Grafovske strukture za vizuelizaciju podataka

Stefana Janićijević¹, Vojkan Nikolić²

¹Information Technology School, Belgrade, Serbia; stefana.janicijevic@its.edu.rs;

²University of Criminal Investigation and Police Studies, Belgrade, Serbia; vojkan.nikolic@kpu.edu.rs

Apstrakt: Mreže su svuda oko nas. Strukture grafova su uspostavljene u jezgru svakog mrežnog sistema, pa se pretpostavlja da se grafovi shvataju kao objekti za vizuelizaciju podataka. Ti objekti rastu od apstraktnih matematičkih paradigmi do informacionih uvida i kanala povezivanja. Izračunate su bitne metrike u grafovima kao što su centralnost stepena (degree centrality), centralnost bliskosti (closeness centrality), centralnost između centralnosti (betweenness centrality) i centralnost ranga stranice i u svima njima opisuju komunikaciju unutar sistema grafa. Osnovni cilj ovog istraživanja je sagledavanje metoda vizuelizacije preko postojećih velikih podataka i predstavljanje novih pristupa i rešenja za trenutno stanje vizuelizacije velikih podataka. Ovaj rad daje klasifikaciju postojećih tipova podataka, analitičkih metoda, tehnika i alata za vizuelizaciju, sa posebnim akcentom na istraživanje evolucije metodologije vizuelizacije poslednjih godina. Na osnovu dobijenih rezultata uočavaju se nedostaci postojećih metoda vizuelizacije.

Ključne reči: graf, veliki podaci, vizuelizacija, mrežna analiza

Graph Structures for Data Visualizations

Abstract: Networks are all around us. Graph structures are established in the core of every network system therefore it is assumed to be understood as graphs as data visualization objects. Those objects grow from abstract mathematical paradigms up to information insights and connection channels. Essential metrics in graphs were calculated such as degree centrality, closeness centrality, betweenness centrality and page rank centrality and in all of them describe communication inside the graph system. The main goal of this research is to look at the methods of visualization over the existing Big data and to present new approaches and solutions for the current state of Big data visualization. This paper provides a classification of existing data types, analytical methods, techniques and visualization tools, with special emphasis on researching the evolution of visualization methodology in recent years. Based on the obtained results, the shortcomings of the existing visualization methods can be noticed.

Keywords: graph, Big data, visualization, network analysis

1. Introduction

Graph is visually represented and it is sensitive to creative interventions in the art perspective. Graph form is unpredictable enough from the aspect of visual forms, but from the point of big data it is very simple to represent it (Benzi et al., 2015; Diestel, 2000). We considered that the basic characteristic of data is information and that the basic property of this visualization is communication. Schulz et al. stated that there are two main representations: network and hierarchy. Graph drawing is focusing on optimized layouts for node-link-representations of networks, but "information visualization" prefers to work on hierarchies focusing on very large structures, different views and interactivity (Schulz and Schumann, 2006; Vitter, 2001).

Data set that was used for calculation is telecommunication data. We research relations and communications between vertices in graphs and calculate main centralities. Vertices are specific objects such as customers and edges are weighted metrics between them.

Graph theory studies mathematical structures called graphs. The graph is represented by ordered pair $G = (V; E)$, where V is a finite, non-empty set of vertices (tops, nodes), and E is a set of two-element subsets of set V , i.e. a set of edges (arcs, branches) (Pardalos and Du, 1999). Examples of graphs are

zero graph, trivial graph, simple graph, undirected graph, directed graph, complete graph, connected graph, K-partite graph, disconnected graph, weighted graph, regular graph, cyclic graph, acyclic graph, star graph, multigraph, planar graph, etc.

Graph representations on a computer are different, and each of them depends on the nature of the problem being solved and the computer resources at its disposal (Pardalos and Du, 1999).

1. The adjacency matrix (Figure 1) represents the elements that provide information about whether there are edges in between vertices corresponding to the indices of these elements. Formally, the neighborhood matrix of dimension $n \times n$ of G is written as:

$$A_G = (a_{ij})_{(i,j) \in V \times V} \quad (1)$$

Figure 1: *Adjacency matrix*

$$\begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 \end{bmatrix}$$

2. A graph can be represented by a list of neighbors such that each of the vertices of the graph is an element to which a list is formed in which the neighbors of that vertex are placed in the graph.

The concept of Big Data was created in order to solve the problem of a large amount of data that is increasing exponentially in our time. Big data characteristics are commonly represented as "3Vs" and refer to volume, velocity and variety, as a significant feature of big data different from traditional data. Volume refers to the size of the dataset. Velocity, where high velocity is meant, refers to the fact that data should be collected and analyzed quickly as well as in a timely manner. Variety refers to different types of big data that include structured, semi-structured, and unstructured data (Ahmed and Ismail, 2020; Hukkeri et al., 2019).

In addition, the National Institute of Standards and Technology (NIST) added the variability feature and introduced it as "4Vs", which represents changes on the three other features that impact data processing. Then NIST also introduced NIST big data reference architecture (NBDRA). NBDRA consists of five main roles: system orchestrator, data provider, big data application provider, big data framework provider and data consumer. Application provider is responsible for data collection, preparation, analysis, visualization, and access. The framework provider provides the infrastructure and data storage and processing platform (Gao et al., 2020; Levin et al., 2015).

Research related to the Big Data concept, which implies increasing availability of huge amounts of data, in the field of scientific approaches, techniques and visualization tools, is very important. The aim of this research was to look at the visualization methods over the existing Big data and to present new approaches and solutions for the current state of Big data visualization. Existing data types, analytical methods, as well as visualization techniques and tools are classified, with special emphasis on the evolution of visualization methodology in recent years.

In addition to the development of technology, the participation of people with the qualities of logical thinking and reasoning is an important factor in processes involving Big Data. Significant human limitations in the process itself are also immediately apparent. Therefore, the concepts of Augmented Reality and Virtual Reality are considered here so that they can be applied in the process of visualizing Big data. What is most important is the placement of the most important data in the central area of the human visual field. In addition, for the visualization process, it is important to obtain significant information in the shortest possible time and without significant data loss due to human perceptual problems.

2. Graph in data science

A graph is an exceptional object located in different fields of creation - it basically comes from discrete mathematics and combinatorics, and goes to mathematical programming and operational research. It represents the data that form the keystone for machine learning and artificial intelligence.

In addition to representing the connections between vertices in graph, there is one important issue of graphs that is widely applied in science and industry. The issue address important question about what are the central vertices in graph and consequently, what are most important vertices in graph. Area where is applied researching of most important vertices is centrality measures graph area. Centrality measures have found their role in various applications such as social network analysis, web mining and biology networks.

The most widely centrality measures applications could be established in social network analysis (SNA) where it is explored social structures through relations between people. In SNA people are vertices and relation between people are edges therefore graph of social network is DAG type of graph. Such graph is sparse, oriented and it is obtained on the basis of big data concepts. There are many different centrality measures, but most important are Degree, Closeness, Betweenness, Page Rank, Eigenvector and all they are trying to research what is the most important vertex in graph or subgraph. Researching of the question of visualization according to graph and most important vertices, it is found that visual representation of centrality measures are most significant for big data insights. It is proposed in literature that if it is strong subpattern around vertex, than the vertex is more central in the graph. Measures such as Degree and Betweenness considers subpatterns like edges, paths and cliques (Riveros and Salas, 2020). Value of centrality measures depend on the fact if the graph is weighted and specifically, if the weight is vertex weighted or edge weighted established. All prominent metrics are generalized for weighted case as well as unweighted case.

Formalization of the centralities established Freeman (1978) according to several features that validate for any vertex in graph. These features are number of edges, possibility to reach all the other vertices quickly, and controlling the flow between other vertices.

Degree is the most used of the vertex centrality measure which is using local structure around vertices. In standard graph the degree is the number of edges a vertex has. In an oriented graph, a vertex may have a different number of outgoing and incoming edges, and therefore, degree is split into out-degree and in-degree, respectively.

Closeness is defined as the sum of distances to all other vertices. The intent behind this measure was to identify the vertices which could reach other vertices quickly. A main limitation of closeness is the lack of applicability to networks with disconnected components: two vertices that belong to different components do not have a finite distance between them. Thus, closeness is generally restricted to nodes within the largest component of a network.

Betweenness measure calculates transactions between other vertices. It is calculated under the shortest paths from vertex and it is scaled with the number of pairs of vertices by summation indices. Considering weighted graph edges are labeled in proportion to their capacity, influence, frequency, or similar characteristic, which adds another dimension of heterogeneity within the graph. Strong vertex in a weighted graph is given by the sum of the weights of its adjacent edges.

Page Rank centrality measure is using successfully according to web graph researching. It is created and implemented by Google and it is based on web nature and edge structure as an indicator of each page value. Google proposed a link from page A to page B as a vote, by page A, for page B. But, Google seeks at more than the absolute volume of votes, or links a page receives.

Eigenvector centrality is a measure of the influence of a vertex in a network. It is an extension of degree centrality, but in Eigenvector there are assigned relative scores to all vertices. It is based on the concept of connections to high scoring vertices to the score of the vertex. A high Eigenvector score means that a vertex is connected to many vertices who themselves have high scores.

There are many types of visualization such as bars, line graphs, charts, plots, etc. They could be classified into 2D and 3D structures (Klein, 2010; Olshannikova et al., 2016). The main advantages of

the applied visualizations are complexity in terms of big data, pattern recognition and information retrieval from the point of visualization. It is important to highlight visual structures in the data and present the most important information.

Graph visualizations reveal the connections between finite representation and large so-called infinite data. It is possible to consider big data as divergent structures that aspire to infinite series of information. The graph as a combinatorial object that approximates continuity since it is a visualization structure that provides all important information points. Big data information is modelling through machine learning algorithms, and the best optimization algorithms are graph based algorithms. Popular graph algorithms are: maximal clique, independent set, dominating set, k coloring, depth first search, breadth first search, etc. (Simonetto et al., 2020; Tsouliaset al., 2020).

The relationship between the finite and the infinite in terms of data is only seen through the visual structures that can best be represented by graphs, since other forms of visual graphs and diagrams mainly evoke aggregation data and reporting results. However, raw data cannot be clearly seen using bar charts, scatter plots, histogram charts, box plots, and other available visual forms, Figure 2 (a), (b). Comparison between those charts and graph representation imply that graph visualization is only possible for row Big data (Figure 2 (c)).

Figure 2: Bar chart (a), scatter plot (b) and graph visualization (c)

Source: Authors

3. Visual information

In this research we used a Telco data set, which is a population of 2.5 million users of telecommunications services that are connected by edges in relation to the use of voice, sms and gprs services (Figure 3 (a)). The visualization was created in Python 3.6, in a Jupyter notebook. In the case of big telecommunications data, it helps to see the connections as well as the type of graphs that are obtained when changing the user sample. By changing the sample in the population, the visualization changes (Figure 3 (b), (c)).

Figure 3: Graph visualization on total population (a), graph visualization on huge sample (b), graph visualization on small sample (c)

Source: Authors

We give an example of telecommunication based graphs Telco data set where users are connected to each other based on the metrics centrality PAGERANK, Degree, Closeness and Betweenness (Diestel, 2000).

PageRank centrality is an algorithm that counts the number and quality of links to a page to determine a rough estimate of how important the website is. It means that more important websites are likely to receive more links from other websites. In Telco data set it assumed that users are similar to web pages, so, the more important user is, he would receive more links from other users (Figure 4).

Figure 4: PageRank centrality

Source: Authors

Degree centrality of a vertex in a graph is the number of edges that are incident to the vertex. In the case of Telco data set which is directed graph, we usually define two separate measures of degree centrality, namely indegree and outdegree (Figure 5).

Figure 5: Degree centrality

Source: Authors

Betweenness centrality is a measure that present considering of an arbitrary vertex in the path of other vertices. In the case of Telco data set vertices with high Betweenness are influencers in graph control information retrieval (Figure 6).

Figure 6: Betweenness centrality

Source: Authors

Closeness centrality present how close a vertex is to all other nodes in the graph. In the case of Telco data set it is calculated as the average of the shortest path length from the vertex to every other vertex in the graph (Figure 7).

Figure 7: *Closeness centrality*

Source: Authors

Significant vertices with their connections present information that are hidden in a big data graph but these information are crucially important for patterns retrieval (Figures 8 (a), (b)).

Figure 8: Pattern retrieval from huge sample (a) and from small sample (b)

Source: Authors

Predictive decisions about vertex locations and about edge directions from a visual perspective are achieved according to a rough and unprocessed data set.

4. Summary and Conclusions

The data science algorithms could be established through different models of graphs that involve information within itself and, like the most intelligent algorithm, processes it. The graph algorithm learns from data, finds patterns and predicts decisions depending on the interaction between vertices. Using graph visual applications helps in advance in investigation of data patterns and data structures. In this paper we present specific view of graph visualization such as visualization of calculated graph metrics. This includes calculation of four centrality measures and presentation in visual form. It is considered that this is meaningful and useful visualization for row data metrics and proposal for exploratory view of pattern from metrics.

This research points recognition challenges for the Big data presentation and visualization. Future investigations will go further into connections calculations and materia understanding of unseen Big data. The next steps of our work will be to predict possible visualizations with machine learning algorithms according to prediction of calculated metrics.

References

1. Ahmed, H., Ismail, M. A. (2020). Towards a Novel Framework for Automatic Big Data Detection. *IEEE Access*, 8, 186304-186322. doi: 10.1109/ACCESS.2020.3030562.
2. Benzi, K., Ricaud, B., Vanderghelynst, P. (2015). Principal Patterns on Graphs: Discovering Coherent Structures in Datasets. *IEEE Transactions on Signal and Information Processing over Networks*.
3. Diestel R. (2000). *Graph Theory*. 2nd Edition, New York: Springer-Verlag.
4. Gao, Y., Chen, X., Du, X. (2020). A Big Data Provenance Model for Data Security Supervision Based on PROV-DM Model. *IEEE Access*, 8, 38742-38752. doi: 10.1109/ACCESS.2020.2975820.
5. Hukkeri, G. S., Goudar, R. H., Kotagi, P. R. (2019). Handling 3vs of Big Data Through Swarm Intelligence, 2019 4th International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT), Bangalore, India, 589-595, doi: 10.1109/RTEICT46194.2019.9016846.
6. Klein, D. (2010). Centrality measure in graphs. *Journal of Mathematical Chemistry*, 47, 1209-1223
7. Levin, O., Boyd, D., Chang, W. (2015). NIST big data interoperability Framework: Framework: Volume 6, reference architecture, Nat. Inst. Std. Tech., Gaithersburg, MD, USA, Tech. Rep. NIST.SP.1500-6.
8. Olshannikova, E., Ometov, A., Koucheryavy, Y., Olsson, T. (2016). Visualizing Big Data. *Big Data Technologies and Applications*, 101-131.
9. Pardalos, P., Du, D. (1999). Handbook of Combinatorial Optimization - The Maximum Clique Problem. Kluwer Academic Publishers.
10. Riveros, C., Salas, J. (2020). A family of centrality measures for graph data based on subgraphs. In 23rd International Conference on Database Theory (ICDT 2020), Schloss Dagstuhl-Leibniz-Zentrum für Informatik.
11. Schulz, H. J., Schumann, H. (2006). Visualizing Graphs - A Generalized View. Tenth International Conference on Information Visualisation (IV'06).
12. Simonetto, P., Archambault, D., Kobourov, S. (2020). Event-Based Dynamic Graph Visualisation. *IEEE Transactions on Visualization and Computer Graphics*, 26 (7), 2373-2386. doi: 10.1109/TVCG.2018.2886901.
13. Tsoulas, K., Palaiokrassas, G., Fragkos, G., Litke, A., Varvarigou, T. A. (2020). A Graph Model Based Blockchain Implementation for Increasing Performance and Security in Decentralized Ledger Systems. *IEEE Access*, 8, 130952-130965. doi: 10.1109/ACCESS.2020.3006383.
14. Vitter, J. S. (2001). External memory algorithms and data structures: dealing with massive data. *ACM Computing Surveys*, 33, 209-271.