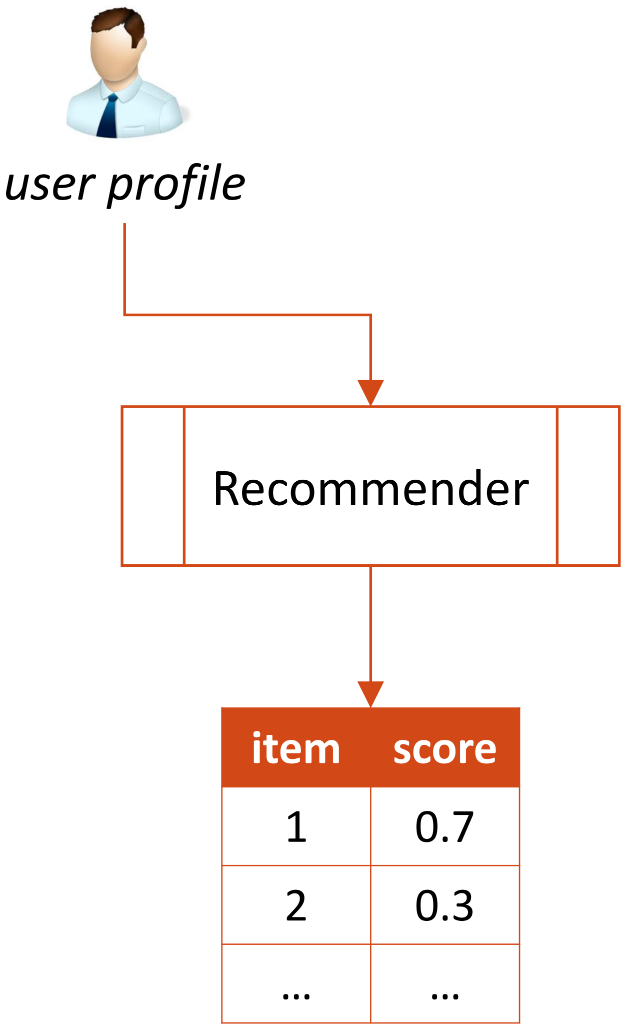
UFMG UNIVERSIDADE FEDERAL DE MINAS GERAIS

Recommender Systems

Content-based Recommendation

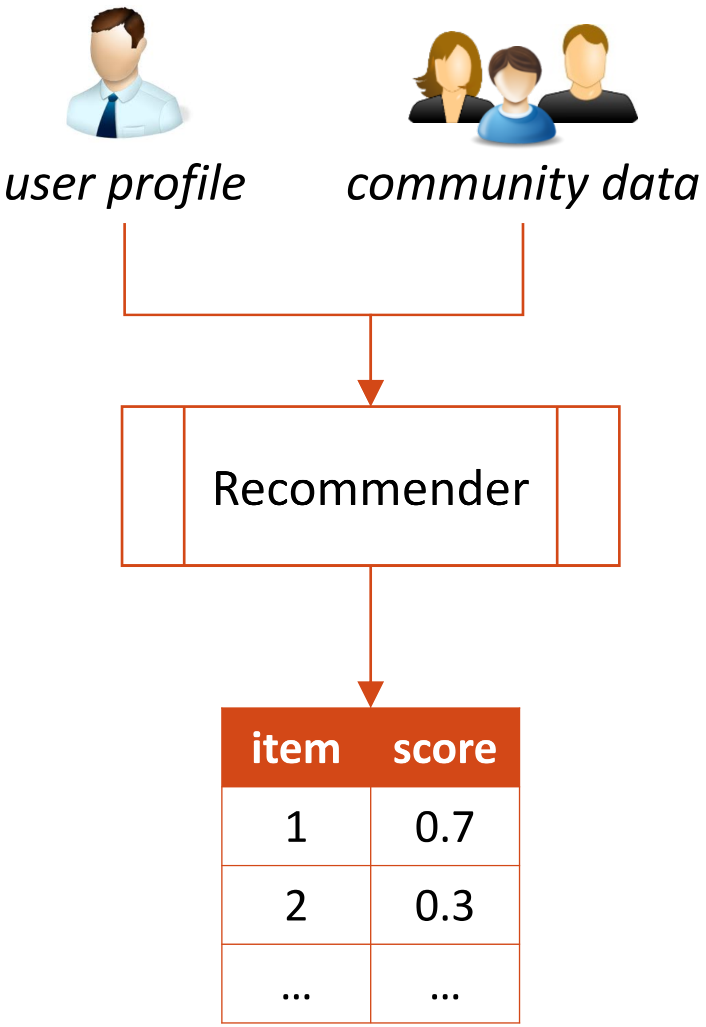
Rodrygo L. T. Santos
rodrygo@dcc.ufmg.br

How to recommend?



item	score
1	0.7
2	0.3
…	…

How to recommend?



user profile

community data

Recommender

item	score
1	0.7
2	0.3
…	…

Collaborative filtering

“tell me what’s popular among my peers”

What if we have new users or items?

The cold-start problem

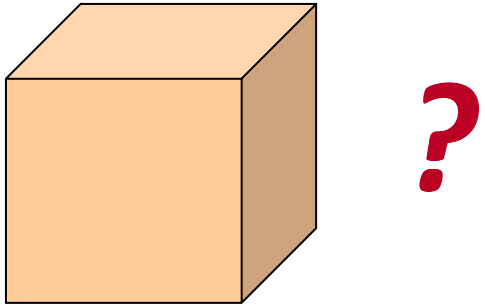
Cold-start user

Sparse user ratings

- Poor predictions

No user ratings

- No personalization



Cold-start item

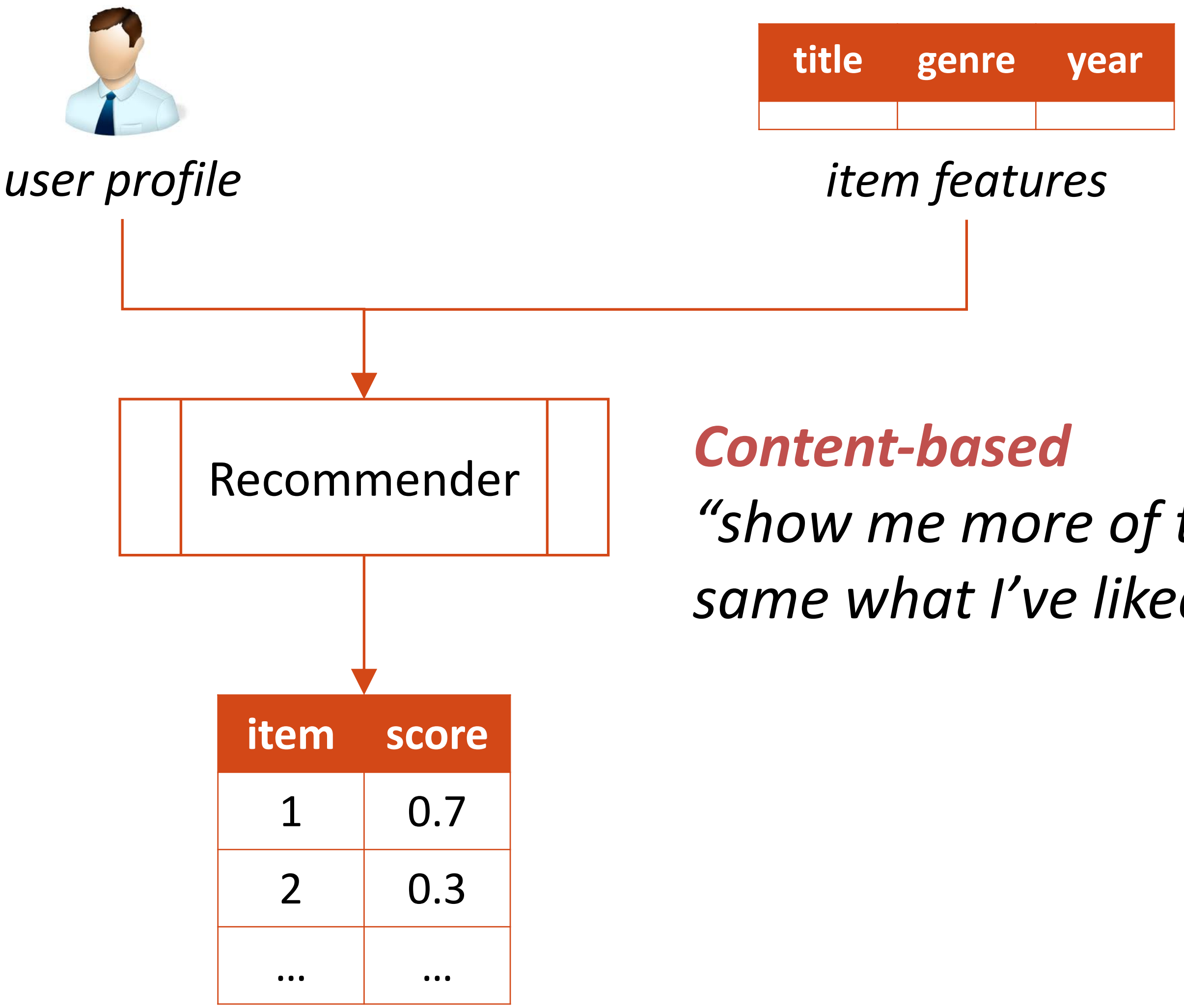
Sparse item ratings

- Poor predictions

No item ratings

- ***Infeasible prediction***

How to recommend?



user profile

title	genre	year

item features

Recommender

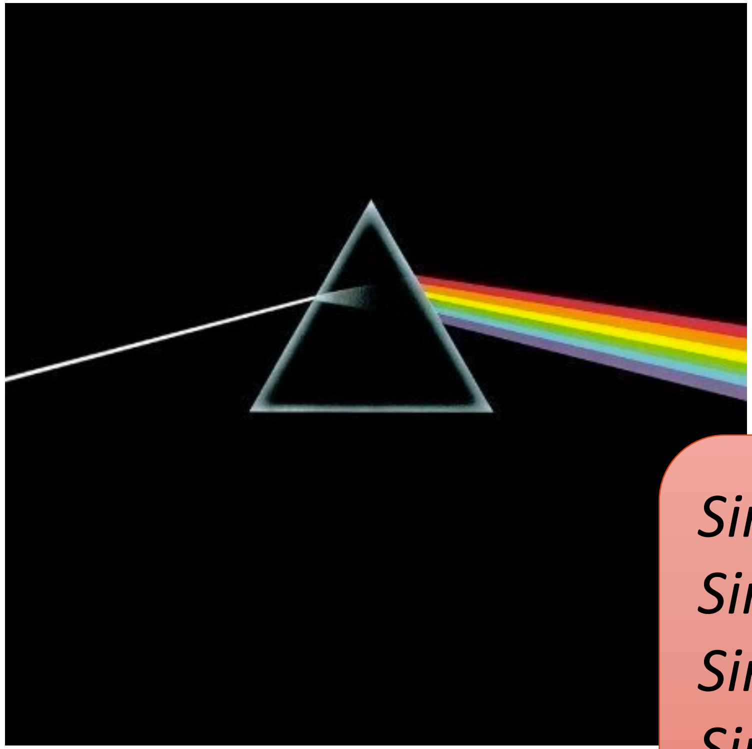
item	score
1	0.7
2	0.3
…	…

Content-based

“show me more of the same what I’ve liked”

Content-based recommendation

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

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

Content-based filtering

- Leverages item content
- Agnostic to item ratings

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

- Able to recommend to users with unique tastes

Able to recommend new and unpopular items

- No first-rater problem

Can provide explanations based on 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
- Harder to find complements than substitutes

What is “content”?

It can be structured text

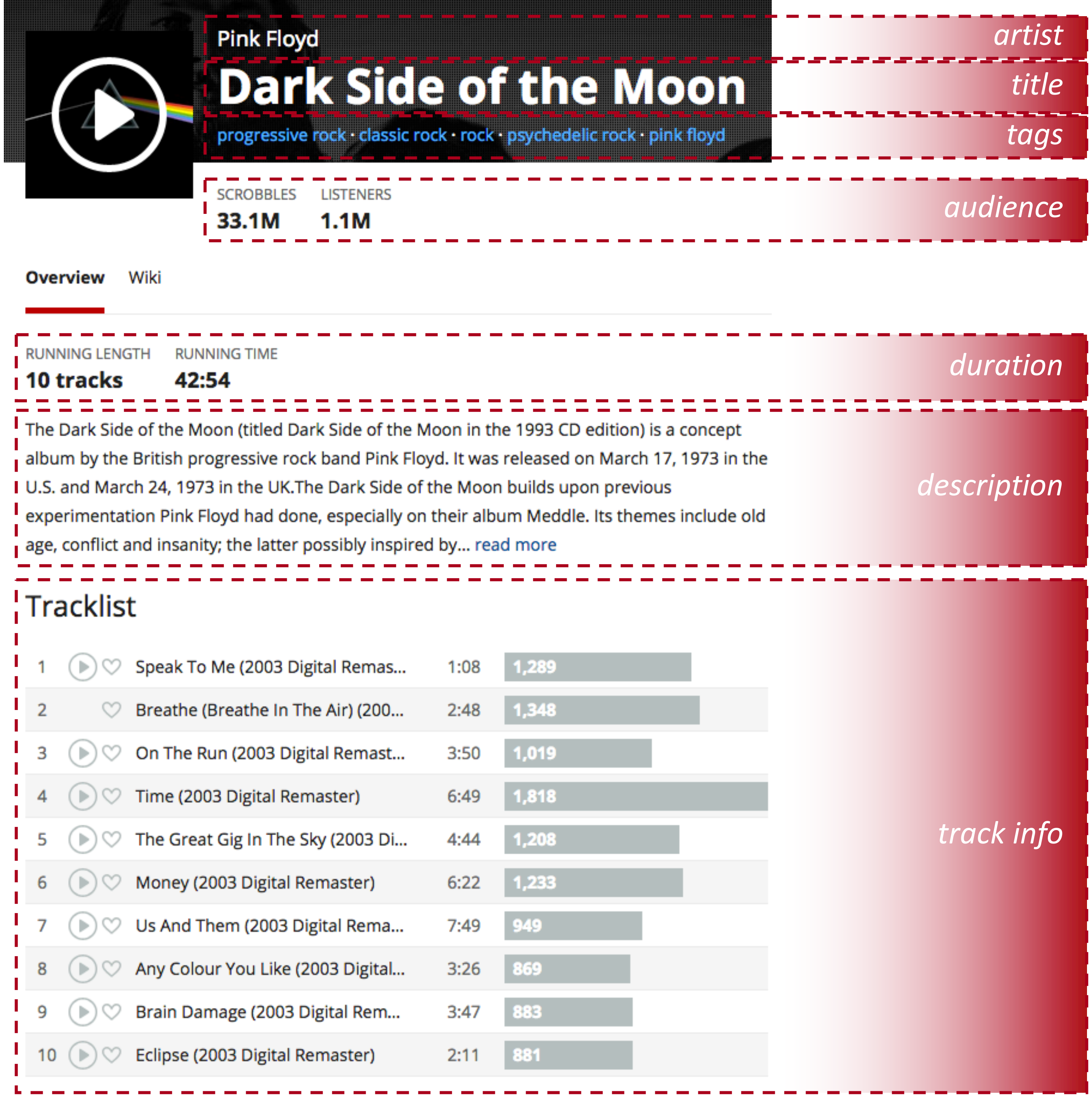
- Artist: Pink Floyd; Genre: Rock; Year: 1973

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



Pink Floyd

Dark Side of the Moon

progressive rock · classic rock · rock · psychedelic rock · pink floyd

SCROBBLES 33.1M

LISTENERS 1.1M

Overview Wiki

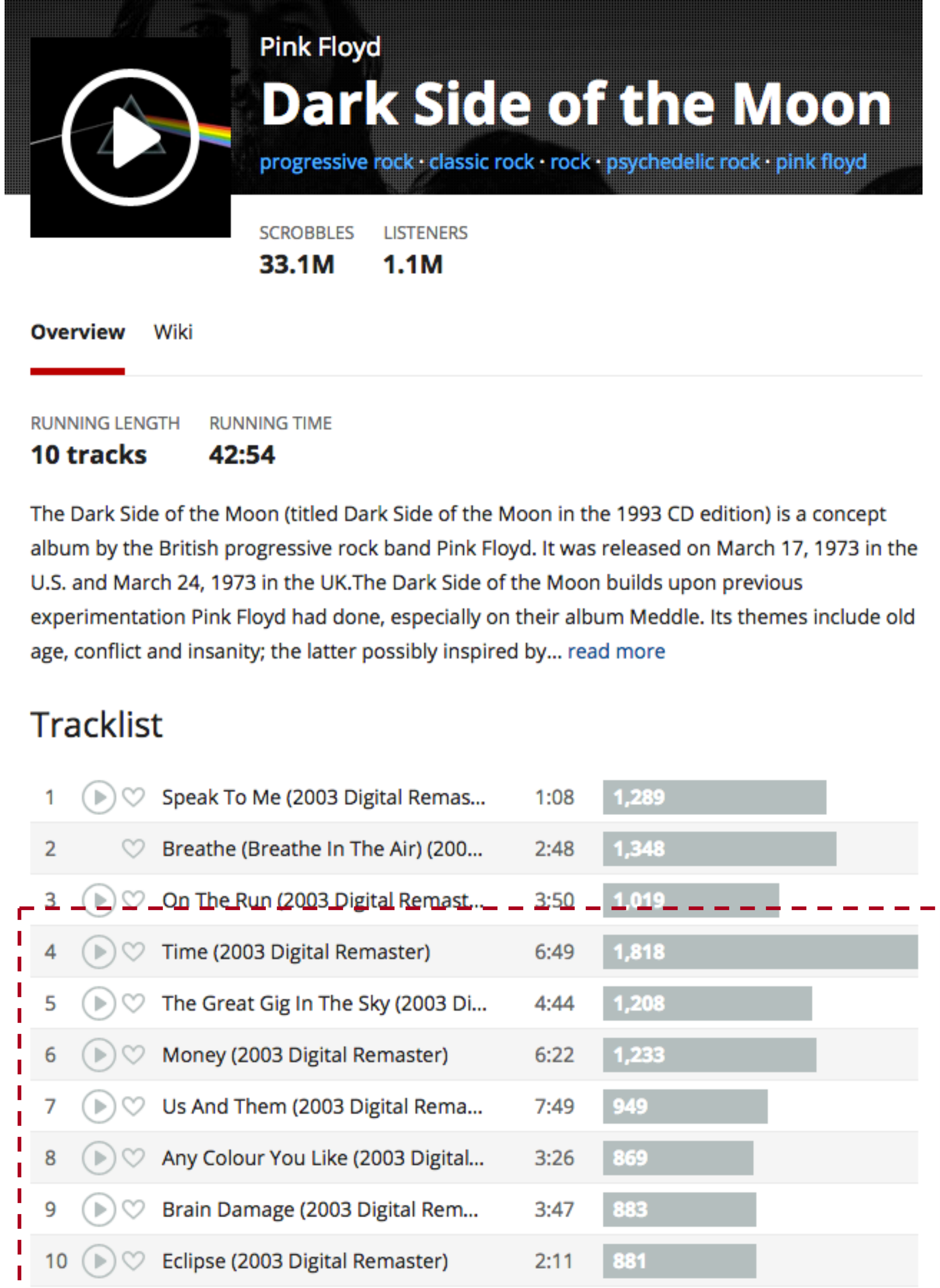
RUNNING LENGTH 10 tracks

RUNNING TIME 42:54

The Dark Side of the Moon (titled Dark Side of the Moon in the 1993 CD edition) is a concept album by the British progressive rock band Pink Floyd. It was released on March 17, 1973 in the U.S. and March 24, 1973 in the UK.The Dark Side of the Moon builds upon previous experimentation Pink Floyd had done, especially on their album Meddle. Its themes include old age, conflict and insanity; the latter possibly inspired by… read more

Tracklist

1	Speak To Me (2003 Digital Remas…	1:08	1,289
2	Breathe (Breathe In The Air) (200…	2:48	1,348
3	On The Run (2003 Digital Remast…	3:50	1,019
4	Time (2003 Digital Remaster)	6:49	1,818
5	The Great Gig In The Sky (2003 Di…	4:44	1,208
6	Money (2003 Digital Remaster)	6:22	1,233
7	Us And Them (2003 Digital Rema…	7:49	949
8	Any Colour You Like (2003 Digital…	3:26	869
9	Brain Damage (2003 Digital Rem…	3:47	883
10	Eclipse (2003 Digital Remaster)	2:11	881



Pink Floyd

Dark Side of the Moon

progressive rock · classic rock · rock · psychedelic rock · pink floyd

SCROBBLES 33.1M

LISTENERS 1.1M

Overview Wiki

RUNNING LENGTH 10 tracks

RUNNING TIME 42:54

The Dark Side of the Moon (titled Dark Side of the Moon in the 1993 CD edition) is a concept album by the British progressive rock band Pink Floyd. It was released on March 17, 1973 in the U.S. and March 24, 1973 in the UK.The Dark Side of the Moon builds upon previous experimentation Pink Floyd had done, especially on their album Meddle. Its themes include old age, conflict and insanity; the latter possibly inspired by… read more

Tracklist

1	Speak To Me (2003 Digital Remas…	1:08	1,289
2	Breathe (Breathe In The Air) (200…	2:48	1,348
3	On The Run (2003 Digital Remast…	3:50	1,019
4	Time (2003 Digital Remaster)	6:49	1,818
5	The Great Gig In The Sky (2003 Di…	4:44	1,208
6	Money (2003 Digital Remaster)	6:22	1,233
7	Us And Them (2003 Digital Rema…	7:49	949
8	Any Colour You Like (2003 Digital…	3:26	869
9	Brain Damage (2003 Digital Rem…	3:47	883
10	Eclipse (2003 Digital Remaster)	2:11	881

comments / reviews

Representing items

	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 ...

Representing users

	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

u_1

Artist	Title	Duration	Listeners	Tags	Description
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 ...

i_3

Artist	Title	Duration	Listeners	Tags	Description
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?

Tokenization

How to split…

- information retrieval?
  - information + retrieval
- 信息检索?
  - 信息 + 检索

We can analyze term statistics

- Probability of segmentation

Term normalization

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” / “retrieving” / “retrieved”* → *“retriev”*

Term frequency

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

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

Long items may yield high frequency terms by chance

- ◦ Content ***length normalization*** may help (next class)

Term proximity

I am interested in *"information retrieval"*

- *i_1* contains *"**information retrieval**"*
- *i_2* contains *"**retrieval** of spatial memory in the brain … recollection asserts that **information** …"*

Once again, ***co-occurrence stats*** may help

- Index *"information retrieval"* as a unit
- Or record the position of each term

Term informativeness

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

Scarcity makes term occurrences more informative

Content structure

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 convey different importance of a term

- ***Field-based term weighting*** may help

Content enrichment

I am interested in “information retrieval”

- i_1 contains *“search engines”*
- i_2 contains *“recommender systems”*

How can they be retrieved?

- Leverage external databases (e.g. knowledge bases)
- Leverage user-generated content (e.g. annotations, implicit and explicit feedback)

Content quality

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 best seller ***(popularity)***
- i_4 is brand new ***(freshness)***

Several a-priori measures of “quality”

- Help distinguish between items with similar topicality

Summary

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 be limited, not automatically extractable

Overspecialization

- Algorithms tend to propose “more of the same”

References

Recommender Systems: An Introduction (Sec. 3.1)

Recommender Systems Handbook (Sec. 3.2)

Recommender Systems: The Textbook (Sec. 4.1-4.2)