Comparative Analysis of Similarity Measures in Heterogeneous Information Network

Vaishali Patil, Dr.Ramesh Vasappanavara, Tushar Ghorpade
Department Of Computer Engineering
Ramrao Adik Institute Of Technology
Nerul, Navi Mumbai, India
vaishalib2703@gmail.com, ramesh.vasappanavara@gmail.com, tushar.ghorpade@gmail.com

***Abstract*— Information network derived from various domains are studied recently. Searching for Similarity is a major task into such types Information Network. Lot of research on computing similar objects is done in Homogeneous Information Network. But real world scenario can be described easily by Heterogeneous Information Network (HIN) which consists of different types of entities and relationship among them. Due to multiple type of entities and links between them in HIN , it is necessary to find the similarities between the nodes of HIN. In Homogeneous Information Network, there is only single type of node and links in between them. There are many existing methods by which similarity among the nodes of Homogeneous Information Network can be calculated. But those methods cannot be applied for the HIN because semantic meaning behind each path cannot be considered .If we want to apply techniques of Homogeneous Information Network on HIN then we need to project HIN into Homogeneous Information Network which causes loss of Information. So there is a need to apply different techniques or similarity measures on HIN to calculate the similarities between nodes in HIN. There are many similarity measures implemented by researchers for HIN. Similarity search basically concentrates on discovering the most similarity objects for a given query entity. In a comparative analysis section,we have discussed some of the measures used for similarity.**

***Keywords*—Information Network, Types Of Information Network, Similarity Measures**

I. INTRODUCTION

Real world is an interconnected. Majority of data or informative objects, items are connected with each other or interacts with each other. So it forms large, interconnected and world-wide networks. Generally networks which are connected with each other of these kinds are called as information networks[1]. Information network forms a real world scenario that can be extracted from multiple domains. Many applications uses similarity search. Basically two objects are considered to be similar if many paths exist between them.

A. Types of Information Network

- Homogeneous Information Network

In this type of network, nodes are objects of the same entity type and links are relationships from the same relation type.

- Heterogeneous Information Network(HIN)

Heterogeneous Information Network consists of multiple types of entities having different types of relationships in between them. Huge information network is created by connecting real world objects with each other that are having some link in between them. Each link has some semantic meaning. These networks can be converted into HIN by arranging objects and relationship into more than one type of entities. Some of the examples of Heterogeneous Information Networks include [1]:

I) Bibliographic Information Network
II) Health care Information Network
III) Twitter Information Network
IV) Flickr Information Network

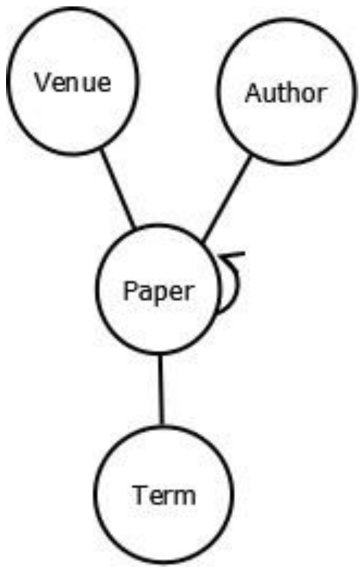

Figure. 1. Schema Of Bibliographic Information Network

Figure 1 [1] shows the network schema of Bibliographic Information Network. Similarly the schema for other types of information network can also be drawn.

B. Searching for similarity: Similarity Search

Basic purpose for Similarity search is to find the most similar information for the query into the huge dataset col-

lection. For example, Let us consider a database in which some people are interested to find the k(any number) neighbors which are most near to it for a specified spatial object in information retrieval. In order to find similar documents for a given document or for a given list of keywords. In a similar manner, it is very important to provide good functions for similarity search in the information networks to find similar object for a given object. Let us consider the bibliographic information network in which a user is required to find most similar authors for a mentioned author, or the most similar location for a given location. Attribute based similarity search based on attributes are different from link based similarity because links forms a important role for measure similarity in information networks, specifically when the entire information about attributes for objects is not easy to obtain [1][2]. The similarity measures used for Information Network of homogeneous type dont take the importance of different types of objects and links into consideration. Use of such measures into heterogeneous networks has some limitations: If user just require to differentiate objects who are of the similar type then referring the paths of different types gives different semantic meaning than actual meaning and it is not useful to mix them and measure the similarity without differentiating their semantics. Similarity Search in a Information Network focuses on most similar node for a given node. Similarity between nodes in information network is derived from the similarity score given by similarity measures.

C. Types of similarity measures

Similarity measures used in the Heterogeneous Information Network are Pathsim,HeteSim,Netsim,JoinSim. Pathsim similarity measure is used to find similar objects of same type in HIN. It is used to find the peer objects in the network. Framework for similarity on the basis of meta path is used to differentiate the semantic meaning of each path. In any Information network if the meta path chosen by the user is different then the values returned for the most similar object pair is different. PathSim measure requires expensive matrix multiplication operations to locate most similar object for a given object because the similarity between query and every other object of same type in a network is needed to be calculated. In order to support online query processing fastlyi.e to find more similar object for a given query for huge network , a method is used in which materialization of the meta paths whose length is small is done partially and then integrate them to get similarity based on meta-path of long length[2]. HeteSim measure is used to find the relevance score of the objects sometimes the relatedness among the objects that are not of same type also needed to be measured. i.e sometimes it is necessary to find how two objects are likely to be related. Relevance search task in heterogeneous information network is , measure the how much related the objects are? Difference between similarity search and relevance search is that similarity search measures the similarity of objects with the similar type but the relevance search measures the relatedness of heterogeneous objects[3]. Netsim measure is used to search the similar objects in a network schema of x-star fashion. X-star network consist of centers having links among them. Each center in the network has some attributes[4]. Joinsim measure is used to find the set of similar objects to join condition specified by user. PathSim measure is able to capture the different meaning among the links, but it is not able to handle efficient similarity join operation in large networks,so there is a need to define a new kind of similarity measure[5].

D. Application of Similarity Search

- Link prediction:

  In social network, many users are connected with each other having many entities. Examples of social networks include the group of all researcher or writer in any particular stream. The line connecting them joins the pairs of author who have co-authored papers. Let us take example that given a scenario of a social network then how it can be predicted or known that which new communication within its users is most probably to occur in the near future? It is similar as finding which two users are more similar? This problem can be stated as a link prediction problem [6].

- Detection of duplicate web documents:

  Near-duplicate web documents are existing in large quantities. Two such kind of near-duplicate web pages varies from each other in a very small portion. Such changes are not so important for web search. There- fore web crawlers efficiency is incremented if it can detect whether a web page which is crawled recently is or not a near-duplicate of a web page which is crawled previously. The research contributions done are as following: Fingerprinting technique proposed by Charikar is good for this purpose. And algorithmic technique for identifying existing f-bit fingerprints that change from a given fingerprint is also proposed. Standard checksumming technique can easily identify duplicate copy of documents, but there is difficulty into identify the documents which are near-duplicate of each other. Such kinds of documents are having same text but there is change in a small part in a document like timestamps, counter, and advertisements. These small changes are not so important for web search[7].

Section II is related work,Section III is the methodologiesi.e similarity measures and section IV is Comparative Analysis of Similarity Measures.

II. RELATED WORK

Yizhou Suny, Jiawei Hany, Peixiang Zhaoy, Zhijun Yiny, Hong Chengz, Tianyi Wu proposed RankClus method[8]. In this paper, the problem of generation of clusters for objects types which is given and information about ranking for objects of all the types is addressed on the basis of these clusters in a multityped information network (i..e heterogeneous). RankClus is a framework used for clustering which generates clusters that are integrated to ranking directly. Ranking is applied individually on the basis of K clusters which are given initially. After this, Decomposition of every object in a vector having K-dimensions is done with the use of mixture

model. Rank distribution measures component coefficient. Re-assignment of the objects is done to the cluster which is most nearer to it in a recent measure space for improving quality of clustering. Because of this, quality of forming clusters and quality of assigning ranks are improved. It means that accuracy of clusters is increased and ranking is also becoming more meaningful. This progressive refinement process repeats until small change is possible.

Chuan Xiao Wei Wang Xuemin Lin Haichuan Shang proposed Top-k set Similarity Join[9]. This method gives top-k pairs of record which is ranked on basis of their similarities. Proposed algorithm uses the Prefix-Filtering principle. Basic idea is that two records are similar to each other if some part overlaps with each other. The main technical challenge is that the value for similarity of the k-th largest pair is not known.

Yasuhiro Fujiwara, Makoto Nakatsuji, Makoto Onizuka , Masaru Kitsuregawa proposed Fast and Exact Top-k Search for Random Walk with Restart method[10]. It gives good proximity scores between the two nodes in graph. This method acts as a basic principle for multiple applications from different domains like system for recommendation, automatic captioning of image, and prediction of link. The aim of this paper is to search nodes that have top-k maximum proximities for the specified node. Previous solution to this problem searches nodes correctly but exactness was very expensive. K-dash solution calculates proximity for a selected node in a efficient manner with the use of sparse matrices and it avoids computations for proximity which are not necessary while searching the top-k nodes. Ideas behind K-dash contains two main concepts as computation of sparse matrix and estimation of tree.

Xiao Yu, Yizhou Sun, Peixiang Zhao, Jiawei Han proposed paper for Query-Driven Discovery of Semantically Similar Substructures in Heterogeneous Networks[11]. A query for subgraph is given and the system finds the subgraphs whose structure is similar to it and its meaning is also same in a huge information network. Since techniques of data mining are used for getting semantically same nodes discovery is used as a term to describe this process. For achieving more efficiency as well as scalability, a framework called filter-and-verification search is designed. It can first generate subgraph candidates which are promising with the use of offline indices and then check candidates with the iterative process of pruning matching.

Zheng, Lei Zou, Yansong Feng, Lei Chen Dongyan Zhao proposed Effcient Top-K SimRank-based Similarity Join technique[12].Time complexity and space complexity for calculating the similarities for all pairs of nodes based on Simrank is more specifically for huge graphs. To solve this problem, in this paper, problem of similarity join based on similarity is studied, which searches k most same pairs of entities with the highest SimRank similarities within all possible pairs. Every node is ciphered as vector by addition of its neighbors as well as convert the computation of the similarity based on Simrank similarity measure among both nodes into calculation of dot product among the vectors corresponding to it. An improved two-step framework is proposed to calculate top-k similar pairs with the use of the vectors. In first phase, a group of candidate entities is identified from which the top-k similar pairs must be computed. In second phase, WAND algorithm based on tree structure is proposed to identify the k most similar pairs among the candidate entities efficiently.

III. METHODOLOGY

Methodology adopted for various methods are described below.

A. PathSim

First a baseline method is proposed which is called as the PathSim-baseline which computes the similarity score and then clustering based pruning method is proposed in which not promising candidate objects are pruned .Basic idea behind the baseline method is given below:

i. Multiplication of a query vector and matrix is done.

ii. Similarity score is calculated by

$$s(i,j)= 2M(i,j)/(M(i,i)+M(j,j)) \quad (1)$$

iii. Similarity score is sorted in descending manner to find most similar objects.

Equation (1) defines the formula for calculating the similarity score. Pruning algorithm is given below:

i. Partially formed Commuting matrix and its co-clusters of size 3x3 and query vector is given.

ii. For object of query, The compression of query vector is done into the query vector of aggregated form whose length is 3.

iii. First Upper bounds is computed for the similarity in between the clusters of target and query on basis of the query in aggregated form and cluster vectors in aggregated form ; Second, the upper bound for similarity score between all three candidates in cluster and query is computed with the use of aggregated vectors for every target clusters if that cluster cannot be pruned; Third, calculation of exact similarity is done with the use of query vector in non-aggregated form as well as candidate vectors if it is not possible to prune candidates[2].

B. HeteSim

Relevance score given by Hetesim measure is based on the number of incoming links of destination node and outgoing links of source node When the idea used in simrank measure is applied to the heterogeneous networks, it faces challenges. First the relatedness of different types of objects is path- constrained. Similarity measure based on paths are required to be designed because relevance path captures the semantics information constrains the walk path. Second,For symmetric or asymmetric path,the measure should be able to calculate relatedness of heterogeneous object pair with a single score[3].

C. NetSim

Meeting probability is important in this measure. Meeting probability is used to calculate the similarity among attributes. Similarity is calculated over the attribute network which is extracted from the entire structure of x-star network. In this technique attribute similarity is calculated by SimRank measure. The similarity between centers is computed online same as the similarities of attributes based on the basic idea that centers which are similar to each other are likely to be linked with similar attributes. Reformalization of importance of link and the importance of relation is done without considering relationship between centers. Online query processing should be done very fast .To support it, algorithm for pruning is developed. It is necessary to build the pruning index to develop the pruning algorithm. Candidate centers which are not useful are pruned by using pruning algorithm[4].

D. JoinSim

For integrating various semantics for paths, a matrix is used to describe the sequence of relation between the source type and destination type of a join path P into this similarity measure. For any node which belong to start or source type, the corresponding row in the matrix is called as feature vector. LSH indexing is widely applied to solve the approximate similarity search problem . Basic idea is that search the family of hashing functions in such a way that, for a hash function choosen randomly from the hash function family, then two objects having more similrity between them are to be hashed into the same bucket with a higher chances. The limitation of the LSH is as explained further.It needs number of hash tables to cover most of nearest neighbors. The storage cost of each hash table is proportional to the data set size. So hash table uses nearby buckets. It discards pairs of objects located in buckets which are far from each other and thus cut down many similarity computations.LSH is improved by BPLSH based PS-join: PS join based on LSH with Buckets Pruning method because it may possible that some pairs within two nearby buckets may or may not be included in final result. To efficiently capture a similar pair of objects that needs to be hashed into separate buckets, LSH table is provided with a extra bucket array. Bucket array stores a set of values for upper bounds and lower bounds for every bucket depending on the distance between its object members and the random hyperplanes which forms LSH, this information is used to prune the buckets which are not needed to be compared. This helps to provide a more better solution to process the top-k similarity join[5].

IV. COMPARATIVE ANALYSIS

Table I describes the table for comparative analysis of the similarity measures discussed above on the basis of parameters. In this section, we are going to discuss about the comparative analysis on the various similarity measures which have discussed above i.e PathSim, HeteSim, NetSim, JoinSim on the basis of parameters such as: Symmetry, Path-based, Triangle Inequality and Pruning.

TABLE I. COMPARATIVE ANALYSIS OF SIMILARITY MEASURES.

	PathSim	HeteSim	NetSim	JoinSim
Symmetry	Yes	Yes	N.A	Yes
Triangle Inequality	No	No	No	Yes
Path Based	Meta path based	Relevance path based	N.A	Join path based
Pruning	By Co-clustering algorithm	N.A	By NetSim Pruning algorithm	By bucket pruning algorithm

- Based on the comparison table, here it can be said that JoinSim method is best if we want to find the top k similar pairs of objects based on the join path specified by user in a HIN. Because JoinSim method satisfies the Triangle Inequality Property, it can be used to find out similarity join in a large network. A triangle inequality is able to find most of the interesting structure on a metric space viz. convergence. For e.g, the fact that any convergent sequence in a metric space is a Cauchy sequence is a direct consequence of the triangle inequality, if any x_n and x_m such that $d(x_n, x) < \epsilon/2$ and $d(x_m, x) < \epsilon/2$ is choosen, where $\epsilon > 0$ is given and arbitrary then by the triangle inequality, $d(x_n, x_m) \leq d(x_n, x) + d(x_m, x) < \epsilon/2 + \epsilon/2 = \epsilon$, so that the sequence x_n is a Cauchy sequence [13].
- PathSim method is able to capture the semantic meaning of the meta-path but it is not able to handle similarity join in large network because it does not satisfy the triangle inequality property and Pathsim measure do not support LSH which is satisfied by the Joinsim measure and this drawback prohibits its application for similarity join in multi dimensional spaces [2].
- Performance of the PathSim method has limited behavior under meta path of infinite length. But it is good for relatively shorter meta path [2].
- Less time is required for the NetSim measure because the attribute network is formed from the entire x-star network and the scale and size of the attribute network is less than actual x-star network[4].
- Netsim method can use the existing techniques for

optimization for further improvement of results but cannot find semantic meaning for relation in proper way[4].

V. CONCLUSION

HIN have drawn more attention as a newly developed network model. Searching for top k pairs of same objects is required in many real life applications. Multiple task of data mining are being explored in Heterogeneous Network, which includes clustering of objects, search for similar objects and classification. Searching for Similarity is a basic and important operation which is required in many applications.Various similarity measures along with its detailed methodology and algorithm are discussed here. We have discussed use of each similarity measure .The comparative analysis of all four similarity measures is done with help of parameters like path-based, symmetry, Triangle Inequaliy and pruning. So here it can be concluded that Joinsim method i.e Similarity measure is best for finding the top k similar pair of objects.

REFERENCES

[1] Y. Sun and J. Han, "Meta-Path-Based Search and Mining in Hetero- geneous Information Networks",TSINGHUA SCIENCE AND TECH- NOLOGY ISSNll1007-0214ll01/10llpp329-338 Volume 18, Number 4,August 2013

[2] Y. Sun, J. Han, X. Yan, P. Yu, and T. Wu, "Pathsim: Meta Path-Based Top-k Similarity Search in Heterogeneous Information Networks," Proc. Intl Conf. Very Large Databases (VLDB), pp. 992-1003, 2011.

[3] C. Shi,X. Kong,Y. Huang,Philip S.Yu,Bin Wu,"HeteSim:A General Framework for Relevance Measure in Heterogeneous Information Net- works" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA EN-ENGINEERING, VOL. 26, NO. 10, OCTOBER 2014

[4] M. Zhang,H. Hu,Z. He,W. Wang,"Top-k similarity search in heterogeneous information networks with x-star network schema",ARTICLE in EXPERT SYSTEMS WITH APPLICATIONS FEBRUARY 2015

[5] Y. Xiong, Y. Zhu, and Philip S. Yu, "Top-k Similarity Join in Hetero- geneous Information Networks" IEEE TRANSACTIONS ON KNOWL- EDGE AND DATA ENGINEERING, VOL. 27, NO. 6, JUNE 2015

[6] David Liben-Nowell,Jon Kleinberg,"The Link Prediction Problem for Social Networks",January 8, 2004

[7] Filip Radlinski,Paul N. Bennett,Emine Yilmaz, "Detecting Web Documents using Clickthrough Data",WSDM11,fEBRUARY 912,2011, Hong Kong,China

[8] Y. Sun, J. Han, P. Zhao, Z. Yin, H. Cheng, and T. Wu. "Rankclus: in- tegrating clustering with ranking for heterogeneous information network analysis". In EDBT09,565576, 2009.

[9] C.Xiao, W. Wang, X. Lin, and H. Shang. "Top-k set similarity joins". In ICDE09, 916927, 2009.

[10] Y.Fujiwara, M. Nakatsuji, M. Onizuka, M. Kitsuregawa,"Fast and Exact Top-k search for Random walk with Restart",Preceedings of the VLDB Endowment Vol.5,2012

[11] X.Yu, Y.Sun, P. Zhao, J. Han,"Query-Driven Discovery of semantically Similar Substructures in Heterogeneous Networks",KDD,Beijing, China 2012

[12] Wenbo Tao Minghe Yu Guoliang Li "Efficient Top-K SimRank-based Similarity Join",Proceedings of the VLDB Endowment, Vol. 8,No. 3, 2014

[13] www.wikipedia.com