SIGNIFICANCE AND COMPARABILITY OF UNSTRUCTURED AND SEMI-STRUCTURED DATA IN THE MODERN WEB

Yavor Tabov

assistant, PhD. Department of Information Technologies and Communications, University of National and World Economy, *e-mail: jtabov@unwe.bg*

Abstract

This document examines the characteristics and features of unstructured and semi-structured data in the modern web. Scales and trends for the growth of these types of data are presented. Popular approaches for analyzing unstructured and semi-structured data are discussed, with a comparison made regarding the different and common approaches. Finally, conclusions are drawn based on the prepared material.

Key words: unstructured data, semi-structured data, Internet, web scraping, machine learning

JEL: C88, L86.

1. Introduction of unstructured and semi-structured data and their place in the modern web

Technological progress has led to a significant increase in data volume in recent years. This growth in big data has brought about changes in the capacity and capabilities of computing systems due to the sheer volume, diversity (including structured, unstructured, and semi-structured data), and speed at which data is generated. A vast volume of data, whether it's unorganized or partially structured, is being produced continually, whether it's on a daily, minute-by-minute, or even second-by-second basis, stemming from a multitude of sources in our everyday lives. The count of internet users is experiencing rapid and substantial growth with each passing day [3]. Extracting valuable information from various data types has become a challenging endeavor, particularly given the vast volume and intricacy of unstructured data [1].

Unstructured data originates from both machine-generated and human-generated sources, and it is generally categorized into two main groups: non-textual and textual. Non-textual unstructured data includes multimedia content such as static images, videos, and audio files. Examples of textual unstructured data encompass items like email messages, collaborative software and instant messages, memos, word processing documents, and PowerPoint presentations. Various standards for handling unstructured data include open XML, SMTP, SMS, CSV [8].

Unstructured data refers to data in an unfamiliar format, which can be challenging and time-consuming to handle. It encompasses various types, including text files, videos, images, and more. Nowadays, organizations possess abundant data resources, but extracting value from unstructured data remains a significant challenge [3]. Early data extraction systems primarily focused on text, which remains the most extensively explored form of data by both commercial systems and the scientific community to this day [11].

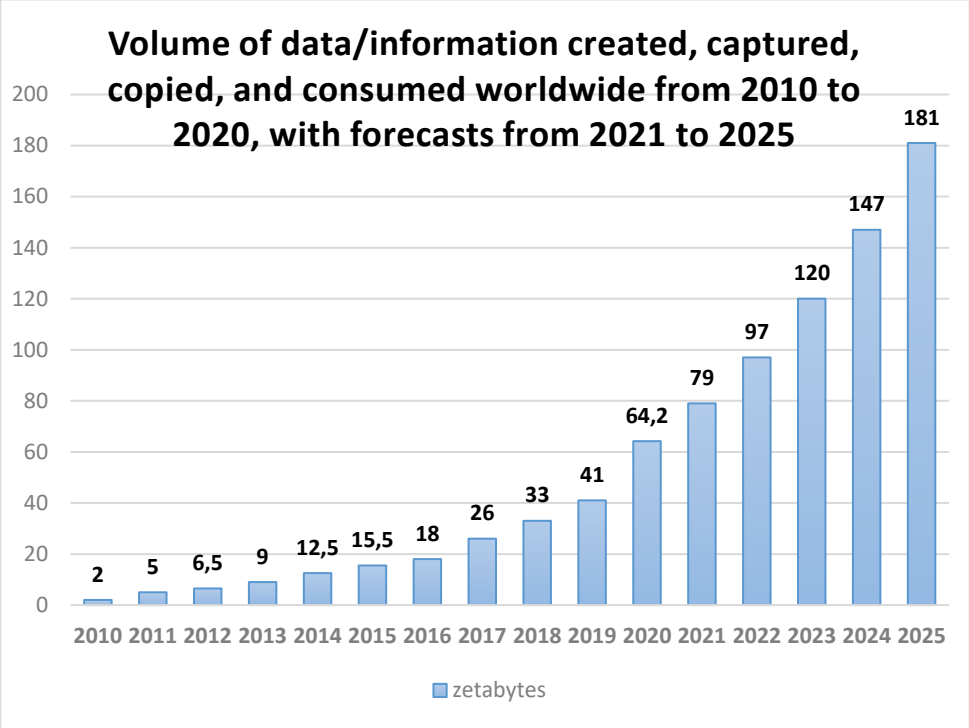
The significance of unstructured data is gaining greater recognition among both scholars and the public. Machine learning paradigms offer a computational advantage through their capacity to derive meaningful patterns from unstructured data for addressing real-world problems. However, it's widely acknowledged that one limitation of these paradigms is the often-inscrutable nature of the models they employ. To casual observers, the data modeling process is concealed by machine learning technologies, which sometimes leads to the perception of AI as a form of magic [7].

Semi-structured data, while considered a subset of structured data, lacks a strict data model or a formal structure. It doesn't require a predefined schema and often includes optional tags or markers

that help delineate semantic elements and establish hierarchies within the data. This type of data has become more prevalent, especially since the internet hosts various data types beyond just full-text documents and databases, and various applications require a means of information exchange. Managing semi-structured data typically involves using markup languages like XML or JavaScript Object Notation (JSON). [8].

Semi-structured data consists of both structured data and unstructured data [3]. It can be seen as a variant of structured data, sharing some characteristics with structured data while not adhering to traditional database models. An instance of this is represented by Common Separated Value (CSV) files [7]. Semi-structured data encompasses formats like emails, XML, and JSON. When dealing with semi-structured data, traditional relational databases are not suitable. Instead, this type of data is represented using concepts like edges, labels, and tree structures [17]. Due to the absence of a predetermined schema definition, semi-structured data offers significant scalability and flexibility [6].

Every day, countless structured, semi-structured, and unstructured documents are generated worldwide [22]. Digital data generated by various digital platforms and devices are growing at an astounding rate globally [15]. To illustrate this increase, one can point to the statistics from the German platform statista.com, specialized in data collection and visualization, which shows that the amount of produced information has grown ninefold from 2013 to 2023. It is anticipated that by 2025, this daily volume will surge to 181 zettabytes [20]. This data growth trend is illustrated in Fig. 1.



Source: statista.com

Figure 1. Data growth trend

2. Approaches for analyzing unstructured and semi-structured data.

Unstructured and semi-structured data pose challenges for professionals in the field of information technology because they often require more time than initially anticipated for structuring and subsequent preparation for analytical processing. These types of data on the Web come in very large volumes. This implies the application of specific approaches when working with them. This document will draw a comparison between unstructured and semi-structured data, examining the most popular approaches for processing these types of data.

In the context of processing unstructured data, various approaches have been formulated in literature. These approaches include, namely:

• Speech Recognition. Speech recognition is the procedure of converting spoken language into a sequence of words using computer programs or algorithms. This field is a fascinating aspect of signal processing [5];

• Machine Translation. Machine translation, often abbreviated as MT (not to be confused with computer-aided translation, machine-aided human translation, or interactive translation), is the automated translation of text from one natural language, like English, to another, such as Ibo, by computer software [16].

Analyzing semi-structured data demands tailored strategies due to their adaptable and variable nature. Popular approaches for analysis and processing of this type of data are:

• XML parsing. XML parsing involves the action of reading an XML document and offering a user application interface to access that document. An XML parser is a software tool that carries out this function. Furthermore, most XML parsers assess the document's proper structure, and many can even validate it against a Document Type Definition (DTD) or an XML schema. By utilizing the parsing interface, the user application can direct its attention to the core application logic without becoming entangled in the intricate specifics of XML [13].

• NoSQL Databases. Relational databases are databases that can be scaled vertically. To manage extensive quantities of semi-structured and unstructured data, NoSQL databases are used. NoSQL databases adhere to the BASE principle, which stands for "Basically, Available, Soft state, and Eventually consistent." Relational databases excel in data integrity, security, and reliable transactions. In contrast, NoSQL databases are suitable for handling large volumes of data in various formats. NoSQL databases manage big data with lower associated costs and minimal overhead. The scalability of NoSQL databases is achieved horizontally by simply adding new servers in a clustered environment. Commodity hardware is utilized for storing substantial data within the cluster [10].

NoSQL databases can be categorized based on the data model used. There are databases of the Document store type. They utilize formats like XML, JSON, BSON, and more [12]. The data within them is semi-structured and comprises pairs of attribute name-value. Retrieving data involves searching for both key values and attribute values. They are well-suited for storing text, XML documents, and other semi-structured data. Prominent examples of this category include MongoDB, Amazon Dynamo DB, Couchbase, and CouchDB [18].

• Web parsing. Web parsing involves the automated retrieval of extensive publicly accessible data from websites and web-based information systems. It streamlines the process of collecting data and transforms the acquired information into diverse formats like HTML, CSV, Excel, JSON, and text. This procedure typically encompasses three main phases: parsing HTML pages, extracting data, and storing it [2]. However, some authors suggest expanding the phases, emphasizing data analysis within the web scrapping process [14]. In the realm of data analysis, proficient programming is crucial. Consequently, many businesses seek to employ developers well-versed in web scraping techniques [2].

There are approaches that are applied to both types of data:

• Natural Language Processing. Natural Language Processing involves various techniques based on linguistic features. It analyzes a given text syntactically using formal grammar rules, and the resulting data is semantically interpreted to extract information from it. Natural Language Processing can be applied deeply, where each part of each sentence is analyzed, and attempts are made to interpret it, or it can be applied superficially by conducting limited semantic analysis on only some passage**s or phrases in sentences. Natural Language Processing allows for techniques to distinguish specific words or word parts. This approach is used by a significant portion of modern information systems designed for automatic text translation [9]. Natural language processing is used to transform semi-structured and unstructured data into structured data;

• Knowledge Discovery in Text. Text mining, also known as Knowledge Discovery from Text (KDT), involves extracting information and identifying patterns from unstructured data. It's particularly useful for analyzing semi-structured or unstructured datasets, such as emails, full-text documents, and HTML files. Text mining, also referred to as Knowledge Discovery from Text (KDT), entails uncovering valuable patterns within extensive text databases to gain knowledge. Text mining applies not only data mining's analytical techniques but also incorporates methods from natural language and information retrieval [21]. The approach of Knowledge Discovery in Text primarily utilizes analytical techniques related to information extraction and machine self-learning. The goal of this approach is to extract models for processing a large number of textual documents. This approach is typically used by contemporary search engines. The Knowledge Discovery in Text approach involves various activities, namely: automatic text classification according to a defined set of categories; grouping texts based on common features; automatic summarization; extracting topics from texts and analyzing thematic trends in text flows [9];

According to the literature in the field of information analysis, Natural Language Processing and Knowledge Discovery in Text are the two main approaches in the context of representing textual data in a structured form used in practice [4];

• Data visualization. Data visualization is a method used to illustrate intricate data through graphical means, making it easier to comprehend. When dealing with structured data, traditional graphical representations are straightforward. However, for unstructured or semi-structured data, achieving effective real-time visualization can be challenging due to their diverse nature [19].

The approaches for analyzing semi-structured and unstructured data are illustrated in Fig. 2.

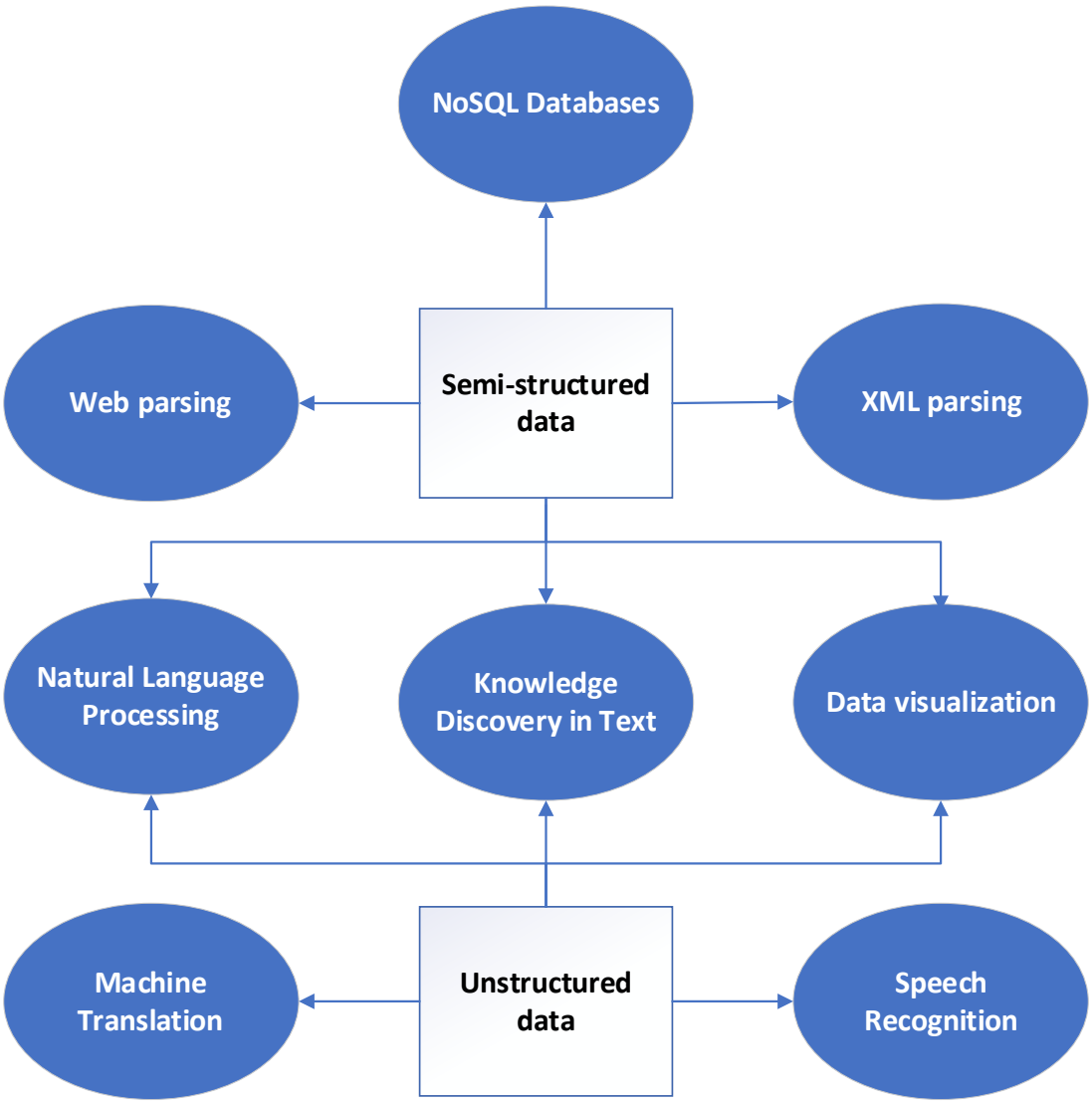


Figure 2. Approaches for analyzing semi-structured and unstructured data

Conclusion

In conclusion from the presented research of the significance and comparability of unstructured and semi-structured data in the modern web, we can summarize the following notes:

- Unstructured and semi-structured data are present in the modern web on a large scale and are growing rapidly.
- Unstructured data is divided into two groups - textual, consisting of text, and non-textual, representing images, video, and audio files.
- Semi-structured data consists of structured and unstructured data. Popular formats for semi-structured data are XML and JSON.
- Popular approaches for analyzing and processing unstructured and semi-structured data include Natural Language Processing, Knowledge Discovery in Text, and Data Visualization. Processing approaches characteristic of unstructured data include Machine Translation and Speech Recognition. Processing approaches characteristic of semi-structured data include Web Parsing, XML Parsing, and NoSQL Databases.

References

1. Adnan. K., Akbar, R. (2019). An analytical study of information extraction from unstructured and multidimensional big data. Journal of Big Data 6, Article number 91, 2019.
2. Britvin, A., Alrawashdeh, J., Tkachuck, R. (2022). Client-server system for parsing data from web pages. Advances in cyber-physical systems, Vol. 7, No. 1, 2022, ISSN: 2524-0382.
3. Chowdhury, A. (2020). Computer & Information Systems into Next Frontier: The Emergence. New Delhi Publishers, 2020.
4. Das, T., Kumar, P. (2013). BIG Data Analytics: A Framework for Unstructured Data Analysis. International Journal of Engineering and Technology (IJET), Vol 5 No 1 Feb-Mar 2013, India, ISSN : 0975-4024.
5. Gaikwad, S, Gawali, B., Yannawar, P. (2010). A Review on Speech Recognition Technique. International Journal of Computer Applications, Volume 10, No. 3, November 2010, ISSN 0975-8887.
6. Halcu, I., Grigoriu, O., Neculoiu, G., Sandulescu, V., Marinescu, M., Marinescu, V. (2013). Converting unstructured and semi-structured data into knowledge. Roedunet International Conference, 2013, ISSN: 2068-1038.
7. Jiang, S., Nocera, A., Tatar, C., Yoder, M., Chao, J., Wiedemann, K., Finzer, W., Rosé, C. (2022). An empirical analysis of high school students' practices of modelling with unstructured data. British Journal of Educational Technology, Volume 53, issue 5, 2022, ISSN 1467-8535.
8. Kanimozhi, K. and Venkatesan, M. (2015). Unstructured Data Analysis-A Survey. International Journal of Advanced Research in Computer and Communication Engineering, Vol. 4, Issue 3, 2015.
9. Kao, A. and Poteet, S. (2005). Text Mining and Natural Language Processing – Introduction for the Special Issue. SIGKDD Explorations, Volume 7, Issue 1, 2005.
10. Khan, W., Ahmed, E., Shahzad, W. (2017). Predictive Performance Comparison Analysis of Relational & NoSQL Graph Databases. International Journal of Advanced Computer Science and Applications, Vol. 8, No. 5, 2017, ISSN 2156-5570.
11. Khder, M. (2021). Web Scraping or Web Crawling : State of Art, Techniques, Approaches and Application. International journal of advances in soft computing and its applications, Vol. 13, No. 3, 2021, ISNN 2074-8523.
12. Kovacheva, M. (2021). Classification, Comparison and Criteria for Choosing NoSQL Databases. ICAICTSEE-2021.
13. Li, C. (2009). XML Parsing, SAX/DOM. Encyclopedia of Database Systems, 2009, ISBN 978-0-387-39940-9
14. Milev, P. (2017). Conceptual Approach for Development of Web Scraping Application for Tracking Information. Economic Alternatives, (3), 475-485.
15. Nti, I., Quarcoo, J., Aning, J., Fosu, G. (2022). A Mini-Review of Machine Learning in Big Data Analytics: Applications, Challenges, and Prospects. Big Data Mining and Analytics, Volume 5, Number 2, June 2022, ISSN 2096-0654.
16. Okpor, M. (2014). Machine Translation Approaches: Issues and Challenges. International Journal of Computer Science Issues, Vol. 11, Issue 5, No 2, September 2014, ISSN 1694-0784.
17. Praveen, S., Chandra, U. (2017). Influence of Structured, Semi-Structured, Unstructured data on various data models. International Journal of Scientific & Engineering Research Volume 8, Issue 12, December 2017, ISSN 2229-5518.
18. Radoev, M. (2017). A Comparison between Characteristics of NoSQL Databases and Traditional Databases. Computer Science and Information Technology, November 2017, ISSN 2331-6071.

19. Sandhu, A. (2022). Big Data with Cloud Computing: Discussions and Challenges. Big Data Mining and Analytics, Volume 5, Number 1, March 2022, ISSN 2096-0654.
20. Statista Research Department. (2021). Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025 [Infographic]. Statista. https://www.statista.com/statistics/871513/worldwide-data-created/
21. Yehia, A., Ibrahim, L., Abulkhair, M. (2016). Text Mining and Knowledge Discovery from Big Data: Challenges and Promise. International Journal of Computer Science Issues, Volume 13, Issue 3, May 2016, ISSN 1694-0784.
22. Zaman, G., Mahdin, H., Hussain, K., Rahman,j A. (2020). Information extraction from semi and unstructured data sources: a systematic literature review. ICIC Express Letters, Volume 14, Number 6, June 2020, ISSN 1881-803X.