News Recommendation with ML

Hauptseminararbeit Master Informatik

Timo Lange

HAW Hamburg
Fakultät Technik und Informatk
Studiendepartment Informatk
Studiengang: Master Informatk

21. September 2018

Die folgende Arbeit dient als Projektvorschlag für den Entwurf eines Recommendation Systems und soll die Rahmenbedingungen und ein mögliches Vorgehen skizzieren. Es werden Forschungsfragen vorgestellt und mit dem Thema verwandte Arbeiten beleuchtet, welche mögliche Methoden zur Beantwortung der Forschungsfrage liefern könnten. Es wird der Datensatz mit welchem gearbeitet werden soll vorgestellt und auf Forschungsrisiken eingegangen. Des Weiteren werden mögliche Lösungswege vorgeschlagen und umrissen. Die Beantwortung der Forschungsfrage ist dabei Gegenstand weiterführender Arbeiten, auf die am Ende ein Ausblick gegeben wird.

1 Einleitung

1.1 Recommendation Systeme (RS)

Recommendation Systeme sind Softwaretools die Vorschläge für Items für Nutzer liefern und diese somit in ihrer (Online) Entscheidungsfindung unterstützen können. Das Ziel von RS ist einfach zugängliche und hoch qualitative Empfehlungen für eine Gemeinschaft von Nutzern zu liefern. Im Zeitalter von *Big Daten* und einer größer werdenden Flut von Informationen mit denen die Nutzer konfrontiert werden, helfen RS den Nutzern mit dem zunehmenden *Information Overload* fertig zu werden. Auf Seite der Unternehmen welche RS einsetzen, können diese die Nutzerzufriedenheit mit dem Dienst und die Einnahmen von Content-Anbietern steigern.

1.2 Deep Learning (in RS)

Im Fokus dieser Arbeit stehen Verfahren aus dem Bereich Deep Learning (DL) Ian Goodfellow u. a., allerdings soll diese Arbeit nicht auf DL begrenzt sein. DL ist ein relativ neues Feld im Bereich der Recommendationsysteme und hat in den letzten Jahren einen Aufschwung erfahren. Dabei können DL Verfahren klassische Verfahren in Bereichen wie etwa Collaborativ Filtering (CF) oder Content Based (CB) ersetzen. Zudem kann DL z.B. dabei Helfen Daten aus verschiedenen Quellen wie Text, Bild oder Video in ein Modell zu integrieren und ist zudem geeignet mit klassischen Verfahren wie etwa Matrix Factorization für CF kombiniert zu werden.

1.3 DPA Datensatz

Ausgangspunkt für das zu entwerfende System ist ein Datensatz der *Deutsche Presse Agentur (DPA)*, welcher in Abschnitt 4.2 näher betrachtet wird. Der Datensatz besteht aus aktuell über 200.000 Artikeln welcher im strukturierten NewsML-G2 Format vorliegt. Dabei sind die Texte von hoher Qualität und enthalten umfangreiche Metadaten. Die von Journalisten verfassten Texte und die dazugehörigen Metadaten sind ein Alleinstellungsmerkmal für diesen Datensatz, welche nach bestem Wissen des Autors nicht in einem öffentlich zugänglichem Datensatz zu finden sind. Somit ist dieser Datensatz besonders interessant, da dieser neue Möglichkeiten zur Entwicklung eines RS eröffnet. Als Ziel der Empfehlungen sind zunächst drei verschiedene Szenarien denkbar:

Recommendation zum jeweiligen Artikel Zu einem Artikel wird eine Empfehlungsliste mit ähnlichen Artikeln gegeben, wobei eine Möglichkeit wäre, verschiedene Empfehlungslisten, jeweils nach bestimmten Aspekten zusammengestellt zu generieren. Diese könnten als Metainformation zum jeweiligen Artikel ausgeliefert werden.

DPA-Select Die Kunden der DPA sind i.d.R. keine Einzelpersonen sondern z.B. Zeitungen und Unternehmen. Diese können einen News-Stream abonnieren, welcher nach gewünschten Themen gefiltert wird. Hier könnte ein RS eingesetzt werden um starre Filterlisten durch individualisierte Empfehlungen für das Unternehmen zu ergänzen oder ein Ranking für die Artikel der Filterliste zu erstellen. Somit könnten Kunden für sie interessante Meldungen erhalten, welche ihnen durch starre Filterregeln evtl. entgangen wären und ein Ranking der Artikel könnte die Arbeit der Redakteure erleichtern.

DPA-Executive Dieses Angebot richtet sich an Einzelpersonen wie Manager und Mitarbeiter die direkt mit aktuellen News versorgt werden. Hier bietet sich eine personalisierte Empfehlung an, wie sie von Nachrichtenseiten, Content-Anbietern wie Netflix oder E-Commerce Seiten wie Amazon bekannt ist.

1.4 Gliederung

Zunächst werden Forschungsfragen gestellt, auf dessen Grundlage im nachfolgenden Abschnitt verwandte Arbeiten beleuchtet und dessen Verfahren vorgestellt werden, welche verwendet werden können um sich einem Ergebnis zu nähern, mit anschließender Einordnung und Abgrenzung der eigenen Arbeit. Hierauf folgt ein Abschnitt in dem das Vorgehen zur Beantwortung der Forschungsfrage erörtert wird, sowie der verwendeten bzw. benötigten Daten und geplanten zu verwendenden Soft- und Hardware. Hierauf wird kurz auf Forschungsrisiken als auch Chancen eingegangen. Die Arbeit schließt mit einem Ausblick auf die weiteren Schritte.

2 Forschungsfragen

Die folgenden Forschungsfragen dienen als Grundlage dieser Arbeit und sollen in nach gelagerten Projekten geklärt werden.

- (RQ1) Wie geeignet sind Deep Learning Modelle für (DPA) News Empfehlungen?
- (RQ2) Welche Algorithmen sind am besten geeignet, um Empfehlungen für den DPA Datensatz zu erzeugen?
- (RQ3) Sind die vorhandenen / noch zu erwartenden Daten von der DPA ausreichend, um gute Empfehlungen zu erzeugen?
- (RQ4) Welche Methoden sind geeignet, aufeinander aufbauend, sukzessive je nach Verfügbarkeit der Daten, ein Empfehlungsmodell zu entwickeln.

2.1 Erwartete Ergebnisse

Zu RQ1 und RQ3 werden die folgenden Ergebnisse erwartet.

- Deep Learning wird sehr nützlich sein, um viele verschiedene Daten für die Empfehlung zu nutzen (Metadaten, Bilder).
- Die Qualität der Empfehlung wird sehr stark davon abhängen, ob viele Nutzerdaten verwendet werden können.

Die Prognose für RQ2 und RQ4 ist mit einer ausgiebigen Literaturrecherche verbunden und wird in Abschnitt 3.1 mittels ausgewählter Arbeiten beleuchtet.

3 Recommendation Systeme

In den folgenden Abschnitten werden zunächst einige ausgewählte, verwandte Arbeiten vorgestellt. Darauf folgt eine kurze Einordnung bzw. Abgrenzung der eigenen Arbeit.

3.1 Related Work / Literatur Review

3.1.1 Embedding-based News Recommendation for Millions of Users

Die Arbeit von Okura u. a. befasst sich mit der personalisierte Empfehlung von Nachrichten für Nutzer der Yahoo! Japan Smartphone Applikation (App). Auf der Startseite der App werden im unteren Bereich individuelle Empfehlungen für Artikel angezeigt, welche nach Relevanz für den Nutzer geordnet angezeigt werden, wobei durch herunterscrollen beliebig viele weitere Artikelempfelungen nachgeladen werden. In der Arbeit wird eine Embedding basierte end-to-end Methoden vorgestellt, welche aus drei Einzelschritten besteht:

1. Erstellung einer *distributed* Repräsentation der Artikel mittels einer Variante von *Denosing Autoencoder*

2. Generieren der Nutzer-Repräsentation mittels eines *Recurrent Neural Network (RNN)* wobei zum einfachen RNN noch die Varianten *Long-Short Term Memory Gated (LSTM)* und *Recurrent Unit (GRU)* getestet wurden. Als Eingabesequenz dient hier der Browsing Verlauf der Nutzer.

3. Erstellung der Empfehlung für die Nutzer mittels Bildung des inneren Produktes der Artikel- und Nutzer-Embeddings. Hierbei erfolgt ein Ranking der Artikel nach Relevanz für den Nutzer und De-Duplikation ähnlicher Artikel.

Zum Vergleich des neuen Modells wurde ein einfaches Bag of Words (BoW) Modell herangezogen. Das neue Modell stellte sich als deutlich performanter heraus, wobei die Nutzer Repräsentation mittels GRU die besten Ergebnisse, gefolgt vom LSTM lieferte. Ein weiterer Vorteil von GRU ist, dass hierbei keine *Gradient Explosion* zu verzeichnen war, wobei das einfache RNN und das LSTM im Training mehrmals wegen der *Gradient Explosion* fehlschlugen und ein *Gradient Clipping* nötig machten.

3.1.2 Dynamic Attention Deep Model for Article Recommendation by Learning Human Editors' Demonstration

Die Arbeit von Wang u. a. befasst sich mit der Empfehlung von Nachrichten für Redakteure. Redakteure müssen aus täglich mehreren Tausend Artikeln aus verschiedenen Quellen passende, qualitativ hochwertige Artikel zur Veröffentlichung auswählen. Die Anzahl der verfügbaren Artikel ist dabei deutlich größer als die Menge, welche das Redakteursteam begutachten kann. Das vorgestellte Modell soll eine Vorauswahl für die Redakteure treffen, sodass möglichst viele unpassende und qualitativ ungeeignete Artikel nicht mehr begutachtet werden müssen. Die Problemstellung ist der Empfehlung der DPA Artikel für deren jeweilige Kunden sehr ähnlich, da auch hier die Nutzer jeweils eine Gruppe von Personen sind und es liegt ein aggregiertes anstatt einem individuellen Interesse vor.

Die Autoren stellen ein *meta-attention* Modell über mehrere DNN(Deep Neural Network) vor, welches die Präferenzen bzw. Auswahlkriterien der Redakteure erfasst. Die Modelle lernen die

Repräsentation der Artikel und die Zusammenhänge mit den Metadaten der Artikel. Es werden über mehrere Tage hinweg verschiedene Modelle trainiert und das *Attention* Modell wählt für das aktuell zu trainierende Modell eines der vorherigen trainierten Modelle aus und lässt dessen "Wissenïn das neue Modell mit einfließen. So wird eine Änderung der Auswahlkriterien der Redakteure adaptiv erfasst. Für das Training eines neuen Modells werden nur die neuen Artikel des jeweiligen Tages verwendet.

Dieses so genannte Dynamic Attention Deep Model (DADM) nutzt die Modellierung des Textes auf Zeichenebene und Convolutional Neural Networks (CNNs) um Muster im Text zu erkennen und so die Auswahl der Redakteure vorherzusagen. Um die unterschiedlichen Veränderungen im Auswahlverhalten der Redakteure zu handhaben, weißt das *Attention* Modell den bereits trainierten Modellen verschiedene Einflussfaktoren, basierend auf dem aktuell trainierten Modell und Artikel zu. Diese Einflussfaktoren sind:

1. Model profile
2. Time profile

Das DADM besteht aus drei Hauptbestandteilen:

Text Representation Learning CNN basiert um die generelle Repräsentation des Dokuments zu erfassen.

Multi-view Categorical Data Modeling Inspiriert vom wide & deep model wird das lineare Modell für die Metadaten und das CNN Modell für die sequentiellen Textdaten kombiniert.

Dynamic Attention Deep Model Erfasst das dynamische Verhalten der Redakteure und nutzt dazu die *speciality* sowie die *timeliness* der über die vorherigen Tage Trainierten Deep Networks.

3.1.3 Multi-Rate Deep Learning for Temporal Recommendation

In dieser Arbeit von (Song u. a.) wird eine neue DNN Architektur vorgestellt, welche die Kombination von statischen *long-term* und temporären *short-term User Preferences* modelliert um Empfehlungen zu verbessern.

Um das Modell auch für große Anwendungsfälle effizient trainieren zu können, wird eine neue *pre-train* Methode vorgestellt, mit welcher die Anzahl freier Parameter signifikant reduziert werden soll.

Für Ihr Modell gehen die Autoren von der Annahme aus, dass sich die *User Preferences* aus zwei Komponenten zusammensetzen. Zum einen die *long-term* welche die relativ stabilen Langzeit Präferenzen der Benutzer darstellen und die temporären Präferenzen, welche die aktuellen Interessen der Nutzer darstellen.

Es wird ein *multi-rate temporal deep learning* Modell vorgestellt, das diese *long-* und *short-term* gleichzeitig optimiert, genannt Multi-Rate Temporal Deep Semantic Structured Model

(MR-TDSSM). Dabei soll der Ausdruck *multi-rate* darauf hindeuten, dass das Modell in der Lage ist, die Nutzerinteressen in unterschiedlicher Granularität zu erfassen, so dass zeitliche Dynamik mit unterschiedlichen Raten gleichzeitig optimiert werden kann. Dabei besteht das vorgestellte Modell aus mehreren Komponenten:

1. Ein Deep Semantic Structured Model (DSSM) um die statischen Nutzerinteressen zu modellieren.
2. Zwei LSTM-based temporal models um die täglichen und wöchentlichen Nutzerinteressen zu erfassen.
3. Ein LSTM temporal model zur Erfassung der globalen Nutzerinteressen.

Die RNNs für die temporale Modellierung unterschiedlicher Raten werden mittels eines *fully-connected* Feedworward Networks verbunden und können so gemeinsam trainiert werden und trotzdem unabhängig in Bezug auf die Feature-Aggregation sein. Der Nachteil dieser Architektur ist die hohe Anzahl an Parametern der RNNs. Diesem Problem begegnen die Autoren mit einer *pre-train* genannten Technik, um die Anzahl freier Parameter zu reduzieren. *pre-train* wird hier ähnlich wie *word embeddings* verwendet um User- und Item Embeddings mittels zusätzlichen Trainingsdaten zu erstellen. Zunächst wird das DSSM trainiert und anschließen werden durch das trainierte Netz die Nutzer- und Item Embeddings generiert die das MR-TDSSM nutzt.

3.1.4 Location-Aware Personalized News Recommendation With Deep Semantic Analysis

In der Arbeit von Chen u. a. (a) werden zwei Verfahren vorgestellt, um Orte in personalisierte News Empfehlungen mit einzubinden. Einmal das rein auf direkt berechneten Wahrscheinlichkeitsverteilungen beruhende Location-aware Personalized news recommendation with Explicit Semantic Analysis (LP-ESA) und das um DNNs erweiterte Location-aware Personalized news recommendation with Deep Semantic Analysis (LP-DSA). Ein öffentlich zugänglicher Twitter Datensatz wird genutzt um eine User-History zu erhalten, wobei nur Tweets aus dem Datensatz genutzt werden, welche eine URL zu einem News Artikel enthalten und der jeweilige Artikel heruntergeladen wird. Wikipedia wird als eine Art Knowladgebase für die Beschreibungen von Orten genutzt, um damit Embeddings für einen Ort bzw. Vektoren von *topics*(Themen/Begriffe) die den jeweiligen Ort beschreiben zu generieren.

LP-ESA bietet eine personalisierte News Empfehlung, basierend auf den persönlichen Interessen des Nutzers und seinem aktuellen Ort. Hierbei wird eine Sammlung von *geo-tagged* Dokumenten als eine Beschreibung des Ortes verwendet. Die Nutzer Historie wird verwendet um die persönlichen Interessen des Nutzers zu modellieren.

LP-ESA projeziert alle Text-Items wie *geo-tagged* Dokumente, Nutzer Historie und News Artikel mittels ESA in einen *topic space* welcher auf Wikipedia-Begriffen basiert.

Die Nutzerinteressen sowie News werden als gewichtete *topic* Vektoren, genannt *general user profiles* und *general news profiles*, dargestellt. Die *topic distributions* für einen Ort werden

mittels der *topic* Vektoren dieses Ortes und *link* Informationen zwischen den betreffenden *topics*(Wikipedia Begriffe) gewonnen. Anschließen wird aus dem *general user profiles* und der *local topic distribution* ein lokalisiertes Nutzerprofil erstellt. Dieses ist ein Vektor, welcher für jedes *topic* eine Wahrscheinlichkeit angibt, dass der Nutzer am gegebenen Ort an diesem *topic* interessiert ist. Das gleiche wird für ein lokalisiertes Newsprofil durchgeführt. Die Relevanz eins Artikels für einen Nutzer an einem gegebenen Ort wird dann anhand der Ähnlichkeit (cosine similarity) der lokalisierten Newsprofile zu einem lokalisierten Nutzerprofil gemessen. Die Empfehlung erfolgt dann durch eine top-k Liste mit den höchsten Relevanzwerten.

LP-ESA wird durch ein DNN zu LP-DSA erweitert, womit dem Problem der hohen Dimensionalität, Redundanz und *sparsity* bedingt durch den Wikipedia-*topic space* begegnet und eine schnelle online Empfehlung ermöglicht werden soll.

Als Eingabe dienen das *General user profile*, die *local topic distribution* und das *general news profile*. Das DNN erzeugt ein *abstract*, *dense* und *low dimensional feature space*, eine Art *Embedding* für das Nuterprofil, den Ort und Newsprofil, genannt *abstract general user profile*, *abstract local topic distribution*, und *abstract general news profiles*. Aus diesen wird anschließend ein *abstract localized user profile* und *abstract localized news profile* erzeugt, bei welchem die lokalisierten *similarities* zwischen Nutzern und den relevanten News maximiert und die von nicht relevanten News minimiert werden. Diesen DNN Teil bezeichnen die Autoren als Deep Semantic Analysis (DSA).

3.1.5 Attentive Collaborative Filtering: Multimedia Recommendation with Item- and Component-Level Attention

Chen u. a. (b) gehen davon aus, dass Nutzerpräferenzen in Multimediaempfehlungen durch implizites Feedback auf Item- und Komponentenebene betrachtet werden müssen. D.h. Items, welche ein Nutzer mag, haben nicht alle die gleiche Gewichtung, sodass einige Items eine höhere Präferenz vom Nutzer haben als andere. Weiterhin präferieren die Nutzer innerhalb eines Items unterschiedliche Komponenten, wie etwa unterschiedliche Ausschnitte eines Bildes oder verschiedene Abschnitte in einem Video.

Bei dem von den Autoren vorgestellte Modell Attentive Collaborative Filtering (ACF) handelt es sich um ein *latent factor model* mit integriertem *neighborhood-based model* und zwei DNNs, welche die zwei Ebenen von implizitem Feedback (Item- und Komponentenebene) in das *neighborhhod model* integrieren.

Dabei ist das Attention Modell ein Neural Network (NN) das aus zwei Modulen besteht. Zum einen das *component-level attention* Modul, welches auf einem beliebigen *content feature extraction network* wie CNN für Bilder oder Videos aufsetzen kann und lernt informative Komponenten der Items auszuwählen. Zum anderen das *item-level attention* Modul, welches lernt die Item Präferenz zu bewerten. Die beiden Attention Module liefern *weights* für die Items bzw. Komponenten der Items, diese Gewichte können wiederum in das *neighborhood* Modell integriert werden.

Um ihre Methoden zu evaluieren nutzen die Autoren zwei Datensätze von Sozialmedia Plattformen, zum einen von Pinterest für Bilddaten und zum anderen von Vine für Videodaten. Beide Datensätze sind öffentlich verfügbar.

3.1.6 Immersive Recommendation: News and Event Recommendations Using Personal Digital Traces

Hsieh u. a. stellen ein Immersive Recommendation genanntes Modell vor, bei welchem Nutzerprofile, Itemprofile und Ratings gemeinsam für Recommendations herangezogen werden. Die Arbeit befasst sich im speziellen mit personalisierter News und Event Recommendation, wofür die öffentlichen Twitter Accounts der Nutzer herangezogen werden. Darüber hinaus werden Datensätze von *Medium.com* und *Meetup.com* Nutzern verwendet. Zusätzlich wird das Modell in einer 33 Personen umfassenden Studie evaluiert, in welcher dessen Twitter, Facebook und Email *traces* verwendet werden.

Der Recommendation Prozess wird in zwei Phasen eingeteilt: *user profiling* und *recommendation*. In der *user profiling* Phase werden die Nutzerprofile, mit Hilfe der Informationen aus ihren *digital traces*, erstellt. Hierfür wird ein Channel(Context)-Aware Latent Dirichlet Allocation (CA-LDA) genanntes Verfahren eingesetzt, um die Themen mit denen sich ein Nutzer beschäftigt, aus den *digital traces* zu erschließen. Das Ergebnis ist eine *topic distribution* bzw. ein *topic embedding* pro Item, wobei CA-LDA mittels verschiedener Kontexte (Quelle des Items) in der Lage ist, *topics* die den Kontext betreffen und nicht die Nutzerinteressen widerspiegeln heraus zu filtern. Die Nuzerprofile werden dann aus aus mehreren nach Relevanz für den Nutzer gewichteten k-dimensionalen *topic embeddings* erstellt, welche aufsummiert werden, woraus sich pro Nutzer ein *topic embedding* ergibt, welches sein Interesse an verschiedenen Topics ausdrückt. Jedes Itemprofil entspricht einem aus seinem Content generierten *topic embedding*. In der *Recommendation* Phase wird ein hybrider Collaborative Filtering(CF) Algorithmus eingesetzt, welcher die User- und Itemprofile sowie Ratings verwendet um eine Empfehlung in Form eines vorhergesagten Ratings pro Item zu geben. Die Empfehlung berechnet sich aus der Ähnlichkeit von Nutzer- und Itemprofil sowie einem *latent offset* für das Nutzerprofi und das jeweilige Item. Dabei ist der *latent offset* ein k-dimensionaler Vektor, welcher durch CF aus den Ratings der Nutzer berechnet wird und zu dem *topic embedding* des Nutzers bzw. des Items addiert wird. Auf diese Weise wird die Relevanz der einzelnen Topics in den Embeddings in Richtung der sich aus den Ratings ergebenen Interessen verschoben.

Das *cold-start-problem* löst das Verfahren, indem bei fehlenden Ratings nur das Nutzer- und Itemprofil zur Empfehlung herangezogen wird. Aus den *topic embeddings* ergibt sich der Vorteil, dass diese im Gegensatz zu abstrakten durch DL Verfahren erzeugte Embeddings, für Menschen verständlich ist und die Empfehlung zum Teil erklärbar macht.

3.2 Einordnung & Abgrenzung der Arbeit

In den vorgestellten Arbeiten sind vier grundlegende Bereiche, die für die Modellbildung besonders relevant sind, auszumachen.

1. Die Content-Eigenschaften
2. Das Nutzerinteresse bzw. Verhalten und daraus resultierende Eigenschaften
3. Die Zeit, in Form sich ändernder Nutzerinteressen (ändernder Eigenschaften von neuem Content)
4. Seiteninformationen wie *knowledge bases* oder Sozial-Media Informationen

Die betrachteten Arbeiten setzen sich in der Regel mit zwei dieser Bereiche auseinander aber nicht mit allen vier. In den dieser Arbeit nach gelagerten Projekten, sollen möglichst alle diese Aspekte beleuchtet und erforscht werden, wie die verschiedenen Verfahren miteinander verknüpfen und kombinieren werden können.

Des Weiteren existieren keine Session-Daten von Nutzern, wie in einigen der vorgestellten Arbeiten. Zudem sind Nutzer im Kontext der DPA jeweils eine ganze Redaktion bzw. eine Zeitung oder ein Unternehmen, so dass Interessen nicht auf ein einzelnes Individuum bezogen werden können.

Weiterhin stehen viele Metainformationen, siehe Abschnitt 4.2, zur Verfügung, welche in den genannten Arbeiten bzw. üblichen Datensätzen nicht bzw. nicht in diesem Umfang zur Verfügung stehen. Hierdurch kann eine umfangreiche auf Content basierende Empfehlung erforscht werden.

4 Methodik

4.1 Vorgehen zur Beantwortung der Forschungsfrage

Um die Forschungsfragen zu klären, soll ein RS mit Hilfe der in Abschnitt 4.3 aufgeführten Software implementiert werden. Wie im vorherigen Abschnitt erwähnt, sollen Content-Eigenschaften, Nutzerinteresse, Zeitverhalten und Seiteninformationen berücksichtigt werden. Damit die Durchführung nicht von Anfang an eine zu große Komplexität aufweist, soll das Empfehlungsmodell sukzessive erweitert werden können. So sollen nicht alle zu testenden Verfahren gleichzeitig implementiert werden, sondern mit einfachen Verfahren begonnen und diese fortführend um aufwendigere erweitert werden.

Da zum jetzigen Zeitpunkt keine Nutzerdaten vorliegen, soll mit der Implementierung rein auf Content basierender Verfahren begonnen werden. Ein Ansatz wäre zunächst Verfahren zu implementieren, mit welchen Repräsentationen wie z.B. Embeddings der Artikel generiert werden können, mit denen die Ähnlichkeit verschiedener Artikel verglichen werden kann. Auch könnten Embeddings für Keywords erzeugt werden und so Artikel ausgemacht werden, welche

nicht mit einem bestimmten Keyword markiert sind , welches aber zu den Artikeln passen würden. Des Weiteren könnte die Zusammenstellung von Artikeln nach bestimmten Aspekten untersucht werden. Ein weiterer Punkt wäre ein System welches mit dem Nutzer interagiert und so Nutzerinformationen erhält. Beispielsweise wäre die Unterstützung eines Nutzers bei der Auswahl von Artikeln für ein Dossier oder ähnlichem denkbar. So könnte ein Nutzer eine Auswahl mehrerer Artikel angeben auf dessen Basis ähnliche oder zusammenhängende Artikel empfohlen werden.

Je nachdem ob Nutzerdaten auf absehbare Zeit beschafft werden können, sollen die genannten Punkte mehr oder weniger intensiv betrachtet werden. Sobald Nutzerdaten vorhanden sind bzw. die Evaluation alternative Nutzerdaten positiv ausfällt, sollen die bereits betrachteten Verfahren um weitere ergänzt werden.

Reine Content basierte Methoden und Verfahren die Nutzerpräferenzen einschließen können sich ergänzen, so können die für Artikel generierte Repräsentation sowohl für Ähnlichkeitsbestimmung der Artikel untereinander als auch zur Bestimmung ob Nutzerpräferenz und Artikel zusammenpassen genutzt werden.

Weiterhin können, bspw. mittels Attention-Mechanismus, Daten aus anderen Quellen in die Modelle eingebunden und diese so mit mehr Informationen angereichert werden. Denkbar wäre die Wikipedia Datenbank, welche viele Begriffe, Bilder und Tags dazu enthält.

Zudem können zu den Artikeln gehörende Bilder ebenfalls genutzt werden, um mehr Features für die Modellbildung zu nutzen.

4.2 Verwendete & Benötigte Daten

4.2.1 Vorhandene Daten

Es stehen über 200.000 hoch qualitative, von Journalisten geschriebene Artikel in deutscher Sprache im *NewsML-G2* Format IPTC (b) zur Verfügung. Dies bedeutet der Text weist wenige Rechtschreib- und Grammatikfehler auf und unterscheidet sich so schon von Datensätzen aus Quellen wie Twitter oder anderen Sozialmediadiensten. Der eigentliche Content steht als *plain text* sowie in Form von formatiertem HTML zur Verfügung. Neben einer kurzen Zusammenfassung enthält jeder News Artikel auch Metainformationen. Diese sind sehr Umfangreich und enthalten Informationen wie, Autor, Daten, Orte, Keywords und Events. Zudem sind die Artikel mit *PTC media topics* IPTC (a) versehen, welche eine Taxonomie mit über 1100 Begriffen darstellt, die in einer Baumstruktur organisiert sind und 17 Toplevel umfasst und eine Tiefe von 5 Leveln aufweist.

4.2.2 Noch erwartete Daten

Zu den bereits vorhandenen Daten werden zusätzlich passende Bilder zu den Artikeln erwartet, welche ebenfalls mit Metadaten im *NewsML-G2* versehen sind und zusätzliche Daten wie Bildunterschrift und abgebildete Personen enthalten.

Des Weiteren ist ein größerer Datensatz mit über 4 Millionen Artikeln plus passenden Bildern zu erwarten, sowie regelmäßige Updates mit neuen Artikeln.

4.2.3 Noch benötigte Daten

Um gute auf Nutzer zugeschnittene Empfehlungen berechnen zu können fehlen noch Nutzerdaten. Idealerweise wären Daten darüber, welcher (anonymisierter) Kunde welche Artikel genau gekauft hat. Um noch genauere Empfehlungen zu geben, wäre nützlich zu wissen, welche News tatsächlich verwendet wurden, etwa im Sinne von einer Quelle für einen Zeitungsartikel.

Alternative Nutzerdaten Sollte es nicht möglich sein Nutzerdaten direkt zu erhalten, wäre eine mögliche Alternative Zeitungsarchive zu durchsuchen und darauf zu prüfen, ob eine bestimme DPA News als Quelle diente bzw. die News direkt veröffentlicht wurde. Dies stellt allerdings einen erheblichen Aufwand dar oder es könnte wegen fehlender Verweise auf eine konkrete DPA News nicht möglich sein den Artikel direkt einer DPA News zuzuordnen.

Eine weitere Alternative wäre, Sozial Media Accounts der Nutzer (Zeitungen) heranziehen um so an Nutzerdaten zu gelangen. Dies wäre wahrscheinlich eine einfacher umzusetzende Möglichkeit als das Durchsuchen von Zeitungsarchiven, allerdings können diese Daten wohl nicht für implezites Feedback herangezogen werden.

4.2.4 Besonderheiten des Datensatzes

Eine für News allgemeine Besonderheit gegenüber Texten wie z.B. Wissenschaftlichen Artikeln ist, dass die Relevanz der Artikel sich stark mit der Zeit verändert, d.h. eine News von heute ist morgen evtl. schon veraltet und nicht mehr für einen Nutzer von Nutzen.

Eine weitere Besonderheit des DPA Datensatzes stellen die Metadaten und Bilder zu den News dar, die sonst nicht in öffentlichen Datensätzen zu finden sind.

Des Weiteren sind die Nutzer keine Endnutzer (außer DPA-Executive) sondern Zeitungen und Unternehmen, daraus ergibt sich dass kein implizites Feedback im herkömmlichen Sinne wie durch Klicks/Views vorhanden ist und es auch keine (Browser) Session History gibt.

4.3 Verwendete Software & Hardware

Das RS soll in Python entwickelt werden, wobei als grundlegendes Framework TensorFlow (TF) Abadi u. a. zum Einsatz kommen soll. TF ist ein machine learning Framework, welches pro Rechner mehrere GPUs und verteilte Berechnungen im Cluster unterstützt. Somit wird ermöglicht, entsprechende Hardware Ressourcen vorausgesetzt, sehr rechenintensive Berechnungen durchzuführen.

Das Hardware Setup für erste Tests besteht aus einer aktuellen Intel Core i7 4-Kern CPU, zwei Nvidia 1080 TI und 64 GB RAM. Für spätere Tests, die u.U. mehr Rechenleistung erfordern, ist ein Cluster Setup mit 32 CPU Kernen, 8 Nvidia P6000 Grafikkarten und 384 GM RAM verfügbar.

5 Forschungsrisiko

Folgende Risiken bestehen bei der Beantwortung der Forschungsfrage:

Keine Daten über Nutzer verfügbar Es ist nicht möglich Nutzerdaten zu erhalten. Hierdurch würden die Möglichkeiten eines RS erheblich eingeschränkt werden.

Zu geringe Menge an Nutzerdaten Wenn Nutzerdaten verfügbar sind, könnte es sein, dass diese nicht ausreichend sind um mit herkömmlichen Verfahren befriedigende Ergebnisse zu erzielen

„Real-World" Daten Die Daten wie sie vorliegen, sind kein speziell aufbereiteter Datensatz um damit Tests durchzuführen. Hieraus könnten sich Probleme, etwa bei der Datenextraktion, ergeben.

Informationsbeschaffung zu RS Anforderungen Wenn es nicht möglich ist, Anforderungen zu erheben, welche die DPA an ein RS stellen würde, könnte ein Entwurf für ein RS zu stark von einem tatsächlich einsetzbarem System abweichen.

Weiterhin sind folgende Frage bzw. Sachverhalt zu klären.

Wenn Nutzerdaten vorhanden sind, repräsentieren diese das tatsächliche Interesse einer Redaktion oder basieren diese nur auf Filterlisten nach bspw. Themen, welche dann redaktionsintern noch weiter gefiltert werden? Wenn dies nicht der Fall ist, wird es möglich sein, zu erkennen, welche News tatsächlich veröffentlicht wurden bzw. als Grundlage für einen Artikel dienten?

Wenn keine Nutzerdaten fürs Training vorliegen, sind auch keine Daten für die Messung der Performance, bzw. Evaluation des Systems vorhanden. Hieraus würde sich die Notwendigkeit ergeben, die eingesetzten Verfahren auf andere Weise zu evaluieren, wobei eine Möglichkeit in Tests oder Umfragen mit Personen besteht. Dies würde zusätzlich einen erheblichen Zeitaufwand zur Folge haben und vermutlich nicht genügen Daten liefern um eine wirklich aussagekräftige Evaluation der RS Verfahren zu ermöglichen.

5.1 Chancen

Zu den möglichen Forschungsrisiken ergeben sich allerdings auch Chancen. Da es sich bei dem DPA Datensatz um „Real-World" Daten handelt, bieten diese die Möglichkeit ein realitätsnahes RS zu entwickeln, welches sich nicht auf künstliche oder aufbereitete Daten verlässt. Zudem werden Metadaten in den meisten Arbeiten nicht umfänglich genutzt, bzw. stehen in nur kleinem Maße zur Verfügung. Hierdurch könnten neue Erkenntnisse gewonnen werden, wie effektiv zusätzliche Metadaten zur Erstellung von Empfehlungen genutzt werden können. Weiterhin sind News RS mit DL Verfahren relativ neu, woraus sich eine Vielzahl an Richtungen ergeben in die weiter geforscht werden kann.

6 Ausblick

Die langfristigen Ziele die durch nachfolgende Projekte erreicht werden sollen sind:

- Der Aufbau einer Pipeline zur Datenvorverarbeitung.
- Die Entwicklung einer Recommendation System Architektur.
- Die Implementation der RS Architektur und Integration in die Verarbeitungspipeline

Zunächst soll mit den vorhandenen DPA Daten die Pipeline für die Vorverarbeitung der Daten aufgebaut werden um dann erste Experimente mit den vorhandenen Daten durchzuführen. Hierbei sind dann nur Methoden die rein auf Content basieren zu evaluieren, da aktuell keine Nutzerdaten vorliegen.

Im weiteren Verlauf, sollen aufbauend auf den vorherigen Schritten, Nutzerdaten in das Modell mit einbezogen werden. Hierfür ist es essenziell, zunächst Nutzerdaten zu beschaffen, wobei natürlich vorrangig Nutzerdaten direkt von der DPA bevorzugt sind. Sollte es nicht möglich sein die Nutzerdaten auf direktem Weg zu erhalten, müssen die alternativen wie in 4.2 und 5 beschrieben erwägt und evaluiert werden.

Literatur

[Abadi u. a.] Abadi, Martín ; Barham, Paul ; Chen, Jianmin ; Chen, Zhifeng ; Davis, Andy ; Dean, Jeffrey ; Devin, Matthieu ; Ghemawat, Sanjay ; Irving, Geoffrey ; Isard, Michael ; Kudlur, Manjunath ; Levenberg, Josh ; Monga, Rajat ; Moore, Sherry ; Murray, Derek G. ; Steiner, Benoit ; Tucker, Paul ; Vasudevan, Vijay ; Warden, Pete ; Wicke, Martin ; Yu, Yuan ; Zheng, Xiaoqiang: TensorFlow: A System for Large-scale Machine Learning. In: *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*, USENIX Association (OSDI'16), S. 265–283. – URL http://dl.acm.org/citation.cfm?id=3026877.3026899. – Zugriffsdatum: 2018-08-14. – ISBN 978-1-931971-33-1

[Chen u. a. a] Chen, C. ; Meng, X. ; Xu, Z. ; Lukasiewicz, T.: Location-Aware Personalized News Recommendation With Deep Semantic Analysis. 5, S. 1624–1638

[Chen u. a. b] Chen, Jingyuan ; Zhang, Hanwang ; He, Xiangnan ; Nie, Liqiang ; Liu, Wei ; Chua, Tat-Seng: Attentive Collaborative Filtering: Multimedia Recommendation with Item- and Component-Level Attention. In: *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM (SIGIR '17), S. 335–344. – URL http://doi.acm.org/10.1145/3077136.3080797. – Zugriffsdatum: 2018-05-31. – ISBN 978-1-4503-5022-8

[Hsieh u. a.] Hsieh, Cheng-Kang ; Yang, Longqi ; Wei, Honghao ; Naaman, Mor ; Estrin, Deborah: Immersive Recommendation: News and Event Recommendations Using Personal

Digital Traces. In: *Proceedings of the 25th International Conference on World Wide Web*, International World Wide Web Conferences Steering Committee (WWW '16), S. 51–62. – URL https://doi.org/10.1145/2872427.2883006. – Zugriffsdatum: 2018-05-30. – ISBN 978-1-4503-4143-1

[Ian Goodfellow u. a.] Ian Goodfellow ; Yoshua Bengio ; Aaron Courville: *Deep Learning*. MIT Press. – URL https://www.deeplearningbook.org/. – Zugriffsdatum: 2018-09-21

[IPTC a] IPTC: *Media Topics*. – URL https://iptc.org/standards/media-topics/. – Zugriffsdatum: 2018-09-21

[IPTC b] IPTC: *NewsML-G2*. – URL https://iptc.org/standards/newsml-g2/. – Zugriffsdatum: 2018-09-21

[Okura u. a.] Okura, Shumpei ; Tagami, Yukihiro ; Ono, Shingo ; Tajima, Akira: Embedding-based News Recommendation for Millions of Users. In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM (KDD '17), S. 1933–1942. – URL http://doi.acm.org/10.1145/3097983.3098108. – Zugriffsdatum: 2018-05-20. – ISBN 978-1-4503-4887-4

[Song u. a.] Song, Yang ; Elkahky, Ali M. ; He, Xiaodong: Multi-Rate Deep Learning for Temporal Recommendation. In: *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM (SIGIR '16), S. 909–912. – URL http://doi.acm.org/10.1145/2911451.2914726. – Zugriffsdatum: 2018-05-20. – ISBN 978-1-4503-4069-4

[Wang u. a.] Wang, Xuejian ; Yu, Lantao ; Ren, Kan ; Tao, Guanyu ; Zhang, Weinan ; Yu, Yong ; Wang, Jun: Dynamic Attention Deep Model for Article Recommendation by Learning Human Editors' Demonstration. In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM (KDD '17), S. 2051–2059. – URL http://doi.acm.org/10.1145/3097983.3098096. – Zugriffsdatum: 2018-05-20. – ISBN 978-1-4503-4887-4