Analysis and visualization in graph database management systems

A S Dubrovin, O V Ogorodnikova, E G Tsarkova, E A Andreeva and T N Kulikova

Voronezh institute of the Russian Federal Penitentiary Service, Russia, Voronezh, Irkutskaya 1a

E-mail: olga-ogorodnikova@yandex.ru

Abstract. The following questions of visual interpretation of the data in the graph models of data are considered in the work with graph database management system (DBMS) with DBMS Neo4j as an example; convenience of data visualization in the form of graphic images as compared with the electronic worksheets was confirmed; a high connectivity of data was revealed for DBMS NEO4j. To do all these algorithms from DBMS NEO4j library are employed. Clusterization is applied for the estimation of the quality of affinity for the predetermined data sets. Here method of silhouette assessment is used. Basing on the performed experimental studies a conclusion was made concerning the quality of revealing the similarity between the predetermined data sets with the use of graph DBMS NEO4j.

1. Introduction

Lately, the use of relational databases results in the difficulties while development of the data models as well as during the work with a great amount of data when the relational DBMS cannot manage these data. Replacement of the preconceived ideas that became understandable and habitual for a multiple of users and experts in the relational DBMS is required due to permanent increase of the data amounts in the areas of data bases for wide applications. As an appropriate alternative for the attainment of this goal the users of data bases more often choose graph DBMS. Graph DBMS use graph data models as the models of data. The use of graph data models and graph DBMS allows increase in their performance and create complicated and flexible requests when working with them. Within the frames of the study of the graph data models the problems of analysis and data visualization represent a serious interest; the problem statement and methodological approaches to their solutions are presented in [1]. Visualization implies transformation of invisible and weakly structured data with the use of computer DBMS program into the visual ones in order to get their further interpretation and analysis. Data visualization promotes and facilitates their interpretation. As a rule, the graphical form of the processed data allows to obtain just the result that is required by the specialists of the objects domains from the usage of the data bases. Further we consider the problems of analysis and visualization of data in the graph DBMS using their typical representative – DBMS Neo4j.
To solve these problems of analysis and data visualization there is a complete and necessary toolkit in DBMS Neo4j. Particularly, algorithms for data processing suitable for the graph models are combined in the library of graph algorithms in DBMS Neo4j. One of these graph algorithms within this library is an algorithm of cosine similarity. It allows determining similarity of the specified data sets at the qualitative level when using a great amount of these data sets. Algorithm of cosine similarity is one of typical means for the analysis and visualization of data in DBMS Neo4j.

The purpose of this work is to consider the following problems of visual data interpretation in the graph models with graph DBMS Neo4j as an example:
- the study of specific capabilities of the graph DBMS with graph DBMS Neo4j as an example;
- confirmation of convenience of the data visualization in the form of graphical images as compared with the spreadsheets;
- revealing of a high connectivity of the data in DBMS Neo4j.

2. Related works
In [2] there were proposed solutions for the development of new generation of the systems for fraud detection with the use of graph data models and graph DBMS. Fraud detection in such perspective systems is based on the fact that in terms of the information on transactions and users it is possible to deduce data on the interrelations between them. With the help of the indexing methods comparative analysis of efficiency for the most widespread graph DBMS was performed in [3]. The purpose of this work was to choose a candidate for the mechanism for keeping graph structure. It was noted that index accelerated retrieval requests and considerably reduced the number of node in a graph. The review of most popular tasks on the graph data bases from the section on the deep analysis of data is presented in [2]. The problem of documents clusterization on the basis of phrase similarity and specified terms with the use of DBMS Neo4j is put forward in [4]. Similarity of documents can be applied for an automatic comparing of a résumé with the description of a job. As the results of experiments it was found that a hybrid similarity provided a greater degree of the accuracy of similarity.

The problem of an estimation for the quality of cluster’s structure obtained in the process of cluster analysis was considered at the International science and technical and science and methodological conference [5].

3. Visualization and similarity estimation
The main fields of application for the graph data models are social systems. Geo-spatial systems, fraud detection systems and search systems [1]. Graph data base keeps them in the form of an intellectual chart allowing to find rather quickly and to build any kind of the relations in the form of a graph with graph edges and nodes. Intellectual chart uses visual images, plots, diagrams and animation. For visualization very often different colors and fonts are used. Graph data bases provide a high connectivity of the stored data. Degree of the data connectivity in a number of the problems of their processing proves to be even more valuable than just the primary data themselves.

The study of capabilities of the graph DBMS on the visualization and data analysis is convenient to perform with the use of rather popular graph DBMS Neo4j, since it has facilities characteristic for the graph data model in the form of the proper graph algorithms integrated in the common library. DBMS Neo4j has the facilities of data visualization in the form of graphical images as well. Such way of visualization is distinguished by a considerably greater convenience and clearness as compared with a tabular way of data visualization that is characteristic for the spreadsheets in the relational data model.

In order to reveal a high data connectivity in DBMS Neo4j let us make corresponding analysis. The primary point in such analysis is an assignment of the key values. For simplicity and at the same time in order to achieve sufficient representability of a key set let us specify two key values. Let us assign different colors to these key values. In fact, colors are used in the graph data model for the clearness of the data visualization, for a convenient visual perception of the users and efficient examination of results of various calculations performed by the requests from the graph DBMS. As a first key value lets us specify the value of “subscriber” and assign him some color that is proper for vision, for
example, the violet one. As a second key value let us specify the value “community” and assign it any other color that is in a good contrast with the color of the first key value at their joint visual perception, for example, yellow color. As a result, corresponding graph is obtained where each node is painted in one of the two chosen colors: violet and yellow.

Concept of the “entity” plays an important role in the graph model as one of its basic ideas. A set of all entities for a certain graph data model forms some hierarchical structure where each entity is related to the properly determined level. The number of levels is not regulated and it can differ for different certain graph data models. The lowest level is considered as the first one. The entities of the first level are not determined through any other entities. Entities of any other level are determined through those ones of the lower levels. Each entity independent on its level has several of its own attributes. Color of the entity can be considered as one of these attributes.

In the example considered above the key values of “subscriber” and “community” are the entities of the first level and they are characterized by the violet and yellow colors, respectively, as if by two different values of their attributes that have the meaning of color. Thus, the key value of “subscriber” is identified as a violet entity, while the key value of “community” is identified as a yellow entity.

One of the characteristic capabilities of the graph DBSM is a qualitative defining of similarity for big data sets. Let us explore this capability with DBMS Neo4j as an example. To do this just after loading of the primary data it is required to study identical subscribers. Let us create a graph of similarity for the pairs of subscribers with the same values of “statement”. This value is chosen as a key one. It is used to realize sampling and then the similarity is revealed. Now we use the algorithm of cosine similarity from the library of the algorithms Neo4j for the graphs.

Two data sets compared for their similarity are presented by vectors $A = (A_i)$ and $B = (B_i)$, where $i = 1, n$, and the same dimensionality $n \in N$, where $n \geq 2$, endowed by Euclidian (Hölder with parameter of 2) norm.

$$
\|A\| = \sqrt{n \sum_{i=1}^{n} A_i^2};
$$

$$
\|B\| = \sqrt{n \sum_{i=1}^{n} B_i^2}.
$$

Note, that certain elements of the data integrated into the sets are presented by the components $A_i \geq 0$ and $B_i \geq 0$ of vectors $A$ and $B$, respectively.

The same dimensionality of vectors and the common Euclidian norm for both of them allows interpreting them as the elements of the common Euclidian space with a dimensionality $n$ and scalar product

$$
A \cdot B = \sum_{i=1}^{n} A_i B_i.
$$

Natural norm generated by the scalar product (2) is

$$
\|A\| = \sqrt{A \cdot A};
$$

$$
\|B\| = \sqrt{B \cdot B}.
$$

and it coincides with the norm (1).

In this Euclidian space scalar product of (2) specifies cosine of the angle between the vectors $A$ and $B$ in the following way:

$$
\text{similarity}(A, B) = \frac{A \cdot B}{\|A\| \|B\|}.
$$
With the account of (1)-(3) equality (4) with the components of $A_i$ and $B_i$ for the vectors $A$ and $B$ takes the following form:

$$similarity(A, B) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}.$$  

(5)

Interpretation of the data sets $A = (A_i)$ and $B = (B_i)$, as the vector elements of the common Euclidian space makes it possible to consider the norms (1), (3), scalar product (2) and cosines (4), (5) as vector operations invariant relative to the linear transformations of coordinates in this space.

An algorithm of cosine similarity from the graphs library of algorithms Neo4j uses formula (5). In this case the informational similarity of the data sets $A$ and $B$ can be estimated by cosine of the angle between the corresponding vectors (5). In general, for the arbitrary elements $A$ and $B$ of Euclidian space inequality of (6) kind is realized

$$-1 \leq similarity(A, B) \leq 1$$

(6)

However, for the real data sets of $A$ and $B$ when performing the informational search inequality (6) is substituted by inequality of the following type

$$0 \leq similarity(A, B) \leq 1$$

(7)

under equality (5) and inequalities $A_i \geq 0$ and $B_i \geq 0$. Inequality (7) means that in the case of the informational search cosine similarity varies within the range of 0 to 1.

Let us make a computational experiment for the assessment of the quality of similarity for the large data sets. It was realized using database of 550 subscribers in one of the social networks that was found in the open access. Then apply procedure for detecting of pairs of the identical subscribers. As a result, for these 550 subscribers’ similarity was detected for 10 012 pairs. Next, apply procedure for creating of similar interrelations between those nodes that have similarity equal to unity. An example of such graph is depicted in figure 1.

![Figure 1. Graph of pairs’ similarity.](image-url)

Now let us run Weakly Connected Components (WCC) algorithm from the library of algorithms Neo4j over this graph of similarities. This algorithm finds the sets of connected nodes in the non-oriented graph. It is often utilized at the preliminary stages of the graph’s analysis in order to understand its structure. Let us use a stream version of the algorithm where the request returns a stream of pairs. As a result of algorithm’s operation connection is created between each node of the graph and its corresponding community. Figure 2 represents grouping of subscribers according to the key value of “statement”.

4. Assessment of the quality in search of similarity

In order to assess the quality in search of similarity we use clusterization. However, it should be noted that nowadays an assessment of the clusterization quality is studied in a less extent than the cluster analysis.

Clusterization aggregates a set of data points into the groups. The main object of clusterization is to increase similarity and difference within the group. Let us run algorithm of clusterization DBSKAN (Density-based spatial clustering of applications with noise). It involves two parameters: vicinity radius $\varepsilon$ and the minimum number of points required for the formation of cluster.

For the direct estimation of the quality in search of similarity a silhouette assessment technique can be applied which is realized in DBSKAN algorithm. Silhouette assessment is applied for the assessment of the quality of cluster finding bearing in mind that this cluster involves the given object. To realize it an assessment is made if and how this object is like the other objects of its own cluster as compared with the objects of the other clusters [4]. With the help of silhouette estimation it is possible to measure separation distance between the obtained clusters. The range of silhouette values is within the limits of $[-1; 1]$. If the calculated silhouette value is close to the value of $-1$, then it follows a conclusion that the object is wrong classified by its membership to its own cluster and is relatively close to the neighboring cluster. If the silhouette value proved to be close to 0, it means that the object can be involved into another neighboring cluster as well as that the object is equally far from both of the clusters. But if the value of silhouette is close to unity it means that the object is far from the neighboring cluster and is relatively close to the considered cluster; it means that the object is well clustered (and thus it is related to its own cluster). Silhouette coefficient is calculated in accordance to the equality

$$ S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}, \quad (8) $$

where

- $a(i)$ – is the mean distance from $i$-th object to all other objects in the same cluster;
- $b(i)$ – is the least mean distance among all of the clusters from $i$-th object to all other objects of this cluster or, in other words, the mean distance from $i$-th object to all other objects of the nearest cluster.

Silhouette estimation means the value of silhouette coefficient (8) averaged over all of the samples. The results of accomplished experiments concerning silhouette estimation are presented in Table 1.

| Experiment | Estimation |
|------------|------------|
| Test 1     | 0.76       |
| Test 2     | 0.78       |
| Test 3     | 0.77       |
From Table 1 it is seen that for all three tests silhouette estimation is close to unity thus implying correct attribution of the object to its own cluster.

Method of AdjustedMutualInformation (AMI) as a technique for the estimation of the quality in the similarity finding can be also applied as an alternative to the method of silhouette assessment. It is based on the calculation of the measure in MutualInformation similarity of two partitions that takes into account probabilities of the objects’ attribution to certain clusters [3]. The measure of MutualInformation is expressed by the following formula:

\[
AMI(U,V) = \frac{MI - E[MI]}{\max\{H(U),H(V)\} - E[MI]},
\]

where 
\(E[MI]\) is the expected value of the measure of MutualInformation;
\(H(U)\) and \(H(V)\) – is the entropy of partitions \(U\) and \(V\), which can be calculated in the following way:

\[
H(U) = \sum_{i=1}^{\|U\|} P(i) \ln P(i);
\]
\[
H(V) = \sum_{i=1}^{\|V\|} P(i) \ln P(i).
\]

For any of the partitions \(U\) and \(V\) the following inequality is implemented [3]:

\[
0 \leq AMI(U,V) \leq 1.
\]

And proximity of the value \(AMI(U,V)\) to unity means the qualitative finding of similarity while proximity to zero means a poor similarity.

Results of the experiments with the use of the trial data concerning the estimation of the measure in partitioning similarity according to the formulas (9), (10) are presented in Table 2.

**Table 2. Results of assessment of the measure in partitioning similarity.**

| Experiment | Estimation |
|------------|------------|
| Test 1     | 0.78       |
| Test 2     | 0.81       |
| Test 3     | 0.76       |

From the Table 2 it is seen that in all of three tests the estimation of the measure in partitioning similarity is close to unity thus indicating at the correct classification of the object.

Thereby, application of both methods (silhouette assessment technique and method of the estimation of measure in partitioning similarity) allows making a general conclusion that the detected similarity with the help of graph DBMS Neo4j is of rather high quality.

5. Conclusion

As a result of the performed experimental studies it was revealed that graph DBMS Neo4j can interpret vision data with a high quality. Data visualization in the form of graphical images is convenient for the perception and interpretation. The use of simple algorithms in the graph DBMS Neo4j makes it possible to obtain connected data. Assessment of the trials employed in graph DBMS Neo4j demonstrated good data connectivity and a high quality of similarity. To attain the same result with the use of relational DBMS it is required to make more requests and more classifying operations. In our further investigations we plan to do the corresponding comparative analysis.

References

[1] Guralnick R I 2016 Some of the challenges for graph databases. *Proceedings of the ISP RAS*
28(4) pp 193-216

[2] Bochkarev P V 2016 Graph data models International Scientific and Technical Journal “Theory. Practice. Innovations”

[3] Steve A 2015 Evaluation Of Graph Databases Performance Through Indexing Techniques International Journal of Artificial Intelligence & Applications (IJIA) 6(5) pp 87-98

[4] Litvinov V L 2017 Rebkovets I V and Yakovlev I V “quality Assessment of algorithms for cluster data analysis” VI International Scientific-technical and Scientific-methodological Conference p 321

[5] Preeti K 2020 Document Clustering based on Phrase and ingle Term Similarity using Neo4j International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN 9(3) pp 3188–3192