Zusammenfassung der wissenschaftlichen Ergebnisse zur Dissertation

Ein generischer Ansatz für Topic-Modelle und seine Anwendung auf virtuelle Gemeinschaften

(Englischer Originaltitel: *A generic approach to topic models and its application to virtual communities*)

der Fakultät Mathematik und Informatik der Universität Leipzig eingereicht von

Dipl.-Ing. Gregor Heinrich

angefertigt in der Abteilung Automatische Sprachverarbeitung

Überblick

Diese Arbeit befasst sich mit einem generischen Modell für Topic-Modelle, um deren Entwurf und Implementierung zu unterstützen. Topic-Modelle sind probabilistische Repräsentationen gruppierter diskreter Daten. Angewandt auf Textdaten, repräsentiert das grundlegende Topic-Modell, *Latent Dirichlet Allocation*, Dokumente als Mischungen von *Topics* – Wahrscheinlichkeitsverteilungen über das Vokabular. Lange ist bekannt, dass zwischen Begriffen, die in einem Topic gemeinsam eine hohe Wahrscheinlichkeit haben, oft ein semantischer Zusammenhang besteht. Dieses auf Wort-Kookkurrenzen beruhende Phänomen kann z.B. für Suchsysteme oder Data Mining ausgenutzt werden, wobei mittlerweile eine Vielzahl von Arbeiten entstanden ist, die das Basismodell durch Hinzufügen neuer Strukturen erweitern. In diesen werden meist Strukturen in den Daten modelliert, die über einfache Kookkurrenz hinausgehen, oder es werden verschiedene Modalitäten in den Daten gemeinsam modelliert, um deren Zusammenhänge zu untersuchen.

Fragestellungen. Während diese Modell-Erweiterungen erfolgreich angewandt werden, existiert bisher noch keine Analyse von Topic-Modellen als eine generische Modell-Klasse. Diese wird in der Dissertation erarbeitet. Die motivierende Vermutung ist dabei, dass sich wichtige Eigenschaften von Topic-Modellen über einen weiten Bereich von Modell-Strukturen verallgemeinern lassen. Damit verbundene Fragestellungen sind, wie dies zu Vereinfachungen in der Herleitung von Modelleigenschaften, Inferenzalgorithmen und schließlich Entwurfsmethoden genutzt werden kann.

Vorgehensweise. Die Vermutung wird in mehreren Schritten überprüft, beginnend mit einer formalen Definition von Topic-Modellen als generische Modell-Klasse. Aufbauend darauf werden die anderen Fragestellungen adressiert, sowohl auf der zuvor definierten generischen Ebene, als auch durch Anwendung auf ein konkretes Szenario.

Anwendung. Als exemplarische Anwendungsdomäne werden virtuelle Gemeinschaften betrachtet, wie sie in großen Organisationen, in der Wissenschaftsgemeinschaft und im “Web 2.0” auftreten. Da deren Daten häufig die Behandlung mehrerer Modalitäten erfordern (z.B. Text, Klassifizierungsinformation, Autorenschaft oder Zitationsnetzwerke, aber auch Nutzungsstatistiken) und die Datenmodelle für verschiedene Anwendungen sehr verschieden sein können (z.B. *Collaborative Filtering* von Autoren gegenüber inhaltsbasierter Expertisesuche), ist die Variationsbreite möglicher Modellstrukturen hoch. Dies macht vereinfachte Entwurfsmethoden, wie sie in der Arbeit untersucht werden, besonders sinnvoll.

Ergebnisse. Die Arbeit leistet Forschungsbeiträge sowohl auf theoretischer Seite – generische Formulierungen verschiedener Aspekte von Topic-Modellen und virtuellen Gemeinschaften –, als auch auf empirischer Seite – durch Anwendung der theoretischen Ergebnisse bei der Implementierung von Inferenzalgorithmen und konkreten Topic-Modellen. Die oben formulierte Vermutung konnte bestätigt werden. Resultierend daraus ist es nun möglich, einen Teil der Komplexität beim Entwurf und der Implementierung von Topic-Modellen vor Anwendern zu verbergen. Mit weniger benötigtem Vorwissen lassen sich Topic-Modelle somit einfacher einsetzen, besonders in Fachgebieten außerhalb des maschinellen Lernens oder der automatischen Sprachverarbeitung.

Im Folgenden werden die Ergebnisse nach thematischen Bereichen zusammengefasst vorgestellt.

Topic-Metamodell und *Networks of Mixed Membership* (NoMMs)

Ein zentraler Forschungsbeitrag ist die Formalisierung von Topic-Modellen als generische Modellklasse in Verbindung mit der Herleitung ihrer wichtigsten Eigenschaften. Hierfür wurde ein Metamodell entwickelt, das auf der Interpretation von Topic-Modellen als Mischverteilungen höherer Ordnung fußt. Die Verbindung dieses Ansatzes mit existierenden Modellen in der Literatur wird illustriert und es wird gezeigt, wie der Anwendungsbereich erweitert werden kann durch Nutzung von Varianten zur typischen diskreten Verteilung mit Dirichlet-Prior. Weiterhin wird als domänenspezifische, kompakte Alternative zu Bayes'schen Netzwerken (der üblichen Darstellung) eine graphische Repräsentation für Topic-Modelle vorgeschlagen: "*Networks of Mixed Membership*" (NoMMs). Der Vorteil dieser neuen Darstellung besteht in der einfacheren Beschreibung typischer Strukturen in Topic-Modellen. Die NoMM-Repräsentation dient als Grundlage für die generische Modellierung in der gesamten Arbeit, wobei sich zeigt, dass verschiedene Modelleigenschaften tatsächlich direkt in NoMM-Strukturen widergespiegelt sind.

Generische Inferenzverfahren

Inferenzverfahren erlauben das Training eines Modells anhand von Daten und bilden die Grundlage für seine praktische Anwendung. Für zwei approximative Bayes'sche Inferenzverfahren – den Gibbs-Sampler und ein variationales Inferenzverfahren – werden die numerischen Aktualisierungsvorschriften und Algorithmen generisch hergeleitet. Für verschiedene Topic-Modelle wird gezeigt, dass es direkte Zuordnungen zwischen Modellstrukturen und Algorithmen gibt. Für konkrete Modelle können daher mit den erarbeiteten NoMM-basierten Verfahren die Inferenzvorschriften direkt angegeben werden, ohne sie von Grund auf für jedes Modell neu herleiten zu müssen (wie in der Literatur üblich). Ein empirischer Vergleich beider Inferenzverfahren für verschiedene Topic-Modelle fällt zugunsten des Gibbs-Samplers aus, wobei Modellqualität (Likelihood bzw. Perplexität eines Testdatensatzes) und Konvergenzzeit betrachtet werden.

Skalierbare Gibbs-Sampler

Da Inferenzalgorithmen vor allem für komplexere Topic-Modelle mit mehreren statistisch abhängigen latenten Variablen nur langsam konvergieren, werden Verfahren zur Verbesserung der Skalierbarkeit betrachtet. Hierfür wird (1) untersucht, inwieweit sich existierende Verfahren für das Basis-Modell *Latent Dirichlet Allocation* auf allgemeinere NoMM-Strukturen erweitern lassen. Es wird (2) eine alternative Methode vorgeschlagen sowie (3) der Einfluss von Abhängigkeiten zwischen latenten Modellvariablen untersucht, da diese ein Hauptfaktor für die Skalierbarkeit sind.

In einer empirischen Studie wird gezeigt, dass sowohl verallgemeinerte serielle, hauptsächlich aber parallele Methoden die Performanz des Gibbs-Samplers besonders bei komplexeren Modellen deutlich verbessern. Ein wichtiges empirisches Ergebnis ist, dass die Auflösung von Abhängigkeiten zwischen latenten Variablen fast keine Nachteile bei der erreichten Modellqualität mit sich bringt, jedoch die absolute Konvergenzzeit zum Teil stark verkürzt. Besonders in Kombination mit seriellen und parallelen Skalierungsverfahren werden auf diese Weise hohe Beschleunigungsfaktoren erzielt. Diese Ergebnisse erweitern die Anwendbarkeit von Topic-Modellen für komplexere Modellstrukturen und größere Datenmengen.

Gibbs-Metasampler

Als eine direkte Anwendung des generischen Inferenzverfahrens wird für die Implementierung von Algorithmen ein Verfahren vorgestellt, das aus der Spezifikation eines NoMMs in einer einfachen domänenspezifischen Sprache den Quellcode eines vollständigen Gibbs-Samplers in den Programmiersprachen Java oder C erstellt. Anders als bestehende Lösungen ist dieser Codegenerator spezialisiert auf die Eigenschaften von Topic-Modellen und kann für Optimierungen einfach erweitert werden. So kann Quellcode für die o.g. Skalierungsverfahren generiert werden. Da das Verfahren Gibbs-Sampler erstellt, wird die Bezeichnung "Gibbs-Metasampler" verwendet. Für eine Anzahl von Beispielmodellen wird die automatische Implementierung mit diesem Codegenerator validiert.

Entwurfsmethode für Topic-Modelle und NoMM-Typologie

Eine Methode zum Entwurf von Topic-Modellen wird erarbeitet. Um die Vorteile der NoMM-Repräsentation bestmöglich zu nutzen, wird untersucht, wie NoMMs in ihre Substrukturen zerlegt und Gibbs-Sampler- und Likelihood-Eigenschaften modularisiert werden können. Anhand einer umfangreichen Literaturrecherche wird eine Typologie von NoMM-Substrukturen entwickelt und in einer "Bibliothek" zusammengefasst. Die Typologie wird durch eine neue Substruktur ergänzt.

Auf Grundlage der Modularisierung und Substrukturen-Bibliothek wird ein Entwurfsverfahren vorgeschlagen, mit dem Designer in einem iterativen Prozess aus gegebenen Datenstrukturen und Modellierungsannahmen NoMM-Strukturen aufbauen können. Da sich beim Aufbau von NoMMs die Eigenschaften des Gesamtmodells direkt nachvollziehen lassen, ist es möglich, mit jedem Konstruktionsschritt die Konsequenzen für das Gesamtmodell abzuschätzen.

Modellierung virtueller Gemeinschaften

Um den Anwendungsbereich virtuelle Gemeinschaften zu charakterisieren, werden wichtige Anwendungsfälle analysiert. Zentraler Gesichtspunkt ist der Zugriff auf Informationen und explizites Wissen sowie die Möglichkeit, Hinweise auf implizites Wissen aufzudecken. Das vorgeschlagene "*AMQ*-Modell" besteht aus einem Graphen mit drei Entitätsklassen: *Actors* (Wissensträger), *Media* (Dokumente und andere Informationsressourcen) und *Qualities* (Einheiten beschreibbaren Wissens). *Qualities* stellen die Verbindung zu bestehenden Wissensrepräsentationsverfahren dar und verallgemeinern Topics. Über Relationen zwischen den Entitäten lassen sich Anwendungsszenarien definieren, die typisch für virtuelle Gemeinschaften sind. Somit lassen sich Problemstellungen einfacher beschreiben und vergleichen.

Fallstudie: Expertisefindung mit getaggten Dokumenten

Eine abschließende Fallstudie baut auf allen anderen Ergebnissen der Dissertation auf und untersucht Expertisefindung in einer wissenschaftlichen virtuellen Gemeinschaft, deren Inhalte durch eine digitale Bibliothek repräsentiert werden. Ausgehend von einer Beschreibung des Szenarios als AMQ-Modell werden mit der vorgeschlagenen NoMM-Entwurfsmethode mehrere Topic-Modelle entworfen, die neben Textinhalten und Autorenschaft von Experten auch semantische Tags von Dokumenten berücksichtigen: *Expert-Tag-Topic Models*. Ein experimenteller Test mit dem Gibbs-Metasampler validiert die Modelle, und es zeigt sich, dass die zusätzlichen Tag-Informationen sowohl die semantische Qualität der gelernten Topics, als auch die Ergebnisse von Beispielanfragen gegenüber dem *Author-Topic-Model* aus der Literatur verbessern.