International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 04 Issue: 01 | Jan -2017 www.irjet.net p-ISSN: 2395-0072

Cross Domain Recommender System using Machine Learning and Transferable Knowledge

Pooja Rawade, Shrushti Dhawale, Sudarshan Jagadale, Kunal Patil, Prof.Shweta Koparde

[1]Pooja Rawade, [2]Shrushti Dhawale, [3]Sudarshan Jagadale, [4]Kunal Patil

Student, Dept. of Computer Engineering, Pimpri Chinchwad College Of Engineering, Maharashtra, India

[5]Prof. Shweta Koparde

Professor, Dept. of Computer Engineering, Pimpri Chinchwad College Of Engineering, Maharashtra, India

---

Abstract - *The useful knowledge from an auxiliary domain can be transferred through the social domain to a target domain. However, sometimes we may suffer from item cold start problem in the target domain. To alleviate this issue we apply cross domain algorithm along with page ranking algorithm. The cross domain algorithm is divided into two stages; In the first stage we apply the TrAdaBoost algorithm to select some items which are being recommended to users in the target domain. Whereas, in the second stage we adopt nonparametric pairwise clustering algorithm to make a decision whether to recommend an item to user or not. The algorithm finds the recommended or not recommended customer groups for one item through the two stages and then with the help of page ranking algorithm we provide relevant and unsearched data to the users.*

Key Words: **Collaborative Filtering, Cross Domain, NonParametric Pairwise Clustering Algorithm, TrAdaBoost Algorithm.**

1.INTRODUCTION

Recommender systems have become an important research area since the appearance of the first papers on collaborative filtering in the mid-1990's.[4][5][6] There has been much work done both in the industry and academia on developing new approaches to the recommender systems over the last decade. The interest in this area still remains high because it constitutes a problem-rich research area and because of the abundance of practical applications that help users to deal with information overload and provide personalized recommendations, content and services to them.[3]

A social networking service is a platform on which users can create and adopt different types of items such as messages, data or images. This huge volume of items generates a problem of information overload.[2] Thus with the development of information technology, we have already entered the era of big data. However, to discover the efficient data in various domains is critical.[7] Most recommender systems [8] encounter cold start problem. Cold start problem not only refers to a new user without any experience but also a new entity with few ratings and entirely a whole system. Cold start problem is challenging because no prior knowledge can be used in recommendation. Collaborative filtering is used to alleviate data cold start problem.[1] Another simple way is to transfer one domains information to the target domain i.e. cross domain recommendation. The data distribution (users, items & their features) in each domain is quite different and the new items in the target domain often suffer from cold start problem.

1.1 Cross Domain :

A cross-domain is a means of information assurance that provides the ability to manually or automatically access or transfer information between two or more differing security domains. They are integrated systems of hardware and software that enable transfer of information among incompatible security domains or levels of classification.

The goal of cross domain system is to allow an isolated critical network to exchange information with others, without introducing the security threat that normally comes from network connectivity. It has three primary elements :

1. Data confidentiality

2. Data integrity

3. Data availability

1.2 Recommender System :

Recommender systems or recommendation systems are a subclass of information filtering system that seek to predict the "rating" or "preference" that a user would give to an item. Recommender systems typically produce a list of recommendations in one of two ways – through collaborative and content-based filtering or the personality-based approach. Collaborative filtering approaches building a model from a user's past behavior (items previously purchased or selected and/or numerical ratings given to those items) as well as similar decisions made by other users. This model is then used to predict items (or ratings for items) that the user may have an interest in. Content-based filtering approaches utilize a series of discrete characteristics of an item in order to recommend additional items with similar properties. The personality-based approach derives product and service preferences from a user's personality.

2. Related Work

Collaborative Filtering :

Collaborative filtering is a method of making automatic predictions (filtering) about the interests of a user by collecting preferences or taste information from many users (collaborating). Many commercial websites use recommender systems to help customers locate products and content. Modern recommenders are based on collaborative filtering: they use patterns learned from users' behavior to make recommendations, usually in the form of related-items lists.

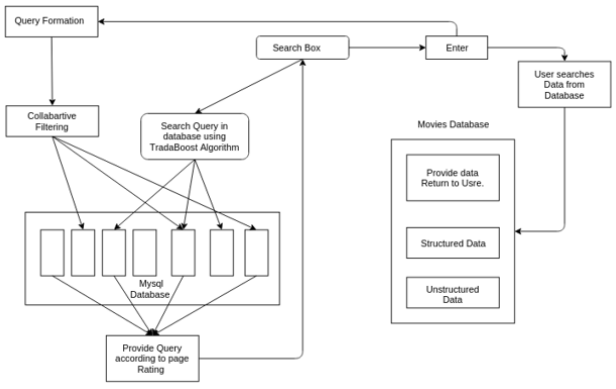
For example, a collaborative filtering recommendation system for television tastes could make predictions about which television show a user should like given a partial list of that user's tastes (likes or dislikes). Note that these predictions are specific to the user, but use information gleaned from many users. This differs from the simpler approach of giving an average (non-specific) score for each item of interest, for example based on its number of votes.

Cold Start Problem :

Cold start is a potential problem in computer-based information systems which involve a degree of automated data modeling. Specifically, it concerns the issue that the system cannot draw any inferences for users or items about which it has not yet gathered sufficient information.

Proposed System

In this section, we will describe our Proposed system architecture and our algorithm in two stages: at the first stage, TrAdaBoost algorithm is used to filter out the invalid data and output some random items which are worthy of being recommended to the users, then at the second stage nonparametric pairwise clustering algorithm is applied to cluster the users into two groups based on an recommended item got from the first stage. Further we apply Page Ranking Algorithm in which the PageRank of a page is defined recursively and depends on the number and PageRank metric of all pages that link to it ("incoming links") and we obtain the unsearched and relevant data content.



TrAdaBoost Algorithm :

It can be used in conjunction with many other types of learning algorithms to improve their performance. The output of the other learning algorithms ('weak learners') is combined into a weighted sum that represents the final output of the boosted classifier. AdaBoost is adaptive in the sense that subsequent weak learners are tweaked in favor of those instances misclassified by previous classifiers.

AdaBoost is sensitive to noisy data and outliers. In some problems it can be less susceptible to the overfitting problem than other learning algorithms. The individual learners can be weak, but as long as the performance of each one is slightly better than random guessing, the final model can be proven to converge to a strong learner.

Algorithm1 : TrAdaBoost

Input : The two labeled data sets D_s and D_{t_train} the unlabeled data set D_{t_test}.

Output : Set the maximum number of iterations M; and set the intial weight vector $W^1 = (W_1^1,\ldots,W^1_{n+m})$.

Loop :

For i=1 , …. , N

1. Set
$$p^t = w^t / (\sum_{i=1}^{n+m} W_i^t)$$
2. Call Learner, providing in the training set X with the distribution p^t over X and the unlabeled data set D_{t_test}. Then get back a hypothesis h_t : X -> Y (or [0,1] by confidence);
3. Calculate the error of h_t on D_{t_train} :
$$err_i = \sum_{i-n+1}^{n+m} \frac{W_i^t |h_t(x) - c(x_t)|}{\sum_{i-n+1}^{n+m} W_i^t}$$
4. Set
$$\beta_i = err_i/(1 = err_i);$$
$$\beta - 1/(1 + \sqrt{2Ln\, n/N};$$
And error required to be less than ½.
5. Update the new weight vector.

$$W_i^{t+1} = \begin{cases} W_i^t \beta_t^{|h_t(x_i)-c(x_i)|} \\ W_i^t \beta_t^{-|h_t(x_i)-c(x_i)|} \end{cases}$$

Output : The hypothesis,

$$H(x) = \begin{cases} 1 & \prod_{t=[\frac{N}{2}]}^{N} \beta_t^{-k_t(x)} \geq \prod_{t=[\frac{n}{2}]}^{N} \beta_t^{-1/2} \\ 0 & \end{cases}$$

Page Ranking Algorithm :

PageRank is a link analysis algorithm and it assigns a numerical weighting to each element of a hyperlinked set of documents, such as the World Wide Web, with the purpose of "measuring" its relative importance within the set. The algorithm may be applied to any collection of entities with reciprocal quotations and references. The numerical weight that it assigns to any given element *E* is referred to as the *PageRank of E* and denoted by PR(E).

A PageRank results from a mathematical algorithm based on the webgraph, created by all World Wide Web pages as nodes and hyperlinks as edges, taking into consideration authority hubs such as cnn.com or usa.gov. The rank value indicates an importance of a particular page. A hyperlink to a page counts as a vote of support. The PageRank of a page is defined recursively and depends on the number and PageRank metric of all pages that link to it ("incoming links"). A page that is linked to by many pages with high PageRank receives a high rank itself.

Nonparametric Pairwise Clustering :

The nonparametric pairwise clustering helps us to cluster the data without giving any specific parameters. Clustering considers the similarity terms between the data points for the purpose of clustering. In two nonparametric classifiers, i.e. the nearest neighbor classifier (NN) and the plug-in classifier (or the kernel density classifier). The generalization error bounds for both unsupervised classifiers are expressed as sum of pairwise terms between the data points.

Algorithm2 : Nonparametric Pairwise Clustering

Input: The output of the first phase { I_1……I_p} and data set D.

Initialize: Set the maximum nearest points of Ij m, the size of users k(U_1…….U_k) and maximum number of iterations N

1) Find out the m nearest neighbors of Ij and define it Up = { I_j, I_1,…….I_m} and we assume U0 is a virtual user that most suitable to be recommended.

2) For each user in D, let $U_{0i} = U_0 \cap U_i$ (I = 1…k, k is the size of users randomly selected in D) be the vector of item insertion that U_0 to all other users.

3) Calculate the proximities (distances) of point U_{0i} to all the other points:

$d_i = (d_{i1}, d_{i2}, \ldots d_{ik})$.

4) Iteration

For t =1,…N

Normalization step. Normalize di by dividing each of its components by $|| d_i ||$.

The resulting normalized vector is thus

$$P_i = (p_{i1}, p_{i2}, \ldots\ldots p_{ik}) = \frac{d_{i1}}{||d_i||}, \ldots\ldots \frac{d_{ik}}{||d_i||}$$

Re-estimation step. Re-estimate the proximity between every pair of data points i and j, so that $d_{ij} = dist(p_i, p_j)$.

End

Output: The result of a matrix A which only contains 0 & 1.

3. CONCLUSIONS

This paper addresses about Collborative Filtering for making predictions. We also reviewed various limitations of the current recommendation methods and discussed possible extensions that can provide better recommendation capabilities. We reconsidered the Cold start problem in a target domain by transfering knowledge from other auxiliary domains.. Therefore, the proposed system reduces cold start problem by using cross domain recommendation algorithm that is mainly divided into two parts TraAdaboost algorithm and Nonparametric Pairwise Clustering. The system filter's the search results and provides users with relevant search queries.

REFERENCES

[1] Panpan Liu,Jingjing Cao, Xiaolei Liang, Wenfeng Li "A Two Stage Cross Domain Recommendation For Cold Start Problem in Cyber-physical System", 2015.

[2] Meng Jiang, Peng Cui, Xumin Chen, Fei Wang, Wenwu Zhu, "Social Recommendation with Cross-Domain Transferable Knowledge", 2015.

[3] Gediminas Adomavicius, Member, IEEE, and Alexander Tuzhilin, Member, IEEE., "Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions", 2005.

[4] W. Hill, L. Stead, M. Rosenstein, and G. Furnas, "Recommending and Evaluating Choices in a Virtual Community of Use," Proc. Conf. Human Factors in Computing Systems, 1995.

[5] P. Resnick, N. Iakovou, M. Sushak, P. Bergstrom, and J. Riedl, "GroupLens: An Open Architecture for Collaborative Filtering of Netnews," Proc. 1994 Computer Supported Cooperative Work Conf., 1994.

[6] U. Shardanand and P. Maes, "Social Information Filtering:Algorithms for Automating 'Word of Mouth'," Proc. Conf. Human Factors in Computing Systems, 1995.

[7] Rajkumar R R, Lee I, Sha L, "Cyber-physical Systems: The next computing revolution", Proceedings of 47th Design Automation Conference. ACM, 2010.

[8] Zhen H, Hu F, "Cyber-physical System For Smart Grid Applications", Cyber-physical Systems: Integrated Computing and Engineering Design, 2013.

[9] Sanjay* and Dharmender Kumar, "Review Paper on Page Ranking Algorithms", International Journal of Advance research in Computer Engineering and Technology(IJARCET), 2015.

[10] Hema Dubey, Prof. B N Roy, "An Improved PageRank Algorithm based on Optimized Normalization Technique", International Journal of Computer Science and Information Technology, 2011.