

# Computer science research seminar: VideoLectures.Net recommender system challenge: presentation of baseline solution

Nino Antulov-Fantulin <sup>1</sup>,  
Mentors: Tomislav Šmuc <sup>1</sup> and Mile Šikić <sup>2 3</sup>

<sup>1</sup>Institute Rudjer Boskovic, Department of electronics, Zagreb, Croatia

<sup>2</sup>Faculty of Electrical Engineering and Computing, Department of Electronic Systems and Information Processing, University of Zagreb, Croatia

<sup>3</sup>Bioinformatics Institute, A\*STAR, Singapore, Republic of Singapore



# Outline

- VideoLectures.Net RS challenge
- Content-based RS
- LSI content-based RS
- Graph-based RS
- Collaborative-based RS
- Hybrid RS
- Learning phase: Stochastic gradient descent

# VideoLectures.Net RS challenge

VideoLectures challenge was organized as an official challenge of the ECML-PKDD 2011 conference, Athens, Greece. VideoLectures.Net challenge is planned to serve as a main use-case for the DMR domain of EU-FP7 e-LICO project.

Organizers:

- Nino Antulov-Fantulin, Matko Bosnjak, Tomislav Smuc - Rudjer Boskovic Institute
- Miha Grcar, Martin Znidarsic, Nada Lavrac, Mitja Jermol, Peter Kese - Jozef Stefan Institute





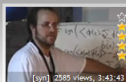
Uvodni nagovor 8.  
nanotehnološki dan 2010

Jadran Lenarčič



Behavioural Learning:  
Inspiration from Nature?

Chris Watkins



Introduction to Kernel Methods

Bernhard Schölkopf



Lecture 10 - Biomolecular  
Engineering: Engineering of  
Immunity (cont.)

W. Mark Saltzman



Telepathy

Senko Rašić



Lecture 21: Case study 3: spatial  
modeling

Duane S. Boning

## CATEGORIES

- Architecture (85)
- Arts (173)
- Astronomy (39)
- Biology (282)
- Biotechnical sciences (3)
- Business (895)
- Chemistry (156)
- Computers (296)
- Computer Science (6364)
- Criminology (8)
- Earth sciences (17)
- Economics (106)
- Education (140)
- Environment (141)

## NEWS

### COIN-PlanetData videos online

Dec. 27, 2011

We published the talks from the COIN-PlanetData school and we invite you to watch the invited talks by Michael Witbrock, Abraham B. Hsuan, and a great tutorial on Complex Event Processing by Roland Stühmer and a great session on Social software by Ugo Negretto presenting the social application that also VideoLectures.NET uses - LIVENETLIFE - Live Contextual Collaboration.

### AAAI Video Competition online now

Dec. 19, 2011

We published the AAAI 5th annual video competition! Its goal is to show the world how much fun AI is by documenting exciting artificial intelligence advances in research, education, and application.

## NEWSLETTER

Subscribe to our newsletter to receive digest of activity

Your Email:

Subscribe

## RECENT EVENTS


MORE

Naslednji val inovacij 2011 - Ljubljana



Naslednji val inovacij: Zapiranje snovnih poti

# TunedIT challenge site



Machine Learning & Data Mining Algorithms  
Automated Tests, Repeatable Experiments, Meaningful Results

News of Biomedical Research Papers ... New challenge is open: Topical Classification of Bio

not logged in - [login](#) | [register](#)

Search in Challenges

[Home](#) [About Us](#) [Research](#) [Challenges](#) [Outsourcing](#) [Wiki](#) [Blog](#) [Forum](#)

[List](#) [Create](#) [About](#)

## Challenges / VideoLectures.Net Recommender System Challenge

Contents

Overview

Summary

News

Tracks

1. Cold start
2. Pooled sequences

Workflow contest

Register

Forum

Overview

The challenge is over now. [Click here to view the Summary.](#)



Welcome to the web page of **ECML/PKDD Discovery Challenge 2011** (European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases).

The tasks of the challenge are focused on making recommendations for video lectures, based on historical data from the **VideoLectures.Net** website. Prize fund of 5,500€ is ensured from the European Commission through the **e-LICO EU project**.

**ECML/PKDD 2011 - Discovery challenge Workshop, Athens, Greece, 5th Sep 2011**

Status Closed

Type Scientific

Start 2011-04-18 10:00:00 CET

End 2011-07-08 11:59:59 CET

Prize 5,500€

Registration is required.

## Aims:

- Source of new computational procedures (workflows) for solving recommender problems
- Source of improvements for current VL.Net recommender
- Contribute a new dataset/problems to research community

## Lecture data:

- Content related: title, type, language, publication date, event identifier, authors
- Collaborative data: number of views, lecture co-viewing statistics, pooled viewing sequences related statistics

## Authors data:

- name, e-mail address, homepage address,
- gender, affiliation, and the respective list of lectures

## Taxonomy lecture data

- Events (Conferences, Workshops, Courses...) are grouped into meta-events and form taxonomy
- Categories (Computer science, Text mining, Mathematics...) are grouped in taxonomy

# Challenge task 1

## **Definition of clickstream:**

Clickstream is a sequence of clicks made by a particular user while web browsing, leading from one item of interest to another.

## **Definition of co-viewing frequencies:**

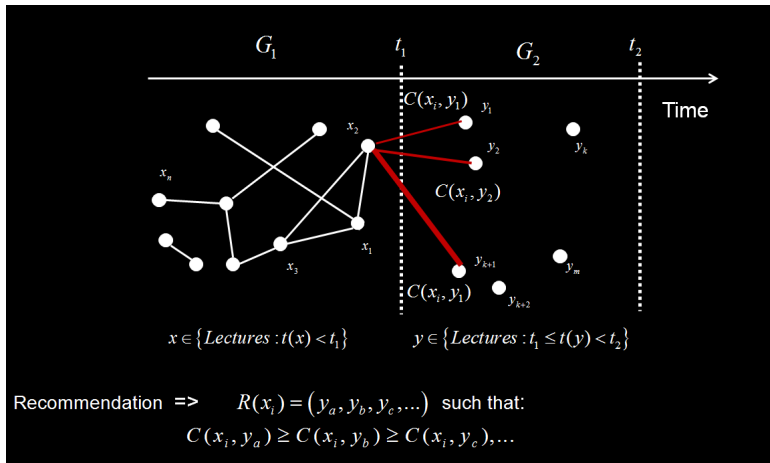
Number of times a pair of lectures has been viewed together in a clickstream is co-viewing frequencies.

## **Challenge task 1:**

Predict ranking of lectures according to withheld lecture co-viewing frequencies in descending order.

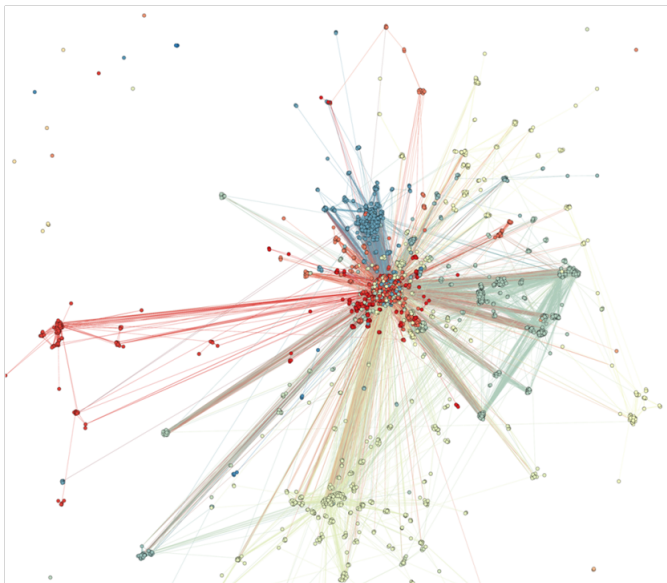


# Challenge task 1



Predict ranking of lectures according to withheld lecture co-viewing frequencies in descending order.

# Data visualization: direct co-viewing sequences



## Evaluation function

The overall score of the submission is mean value over all queries  $R$  (recommended lists  $r$ ) given in the test sets:

$$MARp = \frac{1}{|R|} \sum_{r \in R} AvgRp(r)$$

Average R-precision score -  $AvgRp(r)$  for a single recommended ranked list  $r$  is defined as:

$$AvgRp(r) = \sum_{z \in Z} \frac{Rp@z(r)}{|Z|}$$

where  $Rp@z(r)$  is R-precision at some cut-off length  $z \in Z$  (e.g.  $z \in \{5, 10, 15, 20, 25, 30\}$ ).  $Rp@z(r)$  is defined as the ratio of number of retrieved relevant items and relevant items at the particular cut-off  $z$  of the list:

$$Rp@z(r) = \frac{|relevant \cap retrived|_z}{|relevant|_z} = \frac{|relevant \cap retrived|_z}{\min(m, z)}$$

We will now provide description of baseline recommender systems that were developed prior to the challenge as a baseline solutions.

# Content-based RS

- calculates content distance between all pairs of query lectures and test lectures
- content distance is based on content analysis of textual attributes namely name, description and slide-titles
- we use TF-IDF word vectors
- we provide ranking of lectures according to the ascending value of content distance

# LSI Content-based RS

This recommender system is the same as previous one except it calculates distances in concept space by employing Latent Semantic Indexing or Singular Value Decomposition.

$$A_k = U_k \times S_k \times V_k^T \quad (1)$$

# Graph-based RS

- calculates category or event distance between all pairs of query lectures and test lectures
- we provide ranking of lectures according to the ascending value of category or event distance
- we define the distance between two categories or events as the shortest distance between the nodes that correspond to those categories or events in a graph created from category or event taxonomy.
- category and event taxonomy are a directed graphs, however in order to calculate distances properly, we replaced the directed edges with undirected ones.
- Johnsons algorithm was used to find shortest paths between all pairs of vertices

# Collaborative-based RS

- calculates author distance between all pairs of query lectures and test lectures
- for each author we extract some basic statistics like average view per all his lectures and total view per all his lectures
- for each pair of lectures we calculate distance between corresponding author with respect to some function of average and total view
- provides ranking of lectures according to the ascending value of author distance



# Hybrid combiner RS

This recommender system calculates Manhattan distance between all pairs of query lectures and test lectures in feature space and provides ranking of lectures according to the ascending value of distance. Feature space is spanned by author, event, category and content vectors.

## Temporal hybrid combiner RS

This recommender system is the same as previous one except it feature space incorporates temporal data for lectures. Temporal attribute for each pair of lecture measures the amount of time spent together in the system for all pairs.

# Weighted Temporal hybrid combiner RS

Weighted Manhattan distance between query  $q$  and test  $t$  lecture can be written as:

$$d(q, t, w) = \sum_i w_i * |q_i - t_i| = \sum_i w_i * d_i(q, t)$$

, where vectors  $q$  and  $t$  are elements of feature space  $R^n$ .

Let's denote rank of test lecture  $t$  in recommendation list for some query lecture  $q$  as:  $rank(q, t)$ . We will optimize feature weights  $w_i$  with respect to the cost function as:

$$J(w) = \frac{1}{2m} \sum_{q,t} (d(q, t, w) - \log(rank(q, t)))^2 + \frac{\lambda}{2m} \sum_{j=1}^n w_j^2$$

, where  $m$  is the total number of query-test lectures and  $\lambda$  is the Tikhonov's regularization constant.

## Weighted Temporal hybrid combiner RS

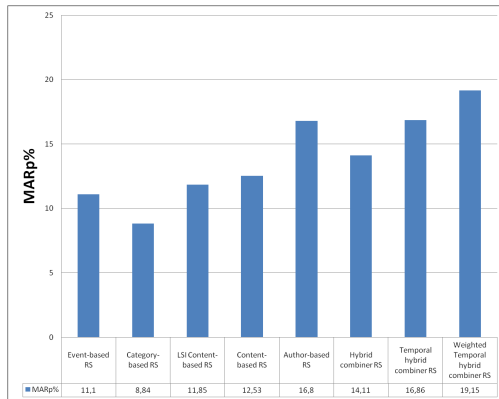
$$J(w) = \frac{1}{2m} \sum_{q,t} (d(q, t, w) - \log(\text{rank}(q, t)))^2 + \frac{\lambda}{2m} \sum_{j=1}^n w_j^2.$$

Gradient vector of the cost function is equal to:

$$\frac{\partial J(w)}{\partial w_k} = \frac{1}{m} \left( \sum_{q,t} (d(q, t, w) - \log(\text{rank}(q, t))) * |q_k - t_k| \right) + \frac{\lambda}{m} w_k.$$

Note that, the scalar value  $d_i(q, t)$  also represents one feature or attribute  $x_i$  of sample  $x$  in pairwise sample space of query-test lectures. Therefore, the Weighted Manhattan distance is just a regularized linear regression model  $\sum_i w_i * x_i$  which approximates logarithmic rank in recommendation list.

# Results



To compare with results from challenge on task 1 see: [http://tunedit.org/challenge/VLNetChallenge/task\\_1?m=leaderboard](http://tunedit.org/challenge/VLNetChallenge/task_1?m=leaderboard)

# Conclusion

- "Learning to rank" is a relatively new very promising research area in machine learning
- Challenge dataset contains rich content descriptions which are publicly available for research purposes  
<http://lis.irb.hr/challenge/index.php/dataset/>
- Winning solution papers from the challenge can be found in proceedings of workshop:  
<http://ceur-ws.org/Vol-770/proceedings.pdf>
- Some of the solutions from the winning submission of the challenge have been built into an improved recommender system of the VideoLectures.Net site: <http://e-lico.videolectures.net>

-  Antulov-Fantulin, N., Bošnjak, M., Šmuc, T., Jermol, M., Žnidaršič, M., Grčar, M., Keše, P., Lavrač, N., *Discovery challenge: "VideoLectures.Net Recommender System Challenge"*, <http://lis.irb.hr/challenge/>
-  Antulov-Fantulin, N., Bošnjak, M., Žnidaršič, M., Grčar, M., Šmuc, T., *ECML/PKDD 2011 Discovery Challenge Overview, In Proc. of ECML-PKDD 2011 Discovery Challenge Workshop, pp 7-20, 2011*
-  Bošnjak M., Antulov-Fantulin N., T. Šmuc, Žnidaršič M., M. Grčar, and Lavrač N.  
Videolectures.net challenge definition.  
Technical report, e-LICO Deliverable D13.2, 2010.
-  Rudjer Boskovic Institute (RBI) Viidea Ltd, Jozef Stefan Institute (JSI).  
Videolectures.net dataset.  
Technical report, e-LICO Deliverable D12.1, 2010.