ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 5, May 2018

Similarity Based and Transfer Learning Based Data Fusion Using Classifier

Yogesh Shinde[1], Nikhil Ghule[1], Rubina Shaikh[1], Prof. Ashish Ramdasi[2]

B. E Students, Department of Computer Engineering, Sinhgad Academy of Engineering, Pune, India[1]

Professor, Department of Computer Engineering, Sinhgad Academy of Engineering, Pune, India[2]

ABSTRACT- We blend the data fusion strategies, gathering them into three classes: the stage-based, the feature level-based, and the semantic meaning-based data fusion methods. The last class of information combination techniques is additionally isolated into four types: multi-view learning-based, similarity-based, probabilistic dependency-based, and transfer learning-based methods. These procedures focus on learning combination as opposed to schema mapping and data consolidate, fundamentally recognizing cross-domain data fusion and traditional data fusion considered in the database gathering. We don't just irregular abnormal state standards of every class of procedures, yet in addition give examples in which these methods are utilized to deal with veritable huge information issues. Likewise, this paper positions existing works in a system, researching the relationship and refinement between various data fusion strategies. This paper will empower an extensive variety of networks answer an answer for data fusion issue in enormous information ventures.

KEYWORDS: cross-domain data mining, matrix factorization, probabilistic graphical models, data fusion, multi-modality data representation, deep neural networks, multi-view learning, Big Data, transfer learning, urban computing.

I. INTRODUCTION

Over couple of years back information combination/data combination was viewed as sensor information combination technique, where diverse sensor would deliver information, data recovered from those sensors would be utilized to detect the other condition circumstance, so to take choice in like manner [5]. Combination comprises in contacting or consolidating data those branches from a several sources and exploiting, blended data in different errands, for example, noting questions, making decisions, numerical estimations for encourage prescient examination, for picking up bits of knowledge of the informationdesign and utilized for recovering valuable deductions. Data combination is process managing affiliation, relationship, and mix of information and data from different sources to achieve refined assessments of parameters, characteristics, events and practices for observed questions in a watched field of view[3]. It is sometime implemented forautomated decision support system. Coordinated data systems furnish clients with a bound together perspective of different heterogeneous sources. Questioning the basic data sources, combine the outcomes, and showing them to client is performed by the integration system [2]. With more datasources effortlessly accessible by means of shoddy system association, either finished the web or in organization intranet, the coveted to get to all these sources through a reliable interface has been the main thrust behind much research in the field of data combination. Amid the most recent three decades various systems that attempt to achieve this objective have been produced, with changing degrees of progress [4].

II. MOTIVATION

Data Mining generally works on the single domain. Recently embrace data is a new technology to works on large amount of information [5]. There are different methodologies which are stage based, feature level based and Semantic meaning-based data fusion method. In semantic meaning based there are four type they are multi view learning based, similarity based probabilistic dependency based and transfer learning-based method [6].

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 5, May 2018

III. OBJECTIVES

This paper consolidates existing data fusion strategies, describing them into three noteworthy categories and giving certifiable examples in each sub-class of methods. This paper researches the relationship and contrasts between various methods, helping individuals find appropriate data fusion techniques to take care of embrace data set issue. Some open multi-methodology datasets have been shared to encourage additionally research into data fusion issues with classifications as K-NN also, Clustering Classification are used.

IV. PROBLEM DEFINITIONS

Cross-domain recommendation gives best result and more varied recommendations leading to the satisfaction of user and solving cold start and sparsely problems but it is less accurate than single domain recommendation system. Also, another issue is the lack of contextual features in cross domain recommender system, which when used can be very efficient in recommendation. Thus, this paper talks about different methodologies such as the recommendation methods are POI and Collaborative filtering.

V. PROPOSED SYSTEMARCHITECTURE

This Document summarizes the methods that can fuse multiple datasets by 2 categories. The first type of data fusion methods uses similarity data sets. Second type uses transfer of data set from source to domain.
These categories are sub category of semantic meanings, which can be further arranged as:

1. Similarity based methods
2. Transfer leaning based Methods

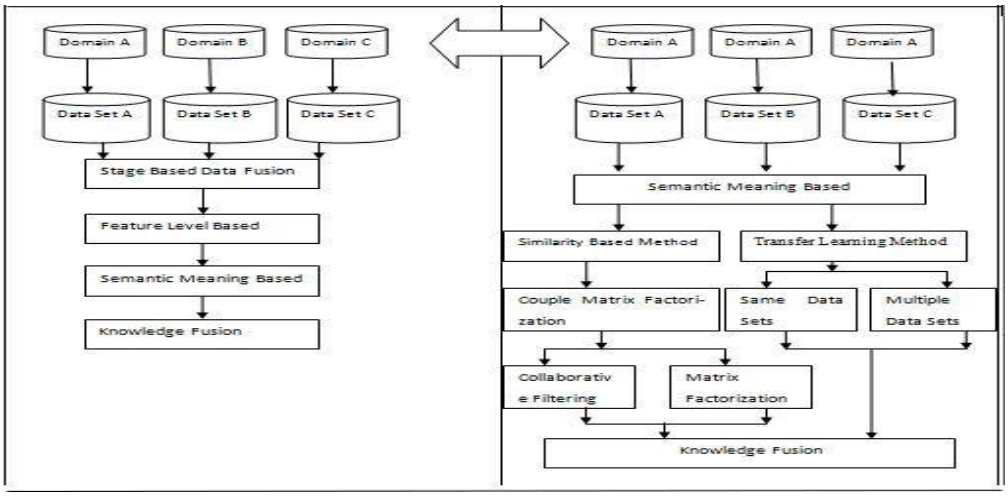


Fig 2: Proposed System Architecture

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 5, May 2018

Similarity based method

Similarity lies between various articles. In the event that we have two articles (X, Y) are comparative in a few aspects, the data of Xcan be utilized by Y when Y is absence of information and the other way around. Whenever X and Y have different datasets individually, we can take in numerous similarity between those articles, every one of which is calculated in light of a couple ofrelating datasets. These similitudes can commonly fortify among each other, merging the relationship between two protests aggregately. The last improves every individual closeness thusly. For instance, the closeness gained from a thick dataset can fortify those got from other inadequate datasets, accordingly helping filling the missing estimations of the last mentioned [12].

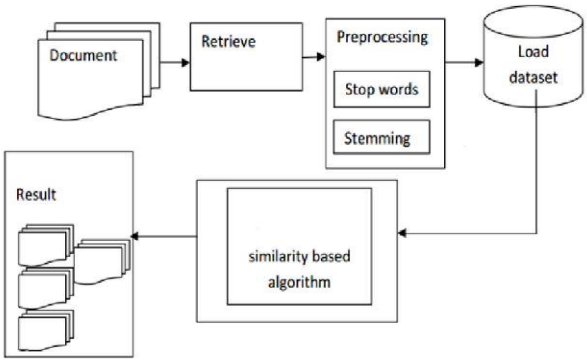


Fig 3: Similarity Architecture

From another point of view, we can state we will probably precisely evaluate the closeness between two questions by consolidating different datasets of them. Therefore, unique datasets can be mixed together in view of similitudes. Coupled matrix factorization and complex arrangement are two sorts of delegate techniques in this classification.

Coupled Matrix Factorization

We have to comprehend two concepts. One is Collaborating Filtering (CF); the other is matrix factorization. The last it can be a productive way to deal with the execution of CF models.

Collaborative Filtering

Collaborative Filtering is an outstanding model generally utilized as a part of recommender systems utilizing numerous angles like POI (point of interest). The general thought behind Collaborative Filtering is that comparative clients make evaluations in a same way for comparable items or say things [11]. along these, if likeness is resolved amongst clients and things or say items, a potential expectation can be made with regards to the rating of a client concerning future things or items. Clients and things are for the most part sorted out by a matrix, where a new entry means a client's evaluating or rating on a thing. The rating can be unequivocal rankings or implicit indications, for example, the quantity of visits to a place or the circumstances that a client has perused a thing. Once defining a matrix, the separation between two rows in the matrix means the similitude between two clients, while the separation between two columns remains for the comparability between two things. Memory-based CF is the most broadly utilized calculations, which figure the estimation of the obscure rating for a client and a thing as a total of the appraisals of some other (more often than not, the N most comparative) clients for a similar thing. There are two classes of memory-based CF models: user/client based and item/thing-based techniques:

$$rpi = \overline{Rp} + d\sum sim(up,uq)\ uq\in U' \times (rqi - \overline{Rq}); \quad (1)$$

$$d = 1\ |U'|\ \sum sim(up,uq)\ uq\in U'\ ; \quad (2)$$

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

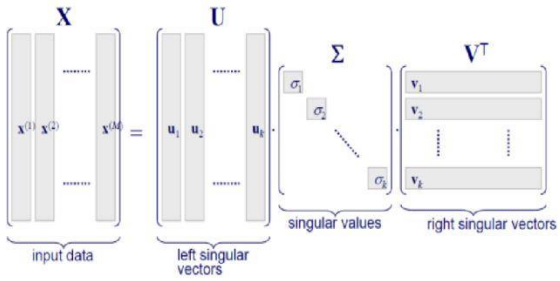
Vol. 6, Issue 5, May 2018

$$Rp^{-} = 1\ |S(Rp)| \sum rpi\ i \in S(Rp)\ ,\ ; \qquad (3)$$

Where (p,) denotes the similarity between user and ; and mean the average rating of and respectively, denoting their rating scale; ( ) representsthe collection of items rated by ; ′ is the collection of users who are the most similar to . − is to avoid rating biases of different users. When the number of users becomes big, computing the similarity between each pair of users is impractical for a real system. Given that the number of items could be smaller than that of users, item based CF, e.g. the Slop One algorithm [42], was proposed to address this issue. When the number of users and number of items are both huge, matrix factorization-based method is employed to implement a CF model.

Matrix factorization

Matrix factorization decomposes a (sparse) matrix into the production of two (low-rank) matrices, which denote the latent variables of users and items respectively. The production of the two matrices can approximate matrix , therefore helping fill the missing values in . There are two widely used matrix factorization methods: Singular Value Decomposition (SVD) and non-negative matrix factorization (NMF)



Transfer learning

A principle supposition in numerous machine learning and data mining calculations is that the preparation and future data must be in a similar element space and have a similardissemination. In any case, in some certifiable applications, this presumption may not hold. For instance, we now and again have a characterization errand in one area of intrigue, yet we just have adequate preparing information in another domain of interest, where the last information might be in an alternate feature space or take after an alternate data conveyance. Different from semi supervised learning, which expect that the dispersions of the labelled and unlabeled data are the same, transfer learning, interestingly, permits the domains, tasks, and circulations utilized as a part of preparing and testing to appear as something else [12]. In reality, we watch numerous cases of transfer learning. For example, figuring out how to perceive tables may help perceive chairs as shown in *Fig 4(a)*.Getting the hang of riding a bicycle mayhelp riding a Moto-cycle as shown in*Fig 4 (b)*.. Such illustrations are likewise broadly seen in the computerized world. For example, by breaking down a client's transaction records in Amazon, we can analyse their interests, which might be moved into another utilization of travel proposal. The information gained from one city's traffic information might be exchanged to another city.

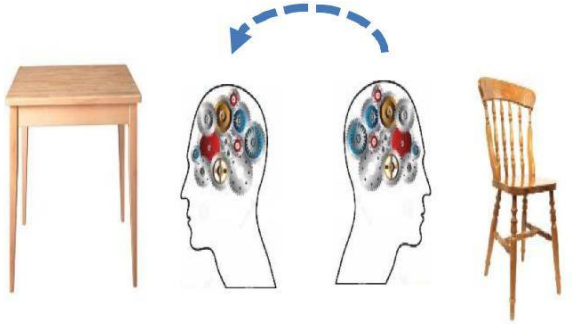


Fig 4 a. Example of transferring knowledge

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 5, May 2018

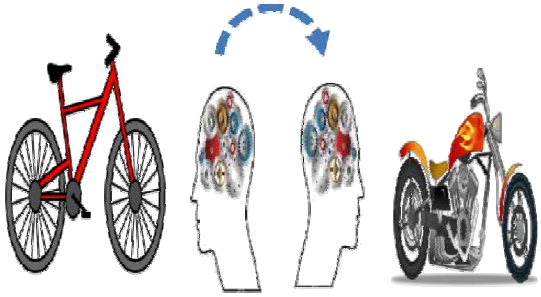


Fig 4 b. Example of transferring knowledge

In the big data period, numerous machine learning tasks need to bridle an assorted variety of information in an area in order to accomplish a superior execution. This calls for new strategies that can exchange the learning of various datasets from a source to a target area. For instance, a major city like Beijing may have adequate datasets, (for example, traffic, meteorological, and human mobility and so forth.) to induce its fine-grained air quality. Be that as it may, while applying the model to another city, we might not have some sort of datasets (e.g. traffic) at all or insufficient perceptions in some datasets (e.g. human mobility). Would we be able to exchange the information gained from numerous datasets of Beijing to another city?

VI. DISCUSSION

It is hard to judge which data fusion method is the best, as different methods behavior differently in different applications. Table 3 presents a comparison among these data fusion methods (list in the first column), where the second column (Meta) indicates if a method can incorporate other approaches as a Meta method. For instance, semantic meaning-based data fusion methods can be employed in a stage-based fusion method. To select a proper data fusion method, we need to consider the following factors:

METHODS	META	POSITION LABELS	GOALS
Stage-based	Y	NA	NA
Direct	N	Flexible	F,P,C,O
DNN	N	Flexible	F,P,A,O
Multiview	Y	Fixed	F,P,O
Probability	N	Fixed	F,P,C,O, A
Similarity	N	Flexible	F,A,0
Transfer	Y	Fixed	F,P,A

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 5, May 2018

When studying a type of objects, e.g. geographical regions, we need to consider whether there are some object instances that can constantly generate labeled data (titled "Fixed" or "Flexible" in the 3rd column "Position") [12]. **For example**, we can have fixed monitoring stations constantly generating air quality data in some regions. The goal, learning approach, and requirement of a machine learning and data mining task. First, goals of fusing multiple datasets include **Filling Missing Values** of a sparse dataset, **Predict Future**, **Causality Inference**, **Object Profiling**, and **Anomaly Detection** etc. [12].

VII. CONCLUSION

The techniques that can discover knowledge from various divergent datasets with fundamental associations. This paper compresses existing data fusion techniques, masterminding them into three noteworthy classes and giving genuine examples in eachsub-characterization of systems. This paper investigates the relationship and contrasts between various strategies, helping people find appropriate information combination techniques to mind enormous information issues. Some open multi-methodology datasets have been shared to support additionally explores into data fusion issues.

REFERENCES

[1] "Location based and Preference Aware Recommendation Using Sparse Geo-Social Networking Data", J.Bao, Y.Zheng and M. F.Mookbel.
[2] "A model of inductive bias learning," J. Baxter.
[3] "Representation learning: A re-view and new perspectives, Y.Bengio, A.Courville, and P.Vincent.
[4] "Christopher M. Bishop, "Chapter 8. Graphical Models," A. Blum and T. Mitchell.
[5] "Co-em support vector learning," U. Brefeld and T.Scheffer,
[6] "Predictive subspace learning for multi-view data: A large margin approach," N. Chen, J. Zhu, and E.P. Xing.
[7] "Context-Dependent Pre-Trained Deep Neural Networks for Large Vocabulary Speech Recognition, E. Dahl, D. Yu, L. Deng, and A. Acero.
[8] "Using collaborative filtering to weave an information tapestry D. Goldberg, N.David, M. O. Brain,T. Douglas.
[9] "Multiple kernel learning algorithms" M. Gonen and E. Alpaydn.
[10] "Investigating causal relations by econometric models and cross-spectral methods" C. W. Granger
[11] "Using collaborative filtering to weave an information tapestry," D. Goldberg, N. David, M. O.Brain, T. Douglas.
[12] "Methodologies for Cross-Domain Data Fusion: An Overview", Yu Zheng
[13] "Similarity based and Transfer learning based data fusion using classifier," Yogesh Shinde, Nikhil Ghule, Rubina Shaikh