Methodology of Knowledge Mapping for Arbitrary Objects and Entities: Knowledge Mining and Spatial Representations – Objects in Multi-dimensional Context

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

Abstract—The research presented in this paper focusses on a new methodology for knowledge based mapping of objects and entities for creating new multi-dimensional context. The mapping to new context can improve complex knowledge mining, discovery, and decision making results. The new context increases the potential for creating new insights. The paper introduces the new methodology used with advanced knowledge mining and provides the results of the present research. Examples from an implementation and a case study on knowledge mining and spatial representations are given. The case study utilises commonly available unstructured data and creates new multi-disciplinary context, especially spatial mapping of entities and integration with data and advanced tools, which can be used for further analysis, e.g., automated and visual analysis. The methodology can employ integrated knowledge resources and services for mapping support and can be applied to any content from arbitrary disciplines. The results of the mapping to new context can be used for knowledge mining workflows, for gaining new insight, and for creating and further improving long-term knowledge resources. The methodology also supports automated learning processes. This research aims on creating required bases for these goals and for new practical mining procedures and algorithms.

Keywords–Data-centric Knowledge Mining; Mapping Objects and Entities; Spatial Mapping and Visualisation; Knowledge Resources; Advanced Computing.

I. INTRODUCTION

It is a truth universally acknowledged, that any knowledge, e.g., based on unstructured and structured data, can contain parts, which may refer to other knowledge but which are not explicitly linked. Further, existing methods promising to deal with lexical and term mapping or ontologies showed deficient and inadequate for coping with challenges of arbitrary knowledge mapping and multi-dimensional context. Methods [1] and implementations for automated mapping [2] are not sufficient, the more as approaches do not span disciplines [3]. Term identification [4] is not suitable for mapping beyond simple context like bibliographic data, too. Available mapping approaches are very limited to non-general knowledge related tasks [5], even when dealing with context [6].

The methodology presented here was developed in order to identify entities inside of or referenced with data and create new context for knowledge objects and entities. Knowledge is an excellent integrator as it can complement, e.g., from factual, conceptual, procedural, and metacognitive knowledge.

Knowledge mapping is the process of creating mappings between two data objects. In that way knowledge mapping contributes significantly to data integration and data sciences methods [7]. The means of referring objects and sub-objects, “entities”, with a new context is considered as “knowledge mapping”. Objects, e.g., a document, a part of a text, or an image may be associated with other objects, by its knowledge, e.g., its factual or conceptual knowledge. For example, creating new spatial context for textual entities in knowledge objects requires to build non-fixed associations, apply a fuzzy spatial locate, and implement a text location to map-mapping.

The procedure enables to automatically create a spatial mapping for possible locations in a document, e.g., Points Of Interest (POI) or other places in a data set or file.

The rest of this paper is organised as follows. Section II introduces the new methodology applied for knowledge mapping. Section III discusses previous work, components, and used resources. Section IV presents the implementation and case study based on the methodology. Section V presents generated interactive dynamical context examples. Section VI summarises the lessons learned, conclusions, and future work.

II. METHODOLOGY

The methodology can be used for creating new object and entity context environments, e.g., in knowledge mining context. The following steps describe the methodology.

1) Start is an arbitrary object.
2) Object / entity analysis.
3) Object / entity mapping.
4) Context creation.
5) Result is an object and/or entity with a new context environment.

Objects can be arbitrary objects, unstructured or structured, un referenced or referenced, e.g., containing different entities of content. The methodology is not limited to any possibly restricted implementation or platform. In case of textual objects and entities, the object can, e.g., be a text document. In case the mapping targets on geo-referencing otherwise non geo-referenced objects or entities, then the mapping can be considered a spatial mapping. With the latter target the context creation can be considered a spatial visualisation.

The methodology of knowledge mapping for arbitrary objects and entities can be schematically summarised (Figure 1).
For example (Figure 2): When the object is a plain text-object and creating spatial visual context is the target, then the steps can be implemented with object and entity analysis, spatial object / entity mapping, and spatial visualisation for creating an object / entity spatial mapping in a new context.

The targets for the case study are spatial visualisation and context. The implementation architecture of mapping arbitrary objects and entities to a new object context environment is shown in Figure 3. Data and modules are provided by Knowledge Resources. The originary resources deliver the data objects and entities, which can be unstructured or structured. The application resources and components contain appropriate modules for the required steps. The object is retrieved, possible object entities are extracted, object data resources are being analysed, objects are being compared, a conceptual mapping is performed on objects, spatial mapping is performed on objects, appropriate spatial media is generated, including media formats and colourisation, and a spatial visualisation is performed. The result is an object / entity instance in a new context environment. The modules and filters perform the analysis and handle the objects and entities, e.g.,

- entities in different context inside an object,
- transcriptions,
- transliterations,
- translations,
- abbreviations,
- acronyms, . . .

In many cases, additional handling of data will be desired, even if not essential for the procedure of a method or the operation of a service. For example, in case of textual objects and entities a number of aspects exist, which contribute to the attainment of a certain quality:

- Differently organised or structured entities per object.
- Sub-entities, multiple entities in a pseudo-entity.
- Inconsistencies in data.
- Errors in data.
- Typographic differences.
- Ambiguous or plurivalent entities.
- Multi-lingual entities.
- Different diction.
- Different syntax.
- Different element ordering in entities.
- Different structures.
- Time dependencies of aspects, mapping, and meaning.
- Different character sets.
- Different formatting.

Any of these and comparable aspects are handled by the modules and appropriate pre- and post-filters. With the case study, for the above aspects respective research was conducted gathering various data and developing suitable methods over several years, data which can be deployed to create filters, which were used for holding the results presented here.

It is required to abstract certain information in many application scenarios, e.g., for generalisation or privacy. Besides any kind of filter, the method also allows to implement fuzziness in a flexible and wide range of ways. For example, on the one hand a precise location can be reduced to city, region, or country. Comparable but different locations can be unified to one different location representing a larger area. On the other hand, location coordinates can be automatically or manually reduced in precision and/or equipped with an offset. With these means, workflows can deliver kind of “Fuzzy Context”, e.g., a fuzzy location, providing a precision level of a public region instead of showing a certain building in a result.

III. PREVIOUS WORK, COMPONENTS, AND RESOURCES

For the implementation of case studies, the modules are built by support of a number of major components and resources, which can be used for a wide range of applications, e.g., creation of resources and extraction of entities.
The facility for consistently describing knowledge is a valuable quality, especially conceptual knowledge, e.g., using the Universal Decimal Classification (UDC). The knowledge resources objects can refer to main UDC-based classes, which for this publication are taken from the Multilingual Universal Decimal Classification Summary (UDCC Publication No. 088) [8] released by the UDC Consortium under the Creative Commons Attribution Share Alike 3.0 license [9] (first release 2009, subsequent update 2012).

Data and objects result from public, commonly available, and specialised Knowledge Resources. The Knowledge Resources are containing factual and conceptual knowledge as well as documentation and instances of procedural and metacognitive knowledge. These resources contain multi-disciplinary and multi-lingual data and context.

Context data for calculations and visualisation also requires cartographic thematic context data. The knowledge resources were integrated with data based on the gridded ETOPO1 1-arc-minute global relief model data [10]. Data can be composed from various sources, e.g., adding Shuttle Radar Topography Mission (SRTM) data [11].

The Network Common Data Form (NetCDF) [12] developed by the University Corporation for Atmospheric Research (UCAR/Unidata), National Center for Atmospheric Research (NCAR) is used for spatial context data. NetCDF is an array based data structure for storing multi-dimensional data. A NetCDF file is written with an ASCII header and stores the data in a binary format, e.g., with a mapping suite.

The Generic Mapping Tools (GMT) [13] suite application components are used for handling the spatial data, applying the related criteria, and for the visualisation.

The visualisation files generated from the mapping results are using the Keyhole Markup Language (KML), an eXtended Markup Language (XML) based format for specifying spatial data and content. KML is considered an official standard of the Open Geospatial Consortium (OGC). The KML description can be used with many spatial components and purposes, e.g., with a Google Earth or Google Maps presentation [14], with a Marble representation [15], using OpenStreetMap (OSM) [16].

Modules are employing Perl Compatible Regular Expressions (PCRE) for specifying common string patterns and Perl [17] for component wrapping purposes with this case study.

IV. IMPLEMENTATION CASE STUDY: SPATIAL CONTEXT

The following sections provide information regarding implemented components (ixloccoord, module for location coordinates) and a practical case study, which was done for demonstrating the methodology of mapping objects and entities, creating new context environments. The case study shows components, which were built for mapping scenarios creating spatial context (Figure 3) and illustrates new insights and relevance for knowledge creation and advanced mining.

A. The components

All the components and modules required for the architecture (Figure 3) were implemented. The following components were created for the practical implementation of the three major central modules, object / entity analysis, mapping, and context creation, demonstrating all steps of the methodology.

- The object / entity analysis modules process objects for entities, which can be fed into a mapping mechanism.
- The pre-filters change, mark, and remove entities before the mapping modules try to create entity mappings.
- The mapping modules do have the task to deliver spatial coordinates for appropriate entities.
- The post-filters change, mark or remove entities after the resolver worked on entities for a spatial mapping.
- The context creation modules deliver the geo-referencing for a spatial application.

The modules can be centralised or distributed, e.g., implemented as a local directory of comparable and resolved entities or an online service. Appropriate directories can be provided by knowledge resources as well as by spatial mapping services.

Change processes in pre- and post-filters can include unification, improvements for resolvability, mapping and so on.

Different application components with different features can be deployed for dynamical and interactive use and visualisation, e.g., GMT, Marble, and Google Maps.
B. Case study: From plain text to spatially linked context

The following passages show some major steps for creating spatially linked context from plain text, which were used in the workflows required for the case studies.

The single data object in Figure 4 contains mostly unstructured text [18], markup, and framing instructions.

Passages not relevant for demonstration were shortened to ellipses. Figure 5 shows the object content after automatically integrated with the Knowledge Resources via a join module.

The Object Entity Mapping can associate relevant objects, e.g., via conceptual knowledge and comparative methods. Table I shows an excerpt of the conceptual data (UDC) used for characteristics and place classification, creating spatial context.

### Table I. Classification references, object/entity analysis and mapping: Characteristics & Place (LX [19]).

<table>
<thead>
<tr>
<th>UDC Code</th>
<th>Description (English, excerpt)</th>
</tr>
</thead>
<tbody>
<tr>
<td>UDC(1)</td>
<td>Place and space in general. Localization. Orientation</td>
</tr>
<tr>
<td>UDC(100)</td>
<td>Universal as to place. International. All countries in general</td>
</tr>
<tr>
<td>UDC:05</td>
<td>Common auxiliaries of persons and personal characteristics</td>
</tr>
<tr>
<td>UDC:057.4</td>
<td>Professional or academic workers</td>
</tr>
<tr>
<td>UDC:378</td>
<td>Higher education. Universities. Academic study</td>
</tr>
</tbody>
</table>

The codes especially reflect the common auxiliaries of general characteristics and place with the analysis of the object and entities, e.g., affiliation and spatial location.

Figure 6 shows an excerpt with possible entities of locations after an object entity analysis and mapping.

After object entity analysis, filters, and mapping, a resolver module can equip the entities with geo-references (Figure 7).

For this result, the pre- and post filters handled all issues as described. The entries are shown in a special 3 column Comma Separated Value (CSV) format. The GMT format for the geo-referenced CSV is straightforward (Figure 8).

The context creation includes the media generation. Figure 9 excerpts a KML representation of the above geo-referenced entities, resulting from the original mapping.

A global view of all resulting entities automatically analyzed and mapped from the single object [18] is shown in Figure 10. The single-object view integrates the new spatial context of the object entities with a high precision topographic-oceanographic thematic view. The bullets are very much oversized for this illustration. The respective components are provided by GMT suite applications, especially pscast and gmtselect [19], which allow a multitude of spatial operations and criteria in context with the entities. Further, KML can be used with many spatial applications, e.g., with Marble and Google Maps. Generators can be configured to mark different types of locations with different markers. It is also possible to automatically mark locations with thumbnail photos being associated with the respective location and so on.
V. FURTHER DYNAMICAL CONTEXT EXAMPLES

Figure 11 is a screenshot of a dynamical, interactive view (Marble), a political map context for above created context.

A consecutive mapping allows to analyse the entities in completely new context. For example, parts of an unstructured document can be put into context with any type of n-dimensional information, e.g., historical and climatological context by using spatial information [20] and mapping for finding links. In this case, data entities can be spatially mapped and associated with multi-dimensional data from many disciplines, and data entities can not only be associated in space but also in time. The data allows to do detailed knowledge mining analysis as well as visual analysis. For the created context, Figures 12 and 13 show screenshots of an interactive globe-view (Marble) with climate zone context and an interactive view (Google Earth) with Earth view context.

Besides the new context of spatial distribution and according algorithms and math, the new context environments build links in order to associate entities with knowledge from arbitrary disciplines and proceed with further analysis.

Due to conceptual attributes of knowledge mapping and spatial algorithms, the implementation allows high grades of scalability and fuzziness. New context can also be kept and used in learning systems components. This, e.g., can provide conditional object / entity aggregation and time sequences.
VI. CONCLUSION

This paper introduces a new methodology for knowledge based mapping of arbitrary objects and entities and creating new multi-dimensional context. The research presented the theoretical background, a successful implementation based on the methodology and a case study. The methodology fulfills the goals of successfully creating new context, knowledge mapping can improve complex knowledge mining and associated tasks as well as it can be beneficial for the development of knowledge resources. A practical case study, evaluated by groups of independent researchers, showed that applying the methodology can create relevant new context for entities in commonly available unstructured data. Mapping to a spatial context is just one of an arbitrary number of possible mappings, which can be created with the methodology. The quality of results can significantly benefit from a training and learning phase, depending on context. Here, with resolving nearly all possible place entities with the used new resources, the creation and learning phase of the modules accumulates to several years. The methodology allowed to implement a data-centric checkpointing, which corresponds to associated learning processes. As shown, in most cases it may be advisable for flexibility to create modular architectures of components instead of monolithic applications. It can further be convenient to consider robustness and reliability of service modules, depending on the architecture of an overall implementation. One means of dealing with infrastructure can be a failure correction, e.g., multiple task runs and check modules.

Analysis and case studies are on the way implementing advanced context generation, e.g., with spatial visualisation, 2D, 3D, route mapping, public transport, animation, and fly-over tours as well as analysing computational requirements, which are widely scalable, depending on implementation of components and computing architectures.

In addition, future work concentrates on further developing and improving the mapping modules and features for closer integration with the multi-disciplinary knowledge resources.

ACKNOWLEDGEMENTS

We are grateful to the “Knowledge in Motion” (KiM) long-term project, Unabhängiges Deutsches Institut für Multidisziplinäre Forschung (DIMF), for partially funding this research, implementation, case study, and publication under grants D2017F1P04708 and D2017F1P04812 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, to Dipl.-Ing. Martin Hofmeister, Hannover, and to Olaf Lau, Hannover, Germany, for fruitful discussion, inspiration, practical multi-disciplinary case studies, and the analysis of advanced concepts. We are grateful to Dipl.-Ing. Hans-Günther Müller, Cray, Germany, for his work and assistance providing practical private cloud and storage solutions 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.

REFERENCES