

Visualisierung konzeptuellen Lernens durch semantische Vernetzung sequenzieller Lehrinhalte

Robin Nicolay¹ Bastian Schwennigcke² Sarah Sahl³ Prof. Alke Martens⁴

Abstract: Akademische Ausbildung soll mehr als die reine Vermittlung von Lernstoff leisten. Das setzt voraus, dass die Lernenden theoriegeleitet arbeiten und die dabei maßgeblichen Konzepte überblicken und anwenden. Wie Lernende die Aneignung wissenschaftlicher Konzepte bewältigen und welche Herausforderungen sich ihnen dabei stellen, ist Teil der hochschuldidaktischen sowie der lerntheoretischen Forschung. Dieser Beitrag beschreibt Herausforderungen bei der Konzeptvermittlung durch Medienrezeption (Printmedien und Vorträge), sowie Aspekte des Konzepterwerbs durch Exploration und Explikation. Aus Sequenzen von Lehrstoff werden Experten-Modelle abgeleitet, die durch Themen und semantische Verbindungen repräsentiert sind. Auf dieser Basis wird ein Ansatz zur Messbarkeit der konzeptuellen Ordnung des Lehrinputs sowie der Differenzen mit den konzeptuellen Rekonstruktionsleistungen Lernender beschrieben. Hierbei kommen Verfahren der Latent-Dirichlet-Allocation und des Topic Modeling zum Einsatz.

Keywords: konzeptuelles Lernen, Konzeptrekonstruktion, Latent-Dirichlet-Allocation, Wissenstransfer, Lehrformen

1 Einleitung und Motivation

Akademische Ausbildung zielt auf den Erwerb konzeptueller Kompetenz. Bereits mit Absolvierung des ersten Studienzyklus sieht der Europäische Qualifikationsrahmen vor, dass Lernende „komplexe und unvorhersehbare Probleme [...] in einem spezialisierten Arbeits- oder Lernbereich“ bearbeiten können und dass sie dabei auf ein „kritische[s] Verständni[s] von Theorien und Grundsätzen“, die innerhalb dieses Arbeitsbereiches einschlägig sind, zurückgreifen. [PR08, Anhang II, C111/5, Niveau 6]

Der Ansatz konzeptueller Kompetenz geht über die Vermittlung und Abfrage von Lernstoffen hinaus. Es wird eine intellektuelle Entwicklung angestrebt [La12, S. 33-36], in deren Verlauf Lernende die Fähigkeit erwerben, Strategien und Lösungsansätze „durch Wahlentscheidungen“ zu motivieren, „bei denen gültiges und signifikantes Wissen eine zunehmende Rolle spielt.“ [Ku10, S. 14] Ein solches Wissen wird

¹ Universität Rostock, Institut für Informatik, Albert-Einstein-Strasse 22, 18059 Rostock, robin.nicolay@uni-rostock.de

² Universität Rostock, Institut für Informatik, Albert-Einstein-Strasse 22, 18059 Rostock, bastian.schwennigcke@uni-rostock.de

³ Universität Rostock, Institut für Informatik, Albert-Einstein-Strasse 22, 18059 Rostock, sarah.sahl@uni-rostock.de

⁴ Universität Rostock, Institut für Informatik, Albert-Einstein-Strasse 22, 18059 Rostock, alke.martens@uni-rostock.de

durch den konzeptuellen Haushalt eines Faches repräsentiert. Zu klären ist, wie die theoretischen und methodischen Konzepte eines Faches Eingang in die „Handlungs-, Operations- und Begriffsschemata“ [OR08, S. 26] findet, die Lernenden bei der Bearbeitung von Problemstellungen zur Verfügung stehen.

Dieser Vorgang ist nicht unmittelbar beobachtbar, was die Einschätzung der Wirksamkeit didaktischer Interventionen erheblich beeinträchtigt. Wichtig ist die Frage, welche Rolle Wissen bei der Auseinandersetzung mit Problemstellungen in Lernumwelten für Lernende spielt. Laurillard unterscheidet [La12]...

1. eine dualistische Wissensauffassung, die Aussagen nach dem richtig-falsch-Schema ordnet und die entsprechende Bewertung von einer Autorität bezieht,
2. eine offene Wissensauffassung, die multiple Perspektiven auf einen Gegenstand zulässt, diese aber als gleichwertig behandelt
3. und eine relativistische Wissensauffassung, die auf Positionierungen in Abgrenzung zu anderen möglichen Positionen im Umgang mit einem Gegenstand beruht.

Geprüft werden kann, ob Veränderungen im Umgang mit Wissen Auswirkungen auf die Rezeption von Wissensmedien haben. Erwartbar ist, dass ein zunehmendes Bewusstsein für Perspektivbildungen und konzeptuelle Differenzierung den sequentiellen Nachvollzug des Narrativs medialer Darstellungen in den Hintergrund treten lässt. Die Rekonstruktion semantischer, assoziativer und argumentativer Relationen gewinnt an Bedeutung. Offen ist, welche Entwicklungsschritte die Lernenden dabei machen, welche typischen Muster der Integration bzw. Ausgrenzung von Inhalten sich finden lassen, wie sich Interventionen auf Ebene der Lernumwelt auswirken und wie inhaltliche Verarbeitungsleistungen aussehen, welche mit den Intentionen der Lehrpersonen (nicht) konvergieren.

2 Technischer Ansatz: Topic Modelling

Erforscht wird, welche semantische Struktur sequenzierten Lehrinhalten wie Vorträgen und durch Lehrpläne verknüpfte Textmedien zugrunde liegt. Ziel ist eine Approximation des zugrundeliegenden Experten-Modells des Lehrenden und eine Abbildung der Interaktionen des Lernenden mit Lehrstoff auf diesem Model. Hierdurch soll die Art der studentischen Verarbeitung und Exploration von Themen und Relationen sichtbar werden. Der folgende Abschnitt beschreibt unseren Ansatz zur Inferenz eines semantischen Netzwerks aus Sequenzen von Vorlesungsmaterialien. Das semantische Netzwerk 1 versieht Sequenzen von Lehrinhalten mit einer Metastruktur über relevante Themen und deren Relationen.

Für die Inferenz einer Metastruktur auf Sequenzen von Lehrinhalten müssen sowohl relevante Themen, als auch semantische Relationen abgeleitet werden. Die folgenden drei Abschnitte beschreiben das LDA-Verfahren vorgestellt von David Blei

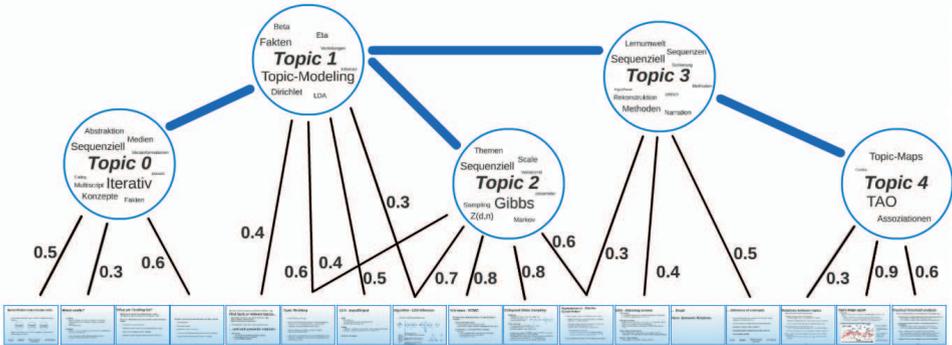


Abb. 1: Vereinfachte Darstellung eines inferierten, semantischen Netzwerks aus Themen und Themenrelationen in Form einer Topic-Map[MMM08]. Occurrences sind eine Sequenzen aus Vortragsfolien. Themen sind definiert durch "Wordclouds"relevanter Begriffe und miteinander durch Lehrverlauf und gemeinsame Occurrences verbunden.

in [BNJ03] und ein daraus entwickeltes Werkzeug als Ansatz zur Inferenz dieser Themen und Relationen.

2.1 Inferenz von Themen aus Lehrinhalten

Zur Ableitung von Themen wurden Sequenzen von Textinhalten durch Verfahren des *Topic-Modelling* analysiert. *Topic-Models* sind statistische Modelle. Sie finden unter anderem Verwendung in der Inferenz abstrakter Themen aus natürlich sprachlichen Textdokumenten. Für den Ansatz dieser Arbeit verwenden wir das Verfahren der "Latent-Dirichlet-Allocation"(kurz LDA). LDA ist eine Weiterentwicklung des "Probabilistic Latent Semantic Indexing"(kurz PLSI) beschrieben in [Ho99]. Ein Hauptmerkmal hierbei sind die hinzugefügten *Dirichlet-Priors*, deren Vorzüge in im Anschluss diskutiert werden.

Zusammengefasst inferiert LDA eine festgelegte Anzahl von Themen $T = t_1, t_2, \dots, t_k$ aus einer Menge von Dokumenten D mit Gesamtvokabular V . Hierbei liefert der Algorithmus eine diskrete Verteilung $\theta(d)$ über das Vorkommen aller Themen $t \in T$ für jedes Dokument $d \in D$. Die Summe der Wahrscheinlichkeiten des Vorkommens aller Themen t in Dokument d ergibt immer 1. Jedes Thema t selbst ist durch eine diskrete Verteilung $\beta(t)$ über die Wahrscheinlichkeit des Vorkommens jedes Wortes $w \in V$ definiert. Auch hier ist die Summe aller Wahrscheinlichkeiten des Vorkommens jedes Wortes $w \in V = 1$

Vortragsfolien arbeiten mit der Anreicherung von Inhalten und besitzen für eine Inferenz nur geringe Mengen von Text. Eine Untersuchung des *Dirichlet-Priors* im LDA-Algorithmus bietet in diesem Kontext viele Vorteile. Die Vorteile des LDA-Priors selbst beschreibt [WMM09] unter anderem, als Möglichkeit zur Anpassung der Inferenz durch Meta-Informationen und eine verbesserte Verarbeitung geringer

Textmengen. Weiterhin unterstützen Priors die Beobachtung von Themenentwicklungen über die Zeit [BL06]. Ein *Dirichlet-Prior* wird durch einen Konzentrationsvektor $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_k)$ mit $k = |T|$ für die Verteilung von Themen auf Dokumente und $k = |V|$ für die Verteilung der Wörter auf Themen festgelegt. Für gleiche $(\alpha_1, \alpha_2, \dots, \alpha_k)$ spricht man von symmetrischen, bei ungleichen von asymmetrischen *Dirichlet-Priors*. Verschiedene Ausprägungen des Konzentrations-Parameters haben folgenden, am Beispiel der Themenverteilung auf Dokumenten $d \in D$, beschriebenen Einfluss auf die LDA-Inferenz.

1. Symmetrisch, $\alpha < 1$: Wenige Themen mit hoher Wahrscheinlichkeit
2. Symmetrisch, $\alpha > 1$: Viele Themen mit ähnlich hoher Wahrscheinlichkeit
3. Asymmetrisch, hohe $u_{x,y,z}$: Themen $t_{x,y,z}$ mit erhöhter Wahrscheinlichkeit

Während sich asymmetrische Priors zur Anpassung der Verteilung von Themen auf Dokumente eignen, reduzieren Fall

2.2 Überführung des technischen Ansatzes: Inferenz von Vortragsthemen

Für erste Versuche zur Inferenz von Themen und abgeleiteten Relationen wurden Vorlesungsfolien aus dem Studiengang Informatik an der Universität Rostock untersucht. Erste Resultate werden am praktische Beispiel der Einführungsveranstaltung zu Webtechnologien mit 49 verwendeten Folien vorgestellt. Für die Durchführung erster Tests wurde eine Python-Implementation [ld15][Da11] des Markov-Chain-Monte-Carlo-Verfahren Gibbs-Samplings[GS04] verwendet. Die Inferenz wurde für 10 Themen mit 1500 Iterationen, und mit symmetrischem Themen-Prior $\alpha = 0.1$ und symmetrischem Wort-Prior $\alpha = 0.01$ (heuristisch für Texte nach [Zh08] optimiert) durchgeführt. Die 49 Vortragsfolien hatten ein Gesamt-Vokabular von $|V| = 840$ unterschiedlichen Wörtern. Zur Schärfung des Vokabulars wurden sukzessiv folgende Filter angewendet.

1. Unifizieren von Groß- und Kleinschreibung ($|V| = 797$)
2. Entfernen von Stoppwörtern [Po15] ($|V| = 726$)
3. Entfernen grammatikalischer Verformungen [Yo15] ($|V| = 691$)
4. Entfernen von Wörtern mit einer Länge kleiner 3 und Zahlen ($|V| = 632$)
5. Entfernen regelmäßig vorkommender Wörter (Beispiel: Fußzeilen) ($|V| = 632$)
6. Entfernen von Bild-Folien ohne Text

Anschließend wurden 10 Themen Berechnet. Im Folgenden ist exemplarisch eine Auflistung einiger Themen mit ihren 5 wahrscheinlichsten Wörtern in absteigender Reihenfolge aufgelistet.

Thema 1	Thema 2	Thema 3	Thema 4	Thema 5
web	business	media	dom	user
technolog	anwend	content	css	neu
konzept	entwickl	honeycomb	java	ajax
vorles	dollar	social	xml	serv
warum	million	user	konzept	respons

Tab. 1: Top 5 Wortstämme inferierter Themen

Tabelle 1 zeigt 5 der 10 inferierten Themen. Je Thema sind die 5 häufigsten Wörter der diskreten Verteilungen über $|V|=632$ Wörter je Thema gelistet. Bereits anhand dieser Wörter sind thematische Schwerpunkte und ein thematischer Verlauf erkennbar. Die Veranstaltung begann mit der Beschreibung der Vorlesungsinhalte, Konzepte und Technologien, über eine Motivation zu späteren Berufszielen und praktischen Fragen des Alltags zu ersten fachlichen Details über Webseiten und Kommunikationstechniken.

2.3 Inferenz gemeinsamer Kontexte

In einem ersten Ansatz wurden gemeinsame Kontexte und damit Relationen zwischen Folien und Themen untersucht. Hierfür wurden Verbindungen zwischen einzelnen Vorlesungsfolien gebildet, die sich ein gemeinsames Thema teilen. Dazu wurden nur Themen in Betracht gezogen, die mit einer Wahrscheinlichkeit von über 65% (manuell festgelegt) für die jeweilige Folie erkannt wurden. Die abschließende Abbildung 2 zeigt nun Verbindungen zwischen Folien, die sich mindestens eines der 10 erkannten Themen teilen. Bereits hier ist zu erkennen, dass aufeinanderfolgende Folien durch gemeinsame Themen verknüpft sind. Teilweise werden im späteren Verlauf der Vorlesung Themen vom Anfang der Vorlesung wieder aufgegriffen. Für einige Folien wurden keine Themen mit einer ausreichend hohen Wahrscheinlichkeit erkannt. Diese besitzen keine Verbindungen.

2.4 Beobachtung studentischer Interaktionen mit Lehrinhalten

Zur Beobachtung der studentischen Verarbeitung von Lehrinhalten kommen verschiedene Methoden in Frage. In voran gegangenen Arbeiten wurde hierzu eine Möglichkeit der Beobachtungen auf Basis von "Coding" [LLG08] und der Assoziation neuer Stimuli mit bekanntem Wissen untersucht [Ni15a] [Ni15b]. Im Fokus stand hierbei die Beobachtung der studentischen Rekonstruktion von Themen und semantischen Relationen im Rahmen der Vorlesungsveranstaltung. Hierzu wurden aktuelle Vortragsfolien während der Vorlesung mit Studenten geteilt und beobachtet, ob eine Annotation durch Markieren von Stichwörtern, Verknüpfen von Folien und der Formulierung von Konflikten, Rückschlüsse auf Lernprozesse erlauben. Erste Untersuchungen zeigten, dass Studenten in Vorlesungssituationen Informationen



Abb. 2: Chorddiagram über die Relationen zwischen einzelnen Vorlesungsfolien einer Vorlesung. Die Vorlesung beginnt oben mittig und schreitet im Uhrzeigersinn voran.

nicht intuitiv verknüpfen. Aktuell werden diese Untersuchungen auf die Beobachtung von Browsingverhalten durch E-Learning Inhalte und Nachbereitung durch Studenten erweitert.

Ein Ähnlicher Ansatz zur Beobachtung der studentischen Exploration von Wissen unter Verwendung des LDA-Verfahrens wurde durch [PK13] untersucht. Hierbei wurden Webseiten entsprechend ihrer Themen durch Experten klassifiziert und mit Hilfe von LDA relevante Klassen-Begriffe extrahiert. Anschließend wurde durch die Beobachtung des Browsingverhaltens von Studenten bei der Exploration von Wissen, beim Besuch passender Webseiten ein Lernzuwachs für entsprechende Themen berechnet. Während bei [PK13] der Lernerfolg von Themen betrachtet wurde, untersuchen wir die Inferenz eines Experten-Modells aus sequenzierten Lehrmaterialien. Hierbei steht die Entwicklung von Relationen zwischen Themen im Verlauf des Lehrens, sowie die Bewegung von Studenten auf diesem in Abbildung 1 skizzierten Netzwerk während des Lernens im Vordergrund.

3 Zusammenfassung und Ausblick

Im vorliegenden Beitrag wurden die technischen Vorüberlegungen zur Anwendung des LDA als Verfahren des Topic Modelling als Möglichkeit zur Ableitung semantischer Relationen aus Lerninhalten vorgestellt und im zweiten Schritt eine Umsetzung anhand von Vorlesungsfolien für die Inferenz relevanter Begriffe umgesetzt. Dabei wurde aufgezeigt, dass das LDA-Verfahren nicht nur für eine Inferenz von relevanten Begriffen auf Webseiten genutzt werden kann, sondern auch eine Analyse der Themen von Vorlesungsfolien möglich ist. Die inferierten Themen wurden zudem in einem ersten Ansatz mittels Chorddiagramm in Relationen gebracht. Im nächsten Schritt soll nun geprüft werden, inwieweit abgeleitete Themen und Themenstrukturen mit den Intentionen von Dozierenden und Lehrveranstaltungen korrelieren. Darüber hinaus soll geprüft werden, inwieweit Vortragsfolien über Meta-Informationen verfügen, die zur Schärfung der Inferenz über Dirichlet-Priors und Textfilter genutzt werden können.

Für die weitere Arbeit ergeben sich verschiedene Fokussierungsschwerpunkte und Fragestellungen. Zum Einen gilt es, die Überlegungen in den Lehr-Lernkontext einzubetten. Mit der Betrachtung der Korrelation von Themenstruktur, Fachwissen und Intention des Dozierenden kann eine Aufmerksamkeit für die inhaltliche Strukturierung von Lehrinhalten angeregt werden. Dazu braucht es auch einen Rückbezug auf bereits vorhandene Studien mit analogen (Visualisierungs-) Methoden, die die Anreicherung konzeptuellen Wissens untersucht haben. Der Gewinn und daraus resultierende Effekte für den Lernprozess sind weitere Fragestellungen. Auf der anderen Seite ist ein empirisches Design, das für die identifizierten Fragestellungen herangezogen werden kann, zu entwickeln. Aus gegenwärtiger Sicht ist eine Interventionsstudie realisierbar, die analoge Methoden und die vorgeschlagene digitale Untersuchung konzeptuellen Wissens miteinander verbindet. Auf diese Weise sollen Differenzen zwischen dem maschinellen Abstraktionsweg und den analogen Strukturierungsleistungen Lernender und Dozierender sichtbar in sichtbar und validiert werden. Dies ermöglicht in weiteren Arbeiten eine Fokussierung auf die Identifizierung von Schwellen im Lernprozess, mit dem Potential die Fragestellung geeigneter Unterstützungen im Lehr-Lern-Prozess zu untersuchen.

Literaturverzeichnis

- [BL06] Blei, David M.; Lafferty, John D.: Dynamic topic models. In: Proceedings of the 23rd international conference on Machine learning. S. 113–120, 2006.
- [BNJ03] Blei, David M.; Ng, Andrew Y.; Jordan, Michael I.: Latent dirichlet allocation. the Journal of machine Learning research, 3:993–1022, 2003.
- [Da11] Darling, W. M.: A Theoretical and Practical Implementation Tutorial on Topic Modeling and Gibbs Sampling. In: In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. 2011.
- [GS04] Griffiths, T. L.; Steyvers, M.: Finding scientific topics. Proceedings of the National Academy of Sciences, 101(Supplement 1):5228–5235, 2004.

- [Ho99] Hofmann, Thomas: Probabilistic Latent Semantic Indexing. In: Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '99, ACM, New York, NY, USA, S. 50–57, 1999.
- [Ku10] Kurtz, Thomas: Der Kompetenzbegriff in der Soziologie. In (Kurtz, Thomas; Pfadenhauer, Michaela, Hrsg.): Soziologie der Kompetenz SE - 1, S. 7–25. VS Verlag für Sozialwissenschaften, Wiesbaden, 2010.
- [La12] Laurillard, Diana: Teaching as a design science: Building pedagogical patterns for learning and technology. Routledge, New York, London, 2012.
- [ld15] lda developers: LDA 1.0.3: Python Package Index: <https://perma.cc/ZJ8W-JR33>. 2015.
- [LLG08] Lee, Hyeon Woo; Lim, Kyu Yon; Grabowski, Barbara L.: Generative Learning: Principles and Implications for Making Meaning. Handbook of research on educational communications and technology, 3:111–124, 2008.
- [MMM08] Marius, Lars; Moore, Garshol; Moore, Graham: ISO 13250-2: Topic Maps — Data Model: <http://perma.cc/MPT4-MVLM>. 2008.
- [Ni15a] Nicolay, Robin; Schwennigcke, Bastian; Sahl, Sarah; Martens, Alke: Learner-Content-Interface as an Approach for self-reliant and student-centered Learning. In: Proceedings of the IADIS International Conference Cognition and Exploratory Learning in Digital Age, S. 327–331. 2015.
- [Ni15b] Nicolay, Robin; Schwennigcke, Bastian; Vetterick, Jonas; Sucharowski, Wolfgang; H. Cap, Clemens: InterLect - Lecture Content Interface. In: 7th International Conference on Computer Supported Education. S. 269–276, 2015.
- [OR08] Oelkers, Jürgen; Reusser, Kurt: Qualität entwickeln – Standards sichern – mit Differenz umgehen (Expertise im Auftrag des BMBF). Bildungsforschung, 27. BMBF, Berlin, 2008.
- [PK13] Pirolli, Peter; Kairam, Sanjay: A knowledge-tracing model of learning from a social tagging system. User Modeling and User-Adapted Interaction, 23(2-3):139–168, 2013.
- [Po15] Porter et al.: Natural Language Toolkit - NLTK 3.0 documentation: Stopwords Corpus: <https://perma.cc/F8L8-KC96>. 2015.
- [PR08] Parlament, Europäisches; Rat, Europäischer: Empfehlungen des Europäischen Parlaments und des Rates zur Einrichtung des Europäischen Qualifikationsrahmens für lebenslanges Lernen. apr 2008.
- [WMM09] Wallach, Hanna M.; Mimno, David M.; McCallum, Andrew: Rethinking LDA: Why priors matter. In: Advances in neural information processing systems. S. 1973–1981, 2009.
- [Yo15] Yoshiki Shibukawa: Snowballstemmer 1.2.1: <https://perma.cc/XPJ6-JNMF>. 2015.
- [Zh08] Zhao, Weizhong: Best Practices in Building Topic Models with LDA for Mining Regulatory Textual Documents: NCTR CTP Working Group: <https://perma.cc/8NGH-GYAL>. 2008.