Building a Large-scale News Evaluation Data Set

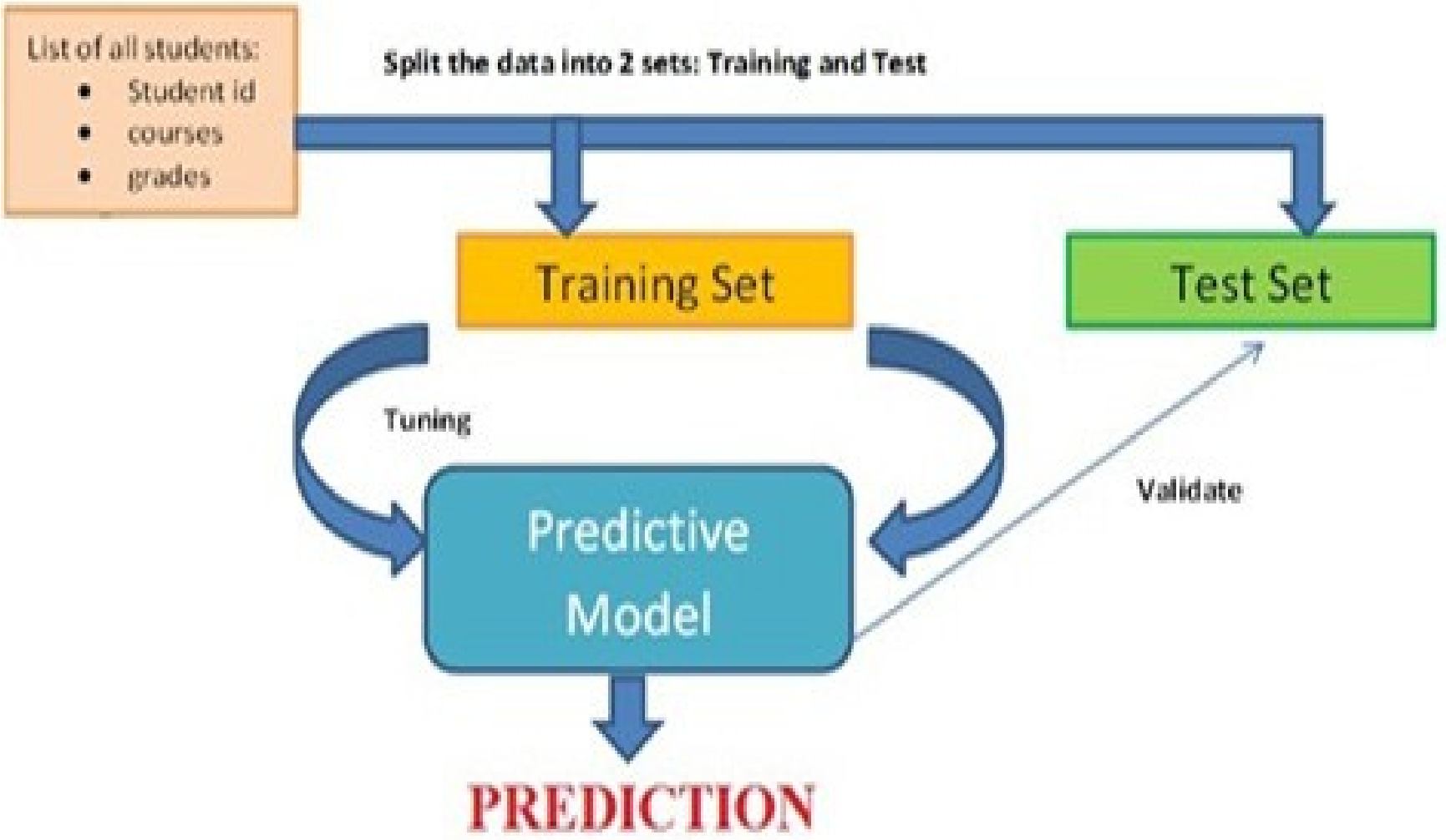
Özlem Özgöbek
Ege University/NTNU

ozlemo@idi.ntnu.no

Data Sets

- Collection of data.
- Real systems work on data.
- Needed to build effective systems.

Data Sets on Recommender Systems



Data Set Examples On Recommender Systems

http://grouplens.org/datasets

MovieLens

(University of Minnesota GroupLens Research Group)

- 100,000 ratings (1-5) from 1000 users on 1700 movies in seven-month period.
  - Each user has rated at least 20 movies.
  - Simple demographic info for the users (age, gender, occupation, zip)
  - Users who had less than 20 ratings or did not have complete demographic information were removed from this data set.

MovieLens

(University of Minnesota GroupLens Research Group)

Users	Movies	Ratings
1000	1700	100.000
6000	4000	1.000.000
72.000	10.000	10.000.000

UserID::MovieID::Rating::Timestamp

```
1::122::5::838985046
1::185::5::838983525
1::231::5::838983392
1::292::5::838983421
1::316::5::838983392
1::329::5::838983392
1::355::5::838984474
1::356::5::838983653
1::362::5::838984885
1::364::5::838983707
1::370::5::838984596
1::377::5::838983834
1::420::5::838983834
1::466::5::838984679
1::480::5::838983653
1::520::5::838984679
1::539::5::838984068
1::586::5::838984068
1::588::5::838983339
1::589::5::838983778
1::594::5::838984679
1::616::5::838984941
2::110::5::868245777
2::151::3::868246450
2::260::5::868244562
```

MovieID::Title::Genres

```
1::Toy Story (1995)::Adventure|Animation|Children|Comedy|Fantasy
2::Jumanji (1995)::Adventure|Children|Fantasy
3::Grumpier Old Men (1995)::Comedy|Romance
4::Waiting to Exhale (1995)::Comedy|Drama|Romance
5::Father of the Bride Part II (1995)::Comedy
6::Heat (1995)::Action|Crime|Thriller
7::Sabrina (1995)::Comedy|Romance
8::Tom and Huck (1995)::Adventure|Children
9::Sudden Death (1995)::Action
10::GoldenEye (1995)::Action|Adventure|Thriller
11::American President, The (1995)::Comedy|Drama|Romance
12::Dracula: Dead and Loving It (1995)::Comedy|Horror
13::Balto (1995)::Animation|Children
14::Nixon (1995)::Drama
15::Cutthroat Island (1995)::Action|Adventure|Romance
16::Casino (1995)::Crime|Drama
17::Sense and Sensibility (1995)::Comedy|Drama|Romance
18::Four Rooms (1995)::Comedy|Drama|Thriller
19::Ace Ventura: When Nature Calls (1995)::Comedy
20::Money Train (1995)::Action|Comedy|Crime|Drama|Thriller
21::Get Shorty (1995)::Action|Comedy|Drama
22::Copycat (1995)::Crime|Drama|Horror|Mystery|Thriller
23::Assassins (1995)::Action|Crime|Thriller
24::Powder (1995)::Drama|Sci-Fi
25::Leaving Las Vegas (1995)::Drama|Romance
26::Othello (1995)::Drama
27::Now and Then (1995)::Drama
28::Persuasion (1995)::Drama|Romance
```

The 2nd International Workshop on Information Heterogeneity and Fusion in Recommender Systems (HetRec 2011)

• They released datasets from Delicious, Last.fm Web 2.0, MovieLens, IMDb, and Rotten Tomatoes.

• These datasets contain social networking, tagging, and resource consuming (Web page bookmarking and music artist listening) information from sets of around 2,000 users.

	Users	
Delicious Bookmarks	1867	105.000 bookmarks
Last.FM	1892	92.800 listening records
MovieLens + IMDb/Rotten Tomatoes	2113	86.000 ratings

Data Set on News Domain:

From VTT Technical Research Centre of Finland, provided courtesy of Arena Partners.

UserId	SessionId	Timestamp	ItemId	ItemCategory
798	4810	238300	1048577	4
798	4810	238365	1048578	4
1744	2420	2560354	1048579	3
580	2096	2579359	1048579	3
1744	2420	2560247	1048580	3
1628	4763	2576025	1048580	3
580	2096	2579128	1048580	3
541	4643	2547761	1048581	3
1744	2420	2560285	1048581	3
1628	4763	2576072	1048581	3
580	2096	2579203	1048581	3
474	5530	693085	1048582	9
474	5530	694934	1048582	9
474	5530	695002	1048582	9
474	5530	693022	1048583	4
474	5530	693036	1048583	4
541	3514	1942401	1048584	3
836	5085	1962486	1048584	3
856	3223	1971168	1048584	3
580	319	1977350	1048584	3
1181	1153	1992660	1048584	3
1645	2118	2017987	1048584	3
1399	5397	117246	1048585	4
1159	2374	2068001	1048586	4
1159	2374	2067979	1048587	4
1159	2374	2068286	1048588	4
1159	2374	2068368	1048589	4
1159	2374	2068489	1048590	4
1159	2374	2068581	1048591	4
1159	2374	2068601	1048592	4
1159	2374	2068895	1048593	6
1159	2374	2068842	1048594	4

No proper open data set:

- News domain
- Norwegian

Without a proper data set it is not possible to develop a useful recommender system.

Our Data Set

Two kinds of data will be collected: Explicit & Implicit

- We will ask users to register the system.
- Some data (mostly demographic information) will be collected during the registration.
- More data will be collected during usage.

On user registration:

- Name
- Birth date (dd/mm/yyyy)
- Sex
- E-mail address
- Occupation
- City
- News category preferences (top 3)

On usage:

• User must see more than one article summary or title on the screen.

• User should be able to rate each article (5 step ratings like "Hate", "Dislike", "Indifferent", "Like", "Love" and it must be clearly described what the user rates for).

• Time spent on each article must be recorded. (<timestamp, user id, article id, client type, location> for all views.)

• News articles must be displayed in the same format (same background color and font types) to avoid the influence of different graphical interfaces on the user.

Two groups of users:

- Selected user group (2 weeks)
- Normal user group (several months)

Selected user group:

- 50-60 people
- 2 weeks
- Regular usage of the system
- Proper ratings for all read articles
- Target: 25.000 ratings

Normal user group:

- Open for everyone to use the system
- Ratings may be missing or wrong
- Noisy data

Expectations from the selected user group:

- Registration data entrance.
- Everyday usage of the system.
- Rating all the articles they read (both liked and disliked).
- 500 articles within 2 weeks.
- They can continue to use the system after 2 weeks.
- A chance of winning a surprise!

The resulting data set can also be used for:

- Customer intelligence for media houses.
- Feedback about the general satisfaction of news articles.
- Internal evaluation.

Outcomes:

- A clean and correct piece of data set (training data)
- A larger but noisy data set (close to real life applications)
- We might use it to evaluate content-based filtering, collaborative filtering and other approaches.
- It will be open to anyone who wants to use the data set.

Not a very large data set for the beginning.
Useful enough to develop and evaluate an effective recommender system.

Thank you!