Patent Recommendation System

TEAM #7
Aaron Lee, Hongfei Xu
Juan Chen, Xiaoting Jin

Motivation

- Patent - Intellectual property provide rights only to the owner to make/use/sell/import an invention for 20 years

- Record number of patents being filed across the world
    - 2017: 3.17 million
    - 2016: 3.13 million

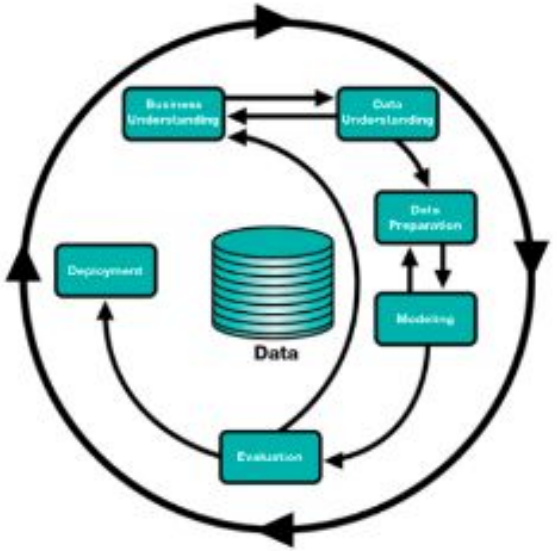
Motivation (continued)

- Current Approach: Companies manually search the following databases:
  - United States Patent and Trademark Office (USPTO)
  - European Patent Office (EPO)
  - Taiwan Intellectual Property Organization (TIPO)
  - World Intellectual Property Organization (WIPO)

- Prevent conflict with Intellectual Property (IP)
  - Reduce wasted resources on irrelevant searches for R&D
  - USPTO examiner to identify and invalidate pre-existing inventions/technologies quicker

How to Approach this Problem?

- Project Plan
  - Problem Understanding -- 1 week
  - Data Gather and Processing -- 2 weeks
  - Model Implementation -- 2 weeks
  - Solution Evaluation -- 1 week
  - Result & Impact Discussion -- 1 week

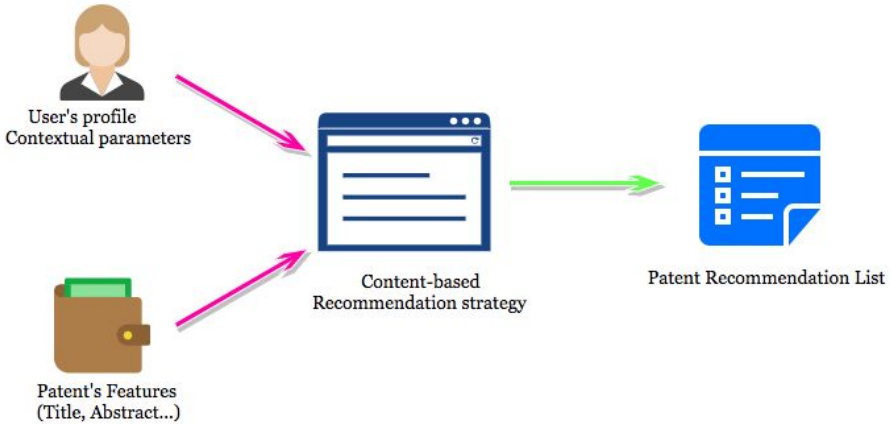


How to Approach this Problem? (Continued)

- Work Done so far
  - Motivation & Problem Discussion
  - Data Understanding & Processing
    - Acquiring data from appropriate sources
    - Exploration on characteristics of the data
    - Dataset Preparation
  - Model Researching
    - Which methods are appropriate?

How to Approach this Problem? (Continued)

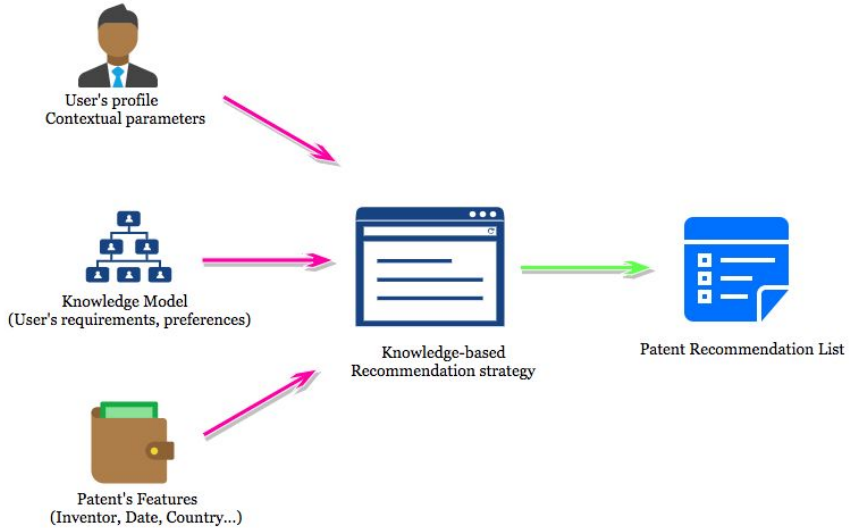
- Work in Progress
  - Model Implementation
    - Content-based Recommendation



Target: Retrieve patents similar to those the user preferred

How to Approach this Problem? (Continued)

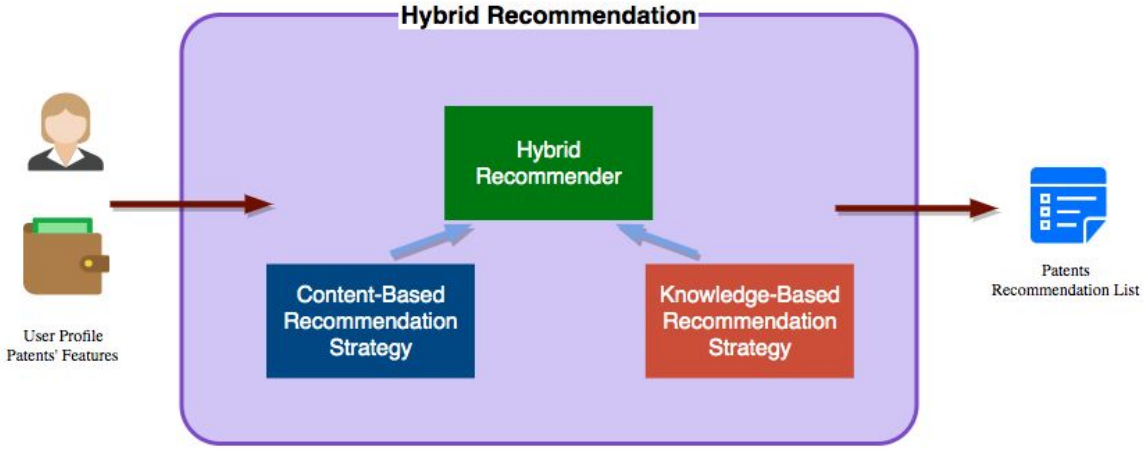
- Work in Progress
  - Model Implementation
    - Knowledge-based Recommendation



Target: Derive a set of patents that fulfill the user's requirements and constraints

How to Approach this Problem? (Continued)

- Work in Progress
  - Model Implementation
    - Monolithic Hybridization



Target: Acquire patent set that satisfy user's requirements with similarity scores

Data Processing

- **Data Sources**

  - Official Website of USPTO

    www.uspto.gov

    USPTO - Federal agency for granting U.S. patents and registering trademarks.
  - API & Bulk Data Provided by PatentsView

    www.patentsview.org

Data Processing (Continued)

- **Data Description**
  - So far we collected data of 6,957,998 patents.
  - There are dozens of attributes for each patent in raw data, including patent ID, title, type, country, abstract, date … …

```
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

Data Processing (Continued)

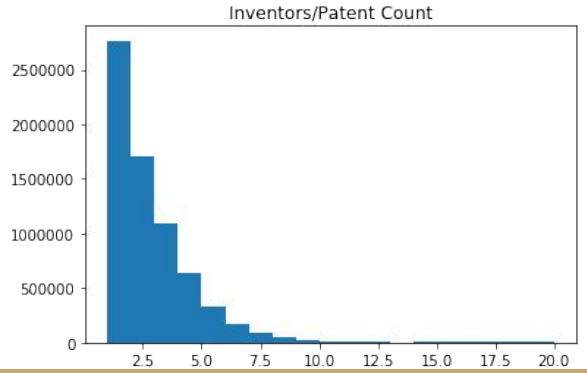
- Feature Exploration

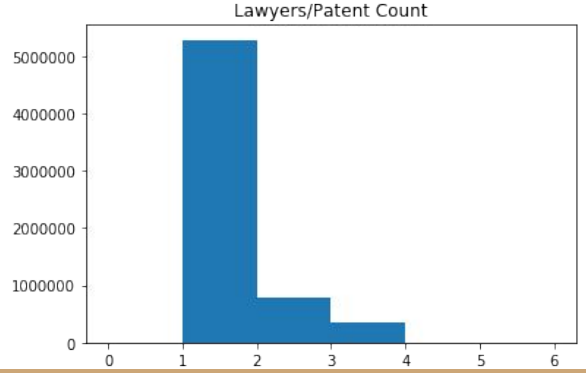
- Feature Extraction
  - Features for Content-Based: Context features such as title, abstract…
  - Features for Knowledge-based: Attributes can be used as constraints such as section, date, inventor, cpc classification, assignee...

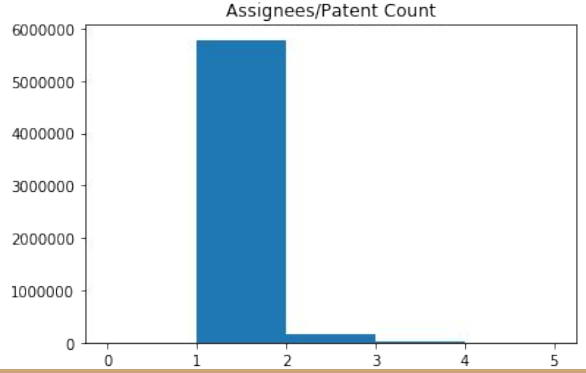
Feature	Description	Value Example
Patnet id	Patent unique ID	1000000
Type	category of patent. Usually "design", "reissue", etc.	utility
Country	country in which patent was granted	US
Date	date when patent was granted	1980-10-2
Abstract	abstract of patent	A golf glove is disclosed h...
Title	title of patent	Golf glove
CPC Section	cpc classification (A = Human Necessitites, B = Performing Operations; Transporting, C = Chemistry; Metallurgy, D = Textiles; Paper, E = Fixed Constructions, F = Mechanical Engineering; Lighting; Heating; Weapons; Blasting Engines or Pumps, G = Physics, H = Electricity, Y = General Tagging of New Technological Developments)	A
Inventor	Inventor(s) of the patent	Joseph C. Marron
Lawyer	Lawyer/Agent for the patent	Birgie E. Morris
Assignee	Assignee of the patent (individuals, companies, government...)	DNH Management LLC

Data Processing (Continued)

- Data Processing
  - Step 1: Cleaning invalid and duplicate data by Pandas
  - Step 2: Feature extraction and dataset preparation
    - Problem : A patent may be assigned to more than one CPC classification
      Solution: Encoding every CPC class by binary number
    - Problem : More than one inventor/lawyer/assignee may correspond to one patent
      Solution: Mapping patent id to every inventor/lawyer/assignee and their own features, followed by creating individual datasets for inventor/lawyer/assignee







Data Processing (Continued)

- Dataset Overview

	id	type	country	date	abstract	title	A	B	C	D	E	F	G	H	Y
0	10000000	utility	US	2018-06-19	A frequency modulated (coherent) laser detecti...	Coherent LADAR using intra-pixel quadrature de...	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
1	10001015	utility	US	2018-06-19	Airfoil and hydrofoil systems include structur...	Drag reduction systems having fractal geometry...	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0
2	10002022	utility	US	2018-06-19	A method, a computer program product, and a co...	Processing interrupt requests	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
3	10003026	utility	US	2018-06-19	A ladder tetrazine polymer is disclosed.	Ladder tetrazine polymers	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0
4	10004044	utility	US	2018-06-19	[Object] To achieve both prevention of harmful...	Communication control apparatus and wireless c...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0

Patent dataset with one-to-one mapping features

Data Processing (Continued)

- Dataset Overview

	patent_id	inventor_id	name_first	name_last
16493759	T998013	T998012-3	Helen C. F.	Su
16493760	T998012	T998012-4	Roy E.	Bry
16493761	T998013	T998012-4	Roy E.	Bry
16493762	T998012	T998012-5	Robert	Davis
16493763	T998013	T998012-5	Robert	Davis

	patent_id	assignee_id	type	name_first	name_last	organization
6158988	8192522	org_S7GmKDtdnrrXsl73aOLr	3.0	NaN	NaN	ET-Energy Corp.
6158989	4671471	org_foOFrge1L1SxWQu6BMaU	2.0	NaN	NaN	Mitchell Wing, Inc.
6158990	8571893	org_wCPzOy5fKzvO4PnNLZEn	2.0	NaN	NaN	Nihon Kohden America, Inc.
6158991	7314014	org_oAqU7fTqmEHjLCh1igiV	3.0	NaN	NaN	Heyring Technologies Pty Ltd.
6158992	D519427	org_klBJJASiYp61UrF4xm9z	2.0	NaN	NaN	DNH Management LLC

	patent_id	lawyer_id	name_first	name_last	organization
7848624	10186949	f096b46361f6143fe8d1cb5050409547	Daniel	Morrris, Esq.	NaN
7848625	4371738	5cb9ef19159e15dddaee5117bc9b2e92	Birgie E.	Morris	NaN
7848626	8374858	d649c64db7754d849acad2c5c8c651a0	Guarav	Mohindra	NaN
7848627	7898229	55199546539b052e1389d0cca8601702	NaN	NaN	Thompson Patent Law Offices, PC
7848628	4296604	efc85be5833eaf435f6c1b91d2eb32a5	NaN	NaN	Finnegan, Henderson, Farabow, et al.

Tables of Inventor (top left), Assignee (top right), and Lawyer (bottom)

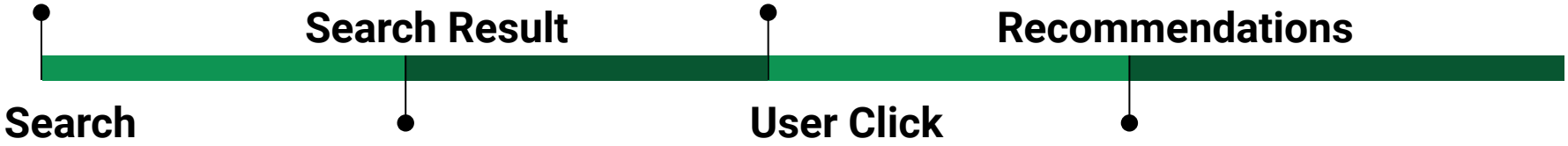
System Design

User Input Search Conditions

Date, Inventor Name, Assignee Name, Inventor Name, Keywords

User Click A Patent

User's click data will be stored and used for recommendations



Return Sorted Patent List

Keywords appear in title and abstract
Different Weight

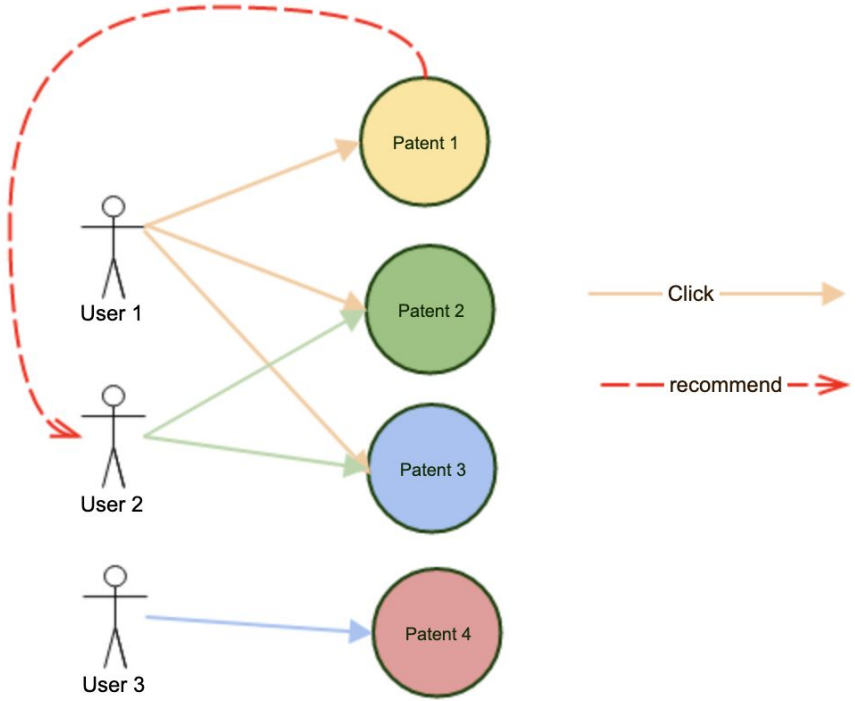
Patent Recommendation Based on User's Click Events

Hybrid Recommenders

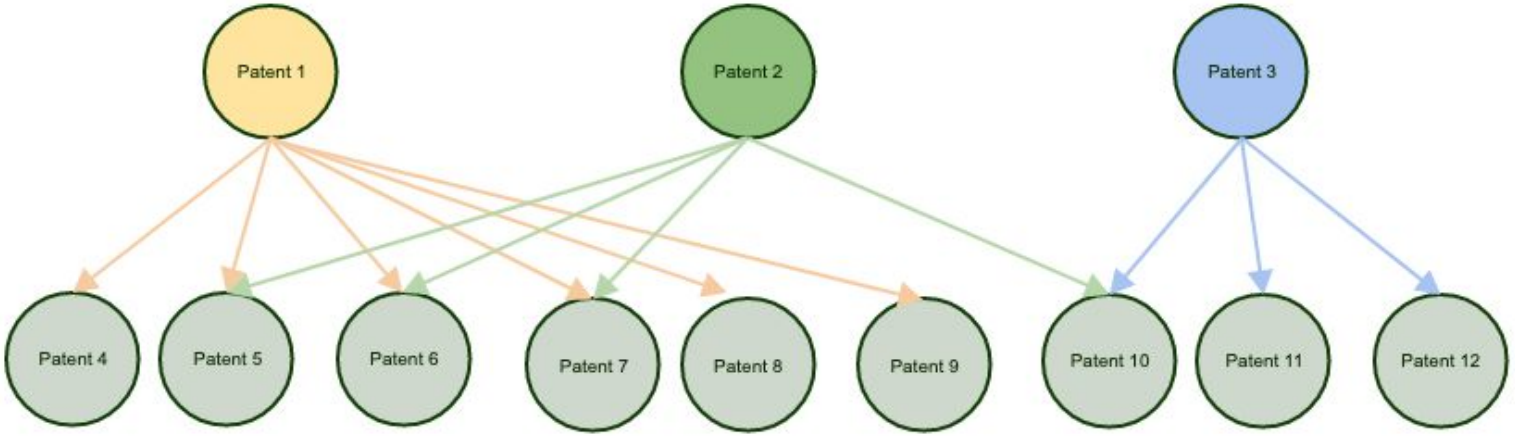
1. Content-based: TF-IDF, similarity between title / abstract
2. User-based Collaborative filtering
   Use user click data
3. Similarity from Citations: Jaccard Similarity

System Design (Continued)

User based filtering



System Design (Continued)



Jaccard Similarity based on Citations

sim(patent1, patent2) > sim(patent2, patent3)

Partial Implementation Result

- Content-based Recommendation System
  - Method: TF-IDF

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

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

  - Applied to features: Abstract, Title
  - Weights: Title (70%), Abstract (30%)

Partial Implementation Result (Continued)

- Content-based Recommendation System
  - Pre-Processing: Removing punctuations, numbers and stop words.
  - Created TF-IDF matrix

```
dftx.head()
```

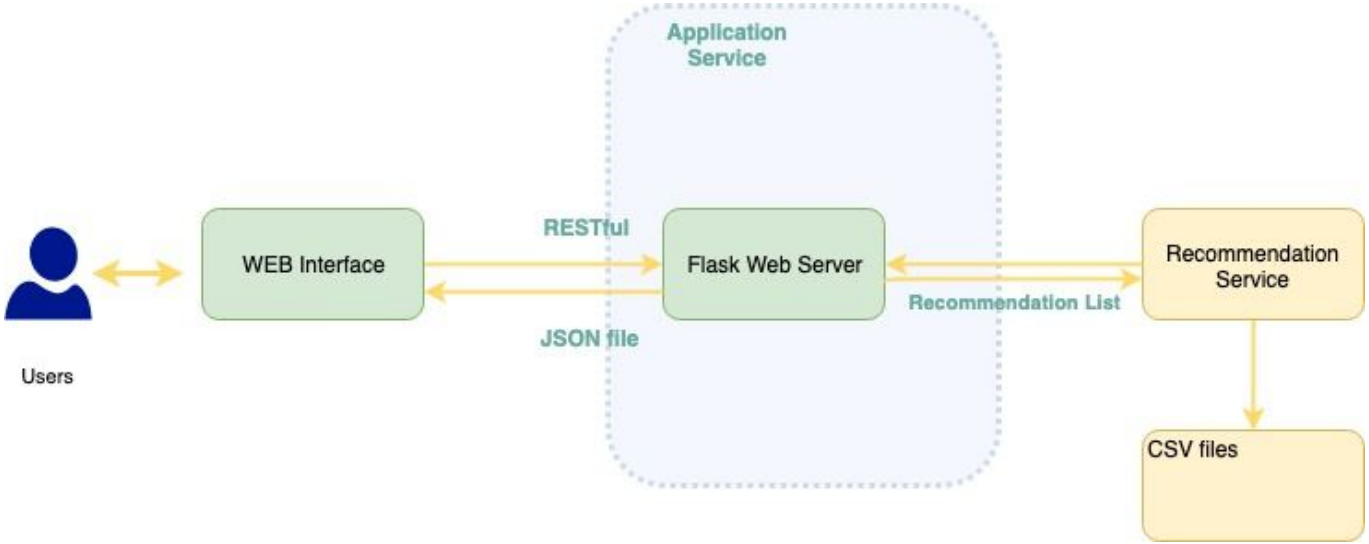
	aa	aad	abbe	abc	abdominal	aberrations	ablatable	ablation	abnormalities	aborting	...	zippering	zippers	zirconia	zone	zones	zoom	zooming	zc
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
2	0.0	0.0	0.0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

Partial Implementation Result (Continued)

- Content-based Recommendation System
  - Test Result
    - Searching by keyword “bicycle”

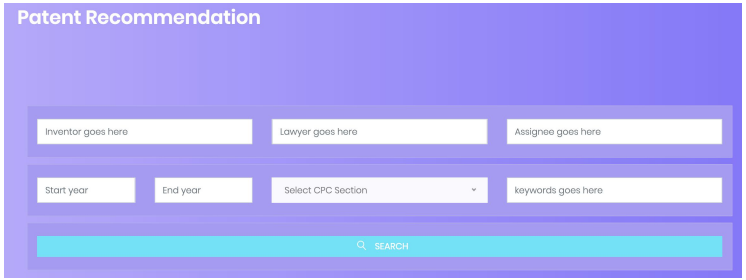
	id	type	country	date	abstract	title	A	B	C	D	E	F	G	H	Y	bicycle
3428	7144024	utility	US	2006-12-05	A, manually, maneuverable,, quick, release, ri...	Ride, along, quick, release, doll, carrier, do...	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.556440
4879	8602279	utility	US	2013-12-10	A, bicycle, rack, configured, mounting, hitch,...	Pivoting, hitch-mountable, bicycle, carrier	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.337376
2737	6450572	utility	US	2002-09-17	A, total, comfort, bicycle, saddle, adapted, p...	Total, comfort, bicycle, saddle	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.129353
4039	7757821	utility	US	2010-07-20	A, bicycle, hydraulic, brake, actuation, devic...	Bicycle, hydraulic, brake, actuation, device	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.106559
577	4277986	utility	US	1981-07-14	, A, stepless, variable, stroke, drive, partic...	Stepless,, variable, stroke, drive, non-rotati...	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.057841

Application / User Interface



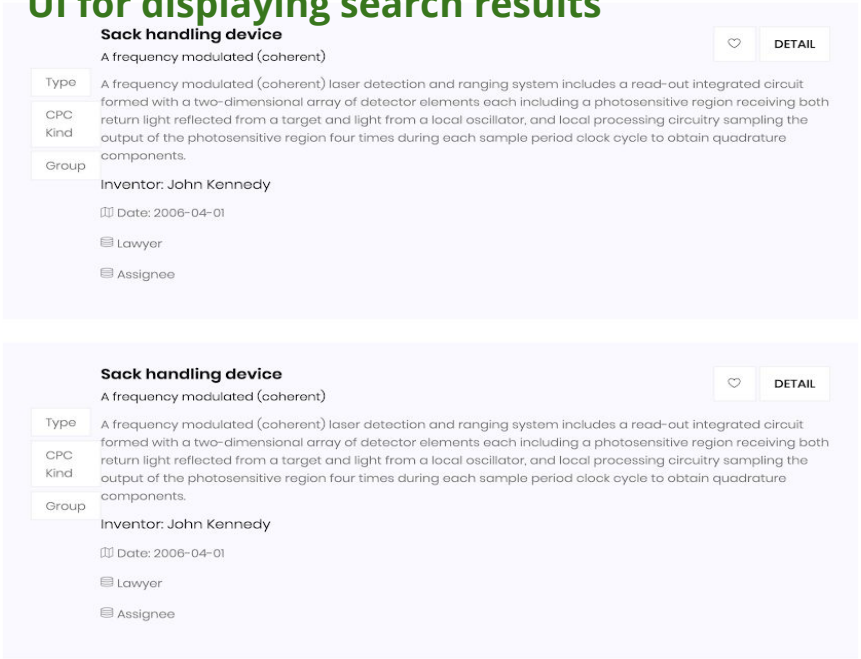
Application / User Interface (Continued)

Search box for User Inputs



https://patten-recommender.herokuapp.com/

UI for displaying search results



Solution Evaluation

- Future work - Solution Evaluation
  - Plan: After completing the recommendation system, an experimental study will be conducted for evaluation purpose. Participants for this study will be selected from students and employees in San Jose.
  - Target: Evaluating our patent recommendation system by metrics - precision and recall. Fine-tuning the parameter of the model and optimizing the system based on evaluation result.

Impact after Implementation

- Less time searching Patent database
- More time for R&D
- Reduce time for patent approval
  - Process can take a year or longer from USPTO
- Searching results will be more precise and efficient
  - Obtaining the patents information that fulfill user requests and preferences
  - Automatically recommending patents based on user behaviours

Thank You