Visual and Statistical Analysis of VideoLectures.NET

Erik Novak
Artificial Intelligence Laboratory
Jožef Stefan Institute
Jamova cesta 39, Ljubljana, Slovenia
+386 31 272 332
erik.novak@ijs.si

Inna Novalija
Artificial Intelligence Laboratory
Jožef Stefan Institute
Jamova cesta 39, Ljubljana, Slovenia
+386 1 4773144
inna.koval@ijs.si

ABSTRACT

This paper presents learning analytics tools for visual and statistical analysis of data from a portal of video lectures. While learning analytics methods traditionally deal with measurement, collection and analysis of data about learners with an aim of improving the learning process, our solutions are targeted at viewers of VideoLectures.NET. The novel VideoLectures Learning Analytics Dashboard and Lecture Landscape tools allow observing, searching and analyzing viewer behavior and, at the same time, efficiently present the information from the portal to the viewers.

General Terms

Learning Analytics, Measurement, Performance, Design.

Keywords

visualization, analytics, data mining, VideoLectures.NET

1. INTRODUCTION

VideoLectures.NET is a free and open access educational video lectures repository. There are over 20.000 lectures by distinguished scholars and scientists at the most important and prominent events like conferences, summer schools, workshops and science promotional events.

Lectures on the VideoLectures.NET portal are categorized into various categories, such as Arts, Biology, Business, Computer Science, Social Sciences, Technology etc.

In this paper we present tools for visual and statistical analysis of VideoLectures.NET portal data – VideoLectures Learning Analytics Dashboard and Lecture Landscape. While the goal of the VideoLectures Learning Analytics Dashboard is to aggregate, harmonize and analyze event data using various data analysis techniques, the Lecture Landscape visualizes the information about videos, categories and authors in an efficient and user-friendly way.

2. RELATED WORK

Tomas & Cook [14] state that visual analysis should be employed as a means to reveal patterns and trends within data, to develop more intuitive perception and to help with in-depth analysis. Conde et al. [8] present a learning analytics dashboard and its application in real-world case studies. Conceptual framework is developed by Bakharia et al. [7].

In our work we adopt both approaches to learning analytics – visual and statistical analysis, which allows having a deep and more intuitive understanding of the obtained results.

3. LEARNING ANALYTICS FOR VIDEOLECTURES

VideoLectures Learning Analytics Dashboard [1] is a tool developed for the analysis of viewer behavior and detecting which lectures are interesting for the users. We present the results of the preliminary work and the functionalities of the dashboard.

Data Processing. The data considered in the analysis is a set of log files from VideoLectures portal that contain raw events on the portal from September 2012 until December 2015. We have processed 11.3 GB of log files that included the ID, timestamp, session, log entry, lecture and other information such as event type, IP address, location (if present) etc.

With the processed raw log files, we established main event types that appeared in the raw logs: *view* (the user accessed the lecture webpage), *download* (the user downloaded presentation, video etc. from the lecture webpage) and *search* (the user performed a search at the portal).

In addition to the raw log files, we have also collected the log files dedicated to the behavior of particular user while watching particular lectures which we call ranges log files. These present the actions of the user while watching videos like moving forward, moving backwards on the player, skipping some video section etc.

Analysis of Log Files. In order to analyze user behavior at VideoLectures.NET portal, we have utilized and developed a set of data analysis techniques. These were used in the development of VideoLectures Learning Analytics Dashboard, which contains both statistical analysis and visual exploration features.

The analysis has been performed from four perspectives: the aggregated perspective for all lectures, perspective of singular lecture, aggregated perspective of all viewers and perspective of singular viewer.

In addition, we have developed a set of metrics for viewers and lectures that provide us an insight into the user behaviour on the VideoLectures.NET portal.

The ***lecture*** metrics measures the number of views and viewers the lecture has, the average and standard deviation of time the lecture was watched (in minutes and percentage) and the average and standard deviation of moves going forwards and backwards through the lecture video (in minutes and percentage).

The ***viewer*** metrics measures the number of lecture views the viewer made, the average and standard deviation of time spent watching (in minutes and percentage) and the average and standard deviation of moves going forwards and backwards through the lecture video (in minutes and percentage).

First we have analyzed a set of 1000 most popular (by views) videos relevant to the Data Science category. Those videos produced 7055427 views and 3020090 downloads in the period from September 2012 until December 2015. The number of visitors for these videos was 1045860. Moreover, there were 866912 searches at the portal. We have then analyzed all of the log files for which the statistics can be found on the Learning Analytics Dashboard.

3.1 VideoLectures Learning Analytics Dashboard

Our interest is not only to make statistical analysis of the viewer behavior but also to have visual exploration features that enables us to see the traffic and activities made on the VideoLectures portal. For this we created graphs that show the number of views, downloads and searches the viewers make. An example can be seen in Figure 1 which shows the view trends.

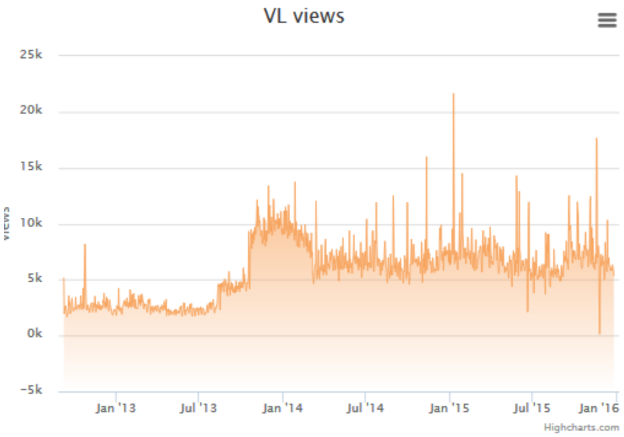
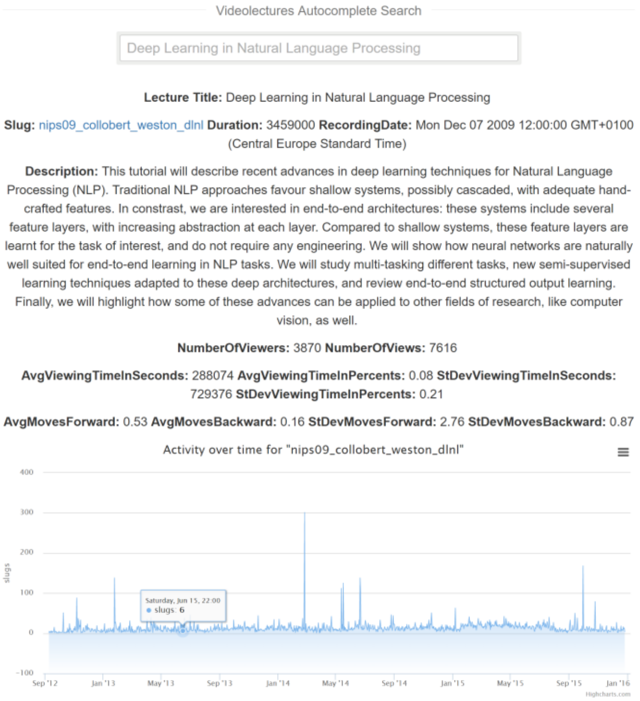


Figure 1. VideoLectures views trend. It shows the total number of lecture views through September 2012 until December 2015.

Statistics for a particular lecture are also available. On the *per lecture* tab one can search for the desired lecture and get its title, description, the measures returned by the lecture metrics and a graph showing the lectures activity. Figure 2 shows the lecture information and measures of “Deep Learning in Natural Langauge Processing” lecture.

The overall statistics of the viewers can also be found under the *all viewers* tab. It shows the distribution of the viewers through countries (see Figure 3) and other statistics measured with the viewer metrics.


Videolectures Autocomplete Search
Deep Learning in Natural Language Processing

Lecture Title: Deep Learning in Natural Language Processing

Slug: nips09_collobert_weston_dlnl **Duration:** 3459000 **RecordingDate:** Mon Dec 07 2009 12:00:00 GMT+0100 (Central Europe Standard Time)

Description: This tutorial will describe recent advances in deep learning techniques for Natural Language Processing (NLP). Traditional NLP approaches favour shallow systems, possibly cascaded, with adequate hand-crafted features. In constrast, we are interested in end-to-end architectures: these systems include several feature layers, with increasing abstraction at each layer. Compared to shallow systems, these feature layers are learnt for the task of interest, and do not require any engineering. We will show how neural networks are naturally well suited for end-to-end learning in NLP tasks. We will study multi-tasking different tasks, new semi-supervised learning techniques adapted to these deep architectures, and review end-to-end structured output learning. Finally, we will highlight how some of these advances can be applied to other fields of research, like computer vision, as well.

NumberOfViewers: 3870 **NumberOfViews:** 7616

AvgViewingTimeInSeconds: 288074 **AvgViewingTimeInPercents:** 0.08 **StDevViewingTimeInSeconds:** 729376 **StDevViewingTimeInPercents:** 0.21

AvgMovesForward: 0.53 **AvgMovesBackward:** 0.16 **StDevMovesForward:** 2.76 **StDevMovesBackward:** 0.87

Figure 2. Lecture information for “Deep Learning in Natural Language Processing” lecture. Displays the basic lecture information, measures and graph of its activity.

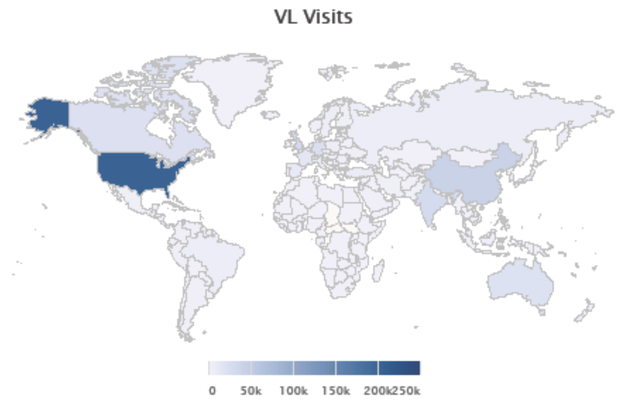


Figure 3. Viewer distribution through countries. Most viewers come from the United States.

In order to implement VideoLectures Learning Analytics Dashboard, we used QMiner [12] for processing data, Node.js [3] for creating the server and dashboard and Highcharts [4] for graphical implementation and dynamics support.

Using the interactive VideoLectures Learning Analytics Dashboard, it is possible to check what are the lectures of broad interest at the VideoLectures.NET portal, how the users of the portal behave through time, where the users come from and how they react at the specific videos.

4. VIDEOLECTURES EXPLORER

Our interest is not only to analyze the viewer behavior but also to see the similarities between lectures. For this we developed VideoLectures Explorer [2], a tool for exploring the lectures published on VideoLectures.NET. The tool enables the user to search through the lectures and find similarities between them, e.g., to find lectures of a specific category or presenter of interest.

Here we explain the method used for visualizing lecture similarities and present the tools functionalities.

4.1 Data Acquisition

The database used for the visualization and basic statistical analysis contains data from all lectures found on VideoLectues.NET. We have acquired data for 23224 lectures, keynotes, interviews, events etc. For each lecture the database contains the lectures title and description, the name of the presenter and his affiliation with city and country, the lectures publication date, video duration, its parent event, number of views and the scientific categories the lecture belongs to. The scientific categories have a hierarchical structure decided by the VideoLectures development team which we use in the visualization process. The database was constructed using the VideoLectures API [5].

4.2 Methodology

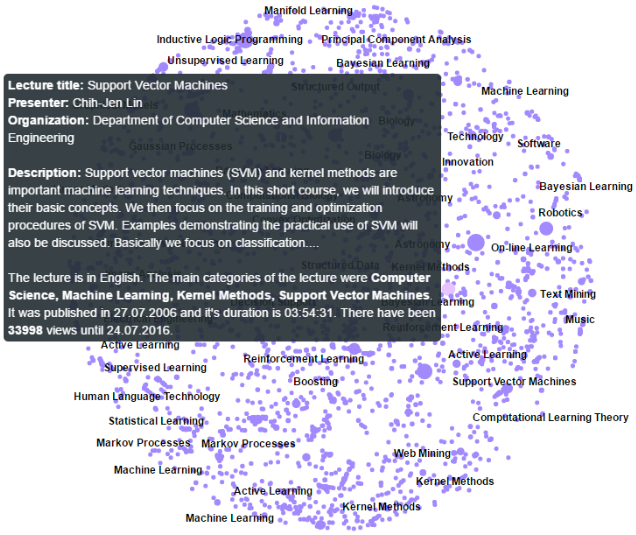
Our objective is to draw the lectures into a two-dimensional vector space to allow plotting on the computer screen, where the similarities of the lectures are maintained. The method we used has been described in [11]. Here is its quick summary:

Using the bag-of-words [13] representation, we represent the lectures as vectors in a high-dimensional vector space, where each dimension corresponds to one word from the vocabulary. These are then used to construct the term-document matrix. Using Latent Semantic Indexing [10] we merge the dimensions associated with the terms that have similar meaning and get the most suitable set of words to describe the corresponding lectures. After that we use Multidimensional Scaling [14] to reduce the dimensionality of the original multidimensional vectors and map them onto two dimensions where the distances between vectors are preserved as well as possible. Once we have the two-dimensional coordinates we can draw the landscape using the preferred visualization library.

The features used for representing the lectures similarities are its title, description, categories and parent event. We used the algorithm described in section 4.2 to calculate the lectures coordinates and draw the landscape using d3.js visualization library [6].

4.3 Explorer Functionality

Each lecture is presented by a point and size mapping to the number of views. Similarity between lectures is mapped to distance between points; more similar lectures are brought closer together. Hovering over a point brings up a tooltip containing information about the lecture: its title and description, the name of the presenters his affiliation, the language in which the lecture was presented, which scientific categories it belongs to, its duration, when it was published and the number of views since the last database update. Any data attribute for which values are not available is omitted from the tooltip. Landmarks show areas populated with lectures that are majorly of the same category. The user can zoom in and out of the landscape to enable a more detailed look of the lectures. An example of the landscape can be seen in Figure 4, which shows the of the *machine learning* lecture landscape.



Lecture title: Support Vector Machines
Presenter: Chih-Jen Lin
Organization: Department of Computer Science and Information Engineering

Description: Support vector machines (SVM) and kernel methods are important machine learning techniques. In this short course, we will introduce their basic concepts. We then focus on the training and optimization procedures of SVM. Examples demonstrating the practical use of SVM will also be discussed. Basically we focus on classification....

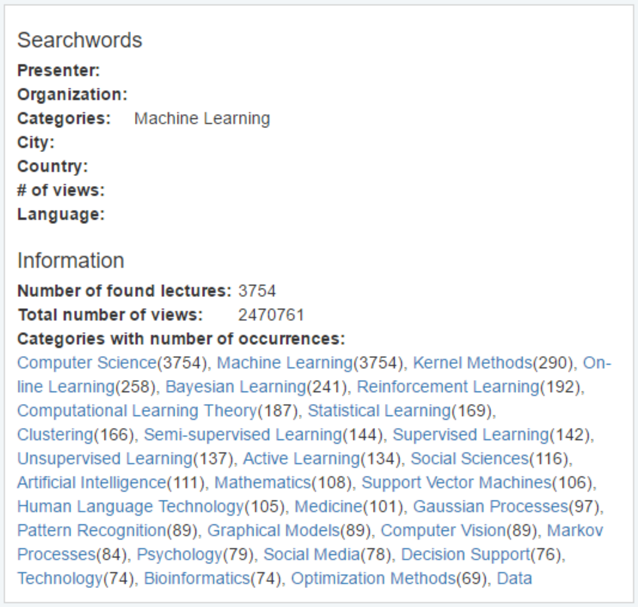
The lecture is in English. The main categories of the lecture were **Computer Science, Machine Learning, Kernel Methods, Support Vector Machines**. It was published in 27.07.2006 and it's duration is 03:54:31. There have been **33998** views until 24.07.2016.

Figure 4. The landscape of machine learning lectures. created using the “Machine Learning” keyword. Lectures that are more similar are brought closer together.

Clicking on the point shows the lectures information. It shows the lectures title and presenter, if the lecture’s public and enabled and a link to the lecture video located on VideoLectures.NET. Clicking on the lecture link opens the corresponding lecture video on the portal.

When the landscape is generated the dashboard also shows an additional information window (see Figure 5). This is in two parts: the first part is the query information, containing the names of presenters, organizations and categories, the minimum and maximum number of views, organization location and the language the user used to query the data.

The second part contains the basic statistics about the queried data, the number of lectures in the queried data, the total number of lecture views and scientific categories with the number of their occurrences in the queried data. Clicking on the category in the dashboard automatically queries the data using the category name, and the information window is updated along the landscape view.



Searchwords

Presenter:
Organization:
Categories: Machine Learning
City:
Country:
# of views:
Language:

Information

Number of found lectures: 3754
Total number of views: 2470761
Categories with number of occurrences:

Computer Science(3754), Machine Learning(3754), Kernel Methods(290), On-line Learning(258), Bayesian Learning(241), Reinforcement Learning(192), Computational Learning Theory(187), Statistical Learning(169), Clustering(166), Semi-supervised Learning(144), Supervised Learning(142), Unsupervised Learning(137), Active Learning(134), Social Sciences(116), Artificial Intelligence(111), Mathematics(108), Support Vector Machines(106), Human Language Technology(105), Medicine(101), Gaussian Processes(97), Pattern Recognition(89), Graphical Models(89), Computer Vision(89), Markov Processes(84), Psychology(79), Social Media(78), Decision Support(76), Technology(74), Bioinformatics(74), Optimization Methods(69), Data

Figure 5. The additional information window created using the "Machine Learning" keyword. It contains the query keywords and the overall statistics about the queried lectures.

5. CONCLUSION AND FUTURE WORK

In this paper we presented learning analytics tools for visual and statistical analysis of data from VideoLectures.NET portal. While learning analytics methods traditionally deal with measurement, collection and analysis of data about learners with an aim of improving the learning process, our solutions are targeted at viewers of VideoLectures. The novel VideoLectures Learning Analytics Dashboard and Explorer tools allow observing, searching and analyzing viewer behavior and, at the same time, efficiently present the information from the portal to the viewers.

The future work will include the development of more efficient learning analytics suggestions for VideoLectures.NET portal and providing the recommendations on how to increase the viewer engagement into the portal.

6. ACKNOWLEDGMENTS

This work was supported by the Slovenian Research Agency and EDSA EU H2020 project (Contract no: H2020-ICT-643937).

7. REFERENCES

[1] Videolectures learning analytics | dashboard, http://learninganalytics.videolectures.net, accessed: 2016-09-09.

[2] Videolectures learning analytics | landscape, http://explore.videolectures.net, accessed: 2016-09-09.

[3] Node.js, https://nodejs.org/en/, accessed: 2016-09-09.

[4] Interactive javascript chart for your website | highcharts, http://www.highcharts.com/, accessed: 2016-09-09.

[5] Swagger ui, http://videolectures.net/site/api/docs/, accessed: 2016-09-09.

[6] D3.js – data-driven documents, http://d3js.org/, accessed: 2016-09-09.

[7] A. Bakharia, L. Corrin, Linda, P. de Barba, G. Kennedy, D. Gašević, R. Mulder, D/ Williams, S. Dawson and L. Lockyer. A conceptual framework linking learning design with learning analytics. *In Proc., Sixth International Conference on Learning Analytics & Knowledge*, ACM, pages 329-338, 2016.

[8] M. Á Conde, F. J. García-Peñalvo, D. A. Gómez-Aguilar and R. Therón. Exploring Software Engineering Subjects by Using Visual Learning Analytics Techniques, *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje*, 10(4), pages 242-252, 2015.

[9] T. F. Cox and M. A. Cox. *Multidimensional Scaling*. CRC press, New York, 2000.

[10] S. T. Dumais. Latent semantic analysis. *Annual review of information science and technology*, 38(1):188-230, January 2004.

[11] B. Fortuna, D. Mladenić and M. Grobelnik. Visualization of temporal semantic spaces. In *Semantic knowledge management*, pages 155-169, Springer, 2009.

[12] B. Fortuna, J. Rupnik, C. Fortuna, M. Grobelnik, V. Jovanoski, M. Karlovcec, B. Kazic, K. Kenda, G. Leban, J. Novljan, M. Papler, L. Rei, B. Sovdat, L. Stopar and A. Muhic. QMiner – Data analytics platform for processing streams of structured and unstructured data. *Software Engineering for Machine Learning Workshop, Neural Information Processing Systems*, 2014.

[13] G. Salton. Developments in automatic text retrieval. *Science*, 253(5023):974-979, August 1991.

[14] J.J. Thomas and K.A.Cook. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*, IEEE CS Press, Los Alamitos, 2005.