Computer and Information Science; Vol. 12, No. 4; 2019
ISSN 1913-8989 E-ISSN 1913-8997
Published by Canadian Center of Science and Education

A Text and Data Analytics Approach to Enrich the Quality of Unstructured Research Information

Otmane Azeroual[1]

[1] German Center for Higher Education Research and Science Studies (DZHW), Berlin, Germany

Correspondence: Otmane Azeroual, German Center for Higher Education Research and Science Studies (DZHW), Schützenstraße 6a, 10117 Berlin, Germany.

Received: October 12, 2019 Accepted: October 29, 2019 Online Published: October 30, 2019

doi:10.5539/cis.v12n4p84 URL: https://doi.org/10.5539/cis.v12n4p84

Abstract

With the increased accessibility of research information, the demands on research information systems (RIS) that are expected to automatically generate and process knowledge are increasing. Furthermore, the quality of the RIS data entries of the individual sources of information causes problems. If the data is structured in RIS, users can read and filter out their information and knowledge needs without any problems. This technique, which nevertheless allows text databases and text sources to be analyzed and knowledge extracted from unknown texts, is referred to as text mining or text data mining based on the principles of data mining. Text mining allows automatically classifying large heterogeneous sources of research information and assigning them to specific topics. Research information has always played a major role in higher education and academic institutions, although they were usually available in unstructured form in RIS and grow faster than structured data. This can be a waste of time searching for RIS staff in universities and can lead to bad decision-making. For this reason, the present paper proposes a new approach to obtaining structured research information from heterogeneous information systems. It is a subset of an approach to the semantic integration of unstructured data using the example of a RIS. The purpose of this paper is to investigate text and data mining methods in the context of RIS and to develop an improvement quality model as an aid to RIS using universities and academic institutions to enrich unstructured research information.

Keywords: research information systems (RIS), heterogeneous databases, unstructured research information, pre-processing, text and data mining, knowledge discovery, data quality, effective decision-making

1. Introduction

The flood of research information that reaches every research data manager is steadily increasing. Multiple and heterogeneous data sources must be processed, created and saved for further use. As a solution to this, special research information systems (RIS) are available to collect, analyze and save research information. RIS is a ***central database*** or a ***specialized federated information system***, it can supply an overview of the research activities and results, capture, process and manage projects and publications, etc. For more details see the papers from (Azeroual & Abuosba, 2017; Azeroual, Saake & Abuosba, 2018; Azeroual, Saake & Schallehn, 2018; Azeroual, Saake & Wastl, 2018; Azeroual et al. 2019a). In addition, they allow for versatile access to research-related information and can meet the information needs of different stakeholders. Institutions can communicate their entire research profile by making the research activities publicly available. This ensures the transparency of research and contributes to the external presentation of the institution. Scientists can also use the RIS to map their research priorities and research achievements. Recently, information on research activities and results in universities and academic institutions in a variety of forms and heterogeneous data sources has been collected, maintained and published through RIS. However, these are mostly unstructured in various forms and media (Azeroual, et al. 2018). Unstructured data is therefore a major challenge for RIS managers, and especially for universities and academic institutions that manage their research information from heterogeneous data sources in research information systems. Unstructured data problems include, for example, *personal*, *publication*, or *project data* in *Word*, *PDFs*, or *XML* (as described in the *Common European Research Information Format (CERIF)* or *Research Core Dataset (RCD)*). Such data in the integration can be stored in various formats and referred to by different technologies (Azeroual et al. 2019b). In addition, this can have a negative impact on RIS managers and users (e.g. managing wrong decisions, increasing costs and reducing employee satisfaction). In

order to obtain clear information from unstructured data, the existing unstructured data must be specially prepared for large and heterogeneous data volumes. Such processing takes place, for example, in the graphical representation of research information. In this way, research information should be made more understandable to users, as graphics are easier to understand than mere data collections. In this processing and management of research information, text mining and data mining represent a process whereby the unstructured data is processed and ultimately displayed to users graphically. Text and data mining processes can make a valuable contribution to improving the data quality of research data management. RIS defines data quality as the suitability of this data for use in certain required uses (Azeroual, Saake & Abuosba, 2018). These must be error-free, complete, correct and consistent (Azeroual & Abuosba, 2017).

The data sources integrated into the RIS are so large and unmanageable that they can no longer be viewed by a RIS employee and therefore no connections can be found without special aids. On the other hand, research information from various stakeholders is in demand, and because of their growth, and in particular textual data in the ever-expanding research environment, it is necessary to integrate text and data mining into RIS. Text mining algorithms are used to extract useful information and relevant knowledge from heterogeneous data with high accuracy (Nahm & Mooney, 2002). Text and data mining methods are concerned with the processing of structured databases (Rajman & Besançon, 1998) and provide good assistance in managing the bulk of existing data.

Research information must be cleaned, corrected and pre-processed before being integrated into RIS. An important task in the pre-processing of the research information is the data cleaning (Azeroual, Saake & Abuosba, 2018). In this, the structure and formats of the research information are standardized. Only then can be integrated on with sufficient data quality for further analysis and use. Subsequently, text and data mining methods are applied to the pre-processed research information. Methods include basic Natural Language Processing (NLP) tasks, information extraction, data classification and clustering. Thereafter, the results can be evaluated and interpreted in knowledge. Figure 1 clarifies the new research information process, including the *NLP*, *information extraction*, *document classification* and *clustering layers*, before being integrated and stored in RIS.

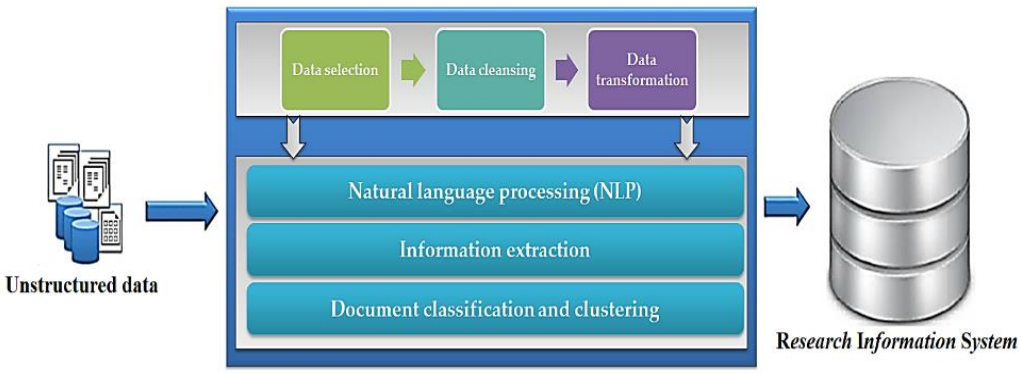

Figure 1. The new research information process and lifecycle integrating new data processing layers (NLP tools, information extraction, document classification and clustering)

The hidden and stored unstructured data and sources in RIS can play an important role in making decisions. After describing the problems of unstructured research information during their acquisition and integration into the RIS, the aim of the paper is to investigate the potentials of using text and data mining methods in the context of RIS and to propose a framework as an aid to RIS users to transform the text sources into structured environments.

The paper is divided into four sections. First, the state of development and the background of text and data mining is shown. It deals with the definition of text and data mining, its interdisciplinarity and practical applications in the context of RIS. It shows how the text and data mining methods can improve the quality of the RIS. This can be illustrated by a developed data quality model. In addition, a case study will be given to demonstrate their usefulness. Finally, the results are summarized in the conclusion.

2. Background

This research focuses on two areas related to RIS: i) data mining; and ii) text mining. Data mining is

interdisciplinary and uses findings from the fields of computer science, mathematics and statistics for the computer-aided analysis of databases. Data mining is the systematic application of computer-aided methods to find patterns, trends or relationships in existing databases (Van der Aalst, 2011). In addition, the relationships are extracted automatically and made available to higher-level goals. The identified patterns can help in decision-making on specific issues. Text mining also known as knowledge discovery from textual databases (Feldman & Dagan, 1995) is a special form of data mining. Using text mining, knowledge of unstructured text data can be extracted and discovered with its multitude of algorithms (Aggarwal & Zhai, 2012).

As (Eler et al. 2018) said, "*Normally, the input data are unstructured and need to be pre-processed before mining tasks. The document pre-processing phase is composed of essential steps for several techniques that deal with textual data, such as text and opinion mining tasks. The pre-processing steps usually filter documents of interest; eliminate irrelevant terms and assign weights to relevant terms*". Text and data mining also helps to improve the search for literature in databases as well as the analysis, storage and availability of information on various websites and search engines are made more efficient and accurate by this technique.

In this emerging research, data mining is used along with text mining algorithms. Text and data mining has become a daily routine for many scientists in a wide range of disciplines as a scientific method. To extract and exploit the unstructured information, over the course of more than thirty years, a variety of text and data mining solutions have been developed that support a wide range of knowledge-gathering processes. There are no studies or concepts on this topic in the context of RIS except the work of (Azeroual, et al. 2018) to demonstrate how to ensure the quality of unstructured data using text and data mining, and which methods of text and data mining can be applied in RIS. It is important to note that this paper refers to the existing and widely used text and data mining methods (e.g. NLP, information extraction, document classification by clustering) and focused on their application in RIS, as these are often discussed and considered in other fields or information systems in the literature. On the other hand, these can be used in the structured collection of documents and put together to new knowledge. In addition, the relevant individual documents are identified in databases with a large number of sources. Therefore, this paper defines the applications of text and data mining in RIS to understand the unstructured research information from heterogeneous systems during integration into RIS and to gain their unknown knowledge. Text and data mining as a new, automated form of using information opens up new opportunities for RIS users. For example, universities and research institutes can benefit from the use of these inspiring research areas (*process mining*) and this can support their decision-making processes. As research information becomes an increasingly important factor for institutions. Only those who have up-to-date, detailed and meaningful information will be able to better position their facility in the long term.

3. Uses of Text and Data Mining Methods in RIS

The introduction of RIS into research institutions means for them that they must provide their required information about research activities and research results in assured quality (Azeroual & Abuosba, 2017; Azeroual et al. 2019a). Poor data quality means that analyzes and evaluations are faulty or difficult to interpret. The occurring quality problems in RIS are on the one hand like spelling mistakes, missing data, incorrect data, wrong formatting, duplicates and contradictions data, etc. and on the other hand like unstructured data formats. These can arise when capturing various independent information systems (such as *external publication databases*, *identifiers* (*ORCID*, *DOIs*, *CrossRef*), *external project data*, etc.) and different standardized exchange formats (e.g., from the *CERIF* or *RCD data model*). To ensure the quality of the data, data quality dimensions (such as completeness, correctness, consistency and timeliness) can be used in RIS to support research decision-making (Azeroual, Saake & Wastl, 2018). For example, low data quality can negatively impact business processes and lead to erroneous decision-making. Text and data mining methods can be applied to unstructured data in RIS. The goal of text mining is similar to that of data mining. Basically, it is about information search and information retrieval from heterogeneous data sources, by finding interesting patterns (He, 2013). Nevertheless, there are many similarities in structure to text mining and data mining. The following steps are required to obtain information from unstructured data in context of RIS (Natarajan, 2005; Feldman & Sanger, 2007):

1. Application of pre-processing routines on heterogeneous data sources.
2. Application of algorithms for the discovery of patterns.
3. Present and visualize the results.

The main difference in terms of data mining is the pre-processing steps used. For text mining, it is necessary to recognize and filter representative features from the natural language heterogeneous data sources and thus to create a structured intermediate form from the texts. In the context of RIS, the following methods can be used in the pre-processing steps: Natural Language Processing (NLP), information extraction, document classification

by clustering.

3.1 Natural Language Processing (NLP)

NLP methods could be used to structure the text documents to be analyzed, with the aim of capturing the meaning of the text being studied. A simple definition for NLP "*is the attempt to extract a fuller meaning representation from free text. This can be put roughly as figuring out who did what to whom, when, where, how and why*" (Kao & Poteet, 2007).

Example of NLP applications methods (Mehler & Wolff, 2005; Miller, 2005):

- Spell checking and correction: By spelling the word and identifying the meaning of a word in context, a correct spelling checker is possible.
- Information gathering: By recognizing syntactic and semantic dependencies, it is possible to extract specific information from a text.
- Question answering: Through syntactic and semantic analysis of a question, a computer can automatically find appropriate answers.
- Machine translation: By clarifying the meaning of words as a single or in context, a correct translation is feasible.

The three main analysis processes of NLP are morphological, syntactic and semantic analysis. In the first step, the text is divided into individual words (tokenization) and these are traced back to their root word and to lemma of the word (stemming). Subsequently, the words are marked, they are annotated. These annotations take Part-of-Speech (POS) taggers, where parts of speech are assigned and parsers that determine the word order in a given sentence. POS tags use dictionaries that capture words and words that they can accept. In the final step, a semantic analysis of the meaning-dependent decomposition and categorization of text is performed. This can be assigned using Named Entity Recognition (NER).

Figure 2 illustrates an example of the NLP features used to analyze a research document before being integrated into the RIS.

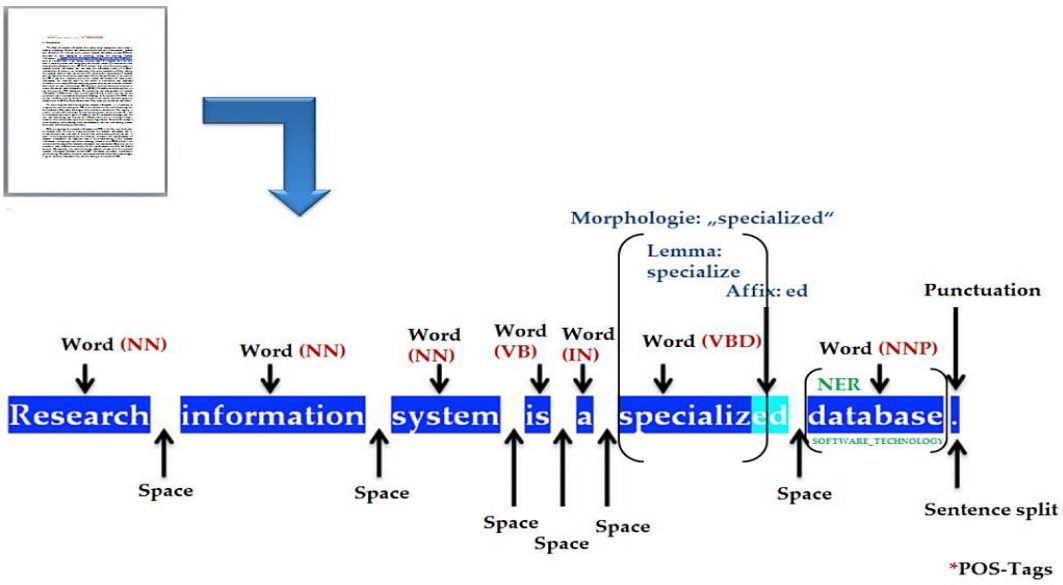

Figure 2. POS data annotation before integration in RIS

3.2 Information Extraction

The main goal of information extraction in RIS is to extract structured information from unstructured or semi-structured text. Relevant information, such as names of authors, locations, or institutions contained in the publication text, are extracted automatically.

This information can be passed directly to a user or other applications such as search engines or databases (Weiss, Indurkhya & Zhang, 2010). Named Entity Recognition (NER) can be the most important task of information

extraction in RIS. A named entity is a word or series of words that designates an object of reality. NER has the task of recognizing these names from a text and assigning them to predefined types. For more details about the functionality of the NER, see the related papers (Collins & Singer, 1999; Cucerzan & Yarowsky, 2002; Asahara & Matsumoto, 2003; McCallum & Li, 2003; Nadeau & Sekine, 2007; Rao, McNamee & Dredze, 2012).

It experimented with 50 selected Web of Science articles containing 152,101 tokens, including 149,890 words and the remaining punctuation. The record contains 204 author names, of which different were organizations and their places. A sequence tag CRF model found in the Stanford NER CRFClassifier was used to extract and annotate named entities. Figure 3 gives an example of entity extraction on a publication text from which organizations and locations were extracted. On the one hand, there is the possibility of annotating and highlighting the entities directly in the text with the corresponding entity type. On the other hand, entities can also be found in a list of annotations, in which the start and end point or starting point and length of the entity in the text and the entity type are noted and returned. In NER, entities in the text are localized (by offset and length or endpoint) and assigned to an entity type. In addition, entity annotations can be supplemented by confidence values which should provide a measure of the precision of the extracted entity.

For the combination of different entity extraction services, the classification schemes used for the entity types play a major role. The most common schemes in the context of RIS are as follows:

- Named Entities (ENAMEX) with PERSON, LOCATION and ORGANIZATION,
- Time expressions (TIMEX) with DATE and TIME,
- Number expressions (NUMEX) with MONEY (financial terms) and PERCENT (percentages).

Unstructured data source

The research information system FACTScience belongs to the manufacturer QLEO Science GmbH. QLEO was founded in 1998 under the name FACT GmbH on the basis of a public private partnership with the Charité - Universitätsmedizin Berlin.

Annotated entities

The research information system FACTScience belongs
to the manufacturer <ORGANIZATION>**QLEO Science GmbH**</ORGANIZATION>.
<ORGANIZATION>**QLEO**</ORGANIZATION> was founded in 1998 under the name
<ORGANIZATION>**FACT GmbH**</ORGANIZATION> on the basis of a public private
partnership with the Charité - Universitätsmedizin <LOCATION>**Berlin**</LOCATION>.

Annotation list	Type	Start	Offset	Confidence
Organization	39	6	0	93
Organization	56	6	0	95
Organization	109	6	0	98
Location	71	8	0	81

Figure 3. Extracting entities from the example of publication text in RIS

3.3 Document classification by Clustering

Clustering algorithms can be used in RIS to quickly find and group similar content of documents or words, as well as to detect duplicates (see Fig. 4).

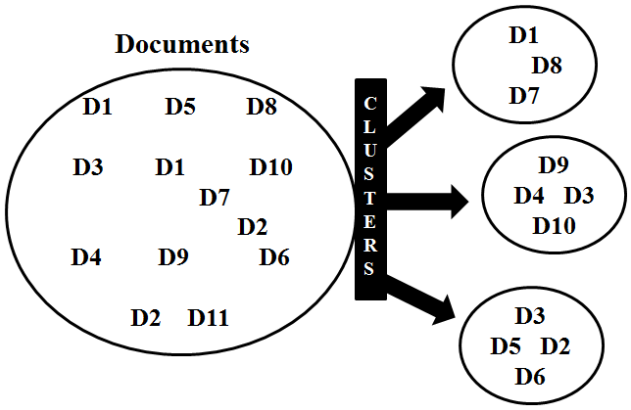

Figure 4. Formation of clusters

The cluster analysis allows building a structure for the objects. Unlike classification, clustering does not use a predefined set of terms or taxonomies that are used to group the documents. Instead, cluster analysis allows to build a structure for the objects. The goal of cluster analysis is to maximize differences between groups and to minimize differences within each group as much as possible.

The process of cluster analysis or document clustering in the context of RIS can be traversed in three phases:

1. Preparation of data
2. Determination of similarities between data objects or document representations
3. Grouping of data objects or document representations

To determine the similarity between documents, different similarity measures are defined. "*A similarity measure is a relation between a pair of objects and a scalar number. Common intervals used to mapping the similarity are [-1, 1] or [0, 1], where 1 indicates the maximum of similarity*" (Cassisi et al. 2012).

In order to consider the similarity between two numbers x and y, the following is assumed (Cassisi et al. 2012):

$$numSim(\mathrm{x},\mathrm{y}) = 1 - \frac{|X - Y|}{|X| + |Y|} \tag{1}$$

Let two time series X=$x1, \ldots, xn$, Y=$y1, \ldots, yn$, some similarity measures are (Cassisi et al. 2012):

Mean similarity defined as:

$$\mathrm{tsim}(\mathrm{X},\mathrm{Y}) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{numSim}(\mathrm{x}_i,\mathrm{y}_i) \tag{2}$$

Root mean square similarity:

$$\mathrm{rtsim}(\mathrm{X},\mathrm{Y}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \mathrm{numSim}(\mathrm{x}_i,\mathrm{y}_i)^2} \tag{3}$$

And peak similarity:

$$\mathrm{psim}(\mathrm{X},\mathrm{Y}) = \frac{1}{n}\sum_{i=1}^{n}\left[1 - \frac{|\mathrm{x}_i - \mathrm{y}_i|}{2\max(|\mathrm{x}_i|,|\mathrm{y}_i|)}\right] \tag{4}$$

There are several algorithms that form classes of documents because of these similarity measures. In the context of RIS, only *k-means* and *hierarchical clustering* are considered, to which many have referred in the literature and used by the author in practice.

K-means is a classic and widely used method of clustering. The basic idea is simply to distribute the amount of documents on *k* clusters of similar documents (see Fig.5).

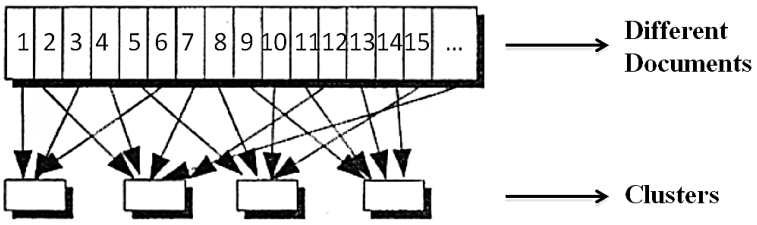

Figure 5. Function of the *k-means* algorithmus

There are 6 steps that describe the *k-means* algorithm:

1. Distribute all documents on *k* clusters
2. Compute the mean vector for each cluster using the following formula

$$E(k) = \sum_{i=1}^{n} \frac{(x^i - m_{ci})^2}{n} \tag{5}$$

3. Compare all documents with the average vectors of all clusters and note the most similar for each document
4. Move all documents into the most similar clusters
5. If no documents have been moved to another cluster, hold; otherwise go to point (2).

Figure 6 is a calculation example for *k-means* clustering. The simplified example shows the algorithm for two clusters. A single number (one-dimensional vector) represents one document each. After three steps, the procedure stops. In each step, their average vectors are calculated for the clusters.

	Cluster 1	Cluster 2
Initial:	1, 5, 2, 4, 5	
Step 1:	1, 5 Mean = 3	2, 4, 5 Mean = 3,67
Step 2:	1, 2 Mean = 1,5	5, 4, 5 Mean = 4,67
Step 3:	1, 2 Mean = 1,5	5, 4, 5 Mean = 4,67

Figure 6. Example calculation with *k-means* clustering

Hierarchical (agglomerative) clustering (HAC) is a popular alternative to *k-means*. Clusters are also created here, but arranged in a hierarchical tree structure. Many different similarity measures can be used, including the average, single/complete link, but also the minimum and maximum spacing of documents within a cluster (Yadav et al. 2019). HAC has made significant studies in the theoretical community and in the application by practitioners. See the related papers (Gan, Wei & Johnstone, 2015; Xu et al. 2016; Tie, et al. 2018; Ieva et al. 2019).

HAC algorithm works in four steps:

6. Start with many clusters, each containing exactly one document
7. Find the most similar pair B and C of clusters that do not have a parent node
8. Combine B and C into a parent cluster A
9. If more than one cluster is left without parents, go to (2).

The end result is a binary tree in which the root represents a cluster of all documents. The children each represent

a division of the parent cluster into two smaller ones. Finally, the leaves contain the smallest clusters, usually with only one document at a time.

There are many different ways to group the clusters in such a binary tree (see Fig. 7). That is, it can additionally process the tree to get a more suitable number of clusters. One way to do this would be to cut off the tree from a certain depth, the result being a fixed number of clusters plus a balanced tree. Another approach is to tailor the tree so that the variance becomes as small as possible.

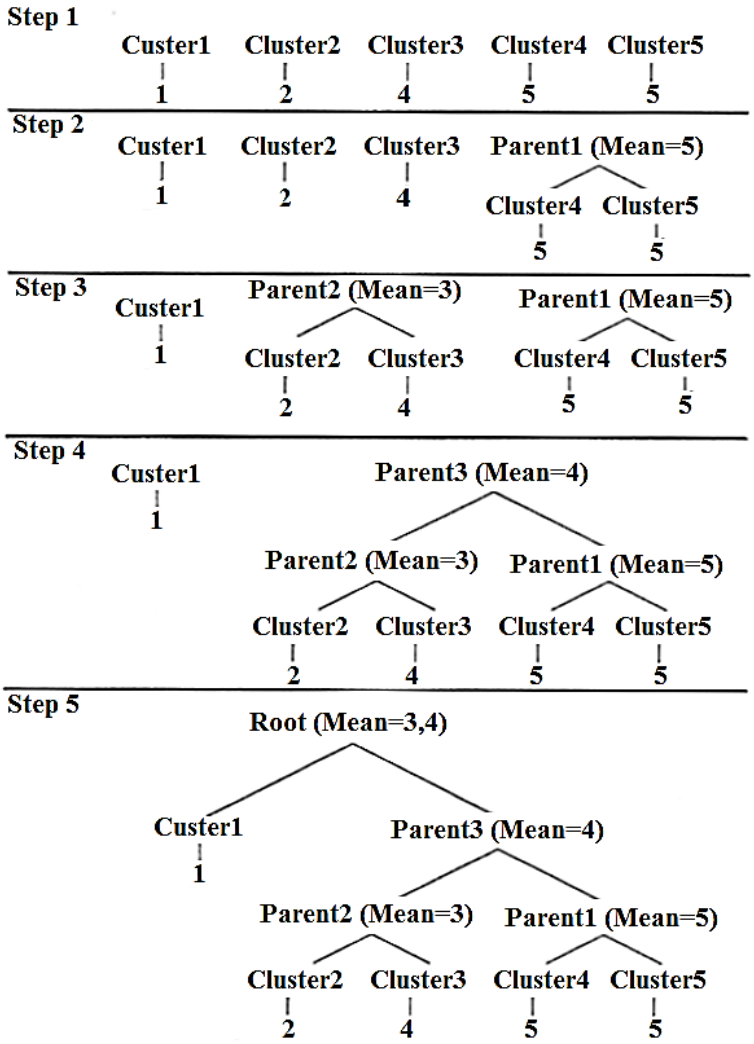

Figure 7. Example calculation of a HAC clustering

The advantage of the HAC is the ability to tailor the resulting binary tree more or less arbitrarily, so a useful and expedient number of clusters can be derived directly from the tree instead of calculating the variance over several runs of different k, as in *k-means* is the case.

In summary, HAC is worthwhile, especially if a hierarchy of documents is required.

If such a hierarchy is not necessary, *k-means* is better suited in many cases. In addition, both algorithms are not only suitable for clustering documents, but also of any data that can be represented as a vector and also used for this purpose.

The benefit of clustering documents in RIS is the combination of the properties of a collection of documents. Since individual documents are analyzed, they can be additionally examined for redundancies and frequencies. Very similar documents such as bug reports with the same problems can be detected. This allows you to count the same objects or prevent redundancies by deleting duplicates. It also reduces the size of the clusters.

The paper is limited to these investigated methods of text and data mining in the context of RIS. These can be served for the purpose of examining and analyzing large amounts of text. The use of the text and data mining

allows the RIS user to plan individual analyzes and perform them interactively. Nowadays, due to the huge amounts of text (articles, project reports, etc.) in research institutions available from different data sources, it is not possible to read them completely and correctly, and to perform a manual analysis. The methods of text and data mining in the institutions are needed to search the texts for keywords and to analyze the sentence structure and the parts of speech, as well as to convert text into data and to discover their new patterns.

Using the text and data mining methods, a simple example from practice is considered as a document-based article and classified according to important information (see Fig. 8).

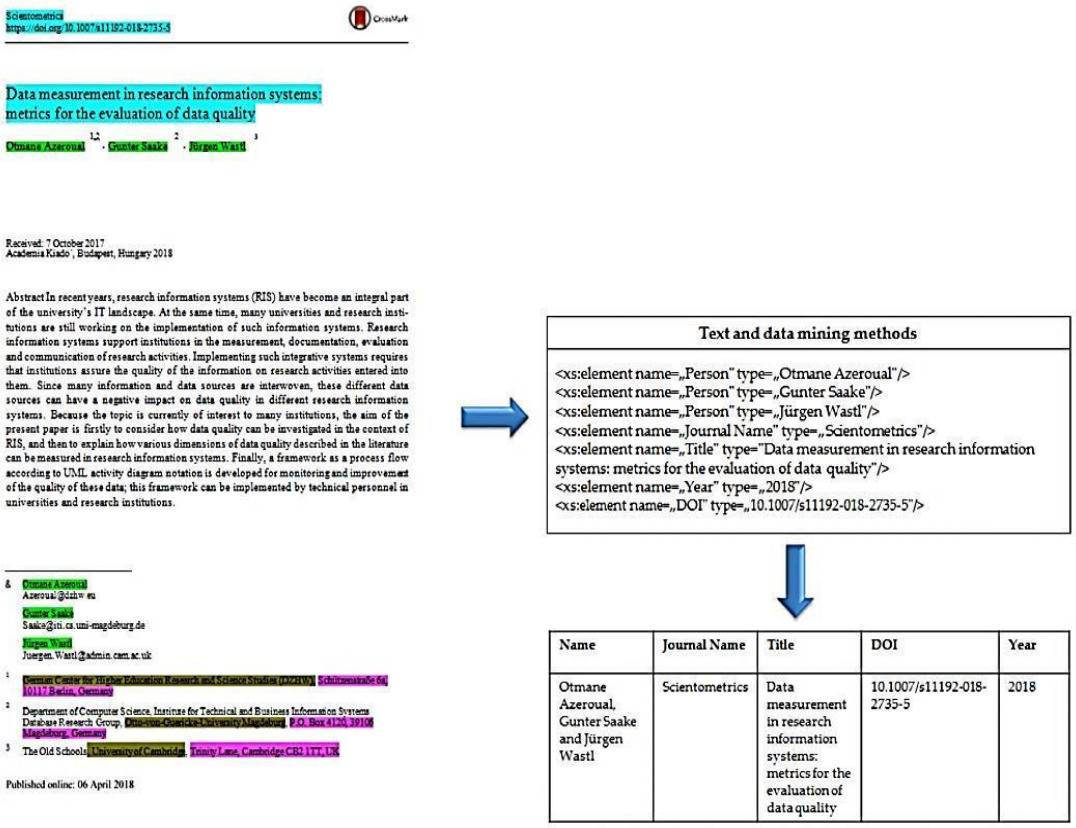

Figure 8. Application of text and data mining during research information integration in RIS

4. Data Quality in RIS including Text and Data Mining Methods

Data quality is a key success factor in text and data mining. The high dimensionality of the data makes manual remediation of data problems difficult. This handling is time consuming, error prone and therefore inefficient and should be avoided if possible. Good data quality ensures that better text and data mining solutions are created. Data quality and text and data mining may be used in an appropriate decision-making strategy in RIS. As well as their goal is to detect, evaluate, explain and finally correct data quality deficiencies in RIS. Discovering the patterns, structures and relationships of data inconsistencies is a task which can be solved very well by text and data mining methods.

Good data quality in RIS, such as text and data mining analysis, can be achieved by developing a model (see Fig. 9). This can support the institutions to permanently carry out the desired quality assurance and gain new knowledge from a variety of sources of unstructured data. This knowledge and its management is the most important success factor and the most crucial resource in a successful academic institution.

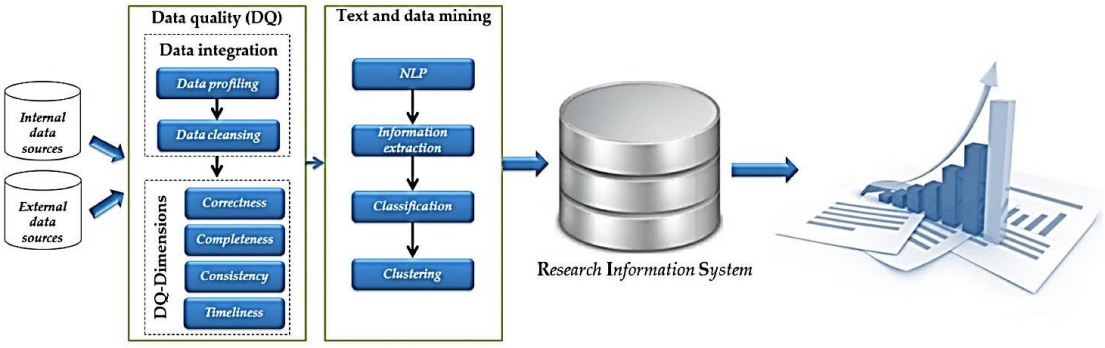

Figure 9. Developed quality model to support using RIS in institutions

The collection and integration of internal and external independent data sources, and in particular the manual input of unstructured research information, requires the high quality of this research information. Better data quality assurance methods should go beyond the steps of *data profiling* (discover data errors), *data cleansing* (correct data errors), and *data quality assessment dimensions* (completeness, correctness, consistency, and timeliness) during the integration phase. These require detailed application knowledge or domain knowledge. Because research information is constantly changing, *data* profiling and data cleansing can be used to identify and manage data errors in RIS (Azeroual, Saake & Abuosba, 2018; Azeroual, Saake & Schallehn, 2018). For a more detailed special application of data cleansing and data profiling in RIS, see both papers (Azeroual, Saake & Abuosba, 2018; Azeroual, Saake & Schallehn, 2018). Using these two methods, duplicates, incorrect and incomplete data, outliers, inconsistent formats etc. in RIS are identified, prioritized, corrected and their causes quickly remedied. Thus, data quality in RIS can be improved and increased. To support the technical tasks of data quality management, today numerous software tools are available that include methods for analyzing (data profiling) and cleansing (data cleansing) the data quality in RIS.

Text and data mining use the methods already described (see Section 3) to extract statements, facts and relationships from textual data and to identify the patterns and relationships between statements which are difficult to recognize. NLP, information extraction and clustering are examples of mining based on an analysis of the text document. These methods operate differently than a simple search. With text and data mining, research information can be searched in great detail, enabling organizations to extract information from large amounts of data that were previously hidden. In recent years, text and data mining methods can be used commercially, to harness the relative complexity of the process and as well as reduce the expenses of the RIS staff and save the cost of the institutions.

5. Conclusion

The paper proposes a new approach to improve existing RIS and gives an overview of the dissemination and potential of text and data mining methods in the context of RIS, so that RIS researchers can take it to the next level. The approach is to integrate text and data mining methods to improve data quality. The main contribution is the generation of knowledge from unstructured data that can be used profitably in RIS. Within the scope of the RIS, there are many possible uses of text and data mining: During the integration of research information numerous textual data are obtained. These can be documents in the form of research papers that the researcher creates in RIS or text documents that RIS managers could create. The methods of text and data mining are able to fully understand and analyze these documents from internal and external data sources. As a result, unstructured research information, structured research information (*metadata*) can be worked out.

The data quality management approach provides the ability to resolve data issues in RIS. By using data quality management to enhance data quality in RIS and for text and data mining analysis in particular by discovering unknown connections and structures, this approach should be integrated into the solution expertise to ensure data quality.

There is a strong dependence of text and data mining with data quality. For example, text and data mining searches for patterns in a data set, so the quality of the data being analyzed is critical to success. This problem is given in every data processing, but text and data mining are particularly affected. Text and data mining works reliably and fully automatically for known, well-understood problems. In many cases it must be ensured for the

analysis that the data is available in the required format and quality. There are various automatic methods for doing this, but in most cases they are not interactive or do without visual support. At this point, the principle of *data wrangling* (sometimes referred to *data munging* or *data crunching*) is used. Data wrangling is a process of unprocessed data from a data source for the analysis to gain usable and valuable data. This process is iterative, that is, a constant repetition of various steps in order to get closer to the best possible solution with the help of the knowledge gained. I will examine this topic in the future work in the context of research information systems or research data management and it is planned to work in cooperation with other scientists in the same field.

References

Aggarwal, C. C., & Zhai, C. X. (2012). *Mining Text Data.* Springer-Verlag, New York, NY, USA. https://doi.org/10.1007/978-1-4614-3223-4

Asahara, M., & Matsumoto, Y. (2003). Japanese named entity extraction with redundant morphological analysis. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, NAACL '03*, Edmonton, Canada, May – June, vol 1, pp.8-15. https://doi.org/10.3115/1073445.1073447

Azeroual, O., & Abuosba, M. (2017). Improving the data quality in the research information systems. *International Journal of Computer Science and Information Security*, *15*(1), 82-86.

Azeroual, O., Saake G., & Abuosba, M. (2018). Data quality measures and data cleansing for research information systems. *Journal of Digital Information Management*, *16*(1), 12-21.

Azeroual, O., Saake G., & Schallehn, E. (2018). Analyzing data quality issues in research information systems via data profiling. *International Journal of Information Management*, *41*(8), 50-56. https://doi.org/10.1016/j.ijinfomgt.2018.02.007

Azeroual, O., Saake G., & Wastl, J. (2018). Data measurement in research information systems: metrics for the evaluation of data quality. *Scientometrics*, *115*(3), 1271-1290. https://doi.org/10.1007/s11192-018-2735-5

Azeroual, O. et al. (2018). Text data mining and data quality management for research information systems in the context of open data and open science. In *Proceedings of 3rd International Colloquium on Open Access - Open Access to Science Foundations, Issues and Dynamics*, Rabat/Morocco, 28-30 November, pp. 29-46.

Azeroual, O. et al. (2019a). Quality of research information in RIS databases: A multidimensional approach. In *Proceedings of 22nd International on Business Information Systems, BIS 2019*, vol 353, pp. 337-349. https://doi.org/10.1007/978-3-030-20485-3_26

Azeroual, O. et al. (2019b). Solving problems of research information heterogeneity during integration – using the European CERIF and German RCD standards as example. *Information Services and Use*, *39*(1-2), 105-122. https://doi.org/10.3233/ISU-180030

Cassisi, C. et al. (2012). Similarity measures and dimensionality reduction techniques for time series data mining. *Advances in Data Mining Knowledge Discovery and Application*. Adem Karahoca, IntechOpen. https://doi.org/10.5772/49941

Collins, M., & Singer, Y. (1999). Unsupervised models for named entity classification. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pp. 100-110.

Cucerzan, S., & Yarowsky, D. (2002). Language independent ner using a unified model of internal and contextual evidence. In *Proceedings of the 6th Conference on Natural Language Learning, COLING '02*, Stroudsburg, PA, USA, vol 20, pp. 1-4. https://doi.org/10.21236/ADA460570

Eler, D. M. et al. (2018). Analysis of Document Pre-Processing Effects in Text and Opinion Mining. *Information*, *9*(4), 100. https://doi.org/10.3390/info9040100

Feldman, R., & Dagan, I. (1995). Knowledge discovery in textual databases (KDT). In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95)*, AAAI Press, Montreal/Canada, 20-21 August, pp. 112-117.

Feldman, R., & Sanger, J. (2007).*The Text Mining Handbook: Advanced approaches in analyzing unstructured data.* Cambridge University Press, New York, NY, USA. https://doi.org/10.1017/CBO9780511546914

Gan, Q., Wei, W. C., & Johnstone, D. (2015). A faster estimation method for the probability of informed trading using hierarchical agglomerative clustering. *Quantitative Finance, Taylor & Francis Journals*, *15*(11), 1805-1821. https://doi.org/10.1080/14697688.2015.1023336

He, W. (2013). Improving user experience with case-based reasoning systems using text mining and Web 2.0. *Expert Systems with Applications, 40*(2), 500-507. https://doi.org/10.1016/j.eswa.2012.07.070

Ieva, C. et al. (2019). Discovering program topoi via hierarchical agglomerative clustering. *IEEE Transactions on Reliability*, *67*(3), 73-80. https://doi.org/10.1109/TR.2018.2828135

Kao, A., & Poteet, S. (2007). *Natural Language Processing and Text Mining, 1*. Auflage, Springer-Verlag, London, pp. 1-7. https://doi.org/10.1007/978-1-84628-754-1

McCallum, A., & Li, W. (2003). Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the seventh Conference on Natural Language Learning at HLT-NAACL 2003, CONLL '03*, Stroudsburg, PA, USA, vol. 4, pp. 188-191. https://doi.org/10.3115/1119176.1119206

Mehler, A., & Wolff, C. (2005). Perspektiven und Positionen des Text Mining. *LDV Forum*, *20*(1), 1-18.

Miller, T. W. (2005). *Data and text mining: A business applications approach.* international ed., Pearson Prentice Hall, Upper Saddle River, NJ, USA.

Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. *Linguisticae Investigationes*, *30*(1), 3-26. https://doi.org/10.1075/li.30.1.03nad

Nahm, U. Y., & Mooney, R. J. (2002): Text mining with information extraction. In *Proceedings of the AAAI-2002 Spring Symposium on Mining Answers from Texts and Knowledge Bases*, Menlo Park, California, pp: 60-68.

Natarajan, M. (2005). Role of text mining in information extraction and information management. *DESIDOC Bulletin of Information Technology, 25*(4), 31-38. https://doi.org/10.14429/dbit.25.4.3663

Rajman, M., & Besan ̧on, R. (1998). Text mining: Natural language techniques and text mining applications. In Proceedings of Data Mining and Reverse Engineering. *IFIP – The International Federation for Information Processing*, Springer, Boston, MA, pp. 50-64. https://doi.org/10.1007/978-0-387-35300-5_3

Rao, D., McNamee, P., & Dredze, M. (2012). Entity linking: Finding extracted Entities in a knowledge base. *Multi-source, Multilingual Information Extraction and Summarization, Theory and Applications of Natural Language Processing*, Springer, Berlin, Heidelberg, pp. 93-115. https://doi.org/10.1007/978-3-642-28569-1_5

Tie, J. et al. (2018).The application of agglomerative hierarchical spatial clustering algorithm in tea blending. *Cluster Computing*, 1-10. https://doi.org/10.1007/s10586-018-1813-z

Van der Aalst, W. (2011). *Process mining: discovery, conformance and enhancement of business processes.* Springer-Verlag Berlin Heidelberg.

Weiss, S., Indurkhya, N., & Zhang, T. (2010). *Fundamentals of predictive text mining.* Springer-Verlag, London. https://doi.org/10.1007/978-1-84996-226-1

Xu, Z. et al. (2016). MICHAC: Defect prediction via feature selection based on maximal information coefficient with hierarchical agglomerative clustering. *IEEE 23rd International Conference on Software Analysis, Evaluation and Reengineering (SANER),* 14-18 March, Suita, Japan, pp. 370-381. https://doi.org/10.1109/SANER.2016.34

Yadav, N. et al. (2019). Supervised hierarchical clustering with exponential linkage. In *Proceedings of the 36th International Conference on Machine Learning*, Long Beach, California, USA, vol 97, pp. 6973-6983.