Fostering Analytics on Learning Analytics Research: the LAK Dataset

Davide Taibi
Institute for Educational Technologies
National Research Council of Italy
Via Ugo La Malfa 153, Palermo
Italy
davide.taibi@itd.cnr.it

Stefan Dietze
L3S Research Center
Appelstr. 9a
30167 Hannover
Germany
dietze@l3s.de

ABSTRACT
This paper describes the Learning Analytics and Knowledge (LAK) Dataset, an unprecedented collection of structured data created from a set of key research publications in the emerging field of learning analytics. The unstructured publications have been processed and exposed in a variety of formats, most notably according to Linked Data principles, in order to provide simplified access for researchers and practitioners. The aim of this dataset is to provide the opportunity to conduct investigations, for instance, about the evolution of the research field over time, correlations with other disciplines or to provide compelling applications which take advantage of the dataset in an innovative manner. In this paper, we describe the dataset, the design choices and rationale and provide an outlook on future investigations.

Categories and Subject Descriptors
H.3.4 [Semantic Web], I.2.4 [Ontologies]

General Terms
Algorithms, Documentation, Design, Experimentation, Standardization.

Keywords
Learning Analytics, Data, Linked Data, Semantic Web, Educational Data Mining

1. INTRODUCTION
As part of an international team of research practitioners consisting of the Society for Learning Analytics Research (SoLAR)¹, ACM², the LinkedUp project³, the Educational Technology Institute of the National Research Council of Italy (CNR-ITD), we have released an unprecedented resource for the Learning Analytics and Educational Data Mining. In order to allow the computational analyses of research publications in the emerging Learning Analytics field, we have published in a machine-readable format a comprehensive set of scientific papers in the field of learning analytics and educational data mining. This can serve as input for studies into scientometrics, investigations into the evolution of the overall discipline or correlations with other fields.

The LAK dataset provides access to documents already available online in unstructured form but also to research works not publicly accessible before at all. The collection includes the proceedings of the International Educational Data Mining Society⁴, as well as the Journal of Educational Technology and Society a special issue on Learning Analytics⁵. In both cases full text has been freely available already but in unstructured format. Content which has previously been accessible to subscribers only includes the “Proceedings of the ACM International Conference on Learning Analytics and Knowledge” edited by ACM. In the framework of our initiative, ACM is providing freely its ACM Digital Library in the Learning Analytics field solely for research purposes.

The following table describes in details the set of papers that have been collected, processed and exposed in a structured form in the Learning Analytics and Knowledge (LAK) dataset:

<table>
<thead>
<tr>
<th>Publication</th>
<th># of papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proceedings of the ACM International Conference on Learning Analytics and Knowledge (LAK) (2011-12)</td>
<td>66</td>
</tr>
<tr>
<td>The open access journal Educational Technology &amp; Society special issue on “Learning and Knowledge Analytics”: Educational Technology &amp; Society (Special Issue on Learning &amp; Knowledge Analytics, edited by George Siemens &amp; Dragan Gašević), 2012, 15, (3), pp. 1-163.</td>
<td>10</td>
</tr>
<tr>
<td>Proceedings of the International Conference on Educational Data Mining (2008-12)</td>
<td>239</td>
</tr>
<tr>
<td>Journal of Educational Data Mining (2008-12)</td>
<td>16</td>
</tr>
</tbody>
</table>

¹ www.solaresearch.org/
² http://acm.org/
³ http://linkedup-project.eu/
⁴ http://www.educationaldatamining.org/proceedings
⁵ http://www.ifets.info/issues.php?id=56
2. FROM SCHOLARLY PAPER TO STRUCTURED DATA

2.1 The extraction process

In order to process and analyze the set of unstructured journals and conferences papers data was transformed into structured data. While each conference proceeding is available on the Web in PDF format, but each collection has its own structure. Even if in some cases the most used format is the ACM template\(^6\), papers not always comply with it entirely, calling for some specifically adapted extraction mechanisms. The overall knowledge extraction process is composed of three main steps:

1. Transforming PDF documents to plain textual representation.
2. Cleaning up and consolidation of the textual information.
3. Extracting structured data from text.

In the first step the PDF file containing the proceeding of a conferences, or the papers of a journal is split up in order to have one document for each paper. Then each PDF file has been elaborated with pdf2text tool in order to have a textual representation for each paper.

In the second step the text files are elaborated in order to transform them in a partially structured format that can be elaborated automatically. In particular at this step tables and figures are removed from the paper, maintaining their captions, that can be useful for text mining processing, footnotes have been also removed from the text, while bulleted or numbered list have been organized using an homogeneous format.

As part of the third step, text files are being processed in order to extract from them the most important sections of the document. Regarding the authors, their name, affiliation, country are represented using the FOAF ontology.

For each paper the following information are collected: title, authors, keywords, abstract, full-text and its relationship with the type of publication or event, (journal or conference proceedings).

It is important to note that beside the common metadata for the learning analytics papers such as: title, abstract, authors and affiliation also the full text of the papers is stored in the dataset. At this stage the full text has been stored without considering its separation in paragraphs and sections, however the elaboration performed at step number 2 has also identified the titles and paragraphs of sections and subsections, thus providing the basis for analyzing full text with further granularity in next versions of the LAK dataset. The referenced papers are also extracted but are not made available in the LAK dataset in this version of the dataset.

2.2 The schema

The schema used to describe the papers in the dataset is based on two established schemas: the Semantic Web Conference (SWC) ontology\(^8\) (already used to describe metadata about publications from the Semantic Web conferences and related events)\(^6\) and the Linked Education schema\(^9\). The Linked Education schema has been developed to represent and catalog both educational and educational related datasets, which are datasets not specifically created for education but that can be used in an educational context. The schema has been used to annotate datasets and resources as part of an integrated dataset\(^10\) which contains educationally relevant resources such as : LinkedUniversities\(^1\) and the meEducator Educational Resources \(^4\) with their Open Educational Resources and materials explicitly related to education, as well as implicitly educationally relevant datasets such as BBC Programmes\(^3\), ACM Library Metadata\(^7\) and Europeana \(^2\) datasets. The main entities collected in the LAK dataset are paper authors, institutions and papers, related to the

\(^6\) http://www.acm.org/sigs/publications/proceedings-templates
\(^7\) http://data.semanticweb.org/ns/swc/ontology
\(^8\) http://data.semanticweb.org/
\(^9\) http://data.linkededucation.org/ns/linked-education.rdf
\(^10\)http://linkdup.l3s.uni-hannover.de:8880/opendf-sesame/repositories/linked-learning-selection?query
\(^11\) http://acm.rkbexplorer.com/
learning analytics area. Authors and institutions have been represented using respectively the classes Person and Organization of the FOAF ontology, while to represent papers, the class InProceedings of the SWRC ontology has been used. The LAK Dataset, at the time of writing, includes 779 authors, connected to 295 institution, and 315 posters, abstract, short and full papers.

2.3 Access Methods

In order to support different access method for the data, the resources of the LAK Dataset have been published in different formats:

- A dump file in zipped RDF/XML file format can be directly downloaded from the SoLaR research web page.
- A version of the dataset in a format that can be elaborated through the R statistic software have been provided
- A Linked Data endpoint with a public SPARQL endpoint has been developed in order to provide access to structured RDF metadata according to LOD principles.

The following SPARQL query on the LAK dataset can be used to extract the full text of all 2011 papers (LAK 2011, and EDM 2011 conferences) in .srx format (XML file which can be opened in any text editor):

```sparql
PREFIX led:<http://data.linkededucation.org/ns/linked-education#>
PREFIX swrc:<http://swrc.ontoware.org/ontology#>

```

On the SoLaR Website some useful examples for querying the SPARQL endpoint of the LAK dataset have been reported. The example queries allow users, for instance, to retrieve: the papers co-authored by two selected authors; all papers published in both EDM and LAK conferences by the authors affiliated to an institution.

3. LAK DATA CHALLENGE

Beyond merely publishing the data, we are actively encouraging its innovative use and exploitation as part of a public LAK Data Challenge sponsored by the European Project LinkedUp. An initial competition is co-located with the ACM LAK13 Conference, Leuven, Belgium (April 2013). The challenge is revolving around the overall question on what insights can be gained from analytics on the LAK corpus about the general discipline of Learning Analytics and its connection to other fields. How can we make sense of this emerging field’s historical roots, current state, and future trends, based on how its members report and debate their research? Challenge submissions should exploit the LAK dataset by covering one or more of the following, non-exclusive list of topics:

- Analysis & assessment of the emerging LAK community in terms of topics, people, citations or connections with other fields
- Innovative applications to explore, navigate and visualise the dataset (and/or its correlation with other datasets)
- Usage of the dataset as part of recommender systems

4. CONCLUSIONS

The LAK Dataset has been a first starting point to allow the analysis of learning analytics as an emerging research discipline and its definition and evolution. Along with the growth of the research works in learning analytics and related fields, we intend to expand the dataset by adding new research publications. In addition, while the dataset currently contains plain metadata and full text of the research publications, it is envisaged to extract and add additional data about contained entities and topics, to provide simple means for assessing, exploring and navigating the data.

5. ACKNOWLEDGMENTS

This work is partly funded by the European Union under FP7 Grant Agreement No 317620 (LinkedUp).

6. REFERENCES


18 http://lakconference2013.wordpress.com/
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 dataset 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.” 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.

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; hereafter referred to as ETS). The RDF representation of the LAK Dataset was processed by a script that extracted for each paper the identifier, venue ∈ {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 purpose. 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 ∈ (2008...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 for each 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 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 2 reveals why: both topics are about student modeling. Topic S generally appears to have several loosely related topics.

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 2 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 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”, considering that understanding and optimization is necessarily based on models of learners and data.

- The topic with the highest mean relevance is “students 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 year.
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 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, 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 $\text{relevance}(k, t) := \frac{1}{|D_t|} \sum_{d \in D_t} \theta_d[k]$, where $D_t$ is the set of documents belonging to time $t$, and $\theta_d$ 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](image-url)
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.

5. D-VITA TOPIC ANALYTICS TOOLKIT

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

Socio-semantic Networks of Research Publications in the Learning Analytics Community
Soude Fazeli, Hendrik Drachsler, Peter Sloep

Open University of the Netherlands (OUNL)
Centre for Learning Sciences and Technologies (CELSTEC)
6401 DL Heerlen, The Netherlands
0031-(0)45-576-2218
{soude.fazeli,hendrik.drachsler,peter.sloep}@ou.nl

ABSTRACT
In this paper, we present network visualizations and an analysis of publications data from the LAK (Learning Analytics and Knowledge) in 2011 and 2012, and the special edition on Learning and Knowledge Analytics in Journal of Educational Technology and Society (JETS) in 2012.

Categories and Subject Descriptors
H.3.3 [Information Search and Retrieval]: Information filtering; K.3.m [computers and education]: Miscellaneous

General Terms
Algorithm, visualizations

Keywords
Network, recommender, visualization, dataset, learning analytics, degree

1. Introduction
The Society for Learning Analytics Research (SOLAR)1 provided a dataset to solicit contributions to the LAK data challenge2 sponsored by the FP7 European Project LinkedUp3. The dataset contains research publications in learning analytics and educational data mining for the years 2010, 2011, and 2012 (Taibi & Dietze, 2013). An overview of the dataset is shown in Figure 1. The dataset contains in total, 173 authors and 76 papers from the LAK (Learning Analytics and Knowledge) conference series in 2011 and 2012, and the special edition on learning and knowledge analytics in the Journal of Educational Technology and Society (JETS) in 2012. We found 24 authors who contributed to all three scientific proceedings.

Having access to a dataset always offers new opportunities, particularly in the educational domain, that lacks public datasets for running experimental studies (Verbert, Drachsler, Manouselis, Wolpers, Vuorikari, & Duval, 2011). Therefore, we used this dataset to present visualization of the authors and papers network, and to carry out a deeper analysis of the generated networks. Our overall aim is to use such a graph of authors and papers to recommend similar items to a target user. In the following sections, we evaluate the suitability of the LAK dataset for this purpose.

2. Motivation
It is often difficult for conference attendees to decide which workshops or sessions are suitable and relevant for them. Therefore, a list of recommended authors and papers based on shared interests could be supportive to plan the conference participation more efficiently and effectively. There already exist several papers published regarding awareness support for researchers (Reinhardt et al., 2012; Fisichella et al., 2010; Ochoa et al., 2009; Henry et al., 2009) and scientific recommender systems (Huang et al., 2002; Wang & Blei, 2010) but none of them has analyzed the Learning Analytics datasets for this purpose yet.

Our overall vision is to support the LAK attendees with a list of LAK authors and papers that are relevant for their own research interests. Such a recommendation could be created based on one or more of their own research papers but also on a short essay or even a tag cloud summarizing the research interest and objectives. Such a priority list can support the awareness of the attendees and empower the network of like-minded authors in the attendees’ particular research focus.

24 51 119
24 42 27
0 20 40 60 80 100 120 140
LAK 2011 LAK 2012 JETS2012

Figure 3. The used datasets
In this paper, then, we aim to explore and identify like-minded authors within the LAK dataset. Supposing that we have a network of all the LAK authors and papers, the main research questions are:
RQ1. How are the authors connected and which authors share more connections and are more central in terms of sharing commonalities with the others?
RQ2. How are the papers connected to each other in terms of similarity?

To answer these questions, we went through two main steps in our analysis: 1. Finding patterns of similarity between authors and

---

1 http://www.solaresearch.org/
2 http://www.solaresearch.org/events/lak/lak-data-challenge/
3 http://linkedup-project.eu/
papers, 2. Visualizing networks of the LAK authors and papers. We will now describe each step in detail.

![Figure 2. The LAK authors’ network](image)

(The Appendix shows a larger version)

3. Data processing

To find relationships between authors, we first computed the similarity of the papers with the TF-IDF\(^4\) algorithm. TF-IDF can create a weighted list of the most commonly used terms in research articles. To generate the TF-IDF matrix for the LAK dataset, we first converted the LAK data from RDF to text files, which is an accepted format for the Mahout\(^5\) system. Then, we ran the default TF-IDF algorithm provided by Mahout on the text files. We removed the stop words by setting the configuration variables within Mahout to 90%. Thus, if a word appears in 90% of the document, it is considered as a stop word (e.g. and, or, the, etc.) and is removed from the similarity matrix. As a final outcome we had:

- A so-called dictionary of all the terms in the LAK dataset
- A binary sequence file that includes the TF-IDF weighted vectors

For computing similarity between the LAK authors, we used the T-index algorithm (Fazeli, Zarghami, Dokoohaki, & Matskin, 2010) as a collaborative filtering recommender algorithm that generates a graph of users. In it the nodes are users and the edges show the relationship between users that originates from similarity of user profiles. The T-index algorithm originally makes recommendations based on the ratings data of users. We extended the T-index algorithm to be able to process tags and keywords extracted from the linked data e.g. RDF files. We used Jena\(^6\) APIs to process RDF files and to handle Ontology Web Language (OWL) files that describe the generated graph of authors and papers. Jena helps to develop semantic Web application and tools.

4. Data visualization

We visualized the generated graphs of authors and papers with the Welkin\(^7\) tool. Welkin takes an OWL file as input and provides visualization of the data as output. We present visualizations of the LAK authors and the LAK papers generated by Welkin in the following sub sections.

4.1. The LAK authors network

Figure 2 presents a network of the LAK authors in which red nodes represent the authors and the edges show the similarity between the publications of two authors. The result shows how the LAK authors are connected in terms of their publications’ commonalities. Moreover, the network shows the users who share more commonalities than do other authors. We call them ‘central authors’. In the next section, we show how they are connected with the other authors in the network.

4.2. The LAK authors’ degree centrality

For some node in the network, the degree centrality shows the total number of incoming and outgoing edges. It is a metric commonly used for Social Network Analysis (SNA) (De Liddo, Buckingham Shum, Quinto, Bachler, & Cannavacciuolo, 2011; Gu’eret, Groth, Stadler, & Lehmann, 2012; Opsahl, Agneessens, & Skvoretz, 2010). In other words, the degree of a node describes how many other nodes are connected to the target node. In fact, it helps to measure how many hubs are in the network. We describe hubs as the nodes that have the most connections to the others in the network. The degree centrality metric may be used to strengthen a network by providing its nodes with more connections. In this data study, degree centrality is used to measure the relevance of an author’s papers to the other authors in the network.

![Figure 3. The degree centrality of the top ten central authors](image)

Figure 3 shows the degree centrality for the first ten authors with the highest similarity degree with respect to the LAK publications. The horizontal axis (x) shows the top ten central users, e.g. u1 is the author whose paper(s) has the highest degree. The vertical axis (y) shows the degree values that describe the number of relationships of a each user shown in the x-axis. Figure 3 also shows degree centrality for two different sizes of nearest neighborhoods (n). Such neighborhoods are commonly used in collaborative filtering recommender algorithms. By increasing the neighborhood size n, the degree of the authors increases accordingly. As a result, we will have a larger number of central authors when n is higher (e.g. n=10). As can be seen in Figure 3, degree for the first central author (u1) is equal to 121 if n=10 and 97 if n=5. These high scores show the high relevancy of u1’s publications to the authors. As a consequence, u1 will appear in the top-n authors recommendations more often than the other authors.

---


\(^6\) [http://jena.apache.org/](http://jena.apache.org/)

\(^7\) [http://simile.mit.edu/welkin/](http://simile.mit.edu/welkin/)
chance to appear in the top papers (higher when considering the highest degree value among the papers). As a consequence, we will have a larger number of top papers and the edges between them represent the similarity of the papers. By finding similar papers, we can recommend the most similar papers to specific authors. This increases the awareness of the authors about papers which are relevant to them and published in their communities.

Figure 4 shows that, some of the papers share more similarity with the others and own a higher degree number. As with the central authors, these papers will appear more often in the top recommendation list than the other papers of the dataset. One may interpret their degree as their popularity. Therefore, the papers with higher degree values are more popular and, presumably, they are more of interests to users. For the publication data, interests of users derives from the words and terms they have used more frequently in their papers.

![Figure 4. The LAK papers network](The Appendix shows a larger version)

### 4.3. The LAK papers network

Figure 4 shows a network of the LAK papers. The red nodes are the papers and the edges between them represent the similarity of the papers. By finding similar papers, we can recommend the most similar papers to specific authors. This increases the awareness of the authors about papers which are relevant to them and published in their communities.

Figure 4 shows that, some of the papers share more similarity with the others and own a higher degree number. As with the central authors, these papers will appear more often in the top recommendation list than the other papers of the dataset. One may interpret their degree as their popularity. Therefore, the papers with higher degree values are more popular and, presumably, they are more of interests to users. For the publication data, interests of users derives from the words and terms they have used more frequently in their papers.

5. **Discussion and conclusions**

The results presented here, allow us to answer our research questions in the following way:

**RQ1. How are the authors connected? Which authors share more connections and are more central in terms of sharing commonalities with the others?**

We presented a visualization of the authors’ network to provide an overview of how they are connected to each other. To justify the authors’ connections and relationships, we evaluated the degree centrality for the first ten, most central authors. Table 1 presents the first ten central authors and their degree to show the authors with the highest relevancy of their publications with others in the network. Table 1 shows the degree of the authors for sizes of neighborhoods equal to 10.

![Figure 5. The degree centrality of Top ten papers](The Appendix shows a larger version)

**Table 1. The first ten central authors**

<table>
<thead>
<tr>
<th>Author</th>
<th>Degree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hendrik Drachsler</td>
<td>116</td>
</tr>
<tr>
<td>Kon Shing Kenneth Chung</td>
<td>87</td>
</tr>
<tr>
<td>Wolfgang Greller</td>
<td>80</td>
</tr>
<tr>
<td>Javier Melenchen</td>
<td>66</td>
</tr>
<tr>
<td>Brandon White</td>
<td>59</td>
</tr>
<tr>
<td>Vania Dimitrova</td>
<td>50</td>
</tr>
<tr>
<td>Erik Duval</td>
<td>45</td>
</tr>
<tr>
<td>Rebecca Ferguson</td>
<td>44</td>
</tr>
<tr>
<td>Anna Lea Dyckhoff</td>
<td>40</td>
</tr>
<tr>
<td>Simon Buckingham Shum</td>
<td>39</td>
</tr>
</tbody>
</table>

**RQ2. How are the papers connected to each other in terms of similarity?**

We presented degree centrality of the LAK papers to give insight in their relationships in the papers’ visualized network. We selected the top ten papers that have the highest similarity with the other papers. To show which papers are placed in the top ten papers’ list, we present the title and authors for each paper.

The top ten papers are not necessarily by the authors who are identified as the central authors. Although most of the central authors also appear in top ten papers’ list (see Table 2), the order is not the same. As we investigated the LAK data, we found out that some of the central authors have more than one paper. For instance, Hendrik Drachsler has contributed to four papers. In this study, similarity is calculated based on all papers of an author. So, it is quite probable that not each and every one of the authors’ papers individually has the highest similarity to the other papers. Although some of the central authors are common to the two
tables, only one of the papers authored by those central authors appears in the top ten papers list shown by Table 2.

Table 2. The Top ten papers

<table>
<thead>
<tr>
<th>Paper</th>
<th>Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learning Dispositions and Transferable Competencies: Pedagogy, Modelling and Learning Analytics</td>
<td>Simon Buckingham-Shum, Ruth Deakin Crick</td>
</tr>
<tr>
<td>The Pulse of Learning Analytics Understandings and Expectations from the Stakeholders</td>
<td>Hendrik Drachsler, Wolfgang Greller</td>
</tr>
<tr>
<td>Social Learning Analytics: Five Approaches</td>
<td>Rebecca Ferguson, Simon Buckingham-Shum</td>
</tr>
<tr>
<td>Multi-mediated Community Structure in a Socio-Technical Network</td>
<td>Dan Suthers, Kar Hai Chu</td>
</tr>
<tr>
<td>Modelling Learning &amp; Performance: A Social Networks Perspective</td>
<td>Walter Christian Paredes, Kon Shing Kenneth Chung</td>
</tr>
<tr>
<td>Teaching Analytics: A Clustering and Triangulation Study of Digital Library User Data</td>
<td>Beijie Xu, Mimi M Recker</td>
</tr>
<tr>
<td>Monitoring Student Progress Through Their Written &quot;Point of Originality&quot;</td>
<td>Johann Ari Larusson, Brandon White</td>
</tr>
<tr>
<td>Learning Designs and Learning Analytics</td>
<td>Lori Lockyer, Shane Dawson</td>
</tr>
<tr>
<td>A Multidimensional Analysis Tool for Visualizing Online Interactions</td>
<td>Eunchul Lee, M’hammed Abdous</td>
</tr>
<tr>
<td>Using computational methods to discover student science conceptions in interview data</td>
<td>Bruce Sherin</td>
</tr>
</tbody>
</table>

Overall, we found that the LAK dataset can help conference attendees to become more aware of their research network, which, in its turn, is useful for sharing knowledge and experiences. However, the current dataset contains no user feedback or evaluations to evaluate either an author or a paper recommender system in terms of common metrics such as prediction accuracy and coverage of the generated recommendations. For future analysis it would be helpful if the LAK dataset also contains references to the papers. The references could be used to identify the top cited authors and papers within the LAK dataset and beyond. As a further step, we are planning to try additional social network analysis measures besides degree, such as betweenness or closeness.

6. References


7. Appendix

7.1. The LAK authors’ network

7.2. The LAK papers’ network
Paperista: Visual Exploration of Semantically Annotated Research Papers

Nikola Milikic  
Faculty of Organizational Sciences, University of Belgrade  
Jove Ilića 154, Belgrade 11000, Serbia  
+381-11-3950853  
nikola.milikic@gmail.com

Uros Krcadinac  
Faculty of Organizational Sciences, University of Belgrade  
Jove Ilića 154, Belgrade 11000, Serbia  
+381-11-3950853  
uros@krcadinac.com

Jelena Jovanovic  
Faculty of Organizational Sciences, University of Belgrade  
Jove Ilića 154, Belgrade 11000, Serbia  
+381-11-3950853  
jejov@gmail.com

Bojan Brankov  
UZROK Labs  
107 Nehruova, Belgrade 10070, Serbia  
+381-63-581879  
bb@uzrok.com

Srdjan Keca  
UZROK Labs  
107 Nehruova, Belgrade 10070, Serbia  
+381-61-311566  
sk@uzrok.com

ABSTRACT
We consider the problem of visualizing and exploring a dataset about research publications from the fields of Learning Analytics (LA) and Educational Data Mining (EDM). Our approach is based on semantic annotation that associates publications from the dataset with Wikipedia topics. We present a visualization and exploration tool, called Paperista (www.uzrok.com/paperista), which presents these topics in the form of bubble and line charts. The tool provides multiple views, thus allowing users to observe and interact with topics, understand their evolution and relationships over time, and compare data originating from different research fields (i.e., LA and EDM). Moreover, users can explore papers to which the presented topics are related to, and make related Web searches to access the papers themselves.

Categories and Subject Descriptors
D.2.2 [Software Engineering]: Design Tools and Techniques - user interfaces

General Terms
Algorithms, Design

Keywords
Learning Analytics, Visualization, Research Papers

1. MOTIVATION
The field of Learning Analytics is emerging in the past few years and attracting more and more researchers from other areas of Technology Enhanced Learning (TEL). It aims to address the current needs in the broad area of education by making use of the latest trends in information technologies where everything is moving towards Big Data and real-time analytics.

Learning Analytics (LA) is defined as “the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the environments in which it occurs” [16]. It is often equated with other similar fields in the TEL area, such as Academic Analytics or Educational Data Mining (EDM) [14]. EDM is a research field that focuses on using computational approaches, namely data mining and machine learning, to analyze educational data in order to facilitate and enhance educational process, and contribute to the overall improvement of students’ learning experience [17]. Even though both LA and EDM are self-contained research fields, they are intertwined and overlap in topics they cover. They share many similarities, but also have some distinct differences as discussed by Siemens and Baker [18].

The Society for Learning Analytics Research (SoLAR) has published LAK dataset1 containing structured data about research publications from Learning Analytics and Knowledge (LAK) Conference, Educational Data Mining Conference, and Journal of Educational Technology & Society (JETS) Special Issue on LAK. The data are represented in the RDF form, which makes them easy to integrate and process by applications.

In this paper, we propose an approach to visualizing and exploring the LAK dataset. It is centered around the topics covered by the papers from the dataset, and is intended to give an overall view of the topics that LA and EDM fields cover. As the focus of researchers and the degree of relevance of particular topics have been changing over years, our approach tries to show a trend of those changes through the whole period the dataset covers, namely from 2008 to 2012. It also allows for topic-based exploration of research papers and easy navigation to them.

2. RELATED WORK
In [11], authors present an interesting work aimed at automating the creation of relations between research areas by using semantically annotated data about research papers in a particular

---

1 www.solaresearch.org/resources/lak-dataset
area. As a continuation of this work, the same authors have created a tool, called Rexplore\(^2\), which, among other things, visualizes authors migration patterns across research areas [15].

In terms of visual representation, we find interesting an approach to visualization of tags (topics) and categories of tags over time. For example, Dubinko et al. [4] consider the problem of visualizing the evolution of Flickr tags. The authors present a new slider-based approach based on a characterization of the most interesting tags. A Flash-based animation in a web browser allows the user to observe and interact with the tags. Zhang et al. [5] present an approach to classification and visualization of temporal and geographic tag distributions. The authors argue that their approach can help humans recognize semantic relationships between tags. Lemma [6] presents the Ebony system, an application for browsing, navigation, and visualization of the DBLP database. Wattenberg [7] introduces arc diagrams for representing complex patterns of repetition in string data. Wattenberg application, the Shape of Song, visualizes music files, creating a static representation of repetition throughout a time series. However, to our knowledge, there has been no (published) research work on the visualization of research topics and publications in the areas of LA and EDM.

### 3. THE PAPERISTA SYSTEM

Our approach is illustrated through a Web application called Paperista. The application visualizes topics associated with research publications from the LAK dataset, allowing users to browse through papers, compare LA and EDM research fields, and make related Web searches. Visualizations are created for each individual year in order to display relevant topics in the LA and EDM fields for a specific year, but also for all years combined in order to give an overall depiction of the topic distribution in these research areas.

#### 3.1 Data Preparation and Analysis

LAK dataset consists of data about conferences and journal papers published in the LA and EDM research fields in the 2008-2012 period. For each paper, the following elements are available: title, author(s), abstract, keyword(s) and full text. Also, basic information about authors is available, such as name and affiliation.

##### 3.1.1 Topic Extraction

Since one of the main features of Paperista is visualization of research topics relevant for the given corpus, the first step in the data preparation process was to extract main topics of the papers encompassed by the LAK Dataset. A straightforward approach was to use keywords associated with the papers. This is because the authors themselves have compiled those keywords, and it is them who know the best which topics describe their work in the most appropriate way. However, the downside of this approach is that those keywords are given as free form text and are not consistent with any existing formal vocabulary. This makes them inconsistent throughout the corpus. Furthermore, the dataset is incomplete in regard to keywords as for conferences EDM 2008, 2009 and 2010 no keywords are provided.

Thus, we decided to employ a service for semantic annotation in order to detect paper topics. We took into consideration two Wikipedia based semantic annotators: TagMe\(^3\) and DBpedia Spotlight\(^4\). The decision to use Wikipedia based annotator was motivated by the fact that Wikipedia is the largest corpus of open encyclopedic knowledge and is often used as a well established large-scale taxonomy [8]. Both annotator services are designed to look for and retrieve recognized Wikipedia concepts from the given text. They can be configured to the specific needs of any particular usage scenario (i.e., corpus). TagMe is designed to identify Wikipedia concepts specifically in short texts. Its REST API\(^5\) allows for configuration of two parameters: i) the rho parameter which refers to the "goodness" of an annotation with respect to the topics of the input text, and ii) the epsilon parameter which is used for fine-tuning the disambiguation process and indicates whether to favor the most-common topics or to take the context more into account [9]. DBpedia Spotlight annotates a given text with concepts from DBpedia, a structured representation of Wikipedia [12]. DBpedia Spotlight REST API\(^6\) exposes two parameters: confidence of the annotation process that takes into account factors such as the topical pertinence and the contextual ambiguity; support parameter specifies the minimum number of inlinks\(^7\) [10]. We used only paper title and abstract for topic extraction, based on an assumption that these two elements contain mentions of the most important and interesting topics a paper is related to. In order to decide which service for semantic annotation to use, the two services were tested with a random sample comprising 5% of all papers and with different parameter settings. The best results were achieved by the TagMe service (rho=0.15; epsilon=0.5). For this reason, TagMe service was employed to annotate all papers in the corpus.

##### 3.1.2 Identifying Popular Topics

Once having all papers associated with topics, we calculated the significance of each topic. Numerical statistic called TF-IDF (Term Frequency – Inverse Document Frequency)\(^8\) was used as it calculates how important a word is to a document in a corpus of documents. This metric was adapted to our case and used to calculate the importance of a topic in a paper. Instead of calculating the frequency of a word, we calculate the frequency of a topic.

Since Paperista allows for visualizing topics in a specific year and overall (in all years, 2008-2012), the significance was calculated for corpora containing papers from each of these different time periods. Accordingly, we had six different corpora and calculated the significance of a topic for each corpus. In order to present only the most significant topics, we have filtered the topic set to only those whose significance for a particular period was over 0.01. This threshold was empirically chosen and presents the best balance between the relevance of topics and their presentation in the Paperista’s visualizations (i.e., assuring easy comprehension by users).

##### 3.1.3 Topic Cleaning

Even though the output of TagMe service consisted of topics that are relevant to the papers’ content, some of them can hardly be considered as relevant research topics in the LA and EDM fields as they are too general. For instance, topics like Methodology,

---

\(^2\) http://technologies.kmi.open.ac.uk/rexplore

\(^3\) http://tagme.di.unipi.it

\(^4\) http://spotlight.dbpedia.org

\(^5\) http://tagme.di.unipi.it/tagme_help.html

\(^6\) http://github.com/dbpedia-spotlight/dbpedia-spotlight/wiki/REST-API

\(^7\) Inlink, or inline link, are incoming links from other DBpedia concepts to the observed DBpedia concept

\(^8\) http://en.wikipedia.org/wiki/Tf-idf
Research, and Experiment can be associated with almost every paper in this corpus. Actually, these topics can be related to research papers from almost any other research area. Similarly, some of the retrieved topics were not relevant to research papers from the LAK dataset. Such topics resulted from imperfection of the TagMe tool (and semantic annotation tools, in general). Some examples of these alien topics include The T.O. Show, Ade Easily, Henry Snapp, etc. For instance, Henry Snapp topic was apparently mistaken with the SNAPP tool\(^9\), a popular learning analytics and visualization tool. Hence, it was important to detect and exclude all these generic and alien topics from the final visualization in order to reduce the noise.

We applied topic cleaning approach similar to [11]. The idea is to identify topics that have little or no relationships with other topics in the corpus. This can be an indicator that a topic is too specific or alien to our set of identified topics and thus can be considered as an exclusion candidate. On the other hand, if a topic has relationships with too many other topics, this can be an indicator that a topic is too generic and again should be considered as an exclusion candidate. In order to detect these outlier topics, we needed a measure of relatedness between topics. To that end, we used the Wikipedia Miner\(^10\) service that calculates semantic relatedness of two topics by finding the corresponding Wikipedia articles, and calculating similarity of those articles by comparing their incoming and outgoing links [13]. Wikipedia Miner has a REST API\(^11\) that allows for retrieving this information programmatically.

Once having relatedness calculated for all the topics in our corpus, we compiled two lists to help us detect removal candidates. In the first list, each topic was associated with the number of other topics that topic is related to. This gave us an insight into which topics can be considered too general/specific (the higher the number of related topics, the more generic the topic is, and vice versa). In the second list, each topic was associated with a sum of its relatedness with all the other topics. This list was meant to complement the first one. The rationale here is that there might be a topic with fair number of relations to other topics, but those relatedness values are weak. This behavior also qualifies a topic to be considered as too specific or alien.

The initial idea with compiling these two lists was that topics to be removed will be at the beginning and the end of the lists (top and bottom 10%), and that they could be removed automatically. However, by examining the lists, among the obvious exclusion candidates, there were also several topics that should not have been excluded. For instance, topics like Online tutoring, Process mining, Educational data mining etc. were at the end of both lists making them removal candidates, even though these topics are obviously highly relevant for LA and EDM fields. The reason for this lays in the nature of Wikipedia itself and the fact that not many other articles in Wikipedia link to these topics. Thus, the topic removal process could not be done completely automatically and an expert in the area was consulted to mark the topics that should not be excluded.

\(^9\) http://www.snappvis.org
\(^10\) http://wikipedia-miner.cms.waikato.ac.nz
\(^11\) http://wikipedia-miner.cms.waikato.ac.nz/services

![Figure 1 - Paperista Interface](image-url)
3.2 Data Visualization and Exploration
The topic visualization applied in Paperista is inspired by the New York Times visualizations *Four Ways to Slice Obama’s 2013 Budget Proposal* [1] and *At the National Conventions, the Words They Used* [2].

The Paperista visualization includes bubble and line charts, allowing users to gain insights into topic trends within the LA and EDM fields. Bubble charts show the importance of a certain topic for the entire dataset, each year, and/or each field. By changing different views, users can watch the changes within the dataset and compare the two fields. Animated transitions between charts help users understand these processes. In addition, since the animation does not show precise changes in topic’s relevancy (calculated using TF-IDF metric, see Sect. 3.1), users are also presented with a relevancy line chart for each topic.

The user interface (Figure 1) consists of an animated bubble cloud, two button sliders, a sidebar, and an optional timeline. The first slider button (All Years / By Year) allows users to choose between the “All Years” and “By Year” views. “All Years” view presents relevant topics for the entire corpus of publications. “By Year” view activates a timeline, showing relevant topics for each year. By using the slider, users can follow the change in topic relevancy through the years the data is available for (2008-2012).

The second slider button (All Topics / Group Topics) allows for grouping and regrouping of topics. “All Topics” view shows one circle-shaped bubble chart. “Group Topics” view divides the chart into two groups of bubbles. The first group presents topics that appear only in the EDM field. The third one shows topics related only to the LA field (i.e., LAK and JETS publications). The group in the middle shows “mixed” topics, i.e., those that appear at least once in both EDM and LAK/JETS. Different views of the bubble chart are presented on Figure 2.

The size of a bubble represents the topic’s relevancy (i.e., TF-IDF value). Two research fields, EDM and LA, are color-coded. Each bubble is divided into two slices the size of which corresponds to the frequency of that topic within publications of each of the two sources. For the years 2008-2010, the dataset contains data only for the EDM research field, so the bubbles are one-colored.

The order of topic bubbles is intended to help users compare the two fields. The leftmost bubbles represent mostly EDM-related topics, while the rightmost bubbles mostly belong to the LA field. Moreover, clicking on a bubble creates a line chart in a sidebar. The line chart shows the growth and decline of a certain topic.

In addition to the visualization, the Paperista application allows users to browse papers by topic. When a user clicks on a particular bubble (topic), a list of papers related to that topic appears in the right sidebar (represented by a title and a list of authors). Clicking on the particular paper opens a link to Google scholar with a name of the article as a search query. Thus, if a paper is available online, a user could easily obtain the paper using the Paperista system.

Figure 2 - Different views of the bubble chart: (1) All Year / All Topic (no highlights); (2) All Years / Group Topics (with highlights); (3) By Year (2009) / All Topics (no highlights); and (4) By Year (2012) / Group Topics (with highlights)
Furthermore, when a user hovers over the paper title, all topics related to that paper become highlighted. By hovering over papers, users can gain quick insight about topic connections between publications. Users can also distinguish papers annotated with highly relevant topics from those marked with insignificant ones. This can show which papers are more related to the fields of EDM and LA, and which can be viewed as “outliers”.

3.3 Paperista Architecture and Dataset API
The Paperista system consists of a Web application and a server application that provides RESTful API for communicating with the dataset. The Web-based visualization is written in D3, a JavaScript library for manipulating documents based on data [3]. We have chosen D3 because of its good performance for animation and interaction within the Web environment. The visualization is available at the following address: www.uzrok.com/paperista.

All data about conference topics and their significance (explained in Section 3.1) is available as a part of Paperista Dataset API. This API supports a REST model for accessing the data and it is available at: http://147.91.128.71:9090/LAKChallenge2013. The Paperista’s Web application calls these operations in order to access data from the dataset (for example, a click on a topic triggers a call to the API, which returns a list of papers).

4. DISCUSSION
When looking at the view displaying topic distribution in all years (Figure 2.1), one can observe that EDM conference dominates in almost all topics. This is due to the fact that EDM conference is being organized longer than the LAK conference (3 years longer), and thus the LAK dataset contains overall more papers coming from the EDM conference.

Filtering topics by years allows for observing the popularity of topics in a particular year and a particular field (LA or EDM). This further enables one to observe the shift in interest for a particular topic by researchers in the LA and EDM fields throughout the years. For instance, one can observe that before 2011, the topic of Learning Analytics was not much popular in the papers from the EDM field; thus this topic is not displayed at all in visualizations for years 2008-2010. In 2011, it boomed in popularity as indicated by the significant rise in the number of papers covering it. In fact, this was the first year the LAK conference was organized, and it immediately occupied the attention of researchers interested in the topic of Learning Analytics. Interestingly, this topic also gained some traction among the researchers publishing in the EDM field. In 2012, the topic’s popularity grew even bigger and the researchers covering it directed their effort toward the LA field. This resulted in papers published within the LA field to almost exclusively cover the topic of Learning Analytics. Similarly, we can observe topics that have kept high popularity in both areas over years. For instance, this is the case with the Data topic, obviously as a consequence of research in both areas concentrating on the analysis of large amounts of data coming from various learning systems and other sources.

The application also allows us to observe that topics such as Intelligent Tutoring System, Prediction and Accuracy and Precision mostly kept their popularity throughout the years and stayed exclusively within the EDM field. On the other hand, one can observe that the large majority of topics have been covered by both fields. This suggests that the similarities between the two fields are significant as they share many research topics.

5. CONCLUSION
In this paper we have presented our approach to visualizing topics and their trends in the LA and EDM fields. Our application allows for easy identification of the main topics researchers in these fields have been focusing on, and also exploration of papers related to those topics.

When compared to other similar tools that provide visualization of research topics, our tool is the most similar to the previously mentioned Rexplore tool. However, while Rexplore is more focused on relations between authors and topics in research areas, Paperista’s focus is on research topics and their trends over time.

Also, Paperista allows for exploring papers related to different topics.

Future work for Paperista will be primarily directed towards extending the system to support other datasets, related to other research areas. Since the LAK dataset is RDF-based, Paperista can easily be expanded to support other RDF-based datasets expressed using the same or related vocabulary, such as the Semantic Web Dog Food corpus [2]. Regarding the interface, we plan to introduce keyword-based search functionality for searching a topic by its name. This would allow for easy navigation to a desired topic and filtering papers related to it. The final goal for Paperista is to become a universal visualization tool for research papers.

6. REFERENCES

12 http://data.semanticweb.org


1st International Conference on Learning Analytics and Knowledge, Banff, Alberta, February 27–March 1, 2011, link https://tekri.athabascau.ca/analytics/


Analysis of the Community of Learning Analytics
Sadia Nawaz
Purdue University
West Lafayette, IN, USA
sadia@alumni.purdue.edu
Farshid Marbouti
Purdue University
West Lafayette, IN, USA
fmarbout@purdue.edu
Johannes Strobel
Purdue University
West Lafayette, IN, USA
jstrobel@purdue.edu

ABSTRACT
The trends of the learning analytics community being presented in this paper are in terms of authors, their affiliation and geographical location. Thus the most influential authors, institutes, and countries who have been actively contributing to this field are brought out. In addition, this paper identifies collaborations among authors, institutes, and countries. The paper also tries to explore the research themes followed by the learning analytics community.

1. DATA AND TOOLS
The data that is analyzed in this paper consists of the conference on Learning Analytics and Knowledge (LAK) 2011–2012, Educational Data Mining (EDM) conference 2008–2012 and the Journal of Educational Technology and Society (JETS) special edition on learning and knowledge analytics. This data was provided on the Society for Learning Analytics Research (SoLAR) website in xml format [1]. The xml data converted to tabular data using an xml to csv convertor [2]. The converted csv files were then processed and merged using macro programming in MS Excel. Later, this data was using NodeXL tool – an open source template for Microsoft® Excel® [3]. It allows the user to work on different worksheets for different operations such as ‘Edges’ worksheet can be used to compute the inter/intra collaboration. ‘Vertices’ worksheet allows the display and computation of individual node properties such as degree, betweenness, centrality etc. Other tools that have been utilized in this paper include NetDraw [4] and IBM’s Many-eyes [5].

2. MOTIVATION
With increase of attention to interdisciplinary field of Learning Analytics, scholars from different disciplines such as education, technology, and social sciences are contributing towards this field [6]. Different authors with different backgrounds, expertise and purpose publish and present their work in Learning Analytics related journals and conferences. To draw a better understanding of who are top collaborators in the field and which institutes and countries are more active in creating and disseminating knowledge, we analyzed the data described in the previous section.

3. AUTHORSHIP TRENDS
Complete summary of various (author related) statistics has been provided in table 1 (detailed definition of these graph theory related terms is available at [7]). Analysis of authors provides information which not only helps in understanding the growth of the field (in terms of publication counts and author counts etc.) but also is used to predict the future of the field e.g.: information such as ‘connected components’ and ‘maximum edges in a connected component’ is showing that the graphs are getting well populated and connected – thus, employing more inclination towards collaboration Overall it can be said that the field itself is growing as apparent from node counts (2008-2012) and article counts (the sum of single and multi-author article counts). Similarly, self-loop count together with single vertex connected component can show how many authors of the single authored publication have / have not collaborated (within this data) e.g., the last column indicates that overall there have been 26 single-authored articles by 25 authors. It was found that 14 of these authors have had no collaborative work in this data. And it was also found that ‘Stephen E. Fancsali’ is the only author with two single authored publications.

<table>
<thead>
<tr>
<th>Graph Metric (graph theory terminologies)</th>
<th>2008</th>
<th>2009</th>
<th>2010</th>
<th>2011</th>
<th>2012</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total unique vertices / nodes (authors)</td>
<td>74</td>
<td>79</td>
<td>151</td>
<td>193</td>
<td>281</td>
<td>623</td>
</tr>
<tr>
<td>Unique edges (edge is loop for single author articles &amp; straight line otherwise)</td>
<td>100</td>
<td>106</td>
<td>208</td>
<td>251</td>
<td>435</td>
<td>938</td>
</tr>
<tr>
<td>Edges with duplicates (i.e., edge weight is greater than 1) (These edges show joint authorship in more than one publication)</td>
<td>17</td>
<td>18</td>
<td>50</td>
<td>42</td>
<td>48</td>
<td>337</td>
</tr>
<tr>
<td>Total edges</td>
<td>117</td>
<td>124</td>
<td>258</td>
<td>293</td>
<td>483</td>
<td>1275</td>
</tr>
<tr>
<td>Self-loop (single author articles)</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>10</td>
<td>8</td>
<td>26</td>
</tr>
<tr>
<td>Multi-author article count</td>
<td>27</td>
<td>31</td>
<td>61</td>
<td>75</td>
<td>96</td>
<td>27</td>
</tr>
<tr>
<td>Connected components (authors forming a cluster based on authorship)</td>
<td>20</td>
<td>22</td>
<td>38</td>
<td>53</td>
<td>79</td>
<td>140</td>
</tr>
<tr>
<td>Single-vertex connected components (Count of the authors of single author articles who did not collaborate)</td>
<td>4</td>
<td>0</td>
<td>3</td>
<td>8</td>
<td>7</td>
<td>14</td>
</tr>
<tr>
<td>Maximum vertices in a connected component</td>
<td>15</td>
<td>7</td>
<td>15</td>
<td>29</td>
<td>22</td>
<td>113</td>
</tr>
<tr>
<td>Maximum edges in a connected component</td>
<td>33</td>
<td>16</td>
<td>36</td>
<td>72</td>
<td>76</td>
<td>370</td>
</tr>
</tbody>
</table>
4. COLLABORATION TRENDS
Collaboration as defined in Oxford dictionary [8] is the ‘action of working with someone to produce something’ and in current context it represents co-authorship of an article by two or more researchers. This term can be extended to institutes and even countries and hence extended collaboration patterns will be extracted between and within institutes and countries respectively. Table 2 shows that there have been 938 pairs of authors who collaborated just once (this number includes single author articles - since in that case a self-loop serves as an edge to itself). Alternatively, it can be stated that 73.57% of all articles have been written by the authors who have collaborated just once. It could either mean that new collaborations are forming or that the authors published just once and then they started working in other research areas, with other authors or they started targeting other venues. Therefore, initiatives such as LAK Data challenge will attract more researchers towards this field and hence may help in further growth and development of authorship networks.

Table 2: Overall collaboration pattern

<table>
<thead>
<tr>
<th>Author Pairs</th>
<th>Article Counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td>15</td>
<td>3</td>
</tr>
<tr>
<td>110</td>
<td>2</td>
</tr>
<tr>
<td>938</td>
<td>1</td>
</tr>
<tr>
<td>1(10)+2(6)+2(5)+10(4)+15(3)+110(2)+938(1) =1275</td>
<td></td>
</tr>
</tbody>
</table>

Table 3 presents some of the top collaborators e.g., N.T. Heffernan had been a co-author with J.E. Beck and Z.A. Pardos in 6 articles. Such analysis can help in finding active researchers and collaborators in this field.

Table 3: Top collaborators based on article count

<table>
<thead>
<tr>
<th>Author</th>
<th>Author</th>
<th>Article Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>S. Ventura</td>
<td>C. Romero</td>
<td>10</td>
</tr>
<tr>
<td>Neil T. Heffernan</td>
<td>Joseph E. Beck, Zachary A. Pardos</td>
<td>6, 6</td>
</tr>
<tr>
<td>Arnon Hershkovitz</td>
<td>Rafi Nachmias</td>
<td>5</td>
</tr>
<tr>
<td>Sujith M. Gowda</td>
<td>Ryan S. J. d. Baker</td>
<td>5</td>
</tr>
</tbody>
</table>

5. DIVERSITY
Diversity in this context is the count of distinct researchers – a given author may have worked with. Table 4 aims at identifying the contributors who have worked with most diverse group of authors e.g., K.R. Koedinger has worked with 34 distinct authors and Ryan Baker has worked with 25 distinct authors. We also extracted the graph of these top contributors (based on degree) i.e., a graph which includes these top authors and all of their collaborators; and it was found that this new graph consists of 128 authors (roughly 21% of the total authors). This percentage shows the significance of the top authors towards EDM, LAK, JETS and in general towards learning analytics.

Table 4: Top 10 authors with highest degree counts

<table>
<thead>
<tr>
<th>Author</th>
<th>Degree</th>
<th>Article Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kenneth R. Koedinger</td>
<td>34</td>
<td>17</td>
</tr>
<tr>
<td>Ryan S. J. d. Baker</td>
<td>25</td>
<td>11</td>
</tr>
<tr>
<td>C. Romero</td>
<td>19</td>
<td>11</td>
</tr>
<tr>
<td>Vincent Aleven</td>
<td>18</td>
<td>5</td>
</tr>
<tr>
<td>S. Ventura</td>
<td>17</td>
<td>11</td>
</tr>
<tr>
<td>Neil T. Heffernan</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>Sujith M. Gowda</td>
<td>15</td>
<td>5</td>
</tr>
<tr>
<td>Mykola Pecheniziky</td>
<td>15</td>
<td>7</td>
</tr>
<tr>
<td>Arthur C. Graesser</td>
<td>14</td>
<td>4</td>
</tr>
<tr>
<td>Jack Mostow</td>
<td>13</td>
<td>12</td>
</tr>
</tbody>
</table>

6. GEOGRAPHICAL LOCATION
Next, the geographical analysis of this dataset is presented which aims to explore the countries that have been extending this field especially through contributions to the venues: EDM, LAK and JETS. There have been contributions from 41 different countries. For extracting this information, all aliases of a country’s name were merged e.g., Netherlands, Netherlands, The_Netherlands etc. were all merged together. The top countries that have had international collaborations are provided in table 5. Clearly, USA and UK are on top of the list. To illustrate the collaboration patterns between countries figure 1 is drawn using ‘NetDraw’. In this figure an edge between two countries depicts the co-authorship between the researchers from these countries. The edge width (also represented by a number) shows the strength of such collaboration. Also, different symbols have been used for different nodes based on their ‘betweenness’ values. ‘Betweenness centrality’ is the “number of times a node acts as a bridge along the shortest path between two other nodes” [9]. Clearly, USA, UK and Germany are on top of this list based on degree and centrality measures. It is apparent that most of the nodes have ‘betweenness’ value of zero as depicted with a ‘+’ symbol. It indicates the peripheral nature of these nodes and thus depicts the birth or growth of this field – in that newer nodes are being added and the graph is currently sparse. Figure 2 illustrates geographical diversity of collaborators. The smaller circles show lesser diversity in terms of collaboration (with researchers from other countries). Similarly, larger circles are indicative of the countries whose researchers have more diverse group of co-authors (from across the world). In this figure a small table at the bottom depicts the count of papers from each continent. Thus it brings out the most active region for research in the area of learning analytics. Clearly, North America and Europe are at the top of this list (complete geographical mapping is available at [5]).

Table 5: Top international collaborators

<table>
<thead>
<tr>
<th>Country</th>
<th>Degree</th>
</tr>
</thead>
<tbody>
<tr>
<td>USA</td>
<td>11</td>
</tr>
<tr>
<td>UK</td>
<td>10</td>
</tr>
<tr>
<td>Australia, Germany</td>
<td>6</td>
</tr>
<tr>
<td>Netherlands</td>
<td>5</td>
</tr>
<tr>
<td>Canada, Belgium, Greece, Spain</td>
<td>4</td>
</tr>
</tbody>
</table>
Figure 1: Collaboration in terms of geographical location

Figure 2: Geographical diversity of collaborators
7. AUTHOR AFFILIATION

Next, the institutional affiliation of authors was analyzed and it was found that there have been contributions from 200 different institutes world-wide. The ranking of the top few institutes in terms of collaboration with other institutes is provided in table 6. The term degree represents count of unique institutes that a given institute may have worked with. This term can be influenced by both the ‘article counts’ and the ‘coauthor counts’. Table 7 provides the institutes with highest count of intra-institute collaboration and table 8 provides the ‘institute – pairs’ that have had highest collaboration. Such analysis is beneficial to research institutes and organizations so that they may collaborate and extend further studies in the field of learning analytics. Figure 3 illustrates trends of collaboration between institutes.

<table>
<thead>
<tr>
<th>Institute</th>
<th>Degree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carnegie Mellon University</td>
<td>20</td>
</tr>
<tr>
<td>University of Cordoba</td>
<td>9</td>
</tr>
<tr>
<td>Stanford University</td>
<td>8</td>
</tr>
<tr>
<td>Fraunhofer Institute for Applied Information Technology</td>
<td>7</td>
</tr>
<tr>
<td>Dept. Computer wetenschappen, KU Leuven</td>
<td>7</td>
</tr>
<tr>
<td>Worcester Polytechnic Institute</td>
<td>7</td>
</tr>
<tr>
<td>Open University of the Netherlands</td>
<td>6</td>
</tr>
<tr>
<td>University of Pittsburgh</td>
<td>6</td>
</tr>
</tbody>
</table>

Table 6: Top institutes with highest counts of distinct collaborators

<table>
<thead>
<tr>
<th>Institute</th>
<th>Self-loop count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Worcester Polytechnic Institute</td>
<td>116</td>
</tr>
<tr>
<td>Carnegie Mellon University</td>
<td>107</td>
</tr>
<tr>
<td>Eindhoven University of Technology</td>
<td>36</td>
</tr>
<tr>
<td>University of Cordoba</td>
<td>33</td>
</tr>
<tr>
<td>University of Memphis</td>
<td>31</td>
</tr>
<tr>
<td>Universitat Oberta de Catalunya (UOC)</td>
<td>31</td>
</tr>
<tr>
<td>University of North Carolina at Charlotte</td>
<td>20</td>
</tr>
<tr>
<td>RWTH Aachen University</td>
<td>16</td>
</tr>
</tbody>
</table>

Table 7: Top institutes with highest count of intra-institute collaboration

8. RESEARCH THEMES

In order to track the research themes being followed by learning analytics society and to see their emergence over time, the authors conducted a keyword based analysis. The information for this analysis has been extracted from the keyword (subject) section of the data provided by Society for Learning Analytics Research (SoLAR) website [1]. However, for initial two years i.e., 2008–2009 this field is empty, similarly some of the articles in later years had this field empty. Therefore, it was decided to use the ‘title’ field for the purpose of keyword extraction. The selection of ‘title’ field rather than the ‘abstract’ field for the purpose of keyword extraction relies on an earlier study by the authors of this paper [10]. Later, Hermetic Word Frequency Counter (HWFC) software [11] was used to parse out top 30 keywords for each year. Some of the common English keywords are already ignored by this software, as available in its stop word list. Other words which are apparent by the nature of the venues EDM, LAK and JETS were then manually eliminated (since they would not bring any insightful information for this analysis) e.g., student, learn, knowledge, education etc. Further refinement was made to merge varying instances of the same word such as ‘visual, visualize, visualization’ etc. Then, IBM’s Many-eyes software utility was used to obtain the Matrix Chart as provided in figure 4. In this figure top 30 keywords for each year have been presented. It should be noted that since the count of articles and venues has also increased over years; therefore, the relative rank or position of keywords will be discussed rather than absolute frequency counts. From this figure, it was found that the usage of some of the keywords such as ‘visualization, intelligent, network*’ is increasing over time. Some keywords such as ‘model*, system*, tutor*’ retain their ranks. The keywords ‘online, collaborat*, performance’ etc. show fluctuating trends. Similarly, other trends can be interpreted. The authors further extracted the context of these keywords: it was found that ‘visualization co-occurs with data-mining’, ‘intelligent appears with tutoring system’. The word ‘online’ has a broader class of co-occurring keywords which includes ‘learning, education, university, assessment systems, tutoring, courses, curriculum’ etc. Interestingly, in 2012 the context changed to ‘online communities, interactions and social learning’ etc. Due to space restriction further analysis cannot be provided in this paper.

CONCLUSION

In this paper the data of past five years of publications related to learning analytics are analyzed. The trends show increasing number of authors and more collaboration between authors as well as institutes. Geographical analysis of authors shows that scholars from different countries have been collaborating and contributing towards this field. Top authors, collaborators, and institutes are identified in this paper. The authors also attempted to bring out the research themes followed by the learning analytics community based on the frequency of the usage of keywords.

The authors plan to extend this study based on author’s disciplinary diversity and on the association between authors and their explored research areas within learning analytics.

<table>
<thead>
<tr>
<th>Institute</th>
<th>Institute</th>
<th>Edge weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Worcester Polytechnic Institute</td>
<td>Carnegie Mellon University</td>
<td>37</td>
</tr>
<tr>
<td>Claremont Graduate University</td>
<td>University of Memphis</td>
<td>18</td>
</tr>
<tr>
<td>University of Belgrade</td>
<td>Simon Fraser University</td>
<td>9</td>
</tr>
<tr>
<td>Northern Illinois University</td>
<td>University of Memphis</td>
<td>9</td>
</tr>
<tr>
<td>Hochschule fur Wirtschaft und Recht</td>
<td>Hochschule fur Technik und Wirtschaft</td>
<td>8</td>
</tr>
<tr>
<td>Beuth Hochschule fur Technik Berlin</td>
<td>Hochschule fur Technik und Wirtschaft</td>
<td>8</td>
</tr>
<tr>
<td>Universidade Federal de Alagoas</td>
<td>Carnegie Mellon University</td>
<td>8</td>
</tr>
<tr>
<td>Fraunhofer Institute for Applied Information Technology</td>
<td>Saarland University</td>
<td>8</td>
</tr>
</tbody>
</table>

Table 8: Top pairs for inter-institute collaboration
- Every connected component has been drawn as a separate group
- The nodes in each group have been presented in a different color and shape
- The smaller groups of connected components have been placed at the bottom
- The size of nodes vary with degree and the width of edges varies with edge weight
- The top 3 connected components of this figure are presented below to explore the main contributors / institutes in further detail.

Figure 3: Trends of collaboration in terms of author affiliation

REFERENCES
Figure 4: Keyword analysis for research theme extraction
Cite4Me: Semantic Retrieval and Analysis of Scientific Publications

Bernardo Pereira Nunes  
PUC-Rio  
Rio de Janeiro, Brazil  
bnunes@inf.puc-rio.br

Besnik Fetahu  
L3S Research Center  
Appelstrasse 9a  
Hannover, Germany  
fetahu@l3s.de

Marco Antonio Casanova  
PUC-Rio  
Rio de Janeiro, Brazil  
casanova@inf.puc-rio.br

ABSTRACT
This paper presents the Cite4Me Web application and its features created for the LAK Challenge 2013. The Web application focuses on two main directions: (i) interlinking of the LAK dataset with related data sources from the Linked Open Data cloud; and (ii) providing innovative search, visualization, retrieval and recommendation of scientific publications from the LAK dataset and related interlinked resources. Our approach is based on semantic and co-occurrence relations to provide new browsing experiences to Web users and an overview of scientific data available. Furthermore, we present a detailed analysis of the LAK dataset along with applications which contributes to the development of the learning analytics field.

1. INTRODUCTION
The volume of information on the Web has been growing steadily over the last decade and has doubled every two years. The vast amount of data available on the Web along with new means of communications have transformed our society, including the way we work, live, relate to each other and learn.

In the midst of change, the Learning Analytics emerges to make sense of the produced educational data reported by learners, professors, institutions and so on. Analyzing and understanding the changes along the past years help us to understand the current state and be aware of the forthcoming trends, enabling a new outlook of the future of learning.

A recent challenge initiative of SOLAR1 and LinkedUp2 project arises to leverage the creation of tools that enables the analysis, visualization, browsing and recommendation of scientific and educational data.

Although the scientific field has fostered the creation of new applications in several areas, such as medical, biology, physics, amongst others, the information access is based mostly on free text search and on hierarchical classification system of the publications3. However, current approaches by main digital library providers, such as ACM Digital Library4 and Elsevier5, do not represent the current state of research on exploring resources using approaches from Information Retrieval, Information Extraction and Semantic Web. Thus, get an overview of research topics, find publications and discover new nomenclatures are an arduous and laborious task that are not always successful.

In this paper, we introduce Cite4Me a novel application for exploratory search, retrieval and visualization of scientific publications. Cite4Me intends to provide to the end users a single point for accessing papers and hence reducing efforts of searching in several data sources. Our system takes advantage of reference datasets, such as DBpedia6, to explore semantic relationships between scientific papers and user queries. Additionally, an analysis of topic coverage and shared concepts from related educational datasets, extracted from the Linked Open Data cloud, will be introduced.

The remaining of the paper is organized as follows. Section 2 presents the approach used for searching, retrieving and recommending papers. Section 3 describes the process of dataset discovery and interlinking and Section 4 shows a brief result analysis of the data discovery. Finally, Section 5 presents related work and Section 6 presents some concluding remarks.

2. CITE4ME
As one of the main goals of the field of “Learning Analytics” is to support students in their learning process, we developed a Web application called Cite4Me7 that assists students in making decisions to find scientific publications and identify relevant research topics.

Cite4Me implements semantic and co-occurrence methods to (a) search and retrieve scientific publications; and (b) recommend scientific publications. Moreover, it provides a Web interface that facilitates the search for publications and may help users on discovering related terms to a given query.

In this section, we provide an overview of the major features of the Web application and its Web interface that assist users to explore scientific data on the Web.

2.1 Search and Retrieval
Cite4Me relies on search functionalities to meet the users needs. Briefly, we implemented standard Information Retrieval (IR) and Semantic Web (SW) approaches to retrieve and recommend scientific papers to the users. We divided this subsection into (i) free text

1Society for Learning Analytics Research - http://www.sola.research.org
2http://linkedup-project.eu/
3http://www.acm.org/about/class/
4http://dl.acm.org
5http://www.elsevier.com
6http://dbpedia.org
7http://www.cite4me.com/
search; (ii) exploratory search; and (iii) semantic search.

2.1.1 Free Text Search

The purpose of the free text search functionality is to offer users the abilities to search for mentions, titles and authors of academic publications contained in the LAK dataset. Even though, this functionality is similar to existing digital libraries, we agree that this is a basic functionality that must be provided by our application. Therefore, we use standard vector space models (tf-idf) for indexing and retrieving documents.

The tf-idf scores were computed for each term extracted from the publication content after applying stemming [14]. Furthermore, the searching functionality offers boolean queries with standard operators, such as 'OR', 'AND', and also a ranking of the matching publications based on the sum of tf-idf scores from the individual query terms.

In summary, our free text search provides to the users publications \( P \) that match query terms and non-matching publications \( P' \), which are related to \( P \) according to a degree of similarity (see Eq. 1), but does not contain the query terms.

The similarity between a matching publication \( P \) and other non-matching publication \( P' \) in the LAK dataset is measured by the standard cosine similarity measure, which is built on top of the computed tf-idf scores.

\[
Sim(P, P') = \frac{P \cdot P'}{|P||P'|}
\]

(1)

where \( P \) and \( P' \) represent the tf-idf scores for the terms in two distinct publications.

2.1.2 Exploratory Search

In this section, we provide detailed insights on the exploratory search functionality of our application. As a preliminary step to provide analytics and information about the actual content and topics coverage, all the scientific publications contained in the LAK dataset are previously enriched. The enrichment process was performed using DBpedia Spotlight API\(^3\), where entities, entity types and their respective categories were extracted.

After the enrichment process, we cluster the publications according to entities and its categories found in each document. The publications are clustered in a tree-based structure over the enrichments. Note that, each node of the tree represents a topic in which a publication under this node covers. Thus, the exploratory search is performed through the topics covered by each publication.

The process of linking publications, categories and extra resources is mediated by DBpedia knowledge graph, where we use the dcterms:subject property to match the resources.

Thus, as a result, the exploratory search provides a way to explore resources through the connections between their topics, which facilitates the search for topically related resources. Figure 1 shows the exploratory search.

2.1.3 Semantic Search

Cite4Me provides also a search engine that assists users to find publications semantically related to the query terms. Analogously to explicit semantic analysis (ESA) technique [5], the relatedness score, is computed between the enriched concepts found in the publications' content.

Basically, the semantic search is an adaptation of the free text search presented in the Section 2.1.1. Instead of computing the tf-idf scores for the words in a text, it computes the tf-idf score for the entities contained in a publication. Finally, the ranking of the results is based on the sum of the tf-idf scores of the matching concepts.

Figure 2 illustrates the semantic search functionality. It also generates a tag cloud from matching publications, showing the most prominent terms for a given query. Specifically, the tag cloud based on the results helps the users to have an insight about the topics and may assist in finding related terms previously unknown by them.

2.2 Paper recommendation

Another key feature of our system is the paper recommendation based on semantic relationships extracted from reference datasets. The recommendation is based on a previous work [12, 11], where we exploit the number of paths and the distance (length of a path) between given entities to compute a relatedness score between extracted entities and associated documents. The first step to measure the relatedness between documents is to compute the semantic connectivity score (SCS\(_r\)) of the entities found in each text (see Eq. 2).

\[
SCS_r(a, b) = \sum_{i=1}^{r} \beta^i \cdot |\text{paths}_{(a,b)}^i| \tag{2}
\]

where \(|\text{paths}_{(a,b)}^i|\) is the number of paths between \( a \) and \( b \) of length \( i \) and \( 0 < \beta \leq 1 \) is a positive damping factor. As in [12, 11], we used \( \beta = 0.5 \) as our damping factor. Furthermore, we also constrained the length of a path to \( \tau = 4 \).

Based on the score for entities, we then define the semantic connectivity score (SCS\(_w\)) between two documents \( W_1 \) and \( W_2 \) as follows:

\[
SCS_w(W_1, W_2) = \left( \sum_{e_i \in E_1} \left( SCS_r(e_i, e_2) + \frac{|E_1 \cap E_2|}{2} \right) \right) \frac{1}{|E_1| \cdot |E_2|} \tag{3}
\]

where \( E_i \) is the set of entities associated with \( W_i \), for \( i = 1, 2 \). Note that documents that contain the same entities receive an extra bonus (the second term on the right-hand side of Eq. 3).

Thus, a list of documents pairs is generated and ranked according the score and suggested to the user. Figure 3 illustrates the paper recommendation process computed based on SCS\(_w\).

3. DATASET DISCOVERY AND INTERLINKING

This section briefly describes the datasets used on automatic related data discovery from DataHub\(^9\) and future steps on dataset discovery and interlinking.

3.1 LAK Dataset

The LAK dataset contains the metadata of the papers published in the proceedings of LAK conference 2011-12, a special issue of Learning and Knowledge Analytics: Educational Technology & Society, the proceedings of the International Conference on Educational Data Mining (2008-12) and the Journal of Educational Data Mining (2008-12). In total, 315 descriptions of papers containing detailed information about authors, institutions, conference venues and the full content of the paper were available.

3.2 Data Analysis

The goal of the data analysis procedure is to align the various publications in the LAK dataset based on mutual information, such

\(^3\)http://dbpedia.org/spotlight

\(^9\)http://www.datahub.io
Figure 1: Preview of the exploratory search functionality.

The learning registry: building a foundation for learning resource analytics
Mari Beekwold, John Breck, Jim Noa
We describe our experimentation with the current implementation of a distribution system used to share descriptive and social metadata about learning resources. The Learning Registry, developed and released in a beta version in October 2011, is...

Learning analytics: envisioning a research discipline and a domain of practice
George Siemens
Learning analytics are rapidly being implemented in different educational settings, often without the guidance of a research base. Vendors incorporate analytics practices, models, and algorithms from datamining, business intelligence, and the...

Figure 2: Preview of the semantic search functionality.

Using learning analytics to assess...
Pavel E. Křtěn

Abstract: There is great interest in assessing student learning in unscripted, open-ended environments, but students’ work can evolve in ways that are too subtle or too complex to be detected by the human eye. In this paper, I describe an automated technique to assess, analyze, and visualize students’ learning computer programming. I log multiple (so far at 17) students’ code during a programming assignment, and I employ different quantitative techniques to extract students’ behaviors and categorize them in terms of programming experience. First, I review the literature on educational data mining, learning analytics, computer vision applied to assessment, and emotion detection, discuss the relevance of the work, and describe one case study with a group of graduate engineering students.

Figure 3: An example of paper recommendation based on $SCS_w$. 
as the topics covered by them. This is achieved using well established datasets like DBpedia\(^{10}\) and Freebase\(^1\), where a reference point for the unstructured textual content of publications is created through an enrichment process.

Again, the enrichment process is carried out using DBpedia Spotlight\(^{12}\) \([10]\) and addresses several issues of significant importance. For instance, it offers several advantages such as: (i) identification of (common) named entities, (ii) disambiguation; and (iii) expansion of the limited dataset and resource descriptions with additional background knowledge.

### 3.3 Data Discovery

Our Web application uses as its starting point the instances in the LAK dataset to automatically explore and recommend to users, datasets that covers similar topics. In order to query, detect and interlink related datasets, we chose the DataHub as a data provider. DataHub serves as a collecting point of datasets from various fields and currently it has over 5000 datasets. Note that, from the large number of datasets, only 300 datasets are provided as Linked Open Data. As the latter is the main focus of our work, the analysis and interlinking process is focused for such datasets.

Briefly, the data discovery is performed using CKAN\(^{13}\) data management framework from DataHub, where based on data analysis and user interests (such as topics covered by a publication/resource) related datasets are suggested.

Additionally, we provide to the user a set of resources, amongst other data analytics, that enables the user to harvest and correlate new information from the discovered resources, considering the LAK dataset as a starting point of such discovery.

This approach presents several advantages such as the adoption and the widespread use of Linked Data principles for publishing scientific papers. Nowadays, many conferences make their proceedings and journals freely accessible, hence our approach would take advantage of such open data and offer users topically relevant papers for a particular resource in the LAK dataset.

1. \[http://dbpedia.org\]
2. \[http://www.freebase.com/\]
3. \[http://spotlight.dbpedia.org/\]
4. \[http://www.ckan.org\]

### 4. Evaluation of Data Analysis and Data Discovery

This section presents an overview of the results obtained by analyzing the LAK dataset with respect to the constructed feature set that describes topics covered by individual publications. Moreover, based on the data analysis procedure and shared information, we show that the establishment of links between the different publications within the LAK dataset and from other datasets in DataHub is possible.

In the following subsections, we show the analysis of the LAK dataset and the discovery of relevant datasets and publications.

#### 4.1 Data Analysis

The data analysis of the LAK dataset focuses mostly on assessing the individual publications for their topic coverage. In this manner, we build a connected data graph consisting of the individual publications and items from the feature set. This step is necessary to provide the exploratory search functionality, where based on the established edges between publications and feature set items, we can navigate through the publications or topics of interest. Therefore, the results obtained with respect to the constructed feature set and LAK dataset graph are shown in what follows.

Table 1 shows the top ranked items for each of the feature sets, along with the number of associations an item has with respect to all publications (entity, category and type items). Figure 5 shows the constructed data graph for the LAK dataset.

#### 4.2 Data Discovery

After creating the feature set based on the information provided from reference datasets, we are able to query for relevant datasets in DataHub. Thus, for the top ranked feature set items, the data discovery for relevant resources is considered. Table 2 shows the discovered resources and datasets for the top-10 entity items. Note that, we focus only on bibliographic datasets, since we aim at recommending topically related scientific publications. Due to the lack of bibliographic datasets, we were not able to find related publications for all entities considered. Table 2 summarizes the discovered resources. The dataset names are represented by their acronyms as follows: b3kat - “Bayerische Staatsbibliothek”, hebis - “Hessisches Bibliotheks Informations System” and npg - “Nature Publishing Group - ALL”.

Additionally, from the set of 96 bibliographic datasets available, only a few of them were offered as Linked Data, thus narrowing our search space for relevant resources.

<table>
<thead>
<tr>
<th>Entity</th>
<th>b3kat</th>
<th>hebis</th>
<th>npg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data</td>
<td>14</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>Learning</td>
<td>5</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Data mining</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Algorithm</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Education</td>
<td>17</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Analysis</td>
<td>42</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Student</td>
<td>7</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Knowledge</td>
<td>11</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Methodology</td>
<td>4</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Statistics</td>
<td>7</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 2: Number of discovered resources from the bibliographic group for the top ranked items from the entity feature set, based on the LAK dataset.
Table 1: Top ranked items from the feature set for the LAK Dataset, from the dataset analysis.

<table>
<thead>
<tr>
<th>Entity</th>
<th>Assoc.</th>
<th>Category</th>
<th>Assoc.</th>
<th>Type</th>
<th>Assoc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data</td>
<td>90</td>
<td>Educational_psychology</td>
<td>161</td>
<td>DBpedia:TopicalConcept</td>
<td>150</td>
</tr>
<tr>
<td>Learning</td>
<td>80</td>
<td>Data_analysis</td>
<td>150</td>
<td>Freebase/book</td>
<td>142</td>
</tr>
<tr>
<td>Data_mining</td>
<td>67</td>
<td>Learning</td>
<td>139</td>
<td>Freebase/book/book_subject</td>
<td>142</td>
</tr>
<tr>
<td>Algorithm</td>
<td>50</td>
<td>Scientific_method</td>
<td>137</td>
<td>Freebase/media_common</td>
<td>138</td>
</tr>
<tr>
<td>Education</td>
<td>49</td>
<td>Neuropsychological_assessment</td>
<td>136</td>
<td>Freebase/media_common/quotatoin_subject</td>
<td>136</td>
</tr>
<tr>
<td>Analysis</td>
<td>48</td>
<td>Greek_loanwords</td>
<td>135</td>
<td>Freebase/computer</td>
<td>125</td>
</tr>
<tr>
<td>Student</td>
<td>46</td>
<td>Data</td>
<td>131</td>
<td>Freebase/education</td>
<td>122</td>
</tr>
<tr>
<td>Knowledge</td>
<td>46</td>
<td>Evaluation_methods</td>
<td>129</td>
<td>Freebase/education/field_of_study</td>
<td>120</td>
</tr>
<tr>
<td>Methodology</td>
<td>42</td>
<td>Computer_data</td>
<td>126</td>
<td>Freebase/computer/software_genre</td>
<td>120</td>
</tr>
<tr>
<td>Statistics</td>
<td>41</td>
<td>Research_methods</td>
<td>124</td>
<td>Freebase/internet</td>
<td>118</td>
</tr>
<tr>
<td>System</td>
<td>37</td>
<td>Systems_science</td>
<td>118</td>
<td>Freebase/internet/website_category</td>
<td>118</td>
</tr>
<tr>
<td>Scientific_modelling</td>
<td>37</td>
<td>Formal_sciences</td>
<td>108</td>
<td>Freebase/award</td>
<td>114</td>
</tr>
<tr>
<td>Prediction</td>
<td>36</td>
<td>Data_management</td>
<td>108</td>
<td>Freebase/media_common/media_genre</td>
<td>105</td>
</tr>
<tr>
<td>Data_set</td>
<td>36</td>
<td>Cognitive_science</td>
<td>107</td>
<td>Freebase/organization</td>
<td>103</td>
</tr>
<tr>
<td>Statistical_classification</td>
<td>30</td>
<td>Statistical_terminology</td>
<td>107</td>
<td>Freebase/award/award/discipline</td>
<td>103</td>
</tr>
<tr>
<td>Evaluation</td>
<td>29</td>
<td>Developmental_psychology</td>
<td>101</td>
<td>Freebase/business</td>
<td>102</td>
</tr>
<tr>
<td>Standard_deviation</td>
<td>29</td>
<td>Intelligence</td>
<td>93</td>
<td>Freebase/organization/organization_sector</td>
<td>99</td>
</tr>
<tr>
<td>Probability</td>
<td>28</td>
<td>Data_mining</td>
<td>91</td>
<td>Freebase/people</td>
<td>99</td>
</tr>
<tr>
<td>Behavior</td>
<td>26</td>
<td>Critical_thinking</td>
<td>87</td>
<td>Freebase/film</td>
<td>94</td>
</tr>
<tr>
<td>Interaction</td>
<td>24</td>
<td>Thought</td>
<td>84</td>
<td>Freebase/book/periodical_subject</td>
<td>93</td>
</tr>
</tbody>
</table>

Figure 5: Topic coverage of LAK data graph for the individual resources.
5. RELATED WORK

Cobo et al. [3] presents an analysis of student participation in online discussion forums using an agglomerative hierarchical clustering algorithm, and explore the profiles to find relevant activity patterns and detect different student profiles. Barber et al. [1] uses a predictive analytic model to prevent students from failing in courses. They analyze several variables, such as grades, age, attendance and others, that can impede the student learning. Kahn et al. [7] present a long-term study using hierarchical cluster analysis, t-tests and Pearson correlation that identified seven behavior patterns of learners in online discussion forums based on their access. García-Solórzano et al. [6] introduce a new educational monitoring tool that helps tutors to monitor the development of the students. Unlike traditional monitoring systems, they propose a facetted browser visualization tool to facilitate the analysis of the student progress. Glass [8] provides a versatile visualization tool to enable the creation of additional visualizations of data collections.

Essa et al. [4] utilize predictive models to identify learners academically at-risk. They present the problem with an interesting analogy to the patient-doctor workflow, where first they identify the problem, analyze the situation and then prescribe courses that are indicated to help the student to succeed. Siadat et al. [13] present the Learn-B environment, a hub system that captures information about the users usage in different softwares and learning activities in their workplace and present to the user feedback to support future decisions, planning and accompanies them in the learning process.

In the same way, McAuley et al. [9] propose a visual analytics to support organizational learning in online communities. They present their analysis through an adjacency matrix and an adjustable timeline that show the communication-actions of the users and is able to organize it into temporal patterns. Bramucci et al. [2] presents Sherpa an academic recommendation system to support students on making decisions. For instance, using the learner profiles they recommend courses or make interventions in case that students are at-risk.

In the related work, we showed how different perspectives and the necessity of new tools and methods to make data available and help decision-makers.

6. CONCLUSION

In this paper we presented the main features of the Cite4Me Web application. Cite4Me makes use of several data sources to provide information for users interested on scientific publications and its applications.

Additionally, we provided a general framework on data discovery and correlated resources based on a constructed feature set, consisting of items extracted from reference datasets. It made possible for users, to search and relate resources from a dataset with other resources offered as Linked Data.

For more information about the Cite4Me Web application refer to http://www.cite4me.com.

7. REFERENCES


Visualizing the LAK/EDM Literature Using Combined Concept and Rhetorical Sentence Extraction

Davide Taibi\(^1\), Ágnes Sándor\(^2\), Duygu Simsek\(^3\), Simon Buckingham Shum\(^3\), Anna DeLiddo\(^3\), Rebecca Ferguson\(^3\)

\(^1\)Institute for Educational Technologies
Italian National Research Council
Via Ugo La Malfa 153
90146 Palermo, Italy
davide.taibi@ild.cnr.it

\(^2\)Parsing & Semantics Group
Xerox Research Centre Europe
6 Chemin de Maupertuis
F-38240 Meylan, France
agnes.sandor@xrce.xerox.com

\(^3\)The Open University
Knowledge Media Institute & Institute of Educational Technology
Milton Keynes, MK7 6AA, UK
firstname.lastname@open.ac.uk

ABSTRACT

Scientific communication demands more than the mere listing of empirical findings or assertion of beliefs. Arguments must be constructed to motivate problems, expose weaknesses, justify higher-order concepts, and support claims to be advancing the field. Researchers learn to signal clearly in their writing when they are making such moves, and the progress of natural language processing technology has made it possible to combine conventional concept extraction with rhetorical analysis that detects these moves. To demonstrate the potential of this technology, this short paper documents preliminary analyses of the dataset published by the Society for Learning Analytics, comprising the full texts from primary conferences and journals in Learning Analytics and Knowledge (LAK) and Educational Data Mining (EDM). We document the steps taken to analyse the papers thematically using Edge Betweenness Clustering, combined with sentence extraction using the Xerox Incremental Parser’s rhetorical analysis, which detects the linguistic forms used by authors to signal argumentative discourse moves. Initial results indicate that the refined subset derived from more complex concept extraction and rhetorically significant sentences, yields additional relevant clusters. Finally, we illustrate how the results of this analysis can be rendered as a visual analytics dashboard.

Categories and Subject Descriptors
K.3.1 [Computers and Education]: Computer Uses in Education

General Terms
Design

Keywords
Learning Analytics, Corpus Analysis, Scientific Rhetoric, Visualization, Network Analysis, Natural Language Processing

1. INTRODUCTION AND MOTIVATION

Our overall aims are to provide users automatically with suggestions about similar papers, about connections between papers, and to present these similarities and connections in ways that are both meaningful and searchable.

In order to achieve this, we integrated three different approaches to linking and analysing a specific dataset of scientific papers (see section 2). These approaches were:

1. network analysis
2. rhetorical analysis
3. visualization of the results

Network analysis yields sets of related papers based on statistical corpus processing (Section 3). In order to improve the precision of information about the content of the connections among the papers, we carried out semantic and rhetorical analysis (Section 4). On the one hand, we extracted similar concepts in order to provide topical similarity indicators (Section 4.1) and, on the other hand, we extracted salient sentences that indicate the main research topics of these papers (Section 4.2). We repeated the statistical analysis of this reduced list of concepts, and of the reduced list of salient sentences. At the end of this paper, we present the design and implementation of the first prototype of an analytics dashboard (Section 5), which is designed to summarize results of the socio-semantic-rhetorical analysis in a way that users will find both meaningful and easy to explore.

2. THE LAK DATASET

We selected the LAK Dataset\(^1\) published by the Society for Learning Analytics Research (SoLAR\(^2\)), which provides machine-readable plain-text versions of the Learning Analytics and Knowledge (LAK) conference proceedings and a journal special issue related to learning analytics, and of the Educational Data Mining (EDM) conferences and journal.

The corpus was extracted using the SPARQL endpoint of the LAK dataset. The corpus comprised the following:

- 24 papers presented at the LAK2011 conference
- 42 papers presented at the LAK2012 conference
- 10 papers from the journal of Educational Technology and Society special issue on learning analytics
- 31 papers presented at the EDM2008 conference
- 32 papers presented at the EDM2009 conference
- 64 papers presented at the EDM2010 conference
- 61 papers presented at the EDM2011 conference
- 52 papers presented at the EDM2012 conference

For each resource, the title, description and keywords properties were used to feed the data mining processes employed in our analysis. At the end of this initial process, a relational database was used to store 305 papers, 599 authors, 448 distinct keywords. After this preliminary phase the entire LAK Dataset

---

\(^1\) LAK Dataset: [http://www.solarresearch.org/resources/lak-dataset](http://www.solarresearch.org/resources/lak-dataset)

Published by SoLAR and made available to the LAK Data challenge of the 3\(^{rd}\) International Conference on Learning Analytics and Knowledge (http://lakconference.org)

\(^2\) [http://www.solarresearch.org](http://www.solarresearch.org)
was analyzed by using the Xerox Incremental Parser (XIP) [1] for concept extraction and rhetorical analysis, a total of 305 papers, from which XIP extracted 7,847 sentences and 40,163 concepts.

3. STATISTICAL ANALYSIS
A preliminary analysis reported the most-used keywords, the most frequently occurring authors and the most-referenced papers. A second phase of analysis was then carried out using the data-mining tool, RapidMiner [2].

3.1 Statistical Data from RapidMiner
A three-step process was developed in order to analyze the corpus using the data-mining tool, RapidMiner:

- **Process documents from file:** this module generates word vectors from the text files.
- **Select attributes:** This allows users to select the attributes to be considered by the analysis. In our case, a threshold was set in order to eliminate less important elements in the word vectors.
- **Data to similarity:** This module was used to calculate a similarity index for the conference papers based on Cosine similarity.

The first block ‘Process Documents from file’ is made up of the following steps:

- **Tokenize:** This operator splits the text of a document into a sequence of tokens.
- **Replace token:** This operator is used to replace tokens, for instance in cases where words are misspelled.
- **Filter tokens (by length):** This operator filters tokens based on their length. In our case, all the words with fewer than three characters were removed.
- **Filter stopwords (English):** This operator filters English stopwords from a document by removing every token that is the same as a stopword from the built-in stopword list.
- **Stem (Snowball):** This operator stems words by applying stemming using the Snowball tool.

At the end of the main process, the ‘Data to Similarity’ step returns two results:

a) The list of the most relevant words (stemmed version) used in the entire corpus
b) The measured similarity index between the papers that make up the corpus.

We employed the similarity relationships between papers to build a network of papers. In this network each node represents a paper, and an edge between two paper is created if the similarity value of a pair of papers overcome a threshold of 0.3.

3.2 Analysing the Network of Papers
The network of papers was then analysed with the yEd tool [4] in order to extract clusters of documents using the algorithm for natural clusters “based on Edge Betweenness Clustering proposed by Girvan and Newman” [3]. This algorithm has been successfully used in Network Analysis to study communities and their aggregations [4]. The yEd tool allows users to balance quality and speed of the cluster algorithm by the use of a slider. When the quality is set at the highest value, the Girvan and Newman algorithm is used in its normal form. At the opposite end, the lowest quality value produces the fastest running time.

In this case it executes a local betweenness calculation following Gregory’s algorithm [5]. When a mid value is chosen for quality and speed, the fast betweenness approximation of Brandes and Pich [6] is applied. In this case, less accurate clustering is balanced by a lower execution time.

The clusters created with yEd have the following properties:

- each node (paper) is a member of exactly one cluster
- each node shares many edges with other members of its cluster, where edges represent the connection between a pair of papers if their similarity values is more than a threshold value (0.3 in our experiment)
- each node shares few or no edges with nodes of other clusters

Figure 1 shows a visualization of the primary clusters. Some of the clusters did seem to have thematic coherence, while others were harder to label:

- Cluster 1: collaborative, learning, social
- Cluster 2: skills, model, slip, guess, parameters
- Cluster 3: causality, variables, model, construct
- Cluster 4: question, fit, grain, school, skill
- Cluster 5: translating, sentences, grinder, corpus

Figure 1: Results of initial LAK paper clustering analysis
The complete list of the papers belonging to the clusters has been reported in the web page [3] associated to this work.

This analysis was word-driven and not concept-driven. The next step was to try and refine this by distilling (1) a richer set of concepts, and (2) a more salient subset of sentences.

4. SEMANTIC ANALYSIS
In order to go beyond full-text statistical analysis and find connections between papers at the level of the claims they make, we processed the corpus using the Xerox Incremental Parser

---

3 http://snowball.tartarus.org
4 http://www.yworks.com
5 http://www.pa.itd.cnr.it/lak-data-challenge.html
A basic observation concerns the distribution of the pairs of similar papers yielded by the three methods. According to the expectations, the most similarity pairs have been yielded by taking into account the full text only in both the LAK and the EDM collection. There are considerable overlaps among the three methods, and there are cases when just one method yields similarity pairs. In subsequent evaluations we aim at evaluating these various cases. As a first step towards a more complete evaluation, we have selected some pairs of papers and checked their similarity according to some independent similarity indicators. We have found that our statistical method is coherent with independent similarity indicators in case of high similarity scores and that in these cases, similarity is found with and without XIP-extracted text. This indicates the validity of our statistical method in these cases for finding related papers. In the case where no independent similarity indicator could be found, but we do have XIP-based similarity pairs, we looked for related key claims or findings in the pairs of papers. In the cases where the similarity score between the two papers was high we did find such interesting related claims in the two papers. However, in cases where the similarity measure is low, we did not find any related claims. This indicates that we might want to define a threshold score. The details of the preliminary tests are reported in the web page.

5. XIP DASHBOARD

The XIP Dashboard was designed to provide visual analytics from XIP output in order to help readers assess the current state of the art in terms of trends, patterns, gaps and connections in the LAK and EDM literature. The dashboard also draws attention to candidate patterns of potential significance within the dataset:

- the occurrence of domain concepts in different metadiscourse contexts (e.g. effective tutoring dialogue in sentences classified as contrast).
- trends over time (e.g. the development of an idea)
- trends within and differences between research communities, as reflected in their publications.

5.1 Implementation

All the papers in the LAK dataset were analyzed using XIP. The output files of the XIP analysis, one per paper, were then imported into a MySQL database, and the user interface was implemented using PHP and JavaScript, making use of Google Chart Tools for the interactive visualizations.\(^6\)

5.2 User Interface

The dashboard consists of three sections, each showing different analytical results in different types of chart.

Section one of the dashboard shows two line charts, representing the LAK and the EDM conferences respectively. Each line chart shows the distribution of the number of salient sentences over time and by rhetorical marker type (see Figure 2 for a list of the types of rhetorical markers). Each coloured line in these line charts indicates how many sentences of a specific rhetorical type were extracted, and how this number changed by year (Figure 3 shows the line chart for the EDM conference).

\(^7\) The related claims have been searched by reading the pairs of sentences. Our long-term goal is to provide the related claims automatically.

\(^8\) https://developers.google.com/chart
Figure 3: Rhetorical sentences graphed by year, for EDM

The second section of the dashboard (Figure 4) allows users to select a combination of the extracted concepts, in order to visualize the occurrence of these concepts in papers within any or all research communities represented in the corpus— that is to say across the whole LAK dataset (EDM plus LAK conference).

Figure 4: Number of papers with rhetorically extracted sentences containing user-selected concepts

The third dashboard section consists of a bubble chart that displays the occurrence of papers within the entire dataset, filtered by user-selected concepts (Figure 5). This visualization can be restricted to display just the LAK or the EDM conference. In Figure 5, each bubble represents a concept that has been selected by the user. This is associated with a specific number of papers and sentences in which that concept has been detected. The colour saturation of each bubble (expressed by the color spectrum shown at the top) represents the ‘density’ of the chosen concept as defined by the number of XIP-extracted sentences in which the concept occurs. The darker the colour, the greater the density.

Figure 5: Concept ‘density’ within XIP sentences, by year and number of papers

When a concept bubble is selected (Figure 6), a pie chart pops up representing the relative distribution of the rhetorical types for that bubble (that is to say for that concept, and across the papers and sentences in which the concept has been detected).

6. SUMMARY

This short paper has summarised an approach to conducting ‘analytics on Learning Analytics’. The LAK Dataset comprising LAK and EDM literature has been analyzed in order to identify clusters of papers dealing with similar topics (conceptual clustering), and in order to identify key contributions of papers in terms of the claims authors make, as signalled by rhetorical patterns. Our preliminary tests are promising, but more thorough testing is needed to validate the method. Finally, we showed how the results of this analysis are beginning to be visualized using an analytics dashboard. All the secondary datasets produced have been published as open data, for further research.

Figure 6: Distribution of rhetorical types in XIP-classified sentences within a selected concept bubble

In the longer term, the aim of this research is to provide users with automatic suggestions about similar papers and about connections between papers, and to present these similarities and connections in ways that are both meaningful and searchable for the users. Future steps will validate the outputs from these analyses with researchers, and test the usability of the dashboard with different end-users (e.g. researchers, educators, students).

7. REFERENCES


Ontology Learning to Analyze Research Trends in Learning Analytics Publications

Amal Zouaq  
Department of Mathematics and Computer Science  
Royal Military College of Canada  
Kingston, ON, Canada  
+1 613 541 6000, Ext. 6478  
amal.zouaq@rmc.ca

Srećko Joksimović  
School of Interactive Arts and Technologies  
Simon Fraser University  
Surrey, BC, Canada  
+1 778 782 7474  
sjoksimo@sfu.ca

Dragan Gašević  
School of Computing and Information Systems  
Athabasca University  
Athabasca, AB, Canada  
+1 604 569 8515  
dgasevic@acm.org

ABSTRACT
In this paper, we show how ontology learning tools can be used to reveal (i) the central research topics that are tackled in the published literature on learning analytics and educational data mining; and (ii) relationships between these research topics and iii) (dis)similarities between learning analytics and educational data mining.

Categories and Subject Descriptors
I.2.7 [Artificial Intelligence]: Natural Language Processing; G.2.2 [Discrete Mathematics]: Graph Theory

General Terms
Algorithms, Measurement, Experimentation

Keywords
Ontology learning, deep parsing, filtering, information retrieval, ranking algorithms, graph theoretic statistics

1. INTRODUCTION
Learning analytics is a new research discipline. Although it attracted a considerable amount of attention in educational research and practice, debate is still very active about the scope of the discipline. The definition of learning analytics offered by the Society for Learning Analytics Research [7], which is commonly used in the literature to date, gives a general framework for the main tasks learning analytics are about. However, given the youth of the discipline, there are generally two open questions:
- What are the central research topics that are tackled in the published literature?
- What are the relationships between the central research topics?
- What are similarities and differences between learning analytics and educational data mining?

To address the above questions, we aimed to analyze systematically textual content available in the LAK Challenge data set. In particular, we used a state-of-the-art ontology learning tool, OntoCmaps, that enabled the automatic (i) parsing of textual content, (ii) creation of conceptual maps based on the extracted concepts and relationships, and (iii) filtering/ranking of the most important concepts and relationships based on measures of information retrieval, graph theory, and voting theory. The concept extraction and their filtering/ranking was done (i) for each edition of the two conferences and the journal special issue (from the LAK 2013 Challenge dataset) individually (i.e., LAK 2011-2012, EDM 2008-2013, and LAK ET&S special issue) to see the emerging trends through the years; and (ii) by creating two subsets – one for the papers presented at the LAK conference editions and another one for the papers presented at the EDM conference editions – in order to compare the two conferences based on concepts and relationships gauged as most important. We also performed analysis based on (a) paper abstracts only and (b) main body of text of the papers.

In this short report, we first describe the data analysis pipeline. This is followed by a very brief discussion of a small fragment of the results we obtained in our analysis. The complete results in the CSV format are available at [8].

2. DATA ANALYSIS PIPELINE
The data analysis relies on our ontology learning tool, OntoCmaps[10]. Ontology learning from text is a multi-layer knowledge extraction task that targets the following components:

Terms and concepts: The first step consists in identifying candidate expressions in texts. These expressions are then ranked using some kind of measure (statistical metrics, graph-based metrics, etc.) to extract those that are relevant for the domain. These filtered relevant expressions are then considered “concepts” in the ontology learning community.

Taxonomy: This step identifies “is-a” links in texts, generally using patterns indicating a taxonomical link in text such as Hearst’s patterns[11], or using the inner structure of multiword expressions. For example, a “carnivorous plant” can be considered a “plant” just by looking at the syntactic structure “Adjective noun” of the expression.

Conceptual relationships: This step uses various techniques (patterns, machine learning, etc.) to identify any kind of transversal relations, with a domain and range.

Axioms: Finally, axioms here mean defined classes, or rules from texts.

OntoCmaps requires a domain corpus as input. As such, LAK and EDM proceedings (the LAK dataset [13]) were an appropriate set of texts to test the ontology learning process. OntoCmaps relies on three main phases to learn a domain ontology: 1) the extraction phase that performs a deep semantic analysis based on dependency patterns; 2) the integration phase that builds concept maps, which are composed of terms and labeled relationships, and uses basic disambiguation techniques. These concept maps form a graph; and finally 3) the filtering phase where various metrics rank the items (terms and relationships) in concept maps.

2.1 The Extraction Phase
In the extraction phase, OntoCmaps builds on a hierarchy of syntactic patterns. Each pattern describes a set of syntactic rela-
tionships that permit the extraction of a “semantic representation”. OntoCmaps does not rely on any predefined domain knowledge. It uses two NLP tools to obtain the syntactic representations: the Stanford Parser along with its dependency module [2] and the Stanford parts-of-speech (POS) Tagger [6]. Given a sentence, the Stanford parser generates syntactic dependency relations between each pair of related words of a sentence. The POS Tagger identifies words’ parts-of-speech. Based on these two inputs, OntoCmaps creates a pattern syntactic format that enriches words in each dependency relation with their parts-of-speech. This enriched representation is then used as input to a pattern recognition task. A recognized pattern fires a rule that applies various transformations on the syntactic representation to obtain a “semantic representation”, in the form of expressions, triples or sets of triples. The patterns are divided into conceptual patterns and hierarchical patterns. Hierarchical patterns concentrate on the extraction of taxonomical links, following the work of [11], but based on the dependency formalism. Conceptual patterns identify the main structures of the language that can be transformed into triples useful for the extraction of conceptual relations. They are organized into a hierarchy from most-detailed patterns (containing the biggest number of dependency relationships) to least detailed. The extraction phase targets deeper levels of the hierarchy first to avoid extracting too abstract or incomplete representations. For instance, if the pattern “nsubj-dobj-xcomp” exists in text, the extractor should fire it instead of firing one of its higher-level counterparts “nsubj-dobj” and “nsubj-xcomp” which contain only a subset of the syntactic relationships of interest. If a pattern is instantiated, then all its parents in the hierarchy are disregarded.

2.2 The Integration Phase
In this integration phase, all the extracted relationships are gathered into concept maps. Some basic term disambiguation tasks are performed at this level mainly: i) lemmatization which considers singular, plural and other forms of the same terms or relationships as referring to a single concept or relationship; ii) basic synonym detection based on abbreviation relations that are generated by the Stanford parser and iii) a kind of co-reference resolution phase that is built in some of the patterns, and that allows for the creation of semantic links between terms in a sentence, even if not direct dependency links existed in the original dependency representation. For example, in the sentence: *carnivorous plants are organisms which eat insects*, the co-reference resolution creates a relation “eat” between the term “carnivorous plants” and the term “insects” while the grammatical representation links the term “plants” to the term “insects”.

All these operations result in concept maps around various terms. For example, if there were a number of statements around the term “*carnivorous plants*” in texts, it is likely that a concept map around “carnivorous plants” will be created. This process is repeated for all identified terms and relationships and results in an aggregation of concept maps through links between various concept maps, thus constituting a graph, with terms representing nodes, and relationships representing edges.

2.3 The Filtering Phase
The third and last phase for learning the domain ontology is the filtering phase, which aims at ranking the items in concept maps (domain terms, taxonomical links, and conceptual links).

2.3.1 Concept Filtering
A number of metrics from graph theory and from information retrieval are used to identify relevant terms. Graph-based metrics were computed using the JUNG framework [3]. These metrics include:
- The Degree centrality of a node which identifies the number of edges from and to a given node.
- The Betweenness centrality, which assigns each node a value that is derived from the number of shortest paths that pass through it;
- The HITS algorithm which ranks nodes according to the importance of hubs and authorities [5]. This resulted in two measures HITS-Hubs and HITS-Authority;
- The PageRank of a node [1];
- We also computed standard information retrieval metrics, mainly term frequency (TF) and TF-IDF.

Finally, using the graph-based metrics, we defined a number of voting schemes with the aim of improving the precision of filtering. All the VS relied on three metrics that were identified as being among the best metrics in previous experiments [10][11]: Degree, Betweenness and HITS-Hubs. The VS include:
- The majority voting scheme, which recognizes a term as an important one if it is chosen by at least k metrics out of n with k>n/2.
- Borda Count Voting Scheme: This method assigns a “rank” to each candidate. A candidate who is ranked first receive n points (n=size of the domain terms to be ranked), second n-1, third n-2 and so on. The “score” of a term for all metrics is equal to the sum of the points obtained by the term in each metric.
- Nauru Voting Scheme: The Nauru voting scheme is based on the sum of the inverted rank of each term in each metric. It is used to put more emphasis on higher ranks.

Table 1 shows the top ranked concepts based on the majority voting scheme. All the base metrics (Betweenness, PageRank, Degree, etc.) and voting schemes have been computed and can be found at [8]. The Web site [8] also features a visualization of the extracted data based on the obtained concept maps. The visualization is performed per venue (EDM/LAK/ETS-SI), per corpus (only abstracts or main texts) and per year (2008-2012).

2.3.2 Relationship Filtering
Similarly, a number of metrics were used to identify important relationships.

The first metric consists of all the relationships that occur between important terms (determined through the voting schemes) as important relationships. This constitutes our voting schemes. All the base metrics (Betweenness, PageRank, Degree, etc.) and voting schemes have been computed and can be found at [8]. The Web site [8] also features a visualization of the extracted data based on the obtained concept maps. The visualization is performed per venue (EDM/LAK/ETS-SI), per corpus (only abstracts or main texts) and per year (2008-2012).

Table 2 shows an excerpt of the top ranked relationships based on the Dice coefficient [9], a standard measure for semantic relatedness.

The second measure ranks relationships based on Edge Betweenness centrality, which is a measure of the importance of edges based on the number of shortest paths which contain them.

The third measure is based on assigning frequencies of co-occurrence weights based on the majority voting scheme for concepts.

Table 2 shows an excerpt of the top ranked relationships based on the Dice coefficient [9], a standard measure for semantic relatedness.

Table 2 shows an excerpt of the top ranked relationships based on the Dice coefficient [9], a standard measure for semantic relatedness.
Table 1. Top ranked concepts based on the majority voting scheme extracted the subsets of the LAK 2013 Challenge dataset

<table>
<thead>
<tr>
<th>Concept</th>
<th>LAK (abstracts)</th>
<th>LAK (paper body)</th>
<th>EDM (abstracts)</th>
<th>EDM (paper body)</th>
</tr>
</thead>
<tbody>
<tr>
<td>student</td>
<td>(0.50)</td>
<td>student (0.75)</td>
<td>student (0.75)</td>
<td>student (0.75)</td>
</tr>
<tr>
<td>datum</td>
<td>(0.45)</td>
<td>datum (0.20)</td>
<td>model (0.38)</td>
<td>model (0.23)</td>
</tr>
<tr>
<td>informal_learn</td>
<td>(0.31)</td>
<td>learner (0.15)</td>
<td>datum (0.37)</td>
<td>datum (0.19)</td>
</tr>
<tr>
<td>learn</td>
<td>(0.31)</td>
<td>course (0.15)</td>
<td>method (0.19)</td>
<td>skill (0.09)</td>
</tr>
<tr>
<td>teacher</td>
<td>(0.29)</td>
<td>analysis (0.12)</td>
<td>paper (0.16)</td>
<td>problem (0.08)</td>
</tr>
<tr>
<td>model</td>
<td>(0.27)</td>
<td>activity (0.11)</td>
<td>system (0.13)</td>
<td>result (0.06)</td>
</tr>
<tr>
<td>learning_analytics</td>
<td>(0.26)</td>
<td>user (0.10)</td>
<td>result (0.12)</td>
<td>method (0.06)</td>
</tr>
<tr>
<td>learner</td>
<td>(0.25)</td>
<td>tool (0.10)</td>
<td>approach (0.11)</td>
<td>parameter (0.05)</td>
</tr>
<tr>
<td>social_factor</td>
<td>(0.21)</td>
<td>learn (0.09)</td>
<td>skill (0.08)</td>
<td>question (0.05)</td>
</tr>
<tr>
<td>social_learn</td>
<td>(0.19)</td>
<td>analytics (0.07)</td>
<td>analysis (0.07)</td>
<td>performance (0.05)</td>
</tr>
<tr>
<td>effective_learn</td>
<td>(0.19)</td>
<td>group (0.07)</td>
<td>intelligent_tutoring_system (0.07)</td>
<td>system (0.05)</td>
</tr>
<tr>
<td>group_learn</td>
<td>(0.17)</td>
<td>system (0.07)</td>
<td>behavior (0.07)</td>
<td>approach (0.04)</td>
</tr>
<tr>
<td>knowledge_professional</td>
<td>(0.17)</td>
<td>teacher (0.06)</td>
<td>tool (0.07)</td>
<td>example (0.04)</td>
</tr>
<tr>
<td>Lak (0.17)</td>
<td>instructor (0.06)</td>
<td>work (0.06)</td>
<td>feature (0.04)</td>
<td></td>
</tr>
<tr>
<td>knowledge (0.17)</td>
<td>network (0.06)</td>
<td>Researcher (0.06)</td>
<td>item (0.04)</td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Top ranked relationships based on the majority voting scheme extracted the subsets of the LAK 2013 Challenge dataset.

Each cell in the table contains a concept-relationship-concept triplet

<table>
<thead>
<tr>
<th>Relationship</th>
<th>LAK (abstracts)</th>
<th>LAK (paper body)</th>
<th>EDM (abstracts)</th>
<th>EDM (paper body)</th>
</tr>
</thead>
<tbody>
<tr>
<td>learner–build–knowledge (1)</td>
<td>course–being recorded as well as to-student (1)</td>
<td>datum–mining–method (1)</td>
<td>model–fit–student (1)</td>
<td></td>
</tr>
<tr>
<td>datum–obtained from–learner (0.81)</td>
<td>datum–break ability to educate student (0.60)</td>
<td>method–linguistics in–paper (0.95)</td>
<td>datum–are collected far from–student (0.96)</td>
<td></td>
</tr>
<tr>
<td>learning_analytics–important step for–teachers_of_tomorrow (0.78)</td>
<td>system–addresses individually–student (0.45)</td>
<td>model–are trained over–datum (0.70)</td>
<td>skill–will have been covered by–student (0.67)</td>
<td></td>
</tr>
<tr>
<td>teachers_of_tomorrow–is a–teacher (0.77)</td>
<td>analysis–have since been moved as–student (0.37)</td>
<td>system–provides–student (0.61)</td>
<td>problem–assign for–student (0.67)</td>
<td></td>
</tr>
<tr>
<td>tool–incorporate functionality to access–datum (0.65)</td>
<td>network–impacting–student (0.31)</td>
<td>student–are represented by–model (0.56)</td>
<td>example–parameterization by–student (0.63)</td>
<td></td>
</tr>
<tr>
<td>model–can be used to inform–student (0.64)</td>
<td>process–finally should promote reflection on–instructor (0.29)</td>
<td>model–can detect–student (0.50)</td>
<td>question–were based–student (0.62)</td>
<td></td>
</tr>
<tr>
<td>datum–obtained from–instructor (0.62)</td>
<td>tool–identify–student (0.27)</td>
<td>datum–derived from–student (0.43)</td>
<td>student–provides useful evidence to–model (0.60)</td>
<td></td>
</tr>
<tr>
<td>learner–generating–datum (0.58)</td>
<td>datum–may be presented to–learner (0.25)</td>
<td>goal–has been investigated by–researcher (0.42)</td>
<td>step–requires–student (0.57)</td>
<td></td>
</tr>
<tr>
<td>student–accessing–online_discussion_forum (0.56)</td>
<td>activity–conducted by–user (0.25)</td>
<td>tutoring_system–is a–system (0.40)</td>
<td>performance–dependent upon–student (0.56)</td>
<td></td>
</tr>
<tr>
<td>model–can be used to inform–teacher (0.51)</td>
<td>group–will contain–student (0.25)</td>
<td>student–study with–intelligent_tutoring_system (0.39)</td>
<td>accuracy–varies across–student (0.48)</td>
<td></td>
</tr>
<tr>
<td>student–flock to–online_service (0.48)</td>
<td>environment–capture–datum (0.24)</td>
<td>skill–studied in–tutoring_system (0.38)</td>
<td>student–is guessing–result (0.48)</td>
<td></td>
</tr>
<tr>
<td>datum–are combined to calculate–likelihood_of_student (0.45)</td>
<td>model–highly accurate on–student (0.22)</td>
<td>intelligent_tutoring_system–are informed by–datum (0.32)</td>
<td>student–collect–datum (0.45)</td>
<td></td>
</tr>
<tr>
<td>instructor–guide–student (0.39)</td>
<td>average–miss–student (0.21)</td>
<td>analysis–reveals–unexpected_result (0.30)</td>
<td>word–uttered by–student (0.44)</td>
<td></td>
</tr>
<tr>
<td>learn–integral to–success_of_community (0.37)</td>
<td>role–are imposed on–student (0.21)</td>
<td>unexpected_result–is a–result (0.30)</td>
<td>datum–were used to build–model (0.44)</td>
<td></td>
</tr>
<tr>
<td>likelihood_of_student–is related to–student (0.36)</td>
<td>information–useful for–student (0.20)</td>
<td>collaborative–learning–interactions_of_student (0.29)</td>
<td>skill–are included in–model (0.41)</td>
<td></td>
</tr>
</tbody>
</table>
We can also notice that we were not always successful in extracting meaningful relationships labels from this corpus. One possible explanation is the type of texts (publications) and the amount of noise in these texts. In fact, OntoCmaps is made to run on clean plain sentences that describe a domain of interest and define it. Parts of research papers such as figure captions, formulas, and references represent noise for OntoCmaps. Additional cleaning of the input texts would be necessary. However, even when the labels were not meaningful, the existence of a link between two concepts (unlabeled relationship) was shedding some light on the domain (see Section 3).

### 3. FINDINGS

In this section, we present only results of the 15-top ranked concepts and relationships according to the Majority Voting Scheme (Betweenness, Degree, and Hits-Hub) as shown in Tables 1-2 (N.B. As can be noticed in the tables, the majority of the terms are lemmatized, that is, we show only their lemma or root. For example, informal_learn for informal learning or datum for data. In few cases, such as learning_analytics, the lemmatizer returned the expression itself). First, we could not possible include all the results of all the metrics we calculated in our experiment (those results are available at [8]). Second, we selected the metrics which were proven to be most accurate in our previous research [10], [11]. Finally, it should be noted that the purpose of our experiment here was not to evaluate the effectiveness of individual metrics, but rather to experiment if ontology learning technology can shed some light on the questions posed in the introduction of relevance to the LAK 2013 Data Challenge.

Concepts reported in Table 1 reveal that papers of both the LAK and EDM conferences have students, data and models as shared concepts. However, it is clear that LAK papers also focus on teachers/instructors, informal learning, and social, networked, and group learning. On the other hand, EDM papers focus on (data mining) methods and approaches, intelligent tutoring systems, features (extraction), and various types of parameters.

Relationships reported in Table 2 further corroborate the observation that the LAK papers are more focused on teachers in order to empower them with learning analytics and to help them guide students. Moreover, there is an emphasis on (promoting) reflection of both students and instructors. Various aspects of social learning such as role playing and impact of communities appear to be highly popular topics in the LAK papers. On the other hand, EDM papers are much more focused on intelligent tutoring systems, accuracy of different types of (predictive) models, and revealing unexpected patterns. Certainly, focus on data is shared by both the LAK and EDM communities, but LAK also seems to be focused on data collected by and for instructors, not only for students. This probably indicates a trend that the LAK community has so far acknowledged the role of instructors in the learning process and aimed at supporting them as much as learners. The EDM community has however focused more on measuring and predicting specific types of skills. This is consistent with their focus on intelligent tutoring systems in which automated assessment of learners’ skills is of paramount importance.

Finally, we were also able to visualize the extracted conceptual graphs. In Figure 1, we show the relationships of concept learning_analytics as extracted from the abstracts of the papers presented at
the LAK conference. This figure further corroborates earlier observations by indicating that learning analytics is an integral part of teaching profession, is an important step for teachers of tomorrow and learners, and offers a new approach. This figure reveals also the nature of learning analytics to promote qualitative understanding of context of information. Learning analytics is also (strongly) related to discourse analytics, which seems to be consistent with the strong emphasis of learning analytics on social learning and which is further confirmed by extracted relationships of discourse learning analytics with sense-making, argumentation and social, all of which are types of skills recognized as important for the modern society.

In future work, we plan to analyze further the research trends over the years for the LAK and EDM communities. Another of our goals is to compare the extractions of an ontology learning system such as OntoCmaps with Linked data Semantic Annotators such as DBPedia Spotlight or Alchemy.

4. CONCLUSION

Funnily, our text analysis tool inferred that EDM is an abbreviation of learning analytics. This probably comes from the open debate reflected in the analyzed papers about the relationships between learning analytics and educational data mining. We hope that this paper sheds some light on the (dis)similarities of the two areas. We also hope that our analysis of the LAK 2013 Data Challenge dataset with the ontology learning tools indicated a high potential of this type of analytics to help the research community of new research discipline define itself and relationships with closest communities. More interesting results are available on our website [8]. For example, those results allow for (i) comparing results of different concept/relationship measures and (ii) chronological trends emerging throughout the years of individual editions of both the conferences. An example of one of the visualizations available at [8] is presented in Figure 2.

Of course, ontology learning tools are not perfectly accurate, and thus, few “strange” concepts and relationships are shown in our tables. An opportunity is however in combining such ontology learning tools as starting points of the concept map development of the learning analytics domain, which can then be refined through crowd sourcing (e.g., in a Wiki-like manner).

5. REFERENCES