Transforming Literature-intensive Research Processes Through Text Analytics - Design, implementation and Lessons Learned

Frank Bensberg[1*], Gunnar Auth[2], Christian Czarnecki[2], Christopher Wörndle[3]

1 Osnabrück University, **2** Hochschule für Telekommunikation Leipzig, **3** Detecon International GmbH, ***** Corresponding author, f.bensberg@hs-osnabrueck.de

Abstract

The continuing growth of scientific publications raises the question how research processes can be digitalized and thus realized more productively. Especially in information technology fields, research practice is characterized by a rapidly growing volume of publications. For the search process various information systems exist. However, the analysis of the published content is still a highly manual task. Therefore, we propose a text analytics system that allows a fully digitalized analysis of literature sources. We have realized a prototype by using EBSCO Discovery Service in combination with IBM Watson Explorer and demonstrated the results in real-life research projects. Potential addressees are research institutions, consulting firms, and decision-makers in politics and business practice.

Keywords: Text analytics, Text mining, Literature review, Research process.

Citation: Bensberg, F., Auth, G., Czarnecki, C., Wörndle, C. (2018, October) *Transforming Literature-intensive Research Processes Through Text Analytics - Design, implementation and Lessons Learned*. Paper presented at the Fifth International Management Information Systems Conference.

Editor: H. Kemal İlter, Ankara Yıldırım Beyazıt University, Turkey

Received: August 19, 2018, **Accepted:** October 18, 2018, **Published:** November 10, 2018

Transforming literature-intensive research processes through text analytics - design, implementation and lessons learned

Frank Bensberg
Osnabrück University of Applied Science, Germany

Gunnar Auth
Christian Czarnecki
Hochschule für Telekommunikation Leipzig, Germany

Christopher Wörndle
Detecon International GmbH, Germany

Abstract

The continuing growth of scientific publications raises the question how research processes can be digitalized and thus realized more productively. Especially in information technology fields, research practice is characterized by a rapidly growing volume of publications. For the search process various information systems exist. However, the analysis of the published content is still a highly manual task. Therefore, we propose a text analytics system that allows a fully digitalized analysis of literature sources. We have realized a prototype by using EBSCO Discovery Service in combination with IBM Watson Explorer and demonstrated the results in real-life research projects. Potential addressees are research institutions, consulting firms, and decision-makers in politics and business practice.

Keywords

Text analytics; text mining; literature review; research process.

1. Problem Definition

The information systems discipline is characterized by a high degree of dynamics with regard to the focused research topics, which are also influenced by hypes and trends (Steininger et al. 2009). This situation results in specific challenges for the scientific research of such topics. For example, working on a new topic requires transparency about the current state of published content in order to identify open research questions. Therefore, a systematic literature review is required, which - especially for dynamic topics - leads to a high number of literature sources to be processed. The interdisciplinary character of information systems further enlarges the search space for relevant literature sources (Sturm et al. 2015). Furthermore, in most cases not only academic literature sources are relevant, but also publications originating from business practice provide empirical findings, which are important for researching information systems. As Bornmann & Mutz (2015) state, the number of scientific publications generally doubles within 24 years with an annual growth rate of about 3 %. Hence, in the future a significant, continuous growth of literature sources is expected.

In view of this development, research practice is facing major challenges with regard to the systematic processing of this increasing amount of literature. So far, tasks of literature search and management are highly digitalized - for example, through the use of established literature databases and literature management programs. However, the content analysis and the synthesis of literature references are largely performed manually (Sturm & Sunyaev 2017).

Text analytics methods, which are able to identify previously unknown patterns and correlations in text data, are suitable for supporting these intellectually demanding research activities (Feldman & Sanger 2006). The spectrum of text analytics methods ranges from simple lexicometric methods (e.g. frequency and concordance analyses for different word forms) to complex, multivariate approaches for segmenting and classifying text documents using machine learning methods (Dann et al. 2017).

In this article, a prototype for text analytics is presented. From a technical perspective it is based on EBSCO Discovery Service, which is used by numerous university libraries as a basic search environment for literature research. In our approach we combine this service with the text analytics system IBM Watson Explorer in order to analyse the literature data. It is available to universities as standard software free of license fees and can support end users in the explorative analysis of technical texts with various methods. For the conceptual foundation of this design oriented approach, the central goals and tasks of literature analysis are presented below. As a first evaluation of our approach, the informational potential of the solution is demonstrated using real data. The article concludes with a discussion of the lessons learned from the use of the prototype.

2. Objectives and Tasks of Literature Analysis

In research processes, literature analyses are basically used to reflect the current state of the art or the historical development in a thematically defined field of research (Bortz & Döring 2016). The taxonomy of Cooper (1988) has become common sense for the configuration of literature analyses. According to this taxonomy, Figure 1 summarizes the general occurrences of research as morphological box.

Characteristic	Categories			
Focus	Research Outcomes	Research Methods	Theories	Practices/ Applications
Goal	Integration	Criticism	Identification of Central Issues	
Perspective	Neutral Representation		Espousal of Position	
Coverage	Exhaustive	Exhaustive with selective citation	Representative	Central
Organization	Historical	Conceptual	Methodological	
Audience	Specialized Scholars	General Scholars	Practitioners or Policy Makers	General Public

Figure 1: Taxonomy for Literature Analyses according to Cooper (1988)

Consequently, the following six characteristics should be used to classify literature analyses:

- The *focus* results from the orientation on different elements of applied research. For example, results, methods, theories and certain fields of application are typical specifics.
- The *goal* can either bring together the available literature (integration), criticize existing results based on ex ante formulated criteria, or identify central problem areas in the research landscape.
- A literature analysis can be conducted from a *neutral perspective*, but can also be characterised by a given *argumentative positioning*.
- The degree of *coverage* shows to which extent the existing literature is examined. Complete coverage is generally only possible in highly specialized research fields, while general research topics often require a restriction to central literature sources (e.g., A/B journals).
- The *organization* characteristic provides anchor points for grouping literature sources in the course of the analysis, for example, on their historical development, the common theoretical-conceptual foundation, and the spectrum of methods used.
- With regard to the *audience*, not only scientific actors but also practitioners, political decision-makers, and the general public could be relevant.

For the operational realization of a literature analysis, a number of activities should be linked. Typically the process begins with the differentiation of a research field, which is then described using a set of thematic keywords (Bortz & Döring 2016). Subsequently, suitable literature databases should be selected (e.g., Google Scholar, EBSCO Discovery Service, SpringerLink). Then suitable database queries are formulated and executed for these literature databases. After successful execution of the queries, the resulted references are exported for further processing - for example, by use of common literature management software. Typically the central bibliographic data and the abstracts are exported, and sometimes full texts are also available in literature databases. Finally, the analysis result (e.g., state of the art) is developed and documented. While the previous activities are mainly search tasks this last activity includes the structuring and evaluation of the content itself, which has a high impact on the quality of the analysis.

In order to create a structured overview of a research field, Webster & Watson (2002) propose the construction of a concept matrix, which is applicable to literature analyses on information systems (e.g., Seifert & Nissen 2016, Labes et al. 2013). In a nutshell, a concept matrix depicts the support of different concepts by the examined literature sources of a research field (cf. Figure 2).

Literature Source	Concepts			
	A	B	C	[…]
Source 1	●	●	●	
Source 2		●		●
[…]			●	

Figure 2: Structure of a Concept Matrix according to Webster & Watson (2002)

To construct the concept matrix, the literature sources must be systematically read, which is nowadays still mainly performed manually by the researchers. After completion of the reading phase, a synthesis of the identified concepts is performed., The results are integrated into a frame of reference. On the linguistic level, concepts are technical terms formed by nouns or more complex noun constructions (e.g., noun sequences or adjective-noun combinations, such as “Smart Service Engineering”). The explorative identification of such linguistic structures can be supported by text analytics methods. Hence, the development of concept matrices can be realized more efficiently and effectively by suitable software. In the following section, we present a prototype, which provides those functionalities by applying text analytics methods.

3. Design and demonstration of the prototype

3.1 Architecture of the prototype

First, we have designed the architecture for the text analytics prototype to support literature-intensive research processes. It is based on a loose, file-based coupling of the two software products EBSCO Discovery Service (EDS) and IBM Watson Explorer (WEX). In our proposed literature research process, the two components are used sequentially. First, a topic-specific literature database is generated with EDS, which is then processed with WEX using text analytics methods.

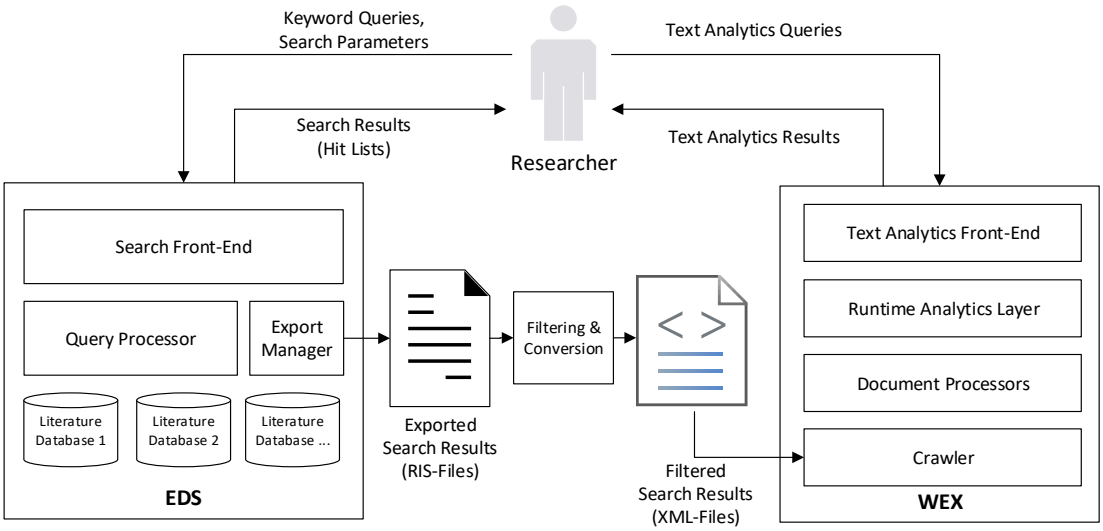

Figure 3: Architecture of the Prototype

Literature data obtained using EDS queries are used as the data basis for the prototype. The resulting references can be exported as files in the Research Information System format (RIS format). The RIS files are filtered and converted to a XML format, which can be further processed by WEX. Filtering is necessary, since EDS results may contain duplicates which have to be eliminated. In addition, literature data without a valid ISSN/ISBN is to be excluded from literature analyses.

We have realized the above architecture by programming the required queries and interfaces. Its demonstration in the exemplary research field *Industry 4.0 in road traffic* is discussed in Subsection 3.2. The

further conceptual design and functionalities of our prototype are described in the following subsections with reference to this example.

3.2 Demonstration scenario for the prototype

As a practical application scenario, the exemplary research field *Industry 4.0 in road traffic* is used. In order to create a database, a corresponding query was formulated. It localises all sources whose abstracts contain the terms *connected*, *car-to-car*, *C2C*, *C2X* or *C2I* in addition to the generic term *car*. The corresponding search query for the EDS literature database is:

```
(AB "car") AND (AB "connected" OR AB "car-to-car"
OR AB "C2C" OR AB "C2I" OR AB "C2X")
```

As a result, 17,926 bibliographic references were identified and exported using the EDS Export Manager. The full texts were not procured due to the volume of the sources. After XML conversion, the WEX text analytics system was used to analyse the literature data. In this context, a simple XML structure is used to import the bibliographic data that contains the following core attributes for each reference:

- type of reference (e.g., journal, book),
- title,
- book title,
- author(s),
- publication year,
- keywords,
- abstract.

Figure 4 shows an exemplary reference in XML format:

```
<?xml version="1.0" encoding="UTF-8" ?>
<item>
<TY>BOOK</TY>
<TI></TI>
<AU>Jaeger, Attila</AU>
<Y1>2016</Y1>
<KW>Vehicular ad hoc networks (Computer networks)
Automobiles--Electronic equipment  COMPUTERS / Online Services
SCIENCE / Environmental Science  TECHNOLOGY + ENGINEERING /
Automotive</KW>
<AB>Attila Jaeger develops an application which notifies a vehicle's
driver of upcoming road weather dangers. This application maps the
information evaluated by in-vehicle sensors in order to draw
conclusions on the current weather condition. Comprehensive data
basis is gained by sharing information with other vehicles using
Car-to-X communication. In order to prove usability of the presented
approaches, the developed application and selected concepts are
implemented and deployed within the context of large scale Car-to-X
field operational trials simTD and DRIVE C2X. Car-to-X communication
is considered as the next major step towards a significant increase
in road safety and traffic efficiency.</AB>
<T1>Weather Hazard Warning Application in Car-to-X Communication :
Concepts, Implementations, and Evaluations</T1>
</item>
```

Figure 4: XML Structure of an Exemplary Reference

After crawling, parsing and indexing the literature data, WEX provides a number of text analytics methods, which are described below.

3.3 Search Queries

The literature data can be queried with WEX by a search engine. Figure 5 shows the truncated search for the keyword *car service* with the corresponding results (n=574).The found keyword is marked in the abstract of the respective source (keyword in context, KWIC analysis). Complex queries can also be formulated using logical operators (NOT, AND, OR), whereby the search can be restricted to individual fields or document

languages (e.g., English, German). With the help of this analysis method, literature data can be checked with regard to terms for a priori defined concepts, so that, for example, a theory-driven examination of the literature data becomes possible.

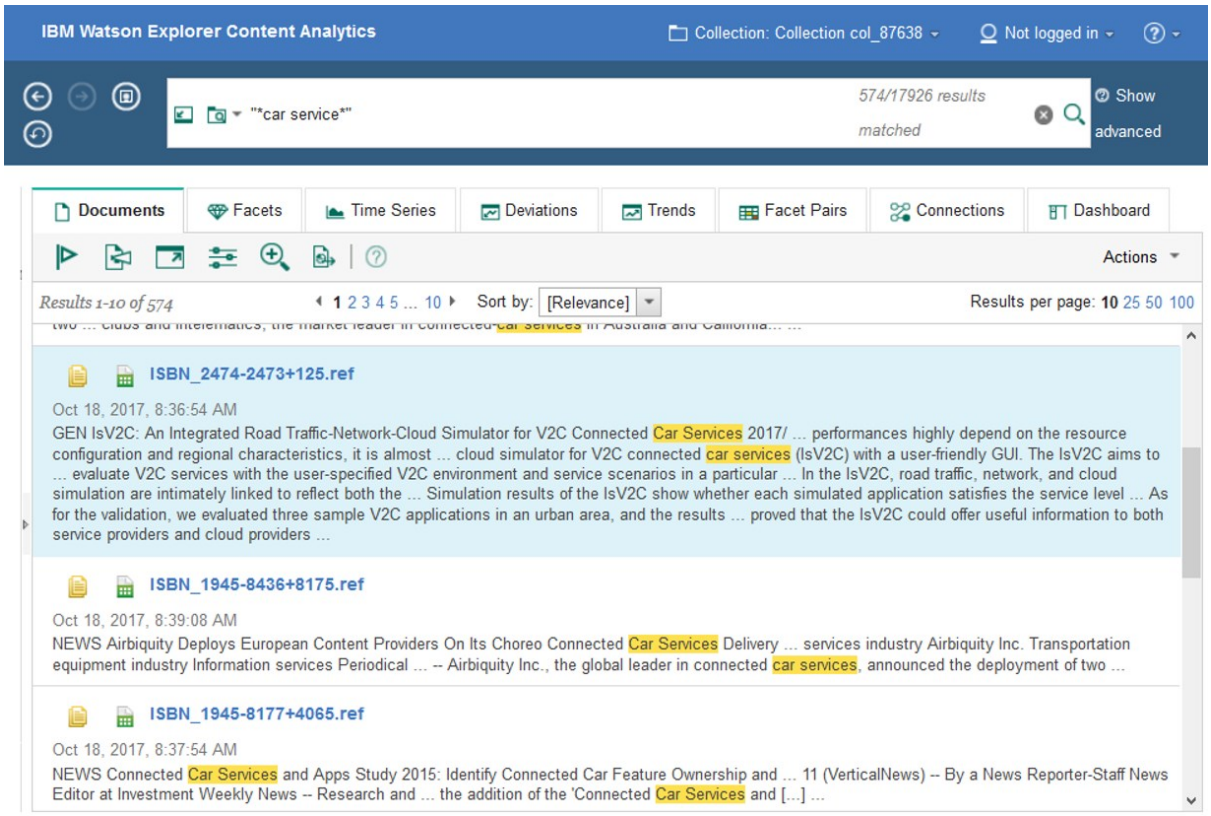

Figure 5: Search Query and Hit List

3.4 Analysis of Word Types and Subgroups

It is also possible to identify different word types (e.g., nouns, verbs) and sentence parts (e.g., noun sequences, verb-substantive/adjective-substantive combinations). Figure 6 shows an excerpt of the noun sequences containing the term *car*, indicating the absolute frequencies (frequency analysis). In this way, frequent concepts (e.g., *car service*, *car platform*) can be identified in the literature database. This is an explorative instrument for inductive derivation of potentially relevant concepts from the references.

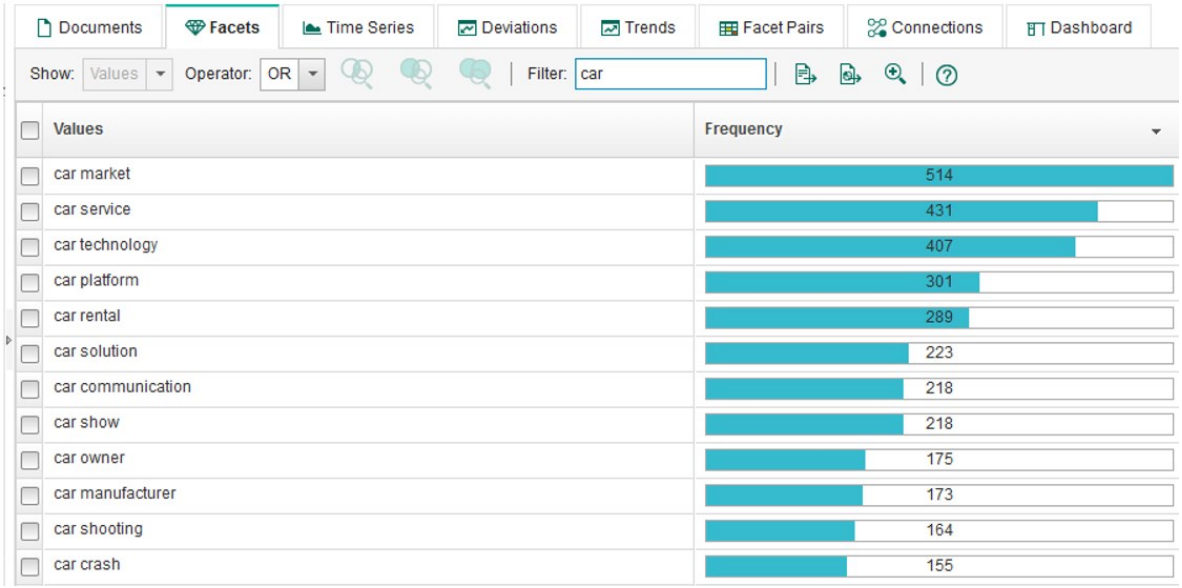

Figure 6: Filtered Frequency Analysis of Noun Sequences

The analysis of word types can also be carried out on a subgroup. In this case, it is possible to identify relationships between concepts in the identified subgroup. Figure 7 shows the noun sequences for all literature sources (n=301) in which the term *car platform* appears. This list is sorted according to the

correlation, which with values greater than 1 indicates a special concentration (density) of the respective noun sequence in the subgroup. The example shows that the concept *car platform* is linked to nouns such as *language understanding*, *voice recognition* and *car service platform*.

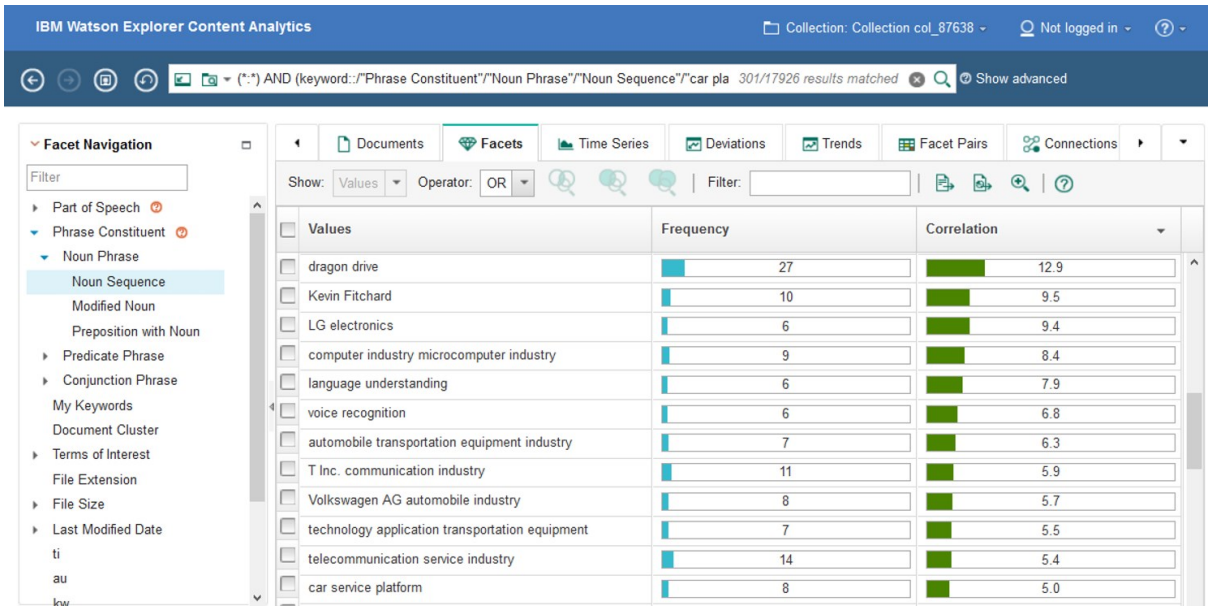

Figure 7: Subgroup Analysis for Literature Sources with the Noun Sequence *Car Platform*

3.5 Time Series Analysis

A time series analysis can be used to answer the question of how publications have developed over time, using the publication year as the time attribute. Any terms can be filtered, so that the temporal diffusion of individual concepts becomes transparent. Figure 8 shows the increase in publications for the research field examined over time.

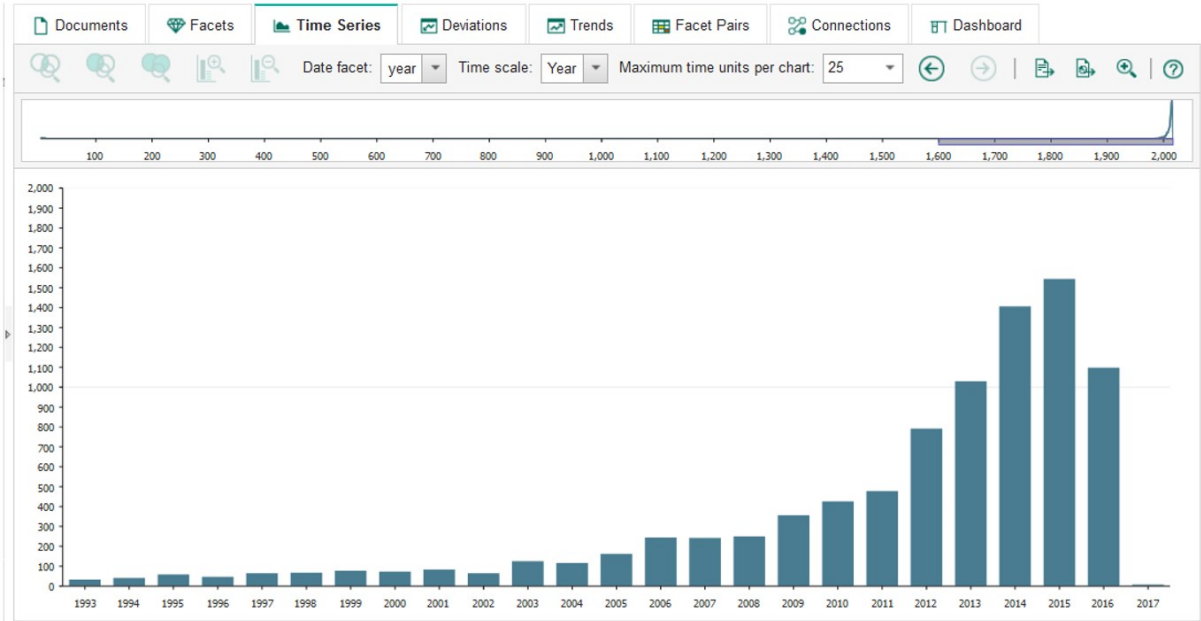

Figure 8: Development of the Literature Base over Time by Year of Publication

3.6 Correlation Analysis

By means of correlation analysis, relationships between different concepts of the domain can be determined. Figure 9 shows a network of highly correlated noun sequences and modified nouns, from which technical connections between the concepts become evident. For instance, the triple *connected car*, *big data*, and *smart city* suggests that those three concepts form a new thematic context. This result can be used for the formulation of new research questions.

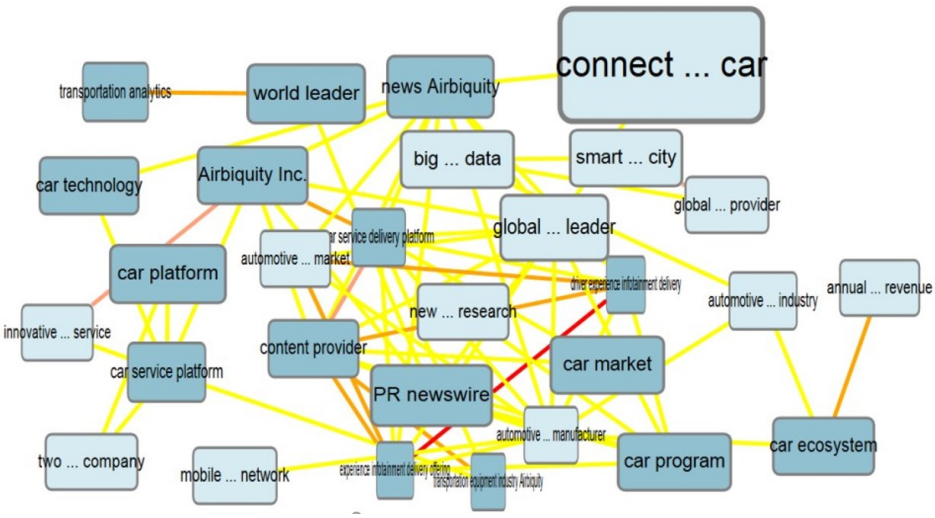

Figure 9: Correlation Analysis for Noun Sequences and Modified Nouns

4. Lessons Learned

We have designed and developed a prototype transforming literature-intensive research processes towards further digitalization. The presented prototype has proven its effectiveness and efficiency in several research projects. One of those research projects is illustrated exemplary in this contribution. In contrast to existing systems, our prototype automates the analysis of published content. Hence, it helps to investigate the status quo of existing and emerging research fields in literature and to identify their central concepts, methods and practical references. With reference to the taxonomy for literature analyses shown in Figure 1, text analytics provides the potential for increasing the coverage of literature analyses. Thus, even extensive collections of literature sources with several hundred thousand abstracts can be examined, which is expected to increase the quality of the resulting research output. Furthermore, the presented functionalities allow the analysis of the literature database according to different criteria. Thus, not only the historical development of concepts can be traced with the help of time series analysis, but also the referencing of conceptual-theoretical foundations and research methods can be substantiated in detail, for example, through search queries and exploration of word parts. Due to the possibility to analyse extensive literature sources automatically, our approach is not only relevant for addressees from science, but also opens up potential to meet the informational needs of decision-makers in politics and business practice.

References

Bornmann, L., Mutz, R. (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *J. Assoc. Inf. Sci. Technol.*, 66(11), 2215-2222.

Bortz, N., Döring, J. (2016). Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften, 5th Ed., Springer, Berlin.

Cooper, H.M. (1988). Organizing Knowledge Syntheses: A Taxonomy of Literature Reviews. *Knowledge in Society*, 1(1), 104-126.

Dann, D., Hauser, M., Hanke, J. (2017). Reconstructing the Giant: Automating the Categorization of Scientific Articles with Deep Learning Techniques. In: Leimeister, J.M. (Ed*.), Proceedings 13. Int. Tagung Wirtschaftsinformatik (WI)*, 1538-1549.

Feldman, R., Sanger, J. (2006). The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press, New York.

Labes, S., Erek, K., Zarnekow, R. (2013). Literaturübersicht von Geschäftsmodellen in der Cloud. *Proceedings Wirtschaftsinformatik 2013*, 443-457.

Seifert, H., Nissen, V. (2016). Virtualisierung von Beratungsleistungen: Stand der Forschung zur digitalen Transformation in der Unternehmensberatung und weiterer Forschungsbedarf. *Proceedings Multikonferenz Wirtschaftsinformatik (MKWI) 2016*, TU Ilmenau 09.-11.03.2016, 1031-1040.

Steininger, K., Riedl, R., Roithmayr, F., Mertens, P. (2009). Fads and Trends in Business and Information Systems Engineering and Information Systems Research – A Comparative Literature Analysis. *Bus Inf Syst Eng.,* doi: 10.1007/s12599-009-0079-7.

Sturm, B., Schneider, S., Sunyaev, A. (2015). Leave No Stone Unturned: Introducing a Revolutionary Meta-search Tool for Rigorous and Efficient Systematic Literature Searches. *Proc. of the 23rd European Conference on Information Systems (ECIS 2015)*, Research-in-Progress Papers, Paper 34.

Sturm, B., Sunyaev, A. (2017). You Can't Make Bricks Without Straw: Designing Systematic Literature Search Systems. *Proc. of the 38th International Conference on Information Systems (ICIS 2017)*.

Webster, J., Watson, R. (2002). Analyzing the Past to Prepare for the Future: Writing a Literature Review. *MIS Quarterly*, 26(2), S. XIII–XXIII.