Graph-based research field analysis by the use of natural language processing: An overview of German energy research

Jan Richarz[1], Stephan Wegewitz[1], Sarah Henn[1], Dirk Müller[1]
[1]*Institute for Energy Efficient Buildings and Indoor Climate (EBC), E.ON Energy Research Center, RWTH Aachen University*

This is the preprint of the corresponding paper:

Richarz, J.; Wegewitz, S.; Henn, S.; Müller, D., 2022. Graph-based research field analysis by the use of natural language processing: An overview of German energy research. Technological Forecasting and Social Change 186, Part B, 122139. doi.org/10.1016/j.techfore.2022.122139

RWTH Aachen University
E.ON Energy Research Center
Institute for Energy Efficient Buildings and Indoor Climate (EBC) Mathieustr.
10, 52074 Aachen
T +49 241 80-49760, F +49 241 80-49769
ebc-office@eonerc.rwth-aachen.de, www.eonerc.rwth-aachen.de/ebc

Graph-based research field analysis by the use of natural language processing: An overview of German energy research

Jan Richarz*, Stephan Wegewitz, Sarah Henn, Dirk Müller

RWTH Aachen University, E.ON Energy Research Center, Institute for Energy Efficient Buildings and Indoor Climate, Mathieustraße 10, 52074, Aachen, Germany

Abstract

To stop climate change caused by the anthropogenic greenhouse effect, the amount of research project funding to develop and demonstrate solutions for carbon emission reduction has increased. Focused topics and procedures of research projects are always presented in text-based descriptions. Natural language processing offers the possibility to process and analyze data like this. In this work, we used natural language processing to extract keywords from descriptions of energy research projects with the algorithms TextRank and TF-IDF which has not been done for this kind of data basis before. A survey-based validation showed TF-IDF to be better suited for our data basis. Extracted keywords were used to conduct a keyword network analysis and calculate static and dynamic indices of the words concerning their recent importance and development over the past two decades. Further insights are shown by allocating the research projects' amounts of funding to the extracted keywords. We found energy research to be more focused on individual components in the past. The last years were characterized by projects on energy systems and the interaction of renewable energy technologies and their integration into existing infrastructure. Finally, the results were compared with governmental research programs and we could analyze a comprehensive agreement.

*Corresponding author
Email address: jricharz@eonerc.rwth-aachen.de (Jan Richarz)

Keywords: Text mining, Keyword extraction, Network analysis, Research trends, Research funding

---

1. Introduction

Large-scale energy research was historically established in the OECD states to decrease oil dependency after the first oil crisis. Nowadays, efforts in energy research focus on saving carbon emissions to contribute to the containment of climate change. As one of the largest conveyor institutions in Germany, the Federal Ministry of Economic Affairs and Climate Action (BMWK) approved more than 1.600 energy research projects in 2019 [1]. With 1.15 billion € (about 1.36 billion US-$), BMWK increased its overall funding in this field by 25% compared to 2014 [1]. Research fields on the energy transition are the consumption sectors (buildings, industry, traffic), energy generation with renewable energy sources, and cross-system research and nuclear safety. Recent key drivers are energy efficiency and the usage of hydrogen [1].

As the volume of funding and research projects has increased and the covered topics have become more and more heterogeneous, an overview of the research field is necessary to identify research focal points and gaps, create transparency about public funding, and provide information for researchers. Furthermore, as Ernst states [2], the analysis of technological change can support political decision-makers to set the focus of R&D (research and development) resources. These analyses realize a review and, if necessary, adjustment of the path towards achieving the national energy research goals.

Text-mining is frequently used in many fields and is a promising method to analyze text-based data automatically. For national energy research, systematically analyzing the status quo and developing topics becomes possible. In this context, keyword extraction and its graph-representation offer a way of semantically transferring the content of published texts into a metric system and investigating trends and focal points. With this, it becomes possible to analyze the current state of research, the chronological development of specific topics and technologies, and their relevance. So-called network analyses offer an intuitive approach to visualize data and the relationships between different topics of the data [3].

For this purpose, we extracted and adapted proven state-of-the-art text-mining methods and applied our developed method to the field of energy research. As a result, one methodical challenge is interpreting the German language instead of mainly analyzing English texts. Therefore, this paper provides an added value in terms of how to deal with non-English texts. Our proposed method can be transferred into other languages.
We created a novel automated approach to analyze research fields based on text-based research project descriptions thematically. Furthermore, we were able to analyze connections between research topics without an external given indicator for relationships (e.g., citations across papers), as mostly being used so far.
Summarized, the central topics this work contributes to are:

- Text-mining approach for a non-English data basis with contemporary tools and increased requirements, due to technical language
- Combination of established keyword extraction methods and graph-based analyses from technology forecasting
- Validation and comparison of the keyword extraction methods TF-IDF and TextRank
- Algorithm-based overview of a research field by the use of short project descriptions
- Analysis of the research field focus topics over time and comparison with the funding institution's issued goals

The paper is structured into a literature review of works and methods for technology forecasting based on text-mining and natural language processing. Furthermore, an explanation of the gathered primary data, a presentation of the developed method for interpreting the data, and comparison and validation of two implemented keyword extraction algorithms are given. Then, the application of the proposed method on German energy research with further specialization in topics and technologies for energy-efficient buildings and neighborhoods

is shown. Finally, the relevance of different research topics today and over the last 20 years is shown based on different static and dynamic characteristic values.

2. Related work

Insights from national technological analyses can, among others, be used by policymakers to evaluate the current state of Research and Development (R&D) and provide a foundation for decisions on future steps [4]. Previous works on forecasting of technological progress were analyzed by Luukkonen-Gronow [5] who found that mainly different kinds of qualitative methods like questionnaires (e.g., Delphi method), interviews, or peer review processes were used. Some heterogeneous studies also used quantitative methods and developed measures to evaluate economic benefit [5]. Porter et al. [6] introduced the so-called Technology Opportunity Analysis, which utilizes bibliometrics, like content analysis, for generating knowledge on possible future technologies.
A widely used quantitative method to analyze content is Text-mining. It is defined as

> *the automated process by which large volumes of unstructured, natural-language text are analyzed in order to pinpoint and extract user-specified information* [7]

With natural-language processing (NLP) being a component of text-mining, summaries and key messages can automatically be generated out of a text. These approaches expanded the research field with insights based on data like patents or papers. Furthermore, the relationship between different text sources through citations, the use of similar extracted topics, or the origin of texts can be analyzed [4].
Especially patents have proven to be a valuable source of knowledge for research results and technical progress [8]. Recent works to be named are, for example, [2, 9, 10] who emphasize the benefits of patents being open access and having the possibility to retrieve large amounts of data through emerging databases systematically. The use of scientific literature in text-mining techniques is also common. In the field of biology and medicine, keyword extraction [11, 12, 13, 14] and topic-clustering methods [15, 16, 17] with various other applications are established. According to Jelodar et al. [16] also fields like software engineering,

geography, and political science use clustering methods.

With the so-called tech-mining, a combination of text-mining and technology management was developed, which proliferated, especially in the analysis of patents and technologies [18]. In addition to the already mentioned bibliometrics, data-mining, network analysis, cluster analysis, technology road maps [19] or hot-spot detection [20] were introduced to analyse patents [21].

Regarding the automatic determination of the key messages of texts, the general approach is to extract and rank keywords with different algorithms. Here, Term Frequency Inverse Document Frequency (TF-IDF) appears to be one of the most used numerical statistics [22], which ranks words according to their frequency of use within a given set of documents. The performance of TF-IDF is influenced by the number and size of the documents in a given domain [23]. It has been shown to be a robust method in many works like [24, 25, 26, 27, 28, 13, 29]. However, by accessing the complete domain, the calculation of the TF-IDF score can need high computational efforts [30].

TextRank on the other hand is a graph-based extraction method, which was developed by [31] and for example used by [32, 33, 34, 35]. It utilizes co-occurrences in a slide of a window within a text and can be used independent of text length and the number of documents [23]. Since the algorithm does not use the whole domain for the computation of the score, it is less intensive in the use of computational resources [30].

In this work, we refer to network analysis as a combination of extracting keywords through natural language processing and building matrices of co-occurring words. Through these matrices, networks of intertwined keywords can be distilled and analyzed. One of the first applications of keyword search was made by Yoon et al. [3] for patent data. In the following years, the method further matured with works from [36, 37, 38] and others while utilizing the increasing performance of computing. It is capable of finding new topics without further algorithm training, but with the disadvantage of finding idiosyncratic or malformed keywords [39].

Yoon et al. [36] showed a combination of extracting keywords with co-word

analysis and generated networks that codify the relations between occurrences of keywords and identifying specific trends from these networks. As defining keywords for emerging technologies is challenging even for experts, they additionally used property functions which are a specific characteristic of a technology [36]. Park et al.[40] describe a patent intelligence system, which combines a Subject–Action–Object (SAO) system with network analysis in order to spot new technologies or potential infringements.

SAO-based extraction has been used extensively because of its additional information through semantic analysis [41, 42, 40], which can be extended with an approach on building morphologies [43, 44]. Lee et al. [45] propose a chunk based text analysis, where noun phrases are broken down to words and their syntactic function, in order to calculate an influence score. A data-driven approach can also expand a morphology analysis, which for example was shown by building a database of the Wikipedia-corpus [46]. Further developments in NLP-enabled methods are the Latent Dirichlet Allocation (LDA) for clustering words within a text [47, 48, 49], the use of ontologies [25, 50, 51] or the use of word embeddings with e.g. Word2vec [12].

In addition to the content of papers or patents themselves, collaborations are a target of recent research. Thus, Yang et al. [52] examined interdisciplinary research relationships in science and technology of a research funding program with the help of network analysis. Campbell et al. [53] and Barbasi et al.[54] studied collaborations between researchers and scientific studies using mapping within a network analysis.

2.1. Contribution

To the best of our knowledge, there has not been a work that analyses the current state of national energy research using proven methods from the field of text-mining and network analysis. Furthermore, in contrast to the reviewed literature, we use research project descriptions as a data basis. This data basis is very homogeneous in text length, degree of detail, and format. By validat-

ing extracted keywords with a questionnaire among researchers of the studied projects, we could choose the keyword-extraction method that best fits our application.

3. Proposed Method

This paper aims to find central aspects of a research field by analyzing a text corpus. This section describes the proposed method and used characteristic values for automatic analysis. Keywords extracted from each document of the data basis are treated as representatives for the central aspects of a document. Afterwards, a graph-based analysis is conducted with these keywords to find connections between different research topics and calculate characteristic values that evaluate the relevance of a keyword.

3.1. Data basis

The data we use within our study is freely available via the *EnArgus* database [55]. With *EnArgus*, the BMWK provides an online portal that holds information on ongoing and completed research projects in the field of *energy research* [55]. Besides background information, like research project start and end dates, research institutions and funding amount, projects descriptions of the central research aspects and methods of the projects are provided for many projects. Moreover, the projects are tagged according to a thematic categorization. The database contains more than 26.000 entries starting from 1968. The research institutions themselves have originally provided the project descriptions within the database. However, the thematic categorization of projects was carried out following the Federal Government's R&D planning system.

Figure 1 shows the number of projects and existing descriptions. It can be observed that the number of starting projects per year increased over the years and peaked in 2017. Up to 2002, project descriptions existed for less than 10% of the projects. Since then, a rising trend can be identified. It should be noted that the year 2020 is not fully represented because not all data was already available at the time of the preparation of this analysis.

As described above, there is a tagging of projects according to the Federal Government's R&D planning system. According to this, there are 156 different

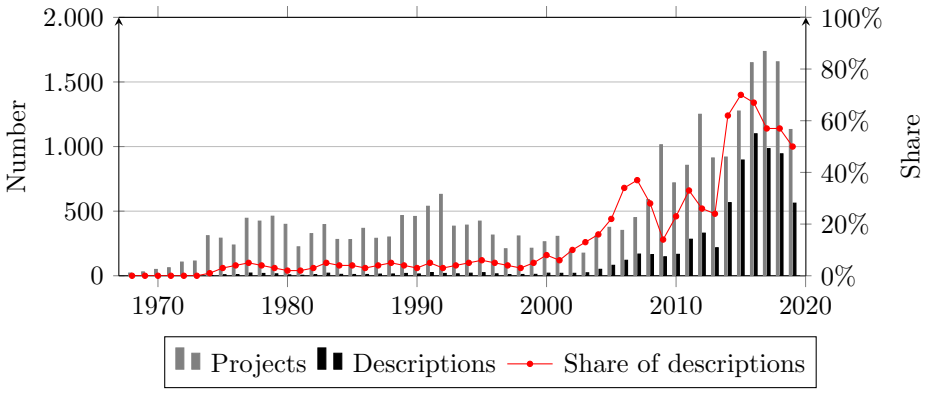

Figure 1: Number of projects and their text-based descriptions since 1968

categories. Each project belongs to one category. 35 categories contain research projects of our special field of interest within this paper (subsection 5.2): *Energy in buildings and neighborhoods*. In the following we refer to these projects with the abbreviation *EWB* (from German *EnergieWendeBauen (EWB)*).
An analysis of the projects, which started within the last five years, by categories is given in Figure 2. In the diagram, the 20 most prevalent categories are shown. We find the two topics *energy optimized buildings* and *local supply systems* of EWB to be under the ten most common categories. Another one, *solar optimized building*, is in the thirteenth rank. The top three categories regarding all projects within the last five years were *alternative engine technologies*, *fundamental research in energy* and *energy-saving industrial processes*.

The development of yearly ongoing funded projects within the categories *energy optimized buildings*, *local supply systems* and *solar optimized building* is shown in Figure 3. The diagram displays the share of projects compared to the total number of ongoing projects per year in percent. From 1994 there is a significant increase of relative project share within the categories *energy optimized buildings* and *solar optimized building*. Research projects within the category of *solar optimized building* reached a peak of 7% of all research projects within energy research projects funded by the BMWK. Research interest in

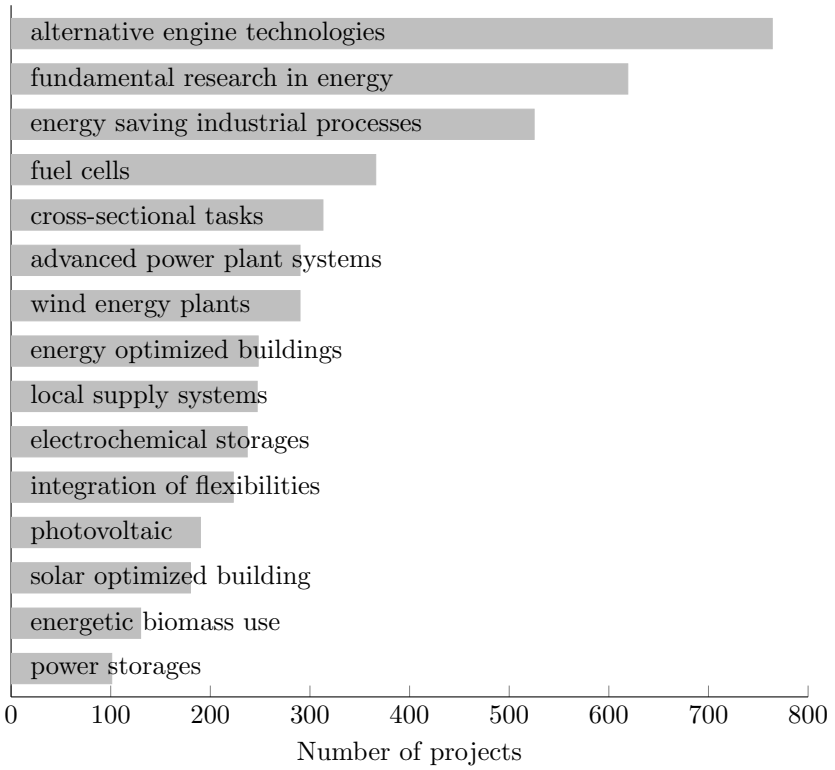

Figure 2: Number of funded projects according to the Federal Government's R&D planning system from 2015 to 2020

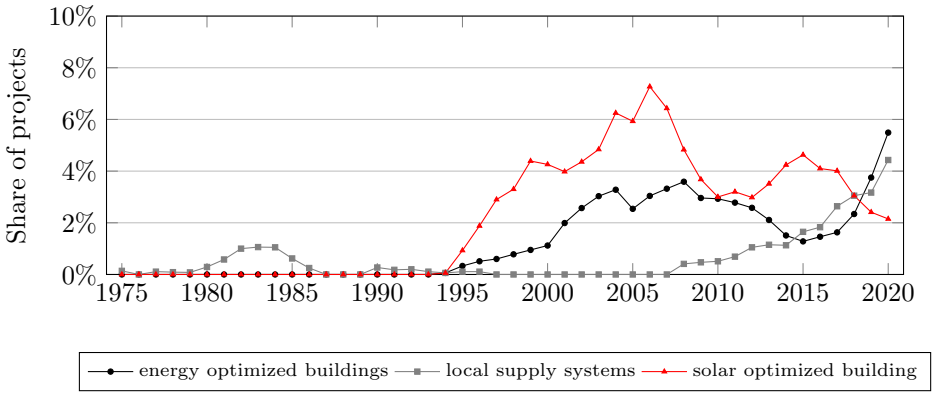

Figure 3: Share of ongoing founded projects per year in categories concerning energy efficient buildings and neighborhoods compared with total number of ongoing projects in energy research

energy optimized buildings showed an upward trend between 2008 and 2015 and had been rising since then. A few projects within the category *local supply systems* took place in the 1980s and 1990s. Since 2007 the interest in research in this category has been continuously increasing.

For our study, we extracted information and project descriptions from 7259 projects for which a description was available in the *EnArgus* database. We use this project data to conduct a keyword extraction and a network analysis, for which the method is described in the following.

3.2. Keyword extraction

For the extraction of keywords, we access the two different extraction methods *Term Frequency Inverse Document Frequency* (TF-IDF) and *TextRank*. Figure 4 shows an overview of our proposed method and workflow from generating the data basis (subsection 3.1) to the pre-processing and extraction of keywords (subsection 3.2) up to their depiction in directed networks (subsection 3.3) and their evaluation (subsection 3.4). For the later analysis, we chose between TextRank and TF-IDF by comparing and validating the results by a questionnaire among the researchers of the studied projects.

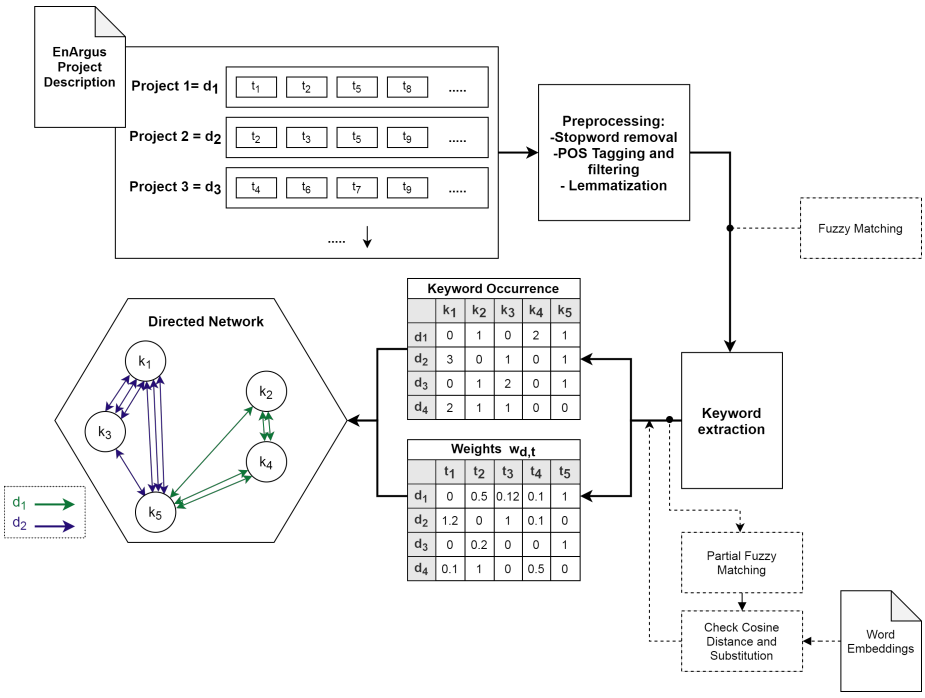

Figure 4: Workflow for creating a directed graph from given corpus

The project descriptions need to be pre-processed, and keywords are extracted and post-processed to gather information from the given text corpus. After that, keyword networks for each year of the data basis can be constructed based on the dates for the beginning and end of the projects.

Before the actual keyword extraction, the given data has to be processed. Adapting the definition by Lee et al. [56], the gathered corpus d consists of text documents d_i for each project, with $d = \{d_0, d_1, d_2, d_3, ..., d_i\}$. Each document d_i contains a number of terms tm, from which a smaller sample of important keywords has to be extracted.

For the natural language processing the two libraries *spaCy* [57] and *nltk* [58] were utilized. With both libraries, we obtained the terms tm of the documents and pre-processed them with tokenization to remove stop words, Part-of-Speech (POS) to identify word categories like nouns or verbs, and lemmatization with *GermaLemma* [59] to group words according to their lemma. Hulth [60] showed that the incorporation of a stop word filter and POS-tagging can highly improve the performance of a keyword extraction algorithm.

Afterwards, we collected all occurrences of a term and its variants in the corpus and compared them to each other with fuzzy string matching. Here, the target is to minimize German variants of words, which the lemmatizer could not target, to reduce the number of individual keywords.

For this purpose, the *Levenshtein*-distance is utilized to calculate word similarity. Through comparison and substitution of similar words, keywords can be reduced. This is especially necessary for scientific terminology and technical expressions in German, like *Wärmepumpensystem (heat pump system)*, which is a combination of *Wärmepumpe (heat pump)* and *System (system)*.

To choose an order for fuzzy replacement, all terms were gathered, ordered by their frequency of use, and replaced by starting from the least frequent one. By going through the frequency order, we ensured that frequent word forms will be the last to be substituted and have a higher possibility to appear in the final corpus.

TextRank (subsubsection 3.2.1) and Term Frequency Inverse Document Frequency (TF-IDF, subsubsection 3.2.2) were chosen as extraction methods for this work. Both have been developed, used and validated in the past and are considered established methods (e.g. [12, 61, 62, 63, 64]). In section 4 we present a survey-based validation of these two methods, concerning our data basis.

Both algorithms are used to compute the weight $w_{k,d}$ of a term $tm_{i,j}$ inside a given document d_i. Analogously to [56], the 15 highest ranked terms are extracted and represent the keywords of a document whereby only nouns are evaluated as candidates. Later on, the weights are used in the network creation process.

After extracting the keywords, fuzzy string matching is applied again because we observed several unintended word variations. Zhang et al. [65] describe the process of multiple rounds of fuzzy matching as *term-clumping*, where words are clustered by their similarity, thesauri are built, and reviewed. We identified that German scientific language is a good use case for this method.

3.2.1. TextRank

As a variation of the algorithm *PageRank* [66], *TextRank* is a ranking model, which extracts importance of the nodes within a graph representing the terms of a document [31, 30].

After tokenization and filtering through POS, each token passes a syntactic filter and is then connected by co-occurrence inside a graph. The algorithm calculates a score for each node, while it *surfs* [31] or *jumps* from node to node inside the graph [30, 31]. For a graph $G = (V, E)$ with nodes V and edges E, with a given weight w, the score WS is computed as follows [31]:

$$WS(V_i) \;=\; (1-\delta) + \delta \cdot \sum_{V_j \in In(V_i)} \frac{w_{ij}}{\sum_{V_k \in Out(V_j)} \cdot w_{jk}} \cdot S(V_j) \qquad (1)$$

δ is a damping factor to describe the probability of jumping between nodes. $In(V_i)$ are nodes, which are pointing to its predecessor node, and $Out(V_j)$ that point to its successor node.

The ranking model itself does not require a deep linguistic knowledge of the given document collection and, in terms of accuracy, can compete with other current algorithms [31]. Therefore it suits the task of evaluating text fragments, which the algorithm never encountered before [31]. This is one reason why we

chose this algorithm. It thereby allows enlarging the corpus at any time, of which TF-IDF is not capable.

3.2.2. Term Frequency Inverse Document Frequency (TF-IDF)

TF-IDF is a method which calculates the term frequency and in-diverse term frequency of a term in a document and weights terms statistically [30, 56]. It calculates a weight w through the term frequency $tf(t,d)$ of a term tm, its frequency f in a document d, the logarithmic gradient of the total number D of documents in the document collection, and the total occurrences $f_{t,D}$ of term t in all documents of the collection [26, 67]:

$$tf(tm,d) = \frac{f_{tm,d}}{|d|} \tag{2}$$

$$idf(tm,D) = log(\frac{|D|}{f_{tm,D}}) \tag{3}$$

$$w_{tm,d} = tf(tm,d) \cdot idf(tm,D) \tag{4}$$

This calculation makes it possible to collect essential terms from documents in a given document collection. TF-IDF proved to be a robust weighting scheme [67] in many different applications (e.g. [68, 28, 27]). Downfalls of TF-IDF are the ignorance of word relations [26] and that the calculated keywords are unique to the given document collection. If changes in the collection occur, there will also be changes in the calculated keywords.

3.3. Network analysis

We used network analysis to analyze shared properties among the keywords and their relation. As our data basis contains one document for each energy research project, we will from now on refer to a document as one *project*. First, we computed a network where the keywords themselves are the nodes. Based on their co-occurrence inside a project, keywords are connected.
The weights of the keywords $w_{tm,d}$ computed by either TF-IDF or TextRank were used to weight the edge $e_{tm,d}$ which is created through co-occurrence of

given keywords within a keyword network. The algorithm with the best performance within the validation (section 4) was then used for the results within section 5.

In the first step, we calculated a network across the complete corpus. With knowledge of the start date and duration of the projects, we then created a network for each year. As long as a project runs, its related keywords are present in a yearly network.

Based on these networks and recommendations in literature ([69, 70, 36]), for analysis at a single point in time t, the following quantitative graph indices were used to calculate the relevance of a keyword inside a network.

3.4. Evaluation of keywords

The keyword's relevance indicates the yearly relevance of a topic presented by this keyword. On the one hand, we calculated the *weighted degree* which is the sum of links between nodes [70] and will be handled according to Equation 5. Another indicator for the relevance of a keyword is the *betweenness centrality*, $c_B(tm)$ (Equation 6) of a set of nodes of terms TM. It describes the number $\sigma(s, l|tm)$ of shortest paths between nodes s and l that go through the node of the term tm [71]. This number is compared to the whole number of shortest paths $\sigma(s, l)$ [71].

$$deg(tm) = \sum_{j=1}^{N} e_{tm,d} \cdot w_{tm,d} \tag{5}$$

$$c_B(tm) = \sum_{s,l \in TM} \frac{\sigma(s, l|tm)}{\sigma(s, l)} \tag{6}$$

We calculated a network for every year between 1974 and 2020. Hence, we can analyze a single point in time and observe the development of the indices between the years. As the development of the absolute relevance of keywords is strongly dependent on the number of projects, we normalized the yearly indices of any given network for year-to-year comparisons. For this normalization, the most minor absolute deviation, also called *L1-normalization*, was chosen.

Our indices are also based on recent literature to evaluate the temporal development of keywords, i.e. topic's relevance. Kim et al. [72] transferred *acceleration* and *skewness* out of the physical theory of motion and from the field of statistics, respectively into network analyses:

Analogously to the theory of motion, velocity v and acceleration a can be defined as the first and second deviation of the change of keyword's weighted degree over a time interval Δt (Equation 7, Equation 8). In our case, this time interval equals one year. According to Kim et al. [72], the velocity is used to describe the change of a keyword over time, with taking into account that every keyword has an individual starting point in its weighted degree. It, therefore, makes it more difficult to compare the indicator of normalized weighted degree on the same basis [72]. Through the second deviation, we can circumvent this problem and only compare the changes in the normalized weighted degree value of several keywords to another, which gives a more nuanced description of their changes over time. [72].

$$v(tm) = \frac{\Delta wdeg(tm)}{\Delta t} \tag{7}$$

$$a(tm) = \frac{\Delta v(tm)}{\Delta t} \tag{8}$$

For the calculation of the skewness, we created a distribution of the changes of the normalized weighted degree for each point in time, starting with t_0. Taking the changes instead of the absolute values into account was necessary because otherwise, the rising number of projects would have automatically directed the skewness in one direction. Therefore, for all time steps, i.e. years $\bar{i} = (t_0, t_1, ..)$, the difference to the initial time step is calculated using Equation 9.

$$F(t, i) = deg(\bar{t}m_t) - deg(tm_{t_0}) \tag{9}$$

Now, the skewness of the distribution is calculated based on Equation 10

with $x \in wdeg(tm_t)$, s as the standard deviation, and μ as mean value of x.

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \mu_x}{s}\right)^3. \tag{10}$$

It shows the distribution's asymmetry and thus whether a keyword is stable in its development or is subject to change over time.

4. Validation of keyword extraction

With TF-IDF and TextRank, we found two promising approaches in literature to extract keywords from texts. These keywords form the basis of our entire method, i.e., all network and keyword evaluations. As the performance of the two algorithms for our data basis (project descriptions) was unknown, a validation with an online questionnaire was conducted. We aimed to extract keywords out of research project descriptions. The researchers of these projects should evaluate how well extracted keywords fit the topics inside the projects. We sent the questionnaire to 1,000 research projects in energy research in buildings and neighborhoods that the BMWK funds. Both algorithms beforehand extracted ten keywords, and the project researchers rated each keyword concerning their project activities between grades 1 (*poor suitable*) and 5 (*very good suitable*). Figure 5 shows the amount of the given grades for the keywords of TF-IDF, TextRank, and for keywords that were extracted by both.

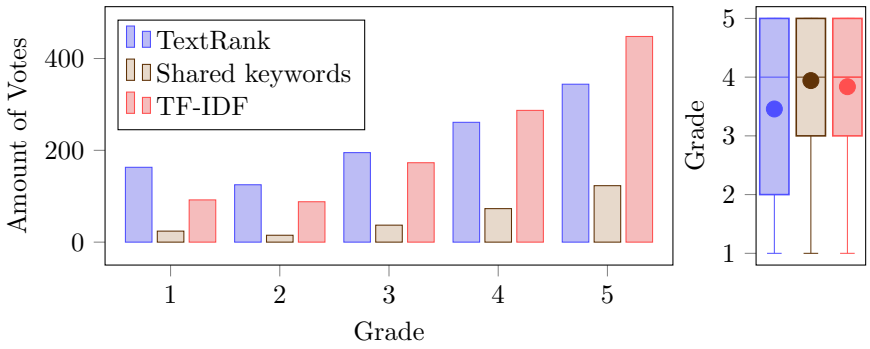

Figure 5: Amount (left) and distribution (right) of the given grades for the keywords extracted by both algorithms.

Based on the increasing votes for the higher grades and the most votes for the grade with the rating *very good suitable (5)*, we observe that both algorithms are rated to perform well for our data basis. However, TF-IDF performs slightly better. This becomes obvious by analyzing the box-plot diagram (right). The mean of the grades, symbolized by the round marker, is higher for TF-IDF than for TextRank. Additionally, the range of the 50%-box is smaller in the case of

TF-IDF.

A more detailed comparison of these two algorithms is presented in Figure 6. The mean of the given grades for keywords of a project is compared within the left diagram. In 66% of the projects, the keywords extracted by TF-IDF got higher or equal grades compared to TextRank.

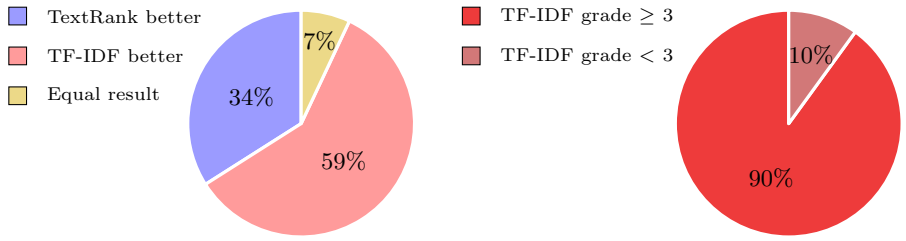

Figure 6: Comparison of algorithms' performance (left) and performance of TF-IDF (right).

Finally, the pie chart on the right shows that the mean grade of TF-IDF was higher than or equal to 3 in 90% of the projects. We can assume that the questionnaire's answers are a representative sample for the entire data basis because all texts have the same structure, and projects are in the same research field and funded by the same conveyor institution (BMWi). In conclusion, TF-IDF performs better for extracting keywords from our data basis. All following keyword networks and evaluations are generated with keywords extracted by this method.

5. Application

In the following, the results of the described method are shown. The validation (section 4) showed TF-IDF to be the better algorithm for our data basis. Therefore, the algorithm was used for the final keyword extraction of the project descriptions. Results are divided into those from the complete network of our data basis (subsection 5.1), a sub-net for projects in the field of energy research in buildings and neighborhoods (subsection 5.2), and overall (subsection 5.3) and recent (subsection 5.4) temporal developments in the relevance of research topics i.e. keywords.

To evaluate the context of the extracted keywords, we compared our findings with federal German research funding goals. Starting from 1970, the BMWK has provided programs of scientific and political ideas and goals. There are seven editions of this program, each of them covering another period. These programs mention key technologies, which should be a focus of energy research. These focal points were compared to the extracted topics and keywords across our analysis.

5.1. Network of the whole data set

The data set contains 7,259 project descriptions from which 15 keywords per description were extracted. In total, 33,660 individual keywords were extracted as nodes. 6,393,030 edges were created between these nodes, within the complete network, spanning from 1974 until 2020.

These keywords were reduced to 29,080 individual keywords through the fuzzy matching post-processing. We aimed to reduce the number of similar words whereby fuzzy matching before and after the keyword extraction showed the best results (subsection 3.2) and maximized the number of edges per individual keyword. Table 1 shows the number of keywords and edges before and after the two fuzzy matching steps. With every use of it, the number of keywords decreases while edges increase. The first fuzzy matching had the most significant

	w/o fuzzy matching	w/ first fuzzy matching	w/ second fuzzy matching
Keywords per description	15	15	15
Total edges	1,524,390	6,393,030	6,393,030
Individual keywords	33,660	33,125	29,671
Share[1] of ind. keywords	0.31	0.30	0.27
Edges per ind. keywords	45	193	215

[1]related to the number of all project descriptions multiplied with keywords per description

Table 1: Effects of using fuzzy matching in the process

effect, in which the edges are quadrupled. Especially the reduction of various word forms in the German language was observed within this step. The second matching improves the results after the extraction by aiming only at certain word forms prescribed before the fuzzy matching to clean up unwanted forms from the first fuzzy matching.

Figure 7 shows the complete visualization of the network across the entire timeline (untranslated German version in appendix: Figure .17). The indicator of weighted degree is used to rank the keywords according to their importance. In addition to the shown networks based on weighted degree, we examined a second one with betweenness centrality being the indicator and found that the keyword ranking behaved similarly.
The corresponding high-ranked keywords are shown in Table .2 of the appendix. The visualization inherits a bigger cluster in the center of the network. It is accompanied by many smaller ones at the edge of the network. Visual analysis shows that the network is heavily interconnected. The density is lower in peripheral areas, and clusters are smaller. This could indicate that the topics are highly interwoven in the research field.

Out of the analysis, keywords can be classified into the following categories:

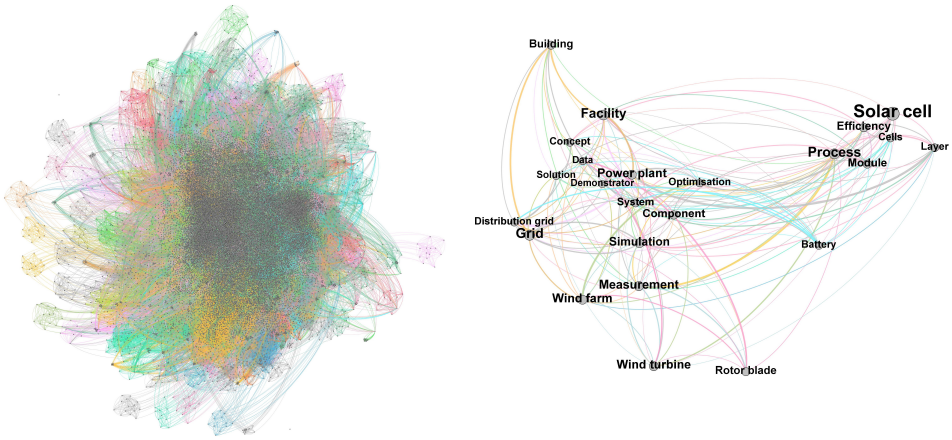

Figure 7: Complete Network on the left and filtered Cluster on right

- Directly named topics and technologies
- Subordinate terms of technologies
- Used research methods
- Unspecific terms in the context of technologies

Keywords like *solar cell*, *wind turbine*, *power plant*, *grid*, and *buildings* belong to the first category. They are under the highest rated keywords because of a high rating by the extraction, the number of occurrences of the keywords, and connections to other keywords of high importance. This could indicate an important role of these technologies in the German energy transition. Furthermore recent or future technologies like *battery* and *hydrogen* are present.
Related technologies i.e. subordinate terms for those mentioned above can also be found under the most relevant words. E.g. *rotor blade* and *wind farms* belong to wind turbines. *Distribution grid*, *inverter*, and *electricity* may belong to the research of the growing electricity *grids* due to integration of renewable energies.
In the third category, keywords which describe research methods like *optimization* and *simulation* or *experiments* and *measurements* can be found.
The last category of terms consists of unspecific words in the context of tech-

nologies and their meaning only becomes apparent within the context of the text. Those terms include descriptions, which are used in the general aspects of recent energy research like *process*, *component* or *efficiency*.
Directly named technologies from the keyword extraction correspond to those found in the governmental research programs. Early on, the federal government set a focus on developing renewable energy technologies, although nuclear and coal energy technologies were regarded as more important due to their economic advantages [73, 74].
Between 1970 and 2000, the research content mainly consisted of the fundamentals for renewable energy technologies in solar thermal plants and photovoltaics and wind, hydro, and geothermal energy. Also, heat pumps, district heating networks, combined heat and power, and biomass were focused. Thematic focal points were set early on building energy efficiency, electrical and thermal storages, and also on nuclear safety [75, 76]. Many of these focal points will be reflected throughout the following in-depth look at the network.

5.2. Network for the sub-research field energy in buildings and neighborhoods

By filtering the graph, a deeper view into specific research fields becomes possible. This chapter focuses on energy in buildings and neighborhoods, which was one of the central aspects of the comprehensive network presented before. Figure 8 (untranslated German version in appendix: Figure .18) and Table .3 show the corresponding network and data. Although this network is smaller than the complete network, no distinctive clusters are visible, and the data indicates a highly dense network. This indicates again that topics within the network are strongly interwoven.

Essential keywords within this sub-net can partly be found within the complete network from subsection 5.1. Important technologies in buildings include *solar thermal energy*, *heat pumps*, and *thermal heat storages*. Compared to the supply types electricity, air, water, and cooling, the supply with *heat* is in the

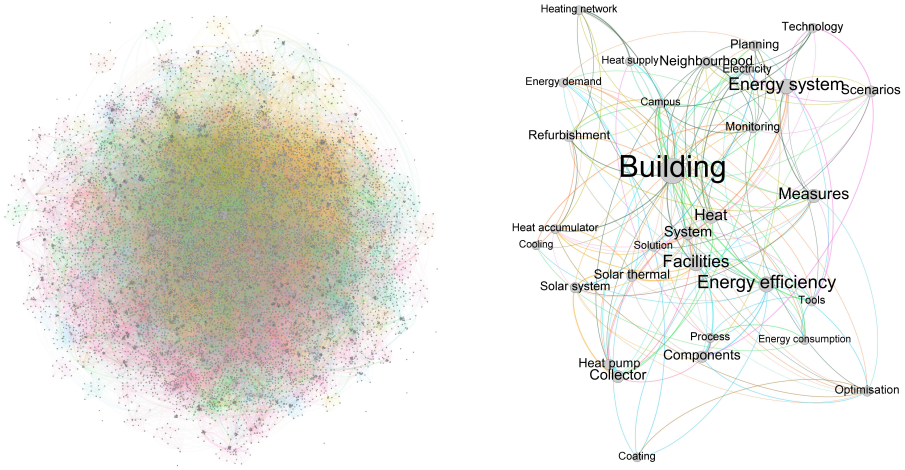

Figure 8: Network of the research field energy in buildings

center of energy research in buildings. This is also recognizable by terms like *heat network* and *heat accumulator*.

Furthermore, considerations of the whole *energy system* of buildings are focused and directly connected to the topics *technologies* and *scenarios*. This correlates to current efforts within the German research, which aim to determine how buildings can be supplied in the more distant future [77]. An important keyword within the sub-net is *refurbishment*, which means the modernization of existing buildings and the integration of renewable energy technologies. This finding is in line with the goal of an increasing renovation rate in Germany [78] and the higher importance of existing buildings compared to new constructions [79].

Neighborhood and *campus* are also found within the highly weighted keywords and show that often constellations of many buildings are part of recent research. In terms of neighborhoods, the BMWK recently started to fund so-called living labs [80]. University campus, on the other hand, are a popular test case for scientific projects [81, 82]. An analogous focus to the complete network is observable concerning the used research methods. Additionally, *monitoring* is an often used building-specific method for the measurement of the energy flow

inside a building and is therefore under the highly weighted keywords.

5.3. Overall temporal development of research topics

In this chapter, the temporal development of different keywords is compared with the goals set in different energy research programs of the last three decades. The underlying program's topic of preventing scarcity of resources was established within early programs [73, 74, 83] and changed towards to the aim of preventing climate change in the later years [84, 75, 76, 80].
After evaluating the complete network, a temporal analysis of keywords for each year within the data set was done. We determined 47 separated networks for each year of the data set, i.e., between 1974 and 2020. Different indicators for these networks were calculated, and their development is shown in Figure 9 and Figure 10.

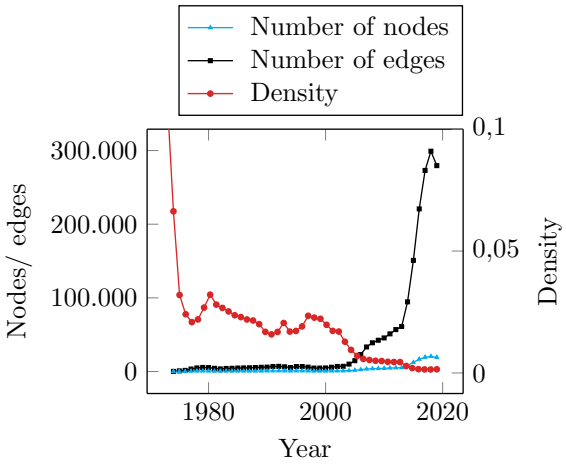

Figure 9: Characteristics of yearly networks

The average weighted degree of the keywords, normalized values, and nodes and edges increase over time. This correlates with the number of project descriptions which increase analogously (Figure 1). The density of the networks decreases over the years. This is probably because keywords got more diverse with the years as research treats more detailed topics. To compare the yearly

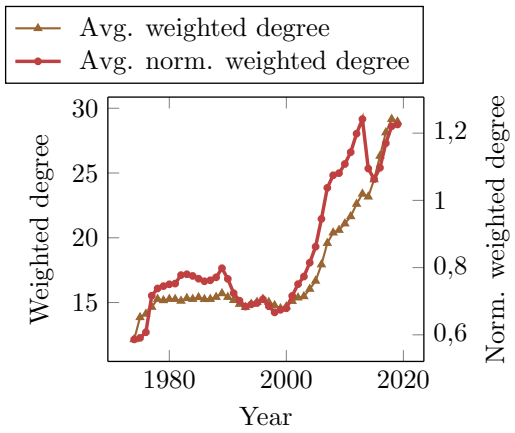

Figure 10: Average values across all networks

networks and related keywords, the weighted degree of each node was normalized (subsection 3.3). According to Figure 10, the average of the weighted degree and the normalized weighted degree overall keywords rise similarly over time. This indicates that after the year 2000, more keywords with a high weighted degree are present, whereby technological trends could be recognizable. We identified some of these trends and they are described in the following. The combination of decreasing density and increasing average weighted degree leads to the assumption that essential keywords are strongly interwoven with each other, while less essential keywords are not connected to other word clusters.
In order to identify trends, the development of the most relevant keywords over the last years can be analyzed with Figure 11. It shows the ten keywords with the highest value of the normalized weighted degree of each year between 2000 and 2020 and thematic groups. Since many project descriptions were available after 2000, TF-IDF performs well for this time horizon. However, also some non-specific keywords like *component*, *devices*, *process* which are often used in technical descriptions, are part of the highest-ranked keywords.

A total of 58 different keywords occur, covering various topics of German energy research. Between 2000 and 2010, the keywords change very rapidly, while

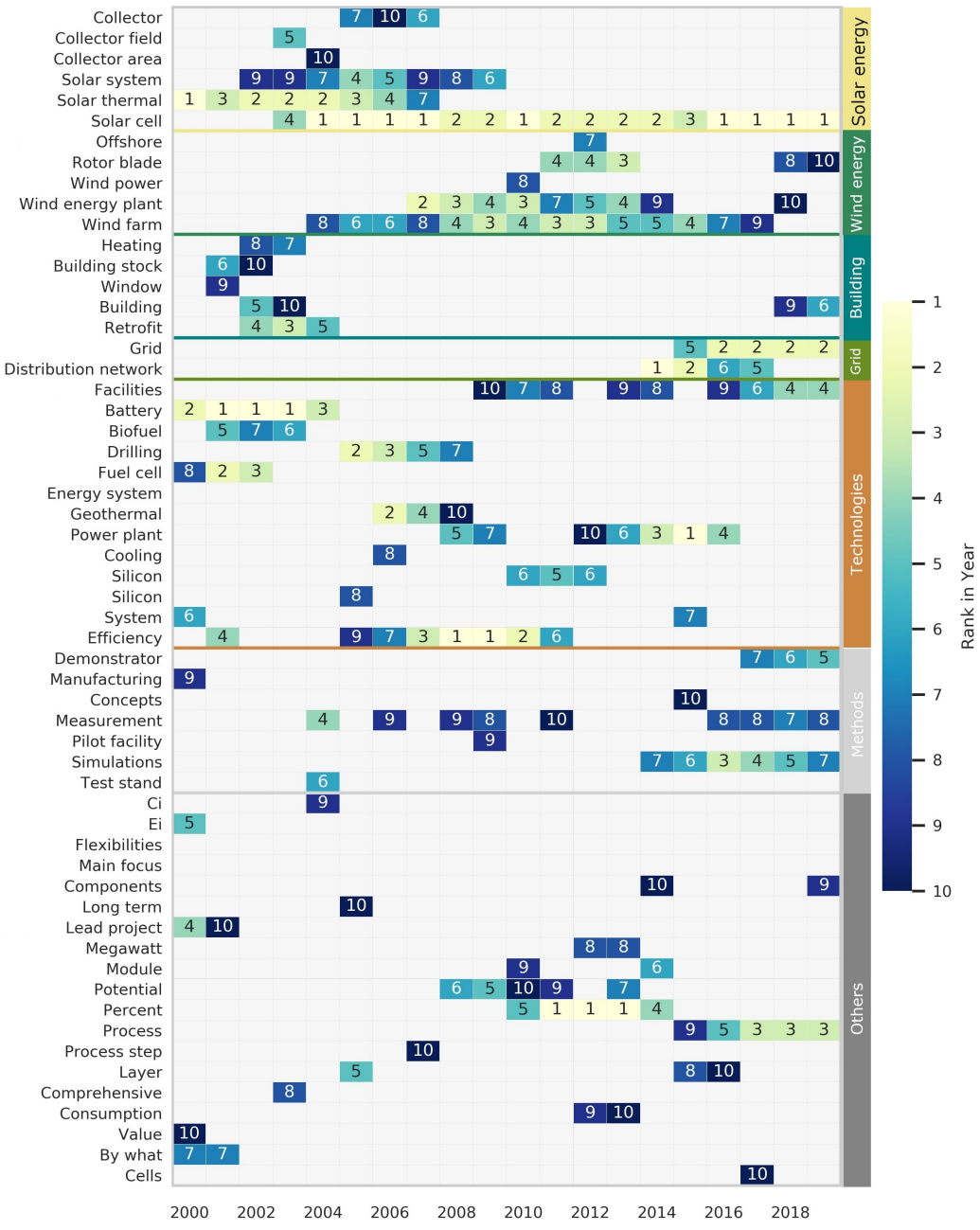

Figure 11: Keywords' occurrence in Top 10 for each year between 2000 and 2020

the rate of change decreases significantly after 2010. This means many topics between 2010 and 2020 have already been present between 2000 and 2010. This indicates that technological trends got stable after 2010, which might be due to a general change in the alignment of energy research at that time. As the leading funding authority of energy research in Germany, the BMWK shifted to more application-based research projects, while in years before 2000, fundamental research, technological efficiency, and resulting marketability were focused [76, 85]. Additionally, a thematic transition to renewable energy technologies was conducted in the last three decades.

Consistently relevant between 2000 and 2020 were *solar cell* and *wind farm* which are the two central technologies of the German energy transition. These findings are in line with other evaluations of energy research programs that found two-thirds of all available subsidies to be used for research in solar, and wind energy [86, 87, 73, 76, 80]. Concerning solar energy, we see *solar thermal energy* to be one of the highest-ranked keywords until 2010, but then it disappears. In contrast *solar cell* i.e., photovoltaics as the competing technology is highly relevant in research until today.

After 2010 *plants* moved into focus. These also include *power plants* until 2017. *Buildings* is the only topic that was in focus around 2000, then disappeared, and moved into focus in recent years. Furthermore, with *grid* and *distribution grid* words from the field of infrastructure have been under the highest-ranked ones from 2014 until today. This thematic focus was introduced in 2010 with the so-called sixth energy research plan [76] because the electricity grid needs to be improved to handle the additional electricity production by solar and wind technologies. In terms of methods, *measurement* is consistently used over the considered time horizon, and *simulations* gained even more relevance after 2014, which is due to the rising computation possibilities.

In the following, the temporal development of some keywords with the highest relevance in the whole network in subsection 5.1 shall be analyzed in detail. Therefore, these keywords were classified into six categories, each represented

in one sub-figure of Figure 12.

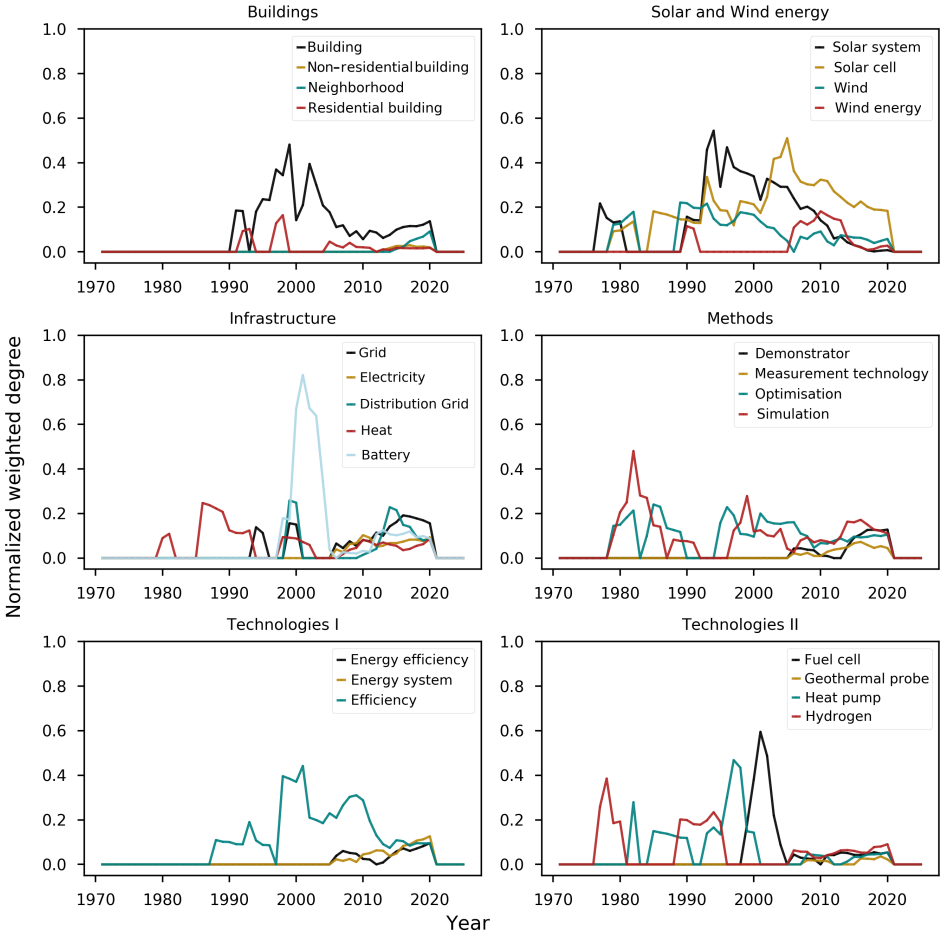

Figure 12: Changes of selected keywords' normalized weighted degree over time

From the created data in Figure 12, we observe that the normalized weighted degree of keywords with high rankings is more volatile before the year 2000, where fewer project descriptions are available.

While *buildings* remain relevant through all years, the focus seemed to be on single residential buildings until approximately 2010. Afterward, non-residential buildings have gained attention. One reason could be that these buildings' energy demand can be significantly higher than that of residentials, which of-

ten leads to a higher energy-saving potential in non-residentials. Additionally, *neighborhoods* seem to have got more important after 2010, which correlates to the trends already mentioned in subsection 5.1. This could be because synergy effects can be used when many buildings are observed together.

As wind and solar energy technologies are key technologies of the German energy transition, different related keywords have high rankings over the last three decades. Solar technologies like solar cells or silicium material have been part of the research topics since 1978. The interest in solar technologies had a peak around 2005 and decreased around 2020. The same can be observed for wind technologies. One reason could be that today the focus is more on integrating these technologies into the existing energy system. This assumption is supported by the analysis of keywords related to the energetic infrastructure, where we see a rising relevance of these keywords after 2010. As mentioned before, this was the year the BMWK introduced this topic with the goal of better handling of the additional renewable electricity [76].

Technologies with a recurring pattern are for example *battery* and *hydrogen*. The use of hydrogen had its highest peak around 1978, returned in 1990 and again in the years before 2000. This corresponds to the focus on funding fundamental research and efficiency research of hydrogen by the BMWK from 1970 onward [73, 74, 83]. *Heat pumps* belong to the returning technologies, which had a strong research period between 1980 and 2000 and started to gain relevance again a few years before 2010. This is in line with the fact that the improvement of this technology was part of early research, and today, its use cases are part of many studies after 2005 [88, 75]. Furthermore, gas-based heat pumps were the focus of any studies around 1980, while today, electricity-based heat pumps are treated as a key technology for the energy transition. *Fuel cells* also had a high relevance within the research between 1996 and 2006, where it was mainly developed and improved for different use cases. Recently it has been part of research for alternative mobile drives, and the integration in buildings [80].

Figure 12 (*Technologies I*) proves the above statements exemplary for some technologies. The improvement of single technologies (*efficiency*) has gradually

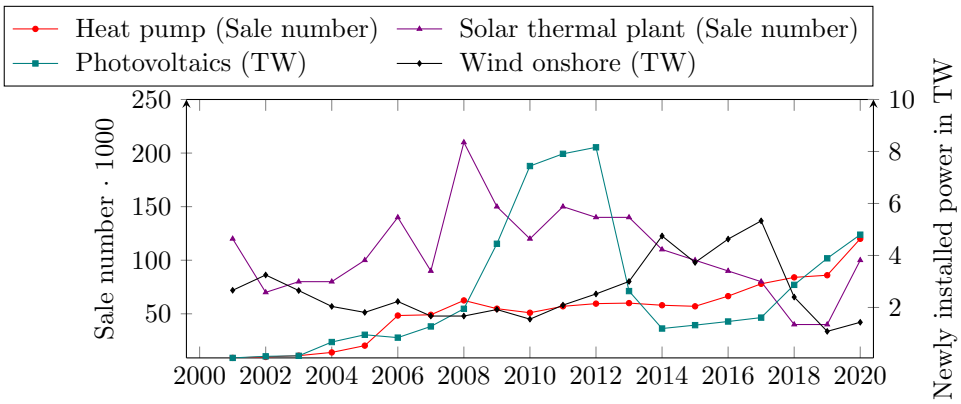

Figure 13: Market development of energy transition related technologies, data: [89, 90, 91, 92, 93]

been replaced with research on whole *energy systems* and their *energy efficiency* since 2005 [75, 76, 85].

We have analyzed projects in the field of applied research. Most efforts of applied research have the aim to develope technologies until they are nearly ready for the market. Therefore, a market analysis can give insights on the dependency between research and market. Figure 13 shows how four central technologies of the energy transition developed on the market over the last twenty years. For heat pumps (data: [89, 90]) and solar thermal energy (data: [91]) which are mainly used for the heat supply in buildings sale numbers are shown. For photovoltaics (data: [92]) that are used on buildings and as solar power plants on open spaces and for wind onshore (data: [93]) the newly installed power is shown.

In terms of solar thermal energy, we mentioned a decreasing relevance in research after 2010 (Figure 11). On the market, sale numbers also decreased few years after 2010. Photovoltaics, which was continuously relevant in research over the last twenty years (Figure 11) but also decreased in relevance in the last years (Figure 12), also decreased in market numbers at the same time as solar thermal energy.

Heat pumps have been highly relevant in research before 1980 and after 2010 (Figure 12). Steadily rising sale numbers after 2010 show a similar behaviour on the market. Wind energy had a high relevance in research over the last twenty years which decreased in the last years. Here, an analogous performance on the market can be observed.

We can conclude that a widely analogous behavior between research on these technologies and their performance on the market can be observed. However, research and market developments are both influenced by further effects like laws and state subsidies.

5.4. Recent temporal development of research topics

Beyond the general development of topics over time, this section describes the recent development in the last years. For this, we use the indicators acceleration and skewness presented in subsection 3.3.

Figure 14 shows the average acceleration of a keyword over its average normalized weighted degree between 2010 and 2020. The average weighted degree represents the general relevance of a keyword in this period, whereby the average acceleration quantifies the change of a keywords' relevance.

Some aspects that were mentioned during the analysis of Figure 11 can be verified with the observation of Figure 14. For example, *rotor blade* and *solar cell* have a high weighted degree, i.e., importance and low acceleration, which indicates that they have constantly been relevant over the whole period.

Beyond the previous findings, we can expect some words with moderate to high weighted degrees, like *DC (direct current)* or *capacitor*, to lose relevance over the following years because of their negative acceleration. Simultaneously, words like *grid*, *battery*, and *wind turbines* have been important and could stay important over the following years because of their high weighted degree and acceleration around zero. Keywords like *neighborhood* and *flexibility* could increase in relevance due to their high acceleration and weighted degree. This corresponds to the trends within governmental programs [76, 80], which emphasize flexibility options within coupling of different energy sectors and the synergy potentials in

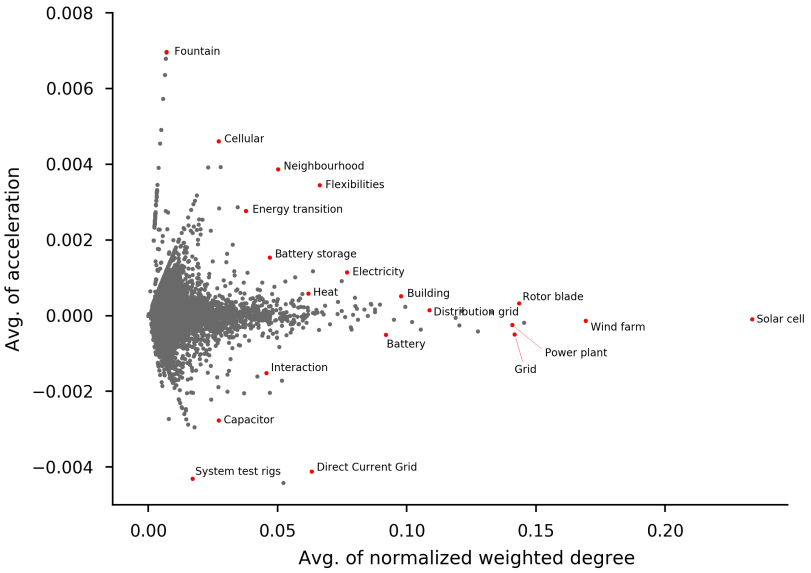

Figure 14: Normalized weighted degree and acceleration of all keywords

neighborhoods.

For a more detailed insight, Figure 15 shows the development of weighted degree, acceleration, and velocity of selected keywords over time. Compared to Figure 14 no average but yearly data is used. Hence, peaks of relevance, for example, *distribution grid* around 2015, become obvious, which is blurred by the average data. The already mentioned stable relevance of *DC*, rising relevance of *neighborhood* and decreasing relevance of *solar cell* are apparent again.

Finally, we present the results of the skewness computation in Figure 16. Like the evaluations of a keyword's acceleration, we aim at calculating the development of a keyword's relevance over time. Thus, the results can be compared to these of the acceleration evaluation.

As explained in subsection 3.3, the skewness describes the asymmetry of a distribution. The distribution consists of a keyword's weighted degree over time in our case. If this distribution has a high skewness, the keyword's relevance will change over time; if not, it is stable. Since the skewness only refers to the change of the relevance and not the relevance itself, it is shown in relation to its

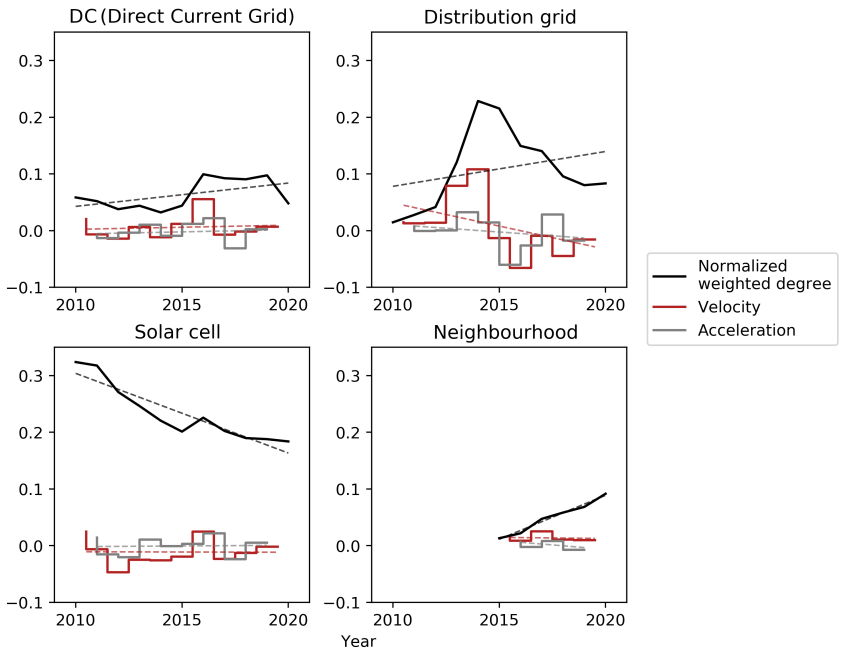

Figure 15: Acceleration, velocity and normalized weighted Degree over time compared for a selection of keywords

weighted degree.

For most of the keywords, similar results were found compared to the acceleration evaluation. For example, *grid*, *power plant*, and *wind farm* have a skewness and acceleration near to zero, which in each case shows that their relevance seems to be stable over time. This is also true for *battery storage*, *grid* which have a positive acceleration and skewness value.

In some cases, different results compared to the acceleration evaluation can be found. For example, *solar cell* has an average acceleration of nearly zero, but its skewness has a high value. This is because outliers of some years have a more significant effect on the skewness of the distribution than on the average acceleration. In Figure 15 it is shown that the relevance of this keyword decreases over time, which is due to a decreasing weighted degree. This fact was not shown in the analysis of the acceleration. Consequently, it is important to

derive effects based on more than one characteristic value.

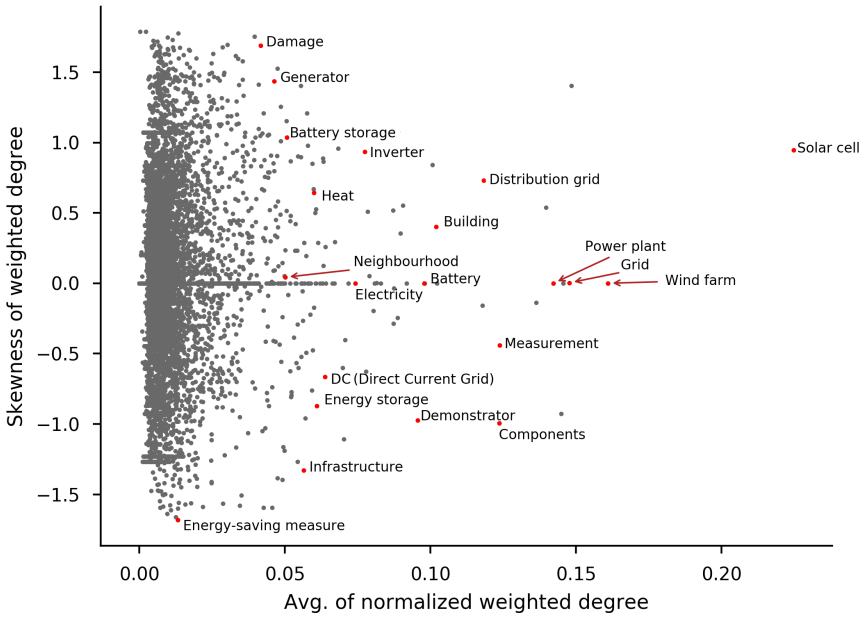

Figure 16: Skewness and normalized weighted degree for all keywords

6. Conclusions

The aim of this work was an automatic analysis of research project descriptions to analyze the relevance of different topics. Keywords were extracted from the project descriptions with TextRank and TF-IDF. Based on a questionnaire under the researchers of the studied projects, we analyzed that TF-IDF performed this keyword extraction better for our data basis. With the extracted keywords, graphical networks were built that show the connection of different research topics. By calculating characteristic values for the relevance of keywords inside these networks, the importance of different research topics at one point in time and how this importance has developed over time were analyzed. Major findings are that energy research before the year 2000 was focused on improving individual components. This is especially the case for solar and wind technologies. After 2000, the focus was directed towards systems integrating renewable energy technologies and improving the related infrastructure. A market analysis showed an analogous behavior between research on the central energy transition technologies and their performance on the market over the last twenty years.

Furthermore, we found that the focus and relevance of different topics and points in time widely reflect the governmental research programs. This indicates that the proposed method can describe the focus and trends of research topics based on project descriptions. Concerning the used characteristic values, the normalized weighted degree of a keyword showed good possibilities for analyzing the keyword's relevance at one point in time. The average acceleration of the weighted degree and the skewness of the distribution of the weighted degree's changes are proposed to analyze the development of the relevance over time. Summarized, we found the need to analyze research priorities and trends using more than one characteristic value.

One scientific novelty of our study is the usage of research project descriptions as a data basis for natural language processing. This provided a sufficiently large data basis to extract a variety of keywords and to find justifiable research focus

and trends. We could observe that the extraction of keywords is especially effective when enough project descriptions of a research field are available. Fewer descriptions led to the extraction of unclear keywords, making it challenging to analyze the relevance of a topic, i.e., keyword. Our work is based on a corpus in German. The language of the data basis is primarily important for the pre-processing before the keyword extraction is done. Most pre-processing steps were successfully transferred from known English approaches into German. However, some characteristics of the German language required individual adaptations. Especially, the stemming and lemmatization of technical terms needed manual revision because they could only be captured with difficulty by currently available German stemmer algorithms. In addition, the extensive occurrence of word compositions made it partly difficult to group technical words, which is why we used a supporting fuzzy matching approach.

For future work, existing German stemmer could be expanded by more technical words of the examined research field, which would decrease manual work. Further manual work could be avoided by using classification or clustering methods that can group word compositions based on their context. Furthermore, applying the proposed method to other research fields that provide project descriptions could give more insights on its generic possibilities to identify thematic focus and trends at one point in time and over time. In addition, cooperations with other countries could show which topics are being addressed internationally in research projects.

Acknowledgement

We gratefully acknowledge the financial support for this project by BMWK (German Federal Ministry for Economic Affairs and Climate Action) under promotional reference 03ET1388A and 03EWB002A. Furthermore, we would like to thank all project managers of the research projects that answered our online questionnaire and thereby helped to validate our keyword extraction methods.

References

[1] Federal Ministry for Economic Affairs and Energy (BMWi), Federal Report Energy Research 2020 - Research funding for the energy transition, Berlin, Germany, 2020.
URL https://www.bmwi.de/Redaktion/DE/Publikationen/Energie/bundesbericht-energieforschung-2020.pdf?__blob=publicationFile&v=16

[2] H. Ernst, Patent information for strategic technology management, World patent information 25 (3) (2003) 233–242. doi:10.1016/S0172-2190(03)00077-2.

[3] B. Yoon, Y. Park, A text-mining-based patent network: Analytical tool for high-technology trend, The Journal of High Technology Management Research 15 (1) (2004) 37–50. doi:10.1016/J.HITECH.2003.09.003.

[4] A. Abbas, L. Zhang, S. U. Khan, A literature review on the state-of-the-art in patent analysis, World Patent Information 37 (2014) 3–13. doi:10.1016/j.wpi.2013.12.006.

[5] T. Luukkonen-Gronow, Scientific research evaluation: a review of methods and various contexts of their application, R&D Management (1987) 207–221doi:10.1111/j.1467-9310.1987.tb00055.x.

[6] A. L. Porter, M. J. Detampel, Technology opportunities analysis, Technological Forecasting and Social Change 49 (3) (1995) 237–255. doi:10.1016/0040-1625(95)00022-3.

[7] R. Cammack, T. Atwood, P. Campbell, H. Parish, A. Smith, F. Vella, J. Stirling, Oxford Dictionary of Biochemistry and Molecular Biology (2 ed.), Oxford University Press, Oxford, 2006. doi:10.1093/acref/9780198529170.001.0001.

[8] Y.-H. Tseng, C.-J. Lin, Y.-I. Lin, Text mining techniques for patent analysis, Information processing & management 43 (5) (2007) 1216–1247. doi:10.1016/j.ipm.2006.11.011.

[9] Y. Liu, Z. Sun, Z. Da, R. Sun, World Traditional Medicine Patent Database and its applications, World Patent Information 36 (2014) 40–46. doi:10.1016/j.wpi.2013.12.001.

[10] J. Kim, S. Lee, Patent databases for innovation studies: A comparative analysis of USPTO, EPO, JPO and KIPO, Technological Forecasting and Social Change 92 (2015) 332–345. doi:10.1016/j.techfore.2015.01.009.

[11] M. A. Sharaf, M. A. Cheema, J. Qi (Eds.), Databases Theory and Applications: 26th Australasian Database Conference, ADC 2015, Melbourne, VIC, Australia, June 4-7, 2015. Proceedings, Vol. 9093, Springer International Publishing, 2015. doi:10.1007/978-3-319-19548-3.

[12] R. Wang, W. Liu, C. McDonald, Using word embeddings to enhance keyword identification for scientific publications, in: Australasian Database Conference, Springer, 2015, pp. 257–268. doi:10.1007/978-3-319-19548-3_21.

[13] X. L. Xiangfeng Luo, N. F. Ning Fang, W. X. Weimin Xu, S. Y. Sheng Yu, K. Y. Kai Yan, H. X. Huizhe Xiao, Experiments study for scientific texts domain keyword acquisition, in: 2006 Semantics, Knowledge and Grid, Second International Conference on, 2006, pp. 45–45. doi:10.1109/SKG.2006.50.

[14] P. K. Shah, C. Perez-Iratxeta, P. Bork, M. A. Andrade, Information extraction from full text scientific articles: where are the keywords?, BMC bioinformatics 4 (1) (2003) 20. doi:10.1186/1471-2105-4-20.

[15] H. Wang, Y. Ding, J. Tang, X. Dong, B. He, J. Qiu, D. J. Wild, Finding

complex biological relationships in recent PubMed articles using Bio-LDA, PloS one 6 (3) (2011) e17243. doi:10.1371/journal.pone.0017243.

[16] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, L. Zhao, Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey, Multimedia Tools and Applications 78 (11) (2019) 15169–15211. doi:10.1007/s11042-018-6894-4.

[17] A. Shvets, D. Devyatkin, I. Sochenkov, I. Tikhomirov, K. Popov, K. Yarygin, Detection of current research directions based on full-text clustering, in: 2015 Science and Information Conference (SAI), IEEE, 2015, pp. 483–488. doi:10.1145/3290605.3300451.

[18] F. Madani, C. Weber, The evolution of patent mining: Applying bibliometrics analysis and keyword network analysis, World Patent Information 46 (2016) 32–48. doi:10.1016/j.wpi.2016.05.008.

[19] S. Lee, S. Lee, H. Seol, Y. Park, Using patent information for designing new product and technology: Keyword based technology roadmapping, R&D Management 38 (2) (2008) 169–188. doi:10.1016/j.techfore.2011.08.015.

[20] L. Ziqiang, Research on the Forecasting Method of Research Hotspots Analysis Based on Time Series Model, Information Studies: Theory & Application (5) (2016) 6.

[21] D. Chiavetta, A. Porter, Tech mining for innovation management, Taylor & Francis, 2013. doi:10.1080/09537325.2013.802933.

[22] J. Beel, B. Gipp, S. Langer, C. Breitinger, Research-paper recommender systems: a literature survey, International Journal on Digital Libraries 17 (4) (2016) 305–338. doi:10.1007/s00799-015-0156-0.

[23] W. Li, J. Zhao, TextRank algorithm by exploiting Wikipedia for short text keywords extraction, in: 2016 3rd International Conference on Information

Science and Control Engineering (ICISCE), IEEE, 2016, pp. 683–686. doi:10.1109/ICISCE.2016.151.

[24] J. Kim, D. Choe, G. Kim, S. Park, D. Jang, Noise removal using TF-IDF criterion for extracting patent keyword, in: Soft Computing in Big Data Processing, Springer, 2014, pp. 61–69. doi:10.1007/978-3-319-05527-5_7.

[25] A. J. Trappey, C. V. Trappey, C.-Y. Wu, Automatic patent document summarization for collaborative knowledge systems and services, Journal of Systems Science and Systems Engineering 18 (1) (2009) 71–94. doi:10.1007/s11518-009-5100-7.

[26] J. Ramos, Using tf-idf to determine word relevance in document queries, in: Proceedings of the first instructional conference on machine learning, Vol. 242, New Jersey, USA, 2003, pp. 133–142.
URL https://www.semanticscholar.org/paper/Using-TF-IDF-to-Determine-Word-Relevance-in-Queries-Ramos/b3bf6373ff41a115197cb5b30e57830c16130c2c

[27] J. Li, K. Zhang, Keyword extraction based on tf/idf for Chinese news document, Wuhan University Journal of Natural Sciences 12 (5) (2007). doi:10.1007/s11859-007-0038-4.

[28] S.-J. Lee, H.-J. Kim, Keyword extraction from news corpus using modified TF-IDF, The Journal of Society for e-Business Studies 14 (4) (2009).

[29] J. Joung, K. Kim, Monitoring emerging technologies for technology planning using technical keyword based analysis from patent data, Technological Forecasting and Social Change (2017) 281–292doi:10.1016/j.techfore.2016.08.020.

[30] A. Onan, S. Korukoğlu, H. Bulut, Ensemble of keyword extraction methods and classifiers in text classification, Expert Systems with Applications 57 (2016) 232–247. doi:10.1016/j.eswa.2016.03.045.

[31] R. Mihalcea, P. Tarau, Textrank: Bringing order into text, in: Proceedings of the 2004 conference on empirical methods in natural language processing, 2004, pp. 404–411.

[32] G. Li, H. Wang, Improved automatic keyword extraction based on TextRank using domain knowledge, in: CCF International Conference on Natural Language Processing and Chinese Computing, Springer, 2014, pp. 403–413. doi:10.1007/978-3-662-45924-9_36.

[33] Z.-H. Wang, Y. Guo, Sentence-Ranking-Enhanced Keywords Extraction from Chinese Patents., Journal of Information Science & Engineering 35 (3) (2019). doi:10.6688/JISE.201905_35(3).0010.

[34] J. Hu, S. Li, Y. Yao, L. Yu, G. Yang, J. Hu, Patent keyword extraction algorithm based on distributed representation for patent classification, Entropy 20 (2) (2018) 104. doi:10.3390/e20020104.

[35] L. Peng, W. Bin, S. Zhiwei, C. Yachao, L. Hengxun, Tag-TextRank: A webpage keyword extraction method based on tags, Journal of Computer Research and Development 49 (11) (2012) 2344–2351.

[36] J. Yoon, S. Choi, K. Kim, Invention property-function network analysis of patents: a case of silicon-based thin film solar cells, Scientometrics 86 (3) (2011) 687–703. doi:10.1007/s11192-010-0303-8.

[37] J. Choi, Y.-S. Hwang, Patent keyword network analysis for improving technology development efficiency, Technological Forecasting and Social Change 83 (2014) 170–182. doi:10.1016/j.techfore.2013.07.004.

[38] E. Corley, J. Melkers, K. Johns, Layered and evolving networks: Innovative evaluation methods for interdisciplinary research in university-based research centers, in: The Atlanta Conference on S&T Policy, Atlanta, GA, 2006.

[39] I. H. Witten, Text Mining, Taylor & Francis, New York, US, 2004, Ch. Enabling Technologies. doi:10.1201/9780203507223.ch14.

[40] H. Park, K. Kim, S. Choi, J. Yoon, A patent intelligence system for strategic technology planning, Expert Systems with Applications 40 (7) (2013) 2373–2390. doi:10.1016/j.eswa.2012.10.073.

[41] S. Choi, J. Yoon, K. Kim, J. Y. Lee, C.-H. Kim, SAO network analysis of patents for technology trends identification: a case study of polymer electrolyte membrane technology in proton exchange membrane fuel cells, Scientometrics 88 (3) (2011) 863–883. doi:10.1007/s11192-011-0420-z.

[42] S. Choi, H. Park, D. Kang, J. Y. Lee, K. Kim, An SAO-based text mining approach to building a technology tree for technology planning, Expert Systems with Applications 39 (13) (2012) 11443–11455. doi:10.1111/j.1467-9310.2012.00702.x.

[43] J. Guo, X. Wang, Q. Li, D. Zhu, Subject–action–object-based morphology analysis for determining the direction of technological change, Technological Forecasting and Social Change 105 (2016) 27–40. doi:10.1016/j.techfore.2016.01.028.

[44] B. Yoon, Y. Park, A systematic approach for identifying technology opportunities: Keyword-based morphology analysis, Technological Forecasting and Social Change 72 (2) (2005) 145–160. doi:10.1016/j.techfore.2004.08.011.

[45] J. Lee, C. Kim, J. Shin, Technology opportunity discovery to R&D planning: Key technological performance analysis, Technological Forecasting and Social Change 119 (2017) 53–63. doi:10.1016/j.techfore.2017.03.011.

[46] H. Kwon, Y. Park, Y. Geum, Toward data-driven idea generation: Application of Wikipedia to morphological analysis, Technological Forecasting and Social Change 132 (2018) 56–80. doi:10.1016/j.techfore.2018.01.009.

[47] M. Kim, Y. Park, J. Yoon, Generating patent development maps for technology monitoring using semantic patent-topic analysis, Computers & In-

dustrial Engineering 98 (2016) 289–299. doi:10.1016/j.cie.2016.06.006.

[48] G. Kim, S. Park, D. Jang, Technology analysis from patent data using latent dirichlet allocation, in: Soft Computing in Big Data Processing, Springer, 2014, pp. 71–80. doi:10.1007/978-3-319-05527-5_8.

[49] H. Chen, G. Zhang, J. Lu, D. Zhu, A fuzzy approach for measuring development of topics in patents using latent Dirichlet allocation, in: IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), IEEE, 2015, pp. 1–7. doi:10.1109/FUZZ-IEEE.2015.7337980.

[50] A. J. Trappey, C. V. Trappey, T.-A. Chiang, Y.-H. Huang, Ontology-based neural network for patent knowledge management in design collaboration, International Journal of Production Research 51 (7) (2013) 1992–2005. doi:10.1080/00207543.2012.701775.

[51] P.-C. Lee, H.-N. Su, T.-Y. Chan, Assessment of ontology-based knowledge network formation by Vector-Space Model, Scientometrics 85 (3) (2010) 689–703. doi:10.1007/s11192-010-0267-8.

[52] C. Yang, H. Park, J. Heo, A network analysis of interdisciplinary research relationships: the Korean government's R&D grant program, Scientometrics 83 (1) (2010) 77–92. doi:10.1007/s11192-010-0157-0.

[53] C. S. Campbell, P. P. Maglio, A. Cozzi, B. Dom, Expertise identification using email communications, in: Proceedings of the twelfth international conference on Information and knowledge management, 2003, pp. 528–531. doi:10.1145/956863.956965.

[54] A.-L. Barabâsi, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, T. Vicsek, Evolution of the social network of scientific collaborations, Physica A: Statistical mechanics and its applications 311 (3-4) (2002) 590–614. doi:https://doi.org/10.1016/S0378-4371(02)00736-7.

[55] L. Oppermann, S. Hirzel, A. Güldner, K. Heiwolt, J. Krassowski, U. Schade, C. Lange, W. Prinz, Finding and analysing energy research funding data: The enargus system, Energy and AI 5 (2021) 100070. doi:https://doi.org/10.1016/j.egyai.2021.100070.
URL https://www.enargus.de/

[56] S. Lee, H.-J. Kim, News keyword extraction for topic tracking, in: 2008 Fourth International Conference on Networked Computing and Advanced Information Management, Vol. 2, 2008, pp. 554–559. doi:10.1109/NCM.2008.199.

[57] M. Honnibal, I. Montani, Spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing (2017).
URL https://spacy.io/

[58] S. Bird, E. Loper, E. Klein, Natural Language Processing with Python (2009).

[59] M. Konrad, GermaLemma (2019).
URL https://github.com/WZBSocialScienceCenter/germalemma

[60] A. Hulth, Improved automatic keyword extraction given more linguistic knowledge, in: Proceedings of the 2003 conference on Empirical methods in natural language processing, 2003, pp. 216–223. doi:10.3115/1119355.1119383.

[61] S. Beliga, A. Meštrović, S. Martinčić-Ipšić, An overview of graph-based keyword extraction methods and approaches, Journal of information and organizational sciences 39 (1) (2015) 1–20.

[62] Z. Liu, X. Chen, M. Sun, Mining the interests of Chinese microbloggers via keyword extraction, Frontiers of Computer Science 6 (1) (2012) 76–87. doi:10.1007/s11704-011-1174-8.

[63] R. G. Rossi, R. M. Marcacini, S. O. Rezende, Analysis of domain independent statistical keyword extraction methods for incremental clustering, Learning and Nonlinear Models 12 (1) (2014) 17–37. doi:10.21528/LNLM-VOL12-NO1-ART2.

[64] M. Habibi, A. Popescu-Belis, Keyword extraction and clustering for document recommendation in conversations, IEEE/ACM Transactions on audio, speech, and language processing 23 (4) (2015) 746–759. doi:10.1109/TASLP.2015.2405482.

[65] Y. Zhang, A. L. Porter, Z. Hu, Y. Guo, N. C. Newman, Term-clumping for technical intelligence: A case study on dye-sensitized solar cells, Technological Forecasting and Social Change 85 (2014) 26–39. doi:10.1016/j.techfore.2013.12.019.

[66] L. Page, S. Brin, R. Motwani, T. Winograd, The pagerank citation ranking: Bringing order to the web., Technical Report 1999-66, Stanford InfoLab, previous number = SIDL-WP-1999-0120 (November 1999).
URL http://ilpubs.stanford.edu:8090/422/

[67] S. Robertson, Understanding inverse document frequency: on theoretical arguments for IDF, Journal of documentation (2004). doi:10.1108/00220410410560582.

[68] C. Wartena, R. Brussee, W. Slakhorst, Keyword extraction using word co-occurrence, in: 2010 Workshops on Database and Expert Systems Applications, IEEE, 2010, pp. 54–58. doi:10.1142/S0218213004001466.

[69] M. E. Newman, Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality, Physical review E 64 (1) (2001) 016132. doi:10.1103/PhysRevE.64.016132.

[70] A. Barrat, M. Barthelemy, R. Pastor-Satorras, A. Vespignani, The architecture of complex weighted networks, Proceedings of the national academy of sciences 101 (11) (2004) 3747–3752. doi:10.1073/pnas.0400087101.

[71] U. Brandes, A faster algorithm for betweenness centrality, Journal of mathematical sociology 25 (2) (2001) 163–177. doi:10.1080/0022250X.2001.9990249.

[72] L. Kim, J. Ju, Can media forecast technological progress?: A text-mining approach to the on-line newspaper and blog's representation of prospective industrial technologies, Information Processing & Management 56 (4) (2019) 1506–1525. doi:10.1016/j.ipm.2018.10.017.

[73] Federal Ministry for Research and Technology, First program energy research and energy technologies 1977 – 1980, Bonn, 1979.
URL https://www.energieforschung.de/lw_resource/datapool/systemfiles/elements/files/92E60D620891561BE0539A695E865013/current/document/Energieforschungsprogramme_der_Bundesregierung_1977_-_2017_.pdf

[74] Federal Ministry for Research and Technology, Second program energy research and energy technologies, Bonn, 1981.
URL https://www.energieforschung.de/lw_resource/datapool/systemfiles/elements/files/92E60D620891561BE0539A695E865013/current/document/Energieforschungsprogramme_der_Bundesregierung_1977_-_2017_.pdf

[75] Federal Ministry for Economics and Labor, Fifth program energy research and energy technologies, Berlin, 2005.
URL https://www.energieforschung.de/lw_resource/datapool/systemfiles/elements/files/92E60D620891561BE0539A695E865013/current/document/Energieforschungsprogramme_der_Bundesregierung_1977_-_2017_.pdf

[76] Federal Ministry for Economic Affairs and Energy, Sixth program energy research and energy technologies, Berlin, 2011.
URL https://www.bmwi.de/Redaktion/DE/Publikationen/Energie/

6-energieforschungsprogramm-der-bundesregierung.pdf?__blob=publicationFile&v=32

[77] V. Bürger, T. Hesse, D. Quack, A. Palzer, B. Köhler, S. Herkel, P. Engelmann, Climate Neutral Building Stock 2050 - Energy Efficiency Potentials and the Impact of Climate Change on the Building Stock, Federal Environment Agency (UBA), 2017.
URL https://www.umweltbundesamt.de/sites/default/files/medien/1410/publikationen/2017-11-06_climate-change_26-2017_klimaneutraler-gebaeudebestand-ii.pdf

[78] German energy agency (DENA), Building report 2019 - Statistics and analyses on energy efficiency in existing buildings, German energy agency (DENA), 2019.
URL https://www.dena.de/fileadmin/dena/Publikationen/PDFs/2019/dena-GEBAEUDEREPORT_KOMPAKT_2019.pdf

[79] W. Canzler, F. Engels, J.-C. Rogge, D. Simon, A. Wentland, From living lab to strategic action field: Bringing together energy, mobility, and information technology in germany, Energy research & social science 27 (2017) 25–35. doi:https://dx.doi.org/10.1016/j.erss.2017.02.003.

[80] Federal Ministry for Economic Affairs and Energy, Seventh program energy research and energy technologies, Berlin, 2018.
URL https://www.bmwi.de/Redaktion/DE/Publikationen/Energie/7-energieforschungsprogramm-der-bundesregierung.pdf?__blob=publicationFile&v=14

[81] L. R. Soleymani, M. Kurrat, Energy efficiency in the campus: Case study of technische universitaet braunschweig, in: International ETG Congress 2017, VDE, 2017, pp. 1–6.

[82] D. Müller, C. A. van Treeck, D. H. Braun, D. Wenner, Schlussbericht - EnEff:Campus RoadMap RWTH Aachen, RWTH Aachen University, E.ON

Energy Research Center, Lehrstuhl für Gebäude- und Raumklimatechnik, Aachen, 2017. doi:http://dx.doi.org/10.2314/GBV:1009837389.

[83] Federal Ministry for Research and Technology, Third program energy research and energy technologies, Bonn, 1991.
URL https://www.energieforschung.de/lw_resource/datapool/systemfiles/elements/files/92E60D620891561BE0539A695E865013/current/document/Energieforschungsprogramme_der_Bundesregierung_1977_-_2017_.pdf

[84] Federal Ministry for Education, Science, Research and Technology, Fourth program energy research and energy technologies, Bonn, 1996.
URL https://www.energieforschung.de/lw_resource/datapool/systemfiles/elements/files/92E60D620891561BE0539A695E865013/current/document/Energieforschungsprogramme_der_Bundesregierung_1977_-_2017_.pdf

[85] K. Kübler, Innovation and research for more energy efficiency - new accents in the federal government's energy research policy (2011).
URL https://www.ffe.de/fachtagung2011/download/FfE-Fachtagung_2011_Vortrag_Dr_Kuebler.pdf

[86] F. Seefeldt, O. Pfirrmann, A. Heimer, B. Weinmann, U. Glöckner, C. Michelsen, H. Schleinitz, S. Knab, Evaluation of the fourth energy research program renewable energies, Prognos AG, Berlin, 2007.
URL https://www.prognos.com/fileadmin/pdf/1190449164.pdf

[87] D. S. Heinrich, F. Seefeldt, G. Gerdes, M. Reichmuth, Evaluation of research funding of the Federal Environment Ministry in the framework of the fifth energy research program - Federal Ministry for Economy and Energy, Bundesministerium für Wirtschaft und Energie, Berlin, 2014.
URL https://www.bmwi.de/Redaktion/DE/Publikationen/Studien/evaluation-der-forschungsfoerderung-des-bundesumweltministeriums-im-rahmen-des-5-energi
pdf

[88] M. Knoll, L. Illge, V. Handke, B. Oertel, F. Korte, D. Mauer, J. Onodera, Final report - evaluation of project funding of the federal ministry for economy and technology in energy research, department energy efficiency in industry, trade and services (ighd) in the framework of the fifth energy research program (2014).
URL https://www.ptj.de/lw_resource/datapool/systemfiles/cbox/1444/live/lw_file/evaluation.pdf

[89] Michael Platt, Stephan Exner, Rolf Bracke, Analysis of the German Heat Pump Market - inventory and trends, Bochum, 2010.
URL https://www.erneuerbare-energien.de/EE/Redaktion/DE/Downloads/Studien/analyse-waermepumpenmarkt.pdf?__blob=publicationFile&v=2

[90] Bundesverband Wärmepumpe (BDH), Sale numbers heat pumps (2021).
URL https://www.waermepumpe.de/presse/zahlen-daten/absatzzahlen/

[91] Bundesverband Solarwirtschaft (BSW), Sale numbers solar thermal plants (2021).
URL https://www.solarwirtschaft.de/

[92] Federal Ministry for Economic Affairs and Energy, Time series on the development of renewable energies in germany (2021).
URL https://www.erneuerbare-energien.de/EE/Navigation/DE/Service/Erneuerbare_Energien_in_Zahlen/Zeitreihen/zeitreihen.html

[93] Deutsche Windguard, Status of onshore wind energy expansion in germany (2021).
URL https://www.windguard.de/jahr-2021.html?file=files/cto_layout/img/unternehmen/windenergiestatistik/2021/Jahr/Status20des20Windenergieausbaus20an20Land_Jahr202021.pdf

Appendix

Appendix

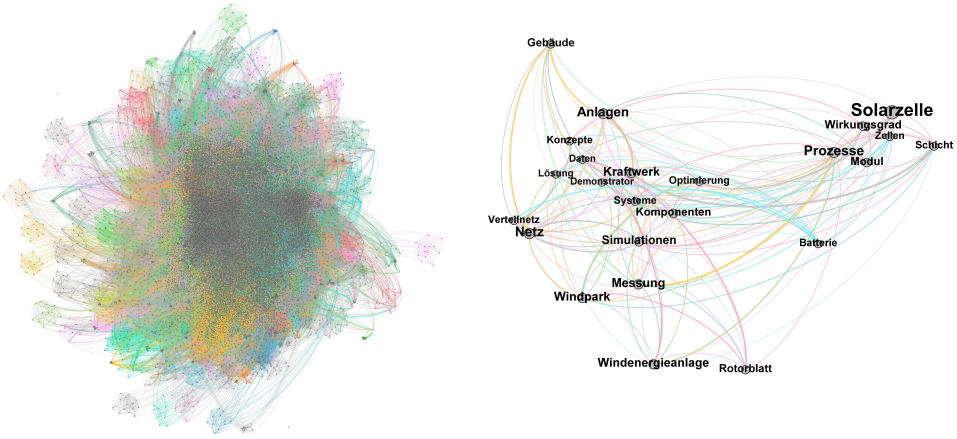

Figure .17: Complete German network (left) and filtered clusters (right)

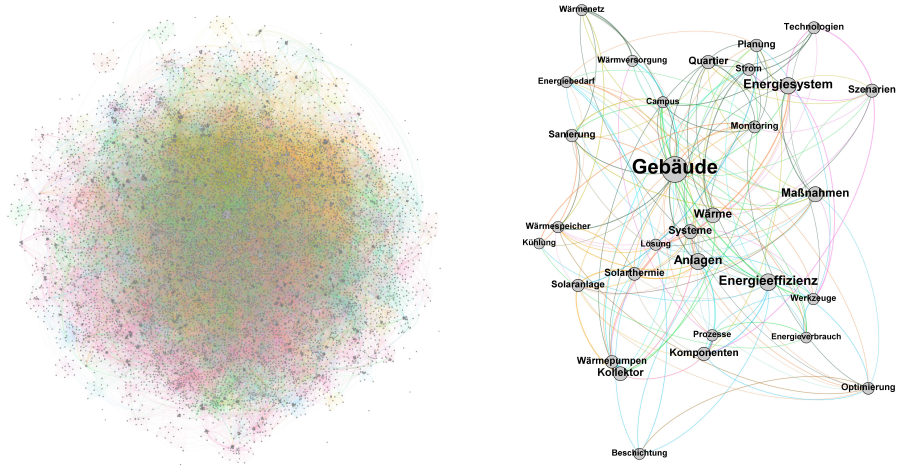

Figure .18: German network (left) and filtered clusters (right) of the research field energy in buildings and neighborhoods

Label	Weighted Degree		Betweeness Centrality	
	Value	Rank	Value	Rank
Solar cell (Solarzelle)	83.354	1	0,0206	1
Process (Prozesse)	61.115	2	0,0186	3
Network/grid (Netz)	60.316	3	0,0174	4
Devices (Anlagen)	58.752	4	0,0199	2
Measurement (Messung)	53.529	5	0,0163	5
Wind farm (Windpark)	53.169	6	0,0116	11
Power plant (Kraftwerk)	52.387	7	0,0112	14
Wind energy plant (Windenergieanlage)	52.257	8	0,0106	16
Simulations (Simulationen)	51.267	9	0,0159	6
Efficiency (Wirkungsgrad)	48.120	10	0,0112	13
Components (Komponenten)	47.601	11	0,0127	8
Module (Modul)	47.155	12	0,0119	10
Rotor blade (Rotorblatt)	46.185	13	0,0084	27
Building (Gebäude)	46.116	14	0,0131	7
Layer (Schicht)	43.046	15	0,0099	19
System (System)	42.520	16	0,0104	18
Concept (Konzepte)	41.774	17	0,0093	22
Optimization (Optimierung)	41.470	18	0,0115	12
Cells (Zellen)	41.408	19	0,0096	20
Battery (Batterie)	41.292	20	0,0096	21

Table .2: Characteristic values of the most important keywords for the complete network from 1974 to 2020

Label	Weighted Degree		Betweeness Centrality	
	Value	Rank	Value	Rank
Building (Gebäude)	41.943	1	0.0716	1
Energy efficiency (Energieeffizienz)	23.365	2	0.0419	2
Energy system (Energiesystem)	22.404	3	0.0220	6
Devices (Anlagen)	21.808	4	0.0290	3
Measure (Maßnahmen)	19.969	5	0.0228	5
Heat (Wärme)	19.219	6	0.0229	4
System (System)	17.492	7	0.0196	8
Collector (Kollektor)	17.363	8	0.0102	26
Component (Komponenten)	16.332	9	0.0190	9
Neighborhood (Quartier)	16.181	10	0.0119	20
Scenario (Szenario)	15.805	11	0.0140	14
Solar thermal energy (Solarthermie)	14.984	12	0.0085	33
Heat pump (Wärmepumpe)	14.837	13	0.0135	16
Planning (Planung)	14.523	14	0.0150	12
Renovation (Sanierung)	14.514	15	0.0088	30
Solar plant (Solaranlage)	13.997	16	0.0073	44
Technologies (Technologien)	13.574	17	0.0133	17
Monitoring (Monitoring)	13.465	18	0.0149	13
Optimization (Optimierung)	13.205	19	0.0202	7
Current (Strom)	13.033	20	0.0109	23

Table .3: Characteristic values of the most important keywords for the EWB sub-net from 1974 to 2020