Creating research topic map for NIMS SAMURAI database using natural language processing approach

Sae Dieb*, Kou Amano#, Kosuke Tanabe#, Daitetsu Sato*, Masashi Ishii# and Mikiko Tanifuji*

*National Institute for Materials Science, Materials Data Platform Center, MaDIS, Namiki, Tsukuba, Japan; #National Institute of Informatics, Research Center for Open Science and Data Platform, Chiyoda-ku, Tokyo, Japan; #TelHi Corporation, Chuo-ku, Tokyo, Japan

ABSTRACT
In this study, we present an approach to create a visual research topic map for materials science researchers from a large collection of archived research papers using natural language processing (NLP). We apply this approach on SAMURAI, a directory service for the researchers of the National Institute for Materials Science (NIMS), Japan. Visualization of research content will support exploratory searches by maximizing the information absorbance and intuitively capturing the research characteristics of each materials science researcher. In addition, a research topic map can connect researchers with similar topics aiming to find potential collaborators. Collaboration can support the advance of scientific research. We analyze all available publications of a researcher using frequent term analysis. In addition, materials science knowledge resources were utilized including dictionaries and automatic extraction for the material names. Noise reduction was implemented using stop word filtering. The topics were then visualized using a word cloud technique for each researcher. An analysis of the topic similarity was conducted to find researchers that share similar topics leading to the creation of a topic map for each researcher. The approach aims at maximizing information absorbance for public knowledge by applying NLP approaches to information mining from materials science research papers. NLP analysis and visualization code are available https://github.com/ThaerDieb/Topic_map.

KEYWORDS
Text and data mining; materials science; word cloud; topic visualization; topic map

CLASSIFICATION
Computer systems and services

1. Introduction
With the load of available research papers published continuously by materials scientists, it is desirable to maximize the absorbance of the extent of information that exists in these papers. The use of natural language processing is getting more attention to analyse scientific publications for knowledge discovery [1,2]. For example, scientific publications clustering [3]. Recently, topic extraction and visualization are being utilized to tackle this issue because of their intuitive display of information that facilitates information access. Additionally, topic visualisation allows to capture the research trend quickly without having to navigate an entire list of publications. Furthermore, it can support the link between researchers with similar research topics to find potential collaboration leading to the construction of a topic map. Such digital infrastructure systems are necessary to facilitate the accumulation and distribution of scientific data.

In this paper, we present an approach to create a research topic map for materials science researchers from a large collection of publications using natural language processing techniques and materials science knowledge resources such as dictionaries. We extract representative topic terms from research publications that are related to materials science and engineering. We use frequent term analysis and automatic extraction of material names in addition to domain knowledge resources. Noise reduction is implemented by removing stop words. A word cloud approach was used to visualize the extracted terms. Finally, a topic similarity analysis was conducted to find researchers who share similar topics, leading to a research topic map for each researcher.

Visualization of topics from an unstructured text has been previously studied using several different approaches. For example, Miller et al. applied wavelet transforms to a custom digital signal constructed from words within a document to create visualisation [4]. Recently, word cloud visualization technique [5,6] has been widely used to present the contents and themes in a text for summary and visualization. Examples include Gottron et al. work on visualizing web documents [7] and ConceptCloud browser, which makes use of a combination of concept lattices and tag clouds, to visually present academic publication data by Dunaiski et al. that also used interactive word cloud visualization [8]. Such a technique is a visual representation of word frequency using different font sizes and colors to represent terms with different levels of
importance. Venetis et al. presented metrics to capture the structural properties of a tag cloud, and a set of tag selection algorithms [9]. Some publishers have implemented a word cloud from research papers on a large scale. An example of this is Elsevier SciVal [10]. Although these studies had a significant impact on the visualization of research content, there has been no focus on the materials science domain. It is necessary to use domain-aware resources to improve the processing for materials science research publications. A few attempts have been made to visualize the research content in materials science; however, they have been restricted to a limited research area such as quantum materials only [11].

There is some work that has been conducted to provide collaborator suggestion such as on ‘ResearchGate’ or ‘Researchmap’; however, these functions are for general science community and do not focus on materials science. In addition, our proposed system provides a wider connected network of potential collaborators. A special feature of this system is that collaborator suggestion is focused on each topic of interest rather than on the general researcher output. For example, a materials scientist seeks to find a collaborator on developing a certain material or a specific measurement method.

A key task in the construction of a topic map is the extraction of key terms [12,13] that represent the researchers’ topics from their publications. Frequent term analysis has been utilized for text mining applications. For example, Nagwani et al. designed a frequent term based text summarization algorithm [14]. The term frequency can reflect the importance of this term in a text of a collection of texts. However, in scientific publications, term frequency analysis based on the general content of the paper is insufficient. It is necessary to study the term frequency for domain-specific terms. For example, in materials science, researchers are interested in searching for publications discussing the synthesis or characterization of certain materials. In this study, we utilized domain knowledge dictionaries and an automatic chemical formula extraction tool to extract materials names and other domain-specific terms from research publications. In addition, the importance of the term can be weighted depending on not only its frequency but also other factors such as its location in the text (title, keywords, abstract).

In this study, we validate the proposed approach on SAMURAI [15], a researcher directory service maintained by the National Institute for Materials Science (NIMS), Japan for its researchers. Publications for each researcher are linked to the NIMS Digital Library [16]. SAMURAI can synchronize the profile information of researchers and their publications self-archived in the NIMS institutional repository systems [17-20]. This database has interoperable functions using the Open Researcher and Contributor ID (ORCID) [21]. It identifies authors of journal articles from NIMS and research members of NIMS through the implementation of a global ID. Each researcher profile is attached to a unique ResearcherID linked to major global databases such as KAKEN (Database of Grants-in-Aid for Scientific Research by National Institute for Informatics (NII)). Through this directory, NIMS is promoting materials research, supporting the management of its researchers’ activities, and introducing NIMS researchers and their work to the public.

Even though we validated this approach on SAMURAI, it can be applied to visualize the content of research papers related to materials science domain in general as it integrates domain knowledge features and it is not dependent on the database itself.

The remainder of this paper is organized as follows. In Sec. 2, we introduce our approach for researcher output visualization and topic map construction. The details of the experiment and results are discussed in Sec. 3. Sec. 4 then provides some concluding remarks and discusses areas of future study.

2. Method

2.1. Outline

The proposed approach consists of 6 main components: in the data collection, a set of materials science publications are collected in XML format for a group of researchers from a database. Next, these publications are preprocessed by extracting certain sections and removing noisy data. Additionally, domain knowledge terms are extracted in a separate step. After that, frequent term analysis is conducted for each extracted section. In the visualization component, a vector representing each researcher’s topics is constructed based on frequent term analysis from the researcher’s publications. This vector is visualized using word cloud technique. In the final step, similarity analysis between different researchers is conducted to construct topic map. Figure 1 illustrates the outline of the system. Details of each component are discussed in the following subsections.

2.2. Data collection

The SAMURAI database holds a list of publications for each researcher that is linked to the NIMS digital library and other data sources. A list of publication digital object identifiers (DOIs) [22] for each researcher was constructed from the researcher’s profile in SAMURAI (in SAMURAI, the DOIs are stored in a PostgreSQL database). Using the DOIs, recent publications, mainly available beginning in 2004, were retrieved from the NIMS text and data mining
platform (TDM-PF). The dataset used for this study consists of 8269 published articles by 1058 SAMURAI researchers in XML format. Publications are the sum of all retrieved publications for each author using the DOIs.

2.3. Data preprocessing

We extracted the plain text from the following sections of all the publications authored by each SAMURAI researcher:

- Title of the paper
- 'Keywords' section
- Abstract of the paper

We used the XML tag name for each section provided by the publisher to extract the section. Because the main body of the publication is not collected, the effect of including the 'Keywords' section is balanced. The text of each section is segmented into tokens using the NLTK word tokenizer. Natural Language Toolkit (NLTK) is an open-source Python package with data sets, supporting research and development in Natural Language Processing [23]. The following processing steps are then conducted on the tokenized texts to remove noisy data:

- Removing numeric values, punctuation marks (for example, “23.5”, “,”, “?”). Such data are not related to the topics discussed in the publications. Even though numerical values are important in some “grey area” topics such as when reporting catalysts with “high yield”, however, it is beyond the scope of this study to tackle these issues.
- Filtering general English language stop-words such as “but”, “an”, “he”. These stop-words frequently occurring in English but do not carry a thematic component or significance themselves [24].
- Physical units such as “mi” (meter) for length measurement, and “K” (Kelvin) for temperature measurement, among others, are frequently found in materials science research publications; however, for our objective in this study, they are not informative regarding the research output. The list was compiled using SI-based and derived units [25].

Other terms can also be filtered out depending on the application. For example, too frequent terms in a certain sub-domain of materials science do not contribute to similarity metrics between researchers in this sub-domain.

2.4. Utilization of domain knowledge resources

2.4.1. Motivation

We aim to construct a topic map for materials science researchers, so they can find potential collaborators. Using general linguistic features only is not sufficient to compute the similarity between researchers’ topics that are focused on materials science. It is necessary to include domain related knowledge terms. It has been proven that utilization of domain knowledge features has leveraged the performance of text mining applications in materials science [26]. In this study, we used two types of domain knowledge resources, one is the chemical compound extractor and the other is a dictionary for measurement related terms.

2.4.2. Extraction of chemical compounds

In materials science research, materials names play an extremely informative role in the content of the publication. Many publications are centered around the synthesis or characterization of a specific material. A segmentation of text using general language tokenization schema might miss the material names. For this reason, it is necessary to extract material names using a domain-specific tokenization schema [27] to provide chemistry-aware tokenization. Because numerous material names are basically chemical compounds, we developed a chemical formula extraction tool to extract materials names from text. This tool uses regular expressions and element information from the periodic table to identify chemical compounds formula.

2.4.3. Extraction of measurement related terms

Measurements of material properties is a fundamental issue in materials science and engineering. Some materials scientists even specialize in measurement science. A significant number of materials science publications contain discussions on measurement and related issues. Identifying measurement related terms such as methods, results, and simulation, is necessary to characterize the researcher output. As discussed in Section 2.4.2, general English language-based tokenization schema might have a low matching
ratio for domain-specific knowledge including measurement related terms. Domain-aware tokenization issue was reported in other domains such as bioinformatics [28,29]. Domain-aware tokenizers were also developed [30]. Matching ratio analysis for some domain-specific terms was conducted [27]. We used a digitized version of the Japanese dictionary of physics and chemistry [31] to extract several measurement related terms translated into English:

- Measurement method dictionary, for example, X-ray diffraction microscopy and Maxam-Gilbert method [32].
- Simulation category such as the Monte Carlo method [33].

### 2.5. Frequent term analysis

We computed term frequency for the terms extracted in Section 2.3 for each section (title, keywords, abstracts) separately. In addition, the term frequency was calculated for the terms extracted in Sections 2.4.2 and 2.4.3. Five sets of (term, frequency) were created and normalized based on the length of the extracted sections for each researcher as follows:

- \( T_{if} \) Title terms extracted from all publications.
- \( K_{if} \) Keywords terms extracted from all publications.
- \( A_{if} \) Abstract terms extracted from all publications.
- \( M_{if} \) Material formulas extracted from all three sections of all research publications.
- \( ME_{if} \) Measurement related terms extracted from all three sections of all research publications.

The research output for a researcher is represented as a set \( R_{if} \) resulting from the merging of five sets. In research publications, sections in the paper structure have different levels of influence on the publication topic [34]. For example, an occurrence in the title is more topic representative than an occurrence in the body. Several studies have used a term weighting strategy for document processing applications [35,36]. For this reason, occurrences of the terms should be weighted differently based on their location in the publication. We use a simple strategy by assigning a heavier weight for terms in the title and the keywords section than those in the abstract. Because domain knowledge terms are extracted separately from other terms, the frequencies of domain knowledge terms are used in the case of double extraction to avoid double-counting. We denote \( T_w, K_w, A_w, M_w, \) and \( ME_w \) as the weights for \( T_{if}, K_{if}, A_{if}, M_{if}, \) and \( ME_{if}, \) respectively. Here \( R_{if} \) is given by Equation (1)

\[
R_{if} = T_w \ast T_{if} \cup K_w \ast K_{if} \cup A_w \ast A_{if} \cup M_w \ast M_{if} \cup ME_w \ast ME_{if}
\]

where the values of the weights \( T_w, K_w, A_w, M_w, \) and \( ME_w \) are determined by experimenting on a small subset of the data and using feedback from materials scientists in SAMURAI and are 1.2, 1.5, 1.0, 1.0 and 1.0 respectively.

### 2.6. Word cloud generation

The word cloud for each researcher is generated using the \( R_{if} \) set of the researcher, i.e. terms and their frequencies multiplied by the designated weights. We use a word cloud generator package in Python language [37] with the following settings: the maximum number of words equals 200, and minimum number of letters for a word to be included is zero.

### 2.7. Topic map construction

Research collaborations are extremely beneficial to advance sciences. Collaborators can learn from each other by exchanging ideas and sharing skills. To find potential collaborators, we want to analyze researchers who share similar topics. Each researcher is represented by a vector \( V \) of discrete variables \( v \in \mathbb{N}, \) where \( v \) indicates the frequencies from the set \( R_{if} \) discussed in Section 2.5. The similarity between the two given researchers are computed using the cosine similarity metric between the two vectors of these researchers, \( V_1, V_2, \) as indicated in (2):

\[
similarity = \cos(\theta) = \frac{V_1 \cdot V_2}{||V_1|| ||V_2||} = \frac{\sum v_i v_j}{\sqrt{\sum v_i^2} \sqrt{\sum v_j^2}}
\]

where \( v_i, v_j \) are components of \( V_1, V_2 \) respectively. When two researchers have a high similarity, the most frequent common term is extracted by intersecting \( R_{if} \) of both researchers. This approach was applied for each SAMURAI researcher in relation to all others to create a relationship based on frequent common terms between researchers. We visualize these relationships as links to create a topic map for the researchers.

### 3. Experiment and results

#### 3.1. Visualization of word cloud

Using the approach discussed in Section 2, we created a word cloud and topics maps for all SAMURAI members. We demonstrate the results here using the three researchers below as examples. The number in the parentheses is the number of the author publications that could be retrieved from the NIMS repository. Coauthored publications by multiple SAMURAI members are counted for each author independently.
Kazuhiro Hono [38], affiliated with the Research Center for Magnetic and Spintronic Materials in NIMS (177).

Yuya Sakuraba [39], also affiliated with the Research Center for Magnetic and Spintronic Materials in NIMS (22).

Hideki Yoshikawa [40], affiliated with the Research and Services Division of Materials Data and the Integrated System, and Research Center for Advanced Measurement and Characterization in NIMS (51).

All the above mentioned researchers have explicitly agreed to have their names, photos, and related data published in this manuscript. Because Kazuhiro Hono and Yuya Sakuraba are affiliated with the same research center, it is expected that they would have similar research topics that differ from those of Hideki Yoshikawa. Figure 2 shows a simplified word cloud visualization for each of these researchers’ topics.

From Figure 2, it can be seen that both Kazuhiro Hono and Yuya Sakuraba have common terms with a large font size such as ‘spin’, ‘magnetics’, and ‘heusler’. This is correlated with the fact that they both belong to the same research center. By contrast, Yuya Sakuraba does not share any of his top terms (‘spin’, ‘Ag’, ‘magnetoresistance’, ‘heusler’) with Hideki Yoshikawa, who belongs to a different center. This indicates that researchers with different affiliations are less correlated with each other.

However, we can also see that there are some irrelevant terms owing to an insufficient tokenization, such as ‘investigated’. Morphological analysis and/or part-of-speech tagging could be used to further improve tokenization efficiency. Considering that our visualization based on an assuming model and limited data is qualitative and projective, a parallel use or model selection would be useful for the visualization of many facets. As pointed out above, the domain knowledge terms are very important aspects for characterizing the research field.

3.2. Research topic map

Using the word cloud generated in Section 3.1, a vector for each researcher was constructed using the method discussed in Section 2.7. Using cosine similarity, researcher vector similarity was checked against all other SAMURAI researchers. When the similarity is high between two researchers, word cloud sets of them were intersected to find the best common term and a link between them was created between their profiles. Figure 3 shows the topic map for each researcher mentioned in Section 3.1. From Figure 3, we can observe the topic terms that a researcher shares with another or multiple researchers. This map can provide insights on potential collaboration on a specific topic.

When publications are authored by multiple authors, the researcher position in the author list can affect his interest topics. For example, the topics from a publication where the researcher is the principal investigator (PI) are closer to his interest than those of a publication where he is not. To verify the differences in extracted topic terms when the researcher is the PI or not, we compare the word cloud. Figure 4 shows Yuya Sakuraba topics terms for his publications as a PI. When comparing Figure 2 (b) and Figure 4, we can notice that even though major terms are similar; however the frequency (importance) of the terms has changed. For example, ‘NiMnSb’ has
gained more importance in publications where the researcher is the first or corresponding author. This approach could be extended to add a weighting factor based on the position of the researcher in the author list.

### 3.3. Effect of domain knowledge resources

In this section, we study how the utilization of domain knowledge resources influence the word cloud visualization. Table 1 shows information about the extracted word cloud terms including terms of domain knowledge resources.

From Table 1, domain knowledge terms make up approximately 5.7% of all the terms in the generated word clouds and 7.8% of the unique terms. Chemical compounds make up approximately 5.5% of all terms and 7.5% of the unique terms. This confirms the importance of the extraction of chemical compounds in representing materials science research because many material names exist in their chemical compounds form. By contrast, measurement related resources in this study did not contribute to representing the output of the researchers. Measurement terms make up less than 0.5% of the total and unique extracted terms. Table 2 shows the most frequent domain knowledge terms in the generated word cloud.

To visualize how the utilization of domain knowledge resources affects the research output representation, we compare the word cloud visualization with

---

**Figure 4.** Word cloud visualization for the researcher Yuya Sakuraba topics as extracted from publications where he is either the first or the corresponding author.

**Table 1.** Terms extracted for word cloud visualization and topic map creation including domain knowledge terms. Publications are the sum of all retrieved publications for each author.

| Researchers | Terms | Publications | Chemical compound | Domain knowledge | Measurement related | Others | Sum |
|-------------|-------|--------------|-------------------|------------------|--------------------|--------|------|
| 1058        | Total | 8269         | 5762              | 189              | 98992              | 104943 |
|             | % of sum | 5.5% | 0.2%             | 94.3%           | 100%              |        |
|             | Unique | 1161 | 40               | 14262           | 15463             |        |
|             | % of sum | 7.5% | 0.3%             | 92.2%           | 100%              |        |

**Table 2.** Top domain knowledge terms extracted for word cloud visualization and topic map creation.

| Chemical compound | Measurement terms |
|-------------------|-------------------|
| In                | X-ray photoelectron spectroscopy |
| Si                | Phase             |
| Fe                | Monte Carlo       |
| Al                | XPS               |

---

**Figure 5.** Comparison of word cloud visualization with and without using domain knowledge resources for the same researcher.
and without using domain knowledge resources for the same SAMURAI researcher. Figure 5 illustrates the differences.

This presented researcher is currently affiliated with both the Electric and Electronic Materials Field, Nano Electronics Device Materials Group, Research Center for Functional Materials, and the Quantum Beam Field, Synchrotron X-ray Group, Research Center for Advanced Measurement and Characterization. From Figure 5, we confirmed that the use of domain knowledge resources generated terms that were unable to be found through a normal frequent terms analysis.

3.4. Researcher correlation analysis

In order to validate the word cloud results, we analyzed the correlation of SAMURAI researchers who belong to different research centers within NIMS using clustering techniques. Researchers who belong to the same center are expected to have a stronger correlation with each other and a weaker correlation with researchers in different centers. We selected two centers, namely, the Research Center for Magnetic and Spintronic Materials in NIMS, and the Research Center for Advanced Measurement and Characterization. SAMURAI members with no publications available in NIMS repositories are not included.

The word cloud for these researchers was generated using the proposed approach described in Section 2. Using these word cloud terms and their relative importance values, a vector for each researcher is created. We used two methods to perform clustering on these vectors. The first is the K-means algorithm [41,42], where a multidimensional dataset is grouped into K number of clusters using the concept in which a cluster center is the arithmetic mean of all its data points. We set K to 2 because we have two different research centers. The second method is the agglomerative clustering, which is a ‘bottom-up’ approach of hierarchical clustering [43], in which each data point starts as a cluster, and merged as it moves up the hierarchy. There are different ways to determine how clusters are merged. In our setting, we use the ward linkage that minimizes the variance of the clusters being merged using the euclidean metric. The number of clusters is also set to 2. Figure 6 shows the clustering results using both methods. To produce a 2D plot, the high dimensionality of the dataset is reduced using the truncated singular value decomposition (T-SVD) method. T-SVD is a matrix decomposition method suitable for applications with a sparse dataset [44] such as our data.

We evaluate the quality of the clustering from internal and external perspectives. An internal evaluation is used to compare the performance of different algorithm on clustering the same dataset, whereas an external evaluation is used to evaluate the clustering results using known class labels for the data points.

We use the three following metrics:

- Davies–Bouldin index [45] is used for the internal evaluation, which is given by Equation (3) as follows:

\[
DBI = \frac{1}{n} \sum_{i=1}^{n} \max_{j \neq i} \left( \frac{\sigma_i + \sigma_j}{\text{distance}(c_i, c_j)} \right)
\]

where \(n\) is the number of clusters, \(c_i\) is the centroid of cluster \(i\), \(\sigma_i\) is the average distance of all elements in cluster \(i\) to the centroid. Clustering algorithm that produces clusters with low intra-cluster distances and high inter-cluster distances will have a lower Davies–Bouldin index, so a smaller Davies–Bouldin index is better.

- The purity (external metric), which measures the extent to which clusters contain a single class [46], is given by Equation (4) as follows:

\[
Purity = \frac{1}{N} \sum_{m \in M} \max_{d \in D} |m \cap d|
\]

where \(N\) is the number of data points, \(M\) is the set of clusters, and \(D\) is a set of known classes for the data points.

- The Adjusted Rand index ARI (external metric) measures how similar the clusters are to

![Figure 6. Clustering results for researchers with two different affiliations in NIMS based on their word cloud. Data points are projected on 2D using truncated singular value decomposition.](image-url)
the known class labels [47]. It is adjusted from RI score for chance. RI is given by Equation (5):

\[
RI = \frac{\text{true positive} + \text{true negative}}{\text{true positive} + \text{false negative} + \text{false positive} + \text{true negative}}
\]  

(5)

Since the true labels are known, true positive, true negative, false positive, and false negative can be computed against the ground truth. Then ARI can be computed from RI using Equation (6)

\[
ARI = \frac{RI - \text{Expected}_RI}{\max(RI) - \text{Expected}_RI}
\]  

(6)

ARI is then will have a value close to 0.0 for random labeling.

Table 3 contains the evaluation metric values for both clustering methods. We can confirm that agglomerative clustering performed better on our dataset from internal evaluation point of view; however, K-means had a better performance from the external evaluation perspective. The high purity in both methods showed a high correlation between researchers affiliated with the same center (same class). The adjusted Rand index of the K-means clustering showed a 70% accuracy, eliminating the possibility that the accuracy occurred simply by chance.

### 4. Conclusion and future studies

In this study, we presented an approach to create a topic map for the materials science researchers using natural language processing (NLP). Publications were collected and preprocessed to remove noisy data. Domain knowledge resources and frequent term analysis were utilized to extract the representative terms. We conducted similarity analysis to find links between researchers who share similar research topics leading to construct a topic map for each researcher. This approach was validated on SAMURAI, a researchers database at the National Institute for Materials Science (NIMS), Japan. We confirmed the stronger correlation between researchers with the same affiliation and a weaker correlation with researchers from different affiliations by performing a clustering experiment.

In the future, we plan to conduct morphological analysis to improve tokenization efficiency. Another extension is to consider a weighting factor for each researcher based on his position in the author list (for example, corresponding author, first author). In addition, we plan to collect and analyze other types of resources such as research notes. We are also implementing an interactive evaluation and adjustment system. Researchers will be allowed to adjust the level of importance of their own research topics or add new terms.

### Acknowledgements

This work was supported by "Research and Services Division of Materials Data and Integrated System (MaDIS)”, fund No. PG4230 from the National Institute for Materials Science (NIMS), Japan. The authors would like to thank the anonymous reviewers for their instructive comments to improve the manuscript.

### Disclosure statement

The authors declare no conflict of interest.

### ORCID

Masashi Ishii [http://orcid.org/0000-0003-0357-2832](http://orcid.org/0000-0003-0357-2832)

### References

[1] Kondo T, Nanba H, Takezawa T, et al Technical trend analysis by analyzing research papers’ titles. In: Vetulani Zeditor. Human language technology. Challenges for computer science and linguistics. Berlin, Heidelberg: Springer Berlin Heidelberg; 2011. p. 512–521.

[2] Poelmans J, Ignatov DI, Viaene S, et al Text mining scientific papers: a survey on fca-based information retrieval research. In: Perner Peditor. Advances in data mining. Applications and theoretical aspects. Berlin, Heidelberg: Springer Berlin Heidelberg; 2012. p. 273–287.

[3] Liu X, Yu S, Janssens F, et al. Weighted hybrid clustering by combining text mining and bibliometrics on a large-scale journal database. J Am Soc Inf Sci Technol. 2010;61(6):1105–1119. Available from: [https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.21312](https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.21312)

[4] Miller NE, Wong PC, Brewster M, et al Topic islands/ sup tm/-a wavelet-based text visualization system. In: Proceedings Visualization ’98 Research Triangle Park North Carolina USA (Cat. No.98CB36276); 1998. p. 189–196.

[5] Cui W, Wu Y, Liu S, et al Context preserving dynamic word cloud visualization. In: 2010 IEEE Pacific Visualization Symposium (PacificVis); Taipei (Taiwan); 2010. p. 121–128.

[6] Lohmann S, Heimerl F, Bopp F, et al Concentri cloud: word cloud visualization for multiple text documents. In: 2015 19th International Conference on Information Visualisation; Barcelona (Spain); 2015. p. 114–120.

[7] Gottron T. Document word clouds: visualising web documents as tag clouds to aid users in relevance decisions. In: Agosti M, Borbinha J, Kapidakis Seditors.
Research and advanced technology for digital libraries.
Berlin, Heidelberg: Springer Berlin Heidelberg: 2009. p. 94–105.

[8] Dunaisky M, Greene GI, Fischer B. Exploratory search of academic publication and citation data using interactive tag cloud visualizations. Scientometrics. 2017;110(3):1539–1571.

[9] Venetis P, Koutrika G, Garcia-Molina H. On the selection of tags for tag clouds. In: Fourth ACM International Conference on Web Search and Data Mining (WSDM 2011). Stanford InfoLab; Hong Kong (China); February 2011. Available from: http://ilpubs.stanford.edu:8090/992/.

[10] SciVal; 2020 November 25. Available from: https://www.scival.com/.

[11] Venugopal V, Broderick SR, Rajan K. A picture is worth a thousand words: applying natural language processing tools for creating a quantum materials database map. MRS Commun. 2019;9(4):1134–1141.

[12] Zhang C, Wang H, Liu Y, et al. Automatic keyword extraction from documents using conditional random fields. J Comput Informat Syst. 2008;4 (3):1169–1180.

[13] Ercan G, Cicikli I. Using lexical chains for keyword extraction. Informat Process Manage. 2007;43 (6):1705–1714. Available from: http://www.sciencedirect.com/science/article/pii/S0306457307000398

[14] Nagwani NK, Verma S. A frequent term and semantic similarity based single document text summarization algorithm. In: J Comput Appl. 2013;17(2):975–887.

[15] SAMURAI. 2020 November 24. Available from: https://samurai.nims.go.jp/.

[16] NIMS Digital Library; 2020 November 24. Available from: https://library.nims.go.jp/.

[17] Takaku M, Tanifuji M. Developing institutional repository at national institute for materials science: researchers directory service “SAMURAI” and research collection library. J Inform Process Manage. 2012;55(1):29–41.

[18] Tanifuji M, Matsuda A, Yoshikawa H. Materials data platform - a FAIR system for data-driven materials science. In: 2019 8th International Congress on Advanced Applied Informatics (IIA-IAI); Toyama (Japan); 2019. p. 1021–1022.

[19] Tanabe K, Matsuda A, Tanifuji M. Designing a vocabulary service for a ‘data-driven’ materials data repository. 2019. Available from: 10.5281/zenodo.3554054.

[20] Ishii M, Takemura T, Tanifuji M. PolylInfo RDF: a semantically reinforced polymer database for materials informatics. ISWC Satell. 2019.

[21] ORCID. 2020 November 25. Available from: https://orcid.org/.

[22] DOI. 2020 November 25. Available from: https://www.doi.org/.

[23] Bird S, Klein E, Loper E. Natural language processing with python. Sebastopol (CA): O’Reilly Media, Inc; 2009. Available from: https://www.oreilly.com/library/view/natural-language-processing/9780596003346/.

[24] Rajaraman A, Ullman JD. Mining of Massive Datasets. Cambridge (England): Cambridge University Press; 2011. p. 1–17.

[25] NIST, SI units. 2020 November 29. Available from: https://physics.nist.gov/cuu/Units/units.html.

[26] Dieb TM, Yoshioka M, Hara S, et al Framework for automatic information extraction from research papers on nanocrystal devices. Beilstein J Nanotechnol. 2015;6:1872–1882.

[27] M Dieb T, Yoshioka M. Extraction of chemical and drug named entities by ensemble learning using chemical ner tools based on different extraction guidelines. Trans MLDM. 2015;8:61–76.

[28] Morgan AA, Lu Z, Wang X, et al. Overview of bio-creative ii gene normalization. Genome Biol. 2008;9 (2):S3.

[29] Leaman R, Gonzalez G. Banner: an executable survey of advances in biomedical named entity recognition. Pac Symp Biocomput. 2008;652-63(PMID): 18229723.

[30] Swain MC, Cole JM. Chemdataextractor: a toolkit for automated extraction of chemical information from the scientific literature. J Chem Inf Model. 2016 10:56 (10):1894–1904.

[31] Nagakura S, Inokuchi H, Ezawa H, et al. Japanese dictionary of physics and chemistry. Tokyo (Japan): Iwanami; 1998.

[32] Maxam AM, Gilbert W. A new method for sequencing dna. Proc Nat Acad Sci. 1977;74(2):560–564. Available from: https://www.pnas.org/content/74/2/560.

[33] Kroese DP, Chan CC. Statistical modeling and computation. USA: Springer-Verlag New York; 2014.

[34] Trotman A. Choosing document structure weights. Informat Process Manage. 2005;41(2):243–264. Available from: http://www.sciencedirect.com/science/article/pii/S0306457305000852.

[35] Lan M, Tan CL, Low HB, et al A comprehensive comparative study on term weighting schemes for text categorization with support vector machines. In: Special Interest Tracks and Posters of the 14th International Conference on World Wide Web; New York, NY, USA. Association for Computing Machinery; 2005. p. 1032–1033; WWW ’05. doi:10.1145/1062745.1062854.

[36] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. Information Process Manage. 1988;24(5):513–523. Available from: http://www.science direct.com/science/article/pii/S0306457388900210.

[37] Andreas M Word cloud python package. 2020 November 30. Available from: https://amueller.github.io/word_cloud/index.html.

[38] Hono K; 2020 December 1. Available from: https://samurai.nims.go.jp/profiles/kazuhiro_hono .

[39] Sakuraba Y; 2020 December 1. Available from: https://samurai.nims.go.jp/profiles/sakuraba_yuya .

[40] Yoshikawa H; 2020 December 1. Available from: https://samurai.nims.go.jp/profiles/yoshikawa_hideki .

[41] Lloyd S. Least squares quantization in pcm. IEEE Trans Inf Theory. 1982;28(2):129–137.

[42] Pelleg D, Moore A. Accelerating exact k-means algorithms with geometric reasoning. In: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; New York, NY, USA. Association for Computing Machinery; 1999. p. 277–281; KDD ’99. doi:10.1145/312129.312248.

[43] Jain AK, Dubes RC. Algorithms for clustering data. USA: Prentice-Hall Inc.; 1988.

[44] Wu Y, Zhang Y, Mao D, et al Sparse super-resolution method based on truncated singular value decomposition strategy for radar forward-looking imaging. J Appl Remote Sens. 2018;12(3):1–20.
[45] Davies DL, Bouldin DW. A cluster separation measure. IEEE Trans Pattern Anal Mach Intell. 1979;PAMI-1(2):224–227.

[46] Sanderson M, Christopher D. Manning, prabhakar raghavan, hinrich schütze, introduction to information retrieval, cambridge university press. 2008. isbn-13 978-0-521-86571-5, xxi 482 pages. Nat Language Eng. 2010;16(1):100–103.

[47] Rand WM. Objective criteria for the evaluation of clustering methods. J Am Stat Assoc. 1971;66 (336):846–850. Available from: https://www.tandfonline.com/doi/abs/10.1080/01621459.1971.10482356.