J. Astrophys. Astr. (2019) 40:23
https://doi.org/10.1007/s12036-019-9590-5

Knowledge discovery through text-based similarity searches for astronomy literature

WOLFGANG E. KERZENDORF[1,2]

[1]Center for Cosmology and Particle Physics, New York University, 726 Broadway, New York, NY 10003, USA.
[2]European Southern Observatory, Karl-Schwarzschild-Strasse 2, 85748 Garching, Germany.
E-mail: wkerzendorf@gmail.com

MS received 4 September 2018; accepted 14 May 2019; published online 13 June 2019

Abstract. The increase in the number of researchers coupled with the ease of publishing and distribution of scientific papers (due to technological advancements) has resulted in a dramatic increase in astronomy literature. This has likely led to the predicament that the body of the literature is too large for traditional human consumption and that related and crucial knowledge is not discovered by researchers. In addition to the increased production of astronomical literature, recent decades have also brought several advancements in computational linguistics. Especially, the machine-aided processing of literature dissemination might make it possible to convert this stream of papers into a coherent knowledge set. In this paper, we present the application of computational linguistics techniques to astronomy literature. In particular, we developed a tool that will find similar articles purely based on text content f rom an input paper. We find that our technique performs robustly in comparison with other tools recommending articles given a reference paper (known as recommender system). Our novel tool shows great power in combining computational linguistics with astronomy literature and suggests that additional research in this endeavor will likely produce even better tools that will help researchers cope with vast amounts of knowledge being produced.

Keywords. Natural language processing—methods: statistical.

1. Introduction

Since the inception of writing, human knowledge has steadily increased as has the number and size of published works. The output of the scientific community has doubled every nine years over the past decades (Bornmann & Mutz 2015).

The computing and internet revolution has made publication and dissemination of these works easy and with the advent of open access channels pluralistic. The public repository *arXiv* has provided open access to almost the entire corpus of publications since 1992 in the physical sciences .

Given the rise of publications each year and the fixed capacity of a human to process information, we shall narrow the specialization range in each field to limit the breadth of the necessary knowledge base or have new tools that filter the available publications.

In astronomy, the NASA Astrophysics Data System (ADS) (Kurtz *et al.* 2000) has provided access (in addition to many other accomplishments such as digitizing old articles) to this large amount of literature with a search interface that captures the traditional way of accessing information (name of first author and year) extremely well. Newer iterations of this system (Chyla *et al.* 2015) have started to branch out and allow not only search algorithms but also provide certain bibliometric statistics as well as a recommender system (named ‘Suggested Articles’). This recommender system is based on citations, text similarity, and co-readership (as described on the ADS 2.0 website and suggested in Henneken & Kurtz 2010; Kurtz 2011). Such recommender systems will be the first step to tackle a world in which the scientific literature has massively outgrown the memory capacity of human brains.

In this paper, we present a new method for article recommendations starting from a reference article or text. We employ the techniques of text similarity and specifically avoid citations. This strict abstention from citation was chosen due to the fact that citations are influenced by many factors and may not provide an unbiased link between publications (several examples in van Wesel *et al.* 2014). A web service, based on the presented tools and techniques, can be found at http://deepthought.space/deepthought.

In Section 2, we describe the data acquisition, initial vetting and processing. Section 3 describes the method used and some statistics. An overview of the framework used in this work and its application to several example papers is given in Section 4. Caveats and possible improvements are discussed in Section 5 and we conclude with an outlook to the future in Section 6.

2. Data processing

For our initial raw corpus, we considered all papers submitted to *arXiv*. Using the bulk data access,[1] we downloaded the entire corpus. After a series of operations (discarding any non-latex submissions), we arrived at individual source directories (a total of 13,01,668). This work focuses currently on the field of astronomy. For all entries, we harvested the metadata through the OAI protocol for metadata harvesting (OAI-PMH) and then selected all *arXiv* papers where one of the posting categories was ‘astro-ph’. This amounts to a total corpus size of 2,32,680 papers (including any cross-listings).

In each source compilations, we tried to identify the main tex file by requiring a single valid \begin{*document*} clause and processed this further with LATEXPAND[2] to a single document that would contain all relevant text content. Not all entries had a uniquely identifiable main *tex-file* (removing ≈5000 papers).

The resulting *tex-files* were further processed by removing the most common environments:

- Figures
  figure, *picture*
- Tables
  table, *deluxetable*
- Equations
  equation, *align*, *subequations*, *eqnarray*, *array*, *matrix*

Then we removed any text before the first section command or if this was not present any text before end abstract.

The final step of the raw reduction process was the removal of latex commands using the OPENDETEX software.[3]

2.1 *Natural language processing*

These raw texts are ready for natural language processing and the following steps use the Natural Language ToolKit (NLTK; Bird *et al.* 2009) tools extensively.

The first step in this process is to break the text into individual words using NLTK.TOKENIZE_WORDS, which splits into individual words and removes all punctuations—except the period (which is treated as a single word at this step).

The next process is to remove stop words such as *these*, *those*, *am*, *is*, *are* (using the English stop words defined in NLTK.CORPUS.STOPWORDS).

The final step of processing is to lemmatize the words, which is the process of grouping together different inflected forms of a word so it can be analysed as a single item. For this process, we use the tool NLTK.WORDNET.MORPHY to bring the words back to their original forms (galaxies maps to galaxy, expanding maps to expand, etc.). We discard words that do not have a corresponding entry in the dictionary provided (WORDNET; Fellbaum 1998).

We ran through these final paper products and removed any manuscript that had less than ten words left (caused by un-closed latex environments that lead to removal of large parts of the paper). After this final step, we are left with a corpus consisting of 2,01,997 papers.

The whole process takes several hours on a server with Intel(R) Xeon(R) CPU E7-4850 with 96 cores.

3. Method

In this work, we will rely entirely on the bag-of-words technique (Harris 1954), that disregards grammar and word order, treating the document just as a collection of words. This is a useful technique for corpora that consists of manuscripts that contain several hundreds to thousands of words (Hinrich Schtze, private communication). The features that we use for our analysis are several statistics based on word frequency. For all feature extraction tasks, we relied on SCIKIT- LEARN (Pedregosa *et al.* 2011).

The first step for any of these methods is building a vocabulary of unique words. This is helped by the

[1] e.g. s3://arxiv/src/arXiv_src_1001_001.tar.

[2] https://www.ctan.org/pkg/latexpand.

[3] https://github.com/pkubowicz/opendetex.

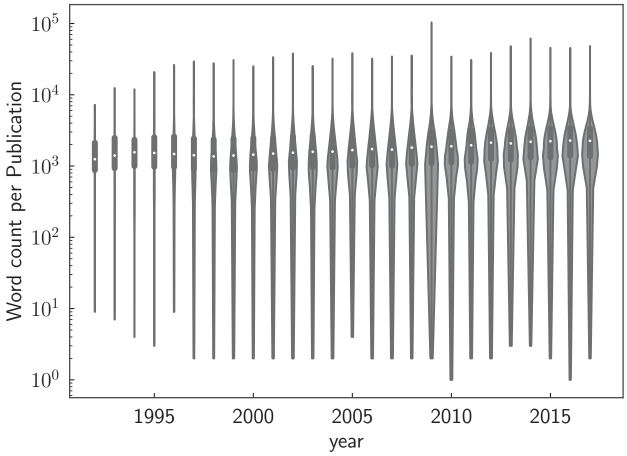

Figure 1. Distributions of word counts without stop words for all papers published (to astro-ph) each year.

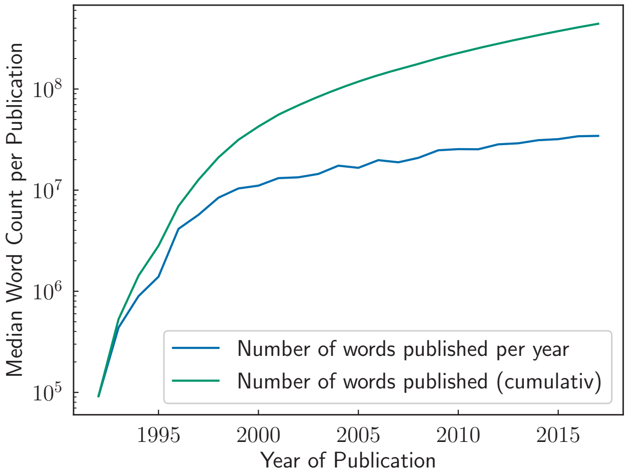

Figure 2. Total number of words without stop words for all papers published (to astro-ph) each year.

fact that we transformed our document by removing any stop words and transforming the words to their simplest form.

This vocabulary consists of ≈32,000 words. This is only slightly larger than the ≈20,000 (Goulden *et al.* 1990) words a well educated native speaker knows and much lower than the ≈1,70,000 words in the Oxford English dictionary (Simpson & Weiner 1989). Our vocabulary can then be used to vectorize (feature extract) the documents to vectors the size of the dictionary (in our case ≈32,000).

The simplest case is the use of a binary statistics for feature extraction which will only encode if a word is present or not present. This statistics can give a rough overview of the content but will de-weight more frequently used words and thus possible to shift the inferred topic of the document in statistical analysis.

The first statistics we have performed on the corpus of documents is a simple count vectorization (using SCIKIT- LEARN.FEATURE_EXTRACTION.COUNT VECTORIZER). This count measure allows us to quantify the growth of literature since the conception of *arXiv*. Figure 1 shows that while the growth in document size (number of words using the processed document word counts as a proxy; no distinction being made between papers and reviews however) has been maybe a factor of 1.5 over the last decade, astronomical literature in total has grown exponentially (see Figure 2) over the same period. This given measure of 'word count', however, has several drawbacks, the most important one being that it increases with document length and thus does not give a useful measure for word importance.

The last vectorization technique is a natural extension of the word counts that aims at emphasizing a word's importance in a text—thus ideal for assessing the content of a paper. In the following, we will use the moniker 'term' and 'word' interchangeably as our analysis uses only one-word terms (unigrams). The term frequency $tf(t, d)$ method normalizes the simple word count by the number of words in the document. This relies on the assumption that the importance (or weight) of a term in documents is proportional to this term frequency (Luhn 1957). In addition to term frequency, we want to quantify the information content a specific term carries. Sparck Jones (1972) have introduced the concept of inverse document frequency $idf(t, d)$. We use the inverse document frequency given in SCIKIT-LEARN as $\log \frac{1+n_d}{1+df(d,t)}$, where n_d is the total number of documents and $df(d, t)$ is the number of documents containing the given term. The combination of both measures gives the well established $tf(t, d) \times idf(t, d)$ (henceforth TFiDF) measure which weigh terms highly that have a high information content due to their rarity. This measure is used in several machine learning tasks including finding similar texts. For an in-depth discussion of TFiDF methods, please refer to (Baeza-Yates *et al.* 2011; Manning *et al.* 2008).

We perform the calculation from the processed and loaded documents to the TFiDF matrix in around seven minutes on a server with a Intel(R) Xeon(R) CPU E7-4850 with 96 cores.

4. Similarity in papers

We study recommender systems for knowledge discovery in this work. Such recommender framework can suggest articles for further reading given a certain publication system. There are two classes of recommender systems: (1) user-based systems that recommend articles that similar users have accessed; (2) item-based recommender systems that recommend similar items. We will focus on an item-based recommender system in our work as this data is available to us. This recommendation system has the advantage that it purely focuses on the content of the paper and not on which kind of users view this content. We, however, do not explore the detailed differences between user-based and item-based systems in this work (a more user based approach can be reviewed at Krstovski *et al.* 2016).

We will evaluate our recommender system by applying it to fields that the author and his colleagues are knowledgeable in. We have chosen three very different manuscripts to cover larger parts of the potential parameter space. The first one being the search for a surviving companion of a supernova, the second focussing on abundances in stellar populations, and the third studying proto-planetary disks.

We first normalize our document vectors using the Euclidean norm $\vec{d}_{\text{norm}} = \frac{\vec{d}}{||\vec{d}||_2}$ before proceeding further, leaving us with the entire sparse matrix (which is available upon request). Here, we present an example of such a matrix (with different document vectors as rows and columns representing different words/terms):

$$A_{\text{TFiDF}} = \begin{array}{c} \\ \text{arXiv} - 1 \\ \text{arXiv} - 2 \\ \vdots \\ \text{arXiv} - n \end{array} \begin{array}{c} \begin{array}{cccc} \text{star} & \text{model} & \cdots & \text{galaxy} \end{array} \\ \begin{pmatrix} 0.021 & \cdots & 0 & \\ 0 & 0.03 & \cdots & 0 \\ 0.019 & 0.016 & \cdots & 0 \\ 0 & 0 & \cdots & 0.023 \end{pmatrix} \end{array}. \tag{1}$$

We simply use the cosine distance by choosing a document that we want to compare and multiply this with the TFiDF matrix $\vec{v}_{\text{similarity}} = A_{\text{TFiDF}} \times \vec{d}_{\text{norm}}$ to measure text similarity. An example that showcases this technique is available at http://deepthought.space/deepthought.

4.1 *Example: SN 1006 companion search*

Here, we will use this approach on a paper that is well-known to the author: 'Hunting for the Progenitor of SN 1006: High-resolution Spectroscopic Search with the FLAMES Instrument' (Kerzendorf *et al.* 2012). This paper describes the failed attempt to find a surviving companion star (often called the donor star) to a supernova (likely caused by a white dwarf), searching in one of the supernova remnants in our galaxy.

We first measure the most important words with the presented algorithm. For this purpose, we inverse sort our $\vec{v}_{\text{similarity}}$ and use the first best 100 matches. We multiply the $\vec{d}_{\text{norm}}$ and look for the highest entries in the resulting vector which should give us the most important words that the algorithm matches. Figure 4 shows the relevant words that one would expect while writing a paper about searching for a donor star in a supernova remnant likely caused by a white dwarf.

In the next step, we compare the TFiDF similarity by choosing other metrics in which the papers are judged similarly. In this case, we choose citations to a paper, references in a paper, and the ADS suggestions given when displaying a certain paper.

All the references in our test paper (12 in total found in our *arXiv* corpus) have a median similarity of 0.38. Only 3% of papers (see Figure 3) in the astronomy corpus are more similar than this median which suggests that as expected the citations are highly relevant.

All the citations in our test paper (30 in total were found in our *arXiv* corpus) have a median similarity of 0.26 and 5% of papers were more similar than this, suggesting that the citations to this article come from a more varied field than the original references (or that references had been forgotten).

Next, we test if the algorithm's top 30 papers (similar to the citations) are neither citations nor references and compare this to the relevance of the other papers. Figure 4 shows the comparison between the median cosine distance of the cited papers, the median cosine distance between the references and the median top 30 results excluding the papers from both these groups.

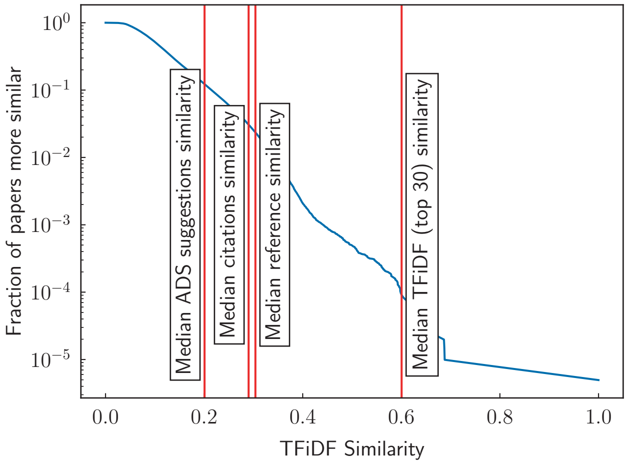

Figure 3. Comparing cumulative similarity to the given TFiDF similarity number. For example, papers that have higher similarity than the median similarity of the references only make up 5% of the entire corpus of papers.

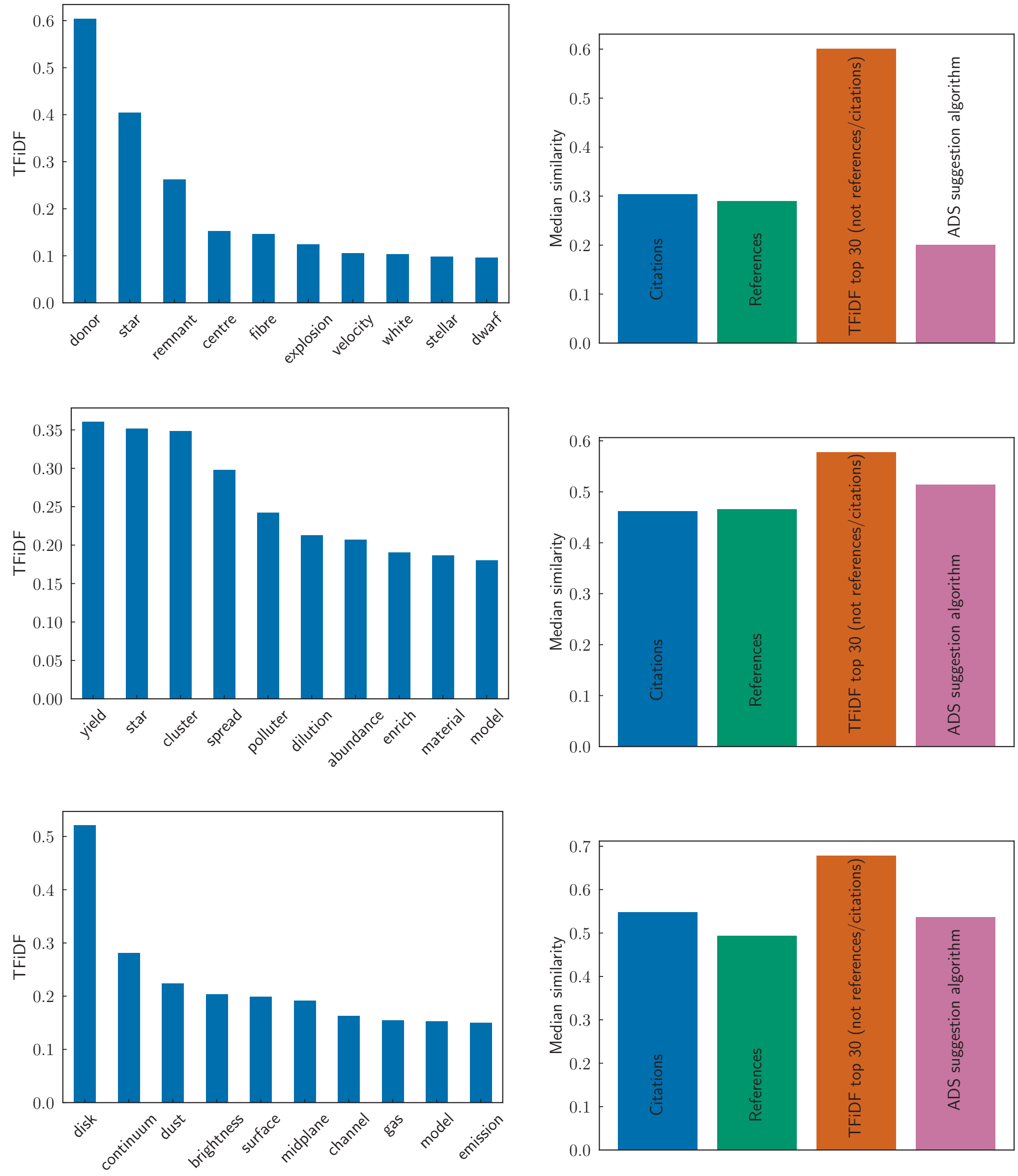

Figure 4. The TFiDF method applied to three distinct papers (*top*) 'Hunting for the Progenitor of SN 1006: High-Resolution Spectroscopic Search with the FLAMES Instrument' (Kerzendorf *et al.* 2012), (*middle*) 'A General Abundance Problem for All Self-Enrichment Scenarios for the Origin of Multiple Populations in Globular Clusters' (Bastian *et al.* 2015) and (*bottom*) 'Unveiling the Gas-and-Dust Disk Structure in HD 163296 using ALMA Observations' (de Gregorio-Monsalvo *et al.* 2013). The left plots show the TFiDF weight for the ten most words with the highest weights while the right plots show the similarity metric applied to several collections associated with these papers.

This demonstrates that such a system can find relevant papers that could easily be missed otherwise.

ADS has also implemented a recommender system (Henneken & Kurtz 2010; Chyla *et al.* 2015). We find that it does not give as relevant matches as the presented algorithm from the astrophysics research perspective (at least for the papers mentioned in this article). There are 30% of papers that are more similar to the document in question compared to the median similarity of the suggested papers. Manual inspection also shows that some of the suggested papers (e.g. 'Maps of Dust Infrared Emission for Use in Estimation of Reddening and Cosmic Microwave Background Radiation Foregrounds' by Schlegel *et al.* 1998) might also be very broadly related but not very relevant. The ADS algorithm uses the citation matrix and user interaction so it is not possible with our current dataset to make a quantitative comparison.

4.2 *Example: Globular cluster*

The second test paper we use is ‘A General Abundance Problem for all Self-Enrichment Scenarios for the Origin of Multiple Populations in Globular Clusters’ by Bastian *et al.* (2015). This paper points out a possible flawed explanation for abundance anomalies in globular clusters. The top matching words (see Figure 4) are highly relevant to a topic that discusses anomalous abundances that might be caused by different pollution system to varying degrees due to their yields. The references are more similar compared to the first paper suggesting that this paper is more focused. The same is true for the citations that also come from more similar papers when compared to the citations in the first paper. The comparison of all median similarity numbers, however, shows a similar pattern when compared to the first paper including dissimilarity of the ADS suggestion algorithm (see Figure 4).

4.3 *Example: ALMA observations of a disk*

The last test paper is titled ‘Unveiling the Gas-and-Dust Disk Structure in HD 163296 using ALMA Observations’ (de Gregorio-Monsalvo *et al.* 2013). This paper describes observations of structure around young massive stars. The important words (see Figure 4) again seem highly relevant. Similar to the second paper, the citations and references have high similarity numbers that might suggest a narrower focus of this paper. In contrast to the other two example papers, in this case, the ADS suggestion algorithm also produced a high similarity index. Only three of the papers from the suggestion algorithm could be found from the *arXiv* corpus (compared to the usual six) which might explain this anomaly.

5. Discussion

This test of the use of natural language processing and machine learning tools NLTK, SCIKIT- LEARN has already shown that even simple techniques result in knowledge discovery. However, there are a number of improvements that can aid in the knowledge discovery part.

Specifically, in the language processing step, there are several steps that might be improved in future versions. We remove all words that are not in the English dictionary during our initial run. This already poses some problems in the lemmatization process as the name ‘Roche’ (as in Roche Lobe) is not recognized and is thus removed. This suggests that there is a need to build a domain-specific lemmatizer (such as the BIOLEMMATIZER for biology Liu *et al.* 2012). In our current approach, we also only consider single words (the so-called unigrams) but terms like ‘white dwarf’ (bigram) suggest that future iterations of this algorithm might find more relevant results if we treat such bigrams separately. Abbreviations are also commonly used in papers and are most often defined at the beginning of most papers. Thus the expanded word enters the word count only once. However, this leads to a misinterpretation of the true importance of the word as all other mentions are discarded.

The next information carrier that is removed are object names which might link papers that are of the same object. However, using this technique already values a certain type of knowledge above another (any study on the object is valued higher than similar studies on other objects for a given paper). This is especially true in our metric as object names will have a very low document frequency (being only mentioned in few papers) and thus will attain very high values in a TFiDF comparison which might not lead to the desired result.

6. Conclusion

We present a new technique for knowledge discovery by using a text similarity approach to find similar papers to a given reference paper. This technique performs robustly and finds relevant papers that are not discovered via citations, references or suggestions from ADS. This metric also seems to be a useful tool when studying if a paper is relevant to a broader field or addresses some detail in a narrower focus. Similar attempts in other fields (e.g. neuroscience; Achakulvisut *et al.* 2016) also suggest that this can be used to provide a powerful method to disseminate papers.

Currently, this allows an additional method to discover knowledge, especially when entering a field that one is unfamiliar with (e.g. using this technique for reviews). Our recommendation method might be further improved by linking our algorithm with citation information and using an algorithm like PAGERANK (popularized by Google; Page *et al.* 1999). This will value highly cited papers more than the lower cited ones. While this technique will help in knowledge discovery by finding relevant papers, our future goal is to identify key measurements and statements in each paper. This will allow a scientist to quickly sift through the vast amount of knowledge and identify the relevant paper by the searched quantities (e.g. the most current mass of the proton) before reading the entire paper and critically evaluating the methodology and statistics used.

Such a machinery would in the first instance help scientists to discover sought-after knowledge (regardless of bias towards certain authors, etc.) but might also allow for additional services. One of these might be the very simple 'fact checking' mechanisms that will aid researchers when compiling a paper by providing the most up-to-date quantities and flagging mistakes (similar to a grammar/spelling checker).

Such a machinery has uses far beyond astronomy and astrophysics. However, among the many academic fields, astronomy exposes the vast majority of papers and data in machine-readable formats (Christine L. Borgman, private communication). This suggests that this field is a good start for the development of such a machinery.

Acknowledgements

The author was supported by an ESO Fellowship and the Excellence Cluster Universe, Technische Universität München, Boltzmannstrasse 2, D-85748 Garching, Germany. He would also like to thank the detailed discussions, encouragement and suggestions from Felix Stoehr and Jason Spyromillio. The support from the library team (Uta Grothkopf, Dominic Bordelon and Silvia Meakins) was invaluable to get an insight into the field of knowledge discovery. Christine Borgman and Bernie Randles (at UCLA) gave suggestions from an Information Sciences point-of-view and the visit would not have been possible if not for the generousity of the UCLA Galactic Center Group (especially Tuan Do and Andrea Ghez). He also thanks Bruno Leibundgut, Kathatrina Immer and Ivan Cabrera-Ziri for testing the algorithm on some well-known papers. Finally, he would like to thank Hinrich Schütze for useful discussion of tools and techniques in the NLP field.

References

Achakulvisut T., Acuna D. E., Ruangrong T., Kording K. 2016, PLoS ONE, 11, e0158423, https://doi.org/10.1371/journal.pone.0158423

Baeza-Yates R., Ribeiro B. d. A. N. *et al*. 2011, Modern Information Retrieval, ACM Press, New York

Bastian N., Cabrera-Ziri I., Salaris M. 2015, MNRAS, 449, 3333, https://doi.org/10.1093/mnras/stv543

Bird S., Klein E., Loper, E. 2009, Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit, O'Reilly Media, Inc.

Bornmann L., Mutz R. 2015, Journal of the Association for Information Science and Technology, 66, 2215

Chyla R., Accomazzi A., Holachek A. *et al*. 2015, in Taylor A. R., & Rosolowsky E., eds., Astronomical Society of the Pacific Conference Series, Volume 495, Astronomical Data Analysis Software an Systems XXIV (ADASS XXIV), 401

de Gregorio-Monsalvo I., Ménard F., Dent W. *et al*. 2013, A&A, 557, A133, https://doi.org/10.1051/0004-6361/201321603

Fellbaum C. 1998, WordNet, Wiley Online Library

Goulden R., Nation P., Read J. 1990, Applied Linguistics, 11, 341

Harris Z. S. 1954, Word, 10, 146

Henneken E., Kurtz M. 2010, in APS March Meeting Abstracts

Kerzendorf W. E., Schmidt B. P., Laird J. B., Podsiadlowski P., Bessell M. S. 2012, ApJ, 759, 7, https://doi.org/10.1088/0004-637X/759/1/7

Krstovski K., Smith D. A., Kurtz M. J. 2016, arXiv e-prints, arXiv:1601.01611.

Kurtz M. J. 2011, Astrophysics and Space Science Proceedings, 24, 23, https://doi.org/10.1007/978-1-4419-8369-5_3

Kurtz M. J., Eichhorn G., Accomazzi A., *et al*. 2000, A&AS, 143, 41, https://doi.org/10.1051/aas:2000170

Liu H., Christiansen T., Baumgartner W. A., Verspoor K. 2012, Journal of Biomedical Semantics, 3, 3

Luhn H. P. 1957, IBM Journal of Research and Development, 1, 309, https://doi.org/10.1147/rd.14.0309

Manning C. D., Raghavan P., Schütze H. 2008, Introduction to Information Retrieval, 100, 2

Page L., Brin S., Motwani R., Winograd T. 1999, The PageRank citation ranking: Bringing order to the web., Tech. rep., Stanford InfoLab

Pedregosa F., Varoquaux G., Gramfort A., *et al*. 2011, Journal of Machine Learning Research, 12, 2825

Schlegel D. J., Finkbeiner D. P., Davis M. 1998, ApJ, 500, 525, https://doi.org/10.1086/305772

Simpson J., Weiner E. S. 1989, Clarendon Press, Oxford. Retrieved March, 6, 2008

Sparck Jones K. 1972, Journal of Documentation, 28, 11

van Wesel M., Wyatt S., ten Haaf J. 2014, Scientometrics, 98, 1601, https://doi.org/10.1007/s11192-013-1154-x