Recommendation System : Leveraging Heterogeneous Information Network

Aakash Patel, Shadab Ahmed, Debarchan Basu, Ugesh Vayugundla[1]

Abstract

Recommendation Systems are essentials for online as well as offline marketing. Traditional approach for recommendation system is non-negative matrix factorization. Heterogeneous information network (HIN) provides flexibility in modeling data heterogeneity, therefor it has been adopted to characterize complex and heterogeneous auxiliary data in recommendation systems. In this project, we are embedding HIN via flexible regularization framework into the recommendation system. It is a matrix factorization-based dual regularization framework which is proposed to flexibly integrate different types of information through adopting users' and items' similarities as regularization on latent factors of users and items.

1. Problem Description

Recently, there is a surge of social recommendation, which leverages social relations among users to improve recommendation performance. However, in many applications, social relations are very sparse or absent. Meanwhile, the attribute information of users or items may be rich. It is challenging to develop effective methods for HIN based recommendation in both extraction and exploitation of the information from HINs. So in this project, our goal is to design a recommendation system which takes additional information about objects and use it to improve the performance of recommendation system.

2. Literature Survey

We did some background study related to recommendation systems and heterogeneous information network.

2.1. Collaborative Filtering

Collaborative filtering (CF) uses the known preferences of a group of users to make recommendations or predictions of the unknown preferences for other users. In CF, the assumption is that if a person X behaves (buy, watch) like another person Y, then he will rate or act on other items similar to Y. There are many problems associated with collaborative filtering like cold start, highly sparse data, privacy, etc. There are two main categories of CF algorithms : memory based, model based. Early generation collaborative filtering systems used user ratings to calculate similarity between different users or different items. And then they use these similarity measures to recommend items to users. This is memory based technique of recommendation system. Since similarity values are based on common items, this technique becomes unreliable when there a lots of items and data is sparse. To overcome these problems, model based CF approaches were used like Markov decision process, etc.

2.2. Non-negative Matrix Factorization

non-negative matrix approximation is a group of algorithms where matrix R is factorized into (usually) two matrices U and V, with the property that all three matrices have no negative elements. User-item rating data can be represented as a matrix where rows represent users, columns represent items and rating is put on the corresponding cell. We can factorize this sparse matrix R into two, such that their multiplication is as close to matrix R as possible for the given entries of R.

$$M_{m*n} \approx W_{M*k} H_{k*n}$$

Where m is number of users, n is number of items and k is a hyper parameter that can be tuned to give best approximation.

To compute matrix U and V, we can choose to reduce root mean square error (RMSE) for the given entries. Hence our optimization function will be following:

$$\begin{aligned} \min_{U,V} L(R,U,V) = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n} I_{ij}(R_{ij} - U_i V_j^T) \\ + \frac{\lambda_1}{2}\|U\|^2 + \frac{\lambda_2}{2}\|V\|^2 \end{aligned} \quad (1)$$

where Matrix I is indicator matrix for available entries in R. Above is regularized objective function for matrix factorization. Adding constraint of U and V to not become negative will be non-negative matrix factorization

[1] Indian Institute of Science. Correspondence to: All authors <aakashpatel@iisc.ac.in, shadabahmed@iisc.ac.in>.

2.3. Heterogeneous Information Network

A heterogeneous information network (HIN) is a special type of information network with underneath data structure as a directed graph, which contains either multiple types of objects or multiple types of links. Specifically, given a schema $S = (A, R)$ which consists of a set of entity types $A = \{A\}$ and a set of relations $R = \{R\}$, an information network is defined as a directed graph $G = (V, E)$ with an object type mapping function $\varphi : V \rightarrow A$ and a link type mapping function $\psi : E \rightarrow R$. If types of objects $\mid A \mid > 1$ or types of relations $\mid R \mid > 1$, the network is called heterogeneous information network.

2.3.1. HIN regularized framework

Different links in network represent different relations between entities. Under HIN based representation, recommendation task can be considered as similarity search over HIN. In HIN based recommendation, we exploit meta-path based similarity. Based on traditional matrix factorization, a dual regularization framework SimMF is proposed to integrate heterogeneous information through adopting similarity information of users and items as regularization on latent factors of users and items. In this approach, two different regularization models, average- and individual-based regularization, can flexibly be used as regularization on users or items. We can define $S_{ij}^{(l)}$ to denote the similarity of two objects u_i and u_j under the given meta-path $\rho^{(l)}$. The similarity(S) is determined by the given meta-path (ρ) and the similarity measure (M). That is S = ρ M. Since similarity measure of different paths are different, therefor they need to be normalized. We can do this by sigmoid function.

$$S_{ij}^{(l)} = \frac{1}{1 + e^{S_{ij}^{(l)} - \bar{S}_{ij}^{(l)}}} \tag{2}$$

Since users (or items) have different similarity under different meta-paths, we consider their similarity on all paths through assigning weights on different paths. we define S^U for the similarity matrix of users on all paths, and S^I for the similarity matrix of items on all paths.

$$S^U = \sum_l W_l^U S^{(l)} \tag{3}$$

$$S^I = \sum_l W_l^I S^{(l)} \tag{4}$$

where W_l^U represents the weight of similarity matrix of users under the path ρ_l and W_l^I represents that of items. Also $\sum_l W_l^U = 1$; $0 \leq W_l^U \leq 1$ and $\sum_l W_l^I = 1$; $0 \leq W_l^I \leq 1$.

Now , we can incorporate this HIN into our traditional matrix factorization model by introducing similarity regularization.

$$Reg^U = \sum_{i=1}^{m} \sum_{j=1}^{m} S_{ij}^U \| U_i - U_j \|^2 \tag{5}$$

$$Reg^I = \sum_{i=1}^{m} \sum_{j=1}^{m} S_{ij}^I \| V_i - V_j \|^2 \tag{6}$$

then our final objective function becomes

$$\begin{aligned} \min_{U,V} L(R, U, V) = \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{n} I_{ij} (R_{ij} - U_i V_j^T) + \frac{\alpha}{2} Reg^U \\ + \frac{\beta}{2} Reg^I + \frac{\lambda_1}{2} \|U\|^2 + \frac{\lambda_2}{2} \|V\|^2 \end{aligned} \tag{7}$$

There is a problem associated with some existing HIN methods. These methods rely on explicit meta-path based reachability. These paths may not be reliable because of noise. It is possible that some paths may have formed accidentally. Hence we are left with two fundamental issues and those are effective information extraction and information exploitation. These issues can be tackled by HIN embedding. We will read the paper and implement HIN embedding in our upcoming work.

3. Dataset and Results

Stemming from the business domain, the widely used Yelp challenge dataset3 [26,28] records users' ratings on local business and also contains social relations and attribute information of business (e.g., cities and categories).
Various relations that are extracted are shown in below:

Relation type	Relations	#relations
Rating	User–business	194,255
Social relation	User–user	150,532
Attribute of business	Business–category	39,406
	Business–location	14,037

3.1. Result

Table 1. Comparison of Performance(in RMSE) on Yelp Data set

Training	User Mean	NMF	NMFR
80%	1.1278	1.0964	1.0904
60%	1.1388	1.1011	1.0908
40%	1.1581	1.1207	1.1163
20%	1.2068	1.1629	1.1596

NMF:Non negative Matrix Factorization
NMFR:Non negative Matrix Factorization with Regularization

References

C. Luo, W. Pang, Z. W. and Lin, C. Hete-cf: Social-based collaborative filtering recommendation using heterogeneous relations. In *ICDM, pages 917-922*, 2014.

Chuan Shi, e. a. Semantic path based personalized recommendation on weighted heterogeneous information networks. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, 2015.

.Chuan Shi, e. a. Integrating heterogeneous information via flexible regularization framework for recommendation. In *Springer-Verlag London*, 2016.

Daniel D. Lee, H. S. S. Algorithms for non-negative matrix factorization. In *14th Annual Neural Information Processing Systems Conference, NIPS 2000 - Denver*, 2000.

Jamali, M. and Lakshmanan, L. V. Heteromf:recommendation in heterogeneous information networks using context dependent factor models. In *WWW, pages 643-653*, 2013.

Lin, C. J. Projected gradient methods for non-negative matrix factorization. in neural computation. In *Neural Computation, pages 2756-2279*, 2007.