CovEx: A User-Centric Recommender System for COVID-19 Scientific Literature

Behnam Rahdari[1], Peter Brusilovsky[1], Khushboo Thaker[1], Hung Kim Chau[1], Daqing He[1], Young Ji Lee[2]

[1]School of Computing and Information, [2]School of Nursing, University of Pittsburgh

ABSTRACT

CovEx system has been deployed online and demonstrated to several target users. The early results indicate that the success of the system to a considerable extent depends on the quality of key phrase extraction. Moreover, the nature of exploratory search calls for special extraction approaches. While we used a relatively powerful approach, it was trained to model gold standard annotation of individual documents in GENIA dataset. We believe, however, that key phrase extraction has to consider the collection as a whole increasing user chances to discover key phrases that could lead to other papers. We are converting CovEx to support clinicians' exploration of Ovarian cancer related literature to further examination of usages of CovEx.

INTRODUCTION

Exploratory search systems form an increasingly popular category of information access and exploration tools. These systems creatively combined search, browsing, and information analysis steps shifting user efforts from recall (formulating a query) to recognition (i.e., selecting a link) and helping them to gradually learn more about the explored domain [1]. In this poster we presenting our attempt to augment the set of search systems focused on COVID-19 research literature [2] with a personalized exploratory search system COVID Explorer - CovEx We hope that CovEx ability to support information discovery, learning-while-searching, and personalization, the system could help a broader set of users to benefit from the assembled collection of COVID-19 resources [3].

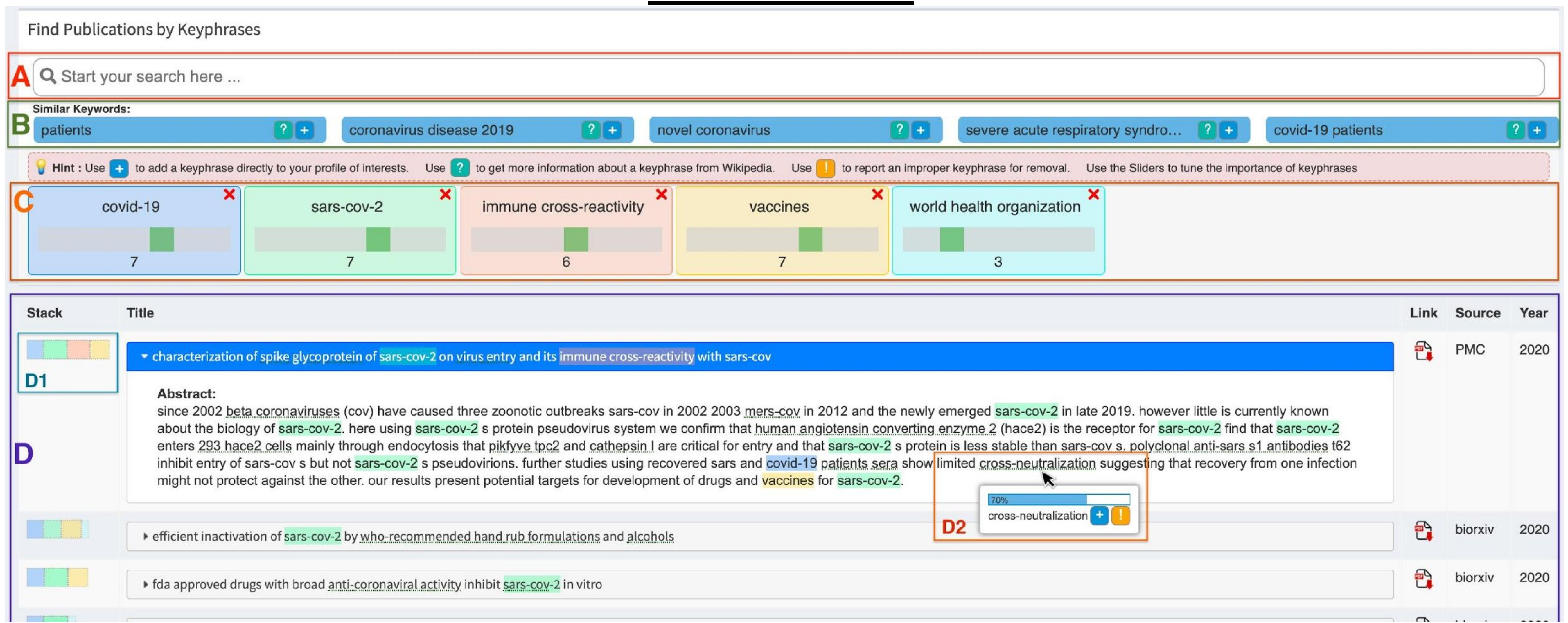
PROFILE-BASED SEARCH

We deploy a two-phase search process to produce the most relevant results based on user interest profile.

Candidate selection: We used the Cypher Querying Language to generate the initial list of candidate publications. At each instance of user interaction with the system (e.g., adding/removing key-words or tuning the sliders), the system considers all publications connected to at least one of the topics of interest in the user profile.

$$RelevanceScore_{(f,A)} = \sum_{i=0}^{|A|} Sim_{(a_i,f)} * w_i$$

INTERFACE DESIGN



Find Publications by Keyphrases

A Start your search here ...

B Similar Keywords:

patients ? + | coronavirus disease 2019 ? + | novel coronavirus ? + | severe acute respiratory syndro... ? + | covid-19 patients ? +

Hint : Use + to add a keyphrase directly to your profile of interests. Use ? to get more information about a keyphrase from Wikipedia. Use ! to report an improper keyphrase for removal. Use the Sliders to tune the importance of keyphrases

C

covid-19	sars-cov-2	immune cross-reactivity	vaccines	world health organization
7	7	6	7	3

D

Stack	Title	Link	Source	Year
D1	▾ characterization of spike glycoprotein of sars-cov-2 on virus entry and its immune cross-reactivity with sars-cov		PMC	2020
	▸ efficient inactivation of sars-cov-2 by who-recommended hand rub formulations and alcohols		biorxiv	2020
	▸ fda approved drugs with broad anti-coronaviral activity inhibit sars-cov-2 in vitro		biorxiv	2020

Abstract:
since 2002 beta coronaviruses (cov) have caused three zoonotic outbreaks sars-cov in 2002 2003 mers-cov in 2012 and the newly emerged sars-cov-2 in late 2019. however little is currently known about the biology of sars-cov-2. here using sars-cov-2 s protein pseudovirus system we confirm that human angiotensin converting enzyme 2 (hace2) is the receptor for sars-cov-2 find that sars-cov-2 enters 293 hace2 cells mainly through endocytosis that pikfyve tpc2 and cathepsin l are critical for entry and that sars-cov-2 s protein is less stable than sars-cov s. polyclonal anti-sars s1 antibodies t62 inhibit entry of sars-cov s but not sars-cov-2 s pseudovirions. further studies using recovered sars and covid-19 patients sera show limited cross-neutralization suggesting that recovery from one infection might not protect against the other. our results present potential targets for development of drugs and vaccines for sars-cov-2.

D2 70% cross-neutralization + !

Figure 1 : Interface Design of CovEx representing different parts of the system.

Search Box: The search box (Figure 1-A) is the gateway to the system. Using an instant search approach, it allows users to discover relevant topics without a fully formulated query. When a user starts typing a query, a series of frequent similar keywords appears, which helps the user to discover a range of matching topics (e.g., cell culture and infected cells).When an item is selected from the list, it will automatically added to the slider area (Figure 1-B). at the same time, an updated list of search results will be presented to the user.

Similar Keywords: When at least one keyword is added to the user's profile, a series of five semantically similar topics appear in the *Similar Keywords* area of the interface (Figure 1-B). Users can add recommended keywords to their interest profiles by clicking on the plus button to the right of each keyword. As the user's profile grows and refines, the set of recommended keywords is updated since the system recommends instances similar to all keywords in the user's profile.

Slider Area: The slider area (Figure 1-C) displays the current interest profile of the user. CovEx implements a content-based recommendation approach, which generates the list of recommended results (Figure 1-D) using the interest profile. To support transparency and controllability of this process, the interest profile is visible and directly editable by the end users.

Search Results: As soon as the user adds the first keyword to the interest profile, a table of the 20 most relevant publications is generated (Figure 1- D). The first column of the table visualizes the combined relevance between key phrases in the user interest profile and each result. The colors in the stacked-bar (Figure 1-D1) are matched with the color of slider in the profile and the size and opacity of each bar expresses the relevance of the result to each profile key phrase.

KNOWLEDGE GRAPH

The knowledge graph consists of three main entities - publications, authors, key phrases and their relationships - extracted from our data set and hosted in a native graph database Neo4j. Figure 2 presents the schematic representation of the knowledge graph. Authors are interconnected by the relation *Co-Author* (based on co-authorship) and connected to papers by the relation *Published*. Papers connected to key phrases using the *Has-Key* relationship. The latter carries a weight that determines the strength of the relationship between each key phrase and the publication.

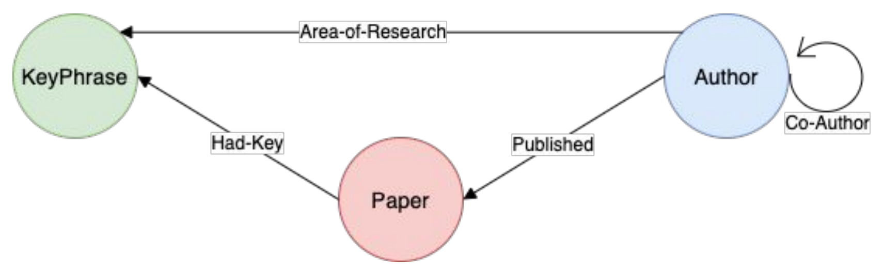


Figure 2 : Knowledge Graph Schema

Reordering the results: After generating the list of candidate results, the system rearranges the results in a way that the most relevant results appear at the top of the list. In order to do that, first a complete list of key phrases that appear in the text (title and abstract) of each publication, alongside with their relevance score (weight) is being generated. Then for every key phrase that exist in the user interest profile, we multiplied its weight with the value of corresponding slider. Finally, the relevance score is assigned to each candidate considering candidate's similarity to each of profile topics and the value of the sliders (Equation above)

DATA SOURCE AND GRAPH STATISTICS

We used COVID-19 Open Research Dataset Challenge (CORD-19) as the main source of data to build the knowledge graph and extract the key phrases. The dataset contains 51078 document, out of which 48251 documents contain either title or abstract.

Labels	No. Nodes	Avg. Properties	Avg. Relations
[Keyword]	211862	3	11.02
[Paper]	48251	12	12.91
[Author]	157589	1	3.65

We approach the key phrase extraction problem as a sequence labeling task. We apply a Bi-LSTM-CRF architecture to perform this task, which has been shown to achieve the best performance across several public datasets. To assign weight for each key phrase extracted from the the document we found the distance of the key phrase from the document in embedding space.

EXPERIENCE AND FUTURE WORKS

CovEx system has been deployed online and demonstrated to several target users. The early results indicate that the success of the system to a considerable extent depends on the quality of key phrase extraction. Moreover, the nature of exploratory search calls for special extraction approaches. While we used a relatively powerful approach, it was trained to model gold standard annotation of individual documents in GENIA dataset. We believe, however, that key phrase extraction has to consider the collection as a whole increasing user chances to discover key phrases that could lead to other papers. We are interested to collaborate with experts on key phrase extraction to develop approaches optimized for exploratory search.

Acknowledgment: This work was supported by the National Library of Medicine (5R01LM013038, PI: Lee, Daqing).