Combining Content with User Preferences for TED Lecture Recommendations

CBMI 2013
Nikolaos Pappas
Andrei Popescu-Belis

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE
idiap INSTITUT DE RECHERCHE
inEvent Accessing Dynamic Networked Multimedia Events

inEvent project (https://www.inevent-project.eu)
Idiap Research Institute, Martigny, Switzerland
Wednesday 19th June, 2013

Motivation

Recommender systems are information filtering systems that seek to predict ratings (preferences) for items that might be of interest to a user.

- divided in *content-based* (CB), *collaborative filtering* (CF) and *hybrid*
- plenty of data available on certain domains (movies, music, etc.)
- fewer for multimedia content (e.g. VideoLectures)

Questions – multimedia recommendations

→ How to perform quantitative experiments with 'objective' measures?

→ Which data to use for evaluation?

→ How important is content vs. collaborative information?

Summary

Recommendation methods for scientific talks

1. studying the merits of CB and CF methods over TED talks
2. evaluating in two different scenarios: cold-start, non-cold-start (absence or presence of collaborative information)

Main contributions

- → Introduction of TED dataset for multimedia recommendations
- → Definition of evaluation tasks over TED
- → Combining content features with user preferences
- → First benchmark scores on this promising dataset

The TED collection

TED is an online repository of lectures (ted.com) which contains:

- audiovisual recordings of talks with extended metadata
- user-contributed material (comments, favorites)

	Total	**Per Talk**		**Per Active User**	
Attribute	Count	Average	Std	Average	Std
Talks	1,149	-	-	-	-
Speakers	961	-	-	-	-
Users	69,023	-	-	-	-
Active Users	10,962	-	-	-	-
Tags	300	5.83	2.11	-	-
Themes	48	2.88	1.06	-	-
Related Videos	3,002	2.62	0.74	-	-
Transcripts	1,102	0.95	0.19	-	-
Favorites	108,476	94.82	114.54	9.89	20.52
Comments	201,934	176.36	383.87	4.87	23.42

We crawled (Apr 2012), formatted and distributed the TED metadata:
https://www.idiap.ch/dataset/ted/ (in agreement with TED)

Ground truth

Typical problem: Given a rating matrix R $(|U| \times |I|)$ where R_{ui} is user's u explicit rating to item i; the goal is to find the value of missing ratings in R.

- Categorical ratings (e.g. good, bad)
- Numerical ratings (e.g. 1 to 5 stars)
- Unary or binary ratings (e.g. favorites or like/dislike)

On TED dataset we deal with unary ratings from user favorites:

$$R_{u,i} = \begin{pmatrix} r_{1,1} & r_{1,2} & \cdots & r_{1,n} \\ r_{2,1} & r_{2,2} & \cdots & r_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{m,1} & r_{m,2} & \cdots & r_{m,n} \end{pmatrix} e.g. \begin{pmatrix} 1 & 1 & ? & ? \\ ? & ? & ? & 1 \\ 1 & 1 & ? & ? \\ 1 & ? & 1 & ? \end{pmatrix}$$

$\rightarrow$ uncertainty about the negative class (one-class problem)
$\rightarrow$ related/similar talks available (TED editorial staff)

Recommendation tasks

1. **Personalized recommendation task**
   Ground-truth: user favorites (binary values), namely "1" for action and "0" or "?" for inaction (not seen, or seen and not liked).

   → Predict the N most interesting items for each user (top-N)
2. **Generic recommendation task**
   Ground-truth: related talks per talk assigned by TED editorial staff.

   → Predict the N most similar items to a given one (top-N)

How to evaluate?

As a top-N ranking problem: train a recommender (ranker) on fragments of user history and evaluate the performance on the held-out ones

→ for each user all items have to ordered based on a scoring function

→ information retrieval metrics to capture the performance (P, R, F1)

Comparison with other collections

Collection	Basic	Sp.	Trans.	Tags	Impl.	Expl.	CC
VideoLectures	✓	✓	✓		✓		
KhanAcademy	✓	✓			✓		
Youtube EDU	✓		✓		✓	✓	
DailyMotion	✓				✓	✓	
TED	✓	✓	✓	✓	✓	✓	✓

Basic: Title, Description
Sp.: Speaker
Tra.: Transcript
Tags: Categories in form of keywords
Impl.: Implicit feedback (e.g. comments or views)
Expl.: Explicit feedback (e.g. ratings, favorites or bookmarks)
CC: Creative Commons Non-Commercial License

Representations of TED talks

Each talk $t_j \in I$ is represented as a feature vector $t_j = (w_1, w_2, ..., w_{ij})$, where each position i corresponds to a word of the vocabulary $w \in V$.
Pre-processing:

$$I \rightarrow \textsc{Tokenization} \rightarrow \textsc{Stop words removal} \rightarrow \textsc{Stemming} \rightarrow V$$

Semantic Vector Space Models
Dimensionality reduction (LSI and RP), topic modeling (LDA) and concept-spaces built with external knowledge (ESA) vs. baseline (TF-IDF).

- diminish the curse of dimensionality effect
- proximity is interpreted as *semantic* relatedness

Comparison of their effectiveness in the recommendation task

Recommendation algorithms

Three types of nearest neighbor (NN) models for a given user u and talk i:
Content-based

$$\hat{r}_{ui} = \sum_{j \in D^k(u;i)} s_{ij}, \tag{1}$$

Collaborative filtering

$$\hat{r}_{ui} = b_{ui} + \sum_{j \in D^k(u;i)} d_{ij}(r_{uj} - b_{uj}), \tag{2}$$

$$b_{ui} = \mu + b_u + b_i, \tag{3}$$

Combined

$$\hat{r}_{ui} = b_{ui} + \sum_{j \in D^k(u;i)} s_{ij}(r_{uj} - b_{uj}), \tag{4}$$

d_{ij}: collaborative similarity of two items computed on the co-rating matrix.
s_{ij}: the content similarity of two items in the given vector space.

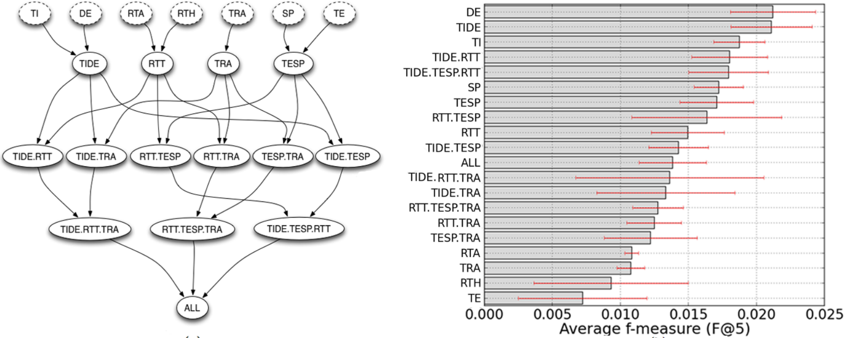
Parameter and feature selection

- → Parameters fixed for all NN models (k=3,$\lambda = 100$)
- → Parameters for VSMs optimized (dimensionality *k* for LSI, RP, LDA and priors α, β for LDA)
- → Features are the words extracted from the metadata

Method	**Optimal Features**	**Performance (%)**		
		P@5	R@5	F@5
LDA (t=200)	Title, desc., TED event, speaker (TIDE.TESP)	1.63	1.96	1.78
TF-IDF	Title (TI)	1.70	2.00	1.83
RP (t=5000)	Description (DE)	**1.83**	**2.25**	**2.01**
LSI (t=3000)	Title (TI)	**1.86**	**2.27**	**2.04**
ESA	Title, description (TIDE)	**2.79**	**3.46**	**3.08**

Table : CB performance with 5-fold c.-v. on the training set.

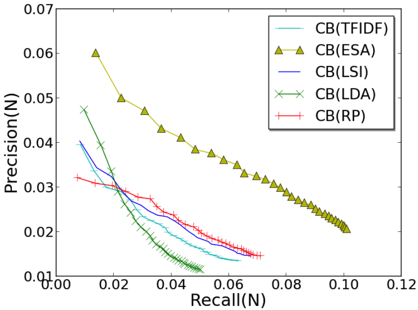
Feature ranking



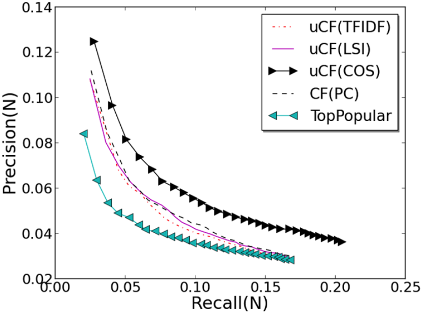
(a) (b)

Ranking based on the average F@5 over all methods with cross-validation.

Experiments on held-out data



(a) Comparison of CB methods in a cold-start setting (collaborative information is unavailable).



(b) Comparison of combined methods (uCF(LSI) and uCF(TFIDF) and CF methods (uCF(COS) and CF(PC)) in a non-cold-start setting (collaborative filtering information is available).

1. semantic spaces outperform keyword-based ones within CB methods
2. combined methods achieve reasonable performance compared to CF ones and they are applicable in both settings with good performance

Conclusions

- New dataset for lecture recommendation evaluation (ground-truth and rich content)
- Two recommendation benchmarks
- First experiments on personalized TED lecture recommendations
- We proposed to combine semantic spaces with CF methods

  → perform well in cold-start settings and can be used reasonably well in non-cold-start settings

  → applicable to multimedia datasets, where new items are inserted frequently (cold-start)

End of presentation

Thank you! Any questions?