Demonstration von thematischen Frames im TopicExplorer-System

Extended Abstract

Alexander Hinneburg[1], Frank Rosner[1], Stefan Peßler[2] und Christian Oberländer[2]

[1]Informatik, Martin-Luther-Universität Halle-Wittenberg
[2]Japanologie, Martin-Luther-Universität Halle-Wittenberg
hinneburg@informatik.uni-halle.de
frank.rosner@student.uni-halle.de
stefan.pessler@japanologie.uni-halle.de
christian.oberlaender@japanologie.uni-halle.de

Themenmodelle bieten sich an, die Inhalte großer Dokumentensammlungen zu erforschen. Thematische Wortlisten präsentieren typische Inhalte. Diese Themen werden automatisch gelernt, ohne das Dokumente manuell annotiert werden müssen. Während des Lernens eines Themenmodells werden die Wörter der Dokumente Themen zugeordnet. Dabei werden zwei gegenläufige Ziele verfolgt: erstens, einem Thema sollen so wenig wie möglich verschiedene Wörter zugeordnet werden und zweitens, ein Dokument soll so wenig wie möglich verschiedene Themen enthalten [2]. Die unüberwachten Lernalgorithmen finden Kompromisslösungen für diese Aufgabenstellung, welche im Fall von Variationsinferenz zu lokalen Extrema der freien Energiefunktion des Modells und im Fall von Gibbs-Samplern zu wahrscheinlichen Zuständen einer Markov-Kette korrespondieren. In keinem Fall garantieren die Algorithmen, dass die berechneten Themen gut durch Menschen interpretierbar sind.

Es ist state-of-the-art die Themen, welche mathematisch gesehen Wahrscheinlichkeitsverteilungen über Wörtern sind, durch die wahrscheinlichsten Wörtern zu repräsentieren. Die Interpretation dieser Wortlisten kann jedoch eine schwierige Aufgabe für den Anwender sein. Eine erfolgreiche Interpretation hängt vom Hintergrundwissen der Person und der Vertrautheit mit dem genutzten Vokabular ab. Zwei wesentliche Probleme können die Interpretation eines Thema beeinträchtigen. Erstens, thematische Wortlisten können komplett aus Substantiven bestehen, deren Beziehungen zueinander mehrdeutig sein können. Ein Beispiel ist ein Thema, dass durch eine Liste von Ländernamen repräsentiert wird. Trotz dessen, dass alle Länder in einer eng umgrenzten Region liegen können, gibt es immer noch mehrere verschiedene Interpretationen, die zu einer solchen Liste passen würden. Deshalb ist sie nicht gut interpretierbar. Ein zweiter Grund kann darin liegen, dass die präsentierten Wörter dem Anwender als unzusammenhängend erscheinen. Dies kann an Wörtern liegen, die der Anwender nicht kennt.

Es ist eine offene Frage, wie durch Themenmodelle berechnete Themen so repräsentiert werden können, dass sie klar und eindeutig durch Menschen interpretiert werden

 In: T. Seidl, M. Hassani, C. Beecks (Eds.): Proceedings of the LWA 2014 Workshops: KDML, IR, FGWM, Aachen, Germany, 8-10 September 2014, published at http://ceur-ws.org

können. Jüngste Forschungen zur Messung von Kohärenz von Themen zeigten, dass sie besser durch Menschen interpretiert werden können, wenn Paare von Wörtern des Themas oft in Dokumenten nahe beieinander stehen [3]. Deshalb kann eine Themenrepräsentation leichter interpretierbar sein, die Paare von wahrscheinlichen Wörtern zeigt, die oft in Dokumenten nahe beieinander stehen. Dieser Ansatz löst jedoch nicht das Problem von Substantivlisten. Eine Schlüsselbeobachtung ist, das Verben in Themenverteilungen oft weniger wahrscheinlich als Substantive sind, weil sie flexibler in verschiedenen Kontexten gebraucht werden können. Deshalb tauchen Verben in nach Themenwahrscheinlichkeit sortierten Wortlisten weiter hinten auf und müssen somit gesondert behandelt werden.

Wort-Kombinationen aus je einem Verb und einem Substantiv können als Basiseinheiten angesehen werden um Inhalte zu transportieren. Minski nannte in den ersten Forschungen zu künstlicher Intelligenz solche Einheiten Frames [1,4]. Deshalb stellt das Auftreten von einem Verb in der Nähe eines Substantivs in einem Dokument eine notwendige Bedingung für das Vorhandseins eines Frames dar. Ein thematischer Frame kann vorhanden sein, wenn ein Themenmodell ein Verb und ein Substantiv, die in einem Dokument nahe zusammenstehen, dem selben Thema zuweist. Unsere Demonstration zeigt anhand mehrerer Beispiele, dass Themen durch die Repräsentation mittels thematischer Frames interpretiert werden können, deren Inhalte allein durch das Zeigen von Wortlisten unklar bleiben würde. Die Entwicklung von Evaluationsmethoden für diese Aufgabe ist jedoch Gegenstand weiterer Forschungen.

Unsere Implementation von thematischen Frames ist in das TopicExplorer System (`http://topicexplorer.informatik.uni-halle.de/`) eingebettet. Wir demonstrieren die thematischen Frames mit Hilfe von verschiedenen Dokumentsammlungen in unterschiedlichen Sprachen. Die Sammlungen müssen allgemein bekannt sein, damit die Themen leicht und ohne Fachwissen zugänglich sind. Deshalb haben wir zu Demonstrationszwecken einen Teil der englischen Wikipedia sowie eine Sammlung deutscher Märchen als Dokumentsammlungen ausgewählt. Weiterhin zeigen wir als echte Anwendung des TopicExplorers und der thematischen Frames die Unterstützung von sozialwissenschaftlicher Forschung bei der Analyse von japanischen Blogs, welche die Auswirkungen der Fukushima-Katastrophe von 2011 und die soziale Verantwortung diskutieren. [Die englische Version des Artikels wird zur CIKM 2014 erscheinen.]

Danksagung

Wir danken Mattes Angelus, Benjamin Schandera und Gert Böhmer für ihre wertvollen Programmierbeiträge zur Code-Basis des TopicExplorer. Weiterhin danken wir der Klaus Tschira Stiftung für die finanzielle Unterstützung des Projektes.

Literatur

1. Allan, K.: Natural language semantics. Wiley (2001)
2. Blei, D.: Topic modeling and digital humanities. Journal of Digital Humanities 2(1) (2012)
3. Lau, J.H., Newman, D., Baldwin, T.: Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In: Proceedings of the European Chapter of the Association for Computational Linguistics (2014)
4. Minsky, M.: Frame-system theory. In: Johnson-Laird, P.N., Wason, P.C. (eds.) Thinking. pp. 355–377. Cambridge: Cambridge University Press (1977)