Comparative Analysis of Data Entities: Knowledge Mining Objects

Claus-Peter Rückemann
Westfälische Wilhelms-Universität Münster (WWU),
Leibniz Universität Hannover,
North-German Supercomputing Alliance (HLRN), Germany
Email: ruckema@uni-muenster.de

Abstract—**This paper presents the research and results on creating means for a comparative analysis of data entities from knowledge resources. The research was started in order to tackle the challenges of data analysis with increasing and complex resources. Comparing data entities is a most ambitious task for increasingly complex data objects, integrated resources, and relations – from the knowledge resources, as well as from the computational perspective. The implementation utilises complementary components, which enable to structure and describe complex knowledge and support an advanced analysis. The paper presents practical examples and discusses the high level view of an implementation and case study. For practical reasons with the comparative analysis, the knowledge resources utilise references to publicly available data resources. The goal of this research is an advanced methodology and modular means for comparative analysis and knowledge mining with information systems and long-term multi-disciplinary knowledge resources.**

***Keywords*–Knowledge Mining; Comparative Analysis; Content Factor; Universal Decimal Classification; Advanced Data-centric Computing.**

I. Introduction

Advanced methods of knowledge mining with information systems and knowledge resources are becoming increasingly important. With that, improving knowledge mining and at the same time integrating larger amounts of data increases the challenges. The core of challenges is the data analysis. Within data analysis, comparing "data" is a central task. Comparing data entities is an even more ambitious task when data objects and relations are becoming more and more complex.

The term data entity in context with knowledge resources refers to any data representing objects of any kind like digital or realia objects, including references, e.g., to objects or conceptual knowledge.

Within this research, special application components were created and implemented in order to provide modular means to be integrated for a comparative analysis, e.g., knowledge resources referring to structured and unstructured data, conceptual data, especially knowledge classification, and methodologies specialised on the before mentioned means, e.g., the Content Factor method (CONTFACT) [1]. The multi-disciplinary knowledge resources and the application of the Content Factor method have enabled new flexible workflows and the creation of new complementary means for both the enhancement of multi-disciplinary knowledge resources and for data-centric knowledge discovery processes [2]. Some of the most widely required means with data entities of knowledge resources are components for a comparative analysis. Comparative Analysis (CA) is defined as an item-by-item comparison of two or more comparable entities.

This paper is organised as follows. Section II summarises the state-of-the-art, motivation, and frame of reference to the ground of comparison. Section III introduces the object and data entities integration with different resources, and implemented for this case study. Section IV provides the computation and analysis results based on the selected resources. Section V discusses the main results and evaluates them in context of the application. Section VI summarises the results and lessons learned, conclusions, and future work.

II. State-of-the-art, motivation, and frame

The elementary way of knowledge mining, practised by the vast majority of approaches and services ignores content quality, document types, and cognitive knowledge. That means, content is handled independently from the creation process and expertise, content from databases, Web pages, and scanned books are not differentiated, and classification of content is disregarded.

CA modules can be used for arbitrary purposes with knowledge mining workflows, e.g., for selecting complementary resources as well as selecting objects supporting decision making processes. The methodology is used with knowledge mining workflows, integrating dedicated knowledge resources and publicly available content, e.g., text documents and books, because of their complementary nature regarding content, structure, and quality. The following sections describe the motivation and the base of the conducted CA.

A. Frame of reference

The significance of integrating different data entities results from the context, in which they are placed. This research presents a methodology of comparing different data entities as referred from objects in advanced knowledge resources.

Objects with higher quality are mostly more complex. Advanced knowledge mining and decision making requires more than one methodology or algorithm for analysis of available objects and their references, data entities, and attributes. A major challenge is the difference of entities, e.g., regarding entity type, original purpose of the entity, and source but also content and structure.

ISBN: 978-1-61208-567-8

Different types of entities cannot be ignored from advanced workflows because they contain unique knowledge and information. In most cases, the knowledge and information can even only be provided by different entities and referred sources. Methodologies should be provided, which are beneficial to be integrated in advanced workflows, especially for analysis, quantisation, and qualification of different entities. The deployed means should allow long-term data-centric applications and intrinsically foster the seamless integration with existing workflows. In addition, the methodologies and architecture of integration should allow the implementation of modular and least invasive components.

B. Grounds for comparison

Besides the complexity, a combination of data entities from different sources and different types was choosen for the following reasons. The rationale behind the choice for knowledge resources and entities from referred objects results from complementary content and context. There is an arbitrary high quality of multi-disciplinary content in the knowledge resources, which are in continuous development [3]. In addition, the knowledge resources can provide an extremely high knowledge and information density. The Gutenberg resources [4] can provide a large number of fully publicly available standard text documents and elaborations for a wide multi-disciplinary context. Both types of resources contain essential amounts of textual content and are continuously extended and improved. The relationship between different entities is the addressed knowledge content with its unique nature. The thesis is, that different entities should neither be left out from advanced workflows nor should their content, the unique knowledge and information, be ignored. The following lens comparison discusses the most important aspects text-by-text, focussing on advanced knowledge resources and referred resources.

III. Integration of resources

The following sections describe how an integration was achieved and which results were gained with the analysis.

A. Data entities

Data entities can be created from many resources. With this research, knowledge objects and data entities were automatically created from ‘Gutenberg documents’. At the time of the case study (January 2017) Project Gutenberg [4] offered 53,855 free ebooks for download. The document files include the text in a version of the respective edition, which can be a revised edition or translation. The text editions are linked as different document files, e.g., plain text files, which can be converted into data entities and integrated with different data entities. Regarding conceptual knowledge, the Gutenberg documents use a flat implementation of the Library of Congress (LoC) classification outline [5]. The ebook links contain some relevant information, too, e.g., the bibliographic record, EBook-No. 25062, a link the LoC Class entries, and the release date of the edition. The original publication date of the source text is contained in the document files.

Data entities from knowledge resources’ collections and containers [1] are used with many knowledge mining applications [3]. Knowledge resources contain multi-disciplinary knowledge objects, which can be used in arbitrary ways for providing factual, conceptual, procedural, and meta-cognitive knowledge. The objects can contain any content and context as well as references, e.g., translations, transliterations, synonyms, associations, references, conceptual knowledge (e.g., UDC), concordances, links, keywords, and Content Factors (of elements). The objects can be based on records, e.g., characters, words, lines, and complex records. Any content and context can be used for analysis and evaluation of an object.

With the created modules the data entities from the Gutenberg resources can be handled in the same way as knowledge resources’ objects, e.g., of different origin. The following case study starts with a knowledge mining request for “Vesuvius” in the context of “volcanology”. The primary Gutenberg result matrix contains a number of documents [6]-[12] provided with the precomputation.

B. Object and data entity integration

Objects and data entities can be integrated with knowledge resources in arbitrary ways, e.g., as a referred object or by creating an instance of an object. The integration allows an analysis and evaluation, e.g., with knowledge resources’ objects. The following excerpt (Figure 1) shows a knowledge resources’ object automatically created from an entity of Gutenberg document 33483 [7] with LoC classification [13].

```
33483-0.txt [Document]:
            ...
            THE
            ERUPTION OF VESUVIUS
            IN 1872, ...
            BY
            PROFESSOR LUIGI PALMIERI,
            _Of the University of Naples; Director of the Vesuvian Observatory._
            ...
            WITH NOTES, AND AN
            _INTRODUCTORY SKETCH OF THE PRESENT STATE OF KNOWLEDGE_
            OF
            TERRESTRIAL VULCANICITY,
            _The Cosmical Nature and Relations of
            Volcanoes and Earthquakes._
            ...
            BY
            ROBERT MALLET,
            _Mem. Inst. C.E., F.R.S., F.G.S., M.R.I.A., &c., &c._
            ...
            WITH ILLUSTRATIONS. ...
            LONDON: ...
            _ASHER & CO._,
            13, BEDFORD STREET, COVENT GARDEN, W.C. ...
            1873.  ...
            W. S. Johnson, Nassau Steam Press, 60, St. Martin’s Lane,
            Charing Cross, W.C.
            ...
```

Figure 1. Automatically created Gutenberg knowledge resources object for document 33483 (geosciences collection, LX, excerpt).

As an example, an object excerpt of an object instance “Vesuvius” from a knowledge resources’ collection, resulting from a knowledge mining process, is shown in Figure 2. The objects can contain any knowledge, e.g., factual and conceptual knowledge. Here, the object carries names and synonyms in different languages, dynamically usable geocoordinates, Universal Decimal Classification (UDC) and so on, including geoclassification (UDC:(37), Italia. Ancient Rome and Italy). The data used here is based on the content and context from the knowledge resources, provided by the LX Foundation Scientific Resources (LX not an acronym) [3].

```
Vesuvius [Volcanology, Geology, Archaeology]:
         (lat.) Mons Vesuvius.
         (ital.) Vesuvio.
         Volcano, Gulf of Naples, Italy.
         Complex volcano (compound volcano).
         Stratovolcano, large cone (Gran Cono).
         Volcano Type: Somma volcano,
         VNUM: 0101-02=,
         Summit Elevation: 1281\UD{m}. ...
         ...
         Syn.: Vesaevus, Vesevus, Vesbius, Vesvius
         s. volcano, super volcano, compound volcano
         s. also Pompeji, Herculaneum, seismology
         ...
         compare La Soufrière, Mt. Scenery, Soufriere
         ...
         %%IML: UDC:[911.2+55]:[57+930.85]:[902]"63"(4+37+23+24)=12=14
         %%IML: GoogleMapsLocation: http://maps.google.de/maps?hl=de&gl=de&vpsrc
          =0&ie=UTF8&ll=40.821961,14.428868&spn=0.018804,0.028238&t=h&z=15
         ...
         ...
         ...
         ...
         ...
         ...
```

Figure 2. Knowledge resources collection object “Vesuvius” (LX resources, geoscientific collection, excerpt).

The LX knowledge resources’ structure and the classification references [14] based on UDC [15] are essential means for the processing workflows and evaluation of the knowledge objects and containers. Both provide strong multi-disciplinary and multi-lingual support. For this part of the research all small unsorted excerpts of the knowledge resources objects only refer to main UDC-based classes, which for this part of the publication are taken from the Multilingual Universal Decimal Classification Summary (UDCC Publication No. 088) [16] released by the UDC Consortium under the Creative Commons Attribution Share Alike 3.0 license [17] (first release 2009, subsequent update 2012).

IV. Computation and Analysis

A. Content Factor computation for data entities

Objects of any kind can be integrated with knowledge resources. Objects can contain instances of data entities and refer to associated knowledge. For an analysis, a number of common information regarding the objects and data entities is required. The following excerpt (Figure 3) illustrates the creation of Content Factor definition sets [1] for the use with data entities. Definition sets are used for both Gutenberg resources and knowledge resources.

```
% (c) LX-Project, 2016, 2017
{Vol}:=[Vv][Oo][Ll][Cc][Aa][Nn]
{Vul}:=[Vv][Uu][Ll][Cc][Aa][Nn]
{Veu}:=[Vv][Ee][Ss][Uu][Vv]
{Vee}:=[Vv][Ee][Ss][Ee][Vv][Oo]
{Kom}:=[Kk][Oo][Mm][Ee][Tt]
{Com}:=[Cc][Oo][Mm][Ee][Tt]
{Met}:=[Mm][Ee][Tt][Ee][Oo][Rr]
{Erd}:=[Ee][Rr][Dd][Bb][Ee][Bb][Ee][Nn]
{Ear}:=[Ee][Aa][rr][Tt][Hh][Qq][Uu][Aa][Kk][Ee]
{Puz}:=[Pp][Uu][Zz][Zz][Oo][Ll]
{Poz}:=[Pp][Oo][Zz][Zz][Oo][Ll]
```

Figure 3. CONTFACT definition set for Gutenberg Project resources and knowledge resources, (LX, excerpt).

Figure 4 shows the Normed Basic Content Factor (NBCF, $\overline{\kappa}_{\mathrm{B}}$) [1] computed for a knowledge resources object reference to the Gutenberg Project document 33483.

```
CONTFACT:BEGIN
CONTFACT:20161227-234624:AU:{Veu}{Veu}{Vul}{Vol}{Ear}{Veu}{Met}{Veu}{Vol}{Met
 }...{Veu}{Veu}{Ear}{Veu}...{Veu}{Veu}{Veu}{Veu}...{Veu}{Ear}{Vol}{Ear}/39843
CONTFACT:20161227-234624:AS:{Ear}...{Veu}{Veu}{Vol}{Vol}...{Vul}{Vul}/39843
CONTFACT:20161227-234624:M:{Vol}:=[Vv][Oo][Ll][Cc][Aa][Nn]
CONTFACT:20161227-234624:M:{Vul}:=[Vv][Uu][Ll][Cc][Aa][Nn]
CONTFACT:20161227-234624:M:{Veu}:=[Vv][Ee][Ss][Uu][Vv]
CONTFACT:20161227-234624:M:{Vee}:=[Vv][Ee][Ss][Ee][Vv][Oo]
CONTFACT:20161227-234624:M:{Kom}:=[Kk][Oo][Mm][Ee][Tt]
CONTFACT:20161227-234624:M:{Com}:=[Cc][Oo][Mm][Ee][Tt]
CONTFACT:20161227-234624:M:{Met}:=[Mm][Ee][Tt][Ee][Oo][Rr]
CONTFACT:20161227-234624:M:{Erd}:=[Ee][Rr][Dd][Bb][Ee][Bb][Ee][Nn]
CONTFACT:20161227-234624:M:{Ear}:=[Ee][Aa][rr][Tt][Hh][Qq][Uu][Aa][Kk][Ee]
CONTFACT:20161227-234624:M:{Puz}:=[Pp][Uu][Zz][Zz][Oo][Ll]
CONTFACT:20161227-234624:M:{Poz}:=[Pp][Oo][Zz][Zz][Oo][Ll]
CONTFACT:20161227-234624:M:STAT:OBJECTELEMENTSDEF=11
CONTFACT:20161227-234624:M:STAT:OBJECTELEMENTSALL=39843
CONTFACT:20161227-234624:M:STAT:OBJECTELEMENTSMAT=356
CONTFACT:20161227-234624:M:STAT:OBJECTELEMENTSCFO=.00900324
CONTFACT:20161227-234624:M:STAT:OBJECTELEMENTSKWO=1
CONTFACT:20161227-234624:M:STAT:OBJECTELEMENTSLAN=0
CONTFACT:20161227-234624:M:INFO:OBJECTELEMENTSOBJ=33483-0.txt
CONTFACT:20161227-234624:M:INFO:OBJECTELEMENTSDCM=(c) LX-Project, 2016, 2017
CONTFACT:20161227-234624:M:INFO:OBJECTELEMENTSMTX=LX Foundation Scientific
 Resources; Object Collection
CONTFACT:20161227-234624:M:INFO:OBJECTELEMENTSAUT=Claus-Peter R\"uckemann
CONTFACT:END
```

Figure 4. NBCF $\overline{\kappa}_{\mathrm{B}}$ computed for knowledge resources object reference to Gutenberg Project document 33483 (LX Resources, excerpt).

Figure 5 shows the NBCF computed for a knowledge resources object reference to the object “Vesuvius”.

```
CONTFACT:BEGIN
CONTFACT:20170205-161508:AU:{Veu}{Vol}{Veu}{Veu}{Vol}{Vol}{Vol}{Vol}{Vol}{Vol}{
 Vol}{Vol}{Vol}/71
CONTFACT:20170205-161508:AS:{Veu}{Veu}{Veu}{Vol}{Vol}{Vol}{Vol}{Vol}{Vol}{Vol}{
 Vol}{Vol}{Vol}/71
CONTFACT:20170205-161508:M:{Vol}:=[Vv][Oo][Ll][Cc][Aa][Nn]
CONTFACT:20170205-161508:M:{Vul}:=[Vv][Uu][Ll][Cc][Aa][Nn]
CONTFACT:20170205-161508:M:{Veu}:=[Vv][Ee][Ss][Uu][Vv]
CONTFACT:20170205-161508:M:{Vee}:=[Vv][Ee][Ss][Ee][Vv][Oo]
CONTFACT:20170205-161508:M:{Kom}:=[Kk][Oo][Mm][Ee][Tt]
CONTFACT:20170205-161508:M:{Com}:=[Cc][Oo][Mm][Ee][Tt]
CONTFACT:20170205-161508:M:{Met}:=[Mm][Ee][Tt][Ee][Oo][Rr]
CONTFACT:20170205-161508:M:{Erd}:=[Ee][Rr][Dd][Bb][Ee][Bb][Ee][Nn]
CONTFACT:20170205-161508:M:{Ear}:=[Ee][Aa][rr][Tt][Hh][Qq][Uu][Aa][Kk][Ee]
CONTFACT:20170205-161508:M:{Puz}:=[Pp][Uu][Zz][Zz][Oo][Ll]
CONTFACT:20170205-161508:M:{Poz}:=[Pp][Oo][Zz][Zz][Oo][Ll]
CONTFACT:20170205-161508:M:STAT:OBJECTELEMENTSDEF=11
CONTFACT:20170205-161508:M:STAT:OBJECTELEMENTSALL=71
CONTFACT:20170205-161508:M:STAT:OBJECTELEMENTSMAT=13
CONTFACT:20170205-161508:M:STAT:OBJECTELEMENTSCFO=.21311475
CONTFACT:20170205-161508:M:STAT:OBJECTELEMENTSKWO=2
CONTFACT:20170205-161508:M:STAT:OBJECTELEMENTSLAN=1
CONTFACT:20170205-161508:M:INFO:OBJECTELEMENTSOBJ=Vesuvius
CONTFACT:20170205-161508:M:INFO:OBJECTELEMENTSDCM=(c) LX-Project, 2016, 2017
CONTFACT:20170205-161508:M:INFO:OBJECTELEMENTSMTX=LX Foundation Scientific
 Resources; Object Collection
CONTFACT:20170205-161508:M:INFO:OBJECTELEMENTSAUT=Claus-Peter R\"uckemann
CONTFACT:END
```

Figure 5. NBCF $\overline{\kappa}_{\mathrm{B}}$ computed for knowledge resources object “Vesuvius” (LX Resources, excerpt).

Both NBCF were computed with the same definition set (Figure 3). The data entities from the referenced Gutenberg resources and knowledge resources both contain multiple matches. The resulting Content Factor for the knowledge resources object is higher due to the higher concentration of relevant elements in the object. The Gutenberg object shows a higher absolute number of matches and multiple hits.

B. Procedures and modules

Two main modules were required with the assistance pre-computation for identifying and selecting objects and data entities from the Gutenberg resources before entering the CA workflow. The preparative assistance data was computed with a module `gutenberganalysis` and the classification was extracted with a module `gutenbergloc`. The implementation case study for the comparative analysis methodology required the creation of several major components and modules. Table I shows a sequence of modules, which allows to create the base for a CA workflow as created with this case study.

TABLE I. COMPARATIVE ANALYSIS WORKFLOW PROCEDURES AND IMPLEMENTED MODULES WITH GUTENBERG RESOURCES.

Procedure	*Module*
Gutenberg interface	`textca_gutenberginterface`
Configuration	
Inconsistencies checker	
Data slicer	
Analysis	`textca_analysis`
Configuration	
Data join	`textca_join`
Configuration	
Visualisation module	`textca_visualisation`
Configuration	
Plotting generator	
Conditional visualisation	
Statistics	`textca_statistics`
Configuration	
Visualisation plotting	`textca_plotting`
Configuration	

Practically, the modules can be implemented with any environment and frameworks. In the case study Perl, Shell, and Gnuplot have been used. In general this means any module could be replaced by a different implementation separately. Any module requires configuration options, which at least can be pre-configured options. In their application, the analysis up to visualisation modules for the Gutenberg resources are identical to the application for the knowledge resources. Therefore, the computations for all data entities were done with the `textca` group of modules.

C. Comparison of data entities in collections

Figure 6 shows the computed CA module result for a case insensitive `vesuv` (`[Vv][Ee][Ss][Uu][Vv]`) target for the above Gutenberg object (Figure 1).

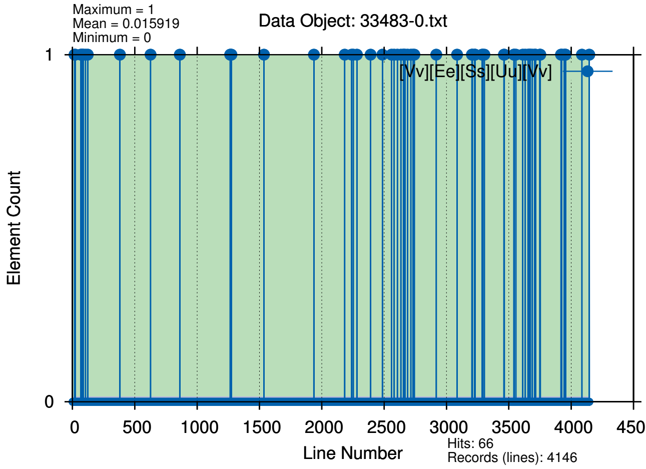

Figure 6. Comparative Analysis module result for a Gutenberg object precomputed by an assistance process for the case insensitive `vesuv` target.

The analysis including the illustration was automatically computed for the respective object.

Figure 7 shows the automatically computed CA module result with the respective target (pattern) for the resulting knowledge resources collection object “Vesuvius” (Figure 2).

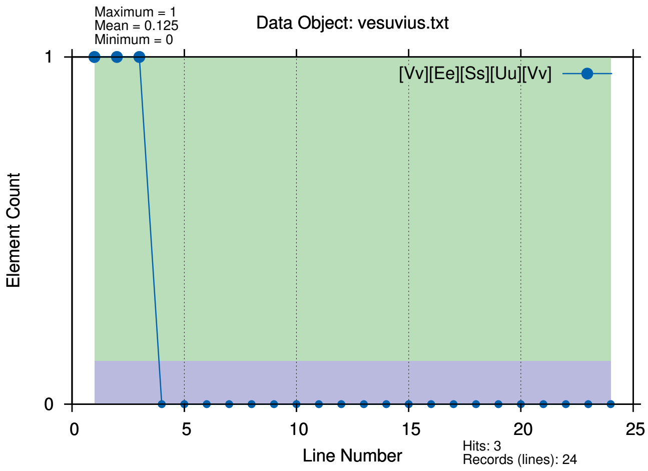

Figure 7. Comparative Analysis module result for the knowledge resources collection object “Vesuvius” (LX resources, geoscientific collection, excerpt).

The result shows some criteria of the object itself in context with the relevant mining pattern. The figure illustrates that the object contains a relevant mining result in the first and several consecutive records (here: lines) with a maximal occurance count of one in a record. The density of relevant occurances in the object is relatively high compared to common texts, even if from comparable special topic documents. Therefore, the mean value is quite high in that case. The computed background shading illustrates the space spanned by the available records (number of lines) and element counts. The mean value is illustrated by the border of the color change.

D. Comparison of data entities in containers

Figure 8 shows the computed CA module result for a Gutenberg object for a case insensitive `vulc/volc` (`[Vv][UuOo][Ll][Cc]`) target.

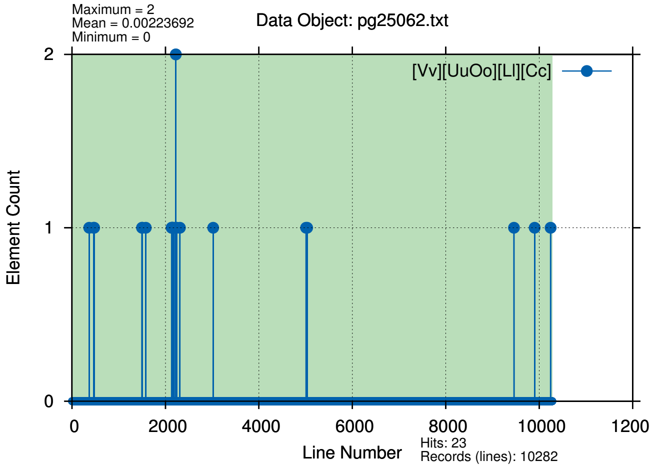

Figure 8. Comparative Analysis module result for a Gutenberg object precomputed by an assistance process for case insensitive `vulc`/`volc` target.

Figure 9 shows the automatically computed CA module result with the respective target (pattern) for the resulting knowledge resources collection object “Vesuvius” (Figure 2).

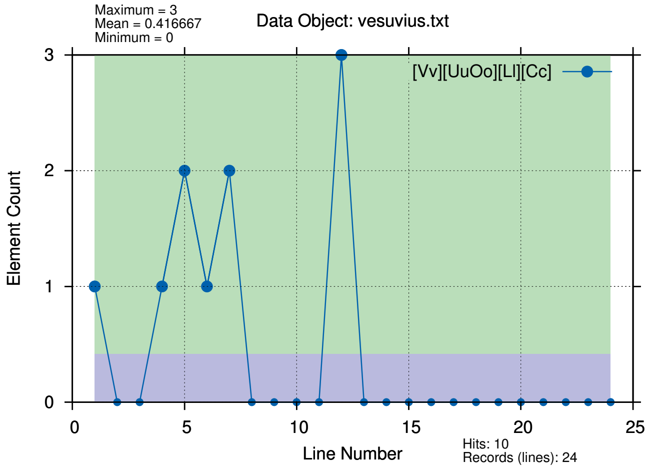

Figure 9. Comparative Analysis module result for the knowledge resources collection object “Vesuvius” (LX resources, geoscientific collection, excerpt).

There is more than one occurance in several lines each, with a maximal occurance count of three in a record. Figure 10 shows the computed CA module result for the volcanological features container for the same target.

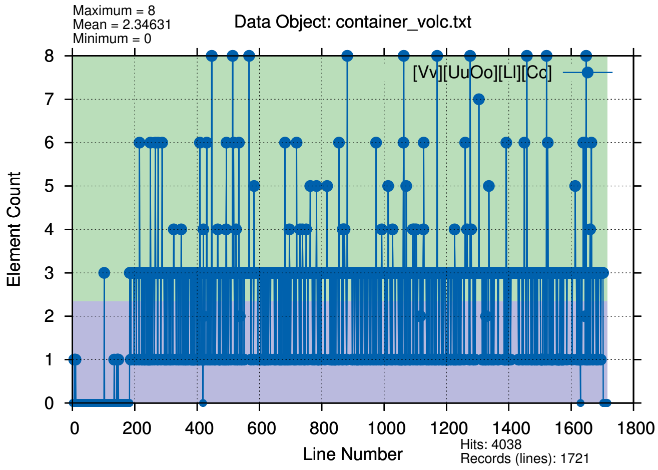

Figure 10. Comparative Analysis module result for the volcanological features container for case insensitive `vulc`/`volc` target (LX resources).

Both figures (Figures 10 and 8) illustrate the very high relevance of the objects. Nevertheless, the structure and density of hits is much higher in the container object than in the Gutenberg object. In addition, the mean value is extremely high for the container object. Also, the central part of the container object does not contain a line without the target. There are even more hits than records.

Even in a top hit Gutenberg object the number of records is much higher and the number of hits is lower. The comparison also reveals that both objects represent different object types, a knowledge resources object and a classical text object. The latter one mostly contains natural language. For any resources, many CA and Content Factor computations are done on a result matrix. With any workflow further information and decision making support can result from computing assistant views for the knowledge entities, e.g., based on their context. The results of the CA, the Content Factor, classification, and any results and attributes from assistant views can be included in an analysis, e.g., if a ranking of results is required for a specific knowledge mining workflow.

V. Discussion

The case study integrates sources of different knowledge entities for knowledge mining workflows, selecting entities by computing advanced analysis criteria.

A. *Integration and comparison*

The selected data compasses over 50,000 Gutenberg documents and more than 50,000 objects from knowledge resources. The selected sizes of objects range from hundreds of bytes to several megabytes.

The limitation for the case study was done for demonstration, due to the fact that the number of available overall knowledge resources objects may easily outnumber the number of Gutenberg documents. With the resources, the classification considers about fifty languages, summing up to about three million descriptions. Conceptual assistance is available for resource and object classification, which allows to automate integration workflows.

For the integration, instances of the objects containing the relevant data entities were automatically computed. It was possible to apply the provided means in the same way to the entities. The computation for the data entities from knowledge resources can be much more fine grained and systematic due to the complex structures and elements. The computation for the Gutenberg data entities can use the same means but some details and structure are not automatically available. The data sizes of the main Gutenberg data entities are most probably larger than these of the average knowledge resources' data entities.

B. *CA mean values*

Table II compares CA mean values from the computation for selected objects and target groups for the integrated resources.

Table II. Selected computation data entities: Objects and target groups sorted by their CA mean values.

Object	*Target / Target-Group*	*CA Mean*
Knowledge res., Vesuvius	`[Vv][Ee][Ss][Uu][Vv]`	0.125
Gutenberg 33483-0	`[Vv][Ee][Ss][Uu][Vv]`	0.015919
Know. res., volc. feat. cont.	`[Vv][Ee][Ss][Uu][Vv]`	0.00291375
Gutenberg 25062	`[Vv][Ee][Ss][Uu][Vv]`	0.000486287
Know. res., volc. feat. cont.	`[Vv][UuOo][Ll][Cc]`	2.34631
Knowledge res., Vesuvius	`[Vv][UuOo][Ll][Cc]`	0.416667
Gutenberg 33483-0	`[Vv][UuOo][Ll][Cc]`	0.0356971
Gutenberg 25062	`[Vv][UuOo][Ll][Cc]`	0.00223692

There are entities with higher and lower mean values, for the Gutenberg resources as well as for the knowledge resources. Higher values indicate a cumulation of relevant terms, e.g., as with the appearance in collections, tabulars, and listings.

Practice showed that for complementing knowledge in the volcanological features container with extended context, relevant entities from the Gutenberg resources with higher mean values can be a primary source for references. Relevant entities from the Gutenberg resources with lower mean values may primarily deliver reference information for collection objects.

C. Ranking

For this scenario, a ranking was built from the rankings for the entities from the Gutenberg entities and from the knowledge resources.

The ranking considers the available information, e.g., classified targets, relevance of targets, references and context. The Gutenberg ranking especially considers the results from the CA, Content Factor and classification (LoC), based on the primary Gutenberg result matrix. The ranking of knowledge resources especially considers the CA, Content Factor, and classification, e.g., UDC and Universal Classified Classification (UCC). The integrated ranking considers the the CA, Content Factor values, and concordances of comparable entities.

An integration for a workflow ranking requires that the means need to be individually choosen for a certain application scenario. In this case, a records base (lines) was an appropriate choice for CA, Content Factor, and conceptual knowledge.

D. Computational trace and context

A common knowledge discovery process integrates a sequence of decision making processes at different levels, e.g., from which resources to which single objects. Each step in a sequence can require to handle millions of objects and references. The access to the Gutenberg resources is not intended to be automated. Therefore, no performance data is available for the Gutenberg resources or for conducting the precomputation for its whole content. The precomputation assistance includes the cached Gutenberg content for the respective mining targets.

Table III shows the computation characteristics relevant with the workflow procedure for an example of the above integrated Gutenberg and knowledge resources case for two objects.

TABLE III. COMPUTATION CHARACTERISTICS WITH THE WORKFLOW PROCEDURE FOR TWO INTEGRATED OBJECTS, WALL TIMES PER CPU.

Workflow Procedure	*Wall Time*
Precomputation assistance	24.8 s
Analysis, resources classification	1.2 s
Analysis, object classification	14.7 s
Comparative Analysis	3.2 s
(Integrative workflow step)	n s

The table times refer to one Central Processing Unit (CPU) per mining process (Intel Xeon, at 2.9 GHz). Due to the complexity of the elementary workflows it is not desirable to have more than one CPU per process involved at the atomic level. Arbitrary practical application scenarios involving many processes with large data resources may be organised to fit the architecture of the available infrastructure. With certain scenarios, where an author wants to integrate complex references, the precomputation assistance can benefit a lot from using many-CPU infrastructures. The higher level workflow step, integrating the aforementioned procedures, will use a lot of intermediate results from procedures and content from resources. There is no general range for the time scale at the higher levels but at these levels the requirements on computation and communication can be extremely high, therefore, the higher level steps are candidates for parallelisation. Anyhow, workflow creators must always be aware that computing requirements can be non-linear, depending on the workflow created by an author for a choosen purpose.

VI. CONCLUSION

The paper presented an advanced methodology for knowledge mining with multi-disciplinary knowledge resources and different data entities. Required modules and algorithms were successfully and efficiently implemented for supporting a Comparative Analysis, integrating different data entities in mining workflows.

This research showed that with the availability of appropriate methodologies different entities neither need to be left out from advanced knowledge mining workflows nor should their content be ignored. It was shown that in result, there is a complementary relationship between objects from knowledge resources and referred objects from external sources, including their data entities.

The Content Factor methodology for data description and analysis is used with all available resources. As was shown, CA methods cannot be replaced by other means like classification or Content Factor because they are based on completely different grounds but these complementary means can be integrated within more complex workflows. CA modules can help optimise the decision making, e.g., with supporting context-spanning Content Factor definition sets. CA modules can be used for delivering additional descriptive information, which can be used for documentation and knowledge mining purposes. CA is much beyond statistics. The significant part of the CA is the visualisation of pattern sequences in entities. The pattern sequences hold relevant parts of the entity characteristics and can also be used for documentation. The statistics are used in addition, for the analysis.

Objects from advanced knowledge resources can provide an excellent data base on knowledge. The knowledge resources can provide high quality object collections and containers with data entities of most reliable and unique content and qualities. Referred objects from external sources can extend the available data base regarding width and depth. Therefore, referred objects and external sources can extend the available data base and content. The best fit targets regarding volcanological features from the resources, including the Gutenberg resources, were automatically analysed.

On the side of the Gutenberg resources, a number of challenges have been found especially with the Gutenberg objects themselves. With the documents, workflow creators face a lot of inconsistencies in structure and marking even regarding major elements. Bibliographic data and versioning could also be improved. Better structured and more complete bibliographic data would be beneficial for any wider and systematic use. A common container format for the Gutenberg documents, handling any data files and associated data in a flexible and 'clean' way would be beneficial.

Besides the purpose laid out with this research, CA modules can be a complementary and supportive methodology applied with a wide range of advanced applications like document identification or plagiarism detection. Future work will be spent on further integrating different resources and creating methods and means for handling data entities and objects.

ACKNOWLEDGEMENTS

We are grateful to the "Knowledge in Motion" (KiM) long-term project, Unabhängiges Deutsches Institut für Multidisziplinäre Forschung (DIMF), for partially funding this implementation, case study, and publication under grants D2014F1P04518, D2014F2P04518, D2016F1P04683, and to its senior scientific members, especially to Dr. Friedrich Hülsmann, Gottfried Wilhelm Leibniz Bibliothek (GWLB) Hannover, to Dipl.-Biol. Birgit Gersbeck-Schierholz, Leibniz Universität Hannover, and to Dipl.-Ing. Martin Hofmeister, Hannover, for fruitful discussion, inspiration, practical multidisciplinary case studies, and the analysis of advanced concepts. We are grateful to Dipl.-Ing. Hans-Günther Müller, Cray, for his work on flexible practical solutions to architectural challenges and excellent technical support. We are grateful to all national and international partners in the Geo Exploration and Information cooperations for their constructive and trans-disciplinary support. We thank the Science and High Performance Supercomputing Centre (SHPSC) for long-term support of collaborative research since 1997, including the GEXI developments and case studies.

REFERENCES

[1] C.-P. Rückemann, "Enhancement of Knowledge Resources and Discovery by Computation of Content Factors," in Proceedings of The Sixth International Conference on Advanced Communications and Computation (INFOCOMP 2016), May 22–26, 2016, Valencia, Spain. XPS Press, 2016, pages 24–31, ISSN: 2308-3484, ISBN-13: 978-1-61208-478-7, URL: http://www.thinkmind.org/download.php?articleid=infocomp_2016_2_30_60047 [accessed: 2017-01-22].

[2] C.-P. Rückemann, Z. Kovacheva, L. Schubert, I. Lishchuk, B. Gersbeck-Schierholz, and F. Hülsmann, Best Practice and Definitions of Data-centric and Big Data – Science, Society, Law, Industry, and Engineering. Post-Summit Results, Delegates' Summit: Best Practice and Definitions of Data-centric and Big Data – Science, Society, Law, Industry, and Engineering, Sep. 19, 2016, The Sixth Symp. on Advanced Computation and Information in Natural and Applied Sciences (SACINAS), The 14th Int. Conf. of Numerical Analysis and Applied Mathematics (ICNAAM), Sep. 19–25, 2016, Rhodes, Greece, 2016, URL: http://www.user.uni-hannover.de/cpr/x/publ/2016/delegatessummit2016/rueckemann_icnaam2016_summit_summary.pdf [accessed: 2017-01-22].

[3] "LX-Project," 2017, URL: http://www.user.uni-hannover.de/cpr/x/rprojs/en/#LX [accessed: 2017-02-05].

[4] "Project Gutenberg," 2017, URL: http://www.gutenberg.org [accessed: 2017-02-05].

[5] "Library of Congress Classification Outline," 2017, Library of Congress (LoC) Classification, URL: https://www.loc.gov/catdir/cpso/lcco/ [accessed: 2017-02-05].

[6] M. Saderra Masó, Catalogue of Violent and Destructive Earthquakes in the Philippines, 2006, Project Gutenberg, eBook, EBook-No.: 18556, Release Date: June 11, 2006, Digitised Version of the Original Publication from 1910, URL: http://www.gutenberg.org/ebooks/18556 [accessed: 2017-02-05], URL: http://www.gutenberg.org/cache/epub/18556/pg18556.txt [accessed: 2017-02-05].

[7] L. Palmieri, The Eruption of Vesuvius in 1872, 2010, Project Gutenberg, eBook, EBook-No.: 33483, Release Date: August 22, 2010, Digitised Version of the Original Publication from 1873, Translator: Mallet, Robert, (1810–1881), URL: http://www.gutenberg.org/ebooks/33483 [accessed: 2017-02-05], URL: http://www.gutenberg.org/files/33483/33483-0.txt [accessed: 2017-02-05].

[8] M. Serao, Sterminator Vesevo (English: Vesuvius the great exterminator), 2014, Project Gutenberg, eBook, EBook-No.: 46658, Release Date: August 22, 2014, Digitised Version of the Original Publication from 1907, Diary of the Eruption of April 1906, URL: http://www.gutenberg.org/ebooks/46658 [accessed: 2017-02-05], URL: http://www.gutenberg.org/cache/epub/46658/pg46685.txt [accessed: 2017-02-05].

[9] C. Davison, A Study of Recent Earthquakes, 2008, Project Gutenberg, eBook, EBook-No.: 25062, Release Date: April 12, 2008, Digitised Version of the Original Publication from 1905, URL: http://www.gutenberg.org/ebooks/25062 [accessed: 2017-02-05], URL: http://www.gutenberg.org/cache/epub/25062/pg25062.txt [accessed: 2017-02-05].

[10] W. Hamilton, Observations on Mount Vesuvius, Mount Etna, and Other Volcanos, 2011, Project Gutenberg, eBook, EBook-No.: 35433, Release Date: March 1, 2011, Digitised Version of the Original Publication from 1774, Editor: Cadell, T., (1742–1802), URL: http://www.gutenberg.org/ebooks/35433 [accessed: 2017-02-05], URL: http://www.gutenberg.org/cache/epub/35433/pg35433.txt [accessed: 2017-02-05].

[11] R. D'Awans, L'Ameublement de l'Hôtel de Pitsembourg au milieu du XVIIe siécle, 2004, Project Gutenberg, eBook, EBook-No.: 11586, Release Date: March 1, 2004, Digitised Version of the Original Publication from 1901, Communication faite en séance du 26 avril 1901, URL: http://www.gutenberg.org/ebooks/11586 [accessed: 2017-02-05], URL: http://www.gutenberg.org/cache/epub/11586/pg11586.txt [accessed: 2017-02-05].

[12] A. H. C. Gelpke, Ueber die schrecklichen Wirkungen des Aufsturzes eines Kometen auf die Erde und über die vor fünftausend Jahren gehabte Erscheinung dieser Art, 2006, Project Gutenberg, eBook, EBook-No.: 18471, Release Date: May 29, 2006, Digitised Version of the Original Publication from 1835, URL: http://www.gutenberg.org/ebooks/18471 [accessed: 2017-02-05], URL: http://www.gutenberg.org/cache/epub/18471/pg18471.txt [accessed: 2017-02-05].

[13] "QE: Science: Geology," 2016, Library of Congress (LoC) Classification, URL: https://www.loc.gov/aba/cataloging/classification/lcco/lcco_q.pdf [accessed: 2017-02-05].

[14] C.-P. Rückemann, "Enabling Dynamical Use of Integrated Systems and Scientific Supercomputing Resources for Archaeological Information Systems," in Proc. INFOCOMP 2012, Oct. 21–26, 2012, Venice, Italy, 2012, pp. 36–41, ISBN: 978-1-61208-226-4.

[15] "UDC Online," 2017, URL: http://www.udc-hub.com/ [accessed: 2017-02-05].

[16] "Multilingual Universal Decimal Classification Summary," 2012, UDC Consortium, 2012, Web resource, v. 1.1. The Hague: UDC Consortium (UDCC Publication No. 088), URL: http://www.udcc.org/udcsummary/php/index.php [accessed: 2017-02-05].

[17] "Creative Commons Attribution Share Alike 3.0 license," 2012, URL: http://creativecommons.org/licenses/by-sa/3.0/ [accessed: 2017-02-05].