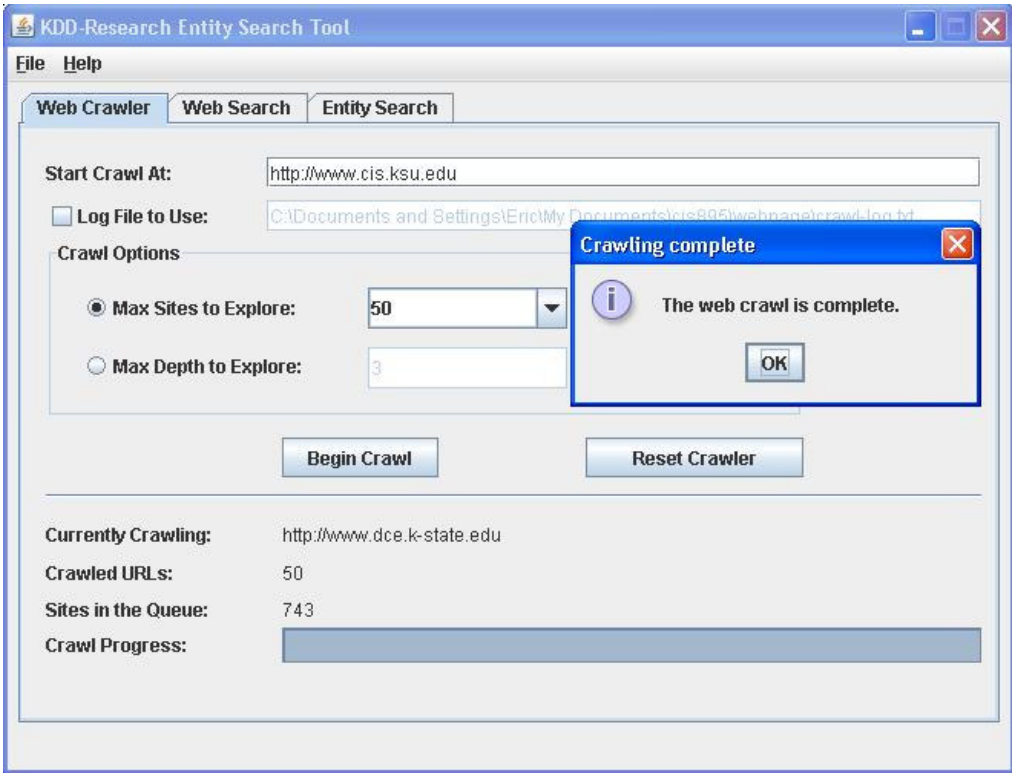

Figure 2: Completed Breadth-First Web Crawl

3.2.2 Depth-First Crawling

This is the type of crawling where you limit the scope of the web crawl by the depth of the websites beyond the start page that you want to explore. First, enter the website that you would like to begin exploring at. After that, make sure that the 'Max Depth to Explore' circle is selected, and enter the maximum depth of websites that you want to have explored. The default depth of 3 can be modified, but keep in mind that increasing it too much can leave the crawler going for a long time!

It is important to note that if the crawler runs out of web pages to explore before it reaches your maximum depth to explore, it will stop crawling. (However, it is extremely rare for this to happen.).

Next, once you are satisfied with the start page and the maximum depth to explore, press the 'Begin Crawl' button. You should see the fields at the bottom of the KREST form start updating with the progress bar moving to tell you how much progress has been made in your web search. When the web crawl is complete the progress will stop moving forward.

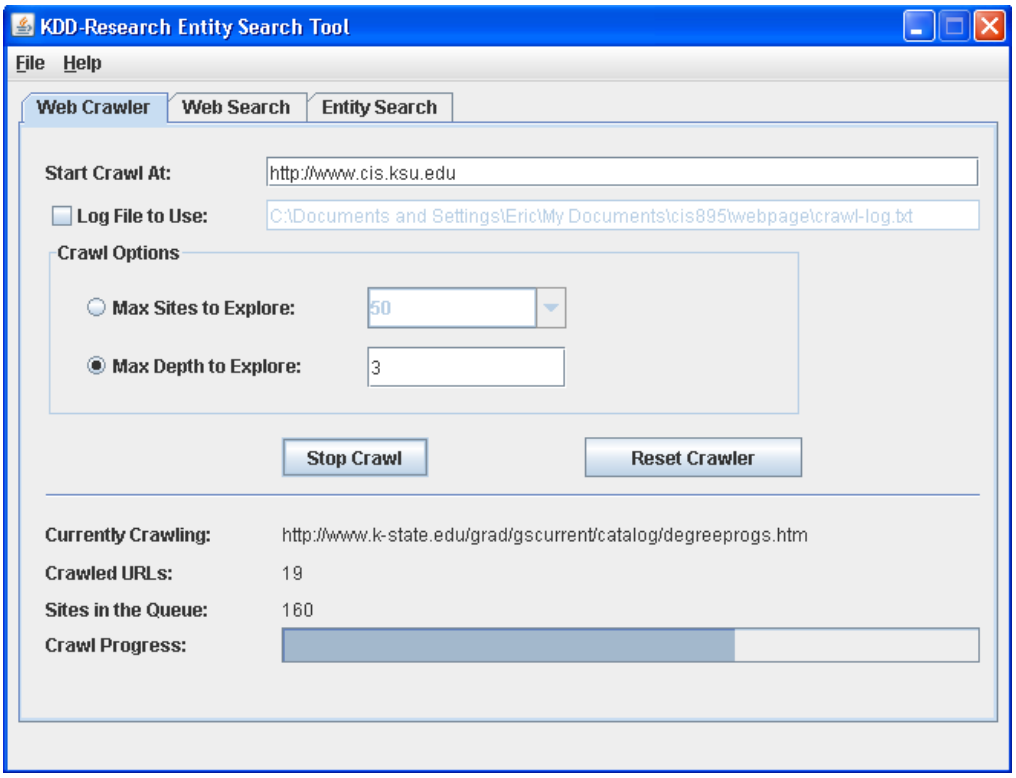

Figure 3: Depth First Crawl in Progress

3.2.3 Saving Web Crawl Information

If you want to save the information about the web crawl, click the box next to the "Log File to Use:" field. You should see the field become editable. Either enter a new file name, or use the one provided. When this box is selected, and the 'Begin Crawl' button is pressed, all information about the web crawl will be written out the file.

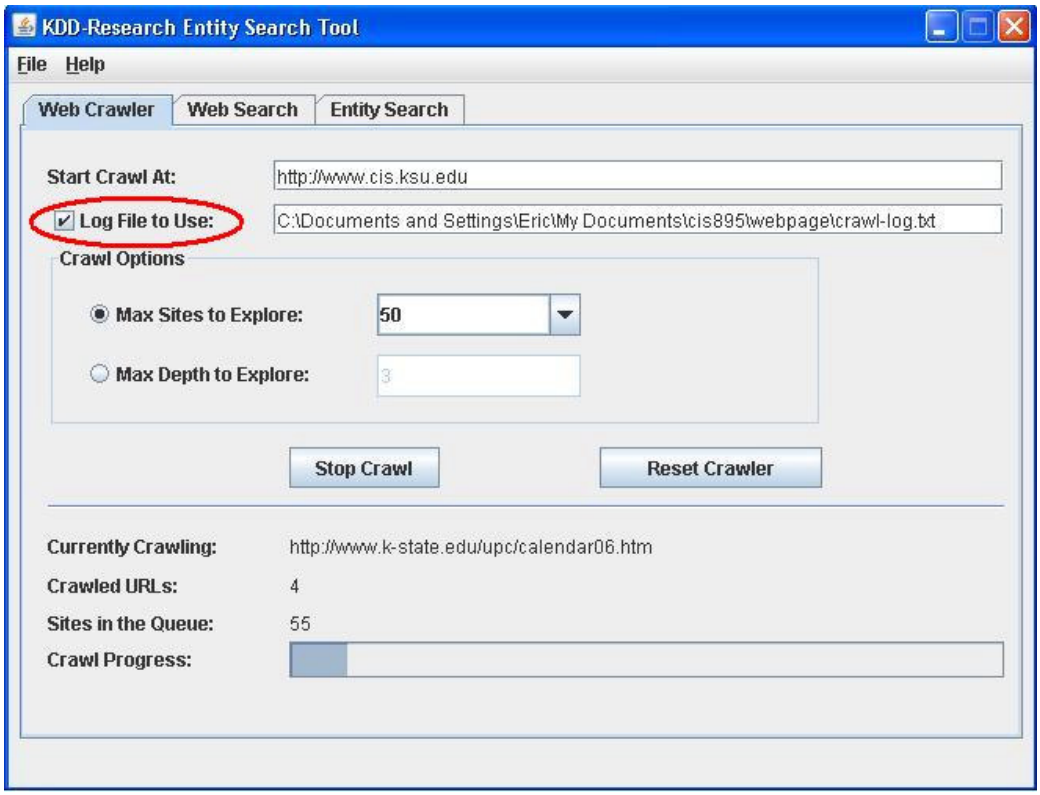

Figure 4: Saving a Web Crawl

3.2.4 Stopping a Web Crawl

Did you make a mistake in the page that you wanted to start crawling from? Is the crawl taking too long, and you just want it to end? Don't worry; you have the ability to stop the web crawl at any point. Once you've started a web crawl, notice that the 'Begin Crawl' button has changed to a 'Stop Crawl' button. Simply press the 'Stop Crawl' button at any point during a web crawl, and the crawl will immediately stop with the status fields being reset to defaults. You may also be interested in the ability to clear crawled pages out of the database, which is detailed in the next section.

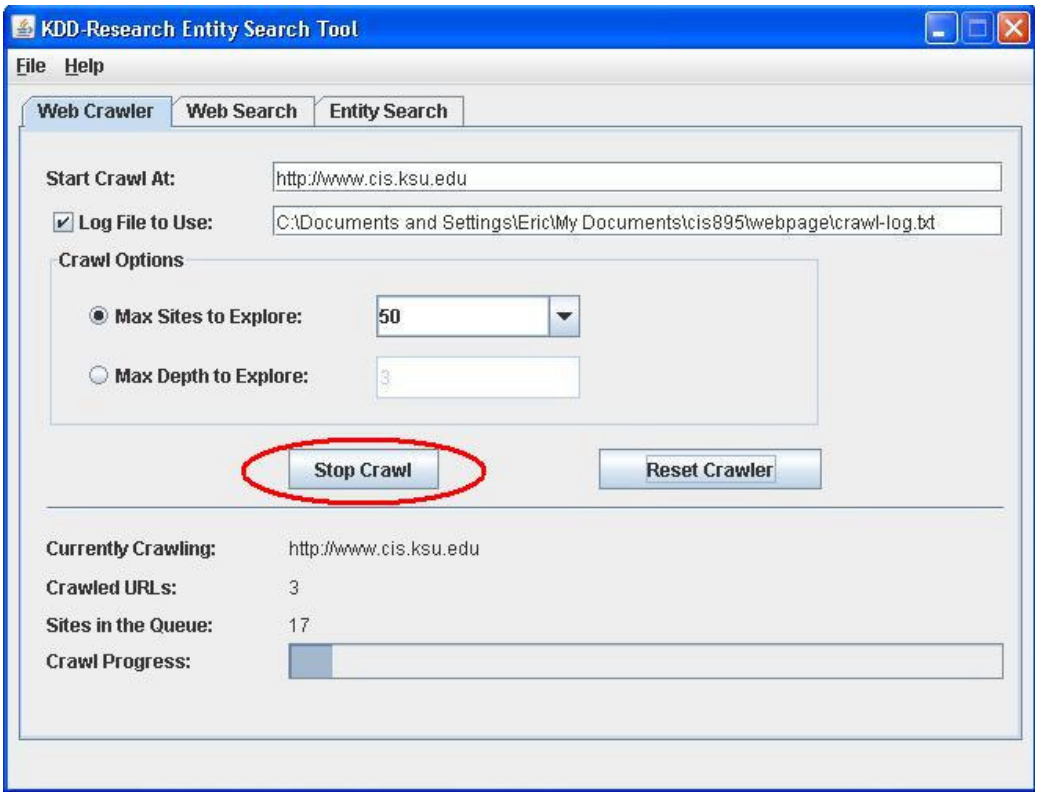

Figure 5: Stopping a Web Crawl

3.2.5 Resetting the Crawled Pages

If you want to start over from scratch after having performed a web crawl, select the 'Reset Crawler' button. It will clear all of the previously crawled web pages out of the database, and reset the fields on the form. If you are in the middle of a web crawl when the 'Reset Crawler' button is pressed, it will stop the web crawl and reset the database. The fields containing information about the crawl will also be reset.

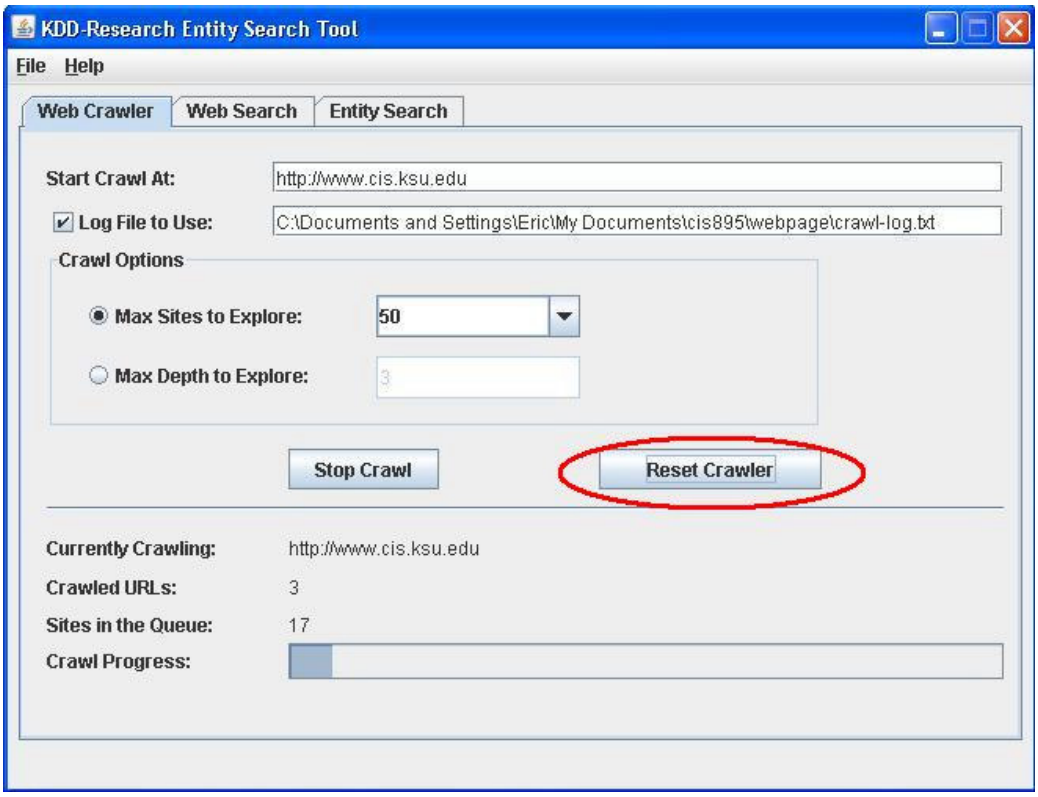

Figure 6: Resetting a Web Crawl

3.3 Performing a Web Search

Performing a web search is simple with KREST. First, you must have either performed a web crawl, or loaded pages through the application. (Loading Data is discussed in Section 3.5). To perform a web search, click on the 'Web Search' tab, enter the term that you would like to search for, and press the 'Begin Search' button. The pages that contained the search terms will be listed in the 'Search Results' table. The matching pages will be ranked according to number of back-links, that is, the number of pages that link to that particular web page.

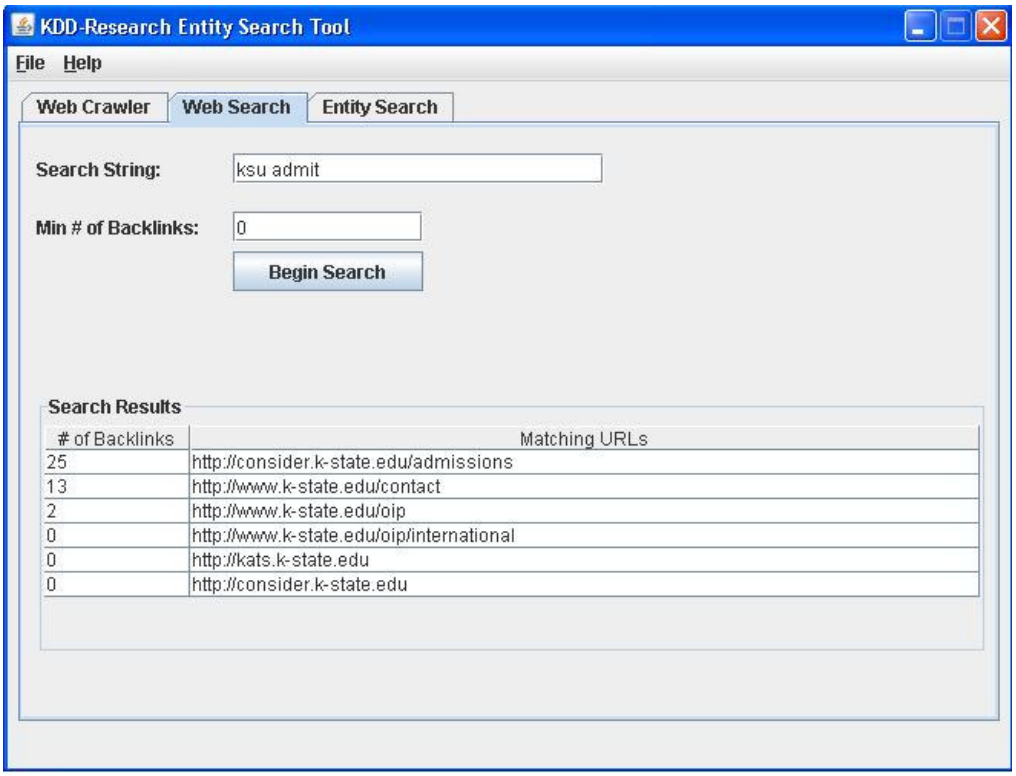

Figure 7: Performing a Web Search

3.3.1 Filtering the Web Search Results

Did you get too many results? Or only want to see the most significant ones? By using the 'Min # of Backlinks' field, you can filter out the results that do not have any other page refer to them. This helps ensure that you get the highest quality results. Simply enter the minimum number of back-links required, and press 'Begin Search' – lesser results will be filtered out automatically.

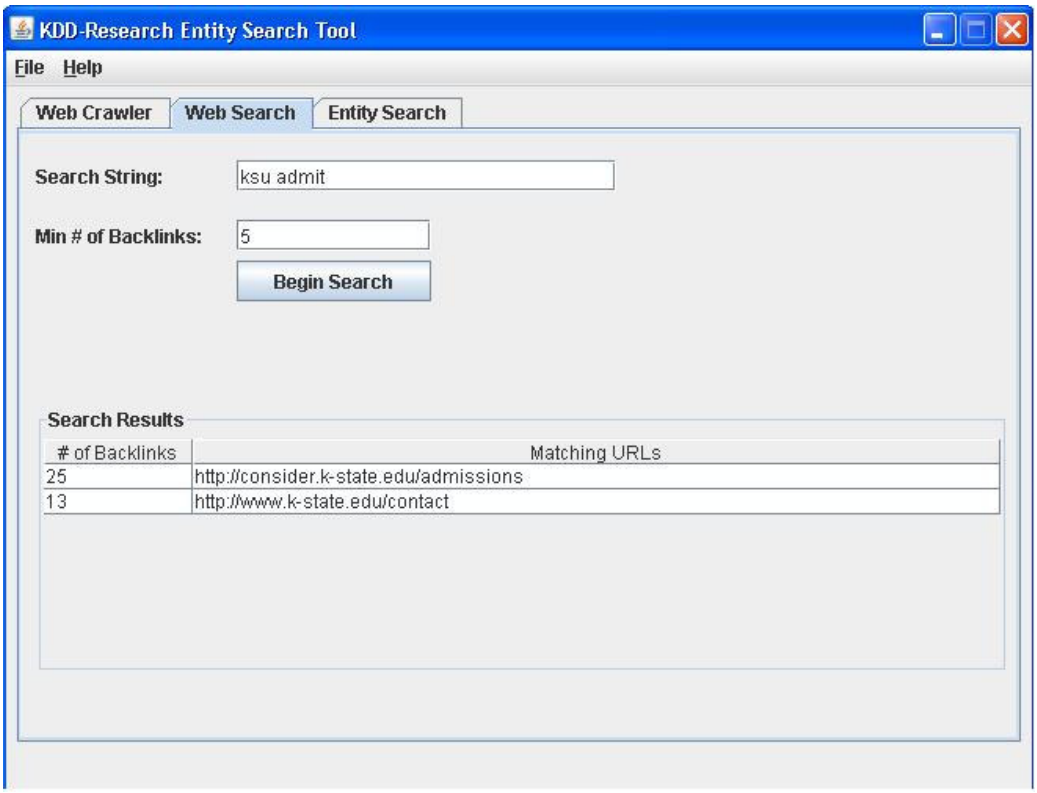

Figure 8: Filtering the Web Search by Back-link Count

3.4 Performing an Entity Search

Performing an entity search is simple with KREST. First, you must have either performed a web crawl, or loaded pages through the application. (Loading Data is discussed in Section 3.5). To perform an entity search, click on the 'Entity Search' tab, enter the term that you would like to search for, following by the entity type that you would like to find and press the 'Begin Search' button. The entity search matches will be returned as well as pages that contain the entities in the 'Search Results' table. The entities found will be ranked according to number of web pages that contained each entity.

To search for an entity, enter the type preceded by the pound (#) sign. Acceptable entity types are Street Addresses (#address), Email Addresses (#email), Phone Numbers (#phone), Fax Numbers (#Fax), and Zip Codes (#Zip). There is also an Overarching entity (#all) that will pick up all entity information. If you do not enter a valid entity type into the search box, a box will pop up notifying you of the valid entity terms.

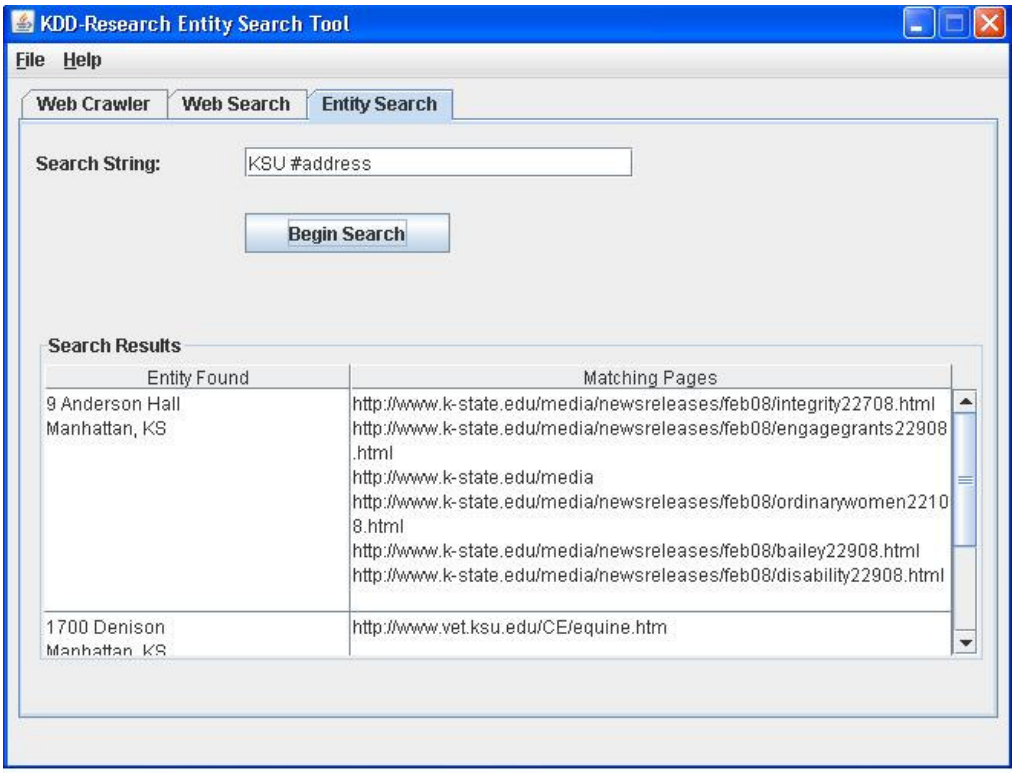

Figure 9: Performing an Entity Search

3.5 Loading Data

Sometimes you'd rather skip the web crawl and look at data that you already have on your computer. In order to load previously crawled data, simply go to the 'File' menu and select 'Load Data'. A file dialog will appear asking you to select the location of the previously crawled data. Once you select the right file, KREST will begin loading – PLEASE NOTE: Loading in data can take a while. Once the file has been loaded, a box will pop up notifying you that loading data is complete.

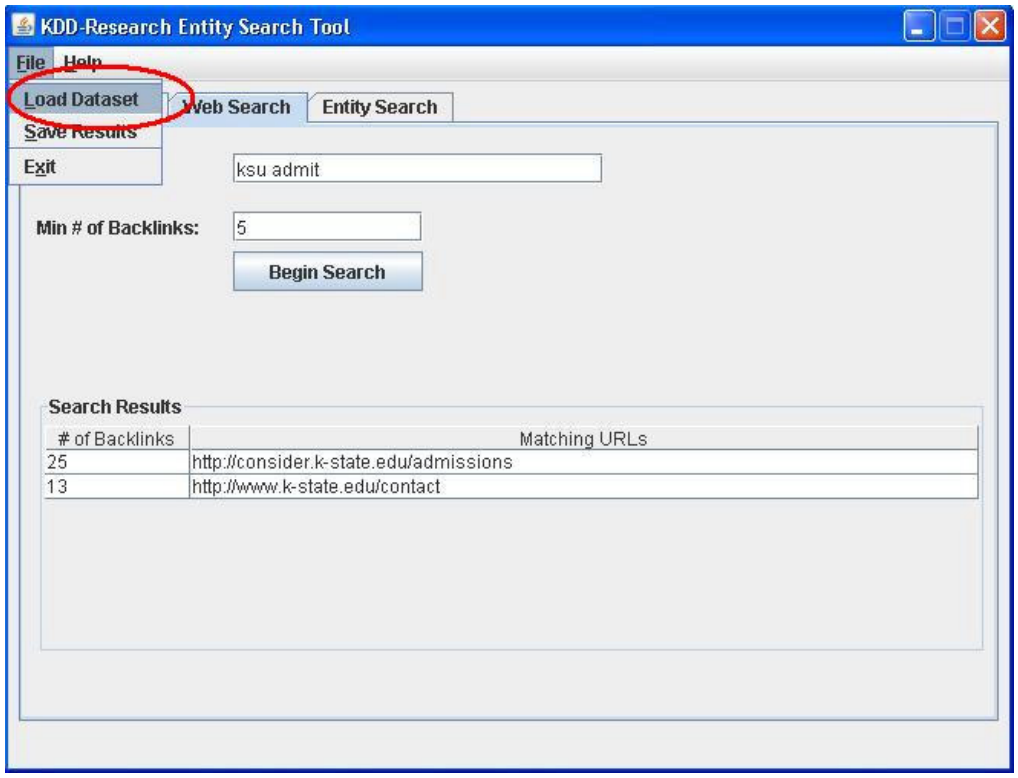

Figure 10: How to Load Data into KREST

3.6 Saving Entity Search Results

Need to save your entity search results out to a file? In order to save the results, complete a web search, then select the 'File' menu and press 'Save Results'. A file dialog will pop up allowing you to select where the results to be saved.

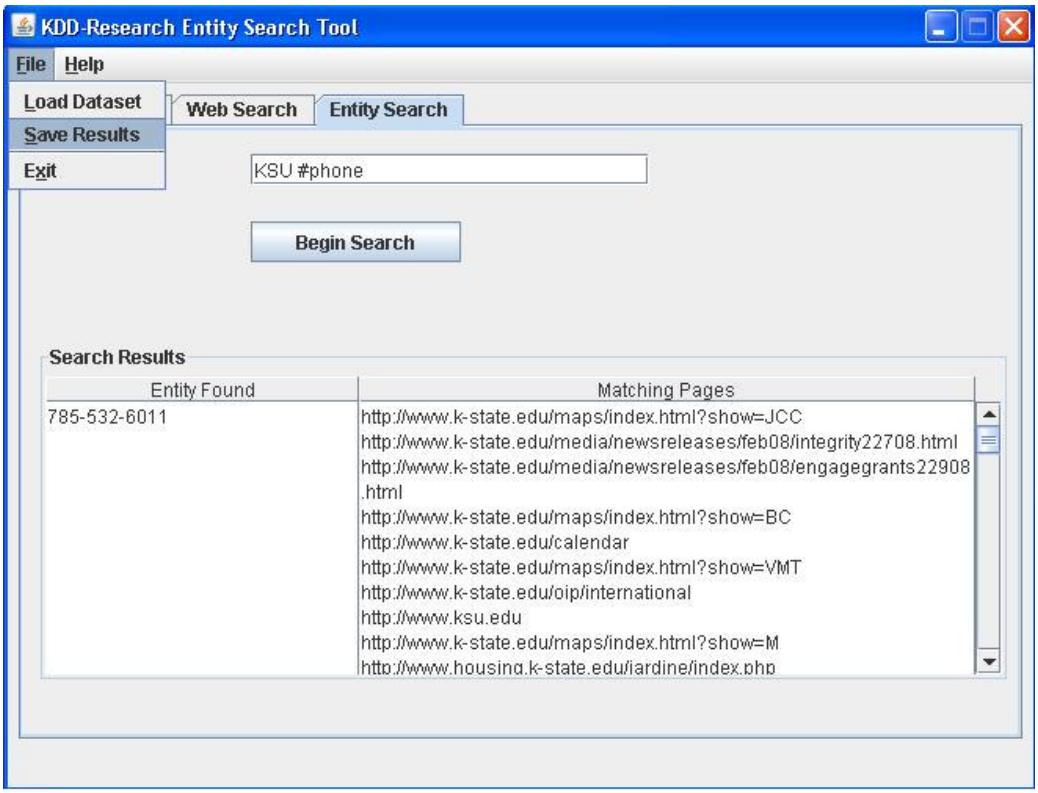

Figure 11: Save Entity Search Results

3.7 Exiting KREST

Leaving so soon? You have two ways that you can shut down the KREST application:

- Click the 'X' button in the upper-right hand corner of the application.
- Go to the 'File' menu and select 'Exit'.

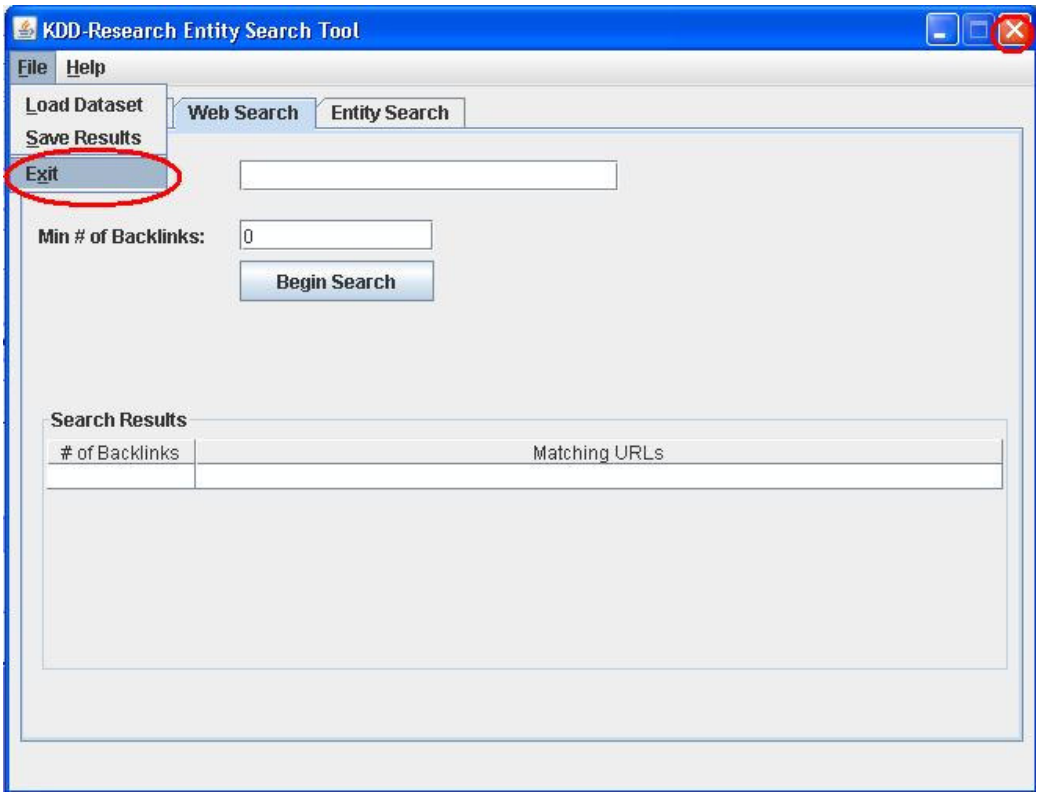

Figure 12: KREST Application with Exit Methods Circled

3.8 Information About KREST

Want to find out who created KREST, and when it was created? Click on the 'Help' menu and select 'About'. You'll see a box pop-up with information on the developer.

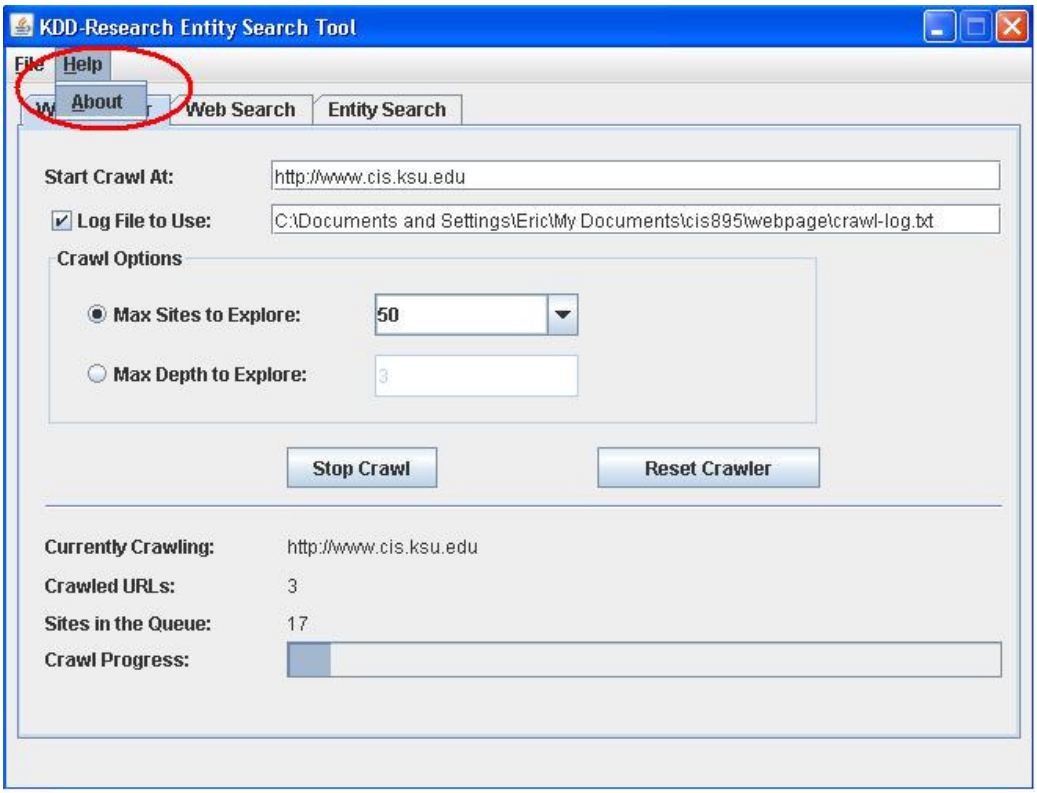

Figure 13: How to Access the Help Menu

3.9 Troubleshooting

Have a problem that wasn't answered elsewhere in the manual? You problem might be answered here.

3.9.1 Crawler Not Getting All Links on a Web Page

The Web Crawler is set to look for all instances of "http://…." in the html of the web page. It is currently unable to extract partial links (such as "/cgi-bin/index.html"). This is a feature that may be implemented in a future build.

3.9.2 Progress Bar Not Updating During Depth-First Crawls

Depth-First crawling works differently that normal Breadth-First crawling. Since the crawling keeps processing until it hits the max depth, there isn't an easy way to track when all of the pages at the max level have been processed. Because of this, the progress bar will sometimes hang at 66%. If it appears that crawling has completed (by the crawled page not changing), it is safe to move on to perform web or entity searches.

3.9.3 Cannot Click on URLs in Web Search Results

The URLs in the Web Search Results area are not clickable URLs. However, if you want to visit one of the URLs that were found, simply

click in the cell and highlight the URL. Copy the text of the URL and paste it into your web browser.

3.9.4 Cannot Click on URLs in Entity Search Results

Ideally, you would not need to click on the URLs in the Entity Search Results area, as the information has already been extracted from the web pages. However, if you really want to see the web page, simply click in the cell and highlight the URL. Copy the text of the URL and paste it into your web browser.

3.9.5 Tried to Load Data, but Received an Error Message

Currently KREST is only able to load datasets downloaded from WebBase (http://dbpubs.stanford.edu:8091/~testbed/doc2/WebBase/). Trying to load any other type of data will result in an error message being displayed.

3.9.6 Tried to Load Data, but Only Loaded X Number of Pages

The KREST application is currently limited to loading in about 32 Megabytes worth of data from a file. This is due to Java's class size restrictions. All pages that were loaded have been loaded properly, and you may perform web searches and entity searches on the loaded pages.

3.9.7 Entity Search Results Don't Match What I Expected for Overarching Results

Overarching results are based on the address. Once the address has been found on a webpage, the other entities will be searched for from that point in the webpage. Nothing before that point in the page will be recorded.

3.9.8 Searching For Multiple Entity Types

KREST is limited to searching for only one entity type at a time. If you want to search for more than one at a time, you will need to combine them all using the "#overarching" entity type. If you try to search for more than one entity type at once, the last one will be used.

3.9.9 Miscellaneous Problem Not Mentioned Above

If you are reading this section after encountering a problem, then you may have found a bug in the application. Please note the bug and email it to the developer at efd3467@ksu.edu (Maintained through May 2008). If the issue is bad enough that it is preventing you from running, shutdown KREST and restart it.