Neue Wahrscheinlichkeitsmodelle und Inferenz-Techniken für Kontextinformationen im World Wide Web[1]

Christoph Kling[2]

Abstract: Thema meiner Dissertation ist die Erkennung von Mustern in Web-Dokumenten mit Metadaten. Schwerpunkt der Arbeit sind sogenannte *Topic Models*. Mithilfe von Topic Models können automatisch Themen in großen Dokumentensammlungen erkannt werden. Dokumente aus dem Web sind oft mit Metadaten wie etwa Zeitstempeln versehen, die den *Kontext* beschreiben, in denen diese erstellt wurden. Diese Kontextinformationen können die Themen-Erkennung durch Topic Models verbessern oder sogar erst ermöglichen. Außerdem erlauben sie die Analyse von Zusammenhängen zwischen Kontext und Themen, etwa an welchen Orten welche Themen populär sind.

In meiner Arbeit werden neuartige Topic Models vorgestellt, die die Einbeziehung von (fast) beliebigen Kontextinformationen erlauben, messbar die Qualität der erkannten Themen verbessern und aufgrund ihrer Struktur eine effiziente Inferenz ermöglichen. Zudem sind die Parameter der Modelle interpretierbar und können auch komplexe Zusammenhänge zwischen Kontext und Themen aufdecken, die mit bisherigen Modellen nicht erkannt werden können.

1 Einleitung

Dokumente aus dem Word Wide Web sind in fast allen Fällen mit Metadaten versehen. Seien es Zeitstempel, Ortsinformationen oder Nutzernamen – fast immer finden sich Informationen über den **Kontext** in dem eine Nachricht verfasst, ein Bild aufgenommen oder ein Nutzerprofil angelegt wurde.

Eine der wichtigsten Analysemethoden für Web-Dokumente ist die Erkennung von Themen mit sogenannten **Topic Models**. Diese erkennen häufig zusammen auftretende Wörter, welche als Themen interpretiert werden können. Moderne Topic Models nutzen Kontextinformationen um die Erkennung von Themen zu verbessern oder sogar erst zu ermöglichen. Zudem erlauben sie die Analyse von Zusammenhängen zwischen Kontext und Themen – etwa die Entwicklung der Popularität von Themen über die Zeit.

Mit der Einbeziehung von Kontextinformationen in Topic Models sind mehrere Fragestellungen verbunden: **(i) Generalisierbarkeit.** Wie kann ein Topic Model beliebige Kontextinformationen berücksichtigen (etwa lineare Zeitinformationen, GPS-Koordinaten auf der Erde oder diskrete Informationen – wie das Geschlecht des Verfassers einer Nachricht), ohne dass das Modell dafür speziell angepasst werden muss? **(ii) Flexibilität.** Zusammenhänge zwischen Kontextinformationen und Themen sind oft komplex – Normalverteilungen oder lineare Zusammenhänge sind in der Regel nicht ausreichend, um die

[1] Englischer Titel der Dissertation: "Probabilistic Models for Context in Social Media: Novel Approaches and Inference Schemes"

[2] GESIS – Leibniz-Institut für Sozialwissenschaften, Christoph.Kling@gesis.org

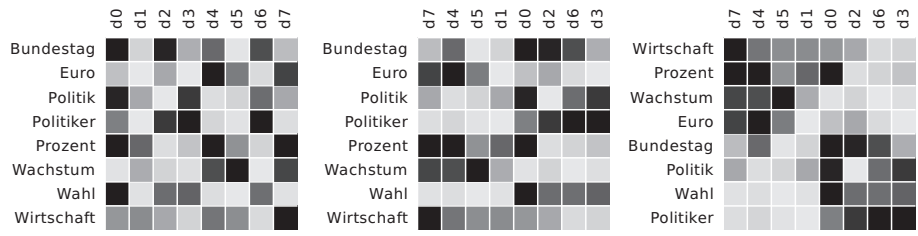


Abb. 1: **Schema der Funktionsweise von Topic Models.** Abb. (a) zeigt eine Dokument-Wort-Matrix (DWM): Zeilen entsprechen Wörtern, Spalten Dokumenten. Zellen zeigen die Häufigkeit der Wörter in Dokumenten, wobei dunklere Felder höhere Worthäufigkeiten bedeuten. Abb. (b) zeigt die Matrix nach Gruppierung ähnlicher Dokument-Wort-Spalten, Abb. (c) nach anschließender Gruppierung ähnlicher Wort-Dokument-Zeilen. Die gefundenen Muster werden von Topic Models genutzt.

Beziehungen zu modellieren. Wie können komplexe Zusammenhänge modelliert werden, ohne etwa unrealistische Unabhängigkeitsannahmen einzuführen? **(iii) Interpretierbarkeit.** Wie kann sichergestellt werden, dass die Modellparameter interpretierbar sind und dass die Funktionsweise der Modelle für Menschen nachvollziehbar und auf Plausibilität überprüfbar ist? **(iv) Skalierbarkeit.** Dokumenten-Sammlungen aus sozialen Medien können erhebliche Ausmaße annehmen. Wie kann die Modellstruktur der Topic Models so gestaltet werden, dass eine effiziente Inferenz möglich ist? **(v) Benutzerfreundlichkeit.** Wie kann die Implementierung gestaltet werden, so dass auch Nicht-Informatiker komplexe Dokumentensammlungen verarbeiten können?

In meiner Arbeit zeige ich, wie Mischungen von Dirichlet-Prozessen für die Modellierung von (fast) beliebigen Kontext-Informationen genutzt werden können. Diese bieten als erste Modelle sowohl eine hohe Generalisierbarkeit und Flexibilität als auch eine direkte Interpretierbarkeit der Modellparameter. Zusätzlich erlaubt die Modellstruktur eine effiziente Inferenz, die in meiner Arbeit vorgestellt wird.

Im Folgenden wird zunächst die grundlegende Funktionsweise von Topic Models und Hierarchischen Dirichlet-Prozessen (HDP) erklärt. Anschließend werden Beiträge aus meiner Dissertation vorgestellt: **Die neuartige Modellierung von Kontextinformationen** in probabilistischen Modellen mittels Kontext-Clustern und nachbarschafts-basiertem Informationsaustausch. Und **eine neue Generalisierung von HDPs, der Hierarchische Multi-Dirichlet Prozess (HMDP)**, der den Informationsaustausch zwischen Clustern ermöglicht. Schließlich werden die Modellierung von Kontexten in HMDP Topic Models und Einsatzmöglichkeiten an Beispielen beschrieben.

2 Verwandte Arbeiten aus dem Berech Topic Modelling

Die Grundidee hinter Topic Models ist die Erkennung von häufig gemeinsam vorkommenden Worten. Beispielsweise könnte eine Analyse von vielen Zeitungsartikeln ergeben, dass die Wortmenge {*Bundestag, Politik, Wahl, Politiker*} und die Wortmenge {*Euro, Prozent,*

Wachstum, Wirtschaft} jeweils häufig gemeinsam in Artikeln vorkommen. Diese Mengen an gemeinsam auftretenden Wörtern werden *Topics* genannt.

Um diese Topics zu erkennen, werden Dokumente meist als ungeordnete Mengen von Wörtern betrachtet. Ein Topic Model versucht nun anhand der Wortmengen Dokumente, die ähnliche Worte enthalten sowie Wörter, die ähnliche Häufigkeiten in den Dokumenten aufweisen, zu denselben Topics zuzuweisen. Abbildung 1 zeigt die Repräsentation einer Dokumentensammlung als Dokument-Wort-Matrix (DWM) und wie eine Neugruppierung die Erkennung von Topics erlaubt.

Frühe Topic Models basierten tatsächlich auf der direkten Mustererkennung in Dokument-Wort-Matrizen.Allerdings sind die Ergebnisse dieser Methoden schwer interpretierbar und erlauben keine Erweiterung des Modells – etwa für Kontextinformationen. Mittlerweile werden hauptsächlich probabilistische Modelle zur Themenerkennung eingesetzt. Diese erkennen für jedes Dokument eine Wahrscheinlichkeitsverteilung über Topics (mit welcher Wahrscheinlichkeit enthält ein Dokument ein gegebenes Topic?) und für jedes Topic eine Wahrscheinlichkeitsverteilung über alle Wörter (mit welcher Wahrscheinlichkeit kommt ein Wort in einem Dokument zu einem Topic vor?). Dokumente können zu mehreren Topics gehören, und eine Wort kann in mehreren Topics eine hohe Wahrscheinlichkeit besitzen (etwa bei Mehrdeutigkeiten).

2.1 Latent Dirichlet Allocation

Das populärste Topic Model ist *Latent Dirichlet Allocation* (LDA) [BNJ03], ein Modell bei dem die Dokument-Topic-Verteilung und die Topic-Wort-Verteilung einer symmetrischen Dirichlet-Verteilung (alle Topics / Wörter sind a-priori gleich wahrscheinlich) mit niedrigen Parametern folgen. Niedrige Dirichlet-Parameter bedeuten, dass Dokumente a-priori jeweils nur wenige Topics enthalten sollen und dass Topics nur wenigen Wörtern eine hohe Wahrscheinlichkeit zuordnen sollen. Praktisch bedeutet das, dass die Trennschärfe bei der Topic-Erkennung besser wird und Mehrdeutigkeiten besser erkannt werden. Das probabilistische Modell von LDA ist wie folgt:

1) Jedes Topic $k \in K$ wird durch eine multinomiale Topic-Wort-Verteilung ϕ_k repräsentiert, welche aus einer symmetrischen Dirichlet-Verteilung mit Parameter β gezogen wird:
$\phi_k \sim Dir(\beta)$

2) Jedes Dokument $m \in M$ wird durch eine multinomiale Dokument-Topic-Verteilung θ_m repräsentiert. Jede Dokument-Wort-Verteilung θ_m wird aus einer symmetrischen Dirichlet-Verteilung mit Parameter α gezogen:
$\theta_m \sim Dir(\alpha)$

3) Für jedes Wort $w_{mi},\ i \in N_m$ in Dokument m (Länge N_m) wird ein Topic-Index z_{mi} aus θ_m gezogen und anschließend das Wort w_{mi} aus dem Topic mit Index z_{mi} gezogen:
$z_{mi} \sim \theta_m; \qquad w_{mi} \sim \phi_{z_{mi}}$

Die Topics können visualisiert werden, indem pro Topic ϕ_k die Wörter mit der höchsten Wahrscheinlichkeit in einer Tabelle dargestellt werden (siehe Tab. 1 auf Seite 8).

2.2 Hierarchische Dirichlet-Prozesse

Eine Einschränkung von LDA ist, dass die Anzahl der Topics K festgelegt werden muss, bevor die Topics gelernt werden. Betrachtet man das Beispiel aus Abb. 1(c), so sieht man, dass in dem Datensatz zwei Topics sinnvoll erkannt werden können. Würde man LDA mit $K = 3$ auf diesen Daten lernen, so würde ein "überflüssiges" Topic erkannt, das kaum zur Erklärung der Dokumente beiträgt.

Um die Anzahl der Topics während der Inferenz zu erkennen, werden Hierarchische Dirichlet-Prozesse (HDP) eingesetzt [Te06]. Dirichlet-Prozesse (DP) sind stochastische Prozesse, deren Realisierungen Verteilungen über Wahrscheinlichkeitsverteilungen sind. Im Fall von Topic Models ergeben Dirichlet-Prozesse Verteilungen über unendlich viele Topics (welche wiederum diskrete Verteilungen über Wörter sind). Formal:

$$G_0 \sim \mathrm{DP}(\gamma, H); \qquad \gamma \in \mathbb{R}^{>0}; \qquad H = Dir(\beta) \tag{1}$$

wobei G_0 die Verteilung über unendlich viele Topics ist, H eine Dirichlet-Verteilung über alle möglichen Topics und γ ein Parameter, der die Ungleichverteilung der Wahrscheinlichkeiten über die Topics beeinflusst. Teilt man die unendlich vielen Topics in Partitionen $A_1, \ldots, A_N$, so folgt die Verteilung über die Topics in den Partitionen einer Dirichlet-Verteilung – eine Eigenschaft, die namensgebend für den Prozess ist:

$$(G_0(A_1), \ldots, G_0(A_N)) \sim Dir(\gamma H(A_1), \ldots, \gamma H(A_N)) \tag{2}$$

Der Dirichlet-Prozess führt typischerweise zu sehr ungleich verteilten Wahrscheinlichkeiten: Wenige Topics bekommen viel Gewicht. Die Intuition hinter dieser Eigenschaft kann durch die Metapher eines Chinese Restaurant-Prozesses dargestellt werden, der in Abb. 2 schematisch abgebildet und beschrieben ist.

Würde man für jedes Dokument Topics aus dem gerade beschriebenen DP ziehen, so würde jedes Dokument eigene Topics verwenden. Um gemeinsame Topics von Dokumen-

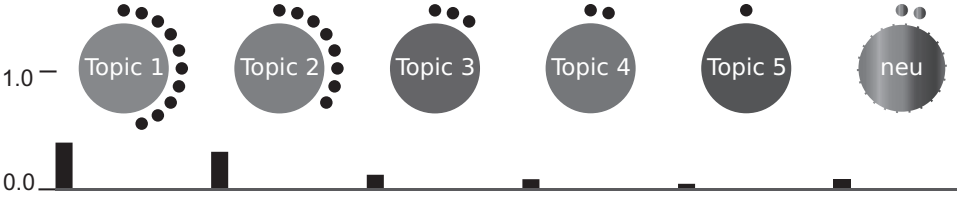


Abb. 2: **Chinese Restaurant-Prozess (CRP).** Der CRP ist eine Metapher für den Dirichlet-Prozess (DP) mit Parametern γ und H. In ein chinesisches Restaurant mit unendlich vielen Tischen kommen Kunden. Jeder neue Kunde setzt sich an einen Tisch, mit einer Wahrscheinlichkeit proportional zu der Anzahl der Kunden, die bereits am jeweiligen Tisch sitzen; Alternativ setzt er sich mit einer Wahrscheinlichkeit proportional zu dem Skalierungs-Parameter γ (hier $\gamma = 2$) an einen neuen Tisch. Jeder Tisch entspricht einer Ziehung aus der Basisverteilung H. Im Fall von Topic Models wird ein neues Topic gezogen. In der Abbildung haben bereits 24 Kunden das Restaurant betreten. Wird dieser Vorgang unendlich oft wiederholt, folgen die Wahrscheinlichkeiten der Tischwahl für neue Kunden (hier dargestellt durch Balken) der Verteilung eines Dirichlet-Prozesses.

ten zu finden, wird eine Ziehung des gerade beschriebenen Dirichlet-Prozesses als Basismaß für einen darunterliegenden, dokument-spezifischen Dirichlet-Prozess verwendet. Jedes Dokument m zieht also für jedes seiner Wörter w_{mi} ein Topic ϕ_{mi} (eine Multinomialverteilung) aus einem eigenen Dirichlet-Prozess mit Skalierungs-Parameter α_0:

$$G_m \sim \mathrm{DP}(\alpha_0, G_0); \qquad \phi_{mi} \sim G_m; \qquad w_{mi} \sim \phi_{mi} \tag{3}$$

Da das Basismaß G_0 eines jeden DP viel Gewicht auf wenige Topics verteilt, wiederholt sich die Wahl des Topics ϕ_{mi}, d.h. unterschiedliche Dokumente verwenden (mit sehr hoher Wahrscheinlichkeit) gleiche Topics. Die optimale Anzahl an global verfügbaren Topics unter den Modellannahmen wird dann während der Parameter-Inferenz herausgefunden.

2.3 Topic Models mit Kontextinformationen

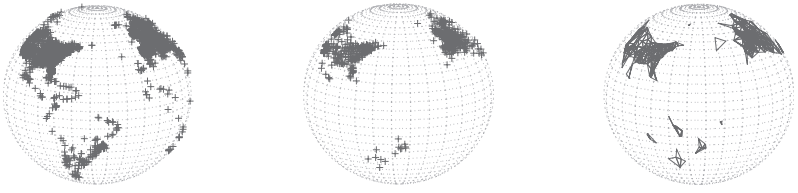
Typische Kontextinformationen für Dokumente aus sozialen Medien sind Zeitstempel, Nutzerinformationen und geographische Koordinaten (etwa die Position, von der aus ein Photo gemacht wurde, oder der Wohnort eines Nutzers). Absolute Zeitstempel beinhalten dabei immer auch Informationen über die Position auf dem Tages- Wochen- und Jahreszyklus. Klassische Kontext-Topic Models werden speziell an den zu bearbeitenden Datensatz angepasst, so dass Wahrscheinlichkeitsverteilungen über die verfügbaren Kontextvariablen (etwa geographische Variablen oder Zeitstempel) hinzugefügt werden. Dieses Vorgehen hat zwei große Nachteile: **i) Die erstellten Modelle sind nur für den gegebenen Datensatz und Datensätze mit identischen Informationen verwendbar** und **ii) Die klassische Modellierung von Kontextvariablen führt unrealistische Annahmen über die Relation zwischen Kontext und Topics ein**. Ein Beispiel für ii) sind Topic Models für geographisch verteilte Dokumente, etwa Fotos mit Tags und GPS-Koordinaten. In [Yi11] werden die Dokumente auf eine zweidimensionale Karte projiziert und mit Topics assoziierte, normalverteilte Regionen über die Daten gelegt. Das Problem ist, dass die Regionen als voneinander unabhängig modelliert werden. Das bedeutet, dass es laut Modell plausibel ist, dass benachbarte Städte völlig unterschiedliche Topicverteilungen haben. In [AHS13] wird daher eine hierarchische Beziehung zwischen den Regionen eingeführt. Dies führt aber auch zu Problemen, da viele geographische Strukturen – etwa Flüsse oder Ländergrenzen – schlecht durch eine Hierarchie aus Normalverteilungen modelliert werden können.

Es existieren bereits Topic Models, die Problem i) behandeln. Diese lernen Regressionen zwischen Kontextinformationen und den beobachteten Topics. Meistzitiertes Beispiel ist [MM08], in dem eine Dirichlet-Multinomiale Regression zwischen Kontext und Topics gelernt wird. Diese Modelle haben zwei Probleme: Erstens können nur Kontextinformationen verwendet werden, die auch Eingabe für eine Regression sein können (also keine geographischen Koordinaten, keine zyklischen Variablen) und zweitens wird ein linearer Zusammenhang zwischen Eingabe und Ausgabe angenommen – eine Annahme, die der Komplexität der Beziehung zwischen Topics und Kontext oft nicht gerecht wird. Erweiterungen wie etwa polynomielle Regressionen führen zu neuen Problemen wie Overfitting oder nicht interpretierbaren Parametern.

3 Hierarchische Multi-Dirichlet-Prozess-Topic Models

Der in meiner Dissertation präsentierte neue Ansatz zur Einbeziehung von Kontext-Informationen *basiert auf einem Clustering der Kontextvariablen und auf Nachbarschaftsbeziehungen*. Die Menge an verfügbaren Daten aus dem Web steigt rasant. Das bedeutet in der Praxis, dass es für viele Kontexte – sprich Orte, Zeitpunkte, usw. – eine große Anzahl von Beobachtungen (Dokumenten) gibt. Daher ist es möglich, Dokumente nach Kontextvariablen zu clustern und dann für jedes Cluster eine cluster-spezifische Topic-Verteilung zu lernen. Wählt man die Anzahl der Cluster hoch genug, können damit auch komplexe Zusammenhänge zwischen Kontext und Topics erkannt werden. Die Anzahl der Cluster wird nur dadurch nach oben beschränkt, dass die Anzahl der enthaltenen Dokumente ausreichen muss, um eine Topic-Verteilung zu bestimmen – wofür in der Praxis ein paar Dutzend Dokumente genügen. Die gefundenen Cluster wären nun unabhängig voneinander. Um Abhängigkeiten zwischen den Clustern herzustellen, werden *Topics zwischen benachbarten Clustern ausgetauscht*. Der Austausch von Topics basiert auf Hierarchischen Multi-Dirichlet-Prozessen und wird im nächsten Abschnitt beschrieben.

Das Clustering von Dokumenten nach Kontextvariablen ist unkompliziert: Für Zeitstempel reicht oft ein einfaches Binning, für geographische Daten können – wie in meiner Dissertation beschrieben – Mises-Fisher-Verteilungen auf der Sphäre [Fi53] auf die Daten angepasst werden, wobei wieder DPs eingesetzt werden können, um die Anzahl der Cluster zu lernen. Und für diskrete Kontexte, wie etwa das Geschlecht eines Nutzers oder eine Nutzer-ID, entsprechen die Cluster direkt den diskreten Ausprägungen der Variablen. Nachbarschaftsbeziehungen zwischen Variablen sind oft leicht zu finden: Existiert eine Ordnung (etwa bei Zeitinformationen oder zyklischen Variablen) so wird jedes Cluster mit dem direkten vorherigen und dem direkt nächsten Cluster in der Ordnung (falls vorhanden) verbunden. Für geographische Cluster kann eine beliebige Nachbarschaftsbeziehung definiert werden, wobei in meiner Arbeit die parameterfreie Delaunay-Triangulierung verwendet wird. Ergebnis ist ein geographisches Netzwerk aus Cluster-Zentroiden. Ein Beispiel für ein Clustering und eine Delaunay-Triangulierung von geographisch verteilten Dokumenten wird in Abb. 3 dargestellt. Um cluster-spezifische Topic-Verteilungen zu lernen, wird der Hierarchische Dirichlet-Prozess um eine Ebene erweitert: Jedes Dokument m zieht seine Verteilung über Topics G_m^d aus einem cluster-spezifischen DP seines zugeordneten Clusters j mit Skalierungsparameter α_1. Dieser DP hat wiederum die globale



(a) Document-Positionen (b) Geographische Cluster (c) Geographisches Netzwerk

Abb. 3: **Erstellung eines geographischen Netzwerks aus Dokument-Clustern.** (a) zeigt die Positionen von 34,707 Fotos, (b) die Cluster-Zentren und (c) die Delaunay-Triangulierung der Cluster.

Topic-Verteilung als Basismaß. Die globale Topic-Verteilung wird aus einem DP mit Basisverteilung $H \sim Dir(\beta)$ gezogen:

$$G_0 \sim \mathrm{DP}(\gamma, H); \quad G_j^c \sim \mathrm{DP}(\alpha_0, G_0); \quad G_m^d \sim \mathrm{DP}(\alpha_1, G_j^c); \quad \phi_{mi} \sim G_m^d; \quad w_{mi} \sim \phi_{mi}. \tag{4}$$

Um die Nachbarschaftsbeziehungen zwischen Clustern zu modellieren und Informationen über die verwendeten Topics zwischen benachbarten Clustern auszutauschen, habe ich eine Generalisierung von Dirichlet-Prozessen eingeführt, den Hierarchischen *Multi-Dirichlet-Prozess* (MDP). Dieser wird ähnlich zum DP definiert (siehe Gleichung 2). Gegeben P Wahrscheinlichkeitsmaße $G_1, \ldots, G_P$ (*"Eltern-Verteilungen"*) mit den Parametern $\alpha_1, \ldots, \alpha_P \in R^{>0}$. Der $MDP(\alpha_1, \ldots, \alpha_P, G_1, \ldots, G_P)$ ist dann ein Wahrscheinlichkeitsmaß G für dessen Realisierung gilt, dass jede Partitionierung $(A_1, \ldots, A_r)$ der (potentiell unendlichen) Menge an Verteilungen Dirichlet-verteilt ist:

$$(G(A_1), \ldots, G(A_r)) \sim \mathrm{Dir}\left(\sum_{p=1}^{P} \alpha_p G_p(A_1), \ldots, \sum_{p=1}^{P} \alpha_p G_p(A_r)\right) \tag{5}$$

In einem HMDP Topic Model wird über die Dokumente eine Abhängigkeit der Topic-Verteilungen benachbarter Cluster eingeführt. Dazu ziehen Dokumente ihre Topics aus einem MDP, dessen Eltern-Verteilungen aus den *Basisverteilungen des eigenen Clusters sowie der benachbarten Cluster mit Gewichtung* η bestehen. Im Falle von mehreren Kontextvariablen werden die *Basisverteilungen der verschiedenen Kontexte mit einer Gewichtung* ζ gemischt. Formal wird die Basisverteilung eines Dokuments dann aus dem folgenden MDP erstellt:

$$G_m^d \sim MDP(\alpha_1 \zeta \eta, G^c); \qquad \zeta \sim \mathrm{Dir}(\varepsilon); \qquad \eta_{fi} \sim \mathrm{Dir}(\delta_f) \tag{6}$$

wobei der Parameter $\alpha_1 \zeta \eta$ eine Abkürzung für die Gewichte der Eltern-Verteilungen ist, α_1 der Skalierungsparameter und G^c die Menge an Eltern-Verteilungen, die wie in Gleichung 4 gezogen werden. Die Gewichtung der Einflüsse von benachbarten Clustern η_{fi} für das $i$te Cluster in Kontext f sowie die Gewichtung der Kontexte ζ werden jeweils aus Dirichlet-Verteilungen mit Parametern ε und δ_f gezogen. Diese Parameter können gelernt werden. Ein Fixpunkt-Schätzer für den Parameter η_{fi} (für wechselnde Anzahlen von Eltern-Verteilungen) wird in meiner Dissertation gegeben.

Für das Lernen der Topics wird eine neue Form der sogenannten *practical stochastic variational inference* [Bl13] für HMDPs hergeleitet. Hierfür zeige ich, wie die dokument- und cluster-spezifischen Topic-Verteilungen G^d und G^c sowie die Mischungsverhältnisse der Kontext-Cluster ζ und η herausintegriert werden können. Hierdurch müssen lediglich die Topic-Zuweisungen der Wörter und die globale Topicverteilung gelernt werden, was die Inferenz deutlich beschleunigt und ein optimiertes verteiltes Berechnen ermöglicht.

4 Anwendungsbeispiele

Dass Topic Models von Kontextinformationen profitieren, lässt sich leicht anhand von sehr kurzen Dokumenten zeigen. Ein Beispiel sind Bilder aus Fotocommunities, die oft

nur einen einzelnen Tag enthalten. *Ohne Kontextinformationen würden hier keine Topics gefunden, mit Kontextmodellen können selbst in Sammlungen von Dokumenten mit jeweils nur einem Wort Topics gefunden werden* [Kl14].

Um die Vorteile des Nachbarschafts-Austauschs von Topics aus Gleichung 6 zu demonstrieren, wurden unter anderem geographisch verteilte, verschlagwortete Fotos von Speisen mit dem HMDP-Topic Model analysiert. Der Datensatz stammt aus dem Paper zu LGTA, dem meistzitierten geographischen Topic Model für regional unterschiedliche Topicverteilungen [Yi11]. In Abb. 1 sieht man am Beispiel der Verteilung des Topics "seafood", dass der *HMDP die komplexe geographische Verteilung von Topics besser erkennt* und dass die *Topics semantisch kohärenter* sind. Diese Beobachtung konnte in mehreren Datensätzen anhand von Vorhersageproblemen gemessen werden.

Zusätzlich wurde eine Nutzerstudie mit 31 Teilnehmern durchgeführt. Um die Qualität der Topics zu messen, wurden pro Topic die fünf Wörter mit der höchsten Wahrscheinlichkeit mit einem fremden Wort vermischt [Ch09]. Aufgabe der Teilnehmer war es, das fremde Wort zu finden. Für Topics von LGTA gelang das in verschiedenen Experimenten bei 57%-67% der Fälle. Bei HMDP-Topics bei 79-82%. Das bedeutet, dass der HMDP besser in der Lage ist, semantisch zusammenhängende Worte als Topics zu identifizieren [Ch09, Kl14].

Die Struktur von HMDP-Topic Models erlaubt nicht nur eine verbesserte Topic-Erkennung, sie ermöglicht auch die *Analyse der Zusammenhänge zwischen Kontextvariablen und Topics*. Dabei können nicht nur komplexe geographische Muster, sondern auch Muster jedes beliebigen Kontextes, etwa über die Zeit beobachtet werden. Dazu können die Topicwahrscheinlichkeiten pro Kontext-Cluster j direkt aus der Verteilung G_j^c abgelesen werden. Als Beispiel wurden in meiner Dissertation alle Emails der Linux-Kernel-Mailingliste seit 1995 analysiert. Dabei wurde herausgefunden, dass es professionelle Topics gibt, die im Wochenzyklus an Werktagen eine höhere Auftretenswahrscheinlichkeit besitzen: Abb. 4 zeigt beispielhaft die Verteilung für das Topic *driver, clock, control, register, device*.

Sind mehrere Kontextvariablen verfügbar, so lernt das HMDP-Topic Model eine *Gewichtung der Kontextvariablen* ζ nach ihrer Vorhersagekraft für die Topics des Modells. Warum

Tab. 1: **Topics von LGTA [Yi11] und dem HMDP [Kl14].** Die Topics des HMDP wurden geordnet, so dass ein Vergleich mit LGTA möglich wird. Die Karte zeigt Dokumentpositionen für Topic 1. LGTA kann im Unterschied zum HMDP den komplexen Verlauf der Küste in Italien nicht modellieren und findet daher dort keine Dokumente zu Topic 1.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Karte für Topic 1
LGTA	fish	chocolate	japanese	vegetarian	wine	chinese	mexican	sushi	
	seafood	cheese	sushi	vegan	italian	chicken	bbq	thai	
	rice	bread	ramen	chocolate	coffee	noodles	chicken	korean	
	shrimp	fish	fish	baking	french	soup	burger	japanese	
	crab	wine	noodle	bread	pizza	rice	sandwich	salmon	
	lobster	tapas	sashimi	cheese	chocolate	vietnamese	fries	rice	
	chicken	orange	noodles	bacon	bakery	dimsum	hamburger	tuna	
HMDP	seafood	chocolate	japanese	salad	wine	chinese	mexican	bbq	
	fish	icecream	sushi	cheese	pizza	thai	tacos	burger	
	lobster	strawberry	fish	tomato	coffee	chicken	taco	fries	
	shrimp	baking	ramen	bread	italian	rice	salsa	hamburger	
	crab	cream	sashimi	chicken	pasta	soup	burrito	grill	
	wine	coffee	rice	fish	cheese	noodles	chicken	chicken	
	salmon	pie	salmon	vegetarian	french	korean	chips	sandwich	

Data: cc-by-SA OSM.org

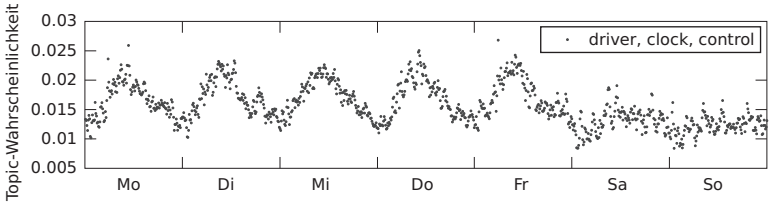


Abb. 4: **Wahrscheinlichkeit eines professionellen Topics in der Linux-Kernel-Mailingliste über den Wochenzyklus.** Die Wahrscheinlichkeit des Topics steigt deutlich zu den Arbeitszeiten.

das eine interessante Information ist, zeige ich anhand von Nutzerprofilen aus einer Online-Community für sexuelle Fetischisten. Die Nutzer der Plattform beschreiben ihre Fetischismen mit frei gewählten Tags, die auch von anderen Profilen kopiert werden können. Insgesamt besteht der Datensatz aus 126.408 Profilen [Fa16], für die Geschlecht, Alter, die Information *ob* der Nutzer in einer Beziehung ist und, falls ja, in welchem Beziehungstyp, die sexuelle Orientierung, Rolle (etwa Herr oder Sklave) und ob der Nutzer Fetish-Events besucht, verfügbar sind. Abb. 5 zeigt, dass Geschlecht, Alter und Sexuelle Orientierung die höchste Vorhersagekraft für die Fetischismen der Nutzer haben. Das ist eine Erkenntnis, die mit herkömmlichen Modellen *nicht* untersucht werden könnte, da eine gegenseitige Abhängigkeit zwischen ζ und den Topics besteht: Wäre ζ anders, würden andere Topics gelernt. Und sähen die Topics anders aus, wäre die Verteilung ζ anders. Nur ein gemeinsames Modell erlaubt eine saubere Analyse.

5 Fazit

Kern meiner Dissertation ist der Hierarchische Multi-Dirichlet-Prozess (HMDP), eine Generalisierung des Hierarchischen Dirichlet-Prozesses. Dieser erlaubt es, beliebige Kontextinformationen (inklusive zyklischer oder geographischer Daten) in probabilistische Modelle zu integrieren, ohne jeweils die Modellstruktur anpassen zu müssen. Alle Parameter des Modells haben eine natürliche Interpretation als Wahrscheinlichkeiten oder Zähler, so dass direkte Visualisierungen und Plausibilitäts-Checks möglich sind. In meiner Arbeit wird eine effiziente Inferenz hergeleitet, mit der auch große Dokumentensammlungen als Datenstrom verarbeitet werden können. Implementierungen aller Modelle mei-

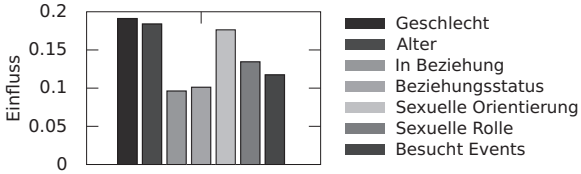


Abb. 5: **Gewichtung demographischer Variablen für den Fetisch-Datensatz.** Dargestellt ist die Wahrscheinlichkeit ζ_f, dass ein Topic aus Cluster-Topic Verteilungen des gegebenen Kontextes f heraus erklärt wird.

ner Dissertation (mit Kommandozeilen-Schnittstelle) sowie Zusatzmaterialien sind unter `http://topicmodels.west.uni-koblenz.de` unter einer freien Lizenz verfübar.

Literaturverzeichnis

[AHS13] Ahmed, Amr; Hong, Liangjie; Smola, Alex: Hierarchical Geographical Modeling of User Locations from Social Media Posts. In: WWW. 2013.

[Bl13] Bleier, A.: Practical Collapsed Stochastic Variational Inference for the HDP. ArXiv e-prints, Dezember 2013.

[BNJ03] Blei, David M.; Ng, Andrew Y.; Jordan, Michael I.: Latent Dirichlet allocation. J. Mach. Learn. Res., 3:993–1022, Marz 2003.

[Ch09] Chang, Jonathan; Boyd-Graber, Jordan; Wang, Chong; Gerrish, Sean; Blei, David M.: Reading Tea Leaves: How Humans Interpret Topic Models. In: Neural Information Processing Systems. 2009.

[Fa16] Fay, D.; Haddadi, H.; Seto, M. C.; Wang, H.; **Kling, C. C.**: An exploration of fetish social networks and communities. In: NetSciX'16. Januar 2016.

[Fi53] Fisher, Ronald: Dispersion on a sphere. Proc. of the Royal Society of London. Series A, Mathematical and Physical Sciences, 217(1130), 1953.

[Kl14] Kling, C. C.; Kunegis, J.; Sizov, S.; Staab, S.: Detecting Non-Gaussian Geographical Topics in Tagged Photo Collections. In: ACM International Conference on Web Search and Data Mining. 2014.

[Kl15] Kling, C. C.; Kunegis, J.; Hartmann, H.; Strohmaier, M.; Staab, S.: Voting Behaviour and Power in Online DemocracyGermany's Pirate Party. In: Proc. Int. Conf. on Weblogs and Social Media. 2015.

[MM08] Mimno, David M.; McCallum, Andrew: Topic Models Conditioned on Arbitrary Features with Dirichlet-multinomial Regression. In: UAI. AUAI Press, S. 411–418, 2008.

[Te06] Teh, Y. W.; Jordan, M. I.; Beal, M. J.; Blei, D. M.: Hierarchical Dirichlet processes. J. of the American Statistical Association, 101:1566–1581, 2006.

[Yi11] Yin, Zhijun; Cao, Liangliang; Han, Jiawei; Zhai, Chengxiang; Huang, Thomas S.: Geographical Topic Discovery and Comparison. In: WWW Conf. S. 247–256, 2011.

Christoph Kling studierte an der Universität Koblenz-Landau Informatik mit Anwendungsfach Wirtschaftsinformatik. Das Diplom-Studium schloss er in Regelstudienzeit mit Note 1.1 ab. Während seiner Promotion war er wissenschaftlicher Mitarbeiter am Institute for Web Science and Technologies. Für das Papier *Voting Behaviour and Power in Online Democracy* [Kl15], dessen Ergebnisse Teil der Dissertation sind, wurde er 2015 auf der AAAI International Conference on Web and Social Media mit dem *Honorable Mention Award* ausgezeichnet. Neben seiner Promotion war er Gründungsmitglied der Hochschulgruppe *Denkfabrik für Humanismus und Aufklärung* an der Universität Mainz, Gründungsmitglied der *GBS Mainz-Rheinhessen e.V. – Gottlose Humanisten*, Gründungsmitglied bei *Intaktiv e.V. – Eine Stimme für genitale Selbstbestimmung* sowie Kandidat auf Listenplatz 2 für die Piratenpartei bei der Mainzer Stadtratswahl 2014.