J. Comput. Educ. (2014) 1(2–3):113–131
DOI 10.1007/s40692-014-0013-y

Using content analysis and domain ontologies to check learners' understanding of science concepts

Oliver Daems · Melanie Erkens · Nils Malzahn · H. Ulrich Hoppe

Received: 2 May 2014 / Revised: 29 June 2014 / Accepted: 11 July 2014 /
Published online: 14 August 2014

Abstract The ongoing EU project JuxtaLearn aims at facilitating the acquisition of science concepts through the creation and sharing of videos on the part of the learners. For several domains of learning threshold concepts have been specified as key elements of target knowledge. Content analysis techniques are used to extract learners' concepts manifested in textual artifacts and to contrast these with the anticipated domain concepts (represented as an ontology). Deviations between student concepts and the ontology can indicate problems of understanding and possibly misconceptions. Two studies explore the potential of (semi-) automated analysis of textual artifacts to identify and characterize the students' comprehension problems around knowledge artifacts. In the first study, protocols from a "flipped classroom" style teacher–student workshop are analyzed. The second study analyses comments to videos from educational video platforms. The "network text analysis" approach was used as a basis for both studies. As an extension of this approach, we have introduced "signal concepts" and their relations to domain concepts as indicators of potential information needs and problems of understanding.

Keywords Conceptual change · Network text analysis · STEM learning · Video-based learning

O. Daems (✉) · M. Erkens · N. Malzahn
Rhine-Ruhr Institute for Applied System Innovation, Duisburg, Germany
e-mail: od@rias-institute.eu

M. Erkens
e-mail: me@rias-institute.eu

N. Malzahn
e-mail: nm@rias-institute.eu

H. U. Hoppe
COLLIDE Research Group, University of Duisburg-Essen, Duisburg, Germany
e-mail: hoppe@collide.info

Introduction

The ongoing European project JuxtaLearn aims at fostering learning and curiosity in different fields of science (or STEM) by combining or "juxtaposing" the understanding of domain concepts with performing. Concretely, the students' performance is substantiated in the form of creative video making and editing activities. We see this way of learning by performing and presenting as a variant of Papert's "constructionism" (Papert and Harel 1991) and as similar to learning by teaching (Gartner et al. 1971). In this context, we are interested in studying the role of video as a medium for learning in different (including passive) forms of usage.

The design of learning activities in JuxtaLearn is guided by previously identified threshold concepts (Meyer and Land 2003). Threshold concepts are the basis for reinforcing deeper understanding and further creative production through scaffold reflections focused on essential elements. To identify such concepts and to explore how they are understood and appropriated by teachers and students, a series of face-to-face workshops have been conducted. After an initial workshop with science teachers, a second workshop also involved a group of six A-level students. As misconceptions surrounding threshold concepts are difficult to pin down, the workshop was structured through a role reversal in which the students taught the teachers, to elicit a deeper understanding of the gaps in the students' knowledge. In this context, learning analytics techniques have been used to extract structured representations of the underlying conceptual relations.

From another angle, we have also tried to identify processes of understanding around videos by analyzing existing web-based learning communities, namely the Khan Academy. Since the videos are also available on YouTube EDU as a less "educationally guided" environment, we were able to compare comments from both contexts. For our analysis, we only look at comments, e.g., questions and answers from discussions on the video page and not the videos themselves. We are particularly interested in extracting information from these texts to shed light on the following aspects:

- Associations of concepts as indicators of the authors' mental constructs (which may be adequate or inadequate from a scientific point of view);
- Concepts that are frequently addressed in questions as indicators for possible origins of problems of understanding;
- Associations between concepts often used in answers as indicators for missing relations in the original understanding.

We use these indicators to infer possible misconceptions or "stumbling blocks." As known from classical learner modeling (cf. Wenger 1987), we have to distinguish between missing knowledge about concepts and/or relations and misconceptions as often idiosyncratic constructions of incorrect knowledge. An example of a misconception (beyond missing knowledge) would be an incorrect association between two or more concepts.

Related work and background

Conceptual change

When learning new and complex concepts in school, students interpret the knowledge taught in relation to their prior knowledge (Krüger 2007). In everyday life, they might simply add newly gained knowledge to their existing knowledge base, if some new information is combinable with the existing knowledge. This process is called *enrichment* (Chi 2008; Inagaki and Hatano 2008; Carey 1985). But the typical school situation is different, here, the students have to face cognitive conflicts, because the corresponding knowledge structures often differ from each other (Vosniadou 2007). Vosniadou and Brewer (1992) offer an illustrative example for such a cognitive conflict and the possible resulting changes of conceptual models: if children learn that the earth is a sphere, this information may be in contradiction with their everyday knowledge, since they do not fall down from earth. This conflict ends up in two ways: either the formation of *misconceptions*, erroneous conclusions (Chi 2008), or a *revision* of prior knowledge, which describes the transformation from wrong into correct knowledge (e.g., Chi 2008; Vosniadou 2007; Hatano and Inagaki 2002; Carey 1985). Vosniadou and Brewer (1992) report that some children, who do not combine but instead just add new information, created a new mental model with two worlds as a result—one flat earth, where people can not fall down from and another one rotating as a solar object in the sky. Consequently, these misconceptions are based on false beliefs, accumulated false beliefs, or ontological miscategorisation (Chi 2008). Concurrently the conflict provides the possibility for a *revision*, which means that students restructure their existing knowledge, in this case with regard to the law of gravity, which leads to the correct knowledge. This revision can be supported by the influence of social processes (Hatano and Inagaki 1997) as they are envisioned in the JuxtaLearn learning process, and by the meta-conceptual awareness of own beliefs (Vosniadou 2003). The question remains how to make the students' conceptual models visible for them or even their teachers? Since domain knowledge in the field of STEM learning has to be constructed in the form of complex relations (e.g., structural, temporal, or causal) between a variety of concepts, we conclude that text networks based on online comments are suitable to reveal the student's knowledge and to represent their problems of understanding.

A network perspective on conceptual models and conceptual change

In various scenarios of learning, knowledge building and knowledge production, humans externalize their knowledge in terms of "knowledge artifacts," which are often represented in the form of texts, and thus susceptible to being analyzed by text mining (Heyer et al. 2006). Content analysis, as a form of artifact analysis, can be used to reduce qualitative textual data into clusters of conceptual categories aiming to unfold patterns and relationships of meaning (Julien 2008). Although several (semi-)automated methods can be used to detect these patterns from content, e.g., statistical methods based on the Vector Space Model (VSM) or Latent Dirichlet

Allocation (Blei 2012) as a probabilistic method, to date these methods have barely been used on learning data (He 2013). Content analysis in the context of learning data has primarily been used for the clustering of resources, e.g., the grouping of e-learning resources according to their similarity (e.g., Hung 2012; Tane et al. 2004). Sherin (2012) found that even without using semantic background knowledge, a VSM-based clustering of spoken word transcripts is an adequate instrument to identify student's concepts and the dynamics of their mental constructs. Additionally, He (2013) provided evidence that similar techniques are suitable for grouping learners' main topics in student-to-teacher online questions and peer-to-peer chat messages related to online video learning lessons.

The aforementioned methods are based on the "bag of word" model, in which the given order of words in a text is of no relevance to the analysis (Blei 2012). A method that takes into account the words' positioning is "Network Text Analysis" (NTA). NTA is a text mining method, which is based upon the assumption that knowledge can be modeled as a network of concepts (Carley et al. 2013a). Against this background, a concept is a single idea, which is represented by one or more words in a network (nodes). The links representing semantic relationships between these words (edges) are differing in strength, directionality, and type based on the words' position to each other in the text (Carley et al. 2013a). The union of all relations builds the semantic network (Carley et al. 2013a), similar to the relational network of a concept map. Similar to text networks, concept maps are networks, in which knowledge is represented by concepts and their relationships to each other. They differ from text-based semantic networks inasmuch as they are arranged hierarchically with the ontological root concepts at the top (Novak and Cañas 2008). In the context of knowledge construction research, concept maps are often used to trace the student's knowledge development (Engelmann and Hesse 2010; Engelmann et al. 2009; Schreiber and Engelmann 2010).

Jacobsen and Kapur (2010) have suggested to conceive learners' mental models or "ontologies" as scale-free networks, which would allow applying known characteristics of such networks to theories of conceptual change. According to Barabási (2009), the evolution of scale-free networks can be explained by the mechanisms of "preferential attachment." Applied to learners' ontologies, preferential attachment means that newly learned concepts are most frequently associated or linked to those concepts that are already more densely connected than others. From this, Jacobsen and Kapur (2010) conclude that such "hubs," i.e., nodes with a relatively high degree centrality, represent root categories of knowledge domains. Hoppe et al. (2012) support this theory in a study in which the volunteers had to create concept maps on the subject of global warming. This study clearly showed a scale-free nature of the maps in terms of an inverse power law degree distribution (a known structural characteristic of scale-free networks). This implies that there are more hubs than to be expected in a randomly connected network. Also, Hoppe et al. (2012) could show that certain graph-theoretical structural measures known, correlate with quality judgements of the maps by independent experts. Interestingly, the "density" measure shows a highly significant negative medium correlation with criterion of map "completeness." Again the scale-free model provides a clear explanation: In a growing scale-free network, the density is anti-proportional to the

size of the network, i.e., the smallest networks will show the highest density. Based on this characterization, the authors hypothesize that newly appearing hubs represent 'hot spots of conceptual change' (Hoppe et al. 2012, p. 297), whereby this change describes a restructuring process, in which learners revise their false beliefs and misconceptions on the relational or ontological level (Chi 2008). If the number of edges around a node is reduced suddenly, this may indicate a qualitative change of understanding or paradigm shift (Hoppe et al. 2012). Viewing concept maps as networks allows for applying a variety of techniques known from Social Network Analysis (cf. Wasserman and Faust 1994). As an example, the betweenness centrality measure is suggested as a possible indicator to identify 'bridge concepts' that link different knowledge domains (Hoppe et al. 2012). It is important to note that the results of NTA (see above) are also networks that can be further analyzed in the same way as concept maps. This would also allow the comparison of textual input (e.g., from student wikis) with concept maps.

Our basic idea and approach is to use content analysis techniques to generate network representations from knowledge artifacts originally created by students or experts and to apply structural and differential (comparative) measures to these representations in order to detect similarities or mismatches. In this approach, expert maps or ontologies can be used as "normative" references for comparison, e.g., to indicate deviations from standard domain knowledge and possibly possible misconceptions. Regarding the evolution of maps, also certain structural features and anomalies, can also be detected.

Network text analysis

Method

We have used the method of "Network Text Analysis" (NTA) as a basis of our analyses of textual artifacts. As defined by Carley et al. (2013a, b), the NTA workflow consists of three main steps: (1) data selection and extraction, (2) text preprocessing and (3) network analysis, which we have applied by the AutoMap/ORA toolset for NTA (Carley et al. 2013a, b). As a first step, the data of interest will be selected, depending on the terms of reference context, for instance, the selection of comments belonging to a certain person or video. In the second step, the preprocessing functions are intended to prepare the textual data for subsequent analyses. Unneeded and unwanted concepts will be reduced removed through simple text cleaning functions such as the removal of extra spaces. Furthermore, this step serves to apply (a) a stemming for reducing words to their root stem by removing suffixes from words, (b) a delete list, which is required for the removal of non-relevant stop words (articles, auxiliary verbs, etc.), and (c) a manually generated thesaurus, used for replacing synonym concepts with the more standard form, for combining *n*-grams and to correct spelling errors. The next step of preprocessing is the identification and classification of concepts. Relevant concepts will be detected by analyzing the words' frequencies based on the following principle: words and *n*-grams that appear more than *x*-times are considered as relevant and will

be included into further analyses, whereby x depends on the size of the corpus. The classification is done by the determination of categories based on the words appearing in a concept list that includes the frequency of all words and can be reduced to the most important key words using a threshold defined by the researcher. After specifying the categories, every single concept will be assigned to one of them. Therefore, an ontology-based meta-thesaurus will be created, which is later used for generating the network. As a result of the processes described before, multimodal networks will be created, whereby the modality of the network depends on the number of categories that can be identified by the researcher.

The analysis process has been applied to the transcripts of an initial role reversal workshop surrounding the STEM topics "moles" in chemistry, "potential difference" in physics and "cells" in biology. In each lesson, two students with different school marks (both excellent/mixed/both middle) taught two teachers. The transcripts were analyzed with the method described above.

Results

As a result of this analysis process, a multimodal network of categorized concepts was generated, graphically represented using ORA. Categories defined are: pedagogical concepts, domain concepts, general concepts (i.e., concepts that are neither domain specific nor pedagogic concepts), tools, and actors, whereby we have declared actors and domain concepts as the most relevant categories. These categories are represented in a meta-thesaurus derived from ontology. Within the thesaurus, the actor category represents all acting persons in the lessons; teachers have been labeled as T1–6, the researcher staff as R1–4 and students as S1–6. The domain concept category represents discipline-specific topics associated with the lesson subjects.

The number of connections between one actor and surrounding topics (also called "degree") indicates the thematic richness of an actor's contributions. Regarding the whole semantic network, the degrees reflect that both physics students (S5 and S6), as well as one of the biology students (S1), who score high in their school marks,also have a higher total degree in the network than the other students with middle school marks. Figure 1 depicts an excerpt of the resulting network, a two-mode network of actors and domain concepts, which contains 16 actor nodes (black circular nodes) and 71 domain concept nodes (rectangular). This excerpt only shows six actors (students) and the domain concepts with a connection to at least one of these actors. In Fig. 1, every subject area and corresponding sub-activity in the workshop (chemistry, biology, and physics) form a cohesive cluster in the overall network as illustrated by gray circles in Fig. 1.

Other relevant structural properties of the extracted network are nodes that represent hubs (nodes with a high total degree centrality) or nodes that bridge over between other concepts or between areas of discourse; which in this case are the domains (high in betweenness centrality). Figure 2 shows the top 6 knowledge items ranked according to their total degree centrality and betweenness centrality in the domain concept $\times$ domain concept network.

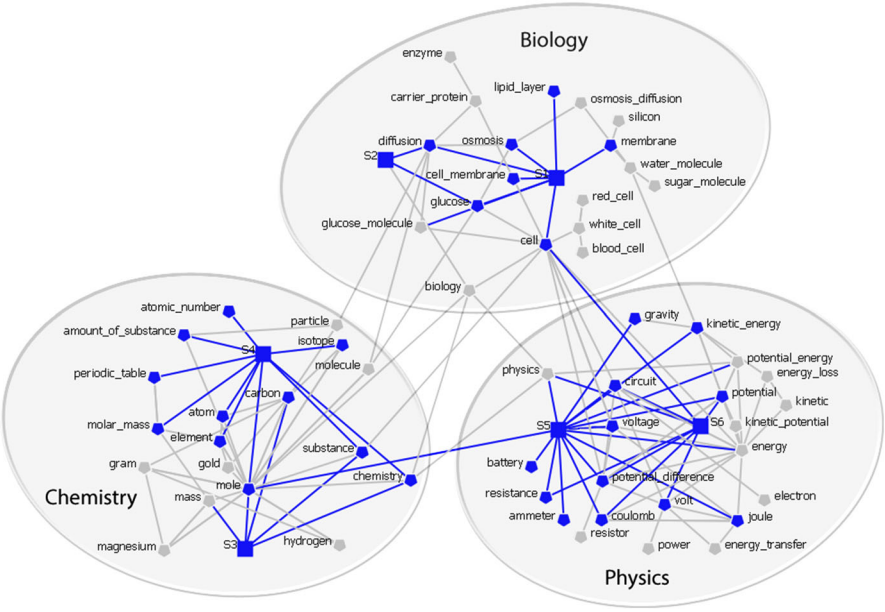

Fig. 1 Partial view of a network generated from a teacher-student workshop, depicting connections of students (S1 to S6) to domain concepts in blue (two-mode network) and relations between domain concepts in grey. (Color figure online)

Concerning total degree centrality, we could find the three main topics of the workshops "mole," "potential difference," and "cell" within the top 6 ranking. Furthermore, the nodes with a high betweenness centrality seem to bridge over different ontological areas. Mole, for instance, builds a fundamental connection between the physics-related cluster and the biology's cluster. Furthermore, cell, second highest in betweenness centrality, also connects the three discipline clusters with each other.

Based on this first example, we claim that text networks can not only represent and characterize the specific foci of the workshops, but also provide ontological information, which is related to the student's conceptualization of specific science topics.

Analysis of video comments

Content analysis

While the NTA workflow described above was conducted on material generated in the context of the JuxtaLearn project, for further content analysis, we focused on material found outside our project, therefore, externalizing our analysis. The approach to content analysis presented here focuses on the attention given to a specific topic in an online discussion around learning materials which in this case

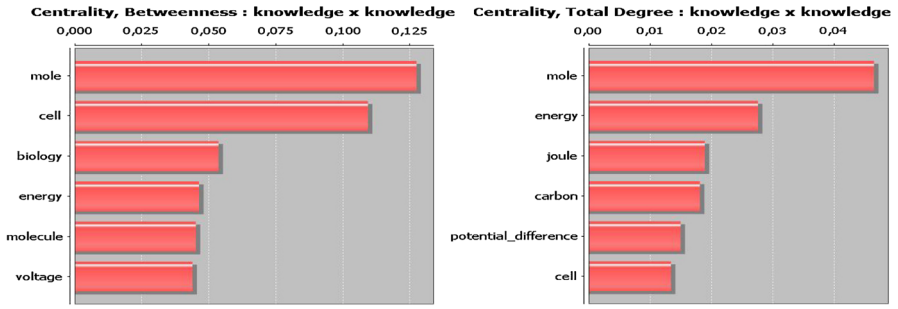

Fig. 2 Domain concepts with highest betweenness centralities (on the *left*) and highest total degree (on the *right*) in a domain concept × domain concept network

are represented by online discussions, composed of comments, and around educational videos. This approach aims to extend the NTA workflow by focusing on the content of single comments instead of the complete discussion.

We used a quantitative approach implementing regular expressions-based matching with the concept lists constructed during the NTA workflow and the preprocessed text base which was generated during this workflow. In principal, this generates a sparsely populated feature vector for each comment based on the occurrence frequency of concepts from each list. We use these vectors to construct a measure we call "semantic richness" which tries to quantify the relation between the domain specific concepts and the general concepts mentioned in each comment. This measure follows the assumption that a comment that uses more domain concepts than general concepts compared to a different comment which uses a lesser ratio of domain and general concepts is more "on topic" thus semantically richer. During our analysis of external data, we derived different methods of evaluating this ratio and the following will present our results.

Khan Academy and YouTube videos

As a starting point for our content analysis, we have chosen Khan Academy's learning videos and the accompanying discussion in the comments section. The Khan Academy website provides a significant amount of videos on different STEM topics and offers the option to enter into a learning dialog with other students, the same scenario which is envisioned for the JuxtaLearn platform. The website offers message boxes below the videos to enter this dialog through a scaffold question/answer construct rather than an open comment section. The videos themselves are hosted on YouTube EDU, which also enables users to comment on the same videos without assistive scaffolding. The library covers science topics such as biology, chemistry, and physics on different levels ranging from junior high school to university and holds more than 4.300 videos with an average length of 10 min (Khan Academy 2013). For this case study, we used both three videos with the titles "The Mole and Avogadro's Number," "Diffusion and Osmosis," and "Voltage - Difference between electrical potential (voltage) and electrical potential energy." In

Table 1 Type and number of extracted comments on STEM videos at Khan Academy's website

Type of artifact	Biology	Chemistry	Physics
Number of questions	184	279	70
Number of answers	312	362	77
Number of comments	496	641	147

Table 2 Subject and number of extracted comments on STEM videos from YouTube EDU

Subject of artifact	Biology	Chemistry	Physics
Number of comments	487	628	86

total, we have extracted 1.284 comments Khan Academy's web service. These textual artifacts have been used as a sample for our analysis. Khan Academy provides the aforementioned scaffold discussion elements and encourages discussion on the topic of the video, which means we can assume that artifacts extracted from the discussion will be about the topic of the video as well. Table 1 gives an overview of the amount of artifacts used in our study. It shows the number of questions and answers for each area the video topic can be assigned to, as well as their sum (number of comments) for easy comparison with other sources.

YouTube EDU is a subsection of YouTube that focuses on educational videos. The idea behind it is to provide everything related to education, spanning from short lessons for supplementing school learning to full courses from universities, and other professional material from educators around the globe. As mentioned above, this includes the Khan Academy videos. For the purpose of comparison, we extracted video comments (using the appropriate Google API) for the same videos. In total we got 1.201 comments from YouTube EDU distributed among the three videos as shown in Table 2. YouTube EDU does not provide any scaffolding for discussion resulting in a mixture of comments and smaller discussion compared to Khan Academy's strict question/answer construct.

Data sampling and first observations

The aforementioned strict construct of questions/answers seems to be enforced by a strong moderation by the staff of Khan Academy, because the discussion artifacts have a very low amount of noise, e.g., spam comments. This is reflected by the high amount of very short and poorly written comments extracted from YouTube EDU. Additionally, the artifacts extracted from Khan Academy seem to focus on the video's topic rather than on the video itself, while the opposite is true for comments from YouTube EDU, which mostly contain short comments that focus on e.g., the quality of the video. This observation can be confirmed by analyzing the content based on the list of domain, and other concepts, we constructed during the NTA workflow. Employing a quantitative approach we counted the occurrences of general and domain concepts in artifacts from both platforms, representing these

Table 3 Results for our semantic richness calculations

Description	Value			
Source from which the basis was extracted	Khan Academy ($N = 1{,}284$)		YouTube ($N = 1{,}201$)	
	M	SD	*M*	SD
Sum of general and domain concepts (length of comment)	23.73	28.04	9.78	11.03
General concepts: with duplicates | only once	9.89 | 7.62	11.88 | 7.43	6.07 | 5.49	6.11 | 5.12
Domain concepts: with duplicates | only once	13.84 | 7.26	17.58 | 5.60	3.71 | 2.51	6.27 | 3.45
(1) Measure: domain concepts/length	0.54 | 0.38	0.23 | 0.20	0.26 | 0.22	0.26 | 0.22
(2) Measure: domain concepts/(general concepts + 1)	1.46 | 1.08	1.48 | 1.02	0.50 | 0.40	0.88 | 0.63

N number of comments extracted

values as feature vectors for general and domain concepts. Table 3 shows the results from this approach.

The two combinatory measures were introduced to create a singular value describing the semantic richness. We introduced two different formulas for both versions of feature vectors (those using all occurrences, therefore, with duplicates/ those using only the first occurrence), the first normalizes the value to a range of 0–1, but favors shorter comments as a comment containing only a domain concept will have a semantic richness of 1. The second favors longer comments, but the value is uncapped, meaning it can grow infinitely when only domain concepts are found (the $+1$ is needed to avoid division by 0).

The two combinatory measures were introduced to create a singular value describing the semantic richness. We introduced two different formulas for both versions of feature vectors (those using all occurrences, therefore, with duplicates/ those using only the first occurrence), the first normalizes the value to a range of 0–1, but favors shorter comments as a comment containing only a domain concept will have a semantic richness of 1. The second favors longer comments, but the value is uncapped, meaning it can grow infinitely when only domain concepts are found (the $+1$ is needed to avoid division by 0).

The numbers clearly indicate both longer and semantically richer artifacts on Khan Academy, which might be a result of the platforms focus on provision of educational content and discussion or simply the suspected strong moderation.

Introduction of "Signal Concepts"

Idea

The idea behind the introduction of signal concepts to supplement our use of general and domain concepts, is the observation that certain general concepts (e.g., related to explanation request) may indicate specific problems of understanding. These signal concepts occur in combination with one or more domain concepts and "signal" a specific relationship either between the author and a domain concept or between two domain concepts. This distinction is reflected by two types of signal concepts: unary and binary. Unary signal concepts refer only to one domain concept and typically express a specific information need or problem of understanding in the part of the author. Binary signal concepts reference two domain concepts in combination or inter-relation. Examples of unary signal concept are "help_needed" or "explain," which may indicate that the author has a problem in understanding the connected domain concept. A typical example for the binary type is "difference_between," which may indicate that the author thinks or inquires about a difference between two domain concepts. Signal concepts provide a resource for teachers interested in learning about potential problems that their students have. Thus, our approach indicates a topic or a domain concept that may be worth focusing on in a future lesson.

Figure 3 gives a partial view of an exemplary network of type (signal concept) $\times$ (domain concept) focusing on "difference_between." This network highlights an

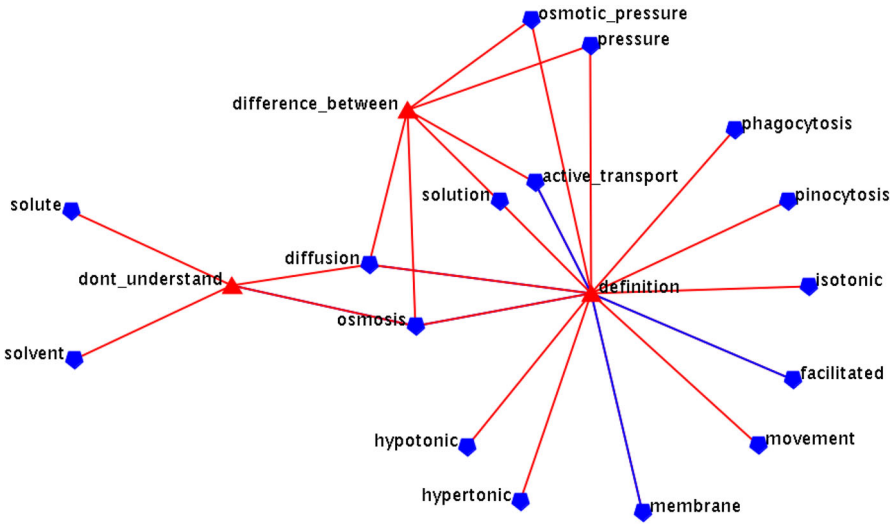

Fig. 3 Network showing an excerpt centered around "difference_between"

inherent problem with analyzing signal concepts through our standard NTA workflow: many instances of "difference pairs" will appear around the signal concept difference_between, but as a consequence of the underlying bi-partite graph representation we cannot formally distinguish these pairs without using semantic background knowledge (e.g., to separate out "osmosis" and "diffusion"). That is, the NTA workflow aggregates all occurrences of a signal concept into one single node in a kind of combinatorial multiplication that does no longer show the original instance pairs.

Approach

To avoid the "combinatorial multiplication" effect, we have replaced the NTA analysis at this stage by a tailored method that maintains the original context in which the pairs of domain concepts associated to a binary signal concept appeared. To visualize these pairs, we introduce "combination nodes" between the signal concept and corresponding pairs of domain concepts. The underlying analysis process is similar to NTA in that it also uses a text window running over the text source (i.e., a preprocessed comment) to detect the co-occurrences in this context. We do not use the whole comment as a context, since some comments are quite long and take up several different issues. For the results presented below, we used a window size of seven words from the preprocessed text, meaning the signal concept and the domain concept(s) must share a context with at most five words between them.

In addition, we have refined and extended this approach by again using the lists of domain concepts and general concepts constructed for the NTA analysis together with a matching algorithm that uses regular expressions to find and highlight

Table 4 Patterns used for digging up signal concepts from artifacts

Pattern	Description
(S)[OG]{0,5}(D)	Matches a unary signal (S) and a domain concept with 0–5 general or other concepts between them
(D)[OG]{0,5}([MT])[OG]{0,5}(D)	Matches a binary signal concept (T or M) that's in the middle of two domain concepts with 0–5 general or other concepts separating the domain and signal concept
(T)[OG]{0,5}(D)[OG]{0,5}(D)	Matches a binary signal concept (T) followed by 0–5 general or other concepts, a domain concept, another 0–5 general or other concepts and a second domain concept

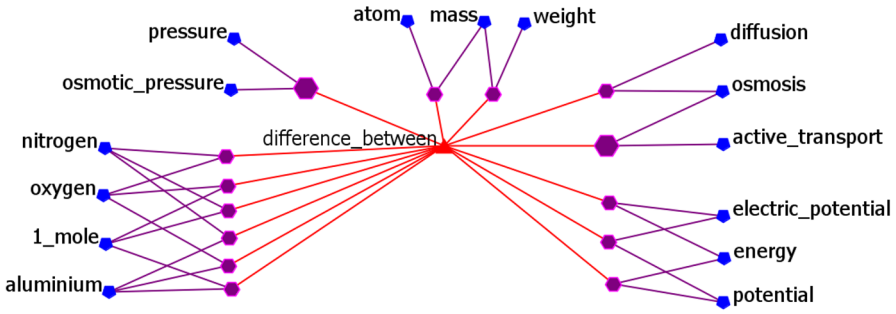

Fig. 4 Network including *blue combination nodes* hovering around "difference_between". (Color figure online)

patterns of domain and signal concepts in the comments. This algorithm transforms the preprocessed text into a compact string representation by replacing each word with a single letter. We used six letters (G, D, S, T, M, and O) which represent general concepts (G), domain concepts (D), unary signal concepts (S), binary signal concepts (T), binary signal concepts that need to be in the middle of two domain concepts (M), and other (O). The patterns, we used a selection of three patterns for the matching on these string representations, are as shown in Table 4, and they can be broken down into two different schemas, for unary and binary signal concepts. The rules reflect different orders of occurrence in the text for these two types of signal concepts.

These patterns were then used to highlight the signal and domain concept(s) in their original context as an additional way of visualizing possible problems in a human readable form.

Results

For the first approach, we generated a new network around the same signal concept "difference_ between" to illustrate the usefulness of our approach and specifically the introduction of the new combination nodes. Figure 4 only shows the "cloud" of concept pairs around the concept difference_between and uses combination nodes to separate the corresponding pairs.

Table 5 Highlighting of signal and domain concepts next to their original context

I finally understand osmosis. Thanks Khan!!
do_understand osmosis thanks khan
How I know if the membrane will allow sugar to diffuse or not? plzany body reply.
explanation i_know if **membrane** be allow sugar diffusion not plzany body reply
KhanAcademy helped me to review a unit on OSMOSIS AND DIFFUSION …
khan_academy **help** review **unit** on osmosis_diffusion …
Still confused about osmotic pressure :/wasted a bit of time..
still **confusion** about **osmotic_pressure**/waste bite time
What is the difference between osmosis and active transport
definition difference_between osmosis active_transport

The red node represents the signal concept that this network focuses on while the purple nodes are the new combination. The size of these combination nodes represents their occurrence in all of the analyzed texts. This means nodes will be larger the more often the combination of the signal concept and both domain concepts occurred in the specified text window. Blue nodes show all the domain concepts that were mentioned along with the signal concept "difference between" in the data we analyzed.

The second approach resulted in a list of comments with co-occurring signal and domain concept(s), containing highlighting for the referenced concepts. An excerpt of this list is shown in Table 5, illustrating the usefulness of this approach. It puts our preprocessed text with highlighting next to the original comment to indicate the detection of a signal and domain concept but also to show the original context both as a means of feedback for us by showing the validity of our approach, as well as a possible source of information in a learning analytics context.

Ontology extraction from Wikipedia

Idea

While the results described above were promising, the NTA workflow relied on automatic extraction, but required a manual classification of domain concepts. For a fully automatic approach, we employed the idea of extracting an ontology based on external sources. We think the approach of automatically extracting an ontology from available resources might be a viable solution. YAGO uses Wikipedia and WordNet (Suchanek et al. 2007), but focuses on describing persons. Therefore, the included taxonomy is based on entities while we require a domain vocabulary instead. While it will be interesting to explore the possibility of extracting a subset of YAGO that suits our needs, instead we decided to implement a very basic approach to evaluate the feasibility of ontology extraction from Wikipedia before evaluating YAGO's suitability.

We decided on using Wikipedia and its structure of categories for the same STEM disciplines as before, specifically: biology, chemistry, and physics, to extract a domain vocabulary, we could use in our content analysis.

Table 6 New results for our semantic richness calculations

Description	Value			
Source from which the basis was extracted	Khan Academy ($N = 3{,}814$)		YouTube ($N = 1{,}338$)	
	M	SD	*M*	SD
Sum of general and domain concepts (length of comment)	27.34	37.56	14.16	12.43
General concepts	24.66	35.29	13.36	11.44
Domain concepts	2.68	3.62	0.80	1.54
(1) Measure: domain concepts/length	0.104	0.098	0.045	0.738
(2) Measure: domain concepts/(general concepts + 1)	0.119	0.125	0.049	0.089

N number of comments extracted

Method

We used a web crawling technique to extract all categories for the three disciplines mentioned and declared the name of the pages related to these categories as words for our new domain concept dictionary. To keep the possibility of expanding this domain dictionary into a domain ontology, we decided to mirror the structure of categories and subcategories in a taxonomy and declared these classes for a domain ontology. The page names we extracted were then treated as instances of these classes, representing the words in our dictionary.

For the results described below, we used the top level category and pages, and then used their subcategories as a seed for a second iteration and did the same twice more. This means we used the category "Biology" and three levels of subcategories and their related pages as the source for our basic ontology.

Results

The extracted ontology contains more than 180.000 page names represented as instances. Using a different set of comment data, again extracted from videos stored on Khan Academy and YouTube, we compared a similar analysis as described before on a total of 5,152 comments. Here, semantic richness is no longer calculated using an interactively (or "manually") defined domain vocabulary but using the automatically extracted much larger domain ontology instead. As for signal concepts, we do not see how to automate the selection process, so we still rely on the same list used before.

The results for semantic richness are shown in Table 6, while Fig. 5 shows the newly calculated network of signal concepts with the combination nodes are sized according to their occurence value.

Since we do not have positive list or dictionary of all general concepts, we consider all words that are not either stop words or represent domain concepts to be general concepts. This new signal concept network was too complex, so it had to be reduced for the visualization to be human readable. We have removed combination nodes (pairs) with less than five occurrences.

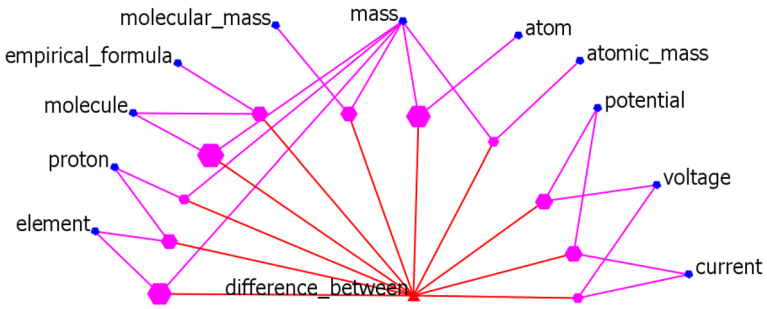

Fig. 5 New (reduced) network with scaled combination nodes around "difference_between"

Table 7 Occurence values for combination nodes

Combination node (Domain concept/domain concept)	Occurence value
Mass/molecule	12
Element/mass	11
Atom/mass	11
Voltage/potential	8
Potential/current	8
Element/proton	7
Molecule/empirical_formula	7
Molecular_mass/mass	7
Mass/atomic_mass	5
Mass/proton	5
Current/voltage	5
Carbonation/hydrogen	4
Atom/carbonation	4
Particle/cass	4
Isotope/element	4
Potential/electron	4
Element/atom	4
Element/molecule	4

Table 7 shows an excerpt of a ranked list of domain concept combinations around "difference_between" found in the comments of physics and chemistry videos.

Conclusion

In our first study, we aimed to find evidence that NTA is a useful instrument to identify and analyze students' conceptual understanding, while learning a specific science topic. Similar to a student's self-created concept map, (semi-)automatically

generated text networks provide information about the learners' conceptions of the domain. As theoretically expected, we could corroborate that hubs found in the ensuing concept networks indeed represent ontological root categories, e.g., the concept of "mole" that was the main topic of a role reversal lesson and consequently reaches a high value for total degree. Furthermore, given betweenness centrality values support the theory that nodes high in this measure indicate bridges between different ontological areas. Overall, we could conclude that the analysis of a network's structural features provides important information on ontology development. This led us to a second study, in which we investigated the impact of differentiating between various categories.

While investigating general concepts of comments in a discussion around publicly available educational videos, we could identify an ontological subcategory of *signal concepts* among frequent general terms high in degree. These words (such as "explanation" or "difference_between") indicate a certain type of concepts, which in conjunction with domain concepts can help to identify problems of comprehension. By means of extending the NTA by a tailored analysis, we receive indicators for drawing conclusions as to which (pre-)knowledge the students might miss or which concepts are particularly difficult to distinguish (such as "osmosis" and "diffusion").

As a next step on our research agenda, we will particularly look at the progression of concept networks over time. Particularly deviations from "normal" progression following the preferential attachment principle (i.e., disappearing hubs or new clusters of high connectivity) will be analyzed to enable a better understanding for the evolution of the students' conceptual models.

Acknowledgments This work was partially funded by the European Community's Seventh Framework Program (FP7/2007-2013) under Grant agreement 317964 JUXTALEARN.

References

Barabási, A.-L. (2009). Scale-free networks: A decade and beyond. *Next Science, 325*(5939), 412–413. doi:10.1126/science.1173299.

Blei, D. M. (2012). Introduction to probabilistic topic models. *Communications of the ACM, 55*(4), 77–84. doi:10.1145/2133806.2133826.

Carey, S. (1985). *Conceptual Change in childhood*. Cambridge, MA: MIT Press.

Carley, K. M., Columbus, D., & Landwehr, P. (2013a). *AutoMap user's guide 2013* (pp. 1–219). Technical Report No. CMU-ISR-13-105. Pittsburgh, PA: Carnegie Mellon University, Institute for Software Research. www.casos.cs.cmu.edu/publications/papers/CMU-ISR-13-105.pdf.

Carley, K. M., Pfeffer, J., Reminga, J., Storrick, J., & Columbus, D. (2013b). *ORA user's guide 2013* (pp. 1–1280). Technical Report No. CMU-ISR-13-108. Pittsburgh, PA: Carnegie Mellon University, Institute for Software Research. http://www.casos.cs.cmu.edu/publications/papers/CMU-ISR-13-108.pdf.

Chi, M. (2008). Three types of conceptual change: Belief revision, mental model transformation, and categorical shift. In S. Vosniadou (Ed.), *International handbook of research on conceptual change* (pp. 61–82). New York: Routledge.

Engelmann, T., Dehler, J., Bodemer, D., & Buder, J. (2009). Knowledge awareness in CSCL: A psychological perspective. *Computers in Human Behavior, 25*(4), 949–960.

Engelmann, T., & Hesse, F. W. (2010). How digital concept maps about the collaborators' knowledge and information influence computer-supported collaborative problem solving. *International Journal of Computer-Supported Collaborative Learning, 5*, 299–319. doi:10.1007/s11412-010-9089-1.

Gartner, A., Kohler, M., & Riessmann, F. (1971). *Children teach Children: Learning by teaching*. Dallas, TX: Harper and Row.

Hatano, G., & Inagaki, K. (1997). Qualitative changes in intuitive biology. *European Journal of Psychology of Education, 12*, 111–130.

Hatano, G., & Inagaki, K. (2002). *Young children's thinking about biological world*. New York: Psychology Press.

He, W. (2013). Examining students' online interaction in a live video streaming environment using data mining and text mining. *Computers in Human Behavior, 29*(1), 90–102. doi:10.1016/j.chb.2012.07.020.

Heyer, G., Quasthoff, U., & Wittig, T. (2006). *Text mining: Wissensrohstoff text*. Bochum: W3L-Verlag, Herdecke.

Hoppe, H. U., Engler, J., & Weinbrenner, S. (2012). The impact of structural characteristics of concept maps on automatic quality measurement. In J. van Aalst, K. Thompson, M. J. Jacobson & P. Reimann (Eds.), *The Future of Learning*: *Proceedings of the 10th International Conference of the Learning Sciences* (*ICLS 2012*, pp. 291–298). Sydney, Australia.

Hung, J. (2012). Trends of E-learning research from 2000 to 2008: Use of text mining and bibliometrics. *British Journal of Educational Technology, 43*(1), 5–16. doi:10.1111/j.1467-8535.2010.01144.x.

Inagaki, K., & Hatano, G. (2008). Conceptual change in naive biology. In S. Vosniadou (Ed.), *International handbook of research on conceptual change* (pp. 240–262). New York: Routledge.

Jacobsen, M. J., & Kapur, M. (2010). Ontologies as scale free networks: Implications for theories of conceptual change. In K. Gomez, L. Lyons & J. Radinsky (Eds.), *Learning in the Disciplines*: *Proceedings of the 9th International Conference of the Learning Sciences* (*ICLS 2010* (pp. 193–194).

Julien, H. (2008). Content analysis. In L. M. Given (Ed.), *Qualitative research methods* (pp. 120–121). Thousand Oaks, CA: SAGE Publications.

Khan Academy. (2013). *A free world-class education for anyone anywhere*. Khan Academy. https://www.khanacademy.org/about.

Krüger, D. (2007). Die conceptual change-Theorie. In D. Krüger & H. Vogt (Eds.), *Theorien in der biologiedidaktischen Forschung* (pp. 81–92). Berlin: Springer.

Meyer, J. H., & Land, R. (2003). Threshold concepts and troublesome knowledge: Linkages to ways of thinking and practising. In C. Rust (Ed.), *Improving student learning—Theory and practice ten years on* (pp. 412–424). Oxford: Oxford Centre for Staff and Learning Development (OCSLD).

Novak, J. D., & Cañas, A. J. (2008). *The theory underlying concept maps and how to construct and use them*. Technical Report IHMC CmapTools 2006-01 Rev 01-2008 No. 2. Pensacola, FL: Florida Institute for Human and Machine Cognition. http://cmap.ihmc.us/Publications/ResearchPapers/TheoryUnderlyingConceptMaps.pdf.

Papert, S., & Harel, I. (1991). *Constructionism*. Norwood, NJ: Ablex Publishing Corporation. http://namodemello.com.br/pdf/tendencias/situatingconstrutivism.pdf.

Schreiber, M., & Engelmann, T. (2010). Knowledge and information awareness for initiating transactive memory system processes of computer-supported ad hoc groups. *Computers in Human Behavior, 26*, 1701–1709. doi:10.1016/j.chb.2010.02.007.

Sherin, B. (2012). Using computational methods to discover student science conceptions in interview data. In *LAK'12*: *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* (pp. 188–197). New York: ACM. doi:10.1145/2330601.2330649.

Suchanek, F. M., Kasneci, G., & Weikum, G. (2007). YAGO: A large ontology from Wikipedia and WordNet. In *World Wide Web Conference 2007—Semantic Web Track*, September 2008 (Vol. 6, Issue 3, pp. 171–240), Banff, Canada.

Tane, J., Schmitz, C., & Stumme, G. (2004). Semantic resource management for the web: an e-learning application. In *Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters* (pp. 1–10). ACM.

Vosniadou, S. (2003). Exploring the relationships between conceptual change and intentional learning. In G. M. Sinatra & P. R. Pintrich (Eds.), *Intentional conceptual change* (pp. 377–406). Mahwah, NJ: Erlbaum.

Vosniadou, S. (2007). Conceptual change and education. *Human Development, 50*(1), 47–54. doi:10.1159/000097684.

Vosniadou, S., & Brewer, W. F. (1992). Mental models of the earth: A study of conceptual change in childhood. *Cognitive Psychology, 24*, 535–585.

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. New York: Cambridge University Press.

Wenger, E. (1987). *Artificial intelligence and tutoring systems: Computational and cognitive approaches to the communication of knowledge*. San Francisco: Morgan Kaufmann Publishers, Inc.

Oliver Daems is a researcher at the "Rhine-Ruhr Institute for Applied System Innovation" (RIAS e.V.). His current research focusses on semantic web technologies, recommender systems and learning analytics.

Melanie Erkens operates at the interface between learning psychology and learning analytics. As a researcher at the "Rhine-Ruhr Institute for Applied System Innovation" (RIAS e.V.) her focus lies on the application of network text analysis to explore conceptual models. Furthermore, she uses text mining techniques and their results 'visualization to improve students' cognitive awareness with regard to structure their learning processes.

Nils Malzahn is a researcher at the "Rhine-Ruhr Institute for Applied System Innovation" (RIAS e.V.) His past and current research concentrates on combining semantic web technologies with social network analysis to build support tools for learning and knowledge management.

H. Ulrich Hoppe holds a full professorship for "Cooperative and Learning Support Systems" in the Department of Computer Science and Applied Cognitive Science at the University of Duisburg-Essen, Germany. He is the founder of the research group COLLIDE (Collaborative Learning in Intelligent Distributed Environments) with which he has been engaged in more than 10 European projects since 1995. His current research interests include the analysis, modelling, and intelligent support of interactive and collaborative learning processes as well as social network analysis and community support.