Recommendation Systems

+ data privacy, and
startup stories

Filip Kaliszan
Stanford University
October 1, 2015

Filip Kaliszan

- studied computer science at Stanford - systems in undergrad, then Master's in human computer interaction;
- learned to build products at: Stanford, Adobe, CourseRank, Chegg, and Guidebook
- work part-time at Life360; advise start-ups; getting ready for my next project
- photography, travel, skiing, and (currently) re-building a house…

CourseRank

- class ratings & reviews website (…CourseCycle, TreeViews)
- started in 2007 in CS 194: Software Project
- 2010: >100k students @ 30 schools, ~30k reviews
- 2010: Chegg acquired CourseRank & integrated tech
- late 2014: Chegg stopped supporting courserank.com

Chegg® NETFLIX

course search

Search

Schedule

Planner

Search Courses

course name, instructor, title, or keywords...

search | show all courses

Invite Friends | **Browse by Subject** | New Questions (10) | New Reviews (10) | Statistics

A,RESEC
AGRICULTURAL AND RESOURCE ECONOMICS

AEROSPC
AEROSPACE STUDIES

AFRICAM
AFRICAN AMERICAN STUDIES

AGR CHM
AGRICULTURAL AND ENVIRON CHEMISTRY

AHMA
ANCIENT HISTORY AND MED. ARCH.

AMERSTD
AMERICAN STUDIES

ETH GRP
ETHNIC STUDIES GRADUATE GROUP

ETH STD
ETHNIC STUDIES

EURA ST
EURASIAN STUDIES

EWMBA
EVE/WKND MASTERS IN BUS. ADM.

FILM
FILM STUDIES

FOLKLOR
FOLKLORE

NATAMST
NATIVE AMERICAN STUDIES

NAV SCI
NAVAL SCIENCE

NE STUD
NEAR EASTERN STUDIES

NEUROSC
NEUROSCIENCE

NSE
NANOSCALE SCIENCE AND ENGINEERING

NUC ENG
NUCLEAR ENGINEERING

CourseRank

Quick Lookup

- Home
- Browse
- Plan
  - Schedule
  - Planner
  - Requirements
- Community

2005-2006	2006-2007	2007-2008	2008-2009

Autumn

units: 20 GPA: 3.75		units: 19 GPA: 3.56		units: 17 GPA: 3.92		units: 0 GPA: N/A
CS 10SC	A+	CS 107	B	CS 140	B	
CS 55N	A	CS 147	A-	CS 145	A+	
ECON 1	A+	CS 194	A+	CS 161	A+	
IHUM 57	B+	ECON 50	CR	CS 202	A+	
MATH 51	B+	MS&E 91SI	CR	CS 247	A	

Winter

units: 18 GPA: 3.43		units: 15 GPA: 3.79		units: 15 GPA: 3.87		units: 16	
CS 106A	A	CME 106	A+	COMM 169	A	CHINLANG 2	5
ECON 1	B	CS 103A	A	CS 399	A+	CS 108	4
IHUM 5A	B+	CS 108	A-	GES 1	B+	CS 221	4
PHYSICS 80N	CR	DANCE 46	CR	MS&E 193	A	CS 223A	3
		EE 108B	B				

Spring

units: 16 GPA: 3.31		units: 1 GPA: 0.00		units: 3 GPA: 2.70		units: 0
CS 106B	B	CS 540	F	CS 155	B-	
CS 196	CR					
IHUM 5B	B+					
PWR 1	A-					

Summer

units: 0 GPA: N/A		units: 15 GPA: 3.53		units: 7 GPA: 3.96		units: 0
		CS 103B	B	ENGR 50	A-	
		CS 194	A+	MUSIC 18B	A+	
		ENGR 40	B			
		PWR 2	A			

Import

Transcript

Export

Adobe PDF

Grades

3.62 overall GPA

146 completed units

162 total units

GERs

Disciplinary Breadth:

Hum: MUSIC 18B

EngrAppSci: CS 106A

Math: CS 103A

NatSci: GES 1

SocSci: COMM 169

Education for Citizenship:
(2 of 4 required)

AmerCul: MUSIC 18B

GlobalCom: n/a

Gender: n/a

EthicReas: n/a

https://www.youtube.com/watch?v=rBEUXEFAo_g

Why focus on recommendations?

// answer in class

- personalized experience for users
- helping you make your decision / choice
- classes you haven't thought of before
- what have people with similar majors taken before?

Why focus on recommendations?

- interesting research (Netflix Prize; rich data set)
- better user experience:
  - helping students discover classes
  - more engaging
  - more time on site
  - easier way to add-classes
- added-classes result in more ratings / reviews

CourseRank Data

Courses ~7,500

(CourseID, DeptID, Num, Title, Description…)

CourseID	DeptID	Num	Title	Descr
1	CS	106A	Pro..	Intro.
2	CS	107	Com…	Funda
3	MATH	51	Calc…	Multi…
4	CME	106	Stati…	Appli..

Students ~6,500

(SuID, Name, Age, GPA, Major…)

SuID	Name	Age	GPA	Major
1	Fred	19	3.8	CS
2	Jenn	18		
3	Matt	20	3.2	MATH
4	Alice	22	3.4	CS

Course History >100,000

(SuID, CourseID, Year, Term, Grade, Rating…)

SuID	CourseID	Year	Term	Grade	Rating
1	2	2015	Winter	A-	5

Fred took CS 107 in Winter 2015, got an A-, and rated it 5 stars.

How would you generate recommendations?

// answer in class

- find similar students and see what classes they took
- follow natural progression of a major
- find classes similar to ones rated highly
- find common combinations between classes
- figure out major / university requirements
- popularity

course search

Search

Schedule

Planner

Search Courses

course name, instructor, title, or keywords...

search | show all courses

Invite Friends | **Browse by Subject** | New Questions (10) | New Reviews (10) | Statistics

A,RESEC
AGRICULTURAL AND RESOURCE ECONOMICS

AEROSPC
AEROSPACE STUDIES

AFRICAM
AFRICAN AMERICAN STUDIES

AGR CHM
AGRICULTURAL AND ENVIRON CHEMISTRY

AHMA
ANCIENT HISTORY AND MED. ARCH.

AMERSTD
AMERICAN STUDIES

ETH GRP
ETHNIC STUDIES GRADUATE GROUP

ETH STD
ETHNIC STUDIES

EURA ST
EURASIAN STUDIES

EWMBA
EVE/WKND MASTERS IN BUS. ADM.

FILM
FILM STUDIES

FOLKLOR
FOLKLORE

NATAMST
NATIVE AMERICAN STUDIES

NAV SCI
NAVAL SCIENCE

NE STUD
NEAR EASTERN STUDIES

NEUROSC
NEUROSCIENCE

NSE
NANOSCALE SCIENCE AND ENGINEERING

NUC ENG
NUCLEAR ENGINEERING

Three Approaches

1. Collaborative Filtering Recommender
2. Content Filtering Recommender
3. Constraint Based Recommender

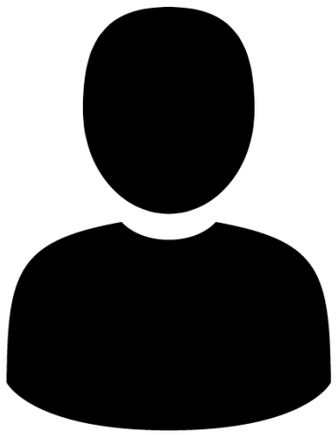
1. Collaborative Filtering

Collaborative Filtering

"…collaborative filtering is a method of making automatic predictions (filtering) about the interests of a user by collecting preferences or taste information from many users (collaborating)"

Wikipedia, Collaborative Filtering, September 30, 2015
https://en.wikipedia.org/wiki/Collaborative_filtering

Collaborative Filtering



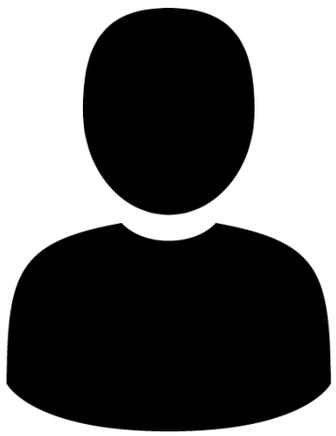
compare

Fred

other students

- 19 years old
- 3.8 GPA
- CS 106A
- CS 107

Collaborative Filtering



compare

students similar to Fred

Fred

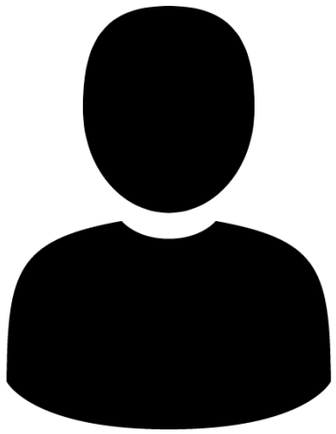
other students

- 19 years old
- 3.8 GPA
- CS 106A
- CS 107

recommend

~~CS 106A~~
~~CS 107~~
MATH 51
CS 108
IHUM 57

Collaborative Filtering



Fred

- 19 years old
- 3.8 GPA
- CS 106A
- CS 107

How to compute similarity…?

Students Age Difference

SuID	Name	Age	Score
2	Jenn	18	
3	Matt	20	
4	Alice	22	

Collaborative Filtering

Jaccard Similarity - statistic for comparing similarity of sets (in this case sets of courses)

$$\textbf{Jaccard Sim(Fred, Jenn)} = \frac{\text{courses taken by Fred } \textbf{and} \text{ Jenn}}{\text{courses taken by Fred } \textbf{or} \text{ Jenn}}$$

$$= \frac{2}{3} = 0.67$$

Fred	**Jenn**
CS 106A	CS 106A
CS 107	CS 107
	MATH 51

SuID	Name	Score
2	Jenn	0.67
3	Matt	0.33
4	Alice	0.50

Collaborative Filtering

How would you optimize it?

// answer in class

- similarity: major, gpa,
- factor in ratings they gave
- sequence

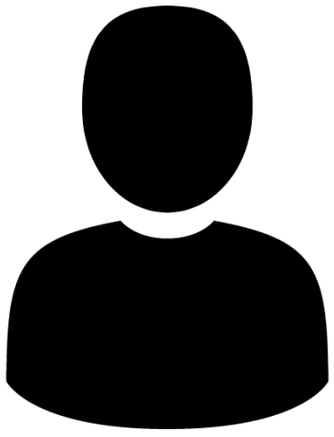
2. Content Filtering

Content Filtering

"…Content-based filtering methods are based on a description of the item and a profile of the user's preference"

Wikipedia, Recommender System, September 30, 2015
https://en.wikipedia.org/wiki/Recommender_system#Content-based_filtering

Content Filtering



Fred

- 3.8 GPA
- CS 106A
- CS 107

Profile / Preferences

- 19 years old / sophomore: **intro classes**
- CS 106A and 107: **computer science**
- CS: part of **school of engineering**

we can also learn more about his specific classes…

Content Filtering

Courses ~7,500

(CourseID, DeptID, Num, Title, Description)

CourseID	DeptID	Num	Title	Descr
1	CS	106A	Pro..	Intro.
2	CS	107	Com…	Funda
3	MATH	51	Calc…	Multi…
4	CME	106	Stati…	Appli..

CS 106A: Programming Methodology

Introduction to the engineering of computer applications emphasizing modern software engineering principles: object-oriented design, decomposition, encapsulation, abstraction, and testing. Uses the Java programming language. Emphasis is on good programming style and the built-in facilities of the Java language. No prior programming experience required. Summer quarter enrollment is limited. Priority given to Stanford students.

Content Filtering

prefix identifies department ⟶

CS 106A: Programming Methodology

Introduction to the engineering of computer applications emphasizing modern software engineering principles: object-oriented design, decomposition, encapsulation, abstraction, and testing. Uses the Java programming language. Emphasis is on good programming style and the built-in facilities of the Java language. No prior programming experience required. Summer quarter enrollment is limited. Priority given to Stanford students.

Content Filtering

prefix identifies department

course numbers often sequential

200- and 300- level are grad courses

CS 106A: Programming Methodology

Introduction to the engineering of computer applications emphasizing modern software engineering principles: object-oriented design, decomposition, encapsulation, abstraction, and testing. Uses the Java programming language. Emphasis is on good programming style and the built-in facilities of the Java language. No prior programming experience required. Summer quarter enrollment is limited. Priority given to Stanford students.

Content Filtering

prefix identifies department

course numbers often sequential

200- and 300- level are grad courses

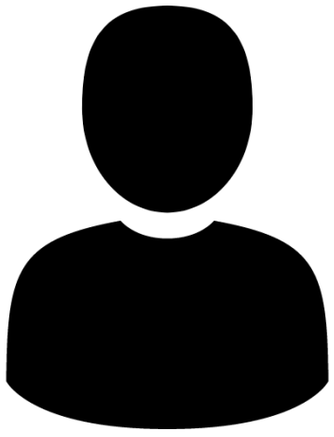
We can also identify **key words** in the text corpus (eg. TF-IDF)

CS 106A: Programming Methodology

Introduction to the **engineering** of **computer applications** emphasizing modern software engineering principles: **object-oriented** design, decomposition, encapsulation, **abstraction**, and testing. Uses the **Java** programming language. Emphasis is on good **programming** style and the built-in facilities of the **Java** language. No prior **programming** experience required. Summer quarter enrollment is limited. Priority given to Stanford students.

** example highlights **not** based on real text analysis **

Content Filtering



Fred

- 3.8 GPA
- CS 106A
- CS 107

Profile / Preferences

- 19 years old / sophomore: **intro classes**
- CS 106A and 107: **computer science**
- CS: part of **school of engineering**

- programming, java, abstraction, engineering, object-oriented

Content Filtering

Search Courses For:

- “CS” department
- 100 number level
- key terms: programming, java, abstraction, engineering, object-oriented

CS 193P: iPhone and iPad Programming

Tools and APIs required to build applications for the iPhone and iPad platforms using the iOS SDK. User interface design for mobile devices and unique user interactions using multi-touch technologies. Object-oriented design using model-view-controller paradigm, memory management, Swift programming language. Other topics include: object-oriented database API, animation, mobile device power management, multi-threading, networking and performance considerations.

Content Filtering

Search Courses For:

- “CS” department
- 100 number level
- key terms: programming, java, abstraction, engineering, object-oriented

CS 193P: iPhone and iPad Programming

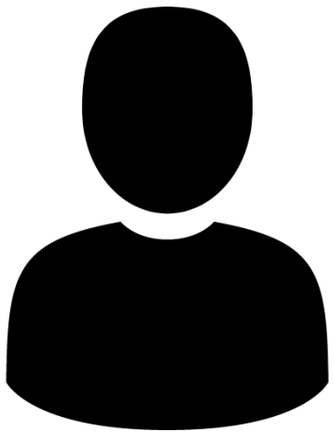
Tools and APIs required to build applications for the iPhone and iPad platforms using the iOS SDK. User interface design for mobile devices and unique user interactions using multi-touch technologies. **Object-oriented** design using model-view-controller paradigm, memory management, Swift **programming** language. Other topics include: **object-oriented** database API, animation, mobile device power management, multi-threading, networking and performance considerations.

3. Constraint Based

Constraint Based

Our constrained-based filtering approach was based on following well-defined selection constraints specific to our application: namely departmental and university-wide graduation requirements.

Constraint Based



Fred

- 3.8 GPA
- CS 106A
- CS 107

Major Undefined:

- recommend GERs
- recommend courses common to most majors

- …then major requirements

Constraint Based

Is this difficult? Why or why not?

// answer in class

- hard to tell if they will major
- people can change majors
- missing data in course history; might recommend things they've already taken
  - "i took it button"
- classes outside of they major
- select different tracks
- planning out requirements over time
- extra curricular (lots of options)
- define the complete set of possible rules

Pros & Cons

Think about each approach in context of scenarios with:

- few students on the system
- sparse course data (missing titles or descriptions)
- many new courses introduced each year
- frequently changing graduation requirements
- poorly documented (or qualitative) major requirements
- self-defined majors

What's missing from each approach? What are its pitfalls?
Which is the best one?

How do you evaluate recommendations?

// answer in class

- users taking the classes
  - what if they were going to take the class anyway
  - have they liked the class (rating)
  - click-through rate

How do you evaluate recommendations?

Personalization metrics at Netflix

- RMSE (predicted rating vs. actual rating)
- % customers at 6 weeks with ≥ 50 ratings
- % customers with ≥ 15 minutes streamed
- % customers with ≥ 6 queue adds in a month (in DVD days)

NETFLIX

Real World Applications

Chegg acquired CourseRank in 2010, forcing growth:

- 30 to 1,500 universities
- 100k to >3mln listed courses
- 100k to “millions” of students

Course ratings & reviews critical to user engagement:

- recommendations: connect students to courses
- clicking courses **much** easier than typing
- real results in ~18 months:
  - 500k to millions of course ratings
  - 30k to >1mln written reviews

Filip Kaliszan

How to reach me?

filip.kaliszan@gmail.com
(650) 796 6302

https://www.facebook.com/filip.kaliszan
flickr.com/kaliszan