Journal of Cleaner Production 190 (2018) 160–168

Contents lists available at ScienceDirect

Journal of Cleaner Production

journal homepage: www.elsevier.com/locate/jclepro

Identifying dominant topics appearing in the Journal of Cleaner Production

Andreas Schober*, Christopher Kittel, Rupert J. Baumgartner, Manfred Füllsack

University of Graz, Institute of Systems Sciences, Innovation and Sustainability Research, Merangasse 18/I, 8010, Graz, Austria

ARTICLE INFO

Article history:
Available online 18 April 2018

Keywords:
Latent semantic analysis (LSA)
Text mining
Sustainability science
Second order science

ABSTRACT

The number of publications in the field of sustainability research has increased rapidly in recent decades and the research topics have multiplied dramatically. It has become difficult to keep track of this highly dynamic field of research. In order to explore the possibilities of computer-aided automated text and meaning capture for the field of sustainability research, we are testing in this paper the method of Latent Semantic Analysis (LSA) with regard to the text corpus published by the Journal of Cleaner Production since 1995. We present the discernible topics identified by this method both in their statistical concept composition and in their temporal evolution and analyze individual, randomly selected contributions in relation to their thematic position in the overall corpus. In particular, the latter gives hope that text mining methods like the here applied LSA could help human readers in the future to maintain an overview in large text corpora and to categorize individual contributions thematically. In this study, as regards content, we identified sustainability education as crucial topic for sustainable development and, additionally, that life-cycle analyses are significantly gaining importance in recent years.

1. Introduction

It is not always easy for researchers to orient themselves and stay up-to-date in a rapidly diversifying scientific field. A great deal of effort often needs to be invested to obtain an overview of the most relevant topics and assess the appropriateness and assignment of publications to these topics. One such highly dynamic research field is sustainability science. 'The ideas of sustainability science are at least two centuries old, but only a decade in practice.' Kates (2012). The scope of activities in this field has become vast during recent years. Some indicators, such as the exponential increase in the numbers of the respective scientific publications[1] (Fig. 1), even seem to indicate that the feedback-driven Matthew effect (Merton, 1968) of sustainability issues is attracting more research as more research is conducted. As a consequence, the field of sustainability research has diversified (Kajikawa et al., 2014) and became even more interdisciplinary (Schoolman et al., 2012), so that topics have become manifold and research orientation has become more difficult. Consequently, mainly within the last decade, several studies tackled the challenge of gauging sustainability science, using several approaches. The research methods applied for these studies were, e. g., literature review (Kajikawa, 2008), bibliometric analysis (Kajikawa, 2008; Bettencourt and Kaur, 2011) and interviews (Miller, 2013). Kajikawa et al. (2014) even rerun and extended the analyses conducted in the former study of Kajikawa (2008), however, still based on the analysis of citation networks. In contrast to these approaches, we decided to directly investigate textual patterns, by use of an automated knowledge discovery method (Witten et al., 2016).

For our investigation we placed the focus on the Journal of Cleaner Production, inspired by the fact that the year 2015 marked the twentieth anniversary of the journal. Thus this journal accompanies a long period of the development of the field of sustainability sciences. With an automated analysis of textual patterns, we undertook what could be called a 'second-order' investigation (Müller and Riegler, 2014) of the debates that have been expressed in publications appearing in this journal. On the one hand, we wanted to spot prolonged and still current foci in this journal and trace back their temporal development. On the other hand, our aim was to test the capabilities of knowledge discovery methods as, e.

* Corresponding author.
E-mail addresses: andreas.schober@uni-graz.at (A. Schober), c.kittel@edu.uni-graz.at (C. Kittel), rupert.baumgartner@uni-graz.at (R.J. Baumgartner), manfred.fuellsack@uni-graz.at (M. Füllsack).

[1] Searching for "sustainability" in article titles, abstracts and keywords on http://www.scopus.com/.

https://doi.org/10.1016/j.jclepro.2018.04.124
0959-6526/

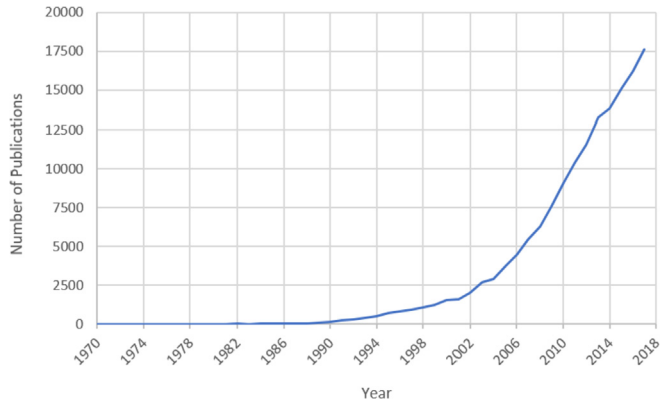

Fig. 1. Number of documents published per year in the field of sustainability science according to Scopus.

g., quickly screening a field and identify thematic focal points of interest.

The paper is organized as follows: in section 2, the applied method is outlined with respect to our analysis of publications appearing in the Journal of Cleaner Production. The major results of this inquiry are presented in sections 3, 4 and 5. In section 3, the ten most dominant topics in the journal are identified and the results of an analysis of their terminological composition and temporal development are provided. In section 4, the results of an analysis of the terminological context of the term *sustainability* as it was used in publications in the journal are presented. In section 5, the results of an analysis of the topical orientation of individual publications are given. An interpretation of all these results is provided in section 6. Section 7 finally draws conclusions from the here presented work.

2. Method

The corpus of publications under analysis comprised 4862 scientific papers with an average length of 9000 words. Simply re-reading, excerpting and categorizing themes from each of these papers was not an option, so we used digital techniques and analytical methods for our analysis. One of these methods, a so-called 'expert system' that can be used to reveal word clusters that represent semantic topics, is Latent Semantic Analysis (LSA) (Deerwester et al., 1990). First developed during the late 1980s, this method allows researchers to conduct a computer-based retrieval of thematic focal points and context information that is comprised in large text corpora.

We applied the LSA method in this way to gather information about the distribution of topics in the corpus of the JCLP. Note that the term *topic* is a *terminus technicus* (Deerwester et al., 1990) used in information retrieval and cannot be viewed as a synonym of a *research topic* in the usual sense. In the context of the method applied in our research, a *topic* consists of the terms that are identified as having a statistical correlation in the analyzed documents using LSA. These groups of terms (which we define as *topics*) were indicators of research areas that could be identified and labeled by the authors after the computational analysis had been completed.

The basis of LSA is a model known as the bag-of-words, which was built on the assumption that grammar and word order can be ignored on the level of analysis at which LSA is conducted. Therefore, researchers adhering to the bag-of-words concept treat the order of words in a text as insignificant with respect to certain aspects of information. Interestingly, this possibility was already mentioned in 1954 by the linguist Zellig S. Harris, long before automatic information retrieval methods were developed (Harris, 1954). On the base of this concept, it was proposed that the nouns appearing in a text represent topics that are covered in the text (Nevin, 2002). The method LSA was built on this assumption and can be used to investigate huge text corpora, such as the one of JCLP, in order to identify the topics covered in a journal as well as their development over time (Sidorova et al., 2007).

In essence, LSA is a structured application that includes different mathematical and statistical methods, all of which are used widely in science and technology. It is closely related to the more widely-known Principal Component Analysis (PCA) methods, which is used in the field of statistics. Like PCA, LSA can be used to reveal specific types of information by analyzing the structure of a dataset in terms of its principal components. In the case of the LSA, the principal components are sets of terms that frequently co-occur in documents of the text corpus. These sets of terms are then considered as *topics*.

The overall framework of a data mining process such as LSA is based on several distinct steps. These steps are data selection, preprocessing, transformation, data mining and interpretation, regardless of the particular application. In the following sections, we explain which steps were performed using the analytical software and how the results were interpreted during the course of the present study. The programming code for the analysis was implemented in python 2.7 (www.python.org), using several data mining and data analysis modules.

2.1. First step: data selection

The corpus[2] of texts (articles and book reviews) that were available online as electronically-accessible files (i.e., PDF) and were published in the Journal of Cleaner Production from 1995 to 2015 represented the initial dataset. The hard copy texts that appeared during the early years of the journal (articles and book reviews that were published up until volume 3, issue 4 in 1995) could not extracted in an automated fashion and were excluded from the study. These electronically-accessible publications, in total 4895 documents, made up the dataset which we used for our analysis.

2.2. Second step: preprocessing

Our goal was to extract the nouns occurring in the text of each publication, which would allow us to investigate the topics occurring in the dataset. Initially, it was necessary to carry out certain preprocessing steps to conduct a semantic analysis of the journal's corpus using computerized methods, specifically part-of-speech (POS) tagging. During this tagging process, each word in the text is assigned to a part of speech (e.g., car - noun, drive – verb). Words that can be assigned to several parts of speech, for example the word *control*, which can be used as both a noun and a verb, are classified by an algorithm depending on the textual context (e.g., *control* followed by an article is classified as a verb while the word is classified as noun if an article precedes it). Different approaches are used during the tagging process, which fall into two groups: a rule-based and stochastic. Each of these approaches serves as a basis of a variety of programs known as POS taggers. For this study, we implemented the Stanford POS Tagger[3] (Manning, 2011), which applies a stochastic Maximum Entropy Markov Model during its tagging process. This is one of the POS Taggers with the highest accuracy (about 97% of words are classified correctly) and it has a

[2] http://www.sciencedirect.com/science/journal/09596526.
[3] http://nlp.stanford.edu/software/tagger.shtml.

Table 1
Excerpt from the term-document matrix that counts the frequency of each noun in each publication.

	1995_3_4_181	1995_3_4_189	1995_3_4_201	1995_3_4_215	1995_3_4_221	1995_3_4_225	…
distanc	1	0	7	0	0	0	…
consider	1	3	1	0	0	0	…
chain	1	2	3	0	0	0	…
control	2	0	1	0	3	0	…
consum	1	0	2	0	0	0	…
burden	5	0	0	0	1	0	…
product	124	20	90	16	6	0	…
paper	8	8	1	0	4	4	…
…	…	…	…	…	…	…	…

user-friendly interface that allows easy implementation.

Several preprocessing steps had to be implemented for each of the downloaded PDF files. First, the plain text was extracted from the PDF, and the python-modul pdfminer[4] was used for this purpose. Then, part-of-speech tagging of the words in the extracted text (using Stanford POS Tagger) was performed. Subsequently, all nouns were derived from the tagged text (whereby proper names were also treated as nouns). Finally, a method called stemming was applied, which reduces nouns to their word stems. For example, words such as *substitute* and *substitution* are reduced to their stem *substitute-*. In this way, the analysis software can treat words with the same stem as equal. We used the Natural Language Toolkit (NLTK) Snowball (Porter 2) Stemmer[5] for stemming. Compared with other stemmers, this one is known to combine high levels of processing speed and result accuracy.

By taking preprocessing steps, we obtained 4895 bag-of-words entities that consisted of only nouns and proper names, with each entity being related to one publication in the journal. All nouns were reduced to their stems and the terms *noun* and *stem* are used synonymously in subsequent sections.

2.3. Third step: transformation

Quantification of the preprocessed data (of the bag-of-words) is required to facilitate mathematical evaluation. This quantification is achieved by transforming the preprocessed data into a term-document matrix (Deerwester et al., 1990). This matrix is generated by counting each noun that appears in each publication. Consequently, the matrix can be understood to represent a countable list of terms. The rows of the matrix represent the terms (nouns), and the columns of the matrix represent the documents (i.e., the publications in the journal). Table 1 exemplarily lists the first eight rows and six columns of the term-document matrix obtained. The column names have the format year_volume_issue_first-page of the corresponding publications.

At this point, in order to obtain more accurate results when the actual Latent Semantic Analysis is conducted, an additional preprocessing step is applied to the term-document matrix. This additional step consists of three parts:

1) Terms are dropped if they only occur in few documents or occur in each document of the overall corpus. Furthermore, terms can also be discriminated with respect to the frequency with which they occur within a text (e.g., through dropping terms, which only occur a few times). On the one hand, dropping these terms reduces background noise and improves the final results. On the other hand, this step saves computation time and memory space. The measure is justified by the fact that including either terms that occur rarely or terms that appear in each document has not been shown to significantly influence the result of the analysis (Sidorova et al., 2007, 2008; Valle-Lisboa and Mizraji, 2007).
2) Weighting of all entries in the term-document matrix leads to a discrimination of terms that occur frequently and, in turn, places an emphasis on terms that occur less frequently. Several approaches of weighting algorithms exist and, depending on the condition of the dataset, the appropriate algorithm must be selected. Various studies have addressed the effects of weighting in information retrieval and demonstrated that weighting (and in particular term frequency–inverse document frequency (TF-IDF) and log-entropy) results in a better performance (Dumais, 1991; Ibrahim and Landa-Silva, 2015).
3) In the last step, each vector in the term-document matrix is normalized (Salton and Buckley, 1988) by applying the Euclidian norm, whereby the absolute value of each vector, which is also referred to as its length, becomes one. This measure of normalization allows documents to be weighted equivalently during the final mathematical evaluation step of the LSA.

To examine the topics that occurred most frequently in the journal, we chose to apply rigid term frequency limits in our analysis. These limits were established in an iterative approach of testing the computational performance and accuracy of the results for several values. As a result, we set the minimum frequency of the terms in a topic to five. Furthermore, we only included terms in the analysis if they occurred in at least ten documents. Weighting was implemented with the TF-IDF weighting factor and normalization was applied. We used the TF-IDF weighting factor after carrying out several tests of various factors, since its application yielded the most coherent results. The log-entropy weighting factor was also tested, but the results generated were insufficient and difficult to interpret, so this factor was neglected.

These additional steps of preprocessing resulted in the production of a weighted and normalized term-document matrix M, which could be directly evaluated using LSA.

2.4. Fourth step: data mining

The core mathematical analytical method using in LSA is Singular Value Decomposition (SVD) (Golub and Kahan, 1965), a method also applied in linear algebra. In general, SVD can be applied to approximate a matrix by producing matrices of lower rank (dimensionality reduction) and analyze the strongest dimensions of a matrix and examine the relationship between the rows and columns of that matrix (latent factor analysis). In LSA, SVD is initially applied to identify the co-occurrence of words in a textual corpus and, thus, can be used to identify latent structures (factors) in that corpus. These structures then are interpreted as semantic topics that are made up of specific clusters of terms. The

[4] https://pypi.python.org/pypi/pdfminer/.
[5] http://www.nltk.org/_modules/nltk/stem/snowball.html.

applied SVD provides the basis of the analysis by revealing the importance of these topics and the relationship between the topics and the analyzed documents. Mathematically expressed, SVD is used to perform the following operation.

$$M = U \times \Sigma \times V^T. \quad (1)$$

In prosaic terms, the TF-IDF-weighted matrix M (mentioned above) is decomposed into three fundamental matrices, U, Σ and V^T (Eq. (1)). Matrix U represents the relationship between the terms (rows of U) and factors representing latent structures (columns of U). Matrix Σ is a diagonal matrix that contains the latent structures in descending order. Matrix V^T represents the relationship between the latent structures (rows of V^T) and documents (columns of V^T).

The document loadings $V\Sigma$, which is the matrix product $V \times \Sigma$ (note that V is the transpose of V^T), is used to generate timelines that revealed the development of topics in the journal over time. The topic composition, determined by the co-occurrence of words, was derived from the matrix product $U\times \Sigma$, which is generally referred to as the term loading matrix $U\Sigma$ (Evangelopoulos et al., 2012).

2.5. Fifth step: interpretation

The interpretation of the values in $V\Sigma$ and $U\Sigma$ was not straightforward due to the fact that both matrices contained positive and negative values. We interpreted the values as probabilistic representations of similarity, whereby positive values indicated similarity and negative values indicated dissimilarity. Consequently, a positive value in $V\Sigma$ indicated the existence of a relationship between the document and topic, whereas a negative value indicated that such a relationship was unlikely. This also was the case for the relationships between terms and topics ($U\Sigma$). For each topic represented in $V\Sigma$, the summation of values related to the document allowed us to evaluate its significance.

In the next section, we present the results of our analysis with regard to the ten semantic topics with the highest significance in the journal.

3. Results of the topical analysis

By applying the LSA method, we could reveal the composition and temporal development of semantically-related word constellations (i.e., *terms* and *topics*) that appeared in publications in the Journal of Cleaner Production from 1995 to 2015. We identified the ten topics that were the most characteristic for this journal, illustrating them as word-clouds (Figs. 3–12) to clearly show the topic composition (i.e., showing the nouns that had the highest contribution to each topic). We limited the word-clouds to twenty-five terms per topic, which we considered sufficient to allow for interpretation and labeling. The larger the contribution of a noun to a topic, the larger the font size of the word in the word-cloud. The word-clouds are sorted by the importance of topics in descending order. Labels, which were based on the identified co-occurrence of nouns in each specific word-cloud, were manually applied as interpretations.

In Fig. 2, the relative importance of each of the ten identified topics for each year is plotted. The y-axis cannot be interpreted in a straightforward way, since the values indicate whether a topic tends to occur or not. A positive value along the y-axis implies that the a topic is an important component in the journal's publications during a year, while a negative value implies that a topic did not play an important role in the journal's publications during that year. The development of these topics and their composition is discussed in the subsequent section. Viewed as a whole, the plot reveals a change in the journal's topical richness. Up until 2006, the curves showed a high degree of variance that was absent from 2007 on.

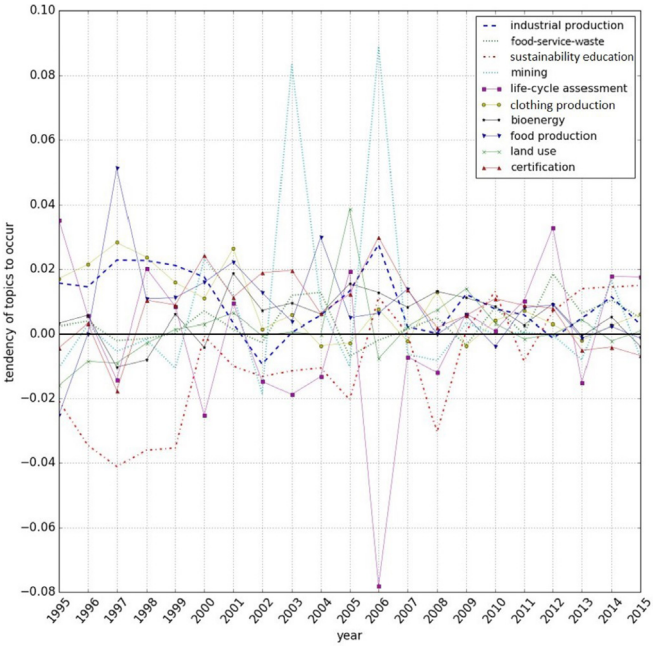

Fig. 2. Temporal development of the ten main topics in the Journal of Cleaner Production from 1995 to 2015.

3.1. Topics in the Journal of Cleaner Production

The results of our analysis, on the one hand, allowed us to identify certain distinct topics such as industrial production (Fig. 3) and mining (Fig. 6), with all of the listed terms clearly referring to the applied label. However, on the other hand, some topics such as land use (Fig. 10) were mainly composed of terms that refer to tourism (e.g., *tourism*, *tour*, *travel*, *hotel*) but also included terms such as *pulp*, *emergy* (Scienceman, 1987) and *network*. These latter, rather indistinct topics allowed for considerable room for interpretation. Therefore, our labeling is preliminary. The ten topics identified as the most characteristic for the journal are discussed in detail below.

According to the results of our analysis, the semantic relation labeled as *industrial production* (Fig. 3) was identified as most characteristic topic appearing in the Journal of Cleaner Production from 1995 to 2015 in total. This topic is composed by terms like

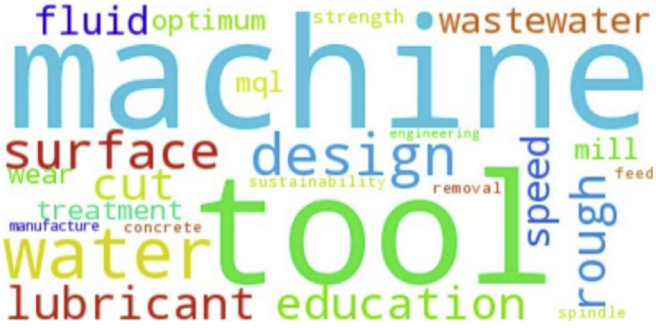

Fig. 3. Industrial Production. The label of this topic is indicated by the terms in the word cloud (e. g. “machine”, “tool”, “wastewater”, etc.). According to our analysis it is the most characteristic topic in the Journal of Cleaner Production.

machine and *tool* as well as additional terms that were clearly related to industrial production. The word-cloud also includes the terms *water* and *wastewater*. This may be an indication for the manner in which sustainability is primarily regarded by papers on industrial production in the journal. The term *sustainability* itself is also among the terms most frequently assigned to this topic, but has minor relevance. Despite the overall relevance of industrial production, its relative importance appeared to decrease over the years (Fig. 2). This decrease might be related to the increase in the topical richness of the journal during this time.

In contrast to industrial production, the topic-cluster of food-service-waste became more frequent in recent years, in particular after 2011. The composition of the word-cloud (Fig. 4) leaves some scope for interpretation. However, the LSA method allowed us to identify this subject as based on agricultural/food production components (e.g., *food*, *farm*, *feed*) and waste components (e.g., *waste*, *landfill*, *compost*) mainly. In addition to terms that clearly referred to food and waste, terms such as *palm*, *oil* and *biodiesel* were associated with this topic. This pointed to the joint production of food and energy sources. Furthermore, the term *service* played a significant role in outlining this topic, which suggested that the food service industry is particularly addressed when the subjects of food and waste are raised.

Sustainability education evolved into a crucial issue in the journal during the years covered by the analysis, especially during 2012–2015. The terms *education*, *university* and *sustainability* were central to this topic, and terms related to energy consumption (e. g., *energy*, *heat*, *biomass*) and the greenhouse effect (e.g., CO_2, *climate*, *greenhouse*) made up this topic (Fig. 5). With *China* and *USA*, the names of the world's main greenhouse gas emitters also appeared among the top terms in this subject. This indicated that sustainability education, as expected, is strongly oriented towards global contexts.

Interestingly, mining (Fig. 6) was the dominant topic in 2003 and 2006, and was somewhat frequently used in 2000 and 2014 as well, but was rather insignificant in all the other years. A closer look at the journal reveals the reason for this: Volume 11, Issue 2 in 2003 was a special issue on the topic of “environmental management in the small-scale mining industry”, and Volume 14, Issues 3–4 and 12–13 focused on “improving environmental, economic and ethical performance in the mining industry”. In 2014, a special issue on mining was published by the JCLP with the title “the sustainability agenda of the minerals and energy supply and demand network”. The subject mining itself mainly builds on mineral names, however, the composition of this topic discloses indications of how sustainability is addressed with respect to this topic in the journal as well. Thus, corporate social responsibility (*csr*) as well as *rehabilitation* appear as parts of this word-cloud in Fig. 6.

Fig. 4. Food-Service-Waste, it is very characteristic for this journal, in particular in recent years. The figure illustrates the most characteristic terms for this topic.

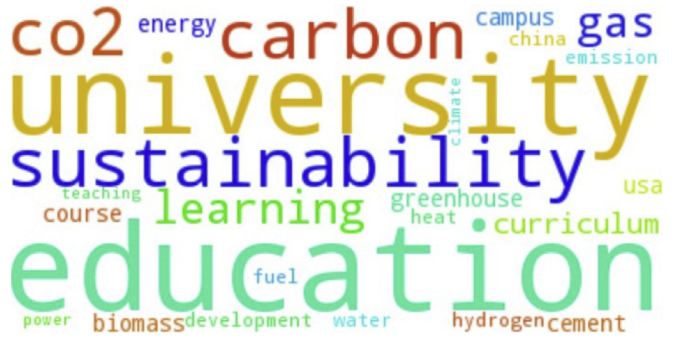

Fig. 5. Sustainability Education. This topic is crucial for the journal, with an increased meaning in the last three years. The figure illustrates the most characteristic terms for this topic.

Fig. 6. Mining is a topic that is discussed in the journal in particular in special issues. This topic was very characteristic in the years 2003 and 2006. The figure illustrates the most characteristic terms for this topic.

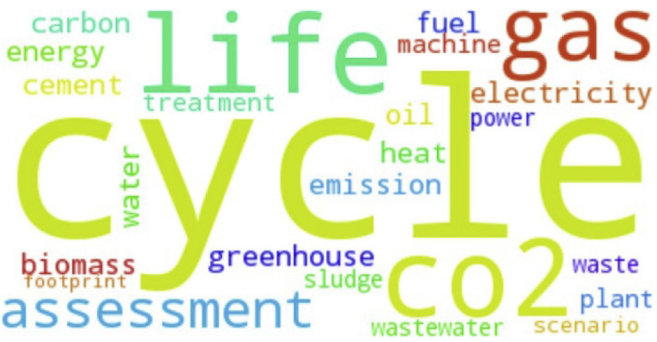

Fig. 7. Life-Cycle-Assessment also is a central research topic in the Journal of Cleaner Production with an increased meaning in recent years. The figure illustrates the most characteristic terms for this topic.

The quantification, comparison and reduction of the environmental impact of products is an essential measure for sustainable development and, hence, is an important topic in the journal. The terms that were related to the topic were labeled as life-cycle-assessment[6] (LCA) (Fig. 7), revealing the areas in which LCA was a subject of discussion. The co-occurrence of the terms *energy*, *greenhouse*, *gas*, *emission* and *electricity* suggests that one area of application is life-cycle greenhouse-gas emissions. Moreover, cement, which is a notoriously large source of carbon dioxide emissions, is part of this topic. However, in total, the topic life-cycle

[6] Note that these nouns as well as all other nouns in our analysis were identified and counted as single words and not as word compositions.

assessment is focused on energy sources. The relative contributions of papers related to LCA were remarkably high in four out of the five last years included in the analysis. Interestingly, the topic seemed to play a conspicuous, tangential role in 2006, at least with its focus on energy sources.

3.1.1. Clothing production

Materials (e.g., *cotton*, *textile*, *leather*) and tools (e.g., *acid*, *sodium*, *tan*) principally formed the topic we labeled as clothing production. The word-cloud on this topic (Fig. 8) includes *pollution*, *prevention* and *remanufacture*. These terms indicate which strategies of sustainable production were primarily mentioned in this field. Furthermore, CO_2 and *China* were terms that appeared with this topic. The timeline of the topic in Fig. 2 illustrates that clothing production was considered particularly relevant during the early years of the journal up until 2001.

The advancement of bioenergy (Fig. 9) was an important topic of political and scientific debates, especially during the early years of the 21st century. In JCLP, discussions that focused on this subject frequently appeared during this period as well. Thus, bioenergy played a steady role in the journal from 2001 to 2012. Interestingly, the frequency of this topic decreased during the most recent years (e.g., from 2012 to 2015). The composition of the topic itself mainly built on the keywords *oil*, *biodiesel* and *biomass*. Furthermore, *hydrogen* and *ethanol* also contributed significantly to the topic. In addition to these terms, which outline the subject, a few other terms suggested what related papers discussed in a wider context. These terms identified were on the one hand *supply* and *chain* and, interestingly, *iso* (for International Standardization Organization) on the other.

3.1.2. Food production

Between 1995 and 2015, food production (Fig. 10) was repeatedly discussed in the JCLP. In total, however, its importance decreased. In 1997, 2004, this subject had the biggest impact in the journal. In 2004, there was a special issue (Volume 12, Issue 5) on sustainable agriculture. Volume 5, Issue 1–2 in 1997 was a special issue on industrial ecology, however, one-third of the publications in this issue referred to *food*. Regarding its composition, the topic food production was one of the more indistinct topics we identified. It is composed by *food*, *farm*, *milk* and similar terms that clearly refer to food production and, moreover, the share of the term *food* is by far the largest in this topic. However, the terms *heat* and *leather* appeared in this topic as well. This co-appearance suggests that these three subjects (*food*, *heat* and *leather*) have similar textual components in publications.

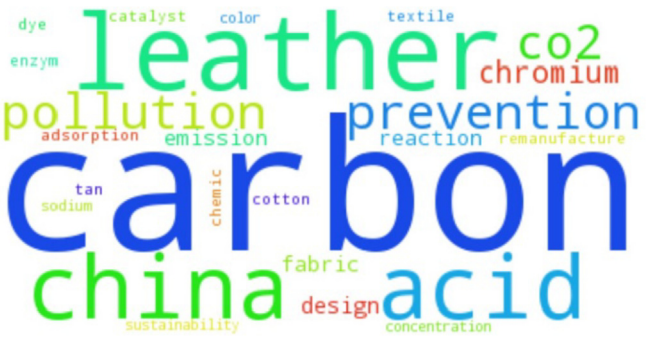

Fig. 8. Clothing Production as research topic was discussed particularly in the early years of the journal. The figure illustrates the most characteristic terms for this topic.

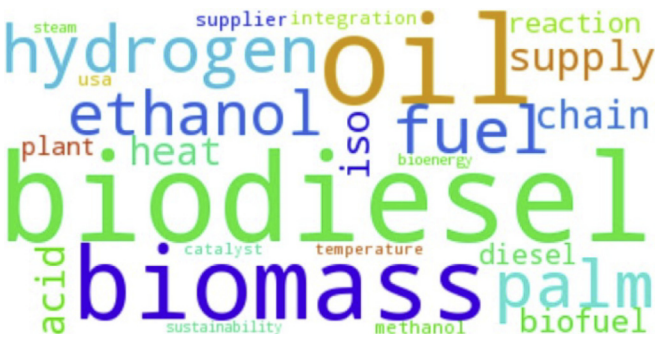

Fig. 9. Bioenergy like Clothing Production is a topic that is less characteristic for discussions in JCLP after 2011, but was of bigger meaning for the journal before. The figure illustrates the most characteristic terms for this topic.

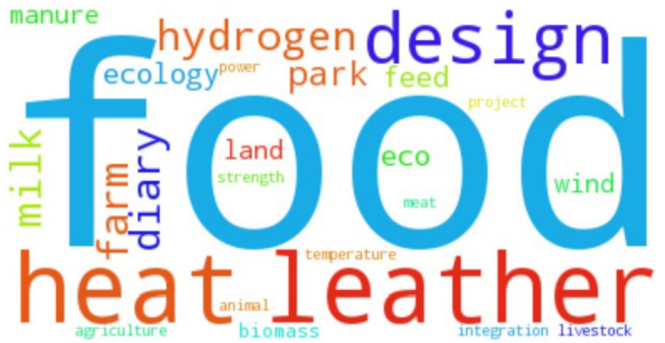

Fig. 10. Food Production is a topic that is not discussed constantly but repeatedly. In 2004, the journal even organized a special issue on this topic. The figure illustrates the most characteristic terms for this topic.

3.1.3. Land use

Land use is a topic that has unclear boundaries. Its composition suggests that tourism (e.g., *tourism*, *travel*, *hotel*) and pulp industry (e.g., *pulp*, *wood*, *tree*) have a joint context. Actually, these industries compete for the same resource of land use. Moreover, subjects related to energy are discussed in this context. Indications for this are the terms *wind*, *emergy* and *power*. Interestingly, the term *forest* is not among the words in the word-cloud (Fig. 11). However, some measures for a sustainable production or usage are raised: *pollution-prevention*, *remanufacture* and *degrowth*. The co-occurrence of the previously mentioned subjects *tourism*, *pulp* and *emergy* led us to conclude that land use was the proper umbrella term for the supporting topic. The results of our analysis indicated that this topic was only of greater importance in 2009 and particularly so in 2005. Indeed, Volume 13, Issue 2 in 2005 was a special issue on sustainable tourism.

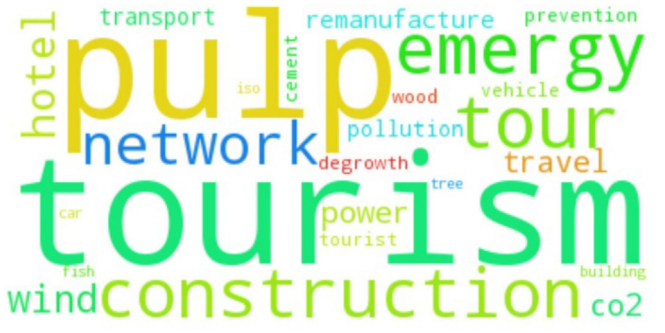

Fig. 11. Land Use as research topic is not a constant factor in the JCLP, according to pour analysis. Nevertheless, it is made a subject of discussion every now and then. The figure illustrates the most characteristic terms for this topic.

Fig. 12. Certification in general is an important issue for the transformation to a sustainable society and so it is also discussed in this very journal. The figure illustrates the most characteristic terms for this topic.

3.1.4. Certification

The last topic heavily depends on the term *iso*. Additionally, the terms *certification*, *audit* and *standard* appear in the corresponding word-cloud (Fig. 12). Fuel production (*biodiesel*, *fuel*, *energy*) and construction industry (*construction*, *building*) are apparently among the main areas that are discussed in the journal with respect to (iso-)certification. The topic certification covers the buzzwords *improvement*, *innovation* and *efficiency*, which indicate the impact of certification efforts on sustainable development. However, the temporal development of this topic indicates that it had already attained a greater importance in 2000 and 2006.

4. Results of analyzing topics in relation to the term *sustainability*

Once we had identified the most significant topics that appeared in the journal's publications and examined how the relevance of these topics changed over time, we analyzed the textual context that frames the term *sustainability* in the journal[7] as a distinct example of how the LSA method can be used.[8] To identify this context and generate a wordnet of relevant terms related to the term *sustainability*, we built on the term-loading matrix, which, as described in section 2.4, expresses the relationship between terms and topics. We considered the complete set of relationships in the corpus (not only the ten most dominant topics with the twenty-five most contributing terms, as in section 3). On the basis of these relationships, a network matrix was calculated. Fifty words that were most strongly related to *sustainability* were selected from this matrix along with the interrelationships among these words. Using this data, we generated the wordnet shown in Fig. 13. The node diameters in this figure correlate to the calculated significance (e.g., the accumulated term loadings of a term that co-occurred with *sustainability* in a topic) of the corresponding term in the journal. The edges between the nodes represent the co-occurrence of the terms. The widths of these edges correlate to the frequency of the co-occurrence. To establish these widths, the term loadings (see section 2.4) of the two involved terms for each topic were multiplied, and the results were summed up. This procedure yielded comparable numbers that could be interpreted as relative

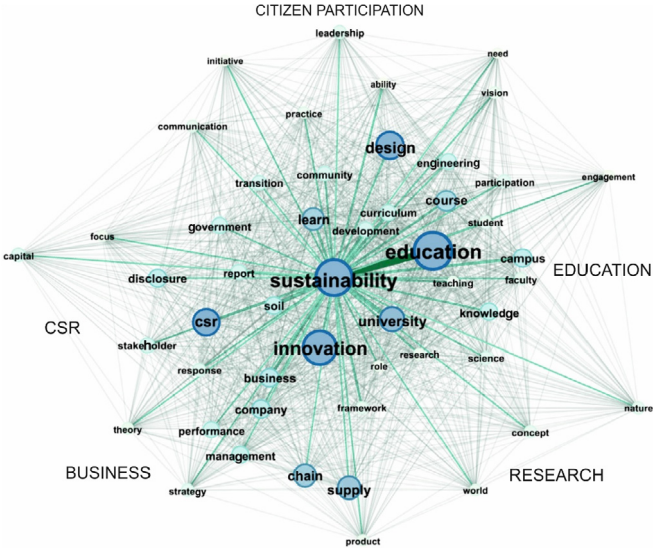

Fig. 13. Wordnet for the contextual relationships between the term sustainability and other terms in JCLP publications.

representations of the frequencies of the co-occurrences of terms.

Fig. 13 illustrates that the term *sustainability* had a strong contextual relationship to the term *education* in the publications appearing in the JCLP from 1995 to 2015. This relationship is not only indicated by the co-occurrences of the term, but also emphasized by other terms in the wordnet that are grouped in context (e.g., *learn*, *course*, *university*). Moreover, six of the ten terms that most frequently co-occur with *sustainability* refer to education. The other four terms of these ten are *soil*, *development*, *report* and *business* in descending order. The frequent appearance of the term *soil* suggests that *soil sustainability* was another characteristic topic in the JCLP during this period. Furthermore, sustainable development and diverse forms of sustainability in business determined the content of the journal, as did the terms *innovation*, *design*, *csr*, *supply* and *chain*. The latter were, however, less strongly associated with sustainability.

In summary, Fig. 13 reveals the co-occurrence of the term *sustainability* with about five different subject areas: *education*, *research*, *business*, *corporate social responsibility* and *citizen participation*. The term *research*, which partly overlaps with *education*, is mainly outlined by the terms *university* and *knowledge*, but has a strong connection to *innovation* as well. *Innovation*, moreover, is among the terms that define the subject area that we labeled as *business*. In addition to *innovation*, *supply chain* was also identified as an important field for discussions related to sustainability in the journal. The topic *business* is associated with both *research* and *csr*. *CSR* is closely related to the terms *stakeholder* and *soil*. As a fifth topic, we identified *citizen participation*, which is the largest but most indistinct one. Terms such as *initiative*, *community* and *participation* led us to label this topic in this way.

5. Results of identifying the topical orientation of individual papers

In the third part of the investigation, we applied LSA to ten randomly chosen (but recent[9]) individual JCLP publications in order to identify their topical orientation with respect to the general topical foci of the journal. The results are shown as a relative

[7] We chose not to focus on the possibly more obvious expression *cleaner production* because it was not possible to reliably separate the occurrences of this expression in relevant texts and in the journal's name, which appeared in the framework information of the publications. Conversion of the PDF to plain text maintained the journal name at the top of each page. Other text-conversion tools may provide better options.

[8] This algorithm, once developed, can be applied to every other term in the corpus.

[9] In order to cover the journals' topical development, we focused on papers published in 2015.

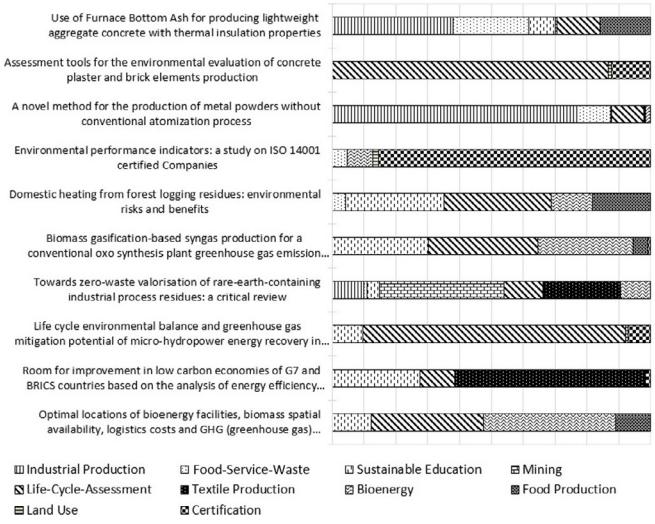

Fig. 14. Results of topical orientation analysis of ten randomly chosen (recent) JCLP publications with respect to the general topical foci of the journal (titles of publications as subscripts, main journal topics below).

composition of topics that are likely to appear as the subject of a publication. Although the publications may comprise more topics, only the topics discussed in section 3 were considered during this assessment.

The LSA was applied to the full text of papers. Nearly all analyzed texts showed a topical orientation that matched the content of information as it appeared in the titles of the papers (Fig. 14). Some results, such as the third, seventh and ninth publication shown, distinctly matched with one primary journal topic. However, this match was less distinct if more topics than the journal's ten most characteristic topics were considered. Other results were broader in range, indicating that the title showed a general affinity to the journal's thematic orientation but was less distinctly specialized with respect to one of its topics, which was maybe due to the multidisciplinary nature of sustainability science. What may seem quite obvious in this case, the match between title and content, may not be true for all publications.

The analysis in this part suggests that a comparison of the JCLP topics identified with the main subject areas as reflected by the topics represented by editorial board members should be made. Several corresponding elements were identified: the area of the special editor for “Industrial Applications and Sustainable Engineering”[10], for instance, seems to match well with the topic *industrial production*. Furthermore, the subject areas “Education and Organizational Change Management for Sustainability” and “Life Cycle and Sustainability Assessment Tools” seem to correspond well with the topics *sustainability education* and *life-cycle-assessment*. However, several subject areas which are considered in the editorial board, such as *eco-design* and *product service system*, could not immediately be detected in our analysis. Only through variation of the input parameters in our analysis software these topics eventually could be mapped as well, which clearly demonstrates the need for fine-tuning and further testing and developing this method.

6. Discussion

Using the method of Latent Semantic Analysis, we examined the textual corpus of publications appearing in the Journal of Cleaner Production from 1995 to 2015 in three ways. We mapped the main factors with respect to topical structure as well as the temporal development of the identified topics over this period. The results of our analysis suggest that the journal had a rather narrow focus during the early years and concentrated on relatively few topics such as *industrial production*, *clothing production* and *food production*. The journal's content clearly focused on (cleaner) production. In later years, however, and particularly after 2007, the thematic focus broadened and topics that had appeared frequently during earlier years began to play a less important role. In turn, topics such as *bioenergy* and *sustainability education* began to appear more frequently. In recent years, particularly subjects related to *life-cycle assessment*, *food-service-waste* and *sustainability education* have taken a dominant role in the journal's contributions. In general, our finding of a very much diversifying topical scope of the journal over time is in a sense in compliance with the result for the whole scientific field of sustainability science, as brought forth by Kajikawa et al. (2014). Both, in the journal and in the whole field, particular the topic *sustainability education* has gained importance (see alo Barth and Michelsen (2013). In contrast to this so sayed compliance, *life-cycle assessment* (LCA) seems to present some kind of a distinguishing feature of the Journal of Cleaner Production. According to our analysis, LCA has become a dominant subject of discussion, particularly in recent years, although the studies we compared our results with (primarily Kajikawa (2008), Bettencourt and Kaur (2011), Miller (2013) and Kajikawa et al. (2014)) didn't identify or discuss this topic. Only Bettencourt and Kaur (2011) mentioned in their supporting information with “sustainability assessment” an umbrella term for methods like LCA (Ness et al., 2007), as frequent bigram in titles of sustainability-related publications. The nonappearance of these terms in the mentioned studies on the one hand may be due the fragmentation of sustainability science, on the other hand it could also be induced by a lower resolution of the applied research methods.

With respect to the second part of our analysis, the identification of the context in which the term *sustainability* is embedded in the journal, we determined, which terms co-occurred most frequently with the term *sustainability*. These terms were classified in five subject areas, of which the most important was, interestingly, *education*. The related terms were also analyzed with regard to their interrelationships, which were revealed by examining certain terminological clusters according to the strength of their relationships. In this way, the clusters *education*, *research*, *business*, *corporate social responsibility* and *citizen participation* were identified. More indirectly, the identification of keywords such as *pollution prevention*, *remanufacture*, *rehabilitation*, *eco/ecology*, *degrowth*, *efficiency* and *LCA* indicated their relationships with the term *sustainability*. However, our results also revealed that topical boundaries are in part vague and transitions are fluent in this respect. This is no wonder, as sustainability has become a highly interdisciplinary scientific field, asking for interdisciplinary approaches, as Schoolman et al. (2012) argued. In this light our results seem to indicate that the Journal of Cleaner Production has well adapted to this demand in the course of its evolution. Moreover, our results also reveal a very strong connection between the terms *sustainability* and *education* in the textual corpus of the journal. Indeed, education and its advancement early has been considered as an essential measure for sustainability transformation (Huckle, 1991). However, education as well has been discussed and scrutinized with increasing frequency to this day.

During the third part of our investigation, our analysis focused on identifying the topical orientation of individual JCLP publications that were related to the main JCLP topics. Moreover, this part also highlights the importance of sustainability education and life-

[10] http://www.journals.elsevier.com/journal-of-cleaner-production/editorial-board, accessed on 10-20-2016.

cycle assessment, which were already identified as dominant subjects of discussion in the journal.

7. Conclusions

With a computerized knowledge discovery method we started out, on the one hand, to spot both currency and temporal development of foci in sustainability science, on the basis of the textual corpus of the Journal of Cleaner Production. On the other hand, we had also the aim to test the limits and demonstrate the capabilities of these automatized methods.

From the viewpoint of methodology, the results of our analysis show that a knowledge discovery method like the here applied LSA can be used to screen and identify structural elements within large textual collections. The method proved sufficient for answering our research questions concerning the identification of dominant topics in the Journal of Cleaner Production. Furthermore, we could show that compared to other studies with similar research questions but different research methods, the resolution of the results yielded by our analysis is relatively high. Additionally, the method deployed allows for a degree of topic classification that may provide a base for further developing tools for automatically identifying a paper's topical orientation and suggesting or assigning it to interested or, like in the case of reviewers, specialized readers.

In terms of content, our results picture a diversified but entangled network of themes that evolved in over two decades in the Journal of Cleaner Production. This enmeshment of research foci and topics reflects the complex framework that is marked out by sustainability science today. However, out of this network sustainability education has to be stressed as noteworthy topic. Apparently, education that intrinsically raises awareness for sustainability issues is of particular importance for sustainable development, at least as perceived from the scientific community. Moreover, this is very plausible for cleaner production as well, as it often requires highly specialized production processes that, for instance, make strong demands on the qualifications of operators and therefore require specialized education or training. Apart from the education topic, which has been made a subject of discussion since the advent of sustainability science, in particular life-cycle analysis is a subject that became increasingly important in recent years in the Journal of Cleaner Production. This may be seen as good indicator for the sustainable transformation to take place in the industries.

While the Journal of Cleaner Production has a broad focus that allows an interdisciplinary discourse, it still covers only a fraction of sustainability science. In further research, we want to deploy similar analyses also on eco-system and management-focused journals to contrast the results with those of the analysis in this paper. Moreover, we want to upscale our analyses and take into account a bigger corpus of publications. By doing so, it may become possible to investigate the effects of political objectives like, e. g., the UN Sustainable Development Goals on research efforts.

References

Barth, M., Michelsen, G., 2013. Learning for change: an educational contribution to sustainability science. Sustain Science 8 (1), 103–119.

Bettencourt, L.M., Kaur, J., 2011. Evolution and structure of sustainability science. Proc. Natl. Acad. Sci. Unit. States Am. 108 (49), 19540–19545.

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R., 1990. Indexing by latent semantic analysis. J. Am. Soc. Inf. Sci. 41, 391–407. https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2–9.

Dumais, S., 1991. Improving the retrieval of information from external sources. Behav. Res. Methods, Instruments. Comput. Times 23, 229–236. https://doi.org/10.3758/BF03203370.

Evangelopoulos, N., Zhang, X., Prybutok, V.R., 2012. Latent Semantic Analysis: five methodological recommendations. Eur. J. Inf. Syst. 21, 70–86. https://doi.org/10.1057/ejis.2010.61.

Golub, G., Kahan, W., 1965. Calculating the singular values and pseudo-inverse of a matrix. J. Soc. Ind. Appl. Math. B Numer. Anal. 2, 205–224. https://doi.org/10.1137/0702016.

Harris, Z.S., 1954. Distributional structure. Word 10, 146–162. https://doi.org/10.1007/978-94-009-8467-7_1.

Huckle, J., 1991. Education for sustainability: assessing pathways to the future. Aust. J. Environ. Educ. 7, 43–62.

Ibrahim, O.A.S., Landa-Silva, D., 2015. Term frequency with average term occurrences for textual information retrieval. Soft. Comput. 1–17. https://doi.org/10.1007/s00500-015-1935-7.

Kajikawa, Y., 2008. Research core and framework of sustainability science. Sustain. Sci. 3 (2), 215–239.

Kajikawa, Y., Tacoa, F., Yamaguchi, K., 2014. Sustainability science: the changing landscape of sustainability research. In: Sustainability Science, vol. 9, p. 431. https://doi.org/10.1007/s11625-014-0244-x.

Kates, R.W., 2012. From the Unity of Nature to Sustainability Science: Ideas and Practice. In Sustainability Science, pp. 3–19.

Manning, C.D., 2011. Part-of-speech tagging from 97% to 100%: is it time for some linguistics? In: Gelbukh, A.F. (Ed.), Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer Berlin Heidelberg, pp. 171–189. https://doi.org/10.1007/978-3-642-19400-9_14.

Merton, R.K., 1968. The Matthew effect in science: the reward and communication systems of science are considered. Science 159 (3810), 56–63.

Miller, T.R., 2013. Constructing sustainability science: emerging perspectives and research trajectories. Sustain.Sci. 8 (2), 279–293.

Müller, K.H., Riegler, A., 2014. Second-order science: a vast and largely unexplored science frontier. Construct. Found. 10 (1), 7–15. http://constructivist.info/10/1/007.

Ness, B., Urbel-Piirsalu, E., Anderberg, S., Olsson, L., 2007. Categorising tools for sustainability assessment. Ecol. Econ. 60 (3), 498–508.

Nevin, B.E., 2002. The legacy of Zellig harris: language and information into the 21st century, current issues in linguistic theory. J. Benjamins, amsterdam; Salton, G., Buckley, C., 1988. Term-weighting approaches in automatic text retrieval. Inf. Process. Manag. 24, 513–523. https://doi.org/10.1016/0306-4573(88)90021-0.

Scienceman, D.M., 1987. Energy and Emergy. Environ. Econ. - Anal. A Major Interface, pp. 257–276.

Schoolman, E.D., Guest, J.S., Bush, K.F., Bell, A.R., 2012. How interdisciplinary is sustainability research? Analyzing the structure of an emerging scientific field. Sustain. Sci. 7 (1), 67–80.

Sidorova, A., Evangelopoulos, N., Valacich, J.S., Ramakrishnan, T., 2008. The intellectual core of the is discipline. MIS Q. 32, 467–482.

Sidorova, A., Evangelopoulos, N., Ramakrishnan, T., 2007. Diversity in is research: an exploratory study using latent semantics. Int. Conf. Inf. Syst.

Valle-Lisboa, J.C., Mizraji, E., 2007. The uncovering of hidden structures by Latent Semantic Analysis. Inf. Sci. 177, 4122–4147. https://doi.org/10.1016/j.ins.2007.04.007.

Witten, I.H., Frank, E., Hall, M.A., Pal, C.J., 2016. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann.