A Dynamic Topic Model of Learning Analytics Research

Michael Derntl
RWTH Aachen University
Advanced Community Information Systems (ACIS)
Aachen, Germany
derntl@dbis.rwth-aachen.de

Nikou Günnemann
RWTH Aachen University
Advanced Community Information Systems (ACIS)
Aachen, Germany
nikou@dbis.rwth-aachen.de

Ralf Klamma
RWTH Aachen University
Advanced Community Information Systems (ACIS)
Aachen, Germany
klamma@dbis.rwth-aachen.de

ABSTRACT

Research on learning analytics and educational data mining has been published since the first conference on Educational Data Mining (EDM) in 2008 and gained momentum through the establishment of the Learning Analytics and Knowledge (LAK) conference in 2011. This paper addresses the LAK Data Challenge from the perspective of visual analytics of topic dynamics in the LAK Dataset between 2008 and 2012. The data set was processed using probabilistic, dynamic topic mining algorithms. To enable exploration and visual analysis of the resulting topic model by LAK researchers and stakeholders we developed and deployed D-VITA, a web-based browsing tool for dynamic topic models. In this paper we explore answers to the questions about past, present, and future of LAK posed in the Data Challenge based on a topic model of all papers in the LAK Dataset. We also briefly describe how users can explore the LAK topic model on their own using D-VITA.

1. OBJECTIVES

The LAK Data Challenge called for contributions to make sense of the field of learning analytics including its "roots, current state, and future trends, based on how its members report and debate their research"[1]. This paper tackles the challenge by presenting facts obtained from statistical analyses of the paper full texts included in the provided LAK Dataset [7]. The main contributions are as follows:

1. A dynamic topic model was computed using the approach presented in [3]. Using this dynamic topic model we explore in Section 4 three questions about the evolution of topics in the LAK Dataset to distill knowledge about past, present and future of LAK research.
2. In Section 5 we describe the visual analytics application D-VITA[2], which puts the toolkit to answer the questions posed in the LAK Data Challenge into the user's hands. D-VITA is a web-based tool that offers topic-based views on the LAK Dataset using a point-and-click metaphor and simple visualizations.

[1] http://solaresearch.org/events/lak/lak-data-challenge/
[2] http://monet.informatik.rwth-aachen.de/DVita/?id=16

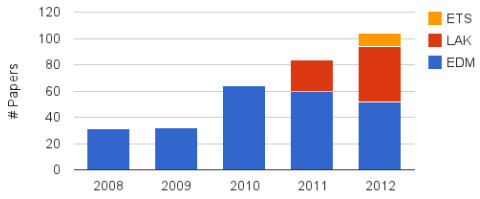


Figure 1: Yearly distribution of papers over venues

2. DATASET AND PREPROCESSING

The LAK Dataset underlying the analyses presented in this paper includes the EDM conference proceedings 2008–2012 (239 papers), the LAK conference proceedings 2011–2012 (66 papers), and the papers of the 2012 Special Issue on Learning Analytics in the Educational Technology and Society journal (10 papers; herafter referred to as ETS). The RDF representation of the LAK Dataset was processed by a script that extracted for each paper the identifier, venue $\in$ {LAK, EDM, ETS}, year of publication, title, authors, abstract, full text, and hyperlink to the full RDF description on `data.linkededucation.org`. The distribution of the 315 papers over time and venues is given in Figure 1.

In the next preprocessing step the paper records were cleaned by removing stopwords and by applying stemming methods on the included word sets. For word stemming we used the Porter Stemming technique [6], which is well established for this purpuse. As a result, close to 5000 distinct word stems were identified as being used in the 315 papers.

3. DYNAMIC TOPIC MINING

From a text mining perspective the LAK Dataset represents a text corpus in which a set of words is used in a set of papers. To identify what is relevant to LAK research, we used the dynamic topic modeling approach described in [3] to obtain the distribution of words over a pre-defined number of topics. This is a probabilistic, unsupervised machine learning approach that has been gaining increasing prominence recently [2]. In these probabilistic topic models a topic is a distribution of words, so each topic is typically represented

by its most frequently occurring words. Topic mining also obtains the distribution of these topics over the papers in the data set. *Dynamic* topic mining applies these analysis steps using several consecutive time slices in the data set. For the LAK Dataset, we chose the five calendar years $\in \{2008 \ldots 2012\}$ as time slices. The results will thus reveal the evolution of topics over documents during those discrete time slices, and the evolution of words used in the papers for each topic over time.

Dynamic topic mining requires the analyst to pre-set the number of topics. Based on previous experiments with varying numbers of topics in paper collections in well-defined subject areas, we decided to run the analysis of the LAK Dataset with a set of 20 topics. This number, while somewhat arbitrary, shall provide for sufficient discriminatory power for both the distribution of topics over papers and the distribution of words over topics. With fewer topics, terms like 'learning', for instance, are more likely to be present with relatively high relevance in many topics, while a larger preset would increase the number of topics exposed in each paper. Both situations would impede reasonable interpretation and visualization of the results.

A word of explanation regarding the labels used to refer to topics in this paper: mathematically each topic is a distribution over words. In a dynamic topic model this distribution changes over time, i.e. a specific word may rise or fall in relevance for a topic. In the rest of the paper we will therefore label each topic with an ordered tuple representing those words with the highest mean relevance for this topic over time. In topic modeling literature we found that four words is a good number to form a topic label. For instance, for topic "`students model parameters skill`" the most relevant word *on average* is student followed by model, parameters, and skill. For illustration, based on the word distribution for this topic in 2008 only, the label would be "`model student skill learning`". Often, such word tuples are rephrased as more expressive labels; for instance "student modeling" could be appropriate in our example.

The obtained topic model including 20 topics was analyzed to see whether the topics have sufficient discriminatory power. To this end, we used the ten most important words for each topic and the corresponding probability distributions to compute a dissimilarity measure of the distributions by using the Jensen-Shannon divergence measure [5]. The matrix with pairwise divergence values is displayed in Figure 2. The maximum Jensen-Shannon divergence value is $ln(2) \approx .69$. The darker the cell color, the lower the divergence, thus the higher the similarity. The matrix is generally "light-colored", indicating that the topics' word distributions diverge to a high degree. Topic pair (A, S) has the lowest dissimilarity value, and Figure 3 reveals why: both topics are about student modeling. Topic S generally appears to have several loosely related topics.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	
A	.00	.38	.49	.41	.34	.56	.49	.39	.32	.38	.54	.44	.37	.53	.50	.49	.44	.69	.22	.63	A
B	.38	.00	.50	.49	.36	.58	.56	.47	.47	.54	.63	.55	.46	.59	.57	.52	.54	.69	.36	.69	B
C	.49	.50	.00	.54	.46	.58	.48	.40	.56	.55	.64	.55	.50	.53	.55	.47	.60	.69	.48	.69	C
D	.41	.49	.54	.00	.49	.59	.50	.52	.57	.59	.64	.63	.51	.62	.58	.56	.59	.69	.40	.64	D
E	.34	.36	.46	.49	.00	.52	.53	.35	.43	.48	.53	.42	.38	.53	.48	.47	.47	.69	.40	.63	E
F	.56	.58	.58	.59	.52	.00	.63	.52	.62	.61	.37	.61	.52	.69	.39	.62	.49	.61	.60	.62	F
G	.49	.56	.48	.50	.53	.63	.00	.49	.59	.63	.66	.47	.55	.57	.59	.53	.62	.69	.54	.57	G
H	.39	.47	.40	.52	.35	.52	.49	.00	.47	.48	.55	.45	.40	.47	.42	.41	.52	.69	.44	.64	H
I	.32	.47	.56	.57	.43	.62	.59	.47	.00	.43	.59	.46	.45	.55	.55	.55	.47	.69	.29	.63	I
J	.38	.54	.55	.59	.48	.61	.63	.48	.43	.00	.56	.48	.48	.56	.51	.61	.49	.69	.46	.62	J
K	.54	.63	.64	.64	.53	.37	.66	.55	.59	.56	.00	.52	.51	.65	.36	.65	.51	.59	.64	.57	K
L	.44	.55	.55	.63	.42	.61	.47	.45	.46	.48	.52	.00	.50	.49	.53	.54	.53	.65	.52	.62	L
M	.37	.46	.50	.51	.38	.52	.55	.40	.45	.48	.51	.50	.00	.56	.43	.51	.48	.69	.44	.62	M
N	.53	.59	.53	.62	.53	.69	.57	.47	.55	.56	.65	.49	.56	.00	.61	.51	.59	.47	.55	.69	N
O	.50	.57	.55	.58	.48	.39	.59	.42	.55	.51	.36	.53	.43	.61	.00	.56	.38	.63	.56	.61	O
P	.49	.52	.47	.56	.47	.62	.53	.41	.55	.61	.65	.54	.51	.51	.56	.00	.59	.69	.50	.69	P
Q	.44	.54	.60	.59	.47	.49	.62	.52	.47	.49	.51	.53	.48	.59	.38	.59	.00	.62	.52	.57	Q
R	.69	.69	.69	.69	.61	.69	.69	.69	.69	.69	.59	.65	.69	.47	.63	.69	.62	.00	.69	.69	R
S	.22	.36	.48	.40	.40	.60	.54	.44	.29	.46	.64	.52	.44	.55	.56	.50	.52	.69	.00	.69	S
T	.63	.69	.69	.64	.63	.62	.57	.64	.63	.62	.57	.62	.62	.69	.61	.69	.57	.69	.69	.00	T
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	

Figure 2: Overview of topic divergence

4. ANALYSIS OF LAK TOPIC DYNAMICS

In this section we explore three questions about the LAK Dataset, intending to shed some light on the past and present topics of learning analytics research, along with a cautious glimpse into the future.

Question 1: What have been the most relevant topics overall in the LAK data set?

This question addresses the LAK Data Challenge aspects of roots and current state of learning analytics. Figure 3 shows an overview chart of the 20 topics identified in the LAK Dataset. The horizontal axis reflects the rank of mean relevance of each topic and the vertical axis reflects the rank of stability[3] over the five time slices in the dataset. The size of each bubble reflects the relevance of the topic in 2012, the most recent period. We make several observations:

- The most relevant topics most prominently feature the terms students/learners, model, and data. This aligns well with SoLAR's definition of learning analytics as "the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs,"[1] considering that understanding and optimization is necessarily based on models of learners and data.

- The topic with the highest mean relevance is "`student model parameters skill`" (A); this topic also has the highest variance in relevance.

- In the top-right quadrant we find topic "`model data features prediction`" (B) which has a strong relevance in 2012, high mean relevance rank over all years and a high stability. As such, it can be considered as one of the core topics in the LAK Dataset. In 2012 the distribution of words in this topic would advocate the label "`prediction model data students`", i.e. prediction is currently most relevant for this topic.

- Topic "`network community discussion analysis`" (R) is also worth looking at. While it is relatively irrelevant and volatile, it is among the relevant topics in 2012 (cf. the bubble size). The topic evolution chart in Figure 4 reveals that this topic accumulated most of its relevance in 2011, the year of the first LAK conference. Also, 8 of the 10 papers with the strongest focus on this topic in 2011 were published in the LAK conference (see bottom portion of Figure 4) although EDM published 2.5 times the number of papers in that

[3]Stability was computed by inverting the variance of the topic's relevance over time

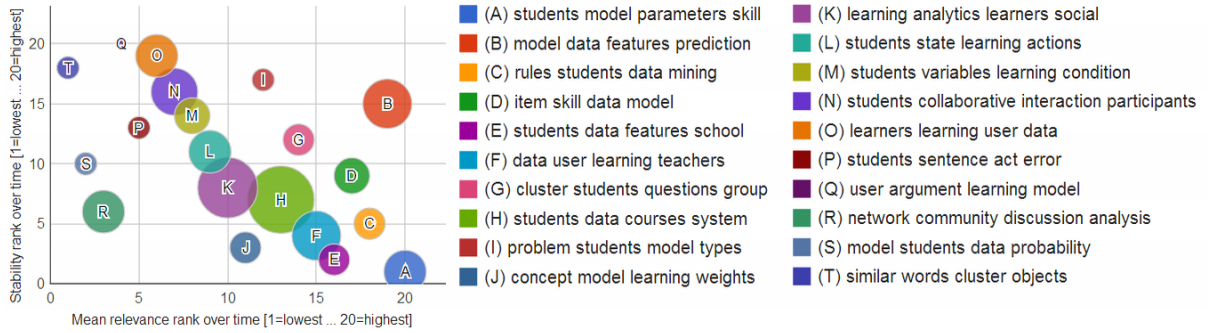
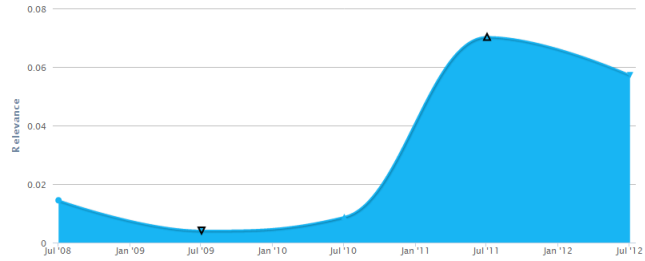


Figure 3: Topic stability plotted against average topic relevance over time.



#.	Title
1	[LAK] A Unified Framework for Multi-Level Analysis of Distributed Learning
2	[LAK] Social and Semantic Network Analysis of Chat Logs
3	[LAK] Learning Analytics To Identify Exploratory Dialogue within Synchronous Text Chat
4	[LAK] Discourse-Centric Learning Analytics
5	[EDM] Analyzing Participation of Students in Online Courses Using Social Network Analysis Techniques
6	[EDM] Combining study of complex network and text mining analysis to understand growth mechanism of communities on SNS
7	[LAK] Generating Predictive Models of Learner Community Dynamics
8	[LAK] SNAPP: A Bird's-Eye View of Temporal Participant Interaction
9	[LAK] iSpot Analysed: Participatory Learning and Reputation
10	[LAK] The Value of Learning Analytics to Networked Learning on a Personal Learning Environment

Figure 4: Evolution (top) and most representative papers in 2011 (bottom) of topic "network community discussion analysis"

year. This topic, in 2012 represented by the word order "`network community social user`", therefore appears to be a genuine LAK topic which was previously rather irrelevant for the EDM conference.

Question 2: What changes in topic dynamics did the first LAK conference in 2011 bring about?

This question aims to reveal whether and how the LAK community relates to the EDM community in terms of topics covered by their papers. To explore this we look (a) at at the overall distribution of topics over time and (b) at the relative change of topic relevance between 2010 and 2011.

The evolution of the overall distribution of topics is illustrated in the ThemeRiver in Figure 5. In a ThemeRiver [4] the horizontal axis represents the points in time to which the documents in a dataset belong (in the LAK Dataset that is the publication date), and the vertical axis represents the relevance of the topic. Each current in the ThemeRiver therefore presents the dynamic development of a selected topic over time. The wider the current, the more relevant is the topic, i.e. the more documents expose this topic. Since each document exposes different topics to varying degrees the relevance of topic k at time t is formally defined as $relevance(k,t) := \frac{1}{|D_t|}\sum_{d\in D_t}\theta_d[k]$, where D_t is the set of documents belonging to time t, and θ_d is the topic distribution for document d. Observing the ThemeRiver in Figure 5 it is evident that there were some shifts in topic focus during the years 2008 and 2010, where we have only the EDM papers in the dataset. Between 2010 and 2011 we identify the strongest turbulence, presumably based on substantial shifts in topic foci introduced by the 2011 LAK conference. Interestingly the topic distribution remains rather stable during the last time slice, in which LAK 2012, EDM 2012 and the ETS special issue are included. This might suggest that these three publication venues propelled the convergence of LAK research as represented in the LAK Dataset.

To see which topics rose in relevance between 2010 and 2011 we filter for topics and zoom into the transition between 2010 and 2011 as illustrated in Figure 6. Those three topics that have their absolute highest relevance in 2011 are marked with an up-pointing triangle with a solid-black outline. These are "`model students data probability`", "`network community discussion analysis`", and "`problem students model types`", indicating an increased focus on student modeling as well as community and network analysis through the first LAK conference in 2011.

Question 3: What topics rose the most in 2012, the most recent time slice in the data set?

This question looks into what the dynamic topic model of the LAK Dataset suggests as rising topics over the next year(s). We try to answer this by identifying those five topics that had the highest rise in relevance between 2011 and 2012. The topic labels represent the word distribution in 2012, and the number in parentheses indicates the absolute gain in relevance:

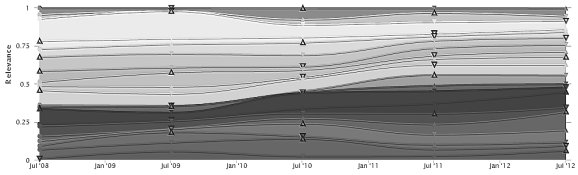


Figure 5: Overall distribution of topic relevance between 2008 and 2012

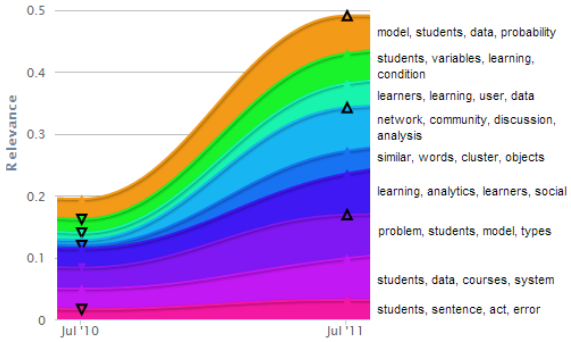


Figure 6: Topics with rising relevance in 2011

1. `students data courses system` (+.054)
2. `students interaction participants analysis` (+.036)
3. `learning analytics social learners` (+.035)
4. `students actions learning state` (+.025)
5. `data user learning dataset` (+.013)

In sum these five topics have accumulated a share of 42% of the topic distribution by 2012, starting from 11% in 2008 (cf. Figure 7). These developments indicate a strong increase in focus on the students' activities and actions in courses as well as social and interaction analytics.

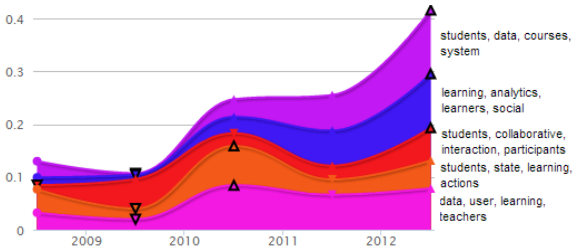


Figure 7: Cumulative relevance of the top-five rising topics 2012 over all years

5. D-VITA TOPIC ANALYTICS TOOKIT

Except for Figures 1 and 3 all figures were produced using D-VITA, a web-based visual analytics tool we developed and deployed for visual analytics of dynamic topic models. The tool allows users to visually interact with the output of the dynamic topic mining algorithms on the LAK Dataset. The application window shown in Figure 8 has three panels:

The *Topics Panel* shows the list of topics obtained by the dynamic topic modeling algorithm; topics can be sorted by rising, falling and mean relevance, as well as variance of relevance. The topics can be filtered using keywords; in the screen shot the keyword "visual" is used as a filter. The topic list thus only includes topics whose set of relevant words includes this word stem. Topics checked by the user will be visualized in the ThemeRiver in the Topic Evolution Panel.

The *Topic Evolution Panel* shows a ThemeRiver of evolution of relevance of the topics selected in the Topics Panel. Data points for each topic and time slice, respectively, can be clicked, which will trigger the display of detailed information on the clicked topic at the selected time slice in the Document and Word Evolution Panel.

The *Document and Word Evolution Panel* shows for the selected topic an ordered list of the most relevant papers in the "Relevant Documents" tab. The icons next to each document allow showing the topic pie for the document and its content, respectively. The "Similar Docs" icon will bring up the Document Browser with a list of similar documents. Under the "Word Evolution" tab the user will find a ThemeRiver illustrating the evolution of the distribution of words in the selected topic over time.

D-VITA also offers a *Document Browser* to perform keyword-based search, explore the topic distribution of documents, and navigate documents based on similarity.

6. CONCLUSION

In a nutshell, we discovered the following: Regarding the *past*, we found that LAK and EDM do have a substantial shared topic foundation including themes like student modeling, data classification, and clustering. We also found that the EDM conference series had some turbulence in topical focus between 2008 and 2010, the time window when only EDM papers are present in the dataset.

Regarding the *present* we found that the LAK Dataset exposes a strong emphasis on learner modeling, data modeling, analysis and prediction. The first LAK conference in 2011 also brought some considerable shifts in topic focus; e.g. LAK 2011 has visibly strengthened network and social analysis aspects on top of EDM topics.

Regarding the near *future* we found that the shifts in the topics' proportions in 2012 appear rather moderate, thus indicating a phase of convergence of LAK research topics. Projecting recent topic shifts into the future, we can expect increased emphasis on social and interaction aspects and a sustained, strong role of students as research subjects.

7. ACKNOWLEDGMENTS

This work was supported by the European Commission through the the support action *TEL-Map* (FP7-257822) and the integrated project *Layers* (FP7-318209).

8. REFERENCES

[1] About SoLAR, 2012. `http://www.solaresearch.org/mission/about/`.

[2] D. M. Blei. Probabilistic topic models. *Commun. ACM*, 55(4):77–84, 2012.

[3] D. M. Blei and J. D. Lafferty. Dynamic topic models. In *ICML*, pages 113–120, 2006.

[4] S. Havre, E. G. Hetzler, P. Whitney, and L. T. Nowell. Themeriver: Visualizing thematic changes in large document collections. *IEEE Trans. Vis. Comput. Graph.*, 8(1):9–20, 2002.

[5] J. Lin. Divergence Measures Based on the Shannon Entropy. *IEEE Transactions on Information Theory*, 37(1), 1991.

[6] M. F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[7] D. Taibi and S. Dietze. Fostering analytics on learning analytics research: the LAK dataset, Technical Report, 03/2013, 2013. `http://resources.linkededucation.org/2013/03/lak-dataset-taibi.pdf`.

Figure 8: Application window showing ThemeRiver and document list (rotated image)