Final Project for Comp 4601

“Crafty - A Solution for Finding Cheaper Prices on Art Supplies”

Brittny Lapierre SID: 100922938
Kelly Maclauchlan SID: 100927176

Abstract

Our project, Crafty, is a product catalog containing data scraped from five major craft and art supply stores. With Crafty, you are able to search across these platforms and receive the basic information on the products such as price, brand, and store along with a link back to the original site if you wish to purchase the item. Crafty was developed to be a resource that can be used by people like ourselves who enjoy buying arts and craft supplies but don’t want to check prices on each individual site.

Introduction

When thinking of an idea for a final project we knew we wanted to create something that we could see ourselves using. Because we both enjoy making art and doing crafts, we came up with a web app to help us find reasonably priced crafting supplies. Our app, Crafty, allows a user to search through products from multiple art supply store websites at the same time. Users can also add products to Crafty’s catalogue manually through Crafty’s ‘add product’ feature. Our project contains various concepts from the semester, some of them being web crawlers, search engines, and recommendation systems.

The following paper will describe how we created Crafty and its functionality in more detail. First, we will go over the tools used to create the project and the concepts that we took from the course. We will then go into some related systems that currently exist and our process of how we tackled the problem. After this, we will move on to discuss what we achieved and what did not work as well as we had hoped. Finally, we will have a conclusion that will go over the future possibilities of this work.

Background

Web crawler
A web crawler is a program that goes through web pages with the intent of gathering information. They can be used to perform a number of tasks but the one we have focused on is web scraping. With web scraping the main goal is to retrieve information from the page and save it in a defined data type that can be used to do something with the data you have collected.

Clustering
Clustering involves organizing items in a set into groups which share common features that can be analyzed or exploited. One of the techniques used to group similar users was clustering using K-Means. With this technique cluster centers are randomly generated in an N-dimensional space corresponding to the dimensions of the data analyzed. This the algorithm looks to assign each data-point to the nearest cluster center and adjust each center using the points surrounding it. This process continues for multiple iterations until the groups become stabilized, giving the final clusters for the data.

Recommender System
A recommender system serves the purpose of predicting user sentiment towards items in any given dataset in order to suggest certain items from the dataset more than others.

Search Engine
Search Engines are systems which allow users to query collections of documents to find ones which can be said to be related to their query.

Related work

The concept of a multi-site search platform has been seen in many different areas, but a common application would be for travel. There are a number of websites that take information from various travel sources. By amalgamating this data into a single search platform users can find the best deals while visiting fewer domains.

Development Methodology

To start off this problem we first decided a list of stores that we wanted to scrape to gather information from. After looking through our original list of nine stores we realized that some were not feasible choices to gather information from. Some of the sites had too many products on them such as Amazon and Walmart that we did not feel we could target the scraper to a reasonable section of the site. Other Sites like Staples have user interface choices that make scraping difficult due to infinite scrolling features. We also considered using Michaels Canada but they only list the prices for their products on their American website, and the price differences are not exactly equivalent to the exchange rate.

In the end, we Scraped 5 stores Deserres[1], Above Ground Art Supplies [2], Art Shack[3], Jerry's Artarama [4], and Wallacks [5]. To scrape these websites we created a python web scraper that went through various sections of these websites and collected the information into a common data structure. The data structure contained a number of elements such as the name, price and original link for the product. These results were all saved to JSON files that were then read in by our Java application to our database. While reading we indexed the products using Lucene for optimized searching.

The Java web application was build upon Jersey, which we had studied in class. We extended it by using a library called Jade4J which allowed us to use the popular template processor Jade to create the front end of our application. In the Java web application, one can find a number of endpoints accessible through the url http://localhost:8080/COMP4601_FINAL_SERVER/crafty/.

- *Parse:* reads in the data collected that has been stored in files by the python scraper and saves them to the database.
- *Query:* accepts a list of terms that can be used to generate search results from the data stored in the database and returns a list of products that match the search.
- *QueryRecommend:* accepts a list of terms and the user's name and generates search results but with the recommended results are shown first.
- *Watching:* adds a query to the users 'Watched Items' list. Users can then see the top 5 matches for each of their watched items when navigating to their dashboard.
- *Add-Product:* accepts a form that contains the parameters to create a new product that will be added to the database
- *Unwatch:* removes a query form the users 'Watched Items' list.
- *Viewed:* adds a product to the users viewed history. This is helpful for gaining insights on what kinds of products our users are interested in.
- *Create-Account:* accepts a form that contains the parameters to create a new user.
- *GenerateUsers:* generates a number of users and the number of products they have viewed. This is currently set at 500 users and 100 randomly determined products each. In order to test the system, we decided to have fewer users than in previous assignments so that the program would run faster. As it is this generation of users takes approximately 5 minutes to be generated and saved to the database.
- *GenerateCommunties:* is used to separate the users into various communities using K-Means. This was originally going to be used to generate recommendations for products

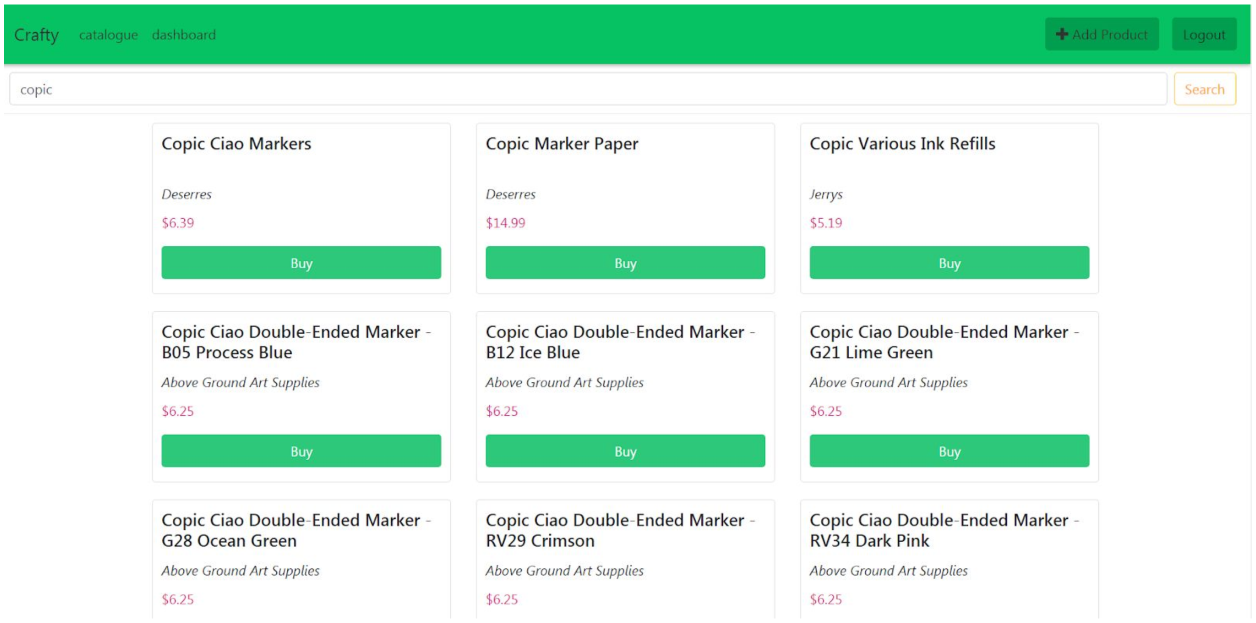
but due to the small number of users and viewed items these results did not produce even groupings and in most cases simply put all of the items into one community. This is most likely because of how we created the vector representation of the user for K-Means. This vector has ten values: five relating to the store that sells the products, and the others relate to the price that these products are viewed. These values are counted while the users are randomly generated, and the percentage decimal of each is taken in relation to the total number of items they have viewed. With five options for each of these, it is common to have the majority of the values around 0.2. Due to this, it was very hard to find a good value to use for the centroids of the clusters.

- *Recommended:* takes in a user and returns ten items that are recommended to them. To generate these items there is first a list of 200 products randomly selected. Then the list is traversed to find items that match the preferred store of the user and that fall within the preferred price range of the user. Both of these values are determined in relation to what the user has previously viewed. To determine the prices that should be considered the prices are categorized based on the price ranges $0-$20, $20-$50, $50-$100, $100-$300, and $300+. For these recommendations, this list is returned once the desired number of items has been added to the list. The process to generate recommendations in a query is very similar. However, instead of using a randomly generated list of products, the returned search results are used as a basis for recommending products. From the search result items, we take the items that match the preferred store and price and put those items the front of the returned results.
- And finally, endpoints also exist for serving up our static html generated using Jade4J.

User Interface

The user interface contains five main pages. These pages include: a login page, the catalogue page, the user dashboard page, and the create product page. As mentioned earlier, we used the Jade4J template engine to create these pages. We were also able to further break down these pages into separate components which could be included into each of the templates. For example, instead of writing the same html code for the navigation bar four times, we simple include another jade file which contained the html code for the navigation bar in each of the page files.

Upon loading the app the user will find themselves on the catalogue page. From there they have the option to search our catalogue for items, add an item to Crafty's catalog, or log into their account to take advantage of more features of the app.



Copic Ciao Markers
Deserres
$6.39
Buy

Copic Marker Paper
Deserres
$14.99
Buy

Copic Various Ink Refills
Jerrys
$5.19
Buy

Copic Ciao Double-Ended Marker - B05 Process Blue
Above Ground Art Supplies
$6.25
Buy

Copic Ciao Double-Ended Marker - B12 Ice Blue
Above Ground Art Supplies
$6.25
Buy

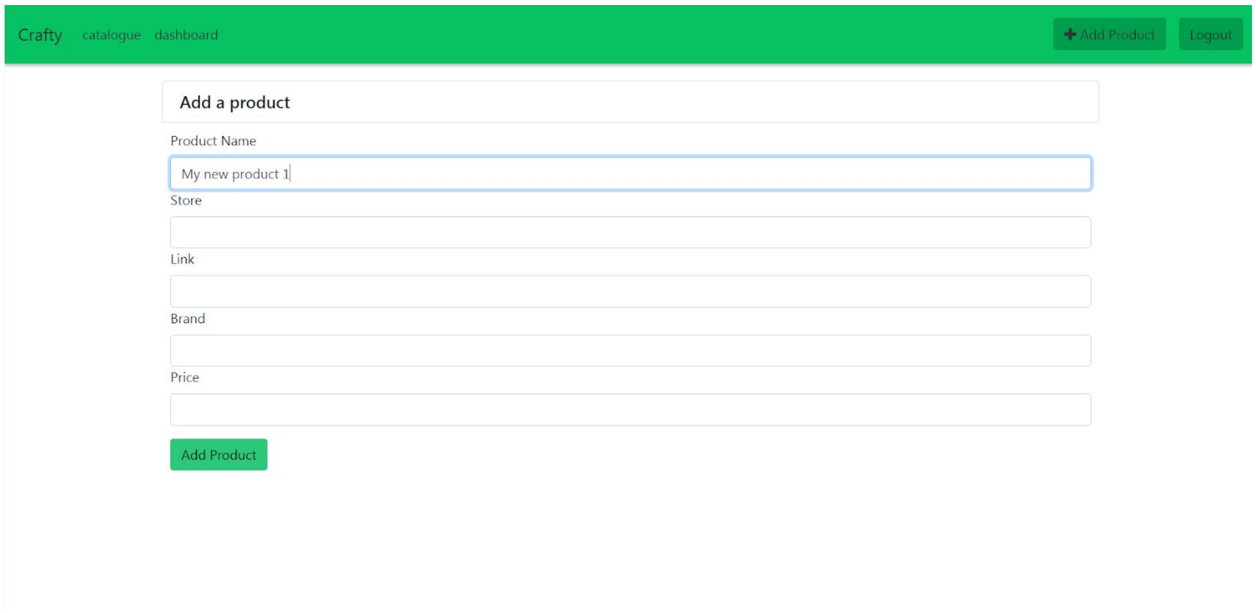
Copic Ciao Double-Ended Marker - G21 Lime Green
Above Ground Art Supplies
$6.25
Buy

Copic Ciao Double-Ended Marker - G28 Ocean Green
Above Ground Art Supplies
$6.25

Copic Ciao Double-Ended Marker - RV29 Crimson
Above Ground Art Supplies
$6.25

Copic Ciao Double-Ended Marker - RV34 Dark Pink
Above Ground Art Supplies
$6.25

Figure 1: The Crafty Catalogue



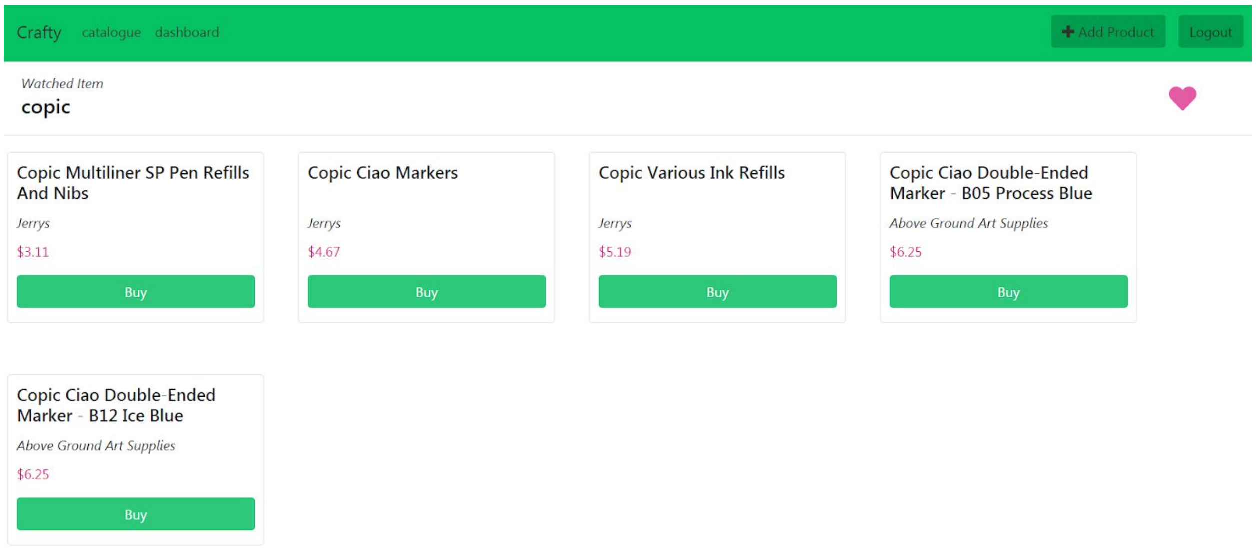
Add a product

Product Name: My new product 1
Store:
Link:
Brand:
Price:

Add Product

Figure 2: Add a New Product to the Catalogue

If users log into Crafty, we can then track what products they are viewing, and automatically order their search results based upon data we collect about them. When users are logged in they also have the ability to save queries as ‘watched Items’ to their dashboard. Their dashboard will then display the top 5 cheapest products matching the saved query. Their dashboard will also display the top 10 recommended products for the users.


Crafty catalogue dashboard

Add Product Logout

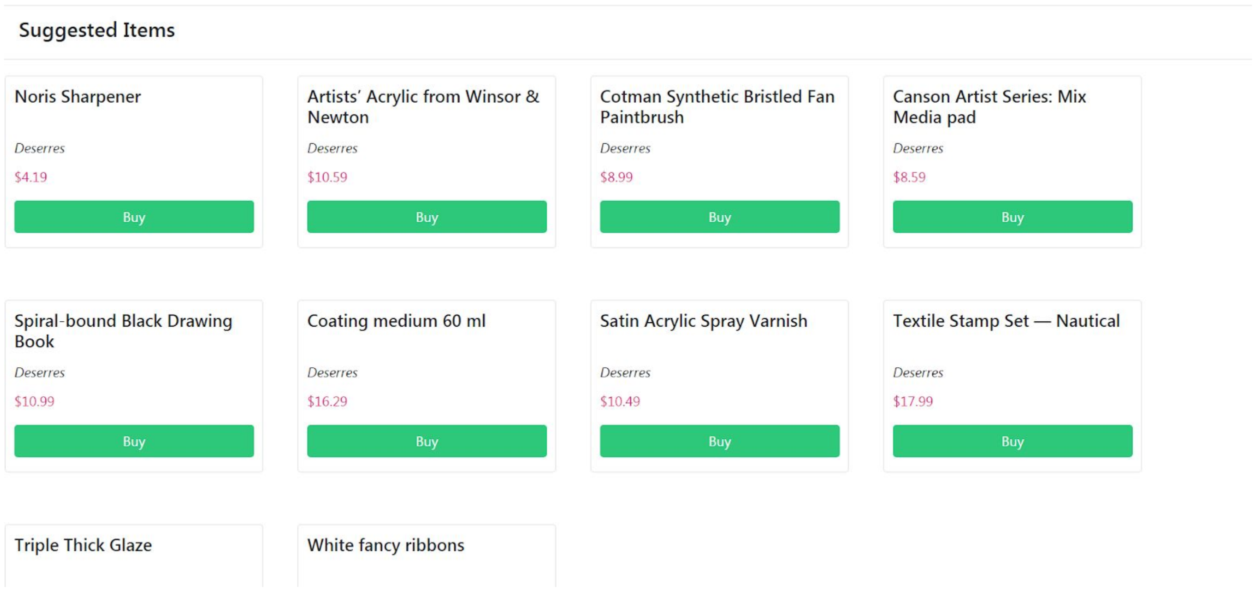
Watched Item

copic

Copic Multiliner SP Pen Refills And Nibs	Copic Ciao Markers	Copic Various Ink Refills	Copic Ciao Double-Ended Marker - B05 Process Blue
Jerrys	Jerrys	Jerrys	Above Ground Art Supplies
$3.11	$4.67	$5.19	$6.25
Buy	Buy	Buy	Buy

Copic Ciao Double-Ended Marker - B12 Ice Blue
Above Ground Art Supplies
$6.25
Buy

Figure 3: View 'Watched Items' on your User Dashboard


Suggested Items

Noris Sharpener	Artists' Acrylic from Winsor & Newton	Cotman Synthetic Bristled Fan Paintbrush	Canson Artist Series: Mix Media pad
Deserres	Deserres	Deserres	Deserres
$4.19	$10.59	$8.99	$8.59
Buy	Buy	Buy	Buy

Spiral-bound Black Drawing Book	Coating medium 60 ml	Satin Acrylic Spray Varnish	Textile Stamp Set — Nautical
Deserres	Deserres	Deserres	Deserres
$10.99	$16.29	$10.49	$17.99
Buy	Buy	Buy	Buy

Triple Thick Glaze	White fancy ribbons

Figure 4: List of the User's Top Ten Recommended Products

Discussion

By using the preferred store of the user and their preferred price range to generate recommendations, we are hoping to show the user items that are similar to what they have viewed in the past but still reflect what they are currently looking for. We were, however, unable to generate realistic communities of users. If we had achieved this we would have had the opportunity to rank the user's search results not only based on what the user's preferences are but also the frequency of the products that other similar users had viewed.

We did succeed in creating a project that gathered data from a variety of sources and use that data to collectively search it. We also implemented a simple recommendation algorithm for users who log into their accounts. Finally, we were able to add functionality which allowed users to add custom products to Crafty's catalogue. This feature allows for greater coverage of different art supply stores, and also makes it easy for users to add in products their local art supply stores which might not have a website. Crafty's convenient features and the intuitive user interface make it a useful tool for people like us to quickly find the products we want for the best price.

Conclusion

In conclusion, there are a number of different ways that Crafty can be improved upon to make it more useful and easy to use. One such improvement would be to add product photos and increasing the number of sites scraped into our database automatically. We could also add more community features such as ratings for products and stores. The products themselves could be also tagged and used to form associations and recommendations as well, similar to what we saw in our second assignment when we categorized the Movies.

References

The sites that were crawled:
[1] www.deserres.ca
[2] store.abovegroundartsupplies.com
[3] www.artshack.ca
[4] www.jerrysartarama.com
[5] wallacks.com