Konzept und Evaluierung eines Recommender Systems für eine Frage-Antwort-Plattform

Daniel Becker

Hochschule Darmstadt – Fachbereich Mathematik & Naturwissenschaften und Fachbereich Informatik

Problemstellung

Frage-Antwort-Plattformen haben im Internet eine sehr hohe Beliebtheit, da dort Benutzer bei einem Problem ihre Frage stellen können und schnell eine Antwort zur Lösung ihres Problems bekommen. Dies findet mittlerweile auch für Firmen als interne Lösung Interesse, um Mitarbeiter bei Problem schnell helfen zu können und eine interne Wissensdatenbank aufzubauen. In Zusammenarbeit mit der PRODYNA SE wurde für eine solche Plattform ein Konzept zur Implementierung eines Recommenders Systems erstellt, welche Fragen anhand ihres Inhaltes und Benutzer anhand ihres Interesses zusammenführen soll.

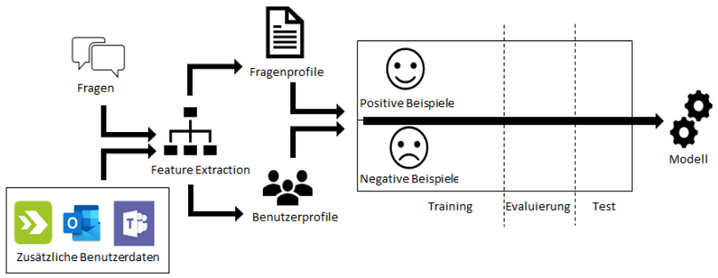

Abbildung 1: Konzept zur Implementierung des Recommender Systems

In der Abbildung 1 wird eine Gesamtübersicht des Konzepts gezeigt. Es wird zuerst ein Modell verwendet, welches Textinhalte vorverarbeitet und in einen Vektor transformiert und diese als Frageprofile verwendet. Anhand der beantworteten Fragen von Benutzer können daraus die Benutzerprofile ermittelt werden. Anschließend wird das Recommender System als Klassifizierungsmodell trainiert, welches Vorhersagen soll, ob eine Übereinstimmung zwischen einem Frage- und Benutzerprofil als Empfehlung besteht.

Daten

Zur Evaluierung des Systems wurden Fragen und Antworten der öffentliche Plattform Stack Overflow verwendet, sowie im Profil der Benutzer hinterlegten Webseiten als zusätzliche Benutzerdaten. Es wurden dabei drei verschiedene Datensätze erstellt, welche verschieden Szenarien einer Frage-Antwort-Plattform darstellen sollen.

Fragenprofile

Die Frageprofile sollen eine inhaltliche Repräsentation der Frage als Vektor darstellen. Für die Ermittlung der Frageprofile werden der Titel, Text und die hinterlegten Hashtags verwendet. Diese werden mit einer Vorverarbeitung (z.B. Kleinschreibung, Lemmatizierung oder entfernen von Stoppwörtern) zuerst bereinigt und anschließend mit einem weiteren Modell die Features extrahiert, um diese als Vektor darstellen zu können. Als Modelle wurden sowohl Tf-idf und verschiedene Varianten von Doc2Vec getestet.

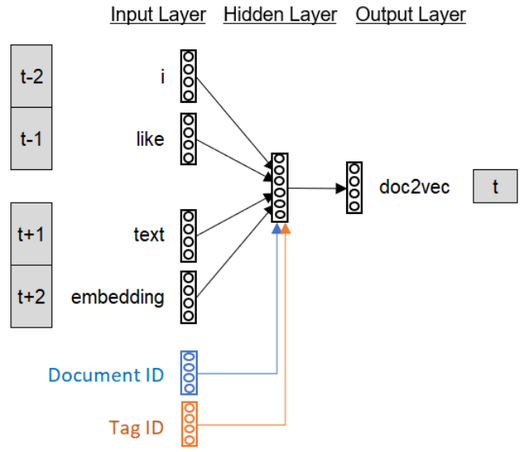

Abbildung 2: Doc2Vec PV-DM Modell mit zusätzlichen Trainingsvektor

Besonders die in Abbildung 2 verwendete Variante von Doc2Vec erzielte gute Ergebnisse. Dies baut auf dem Word2Vec Verfahren auf, welches im oberen Bereich des neuronalen Netzes abgebildet wird und die Word Embeddings trainiert. Durch die Verwendung des zusätzlichen Dokumentenvektor, entsteht das Doc2Vec Modell. Dies wurde noch ergänzt um einen Eingabevektor, welcher die hinterlegten Hashtags von Fragen besitzt, wodurch bei der Evaluierung des Recommender Systems eine höhere Genauigkeit erzielt werden konnte gegenüber der Standardvariante.

Benutzerprofile

Die Benutzerprofile sollen das Interesse eines Benutzers zu Themen darstellen in Form eines Vektors und ermitteln sich anhand der beantworteten Fragen eines Benutzers. Die einfachste Variante zur Berechnung des Benutzerprofils ist die Verwendung des arithmetischen Mittelwerts über alle Fragenprofile des Benutzers. Dies berücksichtigt allerdings ältere und jüngere beantwortete Fragen gleich. Um zu beachten, dass sich das Interesse eines Benutzers an Themen ändern kann und dementsprechend jüngere Fragen stärker berücksichtigt werden sollen, wurde ein exponentieller Zerfall in Kombination mit dem gewichteten arithmetischen Mittelwert verwendet, wodurch ältere Fragen weniger gewichtet werden für das Benutzerprofil aber dennoch nicht komplett vergessen werden. In Abbildung 3 ist dies an einem Beispiel gezeigt, indem der jeweilige Fragevektor (1) mit seinem entsprechenden Gewicht (2) multipliziert wird, die gewichteten Werte pro Dimension summiert werden (3), um daraus das Benutzerprofil (4) zu berechnen.

Original Werte

t	Tok1	Tok2	Tok3
4	0.600	0.100	0.200
3	0.500	0.200	0.450
2	0.550	0.550	0.450
1	0.550	0.550	0.400
0	0.150	0.650	0.300
Summe			
	(1)		

x

f(v)
0.162
0.179
0.198
0.219
0.242
1.000
(2)

=

Gewichtete Werte

Tok1	Tok2	Tok3
0.097	0.016	0.032
0.090	0.036	0.081
0.109	0.109	0.089
0.120	0.120	0.088
0.036	0.157	0.073
0.452	*0.439*	*0.362*
(3)		

sum =>

Benutzerprofil

Tok1	Tok2	Tok3
0.452	0.439	0.362
(4)		

Abbildung 3: Berechnung Benutzerprofil mit gewichtetem arithmetischem Mittelwert

Ergebnisse

In dieser Arbeit wurde ein Recommender System vorgestellt, welches Benutzer auf einer Frage-Antwort-Plattform unterstützen soll mit Empfehlungen von Fragen, welche sie interessieren könnte zum Beantworten. Die Ergebnisse der Evaluierung zeigen, dass das Doc2Vec Verfahren eine gute Möglichkeit bietet, um die Fragen als Vektoren darzustellen. Vor allem mit der Verwendung der Hashtags als zusätzlicher Trainingsvektor konnte eine höhere Genauigkeit bei den Empfehlungen erreicht werden. Auch die Verwendung des gewichteten arithmetischen Mittelwertes und einem exponentiellen Zerfall für die Bestimmung der Benutzerprofile ist eine gute alternative, um besser das aktuelle Benutzerinteresse zu berücksichtigen. So konnte dadurch gegenüber dem Tf-idf Modell und dem normalen arithmetischen Mittelwert als Ausgangsbasis, eine Steigerung der Genauigkeit von bis zu 5,79 Prozentpunkte erzielt werden.

Ausblick

Die Evaluierung des Recommender Systems fand in dieser Arbeit auf einem extrahierten offline Datensatz statt. In einem nächsten Schritt sollte das Recommender System aktiv gesetzt werden in die bestehende Frage-Antwort-Plattform und als Experiment Online weiter beobachtet werden. Dadurch wird es möglich zu evaluieren, wie gut dies auch in einem Echtzeitsystem funktioniert und es können weitere Erkenntnisse erlangt werden, um die Empfehlungen weiter zu optimieren. Auch die Verwendung von zusätzlichen Features (z.B. Qualität der Antwort, Aktivitätenlevel oder Motivation des Benutzers), welche in bekannte Ansätze verwendet werden, können sinnvoll sein, um das Recommender System um neue Fragestellungen zu erweitern und herauszufinden, welche Benutzer gute und/oder schnelle Antworten geben.