2011

Generating Predictive Models of Learner Community Dynamics

C. Teplovs

N. Fujita
University of Windsor, nfujita@uwindsor.ca

R. Vatrapu

Generating Predictive Models of Learner Community Dynamics

Chris Teplovs[1,2], Nobuko Fujita[1,2] and Ravi Vatrapu[2]
[1] University of Toronto [2]Copenhagen Business School
chris.teplovs@gmail.com, nobuko.fujita@utoronto.ca, vatrapu@cbs.dk

ABSTRACT
In this paper we present a framework for learner modelling that combines latent semantic analysis and social network analysis of online discourse. The framework is supported by newly developed software, known as the Knowledge, Interaction, and Social Student Modelling Explorer (KISSME), that employs highly interactive visualizations of content-aware interactions among learners. Our goal is to develop, use and refine KISSME to generate and test predictive models of learner interactions to optimise learning.

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

General Terms
Design, Theory, Analysis.

Keywords
Information visualization, latent semantic analysis, social network analysis, learner models, game theory.

1. INTRODUCTION
The nascent field of Learning Analytics focuses on "the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the environments in which it occurs"[1]. One approach to learning analytics is social network analysis, which examines the patterns of interaction among learners. Social network analysis of, in particular, e-learning is facilitated by the availability of digital data that are amenable to such analysis. Considerably less attention has been paid to the content of the artifacts around which the learners are interacting. Content analysis is time-consuming, pain-staking, and detailed work. Without content analysis, however, claims about the nature of the dynamics among learners are left wanting. Understanding learning, it seems, requires digging deeply into the data that are available.

In this paper we introduce a framework that interweaves social network analysis, semi-automated content analysis, information visualization, and applied economic theory to help us understand and optimise learning. We are interested in investigating research questions such as: Can we "predict" when particular interactions will result in learning? What are some characteristics of interactions of effective learning?

This paper begins with a brief introduction and survey of relevant literature using social network and latent semantic network analysis (LSA) to analyze online discourse. Next, a description of the prototypic software environment (the Knowledge Space Visualizer or KSV) on which the new software (the Knowledge, Interaction and Semantic Student Model Explorer, or KISSME) is being developed is presented. The use of LSA in the generation of student models suitable for studies of collaborative learning is then proposed. Finally, we present a theoretical framework for understanding the dynamics of collaborative learning in terms of examining the outcomes of social and semantic interactions among participants.

LAK'11, February 27-March 1, 2011, Banff, AB, Canada.

2. BACKGROUND
Wasserman and Faust [1] describe social network analysis (SNA) as a methodology that focuses on relationships and patterns of relationships. As such it "requires a set of methods and analytic concepts that are distinct from the methods of traditional statistics and data analysis" (p. 3). They cast SNA in the broader list of topics that have been studied using network analytic methods, including community [2], group problem solving [3-5], diffusion and adoption of innovations [6-8], and cognition [9, 10]. No matter what the objective of the study, though, network analysis focuses on the relations between units.

Studies have explored the application of SNA to explore learning and knowledge construction in Networked Learning/Computer-Supported Collaborative Learning (NL/CSCL) environments. However, researchers have yet to achieve consensus on what methods to use. For example, de Laat, Lally, and Lipponen, [11] used content analysis, critical event recall and SNA to study interaction patterns. They suggest that SNA can be used to complement content analysis [12, 13] to describe and understand patterns of interaction in NL/CSCL. Of the various network metrics that are available (see [1]), these researchers focus on density and centrality. In contrast, Reffay and Chanier [14] applied SNA to determine the cohesion of groups engaged in CSCL. They argue that embedding tools that perform such analyses in the design of the learning environment itself may be more effective than time-consuming content analysis to support teaching and learning.

The importance of time-based analyses has also been noted [15][16] The study by de Laat et al [11] was the first application of using SNA to illustrate how patterns change over time and the relationship of those patterns to teaching and learning. An important generalization from the literature is that the essential

[1] https://tekri.athabascau.ca/analytics/call-papers

features to conduct SNA are two or more units, usually learners and the elucidation of the relationship between them. But there is another equally important type of network analysis to be considered in learning analytics and knowledge work: the network of ideas. Ideas, unfortunately, are difficult to delineate.

2.1 Latent Semantic Analysis

Latent semantic analysis (LSA) represents both a statistical technique and a model of human knowledge acquisition. Landauer and Dumais [17] propose LSA as a model that could answer the question, how do individuals know so much given as little information as they get? This problem is variously known as Plato's Problem, the "Problem of Induction", the "poverty of the stimulus", or "the problem of the expert". (Plato's solution was that individuals possess innate knowledge and only need some stimulation to reveal it.)

LSA provides a high-dimensional representation of the associations between words and the documents containing those words. The final output from LSA is a series of measures that describe the relationships between units such as words, documents, or words-and-documents. In LSA, each document or word is represented by a vector in high-dimensional latent semantic space. The vector is calculated by examining patterns of co-occurrence of words in a term-by-document matrix, which is subsequently simplified using Singular Value Decomposition (SVD). Thus, each document is represented by a vector of numbers, typically numbering between 100 and 300 elements. Whereas dimensions resulting from the application of SVD to data can typically be interpreted (e.g. the dimensions from Principal Components Analysis), the dimensions resulting from LSA are not typically interpretable. This limitation has made the interpretability of LSA-based analyses difficult in the past.

Information visualization techniques seem to be a natural next step in interpreting LSA, and can be used to create meaningful representations of ongoing learning processes. Visualization of LSA-derived similarities may be problematic, though, due to an unacceptable reduction of dimensionality to two or three dimensions suitable for visualization from that which is optimal for LSA (typically around 300) [18].

3. SOFTWARE

In this section we describe software designed to support the visualization of learner models based on social and semantic networks. We present a description of the Knowledge Space Visualizer (KSV), a prototypic software system on which our new software, KISSME, is based.

3.1 The Knowledge Space Visualizer (KSV)

KISSME extends the Knowledge Space Visualizer, which was developed by the first author for his doctoral dissertation. The KSV was designed to allow researchers to use computer-assisted two-dimensional visualization of learner-generated contributions to an online discourse space. In its simplest form this generates a graph in which nodes are contributions and links are relationships between those contributions such as "reply", "reference" and "annotate" (see Figure 1).

These explicit relationships between contributions are based on the behaviours of the contributors. A learner, for example, can intentionally choose to make a contribution that is a reply to another learner's contribution. In the resulting graph the links are based on these behavioural relationships. Content is not considered.

In addition to the explicit linkages defined by behaviours such as replying, referencing and annotating there exist implicit linkages between contributions to the discourse space. These implicit linkages concentrate on the similarity of the content of the contributions. Whereas human raters can evaluate the similarity between documents reliably and with good validity, it is very tedious and time-consuming work. There are a variety of automated and semi-automated techniques that can be used to determine the similarity of text-based contributions. One powerful technique is LSA, described above.

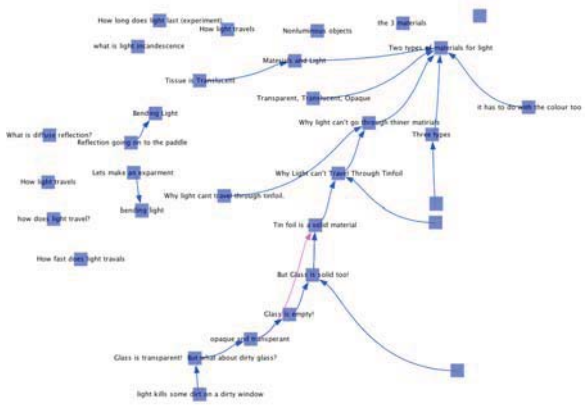

Figure 1. Structural relationships between contributions. Blue lines indicate "build-on" or "reply-to" relationships. Magenta lines indicate "reference" links

The preceding examples are based on the use of a force-directed layout algorithm to position the nodes in to respect the strength of the ties between them while minimizing the distortion of the network of the relationships between the nodes. Other types of layouts are also possible. For example, other researchers [19] have highlighted the importance of chronology when studying the dynamics of learning communities. The KSV supports this sort of inquiry by facilitating the positioning of notes chronologically. More generally, the KSV supports the use of any categorical, ordinal, or continuous variable from the data set to define either of the axes for the display. So in addition to the use of a continuous chronological scale to define the horizontal axis, authorship can be used to define the vertical axis. An example of the resulting learner-time display is shown in Figure 2.

Once contributions are positioned on whatever set of operationally defined axes the analyst has chosen, links between nodes can be overlaid without affecting the positioning of the nodes. For example, the behavioural links can be overlaid on the learner-time display to show how patterns of interaction change over time. An example of this overlay is shown in Figure 3.

In a similar way, links between contributions based on latent semantic analysis can be overlaid on the same learner-time display to show the degree to which contributions are similar over time and authorship. More computationally intensive measures

can also be visualized. For example, one can determine which contributions were opened (and possibly read) by a learner within some specified time interval before that contributor added a new contribution to the discourse space. An example of this sort of "recency influence" diagram is shown in Figure 4.

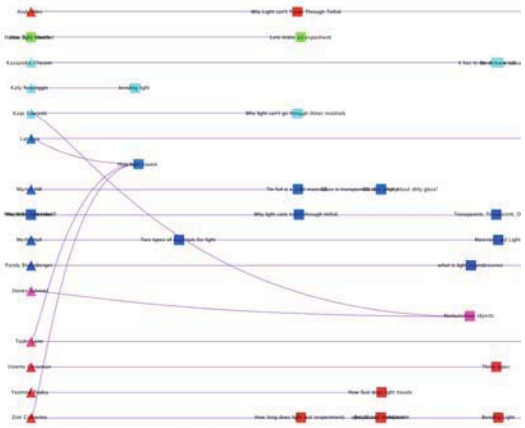

Figure 2. Chronological-authorial layout of contributions

Perhaps some of the most interesting diagrams that can be produced using the KSV are based on the superposition of different link types on the same layout. For example, one can overlay links of LSA-based semantic similarity atop those based on "recency influence" to investigate the degree to which the content of recently opened (read) contributions is reflected in new contributions.

The KSV also allows the user to constrain the analysis by specifying beginning and end dates for the analysis. Rather than specifying the dates a priori, the user can manipulate the beginning and end dates with specially designed slider. In addition to being able to manipulate the beginning and end dates independently of one another, the user can manipulate both dates simultaneously, effectively providing time slices of the network graph.

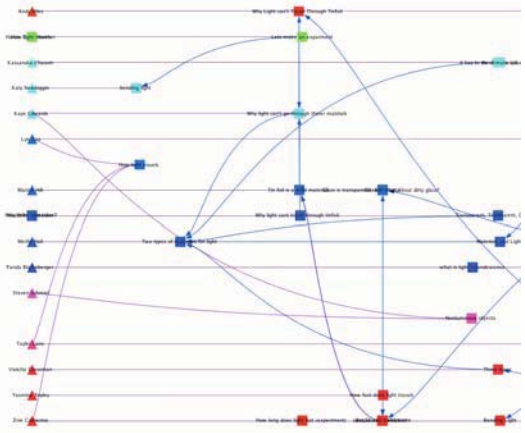

Figure 3. Chronological-authorial layout of contributions overlaid with structural links

One of the key innovations of the KSV was the use of flexible thresholds in the creation of network representations. This is what allowed us to create visualizations of LSA-based representations of texts. Rather than attempting to provide a two-dimensional layout based on the first few dimensions resulting from the matrix decomposition used in LSA, our approach has been to determine the similarities between documents based on the cosines between the vectors representing documents. A graph is then created in which the nodes correspond to the documents and the edges correspond to the LSA-based similarities between them. A force-directed layout algorithm is then applied to the graph such that the positions of nodes in the two-dimensional representation minimize the distortion of the (very low dimensional) representation. This representation of a maximally connected graph typically lacks clarity, and in typical cases where there are tens or hundreds of nodes the graph is essentially unintelligible due to the large number of edges.

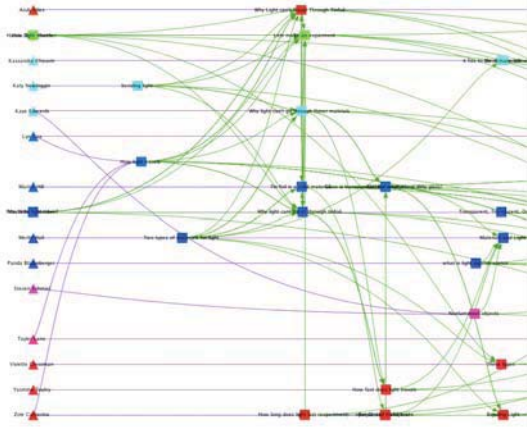

Figure 4. Chronological-authorial layout with overlaid with structural and recency links

This problem of overly connected graphs also presents a conceptual problem: does it make sense to connect two document nodes if their LSA-based similarity is very low? Other researchers [20] have attempted to address the "threshold problem" but heir research suggests that no typical value of cosine threshold for determining document similarity exists. Our approach to tackle this problem is to provide the end user with control over the choice of threshold to use. We do so by providing a slider control in the software that allows the user to specify the cosine value below which edges are not drawn between document nodes. The dynamic nature of this control allows the user, for example, to examine patterns of cluster formation as the similarity threshold is varied.

This provides an example of how visual approaches to learning analytics can provide solutions to previously intractable problems. The answer to the question of "when are two documents (or ideas) different" is typically "it depends on what you're looking for". Given a collection of documents generated by students on, for example, the physics of light. At the most permissive level of similarity threshold, all documents are related by virtue of being in the same language. This corresponds to a similarity threshold of zero. At a value slightly higher than zero, one could imagine the documents cluster into two groups: one about colours of light and one about reflection. As one raised the threshold higher yet one could imagine the colours cluster fragmenting into smaller clusters of related notes about topics such as rainbows, wavelength, and so on. The interactive nature of being able to manipulate the threshold supports this broad range of possibilities

for determining the diversity of ideas that are present in discourse space.

The Knowledge Space Visualizer, while providing powerful visualizations of multi-dimensional networks, has several limitations. First, it relies on the end user having an functional installation of a recent version of Java. Recent advances in browser-based technology -- specifically the widespread adoption of HTML5 -- has enabled the production of highly interactive browser-based visualizations. Perhaps more significantly, the KSV was limited by its focus on document-based networks. The KSV enables the visualization of relationships between documents, based on both explicit and implicit linkages, but other than examining patterns of authorship and co-authorship it was not particularly good at generating visualizations of author-based networks. We are working on creating next-generation software that will facilitate the examination of networks of authors. In its earliest versions, the KSV was highly tuned to data from Knowledge Forum. The KSV was recently enhanced to allow the importation of data from almost any data source that provided indications of authorship, chronology and content. The KSV was released as open source code and is maintained on Google Code at http://code.google.com/p/ksv.

3.2 Visualizing Student Models: The Knowledge, Interaction and Semantic Student Model Explorer (KISSME)

Recent work has led to the implementation of a learner model based on interactions with other learners. The functionality of the KSV, in terms of being able to manipulate the threshold at which two nodes are considered similar enough to be joined by visible edges, was extended from document nodes to learner nodes. Put another way, a learner model based on social network analysis was created in the KSV and the implementation of a flexible threshold (based on the intensity of the interaction between any two learners) allowed researchers to investigate patterns of interaction. The KSV allowed the analyst to exercise considerable control over various parameters such as the intensity of interaction necessary to establish a social link between participants, as well as the date at which the social network was analysed. The ability of the analyst to vary these parameters allowed the detection of patterns of interaction that were previously obscured [21]. However, the network between authors was based solely on their patterns of interaction. No information about the content of their contributions was used in the generation of the graphs.

The ability to model students or other participants and then to visualize those models in an interactive visualization environment offers the potential to gain insights into the nature and outcomes of interactions between learners. In the work with the STEF lab we constrained our analyses to focus on the social networks that formed among learners. While this approach revealed interesting patterns of interaction, we felt the results were incomplete because no attention was paid to the content of the learners' contributions to the online discourse space.

Other researchers have conducted studies that meld automated interaction analysis with manual content analysis [11, 16]. However, manual content analysis represents the rate-limiting step in this sort of analysis. Because manual content analysis takes so long it is incommensurable with real-time analysis, which is one of our goals. Therefore, we are interested in using some sort of automated or semi-automated content analysis. For reasons specified earlier we have chosen to use latent semantic analysis to help us conduct automated content analysis. For our purposes, all that we are using LSA for is to generate mathematical representations of the participants' contributions to the discourse space. We can then use those mathematical representations in a variety of ways. LSA uses a vector representation of text. One characteristic of these vectors is that they are additive: the vectors of two documents can be added together to get the vector of the combined documents. We can extend this property to generate latent semantic models of participants by adding together the vector representations of all their contributions to the discourse space.

This is not the first application of LSA to student modelling. Other researchers [22-24] have used LSA in student modelling but they have not focused on the collaborative nature of learning. Still others have extended techniques from earlier research on LSA to apply to e-learning contexts [25-27]. Zampa and Lamaire's recent work [23] builds on the notion of matching students to text based on the Vygotsky's Zone of Proximal Development. However, theirs is an individualistic model: the selection of "stimuli" is meant to effect individualized optimization of learning.

Our approach is somewhat different: we are interested in combining information about patterns of interaction among participants with information about the content of those contributions. We too take a Vygotskian approach: that optimal learning will take place when interactions occur between individuals who are neither too similar nor too dissimilar from each other, based on the semantics of what they have written. This approach of combining social network analysis and latent semantic network analysis is an example of the sort of "multi-dimensional" network championed by Noshir Contractor [28].

Our current work includes the implementation of software that will allow us as researchers to examine the interplay of interactions between learners and the latent semantic models of those learners. We are interested in testing the Vygotskian hypothesis that uptake [29] is most likely to occur when the semantic relatedness of the corresponding contributor models is neither too high nor too low. We are also interested in simulations of learner interactions that take into consideration both interactions and semantic relatedness. This, we believe, would allow us to generate models of community dynamics in collaborative learning. Once we have simulation data that incorporates interaction and content we can make inferences about the characteristics result in the success (broadly defined) of some learning communities.

4. GAME THEORETICAL APPROACHES TO UNDERSTANDING THE LEARNER'S GROUP DYNAMICS

Our approach to understanding community dynamics is based on understanding the nature of the interaction between members of that community. We are examining a variety of theoretical approaches but one that seems particularly promising is the application of game theory [30] to interactions between users. This approach requires us to consider the outcomes of interactions between users in terms of "payoffs" to each player. Of course, different players can employ different strategies. We consider

this to be part and parcel of learning: our hypothesis is that as learners gain expertise, they enhance their repertoire of learning strategies, and through experience they learn when to employ particular strategies.

5. SUMMARY

We have proposed a framework that combines social network analysis and latent semantic analysis of online discourse. The proposal is speculative: previous work with latent semantic analysis has yielded promising results that may help us understand the nature of interactions among learners. Examining those interactions using a framework such as game theory may allow us to gain insight into the nature of community dynamics.

6. REFERENCES

[1] Wasserman, S., Faust, K.: Social network analysis: Methods and applications. Cambridge University Press, Cambridge, UK (1997)

[2] Wellman, B.: The community question: The intimate networks of East Yorkers. American Journal of Sociology 84, 1201-1231 (1979)

[3] Bavelas, A.: Communication patterns in task-oriented groups. Journal of the Acoustical Society of America 22, 271-282 (1950)

[4] Bavelas, A., Barrett, D.: An experimental approach to organizational communication. Personnel 27, 366-371 (1951)

[5] Leavitt, H.J.: Some effects of communication patterns on group performance. Journal of Abnormal and Social Psychology 46, 38-50 (1951)

[6] Coleman, J.S., Katz, E., Menzel, H.: The diffusion of an innovation among physicians. Sociometry 20, 253-270 (1957)

[7] Coleman, J.S., Katz, E., Menzel, H.: Medical Innovation: A diffusion study. Bobbs-Merrill, Indianapolis (1966)

[8] Rogers, E.M.: Network analysis of the diffusion of innovations. In: Holland, P.W., Leinhardt, S. (eds.) Perspectives on Social Network Research, pp. 137-164. Academic Press, New York, NY (1979)

[9] Freeman, L.C., Romney, A.K., Freeman, S.C.: Cognitive structure and informant accuracy. American Anthropologist 89, 310-325 (1987)

[10] Krackhardt, D.: Cognitive social structures. Social Networks 9, 109-134 (1987)

[11] de Laat, M., Lally, V., Lipponen, L., Simons, R.-J.: Investigating patterns of interaction in networked learning and computer-supported collaborative learning: A role for Social Network Analysis. International Journal of Computer-Supported Collaborative Learning 2, 87-103 (2007)

[12] Henri, F.: Computer conferencing and content analysis. In: Kaye, A.R. (ed.) Collaborative learning through computer conferencing. Springer, London (1992)

[13] Hara, N., Bonk, C.J., Angeli, C.: Content analyses of on-line discussion in an applied educational psychology course. Instructional Science 28, 115-152 (2000)

[14] Reffay, C., Chanier, T.: How social network analysis can help to measure cohesion in collaborative distance-learning. In: Designing for change in networked learning. Proceedings of the international conference on Computer Supported Collaborative Learning 2003., pp. 343-352. Kluwer Academic Publishers, (Year)

[15] Haythornthwaite, C.: Exploring multiplexity: social network structure in a computer-supported distance learning class. The Information Society 17, 211-226 (2001)

[16] Martínez, A., Dimitriadis, Y., Rubia, B., Gomez, E., de la Fuente, P.: Combining qualitative evaluation and social network analysis for the study of classroom social interactions. Computers & Education 41, 353-368 (2003)

[17] Landauer, T.K., Dumais, S.T.: A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. Psychological Review 104, 211-240 (1997)

[18] Landauer, T.K., Laham, D., Derr, M.: From paragraph to graph: Latent semantic analysis for information visualization. PNAS 101, 5214-5219 (2004)

[19] Reimann, P.: Time is precious: Variable- and event-centred approaches to process analysis in CSCL research. International Journal of Computer-Supported Collaborative Learning 4, 239-257 (2009)

[20] Penumatsa, P., Ventura, M., Graesser, A.C., Louwerse, M.M., Hu, X., Cai, Z., Franceschetti, D.R.: The right threshold value: What is the right threshold of cosine measure when using latent semantic analysis for evaluating student answers? International Journal on Artificial Intelligence Tools 15, 767-778 (2006)

[21] Reffay, C., Teplovs, C., Blondel, F.-M.: Productive re-use of CSCL data and analytic tools to provide a new perspective on group cohesion. CSCL2011, Hong Kong (submitted)

[22] Dessus, P., Mandin, S., Zampa, V.: What is teaching? Cognitive-based tutoring principles for the design of a learning environment. In: Tazi, S., Zreik, K. (eds.) Common innovation in e-learning, machine learning and humanoid (ICHSL.6) pp. 49-55. Europa/IEEE, Paris (2008)

[23] Zampa, V., Lemaire, B.: Latent Semantic Analysis for User Modeling. J. Intell. Inf. Syst. 18, 15-30 (2002)

[24] Dessus, P.: An Overview of LSA-Based Systems for Supporting Learning and Teaching. Proceeding of the 2009 conference on Artificial Intelligence in Education: Building Learning Systems that Care: From Knowledge Representation to Affective Modelling. IOS Press (2009)

[25] Kintsch, E., Caccmise, D., Franzke, M., Johnson, N., Dooley, S.: Summary Street®: Computer-guided summary writing. In: Landauer, T.K., McNamara, D.S., Dennis, S., Kintsch, W. (eds.) Handbook of Latent Semantic Analysis. Lawrence Erlbaum Associates, Mahwah, NJ (2007)

[26] Rehder, B., Schreiner, M.E., Wolfe, M.B., Laham, D., Landauer, T.K., Kintsch, W.: Using latent semantic analysis to assess knowledge: Some technical considerations. Discourse Processes 25, 337-354 (1998)

[27] Wolfe, M.B., Schreiner, M.E., Rehder, B., Laham, D., Foltz, P.W., Kintsch, W., Landauer, T.K.: Learning from text: Matching readers and text by latent semantic analysis. Discourse Processes 25, 309-336 (1998)

[28] Contractor, N.: The emergence of multidimensional networks. Journal of Computer-Mediated Communication 14, 743-747 (2009)

[29] Suthers, D., Dwyer, N., Medina, R., Vatrapu, R.: A framework for conceptualizing, representing, and analyzing distributed interaction. International Journal of Computer-Supported Collaborative Learning 5, 5-42 (2010)

[30] Rasmusen, E.: Games and information: An introduction to game theory. Blackwell, Malden, MA (2007)