KoExPubMed: A Tool for Effective and Customized Knowledge Extraction from PubMed

Kashfia Sailunaz[1] Gabi Jurca[1] Deniz Beştepe[2] Buşra Karatay[3] Lama Alhajj[4]
Tansel Özyer[5] Jon Rokne[1] Reda Alhajj [1,2,6]
[1]Department of Computer Science, University of Calgary, Alberta, Canada
[2]Department of Computer Engineering, Istanbul Medipol University, Istanbul, Turkey
[3] Yapı Kredi Teknoloji, R&D and Special Projects Department, Istanbul, Turkey
[4]International School of Medicine, Istanbul Medipol University, Istanbul, Turkey
[5]Department of Computer Engineering, Ankara Medipol University, Ankara, Turkey
[6]Department of Health Informatics, University of Southern Denmark, Odense, Denmark

***Abstract*—An exponential growth in the literature in general and the medical literature in particular raises a need for effective intelligent analysis strategies and tools to provide valuable insights to researchers about the current evolving literature. While existing applications provide more specific approaches to the problem, such as focusing on particular genome or protein information, in this paper, the proposed application provides effective and detailed analysis of PubMed. The developed tool, named KoExPubMed, follows a more generalized and holistic way by taking into consideration different types of information such as authors, countries, genes, and the interactions between them. The developed application consists of four main components; (1) keyword search and ID extraction, (2) PubMed article information and abstract retrieval, (3) country and address extraction, and (4) gene information extraction. In addition to the fundamental components, the tool provides a variety of visualization options for showing the extracted information and the related associations, including line charts for densities and countries, chord charts for collaborations of authors, network graphs for the genes mentioned together, bubble charts for gene frequencies, etc. By addressing the need for a generalized data mining tool, we propose a comprehensive application which is capable of employing data mining and machine learning techniques to extract from PubMed knowledge valuable to researchers and practitioners who are interested in closely investigating the achievements of others.**

***Index Terms*—literature analysis, knowledge extraction, gene interaction, data mining, data visualization**

1. Introduction

The development in technology and techniques for intelligent data analysis since the last decade of the previous century has facilitated effective knowledge discovery in data. Previously, data repositories were merely used for data storage and retrieval with limited processing beyond that. Thanks to the emergence of various data mining techniques and the developed of Application Programming Interfaces (APIs) for retrieving content from existing open access data repositories which have greatly influenced the way researchers approach the investigated problems. When combined with Natural Language Processing (NLP), it leads to detect and extract more relevant and domain specific knowledge for more rich and informative decision making based on a wider coverage and perspective as compared to the limited perspective which dominated in the previous era when the process was possible only manually at a limited scale. One of the most important and popular sources of biomedical and healthcare research is PubMed [1], which is an open access search engine to access the MEDLINE database [12] for publications on biomedical topics. Accordingly, it is the focus for the data extraction process accomplished by the tool described in this paper.

Due to the enormous amount of available articles in PubMed and the wide range of themes covered by these articles, PubMed may be named as the most valuable resource for researchers, mostly in multidisciplinary, interdisciplinary and transdisciplinary research efforts which involved some biomedical or healthcare flavor. Realizing this, several researchers from computer science have been working on developing automated knowledge extraction tools to organize and specify the PubMed searches in a more customized manner for domain specific researchers and practitioner. As the process mainly involves mining knowledge from publications, there are several tools named PubMiner with different features and functionalities. For instanve, Eom et. al [9] proposed a text mining system named PubMiner that used NLP on PubMed by training the model on Genome databases and protein datasets for gene information mining. Another tool named PubMiner [10] was proposed to extract tables and demographic sentences from PubMed data by employing feature engineering and sentence mining. Recently, Botsis et. al [5] proposed another tool named PubMiner which can extract information like cancer type, needed therapy, and cancer genomic information using NLP.

pubmed.mineR [8] is a R package built on the idea of PubMed mining for the relevance between terms with pattern mining by creating association matrices for existing terms and new terms to show their connections. BioReader [6] is another PubMed based text mining tool that classifies the extracted data into positive and negative classes based on the their relevance to the topic. DEBBIE [13] is a PubMed based open access text mining system that provides specified information on biomaterial objects like implants, cell scaffolds, etc. It uses the pipeline for relevant concept recognition. GenCLiP 3 [14] is another recently proposed web server that can mine data from PubMed for gene network searching and analysis by extracting information from publication abstracts collected from PubMed. As seen in the literature, most existing mining applications working on PubMed focused on specific genome information, or other particular issues rather than providing the research overview on a particular topic and the associated researchers [7], [11].

In this paper, we propose a novel application tool called KoExPubMed for effective and customized knowledge extraction from the PubMed repository. KoExPubMed has a sophisticated user interface which allows users to properly visualize the results. It applies a variety of machine learning techniques for comprehensive analysis of PubMed content to extract and convey different types of discoveries related to publications, authors, genes, countries, etc.

2. KoExPubMed: Knowlewdge Extraction from PubMed

KoExPubMed has been developed as a standalone application which is capable of extracting from PubMed data relevant for any project within a domain covered by the articles available in repository. The user is given the opportunity to specify certain parameters which will guide the retrieval process and will help in narrowing down the focus to the benefit of the user. This way, the outcome from KoExPubMed will be concise and descriptive.

KoExPubMed is a native Windows application (.Net) written in C#, developed based on Visual Studio, and uses a MySql database at the backend to keep all the necessary data in a relational database. The main characteristics and features of the developed KoExPubMed tool are discussed in the following subsections.

2.1. Summary of the novel characteristics

The KoExPubMed tool described in this paper contains some novel characteristics compared to the counterparts the relevant researches on data mining from PubMed data. The existing PubMed based data mining tools described in the literature are mostly personalized to serve a narrow community, i.e., they are focused on a particular gene based information extraction, and apply limited set of visualization techniques to convey the results. Some other works concentrated on specifying the relevance of the abstracts extracted from PubMed for a specific biomedical and/or biomaterial topic.

The KoExPubMed tool proposed in the paper has some distinguishing novel features mainly focusing on (1) managing keywords by allowing the user to specify the scope for the target project without limiting the spectrum as it is the case with the other projects described in the literature, (2) extracting from PubMed abstracts relevant to the project to be investigated, (3) extracting other relevant information for the downloaded abstracts, (4) extracting from the downloaded text gene information and the connections between different genes, and (5) extracting countries contributing to the research topic by analyzing author affiliations and generating country based contributions to the research topic.

2.2. List of functions and features

The functions and features of the developed KoExPubMed tool are described below.

Home Page: As shown in Figure 1, the homepage that appears after running the application has some basic options like creating a new project, opening an existing project, saving projects, generating figures, accessing various options in settings, and closing the application. The home page also provides the following four major functions of the application-

- Keywords - managing keywords of the project and extracting unique IDs of abstracts from relevant publications,
- PubMed - extracting all information from PubMed by considering the unique IDs related to the keywords from the previous step,
- Becas - extracting gene information from the abstracts using Becas and UniPort, and
- Google Maps - extracting countries of the authors of the publications retrieved from PubMed.

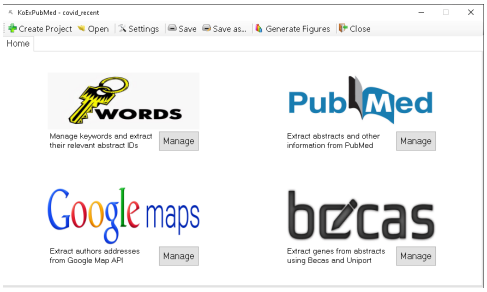

Figure 1. The Main Page of KoExPubMed

Keywords: The user can either create a new project or open an existing project, and then the user may choose the keywords option. When a keyword is added, the system queries PubMed and gets related IDs for the abstracts containing the keywords, show the count of the keywords and the IDs of the abstracts. The user can also add and delete keywords. Furthermore, the user can view the abstract IDs.

A sample example related to a Covid-19 project is shown in Figure 2.

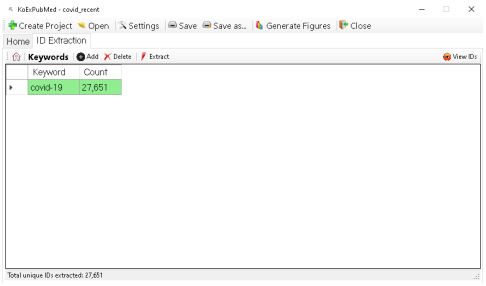

Figure 2. The Keywords option of KoExPubMed

PubMed: As shown in Figure 3, this option is used to extract from PubMed all relevant information related to the investigated topic. It will process the abstract IDs obtained from the keyword phase. For the queries on the IDs, it shows the progress with processed IDs and the processed queries. It will also give the user a complete error panel with the categories- publication title, abstract, journal, date, keywords and authors.

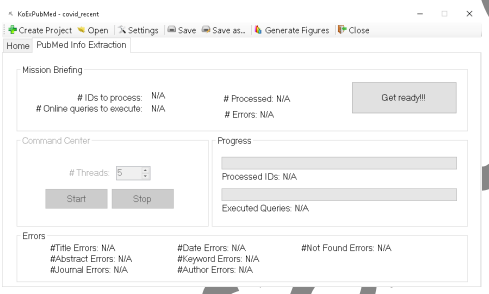

Figure 3. The PubMed option of KoExPubMed

Becas: Figure 4 shows the functionalities of the Becas option.Selecting Becas will lead to extracting the genes from the abstracts extracted using two sources, namely Becas [3] and UniPort [4] APIs. The user can set the contact information. The system will show the progress with articles and threads. It will also show the summary of the query execution with the processed and pending articles.

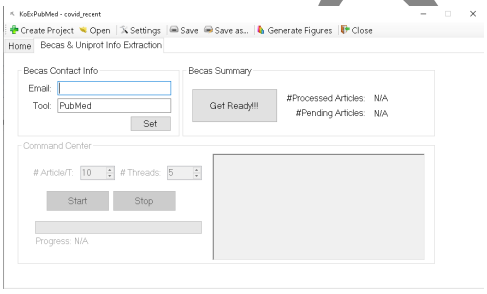

Figure 4. The Becas option of KoExPubMed

Google Maps: KoExPubMed also offers the extractions of the authors' countries. As shown in Figure 5, Google Cloud Maps Platform Geocoding API [2] is used for getting address information from affiliations and for extracting countries from addresses of the author. The system will find the affiliations associated with the extracted publications, then using the API it will extract the countries of the authors. It will show the progress of the query execution, valid and invalid addresses, and the number of records processed at each round. The user can also export and import the address cache.

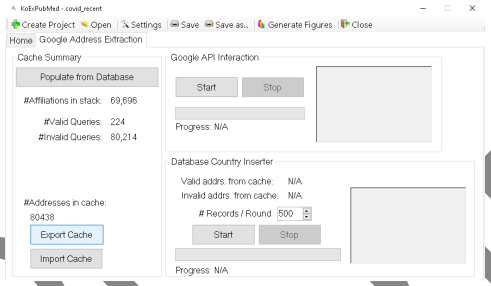

Figure 5. The Google Maps option of KoExPubMed

Visualization: The application also provides various user-friendly visualization options of the results. The visualizations are based on different characteristics of the search results. There are some more generalized results like the publication densities and countries as shown in Figure 6 and Figure 7. The number of published articles in each year is shown in a line chart. The user can select if the results will be cumulative or not, if the latest year will be included or not, if all articles will be counted or only the ones that mention the genes, etc. The user can save the results with a name and add/delete data files to include/exclude them from the results.

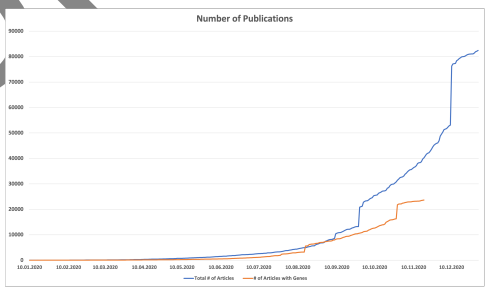

Figure 6. The Visualization - Yearly Articles component of KoExPubMed

The country visualization results show the top countries based on the authors affiliations. The user can choose all the articles or only the articles that mention genes and the number of top countries they want as the output; the system will show them the list of countries based on the number of articles published using the country of the authors. Some articles can be counted in multiple countries if the research reflects collaboration between authors who are affiliated with different countries.

The user can also view the collaborations and affiliations of authors, countries, etc. Figure 8 shows the bar chart visu-

Country	Number of Articles	Number of Articles with Genes	% of Articles with Genes
United States	31354	8648	27,58
China	10082	3811	37,80
United Kingdom	9054	1955	21,59
Italy	8236	2453	29,78
India	4994	1588	31,80
Canada	3649	938	25,71
France	3397	1010	29,73
Germany	3273	1214	37,09
Spain	3042	989	32,51
Australia	2816	638	22,66
Brazil	2362	713	30,19
Iran	1893	622	32,86
Switzerland	1660	459	27,65
Japan	1518	504	33,20
Netherlands	1511	438	28,99
Turkey	1468	428	29,16
Mexico	1290	370	28,68
Singapore	1259	247	19,62
Saudi Arabia	1168	409	35,02
South Korea	1137	399	35,09

Figure 7. Visualization - Top Countries.

alization of the collaborations of the authors. The results are extracted based on the collaborations between the starting year and final year entered by the user. The bar chart shows the affiliations between authors from different countries for the specified period. The supporting data table can also show details on the year of publication, all authors, all articles, the processed articles, affiliations and authors, etc. A similar idea with chord chart on affiliations is shown in Figure 9; it provides the strength of the collaborations to intensify the collaboration outputs. The chord chart shows the density of publication collaborations between different countries. The chord color denotes the width of the chords. The darker red it is, the wider the chord. The chord widths are based on the number of abstracts between two countries of the affiliations. The chord shows if there is at least one authors who has affiliations with both countries at the endpoints, or at least one author of the article has affiliation of one country and another author of that publication has affiliation with the second country.

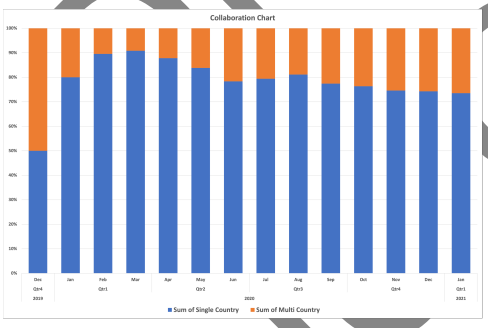

Figure 8. Visualization - Collaboration.

The gene data has the visualization option for showing the genes that appear together most frequently. The graph will show which genes are frequently mentioned together in the publications. The user can also generate the year and country wise results. Figure 10 shows a sample output for frequent genes for the Covid-19 project. Figure 11 also shows another gene based visualization result. The system can create a network graph for the genes that are mentioned together in the same abstract. The gene nodes mentioned

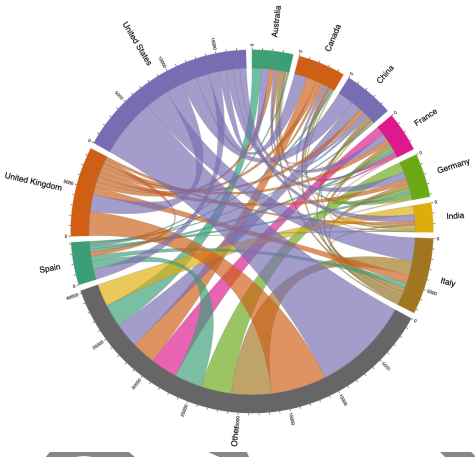

Figure 9. Visualization - Collaboration.

in the same abstract will be connected to each other and the size of the node will be equivalent to the number of mentions. A bubble chart can also be generated for the most frequent genes mentioned in abstracts of articles and the author affiliations as shown in Figure 12. The user can choose to view the overall results or results for each country separately. Bubble color will be based on the country, which means the genes mentioned in the abstracts affiliated with the same countries will have the same bubble color and the bubble label will mention the name of the gene. The size of the bubble will be based on the number of articles mentioning that gene. The user can also select how many top genes they want to include and how many top countries they want to add to the results.

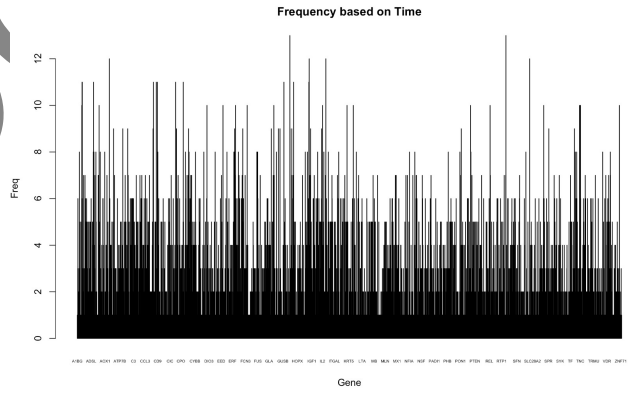

Figure 10. Visualization - Frequency based on time period.

Settings: The settings option provides a connection string builder as shown in Figure 13. The user can update the project name; create a connection string by providing host name, username, database name and password; test the connection; install different configurations; and save the connection.

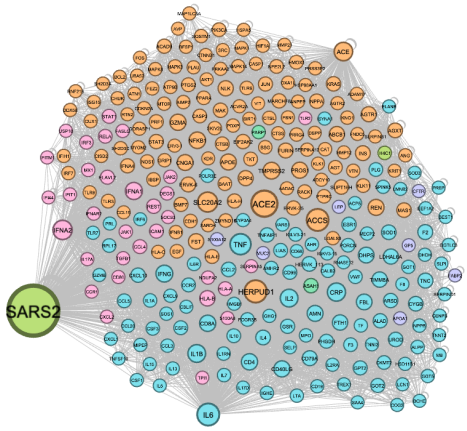

Figure 11. Visualization - Network Analysis.

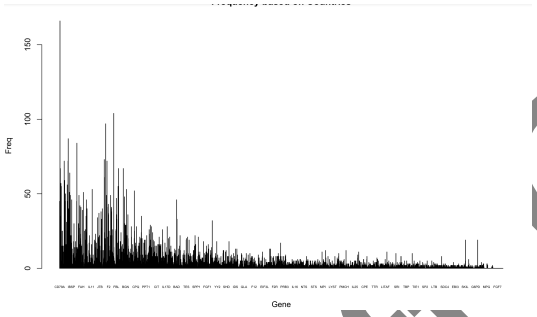

Figure 12. Visualization - frequency based on country.

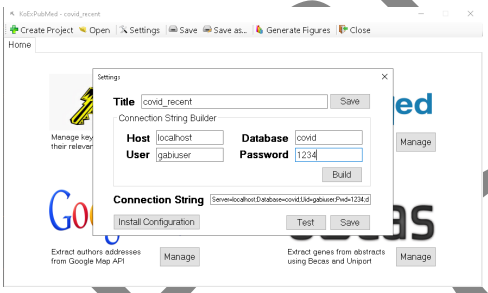

Figure 13. The Settings component of KoExPubMed

3. Conclusion

KoExPubMed has been developed to fill a gap which necessitated the need for a tool capable of automating the processing of literature analysis based on certain prespecified keywords for effective knowledge discovery which cannot be accomplished by following a traditional process. The simple user-interface of KoExPubMed is expected to increase its usage at a wide range from the research community. This is supported by the comprehensive set of visual results which could be retrieved by using KoExPubMed. Such results will guide and increase the confidence of researchers in the outcome/

References

[1] National Library of Medicine, *PubMed*. URL https://pubmed.ncbi.nlm.nih.gov/, 2023.

[2] Google, *Geocoding API*. URL https://developers.google.com/maps/documentation/geocoding/overview, 2023.

[3] Becas, *becas API*. URL https://bioinformatics.ua.pt/becas/#!/api, 2023.

[4] UniPort, *Programmatic access*. URL https://www.uniprot.org/help/programmatic_access, 2023.

[5] T. Botsis, J. Murray, L. E. Alessandro, D. Palsgrove, W. A. Wei, J. R. White, V. E. Velculescu, and V. Anagnostou, *Natural language processing approaches for retrieval of clinically relevant genomic information in cancer*, Studies in health technology and informatics, 295, p.350, 2022.

[6] C. Simon, K. Davidsen, C. Hansen, E. Seymour, M. B. Barnkob, and L. R. Olsen, *BioReader: a text mining tool for performing classification of biomedical literature*, BMC bioinformatics, 19, pp.165-170, 2019.

[7] Z. Lu, *PubMed and beyond: a survey of web tools for searching biomedical literature*, Database, p.baq036, 2011.

[8] J. Rani, A. R. Shah, and S. Ramachandran, *pubmed.mineR: An R package with text-mining algorithms to analyse PubMed abstracts*, Journal of biosciences, 40, pp.671-682, 2015.

[9] J. H. Eom, and B. T. Zhang, *PubMiner: machine learning-based text mining for biomedical information analysis*, Genomics & Informatics, 2(2), pp.99-106, 2004.

[10] J. Bockskopf, Y. uning Chen, D. avid Dowey, B. Gao, A. Garza, and I. Smith, *PubMiner: An Interactive Tool for Demographic-enriched PubMed Searches*.

[11] Y. Han, S. A. Wennersten, and M. P. Lam, *Working the literature harder: what can text mining and bibliometric analysis reveal?*, Expert review of proteomics, 16(11-12), pp.871-873, 2019.

[12] T. Greenhalgh, *How to read a paper: the Medline database*, Bmj, 315(7101), pp.180-183, 1997.

[13] J. O. Corvi, A. McKitrick, J. M. Fernández, C. V. Fuenteslópez, J. L. Gelpí, M. P. Ginebra, S. Capella-Gutierrez, and O. Hakimi, *DEBBIE: The Open Access Database of Experimental Scaffolds and Biomaterials Built Using an Automated Text Mining Pipeline*, Advanced Healthcare Materials, p.2300150, 2023.

[14] J. H. Wang, L. F. Zhao, H. F. Wang, Y. T. Wen, K. K. Jiang, X. M. Mao, Z. Y. Zhou, K. T. Yao, Q. S. Geng, D. Guo, and Z. X. Huang, *GenCLiP 3: mining human genes' functions and regulatory networks from PubMed based on co-occurrences and natural language processing*, 2020.