Hans Pongratz, Reinhard Keil (Hrsg.): Die 13. E-Learning Fachtagung Informatik,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2015 219

OSAI: Ein Tool zur Themenanalyse in Lernraum-Foren

Oliver Blum[1], Sascha Feldmann[2], Agathe Merceron[3] und Ilse Schmiedecke[4]

Abstract: Die Foren von Online-Lernräumen konzentrieren sich typischerweise um Themen, die im Kurs besondere Aufmerksamkeit erfordern. Auch nach Abschluss der Diskussion finden Studierende wie Lehrende hier wertvolle Informationen. In diesem Beitrag wird ein Tool für die Themenanalyse in Lernraum-Foren vorgestellt. Das modulare Tool bietet eine Gesamtlösung an, die für verschiedene Lernraum-Systeme einsetzbar ist und in andere Analyse-Tools integriert werden kann. Die bisher verfügbaren Module leisten die Anonymisierung, die Eliminierung von Quelltext und die Extraktion und Visualisierung der Themen aus Foren.

Keywords: Foren, Themenanalyse, Tool, Anonymisierung, Quellcodeeliminierung, Themenextraktion, Visualisierung.

1 Einleitung

Eine besondere Stärke von Online-Lernraum-Systemen sind die Diskussionsforen. Gerade in der Online-Lehre ist diese asynchrone Kommunikationsform unabdingbar. Sie bringt den zusätzlichen Vorteil mit sich, dass die Diskussionen auch nach ihrem Abschluss verfügbar bleiben. Studierende können so auch im Nachhinein von den unterschiedlichen Erklärungen ihrer Kommilitonen profitieren, während Lehrende z.B. erkennen können, wo zusätzlicher Erläuterungsbedarf besteht. Online-Foren werden umfangreich analysiert, siehe [MEF13] für einen kurzen Überblick.

Dass das Potenzial vor allem im Nachhinein praktisch nicht genutzt wird, erkennt man z.B. daran, dass die Suchfunktion an der BHT hochschulweit nur ca. 1500x im Semester genutzt wird. Das liegt möglicherweise am Erschließungsaufwand; vor allem für Lehrende sind primär die Diskussionsthemen interessant. Da diese häufig vom Threadtitel abweichen, muss jeder Thread einzeln geöffnet und überflogen werden, um herauszufinden, worum es darin geht. Eine automatische Extraktion der Themen, wie sie in diesem Papier vorgeschlagen wird, kann die nachträgliche Forennutzung deutlich praktikabler machen und auch semesterübergreifende Untersuchungen unterstützen.

[1] Beuth Hochschule für Technik, Fachbereich Medieninformatik, Luxemburgerstrasse 19, 13353 Berlin, oliverblum1987@googlemail.com
[2] Beuth Hochschule für Technik, Fachbereich Medieninformatik, Luxemburgerstrasse 19, 13353 Berlin, sascha.feldmann@gmx.de
[3] Beuth Hochschule für Technik, Fachbereich Medieninformatik, Luxemburgerstrasse 19, 13353 Berlin, merceron@beuth-hochschule.de
[4] Beuth Hochschule für Technik, Fachbereich Medieninformatik, Luxemburgerstrasse 19, 13353 Berlin, schmiedecke@beuth-hochschule.de

In dieser Arbeit wird das modulare Werkzeug OSAI vorgestellt, das die Themenextraktion in eine Gesamtlösung einbettet, in der die Forenbeiträge vorab gezielt aufbereitet werden. Das Verfahren kann auf verschiedene Lernraum-Systeme angewendet und auch in Analyse-Tools wie z.B. LeMo [BEF13] integriert werden.

2 Die Module des Tools OSAI

Zu OSAI gehören die Module Anonymisierung, Quelltexteliminierung, Themenextraktion und Visualisierung, sowie ein Präprozess zur Bereinigung der Beiträge von URLs und Email-Adressen. Alle Module sind separat ausführbar, und die meisten sind bereits fest in das Tool integriert.

Anonymisierung ist erforderlich, weil Foren in der Online-Lehre nicht öffentlich sind und daher gesonderte Datenschutzrichtlinien für die Speicherung der extrahierten Daten gelten. Getestet wurde OSAI auf Foren aus der Informatik-Lehre. Hier besteht die Gefahr, dass Quellcodezitate die Themenanalyse "irreführen". Daher enthält OSAI Komponenten, um diese vor der semantischen Analyse auszublenden. Für die Themenextraktion wurden zwei unterschiedliche Verfahren realisiert. Und auch für die Ergebnispräsentation gibt es bisher zwei verschiedene Visualisierungsansätze.

2.1 Anonymisierung

Die Anonymisierung erfolgt zunächst durch Klassifikation und Entfernung von Personen-Entitäten durch Named Entity Recognition (NER). Dabei kommt die Bibliothek „Stanford NLP" zum Einsatz, deren Trainingsdaten für die deutsche Sprache durch Analyse des "Huge German Corpus" (HGC) entstanden [CoN03].

Hallo, ich habe angefangen mit ein Adreßbuch zu basteln. Dabei habe ich ein Problem mit einer Methode und der Verwendung in anderen Klassen. Ich habe in einer Klasse ein Attriut das vom Typ einer anderen Klasse ist. In einer Methode zu Ausgabe der UID wird die Methode der Klasse aufgerufen dabei gibt es allerdings eine Nullpointer exception. Ich sitze da jetzt schon länger dran und finde den Fehler nicht. Vielleicht hat ja einer von euch eine Idee. Gruß und Dank import java. util. Date; import java. text.*; natPerson import java. util. UUID; UID

/* Hallo Author4, könntest Du noch die Superklassen mitschicken? Dann könnte ich das auch mal testweise laufen lassen. Aber zumindest Deine Klasse Person fehlt mir. Zunächst eine Kleinigkeit: Klassennamen per Konvention immer mit einem Großbuchstaben anfangen, also besser NatPerson. Zum eigentlichen Thema: ich kann es wie gesagt so nicht testen, aber versuch doch mal ein paar System. out. println("Methodenname") in Deiner UID-Klasse einzubauen. Ich könnte mir denken, dass da nicht die toString()-Methode aufgerufen wird, die Du eigentlich aufrufen willst oder, dass es Probleme beim Überschreiben von UID gibt (es gibt eine Klasse UID im Java-API) - falls das unbeabsichtigt ist, würde ich einfach mal Deine Klasse umbenennen. Dann würde ich erst einmal die eleganten Einzeiler vorübergehend für die Fehlersuche zu Mehrzeilern machen, z. B.: */ String // Viele Grüße

Hallo, hier ist die Person Klasse, da passiert aber icht viel. Person Danke schonmal und einen schönen Tag

Abb. 1: Anfang eines Threads nach Quellcode-Eliminierung

Trotz Verwendung der deutschen und englischen Stanford-NER-Tagger ließ sich so nur ein Anonymisierungsgrad von 50% erreichen. Die in Foren vielfach verwendeten Namenskürzel und -akronyme sowie viele internationale Studierendennamen blieben unerkannt. Der Namenskorpus wurde daher durch eine forenspezifische Lernliste ergänzt, die durch zwei Heuristiken gebildet wird: Eine Grußformel-Mustererkennung, steigert den Anonymisierungsgrad auf 70%; Die Hinzunahme einer Teilnehmer-Namensliste liefert eine 99%-ige Anonymisierung, die die Datenschutzauflage erfüllt.

2.2 Quellcode-Eliminierung

Die Quellcode-Eliminierung orientiert sich an der Programmiersprache Java. Sie muss differenziert erfolgen, da einzelne Programmausdrücke wie „toString“ Themen sein können. Sie erfolgt durch eine Kombination der Filterung der Klammerführung im Quelltext und das Filtern von Sprachgerüsten über reguläre Ausdrücke. Da im Allgemeinen Quellcode per copy/paste eingefügt wird sind die geschweiften Klammern nicht immer korrekt gepaart. Der Algorithmus eliminiert rekursiv den maximalen Teil zwischen der letzten offenen und der richtig gepaarten geschlossenen Klammer. Programmteile, die sich außerhalb der Klammern befinden, werden durch reguläre Ausdrücke separat eliminiert. Auf diese Weise werden einzelne Programmierkonstrukte im Text wie toString() in *„Ich habe ein wenig mit der toString() Methode gespielt, kam aber zu keinem Ergebnis.”* nicht entfernt. Die Quellcode-Filterung ist konservativ: Im Zweifel wird Quellcode belassen statt Inhalt aus der Diskussion zu eliminieren. Experimente weisen eine Quellcode-Eliminierung von 88% auf. Abb. 1 zeigt die drei ersten Nachrichten eines Threads nach der Quellcode-Eliminierung.

Durch Anpassung oder Ergänzung der regulären Ausdrücke lassen sich auch andere Programmiersprachen berücksichtigen. Das Modul kann grundsätzlich auch durch die Implementierung eines anderen Verfahrens ersetzt werden, wie z.B. das Verfahren aus [KK13] wenn verfügbar.

2.3 Direkte Themenextraktion

Angelehnt an [She12] wurde die Themenextraktion in mehreren intuitiv zu verstehenden Schritten implementiert. Themen werden als häufig benutzte Schlüsselwörter verstanden. Hierfür kommen nur Substantive, Adjektive und Verben in Betracht. Zuerst werden die Wortarten mit dem Lexicalized PCFG Parser der Bibliothek CoreNLP ermittelt. Damit Wörter wie Programm und Programme das gleiche Thema repräsentieren, setzt die Anwendung einen Algorithmus zum Stemmen von Wörtern für die deutsche Sprache ein [ger15]. Schließlich werden die häufigsten Wörter mit Hilfe der Bibliothek Apache Lucene extrahiert. Damit wichtige Begriffe nicht durch allzu häufige verdrängt werden, wird die Häufigkeit n durch die Formel log(n) + 1 gedämpft. Für eine bessere Lesbarkeit werden die gestemmten Wörter wieder in eine Ursprungsform umgewandelt. So wird die folgende Liste für den Thread der Abbildung 1 zurückgegeben: Klassen, Methode, Per-

son, UID, gibt, abstrakten, anderen, Fehler, Gruß, Anwendung.

Da die Implementierung modular ist, kann auch das Verfahren der Themenextraktion durch ein anderes ersetzt werden, z. B. Latent Dirichlet Allocation (LDA) [BNJ03].

2.4 Themenextraktion mit Hilfe semantischer Netze

Als alternative Methode wurde eine Themenextraktion mithilfe semantischer Netze realisiert, um übergeordnete Konzepte als Themen finden zu können. Mit einer strukturellen Basisontologie, in diesem Fall der Struktur der exportierten Forendaten, und einer oder mehreren NER- und POS-Taggern auf Basis geeigneter Korpora (z.B. HGC für die deutsche Sprache) lässt sich automatisch ein semantisches Netz erzeugen, in dem jeder Knoten einem relevanten Begriff entspricht und der Vernetzungsgrad die relative Relevanz des Begriffs spiegelt. OSAI verwendet Apache Jena, um das Themennetz in der Sprache Resource Description Framework (RDF) zu speichern.

Mit RDF-Netze können dynamische Anfragen in der Sprache SPARQL verarbeitet und die Ergebnis-Graphen visualisiert werden. Das Verfahren ist als integriertes Tool für deutsche und englische Texte vollständig automatisiert. Anzugeben sind die bereinigten Forendaten und die zu verwendeten Tagging-Verfahren, hier der TopicZoom-Webservice [TZ12], POS- und NER-Tagging. Eine partielle Visualisierung des Threads Abb. 1 als Netz findet sich in Abb. 2.

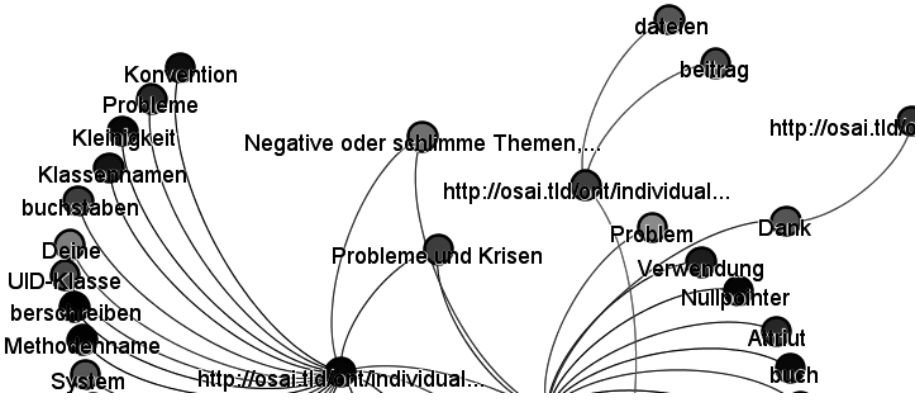

Abb. 2: Ausschnitt aus dem mit Gephi erstellten Themen-Netz

In ersten Experimenten wurden übergeordnete Konzepte kaum gefunden. Es wird untersucht, in wie weit spezifischere Korpora durch Synonym- und Begriffsraumerkennung eine bessere Kategorisierung bringen.

2.5 Visualisierung

Für die schnelle Erschließung der Ergebnisse stehen zur Zeit zwei Visualisierungen zur Verfügung. Für die Semantischen Netze liefert das Werkzeug Gephi [BAS09] eine Visualisierung als Netzwerk, in dem die Knoten die Themen und die Posts des Threads repräsentieren. So kann durch den Themen-Graph navigiert werden und untersucht werden, in welchen Posts Themen vorkommen, siehe Abb. 2.

Abb. 3: Themen eines Threads als Tag Cloud

Die direkte Themenextraktion wird als Tag Cloud dargestellt: die numerische Gewichtung jedes Wortes wird über Größe, Position und Transparenzwert visualisiert. Abb. 3 zeigt die Tag Cloud aus den extrahierten Themen eines einzelnen Threads. Die wichtigsten Wörter „Klassen“, „Methoden“, „Person“ und die zusätzlichen Wörter „UID“ und „abstrakten“ geben eine Zusammenfassung des darin diskutierten Problemschwerpunktes, die einem Vergleich zwischen Jahrgängen ermöglicht.

3 Fazit

Die Stärke des vorgestellten Werkzeugs liegt darin, dass es als geschlossene und trotzdem integrierbare Experimentierplattform konzipiert ist. Es erlaubt Foren-Importe in verschiedenen Formaten, eine freie Kombination der vorhandenen Module und eine

Ergänzung weiterer Module für die Analyse. Die Ergebnisse erster Experimente sind ermutigend, auch wenn die jetzige Themenextraktion mit semantischen Netzen nicht wie gehofft die übergeordneten Konzepte liefern konnte. Als nächster Schritt soll die Verwendung fachspezifischer Korpora für die Analyse mit semantischen Netzen untersucht werden. Es wird eine umfassende Evaluierung des Tools mit Lehrenden und Lernenden folgen. Insgesamt hoffen wir, dass ein Tool zur Forenerschließung entsteht, das Lehrende wie Lernende darin unterstützt, die dort gesammelte Information für sich zu nutzen.

Literaturverzeichnis

[BEF13] Beuster, L.; Elkina, M.; Fortenbacher, A.; Kappe, L.; Merceron, A.; Pursian, A.; Schwarzrock, S.; Wenzlaff, B.: Learning Analytics und Visualisierung mit dem LeMo-Tool. In Proceedings of DeLFI 2013 , Bremen (Germany), September 09-11, Gesellschaft für Informatik Publisher, p. 245-250, 2013.

[BHJ09] Bastian, M.; Heymann S.; Jacomy M.: Gephi: an open source software for exploring and manipulating networks. Proc. Int. AAAI Conference on Weblogs and Social Media, 2009.

[BNJ03] Blei, D.M.; Ng, A.Y.; Jordan, M.I.. Latent Dirichlet Allocation. Journal of Machine Learning Research 3, 993-1022, 2003.

[CoN03] CoNLL. Language-Independent Named Entity Recognition (II). http://www.cnts.ua.ac.be/conll2003/ner/. Letzter Zugriff am: 02-26-2015.

[FT09] Fujita, N.; Teplovs, C.: Automating the analysis of collaborative discourse: Identifying idea clusters. In C. O'Malley, D. Suthers, P. Reimann & A. Dimitracopoulou (Eds.), Proceedings of the Int. Conf. CSCL 2009, Rhodes: ISLS, pp 162-164, 2009.

[ger15] https://code.google.com/p/mauiindexer/source/browse/Maui1.2/src/maui/stemmers/GermanStemmer.java?r=54 letzter Zugriff am: 28.02.2015.

[KK13] Khayyamian M.; Kim, J.: An intelligent web-based Interface for programming content in forums. In proceedings of the companion publication of the international conference on intelligent user interface companion, IUI' 13, ACM, New York, NY, USA, pp. 67-68, 2013.

[MEF13] Merceron, A.; Elkina, M.; Fortenbacher, A.: Learning Analytics und Foren In Proceedings of the Pre-Conference Workshops der 11. e-Learning Fachtagung Informatik – DeLFI 2013, A. Breiter, D. Meier, C. Rensing (Eds), Bremen (Germany), 8.09.2013, Logos Verlag Berlin, p. 139-144, 2013.

[RTZ11] Rabbany, R. K.; Takaffoli, M.; Za¨iane, O.R: Analyzing Participation of Students in Online Courses Using Social Network Analysis Techniques. In (Pechenizkiy, M., Calders, T., Conati, C., Ventura, V., Romero, C., Stamper, J. Hrsg.): Proceedings of the 4th International Conference on Educational Data Mining. (Eindhoven, Netherlands, July 6-8). EDM'11. http://www.educationaldatamining.org/EDM2011/, 253-257, 2011.

[RG14] Ramesh, A.; Goldwasser, D.; Huang, B.; H. Daume III H.; L. Getoor, L.: Understanding MOOC Discussion Forums using Seeded LDA. In Proc. of 9th Workshop on Innovative Use of NLP for Building Educational Applications, pages 28–33. ACL, 2014.

[She12] Sherin, B.: Proceedings of the 2nd International Conference on Learning Analytics and Knowledge. (Eindhoven, Netherlands, July 6-8). LAK'12.

[TZ12] http://www.topiczoom.de/wp-content/uploads/2012/01/Whitepaper-Navigation.pdf letzter Zugriff am: 05.06.2015.