Patent Recommendation System Report

CMPE 256 - Summer 2019
Team Project

Team 7 Members
Aaron Lee (ID: 009085596)
Hongfei Xu (ID:011833978)
Juan Chen (ID: 012483250)
Xiaoting Jin (ID: 013842192)

Index

1. Motivation and Problem Understanding

1.1. Motivation

The World's Intellectual Property Organization (WIPO) describes the intellectual property as the 'creations of the mind, such as inventions; literary and artistic works; designs; and symbols, names and images used in commerce.' People who hold patents have the rights to the invention for 20 years in the United States and they can use it to their benefit for their own financial gains.

Applying for patents, developing a new invention, and getting it approved can be a long and daunting task. According to WIPO, there have been over 6 million patents that were filed during the years of 2016 and 2017. Verifying that you are developing a product or have an idea that has not been proposed yet will take a long time to validate. When the patent gets audited, the United States Patent and Trademark Office also needs to approve it by also searching their databases and verifying that there is not another idea or product that has it copyrighted.

A patent recommendation system will be a good tool to search the database. It will give recommendations on patents that fit the need of the users based on their requirements and also personalized to fit their profile. As a result, this system's goal is to decrease the time used in searching for patents for both the applicants, people searching for patents, and the patent approval auditors.

1.2. Approach

For this patent recommender system, the data that makes up this system will be extracted from the United States Patent and Trademark Office (USPTO) and PatentsView. With this data, a number of features and attributes from each patent are extracted to fit the recommendation model.

After collecting data, we start designing a pipelined hybridization model. In a hybrid recommendation system, it uses a combination of different inputs and recommendation mechanisms -- in this case, content-based, knowledge-based, pipeline hybridization and collaborative filtering strategies.

The content-based strategy is based on the TF-IDF of the title and abstract of each patent. The knowledge-based recommendation will be based on the users'

external inputs. By using the search feature and explicitly selecting a CPC classification category, the knowledge-based and content-based models are cascaded that make up the pipelined hybrid model. The algorithm that was created will take these features and output a list of patents to recommend to the users.

From the list of patents that are generated, more patents can be recommended based on content-based or item-based collaborative filtering depending if a user selects that option on the web application. If a user is logged into an account, the user-based collaborative filtering technique is used to give a set of recommendations, based off of the user's previous history.

2. Dataset

2.1. Data Collection

The dataset that was used comes from www.patentsview.org, which is sourced from USPTO. To access and export the information of all the patents, PatentsView provides bulk data to be downloaded and exported into a tsv file. The following datasets were downloaded:

- ❏ Patent
- ❏ Patent Inventor
- ❏ Patent Assignee
- ❏ Patent Lawyer
- ❏ Inventor
- ❏ Assignee
- ❏ Lawyer
- ❏ Location
- ❏ US Patent Citation
- ❏ CPC_Current

There are 55 relevant datasets in total and in this project, according to our search application design, we mainly focus on patent data, assignee data, inventor data, lawyer data, and US Patent Citation data. The raw data resulted in 6,957,998 patents with many different attributes.

2.2. Data Processing

Data preparation needed to be done to remove duplicate and invalid data using Python's pandas library. Further processing of the data was done to fit the various models.

The following figures indicate the basic information of "patent.tsv" and "cpc_current.tsv". Since the "patent.tsv" raw data contains most of the critical information of patents, we took it as our main dataset and processed other preparation steps on top of it.

```
df.tail()
```

	id	type	number	country	date	abstract	title	kind	num_claims	filename	withdrawn
6957994	T998013	defensive publication	T998013	US	1980-09-02	NaN	Protection of insect pheromones from degradati...	I4	1.0	pftaps19800902_wk36.zip	0.0
6957995	T998014	defensive publication	T998014	US	1980-09-02	NaN	Thiazolyl couplers, coupler compositions and p...	I4	3.0	pftaps19800902_wk36.zip	0.0
6957996	T999001	defensive publication	T999001	US	1980-10-07	NaN	Sack handling device	I4	1.0	pftaps19801007_wk41.zip	0.0
6957997	T999002	defensive publication	T999002	US	1980-10-07	NaN	Application of polymeric powders to a substrate	I4	7.0	pftaps19801007_wk41.zip	0.0
6957998	T999003	defensive publication	T999003	US	1980-10-07	NaN	Shifted photographic dyes and compositions, el...	I4	3.0	pftaps19801007_wk41.zip	0.0

```
cpc_current.head()
```

	uuid	patent_id	section_id	subsection_id	group_id	subgroup_id	category	sequence
0	01bid55nrangfpc2alel0ph2c	4205226	H	H01	H01J	H01J49/025	inventional	1
1	mhl6idj9h404kix2s1480wqfq	4205226	H	H01	H01J	H01J49/482	inventional	2
2	8tzyinp2t0ty6doksz56kx4e7	4205227	H	H01	H01L	H01L27/153	inventional	0
3	wkhck9vs2a1ntcy1d3x4abyja	4205227	G	G08	G08B	G08B13/19	inventional	1
4	x17dr1r14jwfn1iqxuavpbz4m	4205227	H	H01	H01L	H01L31/147	inventional	2

Figure 1: Patent Basic Information and CPC Sections

The CPC section is an important categorization indicator that represents the classification of the patent, and there are nine sections of A-H and Y, so we needed to merge it into the main dataset. However, we found that a patent may belong to multiple CPC sections, so we can't just use one letter to represent the value of this attribute. To solve this problem, we treated the 9 sections as separate attributes and labeled them with dummy variables, then used the pivot table to expand and merged them into the main dataset.

	id	A	B	C	D	E	F	G	H	Y
0	3930271	1.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	3930272	1.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.0
2	3930273	1.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	3930274	NaN	1.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	3930275	1.0	1.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Figure 2: CPC Section Values after Binary Encoding

Another important piece of information is patent-related personnel and organizations, including inventors, lawyers, and assignees. According to our statistics, a patent may have multiple inventors, lawyers, and assignees. Similarly, these groups may also correspond to multiple patents. Hence we created datasets corresponding to the patent ID for these individuals' information, in order to facilitate patent search and reverse search.

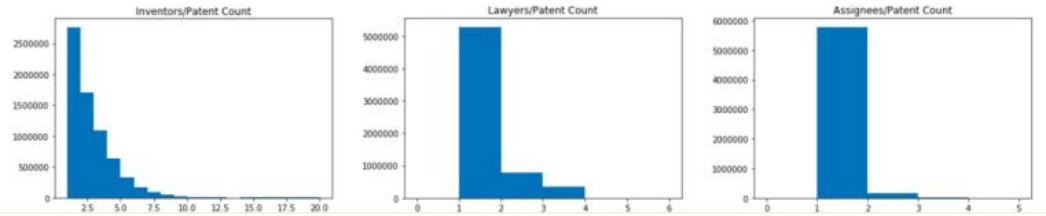

Figure 3: Ratios of Patents and Inventors/Lawyer/Assignees

2.3. Feature Extraction

By using the group counting function provided by Pandas Dataframe, it is found that the values of attributes "country" and "type" are fixed ("US" and "utility"), therefore they can't be applied in our system since they are not distinguishable for patents.

In addition, "filename", "withdrawn" and "number of claims" properties have no cognitive meaning for the public, so we also removed them from our dataset to improve the processing speed of systems.

After that, we set patent ID as index and extracted the names of the inventors, lawyers and assignees corresponding to each ID, followed by merging them into the main dataframe in the form of lists.

id	date	abstract	title	kind	num_claims	A	B	C	D	E	F	G	H	Y	inventor_name	lawyer_name	assignee_na
10177267	2019-01-08	An UV photodetector includes: a substrate, a t...	Photodetector	B2	14.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.0	1.0	['Jianping Zhang', 'Ling Zhou', 'Ying Gao']	[' ,J.C. Patents']	[' ,BOLB IN
10182289	2019-01-15	An earpiece (100) and acoustic management modu...	Method and device for in ear canal echo suppre...	B2	12.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.0	NaN	['Steven Goldstein', 'John Stanley Usher', 'Ma...	['Peter A. Chiabotti,', 'Mammen (Roy) P. Zacha...	[' ,Sta Techiya, LL
10184301	2019-01-22	An embodiment includes a downhole tool with fi...	Downhole drilling tools and connection system ...	B2	30.0	NaN	NaN	NaN	NaN	1.0	NaN	NaN	1.0	NaN	['Guy A. Daigle', 'Daniel E. Burgess', 'Carl A...	[' ,Offit Kurman, P.A.', 'Gregory A. Grissett,']	[' ,A Technolc In
10185309	2019-01-22	In one embodiment, a tangible, non-transitory ...	Systems and methods for recommending component...	B2	23.0	NaN	NaN	NaN	NaN	NaN	1.0	1.0	1.0	1.0	['David A. Vasko', 'Matthew R. Ericsson', 'Kel...	[' ,Fletcher Yoder P.C.']	[' ,Rocku Automat Technolog In

Figure 4: Overview of Dataset Combining Inventor, Lawyer and Assignee Information

3. Solution Implementation

The following figure shows the system design for the whole project. On the user interface, users can search and get the recommendation list by our pipelined hybridization system, which consists of knowledge-based strategy and content-based in order. After getting the recommended results, the user can also query similar items for each result patent separately by the approaches of content-based cosine similarity of item-based collaborative filtering.

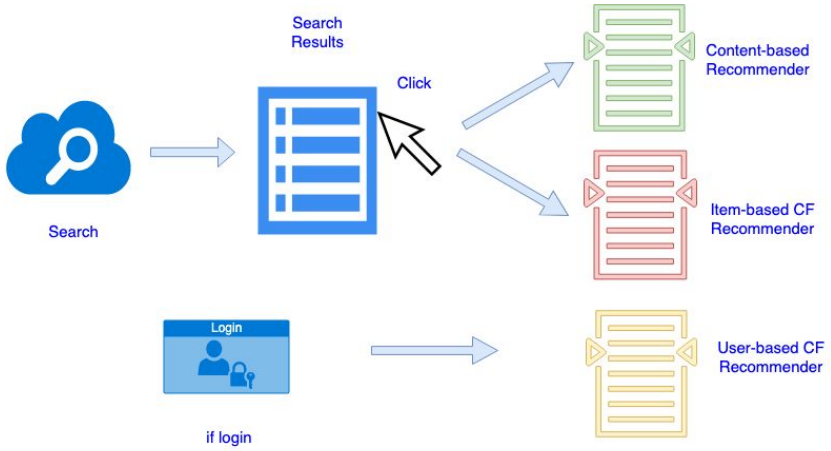

Figure 5: Whole System Design

3.1. Hybridization System

As mentioned earlier, our hybrid system consists of two recommendation strategies - knowledge-based and content-based. In the following part, we will

describe the implementation of these two modules and the hybridization process.

3.1.1. Content-based Recommender

Content-Based recommender is designed to find the most relevant patents related to the keyword search that the user inputs on the UI. The algorithm for this recommender system will go through the title and abstract of each patent and generate a TF-IDF value for each word.

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

Before vectorizing all the vocabulary, we dealt with the target contextual attributes - title and abstract, including removing punctuation, unnecessary numbers, and stop-words of English. In this process, we used NLTK (Natural Language Toolkit) package.

	id	date	abstract	title
0	10000000	2018-06-19	A frequency modulated (coherent) laser detecti...	Coherent LADAR using intra-pixel quadrature de...
1	10001015	2018-06-19	Airfoil and hydrofoil systems include structur...	Drag reduction systems having fractal geometry...
2	10002022	2018-06-19	A method, a computer program product, and a co...	Processing interrupt requests
3	10003026	2018-06-19	A ladder tetrazine polymer is disclosed.	Ladder tetrazine polymers
4	10004044	2018-06-19	[Object] To achieve both prevention of harmful	Communication control apparatus and wireless c

	id	date	abstract	title
0	10000000	2018-06-19	frequency, modulated, (coherent), laser, detec...	coherent, ladar, using, intra-pixel, quadratur...
1	10001015	2018-06-19	airfoil, hydrofoil, systems, include, structur...	drag, reduction, systems, fractal, geometry/ge...
2	10002022	2018-06-19	method,, computer, program, product,, computer...	processing, interrupt, requests
3	10003026	2018-06-19	ladder, tetrazine, polymer, disclosed.	ladder, tetrazine, polymers
4	10004044	2018-06-19	[object], achieve, prevention, harmful, interf...	communication, control, apparatus, wireless, c...

Figure 6: Comparison of Original Texts and Texts after Processing

The tool we used for calculating TF-IDF value is scikit's TfidfVectorizer python module. Every patent in TF-IDF matrix is a vector consists of all words on their own axis.

	aa	aad	abbe	abc	abdominal	aberrations	ablatable	ablation	abnormalities	aborting	...	zippering	zippers	zirconia	zone	zones	zoom	zooming	zz
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	

Figure 7: Head Part of Title's TF-IDF Matrix

The values of TF-IDF matrix work in two functions of our whole system. In hybridization system, our content-based strategy calculates the sum of TF-IDF values for all the keywords from both title and abstract matrix with different weights (0.8 for title and 0.2 for abstract), which can be expressed as:

$$Content\ Score = 0.8 \cdot \sum_{i}^{i \in title} tfidf\_value + 0.2 \cdot \sum_{i}^{i \in abstract} tfidf\_value$$

The content scores will be used for assigning scores in our cascade pipeline system.

```
tfidf_weighted_words('data analytics')
```

	id	date	abstract	title	kind	num_claims
3636	7352993	2008-04-01	A data reproducing system of the invention inc...	Data reproducing apparatus and data reproducin...	B2	20.0
3938	7656400	2010-02-02	A method and a device for converting data of a...	Image data editing device and method, and imag...	B2	24.0
3324	7039656	2006-05-02	A system for synchronizing data records betwee...	Method and apparatus for synchronizing data re...	B1	13.0
2951	6665283	2003-12-16	A wireless communication system transmits data...	Method and apparatus for transmitting data in ...	B2	17.0

Figure 8: Test Case of Searching Keywords ("data analytics") with TF-IDF Values

The other function we make use of TF-IDF matrix is recommending similar items for selected patent. Using the TF-IDF values we generated, the sum of cosine similarity is calculated by the same weights as hybridization system:

$$similarity\ (a, b) = 0.8 \cdot cosine\_sim\ (a, b) + 0.2 \cdot cosine\_sim\ (a, b)_{abstract}$$

This function will return top five similar patents ranked by similarity scores, which represent the most relevant patents related to a certain patent selected by user.

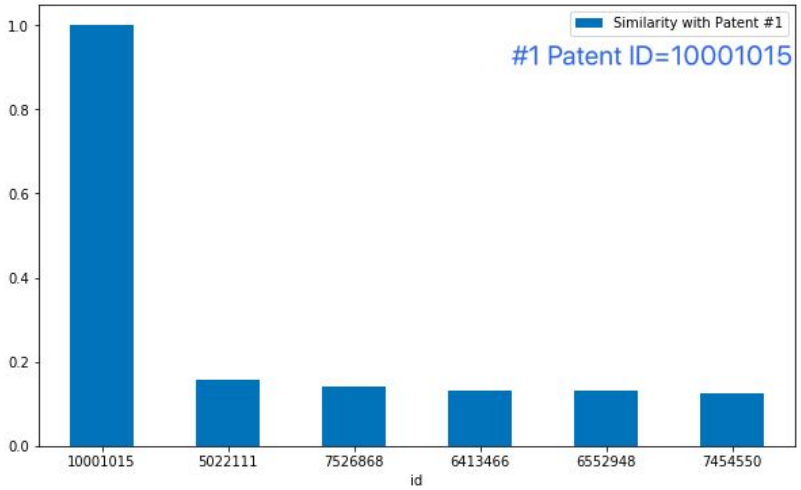

Figure 9: Test Case: Top 5 Similar Patents of Patent(id=10001015)

3.1.2. Knowledge-Based Recommender

In this module, the features we selected are date, CPC section, and information about patents and organizations (including inventors, lawyer and assignees). Specifically, our system uses constraint-based strategy, of which the formula can be expressed as:

$$CSP = (X_i \cup X_u,\ D,\ SRS \cup KB \cup I)$$

In order to interface the user requirements with the patent features, we have built a set of models: we use the Epoch/Unix timestamp numerical comparison to meet the user's request for the registration time of the patent; use the function of searching for the same byte in the Dataframe to recommend a patent-related person/organization that meets the user's standard; and express the CPC section in an easy-to-understand way (A=Human necessities , B=Performing operations or transporting, C=Chemistry or metallurgy, D=Textiles, E=Fixed constructions, F=Mechanical engineering or lighting or heating or weapons or blasting engines or pumps, G=Physics, H=Electricity, Y=General tagging of new technological developments), which is convenient for the user to select.

Finally, the constraint-based function will look for a patent intersection subset in the domain that meets all of the user's requirements and return it as a list of recommendations.

```
search_patent(None, '1980-10-1', '2008-01-01', 'B2', 'G', 'Robert', None, None)
```

	id	date	abstract	title	kind	num_claims	A	B	C	D	E	F	G	H	Y	inventor_name	lawyer_name	assig
2732	6445553	2002-09-03	A system and method for providing a device for...	Method and system for fabricating a high densi...	B2	19.0	NaN	1.0	NaN	NaN	NaN	NaN	1.0	NaN	NaN	['Robert Rottmayer', 'Ronald A. Barr']	[' ,Sawyer Law Group LLP']	[' Co
3020	6734490	2004-05-11	The memory cell is formed in a body of a P-typ...	Nonvolatile memory cell with high programming ...	B2	6.0	NaN	NaN	NaN	NaN	NaN	NaN	1.0	1.0	NaN	['Roberto Bez', 'Alberto Modelli', 'David Esse...	[' ,Seed IP Law Group PLLC', 'Lisa K. Jorgenso...	,STMicro
3025	6739502	2004-05-25	A computer program for managing property using...	Property management system	B2	23.0	NaN	NaN	NaN	NaN	NaN	NaN	1.0	NaN	NaN	['Robert Michael Gruber']	['David S. Kalmbaugh,']	[' , States as re
3200	6915215	2005-07-05	Embodiments of the invention generally encompa...	Integrated low power digital gyro control elec...	B2	30.0	NaN	NaN	NaN	NaN	NaN	NaN	1.0	NaN	NaN	['A. Dorian Challoner', "Robert Thomas M'Closk...	['Bradley K. Lortz,', ' ,Origin Law']	[' Technolo
3312	7027623	2006-04-11	A method and system for providing owners, pote...	Apparatus for capturing an image	B2	82.0	NaN	NaN	NaN	NaN	NaN	NaN	1.0	NaN	NaN	['Richard M. Sutherland', 'Richard Peter McWil...	[' ,Pillsbury Winthrop Shaw Pittman LLP']	[' ,The U Comp

Figure 10: Test Case: Get Recommendation List by explicit constraints

3.1.3. Pipelined Hybridization

We adopted a cascade model for our pipelined hybrid system, which means we use the first strategy to reject items that don't fulfill the user's preferences, followed by assigning scores by successor strategy. The following figure shows the workflow of our hybridization recommender.

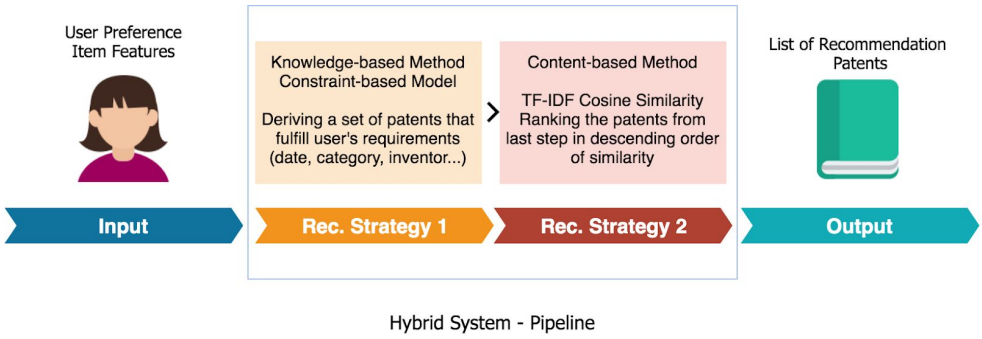

Figure 11: Design of Hybrid Pipeline System

As shown in the workflow, out constraint-based function stated in part 3.1.2 will work as a filter to obtain patents meets the user's requirements on the granted date, individual/organization name, and CPC classification. The fulfilled subset will then be passed to the content-based function and assigned a score based on the sum of TF-IDF values. The final result of this pipelined hybrid system is a set of filtered recommendation patents sorted by content-based score.

```
hybrid_pipe(None, '1995-01-01', None, None, None, 'Joseph', None, None, 'laser')
```

	id	date	abstract	title	kind	num_claims	A	B	C	D	E	F	G	H	Y	inventor_name	lawyer_name	assigne
2614	6327090.0	2001-12-04	A system which is preferably employed in a la...	Multiple laser beam generation	A	25.0	NaN	NaN	NaN	NaN	NaN	NaN	1.0	1.0	NaN	['Joseph F. Rando', 'Timothy J. Litvin']	['Thomas M. Freiburger,']	[' Tec
2919	6633196.0	2003-10-14	An integrated circuit die includes a bond pad ...	Device and method for limiting the extent to w...	B2	31.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.0	NaN	['Joseph C. Sher']	[' ,TraskBritt']	[' Tec
3448	7164096.0	2007-01-16	A method for the fabrication of large metal ma...	Continuous metal matrix composite consolidation	B1	13.0	NaN	1.0	1.0	NaN	NaN	NaN	NaN	NaN	1.0	['Brian E. Joseph', 'Brian L. Gordon', 'James ...	['Philip D. Lane,']	[' ,Tou F Lat
0	10000000.0	2018-06-19	A frequency modulated (coherent) laser detecti...	Coherent LADAR using intra-pixel quadrature de...	B2	20.0	NaN	NaN	NaN	NaN	NaN	NaN	1.0	NaN	NaN	['Joseph C. Marron']	[' ,Munck Wilson Mandala, LLP']	[' ,R Co

Figure 12: Test Case: Get Recommendation List by Content "laser" with requirements of registering after 1995 and Inventor Name "Joseph"

3.2. Collaborative Filtering

3.2.1. User-Based Recommender

If the user has logged in, our system could save users' click history data and we would use click data to empower user-based recommender. The idea behind it is very straightforward, if the click histories of two users are highly overlapped, they are considered to be more similar. To measure the similarity, we use Pearson Correlations. After calculating the similarity matrix, we can get the nearest neighbor of the user and will recommend those patents that have been clicked by the nearest neighbor and have not been clicked by this user.

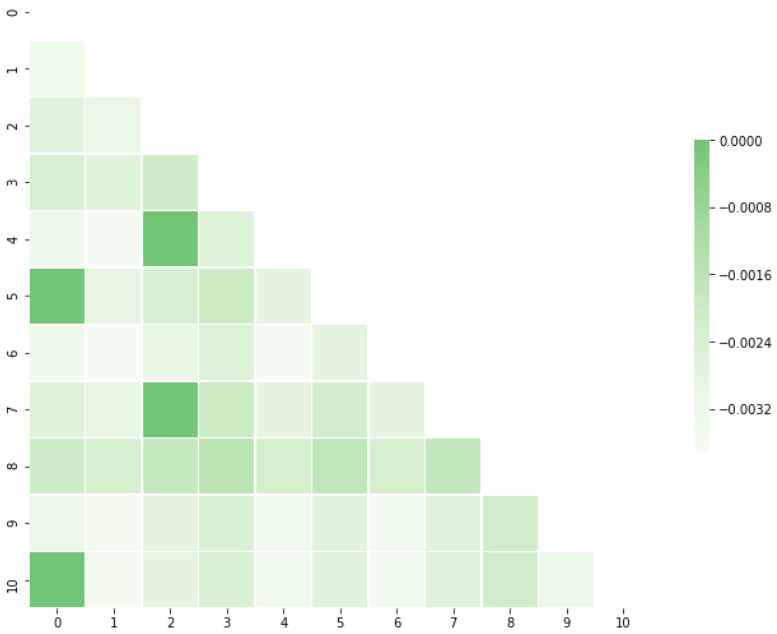

Figure 13: Pearson Correlation of Nine Sample Users in our System

3.2.2. Item-Based Recommender

We use classic Jaccard index to measure the similarity based on citations of two patents.

$$J_{(A,B)} = |citations(A) \cap citations(B)| / |citations(A) \cup citations(B)|$$

One-hot encoding is used where each patent is a dimension and a binary vector represents a patent and its citations. However, the massive number of total cited patents will result in very high dimension and the data is extremely sparse. So sparse matrix is used to store data and perform matrix operations. The sparse matrix is pre-generated and stored to save execution time.

```
citations of patent: 4079741
```

	patent_id	citation_id
61069	4079741	2004581
61070	4079741	2900661
61071	4079741	1743590
61072	4079741	2083380
61073	4079741	1232617

```
citations of patent: 3943599
```

	patent_id	citation_id
62757	3943599	1452098
62758	3943599	2004581

```
: index1 = patent_index[patent_index.patent_id == patent_id_1].iloc[0]['patent_index']
  index2 = patent_index[patent_index.patent_id == patent_id_2].iloc[0]['patent_index']
  print("similarity between patent: %d and patent: %d is:" %(patent_id_1, patent_id_2))
  print(sims[index1, index2])

  similarity between patent: 3943599 and patent: 4079741 is:
  0.16666666666666666
```

Figure 14: Test Case: Similarity between Patent 4079741: and Patent: 3943599

4. User Interface (Application)

4.1. Overview

We separated the application into two major components: The web app and the recommendation service.

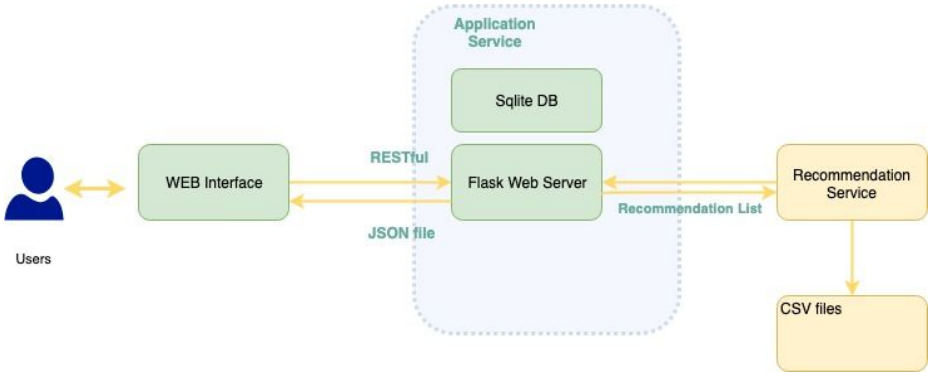

Figure 15: System Design of Web Application

4.2. Web Component

The web components read user inputs and display the recommendation result using dynamically generated HTML pages. The application will import the Recommendation Components as a separate Python Package for better extensibility in the future.

WEB Components: Composed of User Interface, Database, Database ORM, Web Framework, Templating Engine. In our case, we used Sqlite for database, Flask as the web framework, Flask-Sqlalchemy for ORM, Jinja for Template.

4.3. Database

The data stored in DB including two major parts: User account information and User clicking history data. Database schemas are as follows:

UserAccount Table

Username(PK)	password

Primary Key: username or email

User clicking History Data Table

id (PK)	username(FK)	patent_id

Primary Key: id
Foreign Key: username

4.4. Screenshots of the User Interface

4.4.1. Hybrid Recommendation- Pipeline Cascade

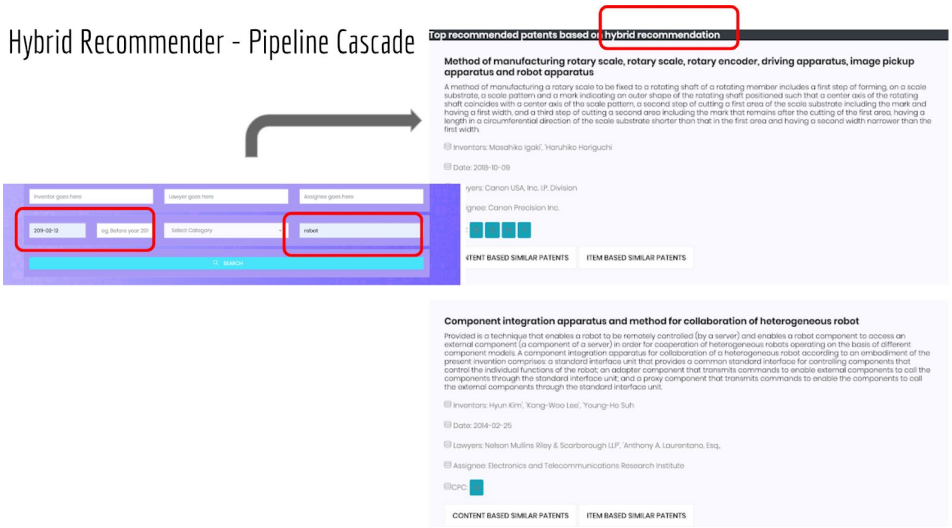

Figure 16: Example of Hybrid Recommendation

4.4.2. User Based CF Recommendation Based on Users' Clicking History

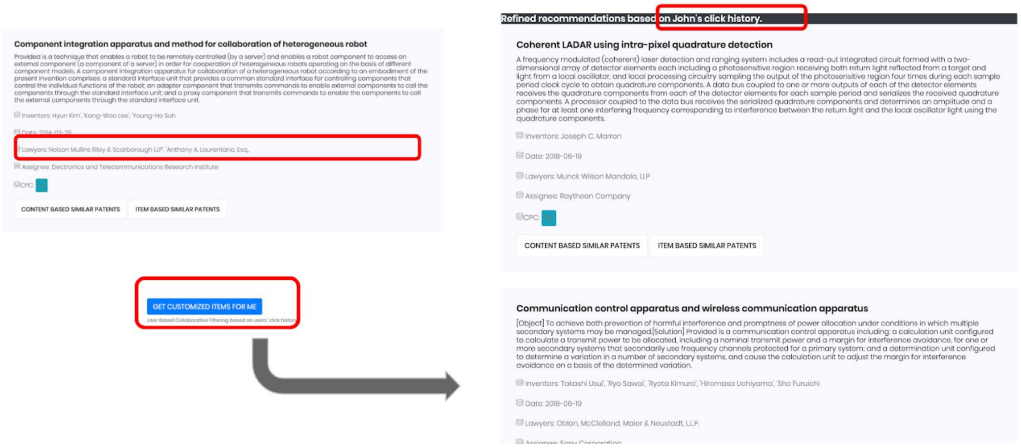

Figure 17: Example of Collaborative Filtering Recommendation

Prerequisite: Registered User and Logged In

Step 1: Log in using username and password

Step 2: View details of patent by clicking on patent title

Step 3: Click "GET CUSTOMIZED" button on the bottom of the page to store click history and get customized user-based recommendations

4.4.3. Content-Based Recommender (TF-IDF)

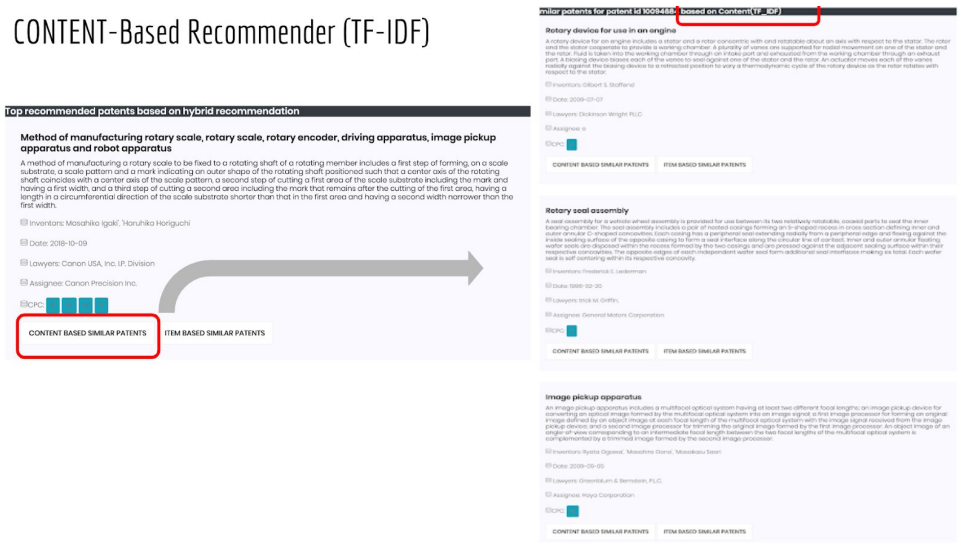

Figure 18: Example of Content-Based Recommendation

4.4.4. Item-Based Collaborative Filtering

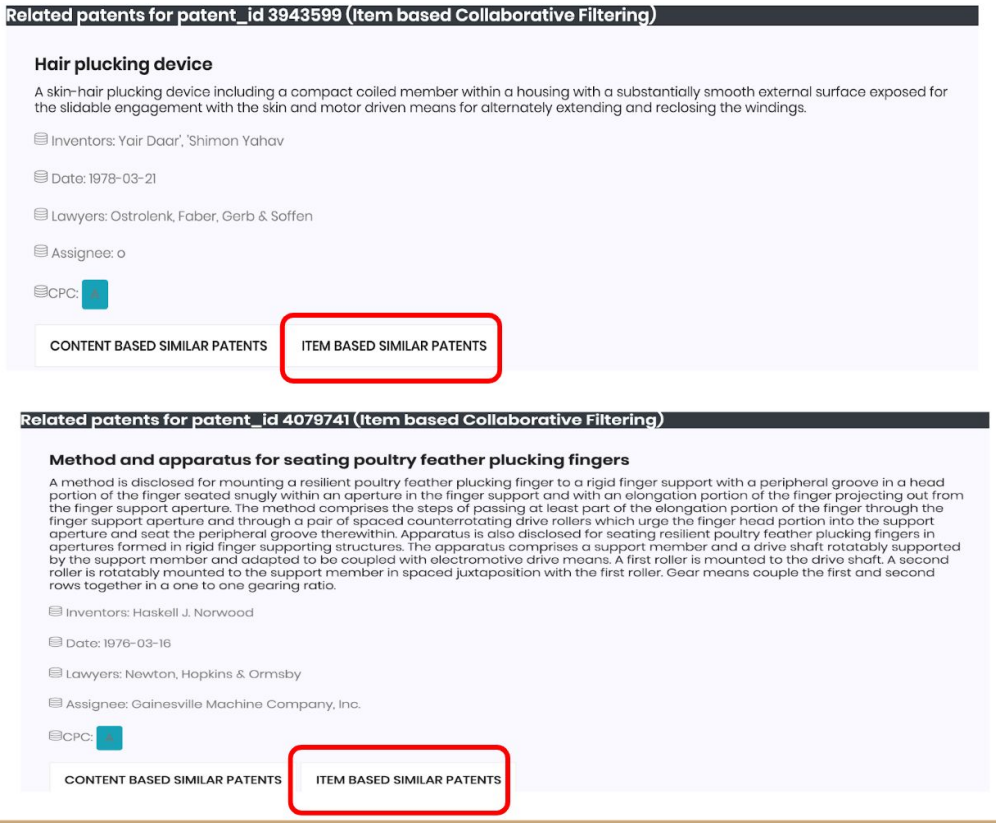

Figure 19: Example of Item-Based Collaborative Filtering

5. Evaluation

To develop this Patent Recommendation System, various recommending techniques that we learned over this course were applied. The methods used included content-based, knowledge-based, item-based, and user-based recommendations. With these recommendations, this hybrid pipelined system is able to produce recommendations for the users based on their requirements and their personal preferences.

With a user-friendly interface in a web application, it is easy to navigate and it includes some key details needed to understand the recommended patents.

Drawbacks:
If one individual is running a recommendation while logged in to someone's profile, with a user's prior click and search history, recommendations will be generated based off of their profile rather than another individual's personal taste. Since our data of users is only obtained from the registered individuals in our app database, while our application is not yet available to the public, so the User-based results may lack accuracy.
Another drawback is that the citation data of patents doesn't work very well for item-based CF model since the common citations between any pair of patents are very few, and the citations of most patents do not intersect, so users may sometimes get zero result through item-based CF model.

6. Impact

The patent application process can take on average, 32 months, according to Erickson Law Group. The goal for this Patent Recommendation Systems is to reduce the time used in researching and also making searches, whether its by patent approvers, researchers, or patent applicants.

With the recommendation system, the generated recommendations provide a brief overview on what the patent is about. Rather than having to go through full patents individually and trying to figure out what each design or invention protects, this system will programmatically find a correlation based on our model and filter out patents that have low correlation.

Links

GitHub: https://github.com/IdleDust/patent-recommend
YouTube: https://youtu.be/c1Gmq4IN48c