ID: Crawler-1

Priority: 10

Cost: 20

Name: Basic Web Crawler

Description: The system should be able to visit all of the pages on the website.

Precondition: Report-1 (A scan has been initiated)

Primary flow:

1. The web crawler will initialize a visited list and a pending queue.

2. The web crawler will retrieve the root page of the domain.

3. The web crawler will parse the retrieved page for resource URLs.

4. The web crawler will create a page object with the URL, a null parent URL, and a list of the child resource URLS that were retrieved.

5. The web crawler will place this page object in the pending queue.

6. If there is a page object in the pending queue, it will be popped, and become the current page. If there is not, the scan is finished, and the crawler exits.

7. The scanner will iterate over all of the child resources in the current page object.

8. If the current child resource is a page object in the visited list or the pending queue, continue with 7.

9. The scanner will create a page object using the current page as the parent.

10. The current child resource will be retrieved.

11. If a page is retrieved, resources will be parsed from the page and appended to the child resource list of the new page object.

12. The new page object will be added to the pending queue.

13. Return to 6.

Alternate flow:

1. The web crawler will initialize a visited list and a pending queue.

2. The web crawler will retrieve the root page of the domain.

3. If the root page cannot be retrieved, the scan exits.

Postconditions: Every URL referenced in any page reachable from the root page will be visited. This does NOT guarantee that it will be retrieved, as it may not exist.

ID: Crawler-2

Priority : 10

Cost: 20

Name: Record Crawler Results

Description: The results of crawling the website will be recorded in a database so that they can be used later.

Preconditions: The webcrawler must be running. The webcrawler must have generated a record in the "scans" table and retrieved a scan ID.

Primary flow:

1. When a page object is created, a matching entry in the database will be created, including the scan id, a link to the parent page (this can be null for the root of the scan), the current URL, the retrieval time, and the HTTP response code.

2. The resource_children table will be updated to create a link between the parent and the new child object.

3. The resource id of the newly created record will then be stored in the page object.

Alternate flow:

1. When a child resource is found to already have an associated page object in the pending or visited queue, a link will be created in the resource_children table between the current page object and the one that was found.

Postconditions: The resources table will contain a complete list of all reachable nodes on a graph of the website. The resource_children table will contain a complete list of the reachable edges of a graph of the website. The resources table will also contain results codes, retrieval times and a spanning tree of the graph of the website. As the response codes will be recorded and records will be created even for pages that are not accessible, this will be a record of broken links as well as working ones.

ID: Report-1

Priority : 10

Cost: 20

Name: Crawler Interaction

Description: The user must be able to start new scans and list the completed scans.

Preconditions: ID-13 (Login)

Primary flow:

1. The system will display a button to start a new scan.

2. The system will retrieve a list of scans that have been completed.

3. The system will display the list of completed scans.

4. The user may click on the "Start Scan" button.

5. When the button is clicked, a scan will be dispatched via a call to exec();

Alternate flow:

1. The system will display a button to start a new scan.

2. If there are no scans completed, none will be listed.

3. The user may click on the "Start Scan" button.

4. When the button is clicked, a scan will be dispatched via a call to exec();

Postconditions: The user will know what scans have been completed. If the button was depressed, a scan will have been started.

ID: Report-2

Priority : 10

Cost: 30

Name: Show Scan Results

Description: As a user, I must be able to view reports of the scans, showing all irretrievable resources under my domain. I should be able to do this from a website.

Preconditions: Crawler-1, Crawler-2 (A scan must have been recorded), **Report-1** (the user must have a list of scans)

Primary flow:

1. The user will select a scan that has been run.

2. The scan container will open.

3. The scan container will request the scan data from the webserver.

4. The webserver will retrieve all page resources that match this scan id, format them as JSON, and return them to the scan container.

5. The scan container will then display the JSON results.

Alternate flow: None

Postconditions: The scan will be displayed.

ID:13

Name: Login

Priority: 3

Cost: 2

Description: This is a security feature that prevents unauthorized users form making the product into a Denial of Service attack platform or find information about the target website that might be exploitable. This login will be set on creation of the system. The login will require a user name and simple password. The login information will be stored in the data base.

Preconditions: Report-2

Standard flow: User will be prompted for user name and password upon arrival to the website. The user will then type in a user name and password into there appropriate blanks and press the login button or enter.

Alternative flow: IF the user does not enter a field such as password or user name the login will fail. If user enters a wrong user name or password the login will throw a invalid login error. Should the user try and bypass this step web pages beyond this will not be displayed.

Post Conditions: After a successful login the user will be able to go beyond the login page and view or start scans.

Considerations or issues: This is a low priory by the customer however for security reasons needs to be implemented. There also is no registration now and requires the system admin to add new login entries.