0619 ინფორმაციისა და კომუნიკაციის ტექნოლოგიები
INFORMATION AND COMMUNICATION TECHNOLOGIES (ICTS)

Knowledge extracting from big textual datasets

Irakli Khachidze
Samtskhe-Javakheti State University,
E-Mail: irakli.khachidze3@gmail.com

Abstract

Generally, data mining seeks to uncover, examine, and analyse significant information from large data sources using various techniques and algorithms. However, with big data, processing and extracting knowledge using traditional methods and algorithms presents substantial challenges. Knowledge mining from big data involves extracting insights and patterns from vast data sets, utilizing the lambda architecture and the parallel processing paradigm of big data. The article discusses a new approach to big data mining. The originality of this work is based on the representation of knowledge in the form of a graph model, as well as the assembling of a single knowledge model from individual graph fragments of knowledge. Combining these two concepts within the context of lambda architecture enables efficient execution of large-scale mining tasks in parallel processing of big data.

Keywords: big data mining, knowledge representation, artificial intelligence, lambda architecture

1. **Introduction to the problematics.** The term "Big Data" stems from the enormous volumes of data that have become difficult to analyse, store, and process using conventional methods due to their complexity. Data mining (DM) involves extracting valuable and hidden insights and patterns from obscure data, aiding business decision-making. Data mining is also referred to as knowledge discovery in data (KDD) [1].

Data mining is defined as the process of discovering patterns within large datasets and making predictions for new data. Data science typically employs a structured methodology to address problems and involves working with data, algorithms, and models. This methodology outlines the knowledge discovery process from structured data. Since the late 20th century, research in data science has aimed to define knowledge discovery (or knowledge discovery in databases [KDD]) as synonymous with the knowledge discovery process (KDP) [3]. Data mining is a step within the broader knowledge acquisition process. Although KDD (or KDP) and data mining are often used interchangeably, data mining is merely a subset of the entire process, concentrating on the use of algorithms to extract patterns from data.

Modern information technology defines information using the following concepts: the DIKW Pyramid (DIKW pyramid – understanding the difference between Data, Information, Knowledge, and Wisdom) [4,5].

2. Data lakes as an object of research. One use to store all types of big data, whether Structured, Semi-Structured, or Unstructured from multiple data sources into a Data Lake. In this paper, we explore knowledge engineering within the context of data lakes, big data analytics, and information integration and knowledge management. It is widely recognized that knowledge engineering, a branch of artificial intelligence, involves a substantial amount of knowledge, including metadata and information about data objects, which describes their content, structure, and processes.

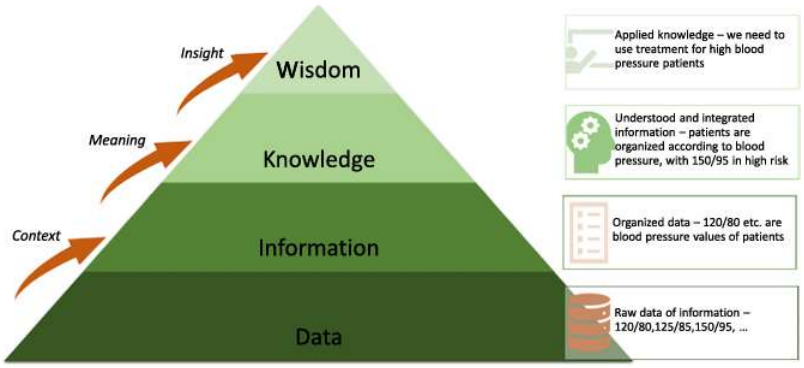


Figure 1.

Because knowledge is derived primarily from textual data, consider so-called big data stores - Data lakes still have limitations, such as lacking transaction support (making it difficult to update data lakes) and ACID compliance (hindering concurrent reads and writes). Unlike data warehouses, data lakes store raw data in various formats that can be utilized for both current and future use cases [6]. The large variety of interaction types involved in knowledge management processes reflects the variety of textual fragments associated with data lakes.

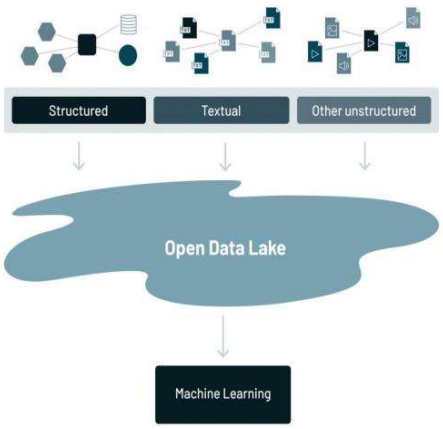


Figure 2

Moreover, in the following years, the development of big data, especially the paradigm of data lakes, brought the necessity of new efforts to acquire knowledge, which was caused by the urgency of developing completely different approaches. Besides, Machine learning is primarily focused (more like limited) on solving clustering or classification tasks. In this context, clusters do not represent knowledge but rather datasets, whereas knowledge should be viewed more in the form of a graph.

3. A new understanding of knowledge engineering. This article explores a novel approach centered on the synergistic hierarchical representation of knowledge. Knowledge representation models differ across levels, distinguishing between external reflections of knowledge and internal (actual) models. External reflection is explicit and observable. External perception can be divided into linguistic and ontological types. When we perceive a word or concept as a sequence of phonemes during language acquisition, it creates an ontological model with relevant attributes. Consequently, when we read or hear this sequence in a specific language, it aligns with this model in our consciousness. However, an important nuance is that the order of phonemic sequences significantly changes the ontological model. For example, two words made up of the same letters but in a different order have completely distinct ontological models with their respective attributes. Therefore, both the composition and sequence of letter-sounds are crucial. On a micro level, any real-world entity (such as an object, event, or other item) is connected to attributes perceived in consciousness through sensory channels. These attributes create a unique neural network that represents an ontological semantic entity, independent of linguistic

aspects. At the phrase or sentence level, this results in a structure made up of a sequence of words. Ontology, as widely acknowledged, provides a formal framework for representing knowledge [7].

Generally, each idea (phrase or sentence) can be structured as: ⟨subject, predicate, object⟩, tense. The subject acts as the distinct identifier of the event, while the predicate and object describe its properties. Adding a timestamp clarifies the event's dynamics. Each idea or concept is governed by the criterion of integrity, defined by the semantic compatibility or synergy among its components. Integrity is a crucial criterion in knowledge engineering, especially in the construction of knowledge, which is closely linked to the concept of synergy.

At the next level, the sentence itself functions as an entity (node) with its graph serving as the foundational knowledge model. At the macro level, a specific knowledge model is seen as a semantic macro entity, which in turn creates a hypergraph of meta-awareness and extends further in the hierarchy. [8].

At all epistemological levels of the knowledge system, a necessary condition applies. Specific knowledge graphs, in various configurations, can be seen as nodes within the knowledge graph at higher levels. According to fractal principles, these epistemological layers of foundational knowledge exist across different dimensions, collectively forming a unified knowledge system.

Currently, all existing approaches are preliminary approximations of natural artificial intelligence. Therefore, it is essential to develop a new generation of artificial intelligence founded on an innovative paradigm of knowledge representation.

4. The role of Machine learning into knowledge management. As for machine translation approaches are far from perfect because a computer cannot adequately translate textual information. In data mining, when tackling clustering or classification tasks, machine learning typically views knowledge merely as a collection of data. This approach persists because modern computers seldom employ knowledge representation models such as semantic networks or framework systems, which only crudely approximate natural intelligence. Machine learning primarily focuses on solving clustering or classification tasks, where a data cluster represents data rather than knowledge. Knowledge, in contrast, should be conceptualized more akin to a graph or hypergraph.

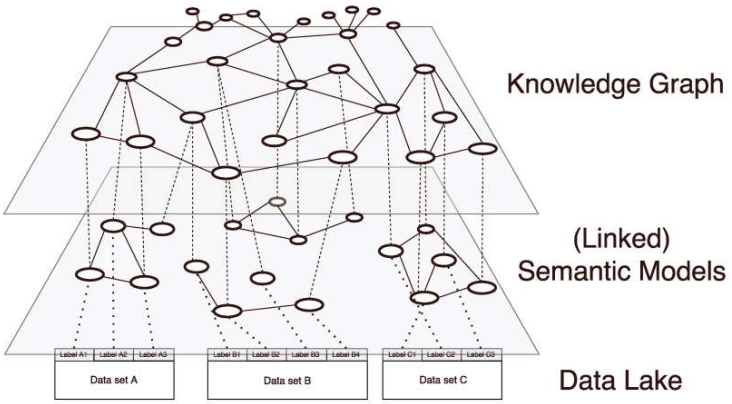


Figure 3

This article examines knowledge engineering within contemporary information systems, encompassing artificial intelligence, big data analytics, information integration, and knowledge management. It is grounded in a novel approach to knowledge representation. [9]. Machine translation approaches are far from perfect because a computer cannot adequately translate textual information. In data mining, when addressing clustering or classification tasks, machine learning often perceives knowledge simply as a dataset. This perspective stems partly from the fact that modern computers seldom employ knowledge representation models like semantic networks or framework systems, which are only a basic approximation of natural intelligence.

5. The phases of knowledge extraction process from big textual datasets (our approach)

Let’s consider the knowledge extraction process from big textual datasets using lambda architecture. According to our approach, the process of knowledge extraction from a text dataset consists of four phases as shown in Figure 4:

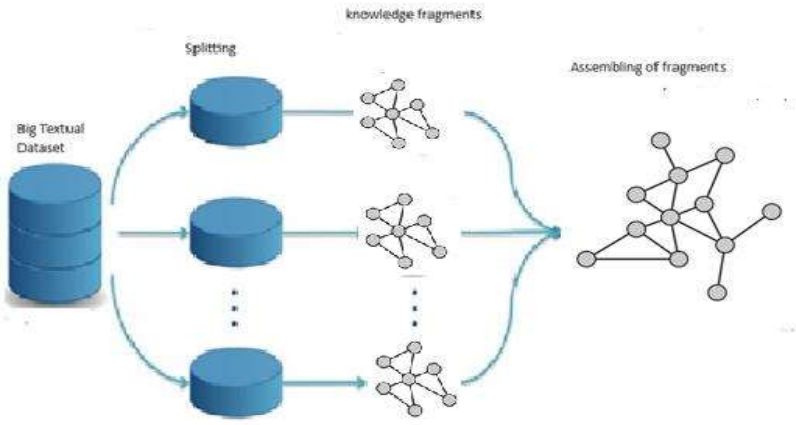


Figure 4

- Splitting phase, when a large text data set is fragmented.
- The second phase represents the KDD process, when the formation of graph fragments of knowledge from textual fragments (like figure 4).
- In the third phase, the process of assembling the knowledge model from the graph fragments of knowledge is completed, according to the KDD processes (like figure 5). The model of the KDD process consists of the following steps (input of each step is output from the previous one), in an iterative (analysts apply feedback loops if necessary) and interactive way [10].

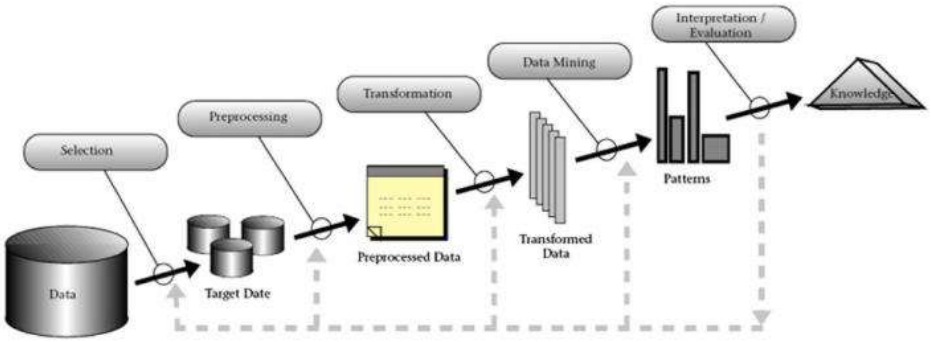
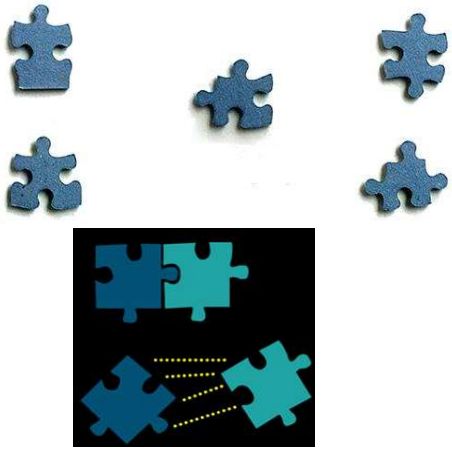


Figure 5. The KDD processes

- In the fourth phase, when the knowledge graph model is optimized through synergistic and complementary (docking) approaches.

One method to accelerate big data mining is by distributing the training process across multiple machines simultaneously. The framework is designed to handle very large textual data in parallel by breaking the work into various independent tasks. To extract knowledge from extensive textual data, algorithms utilizing parallel computing are employed, where a large data set is divided into several subsets, and mining algorithms are then applied to these subsets. As a result of the processing of each subgroup, a certain fragment of knowledge is obtained. This phase is the construction of an overall knowledge model from these fragments, based on parallel computing [2].

And finally, the knowledge graph model derived from large text data mining is optimized with the goal of entropy minimization rather than elimination. (This phase is not discussed in this article). Optimization of the assembled knowledge model is a separate scientific task that goes beyond the scope of this article. Thus, let's limit ourselves to the task of assembling knowledge from fragments, which has certain analogies in form of jigsaw puzzles (figure 6 a, b) and molecular or drug design processes. Unlike puzzles shapes (figure 6 a), the number of pieces of knowledge is incredibly large, although they can also be viewed as an axon-dendritic terminal model. And also, the main criterion of assembly is the maximization of compatibility or synergy of fragments.



• b)

Figure 6

As for the second analogy, here we have much more diverse forms of their axon-dendritic models in the form of molecules, which greatly increases the variation in the selection of fragment conections. Of course, it remains the same main criterion of assembly the maximization of compatibility or synergy of fragments [11].

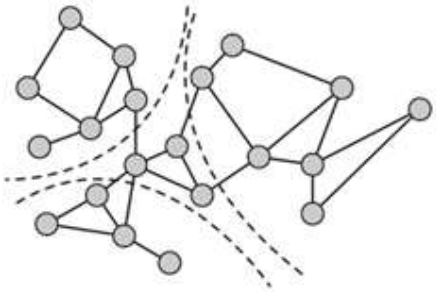


Figure 7

In this article, we tried to present a fragment of textual data in the form of an "information molecule". First, we identified the problem of descriptors, which are sets of textual data. It is through them that the graph based model of knowledge should be built. The main problem in this task is the problem of fitting fragments. Initially, we planned to use a convolutional neural network (CNN), where we initially tried to add several convolutional layers. In general, CNN works well with high-dimensional tensors, but since our data fragments are high-dimensional graphs, this shows the inefficiency of complex models like CNN. This paper distinguishes two different types of knowledge: synergistic knowledge and complementary (docking) knowledge, which separately lead to different outcomes in terms of knowledge creation [12,14].

6. Entropy in thermodynamics and information theory. The statistical definition of thermodynamic entropy aligns with Clausius' classical entropy, typically denoted as S, of a physical system. This article delves into the connections between these two concepts and explores the extent to which they are interconnected. [13].

$$dE = -pdV + TdS$$

Where:

$$S = -k \sum_i p_i \log p_i,$$

Where: p_i is the probability of themicrostate i taken from an equilibrium ensemble.
and the defining expression for entropy in the theory of information established by E. Shannon is of the form:

$$H = -\sum_i p_i \log p_i,$$

Where: p_i is the probability of the message m_i taken from the message space M.

The statistical definition of thermodynamic entropy in molecular modeling coincides with the classical Clausius entropy. As for the new approach to these two main strategies of knowledge assembly, it assumes their joint or complex use. In contrast to molecular docking, where, from the point of view of connecting molecules, the energy minimum is used as a criterion according to the first law of thermodynamics, in our case, that is, in the case of connecting text fragments, we propose an informative variant of Shannon entropy.

This paper presents a novel and dynamic approach to defining the ultimate goal of strategic alliances: synergy. The importance of knowledge synergy is clear in relation to the integration of fragments; However, knowledge engineering is mostly concerned with knowledge compatibility and knowledge complementarity (docking).

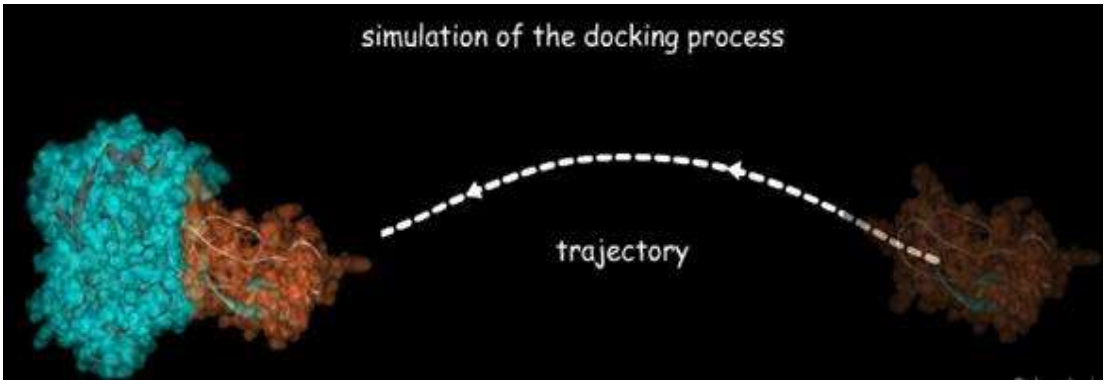


Figure 8

The purpose of assembling a knowledge model from knowledge fragments is to discuss compatibility and complementarity in knowledge management (KM). Nevertheless, they are two complementary processes if they are well integrated.

7. Conclusions

The article discusses a new approach to big data mining, particularly, the knowledge extraction process from big textual datasets using lambda architecture. According to our approach, the process of knowledge extraction from a text dataset consists of follow phases: splitting of large text data set, the formation of graph fragments of knowledge from textual fragments and assembling the knowledge model from the graph fragments of knowledge. The main particularity of this approach consists in the representation of knowledge in the form of a graph model, as well as the assembling of a single knowledge model from individual graph fragments of knowledge. Combining these two concepts within the context of lambda architecture enables efficient execution of large-scale mining tasks in parallel processing of big data.

References

1. Jinlong Wang, Jing Liu, Russell Higgs, Li Zhou, Chuanai Zhou (2017) “The Application of Data Mining Technology to Big Data”, IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC).
2. AS Hashmi, T Ahmed (2016) “Big Data Mining: Tools & Algorithms”, International Journal of Recent Contributions from Engineering, Science & IT (iJES)
3. Valentina Janev. Damien Graux. Hajira Jabeen. Emanuel Sallinge. “Knowledge Graphs and Big Data Processing”. ISBN 978-3-030-53198-0 https://doi.org/10.1007/978-3-030-53199-7
4. Ackoff, Russell .L. “From Data to Wisdom,” Journal of Applied Systems Analysis 16 (1989): 3–9.
5. Nikhil Sharma, “The Origin of Data Information Knowledge Wisdom (DIKW) Hierarchy”, (Google Inc, February 2008).
6. Jinlong Wang, Jing Liu, Russell Higgs, Li Zhou, Chuanai Zhou (2017) “The Application of Data Mining Technology to Big Data”, IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC).
7. AS Hashmi, T Ahmed (2016) “Big Data Mining: Tools & Algorithms”, International Journal of Recent Contributions from Engineering, Science & IT (iJES)
8. Valentina Janev. Damien Graux. Hajira Jabeen. Emanuel Sallinge. “Knowledge Graphs and Big Data Processing”. ISBN 978-3-030-53198-0 https://doi.org/10.1007/978-3-030-53199-7

9. Ackoff, Russell .L. “From Data to Wisdom,” Journal of Applied Systems Analysis 16 (1989): 3–9.
10. Nikhil Sharma, “The Origin of Data Information Knowledge Wisdom (DIKW) Hierarchy”, (Google Inc, February 2008).
11. Joseph Reis, Matthew Housley,“Fundamentals of Data Engineering” (2022). Published by O’Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472. ISBN 978-1-098-10830-4. Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC).
12. Meparishvili, B., Janelidze, G., Khachidze. Ir, “Some new aspects of Knowledge Engineering”. International scientific-practical conference "Modern challenges and achievements in information and communication technologies“ October 12-13, 2023. GTU.
13. Ediberidze A., Meparishvili B., Janelidze G, (2008). New Approaches to a Modeling of Knowledge. IFAC 9 th Workshop on Intelligent Manufacturing Systems (IMS’08), Szczecin, Poland, October 9-10, 2008. 99-103 pp.
14. Meparishvili, B. Gachechiladze, T. Janelidze, G. NATO Science for Peace and Security Series “Complexity and Security”, 2007, ISSN 1874-6276. 379-388 pp.
15. Jinlong Wang, Jing Liu, Russell Higgs, Li Zhou, Chuanai Zhou (2017) “The Application of Data Mining Technology to Big Data”, IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)
16. Kervalishvili, P., Meparishvili,B., (2008). “Molecular Machines-Modeling Approaches”. ERA-2 Proceedings The Contribution Of Information Technology Science, Economy, Society and Education. T.E.I. of PIREAUS: 453-460 pp.
17. Zhiting, Z.: Synergistic Learning: A New Learning Paradigm in Connected Age. Keynotes on Advanced Seminar of 1st Global ET Summit Conference. Shanghai, China. (July 30, 2006)
18. Zorich, V. Entropy in Thermodynamics and in Information Theory. *Probl Inf Transm* 58, 103–110 (2022). https://doi.org/10.1134/S0032946022020016
19. Nielsen, B. B. (2000). Synergies in Strategic Alliances: Motivation and Outcomes of Complementary and Synergistic Knowledge Networks. Copenhagen Business School. Working Paper / Department of International Economics and Management, Copenhagen Business School No. 4-2000.

ცოდნის ამოღება დიდი ტექსტური მონაცემთა ნაკრებიდან

ირაკლი ხაჩიძე

რეფერატი

ზოგადად, მონაცემთა მოპოვება მიზანმიმართულია აღმოაჩინოს, შეისწავლოს და გააანალიზოს მნიშვნელოვანი ინფორმაცია დიდ მონაცემთა წყაროებიდან სხვადასხვა ტექნიკისა და ალგორითმის გამოყენებით. თუმცა, დიდი მონაცემების შემთხვევაში, ტრადიციული მეთოდებისა და ალგორითმების გამოყენებით ცოდნის დამუშავება და მოპოვება მნიშვნელოვან გამოწვევებს უკავშირდება. დიდი მონაცემებიდან ცოდნის ამოღება მოიცავს ინფორმაციისა და შაბლონების აღმოჩენას მონაცემთა დიდი ნაკრებიდან, ლამბდა არქიტექტურისა და დიდი მონაცემების პარალელური დამუშავების პარადიგმების გამოყენებით. სტატიაში განხილულია ახალი მიდგომა დიდ მონაცემთა მოპოვებაში. ნაშრომის ორიგინალურობა მდგომარეობს იმაში, რომ ცოდნის წარმოდგენა განხილულია, როგორც გრაფული მოდელი და ასევე მოიცავს ცოდნის ცალკეული გრაფული ფრაგმენტებიდან ერთიანი ცოდნის მოდელის აწყობას. ამ ორი ცნების გაერთიანება ლამბდა არქიტექტურის კონტექსტში შესაძლებელს ხდის მონაცემთა აღმოჩენის ფართომასშტაბიანი და კომპლექსური ამოცანების ეფექტურ გადაწყვეტას დიდი მონაცემების პარალელური დამუშავებით.

საკვანძო სიტყვები: დიდი მონაცემების მოპოვება, ცოდნის წარმოდგენა, ხელოვნური ინტელექტი, ლამბდა არქიტექტურა.