SeerSuite at Virginia Tech

Robert Wagner, Brian Lieberman

CS 4624 - Multimedia/Hypertext

Virginia Tech – Computer Science

Blacksburg, VA

Client - Tarek Kanan

05/08/2013

Table of Contents

Abstract

Problem Statement

A digital library has computer-managed collections stored in digital formats. Although scientists have researched and developed digital libraries since 1991, there has been limited research on multilingual digital libraries. This project seeks to research the anticipated needs for digital library infrastructure to support multilingual information and how can this best proceed for both Arabic and English digital content. It also will recommend and assemble the necessary tools, SeerSuite and its dependencies, to establish the digital library for the crawled data. We then will display results using a web interface.

Project Procedures

· The first thing we had to do was to get a working Linux machine running and ready to install SeerSuite. We started with Ubuntu, moved to Red Hat, and after many problems with Java we ultimately switched to CentOS 6.3. Red Hat is recommended by the Seersuite developers, but we did not have a valid license so Java would not work.

· After installing the operating system, we then had to configure dependencies for the OS. We installed Java, Perl, and MySQL and configured the variables for system access to these resources.

· We then began installing SeerSuite dependencies such as Apache Tomcat, Apache Solr, Apache Ant, and Apache Axis2. These are all required before installing the SeerSuite package.

· After all the dependencies were working, and many weeks of fighting with Solr, we began installation of SeerSuite. Some of the problems we faced are detailed later in this document as well as their associated solutions The installation worked pretty well and we had the web interface, CiteseerX running then.

· The next step was to import pre-parsed data. We tried for many days to get the data imported and searchable, but we could not ultimately get the data to be searchable.

Results

· We were unable to do much research on multilingual support since we could not get any data imported or searchable.

· We have learned that installing SeerSuite is quite difficult and it would have been nice to have Steve Carman help us, but his help was limited and not as frequent as we would have liked. Due to him not replying to emails and us having issue after issue, this installation was much more painful than it should have been.

· We had our machine compromised and had to start from scratch towards the end of our project. We then worked day and night for many days and restored the machine to a working state, which shows we have learned how to install these programs, and we have laid out the plans for repeating our work in this document.

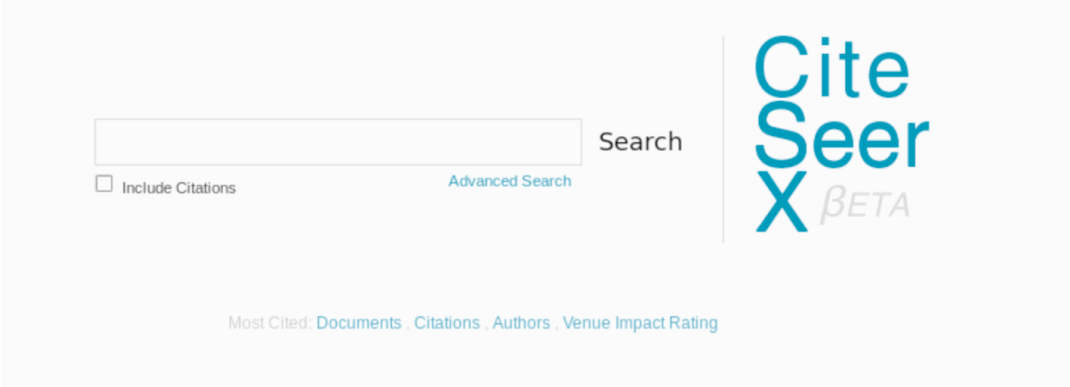
· We were able to successfully construct a system capable of ingesting documents, and crawling them. These documents then are loaded into the server's database and become searchable from the web application's interface. The details of each aspect of the application are broken down in the developer and user guides.

Conclusion

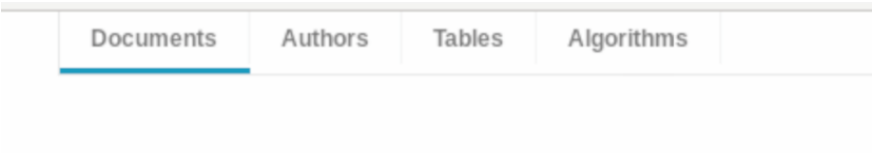
With more time, and a responsive contact that is knowledgeable about SeerSuite, we would have been able to complete this installation. We have put a lot of effort into this, and documented all the issues we had and their solutions. With this new documentation, we are confident that future teams will be able to complete our work with more ease.

User's Manual

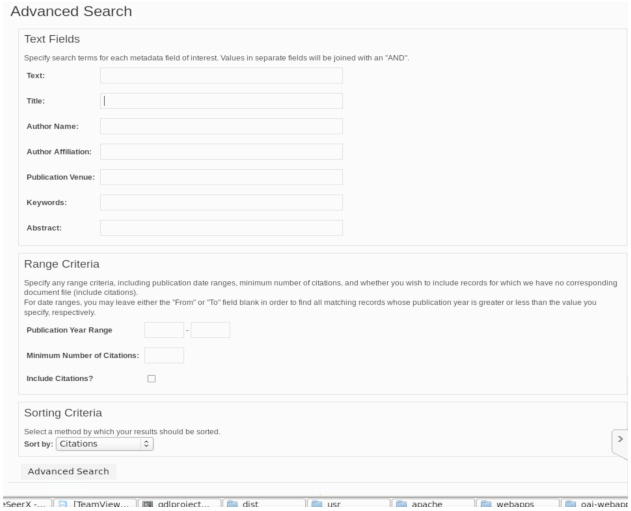
To search for data, simply navigate to the citeseerx web application (by default localhost:8080/citeseerx) and type what you wish to search for in the search box. Citeseerx is the web application front end for SeerSuite. SeerSuite is the backend crawling and processing suite which we will explain how to install later in this document. Clicking the search button will return the relative indexed documents if any are found.



To search by authors, algorithms, or tables, click the tabs found at the top of the web application and the blue highlight will change to the selected item. You can then search based on the desired method if the plugins are installed in the web application to support them. If the plugins are not installed you may receive an error when attempting to use the feature stating "contact site administrator". Currently, the installation at Virginia Tech only supports searching for documents.



Advanced search allows more fine grained control over search parameters such as specific text contained within the title of the document, keywords, and abstract. It also allows you to filter by publication year and the number of times the document was cited.



Optionally, you may follow the link on the homepage to submit additional documents for inclusion in the digital library. Currently the system does not support direct uploading of documents, however it will email the site's administrators with the document so they can review it and include it in the digital library.

Developer's Manual

Installation

Requirements:
CentOS (6.4) or preferably Red Hat if you have a license
Java SDK 5.0 or higher
Perl 5.8 or higher
MySQL 5.0 or higher
Apache Tomcat 6
Apache Solr 1.4.1
Apache Axis2
Apache Ant
Latest SeerSuite Distribution (currently v0.12)

Note: Later versions of this software can be used but have not been tested.

1. Install CentOS on the machine to be used for the server. You may select options for server configurations to include things such as MySQL or can install them later from the package manager. The package manager for CentOS is accessed by using the command "yum install 'packagename'" in the system terminal.

2. Install the SQL databases
The SQL databases can be configured to be accessed either remotely or locally on the current server. Our installation accesses all resources via localhost. MySQL can be installed if it is not already by running "**yum install mysql-server mysql php-mysql**".

Start MySQL by running:
> /etc/init.d/mysqld start

Configure MySQL:

1. **Set the MySQL service to start on boot**

   >chkconfig --levels 235 mysqld on

2. **Start the MySQL service**

   >service mysqld start

3. **Log into MySQL**

> mysql -u root

4 **Set the root user password for all local domains**

> SET PASSWORD FOR 'root'@'localhost' = PASSWORD('*new-password*');

> SET PASSWORD FOR 'root'@'localhost.localdomain' = PASSWORD('*new-password*');

> SET PASSWORD FOR 'root'@'127.0.0.1' = PASSWORD('*new-password*');

5 **Drop any users**

> DROP USER ''@'localhost';

> DROP USER ''@'localhost.localdomain';

6 **Exit MySQL**

> exit

Set up the required databases by navigating to the “seersuite /install” directory and executing:

> perl installdb.pl

When asked for which databases to install, choose ALL. It will prompt you for the login information you created in the previous step as well as domain numbers for some of the databases. You can simply use the value 1 for the domain numbers.

3. Install Apache Solr

If the Solr distribution is a source distribution, navigate to the Solr directory and type:

> ant dist

This command will build and compile the Solr.war file. Now copy the compiled /SOLR_Directory/apache-solr-x.x.x.war to /$CATALINA_HOME/webapps/apache-solr-x.x.x.war

NOTE: $CATALINA_HOME should be an environment variable set to Tomcat’s root directory when Tomcat was installed.

Copy the files in resources/solr/ to your Solr conf/ directory in order to enable the index structure expected by CiteSeerX.

Restart Tomcat and you should now see an entry in the Tomcat GUI manager for Solr.

4. Create a repository for the crawled data to be stored. To start, a local directory in the location of your choosing will suffice.

5. Configure installation

In the citeseerx/conf directory, copy *csx.config.properties.template* to *csx.config.properties*. **Do not simply change the name.** Edit the new file using the values for the configurations you have set up thus far. This includes the database login information and the URLs to Solr (typically localhost:8080/solr). The mail and smtp information can be left alone as it will not be used as of yet.

The repository mapping must be configured separately. In the

files conf/applicationContext-csx-jdbc.xml,

web/citeseerx_webapp/WEB-INF/applicationContext-csx-jdbc.xml, and

web/citeseerx_oaiwebapp/WEB-INF/applicationContext-csx-jdbc.xml

find the section that starts with

```
<bean id="repositoryMap" ...
```

and enter the repository path(s) you created in Step 2 in the

following format:

```
<property name="repositoryMap">
  <map>
    <entry key="KEY1" value="DIRECTORY1"/>
```

```
      <entry key="KEY2" value="DIRECTORY2"/>
      ...
    </map>
  </property>
```

where KEY fields are arbitrary names you choose (e.g., "rep1")
and DIRECTORY fields specify the absolute path to the repository
directories, e.g.

```
     <entry key="rep1" value="/repositories/rep1"/>
```

Next, edit the conf/applicationContext-updates.xml to supply the
key of the repository that you would like to use for imports. You
may change this value at any later time, but only one repository
can be active for importing resources at any given time. Find the
section that starts with

```
   <bean id="fileIngester" ...
```

and edit the "repositoryID" property value to reflect the key of
the repository you would like to use (e.g., "rep1").

Copy the csx-aoiconfig.properties.template to csx-aoiconfig.properties.
Edit the csx-aoiconfig.properties file

to supply values appropriate to your installation

See, INSTALL_EXTERNAL_METADATA.txt for instructions about how to setup

the external metadata database and the programs.

6. Install the web application

For the CiteSeerX web application to work properly, you will need

the following jar files on your application container's class path

in addition to the standard J2EE jars:

log4j

mail

mysql-connector-java

To build: navigate to the seersuite project directory and type:

>ant dist

This will build the citeseerx.war file. This file can then be copied to $CATALINA_HOME/webapps. You should now be able to browse the application at localhost:8080/citeseerx.

NOTE: The application may fail to start with various errors in the Tomcat manager such as “out of memory”. To fix this you can incerase the heap size of the JVM by editing the catalina.sh file in $CATALINA_HOME/bin directory to include the line:

export JAVA_OPTS = “-Xmx1024m”

7. Import pre-crawled data

To import the data, navigate to the seersuite/bin directory and run the following commands:

> ./batchImport “$Import_path”

($import_path is the path to the data to be imported)

> ./updateInference

>./updateIndex

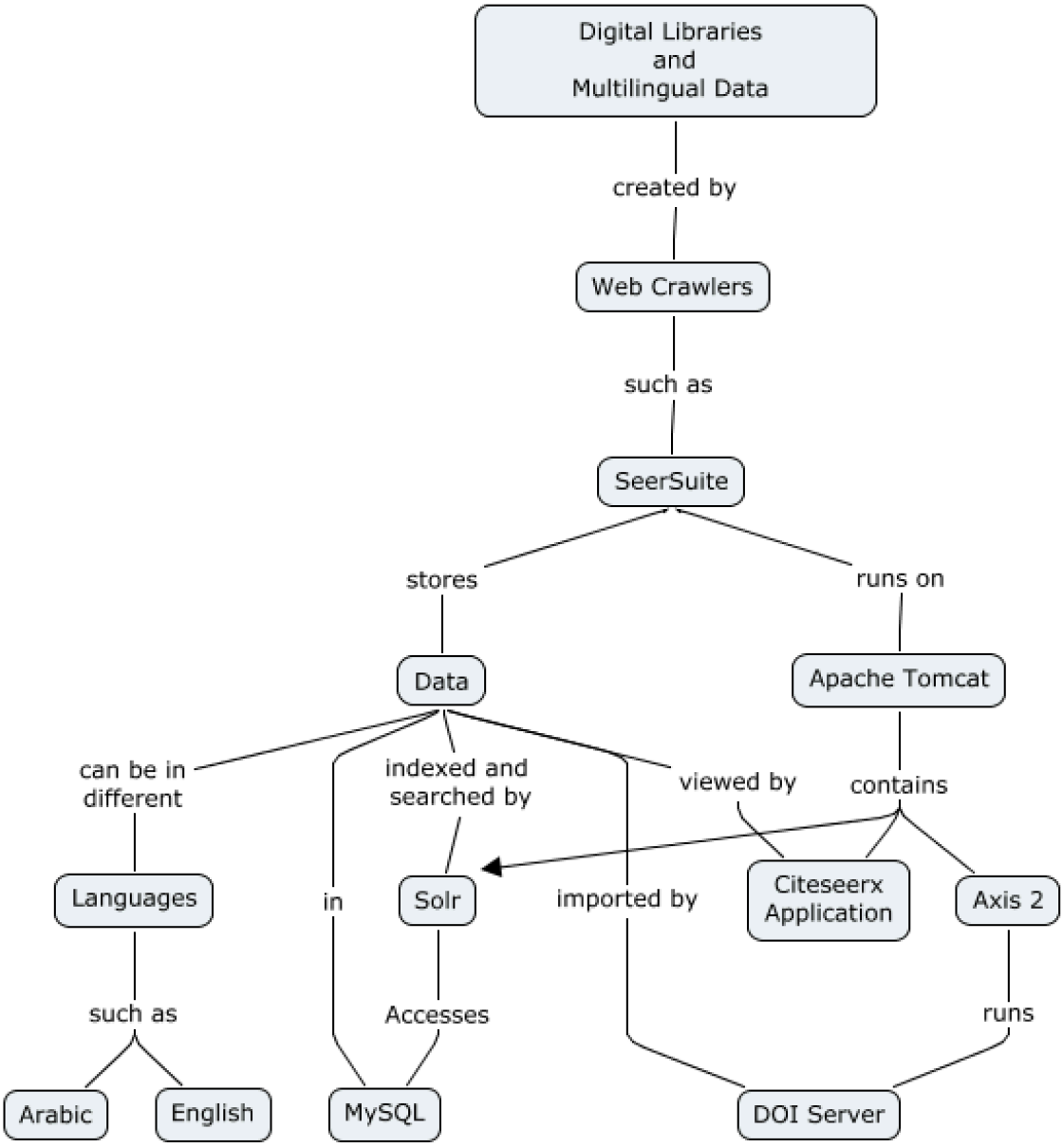
These commands will import the data into Solr and the MySQL databases and then update Solr’s index files with the new data. You should now be able to interact with the web application at localhost:8080/citeseerx and search for your imported data.

8. **Securing the server**

Precautions should be taken to secure the server to avoid further compromise and attack. A step we have taken is to configure iptables so only incoming connections are allowed from VT IP addresses. Below is the configuration for iptables:

```
*filter
:INPUT ACCEPT [0:0]
:FORWARD ACCEPT [0:0]
:OUTPUT ACCEPT [0:0]
-A INPUT -m state --state ESTABLISHED,RELATED -j ACCEPT
-A INPUT -p icmp -j ACCEPT
-A INPUT -i lo -j ACCEPT
-A INPUT -i eth0 -s 128.173.0.0/16 -m state --state NEW -m tcp -p tcp --dport 22 -j
ACCEPT
-A INPUT -i eth0 -s 198.82.0.0/16 -m state --state NEW -m tcp -p tcp --dport 22 -j
ACCEPT
-A INPUT -i eth0 -s 128.173.0.0/16 -m state --state NEW -m tcp -p tcp --dport 80 -j
ACCEPT
-A INPUT -i eth0 -s 198.82.0.0/16 -m state --state NEW -m tcp -p tcp --dport 80 -j
ACCEPT
-A INPUT -i eth0 -s 128.173.0.0/16 -m state --state NEW -m tcp -p tcp --dport 443 -j
ACCEPT
-A INPUT -i eth0 -s 198.82.0.0/16 -m state --state NEW -m tcp -p tcp --dport 443 -j
ACCEPT
-A INPUT -j REJECT --reject-with icmp-host-prohibited
-A FORWARD -j REJECT --reject-with icmp-host-prohibited
COMMIT
```

Concept Map of System Structure



Inventory of Data and Program Files

Apache Tomcat Installation: /usr/apache
Server responsible for hosting and executing web applications
Manager- password can be configured in server.xml found in the Tomcat install directory

Axis2 Installation: /usr/axis2
Responsible for executing the DOI server for document ingestion and processing

Apache Solr Installation: /opt/solr
Responsible for indexing and searching documents

SeerSuite project: /usr/seersuite

MySQL: login info configured in previous installation steps

CentOS: Tarek Kanan has login information for the current system.

Remote Access: SSH is enabled using the current users for the CentOS system we have configured

Lessons Learned

Timeline

- **2/11 Set up Red Hat Linux environment and server**
- **2/13 - 2/20 Install CENTOS because Red Hat did not work and get dependencies for Java, etc running.**
- **2/21-3/25 Install all dependencies for SeerSuite, troubleshoot with Solr.**
- **3/25 Midterm Presentation**
- **3/25-4/15 Work with Steve to get Solr running, and finish installing SeerSuite to get it running.**
- **4/20-4/30 Parse sample data**
- **05/02-05/04 Attempt to crawl some documents, and write up documentation.**
- **5/06 Final Presentation**
- **5/08 All Project materials are due**
- **5/10 System restored after compromise**

Problems

- **Solr**
  - Had trouble getting Apache Solr installed and deployed under Tomcat as servlet.
  - Did not know configuration variable names to put in the SeerSuite config
- **Java**
  - Version of Red Hat we were using suffered from bugs that were in a deprecated version of Java
- **Red Hat**
  - Issues getting dependencies installed and working
- Communication
  - Steve was our expert and rarely replied to us
  - Getting access to the lab through Tarek was difficult
- **MySQL**
  - Login information incorrectly configured caused errors
- **Parsing Data**
  - Loading in sample data did not work, despite many attempts
- **Java Heap Size**
  - Out of memory error in Tomcat when trying to start citeseerx
- **Security**
  - Our machine was hacked and a bot.irc malware program corrupted our machine. Could have been caused by using Solr 1.4.1 and old programs

Solutions

- **Solr**
    - Installed Solr using updated documentation to fix our installation errors. This was the official Solr documentation given by Apache.
    - We talked with Steve and got the variable names and configuration info we needed to put in the CiteseerX config file.
- **Java**
    - Resolved with switch to CentOS 6.3 which did not have the Java installation issues.
- **Red Hat**
    - Switched to supported and current version of CentOS 6.3
- **Communication**
    - Got Steve's Skype and Cell phone to try to contact him more, but it didn't help much
    - Set up remote access so we could work from anywhere and didn't have to contact Tarek to get in the lab
- **MySQL**
    - Reconfigured Citeseerx with correct login information which was username=root and password=Qdl2013
- **Parsing Data**
    - We could not find a solution to loading in data. We tried many things like moving the folder, manually uploading to Solr, messing with MySQL, and nothing we tried worked.
- **Java Heap Size**
    - Added export JAVA_OPTS = "-Xmx1024m" to catalina.sh in Tomcat's install directory to expand heap size.
- **Security**
    - We had to reinstall CentOS and configure everything again, and we will attempt to get everything reinstalled. We did ipconfig to try to secure this more. During our reinstall we updated to the current version of Solr and this may fix some security issues as well.

We learned many lessons working on this project. First and foremost, we learned that software packages do not all work well together. We had to spend many weeks just getting a stable OS and Java combination that worked before we could start installation. Secondly, we learned that using outdated versions of software, like Solr 1.4.1, makes the documentation on install not completely clear. From our problems with Solr we learned that communication with experts can be very helpful, and we resolved our problems by talking with Steve Carman, one of the experts on Seersuite. Another big lesson we learned was without good documentation, installation can be very hard. The Seersuite documentation is lacking a lot of necessary details and we had many issues

getting it installed. We have tried to remedy this by updating the documentation and illustrating solutions to questions we had in our own, revamped developer's manual.

An unforeseeable event occurred with only a few days until the project deadline. Our machine was hacked and we had a bot.irc malware virus. The only solution was to reinstall CentOS and start from scratch and hope to be able to reinstall Seersuite in the small amount of time we had remaining. These issues of being hacked were likely caused by, or at least not helped by, using extremely old version of Tomcat, Solr, etc which would have many vulnerabilities. The Seersuite experts and developers from Penn State recommended these versions of software. They were considered to be known working versions and despite their end of lifecycle, they did work. During our reinstall we updated to the current version of Solr and this may fix some security issues as well.

Acknowledgements

Tarek Kanan - Our Client
tarekk@vt.edu

Steve Carman - Our SeerSuite Contact (old)
shc5011@ist.psu.edu
Cell: (267) 240-0362
Skype: hntd187

Kyle Williams – Our SeerSuite Contact (new)
kiw5209@psu.edu

Edward A. Fox - Our Professor
Computer Science
114 McBryde Hall, M/C 0106
Virginia Tech
Blacksburg, VA 24061

E-mail: fox@vt.edu
Work: (540) 231-5113

References

"Apache Axis2 Installation Guide." *Apache Axis2*. Apache, 17 Apr. 2012. Web. 04 May 2013 <https://axis.apache.org/axis2/java/core/docs/installationguide.html>.

"Apache Tomcat 6.0." *Apache Tomcat 6.0 (6.0.37)*. Apache, 29 Apr. 2013. Web. 04 May 2013 <https://tomcat.apache.org/>.

"MySQL." *How to Install on CentOS RedHat Linux, How to Configure on CentOS RedHat Linux*. Antoine Solutions, Web. 04 May 2013 < http://dev.antoinesolutions.com/mysql>.

"SeerSuite Project Hosting" *SeerSuite Project Sourceforge Page*. Pennsylvania State University, Web. 04 May 2013 < http://citeseerx.sourceforge.net/>.

"SolrTomcat." - *Solr Wiki*. Apache, 06 Jan. 2013. Web. 04 May 2013 < https://wiki.apache.org/solr/SolrTomcat>.