ABSTRACT

Community question answering are very popular on the Internet, where users can ask their questions and get a quick answer to solve their problem. This is also of interest for companies as an internal solution to help employees quickly with problems and to build up an internal knowledge database. By using a recommender system, the existing system should be supported to match questions based on their content and users based on their interest.

In this thesis a concept to implement a recommender system for a community question answering forum for the company PRODYNA SE will be presented. Different approaches to modify and optimise the Recommender System will be used. Three different data sets will be applied in order to consider different use cases and to evaluate the recommender system.

Using Doc2Vec is a good way to present the text of the used questions as vectors. Compared to common methods such as BoW or Tf-idf, this has the additional advantage that a smaller vector is needed for a question, which also results in a faster runtime for further calculations.

The calculation of user profiles, which are used to match questions and users, is limited to the interest of the user in this thesis. Since the interest of a user can change over time, approaches for this behavior are presented and their effects are evaluated. By using a weighted arithmetic mean for the user profiles, which looks more at recent answered questions, and by using a Doc2Vec model for the question profiles, an increase in accuracy of up to 5.79 percentage points is possible compared to the original Tf-idf model and the normal arithmetic mean.

Keywords: *Natural Language Processing, Recommender, Doc2Vec, Similarity, Stack Overflow, Community-Question-Answering*

KURZFASSUNG

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. Durch die Verwendung eines Recommender Systems, soll das bestehende System dabei unterstützt werden, dass Fragen anhand ihres Inhaltes und Benutzer anhand ihres Interesse zusammengeführt werden.

In dieser Arbeit wird ein Konzept zur Realisierung eines Recommender Systems für eine Frage-Antwort-Plattform der Firma PRODYNA SE erstellt. Dabei werden auf die verschiedenen Phasen zu Erstellung des Recommender Systems eingegangen und Ansätze vorgestellt, um in diese einzugreifen und zu optimieren. Zur Berücksichtigung von verschiedenen Anwendungsszenarien werden drei unterschiedliche Datensätze verwendet, anhand den eine Evaluierung des Recommender Systems vorgenommen wird.

Die Verwendung von Doc2Vec ist eine gute Möglichkeit um die Texte der verwendeten Fragen als Vektor darzustellen. Gegenüber verbreiteten Varianten wie BoW oder Tf-idf entsteht dadurch zusätzlich der Vorteil, dass eine kleinerer Vektor benötigt wird für eine Frage, wodurch auch eine schnelle Laufzeit bei weiteren Berechnungen entsteht.

Bei der Ermittlung von Benutzerprofile, welche zur Übereinstimmung zwischen Fragen und Benutzer dienen, wird sich in dieser Arbeit auf das Interesse des Benutzers beschränkt. Da sich das Interesse eines Benutzer über die Zeit verändern kann, werden für dieses Verhalten Ansätze vorgestellt und deren Auswirkung evaluiert. Durch die Verwendung eines gewichteten arithmetischen Mittelwert für die Benutzerprofile, welches kürzlich beantwortete Fragen stärker betrachtet und der Verwendung eines Doc2Vec Modell für die Frageprofile ist eine Steigerung der Genauigkeit von bis zu 5.79 Prozentpunkte möglich gegenüber dem ursprünglichen Tf-idf Modell und dem normalen arithmetischen Mittelwert.

Schlagwörter: *Natural Language Processing, Empfehlungssystem, Doc2Vec, Ähnlichkeiten, Stack Overflow, Frage-Antwort-Plattform*