UFMG UNIVERSIDADE FEDERAL DE MINAS GERAIS

Recommender Systems
Content-Based

Rodrygo Santos
rodrygo@dcc.ufmg.br

How to recommend?

UFMG COMPUTER SCIENCE

user profile

Recommender

item	score
1	0.7
2	0.3
…	…

How to recommend?

UFMG COMPUTER SCIENCE

user profile *community data*

Recommender

Collaborative filtering
"tell me what's popular among my peers"

item	score
1	0.7
2	0.3
…	…

What if we have new users or items?

The cold-start problem

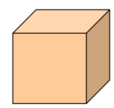
UFMG COMPUTER SCIENCE

Cold-start user

?

- Sparse user ratings
  - Poor predictions
- No user ratings
  - No personalization

Cold-start item



?

- Sparse item ratings
  - Poor predictions
- No item ratings
  - ***Infeasible prediction***

How to recommend?

UFMG COMPUTER SCIENCE

user profile

title	genre	year

item features

Recommender

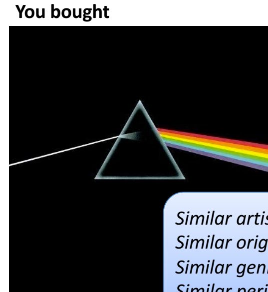
Content-based
"show me more of the same what I've liked"

item	score
1	0.7
2	0.3
…	…

Content-based recommendation

UFMG COMPUTER SCIENCE

You bought



You may like

Similar artist: Pink Floyd
Similar origin: England
Similar genre: Rock
Similar period: 1970s

Content-based recommendation

Collaborative filtering
- Leverages item ratings
- Agnostic to item content

Content-based filtering
- Leverages item content
- Agnostic to item ratings

*Applicable **to any kind of item** (e.g., text, audio, video, food)*

*Applicable even in **extreme cold-start** scenarios*

Same basic idea

- Stable preferences
  – News: I prefer technology, travel
  – Music: I prefer rock, grunge, folk
  – Clothing: I prefer cotton, casual
  – Movies: I prefer sci-fi, thrillers

Advantages

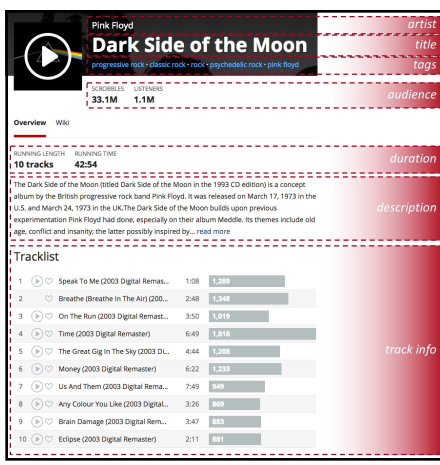
- No need for data on other users
  – No cold-start or sparsity problems
  – Able to recommend to users with unique tastes
- Able to recommend new and unpopular items
  – No first-rater problem
- Can provide explanations of recommended items by relevant content features
  – More on explanations later in the course

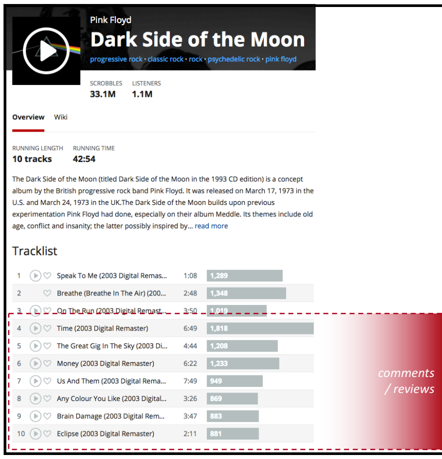
Challenges and drawbacks

- Content-based techniques in general...
  – Depend on well-structured attributes that align with preferences (consider paintings)
  – Depend on having a reasonable distribution of attributes across items (and vice versa)
  – Unlikely to find surprising connections (e.g., chili peppers or lemon with chocolate)
  – Harder to find complements than substitutes

What is "content"?

- It can be structured text
  – Artist: Pink Floyd
  – Genre: Rock
  – Year: 1973
- It can be unstructured text
  – Several techniques to extract content features
  – Several techniques to compute item similarity
- It can be derived from binary data
  – Audio, video, image





UFMG COMPUTER SCIENCE

Representing items

UFMG COMPUTER SCIENCE

	Artist	Title	Duration	Listeners	Tags	Description
i_1	Pink Floyd	Dark Side of the Moon	42:54	1.1M	progressive classic psychedelic pink floyd	The Dark Side of the Moon (titled Dark Side of the Moon in the 1993 CD edition) is a concept album by the British band Pink Floyd ...
i_2	Pink Floyd	The Wall	87:15	480K	70s classic progressive concept	The Wall is a rock opera presented as a double album by the English progressive rock band Pink Floyd, released on ...

Representing users

UFMG COMPUTER SCIENCE

	Artist	Title	Duration	Listeners	Tags	Description
i_1	pink floyd	dark side of the moon	42:54	1.1M	progressive classic psychedelic pink floyd	the dark side of the moon (titled dark side of the moon in the 1993 cd edition) is a concept album by the british band pink floyd ...
i_2	pink floyd	the wall	87:15	480K	70s classic progressive concept	the wall is a rock opera presented as a double album by the english progressive rock band pink floyd, released in november 1979 ...

	Artist	Title	Duration	Listeners	Tags	Description
u_1	pink floyd	dark side of the moon the wall	65:04	790K	progressive classic psychedelic 70s	dark side moon concept album british band pink floyd wall rock november 1979 ...

Making predictions

UFMG COMPUTER SCIENCE

	Artist	Title	Duration	Listeners	Tags	Description
u_1	pink floyd	dark side of the moon the wall	65:04	790K	progressive classic psychedelic 70s	dark side moon concept album british band pink floyd wall rock november 1979 ...

	Artist	Title	Duration	Listeners	Tags	Description
i_3	led zeppelin	led zeppelin iv	44:38	888.6K	classic rock rock hard rock 70s	led zeppelin iv is the common, but unofficial name of the untitled fourth album of english rock band led zeppelin release in ...

- Simple solution
  - Keyword overlap (e.g. Dice coefficient)

$$sim(u_1, i_3) = \frac{2|k(u_1) \cap k(i_3)|}{|k(u_1)| + |k(i_3)|}$$

Are we done yet?

UFMG COMPUTER SCIENCE

Tokenization

UFMG COMPUTER SCIENCE

- How to split...
  - information retrieval?
    - information + retrieval
  - 信息检索?
    - 信息 + 检索
- We can analyze term statistics
  - Probability of segmentation
- Also applicable in other scenarios
  - Domain names, hashtags, etc.

Term normalization

UFMG COMPUTER SCIENCE

- I am interested in *"information retrieval"*
  - i_1 contains *"retrieval"*
  - i_2 contains *"retrieving"*
  - i_3 contains *"retrieved"*

    …
- ***Stemming*** reduces words to a root form
  - *"retrieval"* → *"retriev"*
  - *"retrieving"* → *"retriev"*
  - *"retrieved"* → *"retriev"*

Term frequency

UFMG COMPUTER SCIENCE

- I am interested in *"information retrieval"*
  - i_1 contains *"information retrieval" once*
  - i_2 contains *"information retrieval" ten times*
- Intuitively, ***term frequency*** denotes how much the item is about the particular term
  - Also applicable to n-grams

Term frequency

UFMG COMPUTER SCIENCE

- I am interested in *"information retrieval"*
  - i_1 contains *"information retrieval" once*
    - i_1 has a total of 10 terms
  - i_2 contains *"information retrieval" ten times*
    - i_2 has a total of 100,000 terms
- Intuitively, long items may yield high frequency terms by chance
  - Content ***length normalization*** may help (next class)

Term proximity

UFMG COMPUTER SCIENCE

- I am interested in *"information retrieval"*
  - i_1 contains *"**information retrieval**"*
  - i_2 contains *"**retrieval** of spatial memory in the hippocampus … the theory of constructive recollection asserts that **information** …"*
- Once again, ***co-occurrence stats*** may help
  - Index *"information retrieval"* as a unit
  - Or record the position of each term
- Alternatively, we can identify ***concepts***

Term informativeness

UFMG COMPUTER SCIENCE

- I am interested in *"information retrieval"*
  - i_1 contains *"information"*
  - i_2 contains *"retrieval"*
- Which item should be ranked first?
  - "information" occurs in 35% of all items
  - "retrieval" occurs in 0.1% of all items
- Intuitively, the ***scarcity*** of a term makes its occurrence more informative

Content structure

UFMG COMPUTER SCIENCE

- I am interested in *"information retrieval"*
  - i_1 contains *"information retrieval"* in the title
  - i_2 contains *"information retrieval"* in the body
  - i_3 contains *"information retrieval"* in the URL
- Different fields may convey a different measure of the informativeness of a term
  - ***Field-based term weighting*** may help
  - Typically a machine learning task

Content enrichment

UFMG COMPUTER SCIENCE

- I am interested in “information retrieval”
  - i_1 contains *“search engines”*
  - i_2 contains *“recommender systems”*
- How can they be retrieved?
  - Leverage external databases
    - Lexical databases, knowledge bases
  - Leverage user-generated content
    - Tags, anchor-text (user annotations)
    - Views, clicks, purchases (user feedback)

Content quality

UFMG COMPUTER SCIENCE

- I am interested in *“information retrieval”*
  - i_1 is a book by Manning et al. ***(authority)***
  - i_2 is an entry in Wikipedia ***(readability)***
  - i_3 is a spam page ***(trustworthiness)***
  - i_4 is a best seller ***(popularity)***
  - i_5 is brand new ***(freshness)***
- Several measures of “quality”
  - A-priori notion of relevance, helping distinguish between items with similar topicality

Summary

UFMG COMPUTER SCIENCE

- CB recommendation works for new items
  - Not for new users (still need ratings)
- Keywords alone may not suffice
  - Freshness, usability, aesthetics, writing style
  - Content may also be limited / too short
  - Content may not be automatically extractable
- Overspecialization
  - Algorithms tend to propose “more of the same”

Writing Assignment #4
Due May 1st @ 23:55

UFMG COMPUTER SCIENCE

- Write a one-page summary describing the following paper:
  - Performance of recommender algorithms on top-n recommendation tasks (RecSys 2010) by Paolo Cremonesi, Yehuda Koren, Roberto Turrin