Programmatically Ranking and Sorting Research Articles for Reviews (ALIGN)

Maxim K. Gorshkov
School of Computer Science, Faculty of Science,
McGill University
Unit 406, 3622 Rue Durocher
Montreal, Quebec, Canada H2X2E8

Stella S. Daskalopoulou, MD, PhD
Department of Medicine, Faculty of Medicine,
McGill University
C2.101.4 1650 Ave Cedar
Montreal, Quebec, Canada H3G1A4

ABSTRACT
In an emerging trend to automate the world and daily interactions, academia is not exempt. Systematic reviews have been used in clinical research for decades and in practice, the process involves at least a double-blind sorting of articles in order to reach conclusions on a specific topic. With over 13 different values, a researcher will seldom need to consult other external resources to assess the quality of the paper. As such, ALIGN provides a nearly self-contained research application, which can be used to simplify and streamline the process of writing systematic reviews, while ensuring accuracy and quality. In general, by adding about 40-60 seconds of computing time per paper, researchers can begin to access objective measurements of the paper. This paper explores the different factors that go into evaluating a paper in general, the amalgamation of different resources to summarize useful information about the papers found in the search strategy, as well as the implications and limitations of the process on academia.

General Terms
Research Paper Quality Assessment, Data Mining, Open Access, Data Driven Research, Data-User Interaction.

Keywords
Research Paper Quality Assessment, Inter-Modal Data Mining, Open Access, Data Driven Research, Data-User Interaction Enhancement.

1. INTRODUCTION
Systematic reviews have come to play an important role in research, primarily in the health care field. Formal guidelines for the process of constructing these reviews have been set and refined since 1996. Most recently, the PRISMA (Preferred Reporting Items of Systematic reviews and Meta-Analyses) set of guidelines were established in 2009 and have become a way for researchers to retrieve up to date and relevant information, especially in an ever-changing technological age [1].

In practice, the process involves at least a double-blind sorting of articles in order to reach conclusions on publication eligibility on a specific topic. With many irrelevant articles, researchers need to invest significant amounts of time reading and assessing articles, which may end up being discarded. Although the implementation of data mining, amongst other methods of scraping information from the internet has only come to fruition in the last few years, there already exist many other institutional and private resources available online, which can provide more information about research articles.

Through various resources, articles can be ranked and pre-screened before they are put on deck to be read by the researchers. As such, irrelevant articles can be disregarded without the researchers reading these articles in full, and more time can be spent drawing conclusions from more relevant articles. The intention is to provide researchers with an access to higher quality information about each paper. With over 13 different values, a researcher will seldom need to consult other external resource to assess the quality of the paper. As such, ALIGN provides a nearly self-contained research application, which can be used to simplify and streamline the process of writing systematic reviews, while ensure accuracy and quality.

2. WORKFLOW
ALIGN is intended to work with the general protocol for Systematic Reviews. By feeding data from the search strategy into the EndNote, the proper file format can be exported from EndNote – XML (Extensible Markup Language). Ultimately, the researcher only needs to perform a few extra steps in order to have access to the detailed information about the papers. No complicated software is required. The workflow for including ALIGN in the review process is shown in *Figure 1*.

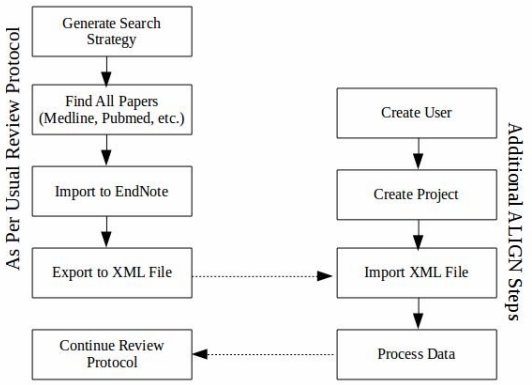


Figure 1: Overall workflow

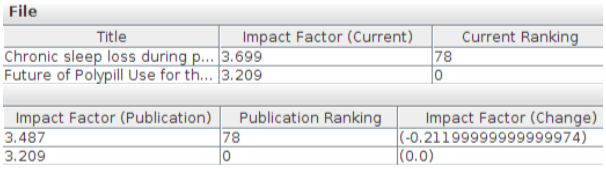
3. DATA RETRIEVED
From the sources currently available, these factors were integrated into the current version of ALIGN as they were deemed to provide information which would aid the researcher. The definition, source, and implementation of each are described in this section.

Throughout the paper, a running example shows real-time values returned for each section for *Independent Risk Factors for Atrial Fibrillation in a Population-Based Cohort* [2] in a shaded table such as this.

3.1 Impact Factor

The impact factor (IF) is a measure of the frequency with which the "average article" in a journal has been cited in a particular year or period [3]. In practice, a single, discrete, value represents the last 2 years of a particular journal. As such, it is important to consider the IF at the time of publication, the most recent value, as well as to track the change in these values. The values for the IF were taken from the Web of Science JCR database [4]. Values in the database total 102,391 entries and cover most journals between 1997 through 2012. At running time, these values are accessed from the internal database.

The IF on its own can only guide the researcher to an extent. In order to understand the quality of the journal within the context of the quality of the research, the IF needs to be compared amongst those in the specific field. By using the reverse search of SJR [5], the journal is ranked for both the year that it was published and for the current rank of the journal. *Figure 2* shows the 5 fields returned related to the impact factor of the paper, *Table 1* shows all 5 results for the example paper.


File

Title	Impact Factor (Current)	Current Ranking
Chronic sleep loss during p...	3.699	78
Future of Polypill Use for th...	3.209	0

Impact Factor (Publication)	Publication Ranking	Impact Factor (Change)
3.487	78	(-0.21199999999999974)
3.209	0	(0.0)

Figure 2: IF results

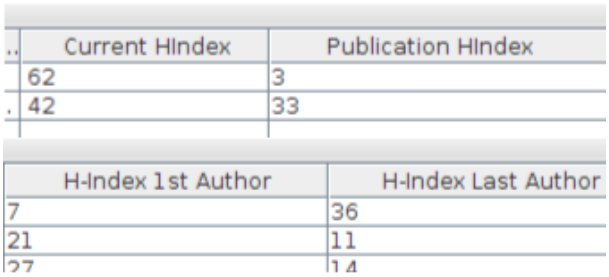
Table 1: Results pertaining to IF for example

Impact Factor (Current)	30.387
Current Ranking	51
Impact Factor (Publication)	6.863
Publication Ranking	38 (First rank available in 1999)
Impact Factor (Change)	+23.524

3.2 H-Index

An H-Index (HI) has the simple premise to quantify the scientific output of an individual researcher, or of a journal as a whole. It is defined as the index h where the number of papers with citation number ≤ h [6]. By retrieving the HI of the author, the researcher can assess the scientific credibility an author may have. The most recent value for the author is used, and given for both the first and last author, as the quality of the paper seldom relies on just a single researcher. The HI for the authors are taken from Scopus [4].

Similarly, the HI of a Journal as a whole can be calculated. Although it is not used very frequently, the HI of a journal is another way to assess the relevance of a journal amongst all journals retrieved by the search strategy. The HI of the Journals are taken from SJR [5] are only presented if the specific journal is found by full name. The fields returned for HI are shown in *Figure 3* and for the example paper in *Table 2*.



Current HIndex	Publication HIndex
62	3
42	33

H-Index 1st Author	H-Index Last Author
7	36
21	11
27	14

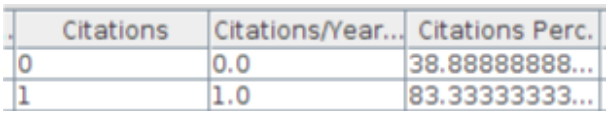
Figure 3: HI data

Table 2: Results pertaining to HI for example

Current h-Index	491
Publication h-Index	491 (first rank available in 1999)
h-Index 1st Author	94
h-Index Last Author	81

3.3 Citations

The overall number of citations that a paper has can be useful, but only when it is examined in the context with the other factors. The value retrieved for the total number of citations represents the citations from every available version of the article, as available on Google Scholar. Furthermore, the average number of citations per year is calculated, along with the percentile of the citations in the context of other papers examined in ALIGN. The number of citations is retrieved from Google Scholar [7]. The data regarding citations are shown in *Figure 4*, and for the example paper in *Table 3*.



Citations	Citations/Year...	Citations Perc.
0	0.0	38.88888888...
1	1.0	83.33333333...

Figure 4: Citation data

Table 3: Results pertaining to citations for example

Citations	1827
Citations/Year Average	91.35
Citations Percentile	*Irrelevant consider only a single paper.*

3.4 Keyword Trends

Keywords describing a research article can be used to quantify the novelty of the research done in a specific paper. Every keyword of a paper is plotted with the number of publications including that keyword from 1890 until now. The search for keywords can be refined to include only a subsection of the results, allowing the data that is returned to be closer to the intended research question, and thus more useful for a researcher. The data are retrieved from the MESH search on PubMed [8]. An example of a graph with all keywords of a paper is shown in *Figure 5*.

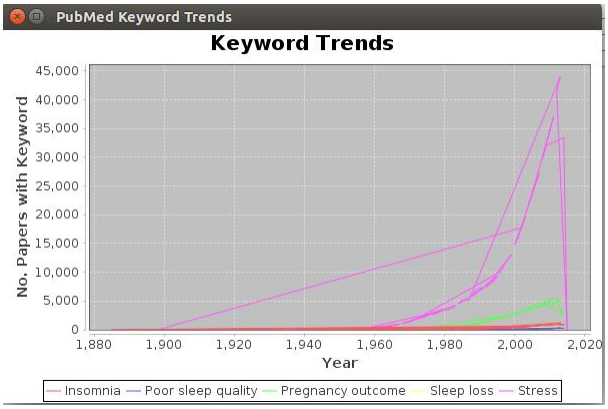


Figure 5: Keyword graph trends

4. DISCUSSION

The overall goal of ALIGN is to retrieve as much information as possible, that the user would not normally have easy access to. This is with the intention of providing universal access to higher quality research materials, resulting in higher quality research articles with minimal extra effort by the researcher. Currently, in the best case, by looking on just PubMed, the user can get a total of 6 pieces of information about the article at first glance: Title, Authors, Journal, Publication Date, Keywords, and Abstract. The values that PubMed provides are specific to the paper itself and do not show the article in the context of any other papers. While comparing papers for accuracy, applicability, and overall relevance, this basic information is not usually very helpful aside from the abstract. On the other hand, ALIGN gives the researcher access to 13 different values: IF at the time of publication and current, ranking at publication time and current, change in IF, HI at publication and current, HI of first and last author, keyword graph, total citations, average citations per year, and the citation percentile. In addition to what is currently available on PubMed, these 13 different values provide a more complete snapshot of the strength of the article. The process of using ALIGN adds only an extra 40-60 seconds of computing time per paper, giving researchers exceptionally rapid access to objective measurements of the paper.

4.1 Limitations

The use of ALIGN is limited by certain technological boundaries as well as the process by which academic articles are ranked.

Technologically, computational time and connectivity plays a large role in the usefulness and the overall accuracy of the program. Without internet connectivity, there are very few features that can be used. Although this may pose as a burden to users without direct internet access, this was done intentionally to provide the user with the most up to date information about papers with a specific keyword or author. However, this feeds into the problem of the computational requirements. Since, essentially, a lot of the data are being read in real time, many of the values take a few seconds to compute providing the user with a significant waiting time. This is further expanded by the need of certain websites to require institutional specific authentication in order to access certain data stores.

In the rare case that users do not have a consistent and considerable internet connection, 5Mbps and above, certain caching techniques can be used in order to not re-download data which were already acquired in the past. However, computational time may be reduced if a user is holding previously acquired information in a temporary store on their device.

Since the IF of an academic journal is only updated every 2 years, and for the most part is overseen by a single group, there is an inherent bias towards some papers. For example, older studies that have been cited less may end up with a higher IF than a new study that has been cited by several sources. Furthermore, since the ratings for journals are not updated until June of a given year, there is a large gap between the first and next IF of a new paper. Finally, by the sheer number of journals and papers now available today compared to a few decades ago, many papers are lost in the mass void of scholar knowledge

To fix inconsistencies with biases towards certain papers, the number of citations per year, as well as the percentiles in the IF can be used in order to give a more accurate representation of the overall quality of the paper. However, while the IF seems to be the most highly regarded ranking for a journal, few other changes can be made.

5. CONCLUSION

With the high turnover rate of researchers and articles published in the field, it is important to consider that some of the work in rating papers should be done through an automated process. With the overwhelming amount of article contribution from researchers all around the world, it is often not feasible for a researcher to look through the full initial results from a search strategy. Each year more than 1.3 million learned articles are published in peer reviewed journals [9]. As this number will only increase, the need for a program like ALIGN will continue to grow. Already today, more and more research is being focused on developing automated programs, such as IBM's Watson, which can process 500 gigabytes, the equivalent of a million books, per second. [10]

In the short term, many of the operations done on a user's machine can be extended to work on a server or cloud service. As such, by having access to more computing power, the user would not have to wait as long to reap the benefits of the extra information relating to each paper. Furthermore, in using a cloud service the user is able to submit a list of jobs, and while ALIGN is retrieving the results, they can work on other projects.

On a grand scale, with the proper implementation of automated methods and with the use of machine learning techniques, the whole process of the systematic review could be made significantly easier. Aside from developing the search strategy and writing the final manuscript, papers could eventually be ranked and chosen by the factors described above, among many others. In this sense, the total number of papers can significantly be reduced from potentially thousands at the start, to the final few dozen that are more relevant, and need to be read and summarized. For now, ALIGN is available to the modern researcher, able to incorporate the advances of today to enable the progress of tomorrow.

6. ACKNOWLEDGMENTS

The authors would like to thank Dr. Jérôme Waldispühl for his guidance regarding the programming portion of this project.

7. REFERENCES

[1] Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gotzsche PC, Ioannidis JP, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. Annals of internal medicine. 2009;151(4):W65-94.

[2] Benjamin EJ, Levy D, Vaziri SM, D'Agostino RB, Belanger AJ, Wolf PA. Independent risk factors for atrial fibrillation in a population-based cohort. The Framingham Heart Study. JAMA : the journal of the American Medical Association. 1994;271(11):840-4.

[3] Eugene G. The Agony and the Ecstasy—The History and Meaning of the Journal Impact Factor. Retrieved December. 2005;28:2006.

[4] Goodman D. Web of Science (2004 version) and Scopus. The Charleston Advisor. 2005;6(3):5-.

[5] Guerrero-Bote VP, Moya-Anegón F. A further step forward in measuring journals' scientific prestige: The SJR2 indicator. Journal of Informetrics. 2012;6(4):674-88.

[6] Bornmann L, Mutz R, Daniel HD. Are there better indices for evaluation purposes than the h index? A comparison of nine different variants of the h index using data from biomedicine. Journal of the American Society for Information Science and Technology. 2008;59(5):830-7.

[7] Jacsó P. Google Scholar: the pros and the cons. Online information review. 2005;29(2):208-14.

[8] Falagas ME, Pitsouni EI, Malietzis GA, Pappas G. Comparison of PubMed, Scopus, web of science, and Google scholar: strengths and weaknesses. The FASEB Journal. 2008;22(2):338-42.

[9] Bjork B-C, Roos A, Lauri M. Scientific journal publishing: yearly volume and open access availability. Information Research: An International Electronic Journal. 2009;14(1).

[10] Zhu W-DJ, Foyle B, Gagné D, Gupta V, Magdalen J, Mundi AS, et al. IBM Watson Content Analytics: Discovering Actionable Insight from Your Content: IBM Redbooks; 2014.

8. APPENDIX

8.1 Programming specifications

All code, including sample projects are available on GitHub: https://github.com/mkgorshkov/COMP396-SUMMER

All code was developed using native Oracle Java 1.7 libraries with the exception of those below, used will explicit permission or by complying with the respective licenses, and is intended to work on Linux, Windows, and Mac OS machines so long as the licensing in the next appendix is considered.

Non-standard libraries used: GSON 2.2.4, HtmlUnit 2.1.5, JCommon 1.0.23, JFreeChart 1.0.19, SQLite JDBC 3.7.2

8.2 Program licensing