International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 04 Issue: 12 | Dec-2017 www.irjet.net p-ISSN: 2395-0072

Cross Domain Data Fusion

Amit Wavhal[1] , Prof. Suhasini Itkar[2]

[1,2] *Department of Computer Engineering P.E.S Modern College of Engineering*

Abstract—In recent years we have seen explosion of data on the World Wide Web front. Most of the information is in the unstructured format and very few information available is in relational form or in some kind of structured form .Data mining has come a long way in extraction of information using such big information of data. This huge data also called Big Data in todays scenario. But need of the hour is not only extract information from one viewpoint or single dimension. Seeing information from one perspective does not always give the right picture, for any organization using BI or business intelligent system decisions are taken using this single dimension of data. But as world is changing fast and to stay relevant in the market or remain market leader one need to see the unseen dimension of data, to answer to this question Data Fusion Technique comes into picture. Then analysis is done for each step from load balancing, accuracy and complexity aspects. Depending on the performance of the algorithms the best solution is considered.

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

I. INTRODUCTION

Over couple of years back data fusion information fusion was seen as sensor data fusion technique, where various sensor would produce data, information retrieved from those sensor would be used to sense the external environment situation, so to take decision accordingly. Fusion consists in touching or merging information that branches from several sources and exploiting, merged information in various task such as answering questions, making decisions, numerical estimations for further predictive analysis, for gaining insights of the data pattern and used for retrieving useful inferences.

Information fusion is process dealing with association, correlation, and combination of data and information from multiple sources to achieve refined estimates of parameters, characteristics, events, and behaviors for observed objects in an observed field of view. It is sometimes implemented for automated decisions support system.

Integrated information systems provide users with a unified view of multiple heterogeneous data source. Querying the underlying data sources, combing the results, and presenting them to user is performed by the integration system.

With more and more information sources easily available via cheap network connection, either over the Internet or in company intranet, the desired to access all these source through a consistent interface has been the driving force behind much research in the field of information integration. During the last three decades many systems that try to accomplish this goal have been developed, with varying degrees of success.

II. REVIEW OF LITERATURE

Data fusion is process consists of mapping source data into target representation, identifying multiple representation of the same real-word object, and finally combining these representation called data fusion, while fusing data, we have to take special care in handling data conflicts, this paper focus on the definition and implementation as in [1], as well as high level understanding, some of the technique which can applied to data fusion in [2].

Different domain data-set is identified first, by applying rules for selection of data-sets to do data fusion[1]. The second step is to identifying the fields that needs to be used for data fusion process. Once data-sets with right number of fields are identified, data mining process is applied individual data-set. Once knowledge is extracted from individual data-set, inference data points are populated which can be used for data fusion, once inferences or knowledge is extracted, then two different domain data-sets are used for fusion. The most important part in data fusion is to discover the association between the different domain data sets. Once this fields which are selected for fusion will discover the association between this different domain fields and knowledge or pattern would be discovered for showing our conclusion.

Integrated information systems must usually deal with diversified representation of data (schemata).In order to present to the user query results in a single unified schema, the schematic Heterogeneity's must be bridged. Data from the data sources must be converted to Conform to the global schema of the information system. Two methods are common to bridge heterogeneity and thus specify data transformation:

schema integration and schema mapping. The former way to deal is driven by the desire to integrate a known set of data sources. Schema integration regards the individual schemata and tries to generate a new schema that is complete and correct with respect to the source schemata, that is minimal, and that is understandable[3]

Data Transformation - The objective of both approaches, schema integration and schema mapping, is the same: convert data of the sources so that it conforms to a common global schema. Given a schema mapping, either to an integrated or to a new schema, finding such a complete and correct transfor-mation is a considerable problem [3]. The data transformation itself, once found, can be performed offline, for instance, as an ETL process for data warehouses; or online, for instance, in virtually integrated federated databases. After this step in the data integration process all objects of a certain type are represented

Duplicate Detection - The next step of the data integration process is that of duplicate detection (also known as record[3] linkage, object identification, reference reconciliation, and many others). The objective of this step is to identify multiple representations of the same real-world object: the basic input to data fusion.

The result of the duplicate detection step is the assignment of an object-ID to each representation. Two representations with the same object-ID indicate duplicates. Note that more than two representations can share the same object-ID, thus forming duplicate clusters. It is the objective of data fusion to fuse these multiple representations into a single one

III. SYSTEM ARCHITECTURE / SYSTEM OVERVIEW

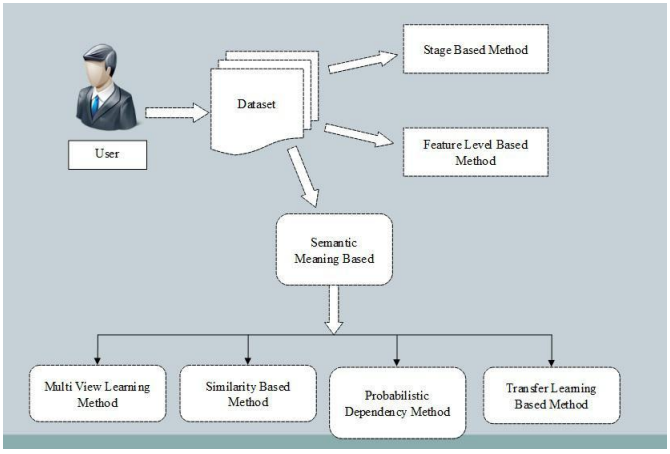

Fig. 1. Architecture Diagram of Cross Domain Data Fusion

Processing Steps The following scheme consists of three basic steps:

1) Cleaning Integration multiple Domain Data
   i. Selection Transformation
   ii. Data Mining (Classification /Clustering)
2) Stage Based Data Fusion Method
3) Feature Level Based Data Fusion Method
4) Probabilistic Dependency and Transfer Learning Based methods
5) Multi-View Learning and Similarity Based Method

1. Stage Based Data Fusion Method- Point of interest is the important thing in this method. The data-set is considered as the nodes and the edges. For friend recommendation system the people and their data are stored in the database. Based on the location and based on the visited places the recommendation have to be done.If the similarity is high means the persons are recommended to the target user. This method is called the stage based data fusion method[14].

2. Feature Level Based Data Fusion Method- The data-set is given as a input to the feature level based data fusion method. From this data-set the features are extracted from the data-set. These features are important terms in the data-set. These features are used to represent the data-set in a fused method. A unified feature representation from disparate data-sets based on DNN. The feature vectors are useful for classification and clustering.

3. Probabilistic Dependency and Transfer Learning Based methods The latent variables are find out by this method. The frequency of the words are the base for this method. Based on the frequent patterns the topics are find out for the given input data. The frequency of the pattern denotes the number of occurrence of a word in a document. The transfer learning is based on the prediction of the terms. Based on the history of the user the future terms for the user is predicted.

4. Multi-View Learning and Similarity Based Method In multi view learning method the KNN classification is used. In KNN classification the distance is the significant term. Based on the distance the unlabeled data are classified. In similarity based method the similarity is very important. For find the similarity the collaborative filtering is used. The similar data are find b out by this collaborative filtering method.

TABLE I

COMPARISON OF DIFFERENT DATA FUSION METHODS

Technique	Vol	Positions	Goals	Training(type)
Stage-Based	NA	NA	NA	NA
Feature-Level	L	FLex	F,P,A,O	U/S
Semantic(probability)	S	Fix	F,P,C,O,A	S/U
Semantic(Transfer-learning)	S	Fix	F,P,A	S/U

IV. SYSTEM ANALYSIS

A. DataSet

The dataset used is created for experimental purpose for application of various techniques of data fusion. The Fields and data is actually mock data which is used by various methods which actually implement this fusion techniques

B. Hardware

1) Memory: 2GB
2) Processor: Intel (R) Pentium (R) CPU B950 @2.10 GHz
3) Hard disk: 64 GB

C. Software

1) Java : Java 1.7.0 Above version
2) Database:
3) Operating System: Windows 7 and above.
4) IDE: Net Beans 7.2.1 and Above
5) vi editor

D. Comparative Parameter

1) Vol = Volume of data can be handled by fusion method
2) Positions = Handling Fixed and Flexible Data object capacity
3) Goals = Goals it can achieve such as
   a) F = Filling Missing Values
   b) P = Predict Future
   c) O = Object Profiling
   d) A = Anomaly Detection
4) Training Type = As supervised or unsupervised

V. CONCLUSION

In the given paper we implemented the solution for Data fusion technique from multiple disparate data-set.

ACKNOWLEDGMENT

Every orientation work has an imprint of many people and it becomes the duty of author to express the deep gratitude for the same. I feel immense pleasure to express deep sense of gratitude and indebtedness to my guide Prof. Suhasni Itkar, for constant encouragement and noble guidance. I also express my sincere thanks to the Computer Department as well as Library of my college. Last but not the least; I am thankful to my friends and my parents whose best wishes are always with me.

REFERENCES

[1] https://www.microsoft.com/en-us/research/publication/methodologies-for-cross-domain-data-fusion-anoverview.

[2] Yu Zheng, Senior Member, Methodologies for Cross-Domain Data Fusion: An Overview, IEEE TRANSACTIONS ON BIG DATA, TBD-2015-05-0037.

[3] Bleiholder, J. and Naumann, Data fusion, ACM Comput. Surv, 41, 1, Article 1 (December 2008),41 pages DOI = 10.1145/1456650.1456651.

[4] Y.Bengio, A.Courville, and P.Vincent, Reprsentation Learning: Review and new prespective, IEEE Transactions on pattern Analysis and Machine Intelligence ,2013:1798-1828.

[5] J.Shang, Y.Zheng, W.Tong, E.Chang and Y.YU, Inferring Gas consumption and Pollution Emission of Vehicles through City, Proc.ACM SIGKDD Conf. Knowledge Discovery and Data Mining pp. 1027-1036 2014.

[6] K.Nigam and R.Ghani, Analyzing the effectiveness and applicability of co-training, Proc. of the ninth international conference on information and knowledge management,PP.86-93,2000.

[7] Y.Zheng,Y.Liu,j.Yuan and X.xie, Urban Computing with Taxicabs, Proc. ACM Conf. Ubiquitous computing (Ubi Comp'11) ,pp.89-98,2011.

[8] N. Sirvastava ,R. Salakhutdinov, Multi Modal Learning with Deep Boltza - mann Machines, Proc. Neural Information Processing Systems, 2012.

[9] Y.LeCun and M. Ranzato,Deep Learning Tutorial, In Tutorials in international Conference on Machine Learning, 2013.

[10] V.W. Zheng, Y. Zheng, X.Xie and Q.Yang, Towards Mobile Intelli- gence: Learning from GPS History Data for Collaborative Recommendation, Artificial Intelligence Journal,pp.17-37,2012.

[11] Y.Sun,Y.Yu and J.Han, Ranging-based clustering of heterogeneous information network with star network schema, Proc the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining ,pp.797-806,2009.

[12] F.Zhang,N .J.Yuan, D. Wilkie, Y.Zheng ,X.Xie, Sensing the Pulse of Urban Refueling Behavior. Perspective from Taxi Mobility, ACM Tans. Intelligent System and Technology, Submitted for publication.

[13] V.W.Zheng ,B.Cao, Y.Zheng, X.Xie, Q.Yang ,Collaborative Filtering Meets Mobile Recommendation: A User Centered Approach, Pro. AAAI Conf. Artificial Intelligent (AAAI'10),pp.236-241,2010.

[14] N.J.Yuan and X.Xie, Segementation of Urban Area Uisng Road Net-works, Technical Report MSR-TR-2012-65,2012.