УДК 004.896

I.V. Bogach
V.A.Kovenko

RECOMMENDATION SYSTEM BASED ON NLP TECHNIQUES

Вінницький національний технічний університет;

Анотація

Розкрито базовий підхід до рекомендаційних систем на основі контенту. Використання Word2Vec, зроблених Google, є неспроможним. Розглянуто можливість використання додаткової логіки побудови

Ключові слова: NLP techniques, Word2Vec, CountVectorizer, cosine similarity, embedding, content based system, content based recommendation.

Abstract

The benchmark approach to content based recommendation systems is exposed in this article. The usage of Word2Vec embeddings made by Google is unleashed. The opportunity of using additional business logic is considered.

Keywords: NLP techniques, Word2Vec, CountVectorizer, cosine similarity, embedding, content based system, content based recommendation.

Introduction

Understanding users' preferences and proposing them the most relevant products is essential for every commercial business which involves the process of interacting with users. As nowadays the web infrastructure is developing rapidly, lots of commercial activities move to the space of the internet. Thus, the demand of recommendation systems arises. Recommendation system is an engine which goal is to recommend relevant items to users. Many world famous companies like Netflix, Amazon, YouTube, etc., use them to attract more people to their websites and increase their income. The recommendation systems can be divided into two groups: content based[1] and user based[2]. Content based recommendation systems focus on the content, its taxonomy and metadata for making predictions, while the user based ones require user interactions like clicks or ratings the user left for items. Nevertheless, user based recommendation systems are much more powerful than content based, they require lots of computational power that can afford working with big data. On the opposite, when building a content based recommendation system, the one is interested only in the catalog of items, and as a rule the number of items is always smaller than the number of users in the system. Content based recommendation system is a nice start for a small company that just appeared on the market of web products. With the developing of the sphere of Natural Language Processing (NLP), the new opportunities for content based recommendation systems appeared. The new approach to recommendation systems is proposed in this article, the problem is stated as measuring the similarities between items' metadata and is addressed as NLP task. The system uses hybrid algorithm based on counting words in a statement and using Word2Vec model provided by Google. The possibility of using additional business logic is considered. Finally, the results are viewed with respect to Movielens dataset.

Data Preparation

Constructed system was validated using famous Movielens[1] dataset, which contains catalog of movies, their metadata and intersections of users with a catalog. The columns that were used are the following: title, movie_id and genres.

The dataset was also enriched with IMDB data for the aim of expanding information about content.

Data Transformation

The MovieLens data is made of strings which describe items' metadata, but for an algorithm to work the transformation of relative columns to the matrix of numbers is needed. For this purpose a hybrid transformation, made of CountVectorizer[3] and Word2Vec[4] model, is used. CountVectorizer simply counts the frequency of each word's appearance. It's then replaces the word with the corresponding frequency.

Word2Vec is a type of neural network which is trained to find the similarity vectors between words. If words often appear in the same context, their vectors will be similar. The Word2Vec vectors are retrieved for each word in the sentence and then averaged to get the semantic vector of the sentence. CountVectorizer is used for fields without semantic meaning, whereas Word2Vec for those that have it.

Algorithm

The algorithm is based on a cosine similarity, that is a mere choice for NLP tasks (1).

$$k(x,y) = \frac{xy^{\top}}{\|x\|\|y\|} \tag{1}$$

The cosine similarity is computed between each column of each item in the dataset, thus for each item we have a matrix of similarities with others by a particular column. Because of the fact that one data column regarding an item can be much more important for the final recommendation than the other one, the additional business logic is added. The additional set of columns' weights that can be configured manually was added. Each attribute/column matrix is multiplied by related weight parameter, that gives an opportunity to decrease/increase contribution of it to the final similarity calculation. This makes the overall system more flexible and extendable towards new logic. Finally, the similarity matrices by columns are averaged to produce a final similarity matrix for an item, Then, top N recommendations can be retrieved using a similarity score.

Summary and further work

Nevertheless, the proposed system is not based on user activity it has lots of advantages. It's flexible, not power consuming, easy to extend and flexible. Addressing the issue of recommendation systems as an NLP task, gives a lift to usage of novel NLP techniques like BERT[5]. On the opposite side, it's much less powerful than user based system and is heavily dependent on the quality of catalog metadata. The particular system can also be used in ensemble with user based recommendation system to construct a hybrid system. To sum up, the proposed algorithm can be used as an alternative to user based system and is an adequate choice for companies which just started their activity on web market.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Content-Based Recommendation System [Electronic reourse] – Electronic data. – Mode of access: https://www.researchgate.net/publication/236895069_Content-Based_Recommendation_Systems – Title from the screen.

2. Recommendation Systems: User-based Collaborative Filtering using N Nearest Neighbors. [Electronic resourse] – Electronic data. – Mode of access: https://medium.com/sfu-big-data/recommendation-systems-user-based-collaborative-filtering-using-n-nearest-neighbors-bf7361dc24e0 – Title from the screen.

3. 10+ Examples for Using CountVectorizer [Electronic resourse] – Electronic data. – Mode of access: https://kavita-ganesan.com/how-to-use-countvectorizer/#.XlF2wHUzaV4 – Title from the screen.

4. Distributed Representations of Words and Phrases and their Compositionality [Electronic resourse] – Electronic data. – Mode of access: https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf – Title from the screen.

5. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [Electronic resourse] – Electronic data. – Mode of access: https://arxiv.org/abs/1810.04805 – Title from the screen.

Богач Ілона Віталіївна, кандидат технічних наук, доцент кафедри автоматизації та інтелектуальних інформаційних технологій, Вінницький національний технічний університет, ilona.bogach@gmail.com.

Ковенко Володимир Андрійович, студент групи 1ІСТ-18б, Факультет комп'ютерних систем та автоматики, Вінницький національний технічний університет, urumipainblackreaper@gmail.com.

Bogach Ilona Vitalyevna, PhD, Associate Professor of the department of automation and intelligent information technologies, Vinnytsia National Technical University, ilona.bogach@gmail.com.

Ковенко Володимир Андрійович, the student of group 1IST-18b, the faculty of computer systems and automation, Vinnytsia National Technical University, urumipainblackreaper@gmail.com.