HAL
archives-ouvertes.fr

Automatic Extraction of IDM-Related Information in Scientific Articles and Online Science News Websites

Oriane Nédey, Achille Souili, Denis Cavallucci

▶ **To cite this version:**

Oriane Nédey, Achille Souili, Denis Cavallucci. Automatic Extraction of IDM-Related Information in Scientific Articles and Online Science News Websites. 18th TRIZ Future Conference (TFC), Oct 2018, Strasbourg, France. pp.213-224, 10.1007/978-3-030-02456-7_18 . hal-02279756

HAL Id: hal-02279756
https://hal.inria.fr/hal-02279756

Submitted on 5 Sep 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution| 4.0 International License

Automatic extraction of IDM-related information in scientific articles and online science news websites

Abstract. Previous studies have made it possible to extract information related to IDM (Inventive Design Method) out of patents. IDM is an ontology-defined method derived from TRIZ. As its mother theory, IDM is primarily based on patent's observation and aims at finding inventive solutions on the basis of contradictions. In this paper, we present a new approach for extracting knowledge, this time out of other types of science-related documents: scientific papers and science news articles. This approach is based on a supervised Machine Learning model that classifies automatically the sentences of a text according to IDM's ontology concepts such as problems, partial solutions and parameters. We use two manually annotated corpora (one for each type of document) for the learning part as well as for the evaluation (with the percentage split method). Moreover, a set of semi-automatically selected linguistic features is involved in the model in order to improve the classification.

Keywords: TRIZ, IDM, Inventive Design, Machine Learning, Knowledge Extraction, Text Mining, NLP

1 Introduction

The World Intellectual Property Organization claims that more than 3 million patents were published worldwide in 2016, that is 8.3% more than in 2015 [1]. Next to this massive source of technical knowledge, scientific articles also carry weight in expert scientific content with at least 2.5 million new scientific papers published each year since 2013 [2]. While patents are legal sources which provide technical information about the innovations they aim to protect, scientific articles provide theoretical as well as technical answers which help the scientific community understanding and gaining control of their environment through diverse applications. With such a huge number of publications and given that they are not always easily accessible, science news websites have developed on the Internet and advertise some important researches and innovative breakthroughs in a condensed and accessible form, making it possible for the scientific community to stay update and learn about advances in other science-related fields.

With innovation going very fast, engineers are facing a challenge to find creative ideas that may lead to innovation. To help them in their innovation process, researches have been made since the 1990s, developing and adapting Altshuller's theory of inventive problem solving (TRIZ) [3]. Yet, though patents seem the best way for engineers to find solutions [4], it doesn't mean that other science-related documents cannot also lead to innovation. In order to determine the significance of this kind of documents on innovation, we need to find out whether these documents include enough content useful for TRIZ theory application. During the last few years, our laboratory has been developing a tool for automated extraction of knowledge related to the ontology of the Inventive Design Method (IDM) from English-language patents [5].

Following these researches, we want to adapt the information extraction from patents to the case of scientific papers and science news articles, i.e. automating the extraction of topics, problems, partial solutions, evaluation and action parameters, as well as values, inside these two alternative types of science-related documents.

This is a challenging project in that both scientific research articles and science news articles are unstructured data, with much more liberty regarding their structure and writing style. Therefore, we need to evaluate the structural, syntactic as well as semantic features implemented in the patent-knowledge extraction tool [5], and then to find new features that are specific to the new document types. This implies the use of machine learning and other advanced methods for Natural Language Processing.

In this article, we will give you an overview of the literature on the Inventive Design Method (IDM) and its tool for patent-knowledge extraction, followed by the literature on scientific research papers and science news articles (Section 2), and then we will present our methodology along with the current advances on the project and an emphasis on our evaluation method (Section 3).

2 State of the art

2.1 The Inventive Design Method

The Inventive Design Method (IDM) is an application method derived from TRIZ, the theory of inventive problem solving. The mother theory was developed in the ex-USSR by Genrich Altshuller and has started to spread worldwide in the 1990s, with multiple applications and tools in the engineering field [5]. The theory is based upon four central notions which are key to an inventive solution without compromise: a few evolution laws that apply to all technical systems and their environment, the principle of increasing ideality, the principle of contradictions – at administrative, technical and physical levels –, and finally the principle of available resources.

The derivate Inventive Design Method describes four steps for innovation [6] : during the first phase, the users must extract knowledge and organize it into a graph. With this graph, they must then formulate a set of contradictions, which will be solved individually in phase 3, and finally, they must choose the most innovative Solution Concepts before investing in it and setting it up.

For the contradiction formulation, the Inventive Design Method offers a formal and practical definition of broad TRIZ contradiction notion, which is very useful for

industrial innovation and introduces other notions linked to IDM-ontology [7]. This contradiction "is characterized by a set of three parameters […] where one of the parameters can take two possible opposite values *Va* and $\overline{Va}$" (**Fig. 1**). The first parameter is called *action parameter* (AP) and is defined with its characteristic to be able to "tend towards two opposite values" and to "have an impact on one or more other parameters". Moreover, "the designers have the possibility of modifying them". The two other parameters in a contradiction are called *evaluation parameters* (EP) and are defined with their capacity to "evolve under the influence of one or more action parameters", thus making it possible to "evaluate the positive aspect of a choice made by the designer".

$$AP \quad \begin{matrix} \\ Va \\ \overline{Va} \end{matrix} \quad \begin{matrix} EP_1 & EP_2 \\ \end{matrix} \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}$$

Fig. 1. Possible representation model of contradictions [7]

Knowledge of how the contradiction must be formulated helps for the knowledge extraction part as well as the graph creation, because the elements to be extracted from the documentation are in priority : problems, partial solutions, evaluation and action parameters as well as their possible values *Va* and $\overline{Va}$. As of problems and partials solutions, research on IDM contains clear definitions, both in the form (syntax and graphical representation) and in its content [8]. A problem (**Fig. 2**) "describes a situation where an obstacle prevents a progress, an advance or the achievement of what has to be done". A partial solution (**Fig. 3**) "expresses a result that is known in the domain and verified by experience.". Cavallucci et al. give more precision about the concept of partial solution [8] :

"It may materialize a tacit or explicit knowledge of one or more members of the design team upon their past experience, a patent filled by the company or a competitor or any partial solution known in the field of competence of the members of the design team. We wish also to remind that a partial solution is supposed to bring the least possible uncertainty about the assertions of its effects on the associated problem. Confusion can appear between a ''solution concept'' (which is the result of an assumption made by a member) and a ''partial solution'', which has been validated by experience, tests, calculations or results known and verified."

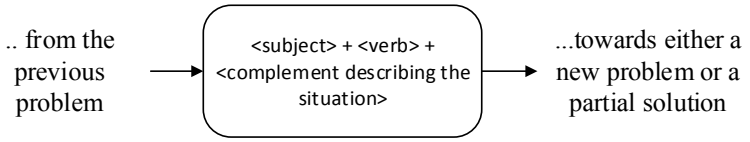

Fig. 2. Graphical representation of a problem

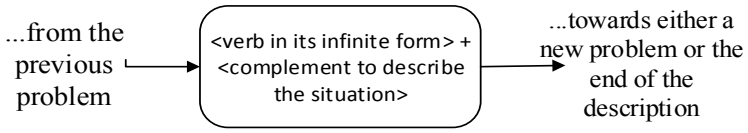

Fig. 3. Graphical representation of a partial solution

2.2 Patent-knowledge extraction

An important principle both for TRIZ and for IDM is that inventive solutions are generally found with analogy, and therefore thanks to solutions or tools which belong to another domain. Patents are documents with very rich technical content, which is generally not to be found elsewhere, like for instance in scientific papers [4]. In order to save engineers a lot of time and help them getting to these analogies, a tool has been developed by Souili et al., which automatically creates a graph of problems, partial solutions and parameters out of English-language patents (**Fig. 4**), selected by a user in a large database [5]. This corresponds to the first phase of the Inventive Design Method, which has been automated thanks to Natural Language Processing tools.

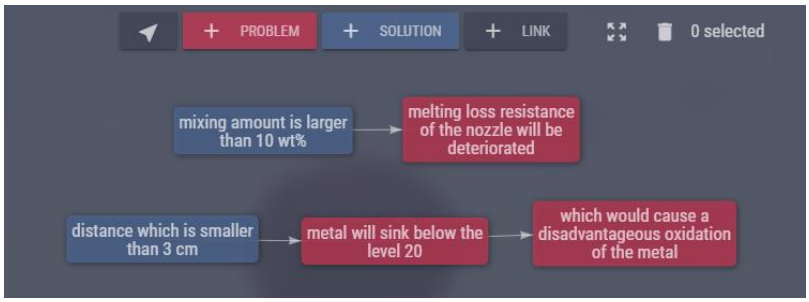

Fig. 4. An example of a problem-solution graph

2.3 Scientific research papers and science news articles

Scientific research papers. Many studies have been conducted on the automated information extraction from scientific research papers. While some of them are strongly related to their domain or a specific application (like biology) [9–13], other present, for general purposes, the extraction of general features of articles such as keywords or key phrases [13–16]. However, scientific research papers have not yet been the topic of research for TRIZ or IDM applications.

One of the main challenges for the adaptation of the patent-knowledge extraction tool is to understand and characterize the differences between the different types of documents: patents, scientific research papers and science news articles.

While all patents have a very similar structure with four sections, and despite the higher degree of liberty in the description section, research papers have much more differences amongst themselves. For instance, title names, even if they often indicate

the same concepts, are not formulated the same way. Moreover, the number and topics of the sections vary a lot, rendering text-mining techniques more difficult to apply.

However, many researches have been conducted on the structure and content of research papers. Pontille, based on a previous research by Bazerman [17], describes a generic structure format named IMRAD (Introduction, Material and methods, Results And Discussion) [18] as a stabile tool for professional practice harmonization; "a particular form of proof expression"[1] and "a standardized argumentative matrix"[2] which has become an official norm with the *Standard for the preparation of scientific papers for written or oral presentation* in 1979 in the USA.

According to Pontille, authors generally define a problem and propose hypotheses in the introduction. Therefore, it will be interesting for our research to study how the main problem is introduced in this section, in order ameliorate its extraction.

From the section about "Material and methods", which explains "the way the study has been conducted"[3] and has become central for the argumentation in the last decades [18], we won't focus on intermediate problems and solutions for our research, since they are often criticized during the article. However, Pontille mentions the occasional presence of a subsection for the description of "variables", which we can link to the IDM-ontology concept of parameter. Therefore, it could be interesting to extract the main parameters and to find a way to know, with the help of their limited context, whether they are evaluation or action parameters, i.e. whether they have an impact on the final output and can be modified, or whether they are central to the evaluation of the output.

The discussion section, as described by Pontille, is also interesting for our extraction tool. By "identifying the practical and theoretical implications of their analysis"[4] and "opening perspectives for futures researches"[5] [18], we could find a repeated version of the main problem, hopefully the main partial solution of the article, and potential unresolved new problems. Moreover, since this section is the interpretation of the results evaluation, it will be interesting to look particularly for evaluation parameters here.

Considering the content, Gosden study on rhetorical themes and more specifically on contextualizing frames throughout research articles [19], with an analysis of their functions and even some links to problems and solutions, is very interesting for the extraction of features which will help us in our information extraction task.

The abstract is also a key part of a research article. According to a study conducted on articles in the field of linguistics and educational technology, three moves seem to be unavoidable in research articles : "Presenting the research", "Describing the methodology" and "Summarizing the findings" [20]. These move are very close to actual

---

1 Our translation. The original sentence is: « Le format IMRAD constitue ainsi une forme particulière d'expression de la preuve » (p.2)

2 Our translation. The original phrase is: « une matrice argumentative standardisée » (p.6)

3 Our translation. The original phrase is : « explicitation de la manière dont l'étude a été conduite » (p.6)

4 Our translation. The original quote is : « ils identifient les implications pratiques et théoriques de leur analyse » (p.7)

5 Our translation. The original quote is : « ouvrir des perspectives de futures recherches » (p.7)

sections of the IMRAD structure, and since it is a summary of the research, the presence of the major problem, partial solution and parameters seem very probable.

Science news articles. Science news articles are not a famous research topic, but media coverage of science is [21–27]. An increasing amount of science-related content is published regularly on the internet, in very diverse forms: online versions of traditional print newspapers and magazines, science section of online newspapers, science blog articles from actual researchers, videos, specialized science news websites, etc. While traditional online sources are declining in terms of audience rates, nontraditional online sources such as blogs and "other online-only media sources for information about specific scientific issues" now get much more attraction for the public [21]. Science news websites therefore seem to be interesting sources for condensed information, with regularly new publications in multiple research fields.

This kind of source is nevertheless contested in the literature because of the trustworthiness of its contents [22]. Since the targeted audience is interested in scientific innovation and breakthrough, but does not have a large scientific background, information is often summarized and the risk is that is diverts from the intention of the original author. In order to prevent engineers from using erroneous extracted data during the IDM process, we will have to compare the results of the extraction from the science news article on the one hand, and from the full original document on the other hand.

3 Methodology

The information extraction tool we aim to create is mainly an adaptation of the existing tool used for the extraction related to IDM ontology. This adaptation requires a certain number of steps, from building a corpus to the evaluation of the final program. As mentioned above, the major challenge for this project is to adapt the features needed to select the candidate sentences for problem or partial solution classification.

3.1 The corpus creation

Annotated corpus. The first step for adapting the extraction tool is to build an annotated corpus that will be used for classification, in order to extract the main features (words or phrases) for each category – problem, partial solution, evaluation parameters, action parameters, values.

Since research articles and science news articles have a very different structure, they should not be treated the same way, but preferably with two separated corpora. Moreover, there are multiple sources publishing articles in diverse fields, with differences in the global structure as well as the length and style of the articles. Therefore, each corpus must contain articles form different sources.

We chose to give the annotation task to 44 students enrolled in a course on Inventive Design Method. With this condition, we decided to build both corpora that corresponds to the level and number of students who will annotate it.

For the annotation task, all articles have been cleaned and transformed into PDF when necessary for reading comfort and homogeneity of annotation procedure, at least for the science news articles that were in html format in the beginning

The corpus of science news articles contains 44 articles from the 7 following sources:

- Machine Design [28] – 6 articles, about 1-2 pages long
- New Atlas [29] – 7 articles, about 1 full page long
- Phys.org [30] – 7 articles, about 1-2 pages long
- Research & Development [31] – 6 articles, about 1 page long
- Science Daily [32] – 6 articles, about 1 full page long
- Science News [33] – 6 articles, about 1-2 pages long
- Science News for Students [34] – 6 articles, about 2 pages long
- The corpus of scientific research articles contains 44 articles from the 4 following sources:
- Accounts of chemical research [35] – 15 articles, about 8-10 pages long
- Annual Review of Condensed Matter Physics [36] – 8 articles, between 21-28 pages long
- Chemistry of Materials [37] – 11 articles, between 6-12 pages long
- Proceedings of the National Academy of Sciences of the USA (PNAS) [38] – 10 articles, about 6 pages long

The annotation procedure is done on the PDF-formatted files with the highlighting tool of Adobe Acrobat, following a given color legend for each element: problem, solution, evaluation parameter, action parameter, value. After collecting all annotations, evaluating and modifying them when necessary, we plan to use Sumnotes, a web application [39], to extract a list for all sequences corresponding to each element.

Clean corpus for the extraction. For the extraction of the different elements from the IDM ontology, we cleaned and transformed the articles into JSON format. Since most of the articles were available in HTML format, both for science news articles and scientific research papers, we used the BeautifulSoup library for Python 3 [40], which is a very efficient library for cleaning data from HTML and XML documents.

However, we are still looking for a way to efficiently clean PDF articles from PNAS, because the GROBID service [41] we used for transforming them into TEI-XML documents mixed up all the sections, therefore making our information extraction task impossible. In the meantime, we had to drop this source for the features extraction, implementation and evaluation steps.

3.2 The extraction and selection of features

The adaptation of the features needed for the sentences, keywords or phrases selection and classification tasks involve three steps, and its goal is to create lists and dictionaries of words, phrases or n-grams that makes the extraction efficient and specific to both scientific research articles and science news articles.

The three steps are:

– Patent features assessment
– Specific features extraction
– Features selection

Evaluation method. The whole features adaptation process involves multiple evaluations which have to be undertaken with the same parameters. The main purpose for the extraction program is to get sufficient results for a further use in the inventive design process following IDM-TRIZ. All the evaluations have to be undertaken with the basis of a single Gold Standard for each element category (e.g. parameters or problems), built from the annotated corpus. The program has a two-step classification process for problems and partial solutions. The first step extracts candidate sentences and the second step confirms or rejects the first choice. The rejected sentences are then classified as "neutral". For a complete evaluation, we will also consider those sentences and add them to the "concepts" category (thus containing all extracted problems, solutions and neutrals). The evaluation criteria are:

- The number of concepts, parameters, partial solutions, neutrals and problems extracted in average per article
- The precision for each category, which is the ratio between the number of relevant elements found automatically in the category and the total number of elements automatically retrieved in the category.
- The recall for each category, which is the ratio between the number of relevant elements retrieved automatically in the category and those found manually in the Gold Standard for this category.
- The rate of misclassification, i.e. the proportion of sentences that should have been classified in another category but is considered useful. We consider as "useful" neutrals, sentences that are really interesting for the research or its future developments, but does not belong to the category of problems or partial solutions.

Patent features assessment. The first step is the assessment of the features from the original program that is made for patents. With a first minimal adaptation of the original program, keeping all the features intact, we assess the features globally, and then individually. The global assessment will be made by evaluating the whole program following the procedure described above. This assessment will serve as a comparison point for the individual assessment that follows, for which the program will run on a loop and evaluate its results, each time ignoring a different feature. This makes it possible to remove the features which don't have any positive impact on the final result. The removing task, however, considering the relatively small number of articles in our corpus, should be done manually in order to maintain the possibility of keeping some features for which we estimate a potential impact with a bigger corpus.

Specific features extraction. The second step is the extraction of new features. This task will involve several tools for advanced Natural Language Processing (NLP) which are still to be defined with testing, as well as a Machine Learning classification task on

the basis of our annotated corpus, using Weka [42], which will also make a ranking of features that helped him guess the results correctly.

Features selection. The third task will be to select the final features from the remaining patent features and the previously extracted new features, then to implement them into the main classification program, and again make an assessment of these features with the same methodology as for the patent features assessment. Depending on the results, it could be necessary to make slight changes, and run a new evaluation again, in order to get the best evaluation results as possible.

4 Conclusion

The automatic extraction of concepts related to IDM ontology in scientific articles and science news articles is mostly an adaptation of the existing tool working with patents. However, keeping the same features as in the original program does not make sense, since the structure, length and style of the three kinds of articles are very different. Therefore, we hope for good results while using the methodology presented in this article, i.e. creating and annotating two corpora and a golden standard, evaluating the features used in the original program, extracting useful new features and implementing the adjusted list of selected features in the program, with potential fine-tuning.

References

1. World Intellectual Property Organization: World intellectual property indicators. WIPO, Geneva (2017)
2. Publish or perish? The rise of the fractional author… - Research Trends, https://www.researchtrends.com/issue-38-september-2014/publish-or-perish-the-rise-of-the-fractional-author/
3. Altshuller, G.: And Suddenly the Inventor Appeared: TRIZ, the Theory of Inventive Problem Solving. Technical Innovation Center, Inc. (1996)
4. Bonino, D., Ciaramella, A., Corno, F.: Review of the state-of-the-art in patent information and forthcoming evolutions in intelligent patent informatics. World Pat. Inf. 32, 30–38 (2010). doi:10.1016/j.wpi.2009.05.008
5. Cavallucci, D.: The theory of inventive problem solving: current research and trends in French academic institutions. Springer Berlin Heidelberg, New York, NY (2017)
6. Cavallucci, D., Strasbourg, I.: From TRIZ to Inventive Design Method (IDM): towards a formalization of Inventive Practices in R&D Departments. 2 (2012)
7. Rousselot, F., Zanni-Merk, C., Cavallucci, D.: Towards a formal definition of contradiction in inventive design. Comput. Ind. 63, 231–242 (2012). doi:10.1016/j.compind.2012.01.001

8. Cavallucci, D., Rousselot, F., Zanni, C.: Initial situation analysis through problem graph. CIRP J. Manuf. Sci. Technol. 2, 310–317 (2010). doi:10.1016/j.cirpj.2010.07.004
9. Andrade, M.A., Valencia, A.: Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families. Bioinformatics. 14, 600–607 (1998). doi:10.1093/bioinformatics/14.7.600
10. Krallinger, M., Valencia, A., Hirschman, L.: Linking genes to literature: text mining, information extraction, and retrieval applications for biology. Genome Biol. 9, S8 (2008). doi:10.1186/gb-2008-9-s2-s8
11. Müller, H.-M., Kenny, E.E., Sternberg, P.W.: Textpresso: An Ontology-Based Information Retrieval and Extraction System for Biological Literature. PLOS Biol. 2, e309 (2004). doi:10.1371/journal.pbio.0020309
12. Yakushiji, A., Tateisi, Y., Miyao, Y., Tsujii, J.: Event extraction from biomedical papers using a full parser. In: Biocomputing 2001. p. 408–419. WORLD SCIENTIFIC (2000)
13. Gelbukh, A., LINK (Online service) éd: Computational linguistics and intelligent text processing: 6th international conference, CICLing 2005, Mexico City, Mexico, February 13-19, 2005: proceedings. Springer, Berlin (2005)
14. Lopez, P., Romary, L.: HUMB: Automatic key term extraction from scientific articles in GROBID. 4
15. Krapivin, M., Autaeu, A., Marchese, M.: Large Dataset for Keyphrases Extraction. University of Trento (2009)
16. Kim, S.N., Medelyan, O., Kan, M.-Y., Baldwin, T.: SemEval-2010 task 5: Automatic keyphrase extraction from scientific articles. 6
17. Bazerman, C.: Shaping Written Knowledge : the genre and activity of the experimental article in science. , Madison, Wisconsin (1988)
18. Pontille, D.: Matérialité des écrits scientifiques et travail de frontières: le cas du format IMRAD. 16
19. Gosden, H.: Discourse functions of marked theme in scientific research articles. Engl. Specif. Purp. 11, 207–224 (1992). doi:10.1016/S0889-4906(05)80010-9
20. Phuong Dzung Pho: Research article abstracts in applied linguistics and educational technology: a study of linguistic realizations of rhetorical structure and authorial stance. Discourse Stud. 10, 231–250 (2008). doi:10.1177/1461445607087010
21. Brossard, D.: New media landscapes and the science information consumer. Proc. Natl. Acad. Sci. 110, 14096–14101 (2013). doi:10.1073/pnas.1212744110
22. Brumfiel, G.: Supplanting the old media, (2009)
23. Puschmann, C.: (Micro)Blogging Science? Notes on Potentials and Constraints of New Forms of Scholarly Communication. In: Bartling, S. et Friesike, S. (éd.) Opening Science. p. 89–106. Springer International Publishing, Cham (2014)
24. Allgaier, J., Dunwoody, S., Brossard, D., Lo, Y.-Y., Peters, H.P.: Journalism and Social Media as Means of Observing the Contexts of Science. BioScience. 63, 284–287 (2013). doi:10.1525/bio.2013.63.4.8

25. Minol, K., Spelsberg, G., Schulte, E., Morris, N.: Portals, blogs and co.: the role of the Internet as a medium of science communication. Biotechnol. J. 2, 1129–1140 (2007). doi:10.1002/biot.200700163
26. Mahrt, M., Puschmann, C.: Science blogging: an exploratory study of motives, styles, and audience reactions. 17
27. Brossard, D., Scheufele, D.A.: Science, New Media, and the Public. Science. 339, 40–41 (2013). doi:10.1126/science.1232329
28. Machine Design, http://www.machinedesign.com/
29. New Atlas, https://newatlas.com/
30. Phys.org - News and Articles on Science and Technology, https://phys.org/
31. Research & Development, https://www.rdmag.com/
32. ScienceDaily: Your source for the latest research news, https://www.sciencedaily.com
33. Science News, https://www.sciencenews.org/
34. Science News for Students | News and feature articles from all fields of science, https://www.sciencenewsforstudents.org/home
35. Accounts of Chemical Research (ACS Publications), https://pubs.acs.org/journal/achre4
36. Annual Review of Condensed Matter Physics | Home, https://www.annualreviews.org/journal/conmatphys
37. Chemistry of Materials (ACS Publications), https://pubs.acs.org/journal/cmatex
38. PNAS, http://www.pnas.org/
39. FiftForce: Sumnotes - Summarize PDF Annotations, https://www.sumnotes.net/
40. Beautiful Soup Documentation — Beautiful Soup 4.4.0 documentation, https://www.crummy.com/software/BeautifulSoup/bs4/doc/
41. Lopez, P.: grobid: A machine learning software for extracting information from scholarly documents. (2018)
42. Frank, E., Hall, M.A., Witten, I.H., Pal, C.J.: The WEKA Workbench. Online Appendix for « Data Mining: Practical Machine Learning Tools and Techniques ». Morgan Kaufmann (2016)