Scientific Journal of Impact Factor (SJIF): 4.14

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 3, Issue 9, September -2016

A Cummulative Study on Sub Graph Matching(Survey Paper)

[1]Ms.Bhati N. Durgade, [2]Mr.K.S.Kadam

[1,2]*Computer Science and Engineering, DKTE Society's Textile and Engineering Institute, Ichalkaranji(An Autonomous)*

Abstract —*Popularity of Social portals have connected user under similar domain sharing similarity in taste and view. Analysis of this complex relation to find similar groups has been active area of research in graph based search. Numerous research works on sub graph matching methodology have developed to identify similar users. Although major research issue is hard rules of similarity detection fail to yield better results. This research manuscript presents sub graph matching technique to recognize user relatedness based on pattern recognition of query. Pearson relation, pattern analysis of frequent items and better reduction techniques enhance system performance. Recursive review on fifteen articles has been done to find out research analysis question and proposed system is based on techniques to solve each and every issue. Comparative examination presented.*

Keywords-*Sub graph matching, Pre-processing, Graph, Data Graph.*

I. INTRODUCTION

Similarity detection based on graph assist in retrieving and mining useful information from network of social websites, biological networks and numerous applications of image mining and computer vision. Finding relatedness with sub graph is principle graph query. Large Graph "g" input graph query "q", sub graph similarity detection retrieves exactly matching sub graphs in "g", with exact similarity in vertex tags and structure of graph.

Open Challenges and Issues:

Problem1: Real world scenario sub-graph matching with exact vertices is not applicable, as a graph vertex might consist of elements with rich features but won't match exactly.

Problem 2: Dynamic sub graph matching is common observed challenge.

Problem 3: Huge relations are generated in graph similarity detection and require reduction techniques to avoid complexity overhead. Pre-processing is widely implemented information reduction technique to avoid over head in data mining. Pre-process enhances effectiveness of processes task. While extracting useful information from unstructured data mining domain urges for better useless data removal techniques. Usually unwanted information leads to loss in precision and accurateness. As such pre-processing is core methodology in any data mining process. In generalized view it's consist of four layers Data cleaning, Integration, transformation and reduction. These techniques are as below.

[1.] Cleaning of data: - Method of collecting missing information, leveling irrelevant information and eradicating boundary items. Cleaning process assist to find correlated items.

[2.] Integration of Data:- Method of collecting information from numerous supplies at one place as to retrieve information easily from only one source.

[3.] Transformation of data: - Method of putting information in more suitable structure as it could be efficient in retrieving method. Transformation uses diverse sub Methods that as normalizing data values, smoothing, simplification and final integration and many more sub process.

[4.] Reduction of Data: Method multifaceted datasets are reduced to simplified structure without evaluating originality of information. Stem reduction of one most commonly used technique to reduce

This method root terms are mined from word in such a way that meaning of word remains same. On same lines of stemming, lemmatization is similar method which has similar work flow with slight different.Base terms are derived for input word with rule set at base, neglecting POS (part of speech), whereas in lemmatization POS and sense of term is initially derived and then base terms are extracted.

Figure 1 represents four layers of pre-processing

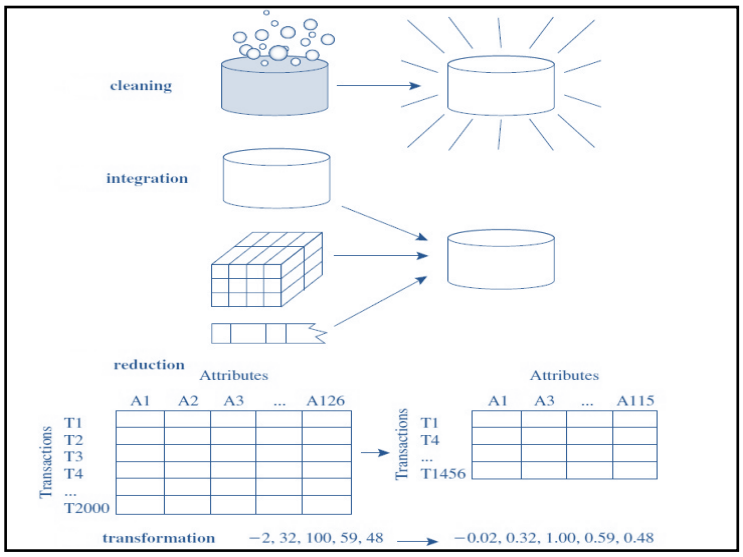

Figure1: Generalized Pre-processing Method[12]

Inspired by above issue and common challenges similarity detection with sub-graph matching framework. Proposed framework maps input query graph, retrieving each and every sub-graph.

Contribution of Manuscript:

- Manuscript introduces open challenges in sub graph matching.
- Background knowledge on graph and Sub graph matching techniques is been introduced.
- Research challenges have been summarized.
- Sub Graph Methodology for large graph search and approximate graph query search is been presented.

II. BACKGROUND

Modeling and solving real life problems objects can be effectively represented with graph. As such knowledge and relationship between entities can be modeled in an easy way. Graph assist in recognizing model based pattern recognition .Research direction in graph based matching are categorized in two classes.

[1.] Inexact Matching

[2.] Exact Matching.

In scenario where one-to one mapping in between two Graph $\boldsymbol{Gd}: \boldsymbol{Data\ Graph} = (Vd, Ed)$ *and* $\boldsymbol{Gm}: \boldsymbol{model\ graph} = (Vm, Em)$It is termed as isomorphism or exact matching. In real time scenarios schematic features may vary i.e no of vertices don't relate objective remains to find non-objective connectivity between Gm and Gd.

Figure 2 presents classification of graph matching techniques.

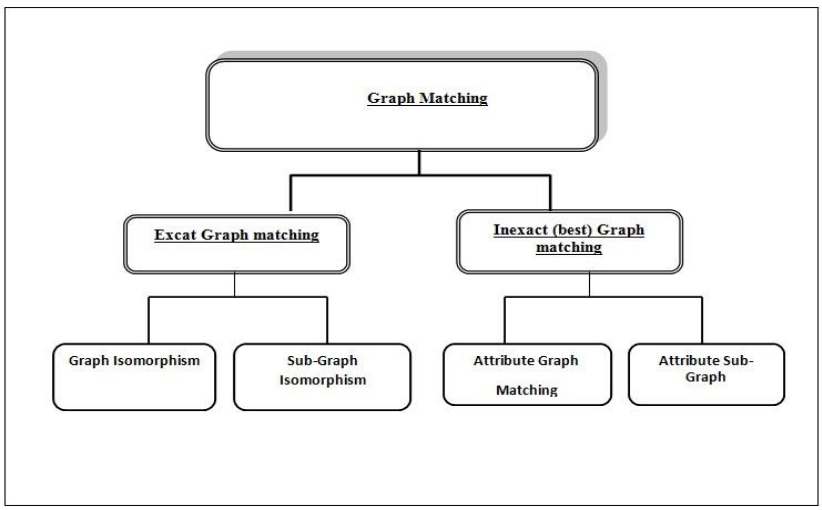

Figure2: classification of Graph based matching techniques [11]

Fitness function is been used to find [11] find best matching graph optimal solution. In Graph matching where Main graph is ***BG***= (V, E) and Small graph ***SG***= (V1, E1), if one to one mapping exists it is termed as sub graph match.

III. LITERATURE REVIEW

A. Survey Technique: Literature review has been carried out scientific portal like ACM, IEEE, and Google Scholar. Key ten articles have been selected based on maximum citation technique. Articles which have max citation score have been surveyed on In exact graph matching papers.
B. Survey Breakdown: every article have been surveyed and breakdown into Algorithm limitations and research scope. Key points have been reviewed and documented.
C.Survey Outcomes: summarized Research Analysis Question (RAQ's) are developed based on survey review .based on this RAQ's proposed Methodology is been selected and innovative algorithm is been proposed.
D.Survey Graph: survey has been done 50% IEEE 25% ACM and 25% other sources.

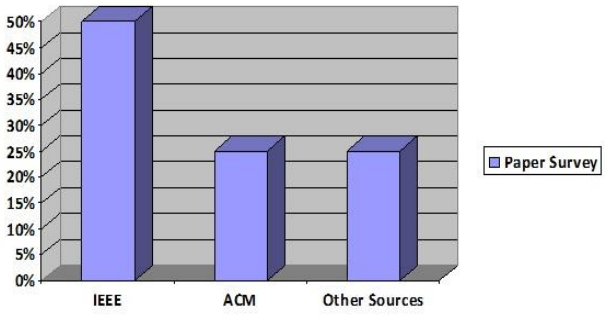

Figure 3: Survey distribution
Figure 3 presents survey distribution of papers used in survey.

E. Survey

[1]Algorithm: Insights of cellular structure can be understood with PPI(protein-protein interaction) network analysis. Author presentsC-GRAAL (Common neighbors based Graph Aligner) heuristic procedure to model relations in networks. It is generalized algorithm can be extended to electric or transportation network. proposed algorithm helps to identify pattern of interaction based on network topology.
Limitation:Result analysis present algorithm is effective to find alignment in between two networks and can transfer information from one domain to other. Even though graph complexity remains challenge and has not be evaluated.
Research Scope: Algorithm has been tested for small amount of data and finding out its performance on Hugh level of data remains scope of work.

[2] Algorithm: Recognizing duplicate events is essential for process mining or provenance querying. Distinctive attributes like displaced traces, difficult names, avert present information integration from methods performing well. In order to solve this issue author presents composite event matching procedure with different reduction and prediction techniques.**Limitation**:Accuracy decrease with increases reduction ration as of absence of structural and statistical information.**Research Scope:**Theoretical bound estimation is scope of future work and has not been studied.

[3] Algorithm: In current scenario Graph data management is required for effective and efficient use of information generated from graph. As such graph pattern recognition algorithms are profoundly required. R-join filter-fetch algorithm for graph pattern recognition is been presented. Cluster join index is been used to fetch patterns from graph code. Further an optimized is achieved with comparison of R-semi joins and R-Joins.**Limitation**:Proposed algorithm can only be implemented for directed and acyclic class of graphs. Work is been tested for small dataset Xmark(16nodes).
Research Scope: Better scalability can be achieved in DP either by reducing I/O cost or any alternative technique remains scope of work.

[4]Algorithm: Huge scale graph-structured information is been generated in network and require efficient query process for handling this information. Searching and retrieving graph structure is major question. SPATH an indexing mechanism is been presented. It assists in space reduction and is flexible in index generation. Here query is been first split in short paths selecting candidates with goodvalue to optimize process.**Limitation**:Network change over time and hence incremental graph approach is required. Noise and failure are need to be addressed. **Research Scope:**Approximate graph querying is unaddressed.Incremental indexing with approximate graph search is future research direction

[5]Algorithm: Information collections frequently contain discrepancy as of variety of reasons and it is pleasing to be able to classify and determine them proficiently. Set similarity assist in cleaning information .numerous research work present set similarity computation techniques. Proposed system is based on TF/IDF with variants similarity techniques for RDBMS. Procedures build effective index and algorithm fro resolving every query. Proposed work is procedure built on three set similarity algorithms.**Limitation**:Only short first algorithm has effective performance and other two need to optimized.**Research Scope:**Top-K processing with parallel processing devised with new similarity measure is scope of work.

[6]Algorithm: Biological information is huge in volume and effectively represented with Graph structure. As such effective indexing is urge and research direction in graph matching. Volume of graph is challenge as previous research focus only on ten vertices and small volume graphs. Proposed work presents new distance measurement presenting frequent substructure idea. GADDI algorithm find sub graph matches and is been optimized with dynamic scheme. System is been tested for scalability.**Limitation**:Work does not highlight limitations. Scalability has been focus but no such computations have been presented. As such testing this system with real time scenario and applications is required. **Research Scope:**Testing system on real time application and measuring scalability of system. Extended work can done on scalability computation.

[7]Algorithm: Effective and efficient graph querying procedures are current state of requirement for handling huge bioinformatics information. Dataset on bioinformatics are incomplete and hence need approximate sub graph matching. Further procedure assist in correct indexing **Limitation**:procedure.Large query matching is still challenge for system and needs better approach in handling.**Research Scope:**Future modular property divide-and-conquer procedure can be implemented. Large queries can be further subdivided to small queries and sub graph matching can be computed against this small block queries.

[8]Algorithm: Graph isomorphism is one of challenge in sub graph matching .proposed system spatial complexity is been reduced in reference to time and space. Work focuses on large graph structure. **Limitation:**Comparative work is been evaluated only in between two algorithms. **Research Scope:**One proposed algorithm and other ulmans algorithm of 1976.Pruning techniques can be used for future reducing sub graph matching.

[9]Algorithm: Sub graph match is active research in relational databases systems. Although small research is present on modeling sub graph match on social networks and biological network. Proposed algorithm retrieves top-k sub graph relative to input query Q based on score function. Balanced tree procedure based on ranked algorithm is been presented.**Limitation:**limitation have been observed in TA matching, prediction about upper bound in left matching achieved is not tight.**Research Scope:**Further pruning techniques can be extended to reduce unnecessary matching's.

[10]Algorithm: Brute force tree search technique is been presented for sub graph matching. Parallel memory is been implemented for isomorphism detection.**Limitation**:Proposed system is tested for very limited class of graph and is very less effective compared to gotheb, corneil.**Research Scope:**Optimization modification can be carried out for better and faster processing algorithm.

[11]Algorithm: Commonly NP hard scenario is been observed in Sub graph matching. Numerous research works try to address sub graph matching in reasonable time with reduction and better index techniques. Still major empirical evaluation exists in numerous research work as no comparative analysis exists on every summarized techniques. Test environments have been evaluated and reimplementation on authors work is been done to avoid discripsincy. **Limitation:**conclusion provides that no such technique is best and hence combined approach or innovative methodology is required**. Research Scope:** new sub graph matching procedure that exploit both good join order selection and selective signature-based pruning.

[14]Algorithm: Graph pattern matching is frequently clear in expressions of sub graph isomorphism,np-complete dilemma. To lesser graph complexity, diverse extensions of graph models have been considered instead. **Limitation:** Proposed work is been done on simulations and requires to be tested in real scenarios. **Research scope:** author focuses to make more strong simulation criteria and develop system on distributed location. Solving cubic time is future scope of work.

[15]Algorithm: Analysis of network and data system can be done efficiently using graph examination. Centralized procedure and time complexity remains major issue. This manuscript focuses on parallel sub graph matching. PSgL a novel algorithm is been proposed for graph matching. Algorithm is based on divide and conquer technique and traverses each node avoiding explicit joins. **Limitation:** algorithm is prototype and can be enhanced for optimized heuristics. **Research Scope:** Framework can be enhanced future to large data graphs.

[16]Algorithm: Difficulty of generating random graphs consistently from set of easy associated graphs having a decided degree series. Author is to offer a procedure designed for sensible use both because of its capability to produce huge large graphs and because it is easy to execute. Research focuseson relatives of heuristics author proposes optimality situation, and demonstrate how this optimization can be arrive in practice. Propose a diverse approach, purposelyconsidered for real-world distributions which outperform other. **Limitations:**Empirical evaluations have been done and comparative analysis is missing.**Research Scope:**Generalized case for graph topology can be achieved in future work.

[17] **Algorithm**:Fewer research work address interactive query based network exploration technique systematically and needs to be addressed in huge networks. Proposed system accepts input graph query and iteratively mines every sub graph matching .filter and verification framework is been devised based on recursive reduction techniques. Proposed work is found to be effective in heterogeneous networks. **Limitations:**proposed procedure is prototype and can be enhanced definitely.**Research Scope:**proposed system can be implemented in recommendation systems and specific application scenarios.

[18] **Algorithm:** In provenance query process recognizing duplicate images is essentially required.Proposed article addresses iterative algorithm for similarity detection .procedure achieves higher performance in terms of accuracy and state of art existing algorithm.**Limitations:**Research work Research Scope: interesting toinvestigate bound of assessment as a future study.**Research Scope:**Algorithm can be enhancedand optimized indistributed system processing.

IV. CONCLUSION

Above Research manuscript presents a literature review work on graph based sub matching. Systematic literature survey has been done on key fifteen survey papers. A survey methodology has been proposed and used is systematic survey. A summarized conclusion is been achieved and problem statement is been devised based on above survey. Conclusion arrived is "Many systems are been introduced to identify the matching sub graphs using similarity between the users. This often yields not much appropriate results due to strict similarity measures. So proposed system uses a technique of identifying correlation between the users for the fired query using pattern identification by incorporating Frequent pattern analysis and Pearson correlation which is been catalyzed by Strong pruning techniques".

REFERENCES

[1.] V. Memisevic and N. Przulj, "C-graal: Common-neighbors-based global graph alignment of biological networks," Integr. Biol.vol. 4, no. 7, pp. 734–743, 2012.
[2.] G. G€ulsoy and T. Kahveci, "Rinq: Reference-based indexing for network queries," Bioinformatics, vol. 27, no. 13, pp. 149–158, 2011.
[3.] R. Di Natale, A. Ferro, R. Giugno, M. Mongiov_ı, A. Pulvirenti, and D. Shasha, "Sing: Subgraph search in non-homogeneous graphs," BMC Bioinformatics., vol. 11, no. 1, p. 96, 2010.
[4.] P. Zhao and J. Han, "On graph query optimization in large networks," Proc. VLDB Endowment, vol. 3, nos. 1/2, pp. 340–351, 2010
[5.] M. Hadjieleftheriou, A. Chandel, N. Koudas, and D. Srivastava, "Fast indexes and algorithms for set similarity selection queries," in Proc. 24th Int. Conf. Data Eng., 2008, pp. 267–276.
[6.] S. Zhang, S. Li, and J. Yang, "Gaddi: Distance index based subgraph matching in biological networks," in Proc. 12th Int. Conf. Extending Database Technol.: Adv. Database Technol., 2009, pp. 192– 203
[7.] Y. Tian, R. C. McEachin, C. Santos, D. J. States, and J. M. Patel, "Saga: A subgraph matching tool for biological graphs," Bioinformatics, vol. 23, no. 2, pp. 232–239, 2007
[8.] L. P. Cordella, P. Foggia, C. Sansone, and M. Vento, "A (sub) graph isomorphism algorithm for matching large graphs," IEEE Trans.Pattern Anal. Mach. Intell., vol. 26, no. 10, pp. 1367–1372, Oct. 2004.
[9.] L. Zou, L. Chen, and Y. Lu, "Top-k subgraph matching query in a large graph," in Proc.ACM 1st PhD Workshop CIKM, 2007, pp. 139–146.
[10.] J. R. Ullmann, "An algorithm for subgraph isomorphism," J. ACM, vol. 23, no. 1, pp. 31–42, 1976.
[11.] Sun, Zhao, et al. "Efficient subgraph matching on billion node graphs." Proceedings of the VLDB Endowment 5.9 (2012): 788-799.
[12.] Y. Shao, L. Chen, and B. Cui, "Efficient cohesive subgraphs detection in parallel," in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2014, pp. 613–624.
[13.] Jinsoo Lee, Wook-Shin Han, "An In-depth Comparison of Subgraph Isomorphism Algorithms in Graph Databases" 39thInternational Conference on Very Large Data Bases, August 26th - 30th 2013, Riva del Garda, Trento, Italy. *Proceedings of the VLDB Endowment, Vol. 6, No. 2.*
[14.] Shuai Ma,Yang Cao, "Capturing Topology in Graph Pattern Matching" 38th International Conference on Very Large Data Bases August 27th - 31st 2012, Istanbul, Turkey. *Proceedings of the VLDB Endowment,* Vol. 5, No. 4
[15.] Yingxia Shao,Bin Cui, "Parallel Subgraph Listing in a Large-Scale Graph" SIGMOD/PODS'14, June 22 - 27 2014, Snowbird, UT, USA.
[16.] Fabien Viger,Matthieu Latapy, "Efficient and simple generation of random simple connected graphs withprescribed degree sequence" Journal of Complex Networks Advance Access published June 9, 2015.
[17.] Xiao Yu,Yizhou Sun, "Query-Driven Discovery of Semantically Similar Substructures in Heterogeneous Networks".
[18.] Xiaochen Zhu, Shaoxu Song, "Matching Heterogeneous Event Data", SIGMOD'14, June 22–27, 2014, Snowbird, UT, USA.Copyright 2014 ACM 978-1-4503-2376-5/14/06 .

