A SURVEY: MACHINE LEARNING USING HETEROGENEOUS INFORMATION GRAPH

Shruthi C J[1], Mouneshachari S[2]

[1]M.tech Student, [2]Associate Professor, Department of CSE, Gsssietw (VTU), Mysore, (India)

ABSTRACT

Most of the information available in the real world are of different type or heterogeneous, when we connect such a kind of information it will form the information network. The extraction of useful knowledge from this information network becomes ubiquitous so, this can be effectively handled by using clustering and ranking approach. When we use clustering and ranking methods for extracting useful knowledge with semantic structure we can get relative information and also it may lead to better understanding of hidden knowledge of the network, as well as particular role of every objects within the cluster. By using the ranking method we can clustered the dataset effectively based on the ranking value of the dataset so ranking methods serves as a good measure for clustering the heterogeneous information. The concept of heterogeneous information graphs has attracted in the field of social-networks, social media and machine learning systems. This paper review some of the methods which is used for the processing of heterogeneous information networks for different types of data sets and structure of heterogeneous information networks for extracting semantic information.

Keywords: Heterogeneous Graphs, Semantic Information, Information Extraction

I. INTRODUCTION

Data which is available in the real world consists of multiple type of objects or components, this objects are interconnected to other set of components and the connection of this types of components will form the heterogeneous networks. By using this heterogeneous information networks we can represent and extract hidden information from the source. The extraction of information from heterogeneous information network requires the grouping of relevant knowledge so, this can be achieved by using clustering and ranking function because these functions will give better and effective performance for information extraction.

Yizhou sun and Jiawei Han[1] have discussed about the mining of heterogeneous information networks and this heterogeneous information networks are formed by using the multiple types of components and which is interconnected so it will form the complex network, some times this type of networks also called as semi-structured networks because it will used for the representation of different types of objects with hidden or uncovered information. This kind of information can be effectively mined by clustering, ranking and classification methods along with this meta-path based similarity approach also used for calculating the similarity so the heterogeneous information can be mined effectively.

Yizhou sun , Jiawei Han and Peixiang Zhao[2] have reported different methods for ranking and clustering the heterogeneous information by using the novel clustering frame work that is RANKCLUS. This method rank the data based on the clustering so, it can improve the ranking quality and also it improves the clustering by conditional ranking and by using this technique we can get more accurate and meaningful results. In this paper, Section 2 presents Related work on this topic, Section 3 discussion regarding most frequently used measures or some important methods, Section 4 Conclude the paper.

II. RELATED WORK

Some of the works reported in the literature that focused on grouping the different types of information and measuring the importance of heterogeneous networks for information extraction process. However, some of methods available in the literature are reviewed in this Section.

Yizhou sun, Yintao Ya and Jiawei Han[3] have presented the clustering method for detecting the new clustering problem by using star network schema and which splits the original network into K layers so, this method differs from the current methods. In this approach NetClus a novel ranking method is used and which generates target objects by using the ranking-based probabilistic generate model and this target objects is mapped to the new low dimensional measure by calculating the posterior probability of the objects which is belongs to net-cluster.

Yizhou sun, Xifeng Yan and Jiawei Han[4] discussed new idea about mining the knowledge from the database and other interconnected data as a heterogeneous information. In this approach the data which is present with in the database that will be considered as heterogeneous information for this database Rank-based clustering and classification method are applied for extracting the information and also meta-path based methods are used for finding the similarity and relationship between the data sets, and this relationship strength can be measured through the selection of attributes and integrating user-guided clustering with meta-path selection process so, it will give better results for database knowledge mining.

Rumi Ghosh , Kristina Lerman[5] have reported work on information can be processed easily by using graphs or networks if it has same type of components but when it consists of different types of components it will become difficult so, this can be handled efficiently using mathematical framework that is specifically modularity-maximization method was developed for analyzing and processing the multiple types of entities and their links and it will be most advantageous because it has tunable parameters and the information is processed by using N-mode matrix data structure for representing different classes of entities and relations of information and Bonacich centrality is used for analyzing the structure of the networks. B- Centrality is used for ranking the nodes and based on the ranking values for communication between different communities are identified and the network is balanced.

Jinpeng Wang, Jianjiang Lu[6] have discussed different data source will form heterogeneous information networks and each different networks has different data models, schemas and languages and when we try to extract data or information from this network it need to integrates one networks with other networks for extracting relevant information so, this can be handled by using ontology- based approach. The ontology method is used as mediated schema for representing different data model or data source semantics and RDF graphs patterns are used for the modeling of different source schemas and this can be process efficiently using SPARQL queries.

Yizhou Sun and Jiawei Han[7] have done work on the large- scale heterogeneous information networks consists of multi- typed interconnected objects and it requires findings the similarity for searching the information in the database or in the web search engines based on the different paths between the entities because semantic meanings are behind the links of heterogeneous network entities. Each links between the entities consists of different semantic similarities or it consists of same semantic similarity so, it will be processed by using meta-path- based similarity methods for representing different object types by links. PathSim is a novel similarity measure is used for finding the peer objects in the networks by using queries and based on this query results we can compute the top- K similarity results of objects of same type in the heterogeneous information graphs.

Andre Freitas and Edward Curry[8] have discussed the querying of heterogeneous structured data become trend in many applications like distributed databases but it will become very difficult because of semantic gap present in the information expressed by different users so, this can be efficiently processed by using natural language interface and semantic index for querying information in the linked datasets and this can be achieved by distributional- compositional semantic approach. This approach automatically extract co- occurring words from the large text and it form τ-Space distributional structured vector model and it will compute the semantic matching and expressive natural language queries.

Ming Ji, Yizhou Sun[9] have discussed about heterogeneous information networks transductive classification problems and they have proposed new novel based graph classification i.e, GNetClass methods for better labeling the unknown data in heterogeneous information networks and using this method each link will be consider separately because the semantic meaning of the network is preserved in the links so, better information extraction can be done and also this method will give most efficient accurate classifications of data compared to other classification methods.

Ludwig M. Busse, Peter Orbanz[10], have reported work on clustering the ranked heterogeneous information and the mixture approach is used for the clustering of ranked data and this ranked data is compared based on the probabilistic model for different length of the ranked data so, it will give better analysis of heterogeneous information.

III. METHODS

This section review the most relevant methods for extracting the information from the heterogeneous information graphs.

Mining heterogeneous information from the networks we need to compute the values for links between the different entities for extracting relevant information from large set of data so, Ranking, Clustering and Classifications are the better methods for processing heterogeneous information.

3.1 Ranking- Based Clustering in Heterogeneous Information Networks

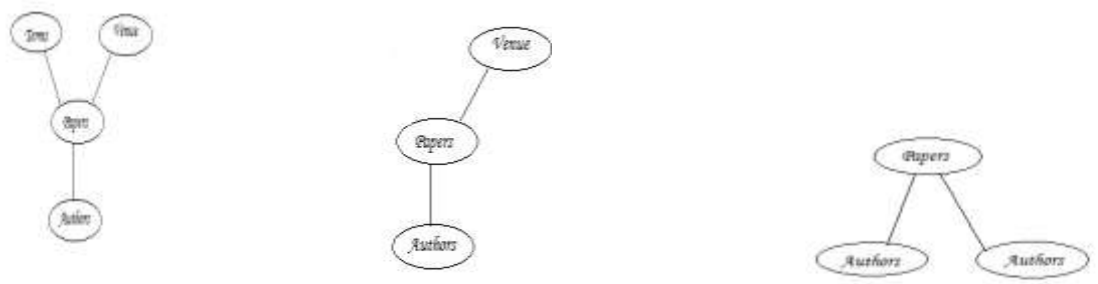
Large- set of information can be connected through the links for heterogeneous types of data and it requires the links between each type of components which is present in the networks and this can be efficiently analyzed by clustering and ranking methods. The ranking and clustering can mutually enhance each other because objects highly ranked in the same cluster. Clustering approach can be understood very efficiently by reading the top-ranked objects in that cluster. The most commonly used ranking- based clustering approach are RankClus and NetClus.

3.2 Ranking- Based Classification in Heterogeneous Information Networks

The knowledge which is present in the heterogeneous information networks can be classified efficiently because the nodes can be linked together are likely to be similar or different type of links have different level of strength so, this ranking values we can classify the heterogeneous information and most commonly used classification methods are GNetMine and RankClass.

3.3 Meta- Path Based Similarity Search and Mining

In heterogeneous information network same type of objects can connect in different links so, for finding similarity and interesting information in the heterogeneous information network can be done by meta- path-based methods and most commonly using method is PathSim for finding the peer objects in the networks and it compares the objects by random- walk.



Network Schema **Meta-Path: APV** **Meta-Path: APA**

Fig 1. Metapath Representation

Table 1. Instance Metapath

	Connection Type I	Connection Type II
Path instance	Jim-P1-Ann Mike-P2-Ann Mike-P3-Bob	Jim-P1-SIGMOD-P2-Ann Mike-P3-SIGMOD- P2-Ann Mike-P4-KDD-P5-Bob
Meta-path	A(uthor)-P(aper)-A	A-P-V(enue)-P-A

IV. CONCLUSION

The main objective of this paper is to highlights the basic methods for mining different types of components based information from the heterogeneous information graphs as well as to provide review report carried out in this area. According to this methods we can get better information from the network and also it will give the useful information about the strong methods used for the mining the data from the heterogeneous information graphs .

REFERENCE

[1]. Yizhou Sun, “Mining Heterogeneous Information Networks: A Structural Analysis Approach”, SIGKDD Explorations Volume 14, Issue 2.

[2]. Yizhou Sun†, Jiawei Han†, Peixiang Zhao, “RankClus: Integrating Clustering with Ranking for Heterogeneous Information Network Analysis”, EDBT 2009, March 24–26, 2009, Saint Petersburg.

[3]. Yizhou Sun Yintao Yu Jiawei Han, “Ranking-Based Clustering of Heterogeneous Information Networks with Star Network Schema”, KDD’09, June 28–July 1, 2009, Paris, France.

[4]. Yizhou Sun, Jiawei Han, Xifeng Yan, “Mining Knowledge from Interconnected Data: A Heterogeneous Information Network Analysis Approach”, Proceedings of the VLDB Endowment, Vol. 5, No. 12, August 27th - 31st 2012.

[5]. Rumi Ghosh, Kristina Lerman,” Structure of Heterogeneous Networks”, arXiv: 0906.2212v1 [cs.CY] 11 Jun 2009.

[6]. Jinpeng Wang, Jianjiang Lu, “Integrating Heterogeneous Data Source Using Ontology”, JOURNAL OF SOFTWARE, VOL. 4, NO. 8, OCTOBER 2009.

[7]. Yizhou Sun, Jiawei Han, “PathSim: Meta Path-Based Top-K Similarity Search in Heterogeneous Information Networks”, 37th International Conference on Very Large Data Bases, Seattle, Washington. Proceedings of the VLDB Endowment, Vol. 4, No. 11, August 29th - September 3rd 2011.

[8]. Andre Freitas, Edward Curry, “Natural Language Queries over Heterogeneous Linked Data Graphs: A Distributional-Compositional Semantics Approach”, IUI’14, February 24–27, 2014, Haifa, Israel.

[9]. Ming Ji, Yizhou Sun, “Graph-based Classification on Heterogeneous Information Networks”, TECHNICAL REPORT, APRIL 2010.

[10]. Ludwig M. Busse, Peter Orbanz, “Cluster Analysis of Heterogeneous Rank Data”, 24th International Conference on Machine Learning, Corvallis, OR, 2007.