User-Centric Pattern Mining on Knowledge Graphs: an Archaeological Case Study

W.X. Wilcke[a,b,*], V. de Boer[b], M.T.M. de Kleijn[a], F.A.H. van Harmelen[b], H.J. Scholten[a]

[a] *Department of Spatial Economics, Vrije Universiteit, Amsterdam, The Netherlands*
[b] *Department of Computer Science, Vrije Universiteit, Amsterdam, The Netherlands*

Abstract

In recent years, there has been a growing interest from the digital humanities in knowledge graphs as data modelling paradigm. Already, this has led to the creation of many such knowledge graphs, many of which are now available as part of the Linked Open Data cloud. This presents new opportunities for data mining. In this work, we develop, implement, and evaluate (both data-driven and user-driven) an end-to-end pipeline for user-centric pattern mining on knowledge graphs in the humanities. This pipeline combines constrained generalized association rule mining with natural language output and facet rule browsing to allow for transparency and interpretability—two key domain requirements. Experiments in the archaeological domain show that domain experts were positively surprised by the range of patterns that were discovered and were overall optimistic about the future potential of this approach.

Keywords: User-Centric, Pattern Mining, Generalized Association Rules, Knowledge Graphs, Digital Humanities, Archaeology

1. Introduction

Digital humanities communities have shown a growing interest in the *knowledge graph* as a data modelling paradigm [1]. Already, this interest has inspired several large-scale international projects—amongst which are Europeana[1], CARARE[2], and ARIADNE[3]—to actively explore the creation and publication of knowledge graphs in their respective domains. These knowledge graphs, and many others like them, have been made available as part of the *Linked Open Data* (LOD) cloud—a vast and internationally distributed network of heterogeneous knowledge—bringing large amounts of structured data within arm's reach of humanities researchers, who are now looking for ways to analyse this wealth of knowledge. This presents new opportunities for data mining [2].

Data mining is the process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [3]. These patterns describe regularities in a dataset which can help researchers gain more insight into their data. Researchers can then use this insight as a starting point to form new research hypotheses, as support for existing ones, or simply to get a better understanding of their data [4]. This entire process can take weeks or even months of hard work in the traditional setting. However, by incorporating data mining into the workflow, much of this time can be saved through the automatic discovery of potentially-relevant patterns. This makes data mining interesting as a support tool for humanities researchers.

Of course, the idea of using data mining as a support tool in the humanities is, in itself, not novel. There have been various attempts before, for instance to classify coins [5] or to cluster cultural heritage [6]. However, the majority of these studies involved mining unstructured data, most commonly in the form of text mining, whereas mining struc-

*Corresponding author
Email addresses: `w.x.wilcke@vu.nl` (W.X. Wilcke), `v.de.boer@vu.nl` (V. de Boer), `mtm.de.kleijn@vu.nl` (M.T.M. de Kleijn), `frank.van.harmelen@vu.nl` (F.A.H. van Harmelen), `h.j.scholten@vu.nl` (H.J. Scholten)

[1] See `www.europeana.eu`

[2] See `www.carare.eu`

[3] See `www.ariadne-infrastructure.eu`

tured data has thus far been largely limited to tabular data and tailored to specific use cases. With the growing popularity of knowledge graphs in the humanities, mining patterns from these structures becomes ever more important to researchers in this domain.

This work present the *MINing On Semantics* pipeline (MINOS) for pattern mining on knowledge graphs in the humanities. Its aim is to support domain experts in their analyses of such knowledge graphs by helping them discover useful and interesting patterns in their data. To this end, the MINOS pipeline places users in the centre by letting them guide the mining process towards their topics of interest and by letting them focus the results via a facet pattern browser.

Under the hood, MINOS employs generalized association rule mining (ARM). An association rule is an implication of the form $X \implies y$, where the presence of a set of items X implies the presence of another item y. These implications are learned by iterating over a dataset of examples, called transactions. Generalized ARM works largely the same, except that the antecedent X holds the item classes rather then the items themselves.

This method was specifically chosen to help overcome two key issues with technological acceptance in the humanities, namely transparency and interpretability [7, 8]. With transparency, we refer to the ease with which a method and its underlying theory can be understood: a black box method, for example, is less transparent than a glass box one. With interpretability, we mean how easily one can interpret the results of a method: it is, for instance, typically more difficult to interpret an n-order tensor than it is to interpret a set of symbolic statements.

Generalized ARM satisfies both of these constraints: it employs basic statistical know-how to produce human-readable rules in an overall deterministic process. A limited background in statistics, which most humanities researchers possess, therefore already suffices to understand how these rules map back to the input data and to check whether they are valid. This allows humanities researchers to put their trust in both the method and its results [9].

Of course, this trust is only gained if the produced rules provide useful and interesting patterns which can help these researchers to get a better understanding of their data. We call this the *effectiveness* of the approach. To assess this effectiveness we conducted experiments in the archaeological domain, specifically on data from various excavations, during which domain experts were asked to evaluate a set of candidate rules on interestingness.

By placing domain experts at the centre of both the pipeline and its evaluation, as opposed to data scientists, we contribute to an as yet largely unexplored niche in this intersecting field of data mining, knowledge graphs, and humanities. Concretely, our main contributions are 1) insight into some of the challenges and possible solutions for introducing data science tools to the humanities, 2) a pipeline design for pattern mining on knowledge graphs which is tailored to domain experts rather than to data scientists, and 3) a user-driven evaluation of our design choices instead of only a data-driven one.

With these contributions, our research aims to add to the interdisciplinary field of the Digital Humanities. For this reason, we will refrain from developing an ARM algorithm from scratch, but instead focus on how we can augment such an algorithm with complementary components to make it into an effective tool for Humanities researchers.

A concise overview of related work is given next, followed by an overview of the pipeline, the dataset, and the experimental setup. This paper then discusses the results from the user-driven evaluation, and concludes with a reflection on the chosen approach in light of these results.

2. Related Work

Studies on data mining in the humanities have thus far largely focussed on unstructured data (text mining), whereas data mining on semi-structured or structured data has been explored less frequently [4, 10]. An example of the latter kind is discussed in [11], in which the authors propose mining association rules from excavation data—sites as transactions, artefacts as items—using the proven Apriori algorithm. This task is similar to that described in this work, but it is executed on a relational database rather than a graph-shaped knowledge base.

Rule mining on (graph-shaped) knowledge bases can take many different forms. Initial efforts primarily focused on Inductive Logic Programming (ILP), due to its natural fit to logic-based systems [12]. Performance issues led to the development of derivatives. Most well-known is arguably AMIE [13], whose main difference is its use of the

Partial Completeness Assumption to cope with the lack of negative examples. In either case however, the rule-generation process is more complex than that of traditional ARM, making it less transparent for non-experts and thus a less suitable choice for the humanities domain.

A handful of studies have focussed on applying ARM algorithms to knowledge graphs. The most straightforward approach simply converts all triples to transactions (as used in traditional ARM) and then feeds these to the Apriori algorithm [14]. This has the downside however, of 1) forcing relational data into an unnatural shape and potentially losing information in the process, and 2) limiting the possible exploitation of both relations (via the graph's structure) and semantics [15]. Fortunately, these caveats can largely be avoided by using an ARM algorithm that is specifically tailored to knowledge graphs.

Several of these tailored ARM algorithms exist, for instance SWApriori; an adaptation of the common Apriori algorithm to knowledge graphs [16]. Its main selling points are its ability to discover patterns which span over multiple triples, i.e., a path, rather than over just a single one, and that it can mine multi-relational patterns. However, all semantic information is disregarded early on for efficiency reasons, hence reducing the dataset to a directed graph.

An alternative that does address this information is SWARM [17], which exploits `RDF` and `RDFS` semantics to generalize patterns. Hereto, SWARM uses the `rdf:type` relation to infer the classes of every resource (item) in a set X, and then computes an inheritance tree for all resources in X using the `rdfs:subClassOf` relation. Resources in the same branch are grouped under their top-most common class, and are therefore said to share the same patterns.

SWARM's ability to exploit common semantics for generalization is unique amongst ARM algorithms for knowledge graphs[4]. For this very reason, we have chosen SWARM for our implementation of the MINOS pipeline (see Section 3).

Other alternatives are discussed in [19] (aligning locations), [20] (constructing ontologies), and [21] (mapping categories), but these are designed to fit a specific task and are therefore less applicable to this work.

Common amongst all these algorithm however, is that they treat resources as items—the things we are trying to find implications between. Another commonality is the focus on the graph's structure—its vertices and edges—whereas the values of its literals are left unaddressed. Therefore, similar values are still treated as completely different items.

Moving from the level of ARM to that of the pipeline, we can draw parallels between the approach presented by Nebot&Berlanga [22, 23] and the MINOS pipeline introduced in this work. Similar to MINOS, the scope of the mining process is tailored to the users' interests provided via a user-defined pattern. However, where our pipeline encodes patterns as triples with optional unbound variables, Nebot&Berlanga ask users to construct a formal SPARQL query using an extended grammar. While this offers more flexibility than our approach, it also increases the difficulty of entering such patterns for anyone not familiar with SPARQL.

Parallels can also be drawn with the approach presented in [24, 25]. Here, the authors perform dimension reduction by eliminating duplicate item sets prior to mining, and by removing unwanted candidate rules afterwards. This latter step is, again, similar to the data-driven filter used in the MINOS pipeline, whereas the former step is implicitly dealt with by the uniqueness property of URIs. A final similarity between both pipelines is the use of *generalized* association rules.

A last but nevertheless important distinction concerns the evaluation of candidate rules: none of the cited work so far has gone beyond a data-driven (objective) evaluation, despite its well known limitations for assessing the interestingness of patterns [26, 27]. Instead, this interestingness can only be assessed properly by combining a data-driven evaluation with a user-driven one. In addition to the common data-driven evaluation, we have therefore asked the local archaeological community to assess a set of candidate rules via an online survey (see Section 5.2).

3. The MINOS pipeline

The MINOS pattern mining pipeline combines an off-the-shelf ARM algorithm with a simple facet rule browser, and a number of crucial pre- and post-processing components. These components enable users to integrate their interests into the process by restricting the search space beforehand, and by filtering the results afterwards. Hereto, the

[4] Note that ARM algorithms that exploit general semantics as background knowledge do exist, for example [18].

pre-processing components translate user-provided target patterns into SPARQL queries, use these queries to retrieve relevant resources from the LOD cloud, and perform context sampling to enrich them with additional information. The post-processing components constrain the mined rules based on the user-specified patterns, and present them to the user for evaluation using the facet rule browser. A flowchart of this structure is provided in Figure 1. The processes depicted in this flowchart will be discussed next.

3.1. Data Retrieval

To start the mining process, users are asked to provide a *target pattern* which describes their current interest. This pattern takes the form of one or more triples and serves as a language bias which restricts the search space to the relevant subgraph [28]. Each triple in this pattern can convey a semantic range by leaving variables uninstantiated: (`_, p, o`) to specify all entities for which (`p, o`) holds, (`_, p, _`) to specify all entities for which `p` holds, (`_, _, _`) to denote the entire graph, etc.

The next step is the automatic translation from the provided target pattern to a SPARQL `CONSTRUCT` query, This query is then used to construct an in-memory copy of the corresponding subgraph. For instance, the pattern [(`_, rdf:type, :Burial_Ground`) ∧ (`_, crm:contains, :Human_Remains`)] results in a graph which holds only instances of burial grounds at which human remains were found. We call these instances the *target entities*. Note that not all triples in a pattern have to be hard constraints: users can specify soft constraints as well. If, for instance, the second triple in the example pattern would be a soft constraint, then also *ContextFind* entities for which this (`p, o`)-pair is unknown are included in the resulting graph. Entities with conflicting relations however, are still excluded.

3.2. Context Sampling

To find only relevant patterns we must first accurately capture the target entities' semantic representations. These representations, which we call their *contexts*, include all information which might be relevant to them during the mining process. Hence, contexts serve a role similar to that of the "individuals" in tradition Data Mining. Capturing these contexts is done using *context sampling*, and involves supplementing all target entities with additional triples, retrieved from the original graph, which are directly or indirectly related to that entity.

In our implementation of the MINOS pipeline we chose a local-neighbourhood-based sampling strategy. This strategy was selected for its simplicity—it only requires a single parameter—and for its ability to produce good approximations without the need for background knowledge. This strategy assumes that the semantic representation of a target entity diffuses with distance: closely-related entities are more relevant than those further away in the graph. To reflect this, a local-neighbourhood strategy samples the neighbouring nodes of a target pattern up to a certain depth.

3.3. Rule Mining

At the heart of the pipeline lies the ARM algorithm. For the reasons explained before—the exploitation of `RDF` and `RDFS` semantics to generalize patterns—we have chosen SWARM for our implementation of the pipeline. To understand how SWARM operates, we will now first expand on our earlier brief introduction to ARM.

An association rule is an implication of the form $X \implies y$, where X is a finite set of items and y is a single item *not* present in X [29]. With generalized ARM, X is reduced to the set of item classes $C_1, C_2, \ldots, C_n$ where $C_i \in X$ if it covers at least $\theta\%$ of the items. An example of such a rule might be $\{Cemetery, BurialGround\} \implies HumanRemains$, implying that human remains are often found at both cemeteries and burial grounds.

SWARM extends the notion of item sets from sets of items with the same type to *semantic item sets*, which contain items (entities) that frequently share (`p, o`)-pairs. Instances of cemeteries and burial grounds, for instance, are likely to share the (`:contains, :Human_Remains`)-pair. Once all semantic item sets in the graph have been computed, SWARM proceeds by joining similar item sets into *common behaviour sets*. In our example case, a likely combination might occur with instances of crypts and catacombs. The rationale underlying this is the assumption that entities with largely overlapping contexts (sets of properties and instances within a set distance) are likely to also share other, as yet-unknown, (`p`, `o`)-pairs. Put differently: these entities follow the same pattern. To quantify this, SWARM introduces the *similarity factor*, which serves as a boundary that separates similar from dissimilar item sets. As a final

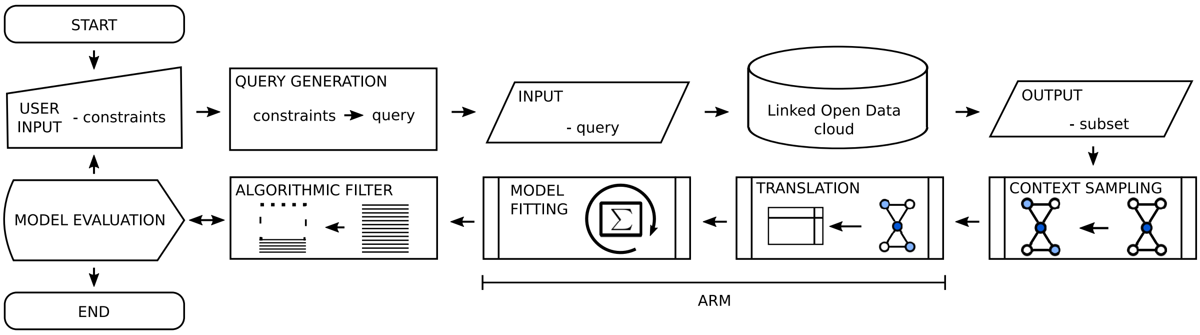


Figure 1: A flowchart of the MINOS pipeline. At start, a SPARQL query, automatically generated from a user-provided target pattern, is executed to retrieve a relevant subset from the LOD-cloud. After capturing the contexts of the target entities in this subset, these contexts are passed to the ARM algorithm. The resulting rules are then first filtered based on a predefined constraint template, with the remainder being presented to domain experts for evaluation.

step, SWARM generalizes the discovered patterns by exploiting class inheritance, ultimately producing rules of the form $\forall \chi (Class(\chi, c) \rightarrow (P(\chi, \phi) \rightarrow Q(\chi, \psi)))$. Here, χ, ϕ, and ψ are entities, c a class, and P and Q predicates.

There are several measures available to assess the interestingness of the resulting rules. For our data-driven evaluation we use the well known *confidence* and *support* measures. The confidence score of a rule $X \implies y$ conveys its strength and equals the proportion of triples (transactions) which satisfy X that also satisfy y. Its support score represents the rule's significance and equals the proportion of triples which satisfy both X and y.

3.4. Dimension Reduction

Association rule mining algorithms commonly produce a large number of candidate rules, resulting in their own knowledge management problem [30]. To combat this, our pipeline includes a data-driven filter which discards unwanted rules based on a predefined set of constraints. These constraints can include minimum and maximum values for the support and confidence scores, but also restrictions on types, on predicates, and on both entire antecedents and consequents.

By default, the set of constraints filters all rules which are too common or too rare, or which are generally unwanted. Typical examples of such unwanted rules are those which include domain-independent relations—`owl:sameAs`, `skos:inSchema`, `dcterms:medium`, etc—which do not contribute to the pattern mining process. Optionally, the default filter can be overridden by providing a custom constraints template at start. This allows us to further reduce the number of candidate rules by tightening the constraints.

Note that the filter is run *after* the rules have been produced. This is a deliberate design choice that allows users to revert any or all of its effects without having to repeat the whole mining process, and therefore offers more flexibility during the analysis of the results. Hereto, the unfiltered result set is kept in memory.

3.5. Rule Browser

Candidate rules which pass the filter are presented to users in an interactive facet browser. To improve their interpretability, these rules are automatically translated to natural language by exploiting the `label`-attributes of both entities and properties. The resulting translations are then inserted into predefined sentence templates; one per language.

To change which rules are shown, users can supply additional information about their interest. This involves modifying the same parameters as those available for the constraints templates discussed previously, and thus includes both restrictions on assigned scores and on contents. Rules which are deemed worthy of saving can be exported to a human-readable file, or can be stored in a binary file which can later be reopened by the rule browser for further analysis and mutation.

For our implementation of the pipeline we opted for a virtual-terminal interface, which offered us the necessary balance between simplicity and flexibility needed during our interactive sessions with domain experts. If desired however, a full-fledged web interface can be used instead by running the rule browser via a client-side environment.

4. The *package-slip* Knowledge Graph

Excavation data is a valuable source of information in many archaeological studies [9]. These studies are therefore likely to benefit from pattern mining on this type of data, and thus make it a suitable choice to base our case study on. In agreement with domain experts, we therefore selected the *package-slip* knowledge graph to run our experiments with.

Package slips are detailed summarizations of entire excavation projects. They are structured as specified by the *SIKB Protocol 0102*, which is a Dutch standard on modelling and sharing excavation data of various degree of granularity[5]. Specifically, package slips capture the following aspects:

- General information about individuals, companies and organisations which are involved, as well as the various locations at which the excavations took place.
- Final and intermediate reports made during the project, as well as different forms of media such as photos, drawings, and videos. Each of these is accompanied by meta-data and (cross) references to their respective file, subject, and mentioning.
- Detailed information about all artefacts discovered during the project, as well as their (geospatial and stratigraphic) relation and the archaeological context in which they were found.
- Fine-grained information about the precise locations and their geometries at which artefacts were discovered, archaeological contexts were observed, and where media was created.

In its specification, the SIKB protocol makes use of the Extensible Markup Language (XML) to define the various concepts that make up a package slip. The limitations of this language for sharing and integrating data motivated the Dutch ARIADNE partners, amongst whom were we, to design an alternative data model based on semantic web standards.

The semantic package slip is primarily built on top of the CIDOC conceptual reference model (CRM) and its archaeological extension CRM English Heritage (CRM-EH). Every excavation project is of the type `DHProject` (*Dutch Heritage Project*, a subclass of CRM-EH's `EHProject`) and links to the discovered artefacts via production events. To classify these artefacts, and all related archaeological contexts, the package-slip graph includes a multi-schema `SKOS` vocabulary which contains more than 7.000 archaeological concepts.

A small and partial example of a package slip is depicted in Figure 2. There, a single artefact (`crmeh:ContextFind`) is shown with its archaeological context (`crmeh:Context`). The artefact also holds several attributes of various types, and is itself part of a unit (`crm:Collection`). These units are *virtual* containers that group artefacts which were found in the same archaeological context. In contrast, bulk finds (`crmeh:BulkFind`) are *physical* containers, often a box, which group artefacts with similar storage requirements (e.g. on humidity, temperature, or oxygen). These are linked to a *Dutch Heritage* project via a production event.

4.1. Knowledge Integration

At present[6], the package-slip knowledge graph[7]contains the aggregated data from little over 70 package slips, totaling roughly 425.000 triples [31]. These package slips were converted from existing XML files[8], and were subsequently integrated into a single coherent graph by using a central triple store which acted as an authority server: hash values were generated for all entities in a package slip—starting from the outer edge of the graph and recursively updating all encountered parents—which were then cross referenced with those on the server. By using this approach, we were able to unify different resources about the same thing, most particular about people, companies, and locations.

The size of the package-slip knowledge graph is expected to grow consistently, as the production of package slips has recently been made mandatory for all Dutch institutes and companies which are involved with archaeological data. This increases the relevancy of this case study for the Dutch archaeological community, and thus adds support for our choice of the package slip.

[5] `www.sikb.nl/datastandaarden/richtlijnen/sikb0102`

[6] Situation on July 2018

[8] The conversion tool is available at `gitlab.com/wxwilcke/pakbonLD`

[8] `pakbon-ld.spider.d2s.labs.vu.nl`

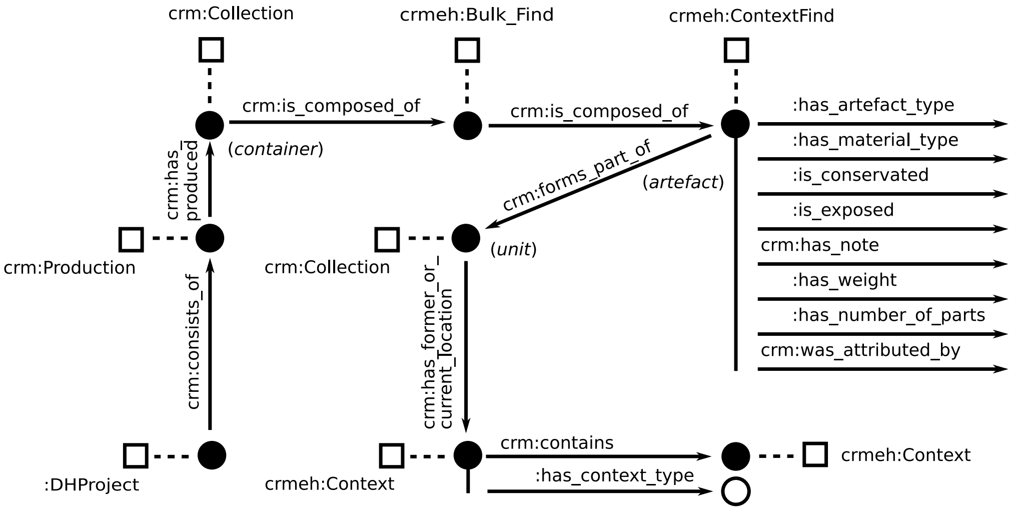


Figure 2: Small and partial example from the package slip knowledge graph. Each project produces (amongst others) one or more containers. These containers are composed of one or more bulk finds, which in turn are composed of one or more artefacts. Each of these artefacts is part of a unit of which its members share a common (multilayered) context. Note that, in this figure, solid and opaque circles represent instance and vocabulary resources, respectively. Three dots represent literals, and opaque boxes represent classes

5. Experiments

To assess the effectiveness of the MINOS pipeline, we have conducted four experiments on the package-slip knowledge graph (see Section 4). Each of these experiments addressed a different granularity of the package-slip graph to investigate the effects of these different granularities on the usefulness of the discovered patterns for domain experts. In order from coarse-grained to fine-grained, these are

Projects, which, amongst other, are of a project class, are held at a specific location, and during which one or more artefacts are discovered.

Artefacts, which, amongst others, are of an artefact class, have dimensions and mass, consist of one or more parts, and are found in a certain archaeological context and under specific conditions.

Archaeological Contexts, which, amongst others, are of a context class, have a geometry, a structure, and a dating, and which consist of one or more subcontexts.

Archaeological Subcontexts, which, amongst others, have a certain shape, colour, and texture, which consist of zero or more nested substructures, and which hold an interpretation of its significance as provided by archaeologists *in situ*.

The bottom three granularities—artefacts, contexts, and subcontexts—roughly correspond to the interests of three domain experts, with whom we had several interviews during the experiment design phase. The most coarse-grained granularity (projects) was added by us to create a balanced cross-section of the graph. From here on, we refer to these granularity levels as topics of interest (ToI).

All experiments were run using our implementation of the MINOS pipeline[9]. Hereto, all four of these experiments were given the same set of hyperparameters (Table 1). Concretely, we set local-neighbourhood context sampling depth (CSD) to a maximum of three hops, and varied the similarity factor (SF) over three values which range from weakly similar to strongly similar. Therefore, each experiment was run three times (CSD×SF) with different hyperparameters. Because different similarity factors *may* result in distinctly different, yet possibly useful, patterns, we aggregated their output to produce a single result set per experiment.

Each experiment produced up to roughly 1800 candidate rules, which was brought down from

[9] `gitlab.com/wxwilcke/MINOS`.

about 36.000 on average using the default filter settings. From these 1800 rules, we created a more manageable evaluation sample by taking the top fifty candidates based on their confidence (first) and support (second) scores. We motivate our choice of confidence as the primary measure due to the emphasis on producing correct, but also yet unknown, patterns.

Five candidate rules are listed in Table 2. These have been hand picked from the evaluation sample, and are representative of the result set as a whole. We will use these five candidates as running examples in our data-driven evaluation.

5.1. Data-Driven Evaluation

To assess the general effectiveness of the pipeline, we have chosen not to incorporate any dataset-specific adjustments into the data-driven evaluation. The reason for this is that we are not interested in how well our approach works on this particular dataset, but rather how effective the support and confidence metrics are as criteria for our data-driven filter and, by extension, for the perceived interestingness of the discovered patterns to domain experts. For the purpose of this evaluation, we limit ourselves to the evaluation sample created earlier.

An inspection of the evaluation sample reveals that many of the candidate rules apply to classes other than the four selected topics of interests. Two examples of such rules are listed in Table 2: both rules `R3` and `R4` apply to entities of the `SiteSubDivision` class, which is not explicitly listed in any of the target patterns. In fact, 16% of all rules in the evaluation sample apply to this class. If we look at the package-slip data model however, we observe that said class lies within three hops of the `DHProject` class. Therefore, the presence of these unexpected classes can likely be attributed to the chosen context sampling strategy. Indeed, the same is the case for all other unexpected classes such as `ContextStuff`, `Collection`, and `Place_Appellation`, which apply to 11.5%, 8%, and 6% of all rules in the evaluation sample.

Another look at the evaluation sample reveals that many of the candidate rules actually describe very similar patterns. This is especially evident with time spans, which are present in nearly half (43%) of the patterns. In fact, of the five examples listed in Table 2, all but one involves a time span as consequent. The reason for their frequent occurrence can likely be traced back to the package-slip vocabulary: many of the artefact classes are in a many-to-one relationship with specific time spans. For instance, all known belly amphora stem from the pre-classical period (612 BCE -– 480 BCE). These predefined relationships are treated as patterns by the ARM algorithm, and, due to overrepresented in the data, score high on confidence and support. This is also true for other predefined relationships, such as those that involve geographical information: map areas are paired with their respective toponymies, and provinces are paired with municipalities.

A final revelation of the evaluation set is that both support and confidence scores alone provide a poor measure of pattern interestingness: over 96% and 99% of the patterns have support and confidence scores close ($\delta \leq 0.05$) or equal to 0 or 1, respectively. This makes it difficult to distinguish between patterns which are actually interesting and those which describe predefined and therefore uninteresting relationships. While this difficulty is a known problem with ARM [30], the extreme form it takes here is likely caused by the nature of the package slip data: nearly all information that might be interesting to domain experts (artefacts, archaeological contexts, etc.) is in a one-to-one relationship with each other. Therefore, this information is not shared within, and also not between, package-slip instances.

5.2. User-Driven Evaluation

To assess the interestingness of the produced rules from a domain expert's point of view we asked the local archaeological community—a Facebook group with over 3000 members from various Dutch academic and commercial organizations—to participate in an online survey (Fig. 3). At the start of this survey, participants were greeted with a concise summary of this research and with a detailed description of the task we asked them to perform: to rate a selection of forty candidate rules on plausibility (*can this be true in the context of the data?*), on relevancy (*can this support you during your research?*), and on newness (*is this unknown to you?*). Once completed, the participants were also asked some final questions about their familiarity with the domain, and whether this approach of rule mining could contribute to their studies.

A five-point Likert scale was offered as answer template throughout the entire survey. To stress the importance of the participants' opinions our Likert scale ranged from *strongly disagree* to

Table 1: Configurations of the four experiments. Column names indicate topic of interest (ToI), target pattern, number of facts in the subset, context sampling depth (CSD), and similarity factors (SFs).

ToI	Target Pattern	# facts	CSD	SFs
Projects	[(_, rdf:type, :DHProject)]	21.9k	3	(0.3, 0.6, 0.9)
Artefacts	[(_, rdf:type, crmeh:ContextFind)]	192.8k	3	(0.3, 0.6, 0.9)
Contexts	[(_, rdf:type, crmeh:Context)]	82.2k	3	(0.3, 0.6, 0.9)
Subcontexts	[(_, rdf:type, crmeh:Context) ∧ (_, pbont:trench_type, _)]	59.5k	3	(0.3, 0.6, 0.9)

Table 2: Five example rules selected from the top-50 candidates of all four experiments. Each rule applies to resources of type **TYPE**, and consists of an antecedent and a consequent in the form of an **IF-THEN** statement. The last two columns list their confidence and support scores.

ID	Semantic Association Rule	Conf.	Supp.

R1	**TYPE** crmeh:ContextFind
	IF :has_artefact_type, :Adze
	THEN crm:has_time-span, [:Neolithic_Period, :Bronze_Age]
		1.00	0.08

R2	**TYPE** crmeh:ContextFind
	IF :has_artefact_type, :Raw_Earthenware_(Nijmeegs/Holdeurns)
	THEN crm:has_time-span, [:Early_Roman_Age, :Late_Roman_Age]
		1.00	0.90

R3	**TYPE** crmeh:SiteSubDivision
	IF :has_location_type, :Base_Camp
	THEN crm:has_time-span, [:Neolithic_Period, :New_Time]
		1.00	0.26

R4	**TYPE** crmeh:SiteSubDivision
	IF :has_location_type, :Flint_Carving
	THEN crm:has_time-span, [:Mesolithic_Period, :Neolithic_Period]
		0.50	0.06

R5	**TYPE** crmeh:Context
	IF :has_trench_type, :Esdek
	THEN :has_sample_method, :Levelling
		1.00	0.71

Table 3: Translation in natural language of the five example rules given in Table 2. The last three columns list their normalized median and (highest) mode scores for plausibility (P), relevancy (R), and novelty (N).

ID	Semantic Association Rule	P_{Mdn} (P_{Mode})	R_{Mdn} (R_{Mode})	N_{Mdn} (N_{Mode})
R1	*For every artefact holds: if it concerns an adze, then it dates from the Neolithic period to the Bronze Age.*	0.75 (0.75)	0.75 (0.75)	0.75 (0.75)
R2	*For every artefact holds: if it concerns raw pottery Nijmeegs/Holdeurns, then it dates from Early Roman to Late Roman period.*	0.75 (0.75)	0.50 (0.75)	0.25 (0.25)
R3	*For every site holds: if it concerns a base camp, then it dates from the early Neolithic period to Recent times.*	0.25 (0.25)	0.25 (0.25)	0.50 (0.75)
R4	*For every site holds: if it concerns flint carving, then it dates from the Mesolithic to Neolithic period.*	0.25 (0.25)	0.25 (0.25)	0.25 (0.25)
R5	*For every context holds: if it concerns an esdek, then the collection method involves levelling.*	0.25 (0.25)	0.25 (0.25)	0.75 (0.75)

strongly agree, where our questions were stated such that a higher agreement corresponds with a higher interestingness. Additionally, if desired, participants could enter further remarks as free-form text.

All forty candidate rules were randomly sampled from the evaluation sample—ten per experiment (stratified)—and automatically[10]translated to natural language using predefined language templates (see Section 3.5). We have listed five of these translated rules in Table 3. Each of these rules is the translation of the rule with the same identifier in Table 2.

5.2.1. Survey Analysis

Twenty-one people participated in our survey. A statistical analysis of their answers is provided in Table 4. For each of the four experiments, this table lists the normalized median and mode scores for plausibility, relevancy, and novelty. Also listed are the overall scores per experiment and per metric, and the score for the entire approach as a whole. For all scores holds that higher is better.

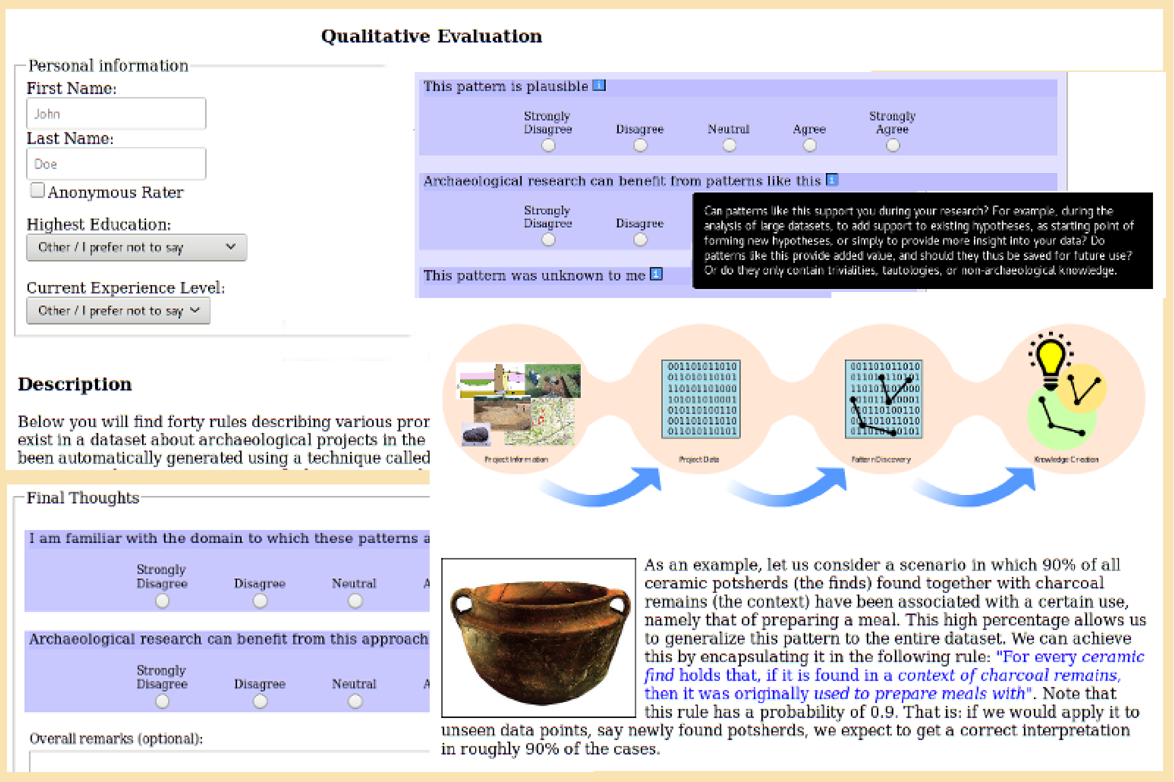
Looking at the four separate experiments it is clear that project and artefact related patterns are rated higher than those of contexts and subcontexts, with an overall median score of `0.50` versus that of `0.25`, respectively. Patterns about artefacts also score relatively high on the mode (`0.75`), exceeding that of all other experiments (`0.25`). This implies a cautious positive opinion towards artefacts related patterns. Contributing to this positiveness are the plausibility and relevancy ratings, on which artefacts score better than any of the other three experiments. Statistical tests (Kruskal-Wallis with $\alpha = 0.01$) indicate that these findings are significant, and suggest a dependency between various metrics and the topics of interest: $p = 7.66 \times 10^{-13}$ for plausibility, $p = 6.97 \times 10^{-5}$ for relevancy, and $p = 1.50 \times 10^{-3}$ for novelty.

The approach as a whole scores best on plausibility and novelty, both of which have a median of `0.50`. Between these two metrics, plausibility takes the cake with a mode of `0.75` versus that of `0.25` for novelty. This implies a cautious positive opinion towards the plausibility of the discovered patterns. More negative are the relevancy ratings with a median and mode of `0.25`. Together, the metrics combine to an overall score with median `0.50` and mode `0.25`, implying a neutral to slightly negative opinion about the approach as a whole. Remarks made by domain experts suggest that this stems from the frequent occurrence of patterns that are either too general, too trivial, or which describe predefined one-to-one or many-to-one relationships. This corresponds with our findings during the data-driven evaluation. Nevertheless, a median and mode score of `0.75` was given to the final separate question about the overall potential usefulness of this approach, implying a more positive standpoint.

Table 5 lists the inter-rater agreements per experiment and per metric. The Krippendorff's alpha (α_K) is used for this purpose, as it allows comparisons over ordinal data with more than two raters. Overall, the ratings are in fair agreement with $\alpha_K = 0.25$. A similar agreement is found between the different topics of interest, which range from 0.18 to 0.30. A more extreme difference is found between the different metrics: a moderate agreement on plausibility ($\alpha_k = 0.41$), and only a slight agreement on both relevancy ($\alpha_K = 0.08$) and novelty ($\alpha_K = 0.09$). This stark difference may be caused by the different familiarities and experiences of the domain experts: whereas plausibility comes down to everyday archaeological knowledge, novelty (and in a lesser degree: relevancy) is far more dependent on one's own view of the domain.

We can get a better understanding of these scores by reading the remarks that have been left by the domain experts. Overall, these remarks suggest a disconnect between how the evaluation task was instructed and how the experts performed it: rather than assessing the patterns within the (limited) context of the data set, our panel of experts appear to have judged the patterns against the knowledge in the archaeological domain as a whole. Numerous patterns have therefore only been scored as implausible because they do not necessary hold outside of the data set. Similarly, remarks on relevance and novelty seems to indicate that the raters only assessed that *exact* pattern instance, and did not consider the potential of such patterns on different data sets. Interestingly, some experts appear to have used this limited window to try to understand the knowledge creation process of fellow archaeologists: whether, for instance, the choice of an unexpected class could indicate an alternate interpretation of the same facts.

[10] In some cases, minor edits were made to improve the flow of the sentence.



Qualitative Evaluation

Personal information

First Name: John

Last Name: Doe

□ Anonymous Rater

Highest Education: Other / I prefer not to say

Current Experience Level: Other / I prefer not to say

This pattern is plausible

Strongly Disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly Agree ○

Archaeological research can benefit from patterns like this

Strongly Disagree ○ Disagree ○

Can patterns like this support you during your research? For example, during the analysis of large datasets, to add support to existing hypotheses, as starting point of forming new hypotheses, or simply to provide more insight into your data? Do patterns like this provide added value, and should they thus be saved for future use? Or do they only contain trivialities, tautologies, or non-archaeological knowledge.

This pattern was unknown to me

Description

Below you will find forty rules describing various pror
exist in a dataset about archaeological projects in the
been automatically generated using a technique called

As an example, let us consider a scenario in which 90% of all ceramic potsherds (the finds) found together with charcoal remains (the context) have been associated with a certain use, namely that of preparing a meal. This high percentage allows us to generalize this pattern to the entire dataset. We can achieve this by encapsulating it in the following rule: "For every *ceramic find* holds that, if it is found in a *context of charcoal remains*, then it was originally *used to prepare meals with*". Note that this rule has a probability of 0.9. That is: if we would apply it to unseen data points, say newly found potsherds, we expect to get a correct interpretation in roughly 90% of the cases.

Final Thoughts

I am familiar with the domain to which these patterns a

Strongly Disagree ○ Disagree ○ Neutral ○ A

Archaeological research can benefit from this approach

Strongly Disagree ○ Disagree ○ Neutral ○ A

Overall remarks (optional):

Figure 3: Collage of various screenshots of the web survey used during the user-driven evaluation. Participants were first asked about their background and experience, and were then given a summary of this research and a detailed description of the task. Next, they were presented with 40 candidate rules and asked to rate these on interestingness. Afterwards, participants were also asked about their opinion on the pipeline as a whole.

Table 4: Normalized separate and overall median and (highest) mode scores for plausibility (P), relevancy (R), and novelty (N) per topic of interest as provided by 21 raters.

ToI	P_{Mdn} (P_{Mode})	R_{Mdn} (R_{Mode})	N_{Mdn} (N_{Mode})	Overall
Projects	0.50 (0.25)	0.25 (0.25)	0.50 (0.75)	0.50 (0.25)
Artefacts	**0.75 (0.75)**	0.50 (0.25)	0.25 (0.25)	**0.50 (0.75)**
Contexts	0.25 (0.25)	0.25 (0.25)	0.50 (0.75)	0.25 (0.25)
Subcontexts	0.25 (0.25)	0.25 (0.25)	0.50 (0.25)	0.25 (0.25)
Overall	**0.50 (0.75)**	0.25 (0.25)	0.50 (0.25)	0.50 (0.25)

Table 5: Inter-rater agreements (Krippendorff's alpha: α_K) over all raters per topic of interest (left) and per metric (right). Overall $\alpha_K = 0.25$.

ToI	α_K
Projects	0.22
Artefacts	**0.30**
Contexts	0.24
Subcontexts	0.18

Metric	α_K
Plausibility	**0.41**
Relevancy	0.08
Novelty	0.09

6. Discussion

Our analysis of the survey's results indicates that the panel of experts was cautiously positive about the plausibility of the produced patterns. This (slight) positiveness does not come as surprise, as association rules *describe* the actual patterns which exist in the data, rather than predict new ones. We can even further explain this observation by our decision to order the candidate rules on confidence—favoring accuracy above coverage—and because the package-slip knowledge graph only contains curated data. Given that this is the case however, we may wonder why the patterns in our evaluation sample were not rated even *more* positively on plausibility.

A possible reason is the presence of errors in the data, but these would likely be incidental and should therefore have only a minor effect on the mean plausibility score. A more likely reason is that we used a suboptimal set of hyperparameters during our experiments. Prime suspects are the chosen similarity factors, which determine how much groups of entities (the semantic item sets) are permitted to overlap before they are combined into more general groups (the common behaviour sets). Concretely, a too low value can result in overgeneralization: two or more item sets are erroneously attributed the same pattern. This results in the generation of rules which do not hold for all members of a set, despite a possible high confidence score implying the opposite, and which are thus likely to score poorly on plausibility in the eyes of domain experts.

We can solve this problem to a certain extent by increasing the similarity factor to a value at which only minor differences in set members are accepted. By doing so however, we risk trading one undesirable situation for another: rather than overgeneralizing, we might undergeneralize such that similar item sets are prevented from forming common behaviour sets. In the worst case, we might even end up joining only very large items sets—the similarity factor is covariant with set size—which have a near perfect overlap, hence favouring sets with (`p`, `o`)-pairs that frequently co-occur across the graph. Examples of such patterns were present in our evaluation set, most particular those that implied relationships between two vocabulary concepts (types of artefact and time spans, provinces and municipalities, etc.) which naturally belong together.

Unfortunately, the optimal value(s) for the similarity factor are unknown beforehand, which is why we varied over three different values—`0.3`, `0.6`, and `0.9`—during our experiments. As explained before, our choice for these specific values came forth from our preliminary findings, which showed that these different values result in distinctly different, yet potentially useful, patterns. Given the occurrence of both too general and too specific patterns in our evaluation set, however, we can now surmise that the outer two values were likely too low and too high, respectively. This suggest that there exists but a narrow range of optimal values in the similarity factor spectrum for which SWARM yields desirable patterns. Determining this optimal range would require a great deal of time and effort from our domain experts, unfortunately, which is why no further preliminary experiments have been run.

While the chosen hyperparameters thus seem to largely correlate with the plausibility score, our results suggests that it are the characteristics of the dataset which are correlated with the relevancy and the novelty scores. On both of these scores, our group of experts were less positive about the presented patterns. The remarks left reveal that many of these patterns were seen as trivial. Others were seen as tautologies or were only thought to be applicable in a specific context. These remarks support our findings made during the data-driven evaluation that the confidence and support metrics are ineffective measures for assessing the interestingness of patterns to domain experts. Both measures are strongly influenced by the size, variety, and quirks of the data. Peculiarities of the package slip data, specifically its inherent hierarchical structure, are likely the reason which made this issue manifest itself even more evidently in this work.

Another reason for the unsatisfying novelty and relevancy scores may be found in the similarity between patterns: many of the produced rules describe variations of the same pattern, with only a different vocabulary concepts in the consequent. One pattern, for example, might imply that some pots are made from red clay, whereas another pattern might imply that other pots are made from grey clay. Our dataset contained over 7.000 of these concepts separated into 22 categories (e.g., material and period). If left unchecked, it is therefore likely for similar patterns to make their way into the result set.

Additionally, many of the potentially interesting data points were encoded via textual or numerical literals. SWARM, as well almost all other methods in this field, lacks the ability to compare literals based on their raw values [15]. Instead, resources are only compared via their URIs, which are unique and therefore always spaced at equal distance from each other in the search space. Because of this limitation, our pipeline was unable to differentiate between closely (or distantly) related literals. This became especially apparent with geometries, measurements, and descriptions, all three of which are abundantly found in archaeological data.

A further reason for the novelty and relevancy scores may lie in with the ontology that is used in the package slip data model: many of the properties defined in this model are expressed via rather long paths (up to five hops). This characteristic is directly inherited from the used `CIDOC CRM` ontology, which specifies that entities and properties are to be linked via various events. Travelling along these long paths is computationally expensive, which is why we had decided to leave such paths unexplored in favor for the more local patterns.

Zooming out from the individual scores can give us an idea about the perceived overall effectiveness of the pipeline. The remarks left by the experts suggest that this effectiveness may have been influenced considerable by the size and scope of the dataset. Indeed, the relatively low number of integrated package slips might not have been enough to allow for well-generalized patterns to emerge, rather than the more-specific patterns which only make sense in a narrow context (e.g., within a single excavation). Similarly, the limited scope of the data—the package slips were supplied by just two companies, each with a particular specialisation—meant that experts with different specialisations (e.g., a different culture or period) might have had insufficient experience to evaluate the discovered patterns on the criteria we asked them to.

A final aspect worthy of noting are the relatively low number of raters we were able to muster, despite our greatest efforts, and the effect this may have had on the outcome of our evaluation: the influence of possible outliers is greater, and the findings are less generalizable beyond our chosen use case. An analysis of the survey's access logs indicates that many potential raters did not actually complete the survey, but instead gave up at an earlier point. From the remarks left by those who *did* complete the survey, we believe that this was likely due to a combination of there being too many patterns to evaluate, and the limited novelty and relevance that these patterns provided to them.

7. Conclusion

In this work, we introduced the user-centric MINOS pipeline for pattern mining on knowledge graphs in the humanities. With this pipeline, we aim to support domain experts in their analyses of such knowledge graphs by helping them discover useful and interesting patterns in their data. Our pipeline therefore emphasizes the importance of these experts and their requirements, rather than those of the usual data scientists. This has led to several design choices, most particular of which is the use of generalized association rules to overcome the lack of transparency and interpretability—two key issues with technological acceptance in the humanities—that persists with many other methods.

To assess the effectiveness of our pipeline we conducted experiments in the archaeological domain. These experiments were designed together with domain experts and were evaluated both objectively (data driven) and subjectively (user driven). The results indicate that the domain experts were cautiously positive about the plausibility of the discovered patterns, but less so about their novelty and their relevance to archaeological research. Instead, a large number of patterns were discarded by our experts for describing trivialities or tautologies. Nevertheless, on average, the experts were positively surprised by the range of patterns that our pipeline was able to discover, and were optimistic about the future potential of this approach for archaeological research.

During our research we encountered several challenges which limited the effectiveness of our pipeline. We were unable to address them at that time and therefore offer these as suggestions for future work. For the most part, these challenges concern the nature of the data and the inability of the mining algorithms to exploit this. Concretely, the inability to 1) exploit common semantics other than `RDF` and `RDFS` (e.g., `SKOS` and, in a lesser degree, `CIDOC CRM`), and 2) cope with knowledge encoded via literal attributes (rather than only via the graph's structure) which make up the majority of knowledge in the humanities.

Solving these challenges would unlock a wealth of additional knowledge which is currently left unused, and which can potentially lead to more useful and more interesting patterns which humanities researchers can use to further their research.

Acknowledgements

We wish to express our deep gratitude to domain experts Milco Wansleeben and Rein van 't Veer for their enthusiastic encouragement and useful critiques during the various steps that have lead to this work. We also wish to thank all domain experts who participated in our survey for their willingness to sacrifice their free time, and without whom we would not have been able to complete this research.

This research has been partially funded by the ARIADNE project through the European Commission under the Community's Seventh Framework Programme, contract no. FP7-INFRASTRUCTURES-2012-1-313193.

References

References

[1] M. Hallo, S. Luján-Mora, A. Maté, J. Trujillo, Current state of linked data in digital libraries, Journal of Information Science 42 (2) (2016) 117–127.

[2] A. Rapti, D. Tsolis, S. Sioutas, A. Tsakalidis, A survey: Mining linked cultural heritage data, in: Proceedings of the 16th International Conference on Engineering Applications of Neural Networks (INNS), ACM, 2015, p. 24.

[3] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, From data mining to knowledge discovery in databases, AI magazine 17 (3) (1996) 37.

[4] J. Hagood, A brief introduction to data mining projects in the humanities, Bulletin of the American Society for Information Science and Technology 38 (4) (2012) 20–23.

[5] C. Velu, P. Vivekanadan, K. Kashwan, Indian coin recognition and sum counting system of image data mining using artificial neural networks, International Journal of Advanced Science and Technology 31 (2011) 67–80.

[6] K. Makantasis, A. Doulamis, N. Doulamis, M. Ioannides, In the wild image retrieval and clustering for 3d cultural heritage landmarks reconstruction, Multimedia Tools and Applications 75 (7) (2016) 3593–3629.

[7] L. Manovich, Trending: The promises and the challenges of big social data, Debates in the digital humanities 2 (2011) 460–475.

[8] B. R. T. Röhle, Digital methods: Five challenges, in: Understanding digital humanities, Springer, 2012, pp. 67–84.

[9] H. Selhofer, G. Geser, D2.1: First report on users needs, Tech. rep., ARIADNE (2014).
URL `http://ariadne-infrastructure.eu/Resources/D2.1-First-report-on-users-needs`

[10] H. Kamada, Digital humanities roles for libraries?, College & Research Libraries News 71 (9) (2010) 484–485.

[11] H.-P. Kriegel, P. Kröger, C. H. Van Der Meijden, H. Obermaier, J. Peters, M. Renz, Towards archaeo-informatics: scientific data management for archaeobiology, in: International Conference on Scientific and Statistical Database Management, Springer, 2010, pp. 169–177.

[12] V. Tresp, M. Bundschus, A. Rettinger, Y. Huang, Towards machine learning on the semantic web., in: Ursw (lncs vol.), Springer, 2008, pp. 282–314.

[13] L. Galárraga, C. Teflioudi, K. Hose, F. M. Suchanek, Fast rule mining in ontological knowledge bases with amie+, The VLDB Journal 24 (6) (2015) 707–730.

[14] T. Anbutamilazhagan, M. K. Selvaraj, A novel model for mining association rules from semantic web data, Elysium Journal 1 (2).

[15] X. Wilcke, P. Bloem, V. de Boer, The knowledge graph as the default data model for learning on heterogeneous knowledge, Data Science (1) (2017) 1–19.

[16] R. Ramezani, M. Saraee, M. Nematbakhsh, Swapriori: a new approach to mining association rules from semantic web data, Journal of Computing and Security 1 (2014) 16.

[17] M. Barati, Q. Bai, Q. Liu, SWARM: An Approach for Mining Semantic Association Rules from Semantic Web Data, Springer International Publishing, Cham, 2016, pp. 30–43. `doi:10.1007/978-3-319-42911-3_3`.
URL `http://dx.doi.org/10.1007/978-3-319-42911-3_3`

[18] R. G. Miani, C. A. Yaguinuma, M. T. Santos, M. Biajiz, Narfo algorithm: Mining non-redundant and generalized association rules based on fuzzy ontologies, Enterprise Information Systems (2009) 415–426.

[19] T. Kauppinen, K. Puputti, P. Paakkarinen, H. Kuittinen, J. Väätäinen, E. Hyvönen, Learning and visualizing cultural heritage connections between places on the semantic web, in: Proceedings of the workshop on inductive reasoning and machine learning on the semantic web (irmles2009), the 6th annual european semantic web conference (ESWC2009), Citeseer, 2009.

[20] J. Völker, M. Niepert, Statistical schema induction, in: Extended Semantic Web Conference, Springer, 2011, pp. 124–138.

[21] J. Kim, E.-K. Kim, Y. Won, S. Nam, K.-S. Choi, The association rule mining system for acquiring knowledge of dbpedia from wikipedia categories., in: NLP-DBPEDIA@ ISWC, 2015, pp. 68–80.

[22] V. Nebot, R. Berlanga, Mining association rules from semantic web data, in: International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, Springer, 2010, pp. 504–513.

[23] V. Nebot, R. Berlanga, Finding association rules in semantic web data, Knowledge-Based Systems 25 (1) (2012) 51–62.

[24] R. G. L. Miani, E. R. H. Junior, Exploring association rules in a large growing knowledge base, International Journal of Computer Information Systems and Industrial Management Applications (2015) 106–114.

[25] R. G. L. Miani, E. R. Hruschka, Analyzing the use of obvious and generalized association rules in a large knowledge base, in: Hybrid Intelligent Systems (HIS), 2014 14th International Conference on, IEEE, 2014, pp. 1–6.

[26] K. McGarry, A survey of interestingness measures for knowledge discovery, The knowledge engineering review 20 (1) (2005) 39–61.

[27] A. A. Freitas, On rule interestingness measures, Knowledge-Based Systems 12 (5) (1999) 309–315.

[28] C. d'Amato, A. G. Tettamanzi, T. D. Minh, Evolutionary discovery of multi-relational association rules from ontological knowledge bases, in: Knowledge Engineer-

ing and Knowledge Management: 20th International Conference, EKAW 2016, Bologna, Italy, November 19-23, 2016, Proceedings 20, Springer, 2016, pp. 113–128.

[29] R. Agrawal, T. Imieliński, A. Swami, Mining association rules between sets of items in large databases, in: Acm sigmod record, Vol. 22, ACM, 1993, pp. 207–216.

[30] M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen, A. I. Verkamo, Finding interesting rules from large sets of discovered association rules, in: Proceedings of the third international conference on Information and knowledge management, ACM, 1994, pp. 401–407.

[31] X. Wilcke, H. Dimitropoulos, D16.3: Final report on data mining, Tech. rep., ARIADNE (2017).
URL `http://ariadne-infrastructure.eu/Resources/D16.3-Final-Report-on-Data-Mining`