CatchStudy: A Software for Users to Retrieve Highly Relevant Information and Select Suitable Journals for Submission of Manuscripts using Keywords to Query PubMed and Algorithm to Sort Abstracts and Journals

Catchstudy: A Software for Extracting Information and Selecting Suitable Journals for Submission of Manuscripts

Chou-Cheng Chen
Department of Business Administration
CTBC Business School
Tainan, Taiwan (ROC)

Abstract:- PubMed database stores more than 34 million citations and abstracts of biomedical literature published in approximately 30,000 Journals. "Best Match" function of the PubMed query website helps users sort selected match publications, but it does not provide suggested journals for users to submit manuscripts. A previous study showed Jane is a website tool with an elegant algorithm that provides journal suggestions and finds related articles for users; however, it merely finds journals that have been indexed from its work. This study thus creates CatchStudy, which is a windows form tool to help users currently extract information from PubMed and find suitable journals fitting the scope of the user. CatchStudy is based on PubstractHelper and adds a new approach to give a fitted score for each abstract, thereby sorting abstracts, and it can significantly help the user extract information from PubMed more quickly than "Best Match" of PubMed ($p < 0.05$). CatchStudy also provides suitable journals for user submission by calculating the sum of the fitted scores in abstracts that are published in the same journal. Users can also provide their title and abstract from their manuscript, and CatchStudy selects highly related nouns to query. CatchStudy uses these selected highly related nouns from using natural language processing (NLP) to token parts of speech and sorts by weight, and then queries PubMed for retrieving abstracts. Suitable journals are provided for the user by summing up fitted scores in abstracts. Comparison between Jane and CatchStudy shows that CatchStudy provides more journals for the user as reference to submit their manuscript. The software can be download from https://drive.google.com/drive/folders/1h65rIt8Udz2KJc3Ass4IE46BJ4E9Qa_t. To the best of my knowledge, CatchStudy provides another efficient method to extract information from querying PubMed and retrieving suitable journals for submission of the user's manuscript.

Keywords:- *Text Mining, Fitted Abstracts, Suitable Journals, Natural Language Processing.*

I. INTRODUCTION

More than 34 million citations and abstracts of biomedical literature and approximately 30,000 Journals are stored in the PubMed database [1, 2]. PubMed thus provides an efficient algorithm, "Best Match", to help users select publications that are sorted by best match from query [3]. The survey of their study also shows the usage of items including "Sort by Best Match", "Sort by TF–IDF (Term Frequency–Inverse Document Frequency)" and "Sort by date" on PubMed website were 39%, 36% and 32% respectively , and usage of "Sort by Best Match" are higher than other factors with statistical significance of 99% confidence level [3]. However, the study does not provide an algorithm for user to choose the most suitable journal for submission. Elsevier uses NLP techniques to deploy JournalFinder, which finds suitable journals with the best scope for submitting a user study via calculating keywords weight in the title and abstract of the manuscript [4]. WILEY, IEEE and Springer also deploy similar website tools such as "Journal Finder Beta", "IEEE Publication Recommender" and "Springer Nature Journal Suggester" which endeavor to find the most suitable journal for manuscript submission respectively [5-7]. Although the tools can help authors find a suitable journal for submission, they focus on their own published journals and do not show related abstracts with manuscripts. Jane is a useful tool with an elegant algorithm for suggesting journals and finding related articles by using the title or abstract from the user manuscript to query [8]. The tool uses a similar k-nearest neighbor approach to produce confidence scores to select suitable journals or related articles [8]. It has indexed 4,171,368 abstracts of articles from 4513 journals in Medline that have been published in the last 10 years [8]. However, the study just finds journals which have been indexed from its computer database [8]. This study thus creates CatchStudy, which is a tool to help the user extract current information from PubMed and find suitable journals for submission of their manuscript.

CatchStudy is based on our previous study, PubstractHelper, which is a website tool and can label sentences with co-occurring keywords in the same sentence via query of user keywords [9]. CatchStudy adds a new approach to provide a fitted score for each abstract and rank it, thereby significantly helping the user extract information from PubMed more quickly than by Best Match ($p < 0.05$). CatchStudy also provides suitable journals for manuscript submission by query of keywords. The suitable journal is selected from sorting by summing up fitted scores of each abstract in the same journal. Users can provide their title or abstract from the manuscript, and CatchStudy selects highly related nouns to query PubMed and retrieve fitted abstracts and suitable journals as the nouns from the article contain key information [10]. The nouns use NLP as token parts of speech, and the weight of each noun is calculated by algorithm [11]. The result shows our software can provide the user with more options of suitably determined journals for submission of their manuscript. Figure 1 shows a screenshot of CatchStudy, and CatchStudy can be downloaded from https://drive.google.com/drive/folders/1h65rIt8Udz2KJc3Ass4IE46BJ4E9Qa_t.

II. TOOL FEATURES

A. Query Method

CatchStudy uses C# to write in Windows form and implements E-utility to retrieve abstracts from PubMed query [12]. The default method of retrieving abstracts uses sort-by-publish date, but the user can select sort-by-best match by checking “Retrieve by Best Match” in the “Setting”. The user can select a maximum of ten groups to query, and the number of keywords that are separated with spaces in each group is not limited, making it possible for the user to also retrieve abstracts up to a maximum of 4,000 and even specifically select the period of publication date.

B. Fitted Score

Fitted score of each abstract is defined as:

$$Score = \frac{\mathcal{O}_a}{\mathcal{K}} + \frac{\mathcal{O}_t}{\mathcal{K}}\mathcal{E}_t + \sum_{i=1}^{n}\frac{\mathcal{O}_i}{\mathcal{K}}\left(1+\frac{i}{n}\right)\mathcal{E}_i$$

$\mathcal{K}$ is defined as the number of total keywords groups, $\mathcal{O}_a$ is defined as the number of keyword groups which occur in the title and abstract. If any keyword of the keyword groups occurs in the title and abstract, this keyword group is then identified as occurring in it. $\mathcal{O}_t$ is defined as the number of keyword groups in the title; put simply, the keyword group is identified as occurring in the title because one of the keywords in it occurs in the title. $\mathcal{E}_t$ is defined as two because all of the keyword groups co-occur in the title, while $\mathcal{E}_t$ is defined as 0.2 because not all of the keyword groups occur in the title. i is defined as the index of the sentence, and n is defined as the number of sentences in the abstract. $\mathcal{O}_i$ is defined as the number of keyword groups in i sentence. The keyword group is identified as occurring in i sentence because one of its keywords occurs in i sentence. $\mathcal{E}_i$ is defined as one because all of the keyword groups co-occur in i sentence, while $\mathcal{E}_i$ is defined as 0.1 because not all of the keyword group occurs in it. Users can choose whether abstracts are sorted by score or not by checking or unchecking “Sort by Score” in the “Setting”.

C. Suitable journals for submission

Suitable journals integrate citation score (CiteScore) from ELSEVIER and JCR categories from THOMSON REUTERS into the data set. Score of suitable journals is summed to a fitted score of each abstract published from the same journal, and suitable journals are sorted by score. The user can choose the best journal for submitting their manuscript because the data set contains suitable rank, citation score and JCR category.

D. Nouns selection from title and abstract

All nouns in the title and abstract are tagged by NLP program from Stanford Parser. The weight of each noun is defined as:

$$Weight = \mathcal{K}_t + \sum_{i=1}^{n}\mathcal{K}_i\frac{i}{n}$$

$\mathcal{K}_t$ is defined as one because this noun occurs in the title, while being defined as zero when not occurring in the title. i is defined as the index of the sentence in the abstract, and n is defined as the number of sentences in the abstract. $\mathcal{K}_i$ is defined as one because this noun occurs in i sentence, while being defined as zero when not occurring in i sentence. The default method of CatchStudy selects the top three nouns sorted by weight to query PubMed, while the user can select up to ten top nouns to query PubMed. Each noun use “AND” of the Boolean method to combine in the E-Utility, and the user can download all nouns with a score tag by using CatchStudy to extract nouns from their title and abstract of the manuscript.

III. COMPARISON WITH OTHER SOFTWARE

This study uses keywords of group one including “software*”, “website*” and “program*”, and of group two including “text*” and “literature*”, and of group three including “mining*”, to query CatchStudy and Best Match. The total number of retrieved abstracts is 1,995, with abstracts of the top 156 rank from CatchStudy being processed in a literature review because CatchStudy finds that all the keywords group co-occur in the title or any of the same sentence. Abstracts of the top-ranked 156 from Best Match of PubMed are also processed in the literature review in comparison with the data set of CatchStudy. The user can get information related to “software of text mining” from 150 abstracts by CatchStudy; from 142 abstracts by Best Match of PubMed; from two contexts of an article by CatchStudy; and from 12 contexts of an article by Best Match of PubMed. Four papers are retrieved by CatchStudy and two papers are retrieved by Best Match of PubMed that are not related to “software of text mining”. Pearson’s Chi-squared test is used to test data via R software (version 4.0.5) and the p value is 0.018. The Bonferroni method was also used for post hoc test via package of “chisq.posthoc.test”. The adjusted p-value is 0.037 using testing comparison between CatchStudy and Best Match of PubMed of getting information from abstracts.

Statistical results of Chi-squared and post hoc tests is shown in Table 1, showing CatchStudy can significantly help the user find key information conveniently. Supplementary Table 1 shows that all PMID (PubMed Unique Identifier) are sorted by CatchStudy and Best Match of PubMed respectively, and it also shows compared different rank differences from each other (https://drive.google.com/drive/folders/1h65rIt8Udz2KJc3Ass4IE46BJ4E9Qa_t). Supplementary Table 1 also shows whether the abstract or context can get key information via literature review. Supplementary Files 1 and 2 show fitted abstracts and suitable journals via CatchStudy execution respectively, and Supplementary File 3 shows abstracts via sorting from Best Match of PubMed. CatchStudy also provides a reference of suitable journals for the user to submit their paper by extracting nouns from title and abstract queries. This study thus compares favorably with the Jane website and shows the advantages of CatchStudy (https://drive.google.com/drive/folders/1h65rIt8Udz2KJc3Ass4IE46BJ4E9Qa_t).

TABLE 1. shows statistical results using Chi-squared and post hoc tests. Key information can be extracted from contexts that from CatchStudy retrieval are less than from Best Match retrieval. It shows that extract of key information from Abstracts using CatchStudy is more efficient than from Best Match.

	Retrieve relative information from Abstract	Retrieve relative information from Context	Without relative information
CatchStudy	150	2	4
Best match	142	12	2
p-value from post hoc test	0.39	0.037	1
p-value = 0.018			

This study used keywords, "(website OR software OR program) AND (text OR literature) AND (mining)" to query and retrieve thirty-three suitable journals from Jane software to compare with 640 suitable journals from CatchStudy retrieval. All retrieved journals from Jane occur in the data set from CatchStudy, and suitable journals from CatchStudy provide extra reference for the user such as "Journal of biomedical semantics", "Journal of the American Medical Informatics Association: JAMIA", "Scientific reports", "Computer methods and programs in biomedicine", etc. This study also uses title and abstract of PMID:35626536 to query and retrieve thirty-two journals from Jane and 308 suitable journals from CatchStudy respectively, and the results show that twenty-four of the retrieved journals from Jane occur in the data set from CatchStudy [13]. The published journal of the article ("Entropy (Basel, Switzerland)") both occur in the retrieved data sets from Jane and CatchStudy. Although some non-occurring journals in results from CatchStudy are suitable journals like "PeerJ. Computer science", "Neural Netw", "IEEE transactions on neural networks and learning systems" and "Medical & biological engineering & computing", CatchStudy does provide extra reference for the user like "BMC bioinformatics", "Bioinformatics (Oxford, England)", "Nucleic acids research", "Journal of biomedical informatics", "Journal of biomedical semantics", "PLoS computational biology", etc. Results of the comparison between Jane and CatchStudy are shown in Supplementary Table 2 (https://drive.google.com/drive/folders/1h65rIt8Udz2KJc3Ass4IE46BJ4E9Qa_t.).

IV. CONCLUSION

To our best knowledge, CatchStudy provides another efficient method to extract information from PubMed queries and retrieve suitable journals for submission of manuscripts by users.

REFERENCES

[1]. PubMed Overview. https://pubmed.ncbi.nlm.nih.gov/about/
[2]. List of All Journals Cited in PubMed. https://www.nlm.nih.gov/bsd/serfile_addedinfo.html
[3]. Fiorini N, Canese K, Starchenko G, Kireev E, Kim W, Miller V, et al. Best Match: New relevance search for PubMed. PLoS Biol. 2018;16:e2005343.
[4]. JournalFinder. https://journalfinder.elsevier.com/
[5]. Journal Finder Beta. https://journalfinder.wiley.com/search?type=match
[6]. IEEE Publication Recommender. https://publication-recommender.ieee.org/home
[7]. Springer Nature Journal Suggester. https://journalsuggester.springer.com/
[8]. Schuemie MJ, Kors JA. Jane: suggesting journals, finding experts. Bioinformatics. 2008;24:727-8.
[9]. Chen CC, Ho CL. PubstractHelper: A Web-based Text-Mining Tool for Marking Sentences in Abstracts from PubMed Using Multiple User-Defined Keywords. Bioinformation. 2014;10:708-10.
[10]. Chen J, Zhuge H. Automatic generation of related work through summarizing citations. Concurrency and Computation: Practice and Experience. 2019;31:e4261.
[11]. Finkel JR, Grenager T, Manning CD. Incorporating non-local information into information extraction systems by gibbs sampling. Proceedings of the 43rd annual meeting of the association for computational linguistics (ACL'05)2005. p. 363-70.
[12]. (US) BMNCfBI. Entrez Programming Utilities Help [Internet]. 2010-. https://www.ncbi.nlm.nih.gov/books/NBK25501/
[13]. Kalouptsoglou I, Siavvas M, Kehagias D, Chatzigeorgiou A, Ampatzoglou A. Examining the Capacity of Text Mining and Software Metrics in Vulnerability Prediction. Entropy (Basel). 2022;24.