Discovering Novelty Patterns from the Ancient Christian Inscriptions of Rome

GIANVITO PIO, FABIO FUMAROLA, ANTONIO E. FELLE, DONATO MALERBA, and MICHELANGELO CECI, University of Bari Aldo Moro

Studying Greek and Latin cultural heritage has always been considered essential to the understanding of important aspects of the roots of current European societies. However, only a small fraction of the total production of texts from ancient Greece and Rome has survived up to the present, leaving many gaps in the historiographic records. Epigraphy, which is the study of inscriptions (epigraphs), helps to fill these gaps. In particular, the goal of epigraphy is to clarify the meanings of epigraphs; to classify their uses according to their dating and cultural contexts; and to study aspects of the writing, the writers, and their "consumers." Although several research projects have recently been promoted for digitally storing and retrieving data and metadata about epigraphs, there has actually been no attempt to apply data mining technologies to discover previously unknown cultural aspects. In this context, we propose to exploit the temporal dimension associated with epigraphs (dating) by applying a data mining method for novelty detection. The main goal is to discover relational novelty patterns—that is, patterns expressed as logical clauses describing significant variations (in frequency) over the different epochs, in terms of relevant features such as language, writing style, and material. As a case study, we considered the set of *Inscriptiones Christianae Vrbis Romae* stored in Epigraphic Database Bari, an epigraphic repository. Some patterns discovered by the data mining method were easily deciphered by experts since they captured relevant cultural changes, whereas others disclosed unexpected variations, which might be used to formulate new questions, thus expanding the research opportunities in the field of epigraphy.

Categories and Subject Descriptors: H.2.8 [**Database Management**]: Database Applications—*Data mining*; H.3.7 [**Information Storage and Retrieval**]: Digital Library—*Epigraphy*

General Terms: Epigraphy

Additional Key Words and Phrases: Epigraphy, evolution discovery, novelty pattern mining

ACM Reference Format:

Gianvito Pio, Fabio Fumarola, Antonio E. Felle, Donato Malerba, and Michelangelo Ceci. 2014. Discovering novelty patterns from the ancient Christian inscriptions of Rome. ACM J. Comput. Cult. Herit. 7, 4, Article 22 (August 2014), 21 pages.
DOI: http://dx.doi.org/10.1145/2629513

1. INTRODUCTION

Epigraphs are inscriptions on buildings, monuments, walls, and jewels, representing invaluable cultural heritage resources that provide us with myriad useful information about our past. They play the role of "time capsules" by allowing us, for example, to shed light on otherwise undocumented historical events or to gain new knowledge about local laws and customs. Epigraphy also indirectly documents

This work is supported by the EU-funded project EAGLE (Europeana network of Ancient Greek and Latin Epigraphy), Grant agreement no. 325122.

Q1 Author's address: M. Ceci (corresponding author), University of Bari Aldo Moro; email: michelangelo.ceci@uniba.it.

DOI: http://dx.doi.org/10.1145/2629513

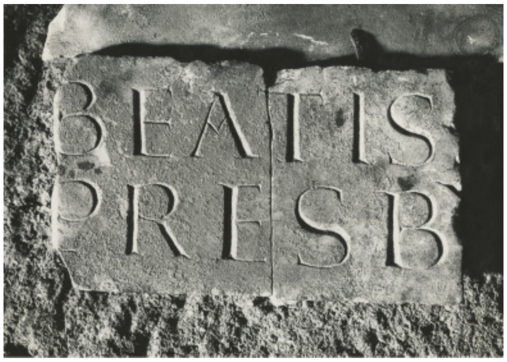


Fig. 1. An epigraph catalogued in the EDB repository (ICVR Volume X, Number 27272). It was found in the Basilica S. Valentino (Via Flaminia, Rome, Italy) and is dated between 376 and 399 AD. It is carved on marble, and its main function is dedicatory. Its transcription reported in EDB is *beatiss[imo martyri —] presby[ter — fecit?]*, which emphasizes the use of diacritical marks to represent the reconstruction of missing textual parts due to the damaging of the support. The picture of the epigraph is kindly provided by the Pontifical Commission for Sacred Archeology.

the evolution of languages and scripts. In some cases, such as that of the Rosetta Stone, it can provide key insights that allow the successful deciphering of an unknown script.

Epigraphic repositories[1] contain large corpora of pictures and the textual document representation thereof, which have been stored and annotated on several levels of interest [Fumarola et al. 2013]. They result from a huge effort made by scholars over decades to give public access to ancient epigraphs (Figure 1). Annotations stored in these databases are related to several aspects, such as dating and cultural contexts, style of writing, location, used material, and other interesting facets. They are inserted by scientists and validated using a peer review method. Although the validation process is strengthened by the knowledge and the experience of the performers, the annotation of old artifacts is still quite demanding due to the deterioration of the support, the ambiguity of abbreviations, and so on.

Recent information systems support epigraphists in various advanced tasks such as geolocalization of epigraphs, interpretation of damaged ancient documents, and transcription of the text. We argue that a new and still unexplored frontier of digital epigraphy projects is that of enabling automatic analysis of information currently stored in epigraphic repositories to extract implicit, previously unknown, and potentially useful knowledge from them. The application of data mining methods may help to reveal interesting relationships among, for instance, linguistic style, positioning, and dating of inscriptions, thus creating new links between different pieces of data. In general, it can help to organize large collections of inscriptions, introduce younger scholars to the field of epigraphy, and identify anomalies that can be explored using more traditional methods, as already done in computational historiography [Mimno 2012].

In this work, we focus our attention on a specific data mining task that can reveal useful information on the evolution of the inscriptions in terms of relevant features, such as language (e.g., the use of metrical texts, the presence of Latin and/or Greek words/symbols), writing style (e.g., minuscule, cursive, uncial), and material (e.g., marble), thus providing epigraphists with useful insights into important aspects of our history. In particular, we propose the application of a data mining algorithm for the identification of *novelty patterns* from annotations on time-stamped epigraphs. Novelty patterns describe a significant variation (in frequency) over time of some particular aspects regarding the

Q2 [1] See, for instance, http://www.eagle-network.eu; http://edh-www.adw.uni-heidelberg.de; http://www.edb.uniba.it; http://eda-bea.es; http://www.epigraphik.uni-hamburg.de; http://usepigraphy.brown.edu; http://www.edr-edr.it.

entities under study (the epigraphs, in this case). In this particular domain, discovering such patterns can significantly help domain experts in (1) finding empirical evidence that confirms hypotheses on (possibly already known) cultural changes and (2) making new hypotheses about previously unknown historical aspects, which are worth being further investigated.

The application of formal methods to the investigation of Latin epigraphs has already been proposed in the seminal work of Borillo [1984], which aimed at predicting unknown dating of epigraphs by analyzing a small amount of heterogeneous data extracted from 59 gravestones of North Africa Roman veterans. In this work, we advocate the application of data mining techniques for the purpose of automatically analyzing large collections of epigraphs, thus contributing to the scalability of formal approaches in epigraphy. In this work, we consider a collection of several thousands of epigraphs stored in Epigraphic Database Bari (EDP), an epigraphic repository, but similar analysis can be performed on other large collections of cultural heritage material, such as the Eagle repository (http://www.eagle-network.eu/) or the whole Europeana repository (http://www.europeana.eu/portal/) and other datasets of annotations extracted from it [Petras et al. 2012]. Moreover, although our focus is on a descriptive task—namely novelty patterns discovery—predictive tasks, such as those considered by Borillo, can enjoy similar benefits.

A specific challenge posed by the analysis of epigraphs is represented by the heterogeneity of available data, such as function, dating, material, textual transcription (with diacritical marks), *signa*, related bibliography, finding place, and conservation place. Among existing methods for novelty pattern discovery, very few are able to work with such heterogeneous data. Indeed, the typical two-way table (or database relation) used to represent a set of independent observations is inadequate for heterogeneous data, which requires more powerful representation formalisms, such as those typically used in relational data mining [Džeroski and Lavrač 2001; Ceci et al. 2007]. Relational data mining aims at designing and developing methods that are able to manipulate and analyze complex and structured data stored in a relational database, thus consisting of a collection of tables. This represents an extension with respect to classical data mining methods which operate on the simple attribute-value data representation. In relational data mining, each table can also be linked to other tables through the so-called foreign key constraints. These logical constraints specify how certain columns in one table can be used to look up information in corresponding columns in the other table, thus relating sets of records in the two tables.

To explain the advantages of the relational approaches, let us refer to the database scheme shown in Figure 2. In this case, a classical, nonrelational approach would be able to analyze the table *epigraphs*, but it would require the application of some aggregation strategies to consider related information stored in other tables (e.g., *included_terms*). For instance, we should consider "the number of terms in the epigraph" instead of the occurrence of a specific term, thus losing relevant information. On the contrary, a relational approach can navigate through the various tables by following the foreign key links, thus avoiding possible loss of information.

Here is an example of a novelty pattern that can be extracted by the relational method proposed in this work:[2]

—*epigraph(E), included_term(E, 'pace')*
$[\mathbf{350} - \mathbf{399}] : [0.25..0.41] \searrow [\mathbf{400} - \mathbf{449}] : 0.18 \quad \mathbf{GR} = 0.55$

This novelty pattern describes a decrease (↘) in the number of epigraphs containing the noun *pace* (peace) in the period from 400 to 449 with respect to the period from 350 to 399. Epigraphists take

[2] We assume that the reader is familiar with some basic notions of computational logic, such as term, atom, literal, clause, and substitution. Readers unfamiliar with this terminology should consult Lloyd [1987].

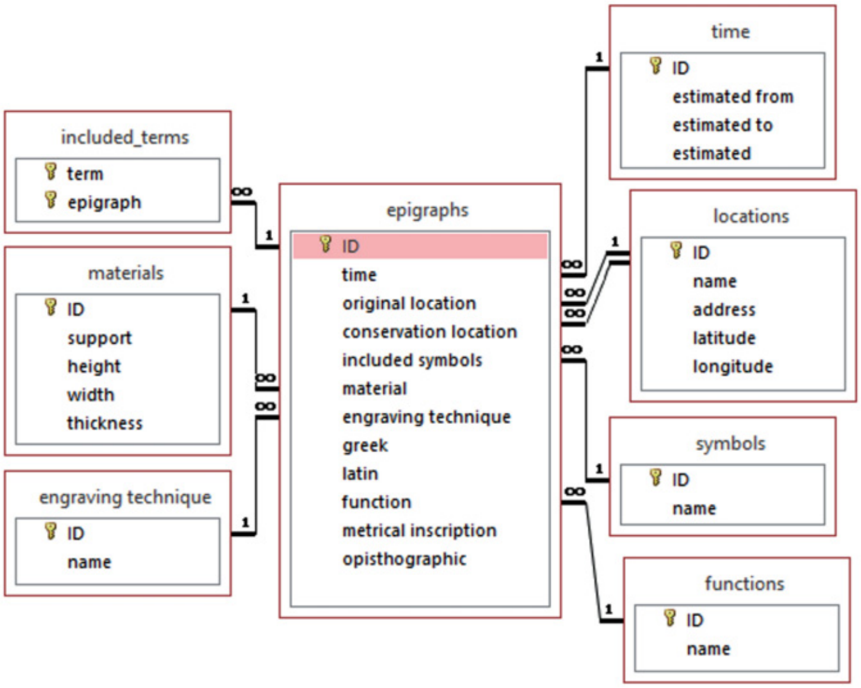


Fig. 2. Database schema used in the task of novelty detection.

some interest in this pattern, since a deeper analysis uncovers that, contrary to the expected evolution of the Latin noun *pax* into *pace*, during the first half of the 5th century, the nominative case (*pax*) reappeared in the inscriptions, after a period in which it was considered obsolete.[3]

The main contributions of this article are as follows:

(1) An innovative use of data mining technologies for the automatic extraction of implicit, previously unknown, and potentially useful knowledge from large collections of inscriptions currently stored in epigraphic repositories

(2) The proposal of a relational data mining method for novelty pattern discovery from heterogeneous data available in epigraphic repositories

(3) A pilot application to a repository of more than 20,000 Latin and Greek inscriptions, with the evaluation/interpretation of discovered patterns in collaboration with epigraphists.

The article is organized as follows. In the next section, we review the related work and briefly introduce the necessary background for this work. In Section 3, we present the proposed approach, describing the representation formalism, the problem definition, and the methodological solution that we propose. Section 4 describes the datasets and the experimental setup and then reports relevant results. Finally, in Section 5, we draw some conclusions and outline some future work.

[3] Henceforth, years will always refer to *Anno Domini*, unless otherwise specified. Moreover, the symbol ↗ (↘) will denote an increase (a decrease) in relative frequency. A more complete description of novelty pattern is provided in Section 3.2.

2. RELATED WORK AND BACKGROUND

In the literature, several information systems that support epigraphists in various advanced tasks have been presented. In particular, in Benefiel [2010], the authors propose a tight integration with a geographic information system to both geolocalize epigraphs and enable epigraph retrieval on the basis of (visual) spatial queries. The authors demonstrate the effectiveness of their solution on a database containing the considerable epigraphical corpus of wall *graffiti* and *dipinti* of Pompeii. Another example is DUGA, a Web information system that facilitates the interpretative process of damaged ancient documents [Roued-Cunliffe 2010]. The main motivations for the development of DUGA are the complexity of the reading process and the difficulty to record and recall how the final interpretation of the document was reached, as well as which competing hypotheses were presented, adopted, or discarded in the process of reading. DUGA employs the Decision Support System technology to facilitate the process of transcribing texts by providing a framework in which scholars can record, track, and trace their progress.

It is noteworthy that the literature contains no attempt to apply automatic tools for extracting knowledge from epigraphs (or epigraphic repositories), although several systems have been proposed for the analysis, also through data mining algorithms, of (general) digital cultural heritage resources and metadata. Typically, they have been developed in the context of research projects, such as MASTER [Le Bourgeois and Kaileh 2004], MEMORIAL [Antonacopoulos and Karatzas 2004], D-SCRIBE [Gatos et al. 2004a, 2004b], CULTURA [Agosti et al. 2013], PROMISE [Gäde et al. 2011], and COLLATE [Frommholz et al. 2003]. However, they neither are tailored to process epigraphs nor consider the temporal dimension which, as stated before, is of particular importance while studying cultural aspects of a historical period.

As for the specific data mining task of novelty detection, several methods have been proposed in the literature. They typically extract patterns that represent new or unknown situations that were never experienced before. In particular, Spinosa et al. [2008] propose a learning method that incrementally clusters data elements (along the time dimension) and identifies novelties with new clusters formed over time. Ma and Perkins [2003] propose to learn a regression function that reflects the normal behavior of a system and define *novelties* as those data elements that significantly differ from the prediction made by the regression function. Keogh et al. [2002] consider a different perspective on the problem and propose a method that discovers patterns whose frequency deviates from the expected value. Finally, Loglisci et al. [2013] propose to identify a sequence of novelty patterns (called *evolution chains*) from dynamic networks by exploiting the concept of "emerging pattern." A review of novelty detection methods is reported in Markou and Singh [2003].

Most data mining approaches to novelty detection do not work in the relational setting. The upgrade of existing novelty detection algorithms to the relational data mining framework is not straightforward, because along the time dimension, newly considered instances (possibly) create new relationships with both newly and previously considered instances. Therefore, one of the main contributions of this article is the proposal of a relational data mining method for the extraction of novelty patterns, represented in first-order logic (FOL), from time-stamped epigraphs.

Data heterogeneity, which characterizes our relational approach, forces us to distinguish between the *reference* (or *target*) table and the *task-relevant* (or *nontarget*) table. The former describes the main subject of the analysis (the reference objects), whereas the latter conveys useful information for the task at hand, as they are (directly or indirectly) linked to the reference table through foreign key constraints. Objects described exclusively by task-relevant tables are referred to as task-relevant objects. In this work, the reference objects are the epigraphs stored in the corresponding database relation, whereas the task-relevant objects are described by all other database relations, such as materials, engraving techniques, and locations (see Figure 2).

3. DISCOVERING NOVELTY PATTERNS

3.1 Data Representation

Data is collected through EDB (http://www.edb.uniba.it), a Web information system that supports epigraphists in storing, annotating, processing, and retrieving information on Christian inscriptions of Rome in Late Antiquity (from 200 to 600 AD). For storage functions, EDB relies on a PostgreSQL database, but only a portion of this database, containing relevant information for the analysis, is considered in the mining process. This portion is shown in Figure 2, where it is possible to identify the table *epigraphs*, which contains the target objects, and additional tables that are used to represent either the time dimension or additional (task-relevant) objects, which provide useful information for the analysis.

For the epigraphs, we consider the following features: *greek*, which indicates the presence of ancient Greek terms or digits; *latin*, which indicates the presence of Latin terms; *metrical inscription*, which indicates the presence of poetical verses; and *opisthographic*, which means that inscriptions are found on more than one side of the used support. Additional attributes represent the connection between epigraphs and other objects, such as locations (both original and conservation locations), symbols (e.g., Signa Christi, such as "crux quadrata"), function (e.g., mortuary), writing technique (e.g., engraved, depicted), material of the support (type, such as marble), and size of the support. The time dimension defines the temporal arrangement of inscriptions (each epigraph is associated with either an estimated or a real specific year).

The textual content is expressed as a set of terms in the table *included_terms*. This representation is equivalent to the classical bag-of-word representation of the text.[4] In particular, the transcription of each epigraph, which is originally stored as a single text field, is preprocessed by applying a whitespace tokenization and by removing stop-words, which, in this work, are limited to Roman numbers (e.g., XV). The removal of Roman numbers is motivated by the fact that in more than 95% of epigraphs, numbers are used to represent dates. Since epigraph dating strictly depends on the dates reported in the inscriptions, by explicitly representing them we would discover trivial patterns stating that the frequency of a specific date changes when the epigraph dating changes. On the contrary, we do not remove a priori small words and conjunctions, as these words also can be subject to the evolution of the language of four centuries.

To focus the analysis only on those terms that are actually relevant for the identification of novelty patterns, we select a subset of terms on the basis of $\chi^2(r, D)$ statistics, where r is a term and D is a subset of epigraphs from the original database. The range of $\chi^2(r, D)$ is $[0; +\infty)$. When the occurrence of r is independent of D—that is, its presence in a given document is not statistically related to the fact that such a document belongs to the subset D, then $\chi^2(r, D)$ is low and the term r can be discarded. On the contrary, by selecting the terms with the highest values of $\chi^2(r, D)$, only those that mostly depend on D are considered. Formally, $\chi^2(r, D)$ is defined as follows [Debole and Sebastiani 2003]:

$$\chi^2(r, D) = \frac{[P(r, D)P(\bar{r}, \overline{D}) - P(\bar{r}, D)P(r, \overline{D})]^2}{P(r, D)P(\bar{r}, \overline{D})P(\bar{r}, D)P(r, \overline{D})}, \quad (1)$$

where

—$P(r, D)$ is the probability that r occurs in an epigraph in D,
—$P(r, \overline{D})$ is the probability that r appears in an epigraph that does not belong to D,

[4] Latin and Greek diacritics can also be used.

—$P(\bar{r}, D)$ is the probability that r does not occur in an epigraph in D, and
—$P(\bar{r}, \overline{D})$ is the probability that r does not occur in an epigraph that does not belong to D.

In the classical χ^2-based feature selection, D is a class of documents. On the contrary, in our case we relate the set of epigraphs D to a time interval (this set is called *data block*, see Section 3.2). In this way, we select the most representative terms for inscriptions dated to a specific period. Since the $\chi^2(r, D)$ statistic of a term is computed for each considered time interval (D), with respect to all other time intervals ($\overline{D}$), we select the n terms that show the highest *average* value of the statistic.

Formally, the set $\mathcal{R}$ of considered terms is defined as follows:

$$\mathcal{R} = toparg_r\left(n, \frac{1}{|\mathcal{D}|}\sum_{D\in\mathcal{D}} \chi^2(r, D)\right), \tag{2}$$

where $\mathcal{D}$ is the set of time intervals and $toparg_r(n, f(r))$ is the function that returns the top n terms according to the function $f(r)$.

The filtering of the n features that maximize $f(r) = \frac{1}{|\mathcal{D}|}\sum_{D\in\mathcal{D}} \chi^2(r, D)$ in Equation (2) is a “global solution” for feature selection. This is an alternative to “local solutions,” which select the top $n/|\mathcal{D}|$ features for each $D \in \mathcal{D}$. In this work, we adopt a global solution because, as empirically proved in Debole and Sebastiani [2003], it is able to outperform local solutions, since the statistics that can be collected from scarcely populated categories (in our case, time intervals) are not robust enough for the local policy to be effective. In these cases, the global policy provides more robust statistics collected over the entire category set (in our case, the entire dataset of epigraphs).

3.2 Problem Definition

Our method is based on concepts developed in the area of temporal data mining, where the input data elements $\mathbf{a_1}, \mathbf{a_2}, \ldots, \mathbf{a_N}$ are associated with time points $\mathbf{t_1}, \mathbf{t_2}, \ldots, \mathbf{t_N}$ (where $\forall i = 1, \ldots, N-1, t_i \leq t_{i+1}$). In this work, data elements are epigraphs, and time points represent the dating of epigraphs.[5] In this way, according to time intervals, it is possible to define data blocks as follows:

Definition 3.1 (*Data block*). Given a time point t and a period p, a corresponding data block B is the list of epigraphs such that their associated dating is in $]t - p, t]$.

It is noteworthy that the period p of B is not related to the number of epigraphs in B. In the degenerate case, the data block B can be an empty list even for large values of p.

Definition 3.2 (*Data partitioning*). Input data elements $\mathbf{a_1}, \mathbf{a_2}, \ldots, \mathbf{a_N}$ are partitioned into k consecutive data blocks $\langle B_1, B_2, \ldots, B_k\rangle$ if and only if the following three conditions hold:

(1) for each pair of distinct blocks B_i and B_j, $i \neq j$, there exists no element $\mathbf{a_l}$, $l = 1 \ldots N$, occurring in both B_i and B_j;
(2) for each data element $\mathbf{a_l}$, $l = 1 \ldots N$, there exists a unique data block B_j, $j = 1 \ldots k$, where $\mathbf{a_l}$ occurs; and
(3) the time interval $]t_j - p_j, t_j]$ associated to B_j precedes the time interval $]t_{j+1} - p_{j+1}, t_{j+1}]$ associated to B_{j+1}.

Definition 3.3 (*Time window*). Let $\langle B_1, B_2, \ldots, B_k\rangle$ be a partition into consecutive data blocks of the data elements $\mathbf{a_1}, \mathbf{a_2}, \ldots \mathbf{a_N}$, and let w be a window size. Then, the time window $W(i, w)$ associated with each B_i ($i \geq w$) is the list of data blocks $\langle B_{i-w+1}, \ldots, B_i\rangle$.

[5] Time points are associated only with target objects and not other objects.

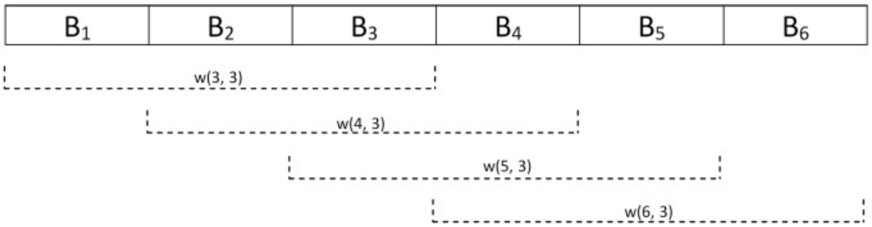


Fig. 3. Sliding windows model.

A time window defines the temporal horizon that we consider to mine relational patterns. However, we are interested in discovering a model that describes changes in data distribution. Therefore, we base our method on the concept of *sliding window*—that is, a fixed-size window that moves along the time dimension. When the sliding window moves forward, new blocks are considered, whereas old blocks are discarded, as shown in Figure 3. The sliding window model is used to discover changes in data distribution over time, thus it is appropriate for novelty pattern discovery.

In this work, the discovered patterns are relational, since they provide a generalization of

(1) properties of epigraphs (reference objects) and task-relevant objects, and
(2) relationships between (reference/task-relevant) objects.

An example of a relational pattern follows:

P2: *epigraph(E), conservation_location(E, L), address(L, 'Via Salaria'), ¬greek(E), function(E, F), name(F, 'Tit. Sepulchralis').*

It identifies (technically speaking, *covers*) any epigraph (denoted by the variable *E*) conserved in Salaria Street, without Greek terms and with function'sepulchralis (mortuary).[6] Here, the literal *conservation_location(E, L)* establishes a relationship between the epigraph and the location (represented by the variable *L*).

Operationally, the set of epigraphs *covered* by the preceding pattern can be found by performing the following query on the original database:

P2^{SQL}:

```
SELECT E.id
FROM epigraphs E, locations L, functions F
WHERE E.conservation_location = L.id AND L.address = 'Via Salaria' AND E.greek = false
   AND E.function = F.id AND F.name = 'Tit. Sepulchralis'
```

By generalizing, we say that an epigraph is *covered* by a relational pattern P if its identifier (see column *ID* in Figure 2) belongs to the result set of the query P^{SQL} associated with P. The number of distinct epigraphs covered by P (or, equivalently, the size of the result set of P^{SQL}) defines the *support* of P. Formally, we follow with the next definition.

Definition 3.4 (*Support*). Given a data block B_i and a pattern P, the *support* of P in B_i, denoted as $s_i(P)$, is the relative frequency of epigraphs covered by the pattern P in the data block B_i—that is,

$$s_i(P) = \frac{|covered(P, B_i)|}{|B_i|},$$

where $covered(P, B_i)$ is the set of epigraphs in B_i that are covered by the pattern P.

[6] In this work, we deal with classical negation, ¬. For the sake of simplicity, we adopt the usual notation *greek(X)* and *¬greek(X)* instead of *greek(X,true)* and *greek(X,false)*, respectively.

A relational pattern P is said to be *frequent* in B_i if its support exceeds a user-defined threshold $minSupp \in [0, 1]$. This notion is useful to define a relational pattern base.

Definition 3.5 (*Relational pattern base*). Given a time window $W(i, w)$, its *relational pattern base* $M(i, w)$ is the set of all relational patterns that are frequent in at least one data block in the time window $W(i, w)$:

$$M(i, w) = \{P \mid \exists j \in \{i - w + 1, \dots, i\} \quad s_j(P) > minSupp\}.$$

A *novelty pattern* is a relational frequent pattern that satisfies additional conditions, according to the following definition.

Definition 3.6 (*Novelty pattern*). Let

(1) $W(i, w) = \langle B_{i-w+1}, B_{i-w+2}, \dots, B_i\rangle$ be a time window composed of w data blocks, where the ending block is B_i;
(2) P be a pattern and $\langle s_{i-w+1}(P), s_{i-w+2}(P), \dots, s_i(P)\rangle$ the list of support values of P on each data block of $W(i, w)$; and
(3) $\Theta_P : [0, 1] \rightarrow \Psi$ be a discretization function that associates a support value of P in the interval $[0, 1]$ with a discrete value $\psi \in \Psi$ (the domain Ψ is defined by the discretization strategy in Section 3.3.3).

Then, P is a *novelty pattern* for the time window $W(i, w)$ if and only if $\Theta(s_{i-w+1}(P)) = \Theta(s_{i-w+2}(P)) = \dots = \Theta(s_{i-1}(P)) \neq \Theta(s_i(P))$.

The preceding definition states that the support of a novelty pattern P remains pretty stable for all blocks of the time window, except for the last one. For instance, the previously mentioned pattern $P2$ is a novelty pattern for the time window $\langle[288 - 299], [300 - 348], [349 - 380]\rangle$ if in the data blocks $[288 - 299]$ and $[300 - 348]$ its support remains in the interval $[0.62, 0.84]$, whereas in the data block $[349 - 380]$ its support drops or rises significantly, falling out of the interval $[0.62, 0.84]$. The change of the support associated with this time window is represented as follows:

$$P2 : [\mathbf{288 - 348}] : [0.62, 0.84] \nearrow [\mathbf{349 - 380}] : 0.85 \quad \mathbf{GR} = 1.16,$$

where *GR* (*Growth rate*) is the ratio between the support of the pattern in the last block and the average support in previous blocks of the same time window. Formally,

$$GR(P, W(i, w)) = \frac{s_i(P)}{\frac{1}{w-1}\sum_{j=(i-w+1)}^{i-1} s_j(P)}. \tag{3}$$

Henceforth, for the sake of brevity, when both the pattern and the time window are already specified, we write GR instead of $GR(P, W(i, w))$.

When $GR > 1$ ($GR < 1$), the pattern describes an increase (decrease) in the number of epigraphs that it covers in the last block, with respect to the number of epigraphs that it covers in previous blocks of the same time window. Obviously, the higher the deviation from 1.0, the more unexpected (and interesting) the corresponding pattern. As stated earlier, $\nearrow$ indicates an increase ($GR > 1$), whereas $\searrow$ represents a decrease ($GR < 1$) of the support of the pattern.

3.3 The Algorithm

Having clarified the type of patterns that we are interested in discovering, we now explain how to efficiently search in the large space of relational patterns. Search is based on the data stream mining algorithm Mr-NoDeS [Ceci et al. 2009], which looks for patterns as soon as a p-sized data block arrives in the stream. Only data blocks falling into a w-sized time window are considered by Mr-NoDeS. The algorithm has two steps. In the first step, the relational pattern base $M(i, w)$ is updated each time a data block B_i is considered, whereas in the second step, patterns in $M(i, w)$ are filtered out to keep only those that represent novelty patterns within the time window $W(i, w)$. In its original formulation, Mr-NoDeS works on generic relational databases where both reference and task-relevant objects arrive as a stream.

In this work, the main differences from the original proposal are threefold. First, Mr-NoDeS operates on a list of time-stamped epigraphs partitioned into data blocks of different size (i.e., the number of epigraphs in each data block can vary). Second, the time dimension is only considered for the target objects (epigraphs), whereas the set of considered task-relevant objects remains the same, independently of the time window under analysis. This difference is motivated by the fact that we do not need to model the evolution of task-relevant objects, but we need to model only the evolution of epigraphs. Third, since we do not work in a stream setting, we do not have to respect strict space and time constraints that lead to limit the number of observations to be analyzed.

Details of relational pattern discovery, pattern base maintenance, and novelty pattern detection are reported in the following.

3.3.1 *Relational Pattern Discovery.* The set of relational patterns can be partially ordered by a generality relation that helps to structure the space of patterns, as well as to design efficient algorithms for its exploration. The choice of the generality relation affects both the algebraic structure of the pattern space and the operators used to explore it. In Mr-NoDeS, the generality order, denoted as $\succeq_\theta$, is based on the θ-subsumption test [Plotkin 1970], which is easily mechanizable. Informally, given two patterns P_1 and P_2, $P_1 \succeq_\theta P_2$ indicates that P_2 can be obtained from P_1 by substituting terms to variables and by adding some literals. When $P_1 \succeq_\theta P_2$, we say that P_1 (P_2) is more general (specific) than P_2 (P_1). For instance, the following patterns:

P3 *epigraph(E).*
P4 *epigraph(E), conservation_location(E, L).*
P5 *epigraph(E), function(E, F).*

are ordered as follows: $P3 \succeq_\theta P4$, $P3 \succeq_\theta P5$.

Interestingly, the generality order $\succeq_\theta$ satisfies the monotonicity property with respect to the support—that is, $P_1 \succeq_\theta P_2$ entails that the support of P_1 is greater than or equal to the support of P_2. As a consequence, patterns more specific than a nonfrequent pattern cannot themselves be frequent. This property provides us a mathematically sound criterion to prune the search space and to gain in computational efficiency. Moreover, it also indicates that the space of relational patterns should be explored level by level, from the most general pattern (i.e., the pattern that contains only the *epigraph*$(\cdot)$ literal) to the most specific ones to benefit this pruning criterion.

Mr-NoDeS iterates between *candidate generation* and *candidate evaluation* phases, similarly to other relational frequent pattern mining algorithms [Mannila and Toivonen 1997; Appice et al. 2005]. In the candidate generation phase, the monotonicity property is used for pruning nonfrequent patterns from the next level; in the candidate evaluation phase, frequencies of candidates are computed with respect to the database.

To improve efficiency, in Mr-NoDeS, the space of candidate patterns is represented as a set of enumeration trees (*SE-trees*) [Ceci et al. 2008]. The idea is to impose an ordering on atoms such that all patterns in the search space are enumerated (Figure 4). Practically, a node g of a SE-tree is represented as a group comprising the *head* ($h(g)$), or the pattern enumerated at g, and the *tail* ($t(g)$), or an ordered set (according to some ordering criterion, such as lexicographic ordering) consisting of the

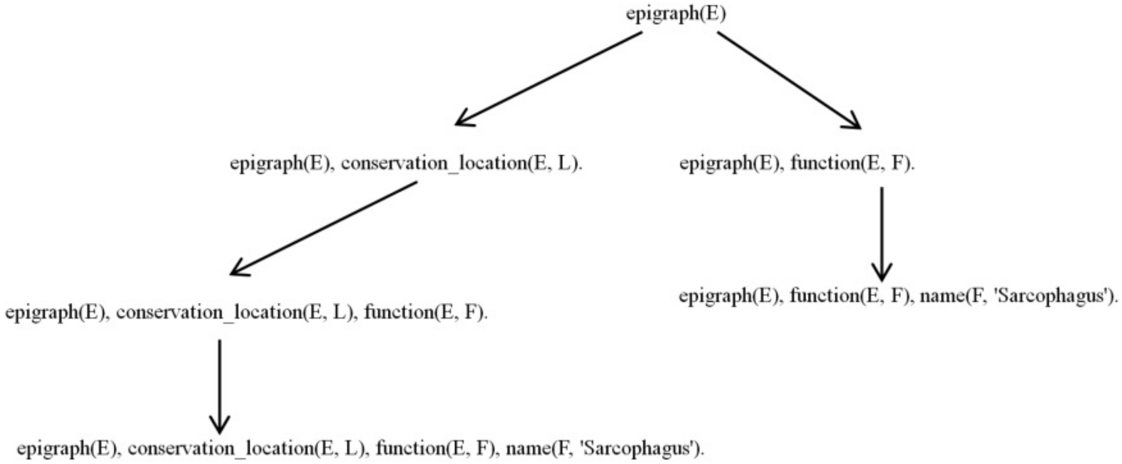

epigraph(E)
epigraph(E), conservation_location(E, L).
epigraph(E), function(E, F).
epigraph(E), conservation_location(E, L), function(E, F).
epigraph(E), function(E, F), name(F, 'Sarcophagus').
epigraph(E), conservation_location(E, L), function(E, F), name(F, 'Sarcophagus').

Fig. 4. An example of enumeration tree. Each node g shows only the head ($h(g)$). According to the linkedness property, the atom $name(F, \text{‘}Sarcophagus\text{’})$ can only be added after the variable F is introduced. Moreover, the lexicographical order of tails forces $conservation\_location(E, L)$ to appear before $function(E, F)$, if both present in the same pattern.

atoms that can potentially be appended to g to form a more specific pattern enumerated by some subnode of g. A child g_c of g is formed by taking an atom $i \in t(g)$ and appending it to $h(g)$. Its tail $t(g_c)$ contains all atoms that follow i in the ordered set in $t(g)$, and, when i is a structural predicate (i.e., a new relation is introduced in the pattern), additional new atoms that can be considered only after i has been appended. Indeed, the preservation of the linkedness property of a clause [Nienhuys-Cheng and de Wolf 1997] prevents the addition of atoms that do not share variables with other atoms already in g_c. Given this child expansion policy, without any pruning of nodes or pattern, the SE-tree enumerates all possible patterns and avoids generation and evaluation of candidates equivalent under $\succeq_\theta$ to some other candidates.

In Mr-NoDeS search proceeds until no candidate pattern is generated in the candidate generation phase. However, it is possible to anticipate the termination by specifying a maximum number of literals ($MaxNumLiterals$) in a pattern—that is, a maximum depth of the SE-trees. This is common to almost all relational data mining methods that tend to balance the complexity and the expressiveness of obtained patterns.

3.3.2 *Pattern Base Maintenance.* The discovery of frequent relational patterns is useful to build the pattern base $M(i, w)$, which is the set of relational patterns that are frequent on at least one block of the time window $W(i, w)$. By definition, for each pattern $P \in M(i, w)$, the list $\langle s_{i-w+1}(P), s_{i-w+2}(P) \ldots, s_i(P)\rangle$ of support values computed for P on each data block in $W(i, w)$ must include at least one value greater than the user-defined threshold for the support ($minSupp$).

The construction of $M(i, w)$ is incremental. It is obtained by updating $M(i-1, w)$ when the new data block B_i arrives. Updating may require the insertion of frequent patterns, the deletion of nonfrequent patterns, and the updating of the support list of patterns already in the base. We distinguish three cases:

(1) $\mathbf{i = 1}$. The pattern base $M(1, w)$ is generated from scratch, using the relational pattern discovery algorithm described earlier: $M(1, w) = \{P_j | s_1(P_j) \geq minSupp\}$.

(2) $\mathbf{i = 2, \ldots, \boldsymbol{w}}$. Relational patterns that are frequent on B_i are discovered and used to construct M_i^+, which is the set of patterns that are frequent on B_i but do not belong to $M(i-1, w)$. $M(i, w)$

is constructed by adding patterns of M_i^+ to $M(i-1, w)$. For each pattern $P \in M(i-1, w)$, the associated support list $\langle s_1(P), \ldots, s_{i-1}(P)\rangle$ is updated by adding $s_i(P)$. For each pattern $P \in M_i^+$, the entire support list $\langle s_1(P), \ldots, s_i(P)\rangle$ is built from scratch.

(3) $\mathbf{i} > \mathbf{w}$. The time window slides from $B_{i-w}, \ldots, B_{i-1}$ to $B_{i-w+1}, \ldots, B_i$, and the data block B_{i-w} is removed from memory. Relational patterns that are frequent on B_i are discovered and used to construct both M_i^+ and M_i^-. M_i^+ is the set of patterns that are frequent on B_i but do not belong to $M(i-1, w)$, whereas M_i^- is the set of patterns that are frequent on B_{i-w} but not in $B_{i-w+1}, \ldots, B_i$. $M(i, w)$ is then constructed from $M(i-1, w)$ by adding patterns of M_i^+ and removing patterns of M_i^-. The support list of each pattern $P \in M(i, w)$ is updated or created from scratch. An update operation is performed when the pattern P is already included in $M(i-1, w)$—that is, $P \in (M(i, w) - M_i^+)$. In this case, the support list associated with P is updated by removing $s_{i-w}(P)$ and by adding $s_i(P)$. A creation operation is performed when $P \in M_i^+$. In this case, the entire support list $\langle s_{i-w+1}(P), \ldots, s_i(P)\rangle$ is built.

It is noteworthy that Mr-NoDeS does not keep in memory the whole data stream, but only data elements in the last w data blocks, since this helps to construct the entire support list $\langle s_{i-w+1}(P), \ldots, s_i(P)\rangle$. This support list is necessary to compute the growth rate of novelty patterns according to Formula (3).

3.3.3 *Time WindowBased Novelty Pattern Detection.*

The pattern base $M(i, w)$ is used to identify novelty patterns for B_i, once the support values have been discretized. The discretization function Θ_P is computed by means of a clustering algorithm (see Algorithm 1). In particular, we use a variant of the density-based clustering algorithm DBSCAN [Sander et al. 1998]. The use of DBSCAN is motivated by its ability to discover arbitrary-shaped clusters with no prior information on the number of clusters, as well as its insensitivity to the ordering of input data.

For each pattern P, the cluster construction starts from a selected data block B_j in $W(i, w)$ called *seed* data block. Then, the neighborhood $N(B_j)$ of B_j is determined[7] and labeled as a cluster c only if it satisfies the following condition of forming a dense region:

$$stdev(S_P(N(B_j))) = \sqrt{\frac{1}{|N(B_j)|-1} \sum_{B_v \in N(B_j)} \left(s_v(P) - \left(\frac{1}{|N(B_j)|} \sum_{B_u \in N(B_j)} s_u(P) \right) \right)^2} \leq maxStD, \qquad (4)$$

where $S_P(N(B_j))$ is the set of support values of the pattern P associated with the set of data blocks $N(B_j)$, and $maxStD$ is a user-defined threshold on the standard deviation ($stdev$) computed on $S_P(N(B_j))$.

When a new cluster is identified (i.e., when Inequation (4) holds), it can then be expanded by merging it with partially overlapping neighborhoods, which are the neighborhoods of blocks belonging to the cluster c. Merging is performed only if the cluster that is obtained after merging still satisfies the condition of a dense region. If a cluster cannot be further expanded, a new seed is selected to evaluate the construction of a new cluster. The strategy adopted to select the seed is the sequential one. The cluster c is labeled with the interval $[minC, maxC]$, where $minC$ ($maxC$) is the minimum (maximum) support value falling in c.

Differently from Mr-NoDeS, in this work we apply the clustering procedure in batch mode—that is, we exploit the whole list of data blocks (and their support values) instead of incrementally analyzing them.[8] This is possible thanks to the low number of data blocks. The batch mode has a twofold

[7] In this work, $N(B_j) = \{B_{j-1}, B_j, B_{j+1}\}$.

[8] In the batch mode, the value k in Algorithm 1 represents the total number of data blocks.

ALGORITHM 1: Clustering algorithm for support values

```
Data: The blocks $B_1, B_2, \ldots B_k$; the support values $s_1(P), s_2(P), \ldots s_k(P)$ associated to the pattern $P$; a threshold
      $maxStD$
Result: A set of clusters built on the support values.
$\mathcal{C} \leftarrow \emptyset$ ;
repeat
    choose a seed data block $B_j$;
    build the neighborhood $N(B_j)$;
    if $stdev(S_P(N(B_j))) \leq maxSD$ then
        foreach $B_l \in N(B_j)$ do
            label $B_l$ as cluster $c$;
        end
        foreach $B_l$ labeled as $c$ do
            build the neighborhood $N(B_l)$;
            foreach $B_q \in N(B_l)$ do
                if $stdev(S_P(\{B_w : B_w \text{ is labeled as } c\} \cup B_q)) \leq maxStD$ then
                    label $B_q$ as $c$;
                end
            end
        end
    else
        label $B_j$ as a new cluster $c$;
    end
    $minC \leftarrow min\_support\_value(c)$;
    $maxC \leftarrow max\_support\_value(c)$;
    assign the label $[minC, maxC]$ to the cluster $c$;
    $\mathcal{C} \leftarrow \mathcal{C} \cup c$;
until all support values are associated to a cluster;
return $\mathcal{C}$;
```

advantage. First, we do not lose information on previous blocks due to the limited time horizon of the incremental approach. Second, we execute the clustering algorithm only once, thus gaining in computational efficiency.

When the clustering algorithm segments the time window $W(i, w)$ of P into only two clusters, namely c_1 and c_2, such that c_1 includes the support values for blocks $B_{i-w+1}, \ldots, B_{i-1}$ and c_2 includes the support value for the data block B_i, then P is marked as a novelty pattern for B_i over $W(i, w)$. In other words, patterns whose support value passes from one cluster to another in B_i are labeled as novelty patterns.

4. APPLICATION TO REAL EPIGRAPHIC REPOSITORIES

The effectiveness of the proposed method has been evaluated on epigraphs stored in the EDB repository, which currently contains more than 30,000 Latin and Greek inscriptions recovered in the area of Rome and published in the *Inscriptiones Christianae Vrbis Romae septimo saeculo antiquiores, nova series* (ICVR) editions (started in 1922).

For the identification of novelty patterns, the temporal dimension is strictly necessary. Thus, all epigraphs with unknown dating have been discarded. Moreover, only the temporal interval from 200 to 500 AD has been taken into account, since it contains most of the stored epigraphs. Indeed, considering also other irrelevant intervals (i.e., containing just a few epigraphs) would have affected the discretization process and/or the significance of the discovered patterns. This preprocessing step has led to a set of 14,874 valid epigraphs. Six data blocks have been defined according to two different

Table I. Dataset Obtained after the Application of EW Discretization (left) and EF Discretization (right)

		All	**Precise Dating**
Block	Interval	Epigraphs	Epigraphs
1	200–249	1,671	5
2	250–299	929	29
3	300–349	7,418	108
4	350–399	2,841	596
5	400–449	1,853	245
6	450–500	162	103

	All		**Precise Dating**	
Block	Interval	Epigraphs	Interval	Epigraphs
1	200–287	2,479	200–359	199
2	288–325	2,479	360–375	189
3	326–348	2,479	376–388	186
4	349–349	2,479	389–401	200
5	350–383	2,479	402–439	183
6	384–500	2,479	440–500	129

discretization strategies applied on the dating of epigraphs—that is, equal width (EW) and equal frequency (EF). The former partitions the whole dating interval into q intervals (bins) of near-equal size. The latter partitions the whole dating interval into q intervals such that each of them contains approximately the same number of epigraphs.

Experiments are performed on two distinct datasets (Table I): one includes all 14,874 valid epigraphs, whereas the other includes the subset of 1,086 epigraphs for which we have a precise dating. Indeed, many epigraphs have an uncertain dating, which is represented by a time interval. In these cases, epigraphs are associated with the median values of their corresponding intervals (estimated dating) to make the application of the method possible. A more precise analysis can be performed by considering only epigraphs with a precise dating, although the lower number of cases that support discovered patterns makes it more difficult to derive all robust findings.

On the basis of preliminary experiments and on the basis of epigraphists' suggestions, we run the algorithm with the following parameters settings: $MaxNumLiterals = 5$, $n = 200$, $w = 3$. For $minSupp$ and $maxStD$, the following configurations have been considered: $minSupp = 0.05$, $maxStD = 0.05$ and $minSupp = 0.1$, $maxStD = 0.1$. This is motivated by the strong dependency between the two parameters.

For all reported experiments, we performed both a quantitative and a qualitative evaluation. Whereas the former aims at measuring the number of the extracted novelty patterns and their significance, the latter aims at showing the usefulness of the patterns from the viewpoint of the domain expert.

To measure the significance of the extracted novelty patterns, we introduce the normalized average change rate (NACR), which represents the average change rate of a set of novelty patterns $\mathcal{P}$ extracted from the time window $W(i, w)$. This measure is defined as follows:

$$NACR(\mathcal{P}, W(i, w)) = \frac{1}{|\mathcal{P}|} \sum_{P \in \mathcal{P}} NCR(P, W(i, w)), \tag{5}$$

where the normalized change rate (NCR) of a single novelty pattern P is defined as follows:

$$NCR(P, W(i, w)) = \begin{cases} 1 - \frac{1}{GR(P, W(i, w))} & if\ GR(P, W(i, w)) \geq 1 \\ 1 - GR(P, W(i, w)) & if\ GR(P, W(i, w)) < 1. \end{cases} \tag{6}$$

Intuitively, the NCR measures how much the growth rate of a novelty pattern is distant from "no change." In particular, NCR (and, consequently, NACR) is defined in the range [0,1], where 0 means that there is no change in the frequency of the pattern (i.e., $GR(P, W(i, w)) = 1$), and 1 means that there is the strongest change, which happens when (1) $GR(P, W(iw, w)) = 0$ (i.e., the pattern totally disappears in the target block but is frequent in the previous blocks of the considered time window), or when (2) $GR(P, W(i, w)) = +\infty$ (i.e., the pattern becomes frequent in the target block).

Table II. Quantitative Analysis of Results for the EW Dataset

				All		Precise Dating	
minSupp/maxStD	**Blocks**	**Time Window**	**Target Block**	**Novelty Patterns**	**NACR**	**Novelty Patterns**	**NACR**
0.05/0.05	1,2,**3**	200–349	300–349	65	0.59	76	0.68
	2,3,**4**	250–399	350–399	87	0.84	46	0.66
	3,4,**5**	300–449	400–449	43	0.69	30	0.48
	4,5,**6**	350–499	450–499	84	0.83	87	0.54
Global				**279**	**0.75**	**239**	**0.60**
0.10/0.10	1,2,**3**	200–349	300–349	172	0.52	35	0.81
	2,3,**4**	250—399	350–399	96	0.59	0	—
	3,4,**5**	300–449	400–449	6	0.20	0	—
	4,5,**6**	350–499	450—499	259	0.67	49	0.65
Global				**533**	**0.60**	**84**	**0.71**

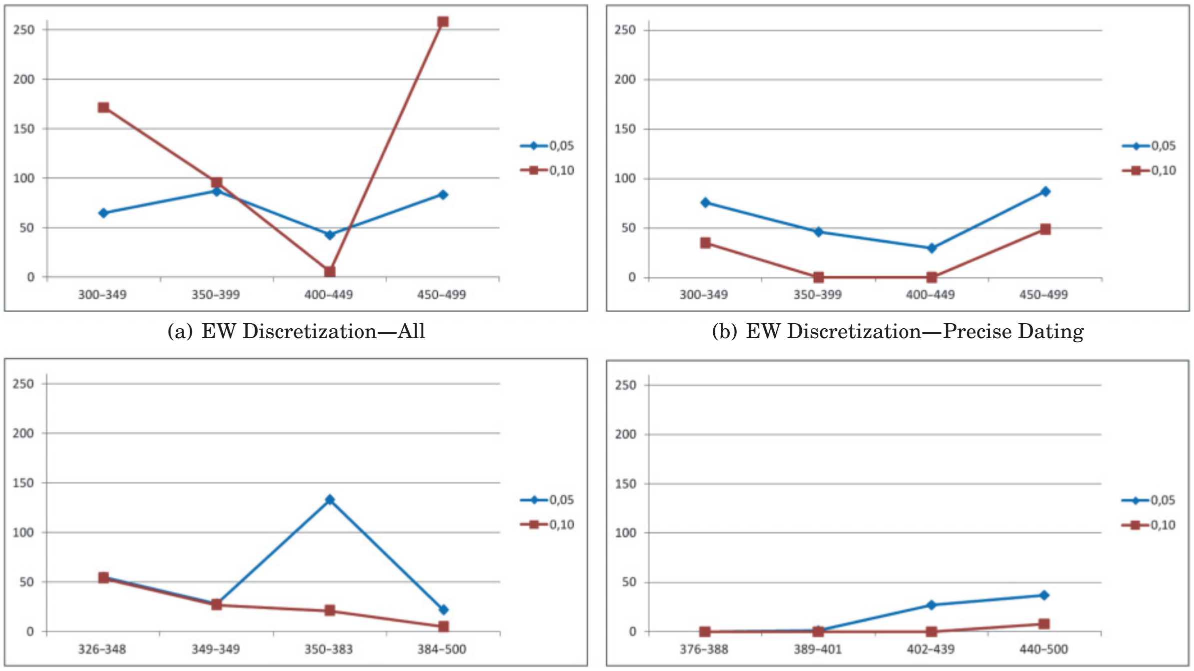
Table III. Quantitative Analysis of Results for the EF Dataset

		All				Precise Dating			
minSupp/ maxStD	**Blocks**	**Time Window**	**Target Block**	**Novelty Patterns**	**NACR**	**Time Window**	**Target Block**	**Novelty Patterns**	**NACR**
0.05/0.05	1,2,**3**	200–348	326–348	55	0.60	200–388	376–388	0	—
	2,3,**4**	288–349	349–349	28	0.59	360–401	389–401	1	0.24
	3,4,**5**	326–383	350–383	133	0.72	376–439	402–439	27	0.51
	4,5,**6**	349–500	384–500	22	0.38	389–500	440–500	37	0.37
Global				**238**	**0.64**			**65**	**0.42**
0.10/0.10	1,2,**3**	200–348	326–348	54	0.43	200–388	376–388	0	—
	2,3,**4**	288–349	349–349	27	0.27	360–401	389–401	0	—
	3,4,**5**	326–383	350–383	21	0.54	376–439	402–439	0	—
	4,5,**6**	349–500	384–500	5	0.17	389–500	440–500	8	0.71
Global				**107**	**0.40**			**8**	**0.71**

The number of identified patterns is reported in Tables II and III, and plotted in Figure 5. In Tables II and III, we also report the NACR value for each time window, as well as the global (microaveraged) NACR computed on the whole set of epigraphs of each considered dataset.

The first observation is that the number of extracted patterns in the case of EF discretization is always lower than in the case of EW. Moreover, in the case of EF, the number of extracted patterns is pretty stable. Although this behavior appears to be preferable with respect to that obtained with the EW discretization, in the evaluation of the results, the size of the considered intervals and the number of epigraphs belonging to them should be taken into account as well. Indeed, on the one hand, the high variability obtained with EW discretization could be due to the highly unbalanced number of epigraphs in the considered intervals (e.g., see block 3 with respect to blocks 1 and 2). On the other hand, the EF discretization could lead to comparing cultural aspects between time intervals of highly unbalanced size. For example, analyzing cultural changes of the target block 4 with respect to blocks 2 and 3, on the dataset containing all of the epigraphs, means identifying changes that take place between a large interval (288–348) and a single year (349). This unbalancing can be motivated by the estimation of the dating of epigraphs, which, in principle, is not free from measurement errors.

By comparing the results obtained with all of the epigraphs with the results obtained with the epigraphs with precise dating, we note that, as expected, the number of patterns extracted in the second case is significantly smaller. This is mainly due to the relatively small number of epigraphs, which



(a) EW Discretization—All

(b) EW Discretization—Precise Dating

(c) EF Discretization—All

(d) EF Discretization—Precise Dating

Fig. 5. Number of discovered novelty patterns. Blue lines refer to results obtained with $minSupp = 0.05$ and $maxStD = 0.05$, whereas red lines refer to results obtained with $minSupp = 0.10$ and $maxStD = 0.10$.

does not allow the system to identify all of the changes that could potentially be captured. Finally, results do not show significant differences in the trend between the different settings of the parameters. However, setting the parameters to conservative values (i.e., $minSupp = 0.10$ and $maxStD = 0.10$) prevents the system from extracting patterns when only epigraphs with precise dating are considered.

The analysis of the NACR results shows that the system extracts relatively significant novelty patterns—that is, novelty patterns whose growth rate tends to $+\infty$ or 0. The best value for NACR is obtained for the complete dataset, when parameters are less conservative ($minSupp = 0.05$ and $maxStD = 0.05$) and the EW discretization is used.

Novelty patterns, which originally were formulated as SQL queries, have been transformed into a logic formalism to support the epigraphists in their interpretation, and have been ranked according to the GR values to identify the most interesting ones. In the following, we report some examples of extracted novelty patterns that have been considered remarkable by the domain expert.

Examples of Novelty Patterns Extracted from All of the Epigraphs

—*epigraph(E), ¬opisthographic(E), material(E, M), support(M, ‘Tabula Marmorea’), engraving_technique(E, T), name(T, ‘Insculptus’)*

$$[\mathbf{200} - \mathbf{299}] : [0.33..0.52] \nearrow [\mathbf{300} - \mathbf{349}] : 0.70 \quad \mathbf{GR} = 1.65$$

This pattern (obtained with EW discretization, $minSupp = 0.10$ and $maxStD = 0.10$) describes a moderate increase of single-sided epigraphs engraved with the insculptus technique on tabula marmorea in the interval [300–349] (block 3) with respect to the interval [200–299] (blocks 1 and 2). This may be due to the greater tolerance for Christians under the Emperor Constantine the Great (306–337 AD) and the consequent diffusion of “official” marble epigraphs in public places. This is a significant change, since the first Christian inscriptions were usually written on tiles and bricks.

—*epigraph(E), ¬opisthographic(E), function(E, F), name(F, ‘Sepulchralis’)*
$[\mathbf{250}-\mathbf{349}] : [0.23..0.35] \nearrow [\mathbf{350}-\mathbf{399}] : 0.72 \quad \mathbf{GR} = 2.48$

This pattern (obtained with EW discretization, $minSupp = 0.10$ and $maxStD = 0.10$) describes a significant increase of single-sided funerary epigraphs in the interval [350–399] (block 4) with respect to the interval [250–349] (blocks 2 and 3). This is mainly due to the mass conversion to Christianity since Constantine (during the 4th century AD).

—*epigraph(E), greek(E), latin(E)*
$[\mathbf{200}-\mathbf{299}] : [0.53..0.65] \searrow [\mathbf{300}-\mathbf{349}] : 0.38 \quad \mathbf{GR} = 0.64$

This pattern (obtained with EW discretization, $minSupp = 0.10$ and $maxStD = 0.10$) describes a moderate decrease of epigraphs containing both Greek and Latin terms in the interval [300–349] (block 3) with respect to the interval [200–299] (blocks 1 and 2). This is coherent with the changing process in the Roman Christian communities, which were much more Latinized with respect to the pre-Constantinian era. Although the pattern emphasizes a change in the language used in the inscriptions in such period, it does not directly clarify which language begins to prevail.

—*epigraph(E), ¬greek(E), engraving_technique(E,T), name(T,‘Insculptus’)*
$[\mathbf{326}-\mathbf{349}] : [0.35..0.53] \nearrow [\mathbf{350}-\mathbf{383}] : 0.65 \quad \mathbf{GR} = 1.48$

This pattern (obtained with EF discretization, $minSupp = 0.10$ and $maxStD = 0.10$) describes a moderate increase of epigraphs made with the insculptus technique without ancient Greek terms in the interval [350–383] (block 5) with respect to the interval [326–349] (blocks 3 and 4). This pattern is coherent with the previous one but is extracted from a following period and gives the possibility to better understand the described phenomenon. In particular, it indicates that in the second part of the 4th century, Greek characters, used until that time, begin to disappear from inscriptions.

—*epigraph(E), ¬metrical_inscription(E), ¬greek(E)*
$[\mathbf{200}-\mathbf{325}] : [0.29..0.45] \nearrow [\mathbf{326}-\mathbf{348}] : 0.70 \quad \mathbf{GR} = 1.89$

This pattern (obtained with EW discretization, $minSupp = 0.10$ and $maxStD = 0.10$) describes an increase in the number of epigraphs without ancient Greek terms or a metrical inscription in the interval [326–348] (block 3) with respect to the interval [200–325] (blocks 1 and 2). Similarly to the previous pattern, this pattern indicates a change in the writing style. In fact, during the first half of the 4th century, inscriptions are less sophisticated. This is due to the phenomenon of mass conversion to Christianity, which required more quantity than quality of inscriptions.

It is noteworthy that preceding patterns are characterized by a relatively low growth rate. This seems to support the hypothesis that societal changes during the reign of Constantine were more gradual than often recognized by historians. This hypothesis is also supported by the edict for the restitution to the Christians of properties, buildings, and *coemeteria* of the emperor Gallienus (260–268; the text is reported in Eusebius, *Historia ecclesiastica* 7, 13), and by the edict of tolerance of Galerius from Serdica (now Sofia) in 311, thus prior to that of Milan, of which recalls the contents. Among the funerary inscriptions, the beginning of the long process of merging between the retrospective traditional model and the emerging model explicitly based on religion can be identified already before Constantine, and precisely during the longa pax (from Gallienus to Diocletian), recalled, as a historical object well known to his community, by pope Damasus (366–384): see ICVR, IV 9513.

Examples of Novelty Patterns on Epigraphs with Precise Dating

Interestingly, but unsurprisingly, the novelty patterns extracted from epigraphs with precise dating provide additional insight into the original location of epigraphs. Some of them are reported next.

—epigraph(E), original_location(E, L), address(L, ‘via Appia’)
$[\mathbf{200}-\mathbf{299}] : [0.00..0.07] \nearrow [\mathbf{300}-\mathbf{349}] : 0.31 \quad \mathbf{GR} = 8.87$

This pattern (obtained with EW discretization, $minSupp = 0.05$ and $maxStD = 0.05$) describes a significant increase of the number of epigraphs in via Appia in the first half of the 4th century. From a historical viewpoint, this pattern appears coherent with the great development of the graveyards in the same zone. This is the case of the catacombs of St. Callixtus, which broadens with the *cubiculum* made by the deacon Severus, dated since the papacy of Marcellinus (296–304), and the area of pope Miltiades (311).

—epigraph(E), original_location(E, L), name(L, ‘coem. Cyriacae ad s. Laurentium’),
address(L, ‘via Tiburtina’)
$[\mathbf{376}-\mathbf{401}] : [0.12..0.16] \nearrow [\mathbf{402}-\mathbf{439}] : 0.26 \quad \mathbf{GR} = 1.90$
—epigraph(E), original_location(E, L), address(L, ‘via Ardeatina’)
$[\mathbf{376}-\mathbf{401}] : [0.10..0.16] \searrow [\mathbf{402}-\mathbf{439}] : 0.04 \quad \mathbf{GR} = 0.34$
—epigraph(E), original_location(E, L), address(L, ‘via Appia’)
$[\mathbf{376}-\mathbf{401}] : [0.27..0.39] \searrow [\mathbf{402}-\mathbf{439}] : 0.19 \quad \mathbf{GR} = 0.65$

In these three patterns (obtained with EF discretization, $minSupp = 0.05$ and $maxStD = 0.05$), it is possible to observe a change of the locations of epigraphs in the period 402–439. In particular, it is possible to see an increase of the number of epigraphs made in via Tiburtina (particularly in the settlement of coem. Cyriacae ad s. Laurentium), and, simultaneously, a decrease of the number of epigraphs made in via Ardeatina and in via Appia. These novelty patterns confirm what already known to scholars. In particular, the increase of the number of epigraphs in via Tiburtina can be related to the construction of new monumental structures for the cult of the martyrs Laurentius and Hippolytus, built respectively on the southern and the northern side of the via Tiburtina; these two martyrial cult centers are historically attested just in the first decades of the 5th century (e.g., see the documented activities of the priest Leopardus and Ilicius). On the contrary, along via Appia and via Ardeatina, some settlements dedicated to the cult of the local martyrs were built in a previous period (i.e., during the activity of pope Damasus, 366–384). The use of such cult settlements decreased in the first decades of the 5th century.

General Discussion of Results

Although these examples do not intend to convey any widespread assumptions about the use, function, and content of Roman inscriptions, because they are based on one specific repository, they provide a close perspective on how epigraphists are beginning to work using the results of the many digital epigraphy projects under development. In particular, novelty patterns can be considered as an innovative tool for the historical and/or archaeological analysis of epigraphs. They provide a way to analyze cultural heritage material by different dimensions of analysis—that is, by considering the amount of epigraphs over time as well as by considering specific historical periods. Novelty patterns represent flexible models for the summarization of the evolution discovered on tens of thousands of epigraphs, for which domain experts do not have the capability of providing an overall view. They can either confirm what is already known by epigraphists or reveal unexpected associations between concepts, thus raising new questions that can stimulate more systematic researches.

5. CONCLUSIONS

In this article, we have demonstrated that it is possible to use data mining technologies to identify interesting patterns in large repositories of epigraphs. These patterns can be used to organize large collections of inscriptions, introduce younger scholars to the field of epigraphy, and identify anomalies that epigraphists can later explore using more traditional methods.

In particular, we have presented a novelty detection algorithm that performs a longitudinal analysis of stored inscriptions and discovers novelty patterns that may convey useful information on the evolution of languages, writing styles, customs, and traditions of a community.

To deal with data heterogeneity, we transformed transcriptions into a set of textual features and resorted to a relational data mining approach to analyze a portion of the database. Indeed, relational data mining methods are able to directly manipulate and analyze complex and structured data consisting of a collection of database relations. The price paid for this extra flexibility is a higher computational complexity and a more sophisticated design of the methods with respect to those that solve the same task on simpler data representations (two-way table or single database relation).

The proposed approach has been used to process ancient Christian inscriptions of Rome, which are stored in the EDB repository created and maintained by the University of Bari, Italy. Together with a quantitative analysis, we have reported and discussed some discovered novelty patterns. In many cases, they can be properly interpreted by epigraphists who take care of the corpus of the inscriptions. For some patterns, however, further historical analysis is still required before considering them as true “nuggets” discovered by the data mining algorithm.

The proposed approach can be profitably exploited to analyze data stored in other epigraphic repositories. In this regard, we plan to conduct a similar analysis on data that will be stored into the federated database that is under development in the context of the EU-funded project EAGLE. Such analysis would provide a much wider view about cultural evolution and would be of interest to a wider community of epigraphists studying epigraphs belonging to different areas and historical periods. Moreover, we will further investigate the historical conclusions reported in this article by taking into account both the formal results presented in this study and other kinds of arguments based on primary sources, including those that do not make use of computational techniques. Finally, we intend to improve the method to take into account the intrinsic uncertainty introduced by the digitization of ancient documents. In particular, we plan to take into account the presence of diacritical marks introduced by epigraphists in the transcription to indicate some damaged/unknown portions of the text, as well as the uncertainty introduced by the dating estimation of the epigraphs.

ACKNOWLEDGMENTS

The authors thank Carlo Carletti (the lead scientist responsible for the EDB project since 1988), Antonella Daniela Agostinelli, Antonio Franzoso, and Anita Rocco, as well as the following past and present collaborators: Cristina Grisanzio, Ruggero Lombardi, Filippo Piazzolla, Marida Pierno, Miriam Ramunni, Domenico Schiraldi, and Carolina Ventura. The authors thank Lynn Rudd for reading through the article.

REFERENCES

Maristella Agosti, Lucio Benfante, and Nicola Orio. 2013. A contribution for the dissemination of cultural heritage content to a wider public. In *Digital Libraries and Archives*, Maristella Agosti, Floriana Esposito, Stefano Ferilli, and Nicola Ferro (Eds.). Communications in Computer and Information Science, Vol. 354. Springer, 195–206.

Apostolos Antonacopoulos and Dimosthenis Karatzas. 2004. Document image analysis for World War II personal records. In *Proceedings of the 1st International Workshop on Document Image Analysis for Libraries (DIAL)*. 336–341.

Annalisa Appice, Margherita Berardi, Michelangelo Ceci, and Donato Malerba. 2005. Mining and filtering multi-level spatial association rules with ARES. In *Foundations of Intelligent Systems*. Lecture Notes in Computer Science, Vol. 3488. Springer, 342–353.

Rebecca R. Benefiel. 2010. Rome in Pompeii: Wall inscriptions and GIS. In *Latin on Stone. Epigraphic Research and Computing*, F. Feraudi-Gruenais (Ed.). Rowman & Littlefield, Lanham, MD, 45–75.

Mario Borillo. 1984. Stratégies de mise à l'épreuve de conjectures historiques. In *Informatique pour les sciences de l'homme*. Bruxelles.

Michelangelo Ceci, Annalisa Appice, Corrado Loglisci, Costantina Caruso, Fabio Fumarola, and Donato Malerba. 2009. Novelty detection from evolving complex data streams with time windows. In *Foundations of Intelligent Systems*. Lecture Notes in Computer Science, Vol. 5722. Springer, 563–572.

Michelangelo Ceci, Annalisa Appice, and Donato Malerba. 2008. Emerging pattern based classification in relational data mining. In *Proceedings of the 19th International Conference on Database and Expert Systems Applications (DEXA'08)*. 283–296.

Michelangelo Ceci, Margherita Berardi, and Donato Malerba. 2007. Relational data mining and ILP for document image understanding. *Applied Artificial Intelligence* 21, 4–5, 317–342.

Franca Debole and Fabrizio Sebastiani. 2003. Supervised term weighting for automated text categorization. In *Proceedings of the 2003 ACM Symposium on Applied Computing (SAC'03)*. ACM, New York, NY, 784–788.

Sašo Džeroski and N. Lavrač. 2001. *Relational Data Mining*. Springer-Verlag.

Ingo Frommholz, Holger Brocks, Ulrich Thiel, Erich J. Neuhold, Luigi Iannone, Giovanni Semeraro, Margherita Berardi, and Michelangelo Ceci. 2003. Document-centered collaboration for scholars in the humanities—The COLLATE system. In *Research and Advanced Technology for Digital Libraries*. Lecture Notes in Computer Science, Vol. 2769. Springer, 434–445.

Fabio Fumarola, Gianvito Pio, Antonio E. Felle, Donato Malerba, and Michelangelo Ceci. 2013. EDB: Knowledge technologies for ancient Greek and Latin epigraphy. In *Bridging Between Cultural Heritage Institutions*. Lecture Notes in Computer Science, Vol. 385. Springer, 29–40.

Maria Gäde, Nicola Ferro, and Monica Lestari Paramita. 2011. CHiC 2011—Cultural heritage in CLEF: From use cases to evaluation in practice for multilingual information access to cultural heritage. In *CLEF (Notebook Papers/Labs/Workshop)*.

Basilios Gatos, Kostas Ntzios, Ioannis Pratikakis, Sergios Petridis, Thomas Konidaris, and Stavros J. Perantonis. 2004a. A segmentation-free recognition technique to assist old Greek handwritten manuscript OCR. In *Document Analysis Systems VI*. Lecture Notes in Computer Science, Vol. 3163. Springer, 63–74.

Basilios Gatos, Ioannis Pratikakis, and Stavros J. Perantonis. 2004b. An adaptive binarization technique for low quality historical documents. In *Document Analysis Systems VI*. Lecture Notes in Computer Science, Vol. 3163. Springer, 102–113.

Eamonn Keogh, Stefano Lonardi, and Bill Y.-C. Chiu. 2002. Finding surprising patterns in a time series database in linear time and space. In *KDD'02: Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, 550–556.

Frank Le Bourgeois and Hala Kaileh. 2004. Automatic metadata retrieval from ancient manuscripts. In *Document Analysis Systems VI*. Lecture Notes in Computer Science, Vol. 3163. Springer, 75–89.

John W. Lloyd. 1987. *Foundations of Logic Programming* (2nd ed.). Springer-Verlag, Berlin.

Corrado Loglisci, Michelangelo Ceci, and Donato Malerba. 2013. Discovering evolution chains in dynamic networks. In *New Frontiers in Mining Complex Patterns*. Lecture Notes in Computer Science, Vol. 7765. Springer, 185–199.

Junshui Ma and Simon Perkins. 2003. Online novelty detection on temporal sequences. In *Proceedings of the 9th International Conference on Knowledge Discovery and Data Mining (KDD'03)*. ACM, New York, NY, 613–618.

Heikki Mannila and Hanno Toivonen. 1997. Levelwise search and borders of theories in knowledge discovery. *Data Mining and Knowledge Discovery* 1, 3, 241–258.

Markos Markou and Sameer Singh. 2003. Novelty detection: A review—part 1: statistical approaches. *Signal Processing* 83, 12, 2481–2497.

David Mimno. 2012. Computational historiography: Data mining in a century of classics journals. *Journal on Computing and Cultural Heritage* 5, 1, Article No. 3.

Shan-Hwei Nienhuys-Cheng and Ronald de Wolf. 1997. *Foundations of Inductive Logic Programming*. Springer, Heidelberg.

Vivien Petras, Nicola Ferro, Maria Gäde, Antoine Isaac, Michael Kleineberg, Ivano Masiero, Mattia Nicchio, and Juliane Stiller. 2012. Cultural heritage in CLEF (CHiC) overview 2012. In *Proceedings of the CLEF 2012 Evaluation Labs and Workshop, Online Working Notes*.

Gordon D. Plotkin. 1970. A note on inductive generalization. *Machine Intelligence* 5, 153–163.

Henriette Roued-Cunliffe. 2010. Towards a decision support system for reading ancient documents. *Literary and Linguistic Computing* 25, 4, 365–379.

Jörg Sander, Martin Ester, Hans-Peter Kriegel, and Xiaowei Xu. 1998. Density-based clustering in spatial databases: The algorithm GDBSCAN and its applications. *Data Mining and Knowledge Discovery* 2, 2, 169–194.

Eduardo J. Spinosa, André Ponce de Leon F. de Carvalho, and João Gama. 2008. Cluster-based novel concept detection in data streams applied to intrusion detection in computer networks. In *Proceedings of the Symposium on Applied Computing (SAC'08)*. ACM, New York, NY, 976–980.

Received August 2013; revised December 2013; accepted March 2014

Queries

Q1: ED/AU: Note that only one AU address/email is given.
Q2: AU: Unable to access "http://eda-bea.es" in Footnote 1. Please check URL.
Q3: AU: Throughout, please confirm use of both space before and after dashes with years in brackets as well as the closed-up style in this article.
Q4: AU: Please confirm use of periods at end of each of the "P" lines.
Q5: AU: Are periods necessary at end of each entry for P3 through P5? If not, please delete them.
Q6: AU: Please review all bulleted entries. First lines are no longer italic. As wanted?
Q7: AU: Please confirm change to "the lead scientist responsible for . . ."