Reconstructing the Giant: Automating the Categorization of Scientific Articles with Deep Learning Techniques

David Dann[1], Matthias Hauser[1], and Jannis Hanke[1]

[1] University of Würzburg, Faculty of Management and Economics, Würzburg, Germany
mail@daviddann.de, {matthias.hauser,jannis.hanke}@uni-wuerzburg.de

Abstract. The present paper is concerned with the automation of the process of conducting literature reviews. Manual reviews are getting more and more difficult as the number of publications increases steeply. Against this backdrop, we investigate the application of deep learning techniques for the automation of the time consuming step of comparing and categorizing large sets of scientific literature. In contrast to prior research, we leverage the potential of the word2vec algorithm that provides a more semantic focus of analysis than common text mining approaches. We evaluate our artifact considering an exemplary document collection comprising 906 articles on Radio Frequency Identification. Our results indicate that our word2vec-based system provides better results than a system based on traditional text mining approaches.

Keywords: Text Mining, Literature Review, Word2vec, Design Science, RFID

1 Introduction

Free text is the most natural form of storing information. Nair and Narayanan [1, 2] show that up to 80% of the world's data is stored in the form of unstructured data such as text documents and that such data is growing at 15 times the rate of structured data. In times of proceeding digitalization, capturing the content of this growing data pool becomes increasingly difficult. As a result, the means to structure and understand unstructured data become more and more important.

This problem is particularly evident in the process of creating literature reviews, which summarize the current state of research in a scientific field and are thus a fundamental component in the process of knowledge creation. The Thomson Reuter's Web of Science, for example, contains about 58 million articles. This large number is clearly illustrated by Van Noorden et al. [3] who outline that printing just one page of every item would lead to a stack of papers that would reach almost to the top of Mt. Kilimanjaro. Moreover, Bornmann and Mutz [4] show that the number of publications doubles approximately every 24 years. Obviously, the large number of articles make it difficult to identify relevant information [5]. We conclude that there is a need for improving the process of conducting literature reviews in order to successfully process this increasing number of articles.

13th International Conference on Wirtschaftsinformatik,
February 12-15, 2017, St. Gallen, Switzerland

Dann, D.; Hauser, M.; Hanke, J. (2017): Reconstructing the Giant: Automating the Categorization of Scientific Articles with Deep Learning Techniques, in Leimeister, J.M.; Brenner, W. (Hrsg.): Proceedings der 13. Internationalen Tagung Wirtschaftsinformatik (WI 2017), St. Gallen, S. 1538-1549

To address this information overload problem, firstly, the literature review process should be structured and follow well established frameworks (e.g., [6–8]). This helps the creation of consistent and comparable reviews and allows researchers to extend easily the work of others and so keep the research community up to date [6]. Secondly, techniques for partial automation of the process should be investigated in order to keep up with the ever growing acceleration of article publications [5].

To facilitate the automation of reviews, a structured review process is a necessary prerequisite. In our research, we consider the structured process for conducting literature reviews described by vom Brocke et al. [6] and investigate the automation of the most labor-intensive process step. Their framework encompasses five phases. In the first phase, the focus of the review is defined. In the next step, one has to gain a broad conception of the selected research topic which covers, for example, the understanding of key concepts and the uncovering of relevant search terms. Afterwards, relevant articles for the review are identified based on these search terms. In the fourth step, the literature is analyzed and synthesized. Here, vom Brocke et al. [6] suggest using a concept matrix to categorize the literature. These matrices were adapted for literature reviews by Webster and Watson [7] and map individual articles to the concepts they belong to. In the last step of the framework, a research agenda is created based on the results from previous phases.

In our research, we aim at automating the fourth phase of the considered framework which is especially complex and time-consuming. To this end, we apply data mining techniques to compare and categorize semantic contents of large amounts of scientific articles. To approach this task, the text documents must first be transformed into a structured data form. For this step, systems described in literature mostly rely on the commonly known bag-of-words model (e.g., [5, 9–11]). This model, however, has several drawbacks and is less suited for categorizing literature owing to loss of word order and the linguistic phenomena of ambiguity and synonymy which are not well handled by the model [5]. In contrast, we develop a system based on the word2vec model introduced by Mikolov et al. [12]. This model uses high quality vector representations of words to express their semantic information value and therefore should be more suitable for comparing and categorizing literature.

2 Related Work

Algorithm-based comparisons of scientific documents and their potentials have been of great interest to researchers (e.g., [5, 9–11, 14–16]). These approaches can be roughly categorized by the way they transform unstructured data into structured data. We differentiate between (i) text-based, (ii) citation-based, and (iii) hybrid approaches. Text-based approaches compare the textual content of articles (e.g., [9–11]), link-based approaches the citation links of articles (e.g., [14, 16]), and finally, hybrid approaches build on both approaches (e.g., [5, 15]).

The text-based approaches we found in literature use the bag-of-words model for the data transformation step but consider different text sections (e.g., keywords, titles, abstracts, full texts) and pursue different objectives (e.g., categorization of documents

or development of recommendation systems). Gulo et al. [10] analyze abstracts of articles and use machine learning techniques and Bayesian classifiers for categorizing these articles. In contrast, Wang and Blei [11] develop a recommendation system for scientific literature. To this end, they process abstracts and titles of scientific articles with topic modelling and collaborative filtering techniques. The document collection of Afonso and Duque [9] contains titles of articles, abstracts, keywords and the first page or column of the introductions. They again aim at categorizing scientific literature and therefore compare different automated text-based clustering approaches.

Carpenter and Narin [14] use citation-based approaches for the data transformation step. For this purpose, they assume that journals in the same scientific domain have similar reference patterns and thus refer primarily to each other. As a result, they base their clustering process on cross-citation links of articles. Chen [16] uses this assumption and develops a citation-based system for analyzing and structuring literature. Deploying principal component analysis, he generates a correlation value which serves as a measurement for the relatedness between scientific papers.

Hybrid approaches were proposed by Bolelli et al. [15] and Aljaber et al. [5]. As with the purely text based approaches, both author groups use the bag-of-words model for the data transformation step. However, in contrast to these approaches they also consider the citations of articles. The first group considers complete article texts and the citation graph spanned by the articles for the categorization of documents.[1] In contrast, the second group considers citation contexts which are sequences of words surrounding citation markers within the full texts of the documents. They argue that including the citation contexts addresses the issue of synonymy, one of the drawbacks of the bag-of-words model. This is because, according to the authors, these contexts provide "relevant synonymous and related vocabulary which will help increase the effectiveness of the bag-of-words representation". The authors furthermore benchmark their model against a purely citation-based and a text-based model that considers the full texts of documents. Their results indicate that citation-based approaches are inferior to text-based approaches for document categorizations.

Counter to Bolelli et al. [15] and Aljaber et al. [5], we choose a purely text-based approach. We do, however, not rely on the bag-of-words model for building word representations. Instead, we use the word2vec model introduced by Mikolov et al. [12], that is able to express the semantic information value of big amounts of data in a way the bag-of-words model is not. In contrast to the hybrid approach developed by Aljaber et al. [5], the word2vec model addresses not only the synonymy issue of the bag-of-words model but also the issue of ambiguity and considers the information value of the word order. This makes a combination of link-based and text-based approaches unnecessary and justifies our purely text-based approach.

[1] A citation graph is a graph with papers as vertices and citations as directed edges between citing and cited documents.

3 Design-oriented Research

We pursue a design-oriented research approach to develop an artifact that automatically compares and categorizes scientific literature. The artifact design process follows the guidelines put forward by Hevner et al. [13]:

- **Problem Relevance:** Conducting a structured literature review is a complex and time-consuming endeavor and is getting more and more difficult as the number of publications increases every day [4]. To cope with this trend, it is essential to examine whether available text and data mining techniques are suitable for automating the process of structuring scientific articles.
- **Research Rigor:** In our research, we utilize deep learning and established data mining techniques. The proposed artifact is based on the word2vec model introduced by Mikolov et al. [12] that is capable of processing large amounts of textual data.
- **Design as a Search Process:** The idea of conducting an automated literature review is not completely new, but has been a field of research for several years [9]. Our paper examines in particular whether the word2vec model is suited to automate the fourth phase of vom Brocke et al.'s [6] literature review framework.
- **Design as an Artifact:** We design an IT artifact consisting of three components to provide an automated categorization of scientific literature. The artifact is implemented using the Python programming language.
- **Design Evaluation:** We evaluate the artifact considering an exemplary document collection comprising 906 articles on Radio Frequency Identification. We benchmark the artifact against a system based on the bag-of-words model.
- **Research Contribution:** We propose an artifact that is able to process large numbers of scientific articles and conceptualize them. Therefore, we apply novel deep learning methods that provide a more semantically focus of analysis than ordinary text processing models.
- **Research Communication:** A system that allows to capture large amounts of scientific literature quickly and effectively is an enrichment for scientists from all research disciplines. In addition, our paper addresses an audience with a more technical focus by explaining in detail the design of the proposed artifact.

4 Artifact Description

The artifact architecture is depicted in Figure 1 and consists of three components. The first component generates a word vector model based on the user-generated document collection (Subsection 4.1). This model allows the representation of each document in a vector space. Subsequently, comparing these vectors enables determining the similarity of documents. The output of the first component is a matrix containing all the similarity values among the documents in the collection and is input to the artifact's second component. This component groups the documents based on this matrix with a hierarchical clustering algorithm (Subsection 3.2). Finally, in the third component each cluster is automatically labelled with meaningful keywords through the application of

a keyword extraction algorithm (Subsection 4.3). The keywords describe the clusters and thus inform users about the predominant topics within the individual clusters.

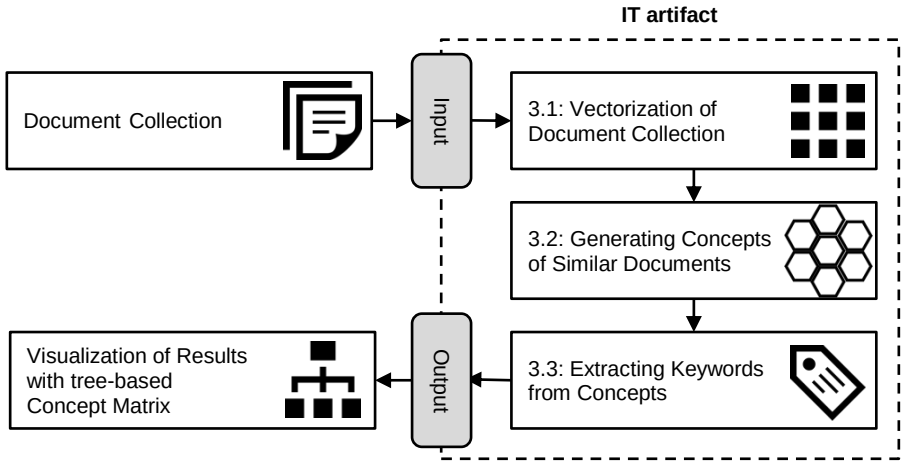


Figure 1. Architecture of the IT artifact

4.1 Vectorization of Document Collection

A user-generated collection of text documents serves as input for the proposed artifact. To represent the single documents as vectors, we train a data model using the Paragraph Vector algorithm introduced by Le and Mikolov [17] which is an extension of Mikolov et al.'s word2vec model [12].

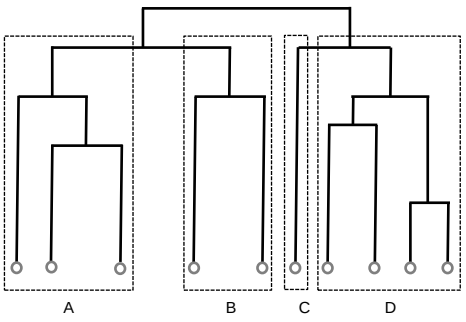
Word2vec aims at training a word vector for each word occurring in the document collection using artificial neural networks. Using this deep learning technique enables us to prevent the drawbacks of the commonly used bag-of-words model. On the one hand, the bag-of-words model does not make use of the information value of the word order which may lead to errors because the model provides identical representations of semantically different sentences in case the same words are used [17]. On the other hand, the model cannot capture the linguistic phenomena of ambiguity and synonymy. The first phenomenon denotes lexically similar but semantically distinct words, the second semantically different but lexically similar words. These weaknesses lead to a more syntactic than semantic focus of the analysis, which makes the bag-of-words model less suitable for comparing and categorizing literature. In contrast, the word2vec model considers the context of words and thus eliminates the problems of traditional text processing models. Mikolov et al. [18] defines the word context as the words that surround a particular word. Leveraging these word contexts allows taking the word order into account. In addition, considering all individual word contexts in the entire document collection results in a high dimensional vector space in which vectors of semantically related words are located in close proximity to each another.

In order to make documents comparable, the Paragraph Vector algorithm builds on the word2vec model to represent each document as a concatenation of its word vectors

in structured form. This vector representation of entire documents allows content-based similarity calculations using traditional distance measures. We apply the cosine distance, which is widely used in text mining applications to quantify the semantic relatedness of documents. We construct a document×document similarity matrix containing similarity values ranging from 0 to 1 with high values indicating similar contents and vice versa. The matrix is input to the artifact's second component.

4.2 Generating Concepts of Similar Documents

The second component automatically groups content-related documents based on the generated similarity matrix. We apply an agglomerative hierarchical clustering approach because it does not require an explicit specification of the number of clusters.[2] We employ the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) algorithm introduced by Sokal and Michener [19]. This algorithm iteratively compares all pairs of the assembled clusters based on the average distance of all elements within them. This allows the construction of a representation of all the documents in the collection in the form of a dendrogram (i.e., a tree-based diagram used for the visualization of clustering results). To receive homogeneous clusters, we automatically merge different branches of the generated dendrogram using the elbow-method introduced by Thorndike [20].[3] Each of the remaining clusters can be considered as columns in Webster and Watson's [7] concept matrix and thus be seen as a distinct concept (see Figure 2). Subsequently, to define the concepts, the artifact's third component applies keyword extraction techniques.



Articles	Concepts			
	A	B	C	D
1	✓			
2			✓	
3				✓
4	✓			
5		✓		
…				

Figure 2. Visualization of clustering results with dendrogram (left) and concept matrix (right)

[2] This approach builds clusters by initially treating each data point as its own cluster. The most similar clusters are then recursively merged until all data points form one single cluster.

[3] The elbow-method aims at finding the optimal number of clusters. The method, therefore, determines the clustering step in which merging two clusters leads to the maximum of variance between all clusters' centroids (which is equal to the maximum value of the second derivative of the average within-cluster distance values function).

4.3 Extracting Keywords from Concepts

The artifact's third component labels each concept with a predefined number of keywords that provide an impression of the documents' content. These keywords are directly drawn from the words occurring in the concept's original texts. To this end, we use the Rapid Automatic Keyword Extraction (RAKE) method developed by Rose et al. [21], which is an unsupervised, domain- and language-independent method for extracting keywords from text collections.[4] RAKE is particularly suited for our system due to its ability to pick highly specific terminology. As a result, our artifact generates meaningful concepts containing content-related scientific documents. As called for by vom Brocke et al. [6], the artifact thus provides a synthesis of scientific documents which facilitates working on the fifth phase of the framework – the generation of a research agenda.

5 Preliminary Evaluation

For evaluation, we instantiate our artifact based on an exemplary collection of scientific articles on Radio Frequency Identification. We consider articles containing the search term "RFID" in title, abstract or the full text.[5] The resulting document collection comprises 906 articles from 39 different journals and conference proceedings.

We benchmark our artifact against a system based on the bag-of-words model.[6] Before feeding the articles into the two systems, the titles, abstracts, keywords and reference sections were removed leaving only the articles' full texts. Based on the previously mentioned elbow-method, each of the systems identified 24 different concepts.

Figure 3 visualizes the systems' results. While the circular dendrogram generated with the bag-of-words-based system (left dendrogram) depicts one very large concept covering almost two thirds of the documents, the dendrogram generated with the word2vec-based system (right dendrogram) shows a more even distribution of concepts.

Figure 3 also zooms into the dendrogram generated with the word2vec model and lists the document titles of two exemplary concepts. Reading the titles of the documents in the concept displayed above the dendrograms suggests that all listed articles are

[4] The algorithm finds representative keywords by first splitting the text into text sequences using delimiters like punctuation or stop words. Then, for individual words in these sequences, word scores are calculated based on word frequency and word degree (i.e., the sum of the length of all sequences the particular word occurs in). The RAKE method then selects the top-scoring words as keywords.

[5] Following suggestions in literature (e.g., [6, 7]) we only consider high-quality articles. Therefore, we relied on the VHB-JOURQUAL 3 ranking [22] and examined articles in the sub-ratings "Operations Research" and "Wirtschaftsinformatik" that were at least B-ranked.

[6] We trained the word2vec model using the genism toolkit (http://radimrehurek.com/gensim/) and the bag-of-words model with the common term-frequency inverse document-frequency weighting using the scikit-learn library (http://scikit-learn.org/). In both cases we relied on the suggested default parameters.

healthcare-related. The titles listed below the dendrograms indicate articles about logistics and supply chain management. Given that the titles of the articles were not part of artifact's input, the results seem promising and indicate a comprehensible conceptualization.

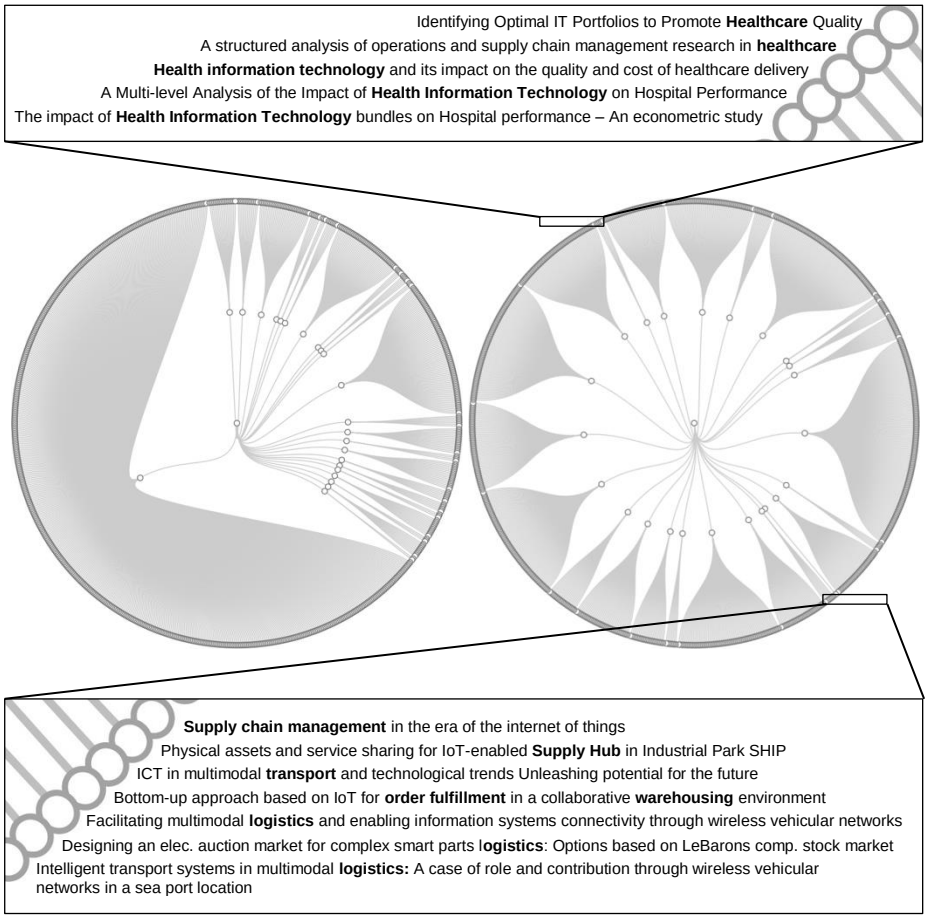


Figure 3. Circular dendrograms generated with bag-of-words-based system (left dendrogram) as well as word2vec-based system (right dendrogram) and two highlighted exemplary concepts

We consider two measures that allow a first quantitative assessment of the two systems' results. Our measures are based on assumptions similar to those introduced by Carpenter and Narin [14] who postulate that "journals which deal with the same subject area will have similar journal referencing patterns" and "journals which deal with the same subject area will refer to each other". Our underlying assumptions are:

1. The content of articles in a specific research domain (e.g., Operations Research) is more likely to be semantically related than content of articles in different research domains.
2. The content of articles in a specific journal (e.g., JAIS, EJOR) is more likely to be semantically related than content of articles in different journals.

We deduce from the first assumption that articles of a specific research domain should be more likely to get grouped into the same concept (research domain score). Second, we conclude from assumption 2 that articles of a specific journal should be more likely to get grouped into the same concept (journal score).

Therefore, we calculate the journal score by

$$\text{journal score} = \sqrt{\sum_{i}^{I}\sum_{j}^{J}\left(\frac{N_j}{N} - \frac{N_{ij}}{N_i}\right)^2 \cdot N_j} \ ,$$

where I is the number of concepts, J the number of journals and conference proceedings, N_j the number of articles in j, N_i the number of articles in i, and N_{ij} the number of j's articles in i. The journal score compares the distribution of journals in individual concepts with the distribution of journals in the entire document collection.[7] An accumulation of one journal's articles within one concept thus leads to a high score.

In analogy, the research domain score is given by

$$\text{research domain score} = \sqrt{\sum_{i}^{I}\sum_{d}^{D}\left(\frac{N_d}{N} - \frac{N_{id}}{N_i}\right)^2 \cdot N_d} \ ,$$

where D is the number of research domains. The research domain score can be considered as a generalization of the journal score. The higher the values of the score, the higher the accumulation of one domain's articles within one concept.

Table 1 summarizes the results for both systems. These preliminary results indicate that the bag-of-words-based system is inferior compared to the word2vec-based system. Although the results are very promising, further evaluations are necessary to fully evaluate the quality of the artifact.

Table 1. Evaluation Results

System	*Concepts*	*Journal score*	*Research domain score*
Bag-of-words-based	24	6.32	13.84
Word2vec-based	24	8.23	20.00

[7] Similar to the statistical measure standard deviation, which considers deviations of single observations, the journal score considers deviations of distributions within concepts. As the individual concepts differ in size, we normalize their squared deviations with their size.

6 Expected Contribution and Future Work

The understanding of an ever-growing number of scientific articles is an increasing challenge for any research domain. Literature reviews are an important instrument for structuring this information. They inform a research community about new findings and enable them to stay up-to-date. However, the number of available publications doubles approximately every 24 years [4]. This can make the process of structuring current research within a research domain very time-consuming, if not impossible. To cope with this trend, the present paper presents an IT artifact that leverages the potential of deep learning techniques to automate the time-consuming fourth phase of the literature framework proposed by vom Brocke et al. [6]. This phase comprises comparing and categorizing large amounts of potentially relevant literature.

Our research has a number of implications for both theory and practice. Regarding theory, we show that text- and data- mining techniques can be applied for this automation step and present an IT artifact capable of automatically categorizing large collections of scientific papers. Furthermore, our preliminary evaluation indicates that the clustering results benefit from the utilization of novel deep learning techniques. Due to the capabilities of Mikolov et al.'s [12] word2vec model to represent linguistic phenomena like ambiguity, synonyms or the information value of word order, it is better suited in the task of processing textual data than the common known bag-of-words model.

For practitioners (in this case scientific researchers), the study provides insights into the capabilities of an automated analysis of scientific document collections of arbitrary size. Implementations of automated literature categorization systems can (i) enable categorizing an amount of scientific literature which is impossible to handle manually, (ii) reduce the time needed to perform the complex but not complicated phase four of document analysis and synthesis and thereby (iii) improve both the researcher's productivity and the quality of the conducted literature review.

We consider the present paper as research in progress. As the evaluation of the artifact is a central activity in conducting rigorous Design Science Research [13], we are currently working on additional possibilities to demonstrate the artifact's capability. Firstly, we are working on including another score for assessing the quality of the results. To this end, we are implementing a measure considering citation patterns of scientific articles based on the assumptions proposed by Carpenter and Narin [14]. These authors identified related articles using a cross-citing matrix and a correlation measure to form article clusters referencing each other. Based on their assumption that articles with similar topics have a higher proportion of citations among themselves, these clusters will allow a direct comparison with the clusters generated by our proposed artifact. In addition, we aim at gathering empirical evidence for the validation of our system leveraging the focus group approach adapted for design science by Tremblay et al. [23]. This quantitative evaluation approach seems especially suitable because it enables direct interactions with domain experts and potential users of our artifact.

References

1. Nair, R., Narayanan, A.: Benefitting from big data: leveraging unstructured data capabilities for competitive advantage. Booz Co. 2, (2012).
2. Nair, R., Narayanan, A.: Getting Results from Big Data-a Capabilities-Driven Approach to the Strategic Use of Unstructured Information. Booz Co. (2012).
3. Van Noorden, R., Maher, B., Nuzzo, R.: The top 100 papers. Nature. 514, 550–553 (2014).
4. Bornmann, L., Mutz, R.: Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references: Growth Rates of Modern Science: A Bibliometric Analysis Based on the Number of Publications and Cited References. J. Assoc. Inf. Sci. Technol. 66, 2215–2222 (2015).
5. Aljaber, B., Stokes, N., Bailey, J., Pei, J.: Document clustering of scientific texts using citation contexts. Inf. Retr. 13, 101–131 (2010).
6. Brocke, J. vom, Simons, A., Niehaves, B., Riemer, K., Plattfaut, R., Cleven, A.: Reconstructing the giant: On the importance of rigour in documenting the literature search process. In: 17th European Conference on Information Systems, ECIS 2009, Verona, Italy, 2009. pp. 2206–2217 (2009).
7. Webster, J., Watson, R.T.: Analyzing the Past to Prepare for the Future: Writing a Literature Review. MIS Q. 26, xiii–xxiii (2002).
8. Schryen, G., Wagner, G., Benlian, A.: Theory of knowledge for literature reviews: an epistemological model, taxonomy and empirical analysis of IS literature. In: International Conference on Information Systems (ICIS) (2015).
9. Afonso, A.R., Duque, C.G.: Automated Text Clustering of Newspaper and Scientific Texts in Brazilian Portuguese: Analysis and Comparison of Methods. J. Inf. Syst. Technol. Manag. 11, 415–436 (2014).
10. Gulo, C.A.S.J., Rúbio, T.R.P.M., Tabassum, S., Prado, S.G.D.: Mining Scientific Articles Powered by Machine Learning Techniques. In: 2015 Imperial College Computing Student Workshop (ICCSW 2015) (2015).
11. Wang, C., Blei, D.M.: Collaborative topic modeling for recommending scientific articles. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. p. 448. ACM Press (2011).
12. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space. CoRR. abs/1301.3781, (2013).
13. Hevner, A.R., March, S.T., Park, J., Ram, S.: Design Science in Information Systems Research. MIS Q. 28, 75–105 (2004).
14. Carpenter, M.P., Narin, F.: Clustering of scientific journals. J. Am. Soc. Inf. Sci. 24, 425–436 (1973).
15. Bolelli, L., Ertekin, S., Giles, C.L.: Clustering Scientific Literature Using Sparse Citation Graph Analysis. In: Fürnkranz, J., Scheffer, T., and Spiliopoulou, M. (eds.) Knowledge Discovery in Databases: PKDD 2006. pp. 30–41. Springer Berlin Heidelberg, Berlin, Heidelberg (2006).
16. Chen, T.T.: The development and empirical study of a literature review aiding system. Scientometrics. 92, 105–116 (2012).

17. Le, Q.V., Mikolov, T.: Distributed Representations of Sentences and Documents. CoRR. abs/1405.4053, (2014).
18. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed Representations of Words and Phrases and their Compositionality. CoRR. abs/1310.4546, (2013).
19. Sokal, R.R., Michener, C.D.: A statistical method for evaluating systematic relationships. Univ. Kans. Sci. Bull. 28, 1409–1438 (1958).
20. Thorndike, R.L.: Who belongs in the family? Psychometrika. 18, 267–276 (1953).
21. Rose, S., Engel, D., Cramer, N., Cowley, W.: Automatic Keyword Extraction from Individual Documents. In: Berry, M.W. and Kogan, J. (eds.) Text Mining. pp. 1–20. John Wiley & Sons, Ltd, Chichester, UK (2010).
22. Verband der Hochschullehrer für Betriebswirtschaft e.V.: VHB-JOURQUAL3. http://vhbonline.org/en/service/jourqual/vhb-jourqual-3/ (2015).
23. Tremblay, M.C., Hevner, A.R., Berndt, D.J.: Focus groups for artifact refinement and evaluation in design research. Commun. Assoc. Inf. Syst. 26, (2010).