VISLA15: 1^{st} international workshop on Visual Aspects of Learning Analytics

organized at the 5th international Learning Analytics and Knowledge conference (LAK15)

The use of visualization techniques for learning is not new. For instance, visualizations have been used in maps and drawings for thousands of years. In a learning analytics context, the application of information visualization techniques can help both teachers and learners to explore and understand relevant user traces that are collected in various (online) environments and to improve (human) learning. The goal of our workshop is to build a strong research capacity around visual approaches to learning analytics. The longer term goal is to improve the quality of learning analytics research that relies on information visualization techniques.

Each contribution to the workshop explicitly addressed the following items:

1. What kind of data is being visualized? What tools were used to clean up the data (if any)?
2. For whom are the visualizations intended (learner, teacher, manager, researcher, other)?
3. How is data visualized? Which interaction techniques are applied? What tools, libraries, data formats, etc. are used for the technical implementations? What workflow and recipe was used to develop the visualization?
4. Why are the chosen visual approaches applied (i.e. rationale behind the application of a visualization)?
5. How has the approach been evaluated or how could it be evaluated?
6. What were the encountered problems and pitfalls during the visualization process?

The workshop is intended for anyone who is using, or is interested in visualization techniques to support learning analytics. The goal of our workshop is to build a strong research capacity around visual approaches to learning analytics. The longer term goal is to improve the quality of learning analytics research that relies on information visualization techniques.

During our 1-day workshop, we aimed to facilitate a very interactive and engaging event where we wanted to avoid death by powerpoint by all means and promote discussion activities over presentational ones. In the first half of the workshop, we therefore asked participants to shortly present the work of another submission and to relate it back to their own work.

During the second half of the workshop, we invited the participants to share their tools, workflows and recipes in a hands-on discussion session so that they could benefit from

Published in CEUR-WS:
Proceedings of the VISLA15 Workshop on Visual Aspects of Learning Analytics (Duval et al., eds), organized at LAK15.

each others' knowledge, apply their visual approaches on either their own dataset or on a dataset that we provided.

Finally, we moved the discussion to the final topic of the workshop, which is the development of the equivalent of the VAST challenge for learning[1], which was linked back with the LAK14 and LAK15[2] data challenge:

> *"The annual Visual Analytics Science and Technology (VAST) challenge provides Visual Analytics researchers, developers, and designers an opportunity to apply their best tools and techniques against invented problems that include a realistic scenario, data, tasks, and questions to be answered. Submissions are processed much like conference papers, contestants are provided reviewer feedback, and excellence is recognized with awards. A day-long VAST Challenge workshop takes place each year at the IEEE VAST conference to share results and recognize outstanding submissions."*

The VISLA15 organizers
Erik Duval, Joris Klerkx, Katrien Verbert, KU Leuven, Belgium
Martin Wolpers, Fraunhofer-Institute for Applied Information Technology FIT, Germany
Abelardo Pardo, University of Sydney, Australia
Sten Govaerts & Denis Gillet, EPFL, Switzerland
Xavier Ochoa, ESPOL, Ecuador
Denis Parra, PUC, Chile

---

[1] K. Cook, G. Grinstein, and M. Whiting. The vast challenge: history, scope, and outcomes: An introduction to the special issue. Information Visualization, 13(4):301–312, 2014.

[2] H . Drachsler, S. Dietze, E. Herder, M. d'Aquin, and D. Taibi. The learning analytics & knowledge (lak) data challenge 2014. In Proceedings of the Fourth International Conference on Learning Analytics And Knowledge, LAK '14, pages 289–290, New York, NY, USA, 2014. ACM

Table of contents

Visualizing Uncertainty in the Prediction of Academic Risk

Xavier Ochoa
Escuela Superior Politécnica del Litoral
Vía Perimetral, Km. 30.5
Guayaquil, Ecuador
xavier@cti.espol.edu.ec

ABSTRACT

This work proposes a generic visual representation to help relevant decision-makers to effectively address the inherent uncertainty present in the prediction of academic risk based on historical data. The three main sources of uncertainty in this type of prediction are visualized: the model predictive power, the data consistency and the case completeness of the historic dataset. To demonstrate the proposed visualization technique, it is instantiated in a real-world scenario where the risk to fail at least one course in an academic semester is predicted and presented in a student-counseling system. This work also proposes how this visualization technique can be evaluated and applied to other Visual Learning Analytics tools.

Categories and Subject Descriptors

K.3.1 [**Computing Milieux**]: Computers and Education-Computer Uses in Education

Keywords

Visual Learning Analytics, Uncertainty Visualization, Academic Risk

1. INTRODUCTION

The main goal of the Learning Analytics field is to provide relevant information to the actors of the learning process (students, instructors and administrators) to help them take better learning-related decisions. A considerable amount of research effort [6] has been invested in find ways to analyze the large amount of traces that are a by-product of the learning process to convert it into that relevant information. An equal important, but lesser researched, area of Learning Analytics explores the best ways in which that relevant information is presented to the final user to maximize its usefulness for decision-making. This second area is often called "Visual Learning Analytics" given that it is very related to the field of Visual Analytics, that focuses on "analytical reasoning facilitated by interactive visual interfaces" [14]. Visual Analytics differentiates from simple data visualization because its purpose is not only presenting the information resulting from a predefined analysis process, but empowering the decision-maker to control the analysis process and interact with the multiple dimensions that the resulting information could have to gain a deep understanding of the implications that those results have in the decision at hand.

Currently, there are very few early examples of Visual Learning Analytics, in contrast to simple visualization of Learning Analytics results: Lemo [7] is a system that use interactive visualization to help instructors understand the activity logs of LMSs. The end-user is capable of exploring the dataset through selecting and filtering the desired information in a variety of visualization options. Gomez et al. [8] also create a system to explore in deeper detail the academic and non-academic data stored in the LMS system through the use of interactive visualizations.

One virtually unexplored avenue of Visual Learning Analytics is how to make explicit the uncertainty that is inherent in any analysis process in a way in which is meaningful for the decision-maker. Moreover, if possible, the decision-maker should also be able to manipulate the analysis process to adjust the uncertainty to a level where he or she finds appropriate. This kind of techniques to present and manage the uncertainty are common in more mature fields such as meteorology (e.g. hurricane path prediction uncertainty [13]), medicine (e.g. uncertainty in the effect of medical interventions [10]) and economy (e.g. uncertainty in the prediction of future growth [13]). There exists, however, some examples of the visualization of uncertainty in Open Learner Models [5] that could be consider a precursor in the field of Visual Learning Analytics.

This work will focus on how Visual Learning Analytics techniques could be used to visualize and control the inherent uncertainty in the prediction of academic risk. The organization of this paper is as follows: Section 2 explores how academic risk is usually obtained and which are the main sources of uncertainty in this type of analysis. Section 3 discusses how the prediction value, together with the main uncertainty values should be visualized. Section 4 presents a case-study where the visualization techniques are instantiated to help counselors give advice about the risk to fail a semester to individual students. Finally, the paper finishes with conclusions about the work and guides for further work

Published in CEUR-WS:
Proceedings of the VISLA15 Workshop on Visual Aspects of Learning Analytics (Duval et al., eds), organized at LAK15.

to evaluate the technique and how to adapt it to other Visual Learning Analytics tools.

2. PREDICTING ACADEMIC RISK

In the context of this work, the term "academic risk" is defined as the probability of a student to reach an unfavorable outcome in their studies. This unfavorable outcome could be as benign as the failure to submit a homework or as costly as dropping-out of a program. As very little can be done once the unfavorable outcome has been already reached, especially for the more costly forms (e.g. failing a course or dropping-out), there is a strong incentive to being able to estimate the academic risk of the student, or what is equivalent, predict the probability that the student will, without intervention, reach the unfavorable outcome. Due to its importance, predicting different forms of academic risk has been one of the oldest forms of Learning Analytics [11].

There are several current examples of systems that seek to estimate different kinds of academic risks: Signals [1] is arguably the poster-boy of learning analytics systems to predict academic risk. Using historical and current information about the behavior of a student in a course, it is able to predict the probability that the student has of fail the course. Another, more simple approach is taken by StepUp! [12] that just compares the activity of a student with the activity of their peers and assigns a ranking value that could be seen as a fuzzy academic risk predictor. Finally, there are several modern drop-out risk predictors from which the work of Dekker et al. [4] could be considered a good representative. This system uses a classification tree trained over historical data in order to obtain rules to assess the risk of a student to dropping-out from a university program.

All of the mentioned systems used data collected from previous or current students to create a prediction model. This model could be built with statistical or data-mining methods. Once the model has been built, it is fed with the information from the student target of the prediction and an estimation of the academic risk is produced. This estimation is normally presented to the instructor, counselor or the student through some form of visualization technique.

In all of the steps of the above-mentioned process there are inherent uncertainties that are propagated and contribute to the uncertainty that is present in the estimated value of academic risk. The following subsection discusses the nature of these sources of uncertainty and their relative importance for the prediction.

2.1 Uncertainty Sources

To facilitate the analysis of the different sources of inherent uncertainty in the prediction of academic risk, they are classified in two group according to their origin: predictive model limitations and dataset limitations. The following subsections sub-classify these two groups into more concise and measurable uncertainty values.

2.1.1 Predictive Model Limitations

Perhaps the most obvious source of uncertainty introduced in any type of prediction is the one introduced by the imperfections of the predictive model. In general, predictive models are built to take as input a group of predictor variables and to produce a predicted value. Given that models are only an approximation and simplification of reality, it is expected that the predicted values differ, in different degrees, from the real values. A whole area of Statistics is devoted to measure the predictive power of different types of models. The best example of the measure of the predictive power is the R-squared statistic used to score regression models. This measurement establishes what percentage of the variance in the real values of the predicted quantity are explained by the model.Different models usually have different predictive power depending on the predictor variables used, the type of algorithm and the amount and quality of data used to build them. It is a common practice to evaluate different competing models and select the one with the best predictive power according to an appropriate scoring function.

2.1.2 Dataset Limitations

Given that most academic risk predictors are built based on historical or current data, the characteristics of the data and its limitations play a major role in the overall uncertainty of the predicted value of that risk. The work of Thomson et al. [15] established a detailed typology for the limitations of data that affect certainty in predictive models: accuracy, precision, completeness, consistency, lineage, currency, credibility, subjectivity and interrelatedness. All these types of limitations are usually defined at the dataset level and their effect in uncertainty is usually propagated into the final predictive power of the model that was built with that dataset.

Given the nature of academic datasets, the most important of these dimensions are consistency and subjectivity. Historical academic data, for example final grades of students, is generally accurate (there is a significant cost of registering a grade wrongly), precise (it has enough resolution to separate passing and failing students), complete (all students should have grades or at least a pass/fail at the end of a course), current (the grades are producing during the course or at least very close to the ending of the course) and credible (the academic institutions will have serious problems if their academic records are not credible). Also, academic records have no major problems with lineage (the grades are rarely processed after the instructor records them) and the records do not suffer from interrelatedness (instructors do not copy the grades from one student to another or among them). However, consistency of academic data could introduce uncertainty in the prediction of academic risk. As academic programs evolve, they also change: the courses offered could change, the grading rules could become more strict or more relaxed, different instructors will imprint their own characteristic in the courses, among other changes. Depending on the nature and magnitude of the changes, the academic records of a current student and one that studied ten years ago could not be comparable or, more dangerously for prediction models, could provide a false sense of similarity when in reality the values in those records are not measuring the same students characteristics. Another possible limitation of historical academic data is its subjectivity. Grades, scores and student evaluations are commonly assigned according to the criteria of the instructor. Even during the same course, students that did a similar level of work could receive different grades. While the effect of consistency errors in the overall prediction uncertainty could be limited by only considering comparable years of the academic program in the dataset, the uncertainty produced by the subjectivity could not be reduced if it is already present in the data.

Due to the fact that most academic risk predictors compare current students to previous similar students that were in a similar context, another type of data limitation plays a role in the overall uncertainty of the prediction: case completeness. For example, predictive model A estimates the academic risk of failing a course based on number of other courses taken at the same time and the GPA of the student; predictive model B estimates the academic risk of failing a course based on the number of courses taken at the same time, the GPA of the student, the fact that the student has an external job, if the student is married, the number of children the student has, the distance from his house to the university and the number of courses taken before the current one. Both models estimate the academic risk of failing the course as the percentage of similar students that have failed the course in the past. A hypothetical prediction power analysis shows that model B is less uncertain that model A. However, this prediction power is calculated for the general population, for some students model A could be less uncertain than model B. Lets suppose that student A is taking 3 other courses, has a GPA of 3.5, has an external job, is married, has 5 children, lives 100 km from the university and has taken just one course before the current one. Lets suppose too that this is a very unusual combination of values for the students of this specific course. If the model A is applied, only the number of other courses that the student is currently taking (3) and his or her GPA (3.5) are considered. These two values, by themselves, are not unusual, so it is probable that there will be several previous students that could be considered similar. The prediction of academic risk for the hypothetical student will be drawn from a large pool of previous experiences. If the model B is applied, due to the unusual values of the rest of variables, the model could only find one other student close enough to be considered similar in the dataset. In this situation, the prediction of academic risk for the student will be 100%, if the previous student failed the course or 0% if he or she passed. While, in general, model B has more predictive power than model A, for this particular student the approximate estimation of model A will be much more less uncertain than the one provided by model B, due to the lack of similar cases in the dataset. The prediction for "outlier" students, that is, students that have few similar students in the dataset, is less certain than the prediction for "mainstream" students, that has a large collection of similiar cases. Simple models have less similarity dimensions, and the number of possible cases is lower than in complex models with larger dimensions sets. The variety and quantity of cases in the dataset, that is the case completeness of the dataset, introduce a uncertainty factor that varies from student to student and depends on the complexity of the model.

3. VISUALIZING UNCERTAINTY

As mentioned in the introduction, the visualization of uncertainty is already an established feature in more mature fields. In Visual Learning Analytics, however there are still no thoroughly evaluated techniques. The most recommended path in this case will be to adapt uncertainty visualization techniques that are common and proved useful [3] in other fields to represent the predicted value, together with the different uncertainty produced by the sources described in the previous section: the model predictive power, the data consistency and the case completeness. The goal of the visualization of those values is to present the most information about the prediction in an interpretable and useful way. The following subsection proposes various techniques for each one of these elements in detail.

3.1 Predicted Risk Value

The value of the academic risk of a student, being just a scalar that can be expressed as an easily interpretable numeric value between 0 and 1 (as probability) or from 0% to 100% (as relative frequency) can be presented using a large variety of visualization techniques such as textual, progress arc, gauge or bullet graphs. Figure 1 shows an example of this type of visualizations. Attached to the visualization of the value, all of these types of visualization present the decision-maker with a pre-defined guide to assess the level of risk described depending on the magnitude of the value. In the case of textual and arch representations, the color of the text or the arch (e.g. green, yellow and red) or an additional iconic representation (e.g. traffic light) could be used to provide an indication of the severity of the risk. In the case of the gauge and bullet graphs, different ranges can be color-coded to also provide this information. Some previous implementations of visualization of academic risk, such as Signals [1], use only an iconic representation (the traffic light approach) to represent the predicted value. Representing only the range in which the value is, instead of the actual value is used to account for the uncertainty of the prediction. However, in most cases, those ranges are crisp, meaning that a single unit change in the predictive value can cause the color to change, defeating the purpose of presenting only ranges in the first place. For example, a student with a risk of 0.50 will be coded with green, while a student with a risk of 0.51 will be coded yellow. With just the iconic representation, there is no way for the decision-maker to establish if the students is closer to green or to red. Moreover, the span of the ranges (what values are considered to be green, yellow or red) is often also unknown to the decision-maker. Using only the iconic representation is discouraged given that this work present other ways to deal with the inherent uncertainty in the prediction.

3.2 Model Predictive Power

Similarly to the predicted risk value, the model predictive power is also an scalar magnitude. Contrary to risk probability, the meaning of the output of the different model-scoring techniques (such as R-squared, BIC, AIC, Brier score, etc.) are far from being easy to interpret by non-statisticians. To effectively communicate the predictive power of the model, or what is the same, the level of uncertainty that a given model will introduce in the prediction, the expert analyst in charge of the academic risk prediction should define a set of iconic representations (e.g. traffic lights, happy-sad faces, plus signs, etc.) to correspond with different values of predictive power. Given that usually there are no model with "bad" power (otherwise it will not be used in the analysis), it is recommended that a plus signs textual representation ("+"

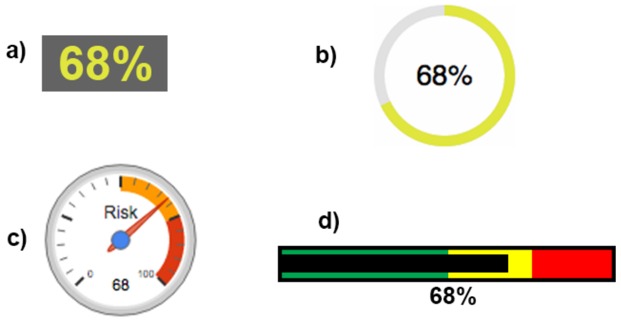


Figure 1: Predicted value visualization: a) textual representation, b) progress arc graph, c) gauge graph and d) bullet graph

for lower scoring models, "++" for medium scoring models and "+++" for the best scoring models) is used to represent different levels of power. The words "Good", "Very Good" and "Excellent" could be complement or replace this visualization. An example of this visualization could be seen in Figure 2.

It is important to note that this visualization is only necessary when the decision-maker can select between different models or the system chooses the model based on the available data. If the predictive risk is using a single model, the value of presenting this extra information is diminished.

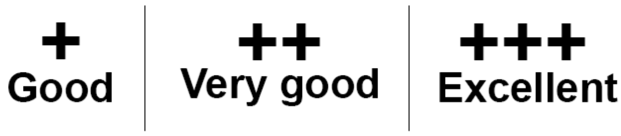


Figure 2: Model predictive power visualization

3.3 Data Consistency

The representation of uncertainty introduce by the data inconsistency is challenging given that there is no way to precisely measure it. In the case of academic datasets, the consistency is related to the changes in different aspects of the study program or course over time. It is expected that the closer in time the historic data is, the greater the level of consistency and the lower the level of uncertainty. If there exists a record of major changes in the academic program (course changes, evaluation policies changes, etc) or the courses (syllabus change, pre-requisites changes, instructor change, etc), they can be plotted in a timeline that span over the whole data range of the historical data. In this way, instructors and counselors that are familiar with the history of the program or course could recognize the changes and adjust their perception of the uncertainty introduced in the prediction, while students or users not familiar with the history of the program or course could just count the number of changes to form their own estimation of the uncertainty in the prediction, although less precise than the ones with previous knowledge. An example of this type of visualization can be seen in Figure 3

3.4 Case Completeness

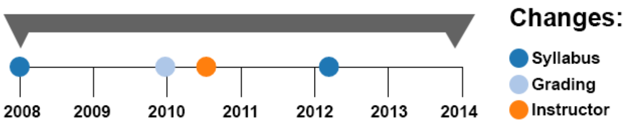


Figure 3: Data consistency visualization for a course historical data

In most predictive models is easy to obtain a measure of how many "similar" elements are considered at the moment of obtaining the predictive value for a given element. In the case of academic data, the case completeness could be measured as the number of records that are directly used to calculate the academic risk of a given student. This number could go from 0 to the total number of records in the dataset. A low value is an indication of a high uncertainty in the predicted value. Higher values, usually larger than 30, are enough to discount the number of cases as a source of uncertainty. The recommended visualization technique for this value is an iconic representation with icons that represent alert states at different number of different cases pre-defined by the expert behind the analysis (e.g. a red stop sign for values between 0 and 5, a yellow exclamation mark for values between 5 and 30 and a green check for values higher than 30). Together with the icon, a textual representation of the number of cases could be included to improve understandability (e.g. This prediction is based only on 3 previous cases). Figure 4 presents an example of this visualization.

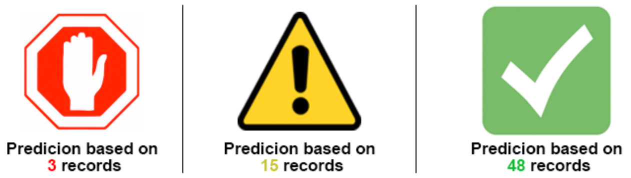


Figure 4: Case completeness visualization based on iconic representation

3.5 Interaction

The visualization described in the previous sub-section could help the decision-maker to better understand the inherent uncertainty of the risk value prediction. However, if the decision-maker is not confortable with the uncertainty of the prediction the only course of action is to discard the prediction. As mentioned in Section 2, the uncertainty of the prediction depends on several factors such as the model used, the length of historical data used and the number of similar cases used by the model to generate the prediction for a given student. The trade-off between these parameters is decided by the expert in charge of the prediction. Usually the model selected will be the one with greatest predictive power and the range of historical data will be selected to maximize this number. This selection is bound to be sub-optimal for some students, specially those with special cases. The use of interactive visualization transfer the control of the analysis parameters to the decision-maker. He or she could adjust them in order to reach the lowest level of uncertainty possible for a given student and the domain knowledge that the decision-maker has about the academic program or course.

Very simple interactive controls could be added to the visualization in order to control the main parameters affecting uncertainty factors. Each time a new value is selected on those controls, the uncertainty visualizations should be updated enabling the exploration of the uncertainty space by the decision-maker. To control the uncertainty resulting from the predictive power of the model, the decision-maker could be presented with a set of widgets where the model algorithm or parameters could be selected. To control the uncertainty resulting from the lack of consistency in the historical records, the timeline where this information is presented could be complemented with a selection bar to select subsets of the whole time period. The uncertainty produced by the lack of similar cases could not be affected directly, but it will change its response to the changes in the model used and the selected time period.

4. CASE-STUDY: RISK TO FAIL

To illustrate the ideas presented in the previous sections, they will be applied to a real-world academic risk prediction application. This application is part of a larger counseling system used regularly by professors and students at a mid-size university in Ecuador. The goal of this application is to determine the academic risk of failing at least in the next semester based on the planned course selection and study load. To produce this prediction the application uses a variety of models that cluster the student and the planned semester with similar students and semesters in the historical dataset. The models calculate the risk based on the previous frequency of similar students in similar semesters that failed at least one course. The counselor could interact with the visual analysis by selecting the courses that the student will take the next semester, the type of clustering that is applied to select similar students and semesters and the time period used to obtain similar cases. The counselor is presented with a prediction of the probability of the student failing the course and the visualization of the uncertainty produced by the model, the data consistency and case completeness. The counselor use the information received to recommend the student to take more or less study load in the coming semester.

4.1 Dataset

The dataset used for this application was built based on a Computer Science program at the target university. All the courses taken by CS students each semester and the grades obtained in those courses were stored since the first semester of 1978 to the second semester 2013. The courses that have changed name were grouped together according to the transition rules during those changes. A total of 30.929 semesters were taken by 2.480 different students.

4.2 Predictions Models

A multi-level clustering approach was used to build different models to find similar students and calculate the academic risk value. Two main variables controlled the generation of the different models: the student similarity and the semester similarity. The students were clustered at three levels: No clustering at all (all the students were considered similar), clustering based on GPA values (five clusters based on range) and clustering based on similarity of grades in the different courses (the Fuzzy C-means (FCM) algorithm [2] was used to create 10 clusters). The semesters were clustered at five levels (all using Fuzzy C-means): Level 1, based on the total load of the courses calculated from their difficulty [9]; Level 2, based on the typology of courses; Level 4, based on the grades that the students obtain in the courses [9]; Level 4, based on the knowledge area of the courses; Level 5, based on the actual name of the courses. The intersection of the level student and semester clustering defines a predictive model. For example, a model is created by finding similar students based on their GPA taking similar semesters based on the difficulty of courses taken (Level 2). The predictive power of the models was obtained computing the Brier score [16] of the forecast made for the last semester (2013-2) with the models built from the data from all the previous semesters.

4.3 Visualizing the Prediction

Figure 5 presents the interactive visualization created for the case-study academic risk prediction application. All the elements discussed in Section 3 are present. The predicted value is presented using a bullet graph with a 0%-100% scale, a yellow interval between 50% and 75% and a red interval between 75% and 100%. The model prediction power is shown with an iconic representation of one, two or three plus signs, together with a textual description. The data consistency is represented with an interactive timeline indicating the major events that changed the Computer Science program during the analyzed period. The case completeness of the dataset for the target student is presented using an iconic representation of group of different amounts of people related to a color (one individual in red to indicate a large amount of uncertainty, few people in yellow to represent middle values and a green crowd to represent low values. Finally, selection boxes are presented to the decision-maker to define the levels of clustering (for students and semesters) that determine the model that will be used for the prediction. All of these visualizations and controls are implemented with easy-to-use D3 Javascript visualization library [1].

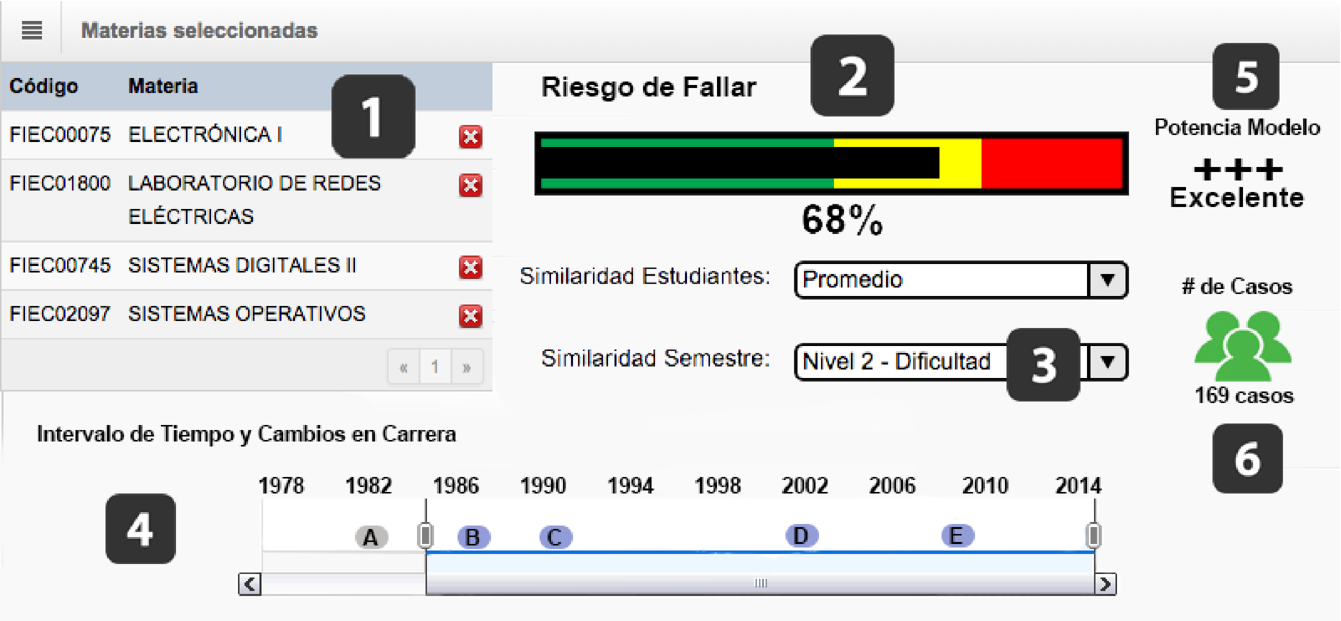
5. CONCLUSIONS AND FURTHER WORK

Visualizing the uncertainty in the prediction of academic risk, specially in an interactive way, has the potential to improve the usefulness of this type of systems. Even simple techniques are able to present to the decision-maker with the information needed to assess the uncertainty of the prediction for different selections of model and historical training data. With an interactive visualization the decision-maker, with their domain-expertise knowledge, becomes a co-designer of the analytic process, instead of a simple user of the results of the analysis. Implementing this visualization in real-world scenarios is simple given that the sources of uncertainty are well understood and could be measured or estimated.

The main task to be completed in this research is the real-world evaluation of the visualization to establish the answers to two main questions: 1) Is the visualization contributing to the understanding of the inherent uncertainty of the prediction of academic risk? and 2) Is the knowledge about the uncertainty helping the decision-maker to make better

---

[1] D3.JS visualization library - http://d3js.org



Código	Materia
FIEC00075	ELECTRÓNICA I
FIEC01800	LABORATORIO DE REDES ELÉCTRICAS
FIEC00745	SISTEMAS DIGITALES II
FIEC02097	SISTEMAS OPERATIVOS

Similaridad Estudiantes: Promedio

Similaridad Semestre: Nivel 2 - Dificultad

Figure 5: Example of visualization integrated in the counseling system: 1) Course selector, 2) Predicted academic risk value visualization, 3) Model selector, 4) Time period selector and consistency visualization, 5) Model predictive power visualization and 6) Case completeness visualization

decisions or to provide better advice? To answer these questions, the tool presented in the case study will be used in two experimental groups of counselors. One group will see the prediction and the uncertainty visualization. The second group will see only the prediction visualization. A third control group will continue to use the counseling system without the academic risk predictor application. The average failure rate for each counselor will be recorded at the end of the semester and compared with the failure rate between experimental and control group and also with the failure rate from previous semesters. Surveys will be conducted just after the counseling sessions in order to establish the level of understanding of the uncertainty in the prediction.

Finally, the ideas presented in this paper could be adapted to other types of Visual Learning Analytics tools, especially those focused on prediction and forecasting. The methodology followed in this paper could be a general framework for these adaptations: 1) exploring the main sources of uncertainty in the analysis, 2) establishing methods to measure or estimate the uncertainty contribution of those sources, 3) using existing visualization techniques to present the uncertainty values in a way that will be easy to interpret by the end-user, 4) provide control to the end-user through interactive visualizations to change the parameters to the models and to select the desired data and 5) evaluate the impact of the visualization. Visualizing the uncertainty is a way to empower the user of Visual Learning Analytics tools, stressing that automatic analysis could support, but not replace, human judgment.

6. ACKNOWLEDGMENTS

The author wants to acknowledge the contribution of Secretaría Nacional de Educación Superior, Ciencia y Tecnología (SENESCYT) in Ecuador and the Fonds Wetenschappelijk Onderzoek - Vlaanderen (FWO) in Belgium through the funding of the “Managing Uncertainty in Visual Analytics” project, to which this work belongs.

7. REFERENCES

[1] K. E. Arnold and M. D. Pistilli. Course signals at purdue: Using learning analytics to increase student success. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, pages 267–270. ACM, 2012.

[2] J. C. Bezdek. *Pattern recognition with fuzzy objective function algorithms.* Kluwer Academic Publishers, 1981.

[3] S. Deitrick and R. Edsall. *The influence of uncertainty visualization on decision making: An empirical evaluation.* Springer, 2006.

[4] G. W. Dekker, M. Pechenizkiy, and J. M. Vleeshouwers. Predicting students drop out: A case study. In *International Conference on Educational Data Mining (EDM)*. ERIC, 2009.

[5] C. Demmans-Epp, S. Bull, and M. Johnson. Visualising uncertainty for open learner model users. In *CEUR Proceedings associated with UMAP 2014*, 2014.

[6] R. Ferguson. Learning analytics: drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, 4(5):304–317, 2012.

[7] A. Fortenbacher, L. Beuster, M. Elkina, L. Kappe, A. Merceron, A. Pursian, S. Schwarzrock, and B. Wenzlaff. Lemo: A learning analytics application focussing on user path analysis and interactive visualization. In *Intelligent Data Acquisition and Advanced Computing Systems (IDAACS), 2013 IEEE 7th International Conference on*, pages 748 – 753,

2013.

[8] D. Gomez, C. Suarez, R. Theron, and F. Garcia. *Advances in Learning Processes*, chapter Visual Analytics to Support E-learning. InTech, 2010.

[9] G. Méndez, X. Ochoa, and K. Chiluiza. Techniques for data-driven curriculum analysis. In *Proceedings of the Fourth International Conference on Learning Analytics And Knowledge*, LAK '14, pages 148–157, New York, NY, USA, 2014. ACM.

[10] M. C. Politi, P. K. Han, and N. F. Col. Communicating the uncertainty of harms and benefits of medical interventions. *Medical Decision Making*, 27(5):681–695, 2007.

[11] C. Rampell. Colleges mine data to predict dropouts. *The chronicle of higher education*, 54(38):A1, 2008.

[12] J. L. Santos, K. Verbert, S. Govaerts, and E. Duval. Addressing learner issues with stepup!: an evaluation. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, pages 14–22. ACM, 2013.

[13] D. Spiegelhalter, M. Pearson, and I. Short. Visualizing uncertainty about the future. *Science*, 333(6048):1393–1400, 2011.

[14] J. Thomas and P. C. Wong. Visual analytics. *IEEE Computer Graphics and Applications*, 24(5):0020–21, 2004.

[15] J. Thomson, E. Hetzler, A. MacEachren, M. Gahegan, and M. Pavel. A typology for visualizing uncertainty. In *Electronic Imaging 2005*, pages 146–157. International Society for Optics and Photonics, 2005.

[16] D. S. Wilks. *Statistical methods in the atmospheric sciences*, volume 100. Academic press, 2011.

Using Sentence Compression to Develop Visual Analytics for Student Responses to Short Answer Questions

Aneesha Bakharia
Queensland University of Technology
Queensland
Australia
aneesha.bakharia@gmail.com

Shane Dawson
University of South Australia
South Australia
Australia
shaned07@gmail.com

ABSTRACT

In this paper, we report on early research to visualize and summarize student responses to short answer questions. Recently published, graph-based multi-sentence compression algorithms have been successfully applied to summarize opinions – a domain area with many similarities to short answer responses. Initial investigations reveal that visual analytics for short answer questions can be derived from the output of graph-based multi-sentence compression algorithms. A proposed open source short answer analytics tool is also briefly discussed, along with an evaluation plan. The proposed analytics tool will allow lecturers and tutors to have a high level overview of how students have responded to questions, identify knowledge gaps and provide feedback to students.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

Algorithms, Measurement, Design, Human Factors.

Keywords

Learning Analytics, Visual Analytics, Sentence Compression, Natural Language Processing, Summarization, Graph Layout.

1. INTRODUCTION

Short answer questions are a useful form of summative and formative assessment as they allow students to explain concepts in their own words without providing prompts to students. Grading and providing feedback for short answer questions, depending upon the number of student responses is a tedious task. Multiple choice questions which are easily automatically graded are predominantly used when student numbers are large such as MOOCs. Within flipped classroom scenarios, students are given pre-lecture readings and required to answer short answer questions, with the lecturer analyzing the student answers and addressing knowledge gaps during the lecture. The need for visual analytics to help lecturers and tutors gain an overview of student responses is therefore becoming increasingly important.

A rapid turnaround between students submitting their responses and the lecturer analyzing the response is required. This research is not focused on automatically grading short answer questions, rather the focus in on providing insight into how students have answered questions and allowing lecturers to easily determine the appropriate feedback and support that students require using visual analytics.

Lecturers and tutors require a way to analyse and visualize student responses so that they can:

- understand how students have responded to a question
- review the vocabulary being used
- identify knowledge gaps
- provide feedback to groups of students with similar knowledge gaps

The visual analytics tool that is proposed in this paper will apply sentence compression to summarize student responses to short answer questions. Graph-based sentence compression algorithms have recently been developed and applied to summarize multiple related sentences [3] and opinions within textual reviews [4]. The fact that student responses are made up of short sentences with common phrases being used amoung students, makes the summarization problem an ideal candidate for sentence compression because there is high similarity and redundancy between student responses. Graph-based approaches have also been applied to automatically grade student responses [4].

2. MULTI-SENTENCE COMPRESSION

The first algorithm that has been investigated is the multi-sentence compression algorithm by Filippova [3]. The Filippova algorithm summarizes similar or related sentences and outputs a single short sentence that summarizes the most salient theme conveyed in the cluster of sentences. The algorithm constructs a word graph and uses an approach based upon the shortest paths between words in the graph to produce a summary sentence. The algorithm is easy to implement because sentences must only be tokenized and part of speech tagged. The Filippova algorithm is the first sentence compression algorithm that does not require "hand-crafted rules, nor a language model to generate reasonably grammatical output".

All of the words contained in the sentences form the nodes in the word graph. A word graph is a directed graph where an edge from word A to word B represents an adjacency relation. It also contains start and end nodes (i.e., punctuation). Part of speech information is used to prevent verbs and nouns from being merged in the word graph which would result in the summarization sentence having ungrammatical sequences. Edges within the word graph are used to connect words that are adjacent in a sentence with the edge weight incremented by 1 each time a word occurs after another in a sequence.

Filippova [3] says that good sentence compression goes through all the nodes which represent important concepts but does not pass the same node several times. This is achieved by inverting the edge weights and finding the *K* shortest paths from the start to the end node in the word graph that don't include a verb. The path

Published in CEUR-WS:
Proceedings of the VISLA15 Workshop on Visual Aspects of Learning Analytics (Duval et al., eds), organized at LAK15.

through the graph with the minimum total weight is selected as the summary sentence. Additional graph scoring and ranking metrics are used to take into consideration strong links between words and determine salient words.

3. VISUAL ANALYTICS BASED ON MULTI-SENTENCE COMPRESSION

Graph-based multi-sentence compression produces *K* candidate summary sentences, with the sentence with the minimum shortest path score being selected as the summary. Within the context of applying the algorithm to develop visual analytics for short answer questions, we propose to use all *K* candidates because difference common pathways are captured and these may have branches that identify different concepts or vocabulary being used by students.

The following 3 approaches are being considered as summarization and visualization tools for short answer questions:

- Approach 1: Display the K candidate sentences that are derived from the Filippova [3] algorithm. This is only a textual display of the sentences.
- Approach 2: Construct a graph from the *K* candidate sentences and use a graph layout algorithm to display the graph is a visual manner [5]. The advantage over the textual display of the sentence is that loops of words and branches between words would be more easily identifiable.
- Approach 3: Display the full word graph and highlight the *K* candidate paths (sentences) on the graph display. This visualization would allow the lecturer/tutor to see the range of words used. Approach 3 is not presented in this paper but will be evaluated in future work.

4. INITIAL INVESTIGATION

An initial investigation on using the Filippova [3] algorithm to produce a summary and visualization of student responses to short answer questions has been conducted using the open dataset provided by Mohler and Mihalcea [6]. The dataset consists of three assignments of seven short answer questions each given to an introductory computer science class at the University of North Texas. Each assignment includes the question, the teachers answer, and the student responses (usually a few short sentences).

An open source implementation of the Filippova [3] multi-sentence compression algorithm, from the Takahe library (https://github.com/boudinfl/takahe) was used. Tokenization and part of speech tagging was done using NLTK [1]. Numerous spelling errors were noted, but were not fixed for the initial investigations. The minimum number of words in the derived the compressions was set to 6. The number of sentence candidate generated for the 2 examples shown in this paper was 10. The visualization for each of the examples was created using the Yifan Hu [5] Layout in Gephi.

The top 10 summary sentences (Approach 1) and the graph visualization of the word graph constructed from the summary sentences (Approach 2) is included for 2 questions from the Mohler and Mihalcea [6] dataset in Section 4.1 and 4.2.

Initial results show that the summary candidate sentences provide a good overview of the common concepts used by students. The graph visualization of the word graph also shows multiple branches and loops.

4.1 Example 1

The summary candidate sentences in the first example shows that most students have included the 2 key similarities between iteration and recursion related to a termination condition and that both can execute infinitely. Students however have not used the word “repetition” but refer to programming code syntax (i.e., control statement).

Question: What are the similarities between iteration and recursion?
Teachers Answer: They both involve repetition; they both have termination tests; they can both occur infinitely.

Table 1. Top 10 candidate summary sentences for Example 1

Score	Candidate Summary Sentence
0.016	both are based on control statement.
0.015	both are based on a control statement.
0.023	they are based on control statement.
0.021	they are based on a control statement.
0.021	both are based on control statement, termination test.
0.02	both are based on control statement both can infinitely.
0.02	both are based on a control statement , termination test.
0.017	both are based on control statement , both involve a termination test.
0.019	both are based on a control statement , both can infinitely.
0.023	based on control statement , both can occur infinitely.

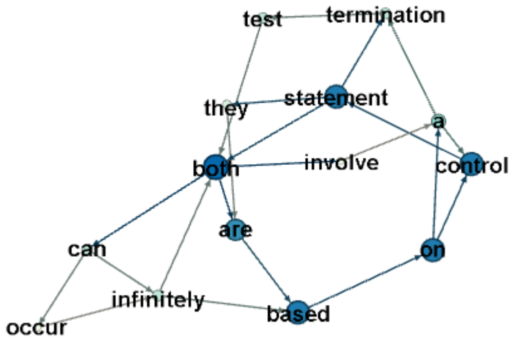


Figure 1. Graph visualization of the top 10 summary sentence candidates in Example 1.

4.2 Example 2

In the second example, very few students include "abstraction" in their answer or concepts that would be associated with “abstraction” such as “encapsulation”. Most students mention reusability and maintenance/debugging but it is actually “abstraction” that leads to easier maintenance/debugging of object oriented programming code. The proposed visualizations would therefore allow the lecturer/tutor to identify the concepts that the students have missed or explained incorrectly and guide the lecturer in providing feedback.

Question: What are the main advantages associated with object-oriented programming?
Teachers Answer: Abstraction and reusability.

Table 2. Top 10 candidate summary sentences for Example 2

Score	Candidate Summary Sentence
0.025	existing classes can be reused program.
0.025	existing classes can be reused program maintenance.
0.023	existing classes can be reused program maintenance and verification are easier.
0.042	objects can be reused program maintenance.
0.04	existing classes can be reused and program.
0.038	existing classes can be reused and program maintenance.
0.05	the classes can be reused program .
0.034	objects can be reused program maintenance and verification are easier.
0.047	the classes can be reused program maintenance.
0.059	objects can be reused and program.

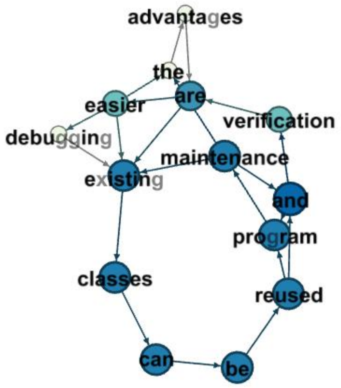


Figure 2. Graph visualization of the top 10 summary sentence candidates in Example 2.

5. TOOL DESIGN AND FUNCTIONALITY

We intend to create an open source tool that incorporates the sentence compression algorithm and the proposed visualizations described in Section 3 that will be made available on Github. The tool will allow lecturers to view student responses that match word graph loops and branches. This will help lecturers to determine context by viewing exemplar student responses. The tool will also allow lecturers to attach feedback to nodes and paths in the word graph as a means of providing specific and targeted feedback to students. Integration with quiz tool export formats from popular Learning Management Systems is also planned.

6. PROPOSED EVALUATION

A between subjects comparative study is being planned. The study will be comprised of two groups. Group A will be required to read all student responses and identify student knowledge gaps. Group B will use the visualizations produced from the output of sentence compression to identify knowledge gaps in the student responses. Identified knowledge gaps from Group A and Group B will then be compared.

Participants in Group B will be shown all 3 approaches described in Section 3 and asked to rate each approach based on principles of visual analytics.

7. CONCLUSION

In this paper, ideas on using graph-based multi-sentence compression as the basis for the visual analysis of student responses to short answer questions were explored. The multi-sentence compression algorithm was introduced and ideas for potential visualizations were discussed. Example visualizations were then presented along with plans to embed visualizations with in an open source tool that is able to integrate with quiz responses from Learning Management Systems. Preliminary results indicate that visualizations derived from the output of sentence compression are able to allow lecturers to identify knowledge gaps and provide feedback to groups of students with similar knowledge gaps. In the future an evaluation of the visualizations will be conducted. The keyphrase extraction algorithm for reranking summary sentences [2] and the Opinosis algorithm [4] will also be evaluated in addition to the Filippova algorithm [3].

8. REFERENCES

[1] Bird, S., Edward L., & Ewan K. 2009. Natural Language Processing with Python. O'Reilly Media Inc.

[2] Boudin, F., & Morin, E. 2013. Keyphrase Extraction for N-best Reranking in Multi-Sentence Compression. In Proccedings of the NAACL HLT 2013 conference.

[3] Filippova, K. 2010. Multi-sentence compression: Finding shortest paths in word graphs. In Proceedings of the 23rd International Conference on Computational Linguistics (pp. 322-330). Association for Computational Linguistics.

[4] Ganesan, K., Zhai, C., & Han, J. 2010. Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions. In Proceedings of the 23rd International Conference on Computational Linguistics (pp. 340-348). Association for Computational Linguistics. Chicago.

[5] Hu, Y. F. 2005. Efficient and high quality force-directed graph drawing. The Mathematica Journal, 10 (37-71).

[6] Mohler, M., & Mihalcea, R. 2009. Text-to-text Semantic Similarity for Automatic Short Answer Grading, in Proceedings of the European Chapter of the Association for Computational Linguistics (EACL 2009), Athens, Greece.

Getting a grasp on tag collections by visualising tag clusters based on higher-order co-occurrences

Katja Niemann, Sarah León Rojas,
Martin Wolpers
Fraunhofer FIT
Schloss Birlinghoven
53754 Sankt Augustin, Germany
{katja.niemann, sarah.leon.rojas,
martin.wolpers}@fit.fraunhofer.de

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

ABSTRACT

Tagging learning resources in repositories or web portals offers a way to meaningfully describe these resources. The more tags there are, however, the more difficult it is to find one's way around the repository, especially when they are user-generated free-text tags. This paper therefore presents a visualisation of tag clusters based on higher-order co-occurrences that allows users of such repositories a plain but simple way of exploring them in an intuitive manner.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Clustering, Information filtering, Search process, Selection process*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing; J.1 [**Administrative Data Processing**]: —*Education*; K.3.1 [**Computers and Education**]: Computer Uses in Education

General Terms

Algorithms, Experimentation, Language

Keywords

clustering, higher-order co-occurrences, tags, technology enhanced learning, visualisation

1. INTRODUCTION

Many educational web portals allow users to manually enrich the offered learning resources with social metadata like comments and free-text tags. It has been shown that tags in particular provide powerful knowledge that can be used to improve the quality of searching and recommendations [4, 7]. Similar to automatically extracted keywords, tags thus offer a way to get a quick grasp on the content or theme of multimedia objects. Especially when dealing with multimedia objects that provide little or no textual context (e.g. photos or videos) tags provide meaningful descriptors of these objects [8].

A common problem, however, when relying on tags is that they are often user-generated and not restricted to a closed vocabulary. Different users can tag the same learning resource with different tags leading to a large collection of rarely used but highly related tags. The use of singular or plural versions of the same word, the same word in different languages or different words with the same meaning, i.e. synonyms, can also lead to problems when relying on tags in order to get an overview on a collection of learning resources. In order to detect unknown relations between tags they therefore need to be contextualised.

Based on an approach of visualising large document collections according to the documents' keywords [6] we suggest to use a visualisation of tag relations that allows users to quickly get a grasp of the resources offered by a learning portal and to dig deeper to get an understanding of certain subject areas. Instead of clustering the learning objects according to their content, however, we cluster the tags according to their higher-order co-occurrences and then present them in a clearly arranged and intuitive manner. The creation of higher-order co-occurrences is a well-known approach in corpus linguistics to discover semantic relations between words based on their usage in text documents [2]. We adapt this approach by analysing the assignments of tags to learning resources instead of the occurrences of terms in sentences or text documents.

The paper is structured as follows. Chapter 2 gives a short overview of related work. Chapter 3 describes the approach of higher-order co-occurrence clustering to group tags with similar meanings, followed by the description of the MACE data set in chapter 4 which is used in this paper. Thereafter, chapter 5 describes the visualisation of the tag clusters and chapter 6 discusses the results. Finally, chapter 7 holds a conclusion and an outlook on future work.

2. RELATED WORK

According to Rivadeneira et al. [5], a meaningful visualisation of tags supports four main functions: (1) search, i.e. tags can be directly included in the search process and, thus, enhance the findability of items, (2) browsing, i.e. the visualisation offers a central entry point for users that know what

Published in CEUR-WS:
Proceedings of the VISLA15 Workshop on Visual Aspects of Learning Analytics (Duval et al., eds), organized at LAK15.

they are looking for but not what exactly to search for, (3) impression formation / gisting, i.e. the visualisation allows users to get a quick grasp on the items' subject areas, and (4) recognition, i.e. the users are offered the possibility to understand different aspects of certain information.

The most common approach to visualise a large number of tags is the creation of tag clouds. Here, the relative size of each tag stands in relation to its frequency in the tag collection. Nowadays, many tools are available that allow an easy integration of personalised tag clouds in web sites, e.g. TagCrowd[1] and Wordle[2]. While there is a huge potential inherent in tag clouds they also suffer from some issues, e.g. the missing semantic between the visualised tags [1, 9]. In order to deal with this, tag clouds have been created that analyse (first-order) co-occurrences between the tags and group tags that often co-occur [11]. Here, *similar* tags do not necessarily reference to the same semantic concept but are linked by the resources they have in common [3]. Another problem of tag clouds is that many frequent tags often dominate the whole tag cloud and less frequent tags and their concepts get lost [1].

This paper presents a clustering approach for tags that is based on higher-order co-occurrences, i.e. a corpus linguistic technique to find semantically related terms [2]. This way we aim to discover and visually cover all subject areas even though it might not be possible to display all single tags.

3. HIGHER-ORDER CO-OCCURRENCE CLUSTERING OF TAGS

The creation of higher order co-occurrences is a corpus-linguistic approach to exploit the usage context of linguistic entities in order to find semantic relations. Two linguistic entities are defined to be co-occurrences if they occur in at least one common usage context, e.g. in a sentence. For example, the word *dog* often co-occurs with the words *bark*, *growl*, and *sniff* among others.

In order to calculate the significance of a co-occurrence statistical association measures are used. Thereafter, the most significant co-occurrences must be selected for each term. Since there is no standard scale of measurement to draw a clear distinction between significant and non-significant occurrences, there are two ways to do so, i.e. by selecting only the n most significant co-occurrences for each resource or by using a threshold.

The significant co-occurrences of an entity form its first-order co-occurrence class and entities which co-occur in first-order co-occurrence classes are second-order co-occurrences. These second-order co-occurrence classes again can be used as input to calculate third order co-occurrences and so forth. When this procedure is repeated several times, the higher-order co-occurrence classes tend to get stable, i.e. their elements do not change any more. This indicates that there exist universal relations between the entities in the remaining classes that induce their aggregation again in each iteration step. In fact, these stable higher-order co-occurrence classes have shown to usually hold semantically related entities. Heyer et al. [2] show this for the co-occurrences of *IBM*, among other words. Their investigations are based on text corpora collected for the portal wortschatz.uni-leipzig.de, the German treasury of words. The first co-occurrence class is rather heterogeneous, and contains words like *computer manufacturer*, *stock exchange*, *global* and so on. After some iterations of computing higher-order co-occurrence classes, however, the classes become more homogenous and stable. The tenth order co-occurrence class only contains names of other computer-related companies like *Microsoft*, *Sony* etc.

In the given scenario we do not have sentences in which the tags occur. However, the tags are assigned to learning resources which can be considered to represent usage contexts. Thus, two tags are co-occurrences if they are assigned to at least one common learning resource. In order to calculate the significance of two tags, the association measure Mutual Information (MI) is used which compares the observed frequency O of a co-occurrence with its expected frequency E, see formula 1.

$$\mathrm{MI} = log_2 \frac{O}{E} \qquad (1)$$

Here, selecting the n most significant co-occurrences for each tag would imply to have a pre-defined cluster size which is not desirable, thus, a threshold is used. Because the calculated significance scores for resource pairs are only comparable if they have one resource in common, a resource-specific threshold is used to distinguish between relevant and non-relevant co-occurrences. Here, this threshold is calculated for each learning resource by averaging the significance values of all its co-occurrences and multiplying the result with a regulation constant α which has a value of 0.95 in the presented experiment.

4. THE MACE DATA SET

The MACE[3] (Metadata for Architectural Contents in Europe) project relates digital learning resources about architecture with each other across repository boundaries to enable a simplified discovery and access [10]. Users are able to search for learning resources and filter the results, e.g. according to their language, the original repository, and the classification terms they hold. Furthermore, the portal offers a social search based on tags, a location search based on the geographical coordinates of buildings represented through learning resources, and a competence search based on the competencies the learning resources aim to impart. Registered and logged-in users are able to rate, tag, and comment on learning resources. Additionally, they can follow the metadata provision activities of other users.

The MACE data set holds 117,907 events on 12,442 learning resources conducted by 630 registered users. 70.8% of the learning resources hold tags in which each tagged learning resource holds on average 6.59 tags. Overall, the users assigned 13,291 distinct tags of which 73% are only used once and only about 4% of the tags are added to more than 10 learning resources.

---

[1] http://tagcrowd.com/

[2] http://wordle.net/

[3] http://mace-project.eu/

5. VISUALISATION

When creating a visualisation of the tag clusters for the MACE data set we decided to not present tags in the visualisation that are assigned to only one or two learning resources. Only clusters that hold more than five tags are selected for presentation. Finally, the two most frequent tags are selected as title for each cluster. If a cluster's most frequent tags significantly overlap, the less frequent one is neglected and the next frequent tag is selected.

After this data processing, the tag clusters and all attached information are written to a JSON file. The visualisation is realised using the Data-Driven Documents D3.js framework[4], 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 the visualisation[5]. The tag clusters are represented by 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. Here, the size of a tag cluster depends on the number of learning objects that are referenced by the tags belonging to it. Additionally to size and position, every cluster has its own color and is labelled with its two most representing tags to enable the users to quickly get a grasp on the clusters' content.

By clicking on a cluster, the view changes and the visualisation zooms into to the chosen cluster for which up to 20 tags become visible. We chose this number to not overload the visualisation. In order to continue the circle approach used for the clusters, we adapted the common usage of font size, coloring and word positioning in tag clouds and used sized and spirally ordered circles for the tags as well. On the right side of the visualisation, a list of all the learning resources that are associated with that cluster is given showing the resources' title, media type, and language additionally to the list of all tags assigned to it. All resource titles link to the original resource.

Clicking on a tag circle results in a new list next to the visualisation in which all resources assigned with that tag are given. By clicking on a specific tag, its circle is highlighted and the object list only displays those resources that are assigned with the highlighted tag, see figure 2.

6. DISCUSSION

This chapter provides an insight on the eleven clusters shown in the visualisation, discusses the topics they cover including their relations, and reference further distinctive features. The following cluster descriptions are ordered by the size of the clusters, i.e. from the outside of the spiral to its center. Whenever needed, the tags' English translations are given in brackets. Here, if two tags hold the same English translation it is only given once.

Cluster 1: cubierta / aislante (cover and insulation). This cluster's tags, which are mainly in Spanish, name methods, objects, and materials used for insulation, e.g. *cobertes (covered)*, *paneles (panels)*, as well as *poliestireno (polystyrene)* and reference 683 distinct resources. The resources' descriptions hold further tags that can be used to orientate in this field. For example, figure 3 shows an excerpt from the list of resources that are assigned with the tag *sandwich*. While this tag might be unexpected at a first glance, the tags it was used with clarify its meaning, i.e. a (panel) structure made of three layers. Overall, 2,190 distinct tags are given in the resource list of this cluster.

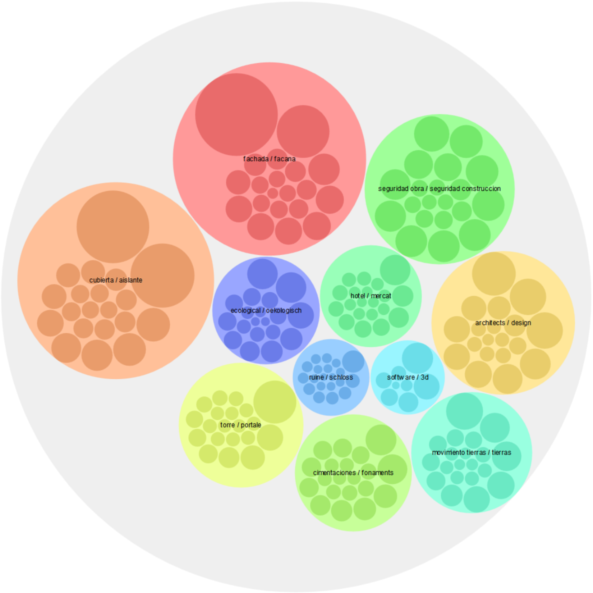


Figure 1: Start view

Cluster 2: fachada / facana (facade). This cluster mainly holds Spanish and Catalan tags that deal with the construction and cladding of buildings, e.g. *sistemas constructivos (building systems), cerramientos (enclosure), gres (stoneware)*, and *constructivos (building)*. Overall, this cluster's tags reference 661 distinct resources that are assigned with 2,080 distinct tags.

Cluster 3: seguridad obra / seguridad construccion (work and construction safety). This cluster holds a mix of Spanish and English tags that deal with security, e.g. *seguridad trabajador (worker safety), construction security, sistemas de seguridad (security systems)*, and *normativa (regulations)*. Overall, it references 532 distinct resources that hold 634 distinct tags.

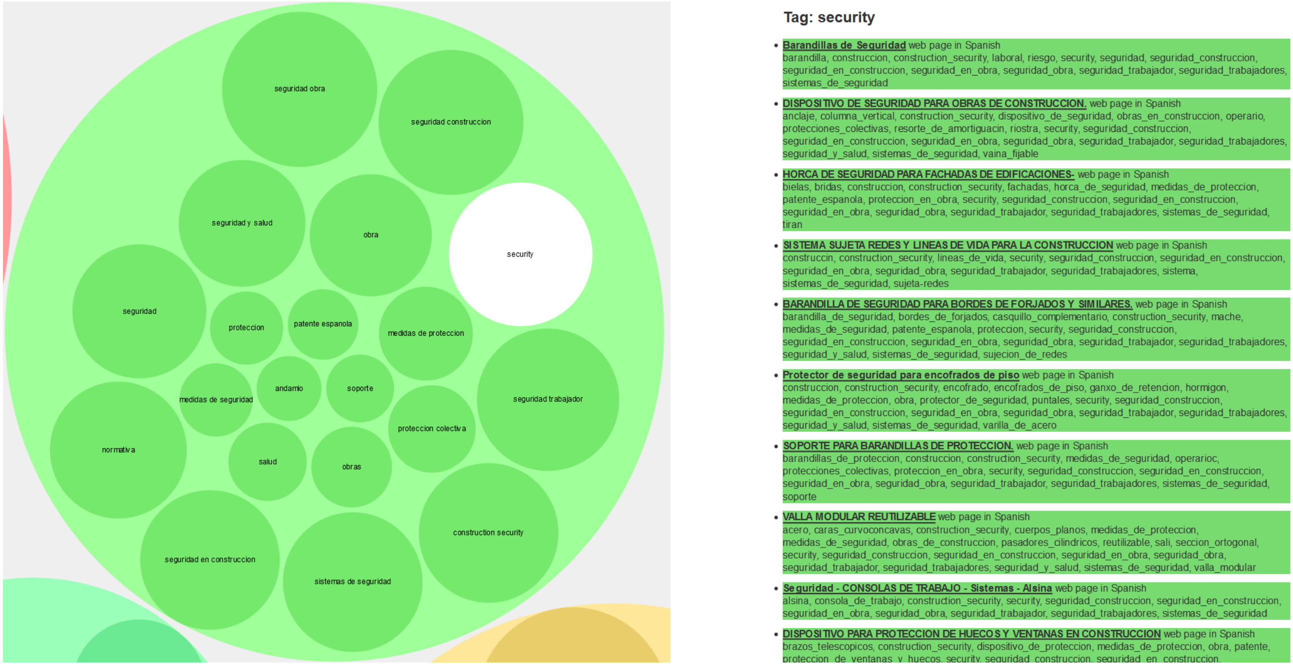
Cluster 4: architects / design. The first cluster that mainly holds English tags and few Spanish ones deals with (green) architecture in the public space, e.g. *architecture, museum, green architecture, architettura (architecture), piazza*, and *bioarchitettura*. It references 296 distinct resources that hold 962 distinct tags.

Cluster 5: movimiento tierras / tierras (land movement). This cluster comprises Spanish tags that deal with the preparation of building zones, e.g. *excavaciones (diggings), maquinaria (machinery), calculo (calculation)*, and

---

[4] http://d3js.org/

[5] The visualisation is available at http://mitarbeiter.fit.fraunhofer.de/˜niemann/VisLA/



Tag: security

- **Barandillas de Seguridad** web page in Spanish
  barandilla, construccion, construction_security, laboral, riesgo, security, seguridad, seguridad_construccion, seguridad_en_construccion, seguridad_en_obra, seguridad_obra, seguridad_trabajador, seguridad_trabajadores, sistemas_de_seguridad
- **DISPOSITIVO DE SEGURIDAD PARA OBRAS DE CONSTRUCCION.** web page in Spanish
  anclaje, columna_vertical, construction_security, dispositivo_de_seguridad, obras_en_construccion, operario, protecciones_colectivas, resorte_de_amortiguacin, riostra, security, seguridad_construccion, seguridad_en_construccion, seguridad_en_obra, seguridad_obra, seguridad_trabajador, seguridad_trabajadores, seguridad_y_salud, sistemas_de_seguridad, vaina_fijable
- **HORCA DE SEGURIDAD PARA FACHADAS DE EDIFICACIONES-** web page in Spanish
  bielas, bridas, construccion, construction_security, fachadas, horca_de_seguridad, medidas_de_proteccion, patente_espanola, proteccion_en_obra, security, seguridad_construccion, seguridad_en_construccion, seguridad_en_obra, seguridad_obra, seguridad_trabajador, seguridad_trabajadores, sistemas_de_seguridad, tiran
- **SISTEMA SUJETA REDES Y LINEAS DE VIDA PARA LA CONSTRUCCION** web page in Spanish
  construccin, construction_security, lineas_de_vida, security, seguridad_construccion, seguridad_en_construccion, seguridad_en_obra, seguridad_obra, seguridad_trabajador, seguridad_trabajadores, sistema, sistemas_de_seguridad, sujeta-redes
- **BARANDILLA DE SEGURIDAD PARA BORDES DE FORJADOS Y SIMILARES.** web page in Spanish
  barandilla_de_seguridad, bordes_de_forjados, casquillo_complementario, construction_security, mache, medidas_de_seguridad, patente_espanola, proteccion, security, seguridad_construccion, seguridad_en_construccion, seguridad_en_obra, seguridad_obra, seguridad_trabajador, seguridad_trabajadores, seguridad_y_salud, sistemas_de_seguridad, sujecion_de_redes
- **Protector de seguridad para encofrados de piso** web page in Spanish
  construccion, construction_security, encofrado, encofrados_de_piso, ganxo_de_retencion, hormigon, medidas_de_proteccion, obra, protector_de_seguridad, puntales, security, seguridad_construccion, seguridad_en_construccion, seguridad_en_obra, seguridad_obra, seguridad_trabajador, seguridad_trabajadores, seguridad_y_salud, sistemas_de_seguridad, varilla_de_acero
- **SOPORTE PARA BARANDILLAS DE PROTECCION.** web page in Spanish
  barandillas_de_proteccion, construccion, construction_security, medidas_de_seguridad, operarioc, protecciones_colectivas, proteccion_en_obra, security, seguridad_construccion, seguridad_en_construccion, seguridad_en_obra, seguridad_obra, seguridad_trabajador, seguridad_trabajadores, sistemas_de_seguridad, soporte
- **VALLA MODULAR REUTILIZABLE** web page in Spanish
  acero, caras_curvoconcavas, construction_security, cuerpos_planos, medidas_de_proteccion, medidas_de_seguridad, obras_de_construccion, pasadores_cilindricos, reutilizable, sali, seccion_ortogonal, security, seguridad_construccion, seguridad_en_construccion, seguridad_en_obra, seguridad_obra, seguridad_trabajador, seguridad_trabajadores, seguridad_y_salud, sistemas_de_seguridad, valla_modular
- **Seguridad - CONSOLAS DE TRABAJO - Sistemas - Alsina** web page in Spanish
  alsina, consola_de_trabajo, construction_security, security, seguridad_construccion, seguridad_en_construccion, seguridad_en_obra, seguridad_obra, seguridad_trabajador, seguridad_trabajadores, sistemas_de_seguridad
- **DISPOSITIVO PARA PROTECCION DE HUECOS Y VENTANAS EN CONSTRUCCION** web page in Spanish
  brazos_telescopicos, construction_security, dispositivo_de_proteccion, medidas_de_proteccion, obra, patente, proteccion_de_ventanas_y_huecos, security, seguridad_construccion, seguridad_en_construccion,

Figure 2: Zoomed cluster view with a selected tag

excavadora (excavator). It references 278 distinct resources that comrpise 1,201 distinct tags.

Cluster 6: cimentaciones / fonaments (foundation). This cluster mainly holds in Spanish and Catalan tags that deal with the construction and anchoring of buildings, e.g. *muro (wall), building, terreno (ground), zapatas (shoes)*, and *anclajes (anchors)*. Overall, this cluster's tags reference 210 distinct learning resources that are assigned with 831 distinct tags.

Cluster 7: torre / portale (tower and portal). The main topic of this cluster is sustainability although its two most frequent tags do not imply it. Further tags are e.g. *bio edilizia (bio building), solar*, and *sostenibilidad (sustainability)*. However, it can be seen in the resource list that the learning resources that are tagged with *torre* or *portale* also deal with this topic, e.g. the insulation of towers. Thus, this cluster exhibits a topical relation to the first one but in contrast, in mainly contains Italian tags. Overall, the cluster references 201 distinct resources and its resource list comprises 661 distinct tags.

Cluster 8: ecological / oekologisch. This cluster also deals with sustainability but with a stronger focus on the generation and recovery rather than on the conservation of energy. Furthermore, it mainly comprises German tags, e.g. *photovoltaikanlage (photovoltaic power station), waermerueckgewinnung (heat recovery)*, and *waermepumpe (heat pump)*. The cluster references 200 distinct resources that are assigned with 825 distinct tags.

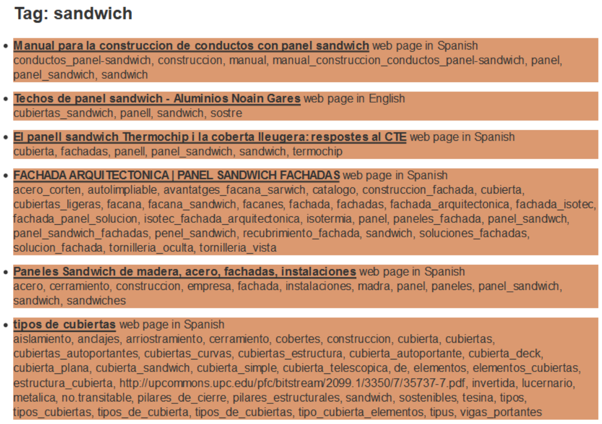
Cluster 9: hotel / mercat (hotel and market). This cluster holds tags that reference resources dealing with (aesthetic) buildings in the in public space like *puente (bridge), rascacielos (skyscraper), puerto (harbour)*, and *hotel arts* as well as famous architects of those buildings, e.g. *Santiago Calatrava Valls* and *Norman Robert Foster*. Overall, the clusters references distinct 86 learning resources that comprise 107 distinct tags.

Cluster 10: software / 3d. The only cluster that contains less than 20 tags deals with the design of buildings using the computer and comprises tags like *cad (computer-aided design), rhino3d (CAD Software), tutorial*, and *programming*. The cluster references 72 distinct resources that are assigned with 274 distinct tags.

Cluster 11: ruine / schloss (ruin and castle). This cluster references learning resources that describe or depict buildings built in the *mittelalter (middle ages)* or *hochmittelalter (high middle ages)* in German regions like *pfealzer wald (Palatinate Forest)* and *rhein-lahn-kreis (Rhine Lahn circle)*. Consequently, all tags are in German. Overall, they reference 51 distinct resources that hold 96 distinct tags.

Concluding, the clusters mostly contain tags that indeed belong to the same subject area, though, they are not completely separated. For example, several clusters deal with sustainability. However, their tags are in different languages and they have different focuses, e.g. the generation vs. the conservation of energy or public vs. private buildings. Furthermore, this shows that sustainability is an important field in architecture. The other clusters reference resources that describe different construction phases (design of buildings, preparation of building zones, as well as construction and cladding of buildings), security issues, and notable buildings as study objects.

In numbers, the tags that hold their own circles in the visualisation reference 2,849 distinct learning resources, i.e. a third of all tagged learning resources in the MACE data set.



Tag: sandwich

- **Manual para la construccion de conductos con panel sandwich** web page in Spanish
  conductos_panel-sandwich, construccion, manual, manual_construccion_conductos_panel-sandwich, panel, panel_sandwich, sandwich
- **Techos de panel sandwich - Aluminios Noain Gares** web page in English
  cubiertas_sandwich, panell, sandwich, sostre
- **El panell sandwich Thermochip i la coberta lleugera: respostes al CTE** web page in Spanish
  cubierta, fachadas, panell, panel_sandwich, sandwich, termochip
- **FACHADA ARQUITECTONICA | PANEL SANDWICH FACHADAS** web page in Spanish
  acero_corten, autolimpliable, avantatges_facana_sarwich, catalogo, construccion_fachada, cubierta, cubiertas_ligeras, facana, facana_sandwich, facanes, fachada, fachadas, fachada_arquitectonica, fachada_isotec, fachada_panel_solucion, isotec_fachada_arquitectonica, isotermia, panel, paneles_fachada, panel_sandwch, panel_sandwich_fachadas, penel_sandwich, recubrimiento_fachada, sandwich, soluciones_fachadas, solucion_fachada, tornilleria_oculta, tornilleria_vista
- **Paneles Sandwich de madera, acero, fachadas, instalaciones** web page in Spanish
  acero, cerramiento, construccion, empresa, fachada, instalaciones, madra, panel, paneles, panel_sandwich, sandwich, sandwiches
- **tipos de cubiertas** web page in Spanish
  aislamiento, anclajes, arriostramiento, cerramiento, cobertes, construccion, cubierta, cubiertas, cubiertas_autoportantes, cubiertas_curvas, cubiertas_estructura, cubierta_autoportante, cubierta_deck, cubierta_plana, cubierta_sandwich, cubierta_simple, cubierta_telescopica, de, elementos, elementos_cubiertas, estructura_cubierta, http://upcommons.upc.edu/pfc/bitstream/2099.1/3350/7/35737-7.pdf, invertida, lucernario, metalica, no.transitable, pilares_de_cierre, pilares_estructurales, sandwich, sostenibles, tesina, tipos, tipos_cubiertas, tipos_de_cubierta, tipos_de_cubiertas, tipo_cubierta_elementos, tipus, vigas_portantes

Figure 3: Excerpt of the resource list for the tag sandwich

While this number seems small at a first glance, it is quite high when considering that only about 3% of the tags hold their own circles. However, this number can be increased by presenting all resources referenced by a tag that was assigned to a cluster in the visualisation. So far, the tags that do not belong to the clusters' 20 most frequents ones are neglected.

Overall, the referenced learning resources are assigned with 6,585 distinct tags (i.e. half of all tags) which are shown in the resource lists. Considering that about 70% of the tags are only used once, this seems to be an acceptable number. Furthermore, it will be increased as well as soon as more resources are displayed.

7. CONCLUSION AND FUTURE WORK

In summary, the visualisation of the tag clusters gives a broad and easily understandable overview on the learning resources' subject areas. Furthermore, it enables the users to explore the data set by zooming into the clusters and browsing the result lists.

This visualisation, though, is not intended to be a standalone tool for the exploration of a data set. It is rather meant to be an additional tool that can be integrated with (already available) search functions like a faceted search or a social search as offered by the MACE portal. This way, the displayed resources could for example be filtered according to their language or media type and the tags in the resources' descriptions could be used to search for resources assigned with one ore more specific tags. Furthermore, the visualisation offers several possibilities for extensions. For example, by clicking on a learning resource in a tag's or a cluster's resource list, all tags that are assigned to this resource but are located in other clusters could be highlighted. This would further enhance the ability to discover relations between tags and, thus, between subject areas. Another option would be to allow the users to browse all tags belonging to one cluster and not only the most frequent ones.

So far, no evaluation has been conducted. In order to do so, the tag cluster visualisation needs to be integrated in a web portal. Thereafter, the acceptance of this visualisation can be evaluated by analysing its usage or by conducting a survey. Furthermore, user studies with control groups can be conducted to investigate if the use of the tag cluster visualisation increases the orientation in the portal or the performance of the students when solving tasks.

8. ACKNOWLEDGMENTS

The work presented in this paper has been supported by the Open Discovery Space that is funded by the European Commissio's CIP-ICT Policy Support Program (Project Number: 297229).

9. REFERENCES

[1] M. A. Hearst and D. Rosner. Tag clouds: Data analysis tool or social signaller? In *Proc. of the 41st Annual Hawaii International Conference on System Sciences*, HICSS '08, pages 160–, Washington, DC, USA, 2008. IEEE Computer Society.

[2] G. Heyer, U. Quasthof, and T. Wittig. *Text Mining: Wissensrohstoff Text. Konzepte, Algorithmen, Ergebnisse.* W3L GmbH, 2006.

[3] O. Kaser and D. Lemire. Tag-cloud drawing: Algorithms for cloud visualization, 2007.

[4] S. Lohmann, S. Thalmann, A. Harrer, and R. Maier. Learner-Generated Annotation of Learning Resources - Lessons from Experiments on Tagging. In *Proc. of the International Conference on Knowledge Management (I-KNOW 2008)*, pages 304–312, 2008.

[5] A. W. Rivadeneira, D. M. Gruen, M. J. Muller, and D. R. Millen. Getting our head in the clouds: Toward evaluation studies of tagclouds. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, pages 995–998, New York, NY, USA, 2007. ACM.

[6] M. Scheffel, K. Niemann, S. Leon Rojas, H. Drachsler, and M. Specht. Spiral me to the core: Getting a visual grasp on text corpora through clusters and keywords. In K. Yacef and H. Drachsler, editors, *Proc. of the Workshops at the LAK 2014 Conference*, volume 1137 of *CEUR Proc.*, Indianapolis, Indiana, USA, 2014.

[7] S. Sen, J. Vig, and J. Riedl. Tagommenders. In *Proc. of the 18th international conference on World wide web (WWW '09)*, pages 671–680, New York, New York, USA, 2009. ACM Press.

[8] B. Sigurbjörnsson and R. van Zwol. Flickr tag recommendation based on collective knowledge. In *Proc. of the 17th international conference on World Wide Web - WWW '08*, pages 327–336, New York, New York, USA, 2008. ACM Press.

[9] J. Sinclair and M. Cardew-Hall. The folksonomy tag cloud: When is it useful? *J. Inf. Sci.*, 34(1):15–29, Feb. 2008.

[10] M. Stefaner, E. D. Vecchia, M. Condotta, M. Wolpers, M. Specht, S. Apelt, and E. Duval. MACE - Enriching Architectural Learning Objects for Experience Multiplication. In E. Duval, R. Klamma, and M. Wolpers, editors, *Proc. of the 2nd European Conference on Technology Enhanced Learning (EC-TEL '07)*, volume 4753 of LNCS, pages 322–336, Berlin, Heidelberg, 2007. Springer.

[11] M. Steinbach, G. Karypis, and V. Kumar. A comparison of document clustering techniques. In *In KDD Workshop on Text Mining*, 2000.

A Network Based Approach for the Visualization and Analysis of Collaboratively Edited Texts

Tobias Hecking
University of Duisburg-Essen
Lotharstraße 63/65
47048 Duisburg, Germany
hecking@collide.info

H. Ulrich Hoppe
University of Duisburg-Essen
Lotharstraße 63/65
47048 Duisburg, Germany
hoppe@collide.info

ABSTRACT

This paper describes an approach for network text analysis and visualization for collaboratively edited documents. It incorporates network extraction from texts where nodes represent concepts identified from the words in the text and the edges represent relations between the concepts. The visualization of the concept networks depicts the general structure of the underlying text in a compact way. In addition to that, latent relations between concepts become visible, which are not explicit in the text. This work concentrates on evolving texts such as wiki articles. This introduces additional complexity since dynamic texts lead to dynamic concept networks. The presented method retains the user information of each revision of a text and makes them visible in the network visualization. In a case study it is demonstrated how the proposed method can be used to characterize the contributors in collaborative writing scenarios regarding the nature of concept relations they introduce to the text.

General Terms

Algorithms, Visualization, Experimentation

Keywords

Network Visualization, Network Analysis, Natural Language Processing, Collaborative Writing, Learning Analytics

1. INTRODUCTION

Network text analysis is the task of extraction and analysis of networks from text corpora. In those networks the nodes are concepts identified from the words in the text and the edges between the nodes represent relations between the concepts. The visualization of concept networks can help to depict the general structure of the underlying text in a compact way. In addition to that, latent relations between concepts become visible, which are not explicit in the text. Thus, approaches for visualizing texts as networks allow analysts to concentrate on important aspects without reading large amounts of the texts. Several network analysis techniques can be applied to identify important concepts, perform concept clustering, as well as comparative analysis of different texts [11].

Existing applications for network text analysis include the identification of key phrases [10], mining of relations between real world entities [6], as well as the extraction of complete concept ontologies and concept maps with labelled edges [18].

This work concentrates on the relations between concepts that can be found in evolving and collaboratively edited texts such as wiki articles. This introduces additional complexity since dynamic texts lead to dynamic concept networks. The presented method retains the user information of each revision of a text which allows for characterizing the contributors in collaborative writing scenarios regarding the nature of concept relations they introduce to the text. The resulting visualization is a concept network with colored edges where each edge color is allocated uniquely to a specific contributor. In further analysis steps, network centrality measures are calculated that give additional information about the contribution of each editor.

The outline of this paper is as follows: Section 2 gives the theoretical background of this work and highlights significant research work in the area of network text analysis. The general idea of our visualization and analysis approach is presented in section 3. Section 4 focuses on the concrete implementation. This incorporates the applied natural language processing chain, as well as the description of network analysis methods.

2. Background

2.1 Collaborative Writing Activities in Education

Collaborative writing activities are a common task in educational scenarios [3, 13]. Users can learn actively by creating artefacts but can also learn passively by consuming artefacts created by others [14].

It could be shown that user generated content is relevant to learners in addition to tutor provided content [13]. With the emergence of online communities such as Wikipedia collaborative knowledge building takes place with open scale in terms of the number of contributors. There is some evidence that individual and collective knowledge co-evolves through collaborative editing of epistemic artefacts in open online environments [9]. In general collaborative writing requires different rhetorical and organizational skills of the editors [8], and thus, the learner generated artefacts are a valuable data source for analysis.

This motivates the development of methods that makes collaborative writing processes visible in order to understand and improve the application of collaborative text writing in educational settings.

2.2 Visualization Approaches for Collaborative Writing

Several methods have been developed to represent evolving texts with multiple editors in a visual way. One of the first approaches

Published in CEUR-WS:
Proceedings of the VISLA15 Workshop on Visual Aspects of Learning Analytics (Duval et al., eds), organized at LAK15.

for the visualization of evolving wiki articles is the History Flow method [17]. In this approach each contributor has assigned a unique color. Each revision of the evolving text is then represented as a sequence of blocks that represent the sections of the document. The blocks are colored according to the author who has edited the section and the size of the block corresponds to the amount of text. This does not only depict the insertion and removal of text sections by the users but additionally allow for the identification of edit wars between authors. In contrast to this page centric view, the iChase method [12] visualizes activities of a set of authors across multiple wiki articles as heatmaps. Southavilay et al. [16] extend the pure depiction of the amount and location of text edits done by a user by incorporating topic modeling. Therefore, they apply latent dirichlet allocation [4] in order to identify the contributions of users to the particular topics covered in a document. Based on the identified topics the evolution of topics as well as collaboration networks of users on particular topics can be analyzed.

2.3 Representing Mental Models as Graphs

Networks are a common representation for relations between entities of various kinds. Schvaneveldt et al. [15] argue that networks between entities based on proximities induced by people have a psychological interpretation. They assume that cognitive concepts such as memory organization and mental categories are reflected in the network structure. The pathfinder algorithm [15] derives a network of concepts from proximity data. Such proximities could be induced, for example, by associations made by a person. In general, it is also possible to derive such proximity data between concepts described in natural language texts [20].

One of the first approaches that utilize computational tools to extract mental models from text has been described by Carley [5]. After the identification of relevant words in a text, the words are linked based on syntactical analysis of the sentences of a text.

This approach has been further developed by Diesner et al. [6] and implemented in the software tool Automap where an analyst can specify a metamatrix of concepts and concept classes. This enables the identification of relations between entities of different types from text corpora, for example, people and organizations.

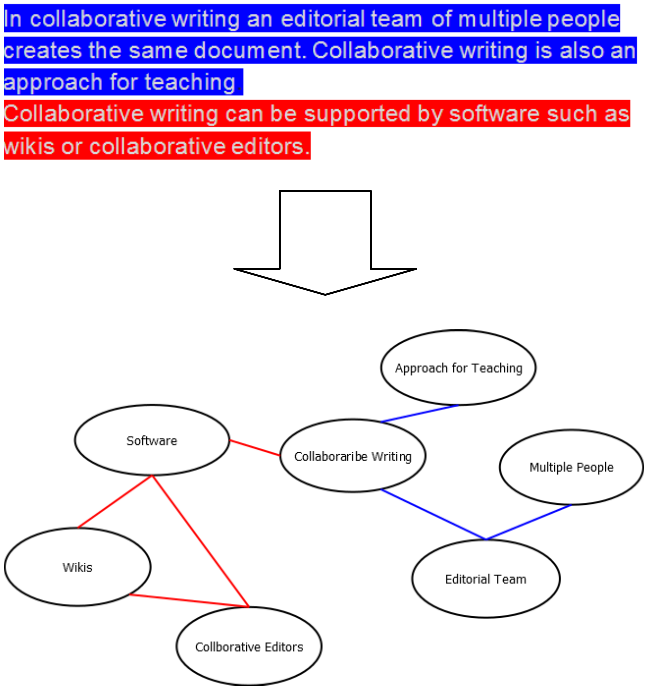
3. Visualization Approach

This paper extends network extraction from texts to dynamically evolving and collaboratively edited documents. When networks extracted from texts are considered as the author's mental model of the domain, as described in section 2.3, the aggregation of the networks extracted from several revisions of a collaboratively edited text can be interpreted as the joint representation of the individual mental models of all authors.

The basic assumption is that different authors introduce different concepts and relations to the text. In order to make these differences visible the author information is additionally incorporated into the network representation.

Each connection between concepts that can be extracted from the text can be labeled with the author who established it. In the small example in Figure 1 the little piece of text was produced by two different authors. Each author has assigned a unique color - in this case blue and red. The edges of the resulting network can then be colored according to the author who was the first who introduced the concept relation in the text.

This not only allows for a characterization of the underlying document in terms of concept relations but also a characterization of the contributors. Central concepts that are used by different authors but linked to different other concepts indicate different associations or views of the authors. Furthermore, the visualization approach additionally depicts which authors concentrate on thematic areas and which authors tend to relate concepts from different sub topics, for example, by writing a summary.



In collaborative writing an editorial team of multiple people creates the same document. Collaborative writing is also an approach for teaching

Collaborative writing can be supported by software such as wikis or collaborative editors.

Figure 1 A concept network extracted from a text edited by two different authors. The authors are represented by color.

By calculating network measures on the concept network a further quantitative characterization of the authors is possible as described in section 4.3.

4. Implementation

This section outlines details of the implementation in two perspectives. In particular, these are word network extraction using natural language processing, and network analysis.

4.1 Extracting Concept Networks from Texts

The extraction of networks from text requires several natural language processing components. In this work the DKPro toolkit [7] was used. It is based on the Apache UIMA[1] framework and provides a large variety of natural language processing algorithms that can be combined in a flexible way. The process of the extraction of word networks from a single document is depicted in Figure 2. First, a preprocessing step is often required for text gathered from the web in order to remove wiki or HTML markup. Further, in this step irrelevant content can be filtered from the document. For example, Wikipedia pages often contain a large reference section and a list of related web resources. These parts are important for the wiki article itself but are a source of noise when the actual content of the article should be analyzed. In the

---

[1] https://uima.apache.org/

second step, the phrases representing concepts in the text have to be identified, and after that, connected to a network by using a proximity measure in step 3. Since the result might contain phrases with slightly different spelling which actually refer to the same semantic concept the entity resolution step merges those candidate phrases to a single concept. Concepts and relations can then be encoded as a network that is used for further processing. In the following the steps 2 to 4 are described in more detail.

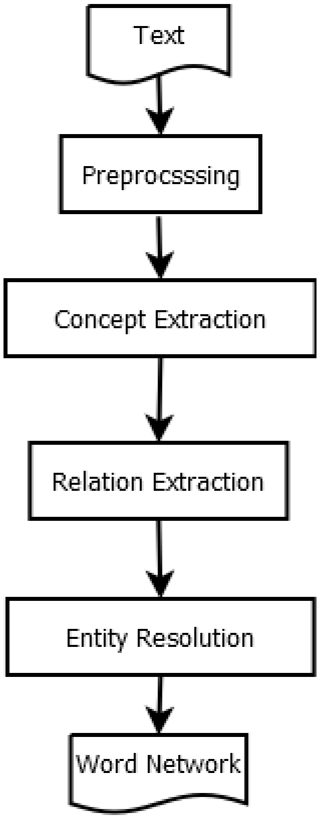


Figure 2 Process chain for the extraction of work networks from texts.

4.1.1 Concept Extraction

For the identification of the concepts in the input text noun phrase chunking was applied. First, the text is segmented into its sentences. Then part-of-speech (POS) tagging (using the Stanford PSO tagger[2]) is applied to label each word according to its function in its sentence. A naive solution for the extraction of concepts from the text would be to take each noun identified by the POS tagging as one concept. However, often one concept is described by more than one word. For example the phrase "Approach [NN] for [for] teaching [NN]" would result in two concepts, namely "Approach" and "Teaching", which does not really reflect the meaning of the phrase. Thus, noun phrase chunking is applied where the POS labeled words are chunked to meaningful noun phrases. This is done with the OpenNLP chunker[3], which identifies noun phrases according to certain rules. For example, the words "Approach [NN] for [for] teaching [NN]" are then identified as one single noun phrase.

4.1.2 Relation Extraction

After all concepts in the text are identified they have to be connected to a concept network according to a certain proximity measure. In this work, an edge between two concepts becomes established if the concepts co-occur in a sliding window of *n* words in at least one of the sentences in the text. This approach is straight forward but works well in practice [6, 10].

4.1.3 Entity Resolution

As already mentioned entity resolution is necessary in order to identify nodes in the network that represent the same concept and to merge them into single nodes. For example the noun phrases "Wiki" and "The Wikis" can be merged to the same concept "Wiki". In order to solve this problem, first all noun phrases have to be normalized using lemmatization. After that the concepts are compared pairwise by substring similarity [1]. If the similarity exceeds a value of 0.7 the concepts are merged and labeled with the shorter label of the two concepts.

4.2 Networks from Different Revisions

In order to extract an aggregated network from different revisions of a collaboratively edited text, the process chain described in section 4.1 is applied to each revision of the text in temporal order from the oldest to the latest revision. Each revision of the text was done by a single author. The edges in the network of the first revision are labeled with the author of this initial revision. Then in the first aggregation step all edges that are part of the network extracted from the second revision but do not exist in the network of the first revision are labeled with the author of the second revision and added to the previously extracted network. This proceeds until each revision has been processed. As described in section 3 the author information attached to the edges can then be visualized by using different colors for each author.

Since the aggregated network contains every noun phrase that has been used by the authors as a concept node, the network can be very large and likely contains concepts that are not relevant for the domain. Those concepts are often not well connected. Thus, in a preprocessing step the *k*-core [2] of the network is computed such that the resulting network contains only concepts with at least *k* connections to other concepts of the core. The resulting network has a reduced number of nodes, and the visualization concentrates on the most important concepts according to the connectedness to other core concepts in the network.

4.3 Quantitative Characterization of Contributors

For quantitative analysis the nodes (concepts) and edges can be ranked according to network centrality measures [19]. In this work concepts are ranked according to eigenvector centrality and betweenness centrality. The eigenvector centrality is a recursive measure and assigns a weight to each node according to the number its neighbors while the connections are weighted according to the centrality of the neighbors. This gives high weight to concepts that have many connections to other important concepts.

Edges are ranked according to the edge-betweenness centrality. The edge-betweenness centrality assigns high weights to edges that often occur on shortest paths between any pair of nodes.

---

[2] http://nlp.stanford.edu/software/tagger.shtml

[3] https://opennlp.apache.org/documentation/1.5.2-incubating/manual/opennlp.html

In order to use the network measures for a characterization of the authors of the document an aggregation is necessary. For the node centric centralities, namely node-betweenness and eigenvector centrality the centrality contribution of an author *A* can be calculated by equation 1:

$$nc\_contrib(A) = \sum_{(c_i,c_j)\in E:lab(e)=A} \frac{cent(c_i, c_j)}{2|(c_i, c_j) \in E: lab(e) = A|} \quad (1)$$

This result is the average centrality of nodes that are incident to edges labeled with author *A*.

The edge-betweenness contribution of author *A* is the average of all edges labeled with author *A* (equation 2):

$$eb_{contrib(A)} = \sum_{e\in E:lab(e)=A} \frac{eb\_cent(e)}{|e \in E: lab(e) = A|} \quad (2)$$

An author with a high contribution in terms of edge-betweenness centralilty could be interpreted as someone who relates different parts of the text and introduces relations between concepts of different sections. This could, for example, be someone who creates a comprehensive summary of a longer wiki article. Authors with high contribution to the eigenvector centrality of the concepts can be those who work on important sections of the text and establish many relations between important domain concepts.

5. Case Study

As a case study the described method was applied to a wiki article on media economy created during a master level university course in a study program on Applied Cognitive Science and Media Science. The relations between the concepts are based on a sliding window with the size of 4 words. Figure 3 depicts the 5-core of the resulting aggregated concept network. The size of the nodes corresponds to the number of connections in order to support the visual discovery of important concepts. It can be directly seen from the visualization that the concept “media combination” is most central. Four of the six authors relate this concept to other concepts as it can be seen by counting the different colors of the incident edges. The highest coverage of the edges has the author who has pink as assigned color. Other contributors relate concepts more according to certain sub topics like communication (see blue edges).

The results for the quantitative characterization of the contributors are presented in Table 1. It is important to mention that reducing the network to its 5-core has mainly presentation purposes. Thus, for more reliable results the calculations were performed on the 2-core of the network in which more concept are present.

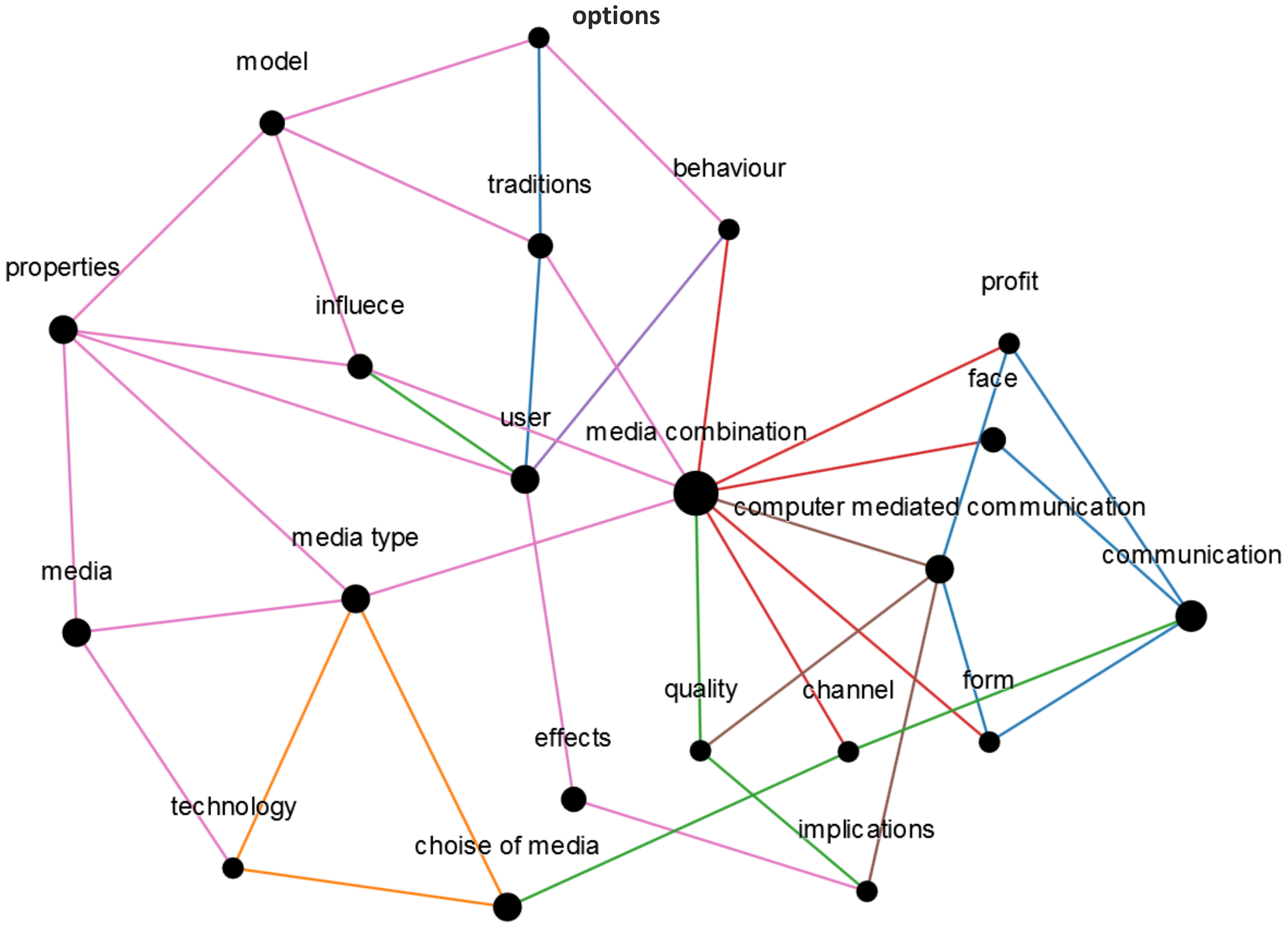


Figure 3 5-core of the aggregated concept networks extracted from a wiki article on media economics.

Table 1 Centrality contributions of the authors. EVC: Eigenvector centrality, NBC: Node-betweenness centrality (normalized), EBC: Edge betweenness centrality.

Author	Color	EVC	NBC	EBC
Student 1	Pink	0.20	0.07	**161.02**
Student 2	Red	**0.71**	**0.16**	95.85
Student 3	Green	0.35	0.07	80.17
Student 4	Blue	0.16	0.05	73.19
Student 5	Orange	0.1	0.04	111.45
Student 6	Brown	0.35	0.08	81.47

Student 1 has by far the highest contribution to the edge betweenness centrality. This is reasonable because this student did a reworking of large parts of the article and was highly involved in the shaping of the particular sections of the text. Student 2 has the highest scores regarding the node based centrality measures. However, the average edge-betweenness centrality is only moderate. This indicates that this student concentrated on the core topic of the article. This can also be seen in Figure 3 where the red edges of student 2 are all incident to the central concept.

6. CONCLUSION AND FURTHER WORK

The research presented in this paper describes an approach for the extraction of concept networks from text that incorporates author information in the visualization. In contrast to other existing visualizations of evolving texts our approach focuses rather on the relations between concepts than on the amount of text that is produced by individual authors. The case study has shown that the method is promising and can contribute to the analysis of collaborative text writing. In educational scenarios the proposed method enables tutors to investigate how students relate important domain concepts, and therefore, gain insights into their (possibly different) mental conceptualization. Thus, different views and focuses of students become visible. In future work the visualization will be integrated in an interactive application that supports the visual exploration of the resulting network through improved node and edge highlighting as well as facilities for data gathering and network reduction using *k*-core analysis. Regarding the interpretation and the analysis of the extracted networks the concept extraction can be adapted in such a way that the concepts and relations can be weighted by an expert according to their importance for the domain. This would result in more compact networks. In further evaluation the student characterizations derived from the colored word network can be related to self-assessment and characterizations made by a tutor.

7. REFERENCES

[1] Bär, D., Zesch, T. and Gurevych, I. DKPro Similarity: An Open Source Framework for Text Similarity. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations.* (Sofia, Bulgaria). Association for Computational Linguistics, 2013, 121-126.

[2] Bader, G. D. and Hogue, C. W. An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics, 4, 1 (2003), 2.

[3] Belanger, Y. and Thornton, J. Bioelectricity: A Quantitative Approach Duke University's First MOOC. (2013) Technical Report, Duke University.

[4] Blei, D. M., Ng, A. Y. and Jordan, M. I. Latent dirichlet allocation. J.Mach.Learn.Res., 3(mar 2003), 993-1022.

[5] Carley, K. and Palmquist, M. Extracting, representing, and analyzing mental models. Social forces, 70, 3 (1992), 601-636.

[6] Diesner, J. and Carley, K. M. Revealing social structure from texts: meta-matrix text analysis as a novel method for network text analysis. *Causal mapping for information systems and technology research: Approaches, advances, and illustrations*, 2005, pp. 81-108.

[7] Eckart de Castilho, R. and Gurevych, I.In *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT*. Association for Computational Linguistics, 2014, 1-11.

[8] Flower, L. and Hayes, J. R. A Cognitive Process Theory of Writing. College Composition and Communication, 32, 4 (1981), pp. 365-387.

[9] Harrer, A., Moskaliuk, J., Kimmerle, J. and Cress, U. Visualizing wiki-supported knowledge building: co-evolution of individual and collective knowledge. In Anonymous *International. Symposium on Wikis.* 2008 19:1-19:9.

[10] Mihalcea, R. and Tarau, P. TextRank: Bringing order into texts. In Proceedings of the EMNLP. Association for Computational Linguistics, (Barcelona, Spain), 2004, 404-411.

[11] Paranyushkin, D. Identifying the pathways for meaning circulation using text network analysis. Technical Report Nodus Labs, Berlin, (2011).

[12] Riche, N. H., Lee, B. and Chevalier, F. iChase: Supporting Exploration and Awareness of Editing Activities on Wikipedia. In *Proceedings of the International Conference on Advanced Visual Interfaces.* (Roma, Italy). ACM, New York, NY, USA, 2010, 59-66.

[13] Sabrina Ziebarth and Hoppe, H. U. Moodle4SPOC: A Resource-Intensive Blended Learning Course. In *Proceedings of the European Conference on Technology Enhanced Learning*. (Graz, Austria), 2014, 359-372.

[14] Scardamalia, M. and Bereiter, C. Computer Support for Knowledge-Building Communities. The Journal of the Learning Sciences, 3, 3 (1993), pp. 265-283.

[15] Schvaneveldt, R. W., Durso, F. T. and Dearholt, D. W. Network structures in proximity data. Psychol. Learn. Motiv., 24 (1989), 249-284.

[16] Southavilay, V., Yacef, K., Reimann, P. and Calvo, R. A. Analysis of collaborative writing processes using revision maps and probabilistic topic models. In *Proccedings of the Learning Analytics and Knowledge Conference*. (Leuven, Belgium), 2013, 38-47.

[17] Viegas, F. B., Wattenberg, M. and Dave, K. Studying cooperation and conflict between authors with history flow visualizations. In *Proceedings of the SIGCHI conference on Human factors in computing systems.* ACM, 2004, 575-582.

[18] Villalon, J. J. and Calvo, R. A. Concept Map Mining: A definition and a framework for its evaluation. In *Proceedings of the International Vonference on Web Intelligence and Intelligent Agent Technology, 2008,* IEEE, 2008, 357-360.

[19] Wasserman, S. and Faust, K. *Social Network Analysis: Methods and Applications.* Cambridge University Press, 1994.

[20] Wild, F., Haley, D. and Bulow, K. Monitoring conceptual development with text mining technologies: CONSPECT. In *Proceedings of eChallenge*, 2010, 1-8.

INSIGHT: a Semantic Visual Analytics for Programming Discussion Forums

Piyush Awasthi
School of Computing, Informatics & Decision Systems Engineering,
Arizona State University,
699 S. Mill Ave., Tempe AZ, USA
Piyush.Awasthi@asu.edu

I-Han Hsaio
School of Computing, Informatics & Decision Systems Engineering,
Arizona State University,
699 S. Mill Ave., Tempe AZ, USA
Sharon.Hsiao@asu.edu

ABSTRACT

This paper presents INSIGHT, a visual analytics web application, designed to induce & inspire programming language learning from discussion forums. The visual analytics, extracts and displays semantic content from 'Stack Exchange' in a form of bubble chart. The bubbles represent summarized semantic concepts from the forum posts and outlines the concept specificity of each individual post. The discussion forum content are modeled as concepts based on an innovative Topic Facet Modeling algorithm (a probabilistic topic model that assumes all words in single sentence are generated from one topic facet), and aimed to provide better understanding and solicitation of the increasing large volume of discussion content. We hypothesize that by navigating and interacting (browsing, sorting, searching etc.) with the Facets, will enhance learning. A comprehensive system design rationales and preliminary qualitative study are reported in this paper.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; K.3.2 [**Computer and Information Science Education**]: Computer Science Education

General Terms

Measurement, Design, Experimentation, Programming, web application.

Keywords

Learning Analytics, discourse analysis, visual analytics, programming, discussion forums, computer-supported collaborative learning.

1. INTRODUCTION

Learning programming involves a variety of complex cognitive activities, from conceptual knowledge construction to basic structural operations, program design, programming understanding, modifying, debugging, and documenting (Lye & Koh, 2014; Piech et al., 2012; Robins, Rountree, & Rountree, 2003). There have been major educational technology advances over the last two decades, centered on understanding the nature of programming skills explicitly using declarative aspects of programmer's knowledge (i.e. program comprehension and generation, required concepts & skills to program). For example: intelligent tutors, auto program feedback generation, collaborative programming support, personalized learning resources, etc. (Aleven, McLaren, Roll, & Koedinger, 2006; Anderson & Skwarecki, 1986; Atkinson & Renkl, 2007; Barnes & Stamper, 2008; Boyer et al., 2011; Hsiao, Sosnovsky, & Brusilovsky, 2010; Lye & Koh, 2014; Piech, Sahami, Koller, Cooper, & Blikstein, 2012; VanDeGrift, 2004) The technology support has evolved from classrooms to online, declarative to exploratory, and individual to social. In teaching and learning programming, students are typically asked to refer to API (Application Programming Interface) or programming textbooks for relevant information (i.e. code examples). The internalization process from forming a question to reaching out to APIs or textbooks is usually not captured in learning programming. From a constructivism point of view, the action of articulating a problem and initiating search or referencing can be a valuable learning activity. There are numerous tools that have been built to make completing programming tasks easier, such as Mica (Stylos & Myers, 2006) (there are more cases reviewed in the literature review section), but less is focused on amplifying learning opportunities.

In the easily accessible Internet era, search engines, index and make the excessive amount of programming problems and solutions available. Because programming problems are usually more complex than a simple sequence of query keywords, dedicated communities such as discussion forums and Q&A sites are the most popular alternatives for problems & solutions. The drastic shift in momentum of learning opportunities from APIs and textbooks to community help is not yet fully comprehended though. Besides, forums or discussion boards usually lack dynamic and extensive content analysis due to large and increasing content volume and high computational cost in discourse analyses. In this work, we aim to research a new technology to facilitate online learning from programming discussion forum. We apply Learning Analytics approach, which has demonstrated promising results in online learning (Siemens & Baker, 2012). However, the majority of learning analytics focuses on visual representations or the system's usefulness, the core should be focused on the visualization impact to improve learning or teaching

Published in CEUR-WS:
Proceedings of the VISLA15 Workshop on Visual Aspects of Learning Analytics (Duval et al., eds), organized at LAK15.

(Verbert, Duval, Klerkx, Govaerts, & Santos, 2013). In this work, we present a new visual analytics system that targets at providing better understanding and solicitation of the increasing large volume of discussion content.

In the rest of the paper, we summarize the related work in learning analytics and other intelligent visual support for programming language learning. We then describe briefly the methodology to extract forum content semantics. In section 4, we present the system design and rationales. A user study and preliminary results are presented in section 5 & 6. Finally, we summarize the work and discuss future work and limitation.

2. LITERATURE REVIEW

2.1 Learning Analytics

Signals project at Purdue University is one of the pioneering examples of the successful application of academic analytics that integrate predictive modeling and report significantly higher grades and retention rates than were observed in control groups (Arnold, 2010). Septris and SICKO project at Stanford School of Medicine utilizes educational simulation games to offer deeper insight into learner's competency and decision making to help prepare doctors well. The game analytics not only help instructors see what choice learners made but also what data was used to make those choices and when they decided to make those choices. (Jamie Tsui, James Lau, Lisa Shieh, 2014). The application has been well received by the learners and instructors with over 32000 usage, 16000 plays and 2500 completions.

Over the decades, discourse analysis on discussion forums has been carried out through various formats, network analyses, topical analyses, interactive explorers, knowledge extraction, etc. (Dave, Wattenberg, & Muller, 2004; Gretarsson et al., 2012; Indratmo, Vassileva, & Gutwin, 2008; Lee, Kim, Cho, & Woo, 2013; Wei et al., 2010). With the rapid growth of free, open, and large user-based online discussion forums, it is essential, therefore, for education researchers to pay more attention to emerging technologies that facilitate learning in cyberspace. For instance, (Sande, 2010) investigated online tutoring forums for homework help by making observations on the participation patterns and the pedagogical quality of the content. (Hanrahan, Convertino, & Nelson,2012; Posnett, Warburg, Devanbu, & Filkov, 2012) studied expertise modeling in such environment. Cohere (Shum,2008) investigates semantic connections by identifying the link types to associate negative, positive, neutral interactions among online discourses. (Wise, Zhao, & Hausknecht,2013) observed the listening behavior, which encapsulates different actions that learners take in relation to others posts (attending, reading etc.), to further describe the discussion engagement.

2.2 Intelligent Visual Support for Programming

In the VL/HCC (IEEE Symposium on Visual Languages and Human-Centric Computing) community, we can see a large amount of research addressing the issue that developers tend to interleave between activities like searching for relevant codes and collecting codes and other information that they believe would be necessary for editing or duplication (Ko, Myers, Coblenz, & Aung, 2006). These tools include navigational shortcuts to the code in IDE (Singer, Elves, & Storey, 2005), leveraging version history data to predict code changes (Zimmermann, Zeller, Weissgerber, & Diehl, 2005), better use of API (Stylos & Myers, 2006), and integration of web search or recommending source code examples in development environment (Brandt, 2010; Holmes & Murphy, 2005; Hsiao, Li, & Lin, 2008; Stylos & Myers, 2006). These systems were designed mainly to extract relevant information from the web to aid in current coding tasks and save time that would otherwise be spent navigating through codes to gather information. Moreover, with the rise of web 2.0, we also see that a variety of technologies (blogs, tags, wikis, recommenders etc.) are emerging to exploit social information foraging (Chi, Pirolli, & Lam, 2007), such as online collaborative programming (social coding in GitHub[1]), Q&A websites, crowdsourcing suggestions, etc. (Bacchelli, Ponzanelli, & Lanza, 2012; Dabbish, Stuart, Tsay, & Herbsleb, 2012; Goldman, Little, & Miller, 2011; Hsiao et al., 2008; Mujumdar et al., 2011; Nasehi, Sillito, Maurer, & Burns, 2012; Treude, Barzilay, & Storey, 2011; Vasilescu, Serebrenik, Devanbu, & Filkov, 2014). However, almost all of these tools are targeted at problem-solving augmentation, reducing coding cognitive overhead when coding, and utility features enhancement (i.e. collaboration). Tools to support learning activities are less evident.

3. TOPIC FACET MODEL

Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003) is an unsupervised algorithm that uses bag of words approach to perform statistical topic modeling, which is a well-established method for uncovering hidden structures in large text corpora. There are several variations of LDA-based topic models to successfully encapsulate large text semantics into topic words, such as online reviews, political opinions, microblog streams, email summaries etc. (Jo & Oh, 2011; Lan, Buntine, & Huidong, 2010; Liu et al., 2012; Wang, Agichtein, & Benzi, 2012). In this work, we present a novel Topical Facets Modeling (TFM) method to capture online forum posts semantics.

The TFM algorithm automatically detect topics from conversational and relatively short amount of texts in each forum post. It is an extension of LDA (Blei et al., 2003) and SLDA (Lan et al., 2010). A topic is a multinomial distribution of words that represents a concept from each forum post. A facet is a multinomial distribution of words that represents a more specific topic in the forum, for instance, extends (a java keyword) is one of the main facets in determining whether a program implemented inheritance concept in Java programming language or not. Thus, Topic Facet Model firstly adopts SLDA (Lan et al., 2010) in the topic model. Essentially, SLDA takes into account the position of each individual word of topic inference. It then forces all words in a sentence are generated from one topic. When a post is topic-specific, short-and-sweet, such as how to write a for loop?, SLDA is supposed to distinctively generate the corresponding topic word - loops. However, as we discussed earlier, an open discussion forums often mix with various complexities of posts. For instance, "Can an array of objects be iterated in enhanced for loop". Given the sentence

---

[1] https://github.com It is an online software repository site, which allows distributed revision control and source code management.

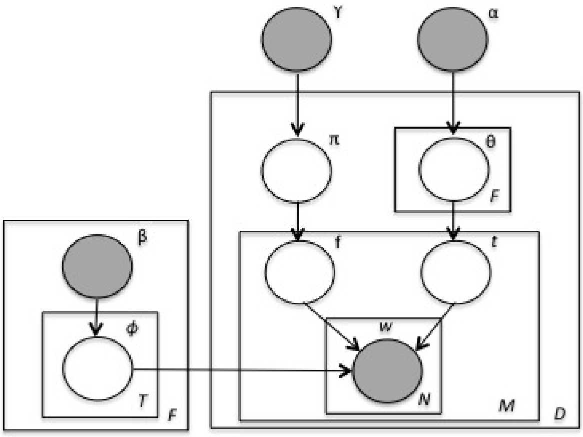


Figure 1: Topic Facet Model.

combines two main concepts, arrays and loops, SLDA will constrain only one topic word to be generated. In this case, the key of the question is about topic arrays (whether one can perform a function with array data structure), however, due to that there are more topic loops related words represented, the SLDA will misinterpret it. This is where the facets come into play, to take into account specificity of a topic in the model. Following the same example, we can specify 'array iteration' as a facet for topic loops (Hsiao,I-H, & Awasthi,P. 2015, to be appeared) . To explain Topic Facet Model algorithmically, Figure 1 shows the plate diagram. The words generative process is explained following.

1. For every pair of topic word t and facet f, draw a word distribution $\phi_{\text{ft}} \sim$ Dirichlet (β_{f})
2. For each document d,
   a. Draw the document's topic word distribution $\pi_{\text{d}} \sim$ Dirichlet (γ)
   b. For each topic word t, draw a facet distribution θ_{df} $\sim$ Dirichlet (α)
   c. For each sentence,
      - Choose a topic word j $\sim$ Multinomial (π_{d})
      - Given topic word j, choose a facet k $\sim$ Multinomial (θ_{dj})
      - Generate words w $\sim$ Multinomial (ϕ_{jk})

4. INSIGHT

In order to provide dynamic intelligent & personalized support for large-scale of online discussion forums, we build a web application and called it INSIGHT (since, it provides an insight on the concepts on which the answer has been built, to help user map his way to proper understanding of it), by using Django, Python and Javascript. The web application (Figure 2) re-structures a discussion forum site into 3 parts: Filters, Analytics Visualizations and Forum Posts. They are represented in the following three UI panels from left to right:

- **Control Panel (Left) -** contains a Search, three links - Inheritance, Loops, Stackoverflow.com. Inheritance and Loop links refreshes the section 3 with the respective posts data. It also refreshes the section 2 TFM bubbles. The search bar performs a normal search against the data on the keywords fed into it.

Table 1: Topic Facet Model notations

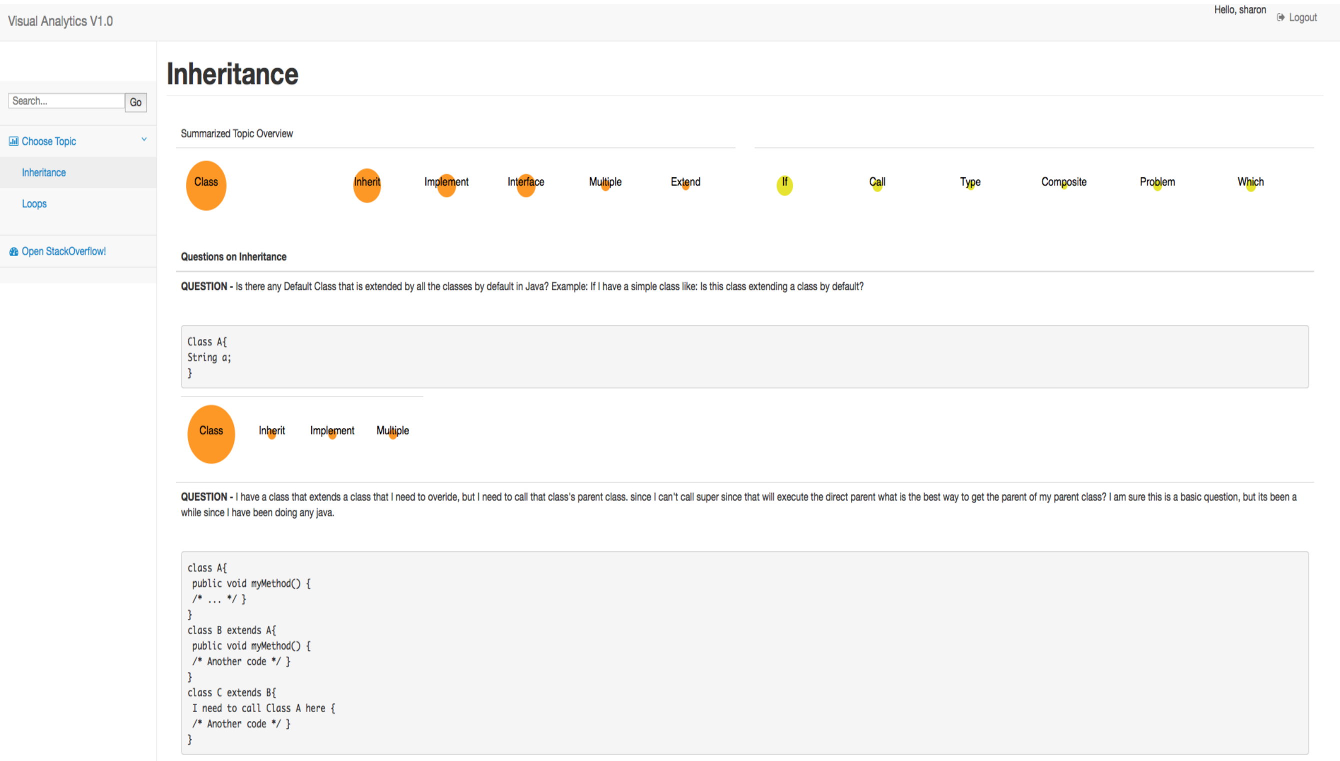
D: number of posts, M: number of sentences, N: number of words, T: number of topic-words, F: number of facets,; ω: word, t: topic-word, f: facet, ϕ: multinomial distribution over words, θ: multinomial distribution over topic-words, π: multinomial distribution over facets, : Dirichlet prior vector for θ, $\beta(_{\text{w}})$, $\beta_{\text{j(w)}}$: Dirichlet prior vector for ϕ(of facet j), $\gamma_{\text{(j)}}$: Dirichlet prior vector for π

- **Analytics panel (Top Right) -** This section provides the result of our TFM model on the data provided in section 3. The results are showcased in form of bubble chart with some words mentioned in the center of each bubble. These words in the bubble chart are the most highlighted topics discovered by our algorithm. The size variation of individual bubbles defines the topic word relevance to the data i.e. Bigger the circle, bigger is relation of data to that topic.

  The bubbles are sectioned into two different color codes - one showing the topic related to the data and the other showing topics which are not related. Our TFM model clearly detects these differences, we call them facets and non facets. The bubble chart also changes according to the link selected on the left i.e. inheritance and loop. For inheritance the bubbles show following TFM facets - class, inheritance, extend, multiple, implement and following non-facets - if, call, type, composite, problem, which. For loops the bubbles show following TFM facets - for, do, loop, instance and following non facets - time, compile, value, optimism, variable. The TFM facets individually are clickable and work like a tag selection. On click, the data in section 3 gets sorted in descending order on the TFM value of the bubble clicked.

- **Forum posts (Bottom Right) -** contains all the forum posts data (question, its accepted answer (if available) and the next top voted answer) on the topic chosen in control panel, i.e., inheritance or loop. Each post contains some text and code, if available. In addition to the texts, each row also contains the TFM bubble, again, the size of the bubble denoting the facet relevance to the content of the post. The purpose of associating each post with its TFM facets value is to help users browse faster to find the related question to their problem. Every row of question is expandable. Once the user finds a related question to his problem, he can click on it to reveal its answer.

INSIGHT has been developed on Django and Python (interpreted languages) therefore, it can be scaled for larger data sets without compromising on the processing time. The architecture of the application has also been optimized to handle larger data sets. Moreover, all the visualization on the application is handled by javascript, therefore, providing the facility to incorporate more chart visualization like d3.js without worrying on the cost of efficiency, as these are



QUESTION - Is there any Default Class that is extended by all the classes by default in Java? Example: If I have a simple class like: Is this class extending a class by default?

```
Class A{
String a;
}
```

QUESTION - I have a class that extends a class that I need to overide, but I need to call that class's parent class. since I can't call super since that will execute the direct parent what is the best way to get the parent of my parent class? I am sure this is a basic question, but its been a while since I have been doing any java.

```
class A{
 public void myMethod() {
 /* ... */ }
}
class B extends A{
 public void myMethod() {
 /* Another code */ }
}
class C extends B{
 I need to call Class A here {
 /* Another code */ }
}
```

Figure 2: Interface of INSIGHT.

well optimized javascripts designed to handle data sets of any size.

4.1 The Analytics

4.1.1 *Implementation*

We implemented user tracking using Javascript on INSIGHTS. Javascript offers a quick and easy way to collect aggregate data on users and is built into INSIGHT. The system as a whole is a comprehensive logging system that tracks user's actions to a specific session. We also wish to provide a debrief of the session to user for improving his learning.

There are multiple third party tracking tools for example google analytics. But they all lack the ability to track an individual user's actions/decisions in chronological order. For example, you could see user clicked on question 1 and 5 to formulate his answer but with GA you cannot determine whether question 1 was clicked first or the question 5. Also, GA doesn't provide all the analytics together and it requires to be combined with other analytical tools to provide the full comprehensive logging system.

Because the order of actions is especially crucial in analytics, we built a new feature for INSIGHT to track all of user's actions in a log, which includes several pieces of information:

- What action they performed
    1. Mouse click on the page
    2. Scrolling up or down
    3. Which buttons were clicked
    4. What text were highlighted
    5. Which keywords were used in search bar
    6. Which questions were expanded for answers
    7. Which TFM bubble did the user click on to sort the data.
- When they performed the action
- On which page they performed the action

Furthermore, we added a tracking feature on User study page as well for all the decisions that the user makes during answering the questions. This feature tracks when a question was answered and which question was answered first.

All of the information from in-application actions and from the user study page are recorded continuously throughout the session. The data is stored for further aggregate analysis and research.

4.1.2 *Benefits and applications*

The tracking/logging feature built in the application allows us to drill down and filter by any of the levels, so we can easily see which actions were performed on which page and at what time. This also allow us to identify the common mistakes and patterns users follow during finding an answer to his coding problem. These mistakes or patterns are then to be addressed with further analysis and research and then built into the application to improve user learning.

The data can also be filtered over date, so further analysis can also be done to study the change of user's understanding over time, which may also be correlated with improvements in learner's knowledge.

5. USER STUDY

5.1 The Design

Table 2: Study Design

Topic 1 - Loops	Topic 2 - Inheritance
Experiment	Control 1
Experiment	Control 0
Control 1	Experiment
Control 0	Experiment

The user study has been designed to test the functional application of INSIGHT and its efficiency against other public online forums. For this particular case study, we use 'Stackoverflow.com' to do comparative study. Table 2. displays our study design with four sets of control environment to test the application thoroughly.

Table defines three control groups - experiment, control 1 and control 0.

- **Experiment** - user will answer the question of the respective section using INSIGHT.
- **Control 1** - user will answer the question of the respective section using help from Stackoverflow.com.
- **Control 0** - user is not allowed to refer Stackoverflow.com for solving the problem.

Though the experiment group and control 0 group use the same approach to answer the problem in hand i.e. referring to visual analytics for help, they have been defined as different set as the experiment group will access programming help through visual analytics interface only and the control group may get access to programming help through Stackoverflow.com depending upon which control group it refers to.

5.2 Study Procedure

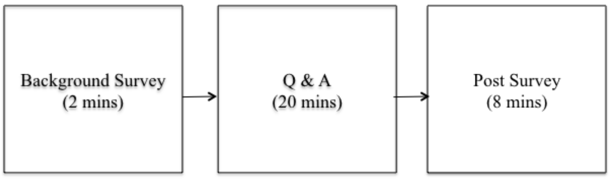


Figure 3: User study process flow.

Figure 3 shows the flow of the study. Every user is asked to go through the following three stages of the study -

- Background Survey
- Q&A
- Post Survey

Background survey is all about knowing the user's knowledge level in the area of coding and also involves asking him how well he is versed with online help i.e. does he uses google and stackoverflow and if he does then how well he is involved in the process.

The Q& A involves asking user two questions - one on topic loops and the other on topic inheritance. The user is required to answer both of the questions by referring to INSIGHT or stackoverflow.com (depending upon which control group the user belongs to). This task is time bound with 10 minutes allotted to each individual question adding it up to 20 minutes in total.

Post completion of the questions, the user is asked to provide a post survey to get their feedback to help improve the application. Post survey is the normal feedback system with users being asked to rate our application on the scale of 1 to 5 (where, 1-very bad and 5-excellent) on satisfaction, ease of use, ease of learning and usefulness. Users are also provided a space to give any comments on how can the application be improved. All of this data with user actions are stored in the database and later will be used to run more experiments using hidden topic markov model to find out how constructive the user responses are(Jeong, Gupta, Roscoe, Wagster, Biswas, Schwartz)

6. EVALUATION

6.1 Data Collection

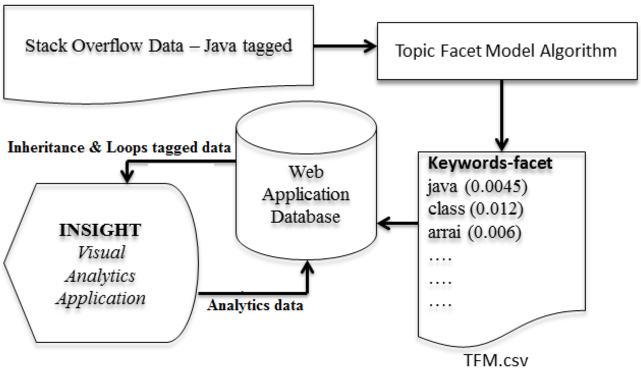

Stack Overflow Data – Java tagged
Topic Facet Model Algorithm
Inheritance & Loops tagged data
Web Application Database
Keywords-facet
java (0.0045)
class (0.012)
arrai (0.006)
….
….
….
INSIGHT
Visual Analytics Application
Analytics data
TFM.csv

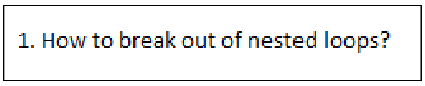
Figure 4: Data Collection.

We sampled one year (year 2013) of forum posts in topic Java from stackoverflow site through StackExchange API. The data pool was selected from the top 10 frequent tagged questions due to most of the posts in this section contained at least one accepted answer. For our case, we only show top 2 frequent tagged out of those 10 i.e. 'Inheritance' and 'Loops'. It will allow us to build a baseline to test INSIGHT on smaller set of data and also it's effectiveness. Later, the application will be scaled up to include all frequent tagged topics and questions.

6.2 User-Study Evaluation

INSIGHT version 1 prototype was recently developed and there are many use cases and further user studies underway. Till now, we have conducted 4 user study testing all four combinations of control environment as shown in Table 2. As the number of users were limited, we provide a qualitative evaluation of our application.

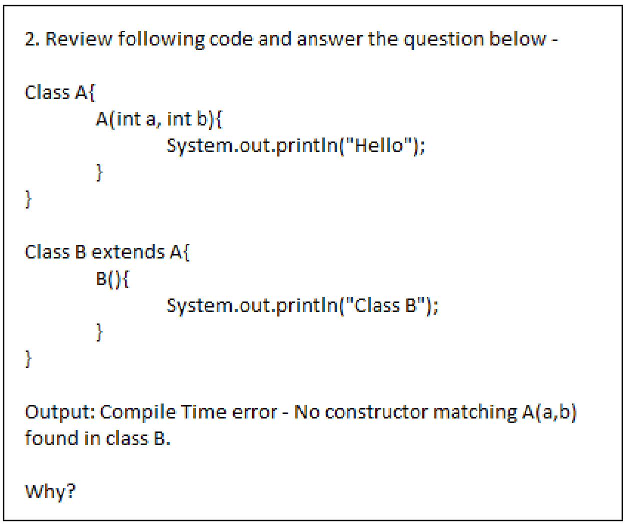
On the base of the background information provided by the users, the users can be clearly divided into two sets - 1. Users with some programming experience and 2. Users with no programming experience. Each set contained 2 users each and these sets were formed completely on the basis of how well they knew coding and how well they are familiar with the online coding forums.



1. How to break out of nested loops?

Figure 5: Loops Question.

The users were presented with same set of two questions - a easy problem on topic loops and a slightly difficult problem on topic inheritance. Figure 5 and 6 shows the snippet of the questions.



2. Review following code and answer the question below -

```
Class A{
        A(int a, int b){
                System.out.println("Hello");
        }
}

Class B extends A{
        B(){
                System.out.println("Class B");
        }
}
```

Output: Compile Time error - No constructor matching A(a,b) found in class B.

Why?

Figure 6: Inheritance Question.

Based on the control group users were required to access the respective resources and answer the questions. Out of the four, two users were able to answer both the question, whereas other two were only able to answer question on topic loops. Moreover, the two users who were able to answer both the question were the users who had some background knowledge of coding and were involved in some online discussion forums i.e. set-1. Because both the set involved one individual case, where the user was allowed to access 'Stackoverflow.com' i.e. Control 1, it comes as no surprise that users with some background knowledge were easily able to browse through the resources (Stackoverflow.com & INSIGHT) and find the solution and the other users weren't.

User's who failed to answer the problem on inheritance belonged to set-2 i.e users with no background knowledge. These users were not able to find the solution either on Stackoverflow.com or on INSIGHT, which provides an intuition whether users require some background knowledge to find the solution or not. It also points to INSIGHT being not so helpful for the users to find solution for the inheritance question. This intuition and deduction has been followed on very small dataset therefore it provides no concrete evidence for whatsoever. To testify for the intuition we require more rigorous testing and user studies.

During the study, users used search and TFM tagging facility extensively to find answer to the questions. Users found TFM bubble chart helpful as it helped them browse through the questions faster. The variation in the sizes of it assisted them to relate to the relevance of the question more easily. Search bar of the application worked in supplement with the TFM bubble chart to help users find related questions, hence the solution. The most frequent words searched for the question loops were - 'break', 'loops' and for question inheritance - 'extends', 'compilation error'.

Though the users liked the ease of use of the application, they felt the need of visually improving the application on the same lines. Collectively, INSIGHT was positively received by the users.

7. FUTURE DIRECTIONS & DISCUSSIONS

With so many variations and wideness in teaching style and technology, finding out ways to make learning effective and interesting becomes quite a task. Here are some of the ways we can lead INSIGHT in directions to make it more personalized:

INSIGHT logs and stores all the user's action on it with individual timestamp of when they were performed. We can filter this data by date, so further analysis can be done over the change of providing an answer by respective users, which may also be correlated with improvement in user's knowledge. Also, providing a dashboard for individual users to track or debrief on there performance by reviewing their answer and action logs can help users to gain deeper insight into their conceptual learning level and also help them review what data they used to formulate the answer. There logs can also be then used for identifying the area of weakness and then can be used by to provide more personalized help.

An ability to provide and instant feedback based on the user's action is very conducive to the improvement in user's knowledge or learning. This will also help users to form a empathetic connection as providing instant feedback stimulates a gesture of more personalization.

8. LIMITATIONS

In this paper we describe a functional prototype of visual analytics tool - INSIGHTS for discourse centric content. Our preliminary results demonstrated that INSIGHTS could be a promising approach to help users really learn and understand the concepts instead just writing the answers to problems but there are several limitations in current implementation. 1) The current version has been only tested for two topics - Loops and Inheritance out of 10 topics that were explicitly chosen to represent Easy & Difficult topics for CS1 course. 2) The user study was only conducted with limited subjects and requires more rigorous testing of our application and also to aggregate quantifiable results to test our hypotheses. 3) We currently only experimented the bubble charts visual representation on the extracted content semantics. We completely ignored the semantic relations, such as the concept causal relations, sequential or network visualizations.

9. REFERENCES

[1] V. Aleven, B. Mclaren, I. Roll, and K. Koedinger. Toward meta-cognitive tutoring: A model of help seeking with a cognitive tutor. *International Journal of Artificial Intelligence in Education*, 16(2):101–128, 2006.

[2] J. R. Anderson and E. Skwarecki. The automated tutoring of introductory computer programming. *Communications of the ACM*, 29(9):842–849, 1986.

[3] K. E. Arnold. Signals: Applying academic analytics. *Educause Quarterly*, 33(1):n1, 2010.

[4] R. K. Atkinson and A. Renkl. Interactive example-based learning environments: Using interactive elements to encourage effective processing of worked examples. *Educational Psychology Review*, 19(3):375–386, 2007.

[5] A. Bacchelli, L. Ponzanelli, and M. Lanza. Harnessing stack overflow for the ide. In *Recommendation Systems for Software Engineering (RSSE), 2012 Third International Workshop on*, pages 26–30. IEEE, 2012.

[6] T. Barnes and J. Stamper. Toward automatic hint generation for logic proof tutoring using historical student data. In *Intelligent Tutoring Systems*, pages 373–382. Springer, 2008.

[7] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[8] K. E. Boyer, R. Phillips, A. Ingram, E. Y. Ha, M. Wallis, M. Vouk, and J. Lester. Investigating the relationship between dialogue structure and tutoring effectiveness: a hidden markov modeling approach. *International Journal of Artificial Intelligence in Education*, 21(1):65–81, 2011.

[9] J. Brandt, M. Dontcheva, M. Weskamp, and S. R. Klemmer. Example-centric programming: integrating web search into the development environment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 513–522. ACM, 2010.

[10] E. H. Chi, P. Pirolli, and S. K. Lam. Aspects of augmented social cognition: Social information foraging and social search. In *Online Communities and Social Computing*, pages 60–69. Springer, 2007.

[11] L. Dabbish, C. Stuart, J. Tsay, and J. Herbsleb. Social coding in github: transparency and collaboration in an open software repository. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pages 1277–1286. ACM, 2012.

[12] K. Dave, M. Wattenberg, and M. Muller. Flash forums and forumreader: navigating a new kind of large-scale online discussion. In *Proceedings of the 2004 ACM conference on Computer supported cooperative work*, pages 232–241. ACM, 2004.

[13] L. Du, W. L. Buntine, and H. Jin. Sequential latent dirichlet allocation: Discover underlying topic structures within a document. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 148–157. IEEE, 2010.

[14] M. Goldman, G. Little, and R. C. Miller. Collabode: collaborative coding in the browser. In *Proceedings of the 4th international workshop on Cooperative and human aspects of software engineering*, pages 65–68. ACM, 2011.

[15] B. Gretarsson, J. O'donovan, S. Bostandjiev, T. Höllerer, A. Asuncion, D. Newman, and P. Smyth. Topicnets: Visual analysis of large text corpora with topic modeling. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(2):23, 2012.

[16] B. V. Hanrahan, G. Convertino, and L. Nelson. Modeling problem difficulty and expertise in stackoverflow. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work Companion*, pages 91–94. ACM, 2012.

[17] R. Holmes and G. C. Murphy. Using structural context to recommend source code examples. In *Proceedings of the 27th international conference on Software engineering*, pages 117–125. ACM, 2005.

[18] I.-H. Hsiao, Q. Li, and Y.-L. Lin. Educational social linking in example authoring. In *Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*, pages 229–230. ACM, 2008.

[19] I.-H. Hsiao, S. Sosnovsky, and P. Brusilovsky. Guiding students to the right questions: adaptive navigation support in an e-learning system for java programming. *Journal of Computer Assisted Learning*, 26(4):270–283, 2010.

[20] Y. Jo and A. Oh. Aspect and sentiment unification model. 2010.

[21] A. J. Ko, B. A. Myers, M. J. Coblenz, and H. H. Aung. An exploratory study of how developers seek, relate, and collect relevant information during software maintenance tasks. *Software Engineering, IEEE Transactions on*, 32(12):971–987, 2006.

[22] Y.-J. Lee, E.-K. Kim, H.-G. Cho, and G. Woo. Detecting and visualizing online dispute dynamics in replying comments. *Software: Practice and Experience*, 43(12):1395–1413, 2013.

[23] S. Liu, M. X. Zhou, S. Pan, Y. Song, W. Qian, W. Cai, and X. Lian. Tiara: Interactive, topic-based visual text summarization and analysis. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(2):25, 2012.

[24] S. Y. Lye and J. H. L. Koh. Review on teaching and learning of computational thinking through programming: What is next for k-12? *Computers in Human Behavior*, 41:51–61, 2014.

[25] D. Mujumdar, M. Kallenbach, B. Liu, and B. Hartmann. Crowdsourcing suggestions to programming problems for dynamic web development languages. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*, pages 1525–1530. ACM, 2011.

[26] S. M. Nasehi, J. Sillito, F. Maurer, and C. Burns. What makes a good code example?: A study of programming q&a in stackoverflow. In *Software Maintenance (ICSM), 2012 28th IEEE International Conference on*, pages 25–34. IEEE, 2012.

[27] C. Piech, M. Sahami, D. Koller, S. Cooper, and P. Blikstein. Modeling how students learn to program. In *Proceedings of the 43rd ACM technical symposium on Computer Science Education*, pages 153–160.

ACM, 2012.

[28] A. Robins, J. Rountree, and N. Rountree. Learning and teaching programming: A review and discussion. *Computer Science Education*, 13(2):137–172, 2003.

[29] S. B. Shum et al. Cohere: Towards web 2.0 argumentation. *COMMA*, 8:97–108, 2008.

[30] G. Siemens and R. S. d Baker. Learning analytics and educational data mining: towards communication and collaboration. In *Proceedings of the 2nd international conference on learning analytics and knowledge*, pages 252–254. ACM, 2012.

[31] M.-A. Storey. Theories, methods and tools in program comprehension: Past, present and future. In *Program Comprehension, 2005. IWPC 2005. Proceedings. 13th International Workshop on*, pages 181–191. IEEE, 2005.

[32] J. Stylos and B. A. Myers. Mica: A web-search tool for finding api components and examples. In *Visual Languages and Human-Centric Computing, 2006. VL/HCC 2006. IEEE Symposium on*, pages 195–202. IEEE, 2006.

[33] C. Treude, O. Barzilay, and M.-A. Storey. How do programmers ask and answer questions on the web?: Nier track. In *Software Engineering (ICSE), 2011 33rd International Conference on*, pages 804–807. IEEE, 2011.

[34] J. Tsui, S. EdTech, and J. Benfield. Brief. 2014.

[35] C. van de Sande and G. Leinhard. Online tutoring in the calculus: Beyond the limit of the limit. *education*, 1(2):117–160, 2007.

[36] T. VanDeGrift. Coupling pair programming and writing: learning about students' perceptions and processes. *ACM SIGCSE Bulletin*, 36(1):2–6, 2004.

[37] B. Vasilescu, A. Serebrenik, P. Devanbu, and V. Filkov. How social q&a sites are changing knowledge sharing in open source software communities. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 342–354. ACM, 2014.

[38] J. Vassileva, C. Gutwin, et al. Exploring blog archives with interactive visualization. In *Proceedings of the working conference on Advanced visual interfaces*, pages 39–46. ACM, 2008.

[39] K. Verbert, E. Duval, J. Klerkx, S. Govaerts, and J. L. Santos. Learning analytics dashboard applications. *American Behavioral Scientist*, page 0002764213479363, 2013.

[40] F. Wei, S. Liu, Y. Song, S. Pan, M. X. Zhou, W. Qian, L. Shi, L. Tan, and Q. Zhang. Tiara: a visual exploratory text analytic system. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 153–162. ACM, 2010.

[41] A. F. Wise, Y. Zhao, and S. N. Hausknecht. Learning analytics for online discussions: a pedagogical model for intervention with embedded and extracted analytics. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, pages 48–56. ACM, 2013.

[42] T. Zimmermann, A. Zeller, P. Weissgerber, and S. Diehl. Mining version histories to guide software changes. *Software Engineering, IEEE Transactions on*, 31(6):429–445, 2005.

Exploring Inquiry-Based Learning Analytics through Interactive Surfaces

Sven Charleer, Joris Klerkx, and Erik Duval
Dept. of Computer Science
KU Leuven
Celestijnenlaan 200A
3001 Leuven, Belgium
sven.charleer@kuleuven.be, joris.klerkx@kuleuven.be, erik.duval@kuleuven.be

ABSTRACT

Learning Analytics is about collecting traces that learners leave behind and using those traces to improve learning. Dashboard applications can visualize these traces to present learners and teachers with useful information. The work in this paper is based on traces from an inquiry-based learning (IBL) environment, where learners create hypotheses, discuss findings and collect data in the field using mobile devices. We present a work-in-progress that enables teachers and learners to gather around an interactive tabletop to explore the abundance of learning traces an IBL environment generates, and help collaboratively make sense of them, so as to facilitate insights.

Categories and Subject Descriptors

H.5.2 [**Information interfaces and presentation**]: User Interfaces; H.5.n [**Information interfaces and presentation**]: Miscellaneous

General Terms

Design, Human Factors, Experimentation

Keywords

interactive surfaces, learning analytics, learning dashboards, collaboration, reflection, awareness, information visualization, sense-making, inquiry-based learning

1. INTRODUCTION

Similar to the Quantified Self [1] movement, which focuses on collecting user traces and using the data for self-improvement, Learning Analytics can help to understand and optimize (human) learning and the environments in which it occurs [12]. However, capturing learner traces can generate an abundance of data, especially in the context of Massive Open Online Courses (MOOCs) that involve tens to thousands of learners whose activities can be tracked in detail. Reflecting on those traces can help learners to understand what is the optimal setting and context in which they learn best. Teachers can, among other things, use the same traces to find out where learners struggle with what content or activity. Dashboards help present this abundance of data in a way that supports both teachers and learners [14].

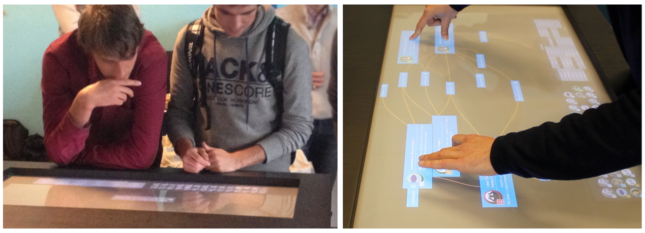


Figure 1: Students gathering around an interactive tabletop, exploring learner traces of a Human-Computer Interaction course.

Teachers show interest in using dashboards collaboratively with learners to discuss their activities, progress and results [3]. Interactive tabletops can facilitate and capture collaboration activities in the classroom [8]. In previous work [2] we explored this platform to visualize learning analytics data (see Figure 1), using the affordances (e.g. large display size, multi-user interaction) of interactive tabletops to create a collaborative sense-making environment [6].

This paper describes our work-in-progress on an interactive tabletop visualization for learner traces that are generated by students in an inquiry-based learning (IBL) environment. Section 2 briefly present the learning environment and the data it generates. Section 3 discusses development details, section 4 explains the design of the tabletop visualization. We discuss our findings and future work in section 5

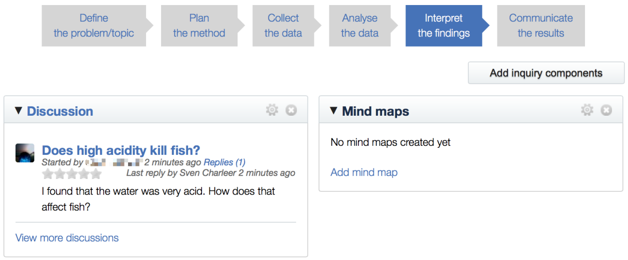
2. IBL LEARNING TRACES

Contrary to a traditional passive role in a classroom, in Inquiry-Based Learning (IBL), learners assume an active role as explorer and scientist with a focus on learning "how to learn". Teachers try to stimulate learners to pose questions and create hypotheses regarding a specific topic, perform independent investigations, gather data to confirm and

---

[1] `http://quantifiedself.com`

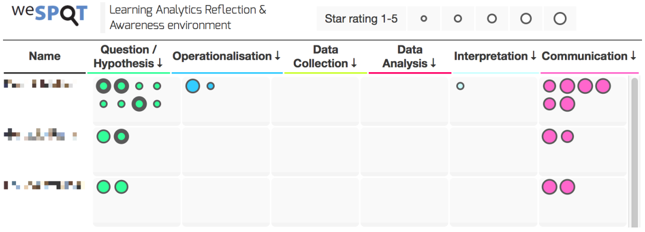
Published in CEUR-WS:
Proceedings of the VISLA15 Workshop on Visual Aspects of Learning Analytics (Duval et al., eds), organized at LAK15.



Does high acidity kill fish?
Started by [redacted] 2 minutes ago Replies (1)
Last reply by Sven Charleer 2 minutes ago
I found that the water was very acid. How does that affect fish?
View more discussions

No mind maps created yet
Add mind map

Figure 2: weSPOT Inquiry Environment, presenting 6 phases and 2 active widgets in phase 5 (Interpretation).



[redacted]

[redacted]

[redacted]

Figure 3: A web-based dashboard for teachers and students providing access to learning analytics data per inquiry.

discuss their findings and generate conclusions. 6 phases of learning activities are often discerned in an IBL process model: problem identification, operationalization, data collection, data analysis, interpretation and communication [9]. As each learner can follow his own route through the IBL process, it is obvious that the sequence and length of these phases differ among students. Individual and collaborative reflection is furthermore vital in every phase. Indeed,*"even at the very beginning when students need to develop a question or a hypothesis, they need to reflect upon the question, and evaluate it before they decide to proceed. They also need to reflect while deciding what kind of data they need to collect, how to proceed to data analysis, and how to communicate their results"* [10].

In the weSPOT Inquiry Environment [2], a teacher can set up an inquiry regarding a specific research topic. For each phase, learners can use a set of widgets (see Figure 2) to e.g. create hypotheses, ask questions, rate and comment on activities, generate mind-maps, etc. By taking pictures, recording videos, entering text and data from measurements through a mobile application, students collect data in the field to support their hypotheses. All activities in the learning environment are logged and stored in a data store and exposed as learning traces through REST services. Teachers and students can access the learning analytics data of a specific inquiry through a web-based dashboard integrated in weSPOT Inquiry Environment 4, and the tabletop application.

3. ITERATIVE DEVELOPMENT

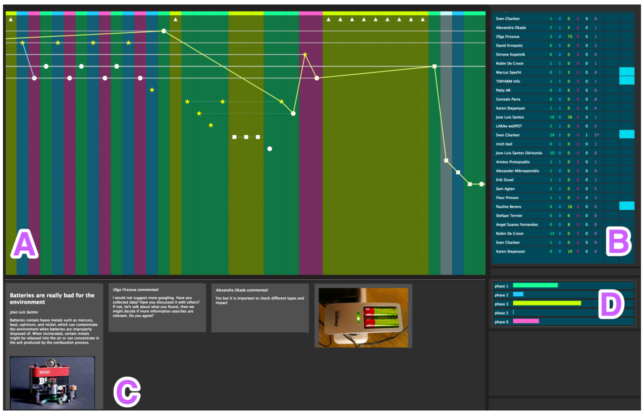


Figure 4: A. The overview of all activities. B. The list of students participating in the inquiry (with student filter options). C. The content behind selected activities. D. Phase filter options.

Following a user-centered rapid prototyping approach, we started from paper prototypes to gather initial feedback on early ideas, gradually developed more functional digital prototypes which have been deployed and evaluated with learners regarding usability.

Web technologies (HTML, CSS3 and JavaScript) facilitate development of quick prototypes and allows us to deploy on most school infrastructures. Interaction is supported through both native browser mouse/touch events and the npTUIClient plug-in [3], allowing the application to run on interactive tabletops, interactive white-boards, tablets, phones and desktop computers. Our interactive tabletop setup currently facilitates up to 5 users.

A centralized filter system using Crossfilter [4] and a modular and event-based architecture facilitates easy creation of new widgets. D3.js [5] and Processing.js [6] help visualize the data. A Node.js [7] back-end generates the web pages while fetching the learning traces from the weSPOT environment.

4. DESIGN

Flexible visual analysis tools must provide appropriate controls for specifying the data and views of interest [5]. Filtering out unrelated information to focus on relevant items is the key control in our learning dashboards due to the abundance of traces learners leave behind. Previous work [3] has shown that there is also a need for context and content to complement the visualized data. We therefore follow the visual information-seeking mantra of "Overview first, zoom and filter, then details-on-demand" [11]: our tabletop visualization presents users with a coordinated set of widgets which contain: (i) a complete overview of all activities (Figure 4.A), (ii) data filters (Figure 4.B/D) and (iii) the content view (Figure 4.C).

---

[2] `http://inquiry.wespot.net/`

[3] `https://github.com/fajran/npTuioClient`

[4] `http://square.github.io/crossfilter/`

[5] `http://d3js.org`

[6] `http://processingjs.org`

[7] `http://nodejs.org`

Published in CEUR-WS:
Proceedings of the VISLA15 Workshop on Visual Aspects of Learning Analytics (Duval et al., eds), organized at LAK15.

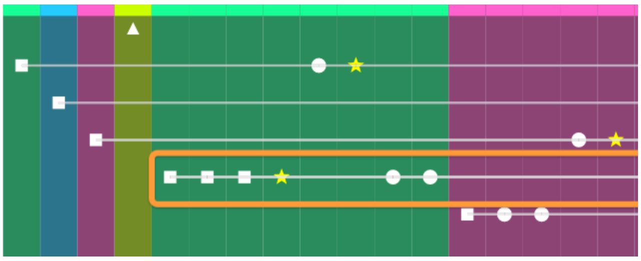


Figure 5: Time-lines per activity thread. The highlighted thread consist of a hypothesis creation followed by 2 edits, a user rating and 2 comments.

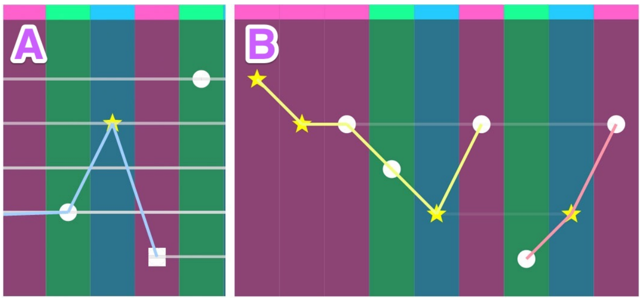


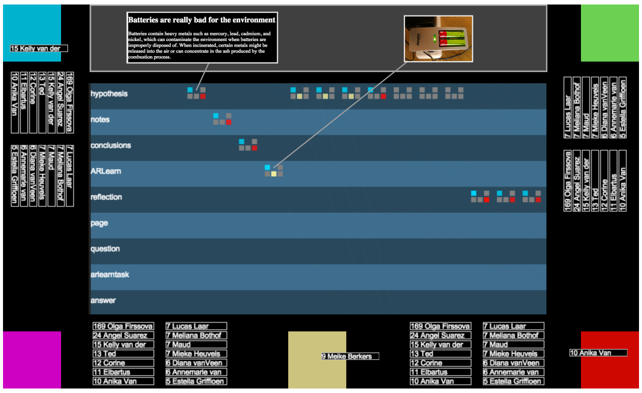
Figure 6: A. The blue path indicates the steps taken by a student. In this case, the student learned something which he then rated. This then lead to the creation of a new hypothesis. B. Visualization limited to a group of 2 students. Individual paths are highlighted. The student indicated by the yellow line has been more active with both commenting and rating activities. The student has also been more active in phase 6 (purple).

4.1 Visualizing IBL Traces

The visualization displays a time-line per *activity thread* (see Figure 5). For instance, the creation of a hypothesis by a learner is followed by every comment on, rating on, and edit of the hypothesis. Squares represent *create* and *edit* events, while circles represent *comment* events. Stars represent a *rating* activity, triangles are *data collection* events. Activities within a single thread are connected by a horizontal line. This enables teachers and learners to see the evolution of an *activity thread*, the comments that may have impacted edits of e.g. the original hypothesis, and the rating trend.

Activities in other *activity threads* can enrich the context of a specific thread. A discussion in one thread might influence the creation of a new hypothesis, or an edit of an existing one. Therefore, every activity is positioned relative in time to all other activities displayed, allowing the users to backtrack through time across multiple threads at once (see Figure 6.A).

IBL phases (see Section 2) in which an activity occurs are indicated by different background colors, matching the colors used of the web dashboard (see Figure 4). The visualization can be panned and zoomed using standard multi-touch interactions.



Batteries are really bad for the environment

Batteries contain heavy metals such as mercury, lead, cadmium, and nickel, which can contaminate the environment when batteries are improperly disposed of. When incinerated, certain metals might be released into the air or can concentrate in the ash produced by the combustion process.

Figure 7: A prototype with 5 filter "drop zones". Dropping a filter value into the blue (top-left) drop zone highlights data points matching the filter result by coloring the top-left part of the glyph.

4.2 Filtering the Data

Using the filter widgets, users can focus on activities by drilling down on one or more phases (see Figure 4.D), or one or more learners (see Figure 4.B). When multiple learners are selected (e.g. a group that works together), the path of each learner can be individually highlighted (see Figure 6.B), in order to provide an overview of work distribution. This can help teachers to find struggling learners in a group. It can also help learners to become aware of uneven work distribution and help to redivide the work. The path can also shed light on the methodology a learner uses to reach a certain result (e.g. Figure 4.A).

The interface of Figure 4 is limited to one person driving the navigation and only supports global filters. To fully use the affordances of the tabletop and create a collaborative sense-making environment, the application must support both individual as well as group work [4]. Figure 7 shows an early prototype that presents 5 participants with individual filtering tools. Global filters result in more tightly coupled collaboration [13], but can disturb individual work. One participant's filter activity could remove data from the visualization another participant is working with. To allow participants to simultaneously filter the data presented on the tabletop, we use the multivariate attributes of a glyph-based visualization [1]. The filter result of each participant is highlighted in the color corresponding to the user interface.

5. CONCLUSION AND FUTURE WORK

Our interactive visualization will be deployed in multiple secondary school pilots [8] across Europe, both on interactive tabletop devices and interactive white-boards. Questionnaires regarding usefulness for both teachers and students will help evaluate our design choices, while interaction logging and video recordings of collaboration sessions can provide insights in whether the application is useful as a sense-making environment.

Our application lets users retrace individual steps taken by (groups of) learner(s), i.e. they can collaboratively (i) re-

---

[8] `http://portal.ou.nl/web/wespot/pilots`

flect on the rationale of a learner's decisions and actions, (ii) (re-)examine past explanations and conclusions, and (iii) (re-)evaluate past evidence data. Students can learn from peers' activities through exploration, discovery and discussion. The application can be used for evaluation purposes, allowing (groups of) learner(s) and teacher(s) to iterate over every step performed from hypothesis to conclusion together. Pilot data can also help IBL researchers with the discussion and refinement of the IBL model.

Enabling multiple learners and teachers to interact with the visualization simultaneously remains the biggest challenge. We shall further explore the possibilities of glyph-based visualizations to provide unobtrusive global filters, use user position tracking through technology such as Kinect to support the dynamic nature of collaborators around a tabletop and explore data lenses (e.g. GeoLens [15, 7]) to facilitate individual exploration of the data on a shared visualization.

6. ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement No 318499 - weSPOT project.

7. REFERENCES

[1] R. Borgo, J. Kehrer, D. H. S. Chung, E. Maguire, R. S. Laramee, H. Hauser, M. Ward, and M. Chen. Glyph-based Visualization: Foundations, Design Guidelines, Techniques and Applications. In M. Sbert and L. Szirmay-Kalos, editors, *Eurographics 2013 - State of the Art Reports*. The Eurographics Association, 2012.

[2] S. Charleer, J. Klerkx, J. L. Santos, and E. Duval. Improving awareness and reflection through collaborative, interactive visualizations of badges. In M. Kravcik, B. R. Krogstie, A. Moore, V. Pammer, L. Pannese, M. Prilla, W. Reinhardt, and T. D. Ullmann, editors, *ARTEL@EC-TEL*, volume 1103 of *CEUR Workshop Proceedings*, pages 69–81. CEUR-WS.org, 2013.

[3] S. Charleer, J. Santos, J. Klerkx, and E. Duval. Improving teacher awareness through activity, badge and content visualizations. In Y. Cao, T. Valjataga, J. K. Tang, H. Leung, and M. Laanpere, editors, *New Horizons in Web Based Learning*, Lecture Notes in Computer Science, pages 143–152. Springer International Publishing, 2014.

[4] C. Gutwin and S. Greenberg. Design for individuals, design for groups: Tradeoffs between power and workspace awareness. In *Proceedings of the 1998 ACM Conference on Computer Supported Cooperative Work*, CSCW '98, pages 207–216, New York, NY, USA, 1998. ACM.

[5] J. Heer and B. Shneiderman. Interactive dynamics for visual analysis. *Queue*, 10(2):30:30–30:55, Feb. 2012.

[6] P. Isenberg, N. Elmqvist, J. Scholtz, D. Cernea, K.-L. Ma, and H. Hagen. Collaborative visualization: Definition, challenges, and research agenda. *Information Visualization*, 10(4):310–326, 2011.

[7] F. Marinho Rodrigues, T. Seyed, F. Maurer, and S. Carpendale. Bancada: Using mobile zoomable lenses for geospatial exploration. In *Proceedings of the Ninth ACM International Conference on Interactive Tabletops and Surfaces*, ITS '14, pages 409–414, New York, NY, USA, 2014. ACM.

[8] R. Martinez-Maldonado, K. Yacef, Y. Dimitriadis, M. Edbauer, and J. Kay. MTClassroom and MTDashboard: supporting analysis of teacher attention in an orchestrated multi-tabletop classroom. In *International Conference on Computer-Supported Collaborative Learning, CSCL 2013*, pages 119–128, 2013.

[9] A. Mikroyannidis, A. Okada, P. Scott, E. Rusman, M. Specht, K. Stefanov, P. Boytchev, A. Protopsaltis, P. Held, S. Hetzner, K. Kikis-Papadakis, and F. Chaimala. weSPOT: A Personal and Social Approach to Inquiry-Based Learning. *Journal of Universal Computer Science*, 19(14):2093–2111, 2013.

[10] A. Protopsaltis, P. Seitlinger, F. Chaimala, O. Firssova, S. Hetzner, K. Kikis-Papadakis, and P. Boytchev. Working environment with social and personal open tools for inquiry based learning: Pedagogic and diagnostic frameworks. *The International Journal of Science, Mathematics and Technology Learning*, 20(4):51–63, 2014.

[11] B. Shneiderman. The eyes have it: a task by data type taxonomy for information visualizations. In *IEEE Symposium on Visual Languages*, pages 336–343. IEEE, 1996.

[12] G. Siemens and P. Long. Penetrating the fog: Analytics in learning and education. volume 46, pages 30–32, Boulder, CO, USA, 2011. EDUCAUSE.

[13] A. Tang, M. Tory, B. Po, P. Neumann, and S. Carpendale. Collaborative coupling over tabletop displays. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '06, pages 1181–1190, New York, NY, USA, 2006. ACM.

[14] K. Verbert, S. Govaerts, E. Duval, J. Santos, F. Van Assche, G. Parra, and J. Klerkx. Learning dashboards: an overview and future research opportunities. *Personal and Ubiquitous Computing*, 18(6):1499–1514, 2014.

[15] U. von Zadow, F. Daiber, J. Schöning, and A. Krüger. GeoLens: Multi-User Interaction with Rich Geographic Information. *Proc. DEXIS 2011*, pages 16–19, 2012.

Uncovering Learning Processes Using Competence-based Knowledge Structuring and Hasse Diagrams

Michael D. Kickmeier-Rust
Graz University of Technology
Knowledge Technologies Institute
8010 Graz, Austria
+43 316 873 30636
michael.kickmeier-rust@tugraz.at

Christina M. Steiner
Graz University of Technology
Knowledge Technologies Institute
8010 Graz, Austria
+43 316 873 30640
christina.steiner@tugraz.at

Dietrich Albert
Graz University of Technology
Knowledge Technologies Institute
8010 Graz, Austria
+43 316 873 30640
dietrich.albert@tugraz.at

ABSTRACT

Learning analytics means gathering a broad range of data, bringing the various sources together, and analyzing them. However, to draw educational insights from the results of the analyses, these results must be visualized and presented to the educators and learners. This task is often accomplished by using dashboards equipped with conventional and often simple visualizations such as bar charts or traffic lights. In this paper we want to introduce a method for utilizing the strengths of directed graphs, namely Hasse diagrams, and a competence-oriented approach of structuring knowledge and learning domains. After a brief theoretical introduction, this paper highlights and discusses potential advantages and gives an outlook to recent challenges for research.

Keywords

Learning analytics, data visualization, Hasse diagram, Competence-based Knowledge Space Theory.

1. INTRODUCTION

Using methods and tools from Learning Analytics (LA) can be considered best practice and is a key factor for making education more personalized, adaptive, and effective. Analyzing a variety of available data to uncover learning processes, strengths and weaknesses, competence gaps undoubtedly is a prerequisite for a formatively-inspired guidance, for changing and adjusting educational measures and teaching, and not least for disclosing and negotiating learner models [4]. Usually, the benefits are seen in the potential to reduce attrition through early risk identification, improve learning performance and achievement levels, enable a more effective use of teaching time, and improve learning design and instructional design [10]. On the basis of available data, ideally large scale data sets, smart tools and systems are being developed to provide teachers with effective, intuitive, and easy to understand aggregations of data and the related visualizations. There is a substantial amount of work going on this particular field; visualization techniques and dashboards are broadly available (cf. [2,4,7]), ranging from simple meter/gauge-based techniques (e.g., in form of traffic lights, smiley, or bar charts) to more sophisticated activity and network illustrations (e.g., radar charts or hyperbolic network trees).

However, LA operates in a delicate and complex area. On the one hand, facing today's classroom realities, we often find technology-lean environments, which do not easily allow or support recording the necessary data. Also, from a socio-pedagogical perspective, learning must be seen as a process of social interaction that not always occurs in front of some electronic. Thus, LA must be based on fewer data. On the other hand, it is rather easy to visualize learning on a superficial level using perhaps the aforementioned traffic lights or bar charts. The added value to the teachers is likely of limited utility to them. To provide a deeper and more formative insight into the learning history and the current state of a learner (beyond the degree to which a teacher might know it intuitively) requires finding and presenting complex data aggregations. This, most often, bears the significant downside that it is hard to understand. Challenges for LA and its visualizations, for example, are to illustrate learning progress (including learning paths) and - beyond the retrospective view - to display the next meaningful learning steps/topics.

In this paper we introduce the method of directed graphs, the so-called Hasse diagrams, for structuring learning domains and for visualizing the progress of a learner through this domain.

2. HASSE DIAGRAMS AND COMPETENCE-BASED KNOWLEDGE SPACES

A Hasse diagram is a strict mathematical representation of a so-called semi-order in form of a directed graph that reads from bottom to top. A semi-order is a type of mathematical ordering of a set of items with numerical values by identifying two items as equal or comparable if the values are within a given interval of error or noise. Semi-orders were introduced in mathematical psychology by Duncan Luce in 1956 [8] in human decision research without the assumption that indifference is transitive. This approach is also crucial for handling human learning and the resulting performance that is prone to all sorts of errors and peripheral aspects (perhaps failing in a test although the learner holds the knowledge due to being tired). A Hasse diagram is one way of displaying such ordering – in our case competences or competency states (which is to be explained in the following section). The technique was invented in the 60s of the last century by Helmut Hasse. The diagram exists of entities (the nodes), which are connected by relationships (indicated by edges).

The mathematical properties of a semi-order and the Hasse diagrams are (i) reflexivity, (ii) anti-symmetry, and (iii) transitivity. Reflexivity refers to the view that an item, perhaps a competency, references itself in a cause/effect sense. Anti-symmetry demands that if one entity is a prerequisite of another, this relationship is not invertible; as an example, if competency x is a prerequisite to develop competency y, y cannot be the perquisite of competency x. Finally, transitivity means that whenever an element x is related to an element y, and y is in turn related to an element z, then x is also related to z. In principle, the direction of a graph is given by arrows of the edges; by

Published in CEUR-WS:
Proceedings of the VISLA15 Workshop on Visual Aspects of Learning Analytics (Duval et al., eds), organized at LAK15.

convention however, the representation is simplified by avoiding the arrow heads, whereby the direction reads from bottom to top. In addition, the arrows from one element to itself (reflexivity property), as well as all arrows indicating transitivity are not shown in Hasse diagrams. The following image (Figure 1) illustrates such a diagram. Hasse diagrams enable a complete view to (often huge) structures. Insofar, they appear to be ideal for capturing the large competence or learning spaces occurring in the context of assessment and learning recommendations (for example, all the competencies involved in the math curriculum for a specific age).

In an educational context, a Hasse diagram can display the non-linear path through a learning domain starting from an origin at the beginning of an educational episode (which may be a single school lesson but could also be the entire semester). Moreover, the elements in the diagram may refer to (latent) competencies, to learning objects or test items. Figure 1 illustrates the simple example of typical learning objects in a certain domain. The beginning of a learning episode is usually shown as { } (the empty set) at the bottom of the diagram. Now a learner might attend three learning objects (K, P, H), which is indicated by the edges; this, in essence, establishes three possible learning paths. After H, as an example, this learner might attend K, or H but not T yet, which in turn opens further three branches for the learning path until reaching the final state, within which all learning objects have been attended.

As claimed initially, in the context of formative LA, a competence-oriented approach is necessary. Thus, a Hasse diagram can be used to identify and display the latent competencies of a learner in the form of so-called competence states. An elaborated theoretical approach to do so is Competence-based Knowledge Space Theory (CbKST). The approach originates from Jean-Paul Doignon and Jean-Claude Falmagne [5, 6] and is a mathematical psychological, set-theoretic framework for addressing the relations among problems (e.g., test items). It provides a basis for structuring a domain of knowledge and for representing the knowledge based on prerequisite relations. While the original Knowledge Space Theory focuses only on performance (the behavior; for example, solving a test item), its extension CbKST [1] introduces a separation of observable performance and latent, unobservable competencies, which determine the performance [1]. This is a psychological learning-theoretical approach, which highlights that competencies (e.g., the ability to add two integers) are unobservable latent constructs and which can only be observed or assessed indirectly.

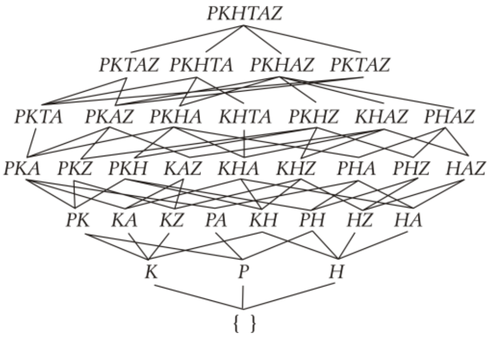

PKHTAZ
PKTAZ PKHTA PKHAZ PKTAZ
PKTA PKAZ PKHA KHTA PKHZ KHAZ PHAZ
PKA PKZ PKH KAZ KHA KHZ PHA PHZ HAZ
PK KA KZ PA KH PH HZ HA
K P H
{ }

Figure 1. A simple Hasse diagram.

We interpret the performance of a learner (e.g., mastering an addition task) in terms of holding or not holding the respective competency. In addition, recent developments of the approach are based on a probabilistic view of having or lacking certain competencies. In our example, mastering one specific addition task allows the conclusion that the person is able to add two numbers (to hold this competency) only to a certain degree or probability. When thinking of a multiple-choice item with two alternatives, as another example, mastering this item allows only to 50 percent that the person has the required competencies/ knowledge.

On the basis of these fundamental views, CbKST is looking for the involved entities of aptitude (the competencies) and a natural structure, a natural course of learning in a given domain. For example, it is reasonable to start with the basics (e.g., the competency to add numbers) and increasingly advance in the learning domain (to subtraction, multiplication, division, etc.). As indicated above, this natural course is not necessary linear, which bears significant advantages over other learning and test theories.

As a result we have a set of competencies in a domain and potential relationships between them. In terms of learning, the relationships define the course of learning and thus which competencies are learned before others. In CbKST such relationships are called prerequisite relations or precedence relations. On the basis of competencies and relationships, in a next step, we can obtain a so-called competence space, the ordered set of all meaningful competence states a learner can be in. As an example, a learner might have none of the competencies, or might be able to add and subtract numbers; other states, in turn, are not included in this space, for example it is not reasonable to assume that a learner holds the competency to multiply numbers but not to add them. By the logic of CbKST, each learner is, with certain likelihood, in one of the competence states.

3. VISUALIZING COMPETENCE SPACES

As claimed, Hasse diagrams are capable of holding a number of important information for an educator to evaluate the learning progress and also to make recommendations. In this paper we want to highlight such advantages.

3.1 Competence States and Levels

As outlined, a competency space is the collection of meaningful states a learner can be in. Depending on the domain, the amount of possible states might be huge. The big advantage, however, is that depending on the degree of structure in the domain, by far not all possible combinations of competencies are reasonable and thus part of the space. When zooming into the diagram, a teacher can exactly identify the set of competencies that is most likely for the learner, by zooming out color-coding can illustrate the most likely locations of a learner within the space. When looking at the entire space, it is obvious at first site at which completion level a learner is approximately (rather at the beginning or almost finished). These zoom levels are shown in Figure 2. Technically, there is a variety of options to achieve the coding, for example, bolding, greying, or color coding, whereas likely states are displayed more distinctly than such with low probability.

Equal to individual states, Hasse diagrams can represent group distributions. Defined by a certain confidence interval of probabilities those states and areas can be made more salient that hold the highest percentage of learners of a group. By this means,

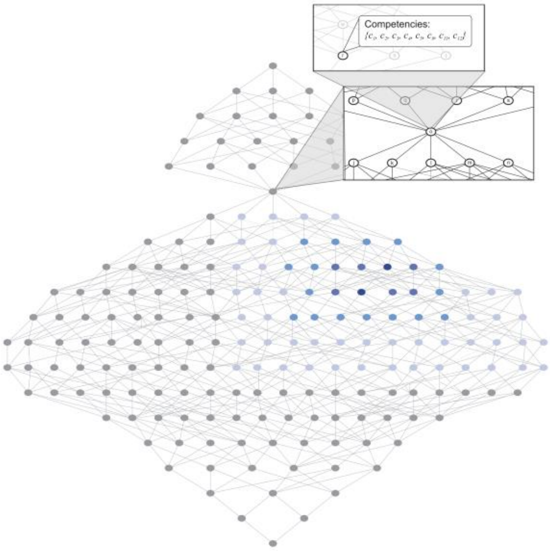


Figure 2. Hasse diagram illustrating the probability distribution over a competence space on three zoom levels.

specific areas in the competency space become apparent within which the most learners are and, in contrast also positive or negative outliners pop out the diagram. A different method was suggested by [9], who altered the size of the nodes to represent the groups' sizes; the larger a node the more learners hold a particular state.

3.2 Learning Paths

In addition to having insight into groups' and individuals' current states of learning, the learning history, the so-called learning paths, are of interested for educators; on the one hand for planning future activities, on the other hand, for negotiation and documenting the achievements of a learning episode (e.g., a semester). Learning paths can be simply displayed by highlighting the edges between the most likely state(s) over time. As for the states, various probable paths can be realized by making more likely paths more intensive (by color coding or line thickness). Figure 3 shows a simple example. A key strength of presenting learning paths, as indicated, is opening up the learner model to the learners (perhaps parents) themselves [9] – to explain where they started at the beginning of a course and how they proceeded during the course and which competencies they hold today. This perhaps can be complemented with comparisons to others or groups. Not least, learning paths can unveil information about the effectiveness and impact of certain learning activities, materials, or the teacher herself.

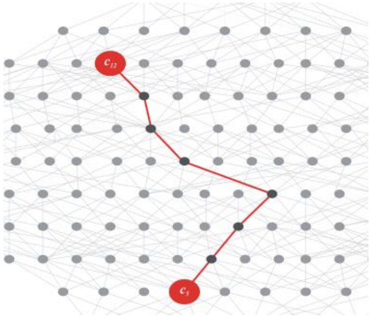


Figure 3. Learning Path. The cutout is part of the structure shown in Figure 2.

3.3 Tests and Recommendations

Hasse diagram offers information about two very distinct concepts, the inner and outer fringes. The inner fringe indicates what a learner can do / knows at the moment. Mathematically it refers to all sets of competencies, which hold all competencies of the current state but one. This inner fringe is a clear hypothesis of which test/assessment items this learner can master within the margins of a certain probability. Such information may be used to generate effective and individualized tests. The test generation can be complemented with group information. If an educator has very clear information in which competency areas of the space most of the learners are, she can generate or select test item covering exactly those competencies. The big advantage of such approach is the effectiveness of a test for identifying competency states or for ranking the learners can be maximized while the efforts for this evaluation (e.g., the number of test items) can be minimized. And of course the test can be optimized to differentiate different learners and the individual capabilities.

On the other hand, the outer fringes determine which competencies should be addressed in a next educational step. Mathematically is refers to all states which include all the competencies of the current state plus one. These fringes provide a clear set of recommendations about the most effective learning activities for a specific individual or a specific group of learners. Moreover, outer fringes, together with learning paths, allow specifically planning the most effective ways of reaching a specific learning goal (which not necessarily is the final stage of the competence space, the full set, and which is not necessarily the same goal for all individual learners).

3.4 Costs and Pace

When supporting teachers with information about learning processes, the concept of costs or learning pace (sometimes referred to as learning trajectories) is of distinct importance. Cost and pace can be considered as the time or any other measure of effort it takes to proceed from one competence state to another. In a Hasse diagram this information can be displayed by varying the length of the edges accordingly. If an educational leap requires a lot of efforts or time the edges are displayed proportionally longer than such that happens rather quickly. This method was introduced initially by [9]; an example is shown in Figure 4. Such information unveils criteria for the effectiveness of certain learning materials or acts of teaching. Particular outliers obviously pop out of the diagram and call educators to action to adapt teaching or teaching materials for a specific individual or a group.

3.5 Subordinate Concepts and General Notions of Achievement, Bottlenecks

A further important aspect in the context of LA is aligning the rather fine grained and low level approach to view competencies on a deeper level of granularity to more general concepts or rather superordinate notions of achievement. A general concept can be considered a higher level cluster of competencies; for example, sub-dividing mathematics into clusters like linear equations, non-linear equations, and vector arithmetic. Lower level competencies can be linked to one or more of those 'chapters'. Equally, one might view learning processes in a domain in terms of maturity. For example, writing skills can be on a low level of maturity, involving certain competencies and abilities, and on a higher one. Such approach is given, for example, in the CEFR language skills (cf. http://en.wikipedia.org/wiki/Common_European_Framework_of_Reference_for_Languages). Finally, teaching might involve the achievement of certain milestones, which should be reached step by step. Hasse diagrams allow identifying such milestones even if they were unclear or unknown initially. Considering that milestones as bottlenecks, i.e. unique competence states, each learning must pass, such bottlenecks immediately pop out in of the diagram. In a formative sense, it is easy for an educator to located their learners in their approach to or exceeding of such milestones (cf. Figure 2). A slightly different variant was introduced by [9] who used additional graphical elements (e.g., intersecting lines) to separate certain levels of maturity (whereas these authors used the CMMI[1] method; cf. Figure 5).

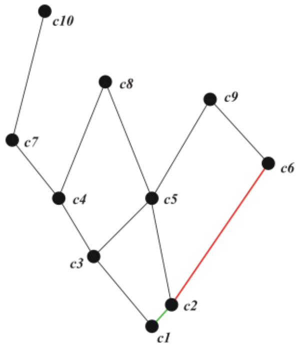


Figure 4. Illustrating learning efforts (as costs or pace). The longer the more efforts/time it took to acquire a further competency.

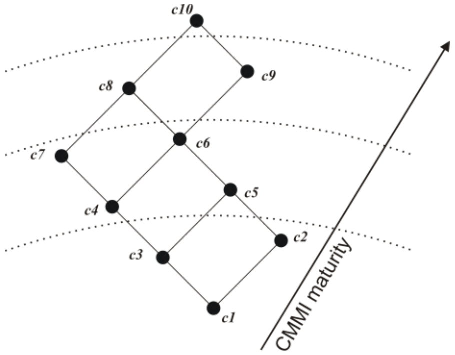


Figure 5. Illustrating maturity levels.

4. WHERE DO DATA COME FROM?

The features of Hasse diagrams and the arising advantages for LA appear all well and good. However, the key question is, where do they data for computing the probabilities of competence states come from. And everything stands or falls with this question. As for all techniques of LA, it depends on a data rich approach to education, the more and the better data exist, the better is the quality of LA conclusions. CbKST and Hasse diagrams are no exception to that. However, the approach of separating latent competencies, which more or less develop and exist in the black box 'human brain', and the performance they determine, bears particular advantages. On the one hand, performance, e.g. test scores, classroom participation, homework, etc., is not only determined by competencies or aptitude; there is a variety of aspects contributing to a certain performance, e.g., motivation, daily constitution, tiredness, external distractors, nutrition, health status, etc. On the other hand, CbKST-ish competence spaces are rather stable, once set up and validated properly. The advantage lays in the fact that performance such as test results, behaviors, achievements, etc. is considered as probability-based indicators for certain competencies. Mathematically this relationship is established in form of interpretation and representation functions [1], which links an arbitrary set of performances/behaviors to one or more competencies, either in an increasing or in a decreasing sense. This, in the end, allows linking all available and perhaps changing data sources to one and the same competence space. It's not about a single test, it's about all available information we can gather, even it is considered being of little importance, all sorts of information may contribute to strengthen the model, the view of the learner. In case the amount or quality of data is weak, CbKST allows conservative interpretations, based on the arising probability distributions, in case there is a richer data basis, the probability distributions are more reliable, valid, and robust. For the educator, and this is important, the uncertainty is mirrored in the degree of likelihood. On a weak data basis, the probabilities of competence states differ substantially less than on the basis of richer data. Such information, however, can change the educator's view and evaluation of a student's achievements. In the end, this approach supports a fairer and more substantiated approach to grading or providing formatively inspired feedback.

5. CONCLUSIONS AND OUTLOOK

There is little doubt that frameworks, techniques, and tools for LA will increasingly be part of a teacher's professional life in the near future. The benefits are convincing – using the (partly massive) amount of available data from the students in a smart, automated, and effective way, supported by intelligent systems in order to have all the relevant information available just in time and at first sight. The ultimate goal is to formatively evaluate individual achievements and competencies and provide the learners with the best possible individual support and teaching. Great. The idea of formative assessment and educational data mining is not new but the hype over recent years resulted in scientific sound and robust approaches becoming available, and usable software products appeared. However, when surveying the educational landscape, at least that of the EU, the educational daily routines are different. We face technology-lean classrooms and schools, we face a lack of proper teacher education in using ICT in schools – not mentioning of using techniques of LA in schools. We face a certain aloofness to use breaking educational technologies and a well-founded pedagogical view that learning ideally is analogous and socially embedded and doesn't occur in front of some kind of

---

[1] CMMI refers to the so-called *Capability Maturity Model Integration* approach which models development processes (e.g., in production) on different predefined levels [3].

electronic device. These are all experiences and results of a large scale European research project named Next-Tell (www.next-tell.eu) that was looking into educationally practices across Europe and that intended to support teachers where exactly they are today with suitable ICT as effective and as appropriately as possible.

The framework of CbKST offers a rigorously competence-based, probabilistic, and multi-source approach that accounts for the latent and holistic abilities of learners and therefore accounts for the recent conceptual change in Europe's educational systems towards a more competence-oriented education including multi-subject competencies and superordinate 21st century (soft) skills.

No matter if data are rich or lean, a teacher is supported to the best possible degree and with a variety of important information about individual and group-based learning processes and performances and not least about the performance of learners and about the educator's own performance. The probabilistic dimension allows teachers to have a more cautious view of individual achievements – it might well be that a learner has a competency but fails in a test; vice versa, a student might luckily guess an answer.

From an application perspective, in the context of European projects we developed and evaluated tools that cover the techniques and approaches described in this paper. In the Next-Tell project, for example, we developed a software tool named ProNIFA, which allowed linking multiple sources of evidence of learning and building CbKST-based learner models. We piloted various school studies and gathered feedback from teachers. In the end, and this can be considered an outlook for future developments, we had to find out that the 'massive' Hasse diagrams are overburdening teachers' understanding and mental models about individual and class-based learning. Moreover, in order to understand the classical Hasse diagrams, it required (too) massive efforts in training teachers to fully utilize the potentials of those diagrams. Large scale surveys yielded that most educators still prefer simple but information-wise shallow visualizations such as traffic lights or bar charts significantly over more information-rich approaches such as Hasse diagrams or, just to mention another interesting approach, parallel coordinates .

Therefore, recent efforts, e.g., in the LEA's BOX (www.leas-box.eu) project, seek to adjust and advance the classical Hasse diagrams to such visualizations that are intuitively understood by educators and, at the same time, hold the same density of information. In particular, focus of research is on an advancement of Hasse diagrams towards specific mental models teachers may hold, such as a starry night sky or organic, biological structures such as cells of a living being. Also, abstraction and simplification techniques are investigated, e.g., fisheye lenses or streamgraphs.

In conclusion, the utility of CbKST-ish approaches to LA, involving a separation of latent competencies and observable behaviors/performance, as well as having a conservative, probabilistic, multi-source approach appears to be a striking classroom-oriented, next-level contribution to LA, learner modelling, and model negotiations.

6. ACKNOWLEDGMENTS

This work is based on the finalized project Next-Tell, which was supported by the European Commission (EC) under the Information Society Technology priority of the 7th Framework Programme for research and development as well as the running LEA's BOX project, contracted under number 619762, of the 7th Framework Programme. This document does not represent the opinion of the EC and the EC is not responsible for any use that might be made of its content.

7. REFERENCES

[1] Albert, D., & Lukas, J. 1999. Knowledge Spaces: Theories, Empirical Research, and Applications. Mahwah, NJ: Lawrence Erlbaum Associates.

[2] Ferguson, R., and Buckingham Shum, S. 2012. Social Learning Analytics: Five Approaches. In Proceedings of the 2nd International Conference on Learning Analytics & Knowledge, 29 Apr - 02 May 2012, Vancouver, British Columbia, Canada.

[3] Forrester, E. C., Buteau, B. L., and Shrum, S. 2009: CMMI for Services. Guidelines for Superior Service. Addison-Wesley.

[4] Dimitrova, V., McCalla, G. and Bull, S. 2007. Open Learner Models: Future Research Directions (Special Issue of IJAIED Part 2), International Journal of Artificial Intelligence in Education 17(3), 217-226.

[5] Doignon, J., & Falmagne, J. 1985. Spaces for the assessment of knowledge. International Journal of Man-Machine Studies, 23, 175–196.

[6] Doignon, J., & Falmagne, J. 1999. Knowledge Spaces. Berlin: Springer.

[7] Duval, E., 2011. Attention Please! Learning Analytics for Visualization and Re-commendation. In Proceedings of the 1st International Conference on Learning Analytics & Knowledge, 27 Feb – 1 March 2011, Banff, Alberta, Canada.

[8] Luce, R. D. 1956. Semiorders and a theory of utility discrimination. Econometric,a 24, 178–191.

[9] Nakamura, Y., Tsuji, H., Seta, K., Hashimoto, K., and Albert, D. 2011. Visualization of Learner's State and Learning Paths with Knowledge Structures. In A. König et al. (Eds.), KES 2011, Part IV. Lecture Notes in Artifical Intelligence 6884, pp. 261-270. Berlin: Springer.

[10] Siemens, G., Gasevic, D., Haythornthwaite, C., Dawson, S., Buckingham Shum, S:, Ferguson, R., Duval, E., Verbert, K., and Baker, R.S..J.D. 2011. Open Learning Analytics: an integrated & modularized platform: Proposal to design, implement and evaluate an open platform to integrate heterogeneous learning analytics techniques. Available online at http://solaresearch.org/OpenLearningAnalytics.pdf

LAK Explorer – A Fusion of Search Tools

Mike Sharkey
Blue Canary
6185 W. Detroit St.
Chandler, AZ 85226 USA
mike@bluecanarydata.com

Mohammed Ansari
Blue Canary
6185 W. Detroit St.
Chandler, AZ 85226 USA
mohammed@bluecanarydata.com

Andy Nguyen
Blue Canary
6185 W. Detroit St.
Chandler, AZ 85226 USA
andy@bluecanarydata.com

ABSTRACT

The LAK Data Challenge asks the question "What do analytics on learning analytics tell us?" One approach to this challenge is not to answer the question, but to provide a simple, user-focused application that allows any user to easily draw their own conclusion. This was Blue Canary's driver for building the LAK Explorer (http://lakexplorer.bluecanarydata.com). Our team combined multiple tools to create a powerful search application. We extracted topics from the papers, used an autocomplete feature in the search bar, added topics as search result metadata, and provided links to similar papers all as part of the user search experience. The value is in the usability. In the same way that Google presents powerful results via a simple interface, LAK Explorer allows for seamless searching, reading, and comparing of over 400 documents. The application is also instrumented to capture user input (search terms, papers viewed) to provide closed-loop analytics in the future.

Categories and Subject Descriptors

- *Computing methodologies~Natural language processing*
- *Computing methodologies~Algorithms*
- *Information systems~User interfaces*

General Terms

Algorithms, Design, Human Factors

Keywords

Search, natural language processing, similarity, document, vector, elastic search, cosine, autocomplete, corpus

1. INTRODUCTION

The LAK Data Challenge asks the question "What do analytics on learning analytics tell us?" The Blue Canary team tackled this question in 2014 by using topic modeling to describe trends in the LAK Corpus [1]. Topic Modeling was a technique used to distill a large corpus of text into a manageable list of topics. While repeating this approach for LAK15 would theoretically yield new results, it wouldn't do much to advance experimentation in the spirit of the LAK Data Challenge.

Building off of previous LAK entries, the Blue Canary team took a different approach. Instead of analyzing the corpus to look for trends and threads, what if we made the corpus more easily searchable so that the analytics community can browse the corpus for meaning?

This was the core of our approach for 2015. The result is the LAK Explorer (http://lakexplorer.bluecanarydata.com). It's an intuitive search application that allows users to search, browse, and find content in the corpus of papers/articles provided by the LAK Data Challenge. Our goal was to automate the processing so that the LAK Explorer could be applied to any corpus, not just specifically tuned to the LAK data.

1.1 Use of Turbo Topics

As will be explained in this paper, the use of Turbo Topics is a key thread to the Blue Canary team's approach. Blei's research [2] allowed the team to use programmatic techniques to extract n-gram topics from the corpus that end up being a much more user-friendly way to digest corpus content.

1.2 LAK Dataset Incomplete

Blue Canary retrieved the LAK Dataset from the Challenge website (http://lak.linkededucation.org/lak/LAK-DATASET-DUMP.rdf.zip). Upon examination, it appeared as if this dataset did not contain the entirety of the updated 2015 content. Of the 579 content tags (<led:body> and <bibo:content>), 108 were empty. Blue Canary inquired about the gaps but at the time of this project submission, the content was not added to the dataset.

2. The LAK Explorer Components

The LAK Explorer is a fusion of existing tools and techniques for interacting with semantic data. From the simple home screen to the detailed neighborhood of papers, each component was used to add utility to the search process. Figure 1 shows how the different tools were used at different points in the search process.

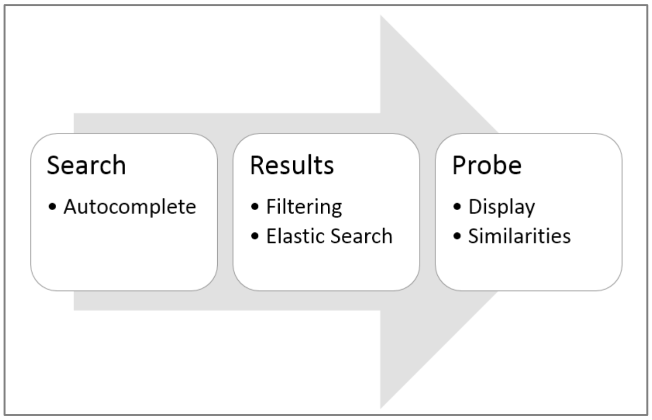


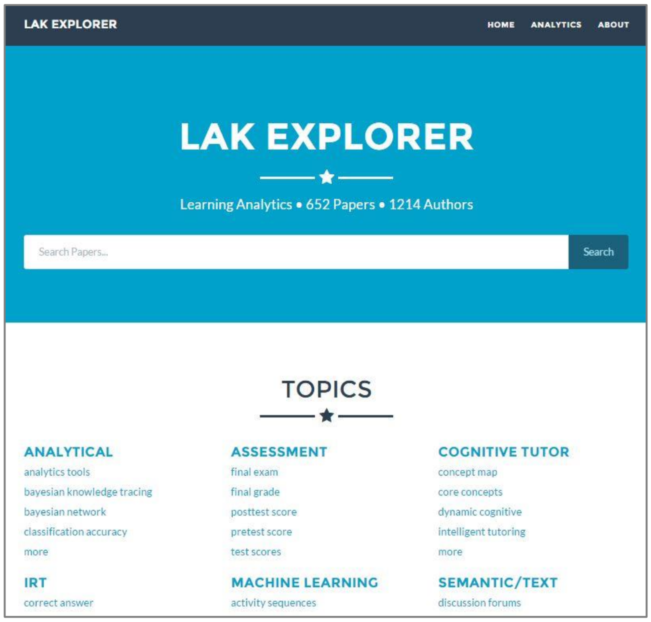
Figure 1. LAK Explorer components

The context for LAK Explorer is similar to that of a search engine. The user comes to the site knowing the universe in which they are searching (corpus of papers) and some idea as to what they want to find out (search terms). However, the user doesn't know exactly what they are looking for. In that way, the presentation of results and related information is vital to improving the utility of the application.

Published in CEUR-WS:
Proceedings of the VISLA15 Workshop on Visual Aspects of Learning Analytics (Duval et al., eds), organized at LAK15.

2.1 Home Page

As with any search engine, the usability of the tool starts with the home page. For LAK Explorer, the Blue Canary team was inspired by the simplicity of Google's ubiquitous home page.



LAK EXPLORER

Learning Analytics • 652 Papers • 1214 Authors

TOPICS

ANALYTICAL	ASSESSMENT	COGNITIVE TUTOR
analytics tools	final exam	concept map
bayesian knowledge tracing	final grade	core concepts
bayesian network	posttest score	dynamic cognitive
classification accuracy	pretest score	intelligent tutoring
more	test scores	more
IRT	**MACHINE LEARNING**	**SEMANTIC/TEXT**
correct answer	activity sequences	discussion forums

Figure 2. LAK Explorer home page

The home page is dominated by a large text entry box. The only other significant feature on the page is a listing of topics that were extracted from the corpus of papers.

2.2 Autocomplete

When a user starts entering text into the search box, the first thing they will notice is the use of autocomplete.

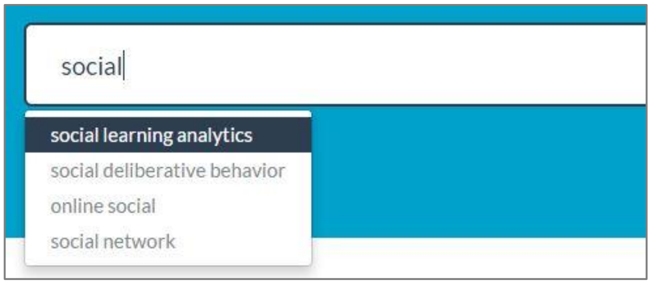


Figure 3. Using autocomplete

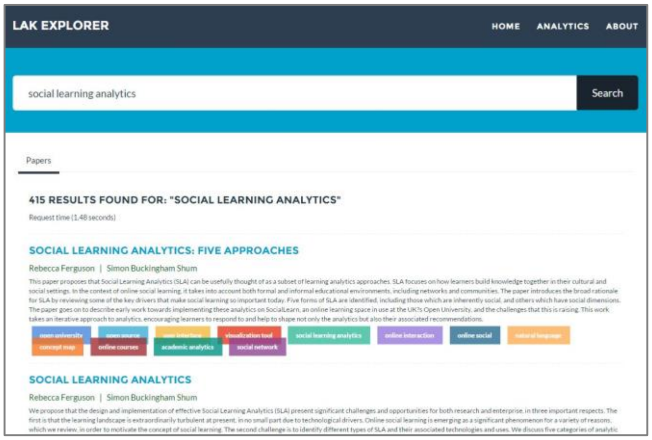
Autocomplete is advantageous since the LAK Explorer deals with a fixed corpus of knowledge. Instead of tying autocomplete to a larger base of content (e.g. dictionary or DBPedia), the team tied it to the topics that were extracted from the corpus using Turbo Topics – the same topics that appear under the search bar. Blue Canary used the Typehead feature from AngularStrap (http://mgcrea.github.io/angular-strap/#/typeaheads#typeaheads) to power this feature.

2.3 Elastic Search

Elastic Search (http://www.elasticsearch.org/) was used to drive the search engine results in LAK Explorer. Blue Canary only used basic features of this tool to drive search results. The text entered in the search box are the inputs to a keyword search algorithm. There are additional features in Elastic Search that could further drive the efficacy of the search and leverage the linked aspect of the LAK data. The search results could be weighted on content found in the content abstract, body, the author(s), and citations.

2.4 Results Filtering

When a user hits the search button, LAK Explorer returns the results similar to the image in Figure 4.



415 RESULTS FOUND FOR: "SOCIAL LEARNING ANALYTICS"

Request time (1.48 seconds)

SOCIAL LEARNING ANALYTICS: FIVE APPROACHES

Rebecca Ferguson | Simon Buckingham Shum

This paper proposes that Social Learning Analytics (SLA) can be usefully thought of as a subset of learning analytics approaches. SLA focuses on how learners build knowledge together in their cultural and social settings. In the context of online social learning, it takes into account both formal and informal educational environments, including networks and communities. The paper introduces the broad rationale for SLA by reviewing some of the key drivers that make social learning so important today. Five forms of SLA are identified, including those which are inherently social, and others which have social dimensions. The paper goes on to describe early work towards implementing these analytics on SocialLearn, an online learning space in use at the UK?s Open University, and the challenges that this is raising. This work takes an iterative approach to analytics, encouraging learners to respond to and help to shape not only the analytics but also their associated recommendations.

SOCIAL LEARNING ANALYTICS

Rebecca Ferguson | Simon Buckingham Shum

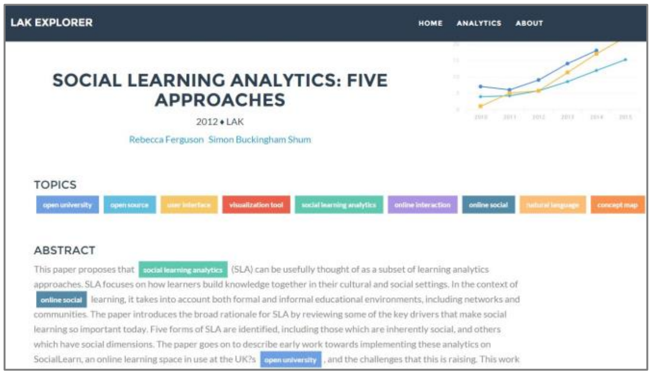
We propose that the design and implementation of effective Social Learning Analytics (SLA) present significant challenges and opportunities for both research and enterprise, in three important respects. The first is that the learning landscape is extraordinarily turbulent at present, in no small part due to technological drivers. Online social learning is emerging as a significant phenomenon for a variety of reasons, which we review, in order to motivate the concept of social learning. The second challenge is to identify different types of SLA and their associated technologies and uses. We discuss five categories of analytic

Figure 4. The search results page

The papers/articles are presented in a fashion similar to an engine like Google Scholar. The title, author(s), and abstract are presented for each result. The most prominent addition is the inclusion of the color-coded topics. LAK Explorer uses the extracted topics as another level of filtering. Upon viewing the results, the user can see what topics are relevant for each paper, click the color-coded topic, and get new search results that are sorted by the frequency of that topic.

2.5 Paper Display

The utility of LAK Explorer is not to just browse search results. Blue Canary wanted to make the tool helpful for actually reading the resulting papers and articles. Therefore, clicking on a paper from the search results gives the display mode shown in Figure 5.



SOCIAL LEARNING ANALYTICS: FIVE APPROACHES

2012 • LAK

Rebecca Ferguson Simon Buckingham Shum

TOPICS

ABSTRACT

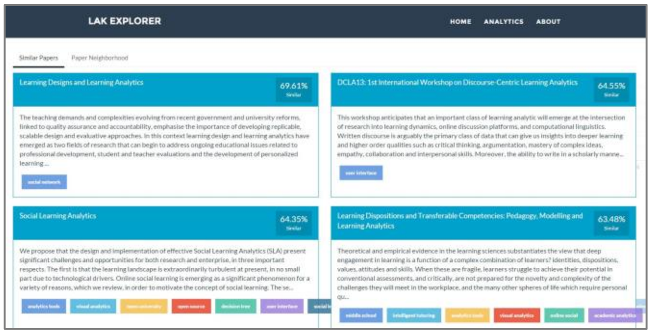
This paper proposes that social learning analytics (SLA) can be usefully thought of as a subset of learning analytics approaches. SLA focuses on how learners build knowledge together in their cultural and social settings. In the context of online social learning, it takes into account both formal and informal educational environments, including networks and communities. The paper introduces the broad rationale for SLA by reviewing some of the key drivers that make social learning so important today. Five forms of SLA are identified, including those which are inherently social, and others which have social dimensions. The paper goes on to describe early work towards implementing these analytics on SocialLearn, an online learning space in use at the UK?s open university , and the challenges that this is raising. This work

Figure 4. The search results page

The paper is displayed in a clear/crisp format for easy online reading. The display is split into three sections. First, the user sees the topics that are most strongly associated with that paper. Then, the abstract is presented followed by the body of the paper. Another key usability point is that the topics are color-coded at the top and the coloring remains intact throughout the body of the paper. This helps draw the reader's attention to topics that may be of particular interest.

2.6 Similar Papers

Perhaps the strongest feature of LAK Explorer is the ability to find similar papers. Keyword and topic searching limits similarity to only papers that share the same frequency of that one term. The similar papers feature uses the entirety of the paper to compare it to other content.



The teaching demands and complexities evolving from recent government and university reforms, linked to quality assurance and accountability, emphasise the importance of developing replicable, scalable design and evaluative approaches. In this context learning design and learning analytics have emerged as two fields of research that can begin to address ongoing educational issues related to professional development, student and teacher evaluations and the development of personalized learning...

This workshop anticipates that an important class of learning analytic will emerge at the intersection of research into learning dynamics, online discussion platforms, and computational linguistics. Written discourse is arguably the primary class of data that can give us insights into deeper learning and higher order qualities such as critical thinking, argumentation, mastery of complex ideas, empathy, collaboration and interpersonal skills. Moreover, the ability to write in a scholarly manne...

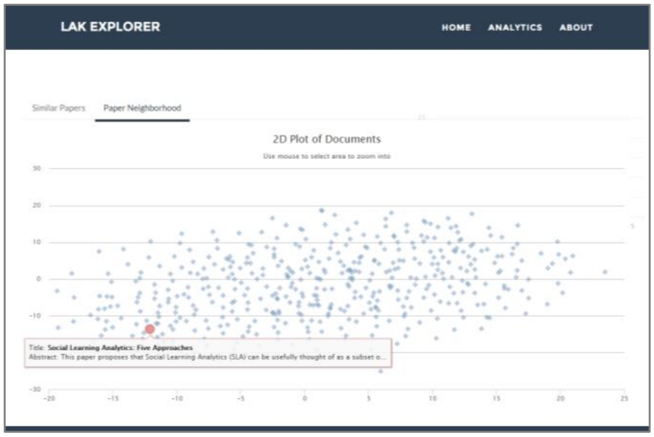
We propose that the design and implementation of effective Social Learning Analytics (SLA) present significant challenges and opportunities for both research and enterprise, in three important respects. The first is that the learning landscape is extraordinarily turbulent at present, in no small part due to technological drivers. Online social learning is emerging as a significant phenomenon for a variety of reasons, which we review, in order to motivate the concept of social learning. The se...

Theoretical and empirical evidence in the learning sciences substantiates the view that deep engagement in learning is a function of a complex combination of learners? identities, dispositions, values, attitudes and skills. When these are fragile, learners struggle to achieve their potential in conventional assessments, and critically, are not prepared for the novelty and complexity of the challenges they will meet in the workplace, and the many other spheres of life which require personal qu...

Figure 5. Browsing similar papers

Blue Canary used a technique called Doc2Vec [3] to generate similarity scores between two papers. This approach condeses the paper into a single vector, and then a cosine similarity measure is used to compare the vector of one paper to all others in the corpus (http://en.wikipedia.org/wiki/Cosine_similarity). The results (Figure 5.) give a match or score percentage showing papers that are most similar to the one currently being read.

Additionally, LAK Explorer displays this same vector similarity in a neighborhood scatter plot (Figure 6.).



Title: **Social Learning Analytics: Five Approaches**
Abstract: This paper proposes that Social Learning Analytics (SLA) can be usefully thought of as a subset o...

Figure 6. Visually representing similar papers

The Paper Neighborhood graph takes the Doc2Vec vector data and uses t-SNE (t-Distributed Stochastic Neighbor Embedding) to give all papers two dimensional Cartesian coordinates [4]. The resulting graph shows the current paper (orange dot) in relation to all other papers (blue dots). While the x- and y-coordinates have no real meaning or definition, the spatial representation of each paper allows the viewer to both browse similar papers and to see how "close" or "far" the current paper is from the rest of the corpus.

3. BENEFITS OF LAK EXPLORER

The Blue Canary team wanted to create an application that LAK researchers and practitioners would find valuable. To that end, the team focused on a few key aspects of LAK Explorer to maximize its contribution to the field.

3.1 Visual usability

The Blue Canary product development team gives significant weight to usability. The team might develop an incredibly useful metric, but if the user can't easily interpret that metric, it is useless. A clean layout, color coding, and simple charts all contribute to the usability of LAK Explorer. These features were consciously added in order to improve ease of use.

3.2 Leveraging Turbo Topics

As the team discovered in our LAK14 submission, extracting topics from the corpus turned out to be a simple yet effective way of absorbing the content of papers. We continued this trend for LAK15 by using the Turbo Topics to aid in the initial search and in the meta-tagging of the search results.

3.3 Search Results Plus Similarities

The LAK Explorer was named due to the fact that users will likely not be looking for a singular result. They will look for the results of their search PLUS explore other papers that are similar to their top search result.

The paper similarity tools allow LAK Explorer users to follow this natural path of exploration:

1. I'm interested in papers related to Topic X
2. Searching for Topic X gives me an ordered list of papers
3. I read through Paper 1 that comes up in search results
4. I am also shown Papers 2, 3, etc. that are like Paper 1

4. FUTURE FEATURES

After integrating the previously described components into the LAK Explorer, the team realized that there were additional features we could add to the tool to further increase its value.

4.1 Tracking User Input

The most impactful feature to add is to track user input to the application. LAK Explorer is already instrumented to capture the terms that users search and also the papers that are viewed. The additional feature would be to expose these tracking metrics to the application's front end so that other users can see how the community is interacting with the tool. For example, a simple listing of "most viewed papers' would add more fidelity to other LAK Explorer visitors.

4.2 Linked Data

The linked aspect of the corpus could be further exploited to improve the search process of LAK Explorer. The linked data discerns between abstracts, body, authors, citations, people, and institutions. These parts can be exposed as metadata in the search results (e.g. view more papers by this author) and the data can also be used to drive search efficacy (e.g. weigh hits to the abstract higher than hits to the paper body).

5. SIMILAR INITIATIVES

This is the third year of the LAK Data Challenge and all of the participants continue to stand on the shoulders of previous contributors. Blue Canary acknowledges that previous entrants such as the ones listed in this section have developed toolsets in the same vein as LAK Explorer

5.1 RekLAK

RecLAK [5] was submitted by a team of researchers from PUC Rio in Brazil for LAK14. RecLAK is a recommendation engine that uses the linked nature of the data to recommend other data sources that have feature similarities to the LAK dataset.

5.2 DEKDIV

DEKDIV [6] is an interactive application that allows the user to drill into different aspects of the LAK dataset. The 'Publications' section of DEKDIV was developed in the same spirit as LAK Explorer – allow the user to look at a paper and understand some of the key concepts.

5.3 Visualizing the LAK/EDM Literature

A team of famous LAK researchers submitted a paper to the first LAK Data Challenge in 2013 [7] where, among other things, they clustered the semantic content of the LAK corpus. While using a different technique, this clustering process achieved the same goal of the LAK Explorer's similarity feature set.

6. ACKNOWLEDGMENTS

As with most initiatives at Blue Canary, this was a product of teamwork. LAK Explorer was made possible by a team of players who each brought additive skills to the table. In addition to the named authors, we'd like to thank Satya Mudiam, Faiz Mohammad, Avinash Narasingam, and Kiran Reddy for their contributions.

7. REFERENCES

[1] Sharkey, M., & Ansari, M. (2014). Deconstruct and Reconstruct: Using Topic Modeling on an Analytics Corpus. In LAK Workshops.

[2] Blei, D. M., & Lafferty, J. D. (2009). Visualizing topics with multi-word expressions. arXiv preprint arXiv:0907.1013.

[3] Le, Q. V., & Mikolov, T. (2014). Distributed representations of sentences and documents. arXiv preprint arXiv:1405.4053.

[4] Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. Journal of Machine Learning Research, 9(2579-2605), 85.

[5] Lopes, G. R., Leme, L. A. P. P., Nunes, B. P., & Casanova, M. A. RecLAK: Analysis and Recommendation of Interlinking Datasets.

[6] Hu, Y., McKenzie, G., Yang, J. A., Gao, S., Abdalla, A., & Janowicz, K. (2014). A Linked-Data-Driven Web Portal for Learning Analytics: Data Enrichment, Interactive Visualization, and Knowledge Discovery. In LAK Workshops.

[7] Taibi, D., Sándor, Á., Simsek, D., Buckingham Shum, S., De Liddo, A., & Ferguson, R. (2013). Visualizing the LAK/EDM literature using combined concept and rhetorical sentence extraction.

Discovering Learning Antecedents in Learning Analytics Literature

Vladimer Kobayashi
CJKR, HRM-OB, ABS
University of Amsterdam
Netherlands
V.Kobayashi@uva.nl

Stefan Mol
CJKR, HRM-OB, ABS
University of Amsterdam
Netherlands
S.T.Mol@uva.nl

Gábor Kismihók
CJKR, HRM-OB, ABS
University of Amsterdam
Netherlands
G.Kismihok@uva.nl

ABSTRACT

We investigated various learning antecedents that have been the research subjects of Learning Analytics (LA) studies and explored the content and quantity of the LA literature with respect to each antecedent through text mining the LAK dataset. Our goal was to simultaneously reveal to what extent do LA researchers address learning antecedents and how they incorporated these in the implementation of LA solutions (e.g. models and software technologies) to facilitate and augment student learning. Instead of taking a pure text mining approach, we undertook a slightly different strategy by (i) identifying antecedents of student learning by examining extant literature on learning and educational theories and (ii) identifying which among the theoretically relevant antecedents are currently reported in LA studies. The analytical techniques we employed were a mix of domain-based analysis and corpus analytics which included association analysis and key-phrase extraction. The results showed that most LA studies are geared toward capturing and measuring student awareness and promoting social learning and less on goal-setting and self-efficacy. Through this work we hope to encourage the LA community to dedicate research efforts to also investigate other relatively neglected yet promising learning antecedents.

Keywords

student learning, corpus analytics, learning analytics

1. MOTIVATION AND OBJECTIVE

The Learning Analytics (LA) field uses analytics to understand and facilitate student learning. Since learning is influenced by various antecedents and circumstances, some LA researchers focus on capturing, measuring, and enhancing these antecedents in an effort to impact student learning. This is especially relevant nowadays with the proliferation of nontraditional venues for learning such as in online learning. Examples of these antecedents include awareness, social learning, and self-regulated learning to name but a few.

As LA studies flourish a need arises to address the question of how LA as a field has contributed so far to our understanding and to the enhancement of student learning. This can be answered in part by characterizing LA studies according to which learning antecedents they tackle. This could help researchers from various education-related disciplines to keep track, compare, and share knowledge and to identify opportunities for further research. It could also provide a basis for the adaption of LA projects and explicating how LA models and software technologies influence learning. How each element of an LA project imparts information or generates and uses data that valuate the determinants for student learning success is a major concern.

Our primary objective was to explore the content and quantity of LA literature that report each learning antecedent. In a parallel manner, we shifted the focus towards the antecedents by finding which antecedents are often addressed and which not. This approach would facilitate a more objective assessment and comparison of whether LA studies have achieved their intended outcomes.

For this study we used the dataset provided by the LAK dataset challenge [9] and other literature on student learning theories to accomplish our objective.

2. METHODOLOGY

As an overview, we used a text mining approach to discover learning antecedents. Although text mining is naturally an inductive approach we supplemented our investigation with domain information. The diagrammatic description of the steps we undertook is illustrated in Figure 1.

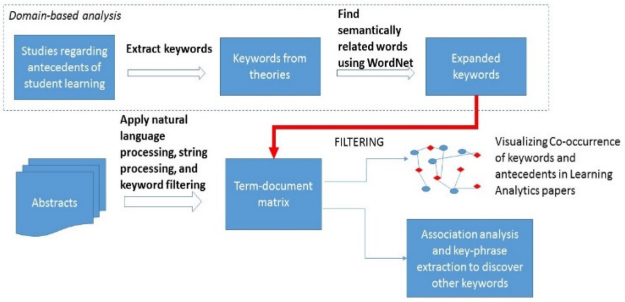

Domain-based analysis
Studies regarding antecedents of student learning
Extract keywords
Keywords from theories
Find semantically related words using WordNet
Expanded keywords
Apply natural language processing, string processing, and keyword filtering
FILTERING
Abstracts
Term-document matrix
Visualizing Co-occurrence of keywords and antecedents in Learning Analytics papers
Association analysis and key-phrase extraction to discover other keywords

Figure 1: Diagrammatic view of the methodological steps followed in this study.

2.1 Domain-based Analysis

We first performed an inquiry regarding the antecedents that influence student learning and learning outcomes. From this inquiry we identified keywords that are usually strongly associated to each antecedent. The keywords represent the vocabulary used to refer to the antecedents that were extracted from existing literature on education and student learning theories. The antecedents are discussed in Section 3.

The list of keywords were further expanded by using a lexical database called WordNet[1] to find semantically similar words. This is a vital step because authors use varying terms to convey the same

---

[1] http://wordnet.princeton.edu/

Published in CEUR-WS:
Proceedings of the VISLA15 Workshop on Visual Aspects of Learning Analytics (Duval et al., eds), organized at LAK15.

concept. An example would be to use "participate" rather than "engage". The expanded keyword list was used in the succeeding steps.

2.2 Corpus Analytics on LAK dataset

Corpus analytics was performed in the following manner.

First, we initially kept matters simple yet meaningful by choosing to perform corpus analytics only on the abstracts of each publication. There might be a downside to this such as missing otherwise important information but in exchange this has kept the analysis manageable. Moreover, this decision is sufficient for our purpose since the abstract contains the gist of the whole article and provides a summary about the paper's objectives, methodology, and conclusion.

Second, we created a corpus containing abstracts of all papers in the LAK dataset. Each document was pre-processed by removing punctuation, removing numbers, transforming upper case letters to lower case, removing stopwords, and selectively stemming specific words. An example of the selective stemming was to treat the words "engaging" and "engagement" as just derivatives of the word "engage". The method of stemming that we applied here is the look-up table method where the look-up table is the expanded keyword list from domain analysis.

Third, a further filtering was implemented to reduce the number of terms. The filtering process was done using the expanded keyword list in conjunction with association analysis so that potentially important words not present in the list could be identified and added.

Fourth and finally, the pre-processing stage culminated in the creation of the document-by-term matrix weighted by raw term frequencies. We were interested in determining which among the theory inspired antecedents (see Section 3) are discussed in each LA study. The document-by-term matrix acted as a springboard from which we explored the construction of other matrices (e.g. co-occurrence matrices) and application of other analytical techniques such as key-phrase extraction.

All analyses were done using the **R** software[2] and the packages **tm**[3], **wordnet**[4], and **igraph**[5].

2.3 Two assumptions

We assumed that the mention of keywords associated to a learning antecedent in the abstract of a paper would indicate that the paper is dealing with that learning antecedent. We anticipate a number of caveats with this assumption. One possible scenario is that the keyword is used in a different sense. An example is the keyword "goal", in some papers the presence of this word does not mean that they are automatically dealing with Goal-setting but it could be the case that the word "goal" here refers to the goal of the study. Thus it is also important to consider the context in which the word is being used. We addressed this by examining other words in the abstract. Using association analysis we noticed that when the word "goal" is used in the sense of Goal-setting words such as performance, achievement, or learning are also encountered.

Another assumption is that the mention of keywords belonging to different learning antecedent in one abstract means that these two learning antecedents are simultaneously addressed and with the same emphasis in that paper. We can see a problem with this since some papers just use the concept but do not develop that concept further. This problem can be addressed by using the information on the raw frequencies of the term. The higher the raw frequency the more importance we can attach to it with respect to a particular paper.

3. NINE ANTECEDENTS OF STUDENT LEARNING

The keywords represent 9 common antecedents that have been reported by educational experts as antecedents for success in learning. The antecedents are: (1) Engagement, (2) Motivation, (3) Self-reflection (including self-assessment and self-regulation), (4) Social Learning (among students and between students and teachers), (5) Assessment (e.g. formatting testing and evaluation), (6) Recommendation (and feedback), (7) Goal-setting, (8) Awareness (social awareness, context awareness), and (9) Self-confidence. These were selected based on our previous content analysis of publications in the area of education and student learning.

Student engagement refers to the quality of effort and level of involvement that students invest in their learning. It has been shown to be positively linked to gains in general abilities, critical thinking, and grades [1]. Therefore it has worthwhile effects on student learning and success in education.

Motivation is a drive, a stimuli, an incentive or desire that causes someone to act or to expend effort to accomplish something [8]. Often, it is manifested when students are attentive, participative and active in class.

Self-reflection occurs when learners evaluate the breadth and scope of their knowledge. It is important in learning because it helps students to identify what they need to learn leading to effective self-regulation [5].

Some researchers view learning as a collaborative process where learners interact and share knowledge. The roles, activities, and behavior that students assume in a social learning context ultimately impact their learning [2].

Testing and assessment in general has long been used to assess whether students have achieved specific learning outcomes. Furthermore, during testing information is stored in the brain for long term retrieval, which in turn is essential for learning transfer (i.e. using information in different contexts) and meaning generation.

Recommendation is seen as a potential antecedent of learning since it helps students track their learning achievement and improve their learning at the same time [3].

Goals direct attention, energize effort and promote persistence. Studies have shown the valuable effect of goal-setting to academic achievement, self-regulation, and deep learning strategies [6].

Awareness provides context for learning since it discloses information about other person's activities and the environment where learning takes place. It has been shown to be crucial to learning and contributes to the quality of active participation [7].

Last is self-efficacy (colloquially termed as self-confidence) which is usually defined as belief in one's own capability to accomplish tasks and achieve goals [4]. It is important in learning since students

---

[2] http://www.r-project.org/

[3] http://cran.r-project.org/web/packages/tm/index.html

[4] http://cran.r-project.org/web/packages/wordnet/index.html

[5] http://cran.r-project.org/web/packages/igraph/index.html

must believe in their own capacity to learn even if the material is difficult.

We added the Analytics to see which LA projects have incorporated advanced analytical tools on top of the basic summarization and visualization features.

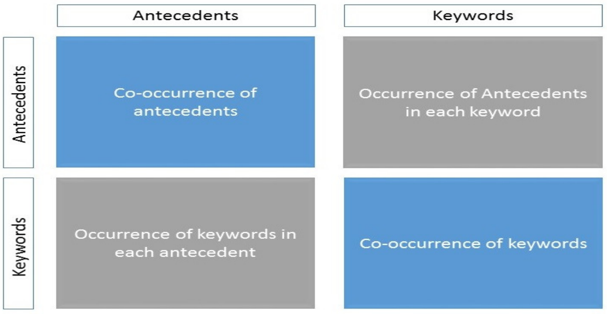
4. MAIN FINDINGs AND DISCUSSION

Combining the keywords obtained from the domain analysis, association analysis, and corpus analytics we obtained the keyword list in Table 1 that are grouped according to the antecedents that are most likely associated to them.

Table 1: Keywords associated to each learning antecedent.

Learning Antecedents	**Keywords**
Engagement	engage, participate, active, access, resource
Motivation	motivate, encourage
Self-reflection	negotiate, self-regulate, self-reflect, self-aware, self-discipline, self-test, reflect, self-report, self-knowledfe
Social Learning	collaborate, network, interact, social, community, graph, connect
Assessment	test, assess
Recommendation	recommend, feedback, intervene
Goal-setting	goal, sub-goal
Awareness	aware, content-aware, track, monitor, compare
Self-Confidence	confidence, self-efficacy
Analytics	model, student model, user model, analytics, analytic, predict, valid, visual, classify

From the document-by-term matrix we identified which among the documents have used analytics and which learning antecedents are addressed in each document. We also constructed 4 co-occurrence matrices (see Figure 2) that reveal which learning antecedents are often treated simultaneously, and which keywords are often mentioned together. A sampling of output is presented in Figure 3.



	Antecedents	Keywords
Antecedents	Co-occurrence of antecedents	Occurrence of Antecedents in each keyword
Keywords	Occurrence of keywords in each antecedent	Co-occurrence of keywords

Figure 2: Four co-occurrence matrices constructed from the term-by-document matrix.

The first subfigure (Figure 3a) shows a bar plot that depicts the number of papers in the LAK dataset that have dealt with each learning antecedent. It can be vividly seen that the focus of many studies are the learning antecedents awareness, social learning, engagement, and assessment. This can be explained by the considerable interest of LA researchers in online learning settings where the capture, measurement, and monitoring of these antecedents are both challenging and crucial. On the other the less often discussed antecedents are goal-setting, motivation, and self-discipline. Although, goal-setting has a slightly higher bar than self-reflection this is because some studies that mention the word “goal” actually referred to the aim or objective of the studies.

Figure 3b depicts both the magnitude of studies that deal with each antecedent and the relationship (in the sense of co-occurrence) among the antecedents. The red circles are the antecedents and the green ones are the keywords. An edge connects a keyword to its associated antecedent and edges between antecedents represent relationship. We include “Analytics” to see which among the antecedents make heavy use of analytics and what type of analytics is commonly employed. It is not difficult to observe that social learning and awareness are the most related in terms of the number of publications that tackled them. It is followed by awareness and assessment, although there is a strong indication that assessment here may imply the students’ assessment of their knowledge, context, peers, and environment and not about test or evaluation.

The last subgraph (Figure 3c) visually represent the relationship among words as well as the quantity of studies that mention each word (as expressed by the size of the circle). It is not surprising to observe that the word “model” is the leading keyword this is because most LA researchers are concerned with creating models to describe some learning-related phenomena, as to be expected from an LA research. Another observation that is worth mentioning is the conspicuousness of the three vertices that represent visual, network, and interact and the interconnections between them. These three are indicative of the social learning antecedent since interactions among students are usually visualized by means of a network structure.

In Table 2, we see the list of words that are highly associated to the keywords of each antecedent. We discovered these with the use of association analysis and key-phrase extraction. The list is incomplete since we just present the ones that were interesting in our opinion. These words could be used to further enrich our original keyword list. Moreover, we unearthed interesting relationships such as the association between “affect” and “engagement”, “assessment” and “scores”, “recommendation” and “similarity”. Some of these associations reveal the kind of techniques used to analyze particular antecedents (e.g. the use of the idea of similarity in recommendation) and the underlying concepts that might govern an antecedent (e.g. the affective state of a student might indicate or influence engagement).

Table 2: Other terms associated to each antecedent.

Engagement	affect, peripheral, discussion, home
motivation	learnograms
self-reflection	cope, personal, health, feelings
Social learning	blackboard, intergroup, intranetwork, cyberlearner
Assessment	scores
Recommendation	similarity
Goal setting	orientation, temporal
Awareness	clues, cope
Self-confidence	Egocentric, high achieving

5. CONCLUSION AND FUTURE WORK

In this study we show how an analysis that combines domain-based information and corpus analytics could be used to uncover and analyse interesting concepts in LA literature. These concepts directly deal with the question of how LA has been used to improve our understanding and control of a number of learning antecedents. We believe that to fully answer that question a more detailed analysis should be undertaken such as investigating the measures and validity of the constructed models as described in the publications. Nevertheless, our approach clears the cloud to expedite such detailed analysis. Our study also highlights the need to study other antecedents that might be critical to student learning but do not yet receive due research attention. From an educator's perspective it is now becoming clearer how LA solutions impact learning and to which aspect the contribution is focused. It is now time that we move LA from a technique-laden endeavor to a more theory driven approach.

If ever, this work will be selected we also show our effort on the temporal analysis of these antecedents such as visualizing the evolution of focus of LA studies on each concept. Moreover, we aim to analyze how publications in educational data mining, learning analytics and technology-enhanced learning differ in this aspect.

6. ACKNOWLEDGEMENT

We gratefully acknowledge the publishers who have contributed to the LAK Dataset: ACM, International Educational Data Mining Society and Journal on Education Technology &Society. We are grateful for the financial support of the Eduworks Marie Curie Initial Training Network Project (PITN-GA-2013-608311) of the European Commissions's 7th Framework Programme.

7. REFERENCES

[1] Aguiar, E. et al. 2014. Engagement vs performance: using electronic portfolios to predict first semester engineering student retention. (2014), 103–112.

[2] Barr, J. and Gunawardena, A. 2012. Classroom Salon: A Tool for Social Collaboration. *Proceedings of the 43rd ACM Technical Symposium on Computer Science Education* (New York, NY, USA, 2012), 197–202.

[3] Bramucci, R. and Gaston, J. 2012. Sherpa: Increasing Student Success with a Recommendation Engine. *Proceedings of the 2Nd International Conference on Learning Analytics and Knowledge* (New York, NY, USA, 2012), 82–83.

[4] Diseth, Å. 2011. Self-efficacy, goal orientations and learning strategies as mediators between preceding and subsequent academic achievement. *Learning and Individual Differences*. 21, 2 (Apr. 2011), 191–195.

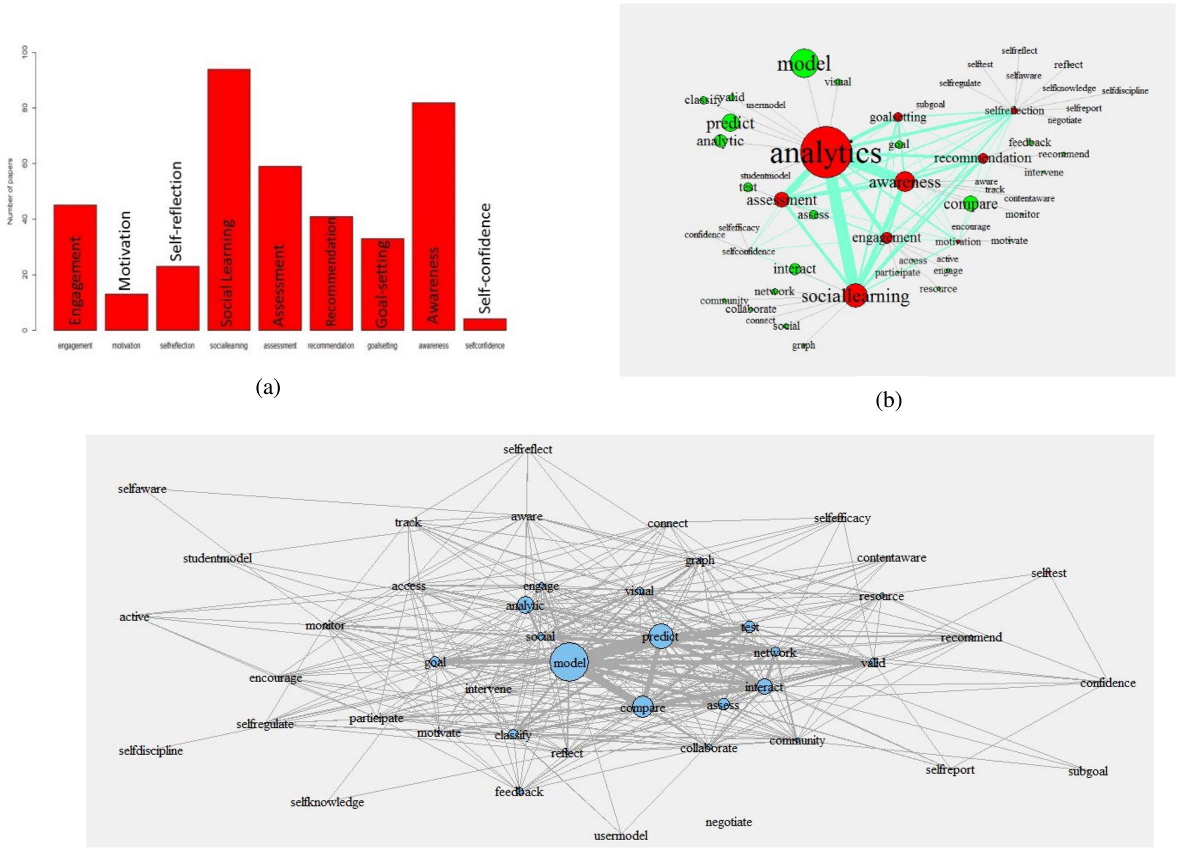
[5] Govaerts, S. et al. 2012. The Student Activity Meter for Awareness and Self-reflection. *CHI '12 Extended Abstracts on Human Factors in Computing Systems* (New York, NY, USA, 2012), 869–884.

[6] Latham, G.P. and Locke, E.A. 2007. New Developments in and Directions for Goal-Setting Research. *European Psychologist*. 12, 4 (Jan. 2007), 290–300.

[7] Pohl, A. et al. 2012. Sensing the classroom: Improving awareness and self-awareness of students in Backstage. *2012 15th International Conference on Interactive Collaborative Learning (ICL)* (Sep. 2012), 1–8.

[8] Schiefele, U. Interest, Learning, and Motivation.

[9] Taibi, D. and Dietze, S. 2013. Fostering analytics on learning analytics research: the LAK dataset. In: *CEUR WS Proceedings Vol 974, Proceedings of the LAK Data Challenge*, held at LAK2013 - 3rd International Conference on Learning Analytics and Knowledge(Leuven, BE, Apr. 2013).



(a)

(b)

(c)

Figure 3: Sampling of the output from the analysis