Knowledge and Information Systems
https://doi.org/10.1007/s10115-024-02139-x

REVIEW

Methods for concept analysis and multi-relational data mining: a systematic literature review

Nicolás Leutwyler[1,2,3,4] · Mario Lezoche[1] · Chiara Franciosi[1] · Hervé Panetto[1] · Laurent Teste[4] · Diego Torres[2,3]

Received: 15 June 2023 / Revised: 14 February 2024 / Accepted: 3 May 2024

Abstract
The Internet of Things massive adoption in many industrial areas in addition to the requirement of modern services is posing huge challenges to the field of data mining. Moreover, the semantic interoperability of systems and enterprises requires to operate between many different formats such as ontologies, knowledge graphs, or relational databases, as well as different contexts such as static, dynamic, or real time. Consequently, supporting this semantic interoperability requires a wide range of knowledge discovery methods with different capabilities that answer to the context of *distributed architectures* (DA). However, to the best of our knowledge there is no general review in recent time about the state of the art of Concept Analysis (CA) and multi-relational data mining (MRDM) methods regarding knowledge discovery in DA considering semantic interoperability. In this work, a systematic literature review on CA and MRDM is conducted, providing a discussion on the characteristics they have according to the papers reviewed, supported by a clusterization technique based on association rules. Moreover, the review allowed the identification of three research gaps toward a more scalable set of methods in the context of DA and heterogeneous sources.

✉ Nicolás Leutwyler
nicolas.leutwyler@univ-lorraine.fr

Mario Lezoche
mario.lezoche@univ-lorraine.fr

Chiara Franciosi
chiara.franciosi@univ-lorraine.fr

Hervé Panetto
herve.panetto@univ-lorraine.fr

Laurent Teste
laurent.teste@snmsf.com

Diego Torres
diego.torres@lifia.info.unlp.edu.ar

1 University of Lorraine, CNRS, CRAN, 54000 Nancy, France

2 LIFIA, CICPBA-Facultad de Informática, UNLP, 1900 La Plata, Buenos Aires, Argentina

3 Dto. CyT, UNQ, 1876 Bernal, Buenos Aires, Argentina

4 SNMSF, 38240 Meylan, France

Published online: 30 May 2024

Keywords Knowledge extraction · Knowledge discovery · Concept analysis · Multi-relational · Semantic interoperability

1 Introduction

The Industry 4.0, defined in [51], is impelling governments and enterprises to adopt practices such as the exploitation of data in order to optimize their processes. Additionally, the growth of Internet of Things (IoT) over the last years raised new challenges related with the large amounts of heterogeneous data it produces. Thus, on the one hand, Information Retrieval is an essential step to consider not only in software development, but also in all the stages of industrial procedures. On the other hand, sharing *knowledge* is important throughout different pieces of software, and enterprises, i.e., semantic interoperability [55, 79]. Consequently, there are many methods for the extraction of knowledge (e.g., K-means, Formal Concept Analysis, Regression Trees, etc.) working in different environments and having different purposes. For example, some of them are *supervised* (the data is previously labeled), others *unsupervised* (the data is unlabeled), and even others can be used in both setups. Moreover, some methods are only suitable for working with single tables, leaving room for multi-relational data mining (MRDM) methods [21], i.e., the ones used when the goal is to extract relationships, from *heterogeneous* and *linked* sources, stored in different tables.

In this context, *knowledge extraction* has become a widely studied topic in the academy, leading to a great number of papers being written about it each year. Moreover, on top of being supervised or non-supervised, these methods have particular characteristics that differentiate them, for instance, being utilized in particular niches, receiving different types of inputs, or producing different types of outputs, i.e., typically several formats of knowledge representation. This heterogeneity can be seen as a desirable characteristic because it gives to the practitioners a rich amount of options when dealing with different kinds of problems. However, not all knowledge extraction methods are suitable for extracting knowledge directly from distributed architecture (DA) environments such as IoT. This is because some of them are designed to work with data stored in a single node, while others are simply not prepared to deal with the typical amounts of data produced by these architectures, e.g., because the output they produce is exponentially bigger than the input they receive. Examples of such methods are Formal Concept Analysis (FCA), Relational Concept Analysis (RCA), Polyadic Concept Analysis (PCA), whose goals are to extract knowledge in the form of hierarchies and relationships between the concepts. Moreover, to the best of our knowledge, while there are some reviews distinguishing data mining methods between their characteristics [24, 48], none of them contrast the aforementioned differences in the fields of IoT, DA and semantic interoperability. Considering this, the objective of this review is to provide evidence on the characteristics of the main methods in these fields in the current literature. Doing so is an important step for the discovery of scientific gaps on the possible improvements and directions toward a more scalable *conceptual knowledge* extraction method.

A systematic literature review (SLR) on methods for knowledge extraction that are based either in Concept Analysis (CA) or in MRDM was conducted for the objective mentioned in the previous paragraph. The purpose is to present an overview of research works, findings, and relations between them in fundamental research and practice. This is necessary to

(1) explore the key features of CA and MRDM methods in different domains,
(2) identify the differences and points in common between MRDM CA-based and non-CA-based methods,

(3) and determine the areas that are commonly covered in both fields and the ones that still require attention.

Additionally, doing so in a systematic fashion provides several advantages, such as a clear way of understanding the set of papers included, and most importantly the *reproducibility*. Thereafter, in order to maximize the discovering of relations between the different metrics gathered during the review, the findings are analyzed using FCA and association rules. Finally, we discuss the characteristics of the methods found, and their applicability in our field of interest, which is the area of DA.

The paper is structured as follows: In Sect. 2, background concepts and related works are presented. In Sect. 3, the followed methodology is introduced in order to provide each step to reproduce the review. In Sect. 4, the results of the applied methodology are detailed. In Sect. 5, the given results are contrasted by extracting association rules using FCA in order to understand the relations between them, while three research gaps found during the assessment of articles are presented. Finally, in Sect. 6 the conclusions and future work are discussed.

2 Background and related work

In this section, firstly, the notions of CA, MRDM, and DA are introduced, as they are the keystone concepts of the discussion. Secondly, the preliminaries of FCA and RCA are presented as examples of CA and MRDM, respectively. Finally, a summary of the articles reviewing similar topics is given.

2.1 Concept analysis

Overall, CA methods aim to extract a set of concepts which can be understood as natural clusters of instances sharing certain properties. These concepts could be partially ordered in a hierarchy that resembles the natural way people think of hierarchies: One concept is a subconcept of the other if it has all its properties and adds some others [91]. Section 2.4 presents this notion more formally.

In this article, CA refers to the ensemble of FCA and all its extensions (e.g., Fuzzy FCA [58], PCA [30], RCA [74]). Its basic notions are those of *formal context* and *formal concept* (detailed in Sect. 2.4). As explained in [28], the word "formal" is meant to highlight the mathematical aspects the framework adopts. Nevertheless, for convenience, they are sometimes referred to as *context* and *concept*, respectively. Furthermore, it can be argued that other frameworks such as principal concept analysis also deal with mathematical aspects, without being called *formal* PCA (not to be confused with Polyadic Concept Analysis, that uses the same acronym). Hence, in the literature, the word "formal" is frequently left aside when it is obvious from the context. For instance, when speaking about the execution of an algorithm that computes a set of *formal concepts*, FCA researchers often refer to them simply as *concepts*.

2.2 Multi-relational data mining

Data mining methods search for interesting patterns in data. Several of these approaches look for patterns in a single data table (or a single type of source). Relational data mining methods look for patterns from multiple tables related to each other from a relational database [21].

MRDM approaches, in this article, will reference the approaches that look for patterns from multiple tables *or* from (heterogeneous) sources related to each other. This concept is relevant because, on the one hand, semantics could be lost in the process of going from multiple tables to only one, and on the other hand, even if it could be done without losing semantics in some cases, usually it is computationally too expensive. Thus, there is the need for methods to mine data from multiple and heterogeneous sources without the necessity of transforming them to only one table with a single relationship type.

2.3 Distributed architectures

In this article, we consider DA to be a collection of autonomous computing elements that appears to its users as a single coherent system [86]. In addition, a DA is often referred to as a distributed system [15, 16], which, in turn, consists of multiple software components that could be on multiple computers, but run as a single system. These computers can either be located in close proximity, connected through a local network, or they can be situated far apart, linked by a wide area network. Distributed systems can take on a variety of configurations, including mainframes, personal computers, workstations, minicomputers, and more. The objective of these architectures is to enable such a network to function as if it were a single, unified computer. Certain systems, such as those in the realm of IoT, inherently possess a distributed nature, while the distribution of others depends on their specific implementation. Examples of such systems are the ones using the MapReduce approach [32, 95] to compute algorithms by doing two operations: *map* (i.e., apply a certain function or transformation in each of the nodes), and *reduce* (i.e., combine the result of the operations previously performed). There are several advantages that DAs offer over centralized ones, such as (1) scalability and (2) redundancy, where (1) means that the system can be extended by incorporating additional machines as required. While (2) means that multiple machines can provide the same services, so if one is unavailable, work does not get interrupted. Additionally, because many smaller machines can be used, this redundancy does not need to be prohibitively expensive.[1]

The impact of DAs in data mining methods could be understood in two ways. Firstly,

(i) a data mining method might be implemented using a DA, allowing it to be more scalable or faster. Secondly,
(ii) it could be a method whose purpose is to perform mining over DAs, regardless of the method itself being implemented in a DA or not.

For instance, [95] implements two FCA algorithms using Twister Iterative *MapReduce* [22], leveraging its properties (*distributed* and iterative) in order to reduce computation time. The first one, called MRGanter, consists of applying a modified version of the NextClosure [28] algorithm to multiple nodes each having a partition of the input in order to maximize the parallelization. Nonetheless, the algorithms are not intended to process tuples as they arrive. In fact, they are meant to divide the *entire* input into several nodes to then start the distributed computation. Hence, although they certainly enhances the computation scalability, they are not designed to mine *from* DAs. For this reason, it would fall into (i).

Additionally, the work presented in [15] entails the implementation of a data mining method on a distributed real-time computation system (i.e., Apache Storm[2]) with the objective of big data stream analysis on smart cities. The presented system uses the stream abstraction

[1] IBM distributed computing definition.

[2] Apache storm website.

provided by Apache Storm, i.e., an unbounded sequence of tuples that is processed and created in parallel in a distributed fashion. Schemas are defined to associate fields in the tuples. A *spout* is a source of a stream in the system. Particularly, spouts read tuples from an external source and emit them to the system (i.e., topology in Apache Storm). The processing operations are executed by *bolts*, which include, for instance, filtering, executing arbitrary functions, aggregating, and so forth. Parallelism is obtained by configuring spouts and bolts to start *executors*, that can be thought as threads being able to run in parallel, each of which can process data by executing different *tasks*. The scalability is acquired by allowing to change the number of executors on run-time. Then, let us suppose several distributed sensors producing data represented by the tuple $\langle x_1, \ldots, x_k \rangle$, $k \in \mathbb{N}$. The tuples will be processed by a broker (e.g., Kafka[3]) and sent to the system, which then will be emitted by a spout with a certain structure depending on the area of the city where the tuple was produced. The tuple arrives to the first bolt that is parallelized depending on the city areas, and whose purpose is to create aggregations based on the specific domain, e.g., given two measures of traffic, consider only the maximum of them. And finally, the second bolt uses the *curated tuples* that continuously arrive to compute the output incrementally. Thus, it would fall into both (i) and (ii).

Concluding the examples, [87] presents an incremental conceptual data mining algorithm, i.e., the conceptual lattice (see Sect. 2.4) is computed as data arrives, instead of having to recompute it. The way the algorithm works is by defining an initial result $\mathcal{L}_0$ when no data has been received. Then, using the function "*add_intent*" to update the conceptual structure $\mathcal{L}_i$ to $\mathcal{L}_{i+1}$ as the i-th object arrives, $i \in \mathbb{N}$. Since each object has a finite amount of attributes, the invocations of "*add_intent*" can be understood as steps in the processing of a data stream of objects and their attributes. However, the presented algorithm is meant to be run in a single node, since it uses shared variables and loops that make it not trivially parallelizable. Therefore, it only falls into the category of (ii).

2.4 Formal concept analysis

FCA, introduced in [91], is a method for extracting knowledge from a dataset called *formal context*, i.e., a table consisting of objects, attributes, and relations between them showing whether an object has an attribute or not. Formally, a formal context $\mathcal{K}$ is a triple (G, M, I), where G is a set of objects, M is a set of attributes, and I is an incidence matrix where iIj if $g_i \in G$ has the attribute $m_j \in M$, and $i \not I j$ otherwise. Let $'$ be the derivation operation on a set of objects $X \subseteq G$ (dually, on a set of attributes $Y \subseteq M$) given by

$$X' = \{m \in M \mid \forall g \in X, gIm\}$$
$$Y' = \{g \in G \mid \forall m \in Y, gIm\}$$

A *formal concept* is a pair $C = (X, Y)$ where $X \subseteq G$, $Y \subseteq M$, $X' = Y$, and $Y' = X$. X is called the extent and Y the intent. The set of all formal concepts and the relation of inclusion of extents form the so-called *concept lattice*, which is a partially ordered set, and is usually noted with the letter $\mathcal{L}$.

It has been successfully used in an environment of DA in [15], but in conjunction with the extensions called Fuzzy Formal Concept Analysis (i.e., an extension of FCA [58] where the incidence matrix is a relation $I \subseteq G \times M \rightarrow [0, 1]$) and Temporal Concept Analysis

[3] Kafka website.

[92]. This work takes advantage of the fuzzy and temporal characteristics of the extensions in order to deal with the exponential growth of the fuzzy lattice.

2.5 Relational concept analysis

While FCA aims to extract formal concepts from a formal context, Relational Concept Analysis (RCA), presented in [74], is used to extract relationships between different formal contexts (MRDM). The input of RCA is named Relational Context Family (RCF) and consists of a tuple (K, R) where K is a set of formal contexts and R is a set of binary relations between objects of the contexts. It has been successfully used in an environment of DA in [13] where the RCA module plays the role of an auxiliary tool.

2.6 Related work

To the best of our knowledge, there have been only two relevant reviews in recent time that discuss the subject of DM methods in DA. The first one is a review for data analytics in IoT applied to the Software Engineering (SWE) best practices [24]. In that review, the authors present a comprehensive systematic literature review by analyzing the current techniques and technologies used in IoT-based systems from the SWE and Big Data Analytics (BDA) perspectives at different domains. Additionally, they introduce a generic architecture for data analytics in IoT with six layers. More specifically, the defined layers are Data Manager, System Resource Controller, System Recovery Manager, SWE Handler, BDA Handler, and Security Manager. Roughly, the Data Manager layer is where the raw data is received from physical IoT objects sources, such as sensors, social media interactions, location-based services, or smart devices. The System Resource Controller is where the reliable and scalable processing environment for IoT data is ensured. The System Recovery Manager is the responsible for forecasting the amount of space needed for the growth of data in the IoT system. The SWE Handler is in charge of ensuring the overall system quality of service (QoS). The BDA Handler is where the data mining is performed, and thus is where the focus of this paper is. And the Security Manager is a transversal layer whose aim is to ensure the data security among the rest of the layers.

Particularly, the BDA handler is composed of four modules:

(1) data aggregation,
(2) data reduction,
(3) data analysis,
(4) data interpretation and visualization.

Regarding the IoT data challenges with respect to this module, they are implied to include the huge volume of different data models support, as well as the analytics latency, accuracy, interpretation, consistency and confidentiality, addressing the 10 Vs of big data [66].

The second one [48] is a systematic literature review on methods and techniques for big data stream analysis. In that review, authors aim to answer the following research questions:

(1) What are the tools and technologies employed for big data stream analysis?
(2) What methods and techniques are used in analyzing big data streams?
(3) What do these tools and technologies have in common and their differences in terms of concept, purpose and capabilities?
(4) What are the limitations and strengths of these tools and technologies?

(5) What are the evaluation techniques or benchmarks used for evaluating big data streaming tools and technology?.

Among these questions, the one most closely related to our work is (2). Furthermore, the results related to that question indicate that the partitioning clustering techniques (such as k-median, k-means, and k-medoid) are unsuitable for the purpose of analyzing big data streams because they require prior knowledge of clusters, which is usually not the case in that context. Moreover, due to concept drift inherent in social media streams, the other algorithms found not to be suitable for big data stream analysis are scalable graph partitioning algorithms because of their tendency toward balanced partitioning. In addition, the most used methods and techniques, according to the findings, are density-based clustering algorithms, threshold-based techniques, and hierarchical and incremental clustering methods. Finally, they found that social media stream processing still lacks research compared to the other big data stream analysis areas.

Both articles *only* analyze data mining methods in the areas of IoT and big data stream analysis, respectively. However, they overlook the methods that have not yet been applied in those areas, despite their potential applicability. Moreover, they do not consider semantic interoperability in their analysis. Hence, this review includes papers presenting improvements in FCA-based methods as an addition to the ones directly being applied in DA, and also considers the dimension of semantic interoperability.

3 Method

This work falls into the classification of Systematic Literature Review (SLR). An SLR aims to answer a very focused and specific question, identify relevant studies, evaluate their quality, and qualitatively (or quantitatively) summarize the findings [72]. The search carried out in this work is based on the guidelines introduced by Petersen et al. [67]. Generally, the methodology consists of defining the research questions that will guide the search. Use the keywords in the questions in order to create a query string that will define the set of articles to consider in the review. Select in which databases the search will take place. Search for articles using the defined query. And finally, analyze the results. One of the reasons why it is interesting to do an SLR is that it leaves evidence on how to replicate the search using the criteria mentioned in it.

Particularly, this work starts with the definition of the research questions 3.1 and then proceeds with the definition of keywords 3.2 that will guide the creation of *search strings*, to then define the databases to be used 3.3. Afterward, the search is performed by querying the selected databases using the previously defined *search strings* and a subset of the resulting articles is selected following specific criteria 3.4. Finally, three more steps are conducted on the final selection of articles: data extraction 3.5, analysis and classification 3.6, and validity evaluation 3.7.

3.1 Literature research questions

In this subsection, the research questions are defined. To guide this study, the main research question is the following: (RQ) What are the approaches on MRDM and CA for knowledge extraction used for semantic interoperability in DA?
Taking RQ into account, we derive the following specific literature-research questions:

- LRQ_1: When and where the articles have been published?

- LRQ_2: What are the methods utilized for MRDM?
- LRQ_3: What are the methods utilized for CA?
- LRQ_4: What are the methods utilized for DA?
- LRQ_5: What are the problematics the papers aim to solve?
- LRQ_6: In which domains the methods are applied?
- LRQ_7: What data format the methods allow extracting and from which ones?
- LRQ_8: What are the characteristics considered in the evaluation of the articles?
- LRQ_9: What are the data formats and methods used in each article for the semantic interoperability problem?

3.2 Keywords and search strings definition

3.2.1 Important concepts

The following concepts will be used toward the rest of the paper,

- *Data Mining* (DM) is the process of discovering knowledge or patterns from massive amounts of data [34].
- *Knowledge Discovery in Databases* (KDD) is an automatic, exploratory analysis and modeling of large data repositories. KDD is the organized process of identifying valid, novel, useful, and understandable patterns from large and complex data sets [57].
- *Knowledge Extraction* (KE) is the creation of knowledge from structured (e.g., relational databases, extensible markup language (XML)) and unstructured (e.g., text, documents, images) sources [84].
- *Information Retrieval* (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers) [59]. This definition differs to the Data Mining one in the sense that, in IR, the goal is to gather information from an unstructured source in order to *satisfy* some *known* need, e.g., browsing in google. Hence, it is important to notice that IR methods might use Data Mining algorithms in their process.
- *Concept Analysis* (CA), in this work, is referred as the ensemble of the FCA method, introduced by Wille [91], and its extensions, e.g., RCA [74], PCA [30].
- *Distributed Architecture* (DA) is a collection of autonomous computing elements that appears to its users as a single coherent system [86].

Considering these definitions, it is our understanding that DM, KDD, and KE are three names for the same concept. Moreover, IR is a concept related to the overall process that might use DM algorithms in its implementation, but also includes steps like data warehousing. Additionally, DA is a type of environment in which both DM and IR algorithms might be implemented in or not. To better represent this idea, Fig. 1 depicts our understanding of the relations between these concepts.

The keywords and the search strings were defined relying on Population, Intervention, Comparison, and Outcomes (PICO, defined in [39]):

- Population: In our context, the population are knowledge extraction or information retrieval studies using some form of CA (e.g., FCA, RCA, Fuzzy FCA) or applied in the MRDM context.
- Intervention: In this work, these are algorithms, methods and tools,

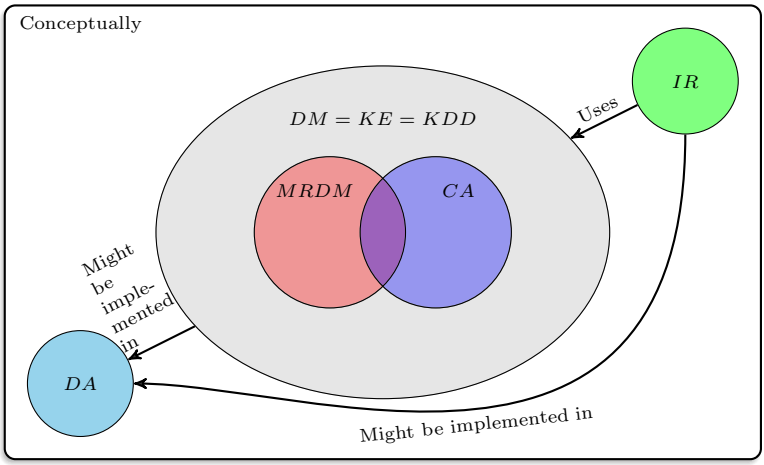


Fig. 1 Venn diagram from a conceptual point of view, showing the relations between *data mining* (DM), *multi-relational data mining* (MRDM), *knowledge extraction* (KE), Knowledge Discovery on Databases (KDD), Concept Analysis (referring to FCA and its extensions, and abbreviated CA), *information retrieval* (IR), and *distributed architectures* (DA)

- Comparison: In this study, we compare the different ways of dealing with the challenges of DA by identifying and analyzing the used algorithms and their applicability in semantic interoperability.
- Outcomes: A classification of the methods in terms of their capabilities.

Considering this, the identified keywords are the following:

- Scoping the search for our domain: "Knowledge Extraction," "Data Mining," "Knowledge Discovery," and "Information Retrieval."
- Related with the population: MRDM, "concept analysis" and their synonyms.
- Terms related with the intervention: "algorithm," "method," and "tool,"
- Search terms related with the comparison: semantic, semantically, interoperability, and interoperable.

3.2.2 Query

Following the four sets of keywords defined in the previous section, we divide the query string into four groups.

S_1: "*algorithm*" OR "*method*" OR "*tool*".

S_2: "*knowledge*" OR "*extraction*" OR "*data*" OR "*mining*" OR "*information*" OR "*retrieval*". This is to look for papers having not only the concepts "knowledge extraction", "data mining", and "information retrieval", but also the combination of words such as "knowledge mining", "data extraction". This group also includes other combinations, such as "mining extraction" or "knowledge data", and the isolated words, but those articles will be filtered out in a later stage (Sect. 3.4).

S_3: "*RCA*" OR "*concept analysis*" OR "*MRDM*" OR "*multi relational*". The purpose of this group is to consider only papers that related to the frameworks FCA or some other form of CA, and MRDM.

Table 1 Query strings used for each search engine

Database	Search
ACM	Title:((method OR algorithm OR tool) AND (knowledge extraction OR data mining OR Information Retrieval) AND (rca OR "concept analysis" OR mrdm OR "multi relational") AND (semantic* OR interop*)) OR Abstract:((method OR algorithm OR tool) AND (knowledge extraction OR data mining OR Information Retrieval) AND (rca OR "concept analysis" OR mrdm OR "multi relational") AND (semantic* OR interop*)) OR Keyword:((method OR algorithm OR tool) AND (knowledge extraction OR data mining OR Information Retrieval) AND (rca OR "concept analysis" OR mrdm OR "multi relational") AND (semantic* OR interop*))
IEEE	(method OR algorithm OR tool) AND (knowledge OR extraction OR data OR mining OR information OR retrieval) AND (rca OR "concept analysis" OR mrdm OR "multi relational") AND (semantic* OR interop*)
Scopus	TITLE-ABS-KEY ((method OR algorithm OR tool) AND (knowledge OR extraction OR data OR mining OR information OR retrieval) AND (rca OR "concept analysis" OR mrdm OR "multi relational") AND (semantic* OR interop*))
Taylor & Francis Online	(method OR algorithm OR tool) AND (knowledge OR extraction OR data OR mining OR information OR retrieval) AND (rca OR "concept analysis" OR mrdm OR "multi relational") AND (semantic* OR interop*)
Web of Science	(method OR algorithm OR tool) AND (knowledge OR extraction OR data OR mining OR information OR retrieval) AND (rca OR "concept analysis" OR mrdm OR "multi relational") AND (semantic* OR interop*)

S_4: "*semantic**" OR "*interop**". This is a group to aim the search toward those papers being about any form of semantic or interoperability. Note that the * means that all types of endings are valid. For example, *semantic* is a valid word as well as *semantically*.

Finally, the query S_1 AND S_2 AND S_3 AND S_4 was performed on the selected databases on the titles, abstracts and keywords, as we show in Table 1.

3.3 Database selection

For the search, we considered the recommendation in [67] that says using IEEE and ACM plus two indexing databases is sufficient. Additionally, we also included a more general database, Taylor & Francis Online, to possibly reach applications in more domains. In summary, we used the databases: ACM, IEEE_Xplore, Scopus, Taylor & Francis Online, and Web of Science. The obtained results can be seen in Table 2. Zotero,[4] a reference management tool, was used in order to delete duplicates and to manage the large amount of references. This study has been conducted in May 2022, so all articles up until the 30th of April were considered during the search.

[4] https://www.zotero.org/.

Table 2 Number of papers obtained on the queries execution on each database

Database	Total	Unique	Duplicated
ACM	169	36	133
IEEE	72	16	56
Scopus	644	644	0
Taylor & Francis online	6	0	6
Web of science	209	12	197
Total	1100	**708**	392

Table 3 Criteria for selecting papers

Inclusion	Exclusion
C1	
Papers written in English	Papers not written in English
Conference Papers	Non-peer-reviewed papers
Journal Papers	Literature reviews
Book Chapters/Sections	Proceedings
C2	
About MRDM or About Formal Concept Analysis	Anything not satisfying the inclusion criteria
Applied to Distributed Architectures or Big Data or Optimization of the method	

3.4 Study selection and quality assessment

From the results, firstly, we selected papers based on the titles and abstracts in a step named first screening. Secondly, we included or excluded papers based on a full-text reading, in a step named second screening. The criteria used for the selection can be seen in Table 3, which is divided in two groups, C_1 being about the type of articles to be included or excluded, and C_2 being their content. It is important to mention that, when in doubt during the title and abstract reading, articles were taken to full-text reading instead of directly excluded. One of the threats to the reliability of the review is that the selection was conducted by only one author, and hence it could have misclassified some articles (more on that in Sect. 3.7). To mitigate this threat, a different author examined a subset of the final selection. Afterward, the full-text and metadata of the references of the articles that passed the first and second screening are filtered with the same criteria in a step called *backward snowballing*, introduced by Jalali and Wohlin [45], and applied in [27]. Similarly, the process of searching in the papers referencing the selection after the second screening is called *forward snowballing*. For these steps, we used the Scolr[5] tool, which allows conducting both screening and snowball steps. The number of papers included/excluded in each of the steps can be seen in Fig. 2.

The quality assessment consisted in comparing the final 69 primary studies with an independent set of papers we knew that ought to be in the final set [4, 15, 16, 19]. Also, the following questions were answered to assess the quality of the selected articles:

- Is the method clearly the core part of the article?
- Are the algorithms, architectures, or methods explicitly defined?

[5] https://scolr.lifia.ar.

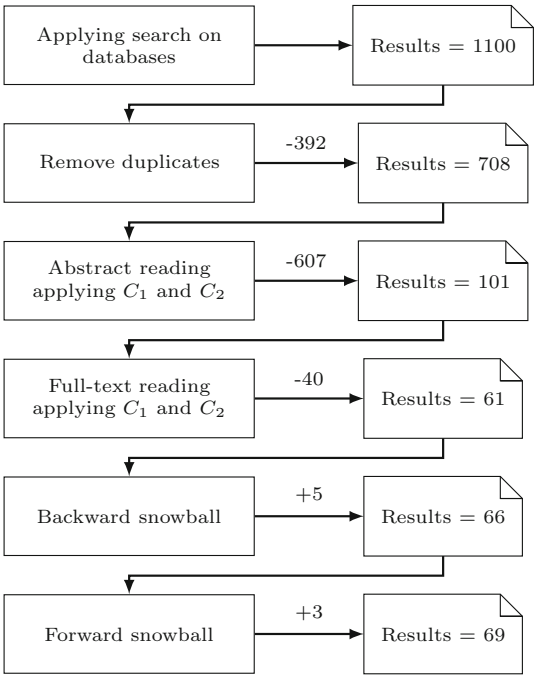


Fig. 2 Number of articles included/excluded in each step of the selection process

- Are experiments or discussions conducted to evaluate the results?

Thus, studies not explicitly defining the methods, algorithms, architectures or not doing experiments or discussions on the results were excluded. The final selection of articles is presented in Table 4.

3.5 Data extraction

The data extraction from the identified primary studies has been done considering Table 5. Data extraction fields have a *data item* and a *value*, which are the name and the description of the fields, respectively. Notice that although the *Article ID* and the *Article title* fields are not related with any LRQ, they were deemed useful, trivially, to differentiate the articles univocally.

3.6 Analysis and classification

The fields extracted from each selected primary article are tabulated and visually illustrated in Sect. 4. The papers were grouped and counted by each of the fields, and since some of them are multiple (e.g., methods), the total sum is greater than the total amount of papers reviewed.

For the field *methods application*, the articles were grouped by how the methods are being used in them, which could be CA, MRDM, or DA. An article is considered to be in CA if the method proposed in it is based on FCA or any of its extensions. Additionally, an article is considered to be in MRDM if it proposes a method for extracting knowledge from any multi-relational data format. Finally, an article is considered to be in DA if the method is

Table 4 Articles by ID

Article ID	Article	Article ID	Article
1	[94]	36	[13]
2	[90]	37	[23]
3	[7]	38	[58]
4	[101]	39	[25]
5	[35]	40	[64]
6	[3]	41	[4]
7	[1]	42	[44]
8	[46]	43	[85]
9	[37]	44	[16]
10	[40]	45	[99]
11	[31]	46	[100]
12	[11]	47	[33]
13	[77]	48	[76]
14	[38]	49	[18]
15	[56]	50	[19]
16	[71]	51	[65]
17	[83]	52	[63]
18	[60]	53	[43]
19	[93]	54	[29]
20	[69]	55	[89]
21	[26]	56	[61]
22	[12]	57	[8]
23	[47]	58	[82]
24	[88]	59	[2]
25	[73]	60	[52]
26	[32]	61	[78]
27	[36]	62	[6]
28	[49]	63	[9]
29	[81]	64	[14]
30	[98]	65	[87]
31	[97]	66	[95]
32	[15]	67	[20]
33	[75]	68	[41]
34	[5]	69	[42]
35	[10]		

implemented using a DA system, i.e., (i) in Sect. 2.3, or if it *extracts knowledge from a DA system*, i.e., (ii) in Sect. 2.3. As an example of how the classification was done, a paper that used FCA for knowledge extraction but implementing a way to use it in DA would have CA and DA in its *method application* field. Furthermore, the analysis was conducted considering the three main groups CA, MRDM, and DA, allowing us to draw conclusions about their different characteristics.

Table 5 Data extraction fields divided in two subcategories: general and process

Data item	Value	RQ
General		
Article ID	DOI/ISBN	
Article Title	Name of the article	
Year of Publication	Calendar year	LRQ_1
Publication Venue	Name of publication venue	LRQ_1
Process		
Methods application	CA | MRDM | DA	LRQ_2 & LRQ_3 & LRQ_4
Addressed in	In which part of the DM process (Fig. 3) the article is addressed	LRQ_5
Domains	In which domains the methods are applied	LRQ_6
Input of the method	Type of input, e.g., relational database, simple table, ontology	LRQ_7
Output of the method	Different types of knowledge representations, e.g., association rules, ontologies	LRQ_7
Evaluated characteristics	Evaluated characteristics of the article, e.g., accuracy, robustness	LRQ_8
Semantic Interoperability	Formats that the article contributed to their semantic interoperability	LRQ_9

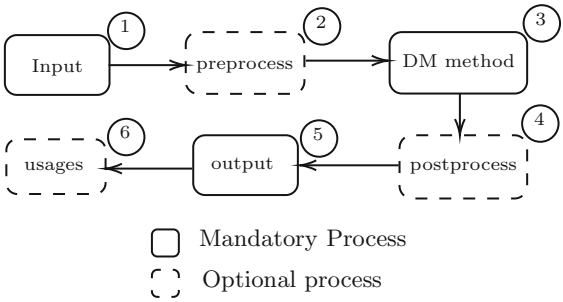


Fig. 3 Generic Data Mining process life cycle

The *addressed in* field reflects where the problematic is addressed regarding the generic DM process lifecycle presented in Fig. 3. This field is multivalued because an article can address the problematic in more than one way. This process is based on the one introduced in [96] under the name of "The steps for data mining process" in Fig. 3. In our diagram,

1. *input* corresponds to the target data where it is needed to extract knowledge (*data* and *target data* in [96]).
2. *Preprocess* is the treatment of the target data in order to match it with the specific input of the DM method to be used (*preprocess* and *transformation* in [96]).
3. *DM method* is the method/algorithm to be used in particular, e.g., k-means, FCA, or Regression Trees (*data mining* in [96]).
4. *Postprocess*, analogously to preprocess, is the treatment of the *direct* output of the DM method in order to achieve the *expected* output, that not always matches with the one it produces.
5. *Output* corresponds to the final knowledge produced (*knowledge* in [96]).
6. *Usages* represents a step in which the produced knowledge is exploited.

The purpose of the *domains* field is to represent in which domain the paper applies the method. The considered domains are

(1) *general* when the article presents a solution that is not bounded to any specific domain (although it could be applied or tested in a specific one, the solution is still presented in a general way),
(2) *semantic web* when the authors aim to contribute with a semantic web technology such as ontologies or Resource Description Framework (RDF),
(3) Machine Learning (*ML*) when the contribution is based to any machine learning method.

The fields *Input of the method* and *Output of the method* represent the method's expected and produced formats, respectively. For instance, if the method used is plain FCA, the input of the method is going to be a Formal Context. The purpose of this field is to determine what are the formats covered to perform data mining on the one hand, and on the other one, what are their respective output formats.

For the *evaluation strategies* that the articles use in order to evaluate their proposed solution, this work makes use of the taxonomy defined in [70]. In particular, the evaluation types considered for categorizing the evaluation strategies of articles are

(1) *Accuracy*: the degree of agreement between outputs of the method and the expected outputs.
(2) *Effectiveness*: the degree to which the method achieves its goal in a real situation.
(3) *Efficacy*: the degree to which the method achieves its goal considered narrowly, without addressing situational concerns.

(4) *Efficiency*: the maximization of the ratio between outputs and inputs of the method.
(5) *Robustness*: the ability of the method to handle invalid inputs or stressful environmental conditions.
(6) *Performance*: the degree to which the method accomplishes its functions within given constraints of time or space. Speed and throughput (the amount of output produced in a given period of time) are examples of time constraints. Memory usage is an example of space constraint.
(7) *Technical feasibility*: evaluates, from a technical point of view, the ease with which a proposed method will be built and operated.
(8) *Operational feasibility*: evaluates the degree to which management, employees, and other stakeholders, will support the proposed method, operate it, and integrate it into their daily practice.
(9) *Learning capability*: the ability of the method to learn.
(10) *Validity*: means that the method works correctly, i.e., correctly achieves its goal.
(11) *Scalability*: the ability of the method to either handle growing amounts of work in a graceful manner, or to be readily enlarged.
(12) *Ease of use*: the degree to which the use of the method by individuals is free of effort.
(13) *Consistency*: the degree of uniformity, standardization, and freedom from contradiction among the elements of the structure of the method.
(14) *Utility*: measures the value of achieving the method's goal, i.e., the difference between the worth of achieving this goal and the price paid for achieving it.

The given definitions come from the appendix in [70].

Finally, the field *semantic interoperability* represents how the articles contributed to the semantic interoperability between different knowledge formats. For example, let us consider an article that develops a method to solve the problem of certain links (edges) not being present in knowledge graphs (KG), by using ML and completing the links with certain predictions. Such an article would have the value *KG* in this field, because it is contributing to the semantic interoperability between two KG, one having more links than the other based on a probabilistic method.

3.7 Validity evaluation

According to the guidelines in [67], these types of validity should be considered: descriptive, theoretical, generalizability, and interpretive validity, all of which are explained and detailed below.

3.7.1 Descriptive validity

Descriptive validity is how accurately and objectively observations are described. Qualitative works have a greater threat to descriptive validity than quantitative ones. In order to reduce this threat, we created a data extraction form in Table 5 and discussed each field in Sect. 3.6 to help objectify the recording of data. Thus, this threat is considered under control.

3.7.2 Theoretical validity

The ability of being able to capture what we intend to capture is called *theoretical validity*. For instance, biases can lead us to select articles we should not and to not select others that we should.

While searching, studies could have been missed. For example, two searches in the same field, could yield different sets of articles. To address this threat, backward and forward snowball sampling of all articles has been done after the full-text reading [45] (see Fig. 2).

Bias is also a threat in the phase of *data extraction and classification*. To address this problem, it is useful that one researcher performs the phase while the other reviews it. However, given the fact that the process involves human judgment, the threat is unavoidable.

3.7.3 Generalizability

[68] introduced a distinction between external and internal generalizability, i.e., between groups or organizations, and within a group, respectively. Internal generalizability does not represent a major threat because of the wide range of articles following the same strategies. In terms of the external one, it is not a major threat as well because the approach takes into account general metrics, defined in Table 5, that can be applied to other fields of study.

3.7.4 Interpretive validity

Interpretive validity refers to the idea of the conclusion being reasonable given the data. A threat to interpretive validity is, again, researcher bias. In our case, the first author's major field of experience is algorithms and efficiency, and that could lead to a bias in the interpretation. Despite this, the rest of the authors are experts in knowledge representation and extraction, semantic interoperability, and industrial engineering, helping reduce this threat.

3.7.5 Repeatability

Repeatability refers to how repeatable the process is. It demands a thorough reporting of the research process. We reported the SLR process followed, and moreover, we explained the different possible threats and our actions to reduce them.

4 Results of the review

The results of the data extraction have been gathered and put together in an open dataset in [53] to facilitate its reproducibility. In this section, using the mentioned dataset, the findings of the systematic literature review are illustrated and presented.

4.1 Frequency of publications ($\mathbf{LRQ_1}$)

Articles were counted by year and publication venue, and the results are shown in Figs. 4 and 5, respectively. The year with more articles was 2016 with 13 articles, more than doubling the second most occurred years 2010, 2019, and 2020 with 6 papers each. Most of the papers were published after 2010. Moreover, the years 1996–2000, 2002, 2003, 2007, and 2008 did not have occurrences.

The growth in papers after 2009 showed that there is still an increasing interest in this topic in the scientific community. Additionally, one of the hypothesis we have about the reason this happens is the increase of challenges the surge of IoT had in those years.

In terms of publication venue, most of the papers were published in conferences and journals with 35 and 26 occurrences each, respectively. Only 6 workshop articles and 2 book

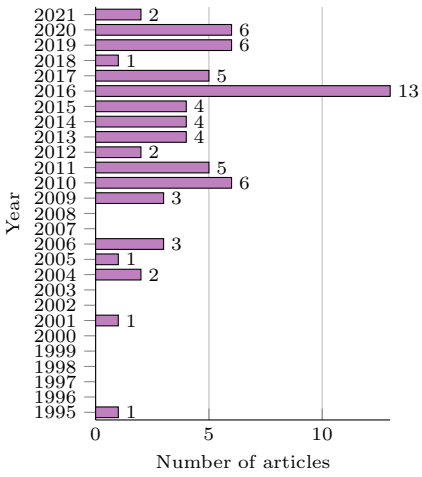


Fig. 4 Articles per year

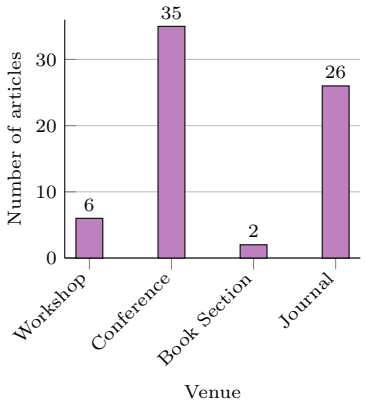


Fig. 5 Articles per publication venue

sections have been included in the study. Particularly, both the journals and conferences in the study are heterogeneous, ranging from computer science specific to engineering and manufacturing ones.

4.2 Methods application (LRQ_2 & LRQ_3 & LRQ_4)

The used methods in the articles were counted and categorized in the main three categories in this study: CA, MRDM, and DA.

In Fig. 6, there is a Venn diagram with the three categories showing the amount of articles in each of the parts: only CA 17, only MRDM 19, and only DA 2. For the mixed used methods, there are 21 CA ∩ MRDM, 8 CA ∩ DA, and 2 in MRDM ∩ DA. Interestingly, there were no articles with methods doing CA ∩ MRDM ∩ DA.

It is important to notice that both CA and MRDM overall had a similar total amount of occurrences: 46 and 42, respectively. The main difference relies on the methods based on DA that are almost unsubstantial in MRDM, with only 2 articles, while CA had 8 occurrences in that regard. Furthermore, in Table 6, the classification on where each DA article stands regarding the items (i) and (ii) is shown.

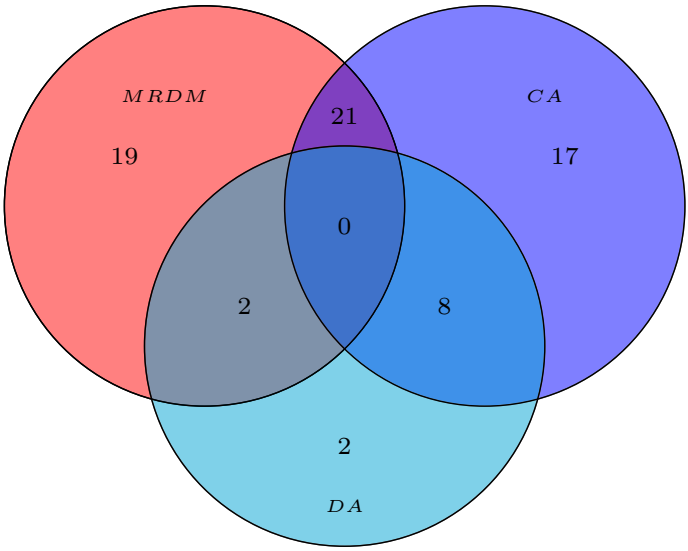


Fig. 6 Methods used in the articles classified in CA, MRDM and DA

Table 6 Subclassification of articles according to whether they are implemented *using* a DA system (i), or they mine *from* a DA system (ii)

Article	(i)	(ii)
Concept discovery from un-constrained distributed context [32]	X	X
Distributed online Temporal Fuzzy Concept Analysis for stream processing in smart cities [15]	X	X
Making sense of cloud-sensor data streams via Fuzzy Cognitive Maps and Temporal Fuzzy Concept Analysis [16]	X	X
Online query-focused twitter summarizer through fuzzy lattice [18]		X
Regression on evolving multi-relational data streams [43]		X
Two FCA-Based Methods for Reducing Energy Consumption of Sensor Nodes in Wireless Sensor Networks [52]		X
xStreams: Recommending Items to Users with Time-evolving Preferences [78]		X
Time Aware Knowledge Extraction for microblog summarization on Twitter [14]		X
AddIntent: A New Incremental Algorithm for Constructing Concept Lattices [87]		X
Distributed Formal Concept Analysis Algorithms Based on an Iterative MapReduce Framework [95]	X	
Mining time-changing data streams [41]		X
Learning model trees from evolving data streams [42]		X

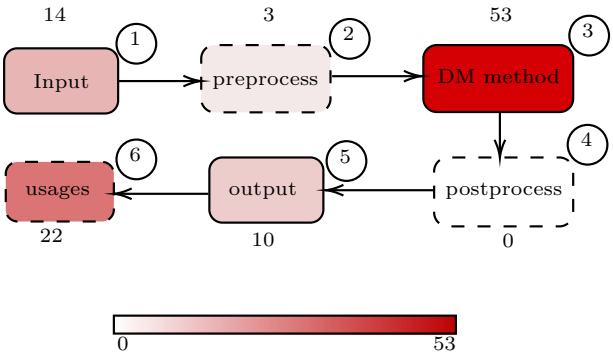


Fig. 7 Heatmap showing the number of articles regarding the DM process life cycle presented in Fig. 3

4.3 Addressed problematics ($\mathbf{LRQ_5}$)

For the addressed problematics, articles were categorized according to their place in the DM process lifecycle depicted in Fig. 3. In Fig. 7, the amount of articles according to the maximum number of papers found in all of the areas has been depicted using a heatmap where completely white means 0 articles, and completely red means 53. We can observe that 14 articles addressed a problematic regarding the input, 3 the preprocess, 53 the DM method, 0 the postprocess, 10 the output, and 22 the usages.

Table 7 shows that from the articles that addressed the problematic by defining a particular DM method, 10 also did it with a specific usage in mind. Only 1 define a particular output while still aiming to a specific use case. And 7 defined the DM method with a specific output in mind, but without the use case.

Additionally, there were 5 articles that addressed their problematic by providing a new type of input to an already defined DM method. 11 also defined a new DM method to consume the defined input. Only 1 defined the entire process with a particular type of input and output. And 2 articles defined a new input to solve a specific use case.

Regarding the articles whose problematic was based on the produced output, only 2 focused entirely on it, whereas 1 did it with a particular use case. Moreover, 3 articles presented a preprocess step with a DM method, and only 1 of them also defined a new output. Finally, there were 5 articles whose problematic was based on a specific use case and they did not have to define anything new but the application.

Notice that there were no articles addressing the problematic with a postprocess solution. One reason why this happened could be the nature of the article databases selected, which are more centered in computer science and not so much in applications, while the postprocess step is probably more common in specific domains that look novel ways to **utilize** DM methods rather than to **improve** them.

4.4 Domains ($\mathbf{LRQ_6}$)

The domains found in the articles by method category are depicted in Fig. 8. From the 46 CA articles, 41 were found to be general solutions, while 4 were applied specifically to semantic web technologies, and 1 to the machine learning domain. Additionally, from the 42 articles in MRDM, 30 were general, 9 on machine learning techniques, and 3 on

Table 7 *Problematic addressed in* multivalued field

Addressed problematic	# of articles
dm method	18
dm method, output	7
dm method, output, usages	1
dm method, usages	10
input	5
input, dm method	11
input, dm method, output	1
input, dm method, usages	2
input, usages	3
output	2
output, usages	1
preprocess, dm method	2
preprocess, dm method, output	1
usages	5

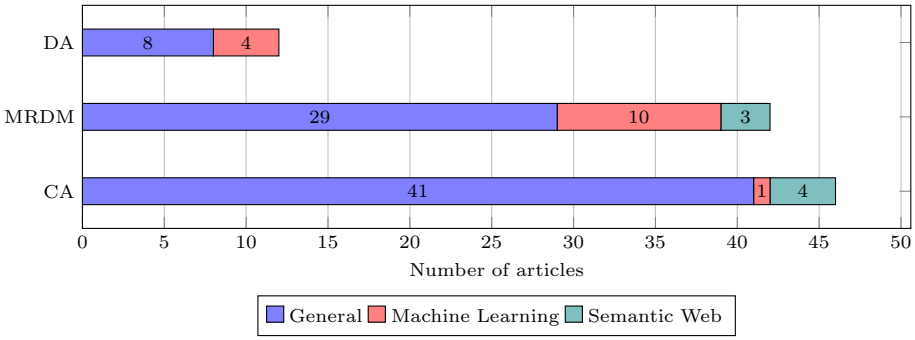


Fig. 8 Domains of articles classified by method category

semantic web technologies. Moreover, from the 12 DA articles, 8 were found to be general, 4 particularly in the domain of machine learning. More granularly, Fig. 9 depicts in a taxonomy the subcategories in each of the three main ones. There were only two subcategories shared by *general* and *machine learning*: *financial* and *social networks*.

4.5 Input and output formats (LRQ_7)

The amount of input formats found in the search was larger than anticipated. In fact, almost every paper used a specific format in it. Thus, for readability purpose, in Fig. 10 only the formats with more than 1 occurrence are displayed. The first thing that stands out of the figure is the fact that only 3 input formats have more than 3 occurrences: Formal Context, RDB, and Data Streams. Then, showing the heterogeneity of scenarios addressed in the articles, RCF has 3 occurrences, and with 2 occurrences there are the following input formats Distributed Formal Context, Fuzzy Formal Context, RDF, Documents (text), Interordinal Formal Context, Knowledge Base, Tweet Stream, and Multi-relational Graph.

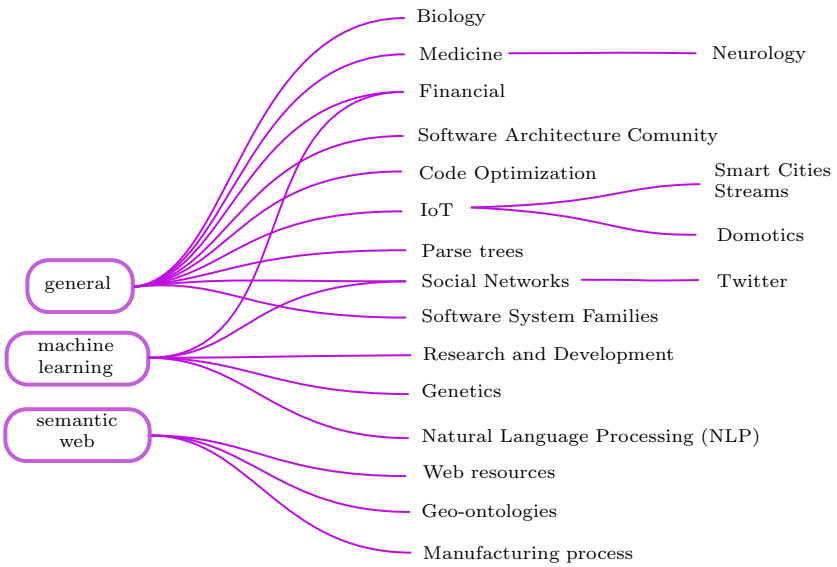


Fig. 9 Taxonomy of the domains found in a granular way

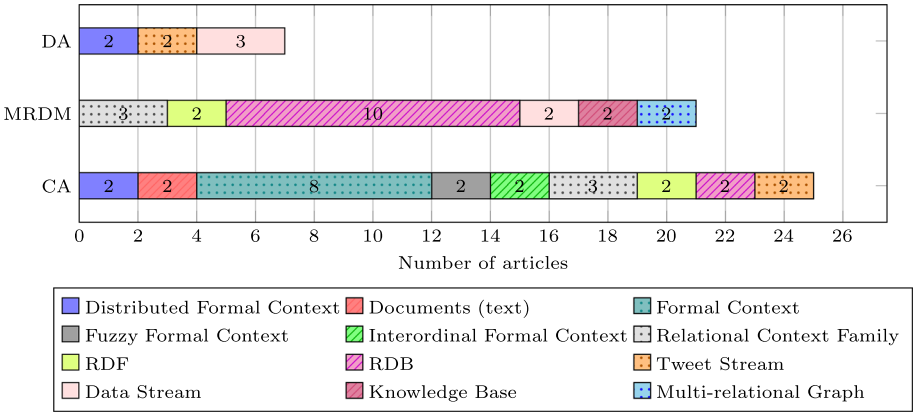


Fig. 10 Input formats with more than 1 occurrence by method category

The amount of output formats found in the search was as numerous as the amount of input formats. For the same reason, Fig. 11 shows only formats with more than 1 occurrence. Differently to the input formats figure, in this one the heterogeneity is slightly more balanced, that is there are 4 dominant format types instead of 2, being *classifier* with 9 occurrences in MRDM, *concepts lattice* with 6 occurrences in CA and 2 in MRDM, *association rules* with 6 in MRDM and 3 in CA, and *ontology* with 6 in CA and 5 in MRDM. In DA, the three output formats with more than 2 occurrences are *formal concepts*, *set of tweets*, and *regression tree* with 2 occurrences each.

In Fig. 12, the relations between input and output formats are depicted. On the left part, there are the input formats with more than 2 occurrences, and on the right hand, the output formats with more than 2 occurrences plus a node *other* representing any output format with

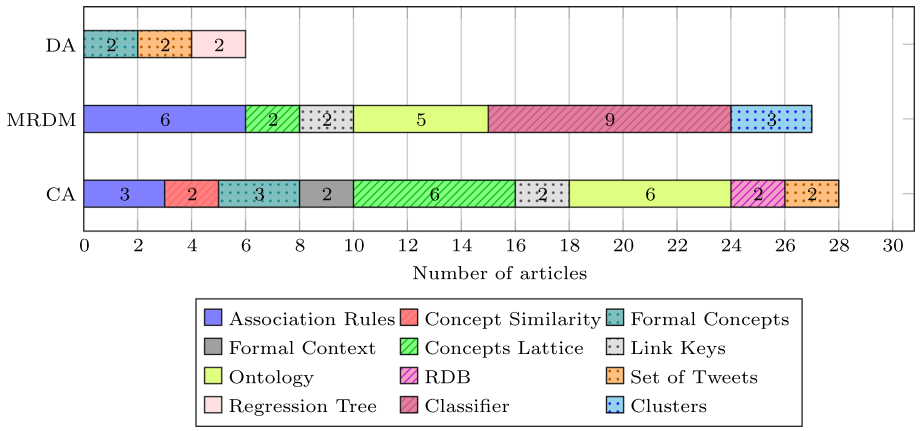


Fig. 11 Output formats with more than 1 occurrence by method category

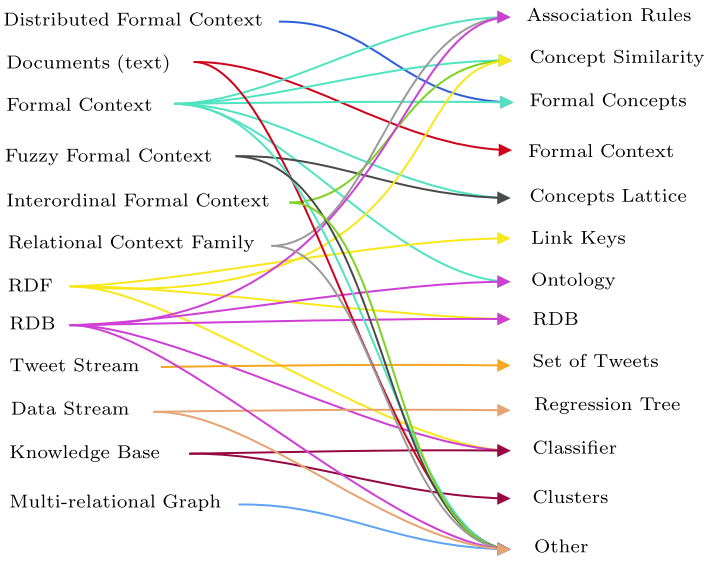


Fig. 12 Relations between input and output formats

only one occurrence. Both sides are connected by arrows conveying that for the specific input format on the arrow, there is at least one article that produces that output format.

4.6 Evaluated characteristics of the methods (LRQ_8)

The evaluated characteristics of the methods in each article have been assessed and counted, yielding the results depicted in Figs. 13, 14, and 15. The top evaluated characteristic was *performance*, having been evaluated in the three method categories with in a similar percentage of all evaluations. The second top evaluated characteristic overall was *effectiveness*; however, it has only been evaluated in 1 of the DA articles. *Accuracy* has been widely evaluated in the three method categories with a considerable percentage, being DA the one in which the

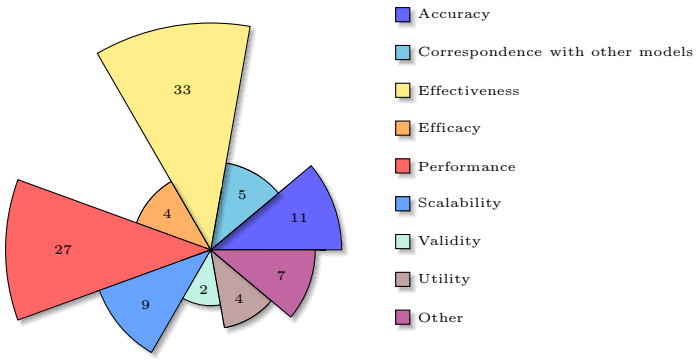
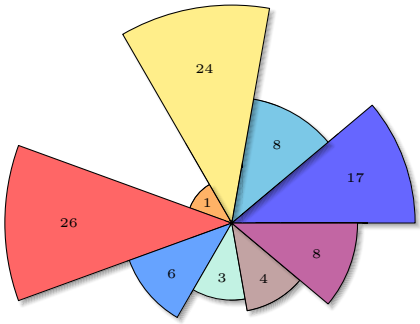


Fig. 13 Evaluation strategy by method in CA

Fig. 14 Evaluation strategy by method in MRDM



metric is the top evaluated, alongside with *performance*. *Scalability* is the next most evaluated characteristic in line, with 9, 6, and 3 occurrences in CA, MRDM, and DA, respectively. *Utility* is a characteristic that, even though it has been evaluated in all method categories, it had more attention in the DA ones in terms of occurrences relative to the amount of articles. *Correspondence with other models* was not so popular between the evaluated characteristics, but it was at least fairly studied in MRDM with 8 occurrences out of the 42 articles. *Validity* and *efficacy* had a similar amount of occurrences in all method categories, nevertheless, they had more relative weight in DA, and a very low impact in CA. Finally, *other* is the category for the evaluated characteristics with 2 or fewer occurrences in all three method categories: *technical feasibility*, *robustness*, *ease of use*, *efficiency consistency*, and *learning capability*.

Different from CA and MRDM, only 1 DA article evaluated effectiveness. This might be related to the fact that DA is still not in a state as mature as that of CA and MRDM. Moreover, this is also similar to the result about having only one DA article evaluating correspondence with other models, while in CA and MRDM there are 5 and 8, respectively. On the other hand, both in the DA and MRDM fields, not all the papers evaluating performance also evaluated scalability, in fact, in those fields, only a small percentage did.

4.7 Methods for semantic interoperability between knowledge formats ($\mathbf{LRQ_9}$)

The problem of semantic interoperability between different knowledge formats has been assessed for each article. From that assessment, several clusters of articles essentially address-

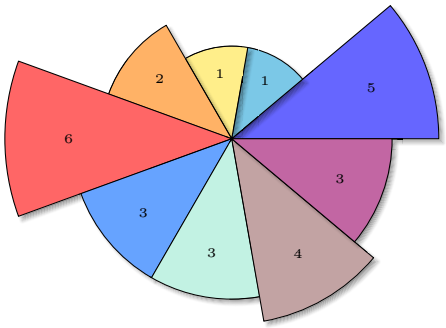


Fig. 15 Evaluation strategy by method in DA

ing the interoperability between the same knowledge formats have been extracted and are depicted in Fig. 16. Following, we summarize each cluster of papers considering whether the methods used were CA, MRDM, or DA and mention the specific methods used in each of them,

1. Tab (tabular data): The articles [4, 73] adopted CA methods to address the semantic interoperability between Formal Contexts, and Comma Separated Value (CSV) files to Formal Contexts. The methods that were utilized were FCA and Mixed FCA.
2. StL (stream to lattice): [15, 16, 32, 87, 95] adopted methods in CA and DA to address the semantic interoperability from Distributed Formal Contexts, Tuples, and Streams, to lattices or some usage of them such as evolution paths. The methods used were FCA, Incremental Fuzzy FCA, and Fuzzy Cognitive Maps.
3. FC (to formal concepts): The papers [9, 26, 38, 58, 65, 77, 88, 90, 93] adopted CA and MRDM methods to address the semantic interoperability from Fuzzy Formal Context, and Formal Concepts, to Formal Concepts. The methods used in these articles were FCA, Granular Computing, Interordinal FCA, Multi-valued FCA, set-coverage, and FCA + Inclusion Degree Theory.
4. TXT (between free text documents): [5, 12, 35, 60] utilized CA and MRDM methods to address the semantic interoperability between documents with free text. The specific methods used were FCA + R-Trees, and FCA + Latent Semantic Analysis.
5. AR (to Association Rules): The articles [20, 63, 71, 76, 82, 100] adopted CA and MRDM methods to deal with the semantic interoperability from Association Rules, Noisy Formal Context, Fuzzy Formal Context, and Relational Context Family (RCF) to Association Rules. The methods used were FCA, Fuzzy C-means Clustering, Fuzzy FCA, RCA-AOC, and Inductive Logic Programming (ILP).
6. Streams (between streams): The papers [14, 18, 41–43, 78] utilized methods in MRDM and in DA to address the semantic interoperability between different kinds of Streams such as tweets, heterogeneous, and general streams. The specific methods used were Fuzzy FCA, Regression, and Decision Trees.
7. Misc (miscellaneous formats): The articles [8, 46, 61, 97] used CA and MRDM methods to deal with the semantic interoperability between Software Programs (source code), and Software System Families and training datasets. The methods used were Convolutional Neural Networks, FCA + Pattern Structures, FCA + Spatial Indexing, and Granular Computing.
8. Onto (to ontologies): The articles [1, 3, 6, 7, 13, 29, 40, 44, 52, 69, 81, 94] adopted CA, MRDM, and DA methods to deal with the semantic interoperability from Formal Contexts, RDBs, ontologies, and web pages, to ontologies. The methods used in the

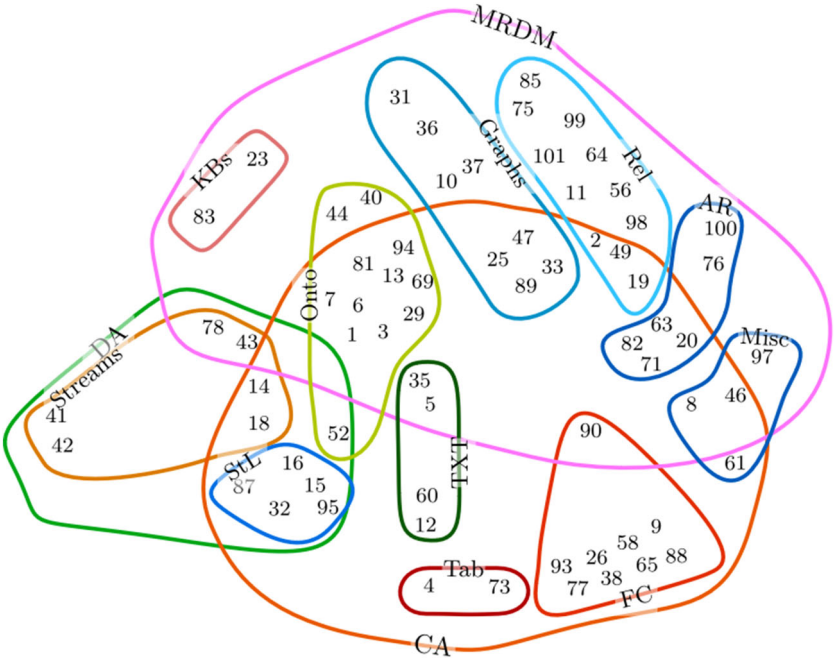


Fig. 16 Clusters of articles by semantic interoperability between knowledge formats

articles were FCA, RCA, Fuzzy FCA, Fuzzy RCA, Rough Set Theory, and Nystörm approximation.

9. Graphs (between graphs): The papers [10, 25, 31, 33, 36, 37, 47, 89] utilized CA and MRDM methods to address the semantic interoperability between different types of graphs such as Conceptual, Knowledge, and multi-relational. The methods used were Graph-FCA, Homophilic FCA, RCA, Deep Learning, Neural Networks, Homophilic FCA, Principal Component Analysis, and Subgraph Matching.
10. Rel (between relational formats): The articles [2, 11, 19, 49, 56, 64, 75, 85, 98, 99, 101] adopted CA and MRDM methods to address the semantic interoperability between RDBs, and RCFs. The methods used were K-means, Semantic tableaux, Naïve Bayes, FCA, Tuple ID propagation, Multi-relational Iceberg-cubes, Bayesian belief network, Multi-relational frequent pattern discovery, and RCA-AOC.
11. KBs (between Knowledge Bases): [23, 83] utilized MRDM methods to address the semantic interoperability between different knowledge bases. The methods used there were evolutionary algorithms, and Fuzzy Clustering.

5 Discussion

This section is structured as follows: First, the results presented in the previous section are analyzed all together in order to find correlations and to provide a discussion that considers the findings not only in isolation. On the other hand, the found scientific gaps and a discussion about them are presented alongside some directions to address them.

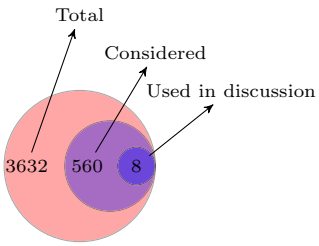


Fig. 17 Venn diagram of association rules, showing the discussed ones in relation to the total

5.1 Association rules with FCA

This subsection serves as a formal supporting element for the subsequent discussion, so that not only isolated results are considered. In Sects. 5.1.1–5.1.3, in order to extract correlations between the different categories, each of which has been discussed in the previous section, the results are interpreted by means of association rules where each article is considered a *transaction*, and each category an *item*. To be more precise, a transaction is a list of items, and an association rule is a pair $X \rightarrow Y$ whose meaning is "a transaction with items X is likely to also have items Y." For instance, if the dataset had only two articles, the first one with the categories {DM method, accuracy, classifier, ...}, and the second one with the categories {DM method, accuracy, regression tree, ...}, one possible association rule would be $r = \{\text{DM method, accuracy}\} \rightarrow \{\text{classifier}\}$, meaning that from all the articles with the categories *DM method* and *accuracy*, some of them also have the category *classifier*. There are two metrics usually considered in order to understand how meaningful these rules are: *support* (σ) and *confidence* (κ), being the amount of transactions including the items in $X \cup Y$ over the amount of all transactions, and the amount of transactions including $X \cup Y$ over the ones including X, respectively. Particularly in this example $\sigma(r) = 50\%$ because half of the articles have all the categories in the rule, and $\kappa(r) = 50\%$ because from the articles having the categories in X, half of them also have the categories in Y. From now on, the association rules will be noted as $X \rightarrow Y_{\sigma=x,\kappa=y}$, where X and Y sets of properties and x and y are the support and confidence, respectively. One of the main difficulties in analyzing these rules is the amount they are in relation to the amount of items and transactions (i.e., exponential). Additionally, several rules can be redundant, for example if $X \rightarrow Z$, we could say that $X \rightarrow Z \backslash X$. In this particular case, as depicted in Fig. 17, the total amount of rules found with Lattice Miner,[6] with minimum $\sigma = 0.1\%$ and minimum $\kappa = 0\%$ were 3632. Moreover, the association rules considered for the analysis were selected from the 560 ones found with Lattice Miner, with a minimum $\sigma = 5\%$ and a minimum $\kappa = 70\%$, i.e., rules that involve at least 5% of the total articles, and that at least 70% of the articles having X also have Y.

Furthermore, three scientific gaps that have been detected during the article assessment are discussed in Sect. 5.2 together with possible research directions and approaches to address them.

5.1.1 About the input treatment

Data Mining methods can be very specific, applied to niche cases with particular characteristics. However, there are methods that are generalized to some degree and allow their

[6] https://sourceforge.net/projects/lattice-miner/.

usage in a variety of different scenarios. For that reason, practitioners sometimes prefer to transform their input to a "standard" one in order to use a general data mining method. In some cases, such transformation is trivial, but in some others, whether it is because of the loss of semantics, performance, or lack of resources, it is not. Thus, there are research articles that address this matter in certain ways, such as providing a more effective transformation function, adapting the DM method in order to be able to process more input formats, or both.

Among the association rules including *input* in their hypothesis, there is the rule $\{\text{DA, input}\} \rightarrow \{\text{DM method}\}_{\sigma \simeq 8\%, \kappa = 100\%}$ implying that *all* articles with a DA method addressing a problematic regarding the input defined or adapted a particular DM method. One of the reasons why this could be happening is that input formats in the DA environment are usually not easily convertible to non-distributed ones, thus improving the interoperability of the DM algorithms in that regard requires adapting the existing algorithms to properly function. Another reason why this could happen is the lack of general DM methods in the DA environment, meaning that with the slightest change in the input, either a modification of an existing method or a completely new method has to be implemented. Another interesting rule is $\{\text{scalability, input}\} \rightarrow \{\text{CA, DM method}\}_{\sigma \simeq 7\%, \kappa = 100\%}$, meaning that *all* articles that evaluated scalability as an important metric, on top of addressing an input problematic, are articles adapting or implementing new CA methods.

5.1.2 About the output treatment

There are several reasons why articles could work on the output of a DM method. The first and most obvious is the translation between a format to another one, considering a particular domain. The second one could be that the output does not comply with the requirements of the problem, such as needing too much time to compute it, or even not being able to load it because of its size. The latter is generally the case of CA methods, that, since most of them rely on FCA, their output is bounded to have to deal with the exponential sized output it could generate.

Among the association rules including *output* in their hypothesis, there is the implication $\{\text{output, effectiveness}\} \rightarrow \{\text{CA}\}_{\sigma \simeq 12\%, \kappa = 100\%}$ which means that all articles whose problematic addressed was in the output and that considered effectiveness as an interesting metric, are articles using CA methods. $\{\text{output, MRDM}\} \rightarrow \{\text{performance}\}_{\sigma \simeq 14\%, \kappa = 90\%}$ shows how important is the performance metric for articles using MRDM methods. Interestingly, $\{\text{output, CA}\} \rightarrow \{\text{MRDM}\}_{\sigma \simeq 11\%, \kappa = 80\%}$, the majority of articles addressing a problematic in output included in the CA method category, are also included in the MRDM method one.

5.1.3 About ML and semantic web

$\{\text{ML, classifier}\} \rightarrow \{\text{MRDM}\}_{\sigma \simeq 10\%, \kappa = 100\%}$ is one of the most interesting rules because it implies that all articles with methods in the domain of ML, providing as an output a classifier, work with MRDM methods. Additionally, this means that no pure CA article worked on the domain of ML and provided a classifier as the final output, although there are known CA methods to do so. Another one in the ML domain with an interesting meaning is $\{\text{ML, classifier, usages}\} \rightarrow \{\text{accuracy}\}_{\sigma \simeq 57\%, \kappa = 100\%}$, conveying that from the mentioned articles, *all* the ones addressing the problematic with a particular usage also consider accuracy as metric to evaluate.

As for Semantic Web, the rule that stands out is $\{\text{semantic web}\} \rightarrow \{\text{CA, effectiveness}\}_{\sigma \simeq 5\%, \kappa = 100\%}$, implying that all semantic web articles in the study use

CA methods and consider effectiveness as an important metric to evaluate. This did not surprise us, since semantic web technologies rely on conceptual structures, and evaluating how effectively the methods can extract those concepts is understandably a metric worth considering.

5.2 Scientific gaps

As depicted in Table 6, most of the DA papers mine *from* DAs (ii) instead of doing so by providing a distributed DM method (i.e., they were not selected because of (i)). Moreover, the vast majority of them used data streams as their input type format, considering certain common constraints. Firstly, the incremental fashion in which data streams usually arrive to be processed [15, 16, 41–43, 78, 87]. Secondly, the temporal characteristic of events [14–16, 41–43, 78]. Finally, the unbounded nature of data streams has been considered in [32] by mining from only a part of it, i.e., a *batch stream* in Apache Spark[7]. In [15] by pruning the formal concepts that did not have an occurrence of an object to them in a certain amount of time based on a parametric threshold λ. And lastly, in [41–43, 78] by considering a bounded part of the stream defined by a sliding window, like in [62].

During the assessment of the articles, we found that MRDM methods are still not enough addressed considering DA constraints and challenges (see the intersection between MRDM and DA in Fig. 6). From the selected DA articles, only [43] and [78] are MRDM methods, both mining data stream as the input data type coming from DAs, but proposing a solution that treats it locally. The rest of the MRDM articles have, for the most part, RDBs as their main input type format (see Fig. 10), which, although they could be used in DAs to store data, these articles have considered them only in non-DA related topics, and have proposed only local algorithms. Moreover, the articles not considering RDBs, worked with RCF, RDF, documents (text), or multi-relational graphs always in a non-distributed fashion, assuming that the whole structure is accessible at all times, which is often not the case in DA.

Furthermore, although CA methods had more involvement in the DA environment research (8 articles considering CA and DA, i.e., [14–16, 18, 32, 52, 87, 95]), they represent still a small portion of the total. More particularly, as depicted in Fig. 12, there were occurrences of distributed formal contexts being treated by methods to produce formal concepts, but only in the binary and traditional FCA. No distributed fuzzy formal context, nor other types of traditional CA inputs with their distributed variation, has been found in the search. In the following subsections, the identified scientific gaps are presented.

5.2.1 General CA method for infinite data streams

In the industry 4.0, several key components work in IoT architectures, which are inherently distributed. A big part of these architectures use data streams as a common format for communication between their components [17]. Moreover, some IoT architectures are designed to run infinitely, creating the need for handling certain tasks such as hot swapping [80]. Toward those lines, in the area of CA methods, there are those that can handle the processing of data streams, usually called *incremental* [15, 16, 87] by only updating the output in the necessary parts with the arrival of each element of the stream. The main problem these methods have is that if the stream never stop producing elements, at some point the update they do will be too costly. This can be solved by using sliding windows, or by removing concepts when they "get

[7] Apache Spark url.

old" (e.g., no occurrence contributed to the extent of the concept in a defined time window) like in [15]. However, the problem these methods introduce is the loss of information, i.e., in the sliding window case, transactions are lost independently of their meaning, while in the later case only the "oldest" are lost.

It is understandably unreasonable to expect a data mining method to be able to process an infinite amount of data without losing ***any*** information. Nevertheless, one of the questions that arises is *what to do when it is needed to recover* ***some*** *of the information lost*. Considering the mentioned incremental methods to process infinite data streams, one of the approaches that could be taken into account to fill the gap is the *merge* of lattices generated in different points in time [54], since this would allow regaining lost information on demand (e.g., depending on the available resources, or depending on the needs), taking advantage from not having to recalculate the stored lattices. This approach has different challenges, from which the highlights are

(1) identify the properties of the data stream that could decrease (or increase) the cost of computing the merge, e.g., whether the attributes are known in advance, what are the structures that are already available to use without any computational cost.
(2) Develop the theory to understand what are the differences between merging the stream parts and then computing the incremental algorithm on the merged data stream and directly merging the lattices.

5.2.2 General CA and MRDM method for infinite data streams

While some CA & MRDM methods directly use FCA and then add certain functions on top of it (e.g., RCA), others base their definition in the idea of *formal concepts*, but their computation is not necessarily related with the one in FCA (e.g., Triadic and Polyadic Concept Analysis). This makes the discussion provided in Sect. 5.2.1 to also hold for CA & MRDM methods, since the development of a *general* CA method for infinite data streams will not carry onto all its extensions naturally. For this reason, a general method on CA and MRDM could be considered an independent gap and has to be addressed in a specific research line, since dealing with heterogeneous sources adds many constraints and challenges to the mix, and an FCA method that allows infinite data stream processing will not be enough to cover all the needs in the MRDM domain regarding CA.

5.2.3 Deal with loss of information and semantics

Another challenge that arose from the article assessment is the standardization of information and semantic loss. On the one hand, in [50], Kuznesov proposed the definition of stability of a formal concept to measure the plausibility of concept-based hypothesis. On the other hand, in [15], De Maio et al. used a function that, together with the support, was used to determine how old a concept is in the case of fuzzy temporal formal concept analysis. These are two examples of how researchers decide how important (or the contrary) a concept is at a given moment. In order to support a flexible method that not only supports processing infinite data streams but also allows recovering certain information, it would be useful to have a way to measure the impact of removing a concept (or an object) from the lattice (or the formal context), as well as the impact of recovering a certain concept (or object).

6 Conclusion and future work

A systematic literature review has been conducted over a well-defined set of articles related to CA, and MRDM methods with the goal of understanding the state of the art regarding the environment of DA. To guide the analysis, a generic Data Mining process lifecycle has been presented by extending and adapting the one introduced in [96].

Additionally, the results of the data extraction have been presented, including the methods utilized in MRDM, CA, and DA, the addressed problematics from the point of view of the introduced process lifecycle, the domains, the input and output formats and the relationships between each other, and finally the evaluated method characteristics. Moreover, a set of association rules regarding the categorization carried out with the set of articles has been presented by considering the input and output treatment in the process lifecycle, and the ML and Semantic Web domain categories. In addition, three gaps have been identified, and some directions have been provided to address them.

In the future, we plan on working in the first and second gaps (Sects. 5.2.1–5.2.2), particularly in the theoretical foundations of **what are the characteristics** an FCA algorithm should have in order **to process infinite data streams effectively**. Following, the plan is to work toward the **implementation of the algorithm** in order to provide a benchmark to measure practical efficiency to researchers and practitioners. Afterward, the goal is to **explore the foundations in the MRDM domain**, aiming to provide also an algorithm with the same characteristics there.

Acknowledgements This work has been funded with the help of the French National Agency for Research and Technology (ANRT) and French National Syndicate of Ski Teachers (SNMSF).

References

1. Akmal S, Batres R (2013) A methodology for developing manufacturing process ontologies. J Jpn Ind Manag Assoc 64:303–316. https://doi.org/10.11221/jima.64.303
2. Albahli S, Melton A (2016) TripleFCA: FCA-based approach to enhance semantic web data management. In: 2016 IEEE 40th annual computer software and applications conference (COMPSAC), pp 625–630. https://doi.org/10.1109/COMPSAC.2016.212
3. Aloui A, Grissa A (2015) A new approach for flexible queries using fuzzy ontologies. In: Azar AT, Vaidyanathan S (eds) Computational intelligence applications in modeling and control. Studies in computational intelligence. Springer, Cham, pp 315–342. https://doi.org/10.1007/978-3-319-11017-2_13
4. Andrews S, Orphanides C (2010) Knowledge discovery through creating formal contexts. In: Proceedings—2nd international conference on intelligent networking and collaborative systems, INCOS 2010. Proceedings—2nd international conference on intelligent networking and collaborative systems, INCOS 201, p 460. https://doi.org/10.1109/INCOS.2010.53
5. Anoop VS, Asharaf S (2017) Extracting conceptual relationships and inducing concept lattices from unstructured text. J Intell Syst. https://doi.org/10.1515/jisys-2017-0225
6. Atencia M, David J, Euzenat J et al (2020) Link key candidate extraction with relational concept analysis. Discrete Appl Math 273:2–20
7. Atencia M, David J, Euzenat J et al (2019) A guided walk into link key candidate extraction with relational concept analysis. In: ISWC 2019—18th international semantic web conference. No commercial editor, Auckland, New Zealand, pp 1–9. https://hal.archives-ouvertes.fr/hal-02984963
8. Carbonnel J, Huchard M, Nebut C (2019) Towards complex product line variability modelling: mining relationships from non-Boolean descriptions. J Syst Softw 156:341–360
9. Chang-sheng Z, Jing R, Hai-long H et al (2013) An algorithm on generating lattice based on layered concept lattice. TELKOMNIKA Indones J Electr Eng 11. https://doi.org/10.11591/telkomnika.v11i8.3063

10. Choudhury S, Holder L, Feo J et al (2013) Fast search for dynamic multi-relational graphs. In: Proceedings of the workshop on dynamic networks management and mining. association for computing machinery, New York, NY, USA, DyNetMM '13, pp 1–8. https://doi.org/10.1145/2489247.2489251
11. Cordero P, Enciso M, Mora A et al (2014) A tableaux-like method to infer all minimal keys. Log J IGPL 22(6):1019–1044. https://doi.org/10.1093/jigpal/jzu025
12. Co V, Taramasco C, Astudillo H (2011) Cheating to achieve Formal Concept Analysis over a large formal context. In: CEUR workshop proceedings
13. De Maio C, Fenza G, Gallo M et al (2014) Formal and relational concept analysis for fuzzy-based automatic semantic annotation. Appl Intell 40(1):154–177. https://doi.org/10.1007/s10489-013-0451-7
14. De Maio C, Fenza G, Loia V et al (2016) Time Aware Knowledge Extraction for microblog summarization on Twitter. Inf Fusion 28:60–74
15. De Maio C, Fenza G, Loia V et al (2017) Distributed online Temporal Fuzzy Concept Analysis for stream processing in smart cities. J Parallel Distrib Comput 110:31–41
16. De Maio C, Fenza G, Loia V et al (2017) Making sense of cloud-sensor data streams via Fuzzy Cognitive Maps and Temporal Fuzzy Concept Analysis. Neurocomputing 256:35–48
17. De Francisci Morales G, Bifet A, Khan L et al (2016) IoT Big data stream mining. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. Association for Computing Machinery, New York, NY, USA, KDD '16, pp 2119–2120. https://doi.org/10.1145/2939672.2945385
18. De Maio C, Fenza G, Loia V et al (2015) Online query-focused twitter summarizer through fuzzy lattice. In: 2015 IEEE international conference on fuzzy systems (FUZZ-IEEE), pp 1–8. https://doi.org/10.1109/FUZZ-IEEE.2015.7337927
19. Dolques X, Le Ber F, Huchard M et al (2016) Performance-friendly rule extraction in large water datasets with AOC posets and relational concept analysis. Int J Gen Syst 45(2):187–210. https://doi.org/10.1080/03081079.2015.1072927
20. Dolques X, Le Ber F, Huchard M (2013) AOC-posets: a scalable alternative to concept lattices for relational concept analysis. CEUR workshop proceedings 1062
21. Džeroski S (2003) Multi-relational data mining: an introduction. ACM SIGKDD Explor Newsl 5(1):1–16. https://doi.org/10.1145/959242.959245
22. Ekanayake J, Li H, Zhang B et al (2010) Twister: a runtime for iterative MapReduce. In: Proceedings of the 19th ACM international symposium on high performance distributed computing. ACM, Chicago Illinois, pp 810–818. https://doi.org/10.1145/1851476.1851593
23. Fanizzi N, d' Amato C, Esposito F (2009) Fuzzy clustering for categorical spaces. In: Rauch J, Raś ZW, Berka P et al (eds) Foundations of intelligent systems. Lecture notes in computer science. Springer, Berlin, pp 161–170. https://doi.org/10.1007/978-3-642-04125-9_19
24. Fawzy D, Moussa S, Badr N (2022) The internet of things and architectures of big data analytics: challenges of intersection at different domains. IEEE Access 10:4969–4992. https://doi.org/10.1109/ACCESS.2022.3140409
25. Ferr S, Cellier P (2016) Graph-FCA in practice, p 107. https://doi.org/10.1007/978-3-319-40985-6_9. https://hal.inria.fr/hal-01405491
26. Formica A (2021) Concept similarity in formal concept analysis with many-valued contexts. Comput Inf 40(3):469–488. https://doi.org/10.31577/cai_2021_3_469
27. Franciosi C, Iung B, Miranda S et al (2018) Maintenance for sustainability in the industry 4.0 context: a scoping literature review. IFAC-PapersOnLine 51(11):903. https://doi.org/10.1016/j.ifacol.2018.08.459
28. Ganter B, Wille R (1999) Formal concept analysis: mathematical foundations. Springer, Berlin. https://doi.org/10.1007/978-3-642-59830-2
29. Gao ZY, Liang YQ, Qiao SH (2016) Relational database ontology discovery method based on formal concept analysis. Atlantis Press, pp 727–735. https://doi.org/10.2991/mme-16.2017.101
30. George Voutsadakis (2002) Polyadic concept analysis. Order 19(3):295–304. https://doi.org/10.1023/A:1021252203599
31. Glorot X, Bordes A, Weston J et al (2013) A semantic matching energy function for learning with multi-relational data. https://doi.org/10.48550/arXiv.1301.3485, arXiv:1301.3485 [cs]
32. Goel V, Chaudhary BD (2015) Concept discovery from un-constrained distributed context. In: Proceedings of the 4th international conference on big data analytics, vol 9498. Springer, Berlin, pp 151–164. https://doi.org/10.1007/978-3-319-27057-9_11
33. Guesmi S, Trabelsi C, Latiri C (2021) Multidimensional community discovering in heterogeneous social networks. Concurr Comput Pract Exp 33(1):e5809
34. Han J (2009) Data mining. In: Liu L, Zsu MT (eds) Encyclopedia of database systems. Springer, Boston, pp 595–598. https://doi.org/10.1007/978-0-387-39940-9_104

35. He W, Li S, Yang X (2015) A hybrid approach for reducing textual formal context based on thesaurus. In: 2015 11th international conference on computational intelligence and security (CIS), pp 146–149. https://doi.org/10.1109/CIS.2015.43
36. Hildebrandt M, Sunder SS, Mogoreanu S et al (2019) Configuration of industrial automation solutions using multi-relational recommender systems. In: Brefeld U, Curry E, Daly E et al (eds) Machine learning and knowledge discovery in databases. Lecture notes in computer science. Springer, Cham, pp 271–287.https://doi.org/10.1007/978-3-030-10997-4_17
37. Hildebrandt M, Sunder SS, Mogoreanu S et al (2019a) A recommender system for complex real-world applications with nonlinear dependencies and knowledge graph context. In: Hitzler P, Fernández M, Janowicz K et al (eds) The semantic web. Lecture notes in computer science, pp 179–193. Springer, Cham. https://doi.org/10.1007/978-3-030-21348-0_12
38. Ho T (1995) An approach to concept formation based on formal concept analysis. IEICE transactions on information and systems
39. Huang X, Lin J, Demner-Fushman D (2006) Evaluation of PICO as a knowledge representation for clinical questions. In: AMIA annual symposium proceedings AMIA symposium 2006, pp 359–363
40. Huang Y, Nickel M, Tresp V et al (2010) A scalable kernel approach to learning in semantic graphs with applications to linked data, pp 3–13
41. Hulten G, Spencer L, Domingos P (2001) Mining time-changing data streams. In: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining. Association for Computing Machinery, New York, NY, USA, KDD '01, pp 97–106. https://doi.org/10.1145/502512.502529
42. Ikonomovska E, Gama J, Džeroski S (2011) Learning model trees from evolving data streams. Data Min Knowl Disc 23:128–168. https://doi.org/10.1007/s10618-010-0201-y
43. Ikonomovska E, Džeroski S (2011) Regression on evolving multi-relational data streams. In: Proceedings of the 2011 Joint EDBT/ICDT Ph.D. Workshop. Association for Computing Machinery, New York, NY, USA, PhD '11, pp 1–7. https://doi.org/10.1145/1966874.1966875
44. Jain N, Krestel R (2020) Learning fine-grained semantics for multi-relational data, p 5
45. Jalali S, Wohlin C (2012) Systematic literature studies: database searches vs. backward snowballing. In: Proceedings of the ACM-IEEE international symposium on Empirical software engineering and measurement. Association for Computing Machinery, New York, NY, USA, ESEM '12, pp 29–38. https://doi.org/10.1145/2372251.2372257
46. Jian Z, Kong L (2016) A novel algorithm for classification rule discovery based on concept granule structure. J Digit Inf Manag 14(2):73–80
47. Khediri N, Karoui W (2017) Community detection in social network with node attributes based on formal concept analysis. In: 2017 IEEE/ACS 14th international conference on computer systems and applications (AICCSA), pp 1346–1353. https://doi.org/10.1109/AICCSA.2017.200
48. Kolajo T, Daramola O, Adebiyi A (2019) Big data stream analysis: a systematic literature review. J Big Data. https://doi.org/10.1186/s40537-019-0210-7
49. Kötters J, Eklund PW (2020) Conjunctive query pattern structures: a relational database model for Formal Concept Analysis. Discrete Appl Math 273:144–171
50. Kuznetsov S (2007) On stability of a formal concept. Ann Math Artif Intell 49:101–115. https://doi.org/10.1007/s10472-007-9053-6
51. Lasi H, Fettke P, Feld T et al (2014) Industry 40. Bus Inf Syst Eng 6(4):239–242
52. Lei Y, Qu M, Lei C et al (2022) Two FCA-based methods for reducing energy consumption of sensor nodes in wireless sensor networks. Scientific Programming 2022. https://doi.org/10.1155/2022/8520447
53. Leutwyler N, Lezoche M, Panetto H et al (2023) Systematic literature review—selected articles—data extraction. https://doi.org/10.5281/zenodo.10036717
54. Leutwyler N, Lezoche M, Torres D et al (2023) Towards a flexible and scalable data stream algorithm in FCA. In: Ojeda-Aciego M, Sauerwald K, Jäschke R (eds) Graph-based representation and reasoning. Springer Nature Switzerland, Cham, Lecture Notes in Computer Science, pp 104–117. https://doi.org/10.1007/978-3-031-40960-8_9
55. Liao Y, Lezoche M, Panetto H et al (2016) Semantic annotations for semantic interoperability in a product lifecycle management context. Int J Prod Res 54(18):5534
56. Liu H, Yin X, Han J (2005) An efficient multi-relational Naïve Bayesian classifier based on semantic relationship graph, pp 39–48. https://doi.org/10.1145/1090193.1090200
57. Maimon O, Rokach L (2005) Introduction to knowledge discovery in databases. In: Maimon O, Rokach L (eds) Data mining and knowledge discovery handbook. Springer, Boston, pp 1–17. https://doi.org/10.1007/0-387-25465-X_1
58. Majidian A, Martin T, Cintra M (2011) Fuzzy formal concept analysis and algorithm. Pages: 7

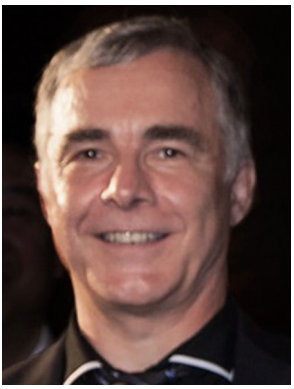
59. Manning CD, Raghavan P, Schütze H (2008) Introduction to information retrieval. Cambridge University Press, Cambridge
60. Martin B, Eklund P (2006) Asymmetric page split generalized index search trees for formal concept analysis. In: Esposito F, Raś ZW, Malerba D et al (eds) Foundations of intelligent systems. Springer, Berlin, Lecture Notes in Computer Science, pp 218–227. https://doi.org/10.1007/11875604_25
61. Martin B, Eklund P (2006) Spatial indexing for scalability in FCA. In: Proceedings of the 4th international conference on Formal Concept Analysis. Springer, Berlin, ICFCA'06, pp 205–220. https://doi.org/10.1007/11671404_14
62. Martin T, Francoeur G, Valtchev P (2020) CICLAD: a fast and memory-efficient closed itemset miner for streams. In: Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining, pp 1810–1818. https://doi.org/10.1145/3394486.3403232, arXiv:2007.01946 [cs.stat]
63. Martynovich V, Vityaev E (2016) Recovering noisy contexts with probabilistic formal concepts
64. Mashhadi NR, Jalali M, Jahan MV (2015) Inference of mobile users' social relationships using Bayesian belief network. In: 2015 international congress on technology, communication and knowledge (ICTCK), pp 232–240. https://doi.org/10.1109/ICTCK.2015.7582676
65. Mouakher A, Ben Yahia S (2016) QualityCover: efficient binary relation coverage guided by induced knowledge quality. Inf Sci 355–356:58–73
66. Patel JA, Sharma P (2020) Online analytical processing for business intelligence in big data. Big Data 8(6):501–518. https://doi.org/10.1089/big.2020.0045
67. Petersen K, Vakkalanka S, Kuzniarz L (2015) Guidelines for conducting systematic mapping studies in software engineering: an update. Inf Softw Technol 64:1–18
68. Petersen K, Gencel C (2013) Worldviews, research methods, and their relationship to validity in empirical software engineering research. In: 2013 joint conference of the 23rd international workshop on software measurement and the 8th international conference on software process and product measurement, pp 81–89. https://doi.org/10.1109/IWSM-Mensura.2013.22
69. Ping Q, Zhongxiang Z, Hualing G et al (2010) Attribute exploration algorithms on ontology construction. In: Shi Z, Vadera S, Aamodt A et al (eds) Intelligent information processing V. Springer, Berlin, IFIP Advances in Information and Communication Technology, pp 234–244. https://doi.org/10.1007/978-3-642-16327-2_29
70. Prat N, Comyn-Wattiau I, Akoka J (2015) A taxonomy of evaluation methods for information systems artifacts. J Manag Inf Syst 32(3):229–267. https://doi.org/10.1080/07421222.2015.1099390
71. Quan TT, Ngo LN, Hui SC (2009) An effective clustering-based approach for conceptual association rules mining. In: 2009 IEEE-RIVF international conference on computing and communication technologies, pp 1–7. https://doi.org/10.1109/RIVF.2009.5174619
72. Robinson P, Lowe J (2015) Literature reviews vs systematic reviews. Aust N Z J Public Health 39(2):103–103. https://doi.org/10.1111/1753-6405.12393
73. Rodriguez-Jimenez JM, Cordero P, Enciso M et al (2016) Concept lattices with negative information: a characterization theorem. Inf Sci 369:51–62
74. Rouane-Hacene M, Huchard M, Napoli A et al (2013) Relational concept analysis: mining concept lattices from multi-relational data. Ann Math Artif Intell. https://doi.org/10.1007/s10472-012-9329-3
75. Seid D, Mehrotra S (2004) Efficient relationship pattern mining using multi-relational iceberg-cubes. In: Fourth IEEE international conference on data mining (ICDM'04), pp 515–518. https://doi.org/10.1109/ICDM.2004.10059
76. Seki H, Honda Y, Nagano S (2010) On enumerating frequent closed patterns with key in multi-relational data. In: Pfahringer B, Holmes G, Hoffmann A (eds) Discovery science. Lecture notes in computer science, pp 72–86. Springer, Berlin. https://doi.org/10.1007/978-3-642-16184-1_6
77. She Y, Wang W, He X et al (2019) A three-valued logic approach to partially known formal concepts. J Intell Fuzzy Syst 37(2):3053–3064
78. Siddiqui ZF, Tiakas E, Symeonidis P et al (2014) xStreams: recommending items to users with time-evolving preferences. In: Proceedings of the 4th international conference on web intelligence, mining and semantics (WIMS14). Association for Computing Machinery, New York, NY, USA, WIMS '14, pp 1–12. https://doi.org/10.1145/2611040.2611051
79. Sowa J (2000) Knowledge representation: logical, philosophical, and computational foundations. In: Knowledge representation: logical, philosophical, and computational foundations. Brooks/Cole Publishing
80. Tabisz W, Jovanovic M, Lee F (1992) Present and future of distributed power systems. In: [Proceedings] APEC '92 Seventh Annual Applied Power Electronics Conference and Exposition, pp 11–18, https://doi.org/10.1109/APEC.1992.228437

81. Tasnim M, Collarana D, Graux D et al (2020) Chapter 8 context-based entity matching for big data. In: Janev V, Graux D, Jabeen H et al (eds) Knowledge graphs and big data processing. Lecture notes in computer science. Springer, Cham, pp 122–146. https://doi.org/10.1007/978-3-030-53199-7_8
82. Touzi AG (2010) Towards a discovering knowledge comprehensible and exploitable by the end-user. In: 2010 second international conference on advances in databases, knowledge, and data applications, pp 126–134.https://doi.org/10.1109/DBKDA.2010.36
83. Tran MD, d'Amato C, Nguyen BT et al (2017) An evolutionary algorithm for discovering multi-relational association rules in the semantic web. In: Proceedings of the genetic and evolutionary computation conference. Association for Computing Machinery, New York, NY, USA, GECCO '17, pp 513–520. https://doi.org/10.1145/3071178.3079196
84. Unbehauen J, Hellmann S, Auer S et al (2012) Knowledge extraction from structured sources. In: Ceri S, Brambilla M (eds) Search computing: broadening web search. Lecture notes in computer science. Springer, Berlin, pp 34–52. https://doi.org/10.1007/978-3-642-34213-4_3
85. Valêncio CR, Oyama FT, Scarpelini Neto P et al (2012) MR-Radix: a multi-relational data mining algorithm. HCIS 2(1):4. https://doi.org/10.1186/2192-1962-2-4
86. van Steen M, Tanenbaum AS (2016) A brief introduction to distributed systems. Computing 98(10):967–1009. https://doi.org/10.1007/s00607-016-0508-7
87. van der Merwe D, Obiedkov S, Kourie D (2004) AddIntent: a new incremental algorithm for constructing concept lattices. In: Eklund P (ed) Concept lattices. Lecture Notes in Computer Science. Springer, Berlin, pp 372–385. https://doi.org/10.1007/978-3-540-24651-0_31
88. Vychodil V (2016) Computing sets of graded attribute implications with witnessed non-redundancy. Inf Sci 351:90–100. https://doi.org/10.1016/j.ins.2016.03.004, arXiv:1511.01640 [cs]
89. Wajnberg M, Lezoche M, Blondin Masse A et al (2018) Semantic interoperability of large systems through a formal method: Relational Concept Analysis. In: 16th IFAC symposium on information control problems in manufacturing, INCOM 2018, Bergamo, Italy, pp 1397–1402. https://doi.org/10.1016/j.ifacol.2018.08.330, https://hal.archives-ouvertes.fr/hal-01813398, issue: 11
90. Wei X, Liang W, Chen Q et al (2010) A calculation method of concept similarity base on inclusion degree theory. In: 2010 2nd IEEE international conference on information management and engineering, pp 501–505. https://doi.org/10.1109/ICIME.2010.5477960
91. Wille R (1982) Restructuring lattice theory: an approach based on hierarchies of concepts. In: Rival I (ed) Ordered sets. Springer, Netherlands, NATO Advanced Study Institutes Series, pp 445–470. https://doi.org/10.1007/978-94-009-7798-3_15
92. Wolff KE (2001) Temporal concept analysis. ICCS-2001 international workshop on concept lattices-based theory. Stanford University, Palo Alto (CA), Methods and Tools for Knowledge Discovery in Databases, pp 91–107
93. Wu X, Zhang J, Lu R (2020) Attribute logic formula description of granule and its application to build concept lattice. IEEE Access 8:12592–12606. https://doi.org/10.1109/ACCESS.2020.2964834
94. Xiao J, He Z (2016) A concept lattice for semantic integration of geo-ontologies based on weight of inclusion degree importance and information entropy. Entropy 18:399. https://doi.org/10.3390/e18110399
95. Xu B, de Frin R, Robson E et al (2012) Distributed formal concept analysis algorithms based on an iterative mapreduce framework. In: Domenach F, Ignatov DI, Poelmans J (eds) Formal concept analysis. Lecture notes in computer science. Springer, Berlin, pp 292–308. https://doi.org/10.1007/978-3-642-29892-9_26
96. Yang A, Zhang W, Wang J et al (2020) Review on the application of machine learning algorithms in the sequence data mining of DNA. Front Bioeng Biotechnol 8. https://www.frontiersin.org/articles/10.3389/fbioe.2020.01032
97. Ye G, Tang Z, Wang H et al (2020) Deep program structure modeling through multi-relational graph-based learning. In: Proceedings of the ACM international conference on parallel architectures and compilation techniques. Association for Computing Machinery, New York, NY, USA, PACT '20, pp 111–123. https://doi.org/10.1145/3410463.3414670
98. Yin X, Han J, Yang J et al (2006) CrossMine: efficient classification across multiple database relations. In: Boulicaut JF, De Raedt L, Mannila H (eds) Constraint-based mining and inductive databases. Lecture notes in computer science. Springer, Berlin, pp 172–195. https://doi.org/10.1007/11615576_9
99. Zhang W (2009) Mining multi-level multi-relational frequent patterns based on conjunctive query containment. In: 2009 WRI global congress on intelligent systems, pp 436–440. https://doi.org/10.1109/GCIS.2009.290
100. Zhang W (2009) Multi-relational data mining based on higher-order inductive logic programming. In: 2009 WRI global congress on intelligent systems, pp 453–458. https://doi.org/10.1109/GCIS.2009.289
101. Zhao D, Liu X (2016) A genetic k-means membrane algorithm for multi-relational data clustering, p 959. https://doi.org/10.1007/978-3-319-31854-7_106

Nicolás Leutwyler earned his Master's degree from the Universidad Nacional de Quilmes in September 2019, with a thesis specializing in algorithms, graph theory, and operations research. Parallel to his studies, he developed professionally in the software industry, working specifically from 2015 to 2021. Currently, Nicolás is pursuing a PhD at the University of Lorraine, CRAN (Centre de Recherche en Automatique de Nancy). His research, titled "Flexible and Scalable methods for Formal Concept Analysis in Distributed Architectures", focuses on creating mathematical models and implementing intelligent methods in the area of unsupervised and supervised learning concerning distributed architectures.

Mario Lezoche is an associate professor at the University of Lorraine, specifically at IUT Hubert Curien of Epinal (Technological University Institut). He teaches object-oriented software engineering, base knowledge systems and database development. He also teaches at the Computer Science Engineering School TELECOM Nancy Artificial Intelligence, ERP configuration, Enterprise 4.0, Blockchains and Semantic programming languages. He is a researcher at CRAN (Research Centre for Automatic Control of Nancy) joint research unit with CNRS. In the CRAN laboratory, he is attached to the ISET department (Department Ingénierie des Systèmes Eco-Techniques - Eco-Techniques Systems Engineering Department) and he is co-managing the research project team Intelligent System and Objects in Interaction (S&O-2I) on the use of Knowledge formalisation (for example ontology, formal concept analysis, relational concept analysis) for optimising models interoperability of production systems. He graduated from Roma TRE University (Italy) in Computer Science Engineering. He received his PhD in Computer Science Engineering in 2009 and his post-Doc in 2011 at the University of Lorraine. In 2021, he achieved the HDR, the French diploma to lead the research. He has a strong experience in Semantic Web research and in models and semantics for systems interoperability. He has long worked in information systems modelling, semantics modelling and discovery, and database development. His research field is based on information systems modelling for enterprise applications and processes interoperability, with applications in enterprise modelling. He is presently working on a conceptualisation approach and the use of Lattice theory through the Formal Concept Analysis (FCA) method for Enterprise Information Systems interoperability. The latest research works are on an improvement of the FCA method called Relational Concept Analysis (RCA) that focuses on tacit knowledge extraction through a clusterization approach on multi contexts data structures. He is moreover a regular reviewer for international journals (EIS, FGCS, CII, IJPR). He is a guest editor of special issues of international journals. He is the author or co-author of more than 75 papers in the field of Automation Engineering, Knowledge formalization and Enterprise systems interoperability. He manages the technical committee "Interoperability on Enterprise Network" (INE) inside the SAGIP association since 2021. He is a member of the IFAC Technical Committee 5.3 on "Enterprise Integration and Networking" since 2016 and the IFIP WP 12 since 2020.

Chiara Franciosi is Associate Professor at Université de Lorraine, France, since September 2023. She teaches at the Faculty of Sciences and Technologies in Nancy, and she carries out research activities at the Modelling and Control of Industrial Systems Department of the Nancy Research Center for Automatic Control (CRAN, UMR CNRS 7039). She graduated cum laude in Management Engineering from the University of Salerno, Italy, in 2014, andobtained her PhD in Industrial Engineering from the same University in 2019 with a thesis entitled "A Conceptual Framework for Measuring Maintenance Impacts on Sustainability". She won the Italian national prize for the best doctoral thesis on maintenance issues in 2020, awarded by the Italian Maintenance-Association (A.I.MAN). She spent 2 years and 9 months as a post-doc and contract lecturer at the Universityof Salerno. In 2022, she moved to CRAN to work as a post-doc in the European projects DIH4CPS, AI-PROFICIENT and MODAPTO. Her current research interests include the role of maintenance for sustainable manufacturing, the assessment of maintenance impacts on sustainability, the knowledge modelling in the field of prognostics and health management of production systems, the conceptualisation of a cognitive digital twin for industrialmaintenance, and the modelling of the circular manufacturing ecosystem.



Hervé Panetto is a Professor of Enterprise Information Systems at University of Lorraine. He teaches Information Systems modelling and development at TELECOM Nancy and conducts research at CRAN (Research Centre for Automatic Control), Joint Research Unit with CNRS where he is managing a research project on the use of neuro-symbolic AI for formalising models related to the interoperability of cyber-physical-social systems. He is a member of the Academia Europaea and a Fellow of the AAIA (Asia-Pacific Artificial Intelligence Association) and Fellow of the AIIA (Artificial Intelligence Industry Alliance). He received his PhD in production engineering in 1991. He has strong experience in information systems modelling, semantics modelling and discovery, and database development. His research field is based on information systems modelling for enterprise applications and processes interoperability. He is working on the cyber-physical systems smart interoperability with neuro-symbolic techniques and cognitive digital twins. He is expert at AFNOR (French National standardisation body), CEN TC310 and ISO TC184/SC4 and SC5. He participated in many European projects including IMS FP5-IST Smart-fm project (awarded by IMS) and the FP6 INTEROP NoE (Interoperability Research for Networked Enterprises Applications and Software). He is serving as expert-evaluator for the European Commission, FNR, AERES and ANR in the domain of ICT. He was visiting Professor in 2013-2015 in the frame of a Science Without Borders PVE project with PUC Parana, Brazil and visiting Professor in 2016 at the UTFPR, Curitiba, Brazil. He is editor or guest editor of books and special issues of international journals. He is author or co-author of more than 200 papers in the field of Automation Engineering, Enterprise Modelling and Enterprise systems integration and interoperability. From 2020 to 2023, he was Chairman of the IFAC French National Member Organization (NMO). After being Chair of the IFAC Technical Committee 5.3 "Enterprise Integration and Networking" from 2008 to 2014 and Chair of the IFAC Coordinating Committee 5 on "Manufacturing and Logistics Systems" from 2014 to 2020, he is now Vice-Chair of the IFAC Technical Committee 9.3 "Control for Smart Cities". He received the IFAC France Award 2013, the INCOSE 2015 Outstanding Service Award and the IFAC 2017 Outstanding Service Award. He is co-organiser of the yearly OTM/IFAC/IFIP EI2N workshop on "Enterprise Integration, Interoperability and Networking" and General Chair of the CoopIS series of conferences. He is Editor-In-Chief of the Annual Reviews in Control, Member of Computers In Industry, the International Journal of Computer Integrated Manufacturing, the International Journal on Universal Computer Science, the scientific journal Facta Universitatis, series Mechanical Engineering, member of the Advisory Board of the Digital Twin International Journal (DTIJ), and a Regional Associate Editor Europe of the international Journal of Intelligent Manufacturing (JIM), Springer, Associate Editor of the Enterprise Information Systems (EIS) journal, Taylor & Francis, the IEEE Internet of Things Journal, the Journal of Industrial Information Integration (JIII), Elsevier, the Engineering Applications of Artificial Intelligence (EAAI), Elsevier, and the Journal SN Computer Science (SNCS), Springer Nature. He is a supporter of ELLIS, and member of DAIRO, TAILOR and CLAIRE Networks.

Laurent Teste earned a Master's degree in Organism and Population Biology, with a specialization in Alpine Ecology, in 1985. The following year, in 1986, he completed a postgraduate diploma in Computer Science with dual competencies. From 1986 to 1999, Laurent worked as an analyst programmer and later as a project manager, focusing on the computerization of medical analysis laboratories. Since 1999, he has been serving as the Information Systems Manager (RSI) for the SNMSF. In 2021, Laurent started participating as a technical advisor in a number of CIFRE PhD, theses related to the usage of AI technologies in the ski e-commerce setting.

Diego Torres holds a PhD in Computer Science from the National University of La Plata (UNLP), Argentina and a PhD from the University of Nantes, France. He has been a researcher at the LIFIA research center since 2001. He is a full professor at the UNLP and a professor at the UNQ. His research interests are knowledge management, semantic web, knowledge graph, adaptive gamification, and applications in open science and citizen science. His main applications are linked to the use of data for decision-making in urban planning, agriculture, and requirements analysis