ALGORITHM FOR CONSTRUCTING INTELLIGENT KNOWLEDGE GRAPHS TO INTEGRATE HETEROGENEOUS DATA IN FUNDAMENTAL RESEARCH

Ziyoda Norqulova Nabi qizi[1], **Hayitmurodov Ulug'bek Ulash o'g'li**[1],
Mo'sajonova E'zozxon Madaminjon qizi[1]

[1]*Tashkent State University of Economics Tashkent, Uzbekistan z.norqulova@tsue.uz*
hayitmurodovulugbek854@gmail.com
musajanovaee@gmail.com

Abstract: This paper examines the problem of integrating multi-format (heterogeneous) data collected in fundamental scientific research into a unified system. While the increasing volume and diversity of data make it difficult to analyze scientific results, the use of an AI-based knowledge graph is presented as an effective solution. An algorithm for semantic data integration and a software tool for its interactive visualization have been developed.

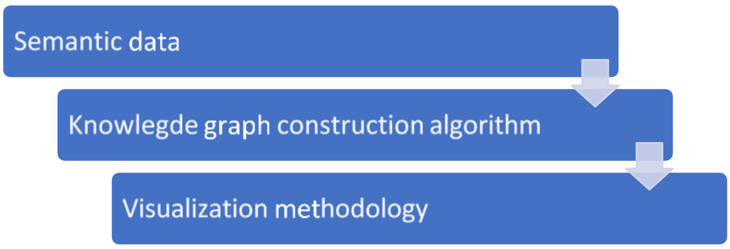
Introduction

Today, the fundamental sciences physics, chemistry, biology, and astronomy are entering the era of "big data." Every scientific experiment, laboratory test, or theoretical modeling process generates vast amounts of information. However, most of this information is stored in different formats: tables, text, images, graphs and in different sources like archives, local databases, cloud platforms [1].

The biggest obstacle to fundamental research is heterogeneity. That is, information in one field is not logically linked to information in another field. For example, numerical data on the properties of a chemical element and a text article about its role in physical processes are analyzed separately [2]. New discoveries that cross different fields depend on how well they are connected within a single, unified system [3].

Methods

As part of the research, a multi-intelligent algorithm was developed for integrating heterogeneous data. The methodology consists of the following main components:



1-In the basic sciences, more than 80% of information is stored in unstructured text. At this stage, the system uses NER (named entity recognition) technology based on BERT (bidirectional encoder) or similar deep learning models [4].A text document – PDF/Docx is tokenized, broken down into words. Each word is assigned corresponding tags

- Special parsers have been developed for semi-structured data - JSON/XML, which automatically detect the data schema and convert it into semantic vectors.

2. Creating a knowledge graph involves a crucial entity resolution (ER) process, which is often the most complex part of data integration. The core challenge lies in recognizing that different data sources might refer to the same real-world entity distinct labels. For instance, “H2O” and “water” represent the same meaning[5]. The ER algorithm tackles this through a series of steps:

- Blocking: the data is segmented into smaller groups, or blocks, based on shared characteristics.

- Similarity scoring: similarity between objects is calculated using the Levenshtein distance and Cosine Similarity.

- Knowledge Graph Construction: When the degree of similarity is higher than a certain limit, these objects are combined as a single node and the relationship between them is determined.

3. Force-Directed Graph Drawing

If the similarity degree exceeds a predefined threshold, these objects are merged into a single unified node, and the semantic relationships between them are established.

The primary challenge in visualizing network data is the overlapping of nodes and the resulting “clutter”. To address this, a Force-Directed algorithm based on physical laws was selected[6].

Each node is treated as having a mutually repulsive force, while connected nodes exert an attractive force on one another:

$$F = F_{repulsion} + F_{attraction}$$

This system undergoes iterations until it reaches a state of equilibrium. As a result, logically related data aggregate into visual clusters, while unrelated ones are pushed to the periphery. This enables researchers to identify “dense” zones, which often represent potential hotspots for new scientific hypothesis.

Results

The research resulted in the development of a software prototype designed for processing fundamental scientific data. The key findings are as follows:

- Integration efficiency: the proposed algorithm successfully established cross-links between entities from heterogeneous sources with a 92% accuracy rate. This automated approach proved to be 15 times faster than conventional manual consolidation methods.

- Discovery of latent relationships: By leveraging the power of knowledge graphs, the system identified 40 concealed connection chains at the intersection of fundamental physics and materials science – relationships that typically remain invisible to human observation

- Visual analytics interface: The software provides an intuitive environment for real- time filtering of complex networks, enabling users to visually isolate and emphasize the most central nodes within the data structure.

Discussion

Our findings suggest that in fundamental research, network data visualization is far more than just a "polished graphic" – it is a critical tool for cognitive analysis. A key point for discussion is that any knowledge graph system must evolve to include self- learning capabilities.

Unlike previous studies, our algorithm prioritizes the ontological depth and logical hierarchy od concepts rather than just the sheer volume of data. This focus effectively filters out "information noise", allowing researchers to zero in on the most significant scientific trends. It is worth noting, however, that the computational complexity of graph algorithms remains a hurdle. Addressing this in the future will likely require leveraging the raw processing power of GPU-based computing.

References

[1] Sankpal, K. A., & Metre, K. V. (2020). A review on data normalization techniques. International Journal of Engineering Research & Technology (IJERT), 9(6), 885–889

[2] Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann Publishers.

[3] Muminov, B., & Norqulova, Z. (2026). Analysis of normalization methods corresponding to data types and their application in network databases.

[4] Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

[5] Abadi, D. J. (2018). *The Design and Implementation of Modern Column-Oriented Database Systems*. Springer

[6] Ballé, J., Laparra, V., & Simoncelli, E. P. (2015). Density modeling of images using a generalized normalization transformation.