IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 1, 2019, accepted July 9, 2019, date of publication July 15, 2019, date of current version August 9, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2928847

Interactive Graph Data Integration System With Spatial-Oriented Visualization and Feedback-Driven Provenance

KULDEEP REDDY
Computer Science and Engineering Institute, Southeast University, Nanjing 211100, China
e-mail: brkuldeep@yahoo.co.in

ABSTRACT This paper proposes new techniques in a visualization-enabled graph data quality assessment and alignment system leveraging feedback-driven provenance with goal of improving scalability, reducing execution time, and increasing relevance. The proposed system consists of following two components. First, graph data quality assessment with spatial-oriented visualization and feedback-driven provenance; in this section, we propose a new paradigm of feedback-driven provenance in order to streamline the collection of run-time provenance information based on user feedback with the goal of reducing execution time and improving relevance. We apply this idea in the context of graph data quality assessment and alignment, in which we propose a system that leverages user feedback on various components on the schema of a graph database during selection of samples of graphs to maintain provenance of representative nodes of samples. We utilize this provenance of representative nodes of samples to improve the effectiveness of future graph samples during quality assessment task. For the visualization component, we propose a solution based on the spatial-oriented approach with the goal of improving scalability, along with statistics and a visualization system based on the notion of heatmaps that involves utilizing quality information of dataset in order to assign to the spatial locations of various graph vertices on the screen varying degree of color intensity pixels. Second, graph data alignment leveraging spatial-oriented visualization and feedback-driven provenance: in this section, we propose a solution based feedback-driven provenance paradigm discussed earlier in context of graph databases by utilizing graph query logs as a feedback to select relevant data of neighborhoods of nodes that were matched during the process of graph alignment at run-time so as to improve its relevance and reduce execution time as well as a spatial-oriented solution to utilize the graph similarity measures in order to allocate spatial pixel positions to graph vertices as a part of a visual analytics tool with a focus on scalability allowing users to compare graphs visually.

INDEX TERMS Data integration, database usability, graph data model, data visualization, data quality assessment, entity resolution, feedback provenance.

I. INTRODUCTION

The core topics of database research improving its performance through various problems such as indexing, query planning, storage mechanisms, data transformation, entity resolution, schema matching etc. However, the research community has come to realize that improving the usability of databases is also an important problem and has received widespread attention from various universities academic research departments as well as industrial research groups. The focus has been on providing a more convenient interface for the users to search and integrate databases. That is, the problem is that of alleviating the burden on the users to comprehend various database query languages and their complexities and instead provide a more accessible way to allow the user to retrieve the information that the user desires as well.

Web Data models Data models that have been primarily proposed in literature to manage web data are XML, RDF and graph data models. A graph database [18] presents a data model with nodes, edges and properties to represent and store data. They offer more benefits than traditional relational

The associate editor coordinating the review of this manuscript and approving it for publication was Ashish Mahajan.

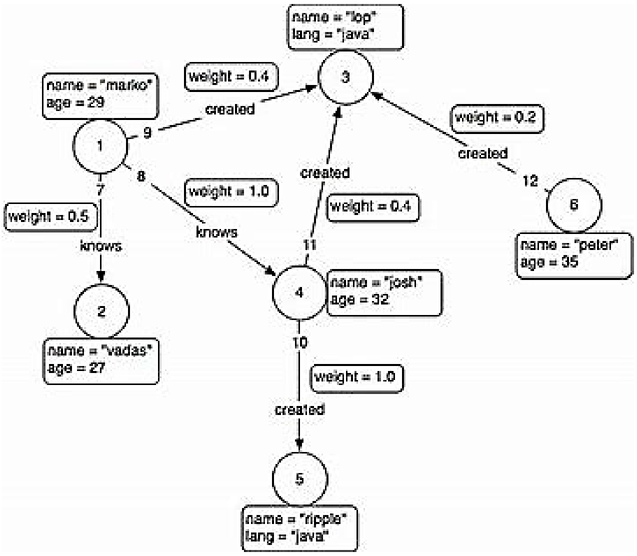

FIGURE 1. Graph data model.

systems in some scenarios by allowing the user to fetch complicated structures which are hard to model in traditional relational model of databases. Specialized graph databases have been designed such as Neo4J which provide their own query languages such as cypher query language.

The figure 1 shown below serves as an example to describe graph data model. The particular example provided is an illustration of a property graph data model, where each nodes and edges can have additional key/value attributes. The nodes represent entities of interest to the user, such as entities with name marko, name apple and edges represent the relationships between entities such as knows created etc. Each edge has a weight associated with it that varies from 0 to 1 which signifies the strength of the relationship. Each node has an additional attribute of age which obviously signifies the age of the entity in question.

Data integration Data integration involves combining data residing in different sources and providing users with a unified view of them [3]. Various components of typical data integration process include - identity resolution, schema mapping, data cleaning and transformation, data quality assessment, data fusion, data provenance assessment, data warehousing etc. Record linkage is the task of finding records in a data set that refer to the same entity across different data sources. Schema matching is the process of identifying that two objects are semantically related while mapping refers to the transformations between the objects. Data provenance documents the inputs, entities, systems, and processes that influence data of interest, in effect providing a historical record of the data and its origins.

Data profiling and quality assessment Data profiling describes the task to constructing a profile or a summary of the databases that allows the user to quickly get a glimpse of the sort of data present [9]. For example, analyzing the database to retrieve a set of integrity constraints or functional dependencies provides the user of a big picture understanding of the database before performing other data integration or querying tasks. Data quality assessment [10] describes the scientific process of trying to understand the quality of the dataset which can ultimately be used to determine which data cleaning task to use to improve the quality of dataset. Reference [8] introduces a web visualization system which provides a case study in which the task to be solved was to research various visual-interactive properties and identify which one of them was more desirable to be used for the case study as hand.

Entity resolution Entity resolution (ER) [11], describes the problem of retrieving and matching the mentions of entities in data. The underlying data can vary from unstructured to structured databases. In relational databases, the problem is often studied in the form of record matching. In unstructured databases, that is text databases, the problem is often studied as an entity linking problem. In graph databases, the problem often appears in the form of graph alignment problem. A number of approaches have been proposed to solve this problem. Approaches include: rule-based methods, pair-wise classification, clustering approaches, and various forms of probabilistic inference. Entity resolution is classified based on the type of input: single-entity ER, in which all mentions belong to a single entity type, relational ER, where real world entities are linked, and multi-entity ER representing the most general problem with potentially linked mentions of different entity types. Similarity join is another fundamental problem in data cleaning and integration closely related to entity resolution which has been implemented as a primitive operator in database systems. Similarity join has been used in various unstructured to structured databases. In unstructured databases, it is known as string similarity join and sometimes set similarity join, where its purpose is to find all pairs of strings whose similarities are above a predetermined threshold according to some distance function such as Cosine, Jaccard, Overlap and Dice, edit distance. In graph databases, the problem often appears in the form of graph similarity join and is solved through various distance metrics particular to graph leveraging its features such as connectivity and various indexing techniques.

Database usability aims to propose ideas in order to make databases easier to use for the users [6]. Examples of visual interfaces include faceted search, template-based search, browsing query history and results. There has been recent work on developing query specification techniques involving just gestures and voice. Text based interface techniques include keyword search and natural language search. Miscellaneous techniques include spreadsheet-based interfaces, query-by-example technique, techniques to handle the empty-result problem with query relaxation or reformulation etc, personalization and diversification techniques.

II. RELATED WORK

Graph data integration We highlight some important ideas that have been proposed in order to integrate graph structured data. Reference [7] makes use of active learning and crowdsourcing techniques in order to integrate two graphs found in different sources. Their algorithm is based on involving a user that repeatedly selects nodes from graphs one after another requesting the user to mark them as matched or unmatched. Reference [4] introduces a unsupervised technique to perform graph alignment. The ideas the paper introduces involve inferring similarity between nodes of various graphs based on structural properties and node edge attributes, another one of which is leveraging locality sensitive hashing to minimize the number of pairwise node comparisons.

Graph visualization We highlight some important works that have been proposed for the problem of graph visualization relevant to the problems that this paper is trying to solve. Reference [1] proposes a visualization system for graph structured RDF data: a spatial-oriented approach for graph visualization; and a disk-based implementation. Reference [2] proposes a new focus + context display in order to enable user friendly large graph exploration. The idea of the system is to first map the node-link diagram on a riemann sphere. This is then projected out to a two-dimensional plane. The work in this paper appeared in a preliminary form at a workshop in [22]. This journal paper is a more detailed version.

Improving usability of data integration systems This section provides an overview of important papers on the problem of improving usability of web data integration systems. Reference [19] proposes a system for visualization for Data Quality Assessment. The purpose of such a system is to act as an aid in the identification and assessment of anomalies in the dataset. It makes use of various techniques such as type inference and data mining to identify data quality issues in tabular data. It then suggests coordinated, multiview visualizations to help an analyst assess anomalies and contextualize them within the larger data set.

Reference [20] proposes a novel interactive system for entity resolution in relational data. The system combines relational entity resolution algorithms with new network visualization techniques that creates a visualization that enables users to make use of contextual information of entities in relational databases to make decisions in matching task. They call such a system D-dupe novel user interface. Reference [21] proposes the idea of creating summaries of entity that is context-aware for which it proposes techniques to mine latent topics from query log in order to model user interest.

Reference [14] introduces techniques to utilize information gleaned from query logs dataset with the aim of finding matches in the schema attributes. Reference [23] introduces techniques to perform ontology mapping which is able to automatically find relevant mapping related to the query that has been reformulated.

Reference [15] proposes a XML schema matching framework based on previous mapping result set. Their system mainly consists of two phases as part of which in the first step XML schemas are processed as schema trees in order to find schema features after which they build a PMRS data structure in order to store the extracted auxiliary information and perform the schema matching tasking based on such a PMRS data structure. The schema matching in the paper is performed utilizing the positions of schema attributes that are present in earlier answers to user queries. It is claimed in the paper that the positions of schema attributes in answers to previous user queries reflect how important they are to the user. Their idea is based on the strategy to collect the statistics about attribute positions from query logs to find correspondences between attributes.

Reference [16] has been proposed to improve the usability of the schema matching process. In particular, it makes use of user preferences. User preferences are used to identify the specific parts of the schemas on which matching process is applied. The benefit of the approach is obvious in that only the relevant parts of the schema are matched thus improving usability. Reference [12] proposes new Query-Driven Approach (QDA) for record matching. The benefit of the approach is that considerably lesser amount of cleaning steps is done which are only relevant to the selected query so that the selected query is answered correctly.

Reference [13] proposes a two-phase schema matching approach by dividing the schema matching task into strong and weak phases. The benefit of this sort of a system is that it is very tunable, although it is not based on any machine learning technique. On the other hand, [17] proposes an active Learning matching strategy which is used to combining matchers based on answers to queries in the active learning process.

III. GRAPH DATA QUALITY ASSESSMENT WITH SPATIAL-ORIENTED VISUALIZATION AND FEEDBACK-DRIVEN PROVENANCE

In this section we describe the solution for the problem we are trying to solve of a building a graph data quality assessment system with spatial-oriented visualization and feedback-driven provenance to improve scalability, reduce execution time and improve relevance usability.

The main idea behind our approach is that of a provenance-driven adaptive graph sampling strategy leveraging user feedback. We take feedback from the user on various portions of the schema. The user marks some components of the schema as more critical and some components as less important and less relevant to the user. The mechanism to do this is explained in subsequent paragraphs. Subsequently, we propose an approach that utilizes this user feedback on schema elements while selecting a representative subgraph of the graph sample under consideration. This representative subgraph is stored in the form of a trie-based index, which represents our proposed paradigm of feedback-driven provenance index. After that, we propose another feature of the system in the form of allocating color intensity of various spatial pixel positions of nodes and vertices of the graphs that would visual estimation of quality of graph data. The process of assigning importance scores to schema components consists of a step where user iteratively marks schema components on a scale from hi-1-2-3 or med-1-2-3 or low-1-2-3, with hi being high importance, med being medium importance and low being least importance on a screen with 1,2,3 being further granular importance within it. These choices are then translated to importance scores between 0.1 to 1.0. Although this process of assigning importance scores seems quite simple, it still produced good results, studying more complex techniques is part of future work.

We leverage the graph sampling techniques [5] in order to also select a representative subgraph of the graph sample along with the actual sample itself. The selected representative subgraph of the sample is stored in the form a provenance trie-based index which not only consists of the representative subgraphs of various samples but also a count of the number of inaccuracies and inconsistencies measured in the sample. The intuition behind the idea to selection representative subgraph is that the nodes that are contained in the representative subgraph be associated with nodes of high importance according to the user feedback on the schema as well as it should be well connected. We build upon the traversal-based technique of graph sampling to select the representative subgraph of the sample to select such a representative subgraph that has high connectivity and high importance score as explained subsequently in the algorithm and the equations. During the process of creating the representative subgraph we also prefer to select the nodes which have not already been selected in the provenance index which requires accessing the provenance index during the representative subgraph construction. The process of computing the estimated inaccuracies and inconsistencies in the sample is based on a number of techniques already proposed in existing work and proposing a new approach for it is beyond the scope of this paper and part of future work.

As shown in

$$S = D * \sum_{1}^{n} S_i$$

, we compute the goodness of the representative nodes selected as the inverse of diameter of the subgraph connecting representative nodes (denoting how well connected the subgraph is) multiplied by the sum of importance scores of relevant schema nodes associated with the representative nodes through the type relationship. If we find that the goodness score of representative nodes does not meet threshold, we repeat the process till we find the adequate set of representative nodes.

For the visualization part, shown in figure 2, we propose a spatial-oriented paradigm for the purposes of our problem of building a tool to visually examine the quality of graph data. There have been few approaches presented in literature which are used to visually examine the quality of data however they suffer from the problem of scalability. To overcome the problem of scalability, we propose a system building upon current graph visualization technique of assigning spatial positions to its various vertices. We incrementally add on to this approach by proposing a system to allocate varying degree of intensity of color to each spatial position based on the number of estimated inaccuracies estimated in its neighborhood. We do it by accessing the provenance index built earlier by estimating the quality of neighborhood of the node under consideration from the information associated with the node in the provenance using the formula presented in the subsequent paragraphs.

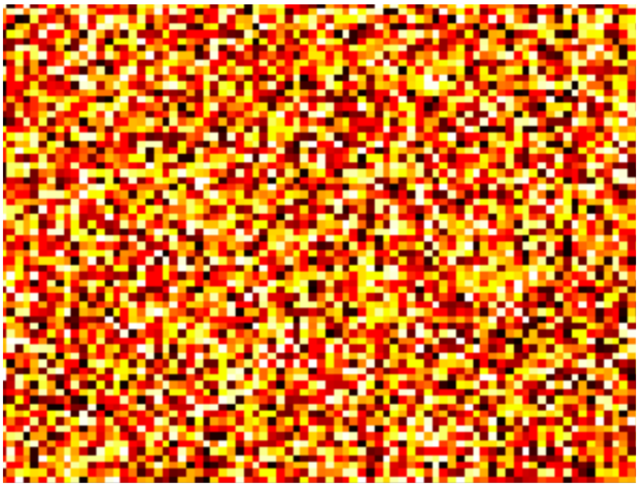

FIGURE 2. Graph data quality assessment with spatial-oriented visualization and feedback-driven provenance.

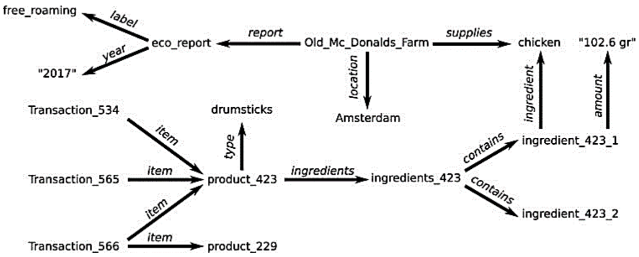

FIGURE 3. Example for graph data quality assessment with spatial-oriented visualization and feedback-driven provenance.

As shown in

$$\sum_{1}^{n} P_i \big/ (n * P_m),$$

the color intensity value is computed as the ratio of the mean of the number of errors found in the vicinity of the node divided by the average number of errors in the entire graph dataset. The scalability is achieved due to the maintenance of spatial indexes and spatial storage mechanisms in the backend for the graph dataset under consideration which were developed in existing for the case of graph visualization, which we have proposed to use for our problem of graph comparison. This results in improvements over the current visual data quality assessment techniques.

As an example shown in figure 3, consider the database of information on hotel in the figure below. The figure shows only the partial information on hotel database, which can potentially be huge. For the purposes of explaining, the figure describes information on the hotel mcdonalds, the foods items prepared in it and its ingredients etc. In the figure, they are classified under the terms ecoreport and oldmcdonaldfarm, which for the purposes of our example consider to be errors that should be corrected to ecological report and mcdonald-farm. If we were to execute the algorithm mentioned below for this task, we would first mark the schema for this database with importance scores for schema elements. Then, we proceed by taking graph samples, and computing the quality scores for them. For instance, for the graph sample collected around the term free-sample, we compute the quality score and select a bunch of representative graph elements utilizing the user importance score on schema elements. These representative nodes and edges are put in a trie-based index. When we take future samples, of say the graph samples around the term amsterdam, we make use of the provenance index to select the modified graph samples which can better reflect the quality of dataset by selecting a more diverse set of graph samples. The quality score is computed by counting potential errors in graph sample by either crowdsourcing or automated approaches through which it can easily be detected that there are errors in terms ecoreport and oldmcdonaldfarm. Correcting these terms to actual terms of ecological report and mcdonalds is not part of this task, as the purpose of this task is just to count potential errors in various samples and display it appropriately. Finally, the visualization is done by improving upon the existing approach in literature by assigning color intensity to spatial positions of graph elements through statistical averages techniques of mean using the provenance representative index of the quality scores, so that the colors of areas around the nodes of science fiction and fantasy appear more intense in the visualization based on quality scores in the provenance representative index.

IV. GRAPH DATA ALIGNMENT LEVERAGING SPATIAL-ORIENTED VISUALIZATION AND FEEDBACK-DRIVEN PROVENANCE

In this section we propose ideas of feedback-based provenance and spatial-oriented visualization techniques in developing a solution to the problem of graph alignment in order to improve scalability, reduce execution time and improve relevance usability. The approach also incorporates user feedback in the form of query logs.

Our approach makes use of a provenance index that aids in solving the problem of more relevant graph alignment. The idea of using a provenance index to improve the quality of entity resolution has shown to be effective for relational databases in current work. However, in our work we realize that the amount of provenance information collected can potentially be huge, so therefore we propose a new paradigm of feedback-driven provenance in order to streamline the collection of provenance information with the goal to reduce execution and improve usability. For the purposes of our problem, we create the provenance index in the form of trie-based index structure. It is basically an index of various k-hop neighborhoods of the previous vertices that were matched during graph alignment. In remainder of the graph alignment process, the index structure is accessed to check whether the neighborhoods of the 2 nodes under consideration now is similar to the neighborhoods of some pair of nodes that were found to be matched earlier.

Algorithm 1 Algorithm for Graph Data Quality Assessment With Spatial-Oriented Visualization and Feedback-Driven Provenance

```
procedure VISUALGRAPHQUALITY(G). Data graph G, Schema graph S
    s1,s2...sn <- [0.1..0.9] // set of schema components importances scores computed from user feedback using graphical screen
    T <- NULL // initialize trie index for provenance
    While(G is not empty)
        select sample g
        score = COMPUTE(Q(g)) // compute the quality score of chosen sample based on automated techniques or crowdsourcing etc
        if (Gi NOT IN T) then r <- Gi // select a node randomly for G if it is not already in provenance index
        While(G is not NULL and goodNess(r)<threshold) // while data graph is not empty to extract representative subgraph consisting of nodes and edges from G based on user importance scores on schema till the goodness of representative graph exceeeds min threshold score
            if(s(Gj) > threshold1 and func(Gj in T) < threshold2) then r <- r U Gj // traverse G starting from nodes in r adding edges and vertices if they are of high importance and not usually found in provenance index
        endloop
        UPDATE(T(r,score)) // insert the representative subgraph and quality score to update the trie-based provenance index
        delete sample g from G
    endloop
    While(T is not fully traversed)
        N <- k-hopneighborhood(Ti) // expand on k-hop neighborhood of Ti in G
        While(N is not full traversed)
            EstimatedQualityScore(Ni) <- EQS(Ni) + SCORE(Ti); // for each neighbor of Ni update its estimated quality score using provenance index
        endloop
        While(G is not empty)
            PixelIntensity(Gi) = EQS(Gi)/Max(EQS(G)) // compute pixel intensity between 0.. 1 for each graph node
        endloop
        Display(G) // display the graph
    endloop
    return
end procedure
```

However, since the main idea of our approach is to streamline the collection of provenance information, we seek to uti-

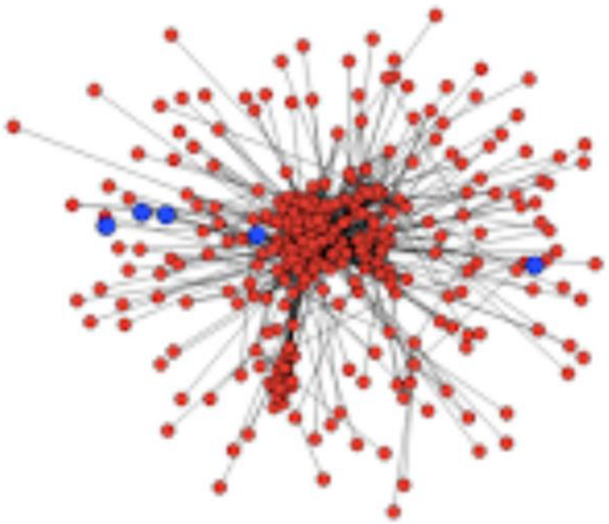

FIGURE 4. **Graph data alignment with spatial-oriented visualization and feedback-driven provenance.**

lize information in a query log which is provided by the user. From the query log, we maintain the neighborhoods index of the nodes and edges occurring in the query log and utilize this index during the construction of the provenance index using various graph matching techniques such as maximum common subgraph technique. As shown in results, this results in considerable reduction in provenance size and execution time.

We compute how good the k-hop neighborhood for graph alignment decision matches the candidate neighborhood from query log using graph similarity measure such as maximum common subgraph, we attempt to find the neighborhood that maximizes the similarity measure.

In the visualization part, shown in figure 4, we seek to apply the technique of spatial-oriented display for graphs for the purposes of our problem of building a system to visually compare 2 graphs. Our idea behind developing this system is that the 2 graphs to be compared are displayed close together which makes it easier for the user to compare the 2 graphs. In order to enable it, we develop a system to allocate spatial positions to the nodes of the second graph using various graph similarity measures. In the system that exists in literature, the system just partitions the graphs and displays them by minimizing the edges between partitions. If we were to simply apply this technique for our problem, it would just generate displays for 2 graphs in which potentially similar nodes can be far from each other impairing usability.

After we find n nearest nodes to the current node in the second graph, we compute its new spatial position as the mean of n nearest nodes in the first graph as shown

$$P(x,y)=\sum_{1}^{n}\left(D_i(x.y)\right)/n$$

Algorithm 2 Algorithm for Graph Data Alignment With Spatial-Oriented Visualization and Feedback-Driven Provenance

```
procedure VISUALGRAPHALIGNMENT(G). Query log Q, Data graphs S1 S2
    QT <- NULL // initialize index for query log neighborhoods
    While(pair of attributed q1,q2 in Q)
        UPDATE(QT,k-hopneighborhood(q1,q2)) // update index with neighborhoods for each pair of attributes in query log
    endloop
    T <- NULL // initialize trie-based provenance index
    While(count(G1,G2) in S1 S2 < threshold) // while number of pairs in graph to be aligned is lesser than threshold
    if(G1,G2 are aligned) then F = EXTRACT(T, k-hopneighborhoodMCS(G1,G2)) // if the nodes are aligned extract the maximum common subgraph of their k-hop neighborhoods
    M = FIND(MCS(F,QT(G1,G2))) // extract the maximum common subgraph of the neighborhood graph of aligned nodes with its counterpart graph in query log neighborhoods index
    UPDATE(T(M)) // insert the representative subgraph to update the trie-based provenance index
    endloop
    While(pair of nodes g1 g2 in S1 and S2 is not empty)
        if (HashMatch(g1,g2) Neighborhoodmatch( g1,g2,T(gx,gy)) > threshold) then Match else NoMatch // to align g1 and g2, in addition to existing approach, match the neighborhoods with the neighborhoods in the provenance index
    While(G is not full traversed)
    ComputeNewpixelposition(g1), ComputeNewpixelposition(g2) // while the graph is not fully traversed compute the new pixel position of the nodes using formula presented earlier
    Updatepixelpositions(remaining nodes in G) // update pixel positions of remaining nodes in G using formula presented earlier
    endloop
    Display    Display(G) // display the graph
    endloop
   return
end procedure
```

Then, we compute the distances all the other nodes in the second graph have to be moved as a consequence of the one of its node being moved, which is shown in

$$C_{(\Delta d)} = P_{(x.y)} - O_{(x.y)}/\left(C_{(x.y)} - O_{(x.y)}\right)$$

that represents the ratio of distance the previous node moved divided by the distance between the other node and the previous node.

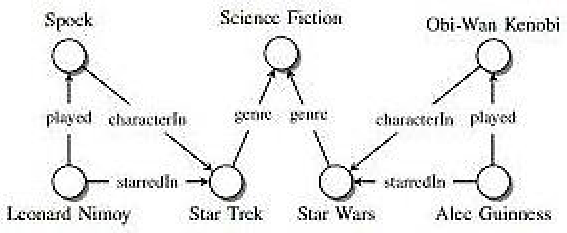

FIGURE 5. Example for graph data alignment with spatial-oriented visualization and feedback-driven provenance.

This process is repeated for all the nodes in the second graph, first we move one of its node based on distance from n nearest nodes in previous graph then after that we move all the other nodes in the second graph in the same direction as the node just moved in the ratio of their closeness. As a result of this, we end up with the 2 graphs to be compared displayed close to each other. The benefit of spatial oriented approach for visual graph comparison is improved scalability, as we are now using spatial indexes and spatial storage mechanisms in the backend for the graph dataset under consideration which were developed in existing for the case of graph visualization, which we have proposed to use for our problem of graph comparison. This results in improvements over the current visual graph comparison techniques.

As an example shown in figure 5, consider the database of information on movies in the figure below. The figure shows only the partial information on movies database, which can potentially be huge. For the purposes of explaining, the figure describes information on the movies inception, star wars, terminator etc. In one of the figure, they are classified as science fiction and in the other figure they are classified under the term fantasy. Although in the figures the common movie names are not explicitly mentioned, however it is not too far-fetched to imagine that two figures could have movie names in common. Since our task is of graph alignment, in out example figures we would like to generate the alignment between the nodes fantasy and science-fiction. If we were to execute the algorithm mentioned below, first we shall extract the pairs of nodes and neighborhoods from the query logs. For instance, if the queries are also about movies, we would extract pairs of nodes such as terminator and arnold and extract the neighborhood graphs around them in query logs graph. Once it is done, we start with the process of alignment of actual data graphs. Since, our idea to maintain provenance index of neighborhood graphs, during the actual execution process, for each pair of aligned nodes in graphs, such as fantasy and science fiction in the example below, we extract the neighborhoods of the nodes. Since the potential neighborhood can be quite large, we restrict its size and extract only the relevant neighborhood based on the neighborhoods extracted from the query log graphs, such as between fantasy neighborhood in query logs and neighborhood of science fiction in the query log. These neighborhoods of aligned nodes in graphs are stored in a trie-based index, which serves to be a provenance index of the graph alignment process. In the remainder of the graph integration phase, we match the neighborhood of terms of fantasy and science fiction with the neighborhood of similar matched nodes earlier in the provenance index. This serves as additional information that augments the task of graph alignment improving its effectiveness. Finally, the visualization is done by improving upon the existing approach in literature by utilizing graph similarity measures in order to allocate spatial pixel locations to various graph components, so that the nodes of science fiction and fantasy appear closer in the visualization.

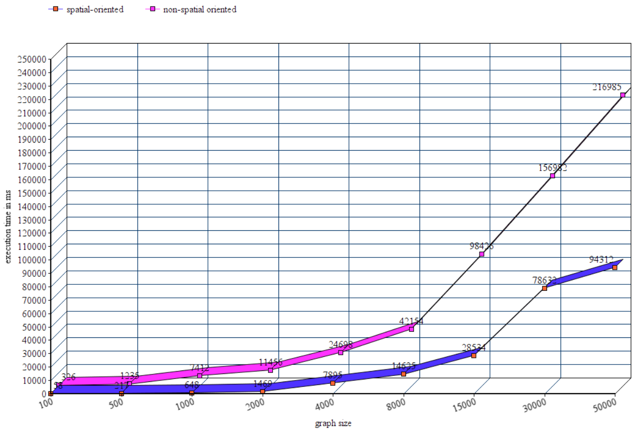

FIGURE 6. Execution times to generate spatial-oriented versus non-spatial oriented display for graph data quality assessment.

V. EXPERIMENTS

A computer running macosx with java installed was used to perform experiments. The datasets which are realworld based are described next. dataset1 - the University of Florida Sparse Matrix Collection (cise.ufl.edu/research) and the Parasol project and KONECT (http://konect.unikoblenz.de/networks) and dataset2(https//old.datahub.io/) - that contains information of various datasets relevant latin america governments and dataset3 - Movielens10m and freebase, which is used in addition to usage data to enrich Movie-Lens10m dataset with more metadata. The results presented are an average of results on the three different datasets. The results clearly show an improvement in performance of proposed techniques over the techniques in existing literature. For comparisons, as the highly cited existing system for visual comparison of graph we chose [25], and highly cited existing system for visual assessment of data quality we chose [24], in the experiments.

As displayed in the figure 6 the describes the execution times in the problem of spatial-oriented display for graph data quality assessment versus non-spatial oriented existing approach for various datasets of graphs with sizes ranging from 100 to 50000 nodes. The results demonstrate a considerable lesser execution time for the proposed approach compared with the existing approach in literature, in fact more than 50 percent lesser execution time.

As displayed in the figure 7 the describes the execution times in the problem of spatial-oriented display for graph data alignment versus non-spatial oriented existing approach for various datasets of graphs with sizes ranging from 100 to 50000 nodes. The results demonstrate a considerable lesser

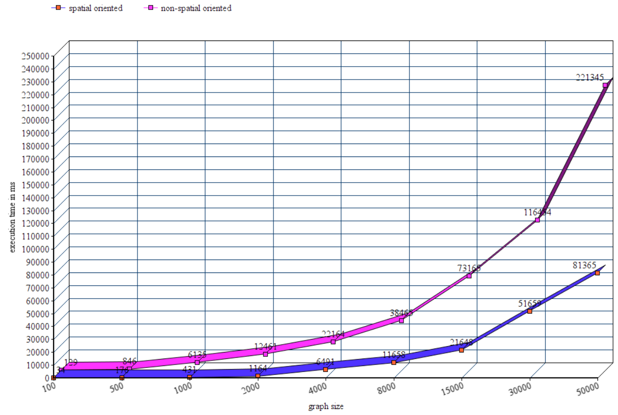

FIGURE 7. Time for execution to generate spatial-oriented display for problem of graph alignment versus non-spatial oriented existing approach.

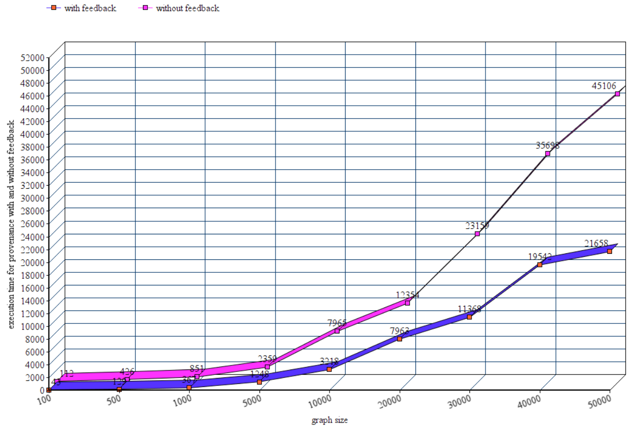

FIGURE 8. Time for execution for feedback-driven provenance for graph data quality assessment versus non-feedback driven provenance.

execution time for the proposed approach compared with the existing approach in literature, in fact more than 60 percent lesser execution time.

As displayed in the figure 8 the describes the execution times in the problem of feedback-driven provenance for graph data quality assessment versus non-feedback driven provenance existing approach for various datasets of graphs with sizes ranging from 100 to 50000 nodes. The results demonstrate that the proposed approach does decrease execution time considerably compared with existing approach.

As displayed in the figure 9 the describes the execution times in the problem of feedback-driven provenance for graph data alignment versus non-feedback driven provenance existing approach for various datasets of graphs with sizes ranging from 100 to 50000 nodes. The results demonstrate that the proposed approach does decrease execution time considerably compared with existing approach.

As displayed in the figure 10 which shows the combined provenance sizes in terms of actual nodes in it for the task of feedback-driven provenance for graph data quality assessment and alignment versus non-feedback driven provenance existing approach for various datasets of graphs with sizes ranging from 100 to 50000 nodes. The results demonstrate that the proposed approach does decreases provenance sized considerably compared with non-feedback approach which resulted in decreased execution time.

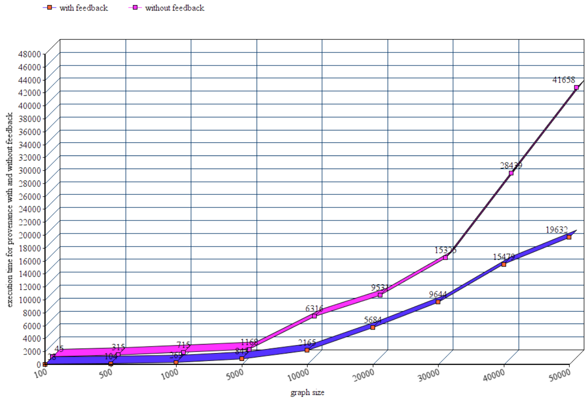

FIGURE 9. Execution times for feedback-driven provenance for graph data alignment versus non-feedback driven provenance existing approach.

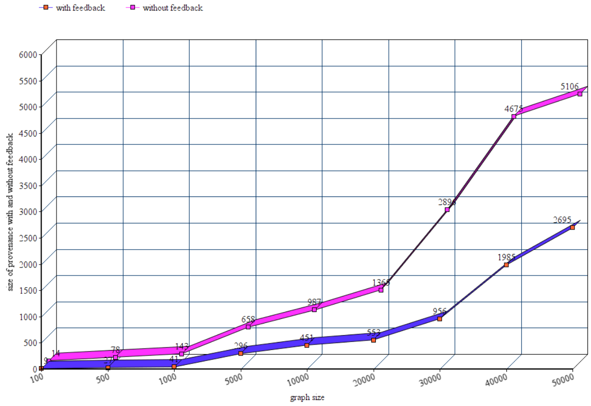

FIGURE 10. Provenance sizes in terms of actual nodes in it for the task of feedback-driven provenance for graph data quality assessment and alignment versus non-feedback driven provenance existing approach.

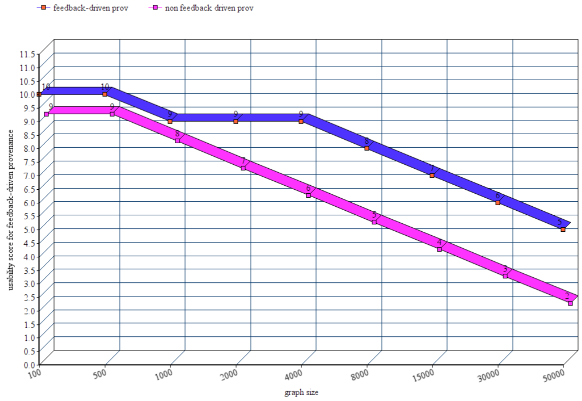

FIGURE 11. Usability score from 25 users for spatial-oriented display for graph data integration versus non-spatial oriented existing approach.

As displayed in the figure 11 which shows score of satisfaction of user for the spatial-oriented display for graph data integration versus non-spatial oriented existing approach for various datasets of graphs with sizes ranging from 100 to 50000 nodes for 25 users from 1- 10. As indicated by the results, the spatial-oriented system shows a considerable improvement in user satisfaction.

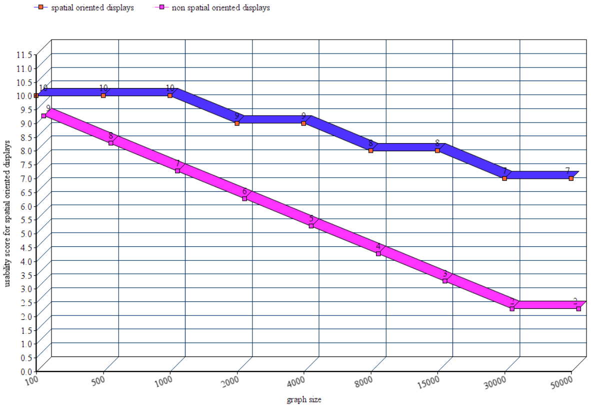

FIGURE 12. Usability score from 25 users for feedback-driven provenance for graph data versus non-feedback driven provenance existing approach.

As displayed in the figure 12 which shows the score of satisfaction of user for the feedback-driven provenance for graph data for various datasets of graphs with sizes ranging from approach for various graph datasets, sizes of which vary 100 to 50000 nodes for 25 users from 1 -10. As indicated by the results, the spatial-oriented system shows a considerable improvement in user satisfaction

VI. CONCLUSION

In conclusion, this paper proposed a spatial oriented visualization and feedback-driven provenance enabled graph data integration system. In particular, Graph data quality assessment leveraging spatial oriented visualization and feedback driven provenance and Graph alignment leveraging spatial oriented visualization and feedback driven provenance where we proposed a solution to the problem seeking to improve the scalability and reducing execution times and increasing relevance and usability scores.

REFERENCES

[1] N. Bikakis, J. Liagouris, M. Kromida, G. Papastefanatos, and T. Sellis, "Towards scalable visual exploration of very large RDF graphs," in *Satellite Events* (Lecture Notes in Computer Science), vol. 9341, F. Gandon, C. Guéret, S. Villata, J. G. Breslin, C. Faron-Zucker, and A. Zimmermann, Eds. Cham, Switzerland: Springer, 2015, pp. 9–13. [Online]. Available: http://dblp.unitrier.de/db/conf/esws/eswc2015s.htmlBikakisLKPS15

[2] F. Du, N. Cao, Y.-R. Lin, P. Xu, and H. Tong, "iSphere: Focus+context sphere visualization for interactive large graph exploration," in *Proc. CHI*, G. Mark, S. R. Fussell, C. Lampe, M. C. Schraefel, J. P. Hourcade, C. Appert, and D. Wigdor, Eds. New York, NY, USA: ACM, 2017, pp. 2916–2927. [Online]. Available: http://dblp.unitrier.de/db/conf/chi/chi2017.htmlDuCLXT17

[3] A. Y. Halevy, G. Weikum, H. Schöning, and E. Rahm, "Data integration: A status report," in *BTW* (LNI), vol. 26. GI, 2003, pp. 24–29.

[4] M. Heimann, W. Lee, S. Pan, K.-Y. Chen, and D. Koutra, "HashAlign: Hash-based alignment of multiple graphs," in *Proc. Pacific–Asia Conf. Knowl. Discovery Data Mining (PAKDD)*, Melbourne, VIC, Australia, Jun. 2018, pp. 726–739.

[5] P. Hu and W. Lau, "A survey and taxonomy of graph sampling," *arXiv.org*, vol. CoRR abs/1308.5865, pp. 1–34, Aug. 2013

[6] H. V. Jagadish, A. Chapman, A. Elkiss, M. Jayapandian, Y. Li, A. Nandi, and C. Yu, "Making database systems usable," in *Proc. SIGMOD*, C. Y. Chan, B. C. Ooi, and A. Zhou, Eds. New York, NY, USA: ACM, 2007, pp. 13–24. [Online]. Available: http://dblp.unitrier.de/db/conf/sigmod/sigmod2007.htmlJagadishCEJLNY07

[7] Y. Kim, W. Jung, and K. Shim, "Integration of graphs from different data sources using crowdsourcing," *Inf. Sci.*, vols. 385–386, pp. 438–456, Apr. 2017. doi: 10.1016/j.ins.2017.01.006.

[8] J. M. B. Josko and J. E. Ferreira, " Vis4DD: A visualization system that supports data quality visual assessment," in *Proc. 32th Brazilian Symp. Databases Demo Appl.*, 2017, pp. 46–51.

[9] F. Naumann, "Data profiling revisited," *ACM SIGMOD Rec.*, vol. 42, no. 4, pp. 40–49, 2013.

[10] P. Woodall, A. Borek, and A. K. Parlikad, "Data quality assessment: The hybrid approach," *Inf. Manage.*, vol. 50, no. 7, pp. 369–382, 2013.

[11] P. Christen, *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Berlin, Germany: Springer, 2012.

[12] H. Altwaijry, D. V. Kalashnikov, and S. Mehrotra, "Query-driven approach to entity resolution," *Proc. VLDB Endowment*, vol. 6, no. 14, pp. 1846–1857, 2013.

[13] N. Bozovic and V. Vassalos, "Two phase user driven schema matching," in *Proc. ADBIS*, in Lecture Notes in Computer Science, vol. 9282, T. Morzy, P. Valduriez, and L. Bellatreche, Eds. Cham, Switzerland: Springer, pp. 49–62, 2015.

[14] H. Elmeleegy, M. Ouzzani, and A. Elmagarmid, "Usage-based schema matching," in *Proc. ICDE*, G. Alonso, J. A. Blakeley, and A. P. Chen, Eds. Washington, DC, USA: IEEE Computer Society, Apr. 2008, pp. 20–29.

[15] H. Fan, J. Liu, W. Luo, and K. Deng, "An efficient schema matching approach using previous mapping result set," in *Proc. DASFAA Workshops*, in Lecture Notes in Computer Science, vol. 9645, H. Gao, J. Kim, and Y. Sakurai, Eds. Cham, Switzerland: Springer, 2016, pp. 285–293.

[16] W. Guél'dria, Z. Bellahsene, and M. Roche, "A flexible approach based on the user preferences for schema matching," in *Proc. 1st IEEE Int. Conf. Res. Challenges Inf. Sci.*, Morocco, Africa, 2007, pp. 21–26.

[17] D. Rodrigues, A. da Silva, R. Rodrigues, and E. dos Santos, "Using active learning techniques for improving database schema matching methods," in *Proc. IJCNN*, Jul. 2015, pp. 1–8.

[18] R. Angles and C. Gutírrez, "Survey of graph database models," *ACM Comput. Surv.*, vol. 40, no. 1, pp. 1–39, Feb. 2008.

[19] S. Kandel, R. Parikh, A. Paepcke, J. M. Hellerstein, and J. Heer, "Profiler: Integrated statistical analysis and visualization for data quality assessment," presented at the Meeting AVI, 2012.

[20] H. Kang, L. Getoor, B. Shneiderman, M. Bilgic, and L. Licamele, "Interactive entity resolution in relational data: A visual analytic tool and its evaluation," *IEEE Trans. Vis. Comput. Graphics*, vol. 14, vol. 5, pp. 999–1014, Sep./Oct. 2008.

[21] J. Yan, Y. Wang, M. Gao, and A. Zhou, "Context-aware entity summarization," in *Proc. WAIM*, in Lecture Notes in Computer Science, vol. 9658, B. Cui, N. Zhang, J. Xu, X. Lian, and D. Liu, Eds. Cham, Switzerland: Springer, 2016, pp. 517–529.

[22] K. Reddy, "Adaptive, efficient and effective graph data integration and search framework," *Procedia Comput. Sci.*, vol. 151, pp. 1255–1260, Jan. 2019.

[23] A. Tian, J. Sequeda, and D. P. Miranker, "On ambiguity and query-specific ontology mapping," in *Proc. 7th Int. Workshop Ontol. Matching*, Vol.946, Boston, MA, USA, Nov. 2012.

[24] J. Sharko, G. G. Grinstein, K. A. Marx, J. Zhou, C.-H. Cheng, S. Odelberg, and H.-G. Simon, "Heat map visualizations allow comparison of multiple clustering results and evaluation of dataset quality: Application to microarray data," in *Proc. 11th Int. Conf. Inf. Vis.*, Jul. 2007, pp. 521–526.

[25] M. Hascoënt and P. Dragicevic, "Interactive graph matching and visual comparison of graphs and clustered graphs," in *Proc. AVI*, 2012, pp. 522–529.

KULDEEP REDDY received the master's degree in computer science from IIT, Madras, India. He is currently pursuing the Ph.D. degree in computer science with Southeast University, Nanjing, China.