THE 16TH EDITION OF THE INTERNATIONAL CONFERENCE
EUROPEAN INTEGRATION
REALITIES AND PERSPECTIVES

# Towards Useful Information from Unstructured Data Mining

**Adina Barila**[1], **Mirela Danubianu**[2], **Corneliu Turcu**[3]

**Abstract**: Continous development of web, social networking, mobile computing and Internet of Things in the last decades has led to a dramatic increase in the volume of data generated and, consequently, to the need for new ways of storing and analyzing of these data. In addition, these data no longer have a structured format but are rather in a semistructured or unstructured format. They contain useful information that companies need to explore. So, the methods of mining structured data had to be developed to mining the semistructured and also the unstructured data. In this paper we describe some of the most fundamental mining tasks.

**Keywords:** structured data; unstructured data; data mining; text mining; opinion mining; web mining

## 1. Introduction

The contemporany world is characterized by rapid advances in development and popularity of information and communication technology. Every year more and more people use computers and mobile devices not only in their professional activity but in all areas of their life. Data has been originally generated by employees but nowadays it has expanded to user-generated and machine-generated level. It results huge amounts of data that need to be stored ana analyzed. If a few decades ago the data to be analyzed were in a structured format, the new generated data have no longer a standard format. They come in the form of text, emails, posts on social networks, images, videos, web-clicks, phone calls, transactions, facial or gestural cues, senzor-generated data and other forms. The amount of unstructured data increased more rapidly then that of structured data. Moreover, in the recent pandemic context the most of activities had to become digital. The collecting and analyzing of unstructurated date may provide organizations new information about their business and can help them develop their competitive advantage and productivity.

In order to explore the huge quantities of data stored in databases and datawarehouses data mining techniques were developed. The methods of mining structured data developed to mining the semistructured and also the unstructured data.

[1] Ștefan cel Mare University of Suceava, Address: Strada Universității 13, Suceava 720229, Romania, Corresponding author: adina@eed.usv.ro.
[2] Integrated Center for Research, Development and Innovation in Advanced Materials, Nanotechnologies, and Distributed Systems for Fabrication and Control (MANSiD), Ștefan cel Mare University, Romania, Address: Strada Universității 13, Suceava 720229, Romania, E-mail: mdanub@eed.usv.ro.
[3] Ștefan cel Mare University of Suceava, Romania, Address: Strada Universității 13, Suceava 720229, Romania, E-mail: cturcu@usm.ro.

This paper aims to present some of the most fundamental mining tasks. Section 2 defines the main characteristics of structured, semistructured and unstructured data. Section 3 introduces data mining tasks. Section 4 presents text mining tasks and make a short comparison between them.

## 2. Structured, Semistructured and Unstructured Data

Structured data have a defined format and lengths, a high degree of organization and are easy to store and analyze. These data are numbers or strings and represent results of repetitive operations (Vespan, 2014). The most common forms of structured data are relational databases. Such data have a schema that defines their attributes and types, constraints on values, and relationships to other tables and attributes. This way of organizing facilitates the query of data (using SQL – Structured Query Language) for extracting and analyzing information. Also, the data in the spreadsheets can be considered structured data.

Semistructured data are data whose structure can change rapidly or unpredictably (Eberendu, 2016). These data don't have a tabular form but still contain labels or markers to separate the semantic elements and to impose hierarchies of records and data fields (Hänig, Schierle, & Trabold, 2010).

Unstructured data don't have a strictly defined format, structure or repetability. Unstructured data are "human information" (Syed, Gillela & Venugopal, 2013) and usually include text documents, images, emails, phone calls, website clicks, IoT sensor data, satellite imagery and other types of data that cannot be tabulated (Feldman & Sanger, 2007).

Many of data mentioned as unstructured could have attributes that make them semistructured. For example, the body of an email is unstructured text, but the header contains data like the name of the sender, the subject that provider a certain structure. Also, posts on a social network contain text in free form or images but include the name of persons who have done.

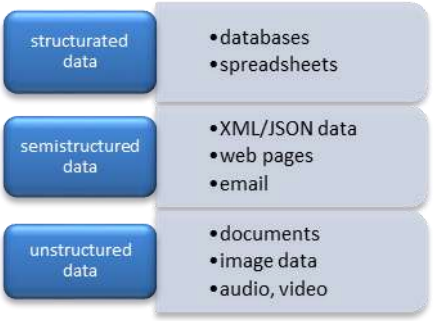


**Figure 4. Types of data**

## 3. Data Mining

Data mining is a non-trivial process of extraction of information which is hidden, previously unknown and is potentially useful, from large databases (Agarwal, 2013). Data mining aims at finding information which are not explicitly available in the database, information that cannot be provided by queries and reports. By applying data mining algorithms two major categories of problems can be solved: prediction and description. The main predictive data mining tasks are classification and regression. Descriptive tasks include deviation detection, cluster analysis, and exploitation of association rules (Danubianu & Barila, 2014).

Classification is two-step process whose goal is to predict the nature of a data based on predetermined classes of data. In the first step, a model defined by the available training dataset is constructed. In the second step, the model is used to assign the class label to subsequent records.

Regression is a task which aims at predicting a range of numeric values given a certain dataset. It analyses the relationship between the dependent (predicted) variable and the independent (predictor) variables. The liniar regresssion and the logistic regression are the most used types for this task.

Deviation detection consists of detecting significant changes in the norm. This task has many similarities with statistical analysis.

Cluster analysis aims at automatically grouping data into several groups so that the samples in a group are similar one to other and there are consistent differences between groups (Danubianu & Barila, 2014). Different measures may be used in determining similarities and differences. The result of cluster analysis is a number of groups (clusters) that form a partition or partition structure of the data set. If each data belongs to a cluster completely or not, it called hard clustering. In soft clustering (or fuzzy clustering) each data belongs to each cluster to a certain degree.

Exploitation of association rules finds interesting association or correlation relationships between a large set of data items. Given a set of transactions, this task identifies association between occurrence of an item based on the occurrences of other items in the transaction. A typical example of the exploitation of association rules is the analysis of the market basket.

## 4. Unstructured Data Mining

According to experts estimates, 80% of companies' information is in unstructured formats in the form of emails, contract documents, notes, reports. Therefore, there is more data stored in unstructured format form than data stored in structured format. In other words, unstructured data dominates modern (commercial) bussiness data. Thus, the need arose to identify effective solutions for exploring and capitalizing on this data.

### 4.1. Text Mining Tasks

The term text mining (or text data mining TDM, or knowledge discovery from text – KDT) refers to the process of extracting quality information from unstructured text data (Vijayarani, & Janani, 2016). It is used to find new, previously unknown information from different unstructured data. The specific tasks of text mining are: information retrieval, information extraction, summarize text, categorizing text and clustering.

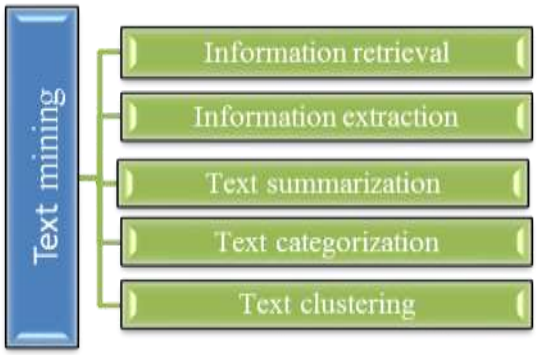


**Figure 5. Specific Tasks of Text Mining**

Information retrieval (IR) is the finding and recognition of information from a collection of unstructured data sets. IR is focusing on facilitating access to information, rather than analyzing information and finding hidden patterns – the main goal of text mining (Allahyari et al., 2017). The most popular IR systems are search engines that identify information or documents relevant to a word data set. Search engine includes site, desktop, enterprise, mobile, social and web searches (Abass & Arowolo, 2017). Desktop search tools search within a user's own computer files. Web search engine is designed to search information, including text documents, image or video files, on World Wide Web while site search tools focuses on the content of a given website. Social search engine searches user-generated content on social media and mobile search enable users to search mobile content available on mobile devices. Enterprise search engine finds information from enterprise-type sources like databases, intranets, emails, documents, file systems.

An extension of IR is document retrieval. This is the process of searching and identifying the possible document that contains the information sought by the user. This system is used by universities, libraries, government and companies to provide access to articles, books, magazines and other documents.

Information Extraction (IE) is the process of automatic extraction of specific information from unstructured or semistructured natural language. This task recognizes entities such as names of people, organization, geographic location, and finds relationships between entities. The extracted information is stored in databases as templates and is available for later use. In general, this activity is based on natural language processing. The IE main tasks include Named Entity Recognition (NER), Coreference Resolution (CO), Relation Extraction (RE) and Event Extraction (EE) (Golshan, Dashti, Azizi, & Safari, 2018) The goal of NER is detection and classification of types of entities, such as person names, organizations, locations, time expressions, currency expressions, quantities, percentages (Golshan, Dashti, Azizi, & Safari, 2018). The performance of this task is affected by certain factors like language, textual genres or domain, types of entities. The modern systems use machine learning based algorithms (Goyal, Gupta & Kumar, 2018). CO is the task of resolving all mentions in a document that refer to the same real world entity (Stylianou & Vlahavas, 2020). It is one of the most difficult tasks in natural language understanding. RE refers to identifying the relationship between named entities in the text and deciding which ones are meaningful for the concrete application or problem. For this task there are used knowledge-based, supervised and self-supervised methods (Konstantinova, 2014). The goal of Event Extraction is to find the existence of an event in text and event-related information such as the "5W1H" about an event (i.e., who, when, where, what, why and how). There are two types of EE tasks: closed-domain and open-domain. In closed-domain EE a particular event structure is predefined, which includes not only event types but also event arguments' roles (Wei & Wang, 2019). Besides detecting event, this task must extract words or phrases to fill in the given event structure. The

open-domain EE task finds an event and extracts keywords related to it. Sometimes, this task clusters similar events.

IE is used for analyzing, filtering and routing emails. IE is used to detect blocks of text (header, signature) to extract and manage information about people (profile, contact).

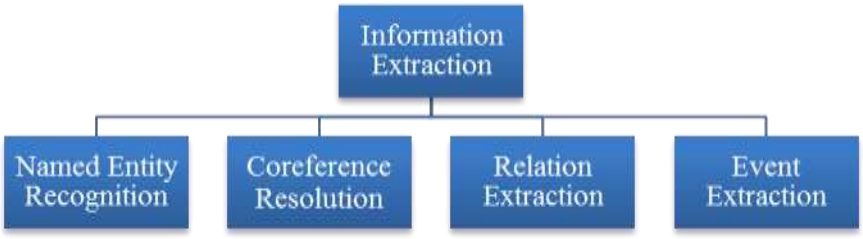


**Figure 6. Information Extraction Tasks**

Text summarization is the process of automatically reducing a text document to create a smaller document that contains the most important aspects of the original document. The summary can determine the relevance of a document. The summarization is generaly a difficult task because the text must be characterized as a whole and its important content must be retained. There are two types of summarization methods: extractive and abstractive. Extractive summarization methods identify important sections of the text and use them to generate the summary. In other words, the summary is generated by extracting sentences from the original text. Abstractive summarization methods examine whole text using advanced natural language techniques and generate a new text containing important information from the original text. At present, automatic extractive methods give better results compared to automatic abstractive methods. In fact, there is no completely abstractive summarization system today (Allahyari, Pouriyeh, et al., 2017).

There are also two types of summarize systems: based on a single source document and based on multiple source documents.

The summary is useful for determining the relevance of a document in a foreign language or for preparing information for use on small mobile devices.

Text categorization is the process by which a category is assigned to a document. In other words, this process (also known as text classification, or topic spotting) sorts a set of documents into categories from a predefined set (Sebastiani, 2003). Text categorization is considered to be a supervised classification technique as a set of pre-classified documents is provided as a training set. There are three types of text categorizing methods: conventional methods, fuzzy logic-based methods, deep learning-based methods (Dhar, Mukherjee, Dash, & Roy, 2021).

Categorization is useful in applications where it is necessary to organize documents, e.g. classification of websites, automatic indexing of scientific articles.

Text clustering is a process of forming groups (clusters) of similar objects from a set of entries. Objects belonging to the same cluster are similar to each other, while objects from two different clusters are different.

Clustering is an unsupervised process by which objects are classified into groups without any prior information.

There are several categories of clustering algorithms such as hierarchical clustering, partitioned clustering, density-based algorithm, self-organizing maps algorithm.

In business, clustering can be used to segment customers into groups for additional analysis and marketing activities.

**Table 2. Comparison between Text Mining Tasks**

| Task | Characteristics | Applications |
|---|---|---|
| Information Retrieval | Finding information from a collection of unstrutured datasets | search engines, searching in digital libraries, access to books, paper retrieval, document retrieval |
| Information Extraction | Extraction of specific information | filtering and routing emails, opinion mining |
| Summarization | Automatically reducing a text document by keeping the meaning | media monitoring, book, customer support, small mobile devices |
| Categorization | Sorting a set of documents into categories from a predefined set | automatic indexing of scientific paper, spam filtering, automatic essay grading, sentiment classification |
| Clustering | Sorting a set of documents into categories without a predefined set | document organization, customer segmentation, web mining |

## 5. Conclusions

The paper presented the most important tasks for mining unstructured data. We focused on text data but useful information can be found applying mining techniques to image, video or other data. In context of continous growing of amount of unstructured data, the organizations need to find and apply methods for extracting useful information from those data. Unstructured data will not replace structured data, but will provide acces to new information that is not available in structured data.

## 6. Acknowledgement

"This work is supported by the project *ANTREPRENORDOC*, in the framework of Human Resources Development Operational Programme 2014-2020, financed from the European Social Fund under the contract number 36355/23.05.2019 HRD OP /380/6/13 – SMIS Code: 123847."

## References

Vespan, D. (2014). *Extragerea cunoştinţelor din documentele electronice/Extracting knowledge from electronic documents*. Bucharest: Pro Universitaria.

Eberendu, A. (2016). Unstructured Data: an overview of the data of Big Data. *International Journal of Computer Trends and Technology (IJCTT),* Vol. 38, No. 1, pp. 46-59.

Hänig, C.; Schierle, M. & Trabold, D. (2010). Comparison of structured vs. unstructured data for industrial quality analysis. *Proceedings of The World Congress on Engineering and Computer Science*.

Syed, A.R.; Gillela, K.; Venugopal, C. (2013). The Future Revolution on Big Data, *International Journal of Advanced Research in Computer and Communication Engineering (IJRACCE)*, Vol. 2, Issue 6, 2446-2451, June 2013.

Feldman, R., Sanger, J. (2007). The text mining handbook: advanced approaches in analyzing unstructured data. *Cambridge University Press.*

Agarwal, S. (2013). Data mining: Data mining concepts and techniques. *Proceedings of 2013 International Conference on Machine Intelligence and Research Advancement*, pp. 203-207, Katra, India, December 2013.

Danubianu, M. & Barila, A. (2014). Big Data vs. Data Mining for Social Media Analytics. *International Conference on Social Media in Academia - Research and Teaching- SMART2014*.

Vijayarani, S. & Janani, R. (2016). Text Mining: Open Source Tokenization Tools – an Ananalysis. *Advanced Computational Intelligence: An International Journal (ACII)*, Vol. 3, No. 1, pp. 37-47.

Allahyari, M.; Pouriyeh, S.; Assefi, M.; Safaei, S.; Trippe, E.; Gutierrez, J. & Kochut, K. (2017). *A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques*.

Abass, O. & Arowolo, O. (2017). Information Retrieval Models, Techniques and Applications. *International Research Journal of Advanced Engineering and Science*, Vol. 2, Issue 2, pp. 197-202.

Golshan, P. N.; Dashti, H. R.; Azizi, S. & Safari, L. (2018). A Study of Recent Contributions on Information Extraction. *Proceedings of 4th National Conference on Distributed Computing and Big Data Processing*, pp. 780-785.

Goyal, A.; Gupta, V. & Kumar, M. (2018). Recent Named Entity Recognition and Classification techniques: A systematic review. *Computer Science Review*, Vol. 29, pp. 21-43.

Stylianou, N. & Vlahavas, I. (2020). *A Neural Entity Coreference Resolution Review*. arXiv:1910.09329v2.

Konstantinova, N. (2014). Review of Relation Extraction Methods: What Is New Out There?. *Analysis of Images, Social Networks and Texts. AIST 2014. Communications in Computer and Information Science*, Vol. 436, Springer, Cham.

Wei, X. & Wang, B. (2019). A Survey of Event Extraction from Text. *IEEE Access*. Vol.7, pp. 173111-173137.

Allahyari, M.; Pouriyeh, S.; Assefi, M.; Safaei, S.; Trippe, E.; Gutierrez, J. & Kochut, K. (2017). *Text Summarization Techniques: A Brief Survey.* arXiv preprint arXiv:1707.02268v3.

Sebastiani, F. (2003). *Text Categorization*.

Dhar, A.; Mukherjee, H.; Dash, N. S. & Roy, K. (2021). Text categorization: past and present. *Artificial Intelligence Review*, Vol. 54, pp. 3007–3054.