Item2vec: Neural Item Embedding for Collaborative Filtering

Oren Barkan
Microsoft, Israel

Noam Koenigstein
Microsoft, Israel

ABSTRACT

Many Collaborative Filtering (CF) algorithms are item-based in the sense that they analyze item-item relations in order to produce item similarities. Recently, several works in the field of Natural Language Processing (NLP) suggested to learn a latent representation of words using neural embedding algorithms. Among them, the Skip-gram with Negative Sampling (SGNS), also known as Word2vec, was shown to provide state-of-the-art results on various linguistics tasks. In this paper, we show that item-based CF can be cast in the same framework of neural word embedding. Inspired by SGNS, we describe a method we name Item2vec for item-based CF that produces embedding for items in a latent space. The method is capable of inferring item-item relations even when user information is not available. We present experimental results that demonstrate the effectiveness of the Item2vec method and show it is competitive with SVD.

1. INTRODUCTION

Computing item similarities is a key building block in modern recommender systems. While many recommendation algorithms are focused on learning a low dimensional embedding of users and items simultaneously [1], computing item similarities is an end in itself.

There are several scenarios where item-based CF methods [2] are desired: in a large scale dataset, when the number of users is significantly larger than the number of items, the computational complexity of methods that model items solely is significantly lower than methods that model both users and items simultaneously. For example, online music services may have hundreds of millions of enrolled users with just tens of thousands of artists (items).

Recent progress in neural embedding methods for linguistic tasks have dramatically advanced state-of-the-art natural language processing (NLP) capabilities [3, 4]. These methods attempt to map words and phrases to a low dimensional vector space that captures semantic and syntactic relations between words. Specifically, Skip-gram with Negative Sampling (SGNS), known also as word2vec [4], set new records in various NLP tasks [4].

In this paper, we propose to apply SGNS to item-based CF. Motivated by its great success in other domains, we suggest that SGNS with minor modifications may capture the relations between different items in collaborative filtering datasets. To this end, we propose a modified version of SGNS named Item2vec. We show that Item2vec can induce a similarity measure that is competitive with an item-based CF using SVD.

2. ITEM2VEC

SGNS is a neural word embedding method that was introduced by Mikolov et. al in [4]. The method aims at finding words representation that captures the relation between a word to its surrounding words in a sentence.

In the context of CF data, the items are given as user generated sets. The application of SGNS to CF data is straightforward once we realize that a sequence of words is equivalent to a set or basket of items. Since we ignore the spatial information, we treat each pair of items that share the same set as a positive example. Therefore, for a given set of items $\{w_i\}_{i=1}^K \subseteq W$, we aim at maximizing the following term:

$$\frac{1}{K}\sum_{i=1}^{K}\sum_{j\neq i}^{K}\log\left(\sigma(u_i^T v_j)\prod_{k=1}^{N}\sigma(-u_i^T v_k)\right) \quad (1)$$

where $u_i \in U(\subset \mathbb{R}^m)$ and $v_i \in V(\subset \mathbb{R}^m)$ are latent vectors that correspond to the target and context representations for the item $w_i \in W$, respectively. $\sigma(x) = 1/1+\exp(-x)$, m is chosen empirically and according to the size of the dataset and N is a parameter that determines the number of negative examples to be drawn per a positive example. A negative item w_i is sampled from the unigram distribution raised to the 3/4rd power.

In order to overcome the imbalance between rare and frequent items, we subsample the data [4]. Specifically, we discard each item w from its set, with a probability $p(discard \mid w) = 1-\sqrt{\rho / f(w)}$ where $f(w)$ is the frequency of the item w and ρ is a prescribed threshold.

U and V are estimated by applying stochastic gradient ascent with respect to the objective in (1). Finally, we use u_i as the representation for the i-th item and the affinity between a pair of items is computed by the cosine similarity.

3. EXPERIMENTAL RESULTS

In this section, we provide qualitative and quantitative results. As a baseline item-based CF algorithm we used item-item SVD. Specifically, we apply SVD to decompose $A = USV^*$, where A is a square matrix in size of number of items. The (i, j) entry in A contains the number of times (w_i, w_j) appears as a positive pair in the dataset, normalized by the square root of the product of its row and column sums. The latent representation is given by the rows of $U_m S_m^{1/2}$. The affinity between items is computed by cosine similarity of their representations.

We evaluate the methods on two different private of datasets. The first dataset consists of user-artist relations that are retrieved from the Microsoft Xbox Music service. This dataset consist of 9M events. Each event consists of a user-artist relation, which means the user played a song by the specific artist. The dataset contains 732K users and 49K distinct artists.

The second dataset contains orders of products from Microsoft Store. An order is given as a basket of items without any information about the user that made it. Therefore, the information in this dataset is weaker in the sense that we cannot bind between users and items. The dataset consists of 379K orders (that contains more than a single item) and 1706 distinct items.

We applied Item2vec and SVD to both datasets. The dimension parameter was set to $m = 40$. We ran item2vec on

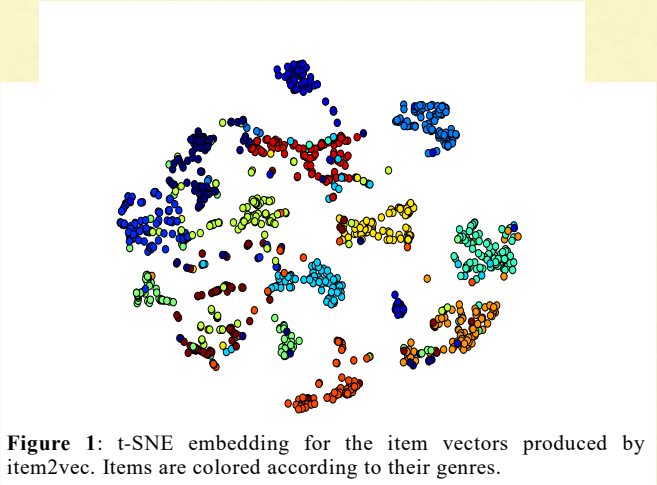

Figure 1: t-SNE embedding for the item vectors produced by item2vec. Items are colored according to their genres.

both datasets for 20 epochs with negative sampling value $N = 15$. We further applied subsampling with ρ values of 1e-5 and 1e-3 to the Music and Store datasets, respectively. The reason we set different parameter values is due to different sizes of the datasets.

The music dataset does not provide genre metadata. Therefore, for each artist we retrieved the genre metadata from the web to form a genre-artist catalog. Then we used this catalog in order to visualize the relation between the learnt representation and the genres. This is motivated by the assumption that a useful representation would cluster artists according to their genre. To this end, we generated a subset that contains the top 100 popular artists per genre for 13 distinct genres. We applied t-SNE [5] with a cosine kernel to reduce the dimensionality of the item vectors to 2. Then, we colored each artist point according to its genre. Figure 1 presents the 2D embedding that was produced by t-SNE, for item2vec. We observe that some of the relatively homogenous clusters in Fig. 1 are contaminated with items that are colored differently. We found out that many of these cases originate in artists that are mislabeled in the web or have a mixed genre.

In order to quantify the similarity quality, we tested the genre consistency between an item and its nearest neighbors. We do that by iterating over the top q popular items (for various values of q) and check whether their genre is consistent with the genres of the k nearest items that surround them. This is done by a simple majority voting. Table 1 presents the results obtained for $k = 8$. We observe that item2vec is consistently better than the SVD model, where the gap between the two keeps growing as q increases. This might imply that item2vec produces a better representation for less popular items than the one produced by SVD. We further validate this hypothesis by applying the same 'genre consistency' test to a subset of 10K unpopular items (the last row in Table 1). We define an unpopular item in case it has less than 15 users that played its corresponding artist. The accuracy obtained by item2vec was 68%, compared to 58.4% by SVD.

TABLE 1: A COMPARISON BETWEEN SVD AND ITEM2VEC ON GENRE CLASSIFICATION TASK FOR VARIOUS SIZES OF TOP POPULAR ARTIST SETS

Top (q) popular artists	SVD accuracy	Item2vec accuracy
2.5k	85%	**86.4%**
5k	83.4%	**84.2%**
10k	80.2%	**82%**
15k	76.8%	**79.5%**
20k	73.8%	**77.9%**
10k unpopular (see text)	58.4%	**68%**

Qualitative comparisons between Item2Vec and SVD are presented in Tables 2-3 for Music and Store datasets, respectively. The tables present seed items and their 4 nearest neighbors (in the latent space). The main advantage of this comparison is that it enables the inspection of item similarities in higher resolutions than genres. Moreover, since the Store dataset lacks any informative tags / labels, a qualitative evaluation is inevitable. We observe that for both datasets, Item2Vec provides lists that are better related to the seed item than the ones that are provided by SVD. Furthermore, we see that even though the Store dataset contains weaker information, Item2Vec manages to infer item relations quite well.

In future we plan to investigate more complex CF models [1] and compare between them and item2vec.

4. REFERENCES

[1] Koren Y, Bell R, Volinsky C. Matrix factorization techniques for recommender systems. Computer. 2009 Aug 1(8):30-7.

[2] Linden G, Smith B, York J. Amazon.com recommendations: Item-to-item collaborative filtering. Internet Computing, IEEE. 2003 Jan;7(1):76-80.

[3] Mnih A, Hinton GE. A scalable hierarchical distributed language model. In Proceedings of NIPS 2009 (pp. 1081-1088).

[4] Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In Proceedings of NIPS 2013 (pp. 3111-3119).

[5] Van der Maaten, L., & Hinton, G. Visualizing data using t-SNE. Journal of Machine Learning Research, (2008) 9(2579-2605), 85.

TABLE 2: A QUALITATIVE COMPARISON BETWEEN ITEM2VEC AND SVD FOR SELECTED ITEMS FROM THE MUSIC DATASET

Seed item (genre)	Item2vec – Top 4 recommendations	SVD – Top 4 recommendations
David Guetta (Dance)	Avicii ,Calvin Harris, Martin Solveig, Deorro	Brothers, The Blue Rose, JWJ, Akcent
Katy Perry (Pop)	Miley Cyrus, Kelly Clarkson, P!nk, Taylor Swift	Last Friday Night, Winx Club, Boots On Cats, Thaman S.
Dr. Dre (Hip Hop)	Game, Snoop Dogg, N.W.A, DMX	Jack The Smoker, Royal Goon, Hoova Slim, Man Power
Johnny Cash (Country)	Willie Nelson, Jerry Reed, Dolly Parton, Merle Haggard	Hank Williams, The Highwaymen, Johnny Horton, Hoyt Axton
Guns N' Roses (Rock)	Aerosmith, Ozzy Osbourne, Bon Jovi, AC/DC	Bon Jovi, Gilby Clarke, Def Leppard, Mtley Cre
Justin Timberlake (Pop)	Rihanna, Beyonce, The Black eyed Peas, Bruno Mars	JC Chasez. Jordan Knight, Shontelle, Nsync

TABLE 3: A QUALITATIVE COMPARISON BETWEEN ITEM2VEC AND SVD FOR SELECTED ITEMS FROM THE STORE DATASET

Seed item	Item2vec – Top 4 recommendations	SVD – Top 4 recommendations
LEGO Emmet	LEGO Bad Cop, LEGO Simpsons: Bart, LEGO Ninjago, LEGO Scooby-Doo	Minecraft Foam, Disney Toy Box, Minecraft (Xbox One), Terraria (Xbox One)
Minecraft Lanyard	Minecraft Diamond Earrings, Minecraft Periodic Table, Minecraft Crafting Table, Minecraft Enderman Plush	Rabbids Invasion, Mortal Kombat, Minecraft Periodic Table
GoPro LCD Touch BacPac	GoPro Anti-Fog Inserts, GoPro The Frame Mount, GoPro Floaty Backdoor, GoPro 3-Way	Titanfall (Xbox One), GoPro The Frame Mount, Call of Duty (PC), Evolve (PC)
Surface Pro 4 Type Cover	UAG Surface Pro 4 Case, Zip Sleeve for Surface, Surface 65W Power Supply, Surface Pro 4 Screen Protection	Farming Simulator (PC), Dell 17 Gaming laptop, Bose Wireless Headphones, UAG Surface Pro 4 Case
Disney Baymax	Disney Maleficent, Disney Hiro, Disney Stich, Disney Marvel Super Heroes	Disney Stich, Mega Bloks Halo UNSC Firebase, LEGO Simpsons: Bart, Mega Bloks Halo UNSC Gungoose
Windows Server 2012	Windows Server Remote Desktop Services 1-User, Exchange Server 5-Client, Windows Server 5-User Client Access, Exchange Server 5-User Client Access	NBA Live (Xbox One) – 600 points Download Code, Windows 10 Home, Mega Bloks Halo Covenant Drone Outbreak, Mega Bloks Halo UNSC Vulture Gunship