ISSN: 0974-2115
www.jchps.com Journal of Chemical and Pharmaceutical Sciences

Meta Path Based Top-K Similarity Join In Heterogeneous Information Networks

S.Sheik Faritha Begum[1*], A.Rajesh[2], M.Vinnarasi[2]
[1]Computer Science and Engineering, Bharath University, India.
[2]Computer Science and Engineering, C.Abdul Hakeem College of Engineering and Technology, Melvisharam, India.
***Corresponding author: E-Mail: sfaritha@gmail.com**

ABSTRACT

Heterogeneous information networks (HINs) has received growing attention in a newly emerging network model. Clustering, classification and similarity search are the data mining tasks explored in HINs. Similarity join operation is required for many problems which is attracting attention from various applications on network data which includes friend recommendation, link prediction and online advertising. Similarity join is studied in homogeneous network but not studied in heterogeneous networks. The heterogeneity of the HINs are ignored completely by the previous research on similarity join which takes different semantic meanings. A Meta-Path-based Similarity join (MPS-join) method to return the top k similar pairs of objects based on the user specified join path in a HIN is proposed and to prune expensive similarity computation by using BPLSH (Bucket Pruning based Locality Sensitive Hashing) indexing. When compared to the existing Link-based Similarity join (LS-join) method, this MPS-join method derives various similarity semantics. The experimental results on real data sets shows the efficiency and effectiveness of the proposed approach.

KEY WORDS: Heterogeneous network, graph, similarity search and similarity join.

1. INTRODUCTION

Heterogeneous information networks are interconnected multiple typed objects. These are the logical network involving multiple typed objects and links with different relations. Bibliographic network, social media network and knowledge network are the few examples of HINs. These networks are a graph data model where the nodes and edges are explained with class and relationships. Yago, DBLP are the large and complex datasets modeled as HINs. Data integration, information retrieval and bio informatics in medical field are the areas where top k similarity join has been applied and studied. The Link based similarity join (LS-join) on networks or graphs is studied previously which ranks the high similarity scored objects. The two sets of records will return the pair of records by their similarities which are more than the threshold using similarity join. Duplicate web page detection, data integration and data mining are the applications of similar joins.

The social networks and the World Wide Web have attentions from the researchers in computer science, social science, physics, and biology for the analysis of information networks. Sequence of node classes and edges between two nodes represent meta paths in a HIS, which is used for information retrieval, decision making and product recommendation. Ranking, community detection and link prediction are the different functions for mining the networks. Meta paths are used to capture semantic relationships in multiple type objects which are a path over the graph. For searching and mining of the network and to analyze the semantic meaning of the object meta paths provide the useful way.

The meta structure of the information networks are explored by mining the HINs. The information objects, individual agents or groups are interconnected with each other which forms large, interconnected networks. These interconnected networks are also known as the information networks. For example, a bibliographic network extracted from DBLP with multiple types of objects including authors, papers, terms and conferences and links between objects correspond to different relations, such as the writing or written by relation between authors and papers.

Related Work: Top K similarity join computes the similar record pairs, where users experiment with various threshold values which has long running time and also have many results. It measures most similar object pairs where multi valued object sets are involved. HINs are the multiple typed objects where the social network, friendship network or web page networks forms homogeneous information network which calculates the semantic of paths. Link-based similarity join which finds the links between the graph and relationship of the sets of nodes so that the nodes with highly similar to each other by Sim Rank (SR) and Personalized Page Rank (PPR) measures. Sim Rank based Join (SRJ) query is used to find the vertex pairs which satisfies the threshold in the graph. It is verified in different types of queries like sub graph search, pattern match query and sim join query. Multi-label classification from the large space of label sets which is exponent to the candidate labels. These classification focuses on the correlations with different class labels that exploits complex linkage information in heterogeneous networks. The discounted hitting time (DHT) is the random walk measure for graph node pairs which has various applications in link prediction and reputation ranking. It returns k lists of n nodes from the groups of nodes.

2. PROPOSED SYSTEM

A Meta Path-based Similarity join (PS-join) method to return the top *k* similar pairs of objects based on any user specified join path in a heterogeneous information network. Study how to prune expensive similarity computation by introducing BPLSH (Bucket Pruning based Locality Sensitive Hashing) indexing. Compared with existing Link-based Similarity join (LS-join) method, MPS-join can derive various similarity semantics. Experimental results on real data sets show the efficiency and effectiveness of the proposed approach.

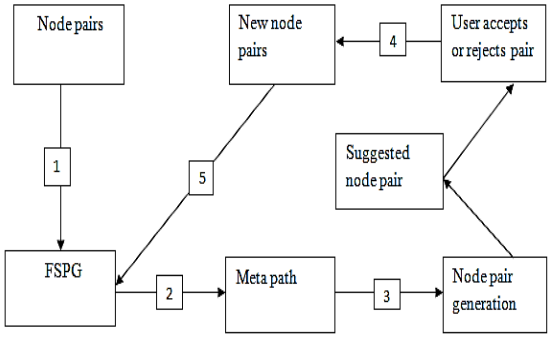

Figure.1. System architecture diagram

In the first step, users provide example pairs to FSPG (Forward Stage wise Path Generation). This algorithm generates meta-paths which are then used to generate new node pairs e.g., via similarity search and join operations. From these node pairs, users will select those pairs that are similar. The accepted node pairs are then input to FSPG to iteratively refine the meta-paths generated. An advantage of this framework is flexibility where users can review and change the meta-paths generated previously, they can also provide new example pairs later after the meta-paths are generated.

Positional Filtering and Suffix Filtering: The positional filtering and suffix filtering technique can be employed here so as to reduce the number of pairs to be verified for their similarity value. Each pair identified by accessing inverted lists is tested under positional filtering and the following suffix filtering before verification is performed. For positional filtering, we use the lowest similarity value of temporary results, sk, if more than k temporary results have been obtained. The similarity value is regarded as a threshold, and the minimum required overlap between x and y is computed for each pair (x, y) to be verified. With the positions of the first common token in both x and y, we can estimate the maximum possible overlap of (x, y), and then compare with the minimum required value. A pair is admitted for verification only if the estimated maximum possible overlap is no smaller than the minimum required value. For suffix filtering, sk is considered as a similarity threshold and converted to a Hamming distance threshold for each pair to be verified. Perform suffix filtering under the Hamming distance constraints and remove the disqualified pairs before performing verification.

Pruning: In this module NBLSH PS-join approach is used which consists of three steps. Firstly, all nodes from the joining set are preprocessed into feature vectors by using relation matrix. Secondly, a random projection based LSH method constructs the indexing structure. Finally, a nearby buckets technique applied to LSH generates candidates and then finds top *k* pairs. The problem of the NBLSH similarity join method is that it needs to check every node pairs from two different buckets within *w* steps. Join Sim needs to be calculated even though some of these node pairs have no chance to become among the top *k* similar pairs. In fact, given a pair of buckets, if the maximum similarity between them is no greater than the current *kth* maximum similarity, can prune them immediately without affecting the final result. Specifically, when we are generating the candidate set, first compute the upper bound of Join Sim between buckets. If the upper bound is less than or equal to the pruning threshold, *z*, then node pairs from these two buckets are discarded.

Optimization: In this module the efficiency of our project by comparing both nearby pruning and Bucket pruning family algorithms exploit the ascending ordering of record sizes and the global ordering of tokens. In order to further optimize the incremental LS join and MPS-join algorithm; exploit the ascending ordering of the lowest similarity value among k temporary results, and the descending ordering of the similarity upper bound.

Algorithms:

Top-K join Algorithm:

Input: R is a collection of records in which each record has been canonicalized by a global ordering O.

Output: Top-k pairs of records (x, y) is ranked by their similarity values.

E ← InitializeEvents(R);
T ←InitializeTempResults(R); /* Store any k pairs as temp results in T */;
Ii ← Ø(1 ≤ i ≤ |U|);

```
while E ≠Ø ; do
(x, px, spx) ←E.pop();
if spx ≤ T[k].sim then
break;      /* stop here */;
w ←x[px];
sk  ←T[k].sim;
for j = 1 to |Iw| do
y ←Iw[j];
if |y| Є [sk|x|, |x|/sk] then    /* size filtering */
sim(x, y)← CalcSimilarity(x, y);
T.add((x, y), sim(x, y));
sk← T[k].sim;
Iw ← Iw U { x };  /* index the current prefix */;
px ←SimilarityUpperBound-Probe(x, px + 1);
E.push(x, px + 1, s′px);
return T
```

BPLSH-MPS Algorithm:

Input: data set D, int k (the number of similar pairs to search for), int m (hash vector length), int t (the number of hash tables), int w (the maximum hamming distance between buckets)

Output: kP Set (the sorted list of top k pairs in decreasing order of Join Sim)

Build advanced LSH indexing for D
Initialize kPSet to; and pruning similarity z to -1;
for each hash table $Th^i{}_m(1 \leq i \leq t)$ do
for each bucket B do
Generate all possible node pairs from B and add them to kPSet if their Join Sim is larger than z;
endfor
for each bucket pair (Bs,Bs′), s ≠ s′ do
if the hamming distance between Bs and Bs′<w
Check each mismatch bit in the hash vectors of Bs and Bs′ and prune (Bs,Bs′) once it satisfies the pruning strategy at a mismatch;
if (Bs ,Bs′) is not pruned then
Generate all possible node pairs from Bs and Bs0 and add them to kPSet if their JoinSim is larger than z;
end if
end if
end for
end for
return kP Set;

3. RESULT

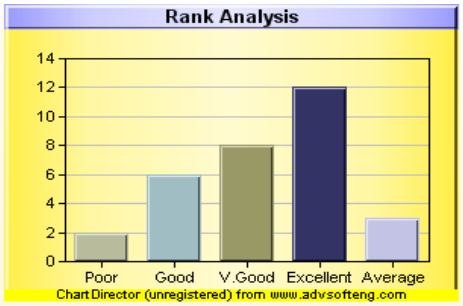

Figure.2. Similarity based rank analysis

In the figure.2, the ranking is done based on similar user reviews (poor, good, excellent, average) for the books dataset which shows the similarity between various heterogeneous searches in the information network. Thus efficiency is measured by running time, which is the time to return the list of top k similar node pairs.

4. CONCLUSION

The problem of finding top k pairs of similar nodes arises in many applications. Two key insights which had not been identified in previous work is presented: (a) To motivate and formulate the new and important problem of path based top-k similarity join in heterogeneous network applications; (b) To develop new pruning and optimization

techniques to solve the similarity join problem in heterogeneous networks based on BPLSH by exploiting the property of the proposed meta path-based similarity measure.

REFERENCES

Bustos B, and Skopal T, Non-metric similarity search problems in very large collections, in Proc. 27th Int. Conf. Data Eng., 2011, 1362–1365.

Kong X, Cao B, and Yu P.S, Multi-label classification by mining label and instance correlations from heterogeneous information networks, KDD, 2013, 614–22.

Kong X, Yu P.S, Ding Y, and Wild D.J, Meta path-based collective classification in heterogeneous information networks, CIKM, 2012, 1567–71.

Kong X, Yu P.S, Ding Y, and Wild D.J, Meta path-based collective classification in heterogeneous information networks, in Proc. 21st ACM Int. Conf. Inf. Knowl. Manage., 2012, 1567– 71.

Lao N, and Cohen W.W, Relational retrieval using a combination of path-constrained random walks, Mach. Learn., 81 (1), 2010.

Lee H, Ng. R.T, and Shim K, Similarity join size estimation using locality sensitive hashing, Proc. VLDB Endowment, 4 (6), 2011, 338–49.

Shi C, Kong X, and Yu P.S, Relevance search in heterogeneous networks, EDBT, 2012, 180–91.

Sun L, Cheng R, Li X, Cheung D.W, and Han J, On link based similarity join, PVLDB, 4 (11), 2011, 714–25.

Sun Y, Han J, Yan X, Yu P.S, and Wu T, Pathsim: Meta pathbased top-k similarity search in heterogeneous information networks, PVLDB, 4 (11), 2011, 992–1003.

Sun Y, Norick B, Han J, Yan X, Yu P.S, and Yu X, Integrating meta-path selection with user-guided object clustering in HINs, KDD, 2012, 1348–56.

Tao Y.F, Yi K, Sheng C, and Kalnis P, Efficient and accurate nearest neighbor and closest pair search in high dimensional space, ACM Trans. Database Syst., 35, 2010, 20.

Xiao C, Wang W, and Lin X, Ed-join: An efficient algorithm for similarity joins with edit distance constraints, VLDB, 2008, 933–44.

Xiao C, Wang W, Lin X, and Shang H, Top-k set similarity joins, ICDE, 2009, 916–27.

Zhang W, Cheng R, and Kao B, Evaluating multi-way joins over discounted hitting time, ICDE, 2014.

Zheng W, Zou L, Feng Y, Chen L, and Zhao D, Efficient sim rank- based similarity join over large graphs, Proc. VLDB Endowment, 6 (7), 2013, 493–504.