**Web Search Engine Project**

Joshua Middleton

Joshuah Braun

Conner Loveless

Logan Rigdon

Andrew Baughman

Web Search Design Team

Submitted to—

Dr. Mahesh Khadka

Department of Electrical and Computer Engineering, Trine University

**Contents:**

**3 - Executive Summary**

**4 - Statement of Problem**

**4 - Objectives**

**5 - Technical Approach**

5 - Identifying Needs of Customers

5 - Identifying Target Specifications

6 - Generating Design Concepts

6 - Selecting Design Concept

**7 - Project Management**

7 - Deliverables

8 - Budget

8 - Communication and Coordination with Sponsor

9 - Team Qualifications

**Executive Summary**

Our goal for this project is to create a web search engine that searches a catalog of Trine University’s domains. While it is possible to find what you need using another search engine, it is not always optimal or accurate. By exclusively making the associated database Trine-related domains, it will make searching for anything related to Trine significantly easier and more centralized.

To accomplish this task, we have developed a design that includes all necessary components to this web search engine system. These main components are web scrapers, a server that will contain the database, and an Azure-hosted website. The web scrapers will continue running automatically, collecting site data and sending it to the database. The database will contain all relevant data that the web scrapers have found. The website is for the end-user and will query the database at the request of the user. The results of this query will be ranked by relevancy and returned to the user in a timely manner.

This web search engine will be available from an outside facing internet device via a web app. It will be capable of calculating search results in a second or less. These results will be displayed to the user on the website using a format we have designed. An algorithm will be used to calculate the relevancy of any website in the database given search parameters.

**Statement of Problem**

Many students and faculty find it difficult to find new information relating to Trine University. While the organized structure of Trine’s various websites helps users to find the content they want, it is sometimes difficult to find that content because people might not know which website to look at or where within a website to look.

Many people use google.com to search for stuff about Trine such as Final Exam Schedules and Degree requirements. While Google is a great resource, it cannot be guaranteed to provide correct results always because it is not restricted to looking just at Trine University information. Since Google searches the entire internet, users might need to dig through results that do not relate to Trine to find the information they want.

**Design Objectives**

This document proposes that we will develop and publish a web app that acts as a search engine to parse through Trine University websites to make a Trine University search engine. We intend for our site to:

* 1. Be able to make a search with an index of approximately 3 words.
  2. Calculate results in a second or less.
  3. Have a relevancy algorithm to return accurate results.
  4. Be accessible from an outside facing internet device.

Our site should be able to be used from any computer that is internet connected, meaning published to a domain. It should be able to provide results that are correlated to the search index that will be input by the front-end user. We need to achieve these objectives because they are necessity; Being able to provide accurate results in a timely manner is the fundamental function of a search engine.

**Technical Approach**

**Identifying Customer Needs**

The customers for this design are perspective students, current students, and staff since all these individuals interact with many pages under the Trine domain.

For current and perspective students, we have personal experience from each of the five team members. This personal experience leads us to the conclusion that pages under the Trine’s domain are too loosely connected and thus information is difficult to find since a search directed at a single point will not return results from all pages under Trine’s domain. As former perspective students and current students, we identify that we need a search engine for Trine’s domain that will return relevant results in a quick manner.

For staff members at Trine, we will need to conduct an informal survey to determine if the needs we perceive as students match the needs of the staff. The gathering of this information will primarily be done through conversations held with certain staff members.

**Identifying Target Specifications**

The first specification to target is the searching mechanism. This specification is targeted by creating JavaScript code that allows communication between the website, the host, and the server database. This code will allow users to provide input and receive a few pages of search results.

The second specification is the calculation of results within one second. This is targeted by creating a lightweight ranking algorithm that will allow results to be calculated quickly on the server side. Compromise will need to be made between relevancy and speed. Since the scope requires that results be returned quickly, relevancy of search results may reduce to allow for quicker computation.

The last specification to target is the ability to access the service from the internet. This can be accomplished using a Domain Name Service. This can either be done by using our server to host not only the database, but also the website. This would require the purchase of a domain name. The other option is to use a cloud service that provides the domain.

**Generating Design Concepts**

The concept generation process included research, discussion, and some experimentation. Team members each did research to grasp the concept of search engines and how one could be designed. Following that process each member provided their own design concept that would meet the scope of the project.

After the initial concepts were generated, with team held a meeting in which the concepts were refined or rejected by the team. After the concepts were narrowed and refined, it was discovered that two primary concepts remained.

The first concept was to hold all resources on premise. This would require the server to not only house the database, but also require that the server host the website. Additionally, a domain name would need to be purchased so that the website could be accessed on the internet.

The second concept was to make used of the Azure cloud service. This would allow the website to be hosted on the Azure platform where the complications of making the website accessible are mitigated. This concept would require communication between the Azure service and our on-premises server since it was determined that the cost of hosting the database in the cloud would not be acceptable.

**Selecting Design Concept**

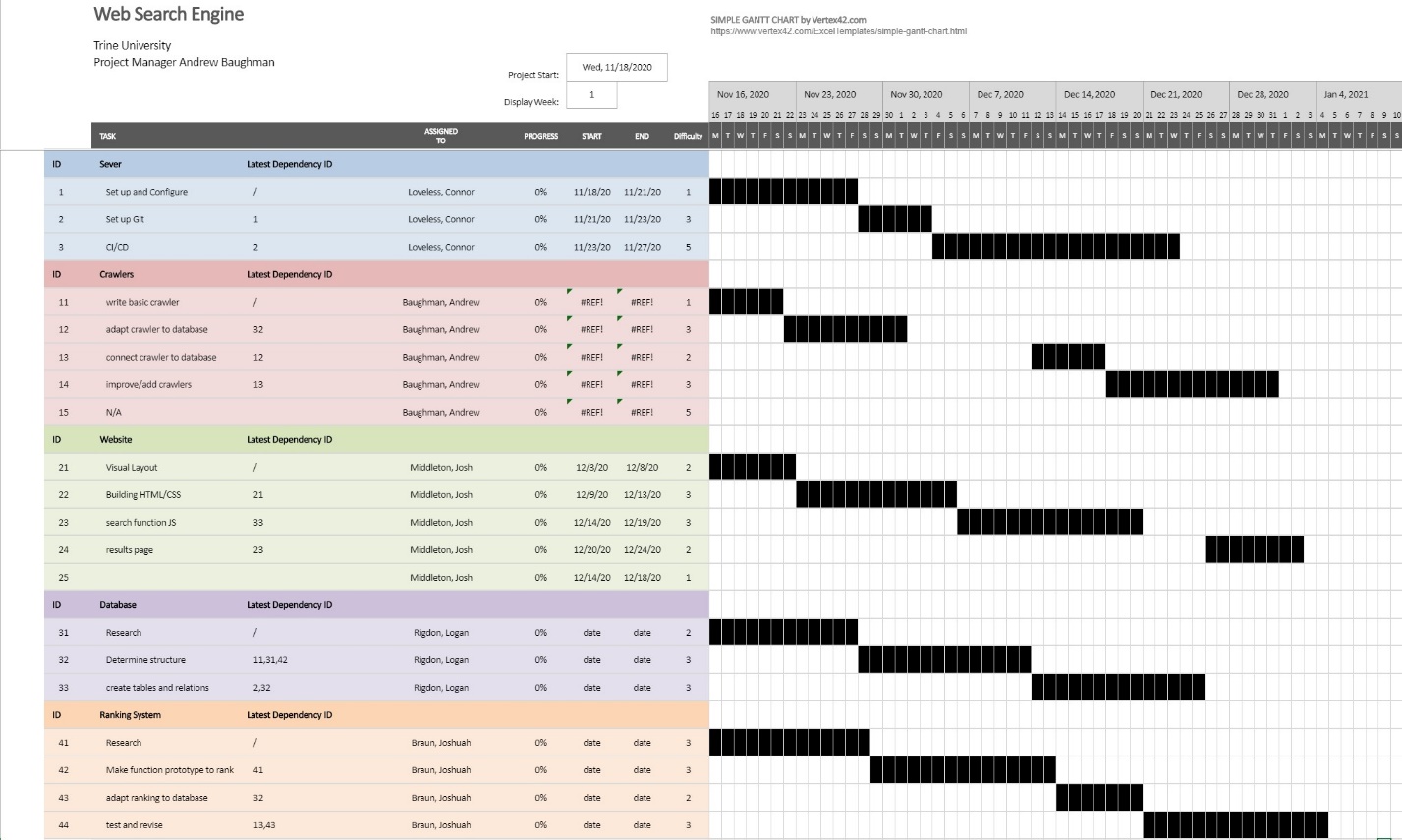
The process of selecting a concept was done through a team meeting in which we weighed the benefits and drawback of the two design concepts that were presented in the last section, “Generating Design Concepts”.

The team decided to select the Azure cloud-based hosting concept as the primary choice. The major benefit of this design was that hosting the site was made easier by adopting the service. As a result of this, the project is made easier to quarantine if the Coronavirus force our team off campus. This is because we do not need to worry about which internet service provider would. This way, contracts that would not need to be resolved to make our website accessible. it was also determined, though research, that it was indeed possible to connect our cloud service to our on-premises database.

Our alternative concept, should the primary one become infeasible, is the fully on-premises approach. This approach would not make use of the Azure service as describe in the previous section, “Generating Design Concepts”.

**Project Management**

Our plan to build a web search engine for Trine University is shown in our Gantt chart below. We plan to finish the project by the expo. The management section will include the deliverables such as hardware to be used in the project, a listed budget with costs of items and suppliers we want to use, communication with the instructor, and team qualifications.



**Deliverables**

When the project is finished, the result will be a website that the customer will be able to access from any web browser. This website will feature a text field and a search button. If the user types a query into the text field and clicks the search button, a webpage will appear that will feature a list of webpages as links with descriptions. The results will be placed so that the most relevant results are at the top of the page.

Other deliverables include the hardware responsible for acquisition and management of data as well as the Microsoft Azure Account that hosts the website. The hardware that acquires and manages the data will be a computer and one or more Raspberry Pi’s for web crawling.

**Budget**

Most items for our project cost very little with the exception of the database server. The two largest cost items aside from the database server will cost quite little. Currently, we can only speculate as to what specifications we’ll need for the database server, so we will not purchase the database server until we know how memory, CPU, and network usage will scale with how much data is present and how many users are connected. Prior to knowing this information, we’ll use a second Raspberry Pi to act as a small-scale database server.

Table 1: Requested items and funds for initial design

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Item | Supplier | Quantity | Unit Price | Total |
| Raspberry Pi 3 B+ | Adafruit | 2 | $35 | $70 |
| Network Cables | The Pi Hut | 3 | $5 | $15 |
| Microsoft Azure Account | Microsoft | 1 | $60/month | $300 |
| Database Server | Microsoft | 1 | $500 | $500 |
| Domain Name | GoDaddy | 1 | $12 (15 for .net, 10 for .org, 12 for .com) | $12 |

**Communication and Coordination with Instructor**

To keep the instructor in the loop, we will be having regular meetings with him that vary from brief to in depth based on what was done during that timeframe. The instructor will help us stay on track as well as keep the project from becoming too large of a task for any member of the group to handle. As a team, we will be submitting our information in a document for the instructor to check as we complete tasks.

**Team Qualifications**

Logan Rigdon has worked with SQL to create and query databases using Oracle SQL Developer.

Josh Middleton has done web development within the Azure cloud service and has experience with SQL, HTML, and JavaScript.

Joshuah Braun has worked with Python-Django to make web applications including PostgreSQL, HTML, JS, and CSS.

Andrew Baughman has developed Django web applications including database management and interaction as well as web-scraping.

Conner Loveless has developed Django web apps that included databasing, Python, and Html and CSS tools.

**Appendix**

**Resumes**



