Spiral me to the core: Getting a visual grasp on text corpora through clusters and keywords

Maren Scheffel, Katja Niemann,
Sarah Leon Rojas
Fraunhofer FIT
Schloss Birlinghoven
53754 Sankt Augustin, Germany
{maren.scheffel, katja.niemann,
sarah.leon.rojas}@fit.fraunhofer.de

Hendrik Drachsler,
Marcus Specht
Open University of the Netherlands
Valkenburgerweg 177
6419 AT Heerlen, The Netherlands
{hendrik.drachsler,
marcus.specht}@ou.nl

ABSTRACT

The amount of literature within a research domain is ever growing, thus making it difficult to stay on top of everything. Getting a grasp on the important topics of and areas within a domain or even knowing where to start is often tough and tedious. This paper therefore presents a visualization, that is a cluster spiral, that offers a fast but plain and simple way of exploring the content of large text collections.

Categories and Subject Descriptors

H.3.1 [**Content Analysis and Indexing**]: Linguistic processing; H.3.3 [**Information Search and Retrieval**]: Clustering, Information filtering; I.2.7 [**Natural Language Processing**]: Text analysis; I.5.3 [**Clustering**]: Algorithms; I.5.4 [**Applications**]: Text processing; I.7.5 [**Document Capture**]: Document analysis

General Terms

Algorithm, Visualization

Keywords

learning analytics, natural language processing, clustering, keyword extraction, visualization

1. INTRODUCTION

One typical aspect of the world of research is the fact that the amount of literature being produced and published is growing every day. The more years pass, the more articles, papers, and books are available. Some research domains might have a slowly but steadily growing literature corpus while others grow rapidly. Looking only at those publications from the last year can be a fairly easy thing to do. But taking several years or even decades of publications into account when trying to get an overview about a chosen domain might prove rather difficult. When wanting to write a literature review within a certain domain of research or about a specific topic, it is thus often difficult to get a grasp on it and to know where to start. One way can be to rely on previous literature reviews. But when a topic spans over several domains, several research communities and a longer period of time, it could be nicer to take all of that into account at the same time in order to get a feel for what one is dealing with. This paper therefore describes a fast and easy way of getting a grasp on a collection of publications, using the LAK Dataset of the LAK Challenge 2014[1] as an example corpus.

2. THE LAK DATASET

The LAK Dataset contains a collection of structured data of several proceedings and journal volumes from the field of learning analytics and educational data mining [11]. The data have been processed according to Linked Data principles[2] and are thus available in machine readable format. As the data set includes the proceedings of the LAK conferences 2011-13, the proceedings of the EDM conferences 2008-13, plus some journal editions (in progress) of Educational Technology & Society and the Journal of Educational Data Mining, it is ideal for our purpose. Currently, there are 462 papers, 853 distinct authors and 272 distinct institutions included in the dataset. The data are available in several formats: RDF/XML, R statistic software compatible, and via a SPARQL endpoint.

The LAK Dataset has previously been used for the first LAK Challenge that took place during LAK2013 [1]. Derntl et al.[2] extract topic models and visualize topic dynamics and evolution over time with a special focus on how the introduction of the LAK conferences changed the topic dynamics of learning analytics and educational data mining. Fazeli et al.[3] look at socio-semantic networks of authors and papers within the learning analytics community in order to provide recommendations to users, e.g. conference attendees. Maturana et al.[4] use their gnoss platform to provide faceted search within the LAK Dataset and provide visualizations of geographical author and organization networks as well as paper evolution and distribution. Another visualization of topic evolution within the LAK and EDM community is presented by Milikic et al.[5] with their tool

[1] http://lak.linkededucation.org/

[2] http://www.w3.org/DesignIssues/LinkedData.html

Paperista. One more social network analysis, this time with a focus on authors and institutions, is presented by Nawaz et al.[6]. The Cite4Me tool by Pereira Nunes et al.[7] offers search and recommendation functionalities within the LAK Dataset as well as reference datasets. Taibi et al.[12] analyze rhetorical patterns over time while Touaq et al.[13] create an ontology of LAK and EDM based on concept mapping in order to compare the two communities.

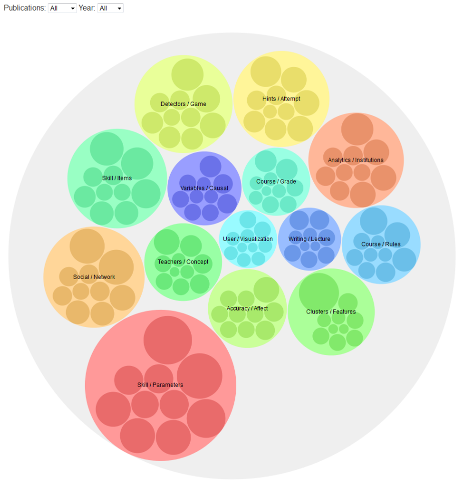
While some of these publications also deal with topic and concept mining, they often either focus on the evolution over time or relations between individual papers, authors, institutions, etc. within the LAK and EDM community when visualizing their results. Our approach, however, focuses on grouping a collection of publications based on their textual content and visualizing that clustered content rather than individual papers in order to get an overall impression of the collection in question. That we use the LAK Dataset for our analysis is one domain example as our approach also works for other large collection of texts.

3. ANALYSIS

Our approach for the analysis and visualization of the LAK Dataset makes use of the RDF version and bases on the following ideas: in order to get a grasp on what a collection of papers is about, keywords play an important role. Keywords offer a superficial but still highly useful semantic representation of a text as they "represent in condensed form the essential content of a document" [9]. For our analysis, a keyword can be one word as well as a sequence of up to three words. Another important means to get an overview over a collection of documents and thus a better grasp on such a collection is clustering. Su et al. [10] define clustering as "a process of partitioning a dataset into groups, or clusters, so that elements of the same cluster are more similar to each other than to elements of different clusters". We therefore employ both methods and combine their results into a visualization that supports users in getting an overview of what large text collections are about.

As we assumed that the keywords already provided within the LAK Dataset by the papers' authors would not be broad enough, that is, they are assigned manually and most likely based on a narrow word range typical for that research domain and thus not properly representative for the texts, we did not want to rely on them. We therefore automatically extracted keywords from all papers' abstracts and bodies using the *AlchemyAPI*[3]. Their algorithm extracts keywords from any given text using statistical algorithms as well as natural language processing techniques and ranks the extracted keywords according to their relevance. Although the *AlchemyAPI* already makes use of a stop word list, e.g. words such as *and*, *to*, *me*, *you*, etc., we created our own stop word list as keywords such as *learning analytics*, *educational data mining*, *data analysis*, *discussion*, *result*, etc. would otherwise quite likely come up as a keyword for the individual papers but would not help forming a distinguishing semantic representation within the given collection. Were our approach to be used for another domain or in a more mixed one, the stop word list could easily be adapted.

In the next step we clustered the paper collection by calling the *carrot2* Java API[4]. Three different clustering algorithms were available: Lingo, STC and k-means. We looked at all three algorithms and liked clusters created by Lingo quite well at first sight. Unfortunately, however, Lingo as well as STC both use soft clustering techniques, that is, they create overlapping clusters with papers possibly being assigned to more than one cluster. As the overlap is not limited to only a few documents but rather a lot, we decided not to use either of the two algorithms but use the bisecting k-means algorithm, i.e. the algorithm starts with $k = 2$ and then always bisects the largest cluster until the final k is reached, offered by *carrot2* instead. The calculation of the clusters is based on the papers' abstracts and main text bodies. Additionally to the clustering of the text collection, the *carrot2* algorithm also calculates labels for every cluster. For our analysis we chose to work with two labels per cluster.



Publications: All Year: All

Figure 1: Start view of the visualization

Finally, in a third step, a JSON file was created combining the keyword extraction results with the clustering results as a source for the visualization: for every cluster, the keywords of its papers are combined and sorted according to their rank. Then the ten keywords with the highest rank are kept for each cluster. A source file thus contains two labels, ten keywords and a list of the respective papers for each cluster. In order to offer users several views and to look at the dataset from different angles, we calculated clusters for several publication-year combinations.

4. VISUALIZATION

When dealing with the analysis of large amounts of text data, visualization is "of crucial importance in facilitating knowledge discovery, as well as providing a *big picture* overview of overwhelmingly large amounts of data" [8]. For our visualization we used the Data-Driven Documents D3.js framework[5], i.e. a JavaScript library, paired with HTML, CSS and JQuery to process the previously created JSON files.

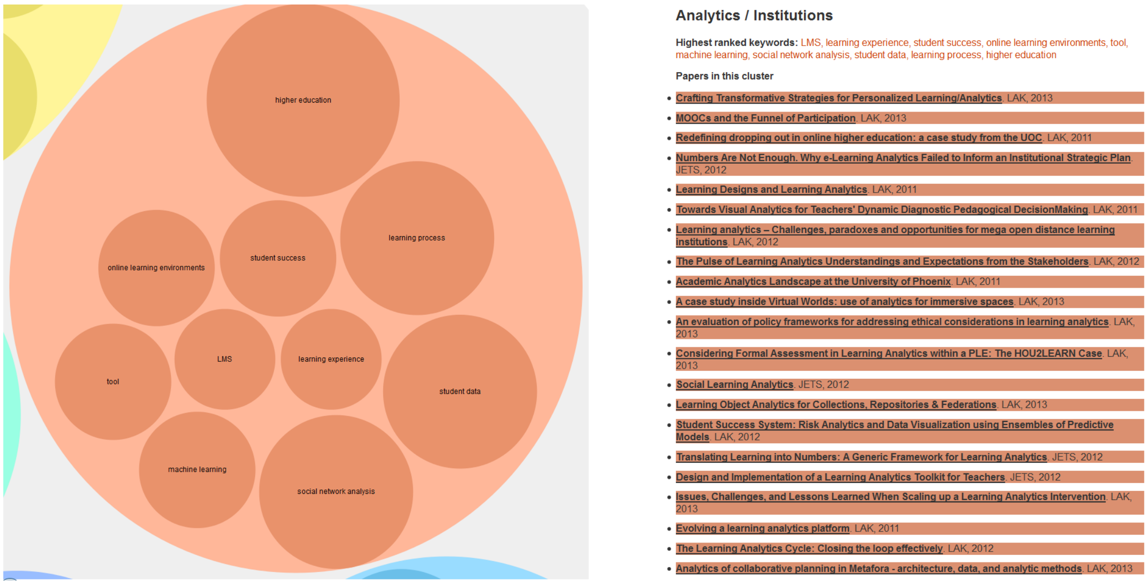
Figure 1 shows the default starting view of our visualization[6]: the clusters for all publications from all years in the

[3] http://www.alchemyapi.com/

[4] http://project.carrot2.org/

[5] http://d3js.org/

[6] The visualization is available at http://mitarbeiter.fit.fraunhofer.de/˜niemann/LAKchallenge2014/



Analytics / Institutions

Highest ranked keywords: LMS, learning experience, student success, online learning environments, tool, machine learning, social network analysis, student data, learning process, higher education

Papers in this cluster

- Crafting Transformative Strategies for Personalized Learning/Analytics. LAK, 2013
- MOOCs and the Funnel of Participation. LAK, 2013
- Redefining dropping out in online higher education: a case study from the UOC. LAK, 2011
- Numbers Are Not Enough. Why e-Learning Analytics Failed to Inform an Institutional Strategic Plan. JETS, 2012
- Learning Designs and Learning Analytics. LAK, 2011
- Towards Visual Analytics for Teachers' Dynamic Diagnostic Pedagogical DecisionMaking. LAK, 2011
- Learning analytics – Challenges, paradoxes and opportunities for mega open distance learning institutions. LAK, 2012
- The Pulse of Learning Analytics Understandings and Expectations from the Stakeholders. LAK, 2012
- Academic Analytics Landscape at the University of Phoenix. LAK, 2011
- A case study inside Virtual Worlds: use of analytics for immersive spaces. LAK, 2013
- An evaluation of policy frameworks for addressing ethical considerations in learning analytics. LAK, 2013
- Considering Formal Assessment in Learning Analytics within a PLE: The HOU2LEARN Case. LAK, 2013
- Social Learning Analytics. JETS, 2012
- Learning Object Analytics for Collections, Repositories & Federations. LAK, 2013
- Student Success System: Risk Analytics and Data Visualization using Ensembles of Predictive Models. LAK, 2012
- Translating Learning into Numbers: A Generic Framework for Learning Analytics. JETS, 2012
- Design and Implementation of a Learning Analytics Toolkit for Teachers. JETS, 2012
- Issues, Challenges, and Lessons Learned When Scaling up a Learning Analytics Intervention. LAK, 2013
- Evolving a learning analytics platform. LAK, 2011
- The Learning Analytics Cycle: Closing the loop effectively. LAK, 2012
- Analytics of collaborative planning in Metafora - architecture, data, and analytic methods. LAK, 2013

Figure 2: Cluster view and paper list

LAK Dataset. On this starting page, the users can choose the publication(s) and the year(s) they want to visualize. They can either choose each publication individually (i.e. LAK, EDM or JETS) or all of them together. When all publications are chosen, the user can choose between individual years or all years combined. If a single publication is chosen, only all of its years can be chosen for display. This adds up to a total of ten possible combinations.

The clusters take the form of circles and are ordered according to their size in the form of a spiral with the largest cluster having the largest circle and being positioned at the outside of the spiral and the smallest cluster being in the middle of the spiral. Additionally to size and position, every cluster also has its own color and is labeled with the two terms calculated by the *carrot2* algorithm.

By clicking on a cluster, the view changes and the visualization zooms into to the chosen cluster. Next to it a list of all the papers in that specific cluster is given, showing the papers' titles and the publications they were taken from. The titles in that list are linked to a Google search for the respective paper so that users can immediately take a closer look at it if needed. Figure 2 shows the cluster labeled *Analytics/Institutions.*

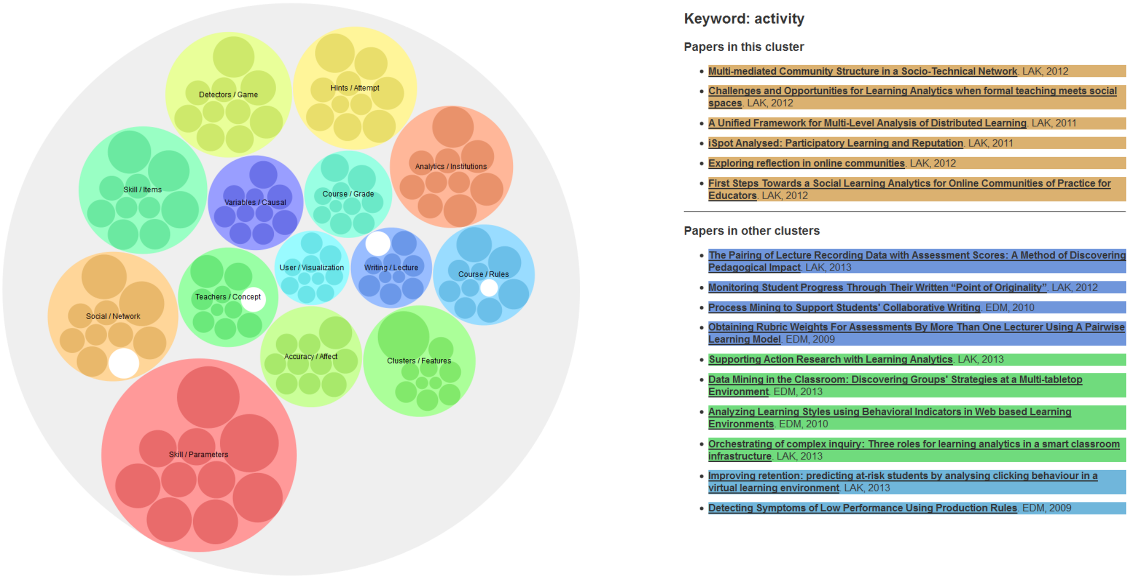
Once the users have zoomed into a cluster, the keywords of that cluster become visible. Figure 2 shows that the *AlchemyAPI* algorithm indeed extracts single words as well as word sequences, e.g. *tool*, *student success*, *online learning environments*, etc. A very common visualization method for keywords are tag clouds as "a tag cloud is highly effective in summarizing large amounts of text in an easily readable, and understandable, visual manner" [8]. In order to continue the circle approach used for the clusters, however, we adapted the common usage of font size, coloring and word positioning in tag clouds and used sized and spirally ordered circles instead: the more often a keyword appears in a cluster, the larger and the further out in the spiral its circle is.

Clicking on a keyword circle results in a new list next to the visualization. All papers represented by that keyword within that cluster are given, followed by a list of papers from other clusters that also have the chosen word as a keyword. The lists are also color-coded and the keyword circle is highlighted in all clusters so as to more easily find the corresponding cluster(s). Figure 3 shows the keyword paper list for the keyword *activity* of the *Social/Network* cluster and three other ones.

5. DISCUSSION AND CONCLUSION

Looking at the visualization of the whole dataset, i.e. all publications from all years are taken into account at the same time, the fourteen clusters and their labels offer a nice overview of what the text collection is about. For example, we can see that *social networks*, *teachers* and *institutions* play an important role, but *skills*, *courses* and *clusters* are important topics within the research area as well. When zooming into the clusters and looking at the different paper lists of the clusters, it is noticeable that many of the lists contain way more papers from the EDM than from LAK. Only two of the clusters are dominated by LAK papers while ten are dominated by those from EDM. This effect, however, is mainly due to the fact that there are about three times more papers from EDM than from LAK. After normalizing the numbers, about half of the clusters are still dominated by one publication type (two by LAK and five by EDM) and the other half is split between them. In general one can say that LAK papers share their topic range quite well with JETS and EDM as only the *Social/Network* and the *Analytics/Institutions* clusters are dominated by LAK papers. Some EDM topics, however, seem to be more exclusive and specific to EDM, e.g. *Skill/Parameters* and *Detector/Game*, as many of the five EDM-dominated clusters contain no or very little papers from LAK or JETS. Topics common to LAK as well as to EDM are, among others, *Clusters/Features*, *Teachers/Concept* and *User/Visualization.*

Another result that the visualization provides becomes clear when inspecting the clusters' keywords more closely. For many clusters the keywords cover aspects of a domain,



Keyword: activity

Papers in this cluster

- Multi-mediated Community Structure in a Socio-Technical Network. LAK, 2012
- Challenges and Opportunities for Learning Analytics when formal teaching meets social spaces. LAK, 2012
- A Unified Framework for Multi-Level Analysis of Distributed Learning. LAK, 2011
- iSpot Analysed: Participatory Learning and Reputation. LAK, 2011
- Exploring reflection in online communities. LAK, 2012
- First Steps Towards a Social Learning Analytics for Online Communities of Practice for Educators. LAK, 2012

Papers in other clusters

- The Pairing of Lecture Recording Data with Assessment Scores: A Method of Discovering Pedagogical Impact. LAK, 2013
- Monitoring Student Progress Through Their Written “Point of Originality”. LAK, 2012
- Process Mining to Support Students' Collaborative Writing. EDM, 2010
- Obtaining Rubric Weights For Assessments By More Than One Lecturer Using A Pairwise Learning Model. EDM, 2009
- Supporting Action Research with Learning Analytics. LAK, 2013
- Data Mining in the Classroom: Discovering Groups' Strategies at a Multi-tabletop Environment. EDM, 2013
- Analyzing Learning Styles using Behavioral Indicators in Web based Learning Environments. EDM, 2010
- Orchestrating of complex inquiry: Three roles for learning analytics in a smart classroom infrastructure. LAK, 2013
- Improving retention: predicting at-risk students by analysing clicking behaviour in a virtual learning environment. LAK, 2013
- Detecting Symptoms of Low Performance Using Production Rules. EDM, 2009

Figure 3: Overview with highlighted keywords and corresponding paper list

an approach, a goal, the data used and the stakeholders involved. Take the *Analytics/Institutions* cluster for example: the keyword *higher education* tells us the domain that is important for this cluster, we can also see that the approaches of *social network analysis* and *machine learning* play a role. As for the goals that this cluster deals with, there are *learning process* and *student success*, and the data analyzed is *student data* coming from *online learning environments* and *LMSs*. Taking the *Course/Grade* cluster as a second example, we can see that it deals with the approaches of *formative evaluation* and *classification algorithms* that are applied to data taken from *online learning activities* in *online courses*, *submissions*, *assignments* and *posts* in order to supply *predictive models* dealing with *final grades* to *instructors*.

These two analyses offer a first step to getting a grasp on the main research topics of the learning analytics and educational data mining literature, including their commonalities and differences. We will use the cluster spiral to delve further into these domains and plan to provide an extensive review that is based on the publications' essential characteristics, e.g. application domain, stakeholders, methodologies, and goals. For new scientists to these communities such a review can offer an entry point to the field. It is also useful to bridge the gap between the LAK and EDM communities and provide researchers from one side insight to the other. A third valuable aspect of a literature review would also be the retrieval of new and important research questions.

6. REFERENCES

[1] M. d'Aquin, S. Dietze, H. Drachsler, E. Herder, and D. Taibi. *Proceedings of the LAK Data Challenge*, volume 974. CEUR Workshop Proceedings, Leuven, Belgium, 2013.

[2] M. Derntl, N. Günnemann, and R. Klamma. A dynamic topic model of learning analytics research. 2013. In [1].

[3] S. Fazeli, H. Drachsler, and P. Sloep. Socio-semantic networks of research publications in the learning analytics community. 2013. In [1].

[4] R. Maturana, M. Alvarado, S. López-Sola, M. Ibañez, and L. Ruiz Elósegui. Linked data based applications for learning analytics research: faceted searches, enriched contexts, graph browsing and dynamic graphic visualisation of data. 2013. In [1].

[5] N. Milikic, U. Krcadinac, J. Jovanovic, B. Brankov, and S. Keca. Paperista: Visual exploration of semantically annotated research papers. 2013. In [1].

[6] S. Nawaz, F. Marbouti, and J. Strobel. Analysis of the community of learning analytics. 2013. In [1].

[7] B. Pereira Nunes, B. Fetahu, and M. Casanova. Cite4me: Semantic retrieval and analysis of scientific publications. 2013. In [1].

[8] A. A. Puretskiy, G. L. Shutt, and M. W. Berry. Survey of text visualization techniques. In M. W. Berry and J. Kogan, editors, *Text Mining: Applications and Theory*, pages 107–127. John Wiley & Sons, Ltd, 2010.

[9] S. Rose, D. Engel, N. Cramer, and W. Cowley. Automatic keyword extraction from individual documents. In M. W. Berry and J. Kogan, editors, *Text Mining: Applications and Theory*, pages 3–20. John Wiley & Sons, Ltd, 2010.

[10] Z. Su, J. Kogan, and C. Nicholas. Constrained clustering with k-means type algorithms. In M. W. Berry and J. Kogan, editors, *Text Mining: Applications and Theory*, pages 81–103. John Wiley & Sons, Ltd, 2010.

[11] D. Taibi and S. Dietze. Fostering analytics on learning analytics research: the lak dataset. 2013. In [1].

[12] D. Taibi, Á. Sándor, D. Simsek, S. Buckingham Shum, A. Deliddo, and R. Ferguson. Visualizing the lakedm literature using combined concept and rhetorical sentence extraction. 2013. In [1].

[13] A. Zouaq, S. Joksimović, and D. Gašević. Ontology learning to analyze research trends in learning analytics publications. 2013. In [1].