

Data Mining

INFS 4203/7203

Miao Xu

miao.xu@uq.edu.au

The University of Queensland, 2020 Semester 2

Last week: text retrieval

- **Retrieve** *relevant* documents in a document collection based on a user's **query**

TF-IDF for documents

For all **documents**:

- Remove stop words, convert to lower case, stemming
- Put words in all documents together, sorting in alphabetical order to have a word list
- Construct TF vector for each document
- Construct IDF vector
- Get the TF-IDF vector for each document
- Normalize the TF-IDF vector to unit length

TF-IDF for query

For the **query**:

- Remove stop words, convert to lower case, stemming
- Use the word list constructed upon **all documents**
- Construct TF vector
- Get the TF-IDF vector by the IDF vector of documents
- Normalize the TF-IDF vector to unit length

- Calculate the **Cosine Similarity** between query vector and document vectors
- The similarity values are then used to **rank** the documents

TF-IDF vs. word embedding

- Both feature extraction methods for text
- Word embedding (usually refer to learning-based feature extraction such as word2vec):
 - Based on deep learning technique
 - Each word is embedded as a “vector”
 - If trained on large datasets, can be reused across different data sets
- TF-IDF
 - Each word is embedded as a “value”
 - Document related, cannot be reused

Latent Semantic Indexing vs. PCA

- Applications on different context
 - LSI used specially in text domain
- Different motivations
 - LSI finds the best subspace in the document-word matrix
 - PCA keeps the features with the greatest variances in data
- Technically, applying PCA to document-word matrix--LSI

Lecture 11: Mining Web Data

Web data

- Web is an unique phenomenon
 - The **scale**, the **distributed** nature of its creation, the **openness** of the underlying platform, and the **diversity** of applications
- Two primary types of data
 - Web content information
 - Document data, linkage data
 - Web usage data
 - Web transactions, ratings, user feedback, web logs

Applications on the web

- Content-centric applications
 - Cluster or classify web documents
 - Web crawling
 - Web search
- Usage-centric applications
 - Recommender systems

Content-centric applications

- Web crawling
- The PageRank algorithm

Usage-centric applications

- Recommender system

Content-centric applications

- Web crawling
- The PageRank algorithm

Usage-centric applications

- Recommender system

Web crawling

- Motivations
 - Resources on the Web are **dispensed** widely across globally distributed sites
 - Sometimes, it is necessary to download all the relevant pages at a **central** location
- Universal crawling
 - Crawl **all** pages on the Web (search engine)
- Preferential crawling
 - Crawl pages related to a **particular** subject or belong to a particular site

A basic crawler algorithm

- Start with a set of seeds, which are a set of URLs given as parameters
 - Seeds are added to a URL **request queue/frontier set**
- Crawler fetches pages from the requested queue/frontier set
- Fetched pages are parsed to find link tags that might contain other useful URLs to fetch
 - **New** URLs are added to the request queue/frontier set
- Continue until no more new URLs or disk full

Select the URL to fetch

- Breadth-first
- Depth-first
- Frequency-based
 - Most universal crawlers are **incremental** crawlers that are intended to refresh previous crawls
 - Choose web pages less visited before
- PageRank-based
 - Choose web pages with high PageRank value

Combating spider traps

- The crawling algorithm maintains a list of previously visited URLs for comparison
 - So, it tends to visit distinct Web pages
- However, many sites create dynamic URLs
 - <http://www.examplesite.com/page1>
 - <http://www.examplesite.com/page1/page2>
 - Limit the maximum size of the URL
 - Limit the number of URLs from a site

Preferential crawling

- Attempts to download only those pages that are about a **particular** topic
- Rely on the “**belief**” that pages about a topic tend to have links to pages about the same topic
 - **Popular** pages for a topic are typically used as seeds
- Crawler uses **text classifiers** to decide whether a page is on topic or not

Preferential crawling-con't

- Only visit those links from a page that are **determined to be relevant** by a pre-trained classifier, or,
- Crawler visits pages with priority based on the **relevance score** assigned by a pre-trained classifier
- How to determine relevance?
 - Parent-based: score of a parent page is distributed to all page it links
 - Anchor-based: score is determined on the anchor text to that page (anchor text: visible words that hyperlinks display when linking to another page)

Content-centric applications

- Web crawling
- **The PageRank algorithm**

Usage-centric applications

- Recommender system

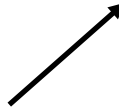
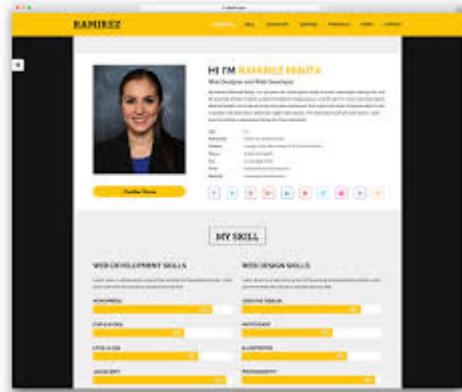
The PageRank algorithm

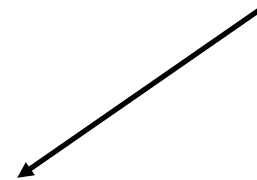
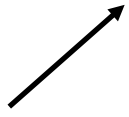
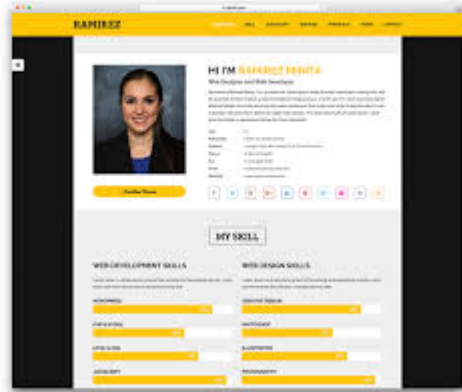
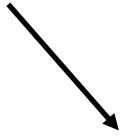
- **First** and **best-known** algorithm used by Google Search to rank web pages
- Larry Page and Sergey Brin developed PageRank at Stanford University in 1996 as part of a research project



Basic idea

- Rank webpages according to importance (called PR value)
- Importance is determined by
 - More pages linked to it, more importance
 - If pages linked to it are important, then the page is important
- Intuitively, it measures the likelihood that a person randomly clicking on links will arrive at any particular page





A simplified PageRank algorithm

- A page p_i has n_i links out, and one of the links is to page p
- There are totally m pages linking to page p
- The Page Rank value $PG(p)$ is defined as

$$PG(p) = \sum_{i=1}^m PG(p_i)/n_i$$

- By assigning some initial PG values to all webpages, the process continues until convergence or nearly convergence (the difference is within a small threshold)

More about Google Search

- Two weeks ago, Google released a documentary film about the 22 years' history of Google Search:

*Trillions of Questions, No Easy Answers:
A (home) movie about how Google
Search works*

- Some fun facts:

- 15% query per day is “never seen before”
- How Google deal with wrong results and offensive search predictions
- ... (explore by yourself)

My dear, here **we** must **run** as fast as **we**
can, just to **stay** in place.

-----Alice in Wonderland





https://www.youtube.com/watch?v=tFq6Q_muwG0

Content-centric applications

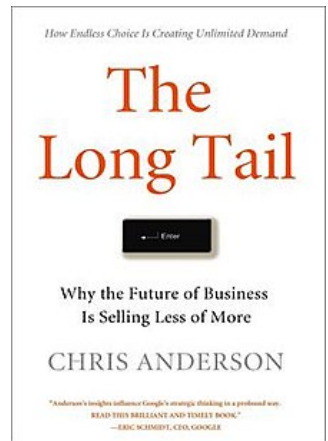
- Web crawling
- The PageRank algorithm

Usage-centric applications

- Recommender system

The Age of Web Search has come to an end

- ...long live the Age of Recommendations
- Chris Anderson in “The Long Tail”
 - *“We are leaving the age of information and entering the age of recommendation”*
- CNN Money, “The race to create a ‘smart’ Google”:
 - *“The Web, they say, is leaving the era of search and entering one of discovery. What’s the difference? Search is what you do when you’re looking for something. Discovery is when something wonderful that you didn’t know existed, or didn’t know how to ask for, **finds you**”*



Everything is personalized

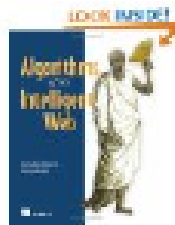


Recommendation in Amazon

Customers Who Bought This Item Also Bought



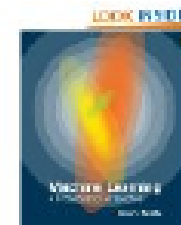
**Recommender Systems
Handbook**
Francesco Ricci
Hardcover
\$167.73



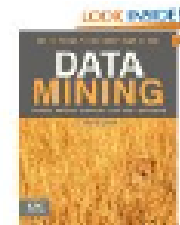
**Algorithms of the Intelligent
Web**
Haralambos Marmanis
★★★★★ (14)
Paperback
\$26.76



**Programming Collective
Intelligence: ...**
➤ Toby Segaran
★★★★★ (91)
Paperback
\$25.20



**Machine Learning: A
Probabilistic ...**
➤ Kevin P. Murphy
★★★★★ (15)
Hardcover
\$81.00




**Data Mining: Practical
Machine Learning ...**
➤ Ian H. Witten
★★★★☆ (29)
Paperback
\$42.61

In e-commerce

Related Items Recommendations – Example 1

Item Bought



BlackBerry Bold 9780 - Black (T-Mobile) Smartphone NIB

Price: **US \$379.99**

Shipping: **FREE** Standard Shipping has more details

Delivery: Estimated between **Wed, Aug. 17** and **Tue, Aug. 23**

Returns: 7 days money back, buyer pays return shipping

eBay Buyer Protection

Seller Info

al8808 (407)

99.3% Positive feedback

Other item info

Item number: 13066073880


Item location: Orem, UT, United States

Ships to: United States

Payments: PayPal, Visa, MasterCard

History: 1 sold

Related Items Recommendations




TWOBLACKBERRY BOLD 9780 BLACK LEATHER CASE POUCH

\$12.77

Buy It Now or Best Offer

Free shipping

See suggestions




Screen Protector LCD for BlackBerry Bold 9780 T-Mobile

\$1.96

Buy It Now

Free shipping

See suggestions



ORIGINAL BLACKBERRY


DEM ORIGINAL BATTERY FOR BLACKBERRY BOLD 9780 T-MOBILE

\$6.93

Buy It Now or Best Offer

Free shipping

See suggestions




MICRO USB HOME CHARGER BLACKBERRY BOLD

\$3.50

Buy It Now

Free shipping

See suggestions




BlackBerry Bold 9780 Bold Mobile Store Black Case

\$12.91

Buy It Now

See suggestions

eBay Inc. confidential



In social media

Who to follow · Refresh · View all

**GNIP, Inc.**  @gnip
 Promoted Follow

**Twitter**  @twitter
Followed by Michael Ekstrand and...
Follow

**Yong Zheng** @irecsys
Followed by sbourke
Follow

Personen, die du vielleicht kennst [Alle anzeigen](#)


4 gemeinsame Freunde
[FreundIn hinzufügen](#)


1 gemeinsame/r FreundIn
[FreundIn hinzufügen](#)

Jobs you may be interested in Beta [Email Alerts](#) [See More »](#)

- **Technical Sales Manager - Europe** 
Thermal Transfer Products - Home office
- **Senior Program Manager (I/m)** 
Johnson Controls - Germany-NW-Burscheid

Groups You May Like [More »](#)

- **Advances in Preference Handling**
 Join
- **FP7 Information and Communication Technologies (ICT)**
 Join
- **The Blakemore Foundation**
 Join

 Picasa™ -Webseiten [Startseite](#) [Meine Fotos](#) [Erkunden](#) [Hochladen](#)

Empfohlene Fotos [Alle anzeigen](#)



Entertainment

Home

My Channel

Subscriptions

History

Watch Later

Purchases2

PLAYLISTS

Favorites

Workout Music

Stime

Show more

SUBSCRIPTIONS

Add channels

Popular on YouTube

Music

Sports

Gaming

Browse channels

Manage subscriptions

HomeSubscriptions

Show ad

Recommended

Hardwell On Air 239
by Hardwell
224,022 views · 2 weeks ago

BEST NEWS BLOOPERS
FEBRUARY 2014
by NewsBeFunny
7,857,720 views · 1 year ago

Bradley Cooper Can't Get Out
of the Way
by The Late Late Show with James
Corden
438,880 views · 1 month ago

GOP Debate: Main Event (Full
Debate) | CNBC
by CNBC
711,147 views · 1 week ago

Truth or Drink (Couples) -
Episode 2: Krystal (Not...
by WatchCut Video
1,537,275 views · 8 months ago

Islam and Stephen Hawking:
Asking big questions about...
by bilalbnrbaash
326,631 views · 5 years ago

Jimmy Kimmel the Uber
Driver
by Jimmy Kimmel Live
4,731,140 views · 1 year ago

What's In Kate Hudson's 'Box
of Lies'?
by Entertainment Tonight
430,148 views · 1 year ago

Show more

The Verge

Recommended channel for you






Subscribe


843,629

X

Point of Interest (POIs)


Just show me:






1. PAPER coffee
Coffee Shop
44 W 28th St (btwn 6th Ave & Broadway), 纽约

[Save](#)


 Bobby S. • May 28, 2018
Serves up Devotion coffee and pastries in a warm space.



2. Birch Coffee
Coffee Shop
134 1/2 E 62nd St (at Lexington Ave), 纽约

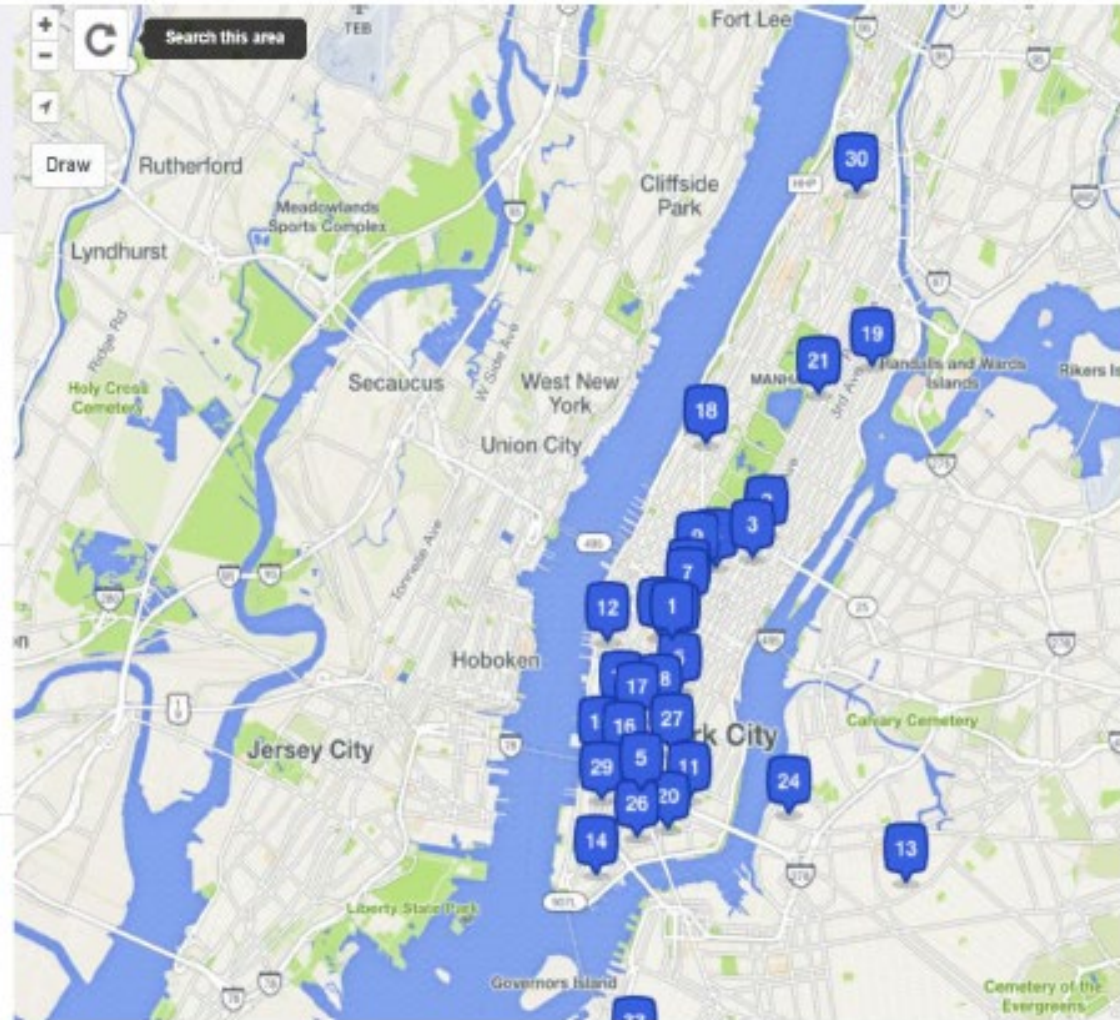
[Save](#)

★ Over 40 tips talk about the cold brew coffee, café, and cortados



3. Little Collins
Coffee Shop • View Menu
667 Lexington Ave (btwn E 55th & E 56th St), 纽约

[Save](#)



Why using recommender systems

- Value for the customer
 - Find things that are interesting
 - Narrow down the set of choices
 - Help explore the space of options
 - Discover new things
 - Entertainment
 - ...

Why using recommender systems

- Value for the provider
 - Personalized service for the customer to increase user experience
 - Increase trust and customer loyalty
 - Increase sales, click through rates...
 - Obtain more knowledge about customers
 - ...

Recommender system

- Given
 - **User data** (with or without demographics, situational context)
 - **Item data** (with or without description of item characteristics)
 - **User-Item Interaction matrix**
- Find
 - Relevance score between user and item, which is used for ranking

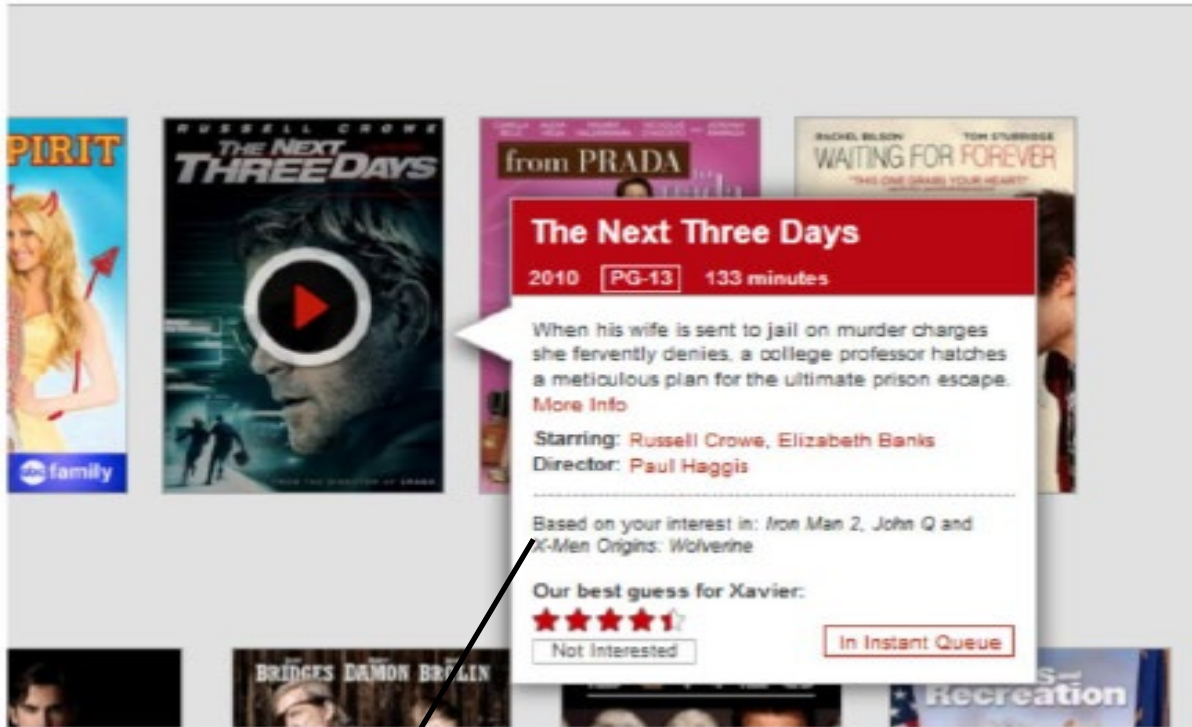
Purpose and success criteria

- Prediction perspective
 - Predict to what degree users like an item
 - Most popular criterion in research (why?)
- Exploration perspective
 - Serendipity
 - Users will be happy but did not know about the existence
- Interaction perspective
 - Convince/persuade users
- ...

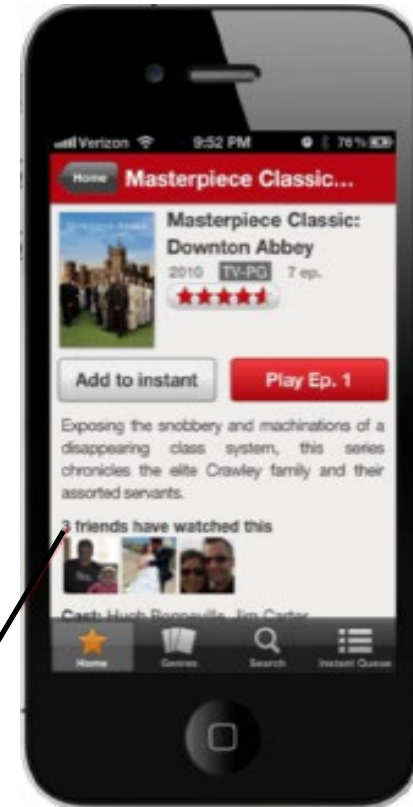
Serendipity

- Unsought finding
- Do not recommend items that user already knows or would have found anyway
- Expand the user's taste into neighbouring areas

Convince/persuade users



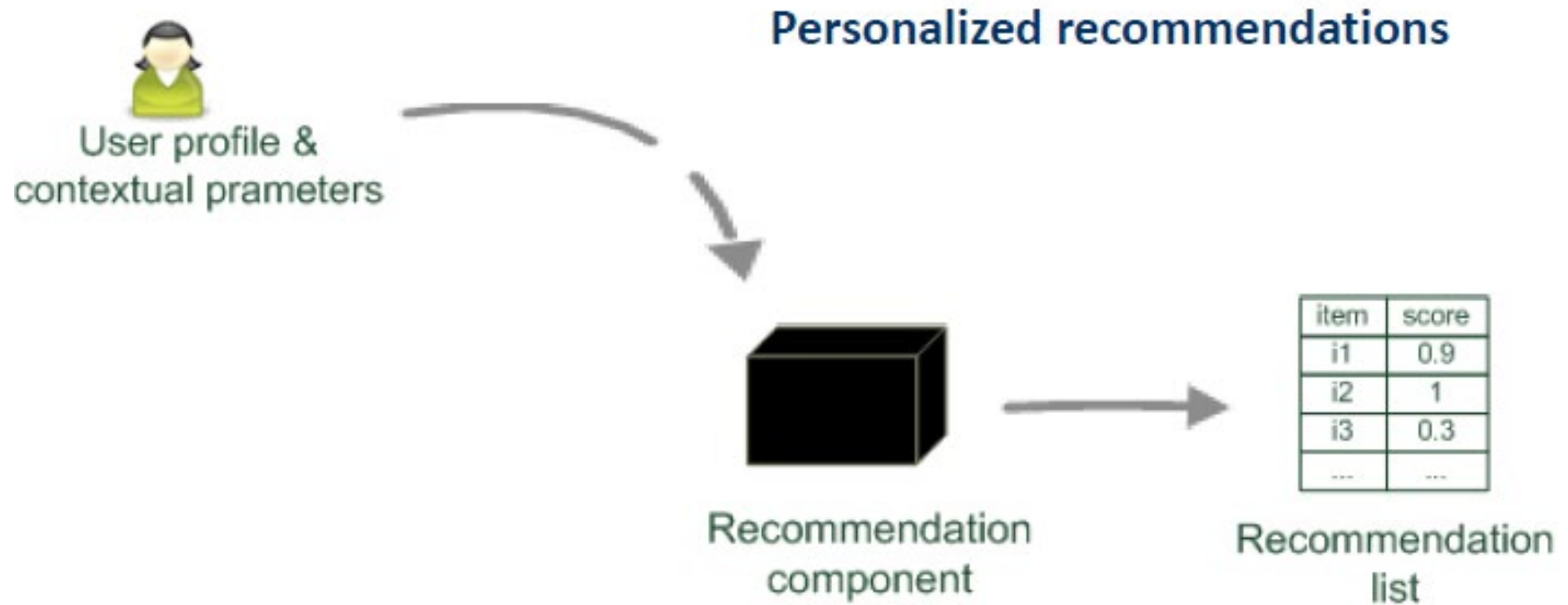
Recommend "*The Next Three Days*" based on your interested in *Iron Man 2*, *John Q* and *X-Men Origins: Wolverine*



Three friends have watched this

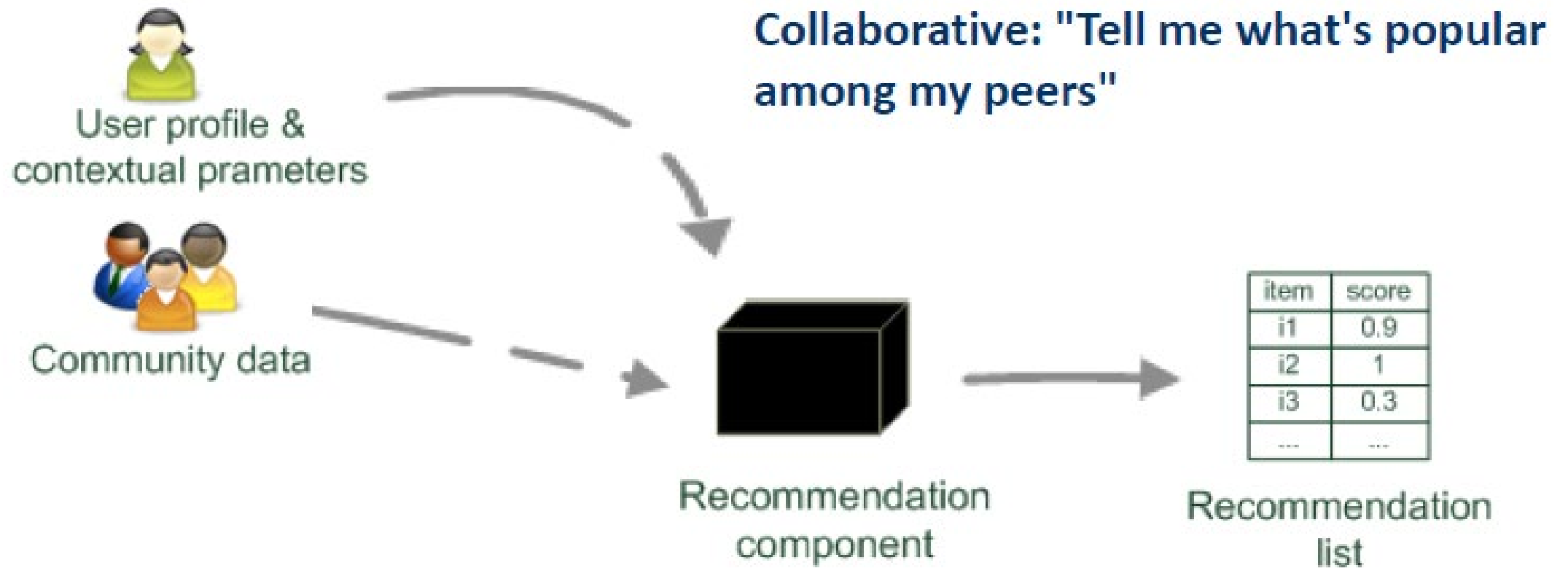
Paradigms of recommender systems

- Personalized recommendation



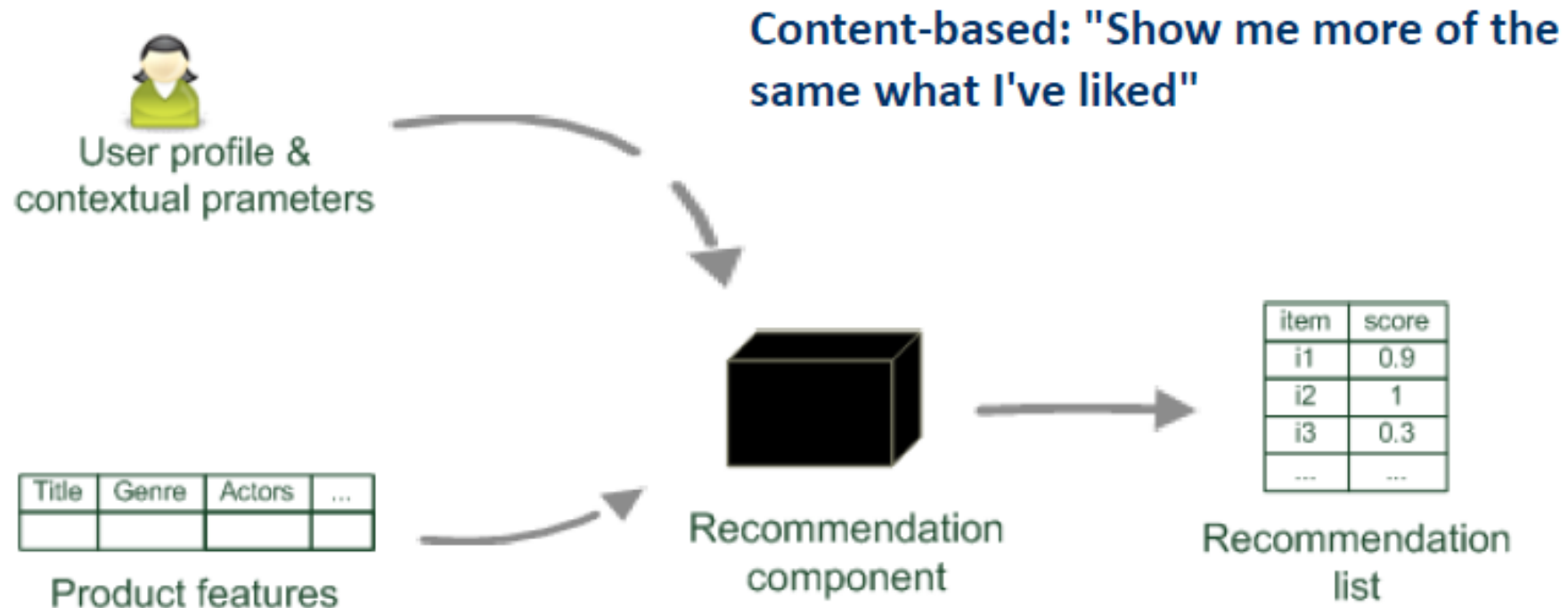
Paradigms of recommender systems

- Collaborative filtering



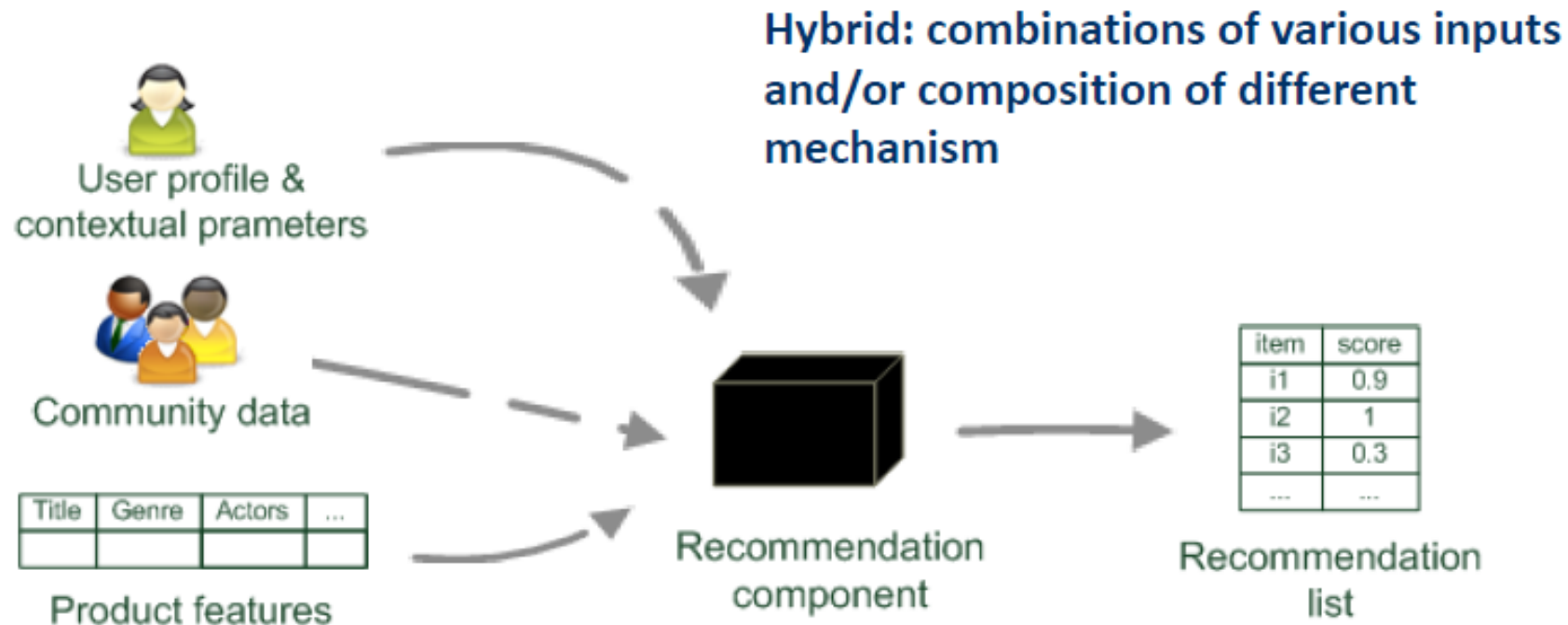
Paradigms of recommender systems

- Content-based recommendation



Paradigms of recommender systems

- Hybrid: combinations of various inputs, and/or composition of different mechanism



Paradigms of recommender systems

	Pros	Cons
Collaborative	Serendipity of results, learns market segments	Requires some form of rating feedback, cold start for new users and new items
Content-based	No community (other users) required, comparison between items possible	Content descriptions necessary, cold start for new users, no surprises

What works

- Depends on the **domain** and particular **problem**
- However, in the general case, it has been demonstrated that **Collaborative Filtering (CF)** is better than content-based
- Other factors impacting the result
 - **Data pre-processing**: outlier removal, denoising, removal of global effects (e.g. individual user's bias...)
 - Combining methods through **Ensemble**

Libraries or tools

- Recommender 101
 - <https://ls13-www.cs.tu-dortmund.de/homepage/recommender101/index.shtml>
- MyMediaLite
 - <http://www.mymedialite.net/>
- RecQ
 - <https://github.com/Coder-Yu/RecQ>

Collaborative Filtering (CF)

- The most prominent approach to generate recommendations
 - Used by large, commercial e-commerce sites in many domains
 - Well-studied, various algorithms exist
- Approach
 - Use the “wisdom of the crowd” to recommend items
 - Only require user-item interaction information
- Basic assumption and idea
 - Customers who had similar tastes in the past, will have similar tastes in the future

Pure CF approach

- Input
 - A matrix of given user-item ratings
- Output
 - A numerical prediction indicating to what degree the current user will like or dislike a certain item (explicit feedback)
 - A top-N list of recommended items (implicit feedback)

User-based collaborative filtering

Given an “active user” Alice:

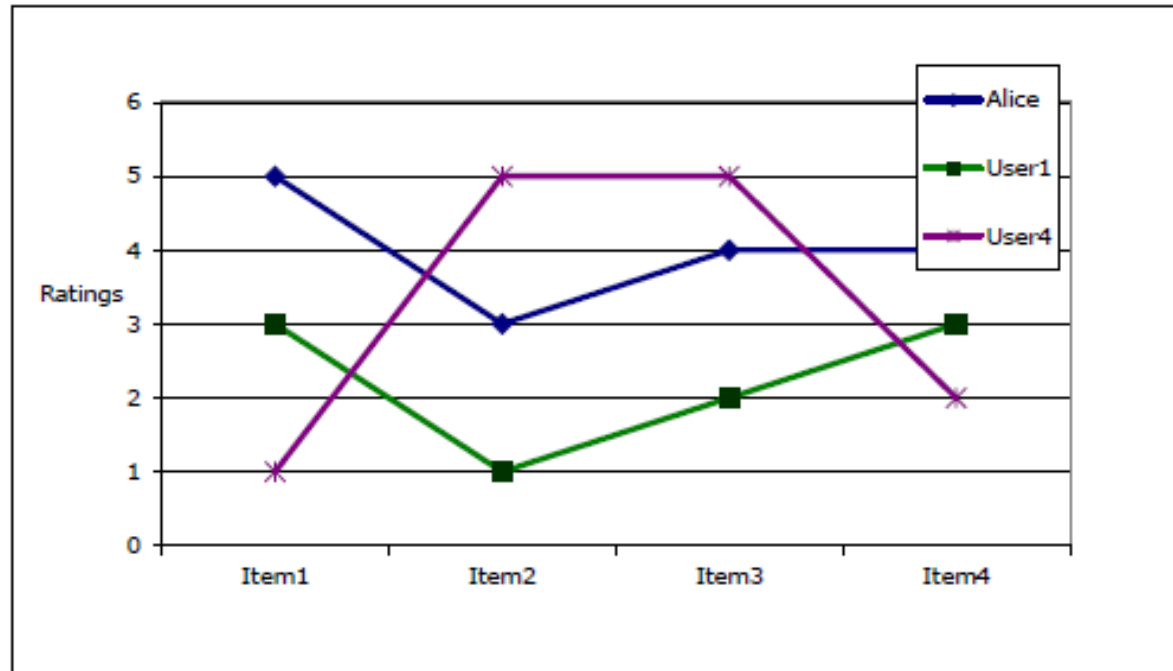
- Find a set of users (peers/nearest neighbours) who liked the same items as Alice in the past
- Find the set of items not rated by Alice but neighbours
- For each item, (weighted) average their ratings in neighbours to predict Alice’s rating
- Recommend the item with the highest ratings/ Rank the items according to the ratings

Example of CF

	Item1	Item2	Item3	Item4	Item5
Alice	5	3	4	4	?
User1	3	1	2	3	3
User2	4	3	4	3	5
User3	3	3	1	5	4
User4	1	5	5	2	1

How to measure the similarity?

- Ratings can be different, but relative ratings may be the same



- Similarity based on relative ratings: **Pearson correlation**

Pearson correlation

- Notation
 - a, b : users
 - $r_{a,p}$: ratings of user a for item p
 - P : set of items rated by both a and b
 - \bar{r}_a : average rating of user a
 - \bar{r}_b : average rating of user b
- Pearson correlation:

$$\text{PearsonCorrelation}(a, b) = \frac{\sum_{p \in P} (r_{a,p} - \bar{r}_a)(r_{b,p} - \bar{r}_b)}{\sqrt{\sum_{p \in P} (r_{a,p} - \bar{r}_a)^2} \sqrt{\sum_{p \in P} (r_{b,p} - \bar{r}_b)^2}}$$

Pearson correlation

Average rating for Alice
on item 1 to item 4: 4

Average rating for User1
on item 1 to item 4: 2.25

	Item1	Item2	Item3	Item4	Item5
Alice	5	3	4	4	?
User1	3	1	2	3	3
User2	4	3	4	3	5
User3	3	3	1	5	4
User4	1	5	5	2	1


Relative rating of Alice: 1, -1, 0, 0

Relative rating of User 1: 0.75, -1.25, -0.25, 0.75

$$\begin{aligned} \text{PearsonCorrelation} &= \frac{1 \times 0.75 + (-1) \times (-1.25) + 0 + 0}{\sqrt{1^2 + (-1)^2} \sqrt{0.75^2 + (-1.25)^2 + (-0.25)^2 + 0.75^2}} \\ &= \frac{2}{1.414 \times 1.658} = 0.85 \end{aligned}$$

Example of CF

	Item1	Item2	Item3	Item4	Item5
Alice	5	3	4	4	?
User1	3	1	2	3	3
User2	4	3	4	3	5
User3	3	3	1	5	4
User4	1	5	5	2	1



sim = 0,85
sim = 0,70
sim = 0,00
sim = -0,79

- How to predicate the rating of Alice for item 5?


How to predicate the rating of Alice for item 5?

- Basic idea
 - Calculate each neighbour's relative rating for item 5
 - Weighted average the relative ratings from neighbours by the Pearson correlation to form the relative rating of Alice for item 5
 - Convert the relative rating back to the absolute rating by adding \bar{r}_a

$$pred(a, p) = \bar{r}_a + \frac{\sum_{b \in NN} sim(a, b) \times (r_{b,p} - \bar{r}_b)}{\sum_{b \in NN} sim(a, b)}$$

Example of prediction

	Item1	Item2	Item3	Item4	Item5
Alice	5	3	4	4	?
User1	3	1	2	3	3
User2	4	3	4	3	5
User3	3	3	1	5	4
User4	1	5	5	2	1



sim = 0,85

sim = 0,70

sim = 0,00

sim = -0,79

Average rating for Alice on item 1 to item 4: 4

Average rating for User1 on item 1 to item 4: 2.25

Average rating for User2 on item 1 to item 4: 3.5

Assume we only consider 2 nearest neighbours here: User 1 and User 2

$$\begin{aligned} \text{pred}(\text{Alice, item 5}) &= 4 + \frac{0.85 \times (3 - 2.25) + 0.70 \times (5 - 3.5)}{0.85 + 0.70} \\ &= 5.0887 \end{aligned}$$

Summary

Content-centric applications

- Web crawling
- The PageRank algorithm

Usage-centric applications

- Recommender system
 - Paradigms of recommender system
 - Collaborative filtering

Final-term review

Lecture1	Introduction to data mining
Lecture2	Association Rule Mining
Lecture3	Clustering 1
Lecture4	Clustering 2
Lecture5	Anomaly detection
Lecture6	Classification 1
Lecture7	Classification 1 - continue
Lecture8	Classification 2
Lecture9	Classification 3
Lecture10	Mining text data
Lecture11	Mining web data

What is data mining

“Data mining is the analysis of (often **large**) observational data sets to find **unsuspected** relationships and to summarize the data in **novel** ways that are both **understandable** and **useful** to the data owner.”

[D. Hand et al. , Principles of Data Mining]



Association rule mining

- Definition
- Application
- Support and confidence
- Apriori property
- Association rule mining algorithm
 - ~~4~~ The Apriori algorithm
 - Rule generation

Clustering

- Definition
- Application
- Distance
 - Manhattan distance, Euclidean distance, Chebyshev distance (at least one of them will be test)
- SSE
- Three methods to do clustering

Anomaly detection

- Definition
- Applications
- Outlier vs. noise
- Four techniques to do anomaly detection


Classification

- Definition
- Missing values/normalization/curse of dimensionality
- Multi-class classification to binary classification
- Generalization and overfitting
- Cross-validation
- Evaluation: accuracy, precision, recall, F1 measurement
- Three classification algorithms and other classification methods based on them

More on the calculation in classification

- Decision tree:
 - If asked to construct a decision tree for the given training set, please describe the final constructed tree in plain language; **NO** procedure required
 - ~~– Calculation based on continuous values~~
 - ~~– Calculation based on Gain ratio~~
 - But related concepts may be tested

- Naïve Bayes

- Briefly show the procedure to do classification (with example later)
- Use Laplacian correction if necessary 
- Calculation on continuous features such as Income/Sales... based on the univariate gaussian distribution will be tested!!!

- k-NN

- Briefly show the procedure to do classification (with example later) with specified distance
- If not explicitly asked, no ~~normalization~~

Mining text data

- Concept
- Measures: precision, recall, F1
- The vector space model
 - TF-IDF vector and retrieval based on TF-IDF vector

Mining web data

- Collaborative filtering
 - Ideas and workflow
 - **NO** calculation tested in final-term

Form

- Online exam, BB test, invigilated by Proctor U
 - It may take up to 30 minutes to connect you with a proctor, however your exam timer does not start until your proctor starts your exam.
 - After an exam is started, you will have the duration of the exam to complete and submit your response.
- Time: Nov. 19th
- Duration: two and a half hours (the duration includes 30 minutes of technical time)
- Allowed materials: calculator/blank scrap paper

- Exam information
- Mock exam
 - 2.5 hours duration
 - Multiple trials are allowed
 - Mimic the situation of network problem/ computer restart... just in case

Mock exam

- Instructions

INSTRUCTIONS

Timed Test	This test has a time limit of 2 hours and 30 minutes.
Timer Setting	This test will save and submit automatically when the time expires.
Force Completion	This test can be saved and resumed at any point until time has expired. The timer will continue to run if you leave the test.
Multiple Attempts	This test allows multiple attempts.
Click Begin to start: Mock Final-Term Exam. Click Cancel to go back. You will be previewing this assessment and your results will not be recorded.	

Click Begin to start. Click Cancel to quit.

Mock exam

- Test information at the head

Test Information	
Description	
Instructions	
Timed Test	This test has a time limit of 2 hours and 30 minutes. This test will save and submit automatically when the time expires. Warnings appear when half the time, 5 minutes, 1 minute, and 30 seconds remain. <i>[The timer does not appear when previewing this test]</i>
Multiple Attempts	This test allows multiple attempts.
Force Completion	This test can be saved and resumed at any point until time has expired. The timer will continue to run if you leave the test.

- Seven questions appear in one page, with several sub questions each
- Additionally, question 8:

QUESTION 8

Please use this space to specify any assumptions you have made in completing the exam and which questions those assumptions relate to. You may also include queries you may have made with respect to a particular question, should you have been able to 'raise your hand' in an examination room.

[illegible]

Question type

- Short answer

QUESTION 1

- [2 marks] Briefly describe the meaning of association rule in association rule mining
- [2 marks] Briefly describe the meaning of strong association rule in association rule mining
- [2 marks] Given the frequent k-itemsets, list the steps to generate the frequent (k+1)-itemsets from frequent k-itemsets.

[illegible]

QUESTION 2

QUESTION 5

Question type

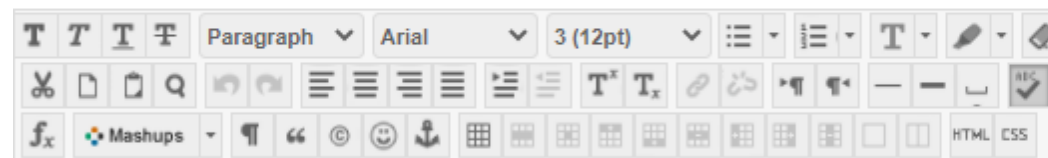
- Problem Solving

We have some data about whether people go to watch a football game. The data includes three attributes: whether the person is a student, whether they are a friend of the star in the game or not:

Tid	Weekend	Friends	Star there	Watch?
1	No	Yes	No	No
2	Yes	Yes	No	No
3	Yes	Yes	Yes	No
4	Yes	Yes	Yes	Yes
5	No	Yes	Yes	Yes
6	No	Yes	No	Yes
7	Yes	Yes	No	No
8	Yes	Yes	Yes	Yes
9	Yes	No	No	Yes
10	Yes	No	No	No
11	No	No	No	Yes

- [5 marks] Construct a decision tree from the provided dataset to predict whether people will watch the football game.
- [2 marks] Briefly discuss the advantage of using Gain Ratio-based splitting criterion compared to Information Gain-based.
- [2 marks] Briefly describe the two pruning methods in decision tree.

Decision tree construction :
Only show the constructed tree. No procedure needed.



Question type

- Calculation

QUESTION 6

Given a data set below, answer the following questions.

Tid	Home Owner	Marital Status	Annual Income	Class
1	Yes	Single	125K	No
2	No	Married	120K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	120K	No
8	No	Married	85K	Yes
9	No	Married	75K	No
10	No	Single	120K	Yes

a. [2 marks] What is the "Attribute conditional independence" assumption for Naïve Bayesian classifier?

b. [4 marks] Predict the class label of a given test record X: (Home Owner: Yes, Marital Status: Single, Annual Income: 80K) using the Naïve Bayesian Classifier.

Naïve bayes prediction problem:

Briefly show the procedure: (only an example of procedure, not the answer to this question)

~~$P(X|Yes)$ is in proportion to~~ $P(Yes)P(HO = Yes|Yes) P(MS = Single|Yes) P(AI = 80K|Yes) =$
 $0.1*0.2*0.3*0.4 = 0.0024$

~~$P(X|No)$ is in proportion to~~ $P(No)P(HO = Yes|No) P(MS = Single|No) P(AI = 80K|No) =$
 $0.5*0.6*0.7*0.8 = 0.168 > 0.0024$

so we predict it as No

Question type

- Calculation

QUESTION 7

Given a two-dimensional data set below, answer the following questions.

(x, y)	Class
(0, 5)	+
(2, 3)	+
(4, 3)	+
(5, 5)	-
(6, 7)	-
(8, 5)	-
(3, 7)	+
(4, 10)	-
(7, 9)	+

a. [1 mark] What is the basic idea of a k-NN method?

b. [2 marks] Classify a new data point $p = (4, 5)$ using Manhattan distance and majority vote by 3- nearest neighbours.

c. [2 marks] Briefly discuss when we need to normalize the attribute values in k-NN classification and list the names of

No need to do normalization without explicitly asked

Briefly show the procedures:



The Manhattan distance from (a, b) to all the 9 points are: 1, 2, 3, 4, 5, 6, 7, 8, 9. The 3-nearest neighbours are (x1, y1) +, (x2, y2) -, (x3, y3) +. By majority voting, (a,b) is classified as +

- Please note that the questions may not be ranked from the easiest to the hardest

CONTACT time

Q&A before the final exam:

- Nov. 17th 4:00 pm to 6:00 pm
- Nov. 18th 4:00 pm to 6:00 pm

Help links besides technical

- Preparing for online exam success

[Library services](#) ▼ [Research tools & techniques](#) ▼ [Collections](#) ▼ [Borrowing & requesting](#) ▼

Preparing for online exam success

[Home](#) / [Library services](#) / [For students](#)

These strategies will help you prepare for and avoid issues during your online exams.

- [Check your IT set up](#)
- [Access the Library resources you need](#)
- [Reading time is included in the exam duration](#)
- [Save your work progress and evidence regularly](#)
- [Time allowance for submitting and uploading your completed online exam](#)
- [eAssessment in Semester 2, 2020 - Inspira Assessment](#)

Check your IT set up

Online supervised (invigilated) exams

[Check what you need to do to prepare before the online supervised \(invigilated\) exam.](#) Also read the [ProctorU FAQs](#).

All online exams

Help links besides technical

- <https://my.uq.edu.au/information-and-services/student-support/health-wellbeing>

Health and wellbeing

We provide a range of programs and counselling services to help improve your confidence and overall physical and mental wellbeing.

Check out some common help topics or keep scrolling for more information.

- > [Support for anxiety](#)
- > [Increasing self-confidence](#)
- > [Support for loneliness](#)
- > [Make an appointment with a UQ counsellor](#)
- > [Mental health & wellbeing at UQ](#)
- > [Mental Health Champions](#)
- > [Request a student welfare check](#)
- > [Wellbeing workshops](#)

thank you

tusind tak
謝謝 dakujem vám
ngiyabonga
dziękuję
merc
baie dankie
धन्यवाद molte grazie
gracias
obrigada
obrigado
teşekkür ederim
tack så mycket
gràcies
tānan
dank u
teşekkür edire
mahalo
suksema
danke
tak

A yellow and black pencil is positioned diagonally on the right side of the image. To its left is a large, fan-shaped pencil shaving and a pile of dark pencil shavings. The background is a sheet of white graph paper with a light blue grid.

Good Luck on your Exam!

Give your best shot on it

I am pretty confident that you can make it

My best wishes are with you!