The Ohio State University / STEM Exploration & Engagement Scholars Program

UPlanner: Data Extraction & Machine Learning to Simplify Locating Campus Opportunities

Stephen Wu, Department of Computer Science Engineering

INTRODUCTION

Problem

As an undergraduate student of a large university, there is a myriad of emails to read, events to attend, people to meet, organizations to join, or competitive opportunities to pursue. The inundation of this information often comes at a price; students, faculty, or staff may face exposure to an excess of information that simply is irrelevant to them. Furthermore, organizers of events, clubs, and scholarships may fail to properly connect information of their event to the proper audience at hand.

Solution

To address this this excess of information sources for campus opportunities, UPlanner was created as a web app to examine various sources to identify, extract, and categorize opportunities on campus. Users can then access this platform to view events catered to their preferences.

UPlanner uses web scraping with machine learning, automated job scheduling, database management systems, and a user approval process and front-end web application to simplify finding campus opportunities.

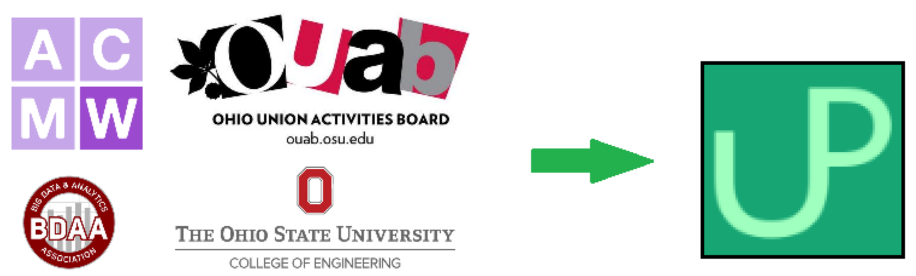


Figure 1. Illustration of information extraction from various sources to UPlanner

Compiling events from sources including:

- Association of Computing Machinery (ACM-W) mailing list
- Ohio Union Activities Board (OUAB) calendar
- Big Data Analytics Association (BDAA) calendar
- STEM Exploration & Engagement Scholars calendar
- Individual department websites

UPlanner seeks to create an everyday service to locate campus events, from movies to lectures to sporting events.

METHODS

Hackathon

UPlanner was first developed at the HackOHI/O 2016 hackathon (Nov 19-20, 2017) with the team of Stephen Wu, Ishan Taparia, Jacob Shoaf, and Stephen Pioro. There, the initial groundwork for the design and features of the product was developed.

Tech Stack

UPlanner is built on Node.js, React, Express, and MySQL. Node.js was chosen due to its increasing popularity and success in building highly scalable web applications. React, Facebook's component-based JavaScript Library, was chosen due to its speed in creating browser-rendered user interfaces. Express is the Node.js go-to tool for website routing and building APIs, and MySQL was chosen due to familiarity and reliability.

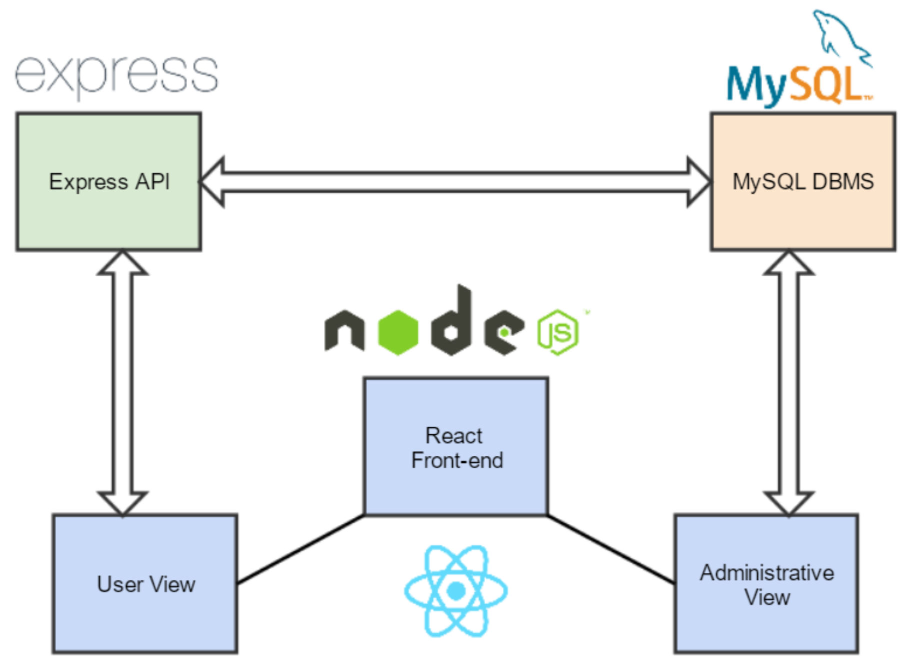


Figure 2. Platform diagram, describing tech stack and relationship between views.

Algorithm

To develop an algorithm to take into account user preferences, subject relationships, and event relevance, the flowchart in Figure 3 was created.

Machine learning in the form of a Naïve Bayes Classifier is to be used to classify events based on their source, text, and location.

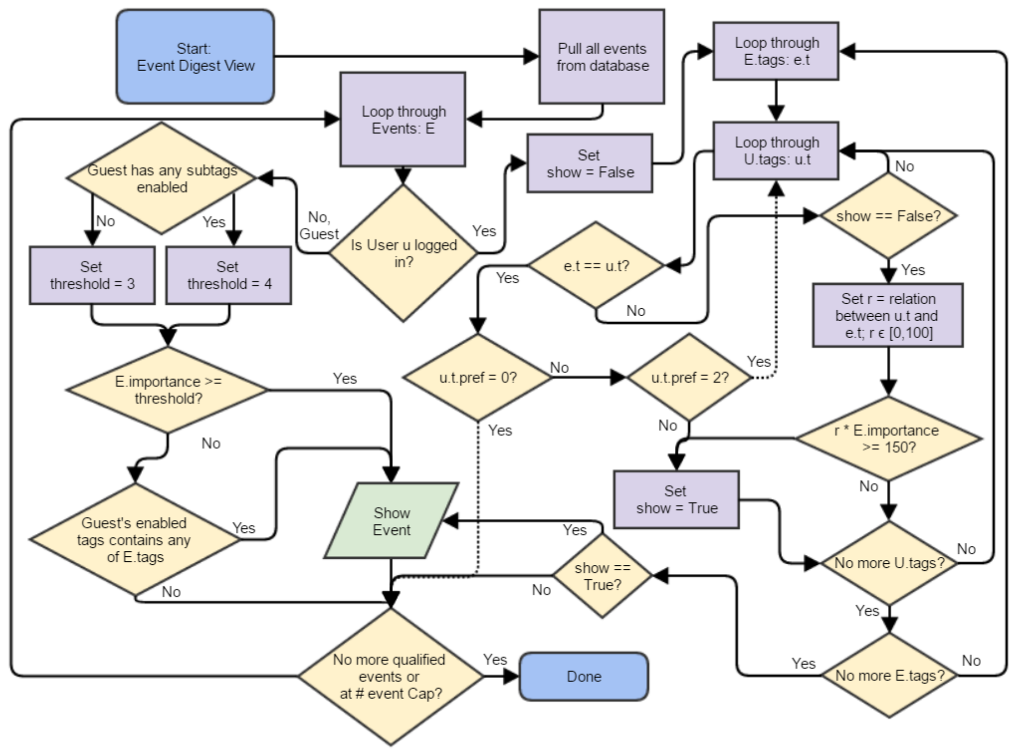


Figure 3. Flowchart depicting the algorithm applied to decide which events a user should see based on their settings

Parameters

E.tags - List of tags that event E belongs to e.g. {11,12,100} for {'Computer Science', 'Electrical Engineering', 'Undergraduates'}

E.importance - Relevance of event, 5 being the most relevant / broad, 1 being the least

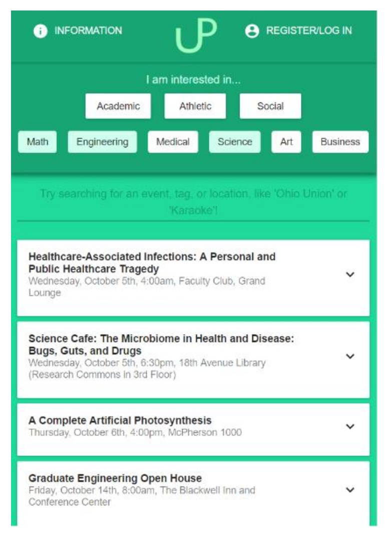
U.tags - List of tag preferences from the registered user. e.g. {[11,2], [12,2] [14,0], [18,2], [100,1], [101,0]} The key-value pairs are tag and preference; 0 = dislike, 1 = default, 2 = like, referenced by "u.t.pref"

r - relationship between two tags These are defined in a table based on a tree-like structure, ranging from [0,100] e.g. Electrical Engineering and Computer Science might have r=75. Music and Business may have r=0.

RESULTS

Hackathon

By the end of HackOHI/O 2016, the prototype included several working API routes, a Material user-interface, and the MySQL database populated from the scraper's events from Teamup calendar.



Healthcare-Associated Infections: A Personal and Public Healthcare Tragedy
Wednesday, October 5th, 4:00am, Faculty Club, Grand Lounge

Science Cafe: The Microbiome in Health and Disease: Bugs, Guts, and Drugs
Wednesday, October 5th, 6:30pm, 18th Avenue Library (Research Commons in 3rd Floor)

A Complete Artificial Photosynthesis
Thursday, October 6th, 4:00pm, McPherson 1000

Graduate Engineering Open House
Friday, October 14th, 8:00am, The Blackwell Inn and Conference Center

Figure 4. Snapshot of mobile demo, March 2017 build

Post-Hackathon

Since the hackathon, various improvements have been made to the project and its algorithms to automate and improve the digest's personalization. The next major step is to implement the Naive Bayes classifier model to determine the tags of each event based on its content and source.

Major post-hackathon improvements include mobile-responsiveness, style changes, working guest subtag navigation, API expansion and caching, improved search, and a Heroku-deployable setup. The guest tag navigation follows the model in the next page to determine whether an event should appear. There is much more to be done in the users backend, web scraping, and classification.

CONCLUSION

The UPlanner prototype demonstrates the prospects of scraping different campus resources to be viewed in a single web app. The next steps of the project are expanding the development team, developing the machine learning model, and obtaining faculty support.

From there, the application will be tested further for its utility at The Ohio State University by enlisting beta testers and feedback. The goal is for the product to be useful without significant administrator or moderator support, so developing a strong scraper and machine learning model is key in accurate predictions. If a beta test is successful, the prototype may be expanded into a fully-fledged product.

ACKNOWLEDGEMENTS

Thanks to the other original hackathon team members--Ishan Taparia, Jacob Shoaf, and Stephen Pioro—for their work and input.
Thanks to the HackOHIO team for putting on a great hackathon and a great place to start a project.
Thanks to faculty members Jorge Eduardo Mendoza and Dr. Tom Mitchell for general career and project advice.